



Χαροκόπειο Πανεπιστήμιο
Τμήμα Πληροφορικής και Τηλεματικής

Εξόρυξη γνώσης από δεδομένα διατροφικών συνηθειών και τρόπου ζωής

Μαρίνα Πότση 21226

Πτυχιακή εργασία
προπτυχιακού επιπέδου

Αθήνα, 30 Σεπτεμβρίου 2016

Σύνοψη

Πτυχιακή εργασία

Εξόρυξη γνώσης από δεδομένα διατροφικών συνηθειών και
τρόπου ζωής

Μαρίνα Πότση 21226

Επιβλέπων καθηγητής:

Ηρακλής Βαρλάμης

Μέλη τριμελούς επιτροπής :

Γιώργος Δημητρακόπουλος

Κωνσταντίνος Τσερπές

Ευχαριστίες

Θα ήθελα να εκφράσω τις ευχαριστίες μου στον επιβλέποντα καθηγητή μου κ. Ηρακλή Βαρλάμη για τις πολύτιμες συμβουλές που μου παρείχε καθ'όλη τη διάρκεια της πτυχιακής εργασίας και γενικότερα των σπουδών μου, καθώς και για την άμεση καθοδήγησή του επί του θέματος.

Επίσης, θα ήθελα να ευχαριστήσω την κ. Γιαννακούλια Μαρία, Αναπληρώτρια Καθηγήτρια του τμήματος Επιστήμης Διαιτολογίας-Διατροφής, για την παραχώρηση των δεδομένων που χρησιμοποιήθηκαν στην πτυχιακή εργασία.

Τέλος, ένα μεγάλο ευχαριστώ στους γονείς μου και την αδερφή μου, για την άμεση υποστήριξή τους και την εμπιστοσύνη που μου έδειξαν όλα αυτά τα χρόνια των σπουδών μου.

Περιεχόμενα

Περιεχόμενα	i
Κατάλογος σχημάτων	iii
Κατάλογος πινάκων	iv
Περίληψη	v
Abstract	vi
1 Εισαγωγή	1
2 Θεωρητικό υπόβαθρο	3
2.1 Σχετικές ερευνητικές εργασίες	3
2.2 Επεξήγηση εννοιών	5
2.2.1 Εξόρυξη γνώσης	5
2.2.2 Μηχανική μάθηση	5
2.2.3 Κατηγοριοποίηση	7
2.2.4 Συσταδοποίηση	8
2.2.5 Ανάλυση πρωταρχικών συνιστωσών (Principal Component Analysis)	8
2.2.6 Μηχανές διανυσμάτων υποστήριξης (Support vector machines)	9
2.3 Εργαλεία	10
2.3.1 Weka	10
2.3.2 Βιβλιοθήκες Python	11
2.4 Μέτρα αξιολόγησης	12
2.4.1 Μετρικές	12
2.4.2 Πίνακας σύγχυσης (Confusion matrix)	14
2.4.3 Καμπύλη ROC (Receiver Operating Characteristic)	14

3	Ανάλυση δεδομένων	17
3.1	Περιγραφή δεδομένων	17
3.2	Προεπεξεργασία δεδομένων	18
3.2.1	Καθαρισμός δεδομένων	18
3.2.2	Μετασχηματισμός δεδομένων	20
3.3	Ασύμμετρο σύνολο δεδομένων (Imbalanced dataset)	21
3.4	Στρωματοποιημένη δειγματοληψία (StratifiedKFold)	23
3.5	Διατροφικοί δείκτες	23
3.5.1	BMI	24
3.5.2	DETERMINE	24
4	Υλοποίηση και αποτελέσματα	26
4.1	Συσχέτιση δεικτών με το BMI	26
4.1.1	Συσχέτιση Pearson	26
4.1.2	InfoGainAttributeEval	31
4.2	PCA ανάλυση με βάση το BMI	33
4.2.1	Περιγραφή της διαδικασίας	33
4.2.2	Εύρεση ιδανικού αριθμού συνιστωσών για την PCA ανάλυση	34
4.2.3	Οπτικοποίηση της απόδοσης	37
4.3	Συσχέτιση χαρακτηριστικών με τον δείκτη Determine	40
4.3.1	CfsSubsetEval	40
4.3.2	InfoGainAttributeEval	42
4.4	PCA ανάλυση με βάση το Determine	45
4.5	Συσχέτιση χαρακτηριστικών με την άνοια	48
5	Συμπεράσματα	51
	Βιβλιογραφία	55

Κατάλογος σχημάτων

2.1	Πίνακας σύγχυσης για δυαδική κατηγοριοποίηση	14
2.2	Παράδειγμα ROC καμπύλης	16
4.1	Γραφική παράσταση δύο χαρακτηριστικών με θετική συσχέτιση, αρνητική συσχέτιση και καμία συσχέτιση	27
4.2	Ιστόγραμμα του δείκτη BMI	34
4.3	Scree plot για την εύρεση του ιδανικού αριθμού συνιστωσών	36
4.4	Διάγραμμα απεικόνισης των μετρικών για κάθε κατηγοριοποιητή, όταν η κλάση είναι το BMI	37
4.5	Πίνακας σύγχυσης για τον αλγόριθμο SVM	39
4.6	Καμπύλη ROC για τον αλγόριθμο SVM	40
4.7	Ιστόγραμμα του δείκτη Determine	46
4.8	Scree test για τον υπολογισμό του ιδανικού αριθμού συνιστωσών	47
4.9	Διάγραμμα απεικόνισης των μετρικών για κάθε κατηγοριοποιητή, όταν η κλάση είναι το Determine	48

Κατάλογος πινάκων

4.1	Αποτελέσματα συσχέτισης Pearson των χαρακτηριστικών με το BMI . . .	29
4.2	Αποτελέσματα μεθόδου InfoGainAttributeEval μεταξύ των χαρακτηρι- στικών και του BMI	32
4.3	Αποτελέσματα μετρικών για κάθε κατηγοριοποιητή, έχοντας ως κλάση το BMI	37
4.4	Υποσύνολο των δεδομένων που προέκυψε από το CfsSubsetEval. Τα χα- ρακτηριστικά συσχετίζονται περισσότερο με το Determine και σχεδόν κα- θόλου μεταξύ τους	42
4.5	Αποτελέσματα της κλάσης InfogainAttributeEval για τη συσχέτιση των χαρακτηριστικών με τον δείκτη υποσιτισμού	44
4.6	Αποτελέσματα μετρικών για κάθε κατηγοριοποιητή, έχοντας ως κλάση το Determine	47
4.7	Αποτελέσματα μεθόδου InfoGainAttributeEval μεταξύ των χαρακτηρι- στικών και της άνοιας	50

Περίληψη

Το θέμα που μελετήθηκε στην παρούσα πτυχιακή είναι η εξόρυξη γνώσης από δεδομένα διατροφικών συνηθειών και τρόπου ζωής. Τα δεδομένα, τα οποία είχαν συλλεχθεί από ερωτηματολόγια, δόθηκαν από την κ. Γιαννακούλια Μαρία, μέλος της ερευνητικής ομάδας HELIAD. Τα χαρακτηριστικά του συνόλου δεδομένων περιέχουν συνήθειες που αφορούν τη διατροφή των ανθρώπων άνω των 65 ετών, καθώς και δύο διατροφικούς δείκτες: το BMI (Δείκτη Μάζας Σώματος) και έναν δείκτη για τον κίνδυνο του υποσιτισμού (Determine).

Στόχος της πτυχιακής είναι η χρήση εργαλείων της εξόρυξης δεδομένων και της μηχανικής μάθησης έτσι ώστε να ανακαλυφθούν πρότυπα για την εξαγωγή χρήσιμων συμπερασμάτων για τα δεδομένα. Έτσι προσδιορίζονται οι συσχετίσεις που υπάρχουν μεταξύ των χαρακτηριστικών και καθενός από τους δείκτες. Αυτό μας βοηθάει να καταλάβουμε ποια χαρακτηριστικά του συνόλου δεδομένων περιέχουν την μέγιστη πληροφορία όσον αφορά τους δείκτες, και τι διατροφικές συνήθειες έχουν οι άνθρωποι που βρίσκονται σε κίνδυνο για υποσιτισμό. Επιπλέον, γίνεται μια προσπάθεια συσχέτισης των διατροφικών συνηθειών και δεικτών με παθήσεις που εμφανίζονται κυρίως στους ηλικιωμένους, όπως για παράδειγμα η άνοια.

Το σύνολο δεδομένων έχει πολλές διαστάσεις, επομένως χρησιμοποιήθηκε η τεχνική της Ανάλυσης Πρωταρχικών Συνιστωσών (PCA), με σκοπό τη μείωση των διαστάσεων του συνόλου, διατηρώντας όσο περισσότερο γίνεται την διασπορά των αρχικών δεδομένων.

Οι αλγόριθμοι κατηγοριοποίησης που συγκρίθηκαν είναι οι SVM, Logistic Regression, Gaussian Naive Bayes, Decision Tree και Random Forest. Τα αποτελέσματα των μετρικών για το SVM είναι καλύτερα σε σχέση με τις μετρικές των υπολοίπων, δηλαδή προβλέπει καλύτερα σε ποια κατηγορία ανήκει ένα καινούριο δείγμα.

Abstract

The subject that was studied in this thesis is the mining of dietary habits and data related to lifestyle. The dataset, that was collected from questionnaires, was provided by Mrs. Yannakoulia Maria, one of the members of the research team HELIAD. The features of the dataset contain not only nutritional habits concerning people over 65 years old, but also two nutritional indices: the first index is BMI (Body Mass Index) and the second is an indicator for the risk of malnutrition (which is called Determine).

The scope of this thesis is to use the tools of data mining and machine learning in order to discover patterns and extract useful conclusions for our data. This is accomplished by identifying any correlation between each one of the features and the class value. In this way, we can understand which and how many features of the dataset contain the highest information gain over the class, as well as which nutritional habits have the people who are at risk for malnutrition. Moreover, it is also important to find correlations between nutritional habits and conditions that affect mainly the elderly people, for example dementia.

The dataset contains multiple dimensions, therefore we used Principal Component Analysis (PCA) in order to reduce a subset of the features, retaining as much as possible the variance of the initial dataset.

The classifiers that were compared on the dataset are SVM, Logistic Regression, Gaussian Naive Bayes, Decision Tree and Random Forest. The evaluation metrics of the SVM algorithm were higher compared to the remainder classifiers, so we presume that SVM can have better results on the prediction of a new sample.

Εισαγωγή

Το θέμα που μελετήθηκε στα πλαίσια της παρούσας πτυχιακής είναι η εξόρυξη γνώσης από δεδομένα που αφορούν τις διατροφικές συνήθειες και τον τρόπο ζωής των ανθρώπων άνω των 65 ετών. Με τον όρο εξόρυξη γνώσης εννοούμε την εύρεση και εξαγωγή χρήσιμης πληροφορίας ή προτύπων από ένα σύνολο δεδομένων, ώστε να κατανοηθεί, να αναλυθεί και να οδηγήσει σε αποφάσεις (Νανόπουλος, 2008).

Στόχος την πτυχιακής είναι η εύρεση τέτοιων προτύπων, που θα μας οδηγήσουν σε κάποια συμπεράσματα για τους διατροφικούς δείκτες και την συσχέτισή τους με διάφορες διατροφικές συνήθειες. Αυτό θα γίνει εφικτό με την εφαρμογή των εργαλείων της εξόρυξης δεδομένων, των αλγορίθμων της μηχανικής μάθησης και της στατιστικής στο πεδίο της διατροφής και διαιτολογίας.

Πιο συγκεκριμένα, οι διατροφικοί δείκτες παρέχουν ένα μέτρο εκτίμησης της ποιότητας της διατροφής σε σχέση με την επάρκεια των θρεπτικών συστατικών που καταναλώνουν οι άνθρωποι. Είναι σημαντικό να προσδιορίσουμε ποια χαρακτηριστικά - διατροφικές συνήθειες έχουν οι ηλικιωμένοι που βρίσκονται σε κίνδυνο για υποσιτισμό και ποιοι διατροφικοί δείκτες οδηγούν σε αυτόν. Αυτό θα μας βοηθήσει να καταλάβουμε αν υπάρχουν συσχετίσεις μεταξύ των χαρακτηριστικών και του υποσιτισμού και πώς μπορούμε να εξαλείψουμε την εμφάνισή του.

Ο υποσιτισμός αναφέρεται ως μια κατάσταση που προκύπτει από τη μη-κατανάλωση ενέργειας ή απαραίτητων θρεπτικών συστατικών για τις ανάγκες του οργανισμού, με αποτέλεσμα να προκαλούνται σοβαρά προβλήματα υγείας. Το 2015 βρισκόταν σε υποσιτισμό το 13% του παγκόσμιου πληθυσμού, ποσοστό σχετικά μικρό συγκρινόμενο με αυτό του 1990, όπου είχε φτάσει το 23% (Skoet and Stamoulis, 2006). Μία από τις ομάδες που πλήττονται περισσότερο σήμερα είναι, μεταξύ άλλων, οι ηλικιωμένοι.

Τα δεδομένα που χρησιμοποιήθηκαν στην πτυχιακή προήλθαν από ερωτηματολόγια

της έρευνας HELIAD (Hellenic Longitudinal Investigation of Aging and Diet) (Dardiotis et al., 2014). Σκοπός της ήταν η εκτίμηση της άνοιας, της ήπιας γνωστικής εξασθένησης και άλλων νευροψυχιατρικών προβλημάτων της γήρανσης στον ελληνικό πληθυσμό, όπως επίσης και η εύρεση συσχετίσεων μεταξύ της διατροφής και των νευροψυχιατρικών ασθενειών που οφείλονται στην διαδικασία της γήρανσης.

Τα χαρακτηριστικά που θέλουμε να προσδιορίσουμε ομαδοποιούνται σε κατηγορίες. Μερικές από αυτές είναι το ατομικό και οικογενειακό ιατρικό ιστορικό των ηλικιωμένων, οι δραστηριότητες ελεύθερου χρόνου (για παράδειγμα περπάτημα, φυσική άσκηση, ενασχόληση με τέχνες και πολιτισμό), οι συνήθειες διατροφής, τα προβλήματα μνήμης, η γηριατρική κλίμακα κατάθλιψης, τα νευροψυχιατρικά συμπτώματα και άλλα. Πιο εκτενής αναφορά στις κατηγορίες των χαρακτηριστικών γίνεται στην υποενότητα 3.1.

Οι διατροφικοί δείκτες που περιέχονται στα χαρακτηριστικά του συνόλου δεδομένων είναι ο BMI και ο Determine. Ο BMI κατηγοριοποιεί έναν άνθρωπο ανάλογα με το βάρος και το ύψος του, ενώ ο Determine ¹ προκύπτει από 10 ερωτήσεις και εκτιμά αν κάποιος βρίσκεται ή όχι σε κίνδυνο υποσιτισμού. Οι δύο αυτοί δείκτες θα περιγραφούν αναλυτικά στην υποενότητα 3.5.

Επομένως με βάση τα παραπάνω καλούμαστε να απαντήσουμε στις εξής γενικές ερωτήσεις:

- Με ποια χαρακτηριστικά συσχετίζεται περισσότερο ο διατροφικός δείκτης Determine και BMI;
- Ποια είναι τα αποτελέσματα από τα μοντέλα πρόβλεψης που βασίζονται πάνω στο Determine και το BMI;
- Συσχετίζεται ο διατροφικός δείκτης Determine με την εμφάνιση ασθενειών όπως η άνοια;

¹<https://www.healthcare.uiowa.edu/igec/tools/nutrition/determineNutrition.pdf>

Θεωρητικό υπόβαθρο

Η ενότητα αυτή χωρίζεται σε τρεις υποενότητες. Η υποενότητα 2.1 περιλαμβάνει αναφορές σε ερευνητικές εργασίες που έχουν γίνει προηγουμένως στον κλάδο της εξόρυξης γνώσης από ιατρικά δεδομένα. Στην υποενότητα 2.2 και 2.3 γίνεται εκτενής αναφορά και επεξήγηση των εννοιών και εργαλείων που χρησιμοποιήθηκαν στην παρούσα πτυχιακή, ενώ η 2.4 περιλαμβάνει τα μέτρα που χρησιμοποιήθηκαν για την αξιολόγηση των κατηγοριοποιητών.

2.1 Σχετικές ερευνητικές εργασίες

Τις τελευταίες δύο δεκαετίες, υπάρχουν αναφορές στην βιβλιογραφία σχετικά με τη χρήση τεχνικών εξόρυξης γνώσης στο πεδίο της διατροφής και της ιατρικής, ενώ παλιότερα αυτές οι τεχνικές σπάνια εφαρμόζονταν σε αυτά τα πεδία.

Συγκεκριμένα το 1997, ένα έργο εξόρυξης δεδομένων έλαβε μέρος στο ιατρικό κέντρο του Πανεπιστημίου Duke στη Βόρεια Καρολίνα της Αμερικής (Prather et al., 1997). Χρησιμοποιήθηκε μια μεγάλη βάση δεδομένων που περιείχε ασθενείς μαιευτικής, με σκοπό τον εντοπισμό των παραγόντων που θα βελτιώσουν την ποιότητα και αποτελεσματικότητα της περιγεννητικής φροντίδας. Για την εξόρυξη γνώσης από τη βάση δεδομένων, χρησιμοποιήθηκε η διερευνητική παραγοντική ανάλυση (exploratory factor analysis), διότι είχε ανακαλύψει επιτυχώς στο παρελθόν πρότυπα σε δεδομένα της μαιευτικής.

Παραγοντική ανάλυση ονομάζεται η στατιστική μέθοδος που προσδιορίζει τα χαρακτηριστικά των δεδομένων που μπορούν να συνδυαστούν, έτσι ώστε να εξηγηθούν οι διακυμάνσεις μεταξύ των διαφορετικών ασθενών. Τα αποτελέσματα έδειξαν ότι υπάρχουν τρεις συνολικοί παράγοντες που καλύπτουν το 48.9% της διασποράς του συνόλου δεδομένων. Οπότε η παραγοντική ανάλυση θεωρήθηκε πολύ χρήσιμη για τον εντοπισμό κλινικών παραγόντων που σχετίζονται με την πρόωρη γέννηση.

Μία ακόμη εφαρμογή της εξόρυξης γνώσης στο πεδίο της διατροφής έγινε το 2002 σε ένα σύνολο διαβητικών ασθενών (Breault et al., 2002). Το σύνολο δεδομένων περιείχε 10 χαρακτηριστικά και συζητήθηκαν εκτενώς δύο μεταβλητές: ο δείκτης συννοσηρότητας μια μετρική του γλυκαιμικού ελέγχου. Ως αλγόριθμος κατηγοριοποίησης χρησιμοποιήθηκε το δέντρο κατηγοριοποίησης (classification tree), έτσι ώστε να βρεθούν σημαντικά πρότυπα για τον διαβητικό έλεγχο. Τα αποτελέσματα έδειξαν ότι η νεαρή ηλικία των ατόμων, και όχι ο δείκτης συννοσηρότητας, σχετίζεται με τον κακό διαβητικό έλεγχο.

Επιπλέον, το 2008 οι Hearty και Gibney χρησιμοποίησαν και σύγκριναν δύο μεθόδους επιβλεπόμενης μάθησης, τα τεχνητά νευρωνικά δίκτυα και τα δέντρα αποφάσεων, σε δεδομένα που σχετίζονται με πρόσληψη τροφίμων και θρεπτικών συστατικών για επτά συνεχόμενες ημέρες. Τα δεδομένα αυτά είχαν παρθεί από 1379 ενήλικες ηλικίας 18-64 χρονών, οι οποίοι ζούσαν στην Ιρλανδία (Hearty and Gibney, 2008).

Παρόλο που και οι δύο μέθοδοι είχαν χρησιμοποιηθεί στο παρελθόν στον ιατρικό κλάδο, δεν υπήρχαν μέχρι τότε αναφορές στην βιβλιογραφία για την χρήση τους στην επιστήμη της διατροφής. Τα αποτελέσματα της έρευνας έδειξαν ότι και τα νευρωνικά δίκτυα και τα δέντρα αποφάσεων μπορούν να χρησιμοποιηθούν για να προβλέψουν αποτελεσματικά την διατροφική ποιότητα.

Για τα ίδια δεδομένα χρησιμοποιήθηκαν και στατιστικές τεχνικές όπως cluster analysis και principal component analysis (PCA). Παρά το γεγονός ότι οι τεχνικές έχουν στατιστικές διαφορές, αναγνωρίζουν παρόμοια μοτίβα όταν εφαρμόζονται στο ίδιο σύνολο δεδομένων και τα μοτίβα αυτά είναι άμεσα συγκρίσιμα μεταξύ τους. Ωστόσο χρειάζεται προσοχή στις αποφάσεις που παίρνουμε όταν χρησιμοποιούμε PCA, διότι μπορεί να έχει άμεσο αντίκτυπο στον αριθμό και στο είδος των διατροφικών συνηθειών που προέκυψαν από τα δεδομένα (Hearty et al., 2009).

Η παρούσα πτυχιακή εργασία χρησιμοποιεί ένα σύνολο από διατροφικά δεδομένα, όπου σε κάθε εγγραφή αντιστοιχίζονται οι προαναφερθέντες δείκτες BMI και Determine. Δοθέντος αυτών των δεικτών, μπορούμε να κατηγοριοποιήσουμε άγνωστα δεδομένα σε καθεμία από τις κλάσεις, ανάλογα με τις τιμές των υπολοίπων χαρακτηριστικών. Αν δεν είχαμε κάποιον από τους δείκτες, δεν θα μπορούσαμε να κατηγοριοποιήσουμε τα δεδομένα, παρά μόνο να χρησιμοποιήσουμε αλγόριθμους συσταδοποίησης, οι οποίοι βρίσκουν κάποια σχέση μεταξύ των χαρακτηριστικών.

2.2 Επεξήγηση εννοιών

Σε αυτήν την ενότητα επεξηγούνται οι όροι εξόρυξη γνώσης και μηχανική μάθηση, καθώς και οι κατηγορίες στις οποίες μπορεί να διακριθεί κάθε ένας από τους όρους.

2.2.1 Εξόρυξη γνώσης

Εξόρυξη γνώσης ονομάζεται η ανακάλυψη μιας ενδιαφέρουσας, άγνωστης μέχρι στιγμής και πιθανώς χρήσιμης πληροφορίας ή προτύπων από μεγάλα σύνολα δεδομένων. Η εύρεση και εξαγωγή της πληροφορίας γίνεται με αλγορίθμους κατηγοριοποίησης ή ομαδοποίησης, κάνοντας χρήση των αρχών της μηχανικής μάθησης, της τεχνητής νοημοσύνης, της στατιστικής και των βάσεων δεδομένων (Chakrabarti et al., 2006). Ο όρος εξόρυξη δεδομένων αναφέρθηκε για πρώτη φορά από τους στατιστικολόγους, για να υποδηλώσουν την εξαγωγή πληροφορίας που δεν υποστηρίζεται από τα δεδομένα. Ωστόσο, σήμερα ο όρος έχει θετική σημασία.

Η βασική ιδέα είναι να βρεθούν πρότυπα στα δεδομένα, που η δομή τους θα είναι κατανοητή για τον άνθρωπο και θα τον οδηγήσει στην εξαγωγή κατάλληλων συμπερασμάτων. Τα πρότυπα που θα εξαχθούν μπορεί να είναι ομάδες από εγγραφές δεδομένων (ομαδοποίηση), ανίχνευση ασυνήθιστων εγγραφών (anomaly detection) ή εξαρτήσεις (εύρεση κανόνων συσχέτισεων).

2.2.2 Μηχανική μάθηση

Η μηχανική μάθηση, ένας κλάδος της τεχνητής νοημοσύνης, αναφέρεται ως το επιστημονικό πεδίο που μελετά και κατασκευάζει αλγορίθμους, οι οποίοι “μαθαίνουν” από τα δεδομένα και κάνουν προβλέψεις πάνω σε άγνωστα δεδομένα¹. Ο Tom M. Mitchell αναφέρει στο βιβλίο του (Learning, 2009) ότι το πεδίο της μηχανικής μάθησης ασχολείται με το πως να κατασκευαστούν προγράμματα υπολογιστών που βελτιώνονται αυτόματα με την εμπειρία, χωρίς την παρέμβαση του ανθρώπου.

Ανάλογα με τον τρόπο που μπορεί να εκπαιδευτεί ένα μοντέλο, η μηχανική μάθηση χωρίζεται σε τρεις κατηγορίες:

- Επιβλεπόμενη μάθηση (Supervised learning)

¹<http://ai.stanford.edu/ronnyk/glossary.html>

- Μη-επιβλεπόμενη μάθηση (Unsupervised learning)
- Ενισχυτική μάθηση (Reinforcement learning)

Επιβλεπόμενη μάθηση (Supervised learning)

Στην επιβλεπόμενη μάθηση, κάθε παράδειγμα από το σύνολο των δεδομένων εκπαίδευσης αποτελείται από ένα σύνολο εισόδου, τα χαρακτηριστικά (features), και μία επιθυμητή τιμή εξόδου, την κλάση (class). Οι αλγόριθμοι επιβλεπόμενης μάθησης αναλύουν τα δεδομένα εκπαίδευσης και έχουν ως στόχο την παραγωγή ενός γενικού κανόνα, ο οποίος χρησιμοποιείται για να χαρακτηρίσει καινούρια, άγνωστα μέχρι τώρα παραδείγματα.

Κάθε σύνολο δεδομένων μπορεί να χωριστεί σε σύνολο εκπαίδευσης (training set) και σύνολο ελέγχου (test set). Το σύνολο εκπαίδευσης περιέχει τις εγγραφές με την τιμή της κλάσης για κάθε μια από αυτές και χρησιμοποιείται για την δημιουργία του γενικού μοντέλου κατηγοριοποίησης. Το σύνολο ελέγχου δεν περιέχει την τιμή της κλάσης για κάθε εγγραφή, οπότε χρησιμοποιείται για την αξιολόγηση του μοντέλου πάνω σε αυτά τα άγνωστα δεδομένα.

Οι μεταβλητές εξόδου μπορεί να παίρνουν συνεχείς ή διακριτές τιμές. Συνεχείς είναι οι μεταβλητές που μπορούν να πάρουν οποιαδήποτε τιμή μέσα σε ένα συνεχές διάστημα, όπως για παράδειγμα το ύψος, το βάρος, η θερμοκρασία. Διακριτές είναι οι μεταβλητές που μπορούν να πάρουν μόνο διακεκριμένες τιμές στο συνεχές διάστημα. Παραδείγματα διακριτών τιμών είναι ο αριθμός παιδιών σε μια οικογένεια, ο πληθυσμός των κατοίκων σε μια περιοχή.

Επομένως ανάλογα με το αν η έξοδος παίρνει διακριτές ή συνεχείς τιμές, η επιβλεπόμενη μάθηση διακρίνεται σε κατηγοριοποίηση (classification) ή παλινδρόμηση (regression). Δηλαδή στην κατηγοριοποίηση, η κλάση που θέλουμε να προβλέψουμε για άγνωστα δεδομένα είναι ένας διακριτός αριθμός, ενώ στην παλινδρόμηση ένας συνεχής αριθμός.

Μη-επιβλεπόμενη μάθηση (Unsupervised learning)

Στην μη-επιβλεπόμενη μάθηση υπάρχουν μόνο οι τιμές εισόδου στο σύνολο δεδομένων και όχι οι τιμές εξόδου. Παραδείγματα τέτοιου είδους συνόλων είναι οι εικόνες, οι ηχογραφήσεις, τα βίντεο, όπου δεν υπάρχει κάποια συγκεκριμένη τιμή που μπορούμε να προβλέψουμε. Οπότε σκοπός του αλγορίθμου είναι να ανακαλύψει συσχετίσεις και δομές μεταξύ των παραδειγμάτων στο σύνολο των δεδομένων εισόδου, δηλαδή να περιγράψει

πως έχουν οργανωθεί τα παραδείγματα.

Παραδείγματα μεθόδων μη-επιβλεπόμενης μάθησης είναι η εύρεση κανόνων συσχέτισης και η συσταδοποίηση (clustering), στην οποία τα δεδομένα οργανώνονται σε ένα σύνολο ομάδων με τέτοιο τρόπο, έτσι ώστε τα αντικείμενα που ανήκουν στην ίδια ομάδα (συστάδα) να είναι παρόμοια μεταξύ τους και λιγότερο όμοια με τα αντικείμενα των υπολοίπων ομάδων.

Ενισχυτική μάθηση (Reinforcement learning)

Στην ενισχυτική μάθηση το σύστημα μαθαίνει από μια σειρά ενισχύσεων-ανταμοιβών ή τιμωριών, δηλαδή μέσα από την άμεση αλληλεπίδραση με το περιβάλλον (Russell et al., 2003). Το σύστημα δεν καθοδηγείται από κάποιον εξωτερικό παράγοντα για το ποια ενέργεια θα πρέπει να ακολουθήσει αλλά πρέπει να ανακαλύψει μόνο του ποιες ενέργειες θα του αποφέρουν το μεγαλύτερο κέρδος (Vlahavas, 2005). Η ενισχυτική μάθηση χρησιμοποιείται κυρίως σε προβλήματα Σχεδιασμού (Planning), όπως για παράδειγμα ο έλεγχος κίνησης ρομπότ.

2.2.3 Κατηγοριοποίηση

Η κατηγοριοποίηση, όπως αναφέρθηκε παραπάνω, είναι μια μέθοδος επιβλεπόμενης μάθησης. Έστω ότι έχουμε ένα σύνολο εγγραφών, το οποίο περιλαμβάνει ορισμένα χαρακτηριστικά, ένα εκ των οποίων είναι η κλάση, δηλαδή το χαρακτηριστικό που θέλουμε να προβλέψουμε. Η κάθε κλάση είναι μια κατηγορία στην οποία ανήκει μια ομάδα αντικειμένων, τα οποία έχουν ένα ή περισσότερα χαρακτηριστικά που μας ενδιαφέρουν. Η κατηγοριοποίηση λοιπόν ορίζεται ως η ανάθεση του συνόλου αυτού σε μία από τις διακριτές τιμές των κλάσεων (Vlahavas, 2005).

Σκοπός της κατηγοριοποίησης είναι η εύρεση ενός μοντέλου που προβλέπει την τιμή της κλάσης από τις τιμές των υπολοίπων χαρακτηριστικών, για δεδομένα του συνόλου ελέγχου.

Η κατηγοριοποίηση διακρίνεται σε δυαδική (binary) ή σε πολλαπλές κλάσεις (multiclass). Στην δυαδική κατηγοριοποίηση εμπλέκονται μόνο δύο κλάσεις, οπότε κάθε αντικείμενο ενός συνόλου δεδομένων μπορεί να κατηγοριοποιηθεί σε μία εκ των δύο κατηγοριών. Στην κατηγοριοποίηση πολλαπλών κλάσεων γίνεται ταξινόμηση των εγγραφών σε μία από πολλές κατηγορίες, όπου ο αριθμός των διαφορετικών κατηγοριών είναι μεγαλύτε-

ρος του δύο. Μία εφαρμογή δυαδικής κατηγοριοποίησης είναι για παράδειγμα η κατάταξη ενός συνόλου ασθενών για μια συγκεκριμένη ασθένεια σε υγιής ή μη υγιής.

Ωστόσο κάποιοι αλγόριθμοι κατηγοριοποίησης δεν λειτουργούν καλά όταν το σύνολο δεδομένων έχει περισσότερες από δύο κλάσεις. Έτσι το πρόβλημα της κατηγοριοποίησης σε πολλαπλές κλάσεις μπορεί να χωριστεί σε πολλές δυαδικές κατηγοριοποιήσεις, κάνοντας χρήση των τεχνικών one-vs-rest ή one-vs-one.

2.2.4 Συσταδοποίηση

Η συσταδοποίηση είναι μια μέθοδος μη-επιβλεπόμενης μάθησης, η οποία προσπαθεί να εντοπίσει κάποια δομή στα δεδομένα. Η δομή μπορεί να εκφραστεί είτε με τη μορφή ομάδων αντικειμένων είτε με ιεραρχία ομάδων (Νανόπουλος, 2008). Η κάθε ομάδα ονομάζεται συστάδα και η κύρια ιδιότητά της είναι ότι τα αντικείμενα που ανήκουν στην ομάδα είναι όσο το δυνατόν πλησιέστερα μεταξύ τους, ενώ τα αντικείμενα διαφορετικών ομάδων είναι όσο το δυνατόν πιο απομακρυσμένα.

Στην εξόρυξη δεδομένων, η ιεραρχία ομάδων ή αλλιώς ιεραρχική συσταδοποίηση έχει ως αποτέλεσμα τη δημιουργία μιας ιεραρχίας από συστάδες και διακρίνεται σε:

- Συσσωρευτική (Agglomerative)

Κάθε παράδειγμα του συνόλου δεδομένων είναι μια μεμονωμένη ομάδα. Καθώς μια ομάδα ανεβαίνει στην ιεραρχία, ζευγάρια συστάδων συγχωνεύονται και δημιουργούν μια καινούρια συστάδα.

- Διαιρετική (Divisive)

Όλα τα παραδείγματα του συνόλου δεδομένων εμφανίζονται ως μια μεγάλη συστάδα. Καθώς ένα παράδειγμα κατεβαίνει στην ιεραρχία, γίνεται διαχωρισμός των παραδειγμάτων και δημιουργούνται καινούριες μικρότερες συστάδες.

2.2.5 Ανάλυση πρωταρχικών συνιστωσών (Principal Component Analysis)

Η ανάλυση πρωταρχικών συνιστωσών (Principal Component Analysis - PCA) είναι μια στατιστική διαδικασία που επινοήθηκε από τον Pearson (1901) και αργότερα αναπτύχθηκε από τον Hotelling (1933), ενώ μια πιο σύγχρονη αναφορά έχει κάνει και ο Jolliffe (2002).

Σκοπός της μεθόδου είναι η μείωση των διαστάσεων του συνόλου δεδομένων, διατηρώντας όσο περισσότερο γίνεται την διασπορά των αρχικών δεδομένων. Γενικώς το

μεγαλύτερο ποσοστό των αλγορίθμων λειτουργεί καλύτερα όταν το πλήθος των διαστάσεων είναι περιορισμένο, δηλαδή τα δεδομένα είναι πιο πυκνά. Όσο μεγαλώνει το πλήθος των διαστάσεων, τόσο αυξάνεται και ο όγκος του χώρου με αποτέλεσμα τα δεδομένα να γίνονται πιο αραιά και οι αλγόριθμοι κατηγοριοποίησης και συσταδοποίησης να μην είναι τόσο αποτελεσματικοί. Αυτό αναφέρεται και ως κατάρα των πολλών διαστάσεων (dimensionality curse).

Η μείωση των διαστάσεων αποσκοπεί στη δυνατότητα οπτικοποίησης συνόλων με υψηλές διαστάσεις, αν ο τελικός αριθμός των διαστάσεων είναι μικρότερος ή ίσος του 3. Τα δεδομένα μπορούν να αναπαρασταθούν σε 2 ή 3 διαστάσεις, διευκολύνοντας έτσι την κατανόηση των αποτελεσμάτων. Επιπλέον, μειώνεται ο χώρος αναζήτησης με αποτέλεσμα να μειώνεται και ο χρόνος εκτέλεσης των αλγορίθμων (Νανόπουλος, 2008).

Πιο αναλυτικά, η ανάλυση πρωταρχικών συνιστωσών χρησιμοποιεί ορθογώνιο μετασχηματισμό ώστε να μετατρέψει τα χαρακτηριστικά του συνόλου δεδομένων που πιθανώς σχετίζονται μεταξύ τους σε γραμμικά ασυσχέτιστα, τα οποία ονομάζονται πρωταρχικές συνιστώσες και είναι συνδυασμός των αρχικών. Μετά τον μετασχηματισμό, λίγες αρχικές συνιστώσες περιέχουν την μέγιστη διασπορά που υπήρχε σε όλα τα αρχικά χαρακτηριστικά. Ο αριθμός των συνιστωσών είναι ίσος ή μικρότερος των αρχικών χαρακτηριστικών που υπάρχουν στο σύνολο δεδομένων (Wold et al., 1987). Η επιλογή του κατάλληλου αριθμού συνιστωσών γίνεται με τεχνικές, οι οποίες θα αναφερθούν και αναλυθούν στην υποενότητα 4.2.2.

2.2.6 Μηχανές διανυσμάτων υποστήριξης (Support vector machines)

Μηχανές διανυσμάτων υποστήριξης ή SVM (Support Vector Machines) ονομάζεται ένας αλγόριθμος επιβλεπόμενης μάθησης που χρησιμοποιείται σε προβλήματα κατηγοριοποίησης και παλινδρόμησης. Έστω ένα σύνολο δεδομένων με δύο κλάσεις, οι οποίες μπορούν να διαχωριστούν γραμμικά. Ένα οποιοδήποτε σημείο του συνόλου αυτού παριστάνεται ως ένα διάνυσμα p -διαστάσεων. Στόχος του αλγορίθμου είναι να ξεχωρίσει τα διάφορα σημεία, χρησιμοποιώντας ένα υπερεπίπεδο (hyperplane) - ή αλλιώς γραμμικό μοντέλο, με $p-1$ διαστάσεις.

Μπορεί να υπάρξουν αρκετά υπερεπίπεδα που να κατηγοριοποιούν τα δεδομένα. Βέλτιστο θεωρείται το υπερεπίπεδο που δημιουργεί το μέγιστο περιθώριο μεταξύ των δύο κλάσεων, οπότε ο αλγόριθμος επιλέγει εκείνο του οποίου η απόστασή του από τα κο-

ντινότερα σημεία σε κάθε πλευρά μεγιστοποιείται (Witten and Frank, 2005). Τα σημεία που βρίσκονται κοντά στο βέλτιστο υπερεπίπεδο, δηλαδή αυτά που έχουν την μικρότερη απόσταση από αυτό, ονομάζονται διανύσματα υποστήριξης.

Ο συγκεκριμένος αλγόριθμος με το υπερεπίπεδο μεγίστου περιθωρίου αποτελεί γραμμικός κατηγοριοποιητής και είχε προταθεί το 1963. Ωστόσο, τρεις δεκαετίες αργότερα προτάθηκε μια παραλλαγή του αλγορίθμου που επιτρέπει τη δημιουργία μη γραμμικών μοντέλων για την διαχώριση των δύο κλάσεων, αντικαθιστώντας κάθε γινόμενο των σημείων με μια συνάρτηση RBF (Radial Basis Function) kernel. Στην παρούσα πτυχιακή ο αλγόριθμος SVM δημιούργησε ένα μη γραμμικό μοντέλο για να διαχωρίσει τις εγγραφές που ανήκουν στην κλάση 0 από αυτές που ανήκουν στην 1, ορίζοντας ως παράμετρο την RBF.

2.3 Εργαλεία

Παρακάτω γίνεται αναφορά στα εργαλεία που χρησιμοποιήθηκαν κατά την διάρκεια της πτυχιακής για την προεπεξεργασία των δεδομένων, την εφαρμογή μοντέλων στο σύνολο εκπαίδευσης και την εξαγωγή συμπερασμάτων. Στην υποενότητα 2.3.1 περιγράφεται το περιβάλλον Weka και στην 2.3.2 οι βιβλιοθήκες Python.

2.3.1 Weka

Το Weka (Waikato Environment for Knowledge Analysis) είναι μια δημοφιλής συλλογή από αλγορίθμους μηχανικής μάθησης, που χρησιμοποιούνται σε εφαρμογές εξόρυξης δεδομένων. Αναπτύχθηκε στο Πανεπιστήμιο Waikato, στη Νέα Ζηλανδία και οι αλγόριθμοι είναι γραμμένοι στη γλώσσα προγραμματισμού Java.

Το Weka περιέχει εργαλεία που αφορούν τα κύρια προβλήματα εξόρυξης δεδομένων, όπως κατηγοριοποίηση, παλινδρόμηση, συσταδοποίηση, εύρεση κανόνων συσχέτισης και επιλογή χαρακτηριστικών. Οι αλγόριθμοι μπορούν να εφαρμοστούν απευθείας σε ένα σύνολο δεδομένων μέσα από το γραφικό περιβάλλον διεπαφής χρήστη (GUI) ή να κληθούν μέσα από κώδικα java. Παρ' όλα αυτά, ένα μεγάλο και σημαντικό βήμα πριν την εφαρμογή των αλγορίθμων, είναι να γνωρίσουμε αρκετά καλά τα δεδομένα μας, το οποίο επιτυγχάνεται με εργαλεία προεπεξεργασίας και οπτικοποίησης δεδομένων που παρέχει το Weka.

2.3.2 Βιβλιοθήκες Python

Η γλώσσα προγραμματισμού Python είναι αρκετά δημοφιλής στον τομέα της ανάλυσης δεδομένων και πιο συγκεκριμένα σε αυτόν της εξόρυξης και οπτικοποίησης δεδομένων. Παρακάτω παρουσιάζονται κάποιες βιβλιοθήκες Python που χρησιμοποιήθηκαν καθόλη τη διάρκεια της πτυχιακής για προεπεξεργασία και οπτικοποίηση των δεδομένων, εκπαίδευση των δεδομένων με τους αλγορίθμους μηχανικής μάθησης και ανάλυση των αποτελεσμάτων.

- NumPy

Το NumPy είναι η θεμελιώδης βιβλιοθήκη της Python για επιστημονικούς υπολογισμούς. Υποστηρίζει μεγάλους σε μέγεθος, πολυδιάστατους πίνακες, μεθόδους γραμμικής άλγεβρας και άλλες υψηλού επιπέδου μαθηματικές λειτουργίες.

- SciPy

Το SciPy είναι μια βιβλιοθήκη ανοιχτού κώδικα που χρησιμοποιείται από επιστήμονες, αναλυτές και μηχανικούς για επιστημονικούς και τεχνικούς υπολογισμούς. Έχει δημιουργηθεί για να υποστηρίζει πίνακες NumPy και υπολογίζει αποδοτικά αριθμητικές λειτουργίες.

- Pandas

Το Pandas είναι μια βιβλιοθήκη που προσφέρει υψηλής απόδοσης εργαλεία για ανάλυση δεδομένων στην Python και δομές δεδομένων που είναι εύκολες στη χρήση.

- Scikit-learn

Η βιβλιοθήκη Scikit-learn περιέχει ένα σύνολο από επιβλεπόμενους και μη-επιβλεπόμενους αλγορίθμους μηχανικής μάθησης και έχει σχεδιαστεί να λειτουργεί με τις βιβλιοθήκες NumPy και SciPy.

- Matplotlib

Το Matplotlib είναι μια βιβλιοθήκη σχεδίασης δισδιάστατων διαγραμμάτων. Με λίγες γραμμές κώδικα παράγει γραφήματα όπως ιστογράμματα, φάσματα ισχύος, ραβδογράμματα, κυκλικά διαγράμματα και άλλα, απεικονίζοντας έτσι την κατανομή των δεδομένων στον χώρο. Επιπρόσθετα πακέτα εργαλείων είναι το mplot3d για την

δημιουργία 3D διαγραμμάτων και η διεπαφή seaborn ² για απεικόνιση στατιστικών δεδομένων.

- Imbalanced-learn ³

Το Imbalanced-learn είναι ένα πακέτο Python, το οποίο προσφέρει τεχνικές δειγματοληψίας που χρησιμοποιούνται συνήθως σε σύνολα δεδομένων για να αντιμετωπιστεί η ανισότητα μεταξύ δύο ή περισσότερων κλάσεων. Περιέχει αλγορίθμους οι οποίοι προσαρμόζουν την κατανομή των κλάσεων σε ένα σύνολο δεδομένων, όπως για παράδειγμα μέθοδοι undersampling, oversampling, ή και συνδυασμός των δύο αυτών μεθόδων.

2.4 Μέτρα αξιολόγησης

Η αξιολόγηση της επίδοσης ενός κατηγοριοποιητή είναι μια θεμελιώδης πτυχή της μηχανικής μάθησης, διότι μας δίνεται η δυνατότητα να καταλάβουμε την ποιότητα του μοντέλου που χρησιμοποιούμε και να συγκρίνουμε τα διάφορα μοντέλα μεταξύ τους.

2.4.1 Μετρικές

Η βασικότερη μετρική αξιολόγησης ενός κατηγοριοποιητή είναι η ευστοχία (accuracy) ενός μοντέλου, η οποία λαμβάνει τιμές στο διάστημα $[0, 1]$ και ορίζεται ως εξής:

$$\text{Accuracy} = \frac{\text{Αριθμός σωστών προβλέψεων}}{\text{Συνολικός αριθμός προβλέψεων}} \quad (2.1)$$

Σε ένα πρόβλημα δυαδικής κατηγοριοποίησης με κλάσεις true και false αντίστοιχα, μπορούμε να διακρίνουμε τις εξής περιπτώσεις:

- Αληθώς θετικές περιπτώσεις (True Positive - TP)

Η εγγραφή ανήκει στην κλάση true και κατηγοριοποιήθηκε ως true

- Αληθώς αρνητικές περιπτώσεις (True Negative- TN)

Η εγγραφή ανήκει στην κλάση false και κατηγοριοποιήθηκε ως false

²<http://web.stanford.edu/~mwaskom/software/seaborn/>

³<http://contrib.scikit-learn.org/imbalanced-learn/>

- Ψευδώς θετικές περιπτώσεις (False Positive - FP)

Η εγγραφή ανήκει στην κλάση false και κατηγοριοποιήθηκε ως true

- Ψευδώς αρνητικές περιπτώσεις (False Negative - FN)

Η εγγραφή ανήκει στην κλάση true και κατηγοριοποιήθηκε ως false

Σε πολλά προβλήματα όμως η ευστοχία μπορεί να μην αποτελεί καλή μετρική για την αξιολόγηση της κατηγοριοποίησης, επειδή μπορεί να νομίζουμε ότι ένας κατηγοριοποιητής είναι καλός, ενώ στην πραγματικότητα δεν είναι. Επίσης είναι πιθανό ένα μοντέλο με μια συγκεκριμένη τιμή ευστοχίας μπορεί να είναι καλύτερο σε σύγκριση με ένα άλλο που έχει υψηλότερη τιμή ευστοχίας.

Αυτό μπορεί να συμβεί στο παράδειγμα της δυαδικής κατηγοριοποίησης, όπου ο αριθμός των παραδειγμάτων που ανήκουν στη μία κλάση διαφέρει σε μεγάλο βαθμό από τον αριθμό των παραδειγμάτων που ανήκουν στην άλλη κλάση. Αυτό το πρόβλημα επεκτείνεται και στη περίπτωση της κατηγοριοποίησης πολλαπλών κλάσεων.

Για τους παραπάνω λόγους, είναι προτιμότερο να λαμβάνουμε υπόψιν και άλλες μετρικές αξιολόγησης οι οποίες ορίζονται ως εξής:

- Ακρίβεια (precision) είναι το ποσοστό των παραδειγμάτων που ο κατηγοριοποιητής έχει κατηγοριοποιήσει ως θετικά και είναι πραγματικά θετικά.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2.2)$$

- Ανάκληση (recall) είναι το ποσοστό των θετικών παραδειγμάτων που κατάφερε να βρει ο κατηγοριοποιητής

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2.3)$$

- F-score θεωρείται ο σταθμισμένος αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης

$$F - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.4)$$

2.4.2 Πίνακας σύγχυσης (Confusion matrix)

Για μια πιο λεπτομερή αξιολόγηση και οπτικοποίηση της επίδοσης μπορούμε να χρησιμοποιήσουμε τον πίνακα σύγχυσης ενός δυαδικού κατηγοριοποιητή. Ο πίνακας έχει διαστάσεις 2X2, δηλαδή περιέχει συνολικά δύο γραμμές και δύο στήλες και χρησιμοποιεί τον αριθμό των true positives, true negatives, false positives και false negatives. Κάθε στήλη αναπαριστά τον αριθμό των παραδειγμάτων στην κλάση που κατηγοριοποιήθηκαν ενώ κάθε γραμμή τον αριθμό των παραδειγμάτων στην πραγματική τους κλάση.

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

Σχήμα 2.1: Πίνακας σύγχυσης για δυαδική κατηγοριοποίηση

Οι σωστές προβλέψεις βρίσκονται στην διαγώνιο του πίνακα, ενώ στα υπόλοιπα κελιά φαίνεται ο αριθμός των λανθασμένων προβλέψεων. Οπότε ιδανικά θα θέλαμε η διαγώνιος να περιέχει μεγάλους αριθμούς, ενώ τα υπόλοιπα κελιά να τείνουν όσο το δυνατόν περισσότερο στο μηδέν (Powers, 2011).

2.4.3 Καμπύλη ROC (Receiver Operating Characteristic)

Η καμπύλη ROC είναι μια ακόμη μέθοδος αξιολόγησης και γραφικής αναπαράστασης της απόδοσης ενός δυαδικού κατηγοριοποιητή στο σύνολο δεδομένων, χωρίς όμως να λαμβάνεται υπόψιν η κατανομή των κλάσεων ή τα κόστη των λαθών της κατηγοριοποίησης. Η καμπύλη δημιουργείται απεικονίζοντας το ποσοστό των true positive (TP rate) στον κάθετο άξονα και το ποσοστό των false positive (FP rate) στον οριζόντιο άξονα.

Το TP rate ορίζει πόσα σωστά θετικά αποτελέσματα προκύπτουν κατά τη διάρκεια του ελέγχου μεταξύ όλων των θετικών δειγμάτων, ενώ το FP rate ορίζει πόσα λανθασμένα θετικά αποτελέσματα προκύπτουν μεταξύ όλων των αρνητικών δειγμάτων.

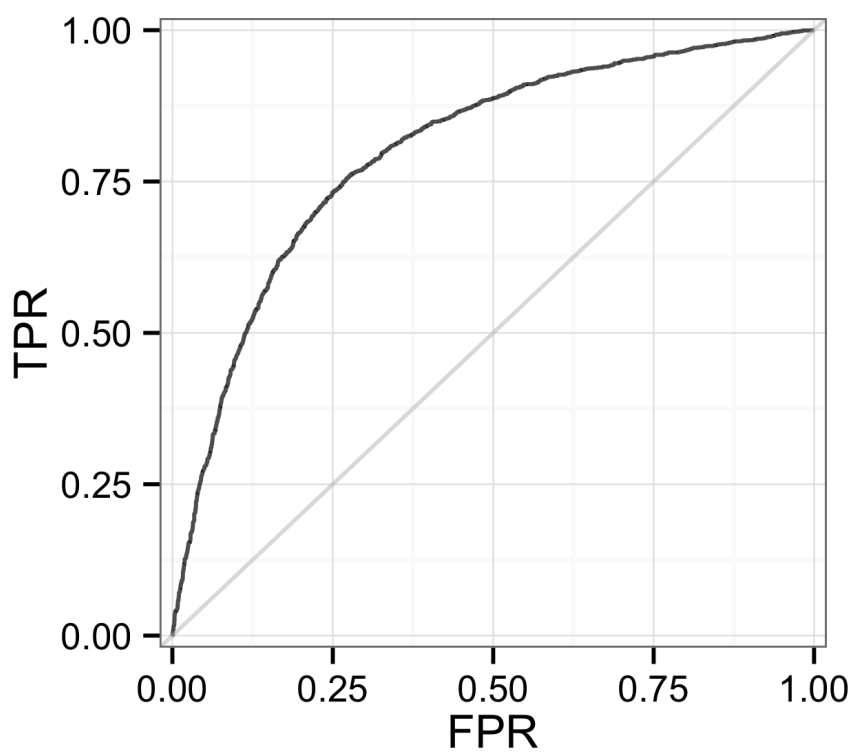
$$\text{TP rate} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \quad (2.5)$$

$$\text{FP rate} = \frac{\text{FP}}{\text{FP} + \text{TN}} \times 100 \quad (2.6)$$

Σημαντικό στοιχείο στις καμπύλες ROC είναι η περιοχή κάτω από την καμπύλη (Area under curve - AUC), δηλαδή το εμβαδόν που περικλείεται από την καμπύλη και τον οριζόντιο άξονα. Η περιοχή αυτή μετράει την ευστοχία (accuracy) του κατηγοριοποιητή, οπότε όσο πιο μεγάλη είναι, τόσο καλύτερη είναι η απόδοση του κατηγοριοποιητή.

Στο Σχήμα 2.2 που δίνεται ως παράδειγμα παρακάτω, η καλύτερη μέθοδος πρόβλεψης θα έδινε ένα σημείο στην πάνω αριστερή γωνία του διαγράμματος με συντεταγμένες (0,1), διότι το ποσοστό των TP και TN θα είναι 100%, ενώ το ποσοστό των FP, FN θα είναι 0%. Ένα οποιοδήποτε σημείο πάνω στην διαγώνιο ευθεία δηλώνει ότι επιλέχθηκε εντελώς τυχαία, όπως για παράδειγμα τα αποτελέσματα που παίρνουμε με τη ρίψη νομίσματος. Οπότε μια καμπύλη που βρίσκεται πάνω στη διαγώνιο δεν θεωρείται καλός κατηγοριοποιητής. Η διαγώνιος χωρίζει τον χώρο σε 2 μέρη: όσα σημεία βρίσκονται στον χώρο πάνω από αυτήν θεωρούνται αποτελέσματα καλού κατηγοριοποιητή, ενώ όσα βρίσκονται στον χώρο κάτω από τη διαγώνιο δηλώνουν ότι τα αποτελέσματα της κατηγοριοποίησης είναι χειρότερα από αυτά της τυχαίας επιλογής.

Επομένως, όσο πιο ψηλά και αριστερά βρίσκεται η καμπύλη, τόσο πιο ακριβείς προβλέψεις θα έχουμε στο σύνολο δοκιμών. Η καμπύλη που φαίνεται στο σχήμα αναπαριστά έναν καλό κατηγοριοποιητή, διότι βρίσκεται πάνω από τη διαγώνιο.



Σχήμα 2.2: Παράδειγμα ROC καμπύλης

Ανάλυση δεδομένων

3.1 Περιγραφή δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα πτυχιακή προήλθαν από το ερωτηματολόγιο του ερευνητικού προγράμματος Hellenic Longitudinal Investigation of Aging and Diet (HELIAD). Στόχος της έρευνας είναι η εκτίμηση της επικράτησης και συχνότητας ορισμένων ασθενειών που σχετίζονται με την γήρανση, όπως για παράδειγμα η νόσος Αλτσχάιμερ, η άνοια, η ήπια γνωστική εξασθένηση και άλλες νευροψυχιατρικές ασθένειες στον ελληνικό πληθυσμό καθώς και η συσχέτισή τους με την διατροφή (Dardiotis et al., 2014).

Οι ερωτήσεις του ερωτηματολογίου συλλέχθηκαν από ανθρώπους που ζουν μόνιμα στην Λάρισα και στο Μαρούσι Αττικής, ηλικίας 43 έως 99 ετών. Το τελικό δείγμα των συμμετεχόντων αποτελείται από περίπου 2000 ανθρώπους, από τους οποίους το 40% είναι άντρες και το 60% γυναίκες. Τα χαρακτηριστικά που συλλέχθηκαν μπορούν να χωριστούν στις εξής κατηγορίες:

- Στοιχεία συμμετέχοντα (ονοματεπώνυμο, ηλικία, φύλο, διεύθυνση, τόπος γέννησης κ.ά)
- Δημογραφικά (οικογενειακή κατάσταση, αριθμός παιδιών, χρόνια εκπαίδευσης, επάγγελμα κ.ά)
- Ιατρικό και οικογενειακό ιστορικό
- Δραστηριότητες ελεύθερου χρόνου - κοινωνικές επαφές
- Λειτουργικότητα, φυσικές δραστηριότητες και σωματική λειτουργικότητα
- Προβλήματα μνήμης

- Ύπνος, σύνδρομο ανήσυχων κάτω άκρων
- Γηριατρική κλίμακα κατάθλιψης (GDS), νοσοκομειακή μέτρηση άγχους
- Νευροψυχιατρικά συμπτώματα (NPI)
- Αδρή νευρολογική εξέταση
- Κλινική εκτίμηση της άνοιας
- Άνοια με σωμάτια Lewy
- Συχνότητα κατανάλωσης τροφίμων
- Εκτίμηση διατροφικού κινδύνου

3.2 Προεπεξεργασία δεδομένων

Η προεπεξεργασία των δεδομένων αποτελεί σημαντικό και απαραίτητο στάδιο της εξόρυξης δεδομένων, πριν την εφαρμογή των αλγορίθμων. Τα δεδομένα συνήθως συλλέγονται με διαδικασίες που μπορεί να μην διευκολύνουν την εξόρυξη γνώσης από αυτά, με αποτέλεσμα να περιέχουν θόρυβο και να επηρεάζεται αρνητικά η ποιότητά τους. Οι διαδικασίες περιλαμβάνουν συλλογή δεδομένων από ερωτηματολόγια, συσκευές μέτρησης όπως αισθητήρες ή ιατρικά όργανα, συνεντεύξεις, παρατηρήσεις ή το διαδίκτυο. Ο θόρυβος και οι ασυνεπείς τιμές μπορεί να προέρχονται από ανθρώπινο λάθος κατά την εισαγωγή τιμών ή από προβλήματα συσκευών. Παράδειγμα ασυνεπών τιμών είναι ο αρνητικός αριθμός για την ηλικία ή όταν ο συνδυασμός Φύλο - Εγκυμοσύνη αντιστοιχίζεται στις τιμές Άντρας - Ναι. Επομένως, η προεπεξεργασία δεδομένων αποτελείται από τα στάδια του καθαρισμού και μετασχηματισμού δεδομένων.

3.2.1 Καθαρισμός δεδομένων

Η επιστήμη της στατιστικής και της μηχανικής μάθησης προσφέρουν εργαλεία πρόβλεψης χρησιμοποιώντας στατιστικά μοντέλα και αλγορίθμους αντίστοιχα. Για την εφαρμογή των στατιστικών μοντέλων στα σύνολα δεδομένων, είναι απαραίτητο τα δεδομένα να είναι καθαρά, δηλαδή να μην έχουν ελλιπείς και άκυρες τιμές. Επειδή η συλλογή των δεδομένων έγινε μέσω των ερωτηματολογίων, είναι φυσικό να περιέχονται τέτοιες τιμές, οπότε ο καθαρισμός τους αποτελεί επιτακτική ανάγκη.

Στα σύνολα δεδομένων, η ύπαρξη ελλιπών τιμών αναφέρεται στην μη αποθήκευση κάποιας συγκεκριμένης τιμής σε ένα πεδίο ή μια εγγραφή. Οι ελλιπείς τιμές είναι ένα συχνό φαινόμενο στα αρχικά σύνολα δεδομένων και μπορεί να έχουν αρνητική επίδραση στα αποτελέσματα που θα προκύψουν. Υπάρχουν αλγόριθμοι που μπορεί να αγνοούν εντελώς τις ελλιπείς τιμές κατά την εκτέλεση τους, άλλοι να τις θεωρούν ως ξεχωριστή κατηγορία (όταν έχουμε κατηγορικές μεταβλητές) ενώ άλλοι να είναι ευαίσθητοι στην ύπαρξή τους.

Οι ελλιπείς τιμές μπορούν κατά κύριο λόγο να προκύψουν από δύο ειδών παράγοντες: ανθρώπινης ή τεχνικής φύσεως. Στην πρώτη περίπτωση προκύπτουν είτε επειδή οι άνθρωποι δεν θέλουν να δημοσιοποιήσουν στοιχεία της προσωπικής τους ζωής, οπότε δεν συμπληρώνουν τα αντίστοιχα πεδία σε ερωτηματολόγια και συνεντεύξεις, είτε επειδή παραλείπουν εν αγνοία τους την συμπλήρωση διαφόρων πεδίων. Στην δεύτερη περίπτωση οι συσκευές καταγραφής δεδομένων, για παράδειγμα οι αισθητήρες, μπορεί να εμφανίσουν πρόβλημα και μην καταγράψουν κάποια τιμή.

Αποτέλεσμα των παραπάνω περιπτώσεων είναι η τιμή μιας μεταβλητής να παραμείνει κενή. Υπάρχουν όμως και εγγραφές που περιέχουν κάποια τιμή, η οποία δεν μπορεί να είναι έγκυρη για το συγκεκριμένο πεδίο. Για παράδειγμα τα πεδία ηλικία, ύψος και βάρος δεν μπορούν να έχουν αρνητικές τιμές ή το πεδίο αριθμός των παιδιών δεν μπορεί να περιέχει δεκαδικούς αριθμούς. Τα λάθη αυτά γίνονται συνήθως κατά την εισαγωγή τιμών στα πεδία και πρέπει να αντικατασταθούν με την σωστή τιμή, αν γνωρίζουμε ποια είναι, ή με διαγραφούν εντελώς από το σύνολο δεδομένων.

Το σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα πτυχιακή ήταν αρχικά σε μορφή *xlsx* και μετατράπηκε σε *csv* (comma separated value) ώστε να αναγνωρίζεται και να φορτώνεται στο Weka. Κάποια χαρακτηριστικά είχαν μεγάλο ποσοστό ελλιπών τιμών ενώ κάποια άλλα σχεδόν καθόλου. Στις περισσότερες εγγραφές, τα πεδία που δεν είχαν απαντηθεί από τους ερωτηθέντες δεν ήταν κενά αλλά είχαν την τιμή -1000, οπότε και διεγράφη η τιμή τους.

Στη συνέχεια, έγινε διόρθωση ή διαγραφή των τιμών που δεν ταίριαζαν στο πεδίο όπου ανήκαν. Τα χαρακτηριστικά που είχαν πάνω από 85% ελλιπείς τιμές διαγράφηκαν εντελώς, επειδή το σύνολο δεδομένων που θα προέκυπτε δεν θα ήταν αντιπροσωπευτικό και δεν θα μας προσέδιδε παραπάνω πληροφορία. Οι ελλιπείς τιμές των υπολοίπων χαρακτηριστικών αντικαταστάθηκαν με την μέση τιμή του κάθε γνωρίσματος, αν το γνώρισμα είναι αριθμητικό, ή με τον μέσο, αν είναι κατηγορικό.

Ο διαδικασία του καθαρισμού των δεδομένων περιλαμβάνει επίσης τη διαγραφή ορισμένων χαρακτηριστικών που δεν προσφέρουν κάποια πληροφορία για εξαγωγή γνώσης, όπως για παράδειγμα το ονοματεπώνυμο, ο κωδικός των συμμετεχόντων, η ημερομηνία, ο τόπος γέννησης και άλλα προσωπικά στοιχεία των συμμετεχόντων, τα οποία μπορεί να θίγουν την ιδιωτικότητα τους.

Οι παραπάνω ενέργειες έγιναν μέσω του προγράμματος Weka, χρησιμοποιώντας τις κλάσεις `ReplaceMissingValues` για αντικατάσταση των ελλিপών τιμών με τη μέση τιμή ή τον μέσο, `Remove` για τη διαγραφή χαρακτηριστικών και `NumericToNominal` για μετατροπή των ποσοτικών χαρακτηριστικών σε κατηγορικά. Συγκεκριμένα, το ποσοστό των χαρακτηριστικών που είχαν ελλείψεις τιμές και αντικαταστάθηκαν με τη μέση τιμή ή τον μέσο ήταν 62.5%, ενώ από αυτά το 43.9% είχε ποσοστό ελλিপών τιμών κάτω από 10%.

3.2.2 Μετασχηματισμός δεδομένων

Μετασχηματισμός δεδομένων ορίζεται ως η εφαρμογή μιας μαθηματικής συνάρτησης σε κάθε σημείο του συνόλου δεδομένων, έτσι ώστε η μορφή που έχουν να μετατραπεί σε μια άλλη πιο χρήσιμη μορφή, σύμφωνα με τις απαιτήσεις του κάθε αλγορίθμου και το πεδίο ορισμού του προβλήματος. Τα δεδομένα μπορεί να περιέχουν χαρακτηριστικά με διαφορετικές κλίμακες και εύρος τιμών. Για παράδειγμα, μία εγγραφή μπορεί να έχει την ίδια αριθμητική τιμή “10” για δύο διαφορετικά χαρακτηριστικά, όμως η μονάδα μέτρησης του πρώτου χαρακτηριστικού να είναι τα γραμμάρια και του δεύτερου τα κιλά. Συνεπώς θα ήταν χρήσιμο να μετατρέψουμε τα δεδομένα, έτσι ώστε να έχουν όλα να ανήκουν στο ίδιο εύρος τιμών.

Οι πιο συχνές τεχνικές μετασχηματισμού δεδομένων είναι οι εξής:

- Τυποποίηση (standardization ή z-score): τα χαρακτηριστικά ακολουθούν την κανονική κατανομή με μέση τιμή $\bar{X} = 0$ και τυπική απόκλιση $S = 1$. Η τυποποίηση για ένα συγκεκριμένο σημείο X γίνεται με βάση τον μετασχηματισμό:

$$X = \frac{X - \bar{X}}{S} \quad (3.1)$$

- Min-max scaling: τα δεδομένα μετασχηματίζονται έτσι ώστε να ανήκουν μέσα σε ένα συγκεκριμένο διάστημα, για παράδειγμα $[0, 1]$. Ο μετασχηματισμός για ένα συγκεκριμένο σημείο X γίνεται ως εξής:

$$X = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \times (X'_{\max} - X'_{\min}) + X'_{\min} \quad (3.2)$$

Όπου $X_{\min}$ και $X_{\max}$ η ελάχιστη και αντίστοιχα η μέγιστη τιμή που είχε αρχικά το X , ενώ $X'_{\min}$ και $X'_{\max}$ είναι η ελάχιστη και μέγιστη τιμή του X στο νέο διάστημα.

Στο συγκεκριμένο σύνολο δεδομένων ο μετασχηματισμός κρίθηκε απαραίτητος, όχι μόνο επειδή τα δεδομένα είχαν διαφορετικές μονάδες, αλλά και επειδή οι αλγόριθμοι που χρησιμοποιούμε παρακάτω (SVM, logistic regression, τεχνική PCA) το επιβάλλουν. Χρησιμοποιώντας την μέθοδο StandardScaler από την βιβλιοθήκη Scikit-learn, τα δεδομένα κανονικοποιήθηκαν έτσι ώστε να είναι συγκεντρωμένα γύρω από την μέση τιμή $\bar{X} = 0$ και να έχουν τυπική απόκλιση $S = 1$.

3.3 Ασύμμετρο σύνολο δεδομένων (Imbalanced dataset)

Έστω ότι έχουμε ένα σύνολο δεδομένων με δύο κλάσεις, για παράδειγμα 0 και 1. Η πιο απλή περίπτωση κατηγοριοποίησης είναι να ταξινομήσουμε κάθε εγγραφή του συνόλου σε μία εκ των δύο κλάσεων, οπότε και έχουμε δυαδική κατηγοριοποίηση. Ιδανικά θα θέλαμε το πλήθος των δεδομένων να είναι συμμετρικό ως προς τις δύο κλάσεις, δηλαδή σε κάθε κλάση να ανήκει περίπου το ίδιο πλήθος εγγραφών. Όμως στα περισσότερα σύνολα που συναντάμε και καλούμαστε να εφαρμόσουμε τεχνικές εξόρυξης δεδομένων, η κατανομή των εγγραφών στις δύο κλάσεων είναι ασύμμετρη, δηλαδή η μία κλάση περιέχει πολλές περισσότερες εγγραφές σε σύγκριση με την άλλη κλάση (Liu et al., 2007).

Ένα παράδειγμα ασύμμετρου συνόλου είναι όταν θέλουμε να ανιχνεύσουμε απάτες μεταξύ συναλλαγών με πιστωτικές κάρτες. Σε αυτήν την περίπτωση, οι περισσότερες συναλλαγές είναι νόμιμες, όμως υπάρχει ένα πολύ μικρό ποσοστό που θεωρούνται απάτες. Η άνιση κατανομή επιφέρει προβλήματα, διότι στην προκειμένη περίπτωση, ο αλγόριθμος κατηγοριοποίησης θα προβλέπει σχεδόν πάντα ότι μια άγνωστη συναλλαγή είναι νόμιμη, με αποτέλεσμα να μην ανιχνεύονται οι απάτες.

Σε τέτοια προβλήματα, η ευστοχία (accuracy) του μοντέλου έχει συνήθως αρκετά υψηλή τιμή, διότι ο κατηγοριοποιητής “μαθαίνει” να προβλέπει σχεδόν πάντα την κλάση με τις περισσότερες εγγραφές και σπάνια την άλλη κλάση. Ένας τέτοιος κατηγοριοποιητής δεν θεωρείται κατάλληλος για τα δεδομένα, παρά την υψηλή τιμή της ευστοχίας που παρουσιάζει. Για αυτόν τον λόγο είναι σημαντικό να αξιολογούμε την επίδοση του μοντέ-

λου χρησιμοποιώντας και άλλες μετρικές, όπως ο πίνακας σύγχυσης (confusion matrix), η ακρίβεια (precision), η ανάκληση (recall), το f-score ή οι καμπύλες ROC, όπως θα δούμε στη συνέχεια.

Έχουν προταθεί πολλές προσεγγίσεις για να λυθεί το πρόβλημα του ασύμμετρου συνόλου δεδομένων. Μια λύση είναι η χρήση μεθόδων αναδειγματοληψίας, η οποία διακρίνεται σε υπερδειγματοληψία (oversampling) και υποδειγματοληψία (undersampling). Στην πρώτη περίπτωση αυξάνεται ο αριθμός των εγγραφών στην κλάση μειοψηφίας (minority class) ενώ στην δεύτερη μειώνεται ο αριθμός των εγγραφών στην κλάση πλειοψηφίας (majority class) του συνόλου, έτσι ώστε να υπάρχει ισορροπία μεταξύ των κλάσεων. Οι μέθοδοι μπορούν να εφαρμοστούν ανεξάρτητα η μία από την άλλη ή να γίνει συνδυασμός και των δύο, ανάλογα με τις απαιτήσεις του κάθε συνόλου δεδομένων.

Στο συγκεκριμένο σύνολο δεδομένων, οι αλγόριθμοι που χρησιμοποιήθηκαν έπαιρναν ως είσοδο στην κλάση τα πεδία BMI, Determine ή την άννοια. Και στις τρεις περιπτώσεις τα δεδομένα έχουν άνιση κατανομή ως προς την κλάση, οπότε η χρήση μεθόδων αναδειγματοληψίας είναι επιτακτική. Συγκεκριμένα, χρησιμοποιήθηκε η κλάση `SmoteTomek`¹, που εκτελεί υπερδειγματοληψία με την χρήση SMOTE και υποδειγματοληψία με την χρήση Tomek Links. Για την καλύτερη κατανόηση της έννοιας SMOTETomek, θα περιγραφούν εν συντομία οι έννοιες SMOTE και Tomek Links.

SMOTE (Synthetic Minority Oversampling Technique) (Chawla et al., 2002) ονομάζεται μια μέθοδος υπερδειγματοληψίας, η οποία αυξάνει την κλάση μειοψηφίας παράγοντας νέα συνθετικά δείγματα που βασίζονται στα πραγματικά. Έχει αποδειχθεί ότι εμφανίζει καλύτερα αποτελέσματα σε σχέση με την μέθοδο της υπερδειγματοληψίας με επανατοποθέτηση, διότι τα δείγματα που δημιουργούνται δεν είναι αντίγραφα του συνόλου δεδομένων, αλλά κάθε δείγμα δημιουργείται από ένα άλλο υπάρχον δείγμα βρίσκοντας τους k κοντινότερους γείτονες του υπάρχοντος δείγματος. Στη συνέχεια επιλέγεται τυχαία ένα σημείο από τα k . Το νέο συνθετικό σημείο είναι επίσης ένα τυχαία επιλεγμένο σημείο, πάνω στο ευθύγραμμο τμήμα που ενώνει το υπάρχον σημείο και τον επιλεγμένο γείτονα.

Το Tomek links είναι μια μέθοδος υποδειγματοληψίας, που αποσκοπεί όχι μόνο στην εξισορρόπηση του συνόλου δεδομένων, αλλά και στην απομάκρυνση των δειγμάτων που περιέχουν θόρυβο. Έστω E_i και E_j δύο παραδείγματα που ανήκουν σε διαφορετικές κλάσεις και $d(E_i, E_j)$ η απόσταση μεταξύ τους. Ως Tomek link ορίζεται το $A(E_i, E_j)$, εάν

¹ Χρησιμοποιήθηκε η κλάση `SMOTETomek` από το `imbalanced-learn` API.

δεν υπάρχει παράδειγμα E_k τέτοιο ώστε $d(E_i, E_k) < d(E_i, E_j)$ or $d(E_j, E_k) < d(E_i, E_j)$. Αυτό σημαίνει ότι τα δύο παραδείγματα απέχουν μεταξύ τους περισσότερο από ότι από κάθε άλλο παράδειγμα στο σύνολο δεδομένων. Έτσι αν δύο παραδείγματα σχηματίζουν ένα Tomek link, τότε το ένα από τα δύο είναι θόρυβος ή και τα δύο αποτελούν οριακά σημεία (Batista et al., 2004). Οπότε με αυτήν την μέθοδο αφαιρούνται από την κλάση πλειοψηφίας τα παραδείγματα που θεωρούνται ως θόρυβος.

3.4 Στρωματοποιημένη δειγματοληψία (StratifiedKFold)

Ο διαχωρισμός του συνόλου δεδομένων σε σύνολο εκπαίδευσης και σύνολο ελέγχου αποτελεί σημαντικό κομμάτι της αξιολόγησης των μοντέλων εξόρυξης δεδομένων. Το σύνολο εκπαίδευσης περιέχει δεδομένα των οποίων η κλάση τους είναι γνωστή και χρησιμοποιείται για την εκπαίδευση του κατηγοριοποιητή, με αποτέλεσμα την παραγωγή ενός μοντέλου. Η επίδοση του μοντέλου εκτιμάται χρησιμοποιώντας το σύνολο ελέγχου, στο οποίο οι τιμές των κλάσεων είναι άγνωστες.

Αν το πλήθος των αντικειμένων διαφέρει σημαντικά μεταξύ των κλάσεων, τότε η απλή τυχαία δειγματοληψία μπορεί να οδηγήσει στην παραγωγή συνόλων όπου δεν αντιπροσωπεύονται ικανοποιητικά όλες οι κλάσεις. Για παράδειγμα, αν έχουμε ένα σύνολο από 100 ανθρώπους όπου οι 90 είναι υγιείς και οι 10 ασθενείς, και ορίσουμε το 90% των δεδομένων ως σύνολο εκπαίδευσης και το 10% ως σύνολο ελέγχου, υπάρχει πιθανότητα το σύνολο εκπαίδευσης να αποτελείται μόνο από υγιείς ανθρώπους. Οπότε το μοντέλο θα προβλέπει κάθε νέο δείγμα ανθρώπου ως υγιή.

Για να αντιμετωπισθεί αυτό το πρόβλημα, χρησιμοποιήθηκε η μέθοδος StratifiedKFold του sklearn, το οποίο χωρίζει το σύνολο δεδομένων σε σύνολο εκπαίδευσης και ελέγχου, διατηρώντας όμως την κατανομή των δειγμάτων σε κάθε κλάση.

3.5 Διατροφικοί δείκτες

Μια βιβλιογραφική ανασκόπηση σχετικά με τους διατροφικούς δείκτες που επικρατούν σήμερα, θα επιστρέψει έναν μεγάλο αριθμό δεικτών. Από αυτούς μόνο τέσσερις θεωρούνται βασικοί δείκτες της διατροφικής ποιότητας που έχουν αναφερθεί και επικυρωθεί εκτενέστερα, ενώ οι υπόλοιποι προέκυψαν από τροποποιήσεις που έγιναν πάνω σε αυτούς (Waijers et al., 2007). Οι τέσσερις αυτοί δείκτες είναι ο Healthy Eating Index

(HEI), ο Diet Quality Index (DQI), ο Healthy Diet Indicator (HDI) και ο Mediterranean Diet Score (MDS).

Ο HEI έχει αναφερθεί ότι συσχετίζεται με ένα ευρύ φάσμα βιοδεικτών, οι οποίοι αντιπροσωπεύουν κυρίως συστατικά από φρούτα και λαχανικά, ενώ ο DQI σχετίζεται μόνο οριακά με τη θρεπτική επάρκεια. Επιπλέον ο HDI συσχετίζεται αντιστρόφως με τη θνησιμότητα των ανδρών από τρεις ευρωπαϊκές χώρες και επίσης με τη γνωστική δυσλειτουργία. Τέλος, ο MDS έχει αναφερθεί ότι συσχετίζεται με αυξημένη επιβίωση των πληθυσμών με μεσογειακή διατροφή. Παρόλα αυτά, η συσχέτιση αυτή είναι πιο δυνατή για τους Έλληνες ηλικιωμένους, παρά για τους ηλικιωμένους της Γαλλίας, Ιταλίας, Ολλανδίας ή Γερμανίας.

Ωστόσο στο συγκεκριμένο σύνολο δεδομένων μελετάμε τους εξής δύο διατροφικούς δείκτες: BMI (Body Mass Index - Δείκτης Μάζας Σώματος) και DETERMINE, οι οποίοι αναλύονται στις υποενότητες [3.5.1](#) και [3.5.2](#).

3.5.1 BMI

Ο BMI (Body Mass Index - Δείκτης Μάζας Σώματος) αποτελεί μία γενική ιατρική ένδειξη για τον βαθμό παχυσαρκίας ενός ατόμου. Υπολογίζεται από την μάζα του σώματος (βάρος) και από το ύψος και ορίζεται ως $\frac{\text{βάρος}}{\text{ύψος}^2}$. Η ευκολία του υπολογισμού του τον καθιστά ευρέως διαδεδομένο στον χώρο της ιατρικής και όχι μόνο και αποτελεί διαγνωστικό εργαλείο των προβλημάτων υγείας που μπορεί να επιφέρει η υπερβολική αύξηση ή μείωση του βάρους. Ο BMI υπολογίζει την μάζα ιστού (μυς, λίπος και κόκκαλο) και ανάλογα με την τιμή του, ένας ενήλικας κατηγοριοποιείται σε μια από τις εξής τέσσερις κατηγορίες:

1. ελλιποβαρής: $\text{BMI} < 18.5$
2. κανονικό βάρος: $18.5 < \text{BMI} < 24.9$
3. υπέρβαρος: $25.0 < \text{BMI} < 29.9$
4. παχύσαρκος: $\text{BMI} > 30.0$

3.5.2 DETERMINE

Ο DETERMINE είναι ένας διατροφικός δείκτης που εκτιμά τον διατροφικό κίνδυνο στους ηλικιωμένους. Ο δείκτης αυτός προκύπτει από την λίστα ελέγχου (checklist) “DETERMINE

Your Nutritional Health”², η οποία δημιουργήθηκε από το US Nutrition Screening Initiative (NSI), με σκοπό τον εντοπισμό της επικράτησης του υποσιτισμού μεταξύ των ηλικιωμένων ανθρώπων (Beck et al., 1999). Η λίστα περιέχει 10 ερωτήσεις στις οποίες έχουν ανατεθεί διαφορετικά βάρη και αφορούν καθημερινές συνήθειες διατροφής. Ο τελικός δείκτης BMI προκύπτει από αυτές τις 10 ερωτήσεις και οι τιμές του κυμαίνονται από 0 έως 17. Όσοι συγκεντρώνουν βαθμό μεγαλύτερο της τιμής 6 θεωρείται ότι διατρέχουν υψηλό κίνδυνο υποσιτισμού.

²<https://www.healthcare.uiowa.edu/igec/tools/nutrition/determineNutrition.pdf>

Υλοποίηση και αποτελέσματα

4.1 Συσχέτιση δεικτών με το BMI

Όπως προαναφέρθηκε, η εξαγωγή γνώσης από τα δεδομένα γίνεται με βάση τους διατροφικούς δείκτες BMI και Determine. Όπως είναι ήδη γνωστό, ο δείκτης BMI υπολογίζεται χρησιμοποιώντας το βάρος και το ύψος των ανθρώπων. Αυτό που έχει ενδιαφέρον όμως να δούμε είναι εάν ο BMI έχει κάποια συσχέτιση με τις ερωτήσεις που καθορίζουν το τελικό Determine και κατά συνέπεια με τον ίδιο τον δείκτη Determine. Επιπλέον, είναι σημαντικό να προσδιορίσουμε αν υπάρχουν άλλα χαρακτηριστικά - διατροφικοί δείκτες, εκτός του βάρους και του ύψους, που σχετίζονται περισσότερο ή λιγότερο με το BMI.

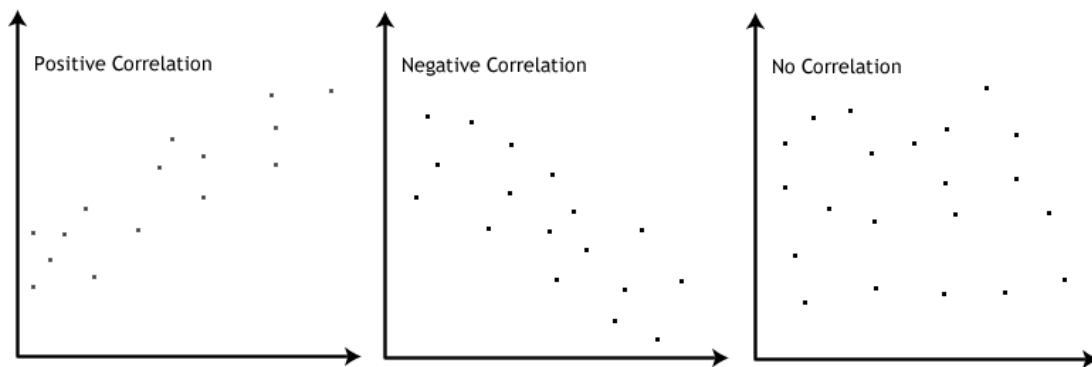
Το σύνολο δεδομένων για το συγκεκριμένο υποερώτημα περιέχει 685 χαρακτηριστικά, εκ των οποίων ως κλάση αναφέρεται ο δείκτης BMI, με εύρος τιμών από 16,2 έως 51,75.

4.1.1 Συσχέτιση Pearson

Για την συσχέτιση των υπολοίπων χαρακτηριστικών με την κλάση BMI χρησιμοποιείται η Συσχέτιση Pearson (Pearson Correlation). Είναι ένα μέτρο που δηλώνει αν υπάρχει γραμμική συσχέτιση μεταξύ δύο χαρακτηριστικών ενός συνόλου δεδομένων, στην συγκεκριμένη περίπτωση μεταξύ της κλάσης BMI και καθενός από τα χαρακτηριστικά του συνόλου. Η συσχέτιση Pearson δύο χαρακτηριστικών X , Y με μέση τιμή $\bar{X}$ και $\bar{Y}$ αντίστοιχα συμβολίζεται ως $\text{corr}(X, Y)$:

$$\text{corr}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (4.1)$$

Η τιμή της συσχέτισης Pearson παίρνει τιμές στο εύρος $[-1, 1]$. Η τιμή -1 δηλώνει αρνητική συσχέτιση, δηλαδή η αύξηση του X επιφέρει μείωση του Y και αντίστροφα, η τιμή $+1$ δηλώνει θετική συσχέτιση, δηλαδή τα χαρακτηριστικά X και Y είναι ανάλογα, ενώ η τιμή 0 δηλώνει ότι τα χαρακτηριστικά είναι ασυσχέτιστα μεταξύ τους. Στο παρακάτω Σχήμα 4.1 παριστάνεται η κατανομή των σημείων στις δύο διαστάσεις σε καθεμία από τις περιπτώσεις που αναφέρθηκαν.



Σχήμα 4.1: Γραφική παράσταση δύο χαρακτηριστικών με θετική συσχέτιση, αρνητική συσχέτιση και καμία συσχέτιση

Έγινε μια προσπάθεια εύρεσης των χαρακτηριστικών και δεικτών που έχουν την μεγαλύτερη συσχέτιση με την κλάση BMI. Αυτό πραγματοποιήθηκε χρησιμοποιώντας την κλάση `CorrelationAttributeEval` του Weka, η οποία μετράει την αξία ενός χαρακτηριστικού, μετρώντας την συσχέτιση Pearson μεταξύ αυτού και της κλάσης. Όμως, πριν την εύρεση της συσχέτισης τα δεδομένα κανονικοποιήθηκαν, ώστε να έχουν μέση τιμή $\mu=0$ και τυπική απόκλιση $\sigma = 1$. Τα αποτελέσματα φαίνονται στον Πίνακα 4.1 για μερικά από τα χαρακτηριστικά του συνόλου, μεταξύ αυτών και ο δείκτης `Determine`.

Στην στήλη `average merit` αναγράφεται η τιμή της συσχέτισης Pearson μεταξύ του κάθε χαρακτηριστικού που αναγράφεται στη δεξιά στήλη και του BMI. Εφόσον αναφε-

ρόμαστε σε πραγματικά δεδομένα τα οποία περιέχουν λάθη καταγραφής, τα ρεαλιστικά όρια της συσχέτισης δεν είναι ακριβώς -1 και +1 αλλά ένα μικρότερο υποσύνολο αυτού. Η στήλη average rank δείχνει την κατάταξη των χαρακτηριστικών σε σχέση με το αποτέλεσμα της συσχέτισης.

Πίνακας 4.1: Αποτελέσματα συσχέτισης Pearson των χαρακτηριστικών με το BMI

Average merit	Average rank	Attribute	Average merit	Average rank	Attribute
0.098 +- 0.004	1 +- 0	F26	0.111 +- 0.006	29.8 +- 5.47	F1
0.081 +- 0.003	2.2 +- 0.4	F15	0.113 +- 0.014	30.2 +- 7.9	L19
0.747 +- 0.004	1 +- 0	WEIGHT	0.11 +- 0.007	30.4 +- 5.35	A11_sex
0.687 +- 0.007	2 +- 0	WAISTCIRCUMFERENCE	0.08 +- 0.007	61.2 +- 7.82	DETERMINE
0.249 +- 0.005	3 +- 0	E2a	0.007 +- 0.006	278.3 +-27.96	L34a
0.235 +- 0.007	4 +- 0	A1_city	0.003 +- 0.004	297.8 +-19.15	L28
0.221 +- 0.01	5.7 +- 0.9	C54_diuretics	0.002 +- 0.005	301.1 +-27.26	C15a
0.219 +- 0.004	5.8 +- 0.6	F24	0 +- 0.006	312.4 +-31.77	L29
0.211 +- 0.006	6.6 +- 0.8	G11	-0.001 +- 0.007	317.7 +-31.14	F40a
0.199 +- 0.006	8 +- 0.45	C1_hypertension	-0.001 +- 0.013	318.4 +-58.46	I1a
0.188 +- 0.005	8.9 +- 0.3	Sum_of_drugs	-0.095 +- 0.007	575.4 +- 4.1	H9
0.168 +- 0.007	10.6 +- 0.8	F20b	-0.101 +- 0.007	579.2 +- 2.52	E1b
0.162 +- 0.009	11.9 +- 1.58	E3b	-0.102 +- 0.005	580.6 +- 4.34	FFQ40
0.157 +- 0.007	14 +- 3	Ave_4_meters	-0.104 +- 0.01	580.7 +- 5.62	hAPAQ14
0.156 +- 0.008	14.1 +- 3.21	C64	-0.103 +- 0.008	580.8 +- 3.63	FFQ5
0.154 +- 0.009	14.9 +- 2.47	L21	-0.11 +- 0.006	583 +- 2	C26e
0.149 +- 0.006	16.1 +- 2.59	E23a	-0.108 +- 0.009	583.5 +- 4.3	FFQ11
0.148 +- 0.007	16.6 +- 2.8	E3a	-0.111 +- 0.009	584.3 +- 4.34	C28
0.147 +- 0.006	16.8 +- 2.52	I11b	-0.11 +- 0.006	584.4 +- 2.33	C26a
0.147 +- 0.008	17.2 +- 3.25	F18	-0.11 +- 0.012	584.5 +- 5.54	FFQ32
0.144 +- 0.01	17.8 +- 2.36	L20	-0.125 +- 0.007	590.3 +- 1.27	E19b
0.145 +- 0.007	18.2 +- 2.93	E11a	-0.131 +- 0.009	591.6 +- 1.02	E19a
0.145 +- 0.01	18.6 +- 2.69	E23b	-0.151 +- 0.012	593.1 +- 1.14	H10
0.132 +- 0.011	22.7 +- 2.87	E2b	-0.154 +- 0.008	593.5 +- 0.5	occupation_type
0.119 +- 0.005	24.7 +- 1.27	G9	-0.191 +- 0.005	595 +- 0	hAPAQ17
0.113 +- 0.011	28.6 +- 5.46	F8	-0.235 +- 0.005	596 +- 0	B8
0.113 +- 0.011	29.3 +- 7.89	I12	-0.254 +- 0.008	597 +- 0	HEIGHT

Από τον πίνακα 4.1 παρατηρούμε ότι τα γνωρίσματα που συσχετίζονται περισσότερο γραμμικά με το BMI είναι τα εξής:

- βάρος
- περιφέρεια της μέσης
- οι δραστηριότητες που κάνει ένας ηλικιωμένος στον ελεύθερό του χρόνο, όπως περπάτημα, αθλητικές δραστηριότητες, ανάγνωση βιβλίων

διότι το average merit είναι μεγαλύτερο του μηδενός και αρκετά κοντά στην τιμή 1. Ο διατροφικός δείκτης Determine έχει βαθμό συσχέτισης 0.08, το οποίο είναι κοντά στο 0, οπότε οι δύο δείκτες δεν συσχετίζονται γραμμικά μεταξύ τους.

Αυτός ο ισχυρισμός έρχεται σε συμφωνία με τον ορισμό του BMI και τον τρόπο που προκύπτει ο δείκτης Determine, καθώς και από τα ευρήματα της αρχικής μελέτης HELIAD, η οποία δεν συσχέτιζε το BMI με την κακή ποιότητα διατροφής και τον υποσιτισμό.

Επιπλέον εμφανίζονται τα χαρακτηριστικά των οποίων η συσχέτιση Pearson έχει αποτέλεσμα κοντά στο 0. Τα χαρακτηριστικά αυτά δεν εμφανίζουν καμία γραμμική συσχέτιση με το BMI και τα περισσότερα από αυτά αποτελούν:

- νευροψυχολογικές διαγνώσεις, όπως άνοια με σωμάτια Lewy (LBD), μετωπο-κροταφική άνοια (FTD), επιληψία, νόσος του Huntington, ψύχωση
- φάρμακα όπως λεκιθίνη, συμπληρώματα οιστρογόνων, φαιτυοΐνη
- γηριατρική κλίμακα κατάθλιψης (GDS)
- νοσοκομειακή μέτρηση άγχους
- παρούσα ψυχιατρική κατάσταση
- νευροψυχιατρικά συμπτώματα (NPI)

Τέλος, υπάρχουν λίγα χαρακτηριστικά τα οποία έχουν αρνητική συσχέτιση με το BMI, δηλαδή όσο αυξάνεται το ένα χαρακτηριστικό τόσο μειώνεται το άλλο και αντίστροφα, όπως για παράδειγμα το ύψος, το οποίο φαίνεται και από τον τύπο υπολογισμού του BMI. Σε αυτήν την κατηγορία ανήκει επίσης και τα εξής χαρακτηριστικά:

- HAPAQ17 (Harokopio Physical Activity Questionnaire), που προκύπτει από ένα ερωτηματολόγιο φυσικής δραστηριότητας του Χαροκοπείου Πανεπιστημίου και σχετίζεται με το περπάτημα τόσο για ψυχαγωγία όσο και για μετακίνηση

- αν ο ηλικιωμένος ετοιμάζει μόνος του το γεύμα
- αν ροχαλίζει συχνά κατά τη διάρκεια του ύπνου
- πόσο συχνά καταναλώνει τσάι ή άλλα αφεψήματα, αρακά, φασολάκια, μπάμιες, αγκινάρες ή δημητριακά

4.1.2 InfoGainAttributeEval

Μία άλλη μέθοδος του Weka που χρησιμοποιήθηκε, το InfoGainAttributeEval(), εκτιμάει την αξία ενός χαρακτηριστικού μετρώντας όχι την γραμμική συσχέτιση, αλλά το κέρδος πληροφορίας που έχει σε σχέση με την κλάση. Στην κατηγοριοποίηση του συνόλου δεδομένων σε σχέση με το BMI δεν είναι σημαντικά όλα τα χαρακτηριστικά του συνόλου. Κύριος στόχος της μεθόδου είναι να βρεθεί ένα υποσύνολο αυτού, που να περιέχει την περισσότερη πληροφορία, δηλαδή να μειώνει τη συνολική εντροπία (αβεβαιότητα) του συνόλου. Η ταξινόμηση των χαρακτηριστικών με βάση το κέρδος πληροφορίας γίνεται με μια ειδική μέθοδο αναζήτησης, που ονομάζεται Ranker.

Το InfoGainAttributeEval υπολογίζεται ως εξής:

$$\text{InfoGain}(\text{Class}, \text{Attribute}) = H(\text{Class}) - H(\text{Class} | \text{Attribute}),$$

όπου $H(\text{Class})$ είναι η εντροπία της κλάσης, $H(\text{Class} | \text{Attribute})$ η εντροπία της κλάσης δοθέντος ενός χαρακτηριστικού και $\text{InfoGain}(\text{Class}, \text{Attribute})$ είναι το κέρδος πληροφορίας του χαρακτηριστικού σε σχέση με την κλάση.

Πίνακας 4.2: Αποτελέσματα μεθόδου InfoGainAttributeEval μεταξύ των χαρακτηριστικών και του BMI

Information Gain	Attributes	Information Gain	Attributes
0.25311	WEIGHT	0.0238	J37i
0.19897	WAISTCIRCUMFERENCE	0.0238	J37f
0.03558	A1_city	0.0238	J34h
0.02804	B8_years_education_self	0.0238	J39i
0.02689	J38g	0.0238	J41f
0.02536	J42	0.02365	J36d
0.0253	J43f	0.02316	J35j
0.02512	J42i	0.0231	J36
0.02486	J41d	0.02302	J43e
0.02486	J32d	0.02302	J40g
0.02486	J33f	0.02297	J42e
0.02486	J33e	0.02288	J32j
0.02486	J34f	0.02272	G11
0.02475	J38j	0.02268	J35e
0.02472	J43	0.02263	J35h
0.02444	J33	0.02261	J36e
0.02439	J42d	0.02243	J43h
0.02392	J43d	0.02238	J35g
0.02389	J38	0.02229	J37d

Από τον Πίνακα 4.2 φαίνεται ότι τα χαρακτηριστικά που περιέχουν την περισσότερη πληροφορία είναι το βάρος, η περιφέρεια μέσης και τα νευροψυχιατρικά συμπτώματα (NPI), όπως παραληρήματα, ψευδαισθήσεις, επιθετικότητα, κατάθλιψη, άγχος, απάθεια, ευερεθιστότητα, παθολογική κινητική συμπεριφορά, διαταραχές συμπεριφοράς τη νύχτα, διαταραχές της διατροφής.

Το γεγονός ότι τα παραπάνω χαρακτηριστικά περιέχουν την περισσότερη πληροφορία σε σχέση με το BMI, σε συνδυασμό με το γεγονός ότι είναι γραμμικά ασυσχέτιστα με το BMI, τα καθιστά κατάλληλα για τον διαχωρισμό των εγγραφών σε μία κατηγορία του BMI.

4.2 PCA ανάλυση με βάση το BMI

Σε αυτήν την υποενότητα θα γίνει μια PCA ανάλυση όλων των χαρακτηριστικών του συνόλου δεδομένων, χρησιμοποιώντας ως κλάση τον δείκτη BMI. Η ανάλυση θα μας βοηθήσει να εντοπίσουμε το πλήθος των συνιστωσών που αντιπροσωπεύουν καλύτερα τα δεδομένα και που περιέχουν το μεγαλύτερο ποσοστό διακύμανσης, μειώνοντας παράλληλα τον αριθμό των χαρακτηριστικών που υπάρχουν στο σύνολο δεδομένων.

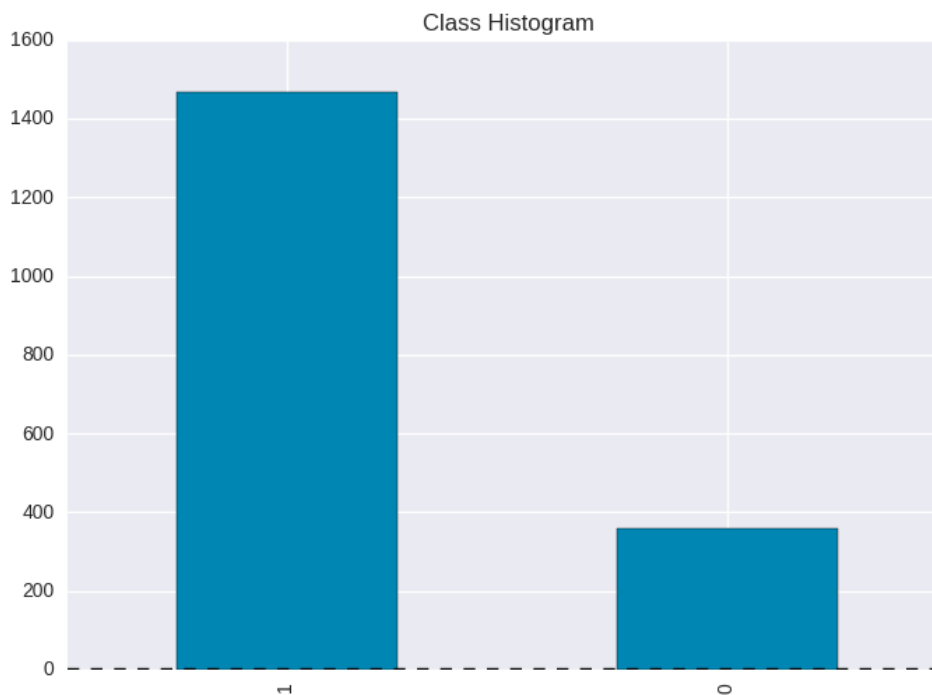
4.2.1 Περιγραφή της διαδικασίας

Το BMI διακρίνεται στις εξής τέσσερις κατηγορίες: ελλιποβαρής, κανονικό βάρος, υπέρβαρος ή παχύσαρκος, αλλά αυτό που μας ενδιαφέρει είναι να εξετάσουμε τις περιπτώσεις όπου ένας συγκεκριμένος άνθρωπος είναι είτε λιποβαρής/κανονικό βάρος είτε υπέρβαρος/παχύσαρκος. Οπότε η μεταβλητή BMI μετατρέπεται σε δυαδική, με τιμές 0 και 1 αντίστοιχα, με αποτέλεσμα να έχουμε δυαδική κατηγοριοποίηση. Επιπλέον από το σύνολο δεδομένων αφαιρέθηκε ο δείκτης Determine, διότι δεν συσχετίζεται με τον BMI, καθώς και τα γνωρίσματα ύψος και βάρος, που γνωρίζουμε εξ ορισμού ότι χρησιμοποιούνται για τον υπολογισμό του συγκεκριμένου δείκτη.

Το ιστόγραμμα της δίτιμης κλάσης στο σύνολο δεδομένων φαίνεται στο Σχήμα 4.2. Παρατηρούμε ότι η κατανομή των εγγραφών σε καθεμία από τις τιμές 0 και 1 της κλάσης BMI είναι άνιση, δηλαδή υπάρχουν πολλά περισσότερα δείγματα που ανήκουν στην κλάση με τιμή 1 (υπέρβαρος) σε σχέση με αυτά που ανήκουν στην κλάση με τιμή 0 (λιποβαρής/κανονικό βάρος). Συγκεκριμένα, η αναλογία των δειγμάτων στις κλάσεις είναι 4:1.

Για να αντιμετωπισθεί αυτό το πρόβλημα, χρησιμοποιήθηκε η μέθοδος SMOTETomek, με την οποία δημιουργήθηκαν συνθετικά δείγματα στην κλάση 0 και διαγράφηκαν όσα δείγματα από την κλάση 1 θεωρήθηκαν Tomek links, δηλαδή όσα αποτελούν θόρυβο ή είναι οριακά σημεία. Το αποτέλεσμα είναι η αύξηση του αριθμού των δειγμάτων που ανήκουν στην κλάση μειοψηφίας (0), οπότε η αναλογία των κλάσεων είναι σχεδόν 1:1. Στην συνέχεια, χρησιμοποιήθηκε η μέθοδος του sklearn StratifiedKFold, η οποία χωρίζει το σύνολο δεδομένων σε σύνολο εκπαίδευσης και ελέγχου, όπου και οι δύο κλάσεις αντιπροσωπεύονται ικανοποιητικά.

Ένα σημαντικό βήμα που απαιτείται πριν την PCA ανάλυση είναι η κανονικοποίηση



Σχήμα 4.2: Ιστόγραμμα του δείκτη BMI

(normalization) των χαρακτηριστικών του συνόλου δεδομένων έτσι ώστε να έχουν μέση τιμή $\mu = 0$ και τυπική απόκλιση $\sigma = 1$, δηλαδή τα δεδομένα να ακολουθούν την κανονική κατανομή. Η τοποθέτηση των τιμών γύρω από την μέση τιμή και η κλιμάκωση πραγματοποιούνται ξεχωριστά σε κάθε χαρακτηριστικό υπολογίζοντας τα παραπάνω στατιστικά μεγέθη στα δείγματα του συνόλου εκπαίδευσης. Για την κανονικοποίηση των χαρακτηριστικών του συνόλου εκπαίδευσης και ελέγχου χρησιμοποιήθηκε η μέθοδος `StandardScaler` από το `sklearn`.

Οπότε για την PCA ανάλυση είναι απαραίτητη η κανονικοποίηση και των δύο συνόλων διότι αν κάποια χαρακτηριστικά έχουν μεγάλη διακύμανση ενώ κάποια άλλα μικρή, ο αλγόριθμος PCA θα επιλέξει τα σημεία εκείνα τα οποία έχουν την μέγιστη διακύμανση και η συμπεριφορά του δεν θα είναι η βέλτιστη.

4.2.2 Εύρεση ιδανικού αριθμού συνιστωσών για την PCA ανάλυση

Για την εύρεση του ιδανικού αριθμού συνιστωσών χρησιμοποιήθηκε ο αλγόριθμος PCA με την μέθοδο `GridSearchCV` του `sklearn`, η οποία οπτικοποιεί δύο παραμέτρους κάνοντας εξαντλητική αναζήτηση σε ένα δισδιάστατο πλέγμα. Ο χρήστης μπορεί να ορί-

σει το χαμηλότερο και υψηλότερο όριο για κάθε παράμετρο, και τον επιθυμητό αριθμό των επαναλήψεων. Η μέθοδος εκτελεί μια γρήγορη αρχική εκτίμηση του πλέγματος χρησιμοποιώντας two-fold cross validation (διασταυρωμένη επικύρωση σε δύο μέρη) στο σύνολο εκπαίδευσης. Ακολουθώντας, το καλύτερο σημείο του πλέγματος ερευνάται πιο προσεκτικά χρησιμοποιώντας αναζήτηση hill-climbing, όπου γίνεται 10-fold cross validation. Εάν στην πορεία κάποιο άλλο σημείο αποδειχθεί καλύτερο, τότε γίνεται το καινούριο κεντρικό σημείο και η διασταυρωμένη επικύρωση συνεχίζεται, ώσπου να μην υπάρχει άλλο καλύτερο σημείο από το ήδη υπάρχον (Witten and Frank, 2005).

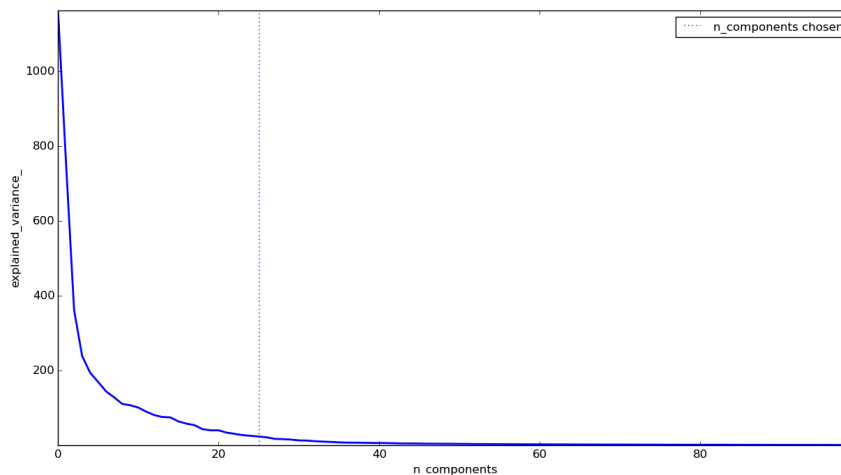
Μία από τις παραμέτρους που είναι σημαντική για την PCA ανάλυση είναι το πλήθος των συνιστωσών (`n_components`), το οποίο πρέπει να είναι τουλάχιστον μικρότερο ή ίσο του συνολικού αριθμού των χαρακτηριστικών που περιέχονται στο σύνολο δεδομένων. Όπως φαίνεται στο παρακάτω κομμάτι κώδικα, το εύρος της μεταβλητής `n_components` είναι από 1 έως 1000. Το μοντέλο πρόβλεψης δημιουργείται στην συγκεκριμένη περίπτωση από τον κατηγοριοποιητή Logistic Regression (λογιστική παλινδρόμηση), το οποίο δημιουργείται χρησιμοποιώντας το σύνολο εκπαίδευσης και στη συνέχεια κατηγοριοποιεί τα άγνωστα δεδομένα του συνόλου ελέγχου. Τέλος, η μέθοδος `GridSearchCV` θα επιλέξει τον κατάλληλο αριθμό συνιστωσών που θα χρησιμοποιηθεί από το PC, επιλέγοντας αυτόν που θα δώσει την μεγαλύτερη ακρίβεια στο σύνολο ελέγχου.

```
1 pipe = Pipeline(steps=[('pca', PCA()), ('logistic', LogisticRegression())])
2 n_components = [1,2,5,10,15,20,25,30,40,50,60,80,100,150,200,300,400,500,800,1000]
3 Cs = np.logspace(-4, 4, 3)
4
5 estimator = GridSearchCV(pipe, dict(pca__n_components = n_components, logistic__C = Cs))
```

Ο αριθμός των συνιστωσών μπορεί να καθοριστεί χρησιμοποιώντας πολλές μεθόδους. Στο παρακάτω γράφημα χρησιμοποιείται το scree test, που είναι μια τεχνική καθορισμού του αριθμού των συνιστωσών που διατηρούνται από μια ανάλυση πρωταρχικών συνιστωσών ή παραγοντική ανάλυση. Η τεχνική PCA υπολογίζει την διακύμανση για κάθε μία συνιστώσα.

Στον οριζόντιο άξονα τοποθετείται το πλήθος των συνιστωσών ενώ στον κάθετο οι τιμές της διακύμανσης για κάθε μία από τις συνιστώσες. Το γράφημα απεικονίζει την σχέση μεταξύ της διακύμανσης και του πλήθους των συνιστωσών. Αρχικά παρατηρούμε ότι η καμπύλη είναι απότομη ενώ στην συνέχεια η κλίση της γίνεται πιο ομαλή, μέχρι που

καταλήγει σε οριζόντια γραμμή. Ο ιδανικός αριθμός των συνιστωσών καθορίζεται από το σημείο εκείνο στο οποίο σταματά η απότομη καμπύλη. Στο Σχήμα 4.3 ο ιδανικός αριθμός συνιστωσών φαίνεται να είναι 25.



Σχήμα 4.3: Scree plot για την εύρεση του ιδανικού αριθμού συνιστωσών

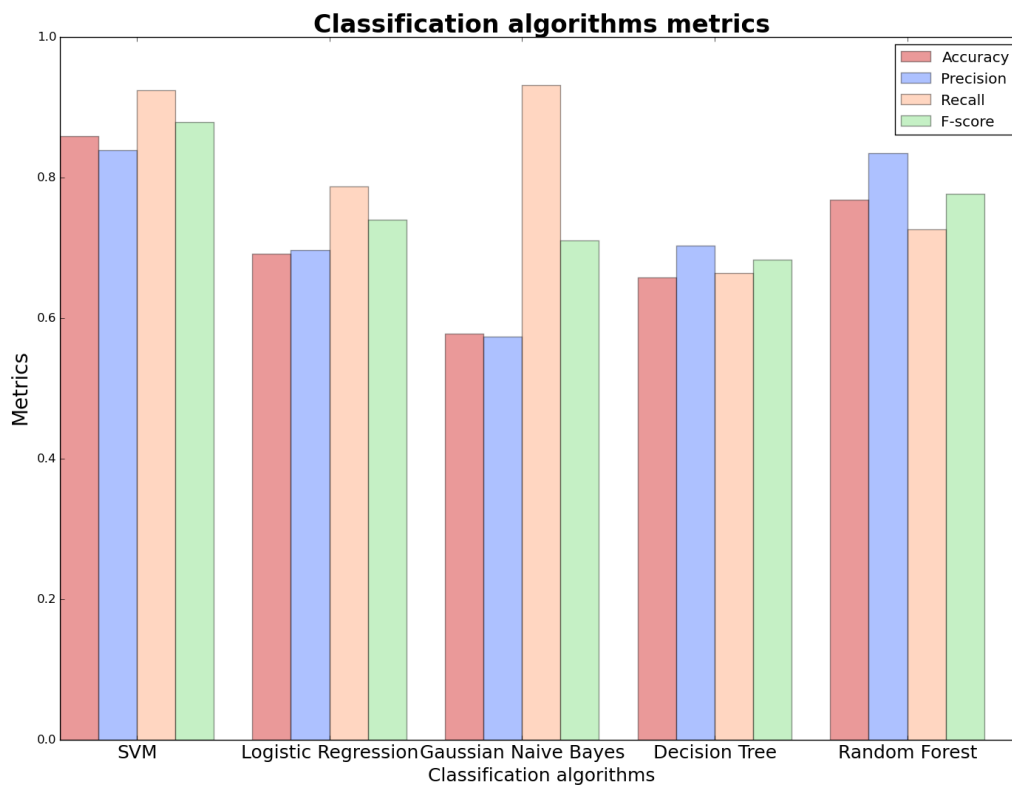
Το επόμενο βήμα είναι να γίνει κατηγοριοποίηση του συνόλου χρησιμοποιώντας τις 25 συνιστώσες που βρέθηκαν με την PCA ανάλυση, αντί για τα 609 χαρακτηριστικά του συνόλου δεδομένων. Οι πέντε κατηγοριοποιητές που θα συγκριθούν είναι οι εξής: SVM, Logistic regression, Gaussian Naive Bayes, Decision tree και Random forest.

Στον Πίνακα 4.3 διακρίνονται οι τιμές των μετρικών αξιολόγησης accuracy, precision, recall και f-score με το διάστημα εμπιστοσύνης στο σύνολο ελέγχου για κάθε έναν από τους παραπάνω αλγορίθμους, με συντελεστή εμπιστοσύνης 95%. Όπως παρατηρούμε, ο SVM και ο Random Forest εμφανίζουν υψηλότερες τιμές στις μετρικές σε σχέση με τους υπόλοιπους τρεις αλγορίθμους, οπότε μπορούν να προβλέψουν με μεγαλύτερο ποσοστό επιτυχίας σε ποιά κλάση θα ανήκει ένα καινούριο δείγμα του συνόλου ελέγχου.

Πίνακας 4.3: Αποτελέσματα μετρικών για κάθε κατηγοριοποιητή, έχοντας ως κλάση το BMI

Algorithm / Metrics	Accuracy	Precision	Recall	F-score
SVM	88.21 $\pm$ 2.56	86.16 $\pm$ 2.74	93.84 $\pm$ 1.91	89.84 $\pm$ 2.4
Logistic regression	71.48 $\pm$ 3.59	72.05 $\pm$ 3.56	79.45 $\pm$ 3.21	75.57 $\pm$ 3.41
Gaussian Naive Bayes	60.46 $\pm$ 3.88	59.05 $\pm$ 3.91	93.84 $\pm$ 1.91	72.49 $\pm$ 3.55
Decision tree	70.72 $\pm$ 3.61	77.17 $\pm$ 3.33	67.12 $\pm$ 3.73	71.79 $\pm$ 3.57
Random forest	80.99 $\pm$ 3.12	88.10 $\pm$ 2.57	76.03 $\pm$ 3.39	81.62 $\pm$ 3.08

Ένα σχηματικό διάγραμμα του παραπάνω πίνακα είναι το Σχήμα 4.4:



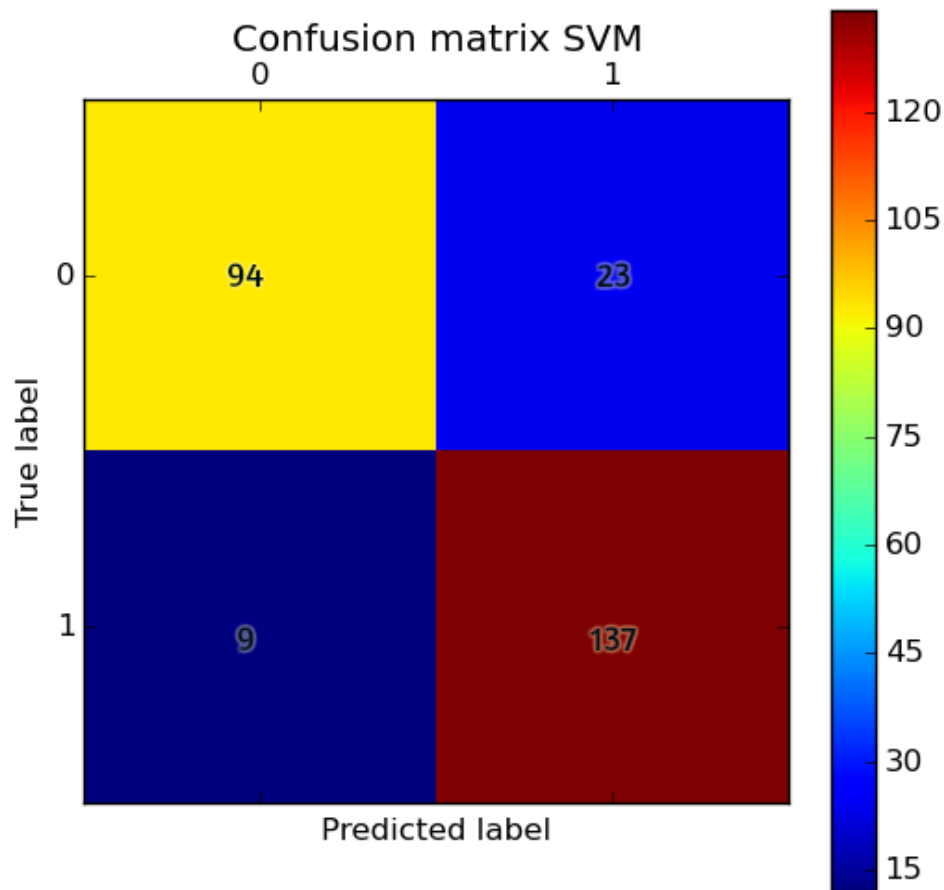
Σχήμα 4.4: Διάγραμμα απεικόνισης των μετρικών για κάθε κατηγοριοποιητή, όταν η κλάση είναι το BMI

4.2.3 Οπτικοποίηση της απόδοσης

Σε αυτήν την υποενότητα θα παρουσιαστούν οι τεχνικές αξιολόγησης και οπτικοποίησης των αλγορίθμων που χρησιμοποιήθηκαν στην παραπάνω ανάλυση. Οι μετρικές αξιο-

λόγησης χρησιμοποιήθηκαν παραπάνω για την σύγκριση των αλγορίθμων μεταξύ τους, έτσι ώστε να βρεθεί ποιος προβλέπει καλύτερα. Δηλαδή ποιος βρίσκει με μεγαλύτερη ποσοστό, σε ποια κλάση ανήκει ένα άγνωστο δείγμα του συνόλου ελέγχου. Ωστόσο, χρειαζόμαστε επιπλέον μέτρα αξιολόγησης, έτσι ώστε να ελέγχουμε την ποιότητα του αποτελέσματος που εξήγαγε ο αλγόριθμος και αν υποπροσαρμόζεται (underfitting) ή υπερπροσαρμόζεται (overfitting) πάνω στα δεδομένα εκπαίδευσης. Η αξιολόγηση γίνεται με τον πίνακα σύγχυσης και την καμπύλη ROC.

Στο Σχήμα 4.5 φαίνεται ο πίνακας σύγχυσης για τον αλγόριθμο SVM. Κάθε γραμμή του πίνακα αναπαριστά τα δείγματα που ανήκουν στην πραγματικότητα στην κλάση 0 και 1 αντίστοιχα, ενώ κάθε στήλη αναπαριστά το πλήθος των δειγμάτων που προβλέφθηκαν ότι ανήκουν σε κάθε μια από τις κλάσεις.

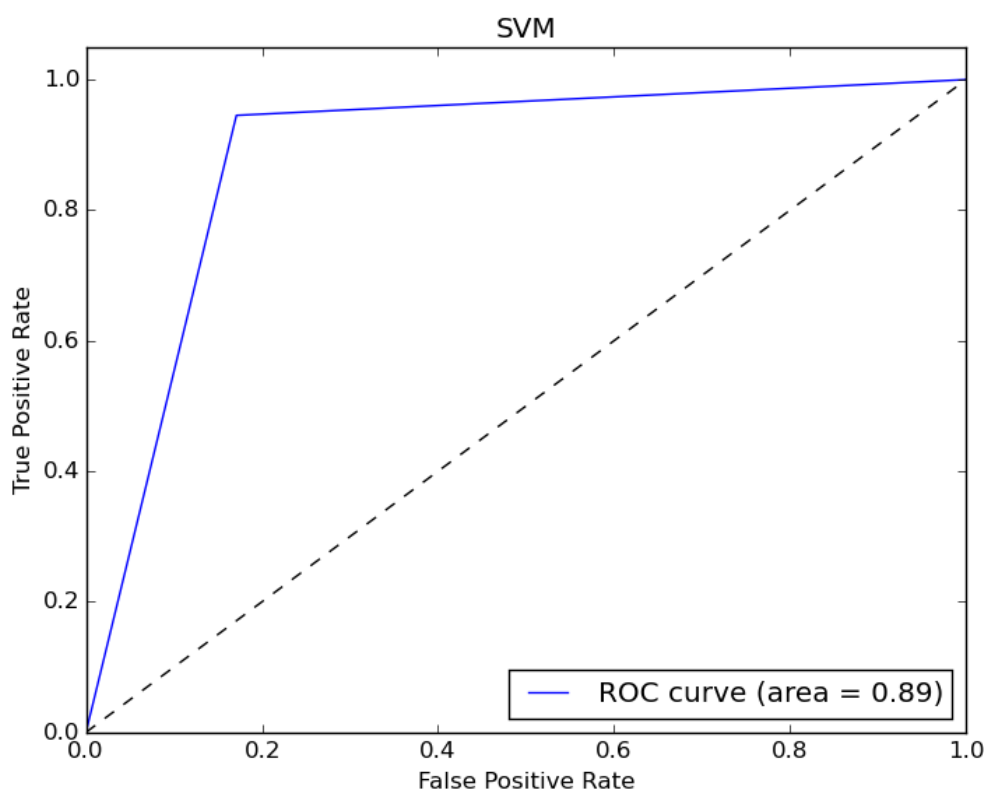


Σχήμα 4.5: Πίνακας σύγχυσης για τον αλγόριθμο SVM

Συνεπώς, προκύπτει ότι 94 από τα δείγματα που ανήκουν στην κλάση 0, κατηγοριοποιήθηκαν όντως στην κλάση 0 και 137 δείγματα που ανήκουν στην κλάση 1 κατηγοριοποιήθηκαν αντίστοιχα στην κλάση 1. Αυτές οι τιμές συγκροτούν τη διαγώνιο του πίνακα και όσο πιο μεγάλες είναι, τόσο καλύτερα είναι τα αποτελέσματα του αλγορίθμου. Αντιθέτως, οι τιμές στα υπόλοιπα δύο κελιά ιδανικά θα θέλαμε να ήταν μηδέν. Στην συγκεκριμένη περίπτωση, ο αλγόριθμος κατηγοριοποίησε λάθος 9 δείγματα του συνόλου στην κλάση 0, ενώ στην πραγματικότητα ανήκουν στην κλάση 1, και αντίστοιχα ότι 23 δείγματα ανήκουν στην κλάση 1, ενώ δεν ανήκουν σε αυτήν.

Η καμπύλη ROC είναι ένα γραφικό διάγραμμα που απεικονίζει την απόδοση ενός δυα-

δικού κατηγοριοποιητή. Δημιουργείται αναπαριστώντας στον οριζόντιο άξονα το False Positive Rate και στον κάθετο το True Positive Rate - ή αλλιώς Recall. Όπως παρατηρούμε στο Σχήμα 4.6, η καμπύλη βρίσκεται πάνω από την διαγώνιο και η περιοχή κάτω από την καμπύλη είναι 0.89, οπότε ο SVM έχει υψηλό ποσοστό TP, TN και χαμηλό ποσοστό FP, FN, γεγονός που τον καθιστά καλό κατηγοριοποιητή για το συγκεκριμένο σύνολο ελέγχου.



Σχήμα 4.6: Καμπύλη ROC για τον αλγόριθμο SVM

4.3 Συσχέτιση χαρακτηριστικών με τον δείκτη Determine

4.3.1 CfsSubsetEval

Σε αυτήν την υποενότητα χρησιμοποιήθηκε η κλάση CfsSubsetEval από το Weka, έτσι ώστε να δούμε με ποια χαρακτηριστικά του συνόλου δεδομένων συσχετίζεται περισσότερο ο δείκτης Determine. Η κλάση αυτή μετράει την αξία ενός υποσυνόλου των χαρακτηριστικών, λαμβάνοντας υπόψιν την ικανότητα πρόβλεψης κάθε ξεχωριστού χαρακτηριστικού καθώς και τον βαθμό πλεονασμού (degree of redundancy) των χαρακτηριστι-

κών μεταξύ τους. Το υποσύνολο που επιλέγεται περιέχει χαρακτηριστικά που σχετίζονται στενά με το Determine, δηλαδή με το εάν ένας ηλικιωμένος βρίσκεται ή όχι σε κίνδυνο υποσιτισμού, αλλά καθόλου μεταξύ τους.

Το υποσύνολο των χαρακτηριστικών που επιλέχθηκε φαίνεται στον Πίνακα 4.4 κατά αλφαβητική σειρά. Τα χαρακτηριστικά αυτά θα χωριστούν στις παρακάτω κατηγορίες, για να διευκολυνθεί η περιγραφή τους.

- Δημογραφικά χαρακτηριστικά του ηλικιωμένου, όπως η πόλη όπου διαμένει, τα χρόνια εκπαίδευσης του ίδιου και των γονέων του, το αν είναι χήρος ή παντρεμένος και για πόσα χρόνια, το αν διαθέτει μόνιμη κατοικία, αυτοκίνητο και αν πηγαίνει για διακοπές.
- Χαρακτηριστικά σχετικά με το ατομικό ιατρικό ιστορικό, όπως αν υπάρχει διάγνωση για στεφανιαία νόσο, τα έτη καπνίσματος όσον αφορά τους καπνιστές, ή αν παίρνουν φάρμακα.
- Δραστηριότητες που κάνουν στον ελεύθερο χρόνο. Αυτές περιλαμβάνουν φυσική άσκηση, ενασχόληση με τις τέχνες και τον πολιτισμό, προσφορά εθελοντικής εργασίας, συμμετοχή σε διάφορα μαθήματα, διατήρηση εργασίας με αμοιβή, επίσκεψη σε μουσεία, συναναστροφή με συγγενείς και φίλους.
- Γηριατρική κλίμακα κατάθλιψης. Περιλαμβάνονται ερωτήσεις προς τον ηλικιωμένο όπως αν νιώθει αβοήθητος, άχρηστος, γεμάτος ενέργεια, αν προτιμάει να μείνει σπίτι από το βγαίνει έξω, αν έχει ψευδαισθήσεις.
- Μέτρηση της σωματικής λειτουργικότητας, εξέταση του κινητικού συστήματος και ιδιοπαθούς τρόμου.
- Συχνότητα κατανάλωσης τροφίμων όπως μοσχάρι, μαρούλι, λάχανο, σπανάκι, μήλο αχλάδι, ελαιόλαδο, και συγκεκριμένα αν γίνεται κατανάλωση και των τριών κυρίων γευμάτων.
- Αν γίνεται χρήση μασέλας στο φαγητό και αν γίνεται κατανάλωση γευμάτων με παρέα
- Περπάτημα για ψυχαγωγία και μετακίνηση, αλλά και αν γίνεται χρήση του υπολογιστή για διάβασμα.

Πίνακας 4.4: Υποσύνολο των δεδομένων που προέκυψε από το CfsSubsetEval. Τα χαρακτηριστικά συσχετίζονται περισσότερο με το Determine και σχεδόν καθόλου μεταξύ τους

A1_city	E8a	L17_2a
B8_years_education_self	E11a	Average_4_meters
B9_years_education_father	E12b	L30c
B10_years_education_mother	E15a	L47
B13a_married	E22b	G6
B13b_married_years	E23a	G9
B13e_widower	E23b	FFQ0
B14c_residence_area	E25a	FFQ18
B15_num_cars	E26	FFQ34
B16_country_house	F29	FFQ38
B17_summer_holidays	J8	FFQ66
B17b_holidays_location	J9	FFQ72
C3_coronary_disease	J12	FFQ73
C32_years_smoker	J13	FFQ74
C78_H-1_blockers_PPI	J14	DENTAL_PLATE
C81_narcotics	J15	EATING_IN_COMPANY
Sum_of_drugs	J29a	HEIGHT
E3b	L12	hAPAQ8
E6a	L16_1a	hAPAQ14
E7b	time_for_first_meter	hAPAQ17

4.3.2 InfoGainAttributeEval

Σε αυτήν την υποενότητα χρησιμοποιήθηκε επίσης η κλάση InfoGainAttributeEval από το Weka, ώστε να βρεθούν τα χαρακτηριστικά που έχουν το μεγαλύτερο κέρδος πληροφορίας σε σχέση με τον δείκτη υποσιτισμού Determine. Το κέρδος πληροφορίας ενός χαρακτηριστικού υπολογίζεται αν από την συνολική εντροπία του συνόλου δεδομένων αφαιρέσουμε την εντροπία του χαρακτηριστικού σε σχέση με το Determine.

Στον Πίνακα 4.5 φαίνεται ένα υποσύνολο των χαρακτηριστικών που βρέθηκαν με την κλάση InfoGainAttributeEval, τα οποία ταξινομούνται σε σχέση με το κέρδος πληροφορίας. Παρατηρούμε ότι και σε αυτήν την περίπτωση, τα χαρακτηριστικά που θεωρούνται

σημαντικά για τον δείκτη υποσιτισμού είναι η χρήση μασέλας κατά τη διάρκεια του φαγητού, η κατανάλωση φαγητού με παρέα, τα φάρμακα που λαμβάνει και τα χρόνια εκπαίδευσης του ηλικιωμένου.

Όλα τα υπόλοιπα χαρακτηριστικά που συσχετίζονται με τον δείκτη υποσιτισμού ανήκουν στην κατηγορία των Νευροψυχιατρικών συμπτωμάτων (NPI). Τα χαρακτηριστικά αυτά αναφέρουν αν ο ηλικιωμένος είχε τον τελευταίο μήνα διαταραχές στην διατροφή του, δηλαδή αυξημένη / μειωμένη όρεξη για φαγητό, απώλεια / αύξηση σωματικού βάρους, αλλαγές στις προτιμήσεις και συνήθειες διατροφής. Επίσης μεγάλο κέρδος πληροφορίας εμφανίζουν τα χαρακτηριστικά που δηλώνουν απάθεια του ηλικιωμένου, δηλαδή αν δείχνει λιγότερο ενδιαφέρον για την οικογένεια και τους φίλους του, αν συμμετέχει λιγότερο στο σπίτι και αν δυσκολεύεται να ανοίξει συζητήσεις.

Άλλα χαρακτηριστικά σε αυτήν την κατηγορία είναι οι διαταραχές συμπεριφοράς που μπορεί να παρουσιάζει τη νύχτα, η παθολογική κινητική συμπεριφορά, η ευερεθιστότητα, οι ψευδαισθήσεις. Πιο συγκεκριμένα, αν μιλάει με ανύπαρκτα πρόσωπα ή βλέπει πράγματα που δεν υπάρχουν ή αν είναι αυτοκτονικός.

Πίνακας 4.5: Αποτελέσματα της κλάσης InfogainAttributeEval για τη συσχέτιση των χαρακτηριστικών με τον δείκτη υποσιτισμού

Average merit	Average rank	Attribute
0.241 +- 0.004	1 +- 0	DENTAL_PLATE
0.158 +- 0.005	2 +- 0	DENTAL_PLATE_FOOD
0.127 +- 0.006	3 +- 0	Sum_of_drugs
0.079 +- 0.002	4.7 +- 0.78	B8_years_education_self
0.079 +- 0.002	5 +- 0.89	hAPAQ14
0.076 +- 0.004	6.5 +- 2.46	EATING_IN_COMPANY
0.071 +- 0.004	6.8 +- 0.75	J43
0.07 +- 0.004	8.4 +- 0.66	J38
0.069 +- 0.004	9.1 +- 1.22	J43f
0.069 +- 0.005	10.7 +- 3.26	J42e
0.068 +- 0.004	11.6 +- 2.33	J38g
0.067 +- 0.004	14.7 +- 3.8	J41
0.067 +- 0.004	16 +- 6.97	J38j
0.067 +- 0.004	16.3 +- 5.83	J42d
0.067 +- 0.004	16.6 +- 3.95	J38d
0.067 +- 0.004	17.8 +- 6.34	J43i
0.067 +- 0.004	17.8 +- 7.59	J33
0.067 +- 0.004	20.8 +- 7.91	J43h
0.066 +- 0.004	20.9 +- 4.16	J38i
0.066 +- 0.004	22.4 +- 5.48	J43d
0.066 +- 0.004	23.5 +- 9.39	J43e
0.066 +- 0.004	23.8 +- 9.69	J33e
0.066 +- 0.004	24.8 +- 7.07	J40
0.067 +- 0.004	24.8 +-18.11	J42
0.066 +- 0.004	25.1 +- 3.08	J32
0.066 +- 0.004	27.4 +- 2.84	J35j
0.066 +- 0.004	28.1 +- 5.03	J33d
0.066 +- 0.004	29.6 +- 4.98	J41i
0.066 +- 0.004	29.6 +- 6.76	J38e
0.066 +- 0.004	30.2 +- 6.1	J35g
0.066 +- 0.004	34.3 +- 8.66	J41e
0.066 +- 0.004	34.5 +-10.35	J35d
0.065 +- 0.004	34.7 +- 5.88	J37e

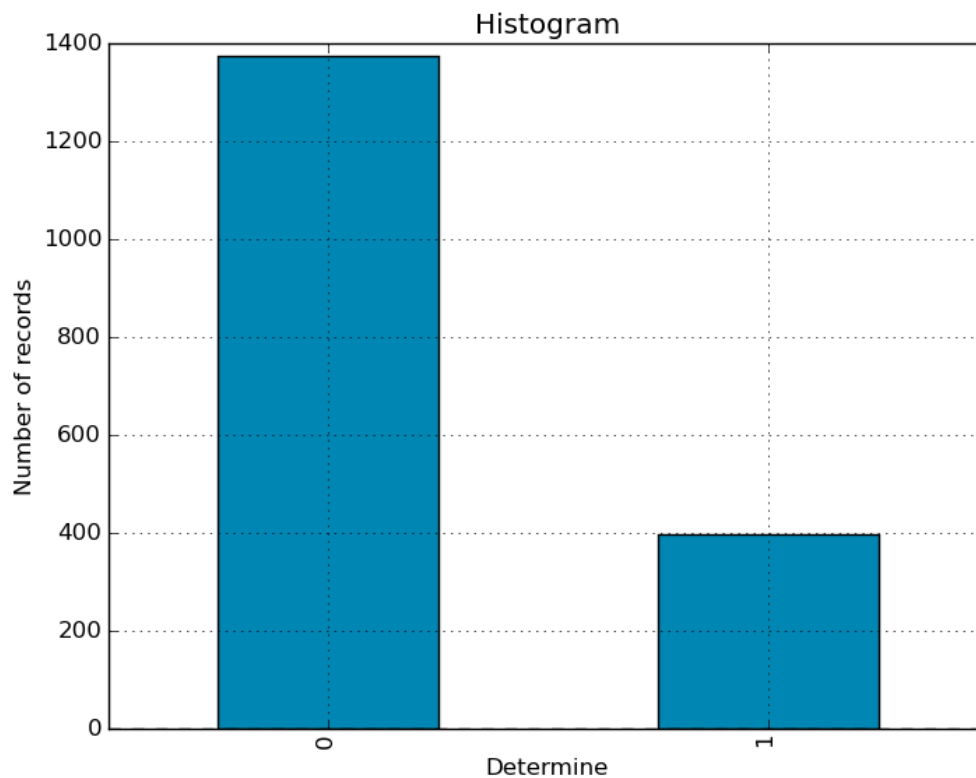
4.4 PCA ανάλυση με βάση το Determine

Στην συγκεκριμένη υποενότητα πραγματοποιείται μια PCA ανάλυση όλων των χαρακτηριστικών του συνόλου δεδομένων με βάση τον διατροφικό δείκτη Determine. Σκοπός της ανάλυσης είναι η εύρεση των βασικών συνιστωσών που καθορίζουν τον υποσιτισμό στον συγκεκριμένο πληθυσμό. Το εύρος των τιμών του δείκτη Determine κυμαίνεται από 0 έως 17 και διακρίνεται στις εξής κατηγορίες ανάλογα με την τιμή που παίρνει:

- 0 - 2: Ο ασθενής δεν διατρέχει διατροφικό κίνδυνο
- 3 - 5: Ο ασθενής βρίσκεται σε μέτριο διατροφικό κίνδυνο
- > 6: Ο ασθενής βρίσκεται σε υψηλό διατροφικό κίνδυνο
-

Στην παρούσα μελέτη όμως μας ενδιαφέρει απλά αν ένας ηλικιωμένος διατρέχει κίνδυνο υποσιτισμού ή όχι, οπότε θα συγχωνευτούν οι τιμές από 0 έως και 5 σε μία. Έτσι ο δείκτης Determine θα μετατραπεί από συνεχής αριθμός σε δυαδικός, λαμβάνοντας τις εξής δύο τιμές: 0 (υποσιτισμένος) και 1 (μη υποσιτισμένος).

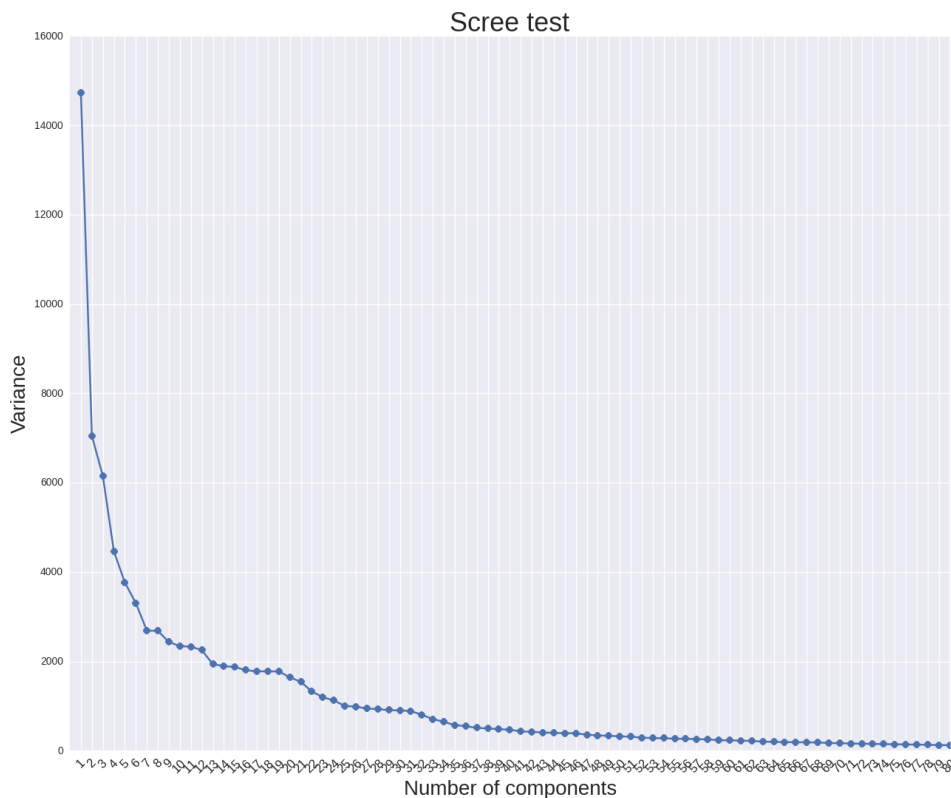
Παρατηρούμε από το ιστόγραμμα στο Σχήμα 4.7 ότι και σε αυτήν την περίπτωση το σύνολο δεδομένων είναι ασύμμετρο, διότι οι εγγραφές που ανήκουν στην κλάση 1 είναι πολύ λιγότερες σε σχέση με αυτές που ανήκουν στην κλάση 0. Επομένως, είναι και πάλι επιτακτική η ανάγκη αύξησης του συνόλου δεδομένων χρησιμοποιώντας την μέθοδο SMOTETomek, για την δημιουργία συνθετικών εγγραφών στην κλάση 1 και διαγραφή των εγγραφών που αποτελούν θόρυβο από την κλάση 0.



Σχήμα 4.7: Ιστόγραμμα του δείκτη Determine

Στη συνέχεια, χωρίζουμε το σύνολο δεδομένων σε σύνολο εκπαίδευσης και ελέγχου με την μέθοδο `StratifiedKFold` της βιβλιοθήκης `Scikit-learn`, ώστε και τα δύο σύνολα να έχουν την ίδια κατανομή των εγγραφών και στις δύο κλάσεις. Όπως και προηγουμένως, και τα δύο σύνολα κανονικοποιήθηκαν χρησιμοποιώντας επίσης την μέθοδο `StandardScaler` της `Scikit-learn`, για να ακολουθείται η κανονική κατανομή.

Από το `scree test` που απεικονίζεται στο παρακάτω Σχήμα 4.8, συμπεραίνουμε ότι ο ιδανικός αριθμός συνιστωσών για την PCA ανάλυση είναι 30, διότι σε αυτό το σημείο η καμπύλη κατεβαίνει με πιο αργό ρυθμό και τελικά καταλήγει σε οριζόντια γραμμή.

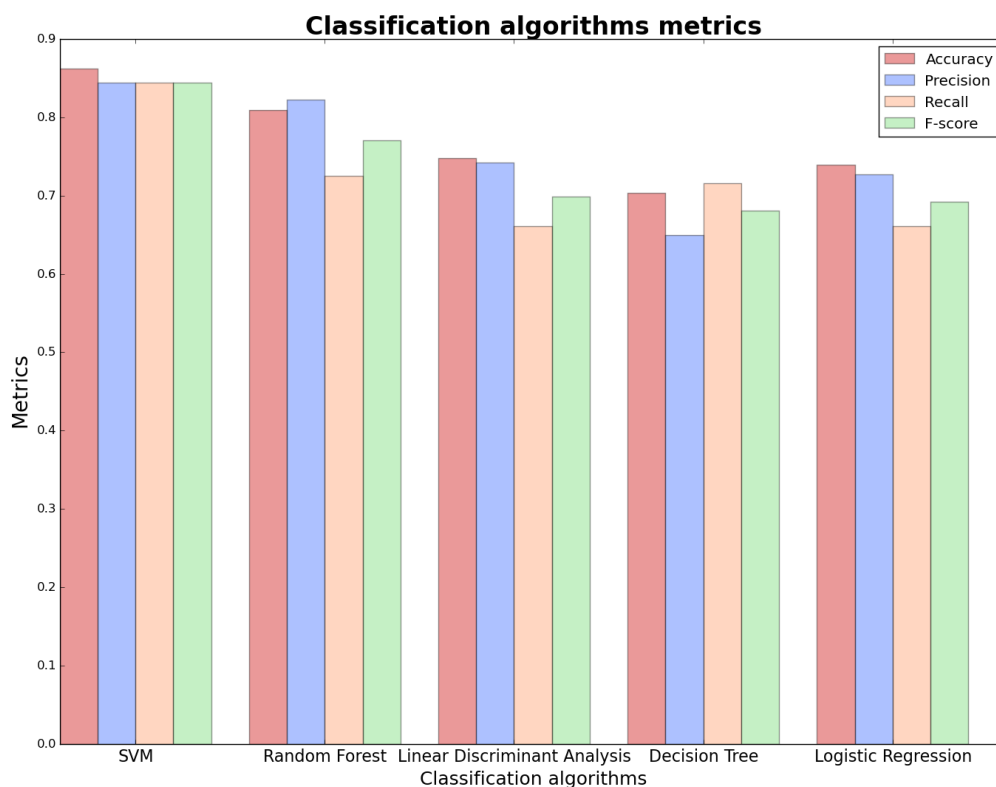


Σχήμα 4.8: Scree test για τον υπολογισμό του ιδανικού αριθμού συνιστωσών

Αφού βρέθηκε ότι ο αριθμός των συνιστωσών είναι 30, κάνουμε δυαδική κατηγοριοποίηση με τις τιμές 0 (υποσιτισμένος) και 1 (μη υποσιτισμένος) χρησιμοποιώντας τους εξής πέντε κατηγοριοποιητές: SVM, Random Forest, Linear Discriminant Analysis, Decision Tree και Logistic Regression. Παρατηρούμε ότι ο SVM, όπως και στην προηγούμενη PCA ανάλυση, έχει τα καλύτερα αποτελέσματα συγκριτικά με τους υπόλοιπους στο σύνολο δοκιμών.

Πίνακας 4.6: Αποτελέσματα μετρικών για κάθε κατηγοριοποιητή, έχοντας ως κλάση το Determine

Algorithm / Metrics	Accuracy	Precision	Recall	F-score
SVM	88.21 ± 2.52	88.46 ± 2.5	84.40 ± 2.84	86.38 ± 2.68
Random Forest	77.64 ± 3.26	80.00 ± 3.0	66.06 ± 3.7	72.36 ± 3.49
Linear Discriminant Analysis	80.08 ± 3.12	80.61 ± 3.09	72.48 ± 3.49	76.33 ± 3.32
Decision tree	74.80 ± 3.39	70.09 ± 3.58	75.23 ± 3.37	72.57 ± 3.49
Logistic Regression	73.58 ± 3.45	76.19 ± 3.33	58.72 ± 3.85	66.32 ± 3.69



Σχήμα 4.9: Διάγραμμα απεικόνισης των μετρικών για κάθε κατηγοριοποιητή, όταν η κλάση είναι το Determine

4.5 Συσχέτιση χαρακτηριστικών με την άνοια

Σε αυτήν την υποενότητα, στόχος είναι να βρεθεί αν υπάρχει συσχέτιση μεταξύ κάποιων χαρακτηριστικών ή διατροφικών δεικτών και ασθενειών που εμφανίζονται κυρίως στους ηλικιωμένους, όπως η άνοια, η νόσος Alzheimer και η νόσος Parkinson. Συγκεκριμένα η άνοια δεν είναι μια απλή ασθένεια, αλλά ένα γενικό σύνδρομο, το οποίο επηρεάζει τη μνήμη, τη γλώσσα και την ικανότητα επίλυσης προβλημάτων. Αποτελεί σοβαρή απώλεια της γενικής νοητικής ικανότητας, όπως απώλεια μνήμης, πέρα από την απώλεια που δημιουργεί η φυσιολογική διαδικασία της γήρανσης. Η άνοια εμφανίζεται πιο συχνά στους ηλικιωμένους, μπορεί όμως σπανιότερα να εμφανιστεί και σε άτομα κάτω των 65 ετών.

Επομένως είναι σημαντικό να γίνει μια εκτίμηση για το ποιες διατροφικές συνήθειες συσχετίζονται άμεσα με την εμφάνιση της άνοιας στους ηλικιωμένους και αν ο δείκτης

Determine περιέχει αρκετή πληροφορία σε σχέση με την άνοια. Παρακάτω φαίνονται τα αποτελέσματα με την μέθοδο InfoGainAttributeEval του Weka, χρησιμοποιώντας 10-fold cross validation.

Αρχικά παρατηρούμε ότι τα χαρακτηριστικά με τη μεγαλύτερη πληροφορία ανήκουν στην κατηγορία των προβλημάτων που οφείλονται σε σωματικά προβλήματα υγείας. Μερικά παραδείγματα είναι τα προβλήματα μνήμης που έχουν οι ηλικιωμένοι όσον αφορά τις λίστες για τα ψώνια, προβλήματα χειρισμού των οικονομικών ή η μειωμένη ικανότητα να χρησιμοποιήσουν τα μέσα μεταφοράς. Περαιτέρω χαρακτηριστικά αφορούν υποκειμενικά προβλήματα μνήμης, όπως για παράδειγμα αν έχουν δυσκολία να θυμηθούν ονόματα ή πράγματα λέξεις που μόλις ειπώθηκαν. Συγκεκριμένα, το χαρακτηριστικό F26 (πρόβλημα μνήμης) εμφανίζεται πρώτο στην κατάταξη του Πίνακα 4.7, με average merit (μέση αξία) 0.098 ± 0.004 , δηλαδή έχει την περισσότερη πληροφορία κατά μέσο όρο σε κάθε ένα από τα δέκα folds.

Στο τέλος της κατάταξης εμφανίζονται οι δέκα ερωτήσεις που σχηματίζουν τον δείκτη Determine καθώς και ο τελικός δείκτης. Η αξία τους όσον αφορά την άνοια είναι πολύ μικρή, κυμαίνεται από 0.001 έως 0.1, οπότε συμπεραίνουμε ότι δεν περιέχουν αρκετή πληροφορία.

Πίνακας 4.7: Αποτελέσματα μεθόδου InfoGainAttributeEval μεταξύ των χαρακτηριστικών και της άνοιας

Average merit	Average rank	Attribute	Average merit	Average rank	Attribute
0.098 +- 0.004	1 +- 0	F26	0.025 +- 0.002	22.4 +- 1.43	F37
0.081 +- 0.003	2.2 +- 0.4	F15	0.024 +- 0.001	22.5 +- 1.63	E17a
0.075 +- 0.004	2.9 +- 0.54	F7	0.025 +- 0.002	22.8 +- 1.47	F4
0.07 +- 0.002	4.4 +- 0.66	F2	0.023 +- 0.002	24 +- 2.32	F31
0.069 +- 0.003	4.9 +- 0.83	F40b	0.021 +- 0.003	28.8 +-10.42	L17_2b
0.065 +- 0.003	6.5 +- 1.43	F12	0.02 +- 0.002	29.5 +- 2.91	L29
0.065 +- 0.003	6.8 +- 0.98	F40a	0.019 +- 0.001	31.1 +- 4.55	F24
0.063 +- 0.003	8.2 +- 0.98	F40d	0.018 +- 0.001	32.2 +- 2.93	F32
0.062 +- 0.003	8.6 +- 1.02	F14	0.017 +- 0.002	36.7 +- 6.91	L28
0.059 +- 0.003	9.5 +- 0.92	F13	0.016 +- 0.002	42.4 +-12.63	DETERMINE10
0.052 +- 0.003	11.3 +- 0.64	F3	0.013 +- 0.001	61.4 +-10.5	DETERMINE_ALL
0.05 +- 0.002	11.9 +- 0.54	F36	0.01 +- 0.001	100.1 +-19.36	DETERMINE8
0.044 +- 0.003	13.3 +- 0.78	F30	0.004 +- 0.001	245.8 +-43.35	DETERMINE5
0.045 +- 0.002	13.5 +- 0.5	F40c	0.003 +- 0.001	294 +-37.41	DETERMINE6
0.033 +- 0.002	15.9 +- 0.94	F5	0.001 +- 0	391.5 +-37.66	DETERMINE9
0.031 +- 0.004	17.3 +- 2.49	A10_age	0.001 +- 0	399.2 +-26.92	DETERMINE2
0.03 +- 0.001	17.8 +- 0.87	F6	0.001 +- 0	424.8 +-28.44	DETERMINE4
0.03 +- 0.004	18.2 +- 2.09	F10	0.001 +- 0	432 +-33.25	DETERMINE1
0.03 +- 0.003	18.4 +- 2.29	C74	0 +- 0	490.6 +-27.27	DETERMINE3
0.029 +- 0.002	18.7 +- 2	F38	0 +- 0	526.7 +-24.48	DETERMINE7

Συμπεράσματα

Αρχικός στόχος της πτυχιακής ήταν, δεδομένου ενός συνόλου με δεδομένα διατροφής, να βρεθεί εάν υπάρχει συσχέτιση μεταξύ των διατροφικών συνηθειών των ανθρώπων άνω των 65 ετών με τους διατροφικούς δείκτες BMI και Determine. Επιπλέον, εφαρμόστηκαν αλγόριθμοι κατηγοριοποίησης στο σύνολο δεδομένων, έχοντας ως κλάση αρχικά το BMI και στη συνέχεια τον δείκτη για τον κίνδυνο υποσιτισμού Determine. Η σύγκριση των αλγορίθμων που χρησιμοποιήθηκαν μας βοηθάει να καταλάβουμε ποιος κατηγοριοποιητής προβλέπει καλύτερα σε ποια κατηγορία ανήκει ένα άγνωστο δείγμα και γιατί.

Αρχικά, χρησιμοποιήθηκε η συσχέτιση Pearson για να βρεθούν ποια χαρακτηριστικά του συνόλου δεδομένων συσχετίζονται γραμμικά με το BMI. Όπως ήδη αναφέρθηκε στην υποενότητα 4.1.1 η τιμή της συσχέτισης Pearson ορίζεται στο διάστημα $[-1, 1]$.

Θετική συσχέτιση έχουμε όταν η τιμή της συσχέτισης Pearson τείνει στο 1. Σε αυτήν την περίπτωση η αύξηση της τιμής του χαρακτηριστικού επιφέρει αύξηση στην τιμή του BMI και αντίστροφα. Τα αποτελέσματα του Πίνακα 4.1 αναφέρουν ότι το βάρος έχει θετική γραμμική συσχέτιση με το BMI με $\text{corr}(\text{βάρος}, \text{BMI}) = 0.747$. Από τον ορισμό του τύπου του $\text{BMI} = \frac{\text{βάρος}}{\text{ύψος}^2}$, φαίνεται ξεκάθαρα ότι όσο αυξάνεται ο αριθμητής του κλάσματος (βάρος), τόσο αυξάνεται και η τιμή του BMI, οπότε τα δύο αυτά χαρακτηριστικά συσχετίζονται θετικώς γραμμικά. Θετική συσχέτιση έχει επίσης η περιφέρεια της μέσης με $\text{corr}(\text{περιφέρεια μέσης}, \text{BMI}) = 0.687$, που σημαίνει επίσης ότι όσο αυξάνεται η περιφέρεια αυξάνεται και ο δείκτης.

Αρνητική συσχέτιση με το BMI έχει το ύψος, διότι η αύξηση του παρανομαστή του κλάσματος προκαλεί μείωση του BMI. Τα αποτελέσματα για αυτά τα χαρακτηριστικά προκύπτουν εξ ορισμού μέσω του τύπου.

Όμως μπορούμε να καταλήξουμε σε επιπρόσθετα συμπεράσματα που δεν προκύπτουν άμεσα από τον τύπο υπολογισμού του BMI. Τα χαρακτηριστικά που αναφέρονται παρα-

κάτω συσχετίζονται μέτρια με το BMI, δηλαδή η σχέση μεταξύ τους δεν μπορεί να περιγραφεί εντελώς γραμμικά. Η τιμή της συσχέτισης τους με το BMI κυμαίνεται μεταξύ 0.102 και 0.249. Για ευκολία θα αναφέρουμε τα περισσότερα από αυτά σε κατηγορίες:

- Οι δραστηριότητες που κάνει ο ηλικιωμένος στον ελεύθερό του χρόνο. Μερικά από τα χαρακτηριστικά είναι το περπάτημα, η κολύμβηση, το γυμναστήριο, η ανάγνωση βιβλίων, η έξοδος για ψώνια, η επίσκεψη σε μουσεία και η συμμετοχή σε διάφορους συλλόγους.
- Προβλήματα που οφείλονται σε σωματικά προβλήματα υγείας. Σε αυτά συμπεριλαμβάνονται αδυναμίες του ηλικιωμένου να κάνει οικιακές δουλειές, προβλήματα στην ούρηση και αφόδευση, η συνήθεια να μιλάει για το παρελθόν, η χρήση μαστουνιού ως βοήθεια για τη βάδιση.
- Συχνότητα κατανάλωσης συγκεκριμένων τροφίμων, όπως άσπρο ψωμί ή φρυγανιά, ελαιόλαδο, τυρί φέτα ή ανθότυρο, αρνί, κατσίκι ή παϊδάκια, πίτες σπιτικές.
- Ιατρικό ιστορικό, για παράδειγμα αν έχει διαγνωστεί με υπέρταση, αρθритικά ή αν έχει κάνει εγχειρήσεις.

Είναι σημαντικό επίσης να τονίσουμε ότι υπάρχουν χαρακτηριστικά που είναι ασυσχέτιστα γραμμικά με το BMI. Δηλαδή η τιμή της συσχέτισης Pearson είναι περίπου ίση με 0 και η σχέση των δύο χαρακτηριστικών δεν μπορεί να περιγραφεί με μια συνάρτηση ευθείας. Τα χαρακτηριστικά που ανήκουν σε αυτήν την κατηγορία είναι συνοπτικά η παρούσα ψυχιατρική κατάσταση του ηλικιωμένου, η κατάθλιψη, το άγχος, οι νευροψυχολογικές διαγνώσεις και τα νευροψυχιατρικά συμπτώματα (NPI), όπως παραληρήματα, ψευδαισθήσεις, επιθετικότητα, απάθεια, παθολογική κινητική συμπεριφορά.

Όσον αφορά τα νευροψυχιατρικά συμπτώματα, τα αποτελέσματα της μεθόδου InfoGainAttributeEval έδειξαν ότι περιέχουν την περισσότερη πληροφορία σε σχέση με το BMI. Αυτή η ιδιότητα, σε συνδυασμό με το ότι είναι γραμμικά ασυσχέτιστα με το BMI, καθιστά αυτά τα χαρακτηριστικά κατάλληλα για τον διαχωρισμό του BMI.

Στο κομμάτι της δυαδικής κατηγοριοποίησης, είδαμε ότι η άνιση κατανομή των δεδομένων μπορεί να επιφέρει προβλήματα και έτσι τα αποτελέσματα να μην είναι ακριβή. Η χρήση μεθόδων αναδειγματοληψίας αποτελεί λύση για τέτοια σύνολα δεδομένων, έτσι ώστε η κατανομή των εγγραφών και στις δύο κλάσεις να είναι περίπου ίδια.

Τα αποτελέσματα από τη σύγκριση των αλγορίθμων έδειξαν ότι ο SVM μπορεί να προβλέψει καλύτερα την κλάση ενός δείγματος του συνόλου ελέγχου. Αυτό συνέβη και στα δύο σύνολα δεδομένων, όπου στο ένα η κλάση ήταν το BMI και στο άλλο ο δείκτης Determine. Ο αλγόριθμος χρησιμοποίησε ως παράμετρο τη συνάρτηση RBF kernel, με αποτέλεσμα να δημιουργηθεί ένα μη γραμμικό μοντέλο. Το μοντέλο αυτό διαχώρισε πολύ καλά τις εγγραφές που ανήκουν σε διαφορετικές κλάσεις.

Όσον αφορά τις μετρικές αξιολόγησης στην δυαδική κατηγοριοποίηση, είναι χρήσιμο να μην λαμβάνουμε υπόψιν μόνο την ευστοχία του κατηγοριοποιητή. Υπάρχουν και άλλες μετρικές τις οποίες θα πρέπει να συγκρίνουμε μεταξύ τους, ώστε να καταλήξουμε σε ένα μοντέλο που να είναι γενικό και να μην υπερπροσαρμόζεται στα δεδομένα εκπαίδευσης. Οι μετρικές αυτές είναι η ακρίβεια, η ανάκληση, το f-score καθώς και ο πίνακας σύγχυσης.

Στη συνέχεια χρησιμοποιήθηκαν οι κλάσεις CfsSubsetEval και InfoGainAttributeEval, ώστε να βρεθεί αν υπάρχει συσχέτιση μεταξύ των χαρακτηριστικών και του διατροφικού δείκτη Determine. Τα αποτελέσματα μας βοήθησαν να συμπεράνουμε ότι ο κίνδυνος υποσιτισμού συσχετίζεται αρκετά με το αν ο ηλικιωμένος χρησιμοποιεί μασέλα κατά τη διάρκεια του φαγητού και αν τρώει μόνος ή με παρέα. Πέρα από αυτά τα δύο χαρακτηριστικά, τα νευροψυχιατρικά συμπτώματα περιέχουν επίσης μεγάλο κέρδος πληροφορίας σε σχέση με τον δείκτη υποσιτισμού, όπως είναι οι διαταραχές της διατροφής και της συμπεριφοράς τη νύχτα, η κατάθλιψη και οι ψευδαισθήσεις.

Τέλος, στην περίπτωση που θέλουμε να προβλέψουμε αν ένα άγνωστο δείγμα έχει άνοια ή όχι, συμπεραίνουμε ότι ο δείκτης Determine δεν θεωρείται σημαντικό γνώρισμα. Αυτό προκύπτει από το γεγονός ότι ο δείκτης, σύμφωνα με τον Πίνακα 4.7 δεν περιέχει πολλή πληροφορία ώστε να μπορέσει να κάνει διάκριση μεταξύ των δύο κλάσεων. Αντιθέτως, τα σημαντικότερα χαρακτηριστικά που βοηθούν στον διαχωρισμό μεταξύ των δύο τιμών της άνοιας είναι τα εξής:

- Προβλήματα του ηλικιωμένου με την μνήμη
- Ικανότητα χειρισμού χρημάτων, πληρωμής λογαριασμών και συναλλαγών
- Προβλήματα μνήμης για γεγονότα που συνέβησαν πρόσφατα
- Ικανότητα χρήσης της τηλεφωνικής συσκευής
- Ικανότητα χρήσης των μέσων μαζικής μεταφοράς

- Υπευθυνότητα στη λήψη φαρμάκων, όσον αφορά τη δοσολογία και τη συχνότητα
- Δυσκολία να θυμηθεί λέξεις που μόλις ειπώθηκαν
- Προβλήματα στον προσανατολισμό και στην αναγνώριση του χώρου όπου βρίσκεται
- Ηλικία

Βιβλιογραφία

- Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, 6(1):20–29, 2004.
- Anne Marie Beck, L Ovesen, and M Osler. The ‘mini nutritional assessment’(mna) and the ‘determine your nutritional health’ checklist (nsi checklist) as predictors of morbidity and mortality in an elderly danish population. *British Journal of Nutrition*, 81(01):31–36, 1999.
- Joseph L Breault, Colin R Goodall, and Peter J Fos. Data mining a diabetic data warehouse. *Artificial intelligence in medicine*, 26(1):37–54, 2002.
- Soumen Chakrabarti, Martin Ester, Usama Fayyad, Johannes Gehrke, Jiawei Han, Shinichi Morishita, Gregory Piatetsky-Shapiro, and Wei Wang. Data mining curriculum: A proposal (version 1.0). Intensive Working Group of ACM SIGKDD Curriculum Committee, page 140, 2006.
- Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- Efthimios Dardiotis, Mary H Kosmidis, Mary Yannakoulia, Georgios M Hadjigeorgiou, and Nikolaos Scarmeas. The hellenic longitudinal investigation of aging and diet (heliad): Rationale, study design, and cohort description. *Neuroepidemiology*, 43(1): 9–14, 2014.
- Áine P Hearty and Michael J Gibney. Analysis of meal patterns with the use of supervised data mining techniques—artificial neural networks and decision trees. *The American journal of clinical nutrition*, 88(6):1632–1642, 2008.

- Aine P Hearty, Michael J Gibney, et al. Comparison of cluster and principal component analysis techniques to derive dietary patterns in irish adults. *British journal of nutrition*, 101(4):590, 2009.
- Machine Learning. Tom mitchell. ISBN: 0-07-042807-7, Publisher: McGraw Hill, 2009.
- Alexander Liu, Joydeep Ghosh, and Cheryl E Martin. Generative oversampling for mining imbalanced datasets. In *DMIN*, pages 66–72, 2007.
- David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.
- Jonathan C Prather, David F Lobach, Linda K Goodwin, Joseph W Hales, Marvin L Hage, and W Edward Hammond. Medical data mining: knowledge discovery in a clinical data warehouse. In *Proceedings of the AMIA annual fall symposium*, page 101. American Medical Informatics Association, 1997.
- Stuart Jonathan Russell, Peter Norvig, John F Canny, Jitendra M Malik, and Douglas D Edwards. Artificial intelligence: a modern approach, volume 2. Prentice hall Upper Saddle River, 2003.
- Jakob Skoet and Kostas G Stamoulis. The State of Food Insecurity in the World 2006: Eradicating World Hunger-Taking Stock Ten Years After the World Food Summit. Food & Agriculture Org., 2006.
- Bassiliades Kokkoras Sakellariou Vlahavas, Kefalas. Artificial Intelligence. University of Macedonia Press / Greece, 2005.
- Patricia MCM Waijers, Edith JM Feskens, and Marga C Ocké. A critical review of predefined diet quality scores. *British Journal of Nutrition*, 97(02):219–231, 2007.
- Ian H Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.
- Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.
- Μανωλόπουλος Νανόπουλος. Εισαγωγή στην Εξόρυξη Δεδομένων και τις αποθήκες Δεδομένων. Εκδόσεις Νέων Τεχνολογιών, 2008.