ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
TECHNICAL UNIVERSITY OF CRETE

Πρόβλεψη απώλειας πελατών με χρήση Νευρωνικών Δικτύων

Customer Churn prediction using Neural Nets

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Κοντοπάνος Εμμανουήλ

Α.Μ: 2017010003

Επιβλέπων καθηγητής: Τσαφαράκης Στέλιος

Χανιά,2023

Ευχαριστίες

Με την ολοκλήρωση της παρούσας διπλωματικής εργασίας κλείνει ένας πολύ όμορφος κύκλος της ζωής μου. Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της διπλωματικής μου εργασίας κ. Τσαφαράκη Στέλιο για την βοήθεια και την εμπιστοσύνη που μου παρείχε. Επιπλέον θα ήθελα να εκφράσω την ευγνωμοσύνη και την εκτίμηση μου στον κ. Κυριακίδη Αναστάσιο για τη συνεισφορά του σε όλα τα στάδια της διπλωματικής εργασίας. Η βοήθεια που μου παρείχε ήταν πολύτιμη για την εκπόνηση της διπλωματικής εργασίας. Παράλληλα, θα ήθελα να ευχαριστήσω τους φίλους μου που ήταν δίπλα μου σε όλες τις στιγμές της φοιτητικής μου ζωής. Τέλος, θα ήθελα να πω ένα τεράστιο ευχαριστώ στην οικογένεια μου και κυρίως τους γονείς μου για την στήριξη που μου παρείχαν όλα αυτά τα χρόνια.

Περιεχόμενα

Περίληψη

Τα νευρωνικά δίκτυα χρησιμοποιούνται όλο και περισσότερο σε διάφορους τομείς, όπως είναι οι επιχειρήσεις, η ιατρική και βιομηχανίες, για την εξόρυξη δεδομένων (data mining) από σύνθετες βάσεις δεδομένων (datasets). Η απώλεια των πελατών (Customer Churn) αποτελεί ένα από τα πιο σημαντικά προβλήματα που καλούνται να αντιμετωπίσουν οι επιχειρήσεις και για αυτό πρέπει να βρουν τρόπους ώστε να αποφευχθεί το συγκεκριμένο πρόβλημα ή τουλάχιστον να ελαχιστοποιηθεί. Διάφορες μέθοδοι Μάρκετινγκ χρησιμοποιούνται από τις εταιρείες για να βρεθούν τα αίτια της απώλειας των πελατών ώστε να μπορούν να διορθώσουν το συγκεκριμένο πρόβλημα πριν αρχίσουν και χάνουν ακόμα μεγαλύτερο όγκο πελατών. Επιπλέον, χρησιμοποιούνται και διάφοροι μέθοδοι ικανοί να προβλέψουν μέσα από τις βάσεις δεδομένων των εταιρίων την απώλεια των πελατών. Στόχος, της συγκεκριμένης εργασίας είναι η ανάλυση ορισμένων μεθόδων Μάρκετινγκ που μπορούν να αποτρέψουν το συγκεκριμένο πρόβλημα και η εκπαίδευση ενός μοντέλου χρησιμοποιώντας νευρωνικά δίκτυα στις βάσεις δεδομένων, ώστε να είναι ικανό να μπορεί να πραγματοποιήσει προβλέψεις όσον αφορά την απώλεια των πελατών. Για αυτό το λόγο, θα γίνει η χρήση των νευρωνικών δικτύων σε μία βάση δεδομένων η οποία απαρτίζεται από τον συνολικό αριθμό των πελατών μιας εταιρείας.

Η βάση δεδομένων αποτελείται από δεδομένα που αφορούν 64,000 πελάτες και προέρχεται από τη βιβλιοθήκη Kaggle. Οι μεταβλητές των βάσεων δεδομένων αφορούν τα δημογραφικά στοιχεία των πελατών, δηλαδή τις περιοχές στις οποίες μένουν, τα χρήματα που έχουν δαπανήσει στην εταιρεία, καθώς και πότε πραγματοποιήθηκε η τελευταία τους αγορά. Επιπλέον, απεικονίζονται οι παροχές που έχουν λάβει από την εταιρεία, αν έχουν δεχτεί δηλαδή προσφορές ή όχι από εκείνη, καθώς και το αν παραμένουν ακόμα πελάτες ή όχι. Με βάση αυτά τα δεδομένα, το μοντέλο εκπαιδεύεται ώστε να βρίσκει τους πελάτες οι οποίοι είναι πιθανό να αποχωρήσουν από την εταιρεία ώστε να μπορεί στο μέλλον να κάνει προβλέψεις.

Το λογισμικό που χρησιμοποιείται στη συγκεκριμένη έρευνα είναι το Orange Data Mining (έκδοση 3.34.0) το οποίο χρησιμοποιεί τη γλώσσα προγραμματισμού Python και τα πνευματικά του δικαιώματα ανήκουν στο πανεπιστήμιο της Λιουμπλιάνας. Περιλαμβάνει διάφορους αλγόριθμους Μηχανικής Μάθησης και τα νευρωνικά δίκτυα που χρησιμοποιούνται στη συγκεκριμένη εργασία.

Αρχικά, η βάση δεδομένων χωρίζεται σε δύο κλάσεις και ως κλάση-στόχος θεωρείται η μεταβλητή conversion, η οποία είναι ο δείκτης μετατροπής και απεικονίζει την διατήρηση ή όχι των πελατών στην εταιρείας.

Ακολούθως, γίνεται η εκπαίδευση του ταξινομητή των νευρωνικών δικτύων για την πρόβλεψη της συμπεριφοράς των πελατών ως προς την εταιρεία, αν θα παραμείνουν δηλαδή ή όχι σε εκείνη. Ο ταξινομητής εκπαιδεύεται για διαφορετικούς συνδυασμούς των υπερπαραμέτρων των νευρωνικών δικτύων, από τις οποίες απεικονίζεται ο συνδυασμός που επιφέρει τα καλύτερα αποτελέσματα όσον αφορά την απόδοσή του. Επειδή όμως στη συγκεκριμένη βάση δεδομένων οι πελάτες που τελικά παραμένουν στην εταιρεία είναι αρκετά περισσότεροι από εκείνους που τελικά αποχωρούν δημιουργείται το πρόβλημα της άνισης κατανομής των κλάσεων (class imbalance problem).

Για την αντιμετώπιση του συγκεκριμένου προβλήματος εφαρμόζονται τεχνικές προεπεξεργασίας (preprocessing) των δεδομένων αλλά και τεχνικές resample με αποτέλεσμα την δημιουργία νέων βάσεων δεδομένων με διαφορετικές προσεγγίσεις. Για όλες τις τεχνικές, το μοντέλο εκπαιδεύεται και έχει διαφορετικές τιμές απόδοσης για την κάθε τεχνική και για διαφορετικούς συνδυασμούς των υπερπαραμέτρων των νευρωνικών δικτύων.

Τέλος, σκοπός της συγκεκριμένης εργασίας είναι η εύρεση της τεχνικής εκείνης και ο συνδυασμός των υπερπαραμέτρων των νευρωνικών δικτύων για τις οποίες το μοντέλο μπορεί να εκπαιδευτεί και να αποδώσει σε τέτοιο βαθμό ώστε οι τιμές των μέτρων πρόβλεψης του να λαμβάνουν τις μεγαλύτερες τιμές τους.

ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ

1.1 Στόχοι της διπλωματικής εργασίας

Με την πάροδο των χρόνων, η επιστήμη των υπολογιστών έχει σημειώσει μεγάλη πρόοδο με αποτέλεσμα την δημιουργία νέων αναγκών, οι οποίες οδήγησαν στην ανάπτυξη διαφόρων μεθόδων για την επεξεργασία δεδομένων και τη μεθοδική λήψη αποφάσεων. Ένας από τους πολλούς τομείς στον οποίο η ανάλυση δεδομένων έχει δώσει λύσεις είναι το Μάρκετινγκ, καθώς από τα δεδομένα εξάγονται χρήσιμες πληροφορίες και μπορούν να παρθούν αποφάσεις σε μια επιχείρηση με μειωμένο ρίσκο. Ένας από τους βασικούς στόχους των πωλήσεων και του Μάρκετινγκ είναι η δημιουργία εμπιστοσύνης των καταναλωτών και η προώθηση της αφοσίωσης των πελατών. Με τα μεγάλα δεδομένα μπορούν να εφαρμοστούν ουσιαστικές διαδικασίες βελτίωσης σε βασικούς τομείς όπως η εμπλοκή πελατών, η αφοσίωση και η βελτιστοποίηση του Μάρκετινγκ.

Τα μεγάλα δεδομένα σε συνδυασμό με τα αναλυτικά στοιχεία μπορούν να παρέχουν πληροφορίες σχετικά με τις ανάγκες, τις τοποθεσίες και τα στοιχεία επικοινωνίας των πελατών βοηθώντας τις εταιρείες να εντοπίσουν και να στοχεύσουν πιθανούς πελάτες.

Ένας τρόπος ανάλυσης των δεδομένων αυτών, είναι με τη χρήση των νευρωνικών δικτύων τα οποία αποτελούν μέθοδο της τεχνητής νοημοσύνης και διδάσκουν στους υπολογιστές να επεξεργάζονται δεδομένα με τρόπο εμπνευσμένο από τη δομή και τη λειτουργία του ανθρώπινου εγκεφάλου. Μπορούν να βοηθήσουν σημαντικά τους υπολογιστές να λάβουν έξυπνες αποφάσεις χωρίς τη συμμετοχή ανθρώπινου παράγοντα, αφού μπορούν να εκπαιδευτούν σε μεγάλα σύνολα δεδομένων και να μοντελοποιήσουν περίπλοκες σχέσεις μεταξύ δεδομένων εισόδου και εξόδου, βγάζοντας συμπεράσματα και πραγματοποιώντας διάφορες γενικεύσεις.

Στην συγκεκριμένη εργασία, λαμβάνεται αρχικά μία βάση δεδομένων που περιέχει χαρακτηριστικά πελατών μιας εταιρείας. Στη βάση δεδομένων, η κλάση-στόχος αποτελεί η μεταβλητή conversion, η οποία δείχνει αν ένας πελάτης παραμένει σε μια εταιρεία ή όχι. Στόχος της εργασίας, αποτελεί η εκπαίδευση ενός μοντέλου με τη χρήση του ταξινομητή των νευρωνικών δικτύων, το οποίο θα είναι ικανό να πραγματοποιεί προβλέψεις σε ικανοποιητικό βαθμό για την απώλεια των πελατών (Customer Churn) ή την διατήρηση τους.

1.2 Δομή της διπλωματικής εργασίας

Αρχικά θα γίνει μια αναφορά στις σύγχρονες μεθόδους Μάρκετινγκ που εφαρμόζουν οι επιχειρήσεις όπως είναι το στοχευμένο Μάρκετινγκ και η διαχείριση των πελατειακών σχέσεων. Επιπλέον, θα αναλυθούν οι έννοιες της ικανοποίησης, αφοσίωσης και απώλειας των πελατών, ενώ θα παρουσιαστούν αποτελέσματα προηγούμενων ερευνών που αφορούν την απώλεια των πελατών με χρήση των νευρωνικών δικτύων.

Στη συνέχεια. θα υπάρξει μια εισαγωγή στα νευρωνικά δίκτυα, στις εφαρμογές τους, στην υλοποίηση τους, στα είδη τους καθώς και στα είδη μάθησης τους. Σημαντική αναφορά γίνεται και στην προεπεξεργασία των δεδομένων ώστε να επιλεχθούν εκείνα που χρειάζονται για ανάλυση. Επιπλέον, αναλύονται οι υπερπαράμετροι τους, τα μέτρα απόδοσης τους και η προσαρμογή τους.

Ακολούθως, αναλύεται το πειραματικό μέρος της έρευνας, το λογισμικό που θα χρησιμοποιηθεί και τέλος παρουσιάζονται τα αποτελέσματα συγκριτικά για διαφορετικές τιμές των υπερπαραμέτρων των νευρωνικών δικτύων, καθώς και τα συμπεράσματα που προκύπτουν από την έρευνα.

Κεφάλαιο 2: Θεωρητικό Υπόβαθρο

2.1 Στοχευμένο Μάρκετινγκ

Η στόχευση στο Μάρκετινγκ είναι μια στρατηγική που διαχωρίζει την ευρύτερη αγορά σε μικρότερα τμήματα. Οι επιχειρήσεις έχουν εύκολη πρόσβαση σε αποθήκες δεδομένων που παρέχουν πληροφορίες για τα δημογραφικά στοιχεία, την ψυχολογία και τη συμπεριφορά των πελατών ώστε να βρεθούν οι κατάλληλοι πελάτες στόχευσης. Η στόχευση γίνεται ώστε να αναγνωριστούν οι πελάτες που είναι πιο πιθανό να αγοράσουν συγκεκριμένο προϊόν ή υπηρεσία. Αυτή η τεχνική επιλέγει πιθανούς πελάτες με μεγαλύτερη επιτυχία από ότι οι κλασσικές μέθοδοι.

Ο τρόπος στόχευσης των πελατών γίνεται με τη δημιουργία ενός μοντέλου βασισμένου στα χαρακτηριστικά τους. Τα νευρωνικά δίκτυα βάση των μεταβλητών τους μπορούν να το μοντελοποιήσουν και να ανιχνεύσουν ποιοι πελάτες ξεχωριστά θα ενδιαφέρονται για συγκεκριμένα προϊόντα και υπηρεσίες. Με αυτόν τον τρόπο γίνεται πιο εύκολη η τμηματοποίηση των πελατών, οι εταιρείες επενδύουν τους πόρους τους στο Μάρκετινγκ πιο αποτελεσματικά και με σχετικά οικονομικό και γρήγορο τρόπο μπορούν να διαχειριστούν τους ήδη υπάρχοντες πελάτες και να αποκτήσουν καινούριους. [1]

2.1.1 Διαδικασία στοχευμένου Μάρκετινγκ

Η διαδικασία του στοχευμένου Μάρκετινγκ γίνεται μέσω τριών σημαντικών βημάτων:

1) Η τμηματοποίηση της αγοράς (**Segmentation**)

2) Η στόχευση της αγοράς (**Targeting**)

3) Η χωροθέτηση του προϊόντος (**Positioning**)

Η παραπάνω διαδικασία είναι γνωστή ως **STP** και καταδεικνύει τους δεσμούς σε μια συνολική αγορά και το πως μια εταιρεία ανταγωνίζεται σε αυτή. Στην αγορά υπάρχουν πολλοί τύποι πελατών όπου ο καθένας έχει ξεχωριστές ανάγκες. Για να αναπτυχθεί μια στρατηγική Μάρκετινγκ, η εταιρεία πρέπει να κατανοήσει τους πελάτες-στόχους της και για αυτό αρχικά πρέπει να ορίσει την αγορά ενδιαφέροντος της. Στην διαδικασία του STP, αρχικά πραγματοποιείται η τμηματοποίηση, μετά η επιλογή ενός ή περισσότερων αγορών στόχων και τέλος εφαρμόζεται η χωροθέτηση.

2.1.2 Τμηματοποίηση

Η τμηματοποίηση αποτελεί ένα τρόπο ομαδοποίησης των πελατών σε μικρότερες ομάδες με κοινά χαρακτηριστικά. Μέσω αυτής προσδιορίζονται τα τμήματα που μπορεί να είναι πιο αποδοτικά ή ελκυστικά για την εταιρεία, ώστε να εφαρμοστεί το κατάλληλο μείγμα Μάρκετινγκ σε αυτά. Με την τμηματοποίηση διαιρείται η αγορά-στόχος σε μικρότερες ομάδες ή τμήματα ώστε η εταιρεία να εφαρμόζει κατάλληλη στρατηγική Μάρκετινγκ για κάθε στόχο. Αυτό έχει ως αποτέλεσμα καλύτερη αντιστοίχιση των αναγκών των πελατών, μεγαλύτερα κέρδη για την επιχείρηση, καλύτερες ευκαιρίες για ανάπτυξη και διατήρηση περισσότερων πελατών. Η τμηματοποίηση μιας αγοράς γίνεται με βάση τις εξής μεταβλητές:

- Με βάση τα Δημογραφικά στοιχεία, τα οποία είναι τα χαρακτηριστικά του πελάτη, όπως είναι η ηλικιακή ομάδα, το φύλο, ο κύκλος ζωής και το εισόδημα
- Με βάση τα Ψυχογραφικά στοιχεία που είναι ο τρόπος ζωής των πελατών
- Με βάση τη Συμπεριφορά των πελατών όπως η ευαισθησία στην τιμή, η εμπιστοσύνη στη μάρκα και τα ζητούμενα οφέλη
- Με βάση τη Γεωγραφική θέση που αφορά την τμηματοποίηση της αγοράς μέσω χωρών, περιοχών, πόλεων και τα λοιπά.

Με τον ορισμό των τμημάτων, μπορούν να αξιολογηθούν ποια τμήματα είναι ελκυστικότερα για την επιχείρηση και να ακολουθήσει η στόχευση των τμημάτων αυτών.

2.1.3 Στόχευση της αγοράς

Η στόχευση είναι το δεύτερο στάδιο του STP και λαμβάνει χώρα όταν καθοριστούν τα τμήματα των αγορών-στόχων. Οι αγορές-στόχοι επιλέγονται αφού εξεταστούν πολλοί παράγοντες με ορισμένους να είναι το μέγεθος της αγοράς, ο ρυθμός ανάπτυξης της και ο ανταγωνισμός. Η εταιρεία επιλέγει ως αγορά-στόχο εκείνη ή εκείνες τις αγορές που είναι πιθανότερο να αγοράσουν το προϊόν της. Οι προσεγγίσεις που αφορούν τη στόχευση είναι οι εξής:

Α) Αδιαφοροποίητο- Μαζικό Μάρκετινγκ που δεν απευθύνονται σε συγκεκριμένο τμήμα της αγοράς

Β) Η εστιασμένη προσέγγιση η οποία στοχεύει σε διαφορετικά τμήματα της αγοράς με συγκεκριμένα μείγματα Μάρκετινγκ ώστε να καλύψουν τις ανάγκες των τμημάτων αυτών. Έχει ως στόχο δηλαδή τη δημιουργία εξειδικευμένου προϊόντος ή υπηρεσίας που απευθύνεται σε μια μικρότερη ομάδα ανθρώπων.

Αυτό το βήμα της διαδικασίας του STP, δείχνει τις ομάδες πελατών που είναι πιθανότερο να αγοράσουν προϊόντα ή υπηρεσίες της εταιρείας, καθώς και το είδος του Μάρκετινγκ και των πωλήσεων που χρειάζεται να αναπτυχθεί.

2.1.4 Χωροθέτηση

Τελευταίο βήμα της διαδικασίας του STP αποτελεί η χωροθέτηση, δηλαδή ο τρόπος με τον οποίο ένα προϊόν ή μια μάρκα εκπροσωπείται σε σχέση με την κατηγορία του και με τα ανταγωνιστικά προϊόντα της κατηγορίας αυτής. Με την χωροθέτηση δημιουργείται μια αντίληψη των προϊόντων στο μυαλό του κοινού-στόχου, κάνοντας έτσι την εταιρεία να φαίνεται διαφορετική ή ότι παρέχει μεγαλύτερη αξία. Σε πλήρη κατανόηση των ομάδων-στόχων των καταναλωτών, το Μάρκετινγκ μπορεί να αναπτύξει τα χαρακτηριστικά και τα οφέλη των καταναλωτών. Η χωροθέτηση δείχνει τη συνολική αξία της εταιρείας, αυτής που δημιουργεί και διατηρεί πελάτες.

Συμπερασματικά, η διαδικασία του STP επιτρέπει σε μια επιχείρηση να πραγματοποιήσει μια στρατηγική Μάρκετινγκ που δένει την εταιρεία, τη μάρκα και τα οφέλη του προϊόντος σε συγκεκριμένα τμήματα πελατών της αγοράς. Χρησιμεύει στο σχεδιασμό της στρατηγικής Μάρκετινγκ ως μια διαδικασία τμηματοποίησης των αγορών, στοχεύοντας συγκεκριμένους πελάτες ώστε να χωροθετηθεί η προσφορά αποτελεσματικά μεταξύ του ανταγωνισμού.[2]

2.2 Διαχείριση Πελατειακών Σχέσεων (CRM)

Για να μπορούν οι εταιρείες να παραμείνουν στη σημερινή αγορά, θα πρέπει να διατηρήσουν μακροχρόνιες σχέσεις με τους πελάτες τους. Σε αυτό, σημαντικό ρόλο διαθέτουν τα συστήματα διαχείρισης πελατειακών σχέσεων ή αλλιώς **CRM** (Customer Relationship Management), καθώς παρέχουν πλατφόρμες σε μονάδες της εταιρείας, ώστε να επικοινωνούν, να κατανοούν και να εκπληρώνουν τις ανάγκες των πελατών.

Η διαδικασία του CRM είναι η κατηγοριοποίηση και ανάλυση των πελατών βάση της σχέσης τους με τον οργανισμό. Οι πληροφορίες που λαμβάνονται από αυτές τις αναλύσεις οδηγούν στη λήψη αποφάσεων που ανταποκρίνονται επικερδώς στις ανάγκες διαφόρων ομάδων πελατών που ταυτοποιούνται κατά τη διαδικασία. Χρησιμοποιείται για να συγκεντρώσει δεδομένα ώστε να διαχωρίσει σε υποομάδες τους πελάτες ανάλογα με τις προτιμήσεις τους και να σχεδιαστούν στρατηγικές Μάρκετινγκ ελκυστικές σε αυτούς. Επιπλέον, σε κάθε πελάτη παρέχεται υποστήριξη μετά την πώληση προϊόντων, οπότε οι εταιρείες κερδίζουν την εμπιστοσύνη των πελατών το οποίο συνεπάγεται την διατήρηση τους.

Το CRM χωρίζεται σε 4 διαφορετικά στάδια:

- Στρατηγικό CRM: Χρησιμοποιείται από τις επιχειρήσεις για τον διαχωρισμό των πελατών σε τμήματα, για την ανάλυση της αξίας τους χρησιμοποιώντας ένα μοντέλο που βασίζεται σε αυτούς, για τον υπολογισμό των κερδών από αυτούς, για την αξιολόγηση του κόστους στο Μάρκετινγκ και τις πωλήσεις και για τη δημιουργία ενός πίνακα που θα προβλέπει τη ροή των πελατών
- E-CRM: Παρέχει πρόσβαση σε πληροφορίες που αφορούν τους πελάτες σε πραγματικό χρόνο μέσω του ίντερνετ δίνοντας την ευκαιρία στους πελάτες και στις επιχειρήσεις να διαθέτουν πληροφορίες των πελατών, του Μάρκετινγκ, των πωλήσεων και των μεταπωλήσεων
- Αναλυτικό CRM: Χρησιμοποιείται για την ενσωμάτωση και επεξεργασία των δεδομένων που χρειάζονται μετατρέποντας τα σε πληροφορίες που είναι χρήσιμες για την διαχείριση των πελατειακών σχέσεων
- Συνεργατικό CRM: Περιλαμβάνει το Μάρκετινγκ, τις πωλήσεις και τις μεταπωλήσεις

Σε μια επιχείρηση το CRM βοηθάει στην ανάπτυξή της, καθώς:

- Βοηθάει στην αναγνώριση του ορθά στοχευμένου τμήματος της επιχείρησης
- Βελτιώνει την ικανοποίηση των πελατών με τις υπηρεσίες που παρέχει
- Βελτιώνει τη συνοχή της εταιρείας ορίζοντας εταιρικούς στόχους που συνδέονται με την ικανοποίηση των πελατών
- Στοχεύει στην αύξηση των πελατών και την διασφάλιση της εμπιστοσύνης τους
- Ενισχύει και επεκτείνει τις σχέσεις των πελατών παράγοντας νέες επιχειρηματικές ευκαιρίες
- Μειώνει το κόστος πωλήσεων
- Αυξάνει την αποτελεσματικότητα της παροχής υπηρεσιών στους πελάτες αφού διαθέτει ολοκληρωμένες πληροφορίες
- Παρέχει πληροφορίες των πωλήσεων και του Μάρκετινγκ σχετικά με τις απαιτήσεις, τις προσδοκίες και τις αντιλήψεις των πελατών

Εν κατακλείδι, το CRM είναι μία από τις λίγες λειτουργίες που έχει βαθύ και άμεσο αντίκτυπο στις επιχειρήσεις και για αυτό είναι κάτι που πρέπει να μελετηθεί εκτενώς από κάθε επαγγελματία. Για την πρόβλεψη απώλειας πελατών θα χρησιμοποιηθεί το Αναλυτικό CRM, καθώς αυτό παρέχει όλες τις πληροφορίες από τα δεδομένα.[3]

2.3 Customer Satisfaction

Η ικανοποίηση των πελατών, αποτελεί το μέτρο εκείνο που δείχνει πως τα προϊόντα μιας εταιρείας, οι υπηρεσίες και γενικά η συνολική εμπειρία των πελατών ανταποκρίνονται στις προσδοκίες των πελατών. Για τις εταιρείες, η έννοια της ικανοποίησης των πελατών μπορεί να δοθεί με διάφορους τρόπους, ωστόσο το βέβαιο είναι ότι υψηλή ικανοποίηση των πελατών σημαίνει πως οι πελάτες είναι ικανοποιημένοι και οι επιχειρήσεις ανθίζουν. Σύμφωνα με έρευνα της Zendesk που πραγματοποιήθηκε το 2022 και αφορά τις εμπειρίες των πελατών, μετά από μια αρνητική εμπειρία ένας πελάτης έχει περίπου 61% πιθανότητα να συνεχίσει τις αγορές του σε ανταγωνιστή και για αυτό είναι πολύ σημαντικό οι εταιρείες να επικεντρωθούν στην ικανοποίησή τους. [4]

Η ικανοποίηση των πελατών μπορεί να οριστεί ως το μέτρο εκείνο με το οποίο οι πελάτες κρίνουν αν το προϊόν ανταποκρίνεται στις προσδοκίες τους και θεωρείται από τις εταιρείες ως ο πιο σημαντικός παράγοντας όσον αφορά τον ανταγωνισμό και την επιτυχία. Στην ουσία είναι η αξιολόγηση του πελάτη ως προς την απόδοση των υπηρεσιών και προϊόντων της εταιρείας. Αποτελεί πολύ σημαντικό παράγοντα στη σημερινή εποχή, καθώς η ικανότητα της παροχής υπηρεσιών για την επίτευξη υψηλού βαθμού ικανοποίησης είναι σημαντική για την διαφοροποίηση του προϊόντος στην αγορά και για τη δημιουργία δυνατών σχέσεων με τους πελάτες.

Μέσω της ικανοποίησης των πελατών, οι πελάτες είναι πολύ πιθανό να μείνουν πιστοί στην εταιρεία, καθώς όλες οι ανάγκες τους ικανοποιούνται και αυτό έχει ως αποτέλεσμα την δημιουργία σχέσεων μεταξύ πελατών και εταιρείας, άρα περισσότερα κέρδη για εκείνη. Για αυτό μια εταιρεία θα πρέπει να επικεντρωθεί στη βελτίωση της ποιότητας των υπηρεσιών που παρέχει και στην χρέωση κατάλληλων τιμών για εκείνες που παρέχει, ώστε να ικανοποιεί τους πελάτες για να μπορέσει να τους διατηρήσει.

Ένα πολύ συχνό φαινόμενο αποτελεί το γεγονός ότι οι υπηρεσίες που προσφέρει μια επιχείρηση και η τιμή η οποία τις χρεώνει να καθορίζουν το βαθμό της ικανοποίησης μεταξύ των πελατών περισσότερο από κάθε άλλο μέτρο. Πολύ σημαντική είναι και η ενασχόληση του πελάτη, καθώς όταν εκείνος θεωρεί το προϊόν σημαντικό και αφιερώσει χρόνο ώστε να λάβει πληροφορίες για αυτό, τότε ενισχύεται σε μεγάλο βαθμό η ικανοποίηση του.

Γενικά, η ικανοποίηση των πελατών έχει θετικό αντίκτυπο στην εταιρεία, καθώς μπορεί ο πελάτης να πραγματοποιήσει και άλλες αγορές, να αγοράσει πολλά προϊόντα, να διαδώσει θετικές φήμες για την εταιρεία και να έχει θέληση να πληρώσει περισσότερα για τη συγκεκριμένη μάρκα. Αντίθετα, αν οι πελάτες δεν είναι ικανοποιημένοι είναι πολύ πιθανό η εταιρεία να χάσει ένα μερίδιο της αγοράς, να

χάσει πελάτες και επενδυτές, κυρίως αν δεν τους ικανοποιεί σε βαθμό που το πράττουν οι ανταγωνιστές της.[5]

2.4 Customer Loyalty

Η αφοσίωση των πελατών μπορεί να βοηθήσει μια επιχείρηση να ανθίσει, αλλά και να επιβιώσει σε δύσκολες στιγμές για εκείνη. Αρχικά, οι επιχειρήσεις μπορούν να αποκτήσουν μια βάση δεδομένων στην οποία θα στηρίζονται με τον καιρό. Οι σταθεροί πελάτες επειδή εμπιστεύονται την επιχείρηση τείνουν να κάνουν περισσότερες αγορές και η διατήρησή τους είναι αρκετά πιο οικονομική από την απόκτηση καινούριων. Επιπλέον, με τους πιστούς πελάτες, μειώνονται τα κόστη της επιχείρησης σε αρκετούς τομείς. Για αυτούς τους λόγους είναι σημαντικό οι επιχειρήσεις να σκέφτονται τους πελάτες τους σε ότι κι αν κάνουν. Ως πιστοί πελάτες, θεωρούνται εκείνοι που πραγματοποιούν τακτικά αγορές, αγοράζουν πολλά προϊόντα και υπηρεσίες, αναφέρουν την εταιρεία σε άλλους και δεν πραγματοποιούν αγορές από κάποιον ανταγωνιστή.

Η αφοσίωση των πελατών δημιουργείται σταδιακά και μέσα από κάποια στάδια αποκτούνται πιστοί πελάτες. Υπάρχουν άνθρωποι οι οποίοι ανήκουν σε διάφορα είδη πελατών Αρχικά, οποιοσδήποτε ενδέχεται να αποκτήσει τα προϊόντα ή τις υπηρεσίες μιας επιχείρησης. Ένας πελάτης αποτελεί προοπτική αν έχει την ανάγκη να αγοράσει τα προϊόντα της επιχείρησης ή της υπηρεσίες της και μπορεί να τα αγοράσει, ενώ αντίθετα δεν αποτελεί προοπτική αν δεν τα χρειάζεται ή δεν μπορεί να τα αγοράσει. Στη συνέχεια υπάρχουν οι πελάτες που έχουν πραγματοποιήσει αγορά μόνο μια φορά οι οποίοι ενδέχεται να αγοράζουν και από ανταγωνιστή αλλά και οι επαναλαμβανόμενοι πελάτες που έχουν πραγματοποιήσει αγορές δύο ή περισσότερες φορές. Επιπλέον, υπάρχουν πελάτες οι οποίοι ψωνίζουν τακτικά από την επιχείρηση, καθώς αγοράζουν από εκείνη οτιδήποτε μπορεί να χρειαστούν. Τέλος, υπάρχουν και πελάτες οι οποίοι ενθαρρύνουν και άλλους να αγοράσουν από την ίδια εταιρεία και την συνιστούν σε όλο και περισσότερο κόσμο.

Ένας πελάτης ο οποίος αγοράζει για πρώτη φόρα με την πάροδο του χρόνου μπορεί να γίνει πιστός πελάτης. Κατά τη διάρκεια κάθε αγοράς, ένας πελάτης διανύει ένα κύκλο αγορών στον οποίο ενημερώνεται για το προϊόν, κάνει μια αρχική επένδυση, στη συνέχεια αξιολογεί το προϊόν που αγόρασε, αποφασίζει να αγοράσει ξανά από την εταιρεία και εντέλει το πράττει. Είναι πολύ σημαντικό για μια επιχείρηση να μετατραπεί σε πιστό πελάτη ένας πελάτης που αγοράζει για πρώτη φορά για τους ακόλουθους λόγους:

- Αυξάνονται οι πωλήσεις, καθώς οι πελάτες αγοράζουν περισσότερα από την επιχείρηση
- Ενισχύεται η θέση της εταιρείας στην αγορά, καθώς προτιμάται έναντι των ανταγωνιστών της

- Τα κόστη του Μάρκετινγκ μειώνονται, καθώς δεν χρειάζεται να δαπανηθούν χρήματα για ήδη υπάρχων πελάτες, οι οποίοι αν είναι ικανοποιημένοι ενθαρρύνουν και άλλους να πραγματοποίησουν αγορές από την επιχείρηση, οπότε δεν αποτελεί ανάγκη η διαφήμιση της
- Πιστοί πελάτες θα παραμείνουν στην εταιρεία και δεν θα πάνε σε κάποιον ανταγωνιστή για διαφορά λίγων χρημάτων
- Ένας ικανοποιημένος πελάτης είναι πολύ πιθανό να δοκιμάσει και τα υπόλοιπα προϊόντα, αυξάνοντας τα κέρδη της εταιρείας

Συμπερασματικά, χρειάζεται χρόνος για να δημιουργηθεί αφοσίωση του πελάτη με την εταιρεία. Για να το καταφέρει αυτό μια εταιρεία, πρέπει να φροντίσει ώστε η επαφή τους από την πρώτη στιγμή με τα προϊόντα της και τις υπηρεσίες που παρέχει να τους ικανοποιήσουν. Οι εταιρείες πρέπει να επιδιώκουν στη διατήρηση μιας βάσης σταθερών πελατών, καθώς εκείνοι τις κάνουν κερδοφόρες.[6]

2.5 Customer Churn

Η αποχώρηση πελατών είναι το φαινόμενο κατά το οποίο οι πελάτες μιας επιχείρησης σταματούν να αγοράζουν από αυτή προϊόντα ή υπηρεσίες και να αλληλοεπιδρούν μαζί της. Ουσιαστικά, πελάτες εγκαταλείπουν μια επιχείρηση και στρέφονται στους ανταγωνιστές της. Στόχος των εταιρειών είναι να αναγνωρίσουν ποιοι πελάτες είναι πιθανό να αποχωρήσουν από την εταιρεία και να λάβουν κάποια μέτρα για να τους διατηρήσουν.

Η ανάλυση της αποχώρησης των πελατών αποτελεί πολύ σημαντικό παράγοντα για την κερδοφορία μίας επιχείρησης, καθώς σύμφωνα με έρευνες η απόκτηση νέων πελατών στη σημερινή εποχή λόγω του ανταγωνισμού που υπάρχει κοστίζει πάνω από 10 φορές περισσότερο από την διατήρηση των ήδη υπάρχων. Αυτή η μέθοδος ανάλυσης χρησιμοποιείται σε περιοχές όπως ο καθορισμός προφίλ των υπάρχων πελατών και για την ανάλυση και εκτίμηση πελατών που θα αποχωρήσουν. Επιπλέον, η αξία των εταιρειών είναι ανάλογη των ενεργών πελατών που διαθέτει, καθώς παράμετροι όπως τα κόστη, η κερδοφορία, το μέγεθος, οι επενδύσεις, η χωρητικότητα, η ροή χρημάτων των εταιρειών, εξαρτώνται από τον αριθμό των πελατών που διαθέτει και συνεπώς από την αφοσίωση τους. Επίσης, οι εταιρείες έχουν μεγάλη κερδοφορία αν διαθέτουν πελάτες πιστούς σε εκείνη για πολλά χρόνια.[7]

Ο δείκτης απώλειας πελατών είναι το ποσοστό του αριθμού των πελατών που έχουν χαθεί προς τον συνολικό αριθμό των πελατών για μια συγκεκριμένη χρονική περίοδο. Είναι πολύ σημαντικό ο δείκτης αυτός να μένει σε χαμηλά ποσοστά, καθώς όσο μεγαλύτερος είναι, τόσο μεγαλύτερο ποσοστό πελατών έχει αποχωρήσει από την επιχείρηση. Αυτό έχει αρνητικό αντίκτυπο για την επιχείρηση, καθώς το κόστος απόκτησης νέων πελατών είναι πολύ μεγαλύτερο από το κόστος διατήρησής τους και οι

υπάρχων πελάτες αν είναι ικανοποιημένοι με την επιχείρηση, το συζητάνε και με τον κοινωνικό τους κύκλο με αποτέλεσμα την αύξηση της βάσης πελατών της εταιρείας χωρίς τη δαπάνη περισσότερων χρημάτων.

Η πρόληψη της απώλειας πελατών έχει τα εξής πλεονεκτήματα:

- Λαμβάνονται πληροφορίες για βελτίωση, καθώς οι δυσαρεστημένοι πελάτες είναι πηγή εποικοδομητικών σχολίων με τα οποία μια επιχείρηση μπορεί να βελτιωθεί
- Μειώνεται το ρίσκο μιας επιχείρησης, καθώς πιθανή απώλεια των πελατών είναι επιβλαβή για την ανάπτυξή της
- Κατανόηση της στοχευμένης αγοράς, αφού η διαρκή προσπάθεια μείωσης της απώλειας πελατών, βοηθάει την επιχείρηση να κατανοήσει καλύτερα τη στοχευμένη αγορά
- Δημιουργία ανταγωνιστικού πλεονεκτήματος στην αγορά, διότι οι επιχειρήσεις μπορούν να διατηρήσουν τους πελάτες τους και ταυτόχρονα να προσπαθούν να αποκτήσουν καινούριους

Συμπερασματικά, η απώλεια πελατών αποτελεί ένα πολύ σημαντικό κομμάτι για τις λειτουργίες της επιχείρησης. Η κατανόηση των λόγων που συμβαίνει, καθώς και η επινόηση αποτελεσματικών στρατηγικών για τον περιορισμό της είναι απαραίτητη για τη μακροχρόνια επιτυχία της.[8]

2.6 Προγενέστερες Έρευνες

Μέσω της ανάλυσης σύνθετων και μεγάλων βάσεων δεδομένων οι επιχειρήσεις είναι σε θέση να κατανοήσουν τις συμπεριφορές των πελατών τους. Με αυτόν τον τρόπο μπορούν να χρησιμοποίησουν πιο αποδοτικές και στοχευμένες τεχνικές Μάρκετινγκ. Σημαντικός στόχος σε μια επιχείρηση είναι η πρόβλεψη του ποσοστού της απώλειας των πελατών ώστε να βρεθούν οι λόγοι για τους οποίους πελάτες αποχωρούν και προκαλείται μεγάλη οικονομική ζημιά στην επιχείρηση. Στις περισσότερες βάσεις δεδομένων παρουσιάζεται το πρόβλημα της άνισης κατανομής των κλάσεων (**class imbalance problem**). Αυτό συμβαίνει διότι συνήθως το ποσοστό των πελατών που αποχωρεί τελικά από την εταιρεία είναι πολύ μικρό σε σχέση με εκείνο που παραμένει. Αυτό έχει ως αποτέλεσμα την αδυναμία των μοντέλων πρόβλεψης να εξετάσουν το ποσοστό των πελατών που τελικά αποχωρεί με αποτέλεσμα να μην είναι εφικτή η εξαγωγή χρήσιμων συμπερασμάτων και αποτελεσμάτων. Οι παρακάτω έρευνες επικεντρώνονται σε τρόπους βελτίωσης της ακρίβειας και ποιότητας των αποτελεσμάτων που λαμβάνονται από τις βάσεις δεδομένων με χρήση νευρωνικών δικτύων και επεξεργασίας δεδομένων.

Οι *CF Tsai και YH Lu* (2009) γνωρίζουν ότι για να διατηρήσουν οι εταιρείες πολύτιμους πελάτες, η ικανότητα πρόβλεψης απώλειας τους είναι απαραίτητη. Στην εργασία αυτή χρησιμοποιούνται δύο διαφορετικοί συνδυασμοί μεθόδων για τη δημιουργία υβριδικών μοντέλων με τη πρώτη να είναι ο συνδυασμός συσταδοποίησης όπως είναι self-organizing maps (SOM) και ταξινόμησης όπως artificial neural networks (ANN) όπου είναι SOM+ANN και ο δεύτερος συνδυασμός είναι δύο τεχνικών ταξινόμησης (ANN), άρα είναι ANN+ANN. Για την αξιολόγηση της απόδοσης των μοντέλων αυτών λαμβάνονται υπόψιν τρία διαφορετικά είδη δοκιμαστικών σετ. Τα αποτελέσματα δείχνουν ότι τα δύο υβριδικά μοντέλα αποδίδουν καλύτερα από το απλό νευρωνικό δίκτυο και συγκεκριμένα το υβριδικό μοντέλο ANN+ ANN αποδίδει καλύτερα από το υβριδικό μοντέλο SOM+ANN. [9]

Οι *A De Caigny et al.* (2020) ερευνούν την επιπρόσθετη αξία των ενσωματωμένων δεδομένων κειμένου στα μοντέλα πρόβλεψης απώλειας πελατών συγκρίνοντας προηγούμενες έρευνες που χρησιμοποιούνται συνελικτικά νευρωνικά δίκτυα (convolutional neural networks ή CNN) με πρόσφατα καλύτερες πρακτικές για ανάλυση δεδομένων κειμένου στη πρόβλεψη απώλειας πελατών. Τα αποτελέσματα δείχνουν πως υπερτερούν οι προηγούμενες έρευνες, καθώς αρχικά η απόδοση πρόβλεψης βελτιώνεται με τη συμπερίληψη δεδομένων κειμένου σε μοντέλα πρόβλεψης απώλειας πελατών. Επιπλέον, τα CNN υπερισχύουν πρόσφατες καλύτερες πρακτικές για εξόρυξη δεδομένων στην πρόβλεψη απώλειας πελατών και τέλος τα μη δομημένα δεδομένα κειμένου δεν μπορούν να δημιουργήσουν μοντέλα πρόβλεψης όπως συμβαίνει με τα δομημένα.[10]

Οι *E. Sivasankar & J. Vijaya* (2019) μελετούν την απώλεια πελατών στις τηλεπικοινωνίες με τη χρήση ενός υβριδικού μοντέλου που χρησιμοποιεί hybrid probabilistic possibilistic fuzzy C-means clustering (PPFCM) με artificial neural network (PPFCM-ANN). Τα δεδομένα που λαμβάνονται ομαδοποιούνται μέσω PPFCM και μετά χρησιμοποιούνται για ταξινόμηση τα ANN. Εφαρμόζονται τρία σετ πειραμάτων όπου αρχικά χρησιμοποιείται ο αλγόριθμος συσταδοποίησης PPFCM, μετά το σετ που αξιολογεί το αποτέλεσμα της ταξινόμησης και το τρίτο σετ που αυθεντικοποιεί την παρουσίαση του προτεινόμενου υβριδικού μοντέλου. Το υβριδικό μοντέλο PPFCM-ANN παρέχει μέγιστη ακρίβεια όταν συγκρίνεται με ένα απλό μοντέλο.[11]

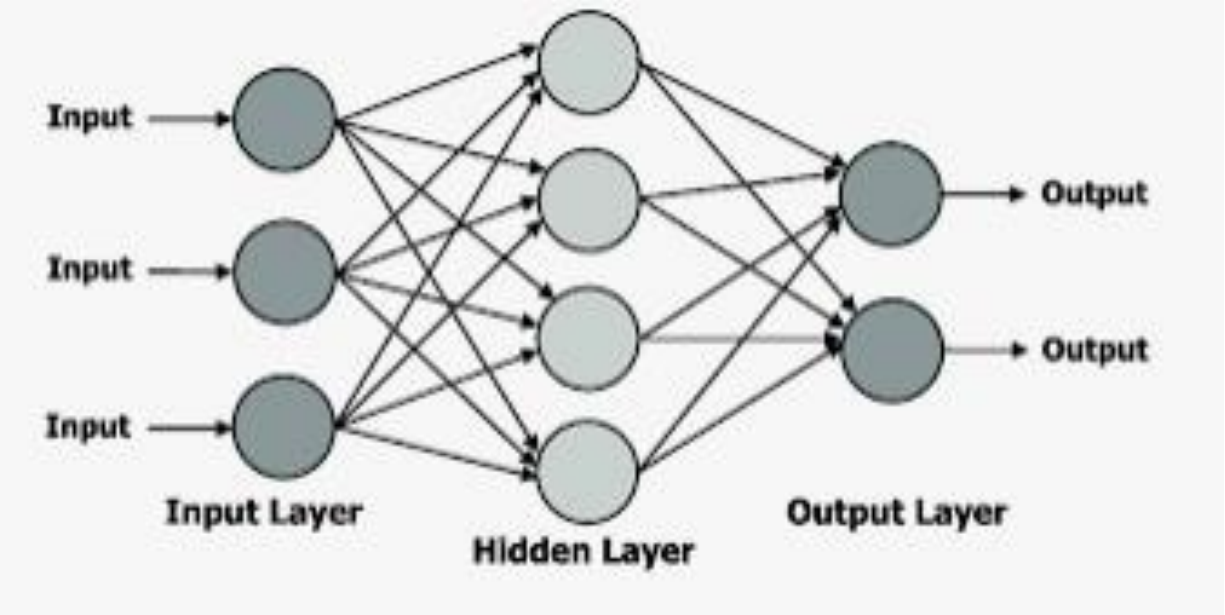
Οι *R. Prashanth et al.* (2017) ερευνούν την πρόβλεψη απώλειας πελατών με μεθόδους στατιστικής και εξόρυξης δεδομένων. Χρησιμοποιούν γραμμικές και μη γραμμικές τεχνικές της Random Forest και αρχιτεκτονική Βαθειάς Μάθησης όπως είναι τα Deep Neural Networks, Deep Belief Networks και Recurrent Neural Network για πρόβλεψη. Παρατηρείται ότι τα μη γραμμικά μοντέλα έχουν καλύτερα αποτελέσματα και τέτοια μοντέλα πρόβλεψης έχουν δυνατότητες χρήσης στη βιομηχανία τηλεπικοινωνίας για λήψη καλύτερων αποφάσεων και διαχείριση πελατών.[12]

Οι *J. Hu et al.* (2019) γνωρίζουν ότι η έγκαιρη πρόβλεψη απώλειας πελατών βοηθάει τους μαρκετίστες να διατηρήσουν υπάρχοντες και πολύτιμους πελάτες, καθώς και να αναπτύξουν ένα μοντέλο πρόβλεψης απώλειας πελατών. Στη συγκεκριμένη έρευνα χρησιμοποιούν Recurrent Neural Network με βάση το προϊόν (pRNN). Το RNN διαθέτει μονάδες μακροπρόθεσμης μνήμης και χρησιμοποιείται για να μάθει διαδοχικά μοτίβα από τα δεδομένα των πελατών τα οποία αλλάζουν με τον καιρό και η λειτουργία του προϊόντος εφαρμόζεται πριν το επαναλαμβανόμενο στρώμα ώστε να μάθει τη διασύνδεση μεταξύ των μεταβλητών. Το pRNN εφαρμόζεται σε πραγματική βάση δεδομένων στις τηλεπικοινωνίες και τα αποτελέσματα της έρευνας δείχνουν ότι αποδίδει καλύτερα από οποιοδήποτε άλλο μοντέλο.[13]

Κεφάλαιο 3: Νευρωνικά Δίκτυα

3.1 Γενικά για τα Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα και κυρίως τα τεχνητά νευρωνικά δίκτυα (ANN) μιμούνται μέσω αλγόριθμων τη λειτουργία του ανθρώπινου εγκεφάλου. Αποτελούνται από στρώματα, τα οποία είναι ένα στρώμα εισόδου, ένα ή περισσότερα κρυφά στρώματα και ένα στρώμα εξόδου. Κάθε στρώμα ενδέχεται να περιέχει ένα ή περισσότερους κόμβους, οι οποίοι συνδέονται με τον επόμενο και αφού τροποποιηθεί μέσω αυτών μεταδίδεται η πληροφορία. Οι κόμβοι διαθέτουν βάρη και τιμές κατωφλίου. Όταν η τιμή εξόδου ενός κόμβου είναι μεγαλύτερη από τη τιμή κατωφλίου ο κόμβος ενεργοποιείται και στέλνει δεδομένα στο επόμενο στρώμα του δικτύου, ενώ αν είναι μικρότερη της τιμής του κατωφλίου δεν στέλνονται δεδομένα στον επόμενο. Για την εκπαίδευση των νευρώνων, τα βάρη ρυθμίζονται ώστε να ελαχιστοποιηθεί η διαφορά μεταξύ της εξόδου και της αναμενόμενης τιμής.



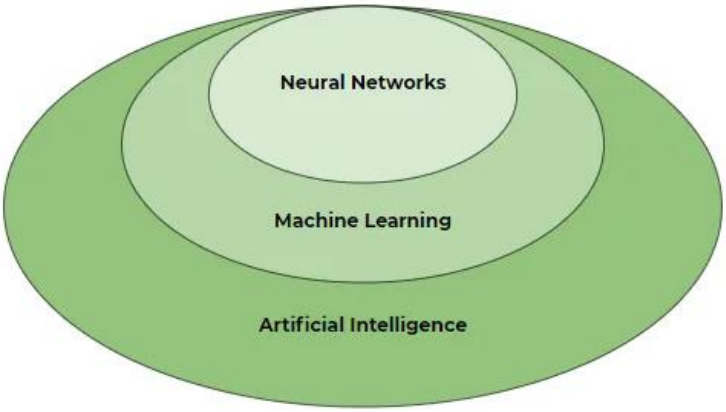
Εικόνα 1) Απεικόνιση ενός νευρωνικού δικτύου

Η Βαθειά Μάθηση αναφέρεται στον αριθμό των στρωμάτων των νευρωνικών δικτύων, καθώς νευρωνικό δίκτυο με πολλαπλό αριθμό κρυφών στρωμάτων μπορεί να θεωρηθεί Βαθειά Μάθηση και με αυτό τον τρόπο να κατανοήσει πιο περίπλοκα προβλήματα.

Η Βαθειά Μάθηση αποτελεί υποκατηγορία της Μηχανικής Μάθησης, καθώς είναι μέθοδος της χρησιμοποιώντας πολλαπλά κρυφά στρώματα. Η Μηχανική Μάθηση, περιλαμβάνει τη δημιουργία υπολογιστικών μοντέλων ικανών να μάθουν και να πραγματοποιούν προβλέψεις ή να παίρνουν

αποφάσεις, βάση των δεδομένων που τους παρέχονται. Με τη χρήση διαθέσιμων βάσεων δεδομένων, ένας αλγόριθμος Μηχανικής Μάθησης υποστηριζόμενος από ένα μαθηματικό μοντέλο μπορεί να πραγματοποιήσει προβλέψεις.

Τέλος, όλα τα παραπάνω αποτελούν υποκατηγορία της της Τεχνητής Νοημοσύνης (AI), η οποία ορίζεται από κάποιους ως έξυπνα συστήματα που λαμβάνουν αποφάσεις για την επίτευξη στόχων, ενώ άλλοι εστιάζουν στο ότι οι μηχανές μιμούνται διάφορες γνωστικές λειτουργίες. Χρησιμοποιεί προβλέψεις και αυτοματοποιήσεις ώστε να βελτιστοποιήσει και να λύσει εργασίες που έχουν πραγματοποιήσει άνθρωποι στο παρελθόν.[14]



Εικόνα 2) Νευρωνικά Δίκτυα- Μηχανική Μάθηση-Τεχνητή Νοημοσύνη

3.1.1 Εξόρυξη Δεδομένων (Data Mining)

Εξόρυξη δεδομένων ορίζεται ως η διαδικασία απόκτησης χρήσιμης πληροφορίας από ένα μεγάλο σε μέγεθος όγκο δεδομένων. Μία αποθήκη δεδομένων είναι μια τοποθεσία στην οποία αποθηκεύεται η πληροφορία. Ο τύπος των δεδομένων που αποθηκεύονται εξαρτάται σε μεγάλο βαθμό από το είδος της εταιρείας, όπου κάποιες αποθηκεύουν όλα τα δεδομένα, ενώ άλλες μόνο εκείνα που θεωρούν σημαντικά.

Για την εξόρυξη δεδομένων μπορούν να χρησιμοποιηθούν νευρωνικά δίκτυα. Υπάρχουν δύο κύριοι τύποι μοντέλων νευρωνικών δικτύων τα επιτηρούμενα και τα μη επιτηρούμενα μοντέλα. Ένα επιτηρούμενο νευρωνικό δίκτυο, χρησιμοποιεί εκπαιδευτικά και ελέγχου δεδομένα για να δημιουργήσει ένα μοντέλο, όπου τα εκπαιδευτικά δεδομένα χρησιμοποιούνται ώστε το μοντέλο να μάθει τον τρόπο για να μπορέσει να κάνει προβλέψεις και τα δεδομένα ελέγχου για την εγκυρότητα του μοντέλου. Τα

νευρωνικά δίκτυα γίνονται πολύ δημοφιλή για την εξόρυξη δεδομένων, λόγω της ικανότητας τους να προβλέψουν σε καλύτερο βαθμό από ότι άλλες στατιστικές μέθοδοι χρησιμοποιώντας αληθινές βάσεις δεδομένων.[15]

3.1.2 Υλοποίηση Νευρωνικών Δικτύων

Σε ένα τεχνητό νευρωνικό δίκτυο πραγματοποιούνται πολλές επεξεργασίες σε διάφορα στρώματα, όπου στο πρώτο στρώμα λαμβάνεται η εισαγωγή της πληροφορίας. Η πληροφορία μετά διαχέεται στα υπόλοιπα στρώματα και τέλος παράγεται η συνολική πληροφορία από το τελευταίο στρώμα. Κάθε στρώμα διαθέτει κόμβους επεξεργασίας οι οποίοι διαθέτουν την πληροφορία, την επεξεργάζονται και η έξοδος ενός κόμβου, αποτελεί την είσοδο για τον επόμενο. Αυτή η διαδικασία ακολουθείται μέχρι το στρώμα εξόδου που δείχνει το αποτέλεσμα του νευρωνικού δικτύου και στο οποίο μπορεί να υπάρχουν ένας ή περισσότεροι κόμβοι εκ των οποίων λαμβάνεται η πληροφορία. Ένα από τα πιο σημαντικά μέρη των κόμβων αποτελεί η συνάρτηση ενεργοποίησης (activation function), η οποία καθορίζει την έξοδο της πληροφορίας από τον κόμβο.

Τα τεχνητά νευρωνικά δίκτυα έχουν την ικανότητα να προσαρμόζονται, καθώς τροποποιούνται και μαθαίνουν διαρκώς μέσω εκπαίδευσης. Κάθε νευρώνας αποτελείται από μία τιμή κατωφλίου (threshold value), όπου λαμβάνει την πληροφορία και ανάλογα το βάρος που εκείνη έχει τη μεταφέρει στον επόμενο. Αν η πληροφορία που έχει πολλαπλασιαστεί με την τιμή του βάρους ξεπερνάει την τιμή κατωφλίου, τότε σαν έξοδο η πληροφορία λαμβάνει την τιμή 1, διαφορετικά έχει την τιμή 0.

Η εκπαίδευση των νευρωνικών δικτύων πραγματοποιείται θέτοντας για όλα τα βάρη μικρές τυχαίες τιμές με αποτέλεσμα το δίκτυο να αρχίζει να αποδίδει τυχαίες τιμές. Αυτό πραγματοποιείται πολλές φορές εκ των οποίων προκύπτουν διαφορετικά αποτελέσματα κάθε φορά. Αρχικά, μετριέται το τετράγωνο της διαφοράς μεταξύ της εξόδου και της επιθυμητής εξόδου. Το άθροισμα όλων των αριθμών που προκύπτουν από τις δοκιμές της εκπαίδευσης του δικτύου ονομάζεται συνολικό σφάλμα και όσο πιο μικρή είναι η τιμή του, τόσο καλύτερα αποδίδει το νευρωνικό δίκτυο. Επιλέγοντας τα βάρη που περιορίζουν το συνολικό σφάλμα λαμβάνεται ένα νευρωνικό δίκτυο ικανό να λύσει οποιοδήποτε πρόβλημα. Με τη μέθοδο της διάδοσης προς τα πίσω (backpropagation), τα βάρη και οι τιμές κατωφλίου αλλάζουν διαρκώς ώστε να ελαχιστοποιηθεί το συνολικό σφάλμα. [16]

Στην πράξη τα νευρωνικά δίκτυα λειτουργούν με συγκεκριμένη μεθοδολογία. Αρχικά γίνεται η επεξεργασία των δεδομένων, τα οποία στη συνέχεια χωρίζονται σε εκπαίδευσης και ελέγχου όπου ένα ποσοστό των δεδομένων θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου και το άλλο για έλεγχο. Στη συνέχεια θα χρησιμοποιηθεί στα νευρωνικά δίκτυα η διαμόρφωση και τροποποίηση των υπερπαραμέτρων (hyperparameter tuning) όπου είναι ο αριθμός των νευρώνων των κρυμμένων

στρωμάτων, η συνάρτηση ενεργοποίησης και ο solver. Τέλος, μέσω του πίνακα συσχέτισης, φανερώνεται το ποσοστό πρόβλεψης του μοντέλου, καθώς και ποια δεδομένα αντιστοιχήθηκαν σωστά ή λανθασμένα.

3.1.3 Τύποι Νευρωνικών Δικτύων

Τα νευρωνικά δίκτυα χρησιμοποιούνται ευρέως στην αναγνώριση εικόνων, στην επεξεργασία ομιλίας, στην αναγνώριση ομιλίας και στην πρόβλεψη των τιμών των μετοχών. Αποτελούνται από διασυνδεδεμένους νευρώνες και έχουν διάφορους τύπους όπως αναλύονται παρακάτω.

Το απλούστερο νευρωνικό δίκτυο είναι εκείνο που περιέχει ένα στρώμα εισόδου και ένα στρώμα εξόδου και συνθέτουν τα perceptrons. Ονομάζονται δίκτυα perceptrons όπου τα perceptrons ορίζουν τιμή ίση με 0 ή 1, ανάλογα το κατώφλι ενεργοποίησης και διαιρούν το σετ σε δύο μέρη.

Το νευρωνικό δίκτυο που διαθέτει πολλαπλά στρώματα, ονομάζεται Layered network ή feed forward στο οποίο η έξοδος του προηγούμενου στρώματος νευρώνων, αποτελεί είσοδο για το επόμενο και η πληροφορία μεταφέρεται προς τα εμπρός. Βρίσκουν εφαρμογή σε περιοχές επιτηρούμενης μάθησης και χρησιμοποιούνται στην αναγνώριση εικόνων, κειμένου και ομιλίας.

Τα νευρωνικά δίκτυα, στα οποία χρησιμοποιείται ανατροφοδότηση των βρόχων όπου το σήμα εξόδου επιστρέφει στους νευρώνες εισόδου, ονομάζονται Recurrent networks (RNN) και μπορούν να παράγουν αλληλουχίες φαινομένων έως ότου η έξοδος σταθεροποιηθεί. Χρησιμοποιούνται για την παραγωγή κειμένου.

Τα Convolutional neural networks (CNN), αποτελούνται από στρώματα εισόδου, splines, βοηθητικά και εξόδου. Ονομάζονται και πλεγμένα δίκτυα και από μη γραμμικές λειτουργίες εξάγεται η πληροφορία που σχετίζεται με τα χαρακτηριστικά της εικόνας. Χρησιμοποιούνται για την ανάλυση μοτίβων και η ακρίβεια τους αυξομειώνεται, βάση τον αριθμό των κρυφών στρωμάτων που επιλέγονται.

Τέλος, τα δίκτυα Gated Recurrent Unit (GRU) και Long short-term memory (LSTM) εκτελούν αναδρομικές εργασίες όπου η έξοδος εξαρτάται από προηγούμενους υπολογισμούς. Διαθέτουν μνήμη, όπου τους επιτρέπει να θυμούνται καταστάσεις δεδομένων σε διαφορετικές φάσεις. Η εκπαίδευση τους διαρκεί περισσότερο και έχουν εφαρμογή στην πρόβλεψη της τροχιάς των αυτόνομων αυτοκινήτων, στη μετατροπή κειμένου σε ομιλία και στη μετάφραση γλωσσών. [17]

3.1.4 Εφαρμογές Νευρωνικών Δικτύων

Τα τελευταία χρόνια τα νευρωνικά δίκτυα έχουν γίνει αρκετά δημοφιλή και βοηθητικά μοντέλα για ταξινόμηση, συσταδοποίηση, αναγνώριση μοτίβων και πρόβλεψη σε πολλούς τομείς. Ένα πλεονεκτήματα εφαρμογής τους είναι ότι αποτελεί ένα πολύ χρήσιμο και καινοτόμο μοντέλο για τη λύση

προβλημάτων και για αυτό χρησιμοποιείται σε πολλούς τομείς. Είναι ικανά να διαχειριστούν προβλήματα στη γεωργία, στις επιστήμες υγείας, στην εκπαίδευση, στα οικονομικά, στη διοίκηση, στην ασφάλεια, στη μηχανική, στη τέχνη και στο εμπόριο. Επιπλέον μπορούν να λύσουν προβλήματα κατασκευαστικά, μεταφορών, ασφάλειας υπολογιστών, τραπεζικά, διοίκησης επιχειρήσεων, μάρκετινγκ, ενέργειας και γενικά προβλήματα που δεν μπορούν να λυθούν με κλασσικές μεθόδους και μαθηματικά.

Γενικά, τα νευρωνικά δίκτυα εφαρμόζονται ευρέως στην ακαδημαϊκή κοινότητα και στις βιομηχανίες. Αρχικά, μπορούν να αντιμετωπιστούν προβλήματα αναγνώρισης φωνής και μοτίβων. Η ικανότητα τους να αντιμετωπίσουν προβλήματα αναγνώρισης φωνής είναι κατανοητή και η ανάπτυξη τους έχει δώσει καινοτόμους τρόπους αντιμετώπισης των προβλημάτων της αναγνώρισης μοτίβων. Επιπλέον, έχουν αντιμετωπιστεί προβλήματα computer vision, το οποίο αποτελεί την ικανότητα των υπολογιστών να λειτουργούν όπως ο ανθρώπινος εγκέφαλος και έχει ως σκοπό την δημιουργία υπολογιστών ικανών να κατανοούν και να επεξεργάζονται δεδομένα που λαμβάνουν από εικόνες και βίντεο. Επίσης, έχει δώσει τη λύση σε προβλήματα αναγνώρισης προσώπων. Τέλος, έχει λύση προβλήματα ανίχνευσης ασθενειών στο τομέα της υγείας, αλλά ανίχνευσης και οικονομικών εγκλημάτων. [18]

3.2 Είδη Μάθησης Νευρωνικών Δικτύων

Ένα νευρωνικό δίκτυο πρέπει να ρυθμιστεί έτσι ώστε ένα σετ εισόδου να παράγει τα επιθυμητά δεδομένα εξόδου. Υπάρχουν πολλοί μέθοδοι να ρυθμιστούν οι συνδέσεις με έναν από αυτούς να είναι να ρυθμίζονται τα βάρη βάση πρότερης γνώσης. Υπάρχει όμως κι ένας εναλλακτικός τρόπος με τον οποίο το νευρωνικό δίκτυο εκπαιδεύεται, τροφοδοτώντας το με διδαγμένα μοτίβα και αφήνοντας το να αλλάξει τα βάρη του, βάση κάποιων κανόνων μάθησης. Τα είδη της μάθησης των νευρωνικών δικτύων είναι η επιτηρούμενη μάθηση (supervised learning), η μη επιτηρούμενη μάθηση (unsupervised learning) και η ενισχυτική μάθηση (reinforcement learning) που αναλύονται παρακάτω.

Στην επιτηρούμενη μάθηση, στο δίκτυο παρέχονται οι είσοδοι και αντιστοιχούνται με μοτίβα εξόδων. Αυτή η αντιστοίχιση παρέχεται από έναν εξωτερικό παράγοντα ή και από το σύστημα το ίδιο που περιλαμβάνει το δίκτυο (self-supervised).

Στη μη επιτηρούμενη μάθηση, η έξοδος εκπαιδεύεται ώστε να ανταποκρίνεται σε συστάδες του μοτίβου στην είσοδο. Με αυτό το τρόπο πρέπει να ανακαλύψει τα κυριότερα χαρακτηριστικά της εισόδου. Σε αντίθεση με την επιτηρούμενη μάθηση, δεν υπάρχει εκ των προτέρων ένα σύνολο στο οποίο τα μοτίβα πρέπει να ταξινομηθούν, αλλά το σύστημα πρέπει από μόνο του να ανακαλύψει την είσοδο.

Η ενισχυτική μάθηση θεωρείται ως μια ενδιάμεση μορφή των δύο προηγούμενων τύπων μάθησης. Σε αυτό το τύπο μάθησης, η μηχανή μάθησης αντιδρά στο περιβάλλον και λαμβάνει η ίδια αντίδραση από αυτό. Το σύστημα μάθησης βαθμολογεί τη δράση του θετικά ή αρνητικά βασισμένο στην αντίδραση του περιβάλλοντος και αντιδρά ανάλογα ρυθμίζοντας τις παραμέτρους του.[19]

3.3 Δεδομένα (Data)

Ως δεδομένα ορίζονται τα γεγονότα ή τα στατιστικά στοιχεία που λαμβάνονται για ανάλυση και το σύνολο των δεδομένων που αποτελεί μια βάση δεδομένων ονομάζεται dataset. Τα δεδομένα αποτελούν μια απλή μορφή γνώσης, τα οποία από μόνα τους δεν έχουν κάποια σημασία και για να έχουν νόημα πρέπει να αναλυθούν, να οργανωθούν και να ερμηνευτούν. Χωρίζονται σε ποσοτικά (quantitative) και ποιοτικά (qualitative).

Τα ποσοτικά δεδομένα αναφέρονται σε πληροφορία η οποία μπορεί να μετρηθεί όπως είναι οι αριθμοί. Η ποιοτική έρευνα βασίζεται στη συλλογή και ερμηνεία αριθμητικών δεδομένων καθώς βασίζεται στη μέτρηση και γενίκευση των αποτελεσμάτων.

Σε αντίθεση με το ποσοτικά δεδομένα, τα ποιοτικά δεδομένα είναι περιγραφικά, συνεπώς δεν μπορούν να μετρηθούν. Αναφέρονται σε λέξεις που περιγράφουν συγκεκριμένα χαρακτηριστικά. Δεν ασχολούνται με τη μέτρηση των δεδομένων, αλλά πιο πολύ στοχεύουν στην περιγραφή τους.

3.3.1 Προετοιμασία Δεδομένων (Data Preprocessing)

Η προεπεξεργασία των δεδομένων αποτελεί πολύ σημαντικό βήμα στη διαδικασία εξόρυξης δεδομένων και αναφέρεται στον καθαρισμό, μετατροπή και ενσωμάτωση των δεδομένων ώστε να είναι έτοιμα για ανάλυση. Στόχος της προεπεξεργασίας των δεδομένων είναι να βελτιώσει την ποιότητα των δεδομένων και να τα κάνει κατάλληλα για την εξόρυξή τους.

Η προεπεξεργασία των δεδομένων είναι μια τεχνική εξόρυξης δεδομένων που χρησιμοποιείται για να μετατρέψει τα απλά δεδομένα σε μια χρήσιμη και αποτελεσματική μορφή. Τα βήματα που περιέχονται στην προετοιμασία των δεδομένων για την εξόρυξη δεδομένων είναι:

1) Καθαρισμός των δεδομένων (**data cleaning**), που χρησιμοποιείται για την συμπλήρωση των δεδομένων που λείπουν, την αναγνώριση των ακραίων και την ομαλοποίηση των θορυβώδη δεδομένων, την διόρθωση ασυνεπή δεδομένων και την επίλυση του πλεονασμού των δεδομένων που προκαλείται από την ενσωμάτωση των δεδομένων .

Α) Δεδομένα που λείπουν και ο χειρισμός τους πραγματοποιείται με διάφορους τρόπου όπως:

- Αγνοώντας τις πλειάδες, που είναι μια προσέγγιση όταν η βάση δεδομένων που διατίθεται είναι αρκετά μεγάλη και πολλαπλές τιμές απουσιάζουν.
- Συμπληρώνοντας τις τιμές που απουσιάζουν, το οποίο γίνεται με ποικίλους τρόπους με έναν από αυτούς να είναι χειροκίνητα είτε και αυτόματα, συνεισφέροντας την πιο πιθανή τιμή. Ωστόσο, αυτό δεν μπορεί να επιτευχθεί για μεγάλου μεγέθους βάση δεδομένων.

Β) Θορυβώδη δεδομένα, τα οποία είναι ασήμαντα δεδομένα που δεν μπορούν να ερμηνευτούν από τις μηχανές και μπορούν να χειριστούν με τους εξής τρόπους:

- Μέθοδος Bining που λειτουργεί σε ταξινομημένα δεδομένα και τμήματα τα οποία χωρίζονται σε ίδια τμήματα για να τα εξομαλύνει. Όλα τα δεδομένα χωρίζονται σε τμήματα ίδιου μεγέθους και διάφορες μέθοδοι πραγματοποιούνται για να ολοκληρωθεί η διαδικασία.
- Παλινδρόμηση όπου εδώ τα δεδομένα εξομαλύνονται προσαρμόζοντάς τα σε λειτουργία παλινδρόμησης που μπορεί να είναι γραμμική ή πολλαπλών ανεξάρτητων μεταβλητών
- Συσταδοποίηση όπου ανιχνεύονται και αφαιρούνται οι ακραίες τιμές.
- Συνδυαστική επιθεώρηση υπολογιστή και ανθρώπου ώστε να εντοπιστούν ύποπτες τιμές και να γίνει έλεγχος από άνθρωπο

2) Μετατροπή των δεδομένων (**data transformation**) που χρησιμοποιείται για μετατροπή των δεδομένων σε πιο κατάλληλες μορφές για εξόρυξη δεδομένων και περιλαμβάνει τους εξής τύπους:

- Ομαλοποίηση που τοποθετεί τις τιμές των δεδομένων σε συγκεκριμένο εύρος (από -1 έως 1 ή από 0 έως 1).
- Aggregation, η μέθοδος η οποία διαθέτει την σύνοψη των δεδομένων και την κατασκευή τμημάτων δεδομένων
- Η εξομάλυνση που αφαιρεί τον ήχο από τα δεδομένα
- Generalization που αποτελεί την έννοια της αναρρίχησης στην ιεραρχία
- Κατασκευή χαρακτηριστικών, όπου νέα χαρακτηριστικά δημιουργούνται από τα ήδη δοσμένα.

3) Ενσωμάτωση των δεδομένων (**data integration**), η οποία είναι μια μέθοδος που ενσωματώνει τα δεδομένα που αποκτούνται από πολλές και ποικίλες πηγές και τα αποθηκεύει όλα μαζί στην ίδια αποθήκη δεδομένων

4) Μείωση δεδομένων (**data reduction**) που αποτελεί σημαντικό βήμα της διαδικασίας εξόρυξης δεδομένων και περιλαμβάνει τη μείωση του μεγέθους του συνόλου δεδομένων, διατηρώντας τη σημαντική πληροφορία, καθώς έτσι βελτιώνεται η αποτελεσματικότητα της ανάλυσης δεδομένων και αποφεύγεται η υπερπροσαρμογή του μοντέλου. Κάποιες μέθοδοι μείωσης δεδομένων είναι:

- Data cube aggregation, όπου στη συγκεκριμένη προσέγγιση κατασκευάζεται ένα πλέγμα δεδομένων εφαρμόζοντας λειτουργίες πρόσθεσης των δεδομένων χωρίς να χαθεί σημαντική πληροφορία
- Με συμπίεση των δεδομένων όπου αφαιρούνται περιττά δεδομένα

- Μείωση του αριθμού των δεδομένων, στην οποία τα δεδομένα συνολικά παρουσιάζονται σε λιγότερο αριθμό με εναλλακτικές μορφές όπου γίνεται με παραμετρικές και μη παραμετρικές μεθόδους,
- Με διακριτοποίηση και την παραγωγή της έννοιας της ιεραρχίας

5) Η διακριτοποίηση των δεδομένων (**data discretization**) στην οποία μειώνεται ο αριθμός για δοσμένα συνεχή χαρακτηριστικά, διαιρώντας το εύρος των χαρακτηριστικών σε διαστήματα και η παραγωγή της έννοιας της ιεραρχίας (concept hierarchy generation) όπου μειώνεται ο αριθμός των δεδομένων αντικαθιστώντας χαμηλότερου επιπέδου δεδομένα με αντίστοιχα υψηλότερου. Διαθέτουν τις εξής τεχνικές:

- Μέθοδος Binning
- Ιστόγραμμα, όπου διαχωρίζονται οι τιμές κουτιά και αποθηκεύονται οι μέσοι όροι για το καθένα.
- Με βάση την εντροπία, στην οποία χρησιμοποιώντας πληροφορίες των κλάσεων, μειώνει το μέγεθος των δεδομένων.
- Συσταδοποίηση, με την οποία διαχωρίζονται τα δεδομένα σε συστάδες και αφαιρεί τις ακραίες τιμές.
- Με τη μέθοδο της συγχώνευσης διαστημάτων (interval merge) μέσω της ανάλυσης χ^2, στην οποία γίνεται η εύρεση των κοντινών διαστημάτων και η συγχώνευση τους για τη δημιουργία μεγαλύτερου διαστήματος όπου αυτό γίνεται μέχρι να προκύψει ένα προκαθορισμένο κριτήριο σταματημού.
- Τμηματοποίηση με φυσικό τρόπο όπου μέσω του κανόνα 3-4-5 τμηματοποιούνται τα αριθμητικά δεδομένα σε φυσικά διαστήματα. [20]

3.4 Διαχωρισμός Δεδομένων

3.4.1 Train/test split

Ο διαχωρισμός των δεδομένων είναι μια τεχνική απαραίτητη για τη μείωση της μεροληψίας που εμφανίζεται στα εκπαιδευτικά δεδομένα, αλλά και για την δημιουργία ενός μοντέλου ικανού να γενικεύει σε άγνωστα δεδομένα.. Αυτή η διεργασία πραγματοποιείται από αναλυτές δεδομένων ώστε να μην υπάρξει υπερπροσαρμογή των αλγόριθμων που θα οδηγήσει σε άσχημα κακή απόδοση των δεδομένων ελέγχου.

Ο διαχωρισμός των δεδομένων είναι η τμηματοποίηση των δεδομένων σε υποκατηγορίες, όπου η εκπαίδευση του μοντέλου και η αξιολόγηση του πραγματοποιούνται ξεχωριστά. Δηλαδή, ένα ποσοστό

της βάσης δεδομένων χρησιμοποιείται για την εκπαίδευση του μοντέλου και το υπόλοιπο για τον έλεγχο της απόδοσης του. Ο διαχωρισμός των δεδομένων, εξαρτάται από το μέγεθος της βάσης δεδομένων και είναι ικανός να επηρεάσει την απόδοση του μοντέλου, αλλά και την ίδια την απόδοση της ταξινόμησης. Για αυτό πρέπει να γίνει ο διαχωρισμός με διαφορετικές αναλογίες ποσοστών, ώστε να βρεθεί η αναλογία εκείνη που διαχωρισμός τη βάση δεδομένων, για τον οποίο το μοντέλο αποδίδει καλύτερα.[21]

3.4.2 k-fold cross validation

Το μοντέλο εκπαιδεύεται με τη μέθοδο cross validation. Πριν την εκπαίδευση του μοντέλου τα δεδομένα χωρίζονται σε εκπαιδευτικά και ελέγχου. Στη διαδικασία αυτή τα εκπαιδευτικά δεδομένα χωρίζονται σε δεδομένα σε δύο τμήματα, όπου το ένα χρησιμοποιείται για την εκπαίδευση του μοντέλου και το άλλο για την αξιολόγηση του. Η μέθοδος k-fold cross validation είναι η πιο γνωστή μέθοδος αυτής της διαδικασίας. Συγκεκριμένα, στη μέθοδο k-fold cross validation η βάση δεδομένων χωρίζεται σε k τμήματα με μια επαναληπτική προσέγγιση. Ο αριθμός της τιμής του k δείχνει τον αριθμό των επαναλήψεων. Αρχικά μία κατηγορία χρησιμοποιείται για έλεγχο και οι υπόλοιπες k-1 για την εκπαίδευση του μοντέλου σε κάθε επανάληψη. Αυτή η διαδικασία γίνεται μέχρι τη χρήση όλων των τμημάτων για τον έλεγχο του μοντέλου. Στο τέλος, σε κάθε επανάληψη έχουν αποθηκευτεί οι αποδόσεις του μοντέλου και ο μέσος όρος τους δίνει τα αποτελέσματα.[22]

3.5 Hyperparameter Tuning – Grid Search

Οι υπερπαράμετροι είναι μεταβλητές του μοντέλου εξωτερικά από αυτό, των οποίων η τιμή δεν μπορεί να καθοριστεί από τη βάση δεδομένων, αλλά καθορίζεται από το χρήστη ή μέσω δοκιμών και λάθους μέχρι να επιτευχθεί μία επιθυμητή ακρίβεια . Τέτοιου είδους παράμετροι αποτελούν: η τιμή C στο Support Vector Machines, η τιμή του k στο k-nearest neighbor και ο αριθμός των κρυφών στρωμάτων στα Νευρωνικά Δίκτυα.

Οι παράμετροι είναι εσωτερικές μεταβλητές του μοντέλου των οποίων οι τιμές μαθαίνονται ή υπολογίζονται από τα δεδομένα. Το μοντέλο χρησιμοποιεί μια προσέγγιση βελτιστοποίησης για τον υπολογισμό των παραμέτρων. Οι μεταβλητές αυτές δεν μπορούν να οριστούν από χρήστη ή ειδικό και χρησιμοποιούνται στη διαδικασία εκπαίδευσης του μοντέλου.

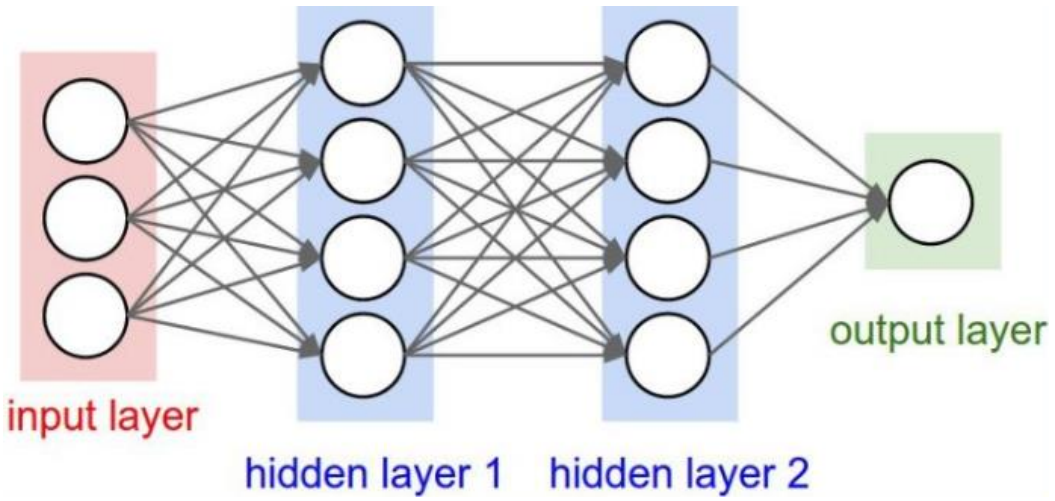
Η μέθοδος Grid-Search χρησιμοποιείται για την εύρεση των βέλτιστων υπερπαραμέτρων του μοντέλου, που θα δώσει σαν αποτέλεσμα τις πιο ορθές προβλέψεις. Υπολογίζει την απόδοση του μοντέλου για κάθε συνδυασμό των τιμών των υπερπαραμέτρων και επιλέγει τη βέλτιστη τιμή τους. Αυτό γίνεται, καθώς θεωρείται πως όλοι οι παράμετροι έχουν την ίδια πιθανότητα να επηρεάσουν την διαδικασία. Είναι μια μέθοδος αξιολόγησης του μοντέλου μέσω της οποίας γίνεται ο συντονισμός των υπερπαραμέτρων ώστε το νευρωνικό δίκτυο να αποδίδει καλύτερα αποτελέσματα. Στα νευρωνικά δίκτυα

υπερπαράμετροι αποτελούν: ο αριθμός των νευρώνων των κρυφών στρωμάτων, η συνάρτηση ενεργοποίησης και ο solver.[23]

3.5.1 Αριθμός Νευρώνων Κρυφών Στρωμάτων

Στα τεχνητά νευρωνικά δίκτυα, τα κρυφά στρώματα βρίσκονται μεταξύ των στρωμάτων εισόδου και εξόδου, όπου οι τεχνητοί νευρώνες λαμβάνουν ένα σύνολο σταθμισμένων εισροών και παράγουν μια εκροή μέσω της συνάρτησης ενεργοποίησης

Τα κρυφά στρώματα νευρωνικών δικτύων στήνονται με αρκετούς διαφορετικούς τρόπους. Είναι πολύ σημαντικό να βρεθεί ο κατάλληλος αριθμός των νευρώνων που χρησιμοποιούνται στα νευρωνικά δίκτυα, καθώς αν υπάρχουν λίγοι νευρώνες είναι πολύ πιθανό να υπάρξει υποπροσαρμογή του μοντέλου, ενώ αν είναι μεγάλος ο αριθμός τους ενδέχεται να υπάρξει το φαινόμενο της υπερπροσαρμογής του μοντέλου. Για την εύρεση του καλύτερου αριθμού των νευρώνων στα κρυφά στρώματα πρέπει να ληφθούν υπόψιν: ο αριθμός των νευρώνων εισόδου και εξόδου, τον αριθμό των εκπαιδευτικών δεδομένων, την ποσότητα του θορύβου στα δεδομένα στόχου, την πολυπλοκότητα της διεργασίας ή της ταξινόμησης που πρέπει να διδαχτεί το μοντέλο, την αρχιτεκτονική του τύπου της συνάρτησης ενεργοποίησης των νευρώνων και τον αλγόριθμο εκπαίδευσης.[24]



Εικόνα 3) Παράδειγμα στρωμάτων εισόδου, εξόδου και κρυφών ενός νευρωνικού δικτύου

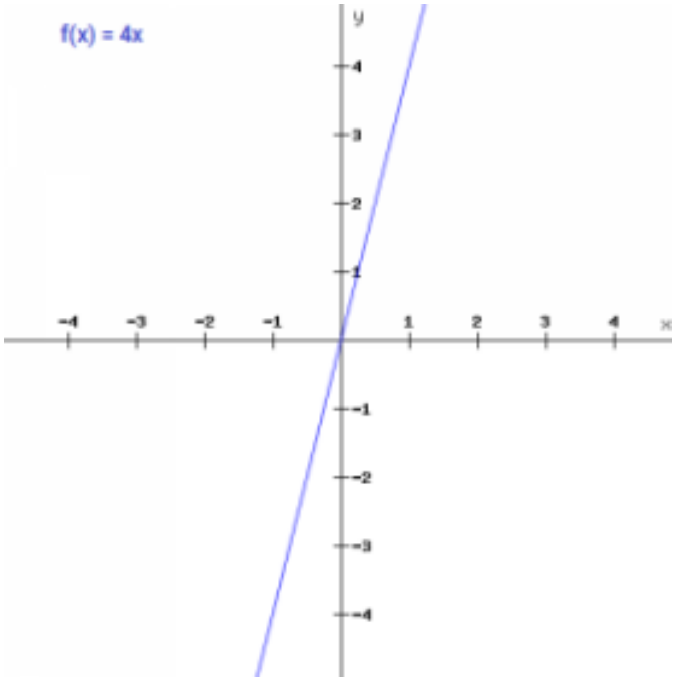
3.5.2 Activation Function (Συνάρτηση Ενεργοποίησης)

Τα νευρωνικά δίκτυα, αποτελούνται από πολλά στρώματα καθένα από το οποίο διαθέτει κόμβους οι οποίοι χρησιμοποιούνται για ταξινόμηση ή πρόβλεψη του μοντέλου από τα δεδομένα που δέχεται ως είσοδο. Ο κάθε κόμβος διαθέτει βάρη, βάση των οποίων επεξεργάζεται η πληροφορία για να μεταδοθεί από ένα στρώμα στο επόμενο. Με τις συναρτήσεις ενεργοποίησης, αποφασίζετε αν ένας νευρώνας θα πρέπει να ενεργοποιηθεί ή όχι, αν η πληροφορία που μεταφέρει είναι σημαντική. Η

ακρίβεια πρόβλεψης του μοντέλου εξαρτάται από τον τύπο της συνάρτησης ενεργοποίησης που χρησιμοποιείται. Αν δεν χρησιμοποιούνταν συναρτήσεις ενεργοποίησης τότε η πληροφορία θα διαχέεται με απλό γραμμικό τρόπο μέσα στο σύστημα οπότε δεν θα είναι εφικτή η χρήση του μοντέλου σε πιο δύσκολες διαδικασίες. Οι τύποι των Συναρτήσεων Ενεργοποίησης παρουσιάζονται παρακάτω:

Α) Γραμμικές Συναρτήσεις Ενεργοποίησης (Linear Activation Function)

Στις γραμμικές συναρτήσεις ενεργοποίησης ή αλλιώς **Identity Activation Function**, η συνάρτηση είναι γραμμική, η έξοδος από τη συνάρτηση δεν βρίσκεται μεταξύ κάποιου εύρους, η μαθηματική πράξη είναι **f(x)=ax** και το εύρος μείον άπειρο έως άπειρο. Η τιμή του a μπορεί να οριστεί από τον χρήστη. Λόγω της πολυπλοκότητας ή της ποικιλίας των παραμέτρων που τροφοδοτούν συνήθως τα νευρωνικά δίκτυα δεν μπορεί να χρησιμοποιηθεί. Εφαρμόζεται κυρίως για απλούστερες διεργασίες.



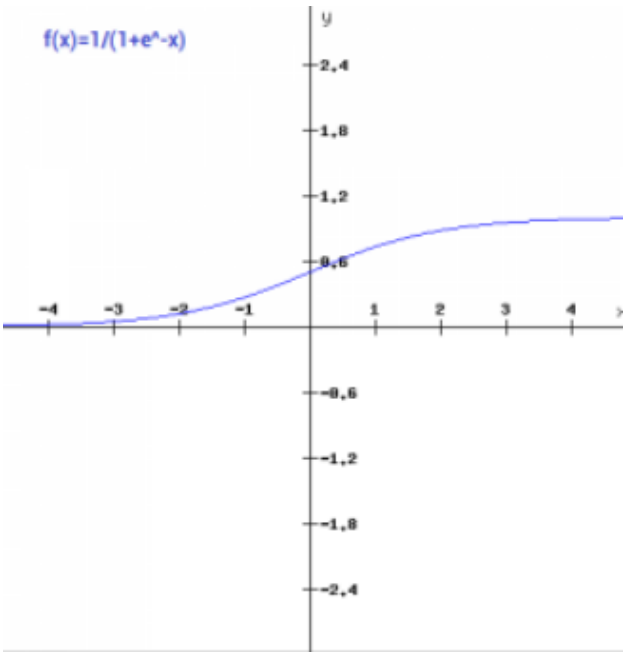
Γράφημα 1) Γραμμική Συνάρτηση Ενεργοποίησης

Β) Μη Γραμμικές Συναρτήσεις Ενεργοποίησης (Non-Linear Activation Function)

Όσον αφορά τις μη γραμμικές Συναρτήσεις Ενεργοποίησης, είναι πιο συνηθισμένη η χρήση τους, καθώς το μοντέλο εύκολα μπορεί να γενικεύει ή να προσαρμόζεται σε μια ποικιλία δεδομένων και να διαφοροποιείται κατά την έξοδο. Είναι πολύ σημαντικό οι συναρτήσεις ενεργοποίησης να είναι

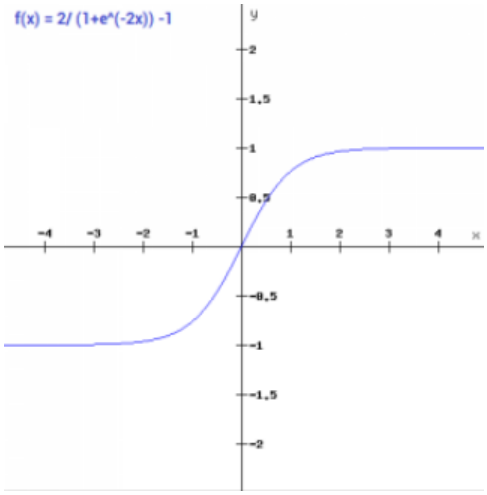
διαφοροποίησιμες, καθώς με αυτό τον τρόπο μπορούν να χρησιμοποιηθούν τεχνικές όπως η backpropagation, οι οποίες ρυθμίζουν τα βάρη για την βελτιστοποίηση του αλγορίθμου. Αποτελούν τις παρακάτω:

1) **Σιγμοειδής ή Λογιστική Συνάρτηση Ενεργοποίησης**: Είναι η πιο διαδεδομένη συνάρτηση ενεργοποίησης, καθώς είναι μη γραμμική. Η καμπύλη έχει σχήμα S, ο κύριος λόγος χρήσης της είναι ότι λαμβάνει τιμές μεταξύ 0 και 1 και χρησιμοποιείται σε μοντέλα όπου πρέπει να προβλεφθεί σαν έξοδο η πιθανότητα. Η λειτουργία είναι συνεχώς διαφορική. Μπορεί να οριστεί ως f(x)= $\frac{1}{e^{-x}}$



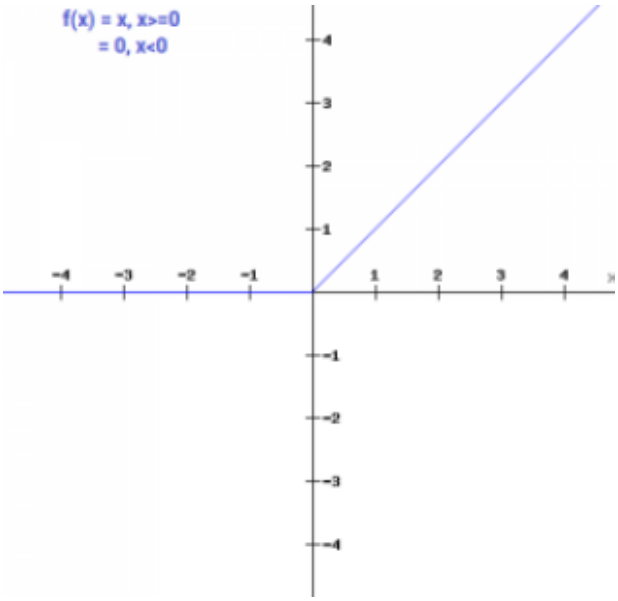
Γράφημα 2) Σιγμοειδής Συνάρτηση Ενεργοποίησης

2)**Tanh ή εφαπτομένης Συνάρτηση Ενεργοποίησης** η οποία μοιάζει με τη λογιστική, αλλά είναι συμμετρική γύρω από την αρχή των αξόνων, είναι συνεχής και διαφορική και έχει εύρος τιμών από -1 έως 1. Το πλεονέκτημα είναι ότι αρνητικές είσοδοι θα χαρτογραφούνται αρνητικά και μηδενικές κοντά στο 0 στο γράφημα tanh. Μπορεί να οριστεί ως f(x)=2sigmoid(2x)-1.



Γράφημα 3) Σύγκριση Tanh με Σιγμοειδή Λειτουργία Ενεργοποίησης

3)**Συνάρτηση Ενεργοποίησης ReLU (Rectified Linear Unit)**, χρησιμοποιείται ευρύτερα από οποιαδήποτε άλλη συνάρτηση ενεργοποίησης. Το πλεονέκτημα της συγκεκριμένης συνάρτησης είναι ότι δεν ενεργοποιούνται όλοι οι νευρώνες ταυτόχρονα το οποίο την κάνει πιο αποτελεσματική. Έχει εύρος τιμών από 0 έως άπειρο, καθώς μπορεί να οριστεί μαθηματικά ως f(x)=max (0,x). Σε κάποιες περιπτώσεις, η τιμή της βαθμίδας είναι μηδενική καθώς τα βάρη δεν έχουν ενημερωθεί κατά τη μέθοδο backpropagation.[25]



Γράφημα 4) Σύγκριση Σιγμοειδούς Λειτουργίας Ενεργοποίησης (αριστερά) με ReLU (δεξιά)

3.5.3 Solver

Στα νευρωνικά δίκτυα η χρήση του Solver έχει ως στόχο τη βελτιστοποίηση του μοντέλου, καθώς με τη χρήση τους, βρίσκονται σχέσεις μεταξύ των τιμών εισόδου και εξόδου Αποτελούνται από τους αλγόριθμους βελτιστοποίησης πρώτης σειράς όπως είναι ο Stochastic Gradient Descent (**SGD**) και ο **ADAM** από αλγόριθμους δεύτερη σειράς όπως είναι ο **L-BFGS-B**.

Μέθοδοι πρώτης σειράς είναι εκείνοι που για την εύρεση του βέλτιστου χρησιμοποιούν μόνο την κλίση ή την πρώτη παράγωγο. Οι πιο χαρακτηριστικοί αλγόριθμοι είναι ο SGD που ακολουθεί την πορεία της αρνητικής βαθμίδας της αντικειμενικής λειτουργίας και ο αλγόριθμος ADAM που είναι πιο περίπλοκος και λιγότερο ευαίσθητος σε θορυβώδης και μικρές κλίσεις. Ο SGD στηρίζεται στο γεγονός ότι η κλίση της λειτουργίας L δείχνει την κατεύθυνση της μέγιστης βελτίωσης, οπότε κινούμενοι στην αντίθετη κατεύθυνση της βαθμίδας μπορεί να επιτευχθεί βελτίωση στη τιμή της στοχευμένης λειτουργίας. Η στοχευμένη λειτουργία (L) μπορεί να γραφτεί ως το άθροισμα των λειτουργιών. Αν και απλή μέθοδος είναι αρκετά επιτυχημένη, ωστόσο ο αλγόριθμος αντιμετωπίζει προβλήματα σε περιοχές που η κλίση είναι μικρή αν και το ποσοστό μάθησης είναι μικρό ή εάν υπάρχει περιοχή με μεγάλη καμπυλότητα όπου πραγματοποιεί μεγάλα βήματα χάνοντας αρκετές πληροφορίες από εκείνη την περιοχή. Αυτά τα προβλήματα έρχεται να επιλύσει ο αλγόριθμος ADAM (Adaptive Moments). Είναι μέθοδος βελτιστοποίησης πρώτης σειράς για στοχαστικές αντικειμενικές λειτουργίες η οποία βασίζεται σε προσαρμοστικές εκτιμήσεις της πρώτης και της δεύτερης σειράς. Είναι πολύ διαδεδομένος αλγόριθμος για την εκπαίδευση των νευρωνικών δικτύων και δείχνει να αποδίδει καλά αποτελέσματα σε διάφορα προβλήματα.

Οι μέθοδοι δεύτερης σειράς χρησιμοποιούν την κλίση για την εύρεση της βέλτιστης τιμής σε συνδυασμό με τον Χεσιανό πίνακα H^2 ή κάποια προσέγγισή του για λογαριασμό της καμπύλης της αντικειμενικής λειτουργίας. Μία μέθοδος δεύτερης σειράς αποτελεί ο αλγόριθμος L-BFGS (Limited-Memory Broyden-Fletcher-Goldfarb-Shanno) ο οποίος είναι παρόμοιος με τον αλγόριθμο BFGS. Στον αλγόριθμο BFGS υπολογίζεται επαναληπτικά ο πίνακας M που προσεγγίζει τον ανάστροφο του Hessian της αντικειμενικής λειτουργίας. Ο προηγούμενος πίνακας M αποθηκεύεται για την επόμενη επανάληψη το οποίο αποτελεί μεγάλο περιορισμό για τα προβλήματα που προσεγγίζονται από αυτόν τον αλγόριθμο, καθώς εκπαιδεύοντας τα νευρωνικά δίκτυα, υπάρχουν πάρα πολλά βάρη και μεροληψίες που κάνουν την αποθήκευση του πίνακα μεταξύ κάθε επανάληψης ακατόρθωτη. Για αυτό χρησιμοποιείται πιο συχνά ο αλγόριθμος L-BFGS για τα νευρωνικά δίκτυα που χρειάζεται λιγότερη μνήμη, καθώς δεν αποθηκεύεται ολόκληρος ο πίνακας. Για αυτό το λόγο, υπάρχει πιθανότητα ο αλγόριθμος να μην είναι το ίδιο ακριβής όπως ο BFGS.[26]

3.6 Μέτρα απόδοσης αλγόριθμων Επιτηρουμένης Μηχανικής Μάθησης

Τα μέτρα απόδοσης είναι συνδεδεμένα με εργασίες μηχανικής μάθησης, καθώς παίζουν σημαντικό ρόλο στην ανάπτυξη, στην επιλογή και στην αξιολόγηση του μοντέλου. Υπάρχουν πολλά μέτρα απόδοσης για ταξινόμηση, όπως παρουσιάζονται παρακάτω. Χρησιμοποιώντας διαφορετικά μέτρα για την αξιολόγηση της απόδοσης η συνολική δύναμη πρόβλεψης του μοντέλου μπορεί να βελτιωθεί, ενώ αν για παράδειγμα βασιστούν μόνο στην ακρίβεια μπορεί να υπάρξουν λανθασμένες προβλέψεις. Ακρίβεια (**Accuracy**) χρησιμοποιείται για να μετρήσει κατά πόσο ο ταξινομητής προβλέπει σωστά. Ορίζεται ως το κλάσμα του συνολικού αριθμού ορθών προβλέψεων προς τον συνολικό αριθμό προβλέψεων.

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

Η ακρίβεια (**Precision**) εξηγεί πόσες από τις θετικά προβλεπόμενες υποθέσεις είναι πραγματικά θετικές. Ορίζεται ως ο αριθμός των δεδομένων που ταξινομήθηκαν ορθώς ως θετικά προς τον αριθμό των θετικά προβλεπόμενων δεδομένων.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Η ανάκληση (**Recall**) εξηγεί σε πόσες από τις πραγματικές θετικές υποθέσεις μπορούσε να γίνει ορθή πρόβλεψη με το μοντέλο. Είναι ο αριθμός των ορθά θετικών καταχωρήσεων διαιρεμένος με το συνολικό αριθμό των πραγματικά θετικών.

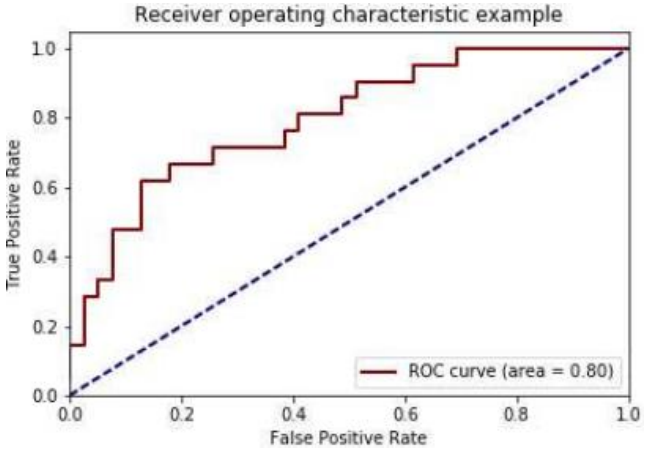
$$\text{Recall} = \frac{TP}{TP+FN}$$

Ο δείκτης **F1 Score** αποτελεί τον αρμονικό μέσο της ανάκλησης και της ακρίβειας. Είναι αποτελεσματικός όταν οι τιμές FP κοστίζουν το ίδιο, όταν προσθέτοντας νέα δεδομένα δεν αλλάζει ενεργά το αποτέλεσμα και όταν το True Negative είναι υψηλό.

$$F1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Ένα άλλο μέτρο απόδοσης είναι το **AUC-ROC**. Αρχικά, το ROC (Receiver Operator Characteristic) έχει ορισμένα πλεονεκτήματα σε σχέση με τα στοιχεία μέτρησης του πίνακα σύγχυσης. Αρχικά, δεν εξαρτάται από ρυθμίσεις του κατωφλίου. Επιπλέον η αλλαγή στη διανομή της κύριας κλάσης δεν το επηρεάζει καθόλου, καθώς εξαρτάται από τους ορθά θετικούς και λανθασμένους δείκτες. Τέλος, η ROC ανάλυση δεν στοχεύει απαραίτητα στην παραγωγή ακριβής πιθανότητας, αλλά πρέπει να διαχωρίσει τις θετικές από τις αρνητικές τιμές. Η καμπύλη ROC, θεωρείται ως συμβιβασμός μεταξύ των

True Positive Rates (TPR) και True Negative Rates (TNR) και είναι η καμπύλη πιθανοτήτων που σχεδιάζεται το TPR με το FPR, δηλαδή με τα πραγματικά θετικά και πραγματικά αρνητικά αντίστοιχα, οπότε δεν μεταβάλλεται αν αλλάξει η διανομή των κλάσεων. Στη καμπύλη ROC η τιμή του άξονα X, δείχνει FPR, ενώ στον άξονα Y TPR και οι διακεκομένες δείχνουν το σημείο στο οποίο το μοντέλο είναι σε θέση να διαχωρίσει αν η βάση δεδομένων είναι καλή ή όχι. Η AUC (Area Under Curve) είναι διαφορετική από τη καμπύλη ROC, καθώς είναι το μέτρο της ικανότητας του ταξινομητή να διαχωρίζεται μεταξύ των κλάσεων. Όσο καλύτερη η τιμή του AUC, τόσο καλύτερη η απόδοση του μοντέλου.σε διαφορετικές τιμές κατωφλίου, καθώς για τιμή ίση με 1, ο ταξινομητής διαχωρίζει ακριβώς τις θετικές και τις αρνητικές κλάσεις, όταν είναι 0 προβλέπει τις αρνητικές για θετικές και το αντίστροφο, ενώ όταν είναι 0.5 δεν μπορεί να διαχωρίσει αυτές τις κλάσεις. Η στατιστική του ορολογία είναι: η AUC ενός ταξινομητή ισούται με την πιθανότητα του ταξινομητή να αξιολογεί ένα τυχαίο θετικά δείγμα με μεγαλύτερη πιθανότητες από ένα τυχαία αρνητικό.

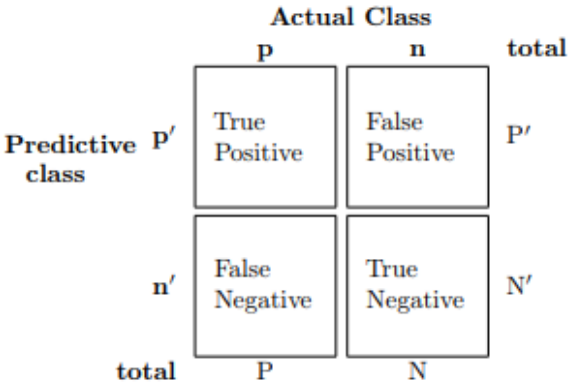


Εικόνα 4) Παράδειγμα καμπύλης ROC

Ο πίνακας σύγχυσης (**Confusion Matrix**) αποτελεί μία μέτρηση απόδοσης για προβλήματα ταξινόμησης όπου η έξοδος είναι δύο ή περισσότερες κλάσεις. Είναι ένας πίνακας συνδυασμών προβλεπόμενων και πραγματικών τιμών. Χρησιμοποιείται συχνά για να περιγράψει την απόδοση ενός μοντέλου ταξινόμησης σε ένα σετ δεδομένων ελέγχου όπου οι σωστές τιμές είναι γνωστές. Τα μέτρα Recall, Precision, Accuracy και του δείκτη F1, βασίζονται στον πίνακα σύγχυσης, καθώς από αυτόν προκύπτουν οι τιμές τους. Ο πίνακας δείχνει ποιες καταχωρήσεις έχουν ταξινομηθεί σωστά στην αντίστοιχη καθώς για δύο κλάσεις έχουμε:

- True Positive (TP): περιπτώσεις που σωστά αντιστοιχήθηκαν σε θετική κλάση
- True Negative (TN): περιπτώσεις που σωστά αντιστοιχήθηκαν σε αρνητική κλάση
- False Positive (FP): περιπτώσεις που λανθασμένα αντιστοιχήθηκαν σε θετική κλάση

- False Negative (FN): περιπτώσεις που λανθασμένα αντιστοιχήθηκαν σε αρνητική κλάση,[27]



Εικόνα 5) Confusion Matrix

3.7 Γενίκευση και Θόρυβος (Generalization and Noise)

Όταν το νευρωνικό σύστημα που διατίθεται εκπαιδεύεται, πρέπει να ληφθεί υπόψιν η γενίκευση του, κατά πόσο δηλαδή το μοντέλο έχει εκπαιδευτεί από τα δοσμένα δεδομένα ώστε να εφαρμόσει σωστή πρόβλεψη σε νέα δεδομένα. Εκπαιδεύοντας ένα νευρωνικό δίκτυο, κάποια δεδομένα χρησιμοποιούνται για την εκπαίδευση του, ενώ κάποια άλλα για τον έλεγχο της απόδοσής του. Εάν το νευρωνικό δίκτυο αποδίδει καλά στα δεδομένα τα οποία δεν έχει εκπαιδευτεί, τότε λέμε ότι στα δοσμένα δεδομένα έχει γενικεύσει καλά.

Η συλλογή των δεδομένων μπορεί να οδηγήσει σε σφάλματα στη βάση δεδομένων τα οποία αναφέρονται ως θόρυβος. Ο θόρυβος των δεδομένων μπορεί να προκαλέσει προβλήματα, καθώς ο αλγόριθμος ερμηνεύει τον θόρυβο σαν μοτίβο και ξεκινάει να γενικεύει από αυτό. Αυτό έχει ως αποτέλεσμα το μοντέλο να μην γενικεύει καλά.

Ωστόσο υπάρχουν τεχνικές που προσθέτουν μικρή ποσότητα θορύβου στο αρχικό τεστ εκπαίδευσης, ώστε το μοντέλο να είναι σε θέση να μπορεί να γενικεύσει καλά. Με αυτό το τρόπο τα δεδομένα εκπαίδευσης μεγαλώνουν, καθώς παράγονται νέα δεδομένα. Αυτό έχει ως αποτέλεσμα την αύξηση του μεγέθους του εκπαιδευτικού συνόλου δεδομένων όπου η εκπαίδευση του ενδέχεται να βοηθήσει στη γενίκευση του μοντέλου. Η εκπαίδευση ενός μοντέλου με χρήση θορύβου έχει ως αποτέλεσμα την βελτίωση της ικανότητας του μοντέλου να αναγνωρίζει δεδομένα που περιέχουν θόρυβο, καθώς και δεδομένα των κλάσεων που δεν ανήκουν στο σετ εκπαίδευσης.[28]

.

3.8 Bias-Variance

Γενικά ένα μοντέλο μηχανικής μάθησης αναλύει τα δεδομένα, βρίσκει μοτίβα σε αυτά και πραγματοποιεί προβλέψεις. Όταν εκπαιδεύεται το μοντέλο, μαθαίνει από αυτά τα μοτίβα στη βάση δεδομένων και τα χρησιμοποιεί στα δεδομένα ελέγχου για πρόβλεψη. Στόχος της μηχανικής μάθησης είναι να μπορεί το μοντέλο να γενικεύει και να αποδίδει καλά σε νέα δεδομένα. Ένα απλό μοντέλο με ένα κρυφό στρώμα έχει υψηλή μεροληψία, ενώ αντίθετα, ένα μοντέλο με πολλαπλά κρυφά στρώματα και βάρη θα έχει χαμηλή μεροληψία, αλλά υψηλή διακύμανση. Αν σε ένα νευρωνικό δίκτυο, γίνει η εξομάλυνση των βαρών των νευρωνικών δικτύων, τότε η διακύμανση θα μειωθεί, αλλά η μεροληψία του δικτύου θα αυξηθεί.[29]

3.8.1 Bias-Variance trade off

Σε σημερινές πρακτικές, ορισμένα μοντέλα όπως τα νευρωνικά δίκτυα εκπαιδεύονται ώστε να προσαρμόζονται ακριβώς στα δεδομένα, γεγονός που δείχνει ότι έχουν υψηλή ακρίβεια στα δεδομένα ελέγχου αν και δημιουργείται το πρόβλημα της υπερπροσαρμογής. Αν το μοντέλο είναι αρκετά απλό μπορεί να υπάρξει η υποπροσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης οπότε να μην είναι ικανό να προβλέψει νέα δεδομένα, καθώς θα έχει υψηλή μεροληψία. Αν το μοντέλο είναι αρκετά σύνθετο, μπορεί να υπάρξει υπερπροσαρμογή στα εκπαιδευτικά δεδομένα με αποτέλεσμα χαμηλή ακρίβεια σε νέα δεδομένα. Στόχος της μηχανικής μάθησης είναι να μπορεί το μοντέλο να αποδίδει σε ικανοποιητικό βαθμό σε νέα δεδομένα. Για αυτό, πρέπει να υπάρχει μια ισορροπία μεταξύ των σφαλμάτων μεροληψίας και διακύμανσης και η ισορροπία αυτή ονομάζεται Bias-Variance trade off, μία ισορροπία δηλαδή μεταξύ της υπερπροσαρμογής και υποπροσαρμογής του μοντέλου. Για ακριβείς προβλέψεις ενός μοντέλου πρέπει να υπάρχει χαμηλή διακύμανση και χαμηλή μεροληψία. Αυτό όμως δεν είναι εφικτό, καθώς είναι όροι αντιστρόφως ανάλογοι, όπου όταν μειώνεται η μεροληψία, αυξάνεται η διακύμανση και το αντίστροφο. Οπότε πρέπει να βρεθεί ένα σημείο στο οποίο η διακύμανση και η διασπορά δημιουργούν το καλύτερο σημείο. Το Bias-Variance trade off σχετίζεται με την εύρεση του σημείου στο οποίο δημιουργείται η ισορροπία μεταξύ σφαλμάτων μεροληψίας και διακύμανσης. Σε γράφημα για την εύρεση του συγκεκριμένου σημείου, δημιουργείται ένα σχήμα σε μορφή U, όπου στον πάτο του βρίσκεται το συγκεκριμένο σημείο, όπου αριστερά από αυτό υπάρχει υποπροσαρμογή του μοντέλου και δεξιά από αυτό η υπερπροσαρμογή του.[30]

3.9 Underfitting-Overfitting

Ένα από τα πιο συνηθισμένα προβλήματα των νευρωνικών δικτύων αποτελεί η υπερπροσαρμογή ή υποπροσαρμογή του μοντέλου, κατά τις οποίες το μοντέλο δεν μπορεί να γενικεύσει, λόγω της χαμηλής απόδοσης των αλγόριθμων μάθησης. Παρακάτω, αναλύονται οι έννοιες της υπερπροσαρμογής και υποπροσαρμογής του μοντέλου, αλλά και κάποιοι τρόποι αντιμετώπισης τους.

Η υπερπροσαρμογή του μοντέλου είναι ένα φαινόμενο που εντοπίζεται όταν ο αλγόριθμος μάθησης ταιριάζει στα εκπαιδευτικά δεδομένα σε τέτοιο βαθμό ώστε να αποθηκεύονται στη μνήμη του μοντέλου, τόσο ο θόρυβος όσο και οι ιδιαιτερότητες των εκπαιδευτικών δεδομένων. Η απόδοση του αλγορίθμου μάθησης πέφτει αν υπάρχει το φαινόμενο υπερπροσαρμογής του μοντέλου όταν ελέγχεται το μοντέλο σε μία άγνωστη βάση δεδομένων. Επιπλέον, η ποσότητα των δεδομένων για τη διαδικασία μάθησης σχετίζεται με το μοντέλο της υπερπροσαρμογής, καθώς όσο πιο μικρή η βάση δεδομένων, τόσο πιο πιθανή είναι να συμβεί υπερπροσαρμογή του μοντέλου. Τέλος στην υπερπροσαρμογή του μοντέλου, παρατηρείται χαμηλή μεροληψία και υψηλή διακύμανση.

Το φαινόμενο της υποπροσαρμογής είναι το ακριβώς αντίθετο από εκείνο της υπερπροσαρμογής, καθώς το μοντέλο διαθέτει υψηλή μεροληψία και χαμηλή διακύμανση. Συμβαίνει όταν το μοντέλο δεν είναι σε θέση να συλλάβει τη μεταβλητότητα των δεδομένων. Ο ταξινομητής δεν είναι καθόλου ικανός να πραγματοποιήσει προβλέψεις και αυτό είναι το αποτέλεσμα της κατανόησης ή της χρήσης ενός αρκετά απλού μοντέλου.

Για να αντιμετωπιστούν τα φαινόμενα υπερπροσαρμογής ή υποπροσαρμογής του μοντέλου χρησιμοποιούνται κάποιες μέθοδοι οι οποίες είναι μέθοδοι ποινών και το έγκαιρο σταμάτημα του αλγορίθμου κατά την εκπαίδευση του.

Οι μέθοδοι ποινών χρησιμοποιούνται για την αποφυγή της υπερπροσαρμογής και έχουν ως στόχο την εύρεση του σημείου εκείνου για το οποίο ελαχιστοποιείται το σφάλμα εκπαίδευσης. Ορισμένες μέθοδοι ποινών παρουσιάζονται παρακάτω:

- Ο χάρτης παρέχει ποινή βασισμένη στο P(H)
- Η αρχή του Minimum Description Length (MDL)
- Ελαχιστοποιημένο ρίσκο δομής
- Γενικευμένο cross-validation

- Hold και cross-validation

Οι μέθοδοι έγκαιρου σταματήματος του αλγόριθμου χρησιμοποιούνται για την αντιμετώπιση της υπερπροσαρμογής και της υποπροσαρμογής του μοντέλου και μπορούν να χρησιμοποιηθούν εκτός από το εκπαιδευτικό σετ δεδομένων, στο σετ ελέγχου και στο σετ εγκυροποίησης. Η υπερπροσαρμογή του μοντέλου μπορεί να παρατηρηθεί κατά την μέθοδο cross validation, όπου τα δεδομένα χωρίζονται σε ελέγχου, εγκυροποίησης και εκπαιδευτικά. Το σφάλμα του σετ εγκυροποίησης παρατηρείται κατά την εκπαίδευση, όπου στην αρχική φάση της εκπαίδευσης μειώνεται. Όταν το σφάλμα στο σετ εγκυροποίησης αρχίζει και αυξάνεται, τότε συμβαίνει το φαινόμενο της υπερπροσαρμογής και ο αλγόριθμος σταματάει και επιστρέφονται τα βάρη που υπήρχαν στο ελάχιστο σφάλμα εγκυροποίησης. Αν το μοντέλο είναι μεγάλο και περιέχει πολλά δεδομένα, είναι ασύμφορη οικονομικά αυτή η μέθοδος και για αυτό ξεκινάνε από μικρή ποσότητα και μεγαλώνουν μέχρι να αρχίσει να χειροτερεύει το σετ εγκυροποίησης.[31]

ΚΕΦΑΛΑΙΟ 4: Πειραματικό μέρος της έρευνας

4.1 Εισαγωγή

Στο συγκεκριμένο κεφάλαιο παρουσιάζεται το λογισμικό που θα χρησιμοποιηθεί στην παρούσα έρευνα, η βάση δεδομένων με τα χαρακτηριστικά της, καθώς και η διαδικασία που ακολουθήθηκε για την επεξεργασία των δεδομένων της βάσης πελατών. Στη συνέχεια, απεικονίζεται το σύνολο των δεδομένων για διαφορετικές τιμές υπερπαραμέτρων των νευρωνικών δικτύων και παρουσιάζεται η απόδοση τους ως προς την πρόβλεψη της απώλειας πελατών.

4.2 Λογισμικό Orange Data Mining

Το λογισμικό που χρησιμοποιήθηκε στην συγκεκριμένη έρευνα είναι το **Orange Data Mining** (έκδοση 3.34.0). Η ιδέα για την δημιουργία του λογισμικού δόθηκε το 1997 από τον Donald Michie, καθώς θεωρούσε πως έπρεπε να υπάρξει μια ανοιχτή βιβλιοθήκη για Μηχανική Μάθηση και για να γίνει η πυροδότηση της ανάπτυξής του πραγματοποιήθηκε ένα συνέδριο που ονομάστηκε WebLab97 στη πόλη Bled της Σλοβενίας. Στο συγκεκριμένο λογισμικό χρησιμοποιείται η γλώσσα Python και τα πνευματικά δικαιώματα ανήκουν στο πανεπιστήμιο της Λιουμπλιάνας. Το λογισμικό χρησιμοποιείται για τις οπτικοποιήσεις των δεδομένων οι οποίες είναι διαδραστικές, παρέχει διάφορα widgets για πληθώρα λειτουργιών όπως είναι ο πίνακας των δεδομένων, διάφορους αλγόριθμους όπως τα νευρωνικά δίκτυα, το test and score που παρουσιάζει τα αποτελέσματα των αλγόριθμων και χρησιμοποιεί cross validation, ο πίνακας σύγχυσης και οι προβλέψεις. Επιπλέον, είναι δυνατή η επιλογή χαρακτηριστικών αλλά και δεδομένων μέσα από γραφήματα, διαγράμματα ή πίνακα δεδομένων και η εξόρυξη τους σε ένα άλλο widget.

4.3 Στατιστικά του dataset

Η βάση δεδομένων της συγκεκριμένης έρευνας δίνεται μέσω της kaggle, η οποία αποτελεί μία πλατφόρμα δεδομένων. Η συγκεκριμένη βάση δεδομένων αποτελείται από δεδομένα 64000 πελατών. Τα δεδομένα αυτά περιλαμβάνουν σύντομες πληροφορίες για τους πελάτες όπως είναι περιοχές που ζουν ή αγορές που έχουν πραγματοποιήσει με την εταιρεία, μεθόδους Μάρκετινγκ που έχει χρησιμοποιήσει η εταιρεία όπως διάφορες προσφορές και αν έχουν γίνει δεκτές από τους πελάτες καθώς και αν συνεχίζουν να είναι πελάτες της επιχείρησης ή όχι. Τα δεδομένα χρησιμοποιούνται για πρόβλεψη της απώλειας πελατών της εταιρείας. [32]

Το σύνολο δεδομένων παρουσιάζεται σε ένα αρχείο σε μορφή csv και αποτελείται από 9 μεταβλητές. Δεν υπάρχουν ελλιπή στοιχεία στη βάση δεδομένων και τα χαρακτηριστικά της απεικονίζονται παρακάτω:

- Recency: Είναι το χρονικό περιθώριο σε μήνες που μεσολαβεί από την τελευταία αγορά του πελάτη
- History: Η συνολική αξία των παρελθοντικών συναλλαγών του κάθε πελάτη με την εταιρεία σε Αμερικανικά δολάρια
- Used Discount: Δείχνει αν ο πελάτης έχει χρησιμοποιήσει έκπτωση στο παρελθόν
- Used Bogo: Δείχνει αν ο πελάτης έχει χρησιμοποιήσει προσφορά ένα συν ένα στο παρελθόν
- Zip Code: Ο τόπος διαμονής του πελάτη και διακρίνεται σε προάστιο (Suburban), αστική περιοχή (Urban) και αγροτική περιοχή (Rural)
- Is referral: Δείχνει αν ο πελάτης αποκτήθηκε μέσω προωθητικών ενεργειών
- Channel: Ο τρόπος επικοινωνίας του πελάτη με την εταιρεία και είναι μέσω ίντερνετ (Web), τηλεφώνου (Phone) ή Multichannel
- Offer: Προσφορές που έχουν λάβει οι πελάτες και διακρίνονται σε έκπτωση (Discount), ένα συν ένα (Buy one get one) και καθόλου προσφορά (No offer)
- Conversion: Είναι η διατήρηση του πελάτη, αν συνεχίζει δηλαδή να πραγματοποιεί αγορές από την εταιρεία ή όχι

Η κλάση-στόχος αποτελεί το **conversion** το οποίο λαμβάνει δύο τιμές οι οποίες είναι 0 ή 1. Επομένως είναι πρόβλημα δυαδικής ταξινόμησης (**binary classification problem**). Από τους 64000 πελάτες, παραμένουν στην εταιρεία (**non churners)** και συνεχίζουν τις αγορές τους οι 54606, ενώ από την εταιρεία αποχώρησαν (**churners**) 9394. Το σύνολο δεδομένων είναι **μη ισορροπημένο (imbalanced)**, καθώς το 85% παραμένει στην εταιρεία (κλάση 0), ενώ το υπόλοιπο 15% αποχώρησε από αυτή.

Η μεταβλητή **recency** δείχνει το χρονικό περιθώριο το οποίο ο πελάτης έχει να πραγματοποιήσει αγορές από την εταιρεία. Οι πελάτες που πραγματοποίησαν αγορές τον τελευταίο μήνα είναι 8952 (14%), ενώ το τελευταίο τρίμηνο συνολικά έχουν πραγματοποιήσει αγορές συνολικά 22393 πελάτες (35%). Από 4-6 μήνες έχουν πραγματοποιήσει αγορές συνολικά 14192 πελάτες (22%) και από 6-12 μήνες έχουν πραγματοποίησει αγορές 27415 πελάτες (43%). Επομένως, παρατηρείται ότι οι περισσότεροι πελάτες έχουν πραγματοποιήσει αγορές εντός 6 μηνών.

Η μεταβλητή **history** αφορά το συνολικό ποσό που έχει ξοδέψει ο κάθε πελάτης για τις συναλλαγές του με την εταιρεία. Οι πελάτες που έχουν ξοδέψει έως 195,79$ αποτελούν την πλειοψηφία

του δείγματος και είναι 36741 (57,4%) ενώ ακολουθούν 13432 (21%) που έχουν ξοδέψει από 195,79$ εώς 361,58$. Από 361,58$ μέχρι 693,18$ έχουν ξοδέψει 9939 πελάτες (15,6%), Οι υπόλοιποι 3888 (6%) πελάτες έχουν ξοδέψει περισσότερα από 693,18$, ενώ μόνο 14 πελάτες έχουν ξοδέψει περισσότερα από 2500$.

Η μεταβλητή **Used Discount** απεικονίζει αν ο πελάτης έχει χρησιμοποιήσει έκπτωση στο παρελθόν με 28734 πελάτες (44,9%) να έχουν χρησιμοποιήσει έκπτωση στο παρελθόν, ενώ οι υπόλοιποι όχι.

Η μεταβλητή **Used Bogo** δείχνει αν οι πελάτες έχουν χρησιμοποίησει προσφορά ένα συν ένα όπου το 45% των πελατών έχουν χρησιμοποιήσει.

Οι περιοχές στις οποίες μένουν οι πελάτες δίνεται από την μεταβλητή **zip code**, όπου το μεγαλύτερο ποσοστό (45%) των πελατών βρίσκεται σε προάστιο, το 40% σε αστική περιοχή και το υπόλοιπο 15% σε αγροτική.

Η μεταβλητή **is referral** δείχνει πόσοι πελάτες αποκτήθηκαν μέσω προωθητικών ενεργειών, όπου 31856 πελάτες (49,78%) αποκτήθηκαν με αυτό το τρόπο.

Ο τρόπος με τον οποίο η εταιρεία επικοινωνεί με τον πελάτη δίνεται μέσω της μεταβλητής **channel**. Οι πελάτες επικοινωνούν με την εταιρεία μέσω ίντερνετ σε ποσοστό 44%, μέσω τηλεφώνου σε ποσοστό 44% και μέσω Multichannel σε ποσοστό 12%.

Οι προφορές που έχουν λάβει οι πελάτες περιγράφονται με την μεταβλητή **offer**. Πελάτες που έχουν λάβει προφορά ένα συν ένα είναι σε ποσοστό 33%, έκπτωση έχουν λάβει πελάτες σε ποσοστό 33%, ενώ 33% δεν έλαβε καμία προσφορά.

4.4 Διαδικασία υλοποίησης της έρευνας

Στη συγκεκριμένη έρευνα που αφορά την πρόβλεψη της απώλειας των πελατών, όπως αναφέρθηκε προηγουμένως, προκύπτει το πρόβλημα της άνισης κατανομής των κλάσεων (**Class Imbalance Problem**). Αυτό δημιουργείται, διότι τα δεδομένα της μειονοτικής κλάσης είναι αρκετά λιγότερα από εκείνα της κύριας με αποτέλεσμα την υπερπροσαρμογή (**Overfitting**) των νευρωνικών δικτύων στην κύρια κλάση. Με αυτό το τρόπο τα νευρωνικά δίκτυα δεν μπορούν να διαχωρίσουν τα δεδομένα της μειονοτικής κλάσης από εκείνα της κύριας με αποτέλεσμα να την αγνοούν και να θεωρούν όλα τα δεδομένα πως ανήκουν στην κύρια. Αυτό προκαλεί μείωση της απόδοσης τους και τα καθιστά ανίκανα να προβλέψουν σωστά το αποτέλεσμα.

Τα προβλήματα της άνισης κατανομής των κλάσεων γενικά μπορούν να αντιμετωπιστούν με τους εξής τρόπους:

1. Μέσω της τεχνικής **Cost-Sensitive Training**, όπου χρησιμοποιούνται αλγόριθμοι μάθησης οι οποίοι έχουν ως ποινή την αύξηση του κόστους για λάθη ταξινόμησης στην μειονοτική κλάση, ενισχύοντας τη σημαντικότητα της
2. Αλλάζοντας το μέτρο απόδοσης, καθώς η ακρίβεια (**Accuracy**) μπορεί να οδηγήσει σε λανθασμένα αποτελέσματα και για αυτό το λόγο λαμβάνονται υπόψιν τα άλλα μέτρα απόδοσης, όπως ο πίνακας σύγχυσης (**Confusion Matrix**), η ακρίβεια (**Precision**), η ανάκληση (**Recall**), ο δείκτης F1 (**F1: Score)** το **AUC** και η περιοχή κάτω από την καμπύλη ROC (**Area Under ROC Curve**)
3. Κατά την προεπεξεργασία των δεδομένων μπορούν να αφαιρεθούν δεδομένα από την κύρια κλάση (**Under-sampling**) ή να προστεθούν δεδομένα στη μειονοτική κλάση (**Oversampling**). Η τεχνική αυτή ονομάζεται **resampling** και είναι από τις πιο διαδεδομένες τεχνικές για την αντιμετώπιση του προβλήματος της άνισης κατανομής των κλάσεων

Όσον αφορά τις τεχνικές resampling, πρέπει να γίνει η εφαρμογή τους με ιδιαίτερη προσοχή καθώς η αύξηση των δεδομένων στη μειονοτική κλάση ενδέχεται να κάνει τον αλγόριθμο να τρέχει αργά, ενώ η μείωση τους στη κύρια κλάση μπορεί να έχει ως αποτέλεσμα να χαθούν σημαντικά για την ανάλυση δεδομένα.

Αρχικά, ο αλγόριθμος έτρεξε στην συγκεκριμένη έρευνα με τα δεδομένα που διέθετε. Ωστόσο προέκυψε το πρόβλημα της άνισης κατανομής των κλάσεων λόγω της υπερπροσαρμογής των δεδομένων στην κύρια κλάση όπως αναφέρθηκε παραπάνω.

Για να αντιμετωπιστεί το πρόβλημα της άνισης κατανομής των κλάσεων πραγματοποιήθηκαν δύο ξεχωριστές διαδικασίες, όπου και οι δύο αφορούν την προεπεξεργασία των δεδομένων. Αρχικά, γίνεται μια προεπεξεργασία των μεταβλητών των δεδομένων με τις μεθόδους **One-hot encoding** και **MinMaxScaler**. Στη συνέχεια χρησιμοποιούνται τεχνικές ανακατανομής των κλάσεων (resampling) οι οποίες είναι οι **Random Under-sampling, Random Oversampling, SMOTE** και **SMOTE-ENN.** Για κάθε τεχνική ξεχωριστά έτρεξε ο αλγόριθμος νευρωνικών δικτύων με σκοπό την εύρεση των βέλτιστων τιμών του μοντέλου ώστε να είναι ικανό να δώσει καλύτερα αποτελέσματα πρόβλεψης. Για να συμβεί αυτό, πραγματοποιήθηκαν διάφορες αλλαγές στις τιμές των υπερπαραμέτρων.

Όπως αναφέρθηκε παραπάνω, σε πρώτη φάση πραγματοποιείται μία προεπεξεργασία των μεταβλητών των δεδομένων με τις μεθόδους One-hot encoding και MinMaxScaler. Με τη χρήση των

μεθόδων αυτών τα δεδομένα μπορούν να ταιριάξουν κατάλληλα στο μοντέλο πριν ξεκινήσει η ανάλυση του.

Τα κατηγορηματικά δεδομένα είναι μεταβλητές που περιέχουν ονομαστικές τιμές για κάθε κατηγορία και όχι αριθμητικές και για αυτό ονομάζονται συχνά ονομαστικά δεδομένα. Επειδή όμως πολλοί αλγόριθμοι δεν είναι ικανοί να λειτουργήσουν με ονομαστικές τιμές, πραγματοποιείται η μετατροπή των κατηγορηματικών δεδομένων σε αριθμητικά και αυτό γίνεται με τη μέθοδο one-hot encoding. Στη συγκεκριμένη μέθοδο τα κατηγορηματικά δεδομένα (categorical data), παρουσιάζονται σαν ακέραιες τιμές. Αυτό γίνεται ιδίως αν δεν υπάρχει κάποια σχέση διάταξης μεταξύ τους. Συγκεκριμένα η κάθε μεταβλητή γίνεται δυαδική όπου κάθε κατηγορία της μεταβλητής λαμβάνει τη τιμή 1 και οι υπόλοιπες την τιμή 0 και αυτό γίνεται για κάθε κατηγορία ξεχωριστά.

Πολλοί αλγόριθμοι αποδίδουν καλύτερα αν οι μεταβλητές των δεδομένων βρίσκονται μεταξύ ενός εύρους τιμών σε μία κλίμακα. Στα νευρωνικά δίκτυα ιδίως, είναι πολύ σημαντικό η στοχευμένη μεταβλητή να ανήκει σε μία κλίμακα τιμών, καθώς με αυτόν τον τρόπο το μοντέλο μαθαίνει πιο εύκολα και μπορεί να πραγματοποιήσει καλύτερες προβλέψεις. Ένας τρόπος τοποθέτησης των δεδομένων σε μια κλίμακα με ένα εύρος τιμών είναι με τη μέθοδο ομαλοποίησης όπου οι τιμές των δεδομένων έχουν εύρος τιμών από 0 έως 1. Η ομαλοποίηση των τιμών απαιτεί την γνώση της μέγιστης και της ελάχιστης τιμής των χαρακτηριστικών των δεδομένων και η βάση δεδομένων μπορεί να ομαλοποιηθεί με τη μέθοδο MinMaxScaler. Στη συγκεκριμένη μέθοδο πραγματοποιείται η μετατροπή των χαρακτηριστικών κλιμακώνοντας κάθε χαρακτηριστικό σε ένα συγκεκριμένο εύρος τιμών μεταξύ 0 και 1. Η μέθοδος μπορεί να χρησιμοποιηθεί ταιριάζοντας τον κλιμακωτή χρησιμοποιώντας τα διαθέσιμα εκπαιδευτικά δεδομένα, εφαρμόζοντας την κλίμακα στα δεδομένα εκπαίδευσης, άλλα και σε καινούρια δεδομένα που χρειάζεται να πραγματοποιηθούν προβλέψεις.

Στην δεύτερη φάση χρησιμοποιούνται οι τεχνικές resample οι οποίες είναι η Random Under-sampling, η Random Over-sampling, η SMOTE και η SMOTE-ENN. Ο αλγόριθμος έτρεξε για κάθε τεχνική ξεχωριστά, όπου για κάθε μία απέδωσε διαφορετικά αποτελέσματα.

Μία μέθοδος για την αντιμετώπιση της άνισης κατανομής των κλάσεων είναι η τυχαία ανακατανομή (random resample) στη βάση δεδομένων εκπαίδευσης. Αυτό μπορεί να επιτευχθεί διαγράφοντας δεδομένα από την κύρια κλάση που ονομάζεται Random Under-sampling ή προσθέτοντας αντίγραφα δεδομένων στη μειονοτική κλάση που ονομάζεται Random Over-sampling. Η τεχνική της τυχαίας ανακατανομής περιλαμβάνει τη δημιουργία μίας νέας τροποποιημένης έκδοσης της εκπαιδευτικής βάσης δεδομένων όπου τα τυχαία επιλεγμένα δεδομένα έχουν διαφορετική κατανομή κλάσης.

Συγκεκριμένα η αλλαγή στην κατανομή των κλάσεων χρησιμοποιείται αποκλειστικά στην εκπαιδευτική βάση δεδομένων, ώστε να γίνεται καλύτερα η προσαρμογή των μοντέλων και όχι στη βάση δεδομένων ελέγχου που χρησιμοποιείται για τον έλεγχο της απόδοσης του μοντέλου. Δύο κύριες τεχνικές της τυχαίας ανακατανομής είναι όπως αναφέρθηκαν παραπάνω η Random Under-sampling και η Random Oversampling. Οι δύο μέθοδοι επαναλαμβάνονται μέχρι να επιτευχθεί η επιθυμητή κατανομή των κλάσεων στην εκπαιδευτική βάση δεδομένων όπου είναι ο ισότιμος διαχωρισμός μεταξύ των κλάσεων και μπορούν να χρησιμοποιηθούν σε προβλήματα δυαδικής ταξινόμησης κλάσεων ή σε προβλήματα ταξινόμησης πολλαπλών κλάσεων. Αναφέρονται ως μέθοδοι αφελής ανακατανομής (naive resample) επειδή δεν υποθέτουν τίποτα για τα δεδομένα και αυτό τους κάνει να είναι γρήγορα εκτελέσιμοι και απλοί στην εφαρμογή το οποίο είναι επιθυμητό για μεγάλες και περίπλοκες βάσεις δεδομένων.

Με τη μέθοδο *Random Under-sample* διαγράφονται τυχαία δεδομένα στην εκπαιδευτική βάση δεδομένων τα οποία επιλέγονται από την κύρια κλάση με αποτέλεσμα τη μείωση των δεδομένων της κύριας κλάσης στη νέα τροποποιημένη έκδοση της εκπαιδευτικής βάσης δεδομένων. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να επιτευχθεί η επιθυμητή κατανομή των κλάσεων όπως είναι ο ίσος αριθμός δεδομένων για κάθε κλάση και ίσως είναι πιο κατάλληλη σε βάσεις δεδομένων που υπάρχει ανισορροπία των κλάσεων παρόλο που η μειονοτική κλάση διαθέτει επαρκή αριθμό δεδομένων, ώστε ένα τόσο χρήσιμο μοντέλο να μπορεί να προσαρμοστεί. Ένας σημαντικός περιορισμός αυτής της μεθόδου είναι ότι διαγράφονται δεδομένα της κύριας κλάσης τα οποία ενδέχεται να είναι χρήσιμα και επειδή διαγράφονται τυχαία δεν υπάρχει τρόπος να εντοπιστούν και να διατηρηθούν χρήσιμες πληροφορίες.

Με τη μέθοδο *Random-Oversample* δημιουργούνται τυχαία αντίγραφα της μειονοτικής κλάσης και τοποθετούνται στην εκπαιδευτική βάση δεδομένων. Δεδομένα από τη μειονοτική κλάση μπορούν να επιλεχθούν πολλαπλές φορές, καθώς επιλέγονται από την αρχική βάση δεδομένων, τοποθετούνται στη καινούρια και επιστρέφουν πάλι στην αρχική ώστε να επιλεχθούν ξανά. Η τεχνική μπορεί να είναι αποτελεσματική σε αλγόριθμους οι οποίοι επηρεάζονται από κατανομές και όπου πολλαπλά αντίγραφα της δοσμένης κλάσης επηρεάζουν την προσαρμογή του μοντέλου. Επιπλέον, μπορεί να είναι χρήσιμος ο συντονισμός της κατανομής της κλάσης στόχου, καθώς σε κάποιες περιπτώσεις στην προσπάθεια να επιτευχθεί ισορροπημένη κατανομή σε μία ανισόρροπη βάση δεδομένων μπορεί να προκληθεί υπερπροσαρμογή των αλγόριθμων στην μειονοτική κλάση με αποτέλεσμα αυξημένο σφάλμα γενίκευσης. Έτσι μπορεί να αποδίδει καλύτερα σε εκπαιδευτική βάση δεδομένων, αλλά να έχει χαμηλή απόδοση στη βάση δεδομένων ελέγχου. Επίσης, η αύξηση των δεδομένων στη μειονοτική κλάση μπορεί να δημιουργήσει αύξηση του υπολογιστικού κόστους όταν προσαρμόζεται το μοντέλο, καθώς παρατηρεί τα ίδια δεδομένα στην εκπαιδευτική βάση δεδομένων ξανά και ξανά.

Μια άλλη μέθοδος Oversampling αποτελεί η μέθοδος *Synthetic Minority Oversampling Technique* ή αλλιώς *SMOTE* της οποίας τα συνθετικά δείγματα παράγονται από τη μειονοτική κλάση. Ο αλγόριθμος αυτός βοηθάει να ξεπεραστεί το πρόβλημα της υπερπροσαρμογής που δημιουργείται από το τυχαίο oversampling. Δημιουργεί ένα συνθετικά ταξικά ισορροπημένο εκπαιδευτικό σετ και μετά εκπαιδεύει τον ταξινομητή και παράγει συνθετικά δείγματα για να ισορροπήσει ανισόρροπες βάσεις δεδομένων ιδιαίτερα στοχεύοντας στη μειονοτική κλάση. Ο τρόπος με τον οποίο λειτουργεί η συγκεκριμένη μέθοδος αναλύεται παρακάτω.

Αρχικά, στήνεται ο συνολικός αριθμός των Oversampling παρατηρήσεων N, ο οποίος γενικά επιλέγεται ώστε η δυαδική κατανομή της κλάσης να είναι 1:1, ωστόσο μπορεί να ρυθμιστεί ανάλογα την ανάγκη. Τότε, η επανάληψη ξεκινάει επιλέγοντας τυχαία ένα δεδομένο της μειονοτικής κλάσης. Στη συνέχεια, υπολογίζεται η απόσταση μεταξύ του τυχαίου δείγματος και του k κοντινότερου του γείτονα (k-nearest neighbor), η οποία διαφορά πολλαπλασιάζεται με ένα τυχαίο αριθμό μεταξύ 0 και 1. Ακολούθως το αποτέλεσμα προστίθεται στη μειονοτική κλάση, και έτσι δημιουργείται ένα νέο συνθετικό δεδομένο. Η διαδικασία συνεχίζεται στον επόμενο πλησιέστερο γείτονα, μέχρι τον αριθμό που έχει ορίσει ο χρήστης. Παρόλο που ο αλγόριθμος είναι χρήσιμος, διαθέτει κάποια μειονεκτήματα, όπως είναι ότι τα νέα δεδομένα βρίσκονται στην ίδια γραμμή με τα ήδη υπάρχοντα το οποίο περιπλέκει την επιφάνεια απόφασης που δημιουργείται από ορισμένους αλγόριθμους ταξινόμησης και επιπλέον ότι η SMOTE δημιουργεί μεγάλο αριθμό θορυβώδη δεδομένων στο χώρο χαρακτηριστικών.

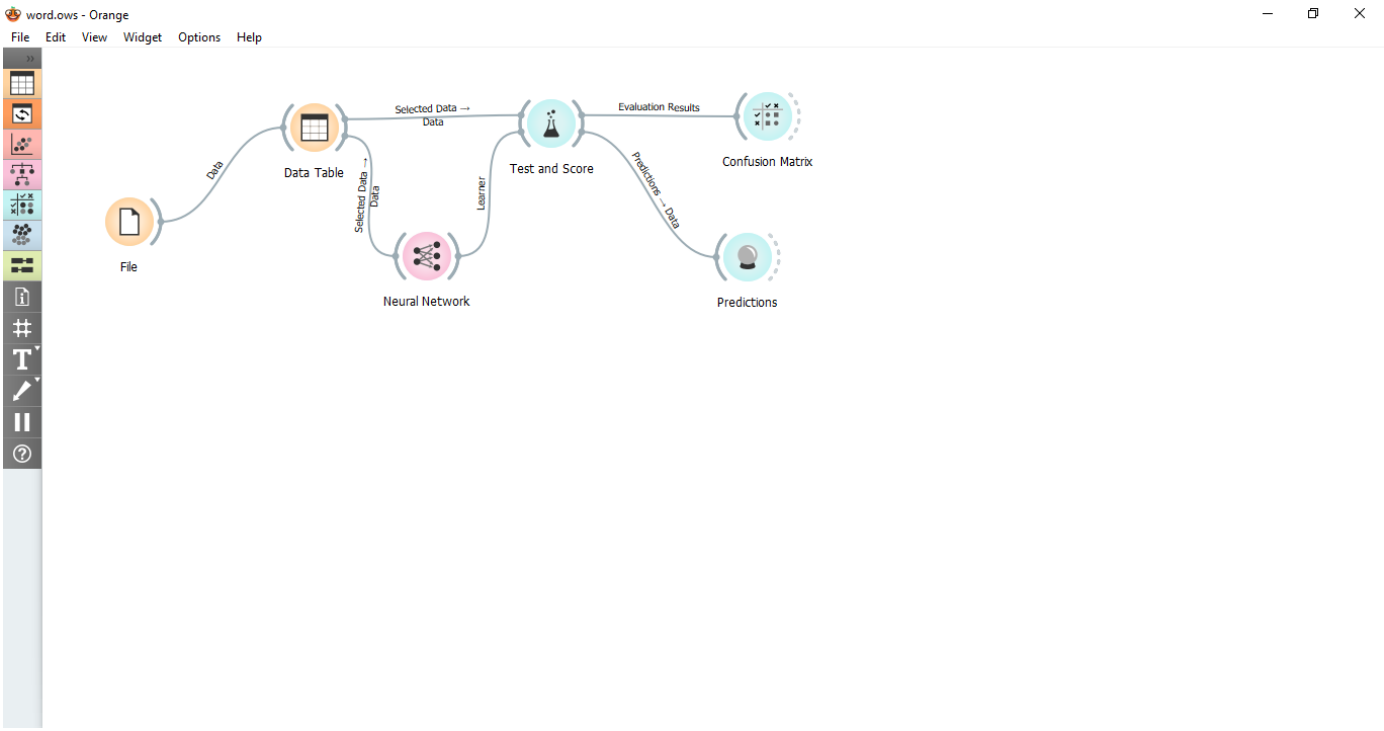
Μία παραλλαγή της μεθόδου SMOTE είναι η μέθοδος *SMOTE-ENN* της οποίας η πλήρης ονομασία είναι *Synthetic Minority Oversampling Technique- Edited Nearest Neighbor* η οποία είναι υβριδική, καθώς χρησιμοποιεί ταυτόχρονα μεθόδους Under-sampling και Oversampling. Με αυτό το τρόπο βελτιώνεται η απόδοση των μοντέλων ταξινόμησης για τα δείγματα που δημιουργούνται με αυτή τη μέθοδο. Αυτή η μέθοδος ξεκινάει αμέσως μόλις ολοκληρωθεί η μέθοδος SMOTE. Συγκεκριμένα, το ENN είναι μέθοδος Under-sampling όπου υπολογίζονται οι κοντινότεροι γείτονες κάθε δεδομένου της κύριας κλάσης. Αν οι κοντινότεροι γείτονες ταξινομήσουν λανθασμένα το συγκεκριμένο δείγμα της κύριας κλάσης, τότε αυτό διαγράφεται. Η ενσωμάτωση της τεχνικής αυτής με Oversampled δεδομένα που δημιουργήθηκαν από τη SMOTE βοηθάει σε εκτεταμένη εκκαθάριση των δεδομένων. Στη μέθοδο αυτή τα δείγματα που ταξινομούνται λανθασμένα από τους κοντινότερους γείτονες τους και από τις δύο κλάσεις διαγράφονται με αποτέλεσμα ένα πιο καθαρό και συνοπτικό διαχωρισμό των κλάσεων.

4.5 Αποτελέσματα

Στο λογισμικό Orange data mining, έτρεξε η αρχική βάση δεδομένων, στην οποία όμως υπήρξε το πρόβλημα της άνισης κατανομής των κλάσεων που αναφέρθηκε και προηγουμένως. Για αυτό

εφαρμόστηκαν και έτρεξαν στο συγκεκριμένο πρόγραμμα ξεχωριστά όλες οι προσεγγιστικές βάσεις δεδομένων που δημιουργήθηκαν με τις τεχνικές resample που αναφέρθηκαν παραπάνω. Τα αποτελέσματα των μεθόδων αυτών αλλά και της αρχικής βάσης δεδομένων, για τα οποία ο ταξινομητής αποδίδει καλύτερα θα παρουσιαστούν παρακάτω.

Αρχικά, πραγματοποιείται η είσοδος μέσω του widget **file** της κάθε βάσης δεδομένων, η οποία ενώνεται με ένα **data table** που απεικονίζει τα δεδομένα της. Εκείνος στη συνέχεια ενώνεται με ένα widget που περιέχει τον ταξινομητή **neural nets** και ένα widget που ονομάζεται **test and score**. Στα νευρωνικά δίκτυα υπάρχουν και οι υπερπαράμετροι όπου για διαφορετικές τιμές τους τρέχει ο αλγόριθμος και δίνει διαφορετικά αποτελέσματα. Το widget neural nets συνδέεται και αυτό με το widget test and score. Από το widget test and score, λαμβάνονται τα μέτρα απόδοσης του αλγόριθμου (AUC, F1 Score, Recall, Precision, Classification Accuracy), τα οποία για διαφορετικές τιμές των υπερπαραμέτρων δίνουν διαφορετικά αποτελέσματα. Το widget αυτό στη συνέχεια συνδέεται με τα widget **confusion matrix** και **predictions**. Το πρώτο διαθέτει τον πίνακα σύγχυσης όπου δείχνει το σύνολο των δεδομένων για τα οποία πραγματοποιήθηκε επιτυχώς η πρόβλεψη και σε ποια όχι, ενώ το δεύτερο παρουσιάζει αναλύτικά όλα τα δεδομένα, σε ποια έγινε σωστή πρόβλεψη, σε ποια όχι, καθώς και τα χαρακτηριστικά που τα διέπουν.



Εικόνα 6) Απεικόνιση του συστήματος στο λογισμικό Orange Data Mining

Κάθε αλγόριθμος έτρεξε και έδωσε αποτελέσματα για διαφορετικές τιμές των υπερπαραμέτρων. Τα αποτελέσματα στα οποία κάθε αλγόριθμος έχει την μεγαλύτερη απόδοση θα παρουσιαστούν εκτενώς παρακάτω. Οι υπερπαράμετροι των νευρωνικών δικτύων είναι:

- ο αριθμός των νευρώνων των κρυφών στρωμάτων (neurons in hidden layers)
- η συνάρτηση ενεργοποίησης (Activation)
- ο Solver

Κάθε αλγόριθμος έτρεξε για όλους τους πιθανούς συνδυασμούς των τριών αυτών υπερπαραμέτρων. Ο αριθμός των νευρώνων των κρυφών στρωμάτων (Neurons in hidden layers) λάμβανε τιμές [(10,5), (20,10), (50,25), (100,50)], η συνάρτηση ενεργοποίησης (Activation) ήταν:[Identity, Logistic, Tanh, ReLU] και τέλος ο Solver που ήταν:[L-BFGS-B, SGD, ADAM]. Ταυτόχρονα, στον ταξινομητή ο αριθμός των επαναλήψεων ήταν σταθερός και ίσος με 200. Τα αποτελέσματα για τους διαφορετικούς συνδυασμούς των υπερπαραμέτρων παρουσιάζονταν στο widget test and score, στο οποίο για καλύτερο έλεγχο της απόδοσης του μοντέλου και μείωση της μεροληψίας επιλέγεται η μέθοδος k-fold cross validation , όπου k=10 και ο διαχωρισμός των δεδομένων πραγματοποιείται σε δύο σετ, όπου ως εκπαιδευτικά λειτουργούν το 80%, ενώ το υπόλοιπο 20% λειτουργεί ως δεδομένα ελέγχου.

4.5.1 Αποτελέσματα Data

Αρχικά, ο αλγόριθμος έτρεξε για όλες τις τιμές των υπερπαραμέτρων στην αρχική βάση δεδομένων και όπως αναφέρθηκε προηγουμένως, επειδή τα δεδομένα της μειονοτικής κλάσης είναι αρκετά λιγότερα από εκείνα της κύριας, υπάρχει το πρόβλημα της άνισης κατανομής των κλάσεων. Αυτό απεικονίζεται και στα αποτελέσματα παρακάτω.

Ο αλγόριθμος έτρεξε για όλους τους συνδυασμούς των υπερπαραμέτρων που αναφέρθηκαν παραπάνω. Εξαιτίας της άνισης κατανομής των κλάσεων, οι τιμές των CA, Recall και F1 παρέμεναν σταθερές για όλους τους συνδυασμούς των υπερπαραμέτρων, καθώς CA=0.853, Recall=0.853 και F1=0.786, ενώ οι τιμές των AUC και Precision είχαν μικρές εναλλαγές τιμών. Οι μεγαλύτερες τιμές για το AUC λαμβάνονται για Neurons in hidden layers (100,50), Activation logistic και τον Solver Adam. Ωστόσο, τα ίδια αποτελέσματα υπάρχουν σχεδόν και για τις άλλες τιμές του αριθμού των νευρώνων των κρυφών στρωμάτων, ενώ για όλες τις άλλες τιμές των υπερπαραμέτρων λόγω της άνισης κατανομής των κλάσεων τα αποτελέσματα είναι παρεμφερή και απεικονίζονται στο Test and Score widget, ενώ παρουσιάζεται και ο πίνακας σύγχυσης (Confusion Matrix).

	0	1	Σ
0	54606	0	54606
1	9394	0	9394
Σ	64000	0	64000

Πίνακας 1) Confusion Matrix για την αρχική βάση δεδομένων

Για τους διαφορετικούς συνδυασμούς των υπερπαραμέτρων, προκύπτουν σχεδόν παρόμοια ή σε χαμηλότερη απόδοση αποτελέσματα. Συμπεραίνεται, ότι ο αλγόριθμος δεν είναι ικανός να διαχωρίσει τη κύρια κλάση από τη μειονοτική, με αποτέλεσμα να προβλέπει τους πελάτες οι οποίοι έχουν αποχωρήσει από την εταιρεία ως πελάτες οι οποίοι συνεχίζουν τις αγορές τους από την εταιρεία. Για αυτό το λόγο, δημιουργήθηκαν και άλλες βάσεις δεδομένων με διαφορετικές προσεγγίσεις με τις τεχνικές resample που παρουσιάστηκαν παραπάνω και τα αποτελέσματα για τα οποία ο αλγόριθμος αποδίδει καλύτερα θα παρουσιαστούν στη συνέχεια.

4.5.2 Αποτελέσματα Random Oversampling

Στη συγκεκριμένη τεχνική, δημιουργήθηκε μία νέα προσεγγιστική βάση δεδομένων με την τεχνική Random Oversample, όπως παρουσιάστηκε παραπάνω, ενώ πραγματοποιήθηκαν και οι τεχνικές προεπεξεργασίας που αναφέρθηκαν πιο πριν, ώστε να αντιμετωπιστεί το πρόβλημα της άνισης κατανομής των κλάσεων. Ο αλγόριθμος αποδίδει καλύτερα στη συγκεκριμένη βάση δεδομένων για τις ακόλουθες τιμές των υπερπαραμέτρων: Neurons in hidden layers (100,50), Activation tanh και Solver adam, όπου λαμβάνει τιμές AUC ίση με 0.732, CA, Recall και F1 ίσα με 0.668 και Precision ίση με 0.669.

	0	1	Σ
0	35144	19462	54606
1	16751	37855	54606
Σ	51895	57317	109212

Πίνακας 2) Confusion Matrix για τη νέα προσεγγιστική βάση δεδομένων με τη μέθοδο Random Oversampling

Αρχικά, παρατηρείται αύξηση του αριθμού της βάσης δεδομένων, καθώς τα δεδομένα που εισέρχονται από 64000, είναι πλέον 109212. Αυτό συμβαίνει, διότι στη νέα προσέγγιση της βάσης δεδομένων, χρησιμοποιήθηκε η τεχνική Random Oversample, η οποία παράγει τυχαία αντίγραφα στη μειονοτική κλάση, με αποτέλεσμα ο αριθμός των δεδομένων της να ισούται με τον αριθμό των δεδομένων της κύριας κλάσης, ώστε να μην υπάρχει το πρόβλημα της άνισης κατανομής των κλάσεων. Ο αλγόριθμος αποδίδει καλύτερα, καθώς αυξάνονται οι τιμές των μέτρων απόδοσης του, ωστόσο τα αποτελέσματα δεν είναι ακόμα ικανοποιητικά. Παρατηρείται ότι από τους συνολικά 54606 πελάτες που

παραμένουν στην εταιρεία, σωστά υπολογίστηκαν οι 35144, άρα το ποσοστό των 64,36%, ενώ από τους 54.606 πελάτες που αποχωρούν, υπολογίστηκε σωστά το 69,32%.

4.5.3 Αποτελέσματα Random Under-sampling

Μία άλλη τεχνική που χρησιμοποιήθηκε για την αντιμετώπιση της άνισης κατανομής των κλάσεων, είναι η Random Under-sampling. Αρχικά και στη συγκεκριμένη τεχνική υπήρξε η προεπεξεργασία των δεδομένων και πραγματοποιήθηκε η τυχαία διαγραφή δεδομένων από την κύρια κλάση της αρχικής βάσης. Αυτό έγινε για την αντιμετώπιση της άνισης κατανομής των κλάσεων και η νέα βάση δεδομένων, αποτελείται από συνολικά 18788. Έτσι, ο αριθμός της κύριας κλάσης ισούται με τον αριθμό της μειονοτικής. Ο αλγόριθμος έχει περίπου την ίδια απόδοση για διάφορους συνδυασμούς των υπερπαραμέτρων και με πολύ μικρή διαφορά αποδίδει καλύτερα για τους αντίστοιχους συνδυασμούς: Neurons in hidden layers (10,5), Activation logistic, Solver adam. Τα αποτελέσματα απόδοσης είναι AUC ίσο με 0.643 και οι τιμές των CA, Precision, Recall, F1 ίσες με 0.605.

	0	1	Σ
0	5368	4026	9394
1	3395	5999	9394
Σ	8763	10025	18788

Πίνακας 3) Confusion Matrix για τη νέα προσεγγιστική βάση δεδομένων με τη μέθοδο Random Undersampling

Από τα παραπάνω παρατηρείται πως ο αλγόριθμος έχει πολύ χαμηλή απόδοση, χαμηλότερη από τον Random Oversample. Τα αποτελέσματα αυτής της τεχνικής δεν είναι ικανοποιητικά και από αυτό μπορεί να γίνει εύκολα κατανοητό, ότι είναι πολύ πιθανό να έγινε η διαγραφή δεδομένων της κύριας κλάσης, τα οποία θα περιείχαν χρήσιμες πληροφορίες για το μοντέλο. Ο αλγόριθμος προβλέπει σωστά το 57,14% των 9394 πελατών που παρέμειναν στην εταιρεία και για εκείνους που αποχώρησαν το 63,86%.

4.5.4 Αποτελέσματα SMOTE

Η μέθοδος SMOTE, αποτελεί μία μέθοδο Oversample, καθώς συνθέτει αντίγραφα των δεδομένων της μειονοτικής κλάσης. Αυτό γίνεται επιλέγοντας ένα τυχαίο δεδομένο αρχικά και στη συνέχεια συνθέτει αντίγραφα του, τα οποία αποτελούν τους k-nearest neighbors. Και σε αυτή τη μέθοδο αρχικά έγινε μια προεπεξεργασία των δεδομένων, όπως στις υπόλοιπες τεχνικές resample. Έτσι, δημιουργείται μία νέα βάση δεδομένων, η οποία αποδίδει καλύτερα για τους εξής συνδυασμούς των υπερπαραμέτρων: Neurons in hidden layers (100,50), Activation tanh, Solver adam. Τα αποτελέσματα δείχνουν ότι το AUC έχει τιμή ίση με 0.726, τα CA, Precision, Recall ίσα με 0.664 και η F1 ίση με 0.663.

	0	1	Σ
0	34435	20171	54606
1	16556	38050	54606
Σ	50991	58221	109212

Πίνακας 4) Confusion Matrix για τη νέα προσεγγιστική βάση δεδομένων με τη μέθοδο SMOTE

Με τη συγκεκριμένη μέθοδο, η νέα βάση δεδομένων αποτελείται από 109212 δεδομένα, όπως ακριβώς και στη μέθοδο Random Oversample. Η διαφορά υφίσταται, στο τρόπο με τον οποίο δημιουργούνται τα νέα δεδομένα με αυτή τη μέθοδο, καθώς δεν δημιουργούνται τυχαία στη μειονοτική κλάση, αλλα πραγματοποιείται η σύνθεση τους με τον τρόπο που επεξηγήθηκε προηγουμένως. Η μέθοδος φέρει καλύτερα αποτελέσματα από τη Random Under-sample, ωστόσο τα αποτελέσματά της είναι παρόμοια με εκείνα της Random Oversample, δηλαδή καλύτερα από τις άλλες μεθόδους, αλλά όχι τα βέλτιστα δυνατά. Ένα μειονέκτημα της μεθόδου αυτής που ενδέχεται να έχει ως αποτέλεσμα την όχι και τόσο ικανοποιητική απόδοση του μοντέλου, οφείλεται στο γεγονός ότι με τη συγκεκριμένη μέθοδο παράγεται μεγάλος αριθμός θορυβώδη δεδομένων με αποτέλεσμα το μοντέλο να επηρεάζεται και να μην αποδίδει σε ικανοποιητικό βαθμό. Εδώ, ο αλγόριθμος υπολόγισε σωστά τους πελάτες που παραμένουν σε ποσοστό 63,06% και εκείνους που αποχωρούν σε ποσοστό 69,68%.

4.5.5 Αποτελέσματα SMOTE-ENN

Η συγκεκριμένη μέθοδος όπως παρουσιάστηκε παραπάνω, αποτελεί παραλλαγή της μεθόδου SMOTE. Είναι υβριδική μέθοδος, καθώς εφαρμόζεται αρχικά η τεχνική SMOTE, με αποτέλεσμα την αύξηση του αριθμού των δεδομένων της μειονοτικής κλάσης και στη συνέχεια με τους edited nearest neighbors, εντοπίζονται οι πλησιέστεροι γείτονες κάθε δεδομένου και στις δύο κλάσεις, το οποίο αν ταξινομηθεί λανθασμένα διαγράφεται. Με αυτό το τρόπο πραγματοποιείται μία εκκαθάριση των δεδομένων. Και εδώ όπως σε όλες τις τεχνικές resample, υπήρξε αρχικά μια προεπεξεργασία των δεδομένων και στη συνέχεια εφαρμόστηκε η συγκεκριμένη τεχνική που είχε ως αποτέλεσμα μια νέα βάση δεδομένων. Ο αλγόριθμος της συγκεκριμένης τεχνικής, αποδίδει καλύτερα με τους εξής συνδυασμούς υπερπαραμέτρων: Neurons in hidden layers(100,50), Activation tanh, Solver adam. Τα αποτελέσματα των μέτρων απόδοσης, είναι AUC ίσο με 0.963 και για τα υπόλοιπα μέτρα απόδοσης CA, Precision, Recall, F1 ίσα με 0.908.

	0	1	Σ
0	21379	2024	23403
1	1996	18201	20197
Σ	23375	20225	43600

Πίνακας 5) Confusion Matrix για τη νέα προσεγγιστική βάση δεδομένων με τη μέθοδο SMOTE-ENN

Στη συγκεκριμένη τεχνική, είναι εμφανές ότι το μοντέλο αποδίδει σε αρκετά ικανοποιητικό βαθμό, καθώς όλα τα μέτρα απόδοσης έχουν βαθμό μεγαλύτερο του 0.9. Η νέα βάση δεδομένων που προκύπτει είναι μικρότερη από την αρχική, καθώς αυτή αποτελείται από 43600 δεδομένα. Αυτό οφείλεται στον καθαρισμό των δεδομένων που προσδίδει σαν αποτέλεσμα η μέθοδος SMOTE-ENN, καθώς αρχικά συνθέτει δεδομένα στη μειονοτική κλάση, ενώ στη συνέχεια διαγράφει δεδομένα και από τις δύο κλάσεις τα οποία δεν χρειάζονται στο μοντέλο. Η μέθοδος αποδίδει καλύτερα από όλες τις άλλες μεθόδους και για αυτό προτιμάται η χρήση της στην εκπαίδευση του συγκεκριμένου μοντέλου, συγκριτικά με τις άλλες μεθόδους. Στη συγκεκριμένη μέθοδο, το μοντέλο προέβλεψε λανθασμένα ότι 2024 αποχώρησαν από την εταιρεία, ενώ κανονικά παρέμειναν όπου είναι σε ποσοστό 8.65%, ενώ προέβλεψε λανθασμένα ότι 1996 πελάτες παρέμειναν στην εταιρεία, ενώ στην πραγματικότητα αποχώρησαν από αυτή και μεταφράζεται σε ποσοστό ίσο με 9.88% . Συνολικά δηλαδή δεν κατάφερε να προβλέψει ορθά 4022 πελάτες, ή αλλιώς σε ποσοστό το 9,22% της συνολικής βάσης δεδομένων.

Σε όλες τις παραπάνω τεχνικές μαζί με την αρχική βάση δεδομένων, εφαρμόστηκαν οι παραπάνω συνδυασμοί των υπερπαραμέτρων, με αριθμό επαναλήψεων (Iterations) ίσο με 200, με σκοπό την εύρεση εκείνων των συνδυασμών, όπου το μοντέλο εκπαιδεύεται καλύτερα και οι αλγόριθμοι αποδίδουν καλύτερα στις βάσεις δεδομένων τους. Όπως αποδείχτηκε για τη βάση δεδομένων SMOTE-ENN, το μοντέλο αποδίδει καλύτερα σε σχέση με τις άλλες βάσεις δεδομένων, καθώς για συγκεκριμένο αριθμό υπερπαραμέτρων η τιμή AUC βρίσκεται στο 0.967, ενώ οι υπόλοιπες 4 στην τιμή 0.908. Γενικά, η δημιουργία των διαφορετικών βάσεων οφείλεται στο γεγονός ότι στην αρχική βάση δεδομένων, εξαιτίας της μεγάλης διαφοράς των δύο κλάσεων που μελετώνται που είναι η αποχώρηση ή η διατήρηση των πελατών το μοντέλο αγνοεί την μειονοτική κλάση με αποτέλεσμα όπως απεικονίστηκε παραπάνω να προβλέπει λανθασμένα πως η εταιρεία θα διατηρήσει όλους τους πελάτες της. Αυτό το αποτέλεσμα προέκυψε σχεδόν για όλους τους συνδυασμούς των υπερπαραμέτρων. Για αυτό το λόγο χρησιμοποιήθηκαν οι παραπάνω τεχνικές, αφού αρχικά πραγματοποιήθηκαν οι δύο μέθοδοι προεπεξεργασίας των δεδομένων και δημιουργήθηκαν οι νέες βάσεις δεδομένων πάνω στις οποίες έτρεξε ο ταξινομητής.

Στη μέθοδο Random Under-sampling, το μοντέλο συνέχιζε να μην αποδίδει καθώς τα μέτρα απόδοσης του αλγορίθμου είχαν αρκετά χαμηλές τιμές. Στις μεθόδους Random Oversampling και SMOTE το μοντέλο αποδίδει καλύτερα, αλλά και πάλι όχι σε ικανοποιητικό βαθμό, όπως αντίθετα συμβαίνει με την τεχνική SMOTE-ENN. Επιπλέον, αξίζει να σημειωθεί πως στις τρεις τελευταίες μεθόδους, για κάθε μέθοδο ξεχωριστά ανεξάρτητα τους διάφορους συνδυασμούς η Συνάρτηση Ενεργοποίησης Identity, η οποία έχει και την πιο απλή μορφή έδινε παρόμοια αποτελέσματα. Συγκεκριμένα, στην μέθοδο Random Over-sampling η τιμή AUC κυμαινόταν περίπου στο 0.647, στη

μέθοδο SMOTE ήταν περίπου 0.645 και στη μέθοδο SMOTE-ENN περίπου 0.789. Τέλος, στην αρχική βάση δεδομένων και σε εκείνη που δημιουργήθηκε με τη μέθοδο Random Under-sampling το μοντέλο παρά τις όποιες αλλαγές των υπερπαραμέτρων δεν σημείωνε μεγάλες διαφορές στις τιμές των μέτρων απόδοσης.

Κεφάλαιο 5: Συμπεράσματα Έρευνας

5.1 Benchmarking μεθόδων resample

Στη συγκεκριμένη έρευνα, χρησιμοποιήθηκε το λογισμικό Orange Data Mining όπου αρχικά υπήρξε ένα σετ δεδομένων το οποίο διαχωρίστηκε σε δύο κλάσεις (**binary classification problem**) σε πελάτες που αποχώρησαν και πελάτες που διατηρήθηκαν σε μια εταιρεία. Στην αρχική βάση δεδομένων υπήρξε το πρόβλημα της άνισης κατανομής των δύο κλάσεων (**Class imbalance problem**) λόγω του γεγονότος ότι η κύρια κλάση διέθετε περισσότερα δεδομένα από την μειονοτική με αποτέλεσμα το μοντέλο να αγνοεί τη μειονοτική κλάση. Για να αντιμετωπιστεί το συγκεκριμένο πρόβλημα χρησιμοποιήθηκαν τεχνικές resample, οι οποίες δημιούργησαν νέες προσεγγιστικές βάσεις δεδομένων πάνω στις οποίες έτρεξε ο ταξινομητής των νευρωνικών δικτύων με σκοπό τη δημιουργία ενός μοντέλου ικανού να διαχωρίζει και να διανέμει τα δεδομένα στις κλάσεις που αντιστοιχούν. Για καλύτερη απόδοση του μοντέλου δοκιμάστηκαν διάφοροι συνδυασμοί των υπερπαραμέτρων των νευρωνικών δικτύων (**Hyperparameter Tuning**). Τα κύρια μέτρα απόδοσης του μοντέλου με τα οποία έγινε η σύγκριση των τεχνικών resample ήταν ο δείκτης **F1** που αποτελεί τον αρμονικό μέσο των δεικτών **Precision** και **Recall**, η τιμή του δείκτη **AUC** που αποτελεί το μέτρο της ικανότητας του ταξινομητή να διαχωρίζει μεταξύ των κλάσεων και ο δείκτης **CA** (Classification Accuracy) που χρησιμοποιείται για να μετρήσει κατά πόσο ο ταξινομητής προβλέπει σωστά. Ο δείκτης Precision εξηγεί πόσες από τις θετικά προβλεπόμενες υποθέσεις είναι πραγματικά θετικές και τέλος ο δείκτης Recall αναδεικνύει σε πόσες από τις πραγματικές θετικές υποθέσεις μπορούσε να γίνει ορθή πρόβλεψη με το μοντέλο.

Από όλες τις τεχνικές resample, εκείνη η οποία έχει ως αποτέλεσμα την καλύτερη απόδοση του μοντέλου με αποτέλεσμα να γίνεται ικανό να πραγματοποιήσει προβλέψεις για συγκεκριμένο συνδυασμό των τιμών των υπερπαραμέτρων, είναι η τεχνική **SMOTE-ENN**. Η τεχνική **Random Under-sampling** έχει αρκετά χαμηλό βαθμό απόδοσης, ενώ οι τεχνικές **Random Oversampling** και **SMOTE**, αποδίδουν καλύτερα από την Random Under-sampling, αλλά όχι σε τόσο μεγάλο βαθμό ώστε να μπορεί το μοντέλο να πραγματοποιήσει προβλέψεις.

Για όλες τις τεχνικές resample, εφαρμόστηκαν αρχικά οι τεχνικές προεπεξεργασίας των δεδομένων one-hot encoding και MinMaxScaler, οι οποίες έχουν ως αποτέλεσμα την μετατροπή των κατηγορικών δεδομένων σε αριθμητικά και τα δεδομένα λαμβάνουν ένα εύρος τιμών μεταξύ 0 και 1. Αυτό γίνεται γιατί ορισμένοι αλγόριθμοι αποδίδουν μόνο σε αριθμητικές τιμές των δεδομένων και επιπλέον άλλοι καλύτερα αν οι τιμές των δεδομένων κυμαίνονται σε ένα εύρος τιμών μεταξύ του 0 και 1.

Για συγκεκριμένους συνδυασμούς υπερπαραμέτρων σε κάθε τεχνική resample ξεχωριστά, οι καλύτερες τιμές των μέτρων απόδοσης κάθε τεχνικής ξεχωριστά απεικονίζονται παρακάτω:

	AUC	CA	F1	Precision	Recall
1.Random Undersampling	**0.643**	**0.605**	**0.605**	**0.605**	**0.605**
2. Random Oversampling	**0.732**	**0.668**	**0.668**	**0.669**	**0.668**
3. SMOTE	**0.726**	**0.664**	**0.663**	**0.664**	**0.664**
4. SMOTE-ENN	**0.963**	**0.908**	**0.908**	**0.908**	**0.908**

Πίνακας 6) Benchmarking τεχνικών resample

5.2 Γενικά Συμπεράσματα

Στη συγκεκριμένη έρευνα πραγματοποιήθηκε αρχικά ο διαχωρισμός των πελατών σε δύο κλάσεις, σε εκείνους που παρέμειναν (**non-churners** με τιμή 0) και εκείνους που αποχώρησαν από την εταιρεία (**churners** με τιμή 1), καθώς η μεταβλήτη conversion αποτελούσε τη μεταβλητή-στόχο. Οι πελάτες που παραμένουν στην εταιρεία είναι πολύ περισσότεροι από εκείνους που τελικά αποχωρούν καθώς αποτελούν το 85.32% της βάσης δεδομένων. Αυτό έχει ως αποτέλεσμα να υπάρχει μία ανισορροπία στις κλάσεις, καθώς το μοντέλο αγνοεί την μειονοτική κλάση και αρχίζει να εκπαιδεύεται με τα δεδομένα της κύριας. Έτσι, το μοντέλο καταφέρνει και έχει υψηλό δείκτη ακρίβειας, καθώς το μοντέλο έχει Classification Accuracy ίσο με 0.853 το οποίο είναι αρκετά ικανοποιητικό ποσοστό και ισούται με το ποσοστό των πελατών της κύριας κλάσης στη βάση δεδομένων. Το μοντέλο όμως, με αυτό το τρόπο δεν είναι ικανό να γενικεύσει, αλλά ούτε και να προβλέψει, καθώς λανθασμένα αγνοεί τη μειονοτική κλάση και προβλέπει πως ακόμα και οι πελάτες που αποχώρησαν, παραμένουν ακόμα στην εταιρεία. Για αυτό τον τρόπο λαμβάνονται υπόψιν άλλα μέτρα απόδοσης, όπως είναι η τιμή του AUC, η τιμή του δείκτη F1, καθώς και ο Confusion Matrix. Σε αυτόν απεικονίζονται όσοι πελάτες που ορθά προέβλεψε το μοντέλο ότι παραμένουν (True Positive ή TP), όσοι πελάτες ορθά προέβλεψε το μοντέλο ότι αποχωρούν (True Negative ή TN), αλλά και όσοι πελάτες αντιστοιχήθηκαν στο μοντέλο σε διαφορετικές κλάσεις από αυτές που τους αναλογούν. Επειδή όμως οι τιμές αυτών των μέτρων απόδοσης ήταν χαμηλές, δημιουργήθηκαν νέες βάσεις δεδομένων με διαφορετικές προσεγγίσεις με τεχνικές resample, ώστε να αντιμετωπιστεί το πρόβλημα της ανισορροπίας των κλάσεων. Με αυτό το τρόπο

δημιουργήθηκαν νέες βάσεις δεδομένων με διαφορετικές προσεγγίσεις, πάνω στις οποίες έτρεξε ο ταξινομητής, για να βρεθεί ο αλγόριθμος στον οποίο ο ταξινομητής μπορεί να αποδώσει καλύτερα.

Για τη δημιουργία των νέων προσεγγιστικών βάσεων δεδομένων αρχικά πραγματοποιήθηκε μια προεπεξεργασία των δεδομένων. Με τη μέθοδο one-hot encoding, έγινε η μετατροπή των κατηγορικών χαρακτηριστικών σε αριθμητικά. Ακολούθως, με τη μέθοδο MinMaxScaler, όλα τα χαρακτηριστικά των μεταβλητών των δεδομένων λαμβάνουν ένα εύρος τιμών μεταξύ 0 και 1. Αυτές οι μετατροπές των μεταβλητών γίνονται διότι ορισμένοι αλγόριθμοι δεν αποδίδουν σε κατηγορικά δεδομένα ή αποδίδουν καλύτερα σε δεδομένα που αποτελούνται από αυτές τις τιμές.

Με τις τεχνικές resample δημιουργούνται νέες προσεγγιστικές βάσεις δεδομένων, οι οποίες αποτελούνται από διαφορετικό αριθμό σε σύγκριση με την αρχική. Στη τεχνική Random Under-sampling, μειώνονται τυχαία τα δεδομένα της κύριας κλάσης ώστε να είναι σε αριθμό ίσα με εκείνα της κύριας, ενώ στη Random Oversampling αυξάνεται τυχαία ο αριθμός των δεδομένων της μειονοτικής κλάσης μέχρι να είναι ίσος με τον αριθμό της κύριας. Το ίδιο συμβαίνει και στην SMOTE, με την διαφορά ότι εδώ πραγματοποιείται η σύνθεση νέων δεδομένων στη μειονοτική κλάση. Τέλος στην SMOTE-ENN, πραγματοποιείται η σύνθεση δεδομένων όπως και στην SMOTE, με κύρια διαφορά ότι αυτή η μέθοδος είναι υβριδική, καθώς μόλις γίνει η σύνθεση νέων δεδομένων στη μειονοτική κλάση ώστε ο αριθμός της να είναι ίδιος με τον αριθμό της κύριας, πραγματοποιείται μια τεχνική Under-sampling, όπου δεδομένα που δεν είναι στην ίδια κλάση με εκείνα που βρίσκονται στους κοντινότερους τους γείτονες διαγράφονται από τη βάση δεδομένων. Όλες οι μέθοδοι έτρεξαν στον ταξινομητή Neural Nets του προγράμματος Orange Data Mining και σε όλα δοκιμάστηκαν διάφοροι συνδυασμοί των υπερπαραμέτρων. Από αυτούς τους συνδυασμούς, κάθε αλγόριθμος αποδίδει καλύτερα για συγκεκριμένους συνδυασμούς υπερπαραμέτρων. Στη μέθοδο Random Under-sampling, ο ταξινομητής δεν αποδίδει σε ικανοποιητικό βαθμό. Με τις μεθόδους Random Oversampling και SMOTE, αποδίδει καλύτερα αλλά και πάλι τα μέτρα απόδοσης έχουν χαμηλή τιμή. Στην τεχνική SMOTE-ENN, ο ταξινομητής αποδίδει σε αρκετά ικανοποιητικό βαθμό, καθώς λαμβάνει τιμές AUC και του δείκτη F1 ίσες με 0.967 και 0.908 αντίστοιχα, οπότε επιλέγεται αυτός ο αλγόριθμος για την πρόβλεψη της απώλειας των πελατών.

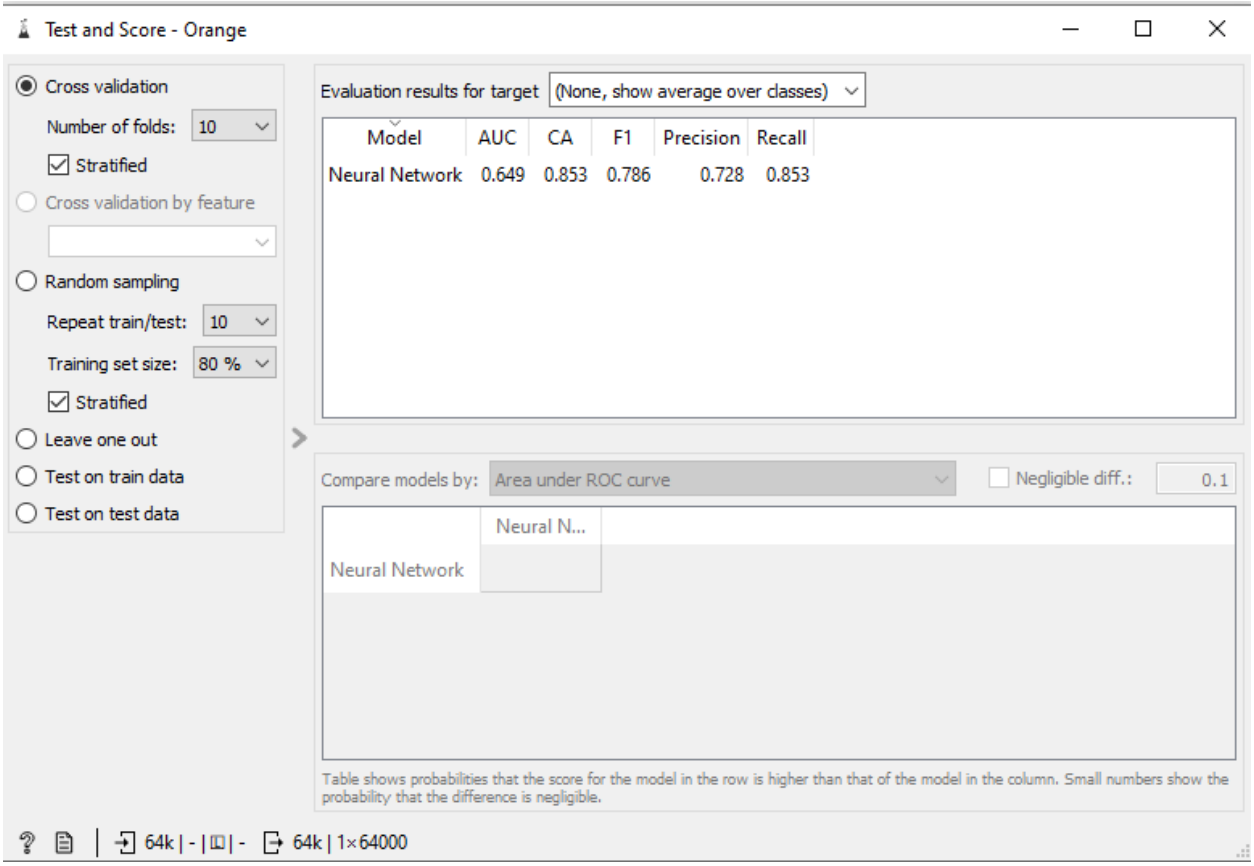
5.3 Αξιοποίηση της παρούσας εργασίας και περαιτέρω έρευνα

Στη συγκεκριμένη εργασία, αναλύθηκαν οι όροι του Στοχευμένου Μάρκετινγκ, της διαχείρισης των πελατειακών σχέσεων, της ικανοποίησης και αφοσίωσης των πελατών, αλλά κυρίως η σημασία της πρόβλεψης της απώλειας των πελατών. Αυτοί οι όροι αποτελούν τη βάση για την ανάπτυξη και εφαρμογή σύγχρονων μεθόδων Μάρκετινγκ από τις επιχειρήσεις. Επίσης, υπήρχε ιδιαίτερη ανάλυση στον όρο νευρωνικά δίκτυα, τον τρόπο υλοποίησης τους, τους τύπους από τους οποίους αποτελούνται, τις υπερπαραμέτρους τους, καθώς και τα μέτρα απόδοσης τους. Επιπλέον, εξηγήθηκε ο όρος των δεδομένων,

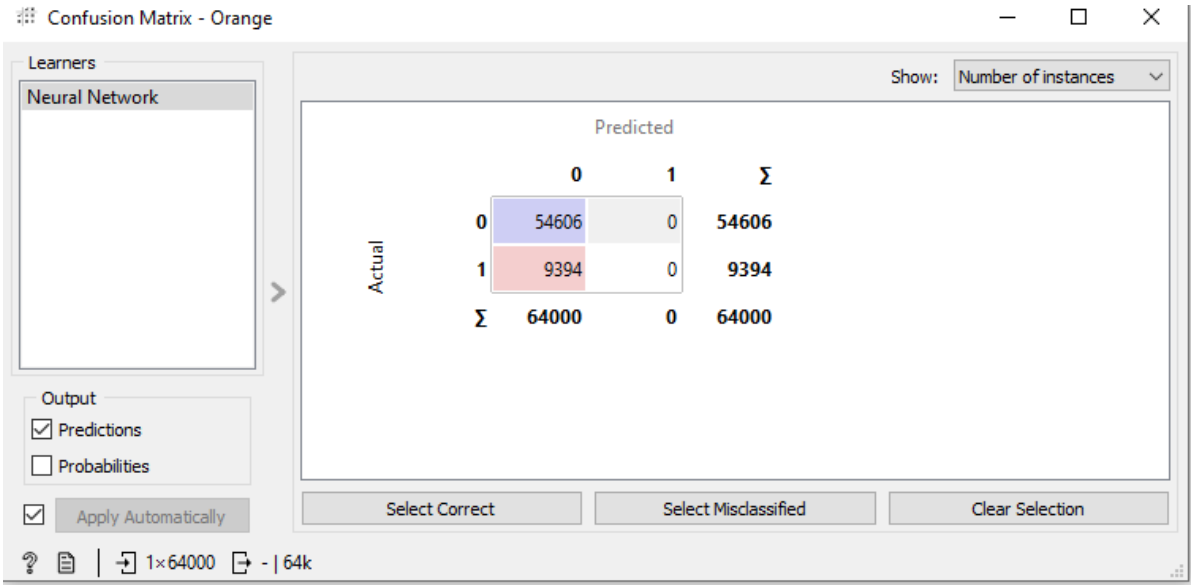
ο τρόπος επεξεργασίας τους, καθώς και οι μέθοδοι διαχωρισμού τους. Στόχος της συγκεκριμένης έρευνας αποτελεί η δημιουργία ενός απλού και ικανοποιητικού μοντέλου ικανού για την πρόβλεψη της απώλειας των πελατών μιας εταιρείας με χρήση νευρωνικών δικτύων.

Η συγκεκριμένη έρευνα αποτελεί μια αρχική προσέγγιση για την χρήση και εφαρμογή των νευρωνικών δικτύων σε θέματα που αφορούν το Μάρκετινγκ και την διαχείριση των πελατών. Οι αλγόριθμοι και ο συντονισμός των υπερπαραμέτρων για τους οποίους το μοντέλο αποδίδει σε ικανοποιητικό βαθμό εφαρμόστηκαν στο λογισμικό Orange Data Mining το οποίο στηρίζεται στην γλώσσα προγραμματισμού Python. Σε συνέχεια της συγκεκριμένης εργασίας, μπορεί να γίνει η χρήση διαφόρων αλγορίθμων **Μηχανικής Μάθησης** για την ταξινόμηση των δεδομένων, όπως είναι οι αλγόριθμοι **Logistic Regression**, **Random Forest**, **k- Nearest Neighbors** και άλλοι. Επιπλέον, μπορεί να γίνει η συσταδοποίηση των δεδομένων (**Clustering**), δηλαδή ο διαχωρισμός των δεδομένων σε συστάδες με κοινά χαρακτηριστικά. Για την επιλογή των χαρακτηριστικών εκείνων που επηρεάζουν περισσότερο το μοντέλο μπορεί να χρησιμοποιηθεί η μέθοδος **Feature Selectio**n και μπορεί να πραγματοποιηθεί και η εξαγωγή κανόνων συσχέτισης (**Association Rules Extraction**), όπου με τον αλγόριθμο **Apriori** δημιουργούνται κανόνες που διέπουν τα χαρακτηριστικά.

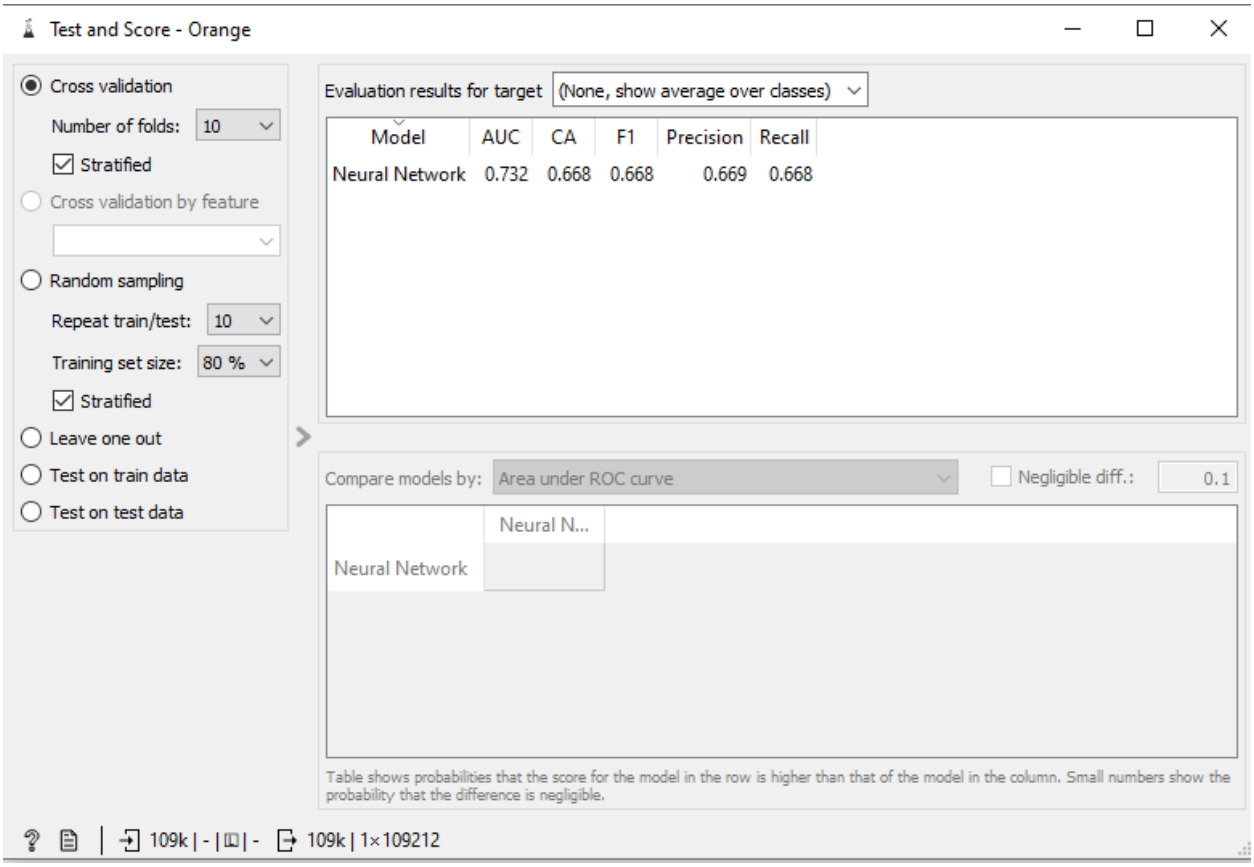
Παράρτημα



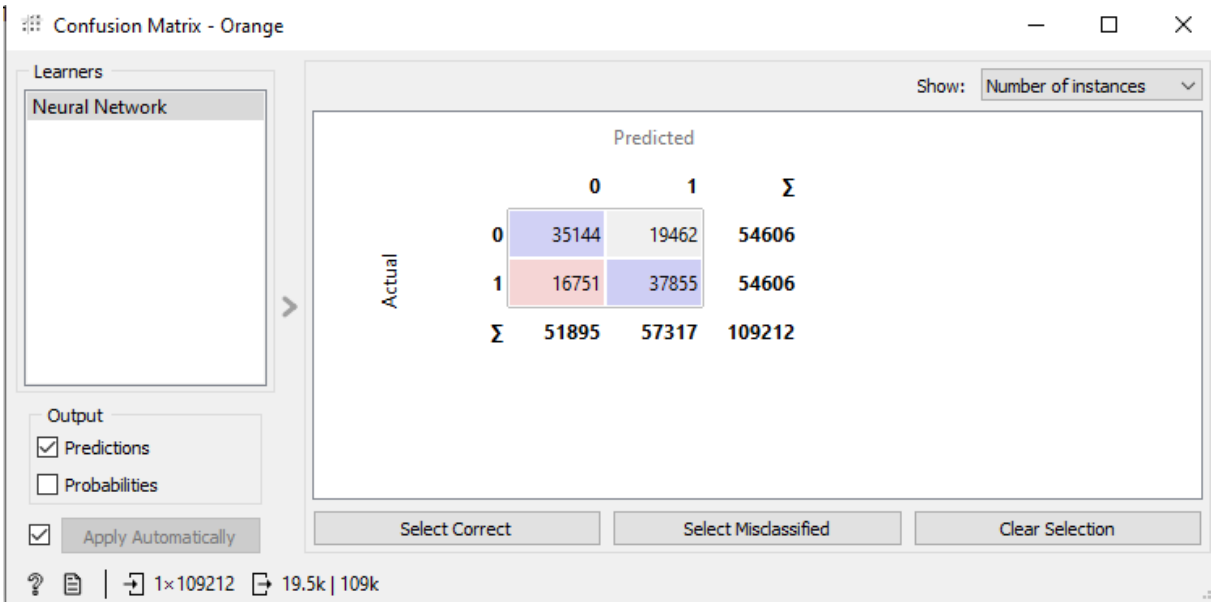
Εικόνα 7) Αποτελέσματα Test and Score για την αρχική βάση δεδομένων



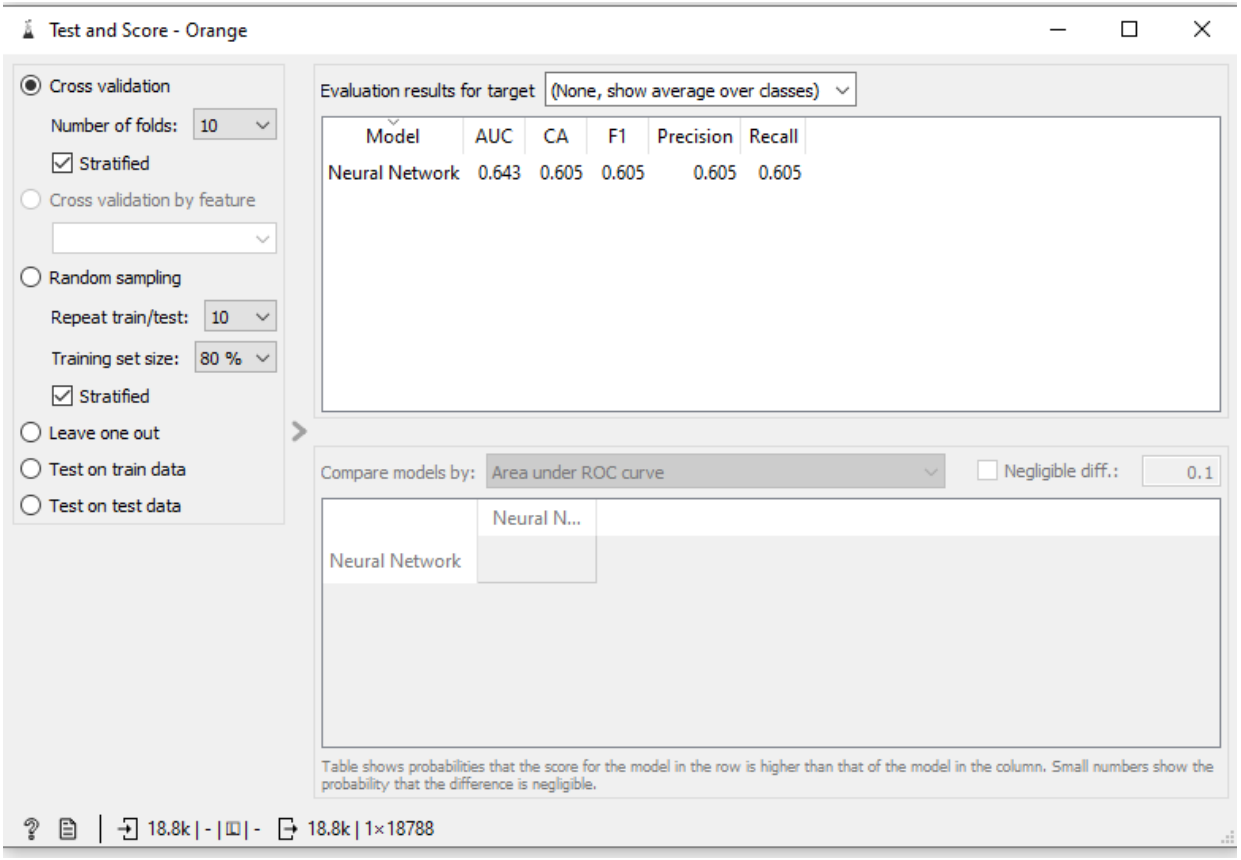
Εικόνα 8) Confusion Matrix για την αρχική βάση δεδομένων



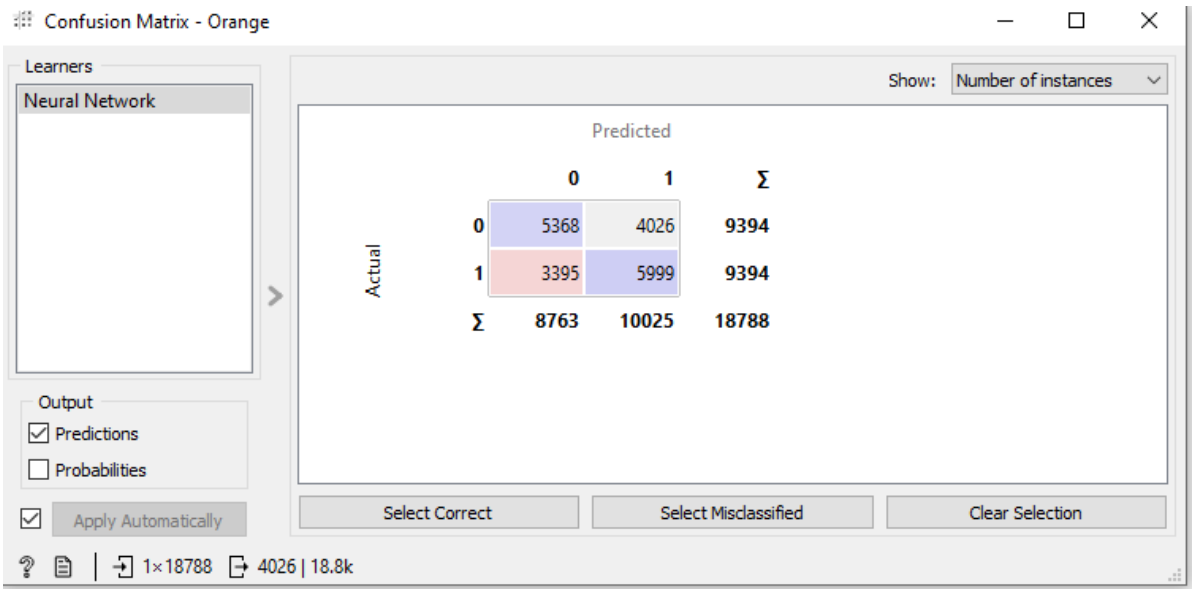
Εικόνα 9) Αποτελέσματα Test and Score για νέα προσεγγιστική βάση δεδομένων Random Oversample



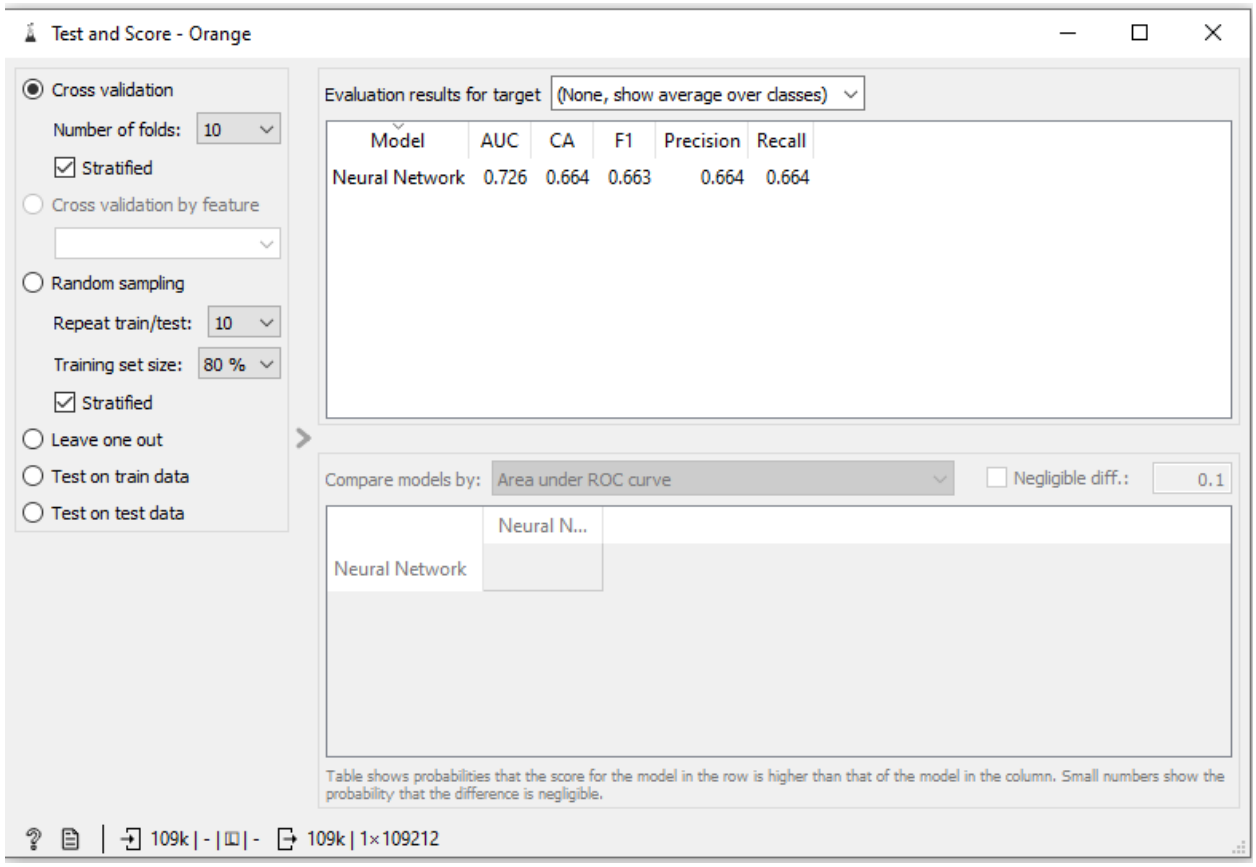
Εικόνα 10) Αποτελέσματα Confusion Matrix για νέα προσεγγιστική βάση δεδομένων Random Oversample



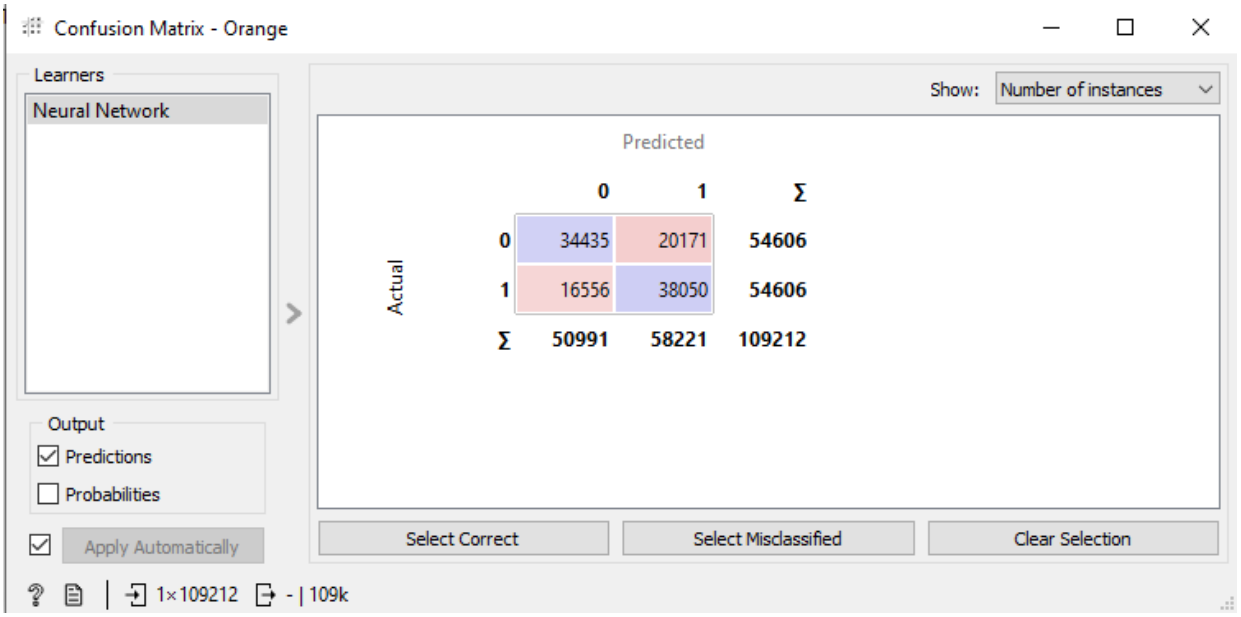
Εικόνα 11) Αποτελέσματα Test and Score για νέα προσεγγιστική βάση δεδομένων Random Under-sampling



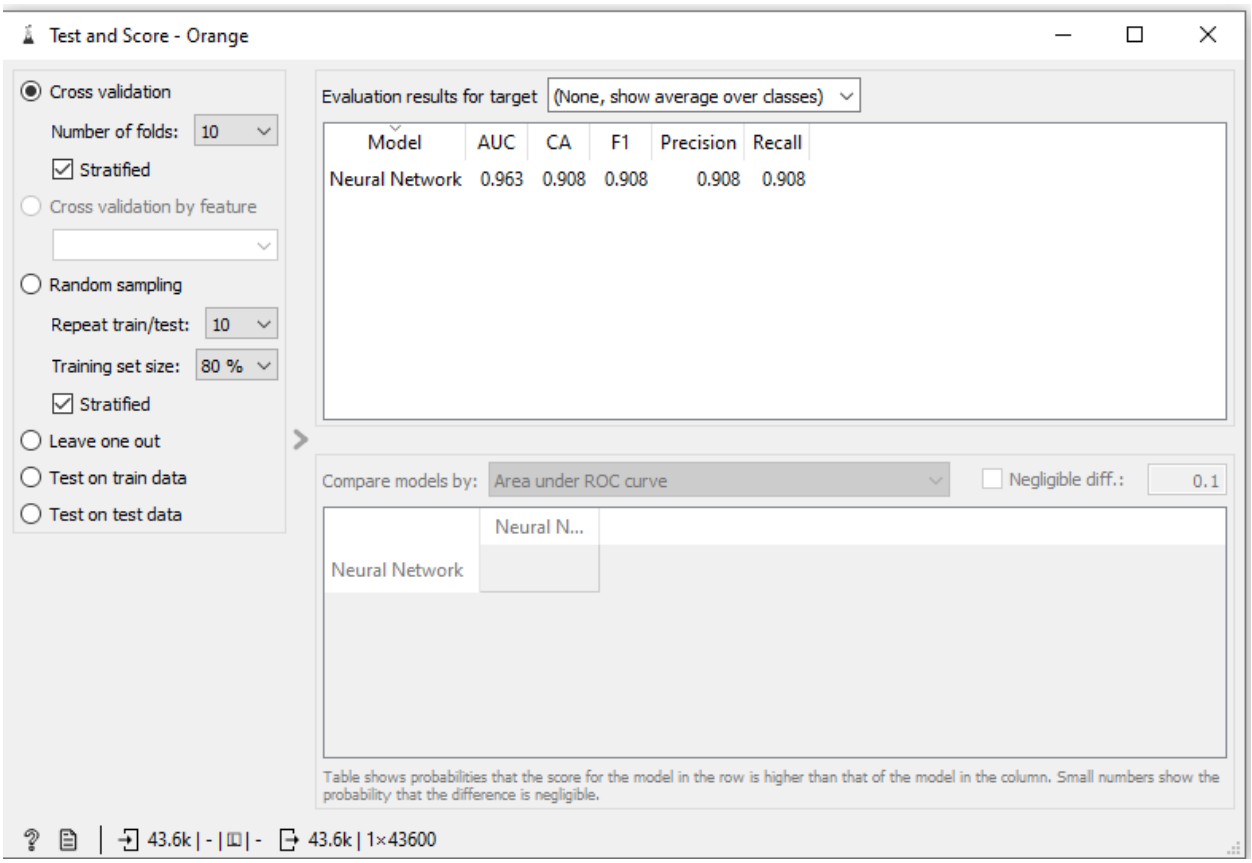
Εικόνα 12) Αποτελέσματα Confusion Matrix για τη νέα προσεγγιστική βάση δεδομένων Random Under-sampling



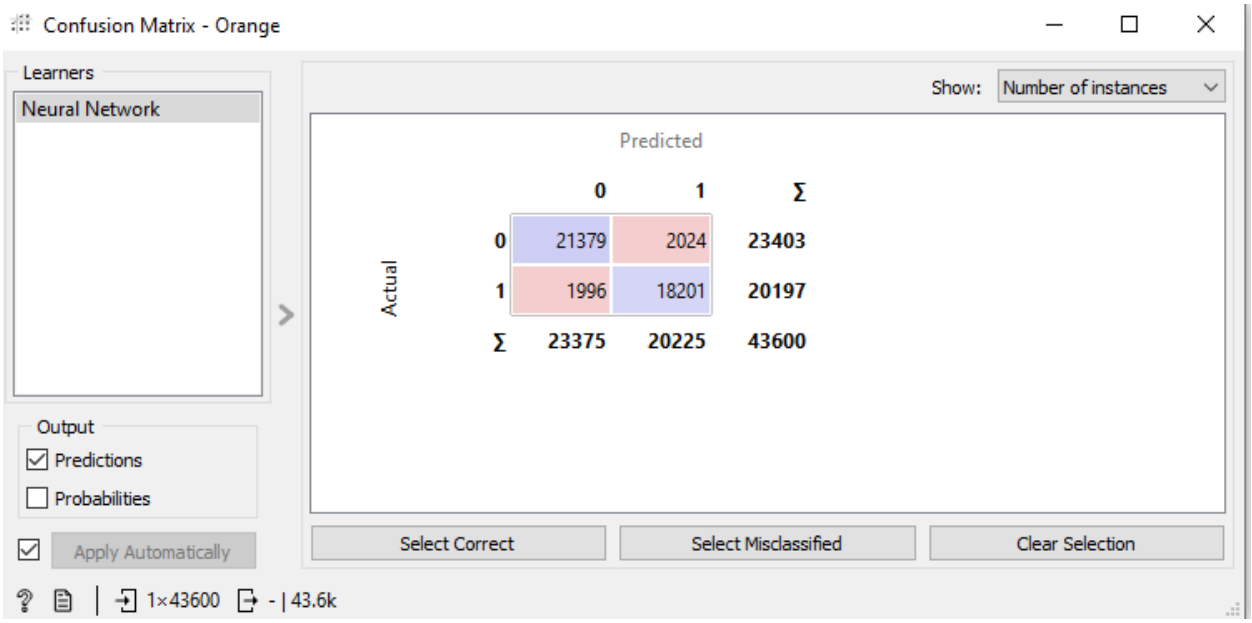
Εικόνα 13) Αποτελέσματα Test and Score για τη νέα προσεγγιστική βάση δεδομένων SMOTE



Εικόνα 14) Αποτελέσματα Confusion Matrix για τη νέα προσεγγιστική βάση δεδομένων SMOTE



Εικόνα 15) Αποτελέσματα Test and Score για τη νέα προσεγγιστική βάση δεδομένων SMOTE-ENN



Εικόνα 16) Αποτελέσματα Confusion Matrix για τη νέα προσεγγιστική βάση δεδομένων SMOTE-ENN

Βιβλιογραφία

[1]Kim, Y. S., & Street, W. N. (2004). An intelligent system for customer targeting: a data mining approach. *Decision Support Systems*, *37*(2), 215–228. Διαθέσιμο στον δικτυακό τόπο: https://doi.org/10.1016/S0167-9236(03)00008-3

[2]Kotler, P. & Keller K. L. (2005), Marketing Management, 12th edition, Pearson Prentice Hall

[3]Chalmeta, R. (2006). Methodology for customer relationship management. *Journal of Systems and Software*, *79*(7), 1015–1024. Διαθέσιμο στον δικτυακό τόπο: https://doi.org/10.1016/J.JSS.2005.10.018

[4] *What is customer satisfaction? Definition + importance*. (n.d.). Διαθέσιμο στον δικτυακό τόπο:https://www.zendesk.com/blog/3-steps-achieving-customer-satisfaction-loyalty/

[5]Riaz, A., Hanif, M., & Hafeez, S. (2010). Factors Affecting Customer Satisfaction. *International Research Journal of Finance and Economics*. Διαθέσιμο στον δικτυακό τόπο: http://www.eurojournals.com/finance.htm

[6]Griffin, J. (n.d.). *Customer Loyalty.* Διαθέσιμο στον δικτυακό τόπο: http://altfeldinc.com/pdfs/Customer%20Loyalty.pdf

[7] Osmanoglu, O. O. (2019). Comparing to Techniques Used in Customer Churn Analysis. Journal of Multidisciplinary Developments, 4(1), 30–38. Διαθέσιμο στον δικτυακό τόπο: https://www.researchgate.net/publication/337103029

[8] *Customer Churn: Definition, Rate, Analysis and Prediction*. (n.d.). Διαθέσιμο στον δικτυακό τόπο: https://www.questionpro.com/blog/customer-churn/

[9] Tsai, C. F., & Lu, Y. H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, *36*(10), 12547–12553. Διαθέσιμο στον δικτυακό τόπο: https://doi.org/10.1016/J.ESWA.2009.05.032

[10] De Caigny, A., Coussement, K., De Bock, K. W., & Lessmann, S. (2020). Incorporating textual information in customer churn prediction models based on a convolutional neural network. *International Journal of Forecasting*, *36*(4), 1563–1578. Διαθέσιμο στον δικτυακό τόπο: https://doi.org/10.1016/J.IJFORECAST.2019.03.029

[11] Sivasankar, E., & Vijaya, J. (2019). Hybrid PPFCM-ANN model: an efficient system for customer churn prediction through probabilistic possibilistic fuzzy clustering and artificial neural network. *Neural Computing and Applications*, *31*(11), 7181–7200. Διαθέσιμο στον δικτυακό τόπο: https://doi.org/10.1007/S00521-018-3548-4/METRICS

[12] Prashanth, R., Deepak, K., & Meher, A. K. (2017). High accuracy predictive modelling for customer churn prediction in telecom industry. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10358 LNAI*, 391–402. Διαθέσιμο στον δικτυακό τόπο: https://doi.org/10.1007/978-3-319-62416-7_28/COVER

[13] Hu, J., Zhuang, Y., Yang, J., Lei, L., Huang, M., Zhu, R., & Dong, S. (2019). PRNN: A Recurrent Neural Network based Approach for Customer Churn Prediction in Telecommunication Sector.

Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018, 4081–4085. Διαθέσιμο στον δικτυακό τόπο: https://doi.org/10.1109/BIGDATA.2018.8622094

[14]Kufel, J., Bargieł-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., Czogalik, Ł., Dudek, P., Magiera, M., Lis, A., Paszkiewicz, I., Nawrat, Z., Cebula, M., & Gruszczyńska, K. (2023). What Is Machine Learning, Artificial Neural Networks and Deep Learning?—Examples of Practical Applications in Medicine. *Diagnostics 2023, Vol. 13, Page 2582*, *13*(15), 2582. Διαθέσιμο στον δικτυακό τόπο: https://doi.org/10.3390/DIAGNOSTICS13152582

[15] Cerny, P. A. (n.d.). *Data mining and Neural Networks from a Commercial Perspective*. Διαθέσιμο στον δικτυακό τόπο: https://www.orsnz.org.nz/conf36/papers/Cerny.pdf

[16]Krogh, A. (2008). What are artificial neural networks? *NATURE BIOTECHNOLOGY*, *26*. Διαθέσιμο στον δικτυακό τόπο: http://www.r-project.org/

[17]Kufel, J., Bargieł-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., Czogalik, Ł., Dudek, P., Magiera, M., Lis, A., Paszkiewicz, I., Nawrat, Z., Cebula, M., & Gruszczyńska, K. (2023). What Is Machine Learning, Artificial Neural Networks and Deep Learning?—Examples of Practical Applications in Medicine. *Diagnostics 2023, Vol. 13, Page 2582*, *13*(15), 2582. Διαθέσιμο στον δικτυακό τόπο: https://doi.org/10.3390/DIAGNOSTICS13152582

[18]Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A. E., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11), e00938. Διαθέσιμο στον δικτυακό τόπο: https://doi.org/10.1016/J.HELIYON.2018.E00938

[19]Singh, D. Y., & Chauhan, A. S. (2005). *NEURAL NETWORKS IN DATA MINING*. Διαθέσιμο στον δικτυακό τόπο: www.jatit.org

[20]Yang, H. (n.d.). *Data Preprocessing-Chapter 3*. Διαθέσιμο στον δικτυακό τόπο: http://cose-stor.sfsu.edu/~huiyang

[21]Muraina, I. O. (n.d.). *IDEAL DATASET SPLITTING RATIOS IN MACHINE LEARNING ALGORITHMS: GENERAL CONCERNS FOR DATA SCIENTISTS AND DATA ANALYSTS*. Διαθέσιμο στον δικτυακό τόπο: https://www.researchgate.net/publication/358284895

[22]Reitermanová, Z. (n.d.). *Data Splitting*. Διαθέσιμο στον δικτυακό τόπο: https://physics.mff.cuni.cz/wds/proc/pdf10/WDS10_105_i1_Reitermanova.pdf

[23]Alibrahim, H., & Ludwig, S. A. (n.d.). *Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization*. Διαθέσιμο στον δικτυακό τόπο: http://www.cs.ndsu.nodak.edu/~siludwig/Publish/papers/CEC2021.pdf

[24]Gaurang, P., & Panchal, D. (2011). Behaviour Analysis of Multilayer Perceptronswith Multiple Hidden Neurons and Hidden Layers. *Article in International Journal of Computer Theory and Engineering*, *3*(2). Διαθέσιμο στον δικτυακό τόπο: https://doi.org/10.7763/IJCTE.2011.V3.328

[25]Sharma, S., Sharma, S., & Athaiya, A. (2020). ACTIVATION FUNCTIONS IN NEURAL NETWORKS. *International Journal of Engineering Applied Sciences and Technology*, *4*, 310–316. Διαθέσιμο στον διακτυακό τόπο: http://www.ijeast.com

[26] *A Comparison of Selected Optimization Methods for Neural Networks OSKAR BONDE LUDVIG KARLSSON KTH ROYAL INSTITUTE OF TECHNOLOGY SCHOOL OF ENGINEERING SCIENCES*. (n.d.).

[27]Gong, M. (2021). A NOVEL PERFORMANCE MEASURE FOR MACHINE LEARNING CLASSIFICATION. *International Journal of Managing Information Technology (IJMIT)*, *13*(1). Διαθέσιμο στον δικτυακό τόπο: https://doi.org/10.5121/ijmit.2021.13101

[28]Sietsma, J., & Dow, R. J. F. (1991). Creating Artificial Neural Networks That Generalize. *Neural ,¥etworks*, *4*, 67–79.

[29]Geman, S., Bienenstock, E., & Doursat, R. (1992). *VIEW~~~~~~~~~~~~~~~~ Communicated by Lawrence Jackel Neural Networks and the Bias/Variance Dilemma*.

[30]Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(32), 15849–15854. Διαθέσιμο στον δικτυακό τόπο: https://doi.org/10.1073/PNAS.1903070116/SUPPL_FILE/PNAS.1903070116.SAPP.PDF

[31]Khalaf Jabbar Rafiqul Zaman Khan, H. D. (2015). *METHODS TO AVOID OVER-FITTING AND UNDER-FITTING IN SUPERVISED MACHINE LEARNING (COMPARATIVE STUDY)*.

[32]*Marketing Promotion Campaign Uplift Modelling*. (n.d.). Διαθέσιμο στον δικτυακό τόπο: https://www.kaggle.com/datasets/davinwijaya/customer-retention