ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΙΑΤΡΙΚΗ

«Μοντελοποίηση και Αξιολόγηση Εκτίμησης Κινδύνων για τη Λήψη Μακροπρόθεσμων Οικονομικών Αποφάσεων»

ΕΥΣΤΑΘΙΑΔΗΣ Θ. ΠΑΝΑΓΙΩΤΗΣ
ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ
Κωνσταντίνος Κολομβάτσος
Επίκουρος Καθηγητής

Λαμία 2022

«Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις που προβλέπονται από τις διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί χωρίς να τα περικλείω σε εισαγωγικά και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.

2. Δέχομαι ότι η αυτολεξεί παράθεση χωρίς εισαγωγικά, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.

3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια.

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ο Δηλών

Περίληψη

Η διαχείριση του κινδύνου και πιο συγκεκριμένα του πιστωτικού κινδύνου αποτελεί ένα εξαιρετικά απαιτητικό και σημαντικό πεδίο στον τομέα των χρηματοοικονομικών. Με την βοήθεια της πληροφορικής έχουν εξελιχθεί οι μέθοδοι και οι τεχνικές που χρησιμοποιούνται σε σχέση με τις πρώτες προσπάθειες που είχαν πραγματοποιηθεί τον τελευταίο αιώνα. Συγκεκριμένα, νέοι αλγόριθμοι έχουν προστεθεί στην φαρέτρα των επιστημόνων που δραστηριοποιούνται στον χώρο των οικονομικών. Μεγάλο μέρος σε όλο αυτό έχει διαδραματίσει η εξέλιξη των υπολογιστικών συστημάτων. Οι μεγάλες βάσεις δεδομένων και η τεχνητή νοημοσύνη που συνεχίζουν να εξελίσσονται συντελούν σε συνεχώς πιο επιτυχή και ακριβέστερα αποτελέσματα. Σήμερα λόγω των συνεχόμενων οικονομικών κρίσεων που έχουν πλήξει την Ευρώπη και τον κόσμο γενικότερα, με ότι αυτό συνεπάγεται για τις οργανωμένες καπιταλιστικές κοινωνίες και τους πολίτες τους, είναι εξαιρετικά σημαντικό να εκτιμηθεί ο κίνδυνος με όσο το δυνατόν καλύτερη ακρίβεια. Στην συγκεκριμένη διατριβή πραγματοποιήθηκαν μελέτες σχετικά με κάποιες από τις πιο σημαντικές τεχνικές διαχείρισης κινδύνου των τραπεζικών ιδρυμάτων και ταυτόχρονα αναλύθηκαν βασικές θεωρίες των χρηματοοικονομικών επιστημών και της στατιστικής.

Περιεχόμενα

1 Εισαγωγή

Η συνεχιζόμενη και αξιοσημείωτη ανάπτυξη του τομέα της εξόρυξης δεδομένων και της ανακάλυψης γνώσης τροφοδοτήθηκε από τη συμβολή διαφόρων παραγόντων: πρώτο, η εκρηκτική αύξηση της συλλογής δεδομένων, δεύτερο η αποθήκευση των δεδομένων σε αποθήκες δεδομένων έτσι ώστε να έχει ολόκληρη η επιχείρηση πρόσβαση σε μια αξιόπιστη τρέχουσα βάση δεδομένων, τρίτο η διαθεσιμότητα αυξημένης πρόσβασης σε δεδομένα από την πλοήγηση στο Web και τα ενδοδίκτυα (intranets), τέταρτο, η ανταγωνιστική πίεση για αύξηση του μεριδίου αγοράς σε μια παγκοσμιοποιημένη οικονομία, πέμπτο, η ανάπτυξη τυποποιημένων συνθέσεων λογισμικού εξόρυξης εμπορικών δεδομένων και, τέλος, η τεράστια αύξηση της υπολογιστικής ισχύος και της χωρητικότητας αποθήκευσης [1].

Ειδικότερα –και κυριότερα–, η εξόρυξη δεδομένων μπορεί να θεωρηθεί ως αποτέλεσμα της φυσικής εξέλιξης της πληροφορικής. Όπως επισημαίνει ο Jiawei Han [3], η εξελικτική αυτή πορεία έχει παρατηρηθεί στη βιομηχανία βάσεων δεδομένων μέσω της προόδου των ακόλουθων λειτουργιών: συλλογή δεδομένων και δημιουργία βάσεων δεδομένων, διαχείριση δεδομένων (συμπεριλαμβανομένης της αποθήκευσης και ανάκτησης δεδομένων και διαδικασία συναλλαγής βάσεων δεδομένων), και ανάλυση και κατανόηση δεδομένων (που περιλαμβάνει αποθήκευση δεδομένων και εξόρυξη δεδομένων). Από τη δεκαετία του 1960, η βάση δεδομένων και η τεχνολογία των πληροφοριών εξελίσσεται συστηματικά από πρωτόγονα συστήματα επεξεργασίας σε εξελιγμένα και ισχυρά συστήματα βάσεων δεδομένων.

Η έρευνα και ανάπτυξη σε συστήματα βάσεων δεδομένων από τη δεκαετία του 1970 οδήγησε στην ανάπτυξη συστημάτων σχεσιακών βάσεων δεδομένων, εργαλεία μοντελοποίησης δεδομένων και τεχνικές ευρετηρίασης και οργάνωσης δεδομένων. Επιπλέον, οι χρήστες απέκτησαν βολική και ευέλικτη πρόσβαση δεδομένων μέσω γλωσσών ερωτημάτων, επεξεργασίας ερωτημάτων και διεπαφών χρήστη. Τέλος, αποτελεσματικές μέθοδοι για on-line επεξεργασία συναλλαγών (OLTP), όπου ένα ερώτημα θεωρείται ως συναλλαγή μόνο για ανάγνωση συνέβαλε ουσιαστικά στην εξέλιξη και την ευρεία αποδοχή της σχεσιακής τεχνολογίας ως σημαντικό εργαλείο για την αποτελεσματική αποθήκευση, ανάκτηση και διαχείριση μεγάλων ποσοτήτων δεδομένων.

Η τεχνολογία βάσεων δεδομένων από τα μέσα της δεκαετίας του 1980 χαρακτηρίζεται από τη δημοφιλή υιοθέτηση σχεσιακής τεχνολογίας και μια αύξηση των δραστηριοτήτων έρευνας και ανάπτυξης σε νέα και ισχυρά συστήματα βάσεων δεδομένων. Γενικότερα, η σταθερή και εκπληκτική πρόοδος της τεχνολογίας υλικού υπολογιστών τις τελευταίες τρεις δεκαετίες οδήγησε σε ισχυρές, προσιτές, και μεγάλες προμήθειες υπολογιστών, εξοπλισμούς συλλογής δεδομένων και μέσων αποθήκευσης. Αυτή η τεχνολογία παρέχει μια μεγάλη ώθηση στη βιομηχανία βάσεων δεδομένων και πληροφοριών και δημιουργεί τεράστιο αριθμό βάσεων δεδομένων και αποθετηρίων πληροφοριών διαθέσιμο για διαχείριση συναλλαγών, ανάκτηση πληροφοριών και ανάλυση δεδομένων. Τα δεδομένα μπορούν τώρα να αποθηκευτούν σε πολλούς διαφορετικούς τύπους βάσεων δεδομένων. Μια αρχιτεκτονική βάσης δεδομένων που εμφανίστηκε πρόσφατα είναι η αποθήκη δεδομένων, ένα αποθετήριο πολλαπλών ετερογενών πηγών δεδομένων, οργανωμένο υπό ένα ενιαίο σχήμα σε έναν ιστότοπο προκειμένου να διευκολυνθεί η λήψη αποφάσεων από τη διαχείριση.

Η αφθονία των δεδομένων, σε συνδυασμό με την ανάγκη για ισχυρά εργαλεία ανάλυσης δεδομένων, έχει περιγραφεί ως μία κατάσταση «πλούσια σε δεδομένα αλλά φτωχή σε πληροφορία». Η ταχέως αναπτυσσόμενη, τεράστια ποσότητα δεδομένων, που συλλέγεται και αποθηκεύεται σε μεγάλες και πολλές βάσεις δεδομένων, έχει ξεπεράσει κατά πολύ την ανθρώπινη ικανότητα κατανόησης χωρίς ισχυρά εργαλεία. Ως αποτέλεσμα, τα δεδομένα που συλλέγονται σε μεγάλες βάσεις δεδομένων γίνονται «τάφοι δεδομένων» –αρχεία δεδομένων που σπάνια επανεξετάζονται. Κατά συνέπεια, σημαντικές αποφάσεις συχνά δεν λαμβάνονται με βάση τα πλούσια σε πληροφορίες δεδομένα που αποθηκεύονται σε βάσεις δεδομένων, αλλά μάλλον βάσει της διαίσθησης του υπεύθυνου λήψης αποφάσεων, απλώς και μόνο επειδή ο υπεύθυνος λήψης αποφάσεων δεν διαθέτει τα εργαλεία για την εξαγωγή των πολύτιμων γνώσεων που βρίσκονται ενσωματωμένες στις τεράστιες ποσότητες δεδομένων. Συνεπώς, το συνεχώς διευρυνόμενο χάσμα μεταξύ δεδομένου και πληροφορίας απαιτεί μία συστηματική ανάπτυξη εργαλείων εξόρυξης δεδομένων που θα μετατρέψουν τους τάφους δεδομένων σε «χρυσά ψήγματα» γνώσης [3].

2 Data mining (εξόρυξη δεδομένων)

2.1 Εισαγωγή

2.1.1 Εξόρυξη δεδομένων: ορισμός

Σύμφωνα με τον Larose (2004) [1], «η εξόρυξη δεδομένων είναι η διαδικασία της ανακάλυψης σημαντικούς νέους συσχετισμούς, μοτίβα και τάσεις, κοσκινίζοντας μεγάλες ποσότητες δεδομένα που αποθηκεύονται σε αποθετήρια, χρησιμοποιώντας τεχνολογίες αναγνώρισης προτύπων καθώς και στατιστικά και μαθηματικές τεχνικές». Σύμφωνα με τον ίδιο υπάρχουν και άλλοι ορισμοί:

> «Η εξόρυξη δεδομένων είναι ένα διεπιστημονικό πεδίο που φέρνει μαζί τεχνικές από μηχανική μάθηση, αναγνώριση προτύπων, στατιστικές, βάσεις δεδομένων και οπτικοποίηση του ζητήματος της εξαγωγής πληροφοριών από μεγάλες βάσεις δεδομένων».

Επίσης,

> «Η εξόρυξη δεδομένων είναι η ανάλυση (συχνά μεγάλων) συνόλων δεδομένων παρατήρησης για εύρεση συσχετίσεων και για συνόψιση των δεδομένων με νέους τρόπους που είναι κατανοητοί όσο και χρήσιμοι για τον κάτοχο των δεδομένων αυτών».

Σύμφωνα με τους Hand κ.ά. (2001) [2], ο προαναφερθείς ορισμός αναφέρεται σε «δεδομένα παρατήρησης», σε αντίθεση με τα «πειραματικά δεδομένα». Η εξόρυξη δεδομένων ασχολείται συνήθως με δεδομένα που έχουν ήδη συλλεχθεί για κάποιο σκοπό εκτός από την ανάλυση εξόρυξης δεδομένων (για παράδειγμα, ενδέχεται να έχουν συλλεχθεί κατά σειρά για να διατηρηθεί ένα ενημερωμένο αρχείο όλων των συναλλαγών σε μια τράπεζα). Αυτό σημαίνει ότι οι στόχοι της άσκησης εξόρυξης δεδομένων δεν παίζουν ρόλο στη στρατηγική συλλογής δεδομένων. Αυτό είναι ένας τρόπος με τον οποίο η εξόρυξη δεδομένων διαφέρει από πολλά στατιστικά στοιχεία, όπου τα δεδομένα συχνά συλλέγονται χρησιμοποιώντας αποτελεσματικές

στρατηγικές για την απάντηση συγκεκριμένων ερωτήσεων. Για αυτόν τον λόγο, η εξόρυξη δεδομένων αναφέρεται συχνά ως «δευτερεύουσα» ανάλυση δεδομένων.

Ο ορισμός αναφέρει επίσης ότι τα σύνολα δεδομένων που εξετάζονται στην εξόρυξη δεδομένων είναι συχνά μεγάλα. Αν συμμετείχαν μόνο μικρά σύνολα δεδομένων, θα συζητούσαμε απλώς την κλασική διερευνητική ανάλυση δεδομένων όπως ασκείται από στατιστικολόγους. Όταν είμαστε αντιμέτωποι με μεγάλα πακέτα δεδομένων, προκύπτουν νέα προβλήματα. Μερικά από αυτά σχετίζονται με θέματα τακτοποίησης σχετικά με τον τρόπο αποθήκευσης ή πρόσβασης στα δεδομένα, ενώ άλλα σχετίζονται με πιο θεμελιώδη ζητήματα, όπως τον τρόπο προσδιορισμού της αντιπροσωπευτικότητας των δεδομένων, τον τρόπο ανάλυσης των δεδομένων σε εύλογο χρονικό διάστημα χρόνου, και πώς να αποφασιστεί εάν μια φαινομενική σχέση είναι απλώς μια πιθανή εμφάνιση που δεν αντικατοπτρίζει καμία υποκείμενη πραγματικότητα.

Συχνά τα διαθέσιμα δεδομένα περιλαμβάνουν μόνο ένα δείγμα από τον πλήρη πληθυσμό (ή, ίσως, από έναν υποθετικό υπερπληθυσμό)· ο στόχος μπορεί να είναι η γενίκευση από το δείγμα στον πληθυσμό. Για παράδειγμα, μπορεί να θέλουμε να προβλέψουμε πώς οι μελλοντικοί πελάτες είναι πιθανό να συμπεριφέρονται ή να προσδιορίσουμε τις ιδιότητες των πρωτεϊνικές δομών που δεν έχουμε ακόμη δει. Τέτοιες γενικεύσεις μπορεί να μην έχουν επιτευχθεί μέσω τυπικών στατιστικών προσεγγίσεων επειδή συχνά τα δεδομένα δεν είναι (κλασική στατιστική) «τυχαία δείγματα», αλλά «εύκολα» ή «ευκαιριακά» δείγματα. Μερικές φορές μπορεί να θέλουμε να συνοψίσουμε ή να συμπιέσουμε ένα πολύ μεγάλο σύνολο δεδομένων με τέτοιο τρόπο ώστε το αποτέλεσμα να είναι πιο κατανοητό, χωρίς καμία έννοια γενίκευσης. Αυτό το ζήτημα θα προέκυπτε, για παράδειγμα, εάν είχαμε πλήρη στοιχεία απογραφής για μία συγκεκριμένη χώρα ή μια βάση δεδομένων που καταγράφει εκατομμύρια μεμονωμένες λιανικές συναλλαγές.

Οι σχέσεις και οι δομές που βρίσκονται σε ένα σύνολο δεδομένων πρέπει, φυσικά, να είναι καινοφανείς. Δεν έχει νόημα να επαναληφθούν οι καθιερωμένες σχέσεις (εκτός εάν η άσκηση στοχεύει στην επιβεβαίωση «υπόθεσης», στην οποία κάποιος προσπαθούσε να προσδιορίσει εάν το καθιερωμένο πρότυπο υπάρχει επίσης σε ένα νέο σύνολο δεδομένων) ή στις απαραίτητες σχέσεις (ότι, για παράδειγμα, όλοι οι έγκυοι ασθενείς είναι γυναίκες). Είναι σαφές ότι η καινοτομία πρέπει να μετρηθεί σε σχέση με τις προηγούμενες γνώσεις του χρήστη.

Σε μια βαθύτερη ανάλυση του ορισμού, οι Hand κ.ά. (2001) [2] υποστηρίζουν ότι ενώ η καινοτομία είναι μια σημαντική ιδιότητα των σχέσεων που επιδιώκουμε, δεν αρκεί για να χαρακτηρίζει χρήσιμες αναζητήσεις σχέσεων. Ειδικότερα, οι σχέσεις πρέπει επίσης να είναι κατανοητές. Για παράδειγμα, οι απλές σχέσεις είναι πιο εύκολα κατανοητές από τις περίπλοκες και, ίσως, προτιμώνται ceteris paribus. Η εξόρυξη δεδομένων ρυθμίζεται συχνά στο ευρύτερο πλαίσιο της «Ανακάλυψης Γνώσεων σε Βάσεις Δεδομένων» ή, αλλιώς, «KDD» (Knowledge Discovery in Databases). Αυτός ο όρος προήλθε στον τομέα της τεχνητής νοημοσύνης (AI).

2.1.2 Εξόρυξη δεδομένων: τι δεν μπορεί να επιτελέσει

Η εξόρυξη δεδομένων είναι ένα εργαλείο. Σύμφωνα με τον Edelstein (1999) [4], δεν θα εγκατασταθεί στη βάση δεδομένων παρακολουθώντας τι συμβαίνει και θα στείλει e-mail όταν βλέπει ένα ενδιαφέρον μοτίβο. Δεν εξαλείφει την ανάγκη να γνωρίζει κάποιος την επιχείρησή του, να κατανοεί τα δεδομένα ή να κατανοεί αναλυτικές μεθόδους. Η εξόρυξη δεδομένων βοηθά τους επιχειρηματικούς αναλυτές να βρουν μοτίβα και σχέσεις στα δεδομένα. Επιπλέον, τα μοτίβα που αποκαλύπτονται από την εξόρυξη δεδομένων πρέπει να επαληθευτούν στον πραγματικό κόσμο. Εν προκειμένω, οι προγνωστικές σχέσεις που εντοπίζονται μέσω της εξόρυξης δεδομένων δεν είναι απαραίτητα αιτίες μιας δράσης ή συμπεριφοράς.

Για να διασφαλίσετε ουσιαστικά αποτελέσματα, είναι σημαντικό να κατανοηθούν τα δεδομένα. Η ποιότητα του αποτελέσματός θα είναι συχνά επιρρεπής σε ακραίες τιμές (τιμές δεδομένων που είναι πολύ διαφορετικές από τις τυπικές τιμές στη βάση δεδομένων), άσχετες στήλες ή στήλες που διαφέρουν μεταξύ τους (όπως η ηλικία και η ημερομηνία γέννησης), στον τρόπο με τον οποίο κωδικοποιούνται τα δεδομένα, τα δεδομένα που επιτρέπονται και τα δεδομένα που εξαιρούνται. Από την άλλη, οι αλγόριθμοι ποικίλλουν ως προς την ευαισθησία τους σε τέτοια ζητήματα δεδομένων. Έτσι, είναι επισφαλές να στηρίζεται η ανθρώπινη σκέψη σε εργαλεία εξόρυξης δεδομένων για την λήψη ορθών αποφάσεων.

Ως εκ τούτου, η εξόρυξη δεδομένων δεν θα ανακαλύψει αυτόματα λύσεις χωρίς καθοδήγηση. Αν και ένα καλό εργαλείο εξόρυξης δεδομένων προστατεύει από τις περιπλοκές των στατιστικών τεχνικών, απαιτεί να κατανοηθεί η λειτουργία των εργαλείων που κάθε φορά επιλέγονται και τους αλγόριθμους στους οποίους βασίζονται. Οι επιλογές που τελούνται κατά τη ρύθμιση του εργαλείου εξόρυξης

δεδομένων και οι βελτιστοποιήσεις που θα επιλέγονται ως βέλτιστες θα επηρεάσουν την ακρίβεια και την ταχύτητα των παραγόμενων μοντέλων. Η εξόρυξη δεδομένων δεν αντικαθιστά ειδικευμένους επιχειρηματικούς αναλυτές ή διαχειριστές, αλλά τους δίνει ένα ισχυρό νέο εργαλείο για τη βελτίωση της εργασίας που κάνουν. Κάθε εταιρεία που γνωρίζει την επιχείρησή της και τους πελάτες της έχει ήδη επίγνωση πολλών σημαντικών, υψηλών αποδόσεων μοντέλων. Αυτό που μπορεί να κάνει η εξόρυξη δεδομένων είναι να επιβεβαιώσει τέτοιες εμπειρικές παρατηρήσεις και να βρει νέα, ανεπαίσθητα μοτίβα που αποφέρουν σταθερή σταδιακή βελτίωση (συν την περιστασιακή σημαντική ανακάλυψη).

2.1.3 Εξόρυξη δεδομένων: σύνολα δεδομένων (datasets)

Εκκινώντας, οφείλουμε να αναλύσουμε τη βασική φύση των συνόλων δεδομένων. Ένα σύνολο δεδομένων είναι ένα σύνολο μετρήσεων που λαμβάνεται από κάποιο περιβάλλον ή διαδικασία. Στην απλούστερη περίπτωση, έχουμε μια συλλογή αντικειμένων και για κάθε αντικείμενο έχουμε ένα σύνολο ίδιων μετρήσεων p. Σε αυτή την περίπτωση, μπορούμε να σκεφτούμε τη συλλογή των μετρήσεων σε αντικείμενα n ως τη μορφή του πίνακα δεδομένων (data matrix) n × p. Οι σειρές n αντιπροσωπεύουν τα αντικείμενα n στα οποία ελήφθησαν μετρήσεις. Τέτοιες σειρές μπορεί να αναφέρονται ως άτομα, οντότητες, περιπτώσεις, αντικείμενα ή αρχεία ανάλογα με τα συμφραζόμενα.

Η άλλη διάσταση του πίνακα δεδομένων περιέχει το σύνολο των μετρήσεων p που πραγματοποιήθηκαν σε κάθε αντικείμενο. Συνήθως, υποθέτουμε ότι οι ίδιες μετρήσεις p γίνονται σε κάθε ένα άτομο αν και αυτό δεν χρειάζεται να συμβαίνει (για παράδειγμα, διαφορετικές ιατρικές εξετάσεις θα μπορούσαν να πραγματοποιούνται σε διαφορετικούς ασθενείς). Παράλληλα, οι στήλες p του πίνακα δεδομένων μπορούν να αναφερθούν ως μεταβλητές, χαρακτηριστικά, στάσεις ή πεδία· πάλι, η γλώσσα εξαρτάται από τα συμφραζόμενα της έρευνας. Σε όλες τις περιπτώσεις η ιδέα είναι η ίδια: αυτά τα ονόματα αναφέρονται στη μέτρηση που αντιπροσωπεύεται από κάθε στήλη.

Τα δεδομένα εμφανίζονται με πολλές μορφές και αυτό είναι ακατάλληλο για την ανάπτυξη μιας πλήρους ταξινόμησης [2]. Πράγματι, δεν είναι καν σαφές ότι μπορεί να αναπτυχθεί μια πλήρης ταξινόμηση, δεδομένου ότι μια σημαντική πτυχή των δεδομένων σε μια κατάσταση μπορεί να είναι ασήμαντη σε άλλη. Ωστόσο, εκεί

βρίσκονται ορισμένες βασικές διακρίσεις τις οποίες πρέπει να αναλύσουμε. Έτσι, μία αφορά τη διαφορά μεταξύ ποσοτικών (quantitative) και κατηγορικών (categorical) μετρήσεων. Μια ποσοτική μεταβλητή μετράται σε αριθμητική κλίμακα και μπορεί, καταρχάς, να πάρει οποιαδήποτε αξία (π.χ. ηλικία, έσοδα κ.λπ.). Αντίθετα, κατηγορηματικές μεταβλητές μπορούν να πάρουν μόνο συγκεκριμένες διακριτές τιμές (π.χ. φύλο, έγγαμος βίος κ.λπ.). Επιπρόσθετα, οι κατηγορηματικές μεταβλητές μπορεί να είναι κανονικές (με φυσική σειρά, όπως στην εκπαιδευτική κλίμακα) ή ονομαστικές (απλώς ονομάζοντας τις κατηγορίες, όπως στην περίπτωση της οικογενειακής κατάστασης).

Οι κλίμακες μέτρησης, όπως αυτές καθορίζονται, βρίσκονται στο κάτω μέρος κάθε ταξινόμησης δεδομένων. Κινούμενοι ανοδικά στην ταξινόμηση, διαπιστώνουμε ότι τα δεδομένα μπορούν να προκύψουν σε διάφορες σχέσεις και δομές. Τα δεδομένα ενδέχεται να προκύψουν διαδοχικά σε χρονοσειρές (time series) και η άσκηση της εξόρυξης δεδομένων μπορεί να αντιμετωπίσει ολόκληρες χρονοσειρές ή συγκεκριμένα τμήματα αυτών των χρονοσειρών. Τα δεδομένα μπορεί επίσης να περιγράψουν χωρικές σχέσεις, έτσι ώστε τα μεμονωμένα αρχεία να αποκτούν την πλήρη σημασία τους μόνο όταν θεωρούνται στο πλαίσιο των άλλων. Χρήσιμο παράδειγμα εδώ αποτελεί ένα σύνολο δεδομένων για ιατρικούς ασθενείς όπου περιλαμβάνονται πολλές μετρήσεις για την ίδια μεταβλητή (αρτηριακή πίεση), διαφορετικοί χρόνοι πραγματοποίησης μετρήσεων (διαφορετικές ημέρες), διαφορετικές μορφές δεδομένων (εικόνες, κείμενα), διαφορετική ιεραρχία μεταξύ των ασθενών (θεράποντες ιατροί, νοσοκομεία, γεωγραφικές τοποθεσίες). Όσο πιο περίπλοκες είναι οι δομές δεδομένων, τόσο πιο περίπλοκα είναι τα μοντέλα εξόρυξης δεδομένων, οι αλγόριθμοι και εργαλεία που πρέπει να εφαρμοστούν. Δοθέντων τούτων, ο πίνακας δεδομένων n × p είναι συχνά μια υπεραπλούστευση ή εξιδανίκευση αυτού που εμφανίζεται στην πράξη [2].

2.2 Τεχνικές εξόρυξης δεδομένων

Η εξόρυξη δεδομένων μπορεί να επιτελέσει πληθώρα εργασιών. Οι κυριότερες από αυτές είναι: περιγραφή, εκτίμηση, πρόβλεψη, συσχέτιση, συσταδοποίηση, ταξινόμηση.

2.2.1 Περιγραφή (description)

Μερικές φορές, ερευνητές και αναλυτές προσπαθούν απλώς να βρουν τρόπους για να περιγράψουν μοτίβα και τάσεις που βρίσκονται στα δεδομένα. Για παράδειγμα, μια δημοσκόπηση μπορεί να αποκαλύψει στοιχεία ότι όσοι έχουν απολυθεί είναι λιγότερο πιθανό να υποστηρίξουν την τρέχουσα κυβέρνηση στις προεδρικές εκλογές [1]. Οι περιγραφές των προτύπων και των τάσεων συχνά προτείνουν πιθανές εξηγήσεις για τέτοια μοτίβα και τάσεις. Τα μοντέλα εξόρυξης δεδομένων πρέπει να είναι όσο το δυνατόν πιο διαφανή (transparent). Δηλαδή, τα αποτελέσματα του μοντέλου εξόρυξης δεδομένων θα πρέπει να περιγράφουν σαφή πρότυπα που είναι επιδεκτικά διαισθητικής ερμηνείας και εξήγησης.

Ορισμένες μέθοδοι εξόρυξης δεδομένων είναι πιο κατάλληλες από άλλες όσον αφορά τη διαφανή ερμηνεία. Για παράδειγμα, τα δέντρα αποφάσεων παρέχουν μία διαισθητική και φιλική προς τον άνθρωπο εξήγηση των αποτελεσμάτων τους. Από την άλλη πλευρά, τα νευρωνικά δίκτυα είναι συγκριτικά πιο αδιαφανή για τους μη ειδικούς, λόγω της μη γραμμικότητας και της πολυπλοκότητας του μοντέλου. Η περιγραφή υψηλής ποιότητας μπορεί συχνά να επιτευχθεί με διερευνητική (exploratory) ανάλυση δεδομένων, μια γραφική μέθοδος διερεύνησης δεδομένων σε αναζήτηση προτύπων και τάσεων.

2.2.2 Εκτίμηση (estimation)

Σύμφωνα με τον Larose (2004), η εκτίμηση είναι παρόμοια με την ταξινόμηση, εκτός από το ότι η υπό ανάλυση μεταβλητή είναι μάλλον αριθμητική παρά κατηγορική (categorical). Τα μοντέλα κατασκευάζονται χρησιμοποιώντας "πλήρεις" εγγραφές, οι οποίες παρέχουν την αξία της μεταβλητής καθώς και των προβλέψεων. Στη συνέχεια, για νέες παρατηρήσεις, εκτιμήσεις της τιμής της μεταβλητής στόχου γίνονται με βάση τις τιμές των προβλέψεων. Για παράδειγμα, μπορεί να μας ενδιαφέρει η εκτίμηση της συστολικής αρτηριακής πίεσης ασθενή νοσοκομείου, με βάση την ηλικία, το φύλο, τον δείκτη μάζας σώματος και τα επίπεδα νατρίου στο αίμα του ασθενούς. Η σχέση μεταξύ συστολικής αρτηριακής πίεσης και των προγνωστικών μεταβλητών θα μας παρείχαν ένα μοντέλο εκτίμησης. Κατ' επέκταση, μπορούμε να εφαρμόσουμε αυτό το μοντέλο σε νέες περιπτώσεις (π.χ.

εκτίμηση των σχολικών εξόδων μιας οικογένειας, εκτίμηση του μέσου όρου ενός φοιτητή κ.ά.).

2.2.3 Πρόβλεψη (prediction)

Η πρόβλεψη είναι παρόμοια με την ταξινόμηση και την εκτίμηση εκτός από το γεγονός ότι τα αποτελέσματα της πρόβλεψης βρίσκονται στο μέλλον. Παραδείγματα εργασιών πρόβλεψης στις επιχειρήσεις και την έρευνα περιλαμβάνουν π.χ. την πρόβλεψη της τιμής ενός αποθέματος τρεις μήνες στο μέλλον, την πρόβλεψη της ποσοστιαίας αύξησης των αυτοκινητιστικών ατυχημάτων το επόμενο έτος, εάν το όριο ταχύτητας αυξηθεί κ.ά. (Larose, 2004). Παράλληλα, οποιεσδήποτε από τις μεθόδους και τις τεχνικές που χρησιμοποιούνται για την ταξινόμηση και την εκτίμηση μπορούν να χρησιμοποιηθούν επίσης, υπό κατάλληλες συνθήκες, για πρόβλεψη. Αυτές περιλαμβάνουν τις παραδοσιακές στατιστικές μεθόδους εκτιμήσεων σημείων (point estimation) και εκτιμήσεων διαστήματος εμπιστοσύνης (confidence interval estimations), απλή γραμμική αναδρομή και συσχέτιση κ.λπ..

2.2.4 Συσχέτιση (association)

Η εργασία της συσχέτισης όσον αφορά την εξόρυξη δεδομένων σχετίζεται με το να βρεθούν ποια χαρακτηριστικά πηγαίνουν μαζί. Πιο διαδεδομένη στον επιχειρηματικό κόσμο, όπου είναι γνωστή ως ανάλυση συνάφειας (affinity analysis) ή ανάλυση καλαθιού αγοράς (market basket analysis), επιδιώκει να αποκαλύψει κανόνες για την ποσοτικοποίηση της σχέσης μεταξύ δύο ή περισσότερων χαρακτηριστικών. Οι κανόνες σύνδεσης έχουν τη μορφή «εάν προηγούμενο, τότε συνεπές» (if antecedent, then consequent) μαζί με την μέτρηση της υποστήριξης και της εμπιστοσύνης που σχετίζονται με τον κανόνα. Για παράδειγμα, ένα συγκεκριμένο σούπερ μάρκετ μπορεί να το βρει ότι από τους 1000 πελάτες που ψωνίζουν την Πέμπτη το βράδυ, 200 αγόρασαν πάνες και από αυτούς τους 200 που αγόρασαν πάνες, 50 αγόρασαν μπύρα. Έτσι, ο κανόνας σύνδεσης θα ήταν «εάν αγοράζετε πάνες, στη συνέχεια αγοράστε μπύρα» με υποστήριξη 200/1000 = 20% και εμπιστοσύνη 50/200 = 25% (Larose, 2004). Περαιτέρω, άλλα παραδείγματα συσχέτισης στις επιχειρήσεις και την έρευνα περιλαμβάνουν τη διερεύνηση του ποσοστού των συνδρομητών στο πρόγραμμα κινητής τηλεφωνίας μιας εταιρείας, την

εξέταση του ποσοστού των παιδιών των οποίων οι γονείς τους διαβάζουν και είναι και οι ίδιοι καλοί αναγνώστες, τον προσδιορισμό του ποσοστού των περιπτώσεων στις οποίες ένα νέο φάρμακο θα παρουσιάσει επικίνδυνες παρενέργειες κ.λπ..

2.2.5 Συσταδοποίηση (clustering)

Σύμφωνα με τον Edelstein (1999), το Clustering χωρίζει μια βάση δεδομένων σε διαφορετικές ομάδες. Εν προκειμένω, ο στόχος της συσταδοποίησης είναι να βρει ομάδες που είναι πολύ διαφορετικές μεταξύ τους, και των οποίων τα μέλη είναι πολύ παρόμοια μεταξύ τους. Σε αντίθεση με την ταξινόμηση, δεν γνωρίζουμε ποια θα είναι τα συμπλέγματα όταν εκκινεί η διαδικασία ή με ποια χαρακτηριστικά θα συγκεντρωθούν τα δεδομένα. Κατά συνέπεια, κάποιος που είναι πεπειραμένος στην επιχείρηση πρέπει να ερμηνεύσει τις συστάδες. Συχνά είναι απαραίτητο να τροποποιηθεί η συσταδοποίηση αποκλείοντας μεταβλητές που έχουν χρησιμοποιηθεί για την ομαδοποίηση συμβάντων, επειδή κατά την εξέταση ο χρήστης τις αναγνωρίζει ως άσχετες ή χωρίς νόημα. Αφού εντοπιστούν τα συμπλέγματα (clusters) που τμηματοποιούν εύλογα τη βάση δεδομένων, αυτά τα συμπλέγματα μπορούν στη συνέχεια να χρησιμοποιηθούν για την ταξινόμηση νέων δεδομένων. Μερικοί από τους κοινούς αλγόριθμους που χρησιμοποιούνται για τη συσταδοποίηση περιλαμβάνουν χάρτες χαρακτηριστικών Kohonen (Kohonen feature maps) και K-μέσα (K-means).

Αξίζει να τονιστεί ότι δεν πρέπει να συγχέεται η συσταδοποίηση με την τμηματοποίηση (segmentation). Η τμηματοποίηση αναφέρεται στο γενικό πρόβλημα του προσδιορισμού ομάδων που έχουν κοινά χαρακτηριστικά. Η συσταδοποίηση είναι ένας τρόπος τμηματοποίησης δεδομένων σε ομάδες που δεν έχουν οριστεί προηγουμένως, ενώ η ταξινόμηση, από την άλλη, είναι ένας τρόπος τμηματοποίησης δεδομένων, εκχωρώντας τα σε ομάδες που έχουν ήδη καθοριστεί. Έτσι, συνοψίζοντας, η εργασία της συσταδοποίησης δεν προσπαθεί να ταξινομήσει, να εκτιμήσει ή να προβλέψει την τιμή μιας μεταβλητής. Αντ' αυτού, οι αλγόριθμοι της συσταδοποίησης επιχειρούν να τμηματοποιήσουν ολόκληρο το σύνολο δεδομένων σε σχετικά ομοιογενείς υποομάδες ή ομάδες, όπου η ομοιότητα των εγγραφών εντός του συμπλέγματος μεγιστοποιείται και η ομοιότητα στις εγγραφές εκτός του συμπλέγματος ελαχιστοποιείται (Larose, 2004). Τέλος, παραδείγματα συσταδοποίησης εργασιών συστήνουν ο λογιστικός έλεγχος με σκοπό την

τμηματοποίηση της οικονομικής συμπεριφοράς σε καλοήθη και ύποπτες, η μείωση διαστάσεων όταν το σύνολο δεδομένων έχει εκατοντάδες χαρακτηριστικά, η συσταδοποίηση έκφρασης γονιδίων, όπου μεγάλες ποσότητες γονιδίων μπορεί να εμφανίσουν παρόμοια συμπεριφορά κ.λπ..

2.2.6 Ταξινόμηση (classification)

Σύμφωνα με τον Edelstein (1999), τα προβλήματα ταξινόμησης στοχεύουν στον εντοπισμό των χαρακτηριστικών που υποδεικνύουν την ομάδα στην οποία ανήκει κάθε περίπτωση. Αυτό το μοτίβο μπορεί να χρησιμοποιηθεί τόσο για την κατανόηση των υπαρχόντων δεδομένων όσο και για την πρόβλεψη της συμπεριφοράς των νέων γεγονότων. Για παράδειγμα, μπορούμε να προβλέψουμε εάν τα άτομα μπορούν να ταξινομηθούν ως πιθανό να ανταποκριθούν σε μια αλληλογραφία ή ως καλούς υποψηφίους για χειρουργική επέμβαση. Η εξόρυξη δεδομένων δημιουργεί μοντέλα ταξινόμησης εξετάζοντας ήδη ταξινομημένα δεδομένα (περιπτώσεις) και επαγωγικά βρίσκοντας ένα προγνωστικό μοτίβο. Αυτές οι υπάρχουσες περιπτώσεις μπορεί να προέρχονται από μια ιστορική βάση δεδομένων, όπως άτομα που έχουν ήδη υποβληθεί σε συγκεκριμένη ιατρική περίθαλψη ή έχουν ανταποκριθεί σε μία αλληλογραφία. Μπορεί να προέρχονται από ένα πείραμα στο οποίο ένα δείγμα ολόκληρης της βάσης δεδομένων δοκιμάζεται στον πραγματικό κόσμο και τα αποτελέσματα χρησιμοποιούνται για τη δημιουργία ενός ταξινομητή (classifier).

Παραπέρα, τα γραφήματα και τα σχέδια είναι χρήσιμα για την κατανόηση των δισδιάστατων και τρισδιάστατων σχέσεων στα δεδομένα. Αλλά μερικές φορές οι ταξινομήσεις πρέπει να βασίζονται σε πολλές διαφορετικές προβλέψεις, που απαιτούν μια πολυδιάστατη πλοκή [1]. Επομένως, πρέπει να στραφούμε σε πιο εξελιγμένα μοντέλα για την εκτέλεση των εργασιών ταξινόμησης. Κοινές μέθοδοι εξόρυξης δεδομένων που χρησιμοποιούνται για την ταξινόμηση είναι ο k-πλησιέστερος γείτονας (k-nearest neighbor), τα δέντρα αποφάσεων (decision tree) και τα νευρωνικά δίκτυα (neural network). Τέλος, ως παραδείγματα εργασιών ταξινόμησης μπορούν να εκληφθούν ο προσδιορισμός του εάν μια συγκεκριμένη συναλλαγή με πιστωτική κάρτα είναι δόλια, η αξιολόγηση του εάν μια αίτηση στεγαστικού δανείου ενέχει ή όχι πιστωτικό κίνδυνο κ.λπ..

2.3 Αλγόριθμοι της εξόρυξης δεδομένων

Δοθέντος του ότι βαίνουμε σε όλο και συνθετότερες ταξινομήσεις (βλ. παραπάνω), η χρήση αλγορίθμων καθίσταται αναγκαία, νοούμενοι ως τα εξελιγμένα εκείνα μοντέλα που θα δώσουν λύση σε σύνθετα προβλήματα. Τα περισσότερα προϊόντα χρησιμοποιούν παραλλαγές αλγορίθμων που έχουν δημοσιευτεί σε επιστημονικά περιοδικά ή σε περιοδικά στατιστικών, με τις συγκεκριμένες εφαρμογές τους να προσαρμόζονται ώστε να ανταποκρίνονται στον στόχο του κάθε πωλητή. Για παράδειγμα, πολλοί προμηθευτές πωλούν εκδόσεις των δέντρων αποφάσεων CART ή CHAID με βελτιώσεις για εργασία σε παράλληλους υπολογιστές. Ορισμένοι προμηθευτές έχουν ιδιόκτητους αλγόριθμους οι οποίοι, ενώ δεν είναι επεκτάσεις ή βελτιώσεις οποιασδήποτε δημοσιευμένης προσέγγισης, μπορεί να λειτουργούν αρκετά καλά.

Γενικότερα, τα περισσότερα από τα μοντέλα και οι αλγόριθμοι μπορούν να θεωρηθούν ως γενικεύσεις του τυπικού –και αξιόπιστου– προϊόντος της μοντελοποίησης, το μοντέλο γραμμικής παλινδρόμησης. Πολλές προσπάθειες έχουν καταβληθεί στις στατιστικές, την επιστήμη των υπολογιστών, την τεχνητή νοημοσύνη και τη μηχανική για να ξεπεραστούν οι περιορισμοί αυτού του βασικού μοντέλου. Παράλληλα, το κοινό χαρακτηριστικό πολλών από τις νεότερες τεχνολογίες είναι ότι ο μηχανισμός εύρεσης προτύπων βασίζεται σε δεδομένα και δεν καθοδηγείται από τους χρήστες. Δηλαδή, οι σχέσεις εντοπίζονται επαγωγικά από το ίδιο το λογισμικό με βάση τα υπάρχοντα δεδομένα αντί να απαιτούν από τον διαμορφωτή να καθορίσει τη λειτουργική μορφή και τις αλληλεπιδράσεις. Αξίζει να αναφερθεί ότι κανένα μοντέλο ή αλγόριθμος δεν μπορεί ή πρέπει να χρησιμοποιείται αποκλειστικά [4]. Για οποιοδήποτε δεδομένο πρόβλημα, η ίδια η φύση των δεδομένων θα κατευθύνει την επιλογή των μοντέλων και των αλγορίθμων. Κατά συνέπεια, απαιτείται μια ποικιλία εργαλείων και τεχνολογιών με στόχο το καλύτερο δυνατό μοντέλο. Έτσι, μεταξύ άλλων τα συνηθέστερα εργαλεία είναι οι: λογιστική παλινδρόμηση, νευρωνικά δίκτυα, δέντρα απόφασης, k-πλησιέστερος.

2.3.1 Λογιστική παλινδρόμηση (logistic regression)

Κατά τον Edelstein (1999) [4], η παλινδρόμηση χρησιμοποιεί υπάρχουσες τιμές για να προβλέψει ποιες άλλες θα προκύψουν. Στην απλούστερη περίπτωση (σε αυτή τη μορφή θα μπορούσε να ενταχθεί στις τεχνικές εξόρυξης δεδομένων που αναλύσαμε παραπάνω), η παλινδρόμηση χρησιμοποιεί τυπικές στατιστικές τεχνικές, όπως η γραμμική παλινδρόμηση. Δυστυχώς, πολλά πραγματικά προβλήματα δεν είναι απλώς γραμμικές προβολές προηγούμενων τιμών. Για παράδειγμα, ο όγκος των πωλήσεων, οι τιμές των μετοχών και τα ποσοστά αποτυχίας ενός προϊόντος είναι πολύ δύσκολο να προβλεφθούν, επειδή μπορεί να εξαρτώνται από πολύπλοκες αλληλεπιδράσεις πολλαπλών μεταβλητών πρόβλεψης. Επομένως, ενδέχεται να απαιτούνται πιο πολύπλοκες τεχνικές για την πρόβλεψη μελλοντικών τιμών όπως είναι η λογιστική παλινδρόμηση. Τέλος, ο ίδιος τύπος μοντέλου μπορεί συχνά να χρησιμοποιηθεί και σε άλλες τεχνικές όπως για παράδειγμα η ταξινόμηση.

2.3.2. Νευρωνικά δίκτυα (neural networks)

Τα νευρικά δίκτυα παρουσιάζουν ιδιαίτερο ενδιαφέρον επειδή προσφέρουν ένα μέσο αποτελεσματικής μοντελοποίησης μεγάλων και πολύπλοκων προβλημάτων στα οποία μπορεί να υπάρχουν εκατοντάδες μεταβλητές πρόβλεψης που έχουν πολλές αλληλεπιδράσεις (τα πραγματικά βιολογικά νευρικά δίκτυα είναι ασύγκριτα πιο περίπλοκα.) Τα νευρωνικά δίκτυα μπορούν να χρησιμοποιηθούν στα προβλήματα ταξινόμησης (όπου η έξοδος είναι κατηγορηματική μεταβλητή) ή της παλινδρόμησης (όπου η μεταβλητή εξόδου είναι συνεχής). Τα βάρη σύνδεσης (W’s) είναι οι άγνωστες παράμετροι που υπολογίζονται με μια μέθοδο εκμάθησης (training method) [4]. Αρχικά, η πιο κοινή μέθοδος εκμάθησης ήταν η αναπαράσταση (backpropagation). Νεότερες μέθοδοι περιλαμβάνουν την κλίση συζευγμένων, quasi-Newton, Levenberg-Marquardt και γενετικούς αλγόριθμους. Κάθε μέθοδος εκμάθησης έχει ένα σύνολο παραμέτρων που ελέγχουν διάφορες πτυχές, όπως αποφυγή τοπικών βέλτιστων ή ρύθμιση της ταχύτητας μετατροπής.

Παραπέρα, η αρχιτεκτονική (ή τοπολογία) ενός νευρονικού δικτύου είναι ο αριθμός των κόμβων και των κρυφών επιπέδων και ο τρόπος με τον οποίο συνδέονται. Κατά το σχεδιασμό ενός νευρονικού δικτύου, είτε ο χρήστης είτε το λογισμικό πρέπει να επιλέξουν τον αριθμό των κρυφών κόμβων και των κρυφών

επιπέδων, τη λειτουργία ενεργοποίησης και τα όρια στα βάρη. Ένας από τους πιο συνηθισμένους τύπους νευρωνικού δικτύου είναι το δίκτυο backpropagation τροφοδοσίας προς τα εμπρός (feed-forward backpropagation). Η εκμάθηση backpropagation είναι απλώς μια έκδοση της καθόδου βασισμένης στην κλίση (gradient descent), ενός τύπου αλγορίθμου που προσπαθεί να μειώσει μια τιμή-στόχο (σφάλμα, στην περίπτωση νευρωνικών δικτύων) σε κάθε βήμα. Ο αλγόριθμος προχωρά ως εξής:

- Προώθηση τροφοδοσίας: Η τιμή του κόμβου (node) εξόδου υπολογίζεται με βάση τις τιμές του κόμβου εισόδου και ένα σύνολο αρχικών βαρών. Οι τιμές από τους κόμβους εισόδου συνδυάζονται στα κρυφά επίπεδα, και οι τιμές αυτών των κόμβων συνδυάζονται για τον υπολογισμό της τιμής εξόδου.
- Backpropagation: Το σφάλμα στην έξοδο υπολογίζεται βρίσκοντας τη διαφορά μεταξύ της υπολογισμένης εξόδου και της επιθυμητής εξόδου (δηλαδή, οι πραγματικές τιμές που βρέθηκαν στο σετ εκμάθησης). Στη συνέχεια, το σφάλμα από την έξοδο αποδίδεται στους κρυφούς κόμβους επιπέδου αναλογικά με τα βάρη. Αυτό επιτρέπει τον υπολογισμό ενός σφάλματος για κάθε κόμβο εξόδου και κρυφό κόμβο στο δίκτυο. Τέλος, το σφάλμα σε κάθε έναν από τους κρυφούς κόμβους και κόμβους εξόδου χρησιμοποιείται από τον αλγόριθμο για να προσαρμόσει το βάρος που εισέρχεται σε αυτόν τον κόμβο μειώνοντας το σφάλμα.

Αυτή η διαδικασία επαναλαμβάνεται για κάθε σειρά στο σετ εκμάθησης. Κάθε πέρασμα από όλες τις σειρές στο σετ εκμάθησης ονομάζεται εποχή (epoch). Το σετ εκμάθησης θα χρησιμοποιηθεί επανειλημμένα, έως ότου το σφάλμα δεν μειώνεται πλέον. Σε αυτό το σημείο το νευρονικό δίκτυο θεωρείται εκπαιδευμένο για να βρει το μοτίβο στο σύνολο δοκιμών. Επειδή μπορεί να υπάρχουν τόσες πολλές παράμετροι στα κρυμμένα στρώματα, ένα νευρονικό δίκτυο με αρκετούς κρυμμένους κόμβους θα ταιριάζει πάντα τελικά με το σετ εκμάθησης εάν αφεθεί να τρέξει αρκετά. Όμως, για να αποφευχθεί ένα υπερβολικό νευρονικό δίκτυο που θα λειτουργεί καλά μόνο στα δεδομένα εκμάθησης, πρέπει να διακόπτεται η εκμάθηση [4]. Ορισμένες εφαρμογές θα αξιολογούν περιοδικά το νευρονικό δίκτυο έναντι των δεδομένων δοκιμής κατά τη διάρκεια της εκμάθησης. Όσο μειώνεται το ποσοστό σφάλματος στο σύνολο δοκιμών, η εκμάθηση θα συνεχιστεί. Εάν το ποσοστό σφάλματος στα δεδομένα δοκιμής αυξηθεί, παρόλο που το ποσοστό σφάλματος στα δεδομένα

εκμάθησης εξακολουθεί να μειώνεται, τότε το νευρικό δίκτυο μπορεί να υπερβαίνει τα δεδομένα.

Τα νευρικά δίκτυα διαφέρουν στη φιλοσοφία από πολλές στατιστικές μεθόδους με διάφορους τρόπους. Πέρα και πάνω από όλα, ένα νευρονικό δίκτυο έχει συνήθως περισσότερες παραμέτρους από ό,τι ένα τυπικό στατιστικό μοντέλο. Επειδή είναι τόσο πολυάριθμες και επειδή τόσοι πολλοί συνδυασμοί παραμέτρων οδηγούν σε παρόμοιες προβλέψεις, οι παράμετροι γίνονται ακατανόητες και το δίκτυο χρησιμεύει ως πρόβλεψη «μαύρου κουτιού». Στην πραγματικότητα, ένα δεδομένο αποτέλεσμα μπορεί να συσχετιστεί με πολλά διαφορετικά σύνολα βαρών. Κατά συνέπεια, τα βάρη του δικτύου γενικά δεν βοηθούν στην κατανόηση της υποκείμενης διαδικασίας που δημιουργεί την πρόβλεψη. Ωστόσο, αυτό είναι αποδεκτό σε πολλές εφαρμογές. Εντούτοις, ένα πλεονέκτημα των μοντέλων νευρωνικών δικτύων είναι ότι μπορούν εύκολα να εφαρμοστούν για να εκτελούνται σε μαζικά παράλληλους υπολογιστές με κάθε κόμβο να κάνει ταυτόχρονα τους δικούς του υπολογισμούς.

Οι χρήστες πρέπει να γνωρίζουν πολλά στοιχεία σχετικά με τα νευρικά δίκτυα: Πρώτον, τα νευρικά δίκτυα δεν ερμηνεύονται εύκολα. Δεν υπάρχει ρητή αιτιολογία για τις αποφάσεις ή τις προβλέψεις που κάνει ένα νευρωνικό δίκτυο. Δεύτερον, τείνουν να υπερκαλύπτουν τα δεδομένα εκμάθησης εκτός και αν χρησιμοποιούνται πολύ αυστηρά μέτρα, όπως μείωση του βάρους ή/και εγκάρσια επικύρωση (cross validation). Αυτό οφείλεται στον πολύ μεγάλο αριθμό παραμέτρων του νευρικού δικτύου που, εάν επιτρέπεται να είναι επαρκού μεγέθους, θα ταιριάζει σε οποιοδήποτε σύνολο δεδομένων αυθαίρετα καλά όταν επιτρέπεται να μαθαίνει συγκλίνοντας. Τρίτον, τα νευρονικά δίκτυα απαιτούν εκτεταμένο χρόνο εκμάθησης εκτός εάν το πρόβλημα είναι πολύ μικρό. Μόλις εκπαιδευτούν, ωστόσο, μπορούν να παρέχουν προβλέψεις πολύ γρήγορα. Τέταρτον, δεν απαιτούν λιγότερη προετοιμασία δεδομένων από οποιαδήποτε άλλη μέθοδο, δηλαδή να απαιτούν πολλή προετοιμασία δεδομένων.

Αξίζει να αναφερθεί πως ένας μύθος των νευρωνικών δικτύων είναι ότι δεδομένα κάθε ποιότητας μπορούν να χρησιμοποιηθούν για την παροχή λογικών προβλέψεων. Οι πιο επιτυχημένες υλοποιήσεις νευρωνικών δικτύων (ή δένδρων αποφάσεων, ή λογιστικής παλινδρόμησης ή οποιασδήποτε άλλης μεθόδου) περιλαμβάνουν πολύ προσεκτικό καθαρισμό δεδομένων, επιλογή, προετοιμασία και προεπεξεργασία. Για παράδειγμα, τα νευρικά δίχτυα απαιτούν όλες οι μεταβλητές να είναι αριθμητικές. Επομένως, τα κατηγορηματικά δεδομένα όπως το "state"

χωρίζονται συνήθως σε πολλαπλές διχοτόμες μεταβλητές (π.χ. "California", "New York"), καθεμία με τιμή "1" (ναι) ή "0" (όχι). Η προκύπτουσα αύξηση των μεταβλητών ονομάζεται κατηγορηματική έκρηξη (categorical explosion). Τέλος, τα νευρωνικά δίκτυα τείνουν να λειτουργούν καλύτερα όταν το σύνολο δεδομένων είναι αρκετά μεγάλο και η αναλογία τόνου σήματος είναι αρκετά υψηλή. Επειδή είναι τόσο ευέλικτα, θα βρουν πολλά ψεύτικα μοτίβα σε περιπτώσεις χαμηλού λόγου σήματος προς θόρυβο.

2.3.3 Δέντρα απόφασης (decision trees)

Όπως εύστοχα παρατηρεί ο Edelstein [4], τα δέντρα απόφασης είναι ένας τρόπος αναπαράστασης μιας σειράς κανόνων που οδηγούν σε μια τάξη ή αξία. Ένα κλασικό παράδειγμα –το οποίο μετέρχεται η ανά χείρας εργασία– συστήνει η ταξινόμηση των αιτούντων δάνειο ως καλούς ή κακούς πιστωτικούς κινδύνους. Εν προκειμένω, όσον αφορά τη δομή των δέντρων, το πρώτο συστατικό είναι ο κόμβος κορυφαίας απόφασης ή, αλλιώς, ριζικός κόμβος (root node) ο οποίος καθορίζει μια δοκιμή που θα πραγματοποιηθεί. Παραπέρα, τα αποτελέσματα αυτής της δοκιμής αναγκάζουν το δέντρο να χωρίσει σε κλαδιά, το καθένα αντιπροσωπεύοντας μία από τις πιθανές απαντήσεις. Ανάλογα με τον αλγόριθμο, κάθε κόμβος μπορεί να έχει δύο ή περισσότερους κλάδους. Για παράδειγμα, το CART δημιουργεί δέντρα με δύο μόνο κλάδους σε κάθε κόμβο. Ένα τέτοιο δέντρο ονομάζεται δυαδικό δέντρο (binary tree). Όταν επιτρέπονται περισσότερα από δύο κλαδιά, ονομάζεται δέντρο πολλαπλών δρόμων (multiway tree). Κάθε κλάδος οδηγεί είτε σε έναν άλλο κόμβο αποφάσεων είτε στο κάτω μέρος του δέντρου, που ονομάζεται κόμβος φύλλων (leaf node).

Επιπρόσθετα, κατά την περιήγηση στο δέντρο αποφάσεων δίνεται η δυνατότητα ορισμού μιας τιμής ή κλάσης σε μια περίπτωση αποφασίζοντας ποιος κλάδος θα ληφθεί, ξεκινώντας από τον ριζικό κόμβο και μεταβαίνοντας σε κάθε επόμενο κόμβο μέχρι να επιτευχθεί ένας κόμβος φύλλων. Κάθε κόμβος χρησιμοποιεί τα δεδομένα από την περίπτωση για να επιλέξει τον κατάλληλο κλάδο. Οπλισμένος με αυτό το δείγμα δέντρου και μια αίτηση δανείου, ένας υπεύθυνος δανείου θα μπορούσε να καθορίσει εάν ο αιτών ήταν καλός ή κακός πιστωτικός κίνδυνος. Τα μοντέλα δέντρων απόφασης χρησιμοποιούνται συνήθως στην εξόρυξη δεδομένων για να εξετάσουν τα δεδομένα και να δημιουργήσουν (induce) το δέντρο και τους κανόνες του που θα χρησιμοποιηθούν για την πραγματοποίηση προβλέψεων.

Γενικότερα, μπορούν να χρησιμοποιηθούν διάφοροι αλγόριθμοι για τη δημιουργία δένδρων αποφάσεων, όπως CHAID (Chi-squared Automatic Interaction Detection), CART (Classification And Regression Trees), Quest και C5.0 [4]. Τα δέντρα αποφάσεων αναπτύσσονται μέσω επαναληπτικού διαχωρισμού δεδομένων σε διακριτές ομάδες, όπου ο στόχος είναι να μεγιστοποιηθεί η απόσταση μεταξύ ομάδων σε κάθε διαχωρισμό. Μία από τις διακρίσεις μεταξύ των μεθόδων του δέντρου αποφάσεων είναι ο τρόπος μέτρησης αυτής της απόστασης. Ενώ οι λεπτομέρειες αυτών των μετρήσεων είναι πέρα από το πεδίο αυτής της εισαγωγής, ο διαχωρισμός διαχωρίζει τα δεδομένα σε νέες ομάδες που είναι όσο το δυνατόν πιο διαφορετικές μεταξύ τους.

Δέντρα αποφάσεων που χρησιμοποιούνται για την πρόβλεψη κατηγορηματικών μεταβλητών ονομάζονται δέντρα ταξινόμησης (classification trees) επειδή τοποθετούν περιπτώσεις σε κατηγορίες ή τάξεις. Τα δέντρα απόφασης που χρησιμοποιούνται για την πρόβλεψη συνεχών μεταβλητών ονομάζονται δέντρα παλινδρόμησης (regression trees) [4]. Από την άλλη μεριά, τα δέντρα μπορούν να γίνουν πολύ περίπλοκα. Αν και ένα δέντρο αποφάσεων μπορεί να εξηγήσει τις προβλέψεις του –το οποίο είναι ένα σημαντικό πλεονέκτημα– ωστόσο, αυτή η σαφήνεια μπορεί να είναι κάπως παραπλανητική. Επιπλέον, τα δέντρα αποφάσεων κάνουν λίγα περάσματα στα δεδομένα (όχι περισσότερο από ένα πέρασμα για κάθε επίπεδο του δέντρου) και λειτουργούν καλά με πολλές μεταβλητές πρόβλεψης. Κατά συνέπεια, τα μοντέλα μπορούν να κατασκευαστούν πολύ γρήγορα, καθιστώντας τα κατάλληλα για μεγάλα σύνολα δεδομένων. Τα δέντρα που αφήνονται να αναπτυχθούν χωρίς δέσμευση χρειάζονται περισσότερο χρόνο για να φτιαχτούν και να γίνουν ακατάληπτα, αλλά το πιο σημαντικό είναι ότι υπερκεράσουν (overfit) τα δεδομένα.

Ως εκ τούτου, το μέγεθος του δέντρου μπορεί να ελεγχθεί μέσω κανόνων διακοπής που περιορίζουν την ανάπτυξη. Ένας κοινός κανόνας διακοπής είναι απλώς ο περιορισμός του μέγιστου βάθους στο οποίο μπορεί να μεγαλώσει ένα δέντρο. Ένας άλλος κανόνας διακοπής είναι να καθοριστεί ένα κατώτερο όριο στον αριθμό των εγγραφών σε έναν κόμβο και να μην χωρίζει κάτω από αυτό το όριο. Παράλληλα, μια εναλλακτική λύση για τη διακοπή των κανόνων είναι να κλαδέψετε το δέντρο. Το δέντρο αφήνεται να αναπτυχθεί στο πλήρες μέγεθός του και στη συνέχεια, χρησιμοποιώντας είτε ενσωματωμένη ευρετική είτε παρέμβαση χρήστη, το δέντρο κλαδεύεται πίσω στο μικρότερο μέγεθος που δεν θέτει σε κίνδυνο την ακρίβεια. Το

CART κλαδεύει τα δέντρα διασταυρώνοντάς τα ώστε να φανεί εάν η βελτίωση της ακρίβειας δικαιολογεί τους επιπλέον κόμβους.

Μια κοινή κριτική για τα δέντρα αποφάσεων είναι ότι επιλέγουν μια διάσπαση χρησιμοποιώντας έναν άπληστο αλγόριθμο στον οποίο η απόφαση για την οποία η μεταβλητή θα χωρίσει δεν λαμβάνει υπόψη οποιαδήποτε επίδραση που μπορεί να έχει η διάσπαση στις μελλοντικές διασπάσεις [4]. Με άλλα λόγια, η απόφαση διαχωρισμού λαμβάνεται στον κόμβο αυτή τη στιγμή και δεν επανεξετάζεται ποτέ. Επιπλέον, όλα τα διαχωριστικά γίνονται διαδοχικά, έτσι κάθε διαχωρισμός εξαρτάται από τον προκάτοχό του. Επομένως, όλες οι μελλοντικές διασπάσεις εξαρτώνται από τον πρώτο διαχωρισμό, πράγμα που σημαίνει ότι η τελική λύση θα μπορούσε να είναι πολύ διαφορετική εάν γίνει διαφορετικός πρώτος διαχωρισμός. Το πλεονέκτημα του να κοιτάς μπροστά για να κάνεις τα καλύτερα χωρίσματα με βάση δύο ή περισσότερα επίπεδα ταυτόχρονα είναι ασαφές.

Επιπλέον, οι αλγόριθμοι που χρησιμοποιούνται για το διαχωρισμό είναι γενικά μονομετάβλητοι (univariate). Δηλαδή, θεωρούν μόνο μία μεταβλητή προβλέψεων κάθε φορά. Και ενώ αυτή η προσέγγιση είναι ένας από τους λόγους για τους οποίους το μοντέλο δημιουργείται γρήγορα –περιορίζει τον αριθμό των πιθανών κανόνων διαχωρισμού για δοκιμή– καθιστά επίσης δυσκολότερη την ανίχνευση σχέσεων μεταξύ των μεταβλητών πρόβλεψης. Τα δέντρα αποφάσεων που δεν περιορίζονται σε διαχωρισμούς μονομεταβλητότητας θα μπορούσαν να χρησιμοποιούν πολλές μεταβλητές πρόβλεψης σε έναν κανόνα διαχωρισμού. Ένα τέτοιο δέντρο αποφάσεων θα μπορούσε να επιτρέψει γραμμικούς συνδυασμούς μεταβλητών, επίσης γνωστοί ως πλάγια δέντρα (oblique trees).

Κλείνοντας, αξίζει να αναφερθεί ότι τα δέντρα αποφάσεων χειρίζονται μη αριθμητικά δεδομένα πολύ καλά [4]. Αυτή η ικανότητα αποδοχής κατηγορικών δεδομένων ελαχιστοποιεί την ποσότητα των μετασχηματισμών δεδομένων και την έκρηξη των μεταβλητών πρόβλεψης που είναι εγγενείς στα νευρονικά δίκτυα. Ορισμένα δέντρα ταξινόμησης σχεδιάστηκαν για –και ως εκ τούτου λειτουργούν καλύτερα– όταν οι μεταβλητές πρόβλεψης είναι επίσης κατηγορηματικές. Οι συνεχείς προβλέψεις μπορούν συχνά να χρησιμοποιηθούν ακόμη και σε αυτές τις περιπτώσεις μετατρέποντας τη συνεχή μεταβλητή σε ένα σύνολο εύρους (binning). Ορισμένα δέντρα αποφάσεων δεν υποστηρίζουν μεταβλητές συνεχούς απόκρισης (δηλαδή, δεν θα δημιουργήσουν δέντρα παλινδρόμησης), οπότε οι μεταβλητές απόκρισης στο σετ εκμάθησης πρέπει επίσης να ενσωματωθούν σε τάξεις εξόδου.

2.3.4 Κ-πλησιέστερος γείτονας (k-nearest neighbor) και συλλογισμός βάσει μνήμης MBR (memory-based reasoning)

Γενικότερα, όταν οι άνθρωποι προσπαθούν να λύσουν νέα προβλήματα εξετάζουν συχνά λύσεις σε παρόμοια προβλήματα που είχαν προηγουμένως επιλύσει. Ο k-πλησιέστερος γείτονας (k-NN) είναι μια τεχνική ταξινόμησης που χρησιμοποιεί μια έκδοση αυτής της ίδιας μεθόδου. Αποφασίζει σε ποια τάξη θα τοποθετήσει μια νέα θήκη εξετάζοντας κάποιον αριθμό –το "k" σε k-πλησιέστερο γείτονα– από παρόμοιες περιπτώσεις ή γείτονες. Μετράει τον αριθμό των περιπτώσεων για κάθε τάξη και εκχωρεί τη νέα θήκη στην ίδια τάξη στην οποία ανήκουν οι περισσότεροι από τους γείτονές της. Το πρώτο πράγμα που πρέπει να δρομολογηθεί είναι να βρεθεί ένα μέτρο της απόστασης μεταξύ των χαρακτηριστικών στα δεδομένα και, στη συνέχεια, να υπολογιστεί. Αν και αυτό είναι εύκολο για αριθμητικά δεδομένα, οι κατηγορηματικές μεταβλητές χρειάζονται ειδικό χειρισμό (όπως, για παράδειγμα, ποια είναι η απόσταση μεταξύ μπλε και πράσινου). Στη συνέχεια, πρέπει να συνοψιστούν τα μέτρα απόστασης για τα χαρακτηριστικά. Μόλις υπολογιστεί η απόσταση μεταξύ των περιπτώσεων, στη συνέχεια επιλέγεται το σύνολο των ήδη διαβαθμισμένων περιπτώσεων που θα χρησιμοποιηθούν ως βάση για την ταξινόμηση νέων περιπτώσεων και αποτιμάται το εύρος της γειτονιάς που θα γίνουν οι συγκρίσεις· τέλος, μετριούνται οι ίδιοι οι γείτονες.

Το K-NN βάζει ένα μεγάλο υπολογιστικό φορτίο στον υπολογιστή, επειδή ο χρόνος υπολογισμού αυξάνεται ως παράγοντας του συνολικού αριθμού πόντων. Αν και είναι μια γρήγορη διαδικασία για την εφαρμογή ενός δέντρου αποφάσεων ή ενός νευρονικού δικτύου σε μια νέα περίπτωση, το k-NN απαιτεί να γίνεται νέος υπολογισμός για κάθε νέα περίπτωση. Για να επιταχυνθεί το k-NN, συχνά όλα τα δεδομένα διατηρούνται στη μνήμη. Η συλλογιστική βάσει μνήμης αναφέρεται συνήθως σε έναν ταξινομητή k-NN που διατηρείται στη μνήμη. Τα μοντέλα K-NN είναι πολύ εύκολα κατανοητά όταν υπάρχουν λίγες μεταβλητές πρόβλεψης. Είναι επίσης χρήσιμα για την κατασκευή μοντέλων που περιλαμβάνουν μη τυπικούς τύπους δεδομένων, όπως κείμενο. Η μόνη απαίτηση για τη δυνατότητα συμπερίληψης ενός τύπου δεδομένων είναι η ύπαρξη κατάλληλης μέτρησης.

2.3.5 Μέτρα αξιολόγησης παραγόμενης γνώσης

Ένα από τα βασικότερα κομμάτια της μηχανικής μάθησης που συνήθως βρίσκεται και στο τελευταίο στάδιο των εργασιών είναι η αξιολόγηση των επιδόσεων των μοντέλων ή των αλγορίθμων μάθησης. Σκοπός είναι να δούμε τις ικανότητες του συστήματος στην πρόβλεψη μελλοντικών αποτελεσμάτων και ποιες είναι οι πιθανότητες λάθους [31]. Η μετρική που αποδεδειγμένα χρησιμοποιείται κατά κόρον είναι η **ακρίβεια ή ευστοχία(Accuracy)**. Η μετρική αυτή ορίζεται ως εξής:

$$Accuracy = \frac{Σωστές\ Προβλέψεις}{Σύνολο\ Προβλέψεων}$$

Οι τιμές της είναι στο διάστημα από [0,1] και το μοντέλο που αγγίζει την τιμή 1, θεωρείται ως το πιο εύστοχο.

Στις περιπτώσεις που έχουμε δυαδικό πρόβλημα ταξινόμησης και οι κλάσεις μας είναι *True or False*, τα αποτελέσματα που παίρνουμε είναι τα παρακάτω.

1. True Positive / TP / Αληθώς θετικές περιπτώσεις.
   Επιτυχημένη πρόβλεψη
2. True Negative / TN / Αληθώς αρνητικές περιπτώσεις
   Επιτυχημένη απόρριψη
3. False Positive / FP / Ψευδώς θετικές περιπτώσεις
   Ψευδές σήμα συναγερμού (False alarm), Υπερεκτίμηση της κατάστασης
4. False Negative / FN / Ψευδώς αρνητικές περιπτώσεις
   Αστοχία πρόβλεψης, υποτίμηση της κατάστασης

Η παραπάνω μετρική αξιολόγησης, accuracy, πολλές φορές μας οδηγεί σε λάθος συμπεράσματα, αλγόριθμοι και μοντέλα που φαινομενικά αξιολογούνται με μεγάλο accuracy τελικώς αποτυγχάνουν σε νέες προβλέψεις και τα αρχικά δείγματα που είχαμε ήταν παραπλανητικά. Σε άλλες περιπτώσεις βάζοντας σε σύγκριση δυο μοντέλα βάσει του accuracy, υπάρχει πιθανότητα το μοντέλο με την χαμηλότερη ευστοχία να είναι αποδοτικότερο και σωστότερο στο τέλος. Μία πιθανή κατάσταση όπου συμβαίνει αυτό το φαινόμενο είναι όταν το αρχείο με τα δεδομένα προς μάθηση

δεν έχει ίση κατανομή των κλάσεων, κατ' επέκταση το φαινόμενο παρατηρείται και όταν έχουμε προβλήματα πολλών κλάσεων προς εκπαίδευση και κατηγοριοποίηση (multi-class problems). Αυτές οι περιπτώσεις μας αναγκάζουν να χρησιμοποιήσουμε και κάποιες εναλλακτικές επιλογές του accuracy, που κινούνται στο ίδιο σκεπτικό περίπου.

Precision (Ακρίβεια). Το ποσοστό των εγγραφών που ο ταξινομητής έχει ταξινομήσει ως θετικά και είναι στην πραγματικότητα θετικά.

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

Recall (Ανάκληση). Το ποσοστό των θετικών εγγραφών που εντόπισε ο ταξινομητής.

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

F-Score ή F-Measure. Μετρική που συνδυάζει το precision και το recall.

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

Κλείνοντας για να έχουμε μία καλύτερη εικόνα των αποτελεσμάτων και της απόδοσης των μοντέλων χρησιμοποιούμε τον **πίνακα σύγχυσης (Confusion Matrix)** [32]. Οι πίνακες σύγχυσης υπολογίζονται χρησιμοποιώντας τις προβλέψεις ενός μοντέλου σε ένα σύνολο δεδομένων. Εξετάζοντας έναν πίνακα σύγχυσης, μπορείτε να κατανοήσετε καλύτερα τα δυνατά και τα αδύνατα σημεία του μοντέλου σας και μπορείτε να συγκρίνετε καλύτερα δύο εναλλακτικά μοντέλα για να καταλάβετε ποιο είναι καλύτερο για την εφαρμογή σας. Παραδοσιακά, ένας πίνακας σύγχυσης υπολογίζεται χρησιμοποιώντας τις προβλέψεις ενός μοντέλου σε ένα συγκρατημένο σύνολο δοκιμών.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Εικόνα 1 Πίνακας Σύγχυσης (Confusion Matrix)

3 Risk Assessment (διαχείριση κινδύνου)

3.1 Εισαγωγή

Στο παρόν κεφάλαιο γίνεται μια εισαγωγή στην έννοια του κινδύνου και συγκεκριμένα της διαχείρισης του κινδύνου. Στη συνέχεια, παρουσιάζονται αναλυτικά οι κατηγορίες του τραπεζικού κινδύνου, όπως για παράδειγμα ο πιστωτικός κίνδυνος και κίνδυνος αγοράς. Επίσης, αναλύεται το σύστημα πιστωτικού κινδύνου και οι βασικοί παράμετροι του τραπεζικού κινδύνου, όπως και η έννοια του Value at Risk. Τέλος, αναλύονται οι μέθοδοι εκτίμησης και περιορισμού πιστωτικού κινδύνου, όπως το credit scoring και γίνεται ανασκόπηση των μεθόδων που χρησιμοποιούνται για τον υπολογισμό του.

Η έννοια του κινδύνου ορίζει οποιαδήποτε αβεβαιότητα μπορεί να προκαλέσει απώλεια. Οι τράπεζες καλούνται καθημερινά να αναλαμβάνουν και να αντιμετωπίζουν κινδύνους. Συγκεκριμένα, ο κίνδυνος εδώ αναφέρεται στη πιθανότητα απόκλισης και ύπαρξης ζημίας στα αναμενόμενα κέρδη ή αποτελέσματα μιας χρηματοοικονομικής θέσης της τράπεζας [5], [6].

Η διαχείριση του κινδύνου στον τραπεζικό τομέα αναφέρεται τόσο στις διαδικασίες όσο και στα μοντέλα, τα οποία μια τράπεζα μπορεί να αξιοποιήσει ώστε να εφαρμόσει πολιτικές και πρακτικές που βασίζονται στον κίνδυνο. Περιλαμβάνει τεχνικές και εργαλεία απαραίτητα για τη μέτρηση, την παρακολούθηση και τον έλεγχο του κινδύνου. Τα μοντέλα και οι διαδικασίες που περιέχει εκτείνονται σε όλων των ειδών κινδύνους: πιστωτικός κίνδυνος, κίνδυνος αγοράς, επιτοκιακός κίνδυνος, κίνδυνος ρευστότητας, κίνδυνος συναλλάγματος και λειτουργικός κίνδυνος [7].

Είναι απαραίτητο, επομένως, οι τράπεζες να μπορούν να αναγνωρίσουν τους κινδύνους και να τους ταξινομήσουν σε μια από τις παραπάνω κατηγορίες με σκοπό να εφαρμόσουν τα ανάλογα μοντέλα για την ποσοτικοποίησή τους. Αυτό γίνεται υπό το πρίσμα ενός ακριβούς πλαισίου που έχει οριστεί από τις εποπτικές αρχές και στοχεύει στην αναγνώριση του εκάστοτε κινδύνου καθώς και στη διαχείρισή του. Στη συνέχεια, πρέπει να γίνει η μέτρηση του κινδύνου, δηλαδή να εκτιμηθεί η αναμενόμενη ζημιά και να αξιολογηθεί εάν η αύξηση της θα οδηγήσει σε ανάλογη αύξηση των κεφαλαιακών απαιτήσεων [7].

Τέλος, μετά το πέρας των παραπάνω διεργασιών, η τράπεζα θα πρέπει να καταλήξει σε συγκεκριμένες αποφάσεις για τη διαχείριση του κινδύνου. Συγκεκριμένα, μπορεί να αναλάβει τον κίνδυνο ή να εφαρμόσει μεθόδους για τη μείωση του [6].

3.2 Κατηγορίες Τραπεζικού Κινδύνου

Η ραγδαία εξέλιξη της τεχνολογίας, η δημιουργία κεντρικών τραπεζών, οι καινούργιοι κανόνες εποπτείας, όπως και η ρευστότητα σε παγκόσμιο επίπεδο είναι μόνο λίγοι από τους παράγοντες που συνετέλεσαν στην εμφάνιση των τραπεζικών κινδύνων [11]. Όλα αυτά σε συνδυασμό με το οικονομικό περιβάλλον (παγκόσμια κρίση), τη παγκοσμιοποίηση και τον αστάθμητο ανθρώπινο παράγοντα φέρνουν τις τράπεζες αντιμέτωπες με μια πληθώρα κινδύνων.

Οι βασικές κατηγορίες τραπεζικού κινδύνου που καλούνται να αντιμετωπίσουν οι τράπεζες σήμερα είναι οι ακόλουθες:

3.2.1 Πιστωτικός Κίνδυνος (Credit Risk)

Ο πιστωτικός κίνδυνος θεωρείται από τους πιο σημαντικούς κινδύνους σε ένα χρηματοπιστωτικό ίδρυμα [7]. Σχετίζεται άμεσα με το κίνδυνο της αθέτησης-αφερεγγυότητας (default risk), δηλαδή την αδυναμία ενός πελάτη να ανταποκριθεί στις υποχρεώσεις που έχει ως προς το πιστωτικό του χρέος. Η αθέτηση αυτή μπορεί να πυροδοτήσει μερική ή ολική απώλεια του ποσού που έχει δανείσει η τράπεζα στον πελάτη. Ο πιστωτικός κίνδυνος μεταφράζεται επίσης ως ο κίνδυνος μείωσης της πιστοληπτικής ικανότητας ενός υπόχρεου ομολόγου ή μετοχής.

Οι τράπεζες καλούνται να αντιμετωπίζουν καθημερινά υψηλά ποσοστά πιστωτικού κινδύνου, αφού η παροχή δανείων αποτελεί μια από τις βασικότερες λειτουργίες της. Για αυτό το λόγο είναι απαραίτητο να μπορεί να αξιολογήσει την πιστοληπτική ικανότητα των πελατών της επιστρατεύοντας μεθόδους και μοντέλα, που θα δούμε στη συνέχεια, και να λαμβάνει τα ανάλογα μέτρα για αποφύγει την πιθανότητα αθέτησης εκ μέρους του πιστούχου. Ο πιστωτικός κίνδυνος μπορεί να επιμεριστεί σε τέσσερις κατηγορίες:

- Κίνδυνος Αθέτησης (Default Risk): Όπως αναφέρθηκε και πιο πάνω αφορά την αδυναμία του πιστούχου να αποπληρώσει τις υποχρεώσεις του. Αυτή η αδυναμία μπορεί να μεταφραστεί σε καθυστέρηση, η οποία μπορεί να αντιμετωπιστεί με πρόληψη, ειδάλλως σε περίπτωση μόνιμης αδυναμίας προκαλεί μεγάλα προβλήματα στη τράπεζα. Το ίδιο συμβαίνει και με την ολική αθέτηση, ενώ σε περίπτωση μονομερής ακύρωσης μιας συμφωνίας μπορεί να γίνει διακανονισμός [12].
- Κίνδυνος Έκθεσης (Exposure Risk): Αφορά το σύνολο του ποσού στο οποίο εκτίθεται μια τράπεζα σε περίπτωση αθέτησης και είναι αρκετά δύσκολο να υπολογιστεί. Οι πιστωτικές κάρτες και τα ανοιχτά επαγγελματικά δάνεια είναι ανοιχτές πιστώσεις που μπορούν να προκαλέσουν αυτό τον κίνδυνο. Η χρήση εγγυήσεων μπορεί να αποτελέσει ένα ανασταλτικό μέτρο αυτού του κινδύνου.
- Κίνδυνος Ανάκτησης (Recovery Risk): Είναι το ποσό που θα μπορέσει να ανακτήσει η τράπεζα σε περίπτωση αθέτησης υποχρέωσης.
- Κίνδυνος Περιθωρίου (Credit Spread Risk). Είναι η διαφορά μεταξύ των αποδόσεων των διαφόρων χρεογράφων.

3.2.2 Κίνδυνος Αγοράς (Market Risk)

Ο κίνδυνος αγοράς αφορά τον κίνδυνο που μπορεί να προκύψει από απώλειες στα κέρδη του εμπορικού χαρτοφυλακίου της τράπεζας. Τέτοιες απώλειες μπορεί να προκύψουν από μεταβολές στα επιτόκια ή τις τιμές των επενδυτικών τίτλων. Ο κίνδυνος της αγοράς επηρεάζεται από τους παρακάτω παράγοντες:

- Επιτοκιακός Κίνδυνος: Αφορά τη μείωση της τιμής του κέρδους λόγω μεταβολής στα επιτόκια.
- Συναλλαγματικός κίνδυνος: Προκύπτει από τις μεταβολές των ισοτιμιών.
- Κίνδυνος από μεταβολές στις τιμές των εμπορευμάτων.
- Κίνδυνος από μεταβολές στις τιμές των χρηματιστηριακών δεικτών αλλά και των μετοχών.
- Λοιποί κίνδυνοι: Πολιτικές αναταραχές, περιβαλλοντικές καταστροφές και τρομοκρατικές επιθέσεις.

Ο κίνδυνος αγοράς διακρίνεται στις εξής κατηγορίες: Κίνδυνος εταιρικών ομολόγων, Διατραπεζικός κίνδυνος, Κίνδυνος κράτους και Κίνδυνος εμπορικού

χαρτοφυλακίου. Για την αντιμετώπιση του κινδύνου της αγοράς οι τράπεζες εστιάζουν αρχικά στον υπολογισμό και την εκτίμηση του ποσού που κινδυνεύει χρησιμοποιώντας τη μέθοδο Value-At-Risk, που θα αναλυθεί αργότερα [6],[7]. Σε περίπτωση που πρέπει να γίνει διαχείριση του κινδύνου οι τράπεζες καλούνται να ρευστοποιήσουν περιουσιακά στοιχεία και τίτλους ή να αξιοποιήσουν εργαλεία αντιστάθμισης κινδύνου.

3.2.3 Επιτοκιακός Κίνδυνος (Interest Rate Risk)

Ο επιτοκιακός κίνδυνος ορίζεται ως ο κίνδυνος μείωσης των κερδών λόγω των κινήσεων των επιτοκίων. Τα περισσότερα από τα στοιχεία των ισολογισμών των τραπεζών δημιουργούν έσοδα και κόστη που εξαρτώνται από τα επιτόκια. Τα επιτόκια είναι ασταθή κάτι που συνεπάγεται ότι και τα κέρδη είναι ασταθή. Επομένως, όποιος δανείζει ή δανείζεται υπόκειται σε κίνδυνο επιτοκίου [7].

3.2.4 Κίνδυνος Ρευστότητας (Liquidity Risk)

Ο κίνδυνος ρευστότητας ορίζεται ως η αδυναμία μιας τράπεζας να καλύψει τις ανάγκες της σε ρευστότητα [10]. Οι κίνδυνοι ρευστότητας μπορούν να διακριθούν στις εξής κατηγορίες: κίνδυνος ρευστότητας από το ενεργητικό ή το παθητικό, αγοράς, χρηματοδότησης και Κεντρικών Τραπεζών. Ένας από τους κυριότερους λόγους εμφάνισης του είναι η ξαφνική έλλειψη εμπιστοσύνης στο τραπεζικό ίδρυμα, για παράδειγμα μαζικές αναλήψεις καταθέσεων, αλλά και μια ξαφνική ανάγκη του ιδρύματος για ρευστό. Συνήθως ο κίνδυνος ρευστότητας είναι πιο πιθανό να εμφανιστεί σε περιόδους κρίσης.

Σε τέτοιες περιπτώσεις τα τραπεζικά ιδρύματα αναγκάζονται πολλές φορές να πουλήσουν τα εύκολα ρευστοποιήσιμα στοιχεία τους, κάτι όμως που ενέχει επιπλέον κινδύνους. Για τον έλεγχο του κινδύνου ρευστότητας είναι σημαντικό να υπάρχει ένας σωστός σχεδιασμός των αναγκών της τράπεζας σε ρευστό, όπως επίσης να υπάρχει πλάνο εύρεσης νέων πηγών χρηματοδότησης σε περιπτώσεις έκτακτης ανάγκης από ρευστό [9].

3.2.5 Λειτουργικός Κίνδυνος (Operational Risk)

Ο λειτουργικός κίνδυνος εσωκλείει όλους εκείνους τους κινδύνους με τους οποίους έρχεται αντιμέτωπη μια τράπεζα κατά τη λειτουργία της. Συγκεκριμένα, περιλαμβάνει οποιαδήποτε δυσλειτουργία των πληροφοριακών συστημάτων και συστημάτων αναφοράς, αναποτελεσματικότητα των εσωτερικών διαδικασιών και συστημάτων ελέγχου για τη διόρθωση και συμμόρφωση με βάση τους εσωτερικούς κανόνες πολιτικής κινδύνου [7]. Μεγάλη συσχέτιση υπάρχει με την κακή λειτουργία που προκύπτει από τον ανθρώπινο παράγοντα [8]. Ο λειτουργικός κίνδυνος μπορεί να προέλθει από τέσσερα διαφορετικά επίπεδα [7]:

1. Ανθρώπινο επίπεδο: Περιλαμβάνει ανθρώπινα λάθη, μειωμένη εμπειρία, απάτη, όπως και έλλειψη συμμόρφωσης στους κανόνες και τις πολιτικές.
2. Επίπεδο Διαδικασιών: Περιλαμβάνει ελλείψεις στις διαδικασίες και τον έλεγχο παρακολούθησης και λήψης αποφάσεων, λάθη στη διαδικασία καταγραφής των συναλλαγών, οργανωτική ανεπάρκεια και κακή εποπτεία κινδύνων.
3. Τεχνικό επίπεδο: Σχετίζεται με σφάλματα στα μοντέλα, στην εφαρμογή και την απουσία επαρκών εργαλείων για τη μέτρηση των κινδύνων.
4. Τεχνολογικό επίπεδο: Σχετίζεται με ελλείψεις του πληροφοριακού συστήματος και αστοχίες του συστήματος.

Για τη εύρυθμη διαχείριση του λειτουργικού κινδύνου είναι σημαντικό να υπάρχει μια συγκεκριμένη οργανωτική δομή, όπως και οργάνωση των διαδικασιών, να υπάρχει ολοκληρωμένος εσωτερικός έλεγχος και σύστημα παρακολούθησης και φυσικά να γίνεται τακτική εκπαίδευση του προσωπικού.

3.3 Συστήματα Εκτίμησης Πιστωτικού Κινδύνου

Το Σύστημα Εκτίμησης Πιστωτικού Κινδύνου (Credit Rating System) σχετίζεται άμεσα με τις αποφάσεις που λαμβάνει μια τράπεζα σχετικά με το αν θα προχωρήσει στη δανειοδότηση ενός ατόμου ή μιας επιχείρησης λαμβάνοντας υπόψη φυσικά και τον κίνδυνο που καλείται να επωμιστεί αν προχωρήσει σε δανεισμό. Για το σκοπό αυτό χρησιμοποιούνται τόσο ποιοτικά όσο και ποσοτικά συστήματα εκτίμησης της πιστοληπτικής ικανότητας των ατόμων και επιχειρήσεων για την αξιολόγηση του πιστωτικού κινδύνου που αναλαμβάνει η τράπεζα σε περίπτωση που

δε μπορούν να ανταπεξέλθουν στις υποχρεώσεις τους. Συγκεκριμένα, οι προσεγγίσεις που προτείνονται για την εκτίμηση του πιστωτικού κινδύνου είναι οι ακόλουθες [13]:

1. Τυποποιημένη προσέγγιση (standardized approach).
2. Προσέγγιση των εσωτερικών αξιολογήσεων (Internal ratings-based approach, IRB). Αφορά κυρίως τα μοντέλα που χρησιμοποιεί η τράπεζα έτσι ώστε να αξιολογήσει την πιθανότητα αθέτησης εκ μέρους του πιστούχου που έχει αιτηθεί δάνειο. Μετά την αξιολόγηση του πιστούχου γίνεται η κατάταξη του σε μια από τις προκαθορισμένες κατηγορίες που υποδεικνύουν και το βαθμό επικινδυνότητας. Όσο πιο υψηλός είναι ο κίνδυνος αθέτησης τόσο πιο δύσκολο είναι να χορηγηθεί δάνειο στον πιστούχο. Η προσέγγιση των εσωτερικών αξιολογήσεων διακρίνεται σε δυο μεθόδους:
    i. Θεμελιώδης προσέγγιση εσωτερικών διαβαθμίσεων
    ii. Προηγμένη προσέγγιση εσωτερικών διαβαθμίσεων (προέκταση της προηγούμενης)

3.3.1 Παράμετροι Πιστωτικού Κινδύνου

Όπως έχει ήδη αναφερθεί είναι σημαντικό τα χρηματοπιστωτικά ιδρύματα να μπορούν να εκτιμήσουν το βαθμό του πιστωτικού κινδύνου που διατρέχουν. Συγκεκριμένα, πρέπει να μπορεί να υπολογιστεί η Αναμενόμενη Ζημιά (Expected Loss-EL) που μπορεί να επιφέρει μια πίστωση. Η αναμενόμενη ζημιά μπορεί να υπολογιστεί από τον παρακάτω τύπο [14] :

$$EL = PD * LGD * EAD$$

Εν προκειμένω, PD είναι η πιθανότητα αθέτησης, LGD η αναμενόμενη ζημιά και EAD η έκθεση σε περίπτωση αθέτησης. Για παράδειγμα, εάν ένας πιστούχος δανειστεί το ποσό των 1000€, η πιθανότητα αθέτησης του είναι 20% και η ζημιά που θα προκύψει σε περίπτωση αθέτησης είναι 50%, τότε η αναμενόμενη ζημιά ανέρχεται στα 100€.

Στη περίπτωση της Θεμελιώδους Προσέγγισης οι παράμετροι LGD και EAD ποσοτικοποιούνται από την Επιτροπή Τραπεζικής Επιθεώρησης της Βασιλείας, ενώ για την παράμετρο PD η τράπεζα προχωράει στην ποσοτικοποίησή της βασιζόμενη στα χαρτοφυλάκιά της. Όσον αφορά την Προηγμένη Προσέγγιση η αξιολόγηση όλων

των παραπάνω παραμέτρων πραγματοποιείται από την ίδια την τράπεζα αξιοποιώντας στοιχεία που έχει συλλέξει. Αναλυτικότερα, ακολουθούν οι βασικοί παράμετροι του πιστωτικού κινδύνου που χρησιμοποιούνται για τον υπολογισμό της αναμενόμενης ζημιάς.

3.3.1.1 Πιθανότητα αθέτησης (Probability of Default-PD)

Αναφέρεται στη πιθανότητα αθέτησης πληρωμής των υποχρεώσεων του πιστούχου κατά τη διάρκεια μιας συγκεκριμένης περιόδου. Οι πιστούχοι που προβαίνουν σε αθέτηση των υποχρεώσεών τους χαρακτηρίζονται ως «defaults», ενώ αυτοί που είναι τυπικοί ως «non-defaults».

Στην περίπτωση αυτή εξετάζεται μια συγκεκριμένη χρονική περίοδος, συνήθως δώδεκα μηνών, η οποία έχει οριστεί από το εποπτικό πλαίσιο της Βασιλείας [15]. Τα μοντέλα που χρησιμοποιούνται είναι ποιοτικά και ποσοτικά και εκτιμάται η πιθανότητα ο πιστούχος να αθετήσει τις υποχρεώσεις του ή ακόμα και να καθυστερήσει την αποπληρωμή. Η αξιολόγησή του βασίζεται στα διάφορα χαρακτηριστικά του, όπως επίσης και στα στοιχεία που έχει η τράπεζα για προηγούμενες αποπληρωμές του [17].

Για τον υπολογισμό του PD χρησιμοποιούνται διάφοροι μέθοδοι [18]. Η πρώτη μέθοδος βασίζεται στην εμπειρία των πιστωτικών αναλυτών της κάθε τράπεζας. Ένα από τα μειονεκτήματα αυτής της πρακτικής είναι ότι η διαδικασία γίνεται με υποκειμενικά κριτήρια κάτι που μπορεί να οδηγήσει σε λανθασμένη αξιολόγηση των αποτελεσμάτων.

Μια ακόμα μέθοδος είναι η χρήση δεδομένων (ratings) από εξωτερικές πηγές και συγκεκριμένα οίκους αξιολόγησης πιστοληπτικής ικανότητας (Credit Rating Agencies), όπως Moody’s, Standard&Poor και Fitch IBCA. Στην περίπτωση αυτή πρέπει να ληφθεί υπόψη ότι οι διάφοροι οίκοι μπορεί να χρησιμοποιούν διαφορετικά κριτήρια εκτίμησης από αυτά της εκάστοτε τράπεζας και επομένως η πιθανότητα αθέτησης να είναι διαφορετική.

Η τρίτη μέθοδος αξιοποιεί τα μοντέλα αξιολόγησης (Credit Scoring Models). Σε αυτή τη μέθοδο αποδίδεται σε κάθε πιστούχο ένα score, που υποδεικνύει την πιθανότητα να μην ανταπεξέλθει στις υποχρεώσεις του. Εάν μια τράπεζα δε διαθέτει επαρκή στοιχεία για ένα πιστούχο, του οποίου τα χαρακτηριστικά είναι γνωστά, και θέλει να αποφασίσει αν θα προβεί σε δανειοδότηση, τότε επιστρατεύει συνήθως τις δύο πρώτες μεθόδους.

Είναι σημαντικό να σημειωθεί ότι σε περιόδους ύφεσης το PD του πιστούχου αυξάνεται και ταυτόχρονα επιδρά και το LGD, οδηγώντας σε αύξηση του ύψους των κεφαλαιακών απαιτήσεων για τις τράπεζες, εξαιτίας της μειωμένης αξίας εξασφάλισης.

3.3.1.2 Εκτίμηση της αναμενόμενης ζημιάς (Loss Given Default-LGD) ή Κίνδυνος ανάκτησης (Recovery risk)

Είναι το ποσό της αναμενόμενης ζημίας, δηλαδή το ποσοστό του κεφαλαίου που αναμένεται να μην αποπληρωθεί από τον πιστούχο λόγω αθέτησης των υποχρεώσεών του. Η αναμενόμενη ζημιά υπολογίζεται αφαιρώντας το ποσό που μπορεί να ανακτηθεί (Recovery Rate-RR) από το ποσό που κινδυνεύει και επομένως το LGD ορίζεται ως εξής:

$$LGD = 1\text{-}RR$$ [7].

Κάποιοι από τους παράγοντες που το επηρεάζουν είναι η κατάταξη του πιστούχου, το είδος χορήγησης, όπως επίσης και οι επαρκείς εξασφαλίσεις. Για παράδειγμα, εάν υπάρχουν εξασφαλίσεις ή εγγυήσεις όταν υπολογίζεται το LGD, τότε κατά την εκτίμηση πρέπει να ληφθεί υπόψη και η πιθανότητα το πιστωτικό ίδρυμα να μη μπορεί να ανακτήσει ή ρευστοποιήσει άμεσα την εξασφάλιση-εγγύηση [7], [15]. Σε αυτή την περίπτωση το LGD για ένα πιστούχο k μια συγκεκριμένη χρονική στιγμή t υπολογίζεται από τον ακόλουθο τύπο:

$$LGD_k\,(t_{DF}) = EAD_k(t_{DF}) - NPV\,(REC_k(t), t \geq t_{DF}) + NPV\,(Costs_k(t), t \geq t_{DF})/EAD_k(t_{DF})$$

Εν προκειμένω, NVP είναι η παρούσα αξία, REC οι ανακτήσεις, Costs τα άμεσα ή έμμεσα έξοδα και DF ο χρόνος αθέτησης. Ο παραπάνω τύπος κάνει εκτίμηση του Workout LGD και αποτελεί την πιο συνηθισμένη μέθοδο υπολογισμού του LGD που εφαρμόζεται από τα περισσότερα τραπεζικά ιδρύματα. Σε περίπτωση απόκλισης από το πραγματικό LGD επηρεάζονται τόσο η αναμενόμενη ζημιά, όσο και οι προβλέψεις και κατ' επέκταση και οι κεφαλαιακές απαιτήσεις της τράπεζας [16].

Όπως αναφέρθηκε και πιο πάνω το PD και LGD συσχετίζονται μεταξύ τους. Ωστόσο, το LGD επηρεάζει τις κεφαλαιακές απαιτήσεις σε μεγαλύτερο βαθμό από το

PD. Για αυτό το λόγο οι τράπεζες επιλέγουν συνήθως να δανειοδοτήσουν πιστούχους που έχουν εξασφαλίσεις.

3.3.1.3 Έκθεση σε περίπτωση αθέτησης (Exposure at Default-EAD)

Αναφέρεται στο ποσό του κεφαλαίου στο οποίο εκτίθεται η τράπεζα από τη στιγμή που ένα άνοιγμα της υπόκειται σε αθέτηση από τον πιστούχο. Οι τράπεζες χρησιμοποιούν εσωτερικά μοντέλα προεπιλογής διαχείρισης κινδύνου για την εκτίμηση των αντίστοιχων συστημάτων EAD.

Η τιμή της EAD υπολογίζεται συνήθως για κάθε δάνειο και εν συνεχεία οι τράπεζες αξιοποιούν τα συγκεκριμένα στοιχεία για να εκτιμήσουν τον κίνδυνο αθέτησης. Το EAD αποτελεί δυναμικό αριθμό και αυτό γιατί κάθε φορά που ο πιστούχος αποπληρώνει τις υποχρεώσεις του ο αριθμός αυτός αλλάζει. Η τράπεζα βασίζει τα στοιχεία της σε δεδομένα που έχει συλλέξει και σε εσωτερικές αναλύσεις, όπως τα χαρακτηριστικά του δανειολήπτη και τον τύπο προϊόντος.

Προκειμένου να αντιμετωπιστεί η πιστωτική κρίση του 2007-2008, το σύνολο των πιστωτικών ιδρυμάτων αποφάσισε να ακολουθήσει διεθνείς κανονισμούς για να ελαττώσει την έκθεσή του σε αθετήσεις. Στόχος της Επιτροπής Τραπεζικής Εποπτείας της Βασιλείας είναι να βελτιώσει την ικανότητα του τραπεζικού τομέα να αντιμετωπίζει οικονομικές πιέσεις. Με τη βελτίωση της διαχείρισης κινδύνων και τη διαφάνεια των τραπεζών, τα χρηματοπιστωτικά ιδρύματα στοχεύουν στην αποφυγή εμφάνισης του φαινομένου της πτώχευσης ντόμινο.

Όταν η ζημιά που προκαλείται ξεπερνάει την αναμενόμενη ζημιά τότε αναφερόμαστε σε μη-αναμενόμενη ζημιά (Unexpected Loss –UL). Η εκτίμηση ενός τέτοιου μεγέθους, όπως και η πρόβλεψή του είναι σχεδόν αδύνατη. Σε αυτές τις περιπτώσεις τα χρηματοπιστωτικά ιδρύματα βασίζονται είτε σε εποπτικά κεφάλαια είτε σε μεθόδους καθορισμού κάποιων επιπέδων του επιτοκίου. Για το λόγο αυτό θα εισάγουμε παρακάτω την έννοια της Αξίας σε Κίνδυνο (Value at Risk -VaR).

3.3.2 Value at Risk

Όπως έχει ήδη αναφερθεί σε περίπτωση που μια ζημιά ξεπερνάει την αναμενόμενη ζημιά, τότε η τράπεζα έχει να αντιμετωπίσει τη μη αναμενόμενη ζημιά. Αν βέβαια η ζημιά ξεπεράσει και τη μη αναμενόμενη ζημιά, τότε η τράπεζα δεν δύναται να την καλύψει ούτε από τα κέρδη της, αλλά ούτε και από τα κεφάλαιά της.

Η έννοια της Αξίας σε Κίνδυνο (Value at Risk-VaR) αποτελεί ένα δείκτη μέτρησης των πιθανών απωλειών και ορίζει ουσιαστικά ένα όριο στο διάστημα εμπιστοσύνης για την έκθεση στα χαρτοφυλάκια και την κεφαλαιακή επάρκεια.

Το διάστημα εμπιστοσύνης προκύπτει από το πλήρες ποσοστό (100%) μείον τη πιθανότητα να μη καλυφθούν οι ζημιές και αντικατοπτρίζει το κίνδυνο που αναλαμβάνει η τράπεζα. Συνήθως, το εύρος διακύμανσης του διαστήματος εμπιστοσύνης είναι 90-99%. Για τον υπολογισμό του VaR επομένως, πολλαπλασιάζεται το επίπεδο εμπιστοσύνης με την τυπική απόκλιση και την αξία του χαρτοφυλακίου.

Η χρήση του VaR από τις τράπεζες αποτελεί μια σημαντική μέθοδο για τη διαχείριση του πιστωτικού κινδύνου. Για τη μέτρηση του υπάρχουν και διάφορες εναλλακτικές μέθοδοι, όπως της Ιστορικής Προσομοίωσης, της Διακύμανσης-Συνδιακύμανσης και της Προσομοίωσης Monte Carlo. Η πρώτη είναι μη παραμετρική μέθοδος, ενώ οι άλλες δυο παραμετρικές.

Συγκεκριμένα στη μέθοδο της Ιστορικής Προσομοίωσης γίνεται μια κατάταξη των ιστορικών αποδόσεων χρησιμοποιώντας ως κριτήριο την καλύτερη απόδοση. Στη μέθοδο Διακύμανσης-Συνδιακύμανσης η μέτρηση της απόδοσης γίνεται αξιοποιώντας την κανονική κατανομή για αυτό υπολογίζεται η τυπική απόκλιση και η μέση απόδοση. Τέλος, στην Προσομοίωση Monte Carlo γίνεται εκτίμηση των μελλοντικών αποδόσεων μέσα από πολλές υποθετικές δοκιμές.

Για την εκτίμηση του VaR χρησιμοποιούνται συνήθως τα μοντέλα Credit Metrics (από JP Morgan) και Credit+ (από Credit Suisse First Boston). Το υπόδειγμα Credit Metrics αποτελεί μια μέθοδο μέτρησης του πιστωτικού κινδύνου των χαρτοφυλακίων. Βασίζεται κυρίως σε ιστορικά δεδομένα και στη πιθανότητα ένας πιστούχος να αλλάξει επίπεδο πιστωτικού κινδύνου σε μια δεδομένη χρονική περίοδο.

Η μέθοδος διακρίνεται σε τρείς φάσεις με την πρώτη να αποτελεί την αξιολόγηση του προφίλ του πιστούχου με βάση την έκθεση του στον κίνδυνο. Στη δεύτερη φάση εκτιμάται η μεταβολή της αξίας εξαιτίας της έκθεσης στον κίνδυνο και τα αίτια που οδήγησαν σε αυτή, ενώ στη τρίτη φάση γίνεται η συσχέτιση των παραπάνω παραμέτρων.

Από την άλλη το υπόδειγμα Credit+ αποτελεί ένα μοντέλο που δε βασίζεται στα αίτια που προκάλεσαν μια αθέτηση, αλλά μόνο στην εκτίμηση της πιθανότητάς της. Διακρίνεται σε δυο φάσεις, με την πρώτη να περιλαμβάνει τον υπολογισμό της

συχνότητας αθέτησης και την εκτίμηση της σοβαρότητας των ζημιών με σκοπό τον προσδιορισμό της κατανομής ζημιών που γίνεται στη δεύτερη φάση. Ουσιαστικά η μέθοδος αυτή διακρίνει τη πιθανότητα αθέτησης σε δυο καταστάσεις, την αθέτηση και μη αθέτηση, που αντιστοιχούν στην αναμενόμενη και μη αναμενόμενη ζημιά, χωρίς να εξετάζει τα αίτια που οδήγησαν στην κάθε μια.

3.4 Μέθοδοι Εκτίμησης και Περιορισμού Πιστωτικού Κινδύνου

Η αύξηση του αριθμού των πτωχεύσεων τα τελευταία χρόνια, ο ανταγωνισμός μεταξύ των χρηματοπιστωτικών ιδρυμάτων και η ραγδαία εξέλιξη της τεχνολογίας είναι μόνο λίγοι από τους λόγους που έχουν κάνει την εκτίμηση του πιστωτικού κινδύνου επιτακτική [27]. Οι τράπεζες επομένως, καλούνται να αξιολογήσουν το προφίλ κάθε πιθανού πελάτη και να αποφασίζουν αν θα προβούν σε δανεισμό και με τι κίνδυνο.

3.4.1 Credit Scoring

Μια βασική μέθοδος μέτρησης του πιστωτικού κινδύνου που αξιοποιείται από την πλειοψηφία των χρηματοπιστωτικών ιδρυμάτων είναι το credit scoring (μέθοδος πιστωτικής βαθμολόγησης). Η μέθοδος αυτή περιλαμβάνει τα διάφορα εργαλεία και μοντέλα που χρησιμοποιεί η τράπεζα, έτσι ώστε να αξιολογήσει τον πιστωτικό κίνδυνο και να λάβει συγκεκριμένες αποφάσεις σχετικά με την πορεία του δανείου [19].

Η μέθοδος αυτή χρησιμοποιείται κυρίως για την αξιολόγηση ιδιωτών ή μικρών επιχειρήσεων που αιτούνται χαμηλά ποσά δανεισμού. Μικρές επιχειρήσεις θεωρούνται αυτές που έχουν τζίρο κάτω από 2.5 εκατομμύρια ευρώ και η μέθοδος αφορά κυρίως τη χορήγηση καταναλωτικών ή στεγαστικών δανείων. Αυτό συμβαίνει καθώς οι τράπεζες προσπαθούν να αυτοματοποιήσουν τη διαδικασία αξιολόγησης μικρής αξίας δανείων σε αντίθεση με τα υψηλής αξίας δάνεια που παρέχονται σε μεγάλες επιχειρήσεις και απαιτούν εξατομικευμένες αξιολογήσεις από την τράπεζα.

Για τη μέθοδο αυτή αξιοποιούνται υπολογιστικά προγράμματα τα οποία τροφοδοτούνται με στατιστικά και ιστορικά δεδομένα και αξιολογούν τον κάθε πελάτη κατατάσσοντάς τον σε μία από τις κατηγορίες κινδύνου. Στόχος είναι η

αυτοματοποίηση της διαδικασίας, με απώτερο σκοπό την μείωση του χρόνου και κόστους αξιολόγησης.

Τα δεδομένα που χρησιμοποιούνται είναι ποσοτικά και ποιοτικά και αφορούν μεγάλο πλήθος πελατών. Τα δεδομένα πρέπει να καλύπτουν ένα μεγάλο χρονικό διάστημα τουλάχιστον 5 χρόνων για να είναι πιο αξιόπιστα τα αποτελέσματα. Τα ποσοτικά δεδομένα αναπαριστούν χρηματοοικονομικούς δείκτες και μετρούνται χρησιμοποιώντας τον ισολογισμό κάθε επιχείρησης έτσι ώστε να παρουσιάζεται μια συνολική εικόνα των οικονομικών της. Όσον αφορά τα ποιοτικά δεδομένα, αυτά σχετίζονται με τη θέση που καταλαμβάνει μια επιχείρηση στη αγορά, την αξιοπιστία της, τη διοικητική της οργάνωση, όπως επίσης και τις δυνατότητες να αναπτυχθεί.

Όλα αυτά τα δεδομένα και οι δείκτες μπορούν να αξιοποιηθούν με τη χρήση μοντέλων και εργαλείων και να κατατάξουν τον εκάστοτε πελάτη σε default και non-default ανάλογα με το score που λαμβάνει. Επίσης, μπορούν να αξιοποιηθούν για την πρόβλεψη της συμπεριφοράς του πιθανού πελάτη με σκοπό να αποφασισθεί εάν θα του χορηγηθεί δάνειο. Τα μοντέλα που μπορούν να χρησιμοποιηθούν σε τέτοιες περιπτώσεις είναι και τα νευρωνικά δίκτυα. Κάποια από τα υποδείγματα που χρησιμοποιούνται επίσης για τον πιστωτικό κίνδυνο είναι τα ακόλουθα [19]:

- Application Scorecards: Είναι μοντέλα που αφορούν την έγκριση ή όχι ενός δανείου και περιλαμβάνει πληροφορίες για τον πελάτη από τη στιγμή που κάνει αίτηση και αφορούν στοιχεία ως προς την αίτηση και προηγούμενα στοιχεία του πελάτη ως προς την αποπληρωμή άλλων δανείων του [22], [25].
- Behavioral Scorecards: Αφορούν μοντέλα που σχετίζονται με τη συμπεριφορά του πιστούχου και εξετάζουν τον πιστωτικό κίνδυνο βασισμένα σε προηγούμενες αποπληρωμές [20], [21].
- Collections Scorecards: Περιλαμβάνει τα μοντέλα συμπεριφοράς σε συνδυασμό με διαδικασίες συλλογής για να αποφασιστεί εάν είναι απαραίτητο ένα άνοιγμα κλήσης από την τράπεζα [24].
- Bureau Score: Αφορά τα μοντέλα συμπεριφοράς που προκύπτουν από πληροφορίες σχετικά με τα ανοίγματα που έχει κάνει ένας πελάτης. Στην Ελλάδα για παράδειγμα υπάρχει το γνωστό ως Τειρεσίας Score [21], [23].

Αναφέρεται, τέλος, ότι στην περίπτωση των μεγάλων επιχειρήσεων χρησιμοποιείται η μέθοδος Credit Rating (μέθοδος πιστωτικής διαβάθμισης). Στη συγκεκριμένη μέθοδο αξιοποιούνται έμπειροι και εξιδεικευμένοι υπάλληλοι της

τράπεζας, για αυτό το λόγο και η αξιολόγηση είναι πιο υποκειμενική. Αφορά μεγάλες χορηγήσεις δανείων και συνήθως παράγει πιο ακριβή αποτελέσματα μιας και βασίζεται και στην εμπειρία των υπαλλήλων [26].;

3.4.2 Ανασκόπηση μεθόδων που αφορούν το Credit Scoring

Τα τελευταία πενήντα χρόνια έχουν αναπτυχθεί και εξελιχθεί αρκετές μέθοδοι με σκοπό τη μείωση του πιστωτικού κινδύνου. Παρακάτω θα αναλύσουμε μερικές από αυτές.

3.4.2.1 Μοντέλο γραμμικής πιθανότητας

Στο συγκεκριμένο μοντέλο χρησιμοποιούνται ιστορικά στοιχεία που αφορούν τη συμπεριφορά ενός πελάτη ως προς την πιστοληπτική του ικανότητα με σκοπό να αξιοποιηθούν για την αξιολόγηση της συμπεριφοράς καινούργιων πελατών για την αποπληρωμή του δανείου τους. Τα ιστορικά αυτά δεδομένα σε συνδυασμό με την πιθανότητα αθέτησης υποστηρίζεται ότι έχουν μια γραμμική σχέση μεταξύ τους.

3.4.2.2 Μονομεταβλητή ανάλυση (Univariate Analysis)

Σκοπός της συγκεκριμένης μεθόδου ήταν η εύρεση ενός δείκτη που θα μπορούσε να εκτιμήσει όσο το δυνατό περισσότερο την κατάσταση μιας επιχείρησης στο μέλλον. Σε μια μελέτη που έγινε από τον Beaver, η οποία χρησιμοποιούσε δεδομένα για ισάριθμο πλήθος επιχειρήσεων που πτωχεύσανε ή παρέμειναν υγιείς, προέκυψε πως κάποιοι βασικοί δείκτες για την διάκριση των επιχειρήσεων είναι οι ακόλουθοι:

1. Ταμειακή ροή και Συνολικές υποχρεώσεις.
2. Καθαρό εισόδημα και Συνολικά περιουσιακά στοιχεία.
3. Συνολικές υποχρεώσεις και Συνολικά περιουσιακά στοιχεία.

Αν και οι συγκεκριμένοι αριθμοδείκτες αποδείχθηκε ότι δεν είναι αξιόπιστοι για τη διάκριση των επιχειρήσεων σε βιώσιμες και μη, παρατηρήθηκε ότι μπορούν να χρησιμοποιηθούν για την κατάταξη των βιώσιμων επιχειρήσεων μεταξύ τους. Το μοντέλο αυτό ήταν αρκετά απλό, με ελάχιστες γνώσεις στατιστικής για την εφαρμογή του, ενώ οι αριθμοδείκτες εμφάνιζαν γραμμικότητα μεταξύ τους κάτι που υποδείκνυε υψηλή συσχέτιση μεταξύ τους.

3.4.2.3 Μοντέλο Logit (λογαριθμικό)

Το μοντέλο Logit χρησιμοποιείται για τον υπολογισμό του πιστωτικού κινδύνου και βασίζεται στη λογιστική παλινδρόμηση Logit, μιας και η πιθανότητα αθέτησης ακολουθεί λογαριθμική κατανομή. Η τιμή της πιθανότητας αθέτησης κυμαίνεται στο διάστημα [0, 1].

3.4.2.4 Μοντέλο Probit

Το μοντέλο Probit είναι παρόμοιο με το μοντέλο Logit διαφέρουν, ωστόσο, στο γεγονός ότι στο μοντέλο Probit η πιθανότητα αθέτησης ακολουθεί κανονική και όχι λογαριθμική κατανομή. Οι δύο μέθοδοι παρουσιάζουν κοινά πλεονεκτήματα και μειονεκτήματα και η μέθοδος Probit δίνει αποτελέσματα που πλησιάζουν αυτά της Logit, ωστόσο λόγω των απαιτήσεών της για περίπλοκους υπολογισμούς και εξειδικευμένη γνώση δεν αξιοποιήθηκε όσο η Logit [7].

3.4.2.5 Διακριτική ανάλυση (Discriminant Analysis)

Το μοντέλο διακριτικής ανάλυσης χρησιμοποιείται κυρίως για την κατάταξη των δανειοληπτών σε μια κατηγορία κινδύνου. Αποτελεί μια μέθοδο ταξινόμησης (classification) και χρησιμοποιεί διακριτικές συναρτήσεις, συχνά μοντέλα παλινδρόμησης. Σκοπός της χρήσης διακριτικών συναρτήσεων είναι η μείωση της απόστασης των ατόμων που ανήκουν στην ίδια ομάδα και η μεγιστοποίηση της απόστασης σε άτομα που ανήκουν σε διαφορετικές ομάδες. Ο υπολογισμός της απόστασης γίνεται χρησιμοποιώντας το μέσο όρο των μεταβλητών ταξινόμησης.

Η διακριτική ανάλυση μπορεί να είναι μονομεταβλητή, δηλαδή να αξιοποιεί μόνο ένα παράγοντα ταξινόμησης, ή πολυμεταβλητή, όπου χρησιμοποιούνται περισσότεροι παράγοντες. Στην περίπτωση αυτή μπορεί να έχουν όλοι την ίδια διακύμανση ή διαφορετική.

Η πιο συχνή διαχωριστική συνάρτηση που χρησιμοποιείται είναι αυτή της γραμμικής παλινδρόμησης. Η διάκριση συνήθως γίνεται μεταξύ default και non-default πιστούχων, έτσι ώστε να εκτιμηθεί η πιθανότητα αθέτησης τής κάθε κατηγορίας. Ορισμένα πλεονεκτήματα αυτής της μεθόδου αποτελούν επί παραδείγματι η ευκολία στην ερμηνεία των αποτελεσμάτων που προκύπτουν μετά το διαχωρισμό και τα ελάχιστα σφάλματα κατά την εκτίμηση. Στα μειονεκτήματα συγκαταλέγονται η ευαισθησία σε ακραίες τιμές (outliers), τα σφάλματα εκτίμησης

όταν οι ομάδες που προκύπτουν είναι σπάνιες και η έλλειψη της υπόθεσης της γραμμικότητας μεταξύ των μεταβλητών [17].

3.4.2.6 Altman Z-score

Χαρακτηριστικό παράδειγμα μοντέλου διακριτικής ανάλυσης είναι το μοντέλο Z-score. Το μοντέλο αυτό που αναπτύχθηκε αρχικά από τον Altman [28] και χρησιμοποιήθηκε έτσι ώστε να μπορέσει να υπολογιστεί η πιθανότητα μια επιχείρηση να χρεωκοπήσει και κατ' επέκταση να αξιολογηθεί και η πιστοληπτική της ικανότητα. Το μοντέλο υπολογίζει τη τιμή του Z-score βασιζόμενη σε δείκτες για να εκτιμηθεί η πιθανότητα χρεωκοπίας.

Οι δείκτες που χρησιμοποιήθηκαν και αξιολογήθηκαν από τον Altman ως οι πιο σημαντικοί κατατάχθηκαν στις εξής κατηγορίες: Ρευστότητας, Αποδοτικότητας, Μόχλευσης, Φερεγγυότητας και Δραστηριότητας. Η εξίσωση που χρησιμοποιήθηκε είναι η ακόλουθη:

$$Z = 0.012\ X_1 + 0.014\ X_2 + 0.033\ X_3 + 0.006\ X_4 + 0.010\ X_5$$

Αναλυτικότερα, όπου X_1 είναι το κεφάλαιο κίνηση/σύνολο ενεργητικού, X_2 τα παρακρατηθέντα κέρδη/σύνολο ενεργητικού, X_3 τα κέρδη προ φόρων και τόκων/σύνολο ενεργητικού, X_4 η τρέχουσα αξία μετοχών/σύνολο υποχρεώσεων, X_5 οι πωλήσεις/σύνολο ενεργητικού.

Να σημειωθεί ότι οι μεταβλητές που χρησιμοποιήθηκαν ως σημαντικές για τον υπολογισμό του Z-score δεν είχαν την ίδια σημαντικότητα στη μονομεταβλητή ανάλυση. Συγκεκριμένα, ο δείκτης Ταμειακή ροή και Συνολικές υποχρεώσεις δεν αξιοποιήθηκε καθόλου.

Επιχειρήσεις που είχαν Z-score μεγαλύτερο του 2.99 κατατάσσονταν στην ασφαλή ζώνη, ενώ αυτές με Z-score μικρότερο του 1.81 είχαν μεγαλύτερες πιθανότητες να χρεοκοπήσουν άμεσα. Τέλος, όσες ανήκαν στο διάστημα $1.81 < Z\text{-score} < 2.99$ δεν ήταν δυνατό να αξιολογηθούν με ακρίβεια. Προκύπτει, επομένως, ότι όσο πιο χαμηλές είναι οι τιμές του Z-score τόσο πιο μεγάλη είναι η πιθανότητα μια επιχείρηση να χρεωκοπήσει.

Τέλος, αναφέρουμε κάποια μοντέλα που χρησιμοποιούνται για credit scoring και αναλύθηκαν στο προηγούμενο κεφάλαιο. Αυτές είναι τα Decision Trees και Random Forests (μια από τις βασικές machine learning τεχνικές), K-Nearest

neighbors (με bootstrap τεχνικές βελτιώνει την ακρίβεια εκτιμήσεων), Support Vector Machines και, φυσικά, τα νευρωνικά δίκτυα.

4 Μοντελοποίηση Πειραματικού Συστήματος

4.1 Εισαγωγή

Ο πιστωτικός κίνδυνος, όπως αναλύθηκε εκτενέστερα στο κεφάλαιο 3.2 παρουσιάζεται από την αδυναμία εκ μέρους ενός αντισυμβαλλομένου μιας τράπεζας, να ανταποκριθεί στις υποχρεώσεις του, κατά συνέπεια η τράπεζα να στερείται των κερδών που είχε στον σχεδιασμό της από τη σύναψη δανείου μαζί του. Σε αυτό το κεφάλαιο λοιπόν δημιουργήθηκε ένα πειραματικό μοντέλο το οποίο είναι σε θέση να αξιολογήσει την ικανότητα ενός πελάτη τραπεζικού ιδρύματος να πραγματοποιεί τις πληρωμές του εντός του χρονοδιαγράμματος που έχει ορισθεί ανάμεσα στις δύο πλευρές. Βάσει των δεδομένων που αξιοποιήθηκαν από το μοντέλο, οι πληροφορίες που παράγονται αναφέρονται στο Γερμανικό κράτος. Χρησιμοποιήθηκαν δεδομένα που συλλέχθηκαν από τα τραπεζικά ιδρύματα της Γερμανίας και αφορούν συγκεκριμένα χαρακτηριστικά μεγάλης γκάμας πελατών. Μετά από επεξεργασία και αλγοριθμικές ακολουθίες, εκπαιδεύτηκαν ταξινομητές που πραγματοποιούν αυτές τις αξιολογήσεις και μπορούν με σχετική ακρίβεια να βοηθήσουν τους αναλυτές των ιδρυμάτων για το αν ο δανειολήπτης σύμφωνα με τα στοιχεία που έχει τροφοδοτήσει την τράπεζα, θα είναι ικανός ή όχι να εκπληρώσει τις υποχρεώσεις του.

4.2 Προετοιμασία Περιβάλλοντος – Εργαλεία

Για την υλοποίηση της πειραματικής διαδικασίας αξιοποιήθηκαν σύγχρονα μέσα, τα οποία χρησιμοποιούνται αποκλειστικά για τέτοιου είδους διαδικασίες. Αρχικά τα δεδομένα μας τα επεξεργαστήκαμε στην εφαρμογή Weka. Το Weka (ver.3) είναι πλατφόρμα που περιλαμβάνει μία συλλογή από εργαλεία εξόρυξης δεδομένων και αλγορίθμους τεχνητής νοημοσύνης. Επίσης το Weka υποστηρίζει το Deep Learning για πολύπλοκες εργασίες και έχει μία σειρά από οδηγούς και μαθήματα που μπορεί να παρακολουθήσει κάποιος έτσι ώστε να εξοικειωθεί με την τεχνητή νοημοσύνη. Η συγκεκριμένη εφαρμογή - βιβλιοθήκη είναι γραμμένη στην γλώσσα Java. Λόγω του εκπαιδευτικού του χαρακτήρα, το Weka, δεν μπορεί να μας δώσει επαγγελματικές λύσεις και εάν γίνει υπέρμετρη χρήση, κυρίως με μεγάλου όγκου δεδομένα, τότε θα μας επιστραφούν αποτελέσματα τα οποία θα είναι αξιόπιστα, αλλά οι χρόνοι θα είναι υπερβολικά μεγάλοι κάτι που είναι μη αποδεκτό ειδικά αν

πρόκειται για αποφάσεις που θα πρέπει να πάρουν οι αναλυτές σε ένα απαιτητικό περιβάλλον εργασίας.

Για αυτό το λόγο στην δεύτερη φάση, μετά την πρώτη ανάλυση των δεδομένων χρησιμοποιήσαμε την πλέον κατάλληλη γλώσσα προγραμματισμού για εργασίες ταξινόμησης και πρόβλεψης την Python. Η Python είναι μία γλώσσα γενικού σκοπού κατάλληλη για τις περισσότερες εργασίες, βασικό της πλεονέκτημα είναι ότι σε σχέση με άλλες σύγχρονες γλώσσες προγραμματισμού μπορεί να εκτελέσει βασικές εντολές με τον ελάχιστο αριθμό γραμμών κώδικα. Αυτό την κάνει ιδιαίτερα δημοφιλή και φιλική προς τον χρήστη. Βρίσκεται πολύ υψηλά στις προτιμήσεις των επιστημόνων των κλάδων των μαθηματικών, της στατιστικής, των οικονομικών και άλλων παρεμφερών κλάδων, όπου οι γνώσεις προγραμματισμού δεν είναι τόσο δυνατές. Ωστόσο με την χρήση των κατάλληλων βιβλιοθηκών που βρίσκονται στην python, τα αποτελέσματα είναι παραπάνω από ικανοποιητικά. Για τη πειραματική διαδικασία που ακολουθήθηκε χρησιμοποιήθηκαν από την Python, αρχικά η βιβλιοθήκη Pandas. Πρόκειται για μία βιβλιοθήκη λογισμικού που αναλαμβάνει με τις κατάλληλες γραμμές κώδικα από τον χρήστη, τον χειρισμό και την ανάλυση δεδομένων. Προσφέρει πιο συγκεκριμένα δομές δεδομένων και λειτουργίες για χειρισμό αριθμητικών πινάκων και χρονοσειρών. Ακόμη πρόκειται για ελεύθερο λογισμικό ο επιστήμονας που την ξεκίνησε είναι ο Wes McKinney και το όνομα της βιβλιοθήκης προέρχεται από ένα λογοπαίγνιο από την φράση «Python Data Analysis». Κάποιες από τις λειτουργίες της είναι η διαχείριση των δεδομένων και αντιμετώπιση των μηδενικών στοιχείων της αποθήκης, η αναδιαμόρφωση και η περιστροφή των συνόλων, οι εισαγωγές ή διαγραφές στηλών κ.α.. Το επόμενο σημαντικό εργαλείο, είναι το Numpy. Η συγκεκριμένη βιβλιοθήκη βοηθάει στην διαχείριση μεγάλων πινάκων και συστοιχιών και περιλαμβάνει μία πολύ μεγάλη συλλογή μαθηματικών συναρτήσεων. Συγκεντρώνει την υπολογιστική ισχύ γλωσσών όπως η C και Fortran στην python. Τέλος χρησιμοποιήθηκαν για την ανάπτυξη και παρουσίαση των αποτελεσμάτων σε κείμενο και με την μορφή γραφικών παραστάσεων άλλες δύο βιβλιοθήκες η Pyplot και η Seaborn.

4.3 Βάση Δεδομένων – Dataset

Τα δεδομένα που θα αξιοποιήσουμε για την διαδικασία των πειραμάτων είναι ένα προϊόν επεξεργασίας και διεργασιών. Πρόκειται για ένα αρχείο αποτελούμενο από χίλιες(1000) εγγραφές. Αυτές οι εγγραφές αντιστοιχούν σε χίλιους(1000) πελάτες των γερμανικών τραπεζών. Την αρχική μελέτη και το σύνολο των δεδομένων έχουν συλλεχθεί από τον Δρ Hans Hoffman του τμήματος Στατιστικής και Οικονομετρίας (Πανεπιστήμιο Αμβούργου). Ο Δρ Hoffman στο dataset που δημιούργησε έχει εστιάσει σε είκοσι(20) ξεχωριστά χαρακτηριστικά τα οποία είναι ικανά να δώσουν ένα προφίλ στον κάθε αιτούντα δανείου και να βοηθήσουν στην ανάλυση και στην τελική πρόβλεψη σχετικά με την ικανότητα αποπληρωμής του δανείου που αιτήθηκε. Το συγκεκριμένο dataset έχει χρησιμοποιηθεί από πολλούς αναλυτές για την εξαγωγή συμπερασμάτων σχετικά με το Machine Learning στον τραπεζικό τομέα.

Για την παρούσα πειραματική διαδικασία αποκτήσαμε την εν λόγω βάση από την ηλεκτρονική πλατφόρμα Kaggle και συγκεκριμένα από τον σύνδεσμο:
https://www.kaggle.com/uciml/german-credit

Στο συγκεκριμένο σύνδεσμο εντοπίσαμε μία τροποποιημένη έκδοση του αρχικού dataset του Δρ Hoffmann, όπου έχουν πραγματοποιηθεί κάποιες αλλαγές. Οι αλλαγές πραγματοποιήθηκαν από ερευνητές, αρχικά το αρχείο έχει μετατραπεί στην μορφή .CSV, ο συγκεκριμένος τύπος είναι αποδεκτός και αρκετά δημοφιλής από όλες τις εφαρμογές που είναι σχετικές με αλγόριθμους ταξινόμησης και προβλέψεων. Επιπλέον, λόγω της μεγάλης πολυπλοκότητας των συμβόλων και των κατηγοριών, αλλά και της ασάφειας ή της μικρής σημασίας κάποιων στηλών, έχουν αφαιρεθεί κάποια στοιχεία. Συνεπώς για την δικιά μας ανάλυση θα χρησιμοποιηθούν τα εξής στοιχεία:

- Πληροφορίες σχετικά με τα χαρακτηριστικά του δανείου (ποσό δανείου, διάρκεια δανείου, σκοπός δανείου)
- Πληροφορίες σχετικά με τον πελάτη την στιγμή που αιτήθηκε το δάνειο (κατάσταση αποταμιεύσεων, τρεχούμενος λογαριασμός τραπέζης, στέγαση)
- Τέλος, δημογραφικές πληροφορίες πελατών (ηλικία, φύλλο, ικανότητα εργασίας)

Πιο αναλυτικά τα πεδία είναι τα εξής:

Τίτλος Στήλης	**Τύπος Στοιχείων**	**Τιμές Δεδομένων**
Ηλικία (Age)	Αριθμητικό	
Φύλλο (Sex)	Κείμενο	Male, Female
Εργασία (Job)	Αριθμητικό	0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled
Διαμονή – Οικία (Housing)	Κείμενο	Own, Rent, Free
Λογαριασμός Αποταμιεύσεων (Savings Account)	Κείμενο	Little, Moderate, Quite Rich, Rich
Λογαριασμός Τρεχούμενος (Checking account)	Αριθμητικό	Τιμή σε Γερμανικά Μάρκα
Ποσό πίστωσης (Credit Amount)	Αριθμητικό	Τιμή σε Γερμανικά Μάρκα
Διάρκεια πίστωσης (Duration)	Αριθμητικό	Τιμή σε μήνες
Σκοπός πίστωσης (Purpose)	Κείμενο	car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others
Risk	Κείμενο	Good, Bad

4.4 Προετοιμασία δεδομένων – Μετασχηματισμοί – Επεξεργασία.

Η διαδικασία της ανάλυσης πραγματοποιήθηκε εξ ολοκλήρου στην Python. Σε αυτό το βήμα έγινε χρήση εντολών σύμφωνα με την ορθή χρήση των βιβλιοθηκών που έχουν αναφερθεί παραπάνω. Τα βήματα που ακολουθήθηκαν είναι τα εξής:

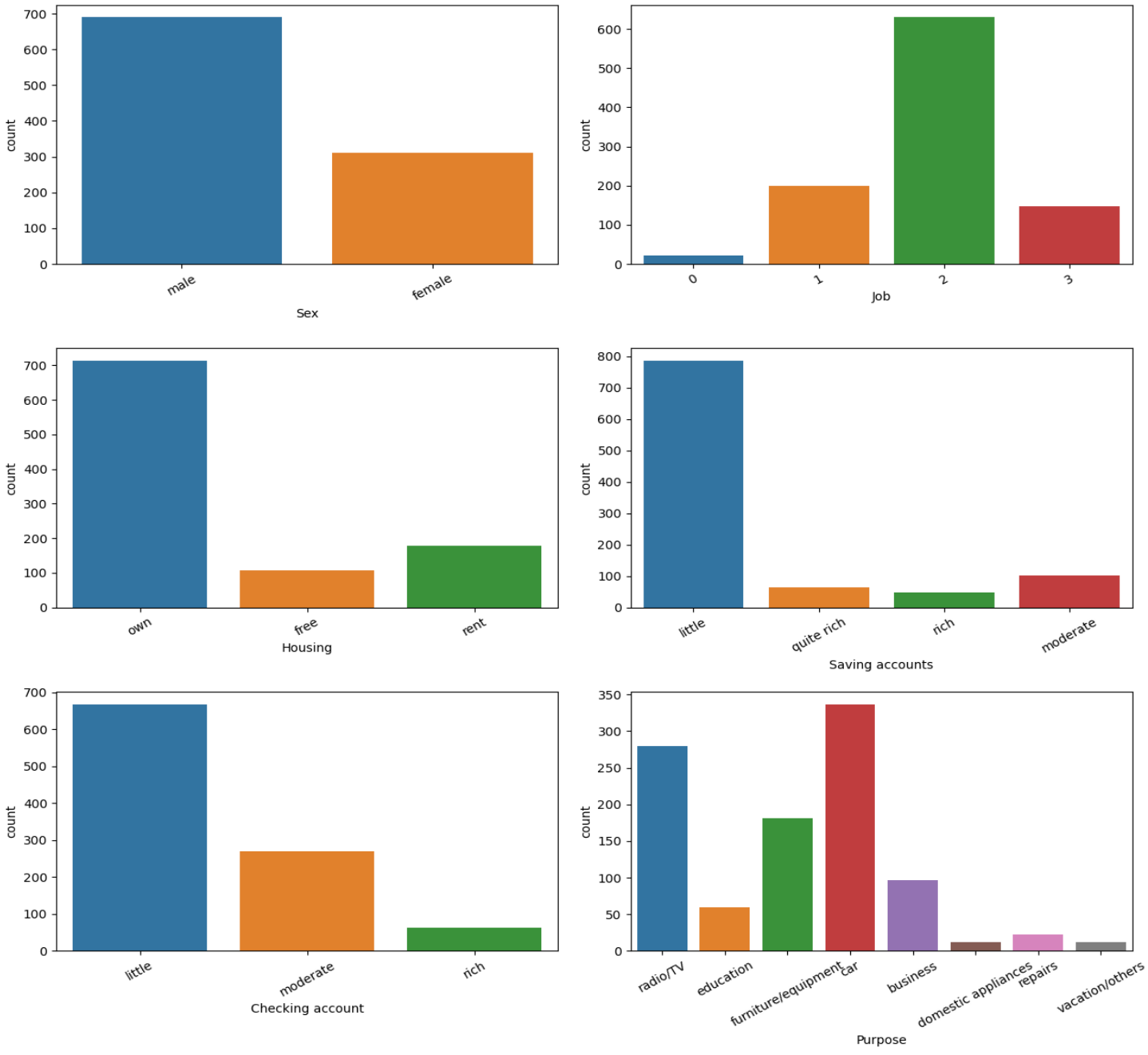
- Φόρτωση αρχείου με την χρήση της βιβλιοθήκης Pandas
- Διερευνητική ανάλυση, τι είδους πληροφορίες περιέχει το αρχείο, κατανομές τιμών των πεδίων, εντοπισμός κλασσικών προβλημάτων όπως missing values, outliers κ.τ.λ.
- Μετασχηματισμός μεταβλητών για την δημιουργία συσχετίσεων
- Μετατροπή τιμών σε ψευδομεταβλητές (dummy variables) όπου χρειάζεται

Οι εντολές που χρησιμοποιήσαμε οδηγούν σε τέσσερις συναρτήσεις της βιβλιοθήκης panda για να πάρουμε τα πρώτα συμπεράσματα. Χρησιμοποιήθηκαν λοιπόν κατά σειρά οι εντολές: **shape()**, **head(), isnull().sum()**, **info()** και **describe()**. Εφόσον αφαιρέσαμε την πρώτη στήλη που ήταν η αρίθμηση (indexing),

παρατηρήθηκαν τα παρακάτω και το αποτέλεσμα μας έδωσε μία πρώτη εικόνα του αρχείου μας.

	NA	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
0	0	67	male	2	own	NaN	little	1169	6	radio/TV	good
1	1	22	female	2	own	little	moderate	5951	48	radio/TV	bad
2	2	49	male	1	own	little	NaN	2096	12	education	good
3	3	45	male	2	free	little	little	7882	42	furniture/equipment	good
4	4	53	male	2	free	little	little	4870	24	car	bad

Εικόνα 1 Χρήση της εντολής head(), από την βιβλιοθήκη pandas. Μας επιστρέφει της πρώτες 5 γραμμές του αρχείου μας.



Εικόνα 2 Ιστογράμματα μπαρών για τα στοιχεία που έχουν αλφαριθμητικές τιμές.

Από τα παραπάνω γραφήματα προκύπτουν τα εξής στοιχεία:

- Οι άνδρες είναι περισσότερες από τις γυναίκες
- Οι περισσότεροι από τους αιτούντες έχουν μία καλή και εξιδεικευμένη στο αντικείμενό τους δουλειά (**skilled**)
- Το μεγαλύτερο ποσοστό των πελατών έχει δικό του σπίτι
- Παρατηρείται στους περισσότερους χαμηλό ύψος καταθέσεων
- Οι πιο πολλοί χρειάζονται τα τραπεζικά προϊόντα για να αγοράσουν αυτοκίνητο και έπειτα ράδιο ή τηλεόραση

```
NA                    0
Age                   0
Sex                   0
Job                   0
Housing               0
Saving accounts     183
Checking account    394
Credit amount         0
Duration              0
Purpose               0
Risk                  0
dtype: int64
```

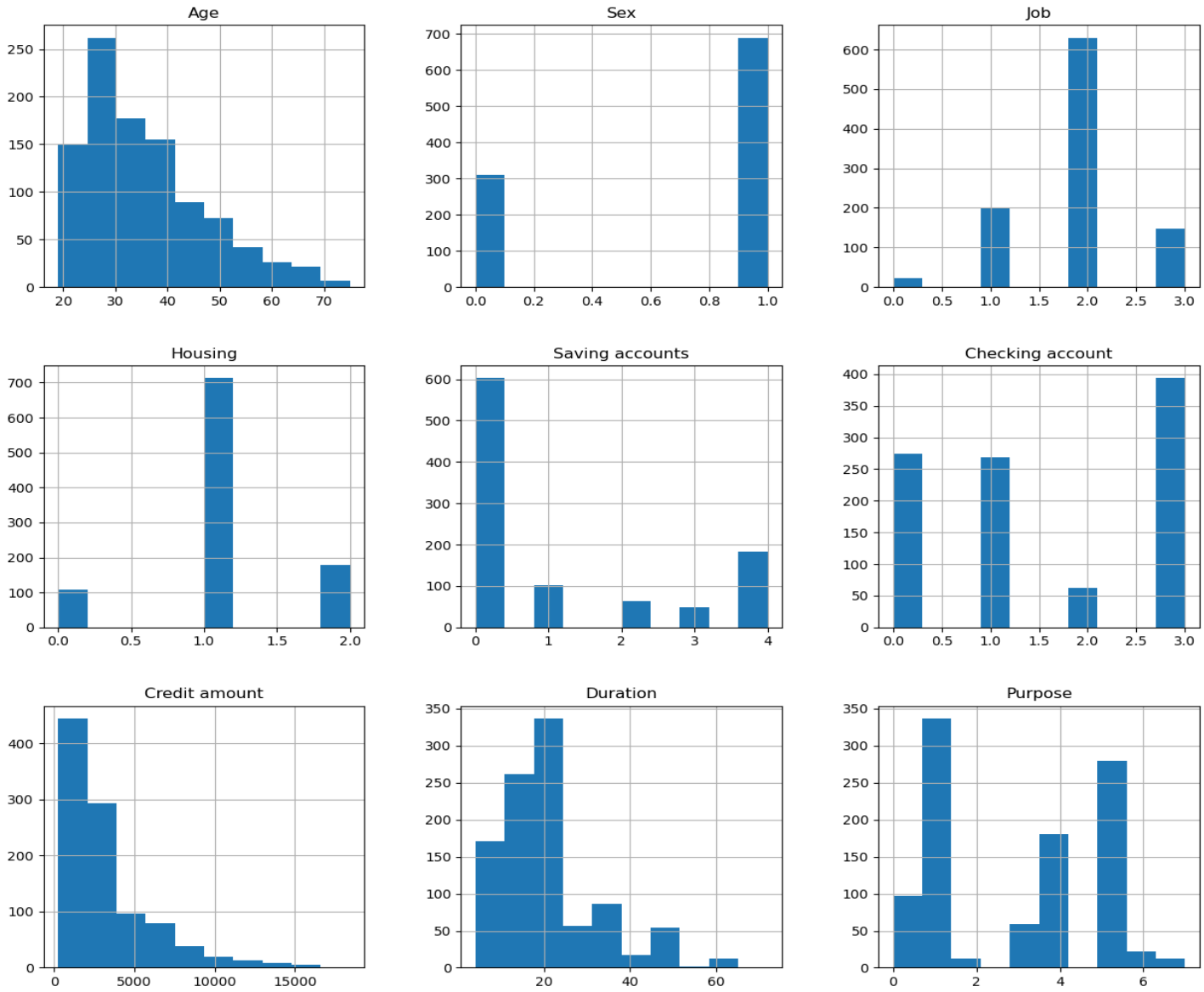
Εικόνα 3 Χρήση της εντολής isnull().sum(), από την βιβλιοθήκη pandas. Μας επιστρέφει το άθροισμα των μηδενικών τιμών που παρατηρείται σε κάθε στήλη του αρχείου μας.

Όσον αφορά τις μετατροπές χρειάστηκε απαλοιφή των μηδενικών στοιχείων στα χαρακτηριστικά *Saving accounts & Checking account* και στην συνέχεια της ανάλυσης προχωρήσαμε στον μετασχηματισμών όλων των αλφαριθμητικών τιμών σε αριθμητικές. Αυτή η ενέργεια μας βελτίωσε την απόδοση του μοντέλου. Η εντολή που χρησιμοποιήθηκε ήταν από την βιβλιοθήκη **sklearn** και το τμήμα preprocessing που έχει δημιουργηθεί αποκλειστικά για την επεξεργασία δεδομένων πριν τις όποιες ενέργειες σχετικές με το machine learning.

	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
0	67	1	2	1	4	0	1169	6	5	1
1	22	0	2	1	0	1	5951	48	5	0
2	49	1	1	1	0	3	2096	12	3	1
3	45	1	2	0	0	0	7882	42	4	1
4	53	1	2	0	0	0	4870	24	1	0

Εικόνα 4 Εντολή head() και εμφάνιση πρώτων πέντε(5) γραμμών του αρχείου έχοντας κάνει μετασχηματισμό με την συνάρτηση LabelEncoder().

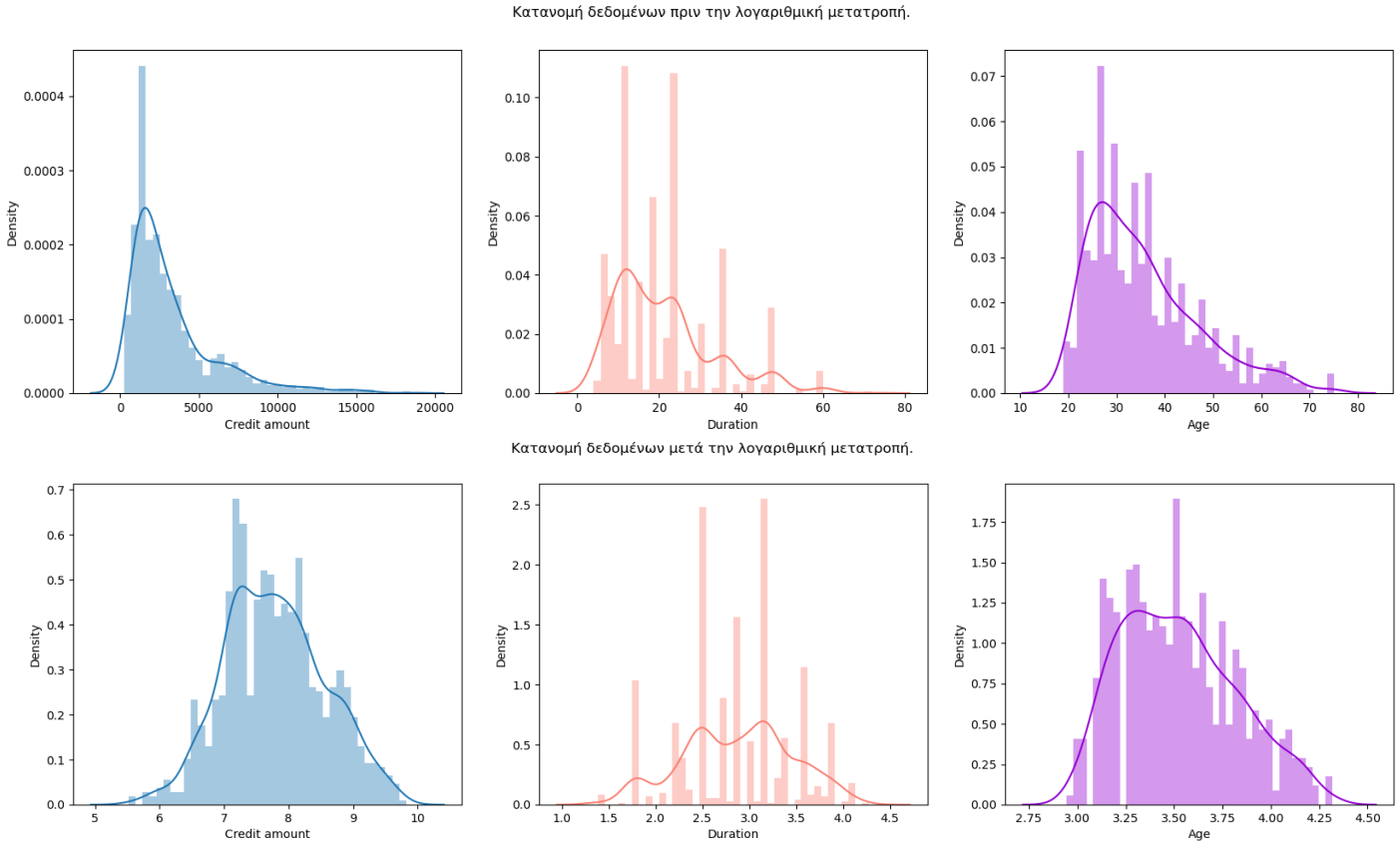
Προχωρώντας με τις αναλύσεις του αρχείου, χρησιμοποιήθηκε η εντολή **hist()**, η οποία δημιούργησε πολλαπλά ιστογράμματα μπαρών και μας ενημερώνουν σχετικά με το κάθε χαρακτηριστικό και την κατανομή του.



Εικόνα 5 Εντολή hist(). Ιστογράμματα με κάθε χαρακτηριστικό του αρχείο και τι περιλαμβάνει.

Για τα αρχικά αριθμητικά χαρακτηριστικά του αρχείου (*Credit Amount*, *Age*, *Duration*) πραγματοποιήθηκαν δοκιμές και με λογαριθμική μετατροπή, οπού κατά περίπτωση επιστρέφει μεγαλύτερη απόδοση στο μοντέλο και καλύτερη ευστάθεια κατά την διάρκεια της διαδικασίας εκμάθησης, έχοντας ως βασικό στόχο την μείωση τον ακραίων τιμών και την διόρθωση της λοξότητας της κατανομής των δεδομένων. Επίσης κάποιοι ερευνητές έχουν επισημάνει και βελτίωση στο accuracy του μοντέλου

αρκετών ποσοστιαίων μονάδων [29]. Συνεπώς εισάγοντας τα παραπάνω αριθμητικά στοιχεία στην συνάρτηση **log()** της βιβλιοθήκης **Numpy** που αναφέρθηκε παραπάνω πήραμε τα σχετικά αποτελέσματα.



Εικόνα 6 Κατανομή αριθμητικών δεδομένων μετά την επεξεργασία.

Μία από τις προεπιλεγμένες μεθόδους επίλυσης καταστάσεων όπου ένα χαρακτηριστικό έχει πολύ μεγαλύτερη διακύμανση από άλλα είναι η χρήση κλιμάκωσης ή τυποποίησης (scaling) [30]. Σύμφωνα με αυτό και για να παραχθούν όσο το δυνατόν καλύτερες συστάδες(clusters) με τους αλγόριθμους που χρησιμοποιήθηκαν στο μοντέλο πραγματοποιήθηκαν δοκιμές επιλέγοντας την χρήση της συνάρτησης **StandardScaler()** που ανήκει στην βιβλιοθήκη **sklearn** και στο ειδικό τμήμα της προ επεξεργασίας των δεδομένων. Ο κύριος στόχος της συνάρτησης είναι να μετασχηματίσει τα αριθμητικά χαρακτηριστικά του αρχείου και να φέρει τις τιμές τους όσο πιο κοντά στους άξονες γίνεται, έτσι ώστε να υπάρξει ομοιομορφία σε όλα τα δεδομένα και να εκπαιδευτεί σωστότερα το μοντέλο μας. Τα αποτελέσματα είναι στην εικόνα 7.

	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose
285	35	0	1	1	0	0	10722	47	1
881	48	1	2	0	4	3	9277	24	1
78	39	1	1	1	4	3	9436	54	1
333	24	0	1	2	1	3	11590	48	1
268	45	1	3	1	0	0	8978	14	1

	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose
285	-0.119403	-2.0	-1.816125	0.342914	-0.841997	-1.353049	-0.461326	0.579834	-0.602626
881	1.038985	0.5	-0.460808	-1.215785	1.433671	1.216032	-1.092328	-1.129932	-0.602626
78	0.237024	0.5	-1.816125	0.342914	1.433671	1.216032	-1.022896	1.100197	-0.602626
333	-1.099577	-2.0	-1.816125	1.901613	-0.273080	1.216032	-0.082288	0.654171	-0.602626
268	0.771664	0.5	0.894509	0.342914	-0.841997	-1.353049	-1.222895	-1.873308	-0.602626

Εικόνα 7 Τα δεδομένα πριν και μετά την χρήση της συνάρτησης StandardScaler()

4.5 Υλοποίηση μοντέλου προβλέψεων

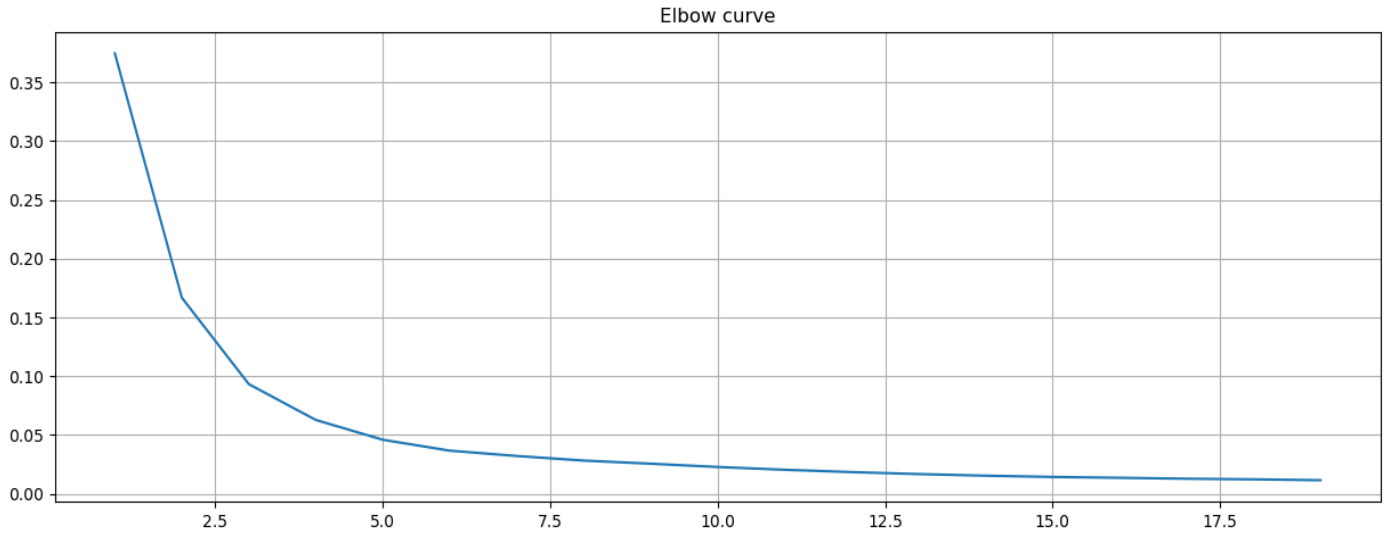
Το πρώτο βήμα στην πειραματική διαδικασία, εφόσον είχε ολοκληρωθεί η διαδικασία της επεξεργασίας των δεδομένων ήταν η δημιουργία συστάδων**(Clustering).** Η python μας δίνει την δυνατότητα να διασπάσουμε το σύνολο των δεδομένων μας σε δύο μέρη. Για το μοντέλο μας χρησιμοποιήθηκε η συνάρτηση **train_test_split()** που ανήκει στην βιβλιοθήκη **sklearn** και διαχώρισε τα δεδομένα για αρχή σε 900 επαφές προς εκπαίδευση και 100 για την φάση της εκτίμησης, αυτό ολοκληρώνεται ορίζοντας την παράμετρο **test_size** στην τιμή 0.1. Έπειτα πραγματοποιήθηκαν μελέτες και δοκιμές με τις δημοφιλέστερες μεθόδους συσταδοποίησης και σύμφωνα με τα [33][34][35] επιλέχθηκαν:

- **K-Means clustering algorithm**
- **Hierarchical clustering analysis** (Agglomerative)
- **DBSCAN** (Density-based spatial clustering of applications with noise)

Επικράτησε για την παρούσα έρευνα ο αλγόριθμος **K-Means**, ένας από τους απλούστερους αλγόριθμους μηχανικής εκμάθησης χωρίς επίβλεψη. Το **K-Means** ψάχνει έναν αριθμό σταθερό (**k**) από συμπλέγματα σε ένα σύνολο δεδομένων. Σύμπλεγμα ορίζουμε μία συλλογή σημείων δεδομένων που έχουν την τάση να συγκεντρώνονται μαζί λόγω κάποιων ομοιοτήτων. Αυτή είναι και στην βάση της η κεντρική μας ιδέα να ομαδοποιήσουμε τους πελάτες του αρχείου μας και να προσπαθήσουμε να ανακαλύψουμε υποκείμενα μοτίβα που θα μας οδηγήσουν σε συμπεράσματα σχετικά με την φερεγγυότητα τους προς το τραπεζικό ίδρυμα. Η πιο διάσημη μέθοδος για την επιλογή του αριθμού **k** είναι η μέθοδος **Elbow**[36].

Πρόκειται για μία ευρετική μέθοδο που χρησιμοποιείται για τον προσδιορισμό των συστάδων. Η ίδια μέθοδος μπορεί να χρησιμοποιηθεί και σε άλλα μοντέλα που βασίζονται σε δεδομένα με σκοπό την επιλογή του αριθμού των παραμέτρων. Στη μέθοδο **Elbow**, για κάθε τιμή του k, υπολογίζουμε το WCSS (In-Cluster Sum of Square). Το WCSS είναι το άθροισμα της τετραγωνικής απόστασης μεταξύ κάθε σημείου και του κέντρου σε ένα σύμπλεγμα. Όταν σχεδιάζουμε το WCSS με την τιμή k, το διάγραμμα μοιάζει με αγκώνα. Καθώς ο αριθμός των συμπλεγμάτων αυξάνεται, η τιμή WCSS θα αρχίσει να μειώνεται. Η τιμή WCSS είναι μεγαλύτερη όταν k = 1. Όταν αναλύουμε το γράφημα μπορούμε να δούμε ότι το γράφημα θα αλλάξει γρήγορα σε ένα σημείο και έτσι θα δημιουργηθεί ένα σχήμα αγκώνα. Από αυτό το σημείο, το γράφημα αρχίζει να κινείται σχεδόν παράλληλα με τον άξονα X. Η τιμή k που αντιστοιχεί σε αυτό το σημείο είναι η βέλτιστη τιμή k ή ένας βέλτιστος αριθμός συστάδων. Η γραφική παράστασης της υλοποίησης μεθόδου Elbow φαίνεται στην επόμενη εικόνα.

Έχοντας χρησιμοποιήσει την μέθοδο κρατήσαμε τους αριθμούς: **3,4 και 5** ως τους



Εικόνα 8 Υλοποίηση μεθόδου Elbow για την επιλογή του αριθμού K.

αριθμούς που θα είναι και οι πιθανές τιμές του **k** και θα μας δώσουν τα καλύτερα αποτελέσματα στην συσταδοποίηση.

Στην συνέχεια του κώδικα, για να ενισχύσουμε το μοντέλο μας και να αποκτήσει μεγαλύτερη ακρίβεια, για κάθε συστάδα που μας επέστρεψε ο αλγόριθμος K-Means δημιουργήσαμε με παρόμοια διαδικασία υποσυστάδες(**sub-clusters**). Χρησιμοποιήσαμε τις εγγραφές που εντάχθηκαν σε κάθε αρχική συστάδα,

εκτελέσαμε πάλι την μέθοδο Elbow και πήραμε τους νέους υποψήφιους αριθμούς k στην συνέχεια εκτελέστηκε σε επανάληψη ο αλγόριθμος K-Means μέχρι να μας δώσει όλες τις νέες υποσυστάδες. Οι τελικοί αριθμοί k επιλέχθηκαν σε συνάρτηση με το τελικό αποτέλεσμα των προβλέψεων του μοντέλου, έχοντας εκτελεστεί αρκετές δοκιμές, επειδή η μέθοδος Elbow δεν ήταν κατατοπιστική.

Μετά από κάθε εκτέλεση του K-Means χρησιμοποιούσαμε την συνάρτηση **predict()** στην οποία δίνουμε ως όρισμα ολόκληρο το σύνολο των εγγραφών που προορίζεται για την εκτίμηση, δηλαδή τις 100 εγγραφές. Η συγκεκριμένη συνάρτηση κατηγοριοποιεί τις επαφές αυτές στις συστάδες που έχουν παραχθεί σε προηγούμενη φάση κάνοντας χρήση του εκπαιδευτικού συνόλου δεδομένων. Στην πράξη αυτή η διαδικασία πραγματοποιείται εφόσον τρέξει ο αλγόριθμός K-Means, δημιουργούνται τα κέντρα(**centroid**) κάθε συστάδας και μετά με το predict υπολογίζεται η ευκλείδεια απόσταση των δεδομένων εκτίμησης από τα κέντρα των συστάδων.

Για την ολοκλήρωση του μοντέλου και την δημιουργία του μηχανισμού προβλέψεων, ορίστηκε ότι για κάθε υποσυστάδα το προβλεπόμενο ρίσκο μιας επαφής είναι **1 (Good),** αν η πλειοψηφία του σύνολού των εγγραφών που απαρτίζουν την υποσυστάδα, από τα δεδομένα της εκπαίδευσης, έχει **Risk = 1,** ενώ αν η πλειοψηφία έχει ως **0(Bad)** το Risk στο σύνολο της τότε όποιο από τα δεδομένα εκτίμησης ταξινομείται με την συγκεκριμένη συστάδα θα παίρνει **Risk = 0** αυτόματα.

4.6 Πειράματα – Αποτελέσματα

Σε αυτό το μέρος παρουσιάζονται οι δοκιμές οι οποίες εκτελέστηκαν και οι διάφορες επιλογές που έγιναν στις παραμέτρους των συναρτήσεων έτσι ώστε να επιτευχθεί η καλύτερη απόδοση στο μοντέλο μας και να εξάγουμε τα τελικά συμπεράσματα μας.

Δοκιμή 1η:

- Αριθμός εγγραφών εκπαίδευσης **900**, Αριθμός εγγραφών εκτίμησης **100 (Test-size:0.1)**
- Ανάμειξη επαφών: Ναι (**Shuffle: True**)
- Καμία αφαίρεση χαρακτηριστικών του αρχείου (**Features: 9**)
- Αριθμός συστάδων: 3 **(k=3)**
- Αριθμός υποσυστάδων: 3 (**k=3**)
- Μετά την πρώτη συσταδοποίηση πραγματοποιήθηκε μετασχηματισμός των αριθμητικών δεδομένων με την συνάρτηση StandardScaler()

Αποτελέσματα:

Accuracy: 81.00

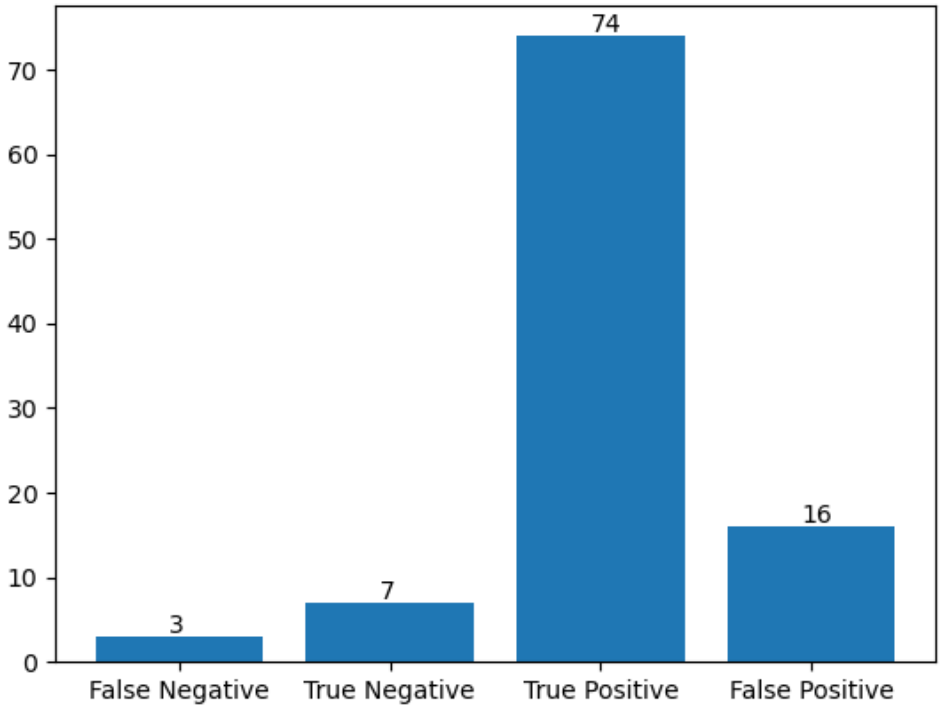
Recall: 96.10

Precision: 82.22

F1 Score: 88.62

	precision	recall	f1-score	support
0	0.70	0.30	0.42	23
1	0.82	0.96	0.89	77
accuracy			0.81	100
macro avg	0.76	0.63	0.66	100
weighted avg	0.79	0.81	0.78	100

Εικόνα 9 Αποτελέσματα από το Classification Report της sklearn. (1η Δοκιμή)



Εικόνα 10 Ιστόγραμμα με τον αριθμό των πελατών και τις σωστές και λάθος προβλέψεις. (1η Δοκιμή)

Δοκιμή 2η:

- Αριθμός εγγραφών εκπαίδευσης **900**, Αριθμός εγγραφών εκτίμησης **100 (Test-size:0.1)**
- Ανάμειξη επαφών: Ναι (**Shuffle: True**)
- Καμία αφαίρεση χαρακτηριστικών του αρχείου (**Features: 9**)
- Αριθμός συστάδων: 3 **(k=3)**
- Αριθμός υποσυστάδων: 3 (**k=3**)
- Τα αρχικά αριθμητικά χαρακτηριστικά - *Credit Amount, Age, Duration* – έχουν υποστεί λογαριθμική μετατροπή

Αποτελέσματα:

Accuracy: 80.00

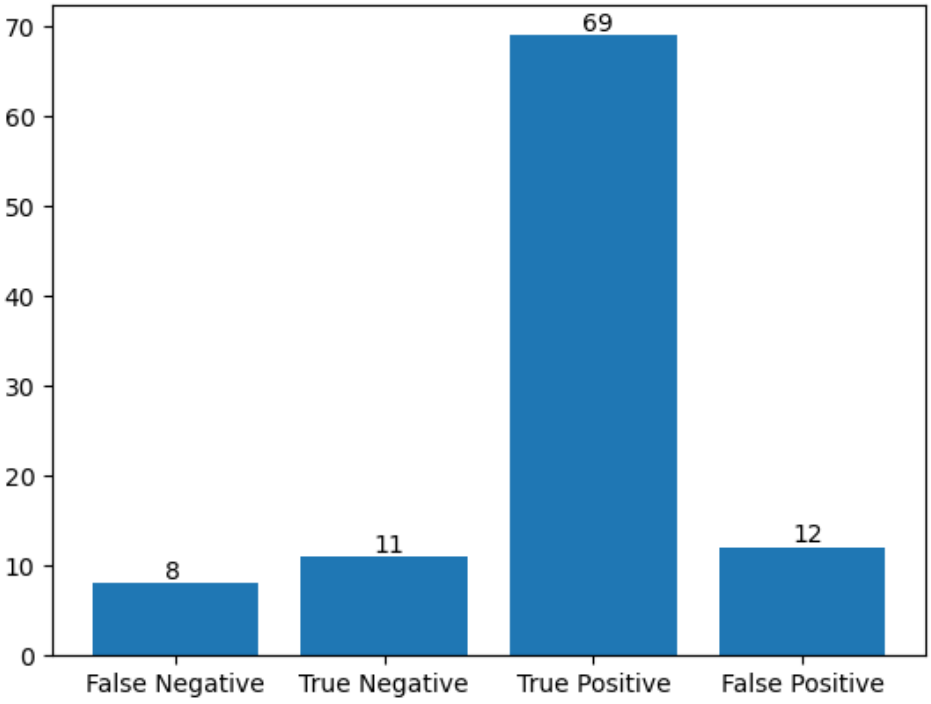
Recall: 89.61

Precision: 85.18

F1 Score: 87.34

	precision	recall	f1-score	support
0	0.58	0.48	0.52	23
1	0.85	0.90	0.87	77
accuracy			0.80	100
macro avg	0.72	0.69	0.70	100
weighted avg	0.79	0.80	0.79	100

Εικόνα 11 Αποτελέσματα από το Classification Report της sklearn. (2η Δοκιμή)



Εικόνα 12 Ιστόγραμμα με τον αριθμό των πελατών και τις σωστές και λάθος προβλέψεις. (2η Δοκιμή)

Δοκιμή 3η:

- Αριθμός εγγραφών εκπαίδευσης **900**, Αριθμός εγγραφών εκτίμησης **100 (Test-size:0.1)**
- Ανάμειξη επαφών: Ναι (**Shuffle: True**)
- Αφαίρεση όλων των αρχικών αλφαριθμητικών χαρακτηριστικών, εκτός του *Job*. Παραμένουν τα *Credit Amount, Age, Duration, Job* (**Features: 4**)
- Αριθμός συστάδων: 4 **(k=4)**
- Αριθμός υποσυστάδων: 3 (**k=3**)
- Μετά την πρώτη συσταδοποίηση πραγματοποιήθηκε μετασχηματισμός των αριθμητικών δεδομένων με την συνάρτηση StandardScaler()

Αποτελέσματα:

Accuracy: 80.00

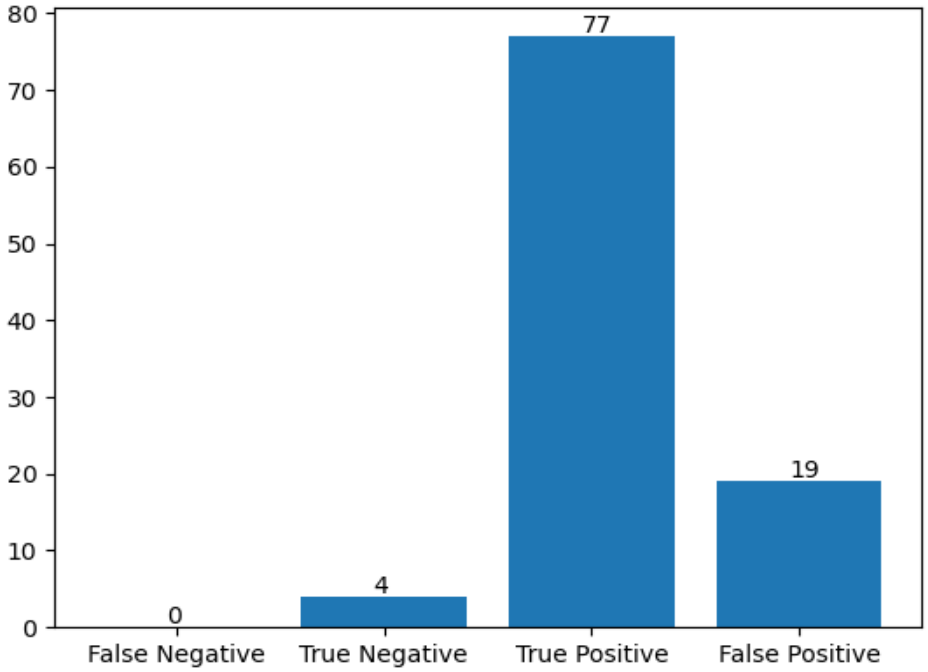
Recall: 100.00

Precision: 80.20

F1 Score: 89.01

	precision	recall	f1-score	support
0	1.00	0.17	0.30	23
1	0.80	1.00	0.89	77
accuracy			0.81	100
macro avg	0.90	0.59	0.59	100
weighted avg	0.85	0.81	0.75	100

Εικόνα 13 Αποτελέσματα από το Classification Report της sklearn. (3η Δοκιμή)



Εικόνα 14 Ιστόγραμμα με τον αριθμό των πελατών και τις σωστές και λάθος προβλέψεις (3η Δοκιμή)

Δοκιμή 4η:

- Αριθμός εγγραφών εκπαίδευσης **900**, Αριθμός εγγραφών εκτίμησης **100 (Test-size:0.1)**
- Ανάμειξη επαφών: Ναι (**Shuffle: True**)
- Αφαίρεση όλων των αρχικών αλφαριθμητικών χαρακτηριστικών. Παραμένουν μόνο τα αριθμητικά *Credit Amount, Age, Duration* (**Features: 3**)
- Αριθμός συστάδων: 4 **(k=4)**
- Αριθμός υποσυστάδων: 6 (**k=6**)
- Μετά την πρώτη συσταδοποίηση πραγματοποιήθηκε μετασχηματισμός των αριθμητικών δεδομένων με την συνάρτηση StandardScaler()

Αποτελέσματα:

Accuracy: 80.00

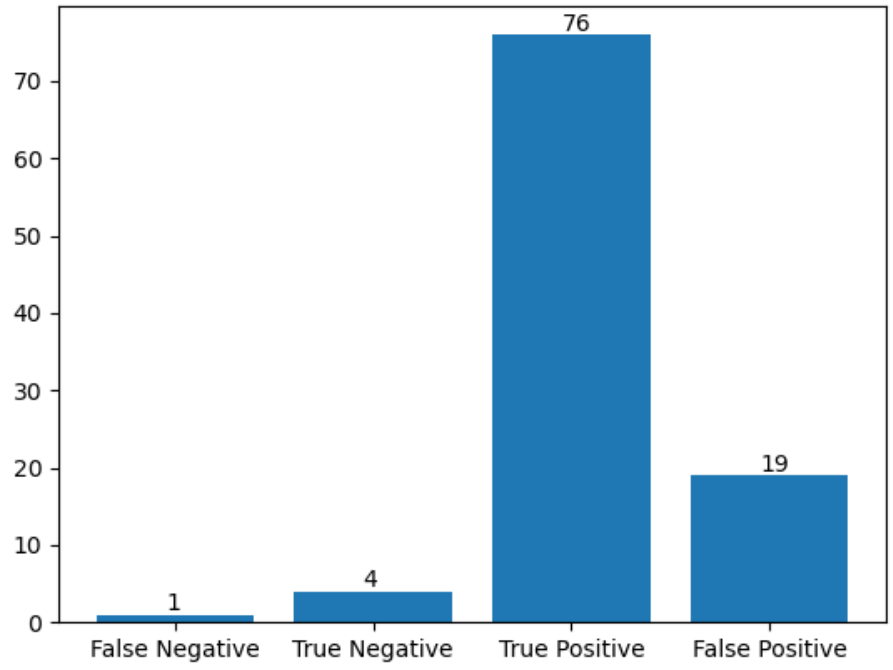
Recall: 98.70

Precision: 80.00

F1 Score: 88.37

	precision	recall	f1-score	support
0	0.80	0.17	0.29	23
1	0.80	0.99	0.88	77
accuracy			0.80	100
macro avg	0.80	0.58	0.58	100
weighted avg	0.80	0.80	0.75	100

Εικόνα 15 Αποτελέσματα από το Classification Report της sklearn. (4η Δοκιμή)



Εικόνα 16 Ιστόγραμμα με τον αριθμό των πελατών και τις σωστές και λάθος προβλέψεις (4η Δοκιμή)

4.6 Συμπεράσματα – Αξιολόγηση Αποτελεσμάτων

Οι δοκιμές που παρουσιάστηκαν παραπάνω, είναι αυτές που έχουν αποφέρει τις μεγαλύτερες βαθμολογίες στις μετρικές αξιολόγησης. Όπως παρατηρήσαμε, οι βασικές αλλαγές, που επηρεάζουν δραματικά την απόδοση του μοντέλου , είναι στους αριθμούς των συστάδων, των υποσυστάδων, στους μετασχηματισμούς των αριθμητικών στοιχείων - Scale & Log Transforamition - και τέλος στις επιλογές χαρακτηριστικών από τα δεδομένα με σκοπό την αποδοτικότερη συσταδοποίηση.

Κατά την εκτίμηση των μοντέλων πιστωτικού κινδύνου, έχουμε 2 τύπους σφαλμάτων, που επηρεάζουν την κερδοφορία και τις πιστωτικές ζημίες του δανειστή και σε κάθε περίπτωση ξεχωριστά πρέπει να ορίζουμε αυτά τα δύο είδη σφαλμάτων σύμφωνα και με τα ζητούμενα του προβλήματος που αντιμετωπίζουμε. Στο δικό μας μοντέλο και σύμφωνα με τους μετασχηματισμούς που κάναμε στην στήλη που αποδίδει το ρίσκο για κάθε πελάτη, το θετικό ρίσκο για το τραπεζικό ίδρυμα πήρε την τιμή 1 (**Good Risk = 1**), ενώ η αρνητική περίπτωση την αντίθετη (**Bad Risk = 0**). Συνεπώς, σύμφωνα με τα [37][38][39] οι τύποι καθορίζονται ως εξής:

1. **Σφάλμα τύπου I (Ψευδώς Θετικό, FP)**. Το μοντέλο προέβλεψε τον πελάτη ως καλοπληρωτή, αλλά στην πραγματικότητα εκείνος αθέτησε και δεν κατάφερε να αποπληρώσει το δάνειο ή την πίστωση . (**Επηρεάζει τις ζημίες και τις προβλέψεις του τραπεζικού ιδρύματος, εφόσον επιλέχθηκαν προς συνεργασία ζημιογόνοι πελάτες**)
2. **Σφάλμα τύπου II (Ψευδώς Αρνητικό, FN):** Το μοντέλο προέβλεψε πως ο πελάτης θα αθετούσε, όμως εκείνος κατάφερε επιτυχώς την αποπληρωμή στο χρόνο που συμφωνήθηκε (**Επηρεάζει την κερδοφορία στο τραπεζικό ίδρυμα καθώς αποκλείονται σοβαροί και φερέγγυοι πελάτες**)

Για να κάνουμε μία ταξινόμηση, έχοντας ως γνώμονα να αποφύγουμε τα ψευδώς θετικά αποτελέσματα, επιλέγουμε την σύνθεση που πραγματοποιήθηκε στην δοκιμή νο 2. Αυτή η επιλογή μας επιστρέφει την μετρική Accuracy παρόμοια με τις υπόλοιπές δοκιμές, ωστόσο έχει εξαιρετικά υψηλά την μετρική **Precision** στο **85.18**. Από την θεωρεία γνωρίζουμε πως το Precision μιλάει για το πόσο ακριβές είναι το μοντέλο ελέγχοντας πόσα από τα προβλεπόμενα θετικά είναι πραγματικά θετικά. Όταν το κόστος των ψευδώς θετικών είναι υψηλό, τότε είναι μία καλή μετρική για αξιολόγηση του μοντέλου. Μετά την δοκιμή νο2 θα ακολουθούσε η δοκιμή νο1 και

μετά οι υπόλοιπες. Καταλήγουμε λοιπόν πως σύμφωνα με τους στόχους που ορίσαμε στο πρόβλημα, που είναι κυρίως να μην ζημιωθεί οικονομικά το τραπεζικό ίδρυμα, σε αρχεία τέτοιου τύπου, όπως το δικό μας, την διαφορά στις λεπτομέρειες κάνει η διαχείριση και ο μετασχηματισμός των δεδομένων έτσι ώστε το μοντέλο να επιτυγχάνει μεγαλύτερο αριθμό σωστών προβλέψεων.

4.7 Μελλοντική Δουλειά – Παράλληλες Δοκιμές

Έχοντας ολοκληρώσει τις βασικές δοκιμές και εφόσον έχουμε εξαγάγει τα συμπεράσματα μας, σε αυτήν την ενότητα αντιμετωπίσαμε το βασικό πρόβλημα της διπλωματικής εργασίας ως ένα δυαδικό πρόβλημα ταξινόμησης **(Binary Classification)**. Οι κλάσεις είναι τα γνωστά από το αρχείο μας **-Good Credit & Bad Credit-** και πραγματοποιήσαμε κάποιες πολύ σύντομες δοκιμές με έτοιμους οδηγούς δημοφιλών εφαρμογών που ασχολούνται με τέτοιου είδους προβλήματα.

Στην πρώτη μας δοκιμή έχουμε, μία νέα σχετικά σε αυτόν τον τομέα, βιβλιοθήκη της Microsoft. Η **ML.Net** όπως ονομάζεται έχει κάποιους σύντομους οδηγούς (Wizards) που αυτοματοποιούν πλήρως την όλη διαδικασία.

Output

Overview Details

ML task: binary-classification
Dataset: C:\Users\Panos Ef\Desktop\b.csv
Column to predict (Label): Risk
Input columns (Features): NA, Age, Sex, Job, Housing, Saving accounts, Checking account, Credit amount, Duration, Purpose
Best Algorithm LightGbmBinary
Best model accuracy: 76%
Training time: 59,7416059 seconds
Models explored (total): 54

Top 5 models explored

Rank	Trainer	Accuracy	AUC	AUPRC	F1-score	Duration
1	LightGbmBinary	0,76	0,806	0,9075	0,8333	1
2	LightGbmBinary	0,7526	0,7224	0,8573	0,831	1,3
3	LightGbmBinary	0,75	0,8125	0,9113	0,8299	0,8
4	LightGbmBinary	0,75	0,806	0,9121	0,8322	1,7
5	LightGbmBinary	0,7474	0,8448	0,8823	0,8261	1,1

Εικόνα 17 Αποτελέσματα εκτέλεσης διαδικασιών μηχανικής μάθησης σε Visual Studio 2019 με ML.Net.

Όπως παρατηρούμε από τα αποτελέσματα, η μέγιστη ακρίβεια (Accuracy) των μοντέλων που δοκιμάσθηκαν ήρθε από το μοντέλο **LightGbmBinary** [40][41], και άγγιξε το **76%** ενώ ο χρόνος εκτέλεσης ήταν περίπου στα **60sec.** Σε αυτήν την εκτέλεση όλη την διαδικασία την ανέλαβε ο αυτοματοποιημένος οδηγός και παρατηρείται ένα μοντέλο χαμηλής ακρίβειας που σίγουρα δεν μπορούμε να το θεωρήσουμε ως αξιόπιστο. Αν είχε προηγηθεί κάποια προ επεξεργασία το δεδομένων από μη αυτοματοποιημένη διαδικασία ή επιλογή των χαρακτηριστικών ίσως ο αλγόριθμος δυαδικής ταξινόμησης να μας επέστρεφε καλύτερα αποτελέσματα.

Η δεύτερη παράλληλη δοκιμή μας ήταν μία εκτέλεση του αλγορίθμου **J48** των δέντρων αποφάσεων (**Decision Trees**) [42], μίας υλοποίησης του αλγορίθμου C4.5 σε γλώσσα Java. Η συγκεκριμένη επιλογή αποτελεί ένα κλασσικό εργαλείο στην πλατφόρμα Weka, όπου εκτελέστηκε αυτό το πείραμα.

```
=== Summary ===

Correctly Classified Instances          74               74      %
Incorrectly Classified Instances        26               26      %
Kappa statistic                          0.2888
Mean absolute error                      0.3309
Root mean squared error                  0.4592
Relative absolute error                 81.4285 %
Root relative squared error            104.1545 %
Coverage of cases (0.95 level)          91      %
Mean rel. region size (0.95 level)      92.5    %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,851    0,577    0,808      0,851    0,829      0,291    0,627     0,797     good
                 0,423    0,149    0,500      0,423    0,458      0,291    0,627     0,416     bad
Weighted Avg.    0,740    0,466    0,728      0,740    0,733      0,291    0,627     0,698

=== Confusion Matrix ===

  a  b   <-- classified as
 63 11 |  a = good
 15 11 |  b = bad
```

Εικόνα 18 Αποτελέσματα εκτέλεσης αλγορίθμου J48 στην πλατφόρμα Weka.

Όπως παρατηρήθηκε από τα πειράματα (Εικόνα 18), ούτε εδώ μπορούμε να ισχυριστούμε ότι η εκτέλεση των αυτοματοποιημένων διαδικασιών είχε τα επιθυμητά αποτελέσματα και είναι πολύ πιθανό πως μόνο ένα μοντέλο κατασκευασμένο ειδικά επάνω στα δικά μας δεδομένα θα μπορούσε να καταφέρει υψηλές αποδόσεις στις μετρικές.

Τέλος ως μελλοντική εργασία θα θεωρούσαμε καλή δοκιμή την δημιουργία ενός προσαρμοσμένου μοντέλου, το οποίο θα βασίζεται αποκλειστικά στα νευρωνικά δίκτυα, σύμφωνα και με τις έρευνες που παρουσιάστηκαν εδώ [43][44], που θα δώσει

με τις σωστές ρυθμίσεις στις παραμέτρους πιο επιτυχημένες προβλέψεις. Ταυτόχρονα μία ακόμη δοκιμή που είναι άξια προσοχής, θα ήταν τα **σχήματα ομαδικής μάθησης (Ensemble Learning Schemes)**, οπού εκεί θα χρησιμοποιούσαμε τα καλύτερα σε επιδόσεις μοντέλα και θα συνδυάζαμε και θα επεξεργαζόμασταν τις εξόδους τους με σκοπό τον καλύτερο δυνατό βαθμό στις μετρικές αξιολόγησης σε σχέση πάντα με την απόδοση που θα είχε ο εκάστοτε αλγόριθμος μόνος του.

Παράρτημα

Βιβλιογραφία

[1] Larose, D. T. (2004). Discovering knowledge in data, Central Connecticut State University, A John Wily & Sons. *Inc., Canada*, 1-5.

[2] Hand, D. J. (2001). Heikki mannila Padhraic Smyth. In *Principles of data mining, Introduction to data mining*. MIT Press.

[3] Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, *5*(4), 83-124.

[4] Edelstein, H. A. (1999). Introduction To Data Mining And Knowledge Discovery, Third Edition. Herbert A. Edelstein.

[5] Αγγελόπουλος, Π. & Ηρειώτης, Ν. & Συριόπουλος, Κ. (2008). Στρατηγική Τραπεζών – Χρηματοοικονομικά Εργαλεία Στήριξης των Ειδικών Μορφών Πίστης. Εκδόσεις ΕΑΠ, Πάτρα.

[6] Συριόπουλος, Κ. (2008). Στρατηγική Τραπεζών – Διαχείριση Τραπεζικού Κινδύνου. Εκδόσεις ΕΑΠ, Β΄Έκδοση Πάτρα.

[7] Bessis, J. (2002). Risk Management in Banking. 3rd Edition, John Wiley & Sons, New York.

[8] Γ. Σαπουντζόγλου και Χ. Πεντότης, (2017), Τραπεζική Οικονομική, Β΄ έκδοση (επικαιροποιημένη), εκδόσεις Μπένου, Αθήνα 2017.

[9] Jorion, P., (2001). The New Benchmark for Managing Financial Risk. Second Edition. Mcgraw Hill.

[10] A. Saunders & M. M. Cornett, (2003). Financial Institutions Management: A Risk Management Approach, 6th edition, McGraw-Hill Education

[11] Προβόπουλος, Γ., Καπόπουλος, Π., (2001), Η Δύναμη του Χρηματοοικονομικού Συστήματος, εκδ. Κριτική

[12] Roesch, Daniel and Scheule, Harald (Harry), A Multi-Factor Approach for Systematic Default and Recovery Risk (2005). Journal of Fixed Income, Vol. 15, No. 2, 2005, pp. 63-75.

[13] Randall, P. & Σπαθαράκης, Δ. (2009). «Βασιλεία ΙΙ: Ο γρήγορος δρόμος για μια νέα εποχή στον τραπεζικό χώρο»

[14] K. Aas (2005). The Basel II IRB approach for credit portfolios: A survey, NorskRegnesentral, Norwegian Computing Center

[15] EU-CRR (2013). REGULATION (EU) No 575/2013 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL, Capital Requirements Regulation – CRR. Official Journal of the European Union.

[16] Engelmann, B., and Rauhmeier, R., (2006). The Basel II Risk Parameters. Springer Berlin.

[17] Brealey, R.A., Myers, S.C. and Allen, F. (2011) Principles of Corporate Finance. 10th Edition, McGraw-Hill/Irwin, New York.

[18] BCBS, (2001). The New Basel Capital Accord, Bank of International Settlements.

[19] Anderson, R. (2007). The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation. Oxford University Press, UK Oxford.

[20] Guegan, D. & Hassani, B. (2018). Regulatory learning: How to supervise machine learning models?An application to credit scoring. The Journal of Finance and Data Science Volume 4, Issue 3, p. 157-171.

[21] Khandani, A.E & Kim, A.J. & Lo, A.W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, Volume 34, Issue 11, November 2010, pp. 2767-2787

[22] Kruppa, J. & Schwarz, A. & Arminger, G. & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. Expert Systems with Applications, Volume 40, Issue 13, 1 October 2013, p. 5125-5131.

[23] Kritzinger, N. & Vuuren, G. (2018). An optimised credit scorecard to enhance cut-off score determination. South African Journal of Economic and Management Sciences. 21. DOI: 10.4102/sajems.v21i1.1571
[24] Rajan, M. & Tulasi, B. (2015). Credit Scoring Process using Banking Detailed Data Store. International Journal of Applied Information Systems, Volume 8, No.6, April 2015
[25] Bao, W. & Lianju, N. & Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. Expert Systems With Applications Volume 128, 15 August 2019, p. 301-315
[26] Αγγελόπουλος, Χρ., Παναγιώτης, (2005), Τράπεζες και Χρηματοπιστωτικό Σύστημα: αγορές προϊόντα, κίνδυνοι, Εκδόσεις Σταμούλης
[27] Saunders Anthony, Allen Linda, (2002) Credit Risk Measurement: New Approaches to Value at Risk and Other Paradigms, 2nd Edition
ΑΡΘΡΑ
[28] Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, vol. 23 no. 4, p. 589–609.
[29] Liu, Y., Li, W., & Li, Y. (2007, August). Network traffic classification using k-means clustering. In *Second international multi-symposiums on computer and computational sciences (IMSCCS 2007)* (pp. 360-365). IEEE.
[30] Mohamad, I. B., & Usman, D. (2013). Standardization and its effects on K-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(17), 3299-3303.
[31] Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, *5*(2), 1.
[32] Beauxis-Aussalet, E., & Hardman, L. (2014). Visualization of confusion matrix for non-expert users. In *IEEE Conference on Visual Analytics Science and Technology (VAST)-Poster Proceedings*.
[33] Zait, M., & Messatfa, H. (1997). A comparative study of clustering methods. *Future Generation Computer Systems*, *13*(2-3), 149-159.
[34] Moore, A. (2001). K-means and Hierarchical Clustering.
[35] Davidson, I. (2002). Understanding K-means non-hierarchical clustering. *Computer Science Department of State University of New York (SUNY), Albany*.
[36] Marutho, D., Handaka, S. H., & Wijaya, E. (2018, September). The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *2018 international seminar on application for technology of information and communication* (pp. 533-538). IEEE.
[37] Abby M. Kern (2017): "Credit Score Analysis", Spring, Southern Illinois University Carbondale, OpenSIUC.
[38] Abdou, H. & Pointon, J. (2011): "Credit scoring, statistical techniques and evaluation criteria: a review of the literature", Intelligent Systems in Accounting, Finance & Management, 18 (2-3), 59-88
[39] Shen, Shi-Wei, Nguyen, Tri-Dung & Ojiako, Udechukwu (2013): "Modelling the predictive performance of credit scoring", Acta Commercii
[40] Ge, D., Gu, J., Chang, S., & Cai, J. (2020, April). Credit card fraud detection using lightgbm model. In *2020 International Conference on E-Commerce and Internet Technology (ECIT)* (pp. 232-236). IEEE.
[41] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.
[42] Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. *Intrenational Journal ofAdvanced Research in Computer Science and Soft-ware Engineering*, *3*(6).
[43] Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, *37*(9), 6233-6239.

[44] Shen, F., Zhao, X., Li, Z., Li, K., & Meng, Z. (2019). A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. *Physica A: Statistical Mechanics and its Applications*, *526*, 121073.