ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ

ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ ΔΙΟΙΚΗΣΗ, ΛΟΓΙΣΤΙΚΗ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

«Εφαρμογή Τεχνικών Εξόρυξης Δεδομένων από τα Χρηματοπιστωτικά Ιδρύματα για την Πρόληψη του Οικονομικού Εγκλήματος»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΣΤΕΛΛΑΣ ΝΤΖΑΒΙΔΑ

Επιβλέπων : Ευστάθιος Κύρκος
Καθηγητής, Διεθνές Πανεπιστήμιο της Ελλάδος

Θεσσαλονίκη, Σεπτέμβριος 2023

Ευχαριστίες

Θα ήθελα να ευχαριστήσω από καρδιάς τον επιβλέποντα καθηγητή μου, κύριο Κύρκο Ευστάθιο, για την αμέριστη βοήθεια και την επιστημονική καθοδήγησή του που συντέλεσαν στην συγγραφή της παρούσας διπλωματικής εργασίας.
Επίσης, θα ήθελα να εκφράσω ένα μεγάλο ευχαριστώ στους γονείς μου, στον σύζυγο και στα παιδιά μου, για την στήριξη, την εμπιστοσύνη και την κατανόηση που έδειξαν κατά τη διάρκεια των σπουδών μου.

Περίληψη

Η παρούσα εργασία αφορά στην εφαρμογή των πιο εξελιγμένων μεθόδων εξόρυξης γνώσης σε ένα πραγματικό σύνολο δεδομένων προκειμένου να ανιχνευτεί πιθανή απάτη κατά το άνοιγμα ενός τραπεζικού λογαριασμού. Η συμβολή της παρούσας εργασίας επικεντρώνεται σε δύο τομείς. Πρώτον, επιχειρεί να αναδειχθούν οι ισχυρότερες μεταβλητές που αποτελούν σημαντικές ενδείξεις για πιθανή διάπραξη δόλιων συναλλαγών. Ο εντοπισμός, ιεράρχηση κατά σπουδαιότητα και αξιολόγηση αυτών των μεταβλητών αποτελεί σημαντικό κριτήριο για την αποδοχή ή μη των εν δυνάμει πελατών από ένα χρηματοπιστωτικό ίδρυμα. Δεύτερον, προτείνει τη δημιουργία αξιόπιστων μοντέλων που να μπορούν να προβλέπουν με υψηλή ακρίβεια την πλειονότητα μελλοντικών περιπτώσεων. Η εργασία οργανώνεται σε τρία μέρη.

Το πρώτο μέρος παρουσιάζει τον ρόλο των πιστωτικών ιδρυμάτων και ιδιαίτερα την λειτουργία της μονάδας Κανονιστικής Συμμόρφωσης στον αγώνα για την καταπολέμηση του οικονομικού εγκλήματος. Αναλύεται η βασική αρχή Know Your Customer και το θεσμικό-νομικό πλαίσιο που ισχύει σήμερα στην Ελλάδα. Παρατίθεται ο ορισμός για το Ξέπλυμα Βρώμικου Χρήματος και περιγράφεται ο κύκλος ενός συστήματος νομιμοποίησής του. Τέλος αναφέρονται οι κυριότερες τεχνολογίες ανάλυσης δεδομένων που χρησιμοποιούνται σήμερα για την καταπολέμηση της απάτης καθώς και βέλτιστες πρακτικές που έχουν προταθεί προς υιοθέτηση στο μέλλον.

Στο δεύτερο μέρος περιγράφονται τα στάδια Ανακάλυψης Γνώσης, οι τέσσερις κατηγορίες μηχανικής μάθησης καθώς και οι σημαντικότερες μέθοδοι επιβλεπόμενης μάθησης. Εκτενής αναφορά γίνεται στο ειδικό θέμα της ανισοκατανομής των κλάσεων, του κόστους σφάλματος και της μέτρησης της απόδοσης των ταξινομητών.

Το τρίτο μέρος παρέχει τη μεθοδολογία της παρούσας εργασίας. Για την επίτευξη των στόχων, χρησιμοποιούνται διάφορες μέθοδοι. Αυτές περιλαμβάνουν την προεπεξεργασία των δεδομένων, την επιλογή χαρακτηριστικών και την εξισορρόπηση της κατανομής των κλάσεων, την ανάπτυξη μοντέλων με αλγορίθμους μηχανικής μάθησης, την επικύρωση των μοντέλων έναντι άγνωστων παρατηρήσεων και τέλος την ταξινόμηση κατά σειρά σημαντικότητας των μεταβλητών που οδηγούν στην παραγωγή των προτύπων. Τέλος παρουσιάζονται τα αποτελέσματα της ανάλυσης, τα συμπεράσματα, οι περιορισμοί και

πιθανές επεκτάσεις για έρευνα. Στο τέλος κάθε μέρους παρατίθεται η σχετική βιβλιογραφία.

Λέξεις Κλειδιά: <<μηχανική μάθηση>>, <<κανονιστική συμμόρφωση>>, <<fraud analytics>>, << AML>>

Abstract

This paper is about applying the most sophisticated data mining methods to a real dataset in order to detect possible fraud when opening a bank account. The contribution of this paper focuses on two areas. First, it attempts to highlight the strongest variables that are important indicators of possible fraudulent transactions. The identification, ranking by importance and evaluation of these variables is an important criterion for the acceptance or not of potential customers by a financial institution. Secondly, it suggests the creation of reliable models that can predict with high accuracy the majority of future cases. The paper is organized in three parts.

The first part presents the role of financial institutions and in particular the function of the Compliance Unit in the fight against financial crime. It analyses the basic Know Your Customer principle and the institutional-legal framework currently in force in Greece. The definition of Money Laundering is given and the cycle of a money laundering scheme is described. Finally, the main data analysis technologies currently used to combat fraud are mentioned, as well as best practices that have been proposed for adoption in the future.

The second part describes the stages of Knowledge Discovery, the four categories of Machine Learning and the most important supervised learning methods. Extensive reference is made to the specific topic of class imbalance, Cost Sensitive Learning and Performance Metrics of classifiers.

The third part provides the methodology of this paper. To achieve the objectives, several methods are used. These include data pre-processing, feature selection and balancing the distribution of classes, developing models with machine learning algorithms, validating the models against unknown observations and finally evaluating and ranking in order of importance the variables leading to the generation of the models. Finally, the results of the analysis, conclusions, limitations and possible extensions for research are presented. At the end of each part, the relevant literature is listed.

Keywords: <<machine learning>>, <<regulatory compliance>>, <<fraud analytics>>, << AML>>

Περιεχόμενα

Περίληψη 1
Abstract 3
ΜΕΡΟΣ Α΄ ΞΕΠΛΥΜΑ ΧΡΗΜΑΤΟΣ ΚΑΙ ΣΥΣΤΗΜΑΤΑ AML 8
Ενότητα 1 Κανονιστική Συμμόρφωση και καταπολέμηση οικονομικού εγκλήματος 8
1.1 Εισαγωγή 8
1.2 Η βασική αρχή «Γνώρισε τον πελάτη σου» 8
KYC «Know Your Customer» 8
1.3 Η λειτουργία της Κανονιστικής Συμμόρφωσης 10
στα Πιστωτικά Ιδρύματα 10
1.4 Νομικό πλαίσιο στην Ελλάδα για την αποτροπή νομιμοποίησης εσόδων από παράνομες δραστηριότητες 11
1.5 Επιπτώσεις μη συμμόρφωσης 14
Ενότητα 2 Πρόληψη και καταστολή Ξεπλύματος Χρήματος και Χρηματοδότησης της τρομοκρατίας 15
2.1 Ξέπλυμα Βρώμικου Χρήματος και Χρηματοδότηση της Τρομοκρατίας 15
2.2 Ο ρόλος των πιστωτικών ιδρυμάτων 18
2.3 Προσέγγιση με βάση τον κίνδυνο (Risk based approach) 21
Ενότητα 3 Συστήματα AML 23
3.1 Εξέλιξη σύγχρονων συστημάτων AML 23
3.2 Βέλτιστες πρακτικές 25
ΒΙΒΛΙΟΓΡΑΦΙΑ ΜΕΡΟΥΣ Α΄ 28
ΜΕΡΟΣ Β΄ ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΠΡΟΗΓΟΥΜΕΝΗ ΕΡΕΥΝΑ 31
Ενότητα 1 Εξόρυξη γνώσης 31
1.1 Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων 31
1.2 Επιβλεπόμενη μάθηση 34
ΒΙΒΛΙΟΓΡΑΦΙΑ ΜΕΡΟΣ Β' ΕΝΟΤΗΤΑ 1 37
Ενότητα 2 Ειδικά θέματα κατηγοριοποίησης 38
2.1 Ανισοκατανομή των κλάσεων 38
2.2 Κόστος σφάλματος 40
2.3 Μέτρηση της απόδοσης των ταξινομητών 42
ΒΙΒΛΙΟΓΡΑΦΙΑ ΜΕΡΟΣ Β΄ ΕΝΟΤΗΤΑ 2 46
ΜΕΡΟΣ Γ΄ ΕΡΓΑΣΤΗΡΙΑΚΟ ΜΕΡΟΣ - WEKA 49

Ενότητα 1 Μεθοδολογία έρευνας 49
1.1 Εισαγωγή 49
1.2 Σύνολο δεδομένων data set 50
1.3 Μοντέλα Κατηγοριοποίησης και άλλοι αλγόριθμοι 53
1.4 Σημαντικότητα μεταβλητών 54
Ενότητα 2 Αποτελέσματα και συμπεράσματα 56
2.1 Αποτελέσματα ανάλυσης κατηγοριοποίησης 56
2.2 Αποτελέσματα ταξινόμησης σημαντικότητας μεταβλητών 58
2.3 Συμπεράσματα, περιορισμοί και προτάσεις για μελλοντική έρευνα 59
Πίνακας Εικόνων 62
ΠΑΡΑΡΤΗΜΑ 63

ΜΕΡΟΣ Α΄ ΞΕΠΛΥΜΑ ΧΡΗΜΑΤΟΣ ΚΑΙ ΣΥΣΤΗΜΑΤΑ AML

Ενότητα 1 Κανονιστική Συμμόρφωση και καταπολέμηση οικονομικού εγκλήματος

1.1 Εισαγωγή

Τα τραπεζικά ιδρύματα έχουν σχεδιάσει αυστηρές πολιτικές δεοντολογίας και συμμόρφωσης οι οποίες αποτελούν ύψιστης σημασίας προτεραιότητα, αφενός γιατί θεωρούνται θετικοί παράγοντες που ενεργά συμβάλλουν στην ανάπτυξή τους αφετέρου για λόγους δημόσιας εικόνας, αποφυγής οικονομικής ζημίας και αποφυγής του ρίσκου επιβολής νομικών και πειθαρχικών κυρώσεων. Η επαγγελματική συμμόρφωση με το θεσμικό πλαίσιο ήταν ανέκαθεν η πλέον βασική αξία στον τραπεζικό κλάδο. Στο εξαιρετικά σύνθετο και δυναμικά μεταβαλλόμενο περιβάλλον στο οποίο λειτουργούν τα τραπεζικά ιδρύματα, η εφαρμογή της βασικής αρχής «Γνώρισε τον πελάτη σου» και η αποτελεσματικότητα των μηχανισμών ανίχνευσης δόλιων συναλλαγών αποτελούν την βάση για την αποτροπή κινδύνων, για την προστασία των καταναλωτών και για την διασφάλιση συνέπειας ως προς τις απαιτήσεις συμμόρφωσης. Εξειδικευμένα πληροφορικά συστήματα βοηθούν τις τράπεζες να διατηρούν τεράστιες βάσεις δεδομένων με τα στοιχεία, τις συνήθειες και τις προτιμήσεις των πελατών τους ενώ ταυτόχρονα πραγματοποιείται πλήρης καταγραφή των συναλλαγών που εκτελούνται από αυτούς.

1.2 Η βασική αρχή «Γνώρισε τον πελάτη σου»

KYC «Know Your Customer»

Η εφαρμογή των διαδικασιών «Γνώρισε τον πελάτη σου» ή KYC – Know Your Customer από ένα τραπεζικό ίδρυμα υιοθετείται κυρίως για δύο σημαντικούς λόγους. Ο πρώτος λόγος

στοχεύει στην ανάπτυξη σχέσεων με τους πελάτες σταθερών και διαχρονικών, που χαρακτηρίζονται από υψηλή ποιότητα.

Το στρατηγικό μάρκετινγκ θεωρεί ότι βασική προϋπόθεση για την ανάπτυξη μιας πελατοκεντρικής εμπορικής προσέγγισης είναι η βαθιά γνώση για τον πελάτη. Η συλλογή και η ανάλυση όσο το δυνατόν περισσότερων στοιχείων του πελατολογίου και της αγοράς παρέχει την κατάλληλη πληροφόρηση για τις ιδιαίτερες ανάγκες, τις συμπεριφορές και τις απαιτήσεις των πελατών. Η επεξεργασία και η διάχυση της πληροφορίας αυτής σε όλα τα επίπεδα ενός οργανισμού οδηγεί στη σωστότερη λήψη αποφάσεων και στην ανάπτυξη κατάλληλων εμπορικών ενεργειών.

Γνωρίζοντας τον πελάτη είναι δυνατή η τμηματοποίηση του πελατολογίου σε κατηγορίες, η καλύτερη στόχευση και η προσωποποιημένη προσέγγιση στις διαφορετικές πελατειακές ομάδες με σκοπό την πιο αποτελεσματική κάλυψη των αναγκών τους. Αναγνωρίζοντας τις προσδοκίες και τους αντικειμενικούς στόχους του πελάτη το μάρκετινγκ γίνεται καλύτερο, προτείνοντας τα κατάλληλα προϊόντα και υπηρεσίες, αυξάνοντας το ποσοστό των σταυροειδών πωλήσεων και το επίπεδο ικανοποίησης των πελατών.

Ο δεύτερος όμως πολύ σημαντικός λόγος είναι γιατί τα τραπεζικά ιδρύματα οφείλουν να λειτουργούν και να συμμορφώνονται σύμφωνα με τους κανόνες και τις συστάσεις που τους υποδεικνύονται σε διαφορετικά επίπεδα δηλαδή σε εθνικό, ευρωπαϊκό αλλά και επιχειρησιακό. Η διαδικασία πιστοποίησης της ταυτότητας των πελατών αποτελεί την βάση όλων των ενεργειών πρόληψης του ξεπλύματος χρήματος και της χρηματοδότησης της τρομοκρατίας. Κατά την έναρξη αλλά και σε όλη την διάρκεια της συνεργασίας συλλέγονται στοιχεία από αξιόπιστες και ανεξάρτητες πηγές που εξακριβώνουν την ταυτότητα των πελατών και διαμορφώνουν το αναμενόμενο συναλλακτικό προφίλ τους, το οποίο αξιολογείται και επικαιροποιείται ανά τακτά χρονικά διαστήματα.

Με αυτόν τον τρόπο αποφεύγεται το ρίσκο επιβολής νομικών ή πειθαρχικών κυρώσεων που μπορεί να εφαρμοστούν στην τράπεζα και στους υπαλλήλους της όπως εσωτερικές πειθαρχικές κυρώσεις, χρηματικά πρόστιμα, αφαίρεση άδειας ακόμα και ποινή φυλάκισης. Τα τραπεζικά ιδρύματα προστατεύουν την εικόνα και την καλή φήμη τους, η οποία μπορεί να επηρεαστεί από την αρνητική δημοσιότητα και που θα έχει ως αποτέλεσμα την απώλεια της εμπιστοσύνης των πελατών και του επενδυτικού κοινού και άρα άμεσο αντίκτυπο στην

πτώση της εμπορικής τους δραστηριότητας. Τέλος με την εφαρμογή των διαδικασιών συμμόρφωσης μπορεί να προληφθεί και να αντιμετωπιστεί εγκαίρως η οικονομική ζημία από διάφορες μορφές απάτης.

1.3 Η λειτουργία της Κανονιστικής Συμμόρφωσης στα Πιστωτικά Ιδρύματα

Ο ρόλος της κανονιστικής συμμόρφωσης στα πιστωτικά ιδρύματα είναι θεσμικός, προβλέπεται από την Πράξη του Διοικητή της Τράπεζας Ελλάδος 2577/2006 που προσδιορίζει *«το πλαίσιο των αρχών λειτουργίας και κριτηρίων αξιολόγησης της οργάνωσης και των Συστημάτων Εσωτερικού Ελέγχου των πιστωτικών και των χρηματοδοτικών ιδρυμάτων»* (ΠΔΤΕ/2577/2006). Έχει βασικό σκοπό την ευθυγράμμιση των επιχειρηματικών στόχων των τραπεζών με τις απαιτήσεις συμμόρφωσης ως προς το θεσμικό πλαίσιο και διασφαλίζει την καλή φήμη, την αξιοπιστία και την ακεραιότητα της λειτουργίας τους. Συνεισφέρει στην προσαρμογή των συστημάτων και όλων των εσωτερικών διαδικασιών ώστε να τηρούνται οι εκάστοτε προβλέψεις των νόμων, των κανονισμών και των διατάξεων των ευρωπαϊκών εποπτικών αρχών.

Αποτελεί έναν από τους τρεις πυλώνες του Συστήματος Εσωτερικού Ελέγχου μαζί με την μονάδα Διαχείρισης Κινδύνου και την μονάδα Εσωτερικού Ελέγχου. Έχει διευρυμένο ρόλο που εμπλέκεται σε πολλά επιπλέον πεδία όπως τον σχεδιασμό νέων προϊόντων, εντύπων και διαδικασιών. Εκτείνεται σε θέματα προστασίας των συναλλασσομένων με τις τράπεζες καθώς και τήρησης του τραπεζικού απορρήτου και διασφάλισης της εμπιστευτικότητας των πληροφοριών των οποίων λαμβάνουν γνώση οι τράπεζες. Τέλος θέτει τον κώδικα ηθικής συμπεριφοράς και επαγγελματικής δεοντολογίας του προσωπικού των πιστωτικών ιδρυμάτων καθώς και την πολιτική υιοθέτησης βέλτιστων πρακτικών για την αποφυγή περιπτώσεων σύγκρουσης συμφερόντων (Ελληνική Ένωση Τραπεζών, 2006).

Η συμμόρφωση με τους κανόνες αποτελεί εταιρική κουλτούρα και ευθύνη της Διοίκησης, εφαρμόζεται σε όλα τα επίπεδα του οργανισμού και αφορά το σύνολο του προσωπικού. Η

μονάδα κανονιστικής συμμόρφωσης διατηρεί στενή συνεργασία με τις νομικές υπηρεσίες του οργανισμού, με τις αντίστοιχες μονάδες κανονιστικής συμμόρφωσης των υπόλοιπων πιστωτικών ιδρυμάτων και με τις εποπτικές αρχές όπως: Τράπεζα της Ελλάδος, Κεντρικές τράπεζες και Επιτροπές Κεφαλαιαγορών, την Ευρωπαϊκή Αρχή Τραπεζών και την Αρχή Καταπολέμησης της Νομιμοποίησης Εσόδων από Εγκληματικές Δραστηριότητες (Financial Intelligence Unit FIU). Η Ελληνική Αρχή Καταπολέμησης συγκροτείται από τον Πρόεδρο που είναι εν ενεργεία ανώτατος εισαγγελικός λειτουργός και 17 μέλη που έχουν διακριθεί για την επιστημονική τους κατάρτιση, το ήθος τους και την εμπειρία τους στον τραπεζικό - νομικό τομέα.

1.4 Νομικό πλαίσιο στην Ελλάδα για την αποτροπή νομιμοποίησης εσόδων από παράνομες δραστηριότητες

Η ΠΔΤΕ 2577/2006 έχει πεδίο εφαρμογής στα πιστωτικά ιδρύματα και στους χρηματοπιστωτικούς οργανισμούς οι οποίοι είναι οι ασφαλιστικές εταιρίες, οι εταιρίες παροχής επενδυτικών υπηρεσιών και οι εταιρίες διαχείρισης αμοιβαίων κεφαλαίων, οι εταιρίες χρηματοδοτικών μισθώσεων, οι εταιρείες κεφαλαίου επιχειρηματικών συμμετοχών, οι εταιρείες πρακτορείας επιχειρηματικών απαιτήσεων τρίτων και τα ανταλλακτήρια συναλλάγματος. Όλοι οι παραπάνω φορείς αποτελούν τις κυριότερες διόδους εισόδου για «ξέπλυμα βρώμικου χρήματος».

Σύμφωνα με τον νόμο 4557/2018 «για την πρόληψη και καταστολή της νομιμοποίησης εσόδων από εγκληματικές δραστηριότητες και της χρηματοδότησης της τρομοκρατίας» εντάσσονται πέραν των πιο πάνω αναφερόμενων κλάδων επιπλέον τομείς δραστηριότητας που εμπίπτουν στις σχετικές διατάξεις. Οι κλάδοι αυτοί ενδέχεται να εμπλακούν στη νομιμοποίηση παράνομων προσόδων και είναι εταιρείες ορκωτών ελεγκτών-λογιστών, εταιρίες παροχής λογιστικών-φοροτεχνικών υπηρεσιών, συμβολαιογράφοι και δικηγόροι όταν ενεργούν για λογαριασμό των πελατών τους σε χρηματοπιστωτικές συναλλαγές ή συναλλαγές ακινήτων, μεσίτες ακινήτων μεγάλης αξίας, καζίνο, στοιχηματικές εταιρίες ή

πρακτορεία που παρέχουν υπηρεσίες τυχερών παιγνίων, οι έμποροι αγαθών μεγάλης αξίας και πολλοί άλλοι.

Οι νόμοι 4816/2021 και 4734/2020 που τροποποίησαν τον ν.4557/2018 και ενίσχυσαν το νομοθετικό πλαίσιο ορίζουν ότι η νομιμοποίηση εσόδων από εγκληματικές δραστηριότητες (ξέπλυμα χρήματος) συντελείται όταν ασκούνται ενέργειες ή παραλείψεις που αποβλέπουν στην μετατροπή ή μεταβίβαση περιουσίας, που προέρχεται από εγκληματική δραστηριότητα, με σκοπό την απόκρυψη ή τη συγκάλυψη της παράνομης προέλευσής της. Η απόκτηση, κατοχή ή ακόμα και διαχείριση τέτοιας περιουσίας και η διευκόλυνση διάπραξης των παραπάνω δραστηριοτήτων αποτελεί παράνομη πράξη για οποιονδήποτε ενέχεται σε αυτές (Παράρτημα, Βασικά Αδικήματα, Πίνακας 1).

Είναι πάγια τακτική των εγκληματιών να χρησιμοποιείται ο χρηματοπιστωτικός τομέας για την τοποθέτηση ή τη διακίνηση μέσω αυτού των εσόδων που προέρχονται από παράνομες δραστηριότητες, με σκοπό να προσδοθεί νομιμοφάνεια στα εν λόγω έσοδα.

Σύμφωνα με το ΠΔΤΕ 2577/2006 οι παραπάνω οργανισμοί έχουν την κύρια αρμοδιότητα και ευθύνη να αξιολογούν τους υποψήφιους και τους υφιστάμενους πελάτες τους ως προς τους κινδύνους που αντιπροσωπεύουν και να αναλαμβάνουν την ανάλογη διαχείρισή τους. Ειδικότερα εφαρμόζονται κριτήρια κατά την προσέλκυση και αποδοχή των πελατών, οι οποίοι ταξινομούνται ανά επίπεδο κινδύνου και στην συνέχεια πραγματοποιείται διαρκής καταγραφή και παρακολούθηση της δραστηριότητάς τους. Απαραίτητη προϋπόθεση είναι η ύπαρξη αναλυτικών εσωτερικών διαδικασιών και η επιλογή πληροφοριακών συστημάτων ικανών για τον άμεσο εντοπισμό των συναλλαγών, οι οποίες δεν συνάδουν με την εικόνα που έχουν σχηματίσει τα πιστωτικά ιδρύματα για τον πελάτη τους και τη συναλλακτική του συμπεριφορά. Οι συναλλαγές αυτές πρέπει να διερευνώνται και εφόσον απαιτείται, να αναφέρονται στις εποπτικές αρχές με την κατάλληλη τεκμηρίωση και επάρκεια.

Τα πιστωτικά ιδρύματα καθώς και οι υπόλοιποι οργανισμοί που εμπίπτουν στις εν λόγω διατάξεις, ελέγχονται για τα μέτρα επιμέλειας τα οποία έχουν υιοθετήσει, ώστε να αντιμετωπίζονται επαρκώς οι κίνδυνοι συμμόρφωσης, και λογοδοτούν για την αποτελεσματικότητά τους. Στην Ελλάδα το έργο της εποπτείας, καθοδήγησης και ελέγχου των πιο πάνω οργανισμών έχει ανατεθεί (ανά κλάδο δραστηριότητας): στην Τράπεζα της Ελλάδος και στην Επιτροπή Κεφαλαιαγοράς (τράπεζες και χρηματοπιστωτικοί οργανισμοί),

στο Υπουργείο Δικαιοσύνης (για συμβολαιογράφους και δικηγόρους), στην Επιτροπή Λογιστικής Τυποποίησης και Ελέγχων (για ορκωτούς ελεγκτές λογιστές), στην Επιτροπή Εποπτείας και Ελέγχου Παιγνίων και στην Ανεξάρτητη Αρχή Δημοσίων Εσόδων Α.Α.Δ.Ε. Επιπρόσθετα Κεντρικός Συντονιστικός Φορέας είναι το Υπουργείο Οικονομικών που επιβλέπει την ομαλή συνεργασία των αρμόδιων αρχών στην Ελλάδα και αναλαμβάνει τη διεθνή εκπροσώπηση της χώρας στις διασκέψεις φορέων της Ευρωπαϊκής Ένωσης, του Συμβουλίου της Ευρώπης και της Financial Action Task Force FATF.

Στο σημείο αυτό οφείλουμε να διευκρινίσουμε ότι σε διεθνές επίπεδο σημαντικό ρόλο διαδραματίζει η Ομάδα Χρηματοπιστωτικής Δράσης ή Financial Action Task Force (FATF) της οποίας μέλος είναι και η Ελλάδα. Είναι ένας ανεξάρτητος διακυβερνητικός οργανισμός που ιδρύθηκε το 1989 με πρωτοβουλία της G7 και απαριθμεί 39 κράτη μέλη και δυο περιφερειακές οργανώσεις. Θέτει τα διεθνή πρότυπα και προωθεί πολιτικές για την προστασία του παγκόσμιου χρηματοπιστωτικού συστήματος. Έχει εκδώσει 40 κατευθυντήριες οδηγίες που αναγνωρίζονται ως οι παγκόσμιες συστάσεις κατά της νομιμοποίησης του ξεπλύματος βρώμικου χρήματος (AML anti-money laundering) και την καταπολέμηση της χρηματοδότησης της τρομοκρατίας (CFT combating financing terrorism), τις οποίες έχουν δεσμευτεί να εφαρμόσουν περισσότερες από 200 χώρες. Στην ετήσια ανασκόπηση που έκδωσε η FATF για το έτος 2022, οι οδηγίες αυτές έχουν υιοθετηθεί σε ποσοστό 76% των χωρών έναντι του αντίστοιχου 36% το 2012 (έτος που παρουσιάστηκαν για πρώτη φορά), γεγονός που αποδεικνύει την τεράστια συμβολή του οργανισμού. Για την χώρα μας ο κίνδυνος ξεπλύματος βρώμικου χρήματος αξιολογείται μέτριος προς υψηλός και με εκτίμηση του 2019 η Ελλάδα έχει εφαρμόσει αποτελεσματικά συστήματα πρόληψης και ειδικότερα τα ελληνικά πιστωτικά ιδρύματα με αποτέλεσμα να επιτυχαίνουν υψηλό ποσοστό συμμόρφωσης (FATF, 2019).

Καταλήγοντας σπουδαίο ρόλο στην Ελλάδα έχει η Επιτροπή Στρατηγικής, ένα διυπουργικό όργανο για την χάραξη της εθνικής στρατηγικής πάνω στο θέμα αυτό. Διενεργεί εκτίμηση των κινδύνων, προτείνει μέτρα και είναι αρμόδια για την διασφάλιση της συμμόρφωσης της Ελλάδας με τα διεθνή πρότυπα σε επίπεδο χωρών.

1.5 Επιπτώσεις μη συμμόρφωσης

Οι επιπτώσεις της μη συμμόρφωσης ή πλημμελούς εκτέλεσης ελέγχων και εποπτείας είναι καταστροφικές για το κύρος και την αξιοπιστία της λειτουργίας των τραπεζών. Ανάλογα με τον βαθμό σπουδαιότητας οι κυρώσεις μπορεί να καταλογισθούν διαζευκτικά ή σωρευτικά τόσο στο πιστωτικό ίδρυμα όσο και στο προσωπικό του, είτε πρόκειται για ανώτερα στελέχη είτε για απλούς υπαλλήλους. Τα διοικητικά πρόστιμα κυμαίνονται από 50.000 ευρώ έως 10 εκατομμύρια για το νομικό πρόσωπο και έως 75.000 ευρώ για τα φυσικά πρόσωπα που εμπλέκονται. Τα πιστωτικά ιδρύματα κινδυνεύουν να κατηγορηθούν για βαριά αμέλεια ή συνέργεια στην πράξη της νομιμοποίησης με πιθανό επακόλουθο την απαγόρευση άσκησης ορισμένων επιχειρηματικών δραστηριοτήτων, τον αποκλεισμό από ενισχύσεις, αναθέσεις έργων και διαγωνισμούς του ελληνικού Δημοσίου, την απαγόρευση ίδρυσης νέων υποκαταστημάτων, την απαγόρευση αύξησης μετοχικού κεφαλαίου και την προσωρινή ή ακόμα οριστική ανάκληση της άδειας λειτουργίας τους. Οι υπάλληλοι που είναι υπόχρεοι για την αναφορά ύποπτων ή ασυνήθιστων συναλλαγών αν παραλείψουν να τις αναφέρουν ή παρουσιάσουν ψευδή και παραπλανητικά στοιχεία τιμωρούνται με ποινή φυλάκισης ως δύο έτη για αμέλεια και ως 10 έτη για συνέργεια. Τα περιουσιακά στοιχεία, όπως λογαριασμοί, θυρίδες και τίτλοι που αποτελούν προϊόντα εγκληματικής δραστηριότητας, δεσμεύονται.

Την τελευταία δεκαετία τεράστια ποσά έχουν διατεθεί για την ενίσχυση των ελέγχων κατά του οικονομικού εγκλήματος. Παρόλα αυτά όμως μεγάλοι τραπεζικοί όμιλοι εξακολουθούν να αποτυγχάνουν παρά το αυστηρό νομοθετικό πλαίσιο και να υφίστανται σοβαρές κυρώσεις και χρηματικά πρόστιμα (International Compliance Association, 2022). Χαρακτηριστικό παράδειγμα είναι το πρόστιμο 107 εκατομμυρίων λιρών που επιβλήθηκε στη Santander από την Αρχή Χρηματοοικονομικής Συμπεριφοράς (FCA) του Ηνωμένου Βασιλείου διότι υπήρχαν σημαντικά και επίμονα κενά και ελλείψεις σε όλο το πλαίσιο ελέγχου (FCA, 2022). Μια άλλη πολύκροτη υπόθεση αφορά στη NatWest με την FCA να επιβάλλει όχι μόνο πρόστιμο 264,8 εκατ. λιρών για παραλείψεις αλλά και να ασκεί ποινικές διώξεις κατά του χρηματοπιστωτικού ιδρύματος. Άλλο ένα μεγάλο πρόστιμο καταλογισμένο επίσης από την FCA, αφορά στην HSBC για παραλείψεις σε σχέση με την παρακολούθηση των συναλλαγών, επιβάλλοντας πρόστιμο ύψους 63,9 εκατομμυρίων λιρών.

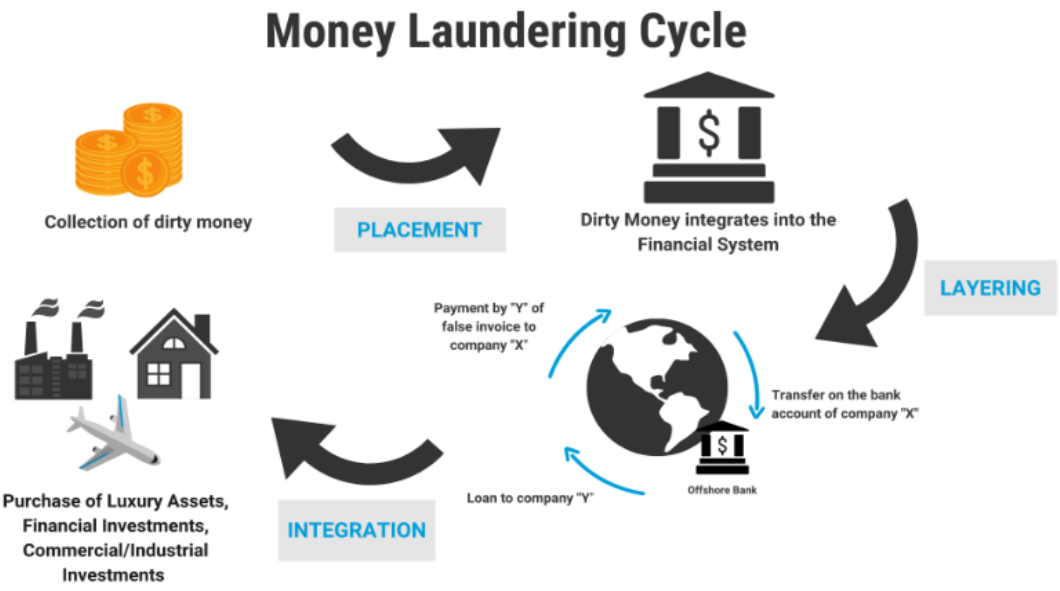
Ενότητα 2 Πρόληψη και καταστολή Ξεπλύματος Χρήματος και Χρηματοδότησης της τρομοκρατίας

2.1 Ξέπλυμα Βρώμικου Χρήματος και Χρηματοδότηση της Τρομοκρατίας

Το «βρώμικο χρήμα» πολλές φορές συγχέεται με την «χρηματοδότηση της τρομοκρατίας». Αποτελούν και τα δύο οικονομικά εγκλήματα με βαρύ αντίκτυπο σε κοινωνικό και οικονομικό επίπεδο. Σκοπός των εγκληματιών είναι να αποκρύψουν την πραγματική προέλευση των χρημάτων προκειμένου να πετύχουν τους σκοπούς τους. Η μεγαλύτερη όμως διαφορά συνίσταται στο κίνητρο που στην περίπτωση της τρομοκρατίας είναι ιδεολογικό ενώ στην περίπτωση του ξεπλύματος βρώμικου χρήματος είναι κερδοσκοπικό. Το ξέπλυμα βρώμικου χρήματος θεωρείται έγκλημα δευτέρου βαθμού, δηλαδή συντελείται ως απόρροια κάποιου άλλου βασικού εγκλήματος (Παράρτημα, Βασικά Αδικήματα, Πίνακας 1).

Στην «χρηματοδότηση της τρομοκρατίας» τα ποσά που διακινούνται είναι συνήθως μικρά, οι συναλλαγές που συνδέονται με αυτά μπορεί να μην είναι σύνθετες και δύναται να προέρχονται τόσο από παράνομες πηγές όσο ακόμα και από νόμιμες όπως για παράδειγμα από φιλανθρωπικές οργανώσεις ή έσοδα από επιχειρήσεις. Η διαδικασία χρηματοδότησης είναι συνήθως γραμμική (συγκέντρωση των πόρων, αποθήκευση, μετακίνηση και χρήση) και τα χρήματα που παράγονται χρησιμοποιούνται για τη συντήρηση των τρομοκρατικών ομάδων και των δραστηριοτήτων τους (ΟΗΕ, 2022).

Αντίθετα ένα σύστημα νομιμοποίησης εσόδων από παράνομη δραστηριότητα είναι κατά κανόνα κυκλικό και τα χρήματα καταλήγουν τελικά στην οντότητα που τα δημιούργησε. Η κύρια ανάγκη των διακινητών είναι να αποκρύπτεται συνεχώς η προέλευση του χρήματος, να διατηρούν διαρκώς τον έλεγχο σε αυτό και να μεταβάλλουν την μορφή του. Στην προσπάθεια αυτή είναι διατεθειμένοι να χάσουν έως και το 40% του αρχικού ποσού προκειμένου να το νομιμοποιήσουν.



Εικόνα 1: Ο κύκλος του Ξεπλύματος Χρήματος

Πηγή: United Nations

https://www.unodc.org/unodc/en/money-laundering/overview.html

Η διαδικασία κατά την οποία οι εγκληματίες προσδίδουν νομιμοφάνεια στα έσοδα από παράνομες δραστηριότητες χωρίζεται σε τρία κύρια στάδια.

- **Τοποθέτηση (Placement):** στο πρώτο στάδιο γίνεται η εισχώρηση αυτών των κεφαλαίων για πρώτη φορά μέσα στο χρηματοπιστωτικό σύστημα με την κατάθεση μετρητών σε κάποιον λογαριασμό. Οι τεχνικές που συνήθως χρησιμοποιούνται είναι η ανάμειξη παράνομων εσόδων με νόμιμα ή η αποδόμηση των καταθέσεων σε ποσά μικρότερα των ορίων που έχουν καθοριστεί για την υποβολή σχετικών αναφορών, πληρωμές σε μετρητά και ανταλλαγές χαρτονομισμάτων. Σε αυτό το στάδιο ο εντοπισμός είναι εύκολος και είναι καθοριστικής σημασίας η λειτουργία ισχυρού AML μηχανογραφικού συστήματος ταυτόχρονα με την επαγρύπνηση και παρατηρητικότητα των υπαλλήλων πρώτης γραμμής του πιστωτικού ιδρύματος.
- **Διαστρωμάτωση (Layering):** στην συνέχεια τα κεφάλαια μετακινούνται και εναλλάσσονται διαρκώς μέσω πολλαπλών χρηματοοικονομικών συναλλαγών προσπαθώντας να χαθεί η αρχική προέλευσή τους και να γίνει δύσκολα ανιχνεύσιμη

η διαδρομή τους. Αυτό επιτυγχάνεται με ανταλλαγές τίτλων για μεγαλύτερα ή μικρότερα ποσά, έκδοση επιταγών, διενέργεια εμβασμάτων από και προς διάφορους λογαριασμούς σε ένα ή περισσότερα πιστωτικά ιδρύματα. Σε αυτό το στάδιο ο εντοπισμός είναι πιο δύσκολος, μόνο με την συνδρομή των AML συστημάτων και την καλή γνώση της αναμενόμενης δραστηριότητας των συναλλασσομένων είναι εφικτός ο εντοπισμός τους.

- **Ολοκλήρωση (Integration):** σε αυτό το τελικό στάδιο πραγματοποιείται η ενσωμάτωση των κεφαλαίων στην νόμιμη οικονομική δραστηριότητα, με αγοροπωλησίες ακινήτων, επενδύσεις σε τίτλους και επιχειρήσεις, με αποπληρωμή δανείων. Σε αυτό το στάδιο ο εντοπισμός των κεφαλαίων είναι σχεδόν αδύνατος.

Τα τρία στάδια δεν είναι πάντα σαφώς διαχωρισμένα χρονικά, μπορεί να συνδυάζονται ή να επαναλαμβάνονται πολλές φορές.

Το ξέπλυμα βρώμικου χρήματος είναι ένα εξαιρετικά σύνθετο φαινόμενο με διεθνικό χαρακτήρα. Αποτελεί τη διάπραξη μιας σειράς ενεργειών, με πολύπλοκα τεχνάσματα και εργαλεία που αρκετές φορές τις καθιστούν δύσκολο να διακριθούν από τις νόμιμες συναλλαγές που προκύπτουν από την κανονική εμπορική δραστηριότητα. Εξελίσσεται με απίστευτη ταχύτητα, χρησιμοποιώντας εκτεταμένους χρηματοοικονομικούς και τεχνολογικούς πόρους, σε διαφορετικές χώρες με περισσότερες από μια δικαιοδοσίες, εκμεταλλευόμενο τυχόν νομοθετικά κενά και παραλείψεις ή αστοχίες κατά την εκτέλεση εποπτείας εκ μέρους του χρηματοπιστωτικού συστήματος (Κωνσταντόπουλος, 2020).

Το μέγεθος του προβλήματος αποτελεί πραγματική πρόκληση δεδομένου ότι λόγω της παγκοσμιοποίησης η ελεύθερη και ταχύτατη διακίνηση κεφαλαίων, η ευρεία χρήση του διαδικτύου και η υιοθέτηση νέων τεχνολογιών στις ήδη πολύπλοκες σύγχρονες οικονομίες είναι ο κανόνας. Η ευελιξία και εφευρετικότητα των εγκληματικών οργανώσεων έχει οδηγήσει σε αναθεωρήσεις των σχετικών διατάξεων για την αντιμετώπιση του φαινομένου (Κούβαρης, 2015).

Δυστυχώς παρά την εκτενή μελέτη του φαινομένου από τους ερευνητές, το αυστηρό νομοθετικό πλαίσιο που έχει τεθεί και τους εξαντλητικούς ελέγχους που διεκπεραιώνονται από τις διωκτικές αρχές, συνεχίζει να αποτελεί κίνδυνο για την ομαλή λειτουργία των σύγχρονων οικονομιών. Λόγω της παράνομης φύσης του δεν μπορεί να αποτυπωθεί ακριβώς

η πραγματική έκταση του φαινομένου. Ωστόσο σύμφωνα με έκθεση του Οργανισμού Ηνωμένων Εθνών εκτιμάται ότι το πόσο των χρημάτων που ξεπλένονται παγκοσμίως σε ετήσια βάση είναι της τάξης του 2 -5 % του παγκόσμιου ΑΕΠ ή 800 δισεκατομμύρια έως 2 τρισεκατομμύρια δολάρια.

2.2 Ο ρόλος των πιστωτικών ιδρυμάτων

Τα πιστωτικά ιδρύματα αποτελούν σήμερα το κυριότερο σύστημα συναλλαγών. Η ιδιαίτερη φύση του υλικού με το οποίο συναλλάσσονται (το χρήμα), οι εμπιστευτικές πληροφορίες που διαχειρίζονται και οι εξουσίες που κατέχουν, τα τοποθετούν σε ένα πλαίσιο το οποίο διέπεται από ειδικούς κανονισμούς σε εθνικό και διεθνές επίπεδο.

Πιο συγκεκριμένα η ευρωπαϊκή και διεθνής νομοθεσία υποχρεώνουν τις τράπεζες να πιστοποιούν και να διαθέτουν ακριβή γνώση των στοιχείων των πελατών, της δραστηριότητας και της οικονομικής τους κατάστασης. Πρέπει να ασκείται διαρκής παρακολούθηση των συναλλαγών οι οποίες εκτελούνται, έτσι ώστε να εντοπίζονται ασυνήθιστες ή ύποπτες συναλλαγές. Εξειδικευμένες μηχανογραφικές εφαρμογές βοηθούν στην ανίχνευση των ύποπτων δραστηριοτήτων, με την συνδρομή εκπαιδευμένων υπαλλήλων που είναι υπεύθυνοι για την αποφυγή ξεπλύματος χρήματος και χρηματοδότησης της τρομοκρατίας και οι οποίοι έχουν στενή συνεργασία με τις τοπικές αρχές ενημερώνοντας για οποιαδήποτε ύποπτη συναλλαγή.

Ειδικότερα η αποτελεσματικότητά της μάχης κατά του ξεπλύματος βρώμικου χρήματος συνοψίζεται στα εξής στάδια:

- <u>Ακριβής πιστοποίηση του πελάτη/συναλλασσόμενου</u>

  Είναι καθήκον να εξακριβώνεται η αληθινή ταυτότητα κάθε πελάτη πριν το άνοιγμα ενός λογαριασμού ή μιας θυρίδας ή την διενέργεια οποιασδήποτε συναλλαγής (όπως ανταλλαγής από ένα νόμισμα σε κάποιο άλλο).

  Όταν είναι εμφανές ότι ο πελάτης δεν ενεργεί για λογαριασμό του είναι απαραίτητο να παρθούν πληροφορίες για τον τελικό κάτοχο του λογαριασμού ή τον πραγματικό

δικαιούχο (beneficiary owner) της συναλλαγής. Για νομικά πρόσωπα, εκτός της περίπτωσης εισηγμένων εταιριών σε οργανωμένη αγορά που ούτως ή άλλως υπόκεινται σε σχετικές γνωστοποιήσεις, είναι απαραίτητο να προσκομίζονται τα προβλεπόμενα νομιμοποιητικά έγγραφα που αφορούν την εταιρεία, τους βασικότερους μετόχους (με ποσοστό άνω του 25%) ή τα φυσικά πρόσωπα που ελέγχουν με οποιοδήποτε τρόπο την λειτουργία της εταιρείας και τον νόμιμο εκπρόσωπο.

- <u>Επαρκής γνώση της δραστηριότητας του πελάτη και της οικονομικής του κατάστασης</u>

  Αυτή η γνώση είναι η μόνη που επιτρέπει τον σχηματισμό του οικονομικού-συναλλακτικού προφίλ του πελάτη και είναι απαραίτητη για την αναγνώριση της κανονικής και αναμενόμενης δραστηριότητας από αυτόν. Έχει εφαρμογή στα φυσικά πρόσωπα σχετικά με τις επαγγελματικές τους δραστηριότητες, τα εισοδήματα και τα περιουσιακά τους στοιχεία. Αντίστοιχα για τις εταιρείες, συγκεντρώνονται δεδομένα που αφορούν τη γεωγραφική ζώνη και το εύρος των δραστηριοτήτων τους καθώς και τις αναλυτικές οικονομικές τους καταστάσεις.

- <u>Παρακολούθηση των συναλλαγών για να οριστούν ασυνήθιστες[1] η ύποπτες[2] συναλλαγές</u>

  Η παρακολούθηση βασίζεται τόσο στη φύση όπως και στον όγκο των συναλλαγών έτσι ώστε να μπορεί να διευκρινιστεί αν οι υπό εκτέλεση ή οι ολοκληρωμένες συναλλαγές είναι συμβατές με το αναμενόμενο προφίλ της δραστηριότητας του εκάστοτε πελάτη. Διεξοδικά εξετάζεται η πηγή προέλευσης και προορισμού των εισερχόμενων κεφαλαίων. Οι τράπεζες έχουν χρέος να αρνούνται οποιαδήποτε συναλλαγή αναλαμβάνεται εκ μέρους πελάτη ή αντισυμβαλλομένου όταν δεν μπορεί να εξακριβωθεί η οικονομική νομιμότητα της συναλλαγής. Η Τράπεζα της Ελλάδος εκδίδει συχνά ενδεικτικό πίνακα με την τυπολογία ύποπτων συναλλαγών.

Υποσημείωση 1: ασυνήθης συναλλαγή είναι αυτή που δεν φαίνεται να σχετίζεται, ως προς την φύση ή το ποσό, με την ακίνητη περιουσία και το εισόδημα του πελάτη ή με τις επαγγελματικές δραστηριότητες ή τις συνήθειές του ή που δεν έχει προφανή σκοπό

Υποσημείωση 2: ύποπτη συναλλαγή είναι αυτή από την οποία προκύπτουν ενδείξεις ή υπόνοιες απόπειρας ή διάπραξης εγκληματικής δραστηριότητας

Σε συνέχεια των πιο πάνω, επιπλέον σημεία που χρίζουν ιδιαίτερης προσοχής είναι τα επόμενα:

i. Οι τράπεζες οφείλουν να έχουν λεπτομερή γνώση ακόμα και των εν δυνάμει πελατών και συνεργατών τους. Η εφαρμογή των διαδικασιών δέουσας επιμέλειας ξεκινά πριν την έναρξη οποιασδήποτε συνεργασίας δεδομένου ότι απαγορεύονται οι σχέσεις με φυσικά ή νομικά πρόσωπα των οποίων οι δραστηριότητες μπορεί να θεωρηθούν παράνομες ή αντίθετες με τις αρχές της τράπεζας. Όλα τα φυσικά ή νομικά πρόσωπα με τα οποία οι τράπεζες έχουν επαγγελματική σχέση πρέπει να πιστοποιούνται, να αξιολογούνται και να εγκρίνονται με την πρόβλεψη της κατάλληλης τεκμηρίωσης. Η έννοια της επαγγελματικής σχέσης καλύπτει κάθε πρόσωπο ή εταιρεία με την οποία υπάρχει ανταλλαγή οικονομικών ροών, όχι μόνο δηλαδή με πελάτες με την αυστηρή έννοια του όρου αλλά επίσης και επαγγελματίες αντισυμβαλλόμενοι, ανεξάρτητοι οικονομικοί σύμβουλοι και πολλοί άλλοι.

ii. Σε πραγματικό χρόνο οφείλεται να πραγματοποιείται έλεγχος και διασταύρωση των στοιχείων πελατών και συναλλασσομένων ως προς τις εκδοθείσες λίστες από τις αρχές με τα ονόματα ατόμων ή οργανισμών που είναι ύποπτοι για τρομοκρατία και ξέπλυμα βρώμικου χρήματος. Το σύστημα ανίχνευσης χρησιμοποιείται και για τις διεθνείς μεταφορές κεφαλαίων όπου επίσης οι τράπεζες είναι υποχρεωμένες να αρνούνται την διεκπεραίωση συναλλαγών και να παγώνουν τα κεφάλαια των οντοτήτων που εμφανίζονται στις επίσημες καταστάσεις των υπόπτων για τρομοκρατία.

iii. Ειδική μέριμνα πρέπει να δίνεται σε πελάτες χωρίς φυσική παρουσία στην τράπεζα, με τους οποίους η επικοινωνία γίνεται τηλεφωνικά ή ηλεκτρονικά. Επίσης συγκεκριμένοι κανόνες έχουν ισχύ για τους περιστασιακούς πελάτες (διερχόμενοι ή πελάτες άλλων καταστημάτων).

Οι τράπεζες οφείλουν να διατηρούν αξιόπιστη βάση για τα δεδομένα του πελατολογίου

τους και να μεριμνούν για τη συνεχή επικαιροποίηση αυτών. Στην Ελλάδα, η Ένωση Ελληνικών Τραπεζών έχει εκδώσει σχετικό φυλλάδιο που εξηγεί στους πελάτες τι χρειάζεται να γνωρίζουν τα τραπεζικά ιδρύματα για αυτούς και γιατί είναι αναγκαία η παροχή αυτών πληροφοριών.

2.3 Προσέγγιση με βάση τον κίνδυνο (Risk based approach)

Είναι γενικά αποδεκτό ότι η πρόληψη της νομιμοποίησης προσόδων από παράνομη δραστηριότητα είναι πιο αποτελεσματική από ότι η δίωξη της δραστηριότητας μετά την διάπραξή της. Σε αυτήν την κατεύθυνση λοιπόν είναι σημαντικό να εντοπίζονται οι κίνδυνοι και να κατανοείται σε βάθος η φύση τους ενώ στην συνέχεια πρέπει να γίνεται προσπάθεια μετριασμού τους. Η προσέγγιση με βάση τον κίνδυνο περιλαμβάνει την αναπροσαρμογή της εποπτικής δράσης με την υιοθέτηση κατάλληλων στρατηγικών ανάλογα με τους εκτιμώμενους κινδύνους. Αυτές σκοπό έχουν να στερήσουν την ευκαιρία από τους εγκληματίες να ξεπλύνουν τα παράνομα έσοδά τους ενώ παράλληλα θα βελτιώσουν την διάχυση και την ποιότητα των πληροφοριών που είναι διαθέσιμες στις εποπτικές αρχές. Ζωτικής σημασίας είναι να διασφαλιστεί ότι η εποπτική δράση θα εστιάσει σε τομείς που ο κίνδυνος απάτης είναι υψηλότερος χωρίς να επιβαρύνει αδικαιολόγητα τις οντότητες και τις δραστηριότητες χαμηλότερου κινδύνου (FATF, 2021).

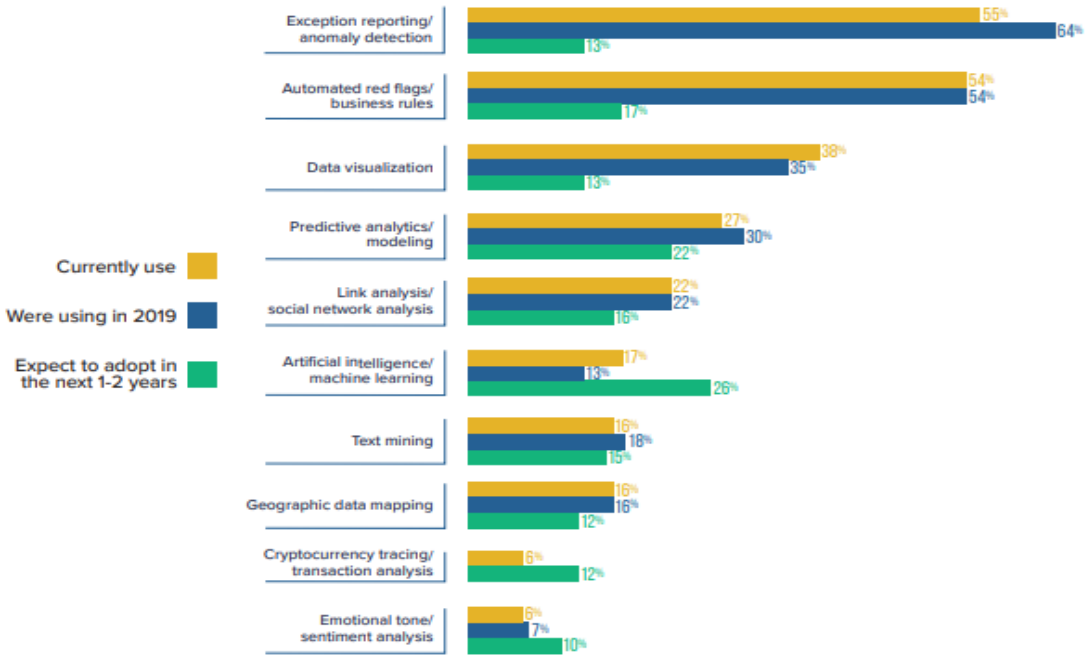
Η μετάβαση από την παραδοσιακή μέχρι σήμερα προσέγγιση «βασισμένη σε κανόνες» σε μια προσέγγιση «βασισμένη στον κίνδυνο» απαιτεί χρόνο, τεχνολογικούς και οικονομικούς πόρους και εξειδικευμένο προσωπικό. Αν ένα έγκλημα δεν παραβιάζει κάποιο προδηλωμένο κανόνα και δεν ανταποκρίνεται σε γνωστό πρότυπο δεν θα ανιχνευτεί. Εισάγεται οπότε ο όρος συνεχής-δυναμική παρακολούθηση (monitoring) για να περιγράψει το πλήρες φάσμα ενεργειών για την παρατήρηση αλλαγών στο προφίλ κινδύνου ή τον εντοπισμό αποκλίνουσας συμπεριφοράς. Αποτελεί μια εξελισσόμενη διαδικασία διότι οι δραστηριότητες των πελατών αλλά και τα προσφερόμενα προϊόντα και υπηρεσίες των τραπεζών μεταβάλλονται διαρκώς. Διαχωρίζεται σαφώς από τον όρο εποπτεία που επικεντρώνεται στην εφαρμογή μέτρων δέουσας επιμέλειας και περιοδικών αναθεωρήσεών

τους. Παραδοσιακά οι πελάτες των πιστωτικών ιδρυμάτων στην Ελλάδα κατατάσσονται σε τρεις κατηγορίες: χαμηλού, κανονικού και υψηλού κινδύνου. Οι πελάτες της τρίτης κατηγορίας αξιολογούνται τουλάχιστον σε ετήσια βάση, στο πλαίσιο αν απαιτείται να διακόψει το πιστωτικό ίδρυμα την μεταξύ τους συνεργασία (Παράρτημα, Κατάταξη νέων και υφιστάμενων πελατών στην κατηγορία υψηλού κινδύνου, Πίνακας 2). Μια πιο ολιστική προσέγγιση υπαγορεύει την εκτίμηση κινδύνου ακόμα και σε επίπεδο οντότητας με έναν συνδυασμό κριτηρίων. Ενδεικτικά για νομικά πρόσωπα θα μπορούσε να ληφθεί υπόψη η σημαντικότητα της οντότητας στον τομέα που δραστηριοποιείται και το μερίδιο αγοράς που κατέχει, το επίπεδο διαφάνειας και το παρελθόν της, η ύπαρξη σύνθετης δομής ιδιοκτησίας, η πολυπλοκότητα των συναλλαγών της σε σχέση με προϊόντα ή υπηρεσίες και κανάλια διανομής, η στρατηγική της (επεκτάσεις σε νέους τομείς ή συγχωνεύσεις και εξαγορές), γεωγραφική εξάπλωση των εργασιών με αξιολόγηση του νομοθετικού πλαισίου στις χώρες στις οποίες έχει παρουσία (χώρες υψηλού κινδύνου, φορολογικοί παράδεισοι, χώρες με διαφορετικές δικαιοδοσίες).

Ενότητα 3 Συστήματα AML

3.1 Εξέλιξη σύγχρονων συστημάτων AML

Με την αλλαγή της χιλιετίας παρατηρήθηκε το φαινόμενο τα δεδομένα να επεκτείνονται ταχύτερα από ότι η απαιτούμενη τεχνολογία για τη διαχείρισή τους. Τα συμβατικά συστήματα KYC και AML που μέχρι πρότινος χρησιμοποιήθηκαν από τα πιστωτικά ιδρύματα και τους περισσότερους οργανισμούς, στήριξαν την λειτουργία τους στην ανίχνευση ανωμαλιών και στον εντοπισμό εξαιρέσεων, καθώς και στην αυτοματοποίηση ελέγχων βάση επιχειρηματικών κανόνων. Επιπλέον περιλάμβαναν πλήθος από χρονοβόρες και χειροκίνητες διαδικασίες, όπως την επαλήθευση της απασχόλησης ή των στοιχείων επικοινωνίας κατά την αποδοχή των νέων πελατών ή ακόμα τον έλεγχο γνησιότητας των συνυποβαλλόμενων εγγράφων.



Εικόνα 2: Χρήση τεχνολογιών ανάλυσης δεδομένων για την καταπολέμηση της απάτης

Πηγή: SAS – ACFE (2022), 2022 Anti-Fraud Technology benchmarking report

Αποδεικνύεται ότι ακόμα και σήμερα η ανίχνευση της απάτης παραμένει μια δύσκολη υπόθεση. Όμως ο εντοπισμός τάσεων ή μοτίβων κινδύνου και ο μετριασμός τους είναι

εφικτός με τη χρήση προηγμένων τεχνολογιών ανάλυσης όπως της μηχανικής μάθησης, της επιστήμης των μεγάλων δεδομένων (big data), της μηχανικής όρασης (ανάλυση δεδομένων από βίντεο και φωτογραφίες) και της ανάλυσης δεδομένων δικτύων. Σύμφωνα με πρόσφατη έρευνα της κορυφαίας εταιρίας παροχής λογισμικού SAS και της ACFE (Association of Certified Fraud Examiners) το ποσοστό των οργανισμών που έχουν υιοθετήσει μεθόδους βασισμένες στην τεχνητή νοημοσύνη είναι 17% ενώ ένα επιπλέον ποσοστό 26% αναμένεται να επενδύσει σε αυτές τις τεχνολογίες στα επόμενα δύο χρόνια (ως τέλος του 2023).

Με την χρήση της τεχνητής νοημοσύνης είναι εφικτή η άντληση και ανάλυση μη δομημένων δεδομένων από πολλαπλές πηγές όπως κυβερνητικά αρχεία, δεδομένα τρίτων μερών και μέσα κοινωνικής δικτύωσης. Με αυτόν τον τρόπο καθίσταται δυνατή η ενδελεχής εξέταση και καλύτερη κατανόηση του συνόλου των διαθέσιμων στοιχείων και χαρτογραφείται ένα δίκτυο συνδέσμων σύμφωνα με την λογική των αρχών KYC και KYCC (Know Your Customer's Customers), δημιουργώντας πολύ-επίπεδα για τις σχέσεις μεταξύ των πελατών και όλων των οντοτήτων με τις οποίες αλληλεπιδρούν (προμηθευτές και λοιποί συνεργάτες). Σκιαγραφείται με λεπτομέρεια το προφίλ κάθε συναλλασσόμενου σε τέτοιο βαθμό ώστε τα συστήματα AML να εκπαιδεύονται και να μοντελοποιούν την αναμενόμενη δραστηριότητα και να προειδοποιούν για άτυπες συμπεριφορές πολύ πριν τα παραδοσιακά συστήματα παρακολούθησης συναλλαγών εντοπίσουν τις εκτός ορίων συναλλαγές.

Οι μέθοδοι εξόρυξης δεδομένων που χρησιμοποιούνται ευρύτατα και εν μέρει θα παρουσιαστούν στο Β΄ Μέρος της παρούσας εργασίας αποτελούν κυρίως αλγόριθμους κατηγοριοποίησης όπως νευρωνικά δίκτυα τύπου Multilayer Perceptron, μηχανές διανυσμάτων υποστήριξης και δέντρα αποφάσεων. Επίσης μέθοδοι μη επιβλεπόμενης μάθησης όπως ανάλυση συστάδων, ανακάλυψη κανόνων συσχέτισης και ανάλυση εξαιρέσεων, βοηθούν στον εντοπισμό ανώμαλης δραστηριότητας. Επιπλέον οι τεχνικές εξόρυξης κειμένου (text mining) και επεξεργασίας φυσικής γλώσσας (NLP) μπορούν να χρησιμοποιηθούν για την ανάλυση μη δομημένων πηγών δεδομένων (όπως νέες κανονιστικές ρυθμίσεις, άρθρα ειδήσεων) με σκοπό την εξαγωγή σχετικών πληροφοριών που θα ενισχύσουν τα δομημένα δεδομένα συναλλαγών με πρόσθετο περιεχόμενο.

3.2 Βέλτιστες πρακτικές

- Υιοθέτηση θετικής νοοτροπίας για συνέργειες και ελεύθερη διάχυση της πληροφορίας

Έρευνες αποδεικνύουν πως όταν οι εταιρείες συνεργάζονται και υπάρχει διάχυση της πληροφόρησης σχετικά με συναλλαγές και προφίλ υψηλού κινδύνου, υπάρχει 25% μεγαλύτερη πιθανότητα να προβλεφθεί η περίπτωση απάτης συγκριτικά με όταν κάθε εταιρεία λειτουργεί μεμονωμένα. Στο πνεύμα αυτό δημιουργήθηκε ερευνητικά η πλατφόρμα ανταλλαγής δεδομένων Integrity Distributed InDi από μια κοινοπραξία εταιριών κολοσσών στην παγκόσμια οικονομία, επαγγελματιών της επιστήμης των δεδομένων, κορυφαίων νομικών συμβούλων και με επικεφαλή ένα μη κερδοσκοπικό οργανισμό του Τεχνολογικού Ινστιτούτου της Μασαχουσέτης MIT. Η πλατφόρμα επιτρέπει στους οργανισμούς επαναλαμβανόμενα να εκπαιδεύουν αλγόριθμους που ανιχνεύουν μοτίβα απάτης και διαφθοράς στους αντίστοιχους κλάδους τους. Αυτούς τους ίδιους αλγόριθμους στην συνέχεια συνεισφέρουν στην κοινοπραξία, δημιουργώντας ένα υπέρ-μοντέλο νευρωνικών δικτύων, διασφαλίζοντας έτσι το απόρρητο και την ανωνυμία των δεδομένων. Το MIT αποκαλεί αυτήν την τεχνική διαχωρισμένη μάθηση (split learning) λόγω του ότι δεν διαμοιράζονται τα αρχικά εμπορικά δεδομένα (Walden, 2023).

- Real time ανάλυση του συνόλου των διαθέσιμων δεδομένων

Η δυναμική ανάλυση οντοτήτων (dynamic entity resolution) είναι ακόμα μια τεχνολογία όπου συνδυάζονται σε πραγματικό χρόνο πολλαπλά σύνολα δεδομένων, που προέρχονται από διαφορετικά συστήματα, εσωτερικά της επιχείρησης ή εξωτερικά όπως το Web 2.0 και τα κοινωνικά δίκτυα. Επιτρέπει ακόμα την επεξεργασία δεδομένων από παλαιότερα συστήματα όπου η ποιότητα των δεδομένων μπορεί να είναι χαμηλή. Σκοπό έχει την δημιουργία δικτύων με την χρήση γραφημάτων που απεικονίζουν τα μοτίβα που καθοδηγούν τις σχέσεις και τη συμπεριφορά και τις προθέσεις των συναλλασσομένων. Έτσι δημιουργείται μια ενοποιημένη εικόνα για κάθε πελάτη που βοηθά στην βαθύτερη κατανόηση των δραστηριοτήτων του (Gross, 2021).

- Χρήση της τεχνητής νοημοσύνης για βελτίωση της ποιότητας και πληρότητας των δεδομένων

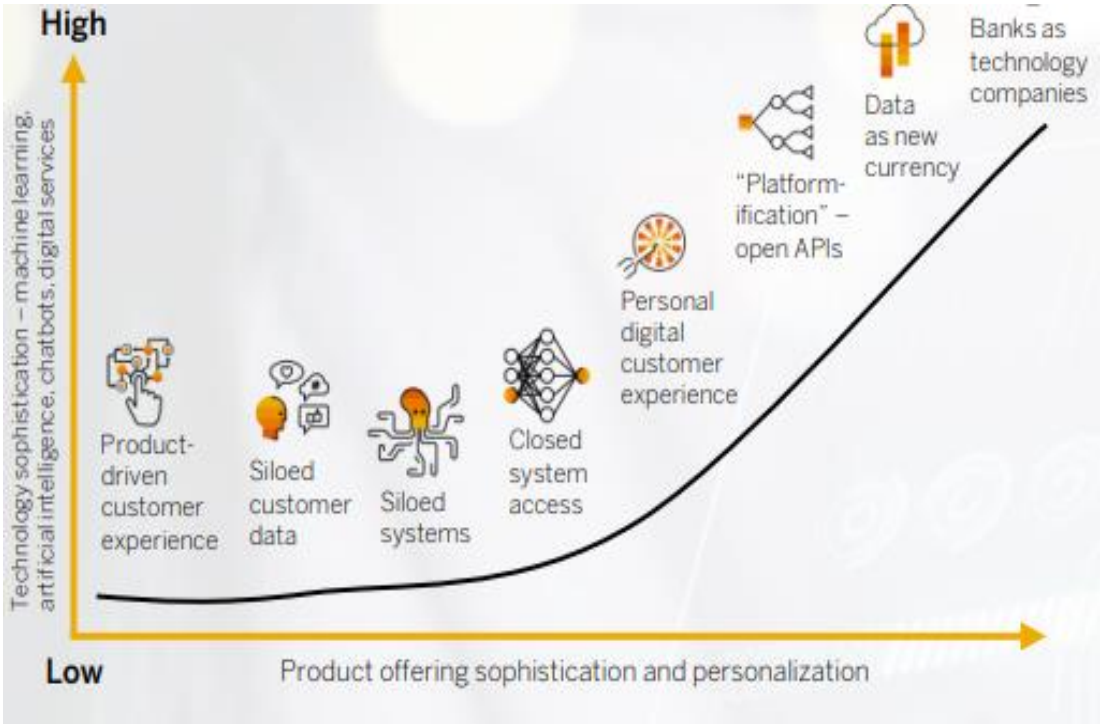
Επιπλέον με την ταχεία επέλαση της τεχνητής νοημοσύνης τα πιστωτικά ιδρύματα είναι σε θέση να αποφασίζουν άμεσα αν θα αποδεχθούν ένα νέο πελάτη, είναι ικανά να ανιχνεύουν την απάτη κατά την υποβολή αιτήσεων σε οποιοδήποτε τύπο καναλιών και να διασφαλίζουν ότι μόνο επαληθευμένα και τυποποιημένα δεδομένα πελατών εισέρχονται στο σύστημα. Πιο συγκεκριμένα με την χρήση μιας μορφής τεχνητής νοημοσύνης, της σημασιολογικής τεχνολογίας (semantic technology) είναι εφικτή η αυτοματοποιημένη αναγνώριση προτύπων στα δεδομένα. Η σημασιολογική τεχνολογία συνδέει τις λέξεις με έννοιες και αναγνωρίζει τις σχέσεις μεταξύ τους. Επιτρέπει στις τραπεζικές πλατφόρμες να συμπληρώνουν τυχόν ελλιπείς ή λανθασμένες τιμές. Αυτό έχει αποτέλεσμα να βελτιώνουν την ποιότητα των δεδομένων που εισάγονται κατά την υποβολή μιας νέας αίτησης πελάτη. Έτσι μειώνεται η πιθανότητα λάθους εκτίμησης κινδύνου και απελευθερώνεται χρόνος και πολύτιμοι πόροι, επιτρέποντας τους αξιολογητές να επικεντρωθούν μόνο σε εκείνους τους τομείς που απαιτείται η ανθρώπινη κρίση και εμπειρία (Maitino, 2020).

- Δημιουργία μιας κοινής γλώσσας

Σε συνέχεια των πιο πάνω έχουν διατυπωθεί ανησυχίες για το πόσο παραγωγική θα μπορούσε να είναι η συνεργασία οργανισμών και η ελεύθερη διάχυση της πληροφορίας μεταξύ τους. Μεγάλο εμπόδιο αποτελεί η πολυπλοκότητα της έρευνας και ο τεράστιος όγκων των στοιχείων προς ανάλυση. Στοιχεία από ένα και μόνο αδίκημα μπορεί να είναι διασκορπισμένα σε πλήθος πληροφοριακών συστημάτων και άλλων πηγών. Επιπλέον το έργο δυσχεραίνουν οι διαφορετικές δικαιοδοσίες και οι λεκτικοί φραγμοί στην επεξεργασία των δεδομένων γιατί προφανώς δεν είναι δυνατό να καταγράφονται σε μια κοινή γλώσσα. Για να ξεπεραστεί αυτός ο περιορισμός ερευνητές προτείνουν την ανάπτυξη οντολογίας, ενός κοινού δηλαδή λεξιλογίου μεταξύ λογισμικών και των χρηστών τους, όπου τα ακατέργαστα δεδομένα θα διαχωρίζονται από τις έννοιές τους και θα μοντελοποιούνται τομείς γνώσης ή λόγου (Carvalho et al., 2015).

- Υπηρεσίες open banking

Με τις υπηρεσίες open banking εγκεκριμένα τρίτα μέρη μπορούν να έχουν ανοικτή πρόσβαση, μέσω ασφαλούς περιβάλλοντος API, σε πληροφορίες όπως τραπεζικά δεδομένα πελατών, με την προϋπόθεση ότι έχουν συναινέσει με την παροχή της απαραίτητης συγκατάθεσής τους, διευκολύνοντας τις εμπορικές συνεργασίες και προάγοντας την διαφάνεια στις συναλλαγές. Για αυτόν τον σκοπό, από το 2018 στην Ελλάδα εφαρμόζεται η ευρωπαϊκή οδηγία Payment Services Directive PSD2 που αφορά τις υπηρεσίες πληρωμών και η οποία ενσωματώθηκε στην ελληνική νομοθεσία με τον Νόμο 4537/2018. Σχετική εργασία για την δημιουργία οντολογίας στην υπηρεσίες ανοικτής τραπεζικής έχει παρουσιαστεί από ομάδα ερευνητών (Paneque et al., 2023).



Εικόνα 3: Μετασχηματισμός του τραπεζικού κλάδου

Πηγή: SAP, 2022
https://www.sap.com/greece/documents/2016/03/58bd8fd0-627c-0010-82c7-eda71af511fa.html

Κορυφαίες εταιρίες κατασκευής λογισμικού και μεγάλοι τραπεζικοί όμιλοι έχουν προσανατολιστεί προς τις παραπάνω κατευθύνσεις. Ενδεικτικά αναφέρουμε το παράδειγμα της εταιρίας SAP η οποία προβλέπει τον μετασχηματισμό των τραπεζικών ιδρυμάτων σε πλατφόρμες τεχνολογίας, ικανοποιώντας τόσο τις απαιτήσεις συμμόρφωσης αλλά κυρίως αναβαθμίζοντας την συνολική εμπειρία και την εξυπηρέτηση που θα απολαμβάνουν οι πελάτες τους, σε χρηματοοικονομικά προϊόντα ή άλλες συναφείς προσωποποιημένες υπηρεσίες που θα προσφέρονται.

ΒΙΒΛΙΟΓΡΑΦΙΑ ΜΕΡΟΥΣ Α΄

ΘΕΣΜΙΚΟ ΠΛΑΙΣΙΟ

ΠΔΤΕ 2577/9.3.2006 «Πλαίσιο αρχών λειτουργίας και κριτηρίων αξιολόγησης της οργάνωσης και των Συστημάτων Εσωτερικού Ελέγχου των πιστωτικών και χρηματοδοτικών ιδρυμάτων και σχετικές αρμοδιότητες των διοικητικών τους οργάνων»

Νόμος 4557/2018 «Πρόληψη και καταστολή της νομιμοποίησης εσόδων από εγκληματικές δραστηριότητες και της χρηματοδότησης της τρομοκρατίας (ενσωμάτωση της Οδηγίας 2015/849/ΕΕ)»

Νόμος 4734/2020 «Τροποποίηση του ν. 4557/2018 (Α’ 139) για την πρόληψη και καταστολή της νομιμοποίησης εσόδων από εγκληματικές δραστηριότητες και της χρηματοδότησης της τρομοκρατίας -Ενσωμάτωση στην ελληνική νομοθεσία της Οδηγίας (ΕΕ) 2018/843 (L 156) και του άρθρου 3 της Οδηγίας (ΕΕ) 2019/2177 (L 334)»

Νόμος 4816/2021 «Πρόληψη και καταστολή της νομιμοποίησης εσόδων από εγκληματικές δραστηριότητες και της χρηματοδότησης της τρομοκρατίας - Τροποποίηση του ν. 4557/2018 - Ενσωμάτωση της Οδηγίας (ΕΕ) 2018/1673 του Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου, της 23ης Οκτωβρίου 2018»

Απόφαση 281/5/17.03.2009 Επιτροπής Τραπεζικών και Πιστωτικών Θεμάτων (ΕΤΠΘ) για «την πρόληψη της χρησιμοποίησης των εποπτευομένων από την Τράπεζα της Ελλάδος πιστωτικών ιδρυμάτων και χρηματοπιστωτικών οργανισμών για τη νομιμοποίηση εσόδων από παράνομες δραστηριότητες και τη χρηματοδότηση της τρομοκρατίας»

Απόφαση 285/6/09.07.2009 Επιτροπής Τραπεζικών και Πιστωτικών Θεμάτων (ΕΤΠΘ) «Ενδεικτική τυπολογία ασύνηθων ή ύποπτων συναλλαγών»

Οδηγία (ΕΚ) 2018/1673: του Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου της 23ης Οκτωβρίου 2018 σχετικά με την καταπολέμηση της νομιμοποίησης εσόδων από παράνομες δραστηριότητες μέσω του ποινικού δικαίου

ΕΛΛΗΝΙΚΗ ΕΝΩΣΗ ΤΡΑΠΕΖΩΝ (2006), *Η νέα Πράξη του Διοικητή της Τράπεζας της Ελλάδος ΠΔ/ΤΕ 2577/9.3.2006 αναφορικά με τα Συστήματα Εσωτερικού Ελέγχου*. Αθήνα.

FATF (2019), Anti-money laundering and counter-terrorist financing measures – Greece, Fourth Round Mutual Evaluation Report, FATF, Paris http://www.fatf-gafi.org/publications/mutualevaluations/documents/mer-greece-2019.html

FATF (2021), Guidance on Risk-Based Supervision, FATF, Paris, www.fatf-gafi.org/publications/documents/Guidance-RBA-Supervision.html

Financial Conduct Authority (2021), Final Notice to HSBC Bank Plc, FCA, London https://www.fca.org.uk/publication/decision-notices/hsbc-bank-plc.pdf

Financial Conduct Authority (2022), Final Notice to Santander UK Plc, FCA, London https://www.fca.org.uk/publication/final-notices/santander-uk-plc-2022.pdf

Carvalho, R., Goldsmith, M., & Creese, S. (2015). Applying Semantic Technologies to Fight Online Banking Fraud. *2015 European Intelligence and Security Informatics Conference*, 61–68.

https://doi.org/10.1109/EISIC.2015.42

Gross, A. (2021). Understanding Connected Data Is the Key to Understanding Your Customer *Transforming Data with Intelligence*. Ανακτήθηκε 8 Απριλίου, 2023, από https://tdwi.org/articles/2021/04/09/bi-all-connected-data-key-to-understanding-customers.aspx

Maitino, P. (2020, 17 Ιανουαρίου). Banking on Semantic Technology: AI-Powered Data Quality Balances Fraud Prevention and Customer Excellence *Transforming Data with Intelligence*. Ανακτήθηκε 8 Απριλίου, 2023, από https://tdwi.org/articles/2020/01/17/diq-all-ai-powered-data-quality.aspx

Paneque, M., Roldán-García, M. del M., & García-Nieto, J. (2023). A Semantic Model for Enhancing Data-Driven Open Banking Services. *Applied Sciences*, *13*(3), Article 3. https://doi.org/10.3390/app13031447

Plenderleith, J. (2022, 19 Δεκεμβρίου). How to prevent AML failures. *International Compliance Association*. Ανακτήθηκε 20 Απριλίου, 2023, από: https://www.int-comp.org/insight/2022/december/how-to-prevent-aml-failures/

Walden, V. (2023, 1 Απριλίου). Unlocking the patterns of corrupt payments through MIT's data-sharing consortium known as Integrity Distributed *Fraud Magazine*. Ανακτήθηκε 8 Απριλίου, 2023, από https://www.fraud-magazine.com/article.aspx?id=4295020637

Κούβαρης Δημήτριος (2015 Πανεπιστήμιο Πειραιώς) Διεθνής τρομοκρατία και οργανωμένο οικονομικό έγκλημα σε ευρωπαϊκό και διεθνές επίπεδο. (n.d.). Retrieved 24 March 2023, from https://freader.ekt.gr/eadd/index.php?doc=43999#p=16

Κωνσταντόπουλος Βασίλειος (2020 Πάντειο Πανεπιστήμιο Κοινωνικών και Πολιτικών Επιστημών) Το ξέπλυμα χρήματος. (n.d.). Retrieved 24 March 2023, from https://freader.ekt.gr/eadd/index.php?doc=48487#p=13

ΜΕΡΟΣ Β΄ ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΠΡΟΗΓΟΥΜΕΝΗ ΕΡΕΥΝΑ

Ενότητα 1 Εξόρυξη γνώσης

1.1 Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων

"Information is the oil of the 21st century and analytics is the combustion engine."
Peter Sondergaard (2011), former vice president and global head of Research at Gartner, Inc

Επιχειρήσεις, άνθρωποι αλλά και μεμονωμένες ηλεκτρονικές συσκευές καθημερινά παράγουν τεράστιο όγκο δεδομένων, ο οποίος αυξάνει εκθετικά κάθε χρόνο. Για την επεξεργασία και την αξιοποίηση αυτών των δεδομένων χρησιμοποιούνται διάφορες επιστημονικές προσεγγίσεις και πλαίσια, πληθώρα από αλγόριθμους και διαδικασίες.

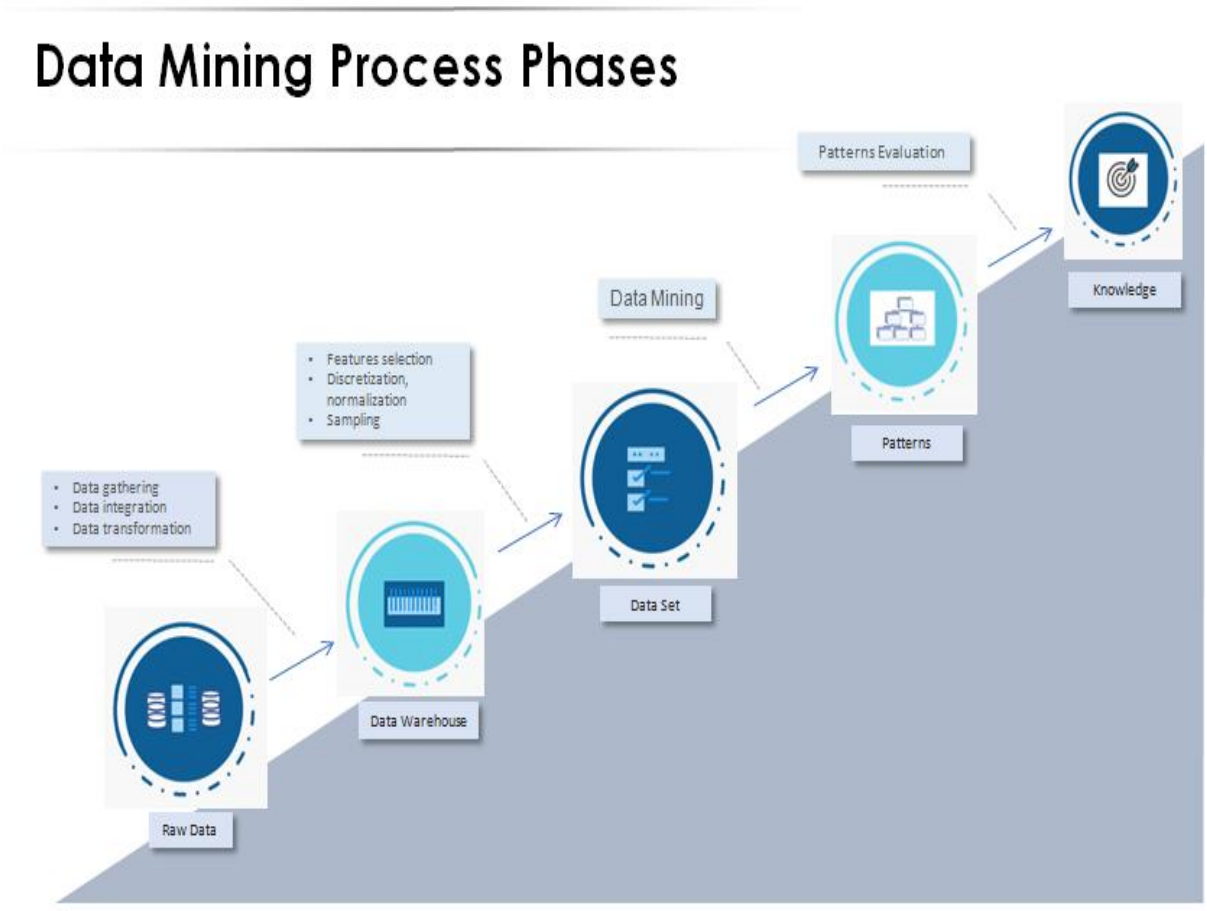
Είναι αμέτρητοι οι ορισμοί που έχουν δοθεί την τελευταία δεκαετία για να περιγράψουν το σύνολο των πρακτικών και των εργαλείων που απαιτούνται για την μετατροπή των ακατέργαστων δεδομένων σε σημαντική πληροφορία (Power et al., 2018). Ταχύτατα αναπτυσσόμενες εφαρμογές της επιστήμης της πληροφορικής έρχονται να συνεισφέρουν τα μέγιστα στο μεγάλο αντικείμενο της επεξεργασίας και ανάλυσης των δεδομένων, παρέχοντας σε επιχειρήσεις και μεγάλους οργανισμούς τους τέσσερις τύπους ανάλυσης και τελικά την πολύτιμη γνώση που χρειάζονται οι υπεύθυνοι λήψης αποφάσεων (Schniederjans et al., 2014).

Πιο συγκεκριμένα η ανάλυση των δεδομένων μπορεί να διακριθεί σε:

- περιγραφική ανάλυση - descriptive: αποτελεί το πρώτο βήμα της ανάλυσης, για να περιγράψει τι υπάρχει σε ένα σύνολο ή σε μια βάση δεδομένων, καταδεικνύοντας ομάδες και ιδιότητες των δεδομένων
- διαγνωστική - diagnostic: αποτελεί το επόμενο επίπεδο ανάλυσης όπου ανακαλύπτονται συσχετίσεις μεταξύ των δεδομένων και εντοπίζονται βαθύτερα αίτια που δημιούργησαν γεγονότα και συγκεκριμένες συμπεριφορές
- προγνωστική - predictive: ανάλυση ιστορικών και τρεχόντων γεγονότων με την χρήση εξελιγμένων μοντέλων πρόβλεψης της μηχανικής μάθησης και εκτίμηση με ακρίβεια της πιθανότητας να συμβεί κάτι στο μέλλον, με την κατασκευή κάποιου μοντέλου

- προδιαγραφική - prescriptive: το τελικό στάδιο ανάλυσης όπου με βάση τα αποτελέσματα των προηγούμενων σταδίων επιλέγονται ποιες ενέργειες πρέπει να γίνουν ώστε να επιτευχθεί το καλύτερο δυνατό αποτέλεσμα

Ο Κύρκος (2016) ορίζει ως επιχειρηματική ευφυΐα ένα σύνολο από μεθόδους, τεχνολογίες, ικανότητες και στρατηγικές με την χρήση των οποίων είναι δυνατή η υψηλού επιπέδου ανάλυση των δεδομένων με σκοπό την εξαγωγή της κατάλληλης πληροφορίας που θα επιτρέπει στους οργανισμούς να αντιλαμβάνονται καταστάσεις, να προβλέπουν τάσεις και γεγονότα, να διατυπώνουν συμπεράσματα, να σχεδιάζουν και να υποστηρίζουν την διαδικασία λήψης αποφάσεων.



Εικόνα 4: Στάδια Ανακάλυψης Γνώσης

Πηγή: Κύρκος, 2016

http://repository.kallipos.gr/handle/11419/1226

Η ανακάλυψη γνώσης είναι μια διαδικασία πολλαπλών και επαναλαμβανόμενων σταδίων που έχει αφετηρία την συλλογή και προ-επεξεργασία των πηγαίων δεδομένων. Στην συνέχεια δημιουργείται το κατάλληλο σύνολο δεδομένων (data set) με επιλογή των σημαντικότερων στηλών (features) ή και γραμμών (instances), ενώ συχνά απαιτείται να μετασχηματιστούν τα δεδομένα (κανονικοποίηση, διακριτοποίηση) ώστε να ανταποκρίνονται στις ιδιαίτερες απαιτήσεις των μεθόδων που θα χρησιμοποιηθούν για την ανάλυση. Η εξόρυξη δεδομένων (data mining) είναι το κυριότερο σημείο στην πορεία για την ανακάλυψη της γνώσης, που επικεντρώνεται στη διερεύνηση μεγάλων συνδυασμένων συνόλων δεδομένων για τον εντοπισμό μοτίβων και συσχετίσεων που μπορούν να οδηγήσουν σε πρότυπα. Στο τελικό στάδιο αξιολογούνται τα πρότυπα και εξάγονται τα ανάλογα συμπεράσματα ή επαναλαμβάνεται η διαδικασία.

Η μηχανική μάθηση είναι η ικανότητα ενός υπολογιστικού συστήματος να μαθαίνει από τα δεδομένα και να βελτιώνει την λειτουργία και την αποδοτικότητά του βάση της εμπειρίας που αποκτά. Η διαδικασία της μηχανικής μάθησης μπορεί γενικά να χωριστεί σε τέσσερις κατηγορίες: επιβλεπόμενη μάθηση, μη επιβλεπόμενη, ημί-επιβλεπόμενη και ενίσχυση.

Κατά την επιβλεπόμενη μάθηση στο σύνολο δεδομένων υπάρχει ένα γνώρισμα-στόχος (κλάση), δηλαδή μια εξαρτημένη μεταβλητή που καθοδηγεί και καθορίζει την διαδικασία της μάθησης. Η ανάλυση έγκειται στην κατασκευή ενός μοντέλου που επιτρέπει την πρόβλεψη της τιμής της εξαρτημένης μεταβλητής από τις ανεξάρτητες μεταβλητές. Αντίθετα στην μη επιβλεπόμενη μάθηση δεν υπάρχει κάποια στήλη στόχος, ούτε προηγούμενη γνώση για την ύπαρξη σχέσεων μεταξύ των δεδομένων και η ανάλυση έγκειται στην ομαδοποίηση και στον εντοπισμό σημαντικών και πυκνών δομών στα δεδομένα. Η ημι-επιβλεπόμενη μάθηση είναι ένας συνδυασμός των δύο μεθόδων. Κυρίως εφαρμόζεται στην περίπτωση μεγάλων συνόλων ακατέργαστων και μη δομημένων δεδομένων όπου εισάγεται μικρή ποσότητα δεδομένων με ετικέτα που φέρουν όμως σημαντική πληροφορία για να υποστηρίξουν την διαδικασία μάθησης στα μη επισημασμένα δεδομένα. Μια άλλη όμως περίπτωση χρήσης ημι-επιβλεπόμενης μάθησης είναι η μάθηση χωρίς επίβλεψη που καθοδηγείται από περιορισμούς και εφαρμόζεται όταν ο αριθμός και η φύση των κλάσεων δεν είναι εκ των προτέρων γνωστές και πρέπει να εξαχθούν από τα δεδομένα. Εκτεταμένη έρευνα για την ημι-εποπτευόμενη μάθηση, τις εφαρμογές και τους περιορισμούς της, έχει παρουσιαστεί

από ομάδα ερευνητών σε συνεργασία με το MIT (Chapelle et al., 2006). Επίσης οι Jespen van Engelen & Hoos (2020) πρόσφατα επιχείρησαν μια συστηματική καταγραφή των σημαντικότερων προσεγγίσεων και αλγορίθμων μάθησης με ημι-επίβλεψη που αναπτύχθηκαν τα τελευταία είκοσι έτη.

Η τέταρτη κατηγορία μηχανικής μάθησης, η ενισχυτική μάθηση (reinforced/enhanced machine learning) δεν βασίζεται σε ένα στατικό σύνολο δεδομένων αλλά λειτουργεί σε ένα περιβάλλον δυναμικό, με την εισαγωγή κανόνων ή επιτρεπτών ενεργειών και αντίστοιχα πιθανών τελικών εκβάσεων. Το μοντέλο ενισχυτικής μάθησης μαθαίνει από το παράδειγμα και την παρατήρηση όταν ο επιθυμητός στόχος είναι σταθερός ή μαθαίνει από την εμπειρία και την ανταμοιβή (η ανταμοιβή είναι μια αριθμητική επιδίωξη του αλγόριθμου) όταν το αποτέλεσμα είναι μεταβλητό. Πρόσφατο παράδειγμα ενισχυτικής μάθησης είναι το πρόγραμμα Alphago της Google που συνδυάζει εξελιγμένο δέντρο αναζήτησης και βαθιά νευρωνικά δίκτυα. Είναι χαρακτηριστικό ότι εκπαιδεύτηκε μέσα σε λίγες μόνο ημέρες, παίζοντας εκατομμύρια παιχνίδια τόσο με ανθρώπους όσο και με διαφορετικές εκδόσεις του εαυτού του, συσσωρεύοντας χιλιάδες χρόνια ανθρώπινης γνώσης.

1.2 Επιβλεπόμενη μάθηση

Οι συνηθέστερα χρησιμοποιούμενες μέθοδοι για την ανίχνευση απάτης χωρίζονται σε δύο κατηγορίες, σε εποπτευόμενες και μη εποπτευόμενες προσεγγίσεις, με την πρώτη να προτιμάται από την πλειονότητα των ερευνητών. Για τους σκοπούς της παρούσας εργασίας και την καλύτερη κατανόηση θα γίνει συνοπτική περιγραφή αλγόριθμων κατηγοριοποίησης οι οποίοι θα δοκιμαστούν στη συνέχεια στο πειραματικό κομμάτι.

Όπως ήδη ειπώθηκε, κατά την επιβλεπόμενη μάθηση κατασκευάζεται ένας μηχανισμός λήψης αποφάσεων που προβλέπει τις τιμές της εξαρτημένης μεταβλητής από ένα σύνολο άλλων γνωρισμάτων. Η κατηγοριοποίηση ως εργασία επιβλεπόμενης μάθησης προβλέπει διακριτές ονομαστικές τιμές, ταξινομώντας τις παρατηρήσεις σε γνωστές εκ των προτέρων κατηγορίες (κλάση). Με παρόμοιο τρόπο λειτουργεί και η παλινδρόμηση προβλέποντας όμως όχι ονομαστικές τιμές αλλά συνεχόμενες (αριθμητικές) τιμές.

Τα Δέντρα Αποφάσεων βασίζονται στον αλγόριθμο ID3 που επινοήθηκε από τον Quinlan το 1986. Ο αλγόριθμος C4.5 που αποτελεί επέκτασή τους βασίζεται στην επιλογή του γνωρίσματος που βέλτιστα χωρίζει τις περιπτώσεις και ο διαχωρισμός γίνεται βάση των τιμών αυτού του γνωρίσματος, χρησιμοποιώντας ένα κριτήριο βασισμένο στην εντροπία, τον Λόγο Κέρδους (Gain Ratio). Τα Δέντρα Αποφάσεων δεν κάνουν αυθαίρετες υποθέσεις για την ανεξαρτησία των μεταβλητών ή την κατανομή των δεδομένων, παράγουν κατανοητά μοντέλα και εκπαιδεύονται γρήγορα. Μειονέκτημά τους είναι ότι μικρές αλλαγές στο σύνολο εκπαίδευσης μπορεί να επηρεάσουν σημαντικά την κατασκευή του μοντέλου όπως και ότι απαιτούν εγκατάσταση του συνόλου των δεδομένων εκπαίδευσης στην μνήμη του υπολογιστή. Τα Νευρωνικά Δίκτυα τύπου Multilayer Perceptron (MLP) ανήκουν επίσης στις μεθόδους επιβλεπόμενης μάθησης. Μιμούμενα την λειτουργία του ανθρώπινου εγκεφάλου αποτελούνται από νευρώνες που συνδέονται μεταξύ τους με συνδέσεις. Η καθεμία από αυτές φέρει μια αριθμητική τιμή το βάρος (weight), με το οποίο πολλαπλασιάζεται το σήμα που περνάει από την σύνδεση. Οι νευρώνες δέχονται τις τιμές εισόδου οι οποίες αθροίζονται και μετασχηματίζονται από την συνάρτηση μετασχηματισμού και στην συνέχεια το συνολικό σήμα διαβιβάζεται στο επόμενο στρώμα νευρώνων. Η εκπαίδευση του νευρωνικού δικτύου συνίσταται στην ρύθμιση και τροποποίηση των βαρών των συνδέσεων έτσι ώστε να ελαχιστοποιείται το μέσο τετραγωνικό σφάλμα μεταξύ της προβλεπόμενης και της πραγματικής κλάσης. Ο πιο διαδεδομένος αλγόριθμος είναι ο αλγόριθμος Αντίστροφης Μετάδοσης Σφάλματος Back Propagation. Τα νευρωνικά δίκτυα πετυχαίνουν πολύ υψηλές αποδόσεις κατηγοριοποίησης έναντι άγνωστων και σύνθετων παρατηρήσεων, μπορούν να χειριστούν θορυβώδη δεδομένα αλλά σημαντικό μειονέκτημα αποτελεί ότι απαιτούν μεγάλο χρόνο εκπαίδευσης, πληθώρα παραμέτρων προσδιορίζεται εμπειρικά από τον χρήστη και ότι τα παραγόμενα μοντέλα δεν είναι εύκολα ερμηνεύσιμα (μαύρο κουτί).

Οι Μηχανές Διανυσμάτων Υποστήριξης Support Vector Machines (SVM) οι οποίες αρχικά διατυπώθηκαν από τον Vapnik το 1992, έχουν σαν βασική ιδέα την κατασκευή ενός υπέρ επιπέδου, σε έναν χώρο περισσότερων διαστάσεων, το οποίο πιθανώς διαχωρίζει γραμμικά την κλάση των παρατηρήσεων και λειτουργεί ως συνάρτηση απόφασης (Cristianini & Ricci, 2008). Για τον υπολογισμό αυτού του βέλτιστου επιπέδου εισάγεται ο όρος περιθώριο (margin) και οι παρατηρήσεις που βρίσκονται στο όριο του μέγιστου περιθωρίου

ονομάζονται διανύσματα υποστήριξης. Οι Μηχανές Διανυσμάτων Υποστήριξης είναι άριστοι ταξινομητές σε περιπτώσεις δυαδικής κλάσης και σε σύνολα δεδομένων με πολλές στήλες και λίγες γραμμές. Δεν παράγουν όμως ερμηνεύσιμα μοντέλα, απαιτούν μεγάλη μνήμη υπολογιστή και μεγάλους χρόνους εκπαίδευσης, όπως επίσης η επιλογή της καταλληλότερης συνάρτησης πυρήνα αποτελεί ακόμα ευρύ πεδίο έρευνας.

Η Λογιστική ή Λογαριθμική Παλινδρόμηση Logistic Regression (LR) είναι η παλαιότερη χρησιμοποιούμενη μέθοδος κατηγοριοποίησης και προέρχεται από την επιστήμη της στατιστικής. Δεν επιτυγχάνει υψηλές επιδόσεις σε σύγκριση με τις προηγούμενες αναφερόμενες μεθόδους αλλά είναι απλή, προβλέπει κλάσεις με περισσότερες από δύο τιμές και δίνει ένα μέτρο της σημαντικότητας των ανεξάρτητων μεταβλητών.

Στην λειτουργία των Δέντρων Αποφάσεων έχει βασιστεί ο αλγόριθμος Random Forest (RF) όπου δημιουργούνται πολλαπλά σύνολα δεδομένων από το αρχικό σύνολο εκπαίδευσης μέσω δειγματοληψίας και αντικατάστασης, επιλέγοντας τυχαία γραμμές και στήλες. Με αυτόν τον τρόπο διασφαλίζεται η διαφοροποίηση των συνόλων δεδομένων και παράγονται πολλαπλά διαφοροποιημένα μοντέλα (ensemble classifiers). Η τελική απόφαση ταξινόμησης μιας νέας παρατήρησης λαμβάνεται με τη συγκέντρωση, συνάθροιση και ψηφοφορία μεταξύ των μεμονωμένων αποφάσεων. Η μέθοδος Random Forest (RF) μπορεί να χειριστεί μεγάλα σύνολα δεδομένων, δεν υπερπροσαρμόζεται και επιτυγχάνει υψηλές επιδόσεις (Breiman, 2001).

ΒΙΒΛΙΟΓΡΑΦΙΑ ΜΕΡΟΣ Β' ΕΝΟΤΗΤΑ 1

Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Chapelle, O., Schölkopf, B., & Zien, A. (Eds.). (2006). *Semi-supervised learning*. MIT Press.

Cristianini, N., & Ricci, E. (2008). Support Vector Machines. In M.-Y. Kao (Ed.), *Encyclopedia of Algorithms* (pp. 928–932). Springer US. https://doi.org/10.1007/978-0-387-30162-4_415

Power, D. J., Heavin, C., McDermott, J., & Daly, M. (2018). Defining business analytics: An empirical approach. *Journal of Business Analytics*, *1*(1), 40–53. https://doi.org/10.1080/2573234X.2018.1507605

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/BF00116251

Schniederjans, M. J., Schniederjans, D. G., & Starkey, C. M. (2014). *Business Analytics Principles, Concepts, and Applications: What, Why, and How*. Pearson Education.

Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373–440. https://doi.org/10.1007/s10994-019-05855-6

Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018). *Random forest for credit card fraud detection*. 1–6. Scopus. https://doi.org/10.1109/ICNSC.2018.8361343

Κύρκος, Ε. (2016). *Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων*. http://repository.kallipos.gr/handle/11419/1226

Ενότητα 2 Ειδικά θέματα κατηγοριοποίησης

2.1 Ανισοκατανομή των κλάσεων

Πολλές φορές συναντώνται σύνολα δεδομένων με ανομοιόμορφη κατανομή των κλάσεων. Αυτό σημαίνει ότι οι παρατηρήσεις που ανήκουν σε μια κλάση υπερτερούν κατά πολύ αριθμητικά σε σχέση με την κλάση των υπόλοιπων παρατηρήσεων, δημιουργώντας πρόβλημα μεροληψίας. Τυπικό παράδειγμα είναι η ανίχνευση απάτης στον τραπεζικό τομέα όπου η πλειοψηφία των περιπτώσεων είναι νόμιμες συναλλαγές αλλά το ενδιαφέρον έγκειται στον εντοπισμό των κακόβουλων συναλλαγών, οι οποίες συνήθως είναι της τάξης του 1,5 – 2% και αποτελούν δηλαδή την μειοψηφούσα κλάση. Όμως τα περισσότερα μοντέλα μηχανικής μάθησης που χρησιμοποιούνται για κατηγοριοποίηση έχουν σχεδιαστεί για να εκπαιδεύονται σε σύνολα δεδομένων με μια ισορροπημένη κατανομή των παρατηρήσεων. Μια σοβαρή ανισοκατανομή των κλάσεων όπως η παραπάνω δημιουργεί σφάλματα στην ταξινόμηση διότι οι αλγόριθμοι θα αγνοήσουν την μικρότερη κλάση, μαθαίνοντας να προβλέπουν με πολύ υψηλά ποσοστά ακρίβειας (>90%) την πλειοψηφούσα κλάση. Αυτό συμβαίνει γιατί έχοντας σαν δεδομένο ότι τα σφάλματα όλων των κλάσεων έχουν το ίδιο κόστος, προσπαθούν να ελαχιστοποιήσουν το συνολικό σφάλμα, στο οποίο ουσιαστικά η μειοψηφική κλάση συμβάλει σε μικρότερο βαθμό προσφέροντας πολύ λίγη πληροφορία για μάθηση (Ganganwar, 2012).

Για την αντιμετώπιση του προβλήματος απαιτείται η αναδιαμόρφωση του συνόλου δεδομένων ή η χρήση υβριδικών αλγόριθμων μάθησης, δίνοντας διαφορετικό κόστος στα παραδείγματα εκπαίδευσης, ώστε να μπορούν να χειρίζονται τη στρεβλή κατανομή των κλάσεων. Σε επίπεδο δεδομένων υπάρχουν εξειδικευμένες τεχνικές για την εξισορρόπηση ανισοβαρών συνόλων δεδομένων με πιο διαδεδομένες την υπέρ-δειγματοληψία της μειοψηφούσας κλάσης (oversampling) ή αντίθετα την τυχαία υπό-δειγματοληψία της πλειοψηφικής κλάσης (random under sampling), την συνθετική δειγματοληψία με παραγωγή τεχνητών δεδομένων (αλγόριθμος Smote) και τις μεθόδους δειγματοληψίας με βάση

συστάδες (cluster-based), (He & Ma, 2013).

Στην υπέρ-δειγματοληψία οι επιστήμονες των δεδομένων αυξάνουν με διάφορες τεχνικές τον αριθμό των σπάνιων γεγονότων. Στην συνθετική υπέρ-δειγματοληψία μειονοτήτων ο αλγόριθμος SMOTE δημιουργεί επιπλέον τεχνητά δεδομένα εκπαίδευσης από την κλάση που μειοψηφεί. Η μέθοδος αυτή είναι δύσκολο να εφαρμοστεί από μόνη της σε μεγάλα σύνολα δεδομένων, λόγω του υπερδιπλασιασμού του όγκου των δεδομένων, αυξάνοντας κατά συνέπεια τον χρόνο εκπαίδευσης και την πιθανότητα υπερπροσαρμογής των μοντέλων λόγω της αντιγραφής των περιπτώσεων. Ερευνητές έχουν προτείνει τον συνδυασμό αυτής ταυτόχρονα με την υπό-δειγματοληψία, αφαιρώντας δηλαδή τυχαία δείγματα από την πλειοψηφική κλάση έως ένα ποσοστό, επιτυγχάνοντας ακόμα καλύτερες επιδόσεις ταξινομητή στην καμπύλη ROC (Chawla et al., 2002). Παρόμοια πρόταση με χρήση υβριδικής τεχνικής έχει προταθεί ειδικά για την περίπτωση ανίχνευσης απάτης, χρησιμοποιώντας ένα συνδυασμό της SMOTE για την υπέρ-δειγματοληψία στα δεδομένα μειοψηφίας που είναι τα δείγματα απάτης και της τυχαίας υπό-δειγματοληψίας στα δείγματα μη απάτης, εξαλείφοντας ταυτόχρονα ακραίες τιμές στις περιοχές μειοψηφίας (Padmaja et al., 2007).

Με την μέθοδο της υπό-δειγματοληψίας προκειμένου να εξισορροπηθούν άνισα σύνολα δεδομένων και να δημιουργηθούν δύο ισομεγέθεις κλάσεις, διατηρούνται όλα τα σπάνια συμβάντα και μειώνεται το πλήθος της κλάσης που πλειοψηφεί. Στην περίπτωση της τυχαίας διαγραφής των περιπτώσεων της πλειοψηφικής κλάσης υπάρχει το ενδεχόμενο να χαθεί δυνητικά σημαντική πληροφορία. Για την αποφυγή του προβλήματος έχουν προταθεί διάφορες τεχνικές ώστε να γίνει προσεκτική επιλογή των περιπτώσεων που θα διαγραφούν. Περιγράφονται στην συνέχεια οι συνηθέστερα χρησιμοποιούμενοι αλγόριθμοι.

- Οι αλγόριθμοι NearMiss εντοπίζουν τα αρνητικά αλλά σχεδόν θετικά δείγματα από την πλειοψηφούσα κλάση με στόχο να διατηρηθούν στο σύνολο δεδομένων. Επιλέγουν με βάση την εγγύτητά τους στην μειοψηφούσα κλάση, με διάφορες παραλλαγές: την μικρότερη, μεγαλύτερη και μέση απόσταση από τους τρεις πλησιέστερους γείτονες στην τάξη της μειονότητας (Zhang & Mani, 2003)
- Αφαίρεση των συνδέσμων Tomek Links (Tomek, 1976). Αρκετοί ερευνητές πιστεύουν πως για περιπτώσεις που βρίσκονται πολύ κοντά στο περιθώριο ταξινόμησης δύο κλάσεων πιθανόν κάποιες από αυτές να αποτελούν θόρυβο οπότε πρέπει να

αφαιρεθούν, (Batista et al., 2004). Οι σύνδεσμοι αυτοί ονομάζονται Tomek Links και με αυτήν τη μέθοδο καθαρίζονται τα δεδομένα και αποφεύγονται επικαλύψεις.

- Edited Nearest Neighbors (ENN): Ο αλγόριθμος ENN λειτουργεί με παρόμοιο τρόπο, χρησιμοποιώντας τον αλγόριθμο k-πλησιέστερων γειτόνων για τον εντοπισμό δειγμάτων στην πλειοψηφική κλάση που έχουν ταξινομηθεί εσφαλμένα και στη συνέχεια τα αφαιρεί (Wilson, 1972).
- Αλγόριθμος Cluster Centroids. Χρησιμοποιώντας τον αλγόριθμο K-means δημιουργούνται συστάδες στην πλειοψηφική κλάση οι οποίες στην συνέχεια αντικαθίστανται από τα κέντρα τους. Σχετική ερευνητική εργασία έχουν παρουσιάσει οι Yen & Lee (2009).

Είναι γεγονός ότι οι περισσότεροι ερευνητές συγκλίνουν στην επιλογή υβριδικών μεθόδων. Ενδεικτικά αναφέρουμε την έρευνα των Miroslav Kubat και Stan Matwin (1997) που με την χρήση των συνδέσμων Tomek και τον αλγόριθμο CNN (μετεξέλιξη του οποίου είναι ο Edited Nearest Neighbors ENN) αφαίρεσαν τα οριακά και περιττά γεγονότα από την πλειοψηφική κλάση. Με συνδυασμό των CNN και ENN εργάστηκε επίσης η J.Laurikkala (2001) αφαιρώντας τις διπλές και διφορούμενες περιπτώσεις δημιουργώντας ένα σύνολο δεδομένων υψηλότερης ποιότητας. Τέλος μια διαφορετική πρόταση παρουσιάζεται στην ιστοθέση της Google (2023) (https://developers.google.com/machine-learning) όπου δημιουργείται ένα νέο υποσύνολο δεδομένων για εκπαίδευση διατηρώντας ατόφια την μειοψηφική κλάση και επιλέγοντας παρατηρήσεις από την πλειοψηφική κλάση, προσθέτοντας ταυτόχρονα βάρη σε αυτές, ανάλογα της υπό-δειγματοληψίας που υφίσταται η κλάση τους.

2.2 Κόστος σφάλματος

Όπως αναλύθηκε παραπάνω τυχόν αποτυχία στην διαχείριση ενός ανισόρροπου συνόλου δεδομένων θα έχει σοβαρό αντίκτυπο στην αποτελεσματικότητα και στην ικανότητα πρόβλεψης των μοντέλων μηχανικής μάθησης. Στα δεδομένα του πραγματικού κόσμου οι αποτυχίες πρόβλεψης δεν έχουν το ίδιο κόστος. Επιλέγεται λοιπόν η εκπαίδευση των μοντέλων να γίνεται με τέτοιο τρόπο ώστε να μειώνεται το συνολικό κόστος των

εσφαλμένων κατηγοριοποιήσεων και η μέθοδος αυτή ονομάζεται εκπαίδευση ευαίσθητη προς το κόστος (cost sensitive learning), (Κύρκος, 2016).

Ένα μοντέλο ευαίσθητο στο κόστος θα προσπαθήσει να μάθει περισσότερα χαρακτηριστικά των δειγμάτων από την κατηγορία με το υψηλό κόστος, δίνοντας μεγαλύτερο βάρος στην εσφαλμένη ταξινόμηση ενός τέτοιου δείγματος (Yen & Lee, 2009). Σε επίπεδο αλγορίθμων εφαρμόζονται ποικίλες λύσεις όπως η ενσωμάτωση του κόστους σφάλματος στην ταξινόμηση με δέντρα αποφάσεων (δέντρα CART Classification and Regression Trees) όπου ο αλγόριθμος χρησιμοποιώντας το κριτήριο Twoing προσπαθεί να μειώσει τις λάθος προβλέψεις της σημαντικής κλάσης, η λύση της βέλτιστης προσαρμογής της συνάρτησης απόφασης στις μηχανές διανυσμάτων υποστήριξης (SVM) και η μέθοδος εκμάθησης από μία κλάση και όχι η μάθηση με βάση τη διάκριση σε δύο κλάσεις (Ganganwar, 2012). Οι Bahnsen κ.α. (2015) έχουν προτείνει έναν αλγόριθμο δέντρου αποφάσεων με ευαισθησία στο κόστος και επιπλέον εισήγαγαν ένα κριτήριο κλαδέματος με βάση το κόστος. Στην εργασία τους εστίασαν στην υπόθεση ότι το κόστος διαφέρει όχι μόνο ανάμεσα στις κλάσεις αλλά εξαρτάται και από το κόστος κάθε παραδείγματος, βελτιώνοντας με αυτόν τον τρόπο την διαδικασία διαγραφής των περιττών κλάδων (κλάδεμα). Το παραγόμενο μοντέλο σχημάτισε μικρότερα δέντρα, απλούστερα και ευκολότερα να αναλυθούν, σε πολύ λιγότερο χρόνο. Οι Sahin κ.ά. πρότειναν επίσης έναν αλγόριθμο δέντρων αποφάσεων ο οποίος ελαχιστοποιεί το άθροισμα του κόστους λανθασμένης ταξινόμησης επιλέγοντας το χαρακτηριστικό διάσπασης σε κάθε μη τερματικό κόμβο του δέντρου (Sahin et al., 2013).

Επιπλέον ο Krawczyk (2016) προτείνει μεθόδους μη επιβλεπόμενης μάθησης όπως την χρήση συστάδων για μια εις βάθος μελέτη της δομής και της φύσης των δεδομένων της μειοψηφούσας κλάσης ή ακόμα και για την αποδόμηση του αρχικού συνόλου δεδομένων σε μικρότερα και την περαιτέρω εξατομικευμένη διαχείρισή τους.

Νεότεροι ερευνητές έχουν εργαστεί με το Imbalance - XGBoost πακέτο Python, εφαρμόζοντας σταθμισμένες cross-entropy και εστιακές απώλειες στη μηχανή ενίσχυσης για δυαδική ταξινόμηση σε ανισσόροπα σύνολα δεδομένων. Η σταθμισμένη cross-entropy απώλεια αυξάνει την τιμωρία στις εσφαλμένες ταξινομήσεις. Η εστιακή απώλεια μειώνει τη σημασία των σωστά ταξινομημένων παρατηρήσεων (Wang et al., 2020). Πολύ σημαντική

είναι και η εργασία των Zhang κ.α. για την ταξινόμηση δεδομένων ανισορροπίας πολλαπλών κλάσεων παρουσιάζοντας στην επιστημονική κοινότητα το Multi-Imbalance, ένα πακέτο λογισμικού ανοικτού κώδικα 18 αλγορίθμων (Zhang et al., 2019).

2.3 Μέτρηση της απόδοσης των ταξινομητών

Σε σύνολα δεδομένων με ανισοκατανομή των κλάσεων ή που οι εσφαλμένες κατηγοριοποιήσεις των κλάσεων έχουν διαφορετικό κόστος, είναι σημαντική η μέτρηση της απόδοσης των ταξινομητών ανά κλάση. Η απόδοση των αλγόριθμων αξιολογείται και απεικονίζεται στον πίνακα σύγχυσης (confusion matrix), όπου οι στήλες είναι η προβλεπόμενη κλάση και οι γραμμές είναι η πραγματική κλάση (Luque et al., 2019).

		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Εικόνα 5: Πίνακας σύγχυσης - confusion matrix

Πηγή: Luque et al., 2019

Όπου οι τιμές στα κελιά του πίνακα είναι:

True Positive (TP): είναι το πλήθος των θετικών περιπτώσεων που ταξινομήθηκαν σωστά

False Positive (FP): είναι το πλήθος των αρνητικών περιπτώσεων που ταξινομήθηκαν λάθος (ταξινομήθηκαν ως θετικά)

False Negative (FN): είναι το πλήθος των θετικών περιπτώσεων που ταξινομήθηκαν λάθος (ταξινομήθηκαν εσφαλμένα ως αρνητικά)

True Negative (TN): είναι το πλήθος των αρνητικών περιπτώσεων που ταξινομήθηκαν σωστά

Σε συνέχεια των παραπάνω, ένα μέτρο για την απόδοση ενός κατηγοριοποιητή σε εξισορροπημένα σύνολα δεδομένων είναι η ακρίβεια (accuracy) η οποία προκύπτει από τον λόγο του πλήθους των σωστών προβλέψεων προς το πλήθος των παρατηρήσεων (ποσοστό ορθών προβλέψεων).

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

Για την αξιολόγηση όμως ανισοβαρών περιπτώσεων έχουν προταθεί επιπλέον μέτρα της απόδοσης όπως: ανάκληση (recall) ή αλλιώς ευαισθησία (sensitivity), εξειδίκευση (specificity), ακρίβεια (precision), F1 Score ή αλλιώς F-measure και ο δείκτης Cohen's Kappa.

Σε ζητήματα δυαδικής ταξινόμησης μετρώνται η ακρίβεια (precision) και η ανάκληση (recall).

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

$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{ή} \qquad \text{Sensitivity} = \frac{TP}{\text{Positive}}$$

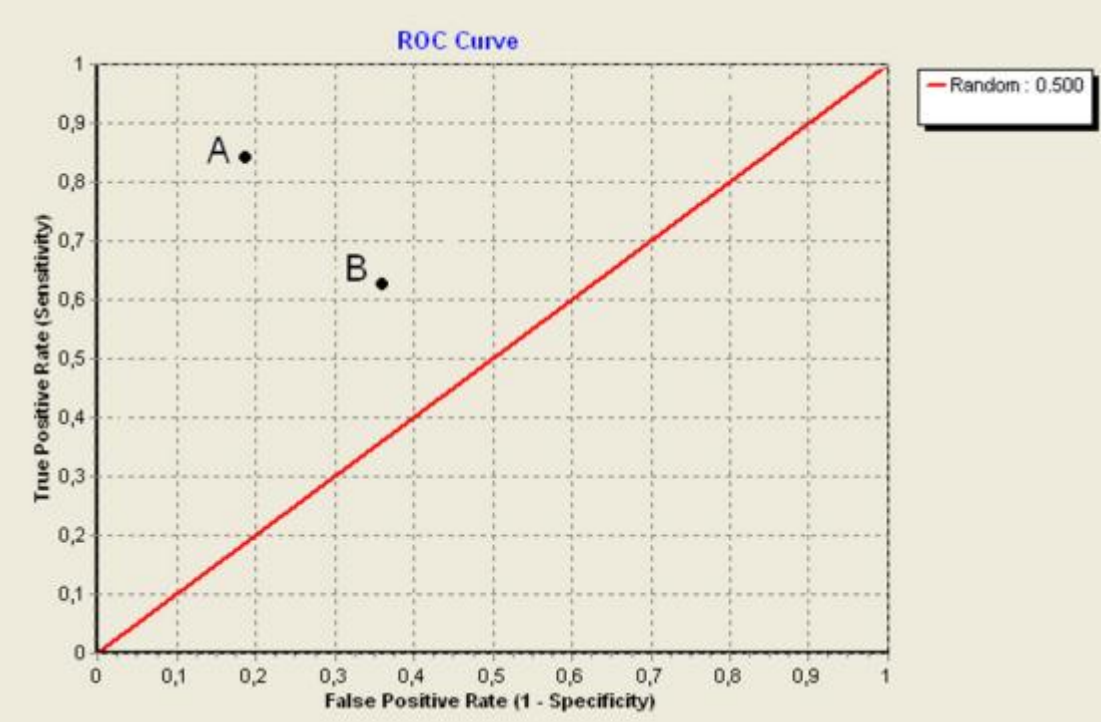
Η ακρίβεια (precision) είναι η ικανότητα του ταξινομητή να μην χαρακτηρίζει ως θετικό ένα δείγμα που είναι αρνητικό ενώ η ανάκληση (recall) ή αλλιώς ευαισθησία (sensitivity), είναι η ικανότητα του ταξινομητή να βρίσκει όλα τα θετικά δείγματα. Το μέτρο F1 Score ή F-measure είναι ένας σταθμισμένος αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης με τιμές από 0-1 (scikit-learn.org).

Στην περίπτωση έντονα ανισοβαρών συνόλων δεδομένων καλύτερο μέτρο της απόδοσης είναι οι καμπύλες ROC, οι οποίες μπορούν να θεωρηθούν ότι αντιπροσωπεύουν καλύτερα τα όρια απόφασης για το σχετικό κόστος των TP και FP, (Chawla et al., 2002). Οι καμπύλες αυτές σχεδιάζονται σε ένα σύστημα αξόνων όπου η ευαισθησία (sensitivity) ή αλλιώς True Positive Rate αποτελεί τον κατακόρυφο άξονα ενώ ο οριζόντιος άξονας εκφράζει το μέγεθος False Positive Rate (ή αλλιώς false alarm rate) και ισούται με

$$\text{False Positive Rate} = 1 - \text{Specificity} = \frac{FP}{\text{Negative}}$$

$$\text{Specificity} = \frac{TN}{\text{Negative}} = \frac{TN}{FP + TN}$$

Κάθε κατηγοριοποιητής παράγει ένα ζεύγος τιμών που αντιστοιχεί σε ένα μόνο σημείο στο γράφημα ROC.



Εικόνα 6: Καμπύλη ROC

Πηγή: Κύρκος, 2016

http://repository.kallipos.gr/handle/11419/1226

Μία καμπύλη ROC απεικονίζει τις σχέσεις ανάμεσα στο όφελος και στο κόστος. Το σημείο (0,0) αντιπροσωπεύει τον κατηγοριοποιητή που δεν εκδίδει ποτέ θετική ταξινόμηση. Αντίθετα το σημείο (1.1) αντιπροσωπεύει τον κατηγοριοποιητή που δίνει πάντα θετικές ταξινομήσεις. Είναι εύλογο ότι το σημείο (0,1) αντιπροσωπεύει την τέλεια ταξινόμηση όπου όλα τα θετικά παραδείγματα ταξινομούνται σωστά και κανένα αρνητικό παράδειγμα δεν έχει ταξινομηθεί ως θετικό. Τα σημεία που έχουν καλύτερες επιδόσεις βρίσκονται υψηλότερα και πιο αριστερά στον χώρο. Κατά συνέπεια, κατηγοριοποιητές που βρίσκονται πιο κοντά στην αριστερή πλευρά του χώρου θα μπορούσαν να χαρακτηριστούν ως συντηρητικοί γιατί δίνουν θετικές ταξινομήσεις μόνο με ισχυρές ενδείξεις (έχουν λίγα ψευδώς θετικά) ενώ

κατηγοριοποιητές στην πάνω δεξιά πλευρά κάνουν θετικές ταξινομήσεις με ασθενή στοιχεία, οπότε και έχουν υψηλά ποσοστά ψευδώς θετικών αποτελεσμάτων (Fawcett, 2006).

Ένα πολύ ισχυρό μέτρο σύγκρισης της απόδοσης των κατηγοριοποιητών είναι η περιοχή κάτω από την καμπύλη ROC, η ονομαζόμενη AUC (Area Under ROC Curve). Όσο μεγαλύτερη περιοχή AUC καλύπτει ένας κατηγοριοποιητής τόσο καλύτερος είναι. Ερευνητές απέδειξαν ότι σε σοβαρές ανισοκατανομές των κλάσεων η απόδοση των ταξινομητών φθίνει δραματικά με εξαίρεση την περιοχή κάτω από την καμπύλη ROC που αποτελεί ένα αξιόπιστο κριτήριο (Jeni et al., 2013).

Τέλος ο δείκτης Kappa του Cohen (1960) είναι ένα στατιστικό μέτρο με εύρος τιμών (-1, 1) που λαμβάνει υπόψη την ανισοκατανομή των κλάσεων. Όσο περισσότερο διαφέρουν οι κατανομές των προβλεπόμενων και των πραγματικών κλάσεων-στόχων, τόσο χαμηλότερη είναι η τιμή του ενώ μία τιμή του άνω του 0,80 αντιπροσωπεύει μια ορθότερη πρόβλεψη στα ανισοβαρή δεδομένα.

ΒΙΒΛΙΟΓΡΑΦΙΑ ΜΕΡΟΣ Β΄ ΕΝΟΤΗΤΑ 2

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Correa Bahnsen, A., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, *42*(19), 6609–6619. https://doi.org/10.1016/j.eswa.2015.04.042

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, *2*, 42–47.

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing Imbalanced Data–Recommendations for the Use of Performance Metrics. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 245–251. https://doi.org/10.1109/ACII.2013.47

He, H.& Ma, Y. (2013). Imbalanced Learning Foundations, Algorithms, and Applications. *John Willey &Sons.*

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Kumbat, M. & Matwin, S. (1997), Addressing The Curse Of Imbalanced Training Sets: One-sided Selection. *International Conference on Machine Learning*

Laurikkala, J.(2001). Improving Identifications of Difficult Small Classes by Balancing Class Distribution.

Artificial Intelligence in Medicine. AIME 2001. Lecture Notes in Computer Science vol 2101. Springer, Berlin, Heidelberg.

Luque, A., Carrasco, A., Martín, A., & De Las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023

Padmaja, T. M., Dhulipalla, N., Bapi, R. S., & Krishna, P. R. (2007). Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection. *15th International Conference on Advanced Computing and Communications (ADCOM 2007)*, 511–516. https://doi.org/10.1109/ADCOM.2007.74

Sahin, Y., Bulkan, S., & Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications*, *40*(15), 5916–5923. Scopus. https://doi.org/10.1016/j.eswa.2013.05.021

Tomek, I. (1976). Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-6*(11), 769–772. https://doi.org/10.1109/TSMC.1976.4309452

Wang, C., Deng, C., & Wang, S. (2020). Imbalance-XGBoost: Leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *Pattern Recognition Letters*, *136*, 190–197. https://doi.org/10.1016/j.patrec.2020.05.035

Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-2*(3), 408–421. https://doi.org/10.1109/TSMC.1972.4309137

Yen, S.-J., & Lee, Y.-S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, *36*(3, Part 1), 5718–5727. https://doi.org/10.1016/j.eswa.2008.06.108

Zhang, J., & Mani, I. (2003). kNN approach to unbalanced data distributions: A case study involving information extraction. In *Proceedings of the ICML'2003 workshop on learning from imbalanced datasets*.

Zhang, C., Bi, J., Xu, S., Ramentol, E., Fan, G., Qiao, B., & Fujita, H. (2019). Multi-Imbalance: An open-source software for multi-class imbalance learning. *Knowledge-Based Systems*, *174*, 137–143. https://doi.org/10.1016/j.knosys.2019.03.001

Κύρκος, Ε. (2016). *Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων*. http://repository.kallipos.gr/handle/11419/1226

ΜΕΡΟΣ Γ΄ ΕΡΓΑΣΤΗΡΙΑΚΟ ΜΕΡΟΣ - WEKA

Ενότητα 1 Μεθοδολογία έρευνας

1.1 Εισαγωγή

Η παρούσα εργασία αφορά στην εφαρμογή των πιο εξελιγμένων μεθόδων εξόρυξης γνώσης σε ένα πραγματικό σύνολο δεδομένων προκειμένου να ανιχνευτεί πιθανή απάτη κατά το άνοιγμα ενός τραπεζικού λογαριασμού.

Η συμβολή της παρούσας εργασίας επικεντρώνεται σε δύο τομείς. Πρώτον επιχειρεί να αναδειχθούν οι ισχυρότερες μεταβλητές που αποτελούν σημαντικές ενδείξεις για πιθανή διάπραξη δόλιων συναλλαγών. Ο εντοπισμός, ιεράρχηση κατά σπουδαιότητα και αξιολόγηση αυτών των μεταβλητών αποτελεί σημαντικό κριτήριο για την αποδοχή ή μη των εν δυνάμει πελατών από ένα χρηματοπιστωτικό ίδρυμα. Δεύτερον προτείνει τη δημιουργία αξιόπιστων μοντέλων που να μπορούν να προβλέπουν με υψηλή ακρίβεια την πλειονότητα μελλοντικών περιπτώσεων.

Για την επίτευξη των στόχων, χρησιμοποιούνται διάφορες μέθοδοι. Αυτές περιλαμβάνουν την προεπεξεργασία των δεδομένων, την επιλογή χαρακτηριστικών και την εξισορρόπηση της κατανομής των κλάσεων, την ανάπτυξη μοντέλων με αλγορίθμους μηχανικής μάθησης, την επικύρωση των μοντέλων έναντι άγνωστων παρατηρήσεων και τέλος την ταξινόμηση κατά σειρά σημαντικότητας των μεταβλητών που οδηγούν στην παραγωγή των προτύπων.

1.2 Σύνολο δεδομένων data set

Συλλογή δεδομένων

Το σύνολο των δεδομένων προέρχεται από το αποθετήριο της Kaggle https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022 Η Kaggle είναι η μεγαλύτερη πλατφόρμα διαγωνισμών data science και αποτελεί μια διαδικτυακή κοινότητα επιστημόνων δεδομένων και επαγγελματιών μηχανικής μάθησης υπό την εποπτεία της Google.

Περιγραφή δεδομένων

Είναι ένα ιδιωτικό σύνολο δεδομένων αποτελούμενο από ένα εκατομμύριο μεμονωμένες αιτήσεις για άνοιγμα τραπεζικού λογαριασμού σε μια μεγάλη εμπορική τράπεζα στην Πορτογαλία. Οι αιτήσεις υποβλήθηκαν ηλεκτρονικά (μέσω φόρμας e-banking) σε χρονικό διάστημα οκτώ μηνών, από Φεβρουάριο έως Σεπτέμβριο.

Λόγω του ότι περιέχονται ευαίσθητα και προστατευμένα χαρακτηριστικά όπως ηλικία, εισόδημα ή επαγγελματική κατάσταση και για να διατηρηθεί ένας βαθμός ιδιωτικότητας έχει προστεθεί ήδη στο αρχικό data set θόλωμα μέσω της εφαρμογής ενός μηχανισμού θορύβου Laplacian. Ένας περιορισμός της έρευνας είναι ότι το data set περιλαμβάνει μόνο τις αιτήσεις εκείνες που έγιναν αποδεκτές από την τράπεζα οπότε για περιπτώσεις αιτήσεων που απορρίφθηκαν λόγω κανονιστικών διατάξεων (για την καταπολέμηση ξεπλύματος βρώμικου χρήματος) ή επιχειρησιακών κριτηρίων που πιθανόν θέτει το χρηματοπιστωτικό ίδρυμα (λόγου χάριν δεν γίνονται αποδεκτοί πελάτες με ηλικία μικρότερη των 18) δεν έχουμε καμία πληροφόρηση.

Αναλυτικά τα πεδία κάθε στήλης είναι (από το Dataset Suite Datasheet):

- **εισόδημα** (αριθμητικό): Ετήσιο εισόδημα του αιτούντος (σε δεκατημόριο). Κυμαίνεται μεταξύ
  [0.1, 0.9].
- **name_email_similarity** (αριθμητικό): Μετρική της ομοιότητας μεταξύ του ηλεκτρονικού ταχυδρομείου και του ονόματος του αιτούντος. Υψηλότερες τιμές αντιπροσωπεύουν μεγαλύτερη ομοιότητα. Κυμαίνεται μεταξύ [0, 1].
- **prev_address_months_count** (αριθμητικό): Αριθμός μηνών στην προηγούμενη καταγεγραμμένη διεύθυνση του αιτούντος, *δηλαδή στην* προηγούμενη κατοικία του αιτούντος, κατά περίπτωση. Κυμαίνεται μεταξύ [-1, 380] μηνών (-1 είναι ελλιπής τιμή).
- **current_address_months_count** (αριθμητικό): Μήνες στην επί του παρόντος εγγεγραμμένη διεύθυνση του αιτούντος. Κυμαίνεται μεταξύ [-1, 429] μηνών (-1 είναι ελλιπής τιμή).
- **customer_age** (αριθμητικό): Η ηλικία του αιτούντος σε έτη, στρογγυλοποιημένη στη δεκαετία. Κυμαίνεται μεταξύ [10, 90] ετών.
- **days_since_request** (αριθμητικό): Αριθμός ημερών που πέρασαν από τότε που έγινε η αίτηση. Κυμαίνεται μεταξύ [0, 79] ημερών.
- **intended_balcon_amount** (αριθμητικό): Αρχικό μεταφερόμενο ποσό για την εφαρμογή. Κυμαίνεται μεταξύ [-16, 114].
- **payment_type** (κατηγορηματικό): Τύπος πιστωτικού προγράμματος πληρωμής. 5 πιθανές (ανωνυμοποιημένες) τιμές.
- **zip_count_4w** (αριθμητικό): Αριθμός αιτήσεων στον ίδιο ταχυδρομικό κώδικα τις τελευταίες 4 εβδομάδες. Κυμαίνεται μεταξύ [1, 6830].
- **velocity_6h** (αριθμητικό): Ταχύτητα των συνολικών αιτήσεων που έγιναν τις τελευταίες 6 ώρες, *δηλαδή*, μέσος αριθμός αιτήσεων ανά ώρα τις τελευταίες 6 ώρες. Κυμαίνεται μεταξύ [-175, 16818].
- **velocity_24h** (αριθμητικό): Ταχύτητα του συνόλου των αιτήσεων που έγιναν τις τελευταίες 24 ώρες, *δηλαδή* μέσος αριθμός αιτήσεων ανά ώρα τις τελευταίες 24 ώρες. Κυμαίνεται μεταξύ [1297, 9586].
- **velocity_4w** (αριθμητικό): Ταχύτητα των συνολικών αιτήσεων που έγιναν τις τελευταίες 4 εβδομάδες, *δηλαδή* μέσος αριθμός αιτήσεων ανά ώρα τις τελευταίες 4 εβδομάδες. Κυμαίνεται μεταξύ [2825, 7020].

- **bank_branch_count_8w** (αριθμητικό): Αριθμός συνολικών αιτήσεων στο επιλεγμένο υποκατάστημα τράπεζας τις τελευταίες 8 εβδομάδες. Κυμαίνεται μεταξύ [0, 2404].
- **date_of_birth_distinct_emails_4w** (αριθμητικό): Αριθμός μηνυμάτων ηλεκτρονικού ταχυδρομείου για αιτούντες με την ίδια ημερομηνία γέννησης τις τελευταίες 4 εβδομάδες. Κυμαίνεται μεταξύ [0, 39].
- **employment_status** (κατηγορηματικό): Κατάσταση απασχόλησης του αιτούντος. 7 πιθανές (ανωνυμοποιημένες) τιμές.
- **credit_risk_score** (αριθμητικό): Εσωτερική βαθμολογία κινδύνου εφαρμογής. Κυμαίνεται μεταξύ
  [-191, 389].

- **email_is_free** (δυαδικό): Τομέας του ηλεκτρονικού ταχυδρομείου της εφαρμογής (είτε δωρεάν είτε επί πληρωμή).
- **housing_status** (κατηγορηματικό): Τρέχουσα κατάσταση κατοικίας για τον αιτούντα. 7 πιθανές (ανώνυμες) τιμές.
- **phone_home_valid** (δυαδικό): Εγκυρότητα του παρεχόμενου οικιακού τηλεφώνου.
- **phone_mobile_valid** (δυαδικό): Εγκυρότητα του παρεχόμενου κινητού τηλεφώνου.
- **bank_months_count** (αριθμητικό): σε μήνες: Πόσο παλιός είναι ο προηγούμενος λογαριασμός (αν τηρείται) σε μήνες. Κυμαίνεται μεταξύ [-1, 32] μηνών (το -1 είναι ελλιπής τιμή).
- **has_other_cards** (δυαδική): Εάν ο αιτών έχει και άλλες κάρτες από την ίδια τραπεζική εταιρεία.

- **proposed_credit_limit** (αριθμητικό): Το προτεινόμενο πιστωτικό όριο του αιτούντος. Κυμαίνεται μεταξύ
[200, 2000].
- **foreign_request** (δυαδικό): Εάν η χώρα προέλευσης του αιτήματος είναι διαφορετική από τη χώρα της τράπεζας.
- **πηγή** (κατηγορική): Διαδικτυακή πηγή της αίτησης. Είτε πρόγραμμα περιήγησης (INTERNET) είτε εφαρμογή (TELEAPP).
- **session_length_in_minutes** (αριθμητικό): Διάρκεια της συνεδρίας του χρήστη στον τραπεζικό ιστότοπο σε λεπτά. Κυμαίνεται μεταξύ [-1, 107] λεπτά.
- **device_os** (κατηγορηματικό): Λειτουργικό σύστημα της συσκευής που έκανε την αίτηση. Πιθανές τιμές είναι: Windows, macOS, Linux, X11 ή άλλες.
- **keep_alive_session** (δυαδικό): Επιλογή χρήστη κατά την αποσύνδεση από τη συνεδρία.
- **device_distinct_emails** (αριθμητικό): Αριθμός διακριτών μηνυμάτων ηλεκτρονικού ταχυδρομείου στον τραπεζικό ιστότοπο από τη χρησιμοποιούμενη συσκευή τις τελευταίες 8 εβδομάδες. Κυμαίνεται μεταξύ [-1, 2].
- **device_fraud_count** (αριθμητικό): Αριθμός δόλιων αιτήσεων με χρησιμοποιημένη συσκευή. Κυμαίνεται μεταξύ [0, 1].
- **μήνας** (αριθμητικό): Μήνας κατά τον οποίο υποβλήθηκε η αίτηση. Κυμαίνεται μεταξύ [0, 7].
- **fraud_bool** (δυαδικό): Εάν η αίτηση είναι δόλια ή όχι.

Η κλάση είναι δυαδική με τιμές 1 για δόλια αίτηση ανοίγματος τραπεζικού λογαριασμού και 0 για νόμιμη αίτηση. Σημειώνεται ότι για τις αιτήσεις που γίνονται δεκτές παρέχεται αυτόματα κάποιο πιστωτικό όριο. Επίσης για όλους τους λογαριασμούς δόθηκε σημαντικό χρονικό διάστημα για να παρακολουθηθεί η δραστηριότητα των πελατών και η κατηγοριοποίησή τους ως απάτη ή όχι λήφθηκε αρκετούς μήνες μετά το άνοιγμά τους.

Προεπεξεργασία δεδομένων

Το αρχικό σύνολο δεδομένων αποτελείται από 1.000.000 αιτήσεις (γραμμές) και 32 χαρακτηριστικά (στήλες) με αποτέλεσμα λόγω του μεγέθους του να μην είναι εύκολα διαχειρίσιμο. Επιπλέον η περίπτωση απάτης αντιστοιχεί σε ποσοστό από 0,85% έως 1,5% οπότε αντιμετωπίζουμε πρόβλημα έντονης ανισοκατανομής των κλάσεων. Αρχικά επιλέχθηκαν τα γνωρίσματα τα οποία θεωρήθηκαν ότι περιέχουν ουσιαστική πληροφορία που σχετίζεται με την παρούσα ανάλυση οπότε οι στήλες μειώθηκαν από 32 σε 19. Παράλληλα για την εξισορρόπηση των κλάσεων και λόγω έλλειψης διαθέσιμου κατάλληλου υπολογιστικού υλικού, επιλέχθηκε η μέθοδος random undersampling σε περιβάλλον python και προέκυψε ένα νέο σύνολο δεδομένων με 22.058 γραμμές και 19 στήλες με αναλογία 1:1 δηλαδή αποτελούμενο από 11.029 περιπτώσεις για κάθε μία από τις τιμές της κλάσης (απάτη ή μη απάτη).

Στη συνέχεια εφαρμόστηκε επιλογή των σημαντικότερων χαρακτηριστικών με βάση την συσχέτιση χρησιμοποιώντας την μέθοδο CFS (Correlation-Based Feature Selection) που διατυπώθηκε από τον Hall (1999). Στόχος της μεθόδου είναι η εύρεση χαρακτηριστικών τα οποία είναι ισχυρά συσχετισμένα (correlated) με τη μεταβλητή της κλάσης αλλά ασθενώς συσχετισμένα μεταξύ τους.

Από την εφαρμογή της προέκυψε ότι από τα 19 χαρακτηριστικά προτείνονται τα 17 άρα όλες οι στήλες είναι σημαντικές. Αυτό είναι αναμενόμενο γιατί από το αρχικό σύνολο δεδομένων ήδη οι 19 στήλες επιλέχθηκαν εκ των 32 αρχικών ως πιο σημαντικές κατά την υποδειγματοληψία. Επίσης κατά την δοκιμή των αλγορίθμων με ή χωρίς την επιλογή CFS δεν προκύπτουν μεγάλες αποκλίσεις στην απόδοση των μοντέλων (Παράρτημα, Πίνακες 3 & 4).

1.3 Μοντέλα Κατηγοριοποίησης και άλλοι αλγόριθμοι

Για την ανάλυση χρησιμοποιήθηκε το WEKA, ένα ελεύθερο λογισμικό ανοιχτού κώδικα από τα πλέον αναγνωρισμένα για Εξόρυξη Δεδομένων, που παρέχεται μέσω της ιστοσελίδας του Πανεπιστημίου WAIKATO της Νέας Ζηλανδίας.

Ως αναλυτικά εργαλεία επιλέχθηκαν μέθοδοι ταξινόμησης που προέρχονται από τη Μηχανική Μάθηση και εφαρμόστηκαν οκτώ από τους πιο γνωστούς ταξινομητές για την

ανάπτυξη των μοντέλων. Αυτοί είναι: Δέντρα Αποφάσεων (J48), Μπαϋεσιανά Δίκτυα (Bayes Net), η Λογιστική Παλινδρόμηση (LR), Rantom Forests (RF), Νευρωνικά Δίκτυα τύπου Multilayer Perceptron (MLP), οι Μηχανές Διανυσμάτων Υποστήριξης (SMO), K- πλησιέστεροι γείτονες (IBK) και ενισχυμένοι κατηγοριοποιητές που μαθαίνουν από την εμπειρία και τις προηγούμενες λάθος ταξινομήσεις όπως ο AdaBoostM1.

Η επικύρωση των μοντέλων έγινε με τη μέθοδο της Διασταυρούμενης Επικύρωσης 10 τμημάτων ή 10 Fold Cross Validation με σκοπό να εκτιμηθεί η ικανότητά τους να προβλέπουν την κλάση άγνωστων παρατηρήσεων. Ο καθορισμός της ακρίβειας ενός μοντέλου είναι σημαντικός αφενός γιατί προσδιορίζει την δυνατότητα ή μη του μοντέλου να χρησιμοποιηθεί για τη λήψη αποφάσεων στον πραγματικό κόσμο, αφετέρου γιατί επιτρέπει την σύγκριση διαφορετικών μοντέλων και την επιλογή του καλύτερου.

Σύμφωνα με την 10 Fold Cross Validation το σύνολο δεδομένων χωρίζεται σε 10 υποσύνολα. Η επιλογή των υποσυνόλων είναι τυχαία. Ένα από τα υποσύνολα χρησιμοποιείται ως σύνολο επικύρωσης (validation set) και τα υπόλοιπα εννέα συνενώνονται και δημιουργούν το σύνολο εκπαίδευσης (training set). Η διαδικασία επαναλαμβάνεται δέκα φορές, χρησιμοποιώντας κάθε φορά διαφορετικό fold για τη δοκιμή και τα υπόλοιπα εννέα ως σύνολο εκπαίδευσης. Στο τέλος υπολογίζεται η μέση επίδοση του μοντέλου.

1.4 Σημαντικότητα μεταβλητών

Στην συνέχεια διερευνήθηκε ποιες μεταβλητές επηρέασαν την κατασκευή των μοντέλων και εκτιμήθηκε η σημαντικότητά τους κατά σειρά σπουδαιότητας. Για τον σκοπό αυτό εφαρμόστηκαν οι παρακάτω οκτώ μέθοδοι οι οποίες περιγράφονται επιγραμματικά.

ChiSquaredAttributeEval: Αξιολογεί την αξία ενός χαρακτηριστικού υπολογίζοντας την τιμή του στατιστικού chi-squared σε σχέση με την κλάση, έχοντας ως έγκυρες επιλογές: M (αντιμετωπίζει τις ελλείπουσες τιμές ως ξεχωριστή τιμή), B (απλά γίνεται δυαδική ανάλυση αριθμητικών χαρακτηριστικών αντί να τα διακριτοποιήσει ορθά)

GainRatioAttributeEval: Αξιολογεί την αξία ενός χαρακτηριστικού μετρώντας τον λόγο

κέρδους σε σχέση με την κλάση

GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute)

Οι έγκυρες επιλογές είναι οι εξής: M (αντιμετωπίζει τις ελλείπουσες τιμές ως ξεχωριστή τιμή)

PairwiseConsistencyAttributeEval: Εκτιμητής χαρακτηριστικών που αξιολογεί την αξία ενός χαρακτηριστικού i προσθέτοντας τα ποσοστά συνέπειας των υποσυνόλων χαρακτηριστικών που αποτελούνται από το χαρακτηριστικό i και καθένα από τα άλλα χαρακτηριστικά.

PairwiseCorrelationAttributeEval: Αξιολογεί την αξία ενός χαρακτηριστικού i υπολογίζοντας το μέσο όρο των αξιών (χρησιμοποιώντας το CfsSubsetEval) των υποσυνόλων χαρακτηριστικών που αποτελούνται από το χαρακτηριστικό i και κάθε ένα από τα άλλα χαρακτηριστικά. Προτιμώνται χαρακτηριστικά με χαμηλή συσχέτιση με άλλα χαρακτηριστικά και υψηλή συσχέτιση με την κλάση.

OneRAttributeEval: Κλάση για την αξιολόγηση των χαρακτηριστικών ξεχωριστά με τη χρήση του ταξινομητή OneR.

ClassifierAttrributeEval + J48: Αξιολογεί την αξία ενός χαρακτηριστικού χρησιμοποιώντας έναν ταξινομητή που έχει καθορίσει ο χρήστης.

ReliefFAttributeEval: Αξιολογεί την αξία ενός χαρακτηριστικού με επαναλαμβανόμενη δειγματοληψία μιας περίπτωσης και λαμβάνοντας υπόψη την τιμή του συγκεκριμένου χαρακτηριστικού για την πλησιέστερη περίπτωση της ίδιας και διαφορετικής κλάσης. Μπορεί να λειτουργήσει τόσο σε δεδομένα διακριτών όσο και συνεχών κλάσεων.

SignificanceAttributeEval: Αξιολογεί την αξία ενός χαρακτηριστικού υπολογίζοντας την Πιθανολογική Σημασία ως συνάρτηση δύο κατευθύνσεων.

(συσχέτιση κλάσεων-χαρακτηριστικών και κλάσεων-χαρακτηριστικών)

Ενότητα 2 Αποτελέσματα και συμπεράσματα

2.1 Αποτελέσματα ανάλυσης κατηγοριοποίησης

Παρατίθενται παρακάτω οι επιδόσεις καθενός από τα παραπάνω παραγόμενα μοντέλα. Παρατηρούμε ότι η πλειονότητα των οκτώ μοντέλων πέτυχαν υψηλούς ρυθμούς ακρίβειας έναντι του συνόλου εκπαίδευσης, οι οποίοι κυμαίνονται από 75,97% έως 79,01%, με εξαίρεση τον αλγόριθμο IBk που είχε απόδοση μόλις 70.38%.

Παρουσιάζονται αναλυτικά οι επιδόσεις κάθε μοντέλου στο παράρτημα ενώ στον πίνακα παρακάτω πραγματοποιείται σύγκριση συγκεντρωτικά και ανά κλάση.

Πίνακας 5: Μέτρηση απόδοσης ταξινομητών

Method	Accuracy	TP Rate	Precision	Recall	F1score	ROC AUC
J48	**75,97**					
Class 0		75.7	76.1	75.7	75.9	78.8
Class 1		76.2	75.9	76.2	76.0	78.8
Average		76.0	76.0	76.0	76.0	78.8
Bayes net	**78,65**					
Class 0		80.1	77.8	80.1	79.0	86.8
Class 1		77.2	79.5	77.2	78.3	86.8
Average		78.7	78.7	78.7	78.6	86.8
Logistic	**78,94**					
Class 0		79.4	78.7	79.4	79.0	87.0
Class 1		78.5	79.2	78.5	78.8	87.0
Average		78.9	78.9	78.9	78.9	87.0
SMO	**79,01**					
Class 0		80.0	78.4	80.0	79.2	79.0
Class 1		78.0	79.6	78.0	78.8	79.0
Average		79.0	79.0	79.0	79.0	79.0
IBk	**70,38**					
Class 0		71.2	70.0	71.2	70.6	70.4
Class 1		69.5	70.7	69.5	70.1	70.4
Average		70.4	70.4	70.4	70.4	70.4

Πίνακας 5: Μέτρηση απόδοσης ταξινομητών

Method	Accuracy	TP Rate	Precision	Recall	F1score	ROC AUC
R.Forest	**78,39**					
Class 0		78.9	78.1	78.9	78.5	86.0
Class 1		77.9	78.7	77.9	78.3	86.0
Average		78.4	78.4	78.4	78.4	86.0
AdaboostM1	**77,29**					
Class 0		80.2	75.8	80.2	77.9	85.6
Class 1		74.4	79.0	74.4	76.6	85.6
Average		77.3	77.4	77.3	77.3	85.6
MLP	**77,87**					
Class 0		77.9	77.9	77.9	77.9	85.5
Class 1		77.9	77.9	77.9	77.9	85.5
Average		77.9	77.9	77.9	77.9	85.5

Τα ευρήματα αναδεικνύουν ως καλύτερους κατηγοριοποιητές την μέθοδο SMO με ακρίβεια 79,01%, ορθή πρόβλεψη της ακριβής κλάσης 78% και περιοχή κάτω από την καμπύλη ROC AUC 79%. Εξίσου πολύ καλή απόδοση έχει η Logistic Regression με ακρίβεια 78,94%, ορθή πρόβλεψη της ακριβής κλάσης 78,5% και ROC AUC 87%. Έναντι των υπολοίπων μεθόδων υπερτερούν τα Μπαυεσιανά Δίκτυα με 78,65% ακρίβεια και ROC AUC 86,8% και τα Δέντρα Αποφάσεων με 78,39% ακρίβεια και ROC AUC 86%.

2.2 Αποτελέσματα ταξινόμησης σημαντικότητας μεταβλητών

Η ανίχνευση των σημαντικών χαρακτηριστικών που επηρεάζουν την εκχώρηση ετικέτας στις περιπτώσεις είναι ένας άλλος κύριος στόχος της παρούσας εργασίας. Εφαρμόστηκαν 8 μέθοδοι αξιολόγησης, οι οποίες παρουσιάστηκαν συνοπτικά στην προηγούμενη ενότητα και τα αναλυτικά αποτελέσματα παρατίθενται στο παράρτημα ανά μέθοδο. Υπολογίζουμε ένα τελικό σκορ για τη σημαντικότητα κάθε μεταβλητής που προκύπτει από την σειρά κατάταξης κάθε μεθόδου. Ο αριθμός αυτός υποδεικνύει την κατάταξή της υψηλότερα με μέτρο την σημαντικότητα.

Πίνακας 6: Σύνοψη σημαντικότητας χαρακτηριστικών

#	Attribute	Score
1	housing_status	8
2	device_os	18
3	credit_risk_score	43
4	keep_alive_session	43
5	prev_address_months_count	49
6	customer_age	56
7	income	67
8	has_other_cards	68
9	employment_status	76
10	date_of_birth_distinct_emails_4w	77
11	phone_home_valid	82
12	name_email_similarity	90
13	payment_type	90
14	device_distinct_emails_8w	101
15	email_is_free	108
16	month	127
17	foreign_request	130
18	phone_mobile_valid	135

Είναι αξιοσημείωτο ότι το χαρακτηριστικό «housing_status» (τρέχουσα κατάσταση κατοικίας) ψηφίζεται από το σύνολο των μεθόδων ως πιο σημαντικό με αποτέλεσμα να αποτελεί την μεταβλητή η οποία επηρεάζει σε μεγαλύτερο βαθμό το αποτέλεσμα της κατηγοριοποίησης. Αυτό το εύρημα συνάδει με την πιστοδοτική πολιτική που συνήθως

ακολουθούν τα πιστωτικά ιδρύματα και προϋποθέτει την ύπαρξη σταθερής και μόνιμης κατοικίας για την σύναψη σχέσης και ιδιαίτερα για την χορήγηση οποιουδήποτε ορίου πίστωσης.

Μία ακόμα πολύ ενδιαφέρουσα στήλη στο σύνολο δεδομένων αποδεικνύεται η δεύτερη στην κατάταξη «device_os» και αναφέρεται στο λειτουργικό σύστημα της συσκευής που έκανε την αίτηση με πιθανές τιμές: windows, macOS, Linux, X11 ή άλλες.

Η προσεκτική εξέταση των αποτελεσμάτων σημαντικότητας αποκαλύπτει και άλλες ενδιαφέρουσες σχέσεις. Η τέταρτη μεταβλητή «keep_alive_session» αφορά την επιλογή του χρήστη να μπορεί να αποσυνδεθεί από την συνεδρία και η πέμπτη μεταβλητή «prev_address_months_count» που αφορά στον αριθμό μηνών στην προηγούμενη καταγεγραμμένη διεύθυνση του αιτούντος επανέρχεται πάλι στο βασικό κριτήριο της μόνιμης κατοικίας.

Οι επόμενες μεταβλητές «customer_age», «income», «has_other_cards» και «employment_status» αποτελούν μερικά από τα βασικότερα στοιχεία πιστοποίησης των πελατών και συναλλασσομένων με τα πιστωτικά ιδρύματα, οπότε είναι αναμενόμενα να βρίσκονται υψηλά στην ταξινόμηση.

2.3 Συμπεράσματα, περιορισμοί και προτάσεις για μελλοντική έρευνα

Η παρούσα εργασία σκοπό έχει να αναδείξει την τεράστια συμβολή των τεχνικών Εξόρυξης Δεδομένων για σκοπούς αποφυγής Ξεπλύματος Βρώμικου Χρήματος και νομιμοποίησης δόλιων συναλλαγών. Χρηματοπιστωτικοί οργανισμοί, ασφαλιστικές, χρηματιστηριακές αλλά και μεγάλες εμπορικές επιχειρήσεις αξιοποιούν καθημερινά τις σύγχρονες τεχνολογίες τεχνητής νοημοσύνης προκειμένου να διασφαλίσουν την συμμόρφωση με τους κανόνες δεοντολογίας αλλά και την πραγματοποίηση των συναλλαγών τους με διαφάνεια μειώνοντας ταυτόχρονα τη δυνατότητα απάτης.

Ως κύριο πεδίο εφαρμογής της εργασίας επιλέγεται η πρώτη επαφή μιας οντότητας με

ένα τραπεζικό ίδρυμα (συγκεκριμένα κατά το άνοιγμα λογαριασμού), η οποία είναι ιδιαίτερα κρίσιμη για την μετέπειτα εξέλιξη της συνεργασίας τους. Επιπλέον το συγκεκριμένο κομμάτι δεν έχει ερευνηθεί συστηματικά και στον ίδιο βαθμό συγκριτικά με την ευρύτερη τραπεζική-οικονομική απάτη όπως λόγου χάρη ύποπτες συναλλαγές σε λογαριασμούς, απάτη με χρήση πιστωτικών καρτών ή πιστοληπτική ικανότητα δανειοληπτών.

Επιλέγεται ο τομέας των τραπεζικών ιδρυμάτων διότι η λειτουργία τους υπόκειται σε πολύ αυστηρούς κανονισμούς αλλά και συγκεκριμένες επιχειρησιακές πολιτικές, που καθιστούν αναγκαία την σωστή επιλογή υποψήφιων πελατών και συνεργατών, προς αποφυγή κυρώσεων, οικονομικής ζημίας αλλά και πιθανότητας δυσφήμισης. Επιπλέον το σύγχρονο ψηφιακό πλαίσιο λειτουργίας και η δυνατότητα απομακρυσμένης εξυπηρέτησης (online αίτηση και άνοιγμα τραπεζικού καταθετικού λογαριασμού χωρίς προσωπική επαφή με τραπεζικό υπάλληλο) αποτελούν μια ακόμα πρόκληση στο δυναμικά μεταβαλλόμενο τραπεζικό περιβάλλον, εγκυμονούν ιδιαίτερους κινδύνους (όπως αδυναμία ελέγχου ακρίβειας και πληρότητας δεδομένων που υποβάλλονται ηλεκτρονικά) και εντείνουν την επιφυλακή στην προσπάθεια ανίχνευσης της απάτης σε πρώιμο στάδιο. Με την χρήση τεχνολογιών τεχνητής νοημοσύνης είναι εφικτή η ανάλυση δεδομένων μεγάλου όγκου σε πραγματικό χρόνο (real time alerts).

Ένας από τους πρωταρχικούς στόχους της παρούσας εργασίας είναι η ανάπτυξη αξιόπιστων μοντέλων, ικανών να προβλέπουν την πιθανότητα απάτης κατά το άνοιγμα ενός τραπεζικού λογαριασμού. Όλα τα μοντέλα πρόβλεψης που εφαρμόζονται, με εξαίρεση τους IBK πλησιέστερους γείτονες, είναι επιτυχή. Τα επιτευχθέντα ποσοστά ακρίβειας είναι υψηλά και κυμαίνονται μεταξύ 75,97% έως 79,01%. Οι Μηχανές Διανυσμάτων Υποστήριξης και η Λογιστική Παλινδρόμηση υπερτερούν των άλλων ταξινομητών σε όλες τις χρησιμοποιούμενες μετρήσεις.

Όσον αφορά τον εντοπισμό των σημαντικότερων παραγόντων για την κατηγοριοποίηση μιας αίτησης ως απάτη ή μη εφαρμόστηκαν οκτώ μέθοδοι από τις οποίες αναδείχτηκαν οι σπουδαιότεροι παράγοντες που δύναται να συνηγορούν για περιπτώσεις δόλιων αιτήσεων. Η μεταβλητή η οποία τοποθετείται υψηλότερα στην ιεράρχηση αυτών των παραγόντων είναι η ύπαρξη μόνιμης διεύθυνσης κατοικίας του αιτούντα.

Η παρούσα εργασία όπως ήδη αναφέρθηκε έχει αρκετούς περιορισμούς. Ο

σημαντικότερος είναι ότι θα μπορούσαν να διερευνηθούν και να εξεταστούν πολλές άλλες προσεγγίσεις για την αντιμετώπιση της ανισόρροπης κατανομής των κλάσεων, κάτι που δεν κατέστη δυνατό λόγω του υπερβολικά μεγάλου μεγέθους του data set. Εξαιρετικά μεγάλα σύνολα δεδομένων χρειάζονται μεγάλους χρόνους εκπαίδευσης και υψηλές απαιτήσεις σε μνήμη υπολογιστή.

Σχετικά με το αρχικό σύνολο δεδομένων οι κυριότεροι περιορισμοί είναι δύο. Ο πρώτος αφορά στην εμπιστευτική φύση των δεδομένων που αντιμετωπίστηκε ήδη από την αρχική ομάδα ερευνητών της Kaggle με προσθήκη θορύβου. Ο δεύτερος περιορισμός σχετίζεται με τον τρόπο απόκτησης των πληροφοριών. Τα περισσότερα πεδία σε κάθε αίτηση έχουν συμπληρωθεί από τον ίδιο τον αιτούντα κατά την καταχώριση της αίτησης στην ηλεκτρονική πλατφόρμα της τράπεζας με αποτέλεσμα να υπάρχουν εσφαλμένες καταχωρήσεις είτε σκόπιμα και κακόβουλα από απατεώνες είτε από αμέλεια από νόμιμους πελάτες. Δεν υπάρχει κάποια λύση για αυτόν τον περιορισμό. Επίσης όπως ήδη αναφέρθηκε, ένας ακόμα περιορισμός της έρευνας είναι ότι το data set περιλαμβάνει μόνο τις αιτήσεις εκείνες που έγιναν αποδεκτές από την τράπεζα οπότε για περιπτώσεις αιτήσεων που απορρίφθηκαν δεν έχουμε καμία γνώση αν ήταν όντως δόλιες.

Καταλήγοντας, στην παρούσα εργασία αποδείχθηκε η χρησιμότητα των προηγμένων αναλυτικών εργαλείων με την πρακτική εφαρμογή τους στα δεδομένα του πραγματικού κόσμου, υποστηρίζοντας την διαχείριση υποθέσεων και την λήψη ανάλογων αποφάσεων.

Πίνακας Εικόνων

Εικόνα 1: Ο κύκλος του Ξεπλύματος Χρήματος 16
Εικόνα 2: Χρήση τεχνολογιών ανάλυσης δεδομένων 23
Εικόνα 3: Μετασχηματισμός του τραπεζικού κλάδου 27
Εικόνα 4: Στάδια Ανακάλυψης Γνώσης 32
Εικόνα 5: Πίνακας σύγχυσης - confusion matrix 42
Εικόνα 6: Καμπύλη ROC 44

ΠΑΡΑΡΤΗΜΑ

Πίνακας 1: ΒΑΣΙΚΑ ΑΔΙΚΗΜΑΤΑ Ν4734/2020

α) η εγκληματική οργάνωση, όπως ορίζεται στο άρθρο 187 ΠΚ,

β) οι τρομοκρατικές πράξεις, η τρομοκρατική οργάνωση, η αξιόποινη υποστήριξή τους (χρηματοδότηση της τρομοκρατίας) και τα τρομοκρατικά εγκλήματα, όπως ορίζονται στα άρθρα 187Α και 187Β ΠΚ και στα άρθρα 32 έως 35 του ν. 4689/2020 (Α' 103),

γ) η δωροληψία και δωροδοκία υπαλλήλου, όπως ορίζονται στα άρθρα 235 και 236 ΠΚ,

δ) η εμπορία επιρροής-μεσάζοντες και δωροληψία και δωροδοκία στον ιδιωτικό τομέα, όπως ορίζονται στα άρθρα 237α και 237β ΠΚ,

ε) η δωροληψία και δωροδοκία πολιτικών προσώπων και δικαστικών λειτουργών, όπως ορίζονται στα άρθρα 159, 159α και 237 ΠΚ,

στ) η εμπορία ανθρώπων, όπως ορίζεται στο άρθρο 323α ΠΚ,

ζ) η απάτη με υπολογιστή, όπως ορίζεται στο άρθρο 386α ΠΚ,

η) η σωματεμπορία, όπως ορίζεται στο άρθρο 351 ΠΚ,

θ) το εμπόριο ναρκωτικών και τα αδικήματα που προβλέπονται στα άρθρα 20 έως και 23 του ν. 4139/2013 (Α' 74),

ι) εμπόριο όπλων, εκρηκτικών μηχανισμών και τα αδικήματα που προβλέπονται στα άρθρα 15 και 17 του ν. 2168/1993 (Α' 147),

ια) τα αδικήματα περί προστασίας αρχαιοτήτων και πολιτιστικής κληρονομιάς που προβλέπονται στα άρθρα 53, 54, 55, 61 και 63 του ν. 3028/2002 (Α' 153),

ιβ) τα αδικήματα που προβλέπονται στις παρ. 1 και 3 του άρθρου 8 του ν.δ. 181/1974 (Α' 347),

ιγ) τα αδικήματα περί μετανάστευσης που προβλέπονται στις παρ. 5 έως και 8 του άρθρου 29 και στο άρθρο 30 του ν. 4251/2014 (Α' 80),

ιδ) τα αδικήματα που προβλέπονται στα άρθρα τέταρτο και έκτο του ν. 2803/2000 (Α' 48),

ιε) τα χρηματιστηριακά αδικήματα που προβλέπονται στα άρθρα 28 έως και 31 του ν. 4443/2016 (Α' 232),

ιστ) τα αδικήματα:

αα) φοροδιαφυγής που προβλέπονται στο άρθρο 66 του ν. 4174/2013 (Α' 170) με την

εξαίρεση του πρώτου εδαφίου της παρ. 5,

ββ) λαθρεμπορίας που προβλέπονται στα άρθρα 155 έως και 157 του ν. 2960/2001 (Α’ 265),

γγ) μη καταβολής χρεών προς το Δημόσιο που προβλέπονται στο άρθρο 25 του ν. 1882/1990 (Α’ 43), με την εξαίρεση της περ. α’ της παρ. 1, καθώς και της μη καταβολής χρεών που προκύπτουν από χρηματικές ποινές ή πρόστιμα που έχουν επιβληθεί από τα δικαστήρια ή από διοικητικές και άλλες αρχές,

ιζ) τα αδικήματα που προβλέπονται στην παρ. 3 του άρθρου 28 του ν. 1650/1986 (Α’ 160),

ιη) κάθε άλλο αδίκημα από το οποίο προκύπτει περιουσιακό όφελος και τιμωρείται με ποινή φυλάκισης.»

Πίνακας 2: Κατάταξη νέων και υφιστάμενων πελατών στην κατηγορία υψηλού κινδύνου

(βάση της απόφασης ΕΤΠΘ 281/5/17.03.09 & ΠΔΤΕ 2652/29.02.12)

- Λογαριασμοί πελατών που δεν κατοικούν μόνιμα στην Ελλάδα.
- Πολιτικώς Εκτιθέμενα Πρόσωπα (PEP Politically Exposed Persons)

  Είναι τα φυσικά πρόσωπα στα οποία έχει ανατεθεί σημαντικό δημόσιο λειτούργημα, οι άμεσοι συγγενείς τους και τα πρόσωπα που είναι γνωστά ως στενοί συνεργάτες τους.
- Εταιρείες με ανώνυμες μετοχές μη εισηγμένες σε οργανωμένες αγορές της Ε.Ε.
- Λογαριασμοί off shore εταιριών και εταιριών ειδικού σκοπού.
- Σχήματα ή οντότητες στερούμενα νομικής προσωπικότητας που διαχειρίζονται κεφάλαια ή άλλες ομάδες περιουσιακών στοιχείων.
- Λογαριασμοί ενώσεων προσώπων μη κερδοσκοπικού χαρακτήρα.
- Έναρξη επιχειρηματικής σχέσης και συναλλαγές χωρίς την φυσική παρουσία του πελάτη.
- Διασυνοριακές σχέσεις τραπεζικής ανταπόκρισης.

  Πιο συγκεκριμένα τα πιστωτικά ιδρύματα απαγορεύεται να συνάπτουν ή να διατηρούν σχέση τραπεζικής ανταπόκρισης με εικονική τράπεζα (shell bank) ή με τράπεζα η οποία είναι γνωστό ότι επιτρέπει να χρησιμοποιούνται οι λογαριασμοί της από εικονικές τράπεζες
- Χώρες οι οποίες δεν εφαρμόζουν επαρκώς τις συστάσεις της FATF
- Επιχειρηματικές σχέσεις και συναλλαγές που ενέχουν αυξημένο κίνδυνο φοροδιαφυγής.
- Στην κατηγορία υψηλού κινδύνου εντάσσονται επίσης ελεύθεροι επαγγελματίες στους λογαριασμούς των οποίων πιστώθηκαν το προηγούμενο ημερολογιακό έτος ποσά άνω των € 200.000 και νομικά πρόσωπα με αντίστοιχο όριο τα € 300.000.

Πίνακας 3: Μέτρηση απόδοσης αλγορίθμων Μηχανικής Μάθησης ΧΩΡΙΣ επιλογή χαρακτηριστικών (CFS)

Method	Accuracy	TP Rate 0	TP Rate 1
J48	75,85	76,4	75,3
Bayes net	78,06	79,5	78,7
Logistic	79,06	79,5	78,7
SMO	79,06	80,0	78,1
IBk	70,12	70,8	69,5
Random Forest	78,60	79,1	78,1
AdaboostM1	77,29	80,2	74,4
MLP	77,18	76,1	78,3

Πίνακας 4: Μέτρηση απόδοσης αλγορίθμων Μηχανικής Μάθησης ΜΕ επιλογή χαρακτηριστικών (CFS)

Method	Accuracy	TP Rate 0	TP Rate 1
J48	75,97	75,7	76,2
Bayes net	78,65	80,1	77,2
Logistic	78,94	79,4	78,5
SMO	79,01	80,0	78,0
IBk	70,38	71,2	69,5
Random Forest	78,39	78,9	77,9
AdaboostM1	77,29	80,2	74,4
MLP	77,87	77,9	77,9

J48

```
Number of Leaves  :     1268

Size of the tree :      2118


Time taken to build model: 10.91 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       16758               75.9724 %
Incorrectly Classified Instances      5300               24.0276 %
Kappa statistic                          0.5194
Mean absolute error                      0.3002
Root mean squared error                  0.4385
Relative absolute error                 60.0488 %
Root relative squared error             87.6968 %
Total Number of Instances            22058

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,757    0,238    0,761      0,757    0,759      0,519    0,788     0,734     0
                 0,762    0,243    0,759      0,762    0,760      0,519    0,788     0,731     1
Weighted Avg.    0,760    0,240    0,760      0,760    0,760      0,519    0,788     0,733

=== Confusion Matrix ===

    a    b   <-- classified as
 8353 2676 |    a = 0
 2624 8405 |    b = 1
```

Bayes net

```
LogScore MDL: -333505.5410463418
LogScore ENTROPY: -333030.47309156583
LogScore AIC: -333125.47309156583


Time taken to build model: 1.64 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       17349               78.6517 %
Incorrectly Classified Instances      4709               21.3483 %
Kappa statistic                          0.573
Mean absolute error                      0.2508
Root mean squared error                  0.3936
Relative absolute error                 50.1518 %
Root relative squared error             78.718  %
Total Number of Instances            22058

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,801    0,228    0,778      0,801    0,790      0,573    0,868     0,865     0
                 0,772    0,199    0,795      0,772    0,783      0,573    0,868     0,866     1
Weighted Avg.    0,787    0,213    0,787      0,787    0,786      0,573    0,868     0,866

=== Confusion Matrix ===

    a    b   <-- classified as
 8834 2195 |    a = 0
 2514 8515 |    b = 1
```

Logistic Regression

```
device_os=x11                          1.0892
keep_alive_session                     2.0459
device_distinct_emails_8w              0.4207


Time taken to build model: 6.31 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       17412               78.9373 %
Incorrectly Classified Instances      4646               21.0627 %
Kappa statistic                          0.5787
Mean absolute error                      0.2917
Root mean squared error                  0.3821
Relative absolute error                 58.3384 %
Root relative squared error             76.4219 %
Total Number of Instances            22058

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,794    0,215    0,787      0,794    0,790      0,579    0,870     0,862     0
                 0,785    0,206    0,792      0,785    0,788      0,579    0,870     0,869     1
Weighted Avg.    0,789    0,211    0,789      0,789    0,789      0,579    0,870     0,866

=== Confusion Matrix ===

    a    b   <-- classified as
 8758 2271 |    a = 0
 2375 8654 |    b = 1
```

SMO

```
Number of kernel evaluations: 655229859 (23.624% cached)



Time taken to build model: 638.27 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       17427               79.0053 %
Incorrectly Classified Instances      4631               20.9947 %
Kappa statistic                          0.5801
Mean absolute error                      0.2099
Root mean squared error                  0.4582
Relative absolute error                 41.9893 %
Root relative squared error             91.6398 %
Total Number of Instances            22058

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,800    0,220    0,784      0,800    0,792      0,580    0,790     0,728     0
                 0,780    0,200    0,796      0,780    0,788      0,580    0,790     0,731     1
Weighted Avg.    0,790    0,210    0,790      0,790    0,790      0,580    0,790     0,729

=== Confusion Matrix ===

    a    b   <-- classified as
 8824 2205 |    a = 0
 2426 8603 |    b = 1
```

IBk

```
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       15526               70.3872 %
Incorrectly Classified Instances      6532               29.6128 %
Kappa statistic                          0.4077
Mean absolute error                      0.2961
Root mean squared error                  0.5441
Relative absolute error                 59.2298 %
Root relative squared error            108.8299 %
Total Number of Instances            22058

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,712    0,305    0,700      0,712    0,706      0,408    0,704     0,648     0
                 0,695    0,288    0,707      0,695    0,701      0,408    0,704     0,650     1
Weighted Avg.    0,704    0,296    0,704      0,704    0,704      0,408    0,704     0,649

=== Confusion Matrix ===

    a    b   <-- classified as
 7857 3172 |    a = 0
 3360 7669 |    b = 1
```

Random Forest

```
RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 38.6 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       17291               78.3888 %
Incorrectly Classified Instances      4767               21.6112 %
Kappa statistic                          0.5678
Mean absolute error                      0.3052
Root mean squared error                  0.3897
Relative absolute error                 61.034  %
Root relative squared error             77.9438 %
Total Number of Instances            22058

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,789    0,221    0,781      0,789    0,785      0,568    0,860     0,851     0
                 0,779    0,211    0,787      0,779    0,783      0,568    0,860     0,853     1
Weighted Avg.    0,784    0,216    0,784      0,784    0,784      0,568    0,860     0,852

=== Confusion Matrix ===

    a    b   <-- classified as
 8704 2325 |    a = 0
 2442 8587 |    b = 1
```

AdaboostM1

```
Time taken to build model: 6.67 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       17050               77.2962 %
Incorrectly Classified Instances      5008               22.7038 %
Kappa statistic                          0.5459
Mean absolute error                      0.3092
Root mean squared error                  0.3931
Relative absolute error                 61.8449 %
Root relative squared error             78.6136 %
Total Number of Instances            22058

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,802    0,256    0,758      0,802    0,779      0,547    0,856     0,852     0
                 0,744    0,198    0,790      0,744    0,766      0,547    0,856     0,852     1
Weighted Avg.    0,773    0,227    0,774      0,773    0,773      0,547    0,856     0,852

=== Confusion Matrix ===

    a    b   <-- classified as
 8847 2182 |    a = 0
 2826 8203 |    b = 1
```

MLP

```
Time taken to build model: 313.7 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       17176               77.8674 %
Incorrectly Classified Instances      4882               22.1326 %
Kappa statistic                          0.5573
Mean absolute error                      0.2769
Root mean squared error                  0.3967
Relative absolute error                 55.3736 %
Root relative squared error             79.3319 %
Total Number of Instances            22058

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,779    0,221    0,779      0,779    0,779      0,557    0,855     0,842     0
                 0,779    0,221    0,779      0,779    0,779      0,557    0,855     0,853     1
Weighted Avg.    0,779    0,221    0,779      0,779    0,779      0,557    0,855     0,848

=== Confusion Matrix ===

    a    b   <-- classified as
 8589 2440 |    a = 0
 2442 8587 |    b = 1
```

Synopsis of Attribute Significance

ChiSquaredAttributeEval	R1
housing_status	1
device_os	2
credit_risk_score	3
customer_age	4
prev_address_months_count	5
income	6
keep_alive_session	7
date_of_birth_distinct_emails_4w	8
employment_status	9
payment_type	10
has_other_cards	11
name_email_similarity	12
phone_home_valid	13
device_distinct_emails_8w	14
email_is_free	15
month	16
foreign_request	17
phone_mobile_valid	18

GainRatioAttributeEval	R2
housing_status	1
prev_address_months_count	2
device_os	3
has_other_cards	4
keep_alive_session	5
device_distinct_emails_8w	6
credit_risk_score	7
employment_status	8
customer_age	9
phone_home_valid	10
date_of_birth_distinct_emails_4w	11
income	12
name_email_similarity	13
foreign_request	14
payment_type	15
email_is_free	16
phone_mobile_valid	17
month	18

PairwiseConsistencyAttributeEval	R3
housing_status	1
device_os	2
credit_risk_score	3
keep_alive_session	4
customer_age	5
prev_address_months_count	6
income	7
date_of_birth_distinct_emails_4w	8
name_email_similarity	9
phone_home_valid	10
payment_type	11
has_other_cards	12
employment_status	13
email_is_free	14
device_distinct_emails_8w	15
month	16
foreign_request	17
phone_mobile_valid	18

PairwiseCorrelationAttributeEval	R4
housing_status	1
device_os	2
prev_address_months_count	3
keep_alive_session	4
credit_risk_score	5
has_other_cards	6
customer_age	7
date_of_birth_distinct_emails_4w	8
employment_status	9
income	10
device_distinct_emails_8w	11
phone_home_valid	12
name_email_similarity	13
payment_type	14
email_is_free	15
foreign_request	16
month	17
phone_mobile_valid	18

OneRAttributeEval	R5
housing_status	1
device_os	2
keep_alive_session	3
credit_risk_score	4
customer_age	5
income	6
prev_address_months_count	7
date_of_birth_distinct_emails_4w	8
phone_home_valid	9
payment_type	10
has_other_cards	11
email_is_free	12
employment_status	13
device_distinct_emails_8w	14
name_email_similarity	15
month	16
phone_mobile_valid	17
foreign_request	18

ClassifierAttrributeEval + J48	R6
housing_status	1
device_os	2
credit_risk_score	3
keep_alive_session	4
customer_age	5
income	6
prev_address_months_count	7
date_of_birth_distinct_emails_4w	8
name_email_similarity	9
phone_home_valid	10
payment_type	11
has_other_cards	12
email_is_free	13
employment_status	14
device_distinct_emails_8w	15
month	16
phone_mobile_valid	17
foreign_request	18

ReliefFAttributeEval	R7
housing_status	1
device_os	2
payment_type	3
employment_status	4
has_other_cards	5
name_email_similarity	6
phone_home_valid	7
income	8
email_is_free	9
month	10
keep_alive_session	11
customer_age	12
phone_mobile_valid	13
credit_risk_score	14
foreign_request	15
date_of_birth_distinct_emails_4w	16
prev_address_months_count	17
device_distinct_emails_8w	18

SignificanceAttributeEval	R8
housing_status	1
prev_address_months_count	2
device_os	3
credit_risk_score	4
keep_alive_session	5
employment_status	6
has_other_cards	7
device_distinct_emails_8w	8
customer_age	9
date_of_birth_distinct_emails_4w	10
phone_home_valid	11
income	12
name_email_similarity	13
email_is_free	14
foreign_request	15
payment_type	16
phone_mobile_valid	17
month	18