ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ



Ανάλυση δεδομένων από online καταστήματα και ανάπτυξη μοντέλων πρόβλεψης

- Πτυχιακή Εργασία -

Καραντζή Παναγιώτα (Α.Μ. Π2015130)

Επιβλέπων: Έξαρχος Θεμιστοκλής

Φεβρουάριος 2020, Κέρκυρα

Πίνακας περιεχομένων

Περίληψη....5
Κεφάλαιο Α: Εισαγωγή....6
Α.1 Εικόνα του ηλεκτρονικού εμπορίου σήμερα....6
Α.2 Πρόβλεψη συμπεριφοράς πελατών....7
Α.2.1 Δεδομένα Συμπεριφοράς....7
Α.2.2 Ανάπτυξη μοντέλου πρόβλεψης....7
Α.3 Οργάνωση εργασίας....9

Κεφάλαιο Β: Βιβλιογραφική Ανασκόπηση....10
Β.1 Κατηγοριοποίηση....10
Β.1.1 Διασταυρούμενη Επικύρωση 10 τμημάτων....11
Β.2 Αλγόριθμοι Μάθησης για Δυαδική Κατηγοριοποίηση....11
Β.2.1 Ο αλγόριθμος C4.5....11
Β.2.2 Ο αλγόριθμος Random Forest....14
Β.2.3 Bayesian Κατηγοριοποίηση και Naive Bayes....16
Β.2.4 Κ Πλησιέστεροι Γείτονες....18
Β.2.5 Νευρωνικά Δίκτυα....19
Β.3 Ανάλυση σχετικής βιβλιογραφίας....21
Β.3.1 Συμπεράσματα ανάλυσης βιβλιογραφίας....28

Κεφάλαιο Γ: Μεθοδολογία....31
Γ.1 Μετρικές Αξιολόγησης....31
Γ.2 Επιλογή Εργαλείου....33

Κεφάλαιο Δ: Δεδομένα....34
Δ.1 Περιγραφή άρθρου άντλησης δεδομένων....34
Δ.2 Περιγραφή Δεδομένων....34
Δ.3 Επεξεργασία Δεδομένων....36
Δ.4 Επιλογή Χαρακτηριστικών....37
Δ.4.1 Ο Best First ως μέθοδος αναζήτησης....38
Δ.4.2 Μέθοδοι επιλογής χαρακτηριστικών στο Weka....39

Κεφάλαιο Ε: Αποτελέσματα Απόδοσης αλγορίθμων....43
Ε.1 Ανάλυση αποτελεσμάτων για imbalanced dataset....43
Ε.1.1 Σύγκριση αλγορίθμων για imbalanced data....46
Ε.2 Ανάλυση αποτελεσμάτων για balanced dataset....47
Ε.2.1 Σύγκριση αλγορίθμων για balanced data....50
Ε.3 Σύγκριση αποτελεσμάτων μεταξύ balanced και imbalanced datasets....50
Ε.4 Αποτελέσματα αλγορίθμων με επιλογή χαρακτηριστικών....51
Ε.4.1 Επιλογή υποσυνόλου χαρακτηριστικών βάσει συσχέτισης....51
Ε.4.2 Επιλογή υποσυνόλου χαρακτηριστικών με Wrapper....52
Ε.4.2 Επιλογή υποσυνόλου χαρακτηριστικών με Ranker – Κέρδους Πληροφορίας....53
Ε.4.2 Επιλογή υποσυνόλου χαρακτηριστικών με Ranker – Συσχέτισης....53
Ε.5 Σύγκριση αποτελεσμάτων με βιβλιογραφία....54

Συμπεράσματα....57
Βιβλιογραφία – Ιστογραφία....59

Ευρετήριο πινάκων

Πίνακας Β1: Σύγκριση άρθρων που μελετήθηκαν.....30
Πίνακας Δ1: Αριθμητικά Χαρακτηριστικά.....35
Πίνακας Δ2: Ονομαστικά Χαρακτηριστικά.....36
Πίνακας Δ3: Χαρακτηριστικά που επιλέγονται με wrapper.....40
Πίνακας Ε1: Αποτελέσματα για J48 σε imbalanced data.....43
Πίνακας Ε2: Confusion Matrix RF για imbalanced.....45
Πίνακας Ε3: Confusion Matrix NB για imbalanced.....45
Πίνακας Ε4: Confusion Matrix KNN για imbalanced.....45
Πίνακας Ε5: Confusion Matrix MLP για imbalanced.....46
Πίνακας Ε6: Αποτελέσματα για imbalanced data.....46
Πίνακας Ε7: Confusion Matrix J48 για balanced data.....48
Πίνακας Ε8: Confusion Matrix RF για balanced.....48
Πίνακας Ε9: Confusion Matrix NB για balanced.....48
Πίνακας Ε10: Confusion Matrix KNN για balanced.....49
Πίνακας Ε11: Confusion Matrix MLP για balanced.....49
Πίνακας Ε12: Αποτελέσματα για imbalanced data.....50
Πίνακας Ε13: Αποτελέσματα με μέθοδο CfsSubsetEval.....51
Πίνακας Ε14: Αποτελέσματα με μέθοδο Wrapper.....52
Πίνακας Ε15: Αποτελέσματα με μέθοδο InfoGain.....53
Πίνακας Ε16: Αποτελέσματα με μέθοδο Ranker Correlation.....54
Πίνακας Ε17: Καλύτερα αποτελέσματα.....55
Πίνακας Ε18: Σύγκριση αποτελεσμάτων.....55

Περίληψη

Αυτή η εργασία αφορά την μελέτη και την ανάπτυξη μοντέλων πρόβλεψης της συμπεριφοράς των πελατών σε ηλεκτρονικά καταστήματα. Δηλαδή, αν ο χρήστης του ηλεκτρονικού καταστήματος έχει την πρόθεση να προβεί σε αγορά ή όχι. Αρχικά, μελετάται η σχετική βιβλιογραφία πάνω στο θέμα και παρουσιάζονται οι αλγόριθμοι που χρησιμοποιούνται στην πορεία της εργασίας. Στη συνέχεια, γίνεται η μελέτη των δεδομένων που χρησιμοποιήθηκαν, η δημιουργία ενός καινούριου ισορροπημένου συνόλου δεδομένων και η επιλογή χαρακτηριστικών μέσω διαφορετικών μεθόδων. Συγκεκριμένα, γίνεται χρήση της τεχνικής του undersampling και των μεθόδων επιλογής χαρακτηριστικών τόσο Wrapper όσο και φίλτρων. Οι βασικές μετρικές αξιολόγησης που χρησιμοποιούνται είναι η ακρίβεια και το F1 Score. Επίσης, το εργαλείο που επιλέχτηκε στην εργασία ήταν το Weka.

Από τα διάφορα μοντέλα που δοκιμάζονται στο τέλος επιλέγονται αυτά με τα πιο ικανοποιητικά ποσοστά μετρικών και συγκρίνονται με αυτά της βιβλιογραφίας. Οι αλγόριθμοι που χρησιμοποιήθηκαν ήταν ο C4.5 ή J48, Random Forest, Naive Bayes, k – NN και Multilayer Perceptron. Τα αποτελέσματα που προκύπτουν είναι αρκετά ικανοποιητικά, με την πλειοψηφία των αλγορίθμων να ξεπερνούν το 80% σε ακρίβεια. Τα καλύτερα αποτελέσματα αποδεικνύεται πως επιφέρουν τα δέντρα απόφασης,

Κεφάλαιο Α

Εισαγωγή

Το περιεχόμενο αυτής της εργασίας, το οποίο είναι η ανάλυση, η κατηγοριοποίηση , και η πρόβλεψη της συμπεριφοράς του πελάτη στον τομέα του ηλεκτρονικού εμπορίου, παρουσιάζονται συνοπτικά σε αυτό το κεφάλαιο.

Α.1 Εικόνα του ηλεκτρονικού εμπορίου σήμερα

Το ηλεκτρονικό εμπόριο μπορούμε να πούμε πως έχει αλλάξει την συνολική εικόνα της αγοράς τα τελευταία χρόνια, τόσο σε εγχώριο επίπεδο, όσο και σε παγκόσμιο. Πολλοί από εμάς στρεφόμαστε καθημερινά στα ηλεκτρονικά καταστήματα για τις αγορές μας. Μερικοί από τους κύριους λόγους που συμβαίνει αυτό είναι η ευκολία που μας προσφέρει, αφού κάνουμε τις αγορές μας από την άνεση του σπιτιού μας, καθώς και η δυνατότητα σύγκρισης τιμών μεταξύ των διάφορων προϊόντων και αναζήτηση των πιο οικονομικών. Επίσης, οι επιχειρήσεις επωφελούνται από την πώληση των προϊόντων τους online. Τόσο οι μεγάλες όσο και οι μικρές επιχειρήσεις μπορούν να χρησιμοποιήσουν το ηλεκτρονικό εμπόριο ώστε να αυξήσουν την έκθεσή τους, αγγίζοντας ακόμα και παγκόσμιο καταναλωτικό κοινό μέσω του web και των μηχανών αναζήτησης. Ένα καλό παράδειγμα της εξέλιξης του ηλεκτρονικού εμπορίου είναι το mobile e-commerce. Δηλαδή, η αγορά προϊόντων στο διαδίκτυο μέσω κινητών συσκευών από οποιοδήποτε μέρος, την οποιαδήποτε στιγμή. Καταλαβαίνουμε, λοιπόν, πως αυτή είναι μια από τις απόρροιες του ηλεκτρονικού εμπορίου που μας δείχνει πως μπορεί αυτό να αλλάξει και να εξελιχθεί (Shiying Yu, Lu Xiong. 2011). Η πρόοδος όμως της τεχνολογίας και του ηλεκτρονικού εμπορίου δεν φαίνονται να σταματούν να εξελίσσονται, και αυτό ίσως συμβαίνει γιατί οι ανάγκες των καταναλωτών αλλάζουν συνεχώς. Γι' αυτό και είναι αναγκαίο για τις ηλεκτρονικές επιχειρήσεις να καταλάβουν σε βάθος την συμπεριφορά των καταναλωτών.

Οι οργανισμοί και οι εταιρίες πάντα προσπαθούσαν να συλλέξουν πληροφορίες από τους πελάτες τους σχετικά με τα προϊόντα ή τις υπηρεσίες που προσφέρουν, με σκοπό την καλύτερη προώθηση αυτών. Μια συνηθισμένη τεχνική προώθησης είναι η διαφήμιση. Σε διάφορα ηλεκτρονικά καταστήματα γίνεται διαφήμιση στη σελίδα των δημοφιλέστερων προϊόντων με σκοπό την αύξηση των πωλήσεων. Επίσης, ο χρήστης προτρέπεται να κάνει εγγραφή στο κατάστημα, ώστε να ενημερώνεται για διάφορες προσφορές. Με την εξέλιξη όμως των τεχνικών της εξόρυξης δεδομένων θα μπορούσαμε να επεκτείνουμε τις παραδοσιακές τεχνικές marketing και

να δημιουργήσουμε νέες πιο αποδοτικές και εξατομικευμένες ως προς τον πελάτη. Τα δεδομένα που συλλέγονται μπορούν να βοηθήσουν τις online επιχειρήσεις να καθορίσουν πως μπορούν να παρέχουν τα καλύτερα προϊόντα και υπηρεσίες που αναζητούν οι καταναλωτές. Έτσι, δημιουργήθηκε η ανάγκη κατανόησης της συμπεριφοράς των πελατών μέσα στα ηλεκτρονικά καταστήματα και η πρόβλεψη των προθέσεων τους.

Α.2 Πρόβλεψη συμπεριφοράς πελατών

Η πρόβλεψη της συμπεριφοράς του χρήστη μέσα στο ηλεκτρονικό κατάστημα, δηλαδή αν απλά ΄σερφάρει΄ ή έχει την πρόθεση αγοράς, είναι κρίσιμη τόσο για τους πωλητές, όσο και για τους πελάτες. Από την οπτική των πωλητών, γνωρίζοντας τις τρέχουσες αγορές των χρηστών μπορεί να αυξήσει τον όγκο των πωλήσεών τους και το περιθώριο κέρδους τους. Η πρόβλεψη της συμπεριφοράς ενός επισκέπτη μπορεί να πυροδοτήσει τις σωστές προτάσεις αγορών για αυτόν και να αυξήσει την πιθανότητα αγοράς. Επίσης, μπορεί να συμβάλλει σε διάφορες στρατηγικές marketing. Για παράδειγμα, ο πωλητής εφόσον γνωρίζει ποιος επισκέπτης του καταστήματος θα κάνει αγορά, μπορεί να του προσφέρει ένα εκπτωτικό κουπόνι για τις μελλοντικές αγορές του, δελεάζοντας τον έτσι να επιστρέψει στο κατάστημα. Επιπλέον, η χρήση ενός τέτοιου μοντέλου πρόβλεψης κάνει τους πωλητές περισσότερο ανταγωνιστικούς, αφού οι πληροφορίες που συλλέγονται μπορούν να ερμηνευτούν και να οδηγήσουν στις απαραίτητες αλλαγές ώστε να βελτιωθεί η εμπειρία του καταναλωτή. Οι χρήστες, λοιπόν, διευκολύνονται στην αναζήτηση των αγορών τους και πολλές φορές αποκτούν προσωποποιημένα προνόμια.

Α.2.1 Δεδομένα Συμπεριφοράς

Τα δεδομένα που συλλέγονται για την ανάλυση της συμπεριφοράς των ηλεκτρονικών πελατών χαρακτηρίζονται ως δεδομένα συμπεριφοράς (behavioral data). Τις περισσότερες φορές αποτελούν ένα συνδυασμό χαρακτηριστικών που αφορούν τα προϊόντα, όπως είναι η τιμή τους και ο κωδικός τους, και χαρακτηριστικών των ενεργειών του χρήστη στο ηλεκτρονικό κατάστημα, όπως το πόσες φορές έγινε προβολή ενός προϊόντος ή η προσθήκη και αφαίρεση αντικειμένων στο καλάθι αγορών. Ορισμένες φορές χρησιμοποιούνται δεδομένα που προέρχονται από ‘τρίτα μέρη’ και θεωρούνται στατικά. Ενώ τα περισσότερα από τα δεδομένα που συλλέγονται εκείνη την στιγμή ονομάζονται δυναμικά.

Α.2.2 Ανάπτυξη μοντέλου πρόβλεψης

Η πρόβλεψη αυτής της συμπεριφοράς θα μπορούσε να αποτυπωθεί ως ένα πρόβλημα κατηγοριοποίησης. Πιο συγκεκριμένα, αποτελεί ένα πρόβλημα δυαδικής κατηγοριοποίησης, όπου κάθε πελάτης του ηλεκτρονικού καταστήματος είτε θα κάνει κάποια αγορά είτε όχι. Έτσι, καταλήγουμε σε δύο κλάσεις παρατηρήσεων. Στην πρώτη κλάση ανήκουν οι συνεδρίες (browsing sessions) που κατέληξαν σε αγορά και μπορεί να χαρακτηριστεί ως η θετική κλάση, αφού περιέχει την επιθυμητή κατάληξη. Ως δεύτερη κλάση ορίζονται οι συνεδρίες που δεν κατέληξαν σε αγορά και μπορεί να χαρακτηριστεί ως η αρνητική κλάση. Στο μοντέλο η θετική κλάση συναντάται συχνά ως '1', 'YES', 'TRUE', ενώ η αρνητική ως '0', 'NO' ή 'FALSE'.

Οι επισκέπτες ενός ιστότοπου αφήνουν περισσότερα ίχνη από ποτέ. Μεγάλα ποσά προσωπικών πληροφοριών, καθώς και δεδομένα clickstream, καταγράφονται κατά τη διάρκεια κάθε επίσκεψης στο ηλεκτρονικό κατάστημα, συλλέγονται, συνδέονται και αποθηκεύονται για ανάλυση με τεχνικές εξόρυξης δεδομένων. Κατά πόσο είναι όμως εύκολη η ανάπτυξη ενός τέτοιου μοντέλου πρόβλεψης; Η δυσκολία στην ανάπτυξη ενός τέτοιου είδους μοντέλου βρίσκεται στην σωστή επιλογή των δεδομένων και των χαρακτηριστικών αυτών. Ακόμη μια πρόκληση στη δημιουργία του μοντέλου είναι πως τα δεδομένα συλλέγονται σε πραγματικό χρόνο και συχνά έχουμε να αντιμετωπίσουμε ανώνυμους χρήστες. Χρήστες, δηλαδή, οι οποίοι δεν έχουν κάποιο ιστορικό αγορών και είναι νέοι πελάτες. Επίσης, η προεπεξεργασία των δεδομένων παίζει σημαντικό ρόλο, αφού τις περισσότερες φορές ο αριθμός των αγοραστών είναι πολύ μικρότερος από τον αριθμό των μη αγοραστών.

Σίγουρα ένας από τους σημαντικότερους ρόλους είναι αυτός της επιλογής του σωστού αλγορίθμου που θα χρησιμοποιηθεί για τον σχηματισμό του μοντέλου. Μπορούμε να διαλέξουμε από μια ποικιλία ειδών αλγορίθμων, όπως είναι τα νευρωνικά δίκτυα, τα δέντρα απόφασης, οι αλγόριθμοι βασισμένοι στην απόσταση και οι βασισμένοι σε κανόνες. Οι δημοφιλέστεροι από αυτούς είναι ο C4.5, ο k πλησιέστερος γείτονας, ο Naive Bayes και ο Multilayer Perceptron. Κάθε ένας από αυτούς μπορεί να λειτουργήσει διαφορετικά και να δώσει καλύτερα αποτελέσματα ανάλογα με τον τρόπο που χρησιμοποιείται. Συχνά σε προβλήματα κατηγοριοποίησης παρατηρούμε πως ένας αλγόριθμος μπορεί να λειτουργήσει καλύτερα για τις περιπτώσεις που κάποιος άλλος δυσλειτουργεί. Με πιο απλά λόγια, τα 'κενά' του ενός, συμπληρώνονται από κάποιον άλλο. Ορισμένες φορές, αν αυτό είναι δυνατό, γίνεται συνδυασμός τεχνικών και αλγορίθμων για το εκάστοτε μοντέλο.

Οι έρευνες που εξετάζονται στη συνέχεια αυτής της εργασίας έχουν δείξει πως η επίδοση και η ακρίβεια του κάθε μοντέλου επηρεάζεται από διάφορους παράγοντες. Ορισμένοι από αυτούς είναι το είδος του ηλεκτρονικού καταστήματος και οι τάσεις στα προϊόντα που αγοράζονται. Επιπλέον, φαίνεται πως η συλλογή δεδομένων από συσκευές αφής, όπως ο αριθμός των taps και

των swipes, μπορούν να παίξουν σημαντικό ρόλο στην δημιουργία ενός μοντέλου εξίσου προσαρμόσιμου σε τέτοιου είδους συσκευές. Τέλος, κρίσιμη είναι και η επιλογή του αλγορίθμου κατηγοριοποίησης, αφού παρουσιάζονται διαφορές μεταξύ των αποτελεσμάτων τους.

Α.3 Οργάνωση εργασίας

Η ενασχόληση, λοιπόν, της συγκεκριμένης εργασίας είναι η δοκιμή διαφορετικών μοντέλων πρόβλεψης με στόχο την πρόβλεψη της συμπεριφοράς του πελάτη ενός ηλεκτρονικού καταστήματος. Με πιο απλά λόγια, να μπορέσουμε να υπολογίσουμε και να συγκρίνουμε με τι ακρίβεια διαφορετικά μοντέλα αλγορίθμων προβλέπουν την πρόθεση κάποιου να ΄αγοράσει΄ ή ΄να μην αγοράσει’. Επίσης, το θέμα εξετάζεται και από βιβλιογραφικής άποψης.

Αναλυτικότερα, στο Κεφάλαιο 2 θα συζητηθούν άρθρα από την βιβλιογραφία και θα παρουσιαστούν οι αλγόριθμοι που θα χρησιμοποιηθούν στο ερευνητικό κομμάτι της εργασίας. Έπειτα, στο Κεφάλαιο 3 θα παρουσιαστούν οι μετρικές αξιολόγησης που λήφθηκαν υπόψη για την αξιολόγηση του μοντέλου, καθώς και λίγα λόγια για το εργαλείο Weka που χρησιμοποιήθηκε. Συνεχίζοντας, στο Κεφάλαιο 4 θα παρουσιαστούν τα δεδομένα που χρησιμοποιήθηκαν κατά την εκπόνηση της εργασίας, ποια ήταν η προεπεξεργασία τους και ποια χαρακτηριστικά λήφθηκαν υπόψη. Ακόμα, στο Κεφάλαιο 5 ακολουθεί μια αναλυτική παρουσίαση της εφαρμογής των αλγορίθμων και των αποτελεσμάτων που προέκυψαν. Τέλος, η εργασία αυτή κλείνει με τα τελικά συμπεράσματα στο Κεφάλαιο 6.

Κεφάλαιο Β

Βιβλιογραφική Ανασκόπηση

Αυτό το κεφάλαιο εξετάζει τα αποτελέσματα και τους αλγορίθμους που χρησιμοποιούνται από μελέτες που ασχολούνται με παρόμοιο δυαδικό πρόβλημα ταξινόμησης με αυτό που πρόκειται να ασχοληθεί η παρούσα εργασία. Ώστε να γίνει ευρέως κατανοητό το πρόβλημα και οι διάφοροι αλγόριθμοι που πρόκειται να χρησιμοποιηθούν στην πορεία της εργασίας, η έννοια της ταξινόμησης και η επεξήγηση των αλγορίθμων θα γίνουν στις παραγράφους που ακολουθούν.

Β.1 Κατηγοριοποίηση

Η εξόρυξη δεδομένων έχει διάφορες εφαρμογές, και η κατηγοριοποίηση αποτελεί μια από αυτές. Η κατηγοριοποίηση είναι εργασία επιβλεπόμενης μάθησης. Σε αυτό το είδος μάθησης προσπαθούμε να βρούμε ποια είναι η σχέση μεταξύ ενός γνωρίσματος ΄στόχου΄, δηλαδή μιας εξαρτημένης μεταβλητής, και ενός συνόλου άλλων γνωρισμάτων, δηλαδή ενός συνόλου μη εξαρτημένων μεταβλητών. Στόχος είναι η πρόβλεψη της εξαρτημένης μεταβλητής με τη χρήση των ανεξάρτητων μέσω της δημιουργίας ενός μηχανισμού, το οποίο ονομάζεται μοντέλο. Να σημειωθεί πως στην κατηγοριοποίηση η εξαρτημένη μεταβλητή έχει διακριτές ονομαστικές τιμές και κάθε μια από αυτές ορίζει την κλάση (class) που ανήκει κάθε αντικείμενο. Στην περίπτωση που οι τιμές είναι αριθμητικές τότε ονομάζεται παλινδρόμηση. Ως δυαδική κατηγοριοποίηση εννοούμε αυτή στην οποία ανήκουν δύο κλάσεις. Στην προκειμένη περίπτωση, ως κλάσεις για το πρόβλημα που διαχειρίζεται η παρούσα εργασία θα μπορούσαμε να ορίσουμε τις εξής: ‘αγοράζει’ και ‘δεν αγοράζει’. Υπάρχουν διάφοροι αλγόριθμοι ή μέθοδοι κατηγοριοποίησης, με τους πιο δημοφιλείς από αυτούς να είναι τα Δέντρα Απόφασης, τα Μπαϋεσιανά Δίκτυα και τα Νευρωνικά Δίκτυα.

Τα τρία στάδια της κατηγοριοποίησης είναι τα εξής:

1. *Επιβλεπόμενη Μάθηση:* Γίνεται ανάλυση των δεδομένων και εντοπίζονται οι σχέσεις μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Επιλέγεται το σύνολο δεδομένων εκπαίδευσης (training set), το οποίο χρησιμοποιείται για την εκπαίδευση του μοντέλου αργότερα.
2. *Επικύρωση (validation) του μοντέλου:* Σε αυτό το στάδιο γίνεται η δοκιμή του μοντέλου και υπολογίζεται η ακρίβεια του μοντέλου. Στη διαδικασία αυτή γίνεται με τη χρήση ενός συνόλου δεδομένων που ονομάζεται σύνολο επικύρωσης (validation set) και περιέχει

αντικείμενα με άγνωστες κλάσεις. Στόχος είναι η σωστή πρόβλεψη της κλάσης από το μοντέλο.

3. *Χρήση του μοντέλου:* Αφού το μοντέλο εκπαιδευτεί και επικυρωθεί, χρησιμοποιείται για την διεξαγωγή προβλέψεων.

Β.1.1 Διασταυρούμενη Επικύρωση 10 τμημάτων

Όπως αναφέρθηκε στην προηγούμενη παράγραφο το δεύτερο στάδιο της κατηγοριοποίησης είναι η επικύρωση. Υπάρχουν διάφορα είδη μεθόδων επικύρωσης, με τα πιο βασικά από αυτά να είναι η μέθοδος holdout, η leave one out, η μέθοδος bootstrap και η διασταυρούμενη επικύρωση k τμημάτων (k fold cross validation). Θα ασχοληθούμε όμως με την 10 fold cross validation, αφού θα είναι και αυτή που θα χρησιμοποιηθεί στην συνέχεια της εργασίας.

Με αυτή τη μέθοδο χωρίζουμε το σύνολο των δεδομένων σε 10 τυχαία υποσύνολα. Ένα από αυτά τα υποσύνολα χρησιμοποιείται ως validation set και τα υπόλοιπα υποσύνολα χρησιμοποιούνται ως training set. Αυτή η διαδικασία επαναλαμβάνεται 10 φορές, χρησιμοποιώντας κάθε φορά διαφορετικό validation set και τα υπόλοιπα εννιά ως training set. Στο τέλος υπολογίζεται η μέση επίδοση του μοντέλου.

Β.2 Αλγόριθμοι Μάθησης για Δυαδική Κατηγοριοποίηση

Στις παρακάτω παραγράφους παρουσιάζονται αναλυτικά οι αλγόριθμοι που θα χρησιμοποιηθούν στην πορεία της συγκεκριμένης εργασίας. Αναλυτικότερα, αυτοί είναι ο C4.5, ο Random Forest, ο Naive Bayes και οι K- πλησιέστεροι γείτονες. Επίσης, θα συζητηθούν η bayesian κατηγοριοποίηση και τα νευρωνικά δίκτυα.

Β.2.1 Ο αλγόριθμος C4.5

Ο C4.5 είναι ένας αλγόριθμος που χρησιμοποιείται για την παραγωγή ενός δέντρου απόφασης. Αποτελεί επέκταση του ID3 αλγορίθμου και προτάθηκε από τον ερευνητή Quinlan το 1993. Ο ίδιος ερευνητής πρότεινε τον C4.5 ως μια βελτίωση του προηγούμενου αλγορίθμου, λόγω των περιορισμών και των προβλημάτων που υπήρξαν. Στον ID3 τα γνωρίσματα που μπορούσε να διαχειριστεί ήταν nominal, το σύνολο των δεδομένων δεν έπρεπε να συμπεριλαμβάνει ελλιπείς

τιμές, και τέλος η προσέγγιση της διάσπασης που χρησιμοποιούσε ο αλγόριθμος μπορούσε να οδηγήσει σε υπερπροσαρμογή.

Αποτελεί αλγόριθμο εποπτευόμενης μάθησης (supervised learning), και γι' αυτό απαιτεί ένα σύνολο από δεδομένα εκπαίδευσης (training set). Κάθε δείγμα μπορεί να θεωρηθεί ως ένα ζευγάρι αντικειμένου εισαγωγής (input object) και μιας επιθυμητής τιμής εξόδου (class). Ο αλγόριθμος αναλύει το training set και κατασκευάζει έναν κατηγοριοποιητή (classifier) που πρέπει να μπορεί να ταξινομήσει σωστά τόσο τα δεδομένα εκπαίδευσης, όσο και αυτά της δοκιμής (test set). Ένα δείγμα δοκιμής λειτουργεί ως αντικείμενο εισαγωγής και ο αλγόριθμος πρέπει να προβλέψει μια τιμή εξόδου, ώστε να αντιστοιχηθεί σε μια κλάση. Ο κατηγοριοποιητής που χρησιμοποιείται από τον C4.5, όπως προαναφέρθηκε, είναι ένα δέντρο απόφασης.

Τα δεδομένα εκπαίδευσης είναι ένα σύνολο S = s1, s2, … από ήδη ταξινομημένα δείγματα, που αποτελούν και το input του αλγορίθμου. Κάθε δείγμα Si = x1, x2, … είναι ένα διάνυσμα όπου τα x1, x2, … αντιπροσωπεύουν τις ιδιότητες ή τα χαρακτηριστικά γνωρίσματα του δείγματος. Επίσης, στα δεδομένα εκπαίδευσης αντιστοιχεί ένα διάνυσμα C = c1, c2, … όπου c1, c2, …cn αντιπροσωπεύει την κλάση στην οποία ανήκει κάθε δείγμα. Το output του αλγορίθμου είναι ένα δέντρο απόφασης ή/και ένα σύνολο από κανόνες που αναθέτει μια κλάση σε ένα καινούριο δείγμα.

Σε κάθε κόμβο, ο C4.5 χρησιμοποιεί το γεγονός ότι κάθε χαρακτηριστικό των δεδομένων μπορεί να χρησιμοποιηθεί για να λάβει μια απόφαση, η οποία χωρίζει τα δεδομένα σε μικρότερα υποσύνολα. Το κριτήριο διαχωρισμού είναι το *ομαλοποιημένο κέρδος πληροφοριών* (information gain). Το χαρακτηριστικό με το υψηλότερο *ομαλοποιημένο κέρδος πληροφοριών* επιλέγεται για να ληφθεί η απόφαση. Ο αλγόριθμος επαναλαμβάνεται για μικρότερες υπολίστες δεδομένων.

Το χαρακτηριστικό με το υψηλότερο *κέρδος πληροφοριών* υπολογίζεται με την χρήση τον παρακάτω εξισώσεων:

Εντροπία: $$E(S)=\sum_{i=1}^{n} -Pr(C_i)*logPr_2(C_i)$$

όπου, E(S) η εντροπία του S, n ο αριθμός των κλάσεων στο S, και Pr(Ci) η συχνότητα της κλάσης Ci στο S.

Κέρδος: $$G(S,A)=E(S)-\sum_{I=1}^{m} Pr(A_i)E(S_{Ai})$$

όπου, G(S,A) το κέρδος του S μετά από το διαχωρισμό του χαρακτηριστικού A, m ο αριθμός των τιμών του χαρακτηριστικού A στο S, Pr(Ai) η συχνότητα των περιπτώσεων όπου έχει τιμή Ai στο S, και E(SAi) το υποσύνολο του S με τα αντικείμενα που έχουν τιμή Ai.

Ο αλγόριθμος έχει τις παρακάτω base cases.

- Όλα τα δείγματα στη λίστα ανήκουν στην ίδια κλάση. Όταν συμβεί αυτό, απλά δημιουργεί έναν κόμβο φύλλου για το δέντρο απόφασης λέγοντας του να επιλέξει αυτή την κλάση.
- Κανένα από τα χαρακτηριστικά δεν παρέχει κέρδος πληροφοριών. Σε αυτή την περίπτωση, ο C4.5 δημιουργεί έναν κόμβο απόφασης υψηλότερα στο δέντρο χρησιμοποιώντας την αναμενόμενη τιμή της κλάσης.
- Εμφάνιση μιας μη προεμφανίσιμης κλάσης. Ξανά, ο C4.5 δημιουργεί έναν κόμβο απόφασης υψηλότερα στο δέντρο χρησιμοποιώντας την αναμενόμενη τιμή της κλάσης.

Οι βελτιώσεις του αλγόριθμου C4.5 σε σχέση με τον ID3 ήταν η διαχείριση αριθμητικών χαρακτηριστικών, το να επιτρέπονται οι ελλιπείς τιμές και να γίνεται κλάδεμα (tree pruning).

Παρακάτω παρουσιάζεται ο ψευδοκώδικας του αλγορίθμου.

ΑΛΓΟΡΙΘΜΟΣ C4.5

1. Έλεγξε για **base cases**
2. Για κάθε χαρακτηριστικό **A**
   - Βρίσκουμε το ομαλοποιημένο κέρδος πληροφοριών από το διαχωρισμό στο **A**
3. Έστω ότι **A_best** είναι το χαρακτηριστικό με το υψηλότερο ομαλοποιημένο κέρδος πληροφοριών
4. Δημιούργησε έναν κόμβο απόφασης **Node** που χωρίζεται στο **A_best**
5. Επανερχόμαστε στις υπολίστες που λαμβάνονται με το διαχωρισμό στο **A_best** και προσθέτουμε αυτούς τους κόμβους ως παιδιά του **Node**.

Τα πλεονεκτήματα του C4.5 είναι πως χτίζει μοντέλα που μπορούν να ερμηνευτούν εύκολα, είναι εύκολος στην εφαρμογή, μπορεί να χρησιμοποιήσει τόσο κατηγορηματικές όσο και συνεχείς τιμές και αντιμετωπίζει τον θόρυβο. Από την άλλη, τα μειονεκτήματα του είναι ότι μικρές διακυμάνσεις των δεδομένων μπορούν να οδηγήσουν σε διαφορετικά δέντρα απόφασης και δεν λειτουργεί καλά με μικρό training set.

B.2.2 Ο αλγόριθμος Random Forest

Ο ταξινομητής Random Forest,όπως υποδηλώνει και το όνομά του, είναι ένας από τους πιο γνωστούς ομοιογενής, συνδυαστικούς ταξινομητές, ο οποίος χρησιμοποιεί πολλά ασυσχέτιστα μεταξύ τους δέντρα αποφάσεων για να πραγματοποιήσει μια εκτίμηση. Ο αλγόριθμος αυτός αναπτύχθηκε από τους Leo Breiman και Adele Cutler το 2001. Αποτέλεσε μια επέκταση της ιδέας του Tin Kam Ho, ο οποίος το 1995 παρουσίασε μια γενική μεθοδολογία για τα τυχαία δάση απόφασης, σε συνδυασμό με την ιδέα του bagging του Breiman.

Κάθε μεμονωμένο δέντρο δείχνει μια πρόβλεψη κλάσης και η κλάση με τις περισσότερες ‘ψήφους’ γίνεται το μοντέλο πρόβλεψης. Είναι ένας αλγόριθμος που είναι αρκετά απλός, αλλά αρκετά ισχυρός. Ουσιαστικά, η θεμελιώδης αρχή πίσω από τον random forest που τον κάνει να λειτουργεί τόσο καλά είναι ο μεγάλος αριθμός σχετικά μη σχετιζόμενων μοντέλων, τα οποία λειτουργούν ως μια ‘επιτροπή’. Η επιτροπή αυτή θα φέρει καλύτερα αποτελέσματα από κάθε ένα από τα επιμέρους μοντέλα. Η βασική ιδέα μεταξύ των μοντέλων είναι η χαμηλή συσχέτιση μεταξύ τους. Σημαντικό είναι πως τα δέντρα προστατεύουν το ένα το άλλο από τα μεμονωμένα λάθη τους. Δηλαδή, ενώ κάποια δέντρα μπορεί να είναι λανθασμένα, πολλά άλλα μπορεί να είναι σωστά, και σαν μια ομάδα θα μπορέσουν να κινηθούν προς την σωστή κατεύθυνση. Επίσης, ο Random Forest στοχεύει στην μείωση του φαινομένου της υπερπροσαρμογής (overfitting) που προκαλεί την προσαρμογή του μοντέλου στον θόρυβο που υπάρχει στα δεδομένα.

Παρακάτω αναλύεται η λειτουργία του αλγορίθμου.
Αρχικά, θέτουμε ότι έστω έχουμε ένα σύνολο δεδομένων με N στοιχεία (instances).

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} & t_1 \\ x_{21} & x_{22} & \cdots & x_{2d} & t_2 \\ \vdots & \vdots & & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} & t_n \end{bmatrix}$$

Από τον παραπάνω πίνακα φαίνεται πως τα διανύσματα εισόδου έχουν διάσταση d. Δηλαδή,

$\overline{x_i} = \left(x_{i1}, x_{i2}, \ldots, x_{id} \right)$

Αρχικά, γίνεται η δημιουργία πολλών νέων συνόλων εκπαίδευσης με την μέθοδο bootstrap. Περιληπτικά, η μέθοδος bootstrap μπορεί να χρησιμοποιηθεί ώστε να εκτιμηθεί μια ποσότητα ενός

πληθυσμού. Αυτό γίνεται με τη συνεχή λήψη μικρών δειγμάτων, υπολογίζοντας τα στατιστικά τους στοιχεία, και παίρνοντας το μέσο όρο από τα υπολογισμένα στατιστικά στοιχεία.

Έστω ότι ο αριθμός των νέων συνόλων είναι T. Από κάθε σύνολο Ti κατασκευάζεται ένα δέντρο απόφασης. Για την κατασκευή κάθε δέντρου παίρνονται υπόψη m χαρακτηριστικά, και όχι όλα τα χαρακτηριστικά d που είναι διαθέσιμα. Ο αριθμός m επιλέγεται έτσι ώστε να ισχύει m << d, όπου ο m είναι σταθερός αριθμός. Οι δημοφιλέστερες τιμές του m είναι οι 12m,m, 2m.

Τέλος, από το σύνολο των δέντρων απόφασης που προκύπτουν, επιλέγεται με ψηφοφορία ο καλύτερος ταξινομητής.

Παρακάτω παρουσιάζεται ο ψευδοκώδικας του Random Forest.

ΑΛΓΟΡΙΣΜΟΣ RANDOM FOREST

1. Τυχαία επέλεξε **m** χαρακτηριστικά από τα συνολικά **d** χαρακτηριστικά, όπου **m << d**
2. Μεταξύ των **m** χαρακτηριστικών, υπολόγισε τον κόμβο **k** χρησιμοποιώντας το καλύτερο σημείο διαχωρισμού.
3. Χώρισε τον κόμβο σε κόμβους παιδιών χρησιμοποιώντας τον **καλύτερο διαχωρισμό.**
4. Επανέλαβε τα βήματα **1-3** μέχρι να φτάσεις σε **l αριθμό κόμβων.**
5. Χτίσε το δάσος επαναλαμβάνοντας τα βήματα **1-4** για **t αριθμό δέντρων.**

Ακολουθεί και ο ψευδοκώδικας πρόβλεψης με την χρήση του Random Forest.

ΑΛΓΟΡΙΘΜΟΣ ΠΡΟΒΛΕΨΗΣ ΜΕ RANDOM FOREST

1. Πάρε το test set και χρησιμοποίησε τους κανόνες για κάθε τυχαία δημιουργημένο δέντρο απόφασης.
   - Πρόβλεψε το αποτέλεσμα και αποθήκευσε το
2. Υπολόγισε τις ψήφους για κάθε προβλεπόμενο αποτέλεσμα.
3. Θεώρησε το αποτέλεσμα με τις περισσότερες ψήφους ως το την τελική πρόβλεψη του αλγόριθμου.

Ο Random Forest έχει παρατηρηθεί πως έχει ποικίλες εφαρμογές. Μερικές από αυτές είναι στον τραπεζικό τομέα, ώστε να βρεθεί ποιος από τους πιθανούς πελάτες μιας τράπεζα θα είναι συνεπείς στις υποχρεώσεις του και ποιος όχι, π.χ. αποπληρωμή δανείου. Στον τομέα της ιατρικής, για να ταυτοποιηθεί η ασθένεια ενός ασθενή αναλύοντας το ιατρικό του ιστορικό. Επίσης, έχει εφαρμοστεί και τον τομέα του ηλεκτρονικού εμπορίου, ώστε να εκτιμηθεί η πιθανότητα του να αρέσουν σε έναν πελάτη τα προτεινόμενα προϊόντα, βασιζόμενοι σε παρόμοια είδη πελατών.
Τα πλεονεκτήματα του συγκεκριμένου αλγόριθμου είναι αρκετά. Τα σημαντικότερα από αυτά είναι πως ξεπερνάται το πρόβλημα της υπερπροσαρμογής λόγω του συνδυασμού αποτελεσμάτων διαφορετικών δέντρων απόφασης, λειτουργεί καλύτερα για μεγάλο αριθμό δεδομένων σε σχέση με τα απλά δέντρα απόφασης, είναι πολύ ευέλικτος αλγόριθμος με υψηλή ακρίβεια, και η ακρίβεια παραμένει υψηλή ακόμα και όταν υπάρχει μεγάλη έλλειψη δεδομένων. Φυσικά, υπάρχουν και μειονεκτήματα, με σημαντικότερα τα ακόλουθα. Η κατασκευή των δασών είναι δυσκολότερη και πιο χρονοβόρα από αυτή των δέντρων απόφασης, είναι λιγότερο εύληπτη διαδικασία στην περίπτωση μεγάλης συγκέντρωσης δέντρων απόφασης και η διαδικασία πρόβλεψης μπορεί να αποβεί χρονοβόρα διαδικασία.

Β.2.3 Bayesian Κατηγοριοποίηση και Naive Bayes

Η Bayesian κατηγοριοποίηση αποτελεί μια κατηγορία μεθόδων δημιουργίας μοντέλων πρόβλεψης και κατηγοριοποίησης δεδομένων. Γίνεται, δηλαδή, μια πρόβλεψη της πιθανότητας ένα δείγμα X να ανήκει σε κάποια κατηγορία (class). Ονομάζεται bayesian επειδή στηρίζεται στην θεωρία κατηγοριοποίησης του Bayes.

Το θεώρημα του Bayes εκφράζεται μαθηματικά από την ακόλουθη εξίσωση:

$$P(A|B)=\frac{P(B|A)\,P(A)}{P(B)}$$

όπου το A και το B είναι γεγονότα και η $P(B) \neq 0$.

Επίσης,

- το P(A|B) αποτελεί μια δεσμευμένη πιθανότητα, δηλαδή την πιθανότητα το γεγονός A να συμβεί δεδομένου ότι το B είναι αληθές, και αντίστροφα.

- P(A) και P(B) είναι οι πιθανότητες το Α και το Β να παρατηρηθούν το ένα ανεξάρτητα από το άλλο. Αυτό το γεγονός είναι γνωστό και ως οριακή πιθανότητα (marginal).

Η πρώτη παραδοχή που πρέπει να γίνει είναι πως τα χαρακτηριστικά είναι ανεξάρτητα. Με άλλα λόγια, πως η παρουσία ενός συγκεκριμένου χαρακτηριστικού δεν επηρεάζει το άλλο. Ως εκ τούτου, αυτή η υπόθεση ονομάζεται αφελής (naive) και γι' αυτό και ο αλγόριθμος που χρησιμοποιεί αυτή την παραδοχή ονομάζεται Naive Bayes. Η δεύτερη παραδοχή που πρέπει να γίνει είναι ότι τα γεγονότα, στη συγκεκριμένη περίπτωση τα χαρακτηριστικά που χρησιμοποιούνται για την πρόβλεψη, έχουν την ίδια επίδραση στο αποτέλεσμα. Για παράδειγμα, λαμβάνοντας υπόψη την παραπάνω παραδοχή μπορούμε να καταλάβουμε εάν ένα χαρακτηριστικό (attribute) είναι συσχετισμένο με το αποτέλεσμα της κλάσης (class). Εάν η ακρίβεια (accuracy) παραμένει ακριβώς η ίδια όταν αφαιρεθεί το χαρακτηριστικό, φαίνεται ότι αυτό το χαρακτηριστικό είναι άσχετο με το αποτέλεσμα της κλάσης. Οπότε, ο κατηγοριοποιητής Naive Bayes αποτελεί μια πρακτική μέθοδο μάθηση που στηρίζεται σε στατιστικά στοιχεία (κατανομές πιθανότητας).

Σύμφωνα με το θεώρημα του Bayes, η πιθανότητα να ανήκει η παρατήρηση X στην κλάση Ci υπολογίζεται ως:

$$P(C_i|X)=\frac{P(X|C_i)*P(C_i)}{P(X)}$$

Ώστε να προβλεφθεί η κλάση μιας άγνωστης παρατήρησης ο Naive Bayes υπολογίζει τις πιθανότητες για την κάθε κλάση και εκχωρεί την παρατήρηση στην κλάση με τη μεγαλύτερη πιθανότητα. Το ζητούμενο είναι να υπολογιστεί η P(X|Ci), το οποίο δίνεται από την παρακάτω εξίσωση:

$$P(X|C_i)=\prod_{k=1}^{n} P(x_k|C_i)$$

όπου k είναι η διάσταση με τιμές x των χαρακτηριστικών X=(X1, X2, ..., Xn).

Αφού υπολογιστούν οι πιθανότητες P(Ci|X) για όλες τις κλάσεις, η παρατήρηση εκχωρείται στην κλάση με την μεγαλύτερη πιθανότητα.

Καταλαβαίνουμε, λοιπόν, πως ο Naive Bayes αποτελεί έναν εύκολο και κατανοητό αλγόριθμο, είναι γρηγορότερος στην πρόβλεψη της κλάσης σε σχέση με άλλους αλγόριθμους κατηγοριοποίησης και μπορεί να εκπαιδευτεί εύκολα όταν χρησιμοποιείται μικρός αριθμός δεδομένων. Βέβαια, η υπόθεση για την ανεξαρτησία των χαρακτηριστικών, μπορεί να θεωρηθεί σχεδόν αδύνατη να συναντηθεί στην πραγματική ζωή.

Β.2.4 Κ Πλησιέστεροι Γείτονες

Ο αλγόριθμος κ-πλησιέστεροι γείτονες (k-nearest neighbors) ή KNN ανήκει σε μια οικογένεια κατηγοριοποιητών που είναι βασισμένοι σε παραδείγματα (Instance Based Classifiers (IBC)), όπου η μάθηση βασίζεται στην αναλογία. Όπως είδαμε και πιο πριν άλλες μέθοδοι κατηγοριοποίησης όπως τα δέντρα απόφασης (αλγόριθμος C4.5, αλγόριθμος random forest), και ο Naive Bayes δημιουργούν κάποιο μοντέλο και αυτό το μοντέλο είναι που χρησιμοποιείται στη συνέχεια για την κατηγοριοποίηση των καινούριων δειγμάτων. Έφοσον στους κατηγοριοποιητές IBC το στάδιο της εκπαίδευσης παραλείπεται, συνεπώς, δεν δημιουργείται κάποιο μοντέλο. Συχνά, χαρακτηρίζονται ως 'τεμπέληδες' (lazy classifiers) εξαιτίας του ότι η κατηγοριοποίηση γίνεται βάσει των ήδη γνωστών παρατηρήσεων. Γι' αυτό και είναι απαραίτητη η αποθήκευση όλων ή μερικών παρατηρήσεων εκπαίδευσης.

Ο αλγόριθμος KNN είναι μια μη παραμετρική μέθοδος που χρησιμοποιείται τόσο για την κατηγοριοποίηση (classification) όσο και για την παλινδρόμηση (regression). Στηρίζεται κυρίως στη χρήση μέτρων βασισμένων στην απόσταση. Η κεντρική ιδέα είναι πως η τιμή της συνάρτησης-στόχου (τιμή κλάσης) για ένα νέο στιγμιότυπο βασίζεται αποκλειστικά στις αντίστοιχες τιμές των k πλησιέστερων στιγμιότυπων εκπαίδευσης, τα οποία αποτελούν τους 'γείτονες' του. Για να καθοριστεί ο αλγόριθμος πρέπει πρώτα να καθοριστούν δύο προϋποθέσεις. Η πρώτη από αυτές είναι να οριστεί η απόσταση μεταξύ των δύο στιγμιότυπων. Η απόσταση αυτή εκφράζει την τιμή πάνω στο χώρο των στιγμιότυπων, που θα εκφράζει την εγγύτητα, δηλαδή την 'ομοιότητα' μεταξύ των στιγμιότυπων. Η πιο γνωστή και συνηθέστερη απόσταση είναι η Ευκλείδεια, η οποία δίνεται από τον τύπο: $d(X,Y) = \sqrt{((x_1 - y_1)^2 + (x_2 - y_2)^2)}$. Μπορούν να επιλεχθούν και άλλα μέτρα ομοιότητας-ανομοιότητας, αντί της Ευκλείδειας. Ορισμένα από αυτά είναι η City block (Manhattan) απόσταση και η Chebychev απόσταση.

Το δεύτερο ζήτημα που πρέπει να αποφασιστεί είναι η τιμή του k. Το k εκφράζει τον αριθμό των γειτόνων που θα ληφθεί υπόψη στις συγκρίσεις. Αν, για παράδειγμα, το k=1, τότε για την εξαγωγή της απόφασης θα συμπεριληφθεί ο ένας 'γείτονας' που βρίσκεται στην πιο κοντινή απόσταση. Αν το k=5, αντίστοιχα θα ληφθούν υπόψη οι 5 κοντινότεροι 'γείτονες'. Σε αυτή την περίπτωση, το στιγμιότυπο θα πάρει την κλάση που έχει η πλειοψηφία από τους 'γείτονες'. Χρησιμοποιώντας μεγαλύτερο αριθμό k γειτόνων δίνονται ομαλότερες περιοχές απόφασης. Όμως, αν τεθεί πολύ μεγάλος αριθμός k γειτόνων παρατηρούνται οι εξής ανεπιθύμητες συμπεριφορές.

Αυξάνεται το υπολογιστικό κόστος και καταστρέφεται η σωστή εκτίμηση των δεδομένων καθώς λαμβάνονται υπόψη παραδείγματα που έχουν μεγάλη απόσταση. Συχνά, για να οριστεί το k χρησιμοποιείται η μέθοδος της διασταυρωμένης επικύρωσης (cross-validation method). Επειδή, όμως, απαιτούνται πολλές πράξεις ορισμένες φορές αυτή η μέθοδος αποδεικνύεται χρονοβόρα. Παρατηρείται ότι για k>1 η διάρκεια εκτέλεσης της μεθόδου δεν είναι ικανοποιητική αφού κάθε δεδομένο πρέπει να αποθηκευτεί και μετά να εξεταστεί για την κατηγοριοποίησή του.

Παρακάτω παρουσιάζεται ο ψευδοκώδικας του KNN

ΑΛΓΟΡΙΘΜΟΣ KNN

1. Φόρτωσε τα training και test data.
2. Επίλεξε την τιμή του K
3. Για κάθε σημείο στα test data
    a) βρες την Ευκλείδεια απόσταση σε όλα τα σημεία των training data
    b) αποθήκευσε τις Ευκλείδειες αποστάσεις σε μια λίστα και ταξινόμησέ τες
    c) επίλεξε τα πρώτα K σημεία
    d) όρισε μια κλάση βασισμένη στην πλειοψηφία των σημείων που εμφανίζονται ως επιλεγμένα σημεία
4. Τέλος

Οι κατηγοριοποιητές του χρησιμοποιούν τον KNN είναι αρκετά αποτελεσματικοί όταν υπάρχουν σύνθετες εξαρτήσεις μεταξύ των μεταβλητών, ο αλγόριθμο είναι σχετικά απλός, και συχνά επιτυγχάνουν υψηλές επιδόσεις. Από την άλλη, βέβαια, η κατηγοριοποίηση νέων παρατηρήσεων διαρκεί πολύ περισσότερο χρόνο, ειδικά στις περιπτώσεις όπου ο αριθμός των 'γειτόνων' είναι μεγάλος. Επίσης, τα αποτελέσματα τους μπορούν να επηρεαστούν σε σημαντικό βαθμό από το πλήθος των γειτόνων k και υπάρχει ευαισθησία σε τοπικά χαρακτηριστικά των δεδομένων.

Β.2.5 Νευρωνικά Δίκτυα

Τα Νευρωνικά Δίκτυα (Neural Networks) πήραν το όνομά τους από το βιολογικό νευρικό σύστημα από το οποίο είναι και εμπνευσμένα, και ειδικότερα από τον ανθρώπινο εγκέφαλο. Έχουν αξιοσημείωτα χαρακτηριστικά, όπως τη δυνατότητα να αναπαριστούν σύνθετες εξαρτήσεις ή την ικανότητα τους να προβλέπουν την κλάση άγνωστων παρατηρήσεων. Υπάρχουν τύποι νευρωνικών δικτύων κατάλληλοι τόσο για επιβλεπόμενη, όσο και για μη επιβλεπόμενη μάθηση.

Παρακάτω παρουσιάζονται τα βασικά συστατικά ενός νευρωνικού δικτύου.

- **Νευρώνας:** αποτελεί στοιχειώδης υπολογιστική μονάδα, η οποία δέχεται τιμές εισόδου και υπολογίζει μια τιμή εξόδου.
- **Συνδέσεις:** κατευθυνόμενα βέλη που συνδέουν τους νευρώνες μεταξύ τους. Μέσω των συνδέσεων ένας νευρώνας δέχεται τιμές εισόδου από άλλους νευρώνες και μπορεί να μεταβιβάσει την τιμή εξόδου του σε άλλους νευρώνες.
    - όταν έχουμε αμφίδρομες συνδέσεις το δίκτυο χαρακτηρίζεται **αναδρομικό**
    - όταν δεν έχουμε αμφίδρομες συνδέσεις το δίκτυο χαρακτηρίζεται **απλής προώθησης**
- **Βάρος:** αριθμητική τιμή που συνοδεύει κάθε σύνδεση και επηρεάζει την επίδραση μεταξύ των συνδεδεμένων νευρώνων. Η τιμή εξόδου ενός νευρώνα πολλαπλασιάζεται με το βάρος της σύνδεσης αυτού του νευρώνα με κάποιον άλλο.

Αρχικά, σε έναν νευρώνα αθροίζονται οι τιμές εισόδου. Οι τιμές εισόδου ισούνται με τις τιμές εξόδου των συνδεδεμένων νευρώνων, πολλαπλασιασμένες με τα βάρη των αντίστοιχων συνδέσεων, όπως φαίνεται στην παρακάτω εξίσωση.

$$S_i = \sum_{j=1}^{n} w_{ij} * u_j$$

όπου, i είναι ο νευρώνας, uj οι τιμές εισόδου που δέχεται ο νευρώνας από άλλους n νευρώνες.

Έπειτα, ακολουθεί το στάδιο του μετασχηματισμού του αθροίσματος των τιμών εισόδου. Η συνάρτηση που χρησιμοποιείται δείχνει την τιμή εξόδου του νευρώνα, και ονομάζεται ‘συνάρτηση μετασχηματισμού’ ή ‘συνάρτηση ενεργοποίησης’. Η παρακάτω αποτελεί την συνηθέστερη συνάρτηση μετασχηματισμού, που είναι η σιγμοειδής συνάρτηση.

$$f(x) = \frac{1}{1+\mathrm{e}^{-x}}$$

Άλλες συναρτήσεις ενεργοποίησης είναι η συνάρτηση ημιτόνου, η συνάρτηση συνημιτόνου, η συνάρτηση υπερβολικής εφαπτομένης.

Όσον αφορά την κατηγοριοποίηση χρησιμοποιούνται κυρίως τα δίκτυα απλής προώθησης. Το δημοφιλέστερο ίσως από αυτά είναι τα δίκτυα πολυεπίπεδων αισθητήρων (**Multilayer**

Perceptron). Στα δίκτυα πολυεπίπεδων αισθητήρων οι νευρώνες είναι οργανωμένοι σε στρώματα (layers). Τα βασικά στρώματα είναι:

- **Στρώμα Εισόδου:** Το πρώτο στρώμα ενός νευρωνικού δικτύου. Δεν μετασχηματίζουν την τιμή.
- **Κρυφό Στρώμα:** Το δεύτερο στρώμα ενός νευρωνικού δικτύου. Δέχονται τις τιμές των νευρώνων εισόδου πολλαπλασιασμένες με τα βάρη των συνδέσεων, τις αθροίζουν και μετασχηματίζουν το άθροισμα σύμφωνα με τη συνάρτηση μετασχηματισμού. Μπορεί να υπάρχουν πάνω από ένα κρυφά στρώματα.
- **Στρώμα Εξόδου:** Το τελικό στρώμα ενός νευρωνικού δικτύου, όπου υπολογίζεται η τελική πρόβλεψη του δικτύου. Δέχεται τις τιμές εξόδου των κρυφών νευρώνων, πολλαπλασιασμένες με τα βάρη των συνδέσεων. Συνήθως υπάρχει ένας νευρώνας εξόδου για κάθε δυνατή τιμή της κλάσης.

Να σημειωθεί πως κάθε νευρώνας διαβιβάζει τιμές σε όλους τους νευρώνες του επόμενου στρώματος και μόνο σε αυτούς. Μια ιδέα που προέκυψε από τα βιολογικά νευρικά κύτταρα είναι η εξής. Στον k-οστό νευρώνα υπάρχει ένα συναπτικό βάρος wk0 με ιδιαίτερη σημασία, που ονομάζεται πόλωση ή κατώφλι (bias or threshold). Η τιμή του κατωφλιού είναι πάντα ίση με 1. Ο νευρώνας ενεργοποιείται, όταν το άθροισμα των τιμών εισόδου είναι μεγαλύτερο από την τιμή του κατωφλιού. Αν, είναι μικρότερο ο νευρώνας μένει ανενεργός.
Τα νευρωνικά δίκτυα έχουν αρκετές εφαρμογές. Χρησιμοποιούνται στην ιατρική διάγνωση για την χορήγηση της κατάλληλης θεραπείας, στις χρηματιστηριακές προβλέψεις για προβλέψεις σχετικά με τις τιμές των μετοχών, στα συστήματα διαχείρισης κινητήρα για την ελαχιστοποίηση κατανάλωσης καυσίμων και σε πολλούς ακόμα τομείς.

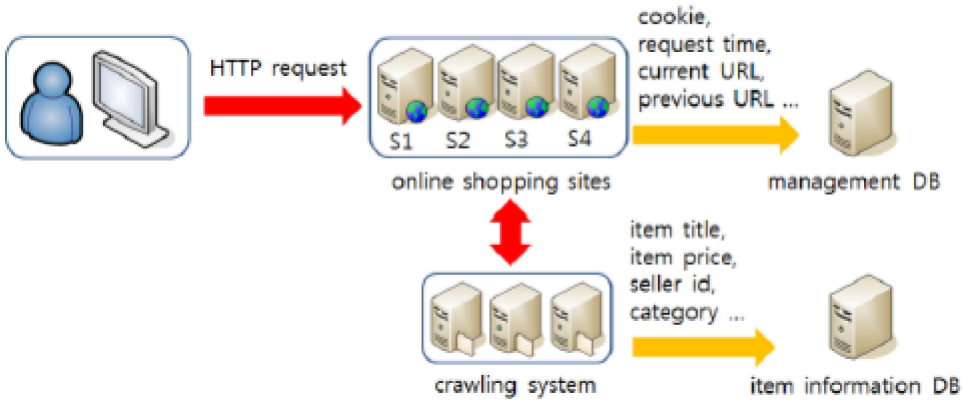
Συμπεραίνουμε, λοιπόν, πως τα πλεονεκτήματα των νευρωνικών δικτύων είναι σημαντικά και αδιαμφισβήτητα. Μπορούν να χειριστούν θορυβώδη και ασυνεπή δεδομένα, ικανά να κατηγοριοποιήσουν αντικείμενα που είναι άγνωστα στο σύστημα και η ύπαρξη των κρυφών στρωμάτων επιτρέπει την ικανοποιητική προσέγγιση σύνθετων συναρτήσεων. Από την άλλη τα μειονεκτήματα δεν παραλείπονται. Συχνά απαιτείται εμπειρικός προσδιορισμός των παραμέτρων τους, ο τρόπος λήψης αποφάσεων είναι συχνά ακατανόητος και απαιτούν μεγάλους χρόνους εκπαίδευσης.

Β.3 Ανάλυση σχετικής βιβλιογραφίας

Στις παραγράφους που ακολουθούν θα γίνει ανάλυση άρθρων σχετικών με το θέμα της εργασίας. Συγκεκριμένα, γίνεται η ανάλυση 8 άρθρων που αφορούν την ταξινόμηση παραδειγμάτων με σκοπό την πρόβλεψη αγοράς ή μη προϊόντων από τον επισκέπτη στο ηλεκτρονικό κατάστημα. Κάθε ένα από τα επόμενα άρθρα χρησιμοποιεί διάφορες τεχνικές εξόρυξης των απαιτούμενων δεδομένων, καθώς και την δοκιμή διαφορετικών αλγορίθμων για την διαδικασία εκπαίδευσης του μοντέλου πρόβλεψης.

Άρθρο 1: Online Footsteps to Purchase: Exploring Consumer Behaviors on Online Shopping Sites (Munyoung Lee et al., 2015)

Στην παρούσα έρευνα χρησιμοποιούνται δεδομένα από τέσσερα διαφορετικά online καταστήματα της Νότιας Κορέας, με σκοπό την ανάλυση της συμπεριφοράς των online πελατών και την κατανόηση των παραγόντων που σχετίζονται με τις αγορές τους, Στόχος είναι η ανάπτυξη ενός μοντέλου πρόβλεψης της επικείμενης αγοράς. Όσο αφορά τα είδη των καταστημάτων που ερευνήθηκαν, μέσα σε αυτά ανήκαν ένα κατάστημα γενικής πώλησης (παρόμοιο με το e-bay), ένα πώλησης ρουχισμού, ένα πώλησης επίπλων και ένα πώλησης αθλητικών ειδών. Για την βαθύτερη κατανόηση της καταναλωτικής συμπεριφοράς οι συγγραφείς, πριν προβούν στην ανάπτυξη του μοντέλου πρόβλεψης, μελέτησαν σε τι ποσοστό οι πελάτες προέρχονται από εξωτερικές πηγές αναζήτησης (π.χ. μηχανές αναζήτησης) και παρατήρησαν την πορεία κάθε πελάτη μέσα στο ηλεκτρονικό κατάστημα. Η έρευνα έδειξε πως αν ένα αντικείμενο προβληθεί τουλάχιστον 3 φορές από τον πελάτη και τοποθετηθεί στο καλάθι αγορών, τότε η πιθανότητα αγοράς του κυμαίνεται μεταξύ 58%-88%.



Εικόνα 1: Άντληση δεδομένων και αποθήκευσή τους

Τα δεδομένα που χρησιμοποιήθηκαν στο μοντέλο χωρίζονται σε τρία datasets και συλλέχθηκαν μέσω τρίτης εταιρείας και ενός crawler από την 1 έως τις 24 Μαρτίου 2014. Το πρώτο αφορά δεδομένα σχετικά με το ιστορικό πρόσβασης του πελάτη (urls, logs κτλ) και το δεύτερο αφορά πληροφορίες σχετικές με τα αντικείμενα προς πώληση (τίτλος, τιμή, κατηγορία κτλ). Το τρίτο dataset που εξετάστηκε αποτελούσε συνδυασμό των δυο πρώτων datasets. Η κατηγοριοποίηση των δεδομένων έγινε με Μηχανή Διανυσμάτων Υποστήριξης (Support Vector Machine), όπου ως συνάρτηση kernel χρησιμοποιήθηκε η RBF (Radial Basic Function), και έγινε 5-cross validation. Τα αποτελέσματα που προέκυψαν έδειξαν πως το μοντέλο που χρησιμοποιήθηκαν μόνο τα χαρακτηριστικά συμπεριφοράς είχε υψηλότερη Ακρίβεια ίση με 82% , σε αντίθεση με το μοντέλο που χρησιμοποιήθηκαν μόνο τα χαρακτηριστικά των αντικειμένων με Ακρίβεια μόνο 58%. Τέλος, για το συνδυαστικό μοντέλο χαρακτηριστικών η Ακρίβεια ήταν επίσης αρκετά υψηλή και έφτασε μέχρι και το 84%.

<u>Άρθρο 2:</u> Predictive Analytics of E-Commerce Search Behavior for Conversion (Xi Niu, Chuqin Li, Xing Yu, 2017)

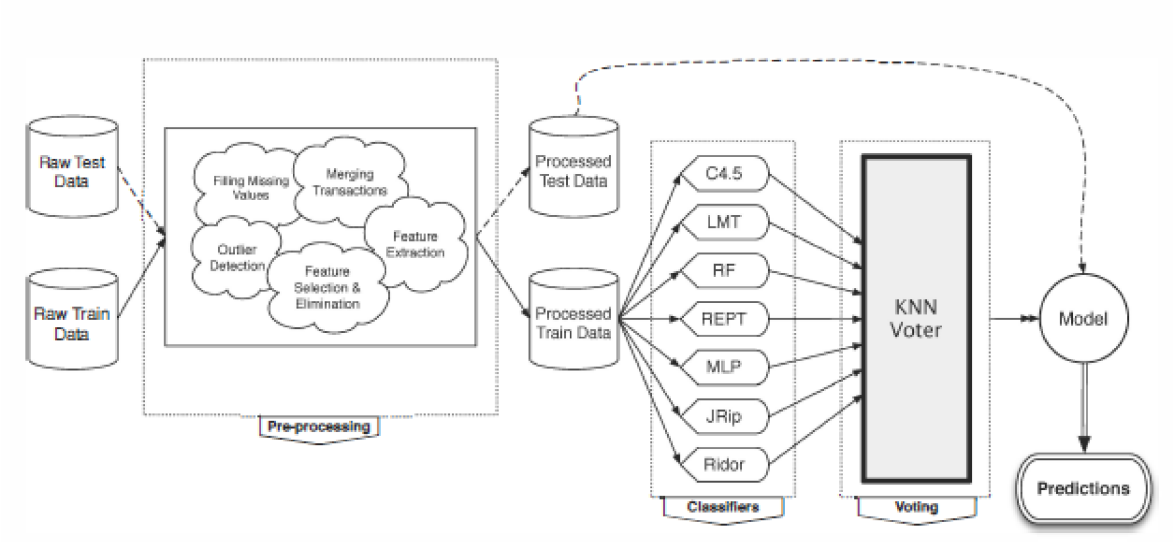
Σε αυτή την έρευνα οι συγγραφείς προσπαθούν να υπολογίσουν το ποσοστό μετατροπής (conversion rate), δηλαδή το ποσοστό των επισκέψεων σε ένα ηλεκτρονικό κατάστημα που κατέληξαν στην πραγματοποίηση κάποιας αγοράς. Τα δεδομένα που χρησιμοποιήθηκαν προήρθαν από τα weblogs του Walmart. Το dataset που χρησιμοποιήθηκε περιείχε 6.944.874 logs, από τα οποία πάρθηκαν μεταβλητές τριών επιπέδων και τεσσάρων κατηγοριών και προέκυψαν 1.530.738 συναλλαγές, από τις οποίες όμως μόνο 4.9% αποτελούσαν συναλλαγές που κατέληξαν σε αγορά. Η στατιστική ανάλυση που έγινε στα δεδομένα έδειξε πως κατά μέσο όρο ο επισκέπτης παραμένει περίπου 3 λεπτά στη σελίδα του αντικειμένου πριν το αγοράσει.

Εξετάστηκαν δύο διαφορετικά μοντέλα πρόβλεψης, το πρώτο με τη χρήση του Random Forest και το δεύτερο με την χρήση της λογικής παλινδρόμησης (logistic regression). Λόγω της μεγάλης ανισότητας των κατηγοριών των δεδομένων εφαρμόστηκε η μέθοδος undersampling για την εξισορρόπησή τους. Όσο αφορά τον Random Forest, χρησιμοποιήθηκαν 15 μεταβλητές και η Ακρίβεια του μοντέλου με την χρήση του 10-cross validation έφτασε το 76%. Από την ανάλυση του μοντέλου προέκυψε πως τα χαρακτηριστικά που παίζουν μεγάλο ρόλο στο αποτέλεσμα είναι ο χρόνος που περνάει ο πελάτης στην σελίδα του αντικειμένου και το πόσο διαφοροποιημένα είναι τα κλικ των χρηστών. Από την άλλη, τα χαρακτηριστικά με την μικρότερη επίδραση στο αποτέλεσμα ήταν αυτά που σχετίζονται με την αναζήτηση γενικά κάποιου όρου ή αντικείμενου. Στη συνέχεια, για την λογική παλινδρόμηση, επιλέχτηκαν 11 μεταβλητές ώστε να αποφευχθεί η

μεγάλη πολυκεντρικότητά τους. Παρατηρήθηκε πως οι καταχωρημένοι χρήστες είχαν 2.28% μεγαλύτερη πιθανότητα να αγοράσουν από τους μη καταχωρημένους. Τέλος, παρατηρήθηκε πως το μοντέλο του Random Forest είχε πολύ καλύτερη απόδοση από αυτό της λογικής παλινδρόμησης.

Άρθρο 3: A Study on Prediction of User's Tendency Toward Purchases in Websites Based on Behavior Models (Boroujerdi et al., 2014)

Στην παρούσα έρευνα χρησιμοποιήθηκαν 500.000 συνεδρίες ως training set και 10.000 συνεδρίες ως test set. Πριν την προεπεξεργασία των δεδομένων τα χαρακτηριστικά ήταν 23, αλλά αφού επεξεργάστηκαν και επιλέχτηκαν τα κατάλληλα έφτασαν τα 45. Για την κατηγοριοποίηση των δεδομένων εξετάστηκαν 3 διαφορετικές κατηγορίες αλγορίθμων και διάφοροι αλγόριθμοι που ανήκουν σε αυτές. Η πρώτη κατηγορία ήταν αυτή των δέντρων αποφάσεων (Functional Trees, C4.5, Alternating Desicion Tree, Logistic Model Trees, RandomForest, RandomTree, REPTree), η δεύτερη ήταν η κατηγορία των συναρτησιακών αλγορίθμων (Logistic Regression, RBFNetwork, MultilayerPerceptron, Spegasos, SMO) και η τρίτη κατηγορία ήταν οι αλγόριθμοι βασισμένοι σε κανόνες (ConjuctiveRule, Jrip, Ridor). Παρατηρήθηκε πως κάθε ομάδα είχε παρόμοια αποτελέσματα ακρίβειας μεταξύ των αλγορίθμων που ανήκαν σε αυτή, αλλά υπήρχαν διαφορές στην ακρίβεια από τη μια ομάδα στην άλλη. Επίσης, υποκατηγορίες του dataset που προβλέπονταν σωστά από τα δέντρα απόφασης, μπορεί να μην προβλέπονταν σωστά από τους αλγόριθμους βασισμένους σε κανόνες, και αντίστροφα. Γι' αυτό, ώστε να αντιμετωπιστεί αυτό το πρόβλημα, χρησιμοποιήθηκε η τεχνική του voting μεταξύ αυτών των δύο οικογενειών αλγορίθμων. Δοκιμάστηκαν δύο διαφορετικοί Voters για τον συνδυασμό των μοντέλων ταξινόμησης. Ο πρώτος ήταν ένας Weighted Voter που αύξησε ελάχιστα την τελική ακρίβεια. Για τον δεύτερο Voter 'χτίστηκε' ένα μοντέλο ταξινόμησης KNN (με k=2) που εφαρμοζόταν στα outputs των ταξινομητών. Τελικά, 7 αλγόριθμοι (C4.5, LMT, RandomForest, REPTree, MultiLayer Perceptron, JRip, και Ridor) δοκιμάστηκαν με τον δεύτερο voter. Αυτό το τελικό μοντέλο voting αλγορίθμου επίτυχε 97.2% Ακρίβεια, ποσοστό τρομερά υψηλό.



Εικόνα 2: Παρουσίαση του μοντέλου με την χρήση voter

Άρθρο 4: Buying or Browsing? : Predicting Real-time Purchasing Intent using Attention-based Deep Network with Multiple Behavior (Long Guo et al., 2019)

Στόχος της συγκεκριμένης έρευνας είναι να αναπτύξει ένα μοντέλο πρόβλεψης της πρόθεσης του πελάτη σε πραγματικό χρόνο. Προτείνεται συγκεκριμένα ένα end-to-end deep Network, το οποίο ονομάζεται DIPN (Deep Intent Prediction Network). Η διαφορά του από τις προϋπάρχοντες προτάσεις μοντέλων είναι πως κάνει χρήση τόσο των χαρακτηριστικών που συλλέγονται από τον περιηγητή (browser), όσο και από τις πληροφορίες που παίρνουμε από την συμπεριφορά του χρήστη που σχετίζονται με τις συσκευές αφής. Γι΄ αυτό και χρησιμοποιούνται δύο dataset, από τα οποία το πρώτο περιέχει πληροφορίες 'περιήγησης' (π.χ. προιόντα που αναζητήθηκαν, ή που προστέθηκαν στο καλάθι) και το δεύτερο περιέχει πληροφορίες σχετικές με την αφή (π.χ. αριθμός swipes, αριθμός taps). Η αρχιτεκτονική του μοντέλου αποτελείται από τέσσερα στρώματα: Embedding Layer, RNN Layer, Hierarchical Attention Layer, Multi Task Layer. Ο τρόπος που χρησιμοποιείται είναι ο εξής. Αρχικά, η εκπαίδευση του μοντέλου έγινε σε ισχυρούς servers στο cloud, και μετά με τη χρήση του AliNN (λύση για την ανάπτυξη μοντέλων μηχανικής μάθησης) συμπιέστηκε και εφαρμόστηκε στις κινητές συσκευές. Τα αποτελέσματα πρόβλεψης στέλνονται στο cloud και η ανταπόκριση του συστήματος είναι μεταξύ 20-50 ms ανάλογα τη συσκευή. Για την αξιολόγηση του μοντέλου χρησιμοποιήθηκε η AUC, η οποία έφτασε έως και 0.84.

Άρθρο 5: Learning Item Temporal Dynamics for Predicting Buying Sessions (Bogina et al., 2016)

Σε αυτή την έρευνα προτείνεται ένα μοντέλο για την πρόβλεψη της κατάληξης μιας επίσκεψης σε ένα ηλεκτρονικό κατάστημα. Σε διαφορά με προηγούμενες έρευνες λαμβάνονται υπόψη η δημοτικότητα του προϊόντος, με άλλα λόγια αν είναι στις τάσεις των αναζητήσεων (trendiness). Το ερώτημα που απαντάται είναι κατά πόσο αυτό το χαρακτηριστικό μπορεί να χρησιμοποιηθεί και να αξιοποιηθεί ώστε να γίνει η πρόβλεψη. 80% από τα δεδομένα χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου και τα υπόλοιπα 20% για το πειραματικό κομμάτι. Επειδή τα δεδομένα ήταν εξαιρετικά άνισα, χρησιμοποιήθηκε ο SMOTE για την εξισορρόπησή τους, ως συνδυασμός των τεχνικών του oversampling και του undersampling. Η τεχνική που χρησιμοποιήθηκε ήταν αυτή του Bagging, και προτιμήθηκε γιατί σε προηγούμενες έρευνες αποδείχτηκε πως οι υβριδικές τεχνικές λειτουργούν καλύτερα και αποδοτικότερα. Οι αλγόριθμοι που χρησιμοποιήθηκαν ήταν ο 1R, ο RepTree και ο NBTree. Για το dataset που περιείχε και το χαρακτηριστικό της τάσης (trendiness), ο 1R είχε την καλύτερη απόδοση σε σχέση με τους υπόλοιπους αλγόριθμους που μελετήθηκαν, όπου η παράμετρος Precision άγγιξε το 0.937. Όσο αφορά τα dataset που δεν περιείχε το χαρακτηριστικό της τάσης, η παράμετρος Presicion μειώθηκε για τον R1 στο 0.64 ενώ για τον RepTree υπήρχε μικρή μείωση στο 0.82 από το 0.85. Τέλος, για τον NBTree η Precision από 0.92 έπεσε στο 0.84.

Άρθρο 6: A Scalable Purchase Intention Prediction System Using Extreme Gradient Boosting Machines with Browsing Content Entropy (Zheng, Liu, 2018)

Η συγκεκριμένη έρευνα επικεντρώνεται στην ανάπτυξη ενός μοντέλου που χρησιμοποιεί την εντροπία των χαρακτηριστικών για την πρόβλεψη της συμπεριφοράς των επισκεπτών σε ένα ηλεκτρονικό κατάστημα. Επίσης, επικεντρώνεται στην τεχνική των gradient boosting μηχανών και στην κατασκευή ενός μοντέλου που ονομάζεται extreme gradient boosting (XGBoost). Τα δεδομένα που χρησιμοποιήθηκαν προέρχονται από μεγάλη Ευρωπαϊκή επιχείρηση πώλησης διάφορων προϊόντων κατανάλωσης. Συνολικά το dataset περιείχε 9.249.729 διαδικτυακές συνεδρίες. Επιλέχτηκε τυχαία το 80% από τις συνεδρίες ως training set και το υπόλοιπο 20% ως test set. Ο XGBoost συγκρίθηκε με άλλους αλγορίθμους, οι οποίοι ήταν ο GBM (gradient boosting machine) και ο AdaBoost. Από τους τρεις αλγορίθμους αυτός που είχε την υψηλότερη Ακρίβεια ήταν ο AdaBoost με ποσοστό 95.5%. Ακολούθησε δεύτερος ο XGBoost με ποσοστό 91.19% και τέλος ο GBM με το χαμηλότερο και μη ικανοποιητικό ποσοστό του 58.54%.

Άρθρο 7: Prediction of User's Purchase Intention Based on Machine Learning (Bing, Yuliang, 2016)

Στην παρούσα εργασία εξετάζεται η κατηγοριοποίηση των επισκεπτών ενός ηλεκτρονικού καταστήματος, με την χρήση του αλγορίθμου C4.5 και του Naive Bayes. Χρησιμοποιήθηκαν συνολικά 15.000 δεδομένα με χαρακτηριστικά όπως η ηλικία του χρήστη, το φύλο, η περιοχή, το δίκτυο πρόσβασης και ο χρόνος που καταναλώθηκε στην σελίδα από τον χρήστη. Χρησιμοποιήθηκε η k-cross validation, όπου τα δεδομένα χωρίστηκαν σε τρεις ισάριθμες ομάδες. Παρατηρήθηκε πως ο Naive Bayes ήταν πιο γρήγορος στον σχηματισμό του μοντέλου κατά 13% σε σχέση με τον C4.5. Τέλος, παρατηρήθηκε πως η ακρίβεια του C4.5 ήταν πιο σταθερή και υψηλότερη από αυτή του Naive Bayes, όταν το σύνολο των δεδομένων αυξανόταν σταδιακά. Εν αντιθέσει, ο Naive Bayes ενώ είχε σχετικά υψηλή ακρίβεια (~78%) για το σύνολο των 5000 δεδομένων, όταν αυτό το σύνολο αυξήθηκε η ακρίβεια άρχισε να πέφτει.

Άρθρο 8: Will this session end with a purchase? Inferring current purchase intent of anonymous visitors (Mokryn et al., 2019)

Στόχος της παρούσας έρευνας είναι να δημιουργήσει ένα μοντέλο πρόβλεψης για πελάτες ηλεκτρονικών καταστημάτων που είναι ανώνυμοι. Δηλαδή, να προβλέπει την πρόθεση αγοράς προϊόντων για πελάτες που δεν έχουν κάποιο προηγούμενο ιστορικό αγορών στο κατάστημα. Επικεντρώνεται στο πόσο δημοφιλές είναι ένα προϊόν στη σελίδα και πως μπορεί αυτή η τάση (trendiness) να επηρεάσει την συμπεριφορά του χρήστη. Όπως, είδαμε και πιο πάνω αυτό το χαρακτηριστικό έχει εξεταστεί ξανά σε άλλη έρευνα (βλέπε 2.3.5). Χρησιμοποιούνται δύο dataset από διαφορετικά ηλεκτρονικά καταστήματα. Το πρώτο αποτελεί κατάστημα ειδών κηπουρικής, εργαλείων και διάφορων ηλεκτρονικών ειδών, ενώ το δεύτερο αποτελεί κατάστημα πώλησης ρουχισμού. Λόγω της μεγάλης ανισότητας στην κατηγοριοποίηση των δεδομένων και στα δύο dataset εφαρμόστηκε η τεχνική SMOTE. Τα χαρακτηριστικά που χρησιμοποιήθηκαν επικεντρώνονται στον αριθμό των click και στην δημοτικότητα (trendiness) του κάθε προϊόντος. Παρατηρήθηκε πως και στα δύο καταστήματα οι συνεδρίες που δεν οδήγησαν σε αγορά ήταν πολύ πιο σύντομες σε διάρκεια από αυτές που τελικά η συνεδρία οδήγησε σε αγορά. Αυτό φαίνεται λογικό, αφού οι πελάτες που είναι αποφασισμένοι να αγοράσουν κάτι θα ξοδέψουν περισσότερο χρόνο στο προϊόν. Χρησιμοποιήθηκαν διάφοροι αλγόριθμοι (Logistic Regression, Bagging, NBTree και XGBoost) για διαφορετικές αγοραστικές περιόδους, έτσι ώστε να παρατηρηθεί για πόσο καιρό η δημοτικότητα ενός προϊόντος συνεχίζει να επηρεάζει τις πωλήσεις του καταστήματος. Παρατηρήθηκε πως ο NBTree ήταν αυτός που είχε το υψηλότερο F1 Score, τόσο όταν το χαρακτηριστικό trendiness συμπεριλαμβανόταν (F1= 0.91), όσο και όταν δεν λαμβανόταν υπόψη (F1= 0.88).

Β.3.1 Συμπεράσματα ανάλυσης βιβλιογραφίας

Συγκρίνοντας τους αλγορίθμους που έχουν χρησιμοποιηθεί στις παραπάνω ερευνητικές εργασίες παρατηρούμε πως πρωταγωνιστικό ρόλο έχουν τα δέντρα απόφασης. Χρησιμοποιούνται σε τέσσερεις από τις οκτώ έρευνες και παρατηρούμε πως τα μοντέλα που προκύπτουν έχουν υψηλά ποσοστά σωστής κατηγοριοποίησης. Στο άρθρο του Boroujerdi et al (2014), έχει χρησιμοποιηθεί πληθώρα διαφόρων δέντρων απόφασης με υψηλότερα ποσοστά να σημειώνουν οι Random Forest και ο REPTree με F1 Score ίσο με 97%, και ο C4.5 με ποσοστό 96.9%. Άλλοι αλγόριθμοι που έχουν σημειώσει αρκετά υψηλό F1 Score είναι ο 1R με 92.9% (Bogina et al, 2016), ο XGBoost με 91.19% και ο AdaBoost με 94.5% (Zheng et al, 2018) και ο NBTree με 91% (Mokryn et al, 2019).

Δεν ήταν όμως μόνο οι διαφορετικότητα των αλγορίθμων που επηρέασαν την απόδοση της κατηγοριοποίησης, αλλά και τα διαφορετικά δεδομένα που χρησιμοποιήθηκαν, τα οποία μπορούν να δείξουν διαφορές στις αποδόσεις. Η Bogina et al. (2016) έδειξε πως ο εμπλουτισμός των δεδομένων με επιπλέον προϊόντα μπορεί να βελτιώσει έστω και ελάχιστα την απόδοση. Επίσης, στις περισσότερες έρευνες προτιμάται να χρησιμοποιούνται τόσο στατικά όσο και δυναμικά δεδομένα. Ο Lee et al. (2015) στα αποτελέσματά του προτείνει πως η χρήση δυναμικών δεδομένων βελτιώνει αισθητά την πρόβλεψη σε σχέση με τη χρήση μόνο πληροφοριών στατικών αντικειμένων. Ένα από τα χαρακτηριστικά που είδαμε να μελετάτε από δύο άρθρα (Bogina et al, Mokryn et al.) ήταν αυτό της δημοτικότητας (trendiness) που έχουν συγκεκριμένα προϊόντα κάποια χρονική περίοδο. Αποδείχθηκε πως ο συνυπολογισμός αυτού του χαρακτηριστικού βελτιώνει την ακρίβεια του μοντέλου και πως οι πωλήσεις του εκάστοτε καταστήματος επηρεάζονται άμεσα από αυτό. Μια ακόμα διαφορά ανάμεσα στα δεδομένα των άρθρων είναι ο όγκος των δεδομένων που χρησιμοποιείται. Για παράδειγμα στο άρθρο του Lee et al. (2015) έγινε χρήση 76.375.439 δειγμάτων ενώ στο άρθρο του Bing et al. (2016) χρησιμοποιήθηκαν μόνο 15.000 δείγματα.

Μια ενδιαφέρουσα παρατήρηση είναι πως όλα τα άρθρα βασίζονται σε δεδομένα που συλλέγονται από τον browser και επικεντρώνονται στο σχηματισμό ενός μοντέλου που θα λειτουργούσε προφανώς με μεγαλύτερη ακρίβεια για χρήστες/ πελάτες κάποιας συσκευής υπολογιστή. Στην έρευνα του Guo et al. (2019) βλέπουμε πως λαμβάνονται υπόψη και δεδομένα που προέρχονται από συσκευές αφής, όπως είναι ο αριθμός των taps και των swipes που κάνει ο χρήστης. Ακόμα, κρίνεται σημαντικό να σημειωθεί πως τα δεδομένα προέρχονται από ηλεκτρονικά καταστήματα που πουλούν διαφορετικά είδη προϊόντων, όπως για παράδειγμα ρούχα, είδη σπιτιού, έπιπλα κτλ. Στο πρώτο άρθρο του Lee et al (2015) γίνεται ανάλυση σε ένα σύνολο δεδομένων που προέρχεται από κατάστημα παρόμοιου χαρακτήρα με το eBay.

Τέλος, παρατηρούμε πως για την ανάπτυξη των μοντέλων δίνεται μεγάλη έμφαση στην επιλογή των χαρακτηριστικών των δεδομένων και στην κατάλληλη εξεργασία τους, αφού σε όλες τις περιπτώσεις η κλάση με τους χρήστες που τελικά αγόρασαν κάποιο προϊόν από το ηλεκτρονικό κατάστημα περιέχει πολύ μικρότερο αριθμό δειγμάτων από την κλάση που περιέχει αυτούς που δεν αγόρασαν.

	Αριθμός δεδομένων	Τύπος δεδομένων	SVM	Naive Bayes	RF	LR	C4.5	MLP	REPTree	NBTree	DIPN	XGBoost
Lee et al.	76,375,439 συναλλαγές	Στατικά Αντικειμ.	Acc = 0,6 AUC = 0,67									
		Δυναμικά Συνεδρίας	Acc = 0,8 AUC=0,78									
		Στατικά+Δυναμικά	Acc = 0,8 AUC=0,79									
Niu et al.	6,944,274 συναλλαγές	Clickstream Στατικά + Δυναμικά			Acc = 0,76 Spec = 0,82	Acc = 0,61 Spec = 0,60						
Boroujerdi et al.	471,237 συναλλαγές	Στατικά Πελάτη + Δυναμικά Συνεδρίας			F1= 0,97	F1=0,94	F1=0,96	F1= 0,96	F1=0,97			
Guo et al.	9,200,000 click sessions	Δυναμικά Αφής + Περιήγησης									AUC = 0,84	
Bogina et al.	9,249,728 συναλλαγές	Clickstream							F1 = 0,80 AUC = 0,88	F1 = 0,75 AUC = 0,81		
		Clickstream + trendiness							F1 = 0,84 AUC = 0,92	F1 = 0,91 AUC = 0,94		
Zheng et al.	9,249,729 συνεδρίες	Δυναμικά Συνεδρίας				Acc = 0,94 F1 = 0,054						Acc = 0,91 F1 = 0,34
Bing et al.	15,000 συνεδρίες	Στατικά+ Δυναμικά		Acc ~ 0,70 t = 0,41 sec			Acc ~ 0,85 t = 0,36 sec					
Mokryn et al.	36,873,690 συνεδρίες	Clickstream Στατικά + Δυναμικά				F1 = 0,73				F1 = 0,88		F1= 0,76
		+ trediness				F1 = 73				F1 = 0,91		F1=0,78

Πίνακας Β1: Σύγκριση άρθρων που μελετήθηκαν

Κεφάλαιο Γ

Μεθοδολογία

Σε αυτή την παράγραφο περιγράφονται οι μετρικές αξιολόγησης που χρησιμοποιήθηκαν, ώστε να συγκριθούν οι διαφορετικοί αλγόριθμοι. Επιπλέον, περιγράφεται και το εργαλείο που χρησιμοποιήθηκε για την διεκπεραίωση του ερευνητικού μέρους της εργασίας.

Γ.1 Μετρικές Αξιολόγησης

Υπάρχουν διάφορες μετρικές για την αξιολόγηση και σύγκριση της απόδοσης των μοντέλων μηχανικής μάθησης σε μια δυαδική διαδικασία κατηγοριοποίησης. Αυτές οι μετρικές βασίζονται στον λεγόμενο πίνακα σύγχυσης (Confusion Matrix), από τον οποίο μπορούμε να αντλήσουμε τις σωστά προβλεπόμενες περιπτώσεις που ονομάζονται True Positive (TP) και True Negative (TN), όπως και τις μη σωστά προβλεπόμενες περιπτώσεις που ονομάζονται False Positive (FP) και False Negative (FN). Πιο αναλυτικά, ως

- *True Positive,* αναφερόμαστε στις παρατηρήσεις που είναι θετικές, και έχουν προβλεφθεί ως θετικές.
- *True Negative,* αναφερόμαστε στις παρατηρήσεις που είναι αρνητικές, και έχουν προβλεφθεί ως αρνητικές.
- *False Positive,* αναφερόμαστε στις παρατηρήσεις που είναι αρνητικές, αλλά έχουν προβλεφθεί ως θετικές.
- *False Negative,* αναφερόμαστε στις παρατηρήσεις που είναι θετικές, αλλά έχουν προβλεφθεί ως αρνητικές.

Όσο αφορά το θέμα που εξετάζουμε αναφερόμαστε στις περιπτώσεις όπου ο επισκέπτης δεν αγόρασε τίποτα και μια μη αγοραστική συνεδρία προέκυψε, και στις συνεδρίες όπου προέκυψε μια αγορά και αυτή προβλέφθηκε, ως TP και TN αντίστοιχα. Επίσης, ως FN είναι οι λανθασμένα προβλεπόμενες περιπτώσεις, και ως FP είναι οι περιπτώσεις όπου προέκυψε μια αγορά αλλά καμία δεν προβλέφθηκε ή όπου δεν έγινε αγορά αλλά προβλέφθηκε μια.

Από τον Confusion Matrix μπορούν να εξαχθούν διάφορες μετρικές απόδοσης. Οι πιο κοινές μετρικές είναι η ακρίβεια (Accuracy) και το σφάλμα (Error). Η ακρίβεια περιγράφει το ποσοστό των σωστών αποτελεσμάτων, ενώ το ποσοστό σφάλματος είναι ο αριθμός των

εσφαλμένων κατηγοριοποιημένων αποτελεσμάτων. Στόχος ενός μοντέλου κατηγοριοποίησης είναι να έχει υψηλό ποσοστό ακρίβειας και χαμηλό ποσοστό σφάλματος. Βέβαια, εδώ να σημειωθεί πως η ακρίβεια δεν είναι πάντα καλό μέτρο αξιολόγησης, ιδιαίτερα όταν έχουμε να κάνουμε με άνισα (imbalanced) δεδομένα. Καλύτερες μετρικές που παρέχουν περισσότερες πληροφορίες είναι η ευαισθησία (Sensitivity ή Recall) και η εξειδίκευση (Specificity), η οποία έρχεται σε αντίθεση με την ευαισθησία . Η ευαισθησία είναι ο αριθμός των true positive τιμών διαιρεμένος με όλες τις θετικές παρατηρήσεις ανάμεσα στα δεδομένα, και η εξειδίκευση μετράει την αναλογία των αρνητικών παρατηρήσεων που ταυτοποιήθηκαν σωστά. Μια ακόμα μετρική είναι η Precision, που ονομάζεται και θετική τιμή πρόβλεψης, είναι ο αριθμός των TP τιμών διά τον αριθμό όλων των θετικά κατηγοριοποιημένων περιπτώσεων. Στο F1 Score τόσο η ευαισθησία όσο και η precision συνυπολογίζονται εξίσου όπως φαίνεται στην εξίσωση 3.6. Παρακάτω δίνονται οι αντίστοιχες εξισώσεις.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (3.1)$$

$$\text{Error Rate} = \frac{FP+FN}{TP+FP+FN+TN} \quad (3.2)$$

$$\text{Sensitivity} = \frac{TP}{FN+TP} \quad (3.3)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (3.4)$$

$$\text{Precision} = \frac{TP}{FP+TP} \quad (3.5)$$

$$\text{F1 Score} = \frac{2}{\frac{FN+TP}{TP}+\frac{FP+TP}{TP}} \quad (3.6)$$

Δεν υπάρχει κάποιος συγκεκριμένος κανόνας ώστε να εκτιμήσουμε ποια από τις μετρικές είναι καλύτερη για κάποιο πρόβλημα κατηγοριοποίησης, αλλά αντιθέτως, εξαρτάται από την περίπτωση. Σε αυτή την εργασία θα χρησιμοποιήσουμε ως μετρικές αξιολόγησης την Ακρίβεια, την Ευαισθησία και την Εξειδίκευση.

Γ.2 Επιλογή Εργαλείου

Στην παρούσα εργασία το εργαλείο που χρησιμοποιήθηκε για το πειραματικό κομμάτι ήταν το software του Weka. Το Weka είναι ένα δωρεάν, μη εμπορικό, open source πρόγραμμα βασισμένο στην γλώσσα προγραμματισμού JAVA. Αποτελεί ένα εύχρηστο και κατανοητό εργαλείο για την μηχανική μάθηση και την εξόρυξη δεδομένων. Το εργαλείο αναπτύχθηκε από το Πανεπιστήμιο του Waikato στη Νέα Ζηλανδία, και μπορεί κανείς να το αποκτήσει μέσω των σχετικών συνδέσμων που υπάρχουν στο διαδίκτυο. Ο πηγαίος κώδικας μπορεί να 'κατεβαστεί' από την επίσημη ιστοσελίδα του εργαλείου. Έχει εφαρμοστεί ευρέως στον τομέα της διδασκαλίας, της έρευνας, όσο και για εμπορικούς λόγους. Περιέχει μια πληθώρα ενσωματωμένων εργαλείων για βασικά καθήκοντα εκμάθησης μηχανών και επιπλέον παρέχει διαφανή πρόσβαση σε γνωστές εργαλειοθήκες όπως τα scikit-learn, R και Deeplearning4j. Το Weka υποστηρίζει διάφορες βασικές διεργασίες εξόρυξης δεδομένων, πιο συγκεκριμένα, προεπεξεργασία δεδομένων, ομαδοποίηση, ταξινόμηση, παλινδρόμηση, απεικόνιση, και την δυνατότητα επιλογής χαρακτηριστικών. Ένας από τους κύριους λόγους που επιλέχθηκε να χρησιμοποιηθεί σε αυτή την εργασία είναι η ευκολία στη χρήση του λόγω των γραφικών διεπαφής χρήστη.

Κεφάλαιο Δ

Δεδομένα

Δ.1 Περιγραφή άρθρου άντλησης δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν προέρχονται από την έρευνα του άρθρου " Real time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks" από τους Sakar, Polat, Katircioglu ,Kastro (2018). Σε αυτήν την εργασίας προτείνουν ένα σύστημα ανάλυση της συμπεριφοράς των ηλεκτρονικών επισκεπτών που αποτελείται από δύο μοντέλα που ταυτόχρονα προβλέπουν την πρόθεση αγορών του επισκέπτη και την πιθανότητα εγκατάλειψης του ιστότοπου. Στο πρώτο μοντέλο, προβλέπουν την πρόθεση αγοράς του επισκέπτη με τη χρήση συγκεκριμένων δεδομένων προβολής της σελίδας (pageview data) που παρακολουθούνται κατά την διάρκεια της επίσκεψης, μαζί με κάποιες συνεδρίες και πληροφορίες χρήστη. Τα χαρακτηριστικά που μαζεύτηκαν, έπειτα τροφοδοτούνται σε Random Forest, Support Vector Machines και σε Multilayer Perceptron. Στη συνέχεια, δοκιμάζουν την τεχνική του oversampling, ώστε να έχουν ένα ισάριθμο σύνολο δεδομένων μεταξύ των κλάσεων που υπάρχουν. Τα αποτελέσματα δείχνουν πως ο MLP έχει υψηλότερη ακρίβεια και F1 Score από τους Random Forest και SVM. Επίσης, παρατηρούν πως ο συνδυασμός χαρακτηριστικών που περιέχουν πληροφορίες για την συνεδρία, σε συνδυασμό με τα clickstream δεδομένα επιφέρουν πιο επιτυχή αποτελέσματα. Στο δεύτερο μοντέλο, εκπαιδεύουν ένα βραχυπρόθεσμο επαναλαμβανόμενο νευρωνικό δίκτυο που δείχνει την προβλεπόμενη πιθανότητα της πρόθεσης ενός επισκέπτη να αφήσει την σελίδα χωρίς να τερματίσει την συναλλαγή του.

Σε αυτό το κεφάλαιο, λοιπόν, θα περιγραφούν τα δεδομένα που χρησιμοποιήθηκαν για την δημιουργία των διαφορετικών μοντέλων, καθώς και οι μέθοδοι που χρησιμοποιήθηκαν για την επιλογή χαρακτηριστικών που έγινε.

Δ.2 Περιγραφή Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν, σχημάτισαν ένα δυαδικό πρόβλημα κατηγοριοποίησης, το οποίο υπολογίζει την πρόθεση του χρήστη να ολοκληρώσει την συναλλαγή. Χρησιμοποιήθηκαν τόσο αριθμητικά, όσο και κατηγορηματικά χαρακτηριστικά. Στο σύνολο των δεδομένων ανήκουν 12.330 συνεδρίες, όπου κάθε συνεδρία ανήκει σε έναν διαφορετικό χρήστη κατά την περίοδο ενός έτους. Από τις 12.330 συνεδρίες το 84,5% (10.422) ήταν αρνητικές

παρατηρήσεις, δηλαδή δεν κατέληξαν σε αγορά, και οι υπόλοιπες 1.908 ήταν θετικές, δηλαδή κατέληξαν σε αγορά. Στους παρακάτω πίνακες φαίνονται τα χαρακτηριστικά που χρησιμοποιήθηκαν και οι περιγραφή τους. Στον πίνακαΔ1 τα χαρακτηριστικά 1 – 6 αναπαριστούν τους τύπους των σελίδων που επιστέφτηκε ο χρήστης στο κατάστημα και τον χρόνο που πέρασε σε αυτές. Τα χαρακτηριστικά 7 – 9 αναπαριστούν μετρικές που προέρχονται από το "Google Analytics". Το Bounce Rate αναφέρεται στο ποσοστό των επισκεπτών που εισέρχονται στη σελίδα και φεύγουν χωρίς να πυροδοτήσουν κανένα άλλο request στον server κατά την διάρκεια της συνεδρίας. Το Exit Value που ασχολείται με τον συνολικό αριθμό εξόδων από μια σελίδα σε σχέση μετ ον συνολικό αριθμό προβολών της σελίδας. Το Page Value παρουσιάζει την μέση τιμή για μια σελίδα που ο χρήστης επισκέφτηκε πριν ολοκληρώσει μια συναλλαγή. Ως Special Day αναφερόμαστε σε μέρες που είναι κοντά σε γιορτές, οπότε μπορεί να επηρεαστούν οι αγορές, όπως είναι ο Άγιος Βαλεντίνος ή η ημέρα της Μητέρας. Στις κοντινότερες μέρες σε τέτοιος ημερομηνίες η τιμή είναι 1. Τέλος, Revenue είναι η τιμή που αντιπροσωπεύει την κλάση. Με TRUE αναφερόμαστε στις συνεδρίες που κατέληξαν σε αγορά και με FALSE σε όσες δεν αγοράστηκε κάποιο προϊόν.

Όνομα Χαρακτηριστικού	Περιγραφή Χαρακτηριστικού	Εύρος τιμής
1. Administrative	Αριθμός σελίδων που επισκέφτηκε ο χρήστης για την διαχείριση λογαριασμού	0 - 27
2. Administrative duration	Συνολικός χρόνος σε sec που πέρασε ο χρήστης για την διαχείριση λογαριασμού	0 -3398
3. Informational	Συνολικός αριθμός σελίδων που επισκέφτηκε ο χρήστης σχετικές με στοιχεία επικοινωνίας, διεύθυνσης καταστήματος κτλ	0 -24
4. Informational duration	Συνολικός χρόνος σε sec που πέρασε ο χρήστης σε σελίδες πληροφοριών	0 -2549
5. Product related	Αριθμός σελίδων που επισκέφτηκε, σχετικές με 'σχετικά προϊόντα'	0 - 705
6. Product related duration	Συνολικός χρόνος σε sec που πέρασε σε σελίδες με 'σχετικά προϊόντα'	0 – 63.973
7. Bounce rate	Μέση τιμή του Bounce rate* των σελίδων που επισκέφτηκε	0 – 0,2
8. Exit rate	Μέση τιμή του Exit rate* των σελίδων που επισκέφτηκε	0 – 0,2
9. Page Value	Μέση τιμή του Page Value* των σελίδων που επισκέφτηκε	0 - 361
10. Special Day	Εγγύτητα της χρονικής περιόδου επίσκεψης της σελίδας σε 'ιδιαίτερη μέρα'	0 ή 1

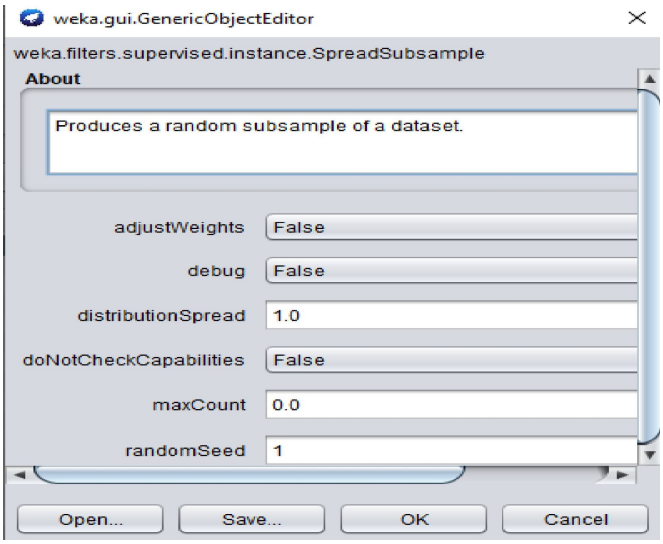
Πίνακας Δ1: Αριθμητικά Χαρακτηριστικά

Όνομα Χαρακτηριστικού	Περιγραφή Χαρακτηριστικού
11. Operating Systems	Λειτουργικό σύστημα χρήστη
12. Browser	Browser χρήστη
13. Region	Γεωγραφική περιοχή που ξεκίνησε η συνεδρία
14. Traffic Type	Πηγή από όπου προήρθε ο χρήστης πχ banner, SMS, απευθείας
15. Visitor Type	Τύπος επισκέπτη: ‘New’, ‘Returning’, ‘Other’
16. Weekend	Boolean μεταβλητή για τον αν είναι σαββατοκύριακο
17. Month	Τιμή μήνα επίσκεψης (1 - 12)
18. Revenue	Τιμή κλάσης που δείχνει πως κατέληξε η συνεδρία

Πίνακας Δ2: Ονομαστικά Χαρακτηριστικά

Δ.3 Επεξεργασία Δεδομένων

Στο σύνολο των δεδομένων που χρησιμοποιήθηκε δεν υπήρχε κάποια τιμή που να λείπει, οπότε δεν χρειάστηκε να συμπληρωθεί ή να υπολογιστεί κάποια. Όμως, όπως αναφέρθηκε και στην προηγούμενη παράγραφο ο αριθμός των παρατηρήσεων που ανήκε σε κάθε κλάση είχε μεγάλη διαφορά, κάνοντας έτσι το σύνολο των δεδομένων αρκετά άνισο (imbalanced). Αυτό μπορεί να οδηγήσει στην υπερπροσαρμογή (overfitting) του μοντέλου. Το μοντέλο κατά την εκπαίδευση θα μάθει να κατηγοριοποιεί καλύτερα τις παρατηρήσεις της αρνητικής κλάσης και ως συνέπεια να αποτυγχάνει στην σωστή κατηγοριοποίηση των παρατηρήσεων της θετικής κλάσης. Αφού, έγινε η παρατήρηση των αποτελεσμάτων για το imbalanced dataset, θα γίνει και η αξιολόγηση για το ισότιμο σύνολο δεδομένων (balanced dataset). Για να γίνει ο αριθμός των παρατηρήσεων των δύο κλάσεων ισάριθμος, θα χρησιμοποιηθεί η μέθοδος του undersampling. Δηλαδή, θα μειωθεί ο αριθμός των παρατηρήσεων της πλεονάζουσας κλάσης, ώστε και στις δύο κλάσεις να ανήκουν από 1908 παρατηρήσεις. Ώστε να επιλέξουμε τυχαία 1908 παρατηρήσεις της πλεονάζουσας κλάσης επιλέγουμε από από το tab Preprocess του Weka το φίλτρο SpreadSubsample και θέτουμε το distributionSpread = 1.0.



Εικόνα 3: Στιγμιότυπο από το Weka, για την εξισορρόπηση των κλάσεων

Δ.4 Επιλογή Χαρακτηριστικών

Κατά την επιλογή χαρακτηριστικών επιλέγουμε ένα υποσύνολο σχετικών χαρακτηριστικών για την χρήση τους στον σχηματισμό του μοντέλου. Αυτό πολλές φορές γίνεται ώστε να μειωθεί το φαινόμενο του overfitting, για να έχουμε πιο απλοποιημένα μοντέλα ή για να έχουμε μικρότερους χρόνους εκπαίδευσης. Η γενική ιδέα πίσω από την επιλογή χαρακτηριστικών είναι ότι τα δεδομένα εμπεριέχουν ορισμένα χαρακτηριστικά που είναι είτε περιττά είτε άσχετα, και έτσι μπορούν να αφαιρεθούν χωρίς να υπάρχει μεγάλη απώλεια πληροφοριών. Αποτελεί έναν συνδυασμό μιας τεχνικής αναζήτησης για τα νέα υποσύνολα, παράλληλα με μια μετρική αξιολόγησης για την εύρεση του ενός υποσυνόλου που ελαχιστοποιεί το error rate. Η επιλογή της μετρικής αξιολόγησης επηρεάζει αρκετά τον αλγόριθμο του μοντέλου. Σε αυτή την εργασία θα γίνει χρήση δύο κύριων κατηγοριών επιλογής χαρακτηριστικού.

Η πρώτη είναι η Wrapper. Αυτή η μέθοδος χρησιμοποιείται συνήθως όταν θέλουμε να χρησιμοποιήσουμε έναν συγκεκριμένο αλγόριθμο για την εκπαίδευση του μοντέλου. Η μέθοδος επιλέγει τα βέλτιστα χαρακτηριστικά για τον καθορισμένο αλγόριθμο. Βασίζεται συνήθως σε ‘άπληστους’ αλγόριθμους αναζήτησης, αφού αξιολογούν όλους τους πιθανούς συνδυασμούς των χαρακτηριστικών και επιλέγουν εκείνον που δίνει το καλύτερο αποτέλεσμα για τον συγκεκριμένο αλγόριθμο αναζήτησης. Η επιλογή γίνεται βάσει του error rate. Αυτό μπορεί να γίνει υπολογιστικά πολύ ακριβό, ειδικά όταν το σύνολο των χαρακτηριστικών είναι πολύ μεγάλο. Επειδή η μέθοδος wrapper βρίσκει το καλύτερο σύνολο χαρακτηριστικών για κάποιον συγκεκριμένο αλγόριθμο, το

μειονέκτημα είναι ότι αυτό το σύνολο χαρακτηριστικών μπορεί να μην είναι το βέλτιστο για κάποιον άλλο αλγόριθμο που θα χρησιμοποιήσουμε στο μοντέλο. Γι' αυτό και στην συνέχεια της εργασίας κάθε Wrapper θα χρησιμοποιεί ως κατηγοριοποιητή της δημιουργίας των υποσυνόλων των χαρακτηριστικών τον αντίστοιχο αλγόριθμο του μοντέλου. Πχ αν στο μοντέλο πρόκειται να χρησιμοποιηθεί ο Random Forest, τότε νωρίτερα στον Wrapper θα χρησιμοποιηθεί πάλι ο Random Forest για την εύρεση των κατάλληλων χαρακτηριστικών. Επίσης, θα χρησιμοποιηθεί ο Best First για την αξιολόγηση των υποσυνόλων.

Η δεύτερη κατηγορία είναι οι μέθοδοι φίλτρου (filter methods). Σε αντίθεση με τους Wrappers στη μέθοδο αυτή τα χαρακτηριστικά επιλέγονται ανεξάρτητα από τον αλγόριθμο που θα χρησιμοποιηθεί στο μοντέλο, αποτελώντας ένα από τα βασικά πλεονεκτήματα της. Επίσης, μπορεί να είναι πολύ γρήγορη σε σχέση με άλλες μεθόδους. Σε αυτές τις μεθόδους αντί για το error rate χρησιμοποιείται ένας 'μεσολαβητής'. Συχνοί μεσολαβητές είναι το Κέρδος Πληροφορίας (Information gain) και ο συντελεστής συσχέτισης του Pearson (correlation coefficient). Εκτός από την χρήση ενός αλγορίθμου εύρεσης για το βέλτιστο υποσύνολο, πολλές φορές χρησιμοποιείται η κατάταξη των χαρακτηριστικών από το πιο σημαντικό στο λιγότερο. Αυτή η μέθοδος αναζήτησης ονομάζεται Ranker. Έτσι, προκύπτει μια αξιολογημένη λίστα των χαρακτηριστικών από την οποία εμείς οι ίδιοι επιλέγουμε τα πιο σημαντικά.

Στην εργασία αυτή θα χρησιμοποιηθεί η μέθοδος επιλογής χαρακτηριστικών Wrapper με αλγόριθμο αναζήτησης τον Best First και η μέθοδος φίλτρου που χρησιμοποιεί τον συντελεστή συσχέτισης με αλγόριθμο αναζήτησης ξανά τον Best First. Επίσης, θα χρησιμοποιηθούν και δύο μέθοδοι φίλτρου που θα χρησιμοποιούν την μέθοδο Ranker. Από αυτές η μια θα χρησιμοποιεί τον συντελεστή συσχέτισης και η άλλη το κέρδος πληροφορίας.

Δ.4.1 Ο Best First ως μέθοδος αναζήτησης

Όπως αναφέρθηκε ο Best First θα είναι ο αλγόριθμος που θα χρησιμοποιηθεί ως μέθοδος αναζήτησης κατά την επιλογή των χαρακτηριστικών. Η αναζήτηση με Best First είναι ένα παράδειγμα αλγόριθμου αναζήτησης γράφου, στο οποίο ένας κόμβος επιλέγεται για επέκταση βάσει της συνάρτησης αξιολόγησης f(n). Τυπικά, ο κόμβος που έχει την χαμηλότερη αξία επιλέγεται για επέκταση επειδή μετριέται η απόσταση από τον στόχο. Μπορεί να αναπαρασταθεί ως μια ουρά προτεραιότητας. Συχνά αναφέρεται πως είναι 'άπληστος' αλγόριθμος γιατί επιλέγει γρήγορα την πιο επιθυμητή διαδρομή μόλις το ευρετικό (heuristic) βάρος της γίνει το πιο επιθυμητό. Τα βήματα που ακολουθεί είναι τα παρακάτω:

1. Πήγαινε στη ρίζα

2. Πήγαινε σε οποιοδήποτε γείτονα της ρίζας, που έχει την μικρότερη απόσταση από τον κόμβο της ρίζας και τοποθέτησε τους σε αύξουσα σειρά στην ουρά
3. Επανέλαβε αυτή την διαδικασία μέχρι να φτάσεις στον επιθυμητό στόχο.

Δ.4.2 Μέθοδοι επιλογής χαρακτηριστικών στο Weka

Παρακάτω παρουσιάζονται οι μέθοδοι που χρησιμοποιήθηκαν για την επιλογή χαρακτηριστικών με τη χρήση του εργαλείου Weka σε αυτήν την εργασία.

1. ***CfsSubsetEval με Best First*:** Αξιολογεί την αξία ενός υποσυνόλου χαρακτηριστικών λαμβάνοντας υπόψη την ατομική ικανότητα πρόβλεψης κάθε χαρακτηριστικού καθώς και τον βαθμό απόλυσης μεταξύ τους. Προτιμούνται υποσύνολα χαρακτηριστικών που συσχετίζονται σε μεγάλο βαθμό με την κλάση ενώ έχουν χαμηλή διασύνδεση.

Μέσω του Weka διαλέγουμε το φίλτρο AttributeSelection όπου θέτουμε ως evaluator: CfsSubsetEval και search:BestFirst, κατά την προεπεξεργασία των δεδομένων. Να σημειωθεί πως εφαρμόζεται στα balanced δεδομένα, αφού αυτά είναι που πέρνουμε τα καλύτερα αποτελέσματα και δεν υπάρχει ο φόβος του overfitting, και τα χαρακτηριστικά που επιλέγονται είναι τρία (BounceRates, PageValues, SpecialDay). Τα χαρακτηριστικά που επιλέγονται με αυτή την μέθοδο είναι μόνο 3 από τα αρχικά 17. Αυτό θα μπορούσε να οδηγήσει σε υπεραπλούστευση του μοντέλου και έτσι να δημιουργηθεί το φαινόμενο του **underfitting**.

2. ***WrapperSubsetEval με Birst First**:* Αξιολογεί τα υποσύνολα των χαρακτηριστικών χρησιμοποιώντας τη μέθοδο wrapper. Χρησιμοποιείται cross validation για την εκτίμηση της ακρίβειας του μοντέλου μάθησης για κάθε υποσύνολο χαρακτηριστικών.

Μέσω του Weka διαλέγουμε το φίλτρο AttributeSelection όπου θέτουμε ως evaluator: WrapperSubsetEval και search:BestFirst. Στον evaluator επιλέγουμε ως κατηγοριοποιητή κάθε φορά τον ίδιο αλγόριθμο με αυτόν που θα χρησιμοποιήσουμε στην συνέχεια στο μοντέλο. Για παράδειγμα, αν στο μοντέλο χρησιμοποιήσουμε τον MLP, τότε και ως κατηγοριοποιητή στον evaluator θα θέσουμε πάλι τον MLP. Να σημειωθεί πως εφαρμόζεται στα balanced δεδομένα, όπως εξηγήθηκε και προηγουμένως, και τα χαρακτηριστικά που επιλέγονται είναι διαφορετικά κάθε φορά, ανάλογα με τον αλγόριθμο που χρησιμοποιείται. Όπως φαίνεται στον παρακάτω πίνακα τα περισσότερα χαρακτηριστικά επιλέγονται για τον Random Forest (11), ενώ για τους υπόλοιπους αλγόριθμους δύο ή τρία χαρακτηριστικά. Η μοναδική περίπτωση που ένα χαρακτηριστικό επιλέγεται είναι αυτή του J48, και γι' αυτό δεν μπορούμε να είμαστε σίγουροι για την εγκυρότητα του μοντέλου που σχηματίζει. Παρόμοια, δεν μπορούμε να είμαστε σίγουρη για τον σχηματισμό

των υπόλοιπων μοντέλων, όπου τα χαρακτηριστικά που επιλέγονται είναι ελάχιστα, αφού και εδώ υπάρχει ο φόβος του underfitting. Η εξαίρεση, βέβαια, είναι ο Random Forest, όπου επιλέγονται 11 από τα 17 συνολικά χαρακτηριστικά.

J48	***Random Forest***	***Naive Bayes***	***K - NN***	***MLP***
• PageValues	• 11 features	• ProductRelated – Duration • BounceRates • PageValues	• PageValues • SpecialDay	• PageValues • BounceRates

Πίνακας Δ3: Χαρακτηριστικά που επιλέγονται με wrapper

3. ***InfoGainAttributeEval με Ranker***: Υπολογίζει το Κέρδος Πληροφορίας (ή αλλιώς εντροπία) για κάθε χαρακτηριστικό για την μεταβλητή εξόδου. Οι τιμές εισόδου κυμαίνονται από 0 (καμία πληροφορία) έως 1 (μέγιστη πληροφορία). Εκείνα τα χαρακτηριστικά που συμβάλλουν περισσότερο στην πληροφορία θα έχουν υψηλότερη τιμή εντροπίας και θα μπορούν να επιλεγούν, ενώ αυτά που δεν προσφέρουν αρκετή πληροφορία θα έχουν χαμηλότερη τιμή εντροπίας και θα μπορούν να αφαιρεθούν.

Μέσω του tab Select attributes διαλέγουμε ως evaluator: InfoGainAttributeEval και ως search method: Ranker. Όπως εξηγήθηκε θα επιλέξουμε τα χαρακτηριστικά που συνεισφέρουν περισσότερο στο τελικό αποτέλεσμα. Γι' αυτό θα επιλέξουμε τις τιμές που είναι πάνω από το 0.05. Έτσι, θα μείνουν 5 χαρακτηριστικά που θα επιλέξω, τα οποία είναι: Page Values, Exit Rates, Product Related Duration, Bounce Rates, Product Related. (Εικόνα 4) Ο αριθμός των χαρακτηριστικών που επιλέγονται ούτε σε αυτή την περίπτωση δεν είναι μεγάλος.

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 18 Revenue):
        Information Gain Ranking Filter

Ranked attributes:
 0.40224    9 PageValues
 0.12235    8 ExitRates
 0.08318    6 ProductRelated_Duration
 0.07507    7 BounceRates
 0.06656    5 ProductRelated
 0.04499    1 Administrative
 0.04437   11 Month
 0.04271    2 Administrative_Duration
 0.04128   15 TrafficType
 0.01841   10 SpecialDay
 0.01528    3 Informational
 0.01491    4 Informational_Duration
 0.0123    16 VisitorType
 0.00835   12 OperatingSystems
 0.00178   17 Weekend
 0         13 Browser
 0         14 Region
```

Εικόνα 4: Αποτελέσματα των βαθμολογημένων χαρακτηριστικών με τον Ranker_InfoGain

4. ***CorrelationAttributeEval με Ranker****:* Υπολογίζει την συσχέτιση (αναφέρεται ως ο συντελεσ΄της συσχέτισης του Pearson) μεταξύ κάθε χαρακτηριστικού και της μεταβλητής εξόδου και επιλέγει μόνο εκείνα τα χαρακτηριστικά που έχουν μέτρια προς υψηλή θετική ή αρνητική συσχέτιση, δηλαδή κοντά στο -1 ή στο 1, και αφαιρεί εκείνα με χαμηλή συσχέτιση, δηλαδή τιμή κοντά στο μηδέν.

Μέσω του tab Select attributes διαλέγουμε ως evaluator: CorrelationAttributeEval και ως search method: Ranker. Όπως εξηγήθηκε στο προηγούμενο κεφάλαιο θα επιλέξουμε τα χαρακτηριστικά που συνεισφέρουν περισσότερο στο τελικό αποτέλεσμα. Γι' αυτό θα επιλέξουμε τις τιμές που είναι πιο κοντά στη μονάδα. Με αυτόν τον τρόπο από τα 17 χαρακτηριστικά που έχουμε θα αποκλείσουμε τα 6 τελευταία (Εικόνα 5) τα οποία είναι: Information Duration, Weekend, Region, Operating System, Browser, Traffic type, και θα συνεχίσουμε με τα 11 εναπομείναντα. Ο αριθμός των χαρακτηριστικών που επιλέγονται από αυτή την μέθοδο φαίνεται να είναι και ο πιο ικανοποιητικός ανάμεσά τους.

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 18 Revenue):
        Correlation Ranking Filter
Ranked attributes:
 0.4456   9 PageValues
 0.3451   8 ExitRates
 0.2621   7 BounceRates
 0.1943   6 ProductRelated_Duration
 0.1814   5 ProductRelated
 0.1769   1 Administrative
 0.1455  10 SpecialDay
 0.1386   2 Administrative_Duration
 0.1255  16 VisitorType
 0.1202   3 Informational
 0.1167  11 Month
 0.0876   4 Informational_Duration
 0.0497  17 Weekend
 0.027   14 Region
 0.0231  12 OperatingSystems
 0.0228  13 Browser
 0.0165  15 TrafficType
```

Εικόνα 5: Αποτελέσματα των βαθμολογημένων χαρακτηριστικών με τον Ranker_Correlation

Στο παρακάτω κεφάλαιο παρουσιάζεται η εφαρμογή αυτών των τεχνικών για την επιλογή χαρακτηριστικών. Ύστερα, γίνεται δοκιμή των συνόλων των χαρακτηριστικών που δίνουν με τα μοντέλα των αλγορίθμων που μελετούνται.

Κεφάλαιο Ε

Αποτελέσματα Απόδοσης αλγορίθμων

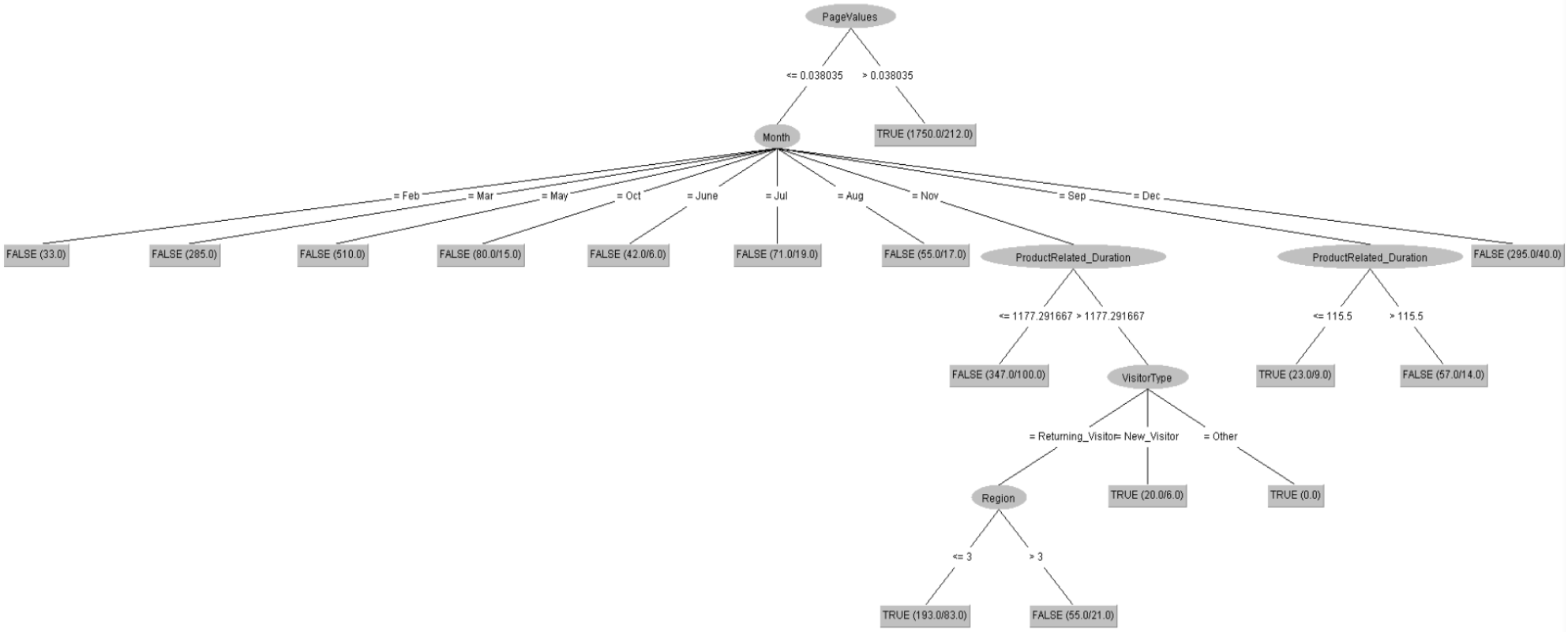
Στις παρακάτω παραγράφους θα αναλυθούν τα αποτελέσματα των αλγορίθμων τόσο για τα imbalanced δεδομένα, όσο και για τα balanced. Θα εντοπίσουμε ποιοι αλγόριθμοι σε κάθε περίπτωση έχουν καλύτερη απόδοση, και θα προχωρήσουμε στην επιλογή χαρακτηριστικών μέσω διαφορετικών τεχνικών. Τέλος, θα αξιολογήσουμε ποια τεχνική αποδίδει καλύτερα και θα συγκρίνουμε τα αποτελέσματα που πήραμε.

Ε.1 Ανάλυση αποτελεσμάτων για imbalanced dataset

Ο πρώτος αλγόριθμος που θα δοκιμαστεί είναι ο C4.5, ο οποίος στο Weka ονομάζεται **J48** και αποτελεί μια open source εκτέλεση σε Java του αλγορίθμου. Πηγαίνοντας στο tab Classify του weka, επιλέγουμε τον αλγόριθμο J48 από τον φάκελο trees. Στο Weka η default τιμή του ελάχιστου αριθμού παραδειγμάτων ανά φύλλο (number of instances per leaf) ισούται με το 2. Εάν διατηρήσουμε αυτή την τιμή, λόγω του μεγάλου αριθμού των δεδομένων που έχουμε (12.330), το δέντρο που προκύπτει είναι πολύ μεγάλο, έχοντας 259 φύλλα και μέγεθος 492. Επίσης, το F1 Score του ισούται με 63,5. Μπορούμε να μειώσουμε το μέγεθος του δέντρου αυξάνοντας τον ελάχιστο αριθμό παραδειγμάτων ανά φύλλο. Οι τιμές που δοκιμάστηκαν ήταν οι 50, 100, 200 και 500. Όπως φαίνεται και στον πίνακα Ε1 όταν ο αριθμός είναι 200 σημειώνεται η υψηλότερη Ευαισθησία (62,5%) και το υψηλότερο F1 Score (65,5%). Γι΄ αυτό και θα επιλέξουμε να προχωρήσουμε με με αυτή την τιμή για τα imbalanced δεδομένα. Το δέντρο που προέκυψε παρουσιάζεται στην Εικόνα 6

J48 _ imbalanced						
NumObj	Acc	Sen	Spec	F1 Score	No of Leaves	Size
50	89,8	59,3	95,4	64,3	27	44
100	89,82	59,4	95,4	64,4	17	24
200	89,8	62,5	94,8	65,5	16	22
500	88,84	53,5	95,3	59,7	4	7

Πίνακας Ε1: Αποτελέσματα για J48 σε imbalanced data



Εικόνα 6: Το δέντρο που προκύπτει από τον J48

Ο επόμενος αλγόριθμος που θα εξεταστεί είναι ο **Random Forest**. Πηγαίνοντας στο tab Classify του weka, επιλέγουμε τον αλγόριθμο RandomForest από τον φάκελο trees. Παρακάτω φαίνεται ο Confusion Matrix που προέκυψε.

Confusion Matrix: Random Forest		
a	b	Classified as
1119	789	a = TRUE
414	10008	b = FALSE

Πίνακας Ε2: Confusion Matrix RF για imbalanced

Από τον παραπάνω πίνακα προκύπτουν πως η Ακρίβεια, η Ευαισθησία, η Εξειδίκευση και το F1 Score είναι 90,2%, 58,6%, 96% και 65%, αντίστοιχα.

Συνεχίζοντας, θα δοκιμάσουμε πόσο καλά αποδίδει ο **Naive Bayes**. Πηγαίνοντας στο tab Classify του weka, επιλέγουμε τον αλγόριθμο NaiveBayes από τον φάκελο bayes. Παρακάτω φαίνεται ο Confusion Matrix που προέκυψε.

Confusion Matrix: Naive Bayes		
a	b	Classified as
1271	637	a = TRUE
1632	8790	b = FALSE

Πίνακας Ε3: Confusion Matrix NB για imbalanced

Από τον παραπάνω πίνακα προκύπτουν πως η Ακρίβεια ισούται με 81,59%, η Ευαισθησία με 66,6%, η Εξειδίκευση με 84,3% και το F1 Score με 52,8%.

Ο τέταρτος αλγόριθμος που θα 'τρέξουμε' είναι ο **k-NN**. Πηγαίνοντας στο tab Classify του weka, επιλέγουμε τον αλγόριθμο IBk από τον φάκελο lazy. Όπως, έχει ειπωθεί και παραπάνω το k παίζει μεγάλο ρόλο στο αποτέλεσμα και μπορεί να το διαφοροποιήσει αρκετά, αν δεν έχει επιλεγεί σωστά. Γι΄ αυτό το λόγο θα δοκιμάζουμε τον αλγόριθμο για διαφορετικά k και θα επιλέξουμε αυτό που δίνει τα πιο ικανοποιητικά αποτελέσματα. Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα που παίρνουμε.

K Nearest Neighbour _ imbalanced				
k =	Accuracy	Sensitivity	Specificity	F1 Score
3	84,43	26,6	95	34,6
5	85,12	21,8	96,7	31,2
7	85,45	18,2	97,8	27,9
9	85,45	15,6	98,2	24,9

Πίνακας Ε4: Confusion Matrix KNN για imbalanced

Από τον παραπάνω πίνακα μπορούμε να παρατηρήσουμε ότι καθώς ο αριθμός του k αυξάνεται, τόσο μεγαλώνουν και τα ποσοστά από της Ακρίβειας και της Εξειδίκευσης. Παρόλα αυτά βλέπουμε, όμως, πως το ποσοστό της Ευαισθησίας, δηλαδή του ποσοστού σωστά ταξινομημένων ατόμων με πρόθεση αγοράς, και του F1 Score μειώνονται σημαντικά ενώ το k αυξάνεται, φτάνοντας τελικά σε 15.61% και σε 24,9%, αντίστοιχα. Έτσι, καταλήγουμε στο συμπέρασμα πως τα καλύτερα αποτελέσματα για τα imbalanced data τα παίρνουμε όταν το **k=3**.

Τέλος, ο αλγόριθμος που θα εξετάσουμε είναι ο **Multilayer Perceptron**. Πηγαίνοντας στο tab Classify του weka, επιλέγουμε τον αλγόριθμο Multilayer Perceptron από τον φάκελο functions. Παρακάτω φαίνεται ο Confusion Matrix που προέκυψε.

Confusion Matrix: MLP		
a	b	Classified as
1062	846	a = TRUE
544	9878	b = FALSE

Πίνακας Ε5: Confusion Matrix MLP για imbalanced

Από τον παραπάνω πίνακα προκύπτουν οι εξής τιμές μετρικών. Η Ακρίβεια είναι 88,7%, η Ευαισθησία 55,7%, η Εξειδίκευση 94,8% και το F1 Score 60,45.

Ε.1.1 Σύγκριση αλγορίθμων για imbalanced data

Στην προηγούμενη παράγραφο εξετάσαμε την απόδοση πέντε διαφορετικών αλγορίθμων για το imbalanced dataset που διαθέτουμε. Παρακάτω παρουσιάζεται ένας συγκριτικός πίνακας από τα αποτελέσματα που λάβαμε.

Results _ Imbalanced _ Dataset				
Αλγόριθμος	Accuracy	Sensitivity	Specificity	F1 Score
J48	89,8	62,5	94,8	65,5
Random Forest	90,2	58,6	96	65
Naive Bayes	81,59	66,6	84,3	52,8
3 - NN	84,4	26,6	95	34,6
Multilayer Perceptron	88,7	55,7	94,8	60,4

Πίνακας Ε6: Αποτελέσματα για imbalanced data

Όπως, παρατηρούμε από τον συγκριτικό πίνακα η Ακρίβεια που έχουν όλοι οι αλγόριθμοι είναι αρκετά ικανοποιητική, αφού όλοι ξεπερνούν το 80%, με υψηλότερο αυτό του Random Forest που αγγίζει το 90,2%. Για να δούμε όμως αν οι αλγόριθμοί μας έχουν όντως τόσο καλή απόδοση θα πρέπει να εξετάσουμε και τις υπόλοιπες μετρικές. Οι δύο μετρικές που θα θέλαμε ιδανικά να είναι υψηλές είναι η Ευαισθησία, αφού μας δείχνει το ποσοστό που αφορά τα θετικά αποτελέσματα, και το F1 Score. Όμως, όπως βλέπουμε η ευαισθησία είναι αρκετά χαμηλή, με το χαμηλότερο ποσοστό να αγγίζει το 26,6% για τον 3 – NN. Επίσης, το χαμηλότερο ποσοστό του F1 Score ανήκει και πάλι στον 3 – NN, όπου είναι μόνο 34,6%. Καταλαβαίνουμε, λοιπόν, πως ο αλγόριθμος των πλησιέστερων γειτόνων είναι αυτός με τα χειρότερα αποτελέσματα. Μια ακόμα παρατήρηση που μπορούμε να κάνουμε είναι πως σε αντίθεση με την ευαισθησία, η εξειδίκευση έχει αρκετά υψηλά ποσοστά. Αυτό προφανώς συμβαίνει γιατί τα θετικά ταξινομημένα παραδείγματα είναι μόνο 1908 , ενώ τα αρνητικά είναι πολύ περισσότερα, 10422. Λόγω αυτής της ανισότητας στα δεδομένα η ευαισθησία παρουσιάζεται μικρή και η εξειδίκευση υψηλή. Οπότε, αν και η ακρίβεια παρουσιάζεται υψηλή, όπως μας επιβεβαιώνει και το F1 Score, τα ποσοστά ακρίβειας που παίρνουμε είναι πλασματικά.

Τέλος, ο αλγόριθμος από όσους δοκιμάστηκαν, που θα μπορούσαμε να πούμε πως έχει τα πιο ικανοποιητικά αποτελέσματα, είναι ο **Naive Bayes**. Αν και η ακρίβειά του είναι η χαμηλότερη (81,59%), έχει το υψηλότερο ποσοστό ευαισθησίας (66,6%) και το χαμηλότερο εξειδίκευσης (84,3%).

Ε.2 Ανάλυση αποτελεσμάτων για balanced dataset

Όπως ειπώθηκε και στο προηγούμενο κεφάλαιο, τα δεδομένα ήταν σε μεγάλο βαθμό άνισα, Γι' αυτό και εφαρμόστηκε η τεχνική του undersampling, ώστε να εξισορροπηθεί ο αριθμός των δεδομένων μεταξύ των δύο κλάσεων. Έτσι, έχουμε για την θετική κλάση (TRUE) 1908 παραδείγματα, και για την αρνητική (FALSE) άλλα 1908.

Με αυτό το balanced dataset θα δοκιμάσουμε αρχικά τον C4.5 ή **J48**. Όπως και για τα άνισα δεδομένα, έτσι και εδώ θα δοκιμάσουμε να μειώσουμε το μέγεθος του δέντρου αυξάνοντας τον ελάχιστο αριθμό παραδειγμάτων ανά φύλλο. Οι διαφορετικές τιμές που θα ελέγξουμε αυτή την φορά είναι οι 10, 20, 50 και 100. Τα αποτελέσματα φαίνονται στον πινάκα που ακολουθεί.

J48 _ balanced						
NumObj	Accuracy	Sensitivity	Specificity	F1-Score	No of Leaves	Size
10	83,93	86	81,9	84,3	31	52
20	84,43	85,1	83,8	84,5	16	22
50	84,51	83,7	85,3	84,4	14	19
100	84,66	82,2	87,1	84,3	2	3

Πίνακας Ε7: Confusion Matrix J48 για balanced data

Μια αρχική παρατήρηση που θα πρέπει να γίνει είναι πως με ελάχιστο αριθμό παραδειγμάτων ανά φύλλο ίσο με 100, το δέντρο που σχηματίζεται έχει 2 φύλλα και μέγεθος 3, γεγονός που δεν καθιστά αυτή την επιλογή καλή. Για τις υπόλοιπες τιμές βλέπουμε ότι η Ακρίβεια αυξάνεται έστω και ελάχιστα όσο αυξάνουμε το minNumObj, όπως και η Εξειδίκευση αυξάνεται σταδιακά. Όσο αφορά την Ευαισθησία, τα ποσοστά της μειώνονται. Το F1 Score έχει μικρές διακυμάνσεις της κλάσης του 0,1. Εκεί που είναι υψηλότερο είναι για τον ελάχιστο αριθμό παραδειγμάτων ανά φύλλο ίσο με 20, όπου αγγίζει το 84,5. Γι΄ αυτό το λόγο, καθώς και επειδή οι τιμές της ακρίβειας και της ευαισθησίας είναι αρκετά υψηλές (84,43% και 85,1%, αντίστοιχα), θα επιλέξουμε να προχωρήσουμε με την τιμή 20, όπου ο αριθμός των φύλλων του δέντρου ισούται με 16 και το μέγεθός του με 22. Αξιοσημείωτο είναι πως το δέντρο που προκύπτει είναι το ίδιο με αυτό για τα imbalanced δεδομένα, και φαίνεται στην Εικόνα 4.

Ο επόμενος αλγόριθμος που θα εξετάσουμε είναι ο **Random Forest**. Παρακάτω φαίνεται ο Confusion Matrix που προέκυψε.

Confusion Matrix: Random Forest		
a	b	Classified as
1613	295	a = TRUE
308	1600	b = FALSE

Πίνακας Ε8: Confusion Matrix RF για balanced

Από τον παραπάνω πίνακα προκύπτουν πως η Ακρίβεια, η Ευαισθησία, η Εξειδίκευση και το F1 Score είναι 84,2%, 84,5%, 83,9% και 84,3%, αντίστοιχα.

Ο τρίτος αλγόριθμος που θα δοκιμάσουμε είναι ο **Naive Bayes**. Παρακάτω φαίνεται ο Confusion Matrix που προέκυψε.

Confusion Matrix: Naive Bayes		
a	b	Classified as
1626	282	a = TRUE
623	1285	b = FALSE

Πίνακας Ε9: Confusion Matrix NB για balanced

Από τον παραπάνω πίνακα προκύπτουν πως η Ακρίβεια ισούται με 76,3%, η Ευαισθησία με 85,2%, η Εξειδίκευση με 67,3% και το F1 Score με 78,2%.

Ο επόμενος αλγόριθμος που θα τεστάρουμε είναι ο **k – NN**. Όπως και πριν θα δοκιμάσουμε διαφορετικά k και θα επιλέξουμε τελικά αυτό που θα μας δώσει τα καλύτερα αποτελέσματα. Τα k που θα δοκιμαστούν είναι τα 3, 5, 7 και 9, όπως και στα imbalanced δεδομένα. Παρακάτω φαίνεται ο πίνακας με τα συγκριτικά αποτελέσματα.

K Nearest Neighbour _ balanced				
k =	Accuracy	Sensitivity	Specificity	F1 Score
3	65,72	67,5	64	66,3
5	66,27	69	63,5	67,2
7	66,77	70,9	62,6	68,1
9	66,98	71,2	62,7	68,3

Πίνακας Ε10: Confusion Matrix KNN για balanced

Από τον παραπάνω πίνακα παρατηρούμε πως όσο αυξάνεται ο αριθμός του k, τόσο αυξάνονται και οι τιμές της Ακρίβειας, της Ευαισθησίας και του F1 Score. Δηλαδή, αυτών τον τιμών που επιθυμούμε να είναι υψηλές. Σε αντίθεση, το ποσοστό της Εξειδίκευσης ελαττώνεται, χωρίς όμως μεγάλες διαφορές. Οπότε, η τιμή του k που φαίνεται να δίνει τα πιο ικανοποιητικά αποτελέσματα είναι αυτή του **9**, και με αυτή θα επιλέξουμε να συνεχίσουμε παρακάτω.

Ο τελευταίος αλγόριθμος που θα εξετάσουμε είναι ο **Multilayer Perceptron**. Παρακάτω φαίνεται ο Confusion Matrix που προέκυψε.

Confusion Matrix: MLP		
a	b	Classified as
1558	350	a = TRUE
452	1456	b = FALSE

Πίνακας Ε11: Confusion Matrix MLP για balanced

Από τον παραπάνω πίνακα προκύπτουν οι εξής τιμές μετρικών. Η Ακρίβεια είναι 79%, η Ευαισθησία 81,7%, η Εξειδίκευση 76,3% και το F1 Score 79,5.

Ε.2.1 Σύγκριση αλγορίθμων για balanced data

Στην προηγούμενη παράγραφο εξετάσαμε την απόδοση πέντε διαφορετικών αλγορίθμων για το balanced dataset που δημιουργήσαμε. Παρακάτω παρουσιάζεται ένας συγκριτικός πίνακας από τα αποτελέσματα που λάβαμε για τα balanced δεδομένα.

Results _ Balanced _ Dataset				
Αλγόριθμος	Accuracy	Sensitivity	Specificity	F1 Score
J48	84,43	85,1	83,8	84,5
Random Forest	84,2	84,5	83,9	84,3
Naive Bayes	76,3	85,2	67,3	78,2
9- NN	66,98	71,2	62,7	68,3
Multilayer Perceptron	79	81,7	76,3	79,5

Πίνακας Ε12: Αποτελέσματα για imbalanced data

Όπως μπορούμε να δούμε την καλύτερη επίδοση έχει ο J48, αφού όλα τα ποσοστά των μετρικών είναι υψηλότερα του 80%. Πολύ κοντά στα αποτελέσματα που έδωσε ο J48 είναι και ο Random Forest, με εξίσου υψηλά ποσοστά. Ακολουθεί ο Multilayer Perceptron με F1 Score 79,5% και ο Naive Bayes με F1 Score 78,2%. Εδώ όμως αξίζει να σημειωθεί πως, αν και ο Multilayer Perceptron έχει υψηλότερο F1 Score, ο Naive Bayes είναι αυτός που έχει το υψηλότερο ποσοστό ευαισθησίας (85,2%) σε σχέση με όλους τους υπόλοιπους αλγόριθμους που δοκιμάστηκαν. Ο αλγόριθμος με τα χειρότερα αποτελέσματα, όπως φαίνεται και από τον πίνακα, είναι ο k – πλησιέστερος γείτονας.

Ε.3 Σύγκριση αποτελεσμάτων μεταξύ balanced και imbalanced datasets

Αρχικά, μπορούμε να παρατηρήσουμε πως και για τα δυο dataset τα δέντρα απόφασης είναι αυτά που δίνουν τα καλύτερα αποτελέσματα. Ο J48 και ο Random Forest και στις δύο περιπτώσεις μας δίνουν ικανοποιητικά ποσοστά ακρίβειας και F1 Score. Επίσης, τα χειρότερα αποτελέσματα και στα δύο σετ δεδομένων τα δίνει ο k –πλησιέστερος γείτονας, όπου η ευαισθησία και το F1 Score είναι τα χαμηλότερα ποσοστά μεταξύ των υπόλοιπων αλγορίθμων. Συγκεκριμένα, για τα imbalanced data το F1 Score αγγίζει μόνο το 34.6%. Επιπλέον, οι Multilayer Perceptron και ο Naive Bayes είναι αυτοί που και στις δύο περιπτώσεις μας δίνουν μέτριες προς καλές τιμές

μετρικών. Αξιοσημείωτο είναι πως και στις δύο περιπτώσεις ο Naive Bayes είναι αυτός που έχει το υψηλότερο ποσοστό ευαισθησίας.

Συμπερασματικά, αν και η ακρίβεια και η εξειδίκευση που δίνουν οι αλγόριθμοι για τα imbalanced δεδομένα είναι υψηλές, τα F1 Score που παίρνουμε δεν είναι ικανοποιητικά, αφού δεν ξεπερνούν σε καμία περίπτωση το 65%, γεγονός που οφείλεται στον χαμηλό αριθμό αρνητικών παρατηρήσεων σε σύγκριση με τις θετικές. Τέλος, για τα balanced δεδομένα οι τιμές των μετρικών είναι αρκετά καλές και ισορροπημένες μεταξύ τους για κάθε αλγόριθμο. Τα ποσοστά της ακρίβειας και του F1 Score ξεπερνούν το 80%, με μόνη εξαίρεση τον k–NN, όπου τα συγκεκριμένα ποσοστά ξεπερνούν το 66%.

Ε.4 Αποτελέσματα αλγορίθμων με επιλογή χαρακτηριστικών

Στην παράγραφο Δ.4.3 περιγράφτηκαν ποιες μέθοδοι επιλογής χαρακτηριστικών χρησιμοποιήθηκαν, καθώς και ποια χαρακτηριστικά επιλέχτηκαν από κάθε μια από αυτές. Στις παρακάτω παραγράφους θα αναλυθούν τα αποτελέσματα που προέκυψαν από την εφαρμογή τους.

Ε.4.1 Επιλογή υποσυνόλου χαρακτηριστικών βάσει συσχέτισης

Παρακάτω υπάρχει ο πίνακας , στον οποίο φαίνονται οι τιμές των μετρικών που λάβαμε για κάθε αλγόριθμο, χρησιμοποιώντας τα χαρακτηριστικά που προέκυψαν με τη χρήση της συγκεκριμένης μεθόδου.

Results _ CfsSubsetEval				
Αλγόριθμος	Accuracy	Sensitivity	Specificity	F1 Score
J48	84,72	80,6	88,8	84,1
Random Forest	81,55	78	85,1	80,9
Naive Bayes	64,59	92,5	36,7	72,3
9- NN	83,25	77,1	89,4	82,2
Multilayer Perceptron	80,92	77,8	84	80,3

Πίνακας Ε13: Αποτελέσματα με μέθοδο CfsSubsetEval

Όπως μπορούμε να παρατηρήσουμε ο αλγόριθμος με τα καλύτερα αποτελέσματα είναι ο J48 με F1 Score 84,1%. Δεύτερος ακολουθεί ο K – NN και έπειτα οι Random Forest, Multilayer Perceptron και ο Naive Bayes. Αξιοσημείωτο είναι πως πριν την επιλογή των συγκεκριμένων

χαρακτηριστικών, ο K – NN ήταν αυτός με τα χαμηλότερα ποσοστά μετρικών (βλ. Πίνακας E12), ενώ με την εφαρμογή του CfsSubsetEval, τα ποσοστά αυξάνονται αρκετά. Η ακρίβεια και το F1 Score αυξάνονται περίπου 15%. Επίσης, βλέπουμε πως για τον Naive Bayes η ευαισθησία αυξάνεται στο 92,5%, που είναι και το μεγαλύτερο ποσοστό ανάμεσα στους αλγορίθμους, ενώ η εξειδίκευση ελαττώνεται υπερβολικά στο 36,7%.

Γενικά, παρατηρούμε ότι μετά την επιλογή των χαρακτηριστικών βάσει συσχέτισης, η ακρίβεια των αλγορίθμων αυξάνεται ελάχιστα, ενώ το F1 Score ελαττώνεται, με εξαίρεση όμως τον K – NN, όπου από τελευταίος (χωρίς την εφαρμογή επιλογής χαρακτηριστικών), έρχεται δεύτερος. Τα αποτελέσματα που πήραμε δεν ήταν άσχημα, αφού η ακρίβεια και το F1 Score ξεπερνούν στην πλειοψηφία το 80%, δεν είναι όμως καλύτερα από αυτά της χρήσης όλων των χαρακτηριστικών (Βλ. Πίνακα). Αυτό, όμως, που είναι ανησυχητικό είναι πως τα χαρακτηριστικά που επιλέγονται με αυτή την μέθοδο είναι μόνο 3 από τα αρχικά 17. Αυτό θα μπορούσε να οδηγήσει σε υπεραπλούστευση του μοντέλου και έτσι να δημιουργηθεί το φαινόμενο του **underfitting**.

E.4.2 Επιλογή υποσυνόλου χαρακτηριστικών με Wrapper

Στον πίνακα που ακολουθεί βλέπουμε τα αποτελέσματα που παίρνουμε από την εφαρμογή της μεθόδου.

Results _ Wrapper				
Αλγόριθμος	Accuracy	Sensitivity	Specificity	F1 Score
J48	84,72	80,6	88,8	84,1
Random Forest	84,72	85,4	84	84,8
Naive Bayes	80,45	72,8	88,1	78,8
9- NN	84,46	79,2	89,7	83,6
Multilayer Perceptron	80,68	78,1	83,2	80,2

Πίνακας E14: Αποτελέσματα με μέθοδο Wrapper

Από τον παραπάνω πίνακα βλέπουμε πως τα αποτελέσματα που παίρνουμε από την χρήση των νέων χαρακτηριστικών είναι αρκετά ικανοποιητικά, αλλά όπως αναφέρθηκε και πιο πάνω δεν μπορούμε να είμαστε σίγουροι για την πραγματική τους επίδοση. Την καλύτερη επίδοση έχει ο Random Forest αγγίζοντας το 85,4% σε ευαισθησία και 84,72% ακρίβεια, ποσοστά αρκετά ικανοποιητικά με τη χρήση 11 χαρακτηριστικών. Ο Random Forest θα είναι και ο μόνος από τους

αλγόριθμους που θα ληφθεί υπόψη στην συνέχεια με τη χρήση της μεθόδου wrapper. Οι υπόλοιποι δεν μπορύν να ληφθούν υπόψη λόγω του μικρού αριθμού χαρακτηριστικών που επιλέγονται.

Ε.4.2 Επιλογή υποσυνόλου χαρακτηριστικών με Ranker – Κέρδους Πληροφορίας

Οι τιμές που επιλέχτηκαν, όπως αναφέρθηκε στο προηγούμενο κεφάλαιο είναι πάνω από το 0.05. Έτσι, θα μείνουν 5 χαρακτηριστικά που θα επιλέξω, τα οποία είναι: Page Values, Exit Rates, Product Related Duration, Bounce Rates, Product Related. (Εικόνα 4) Ο αριθμός των χαρακτηριστικών που επιλέγονται ούτε σε αυτή την περίπτωση δεν είναι μεγάλος, αλλά παρόλα αυτά θα γίνει μια σύγκριση των αποτελεσμάτων που παίρνουμε σε σχέση με τη χρήση όλων των χαρακτηριστικών.

Results _ InfoGain				
Αλγόριθμος	Accuracy	Sensitivity	Specificity	F1 Score
J48	84,61	80,7	88,5	84
Random Forest	83,22	82	84,5	83
Naive Bayes	70,83	89,7	51,9	75,5
9- NN	80,89	78	83,8	80,3
Multilayer Perceptron	80,87	79,2	82,5	80,5

Πίνακας Ε15: Αποτελέσματα με μέθοδο InfoGain

Από τον παραπάνω πίνακα μπορούμε να παρατηρήσουμε πως η επίδοση των αλγορίθμων είναι αρκετά καλή, αφού το F1 Score ξεπερνάει το 80%, με εξαίρεση τον Naive Bayes (75,5%). Επίσης, ο Naive Bayes φαίνεται να είναι ο μόνος που παρουσιάζει τόσο χαμηλή εξειδίκευση (51,9%) και την υψηλότερη ευαισθησία (89,7%). Σε σύγκριση με τα αποτελέσματα που είχαμε πριν την επιλογή χαρακτηριστικών η ευαισθησία και το F1 Score φαίνεται να ελαττώνονται σε μικρό βαθμό. Ο αλγόριθμος που παρουσιάζεται να έχει αρκετά καλύτερα ποσοστά με αυτή την επιλογή χαρακτηριστικών είναι ο k–NN, στον οποίο οι τιμές όλων των μετρικών αυξάνονται. Συγκεκριμένα, η ακρίβεια έχει αύξηση ~ 35%, η ευαισθησία ~ 7% και η εξειδίκευση ~ 20%.

Ε.4.2 Επιλογή υποσυνόλου χαρακτηριστικών με Ranker – Συσχέτισης

Όπως εξηγήθηκε στο προηγούμενο κεφάλαιο θα επιλέξουμε τα χαρακτηριστικά που συνεισφέρουν περισσότερο στο τελικό αποτέλεσμα. Γι’ αυτό θα επιλέξουμε τις τιμές που είναι πιο κοντά στη μονάδα. Με αυτόν τον τρόπο από τα 17 χαρακτηριστικά που έχουμε θα αποκλείσουμε τα 6

τελευταία (Εικόνα 5) τα οποία είναι: Information Duration, Weekend, Region, Operating System, Browser, Traffic type, και θα συνεχίσουμε με τα 11 εναπομείναντα. Ο αριθμός των χαρακτηριστικών που επιλέγονται από αυτή την μέθοδο φαίνεται να είναι και ο πιο ικανοποιητικός ανάμεσά τους. Για αυτό στην συνέχεια της εργασίας τα αποτελέσματα που περιγράφονται παρακάτω θα ληφθούν υπόψη.

Results _ Ranker_Correlation				
Αλγόριθμος	Accuracy	Sensitivity	Specificity	F1 Score
J48	84,74	85	84,5	84,8
Random Forest	84,59	85,1	84,1	84,7
Naive Bayes	75,36	87,7	63	78,1
9- NN	75,52	76,4	74,7	75,7
Multilayer Perceptron	81,21	82,7	79,7	81,5

Πίνακας Ε16: Αποτελέσματα με μέθοδο Ranker Correlation

Μπορούμε να παρατηρήσουμε πως αυτή η τεχνική επιλογής χαρακτηριστικών λειτουργεί αρκετά καλά. Τα αποτελέσματα που παίρνουμε είναι ικανοποιητικά, αφού η ευαισθησία και το F1 Score είναι αρκετά υψηλά. Σε σχέση με τα ποσοστά των μετρικών που είχαμε χωρίς την επιλογή χαρακτηριστικών, μπορούμε να πούμε πως τα καινούρια ποσοστά αυξάνονται αρκετά για την ευαισθησία και το F1 Score, ενώ η εξειδίκευση φαίνεται να μειώνεται, όχι όμως σε σημαντικό ποσοστό. Ο αλγόριθμος που σίγουρα λειτουργεί καλύτερα με την επιλογή των συγκεκριμένων χαρακτηριστικών είναι ο **Multilayer Perceptron**, αφού οι τιμές όλων των μετρικών αυξάνονται και αγγίζουν το 80%.

Ε.5 Σύγκριση αποτελεσμάτων με βιβλιογραφία

Η επιλογή των καλύτερων αποτελεσμάτων για κάθε αλγόριθμο που δοκιμάσαμε θα γίνει βάσει της ακρίβειας και του F1 Score. Η ακρίβεια χρησιμοποιείται συνήθως όταν οι TP και οι TN είναι σημαντικές, και το F1 Score όταν οι FN και οι FP είναι κρίσιμες. Όπως, είδαμε πιο πάνω με τη χρήση διαφορετικών χαρακτηριστικών ο κάθε αλγόριθμος είτε βελτιωνόταν είτε όχι, σε σύγκριση με την επίδοσή του με την χρήση και των 17 χαρακτηριστικών. Παρακάτω παρουσιάζεται ένας πίνακας με τα καλύτερα αποτελέσματα. Στη συνέχεια, αυτά τα αποτελέσματα θα τα συγκρίνουμε με τα αντίστοιχα της βιβλιογραφίας που αναλύθηκε.

Αλγόριθμος	Accuracy	F1 Score	Feature Selection
J48	84,74	84,8	Ranker Correlation
Random Forest	84,72	84,8	Wrapper
Naive Bayes	76,3	78,2	None
9 NN	75,52	75,7	Ranker Correlation
Multilayer Perceptron	81,21	81,5	Ranker Correlation

Πίνακας Ε17: Καλύτερα αποτελέσματα

Από τον παραπάνω πίνακα βλέπουμε πως τα αποτελέσματα που προέκυψαν είναι αρκετά ικανοποιητικά. Οι αλγόριθμοι που έχουν τα υψηλότερα ποσοστά είναι ο J48 και ο Random Forest. Στη συνέχεια, ο επόμενος αλγόριθμος με την καλύτερη επίδοση είναι ο MLP, αφού και αυτός ξεπερνά το 80% γαι την ακρίβεια και το F1 Score. Τέλος, ο Naive Bayes και ο 9NN έχουν καλά αποτελέσματα που ξεπερνούν το 75%. Αξιοσημείωτο είναι πως ο Naive Bayes είναι ο μοναδικός αλγόριθμος που σημείωσε τα καλύτερά του αποτελέσματα χωρίς την εφαρμογή κάποιας μεθόδου επιλογής χαρακτηριστικών.

	J48	**Random Forest**	**Naive Bayes**	**9 NN**	**Multilayer Perceptron**
Niu et al.		Acc = 0,76			
Boroujerdi et al.	F1 = 0,96	F1 = 0,97			F1 = 0,96
Bing et al.	Acc ~ 0,85		Acc ~ 0,70		
Sakar et al.	Acc = 0,823 F1 = 0,82	Acc = 0,822 F1 = 0,81			Acc = 0,87 F1 = 0,86
Καλύτερα αποτελέσματα εργασίας	Acc = 0,847 F1 = 0,848	Acc = 0,847 F1 = 0,848	Acc = 0,763	Acc = 0,75 F1 = 0,75	Acc = 0,812 F1 = 0,815

Πίνακας Ε18: Σύγκριση αποτελεσμάτων

Στον παραπάνω πίνακα φαίνονται τα αποτελέσματα σε σύγκριση με αυτά της βιβλιογραφίας που μελετήθηκε. Έχουν συμπεριληφθεί τα άρθρα που εφάρμοζαν τους ίδιους αλγόριθμους με αυτούς της παρούσας εργασίας. Όπως φαίνεται δεν βρέθηκε κάποιο που να μελετά την δημιουργία του μοντέλου πρόβλεψης με τον αλγόριθμο του k πλησιέστερου γείτονα. Επίσης, δεν μπορούμε να πούμε πως η σύγκριση των αποτελεσμάτων μεταξύ τους είναι απαραίτητα ακριβής, αφού κάθε άρθρο μελετά διαφορετικό αριθμό δεδομένων, διαφορετικά είδη δεδομένων και χρησιμοποιεί διαφορετικές τεχνικές επεξεργασίας και εκπαίδευσης του μοντέλου. Μικρές διαφορές παρατηρούνται και μεταξύ των αποτελεσμάτων της εργασίας και του άρθρου από το οποίο

δανείστηκαν τα δεδομένα. Αυτό πιθανόν να οφείλεται στο γεγονός, πως στο άρθρο χρησιμοποιείται το oversampling, ενώ στην εργασίας το undersampling, καθώς και πως στο άρθρο χρησιμοποιείται η επαλήθευση 80 – 20%, ενώ στην εργασία η 10 fold cross. Ακόμα, οι μέθοδοι επιλογής χαρακτηριστικών διαφέρουν.

Για τον J48 παρατηρούμε πως σε όλες τις έρευνες η ακρίβεια και το F1 Score είναι αρκετά υψηλά. Το ίδιο ισχύει και για τον Random Forest. Υψηλότερο φαίνεται να είναι το F1 Score του δεύτερου άρθρου που μελετήθηκε με 96% για τον J48 και 97% για τον Random Forest. Ο Naive Bayes στην παρούσα εργασία παρουσιάζει υψηλότερη ακρίβεια (76%) σε σχέση με το άρθρο 7. Τέλος, για τον Multilayer Perceptron τα αποτελέσματα που λήφθηκαν ήταν αρκετά καλά, αλλά χειρότερα σε σύγκριση με αυτά του μελετήθηκαν στην βιβλιογραφία.

Συμπεράσματα

Η πρόβλεψη της συμπεριφοράς των ηλεκτρονικών πελατών γίνεται ολοένα και πιο σημαντική καθώς ο κλάδος του ηλεκτρονικού εμπορίου εξελίσσεται. Όπως, παρατηρήθηκε από την βιβλιογραφική έρευνα αυτής της εργασίας το είδος των δεδομένων που θα συλλεχθούν παίζει σημαντικό ρόλο, αφού η χρήση δυναμικών δεδομένων βελτιώνει την επίδοση των μοντέλων. Επίσης, σημαντικός παράγοντας φαίνεται να είναι η επιλογή 'ιδιαίτερων' χαρακτηριστικών, όπως για παράδειγμα αν ένα προϊόν είναι δημοφιλές την συγκεκριμένη χρονική περίοδο ή αν η εποχή που μελετάμε είναι κοντά σε κάποια γιορτή πχ Άγιος Βαλεντίνος, Χριστούγεννα κτλ. Η επιλογή λοιπόν των χαρακτηριστικών μπορεί να επηρεάσει το αναπτυσσόμενο μοντέλο, και γι΄ αυτό η επιλογή τους πρέπει να γίνεται πολύ προσεκτικά.

Ένα από τα σημαντικότερα προβλήματα που συναντάται στον σχηματισμό τέτοιου είδους μοντέλων είναι πως οι κλάσεις που προκύπτουν είναι σε μεγάλο βαθμό άνισες. Σε όλες τις έρευνες που μελετήθηκαν η κλάση που ανήκουν οι συνεδρίες που κατέληξαν σε αγορά, περιέχει πολύ μικρό αριθμό παραδειγμάτων. Ενώ η κλάση που δηλώνει τις συνεδρίες που δεν κατέληξαν σε αγορά έχει μεγάλο αριθμό παραδειγμάτων. Οπότε, προκύπτει το πρόβλημα των imbalanced δεδομένων. Οι τεχνικές που χρησιμοποιούνται για την εξισορρόπηση των κλάσεων είναι το undersampling της κλάσης με τα περισσότερα δείγματα ή το oversampling της κλάσης με τα λιγότερα δείγματα ή ο συνδυασμός των δύο παραπάνω τεχνικών πχ SMOTE. Στην εργασία χρησιμοποιήθηκε η τεχνική του oversampling όπου ο συνολικός αριθμός δειγμάτων και από τις δύο κλάσεις είναι 3816.

Οι συνηθέστεροι αλγόριθμοι που χρησιμοποιούνται για την επίλυση ενός τέτοιου προβλήματος κατηγοριοποίησης είναι τα δέντρα απόφασης, τα νευρωνικά δίκτυα και αλγόριθμοι που χρησιμοποιούν το bagging. Στα πλαίσια της εργασίας μελετήθηκαν οι αλγόριθμοι J48, Random Forest, Naive Bayes, k – NN και Multilayer Perceptron. Μετά την δημιουργία ενός balanced dataset και της εφαρμογής διαφορετικών μεθόδων επιλογής δεδομένων, καταλήξαμε στην επιλογή των καλύτερων αποτελεσμάτων. Χρησιμοποιήθηκαν μέθοδοι επιλογής χαρακτηριστικών τόσο wrapper όσο και φίλτρου. Από τις μεθόδους φίλτρου δοκιμάστηκαν η επιλογή υποσυνόλου χαρακτηριστικών βάσει τη συσχέτιση και τη χρήση του Best First ως αλγόριθμο αναζήτησης, καθώς και με τη χρήση Ranker. Επίσης, με Ranker δοκιμάστηκε και η μέθοδος επιλογής υποσυνόλου βάσει του κέρδους πληροφορίας. Τελικά, η μέθοδος επιλογής χαρακτηριστικών που έδωσε τα καλύτερα αποτελέσματα και ικανοποιητικό αριθμό χαρακτηριστικών ήταν αυτή υου Ranker – Συσχέτισης. Ο αλγόριθμος που παρατηρήθηκε πως δεν είχε καλή επίδοση ήταν ο k – πλησιέστερος γείτονας, ιδιαίτερα στην περίπτωση όπου τα δεδομένα είναι imbalanced. Τα

αποτελέσματα ήταν αρκετά ικανοποιητικά, αφού στις περισσότερες περιπτώσεις η ακρίβεια και το F1 Score ξεπερνούν το 75%, με καλύτερα αυτά των δέντρων απόφασης με ακρίβεια 84,7%.

Συμπερασματικά, η μελέτη της συμπεριφοράς των χρηστών των ηλεκτρονικών καταστημάτων και τον αν έχουν την πρόθεση να προβούν σε αγορά είναι ένα ζήτημα που αφορά άμεσα τους ιδιοκτήτες, αφού αποτελεί έναν τρόπο για την αύξηση των πωλήσεων και της καλύτερης εξυπηρέτησης των πελατών τους. Οι μέθοδοι εξόρυξης δεδομένων και η μηχανική μάθηση μπορούν να προσφέρουν αποτελεσματικές λύσεις σε αυτό το αρκετά περίπλοκο πρόβλημα., που θα εξελίσσεται με την ανάπτυξη του ηλεκτρονικού εμπορίου.

Βιβλιογραφία – Ιστογραφία

[1] M. Lee, T. Ha, J. Han, J. Rha, T. Kwon, *Online Footsteps to Purchase: Exploring Consumer Behaviors on Online Shopping Sites*, WebSci'15, June 28–July 01, 2015

[2] X. Niu, C. Li, X. Yu, *Predictive Analytics of E-Commerce Search Behavior for Conversion,* Twenty-third Americas Conference on Information Systems, 2017

[3] E. Boroujerdi, S. Mehri, S. Garmaroudi, M. Pezeshki, F. Mehrabadi, S. Malakouti, S. Khadivi, *A Study on Prediction of User's Tendency Toward Purchases in Websites Based on Behavior Models,* 6th Conference on Information and Knowledge Technology, 2014

[4] L. Guo, L. Hua, R. Jia, B. Zhao, X. Wang, B. Cui, Buying or Browsing? : *Predicting Real-time Purchasing Intent uing Attention-based Deep Network with Multiple Behavior,* KDD '19, August 4–8, 2019

[5] V. Bogina, T. Kuflik, O. Mokryn, *Learning Item Temporal Dynamics for Predicting Buying*

Sessions, IUI'16, March 07-10, 2016

[6] B. Zheng, B. Liu, *A Scalable Purchase Intention Prediction System Using Extreme Gradient Boosting Machines with Browsing Content Entropy,* IEEE International Conference on Consumer Electronics, 2018

[7] L. Bing, S. Yuliang, *Prediction of User's Purchase Intention Based on Machine Learning*, 3rd International Conference on Soft Computing & Machine Intelligence, 2016

[8] O. Mokryn, V. Bogina, T. Kuflik, *Will this session end with a purchase? Inferring current purchase intent of anonymous visitors,* Elsevier B.V., 2019

[9] Ευστάθιος Γ. Κύρκος, *Επιχειρηματική Ευφυΐα & Εξόρυξη Δεδομένων*, ISBN: 978-960-603-109-0, 2015

[10] S.Neelamegam , Dr.E.Ramaraj, *Classification algorithm in Data mining: An Overview*, International Journal of P2P Network Trends and Technology (IJPTT) - Volume 3 Issue 5 September to October, 2013

[11] S. Yu, L. Xiong, *Research on ecosystem evolution of Mobile E-Commerce,* 10th International Symposium on Distributed Computing and Applications to Business, 2011

[12] Asha K N, R. Rajkumar, *Pre – processing of User behaviour for E – Commerce,* 978-1-5386-0569-1, IEEE, 2017

[13] *Behind the Buzz of Behavioral Data,* https://digitopoly.org/2015/05/14/behind-the-buzz-of-behavioral-data/, 2015

[14] C. Sakar, S. Polat, M. Katircioglu, Y. Kastro, *Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks,* The Natural Computing Applications Forum, Springer, 2018