ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΠΠΣ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ ΜΕΣΟΛΟΓΓΙ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«ΕΦΑΡΜΟΓΗ ΚΑΙ ΣΥΓΚΡΙΣΗ ΑΛΓΟΡΙΘΜΩΝ ΕΠΙΒΛΕΠΟΜΕΝΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΕ ΔΙΑΦΟΡΕΤΙΚΕΣ ΜΕΛΕΤΕΣ ΠΕΡΙΠΤΩΣΗΣ»

Κωνσταντίνος Κούτρας

Α.Μ.: 16658

Μεσολόγγι 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΠΠΣ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ ΜΕΣΟΛΟΓΓΙ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«ΕΦΑΡΜΟΓΗ ΚΑΙ ΣΥΓΚΡΙΣΗ ΑΛΓΟΡΙΘΜΩΝ ΕΠΙΒΛΕΠΟΜΕΝΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΕ ΔΙΑΦΟΡΕΤΙΚΕΣ ΜΕΛΕΤΕΣ ΠΕΡΙΠΤΩΣΗΣ»

Κωνσταντίνος Κούτρας

Επιβλέπων καθηγητής

κ. Αριστογιάννης Γαρμπής

Μεσολόγγι 2021

UNIVERSITY OF PATRAS

SCHOOL OF ECONOMICS & BUSINESS

DEPARTMENT OF MANAGEMENT SCIENCE AND TECHNOLOGY

FORMER DEPARTMENT OF BUSINESS ADMINISTRATION AT MESSOLONGHI

THESIS

«APPLICATION AND COMPARISON OF SUPERVISED MACHINE LEARNING ALGORITHMS IN DIFFERENT CASE STUDIES»

Konstantinos Koutras

Messolonghi 2021

Η έγκριση της πτυχιακής εργασίας από το Τμήμα Διοικητικής Επιστήμης και Τεχνολογίας του Πανεπιστημίου Πατρών δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

Ευχαριστίες

Στο σημείο αυτό, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κ. Αριστογιάννη Γαρμπή για τη συνεργασία μας στην τελική επιλογή θέματος της παρούσας πτυχιακής εργασίας μου. Επιπλέον, τον ευχαριστώ θερμά για τη διαρκή καθοδήγησή του μέχρι και τη διεκπεραίωσή της.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και όλους όσοι με στήριξαν κατά τη διάρκεια των εκπαιδευτικών μου χρόνων, μέχρι και την ολοκλήρωση των σπουδών μου.

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια και σε παγκόσμιο επίπεδο, σε μια κοινωνία ολοένα εξελισσόμενη στην έννοια της πληροφορίας, παρατηρείται μια σημαντική αύξηση στην εφαρμογή «ευφυιών» προγραμμάτων σε υπολογιστικά συστήματα. Όλα αυτά τα προγράμματα περιέχουν ένα υποθετικό βαθμό «αντίληψης», καθιστώντας τα ικανά να προσομοιάσουν τον ανθρώπινο τρόπο σκέψης στην τεχνολογική «εφαρμογή» του, ταξινομώντας, προβλέποντας και διακρίνοντας γεγονότα του πραγματικού κόσμου, με έναν πολύ γρήγορο και διεξοδικό τρόπο.

Όλες αυτές οι τεχνολογίες εμπίπτουν στο πεδίο της «*Μηχανικής Μάθησης*». Ο συγκεκριμένος τεχνολογικός τομέας, εμπλέκει διάφορες προσεγγίσεις και μεθοδολογίες που περιέχονται τόσο στον τομέα της πληροφορικής, όσο και στους τομείς της εξόρυξης πληροφορίας και της στατιστικής. Η βασική ιδέα της μηχανικής μάθησης, αφορά στην κατασκευή προγραμμάτων υπολογιστών που θα είναι σε θέση να επιτελούν διάφορες αυτοματοποιημένες εργασίες, βάσει εμπειρικών καταστάσεων και δεδομένων, που έχουν ήδη «αντικρίσει». Αυτή η προσέγγιση για την δημιουργία τέτοιων προγραμμάτων, αναφέρεται και ως «*επιβλεπόμενη μάθηση*». Ωστόσο, για να επιτευχθεί αυτό, γίνεται χρήση διαφόρων αλγοριθμικών τεχνικών που έχουν ως βάση την ανακάλυψη γνώσης από σύνολα δεδομένων.

Στην παρούσα Πτυχιακή Εργασία, πραγματοποιείται μια διεξοδική ανάλυση των στοιχείων εκείνων που συντελούν, αυτό το ευρύ πεδίο της μηχανικής μάθησης. Επιπροσθέτως, στο πλαίσιο της παρούσας εργασίας, γίνεται παρουσίαση τριών μελετών περιπτώσεων που πραγματοποιήθηκαν με τεχνικές και μεθόδους του πεδίου αυτού της μηχανικής μάθησης που θα μελετηθεί.

Σκοπός της εργασίας αυτής, αποτελεί η λεπτομερής ανασκόπηση όλων των μεταβλητών αυτών που απαρτίζουν το θεωρητικό μέρος της ανάπτυξης προγραμμάτων μηχανικής μάθησης με επιβλεπόμενο κυρίως τρόπο και, η εύρεση των καλύτερων αλγοριθμικών τεχνικών στις τρείς αυτές μελέτες περίπτωσης, κάνοντας χρήση διαφόρων μεθοδολογιών και προσεγγίσεων.

ABSTRACT

In the recent years and in a worldwide level, in all these societies who keep evolving by the term of information, a significant increase in the usage and implementations of intelligent programs in computer systems it is observed. All these programs are meant to have a hypothetical degree of perception, making them capable to emulate the human way of thinking in its technological application, classifying, predicting, and distinguishing facts of the real world, in a very fast and a thorough way.

All these types of technologies fall into the field of "*Machine Learning"*. This specific technological sector contains a variety of approaches and methodologies which can be part of the informatic sector, as well as the sectors of data mining and statistics. The main concept behind machine learning, is the creation of software programs that will be able to do a lot of different automized tasks, according to their experience in previous cases and data, that have already faced. This approach of constructing such programs, is also known as "*supervised learning"*. However, in order for this to happen, there are used a lot of different types of algorithmic techniques, which are based on the knowledge discovery from data sets.

In this present Thesis, a thorough analysis of all those elements regarded in the wide field of machine learning is carried out. In addition, in the content of this present work, there is a presentation of three case studies that were carried out by techniques and methods of this sector of machine learning that is going to be displayed.

The main aim of this project, is the detailed retrospect of all the variables that contribute to the theoretical part of supervised machine learning program development, and the discovery of the best algorithmic techniques in these case studies, implementing different types of methodologies and processing approaches.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

AI	Artificial Intelligence
ANN	Artificial Neural Network
ARFF	Attribute Relation File Format
CLI	Command Line Interface
CSV	Comma Separated Values
IDE	Integrated Development Environment
IoT	Internet of Things
kNN	k-Nearest Neighbors
ML	Machine Leaning
MLP	Multi-Layer Perceptron
SQL	Structured Query Language
SVM	Support Vector Machines
WWW	World Wide Web

ΑΠΟΔΟΣΗ ΟΡΩΝ

Algorithm	Αλγόριθμος
Attribute	Γνώρισμα
Artificial Intelligence	Τεχνητή νοημοσύνη
Classification	Κατηγοριοποίηση
Data	Δεδομένα
Database	Βάση δεδομένων
Data Set	Σύνολο δεδομένων
Entropy	Εντροπία
Framework	Πλαίσιο
Fuzzy Logic	Ασαφής λογική
Input	Είσοδος
Instance	Στιγμιότυπο
Linear	Γραμμικός
Machine Learning	Μηχανική μάθηση
Output	Έξοδος
Port	Θύρα
Programming	Προγραμματισμός
Regression	Παλινδρόμηση
Regularization	Κανονικοποίηση

ΕΙΣΑΓΩΓΗ

Τη σύγχρονη εποχή, εμφανίζονται όλο και περισσότερες ευφυείς εφαρμογές σε υπολογιστικά συστήματα που βοηθούν την ανθρωπότητα στην έγκαιρη πρόβλεψη γεγονότων αλλά και στην λεπτομερής ανάλυση δεδομένων. Όλες αυτές οι εφαρμογές κατασκευάζονται με σκοπό να βελτιστοποιήσουν το επίπεδο βιωσιμότητας, μέσω στατιστικών, λογικών και, αλγοριθμικών μεθόδων καθιστώντας τον άνθρωπο, εξαρτώμενο πλέον από αυτές. Η κατασκευή ενός τέτοιου συστήματος δεν είναι καθόλου απλή υπόθεση, καθώς περιλαμβάνει πολλές παραμέτρους. Οι δύο βασικότερες είναι: α) η σωστή διαχείριση των ειδών δεδομένων που δίνονται με τη μορφή εισόδου στο σύστημα για επεξεργασία και, β) η επιλογή των κατάλληλων αλγοριθμικών τεχνικών που θα εφαρμοστούν, αναλόγως του επιθυμητού στόχου. Συνεπώς, ως πρώτο βήμα πρέπει να γίνει μια λεπτομερής συλλογή δεδομένων από έγκυρες πηγές σε συνδυασμό με τη σωστή προ-επεξεργασία τους, για την αποφυγή λανθασμένων αποτελεσμάτων. Το επόμενο στάδιο αποτελεί η επιλογή της σωστής τεχνικής για την επεξεργασία των δεδομένων ώστε να ληφθεί το πιο «έμπιστο» πιθανό αποτέλεσμα.

Ο πρώτος άνθρωπος που λέγεται ότι έδωσε ορισμούς σχετικά με την υπολογιστική ευφυία ήταν ο φημισμένος μαθηματικός Άλαν Τούρινγκ στα τέλη τις δεκαετίας του 1940 (Γαρμπής και Φωτιάδης, 2015). Ο Τούρινγκ, καθώς και διάφοροι άλλοι επιστήμονες παρεμφερούς πεδίου της τότε εποχής, οραματίζονταν την τέλεια μηχανή η οποία θα μπορούσε να αναπαραστήσει σε μεγάλο βαθμό την ανθρώπινη ευφυία και να είναι σε θέση να κάνει μελλοντικές προβλέψεις. Ο όρος αυτός λοιπόν, που αναπαριστά το γενικό πεδίο της ευφυΐας των υπολογιστών ονομάζεται *τεχνητή νοημοσύνη* (artificial intelligence – AI). Σε αυτό το πεδίο εμπίπτει και οποιαδήποτε διεργασία ενός υπολογιστή που μπορεί να προσομοιάσει τον ανθρώπινο τρόπο σκέψης, να μάθει από εμπειρίες και να κατηγοριοποιήσει καταστάσεις. Ένα από αυτά τα πεδία αποτελεί και η *μηχανική μάθηση* (machine learning – ML).

Στο μεγάλο πεδίο αυτό της μηχανικής μάθησης χρησιμοποιούνται μια πληθώρα από αλγορίθμους εξόρυξης πληροφορίας, οι οποίοι μπορούν να καταστίσουν ένα υπολογιστικό σύστημα ικανό να βελτιώνει τις διεργασίες του μαθαίνοντας από εμπειρικά δεδομένα, αλλά και να είναι σε θέση να επιτελεί προβλέψεις και ταξινομήσεις σε διάφορες καταστάσεις. Η μηχανική μάθηση, σε συνδυασμό με τις τεχνικές της εξόρυξης πληροφορίας, έχουν χρησιμοποιηθεί σε πολλά επιστημονικά πεδία για την εξεύρεση γνώσης σε οποιαδήποτε

μορφής αποθηκευμένα δεδομένα (όπως για παράδειγμα οι βάσεις δεδομένων), βοηθώντας έτσι τους ανθρώπους να ανακαλύψουν πληροφορίες που ίσως να μην είναι εμφανείς με την πρώτη ματιά.

Επιδιώκοντας το καλύτερο αποτέλεσμα κατανόησης και επεξήγησης ως προς τον αναγνώστη, η συγκεκριμένη πτυχιακή εργασία χωρίζεται σε δύο μέρη.

Στο «Μέρος Α», γίνεται μια βιβλιογραφική επισκόπηση των μεγάλων πεδίων της μηχανικής μάθησης και της εξόρυξης πληροφορίας, καθώς και των διαφόρων κατηγοριών που εμπίπτουν στο κάθε ένα. Πριν φτάσουμε στο σημείο αυτό όμως, πραγματοποιείται μια συνοπτική αναφορά στον όρο των δεδομένων και της πληροφορίας, στοιχεία αναπόσπαστα από τα δύο πεδία αυτά που αναφέρθηκαν.

Ως σκοπός του μέρους αυτού της πτυχιακής εργασίας, αποτελεί η κατανόηση προς τον αναγνώστη, μέσω της περιγραφής μερικών από των πιο στοιχειωδών θεωρητικών εννοιών των πεδίων της μηχανικής μάθησης και της εξόρυξης πληροφορίας. Μια γενική γνώση πάνω σε αυτούς τους τομείς κρίνεται απαραίτητη, καθώς σήμερα οι τεχνολογίες αυτές χρησιμοποιούνται σε μια πληθώρα καταστάσεων που μας περιβάλλουν.

Στο «Μέρος Β», γίνεται η παρουσίαση της εφαρμογής και των αποτελεσμάτων αλγορίθμων σε σύνολα δεδομένων τα οποία συλλέχθηκαν από το διαδίκτυο. Οι αλγόριθμοι αυτοί που χρησιμοποιήθηκαν, εμπίπτουν τόσο στο πεδίο της μηχανικής μάθησης, όσο και στο πεδίο της εξόρυξης πληροφορίας. Οι πρακτικές και οι μεθοδολογίες που ακολουθήθηκαν για την διεκπεραίωση των μελετών στα σύνολα δεδομένων, συγχέονται με το πρώτο μέρος της εργασίας όπου και γίνεται η θεωρητική επεξήγησή τους.

Ως σκοπός του δεύτερου μέρους της πτυχιακής εργασίας, αποτελεί η εφαρμογή, η ανάλυση και η αξιολόγηση μερικών γνωστών αλγορίθμων των πεδίων που αναφέρθηκαν σε τρείς διαφορετικές μελέτες περίπτωσης. Επιπλέον, επιδιώχθηκε η χρήση διαφορετικών προσεγγίσεων για την εφαρμογή των αλγορίθμων αυτών, με απώτερο σκοπό την εύρεση αυτού που θα ήταν σε θέση να φέρει τα μεγαλύτερα ποσοστά ακρίβειας και ορθότητας στα αποτελέσματά του.

ΜΕΡΟΣ Α΄

Το παρόν πρώτο μέρος της εργασίας, επικεντρώνεται στη θεωρητική επεξήγηση των στοιχείων εκείνων που θα υποστηρίξουν το εμπειρικό μέρος της. Ειδικότερα, επιδιώκεται η παρουσίαση που αφορά στην ανάλυση και την επεξεργασία των δεδομένων που καταχωρούνται σε ένα σύστημα μηχανικής μάθησης για τη δημιουργία ενός ανεξάρτητου μοντέλου, ικανού να πραγματοποιήσει κατηγοριοποιήσεις και προβλέψεις.

Στο πρώτο κεφάλαιο, γίνεται μια βιβλιογραφική ανασκόπηση σχετικά με στοιχειώδεις ορισμούς που αφορούν στα δεδομένα, καθώς και την επεξήγηση της ιδιότητας τους ως ένα βασικό στοιχείο στη μηχανική μάθηση και την εξόρυξη πληροφορίας. Όπως αναφέρθηκε, η ποιότητα και η προ-επεξεργασία των δεδομένων αποτελούν δύο πολύ σημαντικά στοιχεία για την ανάπτυξη ενός προγράμματος, ικανού να κάνει προβλέψεις. Συνεπώς, κρίθηκε απαραίτητη η αναφορά στους διάφορους τύπους των δεδομένων και στο κρίσιμο στάδιο της προ-επεξεργασίας τους.

Στο δεύτερο κεφάλαιο που θα ακολουθήσει, γίνεται μια λεπτομερής παρουσίαση στους διαφορετικούς τρόπους με τους οποίους ένα υπολογιστικό σύστημα μπορεί να αναπτύξει τις εκάστοτε ικανότητες βάσει των διαδικασιών της μηχανικής μάθησης. Συγκεκριμένα, γίνεται ανάλυση των διάφορων ειδών «μάθησης» στις οποίες μπορεί να υποβληθεί ένα τέτοιο σύστημα. Μεταξύ άλλων, οι πιο βασικές από αυτές είναι η *επιβλεπόμενη* (supervised learning) και η *μη επιβλεπόμενη μάθηση* (unsupervised learning).

Στο τρίτο κεφάλαιο της παρούσης εργασίας, πραγματοποιείται μια συνοπτική ανασκόπηση του πεδίου της εξόρυξης πληροφορίας, ως ένα σημαντικό συστατικό στοιχείο που συγχέεται πολύ στενά με τη μηχανική μάθηση. Παραδειγματικά, η πρόβλεψη και η κατηγοριοποίηση γεγονότων μπορούν να χρησιμοποιηθούν μεν και στα δύο αυτά πεδία, έχοντας όμως μερικές διαφορετικές προοπτικές στη χρήση των αποτελεσμάτων τους. Η εξόρυξη πληροφορίας αποτελεί ένα πολύ σημαντικό επιστημονικό πεδίο το οποίο ουσιαστικά στοχεύει στην ανακάλυψη γνώσης σε σύνολα δεδομένων. Αντίθετα, η μηχανική μάθηση χρησιμοποιεί αλγοριθμικές τεχνικές της εξόρυξης πληροφορίας για τη δημιουργία ενός υπολογιστικού συστήματος που θα μπορεί να αναπαραστήσει ένα επίπεδο αυτονομίας και ευφυΐας. Στο πλαίσιο του κεφαλαίου αυτού, γίνεται επίσης αναφορά σε μερικές βασικές αλγοριθμικές τεχνικές της μηχανικής μάθησης που θα χρησιμοποιηθούν στο Μέρος Β.

1 Δεδομένα και πληροφορία

Ο όρος «*δεδομένα*» συναντάται παντού σήμερα. Είναι ένα από τα δημιουργήματα του ανθρώπου που μας δίνει την ικανότητα να πληροφορηθούμε για τις διάφορες καταστάσεις που μας περιβάλλουν, κάνοντας μας να πράξουμε αναλόγως με την εκάστοτε κατάσταση που αντιμετωπίζουμε. Από τις τεχνολογικές και τις θετικές επιστήμες, μέχρι και τις επιστήμες της οικονομίας και τις διοίκησης και τις επιστήμες υγείας, οι όροι δεδομένα και πληροφορία έχουν παίξει αναπόσπαστο ρόλο για την επιτέλεση διαφόρων εργασιών. Στην καθημερινότητα, είμαστε μάρτυρες κάθε είδους πληροφορίας που βρίσκεται στο περιβάλλον μας. Αυτές οι πληροφορίες που αντλούμε ποικίλουν και, μπορεί να είναι από ειδήσεις σε ένα τηλεοπτικό κανάλι, γεγονότα που συμβαίνουν, έως νέα και ιδέες που αποκτούνται και θεωρούνται γνώση. Οι περισσότεροι άνθρωποι μπορεί να μην το συνειδητοποιούν, αλλά μέχρι και ένα πολύ «μικρό» γεγονός μπορεί να θεωρηθεί ως μια πληροφορία, καθιστώντας μας ικανούς να πράξουμε μια επιλογή ασυνείδητα.

Σήμερα, είναι γενικά αποδεκτό από διάφορα επιστημονικά πεδία, ότι τα περισυλλεγμένα δεδομένα από έρευνες και μελέτες, παραδείγματος χάριν, στην αρχική τους μη κατεργασμένη μορφή συνήθως δεν είναι χρήσιμα και συνεπώς δεν αποτελούν πληροφορία (Ματσατσίνης, 2010). Τι είναι όμως η πληροφορία; Για να εξηγηθεί καλύτερα αυτό το ερώτημα πρέπει να αναλυθεί πιο διεξοδικά αρχικά το τι είναι τα δεδομένα.

Ως *δεδομένο* (data), ορίζεται οποιαδήποτε μετρήσιμη ή υπολογίσιμη τιμή μιας ιδιότητας (Γεωργούλη, 2015). Ένας δεύτερος ορισμός για τα δεδομένα, είναι ότι ορίζονται ως ένα μη αξιολογημένο σύνολο διακριτών στοιχείων μιας αναφοράς, το οποίο «αποτυπώνει» τιμές επί αντικειμένων ή καταστάσεων. Βάσει αυτών, μπορούμε να καταλάβουμε ότι ένα στοιχείο από ένα σύνολο δεδομένων μπορεί να περιέχει διάφορα γνωρίσματα που το «περιγράφουν».

Ο όρος «*πληροφορία*» όμως, έχει μια αρκετά διαφορετική έννοια από ότι τα δεδομένα. Η πληροφορία, απαρτίζεται μεν από δεδομένα, αλλά για να θεωρηθεί ως πηγή πληροφόρησης και «γνώσης» τα δεδομένα αυτά πρέπει να υποστούν μια επεξεργασία και μια μορφοποίηση (Ματσατσίνης, 2010).

Η διαχείρισή των δεδομένων και των πληροφοριών είναι πραγματικά πολύ σημαντικές διεργασίες που επιτελούνται εδώ και πολλά χρόνια. Σε όλα τα επιστημονικά πεδία, υπάρχουν ειδικοί ξεχωριστοί συμβολισμοί και «κανόνες» για την αναπαράσταση των

δεδομένων, των συσχετίσεων τους και τις πιθανές μεθόδους επεξεργασίας τους. Τα δεδομένα, διαχειρίζονται από διάφορες οντότητες στον σημερινό κόσμο, συμπεριλαμβανομένων ανθρώπων και υπολογιστών. Από μια άποψη, εφόσον ο όρος πληροφορίες και δεδομένα κλίνει όλο και περισσότερο προς την ψηφιοποίηση, οι υπολογιστές είναι αυτοί που θα παίξουν τον πρωταρχικό ρόλο καταγράφοντας και επεξεργάζοντας τα. Έτσι, οι άνθρωποι καταλήγουν να κάνουν όλες τις αναλύσεις και επεξεργασίες που χρειάζονται με έναν πιο εύκολο τρόπο μέσω των υπολογιστών. Το κύριο πλεονέκτημα ανάλυσης δεδομένων μέσω των υπολογιστικών συστημάτων, αποτελεί η διοχέτευση μεγάλου όγκο πληροφοριών σε αυτά, η μεγάλη επεξεργαστική ισχύς τους, αλλά και η επίλυση προβλημάτων σε πολύ μικρά χρονικά διαστήματα.

Όσον αναφορά στην αποθήκευση των δεδομένων, σήμερα, χρησιμοποιούνται πολλά και διαφορετικά λογισμικά συστήματα που επιτελούν τη συγκεκριμένη εργασία. Ένα χαρακτηριστικό παράδειγμα από αυτά είναι οι βάσεις δεδομένων. Σε αυτές, μπορούν να αναπαρασταθούν και να αποθηκευτούν μεγάλοι όγκοι πληροφοριών, επιτρέποντας σε κάποιον να ανατρέξει σε αυτές οποιαδήποτε στιγμή αυτός επιθυμήσει (Ταμπακάς, 2017). Πέραν των βάσεων δεδομένων, υπάρχουν και άλλοι τρόποι αποθήκευσης της πληροφορίας στους υπολογιστές σε διάφορα λογισμικά συστήματα και σε διάφορους τύπους αρχείων (όπως πχ τα αρχεία τύπου csv, arff κλπ.).

1.1 Τύποι δεδομένων

Οι «*τύποι των δεδομένων*» (data type), προσδιορίζουν την κατηγοριοποίηση ενός δεδομένου ανάμεσα σε ένα σύνολο δεδομένων, την σημασία του και, τον τρόπο που οι τιμές του μπορούν να αποθηκευτούν (Wiki: Δεδομένα, 2020). Η κύρια ιδιότητα των τύπων των δεδομένων, είναι ότι προσφέρουν διάφορους τρόπους ορισμού, καθώς και συγκεκριμένες τεχνικές για την υλοποίηση και τη χρήση τους. Βέβαια, όλα αυτά τα γνωρίσματα για τους τύπους των δεδομένων είναι απολύτως αντικειμενικά και αφορούν μόνον στον άνθρωπο, μιας και ο υπολογιστής δεν κατανοεί τίποτα παραπάνω από κωδικοποιημένες συμβολοσειρές bit, του «μηδέν» και του «ένα» (Μπέχρουζ, 2015).

Με μια πρώτη διάκριση μπορούμε να διαχωρίσουμε δύο τύπους δεδομένων, τα δεδομένα της *παρατήρησης* και, τα δεδομένα της *πληροφόρησης*[1]. Τα δεδομένα της παρατήρησης πρόκειται για δεδομένα που μπορούν να υποκύψουν σε επεξεργασία ή αξιολόγηση για εύρεση πληροφορίας και γνώσης, έχοντας ένα ορισμένο νόημα με συγκεκριμένη οργάνωση. Ένα τέτοιο παράδειγμα συνόλου δεδομένων είναι τα δεδομένα που έχουν συλλεχθεί από ένα ερωτηματολόγιο ή μια μελέτη. Όσον αφορά στη δεύτερη κατηγορία, τα δεδομένα της πληροφόρησης, πρόκειται για εκείνους τους τύπους δεδομένων που περιγράφουν απλά ένα γεγονός ή μια κατάσταση και δεν περιλαμβάνουν καμία βάση για στατιστική ανάλυση ή περεταίρω ενέργεια. Για παράδειγμα, αυτούς τους τύπους δεδομένων επεξεργάζεται, αποθηκεύει και τροποποιεί ο υπολογιστής. Αξίζει να σημειώσουμε, κάπου εδώ, ότι όλες οι επόμενες κατηγορίες τύπων δεδομένων που θα αναφερθούμε σε αυτήν την ενότητα, καθώς και για ολόκληρη την υπόλοιπη πτυχιακή εργασία, πρόκειται για δεδομένα τύπου παρατήρησης.

Μεταξύ άλλων, τα δεδομένα δεν εγγυώνται ποτέ την ορθότητα τους, καθώς υπάρχουν πολλοί παράγοντες που μπορούν να τα επηρεάσουν. Συνεπώς, με μια δεύτερη διάκριση μπορούμε να χωρίσουμε τα δεδομένα σε *βέβαια-σαφή*, τα οποία περιέχουν ένα επαρκή μέτρο σιγουριάς και εγκυρότητας, και σε *αβέβαια- ασαφή*, τα οποία δεν είναι σαφώς οριοθετημένα και δεν εμπνέουν βεβαιότητα (Γεωργούλη, 2015). Τα υπολογιστικά συστήματα και, πιο συγκεκριμένα τα ευφυή συστήματα υπολογιστών, έχουν κατασκευαστεί με τέτοιον τρόπο ώστε να χειρίζονται τέτοιους τύπους δεδομένων. Αυτή η διαχείριση επιτυγχάνεται κυρίως μέσω αλγοριθμικών διαδικασιών της ασαφούς λογικής: Η «*ασαφής λογική*» (fuzzy logic), είναι ένα πεδίο των μαθηματικών που ασχολείται με τη μελέτη συνόλων και δεδομένων, χωρίς απόλυτα καθορισμένα όρια, εξαρτώμενα το κάθε ένα από ένα διαφορετικό βαθμό συμμετοχής, σε ένα άλλο σύνολο (Ρώτα, 2008:86).

Βάσει της τρίτης και της πιο σημαντικής διάκρισης, η οποία βασίζεται στις βασικές αρχές της επιστημονικής περιοχής της στατιστικής, οι τύποι δεδομένων μπορεί να είναι *ποιοτικοί* ή *ποσοτικοί* (Μητρόπουλος, 2009). Οι *ποσοτικοί* τύποι δεδομένων (quantitative data) μπορούν να διακριθούν στους διακριτούς και στους συνεχείς, ενώ οι *ποιοτικοί* τύποι δεδομένων (qualitive data) μπορούν να διακριθούν στους διατεταγμένους και στους ονομαστικούς.

---

[1] Wiki: Δεδομένα (2020)

- Οι **ποσοτικοί τύποι** δεδομένων χαρακτηρίζονται αυστηρά μόνο από τους πραγματικούς αριθμούς ($\mathbb{R}$) στις μεταβλητές τους, καθιστώντας τις αριθμητικές πράξεις στα μεγέθη τους εφικτές (Γναρδέλλης, 2003). Με τη σειρά τους, οι ποσοτικοί τύποι δεδομένων μπορούν να διακριθούν σε *διακριτά* δεδομένα (discrete data), και σε *συνεχή* (interval data).
    - Οι *διακριτοί* τύποι δεδομένων, μπορούν να πάρουν μόνο ένα ορισμένο αριθμό τιμών (συνήθως παίρνουν ακεραίους αριθμούς, χωρίς να έχουν τη δυνατότητα να πάρουν και άλλες τιμές ενδιάμεσα). Η πιο συνηθισμένη περίπτωση διακριτού δεδομένου είναι αυτά της ποσοτικής αρίθμησης στοιχείων (Γναρδέλλης, 2003).
    - Οι *συνεχείς* τύποι δεδομένων μπορούν να πάρουν οποιαδήποτε τιμή στο εύρος όλων των πραγματικών αριθμών. Αυτός ο τύπος δεδομένων, αποσκοπεί στην ακριβής μέτρηση μιας ποσότητας χωρίς να εμπίπτει στους περιορισμούς των ακέραιων αριθμών. Μερικά παραδείγματα συνεχών τύπων δεδομένων είναι η θερμοκρασία, ο χρόνος, η πυκνότητα, το ύψος κ.λπ.
- Οι **ποιοτικοί τύποι** δεδομένων χαρακτηρίζονται από τιμές ή συμβολοσειρές χαρακτήρων, που αντιπροσωπεύουν την αναπαράσταση των δεδομένων σε κατηγορίες ή σε μια διάταξη και, δεν έχουν ποσοτική αξία (Μητρόπουλος, 2009). Με τη σειρά τους, οι ποιοτικοί τύποι δεδομένων μπορούν να διακριθούν σε *διατεταγμένα* δεδομένα (ordinal data), και *ονομαστικά* (categorical data).
    - Οι *διατεταγμένοι* τύποι δεδομένων είναι οι τύποι των δεδομένων των οποίων οι κατηγορίες ορίζονται βάσει μιας σχέσης διάταξης που υφίσταται μεταξύ τους (Γναρδέλλης, 2003). Για παράδειγμα, διατεταγμένα δεδομένα μπορούν να χαρακτηριστούν τα διαφορετικά επίπεδα αποδοχής σε ένα προϊόν από έναν πελάτη (κακό, μέτριο, καλό, πολύ καλό κλπ.).
    - Οι *ονομαστικοί,* ή αλλιώς *κατηγορηματικοί,* τύποι δεδομένων, αποτελούν τα δεδομένα τα οποία ταξινομούνται σε κατηγορίες σαφώς διαχωρισμένες μεταξύ τους. Ένα τέτοιο παράδειγμα είναι οι τιμές που μπορεί να πάρει ένα δεδομένο ή μια μεταβλητή της αλήθειας-ψέματος, το σωστό-λάθος, της ομάδας αίματος και ούτω κάθε εξής.

Αναλόγως τους τύπους των δεδομένων που είναι καταγεγραμμένα σε ένα σύνολο, μπορούν να πραγματοποιηθούν και διάφοροι υπολογισμοί σχετικά με αυτά, ενημερώνοντας μας για διάφορες σημαντικές πληροφορίες που ίσως να μην είναι ευδιάκριτες με μια πρώτη ματιά. Οι πιο σημαντικές στατιστικές πράξεις που μπορούν να πραγματοποιηθούν σε ένα σύνολο δεδομένων για να μας δείξουν περισσότερες πληροφορίες για κάθε γνώρισμα καταγεγραμμένο σε αυτό, αποτελούν τα «*μέτρα θέσης*» (measures of tendency) και τα «*μέτρα διασποράς*» (measures of dispersion). Πιο συγκεκριμένα, η μέση τιμή (mean), η επικρατούσα τιμή (mode), η διάμεσος (median), το εύρος τιμών (range), η διακύμανση (variance), η και η τυπική απόκλιση (standard deviation), αποτελούν τα πιο σημαντικά μέτρα από αυτά (Han, Kamber και Pei ,2012).

1.2 Σύνολα δεδομένων

Ως ένα *σύνολο δεδομένων* (data set) ορίζεται μια συλλογή αντικειμένων $(x_1, x_2 \dots x_n)$ τα οποία ορίζονται με σαφήνεια και διακρίνονται σαφώς το ένα από το άλλο (Γναρδέλλης, 2003). Ένα τέτοιο σύνολο από δεδομένα απαρτίζεται από διάφορα στοιχεία-μέλη, τα οποία μπορούν να ονομαστούν και ως στοιχεία ή δεδομένα του συνόλου, ενώ οι τύποι που εμπεριέχονται σε αυτό μπορούν να είναι είτε ποσοτικοί είτε ποιοτικοί. Ένας βασικός κανόνας που αφορά τόσο την επεξεργασία όσο και την προ-επεξεργασία των συνόλων δεδομένων είναι ότι, κάθε σύνολο πρέπει να μελετάται με έναν σαφώς καθορισμένο τύπο δεδομένων (ή ποιοτικούς ή ποσοτικούς). Παραδειγματικά, σε περίπτωση ύπαρξης κατηγορηματικών μεταβλητών μέσα σε ένα σύνολο, μπορεί να γίνει μετατροπή τους σε αριθμητικές μεταβλητές, καθιστώντας έτσι την επεξεργασία τους εφικτή ή και πιο αποδοτική.

Η αναπαράσταση ενός συνόλου δεδομένων μπορεί να πραγματοποιηθεί σε πίνακες οι οποίοι απαρτίζονται από γραμμές και από στήλες. Κάθε γραμμή ενός τέτοιου πίνακα αναπαριστά ένα στοιχείο από το σύνολο δεδομένων και, περιέχει επίσης διάφορα γνωρίσματα σχετικά με αυτό το στοιχείο. Σύμφωνα με τους Ρόιγκερ και Γκίατζ (2003) κάθε εγγραφή στοιχείου σε έναν τέτοιο πίνακα λέγεται και «*στιγμιότυπο*» (instance), ενώ κάθε στήλη του πίνακα αντιστοιχεί και σε μια μεταβλητή που περιγράφει -ποσοτικά ή ποιοτικά- κάθε εγγραφή. Μια ονομασία που χρησιμοποιείται αρκετά συχνά για τις στήλες στα σύνολα δεδομένων είναι το «*γνώρισμα*» (attribute). Το πλήθος των γνωρισμάτων ενός συνόλου δεδομένων ορίζεται και ως η διάσταση, ή αλλιώς το μέγεθος του συνόλου αυτού.

Πολλοί επιστήμονες έπειτα από μελέτες και έρευνες, καταλήγουν να έχουν μεγάλα ποσά καταγεγραμμένων στοιχείων σε σύνολα δεδομένων. Η χρήση των κατάλληλων τεχνικών επεξεργασίας στο σύνολο αυτό μπορεί να αποδειχθεί ιδιαιτέρως κερδοφόρα, καθώς σημαντικές πληροφορίες μπορεί να εξαχθούν. Δύο συγχεόμενες επιστημονικές περιοχές που ασχολούνται με το συγκεκριμένο τομέα, της αναζήτησης γνώσης σε δεδομένα δηλαδή, είναι η μηχανική μάθηση και η εξόρυξη πληροφορίας. Για να είμαστε περισσότερο συγκεκριμένοι, η μηχανική μάθηση χρησιμοποιεί αλγόριθμους εξόρυξης πληροφορίας σε σύνολα δεδομένων, με απώτερο σκοπό να καταστίσει ένα υπολογιστικό συστήματα ικανό να βελτιώσει διάφορες εργασίες του, ή ακόμα και να κάνει ευφυείς προβλέψεις σχετικά με διάφορα γεγονότα. Αντίθετα, η εξόρυξη πληροφορίας χρησιμοποιεί διάφορες αλγοριθμικές τεχνικές για την ανακάλυψη σημαντικών πληροφοριών μέσα σε βάσεις και σύνολα δεδομένων. Οι δύο αυτές επιστημονικές περιοχές αποτελούν πραγματικά δύο πολύ σημαντικούς τομείς για την επεξεργασία των δεδομένων και, θα μελετηθούν εκτενέστερα στο δεύτερο και στο τρίτο κεφάλαιο της παρούσας πτυχιακής.

1.3 Η προ-επεξεργασία των δεδομένων

Για την προσθήκη «αξίας» στα δεδομένα αλλά και για την απόκτηση μιας σημαντικής πληροφορίας και γνώσης από αυτά, πρέπει πρώτα να πραγματοποιηθεί μια σειρά διαδικασιών όπως μαθηματικών και στατιστικών αναλύσεων, δομών, οργάνωσης και διόρθωσης λαθών. Αυτές οι διαδικασίες είναι επίσης γνωστές και ως επεξεργασία των δεδομένων και αποσκοπούν στην απόκτηση μιας πιο δομημένης μορφής δεδομένου, που θα μπορέσει να έχει κάποια συμβολή σε μια περαιτέρω σημασιολογία, επεξήγηση ή και διεργασία.

Για να φτάσουμε στο σημείο της επεξεργασίας των δεδομένων με μεθόδους μηχανικής μάθησης, αλλά και για την υλοποίηση οποιουδήποτε αλγορίθμου ή στατιστικής τεχνικής πάνω σε αυτά, αξίζει να σημειωθεί ότι πρέπει να πραγματοποιηθεί μια προ-επεξεργασία και ένας «καθορισμός» του συνόλου των δεδομένων νωρίτερα. Μερικά από τα συνήθη προβλήματα στα μεγάλα σύνολα δεδομένων αποτελούν οι ελλιπείς τιμές, τα ημιτελή δεδομένα ή και οι διπλότυπες εγγραφές. Ως αποτέλεσμα, η επεξεργασία των δεδομένων μας μπορεί να μην είναι έγκυρη και αξιόπιστη.

Σύμφωνα με τον Καραντζιά (2019:11), η *προ-επεξεργασία* των δεδομένων είναι το πιο κρίσιμο αλλά και το πιο χρονοβόρο στάδιο πριν την εφαρμογή τεχνικών για την επεξεργασία των δεδομένων, όπως η μηχανική μάθηση και η εξόρυξη πληροφορίας. Κύριος στόχος της προ-επεξεργασίας αποτελεί η δημιουργία μιας καλύτερης γνωστικής «εικόνας» γύρω από αυτά, καθώς και η μορφοποίηση τους σε μια πιο «επεξεργάσιμη» μορφή. Στη συνέχεια, θα γίνει μια λεπτομερής αναφορά σε τρεις από τις πιο κύριες τεχνικές προ-επεξεργασίας δεδομένων που χρησιμοποιούνται: α) τον *καθορισμό* των δεδομένων (data cleaning), β) τη *μεταμόρφωση* των δεδομένων (data transformation) και γ) τη *μείωση των διαστάσεων* (data reduction).

1.3.1 Ο καθορισμός των δεδομένων

Σύμφωνα με τους Erhard και Hong, ο *καθορισμός*, αποτελεί ίσως το πιο σημαντικό στοιχείο κατά την προ-επεξεργασία των δεδομένων και αφορά τον έλεγχο για την ύπαρξη «θορύβου» στα δεδομένα (noisy values), την ύπαρξη ελλιπών τιμών (missing values) κλπ. Γενικότερα, η συγκεκριμένη μεθοδολογία χρησιμοποιείται για την «καλύτερη δόμηση» ενός συνόλου δεδομένων, ώστε να μπορέσει να καταστεί πιο αποτελεσματική η εφαρμογή αλγορίθμων επεξεργασίας σε αυτό.

- **Ελλιπείς τιμές**

Οι *ελλιπείς τιμές* πρόκειται για «κενά» σημεία στα σύνολα των δεδομένων μας στα οποία δεν έχει γίνει κάποια εγγραφή μιας τιμής. Αυτό μπορεί να συμβαίνει είτε γιατί αυτές οι εγγραφές έχουν αλλοιωθεί ή χαθεί μέσω της αποθήκευσης στον υπολογιστή από λάθος, είτε δεν ήταν συμπληρωμένα εξ αρχής. Μερικές από τις τεχνικές που προτείνουν οι Han, Kamber και Pei (2012) ακολουθούν:

- Αγνόηση ή διαγραφή αυτής της εγγραφής.
- Η χειρωνακτική συμπλήρωση των ελλιπών μερών με τυχαίες τιμές, πράγμα το οποίο καθίσταται ιδιαίτερα δύσκολο σε δεδομένα με πολλές εγγραφές στιγμιότυπων.
- Οι ελλιπείς τιμές για κάθε στιγμιότυπο να συμπληρώνονται με τον μέσο όρο, την διάμεσο ή την επικρατούσα τιμή που ισχύει για την εκάστοτε μεταβλητή.

- Να γίνεται χρήση μιας στατιστικής τεχνικής, όπως για παράδειγμα η γραμμική παλινδρόμηση, για την συμπλήρωση της κενής τιμής με μια τιμή που προβλέφθηκε από το στατιστικό μοντέλο.

- **Θόρυβος στα δεδομένα**

Σύμφωνα με τους Ρόιγκερ και Γκίατζ (2008), ως «*θόρυβος*» σε ένα σύνολο δεδομένων ορίζεται μια τιμή που παίρνει μια μεταβλητή ενός στιγμιότυπου, η οποία απέχει σε μεγάλα ποσοστά από το συνηθισμένο εύρος τιμών άλλων τέτοιων τιμών της ίδιας στήλης. Ένα πράγμα που χαρακτηρίζει τα δεδομένα με θόρυβο αποτελεί η δυσκολία ερμηνείας των τιμών τους συγκριτικά με τις τιμές άλλων παρόμοιων εγγραφών. Η ύπαρξη θορύβου και άλλων τέτοιων «ακραίων» καταγραφών, μπορεί να συμβεί τυχαία είτε να είναι αποτέλεσμα λανθασμένων μετρήσεων ή προγραμματιστικού λάθους. Σε αυτήν την περίπτωση, για την εξομάλυνση των δεδομένων μας (data smoothing), μερικές από τις τεχνικές που μπορούν να πραγματοποιηθούν είναι οι ακόλουθες:

  - Η αντικατάσταση των εγγραφών που περιέχουν θόρυβο με την τιμή της διαμέσου, του μέσου όρου, ή της επικρατούσας τιμής που ισχύει για την κάθε μεταβλητή.
  - Η χρήση της τεχνικής της «*ανάλυσης αποκλεισμού*» (outliner analysis), όπου τα στιγμιότυπα χωρίζονται σε συστάδες βάσει ενός χαρακτηριστικού που τα διαχωρίζει καλύτερα από τα άλλα. Τα στιγμιότυπα τα οποία περιέχουν «θόρυβο» δεν έχουν αντιστοιχισθεί σε κάποια συστάδα και συνεπώς διαγράφονται από το σύνολο των δεδομένων.

1.3.2 Η μεταμόρφωση των δεδομένων

Σύμφωνα με τον Osborne (2002), η *μεταμόρφωση* ή αλλιώς *μετασχηματισμός*, αποτελεί ένα σημείο της προ-επεξεργασίας των δεδομένων, κατά το οποίο, στα δεδομένα εφαρμόζονται διάφορες μαθηματικές πράξεις και τροποποιήσεις, με σκοπό την εξαγωγή ενός καλύτερου αποτελέσματος μετά την επεξεργασία τους. Μεταξύ άλλων, μερικές από τις μεθόδους που απαρτίζουν τη μεταμόρφωση των δεδομένων, αποτελούν η κανονικοποίηση (data normalization), η διακριτοποίηση (data discretization) και η εξομάλυνση των δεδομένων (data smoothing).

- **Κανονικοποίηση**

Η *κανονικοποίηση* (regularization), είναι μια διαδικασία η οποία μετασχηματίζει τις τιμές των αριθμητικών δεδομένων των μεταβλητών ενός συνόλου, σε αριθμητικές τιμές δεδομένων «ίδιας κλίμακας». Η καινούρια αναπαράσταση των δεδομένων που θα δημιουργηθεί, αποτελεί ιδιαίτερη χρησιμότητα, καθώς δε θα χρειαστούμε ούτε εμείς αλλά ούτε και οι αλγόριθμοι της επεξεργασίας των δεδομένων, να διαχειριστούν μεγάλες αριθμητικές τιμές. Η Min-Max, αποτελεί μια μέθοδο κανονικοποίησης, η οποία εκτελεί μια γραμμική μεταμόρφωση για όλες τις τιμές μιας μεταβλητής του συνόλου δεδομένων, με τον εξής τρόπο:

1. Θέτει ως min_a και max_a τη μικρότερη και τη μεγαλύτερη τιμή από μια μεταβλητή, αντίστοιχα.
2. Θέτει ως min'_α και max'_α τη μικρότερη και τη μεγαλύτερη τιμή από το καινούριο διάστημα που θα αναπαρασταθούν οι αριθμοί, για παράδειγμα το 0 και το 1.
3. Τέλος, για κάθε στιγμιότυπο, η αντίστοιχη τιμή της μεταβλητής του (x_i) θα μεταμορφωθεί ως ένα νέο δεδομένο με τη χρήση του εξής τύπου[2]:

$$x'_i = \frac{x_i - min_a}{max_a - min_a} \cdot (max'_\alpha - min'_\alpha) + min'_\alpha$$

Όπου x_i είναι η αρχική τιμή μιας μεταβλητής ενός στιγμιότυπου και x'_i είναι η καινούργια μεταμορφωμένη τιμή που θα του προκύψει.

- **Διακριτοποίηση**

Κατά τη διαδικασία της *διακριτοποίησης*, οι τιμές μια αριθμητικής μεταβλητής κάθε στιγμιότυπου σε ένα σύνολο δεδομένων, αντικαθίστανται ως ένα διαστήματα, ή ως μια αλφαριθμητική τιμή ή και αντίστροφα. Τα διαστήματα αυτά, μπορεί να αποτελέσουν μια ιεραρχία εννοιών για τις αντίστοιχες πρώην αριθμητικές ή κατηγορηματικές μεταβλητές που πλέον αναπαριστούν.

---

[2] Han, Kamber και Pei (2012:114)

- **Εξομάλυνση των δεδομένων**

Η *εξομάλυνση* των δεδομένων, όπως αναφέρθηκε και στον καθορισμό των δεδομένων, πρόκειται για μια διαδικασία που εστιάζει στην εξάλειψη του θορύβου και των περιπτώσεων που κάποιες καταγεγραμμένες τιμές μεταβλητών, απέχουν πολύ από τις άλλες της «τάξης» τους. Ειδικότερα, για την επίτευξη αυτού του σκοπού πραγματοποιούνται διάφορες μέθοδοι όπως η αντικατάσταση αυτών των τιμών με τους αντίστοιχους μέσους όρους, τις διάμεσους κ.λπ.

1.3.3 Μείωση των διαστάσεων

Οι τεχνικές που απαρτίζουν τη «*μείωση των διαστάσεων των δεδομένων*» (data reduction), αποσκοπούν στη δημιουργία ενός μικρότερου σε όγκο συνόλου δεδομένων, το οποίο θα εξακολουθεί να περιέχει τα ίδια χρήσιμα στοιχεία με την αρχική του μορφή. Έτσι, η εφαρμογή οποιασδήποτε τύπου επεξεργασίας από αλγορίθμους, μπορεί να καθίσταται πιο αποτελεσματική (Han, Kamber και Pei, 2012).

Μία πολύ διαδεδομένη τεχνική που επιτελεί αυτήν τη διεργασία της μείωσης των διαστάσεων ονομάζεται «*ανάλυση των κύριων συνιστωσών*» (principal component analysis – PCA). Αν θεωρηθεί ότι ένα σύνολο δεδομένων έχει n γραμμές και m στήλες, ο αλγόριθμος PCA προσπαθεί να βρει k συνιστώσες οι οποίες θα μπορέσουν να περιγράψουν καλύτερα τις m στήλες του συνόλου με $k \leq m$. Συνεπώς, το αρχικό σύνολο δεδομένων ουσιαστικά συμπιέζεται σε ένα μικρότερο χώρο, καταλήγοντας σε μια μείωση των διαστάσεών του, ξεφορτώνοντας έτσι δεδομένα που πιθανώς να μην είναι χρήσιμα. Τα δεδομένα που έχουν πλέον δημιουργηθεί στο καινούριο σύστημα k, δημιουργούν γραμμικούς συνδυασμούς των αρχικών δεδομένων, οι οποίοι περιέχουν το μεγαλύτερο μέρος της διακύμανσής τους. Σύμφωνα με τους Brunton και Kutz (2017), η διαδικασία που ακολουθεί ο PCA αλγόριθμος έχει ως εξής:

1. Υπολογίζεται ο μέσος όρος όλων των μεταβλητών σε ένα σύνολο δεδομένων.
2. Ο μέσος όρος που υπολογίσθηκε αφαιρείται από όλα τα στιγμιότυπα κάθε στήλης που αντιστοιχεί, δημιουργώντας έτσι ένα νέο σύνολο από δεδομένα.
3. Πραγματοποιείται υπολογισμός του πίνακα συνδιασποράς (covariance matrix) βάσει του νέου συνόλου.

4. Υπολογίζονται οι «κύριες συνιστώσες» του πίνακα συνδιασποράς, ως γραμμικοί συνδυασμοί του αρχικού συνόλου.
5. Οι κύριες συνιστώσες ταξινομούνται βάσει του βαθμού σημαντικότητάς τους.
6. Γίνεται αναπαράσταση των δεδομένων σε μια νέα κλίμακα, σύμφωνα με τις πιο σημαντικές συνιστώσες που κατασκευάστηκαν.

2 Μηχανική Μάθηση

Οι άνθρωποι, από τα τέλη της δεκαετίας του 1940 οραματίζονταν την τέλεια μηχανή που θα υποδείκνυε μεγάλα ποσοστά ευφυΐας σαν και των ανθρώπων και, θα μπορούσε να επεξεργαστεί διάφορα δεδομένα όπως ακριβώς και ο ανθρώπινος εγκέφαλος. Βέβαια τότε, η υπολογιστική ισχύς των μηχανημάτων που υπήρχαν, αλλά και τα μεγάλα ποσά χρηματοδοτήσεων που έπρεπε να δαπανηθούν σε έρευνες, δεν επέτρεψαν ποτέ μια τέτοια κατασκευή. Σήμερα όμως, οι υπολογιστές έχουν εξελιχθεί τόσο πολύ που επιτρέπουν πολλούς, αν όχι όλους, από τους οραματισμούς των τότε πατέρων στον χώρο της τεχνητής νοημοσύνης.

Ζούμε σε μια εποχή, όπου τα δεδομένα που μας περιτριγυρίζουν είναι άφθονα, και ως επι των πλείστον βρίσκονται σε ψηφιακή μορφή. Η αξιοποίηση αυτών, μέσω διαφόρων αλγορίθμων και τεχνικών επεξεργασίας, μπορούν να μας οδηγήσουν σε πολύ σημαντικές πληροφορήσεις, και συνεπώς στη γνώση. Η «*μηχανική μάθηση*» (machine learning), πρόκειται για έναν τέτοιον επιστημονικό χώρο, που συμπεριλαμβάνει πολλά στοιχεία από γνωστικά πεδία όπως τη στατιστική, τα μαθηματικά, τη πληροφορική και ιδίως τη τεχνητή νοημοσύνη. Αν μέσω του όρου της τεχνητής νοημοσύνης αναφερόμαστε σε ένα πολύ γενικό πεδίο που αναπαριστά την ευφυία που μπορεί να αναδείξει ένας υπολογιστής, τότε η μηχανική μάθηση μπορεί να οριστεί ως ο «τρόπος» με τον οποίο αυτός ο υπολογιστής θα «αναπτύξει» την εκάστοτε ευφυία και, το πως θα ξεχωρίζει τις διάφορες καταστάσεις που τον περιβάλλουν. Κατά το πέρασμα τον χρόνων έχουν δοθεί πολλοί ορισμοί για το πεδίο της μηχανικής μάθησης. Στην συνέχεια, αναφέρονται μερικοί από αυτούς:

Ο πρώτος που θα γίνει αναφορά, δόθηκε από τον Άρθουρ Σάμιουελ το 1959, όπου όρισε την μηχανική μάθηση ως:

«*το πεδίο μελέτης που δίνει στα υπολογιστικά συστήματα την δυνατότητα να μαθαίνουν και να βελτιώνουν τις αλγοριθμικές τους διαδικασίες, χωρίς απαραίτητα αυτές να έχουν προγραμματιστεί.*»

Ένας δεύτερος ορισμός, αυτήν την φορά από τον Νοέλ Κάρμπονελ το 1987, ανέφερε την μηχανική μάθηση ως:

«τη μελέτη υπολογιστικών μεθόδων για την απόκτηση νέας γνώσης, νέων δεξιοτήτων και νέων τρόπων οργάνωσης της υπάρχουσας γνώσης.»

Ο τρίτος ορισμός που θα αναφερθεί, δόθηκε από τον Τομ Μίτσελ το 1997, περιγράφοντας τη μηχανική μάθηση ως ακολούθως:

«είναι όταν ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από την εμπειρία Ε, σε σχέση με κάποια τάξη εργασιών Τ και μέτρηση απόδοσης Ρ εάν η απόδοσή του σε εργασίες στο Τ, όπως μετράτε από το Ρ, βελτιώνεται με την εμπειρία Ε.»

Με άλλα λόγια, η μηχανική μάθηση μπορεί να οριστεί επίσης ως ένα σύνολο υπολογιστικών μεθόδων, αλγορίθμων και διαδικασιών, όπου επιτρέπουν σε συστήματα υπολογιστών να βελτιώνουν την απόδοσή τους σε μια εργασία ή και να κάνουν προβλέψεις σχετικά με μελλοντικές καταστάσεις (Mohri, Rostamizadeh και Talwalkar, 2018).

Ως πεδίο της τεχνητής νοημοσύνης, η μηχανική μάθηση μελετά την κατασκευή αλγορίθμων που μπορούν να «μαθαίνουν» από πειραματικές πληροφορίες και μοτίβα δεδομένων και να κάνουν διάφορες ευφυείς προβλέψεις σχετικά με αυτά[3]. Για να είμαστε πιο συγκεκριμένοι, όλοι οι αλγόριθμοι της μηχανικής μάθησης δέχονται αυτές τις «πειραματικές πληροφορίες» ως ένα σύνολο δεδομένων, ή αλλιώς χαρακτηριστικών εισόδου (input data), στα οποία θα ζητηθεί να γίνει επεξεργασία. Οι πιθανότητες, η στατιστική, η εξόρυξη πληροφορίας και η αριθμητική βελτιστοποίηση, είναι μερικοί τομείς που συγχέονται αρκετά συχνά με την μηχανική μάθηση, αφού όλες, λίγο ή πολύ, επιτελούν την ίδια διαδικασία, της πρόβλεψης διαφόρων καταστάσεων μέσω κάποιων υπαρχόντων δεδομένων. Η μηχανική μάθηση, εφαρμόζεται σε διάφορες εργασίες υπολογιστών, όπου τόσο ο σχεδιασμός όσο και ο προγραμματισμός αλγορίθμων καθίστανται σχεδόν ανέφικτοι. Για παράδειγμα, ένας αλγόριθμος από μόνος του, δεν θα μπορέσει ποτέ να είναι σε θέση να αναγνωρίσει ποια μηνύματα αποτελούν «σπαμ» σε ένα ηλεκτρονικό ταχυδρομείο. Αντ' αυτού, αν του δοθούν κάποια εμπειρικά δεδομένα και αναλυθούν οι συσχετίσεις που έχουν τα σπαμ[4] μηνύματα σε αντιπαράθεση με τα αποδεκτά, θα είναι πιο εφικτό να γίνουν οι κατάλληλες προβλέψεις και να «φιλτραριστούν» ώστε να μην παρουσιαστούν ξανά στον χρήστη. Συνεπώς, η εκτενής παρουσίαση διαφόρων δεδομένων, ή αλλιώς «εκπαίδευση» των αλγορίθμων όπως

---

[3] Han, Kamber και Pei (2012)

[4] Ως σπαμ, ορίζονται ένας τύπος μηνυμάτων που μπορεί να δεχθεί ο εκάστοτε άνθρωπος σε διαδικτυακές πλατφόρμες, ή και μηνύματα στο κινητό του τηλέφωνο, τα οποία στέλνονται με απώτερο σκοπό την διαφήμιση προϊόντων, υπηρεσιών κλπ., ή ακόμα και την υποκλοπή προσωπικών στοιχείων.

ονομάζεται, σε συνδυασμό με διάφορες άλλες μεθόδους, είναι αυτά που καθορίζουν την εξαγωγή του αποτελέσματος σε ένα σύστημα μηχανικής μάθησης.

Η συνέχεια του συγκεκριμένου κεφαλαίου αποσκοπεί σε μια πιο λεπτομερής ανάλυση της μηχανικής μάθησης και τις λειτουργίες που επιτελούνται κατά αυτήν. Επιπλέον, θα γίνει μια παρουσίαση των διαφορετικών κατηγοριών κατά των οποίων ένας αλγόριθμος μηχανικής μάθησης μπορεί να μάθει να κάνει διακρίσεις σε στοιχεία μέσα σε σύνολα δεδομένων και, να καταλήγει σε συμπεράσματα σχετικά με αυτά.

2.1 Οι κατηγορίες της μηχανικής μάθησης

Όπως ακριβώς σε διάφορους άλλους επιστημονικούς τομείς αλλά και ακόμα και στην ίδια την ζωή, μπορούν να γίνουν διάφορες διακρίσεις ανάλογα με την λειτουργικότητα και τα πεδία εφαρμογής διαδικασιών, έτσι και η μηχανική μάθηση μπορεί να διαχωριστεί σε διάφορες κατηγορίες. Οι Ρόιγκερ και Γκίατζ (2008) αναφέρουν χαρακτηριστικά ότι όλες αυτές οι κατηγορίες *«χρησιμοποιούν την επαγωγική μάθηση (induction-based learning), τη διαδικασία δηλαδή του σχηματισμού εννοιών με την παρατήρηση συγκεκριμένων παραδειγμάτων τους που πρέπει να μαθευτούν».* Εντούτοις, κάθε μια από αυτές τις κατηγορίες έχει μια διαφορετική προσέγγιση στην εκπαίδευση του εκάστοτε αλγορίθμου. Πριν γίνει αναφορά όμως σε αυτές, αξίζει να σημειωθεί ότι σημαντικό ρόλο στην εκπαίδευση ενός αλγορίθμου για να κάνει τις εκάστοτε διακρίσεις, παίζει ρόλο και η δομή του συνόλου των δεδομένων που θα παρουσιαστούν ως είσοδοι στο σύστημα για επιπλέον επεξεργασία. Τα δεδομένα αυτά, μπορούν να διακριθούν σε:

- Ετικετοποιημένα δεδομένα (labeled data) και
- Μη ετικετοποιημένα δεδομένα (unlabeled data).

Τα *ετικετοποιημένα* δεδομένα, πρόκειται για τα στιγμιότυπα σε ένα σύνολο τα οποία περιέχουν μια ξεχωριστή «ειδική» μεταβλητή, η οποία ονομάζεται κλάση και τα αντιπροσωπεύει. Μια άλλη γνωστή ονομασία της κλάσης, είναι το «χαρακτηριστικό εξόδου» (output data). Παραδειγματικά, σε ένα σύνολο δεδομένων χρεοκοπημένων χωρών, η επιθυμητή κλάση θα μπορούσε να ήταν θετική για το αν η χώρα είναι όντως χρεοκοπημένη, ή σε διαφορετική περίπτωση θα μπορούσε να ήταν αρνητική.

Αντίθετα, ως *μη ετικετοποιημένα* δεδομένα ορίζονται τα στιγμιότυπα ενός συνόλου δεδομένων στα οποία δεν αντιστοιχεί κάποια κλάση ή γενικά κάποιο αποτέλεσμα.

Οι αλγόριθμοι της μηχανικής μάθησης ποικίλουν αναλόγως με το είδος αυτών των δεδομένων που επεξεργάζονται. Επειδή επίσης, για κάθε ένα πρόβλημα που καλείται να λύσει το πεδίο της μηχανικής μάθησης υπάρχει και μια διαφορετική προσέγγιση λύσης, οι τύποι της μηχανικής μάθησης ταξινομούνται σε διάφορες κατηγορίες[5]. Σύμφωνα με τους Han, Kamber και Pei (2012) οι πιο σημαντικές από αυτές αποτελούν:

- Η Επιβλεπόμενη μάθηση (supervised learning),
- Η Μη-Επιβλεπόμενη μάθηση (unsupervised learning),
- Η Ημι-Επιβλεπόμενη μάθηση (semi-supervised learning) και η
- Ενεργή μάθηση (active learning).

Μια από τις κύριες διαφορές στις κατηγορίες αυτές είναι ότι τεχνικές της επιβλεπόμενης μάθησης κάνουν χρήση ετικετοποιημένων δεδομένων, αντίθετα με τις τεχνικές της μη-επιβλεπόμενης μάθησης οι οποίες κάνουν χρήση μη-ετικετοποιημένων δεδομένων. Από την άλλη πλευρά, η ημι-επιβλεπόμενη και η ενεργή μάθηση, οι οποίες θεωρούνται και ως «παρακλάδια» της επιβλεπόμενης μάθησης, κάνουν έναν συνδυασμό χρήσης ετικετοποιημένων και μη δεδομένων. Φυσικά, υπάρχουν και άλλες διακρίσεις που μπορούν να γίνουν σε αυτές τις κατηγορίες οι οποίες θα μελετηθούν καλύτερα παρακάτω.

Κάπου εδώ αξίζει να σημειωθεί ότι, αν και στο εμπειρικό μέρος της πτυχιακής εργασίας γίνεται πρακτική χρήση μόνο της επιβλεπόμενης μάθησης, κρίθηκε απαραίτητη η συνοπτική αναφορά και στις υπόλοιπες αυτές κατηγορίες, ως συμπλήρωση μιας «πλήρους γνωστικής εικόνας» και βιβλιογραφικού περιεχομένου γύρω από το μεγάλο πεδίο αυτό της μηχανικής μάθησης.

Στον «Πίνακα 1» που ακολουθεί, γίνεται μια συνοπτική σύγκριση μεταξύ των ειδών αυτών της μηχανικής μάθησης.

---

[5] Οι περισσότερες πηγές αναφέρουν ως σημαντικότερες κατηγορίες της μηχανικής μάθησης την επιβλεπόμενη, την μη-επιβλεπόμενη, την ημι-επιβλεπόμενη και την ενεργή μάθηση. Εντούτοις, υπάρχουν δύο ακόμα κατηγορίες, η ενισχυτική και η βαθιά μάθηση, στις οποίες θα γίνει αναφορά σε έπειτα σημείο της εργασίας.

Πίνακας 1: Σύγκριση των κατηγοριών της Μηχανικής Μάθησης.

Συγκριτική αναπαράσταση των κατηγοριών της Μηχανικής Μάθησης.			
Επιβλεπόμενη Μάθηση	**Μη-Επιβλεπόμενη Μάθηση**	**Ημι-Επιβλεπόμενη Μάθηση**	**Ενεργή Μάθηση**
-Χρήση ετικετοποιημένων δεδομένων.	-Χρήση μη ετικετοποιημένων δεδομένων.	-Χρήση ετικετοποιημένων και μη δεδομένων.	-Χρήση ετικετοποιημένων και μη δεδομένων.
-Χρήση ενός μεγάλου μέρους του αρχικού συνόλου δεδομένων για εκπαίδευση ενός μοντέλου.	-Χρήση ολόκληρου του αρχικού συνόλου των δεδομένων.	- Χρήση ενός μικρού μέρους του αρχικού συνόλου δεδομένων για εκπαίδευση ενός μοντέλου..	- Χρήση ενός μικρού μέρους του αρχικού συνόλου δεδομένων για εκπαίδευση ενός μοντέλου.
-Χρησιμοποιείται για προβλέψεις.	-Χρησιμοποιείται για ανάλυση και περιγραφή.	-Χρησιμοποιείται για προβλέψεις.	-Χρησιμοποιείται για προβλέψεις.

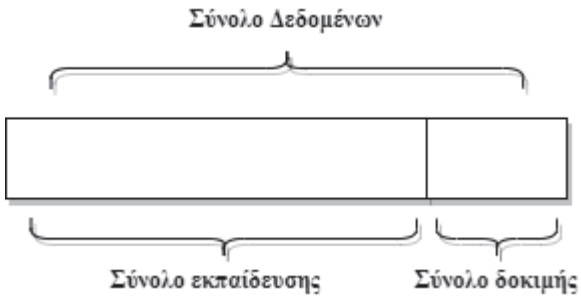
2.2 Επιβλεπόμενη μάθηση

Βάσει των Ρόιγκερ και Γκίατζ (2008), ο κάθε άνθρωπος από μικρή ηλικία χρησιμοποιεί τον όρο της επαγωγής για να δημιουργήσει ορισμούς βασικών καταστάσεων που τον περιτριγυρίζουν, δημιουργώντας τα δικά του μοντέλα κατηγοριοποίησης. Στη συνέχεια, τα μοντέλα αυτά χρησιμοποιούμαι για να μας βοηθήσουν να αναγνωρίσουμε διάφορες καταστάσεις που θα συναντήσουμε κατά την διάρκεια της ζωή μας. Έτσι λοιπόν γίνεται και στην «*επιβλεπόμενη μάθηση*» (supervised learning). Σε αυτήν την κατηγορία μάθησης, δημιουργούνται μοντέλα αλγορίθμων από εμπειρικά σύνολα δεδομένων που δίνονται στο σύστημα.

Πιο συγκεκριμένα, το σύστημα στην περίπτωση της επιβλεπόμενης μάθησης δέχεται ένα σύνολο από n παραδειγματικά δεδομένα της μορφής $([x_1, y_1], [x_2, y_2] \dots [x_n, y_n])$ με το αποτέλεσμα-κλάση που τους αντιστοιχεί (y). Στην συνέχεια, μέσω διαφόρων τεχνικών, οι

οποίες θα μελετηθούν σε μετέπειτα στάδιο της εργασίας, ο αλγόριθμος ψάχνει να βρει μοτίβα στα δεδομένα με σκοπό να μάθει να τα αντιστοιχίζει όσο πιο κατάλληλα γίνεται με το αποτέλεσμά τους. Ως σκοπός, το μοντέλο που θα κατασκευαστεί θα πρέπει να είναι σε θέση, βάσει της «εμπειρίας» του, να αναγνωρίζει τις κλάσεις των μη ετικετοποιημένων στιγμιότυπων που μπορεί να του παρουσιασθούν.

Από το σύνολο των δεδομένων που δίνεται σε ένα σύστημα επιβλεπόμενης μάθησης, επιλέγεται ένα ποσοστό το οποίο χρησιμοποιείται για την «εκπαίδευση» ενός αλγορίθμου. Αυτό το ποσοστό που δίνεται για εκπαίδευση στον αλγόριθμο, είναι γνωστό και ως σύνολο εκπαίδευσης (training set)[6]. Το σύνολο αυτό δίνεται για λεπτομερής εξέταση ως εμπειρικά δεδομένα στον αλγόριθμο, με απώτερο σκοπό την κατάλληλη ρύθμιση διαφόρων παραμέτρων, ώστε η κλάση-ετικέτα του κάθε στιγμιότυπου να μάθει να ξεχωρίζεται όσο καλύτερα γίνεται. Το υπόλοιπο ποσοστό του συνόλου δεδομένων, το λεγόμενο σύνολο δοκιμής (test set), χρησιμοποιείται από τον αλγόριθμο για την αξιολόγηση των προβλέψεων του για τις ετικέτες των στιγμιότυπων του. Αν η αξιολόγηση των προβλέψεων αυτών είναι επιθυμητές βάσει του χρήστη του προγράμματος, το μοντέλο αυτό θα μπορέσει να χρησιμοποιηθεί για πρόβλεψη σε ένα διαφορετικό σύνολο δεδομένων, παρόμοιων μεταβλητών και άγνωστων κλάσεων.



Εικόνα 1: Σύνολο εκπαίδευσης και σύνολο δοκιμής.

Ανάλογα με το χαρακτηριστικό εξόδου που επεξεργάζεται ένας αλγόριθμος επιβλεπόμενης μάθησης, τα προβλήματα μπορούν να διακριθούν σε προβλήματα επιλύσιμα με *αλγόριθμους κατηγοριοποίησης* ή προβλήματα επιλύσιμα με *αλγόριθμους παλινδρόμησης* (Soren, 2017). Στην συνέχεια, θα παρουσιαστούν τα χαρακτηριστικά που συντελούν την κάθε κατηγορία από αυτές, καθώς και διάφοροι αλγόριθμοι που χρησιμοποιούνται:

[6] Σύμφωνα με τον Brownlee (2020), μερικά γνωστά ποσοστά εκπαίδευσης αποτελούν το 67% και το 80% του αρχικού συνόλου που δίνεται στον αλγόριθμο.

- **Κατηγοριοποίηση**

Η *κατηγοριοποίηση* (classification), ή αλλιώς ταξινόμηση, αφορά στη διαδικασία καταμερισμού όλων των δεδομένων ενός συνόλου σε επιμέρους κλάσεις που τους αντιστοιχούν. Ένα βασικό χαρακτηριστικό που ξεχωρίζει την κατηγοριοποίηση από την παλινδρόμηση, αφορά στις διακριτές μόνο τιμές που μπορεί να πάρει η ετικέτα ενός στιγμιότυπου κατά αυτήν.

Σύμφωνα με το χαρακτηριστικό εξόδου, ή αλλιώς την κλάση, ενός στιγμιότυπου, η κατηγοριοποίηση, μπορεί να διακριθεί σε δυαδική κατηγοριοποίηση (binary classification) και κατηγοριοποίηση πολλών κλάσεων (multi-class classification). Η βασική διαφορά των κατηγοριών αυτών είναι το πλήθος των ετικετών ενός συνόλου που μπορούν να επεξεργαστούν σε κάθε μια από αυτές. Στη δυαδική κατηγοριοποίηση για παράδειγμα, γίνεται χρήση μόνο δύο κλάσεων, ενώ στην πολλαπλή κατηγοριοποίηση το χαρακτηριστικό εξόδου μπορεί να πάρει από τρείς έως και περισσότερες διαφορετικές τιμές.

Στην συνέχεια, ακολουθούν μερικοί γνωστοί αλγόριθμοι κατηγοριοποίησης:

- k-Κοντινότεροι Γείτονες (k-nearest neighbors),
- Απλοϊκός Bayes (naïve bayes),
- Αλγόριθμος δένδρων αποφάσεων (decision tree algorithms),
- Λογιστική παλινδρόμηση (logistic regression),
- Τεχνητά νευρωνικά δίκτυα (artificial neural networks) και οι
- Μηχανές διανυσμάτων υποστήριξης (support vector machines).

- **Παλινδρόμηση**

Αντίθετα, στα προβλήματα *παλινδρόμησης* (regression), τα χαρακτηριστικά εξόδου των δεδομένων μπορούν να πάρουν συνεχείς τιμές (Han, Kamber και Pei, 2012). Σκοπός των αλγορίθμων παλινδρόμησης είναι η κατασκευή ενός μοντέλου που θα μπορεί να προβλέπει συνεχείς τιμές εξόδου κάποιου δεδομένου. Αυτό θα επιτευχθεί μέσω των δεδομένων εκπαίδευσης που θα δοθούν στο σύστημα ως είσοδοι και, της εύρεσης της κατάλληλης σχέσης μεταξύ αυτών για την διατύπωση τους σε μια μαθηματική συνάρτηση συνήθως, που θα πραγματοποιεί την πρόβλεψη της τιμής μεταβλητής. Οι εφαρμογές της παλινδρόμησης ποικίλουν. Ο κάθε αλγόριθμος που μπορεί να χρησιμοποιηθεί για παλινδρόμηση κάνει χρήση και μιας διαφορετικής προσέγγισης και μεθοδολογίας για να εκτελέσει τις απαραίτητες

προβλέψεις. Μερικές από τις πιο συνηθισμένες τεχνικές που χρησιμοποιούνται -μεταξύ άλλων- για παλινδρόμηση είναι οι:

- Αλγόριθμοι δένδρων αποφάσεων (decision tree algorithms) και,
- Η γραμμική παλινδρόμηση (linear regression).

2.2.1 Τα βήματα μιας διαδικασίας επιβλεπόμενης μάθησης

Η διαδικασία της αναπαράστασης ενός προβλήματος του κανονικού κόσμου σε ένα σύστημα επιβλεπόμενης μηχανικής μάθησης, αποτελεί ένα πολύ σημαντικό ζήτημα το οποίο πρέπει κάποιος να χειριστεί με ακρίβεια για την αποφυγή τυχών λανθασμένων αποτελεσμάτων. Έτσι, υπάρχει μια συγκεκριμένη διαδικασία επιλεγμένων ενεργειών που πρέπει να πραγματοποιηθούν πριν την εφαρμογή του εκάστοτε αλγορίθμου που θα πραγματοποιήσει την κατηγοριοποίηση ή την παλινδρόμηση. Πιο συγκεκριμένα, τα βήματα που πρέπει να ακολουθηθούν για την διαδικασία της εκπαίδευσης ενός μοντέλου επιβλεπόμενης μάθησης[7] είναι τα ακόλουθα:

1) Η συλλογή δεδομένων,
2) Ο καθορισμός του προβλήματος και του γενικού στόχου της διαδικασίας,
3) Η προετοιμασία και η προ-επεξεργασία των δεδομένων,
4) Ο καθορισμός του συνόλου δεδομένων για εκπαίδευση και για δοκιμή,
5) Η επιλογή και η εκπαίδευση του κατάλληλου αλγορίθμου και,
6) Η αξιολόγηση των αποτελεσμάτων σύμφωνα με το σύνολο δοκιμής.

Τα πιο σημαντικά από αυτά τα στάδια αποτελούν η προ-επεξεργασία των δεδομένων, αλλά και η εκπαίδευση και η αξιολόγηση του κατάλληλου αλγορίθμου. Η προ-επεξεργασία κρίνεται ιδιαίτερα απαραίτητη, κυρίως σε περιπτώσεις που μπορεί να υπάρχουν ελλιπείς τιμές. Έπειτα, η χρήση της κατάλληλης αλγοριθμικής τεχνικής αλλά και η αξιολόγηση των αποτελεσμάτων στο σύνολο δοκιμής, κρίνεται επίσης ιδιαίτερα σημαντική. Αν οι προβλέψεις και οι αξιολογήσεις δεν είναι σύμφωνες με τα επιθυμητά αποτελέσματα καθορισμένα από το χρήστη, η αλγοριθμική διαδικασία μπορεί να συνεχίσει αρκετές φορές μέχρι να επιτευχθεί ένα επιθυμητό ποσοστό ορθότητας του αλγορίθμου, ρυθμίζοντας διάφορες παραμέτρους σε κάθε κύκλο εκπαίδευσης (Ρόιγκερ και Γκίατζ, 2008).

---

[7] Kotsiantis (2007)

Ο Κοτσιάντης (2007), σε περίπτωση ανάπτυξης ενός τέτοιου προγράμματος, προτείνει την συμβουλή και την καθοδήγηση από κάποιον ειδικό, ανάλογα με το σύνολο των δεδομένων που επεξεργαζόμαστε. Βάσει αυτού, μπορούμε με σιγουριά να καθορίσουμε ποια συγκεκριμένα δεδομένα και ποιες μεταβλητές τους θα παίξουν ρόλο στο τελικό αποτέλεσμα. Έτσι, καθίσταται εφικτή η αφαίρεση όλων των εγγραφών από το σύνολο των δεδομένων οι οποίες, τελικώς, δεν θα χρησιμεύσουν κάπου.

2.3 Μη-Επιβλεπόμενη μάθηση

Δεύτερος τύπος της μηχανικής μάθησης αποτελεί η «*μη επιβλεπόμενη μάθηση*» (unsupervised learning). Αντίθετα με την προηγούμενη κατηγορία, η μη επιβλεπόμενη μάθηση δημιουργεί μοντέλα από στιγμιότυπα δεδομένων χωρίς προκαθορισμένες κατηγορίες και χωρίς να γίνεται η διαχείριση κάποιου χαρακτηριστικού εξόδου (Ρόιγκερ και Γκίατζ, 2008). Σε αυτήν την τεχνική πραγματοποιείται μια ομαδοποίηση των δεδομένων βάσει ενός χαρακτηριστικού τους που τα διαχωρίζει καλύτερα. Η συγκεκριμένη προσέγγιση, χρησιμοποιείται, ως επι των πλείστον, σε μεγάλα σύνολα μη ετικετοποιημένων δεδομένων.

Σε αυτό το είδος μάθησης δίνεται στο σύστημα ένα σύνολο δεδομένων $(x_1, x_2 \ \dots \ x_n)$ από μη ετικετοποιημένα δεδομένα, όπου το κάθε ένα από αυτά, απαρτίζεται επιμέρους από διάφορες μεταβλητές και γνωρίσματα. Βάσει μιας διαδεδομένης προσέγγισης επεξεργασίας της μη-επιβλεπόμενης μάθησης, οι αλγοριθμικές διαδικασίες της κατασκευάζουν συστάδες και σχηματισμούς στα δεδομένα, χωρίζοντάς τα σε επιμέρους κατηγορίες, με την κάθε μια από αυτές να έχει και από κάποια χαρακτηρίστηκα που την ξεχωρίζει καλύτερα από τις άλλες (Ρόιγκερ και Γκίατζ, 2008). Χρησιμοποιώντας μια τέτοια μέθοδο για την ανάλυση ενός συνόλου δεδομένων, μπορούμε να εξερευνήσουμε τη δομή των στιγμιότυπων και να εξάγουμε μια σημαντική πληροφορία χωρίς την καθοδήγηση ενός απόλυτου αποτελέσματος. Βάσει αυτού, οι μη-επιβλεπόμενοι αλγόριθμοι δεν αναλύουν κάποιο συγκεκριμένο χαρακτηριστικό εξόδου του συνόλου των δεδομένων και προφανώς, δεν περιέχεται η συμβολή της ανθρώπινης παρέμβασης όπως σε άλλες περιπτώσεις μάθησης.

Τα στιγμιότυπα που θα αποτελέσουν την κάθε συστάδα σχηματίζονται έτσι ώστε να έχουν κάποιες πολύ υψηλές ομοιότητες μεταξύ τους βάσει των γνωρισμάτων-μεταβλητών που τα απαρτίζουν, αλλά και να είναι ανόμοια σε σχέση με άλλα στοιχεία των υπόλοιπων συστάδων.

Ως αποτέλεσμα, μπορούν να σχηματιστούν διάφορες κατηγορίες όμοιων στιγμιότυπων από τα οποία μπορούν να προκύψουν σημαντικοί κανόνες και πληροφορίες.

Οι τεχνικές που χρησιμοποιούνται κατά την μη-επιβλεπόμενη μάθηση διακρίνονται κατά βάση σε *τεχνικές συσταδοποίησης* και σε *τεχνικές συσχέτισης*. Βάσει της *τεχνικής της συσταδοποίησης*, χρησιμοποιούνται διάφοροι αλγόριθμοι για την δημιουργία συστάδων (clusters) σε ένα σύνολο δεδομένων, με το σκεπτικό που αναφέρθηκε νωρίτερα. Μερικοί από τους πιο γνωστούς αλγόριθμους που χρησιμοποιούνται για την τεχνική της συσταδοποίησης αποτελούν:

- Ο K-Means,
- Ο Fuzzy C-Means και,
- Ο DBSCAN.

Όσον αναφορά για στις *τεχνικές συσχέτισης*, η χρήση τους επικεντρώνεται κυρίως στην εύρεση λογικών συσχετίσεων μεταξύ των τιμών των μεταβλητών από ένα σύνολο δεδομένων. Κάποιες από τις πιο απλές και χρησιμοποιημένες τεχνικές που εφαρμόζονται για συσχέτιση, αποτελούν οι:

- Αλγόριθμοι δένδρων αποφάσεων (decision tree algorithms) και, οι
- Αλγόριθμοι κανόνων συσχετίσεων (rule algorithms).

2.4 Ημι-Επιβλεπόμενη και Ενεργή μάθηση

2.4.1 Ημι-επιβλεπόμενη μάθηση

Η «*ημι-επιβλεπόμενη μάθηση*» (semi-supervised learning) συνδυάζει, εν μέρει, την επιβλεπόμενη και την μη επιβλεπόμενη μάθηση (Xiaojin, 2005). Για την εκπαίδευση ενός μοντέλου αλγόριθμου μηχανικής μάθησης, η ημι-επιβλεπόμενη μάθηση χρησιμοποιεί τεχνικές εκπαίδευσης με δεδομένα που μπορεί να είναι είτε ετικετοποιημένα είτε όχι.

Συγκεκριμένα, στο σύστημα δίνεται ένα μικρό σύνολο από n παραδειγματικά δεδομένα $([x_1, y_1], [x_2, y_2] \dots [x_n, y_n])$ μαζί με το αποτέλεσμα-κλάση που τους αντιστοιχεί, καθώς και ένα μεγάλο σύνολο μη ετικετοποιημένων δεδομένων $(x_1, x_2 \dots x_n)$ χωρίς το αποτέλεσμα-

κλάση τους. Ο σκοπός[8] της ημι-επιβλεπόμενης μάθησης είναι να εξάγει τις κλάσεις που αντιστοιχούν στα δεδομένα του μη ετικετοποιημένου συνόλου, βάσει αλγοριθμικών τεχνικών που θα πραγματοποιηθούν στο ετικετοποιημένο σύνολο. Σύμφωνα με τον Xiaojin (2005), οι δύο κυριότερες τεχνικές[9] που χρησιμοποιούνται κατά την ημι-επιβλεπόμενη μάθηση αποτελούν α) το *Self-Training*, και β) το *Co-Training*.

Ειδικότερα:

- **Self-Training**

Κατά την τεχνική του *Self-Training*, γίνεται εκπαίδευση ενός αλγορίθμου κατηγοριοποίησης στο μικρό σύνολο των ετικετοποιημένων δεδομένων, χρησιμοποιώντας τον στην συνέχεια για την πρόβλεψη των χαρακτηριστικών εξόδου των μη ετικετοποιημένων δεδομένων. Τα δεδομένα των οποίων οι προβλέψεις που πραγματοποιήθηκαν είχαν το μεγαλύτερο ποσοστό αξιολόγησης, μεταφέρονται στο σύνολο με τα ετικετοποιημένα δεδομένα. Έτσι, η διαδικασία αυτή επαναλαμβάνεται διαρκώς, κάνοντας τον αλγόριθμο στην ουσία να «εκπαιδεύει» τον εαυτό του. Ο αλγόριθμος τερματίζει σε περίπτωση που συμπληρωθεί ένας συγκεκριμένος αριθμός επαναλήψεων από τον χρήστη, ή όταν το μέτρο της αξιολόγησης των προβλέψεων είναι πολύ χαμηλό.

- **Co-Training**

Κατά την τεχνική του *Co-Training*, κάθε ένα στιγμιότυπο από το σύνολο των ετικετοποιημένων δεδομένων χωρίζεται σε δυο μέρη, και δυο κατηγοριοποιητές χρησιμοποιούνται στο κάθε μέρος ξεχωριστά για εκπαίδευση. Στην συνέχεια, τα καλύτερα αποτελέσματα αξιολόγησης που θα έχει ένας από τους δυο κατηγοριοποιητές στο σύνολο των μη ετικετοποιημένων δεδομένων, θα προστεθούν μαζί με τις ετικέτες τους στο σύνολο εκπαίδευσης του άλλου κατηγοριοποιητή. Όπως και στην τεχνική Self-Training, η όλη αυτή αλγοριθμική διαδικασία συνεχίζεται μέχρις ότου να πραγματοποιηθεί ένας συγκεκριμένος

---

[8] Han, Kamber και Pei (2012)

[9] Η αναφορά των τεχνικών της ημι-επιβλεπόμενης μάθησης στην παρούσα εργασία περιορίζεται στις δυο κυριότερες, την Self-Training και την Co-Training. Εντούτοις, υπάρχουν και άλλες τεχνικές που χρησιμοποιούνται στην ημι-επιβλεπόμενη μάθηση, όπως για παράδειγμα η μέθοδος Graph-Based (Xiaojin, 2005).

αριθμός επαναλήψεων δοσμένος από τον χρήστη, ή μέχρις ότου το μέτρο αξιολόγησης των προβλέψεων σταματάει να βελτιώνεται.

2.4.2 Ενεργή μάθηση

Η «*ενεργή μάθηση*» (active learning), πρόκειται για μια προσέγγιση της μηχανικής μάθησης όπου αφήνεται ο χρήστης του προγράμματος της μηχανικής μάθησης να παίξει έναν σημαντικό ρόλο στην εκπαίδευση ενός μοντέλου αλγορίθμου. Όπως και στην ημι-επιβλεπόμενη μάθηση, έτσι και στην ενεργή, στο σύστημα δίνεται ένα μικρό σύνολο από n ετικετοποιημένα δεδομένα της μορφής $([x_1, y_1], [x_2, y_2] \dots [x_n, y_n])$, και ένα μεγάλο σύνολο από μη ετικετοποιημένα δεδομένα $(x_1, x_2 \ \dots \ x_n)$.

Αρχικά, γίνεται εκπαίδευση ενός αλγορίθμου κατηγοριοποίησης με βάσει του συνόλου των ετικετοποιημένων δεδομένων. Στην συνέχεια, το σύστημα θέτει ερωτήματα (queries) στον χρήστη, ζητώντας του να υποβάλλει το επιθυμητό χαρακτηριστικό εξόδου σε ένα ποσοστό του συνόλου των μη ετικετοποιημένων δεδομένων (Han, Kamber και Pei, 2012). Υπάρχουν διάφορα σενάρια κατά τα οποία ένας χρήστης του προγράμματος μπορεί να απαντήσει στις ερωτήσεις για την ετικετοποίηση και, διάφοροι τρόποι κατά τους οποίους ένα σύστημα μπορεί να επιλέξει τις ερωτήσεις που θα τεθούν. Για παράδειγμα, ο Burr (2009) αναφέρει ότι τα μη ετικετοποιημένα δεδομένα μπορούν να επιλεχθούν και να δοθούν στον χρήστη, σύμφωνα με ένα μέτρο αξιολόγησης τους (pool-based active learning). Έπειτα, ανάλογα με τα δεδομένα που επιλέχθηκαν, ο χρήστης ετικετοποιεί μόνο τα δεδομένα για τα οποία ο αλγόριθμος διαθέτει την λιγότερη βεβαιότητα για την σωστή κλάση που τους αντιστοιχεί[10].

2.5 Άλλα είδη μάθησης

Εκτός από τις τέσσερις κατηγορίες της μηχανικής μάθησης που αναφέρθηκαν, συμπληρωματικά μπορούν να προστεθούν η *ενισχυτική* (reinforcement learning) και η *βαθιά* μάθηση (deep learning). Η επιβλεπόμενη, η μη-επιβλεπόμενη, η ημι-επιβλεπόμενη και η ενεργή μάθηση, χρησιμοποιούνται αρκετά περισσότερο σε τομείς για την ανακάλυψη γνώσης

[10] Η συγκεκριμένη τεχνική αναφέρεται από τον Burr (2009) ως Uncertainty Sampling, και αποτελεί την ποιο συνηθισμένη τεχνική για την απάντηση των ερωτήσεων από τον χρήστη.

σε δεδομένα και στην πρόβλεψη διαφόρων καταστάσεων, χρησιμοποιώντας πολλά στοιχεία των μαθηματικών, της εξόρυξης πληροφορίας και της στατιστικής. Οι δυο προσεγγίσεις της ενισχυτικής και της βαθιάς μάθησης διαφέρουν από τις προηγούμενες που αναλύθηκαν, καθώς εμπίπτουν κατά αρκετά περισσότερο σε ένα πιο ανεξάρτητο πεδίο της τεχνητής νοημοσύνης και τις υπολογιστικής ευφυίας. Στην συνέχεια, θα γίνει μια συνοπτική αναφορά στις λειτουργίες που επιτελούνται σε αυτές τις δυο προσεγγίσεις μάθησης.

2.5.1 Ενισχυτική μάθηση

Ο όρος «*ενισχυτική μάθηση*» (reinforcement learning) αποτελεί έναν συγκεκριμένο τομέα της τεχνητής νοημοσύνης, ο οποίος δεν περιλαμβάνει την έκθεση κάποιου αλγορίθμου σε ένα σύνολο δεδομένων ή την κατασκευή ενός μοντέλου για προβλέψεις. Αντ' αυτού, οι προγραμματιστικές μεθοδολογίες που χρησιμοποιούνται κατά βάσει σε αυτήν τη προσέγγιση μάθησης, είναι αυτές της «*αριθμητικής βελτιστοποίησης*». Συγκεκριμένα, στο πρόγραμμα που κατασκευάζεται, ζητείται να φέρει εις πέρας διάφορες διεργασίες που θα του ζητηθούν, παίρνοντας κάποια «*ανταμοιβή*» (reward) κάθε φορά που θα επιλέγει να κάνει μια επιθυμητή ενέργεια (η οποία είναι ορισμένη από τον χρήστη φυσικά)[11]. Στόχος του προγράμματος αυτού, είναι να μεγιστοποιήσει τα rewards που λαμβάνει, μαθαίνοντας έτσι να φέρει εις πέρας τις διεργασίες που του ζητούνται όσο καλύτερα γίνεται.

2.5.2 Βαθιά μάθηση

Η «*βαθιά μάθηση*» (deep learning), πρόκειται για ένα ακόμα είδος της μηχανικής μάθησης το οποίο εμπίπτει αρκετά στον βιολογικό τρόπο σκέψης. Σύμφωνα με τους Goodfellow, Bengio και Courville (2016), η βαθιά μάθηση χρησιμοποιεί αναπαραστάσεις δεδομένων για την εκπαίδευση διάφορων αλγορίθμων, βγαλμένες από άλλες, πιο απλές αναπαραστάσεις. Για να επιτευχθεί αυτό, πραγματοποιείται μια πληθώρα τεχνικών εκπαίδευσης στους αλγορίθμους, συνδυάζοντας διάφορες προσεγγίσεις όπως η επιβλεπόμενη μάθηση, η ημι-επιβλεπόμενη μάθηση και η μη-επιβλεπόμενη μάθηση.

Για να είμαστε πιο ακριβείς, η βαθιά μάθηση πρόκειται για μια κατηγορία της οποίας ο κύριος σκοπός αποτελεί η προσομοίωση λειτουργίας και σκέψης του ανθρώπινου

[11] Sutton (1999)

εγκεφάλου. Η κύρια αλγοριθμική τεχνική που χρησιμοποιείται για την επίτευξη αυτών που προαναφέρθηκαν, ονομάζεται τεχνητά νευρωνικά δίκτυα (artificial neural networks - ANN)[12]. Αυτού του είδους η τεχνική, προσομοιάζει την λειτουργία του εγκεφάλου των βιολογικών οργανισμών, όπου κάθε νευρώνας διασπά μια πληροφορία σε απλούστερα κομμάτια και την προωθεί για επεξεργασία στους επόμενους του. Έτσι ακριβώς γίνεται και στα τεχνητά νευρωνικά δίκτυα. Η επεξεργασία της πληροφορίας ορίζεται κατά κύριο λόγο ως μια μαθητική συνάρτηση η οποία λαμβάνει χώρα σε κάθε κόμβο ενός νευρώνα, ο οποίος προωθεί το αποτέλεσμα του στο επόμενο επίπεδο του δικτύου για περεταίρω επεξεργασία.
Σε επόμενο κεφάλαιο της εργασίας θα γίνει μια πιο αναλυτική προσέγγιση στον τρόπο λειτουργίας των νευρωνικών δικτύων καθώς και στην αρχιτεκτονική αυτών.

[12] Goodfellow, Bengio και Courville (2016)

3 Εξόρυξη πληροφορίας

Εδώ και αρκετά χρόνια και ιδίως μετά την ανάπτυξη της εποχής των δικτύων των υπολογιστών στα μέσα του 1990, η συλλογή δεδομένων ανεξαρτήτως του τύπου τους αυξάνεται με εκθετικό ρυθμό κάθε μέρα. Τα δεδομένα υπάρχουν παντού πλέον. Από έρευνες, βάσεις δεδομένων νοσοκομείων, στρατιωτικών εγκαταστάσεων και επιχειρήσεων, μέχρι και δεδομένα που μπορούν να συλλεχθούν από τον παγκόσμιο ιστό (World Wide Web – WWW) και IoT δίκτυα (Διαδίκτυο των Πραγμάτων - Internet of Things). Οι αλγόριθμοι εξαγωγής πληροφοριών της μηχανικής μάθησης, πέρα από τη μεγάλη ικανότητά τους να μπορέσουν να καταστήσουν ένα σύστημα υπολογιστή ικανό να βελτιώσει τις αποδόσεις του και να μάθει να ενεργεί αναλόγως εμπειρικών γεγονότων, έχουν επίσης βρει πολλά πεδία εφαρμογής σαν αυτά που αναφέρθηκαν. Η εφαρμογή των αλγορίθμων αυτών σε τέτοιους τομείς της σημερινής κοινωνίας, υπερισχύει σε σχέση με τη διεξαγωγή στατιστικών τεχνικών και τη χρήση γλωσσών ερωτοαπαντήσεων σαν την SQL (Structured Query Language).

Η «*εξόρυξη πληροφορίας*» (data mining - DM), ή αλλιώς εξόρυξη δεδομένων, πρόκειται για έναν τομέα που αποτελείται από όλες αυτές τις αλγοριθμικές τεχνικές για την εξεύρεση πληροφοριών και μοτίβων συσχετίσεων σε μια βάση ή ένα σύνολο δεδομένων, οι οποίες χρησιμοποιούνται επίσης και στις διαδικασίες της μηχανικής μάθησης. Κατά το πέρασμα τον χρόνων, έχουν δοθεί διάφοροι ορισμοί για το πεδίο αυτό της εξόρυξης πληροφορίας. Στη συνέχεια αναφέρονται δύο από αυτούς.

Οι Ρόιγκερ και Γκίατζ (2008) ορίζουν την εξόρυξη πληροφορίας ως:

> «*τη διαδικασία χρήσης μιας ή περισσότερων τεχνικών εκμάθησης υπολογιστών για την αυτόματη ανάλυση και εξαγωγή γνώσεων από δεδομένα που περιέχονται σε μια βάση δεδομένων*».

Ένας δεύτερος ορισμός σχετικά με την εξόρυξη πληροφορίας, σύμφωνα με τους Han, Kamber και Pei (2012) ακολουθεί στην συνέχεια:

> *«Η εξόρυξη πληροφορίας είναι η διαδικασία ανακάλυψης ενδιαφέρον μοτίβων και γνώσης από ένα μεγάλο πλήθος δεδομένων τα οποία είναι αποθηκευμένα σε οποιαδήποτε μορφή αποθήκης δεδομένων».*

Όπως μπορεί να συμπεράνει κανείς από τους προαναφερθέντες ορισμούς σε σύγκριση με τους ορισμούς που δόθηκαν για τη μηχανική μάθηση νωρίτερα, η εξόρυξη πληροφορίας πρόκειται για ένα επιστημονικό πεδίο εξαιρετικά συνδεδεμένο με αυτήν, αφού και στις δυο περιπτώσεις γίνεται χρήση των ίδιων τεχνικών αλλά με διαφορετικούς, βέβαια, σκοπούς. Όπως αναφέρθηκε σε προηγούμενο μέρος της εργασίας, σκοπός του τομέα της εξόρυξης πληροφορίας αποτελεί η εύρεση ενδιαφέρον προτύπων και συσχετίσεων σε μεγάλα σύνολα δεδομένων, βοηθώντας έτσι τον άνθρωπο να πληροφορηθεί κατάλληλα και να πάρει τις βέλτιστες αποφάσεις στην εκάστοτε περίπτωση που αντιμετωπίζει. Η εξόρυξη πληροφορίας μπορεί επίσης να χρησιμοποιηθεί για πρόβλεψη ενός γεγονότος ή για αναλυτική περιγραφή μιας παρούσας κατάστασης[13]. Από την άλλη πλευρά, η μηχανική μάθηση χρησιμοποιεί όλες αυτές τις προβλεπτικές και περιγραφικές τεχνικές εστιάζοντας στην εκπαίδευση, τη διοχέτευση γνώσης και, τη βελτιστοποίηση των διεργασιών ενός υπολογιστικού συστήματος.

Η χρήση των αλγοριθμικών διαδικασιών αυτών που χρησιμοποιούνται κατά την εξόρυξη πληροφορίας, καθίστανται απαραίτητες σήμερα για τη βελτιστοποίηση της κάλυψης των αναγκών των ανθρώπων. Η έγκαιρη διάγνωση καταστάσεων καθώς και η διεξοδική περιγραφή ενός παρόντος γεγονότος, μπορούν να επιφέρουν μόνο καλά αποτελέσματα καθιστώντας τους ανθρώπους, να έχουν ανάγκη πλέον αυτούς τους τύπους μηχανισμών. Οι Han, Kamber και Pei (2012) αναφέρουν χαρακτηρίστηκα τη φράση «*we are living in the information age*», ότι ζούμε δηλαδή ως επί το πλείστον στην εποχή της πληροφορίας και των δεδομένων. Αυτή η έκφραση καθίσταται απόλυτα σωστή αφού οι ρυθμοί που συλλέγονται και αναλύονται τα δεδομένα είναι απίστευτα ραγδαίοι και έχουν βοηθήσει πολλούς σημερινούς τομείς να εξελιχθούν. Συγκεκριμένα, μερικά από τα πεδία εφαρμογής που εφαρμόζεται η εξόρυξη πληροφορίας αναφέρονται στην συνέχεια[14]:

- Η ανακάλυψη γνώσης σε δεδομένα συλλεγμένα από τον παγκόσμιο ιστό,
- Η ανακάλυψη γνώσης σε δεδομένα συλλεγμένα από οποιαδήποτε μορφής δικτύου,
- Η επιχειρηματική ευφυΐα (business intelligence) και η οικονομία,
- Η μηχανική μάθηση και,
- Η δημιουργία διάφορων λογισμικών εφαρμογών (πχ. «έξυπνες εφαρμογές» περιορισμού σπαμ e-mail).

---

[13] Wiki, Εξόρυξη δεδομένων(2020)

[14] Ρόιγκερ και Γκίατζ (2008)

3.1 Μέθοδοι εξόρυξης πληροφορίας

Σύμφωνα με τους Han, Kamber και Pei (2012), οι μέθοδοι εξόρυξης πληροφορίας μπορούν να διακριθούν σε *προβλεπτικές* και *περιγραφικές* μεθόδους. Βάσει αυτών των δύο προσεγγίσεων, υπάρχουν πολλοί αλγόριθμοι οι οποίοι μπορούν να χρησιμοποιηθούν για να πραγματοποιήσουν την εξαγωγή ενός χρήσιμου συμπεράσματος. Στη συνέχεια, θα γίνει αναφορά στη λειτουργία που επιτελεί η κάθε μια από αυτές:

- **Προβλεπτικές μέθοδοι**

Οι *προβλεπτικές μέθοδοι εξόρυξης πληροφορίας* (predictive data mining methods), αφορούν στη χρήση αλγοριθμικών τεχνικών σε ένα σύνολο ή μια βάση δεδομένων για την πρόβλεψη διαφόρων μελλοντικών γεγονότων σχετικά με αυτά. Σύμφωνα με τους Ρόιγκερ και Γκίατζ (2008), η πρόβλεψη αυτή εστιάζει κυρίως σε μια, ή και περισσότερες, μεταβλητές του συνόλου των δεδομένων, οι οποίες επίσης ονομάζονται εξαρτημένες μεταβλητές ή μεταβλητές εξόδου (output data). Οι αλγόριθμοι που συντελούν τις προβλεπτικές τεχνικές μπορούν επίσης να διακριθούν σε δυο μεγάλες κατηγορίες οι οποίες ήδη αναφέρθηκαν νωρίτερα, τους αλγόριθμους *κατηγοριοποίησης* και τους αλγόριθμους *παλινδρόμησης*.

- **Περιγραφικές μέθοδοι**

Οι *περιγραφικές μέθοδοι εξόρυξης πληροφορίας* (descriptive data mining methods), αφορούν στη χρήση αλγοριθμικών τεχνικών σε ένα σύνολο ή μια βάση δεδομένων για τη δημιουργία συσχετίσεων των δεδομένων αλλά και το διαχωρισμό τους βάσει κοινών χαρακτηριστικών. Η διάκριση των αλγοριθμικών τεχνικών που συντελούν τις περιγραφικές μεθόδους μπορούν να διακριθούν βασικά στις τεχνικές της *συσταδοποίησης* και στις τεχνικές *συσχέτισης*.

3.2 Αλγοριθμικές τεχνικές εξόρυξης πληροφορίας και μηχανικής μάθησης

Στη συγκεκριμένη ενότητα θα μελετηθούν οι διάφορες τεχνικές που χρησιμοποιούνται για την κατηγοριοποίηση και την πρόβλεψη στοιχείων ενός συνόλου δεδομένων, σύμφωνα με τις προσεγγίσεις της επιβλεπόμενης μηχανικής μάθησης και της εξόρυξης πληροφορίας. Οι αλγόριθμοι που θα αναλυθούν στην παρούσα ενότητα, αποτελούν ένα σημαντικό

συστατικό στοιχείο του εμπειρικού μέρους της παρούσας πτυχιακής, καθώς στο Μέρος Β΄ θα γίνει πρακτική χρήση τους. Συγκεκριμένα, οι αλγόριθμοι οι οποίοι θα αναλυθούν είναι οι ακόλουθοι:

- k-Κοντινότεροι Γείτονες,
- Μηχανές διανυσμάτων υποστήριξης,
- Τα νευρωνικά δίκτυα,
- Οι αλγόριθμοι δένδρων αποφάσεων,
- Ο απλοϊκός Μπάγιες και,
- Η γραμμική και η λογιστική παλινδρόμηση.

3.2.1 Γραμμική παλινδρόμηση

Η *γραμμική παλινδρόμηση* (linear regression) πρόκειται για μια πολύ απλή μορφή εξίσωσης που μπορεί να χρησιμοποιηθεί για πρόβλεψη. Σύμφωνα με τον Γναρδέλλη (2003), η γραμμική παλινδρόμηση προβλέπει τις τιμές μιας μεταβλητής, η οποία ονομάζεται εξαρτημένη, βάσει μιας ή περισσότερων άλλων μεταβλητών οι οποίες ονομάζονται ανεξάρτητες. Οι τιμές που μπορεί να δώσει η εξίσωση της γραμμικής παλινδρόμησης είναι συνεχείς τιμές, οι οποίες είναι τα αποτελέσματα προβλέψεων της εξαρτημένης μεταβλητής. Ο πιο απλός τύπος της γραμμικής παλινδρόμησης, εκφράζεται από τον ακόλουθο μαθηματικό τύπο[15]:

$$y = a \cdot x + b$$

Το μοντέλο αυτό εκφράζει την εξάρτηση της μεταβλητής y σε σχέση με την μεταβλητή x. Με άλλα λόγια, σύμφωνα με τις τιμές που μπορεί να πάρει η μεταβλητή x μπορούν να προβλεφθούν και οι τιμές της μεταβλητής y. Σε περίπτωση γραφικής αναπαράστασής της συνάρτησης αυτής σε δύο άξονες, x και y , η σταθερά a πρόκειται για την κλίση της ευθείας της εξίσωσης, ενώ το b πρόκειται για το σημείο τομής της στον κάθετο άξονα y (Μητρόπουλος, 2021). Ο υπολογισμός αυτών των δύο μεταβλητών μπορεί να γίνει με διάφορους τρόπους, με τον πιο συνηθισμένο από αυτούς να αποτελεί η χρήση του «*ελάχιστου κριτηρίου τετραγώνων*».

[15] Ρόιγκερ και Γκίατζ, (2008)

Η γραμμική παλινδρόμηση μπορεί να προσαρμοσθεί σε περισσότερες από μια ανεξάρτητες μεταβλητές για την πρόβλεψη της τιμής μιας εξαρτημένης μεταβλητής. Σε αυτήν την περίπτωση, ο τύπος αυτός ονομάζεται *πολλαπλή γραμμική παλινδρόμηση* (multiple linear regression)[16].

3.2.2 Λογιστική παλινδρόμηση

Η *λογιστική παλινδρόμηση* (logistic regression), πρόκειται για μια στατιστική τεχνική δυαδικής κατηγοριοποίησης η οποία χρησιμοποιείται πολύ συχνά στο πεδίο της μηχανικής μάθησης και της εξόρυξης πληροφορίας. Κατά τη λογιστική παλινδρόμηση, η τιμή της εξαρτημένης μεταβλητής μπορεί να κυμανθεί στις τιμές του μηδέν και του ένα, ενώ η συνάρτησή της, γενικά, εκφράζει την πιθανότητα της εξαρτημένης μεταβλητής να ανήκει ή να μην ανήκει σε μια κατηγορία ή μια κλάση. Ειδικότερα, ο μαθηματικός τύπος που την εκφράζει είναι ο ακόλουθος[17]:

$$P = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n + b$$

Όπου το a_n και το b πρόκειται για σταθερούς όρους, ενώ τα x_n αποτελούν τις ανεξάρτητες μεταβλητές που αναλύονται. Η μεταβλητή P, είναι αυτή που αφορά το λόγο της εμφάνισης μιας κατηγορίας που σχετίζεται με $y = 1$, σε σχέση με την εμφάνιση μιας κατηγορίας που σχετίζεται με $y = 0$. Συμβολίζεται ως ο φυσικός λογάριθμος του λόγου της πιθανότητας p να ανήκει μια μεταβλητή σε μια κατηγορία ως προς την πιθανότητα να μην ανήκει. Δίνεται από τον τύπο: $\ln\left(\frac{p}{1-p}\right)$.

Σύμφωνα με τον Γναρδέλλη (2003), ο λόγος της πιθανότητας για την εμφάνιση μιας κλάσης συνδέει την πιθανότητα επιτυχίας p με τις ανεξάρτητες μεταβλητές x_n. Βάσει των λογαριθμικών μοντέλων, ο συγκεκριμένος λόγος ονομάζεται ως $logit$ της y και συμβολίζεται ως $logit(p)$. Δηλαδή:

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \alpha_1 x_1 + \ldots + \alpha_n x_n + b = \frac{e^{\alpha_1 x_1 + \ldots + \alpha_n x_n + b}}{1 + e^{\alpha_1 x_1 + \ldots + \alpha_n x_n + b}}$$

Όπου e είναι η σταθερά της βάσης του φυσικού λογαρίθμου.

---

[16] Boutsikas (2004)

[17] Ρόιγκερ και Γκίατζ (2008)

3.2.3 Αλγόριθμος k-Κοντινότερων Γειτόνων

Ο *αλγόριθμος των k-κοντινότερων γειτόνων* (k-nearest neighbor - kNN), αναφέρθηκε πρώτη φορά τη δεκαετία του 1950 και πρόκειται για έναν αλγόριθμο που εμπίπτει στην κατηγορία αυτή των αλγορίθμων κατηγοριοποίησης (Han, Kamber και Pei, 2012). Ένα από τα κύρια μειονεκτήματα του συγκεκριμένου αλγορίθμου, αποτελεί ο μεγάλος υπολογιστικός χρόνος για την επίτευξη της κατηγοριοποίησης ενός στοιχείου, καθώς και ότι δεν είναι πάντα ακριβής.

Η βασική τεχνική που ακολουθεί ο συγκεκριμένος αλγόριθμος έχει ως ακολούθως: Θεωρείται ότι για οποιοδήποτε στοιχείο το οποίο δίνεται ως μη ετικετοποιημένο δεδομένο στον αλγόριθμο, η κλάση που θα του αντιστοιχεί θα είναι μια από τις κλάσεις των πλησιέστερων ετικετοποιημένων δεδομένων κοντά σε αυτό (Wu, ed al., 2007). Μια σημαντική παράμετρος που δέχεται ο συγκεκριμένος αλγόριθμος από το χρήστη, αποτελεί το πόσα στοιχεία μπορούν να θεωρηθούν κοντινότεροι «γείτονες» σχετικά με το μη ετικετοποιημένο δεδομένο. Αυτή η παράμετρος που αναφέρθηκε ορίζεται με το γράμμα k και, είναι αυτή που απαρτίζει και το αρχικό μέρος του ονόματος του αλγορίθμου.

Πιο συγκεκριμένα, μια απλή εκδοχή του αλγόριθμου kNN εμπεριέχει τα εξής βήματα[18]:

1. Χρήση ενός ετικετοποιημένου συνόλου δεδομένων.
2. Είσοδος στο σύστημα ενός μη ετικετοποιημένου δεδομένου.
3. Επιλογή της μεταβλητής k των πιθανών κοντινότερων γειτόνων.
4. Εύρεση των k κοντινότερων γειτόνων για το μη ετικετοποιημένο δεδομένο χρησιμοποιώντας ένα «μέτρο απόστασης».
5. Επιλογή της κατάλληλης κλάσης του μη ετικετοποιημένου δεδομένου, βάσει αυτής που εμφανίζεται πιο συχνά στο σύνολο των k στοιχείων.

Ως «μέτρο απόστασης» που αναφέρθηκε, ονομάζεται το μέτρο εκείνο που θα δείξει ποια στοιχεία αποτελούν τα πιο κοντινά σε απόσταση από αυτό του δεδομένου εισόδου μας. Υπάρχουν πολλά είδη μέτρων απόστασης που μπορούν να χρησιμοποιηθούν. Τρείς από αυτές τις μετρικές αποτελούν α) η «*Ευκλείδεια Απόσταση*», β) η «*Απόσταση Μανχάταν*» και γ) η «*Απόσταση Μινκόβσκι*».

---

[18]Kotsiantis (2007)

- *Ευκλείδεια Απόσταση (Euclidian Distance):* Πρόκειται για ένα μαθηματικό ορισμό που αφορά στον υπολογισμό της απόστασης δύο σημείων ο οποίος δόθηκε από τον γνωστό μαθηματικό Ευκλείδη. Ο μαθηματικός τύπος της Ευκλείδειας Απόστασης σε ένα χώρο αναπαράστασης των δεδομένων ορίζεται ως εξής[19]:

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Όπου D είναι η απόσταση των δύο σημείων x και y, x_i είναι το υπόλοιπο της αφαίρεσης για την απόσταση των δυο συντεταγμένων x και, y_i είναι το υπόλοιπο της αφαίρεσης για την απόσταση των δυο συντεταγμένων y.

- *Απόσταση Μινκόβσκι (Minkowski Distance):* Το συγκεκριμένο είδος απόστασης δύο σημείων δόθηκε από τον μαθηματικό Χέρμαν Μινκόβσκι και πρόκειται για μια γενίκευση της Ευκλείδειας Απόστασης σε συνδυασμό με την Απόσταση Μανχάταν. Ο μαθηματικός τύπος της συγκεκριμένη απόστασης δίνεται από[20]:

$$D(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

Όπου D είναι η απόσταση των δύο σημείων x και y, x_i είναι το υπόλοιπο της αφαίρεσης για την απόσταση των δυο συντεταγμένων x, y_i είναι το υπόλοιπο της αφαίρεσης για την απόσταση των δυο συντεταγμένων y και, p μια σταθερά η οποία υπολογίζεται βάσει της «*ανισότητας του Μινκόβσκι*».

3.2.4 Μηχανές διανυσμάτων υποστήριξης

Οι *μηχανές διανυσμάτων υποστήριξης* (support vector machines - SVM), πρόκειται για έναν αλγόριθμο κατηγοριοποίησης ο οποίος χρησιμοποιείται σε σύνολα με επιβλεπόμενη μάθηση. Η τεχνική αυτή θεωρείται από τις πιο ακριβείς μεθόδους σε σύγκριση με άλλους

[19] Kotsiantis (2007)

[20] Sharma (2020)

αλγόριθμους κατηγοριοποίησης και εισάχθηκε πρώτη φορά στο πεδίο της μηχανικής μάθησης το 1992.

Σύμφωνα με τους Han, Kamber και Pei (2012), ο συγκεκριμένος τύπος αλγορίθμου αναζητάει τον καλύτερο γραμμικό διαχωρισμό των δεδομένων βάσει των κλάσεων τους. Αυτό πραγματοποιείται μέσω της δημιουργίας μιας ευθείας, ή κυρτής, γραμμής στο χώρο του συνόλου δεδομένων (data space), η οποία θα τα διαχωρίσει καλύτερα βάσει των κλάσεων τους (Wu, ed al., 2007). Σκοπός της συγκεκριμένης τεχνικής είναι η μεγιστοποίηση της τιμής margin, η οποία ορίζεται ως η απόσταση που έχουν τα πιο κοντινά στοιχεία (σημεία υποστήριξης ή αλλιώς support vectors) προς τη γραμμή της γραμμικής διαχώρισης τους[21]. Ως αποτέλεσμα, με τη μεγιστοποίηση της τιμής margin, τα μη ετικετοποιημένα στιγμιότυπα που θα εισέλθουν στο σύστημα, θα έχουν περισσότερες πιθανότητες για τη σωστή ταξινόμηση τους στη σωστή κλάση που τους αντιστοιχεί.

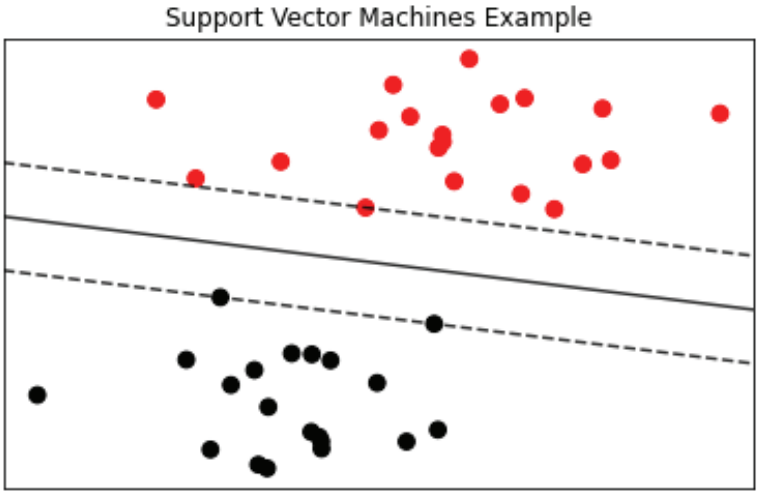
Ένα πολύ βασικό χαρακτηριστικό της συγκεκριμένης τεχνικής, αποτελεί η προσαρμογή της στην εκάστοτε κατάσταση ακόμα και αν τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα. Οι αλγόριθμοι των μηχανών διανυσμάτων υποστήριξης, χρησιμοποιούν μια πληθώρα συναρτήσεων (kernel functions), για τη μετατροπή των παρόντων δεδομένων σε μεγαλύτερες διαστάσεις με την προσθήκη μιας ή και περισσότερων μεταβλητών σε αυτά. Ως αποτέλεσμα, η αναπαράσταση των δεδομένων στην καινούργια κλίμακα διαστάσεων, μπορεί να αποδειχθεί πολύ ωφέλιμη καθώς θα γίνει εφικτός ο διαχωρισμός των δεδομένων και, συνεπώς, η πιο ακριβής κατηγοριοποίηση τους. Αξίζει να σημειώσουμε κάπου εδώ, ότι οι γραμμικές διαχωρίσεις αυτές ονομάζονται hyperplanes και, μπορούν να διαχωρίσουν τις κλάσεις του συνόλου δεδομένων τόσο σε ένα δυσδιάστατο χώρο όσο και σε έναν n-διάστατο.

Έστω, για παράδειγμα, ότι γίνεται μελέτη σε ένα δυσδιάστατο χώρο (x - y), με δύο κλάσεις δεδομένων οι οποίες δεν είναι γραμμικά διαχωρίσιμες. Ο αλγόριθμος αυτός, μπορεί να δημιουργήσει μια νέα μεταβλητή z για την προσθήκη της στην αναπαράσταση των δεδομένων σε έναν τρισδιάστατο πλέον χώρο. Η μεταβλητή αυτή μπορεί να δημιουργηθεί μέσω πολλών σεναρίων, με την πιο απλή περίπτωση να είναι το άθροισμα των τετραγώνων των δυο προηγούμενων μεταβλητών που αναπαριστανόταν (Navlani, 2019). Δηλαδή, οι

[21] Wu, ed al. (2007)

καινούριες μεταβλητές z θα είναι της μορφής: $z_n = x_n^2 + y_n^2$, όπου n ο αριθμός κατάταξης ενός στιγμιότυπου μέσα στο σύνολο.

Για να γίνει πιο κατανοητή η λειτουργία της συγκεκριμένης τεχνικής, στη συνέχεια θα δοθεί ένα παράδειγμα διαχωρισμού που θα επιτελούσε ένας SVM αλγόριθμος σε δυο γραμμικά διαχωρίσιμες μεταβλητές ενός δυσδιάστατου χώρου («Εικόνα 2»).



Εικόνα 2: Παράδειγμα SVM αλγόριθμου.

Στην παραπάνω εικόνα αναπαρίστανται γραφικά δύο κλάσεις δεδομένων (η κόκκινη και η μαύρη), κατανεμημένες σε έναν δυσδιάστατο χώρο. Η μαύρη ευθεία γραμμή, πρόκειται για το hyperplane που διαχωρίζει καλύτερα τις δυο αυτές κλάσεις. Οι διακεκομμένες μαύρες γραμμές, πρόκειται για τις δύο παράλληλες ευθείες στο hyperplane και, αναπαριστούν γραφικά τα δυο κοντινότερα σημεία των δύο κλάσεων αυτών. Όπως παρατηρείται, ο διαχωρισμός στη συγκεκριμένη περίπτωση, έχει πραγματοποιηθεί όσο καλύτερα καταστάθηκε εφικτός βάσει της τιμής margin.

3.2.5 Νευρωνικά δίκτυα

Τα *τεχνητά νευρωνικά δίκτυα* (artificial neural networks), ή απλώς νευρωνικά δίκτυα, αποτελούν ένα μοντέλο βιολογικού νευρωνικού δικτύου που χρησιμοποιείται για πρόβλεψη και κατηγοριοποίηση, όντας σε θέση να γίνει χρήση είτε επιβλεπόμενης είτε μη επιβλεπόμενης μάθησης (autoencoders[22]). Σύμφωνα με τον Hardesty (2017), οι πρώτοι

[22] Savvopoulos & Kalogeras & Anagnostopoulos & Alexakos & Siountas & Kalogeras

άνθρωποι που εφάρμοσαν τη χρήση των νευρωνικών δικτύων ήταν ο Walter Pitt και ο McCulloch Warren το 1944. Ένα βασικό χαρακτηριστικό πλεονέκτημα της συγκεκριμένης τεχνικής, αποτελεί η «ανοχή» που διαθέτουν σε περίπτωση ύπαρξης δεδομένων με θόρυβο (noisy data).

Ένα νευρωνικό δίκτυο αποτελείται από διάφορα επίπεδα με το κάθε ένα από αυτά να αποτελείται από διάφορους κόμβους και νευρώνες συνδεδεμένους μεταξύ τους. Τα επίπεδα που μπορούν να διαχωριστούν στα νευρωνικά δίκτυα αποτελούν το επίπεδο εισόδου, το κρυφό επίπεδο και, το επίπεδο εξόδου (Bishop, 1995). Το επίπεδο εισόδου (input layer) πρόκειται για το επίπεδο στο οποίο εισέρχονται τα δεδομένα για επεξεργασία. Το κρυφό επίπεδο (hidden layer), το οποίο μπορεί να έχει και παραπάνω από ένα στρώματα, επιτελεί την όλη αλγοριθμική διαδικασία για τον υπολογισμό του αποτελέσματος όπου θα εξαχθεί στο επίπεδο εξόδου του δικτύου (output layer). Το κρυφό επίπεδο των νευρωνικών δικτύων αποτελείται από διάφορους νευρώνες (artificial neuron), οι οποίοι επιτελούν τον υπολογισμό μιας μη γραμμικής, συνήθως, συνάρτησης και, προωθούν το αποτέλεσμα τους στον επόμενο νευρώνα από αυτούς για επιπλέον επεξεργασία. Σύμφωνα με τον Ψούνη (2015), ο κάθε νευρώνας από αυτούς αποτελείται από δύο μέρη. Το πρώτο ονομάζεται *αθροιστής* (summation function) και το δεύτερο ονομάζεται *συνάρτηση ενεργοποίησης* (activation function). Αναλυτικότερα:

- Σε γενικές γραμμές, ο *αθροιστής* δέχεται ένα σύνολο από δεδομένα $(x_1, x_2 \ldots\ldots x_n)$ στις εισόδους του τα οποία ονομάζονται και σήματα. Αναλόγως της σημαντικότητας κάθε ενός από αυτών των σημάτων για τον υπολογισμό της εξόδου, τα δεδομένα μεταβάλλονται με μια ξεχωριστή τιμή βάρους $(w_1, w_2 \ldots . w_n)$ το κάθε ένα (Lorenzo, 2017). Ειδικότερα, η δουλειά που επιτελεί ο αθροιστής είναι να υπολογίζει το άθροισμα όλων των γινομένων μεταξύ των δεδομένων εισόδου στον νευρώνα και των βαρών τους. Δίνεται από τον τύπο[23]:

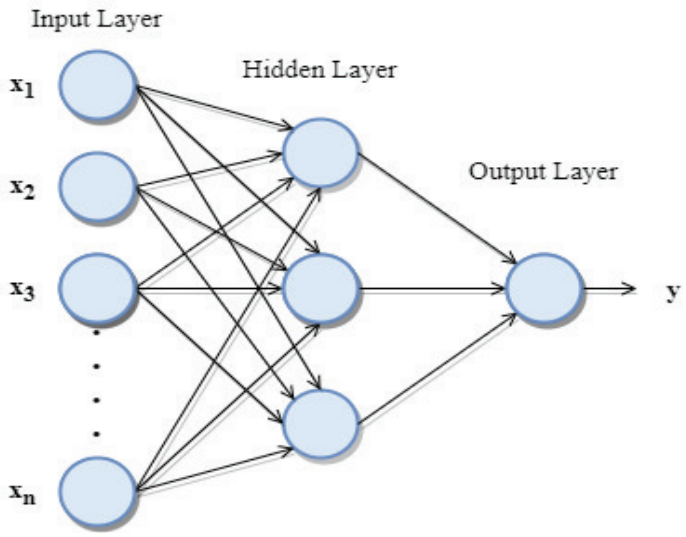
$$V = \sum_{i=0}^{n} w_i x_i = w_0 x_0 + w_1 x_1 + \cdots + w_n x_n$$

[23] Bishop (1995)

- Η *συνάρτηση ενεργοποίησης* δέχεται σαν είσοδο την τιμή που υπολογίστηκε στον αθροιστή (V) και, παράγει βάσει αυτής κάποιο αποτέλεσμα στην έξοδο του νευρώνα. Σύμφωνα με τον Lorenzo (2017), αλλά και τον Ψούνη (2015), τέτοιων ειδών συναρτήσεις αποτελούν οι ακόλουθες: α) η βηματική συνάρτηση (step function), β) η συνάρτηση πρόσημου (sign function), γ) η σιγμοειδής συνάρτηση (sigmoid function), δ) η συνάρτηση ανορθωμένης γραμμικής μονάδας (ReLu) και διάφορες άλλες.

Τα νευρωνικά δίκτυα μπορούν να διαχωριστούν σε κατηγορίες βάσει της κατεύθυνσης ροής των δεδομένων μέσα σε αυτά, βάσει των στρώσεων των κρυφών επιπέδων τους κλπ. Μια κύρια κατηγορία από αυτές αποτελούν τα δίκτυα εμπρός τροφοδότησης (feedforward networks), τα οποία είναι σχεδιασμένα να προωθούν δεδομένα για επεξεργασία στο ακριβώς επόμενο επίπεδο από αυτά (Ρόιγκερ και Γκίατζ, 2008). Ένα πολύ διαδεδομένο εμπρός τροφοδότησης δίκτυο το οποίο αποτελείται από πολλούς εισόδους και μια μόνο έξοδο, αποτελεί το δίκτυο Perceptron (Ψούνης, 2015). Το συγκεκριμένο δίκτυο μπορεί να περιέχει διάφορες στρώσεις από κρυφά επίπεδα (multilayer perceptron - MLP), ενώ οι νευρώνες των επιπέδων αυτών συνδέονται μόνο με τα επόμενα στρώματα και όχι μεταξύ τους.

Η «Εικόνα 3» αναπαριστά ένα εμπρός τροφοδότησης νευρωνικό δίκτυο τύπου Perceptron, το οποίο περιέχει ένα στρώμα κρυφού επιπέδου. Το δίκτυο αυτό, αποτελεί επίσης ένα πλήρως συνδεδεμένο δίκτυο (fully-connected network), αφού όλοι οι κόμβοι και οι νευρώνες του, είναι πλήρως συνδεδεμένοι με τα επόμενα επίπεδα από αυτά.



Εικόνα 3: Νευρωνικό δίκτυο τύπου Perceptron.

Ένας τρόπος επιβλεπόμενης μάθησης σε δίκτυα εμπρός τροφοδότησης σαν τα Perceptron, αποτελεί η «*εκμάθηση με οπισθοδρόμηση*». Σύμφωνα με τους Ρόιγκερ και Γκίατζ (2008), η βασική λογική του συγκεκριμένου αλγορίθμου εκπαίδευσης έχει ως εξής:

1. Στο πρώτο στάδιο επιλέγονται βάρη για κάθε νευρώνα του δικτύου όπου και εισέρχονται δεδομένα με έναν τυχαίο τρόπο.
2. Ως δεύτερο βήμα, τα δεδομένα προωθούνται προς την έξοδο του δικτύου(y_n), όπου και το αποτέλεσμα συγκρίνεται με το επιθυμητό αποτέλεσμα που έχει ορίσει ο χρήστης (d_n) και υπολογίζεται ένα σφάλμα, το οποίο υποδεικνύει κατά πόσο ήταν εσφαλμένος ο υπολογισμός του νευρώνα. Σύμφωνα με τη Γεωργούλη (2015), το σφάλμα ($error$) αυτό δίνεται από τον τύπο: $E(error) = d_n - y_n$
3. Στη συνέχεια, το σφάλμα εξόδου αυτό διαδίδεται στο δίκτυο προς τα πίσω. Κατά τη μετάδοση του σφάλματος κάθε νευρώνας υπολογίζει και ένα νέο βάρος για το δεδομένο εισόδου του, βάσει ενός μαθηματικού τύπου ο οποίος λαμβάνει υπ' όψη το σφάλμα που υπολογίστηκε στην έξοδο, τις τιμές εξόδου των νευρώνων και την παράγωγο της συνάρτησης ενεργοποίησης του.
4. Η διαδικασία αυτή επαναλαμβάνεται μέχρις ότου το σφάλμα να μειωθεί σε σημαντικό βαθμό ή μέχρις ότου πραγματοποιηθεί ένα κριτήριο τερματισμού, όπως για παράδειγμα, η ολοκλήρωση ενός συγκεκριμένου αριθμού επαναλήψεων εκπαίδευσης.

3.2.6 Απλοϊκός Bayes

Ο *απλοϊκός Bayes* (Naïve Bayes), πρόκειται για έναν ισχυρό τύπο κατηγοριοποίησης ο οποίος χρησιμοποιείται ως τεχνική στην επιβλεπόμενη μάθηση, με ένα σημαντικό πλεονέκτημά του να αποτελεί η ανοχή σε δεδομένα με θόρυβο. Ο συγκεκριμένος αλγόριθμος, βασίζεται στις δεσμευμένες πιθανότητες του «*θεωρήματος του Bayes*», υποθέτοντας ότι όλα τα στοιχεία που αναλύονται είναι ανεξάρτητα μεταξύ τους, αλλά με την ίδια όμως σημασία για την εξαγωγή του αποτελέσματος μιας πιθανότητας.

Το θεώρημα αυτό διατυπώθηκε από τον γνωστό Βρετανό μαθηματικό Τόμας Μπάγιες (Thomas Bayes) και, δόθηκε στην δημοσιότητα το 1763. Το συγκεκριμένο θεώρημα διατυπώνεται ως ακολούθως:

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}$$

Όπου $P(Y|X)$ είναι η δεσμευμένη πιθανότητα ενός γεγονότος Υ να εξαρτάται από το X, $P(X|Y)$ είναι η δεσμευμένη πιθανότητα ενός γεγονότος Χ να εξαρτάται από το Υ και, $P(X)$ και $P(Y)$ είναι οι ανεξάρτητες πιθανότητες των γεγονότων X και Y αντίστοιχα.

Βάσει του θεωρήματος αυτού, γίνεται μια αντίστοιχη διατύπωση που χρησιμοποιείται για την εύρεση της πιθανότητας του στιγμιότυπου x να ανήκει στην κλάση y_n. Η συγκεκριμένη διατύπωση ονομάζεται «*απλοϊκός Bayes*» και, είναι μια τεχνική που χρησιμοποιείται στο πεδίο της μηχανικής μάθησης και της εξόρυξης πληροφορίας για κατηγοριοποίηση και πρόβλεψη. Ο υπολογισμός της πιθανότητας αυτής που αναφέρθηκε, πραγματοποιείται για όλες τις κλάσεις του συνόλου, ταξινομώντας έπειτα ένα μη ετικετοποιημένο δεδομένο στην κλάση που έχει την περισσότερη πιθανότητα να ανήκει. Για τον υπολογισμό του $P(x|y_n)$, της πιθανότητας δηλαδή ότι το στιγμιότυπο x να ανήκει στην κλάση y_n, γίνεται αποδοχή της αρχής του θεωρήματος του Bayes, ότι όλα τα δεδομένα αλλά και οι μεταβλητές τους είναι ανεξάρτητα μεταξύ τους. Σύμφωνα με τους Han, Kamber και Pei (2012) υπολογίζεται ως ακολούθως:

$$P(x|y_n) = \prod_{i=1}^{n} P(x_i|y_n) = P(x_1|y_n)P(x_2|y_n)\dots P(x_i|\, y_n)$$

3.2.7 Αλγόριθμοι δένδρων αποφάσεων

Οι *αλγόριθμοι δένδρων αποφάσεων* (decision tree algorithms) αποτελούν ένα αρκετά απλό και διαδεδομένο είδος αλγορίθμου επιβλεπόμενης μάθησης ο οποίος χρησιμοποιείται για κατηγοριοποίηση σε περίπτωση ύπαρξης διακριτών τιμών δεδομένων και, για παλινδρόμηση σε περίπτωση ύπαρξης συνεχών τιμών δεδομένων. Μερικοί από τους πιο γνωστούς αλγορίθμους δένδρων αποφάσεων είναι ο ID3, ο CART και ο C4.5 (Brijain και Kushik, 2014).

Η γενική ιδέα του συγκεκριμένου τύπου αλγορίθμων, είναι η κατασκευή «*κανόνων συσχετίσεων*» των δεδομένων, ξεκινώντας από μια μεταβλητή με τις περισσότερες εμφανίσεις στο σύνολο και, κάνοντάς την «ρίζα» μιας υποτιθέμενης αναπαράστασης ενός δένδρου. Τα δένδρα αυτά απαρτίζονται από τρία στοιχειώδη μέρη: α) τους κόμβους (nodes), β) τα κλαδιά (branches) και γ) τα φύλλα (leaf node). Σύμφωνα με τον Κοτσιάντη (2007), οι κόμβοι αποτελούν την αναπαράσταση μιας μεταβλητής του συνόλου, ενώ τα κλαδιά, τις τιμές που μπορεί να πάρει η μεταβλητή αυτή. Τα κλαδιά με τη σειρά τους μπορεί να

χωρίζονται σε άλλους κόμβους ή σε φύλλα. Τα φύλλα, πρόκειται για την αναπαράσταση των κλάσεων του συνόλου και τις τιμές που παίρνουν αυτές βάσει των προηγούμενων κόμβων και κλαδιών όπου και συνδέονται.

Η βασική αλγοριθμική τεχνική των δένδρων αποφάσεων έχει τα εξής βήματα[24]:

1. Εύρεση της καλύτερης μεταβλητής του συνόλου των δεδομένων και δημιουργία της ρίζας του δένδρου, βάσει αυτής.
2. Δημιουργία κλαδιών σύμφωνα με τις τιμές της μεταβλητής αυτής που επιλέχθηκε.
3. Καταμερισμός των υπόλοιπων μεταβλητών του συνόλου σε τόσα σύνολα όσα και τα υπάρχοντα κλαδιά.
4. Επιλογή επιπλέον συνόλων για την περεταίρω διάσπαση του δένδρου μέχρις ότου όλες οι μεταβλητές να φτάσουν σε ένα φύλλο βάσει του πρώτου βήματος.

Η εύρεση της καλύτερης μεταβλητής στο σύνολο των δεδομένων που αναφέρθηκε στο «βήμα 1», αφορά ένα μαθηματικό υπολογισμό ο οποίος δείχνει το συνολικό όφελος για την επιλογή μιας συγκεκριμένης μεταβλητής ως «κεφαλή» για τους υπόλοιπους διαχωρισμούς του δένδρου. Ο υπολογισμός αυτός μπορεί να πραγματοποιηθεί με τη χρήση της *εντροπίας* και του *πληροφοριακού κέρδους* ή και διάφορων άλλων συντελεστών, όπως για παράδειγμα ο *συντελεστής Gini*.

Η προσέγγιση της *εντροπίας* (entropy) και του *πληροφοριακού κέρδους* (information gain), χρησιμοποιείται ως επι το πλείστον στους αλγόριθμους ID3 και C4.5. Βάσει αυτής, χρησιμοποιείται ένας μαθηματικός τύπος για τον υπολογισμό του κέρδους πληροφορίας ξεχωριστά για την κάθε μεταβλητή. Αυτή η μεταβλητή που έχει τη μεγαλύτερη τιμή που υπολογίσθηκε, χρησιμοποιείται για τη δημιουργία της ρίζας του δένδρου, όπως ήδη αναφέρθηκε.

Σύμφωνα με τους Han, Kamber και Pei (2012), για να υπολογιστεί το κέρδος της πληροφορίας μιας μεταβλητής i, πρέπει πρώτα να υπολογιστεί η *εντροπία* της, η οποία συμβολίζεται ως $E(S_i)$, αλλά και η *συνολική εντροπία διαχωρισμού* της στο σύνολο των δεδομένων, η οποία συμβολίζεται ως $info(A_i)$.

- Η εντροπία $E(S_i)$ μιας μεταβλητής υπολογίζεται από τον τύπο:

$$E(S_i) = -\sum p_i \cdot \log(p_i)$$

[24] Suthaharan (2016)

Όπου p_i είναι η πιθανότητα εμφάνισης μίας κλάσης στο σύνολο των δεδομένων.

- Η συνολική εντροπία διαχωρισμού $info(A_i)$ μιας μεταβλητής στο σύνολο των δεδομένων υπολογίζεται από τον τύπο:

$$info(A_i) = \sum \frac{|D_j|}{|D|} \cdot E(S_i)$$

Όπου $\frac{|D_j|}{|D|}$ είναι ένα μέτρο υπολογισμού για το πλήθος των εμφανίσεων τιμών μίας μεταβλητής στο σύνολο και, $E(S_i)$ η εντροπία της μεταβλητής που μελετάται στην εκάστοτε περίπτωση.

Σύμφωνα με αυτά που αναφέρθηκαν, το κέρδος πληροφορίας μιας μεταβλητής υπολογίζεται ως η διαφορά της εντροπίας της με τη συνολική εντροπία του διαχωρισμού του συνόλου των δεδομένων. Δίνεται από τον τύπο:

$$Gain(S_i, A_i) = E(S_i) - info(A_i)$$

3.3 Αξιολόγηση αποτελεσμάτων επιβλεπόμενης μάθησης

Η *αξιολόγηση των αποτελεσμάτων* αλγορίθμων κατηγοριοποίησης και παλινδρόμησης, πρόκειται για ένα αρκετά σημαντικό στάδιο, κατά το οποίο μελετώνται μετρικές όπως η ακρίβεια των αλγορίθμων στις προβλέψεις τους. Το στάδιο αυτό διαδραματίζεται μετά από κάθε «συνεδρία» επιβλεπόμενης μάθησης, έχοντας μεγάλη χρησιμότητα για την πιθανή επιλογή κάποιου μοντέλου για τη χρήση του ξανά στο μέλλον.

Σύμφωνα με τους Ρόιγκερ και Γκίατζ (2008), μια προσέγγιση ιδιαίτερης χρησιμότητας για την αξιολόγηση των αποτελεσμάτων επιβλεπόμενης μάθησης, αποτελεί ο «*πίνακας σύγχυσης*» (confusion matrix), ή αλλιώς «*μήτρα σύγχυσης*». Ο πίνακας αυτός, αναπαριστά με ένα γραφικό τρόπο τα αποτελέσματα των αλγορίθμων μετά την εκπαίδευση τους, σύμφωνα με τις προβλέψεις τους στο σύνολο ελέγχου (test set). Η κατανομή των αποτελεσμάτων πραγματοποιείται με την κατανομή του πλήθους όλων των σωστών κλάσεων που ανακαλύφθηκαν για τα μη ετικετοποιημένα δεδομένα του συνόλου ελέγχου στην κύρια διαγώνιο του πίνακα. Στη συνέχεια, γίνεται μια ανασκόπηση των στοιχείων αυτών που μπορούν να καταγραφούν σε έναν πίνακα σύγχυσης. Αναλυτικότερα:

- Τα στοιχεία τα όποια προβλέφθηκαν σωστά από τον αλγόριθμο ως θετική κλάση. Αυτού του είδους τα στοιχεία ονομάζονται True Positive (TP).
- Τα στοιχεία τα οποία προβλέφθηκαν σωστά από τον αλγόριθμο ως αρνητική κλάση. Αυτού του είδους τα στοιχεία ονομάζονται True Negative (TN).
- Τα στοιχεία τα οποία προβλέφθηκαν εσφαλμένα από τον αλγόριθμο ως θετική κλάση. Αυτού του είδους τα στοιχεία ονομάζονται False Positive (FP).
- Τα στοιχεία τα οποία προβλέφθηκαν εσφαλμένα από τον αλγόριθμο ως αρνητική κλάση. Αυτού του είδους τα στοιχεία ονομάζονται False Negative (FN).

Για να γίνουν πιο κατανοητοί οι πιο πάνω ορισμοί, στη συνέχεια («Πίνακας 2»), θα δοθεί ένα παράδειγμα ενός απλού πίνακα σύγχυσης. Σε περίπτωση ύπαρξης δύο κλάσεων, της θετικής (y_1) και της αρνητικής (y_2), η καταγραφή των αποτελεσμάτων στον πίνακα αυτόν έχει ως ακολούθως:

Πίνακας 2: Παράδειγμα πίνακα σύγχυσης.

Πίνακας σύγχυσης		
Κλάσεις:	y_1	y_2
Αποτελέσματα αλγορίθμου:	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

Βάσει του πίνακα σύγχυσης και όλων των αριθμητικών δεδομένων που μπορούν να καταγραφούν σε αυτόν, μπορούν να υπολογιστούν διάφορες άλλες μετρικές οι οποίες βοηθούν στη διεξοδική αξιολόγηση του εκάστοτε μοντέλου. Δύο από αυτές αποτελούν η *ορθότητα* (accuracy) και η *ακρίβεια* (precision). Στη συνέχεια, ακολουθεί μια συνοπτική ανασκόπηση της κάθε μίας από αυτές τις μετρικές[25].

- Η *ορθότητα* ενός αλγορίθμου στο πεδίο της μηχανικής μάθησης και της εξόρυξης πληροφορίας, ορίζεται ως το συνολικό μέτρο αξιολόγησης σχεσιακά με τις σωστές και τις λανθασμένες κατηγοριοποιήσεις-προβλέψεις των μη ετικετοποιημένων δεδομένων. Ένα αρκετά καλό ποσοστό ορθότητας των μοντέλων κυμαίνεται από

[25] Shahadat, Arif, Ekramul και Mohamad (2019)

ογδόντα τις εκατό και άνω (≥80%). Η συγκεκριμένη μετρική, χρησιμοποιείται τόσο στην αξιολόγηση των κατηγοριοποιητών, όσο και στους αλγόριθμους τύπου παλινδρόμησης. Στη συνέχεια, δίνεται ο μαθηματικός τύπος της ορθότητας ενός αλγορίθμου.

$$Ορθότητα\ (Accuracy) = \frac{TP + TN}{TP + FN + FP + TN}\ \%$$

- Η *ακρίβεια* ενός αλγορίθμου στο πεδίο της μηχανικής μάθησης και της εξόρυξης πληροφορίας, πρόκειται για το μέτρο αξιολόγησης της αποτελεσματικότητας του αλγορίθμου για τον όσο καλύτερο διαχωρισμό των κλάσεων γίνεται, αλλά και την όσο πιο σωστή πρόβλεψη των τιμών του συνόλου δοκιμής. Συγκεκριμένα, αφορά το ποσοστό των θετικών προβλέψεων που αφορούν όντως τις θετικές εγγραφές κλάσεων του συνόλου. Η μετρική αυτή, βρίσκει πεδίο αξιολόγησης μόνο στους αλγορίθμους τύπου κατηγοριοποίησης. Ο μαθηματικός τύπος που δίνεται για την ακρίβεια ενός αλγορίθμου ακολουθεί στη συνέχεια.

$$Ακρίβεια\ (Precision) = \frac{TP}{TP + FP}\ \%$$

ΜΕΡΟΣ Β΄

Στο παρόν δεύτερο μέρος της πτυχιακής εργασίας, πραγματοποιείται η παρουσίαση τριών διαφορετικών μελετών περιπτώσεων που πραγματοποιήθηκαν με τη χρήση επιβλεπόμενης μηχανικής μάθησης. Ειδικότερα, επιδιώχθηκε η εξόρυξη γνώσης από δεδομένα με αλγόριθμους και, η εύρεση του καλύτερου αλγοριθμικού μοντέλου που θα φέρει τα καλύτερα αποτελέσματα σε προβλέψεις και κατηγοριοποιήσεις.

Το τέταρτο κεφαλαίο, εστιάζει στην ανασκόπηση των εργαλείων που θα χρησιμοποιηθούν για τη διεξαγωγή των μελετών αυτών, συμπεριλαμβανομένης της επεξήγησης των συνόλων δεδομένων και των αλγορίθμων που θα χρησιμοποιηθούν. Συγκεκριμένα, οι μελέτες αυτές πραγματοποιήθηκαν σε τρία διαφορετικά σύνολα δεδομένων που ανακτήθηκαν από το διαδίκτυο. Η εφαρμογή των αλγορίθμων στα σύνολα αυτά, πραγματοποιήθηκε μέσω της γλώσσας προγραμματισμού Python. Η περιεκτικότητα της γλώσσας αυτής σε μια πληθώρα ειδικών προγραμματιστικών εργαλείων, την καθιστούν ιδιαίτερα διαδεδομένη στον ευρύτερο χώρο της ανάλυσης δεδομένων και της μηχανικής μάθησης. Ο κώδικας που κατασκευάστηκε για την εφαρμογή των αλγορίθμων μέσω της γλώσσας αυτής, μπορεί να βρεθεί στα παραρτήματα της παρούσας Πτυχιακής Εργασίας.

Στο πέμπτο, στο έκτο και στο έβδομο κεφάλαιο της εργασίας, δίνεται η μεθοδολογία της εφαρμογής των αλγορίθμων μηχανικής μάθησης στα τρία σύνολα δεδομένων που ανακτήθηκαν. Πιο συγκεκριμένα, στην αρχή του κάθε κεφαλαίου γίνονται εισαγωγικές παρατηρήσεις για τις μεθοδολογίες που θα παρουσιαστούν, ενώ στη συνέχεια γίνεται ανασκόπηση της εφαρμογής και των αποτελεσμάτων των αλγορίθμων στο εκάστοτε από τα σύνολα. Αξίζει να σημειωθεί κάπου εδώ ότι, στο κάθε ένα σύνολο από αυτά, πραγματοποιούνται διαφορετικές προσεγγίσεις προ-επεξεργασίας και, διαφορετικές προσεγγίσεις ποσοστών συνόλων εκπαίδευσης.

Στο όγδοο κεφάλαιο της πτυχιακής εργασίας, παρουσιάζονται οι γενικές αξιολογήσεις και τα συμπεράσματα μετά την εφαρμογή των αλγορίθμων σε αυτές τις τρείς μελέτες περιπτώσεων των συνόλων δεδομένων που χρησιμοποιήθηκαν. Συγκεκριμένα, σε αυτό το κεφάλαιο επιδιώκεται η περιγραφή της πλήρους εικόνας που σχηματίστηκε σχετικά με αυτά, μετά την ολοκλήρωση της εφαρμογής των μεθόδων μηχανικής μάθησης. Στο τέλος αυτού του κεφαλαίου, γίνεται και μια συγκριτική αξιολόγηση μεταξύ των αποτελεσμάτων που δόθηκαν για τα τρία αυτά σύνολα δεδομένων.

4 Περιβάλλοντα υλοποίησης μελετών

Στο πλαίσιο της παρούσας εργασίας, πραγματοποιήθηκε μελέτη σε σύνολα δεδομένων τα οποία συλλέχθηκαν από το διαδίκτυο, όπως ήδη αναφέρθηκε. Τα σύνολα αυτά υπέστησαν επεξεργασία μέσω αλγορίθμων μηχανικής μάθησης οι οποίοι εφαρμόστηκαν με τη γλώσσα προγραμματισμού Python. Η εκπαίδευση των αλγορίθμων που χρησιμοποιήθηκαν, πραγματοποιήθηκε με επιβλεπόμενο τρόπο. Στη συνέχεια, γίνεται μια αναλυτική περιγραφή των περιβαλλόντων υλοποίησης των μελετών, συμπεριλαμβανομένων των εργαλείων και των συνόλων δεδομένων που χρησιμοποιήθηκαν.

4.1 Η γλώσσα Python

Η γλώσσα προγραμματισμού Python πρόκειται για μια διερμηνευτική γλώσσα υψηλού επιπέδου η οποία έκανε πρώτη φορά την εμφάνιση της το 1991. Η ικανότητα της να προσφέρει τον αντικειμενοστραφή και το δομημένο τρόπο προσέγγισης ταυτόχρονα για την επίλυση προβλημάτων, είναι αυτό που την καθιστά ξεχωριστή συγκριτικά με άλλες γλώσσες προγραμματισμού. Το πεδίο των εφαρμογών της γλώσσας αυτής για την ανάπτυξη προγραμμάτων ποικίλει. Σύμφωνα με μια έρευνα που διεξήχθη το 2020 από τη δημοφιλή κοινότητα της Kaggle, η Python πρόκειται για μια γλώσσα η οποία προτιμάται, μεταξύ άλλων, για την ανάλυση δεδομένων και για την κατασκευή προγραμμάτων μηχανικής μάθησης. Τα αποτελέσματα αυτά της έρευνας βρίσκονται διαθέσιμα στον ακόλουθο ηλεκτρονικό σύνδεσμο: https://www.kaggle.com/kaggle-survey-2020.

Όσον αναφορά στα περιβάλλοντα ανάπτυξης (Integrated Development Environments - IDE's) της συγκεκριμένης γλώσσας, υπάρχουν πάρα πολλές επιλογές. Ανάλογα, βέβαια, με τις μεθοδολογίες και τις προτιμήσεις του κάθε προγραμματιστή. Στην περίπτωση της παρούσας Πτυχιακής Εργασίας προτιμήθηκε το περιβάλλον Jupyter NoteBook. Το περιβάλλον αυτό πρόκειται για ένα IDE, το οποίο στοχεύει στη δια-δραστικότητα με το χρήστη με έναν ιδιαίτερο τρόπο, όντας σε θέση να εμφανίζει τμηματικά τα αποτελέσματα του κώδικα. Η υλοποίηση της Python που χρησιμοποιήθηκε στο συγκεκριμένο περιβάλλον ανάπτυξης, είναι η Python 3.8.8. Η εγκατάσταση του περιβάλλοντος αυτού πραγματοποιήθηκε αυτόματα με την εγκατάσταση του πακέτου Anaconda, το οποίο

βρίσκεται διαθέσιμο δωρεάν στην ιστοσελίδα του. Το Anaconda, είναι ένα framework που περιέχει μια πληθώρα σημαντικών εργαλείων για την ανάλυση δεδομένων. Για την είσοδο μας στο περιβάλλον του Jupyter, αρκεί να εισάγουμε την εντολή «jupyter notebook», στην διεπαφή γραμμής εντολών (CLI) του Anaconda. Στη συνέχεια, ο υπολογιστής μας, ξεκινάει μια σύνδεση σε μια συγκεκριμένη θύρα (port) για την εισαγωγή μας στο περιβάλλον αυτό του IDE. Τέλος, το περιβάλλον του Jupyter μπορεί να βρεθεί εύκολα στον localhost του Browser μας, συνήθως στο port:8080.

Η γλώσσα Python περιέχει ένα σύνολο από «βιβλιοθήκες» (libraries) σχετιζόμενες με την ανάλυση δεδομένων και τη μηχανική μάθηση, οι οποίες μπορούν να βοηθήσουν στη γρήγορη προσέγγιση για τη δημιουργία οποιουδήποτε πρότζεκτ. Οι βιβλιοθήκες αυτές, είναι έτοιμα κομμάτια κώδικα τα οποία μπορεί κανείς να εισαγάγει στο πρόγραμμά του, με την πληκτρολόγηση μερικών μόνο εντολών. Οι βιβλιοθήκες που χρησιμοποιήθηκαν για τις εμπειρικές μελέτες της παρούσας εργασίας, αποτελούν όλες κομμάτια του Anaconda. Συνεπώς, δε χρειάστηκε η εγκατάσταση κάποιου περαιτέρω προγράμματος ή πακέτου. Στη συνέχεια γίνεται αναφορά στις βασικότερες βιβλιοθήκες που χρησιμοποιήθηκαν.

- Pandas: Το Pandas είναι μια ανοιχτή βιβλιοθήκη λογισμικού η οποία χρησιμοποιείται σε πολλούς τομείς της ανάλυσης δεδομένων μέσω της γλώσσας Python. Η κύρια λειτουργία της αποτελεί τη δημιουργία ειδικών δομών δεδομένων και λειτουργιών, για τη διαχείριση πινάκων και συνόλων δεδομένων.

- NumPy: Το NumPy είναι άλλη μια ανοιχτή βιβλιοθήκη λογισμικού της γλώσσας Python η οποία χρησιμοποιείται κυρίως για τη συγκέντρωση και διαχείριση μεγάλων συστοιχιών δεδομένων και πινάκων. Επιπλέον, το NumPy προσφέρει ένα σύνολο από έτοιμα μαθηματικά εργαλεία τα οποία μπορούν να χρησιμοποιηθούν για ανάλυση σε πίνακες, λίστες και, σύνολα δεδομένων.

- Matplotlib: Το Matplotlib είναι μια βιβλιοθήκη η οποία παρέχει στο χρήστη τη δυνατότητα γραφικής απεικόνισης και σχεδιασμού διαφόρων ειδών δεδομένων. Μέσω αυτής της βιβλιοθήκης, μπορούν να κατασκευαστούν διάφορες σημαντικές απεικονίσεις στους επιστημονικούς τομείς της ανάλυσης των δεδομένων, όπως για παράδειγμα τα διαγράμματα διασποράς, τα ιστογράμματα κλπ.

- Seaborn: Το Seaborn είναι μια βιβλιοθήκη η οποία έχει ίδια πεδία χρήσης με το Matplotlib. Η χρήση της, εστιάζει κυρίως στο σχεδιασμό και τη γραφική απεικόνιση δεδομένων.

- Scikit-Learn: Το Scikit-Learn είναι μια βιβλιοθήκη της Python, η οποία περιέχει ένα σύνολο από -έτοιμους για χρήση- αλγορίθμους μηχανικής μάθησης και εξόρυξης πληροφορίας.

4.2 Οι αλγόριθμοι στην βιβλιοθήκη Scikit-Learn

Οι αλγόριθμοι που χρησιμοποιήθηκαν από τη βιβλιοθήκη Scikit-Learn της Python περιορίστηκαν μόνο σε αλγορίθμους κατηγοριοποίησης και πρόβλεψης. Αξίζει να σημειωθεί κάπου εδώ ότι, η βιβλιοθήκη αυτή χωρίζεται σε επιμέρους κλάσεις, κατατάσσοντας τους διάφορους αλγόριθμους που περιέχει σε επιμέρους κατηγορίες για την καλύτερη αναζήτηση τους. Στην παρούσα ενότητα θα αναλυθεί ο τρόπος με τον οποίο οι αλγόριθμοι οι οποίοι παρουσιάστηκαν σε προηγούμενα μέρη της εργασίας, μπορούν να αναζητηθούν στη βιβλιοθήκη Scikit-Learn. Επιπλέον, στην παρούσα ενότητα γίνεται αναφορά και στις παραμέτρους που δόθηκαν στον κάθε έναν αλγόριθμο από αυτούς.

Linear Regression

Η γραμμική παλινδρόμηση (linear regression), όπως ήδη αναφέρθηκε, πρόκειται για μια στατιστική συνάρτηση πρόβλεψης η οποία εκφράζει τη σχέση μεταξύ μιας εξαρτημένης μεταβλητής με μια (απλή παλινδρόμηση), ή και περισσότερες (πολλαπλή παλινδρόμηση) ανεξάρτητες μεταβλητές.

Η γραμμική παλινδρόμηση μπορεί να βρεθεί στη βιβλιοθήκη Scikit-Learn ως sklearn.linear_model.linearRegression.

Για τη χρήση του συγκεκριμένου αλγορίθμου, δε χρησιμοποιήθηκαν κάποιες παράμετροι.

Logistic Regression

Η λογιστική παλινδρόμηση (logistic regression), πρόκειται για ένα γραμμικό μοντέλο το οποίο χρησιμοποιείται για δυαδική κατηγοριοποίηση.

Η λογιστική παλινδρόμηση μπορεί να βρεθεί στην κλάση της βιβλιοθήκης Scikit-Learn ως sklearn.linear_model.LogisticRegression.

Η μια και μόνη παράμετρος που δόθηκε στο συγκεκριμένο αλγόριθμο είναι η max_iter. Σύμφωνα με το εγχειρίδιο χρήσης του Scikit-Learn, η παράμετρος αυτή αφορά τον αριθμό επαναλήψεων που θα κάνει ο αλγόριθμος στο σύνολο δεδομένων μέχρι να βρει ένα βέλτιστο πιθανό «αποτέλεσμα». Παραδειγματικά, το αποτέλεσμα αυτό στην λογιστική παλινδρόμηση μπορεί να αφορά την εύρεση των καλύτερων σταθερών a και b που θα αντιπροσωπεύσουν την συνάρτησή της. Η τιμή που δόθηκε στην παράμετρο αυτή ήταν max_iter = 3000.

k-Nearest Neighbors

Ο αλγόριθμος των k-κοντινότερων γειτόνων (k-nearest neighbours), πρόκειται για έναν αλγόριθμο ο οποίος χρησιμοποιείται κατά κόρον τόσο για δυαδική όσο και για πολλαπλή κατηγοριοποίηση.

Για τη χρήση του συγκεκριμένου αλγορίθμου μέσα στο πρόγραμμα, αρκεί να γίνει η αναζήτηση του ως sklearn.neighbors.KNeighborsClassifier.

Οι παράμετροι που χρησιμοποιήθηκαν στον αλγόριθμο αυτόν είναι οι: α) n-neighbors και β) metric. Στη συνέχεια γίνεται επεξήγηση αυτών:

- N-neighbors*:* Η παράμετρος αυτή αφορά στο πλήθος k των γειτονικών στοιχείων που θα αναζητηθούν από τον αλγόριθμο. Οι τιμές που χρησιμοποιήθηκαν για αυτήν την παράμετρο ήταν τυχαίες και, βρίσκονταν στο διάστημα 5 μέχρι 10.
- Metric: Η συγκεκριμένη παράμετρος πρόκειται για τη «μετρική απόστασης» που θα χρησιμοποιηθεί από τον αλγόριθμο κατά τη διάρκεια της λειτουργίας του. Η μετρική απόστασης η οποία χρησιμοποιήθηκε ήταν αυτή της απόστασης του Μινκόβσκι που αναλύθηκε νωρίτερα.

Support Vector Machines

Οι μηχανές διανυσμάτων υποστήριξης (support vector machines), πρόκειται για έναν αλγόριθμο κατηγοριοποίησης ο οποίος επιτελεί ένα γραμμικό διαχωρισμό στις μεταβλητές των δεδομένων ενός συνόλου. Έτσι, οποιοδήποτε μη ετικετοποιημένο δεδομένο δοθεί στον αλγόριθμο θα κατηγοριοποιηθεί στην πιο κατάλληλη κλάση, βάσει των γνωρισμάτων του.

Για τη χρήση του συγκεκριμένου αλγορίθμου μέσα στο πρόγραμμα, αρκεί να γίνει η αναζήτηση του με την ονομασία sklearn.svm.SVC.

Για το συγκεκριμένο αλγόριθμο, δεν δόθηκαν κάποιες παράμετροι.

Neural Networks

Τα νευρωνικά δίκτυα (neural networks) πρόκειται για μια ακόμα αλγοριθμική τεχνική η οποία χρησιμοποιείται για κατηγοριοποίηση. Το νευρωνικό δίκτυο που χρησιμοποιήθηκε από την βιβλιοθήκη Scikit-Learn, μπορεί να αναζητηθεί στην κλάση sklearn.neural_network ως MLPClassifier. Το νευρωνικό δίκτυο αυτό, αποτελεί ένα εμπρός τροφοδότησης δίκτυο με μια έξοδο (δίκτυο τύπου Perceptron).

Οι παράμετροι που δόθηκαν για το συγκεκριμένο αλγόριθμο είναι οι: α) activation, β) hidden_layer_sizes και γ) max_iter. Στη συνέχεια, γίνεται επεξήγηση αυτών:

- Activation: Η παράμετρος αυτή αφορά στο είδος της συνάρτησης ενεργοποίησης που θα χρησιμοποιηθεί από τους νευρώνες. Για την παρούσα πτυχιακή εργασία, χρησιμοποιήθηκε η συνάρτηση ενεργοποίησης ReLu. Σύμφωνα με τον Sharma (2017), η συνάρτηση ReLu δίνει σαν τιμή εξόδου του νευρώνα την τιμή του αθροιστή, αν εφόσον αυτή είναι θετική, ενώ σε διαφορετική περίπτωση δίνει την τιμή 0. Συμβολίζεται ως ακολούθως:

$$R(V) = max(0, V)$$

- Hidden layer sizes: Η συγκεκριμένη παράμετρος αφορά στο πλήθος των νευρώνων που περιέχονται σε ένα κρυφό επίπεδο. Η τιμή που χρησιμοποιήθηκε ήταν τυχαία, στο διάστημα 50 μέχρι 100.
- Max iter: Η τιμή που δόθηκε στην παράμετρο αυτή ήταν max_tier = 3000.

Naïve Bayes

Ο απλοϊκός Bayes (naïve Bayes) πρόκειται για έναν αλγόριθμο κατηγοριοποίησης ο οποίος είναι βασισμένος στο γνωστό θεώρημα του Bayes. Το βασικό σκεπτικό του συγκεκριμένου αλγόριθμου είναι η εύρεση της πιθανότητας ενός στιγμιότυπου να ανήκει ή να μην ανήκει σε μια κλάση και, να το ταξινομήσει εκεί που υπερισχύουν οι πιθανότητες.

Ο αλγόριθμος αυτός συναντάται στη βιβλιοθήκη Scikit-Learn με το όνομα Gaussian Naïve Bayes. Συγκεκριμένα, η αναζήτηση του αλγορίθμου αυτού μπορεί να γίνει ως sklearn.naive_bayes.GaussianNB.

Για το συγκεκριμένο αλγόριθμο δε χρησιμοποιήθηκαν κάποιες παράμετροι.

Decision Tree Algorithm

Οι αλγόριθμοι δένδρων αποφάσεων (decision tree algorithms), είναι μια πολύ διαδεδομένη τεχνική στο χώρο της μηχανικής μάθησης και της εξόρυξης πληροφορίας, η οποία χρησιμοποιείται, ως επι των πλείστων, για κατηγοριοποίηση.

Ο συγκεκριμένος αλγόριθμος μπορεί να βρεθεί στη βιβλιοθήκη Scikit-Learn με την ονομασία Decision Tree Classifier. Η αναζήτηση του στις κλάσεις της βιβλιοθήκης Scikit-Learn, μπορεί να γίνει ως sklearn.tree.DecisionTreeClassifier.

Η βασική παράμετρος που δόθηκε στο συγκεκριμένο αλγόριθμο ονομάζεται criterion και αφορά στη μετρική διαχωρισμού για την κατασκευή του δένδρου. Η τιμή που δόθηκε σε αυτήν την παράμετρο ήταν αυτή της εντροπίας (entropy).

Mini-Max Regularization

Η διαδικασία Mini-Max της κανονικοποίησης (mini-max regularization), πρόκειται για μια τεχνική προ-επεξεργασίας των δεδομένων η οποία χρησιμοποιείται για τη μεταμόρφωση τους σε ένα μικρότερο διάστημα αριθμών. Ο συγκεκριμένος αλγόριθμος καθίσταται ιδιαίτερης σημασίας, καθώς έπειτα, οι τεχνικές επεξεργασίας δε διαχειρίζονται μεγάλες αριθμητικές τιμές.

Η εύρεση του συγκεκριμένου αλγορίθμου στη βιβλιοθήκη Scikit-Learn μπορεί να γίνει ως sklearn.preprocessing.MinMaxScaler.

Στον αλγόριθμο αυτό, δε δόθηκαν κάποιες συγκεκριμένες παράμετροι. Κάπου εδώ αξίζει να σημειωθεί όμως ότι, ως default διάστημα μεταμόρφωσης των τιμών του αλγορίθμου αυτού, ήταν το ανοικτό διάστημα [0,1].

Standard Scaller

Ο αλγόριθμος Standard Scaler πρόκειται επίσης για έναν αλγόριθμο μεταμόρφωσης των δεδομένων. Η βασική ιδέα του συγκεκριμένου αλγορίθμου είναι η κατασκευή ενός συνόλου δεδομένων με μέση τιμή (mean) της κάθε μεταβλητής ίσης με το 0 και, τυπική απόκλιση (standard deviation) ίση με το 1. Η διαδικασία αυτή λαμβάνει χώρα σε όλες τις μεταβλητές του συνόλου που επεξεργαζόμαστε. Σύμφωνα με το εγχειρίδιο χρήσης του αλγορίθμου Standard Scaler από την ιστοσελίδα του Scikit-Learn, ο υπολογισμός της νέας τιμής της μεταβλητής ενός στιγμιότυπου στο σύνολο γίνεται με τον εξής τύπο:

$$z = \frac{(x - u)}{S}$$

Όπου z είναι η καινούργια τιμή της μεταβλητής ενός στιγμιότυπου με αρχική τιμή x και, το u και το S πρόκειται για το μέσο όρο και την τυπική απόκλιση αντίστοιχα της στήλης αυτής που επεξεργαζόμαστε.

Ο συγκεκριμένος αλγόριθμος μπορεί να βρεθεί στη βιβλιοθήκη Scikit-Learn ως sklearn.preprocessing.StandardScaler.

Στο συγκεκριμένο αλγόριθμο δε δόθηκαν κάποιες συγκεκριμένες παράμετροι.

Train Test Split

Το Train Test Split πρόκειται για μια βοηθητική τεχνική της βιβλιοθήκης Scikit-Learn η οποία χρησιμεύει στο διαχωρισμό του αρχικού συνόλου των δεδομένων σε δύο υποσύνολα. Το πρώτο από αυτά, αφορά στο σύνολο εκπαίδευσης (train set) που δίνεται στους αλγόριθμους μηχανικής μάθησης, ενώ το δεύτερο αφορά στο σύνολο δοκιμής τους (test set).

Η συγκεκριμένη τεχνική μπορεί να βρεθεί στη βιβλιοθήκη Scikit-Learn ως sklearn.model_selection.train_test_split.

Οι παράμετροι που δόθηκαν στη συγκεκριμένη τεχνική είναι α) test_size και β) το random_state. Στη συνέχεια, γίνεται επεξήγηση αυτών:

- Test_size*:* Η παράμετρος αυτή αφορά στο ποσοστό διαχωρισμού του αρχικού συνόλου των δεδομένων σε σύνολο εκπαίδευσης και σύνολο δοκιμής. Οι προσεγγίσεις που πραγματοποιήθηκαν διακρίνονται σε δύο διαφορετικές εκδοχές. Η πρώτη αφορά το train_size = 0.2 (δηλαδή χρήση του 80% του αρχικού συνόλου για εκπαίδευση) και η δεύτερη είναι το train_size = 0.33 (δηλαδή χρήση του 67% του αρχικού συνόλου για εκπαίδευση).
- Random_state*:* Η συγκεκριμένη παράμετρος αφορά στην επιλογή τυχαίων στιγμιότυπων για εκπαίδευση και δοκιμή, κάθε φορά που το train_test_split τίθεται σε λειτουργία από το πρόγραμμα. Επιδιώκοντας τη δίκαιη συγκριτική αξιολόγηση μεταξύ των αλγόριθμων, στη συγκεκριμένη παράμετρο δόθηκε ως τιμή random_state = 0. Ως αποτέλεσμα, ο διαχωρισμός του συνόλου γίνεται μία μόνο φορά, ανεξαιρέτως το πόσες φορές θα τρέξει το πρόγραμμα.

Metrics

Το Metrics πρόκειται για ένα σύνολο από βοηθητικά εργαλεία της βιβλιοθήκης Scikit-Learn, τα οποία χρησιμοποιούνται για την αξιολόγηση της εφαρμογής αλγορίθμων μηχανικής μάθησης και εξόρυξης πληροφορίας. Οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση των αποτελεσμάτων των αλγορίθμων στο εμπειρικό μέρος της παρούσας πτυχιακής, ήταν αυτές της ορθότητας (accuracy) και της ακρίβειας (precision). Ειδικότερα:

- Ορθότητα*:* Ο υπολογισμός της ορθότητας ενός αλγορίθμου πραγματοποιείται με υπολογισμούς σύμφωνα με τα αποτελέσματα των αλγορίθμων στο σύνολο δοκιμής. Το ακριβές όνομα της μεθόδου αυτής στην κλάση Metrics του Scikit-Learn, είναι accuracy_score. Οι παράμετροι που δίνονται σε αυτήν τη μέθοδο για την εξαγωγή του αποτελέσματος αποτελούν οι *προβλέψεις* που πραγματοποίησε ο αλγόριθμος για την ταξινόμηση των δεδομένων του συνόλου δοκιμής, έναντι των *πραγματικών αξιών* των στοιχείων αυτών.
- Ακρίβεια: Όπως και στην περίπτωση της ορθότητας, ο υπολογισμός της ακρίβειας ενός αλγορίθμου πραγματοποιείται σύμφωνα με τα αποτελέσματα των αλγορίθμων στο σύνολο δοκιμής. Ως παράμετροι στη συγκεκριμένη μέθοδο, δίνονται οι *προβλέψεις* των κλάσεων του συνόλου δοκιμής, έναντι των *πραγματικών αξιών* τους. Η συγκεκριμένη μέθοδος συναντάται ως precision_score στην κλάση Metrics.

4.3 Σύνολα δεδομένων

Για την εφαρμογή των αλγορίθμων μηχανικής μάθησης επιλέχθηκαν τρία σύνολα δεδομένων τα οποία ανακτήθηκαν από πηγές στο διαδίκτυο. Τα σύνολα αυτά, πρόκειται για σύνολα στα οποία επιδιώκεται να πραγματοποιηθεί η εφαρμογή διάφορων αλγορίθμων κατηγοριοποίησης και, η εφαρμογή του αλγόριθμου της γραμμικής παλινδρόμησης. Στη συνέχεια, ακολουθεί η επεξήγηση αυτών.

4.3.1 Σύνολο δεδομένων: Student Performance Data Set

Το πρώτο σύνολο δεδομένων που επιλέχθηκε ονομάζεται «Student Performance Data Set» και αφορά στη γενική εικόνα της απόδοσης διάφορων μαθητών στις εξεταστικές τους περιόδους βάσει διαφόρων άλλων γνωρισμάτων. Η ανάκτηση του συγκεκριμένου συνόλου πραγματοποιήθηκε δωρεάν από το διαδικτυακό ιστότοπο του UCI Machine Learning Repository. Η ακριβής ηλεκτρονική διεύθυνση του συγκεκριμένου συνόλου δεδομένων, μπορεί να βρεθεί στον ακόλουθο ηλεκτρονικό σύνδεσμο:

https://archive.ics.uci.edu/ml/datasets/Student+Performance.

Το σύνολο αυτό των δεδομένων, συγκεντρώθηκε στο πλαίσιο μιας μελέτης από το Πανεπιστήμιο του Μίνχο, σε δύο ξεχωριστά σχολεία δευτεροβάθμιας εκπαίδευσης της Πορτογαλίας. Σκοπός της συγκεκριμένης μελέτης αποτέλεσε η καταγραφή διάφορων μεταβλητών που επηρεάζουν την απόδοση των μαθητών στις βαθμολογήσεις τριών περιόδων στο μάθημα των Μαθηματικών. Βάσει των συντακτών του άρθρου στον ηλεκτρονικό σύνδεσμο που αναφέρθηκε, οι πιο κρίσιμες μεταβλητές που συμβάλουν στη βαθμολογία της τρίτης περιόδου αποτελούν, κατά βάση, οι βαθμολογίες της πρώτης και της δεύτερης περιόδου. Όλες οι βαθμολογίες αυτές παίρνουν τιμές στο διάστημα από μηδέν μέχρι είκοσι. Στο συγκεκριμένο data set συμπεριλαμβάνονται αρκετές μεταβλητές, όπως για παράδειγμα ο χρόνος που αφιέρωσε ένας μαθητής για να διαβάσει, το φύλο του, την πιθανή εξωσχολική του εκπαίδευση, την οικογενειακή του κατάσταση, τις προηγούμενες αποτυχίες του σε μαθήματα κλπ. Συγκεκριμένα, το σύνολο των δεδομένων αυτό αποτελείται από 33 συνολικά μεταβλητές και 395 εγγραφές στιγμιότυπων.

4.3.2 Σύνολο δεδομένων: Diabetes Dataset

Το δεύτερο σύνολο δεδομένων που επιλέχθηκε για μελέτη, αφορά ένα σύνολο με διάφορες καταγεγραμμένες μεταβλητές που μπορούν να επηρεάσουν στη θετική ή την αρνητική διάγνωση της νόσου του διαβήτη. Το όνομα με το οποίο βρέθηκε στο διαδίκτυο είναι «Pima Indians Diabetes Dataset». Η ανάκτηση του συγκεκριμένου συνόλου πραγματοποιήθηκε από το διαδικτυακό ιστότοπο της Kaggle. Ο διαδικτυακός σύνδεσμος του συγκεκριμένου συνόλου ακολουθεί https://www.kaggle.com/uciml/pima-indians-diabetes-database?select=diabetes.csv.

Το σύνολο των δεδομένων αυτό, προέρχεται από μελέτες που διεξήχθησαν από το Εθνικό Ινστιτούτο Διαβήτη και Παθήσεων του Ύπατος και των Νεφρών (National Institute of Diabetes and Digestive and Kidney Diseases – NIDDK). Σκοπός της δημιουργίας του συγκεκριμένου συνόλου, αποτέλεσε η καταγραφή διάφορων μεταβλητών που μπορούν να επηρεάσουν την εμφάνιση της νόσου του διαβήτη σε γυναίκες, μεγαλύτερες των είκοσι-ενός ετών. Σε αντίθεση με το προηγούμενο σύνολο δεδομένων με την απόδοση των μαθητών, οι μεταβλητές στο Diabetes Data Set είναι αρκετά λιγότερες. Συγκεκριμένα, το σύνολο αυτό περιέχει 9 μεταβλητές και 768 στιγμιότυπα. Οι καταγεγραμμένες μεταβλητές του συγκεκριμένου συνόλου αναφέρονται στη συνέχεια:

- Το σύνολο από τις εγκυμοσύνες που είχε κάποια γυναίκα (Pregnancies),
- Τα επίπεδα γλυκόζης στο αίμα της (Glucose),
- Τα επίπεδα της ινσουλίνης στο αίμα της (Insulin),
- Το δείκτη μάζας του σώματός της (Body Mass Index - BMI),
- Την αρτηριακή πίεση της (Blood Pressure),
- Τη σκληρότητα του δέρματός της (Skin Thickness),
- Την πιθανή ύπαρξη κληρονομικότητας της νόσου από κάποιο συγγενικό πρόσωπο (Diabetes Pedigree Function - DPF) και,
- Την ηλικία της (Age).

Ως στόχος, μέσω της ανάλυσης των προαναφερθέντων μεταβλητών, αποτελεί η πρόβλεψη και η κατηγοριοποίηση των δεδομένων στη σωστή κλάση του συνόλου αυτού. Η κλάση αυτή αποτελεί την $9^{η}$ μεταβλητή του συνόλου και, αφορά στο αν μια γυναίκα είναι θετική ή αρνητική στη νόσο του διαβήτη. Στο σύνολο αυτό των δεδομένων, η κλάση συμβολίζεται ως Outcome.

4.3.3 Σύνολο δεδομένων: Loan Predication Data Set

Το τρίτο σύνολο δεδομένων που επιλέχθηκε για ανάλυση και μελέτη, ονομάζεται «Loan Predication Data Set». Το συγκεκριμένο σύνολο δεδομένων αφορά σε καταγραφές οι οποίες συλλέχθηκαν μέσω ερωτηματολογίων από μια ασφαλιστική εταιρία προς τους πελάτες της. Στόχος ήταν να αυτοματοποιηθεί ο τρόπος με τον οποίο δίνονται δάνεια στους πελάτες, βάσει διαφόρων μεταβλητών και, η ασφαλιστική εταιρία να στοχεύσει «καλύτερα» στους πελάτες που είναι πιο πιθανό να είναι όντως θετική στο να πάρουν κάποιο δάνειο. Το σύνολο δεδομένων αυτό ανακτήθηκε επίσης από την ιστοσελίδα Kaggle, με τον ακριβής ηλεκτρονικό του σύνδεσμο να ακολουθεί: https://www.kaggle.com/ninzaami/loan-predication.

Το συγκεκριμένο σύνολο, αποτελείται από 13 συνολικά μεταβλητές και 615 εγγραφές στιγμιότυπων. Οι 12 από τις μεταβλητές αυτές, χρησιμοποιούνται για την εύρεση της κλάσης του συνόλου, της 13^{ης} μεταβλητής δηλαδή. Η κλάση αυτή του συνόλου, αφορά στο αν το αίτημα ενός ατόμου για λήψη κάποιου ασφαλιστικού δανείου έγινε δεκτό, ή όχι, από την ασφαλιστική εταιρία. Όπως είναι προφανές, το συγκεκριμένο πρόβλημα, όπως και αυτό του Diabetes Data Set, αποτελεί ένα πρόβλημα δυαδικής κατηγοριοποίησης. Στη συνέχεια, παρουσιάζονται οι μεταβλητές που απαρτίζουν το σύνολο αυτό:

- Το φύλο του ατόμου που μελετάται στην εκάστοτε εγγραφή (Gender),
- Το αν το άτομο αυτό είναι παντρεμένο ή όχι (Married),
- Το πλήθος των «εξαρτήσεων» που μπορεί να έχει στη ζωή του (Dependents),
- Το επίπεδο μόρφωσης του (Education),
- Το αν είναι αυτό-απασχολούμενο (Self-Employed),
- Το εισόδημά από την κύρια δουλειά του (Applicant Income),
- Το πιθανό συμπληρωματικό εισόδημα του από άλλους παράγοντες (CoApplicant Income),
- Το ποσό του δανείου που επιθυμεί (Loan Amount),
- Ένας συγκεκριμένος όρος που αφορά στο ποσό ξεχρεώματος του δανείου που θα λάβει (Loan Amount Term),
- Το μοναδικό ID (identification) που του αντιστοιχεί για τη λήψη του δανείου από την εταιρία (Loan ID),
- Το αν το άτομο διαθέτει πιστωτικό ιστορικό με άλλες ασφαλιστικές εταιρίες (Credit History) και,

- Ο τύπος της περιοχής που κατοικεί (Property Area).

Όπως παρατηρείται, όλες αυτές οι μεταβλητές, εκτός από αυτής του Loan ID, μπορούν να παίξουν έναν αρκετά σημαντικό ρόλο για την εύρεση της κλάσης του συνόλου, το αν το άτομο αυτό θα λάβει, ή όχι, το ασφαλιστικό δάνειο (Loan Status) δηλαδή.

5 Μελέτη περίπτωσης: Student Performance Data Set

5.1 Περιγραφή και προετοιμασία μελέτης

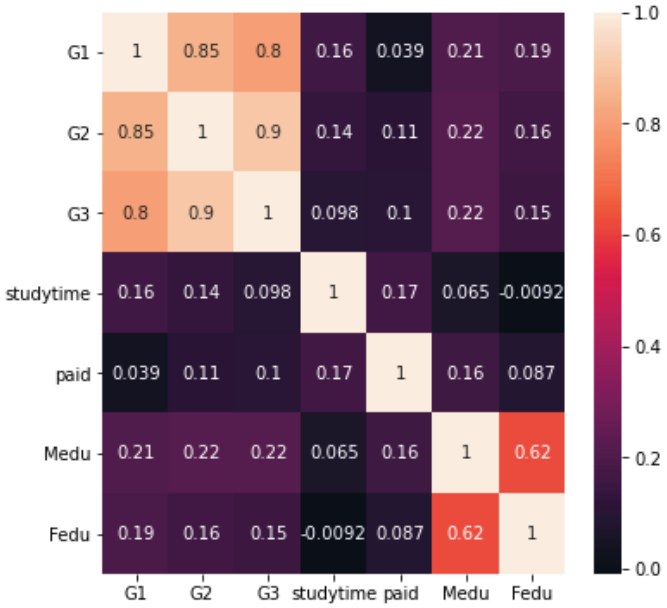
Οι μεταβλητές που περιέχονταν στο σύνολο Student Performance Data Set όπως αναφέρθηκε είναι αρκετές έως και πάρα πολλές. Αυτό το γεγονός μπορεί να αποτελέσει ιδιαίτερο πρόβλημα καθώς, αν διάφορες μεταβλητές που δε σχετίζονται μεταξύ τους χρησιμοποιηθούν για επεξεργασία, μπορεί να εξαχθούν λανθασμένα αποτελέσματα. Συνεπώς, χρησιμοποιήθηκαν μεταβλητές οι οποίες σχετίζονται περισσότερο με την προσωπική απόδοση των φοιτητών στα μαθήματά τους, χωρίς να λαμβάνονται υπόψη ιδιαίτερα οι «εξωγενείς» παράγοντες.

Για την υλοποίηση της μελέτης στο συγκεκριμένο σύνολο δεδομένων, χρησιμοποιήθηκαν οι αλγόριθμοι που αναφέρθηκαν στο «Κεφάλαιο 3» της εργασίας, με σκοπό την εύρεση του καλύτερου μοντέλου γραμμικής παλινδρόμησης και την εύρεση του καλύτερου κατηγοριοποιητή. Η διαδικασία που ακολούθησε πριν την εφαρμογή αλγορίθμων τμηματοποιείται με δύο προσεγγίσεις. Η πρώτη αφορά στη δημιουργία του κατάλληλου συνόλου δεδομένων για την εφαρμογή της γραμμικής παλινδρόμησης, ενώ η δεύτερη αφορά στη δημιουργία του κατάλληλου συνόλου για την εφαρμογή των αλγορίθμων κατηγοριοποίησης. Στη συνέχεια, ακολουθεί η διαδικασία που πραγματοποιήθηκε για τη δημιουργία αυτών.

5.1.1 Χρήση της γραμμικής παλινδρόμησης

Για να εφαρμοσθεί ο αλγόριθμος της γραμμικής παλινδρόμησης έπρεπε να επιλεχθούν οι κατάλληλες μεταβλητές για τη χρήση τους σε μια συνάρτηση η οποία θα επιτελούσε τις προβλέψεις της μεταβλητής G3. Η μεταβλητή G3 είναι η τελική βαθμολογία των μαθητών της τρίτης περιόδου. Η επιλογή των υπόλοιπων ανεξάρτητων μεταβλητών που χρησιμοποιήθηκαν, έγινε βάσει παρατήρησης του «πίνακα συσχέτισης» (correlation matrix) των δεδομένων, ο οποίος δημιουργήθηκε μέσω της βιβλιοθήκης Seaborn. Συγκεκριμένα, ο πίνακας αυτός απεικονίζει τη σημαντικότητα των σχέσεων μεταξύ των μεταβλητών του συνόλου που επεξεργαζόμαστε, βάσει του συντελεστή συσχέτισης Pearson. Αν η συσχέτιση μεταξύ δύο δεδομένων είναι «ισχυρή», παρουσιάζεται στο εκάστοτε σημείο του πίνακα με

ανοιχτό-φωτεινό χρώμα. Σε αντίθετη περίπτωση, δηλαδή αν η συσχέτιση μεταξύ δυο μεταβλητών δεν είναι «ισχυρή», αναπαρίσταται με ένα πιο έντονο και σκουρόχρωμο χρώμα. Τα αποτελέσματα των συσχετίσεων αυτών που βρέθηκαν, αναπαρίστανται μέσω αυτού του είδους πίνακα στην «Εικόνα 4» που ακολουθεί.



Εικόνα 4: Πίνακας συσχέτισης Student Performance Data Set.

Οι μεταβλητές οι οποίες επιλέχθηκαν ως ανεξάρτητες για την πρόβλεψη της G3, είναι οι ακόλουθες:

- G1 (ο βαθμός της πρώτης περιόδου),
- G2 (ο βαθμός της δεύτερης περιόδου),
- Paid (η ύπαρξη παρακολούθησης βοηθητικών μαθημάτων),
- Study Time (ο εβδομαδιαίος χρόνος που αφιέρωνε ένας μαθητής για να διαβάσει),
- Fedu (το επίπεδο μόρφωσης του πατέρα του μαθητή) και
- Medu (το επίπεδο μόρφωσης της μητέρας του μαθητή).

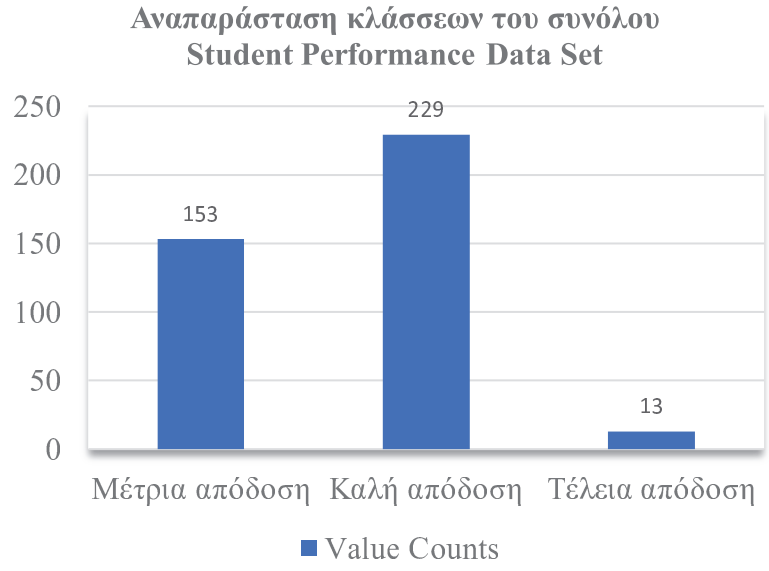
Συνεπώς, βάσει των προαναφερθέντων κεφαλαίων της πτυχιακής εργασίας, το μοντέλο που επιδιώκεται να δημιουργηθεί αποτελεί ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης και είναι της μορφής:

$$G3 = (\alpha_1 \cdot G1) + (\alpha_2 \cdot G2) + (\alpha_3 \cdot Paid) + (\alpha_4 \cdot Studytime) + (\alpha_5 \cdot Fedu) + (\alpha_6 \cdot Medu) + b$$

5.1.2 Χρήση των αλγορίθμων κατηγοριοποίησης

Επειδή το συγκεκριμένο σύνολο δεν περιείχε κάποιο χαρακτηριστικό εξόδου για να εφαρμοστούν οι αλγόριθμοι κατηγοριοποίησης, κατασκευάστηκε μια νέα στήλη η οποία θα αναπαρίστανε την κλάση του συνόλου αυτού[26]. Η στήλη αυτή που κατασκευάστηκε, αφορά την συνολική απόδοση των μαθητών στις σχολικές τους υποχρεώσεις, συμπεριλαμβανομένου των βαθμών όλων των εξεταστικών περιόδων τους. Συγκεκριμένα, η πρώτη τιμή της κλάσης του συνόλου αφορά στο ότι ο μαθητής έχει την τέλεια απόδοση στις σχολικές του υποχρεώσεις. Η δεύτερη, αφορά στη μέτρια απόδοση του μαθητή, ενώ η τρίτη αφορά στη μέτρια-κακή απόδοση. Η ταξινόμηση αυτή έγινε βάσει του μέσου όρου των τριών βαθμολογιών από τις εξεταστικές περιόδους του εκάστοτε μαθητή. Ειδικότερα, η κακή-μέτρια απόδοση των μαθητών κυμαίνεται στους βαθμούς από 0 μέχρι 9.4, η καλή απόδοση από 9.5 μέχρι 17.4 και, η τέλεια απόδοση από 17.5 μέχρι 20.

Στο «Διάγραμμα 1» παρατίθεται γραφικά το πλήθος των μαθητών που είχαν την τέλεια, την καλή και, τη -μέτρια προς κακή- απόδοση στις σχολικές τους υποχρεώσεις.



Διάγραμμα 1: Αναπαράσταση κλάσεων στο σύνολο Student Performance Data Set.

[26] Αξίζει να σημειωθεί κάπου εδώ ότι, παρόμοια μεθοδολογία ακολούθησαν και οι Cortez και Silva. Οι Cortez και Silva πρόκειται για δύο επιστήμονες από το Πανεπιστήμιο του Μίνχο, οι οποίοι κατέγραψαν, αλλά και συνέταξαν έρευνα για το σύνολο δεδομένων αυτό. Τα αποτελέσματα της έρευνας αυτής, μπορούν να βρεθούν στον εξής ηλεκτρονικό σύνδεσμό: http://www3.dsi.uminho.pt/pcortez/student.pdf.

Όπως παρατηρείται, οι περισσότεροι μαθητές έχουν καλή απόδοση στις σχολικές τους υποχρεώσεις, μερικοί είναι αυτοί οι οποίοι τα έχουν πάει άσχημα προς μέτρια, και πολύ λίγοι είναι αυτοί που έχουν αριστεύσει. Συγκεκριμένα, στο σύνολο των στιγμιότυπων που μελετήθηκαν, ο αριθμός των μαθητών που είχε τέλεια απόδοση στα μαθήματα του ήταν 13, καλή απόδοση είχαν 229 μαθητές, ενώ μέτρια απόδοση είχαν 153.

Βάσει του πίνακα συσχέτισης των δεδομένων που παρουσιάσθηκε νωρίτερα, η χρήση των μεταβλητών για την εκπαίδευση των αλγορίθμων της κατηγοριοποίησης είναι οι ίδιες, με αυτές που χρησιμοποιήθηκαν για τον αλγόριθμο της γραμμικής παλινδρόμησης. Η μόνη εξαίρεση αποτελεί η επιπλέον χρήση της μεταβλητής G3 και της κλάσης Performance, όπου και δημιουργήθηκε.

Επειδή το συγκεκριμένο πρόβλημα πρόκειται για ένα πρόβλημα πολλαπλής κατηγοριοποίησης (multi-class classification), δεν κατατέθηκε εφικτή η χρήση της λογιστικής παλινδρόμησης. Συνεπώς, οι αλγόριθμοι οι οποίοι εφαρμόστηκαν σε αυτήν την εκδοχή του συνόλου του Student Performance, είναι οι ακόλουθοι: α) k-Κοντινότεροι Γείτονες, β) Μηχανές Διανυσμάτων Υποστήριξης, γ) Απλοϊκός Μπάγιες, δ) Αλγόριθμος Δένδρου Αποφάσεων και ε) Νευρωνικό Δίκτυο.

5.1.3 Γενική μεθεδολογία της μελέτης

Στη συνέχεια, γίνεται ανάλυση των βημάτων που πραγματοποιήθηκαν για την εφαρμογή των αλγορίθμων στις δύο αυτές εκδοχές του συνόλου δεδομένων του Student Performance που μόλις αναφέρθηκαν.

1. Φόρτωση των συνόλων στην Python*:* Η φόρτωση των συνόλων πραγματοποιήθηκε μέσω της βιβλιοθήκης Pandas, ενώ στη συνέχεια χρησιμοποιήθηκαν κάποιες μέθοδοι της βιβλιοθήκης αυτής για την εύρεση ελλιπών τιμών. Σε καμία από τις δύο εκδοχές των συνόλων δεδομένων δεν υπήρξε κάποιο στοιχείο το οποίο να χρειαζόταν τροποποίηση ή διαγραφή, καθώς όλες οι εγγραφές τους ήταν άρτια συμπληρωμένες.

2. Διαχωρισμός των συνόλων για εκπαίδευση*:* Για το διαχωρισμό των συνόλων εκπαίδευσης έγινε χρήση της μεθόδου train_test_split της βιβλιοθήκης Scikit-Learn. Ειδικότερα, η μέθοδος αυτή χρησιμοποιήθηκε με δυο διαφορετικές προσεγγίσεις ως προς την εκπαίδευση των αλγορίθμων. Η πρώτη αποτελεί τη χρήση του 80% του

συνόλου για εκπαίδευση (train set) και το υπόλοιπο 20% για δοκιμή (test set). Βάσει της δεύτερης προσέγγισης, γίνεται χρήση του 67% του συνόλου για εκπαίδευση και το υπόλοιπο 33% για δοκιμή.

3. Εφαρμογή των αλγορίθμων*:* Η εφαρμογή των αλγορίθμων στα σύνολα αυτά μπορούν να διακριθούν σε μερικές διαφορετικές μεθοδολογίες εφαρμογής. Αρχικά, γίνεται εφαρμογή των αλγορίθμων στα σύνολα χωρίς να χρησιμοποιηθεί κάποιος ειδικός αλγόριθμος προ-επεξεργασίας. Επιδιώκοντας το βέλτιστο αποτέλεσμα των αλγορίθμων ως προς την ακρίβεια και την ορθότητά τους, στη συνέχεια γίνεται η εφαρμογή και η σύγκριση των αποτελεσμάτων με την προ-επεξεργασία του MinMax, του Standard Scaler ή και των δύο σε συνδυασμό.

4. Αξιολόγηση αποτελεσμάτων: Η αξιολόγηση και σύγκριση των αποτελεσμάτων των αλγορίθμων που εφαρμόστηκαν, έγιναν σύμφωνα με την ορθότητα και την ακρίβεια του κάθε ενός. Συγκεκριμένα, τα αποτελέσματα που αξιολογήθηκαν αποτελούν αυτά των προβλέψεων των αλγορίθμων για τα εκάστοτε σύνολα δοκιμής (test sets).

5.2 Εφαρμογή αλγορίθμων

Στη συγκεκριμένη ενότητα θα γίνει παρουσίαση των αποτελεσμάτων των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων του Student Performance. Όπως ήδη αναφέρθηκε, γίνεται χρήση δύο διαφορετικών εκδοχών του συνόλου αυτού. Το κάθε ένα από αυτά τροποποιήθηκε έτσι ώστε: α) να μπορέσει να γίνει εφαρμογή της γραμμικής παλινδρόμησης και β) να γίνει εφαρμογή διάφορων κατηγοριοποιητών.

Σε κάθε μία από αυτές τις περιπτώσεις γίνεται χρήση των αλγορίθμων σε ένα μικρότερο και ένα μεγαλύτερο σύνολο εκπαίδευσης, όπως αναφέρθηκε, καθώς επίσης πραγματοποιείται και μια μελέτη για την ανταπόκριση των αλγορίθμων σε διαφορετικές προσεγγίσεις προ-επεξεργασίας (MinMax και Standard Scaler). Αξίζει να σημειωθεί κάπου εδώ ότι, χρειάστηκε να πραγματοποιηθεί η μετατροπή από κατηγορηματικές και σε αριθμητικές τιμές, των μεταβλητών Paid και Performance.

Στις επόμενες ενότητες παρατίθενται τα αποτελέσματα της εφαρμογής των αλγορίθμων, ξεκινώντας από τη χρήση της γραμμικής παλινδρόμησης και καταλήγοντας στην εφαρμογή των αλγορίθμων κατηγοριοποίησης.

5.2.1 Γραμμική παλινδρόμηση

Τα αποτελέσματα της γραμμικής παλινδρόμησης ταξινομούνται και παρουσιάζονται στην παρούσα ενότητα σύμφωνα με τον πίνακα που ακολουθεί («Πίνακας 3»). Συγκεκριμένα, τα αποτελέσματα που παρουσιάζονται αφορούν στην ορθότητα του εκάστοτε μοντέλου (accuracy), ανάλογα με το σύνολο εκπαίδευσης που χρησιμοποιείται την εκάστοτε φορά.

Πίνακας 3: Αποτελέσματα αλγορίθμου γραμμικής παλινδρόμησης.

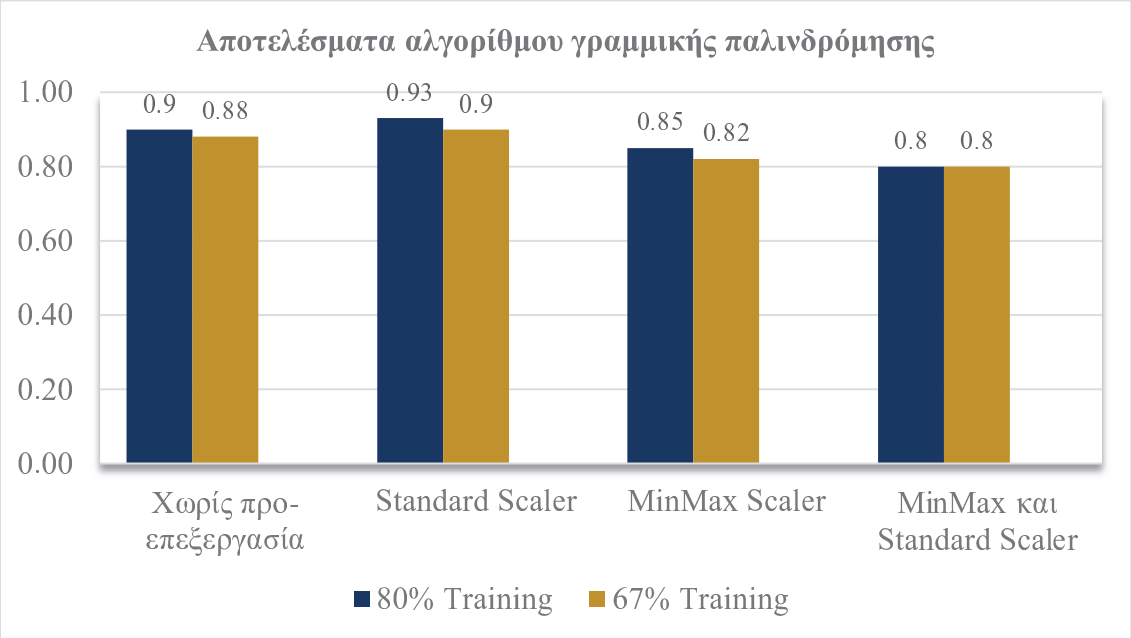
Αποτελέσματα αλγορίθμου γραμμικής παλινδρόμησης		
Προ-επεξεργασία	**Μεγάλο σύνολο εκπαίδευσης (80%)**	**Μικρό σύνολο εκπαίδευσης (67%)**
Χωρίς προ-επεξεργασία	0.90	0.88
Standard Scaler	0.93	0.90
MinMax Scaler	0.85	0.82
MinMax και Standard Scaler	0.80	0.80

Αρχικά, πραγματοποιήθηκε η εφαρμογή του αλγορίθμου της γραμμικής παλινδρόμησης στο σύνολο δεδομένων χωρίς κάποιου είδους προ-επεξεργασίας νωρίτερα. Τα αποτελέσματα της ορθότητας του αλγορίθμου της γραμμικής παλινδρόμησης χωρίς προ-επεξεργασία, έχουν τιμές 0.90 και 0.88 στο μεγαλύτερο και στο μικρότερο σύνολο εκπαίδευσης αντίστοιχα. Στην συνέχεια, έγινε εφαρμογή του αλγορίθμου Standard Scaler στο σύνολο, όπου τα αποτελέσματα που εξήχθησαν έπειτα, ήταν 0.93 για το μεγαλύτερο και 0.90 για το μικρότερο σύνολο εκπαίδευσης. Σαν τρίτη εφαρμογή, έγινε χρήση του αλγορίθμου προ-επεξεργασίας MinMax Scaler, με τον οποίο τα αποτελέσματα της γραμμικής παλινδρόμησης έφεραν ποσοστά 0.85 και 0.82 για το μεγαλύτερο και το μικρότερο σύνολο

εκπαίδευσης αντίστοιχα. Τέλος, έγινε χρήση της διπλής προ-επεξεργασίας MinMax και Standard Scaler στο σύνολο των δεδομένων αυτό. Τα αποτελέσματα της ορθότητας της γραμμικής παλινδρόμησης έπειτα από αυτή την προσέγγιση έφεραν ένα αρκετά μη αναμενόμενο αποτέλεσμα. Συγκεκριμένα, το αποτέλεσμα ήταν 0.80 και για τα δύο διαφορετικά ποσοστά συνόλου εκπαίδευσης, τιμή η οποία, είναι σημαντικά χαμηλότερη από τις προηγούμενες περιπτώσεις.

Όπως παρατηρείται, τα αποτελέσματα του μοντέλου της γραμμικής παλινδρόμησης δείχνουν να φέρουν αρκετά επιθυμητά ποσοστά ορθότητας ακόμα και χωρίς την χρήση κάποιου είδους προ-επεξεργασίας. Ο καλύτερος αλγόριθμος που έδειξε να βελτιώνει την τιμή της απόδοσης τους όμως, όσον αναφορά την ορθότητα, ήταν αυτός της προ-επεξεργασίας Standard Scaler.

Στο ακόλουθο διάγραμμα («Διάγραμμα 2»), αναπαρίστανται τα αποτελέσματα του «Πίνακα 3» με γραφικό τρόπο ως ραβδογράμματα, επιδιώκοντας την καλύτερη ερμηνεία τους μέσω της γραφικής απεικόνισής τους.



Διάγραμμα 2: Αποτελέσματα αλγορίθμου γραμμικής παλινδρόμησης.

5.2.2 Αλγόριθμοι κατηγοριοποίησης

Στις ακόλουθες ενότητες (5.2.3 μέχρι και 5.2.6), παρουσιάζονται τα αποτελέσματα που εξήχθησαν με την εφαρμογή αλγορίθμων κατηγοριοποίησης στις διαφορετικές περιπτώσεις της προ-επεξεργασίας και των διαφορετικών εκδοχών των συνόλων

εκπαίδευσης. Τα αποτελέσματα αυτά αφορούν στην ορθότητα (accuracy) και την ακρίβεια (precision) του εκάστοτε μοντέλου στις διαφορετικές προσεγγίσεις της προ-επεξεργασίας και της εκπαίδευσης. Επιδιώκοντας την καλύτερη κατανόηση και ερμηνεία των αποτελεσμάτων, τα αποτελέσματα των αλγορίθμων παρουσιάζονται γραφικά και ως ραβδογράμματα.

5.2.3 Εφαρμογή κατηγοριοποιητών χωρίς προ-επεξεργασία

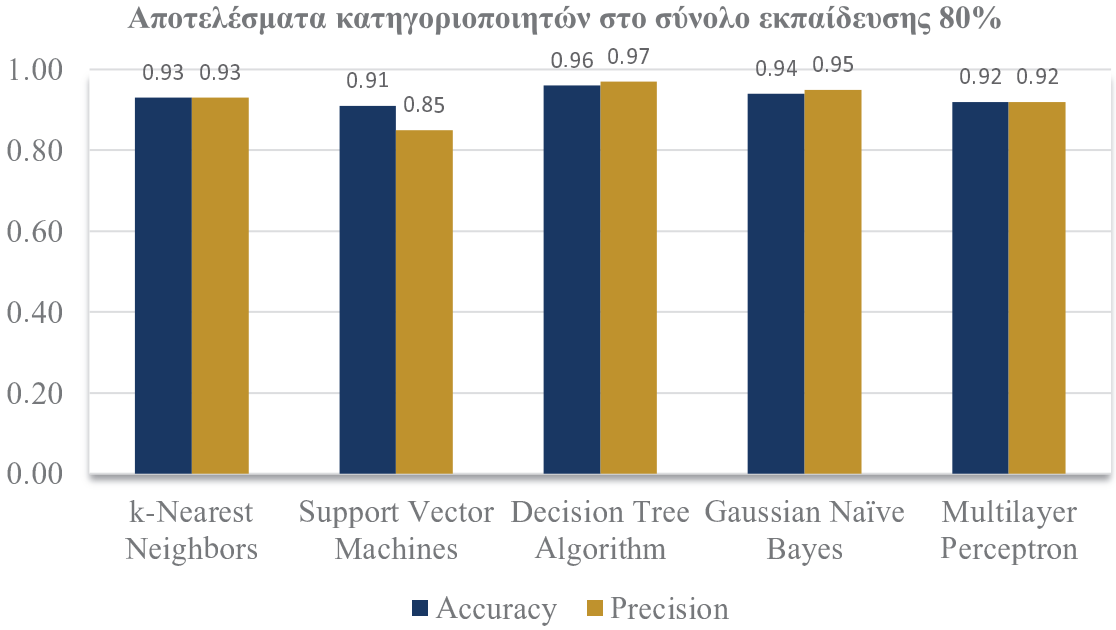
Τα αποτελέσματα των αλγορίθμων κατηγοριοποίησης στο μεγαλύτερο και στο μικρότερο σύνολο εκπαίδευσης χωρίς τη χρήση κάποιου αλγόριθμου προ-επεξεργασίας, παρατίθενται στον ακόλουθο πίνακα:

Πίνακας 4: Αποτελέσματα κατηγοριοποιητών χωρίς προ-επεξεργασία.

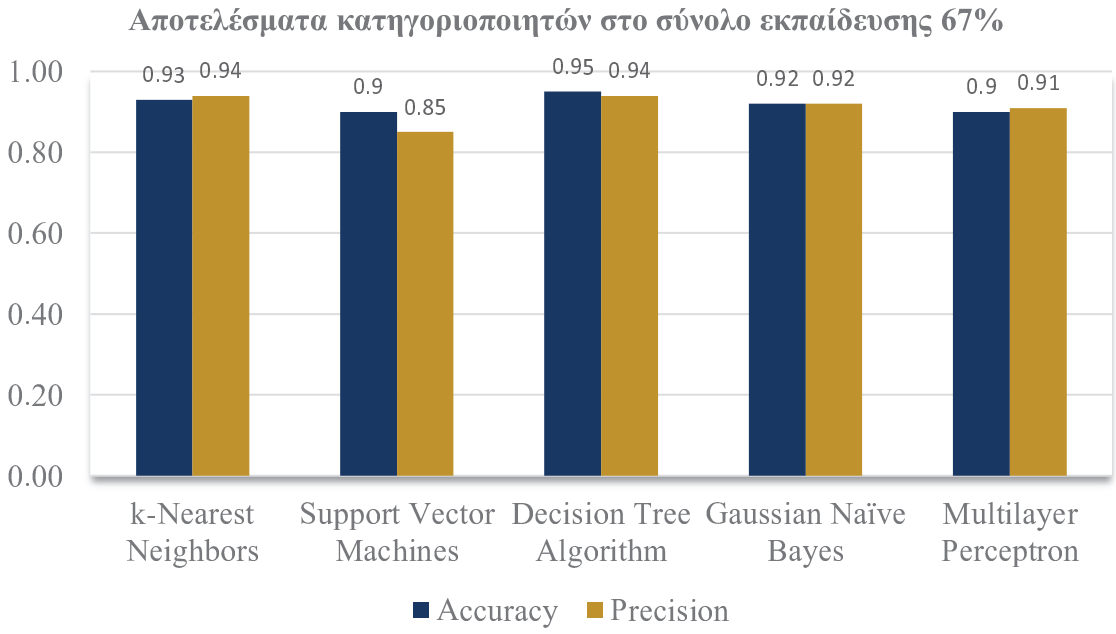
Αποτελέσματα αλγορίθμων κατηγοριοποίησης χωρίς προ-επεξεργασία				
	Μεγάλο Σύνολο Εκπαίδευσης (80%)		**Μικρό Σύνολο Εκπαίδευσης (67%)**	
Κατηγοριοποιητές	**Ορθότητα**	**Ακρίβεια**	**Ορθότητα**	**Ακρίβεια**
k-Nearest Neighbor	0.93	0.93	0.93	0.94
Support Vector Machines	0.91	0.85	0.90	0.85
Decision Tree	0.96	0.97	0.95	0.94
Naïve Bayes	0.94	0.95	0.92	0.92
Multilayer Perceptron	0.92	0.92	0.90	0.91

Όπως παρατηρείται, η μεγαλύτερη ορθότητα και ακρίβεια επιτυγχάνεται από τον αλγόριθμο δένδρου αποφάσεων και στις δύο διαφορετικές προσεγγίσεις των συνόλων εκπαίδευσης, με ποσοστό 0.96 και 0.97, αντίστοιχα. Ο ακριβώς επόμενος αλγόριθμος που δείχνει να φέρει τα καλύτερα ποσοστά στην ορθότητα και στην ακρίβειά του για το μεγαλύτερο σύνολο εκπαίδευσης είναι ο απλοϊκός Bayes, ενώ για το μικρότερο σύνολο εκπαίδευσης είναι ο αλγόριθμος των k-κοντινότερων γειτόνων.

Στα δύο διαγράμματα που ακολουθούν, αναπαρίστανται γραφικά ο προαναφερθέντας πίνακας ως ραβδογράμματα («Διάγραμμα 3» και «Διάγραμμα 4»).



Διάγραμμα 3: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 80% χωρίς προ-επεξεργασία.



Διάγραμμα 4: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 67% χωρίς προ-επεξεργασία.

5.2.4 Εφαρμογή κατηγοριοποιητών με χρήση του Standard Scaler

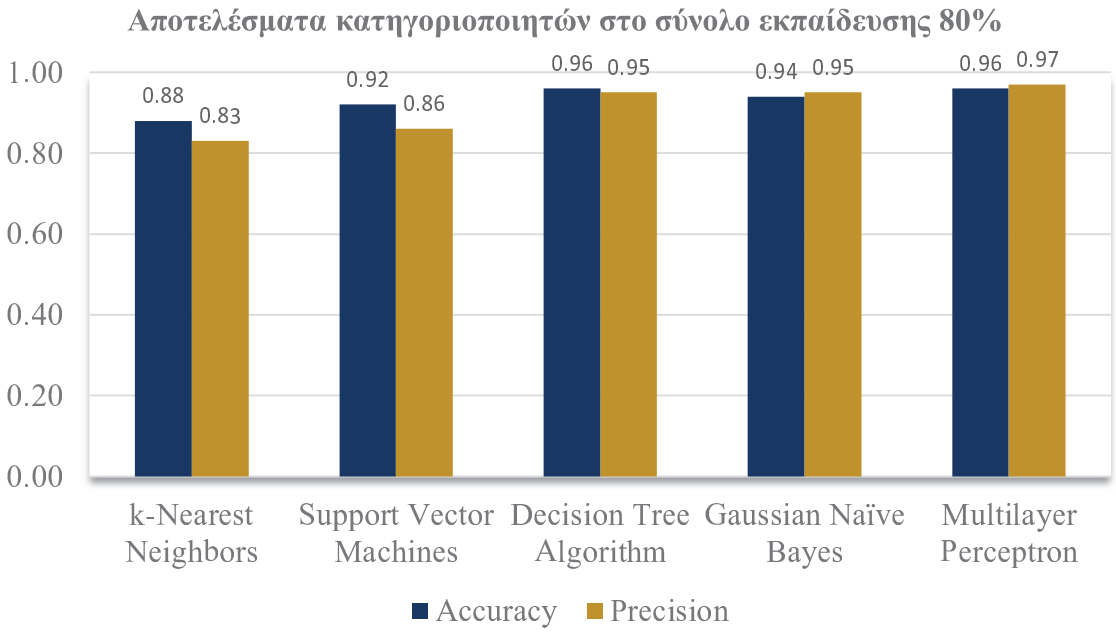
Τα αποτελέσματα των αλγορίθμων της κατηγοριοποίησης για το μικρότερο και το μεγαλύτερο σύνολο δεδομένων που μελετήθηκε μετά τη χρήση του αλγορίθμου Standard Scaler, παρατίθενται στον ακόλουθο πίνακα:

Πίνακας 5: Αποτελέσματα κατηγοριοποιητών με χρήση Standard Scaler.

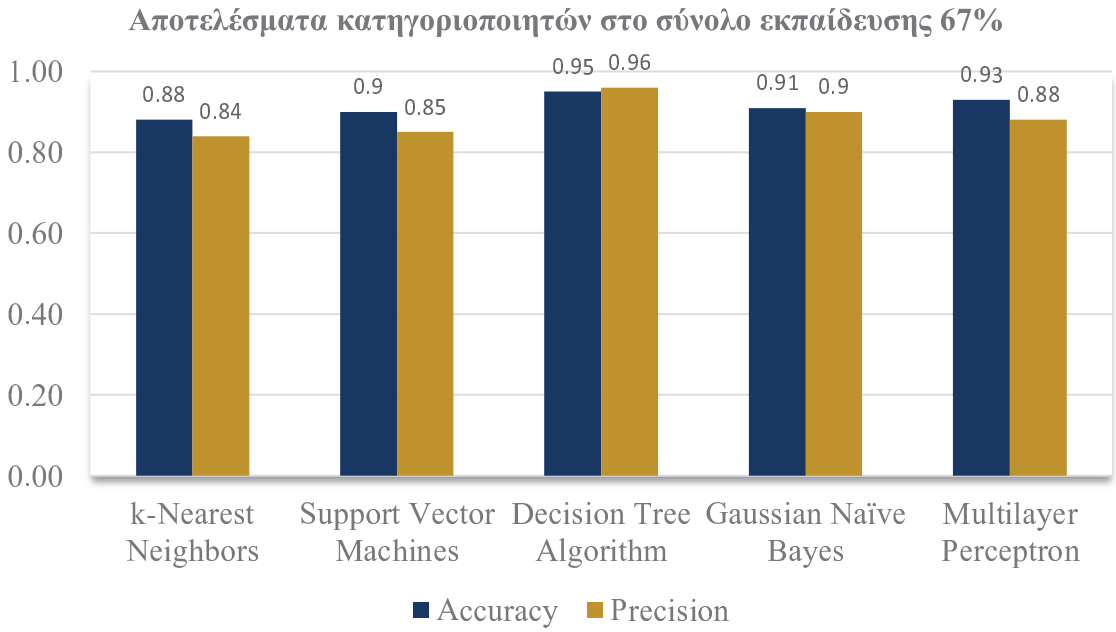
Αποτελέσματα αλγορίθμων κατηγοριοποίησης με χρήση Standard Scaler				
	Μεγάλο Σύνολο Εκπαίδευσης (80%)		**Μικρό Σύνολο Εκπαίδευσης (67%)**	
Κατηγοριοποιητές	**Ορθότητα**	**Ακρίβεια**	**Ορθότητα**	**Ακρίβεια**
k-Nearest Neighbor	0.88	0.83	0.88	0.84
Support Vector Machines	0.92	0.86	0.90	0.85
Decision Tree	0.96	0.95	0.95	0.96
Naïve Bayes	0.94	0.95	0.91	0.90
Multilayer Perceptron	0.96	0.97	0.93	0.88

Όπως παρατηρείται, υπάρχει μια σχετική «στασιμότητα» των αποτελεσμάτων χωρίς κάποια ιδιαίτερη βελτίωση. Ο αλγόριθμος που έφερε τα καλύτερα ποσοστά σχετικά με την ορθότητα και την ακρίβεια στο μεγαλύτερο σύνολο εκπαίδευσης, ήταν το νευρωνικό δίκτυο τύπου Perceptron με τιμές 0.96 και 0,97 αντίστοιχα. Όσον αναφορά στο μικρότερο σύνολο εκπαίδευσης, ο αλγόριθμος του δένδρου αποφάσεων ήταν αυτός που έφερε τα καλύτερα αποτελέσματα με 0.95 ποσοστό ορθότητας και 0.96 ακρίβειας.

Στο «Διάγραμμα 5» και στο «Διάγραμμα 6» που ακολουθούν, αναπαρίστανται γραφικά τα αποτελέσματα των αλγορίθμων στις διαφορετικές προσεγγίσεις των συνόλων εκπαίδευσης με τη χρήση προ-επεξεργασίας του Standard Scaler.



Διάγραμμα 5: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 80% με χρήση του Standard Scaler.



Διάγραμμα 6: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 67% με χρήση του Standard Scaler.

5.2.5 Εφαρμογή κατηγοριοποιητών με χρήση του MinMax Scaler

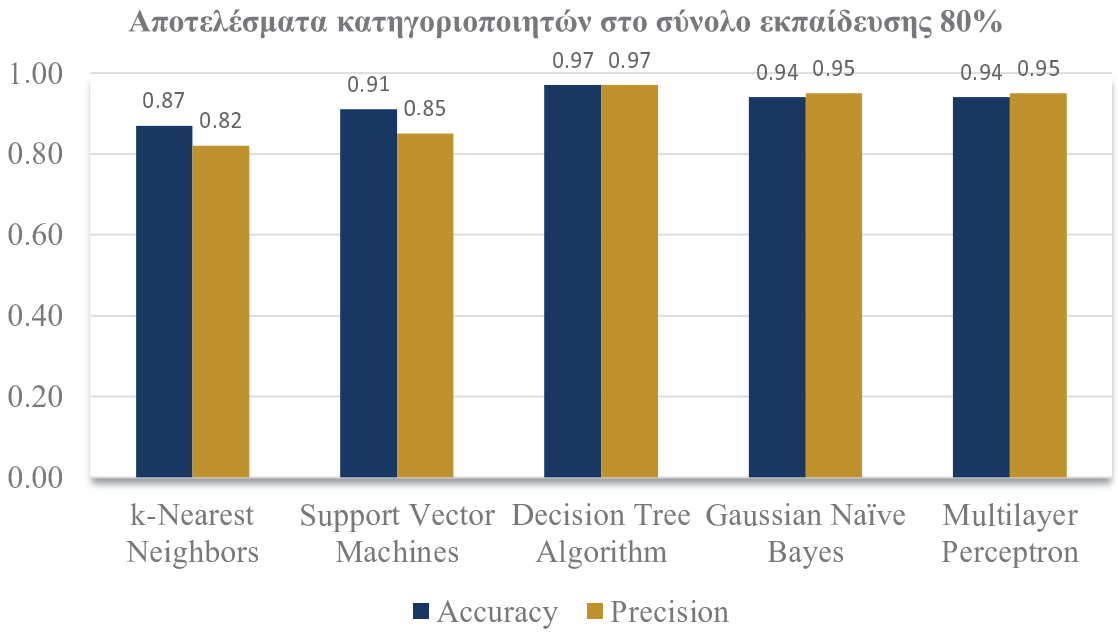
Τα αποτελέσματα των αλγορίθμων της κατηγοριοποίησης για το μικρότερο και το μεγαλύτερο σύνολο δεδομένων μετά τη χρήση του αλγορίθμου MinMax Scaler, παρατίθενται στον ακόλουθο πίνακα:

Πίνακας 6: Αποτελέσματα κατηγοριοποιητών με χρήση MinMax Scaler.

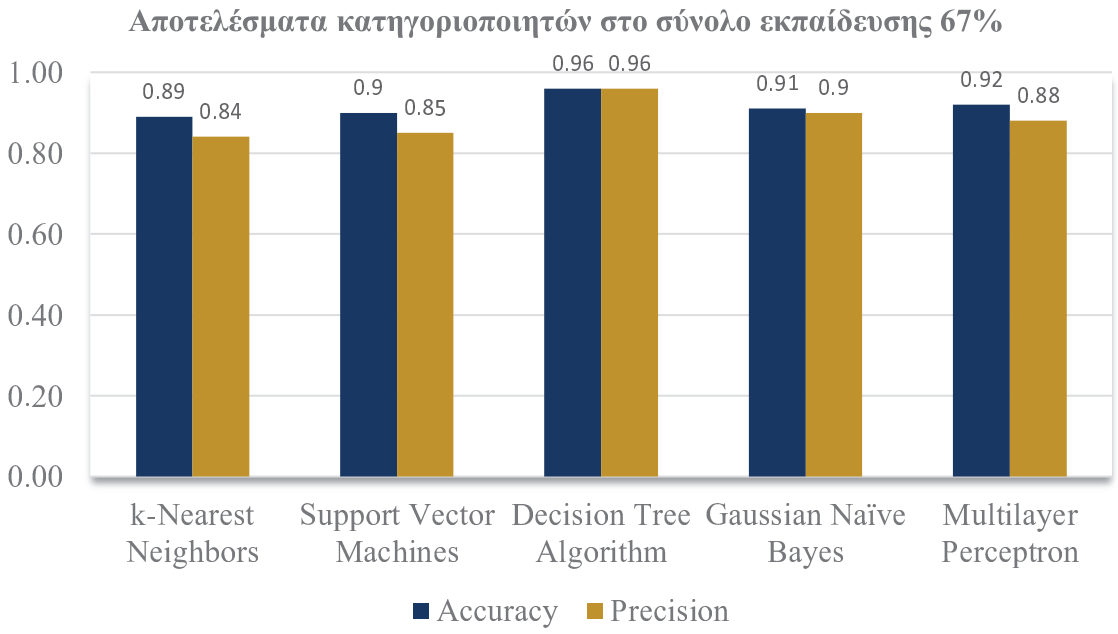
Αποτελέσματα αλγορίθμων κατηγοριοποίησης με χρήση MinMax Scaler				
	Μεγάλο Σύνολο Εκπαίδευσης (80%)		**Μικρό Σύνολο Εκπαίδευσης (67%)**	
Κατηγοριοποιητές	**Ορθότητα**	**Ακρίβεια**	**Ορθότητα**	**Ακρίβεια**
k-Nearest Neighbor	0.87	0.82	0.89	0.84
Support Vector Machines	0.91	0.85	0.90	0.85
Decision Tree	0.97	0.97	0.96	0.96
Naïve Bayes	0.94	0.95	0.91	0.90
Multilayer Perceptron	0.94	0.95	0.92	0.88

Όπως παρατηρείται, η καλύτερη απόδοση από την εφαρμογή των αλγορίθμων στο μεγαλύτερο αλλά και στο μικρότερο σύνολο εκπαίδευσης πραγματοποιείται από τον αλγόριθμο δένδρου αποφάσεων. Η ορθότητα αλλά και η ακρίβεια του συγκεκριμένου αλγορίθμου στο μεγαλύτερο σύνολο ανέρχεται στο 0.97, ενώ για το μικρότερο σύνολο ανέρχεται στο 0.96.

Στα ακόλουθα διαγράμματα («Διάγραμμα 7» και «Διάγραμμα 8»), παρουσιάζονται γραφικά τα αποτελέσματα των αλγορίθμων έπειτα από τη χρήση της προ-επεξεργασίας MinMax Scaler.



Διάγραμμα 7: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 80% με χρήση του MinMax Scaler.



Διάγραμμα 8: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 67% με χρήση του MinMax Scaler.

5.2.6 Εφαρμογή κατηγοριοποιητών με χρήση MinMax και Standard Scaler

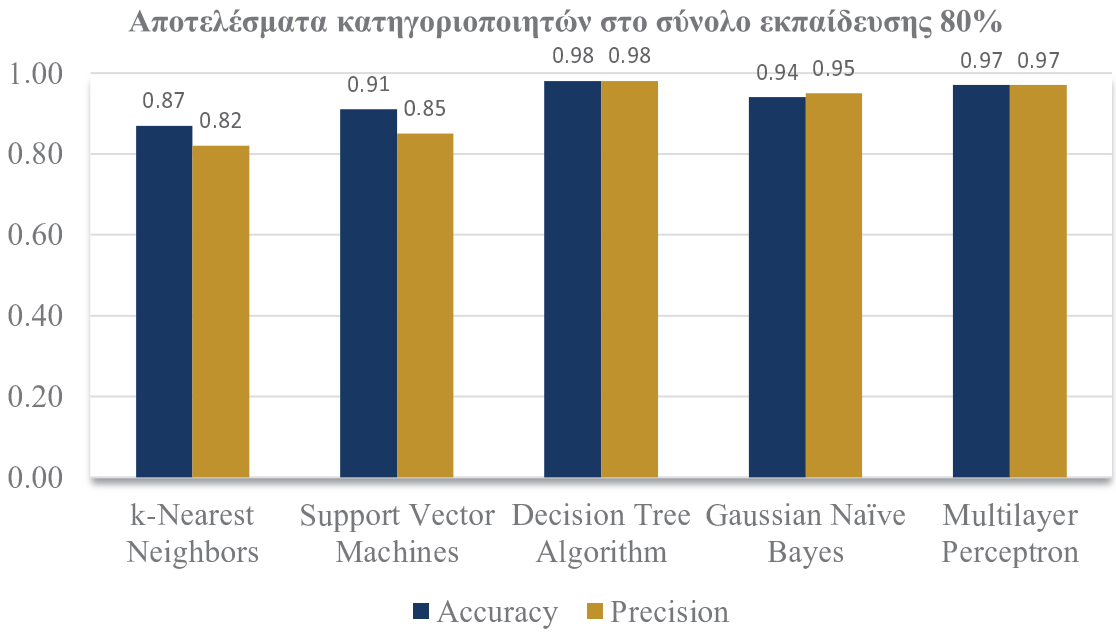
Τα αποτελέσματα των αλγορίθμων της κατηγοριοποίησης για το μικρότερο και το μεγαλύτερο σύνολο δεδομένων μετά τη χρήση των αλγορίθμων προ-επεξεργασίας MinMax και Standard Scaler, παρατίθενται στον ακόλουθο πίνακα:

Πίνακας 7: Αποτελέσματα κατηγοριοποιητών με χρήση MinMax και Standard Scaler.

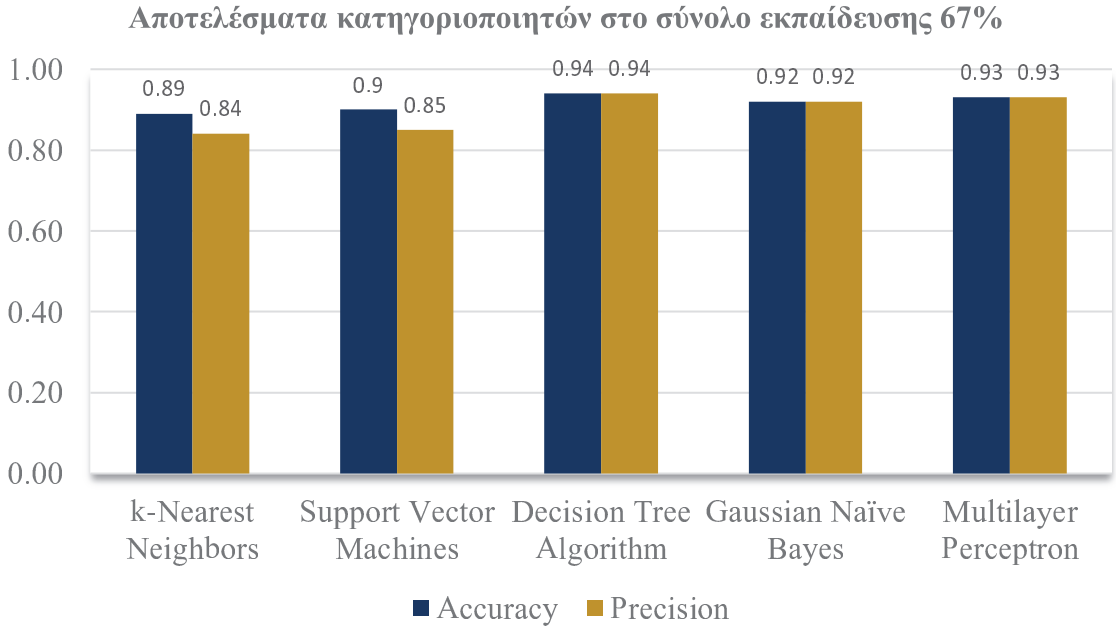
Αποτελέσματα αλγορίθμων κατηγοριοποίησης με χρήση MinMax και Standard Scaler				
	Μεγάλο Σύνολο Εκπαίδευσης (80%)		**Μικρό Σύνολο Εκπαίδευσης (67%)**	
Κατηγοριοποιητές	**Ορθότητα**	**Ακρίβεια**	**Ορθότητα**	**Ακρίβεια**
k-Nearest Neighbor	0.87	0.82	0.89	0.84
Support Vector Machines	0.91	0.85	0.90	0.85
Decision Tree	0.98	0.98	0.94	0.94
Naïve Bayes	0.94	0.95	0.92	0.92
Multilayer Perceptron	0.97	0.97	0.93	0.93

Όπως παρατηρείται, ο αλγόριθμος του δένδρου αποφάσεων έφερε για άλλη μια φορά τα καλύτερα αποτελέσματα στις μετρικές της ορθότητας και της ακρίβειας και στις δύο προσεγγίσεις των συνόλων εκπαίδευσης. Συγκεκριμένα, για το μεγαλύτερο σύνολο οι τιμές της ορθότητας και της ακρίβειάς του ήταν 0.98, ενώ στο μικρότερο σύνολο ήταν 0.94.

Στη συνέχεια, ακολουθούν τα διαγράμματα «Διάγραμμα 9» και «Διάγραμμα 10», τα οποία αναπαριστούν γραφικά τα αποτελέσματα των αλγορίθμων στην προσέγγιση της προ-επεξεργασίας με τους δύο αλγόριθμους MinMax και Standard Scaler.



Διάγραμμα 9: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 80% με χρήση του MinMax και Standard Scaler.



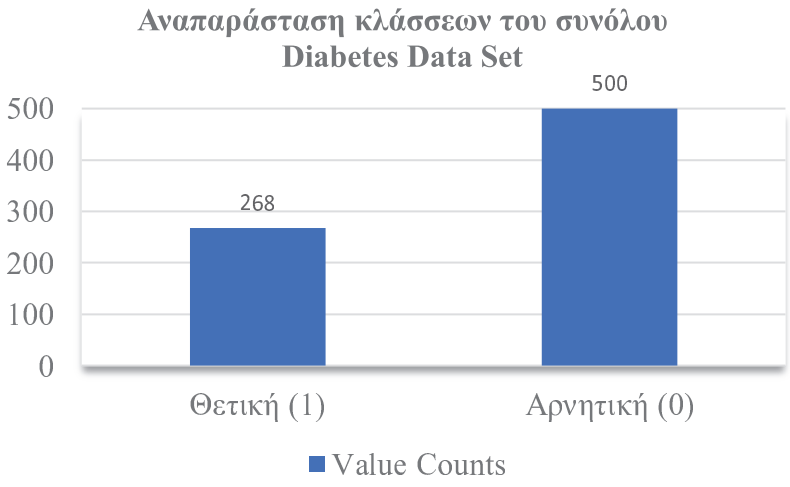
Διάγραμμα 10: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 67% με χρήση του MinMax και Standard Scaler.

6 Μελέτη Περίπτωσης: Diabetes Data Set

6.1 Περιγραφή και προετοιμασία μελέτης

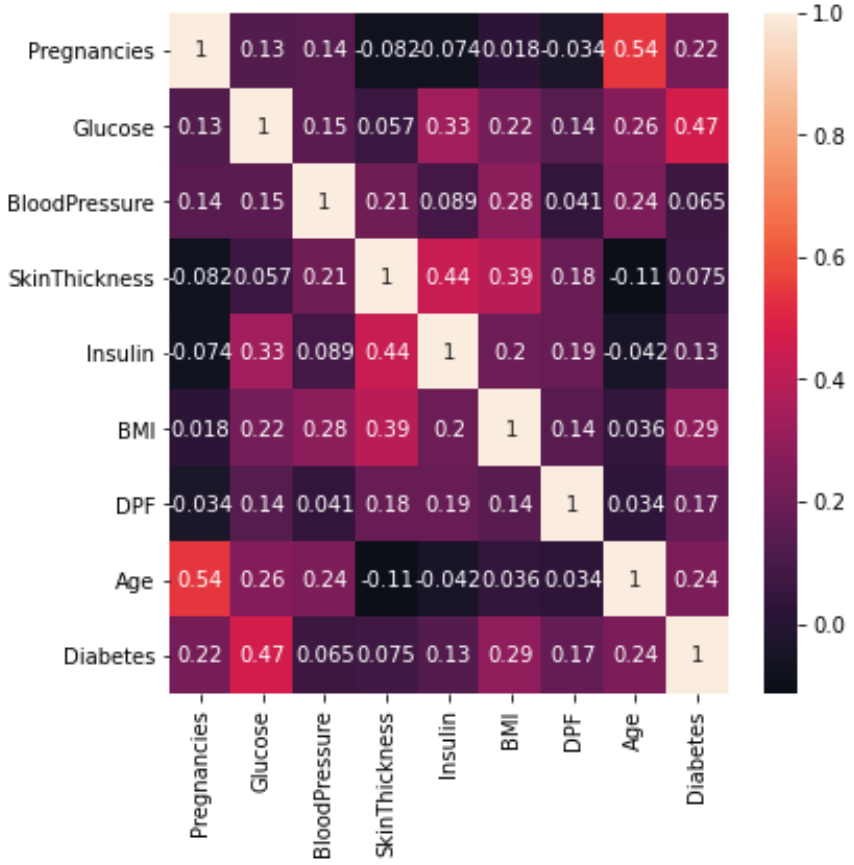
Για την υλοποίηση της μελέτης στο συγκεκριμένο σύνολο δεδομένων, χρησιμοποιήθηκαν όλοι οι αλγόριθμοι που αναφέρθηκαν στο θεωρητικό μέρος της εργασίας («Μέρος Α»), χωρίς όμως να συμπεριλαμβάνεται η γραμμική παλινδρόμηση. Οι μεταβλητές του συγκεκριμένου συνόλου ανέρχονται στις 9, εκ των οποίων η μία από αυτές είναι η κλάση του συνόλου και, 768 είναι οι συνολικές εγγραφές των στιγμιότυπων. Η εφαρμογή των αλγορίθμων πραγματοποιήθηκε με σκοπό την εύρεση του καλύτερου κατηγοριοποιητή από αυτούς, στα αποτελέσματα της ορθότητας και της ακρίβειας των προβλέψεών του για την κλάση του συνόλου.

Το πλήθος των κλάσεων αυτών του συνόλου, αναπαρίστανται στο «Διάγραμμα 11» που ακολουθεί, όπου οι 500 εγγραφές από αυτές είναι αρνητικές στην ύπαρξη της νόσου του διαβήτη, ενώ οι 268 είναι θετικές.



Διάγραμμα 11: Αναπαράσταση κλάσεων του συνόλου Diabetes Data Set.

Για την εφαρμογή των αλγορίθμων της μηχανικής μάθησης, χρησιμοποιήθηκαν όλες οι μεταβλητές οι οποίες ήταν καταγεγραμμένες εξαρχής στο σύνολο, χωρίς να πραγματοποιείται η τροποποίηση κάποιας από αυτές. Στην «Εικόνα 5» παρουσιάζεται ο πίνακας συσχέτισης που δημιουργήθηκε μέσω της βιβλιοθήκης Seaborn για αυτές τις μεταβλητές.



Εικόνα 5: Πίνακας συσχέτισης Diabetes Data Set.

Όπως παρατηρείται, οι μεταβλητές που έχουν τη μεγαλύτερη συσχέτιση αναφορικά με την ύπαρξη της νόσου του διαβήτη, είναι τα επίπεδα γλυκόζης (Glucose) και ο δείκτης μάζας του σώματος (BMI). Σύμφωνα με πολλούς ειδικούς, οι δύο αυτές μεταβλητές μπορούν να επηρεάσουν τη θετική ύπαρξη μιας πληθώρας ασθενειών, καθώς πολλοί άνθρωποι σήμερα ζουν έναν πλήρως ανθυγιεινό τρόπο ζωής.

Η εφαρμογή των αλγορίθμων πραγματοποιήθηκε με τις διαφορετικές προσεγγίσεις α) του ποσοστού του συνόλου εκπαίδευσης και β) των δύο αλγόριθμων προ-επεξεργασίας MinMax και Standard Scaler. Συγκεκριμένα, οι αλγόριθμοι κατηγοριοποίησης που χρησιμοποιήθηκαν είναι οι: α) Λογιστική Παλινδρόμηση, β) k-Κοντινότεροι Γείτονες, γ) Μηχανές Διανυσμάτων Υποστήριξης, δ) Απλοϊκός Μπάγιες, ε) Αλγόριθμος Δένδρου Αποφάσεων και στ) Νευρωνικό Δίκτυο.

6.1.1 Γενική μεθεδολογία της μελέτης

Στην παρούσα ενότητα, θα γίνει ανάλυση των γενικών βημάτων που πραγματοποιήθηκαν για την εφαρμογή των αλγορίθμων μηχανικής μάθησης και εξόρυξης πληροφορίας στο Diabetes Data Set. Στη συνέχεια, παρατίθενται τα βήματα που ακολουθήθηκαν.

- Φόρτωση των συνόλων στην Python: Κατά την αρχή της μελέτης, έγινε φόρτωση του συνόλου με τη βιβλιοθήκη Pandas για περαιτέρω επεξεργασία με τη γλώσσα Python. Στη συνέχεια, πραγματοποιήθηκαν ειδικές μέθοδοι της βιβλιοθήκης αυτής, για την εύρεση ελλιπών τιμών στο σύνολο. Όπως και στην προηγούμενη μελέτη περίπτωσης, δεν υπήρξαν κάποιες ελλιπείς τιμές που να χρειάζονταν συμπλήρωση.

- Διαχωρισμός των συνόλων για εκπαίδευση: Ο διαχωρισμός των συνόλων σε σύνολο εκπαίδευσης και σύνολο δοκιμής, πραγματοποιήθηκε σε δύο διαφορετικές προσεγγίσεις. Η πρώτη αφορά στη χρήση του 80% του συνόλου για εκπαίδευση και 20% για δοκιμή. Η δεύτερη, αφορά στη χρήση του 67% συνόλου για εκπαίδευση και 33% για δοκιμή.

- Εφαρμογή των αλγορίθμων: Η εφαρμογή των αλγορίθμων χωρίστηκε σε διαφορετικές προσεγγίσεις προ-επεξεργασίας, όπως και στην περίπτωση του Student Performance Data Set. Συγκεκριμένα, οι προσεγγίσεις αφορούν στην εφαρμογή των αλγορίθμων στο σύνολο χωρίς καμία προ-επεξεργασία, με προ-επεξεργασία MinMax Scaler και με προ-επεξεργασία Standard Scaler.

- Αξιολόγηση των αποτελεσμάτων: Ως τελευταίο βήμα για την ολοκλήρωσης της μελέτης, πραγματοποιήθηκαν μετρικές αξιολόγησης όλων των διαφορετικών αποτελεσμάτων των αλγορίθμων σχετικά με την ακρίβεια και την ορθότητα, που έφεραν σε όλες αυτές τις διαφορετικές εκδοχές προ-επεξεργασίας και εκπαίδευσης. Τα συγκεκριμένα αποτελέσματα υπολογίστηκαν με τη χρήση του βοηθητικού εργαλείου Metrics της βιβλιοθήκης Scikit-Learn.

6.2 Εφαρμογή αλγορίθμων

Στις ακόλουθες ενότητες (6.2.1 μέχρι και 6.2.4), παρουσιάζονται τα αποτελέσματα που εξήχθησαν με την εφαρμογή αλγορίθμων κατηγοριοποίησης στο σύνολο Diabetes Data Set. Τα αποτελέσματα αυτά αφορούν στην ορθότητα (accuracy) και την ακρίβεια (precision) του εκάστοτε μοντέλου, στις διαφορετικές προσεγγίσεις προ-επεξεργασίας και εκπαίδευσης που προαναφέρθηκαν. Όπως και στην προηγούμενη μελέτη περίπτωσης, γίνεται ξανά η παρουσίαση των αποτελεσμάτων με ραβδογράμματα (bar plots).

6.2.1 Εφαρμογή κατηγοριοποιητών χωρίς προ-επεξεργασία

Τα αποτελέσματα των αλγορίθμων κατηγοριοποίησης στο μικρότερο και στο μεγαλύτερο σύνολο εκπαίδευσης του Diabetes Data Set χωρίς τη χρήση κάποιου είδους αλγόριθμου προ-επεξεργασίας, παρουσιάζονται στον πίνακα που ακολουθεί:

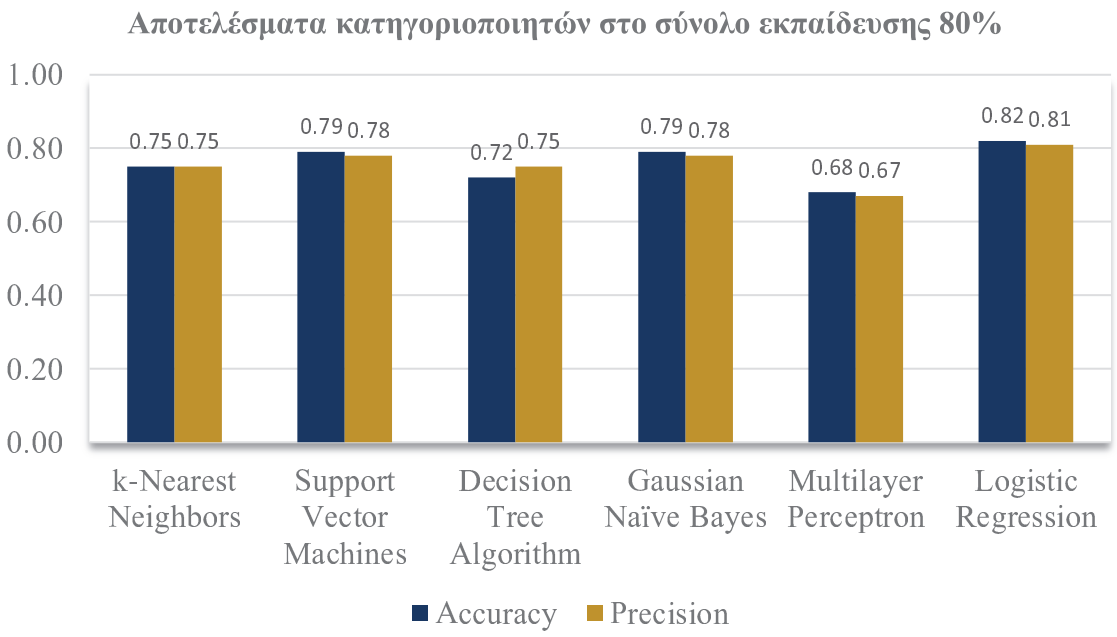
Πίνακας 8: Αποτελέσματα κατηγοριοποιητών χωρίς προ-επεξεργασία.

Αποτελέσματα αλγορίθμων κατηγοριοποίησης χωρίς προ-επεξεργασία				
	Μεγάλο Σύνολο Εκπαίδευσης (80%)		**Μικρό Σύνολο Εκπαίδευσης (67%)**	
Κατηγοριοποιητές	**Ορθότητα**	**Ακρίβεια**	**Ορθότητα**	**Ακρίβεια**
k-Nearest Neighbor	0.75	0.75	0.73	0.72
Support Vector Machines	0.79	0.78	0.74	0.73
Decision Tree	0.72	0.75	0.69	0.69
Naïve Bayes	0.79	0.78	0.74	0.73
Multilayer Perceptron	0.68	0.67	0.70	0.70
Logistic Regression	0.82	0.81	0.77	0.78

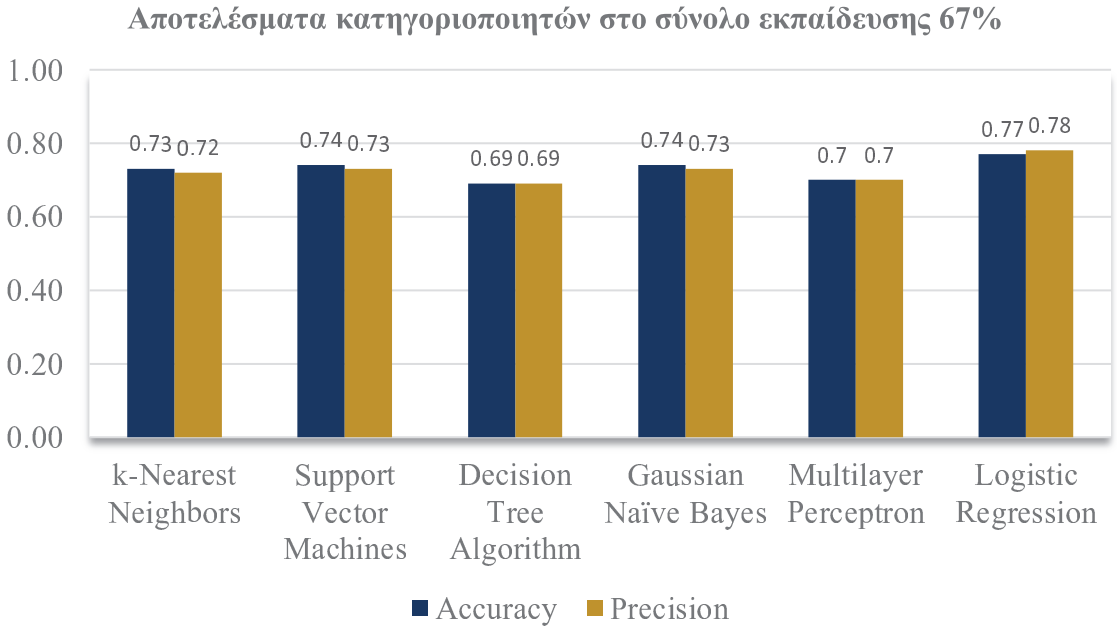
Παρατηρείται ότι τα καλύτερα αποτελέσματα πραγματοποιήθηκαν μέσω του αλγορίθμου της λογιστικής παλινδρόμησης και στις δύο εκδοχές του μικρότερου και του μεγαλύτερου

συνόλου εκπαίδευσης. Οι μεγαλύτερες μετρικές ορθότητας και ακρίβειας εξήχθησαν στο μεγαλύτερο σύνολο και, ήταν 0.82 και 0.81, αντίστοιχα.

Στη συνέχεια, γίνεται γραφική αναπαράσταση των αποτελεσμάτων του «Πίνακα 8» στο «Διάγραμμα 12» και «Διάγραμμα 13» που ακολουθούν.



Διάγραμμα 12: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 80% χωρίς προ-επεξεργασία.



Διάγραμμα 13: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 67% χωρίς προ-επεξεργασία.

6.2.2 Εφαρμογή κατηγοριοποιητών με χρήση του Standard Scaler

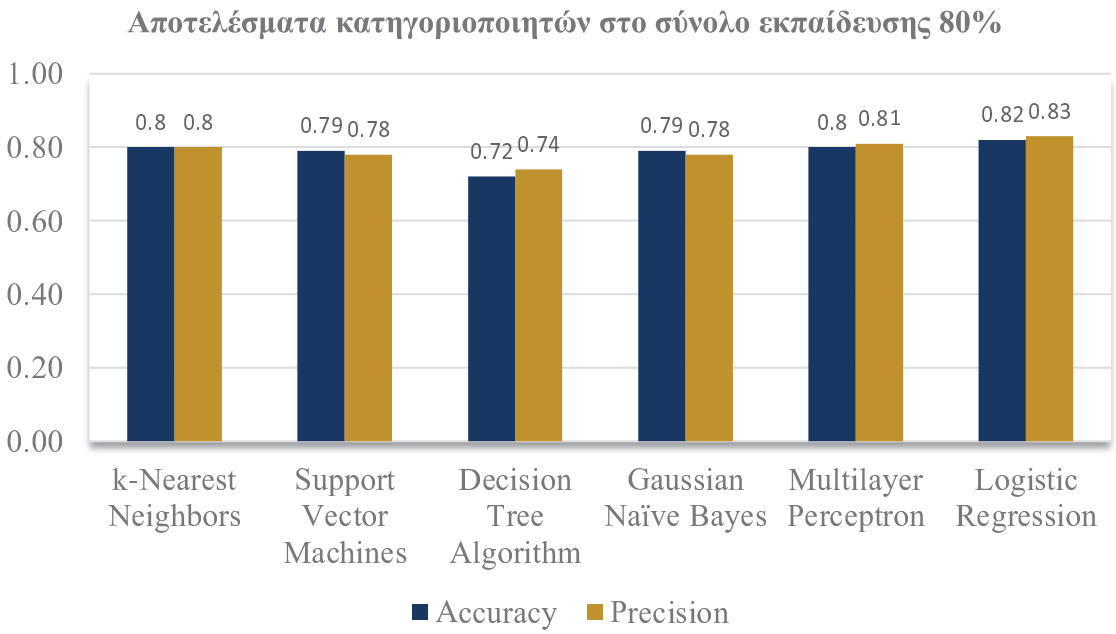
Στον πίνακα που ακολουθεί, παρατίθενται τα αποτελέσματα των αλγορίθμων κατηγοριοποίησης στο μικρότερο και στο μεγαλύτερο σύνολο εκπαίδευσης του Diabetes Data Set, μετά τη χρήση του αλγόριθμου προ-επεξεργασίας Standard Scaler.

Πίνακας 9: Αποτελέσματα κατηγοριοποιητών με χρήση Standard Scaler.

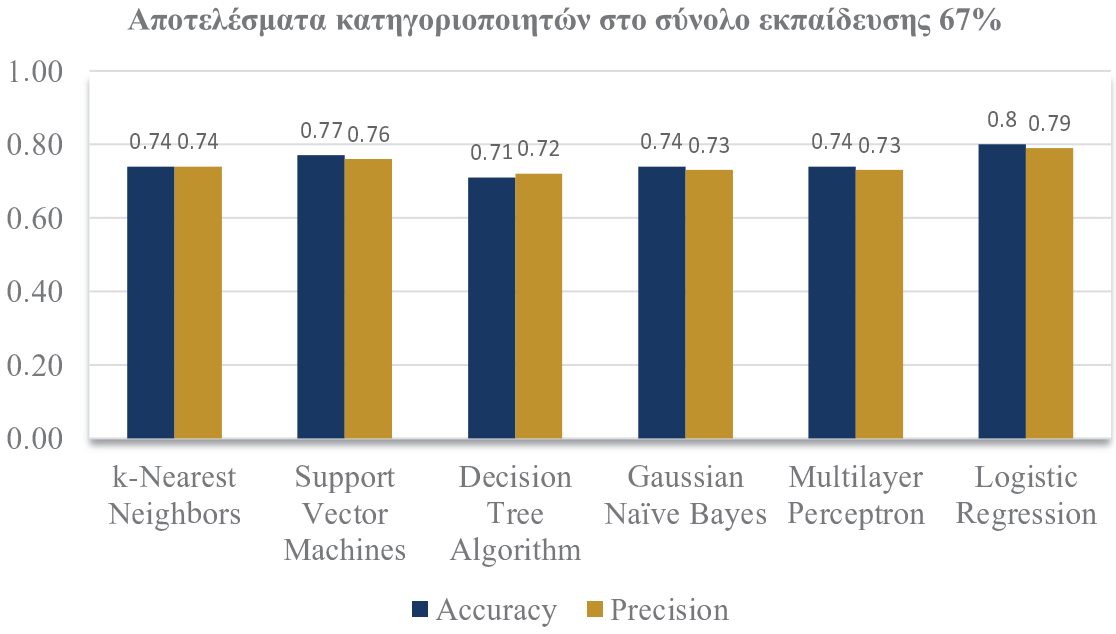
Αποτελέσματα αλγορίθμων κατηγοριοποίησης με χρήση του Standard Scaler				
	Μεγάλο Σύνολο Εκπαίδευσης (80%)		**Μικρό Σύνολο Εκπαίδευσης (67%)**	
Κατηγοριοποιητές	**Ορθότητα**	**Ακρίβεια**	**Ορθότητα**	**Ακρίβεια**
k-Nearest Neighbor	0.80	0.80	0.74	0.74
Support Vector Machines	0.79	0.78	0.77	0.76
Decision Tree	0.72	0.74	0.71	0.72
Naïve Bayes	0.79	0.78	0.74	0.73
Multilayer Perceptron	0.80	0.81	0.74	0.73
Logistic Regression	0.82	0.83	0.80	0.79

Μετά τη χρήση της συγκεκριμένης τεχνικής προ-επεξεργασίας, παρατηρείται ότι αρκετοί ήταν αυτοί οι αλγόριθμοι που έδειξαν μια σημαντική αύξηση στα αποτελέσματά τους. Χαρακτηριστικό παράδειγμα αυτών είναι το νευρωνικό δίκτυο τύπου Perceptron (MLP), στο οποίο αυξήθηκε η ορθότητα και η ακρίβεια του, στην περίπτωση εκπαίδευσης με το μεγαλύτερο σύνολο, κατά 10 τις εκατό περίπου. Ο αλγόριθμος που έφερε για άλλη μια φορά τα καλύτερα αποτελέσματα και στις δύο προσεγγίσεις της εκπαίδευσης ήταν αυτός της λογιστικής παλινδρόμησης. Τα καλύτερα αποτελέσματα αυτού, δόθηκαν ξανά για το μεγαλύτερο σύνολο εκπαίδευσης, με την τιμή της ορθότητας να είναι 0.82 και, την τιμή της ακρίβειας να είναι 0.83.

Το «Διάγραμμα 14» και το «Διάγραμμα 15» που ακολουθούν, αναπαριστούν γραφικά τα αποτελέσματα των αλγόριθμων του προηγούμενου πίνακα. Στο πρώτο διάγραμμα παρουσιάζονται τα αποτελέσματα του μεγαλύτερου συνόλου εκπαίδευσης, ενώ στο δεύτερο διάγραμμα του μικρότερου.



Διάγραμμα 14: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 80% με χρήση του Standard Scaler.



Διάγραμμα 15: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 67% με χρήση του Standard Scaler.

6.2.3 Εφαρμογή κατηγοριοποιητών με χρήση του MinMax Scaler

Στον «Πίνακα 10» που ακολουθεί, αναπαρίστανται τα αποτελέσματα των αλγορίθμων κατηγοριοποίησης για το σύνολο δεδομένων που μελετάται, με τη χρήση της προ-επεξεργασίας του αλγόριθμου MinMax Scaler.

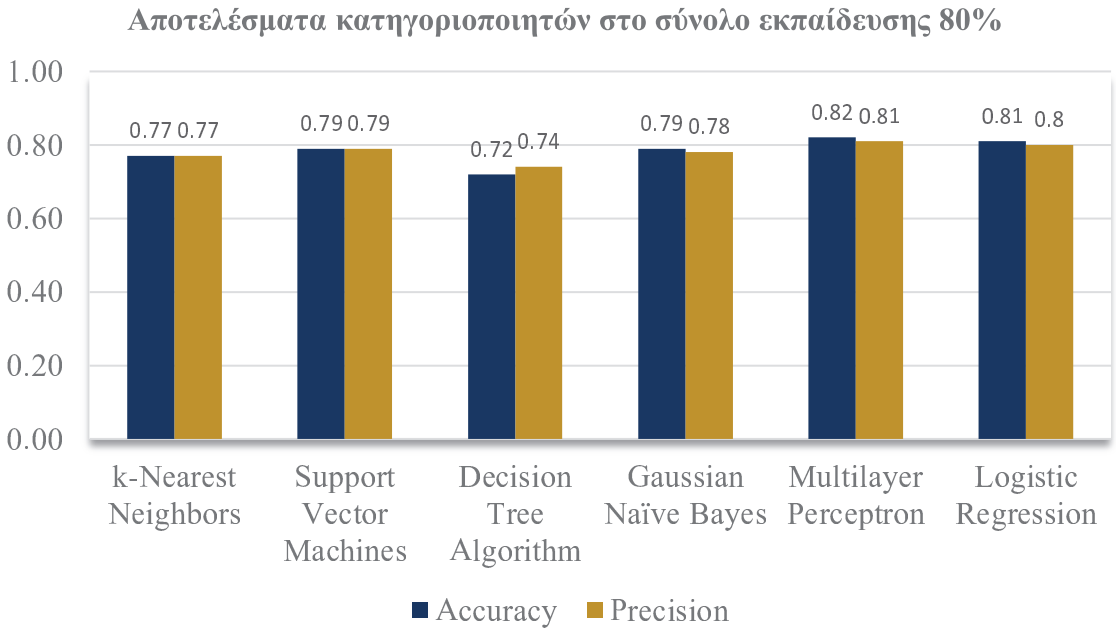
Πίνακας 10: Αποτελέσματα κατηγοριοποιητών με χρήση MinMax Scaler.

Αποτελέσματα αλγορίθμων κατηγοριοποίησης με χρήση του MinMax Scaler				
	Μεγάλο Σύνολο Εκπαίδευσης (80%)		**Μικρό Σύνολο Εκπαίδευσης (67%)**	
Κατηγοριοποιητές	**Ορθότητα**	**Ακρίβεια**	**Ορθότητα**	**Ακρίβεια**
k-Nearest Neighbor	0.77	0.77	0.74	0.74
Support Vector Machines	0.79	0.79	0.77	0.76
Decision Tree	0.72	0.74	0.69	0.70
Naïve Bayes	0.79	0.78	0.74	0.73
Multilayer Perceptron	0.82	0.81	0.74	0.73
Logistic Regression	0.81	0.80	0.79	0.78

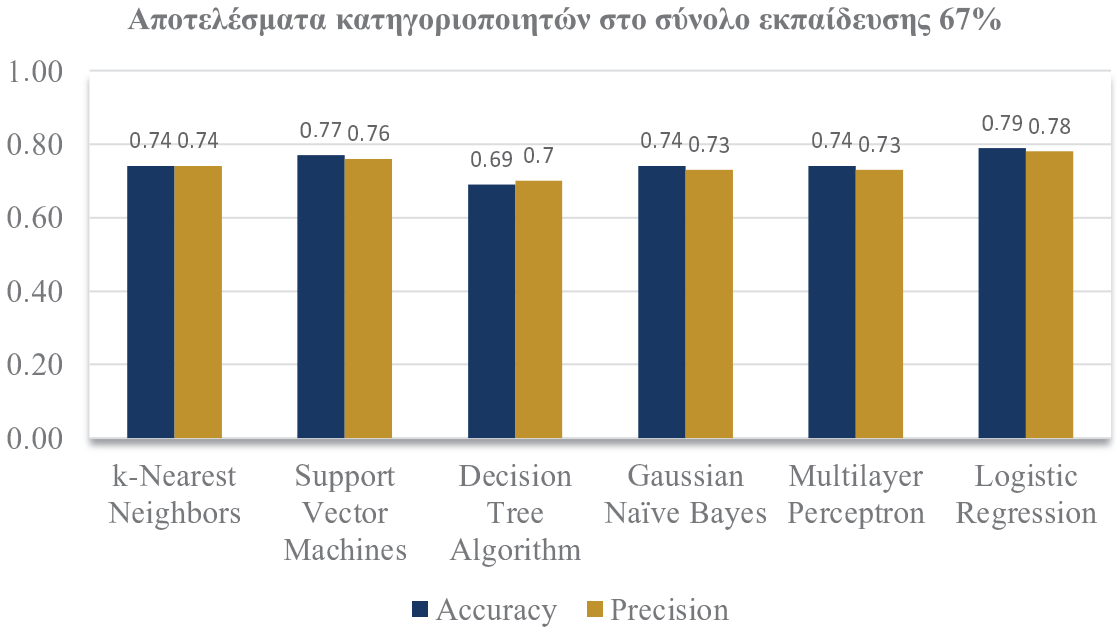
Ο αλγόριθμος που έδωσε τα καλύτερα αποτελέσματα στη συγκεκριμένη προσέγγιση της προ-επεξεργασίας για το μεγαλύτερο σύνολο εκπαίδευσης, είναι το νευρωνικό δίκτυο τύπου Perceptron. Συγκεκριμένα, η τιμή της ορθότητας των προβλέψεών του ήταν 0.82, ενώ η τιμή της ακρίβειάς του 0.81. Όσον αναφορά στο μικρότερο σύνολο εκπαίδευσης, ο αλγόριθμος που έφερε τα καλύτερα αποτελέσματα ήταν αυτός της λογιστικής παλινδρόμησης με τιμή ορθότητας 0.79 και τιμή ακρίβειας 0.78.

Στα δύο διαγράμματα που θα ακολουθήσουν, παρουσιάζονται γραφικά τα αποτελέσματα του πίνακα που μόλις αναφέρθηκε. Ειδικότερα, στο «Διάγραμμα 16» παρουσιάζονται τα

αποτελέσματα για το μεγαλύτερο σύνολο εκπαίδευσης που έγινε η εφαρμογή, ενώ στο «Διάγραμμα 17» παρουσιάζονται τα αποτελέσματα για το μικρότερο σύνολο εκπαίδευσης.



Διάγραμμα 16: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 80% με χρήση του MinMax Scaler.



Διάγραμμα 17: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 67% με χρήση του MinMax Scaler.

6.2.4 Εφαρμογή κατηγοριοποιητών με χρήση MinMax και Standard Scaler

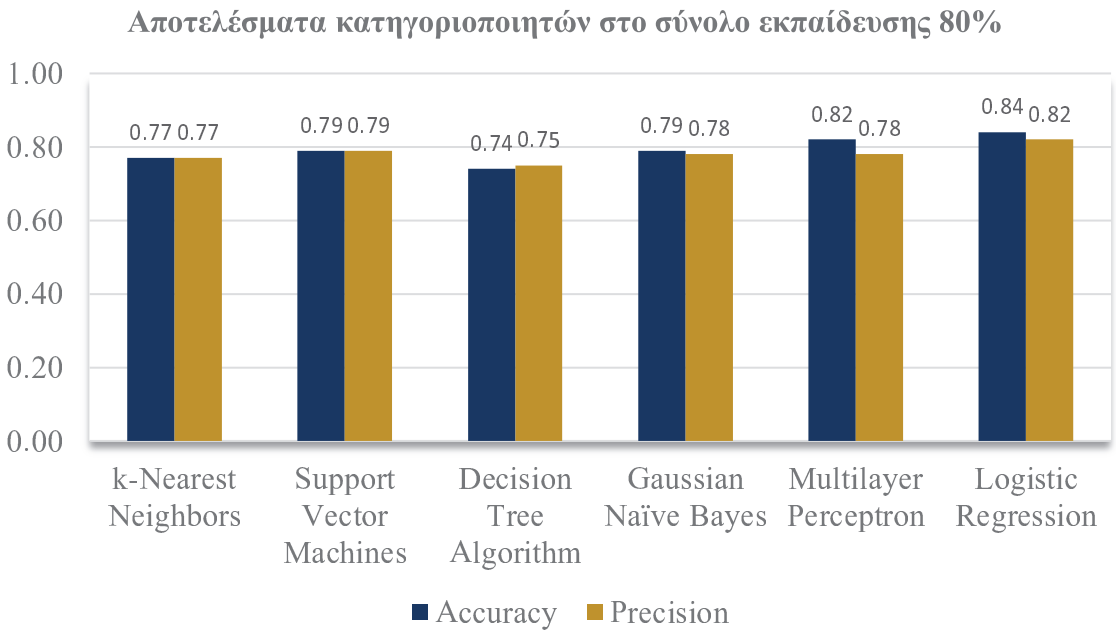
Στον ακόλουθο πίνακα («Πίνακας 11»), παρουσιάζονται τα αποτελέσματα των αλγορίθμων για το σύνολο Diabetes Data Set, μετά την χρήση των αλγορίθμων προ-επεξεργασίας MinMax και Standard Scaler.

Πίνακας 11: Αποτελέσματα κατηγοριοποιητών με χρήση MinMax και Standard Scaler.

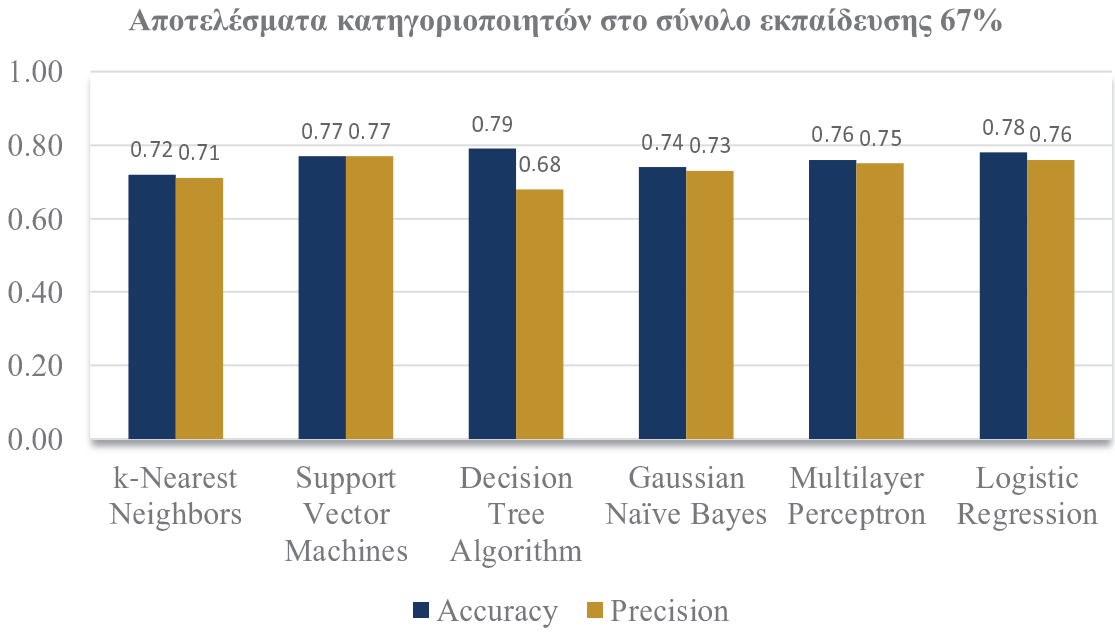
Αποτελέσματα αλγορίθμων κατηγοριοποίησης με χρήση MinMax και Standard Scaler				
	Μεγάλο Σύνολο Εκπαίδευσης (80%)		**Μικρό Σύνολο Εκπαίδευσης (67%)**	
Κατηγοριοποιητές	**Ορθότητα**	**Ακρίβεια**	**Ορθότητα**	**Ακρίβεια**
k-Nearest Neighbor	0.77	0.77	0.72	0.71
Support Vector Machines	0.79	0.79	0.77	0.77
Decision Tree	0.74	0.75	0.79	0.68
Naïve Bayes	0.79	0.78	0.74	0.73
Multilayer Perceptron	0.82	0.78	0.76	0.75
Logistic Regression	0.84	0.82	0.78	0.76

Μετά τη χρήση και των δύο τεχνικών της προ-επεξεργασίας που αναφέρθηκαν, ο αλγόριθμος που έδωσε τα καλύτερα αποτελέσματα για το μικρότερο σύνολο εκπαίδευσης στην ορθότητα, ήταν αυτός του αλγορίθμου δένδρου αποφάσεων με τιμή 0.79, ενώ στην ακρίβεια δόθηκαν από τον SVM με τιμή 0.77. Αντίθετα, στο μεγαλύτερο σύνολο εκπαίδευσης, για άλλη μια φορά τα καλύτερα αποτελέσματα δόθηκαν από τη λογιστική παλινδρόμηση με τιμές ορθότητας και ακρίβειας να είναι 0.84 και 0.82, αντίστοιχα. Όπως και στις προηγούμενες περιπτώσεις που μελετήθηκαν, τα αποτελέσματα των αλγορίθμων δείχνουν να είναι λίγο καλύτερα από μεριάς του μεγαλύτερου συνόλου εκπαίδευσης.

Στο «Διάγραμμα 18» και στο «Διάγραμμα 19», παρουσιάζονται τα αντίστοιχα αποτελέσματα των αλγορίθμων του «Πίνακα 11» γραφικά.



Διάγραμμα 18: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 80% με χρήση MinMax και Standard Scaler.



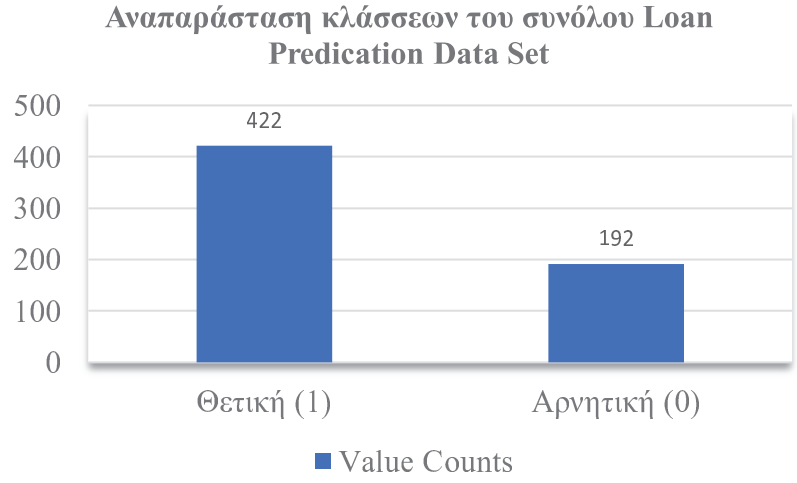
Διάγραμμα 19: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 67% με χρήση MinMax και Standard Scaler.

7 Μελέτη περίπτωσης: Loan Predication Data Set

7.1 Περιγραφή και προετοιμασία μελέτης

Για την υλοποίηση της μελέτης στο σύνολο Loan Predication Data Set, χρησιμοποιήθηκαν όλοι οι αλγόριθμοι κατηγοριοποίησης που αναφέρθηκαν σε προγενέστερα σημεία της εργασίας και, όλες οι μεταβλητές που ήταν εξαρχής καταγεγραμμένες στο σύνολο, εκτός φυσικά της μεταβλητής Loan ID. Η εφαρμογή αυτή τον αλγορίθμων πραγματοποιήθηκε με σκοπό την εύρεση του καλύτερου κατηγοριοποιητή στις διαφορετικές μεθόδους προ-επεξεργασίας του συνόλου, αλλά και στις διαφορετικές προσεγγίσεις χρήσης των συνόλων εκπαίδευσης, όπως ακριβώς και στις προηγούμενες μελέτες περίπτωσης.

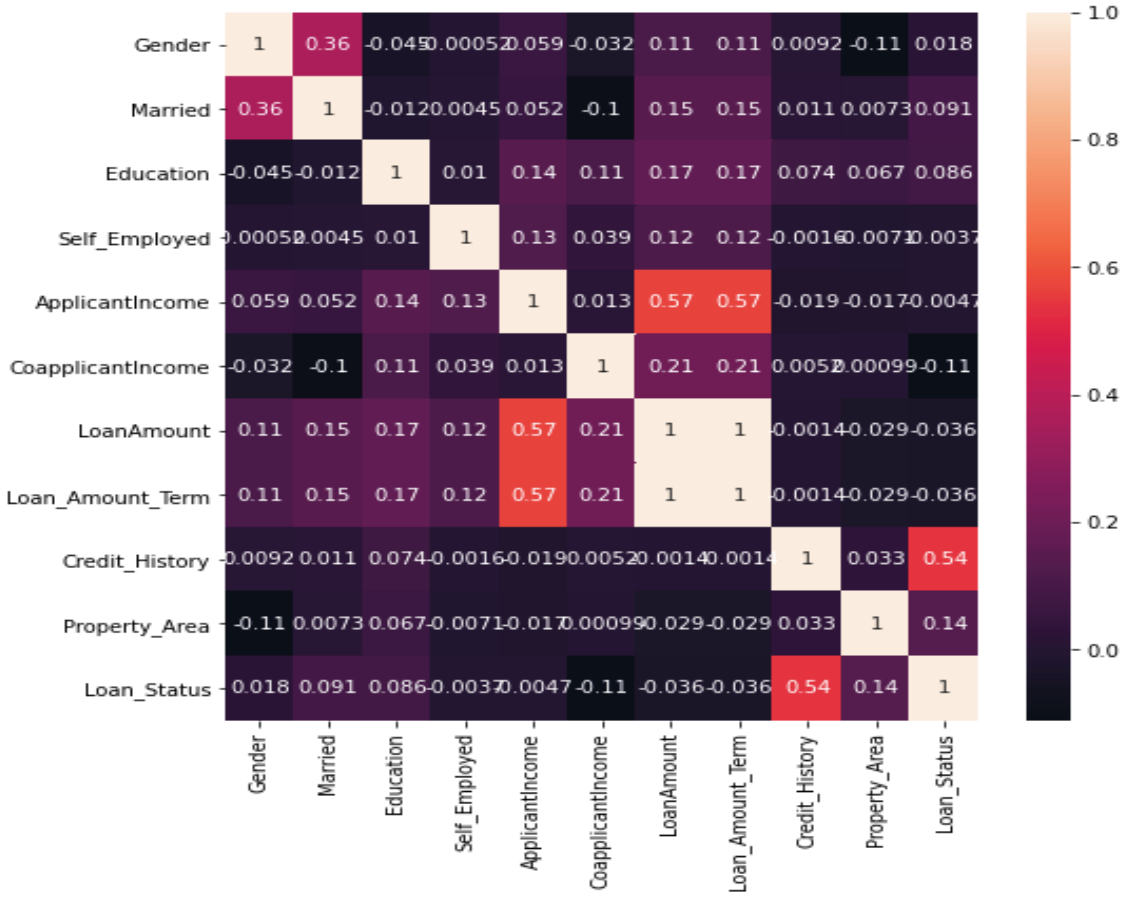
Στο «Διάγραμμα 20», αναπαρίστανται γραφικά το πλήθος των μετρήσεων της κλάσης του συνόλου (Loan Status), όπου η θετική κλάση αφορά στη λήψη του δανείου, ενώ η αρνητική αφορά στην απόρριψη της εταιρίας για λήψη του δανείου.



Διάγραμμα 20: Αναπαράσταση κλάσεων του συνόλου Loan Predication Data Set.

Όπως παρατηρείται, οι περισσότερες κλάσεις του συνόλου αποτελούν αιτήσεις δανείων ασφάλισης προς την εταιρία τα οποία, εν τέλει, εγκρίθηκαν.

Για την καλύτερη κατανόηση των σχέσεων των μεταβλητών του συνόλου, στη συνέχεια παρουσιάζεται και ο πίνακας («Εικόνα 6») συσχέτισης των μεταβλητών ο οποίος δημιουργήθηκε με τη χρήση της βιβλιοθήκης Seaborn.



Εικόνα 6: Πίνακας συσχέτισης Loan Predication Data Set.

Δύο πολύ ευδιάκριτες συσχετίσεις αποτελούν: α) η σχέση μεταξύ της λήψης του δανείου με το πιστωτικό ιστορικό του αιτούντα αλλά και ο τύπος της περιοχής που κατοικεί και, β) οι σχέσεις μεταξύ του ποσού και του όρου του δανείου σε σχέση με το εισόδημα του αιτούντα.

7.1.1 Γενικός έλεγχος του συνόλου

Σε αντίθεση με τις δύο προηγούμενες μελέτες περίπτωσης, το συγκεκριμένο σύνολο δεδομένων, χρειάστηκε μια ιδιαίτερη προσέγγιση στον τομέα της μορφοποίησης των στοιχείων του, για να γίνει έπειτα πιο αποτελεσματική η χρήση των αλγορίθμων κατηγοριοποίησης. Συγκεκριμένα, πραγματοποιήθηκε έλεγχος με διάφορες μεθόδους της βιβλιοθήκης Pandas για την εύρεση ελλιπών τιμών στο σύνολο. Στον «Πίνακα 12», παρουσιάζονται όλες οι ονομασίες των μεταβλητών και, το πλήθος των εμφανίσεων ελλιπών τιμών που βρέθηκαν για την κάθε μια από αυτές.

Πίνακας 12: Προετοιμασία της επεξεργασίας του Loan Predication Data Set.

Μεταβλητή	Εμφανίσεις
Gender	13
Married	3
Dependents	15
Self-Employed	32
Loan Amount	22
Loan Amount Term	14
Credit History	50

Όπως παρατηρείται, οι ελλιπείς τιμές στο συγκεκριμένο σύνολο ήταν αρκετές. Παρουσίας αυτών, ως αποτέλεσμα μετά την επεξεργασία του συνόλου θα μπορούσαν να εξαχθούν μη έγκυρα και λανθασμένα αποτελέσματα. Βάσει αυτού, όλες αυτές οι ελλιπείς τιμές συμπληρώθηκαν με την επικρατούσα τιμή της εκάστοτε μεταβλητής μέσα στο σύνολο.

Κάπου εδώ αξίζει να σημειωθεί ότι, επειδή η μεταβλητή CoApplicant Income είχε αρκετές μηδαμινές τιμές (συγκεκριμένα είχε 273), κρίθηκε απαραίτητη η συμπλήρωση αυτών με το μέσο όρο της στήλης αυτής.

Επιπλέον, πριν την εφαρμογή των αλγορίθμων στο συγκεκριμένο σύνολο, χρειάστηκε να πραγματοποιηθεί μια μετατροπή όλων των κατηγορηματικών μεταβλητών σε αριθμητικές. Συγκεκριμένα, οι μεταβλητές αυτές είναι οι: α) Gender, β) Self-Employed, γ) Education, δ) Married, ε) Property Area και στ) Loan Status.

7.1.2 Γενική μεθοδολογία μελέτης

Στη συγκεκριμένη ενότητα, γίνεται μια συνοπτική ανασκόπηση των βημάτων που ακολουθήθηκαν για την εφαρμογή των αλγορίθμων στο Loan Predication Data Set. Τα βήματα αυτά δίνονται στην συνέχεια:

- Φόρτωση του συνόλου στην Python: Κατά την αρχή της μελέτης, έγινε φόρτωση του συνόλου με τη βιβλιοθήκη Pandas στο IDE της Python για περαιτέρω επεξεργασία. Στη συνέχεια, πραγματοποιήθηκαν διάφορες μέθοδοι για την εύρεση ελλιπών και μηδαμινών τιμών στο σύνολο. Τα αποτελέσματα που εξήχθησαν καθώς και η μεθοδολογία που ακολουθήθηκε, αναφέρθηκε στην προηγούμενη ενότητα της πτυχιακής εργασίας.

- Διαχωρισμός των συνόλων για εκπαίδευση: Όπως και στις προηγούμενες μελέτες περιπτώσεων, ο διαχωρισμός των συνόλων σε σύνολο εκπαίδευσης, πραγματοποιήθηκε με τις δύο διαφορετικές προσεγγίσεις των ποσοστών του 80 και του 67 τοις εκατό.

- Εφαρμογή των αλγορίθμων: Για την εφαρμογή των αλγορίθμων ακολουθήθηκαν οι ίδιες προσεγγίσεις προ-επεξεργασίας που αναφέρθηκαν και στις προηγούμενες μελέτες περίπτωσης. Συγκεκριμένα, πραγματοποιήθηκε εφαρμογή των αλγορίθμων χωρίς προ-επεξεργασία, με προ-επεξεργασία του αλγορίθμου MinMax Scaler, την προ-επεξεργασία του αλγορίθμου Standard Scaler, αλλά και των δύο σε συνδυασμό.

- Αξιολόγηση των αποτελεσμάτων: Οι μετρικές αξιολόγησης όλων των διαφορετικών αποτελεσμάτων των αλγορίθμων σχετικά με την ακρίβεια και την ορθότητα που έφεραν σε όλες αυτές τις διαφορετικές εκδοχές προ-επεξεργασίας και εκπαίδευσης, υπολογίστηκαν σύμφωνα με το εργαλείο Metrics του Scikit-Learn.

7.2 Εφαρμογή αλγορίθμων

Στις ενότητες που ακολουθούν (7.2.1 μέχρι και 7.2.4), γίνεται παρουσίαση των αποτελεσμάτων των αλγορίθμων στις διαφορετικές εκδοχές της προ-επεξεργασίας και των συνόλων εκπαίδευσης. Όπως και στις δύο προηγούμενες μελέτες περιπτώσεων, τα αποτελέσματα αυτά αφορούν στην ακρίβεια και την ορθότητα του εκάστοτε μοντέλου, καθώς επίσης γίνεται και η γραφική αναπαράστασή τους με ραβδογράμματα.

7.2.1 Εφαρμογή κατηγοριοποιητών χωρίς προ-επεξεργασία

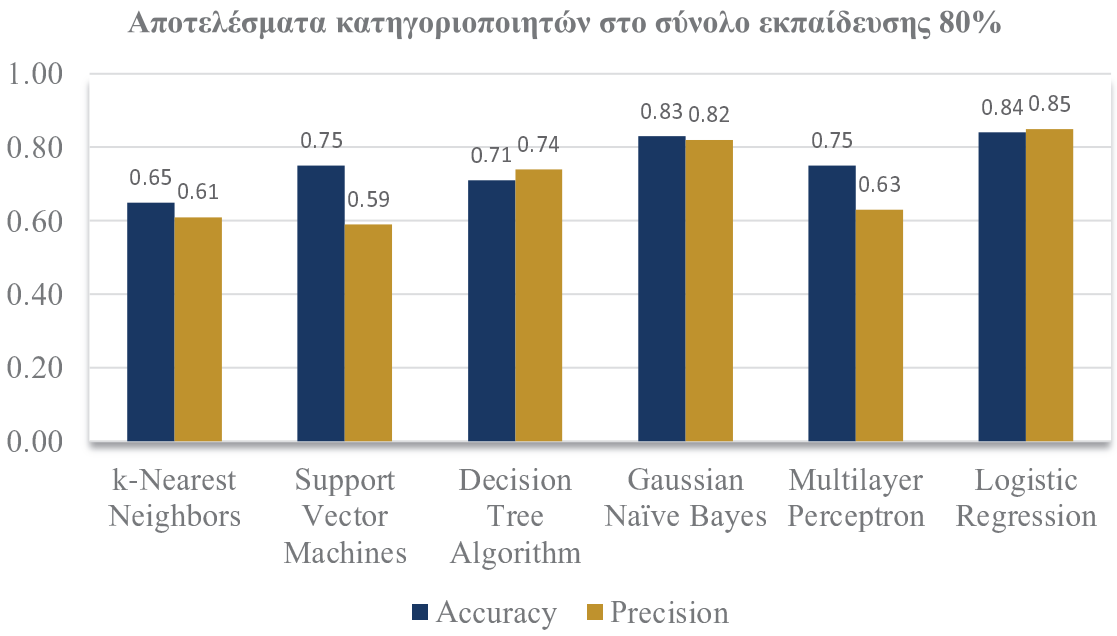
Στον ακόλουθο πίνακα, παρατίθενται τα αποτελέσματα που εξήχθησαν μετά την εφαρμογή των αλγορίθμων κατηγοριοποίησης στο σύνολο δεδομένων χωρίς τη χρήση κάποιου αλγόριθμου για την προ-επεξεργασία και τη μεταμόρφωσή του.

Πίνακας 13: Αποτελέσματα κατηγοριοποιητών χωρίς προ-επεξεργασία.

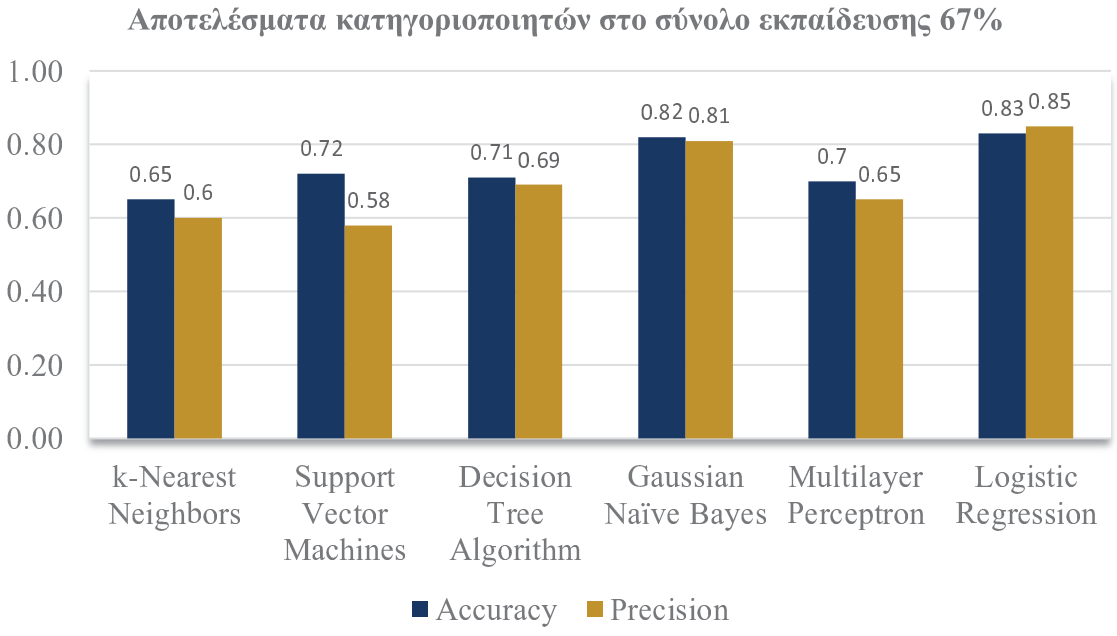
Αποτελέσματα αλγορίθμων κατηγοριοποίησης χωρίς προ-επεξεργασία				
	Μεγάλο Σύνολο Εκπαίδευσης (80%)		**Μικρό Σύνολο Εκπαίδευσης (67%)**	
Κατηγοριοποιητές	**Ορθότητα**	**Ακρίβεια**	**Ορθότητα**	**Ακρίβεια**
k-Nearest Neighbor	0.65	0.61	0.65	0.60
Support Vector Machines	0.75	0.59	0.72	0.58
Decision Tree	0.71	0.74	0.71	0.69
Naïve Bayes	0.83	0.82	0.82	0.81
Multilayer Perceptron	0.75	0.63	0.70	0.65
Logistic Regression	0.84	0.85	0.83	0.85

Ο αλγόριθμος που έφερε τα καλύτερα αποτελέσματα τόσο για το μικρότερο, όσο και για το μεγαλύτερο σύνολο εκπαίδευσης, είναι ο αλγόριθμος της λογιστικής παλινδρόμησης. Συγκεκριμένα, ο αλγόριθμος αυτός πέτυχε τα μεγαλύτερα ποσοστά ορθότητας και ακρίβειας στο μεγαλύτερο σύνολο εκπαίδευσης του 80%, με τιμές 0.84 και 0.85, αντίστοιχα. Η λογιστική παλινδρόμηση, καθώς επίσης και ο απλοϊκός Bayes, σε σύγκριση όλους με τους υπόλοιπους, δείχνουν να φέρουν αρκετά επιθυμητά αποτελέσματα ακόμα και χωρίς τη χρήση κάποιου αλγόριθμου προ-επεξεργασίας. Αυτός ο παράγοντας σχετίζεται κυρίως με τις μεγάλες αριθμητικές τιμές των μεταβλητών του Applicant και CoApplicant Income, καθώς και τη μεταβλητή Loan Amount, τις οποίες δείχνουν να διαχειρίζονται καλύτερα οι δύο αλγόριθμοι αυτοί.

Στα ακόλουθα διαγράμματα αναπαρίστανται γραφικά όλα τα στοιχεία του «Πίνακα 13». Συγκεκριμένα, το «Διάγραμμα 21» αφορά στα αποτελέσματα των αλγορίθμων στο μεγαλύτερο σύνολο εκπαίδευσης, ενώ το «Διάγραμμα 22» αναπαριστά τα αποτελέσματα που εξήχθησαν για το μικρότερο σύνολο εκπαίδευσης.



Διάγραμμα 21: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 80% χωρίς προ-επεξεργασία.



Διάγραμμα 22: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 67% χωρίς προ-επεξεργασία.

7.2.2 Εφαρμογή κατηγοριοποιητών με χρήση του Standard Scaler

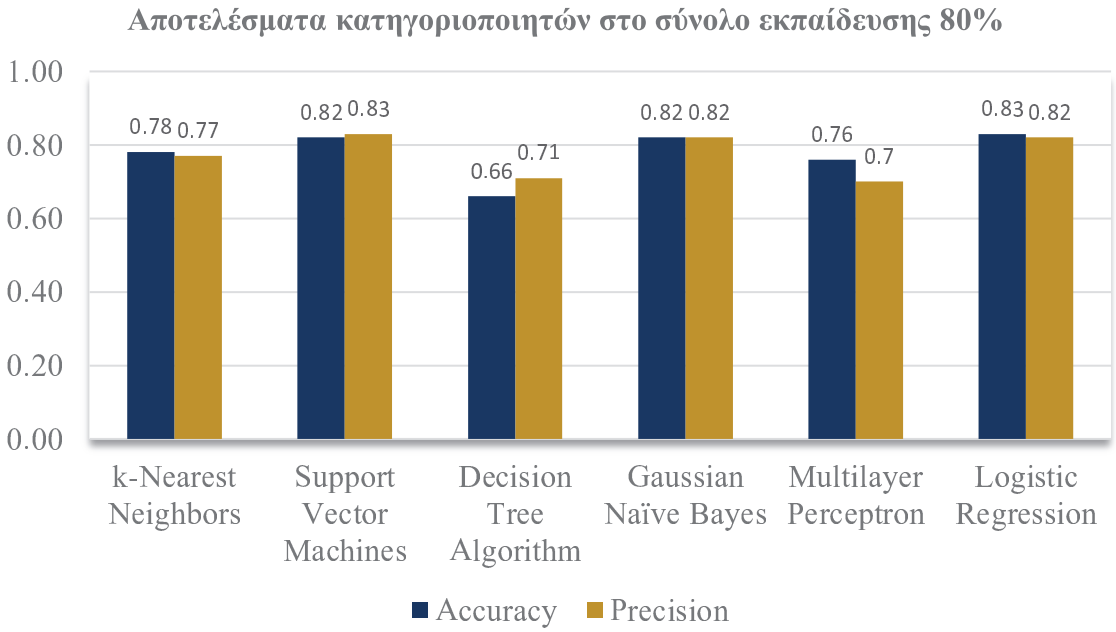
Στον «Πίνακα 14» που ακολουθεί, βρίσκονται καταγεγραμμένα τα αποτελέσματα των αλγορίθμων που εφαρμόστηκαν έπειτα από την προ-επεξεργασία του αλγορίθμου Standard Scaler, στις διαφορετικές περιπτώσεις του συνόλου εκπαίδευσης.

Πίνακας 14: Αποτελέσματα κατηγοριοποιητών με χρήση Standard Scaler.

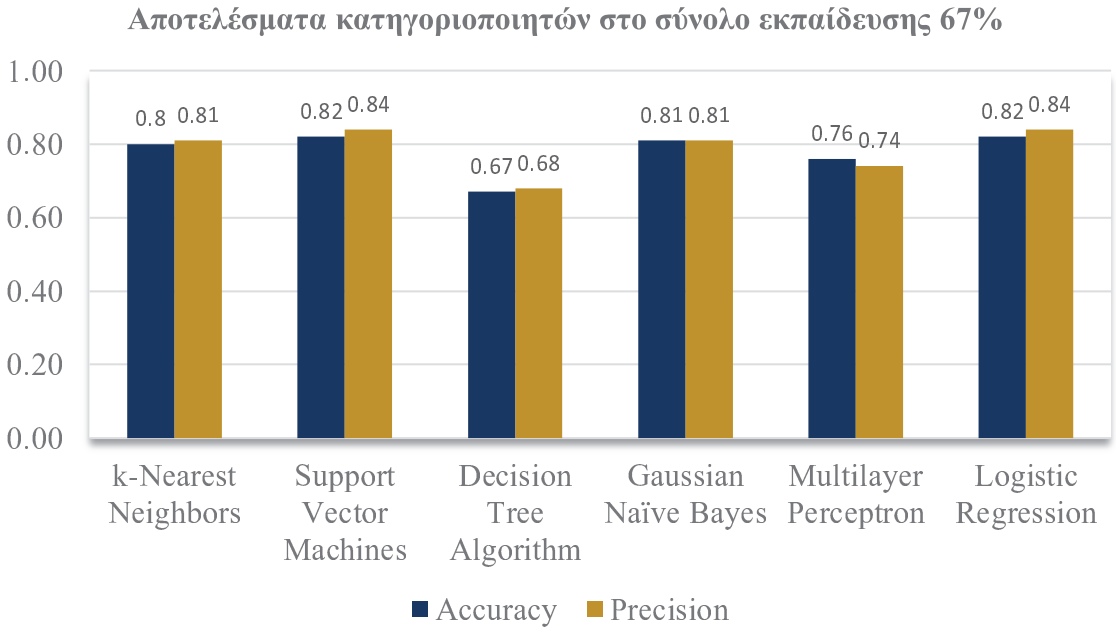
Αποτελέσματα αλγορίθμων κατηγοριοποίησης με χρήση του Standard Scaler				
	Μεγάλο Σύνολο Εκπαίδευσης (80%)		**Μικρό Σύνολο Εκπαίδευσης (67%)**	
Κατηγοριοποιητές	**Ορθότητα**	**Ακρίβεια**	**Ορθότητα**	**Ακρίβεια**
k-Nearest Neighbor	0.78	0.77	0.80	0.81
Support Vector Machines	0.82	0.83	0.82	0.84
Decision Tree	0.66	0.71	0.67	0.68
Naïve Bayes	0.82	0.82	0.81	0.81
Multilayer Perceptron	0.76	0.70	0.76	0.74
Logistic Regression	0.83	0.82	0.82	0.84

Όπως και στην προηγούμενη εφαρμογή αλγορίθμων της προηγούμενης ενότητας, η λογιστική παλινδρόμηση βρίσκεται πάλι να φέρνει τα καλύτερα αποτελέσματα. Συγκεκριμένα, στην περίπτωση του μεγαλύτερου συνόλου εκπαίδευσης ο αλγόριθμος της λογιστικής παλινδρόμησης έφερε αποτελέσματα στην ορθότητα και στην ακρίβεια ίσα με 0.83 και 0.82, αντίστοιχα. Σε γενικές γραμμές, όπως παρατηρείται, όλοι οι αλγόριθμοι μετά από το συγκεκριμένο είδος προ-επεξεργασίας, φέρνουν πιο επιθυμητές τιμές ορθότητας και ακρίβειας στα αποτελέσματά τους. Ο μόνος αλγόριθμος που δείχνει να μην παρουσιάζει βελτίωση αυτών, είναι ο αλγόριθμος του δένδρου αποφάσεων.

Στο «Διάγραμμα 23» και «Διάγραμμα 24», αναπαρίστανται γραφικά τα αποτελέσματα του πίνακα που προηγήθηκε.



Διάγραμμα 23: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 80% με χρήση του Standard Scaler.



Διάγραμμα 24: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 67% με χρήση του Standard Scaler.

7.2.3 Εφαρμογή κατηγοριοποιητών με χρήση του MinMax Scaler

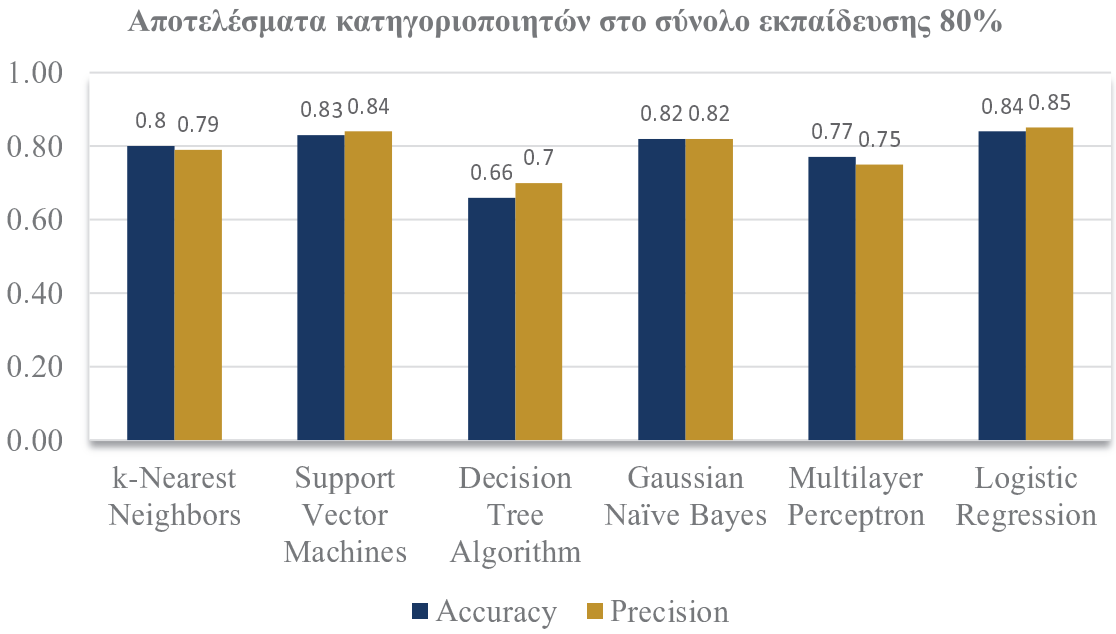
Στον ακόλουθο πίνακα («Πίνακας 15»), παρατίθενται τα αποτελέσματα των αλγορίθμων μετά την χρήση της προ-επεξεργασίας του αλγορίθμου MinMax Scaler:

Πίνακας 15: Αποτελέσματα κατηγοριοποιητών με χρήση MinMax Scaler.

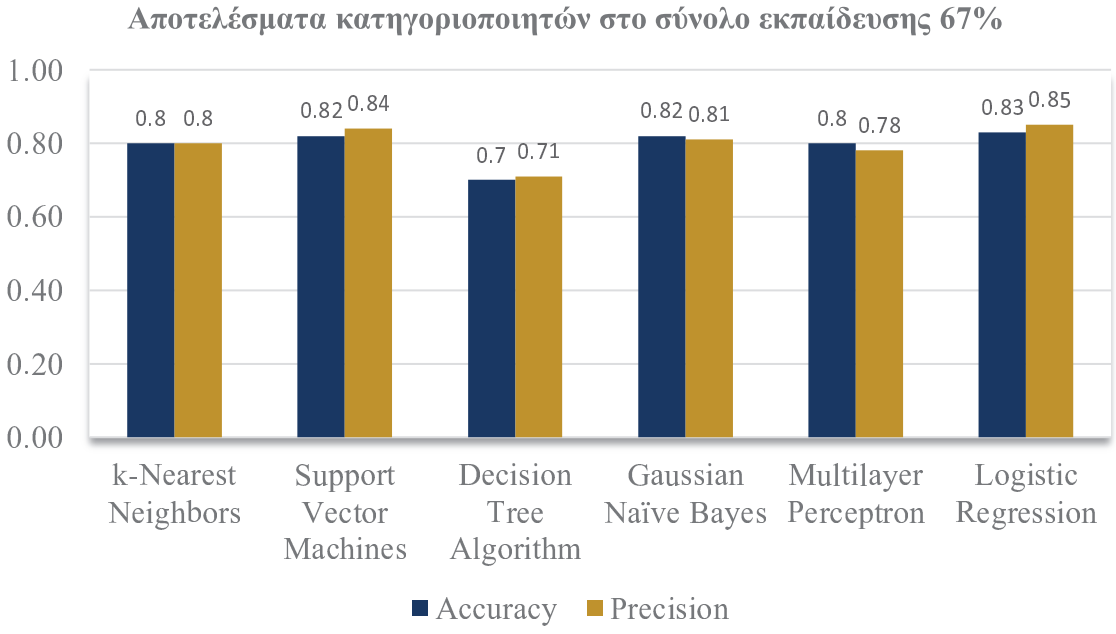
Αποτελέσματα αλγορίθμων κατηγοριοποίησης με χρήση του MinMax Scaler				
	Μεγάλο Σύνολο Εκπαίδευσης (80%)		**Μικρό Σύνολο Εκπαίδευσης (67%)**	
Κατηγοριοποιητές	**Ορθότητα**	**Ακρίβεια**	**Ορθότητα**	**Ακρίβεια**
k-Nearest Neighbor	0.80	0.79	0.80	0.80
Support Vector Machines	0.83	0.84	0.82	0.84
Decision Tree	0.66	0.70	0.70	0.71
Naïve Bayes	0.82	0.82	0.82	0.81
Multilayer Perceptron	0.77	0.75	0.80	0.78
Logistic Regression	0.84	0.85	0.83	0.85

Όπως παρατηρείται, ο αλγόριθμος των μηχανών διανυσμάτων υποστήριξης αλλά και η λογιστική παλινδρόμηση είναι οι δύο αυτοί αλγόριθμοι που έφερναν τα καλύτερα αποτελέσματα στο μεγαλύτερο σύνολο εκπαίδευσης, με τα ποσοστά της ορθότητας και της ακρίβειας τους να κυμαίνονται από 0.83 μέχρι 0.85. Όσον αναφορά στο μικρότερο σύνολο εκπαίδευσης, η λογιστική παλινδρόμηση βρίσκεται για άλλη μια φορά ο καλύτερος αλγόριθμος στα επιδιωκόμενα αποτελέσματα που αναλύουμε. Συγκεκριμένα, η λογιστική παλινδρόμηση στο μικρότερο σύνολο εκπαίδευσης με το ποσοστό του 67%, φέρνει ποσοστά ορθότητας και ακρίβειας 0.83 και 0.85. Η συγκεκριμένη περίπτωση προ-επεξεργασίας δείχνει να λειτουργεί πιο επιθυμητά σε σχέση με αυτήν του Standard Scaler, αφού σχεδόν όλοι οι αλγόριθμοι δείχνουν να φέρνουν καλύτερα αποτελέσματα.

Τα ακόλουθα διαγράμματα («Διάγραμμα 25» και «Διάγραμμα 26»), αναπαριστούν γραφικά τα αποτελέσματα των αλγορίθμων για το μεγαλύτερο και το μικρότερο σύνολο εκπαίδευσης, μετά από αυτήν την προ-επεξεργασία του MinMax Scaler.



Διάγραμμα 25: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 80% με χρήση του MinMax Scaler.



Διάγραμμα 26: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 67% με χρήση του MinMax Scaler.

7.2.4 Εφαρμογή κατηγοριοποιητών με χρήση MinMax και Standard Scaler

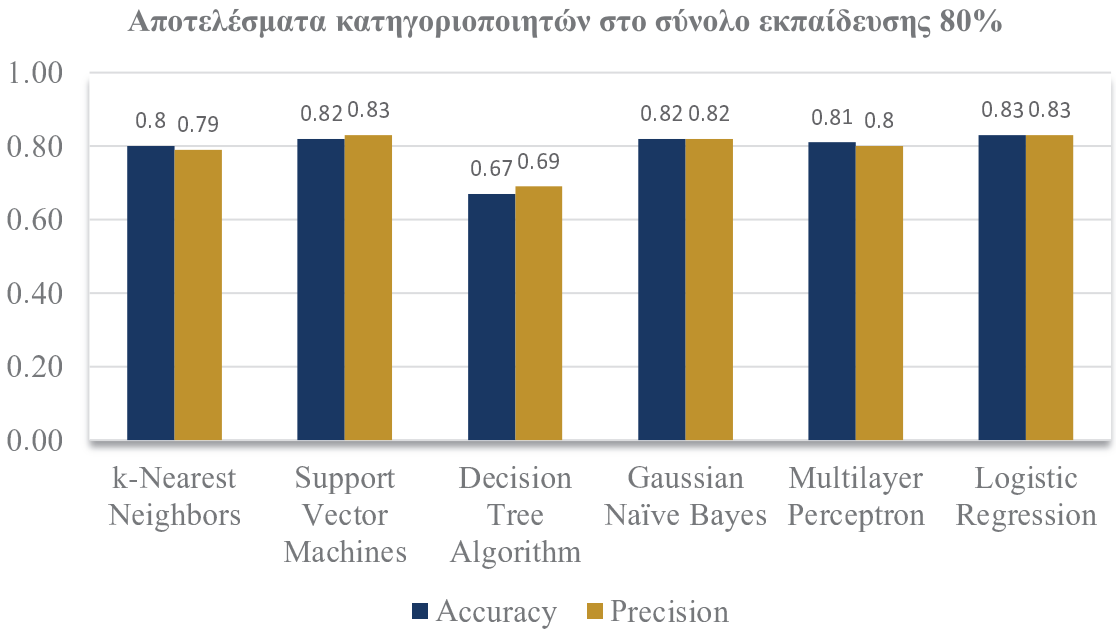
Στον «Πίνακα 16» που ακολουθεί, αναπαρίστανται τα αποτελέσματα των αλγορίθμων για το σύνολο Loan Predication, έπειτα από την εφαρμογή των αλγορίθμων προ-επεξεργασίας των δεδομένων MinMax και Standard Scaler.

Πίνακας 16: Αποτελέσματα κατηγοριοποιητών με χρήση MinMax και Standard Scaler.

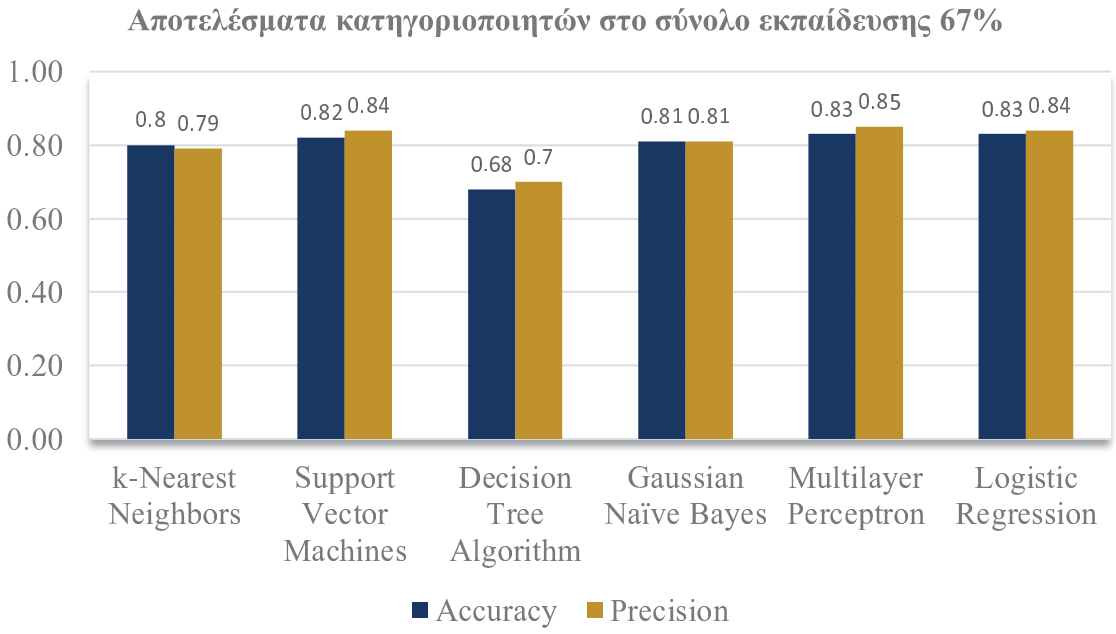
Αποτελέσματα αλγορίθμων κατηγοριοποίησης με χρήση MinMax και Standard Scaler				
	Μεγάλο Σύνολο Εκπαίδευσης (80%)		**Μικρό Σύνολο Εκπαίδευσης (67%)**	
Κατηγοριοποιητές	**Ορθότητα**	**Ακρίβεια**	**Ορθότητα**	**Ακρίβεια**
k-Nearest Neighbor	0.80	0.79	0.80	0.79
Support Vector Machines	0.82	0.83	0.82	0.84
Decision Tree	0.67	0.69	0.68	0.70
Naïve Bayes	0.82	0.82	0.81	0.81
Multilayer Perceptron	0.81	0.80	0.83	0.85
Logistic Regression	0.83	0.83	0.83	0.84

Ο αλγόριθμος που έφερε τα καλύτερα αποτελέσματα στην ακρίβεια και την ορθότητα για το μεγαλύτερο σύνολο εκπαίδευσης, βρίσκεται να είναι ο αλγόριθμος της λογιστικής παλινδρόμησης. Τα αποτελέσματα αυτά του αλγορίθμου, είναι ίσα με 0.83. Όσον αναφορά στο μικρότερο σύνολο εκπαίδευσης, το νευρωνικό δίκτυο δείχνει να φέρνει τα καλύτερα αποτελέσματα με 0.83 στην ορθότητα και, 0.85 στην ακρίβειά του. Η συγκεκριμένη προσέγγιση προ-επεξεργασίας δείχνει να μην λειτουργεί τόσο επιθυμητά όσο αυτή της απλής προ-επεξεργασίας με τον MinMax Scaler. Η μόνη εξαίρεση αποτελεί το νευρωνικό δίκτυο, το οποίο έδειξε μια σημαντική βελτίωση στα αποτελέσματά του σε σχέση με τις προηγούμενες εφαρμογές προ-επεξεργασίας.

Στο «Διάγραμμα 27» και «Διάγραμμα 28» που ακολουθούν, παρατίθενται γραφικά τα αποτελέσματα των αλγορίθμων που εφαρμόστηκαν μετά τη χρήση του συγκεκριμένου είδους προ-επεξεργασίας του MinMax και Standard Scaler.



Διάγραμμα 27: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 80% με χρήση του MinMax και Standard Scaler.



Διάγραμμα 28: Αποτελέσματα αλγορίθμων στο σύνολο εκπαίδευσης 67% με χρήση του MinMax και Standard Scaler.

8 Αξιολόγηση αποτελεσμάτων

Στο παρόν κεφάλαιο, πραγματοποιείται μια ανασκόπηση και επεξήγηση των αποτελεσμάτων, έπειτα από την εφαρμογή των αλγορίθμων μηχανικής μάθησης που πραγματοποιήθηκαν στις τρείς αυτές μελέτες περίπτωσης. Επιπροσθέτως, στην τελευταία ενότητα του παρόντος κεφαλαίου, πραγματοποιείται μια συνολική συγκριτική αξιολόγηση μεταξύ των αποτελεσμάτων αυτών που εξήχθησαν από το κάθε ένα σύνολο δεδομένων που μελετήθηκε.

8.1 Αξιολόγηση Student Performance Data Set

Ο αλγόριθμος της γραμμικής παλινδρόμησης, όπως και οι αλγόριθμοι κατηγοριοποίησης που εφαρμόστηκαν στις δύο εκδοχές του συνόλου Student Performance, δείχνουν να φέρνουν αρκετά υψηλά ποσοστά επιτυχίας στην ορθότητα και στην ακρίβεια των αποτελεσμάτων τους. Ένας βασικός παράγοντας για την επίτευξη τόσο μεγάλων ποσοστών των αποτελεσμάτων, αποτελεί η μορφή του συνόλου των δεδομένων καθώς και οι μεταβλητές που το απαρτίζουν. Στην περίπτωση αυτή που μελετήθηκε, επιλέχθηκαν, όπως αναφέρθηκε, μεταβλητές οι οποίες είχαν τις μεγαλύτερες συσχετίσεις αναφορικά με την απόδοση των μαθητών στις εξεταστικές τους περιόδους. Επιπλέον, το συγκεκριμένο σύνολο δεδομένων, απαρτιζόταν από πολλές μεταβλητές με αρκετά «μικρές» και «απλές» αριθμητικές τιμές. Αυτοί οι δύο παράγοντες που μόλις αναφέρθηκαν, είναι οι δύο κύριοι λόγοι για την επίτευξη μεγάλων αποτελεσμάτων ακρίβειας και ορθότητας των αλγορίθμων.

Η ταξινόμηση της απόδοσης των μαθητών, καθώς και η πρόβλεψη της τελικής βαθμολογίας τους, καθίσταται σχεδόν σίγουρη έπειτα από την ανάλυση και την επεξεργασία του συνόλου αυτού. Συγκεκριμένα, όλα σχεδόν τα αποτελέσματα της ορθότητας και της ακρίβειας, δείχνουν να φέρουν ποσοστά μεγαλύτερα του 0.8 περίπου, πράγμα το οποίο καθιστά σχεδόν βέβαιη την κατηγοριοποίηση ενός μη ετικετοποιημένου δεδομένου στη σωστή κλάση, ή την πρόβλεψη της τιμής μιας μεταβλητής. Κάπου εδώ αξίζει να σημειώσουμε ότι, όλα σχεδόν τα αποτελέσματα δείχνουν να είναι καλύτερα σε έναν βαθμό, με τη χρήση του μεγαλύτερου συνόλου εκπαίδευσης.

Τα καλύτερα αποτελέσματα που βρέθηκαν όσον αναφορά στον αλγόριθμο της γραμμικής παλινδρόμησης, είναι αυτά, μετά την προ-επεξεργασία του συνόλου με τον αλγόριθμο Standard Scaler. Όσον αναφορά στα αποτελέσματα των αλγόριθμων κατηγοριοποίησης στο συγκεκριμένο σύνολο δεδομένων, ο αλγόριθμος του δένδρου αποφάσεων φαίνεται να είναι ο καλύτερος στα αποτελέσματα που αναζητούσαμε από όλους τους υπόλοιπους. Τα πιο επιθυμητά αποτελέσματα από αυτόν τον αλγόριθμο πραγματοποιήθηκαν, μετά την διπλή προ-επεξεργασία με τους αλγόριθμους MinMax και Standard Scaler.

8.2 Αξιολόγηση Diabetes Data Set

Η λήψη αποφάσεων, ειδικά σε τομείς που αφορούν στην υγεία και τη βιωσιμότητα της ζωής των ανθρώπων, αποτελεί ένα πολύ σημαντικό ζήτημα. Με την εφαρμογή των αλγορίθμων στο συγκεκριμένο σύνολο δεδομένων, επιδιώχθηκε, όπως ήδη αναφέρθηκε, η εύρεση του καλύτερου κατηγοριοποιητή για την ταξινόμηση ασθενών στη σωστή κλάση και, η έγκαιρη πρόγνωση της νόσου του διαβήτη. Σε σχέση με το προηγούμενο σύνολο δεδομένων, οι αριθμητικές τιμές των μεταβλητών σε αυτήν την περίπτωση ήταν αρκετά περίπλοκες. Ως συνέπεια, τα αποτελέσματα των αλγορίθμων δεν έφεραν τόσο μεγάλα ποσοστά ακρίβειας και ορθότητας σε σύγκριση με αυτά του Student Performance Data Set.

Συγκεκριμένα, η χρήση των δύο αλγορίθμων της προ-επεξεργασίας, MinMax και Standard Scaler, έδειξαν να βελτιώνουν αρκετά τα αποτελέσματα των αλγορίθμων μετά την εφαρμογή τους στο σύνολο. Ο μόνος αλγόριθμος που δεν έδειξε να παρουσιάζει ιδιαίτερη βελτίωση, είναι ο αλγόριθμος του δένδρου αποφάσεων. Ο συγκεκριμένος αλγόριθμος, έφερε, σχεδόν σε όλες τις περιπτώσεις, τα χαμηλότερα ποσοστά στην ακρίβεια και στην ορθότητα. Όπως και στην προηγούμενη μελέτη περίπτωσης, οι αλγόριθμοι δείχνουν να φέρουν καλύτερα αποτελέσματα με τη χρήση της μεγαλύτερης εκδοχής του συνόλου εκπαίδευσης.

Ο αλγόριθμος που έφερε τα μεγαλύτερα ποσοστά ακρίβειας και ορθότητας, είναι αυτός της λογιστικής παλινδρόμησης. Τα αποτελέσματα αυτά πραγματοποιήθηκαν στο μεγαλύτερο σύνολο εκπαίδευσης, έπειτα από τη διπλή προ-επεξεργασία των δύο τεχνικών του MinMax και Standard Scaler. Τα ακριβώς επόμενα καλύτερα αποτελέσματα, δόθηκαν από τον αλγόριθμο του νευρωνικού δικτύου τύπου Perceptron. Ο αλγόριθμος αυτός είχε

αρκετά μικρές διαφορές στα αποτελέσματά του, σε σύγκριση με αυτόν της λογιστικής παλινδρόμησης.

8.3 Αξιολόγηση Loan Predication Data Set

Ο αυτοματοποιημένος τρόπος λήψης καινοτόμων και στρατηγικών αποφάσεων σε επιχειρήσεις και οργανισμούς, είναι ένα πολύ σημαντικό ζήτημα στο οποίο πλέον μπορεί να επιτευχθεί η βέλτιστη προσέγγιση, με τη χρήση των κατάλληλων στατιστικών και αλγοριθμικών τεχνικών που εμπίπτουν στο πεδίο της μηχανικής μάθησης και της εξόρυξης πληροφορίας. Έτσι, η εφαρμογή των αλγορίθμων σε αυτό το σύνολο των δεδομένων, πραγματοποιήθηκε με σκοπό την εύρεση του καλύτερου κατηγοριοποιητή που θα μπορούσε να προβλέπει τις κλάσεις μη ετικετοποιημένων δεδομένων ενός ασφαλιστικού φορέα. Όπως και στο σύνολο Diabetes Data Set, οι τιμές των μεταβλητών απαρτίζονταν από περίπλοκες και μεγάλες αριθμητικές τιμές. Ως αποτέλεσμα, δεν έγινε εφικτή η εύρεση μιας τόσο καλής τιμής ορθότητας και ακρίβειας όπως στη μελέτη περίπτωσης του Student Performance Data Set, η οποία να ξεπερνούσε το 0.90.

Η εφαρμογή των αλγορίθμων χωρίς προ-επεξεργασία στο συγκεκριμένο σύνολο, έδειξε να φέρει ποσοστά, τα οποία δεν θα ικανοποιούσαν πλήρως τις βλέψεις και τις απαιτήσεις ενός επιχειρηματικού φορέα, σαν αυτόν που μελετήθηκε. Οι μόνες εξαιρέσεις από αυτούς τους αλγόριθμους είναι ο απλοϊκός Bayes και η λογιστική παλινδρόμηση. Οι δύο αλγόριθμοι αυτοί, σε σύγκριση με όλους τους άλλους, έδειξαν να φέρουν αρκετά μεγάλα ποσοστά ακρίβειας και ορθότητας ακόμα και χωρίς τη χρήση κάποιου αλγόριθμου προ-επεξεργασίας. Κάπου εδώ επιβεβαιώνεται ότι αυτά τα δυο πιθανοτικά μοντέλα, αποτελούν, τις καλύτερες τεχνικές για δεδομένα με μεγάλες και περίπλοκες αριθμητικές τιμές.

Σε αυτό το σημείο αξίζει να επισημανθεί, ότι μετά την εφαρμογή των αλγορίθμων MinMax και Standard Scaler, όλοι οι αλγόριθμοι έδειξαν να φέρουν μια σημαντική βελτίωση στα αποτελέσματά τους, με μόνη εξαίρεση να αποτελεί ο αλγόριθμος του δένδρου αποφάσεων. Όπως και στη μελέτη περίπτωσης του Diabetes Data Set, ο συγκεκριμένος αλγόριθμος του δένδρου δείχνει να μην είναι σε θέση να διαχειριστεί μεγάλες και περίπλοκες αριθμητικές τιμές.

Ο αλγόριθμος που έφερε τα καλύτερα αποτελέσματα στα ποσοστά της ακρίβειας και της ορθότητάς του για άλλη μια φορά, είναι αυτός της λογιστικής παλινδρόμησης. Τα αποτελέσματα αυτά πραγματοποιήθηκαν από τον αλγόριθμο αυτόν μετά την εφαρμογή της προ-επεξεργασίας του MinMax Scaler, για το μεγαλύτερο σύνολο εκπαίδευσης. Οι επόμενοι αλγόριθμοι που δείχνουν να φέρνουν τα καλύτερα αποτελέσματα, είναι ο απλοϊκός Bayes, το νευρωνικό δίκτυο τύπου Perceptron και, ο αλγόριθμος των μηχανών διανυσμάτων υποστήριξης.

8.4 Συγκριτική αξιολόγηση αποτελεσμάτων και γενικά αποτελέσματα

Ο ισχυρισμός ότι τα αποτελέσματα των αλγορίθμων μηχανικής μάθησης εξαρτώνται τόσο από τη μορφή όσο και από το περιεχόμενο ενός συνόλου δεδομένων, αποδεικνύεται πολύ εύκολα με μια απλή παρατήρηση όλων των αποτελεσμάτων που δόθηκαν στα προηγούμενα κεφάλαια (5ο, 6ο και 7ο Κεφάλαιο). Συγκεκριμένα, όπως ήδη αναφέρθηκε, τα καλύτερα αποτελέσματα των αλγορίθμων πραγματοποιήθηκαν στο σύνολο δεδομένων Students Performance Data Set. Το συγκεκριμένο σύνολο, αποτελείτο από απλές τιμές στις μεταβλητές του, σε σύγκριση με τα άλλα δύο σύνολα που μελετήθηκαν, το Diabetes και το Loan Predication Data Set δηλαδή. Τα αποτελέσματα που εξήχθησαν στα τρία αυτά σύνολα δεδομένων, είναι αρκετά «επιθυμητά» και, οι αλγόριθμοι που έφεραν τα καλύτερα αποτελέσματα θα μπορούσαν κάλλιστα να χρησιμοποιηθούν στο μέλλον για την αναγνώριση και την ταξινόμηση της κλάσης ενός νέου στιγμιότυπου που μπορεί να εισαχθεί στο πρόγραμμα.

Μια σημαντική παρατήρηση που πρέπει να γίνει κάπου εδώ είναι ότι, όλα τα αποτελέσματα των αλγορίθμων έδειξαν να είναι -σχεδόν σε όλες- τις περιπτώσεις τους, καλύτερα με την χρήση του μεγαλύτερου συνόλου εκπαίδευσης.

Στην περίπτωση της πολλαπλής κατηγοριοποίησης που μελετήθηκε στην πρώτη μελέτη περίπτωσης, του συνόλου Students Performance Data Set δηλαδή, ο αλγόριθμος που έδειξε να προσαρμόζεται καλύτερα και να φέρνει τα πιο επιθυμητά αποτελέσματα, ήταν αυτός του δένδρου αποφάσεων. Σε σύγκριση με τις δύο υπόλοιπες μελέτες περιπτώσεων, οι οποίες αποτελούν προβλήματα δυαδικής κατηγοριοποίησης, ο αλγόριθμος αυτός δείχνει να προσαρμόζεται καλύτερα μόνο στην περίπτωση των μικρότερων αριθμητικών τιμών που μελετήθηκαν κατά βάσει σε αυτήν την πρώτη μελέτη. Οι επόμενοι αλγόριθμοι που

«ξεχωρίζουν» στα αποτελέσματά τους για αυτό το σύνολο δεδομένων, είναι ο απλοϊκός Bayes και το νευρωνικό δίκτυο τύπου Perceptron.

Όσον αναφορά στα υπόλοιπα δύο σύνολα (Diabetes και Loan Predication), ο αλγόριθμος ο οποίος «ξεχωρίζει» στα αποτελέσματά του από τους υπόλοιπους, είναι αυτός της λογιστικής παλινδρόμησης. Όπως επιβεβαιώνεται, ο αλγόριθμος αυτός δείχνει να φέρνει τα καλύτερα αποτελέσματα που μπορούν να επιτευχθούν για τα προβλήματα της δυαδικής κατηγοριοποίησης, «αγγίζοντας» μεγάλα ποσοστά επιτυχίας ακόμα και σε περιπτώσεις χωρίς τη χρήση κάποιας τεχνικής προ-επεξεργασίας για τη μεταμόρφωση των δεδομένων. Οι ακριβώς επόμενοι αλγόριθμοι που δείχνουν να φέρουν τα καλύτερα επιθυμητά αποτελέσματα, όπως και στην πρώτη μελέτη περίπτωσης με την πολλαπλή κατηγοριοποίηση, είναι ο απλοϊκός Bayes και το νευρωνικό δίκτυο τύπου Perceptron.

Οι τεχνικές τις προ-επεξεργασίας που εφαρμόστηκαν στα σύνολα δεδομένων, στην πλειονότητα των περιπτώσεών τους, έδειξαν να βελτιώνουν τα αποτελέσματα των αλγορίθμων, σε κάποιο βαθμό. Συγκεκριμένα, η προσέγγιση της διπλής προ-επεξεργασίας με τους αλγορίθμους MinMax και Standard Scaler ταυτόχρονα, είναι αυτή που «ξεχώρισε» από τις υπόλοιπες για τις δύο πρώτες μελέτες περιπτώσεων. Η εφαρμογή των αλγορίθμων για την τρίτη μελέτη περίπτωσης, του συνόλου Loan Predication δηλαδή, έδειξε να φέρει καλύτερα αποτελέσματα μετά τη χρήση του αλγόριθμου MinMax και μόνο. Σε σύγκριση με τα υπόλοιπα δύο σύνολα, το Loan Predication είχε μεταβλητές με αρκετά μεγάλες αριθμητικές τιμές. Συνεπώς, η μείωση των τιμών αυτών σε ένα μικρότερο διάστημα μέσω του αλγορίθμου MinMax, είναι αυτό, που κατέστησε εφικτή τη βελτίωση των αποτελεσμάτων των αλγορίθμων, χωρίς να υπάρχει ιδιαίτερη επιρροή από τον αλγόριθμο Standard Scaler.

Η μόνη περίπτωση που ο αλγόριθμος Standard Scaler φάνηκε ιδιαίτερης χρησιμότητας, είναι αυτή στην πρώτη μελέτη περίπτωσης, για την εφαρμογή της γραμμικής παλινδρόμησης. Συγκεκριμένα, σε αυτήν την περίπτωση, η τεχνική αυτή της προ-επεξεργασίας του Standard Scaler, έδειξε να κάνει εφικτά τα καλύτερα δυνατά αποτελέσματα για τον αλγόριθμο της γραμμικής παλινδρόμησης στην ορθότητά του.

Συμπεράσματα

Η ανάλυση και η επεξεργασία των δεδομένων με μεθόδους μηχανικής μάθησης και εξόρυξης πληροφορίας, είναι κάτι το οποίο πρέπει να αντιμετωπιστεί με ιδιαίτερη προσοχή, ειδικά για την εξαγωγή των πιο αξιόπιστων και καλύτερων πιθανών αποτελεσμάτων. Στην παρούσα πτυχιακή εργασία παρουσιάστηκαν αρχικά, διάφορα θεωρητικά στοιχεία που συντελούν τους δύο αυτούς τομείς τις μηχανικής μάθησης και της εξόρυξης πληροφορίας. Στη συνέχεια, έγινε η παρουσίαση των αποτελεσμάτων και των αξιολογήσεων τριών διαφορετικών μελετών περιπτώσεων συνόλων δεδομένων με αριθμητικές και κατηγορηματικές μεταβλητές, με σκοπό την εύρεση των καλύτερων προσεγγίσεων προ-επεξεργασίας και, την επιλογή των πιο κατάλληλων μοντέλων.

Μετά τις τρείς διαφορετικές μελέτες περίπτωσης που διερευνήθηκαν, κρίνεται απαραίτητο να αναφερθεί ότι, η μετατροπή ενός συνόλου δεδομένων με τα κατάλληλα εργαλεία, σε μια «κατάλληλα επεξεργάσιμη μορφή», είναι πρώτιστης σημασίας για την εξαγωγή έγκυρων και αξιόπιστων αποτελεσμάτων. Ο όρος «κατάλληλα επεξεργάσιμη μορφή» που μόλις αναφέρθηκε είναι διττός. Για να επιτευχθεί αυτή η μορφή του συνόλου των δεδομένων, θα πρέπει να γίνουν μια σειρά από διάφορες ενέργειες, όπως για παράδειγμα ο έλεγχος για ελλιπείς και για μηδαμινές τιμές, η μεταμόρφωση των τιμών των μεταβλητών αλλά και η πιθανή μείωση των διαστάσεών τους. Όταν επιτευχθεί η πλειονότητα των παραπάνω ενεργειών, οι αλγόριθμοι επεξεργασίας που θα εφαρμοστούν στο εκάστοτε σύνολο θα αποδώσουν αρκετά καλύτερα στα αποτελέσματά τους, καθιστώντας τις αξιολογήσεις και τις λήψεις αποφάσεων, από πλευράς των ανθρώπων, πιο βέβαιες και έγκυρες.

Η αξιοπιστία και η εγκυρότητα των αποτελεσμάτων έπειτα από την εφαρμογή αλγορίθμων μηχανικής μάθησης σε μια συνεδρία επιβλεπόμενης μάθησης, είναι ζητήματα ιδιαίτερης σημασίας, ιδίως σε περιπτώσεις που πρέπει να πραγματοποιηθεί η λήψη μιας «ριψοκίνδυνης» και «δύσκολης» απόφασης κάτω από αβεβαιότητα. Ένα πολύ απλό παράδειγμα βάσει αυτών, είναι η λήψη αποφάσεων και η πρόβλεψη γεγονότων σε τομείς της ιατρικής, όπου και το πιο μικρό λάθος μπορεί να αποβεί μοιραίο. Ένα δεύτερο παράδειγμα αποτελούν οι προβλέψεις και οι κατηγοριοποιήσεις γεγονότων σε οικονομικούς, επιχειρησιακούς και διοικητικούς τομείς, όπου διακυβεύονται μεγάλα χρηματικά ποσά και κεφαλαιουχικά αγαθά.

Οι εφαρμογές των αλγορίθμων που πραγματοποιήθηκαν για την παρούσα πτυχιακή εργασία, θέτουν ένα βασικό θεμέλιο για την περεταίρω ενασχόλησή με τη διεξοδική επεξεργασία και ανάλυση δεδομένων με μεθόδους μηχανικής μάθησης. Συγκεκριμένα, γίνεται εκτίμηση μεγαλύτερων ποσοστών επιτυχίας για τα σύνολα δεδομένων αυτά, σε περίπτωση εφαρμογής ενός αλγορίθμου μείωσης των διαστάσεων, όπως ο αλγόριθμος PCA που αναφέρθηκε, ή σε περίπτωση ενός πιο διεξοδικού ελέγχου για την εξάλειψη δεδομένων με θόρυβο ή διπλότυπων εγγραφών από το σύνολο που μελετάται.

Οι εφαρμογές της μηχανικής μάθησης, φυσικά, δεν περιορίζονται μόνο σε ποσοτικά και ποιοτικά δεδομένα, όπως αυτά που μελετήθηκαν. Το επιστημονικό πεδίο αυτό σήμερα, έχει αναπτυχθεί τόσο πολύ, που πλέον γίνεται χρήση των τεχνικών αυτών για την αναγνώριση της φυσικής γλώσσας του ανθρώπου, την αναγνώριση προτύπων σε φωτογραφίες, κείμενα κλπ. Όλα αυτά τα ευφυή συστήματα υπολογιστών που εμπίπτουν στις τεχνολογίες της μηχανικής μάθησης, δείχνουν να έχουν «κατακλύσει» τον κόσμο μας σήμερα, καθιστώντας διάφορες ανθρώπινες ενέργειες και λήψεις αποφάσεων, πιο εύκολες από ποτέ.

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΕΛΛΗΝΟΓΛΩΣΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

- Γαρμπής, Α. & Φωτιάδης, Δ. (2015). *Εισαγωγή στους Υπολογιστές και την Πληροφορική*, Αθήνα: Πανεπιστημιακές εκδόσεις Αράκυνθος.
- Γεωργούλη, Κ. (2015). *Τεχνητή Νοημοσύνη-Μια εισαγωγική Προσέγγιση*, Αθήνα: Σύνδεσμος Ελληνικών ακαδημαϊκών βιβλιοθηκών, Εθνικό Μετσόβιο Πολυτεχνείο [on line]. Ανακτήθηκε 23 Μαΐου 2021, από https://repository.kallipos.gr/handle/11419/3381
- Γναρδέλλης, Χ. (2003). *Εφαρμοσμένη στατιστική*, Αθήνα: Εκδόσεις Παπαζήση .
- Καραντζιάς, Π. (2019). *Εφαρμογή αλγορίθμων μηχανικής μάθησης σε σύνολα δεδομένων και αποτίμηση αποτελεσμάτων* (Μεταπτυχιακή Εργασία). Πανεπιστήμιο Πειραιά, Πειραιάς.
- Ματσατσίνης, Ν. (2010). *Συστήματα υποστήριξης αποφάσεων*, Αθήνα: Εκδόσεις Νέων Τεχνολογιών.
- Μπέχρουζ, Φ. (2015). *Εισαγωγή στην επιστήμη των υπολογιστών - Τρίτη έκδοση*, Αθήνα: Εκδόσεις Κλειδάριθμος.
- Ρόιγκερ, Ρ. & Γκίατζ, Μ. (2008). *Εξόρυξη πληροφορίας - Ένας εισαγωγικός οδηγός με παραδείγματα*, Αθήνα: Εκδόσεις Κλειδάριθμος.
- Ρώτα, Μ.,Σ. (2008). *Σύγκριση Κλασσικού και Ελέγχου Βασισμένου σε Ασαφή Λογική Ανεμογεννήτριας Μονίμων Μαγνητών* (Μεταπτυχιακή Εργασία). Εθνικό Μετσόβιο Πολυτεχνείο, Αθήνα.
- Ταμπακάς, Β.,Τ. (2017). *Εισαγωγή στις βάσεις δεδομένων*, Πάτρα: Εκδόσεις Γκότσης

ΞΕΝΟΓΛΩΣΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

- Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*, Great Britain, CPI Bath [on line]. Ανακτήθηκε 22 Ιουνίου, 2021, από http://home.elka.pw.edu.pl/~ptrojane/books/Bishop%20-%20Neural%20Networks%20for%20Pattern%20Recognition.pdf.
- Brijain, R.P. & Kushik, K.R (2014). *A Survey on Decision Tree Algorithm for Classification*. India: GEC Modasa, Department of computer engineering [on line].

Ανακτήθηκε 25 Αυγούστου, 2021, από http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=F94ABEED5292A7AE642A0B1CB2CA22E4?doi=10.1.1.673.2797&rep=rep1&type=pdf.

- Brunton, S.L. & Kutz, N.J. (2017). *Data Driven Science & Engineering: Machine Learning, Dynamical Systems, and Control*. USA: University of Washington.
- Cortez, P. & Silva, A., n.d. *Using Data Mining to predict secondary school student performance*. Portugal: University of Minho [on line]. Ανακτήθηκε 25 Αυγούστου, 2021, από http://www3.dsi.uminho.pt/pcortez/student.pdf.
- Erhard, R. & Hong, H., D., n.d. *Data Cleaning: Problems and Current Approaches.* Germany: University of Leipzig [on line]. Ανακτήθηκε 28 Ιουνίου, 2021, από: https://web.archive.org/web/20170809153257/http://lips.informatik.uni-leipzig.de/files/2000-45.pdf.
- Goodfellow, G. & Bengio, Y. & Courville, A. (2016) *Deep Learning*, Massachusetts: The MIT Press.
- Han, J. & Kamber, M. & Pei, J. (2012). *Data mining – Concepts and techniques – Third edition*, USA: Morgan Kaufman Publishers [on line]. Ανακτήθηκε 14 Ιουνίου, 2021, από http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf.
- Klement, P. & Slany, W. (1994). *Fuzzy Logic in Artificial Intelligence,* Austria, Christian Doppler Laboratory for Expert Systems.
- Kotsiantis, S., B. (2007) *Supervised Machine Learning: A Review of Classification Techniques*, University of Peloponnese, Greece [on line]. Ανακτήθηκε 29 Ιουνίου, 2021, από https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf.
- Lorenzo, R. (2017). *Introductory Machine Learning Notes*, Massachusetts Institute of Technology and Istituto Italiano di Tecnologia [on line]. Ανακτήθηκε 25 Αυγούστου, 2021, από http://lcsl.mit.edu/courses/ml/1718/MLNotes.pdf.
- Mohri, M. & Rostamizadeh, A. & Talwalkar, A. (2018). *Foundations of Machine Learning – Second edition*, Massachusetts: MIT Press [on line]. Ανακτήθηκε 14 Μαΐου, 2021, από https://cs.nyu.edu/~mohri/mlbook/.

- Osborne, J., W. (2002). *Notes on the use of Data Transformation*. Miami University [on line]. Ανακτήθηκε 29 Ιουνίου, 2021, από https://www.researchgate.net/publication/200152356_Notes_on_the_Use_of_Data_Transformations
- Savvopoulos, A. & Kalogeras, G. & Anagnostopoulos, C. & Alexakos, C. & Siountas, S. & Kalogeras, A.P., n.d, *Cluster-based Energy Load Profiling on Residential Smart Buildings*. Φυλλάδιο έρευνας (Αλεξάκος, προσωπική επικοινωνία, 16 Ιουνίου, 2021).
- Suthaharan, S. (2016). *Machine Learning Models and Algorithms for Big Data Classification*. New York: Springer [on line]. Ανακτήθηκε 27 Ιουνίου, 2021, από https://link.springer.com/content/pdf/10.1007%2F978-1-4899-7641-3.pdf.
- Sutton, R. (1999). *Reinforcement Learning. Journal of Cognitive Neuroscience* [on line]. Ανακτήθηκε 11 Αυγούστου, 2021, από https://www.researchgate.net/publication/270960086_Reinforcement_learning.
- Wu, X. & Kumar, V. & Quinlan, J. & Ghosh, J. & Yang, Q. & Motoda, H. & McLachlan, G. & Angus, N. & Liu, B. & Yu, P. & Zhou, Z.H. & Steinbach, M. & Hand, D. & Steinberg, D. (2007). *Top 10 algorithms in data mining – Survey Paper* [on line]. Ανακτήθηκε 7 Ιουλίου, 2021, από https://link.springer.com/article/10.1007/s10115-007-0114-2.
- Xiaojin, Z. (2005). *Semi-Supervised Learning Literature Survey* [on line]. Ανακτήθηκε 30 Ιουνίου, 2021, από https://minds.wisconsin.edu/handle/1793/60444.

ΔΙΑΔΙΚΤΥΑΚΕΣ ΑΝΑΦΟΡΕΣ

- *Δεδομένα* (2021). Ανακτήθηκε 23 Ιουνίου, 2021, από το Wikipedia, Wiki: https://el.wikipedia.org/wiki/%CE%94%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CE%B1
- *Εξόρυξη δεδομένων* (2020). Ανακτήθηκε 6 Ιουλίου, 2021, από το Wikipedia, Wiki: https://el.wikipedia.org/wiki/%CE%95%CE%BE%CF%8C%CF%81%CF%85%CE%BE%CE%B7_%CE%B4%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CF%89%CE%BD
- *Θεώρημα Μπάγιες* (2021). Ανακτήθηκε 21 Ιουλίου, 2021, από το Wikipedia, Wiki: https://el.wikipedia.org/wiki/%CE%98%CE%B5%CF%8E%CF%81%CE%B7%CE%BC%CE%B1_%CE%9C%CF%80%CE%AD%CF%85%CE%B6

- Μητρόπουλος, Ι. (2009). *Εισαγωγή στην στατιστική επιχειρήσεων - Κεφάλαιο 2 - Περιγραφικές τεχνικές*. Ανακτήθηκε 23 Ιουνίου, 2021, από Chapter 2 - http://www.ba.teiwest.gr/Nea%20Mathimata/Eisagogi%20stin%20Statistiki%20Epixeiriseon/%CE%9A%CE%B5%CF%86%CE%AC%CE%BB%CE%B1%CE%B9%CE%BF%2002.pdf.
- Μητρόπουλος, Ι. (2021). *Ποσοτικές μέθοδοι στην οικονομία και στην διοίκηση 2 – Απλή γραμμική παλινδρόμηση*. Ανακτήθηκε 22 Αυγούστου, 2021, από https://eclass.upatras.gr/courses/MST142/.
- *Μηχανική Μάθηση* (2016). Ανακτήθηκε 25 Ιουνίου, 2021 από το Wikiversity, Wiki: https://el.wikiversity.org/wiki/%CE%9C%CE%B7%CF%87%CE%B1%CE%BD%CE%B9%CE%BA%CE%AE_%CE%9C%CE%AC%CE%B8%CE%B7%CF%83%CE%B7
- Ψούνης, Δ. (2015). *ΠΛΗ31 ΜΑΘΗΜΑ 3.1 & 3.2 & 4.1.* Ανακτήθηκε 25 Μαΐου, 2021, από http://www.psounis.gr/plh31.html
- Boutsikas, M.V. (2004). «*Στατιστικά Προγράμματα*». Ανακτήθηκε 9 Ιουλίου, 2021, από http://www.unipi.gr/faculty/mbouts/statprog/SPSS_lesson11.pdf
- Brownlee, J. (2020). *Training-Test split for evaluating Machine Learning algorithms*. Ανακτήθηκε 20 Ιουλίου, 2021, από https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/
- Burr, S. (2009). *Active Learning Literature Survey*. Ανακτήθηκε 30 Ιουνίου, 2021, από https://minds.wisconsin.edu/handle/1793/60660
- Cortez, P. & Silva, A., UCI Machine Learning Repository, n.d. *Student Performance Data Set*. Ανακτήθηκε 11 Ιουλίου, 2021, από https://archive.ics.uci.edu/ml/datasets/Student+Performance.
- Hardesty, L. (2017). *Explained: Neural Networks*. Ανακτήθηκε 7 Ιουλίου, 2021, από https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414
- *Herman Minkowski* (2021) Ανακτήθηκε 19 Ιουλίου, 2021, από το Wikipedia, Wiki: https://en.wikipedia.org/wiki/Hermann_Minkowski.
- Kaggle, n.d., n.a. *State of Data Science and Machine Learning 2020*. Ανακτήθηκε 25 Ιουλίου, 2021, από https://www.kaggle.com/kaggle-survey-2020.
- Parjapat, A., Kaggle, n.d. *Loan Predication*. Ανακτήθηκε 6 Αυγούστου, 2021, από https://www.kaggle.com/ninzaami/loan-predication.

- Navlani, A. (2019). *Support Vector Machines with Scikit-Learn*. Ανακτήθηκε 24 Αυγούστου, 2021, από https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python.
- Scikit-Learn, n.d., n.a. *API Reference*. Ανακτήθηκε 25 Αυγούστου, 2021, από https://scikit-learn.org/stable/modules/classes.html.
- Scikit-Lean, n.d., n.a*., SVM: Maximum margin separating hyperplane*. Ανακτήθηκε 27 Αυγούστου, 2021, από: https://scikit-learn.org/stable/auto_examples/svm/plot_separating_hyperplane.html.
- Shahadat, U. & Arif, K. & Ekramul, H. & Mohamad, A.M. (2019). *Comparing different supervised machine learning algorithms for disease prediction*. Ανακτήθηκε 19 Ιουλίου, 2021, από https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1004-8
- Sharma, P. (2020) *Four types of distance metrics in machine learning*. Ανακτήθηκε 26 Αυγούστου, 2021, από https://www.analyticsvidhya.com/blog/2020/02/4-types-of-distance-metrics-in-machine-learning/.
- Sharma, S. (2017) *Activation Functions in Neural Networks.* Ανακτήθηκε 25 Αυγούστου, 2021, από https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6.
- Soren, D. (2017). *Categorising machine learning problems*. Ανακτήθηκε 7 Ιουλίου, 2021, από https://www.practicalai.io/categorizing-machine-learning-problems/
- UCI Machine Learning Repository, Kaggle, n.d. *Pima Indians Diabetes Data Set*. Ανακτήθηκε 14 Ιουλίου, 2021, από https://www.kaggle.com/uciml/pima-indians-diabetes-database?select=diabetes.csv.

ΠΑΡΑΡΤΗΜΑΤΑ

Κώδικας Python

Κώδικας για την αναπαράσταση της «Εικόνας 2»

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.svm import SVC
from sklearn.datasets import make_blobs


# Create 40 separable points and fit SVM to them
X, y = make_blobs(n_samples=40, centers=2, random_state=6)
clf = svm.SVC(kernel='linear', C=1000)
clf.fit(X, y)

# Start creating the plot
plt.figure(figsize=(6.5,4))
plt.scatter(X[:, 0], X[:, 1], c=y, s=80, cmap=plt.cm.flag, marker = 'o')

# Plot the decision function
ax = plt.gca()
xlim = ax.get_xlim()
ylim = ax.get_ylim()

# Create grid to evaluate model
xx = np.linspace(xlim[0], xlim[1], 30)
yy = np.linspace(ylim[0], ylim[1], 30)
YY, XX = np.meshgrid(yy, xx)
xy = np.vstack([XX.ravel(), YY.ravel()]).T
Z = clf.decision_function(xy).reshape(XX.shape)

# Plot decision boundary and margins
ax.contour(XX, YY, Z, colors='black', levels=[-1, 0, 1], alpha=0.8,
linestyles=['--', '-', '--'])

# Plot support vectors
ax.scatter(clf.support_vectors_[:, 0], clf.support_vectors_[:, 1], s=120,
linewidth=5, facecolors='none', edgecolors='face')

# Final plot
plt.title("Support Vector Machines Example")
plt.xticks(())
plt.yticks(())
plt.show()
```

Εισαγωγή και φόρτωση των κατάλληλων βιβλιοθηκών για τις μελέτες περίπτωσης

```
# Importing necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns
import sklearn
import math
import time
from sklearn import model_selection
from sklearn import metrics
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.neural_network import MLPClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.linear_model import LogisticRegression, LinearRegression
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from matplotlib import pyplot as plt
from random import randrange
```

Φόρτωση και προ-ετοιμασία του συνόλου Students Performance Data Set

```
# Import the Data
data = pd.read_csv("../BscThesisDataSets/Students/GradesDS/student-
mat.csv", sep =";")


# Explore the dataset / Check for null values / Check for missing values
data.describe()
data.isnull().sum()
data.isna().sum()


# Create the "Average Grade" column
data["Average Grade"] = (data ["G1"] + data["G2"] + data["G3"]) / 3


# Create the class for the dataset
PerformanceList = []
for value in data["Average Grade"]:
    if value >= 17.5:
        PerformanceList.append("Excellent")
    elif value >= 9.5:
        PerformanceList.append("Good")
    else:
        PerformanceList.append("Average")
data.insert(loc = 0, column = "Performance", value = PerformanceList)


# Heatmap Plot
plt.figure(figsize=(8, 8))
sns.heatmap(data.corr(), annot = True)
plt.show()
```

```
# Correlation Plots
sns.pairplot(data)

# Plot the classes
x_names = ["Excellent", "Good", "Average"]
y_names = data["Performance"].value_counts()
x_plot = np.arange(len(x_names))
plt.figure(figsize=(4, 4))
plt.xticks(x_plot, x_names)
plt.title("Student Performance")
plt.ylabel("Value Counts")
plt.bar(x_plot, y_names, width = 0.5, color = "navy")


# Replace the categorical values with numerical one's
data["paid"].replace(['yes', 'no'],[1,0], inplace = True)
data["Performance"].replace(['Excellent', 'Good', 'Average'],[2, 1, 0],
inplace = True)


# Create the Data Set attribute's which are going to be used for processing
# NOTE: Do not use Performance attribute for Linear Regression
data = data[["G1", "G2", "G3", "studytime", "paid", "Medu", "Fedu",
"Performance"]]
cls = "Performance"
X = data.drop([cls], 1)
y = data[cls]
```

Φόρτωση και προ-ετοιμασία του συνόλου Diabetes Data Set

```
# Import the Data
data = pd.read_csv('../BscThesisDataSets/Health/Diabetes.csv')
data = data.rename(columns = {'DiabetesPedigreeFunction': 'DPF', 'Outcome':
'Diabetes'}, inplace = False)


# Explore the dataset / Check for null values / Check for missing values
data.describe()
data.isnull().sum()
data.isna().sum()


# Heatmap Plot
plt.figure(figsize=(6, 6))
sns.heatmap(data.corr(), annot = True)
plt.show()


# Correlation Plots
sns.pairplot(data)
```

```
# Plot the classes
numbers = data["Diabetes"].value_counts()
x_names = ["Positive", "Negative"]
x_plot = np.arange(len(x_names))
y_names = data["Diabetes"]
plt.figure(figsize = (4, 4))
plt.xticks(x_plot, x_names)
plt.title("Diabetes Outcome")
plt.ylabel("Value Counts")
plt.bar(x_plot, numbers, width = 0.5, color = "navy")


# Create the Data Set attribute's which are going to be used for processing
cls = "Diabetes"
X = data.drop([cls], 1)
y = data[cls]
```

Φόρτωση και προ-ετοιμασία του συνόλου Loan Predication Data Set

```
# Import the Data
data = pd.read_csv('../BscThesisDataSets/Buisness-Economis/Loan.csv')
data = data.drop(["Loan_ID"], 1)


# Explore the dataset / Check for null values / Check for missing values
data.describe()
data.isnull().sum()
data.isna().sum()


# Replace the categorical values with numerical one's
data["Gender"].replace(['Male', 'Female'],[1,0], inplace = True)
data["Self_Employed"].replace(['Yes', 'No'],[1,0], inplace = True)
data["Education"].replace(['Graduate', 'Not Graduate'],[1,0], inplace =
True)
data["Married"].replace(['Yes', 'No'],[1,0], inplace = True)
data["Property_Area"].replace(['Rural', 'Urban', 'Semiurban'],[0, 1, 2],
inplace = True)
data["Loan_Status"].replace(['Y', 'N'],[1, 0], inplace = True)
data["Dependents"] = data["Dependents"].replace( "3+", 3)


# Fill the missing values of each column with: a) values that are viewed
more or b) the mean of each attribute
data["Gender"].value_counts()
data["Dependents"].value_counts()
data["Self_Employed"].value_counts()
data["Credit_History"].value_counts()

data["Gender"] = data["Gender"].fillna(value = 1)
data["Dependents"] = data["Dependents"].fillna(value = 0)
data["Self_Employed"] = data["Self_Employed"].fillna(value = 0)
data["Credit_History"] = data["Credit_History"].fillna(value = 1)
```

```
LoanAmmount_Mean = int(data["LoanAmount"].mean())
LoanAmountTerm_Mean = int(data["Loan_Amount_Term"].mean())

data["LoanAmount"] = data["LoanAmount"].fillna(value = LoanAmmount_Mean)
data["Loan_Amount_Term"] = data["LoanAmount"].fillna(value =
LoanAmountTerm_Mean)


# Check for missing values again
data.isna().sum()


# Fill zero values of CoApplicant Income attribute
def CoapplicantIncome_Mean():
    sumVals = 0
    countVals = 0
    for value in data["CoapplicantIncome"]:
        if (value == 0) or (math.isnan(value) == True):
            pass
        else:
            sumVals = sumVals + value
            countVals+=1
    CoapplicantIncome_Mean = sumVals / countVals
    return CoapplicantIncome_Mean

Mean = CoapplicantIncome_Mean()
data["CoapplicantIncome"] = data["CoapplicantIncome"].replace(0, Mean)


# HeatMap Plot
plt.figure(figsize=(8,8))
sns.heatmap(data.corr(), annot = True)
plt.show()


# Correlation Plots
sns.pairplot(data)


# Plot the classes
x_names = ["Approved", "Not Approved"]
y_names = data["Loan_Status"].value_counts()
x_plot = np.arange(len(x_names))
plt.figure(figsize=(4,4))
plt.xticks(x_plot, x_names)
plt.title("Loan Status")
plt.ylabel("Value Counts")
plt.bar(x_plot, y_names, width = 0.5, color = "navy")


# Create the Data Set attribute's which are going to be used for processing
cls = "Loan_Status"
X = data.drop([cls], 1)
y = data["Loan_Status"]
```

Κώδικας για την εφαρμογή των αλγορίθμων κατηγοριοποίησης

```
# Create the pre-processed Data
scaler = StandardScaler()
mmScaler = MinMaxScaler()
x_Scaled = pd.DataFrame(scaler.fit_transform(X), columns = X.columns)
x_mmScaled = pd.DataFrame(mmScaler.fit_transform(X), columns = X.columns)
x_doubleScaled = pd.DataFrame(mmScaler.fit_transform(x_Scaled), columns =
X.columns)


# Create the algorithms and their parameters
# NOTE: Logistic Regression is not used in the first case study
Algorithms = ['kNN', 'SVM', 'Tree', 'Bayes', 'MLP', 'Logistic']
randomkNN = randrange(5, 10)
randomMLP = randrange(50,100)
kNN = KNeighborsClassifier(n_neighbors = randomkNN, metric = 'minkowski')
Tree = DecisionTreeClassifier(criterion = 'entropy')
MLP = MLPClassifier(activation='relu', hidden_layer_sizes = randomMLP,
max_iter = 3000)
SVC = SVC()
Bayes = GaussianNB()
LogReg = LogisticRegression(max_iter = 3000)
classifiers = [('kNN', kNN),
    ('SVC', SVC),
    ('Tree', Tree),
    ('Bayes', Bayes),
    ('MLP', MLP),
    ('LogReg', LogReg)]


# Plotting results function
# NOTE: Logistic Regression is not used in the first case study
def barPlot(accuracy, precision, split):

    Algorithms = ['kNN', 'SVM', 'Tree', 'Bayes', 'MLP', 'Logistic']
    x_pos = np.arange(len(Algorithms))
    yVals = [0.2, 0.4, 0.6, 0.8, 1.0]

    plt.figure(figsize=(4,4))
    plt.xticks(x_pos, Algorithms)
    plt.yticks(yVals)
    plt.title("Classification Results ({}% training size)".format(split),
    loc="center")
    plt.ylabel("Accuracy - Precision")
    plt.bar(x_pos-0.15, accuracy, 0.3, align = "center", color = "navy",
    label = "Accuracy")
    plt.bar(x_pos+0.15, precision, 0.3, align = "center", color = "orange",
    label = "Precision")
    plt.legend(loc = 8, labelspacing = 1.0)
```

```
# Printing results function
def PrintResults(accuracy, precision, split):
    Results = zip(accuracy, precision)
    output_data = pd.DataFrame(Results, columns = ["Accuracy",
    "Precision"], index = Algorithms)
    print("{}% Results \n \n".format(split), output_data, " \n \n")


# Main Classification processing function
def MachineLearningMain(X, y):

    # For training split 80%
    AccuracyList = []
    PrecisionList = []
    x_train, x_test, y_train, y_test =
    sklearn.model_selection.train_test_split(X.values, y.values, test_size
    =  0.2, random_state = 0)

   for clf, model in classifiers:
        model.fit(x_train, y_train)
        y_true = y_test
        y_pred = model.predict(x_test)
        acc = metrics.accuracy_score(y_true, y_pred)
        AccuracyList.append(acc)
        precision = metrics.precision_score(y_true, y_pred, average =
        "weighted", zero_division = 0)
        PrecisionList.append(precision)

    barPlot(accuracy = AccuracyList, precision = PrecisionList, split =
    '80')
    PrintResults(accuracy = AccuracyList, precision = PrecisionList, split
    = '80')
    time.sleep(2)

    # For training split 67%
    AccuracyList = []
    PrecisionList = []
    x_train, x_test, y_train, y_test =
    sklearn.model_selection.train_test_split(X.values, y.values, test_size
    = 0.33, random_state = 0)

    for clf, model in classifiers:
        model.fit(x_train, y_train)
        y_true = y_test
        y_pred = model.predict(x_test)
        acc = metrics.accuracy_score(y_true, y_pred)
        AccuracyList.append(acc)
        precision = metrics.precision_score(y_true, y_pred, average =
        "weighted", zero_division = 0)
        PrecisionList.append(precision)

    barPlot(accuracy = AccuracyList, precision = PrecisionList, split =
    '67')
    PrintResults(accuracy = AccuracyList, precision = PrecisionList, split
    = '67')
    time.sleep(2)
```

Κώδικας για την εμφάνιση των αποτελεσμάτων αλγορίθμων κατηγοριοποίησης

```
# Print the results depending on the pre-processing technique
# NOTE: Do the following for each case study


print("Machine Learning Algorithm Results \n \n \n" + "Results without pre-
processing: \n")
MachineLearningMain(X = X, y = y)

print("Results with Standard Scaler: \n")
MachineLearningMain(X = x_Scaled, y = y)

print("Results with MinMax Scaler: \n")
MachineLearningMain(X = x_mmScaled, y = y)

print("Results with MinMax and Standard Scaler: \n")
MachineLearningMain(X = x_doubleScaled, y = y)
```

Κώδικας για την εφαρμογή της γραμμικής παλινδρόμησης στο Student Performance Data Set

```
# Function to create scatter plots of actual and predicted G3 values
def plots(y, predictions, split):
    xVals = [0 , 5 , 10 , 15, 20]
    yVals = [0 , 5 , 10 , 15, 20]
    plt.figure(figsize = (4,4))
    plt.xticks(xVals)
    plt.yticks(yVals)
    plt.scatter(y, predictions, color = 'navy', marker = "o")
    plt.xlabel('Actual Values')
    plt.ylabel('Predicted Values')
    plt.title('Actual vs Predicted ({}% training size) '.format(split))
    plt.show()

# Import Linear Regression model
linear = linear_model.LinearRegression()

# Create the Data Set attribute's which are going to be used for processing
data = pd.read_csv("../BscThesisDataSets/Students/GradesDS/student-
mat.csv", sep =";")
data = data[["G1", "G2", "G3", "studytime", "paid", "Medu", "Fedu"]]
predict = "G3"
X = data.drop([predict], 1)
y = data[predict]

# Main Linear Regression processing function
def LinearMain(X, y):

    #For training split 80%
    x_train, x_test, y_train, y_test =
    sklearn.model_selection.train_test_split(X.values, y.values,
    test_size = 0.2,random_state = 0)
```

```
    linear. Fit(x_train, y_train)
    accA = linear.score(x_test, y_test)
    coefA = (linear.coef_)
    interceptA = (linear.intercept_)
    predictionsA = (linear.predict(x_test))
    plots(y = y_test, predictions = predictionsA, split = "80")
    print("1) 20% Test Split.\n"+"Linear Regression Accuracy: ",
    f'{accA}', '%')
    print("Coefficients: {}".format(coefA) , "\n" + "Intercept: ",
    str(interceptA), "\n" )
    time.sleep(2)

    #For training split 67%
    x_train, x_test, y_train, y_test =
    sklearn.model_selection.train_test_split(X.values, y.values,
    test_size=0.33,random_state = 0)
    linear.fit(x_train, y_train)
    accB = linear.score(x_test, y_test)
    coefB = (linear.coef_)
    interceptB = (linear.intercept_)
    predictionsB = (linear.predict(x_test))
    plots(y = y_test, predictions = predictionsB, split = "67")
    print("2) 33% Test Split.\n"+"Linear Regression Accuracy: ",
    f'{accB}', '%')
    print("Coefficients: {}".format(coefB)  , "\n" + "Intercept: ",
    str(interceptB))
    time.sleep(2)
```

Κώδικας για την εμφάνιση των αποτελεσμάτων της γραμμικής παλινδρόμησης στο Student Performance Data Set

```
# Print the results depending on the pre-processing technique
print("Linear Regression Algorithm Results \n \n \n" + "Results without
pre-processing: \n")
LinearMain(X = X, y = y)

print("Results with Standard Scaler: \n")
LinearMain(X = x_Scaled, y = y)

print("Results with MinMax Scaler: \n")
LinearMain(X = x_mmScaled, y = y)

print("Results with MinMax and Standard Scaler: \n")
LinearMain(X = x_doubleScaled, y = y)
```

Πνευματικά δικαιώματα

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν. 1599/1988 και τα άρθρα 2,4,6 παρ. 3 του Ν. 1256/1982, η παρούσα εργασία αποτελεί αποκλειστικά προϊόν προσωπικής εργασίας και δεν προσβάλλει κάθε μορφής πνευματικά δικαιώματα τρίτων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον.

Κωνσταντίνος Κούτρας, 2021