Πανεπιστήμιο Πειραιώς

Τμήμα Ψηφιακών Συστημάτων

Πρόγραμμα Μεταπτυχιακών Σπουδών «Ψηφιακά Συστήματα και Υπηρεσίες»

Κατεύθυνση
«Μεγάλα Δεδομένα και Αναλυτική»

Διπλωματική Εργασία

με θέμα:

«Μελέτη και Σύγκριση Αλγόριθμων Εξόρυξης»

Καθηγητής
Δρ. Μιχαήλ Φιλιππάκης, Αναπληρωτής Καθηγητής

Παναγόπουλος Θεόδωρος (ΜΕ 1723)

Πειραιάς 2020

Περιεχόμενα

Λίστα εικόνων

Λίστα Γραφημάτων

Λίστα Πινάκων

1 Εισαγωγή

Η εξόρυξη δεδομένων έχει αποτελέσει έναν από τις σημαντικότερους τρόπους για την επεξεργασία δεδομένων τα τελευταία χρόνια, με σκοπό την εξαγωγή προτύπων για την παραγωγή νέων δεδομένων-πληροφοριών αλλά και την λήψη αποφάσεων. Πολλές φορές η ανθρώπινη άγνοια αλλά και η αδυναμία περίπλοκης σύνδεσης μεγάλου όγκου δεδομένων, δεν καθιστούν τις αναγκαίες αποφάσεις εμφανείς και εφικτές, αντίστοιχα. Σε κάθε τομέα όπου η μελέτη των δεδομένων του παρελθόντος μπορεί να αναγνωρίσει μελλοντικά περιστατικά, στα οποία χρειάζεται να παρθούν ανάλογες αποφάσεις, η εξόρυξη δεδομένων δίνει το ζητούμενο. Κάθε τομέας περιλαμβάνει μοτίβα, όπως ο χρηματιστηριακός, ο τομέας υγείας, η επιστήμη, ο εργασιακός τομέας, το μάρκετινγκ, οι πωλήσεις, η παρακολούθηση λειτουργίας μηχανημάτων, και άλλοι. Μικρό ποσοστό όμως αυτών χρησιμοποιούν την εξόρυξη δεδομένων για την λήψη μελλοντικών αποφάσεων, αν και όλοι εφαρμόζουν ήδη τρόπους για την καταγραφή μεγάλου όγκου πληροφοριών.

Ο όγκος των πληροφοριών μεγαλώνει με ρυθμούς οι οποίοι τις καθιστούν δύσκολο ο μέσος άνθρωπος να κατανοήσει. Χιλιάδες αισθητήρες, ψηφιακοί (διαδίκτυο, εφαρμογές, κ.α.) ή φυσικοί (Internet Of Things-IOT, κ.α.), έχουν σαν αποτέλεσμα την παραγωγή πάνω από 2,5 εκατομμύρια terabytes δεδομένων κάθε λεπτό, ρυθμός που έχει τη παραγωγή του 90% του όγκου των δεδομένων του πλανήτη να έχει πραγματοποιηθεί τα τελευταία δύο χρόνια, [38]. Ο ραγδαίος αυτός ρυθμός και ο ήδη υπάρχον όγκος δεδομένων, καθιστά πιο αισθητή την ανάγκη για χρήση νέων τεχνικών για την μετατροπή όγκων δεδομένων, σε χρήσιμές πληροφορίες.

Παράδειγμα της μετατροπής όγκου δεδομένων σε χρήσιμη πληροφορία, αποτελεί το πρόγραμμα EOS, Earth Observing System, της Nasa. Στο πρόγραμμα αυτό, το οποίο είναι σε λειτουργία από το 1997 με την επιτυχή εκτόξευση του πρώτου δορυφόρου, ένα σύμπλεγμα δορυφόρων σε τροχιά γύρω από την γη, παράγει δεδομένα σε ρυθμούς gigabyte ανά λεπτό σχετικά με τον πλανήτη. Παρόμοιοι τρόποι και ρυθμοί παραγωγής δεδομένων απαιτούν την χρήση της εξόρυξης δεδομένων με σκοπό την αναγνώριση προτύπων, την οργάνωση δεδομένων και ανάλυση αυτών για την πραγματοποίηση προβλέψεων, [54].

Ακόμα ένα παράδειγμα της μετατροπής δεδομένων σε πληροφορία αποτελεί η εταιρία “The Echo Nest”, [16], όπου συνεχώς αναλύονται χαρακτηριστικά των τραγουδιών που παράγονται με σκοπό την οργάνωσή τους και την αξιοποίησή τους ως μέθοδο κατανόησης της μουσικής αλλά και των θαυμαστών της μουσικής.

Στον τομέα των επιχειρήσεων, η πραγματοποίηση προβλέψεων και ο αντίστοιχος προγραμματισμός είναι μεγάλης σημασίας καθώς ο ανταγωνισμός αυξάνεται. Σε περιβάλλον παραγωγής προϊόντων ειδικότερα, η πρόβλεψη της ζήτησης/πωλήσεων για τον προγραμματισμό της προσφοράς/γραμμής παραγωγής, και λήψης οικονομικών αποφάσεων μπορούν να επωφεληθούν, [7]. Η διαχείριση επιχειρησιακών διεργασιών, στοχεύει τα προβλήματα των διεργασιών, των ανθρωπίνων παραγόντων που τις καταρτίζουν, του προγραμματισμού των υλικών απαιτήσεων μιας παραγωγής, μέσω της χρήσης μεθόδων εξόρυξης δεδομένων προβλέποντας κόστη, και συσχετίσεις που δεν είναι εύκολα διαχωρίσιμες, [22].

Οι παραπάνω τομείς τέθηκαν στο προσκήνιο για την αναγνώριση της σημαντικότητας της χρήσης τεχνικών εξόρυξης δεδομένων. Τέτοια είναι όμως η φύση των τομέων εφαρμογής τους, όπου δεν είναι μόνο αυτοί με τις ιδιάζουσες μορφές δεδομένων που προσφέρουν, αλλά και ο εκάστοτε αναλυτής, που θα κρίνει σκόπιμη τη χρήση της μεθόδου Α, ή της Β.

Οι στόχοι της εξόρυξης δεδομένων τίθενται από τον αναλυτή, όπου σκοπός του είναι η εύρεση κατάλληλης μεθόδου για την επίτευξη αυτών.

1.1 Δομή διπλωματικής εργασίας

Η παρούσα διπλωματική εργασία αποτελείται από πέντε κεφάλαια. Στο πρώτο κεφάλαιο πραγματοποιείται η εισαγωγή, δίνοντας την έννοια της εξόρυξης δεδομένων και του σκοπού της, παρουσίαση κάποιων τομέων εφαρμογής, αλλά και στόχος συνεισφοράς της εργασίας αυτής. Το δεύτερο κεφάλαιο παρουσιάζεται μια βιβλιογραφική επισκόπηση της εξόρυξης δεδομένων, ανάλυση των τεχνικών που στοχεύουν στην πρόβλεψη, και κριτήριων που χρησιμοποιήθηκαν σε άλλες έρευνες. Η μεθοδολογία που ακολουθείται καθώς και τα δεδομένα στα οποία θα εφαρμοσθεί παρουσιάζεται στο τρίτο κεφάλαιο, με το τέταρτο κεφάλαιο να ακολουθεί με την παράθεση των αποτελεσμάτων των τεχνικών που εφαρμόσθηκαν. Στο πέμπτο κεφάλαιο δίνεται μία σύνοψη στα αποτελέσματα, βάση των οποίων πραγματοποιείται σύγκριση των τεχνικών με σκοπό την εξαγωγή συμπερασμάτων, και στοχοθέτηση για περαιτέρω έρευνα.

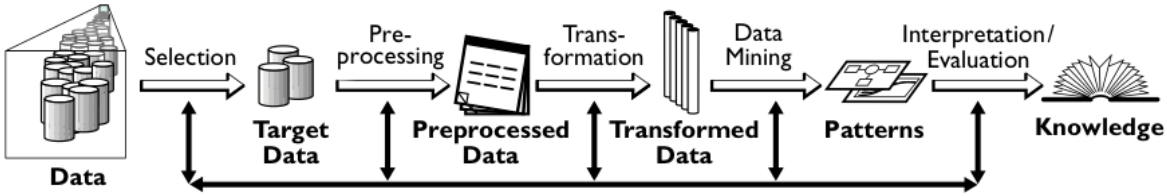
1.2 Ανασκόπηση

Στο εγχειρίδιο “Data Mining Practical Machine Learning Tools and Techniques”, από τους Ian H. Witten και Eibe Frank, [56], η εξόρυξη δεδομένων περιγράφεται ως η διαδικασία αναζήτησης προτύπων σε δεδομένα, διαδικασία η οποία μπορεί να πραγματοποιηθεί αυτοματοποιημένα πλήρως ή μερικώς. Σκοπός του αποτελέσματος της διαδικασίας αυτής είναι να οδηγήσει σε κάποιο πλεονέκτημα συνήθως οικονομικό.

Στο “International Journal for Research In Advanced Computer Science And Engineering”, [53], οι Shruti Upadhyay , Neha Patel, Rakesh Patel, Prateek Kumar Singh, αναφέρουν πως η εξόρυξη δεδομένων συνδέεται σαφώς με την Διαδικασία Ανακάλυψη Γνώσης, “Knowledge Discovery in Databases - KDD”, ως τρόπος διεξαγωγής πληροφοριών που ήταν άγνωστες από τα δεδομένα.

Με όμοιο τρόπο ο Usama M. Fayyad, στο "Data-Mining and Knowledge Discovery: Making Sense Out of Data", [22], δίνει την KDD ως την γενική κατηγορία κάτω από την οποία βρίσκονται όλες οι μέθοδοι των οποίων ο σκοπός είναι η

ανακάλυψη σχέσεων ανάμεσα στα δεδομένα, με την εξόρυξη δεδομένων να αποτελεί ένα από τα στάδια, όπως φαίνεται στην Εικόνα 1.



Εικόνα 1 Διαδικασία KDD

Η πρώτη αναφορά του όρου “Knowledge Discovery in Databases - KDD” παρουσιάζεται από τον Usama M. Fayyad, στο πρώτο συνέδριο με τίτλο “International Conference on Knowledge Discovery and Data Mining”, το 1995, στο Μόντρεαλ.

Τα στάδια από τα οποία αποτελείται η διαδικασία Knowledge Discovery in Databases – KDD, σύμφωνα με τον Fayyad, φαίνονται στο παραπάνω σχήμα. Στο πρώτο στάδιο περιλαμβάνεται η επιλογή ενός συνόλου δεδομένων ή την εστίαση σε ένα υποσύνολο μεταβλητών ή δειγμάτων δεδομένων. Στην συνέχεια τα δεδομένα καθαρίζονται και προεπεξεργάζονται μέσω βασικών λειτουργιών όπως η αφαίρεση θορύβου και ακραίων τιμών αν κρίνεται αναγκαία. Επίσης αντιμετωπίζονται θέματα όπως ελλιπή και άγνωστα δεδομένα καθώς και μετασχηματίζεται ο τύπος δεδομένων. Έπειτα τα δεδομένα υπόκεινται σε μεθόδους μείωσης ή μετασχηματισμών των διαστάσεών τους με στόχο την μείωση του πραγματικού υπό ανάλυση μεταβλητών και κατάλληλη προβολή τους με σκοπό την εύρεση χρήσιμων χαρακτηριστικών για την αναπαράστασή τους σύμφωνα με τον στόχο της ανάλυσης. Έπειτα η επιλογή της λειτουργίας της εξόρυξης δεδομένων περιλαμβάνει την απόφαση του στόχου του μοντέλου που προκύπτει από τον αλγόριθμο εξόρυξης δεδομένων δηλαδή ταξινόμηση, παλινδρόμηση ή ομαδοποίηση κ.α.), και εν συνεχεία επιλέγεται ο αλγόριθμος εξόρυξης δεδομένων όπου περιλαμβάνει την επιλογή μεθόδων αναζήτησης μοτίβων στα δεδομένα (δεν είναι κατάλληλοι όλοι οι μέθοδοι για όλους τους τύπους δεδομένων. Στο στάδιο της εξόρυξης δεδομένων περιλαμβάνεται η αναζήτηση μοντέλων ενδιαφέροντος στα δεδομένα σε μία παραστατική μορφή ή σύνολο γραφικών παραστάσεων, όπως των κανόνων ταξινόμησης ή των δένδρων,

την παλινδρόμηση, την ομαδοποίηση, την τροποποίηση αλληλουχίας, την εξάρτηση και την ανάλυση γραμμής. Η ερμηνεία των ανακαλυφθέντων μοτίβων έχει την ενδεχόμενη επιστροφή σε οποιοδήποτε από τα προηγούμενα βήματα, καθώς και την πιθανή απεικόνιση των μοτίβων, την απομάκρυνση περιττών ή μη σχετικών μοτίβων και τη μετάφραση των χρήσιμων σε όρους κατανοητούς από τους χρήστες. Τέλος η ανακαλυφθείσα γνώση ενσωματώνεται στο σύστημα λήψης δράσεων, αναφέρεται στους ενδιαφερόμενους, καθώς και ελέγχεται η επίλυση δυνητικών συγκρούσεων με τις προηγούμενες γνώσεις ή αποτελέσματα KDD.

Επιπροσθέτως στο "Applied Data Mining-Statistical Methods for Business and Industry", ο Paolo Giudici, [23], παρουσιάζει την ιστορικότητα του όρου Knowledge Discovery in Databases – KDD. Αρχικά παρουσιάζεται ως το σύνολο μεθόδων εύρεσης σχέσεων και προτύπων μεταξύ των δεδομένων, και η ο όρος επεκτείνεται στην συνέχεια να εκφράζει ολόκληρη την διαδικασία εξόρυξης πληροφοριών, με τον όρο εξόρυξη δεδομένων να περιγράφει ένα από τα στάδια της διαδικασίας KDD, στο οποίο πραγματοποιείται η εφαρμογή των αλγόριθμων εκμάθησης.

Η εξόρυξη δεδομένων αποτελείται από περιγραφικές μεθόδους και μεθόδους πρόβλεψης. Οι περιγραφικές μέθοδοι στοχεύουν στην περιγραφή συνόλων δεδομένων, σύνολα τα οποία μπορεί να ήταν ή και όχι γνωστά, όπου οι μεταβλητές μπορεί να συνδέονται μεταξύ τους με τρόπους που μπορεί να μην ήταν γνωστοί στο παρελθόν. Οι μέθοδοι αυτοί είναι συμμετρικοί, και δεν χρειάζονται επιτήρηση (unsupervised), με παραδείγματα να αποτελούν οι κανόνες συσχέτισης , κ.α. Οι μέθοδοι πρόβλεψης στοχεύουν στην πρόβλεψη μίας ή περισσοτέρων μεταβλητών συναρτήσει των υπολοίπων μεταβλητών, είναι ασύμμετροι μέθοδοι, και χρίζουν επιτήρησης (supervised). Μέθοδοι πρόβλεψης έχουν αναπτυχθεί στο πλαίσιο της μηχανικής μάθησης, τομέας που πολλές φορές συγχέεται με την εξόρυξη δεδομένων, όπως τα νευρωνικά δίκτυα. Τεχνικές μέθοδοι πρόβλεψης αποτελούν και τα δέντρα απόφασης και επίσης κλασικά στατιστικά μοντέλα όπως η γραμμική παλινδρόμηση.

Στο πλαίσιο της προβλεπτικής εξόρυξης δεδομένων, η μέθοδος που ακολουθείται είναι όμοια με αυτή του ανθρωπίνου εγκεφάλου, με την διαφορά ότι οι προβλεπτικές μέθοδοι μπορούν να εφαρμοστούν σε μεγάλο όγκο δεδομένων και να «μάθουν» από παρελθοντικά δεδομένα, περιορισμοί που η ανθρώπινη πλευρά παρουσιάζει.

Καθώς η εξόρυξη δεδομένων είναι μια σχετικά νέα κατεύθυνση για την ανάλυση του μεγάλου όγκου δεδομένων που έχει συσσωρευτεί, χώρα λαμβάνουν και προκλήσεις. Ο όγκος των δεδομένων δεν γίνεται μικρότερος, και η επιλογή κατάλληλης μεθόδου εξόρυξης των δεδομένων είναι ζωτικής σημασίας.

Η επιχείρηση δεν γίνεται ευκολότερη αφού τα δεδομένα δημιουργούνται από ένα μη τέλειο κόσμο, και επομένως δεν είναι πλήρη, έχουν διπλοεγγραφές, ή στον αντίποδα είναι μικρά σε μέγεθος για κάποιους σκοπούς. Εργαλεία εξόρυξης δεδομένων υπάρχουν, που όμως λειτουργούν με δομημένα δεδομένα, γεγονός ατυχές καθώς η πλειοψηφία των δεδομένων είναι μη δομημένη, όπως τα δεδομένα του παγκοσμίου ιστού (WWW). Αν και η ύπαρξη του τεράστιου όγκου δεδομένων στον παγκόσμιο ιστό έχει αρχίσει να περιμαζεύεται (scraping) με κατάλληλα εργαλεία (π.χ. Selenium), κανένα ακόμα εργαλείο δεν έχει κατασκευαστεί για την επίδειξη της κατάλληλης μεθόδου εξόρυξης, και της διαχείρισης δυναμικών ή και κακής ποιότητας δεδομένων. Αυτές οι ενέργειες επιβαρύνουν ακόμα και σήμερα τους αναλυτές που αναπτύσσουν μεθόδους εξόρυξης δεδομένων.

1.3 Αναγνώριση προβλήματος - Στόχος εργασίας

Η προβλεπτική εξόρυξη δεδομένων αποτελεί την πτυχή της εξόρυξης δεδομένων με την μεγαλύτερη επιβράβευση, την πρόβλεψη. Όπως επισημάνθηκε παραπάνω σε όλες τις πτυχές της εξόρυξης δεδομένων, η επιλογή της κατάλληλης μεθόδου εξόρυξης σε ένα πακέτο δεδομένων επιβαρύνει εξ' ολοκλήρου τον αναλυτή και τον βαθμό στον οποίο έχει κατανοήσει τον σκοπό της ανάλυσης και την πρώτη ύλη (δεδομένα) που έχει στην κατοχή του.

Η διαδικασία αυτή έχει μεγάλο κόστος υπολογιστικών πόρων και χρόνου, καθώς η κατάλληλη μέθοδος εξόρυξης τελικά επιλέγεται δια της άτοπου απαγωγής, δημιουργώντας την ανάγκη αναγνώρισης της κατάλληλης μεθόδου εξόρυξης για ένα συγκεκριμένο σύνολο δεδομένων.

Με την αναγνώριση της ανάγκης αυτής, η παρούσα εργασία, παρουσιάζει την χρήση πολλαπλής γραμμικής παλινδρόμησης (multiple linear regression MLR), κύρια παλινδρόμηση συνιστωσών (principal component regression, PCR), μερικής ελαχίστων τετραγώνων (Partial Least Squares, PLS), και νευρωνικών δικτύων (Neural Networks) ως μεθόδους πρόβλεψης σε τρία διαφορετικά κατά τα χαρακτηριστικά τους σύνολα δεδομένων, με στόχο την σύγκριση και αξιολόγηση.

Παράλληλα με την σύγκριση των μεθόδων πρόβλεψης, παρουσιάζονται και κάποιες τεχνικές προ-επεξεργασίας των δεδομένων που αποσκοπούν στην αποκάλυψη της φύσης των δεδομένων, που με την σειρά της βοηθούν στην επιλογή της κατάλληλης μεθόδου πρόβλεψης.

Με την εφαρμογή των μεθόδων αυτών αναλύονται τα πλεονεκτήματα και τα μειονεκτήματα τους, αποσκοπώντας στην αναγνώριση της μεθόδου η οποία αποδίδει κατά τον μεγαλύτερο δυνατό βαθμό, στοχεύοντας στην παρακίνηση αναλυτών για την μείωση χρόνου εύρεσης της κατάλληλης μεθόδου εξόρυξης δεδομένων.

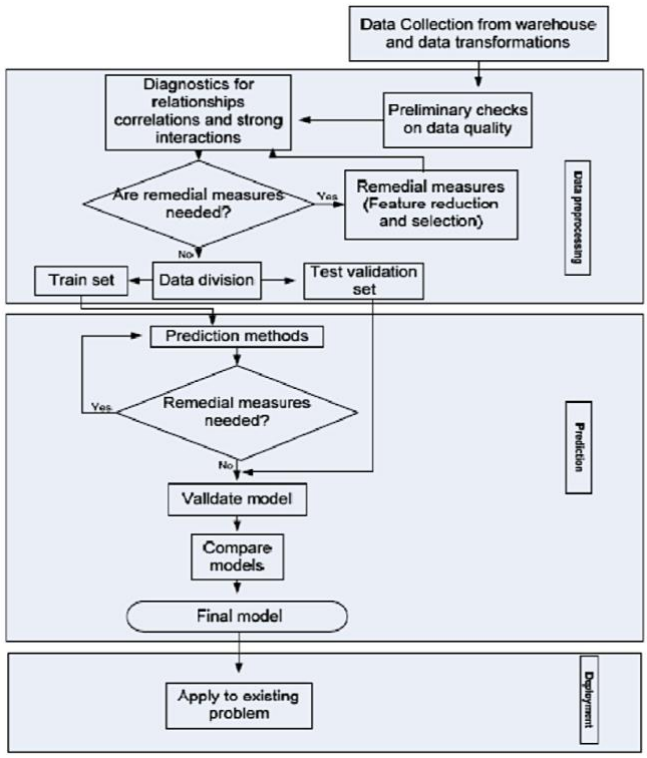
2 Βιβλιογραφική Ανασκόπηση

Η εξόρυξη δεδομένων αποτελεί την εξερεύνηση των ιστορικών δεδομένων με σκοπό την ανακάλυψη ενός ή παραπάνω προτύπων, ή της σχέσης μεταξύ μεταβλητών αυτού. Τα αποτελέσματα συνήθως επικυρώνονται με την εφαρμογή των νεο-ανακαλυφθέντων προτύπων σε υποσύνολα ή όμοιου τύπου δεδομένα, συγκρίνοντας έτσι τον βαθμό ικανότητας πρόβλεψης που πρόβλεψης που προσφέρουν με την πραγματικότητα.

Οι βάσεις της εξόρυξης δεδομένων αποτελούν η στατιστική επιστήμη, η τεχνητή νοημοσύνη, και η μηχανική μάθηση.

Σύμφωνα με τον Usama M. Fayyad, στο “Advances in Knowledge Discovery and Data Mining”, [19], η εξόρυξη δεδομένων μπορεί να διαχωριστεί σε δύο τύπους εργασιών, τις περιγραφικές και τις προβλεπτικές. Η περιγραφή των δεδομένων εκτελείται σε αρκετά υψηλά ικανοποιητικό βαθμό τα τελευταία χρόνια μέσω περιγραφικής εξόρυξης δεδομένων, και χρήσης σχετικά απλών μεθόδων και εργαλείων. Συνέπεια αυτού είναι η εξόρυξη δεδομένων να έχει ως απώτερο σκοπό την πρόβλεψη, καθώς αυτή έχει το μεγαλύτερο εύρος εφαρμογής.

Η προβλεπτική εξόρυξη δεδομένων περιέχει ως πρώτο στάδιο την έρευνα και τους μετασχηματισμούς των δεδομένων που θα χρησιμοποιηθούν. Στην συνέχεια πραγματοποιείται εξερεύνηση των δεδομένων, με επιλογή ή και μείωση των χαρακτηριστικών που θα ληφθούν υπόψιν στην διαδικασία εξόρυξης, αν κρίνεται αναγκαία. Το τρίτο και τελευταίο στάδιο περιέχει την εξόρυξη και τον σχηματισμό του μοντέλου πρόβλεψης, με επεξήγηση, και εφαρμογή του μοντέλου. Η παραπάνω διαδικασία περιγράφεται στο σχεδιάγραμμα που ακολουθεί.



Εικόνα 2 Διαδικασία Προβλεπτικής Εξόρυξης Δεδομένων

Εφαρμογές της προβλεπτικής εξόρυξης δεδομένων πραγματοποιούνται σε ιστορικό αγορών πελατών και ιστορικό χρήσης πλαστικού χρήματος, για την ανακάλυψη προτύπων στην συμπεριφορά πελατών και την ανίχνευση απάτης πιστωτικών καρτών αντίστοιχα, (business intelligence). Επίσης οι βαθμολογίες που πραγματοποιούν θεατές σε διάφορα κανάλια προσφοράς κινηματογραφικών ταινιών, (Netflix, YouTube, κ.α.), χρησιμοποιούνται για την σύσταση παρακολούθησης περιεχομένου που προσφέρουν, αναγνωρίζοντας ομοιότητες μεταξύ των θεατών αλλά και του ίδιου του περιεχομένου, (collaborative filtering). Ακόμα και οι μηχανές αναζήτησης του παγκόσμιου ιστού πραγματοποιούν σε κάποιο στάδιο της αναζήτησης εξόρυξη δεδομένων, φέροντας ως αποτέλεσμα αναφορές του θέματος αναζήτησης με σειρά σχετικότητας. Μεγάλη εφαρμογή και ανάπτυξη παρουσιάζει ο τομέας της υγείας, όπου λαμβάνοντας ιστορικό υγείας ασθενών, και δημογραφικά δεδομένα γίνεται ευκολότερη η διάγνωση του ασθενούς. Ο τομέας μετεωρολογικής πρόβλεψης χρησιμοποιεί μετρήσεις από αισθητήρες, εικόνες από δορυφόρους, και ιστορικά δεδομένα για την συσχέτιση μετεωρολογικών παρατηρήσεων και κατά συνέπεια την αύξηση ακρίβειας των προβλέψεων, Τέλος, ο τομέας των επιχειρησιακών διεργασιών με την χρήση καταγραφών της ροής μιας επιχειρησιακής

διεργασίας, εξερευνά τον λόγο απόκλισης ανάμεσα στην θεωρητική και την πραγματική εκτέλεση μιας διεργασίας, και τρόπους μείωσης της διαφοράς αυτής.

Καθώς οι τομείς εφαρμογής είναι ποικίλοι, καλύπτοντας ένα ευρύ φάσμα, πολλά πλαίσια (frameworks) κατασκευής και εφαρμογής εξόρυξης δεδομένων προτείνονται, βασισμένα αναλόγως με τον τομέα εφαρμογής. Καθώς ένα σχέδιο εξόρυξης δεδομένων περιλαμβάνει, και απαιτεί για την ομαλή εφαρμογή του, τον συντονισμό πολλών τμημάτων σε ένα επιχειρησιακό περιβάλλον, (παράδειγμα εταιρίας όπου διάφοροι ειδικοί, διοικητικά στελέχη, και τμήματα θα πρέπει να συνεργαστούν ανταλλάσσοντας πληροφορίες και γνώσεις πάνω στο θέμα προς ανάλυση), παρουσιάζεται η ανάγκη δημιουργίας σχεδιαγραμμάτων για την οργάνωση της συλλογής των δεδομένων, την ανάλυσή τους, την διάδοση των αποτελεσμάτων, εφαρμογής των αποτελεσμάτων και παρακολούθησης για τυχόν βελτιώσεις της διαδικασίας.

Τα πιο διαδεδομένα πλαίσια εξόρυξης δεδομένων είναι τα Cross Industry Process for Data Mining (CRISP-DM), [10], Define, Measure, Analyze, Improve, Control (DMAIC), [43], και Sample, Explore, Modify, Model, Assess, (SEMMA), [11]:

1. CRISP-DM: Το Cross Industry Process for Data Mining, προτάθηκε από μια κοινοπραξία ευρωπαϊκών επιχειρήσεων το 1990, με σκοπό να χαρτογραφηθεί ο τρόπος εφαρμογής της εξόρυξης δεδομένων. Αποτελείται από έξι φάσεις, κατανόησης του χώρου της επιχείρησης (Business Understanding), κατανόησης των δεδομένων (Data Understanding), προετοιμασίας δεδομένων (Data preparation), μοντελοποίησης (Modeling), αξιολόγησης (Evaluation), και ανάπτυξης (Deployment).

2. DMAIC: Είναι μία από τις μεθοδολογίες κατασκευασμένες από τον Bill Smith, μηχανικό της Motorola, για την βελτίωση διεργασιών, μέσω της μείωσης των ελαττωμάτων, βελτιστοποίησης και σταθεροποίησης τους. Χρησιμοποιείται κυρίως στον τομέα των κατασκευών, την παροχή υπηρεσιών, και την διαχείριση επιχειρησιακών διαδικασιών. Η γενική φύση της μεθοδολογίας αυτής την κάνει εφαρμόσιμη και σε εφαρμογές εξόρυξης δεδομένων. Τα αρχικά του ακρωνυμίου

DMAIC, Define, Measure, Analyze, Improve, Control, αποτελούν και τα στάδια βελτίωσης από τα οποία αποτελείται.

3. SEMMA: Αποτελεί μία λίστα αναπτυγμένη από το ινστιτούτο SAS (Statistical Analysis System), η οποία καθοδηγεί εφαρμογές εξόρυξης δεδομένων. Αποτελείται από παρόμοια στάδια με την CRISP-DM μεθοδολογία, αφήνοντας έξω από την διαδικασία την κατανόηση του χώρου της επιχείρησης (business understanding), γεγονός για το οποίο κατακρίνεται. Να σημειωθεί πως η μεθοδολογία SEMMA έχει σχεδιαστεί εξ αρχής ως βοήθημα των χρηστών του λογισμικού “SAS Enterprise Miner”, με την εφαρμογή του σε εξωτερικά πλαίσια να έχει έμφυτα μειονεκτήματα.

2.1 Απόκτηση / Αποθήκευση Δεδομένων

Ακόμα και με τον υψηλό ρυθμό παραγωγής του, αλλά και τον ήδη υπάρχοντα όγκο τους, η απόκτηση των δεδομένων χρίζει υψηλό βαθμό προσπάθειας και προσοχής. Οι τρόποι με τον οποίο παράγονται τα δεδομένα είναι συνήθως αυτόνομοι, ειδικά στην παρούσα εποχή της συνεχώς αυξανόμενης εφαρμογής του Διαδικτύου των Πραγμάτων, (Internet Of Things, IOT). Με την παραγωγή των δεδομένων γεννιέται η ανάγκη για την καταγραφή τους, και την επεξεργασία τους για κατάλληλη οπτικοποίηση, ανάγκη η οποία αυξάνεται καθώς αυξάνεται και ο όγκος των δεδομένων. Η αποθήκευσή τους γίνεται συνήθως σε αποθήκες δεδομένων (data warehouses), από τις οποίες είναι εύκολη εξαγωγή τους. Τα δεδομένα στην παρούσα εργασία λήφθηκαν από το UCI Machine Learning Repository, (https://archive.ics.uci.edu/ml/index.php), αποθήκη δεδομένων του Πανεπιστημίου της Καλιφόρνια στο Ιρβάιν.

2.2 Προετοιμασία Δεδομένων

Τα δεδομένα στην ακατέργαστη μορφή τους δεν προσφέρονται για άμεση χρήση σε εφαρμογές προβλεπτικής εξόρυξης δεδομένων. Συνήθως περιέχουν θόρυβο, ελλιπείς τιμές και ασυνεπή δεδομένα, απαιτώντας την προεπεξεργασία τους, με σκοπό την βελτίωση της ποιότητάς τους πριν την εφαρμογή της εξόρυξης δεδομένων. Η προετοιμασία τους ώστε να έρθουν σε κατάλληλη μορφή είναι συνήθως αναγκαία, καθώς τεχνικές εξόρυξης δεδομένων συμπεριφέρονται διαφορετικά αναλόγως με τις μεθόδους επεξεργασίας και μετασχηματισμού που εφαρμόστηκαν.

2.2.1 Επεξεργασία Δεδομένων

Καθώς τα δεδομένα δεν είναι συνήθως στην ιδανική κατάσταση για άμεση χρήση σε προβλεπτική εξόρυξη δεδομένων λόγω ύπαρξης θορύβου, ακραίων τιμών, ή ελλιπών τιμών εμφανίζεται η ανάγκη για κατάλληλη επεξεργασία και εξομάλυνσή τους. Η κρίση του αναλυτή στην εφαρμογή ορισμού του θορύβου στα εκάστοτε δεδομένα παίζει καίριο ρόλο, [42]. Υπάρχουν διάφορα μέσα φιλτραρίσματος με σκοπό την εξομάλυνση των δεδομένων:

1) Κινητός Μέσος Όρος (Moving Average): αυτή η μέθοδος χρησιμοποιείται για το φιλτράρισμα υψηλών και χαμηλών παρατηρήσεων, [42, 46]. Χρησιμοποιείται κυρίως στους χρηματιστηριακούς δείκτες. Εφαρμόζεται επιλέγοντας ένα σημείο μιας σειράς δεδομένων, και συνεχίζοντας την πορεία της σειράς, χρησιμοποιώντας την μέση τιμή της παρατήρησης με τις προηγούμενες παρατηρήσεις, αντί της πραγματικής τιμής της παρατήρησης. Με την τεχνική αυτή η διακύμανση των δεδομένων μειώνεται.

2) Εξομάλυνση με χρήση διαμέσου (median filtering): εφαρμόζεται συνήθως σε δεδομένα χρονοσειρών για την αφαίρεση ακραίων τιμών, [46, 60]. Η μέθοδος αυτή διατηρεί τα χαρακτηριστικά των δεδομένων, σε σχέση με την μέθοδο του κινητού μέσου όρου, ενώ παράλληλα απαλλάσσει από θόρυβο και ακραίες τιμές.

3) Κανονικοποίηση: είναι μέθοδος κατά την οποία τιμές δεδομένων αντικαθίστανται συνήθως με άλλες τιμές κυμαινόμενες στο διάστημα [0,1]. Η πλειοψηφία των τεχνικών εξόρυξης δεδομένων λειτουργούν βέλτιστα με κανονικοποιημένα δεδομένα, [42, 46]. Υπάρχουν διάφοροι τύποι αυτής της μεθόδου:

   - η κανονικοποίηση ελάχιστου – μέγιστου, όπου οι αριθμητικές τιμές αντικαθίστανται με άλλες κυμαινόμενες εντός προκαθορισμένης περιοχής τιμών:

$$x' = \frac{x - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

   όπου x'η νέα τιμή, x η υπάρχουσα τιμή, max_A, min_A, η μέγιστη και ελάχιστη τιμή, new_max_A, new_min_A η νέα μέγιστη και ελάχιστη τιμή

   - η κανονικοποίηση τυπικής απόκλισης κατά την οποία αφαιρείται η μέση τιμή των μεταβλητών από κάθε μεταβλητή, και στην συνέχεια διαιρείται από την τυπική απόκλισή τους. Αποτέλεσμα αυτού είναι μεταβλητές με μέση τιμή μηδέν και διακύμανση μονάδα, δίνοντας σε κάθε μία από τις μεταβλητές την ίδια πιθανότητα στην εμφάνιση στο μοντέλο:

$$x' = \frac{x - \mu,}{\sigma}$$

   όπου x'η νέα τιμή, x η υπάρχουσα τιμή, μ η μέση τιμή, και σ η τυπική απόκλιση.

4) Αντιμετώπιση ελλιπών τιμών: μια ελλιπής τιμή είναι πιθανή κατά την μεταφορά των δεδομένων, να μην ήταν διαθέσιμη την στιγμή καταχώρησης, ή αποτέλεσμα κακής διαχείρισης τους μέχρι το στάδιο της εξόρυξης, [42, 46]. Η απώλεια θα πρέπει να διορθωθεί, καθώς επηρεάζει ή πολλές φορές καθιστά την εφαρμογή εξόρυξης ανέφικτη. Κάποια εργαλεία εξόρυξης δεδομένων δεν λαμβάνουν υπόψιν τους τις μεταβλητές αυτές, ή βρίσκουν κατάλληλα υποκατάστατά τους. Σε καμία όμως από τις παραπάνω περιπτώσεις ο αναλυτής δεν έχει τον έλεγχο, αντιμετωπίζοντας την

πιθανότητα λανθασμένης προδιάθεσης στα δεδομένα (bias). Ο αναλυτής λοιπόν θα πρέπει να κατέχει τον έλεγχο της κατάστασης, καθώς το μοτίβο των δεδομένων πρέπει να διατηρείται. Μέθοδοι όπως ο υπολογισμός της μέσης τιμής των μεταβλητών και εισαγωγής αυτής στις ελλείπεις μεταβλητές είναι αποδεκτή, καθώς με τον τρόπο αυτό διατηρείται η μέση τιμή των μεταβλητών, ή εισαγωγή τιμών οι οποίες διατηρούν σταθερή την τυπική απόκλιση, δηλαδή τιμών οι οποίες είναι πολύ κοντά στις πραγματικές τιμές μέσω εφαρμογής κατηγοριοποίησης ή παλινδρόμησης. Η διαγραφή της γραμμής της πληροφορίας σαν μέθοδος αντιμετώπισης, αν και λύνει το πρόβλημα, οδηγεί σε απώλεια χρήσιμης πληροφορίας.

5) Ονομαστικές μεταβλητές (categorical data): αν και η πλειοψηφία των μεθόδων εξόρυξης εφαρμόζονται σε ποσοτικά δεδομένα, υπάρχει και πληθώρα ποιοτικών δεδομένων, [1]. Συνήθως αυτά περιγράφουν μία κατάσταση, και η μετατροπή τους σε ποσοτικά δεδομένα γίνεται ως εξής: αν υπάρχουν n ποιοτικές μεταβλητές, τότε χρειάζονται n-1 ποσοτικές μεταβλητές.

2.2.2 Μείωση Διαστάσεων (Dimension Reduction)

Διαστάσεις ενός συνόλου ονομάζεται το πλήθος των χαρακτηριστικών που περιέχουν τα αντικείμενα που απαρτίζουν το σύνολο δεδομένων. Η μείωση διαστάσεων είναι η συγχώνευση των πληροφοριών ενός μεγάλου συνόλου δεδομένων σε ένα μικρότερο, ευκολότερα διαχειρίσιμο σύνολο. Η διαδικασία αυτή λοιπόν παρέχει σαν αποτέλεσμα ένα σύνολο δεδομένων το οποίο παράγει κατά ένα μεγάλο βαθμό τα ίδια αναλυτικά αποτελέσματα, αλλά είναι πολύ μικρότερο σε μέγεθος από το αρχικό σύνολο δεδομένων. Καθώς το μέγεθος είναι μικρότερο, η διαδικασία της εξόρυξης δεδομένων απαιτεί λιγότερο χρόνο για να ολοκληρωθεί, δίνοντας την ευκαιρία να πραγματοποιηθούν εφαρμογές διαφόρων τεχνικών εξόρυξης με σκοπό την εύρεση της πλέον αποτελεσματικής. Επίσης αφαιρούνται μη σχετικές πληροφορίες, με αποτέλεσμα την μείωση του θορύβου των δεδομένων, και της ευκολότερης η οπτικοποίησης των δεδομένων, οδηγώντας τελικά σε ευκολότερα κατανοητό μοντέλο πρόβλεψης. Κάποιοι μέθοδοι μείωσης διαστάσεων αποτελούν η παλινδρόμηση

(regression), η συσταδοποίηση (clustering), η ανάλυση κυρίων συνιστωσών (Principal Component Analysis, PCA).

2.2.3 Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis, PCA)

Σχεδόν σε κάθε περίπτωση εξόρυξης δεδομένων, η Ανάλυση Κυρίων Συνιστωσών (PCA), [29], μπορεί να πραγματοποιηθεί σαν πρώτο βήμα. Είναι στατιστική μέθοδος κατά την οποία περιορίζεται η πολυπλοκότητα των δεδομένων χωρίς όμως την απώλεια πληροφορίας, [40, 47]. Κατά την μέθοδο αυτή, τα δεδομένα μετασχηματίζεται από ένα σύνολο συγγενικών μεταβλητών σε ένα καινούργιο σύνολο ασυσχέτιστων μεταβλητών, τις κύριες συνιστώσες (principal components). Κάθε μία από τις κύριες συνιστώσες είναι μια γραμμική απεικόνιση των αρχικών μεταβλητών, όπου οι συντελεστές υποδηλώνουν την σπουδαιότητα της μεταβλητής. Στην περίπτωση όπου οι αρχικές μεταβλητές δεν παρουσιάζουν κάποια σημαντική συσχέτιση μεταξύ τους, η εφαρμογή της ανάλυση κυρίων συνιστωσών δεν θα αποφέρει κάποιο κέρδος στην εξόρυξη δεδομένων.

Πριν την εφαρμογή της ανάλυσης κυρίων συνιστωσών, οι μεταβλητές θα πρέπει να είναι σε κατάλληλη κατάσταση, δηλαδή να έχουν παρόμοιες διακυμάνσεις, και να μετρούνται σε συγκρίσιμες μονάδες μέτρησης. Στην περίπτωση που δεν είναι, θα χρειαστεί να πραγματοποιηθεί κανονικοποίηση των μεταβλητών, με απώτερο σκοπό μεταβλητές με μέσο όρο μηδέν, και διακύμανση ίση με ένα. Η διαδικασία της κανονικοποίησης πραγματοποιείται ώστε μεταβλητές με μεγάλο εύρος τιμών να μη υπερισχύσουν έναντι αυτών με μικρό.

2.2.4 Ανάλυση Συντελεστών Συσχέτισης

Η ανάλυση συντελεστών συσχέτισης, Correlation Coefficient Analysis (CCA), δίνει την δυνατότητα εκτίμησης της γραμμικής εξάρτησης μεταξύ δύο τυχαίων μεταβλητών, [13]. Η τιμή της κύριας ανάλυσης συνιστωσών ισούται με την συνδιακύμανση των δύο μεταβλητών διαιρούμενη με την μεγαλύτερη δυνατή συνδιακύμανση. Το εύρος τιμών το οποίο λαμβάνει είναι στο διάστημα [-1,+1], με αρνητικό συντελεστή συσχέτισης να σηματοδοτεί μια αντιστρόφως ανάλογη σχέση, και θετικό συντελεστή μια ανάλογη σχέση μεταξύ των μεταβλητών, [24].

Με δύο μεταβλητές X,Y, οι οποίες έχουν διασπορά σ_x^2, σ_y^2, αντίστοιχα, και συνδιασπορά $\sigma_{xy} = C(X,Y) = E(X,Y) - E(X)E(Y)$, υπάρχει η μήτρα συνδιασποράς:

$$cov(x,y) = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$$

και ο συντελεστής συσχέτισης:

$$p_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Τότε η ανάλυση συνιστωσών συσχέτισης δίνει σαν αποτέλεσμα μήτρα της μορφής:

$$corrcoef(x,y) = \begin{bmatrix} 1 & p_{xy} \\ p_{xy} & 1 \end{bmatrix}$$

Οι τιμή του συντελεστή συσχέτισης όπως αναφέρθηκε, δείχνει τον γραμμικό βαθμό συσχέτισης μεταξύ δύο μεταβλητών. Τιμή συσχέτισης μικρότερη του 0.3 δείχνει μικρή συσχέτιση των μεταβλητών, με μηδέν να εκφράζεται η μη ύπαρξη συσχέτισης, ενώ τιμές μεγαλύτερες του μηδέν και μικρότερες του 0.7, μεγάλη συσχέτιση των μεταβλητών. Η περίπτωση τιμής μεγαλύτερης του 0.7, προσδιορίζει ισχυρή σχέση.

Εύκολο είναι να παρατηρηθεί πως ο συντελεστής συσχέτισης δύο σταθερών μεταβλητών, είναι κοντά στο μηδέν, ακόμα και με την ύπαρξη θορύβου στα δεδομένα.

Καθώς πραγματοποιείται η χρήση της διασποράς των μεταβλητών, είναι ευκόλως αντιληπτό πως μεταβλητές με μεγάλα μεγέθη θα έχουν διαφορετική αντιμετώπιση σε σχέση με τις μεταβλητές με μικρά μεγέθη. Για αυτό τον λόγο είναι θεμιτό η κανονικοποίηση των μεταβλητών πριν την εφαρμογή της μεθόδου, συνήθως με την αφαίρεση της μέσης τιμής και την διαίρεση της κάθε παρατήρησης με την τυπική απόκλιση, όπου:

$$x' = \frac{x_i - \mu_x}{\sigma_{ii}}, \forall i = 1, \dots, n$$

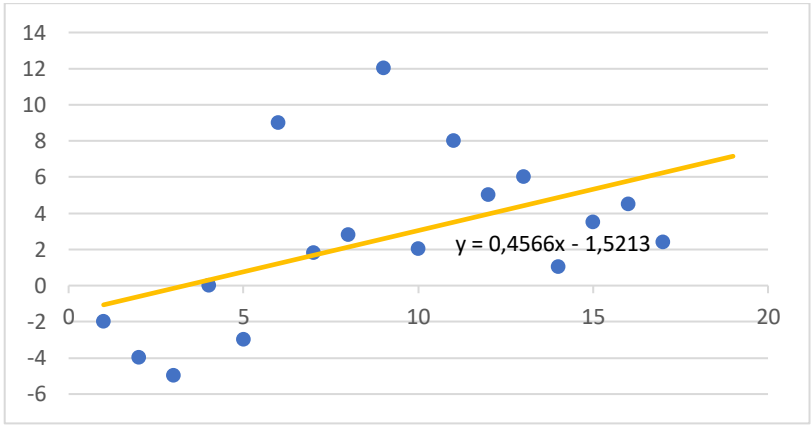
Με τον μετασχηματισμό αυτό ο πίνακας συνδιασποράς του x', ισούται με τον πίνακα συσχέτισης του x. Η μέθοδος έχει σαν αποτέλεσμα την αφαίρεση πιθανής συσχέτισης μεταξύ των συνιστωσών.

2.3 Προβλεπτική Εξόρυξη Δεδομένων

Η προβλεπτική εξόρυξη δεδομένων ακολουθεί της απόκτησης και προεπεξεργασίας τους. Η προεπεξεργασία των δεδομένων δίνει επίγνωση της φύσης των δεδομένων και των σχέσεων που είναι παρών σε αυτά. Υπάρχουν πολλές τεχνικές όπως η παλινδρόμηση, τα νευρωνικά δίκτυα, τα δέντρα απόφασης, η ομαδοποίηση κ.α. Στην παρούσα διπλωματική εργασία αναλύονται τεχνικές παλινδρόμησης, γνωστές και ως γραμμικές, και τα νευρωνικά δίκτυα. Η επιλογή των τεχνικών πραγματοποιήθηκε λόγω της έμφυτης διαφοράς τους. Οι τεχνικές παλινδρόμησης αποτελούν τις δομή σχέσεων μεταξύ των επιλεγμένων τιμών του x και των παρατηρούμενων τιμών του y από τις οποίες η πιο πιθανή τιμή του y μπορεί να προβλεφθεί για οποιαδήποτε τιμή του x. Στον αντίποδα τα νευρωνικά δίκτυα, είναι τεχνική εξόρυξης δεδομένων εμπνευσμένη από τους νευρώνες του εγκεφάλου, και είναι ικανά μάθησης λαμβάνοντας υπόψιν τους παραδείγματα του παρελθόντος τα οποία είναι μη γραμμικά, με υψηλή αντοχή στο θόρυβο που μπορεί να παρουσιάζουν τα δεδομένα.

2.3.1 Παλινδρόμηση (Regression)

Η γραμμική παλινδρόμηση είναι η αρχαιότερη προβλεπτική τεχνική, βασισμένη στις σχέσεις των μεταβλητών εισόδου και της μεταβλητής εξόδου. Ειδικότερα χρησιμοποιεί την γραμμική εξίσωση, $y = a \cdot x + b$, όπου x οι μεταβλητές εισόδου που χρησιμοποιούνται για την πρόβλεψη της μεταβλητής y, και α η κλίση της ευθείας αυτής, παράδειγμα της οποίας φαίνεται στο Γράφημα 1.



Γράφημα 1 y=0,4566x-1,5213

Καθώς η τέλεια σχέση ανάμεσα στις μεταβλητές x,y σπανίζει, κοινώς υπάρχει ένα ποσοστό σφάλματος στην γραμμική σχέση-πρόβλεψη, [54], η γραμμική σχέση μπορεί να εκφραστεί ως $y = a \cdot g(x) + b, όπου\ g(x) = a \cdot x + \varepsilon$, όπου α το βάρος το οποίο γραμμικά συνδέει την μεταβλητή εισόδου x με την πρόβλεψη, και ε το σφάλμα πρόβλεψης, δηλαδή η διαφορά ανάμεσα στην προβλεπόμενη και την πραγματική τιμή τις μεταβλητής y.
Στην περίπτωση τις ύπαρξης πολλαπλών μεταβλητών εισόδου, $x_1, x_2, x_3, \dots, x_n$, η τεχνική ονομάζεται πολλαπλή γραμμική παλινδρόμηση.

Κατά την εφαρμογή παλινδρόμησης, πραγματοποιούνται κάποιες παραδοχές:

1) Θεωρείται υπαρκτή η γραμμική σχέση ανάμεσα τις μεταβλητές εισόδου και τις μεταβλητές εξόδου, [39].
2) Οι όροι σφάλματος είναι τυχαίοι, ακολουθούν κανονική κατανομή με μέση τιμή μηδέν, και δεν υπάρχει συσχέτιση μεταξύ τους, [31, 39].
3) Το πλήθος των ακραίων παρατηρήσεων (outliers), είναι μικρό, [31].
4) Οι μεταβλητές εισόδου, δεν παρουσιάζουν ή παρουσιάζουν μικρές αλληλεπιδράσεις μεταξύ τους, [31, 33].

2.3.1.1 Πολλαπλή Γραμμική Παλινδρόμηση

Η μέθοδος πολλαπλής γραμμικής παλινδρόμησης, (multiple linear regression, MLR), αποτελεί γενίκευση της απλής γραμμικής παλινδρόμησης, καθώς ένα σύνολο $X_1, X_2, \dots, X_{p-1}$, μεταβλητών χρησιμοποιούνται για την πρόβλεψη της μεταβλητής Y, [13, 42]. Με τρόπο ίδιο με αυτό της απλής γραμμικής παλινδρόμησης, το μοντέλο αυτό έχει την μορφή:

$$Y = \alpha_0 + \alpha_1 \cdot X_1 + \alpha_2 \cdot X_2 + \alpha_3 \cdot X_3 + \cdots + \alpha_{p-1} \cdot X_{p-1} + \varepsilon$$

για κάποιες παραμέτρους $\alpha_0, \alpha_1, \dots, \alpha_{p-1}$.

Για την διερεύνηση της σχέσης μεταξύ της μεταβλητής Y, και $X_1, X_2, \dots, X_{p-1}$, λαμβάνεται υπόψιν δείγμα μεγέθους n, όπου καταγράφονται οι τιμές $Y_i, X_{i,1}, X_{i,2}, \dots, X_{i,p-1}, \forall i = 1, \dots, n$ εγγραφή του δείγματος:

$$Y_i = \alpha_0 + \alpha_1 \cdot X_{i,1} + \alpha_2 \cdot X_{i,2} + \alpha_3 \cdot X_{i,3} + \cdots + \alpha_{p-1} \cdot X_{i,p-1} + \varepsilon_i, i = 1, \dots, n$$

όπου τα σφάλματα $\varepsilon_i, i = 1, \dots, n$, θεωρούνται ανεξάρτητες μεταβλητές.

Συνεπώς: $Y = a \cdot X + \varepsilon$, όπου

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, a = \begin{pmatrix} a_0 \\ \vdots \\ a_{n-1} \end{pmatrix}, X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1,p-1} \\ \vdots & X_{21} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{n,p-1} \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_0 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Η παραπάνω σχέση μπορεί να προσδιορίσει την βέλτιστη τιμή των $\alpha_0, \alpha_1, \dots, \alpha_{n-1}$, η οποία προσδιορίζεται όταν το άθροισμα των τετράγωνων των σφαλμάτων ελαχιστοποιείται, με

$$SSE = \sum_{i=1}^{n} (y - \hat{y})^2 = \sum (Y - Xa)^2 = \sum_{i=0}^{n} \varepsilon^2$$

όπου $\hat{y}$, η πρόβλεψη του y, και n αριθμός εγγραφών.

Το Y διάνυσμα θα ακολουθεί πολυδιάστατη κανονική, καθώς το ε αποτελείται από n αριθμό τυχαίων μεταβλητών, και έχει συνάρτηση πυκνότητας πιθανότητας $N(o, \sigma^2 I_n)$ δηλαδή ακολουθεί πολυδιάστατη κατανομή όπου I_n είναι ο μοναδιαίος πίνακας διάστασης n:

$$L(a, \sigma^2) = f(y_1, y_2, \dots, y_n; b, \sigma^2) = \frac{e^{-\frac{1}{2\sigma^2}(y - Xa)^T (y - Xa)}}{(2\pi)^{\frac{n}{2}} (\sigma^2)^{n/2}}$$

Οπότε η μεγιστοποίηση ως προς b, προϋποθέτει την ελαχιστοποίηση του:

$$(Y - Xa)^T(Y - Xa) = e^T e = \sum_{i=0}^{n} \varepsilon^2$$

Συνεπώς:

$$(Y - Xa)^T(Y - Xa) = (Y^T - a^T X^T)(Y - Xa) = Y^T Y - Y^T Xa - a^T X^T Y + a^T X^T Xa$$

Οπότε παραγωγίζοντας ως προς a:

$$\frac{df}{da}(Y - Xa)^T(Y - Xa) = -2X^T Y + 2X^T Xa$$

Η οποία παράγωγος είναι ίση με μηδέν όταν: $X^T Xa = X^T Y$, με το σύστημα εξισώσεων με p αγνώστους να έχει μοναδική λύση όταν υπάρχει ο αντίστροφος του $X^T X$, οπότε η εκτιμήτρια μέγιστης πιθανοφάνειας του $\alpha = \left[\alpha_0, \alpha_1, \dots, \alpha_{p-1}\right]^T$ θα είναι:

$$\hat{a} = (X^T X)^{-1} X^T Y$$

Με αποτέλεσμα οι προβλέψεις της μεταβλητής Y, να είναι:

$$\hat{Y} = X\hat{a} = X(X^T X)^{-1} X^T Y$$

και τα εκτιμημένα σφάλματα, οι διαφορές των προβλέψεων της μεταβλητής Y από τις πραγματικές τιμές της:

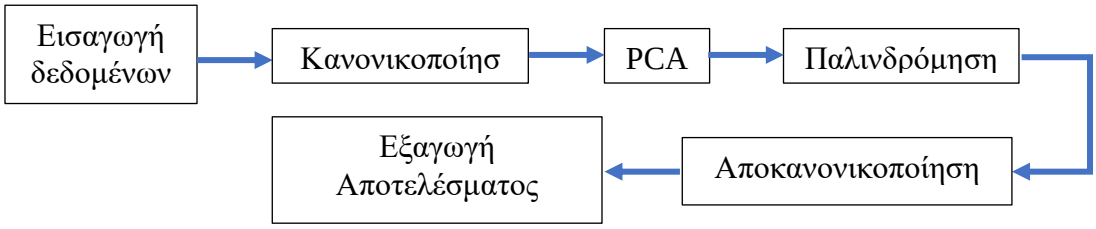
$$\hat{\varepsilon} = Y - \hat{Y} = Y - X(X^T X)^{-1} X^T Y$$

2.3.1.2 Παλινδρόμηση Κυρίων Συνιστωσών

Η παλινδρόμηση κυρίων συνιστωσών, (Principal Component Regression, PCR), είναι τεχνική παλινδρόμησης η οποία πραγματοποιεί χρήση της ανάλυσης κυρίων συνιστωσών, PCA, [29, 59]. Η τεχνική αποτελείται από μια διαδικασία τριών σταδίων:

1) Υπολογισμός των κυρίων συνιστωσών
2) Επιλογή κυρίων συνιστωσών
3) Εφαρμογή πολλαπλής γραμμικής παλινδρόμησης

Η διαδικασία της παλινδρόμησης κυρίων συνιστωσών αποτυπώνεται στο Γράφημα 2:



Γράφημα 2 Διαδικασία Παλινδρόμησης Κυρίων Συνιστωσών

Η εφαρμογή της ανάλυσης κυριών συνιστωσών πραγματοποιείται με την ανάλυση του πίνακα των μεταβλητών με ανάλυση σε ιδιάζουσες τιμές, (singular value decomposition, SVD). Επομένως για πίνακα X mxn:

$$X = U * S * V^T$$

Όπου U ένας ορθοκανονικός πίνακας, ($U^T = U^{-1}$), mxm αποτελούμενος από τα ιδιοδιανύσματα του XX^T, S πίνακας mxn ορθογώνιος διαγώνιος με τα διαγώνια στοιχεία του να αποτελούν τις τετραγωνικές ρίζες των ιδιοτιμών του XX^T, και V ένας nxn ορθοκανονικός πίνακας των ιδιοδιανυσμάτων του X^TX.

Τότε έχουμε την κατασκευή των κυρίων συνιστωσών ως:

$$z = X \cdot V \text{ ή } z = U \cdot S$$

όπου Z, ονομάζεται πίνακας βαθμολόγησης, X ο mxn πίνακας των αρχικών δεδομένων και V ο πίνακας μετασχηματισμού nxn, με m να αναπαριστά τον αρχικό αριθμό διαστάσεων και n τον νέο μειωμένο αριθμό διαστάσεων. Συνεπώς με τον τρόπο αυτό προβάλλονται τα αρχικά δεδομένα σε νέο σύστημα συντεταγμένων V, με χρήση του πίνακα βαθμολόγησης Z. Άρα:

$$X = z_1 v_1^T + z_2 v_2^T + \cdots + z_M v_m^T$$

Με την χρήση της μεθόδου ανάλυσης κυρίων συνιστωσών, ξεπερνούνται τα προβλήματα που συναντά κανείς με την εφαρμογή της απλής πολλαπλής γραμμικής παλινδρόμησης σε συγγραμικά δεδομένα, και εφαρμόζεται παλινδρόμηση σε μειωμένα δεδομένα, χάρη στην χρήση ανάλυσης κυρίων συνιστωσών, τα οποία είναι ανεξάρτητα.

Η απόφαση για τον αριθμό κυρίων συνιστωσών που θα ληφθούν υπόψιν μπορεί να ληφθεί με διαφορετικούς τρόπους, [21, 28, 49]:

1) Με εις άτοπο απαγωγή
2) Επιλογή των συνιστωσών που επεξηγούν το 90% της συνολικής πληροφορίας.
3) Επιλογή του αριθμού των κυρίων συνιστωσών που περιέχουν την μεγαλύτερη μεταβλητότητα, ή το μεγαλύτερο δυνατό ποσοστό πληροφορίας.
4) Από το γράφημα των ιδιοτιμών επιλογή των συνιστωσών πάνω από την απότομη κλίση που παρατηρείται (knee rule)
5) Επιλογή των συνιστωσών των οποίων οι βαθμολογίες u*s είναι συσχετίζονται με την μεταβλητή προς πρόβλεψη, (Best Subset Selection)

2.3.1.3 Μερικών Ελαχίστων τετραγώνων

Η τεχνική μερικών ελαχίστων τετραγώνων, (Partial Least Squares, PLS), αποτελεί μέθοδο μοντελοποίησης με γραμμική παλινδρόμηση του συνόλου μεταβλητών εισόδου (x), μέσω της τροποποίησης του σε ένα νέο σύνολο μεταβλητών εισόδου (t), για την πρόβλεψη τροποποιημένου συνόλου μεταβλητών (u) από το αρχικό σύνολο μεταβλητών προς πρόβλεψη (y), [36]. Με τον τρόπο αυτό απαλείφεται η συγγραμικότητα ανάμεσα στα δεδομένα που χρησιμοποιούνται για την πραγματοποίηση της πρόβλεψης και τα προβλεπόμενα δεδομένα.

Έστω X nxk πίνακας, και Y nxm πίνακας όπου p ο αριθμός των μεταβλητών πρόβλεψης, m ο αριθμός των μεταβλητών προς πρόβλεψη και n ο αριθμός των παρατηρήσεων. Σκοπός της τεχνικής μερικής ελαχίστων τετραγώνων είναι να ληφθεί υπόψιν η μεγαλύτερη δυνατή διακύμανση των μεταβλητών πρόβλεψης και μεταβλητών προς πρόβλεψη, και η μεγιστοποίηση της συσχέτισης ανάμεσα στις διακυμάνσεις αυτές. Τότε:

$$X = TP^T + E$$
$$Y = UC^T + G$$
$$T = U + H$$

Όπου T ο πίνακας βαθμολόγησης που συνοψίζει τις X μεταβολές, P ο πίνακας X-φορτίων, U ο πίνακας βαθμολόγησης που συνοψίζει τις Y μεταβολές, C ο πίνακας Y-φορτίων, και E,G,H πίνακες των σφαλμάτων - υπολοίπων (residuals).

Αναλυτικότερα η μέθοδος αρχικά βρίσκει το νέο σύνολο μεταβλητών $t_a = (a = 1,2,\dots,A)$, οι οποίες είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών $x_k = (k = 1,2,\dots,k)$, για n παρατηρήσεις, με συντελεστές - βάρη $w_{ka}^* = (a = 1,2,\dots,A)$ και e_{ik}τα σφάλματα-υπόλοιπα (residuals):

$$t_{ia} = \sum_k w_{ka}^* x_{ik} \Longrightarrow x_{ik} = \sum_\alpha t_{ia} p_{ak} + e_{ik}$$

Ομοίως για τις Y μεταβλητές:

$$y_{im} = \sum_{\alpha} u_{ia} c_{am} + g_{im}$$

όπου $u_{ia} = (a = 1,2, \dots, A)$ οι συντελεστές – βάρη, και g_{im} τα σφάλματα απόκλισης των προβλέψεων από τις πραγματικές τιμές.

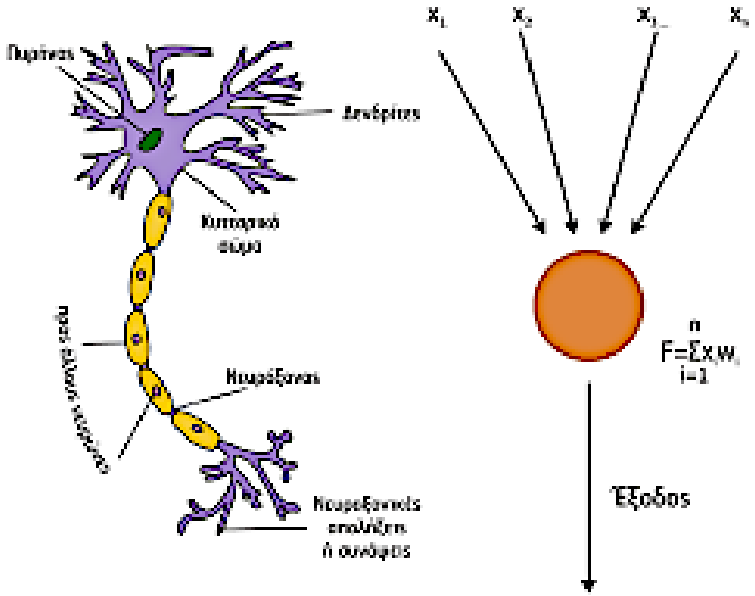
Επομένως η πρόβλεψη αποτυπώνεται ως:

$$y_{im} = \sum_{\alpha} c_{am} \sum_{k} w_{ka}^{*} x_{ik} + f_{im}$$

Είναι εμφανής η ομοιότητα της διαδικασίας που ακολουθείται με αυτή της ανάλυσης κυρίων συνιστωσών, όμως καθώς η ανάλυση κυρίων συνιστωσών βρίσκει υπερεπίπεδα με την μέγιστη δυνατή διακύμανση ανάμεσα στις μεταβλητές εισόδου, η μερικής ελαχίστων τετραγώνων τεχνική αποσκοπεί στην συσχέτιση μέσω μοντέλου γραμμικής παλινδρόμησης της προβολής των μεταβλητών εισόδου και προβλεπόμενων μεταβλητών σε νέα συστήματα συντεταγμένων.

2.3.2 Νευρωνικά Δίκτυα (Neural Networks)

Τα νευρωνικά δίκτυα διαφέρουν των τεχνικών παλινδρόμησης που ακολουθούν τους προκαθορισμένους κανόνες λειτουργίας των υπολογιστών, καθώς τον συνδυάζουν και με τον αφηρημένο τρόπο σκέψης και λειτουργίας του εγκεφάλου, προσπαθώντας να προσομοιώσουν τις πολύπλοκες αντιδράσεις ενός ζωντανού οργανισμού, με την υπολογιστική ταχύτητα του υπολογιστή, [44]. Αποτελούνται, κατά αντιστοίχιση με την σύσταση βιολογικού εγκεφάλου, από νευρώνες, οι οποίοι όμως δεν είναι τόσο πολύπλοκοι στην δομή τους, όπως φαίνεται στην Εικόνα 3.

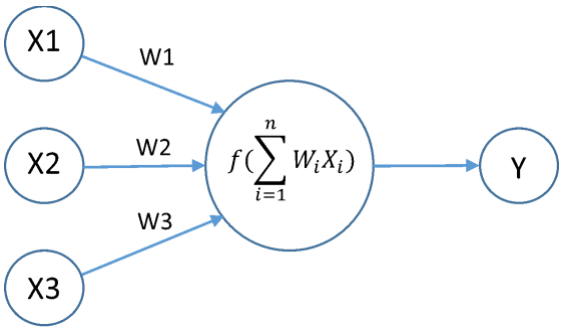


Εικόνα 3 Νευρώνας

Οι νευρώνες από τους οποίους αποτελείται ένα νευρωνικό δίκτυο, προσομοιώνουν την δομή των νευρώνων του εγκεφάλου, των οποίων η δομή είναι μαθηματικοποιημένη:

1) Η συνδέσεις μεταξύ των νευρώνων είναι ίδιες, δηλαδή ο χρόνος που χρειάζεται για την μεταφορά του σήματος από τον νευρώνα i στον j, είναι ίδιος $\forall i, j$
2) Κάθε νευρώνας περιέχει μία συνάρτηση ενεργοποίησης, $f(\sum_1^n w_i x_i)$, η οποία καθορίζει το σήμα εξόδου του νευρώνα, Y, συναρτήσει της έντασης των σημάτων εισόδου. Ο χρόνος που είναι αναγκαίος για τον υπολογισμό της συνάρτησης διαφέρει ανάμεσα στους νευρώνες.

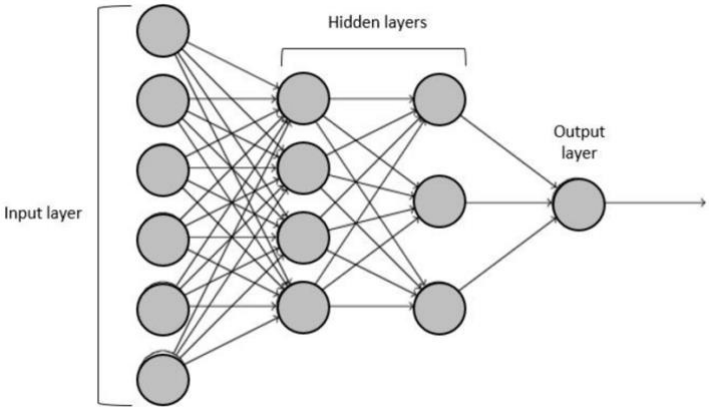
3) Κατά την έξοδο από τον νευρώνα το σήμα πολλαπλασιάζεται με ένα καθορισμένο αριθμό, κατά τον οποίο καθορίζεται η ένταση της εξόδου, και ονομάζεται συναπτικό βάρος, w_i, όπως φαίνεται στην Εικόνα 4.



Εικόνα 4 Νευρώνας Νευρωνικού Δικτύου

4) Η εκπαίδευση των νευρώνων, και κατά επέκταση του νευρωνικού δικτύου το οποίο απαρτίζουν πραγματοποιείται με σκοπό την βελτίωση λειτουργίας του και ικανοποίησης κάποιου κριτηρίου.

Η δομή των νευρωνικών δικτύων αποτελείται συνήθως από επίπεδα, (layers), νευρώνων, όπου τα ενδιάμεσα επίπεδα καλούνται κρυμμένα, (hidden layers). Οι νευρώνες αλληλοεπιδρούν μεταξύ τους διεγείροντας ή αναστέλλοντας την ενεργοποίησή τους, μετά από λήψη του σταθμισμένου αθροίσματος όλων των εισόδων που καταλήγουν σε αυτούς, και παράγουν μέσα από την συνάρτηση ενεργοποίησης, (Εικόνα 5).



Εικόνα 5 Πολυστρωματικό Νευρωνικό Δίκτυο

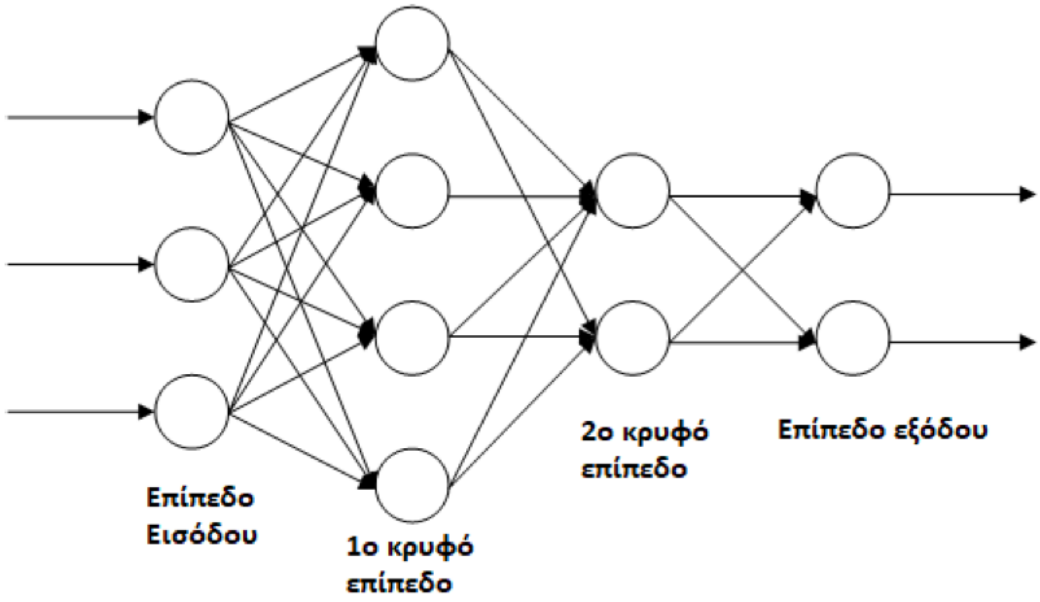
Υπάρχουν διάφορες συναρτήσεις ενεργοποίησης, όπου οι πιο διαδεδομένες αναφέρονται στον Πίνακα 1:

Πίνακας 1 Συναρτήσεις Ενεργοποίησης Νευρώνων

Όνομα	Συνάρτηση	Εύρος
Γραμμική	$F(S) = kS, k \in \Re$	(-∞, +∞)
Ημιγραμμική	$F(S) = \begin{cases} kS, S > o, k \in \Re + \\ 0, S \leq 0 \end{cases}$	[0, ∞)
Σιγμοειδής	$F(S) = (1 + e^{-aS})^{-1}$, $a \in \Re$	(0, 1)
Διπολική Σιγμοειδής	$F(S) = 2(1 + e^{-aS})^{-1} - 1,$ $a \in \Re$	(-1,+1)
Εφαπτομένη Υπερβολής	$F(S) = (e^{aS} - e^{-aS})/(e^{aS} + e^{-aS}),$ $a \in \Re$	(-1, +1)
Εκθετική	$F(S) = e^{-aS}, a \in \Re$	(0, ∞)
Ημιτονοειδής	$F(S) = \sin S$	[-1,+1]
Κλασματική	$F(S) = S/(a + \|S\|), a \in \Re$	[-1,+1]
Κατωφλίου	$F(S) = \begin{cases} 1, S \geq o \\ 0, S < 0 \end{cases}$	[0, 1]
Δυαδικό Κατώφλι	$F(S) = \begin{cases} -1, S \leq -1 \\ S, -1 < S < 1 \\ 1, S \geq 1 \end{cases}$	[-1,+1]

2.3.2.1 Πολυστρωματικό Νευρωνικό Δίκτυο

Το πολυστρωματικό νευρωνικό δίκτυο αποτελεί την πιο συνήθης μορφή ενός νευρωνικού δικτύου. Η ονομασία του αναφέρεται στην πολυστρωματική δομή των νευρώνων από τους οποίους αποτελείται, και τον πρόσθιο τρόπο μετάδοσης της πληροφορίας, όπως φαίνεται στην Εικόνα 6:



Εικόνα 6 Πολυστρωματικό Νευρωνικό Δίκτυο

Επιπλέον χαρακτηριστικά των πολυστρωματικών νευρωνικών δικτύων αποτελούν:

1) Πλήρη διασύνδεση των νευρώνων, δηλαδή κάθε νευρώνας είναι συνδεδεμένος με όλους τους νευρώνες του προηγούμενου επιπέδου
2) Κάθε νευρώνας περιέχει μη γραμμική συνάρτηση ενεργοποίησης, συνήθως τη σιγμοειδή. Η συνάρτηση αυτή είναι παραγωγίσιμη καθώς οι μέθοδοι βελτιστοποίησης/εκπαίδευσης κάνουν χρήση παραγώγων.
3) Δύο είδη σημάτων με τα λειτουργικά σήματα, τα οποία είναι τα σήματα εισόδου στο δίκτυο, τα οποία στην συνέχεια διαδίδονται/υπολογίζονται ως συνάρτηση των εισόδων των νευρώνων για την εμφάνισή τους ως σήμα εξόδου τελικά, και τα σήματα σφάλματος τα οποία δημιουργούνται σε ένα νευρώνα και διαδίδονται προς τα πίσω διαμέσου του δικτύου.
4) Χρήση της τεχνικής οπισθοδιάδοσης του σφάλματος, όπως αναφέρεται στο 3, για εκπαίδευση.

Έστω πολυστρωματικό νευρωνικό δίκτυο, δύο στρωμάτων, το οποίο δέχεται n εισόδους, $x_1, x_2, \dots, x_n$. Τότε αν το πρώτο κρυφό επίπεδο νευρώνων πλήθους p που χρησιμοποιεί την σιγμοειδή συνάρτηση ενεργοποίησης f, έχουν έξοδο

$$y_i = f\left(\sum_{j=1}^{n}\left(w_{ij}x_j + \theta_i\right)\right), \forall i = 1, \dots, p$$

Και συνεπώς το δεύτερο επίπεδο νευρώνων προσθέτει τις ενεργοποιήσεις του πρώτου επιπέδου χρησιμοποιώντας συναπτικά βάρη $\omega_1, \omega_2, \dots, \omega_p$, άρα πραγματοποιείται έξοδος:

$$g(x_1, x_2, \dots, x_n) = \sum_{i=1}^{p} \omega_i y_i - \theta$$

Υπάρχει κατάλληλος ακέραιος p και τιμές ω_i, w_{ij}, ώστε η $g(x_1, x_2, \dots, x_n)$ να είναι δυνατό να προσεγγίσει οποιαδήποτε συνάρτηση $\varphi(x_1, x_2, \dots, x_n)$, $\forall \varepsilon > 0$, όπου ε το σφάλμα.

Η διαδικασία εκπαίδευσης με οπισθοδιάδοση του σφάλματος είναι η διαδικασία ρύθμισης των συναπτικών βαρών. Αποτελείται από δύο περάσματα του νευρωνικού δικτύου, όπου στο εμπρόσθιο πέρασμα πραγματοποιείται η εφαρμογή ενός διανύσματος στην είσοδο του δικτύου και παρακολούθηση της επίδρασης με σταθερά συνοπτικά βάρη, με τελικό στάδιο την έξοδο - απόκριση του δικτύου, και το οπίσθιο πέρασμα όπου τα βάρη μεταβάλλονται σύμφωνα με τον κανόνα διόρθωσης του σφάλματος. Δηλαδή στην διαδικασία αυτή η πραγματική απόκριση του δικτύου αφαιρείται από την επιθυμητή απόκριση, δημιουργώντας ένα σφάλμα κατά το οποίο τα συνοπτικά βάρη προσαρμόζονται με τέτοιο τρόπο ώστε να πραγματοποιηθεί μείωση του σφάλματος.

Έστω ένα πολυστρωματικό νευρωνικό δίκτυο με λ επίπεδα, n εισόδους και m εξόδους, και P διανύσματα εισόδου, και ίδιος πλήθος διανυσμάτων-στόχων. Επίσης έστω $x^p, t^p, p = \{1,2, \dots, P\}$, τα διανύσματα εισόδων και στόχων αντίστοιχα, δηλαδή τα

δεδομένα εκπαίδευσης του δικτύου. Τότε τα διανύσματα εισόδου, εξόδου, και στόχων είναι $x = [x_1, x_2, \dots, x_n]^T, y = [y, y_2, \dots, y_m]^T, t = [t_1, t_2, \dots, t_m]^T$, συνεπώς ιδανικός στόχος της εκπαίδευσης αποτελεί η ισότητα διανυσμάτων στόχων και εξόδου, $y^p = t^p, \forall\, p$. Η βέλτιστη προσέγγιση της επιθυμητής εξόδου έχει ως κόστος το μέσο τετραγωνικό σφάλμα:

$$E = \frac{1}{P}\sum_{p=1}^{P} \|t^P - y^P\|^2 = \frac{1}{P}\sum_{p=1}^{P}\sum_{i=1}^{m}\left(t_i^p - y_i^p\right)^2$$

Η μέθοδος που χρησιμοποιείται για την εκπαίδευση με οπισθοδρόμηση ονομάζεται κατάβαση δυναμικού, και σκοπός της είναι η διόρθωση των συνοπτικών βαρών w_{ij} ώστε να ελαχιστοποιηθεί το μέσο τετραγωνικό σφάλμα Ε.

Η μεταβολή του συνοπτικού βάρους w_{ij} ως προς το χρόνο είναι ίση με το αντίθετο του κλάσματος του μέσου τετραγωνικού σφάλματος ως προς την τιμή του βάρους w_{ij}. Καθώς η μεταβολή του συνοπτικού βάρους ως προς τον χρόνο εξαρτάται από το υπολογιστικό σύστημα στο οποίο εφαρμόζεται, εξομοιώνουμε την εξίσωση στον διακριτό χρόνο k:

$$w_{ij}(\lambda, k+1) - w_{ij}(\lambda, k) = -\beta\frac{\vartheta E}{\vartheta w_{ij}(\lambda, k)}$$

όπου $w_{ij}(\lambda, k)$, το συναπτικό βάρος του j νευρώνα του επιπέδου λ-1, με τον νευρώνα i του επιπέδου λ, β το βήμα εκπαίδευσης και k η χρονική στιγμή.

Το τοπικό σφάλμα του νευρώνα i την χρονική στιγμή k:

$$\delta^k(\lambda) = -\frac{\vartheta E}{\vartheta u_i^k(\lambda)}$$

Και με εφαρμογή του κανόνα αλυσίδας του διαφορικού λογισμού:

$$\frac{\vartheta E}{\vartheta u_i^k(\lambda)} = -\frac{\vartheta E}{\vartheta u_i^k(\lambda)}\frac{\vartheta u_i^k(\lambda)}{\vartheta w_{ij}(\lambda,k)} = -\delta_i^k(\lambda)\frac{\vartheta u_i^k(\lambda)}{\vartheta w_{ij}(\lambda,k)}$$

όπου $u_i^k(\lambda)$η δικτυακή διέγερση του νευρώνα i και ισούται με το άθροισμα των διεγέρσεων των νευρώνων του προηγούμενου επιπέδου πολλαπλασιασμένο με τα συνοπτικά βάρη $w_{ij}(\lambda,k)$, με:

$u_i^k(\lambda) = \sum_{j=1}^{n} w_{ij}(\lambda,k)y_j^k(\lambda-1) + w_{i0}(\lambda,k)$, και

$$y_i^k(\lambda) = f(u_i^k(\lambda))$$

Άρα $\frac{\vartheta u_i^k(\lambda)}{\vartheta w_{ij}(\lambda,k)} = y_j^k(\lambda-1)$, οπότε $\frac{\vartheta E}{\vartheta w_{ij}(\lambda,k)} = -\delta_i^k(\lambda)y_j^k(\lambda-1)$, για $j = 0,1,\dots,n$, για $\lambda = 0,1,\dots,L$

Για την εύρεση του σφάλματος σε κάθε νευρώνα, πραγματοποιούμε εκκίνηση της μεθόδου από το τελευταίο L επίπεδο και φτάνουμε στο πρώτο:

- για το επίπεδο L: $\delta_i^k(L) = -\frac{\vartheta E}{\vartheta u_i^k(L)} = -\frac{\vartheta E}{\vartheta y_i^k(L)}\frac{\vartheta y_i^k(L)}{\vartheta u_i^k(L)} \xrightarrow{\frac{\vartheta E}{\vartheta y_i^k(L)} = -\left(t_i^k - y_i^k\right)} \Longrightarrow \delta_i^k(L) = \left(t_i^k - y_i^k\right)f'(u_i^k(L))$

Δηλαδή το τοπικό σφάλμα είναι η διαφορά της εξόδου του νευρώνα i από την αντίστοιχη επιθυμητή έξοδο, πολλαπλασιασμένη επί την παράγωγο της συνάρτησης ενεργοποίησης που χρησιμοποιεί το νευρώνας.

- για το σφάλμα του επιπέδου λ=1,2,…,L-1 : : $\delta_i^k(\lambda) = -\frac{\vartheta E}{\vartheta u_i^k(\lambda)} =$

$$= -\sum_{\mu=1}^{N(\lambda+1)} \frac{\vartheta E}{\vartheta u_\mu^k(\lambda+1)}\frac{\vartheta u_\mu^k(\lambda+1)}{\vartheta y_i^k(\lambda)}\frac{\vartheta y_i^k(\lambda)}{\vartheta u_i^k(\lambda)} \Rightarrow$$

$$\Rightarrow \delta_i^k(\lambda) = \sum_{\mu=1}^{N(\lambda+1)} \delta_\mu^k(\lambda+1)\, w_{\mu i}^k(\lambda+1) f'(u_i^k(\lambda))$$

όπου $\delta_\mu^k(\lambda+1)$, το σφάλμα ενός νευρώνα μ στο επίπεδο λ+1.

Προκύπτει λοιπόν ότι το σφάλμα σε κάθε νευρώνα στο επίπεδο λ, είναι συνάρτηση των σφαλμάτων του επόμενου επιπέδου λ+1. Συνεπώς η τελική μορφή της μεθόδου οπισθοδρόμησης είναι

$$w_{ij}(\lambda, k+1) = w_{ij}(\lambda, k) + \beta \delta_i^k(\lambda) y_i^k(\lambda-1)$$

Η τιμή του βήματος εκπαίδευσης β, επηρεάζει την ταχύτητα σύγκλισης του δικτύου, αλλά και την συμπεριφορά σύγκλισης του αλγορίθμου. Μια μικρή τιμή του β έχει ως αποτέλεσμα αργή σύγκλιση του αλγορίθμου στο κοντινότερο τοπικό ελάχιστο,(λόγω χρήσης παραγώγου), ενώ μεγάλες τιμές του β πραγματοποιούν πιο γρήγορη σύγκλιση η οποία θα είναι κακή λόγω μεγάλων μεταβολών στα συνοπτικά βάρη τα οποία θα οδηγηθούν πολλές φορές στο ∞.

Ο τερματισμός της εκπαιδευτικής διαδικασίας οπισθοδρόμησης πραγματοποιείται συνήθως με βάση κάποιου ορίου κόστους ή ορίου μεταβολής σφάλματος το οποίο ορίζει ο χρήστης, δηλαδή όταν:

$$E = \frac{1}{P} \sum_{p=1}^{P} \sum_{i=1}^{m} \left(t_i^p - y_i^p\right)^2 < \varepsilon \text{ ή } E(n-1) - E(n) < \varepsilon$$

ή και με ένα προκαθορισμένο αριθμό περασμάτων του δικτύου, δηλαδή ενός πλήθους περασμάτων της μεθόδου οπισθοδρόμησης για την καλύτερη δυνατή προσαρμογή των συναπτικών βαρών.

Η αρχιτεκτονική ενός νευρωνικού δικτύου αποτελεί καθοριστικό παράγοντα της εφαρμογής τους καθώς σε αυτόν βασίζεται σαν πρώτο στάδιο η απόδοσή του. Από την αρχή της σύλληψης των νευρωνικών δικτύων έχει πραγματοποιηθεί πληθώρα

ερευνών κατά τις οποίες ορίζονται κανόνες για τον ορισμό του πλήθους των κρυφών επιπέδων (hidden layers) ενός νευρωνικού δικτύου, και των νευρώνων που περιέχονται σε αυτά. Η έρευνα των K. Gnana Sheela and S. N. Deepa, με τίτλο “Review on Methods to Fix Number of Hidden Neurons in Neural Networks”,[48] , πραγματοποιεί μια πολύ καλή αναφορά σε μεγάλο πλήθος των μεθόδων που ακολουθούνται για τον ορισμό του πλήθους των κρυφών νευρώνων, και καταλήγει με μια νέα.

Ο Timothy Masters, [35], αναφέρει τον κανόνα της γεωμετρικής πυραμίδας ο οποίος και ακολουθείται στην εργασία αυτή σαν πρώτο στάδιο αρχιτεκτονικής στην εφαρμογή των νευρωνικών δικτύων. Σύμφωνα με τον κανόνα αυτό ο αριθμός των νευρώνων σε διαδοχικά επίπεδα νευρωνικού δικτύου ακολουθούν σχήμα πυραμίδας καθώς μειώνεται από την είσοδο του νευρωνικού δικτύου προς την έξοδο. Ακολουθείται μία γεωμετρική αλληλουχία από το πλήθος των νευρώνων σε διαδοχικά επίπεδα, όπου για νευρωνικό δίκτυο με n και m αριθμό νευρώνων εισόδου και εξόδου αντίστοιχα :

- Ένα κρυφό επίπεδο θα έχει πλήθος νευρώνων: $NHN = \sqrt{n \times m}$
- Δύο κρυφά επίπεδα θα έχουν: $NHN_1 = m \times r^2, NHN_1 = m \times r, όπου\ r = \sqrt[3]{n/m}$
- Και ούτω καθεξής.

Όλοι οι μέθοδοι όμως είναι στην πλειοψηφία τους εφαρμοσμένοι σε συγκεκριμένου τύπου δεδομένα, με αποτέλεσμα η επιλογή σε κάθε εφαρμογή των νευρωνικών δικτύων να είναι διαφορετική. Συνεπώς η αρχιτεκτονική ενός πολυστρωματικού νευρωνικού δικτύου είναι αποτέλεσμα συνεχών επαναλήψεων του νευρωνικού δικτύου με σκοπό την εύρεση αυτής που αποδίδει μέγιστα για τα εκάστοτε δεδομένα.

Λοιποί παράγοντες που επηρεάζουν την απόδοση αλλά και την δομή του δικτύου αποτελούν ο ρυθμός εκμάθησης, οι επαναλήψεις που αφήνει ο αναλυτής το δίκτυο ώστε να εκπαιδευτεί, η αρχικοποίηση των συναπτικών βαρών, το μέγεθος των δεδομένων εκπαίδευσης, και άλλοι, που αν και σημαντικοί δεν αποτελούν κομμάτι του πεδίο έρευνας της εργασίας.

2.5 Ανασκόπηση Σύγκρισης Προβλεπτικών Μεθόδων Εξόρυξης Δεδομένων

Με την συνεχώς αύξουσα παραγωγή δεδομένων, η πρόκληση που αντιμετωπίζει κάθε αναλυτής είναι η επιλογή της τεχνικής την οποία θα εφαρμόσει σε αυτά. Πολλές φορές τα δεδομένα δεν συνοδεύονται από την πλήρη επεξήγηση που θα οδηγήσουν στην επιλογή της μεθόδου Α αντί της μεθόδου Β. Στόχος του αναλυτή είναι η κατανόηση των δεδομένων, αλλά και των διαθέσιμων μεθόδων εξόρυξης που έχει στην διάθεσή του, με σκοπό την επιλογή της κατάλληλης γνωρίζοντας τα πλεονεκτήματα και τα μειονεκτήματα της κάθε μίας. Με την εξέλιξη των προβλημάτων από απλά γραμμικά σε μη γραμμικά, και η αύξηση του όγκου των δεδομένων προς εξόρυξη, οδήγησε στην εφαρμογή τεχνικών όπως τα τεχνητά νευρωνικά δίκτυα αντί των γραμμικών μεθόδων.

Πολλές έρευνες έχουν πραγματοποιηθεί που επιδεικνύουν την υπεροχή των νευρωνικών δικτύων έναντι των κλασικών μεθόδων παλινδρόμησης:

- Στον τομέα των περιβαλλοντικών ερευνών:

  - οι Stéphanie Manel, Jean-Marie Dias, και Steve J.Ormerod, “Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird”, [37], στην σύγκριση που πραγματοποίησαν μεταξύ νευρωνικών δικτύων και λογιστικής παλινδρόμησης για την κατανομή μίας κατηγορίας πτηνού, συμπεραίνουν ότι τα νευρωνικά δίκτυα ήταν λίγο καλύτερα στην πρόβλεψη σε σχέση με τις γραμμικές μεθόδους όταν εκπαιδεύθηκαν με όλο το πακέτο δεδομένων, αλλά σε κάθε αλλά σε περίπτωση διαχωρισμού των δεδομένων η γραμμική παλινδρόμηση ξεπέρασε σε απόδοση τα νευρωνικά δίκτυα. Επισημαίνεται επίσης ο χρόνος που ήταν αναγκαίος για την εκπαίδευση των νευρωνικών δικτύων σε σχέση με την εφαρμογή των παλινδρομήσεων, αλλά και η διαφοροποίηση της απόδοσης των μοντέλων παλινδρόμησης βασισμένα σε δεδομένα που παρουσιάζουν υψηλή συσχέτιση.

- Οι Ivaldo da Silva Tavares Júnior, Jonas Elias Castro da Rocha, Ângelo Augusto Ebling, Antônio de Souza Chaves, José Cola Zanuncio, Aline Araújo Farias και Helio Garcia Leite, “Artificial Neural Networks and Linear Regression Reduce Sample Intensity to Predict the Commercial Volume of Eucalyptus Clones”, [50], συμπέραναν την πιο ικανοποιητική απόδοση των νευρωνικών δικτύων σε σχέση με μοντέλα γραμμικής παλινδρόμησης. Το σύνολο δεδομένων που χρησιμοποιήθηκε ήταν σχετικά μικρό με 666 εγγραφές.

- Στην έρευνα των Erdi Tosun, Tayfun Ozgur, Ceyla Ozgur, Mustafa Ozcanli, Hasan Serin, Kadir Aydin, “Comparative analysis of various modelling techniques for emission prediction of diesel engine fueled by diesel fuel with nanoparticle additives”, [52], πραγματοποιείται σύγκριση παλινδρόμησης, και νευρωνικών δικτύων για πρόβλεψη των εκπομπών καυσαερίων κινητήρων χρησιμοποιώντας ντίζελ με πρόσθετα. Τα νευρωνικά δίκτυα απέδωσαν προβλέψεις σε πιο ικανοποιητικό βαθμό από την παλινδρόμηση.

- Στον τομέα των οικονομικών:

  - Οι Qing Cao, Karyl B. Leggio, Marc J. Schniederjans, [9], στην σύγκρισή τους ανάμεσα στην απόδοση των νευρωνικών δικτύων και της γραμμικής παλινδρόμηση για την πραγματοποίηση προβλέψεων στο χρηματιστήριο της Κίνας καταλήγουν στο συμπέρασμα ότι τα νευρωνικά δίκτυα υπερτερούν σε όλες τις προβλέψεις που πραγματοποίησαν έναντι των γραμμικών μεθόδων.

  - Οι Reza Gharoie Ahangar, Mahmood Yahyazadehfar, και Hassan Pournaghshaband, “The Comparison of Methods Artificial Neural Network with Linear Regression Using Specific Variables for Prediction Stock Price in Tehran Stock Exchange”, [3], στην πραγματοποίηση σύγκρισης των νευρωνικών δικτύων με την γραμμική παλινδρόμηση, οδηγούνται στο συμπέρασμα ότι τα νευρωνικά δίκτυα είναι πιο αποδοτικά και πιο επιεική στα σφάλματα, για την εκτίμηση τιμών μετοχών στο χρηματιστήριο της Τεχεράνης.

- Οι Sainful Anwar και Kenji Watanabe, “Performance Comparison of Multiple Linear Regression and Artificial Neural Networks in Predicting Depositor Return of Islamic Bank”, [5], πραγματοποίησαν την σύγκριση της πολλαπλής γραμμικής παλινδρόμησης και των νευρωνικών δικτύων για την πρόβλεψη της επιστροφής του καταθέτη. Η μελέτη τους με χρήση των δεδομένων δέκα ετών, όπου περιέχονταν τέσσερις ανεξάρτητες και μια εξαρτημένη μεταβλητή έδειξε ότι τα νευρωνικά δίκτυα είναι ικανά για προβλέψεις με μεγαλύτερη ακρίβεια.

- Στον τομέα της ιατρικής:

  - Οι P.Abdolmaleki, M.Yarmohammadi και M.City, στην σύγκριση νευρωνικών δικτύων και παλινδρόμησης για πρόβλεψη του αποτελέσματος της βιοψίας καρκίνου του μαστού, “Comparison of logistic regression and neural network models in predicting the outcome of biopsy in breast cancer from MRI findings”, [2], απέδωσαν στα νευρωνικά δίκτυα υψηλότερη απόδοση σε σχέση με την παλινδρόμηση. Όμως η παλινδρόμηση με χρήση στατιστικά σημαντικών χαρακτηριστικών απέδωσε στα ίδια επίπεδα όπως και το νευρωνικό δίκτυο. Σημαντικό στοιχειό αποτελεί ο μικρός όγκος δεδομένων στα οποία πραγματοποιήθηκαν οι μέθοδοι εξόρυξης.

  - Οι Behzad Eftekhar, Kazem Mohammad, Hassan Eftekhar Ardebili, Mohammad Ghodsi και Ebrahim Ketabchi, στην σύγκριση νευρωνικών δικτύων και παλινδρόμησης για πρόβλεψη της θνησιμότητας από τραύμα στο κεφάλι, “Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data”, [16], καταλήγουν στην σημαντική προβλεπτική απόδοση των νευρωνικών δικτύων σε σχέση με την παλινδρόμηση. Ο όγκος δεδομένων ήταν 1271, με 24 ανεξάρτητες και μία εξαρτημένη μεταβλητή.

- Στον τομέα της βιομηχανίας:

  - Οι Mingjun Li και Junxing Wang, στην μελέτη για την πρόβλεψη παραμόρφωσης του τσιμέντου σε φράγματα, “An Empirical Comparison of Multiple Linear Regression and Artificial Neural Network for Concrete Dam Deformation Modelling”, [32], λαμβάνουν την βέλτιστη απόδοση από νευρωνικά δίκτυα σε σύγκριση με την πολλαπλή γραμμική παλινδρόμηση, με χρήση μεγάλου όγκου δεδομένων.

- Άλλοι τομείς:

  - Σε έρευνα του πολεμικού ναυτικού της Αμερικής, ο Bradley Steven Russell, [45], πραγματοποίησε μια σύγκριση των νευρωνικών δικτύων και των μοντέλων παλινδρόμησης για πρόβλεψη της συμπεριφοράς επαναστρατολόγησης του προσωπικού, με σχετικά μικρό πακέτο δεδομένων, (780 εγγραφές) με δεκαεπτά μεταβλητές. Η σύγκριση απέφερε την προβλεπτική απόδοση των νευρωνικών δικτύων να είναι όμοια με αυτή της παλινδρόμησης, αν και το νευρωνικό δίκτυο είχε μικρότερη τιμή R^2. Πραγματοποιείται επίσης αναφορά στην ικανότητα του νευρωνικού δικτύου να μην επηρεάζεται από τον θόρυβο που μπορεί να υπάρχει στα δεδομένα, σε σχέση με την πολλαπλή γραμμική παλινδρόμηση.

Στον αντίποδα υπάρχουν και εφαρμογές όπου οι μέθοδοι της παλινδρόμησης υπερείχαν ή είχαν παρόμοια απόδοση με τα νευρωνικά δίκτυα:

- Η Susan L.King στην έρευνα για πραγματοποίηση πρόβλεψης της προηγούμενης διαμέτρου δέντρων “Neural Networks vs. Multiple Linear Regression for Estimating Previous Diameter”, [30], με χρήση πολλαπλής γραμμικής παλινδρόμησης και νευρωνικών δικτύων, καταλήγει στο συμπέρασμα ότι καμία από τις δύο μεθόδους δεν ξεπέρασε την άλλη σε όλες τις δοκιμές που πραγματοποίησε.

- Οι Feng, C.-X., & Wang, X. το 2002, “Digitizing uncertainty modeling for reverse engineering applications: Regression versus neural networks.”, [20], καταλήγουν στην προβλεπτική υπεροχή της πολλαπλής γραμμικής παλινδρόμησης, έναντι των νευρικών δικτύων.

- Οι Ainslie, A., & Dreze, X το 1996, “Data-mining and choice classic models/neural networks.”, [4], πραγματοποιήσαν σύγκριση της προβλεπτικής ικανότητας της παλινδρόμησης και των νευρωνικών δικτύων, όπου σε μία περίπτωση η παλινδρόμηση είχε καλύτερα αποτελέσματα.

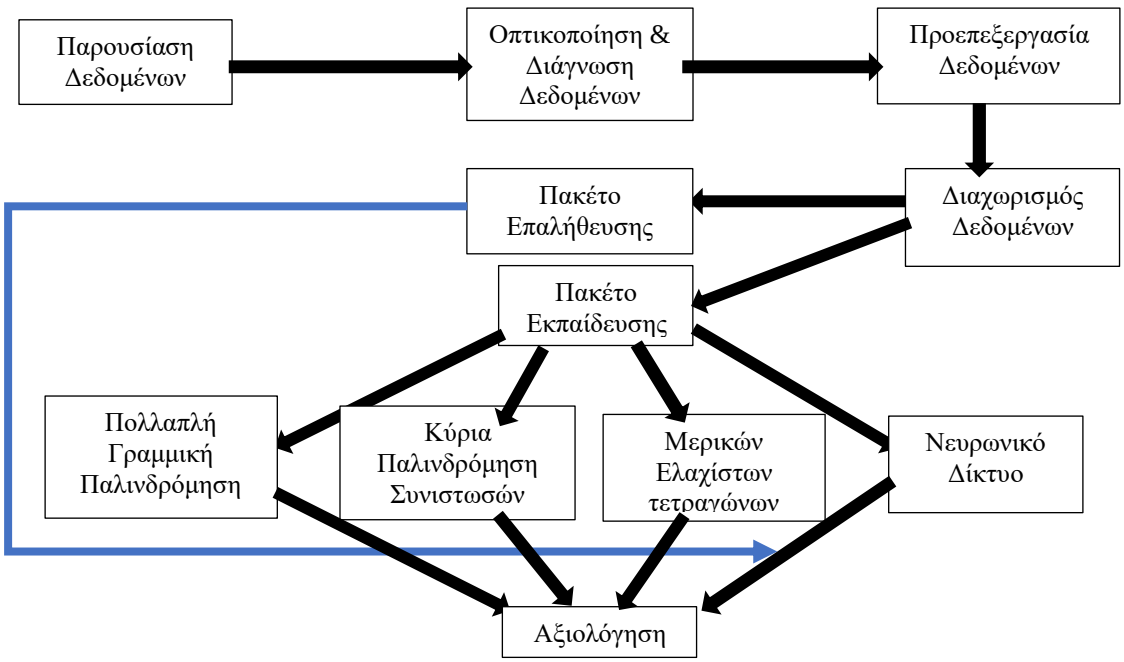
Παρατηρείται ότι αν και τα νευρωνικά δίκτυα έχουν εφαρμογές σε πολλούς τομείς, ξεπερνώντας σε απόδοση τις κλασσικές τεχνικές παλινδρόμησης, υπάρχουν και αποτελέσματα τα οποία επιδεικνύουν όμοια ή και καλύτερη απόδοση των παλινδρομήσεων. Η επιλογή της μεθόδου εξόρυξης δεδομένων επιβαρύνει εξολοκλήρου τον αναλυτή, ο οποίος πολλές φορές επαφίεται στην υπάρχουσα γνώση απόδοσης των μεθόδων από την υπάρχουσα βιβλιογραφία. Η επιλογή μη εξιδεικευμένων πακέτων δεδομένων, στην παρούσα ανάλυση στοχεύει στην γενική καθοδήγηση για την επιλογή της μεθόδου, βάση του μεγέθους των δεδομένων, και των χαρακτηριστικών τους.

3 Μεθοδολογία και Παρουσίαση δεδομένων

Στο παρόν κεφάλαιο αναλύεται η μεθοδολογία αξιολόγησης των προβλεπτικών μεθόδων αξιολόγησης, και παρουσιάζονται τα πακέτα δεδομένων που χρησιμοποιήθηκαν. Επεξηγείται η διαδικασία αξιολόγησης της απόδοσης της κάθε μία από τις τρείς μεθόδους εξόρυξης όταν εφαρμόζεται σε καθένα από τα τρία πακέτα δεδομένων, καθώς και παρουσιάζονται οπτικοποιήσεις των δεδομένων, όπως και τεχνικές αναγνώρισης σχέσεων μεταξύ των μεταβλητών.

3.1 Μεθοδολογία

Πρώτο στάδιο της μεθοδολογίας αποτελεί η παρουσίαση των δεδομένων, ακολουθούμενη από την οπτικοποίησή τους και ενέργειες για την διάγνωση σχέσεων των μεταβλητών, με αποτέλεσμα την λήψη διαφορετικής οπτικής για αυτά. Στο δεύτερο στάδιο πραγματοποιείται η προετοιμασία των δεδομένων, όπως πιθανή κανονικοποίηση. Στο τρίτο στάδιο πραγματοποιείται διαχωρισμός του πακέτου δεδομένων σε πακέτο εκπαίδευσης και πακέτο επαλήθευσης. Στο τέταρτο στάδιο κατασκευάζεται το μοντέλο εξόρυξης δεδομένων με διαφορετικές τεχνικές. Στο πέμπτο στάδιο όλα τα μοντέλα εξόρυξης συγκρίνονται μεταξύ τους. Η διαδικασία επαναλαμβάνεται για καθένα από τα τρία πακέτα δεδομένων.



Γράφημα 3 Μεθοδολογία Εργασίας

3.2 Παρουσίαση Δεδομένων

Η παρούσα διπλωματική εργασία πραγματοποιεί χρήση τεσσάρων πακέτων δεδομένων, τα οποία λήφθηκαν από το UCI Machine Learning Repository, (https://archive.ics.uci.edu/ml/index.php). Τα πακέτα είναι “Divorce Predictors Dataset”, [62], “Skin Segmentation Data set”, [63], και “Wine Quality Data set”, [64]. Η πηγή των δεδομένων επεξηγεί σε ικανοποιητικό βαθμό την φύση των δεδομένων και τι αντιπροσωπεύουν. Η επιλογή των πακέτων δεδομένων πραγματοποιήθηκε με κριτήριο το μέγεθός τους και το πλήθος των μεταβλητών/χαρακτηριστικών τους. Το πακέτο δεδομένων “Divorce Predictors Dataset” έχει 170 εγγραφές με 54 χαρακτηριστικά, το “Skin Segmentation Dataset ” έχει 245.057 εγγραφές και 4 χαρακτηριστικά, και το “Wine Quality Dataset” 4.898 και 12 χαρακτηριστικά.

3.3 Περιγραφή Δεδομένων και προεπεξεργασία

Πριν την εφαρμογή τεχνικών επεξεργασίας των δεδομένων, πραγματοποιήθηκαν γραφικές αναπαραστάσεις των μεταβλητών με σκοπό την κατανόηση του εύρους των δεδομένων και ίσως της σχέσης μεταξύ των μεταβλητών πρόβλεψης και προβλεπόμενων μεταβλητών. Υπολογίζονται οι συντελεστές συσχέτισης των μεταβλητών, με σκοπό την αναγνώριση των συσχετίσεών τους, και για την ανακάλυψη της ύπαρξης μη γραμμικών σχέσεων, η ανάλυση και το γράφημα κυρίων συνιστωσών.

3.3.1 Περιγραφή και επεξεργασία του “Divorce Predictors Dataset”

Το πακέτο δεδομένων “Divorce Predictors Dataset” λήφθηκε από την ιστοσελίδα του UCI Machine Learning Repository, (https://archive.ics.uci.edu/ml/index.php). Αποτελείται από 170 εγγραφές ερωτηματολόγιων σχετικά με την σχέση ζευγαριών που είχαν ή όχι πάρει διαζύγιο, με 54 ερωτήσεις/χαρακτηριστικά τους, όπου βαθμολογούσαν σε κλίμακα ακεραίων [0,4] τις διάφορες πλευρές της σχέσης ως ανεξάρτητες μεταβλητές, και τελικά αναφέρεται αν έχουν πάρει ή όχι διαζύγιο.

Τα δεδομένα περιγράφονται πλήρως στην ιστοσελίδα από την οποία λήφθηκαν, (https://archive.ics.uci.edu/ml/datasets/Divorce+Predictors+data+set#), και λόγω του μεγάλου όγκου των χαρακτηριστικών δεν αναφέρονται στην παρούσα εργασία.

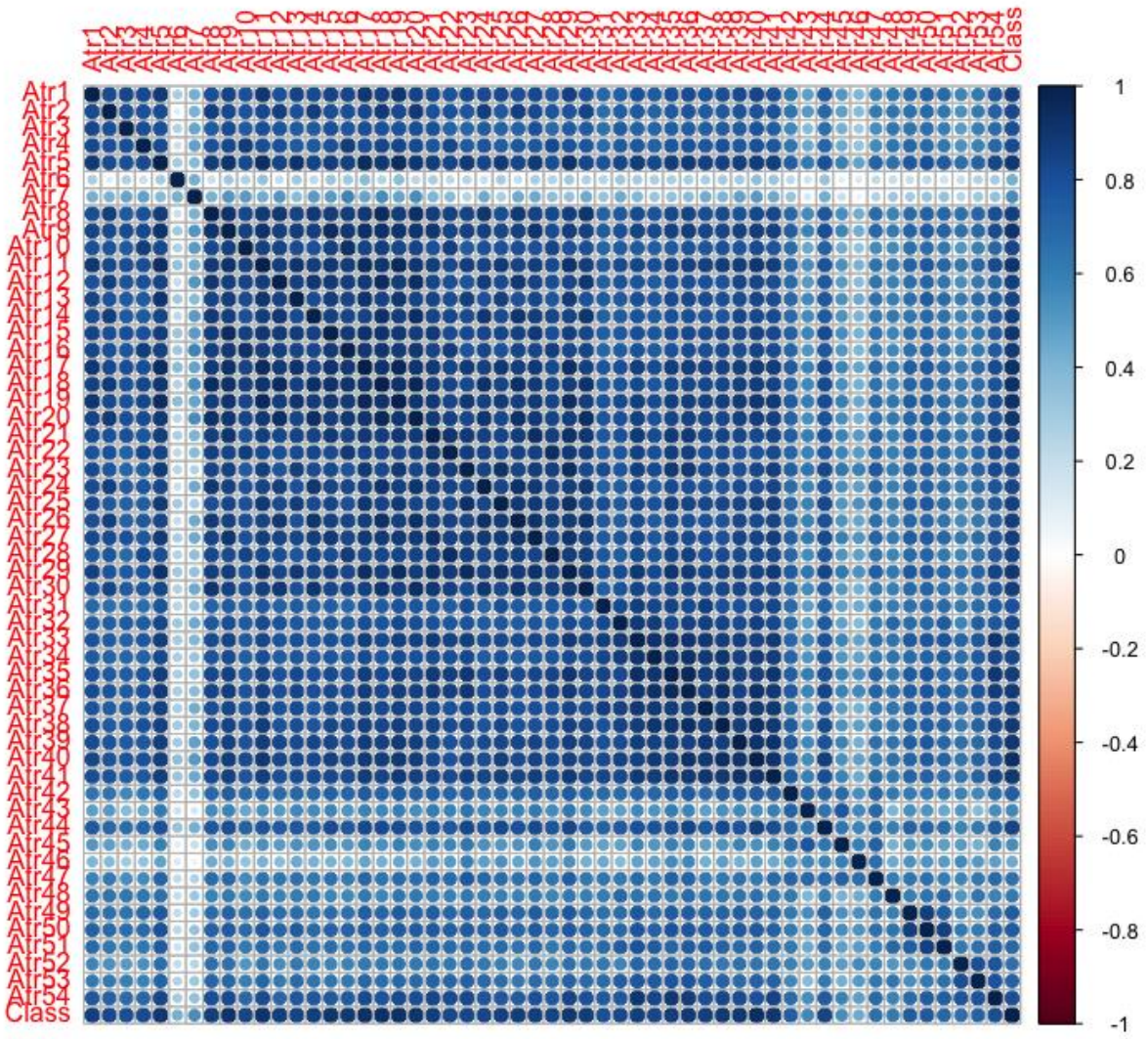
Αρχικά πραγματοποιείται η είσοδος των δεδομένων στο σύστημα και στην συνέχεια έλεγχος για τυχόν ελλιπή δεδομένα, που στην περίπτωση του πακέτου “Divorce Predictors Dataset” δεν υπάρχουν. Στην συνέχεια με την προβολή του γραφήματος θηκογράμματος των μεταβλητών λαμβάνουμε μια εικόνα της κατανομής των τιμών της κάθε μεταβλητής και των ακραίων τιμών της.

Είναι εμφανής η ανάγκη για κανονικοποίηση στα δεδομένα ώστε να έχει κάθε μεταβλητή την ίδια ευκαιρία να εμφανίσει την επίδρασή της στο αποτέλεσμα. Με αυτόν τον τρόπο τα δεδομένα παρουσιάζουν μέση τιμή μηδέν και τυπική απόκλιση ένα. Η προβολή του γραφήματος θηκογράμματος, Γράφημα 4, κάνει εμφανές το αποτέλεσμα της εφαρμογής κανονικοποίησης.

Με τον υπολογισμό συντελεστή συσχέτισης και την προβολή του διαγράμματος συσχέτισης παρατηρούνται οι συσχετίσεις ανάμεσα στις ανεξάρτητες μεταβλητές και την εξαρτημένη μεταβλητή. Υπάρχουν μεταβλητές οι οποίες έχουν υψηλές συσχετίσεις με την εξαρτημένη μεταβλητή, αλλά παρατηρούνται και υψηλές συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών, γεγονός που θα επηρεάσει την αποδοτικότητα κάποιων μεθόδων που θα εφαρμοστούν. Ο βαθμός των συσχετίσεων παρουσιάζεται από το Γράφημα 6, όπως και από τον Πίνακα 2 που δίνει τις υψηλές τιμές των συσχετίσεων των ανεξάρτητων μεταβλητών με την εξαρτημένη.

Γράφημα 4 Θηκόγραμμα Divorce Predictors Dataset

Γράφημα 5 Θηκόγραμμα κανονικοποιημένο Divorce Predictors Dataset



Γράφημα 6 Γράφημα συσχετίσεων Divorce Predictors Dataset

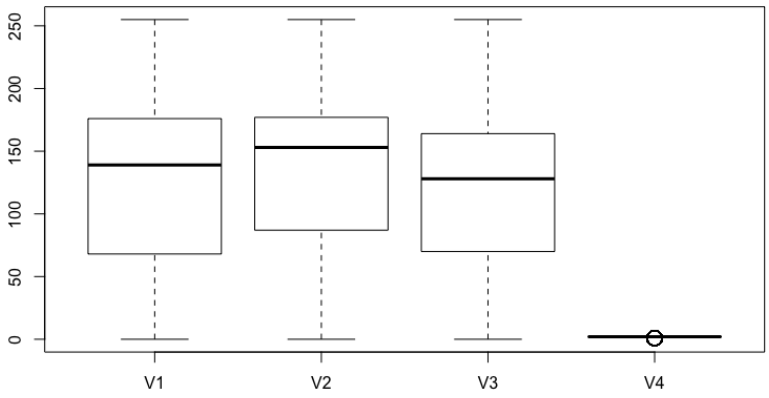
Πίνακας 2 Συντελεστές συσχέτισης ανεξάρτητων με εξαρτημένη μεταβλητή του Divorce Predictors Dataset

Variable	Target Variable	Variable	Target Variable
Atr1	0.861324169432378	Atr25	0.85705246493588
Atr2	0.820774176515677	Atr26	0.872868128336978
Atr3	0.8067085149116	Atr27	0.869788489819457
Atr4	0.819582871908453	Atr28	0.846606097796576
Atr5	0.893179825568263	Atr29	0.892953832675356
Atr6	0.420912543547114	Atr30	0.874530915219955
Atr7	0.54483520017558	Atr31	0.792607196960475
Atr8	0.869569027205944	Atr32	0.829055562995078
Atr9	0.912368228789345	Atr33	0.861328485889379
Atr10	0.83489697795214	Atr34	0.835166742051427
Atr11	0.918385957974587	Atr35	0.86262411242719
Atr12	0.868982725406959	Atr36	0.886497210485215
Atr13	0.844742964443075	Atr37	0.86359674311015
Atr14	0.864316039887655	Atr38	0.88331144020655
Atr15	0.901219664176099	Atr39	0.896179892485547
Atr16	0.886260335835369	Atr40	0.938683632131717
Atr17	0.929346028395035	Atr41	0.894355565439942
Atr18	0.923208317811005	Atr42	0.739629019724654
Atr19	0.928626984410765	Atr43	0.566242199643916
Atr20	0.907007894029731	Atr44	0.847335568861799
Atr21	0.864519285666097	Atr45	0.546449760166005
Atr22	0.825937967948747	Atr46	0.443465065072592
Atr23	0.83750371896621	Atr47	0.656409434579377
Atr24	0.839391866012656	Atr48	0.619830007738496
Atr25	0.85705246493588	Atr49	0.740703970051262
Atr26	0.872868128336978	Atr50	0.755248490505262
Atr27	0.869788489819457	Atr51	0.692680855378161
Atr28	0.846606097796576	Atr52	0.651477889262309
Atr29	0.892953832675356	Atr53	0.711176264075097
Atr30	0.874530915219955	Atr54	0.806765284696907

3.3.2 Περιγραφή και επεξεργασία του “Skin Segmentation Dataset”

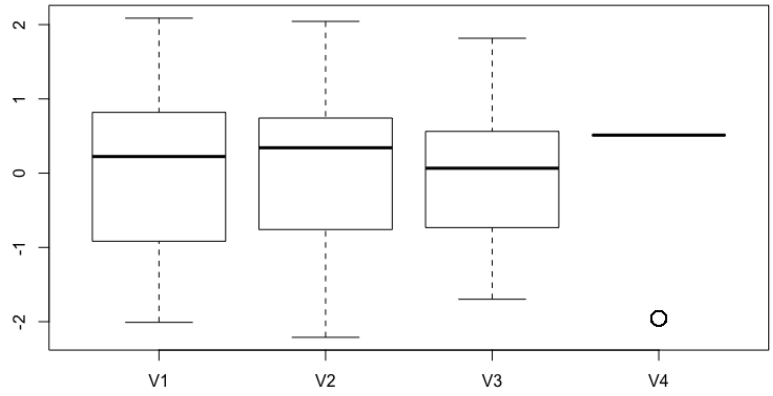
Το πακέτο δεδομένων “Skin Segmentation Dataset” λήφθηκε από την ιστοσελίδα του UCI Machine Learning Repository, (https://archive.ics.uci.edu/ml/index.php). Αποτελείται από 245.057 εγγραφές και 4 μεταβλητές. Τα δεδομένα έχουν παραχθεί από την ανάλυση εικόνων και εξαγωγή του πρότυπου χρώματος RGB, εικόνων προσώπου. Από τις μεταβλητές αυτές τρείς αποτελούν τις ανεξάρτητες με τιμές στο εύρος τιμών RGB, και μία την εξαρτημένη, κατά την οποία κρίνεται είναι δέρμα ανθρώπου ή όχι.

Οι ανεξάρτητες μεταβλητές λαμβάνουν τιμές κυμαίνονται από 0 έως 255, και η εξαρτημένη από 1 έως 2. Η προβολή του θηκογράμματος δείχνει τις διασπορές των μεταβλητών. Το γράφημα δίνει και μια πολύ καθαρή εικόνα για τις ακραίες τιμές των δεδομένων, ειδικά στην εξαρτημένη μεταβλητή, την μεταβλητή του ύψους, και το μέσο φόρτο εργασίας ανά ημέρα. Ακόμα μια πληροφορία του θηκογράμματος του γραφήματος 7 αποτελεί και η ανάγκη για κανονικοποίηση των δεδομένων με αποτέλεσμα μέσης τιμής μηδέν και τυπικής απόκλισης ένα για κάθε μεταβλητή.



Γράφημα 7 Θηκόγραμμα Skin Segmentation Dataset

Με την κανονικοποίηση των δεδομένων λαμβάνουμε κάθε στήλη με μέση τιμή μηδέν και τυπική απόκλιση ένα, δίνοντας σε όλες τις μεταβλητές την ίδια ισχύ σε περιβάλλον ανάλυσης. Το θηκόγραμμα του Γραφήματος 8, μετά την κανονικοποίηση των δεδομένων:



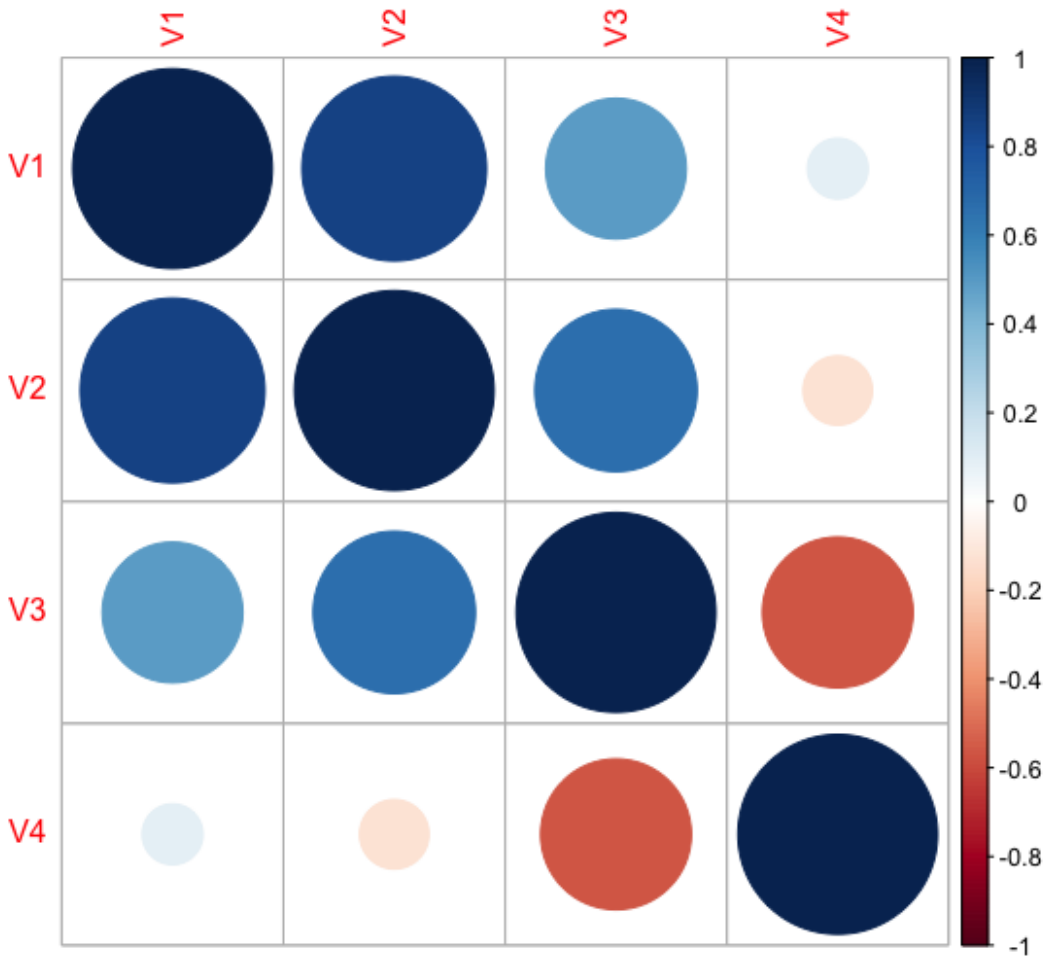
Γράφημα 8 Κανονικοποιημένοο Skin Segmentation Dataset

Ο υπολογισμός των συντελεστών συσχέτισης δίνει το παρακάτω αποτέλεσμα του Πίνακα 3:

Πίνακας 3 Συντελεστές Συσχέτισης Skin Segmentation Dataset

	V1	V2	V3	V4
V1	1	0.8552504203	0.496376343	0.09203009164
V2	0.8552504203	1	0.660097969	-0.1203274404
V3	0.4963763437	0.6600979698	1	-0.5699582232
V4	0.0920300916	-0.120327440	-0.56995822	1

Παρατηρείται πολύ χαμηλή συσχέτιση της ανεξάρτητης μεταβλητής V1, με την εξαρτημένη V4, και σημαντική συσχέτιση της μεταβλητής V3, με την εξαρτημένη V4. Επίσης παρατηρούνται και σχετικά υψηλές συσχετίσεις μεταξύ των μεταβλητών. Το ίδιο αποτέλεσμα δίνει και το γράφημα συσχετίσεων στην Εικόνα 15.



Γράφημα 9 Γράφημα συσχέτισης Skin Segmentation Dataset

3.3.3 Περιγραφή και επεξεργασία του “Wine Quality Dataset”

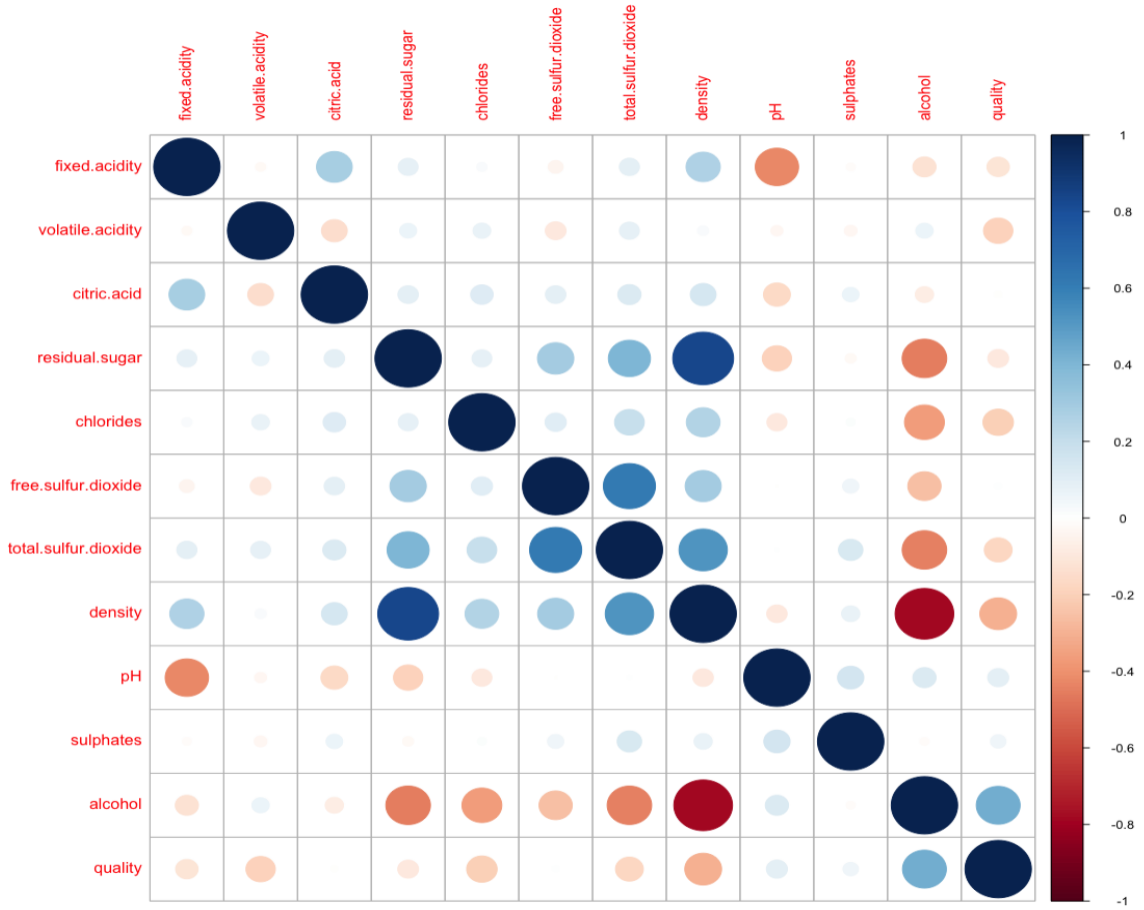
Το πακέτο δεδομένων “Wine Quality Dataset” λήφθηκε από την ιστοσελίδα του UCI Machine Learning Repository, (https://archive.ics.uci.edu/ml/index.php). Αποτελείται από 4.898 εγγραφές και 12 μεταβλητές. Τα δεδομένα αφορούν χαρακτηριστικά οίνων, κόκκινο και λευκό, σε δύο διαφορετικά πακέτα δεδομένων. Στην παρούσα εργασία πραγματοποιείται χρήση του πακέτου δεδομένων με τα λευκά. Οι μεταβλητές είναι:

1) Σταθερή οξύτητα
2) Πτητική οξύτητα
3) Κιτρικό οξύ
4) Υπολειμματική ζάχαρη
5) Χλωριούχα
6) Ελεύθερο διοξείδιο του θείου
7) Σύνολο διοξειδίου του θείου
8) Πυκνότητα
9) Δείκτης pH
10) Θειικά άλατα
11) Επίπεδο αλκοόλ
12) Ποιότητα (βαθμολογία)

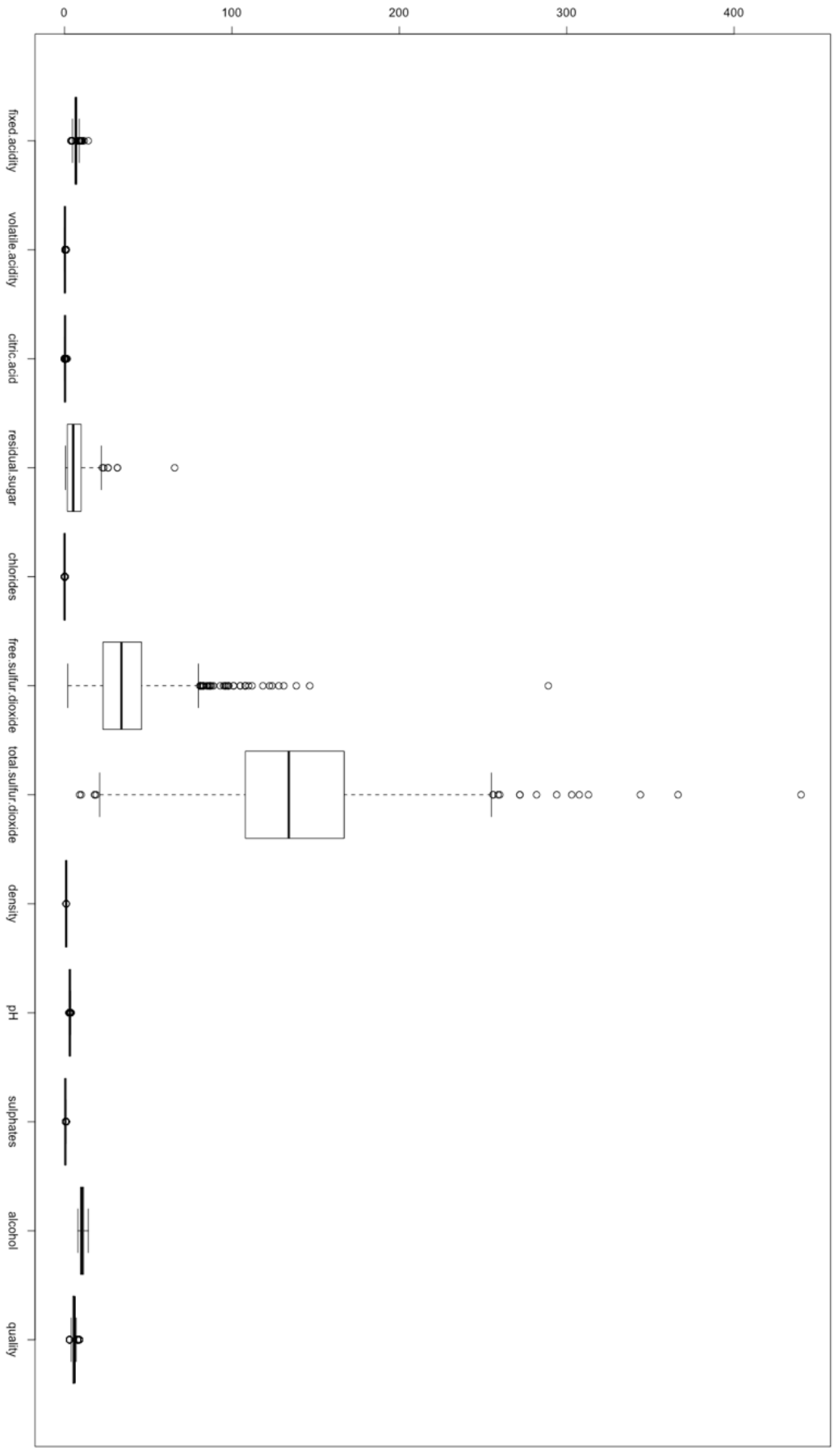
από τις οποίες οι πρώτες 11 είναι οι ανεξάρτητες και η $12_η$ είναι η εξαρτημένη. Σκοπός είναι η πρόβλεψη της βαθμολογίας ενός οίνου σύμφωνα με τα χαρακτηριστικά του.

Οι μεταβλητές παρουσιάζουν εύρος τιμών από 0 έως 440. Η προβολή του θηκογράμματος στο γράφημα 11, παρουσιάζει την διασπορά των μεταβλητών και την αναγκαιότητα για κανονικοποίηση των δεδομένων ώστε να έχουν όλα την ίδια πιθανότητα εμφάνισης μέσα στις μεθόδους εξόρυξης. Με την κανονικοποίηση τα δεδομένα παρουσιάζουν την ίδια συμπεριφορά αλλά με μέση τιμή μηδέν και τυπική απόκλιση ένα. Παρατίθεται και το θηκόγραμμα μετά την κανονικοποίηση, στο γράφημα 12.

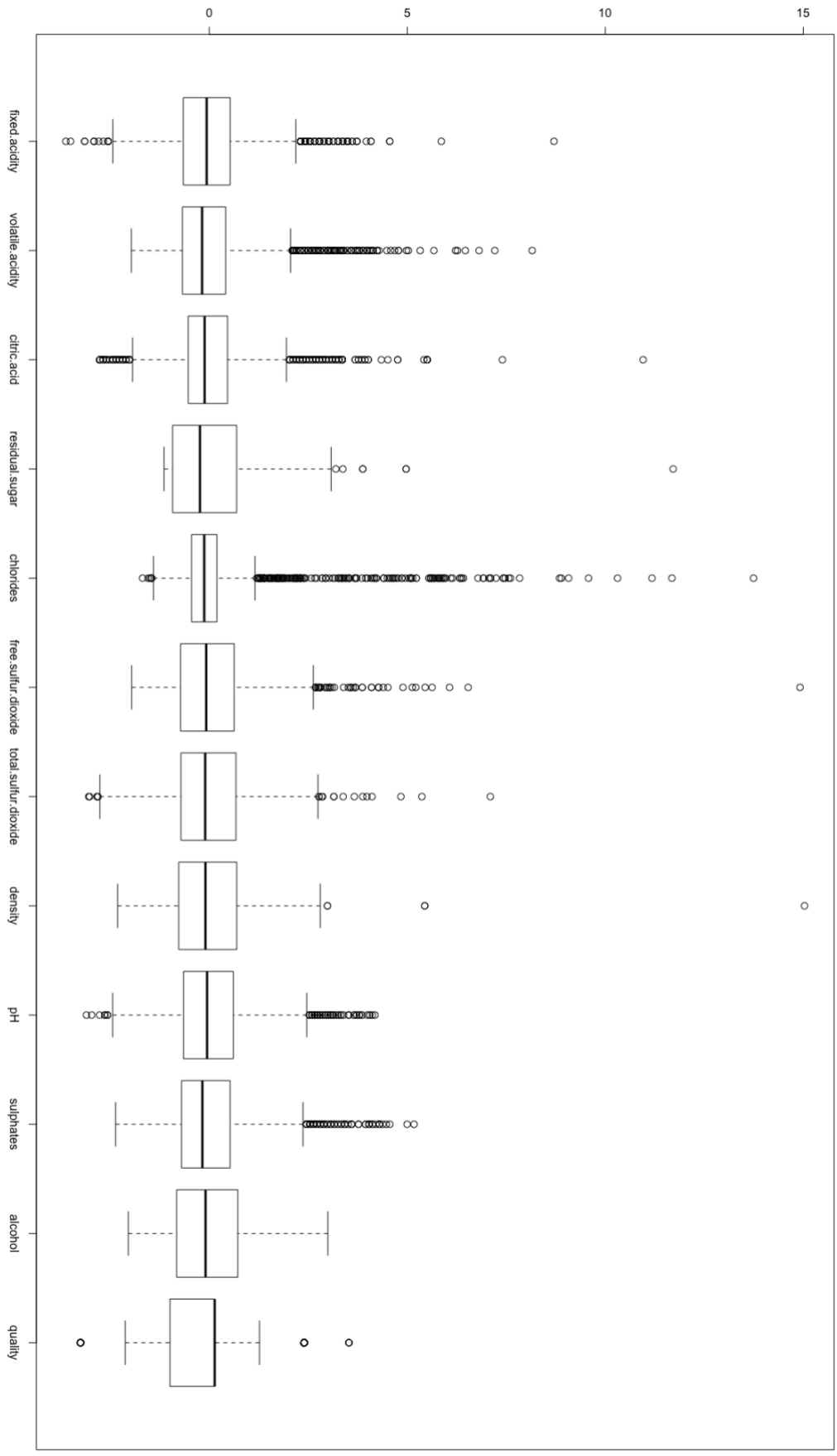
Στην συνέχεια πραγματοποιείται ο υπολογισμός των συντελεστών συσχέτισης των μεταβλητών, και παρατίθενται στον πίνακα 4. Παρατηρούνται κάποιες συσχετίσεις των ανεξάρτητων μεταβλητών με την εξαρτημένη, που όμως δεν μπορούν να χαρακτηρισθούν ως καθοριστικές, όπως η ποσότητα του αλκοόλ με θετική συσχέτιση corrcoeff=0.436 και η πυκνότητα με αρνητική συσχέτιση corrcoeff =-0.307. Στον αντίποδα το πακέτο δεδομένων παρουσιάζει υψηλές συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών, όπως η αρνητική συσχέτιση του αλκοόλ με την πυκνότητα με μεγάλο συντελεστή αρνητικής συσχέτισης -0.780, ή της ποσότητας υπολειμματικής ζάχαρης με την πυκνότητα του οίνου με θετική συσχέτιση 0.839. Μεγάλες συσχετίσεις εξαρτημένων μεταβλητών θα έχουν αρνητικά ή και μεροληπτικά αποτελέσματα. Το ίδιο παρατηρείται για τις ακραίες τιμές των δεδομένων που όμως παραμένουν για την εφαρμογή των προβλεπτικών μεθόδων εξόρυξης, για την μεγαλύτερη δυνατή κάλυψη όλων των περιπτώσεων.



Γράφημα 10 Διάγραμμα Συσχέτισης Wine Quality Dataset



Γράφημα 11 Θηκόγραμμα Wine Quality Dataset



Γράφημα 12 Θηκόγραμμα κανονικοποιημένου Wine Quality Dataset

Πίνακας 4 Συντελεστές Συσχέτισης Wine Quality Dataset

Variable	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
fixed.acidity	1	-0.023	0.289	0.089	0.023	-0.049	0.091	0.265	-0.426	-0.017	-0.121	-0.114
volatile.acidity	-0.023	1	-0.149	0.064	0.071	-0.097	0.089	0.027	-0.032	-0.036	0.068	-0.195
citric.acid	0.289	-0.149	1	0.094	0.114	0.094	0.121	0.15	-0.164	0.062	-0.076	-0.009
residual.sugar	0.089	0.064	0.094	1	0.089	0.299	0.401	0.839	-0.194	-0.027	-0.451	-0.098
chlorides	0.023	0.071	0.114	0.089	1	0.101	0.199	0.257	-0.09	0.017	-0.36	-0.21
free.sulfur.dioxide	-0.049	-0.097	0.094	0.299	0.101	1	0.616	0.294	-0.001	0.059	-0.25	0.008
total.sulfur.dioxide	0.091	0.089	0.121	0.401	0.199	0.616	1	0.53	0.002	0.135	-0.449	-0.175
density	0.265	0.027	0.15	0.839	0.257	0.294	0.53	1	-0.094	0.074	-0.78	-0.307
pH	-0.426	-0.032	-0.164	-0.194	-0.09	-0.001	0.002	-0.094	1	0.156	0.121	0.099
sulphates	-0.017	-0.036	0.062	-0.027	0.017	0.059	0.135	0.074	0.156	1	-0.017	0.054
alcohol	-0.121	0.068	-0.076	-0.451	-0.36	-0.25	-0.449	-0.78	0.121	-0.017	1	0.436
quality	-0.114	-0.195	-0.009	-0.098	-0.21	0.008	-0.175	-0.307	0.099	0.054	0.436	1

4 Αποτελέσματα και Σύγκριση Μεθόδων

Στο προηγούμενο κεφάλαιο πραγματοποιήθηκε η παρουσίαση των δεδομένων, και η προκαταρκτική ανάλυση και επεξεργασία τους. Η διαδικασία αυτή πραγματοποιήθηκε για την αναγνώριση δεν δεδομένων, των μεταβλητών που τα απαρτίζουν και τις σχέσεις που τα διέπουν. Στο κεφάλαιο αυτό εφαρμόζονται οι τεχνικές εξόρυξης στα δεδομένα με σκοπό την αξιολόγηση των προβλεπτικών τους ιδιοτήτων.

4.1 Κριτήρια Αξιολόγησης

Τα κριτήρια σύμφωνα με τα οποία θα αξιολογηθούν οι τεχνικές επιλέχθηκαν με σκοπό την εκτενέστερη αξιολόγησή τους. Σκοπός τους είναι η αξιολόγηση της προβλεπτικής απόδοσής των τεχνικών εξόρυξης, και κατά επέκταση του διαχωρισμού τους για την μελλοντική τους εφαρμογή σε παρόμοιου τύπου δεδομένα.

1) R-square (R^2): αποτελεί κριτήριο της επεξήγησης των δεδομένων από το μοντέλο πρόβλεψης, με τιμές να κυμαίνονται [0,1]
2) R-square adjusted (R^2_{adj}): αποτελεί κριτήριο επεξήγησης των δεδομένων από το μοντέλο πρόβλεψης, με τιμές να κυμαίνονται [0,1], με διαφορά από το R-square την έλλειψη επίδρασης από ακραίες τιμές. Η τιμή του R^2_{adj} αυξάνεται μόνο όταν η πρόσθεση ενός χαρακτηριστικού στο μοντέλο, αυξάνει και το ποσοστό επεξήγησης των δεδομένων.
3) Mean Square Error (MSE), Μέσο Τετραγωνικό Σφάλμα: αποτελεί κριτήριο μέτρησης διαφοράς της προβλεπόμενης τιμής από την πραγματική τιμή. Ένα μοντέλο παράγει καλή ποιότητα προβλέψεων όταν έχει μικρή τιμή μέσου τετραγωνικού σφάλματος, και αντίστροφα.
4) Root Mean Square Error (RMSE), Ρίζα Μέσου Τετραγωνικού Σφάλματος: έχει τιμή όσο η τετραγωνική ρίζα του Μέσου Τετραγωνικού Σφάλματος. Η τιμή του είναι ≥ 0, με μηδέν να υποδεικνύει την πλήρη επεξήγηση των δεδομένων από το μοντέλο.
5) Mean Absolute Error (MAE), Απόλυτο Μέσο Σφάλμα: αποτελεί η μέση τιμή του αθροίσματος των απόλυτων τιμών των σφαλμάτων. Κατέχει πλεονέκτημα

έναντι του μέσου τετραγωνικού σφάλματος καθώς δεν υπερεκτιμά τις ακραίες τιμές των δεδομένων.

6) Αριθμός μεταβλητών: ο αριθμός μεταβλητών που χρησιμοποιήθηκε για την παραγωγή του μοντέλου, συνήθως καθορίζει την προβλεπτική ευστοχία του μοντέλου. Στις περισσότερες περιπτώσεις, μεγαλύτερος όγκος μεταβλητών τείνει να αυξάνει το μέσο τετραγωνικό σφάλμα, καθώς ένα ικανό μοντέλο πρόβλεψης πρέπει να λαμβάνει όσο το δυνατό μεγαλύτερο όγκο των δεδομένων υπόψιν του. Δηλαδή η πρόσθεση μεταβλητών συνήθως αυξάνει την πιθανότητα πρόσθεσης μη αξιόλογης πληροφορίας στο μοντέλο

Αρχικά κάθε πακέτο δεδομένων χωρίζεται τυχαία σε πακέτο εκπαίδευσης και πακέτο αξιολόγησης. Για την παραχώρηση του μεγαλύτερου δυνατού μεγέθους δεδομένων στις μεθόδους πρόβλεψης πραγματοποιείται διαχωρισμός 75:25, όπου το 75% των δεδομένων χρησιμοποιείται για την κατασκευή του μοντέλου και το 25% για την αξιολόγησή του.

Κατά την διαδικασία αυτή οι ανεξάρτητες μεταβλητές του πακέτου εκπαίδευσης χρησιμοποιούνται για την δημιουργία/εκπαίδευση του μοντέλου πρόβλεψης των εξαρτημένων μεταβλητών του πακέτου εκπαίδευσης. Με την εκπαίδευση του μοντέλου, πραγματοποιείται η σύγκριση των προβλέψεων των εξαρτημένων μεταβλητών του πακέτου εκπαίδευσης με τις πραγματικές τιμές των εξαρτημένων μεταβλητών του πακέτου εκπαίδευσης. Αυτό αποτελεί το στάδιο εκπαίδευσης του μοντέλου. Στην συνέχεια, η ικανότητα πρόβλεψης του μοντέλου αποδεικνύεται από την χρήση του ίδιου μοντέλου για την πρόβλεψη των εξαρτημένων μεταβλητών του πακέτου αξιολόγησης, με την χρήση των ανεξάρτητων μεταβλητών του ίδιου πακέτου.

Η αξιολόγηση των προβλέψεων του πακέτου εκπαίδευσης είναι σημαντική καθώς το μοντέλο θεωρητικά αποδίδει το μέγιστο με τα δεδομένα που εκπαιδεύτηκε. Η πραγματική όμως ορθότητα ενός μοντέλου μπορεί να μετρηθεί με την προβλεπτική ικανότητα σε δεδομένα τα οποία δεν χρησιμοποιήθηκαν για την εκπαίδευσή του. Με τον τρόπο αυτό αποφεύγεται η δημιουργία μονόπλευρης συμπεριφοράς, (bias), του μοντέλου.

4.2 Ανάλυση του “Divorce Predictors Dataset”

Το πακέτο δεδομένων “Divorce Predictors Dataset” περιέχει 54 ανεξάρτητες μεταβλητές και μια εξαρτημένη μεταβλητή. Πραγματοποιείται η εφαρμογή τεσσάρων τεχνικών προβλεπτικής εξόρυξης δεδομένων, πολλαπλής γραμμικής παλινδρόμησης με πλήρη χρήση των ανεξάρτητων μεταβλητών και με χρήση της μεθόδου stepwise, κύριας παλινδρόμησης συνιστωσών με χρήση όλων των συνιστωσών και με επιλογή αυτών, μερικής ελαχίστων τετραγώνων με όλες τις συνιστώσες που παράγει η μέθοδος και επιλογή αυτών, και τέλος με νευρωνικά δίκτυα οπισθοδιάδοσης του σφάλματος.

4.2.1 Πολλαπλή Γραμμική Παλινδρόμηση στο “Divorce Predictors Dataset”

Στην τεχνική πολλαπλής γραμμικής παλινδρόμησης εφαρμόστηκαν δύο μέθοδοι, πλήρης και stepwise παλινδρόμησης.

1) Πλήρη Πολλαπλή Γραμμική παλινδρόμηση: Στην τεχνική αυτή λαμβάνονται υπόψιν και οι 54 ανεξάρτητες μεταβλητές του πακέτου δεδομένων για την κατασκευή μοντέλου πολλαπλής γραμμικής παλινδρόμησης με σκοπό την πρόβλεψη της εξαρτημένης μεταβλητής. Δηλαδή:

$$Y_i = \alpha_0 + \alpha_1 \cdot X_{i1} + \alpha_2 \cdot X_{i2} + \alpha_3 \cdot X_{i3} + \cdots + \alpha_{54} \cdot X_{i54} + \varepsilon_i, i = 1, \dots, 170$$

όπου Y_iη εξαρτημένη μεταβλητή και $X_{ij}, i = 1, \dots, 170, j = 1, \dots 54$, η ανεξάρτητη.

Αποτέλεσμα της μεθόδου πολλαπλής γραμμικής παλινδρόμησης είναι 54 συντελεστές που δίνουν την κλίση του υπερεπιπέδου παλινδρόμησης, και μία σταθερά α_0. Επίσης τα σφάλματα της πρόβλεψης ε_i, δηλαδή η διαφορά της προβλεπόμενης από την πραγματική τιμή της εξαρτημένης μεταβλητής λαμβάνουν χώρα στην συνάρτηση πρόβλεψης.

Παρατηρούμε ότι η τιμές των R-squared, και adjusted R-squared είναι μεγάλες, 0.9793 και 0.9638 αντίστοιχα. Δηλαδή το μοντέλο επεξηγεί πάνω από το 96% των εξαρτημένων δεδομένων με τα οποία εκπαιδεύτηκε. Επίσης η τιμή της p-value: < 2.2e-16<0.05, είναι αρκετά μικρή ώστε να υποδεικνύει ότι μπορούμε να βγάλουμε στατιστικά σημαντικό προβλεπτικό συμπέρασμα για την ανεξάρτητη μεταβλητή από όλες τις εξαρτημένες, δηλαδή το μοντέλο στατιστικά είναι αξιόπιστο.

Στην συνέχεια πραγματοποιείται πρόβλεψη για το πακέτο δεδομένων αξιολόγησης με την εισαγωγή των ανεξάρτητων μεταβλητών στο μοντέλο πρόβλεψης, με σκοπό την λήψη των προβλέψεων. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 5:

Πίνακας 5 Πλήρης Πολλαπλή Γραμμική Παλινδρόμηση Divorce Predictors Dataset

MLR	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of Predictors
All predictors	0.9793	0.9638	0.1450159	0.3808096	0.2451010	54

2) Πολλαπλή γραμμική παλινδρόμηση με τεχνική stepwise: Στην τεχνική αυτή εφαρμόζεται αρχικά το πλήρες μοντέλο πολλαπλής γραμμικής παλινδρόμησης και στην συνέχεια εκτελούνται συνεχείς παλινδρομήσεις του πακέτου δεδομένων εκπαίδευσης έως ότι βρεθεί η βέλτιστη. Η επιλογή πραγματοποιείται με βάση το κριτήριο πληροφορίας Akaike, (AIC), κατά το οποίο βελτιστοποιείται η ισορροπία μεταξύ της βελτιστοποίησης του απλούστερου δυνατού μοντέλου και της έλλειψης πληροφορίας που παράγεται με κάθε εκτέλεση της παλινδρόμησης με λιγότερες ανεξάρτητες μεταβλητές. Η μέθοδος κατέληξε σε μοντέλο 32 ανεξάρτητων στατιστικά σημαντικών μεταβλητών, με τιμές των R-squared, και adjusted R-squared 0.9787 και 0.9707 αντίστοιχα, αντιπροσωπεύοντας το 97% της διακύμανσης των εξαρτημένων δεδομένων του πακέτου εκπαίδευσης. Επίσης η τιμή p-value: < 2.2e-16<0.05, είναι αρκετά μικρή ώστε να υποδεικνύει ότι μπορούμε να βγάλουμε στατιστικά ορθό προβλεπτικό συμπέρασμα.

Με την πραγματοποίηση πρόβλεψης των εξαρτημένων μεταβλητών από τις ανεξάρτητες του πακέτου αξιολόγησης έχουμε τα παρακάτω στατιστικά κριτήρια του πίνακα 6:

Πίνακας 6 Πολλαπλές Γραμμικές Παλινδρομήσεις Divorce Predictors Dataset

MLR	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of Predictors
Stepwise	0.9772	0.9701	0.1358568	0.3685875	0.2444359	32
All predictors	0.9793	0.9638	0.1429469	0.3780831	0.2534971	54

Από τον Πίνακα 6, συμπεραίνεται ότι η μέθοδος stepwise πραγματοποίησε μείωση των μεταβλητών που λαμβάνουν μέρος στην παλινδρόμηση για την πρόβλεψη, με μικρές διαφορές στα κριτήρια αξιολόγησης που έχουν τεθεί. Η μείωση των μεταβλητών δεν απέφερε μεγάλη έλλειψη πληροφορίες του μοντέλου παλινδρόμησης, οδηγώντας σε ένα πιο απλό μοντέλο, αντιπροσωπεύοντας το ίδιο ποσοστό δεδομένων εκπαίδευσης με παράλληλη μείωση των σφαλμάτων. Συνεπώς το μοντέλο γραμμικής παλινδρόμησης με την μέθοδο stepwise, απέδωσε καλύτερα από αυτό με χρήση όλων των ανεξάρτητων μεταβλητών.

4.2.2 Κύρια Παλινδρόμηση Συνιστωσών στο “Divorce Predictors Dataset”

Στην τεχνική της Κυρίας Παλινδρόμησης Συνιστωσών εφαρμόστηκαν δύο μέθοδοι, η πλήρης όπου λαμβάνονται υπόψιν όλες οι συνιστώσες για την παλινδρόμηση, και μια η στην οποία επιλέγονται συνιστώσες.

1) Πλήρης Κύρια Παλινδρόμηση Συνιστωσών: Στην τεχνική αυτή εφαρμόζεται η τεχνική ανάλυσης κυρίων συνιστωσών (Principal Component Analysis, PCA), και όλες οι συνιστώσες οι οποίες παράγονται λαμβάνουν μέρος σε πολλαπλή γραμμική παλινδρόμηση με σκοπό την πρόβλεψη της εξαρτημένης μεταβλητής. Τα δεδομένα έχουν κανονικοποιηθεί, ώστε να έχουν όλες οι μεταβλητές την ευκαιρία για εμφάνιση στο μοντέλο, και πραγματοποιείται η ανάλυση σε

ιδιάζουσες τιμές, (Singular Value Decomposition, SVD), για την παραγωγή των κυρίων συνιστωσών.

Με την εκπαίδευση του αλγορίθμου με τα δεδομένα εκπαίδευσης παράγονται 54 κύριες συνιστώσες, κάθε μία από τις οποίες είναι γραμμικός συνδυασμός των αρχικών ανεξάρτητων μεταβλητών. Πάνω σε αυτές τις συνιστώσες πραγματοποιείται η πολλαπλή γραμμική παλινδρόμηση για την πρόβλεψη της εξαρτημένης μεταβλητής, με τα αποτελέσματα του Πίνακα 7.

Πίνακας 7 Παλινδρόμηση Κυρίων Συνιστωσών Divorce Predictors Dataset

PCR	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of PCs
All PCs	0.8540672	0.7425908	0.0803565	0.2834722	0.1770755	54

Παρατηρείται η αρκετά μεγάλη αντιπροσώπευση της διακύμανσης εξαρτημένων δεδομένων από το μοντέλο με τιμές R-squared, και adjusted R-squared 0.8540672 και 0.7425908 αντίστοιχα. Επίσης οι τιμές των σφαλμάτων είναι μικρές δείχνοντας μια καλή ικανότητα πρόβλεψης για τα δεδομένα αξιολόγησης.

2) Κύρια Παλινδρόμηση Συνιστωσών, με επιλογή του πλήθους συνιστωσών: Στην τεχνική αυτή εφαρμόζεται η τεχνική της ανάλυσης κυρίων συνιστωσών αλλά επιλέγονται οι συνιστώσες οι οποίες θα χρησιμοποιηθούν στην δημιουργία του μοντέλου. Από την εκτέλεση του πλήρους μοντέλου της κύριας παλινδρόμησης συνιστωσών, λαμβάνουμε το ποσοστό των ανεξάρτητων μεταβλητών που λαμβάνει υπόψιν του με την επιλογή x πλήθους συνιστωσών, και το ποσοστό των περιπτώσεων των εξαρτημένων μεταβλητών που επεξηγεί.

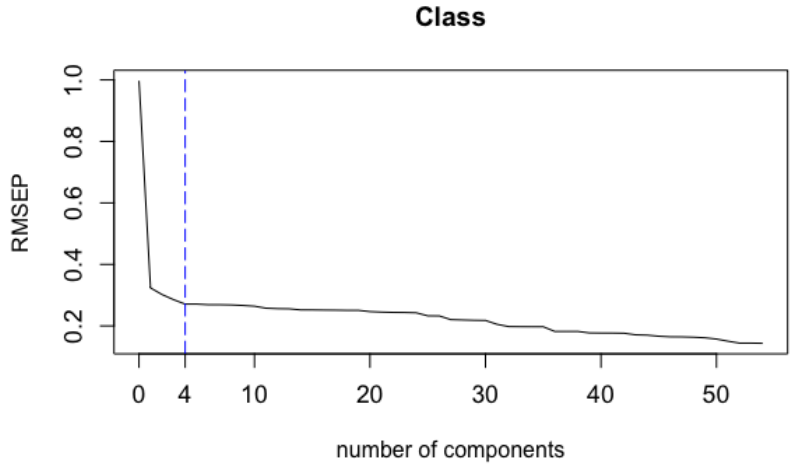
   Τα ποσοστά αυτά παρατίθεται στον παρακάτω πίνακα 8. Είναι μια ισορροπία ανάμεσα στις ανεξάρτητες μεταβλητές που θα χρησιμοποιηθούν για να κατασκευαστεί το μοντέλο, και το ποσοστό των εξαρτημένων που θα είναι δυνατό να προβλεφθούν, καθώς σε ένα μοντέλο πρόβλεψης είναι πάντα απώτερος σκοπός η σωστή πρόβλεψη όσο το δυνατό περισσότερων περιπτώσεων, με τα λιγότερα δυνατά δεδομένα.

Πίνακας 8 Συνιστώσες Παλινδρόμησης Κυρίων Συνιστωσών Divorce Predictors Dataset

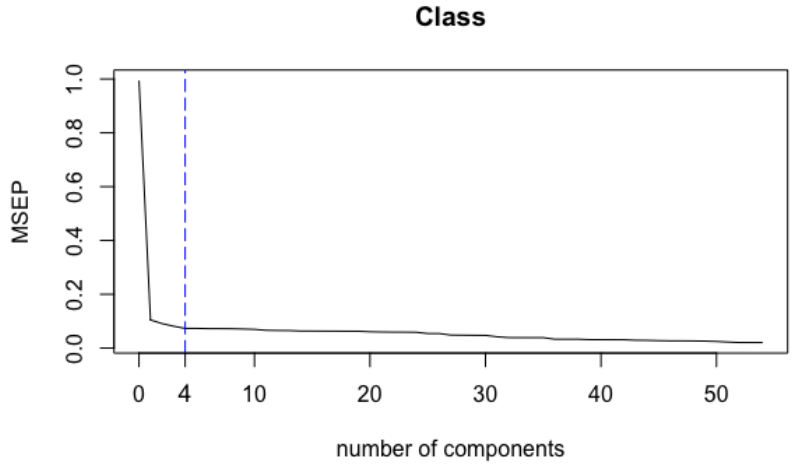
Αριθμός Συνιστωσών	Συνολικό ποσοστό πληροφορίας	Αριθμός Συνιστωσών	Συνολικό ποσοστό πληροφορίας
1	89.47	28	95.16
2	90.83	29	95.21
3	91.82	30	95.23
4	92.62	31	95.77
5	92.62	32	96.06
6	92.72	33	96.08
7	92.73	34	96.09
8	92.76	35	96.09
9	92.86	36	96.67
10	92.99	37	96.67
11	93.33	38	96.67
12	93.41	39	96.84
13	93.43	40	96.85
14	93.61	41	96.86
15	93.62	42	96.88
16	93.64	43	97.06
17	93.66	44	97.10
18	93.68	45	97.21
19	93.69	46	97.28
20	93.89	47	97.29
21	93.97	48	97.32
22	94.02	49	97.38
23	94.02	50	97.51
24	94.08	51	97.73
25	94.55	52	97.91
26	94.57	53	97.91
27	95.11	54	97.93

Παρατηρείται από τον Πίνακα 8, ότι με επιλογή τεσσάρων και πέντε συνιστωσών λαμβάνεται το ίδιο ποσοστό πληροφορίας της εξαρτημένης μεταβλητής, και με αύξηση των συνιστωσών για την κατασκευή του μοντέλου, η διαφορά ποσοστού της πληροφορίας που επεξηγείται παρουσιάζει φθίνουσα συμπεριφορά. Είναι συνεπώς δυνατό να κατασκευαστεί μοντέλο με αρκετά χαμηλό αριθμό συνιστωσών ικανό να προβλέψει το μεγαλύτερο δυνατό ποσοστό της εξαρτημένης μεταβλητής.

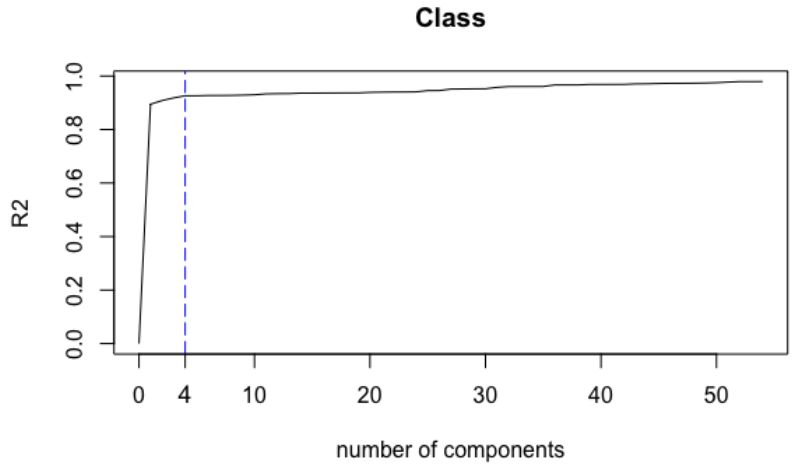
Η παρουσίαση των γραφημάτων της σχέσης του αριθμού των συνιστωσών που χρησιμοποιούνται και των κριτηρίων μέσου τετραγωνικού σφάλματος, (Γράφημα 13), ρίζας μέσου τετραγωνικού σφάλματος,(Γράφημα 14), και R^2,(Γράφημα 15) , δίνει μια εικόνα για την επίδραση του από μια τέτοια επιλογή.



Γράφημα 13 RMSEP - Αριθμός Συνιστωσών PCR - Divorce Predictors Dataset



Γράφημα 14 MSEP - Αριθμός Συνιστωσών PCR - Divorce Predictors Dataset



Γράφημα 15 R^2-Αριθμός Συνιστωσών PCR - Divorce Predictors Dataset

Η επιλογή λοιπόν τεσσάρων συνιστωσών φαίνεται να είναι αρκετή για την μέγιστη δυνατή προβλεπτική ικανότητα με ελάχιστη χρήση κυρίων συνιστωσών, και παράλληλη μείωση των κριτηρίων μέσου τετραγωνικού σφάλματος, ρίζας μέσου τετραγωνικού σφάλματος και R^2. Τα αποτελέσματα της εφαρμογής του μοντέλου με την χρήση τεσσάρων κύριων συνιστωσών φαίνεται στον παρακάτω πίνακα 9, όπως και αυτών του πλήρους μοντέλου:

Πίνακας 9 Παλινδρομήσεις Κυρίων Συνιστωσών - Divorce Predictors Dataset

PCR	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of PCs
4 PCs	0.9317735	0.9295365	0.0677979	0.2603803	0.1363983	4
All PCs	0.8540672	0.7352363	0.0803565	0.2834722	0.1770755	54

Φαίνεται ξεκάθαρη η βελτιστοποίηση της προβλεπτικής ικανότητας του μοντέλου με χρήση τεσσάρων κυρίων συνιστωσών. Η επιλογή αυτή είναι ικανή για πρόβλεψη του 93% των δεδομένων του μοντέλου αξιολόγησης, με παράλληλα απλούστερο μοντέλο σε σχέση με αυτό της χρήσης όλων των κυρίων συνιστωσών. Η συμπεριφορά του αλγορίθμου πρόβλεψης κυριών συνιστωσών σε αυτή την περίπτωση, όπου με μείωση των ανεξάρτητων μεταβλητών, είχαμε καλύτερη πρόβλεψη των εξαρτημένων, εξηγείται από την υψηλή συσχέτιση των αρχικών ανεξάρτητων μεταβλητών. Καθώς οι

αρχικές ανεξάρτητες μεταβλητές παρουσίαζαν υψηλές συσχετίσεις μεταξύ τους, η αναγωγή σε κύριες συνιστώσες έχει ως αποτέλεσμα μικρός αριθμός των συνιστωσών, να περιέχει υψηλό ποσοστό πληροφορίας τους. Αποτέλεσμα αυτού, η επιλογή λιγότερων κυρίων συνιστωσών να οδηγήσει σε καλύτερο μοντέλο πρόβλεψης από το μοντέλο με χρήση όλων των συνιστωσών.

4.2.3 Μερικών Ελαχίστων Τετραγώνων στο “Divorce Predictors Dataset”

Η τεχνική μερικής ελαχίστων τετραγώνων ακολουθεί την ίδια λογική με την τεχνική παλινδρόμησης κυρίων συνιστωσών, με την διαφορά ότι πραγματοποιείται η γραμμική παλινδρόμηση των προβολών, των ανεξάρτητων και των εξαρτημένων μεταβλητών σε νέο σύστημα διαστάσεων. Με την εφαρμογή της τεχνικής αυτής παράγονται όπως και στην μέθοδο παλινδρόμησης κυρίων συνιστωσών, συνιστώσες οι οποίες θα χρησιμοποιηθούν από τον αλγόριθμο για την κατασκευή μοντέλου πραγματοποίησης προβλέψεων. Κατασκευάζονται δύο μοντέλα της μερικής ελαχίστων τετραγώνων μεθόδου, αρχικά με την χρήση όλων των συνιστωσών, και στην συνέχεια με την επιλογή του πλήθους συνιστωσών.

1) Πλήρης τεχνική Μερικών Ελαχίστων Τετραγώνων: Στην τεχνική μερικής ελαχίστων τετραγώνων με χρήση όλων των συνιστωσών, δημιουργούνται 54 συνιστώσες, από τις 54 ανεξάρτητες αρχικές μεταβλητές, για την κατασκευή μοντέλου πρόβλεψης, μέσω της μεγιστοποίησης της συνδιακύμανσης μεταξύ των εξαρτημένων και ανεξάρτητων μεταβλητών.

Τα αποτελέσματα του αλγορίθμου παρατίθενται στον παρακάτω πίνακα 10:

Πίνακας 10 Μερικών Ελαχίστων Τετραγώνων - Divorce Predictors Dataset

PLS	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of PCs
All PCs	0.8540672	0.7352363	0.1319398	0.3632352	0.2330581	54

Παρατηρείται η αρκετά μεγάλη αντιπροσώπευση της διακύμανσης των εξαρτημένων δεδομένων από το μοντέλο με τιμές R-squared, και adjusted R-squared 0.8540672 και 0.7352363 αντίστοιχα. Επίσης οι τιμές των σφαλμάτων είναι σχετικά μικρές δείχνοντας μια καλή ικανότητα πρόβλεψης των δεδομένων αξιολόγησης.

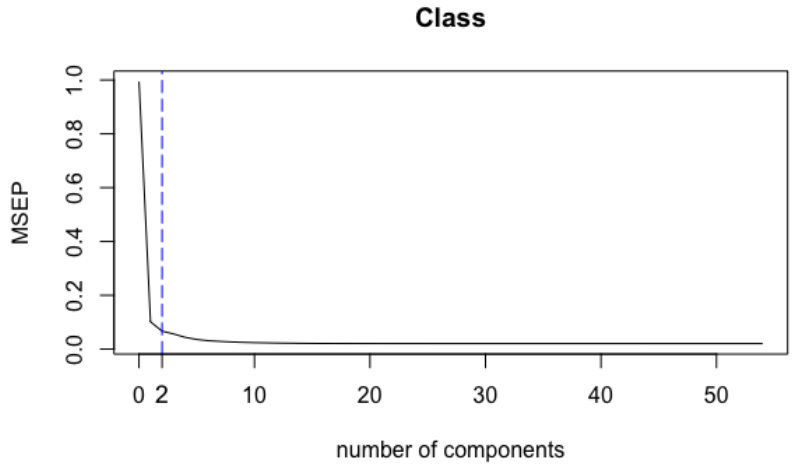
2) Τεχνική Μερικών Ελαχίστων Τετραγώνων με επιλογή συνιστωσών: Στην τεχνική αυτή εφαρμόζεται η μέθοδος της ανάλυσης μερικής ελαχίστων τετραγώνων με επιλογή των συνιστωσών που θα χρησιμοποιηθούν στην δημιουργία του μοντέλου. Από την εκτέλεση του πλήρους μοντέλου, παρατηρείται το ποσοστό των ανεξάρτητων μεταβλητών που λαμβάνει υπόψιν του με την επιλογή x πλήθους συνιστωσών, και το ποσοστό των περιπτώσεων των εξαρτημένων μεταβλητών που επεξηγεί. Τα ποσοστά αυτά παρατίθεται στον παρακάτω πίνακα. Όπως και στην επιλογή των συνιστωσών στην τεχνική κύριας παλινδρόμησης συνιστωσών, χρειάζεται ισορροπία ανάμεσα στις ανεξάρτητες μεταβλητές που θα χρησιμοποιηθούν για να κατασκευαστεί το μοντέλο, και το ποσοστό των εξαρτημένων που είναι δυνατό να προβλεφθούν.

Παρατηρείται στον πίνακα 11 που ακολουθεί ότι η επιλογή από μία έως έξι συνιστώσες είναι ικανές για την επεξήγησης του 89.90% έως και του 96.86% των εξαρτημένων μεταβλητών. Η αύξηση του πλήθους των συνιστωσών πάνω από επτά δεν φαίνεται να αυξάνει το ποσοστό επεξήγησης τόσο ώστε να επιλεγούν για την πρόβλεψη.

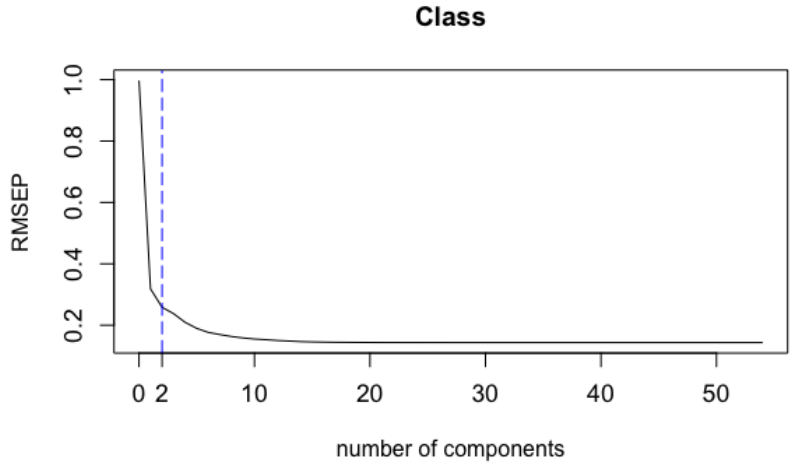
Η προβολή των διαγραμμάτων του μέσου τετραγωνικού σφάλματος, γράφημα 16, ρίζας μέσου τετραγωνικού σφάλματος, γράφημα 17, και R^2, γράφημα 18, με τον αριθμό των συνιστωσών οδηγούν στην επιλογή δύο κυρίων συνιστωσών για την πραγματοποίηση προβλέψεων.

Πίνακας 11 Συνιστώσες Μερικώς Ελαχίστων Τετραγώνων - Divorce Predictors Dataset

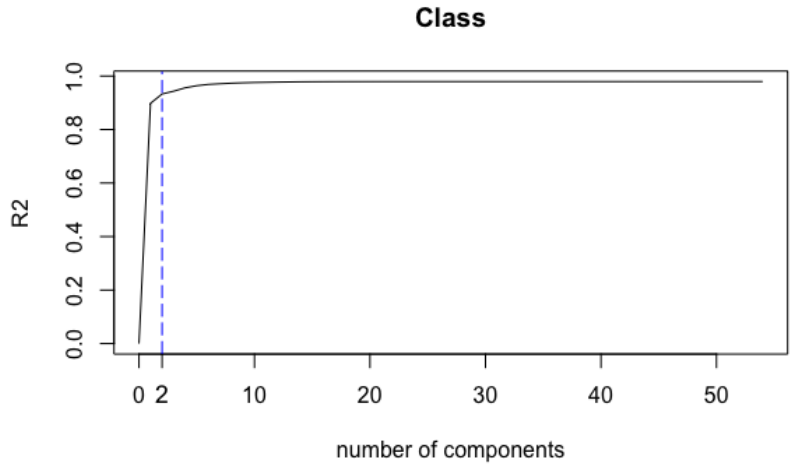
Αριθμός Συνιστωσών	Συνολικό ποσοστό πληροφορίας	Αριθμός Συνιστωσών	Συνολικό ποσοστό πληροφορίας
1	89.80	28	97.93
2	93.28	29	97.93
3	94.33	30	97.93
4	95.59	31	97.93
5	96.38	32	97.93
6	96.86	33	97.93
7	97.11	34	97.93
8	97.32	35	97.93
9	97.47	36	97.93
10	97.58	37	97.93
11	97.65	38	97.93
12	97.73	39	97.93
13	97.78	40	97.93
14	97.84	41	97.93
15	97.87	42	97.93
16	97.89	43	97.93
17	97.90	44	97.93
18	97.91	45	97.93
19	97.92	46	97.93
20	97.92	47	97.93
21	97.93	48	97.93
22	97.93	49	97.93
23	97.93	50	97.93
24	97.93	51	97.93
25	97.93	52	97.93
26	97.93	53	97.93
27	97.93	54	97.93



Γράφημα 16 MSEP - Αριθμός Συνιστωσών PLS - Divorce Predictors Dataset



Γράφημα 17 RMSEP - Αριθμός Συνιστωσών PLS - Divorce Predictors Dataset



Γράφημα 18 R^2 - Αριθμός Συνιστωσών PLS - Divorce Predictors Dataset

Η επιλογή λοιπόν δύο συνιστωσών φαίνεται να είναι αρκετή για την μέγιστη δυνατή προβλεπτική ικανότητα με ελάχιστη χρήση κυρίων συνιστωσών, και παράλληλη μείωση των κριτηρίων μέσου τετραγωνικού σφάλματος, ρίζας μέσου τετραγωνικού σφάλματος και R^2. Τα αποτελέσματα της εφαρμογής του μοντέλου με την χρήση τεσσάρων κύριων συνιστωσών φαίνεται στον παρακάτω πίνακα 12, όπως και αυτών του πλήρους μοντέλου:

Πίνακας 12 Συνολική Μερικών Ελαχίστων Τετραγώνων - Divorce Predictors Dataset

PLS	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of PCs
2 PCs	0.9337532	0.9326847	0.0658306	0.2565747	0.1461756	2
All PCs	0.8540672	0.7352363	0.1319398	0.3632352	0.2330581	54

Φαίνεται ξεκάθαρη η βελτιστοποίηση της προβλεπτικής ικανότητας του μοντέλου με χρήση τεσσάρων κυρίων συνιστωσών. Η επιλογή αυτή είναι ικανή για επεξήγηση του 93% των εξαρτημένων μεταβλητών των δεδομένων αξιολόγησης, παρουσιάζοντας μεγάλη βελτίωση, και κατά πολύ απλούστερο μοντέλο σε σχέση με αυτό της χρήσης όλων των κυρίων συνιστωσών.

Η συμπεριφορά του αλγορίθμου πρόβλεψης μερικώς ελαχίστων τετράγωνων σε αυτή την περίπτωση, όπου με μείωση των ανεξάρτητων μεταβλητών, παρουσίασε καλύτερη πρόβλεψη των εξαρτημένων, εξηγείται από την υψηλή συσχέτιση των αρχικών ανεξάρτητων μεταβλητών, και την παράλληλη συσχέτιση των εξαρτημένων με τις ανεξάρτητες. Καθώς οι αρχικές ανεξάρτητες μεταβλητές παρουσίαζαν υψηλές συσχετίσεις όχι μόνο μεταξύ τους, αλλά και με τις εξαρτημένες μεταβλητές, έχει ως αποτέλεσμα μικρός αριθμός των συνιστωσών, να περιέχει υψηλό ποσοστό πληροφορίας τους. Αποτέλεσμα αυτού, η επιλογή ελάχιστου πλήθους κυρίων συνιστωσών να οδηγήσει σε βελτιστοποίηση του μοντέλου πρόβλεψης σε σχέση με το μοντέλο χρήσης όλων των συνιστωσών.

4.2.4 Νευρωνικά δίκτυα στο “Divorce Predictors Dataset”

Αρχικό στάδιο της κατασκευής του πολυστρωματικού νευρωνικού δικτύου είναι η απόφαση του πλήθους των κρυφών ή ενδιάμεσων επιπέδων νευρώνων που θα έχει και στην συνέχεια το πλήθος των νευρώνων που θα έχει το κάθε επίπεδο. Η αρχιτεκτονική του νευρωνικού δικτύου για το σύνολο δεδομένων “Divorce Predictors Dataset” είναι αποτέλεσμα συνεχών εκτελέσεων, με διαφορετικό πλήθος επιπέδων και πλήθους νευρώνων που εμπεριέχονται σε αυτά.

Μετά από τις συνεχείς εκτελέσεις του νευρωνικού δικτύου, με ένα ή δύο επίπεδα και διαφορετικό πλήθος νευρώνων σε αυτά, καταλήγουμε σε πέντε νευρώνες σε ένα κρυφό επίπεδο το οποίο είχε την βέλτιστη απόδοση. Τα αποτελέσματα του νευρωνικού αυτού δικτύου με από το σύνολο δεδομένων αξιολόγησης παρατίθενται στον παρακάτω πίνακα 13:

Πίνακας 13 Νευρωνικά Δίκτυα - Divorce Predictors Dataset

MLP	R^2	R^2_{adj}	MSE	RMSE	MAE
5 neurons	0.9029246	0.8526326	0.09646549	0.3105889	0.1425251

4.3 Ανάλυση του “Skin Segmentation Dataset”

Το πακέτο δεδομένων “Skin Segmentation Dataset”, περιέχει τρείς ανεξάρτητες μεταβλητές, μια εξαρτημένη μεταβλητή, και 245.057 εγγραφές. Πραγματοποιείται ξανά η εφαρμογή των τεσσάρων τεχνικών προβλεπτικής εξόρυξης δεδομένων, πολλαπλής γραμμικής παλινδρόμησης με πλήρη χρήση των ανεξάρτητων μεταβλητών και με χρήση της μεθόδου stepwise, παλινδρόμησης κυρίων συνιστωσών με χρήση όλων των συνιστωσών και με επιλογή αυτών, μερικής ελαχίστων τετραγώνων με όλες τις συνιστώσες που παράγει η μέθοδος και επιλογή αυτών, και τέλος με νευρωνικά δίκτυα με αλγόριθμο εκπαίδευσης οπισθοδιάδοσης του σφάλματος.

4.3.1 Πολλαπλή Γραμμική Παλινδρόμηση στο “Skin Segmentation Dataset”

Στην τεχνική πολλαπλής γραμμικής παλινδρόμησης εφαρμόστηκαν δύο μέθοδοι, πλήρης και stepwise παλινδρόμησης.

1) Πλήρη Πολλαπλή Γραμμική παλινδρόμηση: Στην τεχνική αυτή λαμβάνονται υπόψιν και οι τρεις ανεξάρτητες μεταβλητές του πακέτου δεδομένων για την κατασκευή μοντέλου πολλαπλής γραμμικής παλινδρόμησης με σκοπό την πρόβλεψη της εξαρτημένης μεταβλητής. Δηλαδή:

$$Y_i = \alpha_0 + \alpha_1 \cdot X_{i1} + \alpha_2 \cdot X_{i2} + \alpha_3 \cdot X_{i3} + \varepsilon_i, i = 1, \dots, 245.057$$

όπου Y_i η εξαρτημένη μεταβλητή και $X_{ij}, i = 1,2,3,\ j = 1, \dots, 245.057$,η ανεξάρτητη.

Αποτέλεσμα της μεθόδου πολλαπλής γραμμικής παλινδρόμησης είναι τρείς συντελεστές που δίνουν την κλίση του υπερεπιπέδου παλινδρόμησης, και η σταθερά α_0. Επίσης τα σφάλματα πρόβλεψης $\varepsilon_i, i = 1, \dots, 245.057$, δηλαδή η διαφορά της προβλεπόμενης από την πραγματική τιμή της εξαρτημένης μεταβλητής λαμβάνουν μέρος στην συνάρτηση πρόβλεψης.

Παρατηρούμε τις τιμές των R-squared, και adjusted R-squared που είναι σχετικά μικρές, 0.511 και οι δύο. Δηλαδή το μοντέλο επεξηγεί περίπου το 51% της διακύμανσης των ανεξάρτητων μεταβλητών με τα δεδομένα που εκπαιδεύτηκε.

Στην συνέχεια πραγματοποιείται πρόβλεψη για το πακέτο δεδομένων αξιολόγησης με την εισαγωγή των ανεξάρτητων μεταβλητών του στο μοντέλο πρόβλεψης, με αποτέλεσμα την λήψη των προβλέψεων των εξαρτημένων μεταβλητών. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 14:

Πίνακας 14 Πλήρης Πολλαπλή Γραμμική Παλινδρόμηση Skin Segmentation Dataset

MLR	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of Predictors
All predictors	0.511	0.511	0.4868090	0.6977170	0.4604120	3

2) Πολλαπλή γραμμική παλινδρόμηση με τεχνική stepwise: Στην τεχνική αυτή εφαρμόζεται αρχικά το πλήρες μοντέλο πολλαπλής γραμμικής παλινδρόμησης και στην συνέχεια εκτελούνται συνεχείς παλινδρομήσεις του πακέτου δεδομένων εκπαίδευσης έως ότι βρεθεί η βέλτιστη. Η μέθοδος στο πακέτο δεδομένων “Skin Segmentation Dataset”, δεν κατέληξε σε μοντέλο πρόβλεψης με αριθμό ανεξάρτητων μεταβλητών μικρότερου του πλήρους μοντέλου. Η εξήγηση του φαινομένου αυτού παρατηρείται στην υψηλή στατιστική σημασία όλων των ανεξάρτητων μεταβλητών για την πρόβλεψη της εξαρτημένης μεταβλητής. Για αυτόν τον λόγο τα αποτελέσματα του μοντέλου πολλαπλής γραμμικής παλινδρόμησης με μέθοδο stepwise, είναι όμοια με αυτά του πλήρους μοντέλου πολλαπλής γραμμικής παλινδρόμησης, όπως φαίνονται στον Πίνακα 15:

Πίνακας 15 Πολλαπλές Γραμμικές Παλινδρόμησεις - Skin Segmentation Dataset

MLR	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of Predictors
Stepwise	0.511	0.511	0.4868090	0.6977170	0.4604120	3
All predictors	0.511	0.511	0.4868090	0.6977170	0.4604120	3

Από τον παραπάνω πίνακα 15, συμπεραίνεται ότι η μέθοδος stepwise δεν πραγματοποίησε μείωση των μεταβλητών που λαμβάνουν μέρος στην παλινδρόμηση για την πρόβλεψη, με αποτέλεσμα να μην υπάρχει διαφορά στα κριτήρια αξιολόγησης που έχουν τεθεί. Καθώς δεν υπήρξε μείωση των εξαρτημένων μεταβλητών, η μέθοδος stepwise δεν πραγματοποίησε βελτίωση του μοντέλου πολλαπλής γραμμικής παλινδρόμησης, αλλά ούτε και απλοποίησή του.

4.3.2 Παλινδρόμηση Κύριων Συνιστωσών στο “Skin Segmentation Dataset”

Στην τεχνική της παλινδρόμησης κύριων συνιστωσών εφαρμόστηκαν δύο μέθοδοι, η πλήρης όπου λαμβάνονται υπόψιν όλες οι συνιστώσες για την παλινδρόμηση, και μια κατά την οποία επιλέγεται ο αριθμός των συνιστωσών

1) Πλήρης Παλινδρόμηση Κύριων Συνιστωσών: Στην τεχνική αυτή εφαρμόζεται η τεχνική ανάλυσης κυρίων συνιστωσών (Principal Component Analysis, PCA), και όλες οι συνιστώσες οι οποίες παράγονται λαμβάνουν μέρος σε πολλαπλή γραμμική παλινδρόμηση με σκοπό την πρόβλεψη της εξαρτημένης μεταβλητής. Τα δεδομένα έχουν κανονικοποιηθεί, ώστε να έχουν όλες οι μεταβλητές την ευκαιρία για εμφάνιση στο μοντέλο, πριν πραγματοποιηθεί η ανάλυση των κυρίων συνιστωσών.

Με την εκπαίδευση του αλγορίθμου με τα δεδομένα εκπαίδευσης παράγονται 3 κύριες συνιστώσες, κάθε μία από τις οποίες είναι γραμμικός συνδυασμός των αρχικών μεταβλητών. Πάνω σε αυτές τις συνιστώσες πραγματοποιείται η πολλαπλή γραμμική παλινδρόμηση για την πρόβλεψη της εξαρτημένης μεταβλητής. Τα αποτελέσματα της παλινδρόμησης κυρίων συνιστωσών παρατίθεται στον παρακάτω πίνακα 16:

Πίνακας 16 Παλινδρόμηση Κυρίων Συνιστωσών Skin Segmentation Dataset

PCR	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of PCs
All PCs	0.5135587	0.5135428	0.6445043	0.8028103	0.5662549	3

Παρατηρείται η σχετικά μικρή αντιπροσώπευση της διακύμανσης της εξαρτημένης μεταβλητής από το μοντέλο με R-squared, και adjusted R-squared 0.5135587 και 0.5135428 αντίστοιχα. Επίσης οι τιμές των σφαλμάτων είναι τέτοιες που παρουσιάζουν όχι τόσο ικανή δυνατότητα πρόβλεψης των δεδομένων αξιολόγησης.

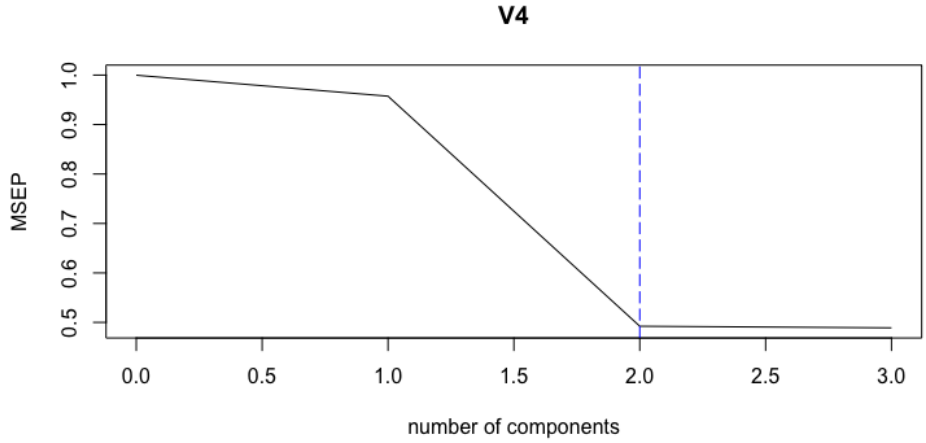
2) Κύρια Παλινδρόμηση Συνιστωσών, με επιλογή του πλήθους συνιστωσών: Στην τεχνική αυτή εφαρμόζεται η τεχνική της ανάλυσης κυρίων συνιστωσών αλλά επιλέγονται οι συνιστώσες οι οποίες θα χρησιμοποιηθούν στην δημιουργία του μοντέλου. Το ποσοστό των ανεξάρτητων μεταβλητών που λαμβάνεται υπόψιν του μοντέλου με την επιλογή x πλήθους συνιστωσών, και το ποσοστό των περιπτώσεων των εξαρτημένων μεταβλητών που επεξηγεί, παρατίθεται στον παρακάτω πίνακα 17:

Πίνακας 17 Συνιστώσες Παλινδρόμησης Κυρίων Συνιστωσών - Skin Segmentation Dataset

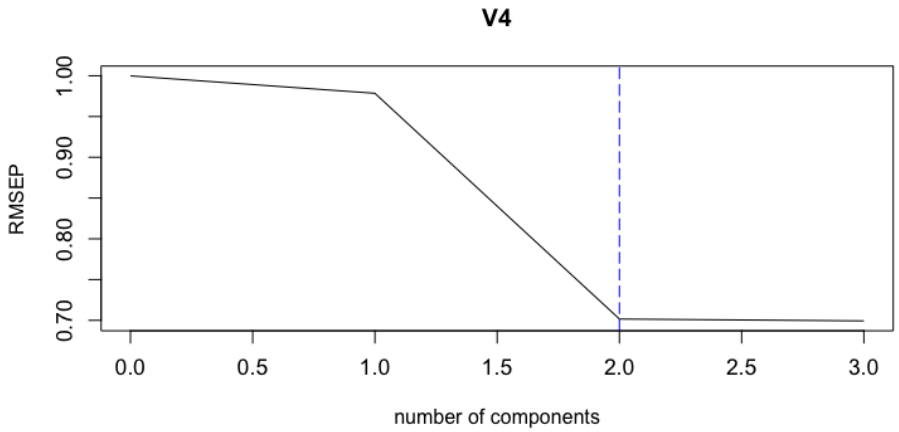
Αριθμός Συνιστωσών	Συνολικό ποσοστό πληροφορίας εξαρτημένων μεταβλητών	Συνολικό ποσοστό πληροφορίας ανεξάρτητων μεταβλητών
1	78.40	4.246
2	96.08	50.77
3	100	51.10

Παρατηρείται ότι με επιλογή δύο συνιστώσες λαμβάνεται σχεδόν το ίδιο ποσοστό πληροφορίας της εξαρτημένης μεταβλητής όπως και με επιλογή τριών συνιστωσών, 50.77%~51.10%. Η πρόσθεση μίας ακόμα συνιστώσας δεν συντελεί στην αύξηση της επεξήγησης πολύ μεγαλύτερου ποσοστού της εξαρτημένης μεταβλητής, αν και λαμβάνεται υπόψιν το 100% των ανεξάρτητων μεταβλητών. Είναι συνεπώς συνετό να κατασκευαστεί μοντέλο με δύο συνιστώσες που θα είναι ικανό να προβλέψει σχεδόν το ίδιο ποσοστό εξαρτημένων μεταβλητών, όσο και αυτό με τρείς.

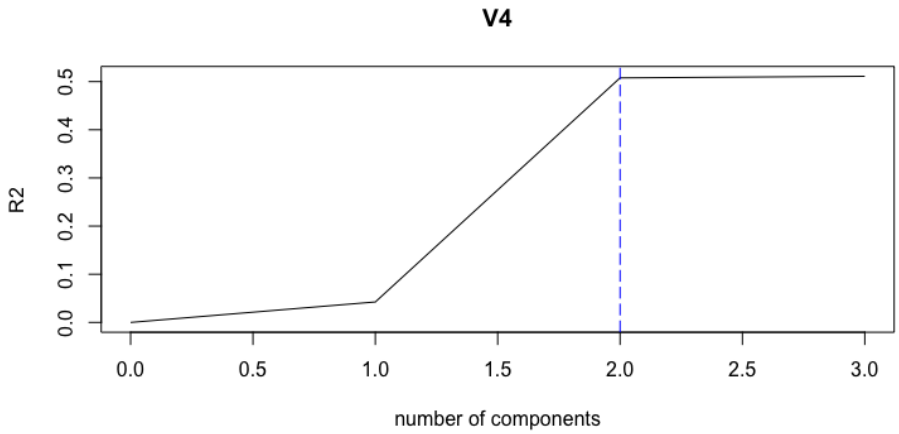
Η παρουσίαση των γραφημάτων της σχέσης του αριθμού των συνιστωσών που χρησιμοποιούνται και των κριτηρίων μέσου τετραγωνικού σφάλματος, (Γράφημα 19), ρίζας μέσου τετραγωνικού σφάλματος, (Γράφημα 20), και R^2, (Γράφημα 21), συνιστούν την ίδια επιλογή.



Γράφημα 19 MSEP - Αριθμός Συνιστωσών PCR - Skin Segmentation Dataset



Γράφημα 20 RMSEP - Αριθμός Συνιστωσών PCR - Skin Segmentation Dataset



Γράφημα 21 R^2 - Αριθμός Συνιστωσών PCR - Skin Segmentation Dataset

Η επιλογή λοιπόν δύο συνιστωσών φαίνεται να είναι αρκετή για την μέγιστη δυνατή προβλεπτική ικανότητα με ελάχιστη χρήση κυρίων συνιστωσών, διατηρώντας το επίπεδο των κριτηρίων μέσου τετραγωνικού σφάλματος, ρίζας μέσου τετραγωνικού σφάλματος και R^2. Τα αποτελέσματα της εφαρμογής του μοντέλου με την χρήση δύο κύριων συνιστωσών φαίνεται στον Πίνακα 18, όπως και αυτών του πλήρους μοντέλου.

Πίνακας 18 Παλινδρομήσεις Κυρίων Συνιστωσών - Skin Segmentation Dataset

PCR	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of PCs
2 PCs	0.5110465	0.5110385	0.7233520	0.8505010	0.6191763	2
All PCs	0.5135587	0.5135428	0.6445043	0.8028103	0.5662549	3

Παρατηρείται η διατήρηση ικανότητας αντιπροσώπευσης της διακύμανσης της εξαρτημένης μεταβλητής με χρήση δύο κυρίων συνιστωσών. Η συμπεριφορά του αλγορίθμου πρόβλεψης εξηγείται από την υψηλή συσχέτιση των αρχικών ανεξάρτητων μεταβλητών. Καθώς οι αρχικές ανεξάρτητες μεταβλητές παρουσίαζαν υψηλές συσχετίσεις μεταξύ τους, η αναγωγή σε κύριες συνιστώσες έχει ως αποτέλεσμα μικρός αριθμός των συνιστωσών, να περιέχει υψηλό ποσοστό πληροφορίας τους. Αποτέλεσμα αυτού, η επιλογή λιγότερων κυρίων συνιστωσών να οδηγήσει σε απλούστερο μοντέλο πρόβλεψης από το μοντέλο με χρήση όλων των συνιστωσών, και με σχεδόν την ίδια απόδοση.

4.3.3 Μερικών Ελαχίστων Τετραγώνων στο “Skin Segmentation Dataset”

Η μέθοδος μερικής ελαχίστων τετραγώνων ακολουθεί την ίδια λογική με την μέθοδο παλινδρόμησης κυρίων συνιστωσών, με την διαφορά ότι πραγματοποιείται η γραμμική παλινδρόμηση των προβολών, των ανεξάρτητων και των εξαρτημένων μεταβλητών σε νέο σύστημα διαστάσεων. Κατασκευάζονται δύο μοντέλα της μερικής ελαχίστων τετραγώνων μεθόδου, με την χρήση όλων των συνιστωσών, και με επιλογή του πλήθους συνιστωσών.

1) Πλήρης τεχνική Μερικών Ελαχίστων Τετραγώνων: Στην τεχνική μερικών ελαχίστων τετραγώνων με χρήση όλων των συνιστωσών, δημιουργούνται 3 συνιστώσες, από τις 3 ανεξάρτητες αρχικές μεταβλητές, για την κατασκευή μοντέλου πρόβλεψης, μέσω της μεγιστοποίησης της συνδιακύμανσης μεταξύ των εξαρτημένων και ανεξάρτητων μεταβλητών.

Τα αποτελέσματα του αλγορίθμου παρατίθενται στον Πίνακα 19, παρακάτω:

Πίνακας 19 Μερικώς Ελαχίστων Τετραγώνων - Skin Segmentation Dataset

PLS	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of PCs
All PCs	0.5135587	0.5135428	0.5478618	0.7401769	0.5348800	3

Παρατηρείται η μέτρια επεξήγηση της διακύμανσης των εξαρτημένων μεταβλητών του πακέτου δεδομένων αξιολόγησης από το μοντέλο με τιμές R-squared, και adjusted R-squared 0.5135587 και 0.5135428 αντίστοιχα. Επίσης οι τιμές των σφαλμάτων είναι μεγάλες παρουσιάζοντας μια έλλειψη ακρίβειας στην ικανότητα πρόβλεψης.

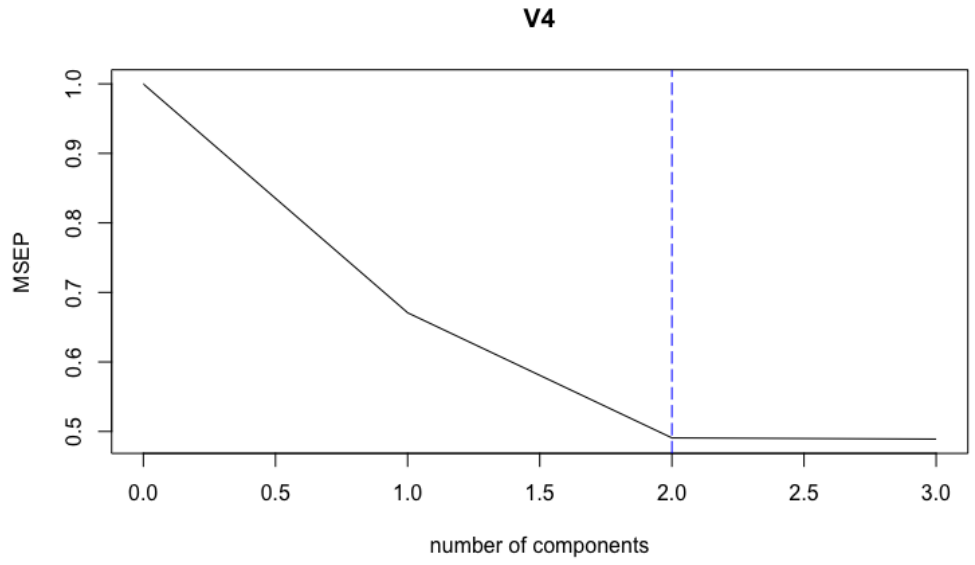
2) Τεχνική Μερικών Ελαχίστων Τετραγώνων με επιλογή συνιστωσών: Στην τεχνική αυτή εφαρμόζεται η τεχνική της ανάλυσης μερικών ελαχίστων τετραγώνων με επιλογή των συνιστωσών που θα χρησιμοποιηθούν στην δημιουργία του μοντέλου. Το ποσοστό των ανεξάρτητων μεταβλητών που λαμβάνεται υπόψιν του μοντέλου με την επιλογή x πλήθους συνιστωσών, και το ποσοστό των περιπτώσεων των εξαρτημένων μεταβλητών που επεξηγεί, παρατίθεται στον παρακάτω Πίνακα 20.

Πίνακας 20 Συνιστώσες Μερικώς Ελαχίστων Τετραγώνων - Skin Segmentation Dataset

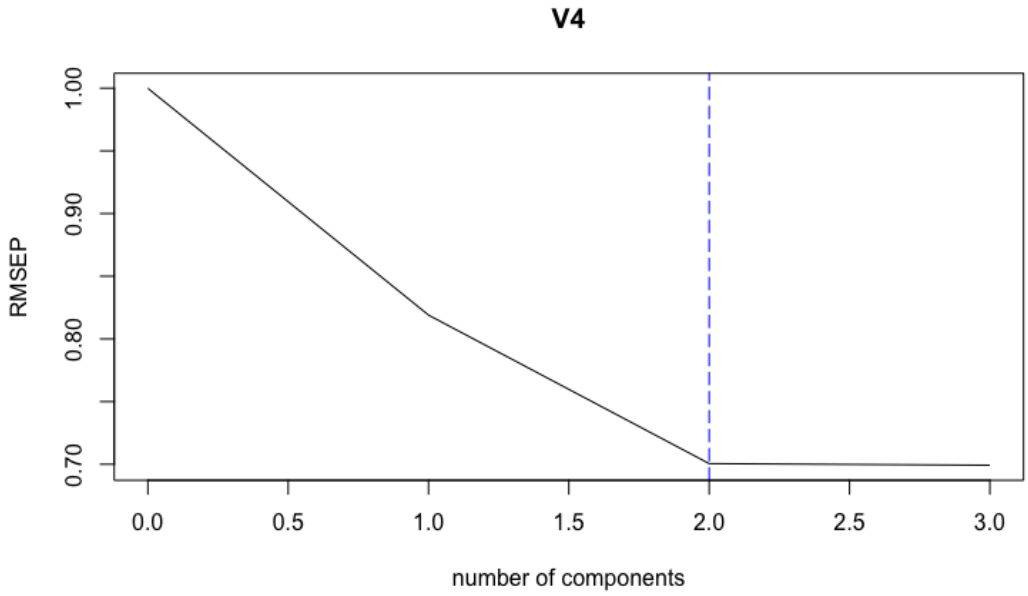
Αριθμός Συνιστωσών	Συνολικό ποσοστό πληροφορίας εξαρτημένων μεταβλητών	Συνολικό ποσοστό πληροφορίας ανεξάρτητων μεταβλητών
1	56.68	32.91
2	96.08	50.92
3	100.0	51.10

Παρατηρείται ότι με επιλογή δύο συνιστώσες λαμβάνεται σχεδόν το ίδιο ποσοστό πληροφορίας της εξαρτημένης μεταβλητής όπως και με επιλογή τριών συνιστωσών, 50.92%~51.10%. Η πρόσθεση μίας ακόμα συνιστώσας δεν συντελεί στην αύξηση της επεξήγησης πολύ μεγαλύτερου ποσοστού της εξαρτημένης μεταβλητής, αν και λαμβάνεται υπόψιν το 100% των ανεξάρτητων μεταβλητών. Είναι συνεπώς συνετό να κατασκευαστεί μοντέλο με δύο συνιστώσες που θα είναι ικανό να προβλέψει σχεδόν το ίδιο ποσοστό εξαρτημένων μεταβλητών, όσο και αυτό με τρείς.

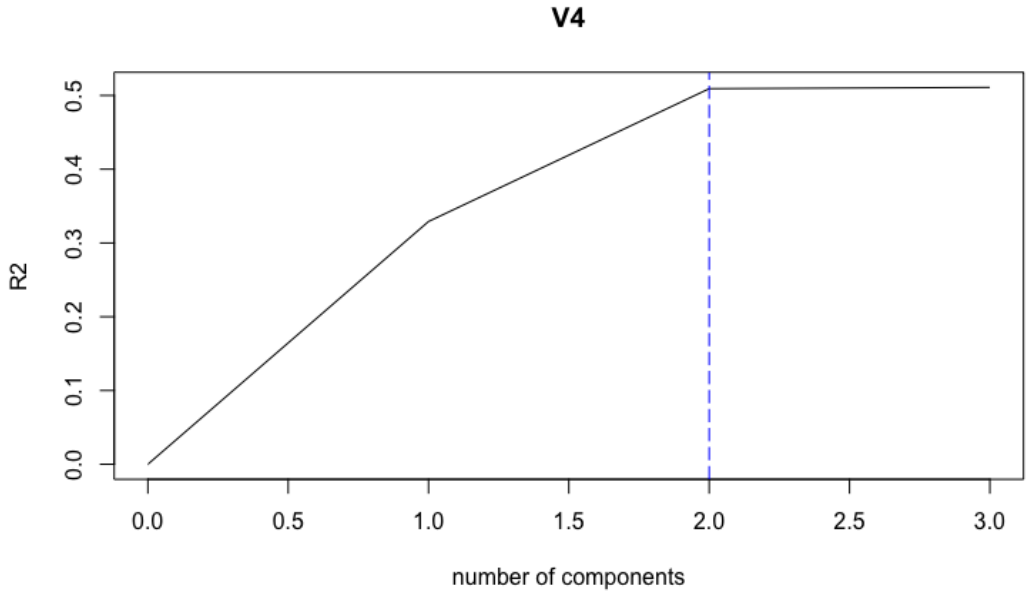
Η παρουσίαση των γραφημάτων της σχέσης του αριθμού των συνιστωσών που χρησιμοποιούνται και των κριτηρίων μέσου τετραγωνικού σφάλματος, ρίζας μέσου τετραγωνικού σφάλματος και R^2, συνιστούν την ίδια επιλογή.



Γράφημα 22 MSEP - Αριθμός Συνιστωσών PLS - Skin Segmentation Dataset



Γράφημα 23 RMSEP - Αριθμός Συνιστωσών PLS - Skin Segmentation Dataset



Γράφημα 24 R^2 - Αριθμός Συνιστωσών PLS - Skin Segmentation Dataset

Η επιλογή λοιπόν δύο συνιστωσών φαίνεται να είναι αρκετή για την μέγιστη δυνατή προβλεπτική ικανότητα με ελάχιστη χρήση κυρίων συνιστωσών, και διατήρηση των κριτηρίων μέσου τετραγωνικού σφάλματος, ρίζας μέσου τετραγωνικού σφάλματος και R^2. Τα αποτελέσματα της εφαρμογής του μοντέλου με την χρήση δύο κύριων συνιστωσών φαίνεται στον παρακάτω Πίνακα 21, όπως και αυτών του πλήρους μοντέλου.

Πίνακας 21 Συνολική Μερικών Ελαχίστων Τετραγώνων - Skin Segmentation Dataset

PLS	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of PCs
2 PCs	0.5123498	0.5123419	0.5783882	0.7605184	0.5721140	2
All PCs	0.5135587	0.5135428	0.5478618	0.7401769	0.5348800	3

Φαίνεται ξεκάθαρη η διατήρηση της προβλεπτικής ικανότητας του μοντέλου με χρήση δύο κυρίων συνιστωσών, σε σχέση με το πλήρες. Πραγματοποιήθηκε απλούστευση του μοντέλου πρόβλεψης, χρησιμοποιώντας λιγότερα δεδομένα, παράγοντας το ίδιο σχεδόν αποτέλεσμα.

4.3.4 Νευρωνικά δίκτυα στο “Skin Segmentation Dataset”

Η αρχιτεκτονική του νευρωνικού δικτύου για το σύνολο δεδομένων “Skin Segmentation Dataset ” είναι αποτέλεσμα συνεχών εκτελέσεων, με διαφορετικό πλήθος επιπέδων και πλήθους νευρώνων να εμπεριέχονται σε αυτά.

Μετά από τις συνεχείς εκτελέσεις του νευρωνικού δικτύου, με ένα ή δύο επίπεδα και διαφορετικό πλήθος νευρώνων σε αυτά, καταλήγοντας σε πέντε νευρώνες στο πρώτο, και δύο στο δεύτερο κρυφό επίπεδο. Τα αποτελέσματα του νευρωνικού αυτού δικτύου με από το σύνολο δεδομένων αξιολόγησης παρατίθενται στον παρακάτω πίνακα:

Πίνακας 22 Νευρωνικά Δίκτυα - Skin Segmentation Dataset

MLP	R^2	R^2_{adj}	MSE	RMSE	MAE
2 layers, with 5 and 2 neurons	0.9948	0.99219	0.00168	0.0411	0.0088

4.4 Ανάλυση του “Wine Quality Dataset”

Το πακέτο δεδομένων “Wine Quality Dataset”, περιέχει έντεκα ανεξάρτητες μεταβλητές, μια εξαρτημένη μεταβλητή, και 4.989 εγγραφές. Επαναλαμβάνεται η εφαρμογή των τεσσάρων τεχνικών προβλεπτικής εξόρυξης δεδομένων.

4.4.1 Πολλαπλή Γραμμική Παλινδρόμηση στο “Wine Quality Dataset”

Στην τεχνική πολλαπλής γραμμικής παλινδρόμησης εφαρμόστηκαν δύο μέθοδοι, πλήρης και stepwise παλινδρόμησης.

2) Πλήρης Πολλαπλή Γραμμική παλινδρόμηση: Στην τεχνική αυτή λαμβάνονται υπόψιν και οι έντεκα ανεξάρτητες μεταβλητές του πακέτου δεδομένων για την κατασκευή μοντέλου πολλαπλής γραμμικής παλινδρόμησης με σκοπό την πρόβλεψη της εξαρτημένης μεταβλητής. Δηλαδή:

$$Y_i = \alpha_0 + \alpha_1 \cdot X_{i1} + \alpha_2 \cdot X_{i2} + \cdots + \alpha_{11} \cdot X_{i11} + \varepsilon_i, i = 1, \dots, 4.989$$

όπου Y_i η εξαρτημένη μεταβλητή και $X_{ij}, i = 1, \dots, 11,\ j = 1, \dots, 4.989$,η ανεξάρτητες.

Αποτέλεσμα της μεθόδου πολλαπλής γραμμικής παλινδρόμησης είναι έντεκα συντελεστές που δίνουν την κλίση του υπερεπιπέδου παλινδρόμησης, και η σταθερά α_0. Επίσης τα σφάλματα πρόβλεψης $\varepsilon_i, i = 1, \dots, 4.989$, δηλαδή η διαφορά της προβλεπόμενης από την πραγματική τιμή της εξαρτημένης μεταβλητής λαμβάνουν μέρος στην συνάρτηση πρόβλεψης.

Παρατηρούμε τις τιμές των R-squared, και adjusted R-squared που είναι σχετικά μικρές, 0.2902 και 0.288 αντίστοιχα. Δηλαδή το μοντέλο επεξηγεί περίπου το 28% της διακύμανσης των ανεξάρτητων μεταβλητών με τα δεδομένα που εκπαιδεύτηκε.

Στην συνέχεια πραγματοποιείται πρόβλεψη για το πακέτο δεδομένων αξιολόγησης με την εισαγωγή των ανεξάρτητων μεταβλητών του στο μοντέλο πρόβλεψης, με αποτέλεσμα την λήψη των προβλέψεων των εξαρτημένων μεταβλητών. Τα αποτελέσματα παρουσιάζονται στον παρακάτω Πίνακα 23:

Πίνακας 23 Πολλαπλή Γραμμική Παλινδρόμηση - Wine Quality Dataset

MLR	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of Predictors
All predictors	0.2871	0.285	0.7405692	0.8605633	0.6614183	11

3) Πολλαπλή γραμμική παλινδρόμηση με τεχνική stepwise: Στην τεχνική αυτή εφαρμόζεται αρχικά το πλήρες μοντέλο πολλαπλής γραμμικής παλινδρόμησης και στην συνέχεια εκτελούνται συνεχείς παλινδρομήσεις του πακέτου δεδομένων εκπαίδευσης έως ότι βρεθεί η βέλτιστη. Η μέθοδος στο πακέτο δεδομένων “Wine Quality Dataset”, καταλήγει σε μοντέλο πρόβλεψης με αριθμό ανεξάρτητων μεταβλητών μικρότερου του πλήρους μοντέλου. Η μέθοδος κατέληξε σε μοντέλο οκτώ ανεξάρτητων στατιστικά σημαντικών μεταβλητών, με τιμές των R-squared, και adjusted R-squared 0.2778 και 0.2764 αντίστοιχα, αντιπροσωπεύοντας το 27% της διακύμανσης των εξαρτημένων δεδομένων του πακέτου εκπαίδευσης.

Πίνακας 24 Πολλαπλές Γραμμικές Παλινδρομήσεις - Wine Quality Dataset

MLR	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of Predictors
Stepwise	0.2869	0.2853	0.7391921	0.8597628	0.6605480	8
All predictors	0.2871	0.285	0.7405692	0.8605633	0.6614183	11

Από τον παραπάνω πίνακα 24, συμπεραίνεται ότι η μέθοδος stepwise πραγματοποίησε μείωση των μεταβλητών που λαμβάνουν μέρος στην παλινδρόμηση για την πρόβλεψη, με σχετική διατήρηση στα επίπεδα των κριτηρίων αξιολόγησης που έχουν τεθεί, πραγματοποιώντας κατασκευή πιο απλοποιημένου μοντέλου με την ίδια ικανότητα πρόβλεψης.

4.4.2 Παλινδρόμηση Κύριων Συνιστωσών στο “Wine Quality Dataset”

Στην τεχνική της παλινδρόμησης κύριων συνιστωσών εφαρμόστηκαν δύο μέθοδοι, η πλήρης όπου λαμβάνονται υπόψιν όλες οι συνιστώσες για την παλινδρόμηση, και μια κατά την οποία επιλέγεται ο αριθμός των συνιστωσών

1) Πλήρης Παλινδρόμηση Κύριων Συνιστωσών: Στην τεχνική αυτή εφαρμόζεται η τεχνική ανάλυσης κυρίων συνιστωσών (Principal Component Analysis, PCA), και όλες οι συνιστώσες οι οποίες παράγονται λαμβάνουν μέρος σε πολλαπλή γραμμική παλινδρόμηση με σκοπό την πρόβλεψη της εξαρτημένης μεταβλητής. Τα δεδομένα έχουν κανονικοποιηθεί, ώστε να έχουν όλες οι μεταβλητές την ευκαιρία για εμφάνιση στο μοντέλο, πριν πραγματοποιηθεί η ανάλυση των κυρίων συνιστωσών.

Με την εκπαίδευση του αλγορίθμου με τα δεδομένα εκπαίδευσης παράγονται 11 κύριες συνιστώσες, κάθε μία από τις οποίες είναι γραμμικός συνδυασμός των αρχικών μεταβλητών. Πάνω σε αυτές τις συνιστώσες πραγματοποιείται η πολλαπλή γραμμική παλινδρόμηση για την πρόβλεψη της εξαρτημένης μεταβλητής. Τα αποτελέσματα της πλήρους παλινδρόμησης κυρίων συνιστωσών παρατίθεται στον παρακάτω πίνακα:

Πίνακας 25 Παλινδρόμηση Κυρίων Συνιστωσών - Wine Quality Dataset

PCR	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of PCs
All PCs	0.2567637	0.2539199	0.8175608	0.9041907	0.6975653	11

Παρατηρείται η σχετικά μικρή αντιπροσώπευση της διακύμανσης της εξαρτημένης μεταβλητής από το μοντέλο με τιμές R-squared, και adjusted R-squared 0.2567637 και 0.2539199 αντίστοιχα. Επίσης οι τιμές των σφαλμάτων είναι τέτοιες που παρουσιάζουν μια όχι τόσο ικανή δυνατότητα πρόβλεψης των δεδομένων αξιολόγησης.

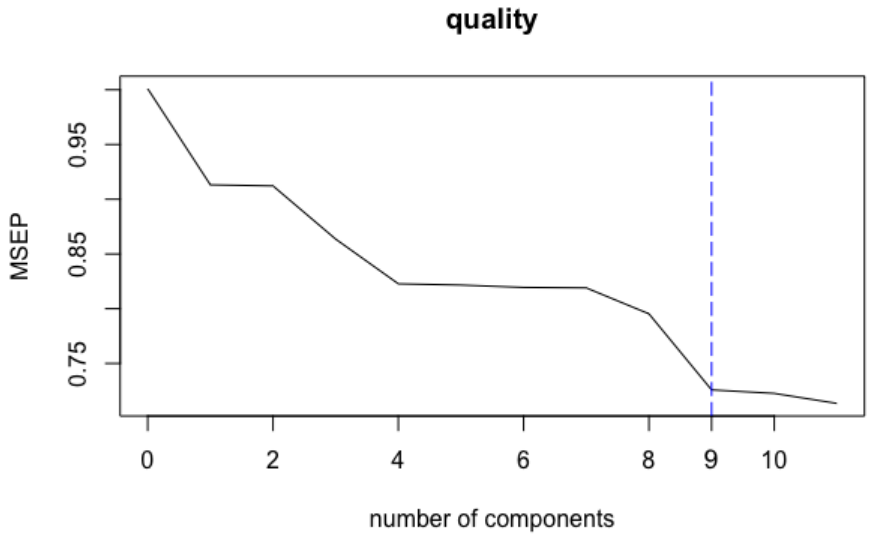
2) Κύρια Παλινδρόμηση Συνιστωσών, με επιλογή του πλήθους συνιστωσών: Στην τεχνική αυτή εφαρμόζεται η τεχνική της ανάλυσης κυρίων συνιστωσών αλλά επιλέγονται οι συνιστώσες οι οποίες θα χρησιμοποιηθούν στην δημιουργία του μοντέλου. Το ποσοστό των ανεξάρτητων μεταβλητών που λαμβάνεται υπόψιν του μοντέλου με την επιλογή x πλήθους συνιστωσών, και το ποσοστό των περιπτώσεων των εξαρτημένων μεταβλητών που επεξηγεί, παρατίθεται στον παρακάτω Πίνακα 26.

Πίνακας 26 Συνιστώσες Παλινδρόμησης Κυρίων Συνιστωσών - Wine Quality Dataset

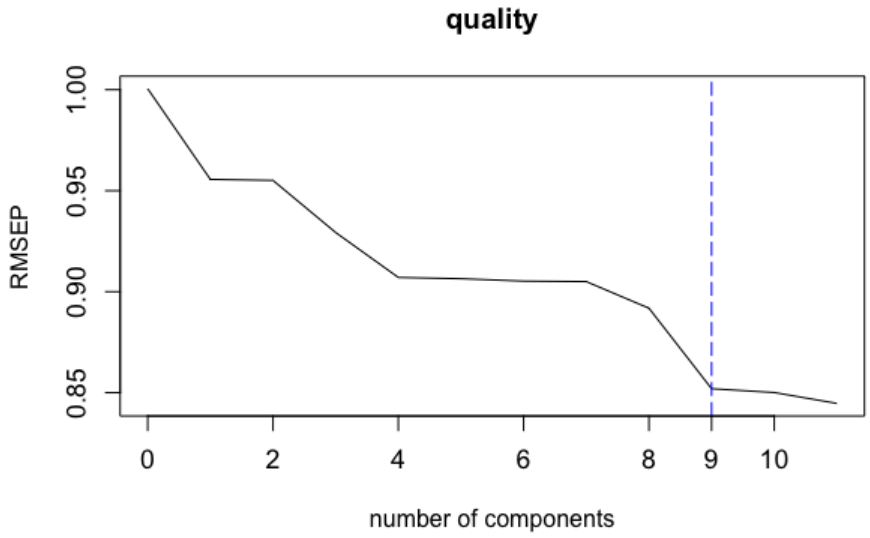
Αριθμός Συνιστωσών	Συνολικό ποσοστό πληροφορίας εξαρτημένων μεταβλητών	Συνολικό ποσοστό πληροφορίας ανεξάρτητων μεταβλητών
1	28.99	8.766
2	43.47	8.867
3	54.73	13.72
4	64.18	17.80
5	73.07	17.92
6	81.48	18.13
7	88.00	18.17
8	93.50	20.54
9	97.21	27.48
10	99.87	27.80
11	100.00	28.71

Παρατηρείται ότι με επιλογή εννιά συνιστώσών λαμβάνεται σχεδόν το ίδιο ποσοστό πληροφορίας της εξαρτημένης μεταβλητής όπως και με επιλογή δέκα συνιστωσών, 27.48%~27.80%, και πως η πρόσθεση μίας ακόμα συνιστώσας το ποσοστό παραμένει σχεδόν αμετάβλητο. Η πρόσθεση μίας ακόμα συνιστώσας δεν συντελεί στην αύξηση της επεξήγησης πολύ μεγαλύτερου ποσοστού της εξαρτημένης μεταβλητής, (από 27.80% με δέκα συνιστώσες, σε 28.71% με έντεκα συνιστώσες), αν και λαμβάνεται υπόψιν το 100% των ανεξάρτητων μεταβλητών. Είναι συνεπώς συνετό να κατασκευαστεί μοντέλο με εννιά συνιστώσες που θα είναι ικανό να προβλέψει σχεδόν το ίδιο ποσοστό εξαρτημένων μεταβλητών, όσο και αυτό με έντεκα.

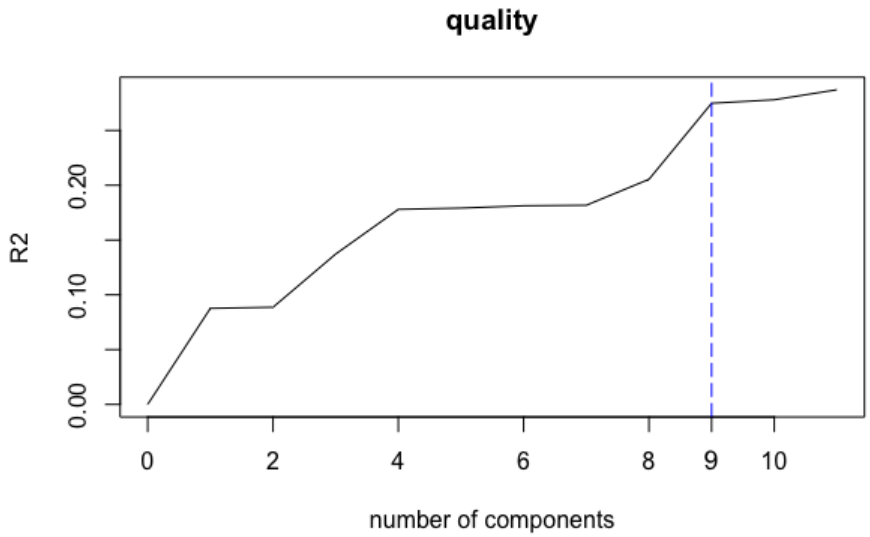
Η παρουσίαση των γραφημάτων της σχέσης του αριθμού των συνιστωσών που χρησιμοποιούνται και των κριτηρίων μέσου τετραγωνικού σφάλματος (γράφημα 25), ρίζας μέσου τετραγωνικού σφάλματος (γράφημα 26), και R^2(γράφημα 27), συνιστούν την ίδια επιλογή. Η μείωση των διαστάσεων αποφέρει μικρή αύξηση μέσου τετραγωνικού σφάλματος και ρίζας μέσου τετραγωνικού σφάλματος, διατηρώντας την τιμή της R^2 σε παρόμοια επίπεδα:



Γράφημα 25 MSEP - Αριθμός Συνιστωσών PCR - Wine Quality Dataset



Γράφημα 26 RMSEP - Αριθμός Συνιστωσών PCR - Wine Quality Dataset



Γράφημα 27 R^2 - Αριθμός Συνιστωσών PCR - Wine Quality Dataset

Η επιλογή λοιπόν πλήθους εννιά συνιστωσών φαίνεται να είναι αρκετή για την μέγιστη δυνατή προβλεπτική ικανότητα με ελάχιστη χρήση κυρίων συνιστωσών, διατηρώντας σχετικά το επίπεδο των κριτηρίων μέσου τετραγωνικού σφάλματος, ρίζας μέσου τετραγωνικού σφάλματος και R^2. Τα αποτελέσματα της εφαρμογής του μοντέλου με την χρήση εννιά κύριων συνιστωσών φαίνεται στον παρακάτω Πίνακα 27, όπως και αυτών του πλήρους μοντέλου:

Πίνακας 27 Παλινδρομήσεις Κυρίων Συνιστωσών - Wine Quality Dataset

PCR	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of PCs
9 PCs	0.2618732	0.2600596	0.7354780	0.8576002	0.6632146	9
All PCs	0.2567637	0.2539199	0.8175608	0.9041907	0.6975653	11

Παρατηρείται η διατήρηση ικανότητας αντιπροσώπευσης της διακύμανσης της εξαρτημένης μεταβλητής με χρήση εννιά κυρίων συνιστωσών. Η συμπεριφορά του αλγορίθμου πρόβλεψης εξηγείται από την συσχέτιση των κάποιων αρχικών ανεξάρτητων μεταβλητών, όπως η υπολειμματική ζάχαρη με την πυκνότητα και την περιεκτικότητα σε αλκοόλ. Αποτέλεσμα αυτού, η επιλογή λιγότερων κυρίων συνιστωσών να οδηγήσει σε απλούστερο μοντέλο πρόβλεψης από το μοντέλο με χρήση όλων των συνιστωσών, και με σχετική βελτίωση της προβλεπτικής απόδοσης.

4.4.3 Μερικών Ελαχίστων Τετραγώνων στο “Wine Quality Dataset”

Η τεχνικής μερικών ελαχίστων τετραγώνων ακολουθεί την ίδια λογική με την μέθοδο παλινδρόμησης κυρίων συνιστωσών, με την διαφορά ότι πραγματοποιείται η γραμμική παλινδρόμηση των προβολών, των ανεξάρτητων και των εξαρτημένων μεταβλητών σε νέο σύστημα διαστάσεων. Κατασκευάζονται δύο μοντέλα της μερικής ελαχίστων τετραγώνων μεθόδου, με την χρήση όλων των συνιστωσών, και με επιλογή του πλήθους συνιστωσών.

1) Πλήρης τεχνική Μερικών Ελαχίστων Τετραγώνων: Στην τεχνική μερικών ελαχίστων τετραγώνων με χρήση όλων των συνιστωσών, δημιουργούνται 11 συνιστώσες, από τις 11 ανεξάρτητες αρχικές μεταβλητές, για την κατασκευή μοντέλου πρόβλεψης, μέσω της μεγιστοποίησης της συνδιακύμανσης μεταξύ των εξαρτημένων και ανεξάρτητων μεταβλητών.

Τα αποτελέσματα του αλγορίθμου παρατίθενται στον παρακάτω Πίνακα 28:

Πίνακας 28 Μερική Ελαχίστων Τετραγώνων - Wine Quality Dataset

PLS	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of PCs
All PCs	0.2567637	0.2539199	0.7448949	0.8630729	0.6661550	11

Παρατηρείται η σχετικά μικρή επεξήγηση της διακύμανσης των εξαρτημένων μεταβλητών του πακέτου δεδομένων αξιολόγησης από το μοντέλο με τιμές R-squared, και adjusted R-squared 0.2567637 και 0.2539199 αντίστοιχα. Επίσης οι τιμές των σφαλμάτων είναι μεγάλες παρουσιάζοντας μια έλλειψη ακρίβειας στην ικανότητα πρόβλεψης.

2) Τεχνική Μερικών Ελαχίστων Τετραγώνων με επιλογή συνιστωσών: Στην τεχνική αυτή εφαρμόζεται η τεχνική της ανάλυσης μερικής ελαχίστων τετραγώνων με επιλογή των συνιστωσών που θα χρησιμοποιηθούν στην

δημιουργία του μοντέλου. Το ποσοστό των ανεξάρτητων μεταβλητών που λαμβάνεται υπόψιν του μοντέλου με την επιλογή x πλήθους συνιστωσών, και το ποσοστό των περιπτώσεων των εξαρτημένων μεταβλητών που επεξηγεί, παρατίθεται στον παρακάτω πίνακα 29.

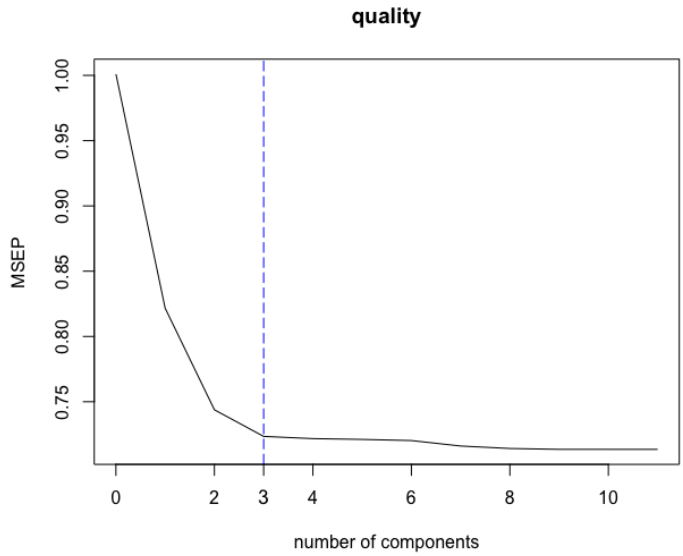
Πίνακας 29 Συνιστώσες Μερικών Ελαχίστων Τετραγώνων - Wine Quality Dataset

Αριθμός Συνιστωσών	Συνολικό ποσοστό πληροφορίας εξαρτημένων μεταβλητών	Συνολικό ποσοστό πληροφορίας ανεξάρτητων μεταβλητών
1	26.30	17.91
2	38.04	25.69
3	44.48	27.72
4	54.40	27.89
5	64.00	27.95
6	72.16	28.04
7	74.58	28.46
8	79.30	28.65
9	28.65	28.71
10	91.54	28.71
11	100.00	28.71

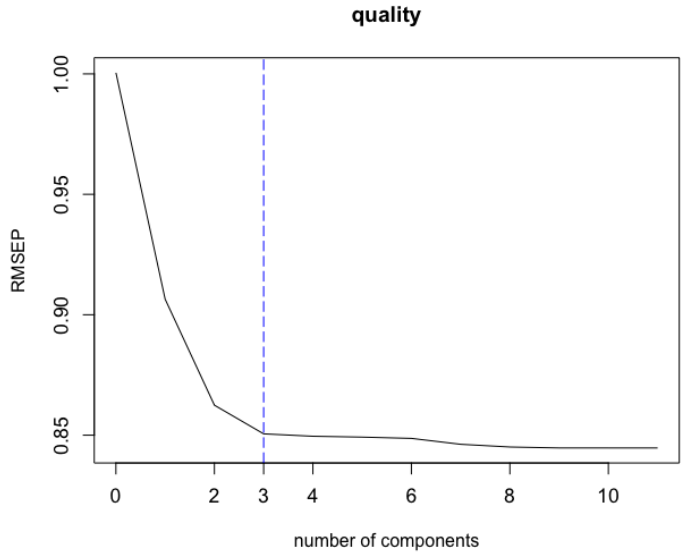
Παρατηρείται ότι με επιλογή πλήθους τριών συνιστωσών λαμβάνεται σχεδόν το ίδιο ποσοστό πληροφορίας της εξαρτημένης μεταβλητής όπως και με επιλογή πέντε συνιστωσών, 27.72%~27.95%. Η πρόσθεση επιπλέων συνιστωσών δεν συντελεί στην αύξηση της επεξήγησης πολύ μεγαλύτερου ποσοστού της εξαρτημένης μεταβλητής, αν και λαμβάνεται υπόψιν μεγαλύτερο ποσοστό των ανεξάρτητων μεταβλητών. Είναι συνεπώς συνετό να κατασκευαστεί μοντέλο με τρείς συνιστώσες που θα είναι ικανό να προβλέψει σχεδόν το ίδιο ποσοστό εξαρτημένων μεταβλητών, όσο και αυτό με έντεκα.

Τα γραφήματα της σχέσης του αριθμού των συνιστωσών που χρησιμοποιούνται και των κριτηρίων μέσου τετραγωνικού σφάλματος, ρίζας μέσου τετραγωνικού σφάλματος και R^2, συνιστούν την ίδια επιλογή. Με τρείς συνιστώσες θα πραγματοποιηθεί μικρή αύξηση του μέσου τετραγωνικού σφάλματος, και της ρίζας

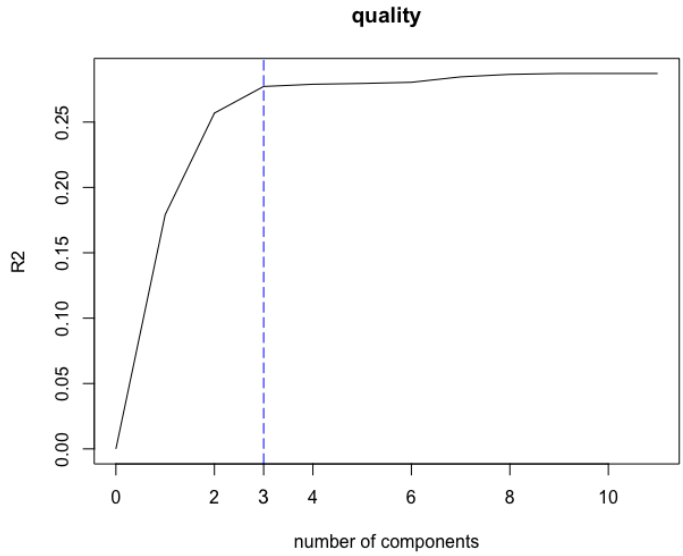
μέσου τετραγωνικού σφάλματος, διατηρώντας την R^2 σε παρόμοια επίπεδα όπως με την χρήση όλων των συνιστωσών. Παρατίθενται τα γραφήματα:



Γράφημα 28 MSEP - Αριθμός Συνιστωσών PLS - Wine Quality Dataset



Γράφημα 29 RMSEP - Αριθμός Συνιστωσών PLS - Wine Quality Dataset



Γράφημα 30 R^2 - Αριθμός Συνιστωσών PLS - Wine Quality Dataset

Η επιλογή λοιπόν τριών συνιστωσών φαίνεται να είναι αρκετή για την μέγιστη δυνατή προβλεπτική ικανότητα με ελάχιστη χρήση κυρίων συνιστωσών, και διατήρηση των κριτηρίων μέσου τετραγωνικού σφάλματος, ρίζας μέσου τετραγωνικού σφάλματος και R^2. Τα αποτελέσματα της εφαρμογής του μοντέλου με την χρήση δύο κύριων συνιστωσών φαίνεται στον παρακάτω πίνακα, όπως και αυτών του πλήρους μοντέλου:

Πίνακας 30 Συνολική Μερικών Ελαχίστων Τετραγώνων - Wine Quality Dataset

PLS	R^2	R^2_{adj}	MSE	RMSE	MAE	Amount of PCs
3 PCs	0.2606322	0.2600277	0.7367145	0.8583208	0.6661071	3
All PCs	0.2567637	0.2539199	0.7448949	0.8630729	0.6661550	11

Φαίνεται ξεκάθαρη η διατήρηση της προβλεπτικής ικανότητας του μοντέλου με χρήση τριών κυρίων συνιστωσών, σε σχέση με το πλήρες. Πραγματοποιήθηκε απλοποίηση του μοντέλου πρόβλεψης, χρησιμοποιώντας λιγότερα δεδομένα, παράγοντας το ίδιο σχεδόν αποτέλεσμα.

4.4.4 Νευρωνικά δίκτυα στο “Wine Quality Dataset”

Η αρχιτεκτονική του νευρωνικού δικτύου για το σύνολο δεδομένων “ Wine Quality Dataset” είναι αποτέλεσμα συνεχών εκτελέσεων, με διαφορετικό πλήθος επιπέδων και πλήθους νευρώνων που εμπεριέχονται σε αυτά.

Μετά από τις συνεχείς εκτελέσεις του νευρωνικού δικτύου, με ένα ή δύο επίπεδα και διαφορετικό πλήθος νευρώνων σε αυτά, δεν επιλέγεται νευρωνικό δίκτυο με ένα ή δύο κρυφά επίπεδα. Η επανάληψη των πειραμάτων με πληθώρα αριθμού νευρώνων στο ένα κρυφό επίπεδο έφερε σχεδόν τα ίδια αποτελέσματα, με αυτά των δύο επιπέδων. Τελικά το βέλτιστο δυνατό λαμβάνεται από τρία κρυφά επίπεδα με 7, 9 και 4 νευρώνες αντίστοιχα.

Τα αποτελέσματα του νευρωνικού αυτού δικτύου με από το σύνολο δεδομένων αξιολόγησης παρατίθενται στον παρακάτω πίνακα:

Πίνακας 31 Νευρωνικά Δίκτυα - Wine Quality Dataset

MLP	R^2	R^2_{adj}	MSE	RMSE	MAE
3 hidden layer with 7,9,4 neurons	0.4836	0.315674	0.60824	0.5986	0.4836

5 Σύνοψη αποτελεσμάτων και συμπεράσματα

Στο κεφάλαιο αυτό πραγματοποιείται μία σύνοψη των αποτελεσμάτων των μεθόδων εξόρυξης που πραγματοποιήθηκαν στο τέταρτο κεφάλαιο και αναλύονται τα συμπεράσματα αυτών. Επίσης παρατίθενται και κάποιες συστάσεις για μελλοντική έρευνα.

Η σύγκριση των μεθόδων πραγματοποιείται βάση πέντε κριτηρίων τα οποία παρατίθενται στο τέταρτο κεφάλαιο, R-square (R^2), R-square adjusted (R^2_{adj}), Μέσο Τετραγωνικό Σφάλμα (MSE), Ρίζα Μέσου Τετραγωνικού Σφάλματος (RMSE), Μέσο Απόλυτο Σφάλμα (MAE), και αριθμός μεταβλητών.

Αρχικά, ανά πακέτο δεδομένων, συγκρίνονται οι τεχνικές παλινδρόμησης από τις οποίες επιλέγεται η πιο αποδοτική και στην συνέχεια αυτή συγκρίνεται με την απόδοση του νευρωνικού δικτύου. Η σύγκριση των γραμμικών μεθόδων παλινδρόμησης πραγματοποιείται αρχικά σύμφωνα με την τιμή του μέσου τετραγωνικού σφάλματος (MSE), με την τιμή του μέσου απολύτου σφάλματος (MAE) να χρησιμοποιείται σε περίπτωση ισοβαθμίας, και στην συνέχεια με τον αριθμό των ανεξάρτητων μεταβλητών που χρησιμοποιήθηκαν για την εξαγωγή του αποτελέσματος. Στην συνέχεια η καλύτερη σε απόδοση γραμμική μέθοδος συγκρίνεται με την απόδοση του νευρωνικού δικτύου σύμφωνα με την τιμή του μέσου τετραγωνικού σφάλματος.

Τα συμπεράσματα κατηγοριοποιούνται στην συνέχεια βάση των χαρακτηριστικών των δεδομένων και της μεθόδου που είχε την καλύτερη απόδοση σε αυτά.

5.1 Σύνοψη αποτελεσμάτων

Στην ενότητα αυτή πραγματοποιείται μια σύνοψη των αποτελεσμάτων των τεχνικών εξόρυξης δεδομένων που εφαρμόστηκαν ανά πακέτο δεδομένων, και αναγνωρίζεται η πιο αποδοτική τεχνική για το κάθε ένα. Στην συνέχεια αναφέρονται τα χαρακτηριστικά των συνόλων δεδομένων και πραγματοποιείται αντιστοίχιση χαρακτηριστικού των δεδομένων και τεχνικής που απέδωσε καλυτέρα σε αυτό.

5.1.1 Σύνοψη Αποτελεσμάτων “Divorce Predictors Dataset”

Για το πακέτο δεδομένων “Divorce Predictors Dataset”, παρατίθεται η απόδοση των τεχνικών εξόρυξης στον παρακάτω πίνακα. Παρατηρείται η μικρότερη τιμή του μέσου τετραγωνικού σφάλματος από την τεχνική μερικών ελαχίστων τετραγώνων, με την επιλογή δύο συνιστωσών. Η παλινδρόμηση κυρίων συνιστωσών παρουσιάζει παρόμοια απόδοση με καλύτερο δείκτη μέσου απόλυτου σφάλματος, αλλά με χρήση τεσσάρων συνιστωσών. Τα νευρωνικά δίκτυα είχαν δείκτη μέσου τετραγωνικού σφάλματος πολύ κοντά με την τεχνική μερικών ελαχίστων τετραγώνων. Συνεπώς καλύτερη απόδοση για το σύνολο δεδομένων “Divorce Predictors Dataset”, έχει η τεχνική μερικών ελαχίστων τετραγώνων.

Πίνακας 32 Σύνοψη Απόδοσης Τεχνικών - Divorce Predictors Dataset

Τεχνική	Μέθοδος	R^2	R^2_{adj}	MSE	MAE	Μεταβλητές
MLR	Stepwise	0.9772	0.9701	0.135856	0.244435	32
	Full	0.9793	0.9638	0.142946	0.253497	54
PCR	4 PCs	0.931773	0.929536	0.067797	0.136398	4
	All PCs	0.854067	0.735236	0.080356	0.177075	54
PLS	2 PCs	0.933753	0.932684	0.065830	0.146175	2
	All PCs	0.854067	0.735236	0.131939	0.233058	54
NN	MLP	0.902924	0.8526326	0.0964654	0.1425251	

5.1.2 Σύνοψη Αποτελεσμάτων “Skin Segmentation Dataset”

Για το πακέτο δεδομένων “Skin Segmentation Dataset”, παρατίθεται η απόδοση των μεθόδων στον παρακάτω πίνακα. Εύκολα συμπεραίνεται η χαμηλή απόδοση των γραμμικών τεχνικών. Παρατηρείται ότι η μικρότερη τιμή του μέσου τετραγωνικού σφάλματος από τις γραμμικές τεχνικές κατέχεται από την τεχνική πολλαπλής παλινδρόμησης, με μέθοδο stepwise, καθώς και οι δύο μέθοδοι (πλήρης και stepwise), έχουν τα ίδια αποτελέσματα. Η απόδοση όμως των νευρωνικών δικτύων ξεπέρασε κατά πολύ των απόδοση των γραμμικών τεχνικών, με τιμή μέσου τετραγωνικού σφάλματος και μέσου απολύτου σφάλματος 0.00168 και 0.0088 αντίστοιχα. Συνεπώς καλύτερη απόδοση για το σύνολο δεδομένων “Skin Segmentation Dataset”, έχει η τεχνική των νευρωνικών δικτύων.

Πίνακας 33 Σύνοψη Απόδοσης Τεχνικών - Skin Segmentation Dataset

Τεχνική	Μέθοδος	R^2	R^2_{adj}	MSE	MAE	Μεταβλητές
MLR	Stepwise	0.511	0.511	0.4868090	0.4604120	3
	Full	0.511	0.511	0.4868090	0.4604120	3
PCR	4 PCs	0.5110465	0.511038	0.723352	0.619176	2
	All PCs	0.513558	0.513542	0.644504	0.566254	3
PLS	2 PCs	0.512349	0.512341	0.578388	0.572114	2
	All PCs	0.513558	0.513542	0.547861	0.534880	3
NN	MLP	0.9948	0.99219	0.00168	0.0088	

5.1.3 Σύνοψη Αποτελεσμάτων “Wine Quality Dataset”

Για το πακέτο δεδομένων “Wine Quality Dataset”, παρατίθεται η απόδοση των μεθόδων στον παρακάτω πίνακα. Εύκολα συμπεραίνεται η χαμηλή απόδοση των γραμμικών τεχνικών. Παρατηρείται ότι η μικρότερη τιμή του μέσου τετραγωνικού σφάλματος από τις γραμμικές τεχνικές κατέχεται από την τεχνική πολλαπλής παλινδρόμησης, με μέθοδο stepwise. Η απόδοση όμως των νευρωνικών δικτύων ξεπέρασε κατά πολύ των απόδοση των γραμμικών τεχνικών, με τιμή μέσου τετραγωνικού σφάλματος και μέσου απολύτου σφάλματος 0.60824 και 0.5986 αντίστοιχα. Συνεπώς καλύτερη απόδοση για το σύνολο δεδομένων “Wine Quality Dataset”, έχει η τεχνική των νευρωνικών δικτύων.

Πίνακας 34 Σύνοψη Απόδοσης Τεχνικών - Wine Quality Dataset

Τεχνική	Μέθοδος	R^2	R^2_{adj}	MSE	MAE	Μεταβλητές
MLR	Stepwise	0.2869	0.2853	0.7391921	0.6605480	8
	All predictors	0.2871	0.285	0.7405692	0.6614183	11
PCR	4 PCs	0.261873	0.260059	0.735478	0.663214	9
	All PCs	0.256763	0.253919	0.817560	0.697565	11
PLS	2 PCs	0.260632	0.260027	0.736714	0.666107	3
	All PCs	0.256763	0.253919	0.744894	0.666155	11
NN	MLP	0.4836	0.315674	0.60824	0.5986	

5.2 Συμπεράσματα

Στο πλαίσιο της παρούσας διπλωματικής πραγματοποιήθηκαν οι εφαρμογές τεσσάρων τεχνικών εξόρυξης σε τρία διαφορετικά σύνολα δεδομένων. Τα σύνολα δεδομένων παρουσίαζαν διαφορετικά χαρακτηριστικά μεταξύ τους, όπως το σύνολο των ανεξάρτητων μεταβλητών και η συσχέτιση μεταξύ τους, η συσχέτιση των ανεξάρτητων μεταβλητών με την εξαρτημένη, η διασπορά τους, και το πλήθος των εγγραφών. Η μόνη ομοιότητα που επιλέχθηκε να παρουσιάζουν αποτελεί η εξαρτημένη μεταβλητή προς πρόβλεψη να είναι μία.

Από την ανάλυση των συνόλων δεδομένων στο τρίτο κεφάλαιο και των αποτελεσμάτων από τις μεθόδους εξόρυξης του τετάρτου κεφαλαίου:

- στο σύνολο δεδομένων “Divorce Predictors Dataset” παρατηρούνται οι πολύ υψηλές συσχετίσεις του παρουσιάζουν οι ανεξάρτητες μεταβλητές με την εξαρτημένη, όπως και οι ανεξάρτητες μεταβλητές μεταξύ τους. Επίσης παρατηρείται μικρή ύπαρξη ακραίων τιμών. Το μέγεθος των δεδομένων είναι μικρό με 170 εγγραφές αλλά με μεγάλο πλήθος, (54), ανεξάρτητων μεταβλητών. Οι τεχνικές παλινδρόμησης κυρίων συνιστωσών και μερικών ελαχίστων τετραγώνων απέδωσαν σε καλύτερο βαθμό από τις άλλες τεχνικές. Τα νευρωνικά δίκτυα απέδωσαν σε συγκρίσιμο βαθμό, χρησιμοποιώντας πολλαπλάσιο χρόνο όμως για την εξαγωγή του αποτελέσματος. Η πολλαπλή γραμμική παλινδρόμηση απέδωσε ικανοποιητικά αλλά με σχεδόν διπλάσιο δείκτη σφάλματος σε σχέση με τις άλλες μεθόδους.

- στο σύνολο δεδομένων “Skin Segmentation Dataset”, παρατηρείται η χαμηλή συσχέτιση της πλειοψηφίας των ανεξάρτητων μεταβλητών με την εξαρτημένη και παράλληλα η υψηλή συσχέτιση των ανεξάρτητων μεταβλητών μεταξύ τους. Ελάχιστη ύπαρξη ακραίων τιμών. Μεγάλο πλήθος εγγραφών, (245.057), με λίγες, (3), ανεξάρτητες μεταβλητές. Τα νευρωνικά δίκτυα απέδωσαν πολύ καλύτερα από τις γραμμικές τεχνικές, οι οποίες απέδωσαν περίπου στα ίδια επίπεδα.

- στο σύνολο δεδομένων “Wine Quality Dataset” παρατηρείται η χαμηλή συσχέτιση της πλειοψηφίας των ανεξάρτητων μεταβλητών με την εξαρτημένη, και επίσης χαμηλή συσχέτιση της πλειοψηφίας των ανεξάρτητων μεταβλητών μεταξύ τους. Κάποιες από τις ανεξάρτητες μεταβλητές δεν παρουσιάζουν συσχέτιση με τις άλλες, και με την εξαρτημένη μεταβλητή παράλληλα. Αρκετές οι ακραίες τιμές. Μέτριο πλήθος εγγραφών, (4.898), με 12 ανεξάρτητες μεταβλητές. Τα νευρωνικά δίκτυα απέδωσαν καλύτερα από τις γραμμικές τεχνικές, όμως η απόδοσή τους δεν θεωρείται υψηλή, καθώς παραμένει υψηλό το μέγεθος του μέσου τετραγωνικού, αλλά και μέσου απόλυτου, σφάλματος.

Καθώς τα δεδομένα επιλέχθηκαν για την μη ειδικευμένη φύση τους, και λόγω της έμφυτης διαφορετικότητάς τους, τα παραπάνω αποτελέσματα μπορούν να γενικευθούν αποτελώντας έναν οδηγό μιας πρώτης επιλογής τεχνικής εξόρυξης σύμφωνα με τα χαρακτηριστικά των δεδομένων.

Η πολλαπλή γραμμική παλινδρόμηση αποδίδει ικανοποιητικά στα δεδομένα όπου το πλήθος των εγγραφών είναι μικρό, και το πλήθος των ανεξάρτητων μεταβλητών μεγάλο με υψηλό βαθμό συσχέτισης με την εξαρτημένη μεταβλητή. Παράλληλα όμως το σφάλμα των προβλέψεων επηρεάζεται από τον υψηλό βαθμό συσχέτισης που παρουσιάζουν οι εξαρτημένες μεταβλητές μεταξύ τους. Η μέθοδος stepwise βοηθά να μειωθεί το σφάλμα μη λαμβάνοντας υπόψιν της εξαρτημένες μεταβλητές που δεν προσφέρουν την αύξηση της γνώσης του μοντέλου. Οι ανάγκες για εκπαίδευση της τεχνικής αυτής είναι μικρές, από άποψη χρόνου και υπολογιστικών πόρων.

Η παλινδρόμηση κυρίων συνιστωσών αποδίδει σε καλύτερο βαθμό στα δεδομένα όπου το πλήθος των ανεξάρτητων μεταβλητών είναι μεγάλο με υψηλό βαθμό συσχέτισης μεταξύ τους. Η μείωση των διαστάσεων που μπορεί να πραγματοποιηθεί με την επιλογή του αριθμό των συνιστωσών που λαμβάνουν μέρος στο μοντέλο, βοήθα στην μείωση του σφάλματος σε σχέση με την πολλαπλή γραμμική παλινδρόμηση που η υψηλή συσχέτιση των ανεξάρτητων μεταβλητών την επηρεάζει. Σε υψηλό αριθμό εγγραφών και μικρό πλήθος ανεξάρτητων μεταβλητών έχει ίδια ή και χειρότερη απόδοση σε σχέση με την πολλαπλή γραμμική παλινδρόμηση. Οι ανάγκες για εκπαίδευση της τεχνικής αυτής είναι σχετικά μικρές, από άποψη χρόνου και υπολογιστικών πόρων.

Η τεχνική μερικώς ελαχίστων τετραγώνων αποδίδει σχεδόν στο ίδιο βαθμό με την παλινδρόμηση κυρίων συνιστωσών, αλλά η ανάλυση σε συνιστώσες των ανεξάρτητων αλλά και των εξαρτημένων μεταβλητών παρέχει ένα πλεονέκτημα. Με λιγότερες συνιστώσες παρέχονται τα ίδια ή και καλύτερα προβλεπτικά αποτελέσματα. Η απόδοσή της αναδεικνύεται από την επιλογή του αριθμού των συνιστωσών που λαμβάνουν μέρος στο μοντέλο πρόβλεψης, αποφεύγοντας έτσι την συσχέτιση που μπορεί να παρουσιάζουν οι ανεξάρτητες μεταβλητές μεταξύ τους. Σε υψηλό αριθμό εγγραφών και μικρό πλήθος ανεξάρτητων μεταβλητών έχει ίδια ή και χειρότερη απόδοση σε σχέση με την πολλαπλή γραμμική ή των κυρίων συνιστωσών παλινδρόμηση. Οι ανάγκες για εκπαίδευση της τεχνικής αυτής είναι σχετικά μικρές, από άποψη χρόνου και υπολογιστικών πόρων.

Τα νευρωνικά δίκτυα αποδίδουν σε υψηλό βαθμό στην πλειοψηφία των περιπτώσεων. Η απόδοσή τους έναντι των γραμμικών μεθόδων αναδεικνύεται σε δεδομένα με υψηλό πλήθος εγγραφών, ακόμα και όταν οι ανεξάρτητες δεν δείχνουν υψηλό βαθμό συσχέτισης με την εξαρτημένη μεταβλητή. Δεν επηρεάζονται από τις ακραίες τιμές του συνόλου. Αν και το αποτέλεσμα πολλές φορές είναι πολλαπλάσια καλύτερο από αυτό των γραμμικών μεθόδων, ο χρόνος εκπαίδευσης που απαιτείται, είναι ανασταλτικός παράγοντας.

5.3 Προτάσεις για μελλοντική έρευνα

Η εργασία αυτή πραγματοποιεί μια προσπάθεια στην απάντηση του ερωτήματος της επιλογής κατάλληλης τεχνικής εξόρυξης δεδομένων. Η αύξηση των τεχνικών μη γραμμικού χαρακτήρα δίνει δυνατότητες για καλύτερα αποτελέσματα σε προβλήματα όπου είχαν βρεθεί οι βέλτιστες λύσεις. Τύποι νευρωνικών δικτύων και υβριδικά μοντέλα προβλεπτικής εξόρυξης δεδομένων κατασκευάζονται συνεχώς παρουσιάζοντας την ανάγκη για την αξιολόγηση τους σε σύνολα δεδομένων τα οποία με γενικά χαρακτηριστικά θα δώσουν το ερέθισμα για την πιθανώς βέλτιστη επιλογή της μεθόδου που είναι κάθε φορά αναγκαία.

Η ανάπτυξη λογισμικών που θα κάνει πλαίσια εφαρμογής πιο προσιτά στο ευρύ κοινό της ανάλυσης και εξόρυξης δεδομένων, είναι ακόμα σε έλλειψη. Η προσπάθεια προσδιορισμού καλύτερης προβλεπτικής μεθόδου σε περιπτώσεις όπου οι εξαρτημένες μεταβλητές είναι παραπάνω από μία, παραμένουν ακόμα εφαρμογές σε ακαδημαϊκό περιβάλλον, με τις εφαρμογές τους έξω από αυτό το πλαίσιο να είναι λίγες. Η λειτουργικότητα των εργαλείων εφαρμογής τεχνικών εξόρυξης παραμένει μονόπλευρη καθώς δεν υποστηρίζονται παράλληλες εκτελέσεις τεχνικών που να υποστηρίζουν πολλαπλά αποτελέσματα. Η ανάπτυξη των κβαντικών υπολογιστών με την υψηλή παράλληλη υπολογιστική δυνατότητα ίσως δώσουν την λύση στο πρόβλημα αυτό. Από την πλευρά της πρώτης ύλης, καθώς τα δεδομένα που υπάρχουν στον παγκόσμιο ιστό είναι δυναμικά, μη ολοκληρωμένα και σποραδικά στον χρόνο παραγωγής τους, είναι γεωμετρικά αυξάνουσα πιο αισθητή η ανάγκη για δημιουργία εργαλείων διαχείρισης της πληροφορίας αυτής.

Ενώ οι παραπάνω προβληματισμοί και ανάγκες θωρούν αξεπέραστα εμπόδια στην παρούσα στιγμή, ο ανθρώπινος παράγοντας που επενδύει συνεχώς στον χώρο δίνει την ελπίδα για την γρήγορη επίλυσή τους.

Βιβλιογραφία

1. Agresti, Alan. Categorical Data Analysis. Hoboken, NJ: John Wiley & Sons, 2003.

2. Abdolmaleki, P., M. Yarmohammadi, and M. City. "Comparison of logistic regression and neural network models in predicting the outcome of biopsy in breast cancer from MRI findings." International Journal Of Radiation Research 1, no. 4 (2004).

3. Ahangar, Reza G., Mahmood Yahyazadehfar, and Hassan Pournaghshaband. "The Comparison of Methods Artificial Neural Network with Linear Regression Using Specific Variables for Prediction Stock Price in Tehran Stock Exchange." International Journal of Computer Science and Information Security 7, no. 2 (2010).

4. Ainslie, A., and X. Dreze. "Data mining and the choice between classical models and neural networks." Decisions Marketing, 1996.

5. Anwar, Saiful, and Kenji Watanabe. "Predicting Future Depositor`s Rate of Return Applying Neural Network: A Case-study of Indonesian Islamic Bank." International Journal of Economics and Finance 2, no. 3 (2010).

6. Berry, Michael J., and Gordon S. Linoff. Data Mining Techniques: For Marketing, Sales, and Customer Support. Hoboken, NJ: Wiley, 1997.

7. Berson, Alex, Stephen Smith, and Kurt Thearling. Building Data Mining Applications for CRM. New York, NJ: McGraw-Hill Companies, 2000.

8. Bishop, Christopher M. Neural Networks for Pattern Recognition. NJ: Oxford University Press, 1995.

9. Cao, Qing, Karyl B. Leggio, and Marc J. Schniederjans. "A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market." Computers & Operations Research 32, no. 10 (2005).

10. Chapman, Pete. CRISP-DM 1.0: Step-by-step Data Mining Guide. NJ2000.

11. Chye Koh, Hian, and Chan Kee Low. "Going concern prediction using data mining techniques." Managerial Auditing Journal 19, no. 3 (2004).

12. "Classifying Spatial Patterns." Neural Networks for Pattern Recognition, 1993. doi:10.7551/mitpress/4923.003.0005.

13. Cohen, Jacob. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3rd Edition. NJ2003.

14. Crawley, Michael J. The R Book. NJ: John Wiley & Sons, 2012.

15. The Echo Nest. Accessed January 20, 2020. https://the.echonest.com.

16. Eftekhar, Behzad, Kazem Mohammad, Hassan E. Ardebili, Mohammad Ghodsi, and Ebrahim Ketabchi. "Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data." BMC Medical Informatics and Decision Making 5, no. 1 (2005).

17. Elisseeti, Isabelle. "An Introduction to Variable and Feature Selection." The MIT Press Journals, 2003.

18. Fausett, Laurene V. Fundamentals of Neural Networks: Architectures, Algorithms, and Applications. NJ: Prentice Hall, 1994.

19. Fayyad, Usama M. Advances in Knowledge Discovery and Data Mining. Cambridge, NJ: Mit Press, 1996.

20. Feng, Chang-Xue & Wang, Xianfeng. (2002). Digitizing uncertainty modeling for reverse engineering applications: Regression versus neural networks. Journal of Intelligent Manufacturing. 13.

21. Ferré, Joan, and F. X. Rius. "Selection of the Best Calibration Sample Subset for Multivariate Regression." Analytical Chemistry 68 (1996).

22. Fayyad, Usama M. "Data mining and knowledge discovery: making sense out of data." IEEE Expert11, no. 5 (1996), 20-25.

23. Giudici, Paolo. Applied Data Mining: Statistical Methods for Business and Industry. NJ: John Wiley & Sons, 2005.

24. Hansen, Per C. "Analysis of Discrete Ill-Posed Problems by Means of the L-Curve." SIAM Review 34, no. 4 (1992).

25. Haykin, Simon S. Neural Networks: A Comprehensive Foundation. NJ: Prentice Hall, 1999.

26. Heaton, Jeff. Artificial Intelligence for Humans, Volume 3: Deep Learning and Neural Networks. NJ: CreateSpace Independent Publishing Platform, 2015.

27. Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας, Η. Σακελλαρίου. Τεχνητή Νοημοσύνη - Γ' Έκδοση, ISBN: 978-960-8396-64-7, Έκδοση/Διάθεση: Εκδόσεις Πανεπιστημίου Μακεδονίας, 2011

28. Jeffers, J. N. "Two Case Studies in the Application of Principal Component Analysis." Applied Statistics 16 (1967), 225.

29. Jolliffe, I.T. Principal Component Analysis. Berlin, NJ: Springer Science & Business Media, 2013.

30. King, Susan L. Neural Networks vs. Multiple Linear Regression for Estimating Previous Diameter. NJ: United States Department of Agriculture, n.d. https://www.ncrs.fs.fed.us/pubs/ch/ch12/CHvolume12page159.pdf.

31. Lewis-Beck, Michael S., and Michael Lewis-Beck. Applied Regression: An Introduction. NJ: SAGE Publications, 1980.

32. Li, Mingjun, and Junxing Wang. "An Empirical Comparison of Multiple Linear Regression and Artificial Neural Network for Concrete Dam Deformation Modelling." Mathematical Problems in Engineering, 2019.

33. Kline, Rex B. Principles and Practice of Structural Equation Modeling, Fourth Edition. New York, NJ: Guilford Publications, 2015.

34. Masashi, Sugiyama. "Estimating the Error at Given Test Input Points for Linear Regression." Neural Networks, 2004.

35. Masters. Practical Neural Network Recipies in C++. Amsterdam, NJ: Elsevier, 2014.

36. Malinowski, Edmund R. "Determination of the number of factors and the experimental error in a data matrix." Analytical Chemistry 49, no. 4 (1977), 612-617.

37. Manel, Stéphanie, Jean-Marie Dias, and Steve J. Ormerod. "Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird." Ecological Modelling 120, no. 2-3 (1999), 337-347.

38. Marr, Bernard. "How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read." Forbes. Last modified September 5, 2019. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/.

39. Montgomery, Douglas C., and George C. Runger. APPLIED STATISTICS AND PROBABILITY FOR ENGINEERS, 3rd Edition. NJ: John Wiley & Sons, 2002.

40. Morrison, Donald F. Multivariate Statistical Methods. NJ: Brooks/Cole, 2005.

41. Perner, Petra. Advances in Data Mining: Applications in Medicine, Web Mining, Marketing, Image and Signal Mining, 6th Industrial Conference on Data Mining, ICDM 2006, Leipzig, Germany, July 14-15, 2006, Proceedings. Berlin, NJ: Springer Science & Business Media, 2006.

42. Pyle, Dorian. Data Preparation for Data Mining. Burlington, NJ: Morgan Kaufmann, 1999.

43. Pyzdek, Thomas. The Six Sigma Handbook, Revised and Expanded, 2nd Edition. NJ2002.

44. Schalkoff, Robert J. Artificial neural networks. New York, NJ: McGraw-Hill Companies, 1997.

45. Russell, Bradley S. "A comparison of neural network and regression models for Navy retention modeling." Master's thesis, 1993.

46. Rem, Olaf, and Marten Trautwein. "Best Practices Report Experiences with Using the Mining Mart System." Mining Mart Techreport, 2002.

47. Seal, Hilary L. Multivariate Statistical Analysis for Biologists. NJ1966.

48. Sheela, K. G., and S. N. Deepa. "Review on Methods to Fix Number of Hidden Neurons in Neural Networks." Mathematical Problems in Engineering 2013 (2013).

49. Sun, Jianguo. "A correlation principal component regression analysis of NIR data." Journal of Chemometrics 9 (1995).

50. Tavares Júnior, Ivaldo, Jonas Rocha, Ângelo Ebling, Antônio Chaves, José Zanuncio, Aline Farias, and Helio Leite. "Artificial Neural Networks and Linear Regression Reduce Sample Intensity to Predict the Commercial Volume of Eucalyptus Clones." Forests 10, no. 3 (2019).

51. Teetor, Paul. R Cookbook: Proven Recipes for Data Analysis, Statistics, and Graphics. NJ: O'Reilly Media, 2011. Lecture.

52. Tosun, Erdi, Tayfun Ozgur, Ceyla Ozgur, Mustafa Ozcanli, Hasan Serin, and Kadir Aydin. "Comparative analysis of various modelling techniques for emission prediction of diesel engine fueled by diesel fuel with nanoparticle additives." European Mechanical Science1, no. 1 (2017).

53. Upadhyay, S., N. Patel, R. Patel, and P. Kumar. "A Survey Paper of Study about Data Mining." *International Journal for Research In Advanced Computer Science And Engineering,*(2016).

54. Walpole, Ronald E., Raymond H. Myers, Keying Ye, and Sharon L. Myers. Probability & Statistics for Engineers & Scientists. NJ2002.

55. Way, J., and E.A. Smith. "The evolution of synthetic aperture radar systems and their progression to the EOS SAR." IEEE Transactions on Geoscience and Remote Sensing29, no. 6 (1991), 962-985.

56. Webster’s Revised Unabridged Dictionary. Springfield, NJ n.d. Revised 1998

57. Witten, Ian H., and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Amsterdam: Elsevier, 2005.

58. Weiss, Sholom M., and Nitin Indurkhya. Predictive Data Mining: A Practical Guide. Burlington, NJ: Morgan Kaufmann, 1998.

59. Xie, Yu-Long, and John H. Kalivas. "Evaluation of principal component selection methods to form a global prediction model by principal component regression." Analytica Chimica Acta 348, no. 1 (1997).

60. Yong Lee, Kassams S. "Generalized median filtering and related nonlinear filtering techniques." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1985.

61. Sugiyama, M. "Estimating the Error at Given Test Input Points for Linear Regression. " Proceedings of the IASTED International Conference on Neural Networks and Computational Intelligence.(2004)

62. "UCI Machine Learning Repository: Divorce Predictors Data Set Data Set." https://archive.ics.uci.edu/ml/datasets/Divorce+Predictors+data+set.

63. "UCI Machine Learning Repository: Skin Segmentation Data Set." https://archive.ics.uci.edu/ml/datasets/Skin+Segmentation.

64. "UCI Machine Learning Repository: Wine Quality Data Set." https://archive.ics.uci.edu/ml/datasets/Wine+Quality.

Παράρτημα Κώδικα

```
library(DMwR)
library(dplyr)
library(factoextra)
library(MASS)
library(neuralnet)
library(nnet)
library(pls)
library(ISLR)
library(NeuralNetTools)


#following r code is for divorce dataset. with small changes can be adapted

set.seed(123)
#path for selected data, have to choose the separator
data<-read.table("path to select data", header = TRUE, sep = ";")
summary(data)
```

```
#check for null and N/As data
if(length(which(is.na(data)))==0){message("its ok")}

boxplot(data)

#normalizing the data z-score style
scaled_data<-as.data.frame(scale(data,center = TRUE, scale = TRUE))

summary(scaled_data)

boxplot(scaled_data)
#view(scaled_data)

#get bigger print size
options(max.print = 10000)

#calculate correlation coefficients
cor(scaled_data, method = c("pearson"))
#plotof correlation coefficients
corrplot::corrplot(cor(scaled_data, method = c('pearson')))

#test and model building
#first train and test dataset creation

#shuflle the data for training and validation purpose
rows<-sample(nrow(scaled_data))

rand_scaled_data<-scaled_data[rows,]

#view(rand_scaled_data)

#create the train and test datasets from the random with split 75% train, 25% test
data_split<-floor(0.75*nrow(rand_scaled_data))
```

```
split_indexes<-sample(seq_len(nrow(rand_scaled_data)),size = data_split)

train_data =rand_scaled_data[split_indexes,]

test_data=rand_scaled_data[-split_indexes,]



#multiple linear regression-MLR!!!!
#predict data model based on all the independent variables (Atr1 to Atr54) example
of divorce dataset
#model<-Class~Atr1 + Atr2 + Atr3 + Atr4 + Atr5 + Atr6 + Atr7 + Atr8 + Atr9 +
#  Atr10 + Atr11 + Atr12 + Atr13 + Atr14 + Atr15 + Atr16 + Atr17 +
#  Atr18 + Atr19 + Atr20 + Atr21 + Atr22 + Atr23 + Atr24 + Atr25 +
#  Atr26 + Atr27 + Atr28 + Atr29 + Atr30 + Atr31 + Atr32 + Atr33 +
#  Atr34 + Atr35 + Atr36 + Atr37 + Atr38 + Atr39 + Atr40 + Atr41 +
#  Atr42 + Atr43 + Atr44 + Atr45 + Atr46 + Atr47 + Atr48 + Atr49 +
#  Atr50 + Atr51 + Atr52 + Atr53 + Atr54

#general format for n independent variables and 1 dependent
model<-dependent_variable~independent_variable_1+....+independent_variable_n

#full multiple linear regression
full_mlr<-lm(model,train_data)
summary(full_mlr)

#test predictions
#create dataframe with test idependent variables
test_variables<-test_data[1:(length(data[1,])-1)]

#create dataframe with test dependent variables
test_target_variable<-test_data[length(data[1,])]

##predict!!
predictions_full<-predict.lm(full_mlr,test_variables)
```

```
predictions_dF_full<-as.data.frame(predictions_full)

DMwR::regr.eval(test_target_variable$Class, predictions_dF_full$predictions)
full_criteria_mlr<-as.data.frame(DMwR::regr.eval(test_target_variable$Class,
predictions_dF_full$predictions))

#------------------------#

#MLR---->stepwise model mlr creation!!

step<-stepAIC(full_mlr,direction = c("backward"))

#apply the step
step_final<-lm(step$terms, train_data)
summary(step_final)

##predict with final stepwise model!!!
predictions_step<-predict(step_final,test_variables)
predictions_df_step<-as.data.frame(predictions_step)
DMwR::regr.eval(test_target_variable$Class, predictions_df_step$predictions)
step_criteria_mlr<-as.data.frame(DMwR::regr.eval(test_target_variable$Class,
predictions_df_step$predictions))

#----------------------------------------------------------#

#principal component regression-PCR

#first train with train data
pcr_test_x<-model.matrix(model, test_data)[,-1]
pcr_train_x<-model.matrix(model, train_data)[,-1]

pcr_train_y<-train_data$Class
pcr_test_y<-test_data$Class
```

```
full_pcr<-pcr(model, data=train_data)
summary(full_pcr)

##predict with pcr full
full_pcr_predict<-predict(full_pcr, pcr_test_x)

#_____________________________________________#

#criteria of full pcr model with train data
DMwR::regr.eval(full_pcr_predict, pcr_test_y)
full_criteria_pcr<-as.data.frame(DMwR::regr.eval(full_pcr_predict, pcr_test_y))

R2(full_pcr,estimate = "test", newdata = test_data)
full_pcr_stats<-mvrValstats(full_pcr, estimate = "test", newdata = test_data)

#sse of the model with all the pcs
full_pcr_sse<-full_pcr_stats$SSE[length(data[1,])]

#calculation of r2 and r2adj
r2fullpcr<-1-(full_pcr_sse/full_pcr_stats$SST)
r2adjfullpcr<-1-(1-r2fullpcr)*((length(train_data[,1]))/(length(train_data[,1])-
length(full_pcr_stats$comps)-1-1))
1-(1-r2fullpcr)*((length(train_data[,1]))/(length(train_data[,1])-
length(full_pcr_stats$comps)-1-1))

#-------------------#
##validation plots to select number of PCs
validationplot(full_pcr, val.type="RMSEP", cex.axis=1.1)
axis(side = 1, at = c(4), cex.axis=1.1)
abline(v = 4, col = "blue", lty = 5)

validationplot(full_pcr,val.type  = "MSEP")
axis(side = 1, at = c(4), cex.axis=1.1)
```

```
abline(v = 4, col = "blue", lty = 5)

validationplot(full_pcr,val.type  = "R2")
axis(side = 1, at = c(4), cex.axis=1.1)
abline(v = 4, col = "blue", lty = 5)

#------------#
#predict with the model using 4 PCs
limited_pcr_pred<-predict(full_pcr, pcr_test_x,comps =1:4)
DMwR::regr.eval(limited_pcr_pred, pcr_test_y)
limit_criteria<-as.data.frame(DMwR::regr.eval(limited_pcr_pred, pcr_test_y))

R2(full_pcr,estimate = "test", newdata = test_data, comps =1:4)

limit_pcr_stats<-mvrValstats(full_pcr, estimate = "test", newdata = test_data,comps
=1:4)

r2limitpcr<-1-(limit_pcr_stats$SSE[1]/limit_pcr_stats$SST)
1-(limit_pcr_stats$SSE[1]/limit_pcr_stats$SST)
r2adjlimitpcr<-1-(1-r2limitpcr)*((length(train_data[,1])-1)/(length(train_data[,1])-
length(limit_pcr_stats$comps)-1))
1-(1-r2limitpcr)*((length(train_data[,1])-1)/(length(train_data[,1])-
length(limit_pcr_stats$comps)-1))
#view(r2adjlimitpcr)

#_____________________________________________________#

##partial least squares
#first train with train data
pls_test_x<-model.matrix(model, test_data)[,-1]
pls_train_x<-model.matrix(model, train_data)[,-1]

pls_train_y<-train_data$Class
pls_test_y<-test_data$Class
```

```
full_pls<-plsr(model, data=train_data)
summary(full_pls)
write.csv(full_pls$coefficients,file = "fullplscoef.xls")

##predict with pls full model the test data
full_pls_predict<-predict(full_pls, pls_test_x)

#criteria of full pls with train data
DMwR::regr.eval(full_pls_predict, pls_test_y)
full_criteria_pls<-as.data.frame(DMwR::regr.eval(full_pls_predict, pls_test_y))

R2(full_pls,estimate = "test", newdata = test_data)

full_pls_stats<-mvrValstats(full_pls, estimate = "test", newdata = test_data)

#sse of the model with all the pcs
full_pls_sse<-full_pls_stats$SSE[length(data[1,])]

#calculation of r2 and r2adj
r2fullpls<-1-(full_pls_sse/full_pls_stats$SST)
1-(full_pls_sse/full_pls_stats$SST)

r2adjfullpls<-1-(1-r2fullpls)*((length(train_data[,1]))/(length(train_data[,1])-
length(full_pls_stats$comps)-1-1))
1-(1-r2fullpls)*((length(train_data[,1]))/(length(train_data[,1])-
length(full_pls_stats$comps)-1-1))
#view(r2adjfullpls)

##validation plots to select number of PCs for pls
validationplot(full_pls, val.type="RMSEP", cex.axis=1.1)
axis(side = 1, at = c(2), cex.axis=1.1)
abline(v = 2, col = "blue", lty = 5)
```

```
validationplot(full_pls,val.type  = "MSEP")
axis(side = 1, at = c(2), cex.axis=1.1)
abline(v = 2, col = "blue", lty = 5)

validationplot(full_pls,val.type  = "R2")
axis(side = 1, at = c(2), cex.axis=1.1)
abline(v = 2, col = "blue", lty = 5)


#predict with the model using number of PCs for pls
limited_pls_pred<-predict(full_pls, pls_test_x,comps =1:2)
DMwR::regr.eval(limited_pls_pred, pls_test_y)
limit_criteria<-as.data.frame(DMwR::regr.eval(limited_pls_pred, pls_test_y))

R2(full_pls,estimate = "test", newdata = test_data, comps =1:2)

limit_pls_stats<-mvrValstats(full_pls, estimate = "test", newdata = test_data,comps
=1:2)

r2limitpls<-1-(limit_pls_stats$SSE[1]/limit_pls_stats$SST)
1-(limit_pls_stats$SSE[1]/limit_pls_stats$SST)
r2adjlimitpls<-1-(1-r2limitpls)*((length(train_data[,1])-1)/(length(train_data[,1])-
length(limit_pls_stats$comps)-1))
1-(1-r2limitpls)*((length(train_data[,1])-1)/(length(train_data[,1])-
length(limit_pls_stats$comps)-1))
#view(r2adjlimitpcr)

#_____________________________#
#neural networks!!!!!###


#testing....with multiple neurons on hidden layers
neuralnetwork1 <- neuralnet(model,data = train_data, hidden = 5, linear.output =
TRUE, algorithm ="backprop", learningrate = 0.00001,startweights = NULL, rep = 20)
```

```
pred_nn1<-compute(neuralnetwork1,test_variables, rep = 1)
mse.nn1<-sum((test_target_variable -
pred_nn1$net.result)^2)/nrow(test_target_variable)
sum((test_target_variable - pred_nn1$net.result)^2)/nrow(test_target_variable)
ssr1<-sum((test_target_variable - pred_nn1$net.result)^2)
sum((test_target_variable - pred_nn1$net.result)^2)
RMSE.NN1 = sqrt(sum((test_target_variable - pred_nn1$net.result)^2) /
nrow(test_target_variable))
sqrt(sum((test_target_variable - pred_nn1$net.result)^2) /
nrow(test_target_variable))
mae.nn1<-sum(abs(test_target_variable -
pred_nn1$net.result))/nrow(test_target_variable)
sum(abs(test_target_variable - pred_nn1$net.result))/nrow(test_target_variable)
ssres.nn1<-sum((test_target_variable - pred_nn1$net.result)^2)
sstot.nn1<-sum((test_target_variable-
sum(test_target_variable)/nrow(test_target_variable))^2)
r2.nn1<-1-(ssres.nn1/sstot.nn1)
1-(ssres.nn1/sstot.nn1)
r2adj.nn1<-1-(1-r2.nn1)*((length(train_data[,1])-1)/(length(train_data[,1])-
length(pred_nn1$net.result)-1))
1-(1-r2.nn1)*((length(train_data[,1])-1)/(length(train_data[,1])-
length(pred_nn1$net.result)-1))
```