ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Π.Μ.Σ. ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ

ΚΑΤΕΥΘΥΝΣΗ: ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«ΣΥΓΚΡΙΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΕ ΔΕΔΟΜΕΝΑ ΑΣΘΕΝΩΝ ΜΕ ΔΙΑΒΗΤΗ»

ΧΑΡΑΛΑΜΠΟΣ ΤΗΛΛΥΡΟΣ

ΑΜ: ΜΕ1734

ΕΠΙΒΛΕΠΟΥΣΑ ΚΑΘΗΓΗΤΡΙΑ

ΚΑΘΗΓΗΤΡΙΑ ΑΝΔΡΙΑΝΑ ΠΡΕΝΤΖΑ

ΠΕΙΡΑΙΑΣ 2019

UNIVERSITY OF PIRAEUS

DEPARTMENT OF DIGITAL SYSTEMS

MSc on INFORMATION SYSTEMS AND SERVICES

Area of Study: BIG DATA AND ANALYTICS

MASTER THESIS

"A COMPARATIVE EVALUATION OF MACHINE LEARNING ALGORITHMS IN PATIENT DATA WITH DIABETES"

CHARALAMBOS TYLLIROS

ME1734

SUPERVISOR

PROFESSOR ANDRIANA PRENTZA

PIRAEUS 2019

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα πτυχιακή εργασία εκπονήθηκε από το μεταπτυχιακό φοιτητή Τήλλυρο Χαράλαμπο του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς, κατά το ακαδημαϊκό έτος 2018 – 2019, υπό την επίβλεψη της καθηγήτριας Πρέντζα Ανδριάνα.

Θα ήθελα να εκφράσω τις ευχαριστίες και την ευγνωμοσύνη μου στην καθηγήτρια μου για την ανάθεση του θέματος, για τις πολύτιμες γνώσεις που μοιράστηκε μαζί μου, την ανεκτίμητη βοήθειά της, ως επίσης το ενδιαφέρον της, καθώς και το χρόνο που διέθεσε, ώστε να καταστεί δυνατή η από μέρους μου διεκπεραίωση της πτυχιακής εργασίας.

Ιδιαίτερες ευχαριστίες εκφράζω και στην οικογένεια μου για όλα όσα μου έχουν προσφέρει κατά την διάρκεια των φοιτητικών μου χρόνων αλλά και για τη συνεχόμενη υποστήριξη τους.

ΠΕΡΙΛΗΨΗ

Η ραγδαία ανάπτυξη της τεχνολογίας τα τελευταία χρόνια σε συνδυασμό με την καθημερινή αποθήκευση μεγάλων όγκων δεδομένων, καθώς και η ανάγκη για σωστή κατηγοριοποίηση των δεδομένων αυτών, οδήγησε στην υλοποίηση διαφόρων αλγορίθμων με σκοπό την καλύτερη κατηγοριοποίηση των δεδομένων αυτών. Η κατηγοριοποίηση εφαρμόζεται σε πολλά επιστημονικά πεδία όπως η ιατρική, η οικονομία, η μετεωρολογία και άλλα πολλά. Στο πεδίο της ιατρικής η σωστή και άμεση πρόβλεψη μιας μεταβολικής ασθένειας όπως ο διαβήτης, παίζει πολύ σημαντικό ρόλο. Οι ασθένειες αυτές μπορούν να προκαλέσουν διάφορες άλλες σοβαρότερες επιπλοκές, έτσι γίνεται αντιληπτό ότι η ανάπτυξη συστημάτων, τα οποία θα μπορούν να προβλέψουν με αρκετά υψηλή ακρίβεια τέτοιου είδους ασθένειες, είναι πολύ σημαντική.

Η παρούσα διπλωματική εργασία με τίτλο «Συγκριτική Αξιολόγηση Αλγορίθμων Μηχανικής Μάθησης σε Δεδομένα Ασθενών με Διαβήτη» αναφέρεται στη συγκριτική αξιολόγηση της επίδοσης αλγορίθμων μηχανικής μάθησης και συγκεκριμένα της επιβλεπόμενης μάθησης για την πρόβλεψη του διαβήτη. Τέτοιοι αλγόριθμοι είναι ο απλοϊκός Bayes, η λογιστική παλινδρόμηση, τα νευρωνικά δίκτυα, οι μηχανές διανυσμάτων στήριξης, τα δένδρα απόφασης, η συλλογική μάθηση και ο "Κ κοντινότεροι γείτονες". Γίνεται συνοπτική παρουσίαση των διαφόρων αλγορίθμων που επιλέχτηκαν, των αποτελεσμάτων από άλλες διεθνείς μελέτες που αφορούν το ίδιο θέμα, και στη συνέχεια παρουσιάζονται τα αποτελέσματα εφαρμογής των αλγορίθμων σε δύο σύνολα δεδομένων, διαθέσιμα δωρεάν για μελέτη από το διαδίκτυο. Η επεξεργασία πραγματοποιήθηκε με τη χρήση της γλώσσας προγραμματισμού Python.

Λέξεις κλειδιά: διαβήτης, μηχανική μάθηση, απλοϊκός Bayes, λογιστική παλινδρόμηση, νευρωνικά δίκτυα, μηχανές διανυσμάτων στήριξης, δένδρα απόφασης, συλλογική μάθηση, Κ κοντινότεροι γείτονες.

ABSTRACT

The rapid development of technology in recent years, coupled with the daily storage of large volumes of data, and the need for a proper classification of these data, has led to the implementation of various algorithms for their better classification. Classification takes place in many scientific fields such as medicine, economics, meteorology and much more. In the field of medicine, the correct and immediate prediction of a metabolic disease such as diabetes plays a very important role. These diseases can cause several other more serious complications, so it is conceivable that developing systems where they can predict these diseases with big accuracy is very important.

This dissertation titled "A Comparative Evaluation of Machine Learning Algorithms in Patient Data with Diabetes" refers with the comparative evaluation of the performance of mechanical learning algorithms and in particular supervised learning for predicting diabetes. Such algorithms are Bayesian simplistic, logistic regression, neural networks, support vector machines, decision trees, collective learning and the “K closest neighbors”. A summary of the different algorithms selected, the results of other international studies on the same topic, are presented, and then the results of applying the algorithms to two sets of data, available for free study online. The editing was done using the Python programming language.

Keywords: diabetes, machine learning, naïve Bayes, logistic regression, neural networks, support vector machines, decision trees, ensemble learning, k nearest neighbors.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Πίνακας Πινάκων

Πίνακας Εικόνων

Πίνακας Εξισώσεων

Πίνακας Γραφικών Παραστάσεων

Πίνακας Συντομεύσεων

Diabetic Mellitus	DM
False Negative	FN
False Negative Rate	FNR
False Positive	FP
False Positive Rate	FPR
Insulin Dependent Diabetes Mellitus	IDDM
k Nearest Neighbors	kNN
Least Absolute Shrinkage and Selection Operator	Lasso
Linear Discriminant Analysis	LDA
Non Insulin Dependent Diabetes Mellitus	NIDDM
Partitioning Around Medoid	PAM
Pima Indian Diabetes Dataset	PIDD
Positive Predictive Value	PPV
Principal Components Analysis	PCA
Receiver Operating Characteristic	ROC
Support Vector Machines	SVM
True Negative	TN
True Negative Rate	TNR
True Positive	TP
True Positive Rate	TPR
World Health Organization	WHO

Κεφάλαιο 1: Εισαγωγή

1.1 Εισαγωγή

Η εξόρυξη δεδομένων χρησιμοποιείται σε πολλά συστήματα, αφού μπορεί να μετατρέψει τα ακατέργαστα δεδομένα σε χρήσιμη πληροφορία. Ένα από αυτά τα συστήματα είναι το σύστημα υγείας, αφού τα ιατρικά δεδομένα τα οποία παράγονται καθημερινά και η εξαγωγή χρήσιμης πληροφορίας από αυτά με τη χρήση της εξόρυξης δεδομένων, μπορεί να βοηθήσει στη βελτίωση της διάγνωσης και θεραπείας των ασθενών. [1]

1.2 Πρόβλημα

Μια από τις κυριότερες αιτίες πρόωρων ασθενειών και θανάτων παγκοσμίως είναι η εμφάνιση του διαβήτη. Σε αρκετές χώρες οι επιπλοκές και η νοσηρότητα αυξάνονται εκθετικά αφού δεν γίνεται έγκαιρη διάγνωση και κατάλληλη θεραπεία σε αρκετούς από τους ασθενείς που διαγιγνώσκονται με διαβήτη[2]. Ο διαβήτης είναι μια ασθένεια η οποία συνδέεται κυρίως με την αύξηση του επιπέδου της γλυκόζης στο αίμα. Το φαινόμενο αυτό ονομάζεται υπεργλυκαιμία, η οποία χωρίζεται σε δύο τύπους. Ο πρώτος τύπος ονομάζεται διαβήτης τύπου 1 και οφείλεται στην ανεπαρκή ινσουλίνη στο πάγκρεας, αφού τα βήτα κύτταρα του παγκρέατος αδυνατούν να παράγουν αρκετή ινσουλίνη. Ο δεύτερος τύπος διαβήτη, ο οποίος είναι ο πιο κοινός, ονομάζεται διαβήτης τύπου 2, οφείλεται στο ότι ο οργανισμός δεν μπορεί να χρησιμοποιήσει αποτελεσματικά την παραγόμενη ινσουλίνη.

1.3 Δομή διπλωματικής εργασίας

Στα πλαίσια της διπλωματικής εργασίας μελετήθηκαν διάφορες διεθνείς δημοσιεύσεις που αφορούν την πρόβλεψη διαβήτη τύπου 2 με χρήση αλγορίθμων μηχανικής μάθησης. Συγκεκριμένα, έγινε η μελέτη των αποτελεσμάτων κατηγοριοποίησης ενός ασθενούς αν είναι διαβητικός ή όχι, με τη χρήση διαφόρων αλγορίθμων μηχανικής μάθησης, όπως ο απλοϊκός Bayes (Naïve Bayes), μηχανές διανυσμάτων υποστήριξης (Support Vector Machines), η μέθοδος της παλινδρόμησης (Regression) , τα νευρωνικά δίκτυα (Neural Networks), τα δέντρα απόφασης (Decision Trees) και αλγόριθμοι συλλογικής μάθησης, όπως ο AdaBoost, τα τυχαία δάση (Random Forests) και GradientBoost.

Στο κεφάλαιο 2 γίνεται βιβλιογραφική ανασκόπηση σχετικά με το σακχαρώδη διαβήτη και τη μηχανική μάθηση. Αρχικά αναλύεται η έννοια του σακχαρώδους διαβήτη και τα προβλήματα που μπορεί να υποστεί ένα άτομο το οποίο πάσχει από διαβήτη. Στη συνέχεια

γίνεται αναφορά στην ινσουλίνη, την ορμόνη από όπου επηρεάζεται και δημιουργείται ο διαβήτης. Επιπρόσθετα, δίνονται τα ποσοστά γλυκόζης, τα οποία πρέπει να βρίσκονται στο αίμα πριν και μετά το φαγητό. Παρουσιάζεται επίσης μια στατιστική αναφορά, η οποία βασίζεται σε στοιχεία του Παγκόσμιου Οργανισμού Υγείας (World Health Organization – WHO). Στη συνέχεια δίνεται ο ορισμός της μηχανικής μάθησης και τα πλεονεκτήματα τα οποία παρέχει. Αναλύονται τα διάφορα είδη μηχανικής μάθησης που υπάρχουν, όπως επιβλεπόμενη, μη επιβλεπόμενη και ενισχυτική μάθηση, καθώς και οι τρόποι αξιολόγησης ενός μοντέλου (ορθότητα μοντέλου, μέτρα αποτελεσματικότητας κατηγοριοποίησης, πίνακας σύγχυσης, μέθοδος K – Folds Cross Validation και καμπύλη χαρακτηριστικών λειτουργίας δέκτη – Receiver Operating Characteristic curve (ROC curve)). Αναλύονται επίσης η κατηγοριοποίηση, η παλινδρόμηση, ο μετασχηματισμός και η συσταδοποίηση αντίστοιχα. Τέλος γίνεται μια συνοπτική αναφορά των αλγορίθμων κατηγοριοποίησης και προεπεξεργασίας των δεδομένων που χρησιμοποιήθηκαν για την υλοποίηση της μεταπτυχιακής διπλωματικής εργασίας.

Στο κεφάλαιο 3 παρουσιάζεται η μεθοδολογία, η οποία αναπτύχθηκε στην παρούσα μεταπτυχιακή διπλωματική εργασία. Παρουσιάζονται τα εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση της εργασίας καθώς και οι βιβλιοθήκες οι οποίες χρησιμοποιήθηκαν. Περιγράφονται τα σύνολα δεδομένων που χρησιμοποιήθηκαν και τα διάφορα χαρακτηριστικά τους. Παρουσιάζεται μια σύντομη αιτιολόγηση των σημαντικότερων αλγορίθμων μηχανικής μάθησης που επιλέχτηκαν καθώς και οι παράμετροι που χρησιμοποιήθηκαν στον κάθε αλγόριθμο, κατά τη φάση της προεπεξεργασίας αλλά και της επεξεργασίας των δεδομένων. Επίσης, παρουσιάζονται τα βήματα τα οποία ακολουθήθηκαν για την υλοποίηση της διπλωματικής εργασίας.

Στο κεφάλαιο 4 αναφέρονται τα αποτελέσματα από διάφορες διεθνείς μελέτες. Έπειτα παρουσιάζονται τα αποτελέσματα από την εφαρμογή των αλγορίθμων που αναλύθηκαν, επιλέχτηκαν και εφαρμόστηκαν βάση των βημάτων υλοποίησης που παρουσιάστηκαν στο κεφάλαιο 3. Πιο συγκεκριμένα γίνεται η παρουσίαση των πιο πάνω αποτελεσμάτων χρησιμοποιώντας τις μεθόδου Split και K – Folds Cross Validation. Σε κάθε υποκατηγορία των μεθόδων αυτών παρουσιάζονται γραφικές παραστάσεις όπου δείχνουν τα αποτελέσματα της ορθότητας, της ακρίβειας, της ανάκλησης και του αρμονικού μέσου και εξάγονται κάποια συμπεράσματα. Τέλος, γίνεται η συγκριτική αξιολόγηση των αλγορίθμων στα δύο σύνολα

δεδομένων και επιλέγεται ο καλύτερος αλγόριθμος για την κατηγοριοποίηση του ασθενή με διαβήτη τύπου 2.

Στο κεφάλαιο 5 παρουσιάζονται τα συμπεράσματα. Επίσης, προτείνονται διάφοροι νέοι τρόποι για μελλοντική μελέτη στο πρόβλημα της κατηγοριοποίησης του ασθενή με διαβήτη τύπου 2.

Στο κεφάλαιο 6 παρουσιάζονται οι πηγές που χρησιμοποιήθηκαν για τη διεκπεραίωση της διπλωματικής εργασίας.

1.4 Συνεισφορά

Ο διαβήτης αποτελεί μια από τις σημαντικότερες χρόνιες ασθένειες και πηγές θανάτου παγκοσμίως, προκαλώντας πολλά και σοβαρά προβλήματα στον άνθρωπο. Η συγκεκριμένη μεταπτυχιακή διπλωματική εργασία, προσπαθεί να αναλύσει και να εντοπίσει το καλύτερο μοντέλο αλγορίθμου μηχανικής μάθησης. Το μοντέλο αυτό θα μπορεί αυτόματα, με ένα υψηλό ποσοστό ακρίβειας, να προβλέπει και να προειδοποιεί έγκαιρα τους ασθενείς αν έχουν διαβήτη ή όχι. Η έγκαιρη πρόβλεψη της οποιασδήποτε ασθένειας, συμπεριλαμβανομένου και του διαβήτη, βοηθάει στην αποτελεσματικότερη και έγκαιρη αντιμετώπισή της.

Κεφάλαιο 2: Βιβλιογραφική Επισκόπηση

Στο δεύτερο κεφάλαιο γίνεται μια γενική βιβλιογραφική ανασκόπηση της διπλωματικής εργασίας. Αρχικά παρουσιάζεται ο σακχαρώδης διαβήτης, τα προβλήματα τα οποία μπορεί να έχει κάποιος ο οποίος έχει διαβήτη, τους τύπους διαβήτη που υπάρχουν και που οφείλονται. Παρουσιάζονται επίσης κάποια στατιστικά στοιχεία του διαβήτη από τον Παγκόσμιο Οργανισμό Υγείας , τα οποία κατατάσσουν το διαβήτη σε μία πηγή θανάτου. Στη συνέχεια γίνεται αναφορά στη μηχανική μάθηση, στα είδη μηχανικής μάθησης που υπάρχουν καθώς και στις διάφορες μεθόδους αξιολόγησης του μοντέλου. Τέλος, αναφέρεται το θεωρητικό υπόβαθρο των βασικότερων αλγορίθμων μηχανικής μάθησης.

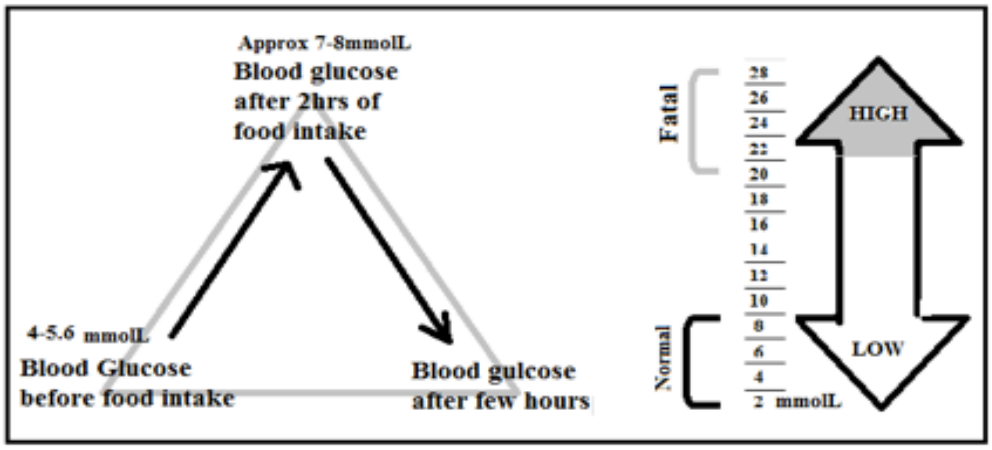
2.1 Σακχαρώδης διαβήτης

Ο σακχαρώδης διαβήτης (Diabetic Mellitus – DM) είναι μια χρόνια ασθένεια η οποία χαρακτηρίζεται από ψηλά επίπεδα γλυκόζης. Αποτελεί ένα πολύ μεγάλο κίνδυνο για την υγεία στις μέρες μας, αφού κατηγοριοποιείται σαν μη μεταδοτική ασθένεια, η οποία επηρεάζει τους ανθρώπους και τους οδηγεί σε διάφορα προβλήματα. Ένα άτομο το οποίο πάσχει από σακχαρώδη διαβήτη, εκτός από προβλήματα στην καρδιά, μπορεί να υποστεί εγκεφαλικό επεισόδιο, καταρράκτη, τύφλωση, νεφρική ανεπάρκεια καθώς και καταστροφή των αιμοφόρων αγγείων[2]-[4]. Σύμφωνα με μελέτες, περισσότερο από 80% του ποσοστού των ανθρώπων που έχουν διαβήτη, πεθαίνουν λόγω καρδιάς ή καταστροφής των αιμοφόρων αγγείων [3] [4].

Ο άνθρωπος επηρεάζεται από το διαβήτη μέσω μιας φυσικής ορμόνης, την ινσουλίνη, η οποία βρίσκεται στο αίμα. Η ινσουλίνη προστατεύεται από το πάγκρεας, το οποίο χρησιμοποιείται από τον οργανισμό στην απορρόφηση της ζάχαρης, του άμυλου και των μορίων των τροφίμων από τα κύτταρα, έτσι ώστε να παραχθεί η ενέργεια που απαιτείται στην καθημερινή ζωή [5]. Υπάρχουν τρείς κύριοι τύποι σακχαρώδους διαβήτη. Ο πρώτος τύπος ονομάζεται διαβήτης τύπου 1, ο οποίος μπορεί να προκληθεί λόγω της αδυναμίας των βήτα κυττάρων του παγκρέατος να παράγουν την κατάλληλη ποσότητα ινσουλίνης [6]-[8]. Ο τύπος αυτός ονομάζεται επίσης και σακχαρώδης διαβήτης εξαρτώμενος από την ινσουλίνη (Insulin Dependent Diabetes Mellitus – IDDM). Ο δεύτερος τύπος ονομάζεται διαβήτης τύπου 2. Όπως και ο διαβήτης τύπου 1, έτσι και ο διαβήτης τύπου 2, οφείλεται και αυτός στα βήτα κύτταρα του σώματος, με τη διαφορά ότι τα κύτταρα αυτά δεν μπορούν να χρησιμοποιήσουν σωστά την ινσουλίνη η οποία παράγεται. Ο τύπος αυτός είναι γνωστός ως μη ινσουλινοεξαρτώμενος

σακχαρώδης διαβήτης (Non Insulin Dependent Diabetes Mellitus – NIDDM). Τελευταίος και σημαντικότερος τύπος διαβήτη είναι ο διαβήτης κύησης, ο οποίος μπορεί να προκληθεί λόγω της ανάπτυξης υψηλού σακχάρου στο αίμα σε εγκύους γυναίκες, χωρίς όμως να υπάρχει προηγούμενη διάγνωση σακχαρώδη διαβήτη. Ο τύπος αυτός μπορεί να κατηγοριοποιηθεί και σαν διαβήτης τύπου 2 [6]. Συνήθως ο διαβήτης τύπου 1 διαγιγνώσκεται από τη μέρα που γεννιέται το άτομο, ενώ ο διαβήτης τύπου 2 εμφανίζεται κυρίως σε ενήλικες άνω των 40 ετών [18].

Το φυσιολογικό εύρος του επιπέδου της γλυκόζης η οποία βρίσκεται στο αίμα, κυμαίνεται μεταξύ 4,0 – 5,6 mmol/L. Μετά την κατανάλωση ενός γεύματος, το επίπεδο γλυκόζης στο αίμα αυξάνεται φτάνοντας μέχρι και 7,8 mmol/L. Οποιαδήποτε τιμή υψηλότερη από αυτές δείχνει την ύπαρξη διαβήτη. Μετά τις δύο ώρες από την ώρα του γεύματος, το επίπεδο γλυκόζης στο αίμα πέφτει και πάλι. Μια ενδιάμεση κατάσταση η οποία παρουσιάζεται, ονομάζεται προ-διαβήτης, όπου το επίπεδο γλυκόζης στο αίμα είναι υψηλότερο από το φυσιολογικό εύρος, αλλά δεν είναι αρκετά υψηλό ώστε να μπορεί να χαρακτηριστεί ως διαβήτης.



Εικόνα 1: Επίπεδα κανονικών και υψηλών επιπέδων γλυκόζης στο αίμα.

[5]

Υπάρχουν τρεις κατηγορίες ανθρώπων ανάλογα με το επίπεδο γλυκόζης που υπάρχει στο αίμα. Οι κατηγορίες αυτές ονομάζονται υγιείς, προ-διαβητικοί και διαβητικοί. Ο επόμενος πίνακας παρουσιάζει το επίπεδο γλυκόζης για κάθε ομάδα ανθρώπων που πρέπει να βρίσκεται στο αίμα πριν και μετά το γεύμα.

ΚΑΤΑΣΤΑΣΗ ΥΓΕΙΑΣ	ΕΠΙΠΕΔΑ ΓΛΥΚΟΖΗΣ ΠΡΙΝ ΑΠΟ ΤΟ ΓΕΥΜΑ	ΕΠΙΠΕΔΑ ΓΛΥΚΟΖΗΣ ΔΥΟ ΩΡΕΣ ΜΕΤΑ ΤΟ ΓΕΥΜΑ
ΥΓΙΗΣ	4.0 – 5.6 mmol/L	<7.8 mmol/L
ΠΡΟ – ΔΙΑΒΗΤΙΚΟΣ	5.6 – 7.0 mmol/L	<7.8 mmol/L
ΔΙΑΒΗΤΙΚΟΣ	>7 mmol/L	≥7.8 mmol/L

Πίνακας 1: Κατάσταση υγείας ασθενών βάση του επιπέδου γλυκόζης δύο ώρες πριν και δύο ώρες μετά το γεύμα.

Ο διαβήτης είναι μια σιωπηρή επιδημία και σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας επηρεάζει 246 εκατομμύρια ανθρώπους στον κόσμο. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας, προβλέπεται ότι οι αναπτυσσόμενες χώρες στον 21ον αιώνα θα επιβαρυνθούν από αυτή την επιδημία, ενώ μέχρι στιγμής περισσότερο από το 70% των ατόμων που έχουν διαβήτη ζουν σε χώρες οι οποίες έχουν χαμηλό και μεσαίο εισόδημα, αντιπροσωπεύοντας έτσι σχεδόν το 6% του ενήλικου πληθυσμού στον κόσμο. Ο αριθμός του ασιατικού πληθυσμού που υποφέρει από διαβήτη είναι πενταπλάσιος από τον αριθμό του λευκού πληθυσμού, αφού αν αναλογιστεί κάποιος ότι ο αριθμός μόνο στην Ινδία ανέρχεται στα 40 εκατομμύρια και αναμένεται ότι μέχρι το 2025 ο αριθμός αυτός θα φτάσει τα 70 εκατομμύρια, γεγονός που καθιστά την Ινδία ως η πρωτεύουσα του διαβήτη. Ο διαβήτης προκαλεί 6 θανάτους κάθε λεπτό και ένας στους 20 θανάτους στον κόσμο οφείλεται σε αυτόν. Κάθε χρόνο εκτιμάται ότι 3,2 εκατομμύρια άνθρωποι στον κόσμο πεθαίνουν λόγω του διαβήτη ή κάποιας ασθένειας η οποία σχετίζεται με το διαβήτη. Από έρευνες που έγιναν έχει διαπιστωθεί ότι ο διαβήτης είναι η τέταρτη κύρια αιτία θανάτου παγκοσμίως [8].

2.2 Μηχανική μάθηση

2.2.1 Τι είναι η μηχανική μάθηση

Ζούμε σε μια εποχή όπου τα δεδομένα τα οποία παράγονται είναι άφθονα και η αξιοποίησή τους με χρήση διαφόρων αλγορίθμων από το πεδίο της μηχανικής μάθησης μπορεί να μετατρέψει αυτά τα δεδομένα σε σημαντική γνώση. Η μηχανική μάθηση ή αλλιώς προγνωστική αναλυτική είναι ένα πεδίο έρευνας το οποίο προέρχεται από την διασταύρωση τριών πεδίων: της στατιστικής, της τεχνητής νοημοσύνης και της επιστήμης των υπολογιστών [9]. Το 1959 ο Arthur Samuel ορίζει τη μηχανική μάθηση ως "*Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί.*" [10][11]Ένας άλλος ορισμός που αφορά τη μηχανική μάθηση, αυτή την φορά από τον Tom M. Mitchell το

1997, αναφέρει ότι "*ένα πρόγραμμα υπολογιστών λέγεται ότι μαθαίνει από την εμπειρία E, σε σχέση με κάποια τάξη εργασιών T και μέτρηση απόδοσης P (Performance Measure) εάν η απόδοσή του σε εργασίες στο T, όπως μετράτε από το P, βελτιώνεται με την εμπειρία E.*" [10][11].

Με άλλα λόγια, η μηχανική μάθηση μπορεί να θεωρηθεί ως μια συλλογή από μεθόδους που μπορούν αυτόματα να αναγνωρίσουν διάφορα μοτίβα στα δεδομένα και στη συνέχεια με βάση αυτά τα μοτίβα να προβλέψουν μελλοντικά αποτελέσματα ή να πάρουν αποφάσεις κάτω από συγκεκριμένες καταστάσεις. Όλα αυτά έχουν τη δυνατότητα να αξιοποιηθούν χρησιμοποιώντας διάφορους αλγορίθμους που επιτρέπουν στις μηχανές να καταλαβαίνουν διάφορες καταστάσεις και βασισμένες σε αυτές να παίρνονται οι αποφάσεις [6].

Η μηχανική μάθηση είναι χρήσιμη σε όλους τους τομείς αφού παρέχει διάφορα πλεονεκτήματα, όπως:

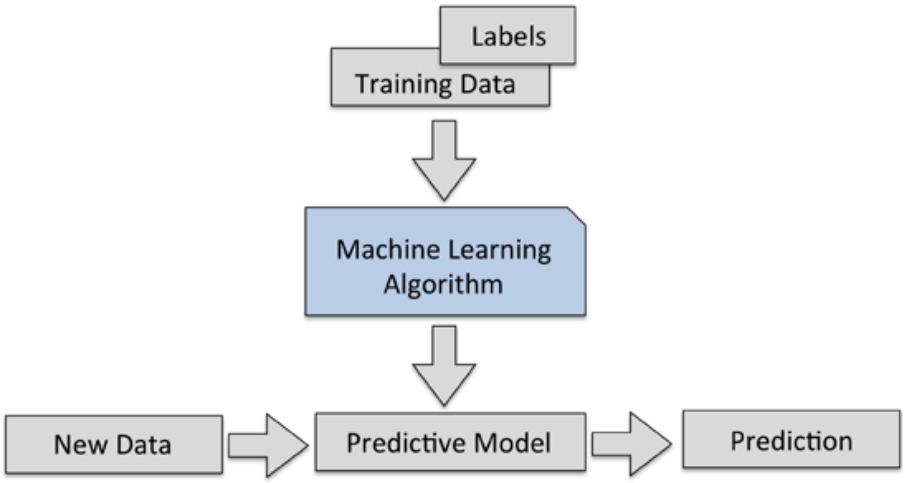
- Γρήγορη απόφαση: Η μηχανική μάθηση παρέχει γρήγορα τα καλύτερα αποτελέσματα.
- Ικανότητα προσαρμογής: Έχει τη δυνατότητα να προσαρμόζεται γρήγορα στο νέο περιβάλλον, το οποίο μεταβάλλεται συνεχώς, αφού τα δεδομένα ενημερώνονται συνεχώς.
- Καινοτομία: Με τη χρησιμοποίηση προηγμένων αλγορίθμων βελτιώνεται η ικανότητα λήψης αποφάσεων, βοηθώντας έτσι στην ανάπτυξη καινοτόμων επιχειρηματικών υπηρεσιών και μοντέλων.
- Διορατικότητα: Γίνεται η κατανόηση μοναδικών προτύπων δεδομένων και με βάση αυτών βασίζονται στις ενέργειες που μπορούν να παρθούν.
- Επιχειρηματική ανάπτυξη: Η συνολική επιχειρηματική διαδικασία και η ροή εργασίας είναι ταχύτερες, βοηθώντας έτσι στην επιχειρηματική ανάπτυξη.
- Καλό αποτέλεσμα: Το αποτέλεσμα θα βελτιώνεται σε αντίθεση με τη πιθανότητα σφάλματος, η οποία θα μειώνεται.

Υπάρχουν τρεις τύποι μηχανικής μάθησης, η επιβλεπόμενη μάθηση (Supervised Learning), η μη επιβλεπόμενη μάθηση (Unsupervised Learning) και η ενισχυτική μάθηση (Reinforcement Learning). Η επιβλεπόμενη μάθηση έχει σαν κύριες μεθόδους την κατηγοριοποίηση και την παλινδρόμηση, ενώ η μη επιβλεπόμενη μάθηση το μετασχηματισμό και τη συσταδοποίηση, έννοιες οι οποίες αναλύονται στη συνέχεια. Και στις δύο περιπτώσεις, τα δεδομένα εισόδου πρέπει να έχουν σωστή αναπαράσταση για να μπορεί να τα καταλάβει

ένας υπολογιστής. Η ενισχυτική μάθηση ασχολείται κυρίως με διάφορες οντότητες που ονομάζονται πράκτορες, οι οποίοι παίρνουν τις αποφάσεις τους από το περιβάλλον, με σκοπό να εκτελέσουν κάποια ενέργεια.

2.2.2 Επιβλεπόμενη μάθηση

Η επιβλεπόμενη μάθηση είναι μια από τις πιο κοινές και επιτυχημένες μεθόδους που χρησιμοποιούνται στη μηχανική μάθηση, η οποία χρησιμοποιείται όταν θέλουμε να προβλέψουμε ένα σίγουρο αποτέλεσμα από μία δεδομένη είσοδο [9]. Ονομάζεται επιβλεπόμενη μάθηση γιατί το μοντέλο μας μαθαίνει από ένα σύνολο εκπαίδευσης δημιουργώντας έτσι ένα άλλο μοντέλο, όπου με βάση αυτό εφαρμόζεται στο νέο σύνολο δεδομένων για να προβλέψει τα αποτελέσματα [6]. Υπάρχουν δύο υποκατηγορίες επιβλεπόμενης μάθησης, η κατηγοριοποίηση και η παλινδρόμηση. Αλγόριθμοι οι οποίοι εφαρμόζουν επιβλεπόμενη μάθηση είναι η γραμμική παλινδρόμηση (Linear Regression), τα νευρωνικά δίκτυα (Neural Networks), οι μηχανές διανυσμάτων στήριξης (Support Vector Machines – SVMs), η μάθηση κατά Bayes (Bayesian Learning), τα δένδρα απόφασης (Decision Trees), ο k πλησιέστεροι γείτονες (k Nearest Neighbors – kNN), η λογιστική παλινδρόμηση (Logistic Regression) και τα τυχαία δάση (Random Forests).



Εικόνα 2: Λειτουργία επιβλεπόμενης μάθησης.

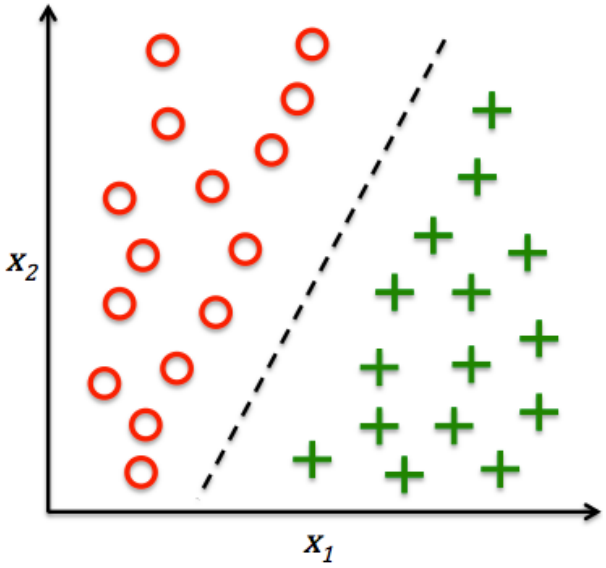
[12]

Η πιο πάνω εικόνα παρουσιάζει τη λειτουργία της επιβλεπόμενης μάθησης. Πιο συγκεκριμένα, υπάρχει το σύνολο εκπαίδευσης (Training Data) μαζί με την ετικέτα κατηγοριοποίησης (Label) και στη συνέχεια η επιλογή του αλγορίθμου μηχανικής μάθησης (Machine Learning Algorithm), ο οποίος θα χρησιμοποιηθεί για την εκπαίδευση. Αφού ολοκληρωθεί η εκπαίδευση, τα νέα δεδομένα (New Data), εφαρμόζονται στο μοντέλο

πρόβλεψης (Predictive Model) το οποίο βασίζεται στον προηγούμενο αλγόριθμο μηχανικής μάθησης και παράγει το αποτέλεσμα πρόβλεψης (Prediction).

2.2.2.1 Κατηγοριοποίηση

Η κατηγοριοποίηση (Classification) είναι μια από τις βασικότερες εργασίες της μηχανικής μάθησης. Κύριος στόχος της κατηγοριοποίησης είναι η πρόβλεψη μιας κατηγοριοποιημένης ετικέτας κατηγοριών νέων περιπτώσεων βασισμένη σε προηγούμενες παρατηρήσεις. Ως επί το πλείστον, η κατηγοριοποίηση είναι χωρισμένη σε δυαδική κατηγοριοποίηση, όπου ο αλγόριθμος μαθαίνει μια σειρά από κανόνες για τη διάκριση των ετικετών μεταξύ δύο πιθανών κατηγοριών, και σε πολλαπλή κατηγοριοποίηση (Multiclass Classification), η οποία κατηγοριοποιεί τα δεδομένα σε περισσότερες από δύο κατηγορίες [9][12]. Η κατηγοριοποίηση μπορεί να περιγραφεί ως μια διαδικασία με δύο στάδια, την εκμάθηση και την κατηγοριοποίηση / εκπαίδευση [13]. Παράδειγμα τέτοιων αλγορίθμων είναι ο k κοντινότεροι γείτονες, η λογιστική παλινδρόμηση, ο απλοϊκός Bayes (Naïve Bayes), οι μηχανές διανυσμάτων υποστήριξης, τα δένδρα απόφασης και τα νευρωνικά δίκτυα.



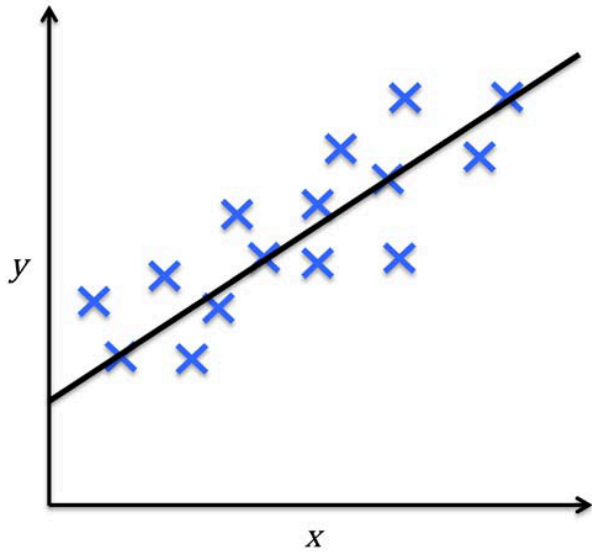
Εικόνα 3: Αναπαράσταση κατηγοριοποίησης.

[12]

Στο παράδειγμα της Εικόνας 3, η κατηγοριοποίηση είναι δυαδική και το σύνολο δεδομένων αποτελείται από τριάντα δείγματα. Η διαγώνια γραμμή είναι η διαχώριση που γίνεται μεταξύ των δειγμάτων στην κατηγοριοποίηση. Παρατηρούμε ότι τα μισά είναι θετικά και τα άλλα μισά αρνητικά.

2.2.2.2 Παλινδρόμηση

Δεύτερος τύπος της επιβλεπόμενης μάθησης είναι η μέθοδος της παλινδρόμησης (Regression), η οποία χρησιμοποιείται για την πρόβλεψη συνεχών αποτελεσμάτων. Στόχος αυτής της μεθόδου είναι να προβλεφθεί ένας συνεχής αριθμός, αφού αρχικά δοθεί ένας αριθμός από τον παράγοντα της μεταβλητής καθώς και μια συνεχής μεταβλητή αποτελέσματος. Με βάση τα πιο πάνω προσπαθούμε να βρούμε μια σχέση μεταξύ των μεταβλητών, η οποία θα μας επιτρέψει να προβλέψουμε το αποτέλεσμα [13]. Αλγόριθμοι που μπορούν να εφαρμόσουν παλινδρόμηση είναι η γραμμική παλινδρόμηση (Linear Regression), τα δένδρα απόφασης, τα νευρωνικά δίκτυα και τα τυχαία δάση.



Εικόνα 4: Αναπαράσταση ευθείας παλινδρόμησης.

[12]

Η Εικόνα 4 απεικονίζει μια ευθεία γραμμικής παλινδρόμησης, όπου γνωρίζοντας τις τιμές χ και ψ αντίστοιχα, εφαρμόζεται η ευθεία γραμμή με σκοπό τη μείωση της απόστασης [12].

2.2.3 Μη επιβλεπόμενη μάθηση

Δεύτερος τύπος μηχανικής μάθησης είναι η μη επιβλεπόμενη μάθηση. Σε αντίθεση με την επιβλεπόμενη μάθηση, όπου γνωρίζαμε εξαρχής τη σωστή απάντηση, στη μη επιβλεπόμενη μάθηση ερχόμαστε αντιμέτωποι με μη κατηγοριοποιημένα δεδομένα ή με δεδομένα με άγνωστη δομή. Χρησιμοποιώντας αυτή την τεχνική, μπορούμε να εξερευνήσουμε τη δομή των δεδομένων και να εξάγουμε σημαντική πληροφορία χωρίς την καθοδήγηση ενός γνωστού αποτελέσματος [13]. Στη μη επιβλεπόμενη μάθηση, ο αλγόριθμος δέχεται απλά τα δεδομένα και ζητείται η εξαγωγή γνώσης από τα δεδομένα αυτά. Όπως και στην επιβλεπόμενη

μάθηση, έτσι και στη μη επιβλεπόμενη μάθηση υπάρχουν δύο υποκατηγορίες, οι οποίες ονομάζονται μετασχηματισμός δεδομένων και συσταδοποίηση. Αλγόριθμοι που εφαρμόζουν μη επιβλεπόμενη μάθηση είναι η ανάλυση κυρίων συνιστωσών (Principal Components Analysis – PCA), η γραμμική διαχωριστική ανάλυση (Linear Discriminant Analysis – LDA), η τυποποίηση (Standardization), η κανονικοποίηση (Normalization), ο K-Means, ο Fuzzy C-Means, ο PAM (Partitioning Around Medoids), ο BIRCH και ο DBSCAN.

2.2.3.1 Μετασχηματισμός

Η χρησιμοποίηση αλγορίθμων για τη δημιουργία νέων αναπαραστάσεων του συνόλου δεδομένων μας ονομάζεται μετασχηματισμός (Transformation) δεδομένων. Οι αναπαραστάσεις αυτές είναι ευκολότερες στην κατανόηση από τους ανθρώπους ή από τους αλγορίθμους μηχανικής μάθησης, έτσι ώστε να είναι εφικτή η καλύτερη σύγκριση των δεδομένων από την αρχική [9]. Η πιο κοινή εφαρμογή αλγορίθμων μετασχηματισμού είναι η μείωση διαστάσεων (Dimensionality Reduction). Αυτό οφείλεται στο ότι τα αρχικά δεδομένα τα οποία χρησιμοποιούμε και επεξεργαζόμαστε έχουν αρκετές διαστάσεις, πράγμα το οποίο μπορεί να παρουσιάσει πρόκληση όταν υπάρχει περιορισμένος χώρος αποθήκευσης και υπολογιστικής απόδοσης των αλγορίθμων μηχανικής μάθησης. Η χρήση της μείωσης διαστάσεων γίνεται κυρίως κατά τη διαδικασία της προεπεξεργασίας όπου αφαιρείται ο θόρυβος από τα δεδομένα, καθώς και η συμπίεση των δεδομένων σε ένα μικρότερο διάστημα τιμών, διατηρώντας όμως την περισσότερη σχετική πληροφορία [13]. Αλγόριθμοι οι οποίοι χρησιμοποιούνται για μείωση διαστάσεων είναι η ανάλυση κυρίων συνιστωσών, η γραμμική διαχωριστική ανάλυση, η κανονικοποίηση και η τυποποίηση.

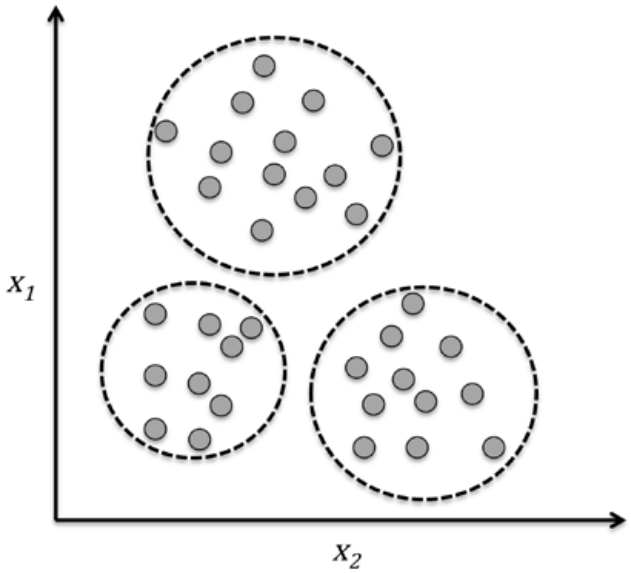
2.2.3.2 Συσταδοποίηση

Δεύτερος τύπος μη επιβλεπόμενης μάθησης είναι η μέθοδος συσταδοποίησης (Clustering), μιας διερευνητικής τεχνικής ανάλυσης δεδομένων, που μας επιτρέπει να οργανώσουμε μια στοίβα από πληροφορίες σε σημαντικές υποομάδες (clusters), χωρίς όμως να υπάρχει κάποια προηγούμενη γνώση των μελών της ομάδας τους. Κάθε υποομάδα (cluster) η οποία μπορεί να προκύψει κατά τη διάρκεια της ανάλυσης, ορίζεται από μια ομάδα αντικειμένων που έχουν κάποια κοινά χαρακτηριστικά, αλλά είναι και περισσότερο ανόμοια με αντικείμενα από τις άλλες υποομάδες (clusters) [13]. Ανάλογα με τη μέθοδο που θα χρησιμοποιηθεί για τον καθορισμό των συστάδων, οι αλγόριθμοι κατηγοριοποιούνται σε έξι είδη:

- Διαιρετική συσταδοποίηση (Partitional Clustering).

- Ιεραρχική συσταδοποίηση (Hierarchical Clustering).
- Συσταδοποίηση βασισμένη στην πυκνότητα (Density-based Clustering).
- Συσταδοποίηση βασισμένη σε πλέγμα (Grid-based Clustering).
- Συσταδοποίηση υποχώρων (Subspace Clustering).
- Ασαφής συσταδοποίηση (Fuzzy Clustering). [13]

Σε αντίθεση με την κατηγοριοποίηση, η συσταδοποίηση δεν βασίζεται σε προκαθορισμένες κατηγορίες. Παραδείγματα τέτοιων αλγόριθμων είναι ο SVD, ο K-Means, ο Fuzzy C-Means, ο PAM, ο BIRCH και ο DBSCAN.



Εικόνα 5: Αναπαράσταση συσταδοποίησης.

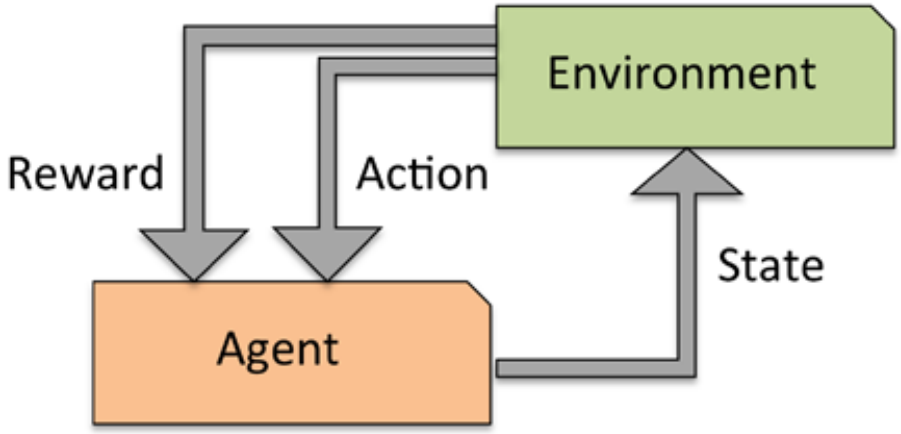
[12]

Η Εικόνα 5 παρουσιάζει την εφαρμογή της συσταδοποίησης σε ένα σύνολο δεδομένων. Οργανώνει τα δεδομένα αυτά σε τρεις υποομάδες, με κοινό χαρακτηριστικό μεταξύ τους τις τιμές των x_1 και x_2.

2.2.4 Ενισχυτική μάθηση

Εκτός από την επιβλεπόμενη και μη-επιβλεπόμενη μάθηση, οι οποίες είναι τα σημαντικότερα είδη μάθησης, άλλο ένα είδος μάθησης είναι η ενισχυτική μάθηση (Reinforcement Learning). Σκοπός του συστήματος είναι η μεγιστοποίηση της αριθμητικής τιμής της ανταμοιβής. Η μέθοδος αυτή χρησιμοποιεί τρεις μεθόδους (Components), τον πράκτορα (Agent), το περιβάλλον (Environment) και την ενέργεια (Action). Ο πράκτορας είναι η οντότητα η οποία μαθαίνει και παίρνει διάφορες αποφάσεις, ενώ οτιδήποτε άλλο είναι το περιβάλλον. Υπάρχει μια συνεχείς αλληλεπίδραση μεταξύ πράκτορα και περιβάλλοντος, όπου

ο πράκτορας επιλέγει διάφορες ενέργειες για να πραγματοποιήσει το περιβάλλον, παρουσιάζοντας του έτσι καινούριες καταστάσεις. Το περιβάλλον δίνει στον πράκτορα διάφορες ανταμοιβές, όπου με βάση και το σκοπό του συστήματος προσπαθεί να μεγιστοποιήσει μακροπρόθεσμα [10].



Εικόνα 6: Διαδικασία ενισχυτικής μάθησης.

[3]

Η Εικόνα 6 παρουσιάζει την αλληλεπίδραση που έχει ο πράκτορας με το περιβάλλον. Πιο συγκεκριμένα, ο πράκτορας βρίσκεται σε ένα περιβάλλον το οποίο περιγράφεται από διάφορες πιθανές καταστάσεις (States), όπου κάθε φορά που εκτελεί μια ενέργεια σε αυτή την κατάσταση, λαμβάνει την ανταμοιβή του (Reward). [10] Το καλύτερο παράδειγμα για την ενισχυτική μάθηση είναι τα παιχνίδια, όπως το σκάκι, όπου ο πράκτορας αποφασίζει για διάφορες κινήσεις, οι οποίες θα του επιφέρουν την ανταμοιβή αν κερδίσει ή αν χάσει [12].

Ο Πίνακας 2 που ακολουθεί είναι ένας συνοπτικός συγκριτικός πίνακας των ειδών μηχανικής μάθησης.

ΕΠΙΒΛΕΠΟΜΕΝΗ – ΜΗ ΕΠΙΒΛΕΠΟΜΕΝΗ – ΕΝΙΣΧΥΤΙΚΗ ΜΑΘΗΣΗ		
ΕΠΙΒΛΕΠΟΜΕΝΗ ΜΑΘΗΣΗ	**ΜΗ ΕΠΙΒΛΕΠΟΜΕΝΗ ΜΑΘΗΣΗ**	**ΕΝΙΣΧΥΤΙΚΗ ΜΑΘΗΣΗ**
• Τα δεδομένα εισόδου επισημαίνονται. • Χρησιμοποιεί σύνολο εκπαίδευσης.	• Τα δεδομένα εισόδου δεν επισημαίνονται. • Χρησιμοποιεί μόνο το σύνολο δεδομένων.	• Παίρνει αποφάσεις με βάση την πείρα που έχει αποκτήσει.

• Χρησιμοποιείται για πρόβλεψη. • Κατηγοριοποίηση και παλινδρόμηση.	• Χρησιμοποιείται για ανάλυση. • Συσταδοποίηση και μετασχηματισμός δεδομένων.	• Δύσκολες αποφάσεις για περίπλοκα προβλήματα. • Συστήματα ανταμοιβής.

Πίνακας 2: Συγκριτικός πίνακας επιβλεπόμενης, μη επιβλεπόμενης και ενισχυτικής μάθησης.

2.2.5 Μέθοδοι αξιολόγησης μοντέλου

Για την αξιολόγηση του μοντέλου μας δεν μπορούμε να χρησιμοποιήσουμε τα δεδομένα που χρησιμοποιήσαμε για την κατασκευή του μοντέλου της μηχανικής μάθησης. Αυτό οφείλεται στο ότι το μοντέλο το οποίο δημιουργείται, έχει τη δυνατότητα να θυμάται ολόκληρο το σύνολο εκπαίδευσης, επομένως θα κατηγοριοποιεί συνεχώς σωστά τα οποιαδήποτε δεδομένα από το σύνολο εκπαίδευσης. Η εμφάνιση όμως νέων δεδομένων χωρίς την ετικέτα κατηγοριοποίησης βοηθάει στην αξιολόγηση της απόδοσης του μοντέλου. Συγκεκριμένα, γίνεται διαχωρισμός των δεδομένων που συλλέγονται σε δύο μέρη, όπου το ένα μέρος χρησιμοποιείται για την κατασκευή του μοντέλου μηχανικής μάθησης και ονομάζεται σύνολο εκπαίδευσης, ενώ τα δεδομένα τα οποία θα χρησιμοποιηθούν για να εκτιμηθεί το ποσοστό αποτελεσματικότητας του μοντέλου ονομάζεται σύνολο ελέγχου [9].

Το βασικότερο μέτρο το οποίο χρησιμοποιείται για την αξιολόγηση της αποτελεσματικότητας ενός συστήματος είναι η ορθότητα (Accuracy) [14]. Άλλα μέτρα τα οποία είναι χρήσιμα για την αξιολόγηση του μοντέλου είναι η ακρίβεια (Precision), η ανάκληση (Recall) και ο αρμονικός μέσος (F measure). Για τον υπολογισμό των μέτρων αυτών σημαντικό ρόλο παίζει η εύρεση του πίνακα σύγχυσης (Confusion Matrix), αφού διευκολύνεται ο υπολογισμός τους.

2.2.5.1 Ορθότητα μοντέλου

Η εύρεση της ορθότητας του μοντέλου κατηγοριοποίησης είναι το σημαντικότερο κριτήριο το οποίο χρησιμοποιείται στη διαδικασία κατηγοριοποίησης, αφού επιτρέπει σε αυτόν που έφτιαξε το μοντέλο μηχανικής μάθησης να αξιολογήσει το ποσοστό ακρίβειας κατηγοριοποίησης των μελλοντικών δεδομένων [9].

Η ορθότητα του μοντέλου σε ένα σύνολο δεδομένων είναι το ποσοστό των δειγμάτων τα οποία κατηγοριοποιήθηκαν σωστά στο μοντέλο εκπαίδευσης είτε η πρόβλεψη είναι θετική

είτε είναι αρνητική. Σε περίπτωση που η ακρίβεια του μοντέλου που κατασκευάστηκε είναι αποδεκτή, τότε το μοντέλο αυτό μπορεί πλέον να χρησιμοποιηθεί για την κατηγοριοποίηση νέων αγνώστων δεδομένων [13].

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

Εξίσωση 1: Τύπος ορθότητας

Με άλλα λόγια, η ορθότητα μπορεί να οριστεί σαν ο αριθμός των σωστών προβλέψεων θετικών και αρνητικών δειγμάτων, διαιρούμενος με το συνολικό αριθμό όλων των δειγμάτων.

2.2.5.2 Μέτρα αποτελεσματικότητας κατηγοριοποίησης

Τα σημαντικότερα μέτρα αποτελεσματικότητας ενός αλγορίθμου κατηγοριοποίησης, όσον αφορά την ακρίβεια των προβλέψεών του είναι η ακρίβεια, η ανάκληση και ο αρμονικός μέσος.

Η ακρίβεια ορίζεται ως μέτρο αποτελεσματικότητας, όταν ο στόχος μας είναι ο περιορισμός των ψευδών θετικών κατηγοριοποιημένων στοιχείων. Είναι σημαντικό το μοντέλο να έχει υψηλή ακρίβεια, πράγμα που σημαίνει ότι δεν υπάρχει παραγωγή πολλών ψεύτικων θετικών κατηγοριοποιημένων στοιχείων. Η ακρίβεια είναι γνωστή και ως θετική τιμή πρόβλεψης (Positive Predictive Value - PPV).

Η ανάκληση ορίζεται ως μέτρο αποτελεσματικότητας, όταν ο στόχος μας είναι ο εντοπισμός όλων των θετικών δειγμάτων. Άλλα ονόματα για την ανάκληση είναι η ευαισθησία (Sensitivity), ποσοστό επιτυχίας ή πραγματικό θετικό ποσοστό (True Positive Rate - TPR) [9].

Πολλές φορές τα αποτελέσματα της ακρίβειας και της ανάκλησης δεν μας δίνουν ακριβή εικόνα για τα αποτελέσματα της κατηγοριοποίησης. Ένα άλλο μέτρο που μας δίνει καλύτερα αποτελέσματα είναι ο αρμονικός μέσος, ο οποίος συνδυάζει τα δύο πιο πάνω μέτρα.

Ο Πίνακας 3 παρουσιάζει τους τύπους της ορθότητας, ακρίβειας, ανάκλησης και αρμονικού μέσου.

ΤΥΠΟΙ ΥΠΟΛΟΓΙΣΜΟΥ	
Ορθότητα	$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$
Ακρίβεια	$Precision = \frac{TP}{TP + FP}$
Ανάκληση	$TPR = \frac{TP}{TP + FN}$
Ειδικότητα	$1 - FPR = (1 - \frac{FP}{FP + TN})$
Αρμονικός μέσος	$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$

Πίνακας 3: Συγκεντρωτική παρουσίαση των μέτρων των τύπων που υπολογίζονται με τη χρήση του πίνακα σύγχυσης.

2.2.5.3 Πίνακας σύγχυσης

Ο πίνακας σύγχυσης (Confusion Matrix) παρέχει πληροφορίες σχετικά με τις πραγματικές και προβλεπόμενες κατηγοριοποιήσεις που πραγματοποιούνται από ένα σύστημα κατηγοριοποίησης. Η απόδοση αξιολογείται χρησιμοποιώντας τα δεδομένα στη μήτρα, τα οποία συμβολίζονται TP (True Positive), FN (False Negative), FP (False Positive), TN (True Negative) όπου το κάθε ένα συμβολίζει:

- TP: πλήθος θετικών δειγμάτων που έχουν προβλεφθεί σωστά από το μοντέλο.
- FN: πλήθος θετικών δειγμάτων που έχουν προβλεφθεί εσφαλμένα ως αρνητικά.
- FP: πλήθος αρνητικών δειγμάτων που έχουν προβλεφθεί εσφαλμένα ως θετικά.
- TN: πλήθος αρνητικών δειγμάτων που έχουν προβλεφθεί σωστά από το μοντέλο.

Οι μετρήσεις του πίνακα μπορούν να εκφραστούν με ποσοστά. Τα ποσοστά αυτά ορίζονται ως ακολούθως:

- $TPR = \frac{TP}{TP+FN}$ ποσοστό θετικών δειγμάτων που έχουν προβλεφθεί σωστά.
- $TNR = \frac{TN}{TN+FP}$ ποσοστό αρνητικών δειγμάτων που έχουν προβλεφθεί σωστά.
- $FPR = \frac{FP}{TN+FP}$ ποσοστό αρνητικών δειγμάτων που έχουν προβλεφθεί ως θετικά.
- $FNR = \frac{FN}{TP+FN}$ ποσοστό θετικών δειγμάτων που έχουν προβλεφθεί ως αρνητικά.

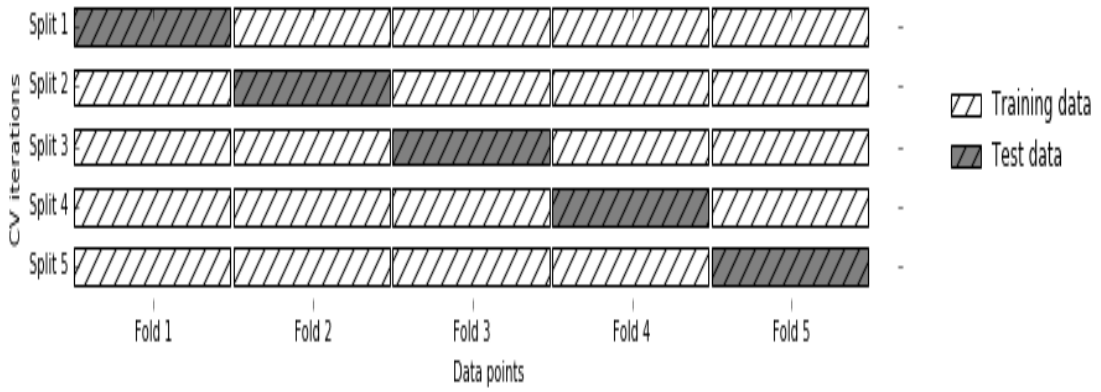
Με βάση τον πίνακα σύγχυσης μπορεί να υπολογιστούν η ορθότητα, η ακρίβεια, η ανάκληση, ο αρμονικός μέσος και η ειδικότητα. Η ακρίβεια, όπως αναφέρθηκε πιο πάνω, είναι το μέτρο που αξιολογεί το ποσοστό των δειγμάτων που κατηγοριοποιούνται ως θετικά και τα οποία είναι πραγματικά θετικά. Η ανάκληση αξιολογεί πόσο καλά ο κατηγοριοποιητής μπορεί να αναγνωρίσει τα θετικά δείγματα και η ειδικότητα μετρά πόσο καλά ο κατηγοριοποιητής μπορεί να αναγνωρίσει τα αρνητικά δείγματα. Τέλος, ο αρμονικός μέσος υπολογίζεται με βάση τα αποτελέσματα της ακρίβειας και της ανάκλησης. Στον Πίνακα 4 που ακολουθεί, παρουσιάζεται ο πίνακας σύγχυσης για δυαδική κατηγοριοποίηση.

ΠΙΝΑΚΑΣ ΣΥΓΧΥΣΗΣ			
	Θετικά	**Αρνητικά**	**Κλάση**
Πραγματικά	TP (αληθή θετικά)	FN (ψευδή αρνητικά)	Θετικά
	FP (ψευδή θετικά)	TN (αληθή αρνητικά)	Αρνητικά

Πίνακας 4: Παρουσίαση πίνακα σύγχυσης για δυαδική κατηγοριοποίηση.

2.2.5.4 K – Fold Cross Validation

Στη μέθοδο K – Fold Cross Validation το αρχικό σύνολο δεδομένων χωρίζεται σε k υποσύνολα, τα οποία ονομάζονται “folds”. Αυτά τα υποσύνολα είναι αμοιβαία αποκλειόμενα και έχουν περίπου το ίδιο μέγεθος, όπου ο κατηγοριοποιητής εκπαιδεύεται επαναληπτικά και εξετάζεται k φορές. Σε κάθε επανάληψη, το υποσύνολο διατηρείται ως σύνολο δοκιμής ενώ τα υπόλοιπα υποσύνολα χρησιμοποιούνται για να εκπαιδεύσουν τον κατηγοριοποιητή. Ο πιο συχνός αριθμός k που καθορίζεται από τον χρήστη είναι συνήθως 5 ή 10.



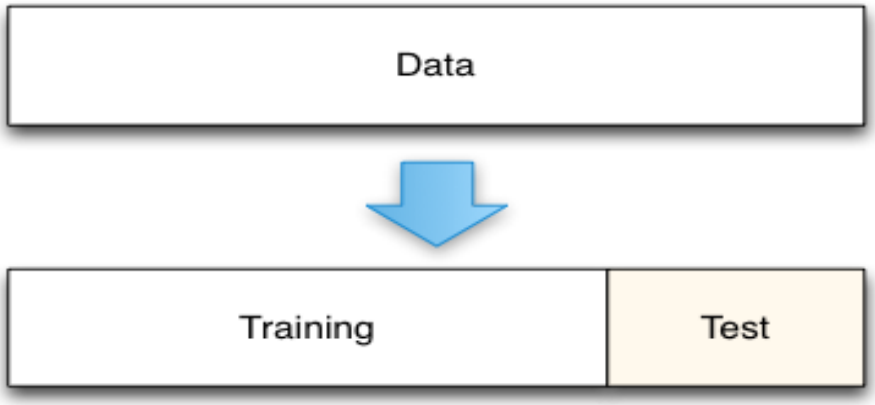
Εικόνα 7: Παράδειγμα μεθόδου K - Fold Cross Validation με αριθμό folds = 5.

[9]

Η Εικόνα 7 παρουσιάζει ένα σύνολο δεδομένων που χωρίζεται σε πέντε υποσύνολα, όπου την κάθε φορά το ένα υποσύνολο από τα υπόλοιπα χρησιμοποιείται για σύνολο εκπαίδευσης.

2.2.5.5 Split

Η μέθοδος Split είναι η πιο κλασική και δημοφιλής προσέγγιση για το διαχωρισμό των δεδομένων σε σύνολο εκπαίδευσης και σύνολο δοκιμής. Ονομάζεται επίσης και Hold – out μέθοδος. Αν για παράδειγμα η μέθοδος Split είναι 80%, αυτό σημαίνει ότι ένα τυχαίο ποσοστό του 80% από το σύνολο δεδομένων θα χρησιμοποιηθεί για σύνολο εκπαίδευσης και το υπόλοιπο 20% για σύνολο δοκιμής.



Εικόνα 8: Διαχωρισμός συνόλου δεδομένων με τη μέθοδο Split.

2.2.5.6 Καμπύλη χαρακτηριστικών λειτουργίας δέκτη

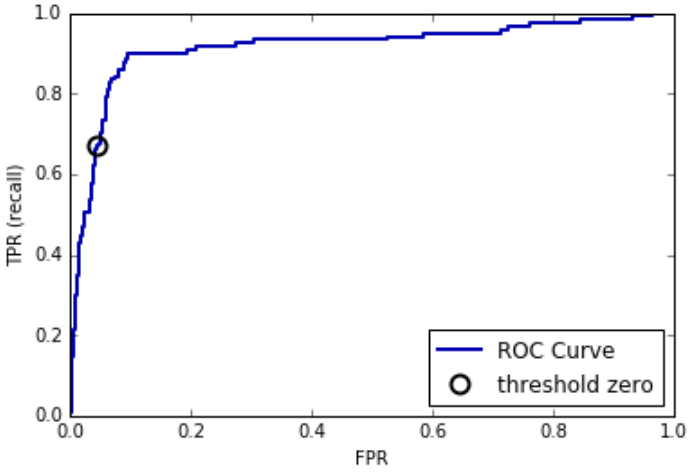
Το γράφημα καμπύλης χαρακτηριστικών λειτουργίας δέκτη (Receiver Operating Characteristic - ROC) είναι ένα χρήσιμο εργαλείο το οποίο χρησιμοποιείται για την ανάλυση της απόδοσης του κατηγοριοποιητή, υπολογίζοντας τις ψευδείς θετικές και τις πραγματικές θετικές τιμές με τη μετατόπιση του κατωφλίου απόφασης του κατηγοριοποιητή. Η καμπύλη ROC παρόλο που εξετάζει όλα τα κατώτατα όρια για ένα δεδομένο κατηγοριοποιητή, δεν επιστρέφει την ακρίβεια και την ανάκληση, αλλά δείχνει την ψευδή θετική συχνότητα (False Positive Rate) έναντι του πραγματικού θετικού ρυθμού (True Positive Rate).

Η διαγώνιος της καμπύλης ROC ονομάζεται τυχαία εικασία και τα μοντέλα τα οποία είναι κάτω από αυτή θεωρούνται χειρότερα, σε σχέση με αυτά που είναι πάνω από αυτή. Όσο ένας κατηγοριοποιητής βρίσκεται πιο κοντά στην αριστερή πάνω γωνία της καμπύλης, τόσο καλύτερος είναι αφού έχει πραγματική θετική τιμή 1 και ψευδείς θετική τιμή 0 [13]. Αντίθετα,

όσο πιο κοντά βρίσκεται το μοντέλο στη διαγώνιο, τόσο λιγότερο ακριβές είναι το μοντέλο, ενώ αν πέσει ακριβώς πάνω στη διαγώνιο τότε είναι εντελώς τυχαίο.

Για την κατασκευή της καμπύλης ROC ακολουθείται η πιο κάτω διαδικασία:

1. Ταξινομούνται οι εγγραφές ελέγχου σε αύξουσα σειρά υποθέτοντας ως έξοδο τιμών τη θετική κατηγορία.
2. Επιλέγεται η χαμηλότερη σε σειρά εγγραφή ελέγχου.
3. Επιλέγεται η επόμενη εγγραφή. Αν η εγγραφή ανήκει στη θετική κατηγορία το TP μειώνεται και το FP παραμένει το ίδιο, ενώ αν η εγγραφή που επιλέχτηκε ανήκει στην αρνητική κατηγορία τότε το FP μειώνεται και το TP μένει το ίδιο.



Εικόνα 9: Καμπύλη χαρακτηριστικών λειτουργίας δέκτη.

[9]

2.3 Θεωρία αλγορίθμων μηχανικής μάθησης

2.3.1 Απλοϊκός Bayes

Ο κατηγοριοποιητής απλοϊκός Bayes (Naïve Bayes) κατηγοριοποιεί ένα δείγμα χρησιμοποιώντας το απλοϊκό μοντέλο Bayes, υποθέτοντας ότι τα γνωρίσματα είναι κατά συνθήκη ανεξάρτητα μεταξύ τους, με δεδομένη την κατηγορία [15]. Έστω ότι υπάρχει ένα σύνολο δεδομένων D το οποίο αναπαρίσταται από ένα n-διάστατο διάνυσμα $X = (x_1, x_2, \ldots, x_n)$ που είναι οι μετρήσεις στο κάθε δείγμα του συνόλου για n χαρακτηριστικά. Υποθέτουμε ότι υπάρχουν m κατηγορίες $C_1, C_2, \ldots, C_m$ μιας κλάσης C, η οποία πρέπει να προβλεφθεί η κατηγοριοποίησή της. Για την κατηγοριοποίηση ενός δείγματος X υπολογίζουμε τις πιθανότητες $P(C_1 \mid X), P(C_2 \mid X), \ldots, P(C_m \mid X)$, δηλαδή ποια η πιθανότητα του δείγματος να ανήκει στην κατηγορία $C_1, C_2, \ldots, C_m$ αντίστοιχα. Το δείγμα X κατηγοριοποιείται σε εκείνη την κατηγορία της οποίας η πιθανότητα $P(C \mid X)$ είναι η μέγιστη. Για την εύρεση της

πιθανότητας ενός δείγματος X να ανήκει σε μια κατηγορία C_i, χρησιμοποιείται το θεώρημα Bayes οπότε έχουμε:

$$P(C_i/X) = \frac{P(X/C_i) \cdot P(C_i)}{P(X)}$$

Εξίσωση 2: Θεώρημα Bayes

Το P(X) είναι το ίδιο σε όλες τις κατηγορίες, οπότε εστιάζουμε μόνο στον αριθμητή $P(X/C_i) \cdot P(C_i)$. Για τον υπολογισμό της πιθανότητας P(X | C_i) κάνουμε την "απλοϊκή" παραδοχή ότι η επίδραση ενός χαρακτηριστικού σε μια κατηγορία C_i είναι ανεξάρτητη από τις τιμές των άλλων χαρακτηριστικών. Άρα έχουμε:

$$P(X/C_i) = \prod_{k=1}^{n} P(X_k/C_i) = P(X_1/C_i)P(X_2/C_i)\dots P(X_n/C_i)$$

Εξίσωση 3: Απλοϊκός Bayes

Η μάθηση του απλοϊκού μοντέλου Bayes αποδίδει πολύ καλά σε ένα μεγάλο εύρος εφαρμογών, αφού είναι ένας από τους πιο αποτελεσματικούς αλγόριθμους μηχανικής μάθησης, ο οποίος μπορεί να κλιμακωθεί καλά σε πολύ μεγάλα προβλήματα. Τέλος, τα δεδομένα τα οποία περιέχουν κάποιο θόρυβο δεν επηρεάζουν τον αλγόριθμο, δίνοντας έτσι απαντήσεις στο πρόβλημα της κατηγοριοποίησης [15].

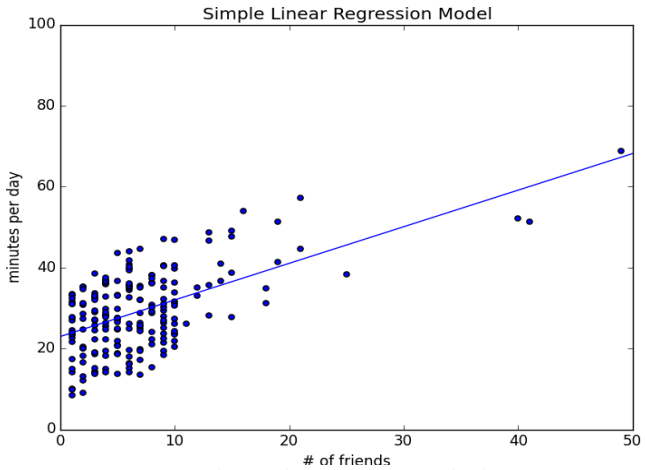
2.3.2 Γραμμική παλινδρόμηση

Ένα μοντέλο της μορφής $Y = \beta_0 + \beta_1 x$, το οποίο περιγράφει ένα τρόπο πρόβλεψης της μεταβλητής X μέσω μιας γραμμικής συνάρτησης του X, λέγεται μοντέλο γραμμικής παλινδρόμησης του Y πάνω στη μεταβλητή X. Με την ανάλυση παλινδρόμησης εξετάζουμε τη σχέση μεταξύ δύο ή περισσότερων μεταβλητών, με σκοπό την πρόβλεψη των τιμών της μιας μέσω των τιμών της άλλης. Υπάρχουν δύο είδη μεταβλητών:

- Ανεξάρτητες μεταβλητές: Η τιμή αυτής της μεταβλητής είναι ελεγχόμενη από εμάς.
- Εξαρτημένες μεταβλητές: Η τιμή προσδιορίζεται από την τιμή που έχει πάρει η ανεξάρτητη μεταβλητή.

Σκοπός είναι η εύρεση μιας μαθηματικής σχέσης. Αν θεωρήσουμε ότι X είναι η ανεξάρτητη μεταβλητή και Y η εξαρτημένη μεταβλητή, τότε μια σχέση μπορεί να είναι της μορφής Y=f(X). Αυτό σημαίνει ότι τη δεδομένη στιγμή x η συνάρτηση αυτή μας δίνει την αντίστοιχη τιμή y της μεταβλητής Y. Όμως, η διάκριση μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών δεν είναι πάντα σαφής. Για παράδειγμα, αν θεωρήσουμε βάρος και ύψος ατόμου ως μεταβλητές δεν είναι σαφές ποια θα θεωρηθεί εξαρτημένη και ποια ανεξάρτητη. Αυτό οφείλεται στο ότι αν επιλέξουμε άτομα με συγκεκριμένα βάρη και εξετάσουμε τα ύψη τους ως

ανεξάρτητη μεταβλητή παίρνουμε το βάρος και ως εξαρτημένη το ύψος. Ενώ αν επιλέξουμε άτομα με συγκεκριμένα ύψη και εξετάσουμε το βάρος τους, ισχύει το αντίθετο. Η πρώτη εφαρμογή του μοντέλου της γραμμικής παλινδρόμησης έγινε από τον Legendre το 1805 και από τον Gauss το 1809.



Εικόνα 10: Ευθεία απλής γραμμικής παλινδρόμησης.

[23]

Η εικόνα 10 παρουσιάζει την ευθεία γραμμικής παλινδρόμησης μεταξύ των μεταβλητών: αριθμός φίλων (# of friends) και λεπτά την ημέρα (minutes per day).

2.3.2.1 Λογιστική παλινδρόμηση

Η λογιστική παλινδρόμηση (Logistic Regression) δημιουργήθηκε το 1958 από τον στατιστικολόγο David Cox. Είναι ένα γραμμικό μοντέλο κατηγοριοποίησης το οποίο χρησιμοποιείται ευρέως στον τομέα της εξόρυξης δεδομένων, όπου στις περισσότερες περιπτώσεις βοηθάει στην επίλυση προβλημάτων δύο κατηγοριών, όπως η αυτόματη διάγνωση ασθενειών και η οικονομική πρόβλεψη. Ο αλγόριθμος της λογιστικής παλινδρόμησης βασίζεται στο μοντέλο της γραμμικής παλινδρόμησης και εκφράζεται ως εξής:

$$P = a + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m$$

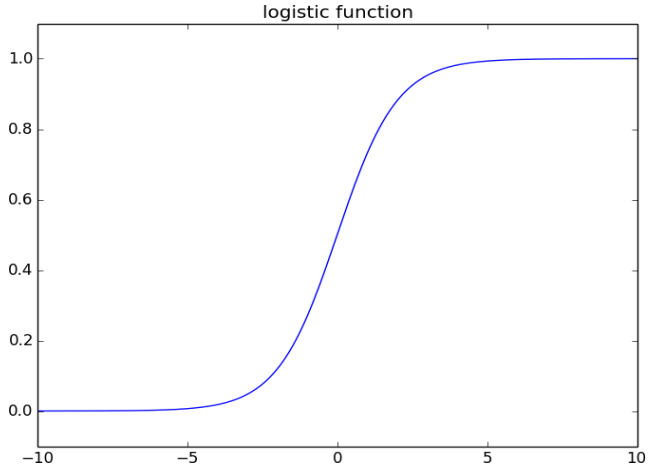
Εξίσωση 4: Λογιστική παλινδρόμηση

Η λογιστική παλινδρόμηση μοιάζει πολύ με τη γραμμική παλινδρόμηση, αλλά δεν μπορεί να πραγματοποιηθεί μέσω αυτής, λόγω της τιμής η οποία υπολογίζεται για μια μεταβλητή. Στη γραμμική παλινδρόμηση η τιμή της μεταβλητής, η οποία υπολογίζεται είναι συνεχής, σε αντίθεση με την τιμή στη λογιστική παλινδρόμηση όπου η τιμή είναι διακριτή [15]. Η τιμή αυτή κυμαίνεται μεταξύ των τιμών 0 και 1, όμως με τη χρήση της γραμμικής παλινδρόμησης η τιμή αυτή μπορεί να είναι μεγαλύτερη του 1 ή μικρότερη του 0. Σε αντίθεση με τη γραμμική παλινδρόμηση, η λογιστική παλινδρόμηση μειώνει το εύρος πρόβλεψης και η τιμή που παίρνει είναι πάντα μεταξύ του 0 και του 1 [7].

Για την ανάλυση της λογιστικής παλινδρόμησης υπολογίζεται αρχικά ο λόγος πιθανοτήτων που ονομάζεται odds. Αν θεωρήσουμε ότι p είναι η πιθανότητα επιτυχίας εμφάνισης του γεγονότος και 1-p η πιθανότητα αποτυχίας εμφάνισης του γεγονότος, τότε ο λόγος πιθανοτήτων υπολογίζεται από τον τύπο $odds = \frac{p}{1-p}$. Τέλος, καθορίζεται η λειτουργία logit, η οποία είναι ο φυσικός λογάριθμος του λόγου πιθανότητας έτσι ώστε να μπορεί να ενσωματωθεί στο μοντέλο παλινδρόμησης και συμβολίζεται σαν:

$$logit(p) = ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_n X_n}}$$ [5]

Εξίσωση 5: Φυσικός λογάριθμος του λόγου πιθανότητας



Εικόνα 11: Σιγμοειδής συνάρτηση.

[17]

2.3.3 Νευρωνικά δίκτυα

Τα τεχνητά νευρωνικά δίκτυα (Neural Networks) ήταν ο πρωταρχικός στόχος της τεχνητής νοημοσύνης. Είναι μια εναλλακτική τεχνική μάθησης, η οποία μοιάζει πολύ με τη λειτουργία του εγκεφάλου στην προσπάθεια επίλυσης περίπλοκων προβλημάτων. Πρώτοι που εφάρμοσαν τη χρήση των νευρωνικών δικτύων ήταν ο Warren McCulloch και Walter Pitt το 1940. Εναλλακτικά ονόματα των νευρωνικών δικτύων είναι ο συνδεσμισμός (Connectionism), παράλληλη κατανεμημένη επεξεργασία (Parallel Distributed Processing) και νευρωνικός υπολογισμός (Neural Computation)[12][15]. Τα τεχνητά νευρωνικά δίκτυα είναι πολύπλοκες μη γραμμικές συναρτήσεις με πολλές παραμέτρους. Είναι ιδιαίτερα δημοφιλή στην επίλυση προβλημάτων κατηγοριοποίησης, πρόβλεψης, αποτίμησης (Assesment) και αναγνώρισης (Recognition). Σημαντικότερο πλεονέκτημα που παρέχουν τα νευρωνικά δίκτυα, σε σχέση με άλλους αλγορίθμους, είναι ότι παρουσιάζουν ανοχή σε δεδομένα εκπαίδευσης με θόρυβο, ενώ

το μειονέκτημά τους είναι ότι αδυνατούν να εξηγήσουν ποιοτικά τη γνώση που μοντελοποιούν. Για την κατασκευή ενός νευρωνικού δικτύου ακολουθούνται τα εξής βήματα:

- Αναγνώριση των χαρακτηριστικών εισόδου και εξόδου.
- Κατασκευή δικτύου με την κατάλληλη τοπολογία.
- Εκπαίδευση δικτύου με βάση σύνολο δεδομένων κατάλληλα διαμορφωμένο ώστε το δίκτυο να μπορεί να αναγνωρίσει τα πρότυπα.
- Έλεγχος δικτύου χρησιμοποιώντας το σύνολο ελέγχου το οποίο είναι ανεξάρτητο από το σύνολο εκπαίδευσης.

Οι νευρώνες οργανώνονται σε τρία κύρια επίπεδα τα οποία είναι το επίπεδο εισόδου (Input Layer), το κρυφό επίπεδο (Hidden Layer) και το επίπεδο εξόδου (Output Layer). Το επίπεδο εισόδου αποτελείται από τους νευρώνες, οι οποίοι θα χρησιμοποιηθούν για την εισαγωγή των δεδομένων μας στο νευρωνικό δίκτυο. Στο κρυφό επίπεδο ένα νευρωνικό δίκτυο μπορεί να παρέχει ένα ή περισσότερα κρυφά επίπεδα. Τέλος, το επίπεδο εξόδου αποτελείται από την έξοδο του νευρωνικού δικτύου όπου υπάρχει ένα κόμβος για κάθε κατηγορία.

Αρχικά, οι νευρώνες πολλαπλασιάζουν κάθε είσοδο τους με το αντίστοιχο βάρος υπολογίζοντας το ολικό άθροισμα το οποίο δίνεται από τη σχέση:

$$in_i = \sum_{j=0}^{n} W_{j,i}\alpha_j$$

Εξίσωση 6: Συνολικό άθροισμα νευρώνα

Από τον πιο πάνω τύπο προκύπτει ότι το $W_{j,i}$ είναι το αριθμητικό βάρος (Weight), το οποίο προσδιορίζει την ισχύ και το πρόσημο της σύνδεσης και το α_j είναι η διάδοση ενεργοποίησης (Activation) από το j στο i. Στη συνέχεια το άθροισμα τροφοδοτεί τη συνάρτηση ενεργοποίησης, η οποία μπορεί να είναι διαφορετική σε κάθε νευρώνα, για την παραγωγή της εξόδου. Η συνάρτηση ενεργοποίησης υπολογίζεται από τον τύπο:

$$a_i = g(in_i) = g(\sum_{j=0}^{n} w_{i,j}a_j)$$

Εξίσωση 7: Συνάρτηση ενεργοποίησης

Τέτοια συνάρτηση μπορεί να είναι η βηματική συνάρτηση ή συνάρτηση κατωφλίου (Threshold), η συνάρτηση προσήμου (Signum Function), η σιγμοειδής συνάρτηση (Sigmoid Function) και η γραμμική συνάρτηση (Linear Function). Οι τιμές που μπορούν να πάρουν τα σήματα εξόδου είναι ανάλογες με τη συνάρτηση ενεργοποίησης που εφαρμόζεται. Όταν η συνάρτηση ενεργοποίησης είναι η βηματική συνάρτηση, η έξοδος μπορεί να είναι 0 ή 1. Στην περίπτωση που χρησιμοποιείται η συνάρτηση προσήμου, η έξοδος μπορεί να είναι 1 ή -1.

Τέλος, όταν χρησιμοποιείται η σιγμοειδής συνάρτηση, η έξοδος μπορεί να είναι μεταξύ του 0 και του 1. Οι νευρώνες του επιπέδου εισόδου δεν έχουν συνάρτηση ενεργοποίησης.

Τα γνωστότερα είδη νευρωνικών δικτύων παρουσιάζονται στον Πίνακα 7 που ακολουθεί:

ΟΝΟΜΑ	ΚΑΤΑΣΚΕΥΑΣΤΗΣ	ΕΤΟΣ	ΤΡΟΠΟΣ ΕΚΠΑΙΔΕΥΣΗΣ
Perceptron	Rosenblatt	1957 - 1962	Με επίβλεψη
Adaline / Madaline	Widrow	1960 -1962	Με επίβλεψη
Back - propagation	Werbow, Rumelhart etal	1974 - 1986	Με επίβλεψη
Self-organizing map	Kohonen	1981	Χωρίς επίβλεψη
Hopfield net	Hopfield	1982	Με επίβλεψη
Boltzmann machine	Hinton, Hopkins, Szu	1985 - 1986	Με επίβλεψη

Πίνακας 4: Διάφορες πληροφορίες σχετικά με τα γνωστότερα νευρωνικά δίκτυα.

2.3.3.1 Αρχιτεκτονικές νευρωνικών δικτύων

Οι νευρώνες που αποτελούν τα επίπεδα ενός νευρωνικού δικτύου μπορεί να συνδέονται με όλους τους νευρώνες του επόμενου επιπέδου ή να υπάρχουν νευρώνες που να μην συνδέονται με όλους τους νευρώνες του επόμενου επιπέδου. Στην πρώτη περίπτωση, τα νευρωνικά αυτά δίκτυα ονομάζονται πλήρως συνδεδεμένα (Fully Connected), ενώ στη δεύτερη περίπτωση ονομάζονται μερικώς συνδεδεμένα (Partially Connected).

Ανάλογα με το πώς ένας νευρώνας τροφοδοτεί το επόμενο επίπεδο, τα νευρωνικά δίκτυα διαχωρίζονται σε 2 κύριες κατηγορίες, της πρόσθιας τροφοδότησης (Feed Forward) και ανατροφοδότησης ή αναδρομικά (Feed Backward ή Reccurent). Στην πρόσθια τροφοδότηση οι νευρώνες είναι οργανωμένοι σε διαφορετικά επίπεδα, ώστε οι νευρώνες ενός επιπέδου να τροφοδοτούν τους νευρώνες του επόμενου επιπέδου. Στην περίπτωση αυτή δεν υπάρχει σύνδεση μεταξύ νευρώνων ενός επιπέδου και νευρώνων προηγούμενου επιπέδου. Ουσιαστικά, η πληροφορία ξεκινάει από το επίπεδο εισόδου και καταλήγει στο επίπεδο εξόδου. Σε αντίθεση με τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης, στα αναδρομικά νευρωνικά δίκτυα οι νευρώνες ενός επιπέδου μπορούν να τροφοδοτήσουν νευρώνες ίδιου ή και προηγούμενου

επιπέδου. Αν οι νευρώνες τροφοδοτούν κόμβους στο ίδιο επίπεδο τότε τα δίκτυα ονομάζονται αυτοσυσχετιζόμενες μνήμες (Autoassociated Memories), ενώ αν τροφοδοτούν κόμβους σε προηγούμενο επίπεδο ονομάζονται ετεροσυσχετιζόμενες μνήμες (Heteroassociated Memories).

2.3.3.2 Νευρωνικό δίκτυο πολλαπλών στρώσεων

Ο όρος αισθητήρας (Perceptron) ορίστηκε από τον Rosenbatt το 1962, όπου και χρησιμοποιήθηκε για νευρωνικά δίκτυα πολλών εισόδων και μόνο μίας εξόδου. Υπάρχει μόνο ένα επίπεδο απλών νευρώνων, το οποίο μπορεί να χρησιμοποιηθεί ως είσοδος αλλά και ως έξοδος. Η μάθηση κάθε νευρώνα γίνεται ανεξάρτητα από τους υπόλοιπους νευρώνες, αφού ο κάθε νευρώνας είναι ανεξάρτητος.

Στη περίπτωση που το νευρωνικό δίκτυο είναι πολλαπλών επιπέδων (Multilayer Perceptron), οι συνδέσεις στο ίδιο επίπεδο, καθώς και οι απευθείας συνδέσεις μεταξύ εισόδου και εξόδου, δεν υπάρχουν, ενώ υπάρχει πλήρως σύνδεση μεταξύ επιπέδων. Το πλήθος των εξόδων είναι ανεξάρτητο από το πλήθος εισόδων, όπως και το πλήθος κόμβων είναι ανεξάρτητο ανά επίπεδο. Τέλος, κάθε μονάδα είναι ένας αισθητήρας.

Για την εκπαίδευση ενός τέτοιου νευρωνικού δικτύου πρέπει να γίνει η επιλογή των κρυφών επιπέδων, του πλήθους νευρώνων, της παραγωγίσιμης συνάρτησης ενεργοποίησης, να οριστεί η συνάρτηση λάθους και να γίνει αναζήτηση των συνοπτικών βαρών που ελαχιστοποιούν το σφάλμα χρησιμοποιώντας μεθόδους βελτιστοποίησης. Ο υπολογισμός της εξόδου ενός νευρωνικού δικτύου πολλαπλών επιπέδων καθορίζεται με βάση τα επόμενα πέντε βήματα:

1. Καθορίζονται τυχαία συνοπτικά βάρη.
2. Προωθούνται τα δεδομένα, από το επίπεδο εισόδου προς το επίπεδο εξόδου, ώστε να δημιουργηθεί η έξοδος.
3. Με βάση την έξοδο, υπολογίζεται το σφάλμα για κάθε νευρώνα, το οποίο είναι η διαφορά μεταξύ υπολογιζόμενου και επιθυμητού αποτελέσματος. Στη συνέχεια, γίνεται η αλλαγή των βαρών εισόδου.
4. Αντίθετα με το δεύτερο βήμα, σε αυτό το βήμα ξεκινάμε από το επίπεδο εξόδου με κατεύθυνση το επίπεδο εισόδου, όπου υπολογίζεται για κάθε νευρώνα η συμμετοχή του στα σφάλματα των νευρώνων εξόδου και γίνεται η αλλαγή των βαρών στην είσοδό του.

5. Επανάληψη διαδικασίας μέχρι το συνολικό σφάλμα να σταματήσει να μειώνεται ή μέχρι να ολοκληρωθεί ένας αριθμός επαναλήψεων ή μετά από ένα συγκεκριμένο χρονικό διάστημα.

Το πραγματικό σφάλμα ενός νευρώνα εξόδου k ενός παραδείγματος p υπολογίζεται από τον τύπο:

$$E_k = (a_{k,p} - o_{k,p})$$

Εξίσωση 8: Πραγματικό σφάλμα ενός νευρώνα εξόδου

Το ο συμβολίζει την επιθυμητή έξοδο του νευρώνα. Στη συνέχεια το σφάλμα αυτό πολλαπλασιάζεται με την παράγωγο της συνάρτησης ενεργοποίησης στο νευρώνα k (U_k), με βάση τον κανόνα δέλτα, υπολογίζοντας έτσι το προσαρμοσμένο σφάλμα νευρώνα:

$$\delta_\kappa = (a_{k,p} - o_{k,p})g'(u_k)$$

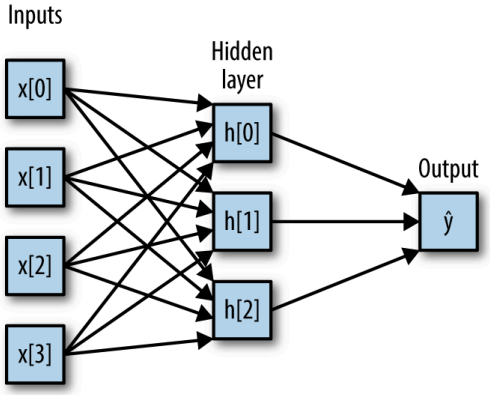
Εξίσωση 9: Προσαρμοσμένο σφάλμα νευρώνα

Τέλος, το αντίστοιχο σφάλμα σε ένα νευρώνα κρυφού επιπέδου i υπολογίζεται από τα προσαρμοσμένα σφάλματα στους κ νευρώνες του επόμενου επιπέδου με τις οποίες ο νευρώνας συνδέεται με βάρη w_{ik}, με τον εξής τύπο:

$$\delta_i = g'(u_i)\sum_{1}^{k} w_{ik}\delta_\kappa$$

Εξίσωση 10: Σφάλμα νευρώνα κρυφού επιπέδου

Ο αλγόριθμος που εφαρμόζεται στο βήμα 5 ονομάζεται αλγόριθμος οπισθοδρόμησης (backpropagation). Τα πλεονεκτήματα που παρέχει είναι η ευκολία του στη χρήση και μπορεί να εφαρμοστεί σε ευρεία περιοχή δεδομένων. Τα μειονεκτήματά του είναι η αργή εκμάθηση, τα νέα στοιχεία θα αντικαταστήσουν τα παλαιά και δεν μπορεί να εγγυηθεί γενίκευση ακόμα και εάν έχει υπολογιστεί το ελάχιστο σφάλμα.



Εικόνα 12: Νευρωνικό δίκτυο πολλαπλών επιπέδων με μόνο ένα κρυφό επίπεδο.

[9]

Η πιο πάνω εικόνα παρουσιάζει ένα νευρωνικό δίκτυο πολλαπλών στρώσεων το οποίο αποτελείται από τρία επίπεδα: επίπεδο εισόδου (Inputs), κρυφό επίπεδο (Hidden layer) και

επίπεδο εξόδου (Output). Κάθε μονάδα μεταξύ των επιπέδων είναι πλήρες συνδεδεμένη με τα άλλα επίπεδα.

2.3.4 Μηχανές διανυσμάτων υποστήριξης

Οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines) δημιουργήθηκαν από τον Hava Siegelmonn και τον Vladimir Vapnik το 1992. Είναι μια μέθοδος διανυσματικής μάθησης η οποία σε σχέση με άλλους ταξινομητές, όπως ο αλγόριθμος Multilayer Perceptron ή ο Naive Bayes που εντοπίζουν έναν οποιοδήποτε γραμμικό διαχωριστή ή αναζητούν τον καλύτερο δυναμικό γραμμικό διαχωριστή με βάση κάποιο κριτήριο αντίστοιχα, έχουν ως στόχο τον εντοπισμό ενός ορίου απόφασης μεταξύ των κλάσεων. Το όριο αυτό πρέπει να βρίσκεται στη μέγιστη δυνατή απόσταση από οποιοδήποτε σημείο των δεδομένων εκπαίδευσης [14]. Έστω η μηχανή εκπαίδευσης f(x,α) όπου το α είναι το σύνολο των παραμέτρων της συνάρτησης βασισμένο σε σημεία τα οποία αποτελούνται από το υποσύνολο των δεδομένων όπου ορίζεται η θέση του διαχωριστή. Τα σημεία αυτά ονομάζονται διανύσματα υποστήριξης.

Στόχος των μηχανών διανυσμάτων υποστήριξης δεν είναι η ελαχιστοποίηση του εμπειρικού κινδύνου, αλλά η ελαχιστοποίηση του ανώτερου ορίου σφάλματος γενίκευσης. Για την επίτευξη αυτού του στόχου χρειάζεται το όριο απόφασης της μηχανικής εκπαίδευσης να έχει τη μέγιστη ελάχιστη απόσταση από το πιο κοντινό σημείο εκπαίδευσης. Το αναμενόμενο σφάλμα ελέγχου ορίζεται από τον τύπο

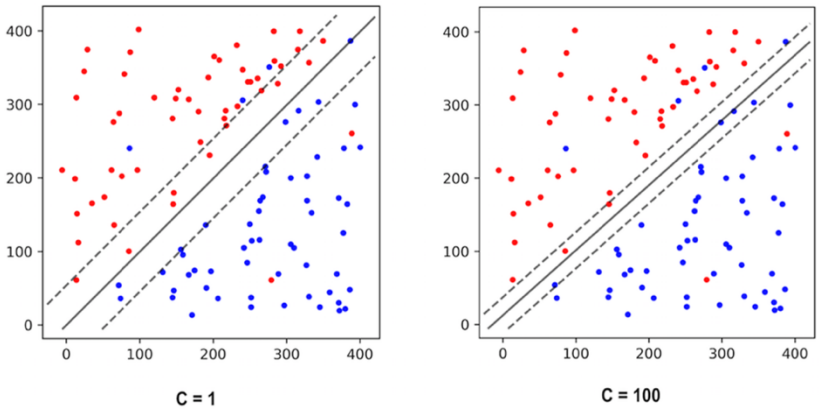
$$R(\alpha) = (\frac{1}{2}\int |y - f(z,\alpha)dP(x,y)|)$$

Εξίσωση 11: Αναμενόμενο σφάλμα ελέγχου

όπου R(α) ονομάζεται αναμενόμενος κίνδυνος ή απλός κίνδυνος. Αντίστοιχα, εμπειρικός κίνδυνος ορίζεται το ποσοστό σφάλματος πάνω στο σύνολο εκπαίδευσης και ορίζεται από τον τύπο:

$$R_{emp}(\alpha) = (\frac{1}{2}\sum_{i=1}^{n}|y_i - f(x_i,\alpha)|)$$

Εξίσωση 12: Εμπειρικός κίνδυνος



Εικόνα 13: Μηχανές διανυσμάτων υποστήριξης.

Η πιο πάνω εικόνα παρουσιάζει το πως επηρεάζεται ο αλγόριθμος SVM αλλάζοντας τη παράμετρο C όπου εξαρτάται το όριο απόφασης, από ένα σε εκατό. Παρατηρείται ότι αν η τιμή του C είναι χαμηλή τότε το όριο απόφασης είναι μεγάλο, ενώ αν το C είναι υψηλό τότε το όριο απόφασης θα είναι μικρό προσπαθώντας έτσι να ελαχιστοποιήσει τις λάθος κατηγοριοποιήσεις.

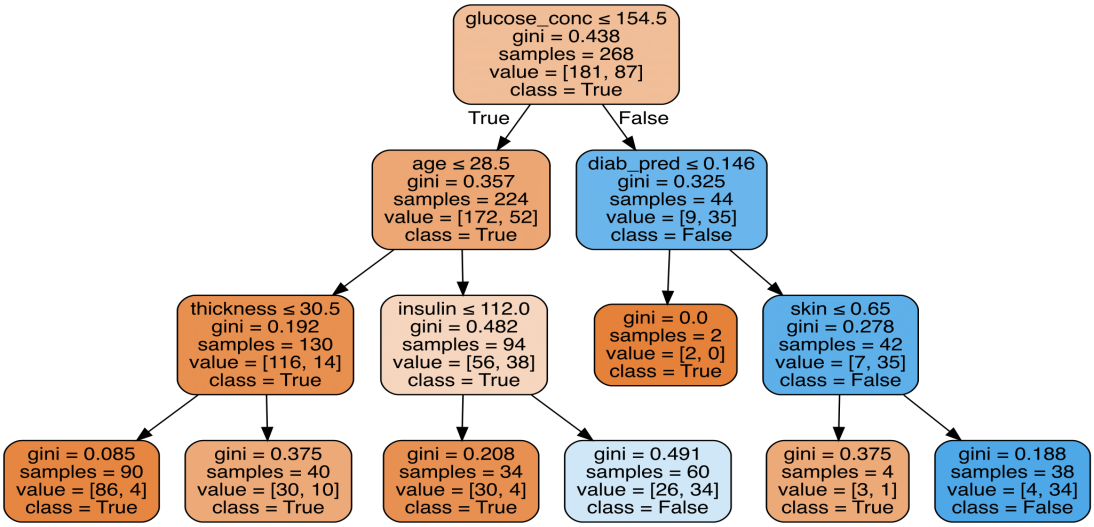
2.3.5 Δένδρα απόφασης

Τα δένδρα απόφασης είναι ο απλούστερος και ευκολότερος τρόπος κατηγοριοποίησης και πρόβλεψης. Είναι πολύ εύκολα ως προς την κατανόηση και την ερμηνεία. Σε αντίθεση με άλλα μοντέλα κατηγοριοποίησης, τα δέντρα απόφασης μπορούν να έχουν συνδυασμό αριθμητικών και κατηγορικών χαρακτηριστικών, αλλά και να κατηγοριοποιήσουν ελλιπή χαρακτηριστικά [17]. Η διαδικασία με την οποία καταλήγουν στην κατηγοριοποίηση γίνεται με τον εξής τρόπο:

- Κάθε εσωτερικός κόμβος ενός δένδρου ονοματίζεται με το όνομα του χαρακτηριστικού.
- Κάθε κλαδί - σύνδεση δύο κόμβων ονοματίζεται με μία συνθήκη ή τιμή για το χαρακτηριστικό του γονικού κόμβου.
- Κάθε φύλλο σχετίζεται με το όνομα μιας κλάσης.

Αρχικά το δένδρο δέχεται ως είσοδο ένα σύνολο εκπαίδευσης με τα διάφορα χαρακτηριστικά που το χαρακτηρίζουν. Τα κλαδιά του δένδρου περιέχουν τις τιμές ελέγχου για κάθε γνώρισμα. Τα φύλλα του δένδρου αντιστοιχούν στις τιμές των κατηγοριών που έχουν οριστεί. Τα χαρακτηριστικά εισόδου μπορούν να είναι διακριτά ή συνεχή, όπως και τα χαρακτηριστικά της τιμής εξόδου. Στην περίπτωση που η τιμή εξόδου είναι διακριτή τιμή τότε έχουμε κατηγοριοποίηση, ενώ όταν η τιμή εξόδου είναι συνεχής συνάρτηση τότε έχουμε

παλινδρόμηση. Το πιο σημαντικό είναι το δέντρο που έχει δημιουργηθεί να μην υπερφορτωθεί. Πιο γνωστοί αλγόριθμοι δένδρων απόφασης είναι ο ID3, ο C4.5, ο CART, ο SLIQ και ο SPRINT.



Εικόνα 14: Δένδρο απόφασης του συνόλου δεδομένων Pima που χρησιμοποιήθηκε.

2.3.5.1 ID3

Ο αλγόριθμος ID3 λειτουργεί με βάση τις έννοιες της εντροπίας και του πληροφοριακού κέρδους, προκειμένου να γίνει η επιλογή του γνωρίσματος που θα πραγματοποιηθεί ο έλεγχος σε κάθε κόμβο. Το κέρδος πληροφορίας υπολογίζεται από τον τύπο Gain(S,A) = E(S) - I(S,A). Το I(S,A) υπολογίζεται από τον τύπο:

$$I(S,A) = \sum \frac{|S_j|}{|S|} * E(S)$$

Εξίσωση 13: Εντροπία διαχωρισμού

Το S_j συμβολίζει το πλήθος των στοιχείων του δείγματος J. Το E(S) συμβολίζει τη συνολική εντροπία και υπολογίζεται από τον τύπο:

$$E(S) = -\sum p_i log(p_i)$$

Εξίσωση 14: Συνολική εντροπία

Το p_i συμβολίζει την πιθανότητα εμφάνισης της κλάσης i στο j. [2][1] Ο αλγόριθμος ID3 λειτουργεί με τα παρακάτω βήματα:

1. Υπολογίζεται το κέρδος πληροφορίας κάθε μεταβλητής.
2. Θέτει ως ρίζα τη μεταβλητή εκείνη με το μεγαλύτερο πληροφοριακό κέρδος.
3. Δημιουργεί τόσα κλαδιά όσες οι διακριτές τιμές της μεταβλητής.
4. Χωρίζει το σύνολο των δεδομένων σε τόσα υποσύνολα όσα και οι διακριτές τιμές της μεταβλητής που επιλέχτηκε.
5. Επιλέγει μια τιμή - υποσύνολο που δεν έχει επιλεγεί. Αν σε αυτή την τιμή αντιστοιχεί μια μόνο κλάση, πάει στο βήμα 6, διαφορετικά στο βήμα 7.

6. Βάζει την τιμή κλάσης ως φύλλο και προχωρεί στην επόμενη τιμή μεταβλητής και πηγαίνει στο βήμα 5.
7. Υπολογίζει το πληροφοριακό κέρδος των υπολοίπων μεταβλητών για το συγκεκριμένο υποσύνολο.
8. Επιλέγει τη μεταβλητή με το μεγαλύτερο πληροφοριακό κέρδος και προσθέτει ένα ακόμα κόμβο στο κλαδί που αντιστοιχεί στην τρέχουσα τιμή.

Ένας δεύτερος τρόπος κατασκευής δένδρου απόφασης είναι το GiniIndex, όπου οι τιμές κυμαίνονται μεταξύ του 0 και του 1, όπου το 0 δηλώνει πλήρη ισότητα και το 1 πλήρη ανισότητα. Υπολογίζεται από τον τύπο:

$$gini(S) = 1 - \sum_{j=1}^{k} p_j^2$$

Εξίσωση 15: GiniIndex

Από τον πιο πάνω τύπο γνωρίζουμε ότι το p_j είναι η πιθανότητα εμφάνισης της κλάσης j στο σύνολο δεδομένων S. Αν το S διαχωριστεί σε S1 και S2, τότε η σχέση μας γίνεται:

$$gini(S) = \frac{n_1}{n} g(S1) + \frac{n_2}{n} g(S2)$$

Εξίσωση 16: Διαχωρισμός S στο GiniIndex

Στη περίπτωση αυτή το n_1 είναι το σύνολο των δειγμάτων στο S1 και το n_2 στο S2. Το πλεονέκτημα αυτής της μεθόδου σε σχέση με την προηγούμενη μέθοδο είναι ότι για τον υπολογισμό απαιτείται μόνο ο διαχωρισμός των κλάσεων σε ένα υποσύνολο, όπου καλύτερο χαρακτηριστικό είναι αυτό με τη μικρότερη τιμή Gini.

2.3.5.2 C4.5

Ο αλγόριθμος C4.5 βασίζεται και αυτός όπως και ο ID3 στο κέντρο πληροφορίας και στο δείκτη Gini. Τα γνωρίσματα κάθε κόμβου του δένδρου C4.5 μπορούν να έχουν συνεχείς τιμές. Για τη σωστή λειτουργία χρειάζονται ολοκληρωμένα δεδομένα. Σε περίπτωση που το σύνολο δεδομένων είναι αρκετά μεγάλο, τότε ο αλγόριθμος C4.5 κρίνεται ακατάλληλος, αφού η ακρίβεια που παρουσιάζει είναι πολύ μικρή [13].

2.3.6 Συλλογική μάθηση

Με τον όρο συλλογική μάθηση (Ensemble Learning) εννοούμε το συνδυασμό πολλαπλών μοντέλων μηχανικής μάθησης με σκοπό τη δημιουργία ισχυρότερων μοντέλων. Σε ένα φάσμα αλγορίθμων που ανήκουν σε αυτή την κατηγορία, οι σημαντικότεροι αλγόριθμοι που έχουν αποδειχθεί αποτελεσματικότεροι σε σχέση με τους άλλους για κατηγοριοποίηση και

παλινδρόμηση είναι τα τυχαία δάση (Random Forests) και τα ενισχυόμενα δέντρα απόφασης (Boosted Recision Trees).

Οι αλγόριθμοί που χρησιμοποιούν τη μέθοδο της ενδυνάμωσης (boosting) είναι βασισμένοι στη λειτουργία του σταθμισμένου συνόλου εκπαίδευσης. Το σταθμισμένο σύνολο εκπαίδευσης που χρησιμοποιείται έχει ένα βάρος $w \geq 0$. Το βάρος αυτό αντικατοπτρίζει την πιθανότητα το συγκεκριμένο χαρακτηριστικό να επιλεγεί στο σύνολο εκπαίδευσης. Όσο μεγαλύτερο είναι το βάρος, τόσο μεγαλύτερη είναι η σημασία του προς την μάθηση μιας υπόθεσης.

Τα βήματα που ακολουθούνται κατά τη διαδικασία της ενδυνάμωσης είναι:

1. Το αρχικό βάρος για όλα τα χαρακτηριστικά είναι 1.
2. Δημιουργία πρώτης υπόθεσης όπου θα γίνει η κατηγοριοποίηση του χαρακτηριστικού σωστά ή εσφαλμένα.
3. Αύξηση βαρύτητας των χαρακτηριστικών που κατηγοριοποιήθηκαν λάθος και μείωση βαρύτητας σωστά κατηγοριοποιημένων χαρακτηριστικών. Σκοπός του βήματος αυτού είναι η επόμενη υπόθεση να έχει καλύτερο αποτέλεσμα σε σχέση με την πρώτη.
4. Γίνεται επανάληψη της πιο πάνω διεργασίας μέχρι τη δημιουργία Κ υποθέσεων, με την κάθε υπόθεση να αντιστοιχεί στο πόσο καλά απέδωσε.

2.3.6.1 AdaBoost

Ο αλγόριθμος AdaBoost είναι ένας από τους αλγορίθμους που χρησιμοποιεί τη μέθοδο της ενδυνάμωσης. Σημαντική ιδιότητα που έχει ο αλγόριθμος αυτός είναι ότι αν ο αλγόριθμος που δέχεται σαν είσοδο είναι ασθενούς μάθησης (Weak Learning), τότε ο αλγόριθμος θα κατηγοριοποιήσει πολύ καλά έως τέλεια τα δεδομένα εκπαίδευσης, για ένα αρκετά μεγάλο σύνολο υποθέσεων Κ. Συνεπώς, ο αλγόριθμος AdaBoost ενδυναμώνει την ακρίβεια του αλγορίθμου. Δεν επηρεάζεται από το πόσο μη-εκφραστικός είναι ο χώρος υποθέσεων και πόσο πολύπλοκη είναι η συνάρτηση. Με λίγα λόγια ο αλγόριθμος AdaBoost χρησιμεύει στην κατασκευή ενός ισχυρού κατηγοριοποιητή βασισμένου σε ένα αδύνατο κατηγοριοποιητή. Υπολογίζεται από τον τύπο

$$f(x) = \sum_{k=1}^{K} a_k h_k(x)$$

Εξίσωση 17: Τύπος υπολογισμού AdaBoost

Μερικά από τα πλεονεκτήματα τα οποία παρέχει ο αλγόριθμος AdaBoost είναι ότι είναι εύκολος, απλός, γρήγορος και η μοναδική παράμετρος που προσαρμόζεται είναι το πλήθος των αδύναμων κατηγοριοποιητών, έτσι καθορίζεται και το πλήθος των επαναλήψεων. Επίσης, δεν απαιτείται από πριν η γνώση του αδύναμου κατηγοριοποιητή, άρα μπορεί να εφαρμοστεί με οποιαδήποτε μέθοδο κατηγοριοποίησης. Τέλος, οι τιμές οι οποίες έχουν μεγαλύτερο βάρος είναι συνήθως ακραίες τιμές, αφού ο αλγόριθμος δίνει σημασία σε τιμές χαρακτηριστικών που κατηγοριοποιούνται δυσκολότερα, εντοπίζοντας έτσι τις ακραίες τιμές. Το σημαντικότερο μειονέκτημα που έχει ο AdaBoost είναι ότι στρέφεται κυρίως σε δεδομένα τα οποία κατηγοριοποιούνται λάθος με αποτέλεσμα να μην είναι ακριβής σε δεδομένα τα οποία περιέχουν θόρυβο.

2.3.6.2 Τυχαία δάση

Τα τυχαία δάση (Random Forests) είναι ένας αλγόριθμος επέκτασης των δέντρων αποφάσεων ο οποίος χρησιμοποιείται και αυτός για κατηγοριοποίηση ή παλινδρόμηση, αποτελούμενος από μια συλλογή αρκετών δέντρων απόφασης. Σκοπός του αλγορίθμου είναι η δημιουργία ενός δάσους από ένα τυχαίο αριθμό δέντρων αποφάσεων, έτσι προκύπτει και το όνομα του αλγορίθμου. Κάθε δέντρο το οποίο υπάρχει στο τυχαίο δάσος είναι ελαφρώς διαφορετικό από το άλλο. Η κύρια ιδέα πίσω από τη δημιουργία του τυχαίου δάσους είναι ότι ένα δέντρο απόφασης μπορεί να παράγει σχετικά καλά αποτελέσματα όσο αφορά την πρόβλεψη, αλλά έτσι δημιουργείται η πιθανότητα να έχουμε υπερφόρτωση (overfitting) των δεδομένων μας. Στην περίπτωση που δημιουργηθούν αρκετά δέντρα απόφασης, τα οποία λειτουργούν καλά, αλλά παρατηρείται το φαινόμενο της υπερφόρτωσης, το μέγεθος της υπερφόρτωσης αυτής μπορεί να μειωθεί υπολογίζοντας το μέσο όρο των αποτελεσμάτων τους. Για την υλοποίηση της πιο πάνω μεθόδου, θα πρέπει να δημιουργηθούν αρκετά δέντρα αποφάσεων. Το κάθε δέντρο θα είναι διαφορετικό από το άλλο και θα υπολογίζει ένα δικό του αποτέλεσμα. Με πιο απλά λόγια ένα τυχαίο δάσος δημιουργεί πολλαπλά δέντρα απόφασης και τα ενώνει μεταξύ τους με σκοπό να έχει πιο ακριβή και σταθερή πρόβλεψη.

Οι κύριοι μέθοδοι που εφαρμόζονται σε ένα τυχαίο δάσος είναι η επιλογή των σημείων δεδομένων που χρησιμοποιούνται για την κατασκευή ενός δέντρου και η επιλογή των χαρακτηριστικών από κάθε διαιρούμενο σύνολο [9].

Για τη δημιουργία ενός τυχαίου δάσους ακολουθούνται τα εξής βήματα:

- Εισαγωγή δεδομένων εκπαίδευσης.

- Δημιουργία πολλαπλών συνόλων εκπαίδευσης.
- Κατασκευή δέντρων απόφασης με τυχαία επιλογή χαρακτηριστικών σε κάθε κόμβο.
- Κατηγοριοποίηση των δεδομένων.

Η σχέση που υπάρχει μεταξύ του αριθμού των δέντρων που υπάρχουν στο δάσος και στα αποτελέσματα είναι ότι, όσο μεγαλύτερος είναι ο αριθμός των δέντρων, τόσο πιο ακριβές είναι το αποτέλεσμα. Χρησιμοποιεί ένα γράφημα τύπου δέντρου για να δείξει τις πιθανές συνέπειες. Εάν γίνει εισαγωγή ενός συνόλου δεδομένων κατάρτισης με στόχους και χαρακτηριστικά στο δέντρο αποφάσεων, τότε θα δημιουργήσει ένα σύνολο κανόνων. Οι κανόνες αυτοί μπορούν να χρησιμοποιηθούν για την εκτέλεση προβλέψεων. Η διαδικασία υπολογισμού των διαφόρων κόμβων και της διαδικασίας διαμόρφωσης κανόνων χρησιμοποιεί τη μέθοδο κέρδος πληροφορίας και στο δείκτη Gini.

Υπάρχουν δύο κύριες διαφορές του αλγορίθμου Random Forest και του C4.5. Πρώτη διαφορά είναι ότι οι διαδικασίες της εύρεσης του κόμβου ρίζας και του διαχωρισμού των κόμβων χαρακτηριστικών γίνονται τυχαία. Δεύτερη διαφορά είναι ότι στα δέντρα απόφασης παρατηρείται το φαινόμενο της υπερφόρτωσης. Στο τυχαίο δάσος, η υπερφόρτωση αποτρέπεται δημιουργώντας τυχαία υποσύνολα των χαρακτηριστικών και δημιουργώντας μικρότερα δέντρα χρησιμοποιώντας αυτά τα υποσύνολα. Αυτό όμως δεν λειτουργεί κάθε φορά. Η πιο πάνω διάσπαση του συνόλου των χαρακτηριστικών σε μικρότερα υποσύνολα, καθώς και η αναλογία των δέντρων που δημιουργούνται στο δάσος, καθιστούν τον υπολογισμό πιο αργό.

2.3.6.3 GradientBoost

Ο αλγόριθμος GradientBoost χρησιμοποιείται για προβλήματα παλινδρόμησης και κατηγοριοποίησης, συνδυάζοντας πολλαπλά δέντρα αποφάσεων. Σε αντίθεση με τα τυχαία δάση, το κάθε δέντρο στη μέθοδο GradientBoost προσπαθεί να διορθώσει τα λάθη του προηγούμενο. Για την επίτευξη του πιο πάνω ο αλγόριθμος κατασκευάζει διάφορες δομές με ένα σειριακό τρόπο. Το βάθος των δέντρων είναι πολύ μικρότερο σε σχέση με τα τυχαία δάση, από ένα έως πέντε, κάνοντας έτσι το μοντέλο μικρότερο και βοηθώντας το έτσι στο να κάνει γρηγορότερες προβλέψεις [9].

2.3.7 Κ κοντινότεροι γείτονες

Τελευταίος αλγόριθμος επιβλεπόμενης μάθησης είναι ο αλγόριθμος k κοντινότεροι γείτονες (k Nearest Neighbors – kNN), ο οποίος παρουσιάζει αρκετό ενδιαφέρον στον τρόπο που ακολουθεί για την κατηγοριοποίηση των δεδομένων, αφού λειτουργεί διαφορετικά από τους προαναφερθέντες αλγορίθμους. Ονομάζεται επίσης και τεμπέλης αλγόριθμος (lazy algorithm) λόγω του ότι απομνημονεύει όλο το σύνολο εκπαίδευσης στη μνήμη. Το πλεονέκτημα αυτής της προσέγγισης είναι ότι ο κατηγοριοποιητής κατηγοριοποιεί αμέσως τα νέα δεδομένα εκπαίδευσης που συλλέγονται. Σημαντικό μειονέκτημα είναι ότι η εισαγωγή νέων δεδομένων για κατηγοριοποίηση αυξάνει γραμμικά την υπολογιστική πολυπλοκότητα [3].

Η βασική ιδέα του αλγορίθμου kNN ως προς την κατηγοριοποίηση ενός στοιχείου είναι οι ιδιότητες κάθε συγκεκριμένου στοιχείου που δίνεται σαν είσοδο στον αλγόριθμο να είναι παρόμοιες με τις ιδιότητες που έχουν τα άλλα σημεία, σε μια συγκεκριμένη απόσταση από αυτό. Η απόσταση αυτή ονομάζεται και "γειτονιά", από όπου προκύπτει και το όνομα του αλγορίθμου. Τα βήματα τα οποία ακολουθούνται για την υλοποίηση του αλγορίθμου kNN είναι τα εξής:

1. Επιλογή των αριθμών των γειτόνων και το μέτρο απόστασης.
2. Εύρεση του γείτονα αυτού που πρέπει να κατηγοριοποιηθεί το στοιχείο.
3. Κατηγοριοποίηση του νέου στοιχείου.

Προκειμένου να μετρήσουμε την ομοιότητα ή την απόσταση μεταξύ σημείων θα πρέπει να χρησιμοποιηθεί κάποιο μέτρο απόστασης $D(x_1, x_2)$. Τέτοια μέτρα απόστασης μπορεί να είναι η Ευκλείδεια απόσταση, η απόσταση Manhattan, η απόσταση Minkowski, η απόσταση Chebyshev και η απόσταση Hamming, οι οποίες δίνονται από τους εξής τύπους:

- Ευκλείδεια απόσταση: όταν το r = 2, τότε η απόσταση ονομάζεται ευκλείδεια απόσταση (Euclidean distance) και υπολογίζεται από τον τύπο:

$$d(x, y) = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}$$

Εξίσωση 18: Ευκλείδεια απόσταση

- Απόσταση Manhattan: όταν το r = 1 τότε η απόσταση ονομάζεται απόσταση Manhattan και υπολογίζεται από τον τύπο:

$$d(x,y) = \sum_{j=1}^{n} |x_j - y_j|$$

Εξίσωση 19: Απόσταση Manhattan

- Απόσταση Minkowski: είναι μια γενίκευση της ευκλείδειας απόστασης και της απόστασης Manhattan και υπολογίζεται από τον τύπο:

$$d(x,y) = \sqrt[p]{\sum_{j=1}^{n} |x_j - y_j|^p}$$

Εξίσωση 20: Απόσταση Minkowski

- Απόσταση Chebyshev: όταν το r τείνει στο άπειρο τότε η απόσταση ονομάζεται απόσταση Chebyshev και υπολογίζεται από τον τύπο:

$$d(x,y) = \max_{j=1,\dots,n} |x_j - y_j|$$

Εξίσωση 21: Απόσταση Chebyshev

- Απόσταση Hamming: τελευταία απόσταση είναι η απόσταση Hamming, η οποία χρησιμεύει για υπολογισμό διακριτών χαρακτηριστικών όταν το r = 0. Ο τύπος υπολογισμού της είναι:

$$d(x,y) = \sum_{j=1}^{n} 1_{x_j} \neq y_j$$

Εξίσωση 22: Απόσταση Hamming

Για τον προσδιορισμό του k πλησιέστερου γείτονα κάποιου σημείου του συνόλου δεδομένων, πρέπει να υπολογιστεί ένα από τα πιο πάνω μέτρα απόστασης.

2.3.8 Προεπεξεργασία δεδομένων

Τα δεδομένα τα οποία έχουμε στη διάθεσή μας για ανάλυση αρκετές φορές δεν είναι πλήρη. Συγκεκριμένα, αρκετά δεδομένα είναι ελλιπή, οι τιμές τους δεν σχετίζονται με τις άλλες τιμές του χαρακτηριστικού ή οι τιμές δεν συλλέγονται σωστά τη στιγμή που έχουν καταγραφεί τα δεδομένα. Οι τιμές που λείπουν από ένα σύνολο δεδομένων μπορούν να επηρεάσουν την ακρίβεια και την απόδοση του κατηγοριοποιητή, γεγονός που δυσχεραίνει την εξαγωγή σημαντικών πληροφοριών από το σύνολο δεδομένων, λόγω της απώλειας της αποτελεσματικότητας που υπάρχει [6]. Υπάρχουν 3 είδη κατηγοριών όπου κατηγοριοποιούνται οι τιμές αυτές:

- Ημιτελή: παρουσιάζουν ελλιπή στοιχεία ή έλλειψη τιμών οι οποίες σχετίζονται άμεσα με το χαρακτηριστικό.
- Θόρυβος: περιέχουν λάθη ή οι τιμές είναι πολύ μεγάλες σε σχέση με τα άλλα δεδομένα.

- Ασυνεπή: περιέχουν αποκλίσεις στις κωδικοποιήσεις που χρησιμοποιούνται για τη κατηγοριοποίηση των δεδομένων [13].

Υπάρχουν δύο κύριες τεχνικές προεπεξεργασίας οι οποίες προσπαθούν να βελτιώσουν την ποιότητα των δεδομένων. Η πρώτη τεχνική προσπαθεί να μειώσει τις τιμές των χαρακτηριστικών σε ένα εύρος τιμών μεταξύ του 0 και του 1, ή να δημιουργήσει μια κανονική κατανομή με μηδενικό μέσο και σταθερή τυπική απόκλιση. Πολλές φορές αρκετά από τα χαρακτηριστικά έχουν μεγάλη συσχέτιση μεταξύ τους ενώ κάποια όχι. Η δεύτερη τεχνική είναι η μείωση των διαστάσεων αφού χρησιμοποιείται προκειμένου να μειωθούν τα χαρακτηριστικά του συνόλου δεδομένων, επιλέγοντας τα σημαντικότερα, δηλαδή αυτά που έχουν μεγαλύτερη συσχέτιση μεταξύ τους [12][13]. Αυτό έχει ως πλεονέκτημα ο χώρος αποθήκευσης που χρειάζεται να μειώνεται και ο αλγόριθμος μπορεί να τρέξει πολύ πιο γρήγορα [12].

2.3.8.1 Δεδομένα ίδιας κλίμακας

Μία από τις κυριότερες μεθόδους προεπεξεργασίας των δεδομένων είναι η τροποποίηση των τιμών των δεδομένων στο ίδιο μέγεθος κλίμακας. Υπάρχουν δύο κύριες προσεγγίσεις που έχουν τη δυνατότητα αυτή, η κανονικοποίηση (normalization) και η τυποποίηση (standardization).

Η κανονικοποίηση είναι χρήσιμη όταν χρειαζόμαστε τιμές σε ένα ορισμένο διάστημα αφού χρησιμεύει στην αναδιάταξη των δεδομένων μας σε ένα εύρος τιμών μεταξύ 0 και 1. Για την εύρεση αυτού του εύρους εφαρμόζεται η μέθοδος min-max σε κάθε στήλη, η οποία υπολογίζεται από τον τύπο:

$$X_{norm}^{i} = \frac{x^{i} - x_{min}}{x_{max} - x_{min}}$$

Εξίσωση 23: Τύπος κανονικοποίησης

Από τον πιο πάνω τύπο γνωρίζουμε ότι το x^i είναι η τιμή του συγκεκριμένου δείγματος, x_{min} είναι η ελάχιστη τιμή του δείγματος και x_{max} η μέγιστη τιμή του δείγματος.

Η μέθοδος της τυποποίησης είναι αρκετές φορές καλύτερη και πιο πρακτική σε πολλούς αλγορίθμους μηχανικής μάθησης, κυρίως όταν χρησιμοποιούνται γραμμικά μοντέλα, όπως λογιστική παλινδρόμηση και μηχανές υποστήριξης διανυσμάτων. Αυτό οφείλεται στο γεγονός ότι οι αλγόριθμοι αυτοί αρχικοποιούν τα βάρη σε τιμές ίσες ή κοντά στο 0. Η τυποποίηση κρατάει το μέσο όρο των στηλών στο 0 και την τυπική απόκλιση στο 1, έχοντας

τη μορφή της κανονικής κατανομής, βοηθώντας έτσι στην καλύτερη εκμάθηση των βαρών. Υπολογίζεται από τον τύπο:

$$X_{std}^{i} = \frac{x^{i} - \mu_x}{\sigma_x}$$

Εξίσωση 24: Τύπος τυποποίησης

Το x^i είναι η τιμή του συγκεκριμένου δείγματος, το μ_x ο δειγματικός μέσος και σ_x η τυπική απόκλιση [12].

2.3.8.2 Μείωση διαστάσεων

2.3.8.2.1 Ανάλυση κυρίων συνιστωσών

Η μέθοδος ανάλυσης κυρίων συνιστωσών (Principal Components Analysis - PCA) είναι μια τεχνική γραμμικού μετασχηματισμού, η οποία χρησιμοποιείται για μείωση διαστάσεων του συνόλου δεδομένων κατά τη φάση της προεπεξεργασίας, με σκοπό ο αλγόριθμος κατηγοριοποίησης να γίνει πιο αποτελεσματικός [3]. Είναι επίσης μια δημοφιλής μέθοδος για την εξαγωγή των σημαντικών χαρακτηριστικών από τα δεδομένα εκπαίδευσης που χρησιμοποιούνται για την εκμάθηση ενός μοντέλου μηχανικής μάθησης. Αν θεωρήσουμε ότι ένα σύνολο δεδομένων έχει σύνολο αριθμό γραμμών N και αριθμό στηλών M, τότε με τη μέθοδο PCA βρίσκουμε ένα σύστημα K κάθετων διανυσμάτων, το οποίο είναι μικρότερο από το συνολικό αριθμό των στηλών του αρχικού συνόλου δεδομένων. Στη συνέχεια προβάλλουμε τα δεδομένα στο νέο μας σύστημα K, δημιουργώντας έτσι γραμμικούς συνδυασμούς των αρχικών μεταβλητών, οι οποίοι είναι ασυσχέτιστοι μεταξύ τους και περιέχουν το μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών.

Διαδικασία PCA:

- Για κάθε στήλη υπολογίζεται η μέση τιμή και αφαιρείται από όλες τις τιμές της. Έτσι η μέση τιμή κάθε στήλης γίνεται μηδέν.
- Υπολογίζεται ο πίνακας συνδιασποράς για τις νέες τιμές του πίνακα.
- Υπολογίζονται τα ιδιοδιανύσματα του πίνακα συνδιασποράς, τα οποία έχουν μήκος ίσο με ένα. Τα ιδιοδιανύσματα ονομάζονται και κύριες συνιστώσες και τα αρχικά δεδομένα μπορούν να εκφραστούν ως γραμμικοί συνδυασμοί των κυρίων συνιστωσών.
- Τα ιδιοδιανύσματα ταξινομούνται με βάση τις ιδιοτιμές τους, οι οποίες αποτελούν μέτρο σημαντικότητας των ιδιοδιανυσμάτων, αφού όσο μεγαλύτερη είναι η ιδιοτιμή, τόσο σημαντικότερο είναι το ιδιοδιάνυσμα, δηλαδή η συνιστώσα αυτή περιλαμβάνει περισσότερη πληροφορία σχετικά με τη διασπορά των δεδομένων.

- Ταξινόμηση των βασικών συνιστωσών σε φθίνουσα σειρά και αφαίρεση λιγότερων σημαντικών.
- Τα δεδομένα επανακαθορίζονται με βάση το νέο σύστημα αξόνων. Στην περίπτωση που διατηρούμε όλες τις βασικές συνιστώσες, τότε γίνεται μια ανακατασκευή των δεδομένων με τα αρχικά δεδομένα, προβάλλοντάς τα στο αρχικό σύστημα αξόνων. Στην περίπτωση που επιλεγούν οι σημαντικότερες συνιστώσες, τότε κατασκευάζεται μια ικανοποιητική προσέγγιση των αρχικών δεδομένων στο νέο σύστημα αξόνων.

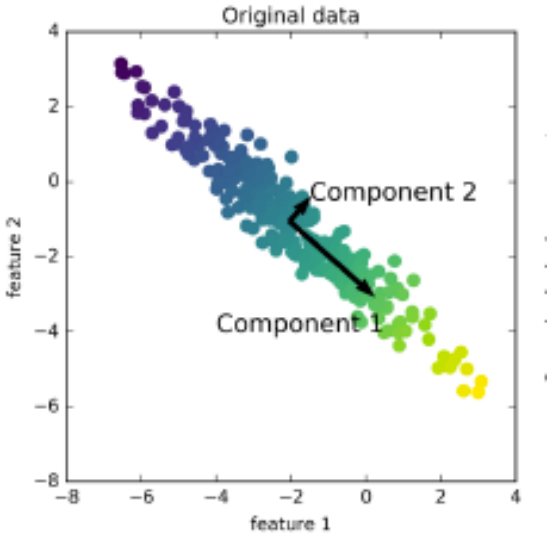
Η κάθε κύρια συνιστώσα ορίζεται ως ένας γραμμικός συνδυασμός των αρχικών μεταβλητών και έχει τη μορφή:

$$PC_i = \alpha_{i1}x_1 + \alpha_{i2}x_2 + \ldots + \alpha_{in}x_n$$

Εξίσωση 25: Κύρια συνιστώσα

Ο συντελεστής α_{ik} καθορίζει τον βαθμό στον οποίο κάθε μεταβλητή επηρεάζει την εκάστοτε βασική συνιστώσα.

Σε μια πιο γενική μορφή, η μέθοδος PCA είναι μια τεχνική γραμμικού μετασχηματισμού και αποτελεί μια μέθοδο μείωσης των διαστάσεων των δεδομένων και την προβολή αυτών σε ένα χώρο λιγότερων αλλά διαφορετικών διαστάσεων και όχι μια μέθοδο επιλογής σημαντικών χαρακτηριστικών.



Εικόνα 15: Μείωση διαστάσεων με τη μέθοδο ανάλυσης κυρίων συνιστωσών.

[9]

Η πιο πάνω εικόνα παρουσιάζει τα διάφορα σημεία με διαφορετικό χρώμα με σκοπό την διαχώριση τους. Αρχικά ο αλγόριθμος βρίσκει την κατεύθυνση της μέγιστης διακύμανσης με την

ένδειξη Component 1. Στη συνέχεια ο αλγόριθμος βρίσκει την κατεύθυνση εκείνη που είναι ορθογώνια ως προς την πρώτη κατεύθυνση και περιέχει τις περισσότερες πληροφορίες (Component 2).

2.3.8.2.2 Γραμμική διαχωριστική ανάλυση

Ο Ronald Fisher το 1963, αναφέρει ότι *"Ο μηχανισμός επεξεργασίας που οικοδομήθηκε σε εφαρμογές απείρων δεδομένων δεν είναι αρκετά ακριβής για απλά εργαστηριακά δεδομένα. Μόνο με συστηματική επιλογή προβλημάτων με λίγα δείγματα, ανάλογα με τα ιδιαίτερα χαρακτηριστικά τους, μπορούμε να έχουμε ακριβή τεστ σε πρακτικά δεδομένα"*. Η μέθοδος της γραμμικής διαχωριστικής ανάλυσης (Linear Discriminant Analysis - LDA), όπως και η μέθοδος PCA, είναι μια τεχνική γραμμικού μετασχηματισμού που χρησιμοποιείται για την εξαγωγή χαρακτηριστικών από ένα σύνολο, έτσι ώστε να αυξηθεί η υπολογιστική αποδοτικότητα και να μειωθεί η υπερφόρτωση των δεδομένων. Είναι μια μέθοδος επίβλεψης, σε αντίθεση με τη μέθοδο PCA, η οποία είναι μέθοδος μη επιβλεπόμενη [12]. Σκοπός του αλγορίθμου είναι ο διαχωρισμός δειγμάτων σε ομάδες μεγιστοποιώντας τη διαχωρισιμότητα μεταξύ των κλάσεων και τη μεταβλητότητα εντός της κλάσης, μειώνοντας έτσι τις διαστάσεις, διατηρώντας όμως τις κλάσεις όσο το δυνατόν πιο διακριτές. Γενική ιδέα του αλγορίθμου είναι η εύρεση του κατάλληλου υποχώρου με χαρακτηριστικά που θα βελτιστοποιεί τη διαχώριση των κλάσεων.

Η διαχωρισιμότητα μεταξύ των κλάσεων υπολογίζεται από τον τύπο:

$$S_b = \sum_{j=1}^{c} (\mu_j - \mu)(\mu_j - \mu)^T$$

Εξίσωση 26: Διαχωρισιμότητα μεταξύ κλάσεων.

Το μ_j συμβολίζει το μέσο της κάθε κλάσης. Για τον υπολογισμό της μεταβλητότητας εντός της κλάσης χρησιμοποιείται ο τύπος :

$$S_w = \sum_{j=1}^{c} \sum_{i=1}^{N_j} (x_i^j - \mu_j)(x_i^j - \mu_j)^T$$

Εξίσωση 27: Μεταβλητότητα εντός της κλάσης.

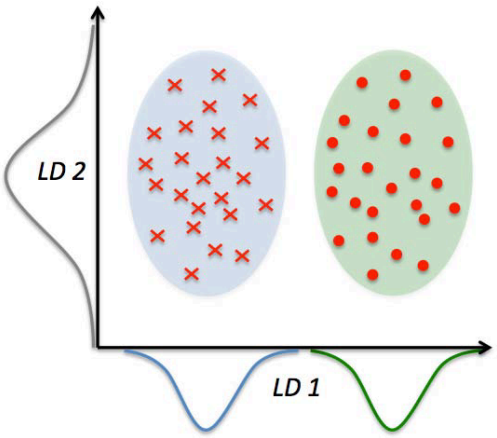
Από τον τύπο της μεταβλητότητας, το x_i^j συμβολίζει το i-οστό δείγμα της κλάσης j, το μ_j είναι ο μέσος της κλάσης j, ο c είναι ο αριθμός των κλάσεων και N_j ο αριθμός των δειγμάτων μέσα στην κλάση. Για την επίτευξη του πιο πάνω σκοπού, πρέπει να υπολογιστεί ο πίνακας της ορίζουσας που μεγιστοποιεί το λόγο της ορίζουσας S_b ως προς την ορίζουσα S_w

$$Plda = \frac{det|S_b|}{det|S_w|}$$

Εξίσωση 28: Πίνακας ορίζουσας.

Το πλεονέκτημα χρησιμοποίησης του λόγου αυτού είναι ότι αποδεικνύεται πώς εάν ο S_w είναι πίνακας με διακρίνουσα, τότε ο λόγος που μεγιστοποιείται ο πίνακας Plda συνθέτεται από τα μέγιστα ιδιοδιανύσματα του πίνακα $S_w^{-1}S_b$ [18].

Οι αλγόριθμοι LDA (Linear Discriminant Analysis) και PCA αποτελούν τους δύο πιο κύριους αλγορίθμους προεπεξεργασίας δεδομένων. Και οι δύο αλγόριθμοι χρησιμοποιούνται για τη μείωση διαστάσεων του συνόλου δεδομένων, όμως ο τρόπος με τον οποίο λειτουργούν είναι διαφορετικός. Οι δύο αυτοί αλγόριθμοι χρησιμοποιούνται βέλτιστα όταν τα χαρακτηριστικά που περιέχει ένα σύνολο δεδομένων είναι πάρα πολλά, με αρκετά από αυτά να μην χρειάζονται για τη σωστή κατηγοριοποίηση. Σε αντίθετη περίπτωση όπου τα χαρακτηριστικά είναι λίγα, οι δύο αυτοί αλγόριθμοι ενδέχεται να μην βοηθήσουν στην κατηγοριοποίηση των δεδομένων.



Εικόνα 16: Μείωση διαστάσεων με χρήση της μεθόδου γραμμικής διαχωριστικής ανάλυσης.

[12]

Από τη πιο πάνω εικόνα παρατηρούμε ότι η διαχώριση των δύο κανονικά κατανεμημένων κατηγοριών είναι καλύτερη με βάση τον άξονα x (LD1).

Κεφάλαιο 3: Μεθοδολογία

Σε αυτό το κεφάλαιο παρουσιάζεται η μεθοδολογία που ακολουθήθηκε για την υλοποίηση της μεταπτυχιακής διπλωματικής εργασίας. Αρχικά γίνεται αναφορά στο περιβάλλον υλοποίησης, τη γλώσσα προγραμματισμού και στις βιβλιοθήκες που χρησιμοποιήθηκαν για την υλοποίηση της μεταπτυχιακής διπλωματικής εργασίας. Ακολούθως παρουσιάζονται τα σύνολα δεδομένων τα οποία χρησιμοποιήθηκαν καθώς και διάφορα χαρακτηριστικά τους. Στη συνέχεια, παρουσιάζονται οι αλγόριθμοι που χρησιμοποιήθηκαν μαζί με τις παραμέτρους τους και παρουσιάζονται τα βήματα υλοποίησης.

3.1 Περιβάλλον υλοποίησης

Για την υλοποίηση των αλγορίθμων της μεταπτυχιακής διπλωματικής εργασίας χρησιμοποιήθηκαν οι βιβλιοθήκες Pandas, Numpy, Matplotlib, Scipy και Scikit-Learn. Οι βιβλιοθήκες αυτές είναι οι βασικές βιβλιοθήκες της γλώσσας προγραμματισμού Python στον τομέα της μηχανικής μάθησης. Το περιβάλλον προγραμματισμού το οποίο χρησιμοποιήθηκε για την υλοποίηση είναι το JetBrain Pycharm Community Edition 2018.2, το οποίο βρίσκεται διαθέσιμο δωρεάν στην επίσημη ιστοσελίδα της Jetbrains και πιο συγκεκριμένα στο σύνδεσμο https://www.jetbrains.com/pycharm/. Οι βιβλιοθήκες που χρησιμοποιήθηκαν είναι οι ακόλουθες:

- Pandas: Η βιβλιοθήκη Pandas είναι μια ανοιχτή βιβλιοθήκη υψηλής απόδοσης, η οποία είναι βασισμένη στη βιβλιοθήκη Numpy [12], βοηθάει την Python στην ανάλυση των δεδομένων και στη μοντελοποίησή τους. Είναι βασισμένη πάνω σε μια δομή δεδομένων που ονομάζεται DataFrame και η οποία μοιάζει με πίνακα. Έτσι μπορούν να εκτελεστούν διάφορα SQL ερωτήματα καθώς και ενώσεις πάνω στο DataFrame.
- Numpy: Όπως και η βιβλιοθήκη Pandas, έτσι και η Numpy είναι ανοιχτή βιβλιοθήκη για την Python. Περιέχει διάφορες έτοιμες μεθόδους για πολυδιάστατους πίνακες και μαθηματικές συναρτήσεις. Επιπρόσθετα, έχει τη δυνατότητα να παράγει ψευδοτυχαίους αριθμούς [9]. Χρησιμοποιεί επίσης αυθαίρετους τύπους δεδομένων επιτρέποντας να ενσωματώνεται χωρίς προβλήματα και με ταχύτητα σε αρκετές βάσεις δεδομένων.
- Matplotlib: Η βιβλιοθήκη Matplotlib είναι η βασική βιβλιοθήκη για σχεδιασμό στην Python. Μέσω αυτής της βιβλιοθήκης δίνεται η δυνατότητα να κατασκευαστούν γραφήματα, ιστογράμματα, γραφήματα διασποράς, γραμμικά διαγράμματα κ.α.

- Scipy: Είναι μια συλλογή μαθηματικών αλγορίθμων και λειτουργιών που βασίζονται στη βιβλιοθήκη Numpy.
- Scikit-Learn: Είναι η βιβλιοθήκη της Python η οποία είναι κτισμένη πάνω στη Scipy και περιέχει τους αλγορίθμους της μηχανικής μάθησης.

3.2 Σύνολα δεδομένων

3.2.1 Σύνολο δεδομένων Pima Indian Diabetes Dataset

Η ανίχνευση του διαβήτη έχει αρκετά μεγάλη σημασία, κυρίως όσον αφορά τις επιπλοκές του. Αρκετές έρευνες έχουν πραγματοποιηθεί για την αναγνώριση του διαβήτη, βασισμένες στο σύνολο δεδομένων Pima Indian Diabetes Dataset [19]. Το σύνολο δεδομένων Pima Indian Diabetes Dataset ανήκει στο National Institute of Diabetes and Digestive and Kidney Diseases of the NIH [2][19], ξεκίνησε την καταγραφή των δεδομένων από το 1965, λόγω του ότι παρατηρήθηκε πως ο ρυθμός αύξησης του διαβήτη ήταν αρκετά ψηλός [19]. Είναι διαθέσιμο δωρεάν στη βάση του University of California, Irvine Machine Learning Repository.

Συνολικά υπάρχουν 768 καταγραφές ασθενών στο σύνολο δεδομένων. Οι εγγραφές αυτές αφορούν γυναικείο πληθυσμό ο οποίος βρίσκεται κοντά στο Phoenix της Αριζόνας των Ηνωμένων Πολιτειών και η ηλικία τους είναι μεγαλύτερη των 21 ετών [6]. Από τις 768 καταγραφές των ασθενών, οι 268 έχουν κατηγοριοποιηθεί ως θετικές ενδείξεις διαβήτη ενώ οι 500 έχουν καταγραφεί ως αρνητικές. Κάθε καταγραφή περιλαμβάνει στο σύνολο 9 αριθμητικές τιμές. Οι πρώτες 8 είναι τα χαρακτηριστικά τα οποία περιλαμβάνουν τα προσωπικά δεδομένα υγείας καθώς και αποτελέσματα εξετάσεων [7]. Η τελευταία τιμή είναι δυαδική και συμβολίζει την τιμή απόφασης, δηλαδή την κατηγοριοποίηση του ασθενή, αν είναι διαβητικός ή όχι. Στον πίνακα που ακολουθεί παρουσιάζονται τα πιο πάνω χαρακτηριστικά και μια σύντομη στατιστική ανάλυσή τους, όπως η μέση τιμή των τιμών των χαρακτηριστικών, η τυπική απόκλιση και η ελάχιστη και μέγιστη τιμή του κάθε χαρακτηριστικού.

ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ PIMA INDIAN DIABETES			
ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ	**ΜΕΣΟΣ**	**ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ**	**ΕΛΑΧΙΣΤΟ / ΜΕΓΙΣΤΟ**
Πόσες φορές ήταν έγκυος	3.8	3.4	0 / 17

Συγκέντρωση γλυκόζης πλάσματος και 2 ώρες ανοχής γλυκόζης διά στόματος	120.9	32.0	0 / 199
Διαστολική αρτηριακή πίεση (mm Hg)	69.1	19.4	0 / 122
Πάχος πτυχών δέρματος (mm)	20.5	16.0	0 / 99
Ινσουλίνη ορού 2 ωρών (mu U/ml)	79.8	115.2	0 / 846
Δείκτης μάζας σώματος (βάρος σε kg και ύψος σε m)	32.0	7.9	0 / 67.1
Διαβήτης γενεαλογικής λειτουργίας (DPF)	0.5	0.3	0.078 / 2.42
Ηλικία (έτη)	33.2	11.8	21 / 81
Μεταβλητή κλάσης (0 ή 1)			0 / 1

Πίνακας 5: Περιγραφή χαρακτηριστικών συνόλου δεδομένων Pima

Ο διαβήτης γενεαλογικής λειτουργίας (Diabetes Pedigree Function - DPF) αναφέρεται στο γενετικό παράγοντα του διαβήτη και υπολογίζεται από τον τύπο:

$$DPF = \frac{\sum_i K_i(88 - ADM_i) + 20}{\sum_i K_j(ALC_j - 14) + 50}$$

Εξίσωση 29: Διαβήτης γενεαλογικής λειτουργίας

Από τον πιο πάνω τύπο προκύπτει ότι το i κυμαίνεται μεταξύ των συγγενών που είχαν αναπτύξει διαβήτη, ενώ το j το αντίθετο. Ο συντελεστής K_x ισούται με το ποσοστό των γονιδίων που μοιράζονται με το αντίστοιχο x, και παίρνει τιμές 0.5, 0.25 και 0.125. Επίσης το ADM αντιπροσωπεύει την ηλικία του ασθενούς όταν διαγνώστηκε με διαβήτη. Τέλος, το ALC αναφέρεται στην ηλικία κατά την οποία πραγματοποιήθηκε η τελευταία μη διαβητική εξέταση.[19]

3.2.1.1 Χαμένες τιμές στο σύνολο δεδομένων

Ένα από τα προβλήματα που υπάρχουν στα διάφορα σύνολα δεδομένων είναι οι χαμένες τιμές ή αλλιώς μη λογικές τιμές, που αντιστοιχούν στα διάφορα χαρακτηριστικά του

συνόλου δεδομένων. Στο σύνολο δεδομένων που χρησιμοποιήθηκε, αρκετές τιμές διαφόρων χαρακτηριστικών ήταν μηδενικές. Σε έναν ασθενή το επίπεδο γλυκόζης που έχει στο αίμα είναι αδύνατο να είναι ίσο με μηδέν ή η αρτηριακή του πίεση να είναι και αυτή ίση με μηδέν. Ο Πίνακας 6 παρουσιάζει τα χαρακτηριστικά και το σύνολο των χαμένων τιμών που υπάρχουν στο σύνολο δεδομένων.

ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ PIMA INDIAN DIABETES	
ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ	**ΑΡΙΘΜΟΣ ΜΗΔΕΝΙΚΩΝ ΤΙΜΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟΥ**
Πόσες φορές ήταν έγκυος	0
Συγκέντρωση γλυκόζης πλάσματος και 2 ώρες ανοχής γλυκόζης διά στόματος	5
Διαστολική αρτηριακή πίεση (mm Hg)	35
Πάχος πτυχών δέρματος (mm)	227
Ινσουλίνη ορού 2 ωρών (mu U/ml)	374
Δείκτης μάζας σώματος (βάρος σε kg και ύψος σε m)	11
Διαβήτης γενεαλογικής λειτουργίας (DPF)	0
Ηλικία (έτη)	0

Πίνακας 6: Συνολικός αριθμός μηδενικών τιμών που υπάρχουν σε κάθε χαρακτηριστικό στο σύνολο δεδομένων.

3.2.2 Σύνολο δεδομένων από τον Δρ. John Schorling

Το δεύτερο σύνολο δεδομένων το οποίο χρησιμοποιήθηκε για την εύρεση του καλύτερου αλγορίθμου κατηγοριοποίησης του διαβήτη τύπου 2, προέρχεται από το τμήμα ιατρικής σχολής του πανεπιστημίου της Βιρτζίνιας. Το σύνολο δεδομένων δόθηκε από τον Δρ. John Schorling και περιέχει 15 μεταβλητές με 403 περιπτώσεις, από αυτές οι 338 περιπτώσεις κατηγοριοποιήθηκαν ως μη διαβητικές, ενώ οι υπόλοιπες 65 ως διαβητικές περιπτώσεις. Διατίθεται δωρεάν στο σύνδεσμο http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets.

ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ PIMA INDIAN DIABETES			
ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ	**ΜΕΣΟΣ**	**ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ**	**ΕΛΑΧΙΣΤΟ / ΜΕΓΙΣΤΟ**
Συνολική χοληστερόλη	207.85	44.45	78 / 4443
Σταθεροποιημένη χοληστερόλη	106.67	53.07	48 / 385
Λιποπρωτεΐνη υψηλής πυκνότητας	50.45	17.26	12 / 120
Δείκτης χοληστερόλης / HDL	4.52	1.73	1.5 / 19.3
Γλυκοζολιωμένη αιμοσφαιρίνη	5.59	2.24	2.68 / 16.11
Ηλικία	46.86	16.31	19 / 92
Φύλο	-	-	Male / Female
Ύψος	66.02	3.92	52 / 76
Βάρος	177.59	40.34	99 / 325
Πρώτη συστολική αρτηριακή πίεση	136.9	22.74	90 / 250
Πρώτη διαστολική αρτηριακή πίεση	83.32	13.59	48 / 124
Δεύτερη συστολική αρτηριακή πίεση	152.38	21.71	110 / 238
Δεύτερη διαστολική αρτηριακή πίεση	92.52	11.56	60 / 124
Μέση	37.90	5.73	26 / 56
Ισχίο	43.03	5.65	30 / 64

Πίνακας 7: Περιγραφή χαρακτηριστικών συνόλου δεδομένων από τον Δρ. John Schorling.

Στο σύνολο δεδομένων δημιουργήθηκε μια νέα στήλη, η οποία κατηγοριοποιεί τα δεδομένα. Πιο συγκεκριμένα, όταν η τιμή της γλυκοζυλιωμένης αιμοσφαιρίνης είναι μεγαλύτερη ή ίση με 6.5 τότε κατηγοριοποιείται σαν θετική ένδειξη διαβήτη, ενώ όταν είναι μικρότερη από 6.5 τότε κατηγοριοποιείται σαν αρνητική ένδειξη διαβήτη. Επίσης έγινε η διαγραφή των εγγραφών στα οποία η τιμή της γλυκοζυλιωμένης αιμοσφαιρίνης δεν υπάρχει.

3.3 Επιλογή αλγορίθμων

Όλοι οι αλγόριθμοι οι οποίοι χρησιμοποιήθηκαν βρίσκονται υλοποιημένοι στη βιβλιοθήκη της Python, την Scikit-Learn. Για την επιλογή των πιο πάνω αλγορίθμων, σημαντικό ρόλο έπαιξαν πολλοί παράγοντες, διαφορετικοί σε κάθε περίπτωση. Οι λόγοι που οδηγήθηκαν στην επιλογή αυτών των αλγορίθμων είναι οι εξής:

- Λογιστική παλινδρόμηση: Η λογιστική παλινδρόμηση είναι ο απλούστερος μη γραμμικός κατηγοριοποιητής με γραμμικό συνδυασμό παραμέτρων και μη χρήση της σιγμοειδούς συνάρτησης για δυαδική κατηγοριοποίηση. Το σύνολο μας αποτελείται από δύο κατηγορίες, επομένως η χρήση της λογιστικής παλινδρόμησης αναμένεται να έχει καλά αποτελέσματα. Για τη μεταπτυχιακή διπλωματική εργασία χρησιμοποιήθηκε ο αλγόριθμος LogisticRegression().
- Δένδρα απόφασης: Τα δένδρα απόφασης είναι αρκετά παρόμοια με τη διαδικασία λήψης απόφασης από τον άνθρωπο καθώς και σε συνδυασμό ότι η ερμηνεία τους είναι εύκολη, αποτελεί μια καλή λύση του προβλήματος κατηγοριοποίησης. Επίσης, εκτός από μέθοδος κατηγοριοποίησης χρησιμοποιείται και για παλινδρόμηση. Χρησιμοποιούνται κυρίως σε συνδυασμό, όπως το τυχαίο δάσος ή σε αλγόριθμους ενίσχυσης. Έτσι οι αλγόριθμοι οι οποίοι επιλέχτηκαν είναι ο DecisionTreeClassifier() για δένδρο απόφασης κατηγοριοποίησης, AdaBoostClassifier() και GradientBoostingClassifier() για αλγορίθμους ενίσχυσης και RandomForestClassifier() για τυχαίο δάσος.
- KNN: Ο αλγόριθμος KNN είναι και αυτός από τους απλούστερους αλγορίθμους για κατηγοριοποίηση, ο οποίος βασίζεται σε διάφορα μέτρα απόστασης τα οποία αναφέρθηκαν στο κεφάλαιο 2. Τα μέτρα αυτά δίνουν τη δυνατότητα στον αλγόριθμο KNN να παράγει διαφορετικά αποτελέσματα. Ο αλγόριθμος ο οποίος χρησιμοποιήθηκε είναι ο KNeighborsClassifier().
- Νευρωνικό Δίκτυο: Τα νευρωνικά δίκτυα είναι η νέα γενιά των αλγορίθμων που εφαρμόζονται για την επίλυση πολλών προβλημάτων. Τα νευρωνικά δίκτυα

λειτουργούν όπως ο εγκέφαλος του ανθρώπου με αποτέλεσμα να μαθαίνουν μόνα τους. Ο αλγόριθμος που χρησιμοποιήθηκε για το νευρωνικό δίκτυο είναι ο MLPClassifier().

- Μηχανές Διανυσμάτων Στήριξης: Πολλές φορές τα δεδομένα δεν μπορούν να διαχωριστούν με μία ευθεία γραμμή, έτσι χρειάζεται η ανάλυσή τους σε ένα μεγαλύτερο χώρο διαστάσεων χρησιμοποιώντας τους πυρήνες. Στη συγκεκριμένη περίπτωση ο αλγόριθμος που χρησιμοποιήθηκε είναι ο LinearSVC().
- Δεδομένα ίδια κλίμακας: Στα δεδομένα ίδιας κλίμακας οι κύριοι αλγόριθμοι οι οποίοι χρησιμοποιούνται είναι ο αλγόριθμος MinMaxScaler() και StandardScaler(), για κανονικοποίηση και τυποποίηση αντίστοιχα.
- Μείωση διαστάσεων: Τέλος, στη μείωση διαστάσεων και συγκεκριμένα στην ανάλυση κυρίων συνιστωσών χρησιμοποιήθηκε ο αλγόριθμος PCA, αφού είναι ο καλύτερος αλγόριθμος στη μείωση διαστάσεων με ελάχιστη απώλεια πληροφοριών.

3.3.1 Παράμετροι αλγορίθμων

Για τη λειτουργία του κάθε αλγορίθμου ο οποίος αναφέρθηκε και χρησιμοποιήθηκε δέχεται διάφορους παραμέτρους. Οι παράμετροι αυτοί παίζουν σημαντικό ρόλο ως προς τη συμπεριφορά του κάθε αλγορίθμου ξεχωριστά. Για τον κάθε αλγόριθμο χρησιμοποιήθηκαν οι πιο κάτω παράμετροι:

LogisticRegression:

Οι παράμετροι οι οποίες εφαρμόστηκαν κατά την υλοποίηση του αλγορίθμου Logistic Regression είναι οι penalty, dual, tol και C. Η παράμετρος penalty χρησιμοποιείται για τον καθορισμό της τιμωρίας. Οι τιμές που λαμβάνει είναι str, l1, l2, elasticnet ή none. Οι βασικές τιμές που χρησιμοποιούνται είναι η l1 και l2. Η διαφορά μεταξύ αυτών των δύο τιμών είναι η τεχνική της κανονικοποίησης που θα εφαρμοστεί, αφού η τιμή l1 χρησιμοποιεί την τεχνική παλινδρόμηση Lasso (Least Absolute Shrinkage and Selection Operator), ενώ η τιμή l2 χρησιμοποίει την παλινδρόμηση Ridge. Η παλινδρόμηση Lasso χρησιμοποιεί την απόλυτη τιμή του μεγέθους του συντελεστή ως όρο τιμωρίας, ενώ η παλινδρόμηση Ridge χρησιμοποιεί το τετραγωνικό μέγεθος του συντελεστή. Οι πιο πάνω παλινδρομήσεις υπολογίζονται από τους εξής τύπους:

$$RidgeRegression = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Εξίσωση 30: Ridge παλινδρόμηση

$$LassoRegression = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Εξίσωση 31: Lasso παλινδρόμηση

Η παράμετρος dual χρησιμοποιείται μόνο όταν η τιμωρία είναι l2. Παίρνει τιμές False και True. Η tol είναι η παράμετρος που δηλώνει την ανοχή για τον τερματισμό των κριτηρίων. Τέλος, η παράμετρος C πρέπει να είναι ένας θετικός αριθμός.

Για τις πιο πάνω παραμέτρους χρησιμοποιήθηκαν οι ακόλουθες τιμές:

- **penalty:** l2
- **dual:** False
- **tol:** 1e-4
- **C:** 1.0

KNeighborsClassifier:

Στον αλγόριθμο KNeighborClass σημαντικό ρόλο παίζει ο καθορισμός των παραμέτρων n_neighbors, weighs και metric. Η παράμετρος n_neighbors είναι ο αριθμός γειτόνων με προκαθορισμένη τιμή 5. Η weighs είναι η συνάρτηση βάρους που χρησιμοποιείται για την πρόβλεψη. Οι τιμές που δέχεται είναι uniform, distance ή callable. Τέλος η metric είναι η απόσταση που χρησιμοποιείται για την εύρεση του κοντινότερου γείτονα.

Οι τιμές που χρησιμοποιήθηκαν για τις πιο πάνω παραμέτρους είναι οι εξής:

- **n_neighbors:** 5
- **weighs:** uniform, όπου όλα τα σημεία στον κάθε γείτονα είναι ίσα.
- **metric:** minkowski, Euclidean, hamming

LinearSVC:

Ο αλγόριθμος LinearSVC προέρχεται από τον αλγόριθμο SVC με την παράμετρο της συνάρτησης του πυρήνα να είναι γραμμική (kernel = linear). Σε αντίθεση με τον αλγόριθμο SVC ο οποίος είναι κτισμένος στη βιβλιοθήκη libsvm (Library Support Vector Machines), ο αλγόριθμος LinearSVC βασίζεται στην βιβλιοθήκη liblinear (Library Large Linear Classification), δίνοντάς του έτσι την ευελιξία στην επιλογή των ποινών. Οι παράμετροι που δέχεται ο αλγόριθμος είναι penalty, tol και C. Η παράμετρος penalty παίρνει τιμές l1 ή l2, με προεπιλεγμένη την τιμή l2, η tol είναι η παράμετρος που δηλώνει την ανοχή για τον τερματισμό

των κριτηρίων και η παράμετρος C πρέπει να είναι ένας θετικός αριθμός. Για τον αλγόριθμο αυτό χρησιμοποιήθηκαν οι εξής παράμετροι:

- **penalty:** παίρνει τιμές l1 ή l2, με προεπιλεγμένη την τιμή l2, η οποία χρησιμοποιήθηκε στη μεταπτυχιακή διπλωματική εργασία.
- **tol:** 1e-4
- **C:** 1.0

DecisionTreeClassifier:

Σημαντικοί παράμετροι κατά την υλοποίηση του του αλγορίθμου DecisionTreeClassifier είναι η παράμετρος criterion και max_depth. Η παράμετρος criterion είναι η συνάρτηση κριτηρίων διαχωρισμού η οποία παίρνει τιμές Gini και entropy και η max_depth είναι το μέγιστο βάθος του δέντρου στο οποίο αν δεν υπάρχει κάποια τιμή, τότε το δέντρο επεκτείνεται μέχρι το τέλος. Οι τιμές που έλαβε η κάθε παράμετρος είναι:

- **criterion:** Gini.
- **max_depth:** Δεν χρησιμοποιήθηκε κάποια τιμή.

RandomForestClassifier:

Ο αλγόριθμος RandomForest έχει ως παραμέτρους το n_estimators και criterion. Η n_estimators είναι ο αριθμός των δέντρων αποφάσεων που υπάρχουν μέσα σε ένα τυχαίο δάσος και το criterion είναι η συνάρτηση κριτηρίων διαχωρισμού το οποίο παίρνει τιμές Gini και entropy, όπως στον αλγόριθμο DecisionTreeClassifier. Κάθε παράμετρος πήρε τις εξής τιμές:

- **n_estimaros:** 10.
- **criterion:** Gini.

AdaBoostClassifier:

Για τον αλγόριθμο ενίσχυσης AdaBoostClassifier χρησιμοποιήθηκαν οι παράμετροι base_estimator και n_estimators. Ο base_estimator είναι ο εκτιμητής βάσης από όπου είναι κτισμένο το νέο σύνολο δεδομένων. Στη περίπτωση που δεν υπάρχει κάποια τιμή τότε ο base_estimator δέχεται την τιμή DecistionTreeClassifier με max_depth=1. Τέλος η παράμετρος n_estimators είναι ο μεγαλύτερος αριθμός εκτιμητή που τερματίζεται η ενίσχυση. Οι τιμές που χρησιμοποιήθηκαν για την κάθε παράμετρο είναι:

- **base_estimator:** DecisionTreeClassifier με max_depth = 1.

- **n_estimators:** 50.

GradientBoostClassifier:

Άλλος ένας αλγόριθμος ενίσχυσης, ο οποίος χρησιμοποιήθηκε στα πλαίσια της μεταπτυχιακής διπλωματικής εργασίας, είναι ο GradientBoostClassifier, όπου χρησιμοποιήθηκαν οι παράμετροί loss και n_estimators. Οι τιμές που δέχεται η παράμετρος loss είναι οι deviance και exponential. Η τιμή deviance αναφέρεται στη βελτίωση της τυπικής απόκλισης χρησιμοποιώντας έτσι τη λογιστική παλινδρόμηση, ενώ η τιμή exponential χρησιμοποιεί τον αλγόριθμο AdaBoost. Η παράμετρος n_estimators είναι ο αριθμός των φάσεων της ενίσχυσης που πρέπει να εκτελεστούν, με ένα μεγάλο αριθμό να έχει συνήθως καλύτερη απόδοση. Έτσι χρησιμοποιήθηκαν οι παρακάτω τιμές:

- **loss:** deviance.
- **n_estimators:** 100.

MLPClassifier:

Όσο αφορά το νευρωνικό δίκτυο το οποίο χρησιμοποιήθηκε για τον αλγόριθμο MLPClassifier, χρησιμοποιήθηκαν οι παραμέτροι hidden_layers_sizes, activation και solver. Η παράμετρος hidden_layers_sizes είναι ο συνολικός αριθμός των νευρώνων σε κάθε κρυμμένο επίπεδο. Επίσης, η παράμετρος activation είναι η συνάρτηση ενεργοποίησης του κρυμμένου επιπέδου. Οι τιμές οι οποίες δέχεται η παράμετρος activation είναι identity, logistic, tanh και relu. Τέλος, η παράμετρος solver καθορίζει τη βελτιστοποίηση του βάρους. Οι τιμές που δέχεται είναι lbfgs, sgd και adam. Η προκαθορισμένη τιμή adam δουλεύει καλύτερα όταν έχουμε πολύ μεγάλο σύνολο δεδομένων, όσο αφορά την ταχύτητα και το αποτέλεσμα επαλήθευσης. Για μικρότερα σύνολα δεδομένων χρησιμοποιείται η τιμή lbfgs. Οι τιμές για τον αλγόριθμο MLPClassifier που χρησιμοποιήθηκαν είναι:

- **hidden_layers_sizes:** Τυχαία τιμή, καθορισμένη από τον υπολογιστή σε εύρος 0 έως 100.
- **activation:** relu.
- **solver:** lbfgs

PCA:

Για τον αλγόριθμος που χρησιμοποιήθηκε για τη μείωση των διαστάσεων, PCA, η μόνη παράμετρος που ορίστηκε είναι η παράμετρος n_components, δηλαδή ο αριθμός των διαστάσεων που θα παραμείνουν στο σύνολο δεδομένων. Το εύρος τιμών που χρησιμοποιήθηκε είναι:

- **n_components:** από 2 έως 7.

MinMaxScaler:

Ο αλγόριθμος MinMaxScaler δέχεται σαν είσοδο την παράμετρο feature_range, η οποία υποδηλώνει το εύρος των μετασχηματισμένων δεδομένων. Χρησιμοποιήθηκαν οι εξής τιμές:

- **feature_range:** 0 και 1.

StandardScaler:

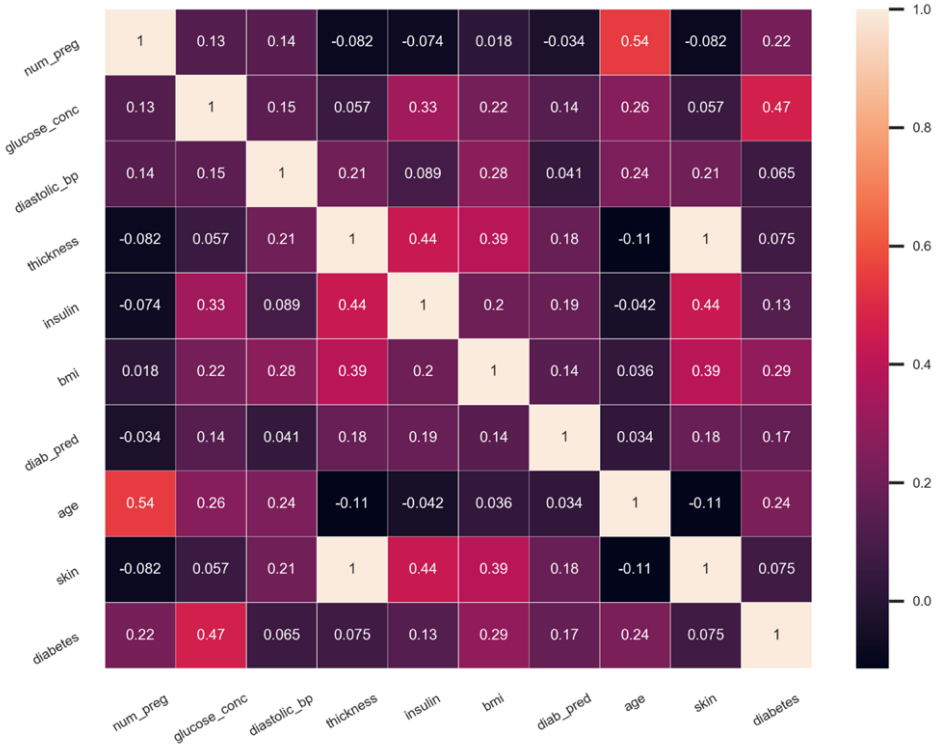
Τελευταίος αλγόριθμος, ο αλγόριθμος StandardScaler, δέχεται δύο παραμέτρους, την with_mean και with_std. Η with_mean δέχεται τιμές True ή False και σε περίπτωση που η τιμή είναι True τότε τα δεδομένα κεντράρονται πριν την κλιμάκωση. Τέλος, η παράμετρος with_std δέχεται και αυτή τις τιμές True ή False. Σε περίπτωση που η τιμή είναι True τότε μεταβάλλεται η διακύμανση των δεδομένων σε μονάδα. Οι τιμές που πήραν οι παράμετροι είναι:

- **with_mean:** True.
- **with_std:** True.

3.4 Βήματα υλοποίησης

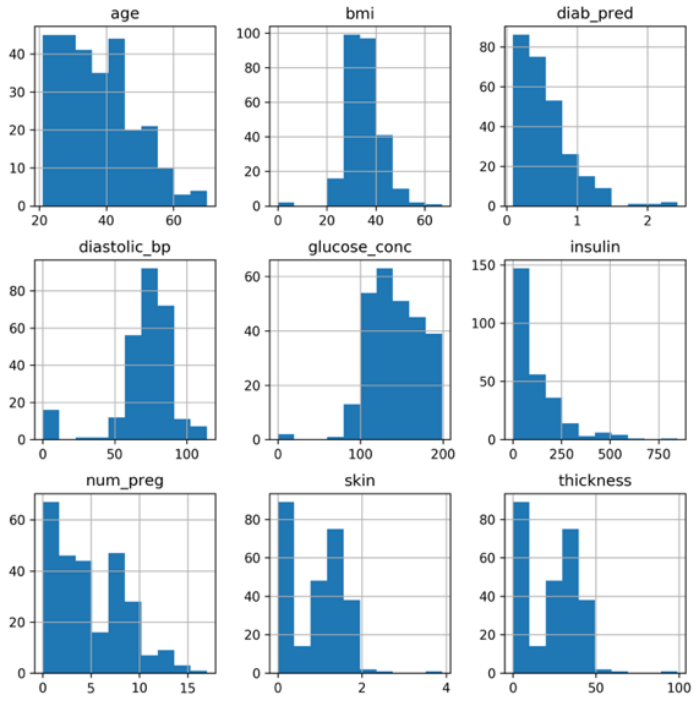
Για την υλοποίηση της μεταπτυχιακής διπλωματικής εργασίας χρησιμοποιήθηκαν οι πιο πάνω αλγόριθμοι, με σκοπό την εύρεση του καλύτερου αλγόριθμου για την κατηγοριοποίηση του συγκεκριμένου συνόλου δεδομένων. Πιο συγκεκριμένα ακολουθήθηκε η εξής διαδικασία:

1. Ανάγνωση αρχείου pima_data_diabates.csv με τη χρήση της βιβλιοθήκης pandas και αποθήκευση των τιμών αυτών σε ένα DataFrame.
2. Εκτύπωση του πίνακα συσχέτισης και ιστογραμμάτων για την πρώτη επαφή με τα δεδομένα με τη χρήση του Matplotlib.



Εικόνα 17: Πίνακας συσχέτισης συνόλου δεδομένων Pima Indian.

Ο πίνακας συσχέτισης δείχνει τη συσχέτιση που υπάρχει μεταξύ των μεταβλητών στο σύνολο δεδομένων. Όσο πιο έντονο είναι το χρώμα τόσο λιγότερη είναι η συσχέτιση μεταξύ των στοιχείων, ενώ όσο πιο άτονο είναι τόσο μεγαλύτερη είναι η συσχέτιση μεταξύ τους. Εάν η τιμή είναι κίτρινο, τότε η συσχέτιση μεταξύ των στοιχείων είναι απόλυτη.



Εικόνα 18: Ιστογράμματα συνόλου δεδομένων Pima Indian.

3. Φόρτωση και δημιουργία λίστας με όλους τους αλγορίθμους που θα χρησιμοποιηθούν για την κατηγοριοποίηση από τη βιβλιοθήκη Sklearn.
4. Εφαρμογή των πιο πάνω αλγόριθμων χρησιμοποιώντας το αρχικό DataFrame, δηλαδή το σύνολο δεδομένων στην αρχική του μορφή χωρίς να έχει υποστεί την οποιαδήποτε επεξεργασία.
5. Αντικατάσταση των χαμένων – μηδενικών τιμών του συνόλου δεδομένων με τη μέση τιμή, δημιουργώντας ένα καινούριο DataFrame. Έπειτα γίνεται η εκτέλεση των αλγορίθμων βασισμένη σε αυτό το DataFrame.
6. Διαγραφή εγγραφής η οποία περιέχει έστω μια μηδενική τιμή από το σύνολο δεδομένων, αποθηκεύοντάς το σε ένα διαφορετικό DataFrame. Γίνεται επανάληψη των αλγορίθμων.
7. Δημιουργία δύο νέων DataFrame βασισμένα στο αρχικό, τα οποία έχουν υποστεί κάποια προεπεξεργασία στα δεδομένα τους με χρήση των αλγορίθμων MinMaxScaler και StandardScaler. Εφαρμόζονται οι αλγόριθμοι σε αυτά τα DataFrame και στη συνέχεια ακολουθούνται τα βήματα 5 και 6 σε κάθε ένα από αυτά τα DataFrame.
8. Τελευταία μέθοδος που εφαρμόστηκε είναι η μείωση διαστάσεων από δύο διαστάσεις μέχρι οκτώ στο αρχικό σύνολο δεδομένων. Όπως και στο προηγούμενο βήμα, εφαρμόζονται και πάλι τα βήματα 4, 5 και 6.

Όλα τα βήματα εφαρμόζονται για τις μεθόδους K – Fold Cross Validation και Split. Στη μέθοδο K – Fold Cross Validation η τιμή των folds είναι 10, ενώ στο Split είναι 80%. Για καλύτερα αποτελέσματα η κάθε διαδικασία εκτελείται 10 φορές. Οι τιμές των αποτελεσμάτων αποθηκεύονται σε μία λίστα και στη συνέχεια υπολογίζεται ο μέσος όρος τους.

Κεφάλαιο 4: Αποτελέσματα

Στο κεφάλαιο αυτό αρχικά γίνεται μια συνοπτική αναφορά των αποτελεσμάτων από διάφορες διεθνείς μελέτες. Παρουσιάζονται τρεις συνοπτικά μελέτες όπου ο σκοπός τους ήταν η εύρεση του καλύτερου αλγορίθμου με τη μεγαλύτερη ακρίβεια στην πρόβλεψη του σακχαρώδη διαβήτη. Και στις τρεις μελέτες χρησιμοποιήθηκε το σύνολο δεδομένων Pima Indian, ενώ στη τρίτη μελέτη χρησιμοποιήθηκε και το σύνολο δεδομένων από τον Δρ. John Schorling. Τα δύο σύνολα δεδομένων παρουσιάστηκαν στο κεφάλαιο 3, αφού είναι τα σύνολα δεδομένων που χρησιμοποιούνται στη παρούσα μεταπτυχιακή διπλωματική εργασία. Έπειτα, γίνεται η αναφορά όλων των αποτελεσμάτων τα οποία πάρθηκαν. Επίσης υπάρχει μια συγκριτική αξιολόγηση αποτελεσμάτων της ορθότητας, ακρίβειας, ανάκλησης και αρμονικού μέσου μεταξύ των δύο συνόλων δεδομένων. Τέλος γίνεται η επιλογή του καλύτερου αλγορίθμου.

Για να έχουμε καλύτερα αποτελέσματα, η εφαρμογή των αλγορίθμων μόνο μια φορά δεν δίνει μια ξεκάθαρη εικόνα. Έτσι, εφαρμόστηκαν οι αλγόριθμοι 10 φορές ο κάθε ένας. Σαν αποτέλεσμα πάρθηκε ο μέσος όρος των αποτελεσμάτων. Τα αποτελέσματα που προκύπτουν είναι τα εξής:

4.1 Αποτελέσματα από προηγούμενες μελέτες

4.1.1 1η Μελέτη

Η μελέτη των Sidong Wei, Xuejiao Zhao και Chunyan Miao με τίτλο *A Comprehensive Exploration to the Machine Learning Techniques for Diabetes Identification* [19] χρησιμοποίηση το σύνολο δεδομένων Pima Indian και είχε ως σκοπό τη σύγκριση της ακρίβειας των ακολούθων κατηγοριοποιητών:

- Λογιστική Παλινδρόμηση
- Βαθύ Νευρωνικό Δίκτυο (Deep Neural Network – DNN)
- Support Vector Machines
- Δένδρο απόφασης (Decision Tree)
- Απλοϊκός Bayes

Κατά τη διάρκεια της προεπεξεργασίας εφαρμόστηκαν οι τεχνικές αντικατάστασης χαμένων τιμών με τη μέση τιμή, κανονικοποίηση, τυποποίηση, καθώς και ο συνδυασμός των πιο πάνω τεχνικών.

Έπειτα εφαρμόστηκε η μέθοδος K – Fold Cross Validation, με αριθμό K να είναι 10. Αυτό σημαίνει ότι το σύνολο διαιρέθηκε σε 10 υποσύνολα, από τα οποία τα 9 χρησιμοποιήθηκαν για την εκπαίδευση του συνόλου δεδομένου και το 10ο για την επαλήθευση των δεδομένων.

Τα αποτελέσματα της ακρίβειας παρουσιάζονται στον Πίνακα 8:

ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗΣ	ΑΚΡΙΒΕΙΑ
Λογιστική Παλινδρόμηση	0.77474
DNN	0.778646
SVM	0.776042
Decision Tree	0.763021
Απλοϊκός Bayes	0.757861

Πίνακας 8: Αποτελέσματα 1ης μελέτης

Με βάση τα πιο πάνω αποτελέσματα, παρατηρούμε ότι καλύτερα αποτελέσματα προκύπτουν από την εφαρμογή του βαθύ νευρωνικού δικτύου.

4.1.2 2η Μελέτη

Η μελέτη *A comparative study on the pre-processing and mining of Pima Indian Diabetes Dataset* των Amatul Zehra, Tuty Asmawaty και M.A M. Aznan [5] συγκρίνει την ακρίβεια των κατηγοριοποιητών, τόσο σε μη επεξεργασμένα μοντέλα, όσο και σε δεδομένα που έχουν υποστεί κάποια επεξεργασία. Το σύνολο δεδομένων το οποίο χρησιμοποιήθηκε για τον υπολογισμό της ακρίβειας ήταν το Pima Indian. Οι κατηγοριοποιητές που χρησιμοποιήθηκαν είναι οι ακόλουθοι:

- Απλοϊκός Bayes
- Multilayer Perceptron
- Decision Table
- J48
- Simple Cart

Αρχικά, στην πρώτη φάση επεξεργασίας των δεδομένων, αφαιρέθηκαν όλες οι εγγραφές που είχαν έστω μια μηδενική τιμή. Έπειτα, εφαρμόστηκε η τεχνική διακριτικοποίησης (discretization), η οποία μετατρέπει τις συνεχείς τιμές σε ένα πεπερασμένο σύνολο δεδομένων.

Τα αποτελέσματα της ακρίβειας των κατηγοριοποιητών πριν από την επεξεργασία καθώς και μετά την επεξεργασία παρουσιάζονται στον Πίνακα 9:

ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗΣ	ΑΚΡΙΒΕΙΑ ΠΡΙΝ ΑΠΟ ΤΗΝ ΕΠΕΞΕΡΓΑΣΙΑ	ΑΚΡΙΒΕΙΑ ΜΕΤΑ ΑΠΟ ΤΗΝ ΕΠΕΞΕΡΓΑΣΙΑ
Λογιστική Παλινδρόμηση	76.3%	80.3%
Multilayer Perceptron	75.39%	81%
Decision Table	71.22%	85.2%
J48	73.82%	80%
Simple Cart	75.13%	79.6%

Πίνακας 9: Αποτελέσματα 2ης μελέτης

Καλύτερα αποτελέσματα παρουσιάζονται με την εφαρμογή προεπεξεργασίας των δεδομένων και με τη χρήση του αλγορίθμου Decision Table.

4.1.3 3η Μελέτη

Τελευταία μελέτη, η *Type 2 diabetes mellitus prediction model based on data mining* των Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang [7] συγκρίνει την ακρίβεια της λογιστικής παλινδρόμησης χρησιμοποιώντας δύο διαφορετικά σύνολα δεδομένων, το σύνολο δεδομένων Pima Indian και το σύνολο δεδομένων του Δρ. John Schorling.

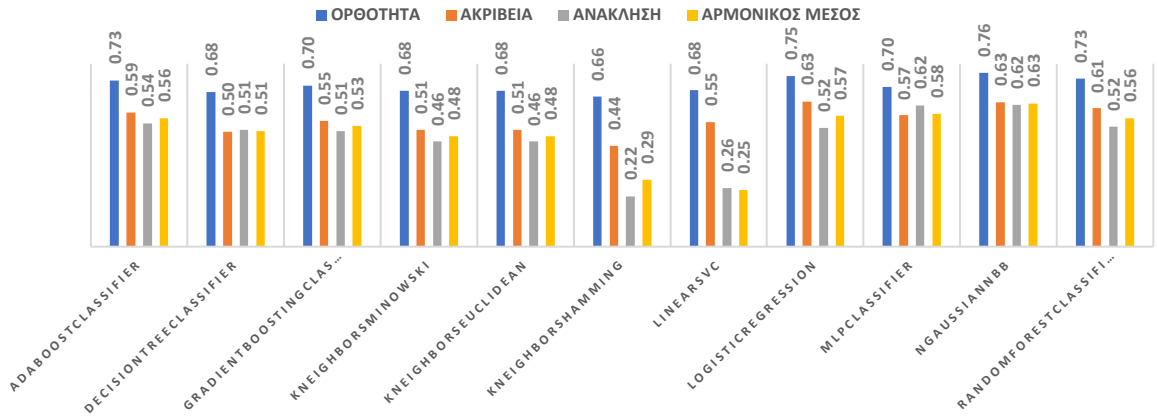
Στη φάση της προεπεξεργασίας, γίνεται η εφαρμογή της αντικατάστασης της χαμένης τιμής με τη μέση τιμή. Έπειτα, εφαρμόστηκε ο αλγόριθμος συσταδοποίησης K Means και τέλος η λογιστική παλινδρόμηση. Τα αποτελέσματα της ακρίβειας που πάρθηκαν στα δύο σύνολα δεδομένων είναι 95.42% στο σύνολο δεδομένων Pima Indian και 93.7% στο σύνολο δεδομένων του Δρ. John Schorling.

4.2 Αποτελέσματα με χρήση της μεθόδου Split

Οι επόμενες γραφικές παραστάσεις παρουσιάζουν όλα τα αποτελέσματα τα οποία προκύπτουν εφαρμόζοντας τους αλγορίθμους μηχανικής μάθησης στο σύνολο δεδομένων Pima Indian, όπου η διαχώρισή τους σε σύνολο εκπαίδευσης και σύνολο δοκιμής είναι η μέθοδος Split.

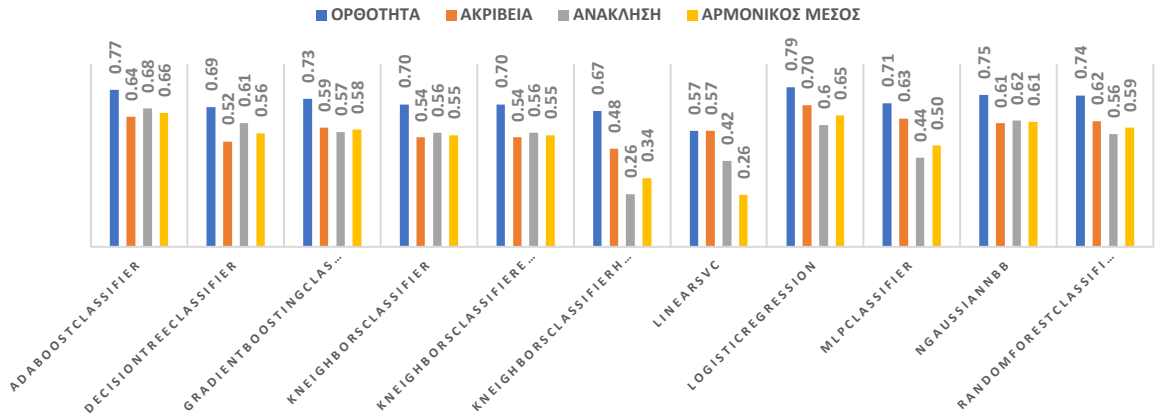
4.2.1 Εφαρμογή αλγορίθμων χωρίς προεπεξεργασία

4.2.1.1 Χρήση αρχικού συνόλου δεδομένων



Χρησιμοποιώντας το αρχικό σύνολο δεδομένων χωρίς κάποια επεξεργασία παρατηρούμε ότι καλύτερη ορθότητα στην κατηγοριοποίηση των δεδομένων την κάνει ο απλοϊκός Bayes με ποσοστό 0.76. Επίσης, εκτός από καλύτερη ορθότητα, ο απλοϊκός Bayes έχει και τα καλύτερα αποτελέσματα όσο αφορά την ακρίβεια, την ανάκληση και τον αρμονικό μέσο με 0.63, 0.62 και 0.63 αντίστοιχα.

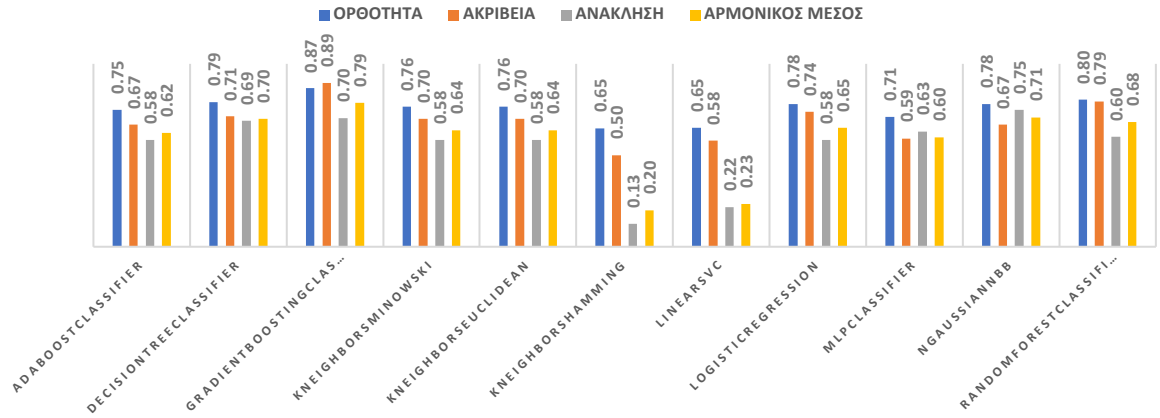
4.2.1.2 Αντικατάσταση μηδενικής τιμής με τη μέση τιμή



Αντικαθιστώντας τη μηδενική τιμή με τη μέση τιμή, καλύτερο αποτέλεσμα στην ορθότητα και στην ακρίβεια παρουσιάζει η λογιστική παλινδρόμηση με 0.79 και 0.7

αντίστοιχα. Καλύτερο αποτέλεσμα την ανάκληση και στον αρμονικό μέσο παρουσιάζει ο αλγόριθμος AdaBoost με αποτέλεσμα 0.68 και 0.66 αντίστοιχα.

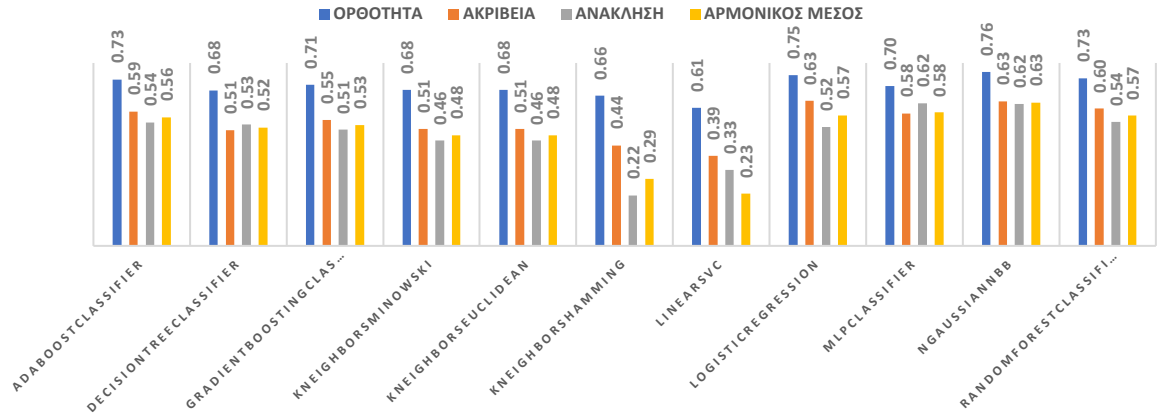
4.2.1.3 Διαγραφή εγγραφής η οποία περιέχει μία μηδενική τιμή



Διαγράφοντας κάθε εγγραφή η οποία έχει έστω και μια μηδενική τιμή ίση με το μηδέν καλύτερα αποτελέσματα δίνει ο αλγόριθμος GradientBoost όσο αφορά το αποτέλεσμα της ορθότητας, της ακρίβειας και του αρμονικού μέσου με αποτέλεσμα 0.87, 0.89 και 0.79, αντίστοιχα. Ο απλοϊκός Bayes δίνει καλύτερα αποτελέσματα στην ανάκληση με 0.75.

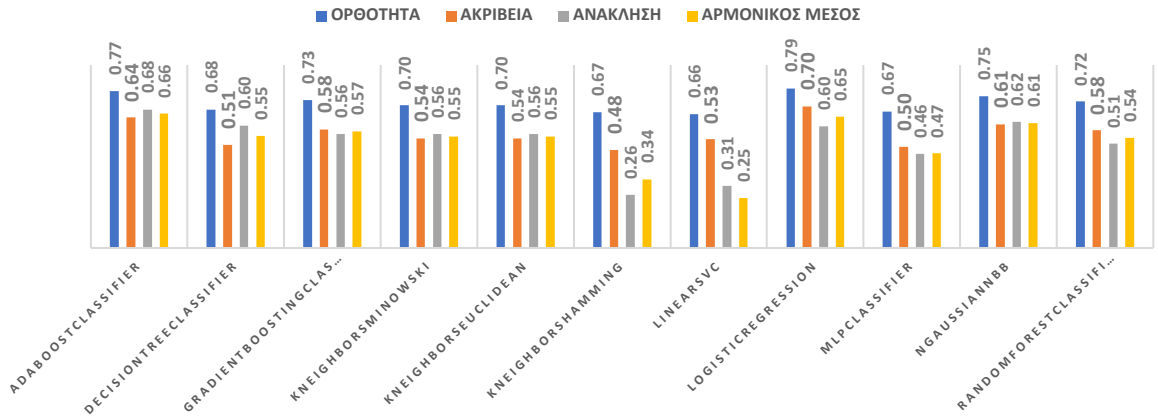
4.2.2 Εφαρμογή MinMaxScaler

4.2.2.1 Εφαρμογή MinMaxScaler στο αρχικό σύνολο δεδομένων



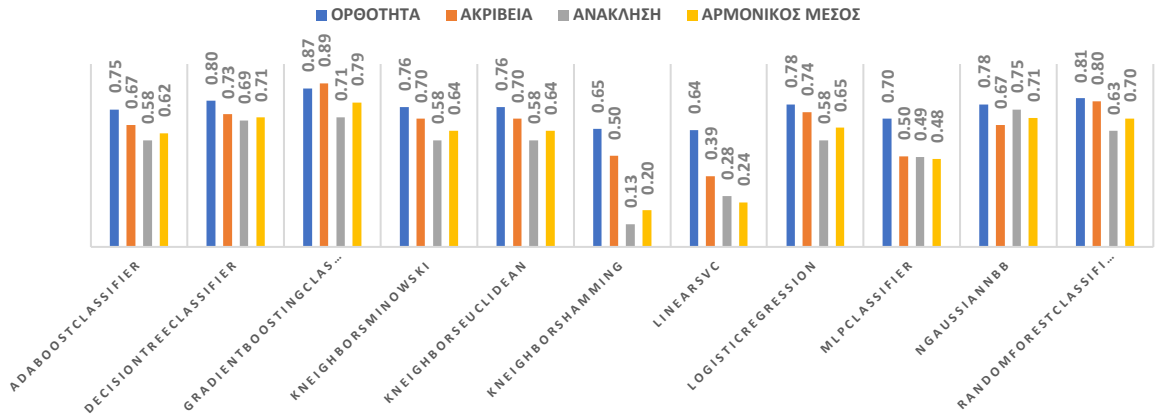
Εφαρμόζοντας τον αλγόριθμο MinMaxScaler στο αρχικό σύνολο δεδομένων, τα καλύτερα αποτελέσματα όσο αφορά την ορθότητα, την ακρίβεια, την ανάκληση και τον αρμονικό μέσο δίνονται από τον απλοϊκό Bayes με τιμές 0.76, 0.63, 0.62 και 0.63 αντίστοιχα.

4.2.2.2 Εφαρμογή MinMaxScaler και αντικατάσταση μηδενικής τιμής με μέση τιμή



Αφού εφαρμοστεί ο αλγόριθμος MinMaxScaler και στη συνέχεια γίνει η αντικατάσταση της μηδενικής τιμές με τη μέση τιμή, καλύτερα αποτελέσματα στην ορθότητα και στην ακρίβεια δίνει η λογιστική παλινδρόμηση με 0.79 και 0.70 αντίστοιχα. Σχετικά με την ανάκληση και τον αρμονικό μέσο καλύτερα αποτελέσματα έχουμε από τον αλγόριθμο AdaBoost με τιμές 0.68 και 0.66 αντίστοιχα.

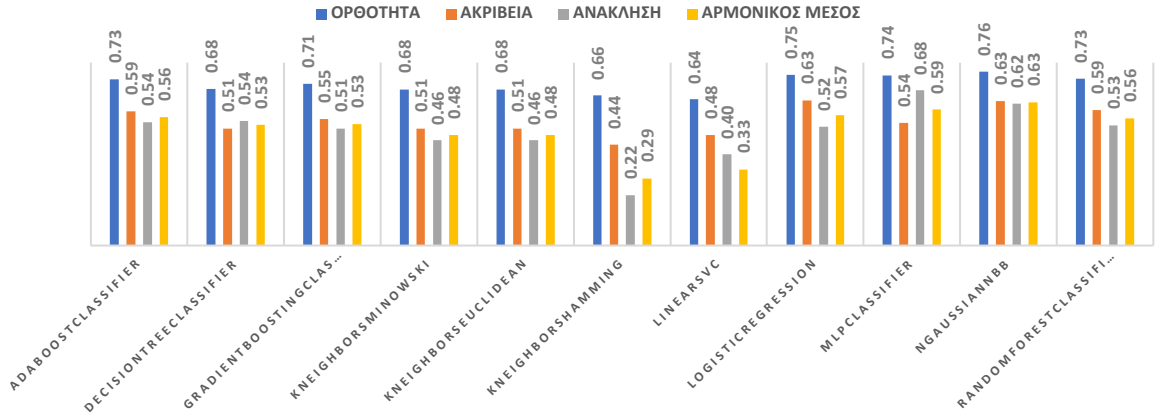
4.2.2.3 Εφαρμογή MinMaxScaler και διαγραφή εγγραφής με μια μηδενική τιμή



Εφαρμόζοντας τον αλγόριθμο MinMaxScaler και στη συνέχεια διαγράφοντας την εγγραφή η οποία περιέχει έστω και μια μηδενική τιμή, καλύτερα αποτελέσματα στην ορθότητα, στην ακρίβεια και στον αρμονικό μέσο έχουμε από τον GradientBoost με 0.87, 0.89 και 0.79 αντίστοιχα, ενώ καλύτερο αποτέλεσμα στην ανάκληση έχει ο απλοϊκός Bayes με 0.75.

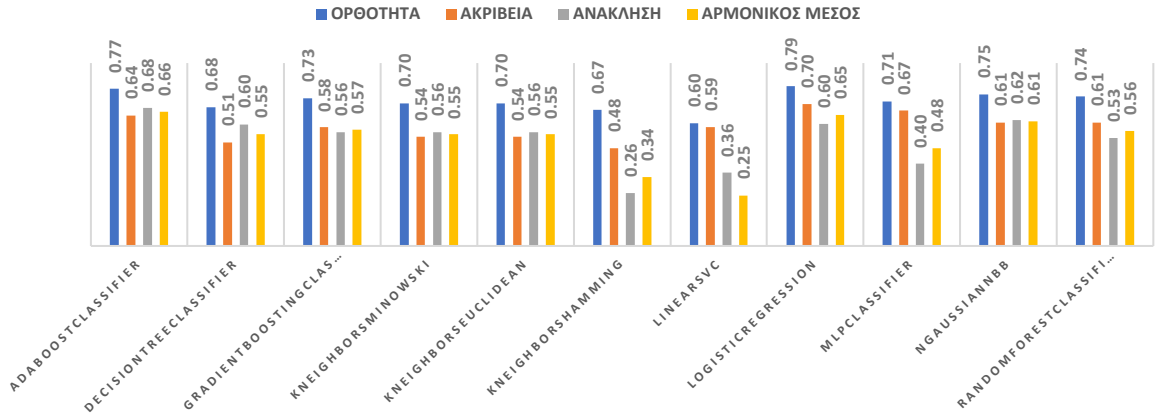
4.2.3 Εφαρμογή StandardScaler

4.2.3.1 Εφαρμογή StandardScaler στο αρχικό σύνολο δεδομένων



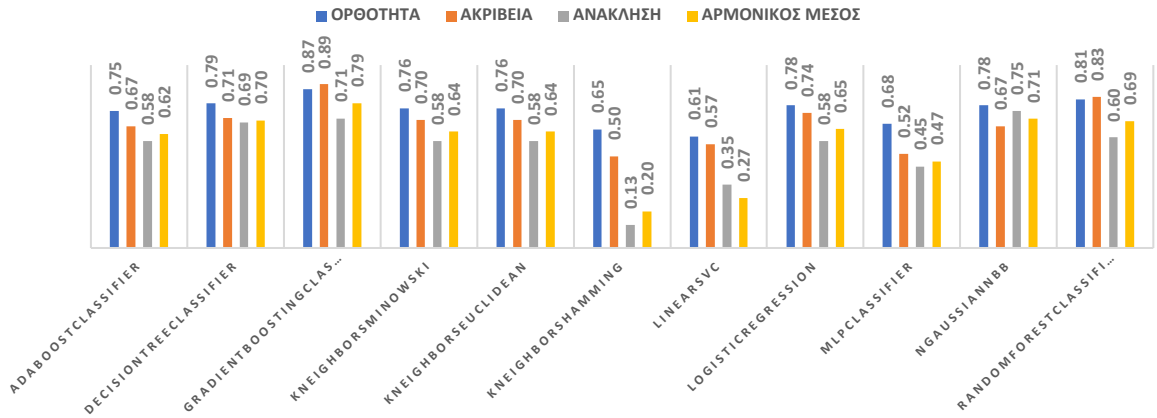
Εφαρμόζοντας τον αλγόριθμο StandardScaler στο αρχικό σύνολο δεδομένων, τα καλύτερα αποτελέσματα σχετικά με την ορθότητα, την ακρίβεια και τον αρμονικό μέσο δίνονται από τον απλοϊκό Bayes με τιμές 0.76, 0.63 και 0.63 αντίστοιχα. Το νευρωνικό δίκτυο παρέχει τα καλύτερα αποτελέσματα για την ανάκληση με 0.62.

4.2.3.2 Εφαρμογή StandardScaler και αντικατάσταση μηδενικής τιμής με μέση τιμή



Αφού εφαρμοστεί ο αλγόριθμος StandardScaler και στη συνέχεια γίνει η αντικατάσταση της μηδενικής τιμής με τη μέση τιμή, καλύτερα αποτελέσματα στην ορθότητα και στην ακρίβεια δίνει η λογιστική παλινδρόμηση με 0.79 και 0.70 αντίστοιχα. Όσο αφορά την ανάκληση και τον αρμονικό μέσο καλύτερα αποτελέσματα έχουμε από τον αλγόριθμο AdaBoost με τιμές 0.68 και 0.66 αντίστοιχα.

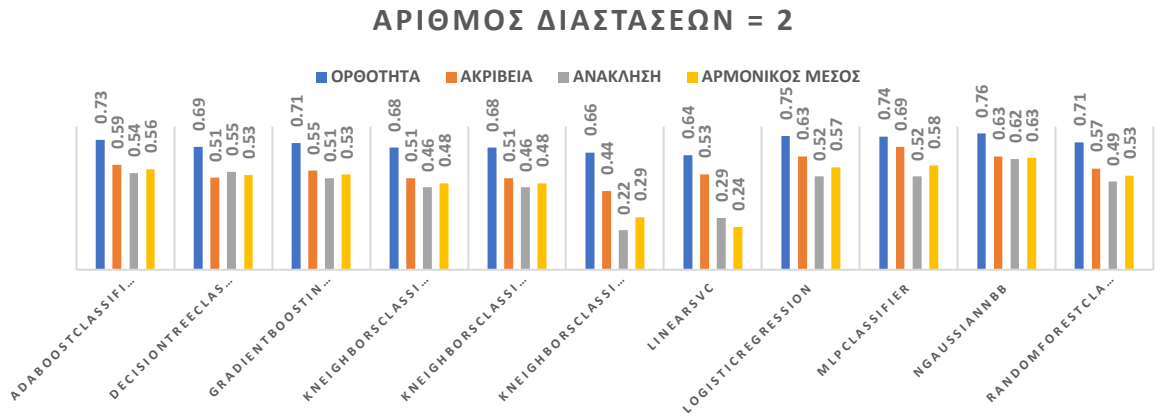
4.2.3.3 Εφαρμογή StandardScaler και διαγραφή εγγραφής με μια μηδενική τιμή

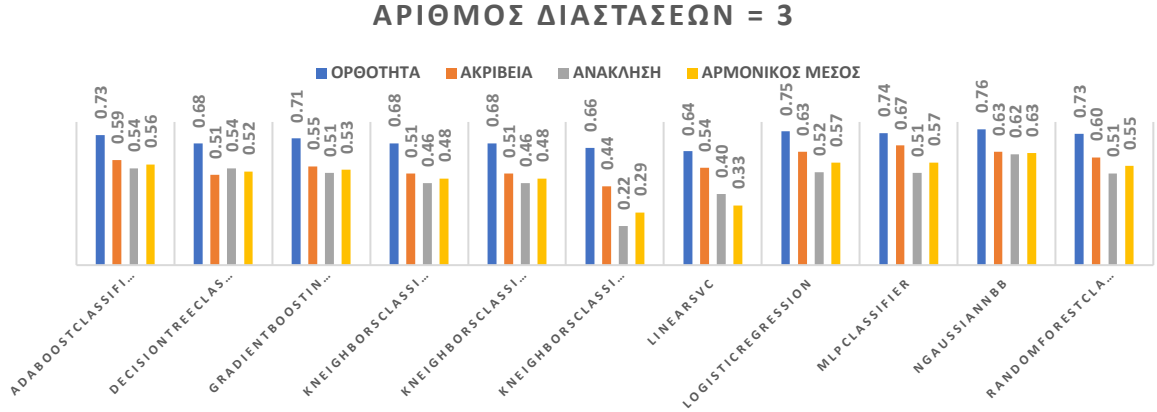


Εφαρμόζοντας τον αλγόριθμο StandardScaler και στη συνέχεια διαγράφοντας την εγγραφή η οποία περιέχει έστω και μια μηδενική τιμή, καλύτερα αποτελέσματα στην ορθότητα, στην ακρίβεια και στον αρμονικό μέσο έχουμε από τον GradientBoost με 0.87, 0.89 και 0.79 αντίστοιχα ενώ καλύτερο αποτέλεσμα στην ανάκληση έχει ο απλοϊκός Bayes με 0.75.

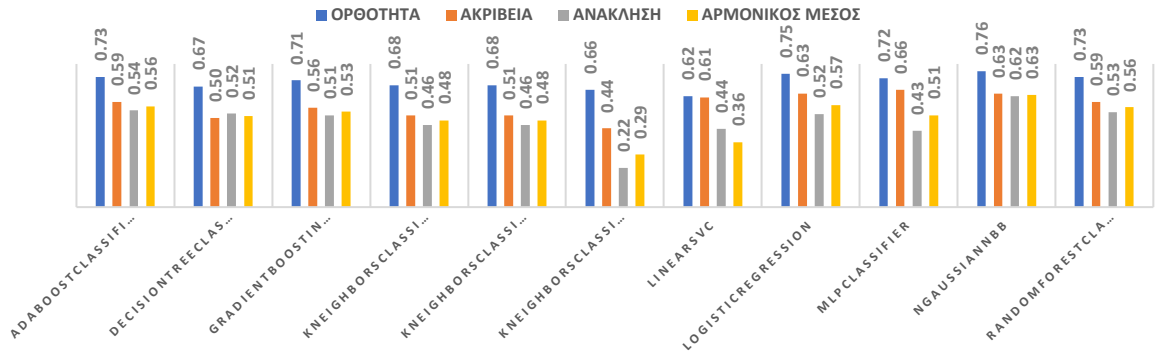
4.2.4 Εφαρμογή PCA

4.2.4.1 Εφαρμογή PCA στο αρχικό σύνολο δεδομένων

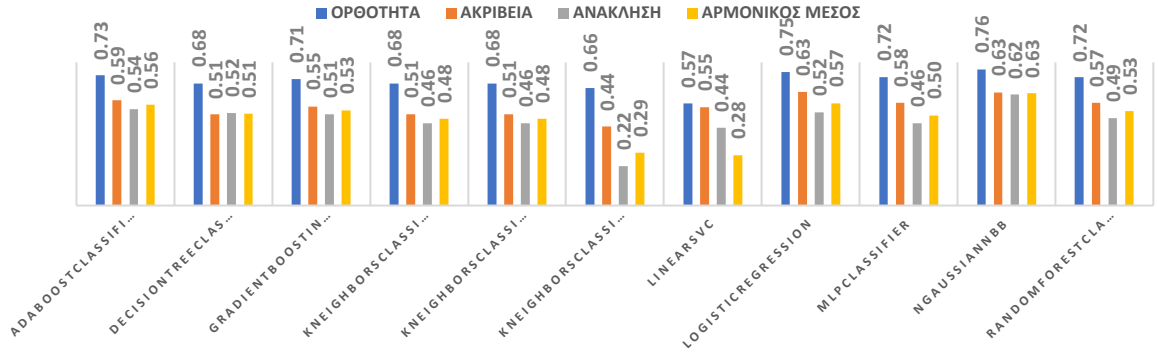




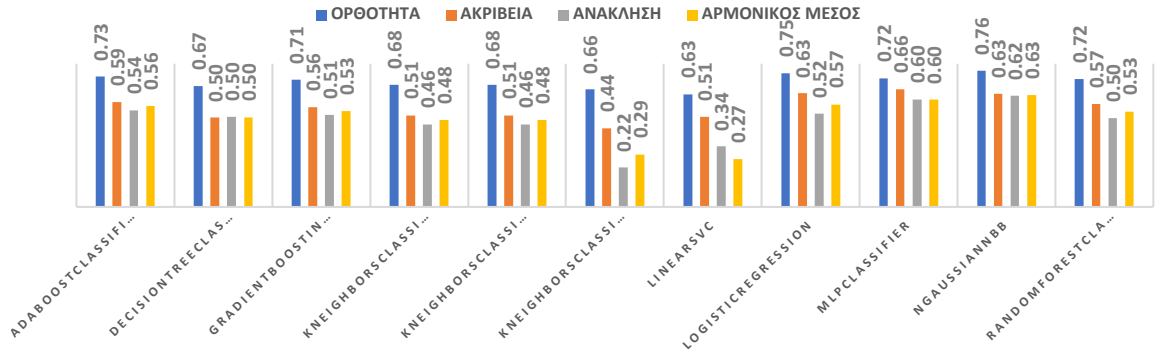
ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 4

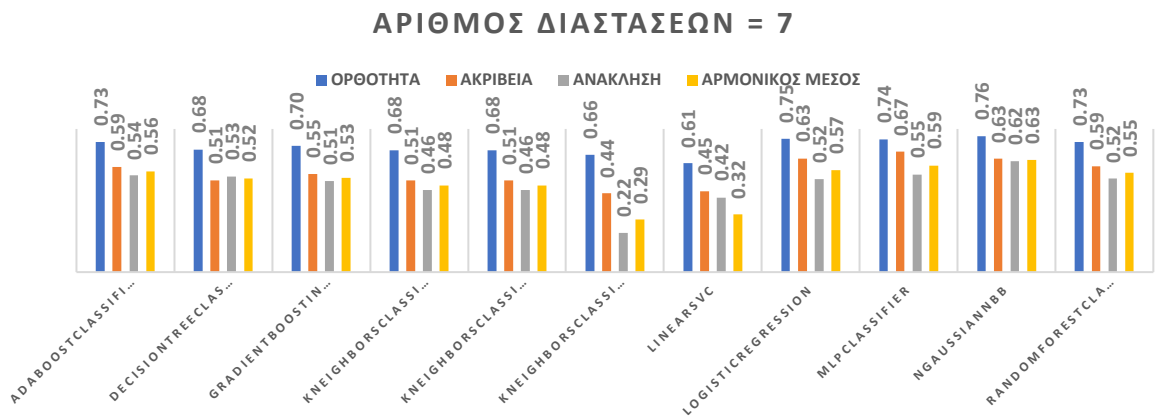


ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 5



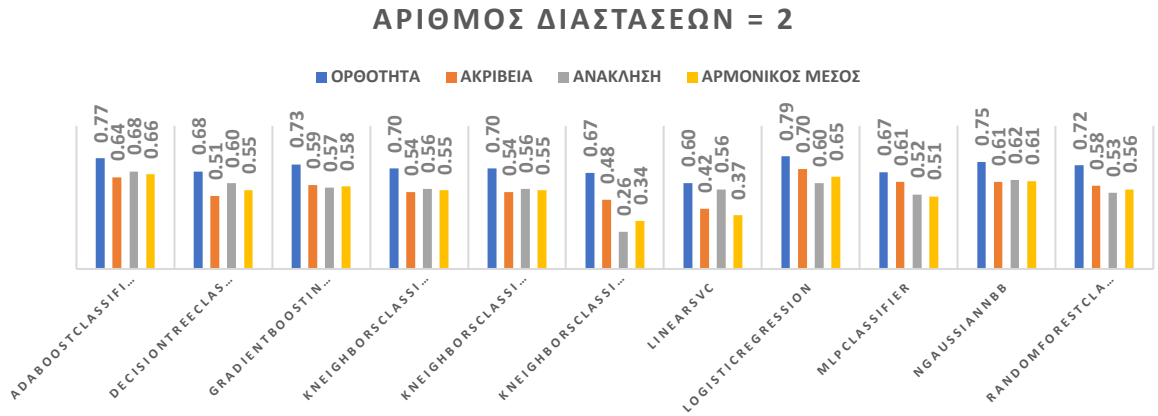
ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 6

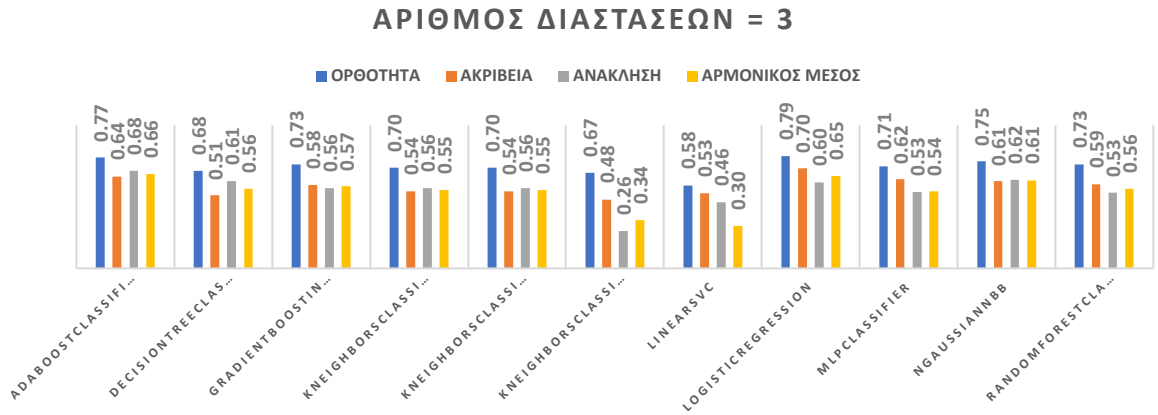


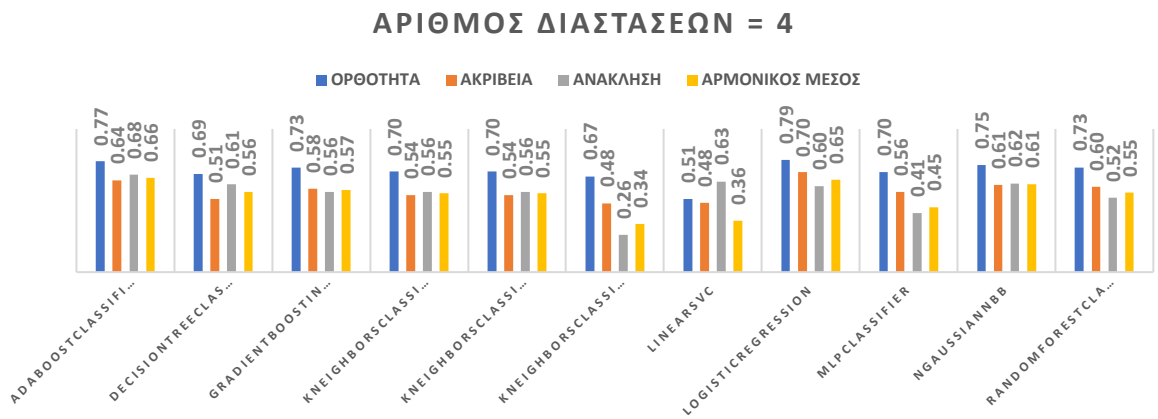


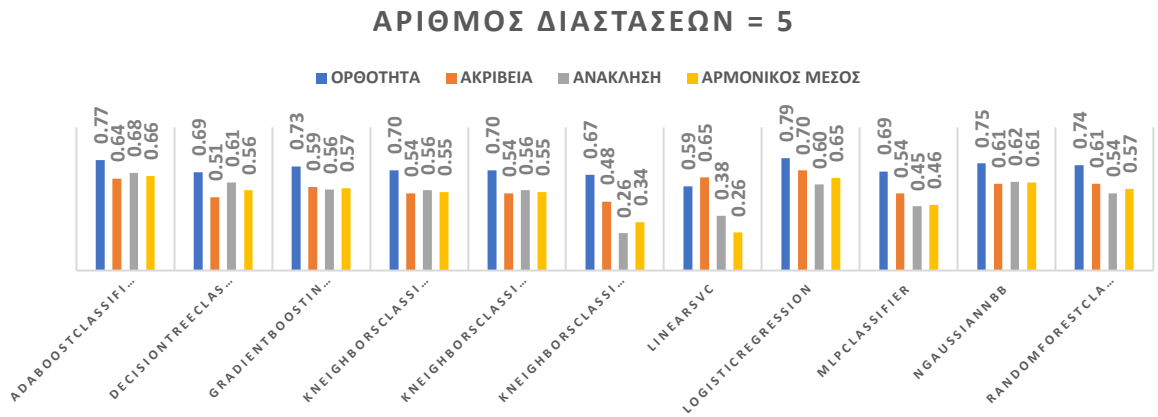
Εφαρμόζοντας τον αλγόριθμο PCA για τη μείωση των διαστάσεων στο αρχικό σύνολο δεδομένων καλύτερα αποτελέσματα έχει ο απλοϊκός Bayes στην ορθότητα, την ανάκληση και στον αρμονικό μέσο, με τιμές 0.76, 0.63 και 0.63 αντίστοιχα. Την καλύτερη τιμή στην ακρίβεια την έχει το νευρωνικό δίκτυο, όταν οι διαστάσεις του συνόλου δεδομένων είναι 2 με τιμή 0.69.

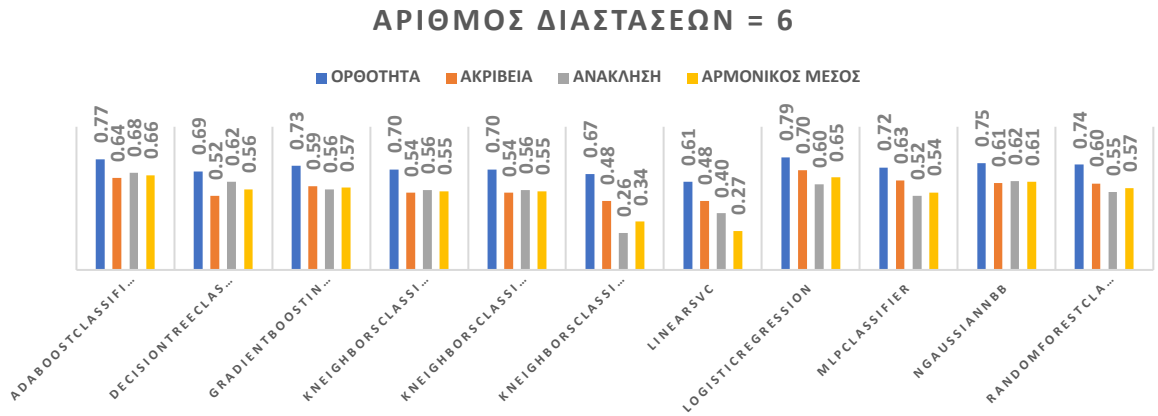
4.2.4.2 Εφαρμογή PCA και αντικατάσταση μηδενικής τιμής με μέση τιμή

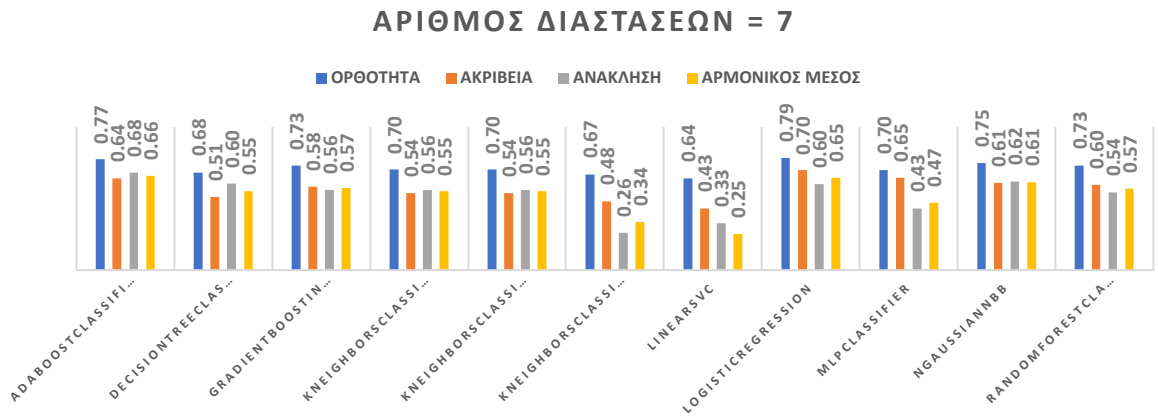





ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 4
ΟΡΘΟΤΗΤΑ
ΑΚΡΙΒΕΙΑ
ΑΝΑΚΛΗΣΗ
ΑΡΜΟΝΙΚΟΣ ΜΕΣΟΣ
0.77 0.64 0.68 0.66
0.69 0.51 0.61 0.56
0.73 0.58 0.56 0.57
0.70 0.54 0.56 0.55
0.70 0.54 0.56 0.55
0.67 0.48 0.26 0.34
0.51 0.48 0.63 0.36
0.79 0.70 0.60 0.65
0.70 0.56 0.41 0.45
0.75 0.61 0.62 0.61
0.73 0.60 0.52 0.55
ADABOOSTCLASSIFI...
DECISIONTREECLAS...
GRADIENTBOOSTIN...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
LINEARSVC
LOGISTICREGRESSION
MLPCLASSIFIER
NGAUSSIANNBB
RANDOMFORESTCLA...

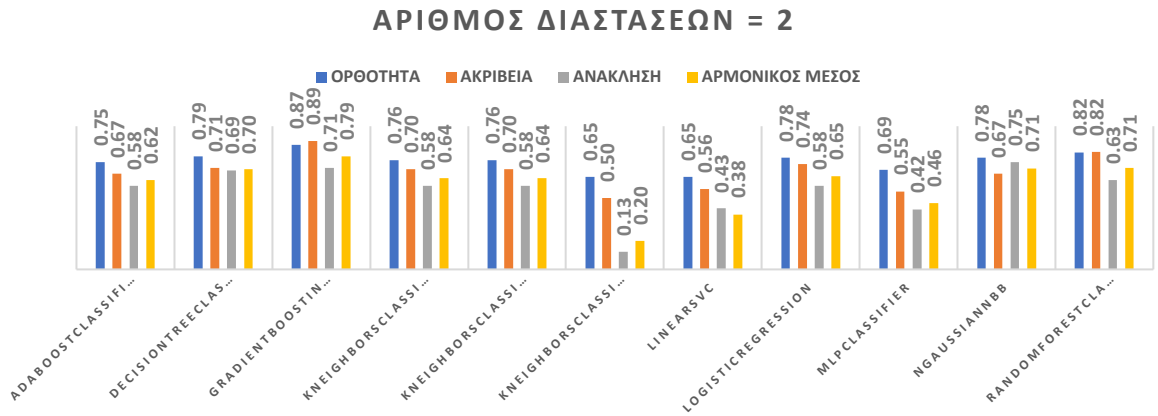

ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 5
ΟΡΘΟΤΗΤΑ
ΑΚΡΙΒΕΙΑ
ΑΝΑΚΛΗΣΗ
ΑΡΜΟΝΙΚΟΣ ΜΕΣΟΣ
0.77 0.64 0.68 0.66
0.69 0.51 0.61 0.56
0.73 0.59 0.56 0.57
0.70 0.54 0.56 0.55
0.70 0.54 0.56 0.55
0.67 0.48 0.26 0.34
0.59 0.65 0.38 0.26
0.79 0.70 0.60 0.65
0.69 0.54 0.45 0.46
0.75 0.61 0.62 0.61
0.74 0.61 0.54 0.57
ADABOOSTCLASSIFI...
DECISIONTREECLAS...
GRADIENTBOOSTIN...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
LINEARSVC
LOGISTICREGRESSION
MLPCLASSIFIER
NGAUSSIANNBB
RANDOMFORESTCLA...

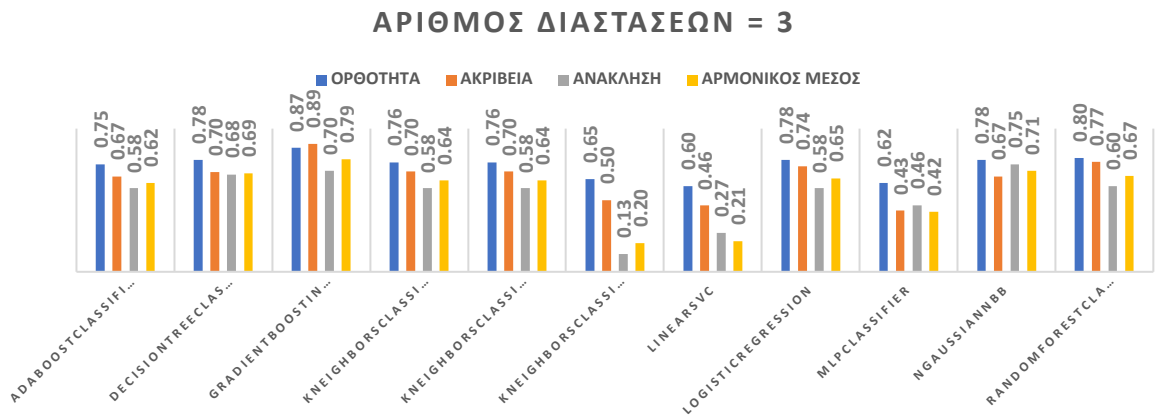

ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 6
ΟΡΘΟΤΗΤΑ
ΑΚΡΙΒΕΙΑ
ΑΝΑΚΛΗΣΗ
ΑΡΜΟΝΙΚΟΣ ΜΕΣΟΣ
0.77 0.64 0.68 0.66
0.69 0.52 0.62 0.56
0.73 0.59 0.56 0.57
0.70 0.54 0.56 0.55
0.70 0.54 0.56 0.55
0.67 0.48 0.26 0.34
0.61 0.48 0.40 0.27
0.79 0.70 0.60 0.65
0.72 0.63 0.52 0.54
0.75 0.61 0.62 0.61
0.74 0.60 0.55 0.57
ADABOOSTCLASSIFI...
DECISIONTREECLAS...
GRADIENTBOOSTIN...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
LINEARSVC
LOGISTICREGRESSION
MLPCLASSIFIER
NGAUSSIANNBB
RANDOMFORESTCLA...



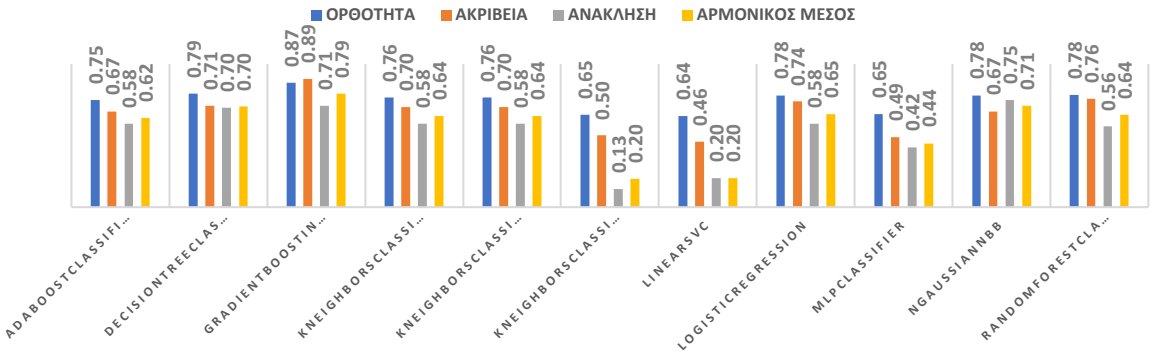
Εφαρμόζοντας τον αλγόριθμο PCA για τη μείωση των διαστάσεων στο αρχικό σύνολο δεδομένων και στη συνέχεια αντικατάσταση των μηδενικών τιμών με τη μέση τιμή, καλύτερα αποτελέσματα για όλες τις διαστάσεις στην ορθότητα και την ακρίβεια τα έχει η λογιστική παλινδρόμηση με τιμές 0.79 και 0.70 αντίστοιχα. Τα καλύτερα αποτελέσματα στην ανάκληση και τον αρμονικό μέσο τα έχει ο AdaBoost σε όλες τις διαστάσεις με τιμές 0.68 και 0.66 αντίστοιχα.

4.2.4.3 Εφαρμογή PCA και διαγραφή εγγραφής με μηδενική τιμή

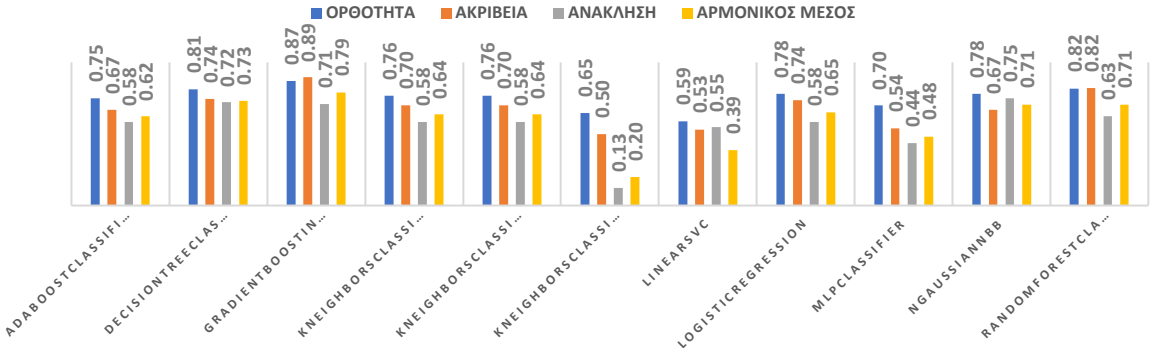




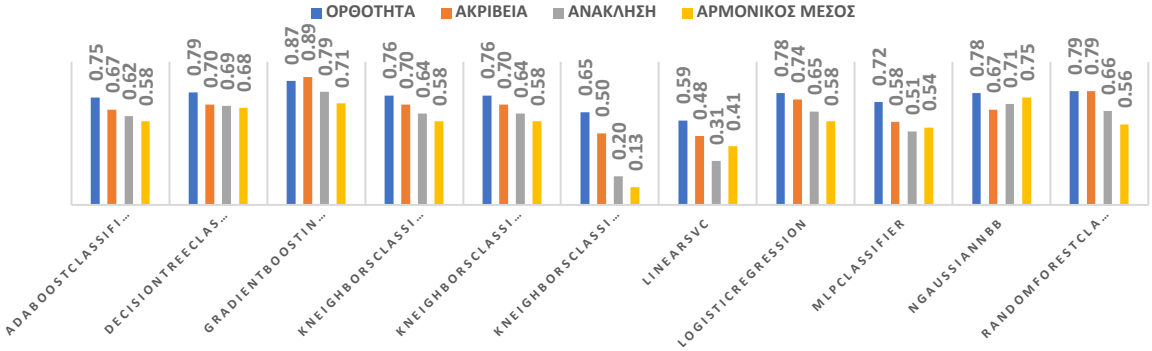
ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 4

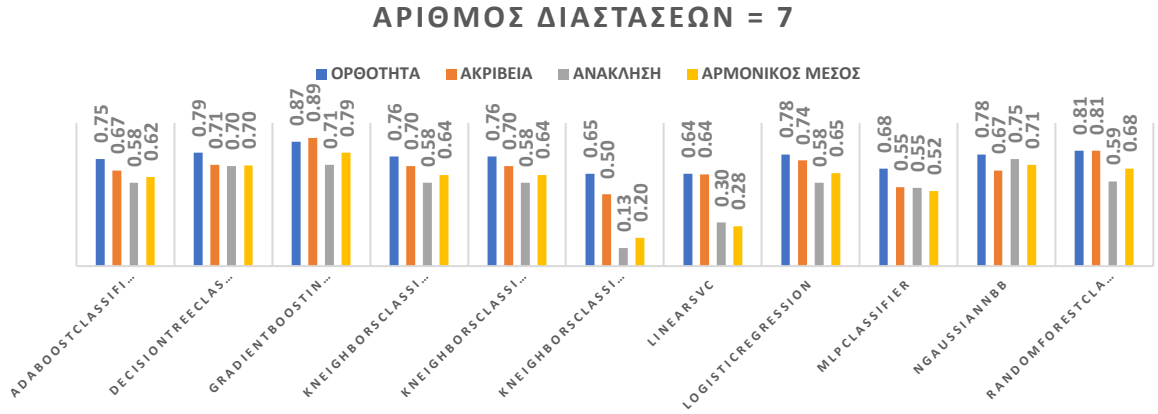


ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 5



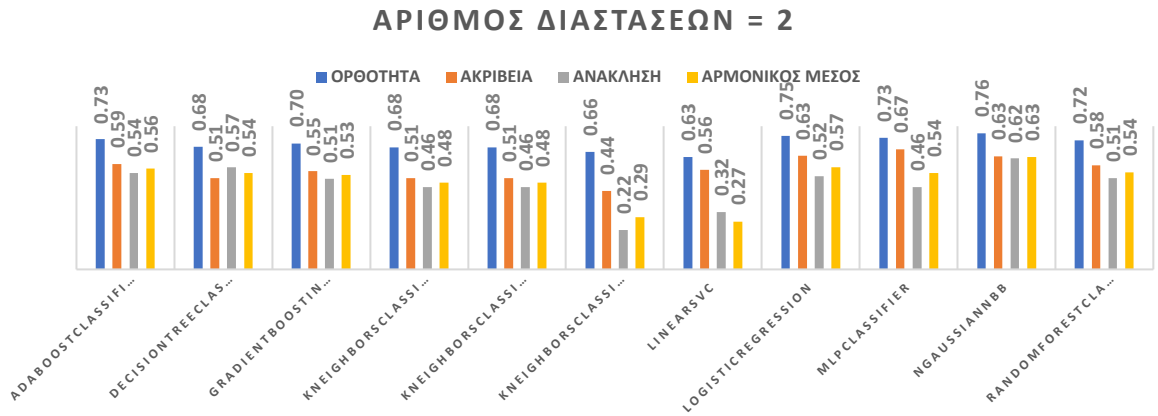
ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 6

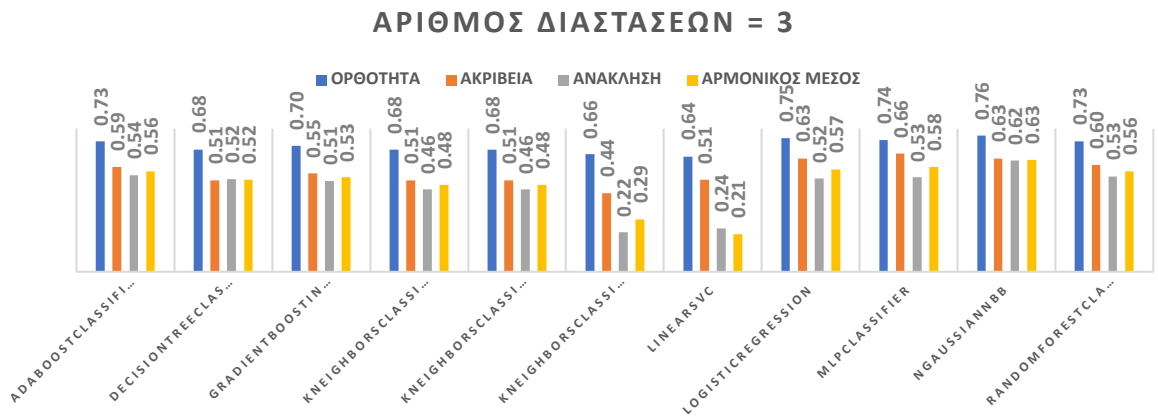




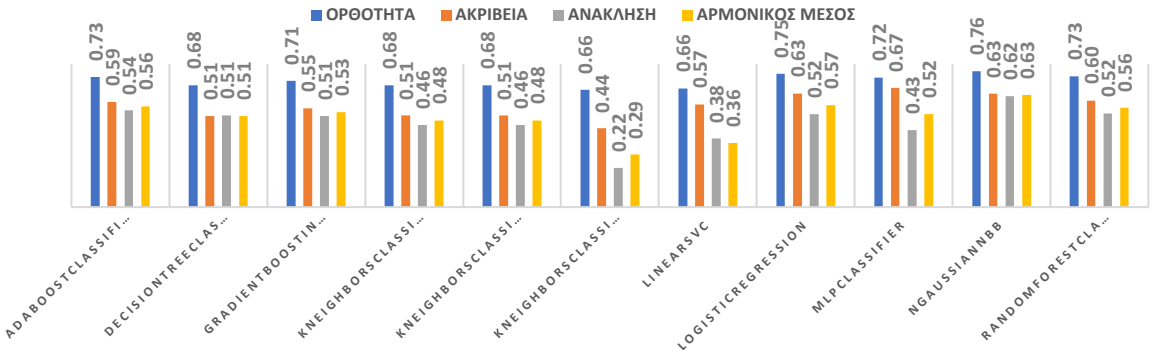
Εφαρμόζοντας τον αλγόριθμο PCA για τη μείωση των διαστάσεων στο αρχικό σύνολο δεδομένων και στη συνέχεια διαγραφή των εγγραφών με μια μηδενική τιμή, καλύτερα αποτελέσματα για όλες τις διαστάσεις στην ορθότητα, στην ακρίβεια και στην ανάκληση τα έχει ο GradientBoosting με τιμές 0.87, 0.89, 0.71 αντίστοιχα. Όσο αφορά τα αποτελέσματα του αρμονικού μέσου, καλύτερα αποτέλεσμα έχει και πάλι ο GradientBoosting αλγόριθμος, αλλά μόνο όταν το σύνολο δεδομένων έχει διαστάσεις 2, 3, 4 και 7 με αποτέλεσμα 0.79.

4.2.4.4 Εφαρμογή PCA και MinMaxScaler στο αρχικό σύνολο δεδομένων

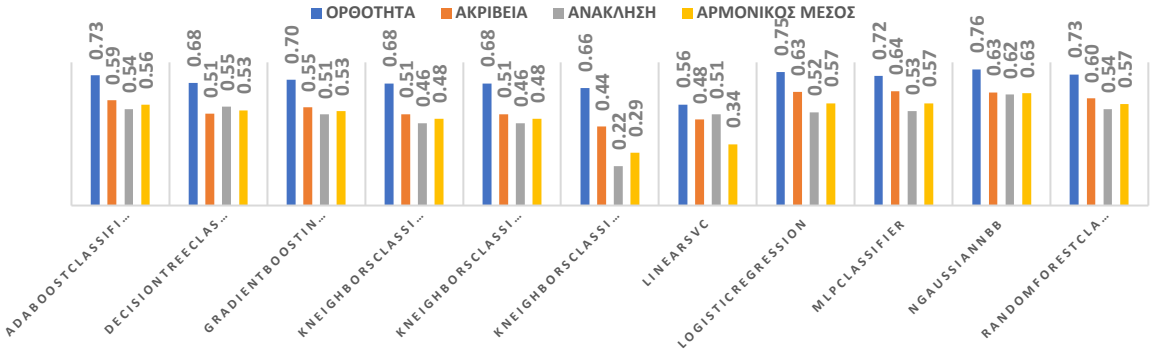




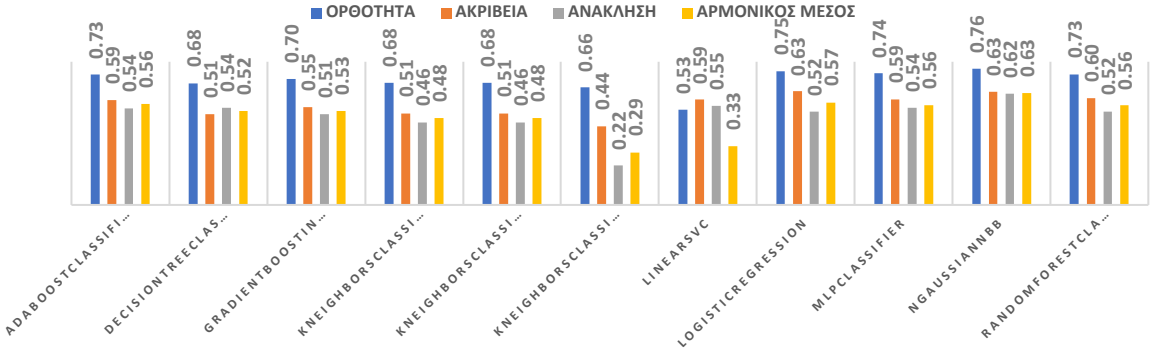
ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 4

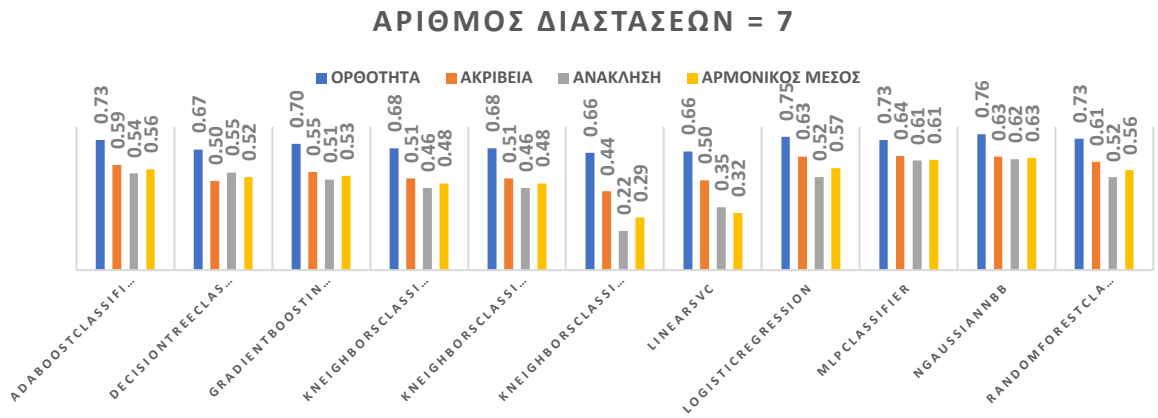


ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 5



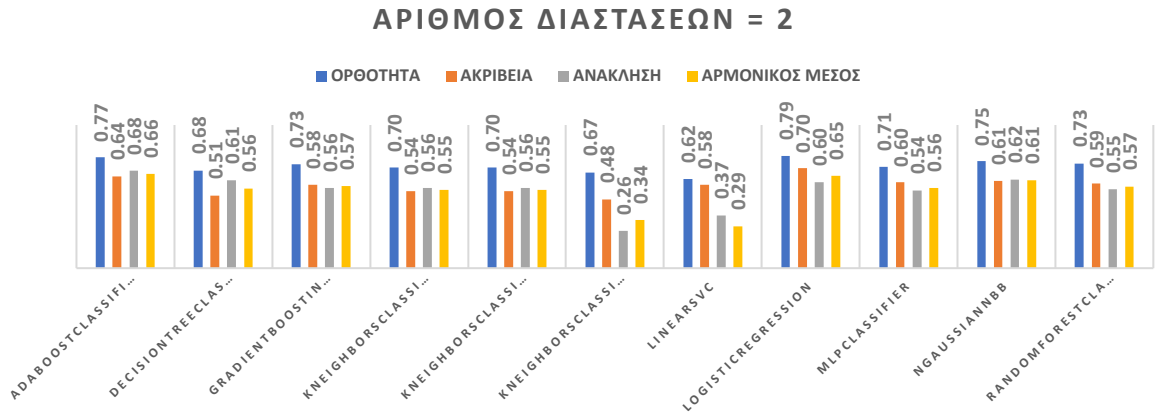
ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 6

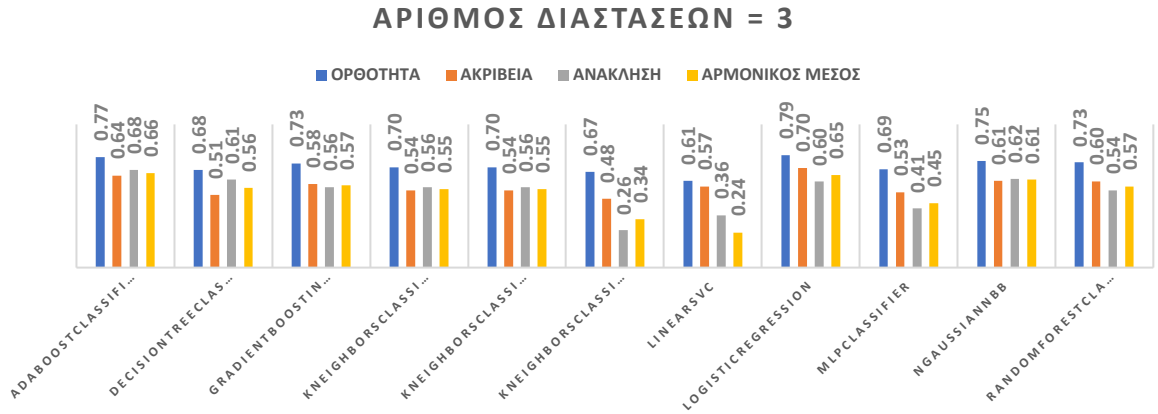


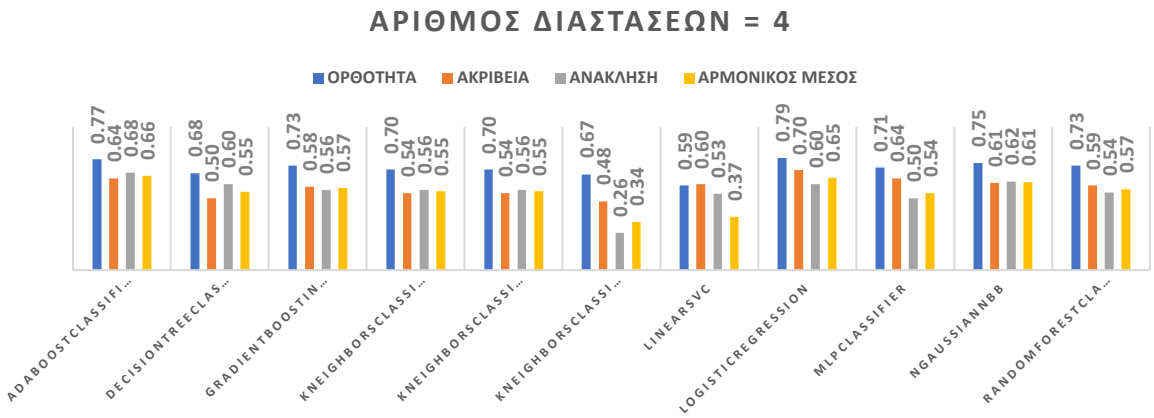


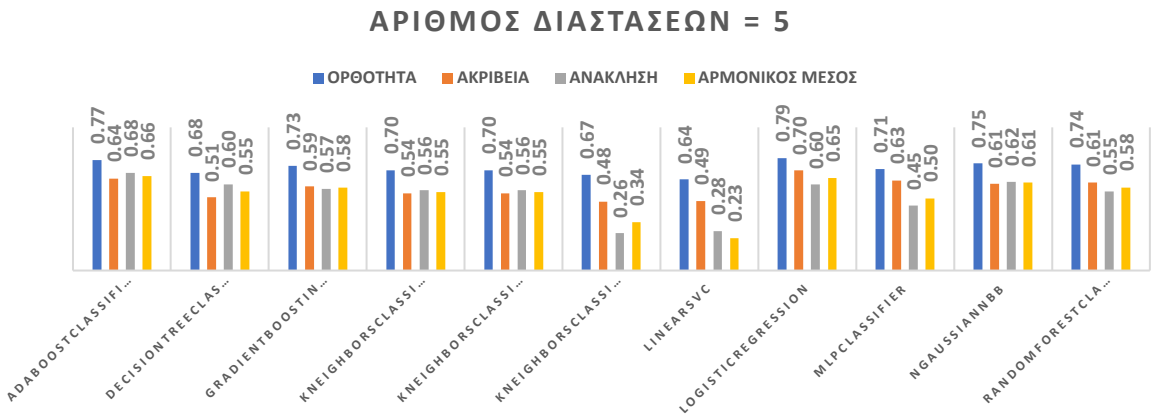
Εφαρμόζοντας τον αλγόριθμο PCA και στη συνέχεια MinMaxScaler, καλύτερα αποτελέσματα για όλες τις διαστάσεις στην ορθότητα, στην ανάκληση και στον αρμονικό μέσο τα έχει ο απλοϊκός Bayes με τιμές 0.76, 0.63 και 0.63 αντίστοιχα. Καλύτερα αποτελέσματα για την ακρίβεια τα έχει το νευρωνικό δίκτυο, όταν οι διαστάσεις του συνόλου είναι 2 με αποτέλεσμα 0.67.

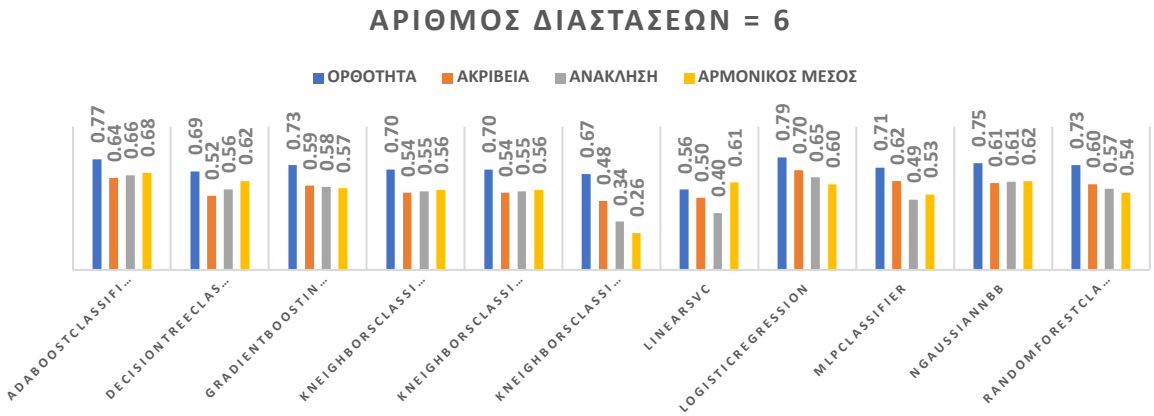
4.2.4.5 Εφαρμογή PCA, MinMaxScaler και αντικατάσταση μηδενικής τιμής με μέση τιμή

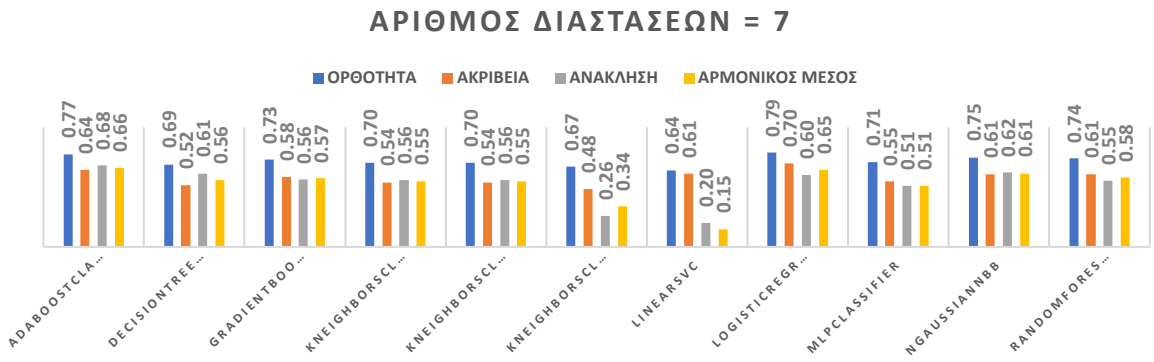





ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 4
ΟΡΘΟΤΗΤΑ
ΑΚΡΙΒΕΙΑ
ΑΝΑΚΛΗΣΗ
ΑΡΜΟΝΙΚΟΣ ΜΕΣΟΣ
0.77 0.64 0.68 0.66
0.68 0.50 0.60 0.55
0.73 0.58 0.56 0.57
0.70 0.54 0.56 0.55
0.70 0.54 0.56 0.55
0.67 0.48 0.26 0.34
0.59 0.60 0.53 0.37
0.79 0.70 0.60 0.65
0.71 0.64 0.50 0.54
0.75 0.61 0.62 0.61
0.73 0.59 0.54 0.57
ADABOOSTCLASSIFI...
DECISIONTREECLAS...
GRADIENTBOOSTIN...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
LINEARSVC
LOGISTICREGRESSION
MLPCLASSIFIER
NGAUSSIANNBB
RANDOMFORESTCLA...

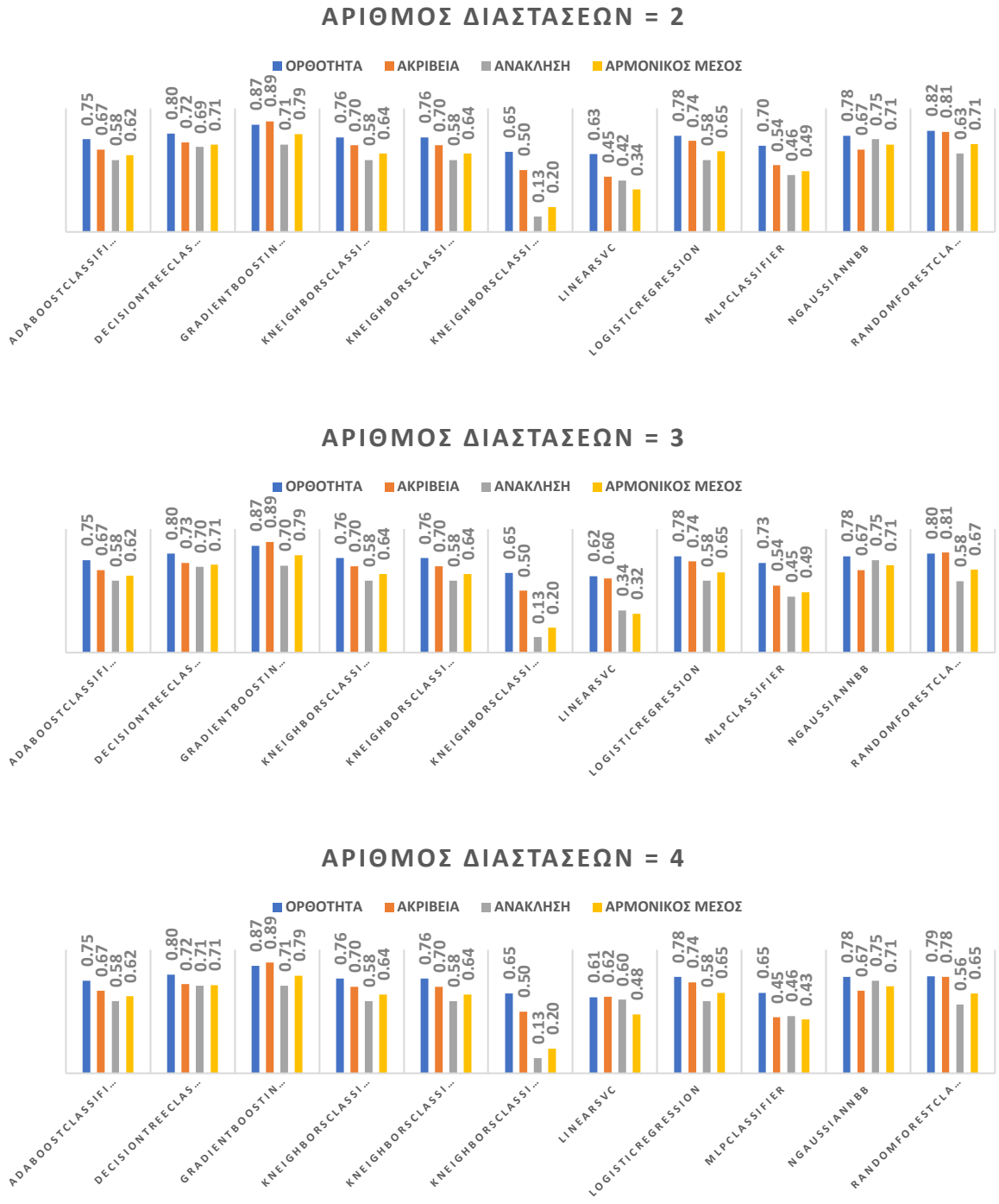

ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 5
ΟΡΘΟΤΗΤΑ
ΑΚΡΙΒΕΙΑ
ΑΝΑΚΛΗΣΗ
ΑΡΜΟΝΙΚΟΣ ΜΕΣΟΣ
0.77 0.64 0.68 0.66
0.68 0.51 0.60 0.55
0.73 0.59 0.57 0.58
0.70 0.54 0.56 0.55
0.70 0.54 0.56 0.55
0.67 0.48 0.26 0.34
0.64 0.49 0.28 0.23
0.79 0.70 0.60 0.65
0.71 0.63 0.45 0.50
0.75 0.61 0.62 0.61
0.74 0.61 0.55 0.58
ADABOOSTCLASSIFI...
DECISIONTREECLAS...
GRADIENTBOOSTIN...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
LINEARSVC
LOGISTICREGRESSION
MLPCLASSIFIER
NGAUSSIANNBB
RANDOMFORESTCLA...

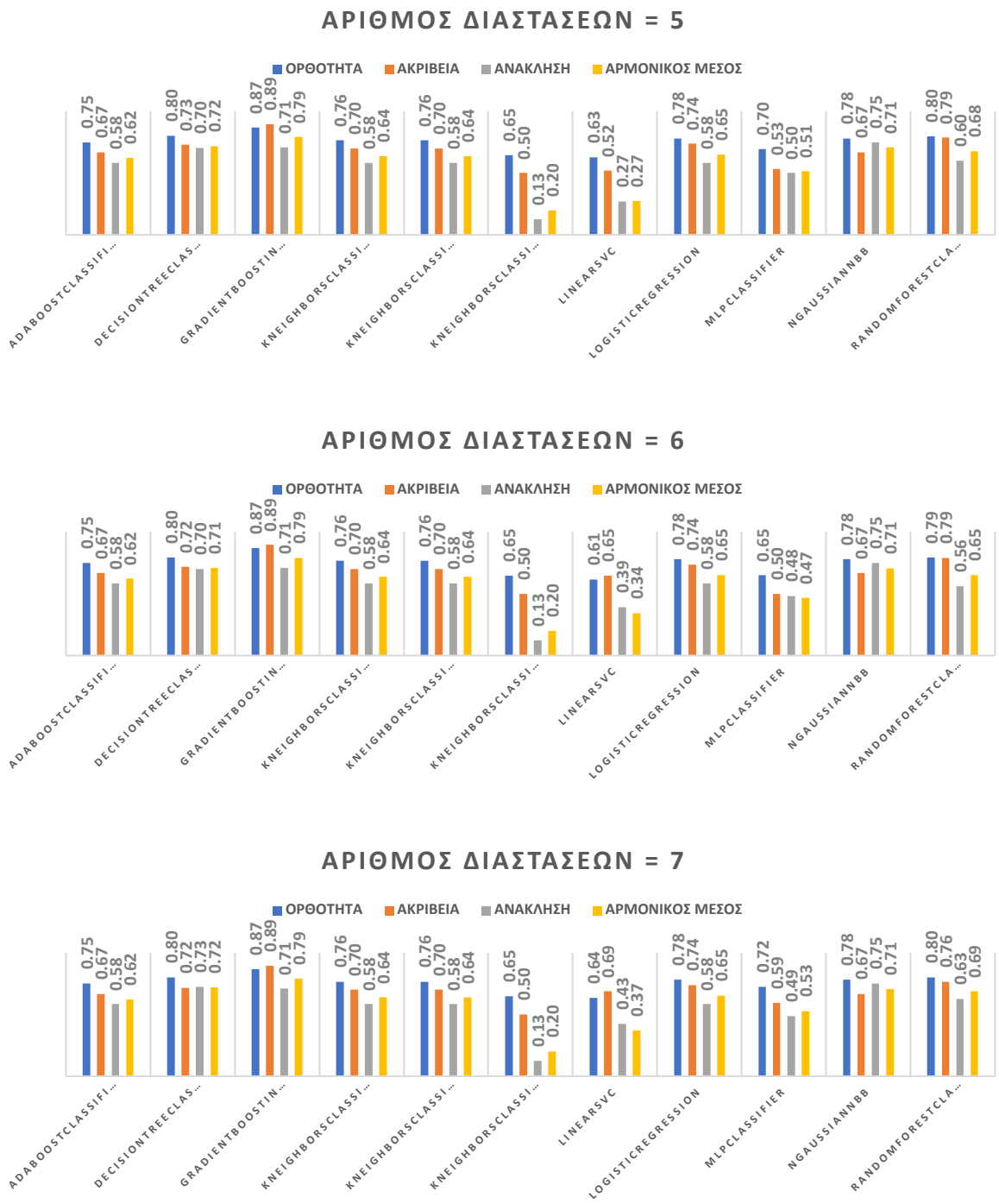

ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 6
ΟΡΘΟΤΗΤΑ
ΑΚΡΙΒΕΙΑ
ΑΝΑΚΛΗΣΗ
ΑΡΜΟΝΙΚΟΣ ΜΕΣΟΣ
0.77 0.64 0.66 0.68
0.69 0.52 0.56 0.62
0.73 0.59 0.58 0.57
0.70 0.54 0.55 0.56
0.70 0.54 0.55 0.56
0.67 0.48 0.34 0.26
0.56 0.50 0.40 0.61
0.79 0.70 0.65 0.60
0.71 0.62 0.49 0.53
0.75 0.61 0.61 0.62
0.73 0.60 0.57 0.54
ADABOOSTCLASSIFI...
DECISIONTREECLAS...
GRADIENTBOOSTIN...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
LINEARSVC
LOGISTICREGRESSION
MLPCLASSIFIER
NGAUSSIANNBB
RANDOMFORESTCLA...


ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 7
ΟΡΘΟΤΗΤΑ
ΑΚΡΙΒΕΙΑ
ΑΝΑΚΛΗΣΗ
ΑΡΜΟΝΙΚΟΣ ΜΕΣΟΣ
0.77 0.64 0.68 0.66
0.69 0.52 0.61 0.56
0.73 0.58 0.56 0.57
0.70 0.54 0.56 0.55
0.70 0.54 0.56 0.55
0.67 0.48 0.26 0.34
0.64 0.61 0.20 0.15
0.79 0.70 0.60 0.65
0.71 0.55 0.51 0.51
0.75 0.61 0.62 0.61
0.74 0.61 0.55 0.58
ADABOOSTCLA...
DECISIONTREE...
GRADIENTBOO...
KNEIGHBORSCL...
KNEIGHBORSCL...
KNEIGHBORSCL...
LINEARSVC
LOGISTICREGR...
MLPCLASSIFIER
NGAUSSIANNBB
RANDOMFORES...

Εφαρμόζοντας τον αλγόριθμο PCA, στη συνέχεια MinMaxScaler και τέλος αντικαταστώντας τις μηδενικές τιμές με τη μέση τιμή, καλύτερα αποτελέσματα για όλες τις διαστάσεις στην ορθότητα και στην ακρίβεια τα έχει η λογιστική παλινδρόμηση με τιμές 0.79 και 0.7 αντίστοιχα. Καλύτερα αποτελέσματα σε όλες τις διαστάσεις για την ανάκληση τα έχει ο AdaBoost με 0.68. Τέλος, καλύτερα αποτελέσματα για τον αρμονικό μέσο τα έχει πάλι ο AdaBoost μόνο όταν οι διαστάσεις του συνόλου δεδομένων είναι 5 με αποτέλεσμα 0.68.

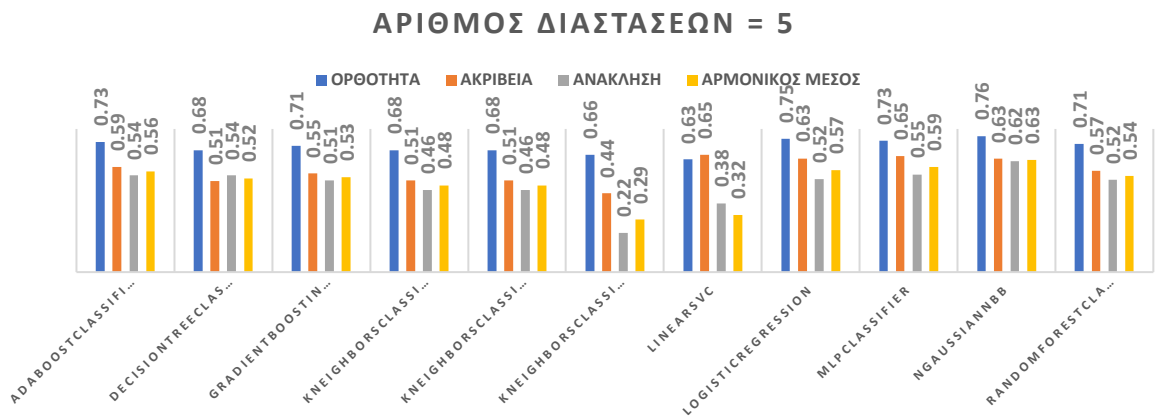
4.2.4.6 Εφαρμογή PCA, MinMaxScaler και διαγραφή εγγραφής με μηδενική τιμή

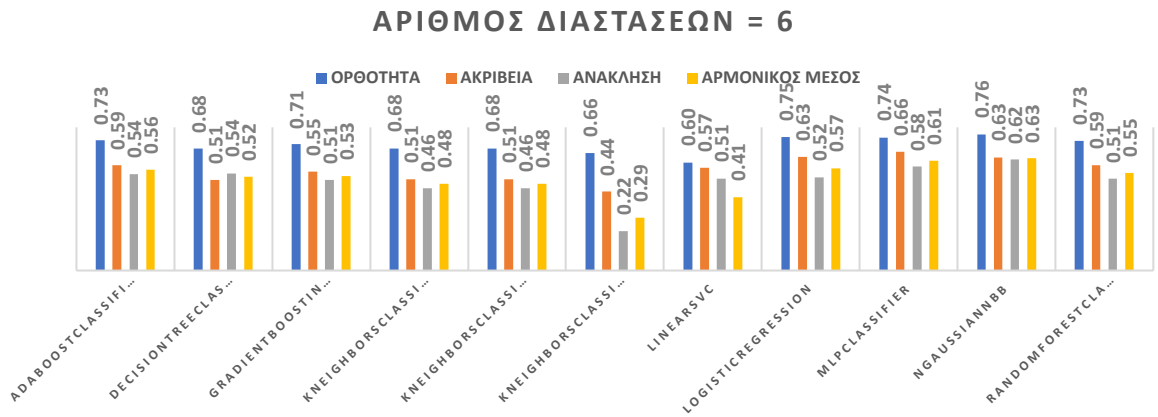


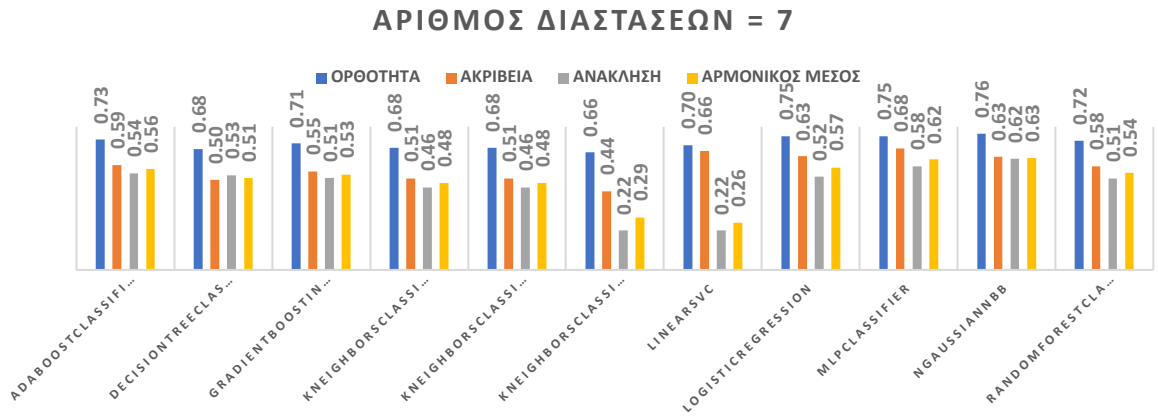


Εφαρμόζοντας τον αλγόριθμο PCA, MinMaxScaler και τέλος διαγράφοντας τις εγγραφές με έστω μια μηδενική τιμή, καλύτερα αποτελέσματα για όλες τις διαστάσεις στην ορθότητα, στην ακρίβεια και του αρμονικού μέσου τα έχει ο GradientBoost με τιμές 0.87, 0.89 και 0.79 αντίστοιχα. Καλύτερα αποτελέσματα σε όλες τις διαστάσεις για την ανάκληση τα έχει ο απλοϊκός Bayes με 0.75.

4.2.4.7 Εφαρμογή PCA και StandardScaler στο αρχικό σύνολο δεδομένων

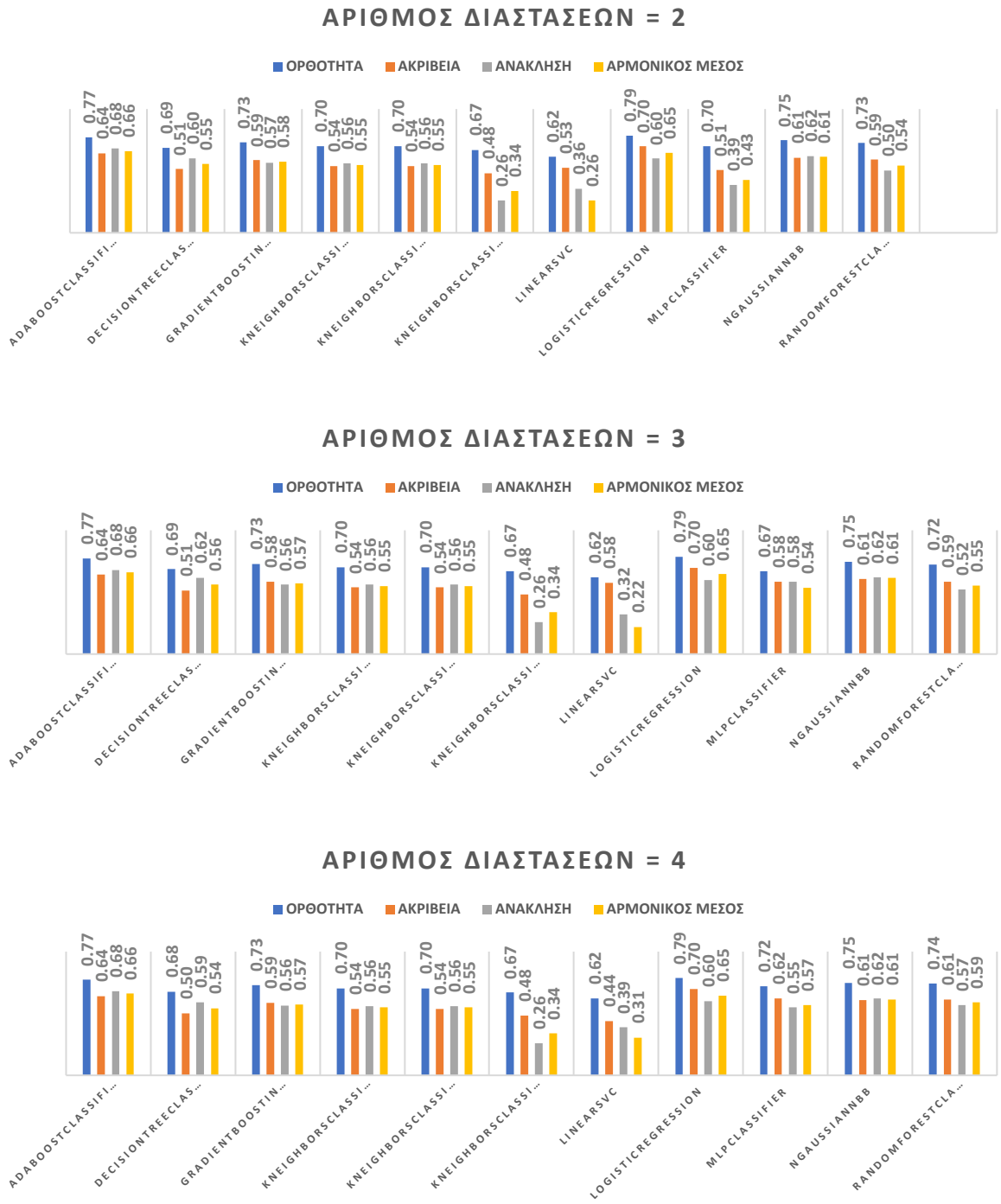


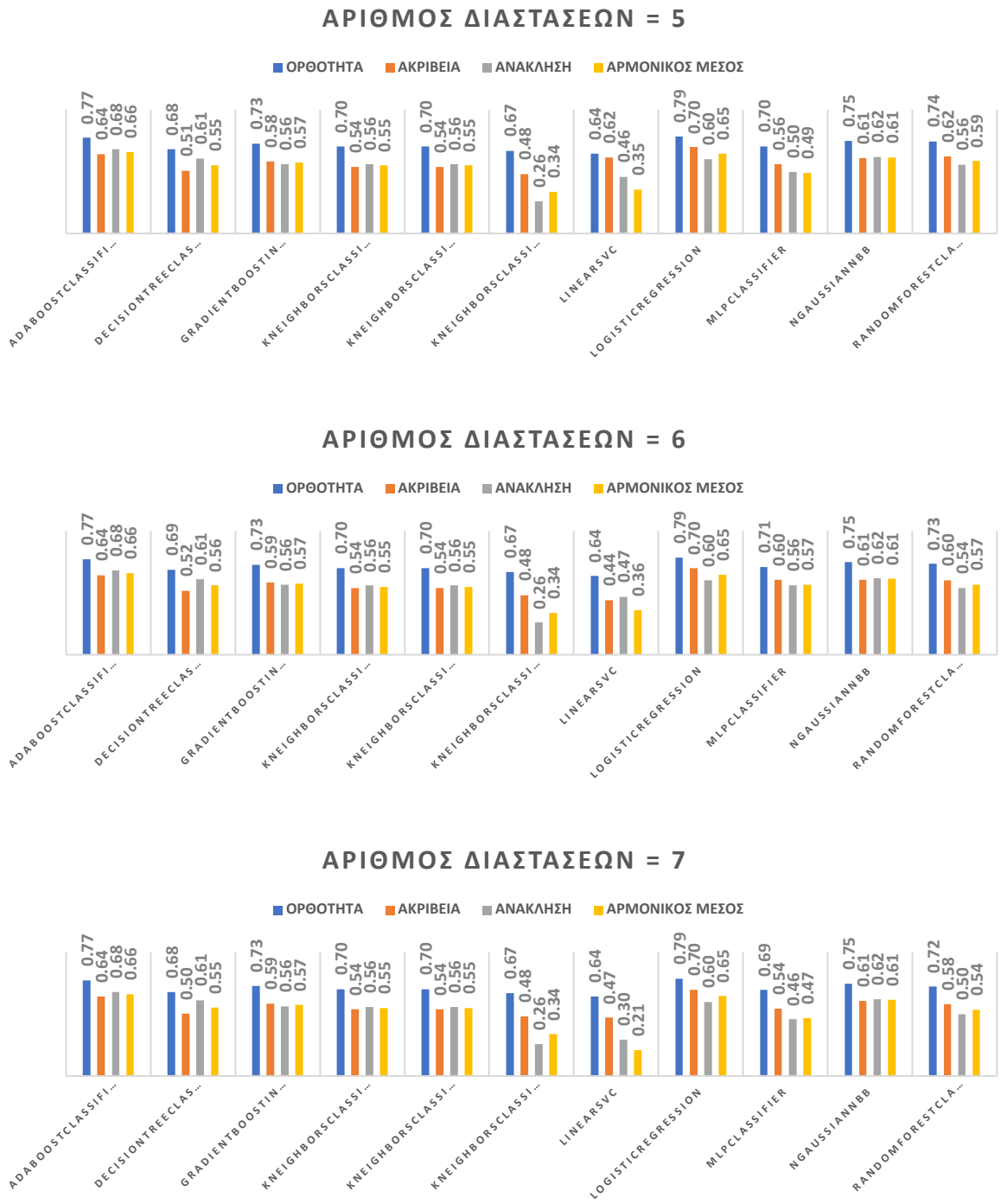




Εφαρμόζοντας τον αλγόριθμο PCA και στη συνέχεια StandardScaler, καλύτερα αποτελέσματα για όλες τις διαστάσεις στην ορθότητα, στην ανάκληση και στον αρμονικό μέσο τα έχει ο απλοϊκός Bayes με τιμές 0.76, 0.63 και 0.63 αντίστοιχα. Καλύτερα αποτελέσματα για την ακρίβεια τα έχει το νευρωνικό δίκτυο, όταν οι διαστάσεις του συνόλου είναι 2 με αποτέλεσμα 0.65.

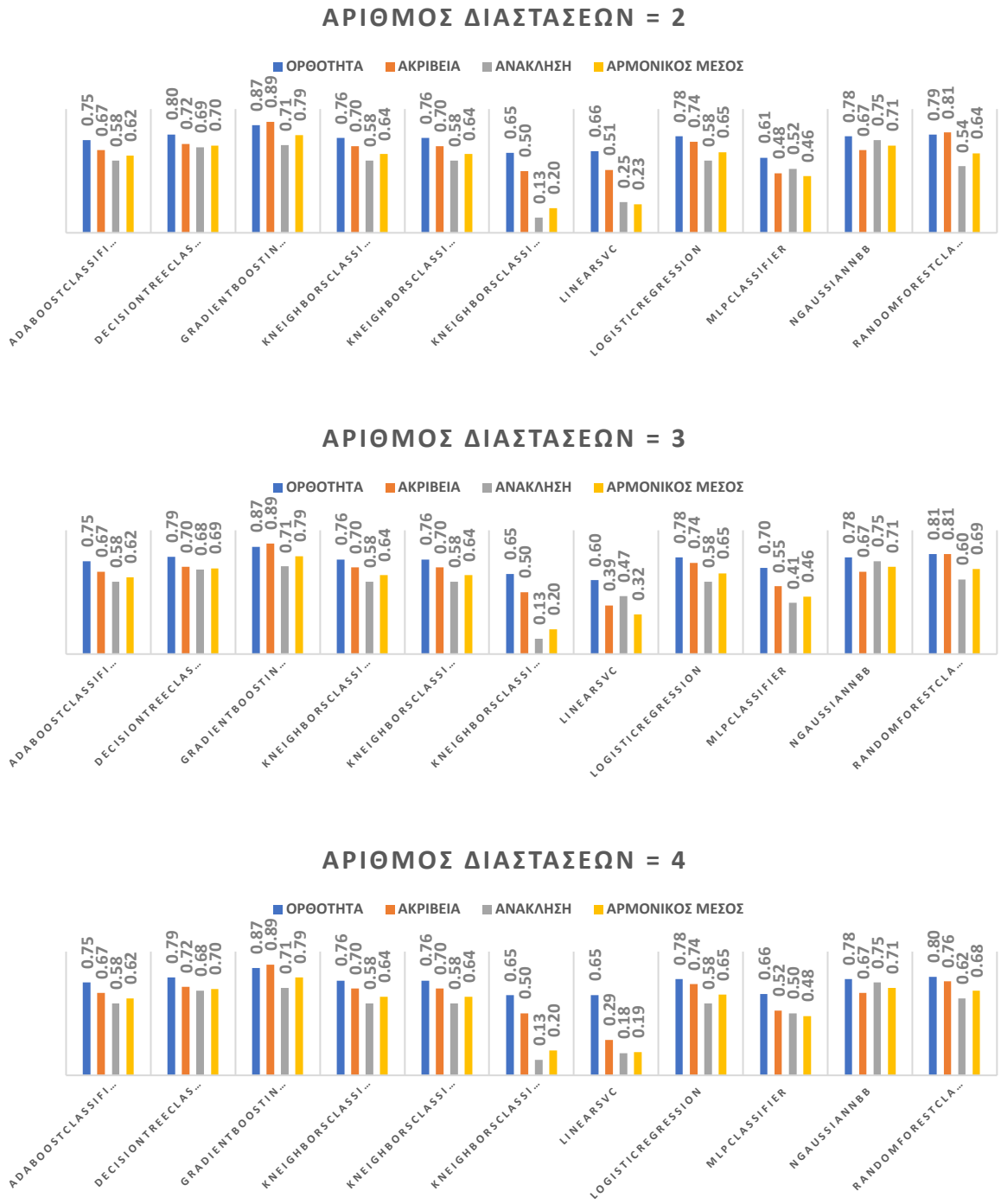
4.2.4.8 Εφαρμογή PCA, StandardScaler και αντικατάσταση μηδενικής τιμής με μέση τιμή

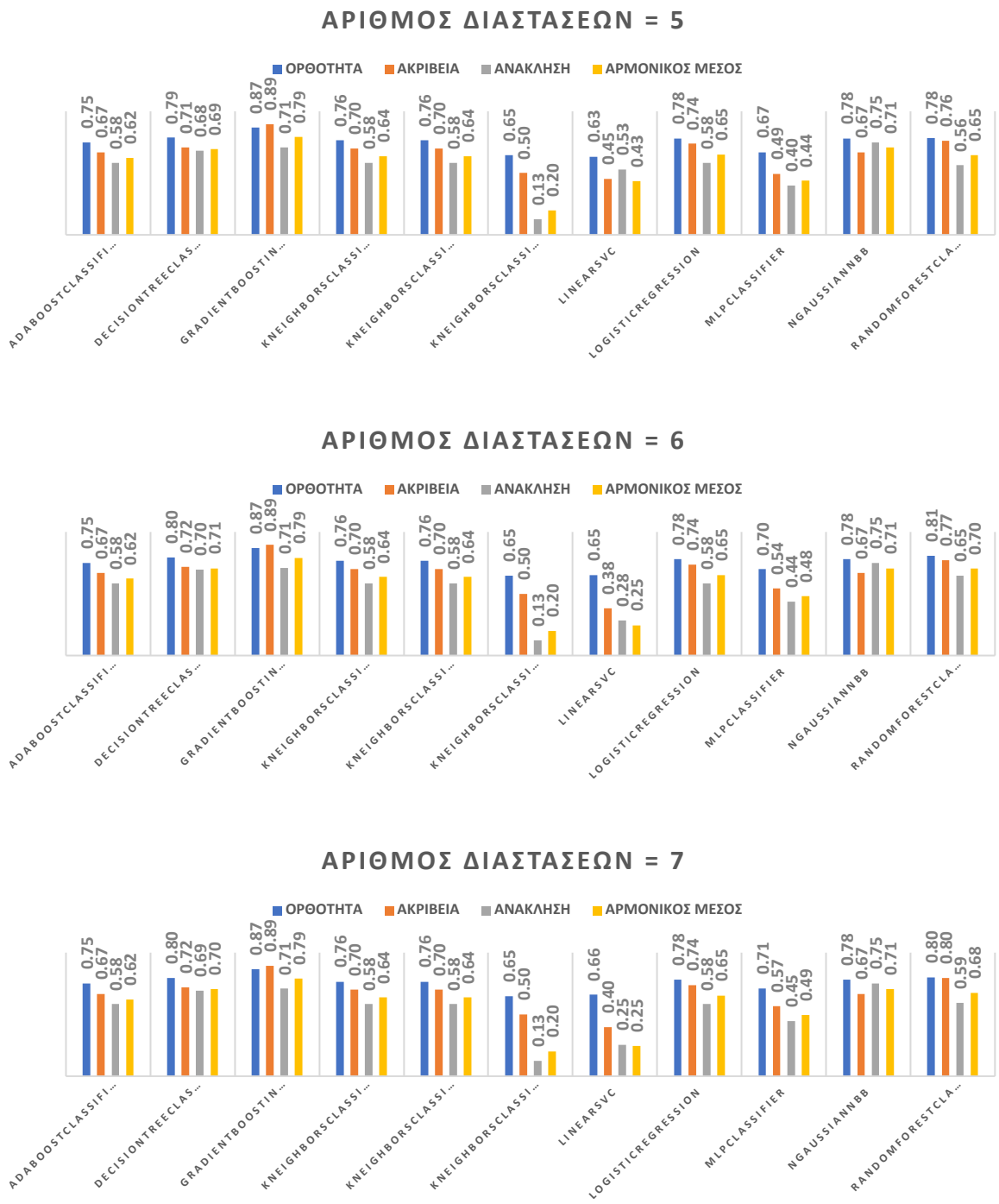




Εφαρμόζοντας τον αλγόριθμο PCA, στη συνέχεια StandardScaler και τέλος αντικαταστώντας τις μηδενικές τιμές με τη μέση τιμή, καλύτερα αποτελέσματα για όλες τις διαστάσεις στην ορθότητα και στην ακρίβεια τα έχει η λογιστική παλινδρόμηση με τιμές 0.79 και 0.7 αντίστοιχα. Καλύτερα αποτελέσματα σε όλες τις διαστάσεις για την ανάκληση και τον αρμονικό μέσο τα έχει ο AdaBoost με 0.68 και 0.66 αντίστοιχα.

4.2.4.9 Εφαρμογή PCA, StandardScaler και διαγραφή εγγραφής με μηδενική τιμή





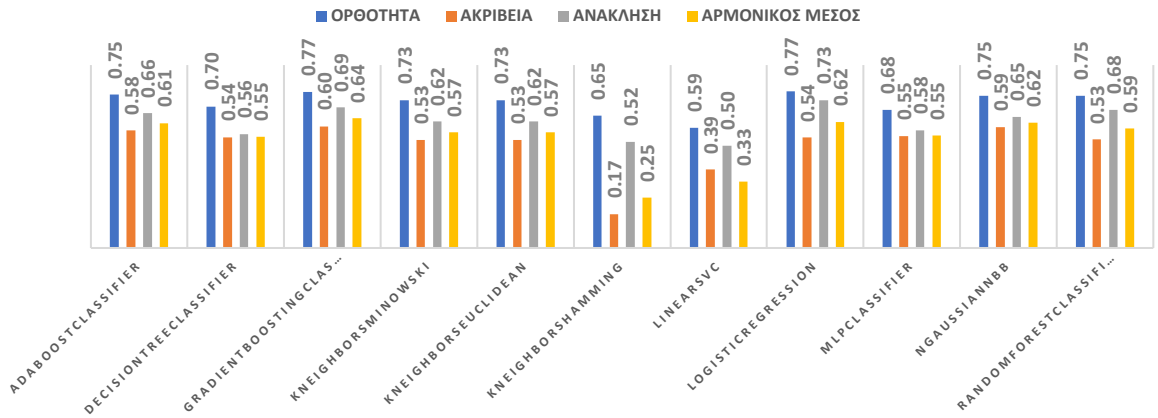
Εφαρμόζοντας τον αλγόριθμο PCA, StandardScaler και τέλος διαγράφοντας τις εγγραφές με έστω μια μηδενική τιμή, καλύτερα αποτελέσματα για όλες τις διαστάσεις στην ορθότητα, στην ακρίβεια και του αρμονικού μέσου τα έχει ο GradientBoost με τιμές 0.87, 0.89 και 0.79 αντίστοιχα. Καλύτερα αποτελέσματα σε όλες τις διαστάσεις για την ανάκληση τα έχει ο απλοϊκός Bayes με 0.75.

4.3 Αποτελέσματα με χρήση της μεθόδου K-Fold Cross Validation

Στις επόμενες γραφικές παραστάσεις παρουσιάζονται όλα τα αποτελέσματα τα οποία προκύπτουν εφαρμόζοντας τους αλγορίθμους μηχανικής μάθησης στο σύνολο δεδομένων Pima Indian, όπου η διαχώρισή τους σε σύνολο εκπαίδευσης και σύνολο δοκιμής είναι η μέθοδος K-Fold Cross Validation με αριθμό K = 10.

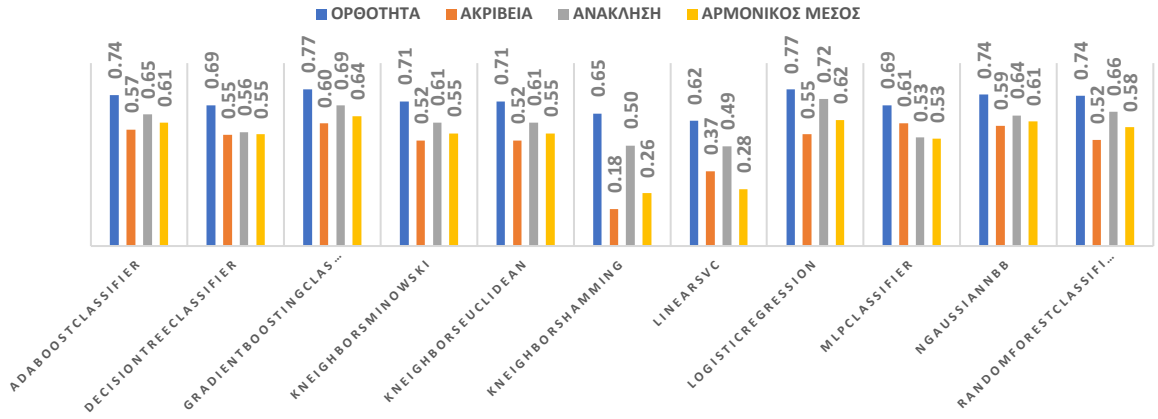
4.3.1 Εφαρμογή αλγορίθμων χωρίς επεξεργασία

4.3.1.1 Χρήση αρχικού συνόλου δεδομένων



Χρησιμοποιώντας το αρχικό σύνολο δεδομένων χωρίς κάποια επεξεργασία παρατηρούμε ότι καλύτερη ορθότητα και ανάκληση στην κατηγοριοποίηση των δεδομένων την κάνει η λογιστική παλινδρόμηση με 0.77 και 0.73 αντίστοιχα. Ο GradientBoost αλγόριθμος έχει τα καλύτερα αποτελέσματα στην ακρίβεια και στον αρμονικό μέσο με 0.6 και 0.64 αντίστοιχα .

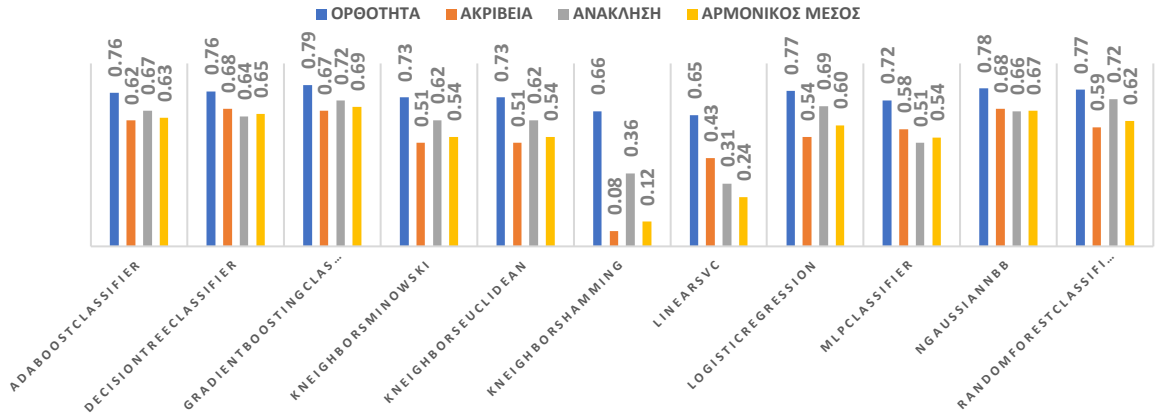
4.3.1.2 Αντικατάσταση μηδενικής τιμής με τη μέση τιμή



Αντικαθιστώντας τη μηδενική τιμή με τη μέση τιμή, καλύτερα αποτελέσματα στην ορθότητα και στον αρμονικό μέσο παρουσιάζει ο GradientBoost με 0.77 και 0.64 αντίστοιχα. Το νευρωνικό δίκτυο έχει το καλύτερο αποτέλεσμα στην ακρίβεια με το αποτέλεσμα να είναι

0.61. Τέλος, καλύτερο αποτέλεσμα στην ανάκληση παρουσιάζει η λογιστική παλινδρόμηση με αποτελέσματα 0.72.

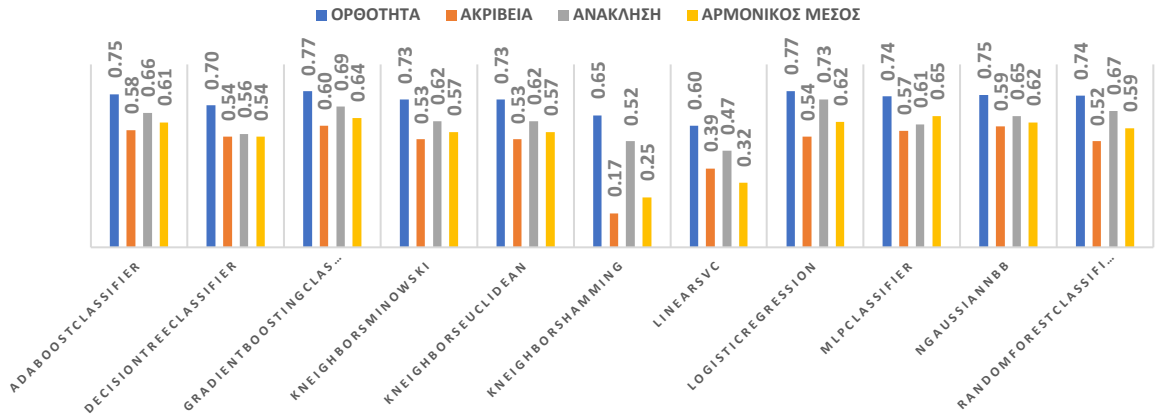
4.3.1.3 Διαγραφή εγγραφής η οποία περιέχει μία μηδενική τιμή



Διαγράφοντας κάθε εγγραφή η οποία έχει έστω και μια μηδενική τιμή ίση με το μηδέν, καλύτερα αποτελέσματα δίνει ο αλγόριθμος GradientBoost όσο αφορά το αποτέλεσμα της ορθότητας και του αρμονικού μέσου με αποτελέσματα 0.79 και 0.69 αντίστοιχα. Τα δέντρα απόφασης και συγκεκριμένα ο DecisionTreeClassifier δίνει τα καλύτερα αποτελέσματα όσο αφορά την ακρίβεια με τιμή 0.68. Τέλος τα τυχαία δάση έχουν τιμή 0.72 στην ανάκληση που είναι η καλύτερη.

4.3.2 Εφαρμογή MinMaxScaler

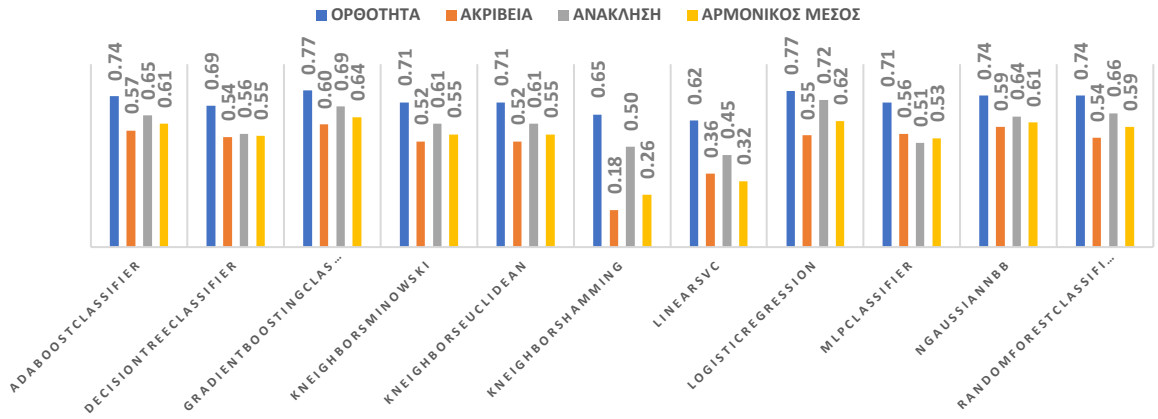
4.3.2.1 Εφαρμογή MinMaxScaler στο αρχικό σύνολο δεδομένων



Εφαρμόζοντας τον αλγόριθμο MinMaxScaler στο αρχικό σύνολο δεδομένων, τα καλύτερα αποτελέσματα όσο αφορά την ορθότητα και την ανάκληση δίνονται από τη λογιστική παλινδρόμηση με τιμές 0.77 και 0.73 αντίστοιχα. Καλύτερη ακρίβεια δίνει ο

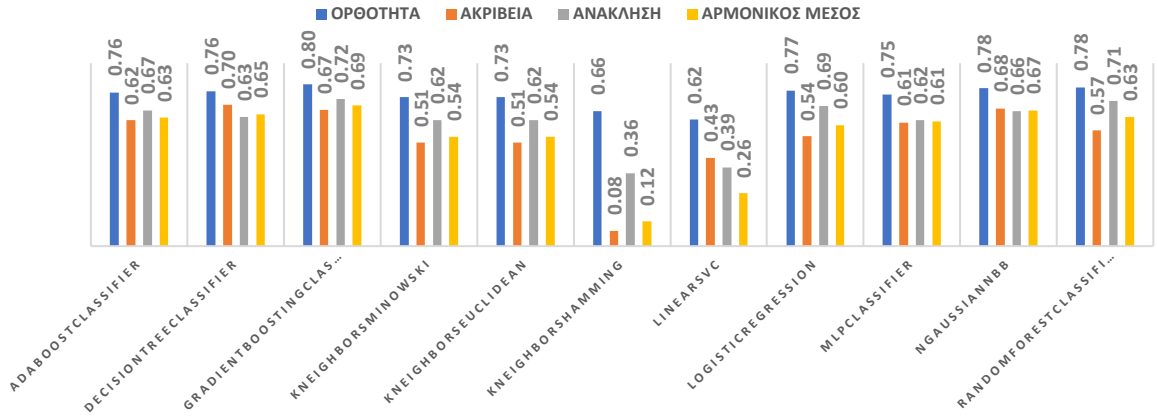
GradientBoost, ενώ καλύτερο αρμονικό μέσο το νευρωνικό δίκτυο με τιμές 0.69 και 0.65 αντίστοιχα.

4.3.2.2 Εφαρμογή MinMaxScaler και αντικατάσταση μηδενικής τιμής με μέση τιμή



Αφού εφαρμοστεί ο αλγόριθμος MinMaxScaler και στη συνέχεια γίνει η αντικατάσταση της μηδενικής τιμής με τη μέση τιμή καλύτερα αποτελέσματα στην ορθότητα, στην ακρίβεια και στον αρμονικό μέσο δίνει ο GradientBoost με 0.77, 0.69 και 0.64 αντίστοιχα. Όσο αφορά την ανάκληση καλύτερα αποτελέσματα έχουμε από τη λογιστική παλινδρόμηση με 0.72.

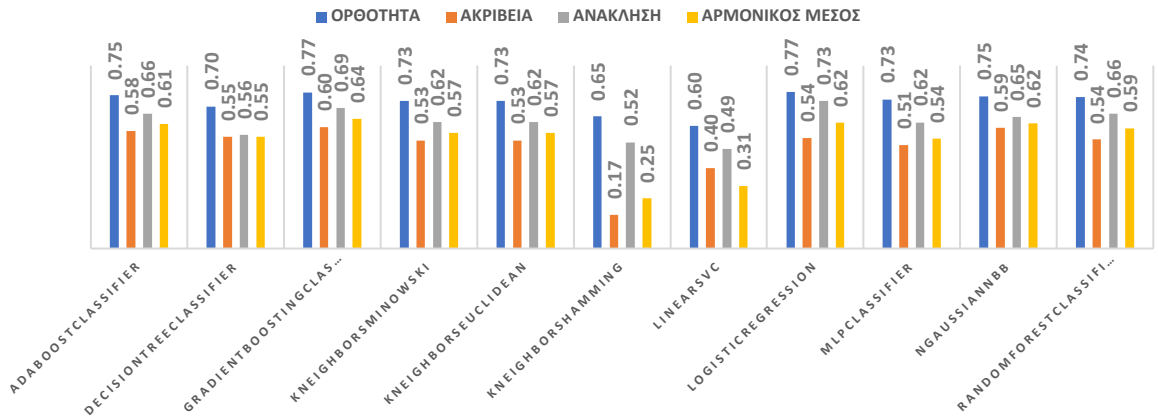
4.3.2.3 Εφαρμογή MinMaxScaler και διαγραφή εγγραφής με μια μηδενική τιμή



Εφαρμόζοντας τον αλγόριθμο MinMaxScaler και στη συνέχεια διαγράφοντας την εγγραφή η οποία περιέχει έστω και μια μηδενική τιμή καλύτερα αποτελέσματα στην ορθότητα, στην ανάκληση και στον αρμονικό μέσο δίνει ο GradientBoost με 0.80, 0.72 και 0.69 αντίστοιχα. Όσο αφορά την ακρίβεια καλύτερα αποτελέσματα έχουμε από τα δέντρα απόφασης και συγκεκριμένα από τον DecistionTreeClassifier με 0.7.

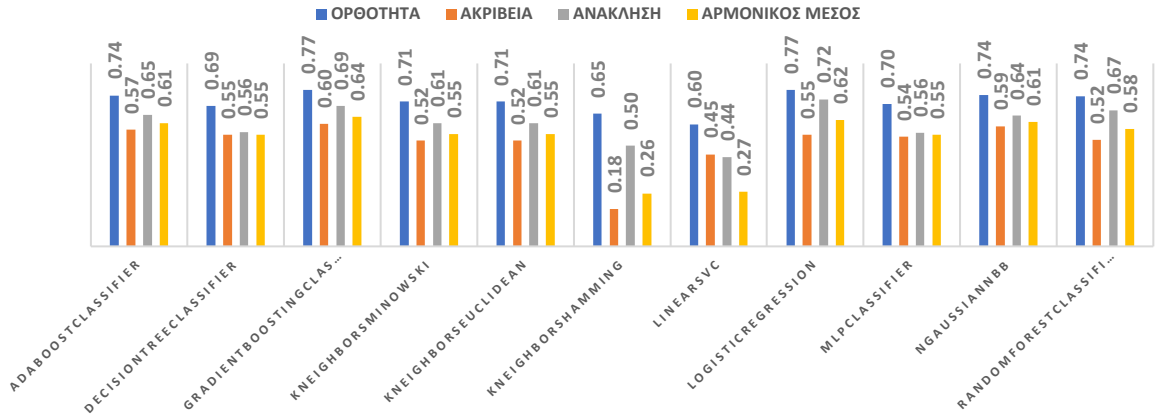
4.3.3 Εφαρμογή StandardScaler

4.3.3.1. Εφαρμογή StandardScaler στο αρχικό σύνολο δεδομένων



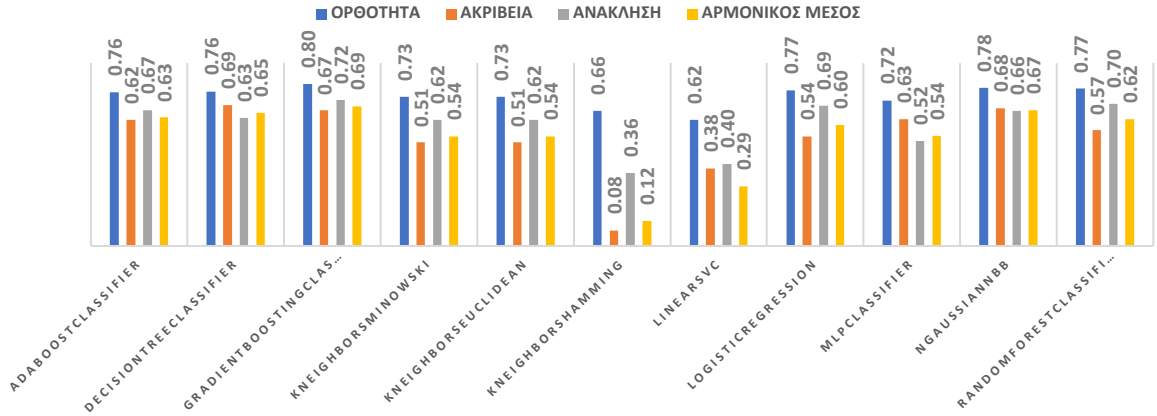
Εφαρμόζοντας τον αλγόριθμο StandardScaler στο αρχικό σύνολο δεδομένων, τα καλύτερα αποτελέσματα όσο αφορά την ορθότητα και την ανάκληση προκύπτουν με τη χρήση της λογιστικής παλινδρόμησης με τιμές 0.77 και 0.73 αντίστοιχα. Ο GradientBoost δίνει καλύτερα αποτελέσματα στην ακρίβεια και στον αρμονικό μέσο με 0.6 και 0.64, αντίστοιχα.

4.3.3.2 Εφαρμογή StandardScaler και αντικατάσταση μηδενικής τιμής με μέση τιμή



Αφού εφαρμοστεί ο αλγόριθμος StandardScaler και στη συνέχεια γίνει η αντικατάσταση της μηδενικής τιμής με τη μέση τιμή, καλύτερα αποτελέσματα στην ορθότητα, στην ακρίβεια και στον αρμονικό μέσο δίνει ο GradientBoost με 0.77, 0.6 και 0.64 αντίστοιχα. Όσο αφορά την ανάκληση καλύτερα αποτελέσματα έχουμε από τον αλγόριθμο της λογιστικής παλινδρόμησης με τιμή 0.72.

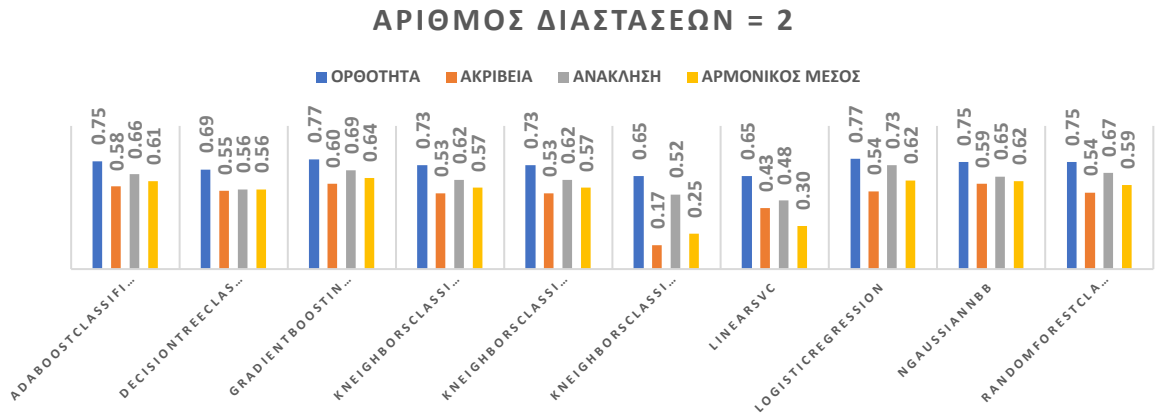
4.3.3.3 Εφαρμογή StandardScaler και διαγραφή εγγραφής με μια μηδενική τιμή

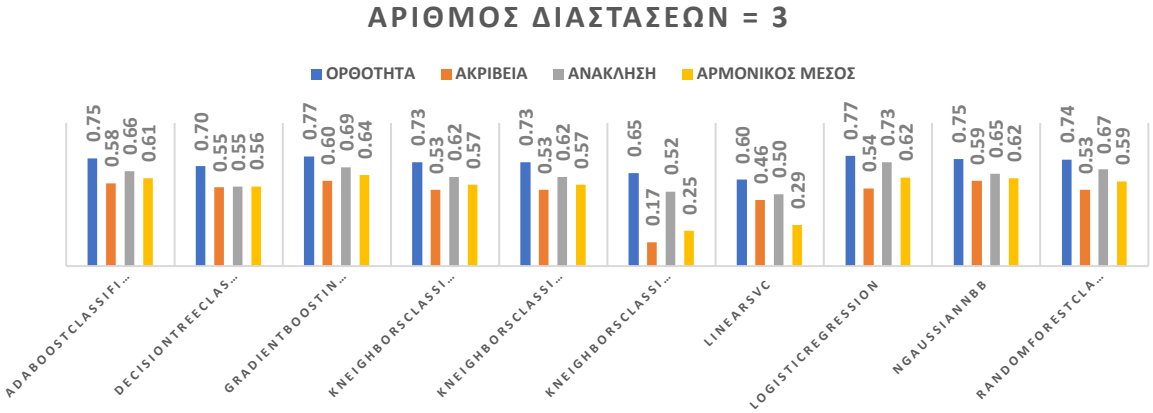


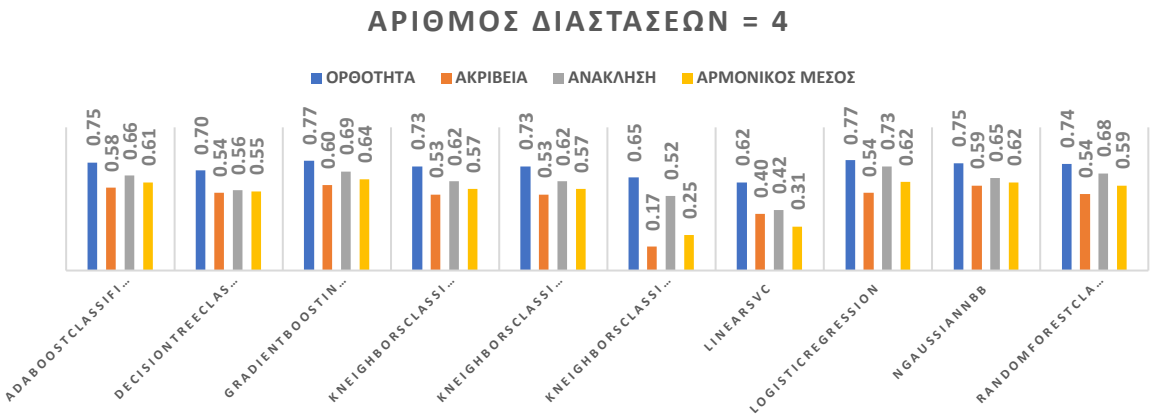
Εφαρμόζοντας τον αλγόριθμο StandardScaler και στη συνέχεια διαγράφοντας την εγγραφή, η οποία περιέχει έστω και μια μηδενική τιμή, καλύτερα αποτελέσματα στην ορθότητα, στην ανάκληση και στον αρμονικό μέσο έχουμε από τον GradientBoost με 0.8, 0.72 και 0.69 αντίστοιχα, ενώ καλύτερο αποτέλεσμα στην ακρίβεια έχουν τα δέντρα απόφασης και συγκεκριμένα ο DecisionTreeClassifier με 0.69.

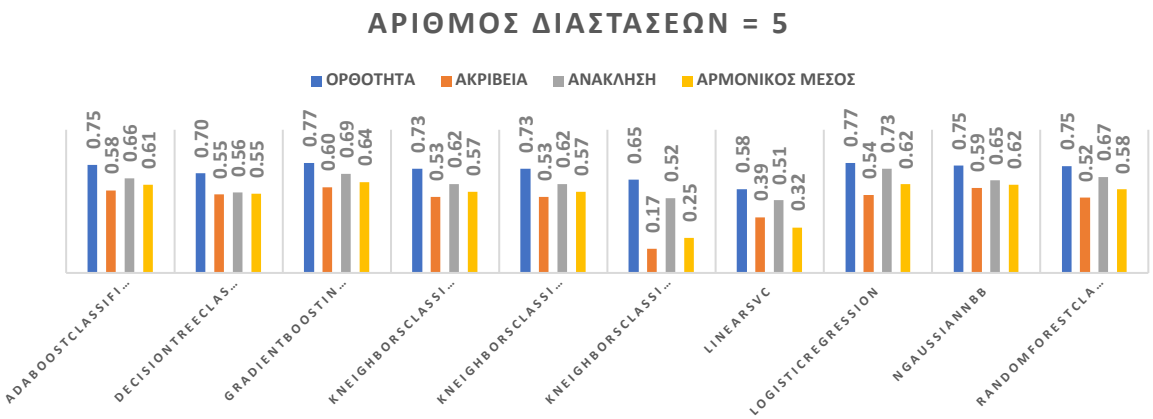
4.3.4 Εφαρμογή PCA

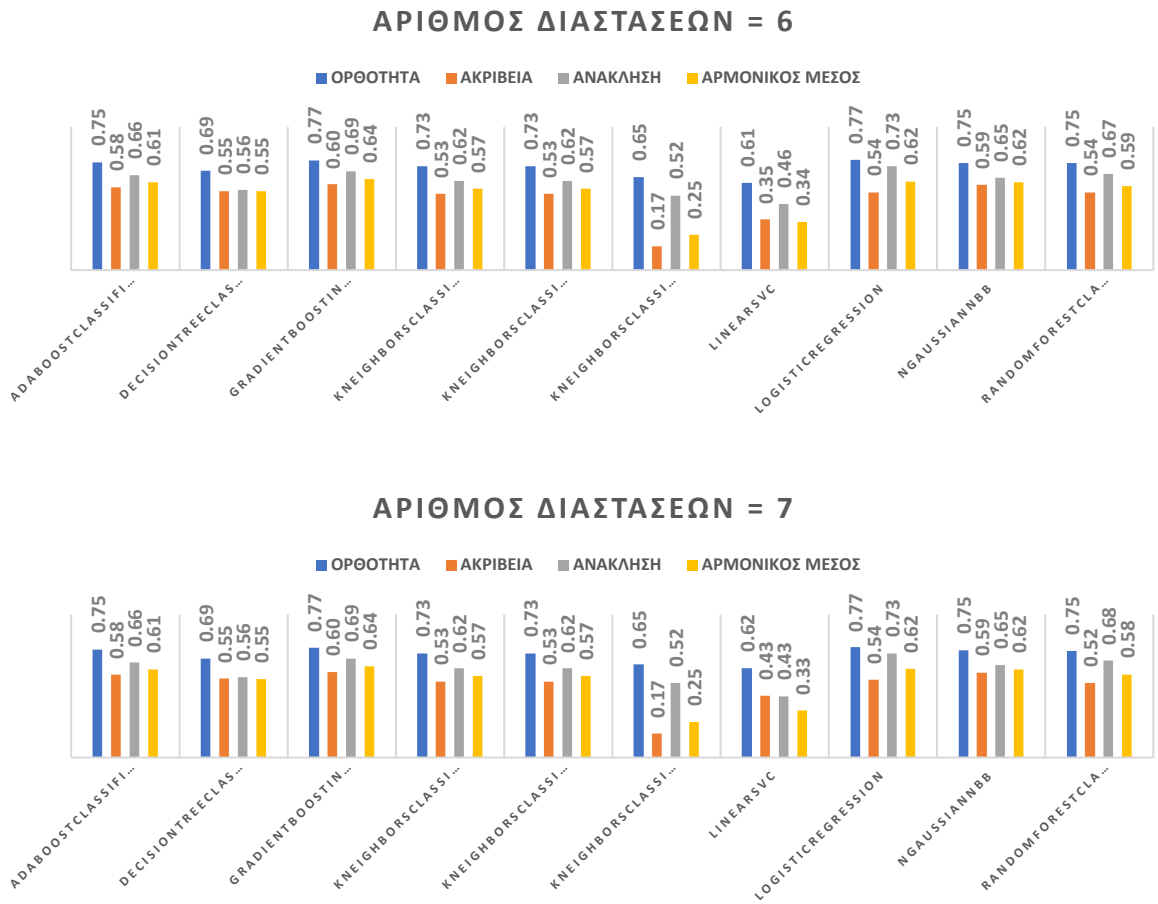
4.3.4.1 Εφαρμογή PCA στο αρχικό σύνολο δεδομένων




ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 3
ΟΡΘΟΤΗΤΑ
ΑΚΡΙΒΕΙΑ
ΑΝΑΚΛΗΣΗ
ΑΡΜΟΝΙΚΟΣ ΜΕΣΟΣ
0.75 0.58 0.66 0.61
0.70 0.55 0.55 0.56
0.77 0.60 0.69 0.64
0.73 0.53 0.62 0.57
0.73 0.53 0.62 0.57
0.65 0.17 0.52 0.25
0.60 0.46 0.50 0.29
0.77 0.54 0.73 0.62
0.75 0.59 0.65 0.62
0.74 0.53 0.67 0.59
ADABOOSTCLASSIFI...
DECISIONTREECLAS...
GRADIENTBOOSTIN...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
LINEARSVC
LOGISTICREGRESSION
NGAUSSIANNBB
RANDOMFORESTCLA...

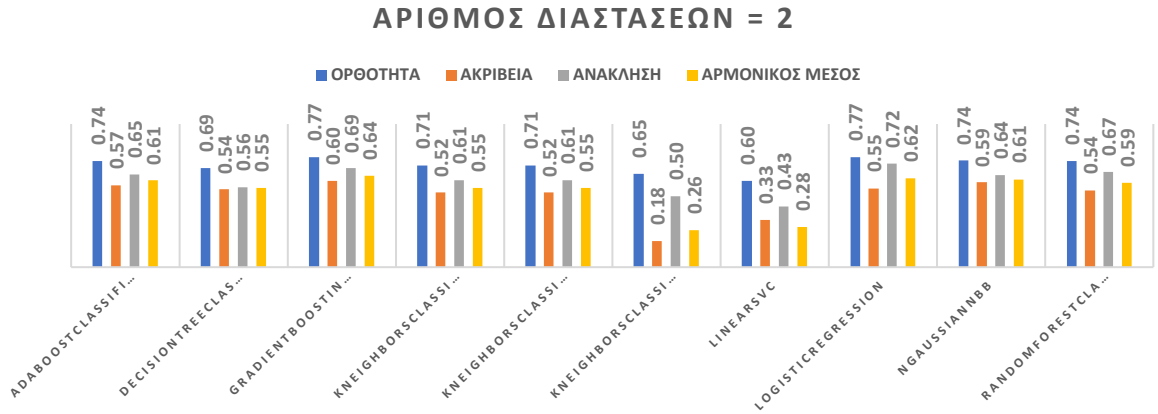

ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 4
ΟΡΘΟΤΗΤΑ
ΑΚΡΙΒΕΙΑ
ΑΝΑΚΛΗΣΗ
ΑΡΜΟΝΙΚΟΣ ΜΕΣΟΣ
0.75 0.58 0.66 0.61
0.70 0.54 0.56 0.55
0.77 0.60 0.69 0.64
0.73 0.53 0.62 0.57
0.73 0.53 0.62 0.57
0.65 0.17 0.52 0.25
0.62 0.40 0.42 0.31
0.77 0.54 0.73 0.62
0.75 0.59 0.65 0.62
0.74 0.54 0.68 0.59
ADABOOSTCLASSIFI...
DECISIONTREECLAS...
GRADIENTBOOSTIN...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
LINEARSVC
LOGISTICREGRESSION
NGAUSSIANNBB
RANDOMFORESTCLA...

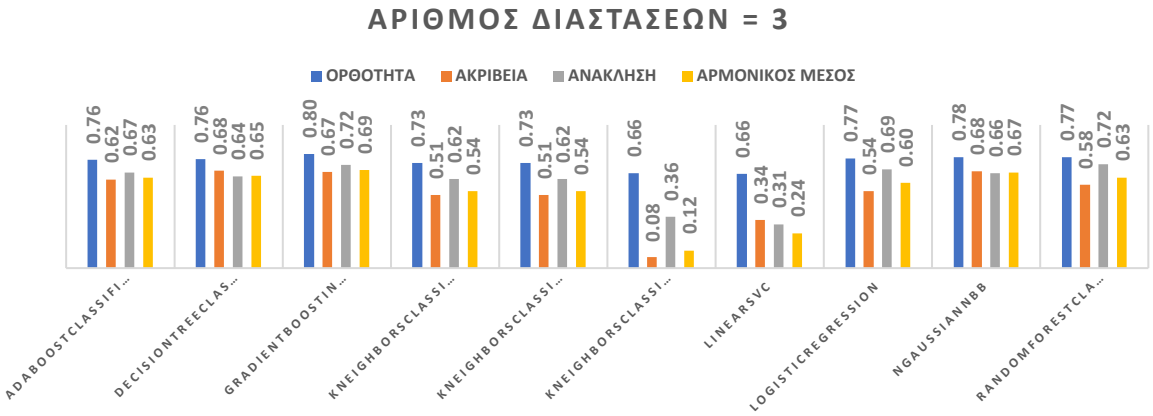

ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 5
ΟΡΘΟΤΗΤΑ
ΑΚΡΙΒΕΙΑ
ΑΝΑΚΛΗΣΗ
ΑΡΜΟΝΙΚΟΣ ΜΕΣΟΣ
0.75 0.58 0.66 0.61
0.70 0.55 0.56 0.55
0.77 0.60 0.69 0.64
0.73 0.53 0.62 0.57
0.73 0.53 0.62 0.57
0.65 0.17 0.52 0.25
0.58 0.39 0.51 0.32
0.77 0.54 0.73 0.62
0.75 0.59 0.65 0.62
0.75 0.52 0.67 0.58
ADABOOSTCLASSIFI...
DECISIONTREECLAS...
GRADIENTBOOSTIN...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
LINEARSVC
LOGISTICREGRESSION
NGAUSSIANNBB
RANDOMFORESTCLA...

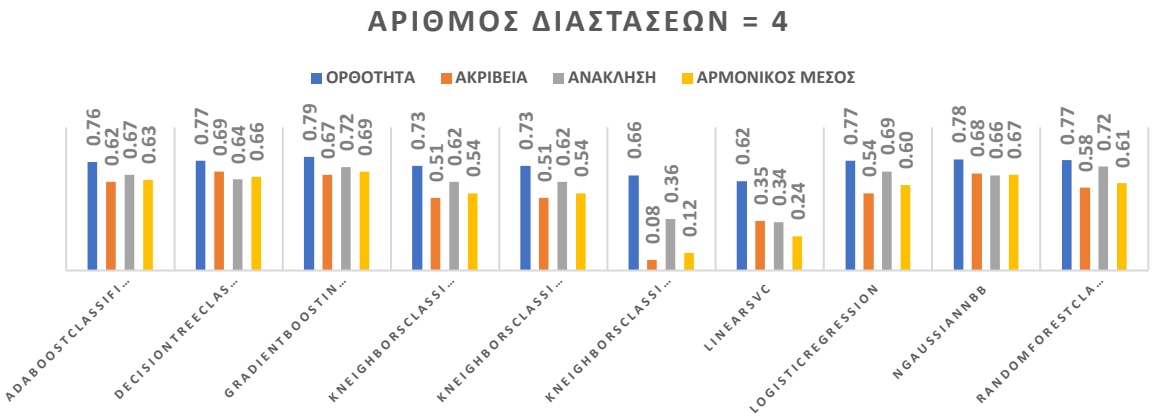


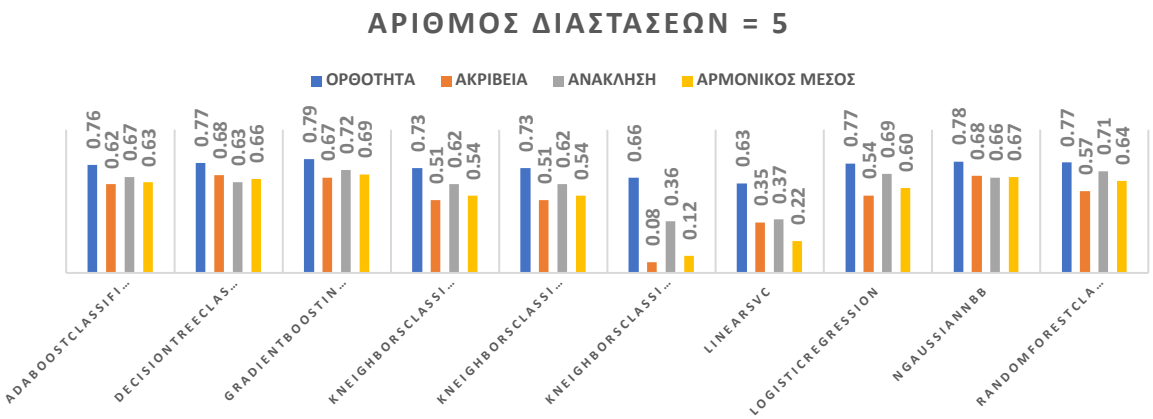
Εφαρμόζοντας τον αλγόριθμο PCA και στη συνέχεια MinMaxScaler, καλύτερα αποτελέσματα για όλες τις διαστάσεις στην ορθότητα και στην ανάκληση τα έχει η λογιστική παλινδρόμηση με τιμές 0.77 και 0.73 αντίστοιχα. Καλύτερα αποτελέσματα για την ακρίβεια και τον αρμονικό μέσο έχει ο GradientBoost όταν διαστάσεις του συνόλου είναι 5 με αποτέλεσμα 0.6 και 0.64.

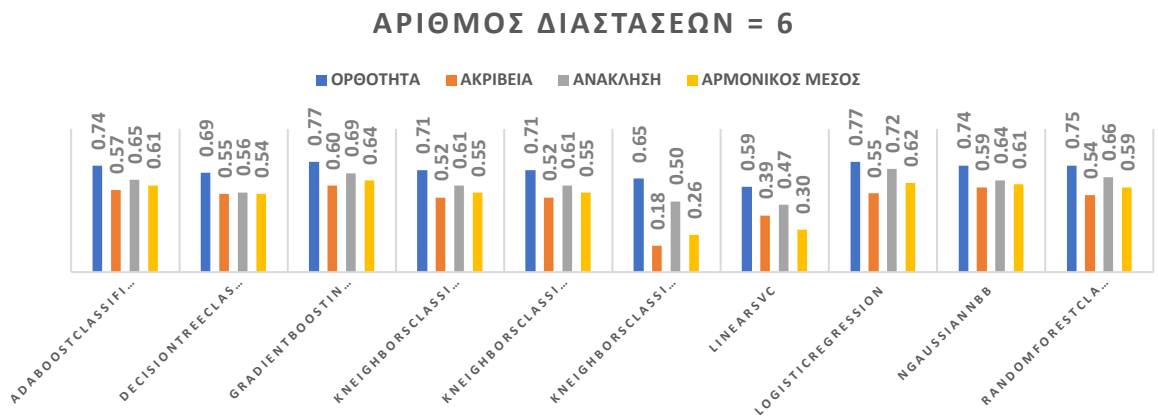
4.3.4.2 Εφαρμογή PCA και αντικατάσταση μηδενικής τιμής με μέση τιμή

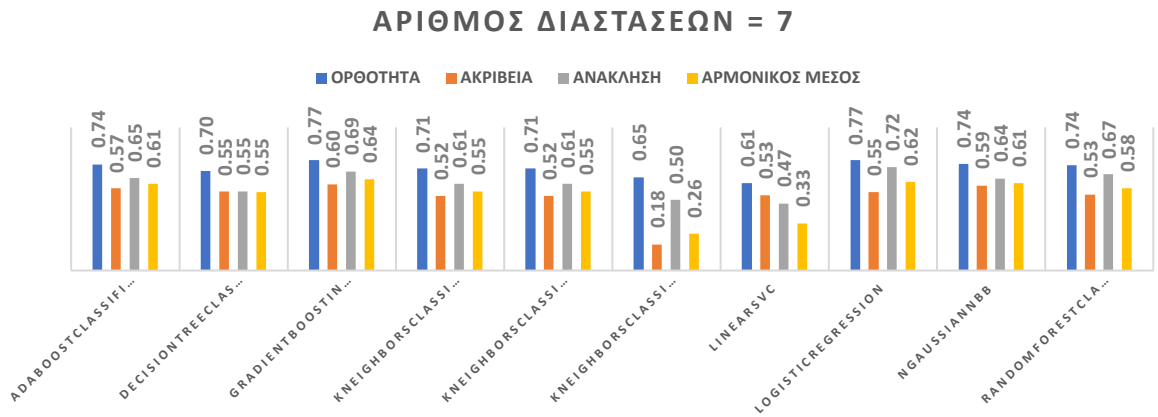



ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 3
ΟΡΘΟΤΗΤΑ
ΑΚΡΙΒΕΙΑ
ΑΝΑΚΛΗΣΗ
ΑΡΜΟΝΙΚΟΣ ΜΕΣΟΣ
0.76 0.62 0.67 0.63
0.76 0.68 0.64 0.65
0.80 0.67 0.72 0.69
0.73 0.51 0.62 0.54
0.73 0.51 0.62 0.54
0.66 0.08 0.36 0.12
0.66 0.34 0.31 0.24
0.77 0.54 0.69 0.60
0.78 0.68 0.66 0.67
0.77 0.58 0.72 0.63
ADABOOSTCLASSIFI...
DECISIONTREECLAS...
GRADIENTBOOSTIN...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
LINEARSVC
LOGISTICREGRESSION
NGAUSSIANNBB
RANDOMFORESTCLA...


ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 4
ΟΡΘΟΤΗΤΑ
ΑΚΡΙΒΕΙΑ
ΑΝΑΚΛΗΣΗ
ΑΡΜΟΝΙΚΟΣ ΜΕΣΟΣ
0.76 0.62 0.67 0.63
0.77 0.69 0.64 0.66
0.79 0.67 0.72 0.69
0.73 0.51 0.62 0.54
0.73 0.51 0.62 0.54
0.66 0.08 0.36 0.12
0.62 0.35 0.34 0.24
0.77 0.54 0.69 0.60
0.78 0.68 0.66 0.67
0.77 0.58 0.72 0.61
ADABOOSTCLASSIFI...
DECISIONTREECLAS...
GRADIENTBOOSTIN...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
LINEARSVC
LOGISTICREGRESSION
NGAUSSIANNBB
RANDOMFORESTCLA...


ΑΡΙΘΜΟΣ ΔΙΑΣΤΑΣΕΩΝ = 5
ΟΡΘΟΤΗΤΑ
ΑΚΡΙΒΕΙΑ
ΑΝΑΚΛΗΣΗ
ΑΡΜΟΝΙΚΟΣ ΜΕΣΟΣ
0.76 0.62 0.67 0.63
0.77 0.68 0.63 0.66
0.79 0.67 0.72 0.69
0.73 0.51 0.62 0.54
0.73 0.51 0.62 0.54
0.66 0.08 0.36 0.12
0.63 0.35 0.37 0.22
0.77 0.54 0.69 0.60
0.78 0.68 0.66 0.67
0.77 0.57 0.71 0.64
ADABOOSTCLASSIFI...
DECISIONTREECLAS...
GRADIENTBOOSTIN...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
KNEIGHBORSCLASSI...
LINEARSVC
LOGISTICREGRESSION
NGAUSSIANNBB
RANDOMFORESTCLA...



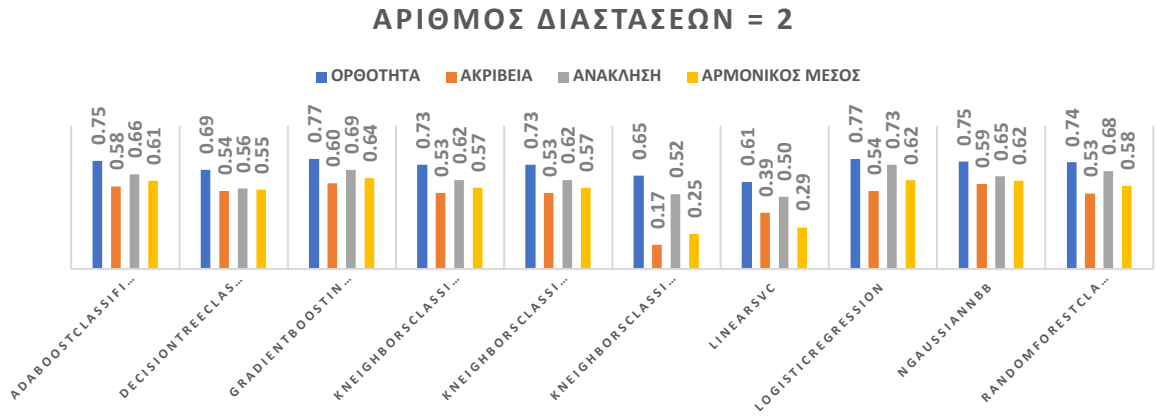


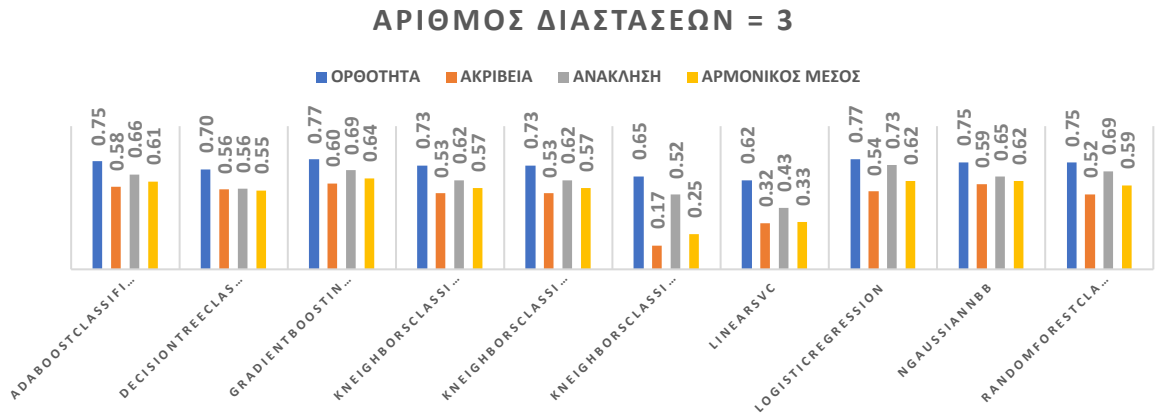
Εφαρμόζοντας τον αλγόριθμο PCA, στη συνέχεια MinMaxScaler και τέλος αντικαθιστώντας τις μηδενικές τιμές με τη μέση τιμή, καλύτερο αποτέλεσμα στην ορθότητα, στην ακρίβεια και στον αρμονικό μέσο το έχει ο GradientBoost με τιμές 0.79, 0.67 και 0.64 αντίστοιχα. Οι πιο πάνω τιμές πάρθηκαν όταν ο αριθμός των διαστάσεων στο σύνολο δεδομένων για την ορθότητα ήταν 4, για την ακρίβεια ήταν 5 και για τον αρμονικό μέσο 2. Τέλος, καλύτερα αποτελέσματα για την ανάκληση δίνει η λογιστική παλινδρόμηση σε όλες τις διαστάσεις με τιμή 0.72.

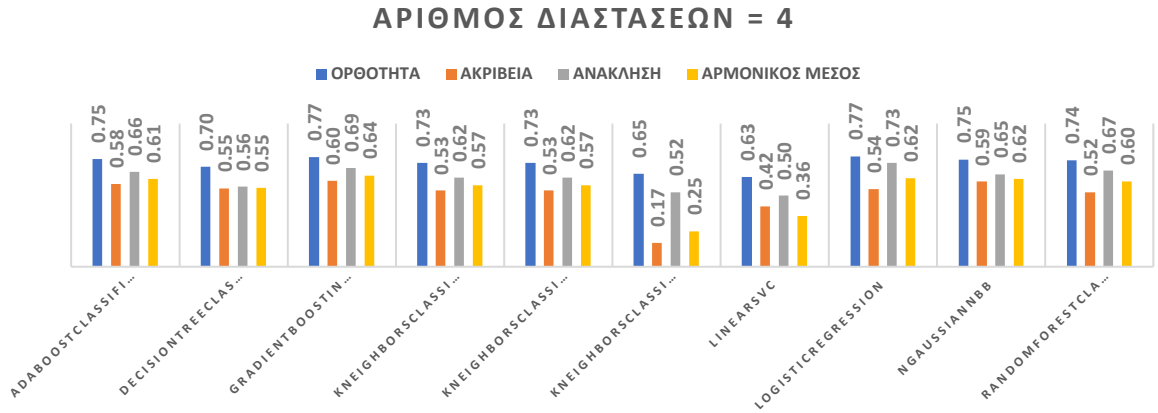
4.3.4.3 Εφαρμογή PCA και διαγραφή εγγραφής με μηδενική τιμή

Εφαρμόζοντας τον αλγόριθμο PCA, MinMaxScaler και τέλος διαγράφοντας τις εγγραφές με έστω μια μηδενική τιμή, καλύτερο αποτέλεσμα για την ορθότητα είναι όταν το πλήθος των διαστάσεων είναι ίσο με 2 και δίνεται από τον GradientBoost με 0.8. Όσο αφορά την ακρίβεια, τα δέντρα απόφασης και συγκεκριμένα ο DecisionTreeClassifier και ο αριθμός διαστάσεων στο σύνολο δεδομένων να είναι ίσο με 2, δίνει τα καλύτερα αποτελέσματα με 0.69. Το τυχαίο δάσος όταν οι διαστάσεις του συνόλου είναι 3 δίνει το καλύτερο αποτέλεσμα στην ανάκληση με 0.72. Τέλος, ο GradientBoost έχει το καλύτερο αποτέλεσμα στον αρμονικό μέσο όταν οι διαστάσεις είναι 3 με τιμή 0.69.

4.3.4.4 Εφαρμογή PCA και MinMaxScaler στο αρχικό σύνολο δεδομένων

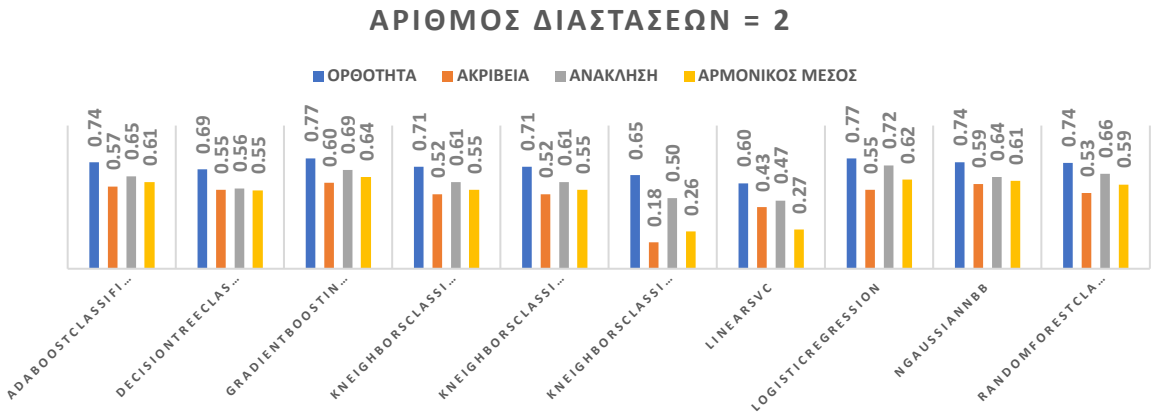


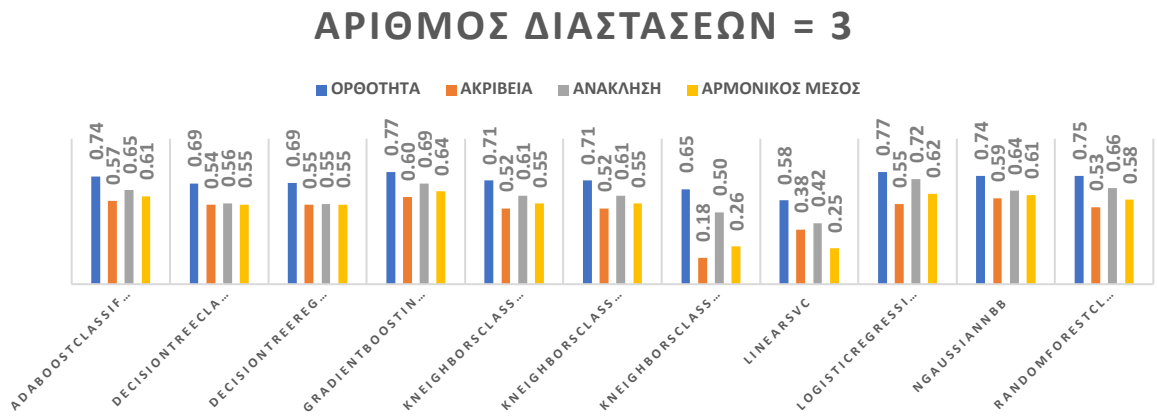


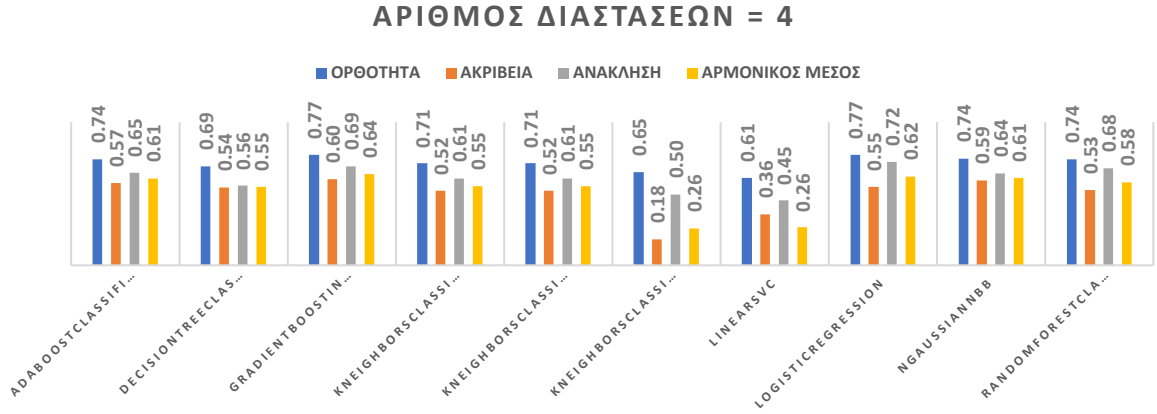


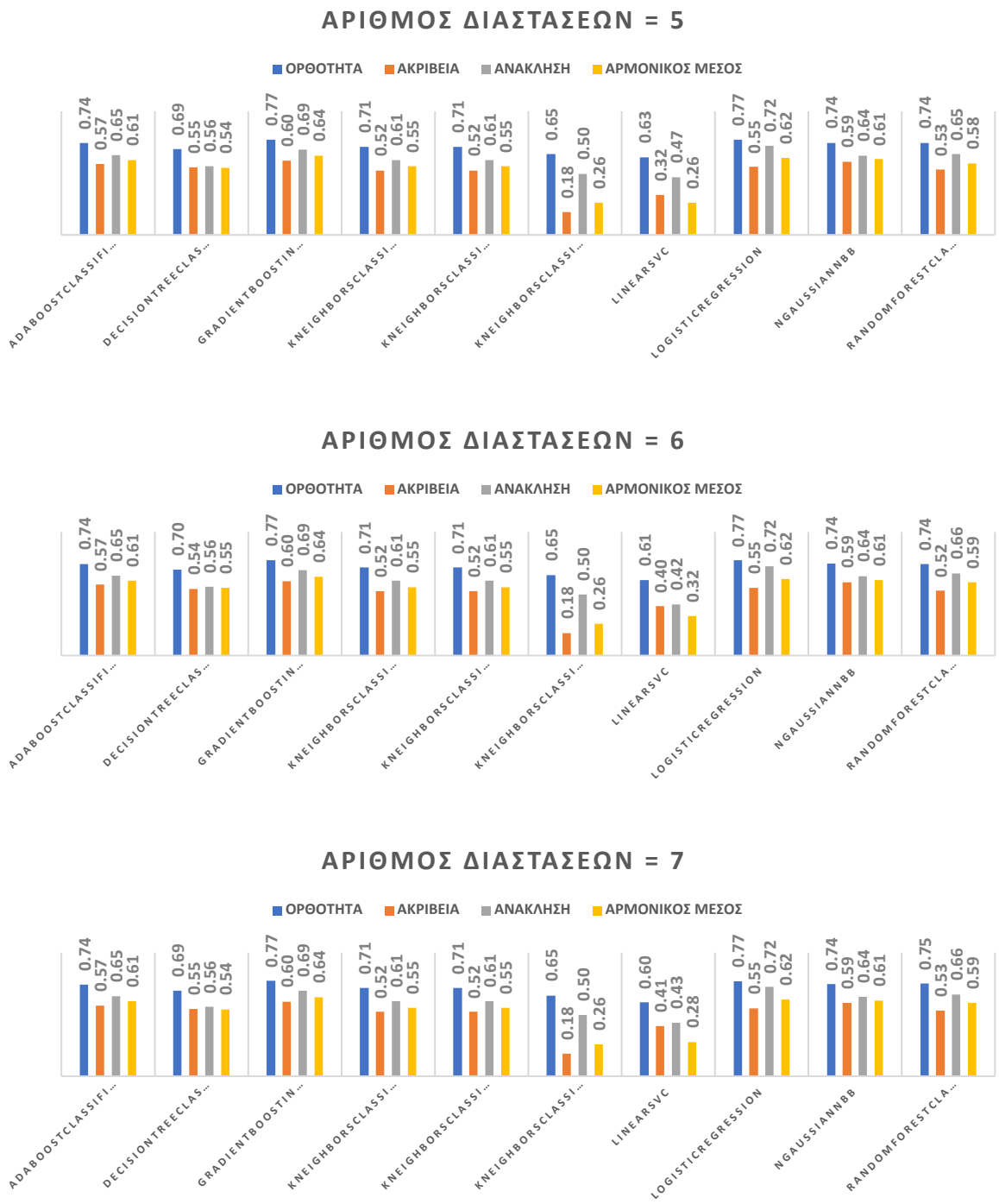
Εφαρμόζοντας τον αλγόριθμο PCA και στη συνέχεια MinMaxScaler, καλύτερα αποτελέσματα για όλες τις διαστάσεις στην ορθότητα, στην ανάκληση τα έχει η λογιστική παλινδρόμηση με τιμές 0.77 και 0.73. Ο GradientBoost έχει την καλύτερη απόδοση για την ακρίβεια με 0.6 και τον αρμονικό μέσο με 0.64, όταν οι διαστάσεις είναι 2 και 7 αντίστοιχα.

4.3.4.5 Εφαρμογή PCA, MinMaxScaler και αντικατάσταση μηδενικής τιμής με μέση τιμή







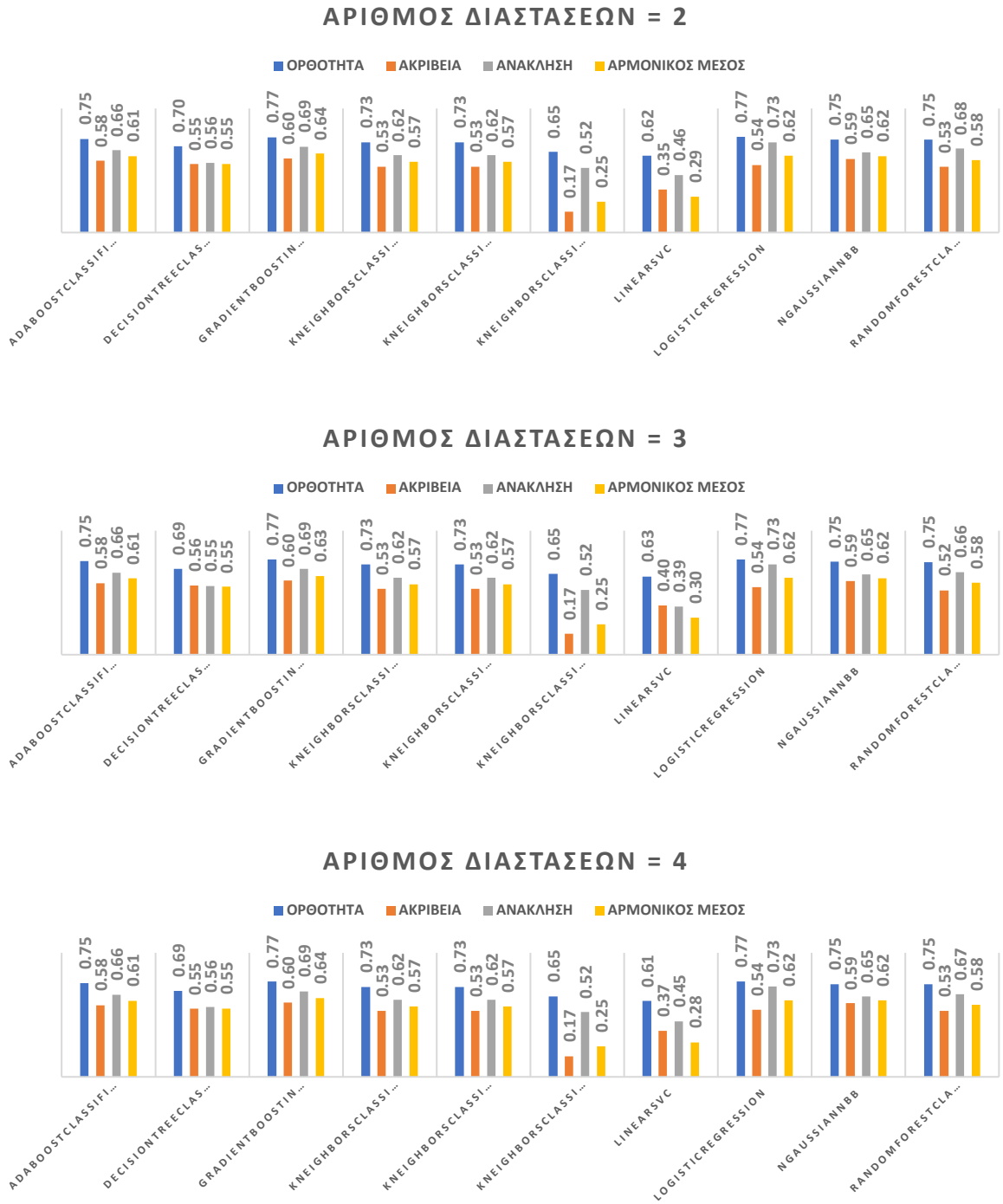


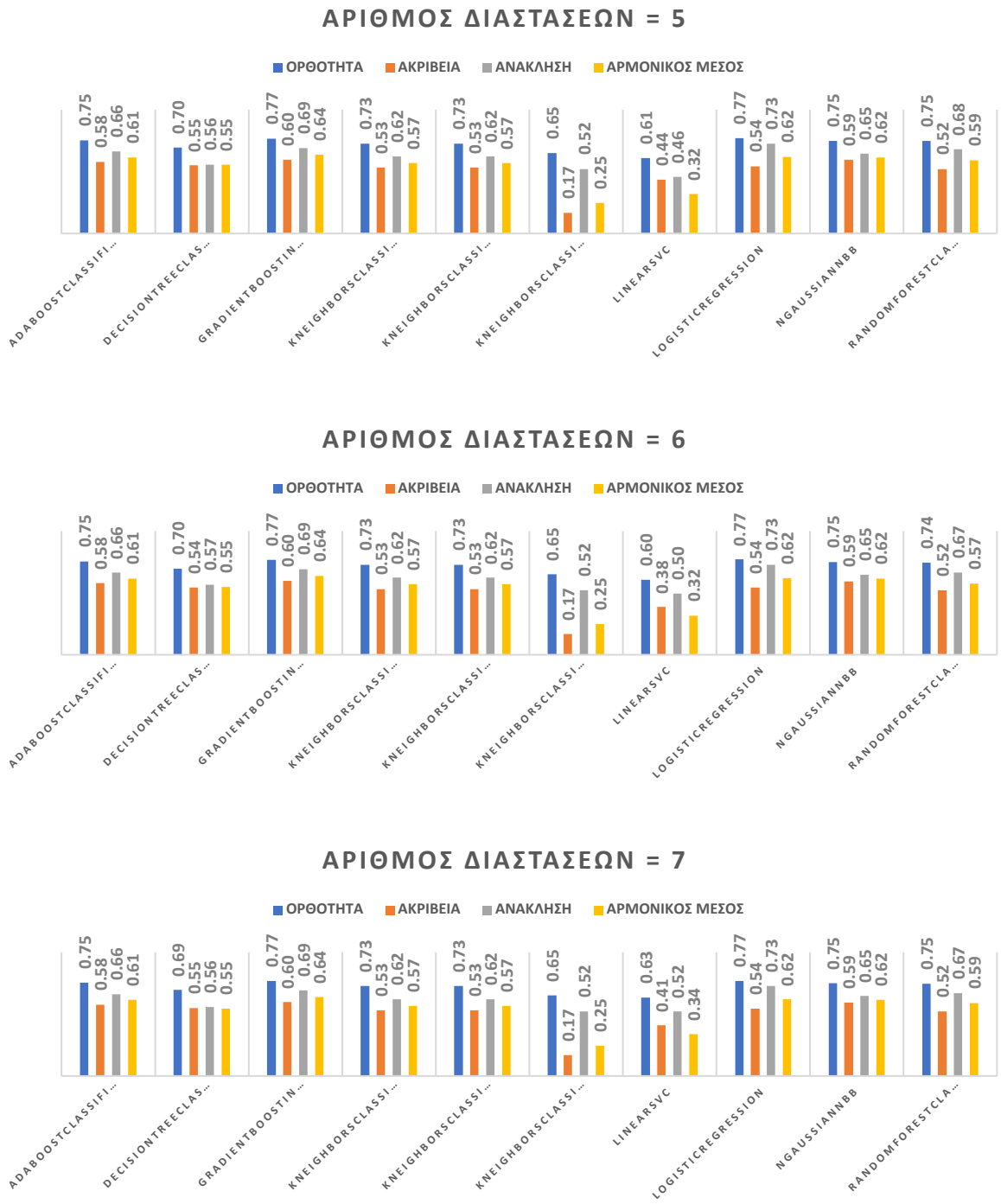
Εφαρμόζοντας τον αλγόριθμο PCA, στη συνέχεια MinMaxScaler και τέλος αντικαθιστώντας τις μηδενικές τιμές με τη μέση τιμή, καλύτερα αποτελέσματα όταν οι διαστάσεις είναι 5 τις δίνει ο GradientBoost με 0.77 και 0.6 για την ορθότητα και την ακρίβεια αντίστοιχα. Η λογιστική παλινδρόμηση για όλες τις διαστάσεις δίνει το καλύτερο αποτέλεσμα στην ανάκληση με 0.72, ενώ ο GradientBoost όταν οι διαστάσεις είναι 4 δίνει το καλύτερο αποτέλεσμα στον αρμονικό μέσο με 0.64.

4.3.4.6 Εφαρμογή PCA, MinMaxScaler και διαγραφή εγγραφής με μηδενική τιμή

Εφαρμόζοντας τον αλγόριθμο PCA, MinMaxScaler και τέλος διαγράφοντας τις εγγραφές με έστω μια μηδενική τιμή, καλύτερο αποτέλεσμα για την ορθότητα είναι όταν το πλήθος των διαστάσεων είναι ίσο με 2 και δίνεται από τον GradientBoost με 0.79. Όσο αφορά την ακρίβεια, τα δέντρα απόφασης και συγκεκριμένα ο DecisionTreeClassifier και ο αριθμός διαστάσεων στο σύνολο δεδομένων να είναι ίσο με 4, δίνει τα καλύτερα αποτελέσματα με 0.69. Το τυχαίο δάσος όταν οι διαστάσεις του συνόλου είναι 3 δίνει το καλύτερο αποτέλεσμα στην ανάκληση με 0.73. Τέλος, ο GradientBoost έχει το καλύτερο αποτέλεσμα στον αρμονικό μέσο όταν οι διαστάσεις είναι 2 με τιμή 0.69.

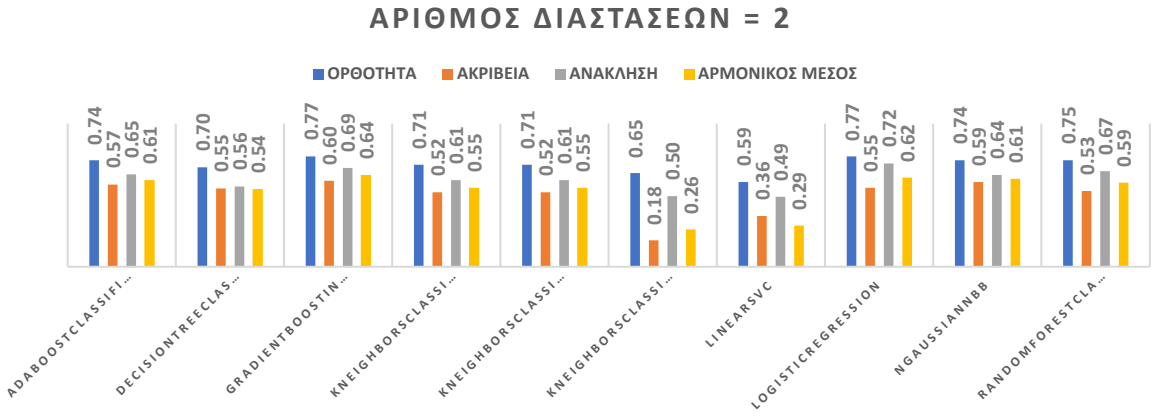
4.3.4.7 Εφαρμογή PCA και StandardScaler στο αρχικό σύνολο δεδομένων

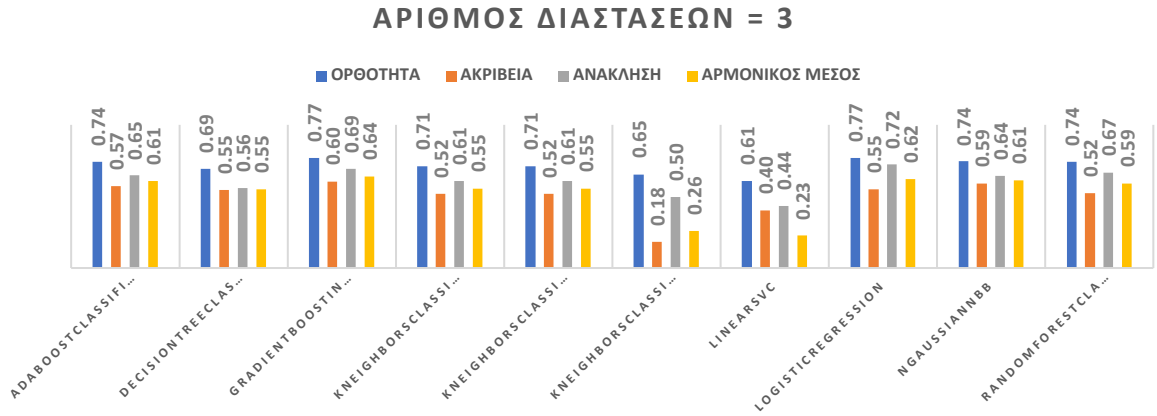


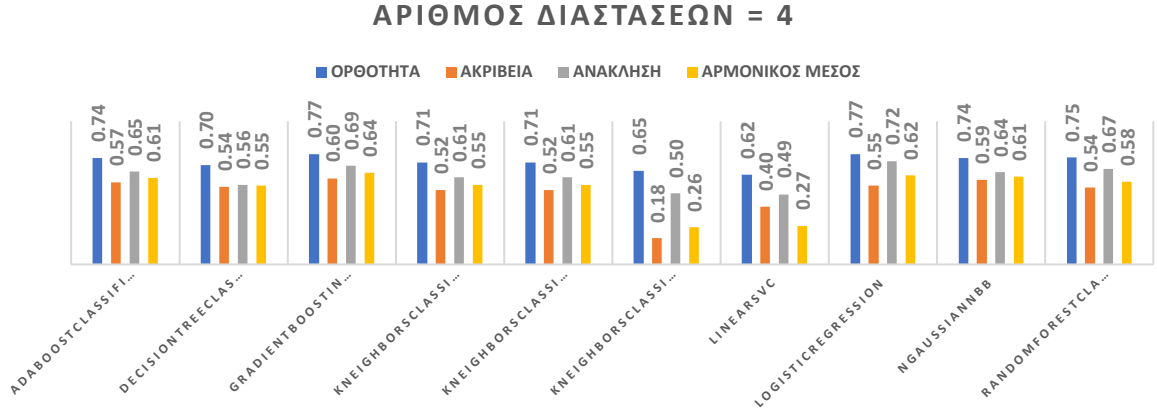


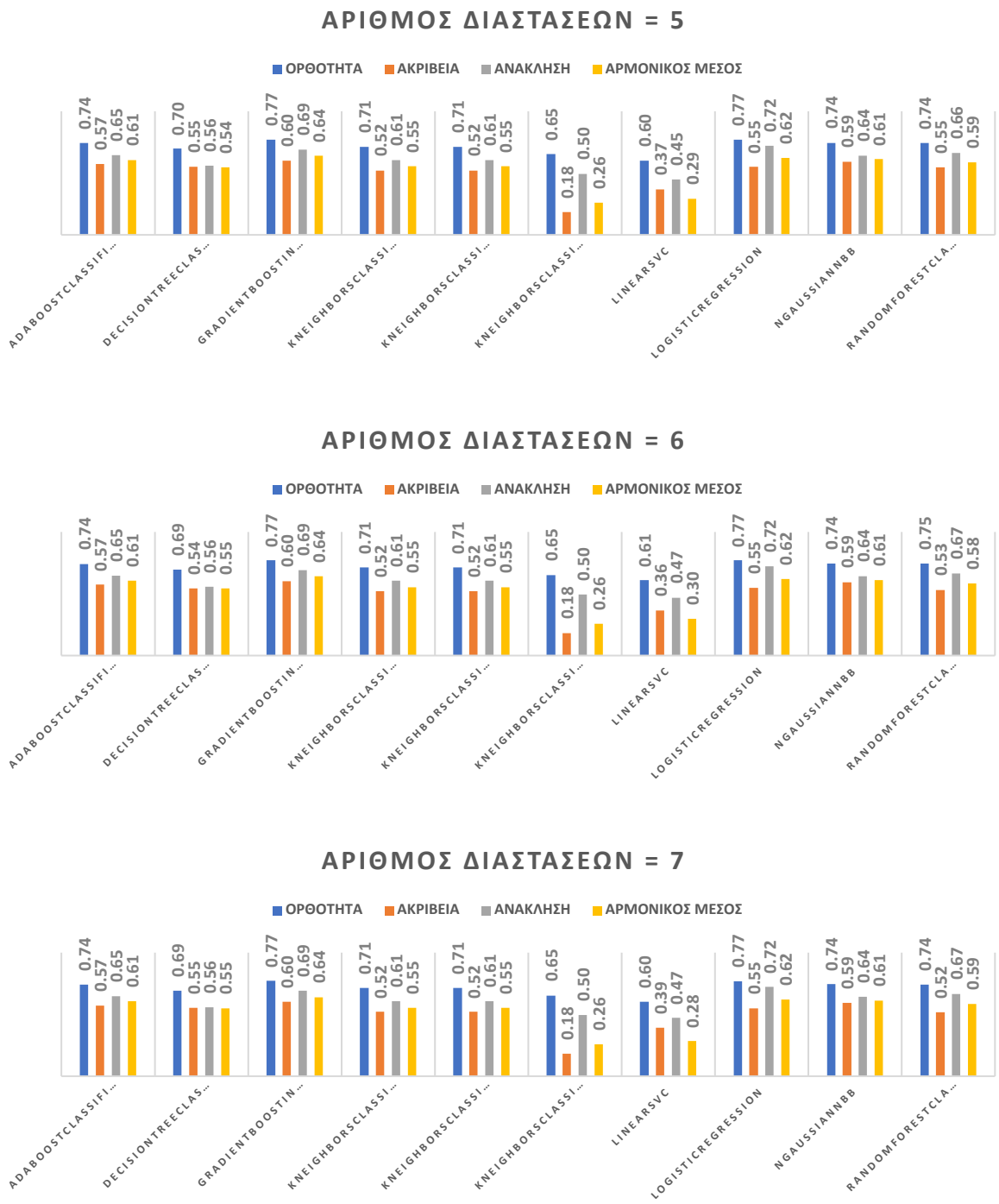
Εφαρμόζοντας τον αλγόριθμο PCA και στη συνέχεια StandardScaler, καλύτερα αποτελέσματα για όλες τις διαστάσεις στην ορθότητα και στην ανάκληση τα έχει η λογιστική παλινδρόμηση με τιμές 0.77 και 0.73. Ο GradientBoost έχει την καλύτερη απόδοση για την ακρίβεια με 0.6 και τον αρμονικό μέσο με 0.63, όταν οι διαστάσεις είναι 2 και 5 αντίστοιχα.

4.3.4.8 Εφαρμογή PCA, StandardScaler και αντικατάσταση μηδενικής τιμής με μέση τιμή



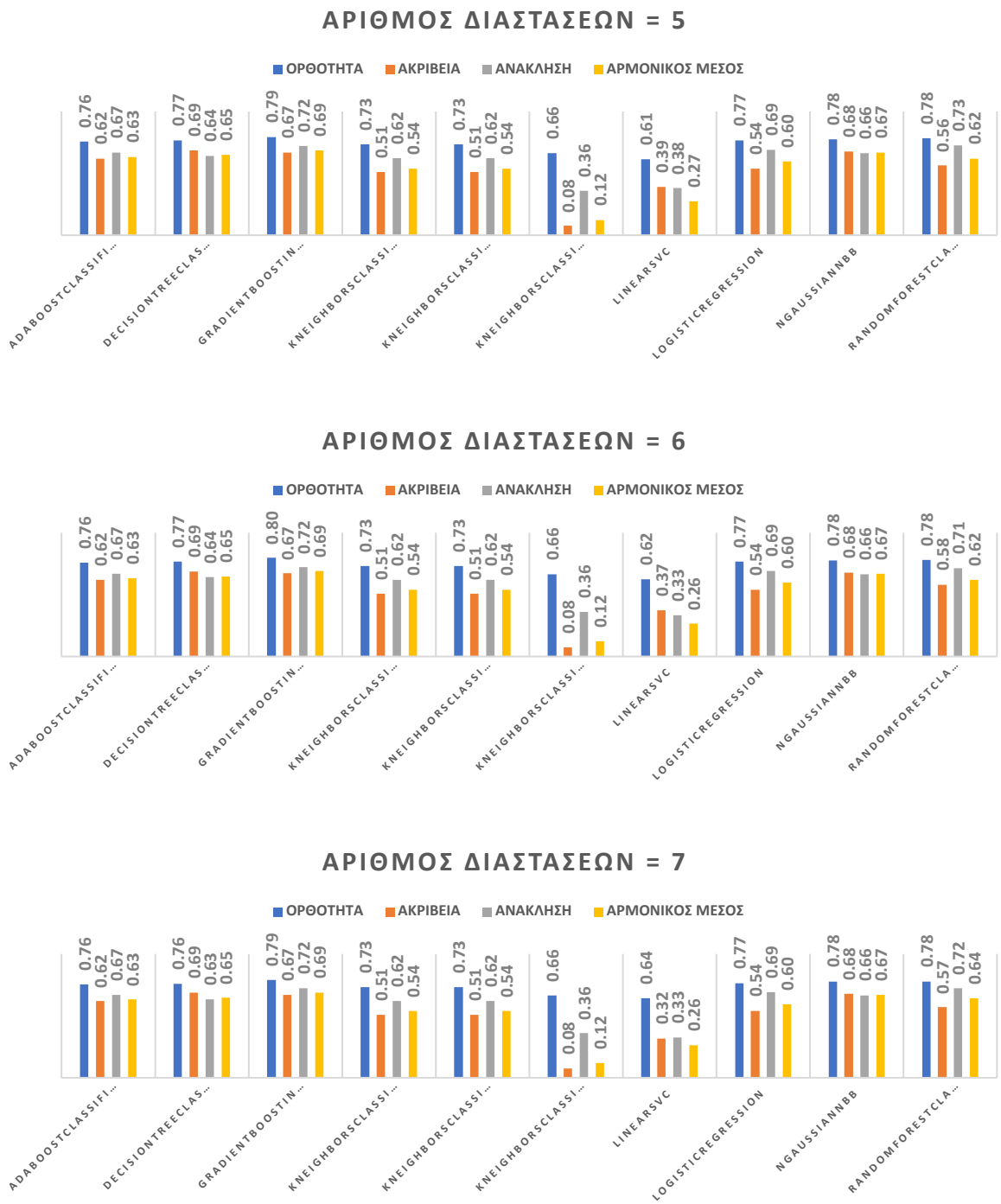






Εφαρμόζοντας τον αλγόριθμο PCA, στη συνέχεια StandardScaler και τέλος αντικαθιστώντας τις μηδενικές τιμές με τη μέση τιμή, καλύτερα αποτελέσματα έχει ο GradientBoost όταν οι διαστάσεις είναι 5 και 6 με 0.77 για την ορθότητα, 2 για την ακρίβεια με 0.6 και 5 για τον αρμονικό μέσο με 0.64. Η λογιστική παλινδρόμηση για όλες τις διαστάσεις δίνει το καλύτερο αποτέλεσμα στην ανάκληση με 0.72.

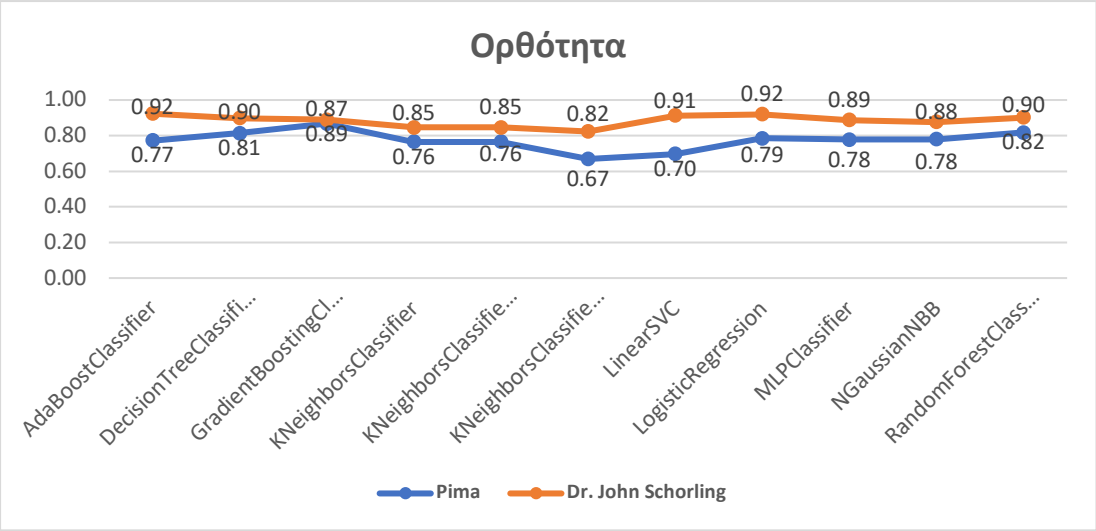
4.3.4.9 Εφαρμογή PCA, StandardScaler και διαγραφή εγγραφής με μηδενική τιμή



Εφαρμόζοντας τον αλγόριθμο PCA, StandardScaler και τέλος διαγράφοντας τις εγγραφές με έστω μια μηδενική τιμή, καλύτερο αποτέλεσμα για την ορθότητα είναι όταν το πλήθος των διαστάσεων είναι ίσο με 6 και δίνεται από τον GradientBoost με 0.8. Όσο αφορά την ακρίβεια, τα δέντρα απόφασης και συγκεκριμένα ο DecisionTreeClassifier και ο αριθμός διαστάσεων στο σύνολο δεδομένων να είναι ίσο με 3, δίνει τα καλύτερα αποτελέσματα με 0.68. Το τυχαίο δάσος όταν οι διαστάσεις του συνόλου είναι 5 δίνει το καλύτερο αποτέλεσμα στην ανάκληση με 0.73. Τέλος, ο GradientBoost έχει το καλύτερο αποτέλεσμα στον αρμονικό μέσο όταν οι διαστάσεις είναι 3 με τιμή 0.69.

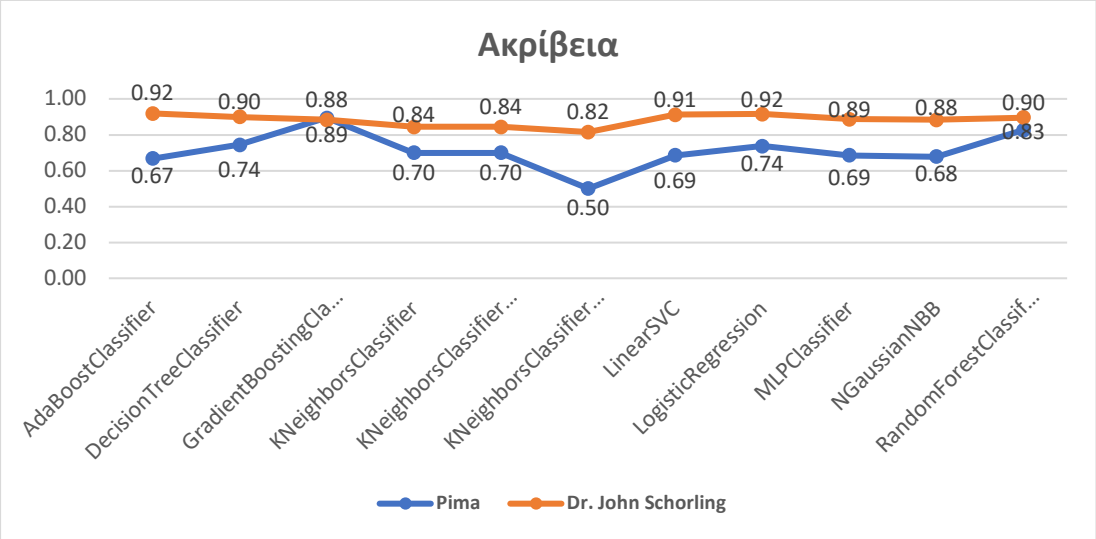
4.4 Συγκριτική αξιολόγηση των αλγορίθμων στα δύο σύνολα δεδομένων

Οι προαναφερθέντες αλγόριθμοι και διαδικασίες εφαρμόστηκαν και στο νέο σύνολο δεδομένων. Ακολούθως, όπως και στο αρχικό σύνολο δεδομένων Pima, έτσι και στο νέο σύνολο δεδομένων, έγινε η καταγραφή των αποτελεσμάτων της ορθότητας, της ακρίβειας, της ανάκλησης και του αρμονικού μέσου. Στις επόμενες γραφικές παραστάσεις παρουσιάζονται τα καλύτερα αποτελέσματα των πιο πάνω μέτρων αποτελεσματικότητας μετά την εφαρμογή της διαδικασίας και στα δύο σύνολα δεδομένων.



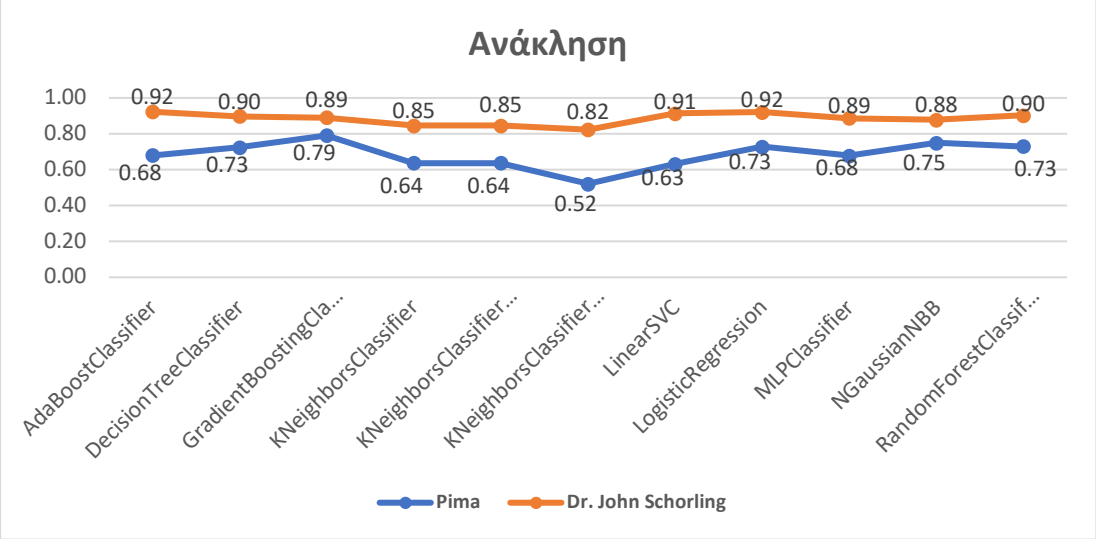
Γραφική Παράσταση 1: Συγκριτική αξιολόγηση ορθότητας μεταξύ των δύο συνόλων.

Καλύτερα αποτελέσματα στην ορθότητα για το σύνολο δεδομένων Pima έχουμε με τη χρήση του αλγορίθμου GradientBoostingClassifier με 0.89, του RandomForestClassifier με 0.82, του DecisionTreeClassifier με 0.8, του LogisticRegression με 0.79 και του NGaussianNBB με 0.78. Όσο αφορά το σύνολο δεδομένων του Δρ. John Schorling, το οποίο χρησιμοποιήθηκε, παρατηρούμε ότι καλύτερα αποτελέσματα παράγονται με την εφαρμογή του αλγορίθμου AdaBoostClassifier με 0.92, του LogisticRegression με 0.92, του LinearSVC με 0.91, του RandomForestClassifier με 0.90 και του DecisionTreeClassifier με 0.90.



Γραφική Παράσταση 2: Συγκριτική αξιολόγηση ακρίβειας μεταξύ των δύο συνόλων.

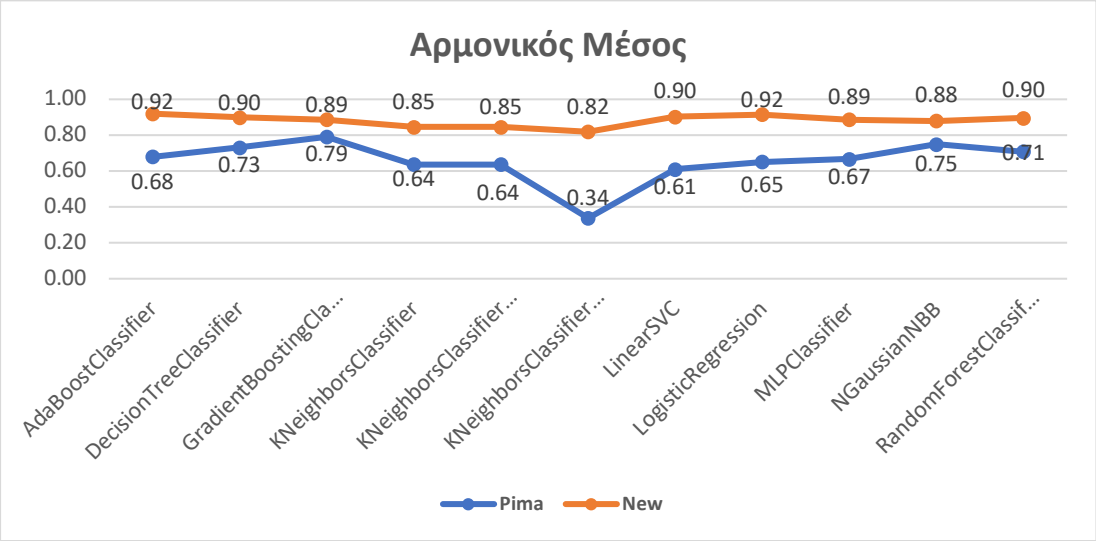
Τα καλύτερα αποτελέσματα που πάρθηκαν στο σύνολο δεδομένων Pima για την ακρίβεια ήταν με την εφαρμογή των αλγορίθμων GradientBoostingClassifier με 0.89, RandomForestClassifier με 0.83, DecisionTreeClassifier με 0.74, LogisticRegression με 0.74 και KNeighborsClassifier με απόσταση Minowski και Euclidean με 0.7. Για το σύνολο δεδομένων του Δρ. John Schorling, καλύτερα αποτελέσματα στην ακρίβεια προκύπτουν με την εφαρμογή των αλγορίθμων AdaBoost με 0.92, LogisticRegression με 0.92, LinearSVC με 0.91, DecisionTreeClassifier με 0.9 και RandomForestClassifier με 0.9.



Γραφική Παράσταση 3: Συγκριτική αξιολόγηση ανάκλησης μεταξύ των δύο συνόλων.

Όσο αφορά τα αποτελέσματα της ανάκλησης, καλύτερα αποτελέσματα στο σύνολο δεδομένων Pima έχει ο αλγόριθμος GradientBoostingClassifier με 0.79, NGaussianNBB με 0.75, RandomForestClassifier με 0.73, DecisionTreeClassifier με 0.73 και LogisticRegression με 0.7273. Από το σύνολο δεδομένων του Δρ. John Schorling προκύπτει ότι καλύτερα αποτελέσματα της ανάκλησης προκύπτουν εφαρμόζοντας τους αλγορίθμους AdaBoost με

0.92, LogisticRegression με 0.92, LinearSVC με 0.91, RandomForestClassifier με 0.90 και DecisionTreeClassifier με 0.9.



Γραφική Παράσταση 4: Συγκριτική αξιολόγηση αρμονικού μέσου μεταξύ των δύο συνόλων.

Τέλος, καλύτερα αποτελέσματα στον υπολογισμό του αρμονικού μέσου στο σύνολο δεδομένων Pima παρουσιάζουν οι αλγόριθμοι GradientBoostingClassifier με 0.79, NGaussianNBB με 0.75, DecisionTreeClassifier με 0.73, RandomForestClassifier με 0.71 και AdaBoost με 0.68. Στο σύνολο δεδομένων του Δρ. John Schorling καλύτερα αποτελέσματα είχαμε με τη χρήση του αλγορίθμου AdaBoost με 0.92, LogisticRegression με 0.92, LinearSVC με 0.90, DecisionTreeClassifier με 0.90 και RandomForestClassifier με 0.90.

Οι καλύτεροι αλγόριθμοι και για τα δύο σύνολα δεδομένων αναγράφονται στον επόμενο πίνακα αλφαριθμητικά:

ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ PIMA	**ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ ΔΡ. JOHN SCHORLING**
• AdaBoost • DecisionTreeClassifier • GradientBoostingClassifier • LogisticRegression • RandomForestClassifier	• AdaBoost • DecisionTreeClassifier • LinearSVC • LogisticRegression • RandomForestClassifier

Πίνακας 10: Καλύτεροι αλγόριθμοι στα δύο σύνολα δεδομένων.

Παρατηρούμε ότι και στα δύο σύνολα δεδομένων οι τέσσερις από τους πέντε αλγορίθμους, οι οποίοι παρουσιάζουν τα καλύτερα αποτελέσματα είναι οι ίδιοι. Αυτοί είναι οι

αλγόριθμοι AdaBoost, DecistionTreeClassifier, RandomForestClassifier και LogisticRegression. Ο αλγόριθμος ο οποίος διαφέρει στα σύνολα δεδομένων είναι ο GradientBoostingClassifier στο σύνολο δεδομένων Pima και LinearSVC νέο σύνολο δεδομένων.

Για τα αποτελέσματα της ορθότητας, ο αλγόριθμος GradientBoostingClassifier παράγει τα αποτελέσματά του, αφού πρώτα εφαρμοστεί η μέθοδος της μείωσης διαστάσεων PCA. Οι αλγόριθμοι RandomForestClassifier, DecisionTreeClassifier και NGaussianNBB παράγουν τα αποτελέσματά τους, αφού διαγραφούν οι εγγραφές με έστω μια μηδενική τιμή. Τέλος, ο αλγόριθμος LogisticRegression, τα καλύτερα του αποτελέσματα τα παράγει αφού εφαρμοστεί η τεχνική της αντικατάστασης της μηδενικής τιμής με τη μέση τιμή του πεδίου.

Τα καλύτερα αποτελέσματα της ακρίβειας, ο αλγόριθμος GradientBoostingClassifier, όπως και ο LinearRegression, παράγουν τα αποτελέσματα τους εφαρμόζοντας τη μέθοδο μείωσης διαστάσεων PCA, τη διαγραφή της μηδενικής τιμής και εφαρμόζοντας την μέθοδο MinMaxScaler ή StandardScaler. Ο αλγόριθμος RandomForestClassifier παράγει τα καλύτερα του αποτελέσματα με τη διαγραφή της εγγραφής με μια μηδενική τιμή και εφαρμογή της μεθόδου StandardScaler. Τέλος, ο αλγόριθμος DecisionTreeClassifier έχει καλύτερα αποτελέσματα εφαρμόζοντας τη μέθοδο μείωσης διαστάσεων PCA και διαγραφή εγγραφής με μια μηδενική τιμή.

Για το μέτρο της ανάκλησης, ο αλγόριθμος GradientBoostingClassifier παράγει τα καλύτερα του αποτελέσματα, διαγράφοντας την εγγραφή που περιέχει μια μηδενική τιμή. Οι αλγόριθμοι DecisionTreeClassifier και RandomForestClassifier, έχουν καλύτερα αποτελέσματα όταν εφαρμόζεται η τεχνική μείωσης διαστάσεων PCA, η διαγραφή εγγραφής με μια μηδενική τιμή και η εφαρμογή MinMaxScaler. Τέλος, ο αλγόριθμος LogisticRegression έχει καλύτερα αποτελέσματα στην ανάκληση, όταν εφαρμοστεί η μέθοδος MinMaxScaler ή StandardScaler.

Τέλος, τα καλύτερα αποτελέσματα για τον αρμονικό μέσο, παράγονται από τον GradientBoostingClassifier όταν εφαρμοστεί η μέθοδος μείωσης διαστάσεων PCA, η διαγραφή της μηδενικής τιμής και η εφαρμογή της μεθόδου MinMaxScaler ή StandardScaler. Οι αλγόριθμοι DecisionTreeClassifier και RandomForestClassifier, έχουν καλύτερα αποτελέσματα όταν εφαρμόζεται η τεχνική μείωσης διαστάσεων PCA, η διαγραφή εγγραφής

με μια μηδενική τιμή και η εφαρμογή MinMaxScaler. Τελευταίος αλγόριθμος, ο αλγόριθμος AdaBoost ο οποίος τα καλύτερα του αποτελέσματα τα παράγει αφού εφαρμοστεί η τεχνική της αντικατάστασης της μηδενικής τιμής με τη μέση τιμή του πεδίου και εφαρμόζοντας τη μέθοδο MinMaxScaler.

4.5 Επιλογή καλύτερου αλγορίθμου

Εφαρμόζοντας τη διαδικασία η οποία περιγράφηκε πιο πάνω, παρατηρούμε ότι καλύτερα αποτελέσματα στην ορθότητα για το σύνολο δεδομένων Pima Indian, δίνονται από τον αλγόριθμο GradientBoostingClassifier εφαρμόζοντας την τεχνική διαγραφής της εγγραφής στη φάση της προεπεξεργασίας, όπου περιέχει μια μηδενική τιμή. Αποτέλεσμα αυτού είναι η μείωση του συνόλου δεδομένων. Όμως, από την προεπεξεργασία αυτή, το σύνολο δεδομένων που παράγεται είναι αρκετά μικρό, με αποτέλεσμα να μην είναι αξιόπιστο κατά τη φάση της δοκιμής. Δεύτερος καλύτερος αλγόριθμος είναι η λογιστική παλινδρόμηση, όπου έχει καλύτερα αποτελέσματα εφαρμόζοντας τη τεχνική αντικατάστασης της χαμένης τιμής με τη μέση τιμή.

Στο σύνολο δεδομένων του Δρ. John Schorling, καλύτερα αποτελέσματα όσο αφορά την ορθότητα, έχει ο αλγόριθμος ενίσχυσης AdaBoost με τον αλγόριθμο της λογιστικής παλινδρόμησης να είναι και πάλι ο δεύτερος καλύτερος. Η διαφορά μεταξύ των δύο αλγορίθμων είναι 0.0025, η οποία είναι πολύ ελάχιστη.

Συγκρίνοντας έτσι τα πιο πάνω αποτελέσματα, καταλήγουμε στη επιλογή ότι καλύτερος αλγόριθμος είναι η λογιστική παλινδρόμηση σε συνδυασμό με την αντικατάσταση της χαμένης τιμής με τη μέση. Ο συνδυασμός αυτός, δίνει ένα ποσοστό ορθότητας της κατηγοριοποίησης του ασθενή στο 79%.

Κεφάλαιο 5: Συμπεράσματα και Μελλοντικές Κατευθύνσεις

5.1 Συμπεράσματα

Οι επιπτώσεις που υπάρχουν με την ανάπτυξη διαφόρων συστημάτων και εργαλείων μηχανικής μάθησης σε συνδυασμό με την εξόρυξη δεδομένων από διάφορους τομείς μπορούν να επιλύσουν διάφορα προβλήματα. Ένας τέτοιος κρίσιμος τομέας είναι ο ιατρικός τομέας, στον οποίο η κάθε λήψη απόφασης για την πρόβλεψη κάποιας ασθένειας, έχει μεγάλη σημασία.

Με την παρούσα μεταπτυχιακή διπλωματική εργασία, έγινε η μελέτη διαφόρων ήαλγορίθμων μηχανικής μάθησης και ακολούθως η σύγκριση τους στην ακριβή πρόβλεψη της ύπαρξης ή όχι σακχαρώδη διαβήτη σε ένα ασθενή. Οι αλγόριθμοι που μελετήθηκαν, υλοποιήθηκαν και πραγματοποιήθηκε η συγκριτική τους αξιολόγηση ως προς την ορθότητα, την ακρίβεια, την ανάκληση και τον αρμονικό μέσο είναι ο απλοϊκός Bayes, η λογιστική παλινδρόμηση, τα νευρωνικά δίκτυα, οι μηχανές διανυσμάτων υποστήριξης, τα δέντρα απόφασης, οι αλγόριθμοι συλλογικής μάθησης και ο Κ κοντινότεροι γείτονες.

Τα σύνολα δεδομένων τα οποία χρησιμοποιήθηκαν είναι τα ίδια με σύνολα δεδομένων από τις μελέτες που παρουσιάστηκαν στο κεφάλαιο 4. Ως κύριο σύνολο δεδομένων χρησιμοποιήθηκε το Pima Indian, ενώ το σύνολο δεδομένων από τον Δρ. John Schorling χρησιμοποιήθηκε προκειμένου να υπάρχουν περισσότερα στοιχεία ως προς την ακρίβεια του μοντέλου που επιλέχθηκε.

Η πρώτη μελέτη χρησιμοποίησε 5 αλγορίθμους για υπολογισμό της ακρίβειας, καθώς επίσης εφαρμόστηκαν και μέθοδοι προεπεξεργασίας των δεδομένων με τεχνικές αντικατάστασης χαμένων τιμών με τη μέση τιμή, κανονικοποίηση και τυποποίηση. Από τα προεπεξεργασμένα δεδομένα προκύπτουν καλύτερα αποτελέσματα. Το σύνολο που χρησιμοποιήθηκε ήταν το Pima Indian και η τεχνική για το διαχωρισμό του συνόλου δεδομένων σε σύνολο εκπαίδευσης και δοκιμής ήταν η K – Folds Cross Validation με αριθμό K = 10. Στη συγκεκριμένη μεταπτυχιακή διπλωματική εργασία, χρησιμοποιήθηκαν τέσσερις από τους πέντε ίδιους αλγορίθμους για το σύνολο των δεδομένων Pima, από όπου παράγονται αποτελέσματα με τον αλγόριθμο λογιστικής παλινδρόμησης με 0.79. Αυτό οφείλεται στο ότι εφαρμόστηκαν και οι δύο μέθοδοι διαχωρισμού του συνόλου δεδομένων, με καλύτερα αποτελέσματα να δίνει η μέθοδος Split σε σχέση με τον K – Folds Cross Validation.

Στη δεύτερη μελέτη χρησιμοποιήθηκαν και πάλι πέντε αλγόριθμοι, όπου έγινε η σύγκρισή τους πριν και μετά τη φάση της προεπεξεργασίας. Κατά τη φάση της προεπεξεργασίας έγινε η διαγραφή όλων των εγγραφών με έστω μια μηδενική τιμή. Καλύτερα αποτελέσματα πάρθηκαν με τα προεπεξεργασμένα δεδομένα και με τον αλγόριθμο Decision Table με 85.2% ποσοστό ακριβείας. Στη διπλωματική εργασία, εφαρμόστηκαν οι τέσσερις από τους πέντε αλγόριθμους. Τα καλύτερα αποτελέσματα που παράγονται με τη εφαρμογή της ίδιας τεχνικής κατά τη φάση της προεπεξεργασίας, είναι με τον αλγόριθμο GradientBoost με ποσοστό ακρίβειας 87%.

Η τρίτη μελέτη, όπως και στις πιο πάνω μελέτες, χρησιμοποίησε πέντε αλγορίθμους και έγινε η σύγκρισή τους πριν και μετά τη φάση της προεπεξεργασίας, όπου εφαρμόστηκε ένας γενετικός αλγόριθμος για τη μείωση των διαστάσεων του συνόλου δεδομένων. Εφαρμόστηκε η τεχνική K – Fold Cross Validation με αριθμό K = 10 με καλύτερα αποτελέσματα να προκύπτουν στα προεπεξεργασμένα δεδομένα με τον αλγόριθμο Multi – Objective NSGA II με 83.04% ποσοστό ακριβείας. Η μελέτη αυτή παράγει καλύτερα αποτελέσματα σε σχέση με τις μεθόδους που εφαρμόστηκαν στη παρούσα διπλωματική εργασία.

Η τελευταία μελέτη συνδυάζει μη επιβλεπόμενη μάθηση και επιβλεπόμενη μάθηση. Αρχικά χρησιμοποιεί την τεχνική της αντικατάστασης των χαμένων τιμών με τη μέση τιμή. Στη συνέχεια εφαρμόζει τον αλγόριθμο συσταδοποίησης K Means ο οποίος πρόκειται για αλγόριθμο μη επιβλεπόμενης μάθησης. Τέλος, γίνεται η εφαρμογή της λογιστικής παλινδρόμησης με ποσοστό ακριβείας 95.42% στο σύνολο δεδομένων Pima Indian και 93.7% στο σύνολο δεδομένων του Δρ. John Schorling. Συγκρίνοντας τα αποτελέσματα από το Pima Indian, η μελέτη αυτή έχει πολύ καλύτερα αποτελέσματα σε σχέση με τα αποτελέσματα της συγκεκριμένης διπλωματικής εργασίας αφού το καλύτερο ποσοστό ακρίβειας είναι 79% στον αλγόριθμο της λογιστικής παλινδρόμησης. Στο σύνολο δεδομένων του Δρ. John Schorling όμως καλύτερα αποτελέσματα έχει η παρούσα διπλωματική εργασία στη λογιστική παλινδρόμηση με ποσοστό ακριβείας στο 92%.

Παρουσιάζοντας τις διάφορες διεθνείς μελέτες οι οποίες έγιναν, καθώς επίσης και με τα όσα παρουσιάστηκαν και εφαρμόστηκαν στην παρούσα διπλωματική εργασία γίνεται εύκολα αντιληπτό ότι η φάση της προεπεξεργασίας παίζει σημαντικό ρόλο στα αποτελέσματα που προκύπτουν από τον κάθε αλγόριθμο. Τέλος, η τεχνική η οποία θα εφαρμοστεί κατά τη

φάση του διαχωρισμού του συνόλου δεδομένων σε σύνολο εκπαίδευσης και σύνολο δοκιμής παίζει και αυτή σημαντικό ρόλο, αφού παρατηρούμε ότι η ίδια τεχνική προεπεξεργασίας δίνει διαφορετικά αποτελέσματα με βάση το διαχωρισμό που θα γίνει στα δύο υποσύνολα.

5.2 Μελλοντικές κατευθύνσεις

Η παρούσα διπλωματική εργασία σκοπό είχε τη συγκριτική αξιολόγηση των κατηγοριοποιητών για την πρόβλεψη του σακχαρώδη διαβήτη τύπου 2.

Επίσης, ο συνδυασμός των δύο μαθήσεων, επιβλεπόμενης και μη επιβλεπόμενης, όπως προτάθηκε στην τέταρτη μελέτη που παρουσιάστηκε στο κεφάλαιο 4 από τους Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang καθώς επίσης και από τους Savvas Karatsiolis και Christos N. Schizas στη μελέτη τους με τίτλο *Region based Support Vector Machine Algorithm for Medical Diagnosis on Pima Indian Diabetes DataSet* [20], πιθανόν να δώσει καλύτερα αποτελέσματα.

Τέλος η μελέτη περισσότερων αλγορίθμων κατηγοριοποίησης και διαφόρων τεχνικών προεπεξεργασίας των δεδομένων θα μπορούσε να βελτιώσει περισσότερο την ακρίβεια της διάγνωσης της ασθένειας του σακχαρώδη διαβήτη. Μια εναλλακτική λύση στην αντικατάσταση της τιμής, θα ήταν η χρήση ενός μοντέλου μηχανικής μάθησης για πρόβλεψη της χαμένης τιμής, όπως η γραμμική παλινδρόμηση ή τα δέντρα παλινδρόμησης.

Κεφάλαιο 6: Βιβλιογραφία

[1] PAVLYSHENKO, B. 2016. MACHINE LEARNING, LINEAR AND BAYESIAN MODELS FOR LOGISTIC REGRESSION IN FAILURE DETECTION PROBLEMS. [Online] pp. 2046 - 2050. Available from: https://ieeexplore.ieee.org/document/7840828 [Accessed 28th December 2018].

[2] SELVAKUBERAN, K., KAYATHIRI, D., HARINI, B., DEVI, INDRA M., 2011. AN EFFICIENT FEATURE SELECTION METHOD FOR CLASSIFICATION IN HEALTH CARE SYSTEMS USING MACHINE LEARNING TECHNIQUES. [Online] pp. 223 - 226. Available from: https://ieeexplore.ieee.org/document/5941891 [Accessed 28th December 2018].

[3] POLAT, K., GUNES, S., 2007. AN EXPERT SYSTEM APPROACH BASED ON PRINCIPAL COMPONENT ANALYSIS AND ADAPTIVE NEURO-FUZZY INFERENCE SYSTEM TO DIAGNOSIS OF DIABETES DISEASE. [Online] 17.(4), pp. 702 - 710. Available from: https://www.sciencedirect.com/science/article/pii/S1051200406001370 [Accessed 28th December 2018].

[4] POLAT, K., GUNES, S., ARSLAN, A., 2008. A CASCADE LEARNING SYSTEM FOR CLASSIFICATION OF DIABETES DISEASE: GENERALIZED DISCRIMINANT ANALYSIS AND LEAST SQUARE SUPPORT VECTOR MACHINE. [Online] 34. (1), pp. 482 - 487. Available from: https://www.sciencedirect.com/science/article/pii/S0957417406002995 [Accessed 28th December 2018].

[5] ZEHRA, A., ASMAWATY, T., AZNAN, M.A. M., 2013. A COMPARATIVE STUDY ON THE PRE-PROCESSING AND MINING OF PIMA INDIAN DIABETES DATASET. [Online] pp. 1 - 10. Available from: http://umpir.ump.edu.my/id/eprint/5035/1/31-UMP.pdf [Accessed 28th December 2018].

[6] KALYANKAR, G. D., POOJARA, SHIVANANDA R., DHARWADKAR, NAGARAJ V., 2017. PREDICTIVE ANALYSIS OF DIABETIC PATIENT DATA USING MACHINE LEARNING AND HADOOP. [Online] pp. 619 - 624. Available from: https://ieeexplore.ieee.org/document/8058253 [Accessed 28th December 2018].

[7] WU, H., YANG, SHENGQI., HUANG, ZHANGQIN., HE, JIAN., WANG, XIAOYI., 2018. TYPE 2 DIABETES MELLITUS PREDICTION MODEL BASED ON DATA MINING. [Online] 10. pp. 100 - 107. Available from:

https://www.sciencedirect.com/science/article/pii/S2352914817301405 [Accessed 28th December 2018].

[8] BARAKAT, N., ANDREW, BRADLEY P., BARAKAT, MOHAMED N., 2010. INTELLIGIBLE SUPPORT VECTOR MACHINES FOR DIAGNOSIS OF DIABETES MELLITUS. [Online] 14. (4), pp. 1114 - 1120. Available from: https://ieeexplore.ieee.org/document/5378519 [Accessed 28th December 2018].

[9] MULLER, A. C., GUIDO, SARAH (2017). *INTRODUCTION TO MACHINE LEARNING WITH PYTHON: A GUIDE FOR DATA SCIENTISTS,* O'REILLY MEDIA, INC., 1005 GRAVENSTEIN HIGHWAY NORTH, SEBASTOPOL, CA 95472.

[10] MITCHELL, T. M. (1997). *MACHINE LEARNING,* MCGRAW-HILL SCIENCE/ENGINEERING/MATH.

[11] AWAD. MARIETTE, K. R. (2015). *EFFICIENT LEARNING MACHINES: THEORIES, CONCEPTS, AND APPLICATIONS FOR ENGINEERS AND SYSTEM DESIGNERS,* APRESS, BERKELEY, CA.

[12] RASCHKA, S. (2015). *UNLOCK DEEPER INSIGHTS INTO MACHINE LEARNING WITH THIS VITAL GUIDE TO CUTTING-EDGE PREDICTIVE ANALYTICS,* PACKT.

[13] ΧΑΛΚΙΔΗ, Μ., ΒΑΖΙΡΓΙΑΝΝΗΣ, Μ (2005). *ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ,* ΤΥΠΩΘΗΤΩ.

[14] CHRISTOPHER D. MANNING, P. R., HINRICH SCHUTZE (2008). *INTRODUCTION TO INFORMATION RETRIEVAL,* UNITED STATES OF AMERICA BY CAMBRIDGE UNIVERSITY PRESS, NEW YORK.

[15] RUSSEL, S., NORVIG, PETER (2005). *ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ ΜΙΑ ΣΥΓΧΡΟΝΗ ΠΡΟΣΕΓΓΙΣΗ ΔΕΥΤΕΡΗ ΕΚΔΟΣΗ,* ΚΛΕΙΔΑΡΙΘΜΟΣ.

[16] BHAT, V. H., RAO, PRASANTH G., SHENOY, P. D., VENUGOPAL, K.R., PATNAIK, L.M., 2009. AN EFFICIENT PREDICTION MODEL FOR DIABETIC DATABASE USING SOFT COMPUTING TECHNIQUES [Online] 5908. pp. 328 - 335. Available from: https://link.springer.com/chapter/10.1007/978-3-642-10646-0_40 [Accessed 28th December 2018].

[17] GRUS, J. (2015). *DATA SCIENCE FROM SCRATCH,* O'REILLY MEDIA, INC., 1005 GRAVENSTEIN HIGHWAY NORTH, SEBASTOPOL, CA 95472.

[18] MARTINEZ, A. M. 2001. PCA versus LDA. [Online] 23. pp. 228 - 233. Available from: https://ieeexplore.ieee.org/abstract/document/908974 [Accessed 15th March 2018].

[19] WEI, S., ZHAO, XUEJIAO., MIAO, CHUNYAN 2018. A COMPREHENSIVE EXPLORATION TO THE MACHINE LEARNING TECHNIQUES FOR DIABETES IDENTIFICATION. [Online] pp. 291 - 295. Available from: https://ieeexplore.ieee.org/document/8355130 [Accessed 28th December 2018].

[20] KARATSIOLIS, S., SCHIZAS, CHRISTOS N., 2012. REGION BASED SUPPORT VECTOR MACHINE ALGORITHM FOR MEDICAL DIAGNOSIS ON PIMA INDIAN DIABETES DATASET [Online] pp. 139 - 144. Available from: https://ieeexplore.ieee.org/document/6399663 [Accessed 28th December 2018].

[21] BRADLEY, A. P. 1996. THE USE OF THE AREA UNDER THE ROC CURVE IN THE EVALUATION OF MACHINE LEARNING ALGORITHMS. [Online] 30. (7), pp. 1145 - 1159. Available from: https://www.sciencedirect.com/science/article/pii/S0031320396001422 [Accessed 28th December 2018].

[22] KAYAER, K., YILDIRIM, T., MEDICAL DIAGNOSIS ON PIMA INDIAN DIABETES USING GENERAL REGRESSION NEURAL NETWORKS. [Online] Available from: https://pdfs.semanticscholar.org/ef31/2e378325707b371c4727f6b1f9225fc03a9f.pdf [Accessed 28th December 2018].

[23] BRZEZINSKI, J. R., KNAFL, G.J., 1999. LOGISTIC REGRESSION MODELING FOR CONTEXT-BASED CLASSIFICATION. [Online] Available from: https://ieeexplore.ieee.org/document/795279 [Accessed 28th December 2018].

[24] VAISHALI, R., SASIKALA, R., RAMASUBBAREDDY, S., REMYA, S., NALLURI, S., 2017. GENETIC ALGORITHM BASED FEATURE SELECTION AND MOE FUZZY CLASSIFICATION ALGORITHM ON PIMA INDIANS DIABETES DATASET [Online] Available from: https://ieeexplore.ieee.org/document/8123815 [Accessed 28th December 2018].

[25] CHOUBEY, D. K., PAUL, SANCHITA., KUMAR, SANTOSH., 2017. CLASSIFICATION OF PIMA INDIAN DIABETES DATASET USING NAIVE BAYES WITH GENETIC ALGORITHM AS AN ATTRIBUTE SELECTION. [Online] pp. 451 - 455. Available from: http://www.academia.edu/31707552/Classification_of_Pima_indian_diabetes_dataset_using_naive_bayes_with_genetic_algorithm_as_an_attribute_selection [Accessed 28th December 2018].

[26] BARALE, M. S., SHIRKE, D.T., 2016. CASCADED MODELING FOR PIMA INDIAN DIABETES DATA. [Online] 139. (11), pp. 1 - 4. Available from:

https://www.researchgate.net/profile/Mahesh_Barale/publication/301335647_Cascaded_Modeling_for_PIMA_Indian_Diabetes_Data/links/59c0ba97a6fdcca8e5724d61/Cascaded-Modeling-for-PIMA-Indian-Diabetes-Data.pdf [Accessed 28th December 2018].

[27] KAREGOWDA, A. G., MANJUNATH, A.S., JAYARAM, M.A., 2011. APPLICATION OF GENETIC ALGORITHM OPTIMIZED NEURAL NETWORK CONNECTION WEIGHTS FOR MEDICAL DIAGNOSIS OF PIMA INDIANS DIABETES. [Online] 2. (2), pp. 15 - 23. Available from: https://www.researchgate.net/publication/228983015_Application_of_Genetic_Algorithm_Optimized_Neural_Network_Connection_Weights_for_Medical_Diagnosis_of_PIMA_Indians_Diabetes [Accessed 28th December 2018].