



Ανώτατο Εκπαιδευτικό Ίδρυμα Ιονίου Πανεπιστημίου

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πτυχιακή Εργασία:

**ΑΝΙΧΝΕΥΣΗ ΑΠΑΤΩΝ ΣΕ ΠΙΣΤΩΤΙΚΕΣ ΚΑΡΤΕΣ
ΜΕ ΧΡΗΣΗ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ**

Επιβλέπων Καθηγητής: κος Θέμης Έξαρχος

Φοιτητής: Σιδηρόπουλος Γεώργιος

Αριθμός Μητρώου: Π2015057

Πρόλογος

Η παρούσα πτυχιακή εργασία με τίτλο “Ανίχνευση απατών σε πιστωτικές κάρτες με χρήση μηχανικής μάθησης”, εκπονήθηκε στα πλαίσια της ολοκλήρωσης των προϋποθέσεων για την λήψη πτυχίου από το Ανώτατο Εκπαιδευτικό Ίδρυμα Ιονίου τμήματος Πληροφορικής. Το θέμα λήφθηκε τον Νοέμβριο του 2019, με υπεύθυνο καθηγητή τον κ. Θέμη Έξαρχο. Η ολοκλήρωση της πραγματοποιήθηκε τον Σεπτέμβρη του 2020. Σκοπός μου κατά την υλοποίηση της πτυχιακής εργασία ήταν η κατανόηση του προβλήματος και πως το προσεγγίζουμε. Επίσης, έγινε σημαντική προσπάθεια έτσι ώστε το περιεχόμενο να είναι κατανοητό από όλους, γι αυτό θα δείτε αρκετούς πίνακες και διαγράμματα.

Ευχαριστίες

Πρώτα από όλους θα ήθελα να ευχαριστήσω τον καθηγητή μου κ. Έξαρχο ο οποίος με βοήθησε σε σημαντικά σημεία της εργασίας με τις υποδείξεις του, έδειξε υπομονή με τις ερωτήσεις και τις απορίες που είχα και για αυτό τον λόγο ότι είμαι ευγνώμον. Έπειτα θα ήθελα να ευχαριστήσω την οικογένεια μου που με στήριξε απο την αρχή έως και το τέλος αυτή της διαδρομής, μιας διαδρομής που πέρασε από πολλά κύματα. Τέλος, θέλω να πω ένα μεγάλο ευχαριστώ στον Μاريو, τον Παύλο, τον Μιχάλη, τον Φάνη, τον Αριστοτέλη, τον Ξενοφώντα, τον Αρσένη, τον Θάνο και τον Βασίλη που ήταν πάντα εκεί για μένα και έκαναν αυτην την διαδρομή διασκεδαστική. Σας ευχαριστώ όλους.

Πρόλογος	3
Ευχαριστίες	3
Κεφάλαιο 1ο	5
Εισαγωγή	5
Τι είναι το Credit Card Fraud Detection και ποιά η ανάγκη για DSS στο Credit Card Fraud Detection	6
Κεφάλαιο 2ο	7
State of the Art	7
Σύνοψη της όλης εργασίας	43
Κεφάλαιο 3ο	44
Περιγραφή Data Set	44
Περιγραφή Μεθοδολογίας	45
Περιγραφή Αλγορίθμων	45
Κεφάλαιο 4ο	51
Αποτελέσματα	51
Κεφάλαιο 5ο	53
Πίνακας Σύγκρισης Εργασιών	53
Σύγκριση Εργασιών	57
Συμπεράσματα	61
Κεφάλαιο 6ο	63
Βιβλιογραφία	63
Εργαλεία και Πηγές	65

Κεφάλαιο 1ο

Εισαγωγή

Με την ευρεία εξάπλωση του διαδικτύου καθώς και με την αύξηση των ατόμων που έχουν πρόσβαση σε έναν ηλεκτρονικό υπολογιστή ο οποίος με την σειρά του είναι συνδεδεμένος στο διαδίκτυο, μπορούμε να πούμε ότι μπαίνουμε σε μία νέα μορφή κοινωνίας. Όπως και στον φυσικό κόσμο έτσι και στον ηλεκτρονικό η διαδικασία της αγοράς προϊόντων και υπηρεσιών, αποτελεί έναν από τους ακρογωνιαίους λίθους της . Πάραυτα παρατηρείται ότι πολλές εταιρείες και καταναλωτές σε όλο τον κόσμο ζημιώνονται μέσα από αυτές τις ηλεκτρονικές συναλλαγές. Ποιός είναι όμως ο λόγος για αυτές τις απώλειες ; Η αξιοποίηση των ικανοτήτων ορισμένων ατόμων με σκοπό την υποκλοπή χρημάτων από λογαριασμούς αγνώστων σε αυτούς καταναλωτών. Για να γίνει κατανοητό το μέγεθος του προβλήματος αρκεί να πούμε ότι το 2015 παρήχθησαν περισσότερα από 31 τρισεκατομμύρια δολάρια μέσα από διαδικτυακές συναλλαγές, ποσοστά που αυξήθηκαν κατά 7.3% συγκριτικά με το 2014.Βλέπουμε δηλαδή ότι η δυνατότητα των ανθρώπων να πραγματοποιούν αγορές μέσω του διαδικτύου αξιοποιείται σε πάρα πολύ υψηλό βαθμό. Δυστυχώς όμως, την ίδια χρονιά οι μετρήσεις που πραγματοποιήθηκαν πάνω στις απώλειες που οφείλονται στις απάτες οι οποίες στοχεύουν πιστωτικές κάρτες, έφτασαν τα 21 δισεκατομμύρια και πιθανολογείτε ότι θα φτάσουν τα 31 δισεκατομμύρια το 2020. Παρατηρώντας λοιπόν τα νούμερα αυτά μπορούμε να καταλάβουμε ότι υπάρχει μεγάλη ανάγκη για την ανάπτυξη συστημάτων για την αντιμετώπιση αυτού του είδους απατών. Αυτό είναι και το σημείο το οποίο καθιστά την ανίχνευση των απατών σε πιστωτικές κάρτες θέμα που απαιτεί ιδιαίτερη προσοχή

Τι είναι το Credit Card Fraud Detection και ποιά η ανάγκη για DSS στο Credit Card Fraud Detection

Η απάτη στις πιστωτικές κάρτες μπορεί να χαρακτηριστεί ως η χρήση της πιστωτικής κάρτας από κάποιον άλλο, ο οποίος δεν είναι ο πραγματικός της κάτοχος αλλά ούτε και εξουσιοδοτημένο μέλος όσον αφορά την χρήση της. Πιο συγκεκριμένα, τέτοιου είδους απάτες πραγματοποιούνται όταν κάποιος χρησιμοποιεί την πιστωτική κάρτα κάποιου άλλου για προσωπικούς λόγους χωρίς την άδεια του πραγματικού κατόχου της, με σκοπό είτε να αγοράσει διαδικτυακά προϊόντα για τον ίδιο , για άλλους ή ακόμα και για να μεταπωλήσει ορισμένα προϊόντα. Ο μηχανισμός ανίχνευσης απάτης με πιστωτικές κάρτες εξαρτάται από τον ίδιο τον μηχανισμό απάτης. Υπάρχουν διάφοροι φορείς με τους οποίους μπορεί να πραγματοποιηθεί η απάτη . Ωστόσο, αυτά μπορούν να ταξινομηθούν σε δύο κύριες κατηγορίες. Πρώτη από αυτές είναι οι απάτες που προκαλούνται από την παράνομη κατοχή της φυσικής κάρτας. Αυτό το είδος απάτης δεν μπορεί να υλοποιηθεί αποκλειστικά και μόνο μόνο με την κλοπή της κάρτας από τον πραγματικό ιδιοκτήτη πριν ή μετά την παράδοση, αλλά με πολλές άλλες μεθόδους, επίσης, όπου δημιουργείται μια νέα κλωνοποιημένη κάρτα που μπορεί να λειτουργήσει ως πλαστό για την πραγματική κάρτα. Η άλλη κατηγορία απάτης προκαλείται από την παράνομη λήψη πληροφοριών πιστωτικής κάρτας. Αυτό μπορεί να γίνει παίρνοντας αποτυπώματα καρτών σε ξενοδοχεία και phishing μεταξύ άλλων. Τα Fraud Detection Systems αποτελούν αυτοματοποιημένες λύσεις βασισμένες στην μηχανική μάθηση, τις οποίες λύσεις επιστρατεύουν οι εταιρίες προκειμένου να ανακαλύψουν απάτες προτού καν ο χρήστης μπει στην διαδικασία ενημέρωσης της τράπεζας για κάτι τέτοιο. Είναι πολύ σημαντικό να βρισκόμαστε πάντα σε θέση να ανιχνεύουμε πιθανές απάτες έτσι ώστε να υπάρχει μια αίσθηση ασφάλειας για όλον τους διαδικτυακούς αγοραστές. Ένα ιδανικό σύστημα ανίχνευσης απάτων θα πρέπει να είναι σε θέση να ελαχιστοποιεί τις λανθασμένες ενδείξεις σε συναλλαγές που πραγματοποιούνται από μία πιστωτική κάρτα οι οποίες είναι πραγματικά ασφαλή και όχι προϊόν απάτης. Όπως έχει προαναφερθεί από το κομμάτι της εισαγωγής κίολας, οι απάτες σε πιστωτικές κάρτες προκαλούν τεράστιες ζημιές στο οικονομικό κομμάτι των επιχειρήσεων ανα τον κόσμο και κυρίως στο απλό καταναλωτή καθώς είναι περισσότερο ευάλωτος σε τέτοιου είδους απάτες. Καταλαβαίνουμε λοιπόν την ανάγκη για ύπαρξη συστημάτων, τα οποία θα είναι δομημένα με τέτοιο τρόπο ώστε να σχηματίζουν καταναλωτικά προφίλ για τον κάθε πελάτη και κατ' επέκταση θα καταχωρούν αυτά μαζί με τις πληροφορίες που τα ίδια διαθέτουν σε μία βάση δεδομένων, μέσω της οποίας θα αντλούν αυτά τα δεδομένα έτσι ώστε κάθε φορά που ο πελάτης αυτός θα πραγματοποιεί μία συναλλαγή το σύστημα θα ελέγχει το συναλλακτικό του προφίλ και ανάλογα αν ταιριάζει σε αυτό που έχει ήδη αποθηκεύσει καθώς και σε ορισμένες συνιστώσες, όπως για παράδειγμα ο χρόνος και το ποσό που σκοπεύει να επενδύσει, θα στέλνει μήνυμα στην εκάστοτε υπεύθυνη εταιρεία ότι πρόκειται για τον πραγματικό κάτοχο της κάρτα ή για κάποιον παραβάτη. Βέβαια όπως μπορεί να γίνει αντιληπτό, πάντα θα υπάρχουν και περιπτώσεις που ο ίδιος ο κάτοχος της κάρτας μπορεί να αλλάξει το καταναλωτικό του αυτό προφίλ, για παράδειγμα αν αγοράσει με την κάρτα του κάποιο άλλο μέλος της οικογένειας του ή αλλάξει αυτός τις καταναλωτικές του συνήθειες, αλλά τό μόνο που μπορεί να προκύψει από μία τέτοια περίπτωση είναι η ακύρωση της συναλλαγής αυτής και η επικοινωνία της υπεύθυνης για αυτόν τράπεζα μαζί του για εξακρίβωση. Αυτές οι

ατέλειες όμως είναι και αυτές που δεν μπορούν να προβλεφθούν σε μεγάλο βαθμό, πράγμα που δεν τις καθιστά ανησυχητικές. Μπορούμε επίσης να τις δούμε ως κάτι θετικό, διότι το εκάστοτε σύστημα έκανε την δουλειά του όπως έπρεπε και εμπόδησε μια άγνωστη για αυτό συμπεριφορά του λογαριασμού. Σε επόμενο στάδιο θα μπορούσε να προστεθεί και αυτή η συμπεριφορά στην βάση δεδομένων ώστε να φτάσουμε σε σημείο να μην υπάρχουν τέτοιου είδους λάθη. Γεγονός που όπως είπα είναι αναγκαίο για να καταπολεμηθεί το πρόβλημα με τις απάτες.

Κεφάλαιο 2ο

State of the Art

Το άρθρο [1] μιλάει για την χρήση του αλγόριθμου Random Forest με σκοπό την ανίχνευση απάτης στις πιστωτικές κάρτες. Το σετ δεδομένων που χρησιμοποιείται είναι ένα πραγματικό B2C(Business to consumer) σετ δεδομένων. Αρχικά μας ενημερώνει ότι υπάρχουν δύο μέθοδοι που χρησιμοποιούνται για τον εντοπισμό απατών στις κάρτες: 1)misuse detection και 2)anomaly detection. Επικεντρώνεται στην πρώτη μέθοδο που είναι το misuse detection. Το misuse detection χρησιμοποιεί μεθόδους classification με σκοπό να κατανοήσει αν μια συναλλαγή είναι frauded ή όχι. Εδώ χρησιμοποιείται ο Random Forest ώστε να εκπαιδεύσει τόσο φυσιολογικές όσο και καταστάσεις με απάτη. Στην συνέχεια μας μιλάει παραπάνω για τον Random Forest και μάς λέει ότι είναι ένας συνδυασμός πολλαπλών προγνωστικών δέντρων, έτσι ώστε κάθε δέντρο να εξαρτάται από ένα τυχαίο ανεξάρτητο σύνολο δεδομένων και όλα τα δέντρα στο δάσος να έχουν την ίδια κατανομή. Η χωρητικότητα του Random Forest μας λέει ότι εξαρτάται όχι μόνο από τη δύναμη του μεμονωμένου δένδρου αλλά και από τη συσχέτιση μεταξύ διαφορετικών δέντρων. Όσο ισχυρότερη είναι η δύναμη ενός δέντρου και όσο λιγότερη είναι η συσχέτιση διαφορετικών στρώσεων, τόσο καλύτερη είναι η απόδοση τυχαίων δασών. Προσθέτει ότι, η παραλλαγή των δένδρων προέρχεται από την τυχαία σειρά τους, η οποία περιλαμβάνει δείγματα με bootstrapped και εκλέγει τυχαία ένα υποσύνολο χαρακτηριστικών δεδομένων. Παρόλο λέει, που ενδέχεται να υπάρχουν κάποιες λανθασμένες ενδείξεις στο σύνολο δεδομένων, ο Random Forest εξακολουθεί να είναι ανθεκτικός στο θόρυβο και στις ακραίες τιμές. Μας παραθέτει κάποια αποτελέσματα αναφορικά με το πρώτο από τα δύο πειράματα που πραγματοποιήθηκαν, το οποίο είναι πείραμα πάνω σε δύο μορφές του Random Forest έτσι ώστε να επιλέξει τον κατάλληλο για το τελικό πείραμα. Τα αποτελέσματα ήταν:

Models\Measure	Accuracy	Precision	Recall	F-Measure
Random Forest I	91.96%	90.27%	67.89%	0.7811
Random Forest II	96.77%	89.46%	95.27%	0.9601

TABLE 1. Results of two kinds Random Forests

Ratio of legal and fraud transactions	Legal Transactions	Fraud Transactions	Under-Sampling ratio of legal transactions
1:1	38000	38454	0.81%
1:2	76000	38454	1.61%
1:3	114000	38454	2.42%
1:4	152000	38454	3.23%
1:5	190000	38454	4.03%
1:6	228000	38454	4.84%
1:7	266000	38454	5.64%
1:8	304000	38454	6.45%
1:9	342000	38454	7.26%
1:10	380000	38454	8.06%

TABLE 2. Overview of the original dataset

Στο παραπάνω σχήμα βλέπουμε τα αποτελέσματα από κάθε εκδοχή του Random Forest που χρησιμοποιήθηκε (Πάνω Πίνακας). Στον δεύτερο πίνακα βλέπουμε τα 10 subsets που χωρίστηκε το αρχικό data set καθώς και το ratio legal και fraud συναλλαγών(Κάτω Πίνακας)

Στην συνέχεια μας λέει ότι επιλέγει για να χρησιμοποιήσει για το τελικό πείραμα τον Random Forest 2, διότι μπορεί να έχει μικρότερο precision αλλά όλες οι άλλες μετρήσεις είναι υψηλότερες. Τέλος μας παραθέτει τα αποτελέσματα του δεύτερου και τελικού πειράματος τα οποία παρουσιάζονται στον παρακάτω πίνακα:

Measure\Models	Random Forest Model
----------------	---------------------

Accuracy of Transaction	98.67%
Precision of Transaction	32.68%
Recall of Transaction	59.62%
Intervention rate of Transaction	1.48%
Coverage rate of customer	34.09%

TABLE 3. Result of Fraud Detection model

Το άρθρο[2] μας μιλάει γενικά για την απόδοση των αλγορίθμων μηχανικής μάθησης όταν αυτοί χρησιμοποιούνται για ανίχνευση απατών σε πιστωτικές κάρτες. Αρχικά , μας αναφέρει πως ένα ιδανικό Fraud Detection System(FDS) θα είναι σε θέση να ελαχιστοποιεί τις λανθασμένες ενδείξεις σε συναλλαγές οι οποίες είναι πραγματικά ασφαλή και όχι frauded .Ξεκινάει παραθέτοντας ποιους supervised αλγόριθμους ελέγχει οι οποίοι είναι :

- 1) Random Forest
- 2)Neural Networks (NN) and Artificial Neural Networks (ANN)
- 3)Deep Learning (DL)
- 4) Support Vector Machine (SVM)
- 5) Naive Bayes (NB)
- 6) Logistic Regression (LR)
- 7)Extended Gradient Boosted Tree
- 8)Quadratic Discriminant Analysis (QDA)
- 9)The K-Nearest Neighbour (KNN)
- 10) Hybrid Supervised Approaches

Συνεχίζει προσθέτοντας τους unsupervised αλγορίθμους και είναι οι :

- 1) Self-Organizing Maps (SOM)
- 2)K-means
- 3) Isolation Forest
- 4) The local outlier factor

Μας ενημερώνει ότι το σετ δεδομένων που χρησιμοποιείται έχει 284.807 συναλλαγές, από τις οποίες οι 492 είναι frauded, συνεπώς μιλάμε για ένα αρκετά ανισόρροπο σετ δεδομένων. Για την αξιολόγηση χρησιμοποιήθηκαν οι τιμές: 1)Positive Predictive Value(PPV),2) Negative Predictive Value(NPV), 3) Specificity, 4) Prevalence , 5)Balance Accuracy και 6) Diagnostic Odd Ratio(DOR). Τα αποτελέσματα για κάθε αλγόριθμο μπορούμε να τα δούμε στους παρακάτω πίνακες.

Techniques Used	Precision	Negative Predictive Value	Prevalence	Specificity	True Positive Rate	Diagnostic Odd Ratio
NB	0.06	0.99	0.001	0.97	0.82	6.319
RF	0.99	0.16	0.998	0.91	0.99	18.857
K-NN	NaN	0.99	0.001	1.0	0.0	0
LR	0.99	0.63	0.99	0.87	0.99	168.56
XGBT	0.99	0.92	0.99	0.81	0.99	1138.5
SVM	NaN	NaN	NaN	0.92	0.93	0
ANN	0.99	0.84	0.99	0.77	0.99	462
DL	0.98	0.93	0.86	0.98	0.91	651
QDA	0.97	0.89	0.97	0.42	0.88	261.60
NN	0.99	NaN	0.99	0.0	1.0	0

TABLE 4. Performance Evaluation of Machine Learning Algorithms

Στον πίνακα 4, βλέπουμε τα αποτελέσματα για τους supervised αλγόριθμους

Techniques Used	Precision	Negative Predictive Value	Prevalence	Specificity	True Positive Rate	Diagnostic Odd Ratio
Simple Model	0.99	0.82	0.84	0.99	1.0	726
Model Outliers	0.99	0.82	0.82	1.0	1.0	726
Ensemble Model	0.99	0.82	0.84	1.0	1.0	726

TABLE 5. Performance Evaluation of Hybrid Machine Learning Algorithms

Techniques Used	Precision	Negative Predictive Value	Prevalence	Specificity	True Positive Rate	Diagnostic Odd Ratio
SOM Hybrid	0.92	0.84	0.99	0.83	0.92	60.375
Isolation Forest	0.99	0.99	0.99	1.0	1.0	9801
Local Outlier Factor	0.99	0.99	0.998	1.0	1.0	9801
K-Means	0.99	NaN	0.998	1.0	0.0	0

TABLE 6. Performance Evaluation of Unsupervised Machine Learning Algorithms

Στον πίνακα 5, βλέπουμε τα αποτελέσματα των υβριδικών αλγορίθμων, ενώ στον πίνακα 6 βλέπουμε τα αποτελέσματα των unsupervised αλγορίθμων.

Τέλος μας αναφέρει ότι όλοι οι αλγόριθμοι έχουν τα θετικά και τα αρνητικά τους μέρη, αλλά επειδή το σετ δεδομένων που χρησιμοποιήθηκε ήταν ανισόρροπο και κυρίως για αυτού του είδους σετ δεδομένων, οι unsupervised αλγόριθμοι φαίνεται να λειτουργούν καλύτερα.

Το άρθρο [3] εξετάζει και σχολιάζει διάφορες τεχνικές αιχμής που χρησιμοποιούνται για τον ακριβή εντοπισμό της απάτης με πιστωτικές κάρτες.

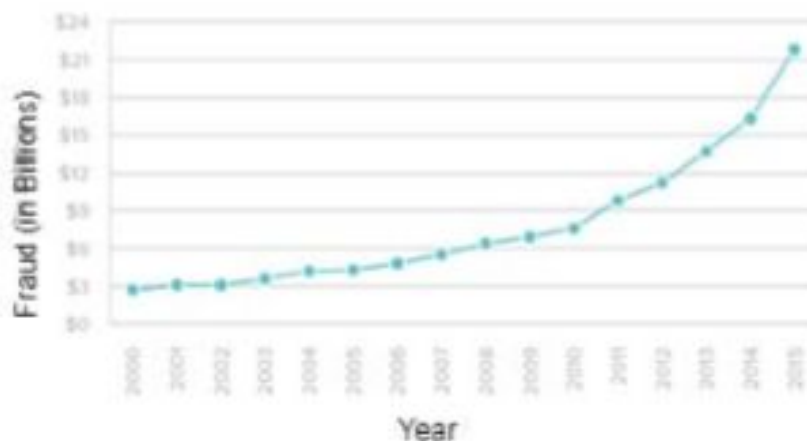


Figure 1. Annual Global Fraud Losses

Το παραπάνω γράφημα δείχνει τα ποσά που χάνονται εξαιτίας των απατών στις πιστωτικές κάρτες.

Αναφέρει τα κύρια προβλήματα που αντιμετωπίζει κανείς όταν πραγματοποιεί μία έρευνα σε αυτό το κομμάτι και είναι :

α)Οι περισσότερες συναλλαγές με πιστωτικές κάρτες είναι non-fraudulent και γίνονται από πραγματικούς πελάτες της εν λόγω τράπεζας . Όταν λέμε οι περισσότερες εννοούμε ότι το 99.9% όλων των συναλλαγών είναι πραγματικές συναλλαγές. Αυτό δημιουργεί ένα σετ δεδομένων το οποίο δεν μπορεί να χρησιμοποιηθεί άμεσα ώστε να σχηματιστεί ένα μοντέλο για μια τέτοια πρόβλεψη.

β)Το δεύτερο πρόβλημα που αναφέρεται στο άρθρο είναι η μη διαθεσιμότητα δεδομένων σχετικά με τις απάτες πιστωτικών καρτών . Αυτό οφείλεται στον εμπιστευτικό χαρακτήρα των συναλλαγών που πραγματοποιούνται μέσω πιστωτικών καρτών.

Έπειτα αναφέρεται στις παραμέτρους που θα χρησιμοποιηθούν για τις μετρήσεις και είναι οι : Sensitivity, Specificity, PPV(Positive Predictive Value), NPV, Precision, Recall, F1, Prevalence, Balanced Accuracy. Προσθέτει ότι χρησιμοποιείτε ένα υβριδικό σχήμα over-sampling και cost-sensitive αλγορίθμων. Συνεχίζει παραθέτοντας όλους τους αλγορίθμους που χρησιμοποιούνται καθώς και τους περιορισμούς και τα πλεονεκτήματα του καθένα

System	Advantaged	Limitations
Genetic Algorithm	1.False alerts are minimized 2.Learns from itself 3.Improves after every iteration	1.Defines the misclassification of transactions thus may lead to problems 2.Hard to build and train the model
Sparse autoencoders	Generalizes data very well	-
Hidden Markov Model	Involves some kind of encryption which provides some security to the details of the customer	Making observational symbols is time consuming and need to upload after every transaction
ANN Using Merchant Side Data	1. Good success rate 2.Beneficial to merchant	Transaction data are provided by the merchant hence limits the fraud detection capability
ANN with Whale Optimization Algorithm	Converges to optimal solution very efficiently	-
Cost Sensitive Modeling	1.False negative alerts are	High algorithmic

using Neural Network	less 2.Operational costs are less	complexity
----------------------	--------------------------------------	------------

TABLE 7. Advantages and Limitations of the Algorithms

Ο παραπάνω πίνακας 7 μας παρουσιάζει τα πλεονεκτήματα και του περιορισμούς των αλγορίθμων. Ο πίνακας 8 που ακολουθεί ,μας παρουσιάζει την σύγκριση μεταξύ των αλγορίθμων

Method	Accuracy	Resources Used	Dataset Used	Parameters
Genetic Algorithm	Avg	High	Synthetically generated data	Time card was used,location,current balance,average spending,habit of customer
Sparse autoencoders	Avg	Avg	German Credit dataset	Input transactions
Hidden Markov Model	Low	Low	Synthetically generated data	Spending behaviour, transactions amount
ANN Using Merchant Side Data	Avg	Avg	Merchant side dataset	Shipping address,transaction date,time,previous data
ANN with Whale Optimization Algorithm	High	Avg	Kaggle dataset	-
Cost Sensitive Modeling using Neural Network	High	Avg	Synthetically generated data	-

TABLE 8. Comparison Between the Algorithms

Παρατηρείται ότι αντί για αριθμητικές τιμές έχουμε High,Average και Low. Σύμφωνα με το άρθρο αυτό γίνεται για να αποφευχθεί η παροχή παραπλανητικών πληροφοριών. Τέλος αναφέρει τα θετικά και τα αρνητικά σημεία κάθε μεθόδου

καθώς και ότι δεν χρησιμοποιούνται απόλυτες τιμές για την αναπαράσταση της ακρίβειας των συστημάτων, καθώς δεν υπάρχουν τυποποιημένα δείγματα αναφοράς για σκοπούς δοκιμών

Το άρθρο [4] έχει ως σκοπό να προσδιορίσει τις αδυναμίες των υπαρχόντων προσεγγίσεων για την ανισορροπία των κλάσεων. Πιο συγκεκριμένα, στις μεθόδους μηχανικής μάθησης. Η μελέτη, μας λέει το άρθρο ότι, πραγματοποιείται με τις πλέον σύγχρονες τεχνολογίες οι οποίες χρησιμοποιούνται σε εφαρμογές με σκοπό να δοθεί λύση στο πρόβλημα του imbalance classification. Στόχος του άρθρου είναι να εντοπιστούν τα προβλήματα που πρέπει να λυθούν προκειμένου να έχουμε μια εξαιρετικά αποδοτική λύση για το πρόβλημα της ανισορροπίας κλάσεων Στην συνέχεια αναφέρεται στους αλγόριθμους μηχανικής μάθησης που χρησιμοποιούνται για την ανίχνευση απάτης σε πιστωτικές κάρτες, τους οποίους αλγόριθμους μετά το άρθρο αξιολογεί και συγκρίνει. Οι αλγόριθμοι ταξινόμησης που πραγματεύονται στο άρθρο είναι ονομαστικά οι :

1. C5.0 ALGORITHM
2. SUPPORT VECTOR MACHINE (SVM)
3. Artificial Neural Network (ANN)
4. NAÏVE BAYES (NB)
5. BAYESIAN BELIEF NETWORK (BBN)
6. LOGISTIC REGRESSION (LR)
7. ARTIFICIAL IMMUNE SYSTEMS (AIS)

Έπειτα μας παρουσιάζονται οι τεχνικές που χρησιμοποιούνται για την προσέγγιση ενός προβλήματος imbalanced classification. Η πρώτη τεχνική εφαρμόζεται στα δεδομένα ως ένα από τα βήματα της προ επεξεργασίας ενώ η δεύτερη τεχνική χρησιμοποιείται μέσα σε κάποιον αλγόριθμο όπως Cost-Sensitive (CS) προσεγγίσεις ή One-Class Classification (OCC). Οι τεχνικές αυτές είναι :

1. RANDOM OVERSAMPLING (RO)
2. ONE-CLASS CLASSIFICATION (OCC)

Στην συνέχεια το άρθρο μιλάει για τον τρόπο με τον οποίο είναι στημένη η πειραματική μελέτη και είναι ως εξής :Στην πρώτη φάση συγκρίνονται 8 μέθοδοι ταξινόμησης. Η σύγκριση έγινε βασισμένη σε τρεις παραμέτρους: Accuracy , sensitivity και Area Under Precision-Recall Curve (AUPRC). Η σύγκριση αυτή καταλήγει στην επιλογή των καταλληλότερων αλγορίθμων. Σε αυτή περιέχονται και οι αλγόριθμοι :C5.0,SVM και ANN. Στη δεύτερη φάση , οι επιλεγμένοι αλγόριθμοι χρησιμοποιούνται για την σύγκριση επιλεγμένων imbalance classification προσεγγίσεων όπως Random Oversampling,One-Class Classificationand και Cost Sensitive.

Μας αναφέρει επίσης ότι το σετ δεδομένων περιέχει 6.178.231 δηλαδή το 61.7% συναλλαγές από άνδρες και 3.821.769 συναλλαγές από γυναίκες δηλαδή το 38.3%.Επίσης 596.014 (5.96%) είναι συναλλαγές από απάτες και 9.403.986 (94.04%) νόμιμες συναλλαγες. Το 70% του σετ δεδομένων χρησιμοποιήθηκε ως σετ εκπαίδευσης και το 30% ως σετ ελέγχου για την μελέτη της απόδοσης.Μας λέει ότι

χρησιμοποιήθηκε το 2% του αρχικού σετ δεδομένων αλλά με τον ίδιο λόγο ανισορροπίας και αριθμό συναλλαγών.

Method	Accuracy	Sensitivity	AUPRC	Performance
C5.0	96%	43%	0.6	Highest Accuracy
SVM	96%	39%	0.63	Highest Accuracy
ANN	96%	47%	0.62	Highest Accuracy
LR	96%	49%	0.66	Highest AUPRC
NB	93%	56%	0.5	Highest Sensitivity
BBN	94%	15%	0.64	High AUPRC
KNN	95%	45%	0.29	Lowest AUPRC
NSA	92%	51%	0.29	Lowest AUPRC

TABLE 9. Performance of Different Methods

Ο παραπάνω πίνακας 9 μας δείχνει την απόδοση του κάθε αλγορίθμου

Τέλος το άρθρο αναφέρει ότι, σύμφωνα με τα αποτελέσματα, βρέθηκε ότι για όλους τους αλγορίθμους ότι η ακρίβεια τους είναι μεγαλύτερη απο 90% με διαφορετική ευαισθησία και AUPRC. Αλλά, βασιζόμενοι στην μελέτη και με τα παραπάνω αποτελέσματα , καταλήγουμε στο γεγονός ότι από όλους τους αλγορίθμους οι C5.0, SVM και ANN είναι οι καλύτεροι όταν χρειάζεται να αξιολογήσουμε την απόδοση των προσεγγίσεων για ανισόρροπη ταξινόμηση.

Το άρθρο [5] μας λέει ότι εξάγει συναλλακτικές συμπεριφορές των χρηστών χρησιμοποιώντας τόσο το ιστορικό συναλλαγών του ίδιου αλλά και δεδομένα κάποιων χρηστών με παρόμοια συμπεριφορά. Επιπλέον, προτείνει έναν μηχανισμό ανατροφοδότησης ο οποίο μπορεί να προσαρμοστεί στη συναλλακτική συμπεριφορά ενός κατόχου πιστωτικής κάρτας. Ξεκινάει με την διαδικασία προ επεξεργασίας των δεδομένων. Πρώτο βήμα της προ επεξεργασίας είναι όπως μας λέει το Cardholder Clustering κατά το οποίο χρησιμοποιεί την μέθοδο k-means ώστε να χωρίσει όλους τους κατόχους σε τρεις παρόμοιες ομάδες οι οποίες με την σειρά τους χαρακτηρίζονται ως high (h), medium (m) και low (l), ανάλογα με το ποσό της

συναλλαγής. Το επόμενο βήμα είναι το Sliding Window. Το Sliding window είναι ένας τύπος αλγορίθμου ο οποίος μπορεί να συγκεντρώσει τις συναλλαγές και στη συνέχεια να αντλήσει ορισμένα νέα χαρακτηριστικά που σχετίζονται με το ποσό / χρόνο από τα συγκεντρωτικά δεδομένα, προκειμένου να χαρακτηρίσει πιο ακριβή τα πρότυπα συμπεριφοράς ενός κατόχου κάρτας. Τελευταίο βήμα στην διαδικασία της προ επεξεργασίας είναι το Feature Extraction κατά το οποίο εξάγονται νέα χαρακτηριστικά. Στο κάτω σχήμα βλέπουμε τον αλγόριθμο που χρησιμοποιήθηκε για την εξαγωγή των χαρακτηριστικών

Algorithm 1: Sliding-window-based algorithm for aggregating transactions and deriving features

Input: The unique id of a cardholder, a transaction sequence t^{id} and the size of sliding window p ;
Output: The feature sets of the cardholder G^{id} and F^{id}

```

1  $n :=$  the length of  $t^{id}$  ;
2  $G^{id} := \emptyset$  ;
3  $F^{id} := \emptyset$  ;
4 for ( $i:=0$ ;  $i < n - p + 1$ ;  $i++$ ) do
5    $T_i^{id} := \emptyset$  ;
6   //WINDOW Function
7   for ( $j := i$ ;  $j \leq i + p - 1$ ;  $j++$ ) do
8      $T_i^{id} = T_i^{id} \cup t_j^{id}$  ;
9   end
10   $x_{i1}^{id} = MAX\_AMOUNT(T_i^{id})$  ;
11   $x_{i2}^{id} = MIN\_AMOUNT(T_i^{id})$  ;
12   $x_{i3}^{id} = AVG\_AMOUNT(T_i^{id})$  ;
13   $x_{i4}^{id} = AMOUNT(T_i^{id})$  ;
14   $f_i := 4$  ; //  $f_i$  is a counter
15  for ( $j := i + 1$ ;  $j \leq i + p - 1$ ;  $j++$ ) do
16     $f_i := f_i + 1$  ;
17     $x_{if_i}^{id} = TIME(t_j^{id}) - TIME(t_{j-1}^{id})$  ;
18    //  $TIME(t_j^{id})$  denotes the transaction time of  $t_j^{id}$ 
19  end
20   $X_i^{id} = (x_{i1}^{id}, \dots, x_{if_i}^{id})$  ;
21   $y_i^{id} = LABEL(T_i^{id})$  ;
22  if  $y_i^{id} = 0$  then
23     $G^{id} = G^{id} \cup X_i^{id}$  ;
24  else
25     $F^{id} = F^{id} \cup X_i^{id}$  ;
26  end
27 end

```

Figure 2. Sliding-window-based algorithm for aggregating transactions and deriving features

Μετά την εξαγωγή νέων χαρακτηριστικών για το κάθε “παράθυρο”, το X^{id} μπορεί να θεωρηθεί ως το μοτίβο της συμπεριφοράς ενός κατόχου πιστωτικής κάρτας και το G^{id} είναι το σύνολο όλων των μοτίβων συμπεριφοράς του κατόχου ο οποίο έχει id . Μας λέει επίσης ότι, στον πραγματικό κόσμο είναι πολύ δύσκολη δουλειά να ταξινομηθούν αυτά τα φυσικά σετ χαρακτηριστικών σε συγκεκριμένα μοτίβα συμπεριφοράς. Ο λόγος είναι ότι ο ορισμός ενός μοτίβου συμπεριφοράς μπορεί να είναι μέχρι και ασαφής όταν αυτά συνάπτονται χρησιμοποιώντας γνώσεις που προέρχονται από τον ανθρώπινο τομέα. Βέβαια είναι πιο χρήσιμο να χρησιμοποιηθεί μία μέθοδος ομαδοποίησης έτσι ώστε να λυθεί αυτό το unsupervised πρόβλημα το οποίο μπορεί να οργανώνει αυτόματα υψηλού επιπέδου αφηρημένη γνώση.

Αναφέρει επίσης ότι , η True Label μία συναλλαγής σε δοκιμαστικό σετ δεδομένων δεν χρησιμοποιήθηκε ποτέ για να ενημερώσει το προφίλ ενός κατόχου πιστωτική κάρτας. Να σημειωθεί ότι η πληροφορία από ένα True Label είναι πολύ σημαντική καθώς αντικατοπτρίζει τις αλλαγές της συναλλακτικής συμπεριφοράς ενός κατόχου έμμεσα. Ακριβώς γι' αυτό , η προτεινόμενη μέθοδος του άρθρου χρησιμοποιεί έναν μηχανισμό ανατροφοδότησης έτσι ώστε να ενημερώνει το προφίλ του κάθε κατόχου όταν μια νέα συναλλαγή έρχεται. Όταν η μέθοδος αυτή παράγει μια λανθασμένη πρόβλεψη , θεωρεί ότι οι νέες συναλλαγές δεν συμμορφώνονται με το ανανεωμένο προφίλ του κατόχου. Η True Label μιας συναλλαγής θα χρησιμοποιηθεί προκειμένου να μετατραπεί η μέτρηση της βαθμολογίας του classifier και η μέθοδος του άρθρου προσπαθεί να βρει τον καταλληλότερο classifier για το νέο προφίλ του κατόχου.

Έπειτα μας αναφέρει ένα από τα προβλήματα που συναντάται συχνά σε τέτοιου είδους έρευνες και είναι τα δεδομένα. Ποιό συγκεκριμένα αναφέρει ότι είναι πολύ δύσκολο να πραγματοποιηθεί η προτεινόμενη μέθοδος χρησιμοποιώντας πραγματικά δεδομένα , καθώς οι τράπεζες δεν μοιράζονται με ερευνητές τέτοιου είδους δεδομένα. Για αυτό τον λόγο το άρθρο χρησιμοποιεί έναν simulator για να παράγει δεδομένα συναλλαγών. Επίσης για τους σκοπούς της έρευνας υποθέτει πως αυτός που πραγματοποιεί την απάτη έχει οικειοποιηθεί με τα ποσά των συναλλαγών του κατόχου καθώς μπορεί να αποκτήσει πληροφορίες σχετικά με το ιστορικό της πιστωτικής κάρτας. Με αυτά κατά νου, το ποσό της fraudulent συναλλαγής είναι παρόμοιο με τις φυσιολογικές συναλλαγές, αλλά η μόνη διαφορά ανάμεσά τους είναι συχνότητα των συναλλαγών σε μια χρονική στιγμή, διότι ένας απατεώνας θέλει να καταναλώνει όσο πιο γρήγορα γίνεται .Ωστόσο , η παραπάνω simulator κατασκευάζει μόνο συναλλακτικά ποσά. Χρειαζόμαστε και άλλα χαρακτηριστικά όπως εβδομαδιαία και χρονικά μοτίβα .Χρησιμοποιείται στο άρθρο μια διαδικασία Poisson ώστε να πάρει τέτοιες καταστάσεις. Υποθέτει πως ο κάθε κάτοχος έχει 50 συναλλαγές τον μήνα και ο κάθε μήνας έχει 4 εβδομάδες. Έτσι ο αριθμός των συναλλαγών αλλάζει κάθε εβδομάδα.

Ποιό συγκεκριμένα σχετικά με το πειραματικό κομμάτι αναφέρει ότι: το p κυμαίνεται από 25 έως 100 με βήμα 4. Οι τιμές κατωφλίου ορίζονται από 10%, 30%, 50%, 70% και 90%. Το k κυμαίνεται από 1 έως 11 με βήμα 6. Έτσι, υπάρχουν $4 \times 6 = 24$ πιθανοί συνδυασμοί. Οι παρακάτω πίνακες μας δείχνουν την μεταβολή του accuracy και του recall ανάλογα με το p και το k .

Threshold (%)	Accuracy average over all the 6 kinds of k for different sliding window sizes				Recall average over all the 6 kinds of k for different sliding window sizes			
	25	50	75	100	25	50	75	100
10	0.6852	0.5889	0.6183	0.5757	0.6131	0.7418	0.7390	0.7556
30	0.7319	0.6954	0.7021	0.6571	0.4546	0.6055	0.5839	0.6123
50	0.7859	0.8181	0.7844	0.7891	0.3297	0.4174	0.3746	0.3874
70	0.8506	0.8782	0.8539	0.8732	0.1422	0.1531	0.1395	0.1093
90	0.8951	0.8997	0.8969	0.900	0.0441	0.04444	0.0346	0.0262

TABLE 10. Variation of Accuracy and Recall for Different Sliding Window Sizes

Thresh old(%)	Accuracy average over all the 4 sliding window sizes for different k						Recall average over all the 4 sliding window sizes for different k					
	1	3	5	7	9	11	1	3	5	7	9	11
10	0.8973	0.8542	0.7094	0.5260	0.4012	0.3033	0.1229	0.6931	0.8114	0.8601	0.8923	0.9002
30	0.9032	0.8666	0.7700	0.6540	0.5249	0.4483	0.0651	0.4670	0.6291	0.6933	0.7661	0.7835
50	0.9032	0.8815	0.8420	0.7779	0.7185	0.6524	0.0396	0.2748	0.4000	0.4647	0.5342	0.5682
70	0.9018	0.8962	0.8782	0.8568	0.8455	0.8163	0.0164	0.1014	0.1395	0.1556	0.1541	0.2340
90	0.9006	0.9021	0.8999	0.8975	0.8973	0.8921	0.0035	0.0330	0.0403	0.0468	0.0432	0.0535

TABLE 11. Variation of Accuracy and Recall for Different k

Τέλος μας παρουσιάζει τις μεθόδους. Για διευκόλυνση, οι δύο μέθοδοι και η μέθοδος του άρθρου παρουσιάζονται ως: 1) Logistic Regression με raw data (RawLR). 2) Random Forest με aggregated data (AggRF). 3) Random Forest και feedback technique με aggregated data (AggRF+FB). Η τρίτη είναι και η μέθοδος του άρθρου. Καταλήγει ότι τα αποτελέσματα δείχνουν την απόδοση και την αποτελεσματικότητα της μεθόδου που προτείνεται καθώς, σε σύγκριση με άλλες δύο μεθόδους, όλες

μπορούν να επιτύχουν ακρίβεια 80% κατά τον εντοπισμό των συναλλαγών. Ωστόσο, το AggRF αποδίδει καλά μόνο στο FFDR(Fraud Features Detection Rate) και το RawLR αποδίδει καλά μόνο στο CDDR(Concept Drift Detection Rate). Η προτεινόμενη μέθοδος μας μπορεί να βελτιώσει την απόδοση τόσο σε FFDR όσο και σε CDDR και να αυξήσει τη μέση ανάκληση και ακρίβεια.

Ο σκοπός του άρθρου [6] είναι να κάνει μια συγκριτική ανάλυση τεχνικών ανίχνευσης απατών σε πιστωτικές κάρτες χρησιμοποιώντας : naive bayes, k-nearest neighbor και logistic regression τεχνικές σε highly skewed data βασισμένα στις μετρικές: accuracy, sensitivity, specificity and Matthews's correlation coefficient (MCC). Το ανισόρροπο σετ δεδομένων που χρησιμοποιείται σε αυτή την μελέτη η οποία περιέχει περίπου 0.172% απάτες σε συναλλαγές είναι δείγμα υβριδικής προσέγγισης . Η θετική κλάση (Απάτη) είναι over-sampled και η αρνητική κλάση (νόμιμη) είναι under-sampled από τον ίδιο αριθμό ώστε να επιτευχθούν δύο εκδόσεις των 34 προς 66 και των 10 προς 90 .Οι 3 παραπάνω τεχνικές εφαρμόζονται στα δεδομένα. Η σύγκριση της απόδοσής τους πραγματοποιείται βασισμένη στην accuracy, sensitivity, specificity, Matthews Correlation Coefficient (MCC) και balanced classification rate.Στην συνέχεια μας παρουσιάζει τους 3 classifiers που χρησιμοποιήθηκαν, κατ ονόματι : Naive Bayes , k-Nearest Neighbors & Logistic regression techniques .Να σημειωθεί ότι τα πειράματα αξιολογούνται με την χρήση μετρήσεων True Positive, True Negative, False Positive, False Negative. Στην συνέχεια μας μιλάει για το σετ δεδομένων που χρησιμοποιείται, το οποίο περιέχει συναλλαγές πιστωτικών καρτών που έγιναν από Ευρωπαίους ιδιοκτήτες καρτών τον Σεπτέμβριο του 2013 . Μας παρουσιάζει τις συναλλαγές που πραγματοποιήθηκαν μέσα σε δύο μέρες,με συνολικό αριθμό 284,807.Η θετική κλάση αποτελεί το 0.172% του συνόλου των συναλλαγών. Να σημειωθεί 'το το παρών σετ δεδομένων έχει μόνο αριθμητικές μεταβλητές λόγο PCA πράγμα που οδηγεί σε 28 κύρια συστατικά. Επομένως, χρησιμοποιούνται συνολικά 30 input features.Μας ενημερώνει ότι προκειμένου να αξιολογήσουμε αυτά τα μοντέλα, 70% του dataset χρησιμοποιείται για training και το 30% για επικύρωση και δοκιμές.

Metrics	Classifiers		
	Naive Bayes	K-Nearest Neighbour	Logistic Regression
Accuracy	0.9737	0.9691	0.9824
Sensitivity	0.8072	0.8835	0.9767
Specificity	0.9471	0.9711	0.9824
Precision	0.0505	0.4104	0.0873
MCC	+0.1979	+0.5903	+0.2893
Balanced Classification Rate	0.8907	0.9273	0.9796

TABLE 12. Accuracy Result for Un-sampled Data Distribution

Ο παραπάνω πίνακας 12 δείχνει την ακρίβεια για την διανομή των δεδομένων χωρίς δείγμα

Metrics	Classifiers		
	Naive Bayes	K-Nearest Neighbour	Logistic Regression
Accuracy	0.9752	0.9715	0.3639
Sensitivity	0.8210	0.8285	0.7155
Specificity	0.9754	1.0000	0.2939
Precision	0.0546	1.0000	0.1678
MCC	+0.2080	+0.8950	+0.0077
Balanced Classification Rate	0.8975	0.9143	0.5047

TABLE 13. Accuracy Results for 10:90 Distribution

Ο παραπάνω πίνακας 13 δείχνει την ακρίβεια για την διανομή των 10:90 δεδομένων

Metrics	Classifiers		
	Naive Bayes	K-Nearest Neighbour	Logistic Regression
Accuracy	0.9769	0.9792	0.5486
Sensitivity	0.9514	0.9375	0.5833
Specificity	0.9896	1.0000	0.5313
Precision	0.9786	1.0000	0.3836
MCC	+0.9478	+0.9535	+0.1080
Balanced Classification Rate	0.9705	0.9688	0.5573

TABLE 14. Accuracy Results for 34:66 Data Distribution

Ο παραπάνω πίνακας 14 δείχνει την ακρίβεια για την διανομή 34:66 δεδομένων

Τέλος το άρθρο καταλήγει στο γεγονός ότι τα αποτελέσματα από το πείραμα δείχνουν ότι το kNN εμφανίζει σημαντική απόδοση για όλες τις μετρήσεις που αξιολογήθηκαν, εκτός από την ακρίβεια στην κατανομή δεδομένων 10:90. Αυτή η μελέτη δείχνει την επίδραση της υβριδικής δειγματοληψίας στην απόδοση της δυαδικής ταξινόμησης των ανισορροπων δεδομένων.

Οι δύο τύποι απατών που μπορούν να αναγνωριστούν είναι οι Card-not-present (CNP) frauds και Card-present (CP) frauds. Το παρόν άρθρο [7] ασχολείται με την CNP κατηγορία απατών και προτείνει μια μέθοδο για ανίχνευση αυτών σε πραγματικό χρόνο. Τα δεδομένα τα οποία χρησιμοποιούνται σε αυτή την μελέτη αναλύονται σε δύο μέρη: ως categorical και numerical δεδομένα. Τα πρωτογενή δεδομένα μπορούν να παρασκευαστούν με τον καθαρισμό δεδομένων και άλλες βασικές τεχνικές προεπεξεργασίας. Πρώτα, τα κατηγορηματικά δεδομένα μπορούν να μετατραπούν σε αριθμητικά δεδομένα και στη συνέχεια να εφαρμοστούν κατάλληλες τεχνικές για την αξιολόγηση τους. Δεύτερον, τα κατηγορηματικά δεδομένα χρησιμοποιούνται στις τεχνικές μηχανικής εκμάθησης για την εύρεση του βέλτιστου αλγορίθμου. Στην συνέχεια το άρθρο μιλάει για το σετ δεδομένων και μας λέει ότι εξαιτίας της προστασίας προσωπικών δεδομένων μερικές ευαίσθητες ιδιότητες έχουν υποστεί κρυπτογράφηση με συναρτήσεις hash. Επίσης μας λέει ότι κατά την αξιολόγηση τους σετ δεδομένων, το σχήμα των δεδομένων ήταν πολύ ασαφές λόγω του ανισόρροπου αριθμού νόμιμων συναλλαγών και δόλιων περιστατικών. Το αρχείο με τις περιπτώσεις απάτης είχε 200 εγγραφές, ενώ το αρχείο καταγραφής συναλλαγών είχε 917781 εγγραφές.

Field Name	Description
CARD_NO	Credit cardholders hashed card number
DATE	Date of the transaction
TIME	Time of the transaction
TRANSACTION AMOUNT	Transaction amount
MERCHANT_NAME	Merchant name relevant to the transaction
MERCHANT_CITY	Registered city of the merchant
MERCHANT_COUNTRY	Registered country of the merchant
RESPONSE_CODE	ISO response code related to the transaction
MERCHANT_CATEGORY_CODE	Category code of the merchant
APPROVED/ NOT APPROVED	Status of the transaction

TABLE 15. Attributes of Genuine Transactions Log

Στον πίνακα 15 παρουσιάζονται τα χαρακτηριστικά των πραγματικών αρχείων καταγραφής συναλλαγών

Field Name	Description
CARD_NO	Credit cardholders hashed card number
DATE	Date of the transaction
TIME	Time of the transaction
SEQ	A unique sequence number which was given for frauds
NATURE OF THE FRAUD	Nature of the frauds as card present or not present
MCC	Category code of the merchant
AMOUNT OF THE FRAUD	Transactions amount
REVERSAL	Bank related field

TABLE 16. Attributes of Fraudulent Transactions Log

Σε αυτόν τον πίνακα 16 παρουσιάζονται τα χαρακτηριστικά των μη νόμιμων αρχείων καταγραφής συναλλαγών

Έπειτα συνεχίζει μιλώντας για το κομμάτι της επεξεργασίας των δεδομένων καθώς και για τον τρόπο με τον οποίο θα γίνει η αξιολόγηση. Τα ακατέργαστα δεδομένα αρχικά διαχωρίστηκαν σε 4 σετ δεδομένων ανάλογα με την ακολουθία της απάτης. Αυτή η διαδικασία έγινε με τα δεδομένα που αποκτήθηκαν από την τράπεζα. Τα 4 αυτά σετ δεδομένων είναι:

- 1) Transactions with Risky Merchant Category Code (MCC).
- 2) Transactions larger than \$100.
- 3) Transactions with risky ISO Response code.
- 4) Transactions with unknown web addresses.

Αυτά τα 4 σύνολα δεδομένων χρησιμοποιήθηκαν με δύο διαφορετικούς τρόπους. Μετατρέποντας τα ανεπεξέργαστα δεδομένα σε αριθμητική μορφή (Τύπος Α). Με την προετοιμασία ακατέργαστων δεδομένων κατηγορηματικά χωρίς να γίνει μετασχηματισμός (Τύπος Β). Οι δύο πηγές δεδομένων χαρακτηρίζονται η υψηλή δυσαναλογία στην κατανομή των παραδειγμάτων στις κλάσεις. Για να ξεπεραστεί αυτό το πρόβλημα, πραγματοποιήθηκε undersampling και oversampling μειώνοντας τα περιστατικά πλειοψηφίας και αυξάνοντας τα περιστατικά μειοψηφίας αντίστοιχα. Για oversampling, χρησιμοποιήθηκαν τεχνικές δειγματοληψίας συνθετικών μειονοτήτων (SMOTE) και για undersampling, χρησιμοποιήθηκαν συμπυκνωμένοι πλησιέστεροι γείτονες (CNN) και τυχαία undersampling (RUS). Χρησιμοποιήθηκαν 4

αλγόριθμοι μηχανικής μάθησης στην ανάλυση. Αυτοί είναι: Support Vector Machine, Naive Bayes, K-Nearest Neighbor and Logistic Regression. Κατά την επιλογή μοντέλων μηχανικής μάθησης που μπορούν να αιχμαλωτίσουν κάθε απάτη ή ακρίβεια και η απόδοση του κάθε μοντέλου λήφθηκαν υπόψη. Κάνει αναφορά στα συστήματα ανίχνευσης απατών σε πραγματικό χρόνο και μας λέει ότι είναι ουσιαστικά η εκτέλεση μοντέλων ανίχνευσης απατών την στιγμή που μία online συναλλαγή λαμβάνει χώρα. Κατ' αυτό τον τρόπο το σύστημα είναι ικανό να ανιχνεύσει τις απάτες. Ειδοποιεί την τράπεζα σχετικά με την ακολουθία της απάτης και την ακρίβεια της, έτσι ώστε να είναι εύκολο στις ομάδες που ασχολούνται με αυτό το κομμάτι να πάνε στο επόμενο βήμα χωρίς καθυστέρηση. Κλείνει λέγοντας ότι υπάρχει σημαντικός αντίκτυπος από τη χρήση τεχνικών δειγματοληψίας για τη λήψη συγκριτικά υψηλότερης απόδοσης από τον ταξινομητή. Τα μοντέλα μηχανικής εκμάθησης που κατέλαβαν τα τέσσερα μοτίβο απάτης με τα υψηλότερα ποσοστά ακρίβειας είναι LR, NB, LR και SVM (Τα ονόματα αναφέρθηκαν πιο πάνω). Επιπλέον, τα μοντέλα έδειξαν ποσοστά ακρίβειας 74%, 83%, 72% και 91% αντίστοιχα.

Το άρθρο [8] ασχολείται με την ενσωμάτωση unsupervised τεχνικών με supervised ταξινομητές απάτης πιστωτικών καρτών. Συγκεκριμένα, παρουσιάζονται ορισμένα κριτήρια για τον υπολογισμό βαθμολογία outlier σε διαφορετικά επίπεδα διακριτότητας και γίνεται εκτίμηση της προστιθέμενης αξίας τους ως προς την ακρίβεια όταν ενσωματωθούν ως χαρακτηριστικά σε μια εποπτευόμενη στρατηγική μάθησης. Το άρθρο συγκεντρώνει τους κατόχους καρτών σε δύο ομάδες βάσει των μέσων ποσών δαπανών τους. Το πρώτο σύμπλεγμα θα περιέχει κατόχους καρτών με υψηλές μέσες δαπάνες (CH1 και CH2) και το δεύτερο θα περιέχει κατόχους καρτών με χαμηλές μέσες δαπάνες (CH3, CH4 και CH5). Τα μέσα ποσά συναλλαγών είναι 99,8 στην ομάδα υψηλών δαπανών και 39,1 στην ομάδα χαμηλών δαπανών. Σε αυτήν την περίπτωση, ο κάτοχος της κάρτας CH2 ειδοποιείται και πάλι, καθώς η συναλλαγή του στο ποσό των 185 αποκλίνει περισσότερο από το μέσο ποσό του συμπλέγματος (99,8).

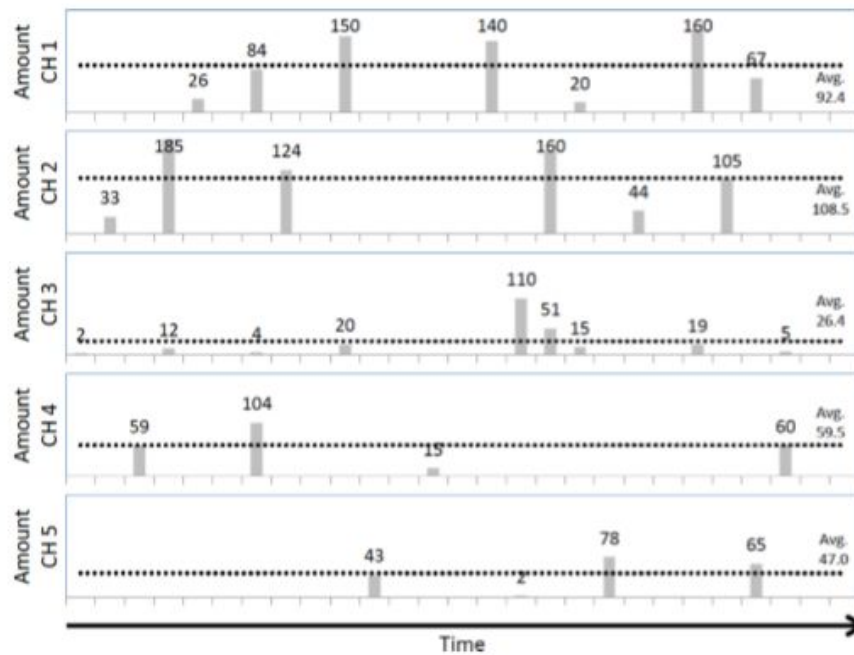


Figure 3. This illustrative examples presents the behaviours of five card holders over a fixed period of time. The dotted line represent the amount of money spent in each transaction

Το παραπάνω σχεδιάγραμμα μας παρουσιάζει ένα παράδειγμα καταναλωτικών συμπεριφορών για τους κατόχους καρτών που αναφέρθηκαν παραπάνω

Στο κομμάτι της έρευνας το άρθρο επιλέγει να χρησιμοποιήσει τον αλγόριθμο k-means καθώς είναι εύκολο να ερμηνευτεί, τρέχει γρήγορα σε μεγάλα σύνολα δεδομένων και μας προσφέρει τη δυνατότητα να αποφασίσουμε εκ των προτέρων έναν αυθαίρετο αριθμό συστάδων που θα αναγνωριστούν. Το ζήτημα της επιλογής των λειτουργιών στις οποίες θα πραγματοποιηθεί η ομαδοποίηση δεν είναι ασήμαντο. Αυτή η απόφαση μπορεί να έχει σημαντικό αντίκτυπο στην τελική ακρίβεια. Εξετάστηκαν δύο διαφορετικά σύνολα χαρακτηριστικών: το πρώτο περιγράφει τη συμπεριφορά του κατόχου της κάρτας και το δεύτερο συνοψίζει τα προσωπικά δεδομένα του κατόχου της κάρτας. Στην πρώτη περίπτωση, εξετάζονται οι μέσες δαπάνες συναλλαγών και ο συνολικός αριθμός συναλλαγών των κατόχων πιστωτικών καρτών τις τελευταίες 24 ώρες. Στη δεύτερη περίπτωση εξετάζεται η ηλικία, η εθνικότητα και το φύλο του κατόχου της κάρτας. Δεδομένου ότι η συμπεριφορά των κατόχων καρτών οδηγεί σε καλύτερη ακρίβεια, παρουσιάζονται στο άρθρο μόνο ομάδες που έχουν δημιουργηθεί σύμφωνα με τις λειτουργίες συμπεριφοράς των κατόχων καρτών στο πειραματικό μέρος. Η κύρια παράμετρος υπερ του αλγορίθμου k-means είναι ο αριθμός k των συστάδων που θα δημιουργηθούν. Αυτή η υπερ-παράμετρος είναι επίσης σημαντική για τη μελέτη περίπτωσης του άρθρου, καθώς καθορίζει το επίπεδο λεπτομέρειας του βαθμού outlier. Στα πειράματά

επιτρέπεται στην υπερ-παράμετρο να κυμαίνεται από ελάχιστο χαμηλό βαθμό λεπτομέρειας 10 συστάδων έως μέγιστο υψηλό βαθμό λεπτομέρειας 5000 συστάδων. Συνεχίζει παραθέτοντας πληροφορίες σχετικά με το σετ δεδομένων που χρησιμοποιήθηκε. Το σύνολο δεδομένων που χρησιμοποιήθηκε για τα πειράματα περιλαμβάνει πληροφορίες από συναλλαγές 334 ημερών που καταγράφηκαν από τις 2 Φεβρουαρίου έως τις 31 Δεκεμβρίου 2016. Αυτό το σύνολο δόθηκε από την εταιρία Worldline και περιλαμβάνει 76 εκατομμύρια συναλλαγές. Το ποσοστό μη νόμιμων συναλλαγών σε αυτό το σύνολο δεδομένων είναι 0,36%. Οι συναλλαγές έως τις 30 Σεπτεμβρίου χρησιμοποιούνται για τα εκπαιδευτικά δεδομένα, ενώ αυτές που πραγματοποιήθηκαν μεταξύ 8 Οκτωβρίου και 31 Δεκεμβρίου χρησιμοποιούνται ως δεδομένα δοκιμαστικού συνόλου. Η εβδομάδα από 1 Οκτωβρίου έως 7 Οκτωβρίου 2016 αντιπροσωπεύει την περίοδο καθυστέρησης επαλήθευσης και τα δεδομένα από αυτήν την περίοδο δεν χρησιμοποιούνται για σκοπούς εκπαίδευσης ή δοκιμών. Το άρθρο εξετάζει δύο μετρήσεις: το Topn Precision και το AUC-PR και αυτές χρησιμοποιούνται τόσο για παγκόσμιες όσο και για τοπικές προσεγγίσεις. Τελειώνει αναφέροντας πως τα αποτελέσματα δεν είναι πειστικά όσον αφορά τις παγκόσμιες και τοπικές προσεγγίσεις. Η ερμηνεία που δίνει το άρθρο είναι ότι και οι δύο προσεγγίσεις δεν έχουν το σωστό επίπεδο λεπτομέρειας απαραίτητο για να επωφεληθούν από τις μη εποπτευόμενες πληροφορίες. Ένα πιο ελπιδοφόρο αποτέλεσμα επιτυγχάνεται μέσω της προσέγγισης συμπλέγματος (ιδίως όσον αφορά το AUC-PR), αν και φαίνεται ότι η αύξηση των συνόλων δεδομένων με πάρα πολλές βαθμολογίες θα μπορούσε να είναι επιζήμια λόγω προβλημάτων υπερβολικής και διακύμανσης. Τα ληφθέντα αποτελέσματα ανοίγουν το δρόμο για πολλές πιθανές ερευνητικές κατευθύνσεις

Το άρθρο [9] προτείνει ένα νέο μέτρο εξοικονόμησης που βασίζεται στη σύγκριση του οικονομικού κόστους ενός αλγορίθμου σε σχέση με τη χρήση κανενός απολύτως μοντέλου. Έπειτα προτείνεται μια διευρυμένη έκδοση της στρατηγικής συγκέντρωσης συναλλαγών, συμπεριλαμβάνοντας έναν συνδυασμό κριτηρίων κατά την ομαδοποίηση συναλλαγών. Αυτό επιτρέπει την ύπαρξη πλουσιότερου χώρου. Επιπλέον, προτείνεται μια νέα μέθοδος ώστε να εξαχθούν περιοδικά χαρακτηριστικά με σκοπό να υπολογιστεί αν ο χρόνος μιας νέας συναλλαγής είναι μέσα στα διαστήματα εμπιστοσύνης προηγούμενων χρόνων συναλλαγής. Το κίνητρο είναι ότι ένας πελάτης αναμένεται να κάνει συναλλαγές σε παρόμοιες ώρες. Η προτεινόμενη μεθοδολογία βασίζεται στην ανάλυση της περιοδικής συμπεριφοράς ενός χρόνου συναλλαγής, χρησιμοποιώντας τη διανομή von Mises. Αναφέρεται στο γεγονός ότι, η πληροφορία μιας μόνο συναλλαγής δεν είναι αρκετή για να εντοπιστεί μια fraudulent συναλλαγή, καθώς χρησιμοποιώντας μόνο τα βασικά χαρακτηριστικά, παραλείπονται σημαντικές πληροφορίες όπως για παράδειγμα το καταναλωτικό προφίλ του χρήστη. Για να αντιμετωπιστεί το παραπάνω, το άρθρο προτείνει ένα νέο σετ

χαρακτηριστικών έτσι ώστε οι πληροφορίες της τελευταίας συναλλαγής που πραγματοποιήθηκε με την ίδια πιστωτική κάρτα να χρησιμοποιούνται επίσης για την πρόβλεψη. Ο στόχος είναι, να είναι σε θέση να ανιχνευθούν υπερβολικά ανόμοιες συνεχείς συναλλαγές μέσα από τις αγορές του καταναλωτή. Το νέο σετ χαρακτηριστικών που προτείνεται περιέχει : χρόνος από την τελευταία συναλλαγή, το προηγούμενο ποσό της συναλλαγής, την προηγούμενη χώρα της συναλλαγής. Παρόλα αυτά , αυτά τα χαρακτηριστικά δεν λαμβάνουν υπόψη την συμπεριφορά των καταναλωτών εκτός από την τελευταία συναλλαγή του, πράγμα που οδηγεί σε ανολοκλήρωτο προφίλ των καταναλωτών. Ένας καλύτερος τρόπος για να ληφθεί υπόψη το καταναλωτικό προφίλ ενός πελάτη είναι να αντληθούν ορισμένα χαρακτηριστικά χρησιμοποιώντας μια στρατηγική συγκέντρωσης συναλλαγών. Η παραγωγή των χαρακτηριστικών συγκέντρωσης απαρτίζεται από την συγκέντρωση των συναλλαγών που πραγματοποιήθηκαν κατά την διάρκεια των τελευταίων ωρών , αρχικά από την κάρτα ή τον αριθμο λογαριασμού , έπειτα από το είδος της συναλλαγής, ομάδα εμπορών, χώρα ή άλλη, ακολουθούμενη από τον υπολογισμό του αριθμού των συναλλαγών ή του συνολικού ποσού που δαπανήθηκε για αυτές τις συναλλαγές. Για το συγκεκριμένο άρθρο χρησιμοποιείται ένα σετ δεδομένων το οποίο παρέχεται από μία μεγάλη Ευρωπαϊκή εταιρία επεξεργασίας πιστωτικών καρτών. Οι συναλλαγές έγιναν ανάμεσα στην περίοδο Ιανουαρίου του 2012 και Ιουνίου 2013. Συνολικά περιέχει 120.000.000 συναλλαγές, κάθε μία από τις οποίες έχει 27 ιδιότητες καθώς και μία ετικέτα η οποία μας ενημερώνει ότι η τρέχουσα συναλλαγή είναι *frauded*. Σε αυτό το σετ δεδομένων έχουμε συνολικά 40.000 τέτοιες συναλλαγές (*frauded*) κάτι που οδηγεί στο να έχουμε 0.025% δείκτη απάτης. Επίσης μέσα από τις μεθοδολογίες που χρησιμοποιήθηκαν για εξαγωγή χαρακτηριστικών προκύπτουν 293 χαρακτηριστικά. Επιπλέον , για τα πειράματα, επιλέχθηκε ένα μικρότερο υποσύνολο των συναλλαγών με μεγαλύτερο δείκτη απάτης το οποίο απαντά σε συγκεκριμένη ομάδα συναλλαγών. Αυτό το υποσύνολο περιέχει 236.735 συναλλαγές με δείκτη απάτης 1.5%. Σε αυτό το σετ δεδομένων τα συνολικά χρήματα που κλάπηκαν είναι 895.154 ευρώ. Αυτό το υποσύνολο επιλέχθηκε καθώς είναι και αυτό που παρατηρούνται τα περισσότερα κρούσματα απάτης. Από το σύνολο του σετ δεδομένων, εξάγονται 3 διαφορετικά σύνολα δεδομένων για: εκπαίδευση, επικύρωση και δοκιμή. Κάθε μία περιέχει 50%, 25% και 25% των συναλλαγών αντίστοιχα.

Για τα πειράματα χρησιμοποιήθηκαν τρεις αλγόριθμοι ταξινόμησης χωρίς ευαισθησία κόστους και είναι οι εξής : 1) decision tree (DT), 2) logistic regression (LR), 3) random forest (RF), 4) Bayes minimum risk (BMR) model

Τελειώνει λέγοντας πως μέσα από την προεπεξεργασία των δεδομένων προκειμένου να συμπεριληφθεί η πρόσφατη συμπεριφορά των καταναλωτών, η απόδοση αυξάνεται κατά περισσότερο από 200% σε σύγκριση με τη χρήση μόνο των πρώτων πληροφοριών συναλλαγής. Το νέο προτεινόμενο σύνολο χαρακτηριστικών αυξάνει την απόδοση κατά 252% και 287%, αντίστοιχα. Ωστόσο σημειώνει ότι, επειδή αυτή η

μελέτη πραγματοποιήθηκε χρησιμοποιώντας ένα σύνολο δεδομένων από ένα χρηματοπιστωτικό ίδρυμα, δεν ήταν εφικτό να συζητηθούν βαθιά τα ειδικά χαρακτηριστικά που δημιουργήθηκαν αλλά ούτε και ο ατομικός αντίκτυπος της κάθε δυνατότητας

Στο άρθρο [10] δημιουργείται ένα μοντέλο νευρωνικού δικτύου το οποίο κάνει classification των συναλλαγών με πιστωτική κάρτα βασισμένο σε πραγματικά ιστορικά δεδομένα. Στην συνέχεια μας μιλάει για την σύνδεση της ανίχνευσης των απατών στις πιστωτικές κάρτες με τα νευρωνικά δίκτυα αναφέροντας πως η ανίχνευση απατών σε πιστωτικές κάρτες με χρήση νευρωνικών δικτύων είναι ικανή λόγω της δυνατότητας των νευρωνικών δικτύων να λύνουν classification problems. Ο ανθρώπινος εγκέφαλος μαθαίνει μέσα από την εμπειρία. Τα νευρωνικά δίκτυα κατα την ανίχνευση απατών στις πιστωτικές κάρτες δουλεύουν με την ίδια λογική. Όταν ένας πελάτης χρησιμοποιεί την πιστωτική του κάρτα, δημιουργεί μία καθορισμένη ακολουθία. Η ακολουθία αυτή έπειτα χρησιμοποιείται στο classification. Όταν μία πιστωτική κάρτα χρησιμοποιείται, το σύστημα συγκρίνει την πληροφορία που δίνει η εκάστοτε συναλλαγή με τις πληροφορίες που έχει αποθηκευμένες από προηγούμενες συναλλαγές του ίδιου του χρήστη. Αν τα δεδομένα αντιστοιχούν στην ακολουθία, τότε είμαστε σχεδόν σίγουροι ότι η πιστωτική κάρτα χρησιμοποιείται από τον πραγματικό της κάτοχο. Αν δεν υπάρχει καμία αντιστοιχία τότε η πιθανότητα να είναι προϊόν απάτης αυξάνεται δραματικά. Προχωράει κάνοντας αναφορά στον τρόπο διαχείρισης ανισορροπων σετ δεδομένων, γιατί πρέπει να υπάρξει προ επεξεργασία σε αυτά αλλά και τους δείκτες που πρέπει να χρησιμοποιηθούν για την αξιολόγηση. Πιο συγκεκριμένα λέει ότι: «α ν έχουμε ένα ανισόρροπο σετ δεδομένο, που περιέχει 1% από την κλάση μειονότητας και 99% από την κλάση πλειοψηφίας, τότε ένας αλγόριθμος μπορεί να προβλέψει όλες τις υποθέσεις σαν να ανήκουν στην κλάση πλειοψηφίας. Επομένως η ακρίβεια του αλγορίθμου θα είναι 99% πράγμα που φαίνεται εντυπωσιακό, είναι όμως; Επειδή η ακρίβεια δεν είναι ο καλύτερος τρόπος μέτρησης της απόδοσης ενός ανισόρροπου σετ δεδομένων, χρησιμοποιούνται άλλοι δείκτες. Αυτοί που χρησιμοποιούνται σε αυτό το άρθρο είναι: False Positive (FP), True Positive (TP), False Negative (FN), True Negative (TN).

Έπειτα μας παρουσιάζει το σετ δεδομένων που χρησιμοποιήθηκε, καθώς και τα χαρακτηριστικά του. Περιέχει συναλλαγές με πιστωτικές κάρτες τον Σεπτέμβριο του 2013 από ευρωπαίους κατόχους καρτών. Αυτό το σύνολο δεδομένων παρουσιάζει συναλλαγές που πραγματοποιήθηκαν σε δύο ημέρες, όπου έχουμε 492 απάτες από 284 807 συναλλαγές. Το σύνολο δεδομένων είναι εξαιρετικά ανισόρροπο, η θετική τάξη (απάτες) αντιστοιχεί στο 0,172% όλων των συναλλαγών. Επιπλέον μας λέει ότι θα χρησιμοποιηθούν τεχνικές υπο-δειγματοληψίας διότι παίρνοντας ένα υποσύνολο 900 συναλλαγών από το σύνολο δεδομένων καθώς και αφαιρώντας τυχαίες εγγραφές από την κατηγορία πλειοψηφίας προκύπτει τελικά ένα υποσύνολο που περιέχει 450 κανονικές και 450 συναλλαγές απάτης.

Προχωράει μιλώντας για τον τρόπο δόμησης του μοντέλου που θα χρησιμοποιηθεί. Η απόφαση σχετικά με τον αριθμό των νευρώνων εξόδου είναι κάπως πιο απλή, καθώς υπάρχουν επιτακτικοί λόγοι για να χρησιμοποιούνται πάντα έναν ή δύο νευρώνες εξόδου. Χρησιμοποιούμε δύο νευρώνες εξόδου. Όταν εκπαιδεύουμε πολυεπίπεδα δίκτυα, η συνήθης τακτική είναι αρχικά να διαχωριστεί το σετ δεδομένων σε

υπο-σύνολα. Τα διανύσματα εισόδου και τα διανύσματα στόχου χωρίζονται τυχαία σε τρία σύνολα. Το πρώτο υπο-σύνολο του 70% των δεδομένων και χρησιμοποιείται ως training set για τον υπολογισμό της διαβάθμισης και την ενημέρωση των “βαρών” και των “προκαταλήψεων” του δικτύου. Το δεύτερο υπο-σύνολο του 15% των δεδομένων χρησιμοποιείται ως validation set χρησιμοποιείται για να επιβεβαιώσει ότι το δίκτυο γενικεύεται και για να σταματήσει την προπόνηση πριν από την υπερβολική τοποθέτηση. Το τελευταίο υπο-σύνολο του υπόλοιπου 15% των δεδομένων χρησιμοποιείται ως εντελώς ανεξάρτητος έλεγχος της γενίκευσης του δικτύου.

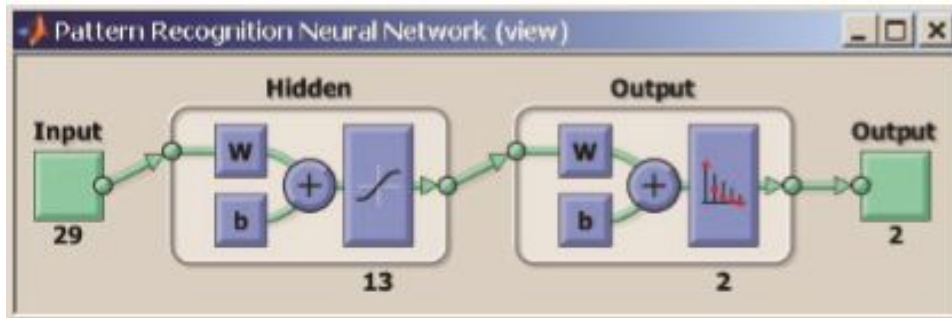


Figure 4. Architecture of trained pattern recognition neural network

Η παραπάνω φωτογραφία μας δείχνει την αρχιτεκτονική του νευρωνικού αυτού δικτύου



Figure 5. Confusion Matrices for trained network

Ο πάνω πίνακας μας δείχνει όλους του πίνακες σύγχυσης για το εκπαιδευμένο δίκτυο

Τέλος μας λέει ότι τα δίκτυα αναγνώρισης προτύπων με 29 νευρώνες εισόδου, ένα κρυφό στρώμα με 13 νευρώνες και ένα επίπεδο εξόδου με 2 νευρώνες κατασκευάζονται με την εργαλειοθήκη νευρωνικού δικτύου στο Matlab. Η ακρίβεια του δικτύου είναι 94,2%. Η βαθμολογία ακρίβειας είναι 91,6%, η βαθμολογία ανάκλησης είναι 97,3%, η βαθμολογία F1 είναι 94,40% και η βαθμολογία Matthew Coefficient είναι 88,62%. Η βαθμολογία F1 και η βαθμολογία MCC δείχνουν ότι το μοντέλο λειτουργεί πολύ καλά.

Ο σκοπός του άρθρου [11] είναι να ανιχνεύσει την ακρίβεια των δόλιων συναλλαγών χρησιμοποιώντας τον αλγόριθμο Random Forest. Στο προτεινόμενο σύστημα, το συγκεκριμένο άρθρο, χρησιμοποιεί τον αλγόριθμο Random Forest και τα Νευρωνικά Δίκτυα για ταξινόμηση και παλινδρόμηση του σετ δεδομένων. Τα δεδομένα που χρησιμοποιούνται στο άρθρο βασίζονται σε πραγματικές συναλλαγές από μία μεγάλη Ευρωπαϊκή εταιρεία. Προσωπικές πληροφορίες των πελατών παραμένουν εμπιστευτικές. Η ακρίβεια του αλγορίθμου είναι περίπου 50%. Συνεπώς η ακρίβεια των αποτελεσμάτων που αποκτήθηκαν από τις μεθόδους συγκρίνονται στο ελάχιστο με το προτεινόμενο σύστημα. Αρχικά, συλλέγονται τα δεδομένα από το σετ δεδομένων και πραγματοποιείται μία ανάλυση πάνω σε αυτά. Έπειτα πραγματοποιείται ένας “καθαρισμός” του του σετ δεδομένων ώστε να αφαιρεθούν από αυτό τιμές οι οποίες επαναλαμβάνονται αλλά και μηδενικές τιμές. Στην συνέχεια χωρίζει το σετ δεδομένων σε δύο κατηγορίες ως σετ δεδομένων εκπαίδευσης και ελέγχου, με σκοπό να πραγματοποιηθεί σύγκριση και ανάλυση του σετ δεδομένων. Μετά τον διαχωρισμό του σετ δεδομένων εφαρμόζεται ο RFA ο οποίος θα δώσει καλύτερη ακρίβεια για τις συναλλαγές με πιστωτική κάρτα. Εφαρμόζοντας τον RFA το σετ δεδομένων θα κατηγοριοποιηθεί σε 4 κατηγορίες που θα αποκτηθούν στην μορφή ενός πίνακα σύγχυσης. Στον πίνακα σύγχυσης το σύνολο δεδομένων θα χωριστεί σε τέσσερα μπλοκ: True Positive (TP), True Negative (TN), False Positive (FP) και False Negative (FN). Έτσι, για να υπάρξει καλύτερη ακρίβεια, χρησιμοποιείται ο αλγόριθμος Random Forest όπου παίρνει όλες τις τιμές του γραφήματος και επιστρέφει μόνο τις απαραίτητες τιμές με την καλύτερη ακρίβεια σε σύγκριση με όλους τους άλλους αλγόριθμους. Βασιζόμενοι στην παραπάνω κατηγοριοποίηση της απόδοσης των δεδομένων θα πραγματοποιηθεί η ανάλυση.

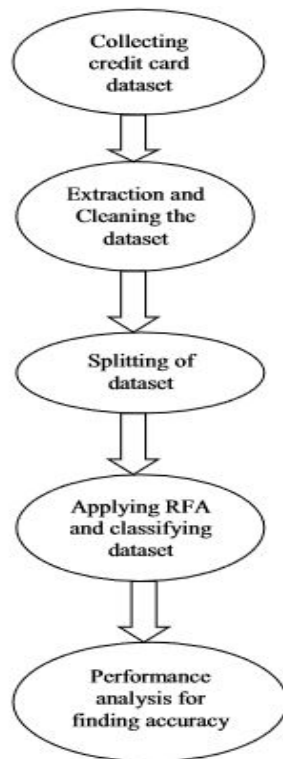


Figure 6. Dataset and RFA implementation process

Το παραπάνω σχήμα μας δείχνει ένα ένα τα στάδια που μόλις αναφέρθηκαν

Αναφέρει ότι αυτός ο RFA που χρησιμοποιείται βασίζεται σε επιβλεπόμενη μάθηση και το μεγαλύτερό του πλεονέκτημα είναι ότι μπορεί να χρησιμοποιηθεί τόσο για ταξινόμηση και παλινδρόμηση. Δίνει καλύτερη ακρίβεια συγκριτικά με όλα τα υπόλοιπα υπάρχοντα συστήματα και είναι ο πιο συχνά χρησιμοποιούμενος αλγόριθμος. Σε αυτό το άρθρο η χρήση του RFA για την ανίχνευση απατών σε πιστωτικές κάρτες μπορεί να δώσει 90-95% ακρίβεια.

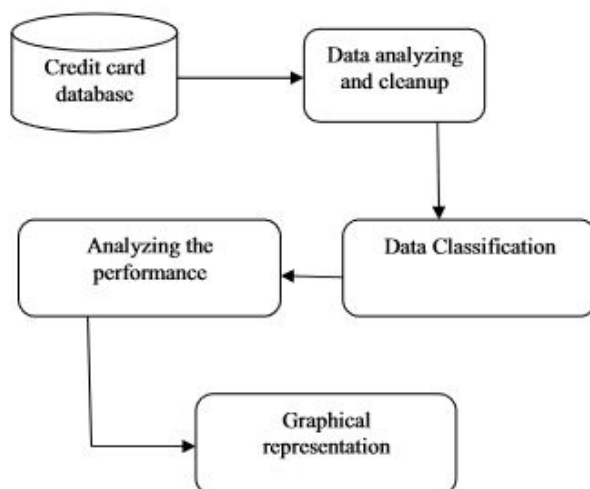


Figure 7. System Architecture

Στο παραπάνω σχήμα βλέπουμε την αρχιτεκτονική του συστήματος που πραγματεύεται το άρθρο

Το άρθρο [12] ,ουσιαστικά αποτελεί μία μελέτη σχετικά με την ανίχνευση απάτης με πιστωτικές κάρτες με χρήση αλγορίθμων μηχανικής μάθησης έχει παρουσιαστεί σε αυτό το έγγραφο.Οι αλγόριθμοι κυμαίνονται από τυπικά νευρωνικά δίκτυα έως μοντέλα βαθιάς μάθησης. Αξιολογούνται χρησιμοποιώντας τόσο σετ δεδομένων πιστωτικών καρτών συγκριτικής αξιολόγησης όσο και σετ δεδομένων που αντιστοιχεί στον πραγματικό κόσμο. Επιπλέον, οι μέθοδοι AdaBoost και majority voting εφαρμόζονται για τη διαμόρφωση υβριδικών μοντέλων. Για να εκτιμηθεί περαιτέρω η στιβαρότητα και η αξιοπιστία των μοντέλων, προστίθεται θόρυβος στο πραγματικό σύνολο δεδομένων.Στην συνέχεια μιλάει για το πως χρησιμοποιεί τις δύο αυτές μεθόδους. Το Majority Voting χρησιμοποιείται αρκετά συχνά για ταξινόμηση δεδομένων, εμπεριέχει έναν συνδυασμό με τουλάχιστον δύο αλγόριθμους. Κάθε αλγόριθμος πραγματοποιεί την δικιά του πρόβλεψη για κάθε δείγμα. Το τελικό αποτέλεσμα είναι για τον ένα που λαμβάνει την πλειοψηφία των ψήφων. Το Adaptive Boosting ή το AdaBoost χρησιμοποιείται σε συνδυασμό με διαφορετικούς τύπους αλγορίθμων για τη βελτίωση της απόδοσής τους. Οι έξοδοι συνδυάζονται χρησιμοποιώντας ένα σταθμισμένο άθροισμα(weighted sum), το οποίο αντιπροσωπεύει τη συνδυασμένη έξοδο του ενισχυμένου ταξινομητή.Έπειτα μας παρουσιάζει τους αλγόριθμους που χρησιμοποιήθηκαν καθώς και έναν πίνακα με τα πλεονεκτήματα και τους περιορισμούς της κάθε μεθόδου.Οι αλγόριθμοι είναι οι:1)Naïve Bayes (NB),2) Decision Trees(DT),3)Random Forest(RF),4)Gradient Boosted Tree(GBT),5)Decision Stump(DS),6)Random Tree(RT),7)Multilayer Perceptron(MLP),8)Feed-Forward Neural Network(NN),9)Deep Learning (DL),10)Linear Regression(LIR),11)Logistic Regression(LOR),12)Support Vector Machine(SVM)

Model	Strengths	Limitations
Bayesian	Good for binary classification problems, efficient use of computational resources, suitable for real-time operations	Need good understanding of typical and abnormal behaviours for different types of fraud detection
Trees	Easy to understand and implement, the procedures require a low computational power,suitable for real-time operations	Potential of over-fitting if the training set does not represent the underlying domain information, retraining is required for new types of fraud cases
Neural Network	Suitable for binary classification problems and widely used for fraud detection	Need a high computational power, unsuitable for real-time operations,retraining is required for new types of fraud cases

Linear Regression	Provide optimal results when the relationship between independent and dependent variables are linear	Sensitive to outliers and limited to numeric values only
Logistic regression	Easy to implement and historically used for fraud detection	Poor classification performance as compared with other data mining methods
Support Vector Machine	Able to solve nonlinear classification problems, low computational power, suitable for real-time operations	Not easy to process the results due to transformation of the input data

TABLE 17. Strengths and Limitations of Machine Learning Methods

Το πρώτο σετ δεδομένων που ελέγχονται οι αλγόριθμοι περιέχει συνολικά 284.807 συναλλαγές που πραγματοποιήθηκαν τον Σεπτέμβριο του 2013 από ευρωπαίους κατόχους καρτών. Το σύνολο δεδομένων περιέχει 492 συναλλαγές απάτης, γεγονός που σημαίνει ότι υπάρχει υψηλή ανισορροπία. Λόγω του ζητήματος εμπιστευτικότητας, παρέχονται συνολικά 28 βασικά τα οποία όμως έχουν μετασχηματιστεί . Μόνο τα δεδομένα χρόνου και ποσότητας δεν μετασχηματίζονται και παρέχονται ως έχουν. Τα αποτελέσματα τόσο για το AdaBoost όσο και για το majority voting παρέχονται από τους παρακάτω πίνακες.

Model	Accuracy	Fraud	Non-Fraud	MCC
NB	98.038%	82.520%	98.064%	0.235
DT	99.919%	81.089%	99.951%	0.775
RF	99.889%	42.683%	99.988%	0.604
GBT	99.903%	81.707%	99.935%	0.747
DS	99.906%	66.870%	99.963%	0.711
RT	99.866%	32.520%	99.982%	0.497
DL	99.915%	79.878%	99.950%	0.765
NN	99.933%	81.301%	99.965%	0.807
MLP	99.933%	80.894%	99.966%	0.806
LIR	99.907%	54.472%	99.985%	0.686
LOR	99.926%	79.065%	99.962%	0.786
SVM	99.927%	82.317%	99.957%	0.796

TABLE 18. Results of AdaBoost

Model	Accuracy	Fraud	Non-Fraud	MCC
DS+GBT	99.848%	11.992%	100.000%	0.343
DT+DS	99.850%	14.024%	99.998%	0.361
DT+GBT	99.920%	60.366%	99.988%	0.737
DT+NB	99.932%	72.967%	99.978%	0.788
NB+GBT	99.919%	66.463%	99.976%	0.788
NN+NB	99.941%	78.862%	99.978%	0.823
RF+GBT	99.865%	23.780%	99.996%	0.468

TABLE 19. Results of Majority Voting

Έπειτα το άρθρο χρησιμοποιεί ένα δεύτερο σετ δεδομένων αυτή την φορά παρεχόμενο από από ένα χρηματοπιστωτικό ίδρυμα στη Μαλαισία.Βασίζεται σε κατόχους καρτών από την περιοχή της Νοτιοανατολικής Ασίας από τον Φεβρουάριο έως τον Απρίλιο του 2017. Καταγράφονται συνολικά 287.224 συναλλαγές, με 102 από αυτές να κατατάσσονται ως περιπτώσεις απάτης. Τα δεδομένα αποτελούνται από μια χρονική σειρά συναλλαγών. Για να συμμορφωθεί με τις απαιτήσεις εμπιστευτικότητας των πελατών, δεν χρησιμοποιούνται προσωπικές πληροφορίες ταυτοποίησης.Τα αποτελέσματα παρέχονται ξανά από τους παρακάτω πίνακες.

Model	Accuracy	Fraud	Non-Fraud	MCC
NB	100.000%	100.000%	100.000%	1.000
DT	99.999%	100.000%	99.999%	0.990
RF	100.000%	100.000%	100.000%	1.000
GBT	99.999%	100.000%	99.999%	0.986
DS	99.999%	100.000%	99.999%	0.990
RT	100.000%	100.000%	100.000%	0.995
DL	99.994%	96.078%	99.995%	0.917%
NN	99,998%	100.000%	99,998%	0.967
MLP	99.996%	100.000%	99.996%	0.950

LIR	99.992%	94.118%	99.994%	0.890
LOR	99.999%	100.000%	99.999%	0.981
SVM	99.959%	1.961%	99.994%	0.044

TABLE 20. Results of AdaBoost

Model	Accuracy	Fraud	Non-Fraud	MCC
DS+GBT	100.000%	100.000%	100.000%	0.995
DT+DS	100.000%	100.000%	100.000%	0.995
DT+GBT	100.000%	100.000%	100.000%	1.000
DT+NB	99.999%	98.039%	100.000%	0.990
NB+GBT	99.999%	98.039%	100.000%	0.990
NN+NB	99.998%	95.098%	100.000%	0.975
RF+GBT	99.999%	100.000%	99.999%	0.990

TABLE 21. Results of Majority Voting

Τέλος η μέτρηση MCC έχει υιοθετηθεί ως μέτρο απόδοσης, καθώς λαμβάνει υπόψη τα TP,TN,FP,FN προβλεπόμενα αποτελέσματα.Για το πρώτο σετ δεδομένο η υψηλότερη μέτρηση ήταν 0.823 η οποία αποκτήθηκε με majority voting. Για το δεύτερο σετ δεδομένων η υψηλότερη μέτρηση ήταν το 1 τόσο στο AdaBoost όσο και στο majority voting.Έπειτα προστέθηκε θόρυβος 30% στο σετ δεδομένων και προέκυψε μία μέτρηση MCC της τάξης του 0.942% για το majority voting. Αυτό δείχνει, σύμφωνα με το άρθρο, ότι το majority voting έχει σταθερή απόδοση κατά την παρουσία θορύβου.

Το άρθρο [13] ασχολείται με μια υβριδική προσέγγιση για τον εντοπισμό ομάδων πελατών με βάση τα προφίλ συμπεριφοράς των κατόχων πιστωτικών καρτών και στη συνέχεια τη δημιουργεί συγκεκριμένους classifiers για την κάθε ομάδα ξεχωριστά . Η προσέγγισή αυτή έχει τρία βήματα: 1) δημιουργία προφίλ λογαριασμών 2) ομαδοποίηση λογαριασμών και 3) οικοδόμηση των ταξινομητών. Όσον αφορά το σετ δεδομένων παρέχεται από μια τράπεζα και περιέχει περίπου 80 εκατομμύρια συναλλαγές από 1.1 εκατομμύρια χρήστες .Αυτές οι συναλλαγές καταγράφηκαν μέσα από μία περίοδο 8 μηνών και έχουν 69 χαρακτηριστικά. Η κυρίως πρόκληση με αυτό το σετ δεδομένων είναι η σημαντική ανισορροπία της κλάσης. Μόνο το 0.136% των συνολικών συναλλαγών ανήκει στις απάτες. Επιπλέον , μόνο 26.000 λογαριασμοί έχουν μη νόμιμες συναλλαγές συσχετισμένες με αυτούς .Τα δεδομένα περιέχουν numeric και categorical στοιχεία, εκ των οποίων και τα δύο περιέχουν missing values.

Όλα τα χαρακτηριστικά που περιέχουν missing values για περισσότερο από το 80% των συναλλαγών, αφαιρέθηκαν έτσι ώστε να αποφευχθεί η πιθανότητα τα συνθετικά δεδομένα να επηρεάσουν τα αποτελέσματα. Όλα τα υπόλοιπα δεδομένα που λείπουν υπολογίζονται με μέσα για αριθμητικά χαρακτηριστικά και λειτουργία για κατηγοριοποιημένες δυνατότητες. Καμία συναλλαγή δεν αφαιρέθηκε κατά την διαδικασία. Προσθέτει επίσης ότι πέντε δείγματα δημιουργήθηκαν από το αρχικό σετ δεδομένων για αυτή την μελέτη. Κάθε δείγμα περιέχει 110.000 λογαριασμούς και 5 εκκ συναλλαγές . Όλοι οι λογαριασμοί που είναι frauded συμπεριλαμβάνονται σε κάθε δείγμα και δίνουν κοντά στο 2% fraud rate. Οι υπόλοιποι 84.000 λογαριασμοί επιλέγονται τυχαία από λογαριασμούς που δεν περιέχουν απατηλές συναλλαγές. Αυτή η ερευνητική μελέτη χρησιμοποιεί μια πιο πολύπλοκη προσέγγιση , στην οποία οι λειτουργίες πρώτα κατασκευάζονται χρησιμοποιώντας δεδομένα επιπέδου λογαριασμού αντί για δεδομένα επιπέδου συναλλαγής. Αυτοί οι λογαριασμοί συγκεντρώνονται σε ξεχωριστά group ανάλογα με την ακολουθία που έχουν στην συμπεριφορά τους , γεγονός που βοηθάει ώστε να αναγνωριστούν μοναδικά γκρούπ πελατών τα οποία πιο πριν μπορεί να ήταν άγνωστα. Classifiers μετά δημιουργούνται για κάθε συγκέντρωση λογαριασμών χρησιμοποιώντας το εκάστοτε σετ συναλλαγών ώστε να εκπαιδεύσουν τον classifier. Επίσης αναπτύχθηκε ένα αναλυτικό framework για προφίλ λογαριασμών και ταυτοποίηση των τάσεων συμπεριφοράς σε επίπεδο λογαριασμού το οποίο έχει 4 διαστάσεις – Spend, Spread, Safety and Sketch. Για την ταξινόμηση χρησιμοποιήθηκαν οι μέθοδοι Random Forest και XGBoost. Από την στιγμή που το fraud rate είναι περίπου 2%, κατάλληλα class weights παρέχονται στους αλγορίθμους ώστε να εξισορροπήσουν τον αντίκτυπο της ανισορροπίας. Και για τα δύο μοντέλα , 70 δέντρα παρήχθησαν, καθένα με μέγιστο βάθος 6 διαχωρισμούς και κάθε διαχωρισμός θα μπορούσε να είναι το 50% του συνόλου χαρακτηριστικών. Συνεχίζει λέγοντας ότι αφού υπολογίστηκε το άθροισμα των παραλλαγών σε κάθε αριθμό συστάδων από το πέντε έως το τριάντα, ο κατάλληλος αριθμός συστάδων που χρησιμοποιήθηκαν προσδιορίστηκε να είναι 10. Αυτές οι 10 είναι οι εξής :

1. Global Spenders
2. Diverse Spenders
3. Low Spenders
4. Frequent Spenders
5. Big Spenders
6. Loyal Spenders
7. General
8. Local Spenders
9. Big and Infrequent Spenders
10. Global Spenders

Customer Groups	Average Fraud Rate	Standard Deviation
Global Spenders	6.36%	0.79%
Diverse Spenders	1.21%	0.11%
Low Spenders	2.17%	0.80%
Frequent Spenders	1.64%	0.33%
Big Spenders	4.74%	0.08%
Loyal Spenders	2.20%	0.74%
General	1.54	0.30%
Local Spenders	1.34%	0.05%
Big and Infrequent Spenders	2.95%	0.08%
Glocal Spenders	2.93%	1.10%

TABLE 22. Fraud Rates Across Customer Groups

Ο παραπάνω πίνακας 22 μας δείχνει τα ποσοστά απάτης για κάθε μία από τις παραπάνω ομάδες

Κλείνει λέγοντας πως τα τρέχοντα αποτελέσματα επισημαίνουν τη δυνατότητα των ταξινομητών να ποικίλλουν ανάλογα με την συστάδα , υποδηλώνοντας έτσι ότι αυτοί οι ταξινομητές μπορούν να ενισχύσουν τη συνολική απόδοση εντοπισμού απάτης όταν αξιολογούνται με τη χρήση ομαδοποίησης. Επιπλέον, τα χαρακτηριστικά λογαριασμού και συναλλαγής κάθε συστάδας θα πρέπει να διερευνηθούν περαιτέρω για να κατανοήσουν ποιες λειτουργίες είναι χρήσιμες κατά τη διαίρεση των πελατών. Συγκεκριμένα, οι συστάδες που δεν μπορούν να διαφοροποιηθούν μεταξύ τους πρέπει να διερευνηθούν περαιτέρω για να κατανοήσουν καλύτερα τις συμπεριφορές των πελατών στους οποίους αντιστοιχούν.

Το άρθρο [14] μιλάει για fraud detection systems(FDS) στον πραγματικό κόσμο.Επίσης μας λέει ότι σε ένα FDS που ανταποκρίνεται στον πραγματικό κόσμο, η τεράστια παροχή των αιτήσεων για πληρωμή ελέγχεται γρήγορα από αυτοματοποιημένα εργαλεία τα οποία αποφασίζουν ποιες συναλλαγές θα εξουσιοδοτηθούν.Στην συνέχεια αναφέρεται στα στάδια ελέγχου ενός FDS τα οποία είναι: 1) Terminal,2) Transaction-Blocking Rules,3) Scoring Rules,4) Data Driven Model (DDM),5) Investigators.Προσθέτει επίσης τις μεθόδους προσέγγισης που χρησιμοποιούνται.Οι μη εποπτευόμενες μέθοδοι συνίστανται σε τεχνικές εντοπισμού

outlier / ανωμαλίας που θεωρούν ως απάτη κάθε συναλλαγή που δεν συμμορφώνεται με την πλειοψηφία. Είναι αξιοσημείωτο ότι ένα μη εποπτευόμενο DDM σε ένα FDS μπορεί να ρυθμιστεί απευθείας από συναλλαγές χωρίς σήμανση. Οι εποπτευόμενες μέθοδοι είναι μακράν οι πιο δημοφιλείς στην ανίχνευση απάτης και εκμεταλλεύονται επισημασμένες συναλλαγές για την εκπαίδευση ενός ταξινομητή. Οι απάτες εντοπίζονται με την ταξινόμηση διανυσμάτων χαρακτηριστικών των εγκεκριμένων συναλλαγών ή πιθανώς με την ανάλυση του οπίσθιου μέρους του ταξινομητή. Σε αυτό το σημείο προσθέτει τις προκλήσεις με τις οποίες έρχεται κανείς αντιμέτωπος όταν σχεδιάζει ένα FDS και αυτές αναφορικά είναι: 1) Class Imbalance, 2) Concept Drift, 3) Alert-Feedback Interaction and Sample Selection Bias. Έπειτα παρουσιάζει το σετ δεδομένων καθώς και την τεχνική μάθησης με την μορφή αλγορίθμου. Όσον αφορά το σετ δεδομένων: Χρησιμοποιούνται δύο μεγάλα σετ δεδομένων από διαδικτυακές συναλλαγές ευρωπαϊκών κατόχων πιστωτικών καρτών. Παρέχονται πληροφορίες σχετικά με αυτά τα σετ δεδομένων, τα οποία δηλώνονται το πρώτο από το 2013 και το δεύτερο από 2014-2015. Εκφράζεται η εξαιρετικά υψηλή ανισορροπία κλάσης καθώς μόνο το 0.2% αποτελούνται από λογαριασμούς απάτης.

Id	Start Day	End Day	Days	Instances	Features	Fraud Transactions (%)
2013	05-09-2013	18-01-2014	136	21,830,303	51	0.19%
2014+2015	05-08-2014	31-05-2015	296	54,764,384	51	0.24%

TABLE 23. Data Sets

Ο παραπάνω πίνακας δείχνει πιο αναλυτικά τα στοιχεία των σετ δεδομένων. Ο κάτω πίνακας μας δείχνει την προτεινόμενη στρατηγική μάθησης

Algorithm 1 Proposed Learning Strategy

Require: M and Q , i.e., the number of days of delayed samples and feedbacks to use, respectively; $\mathcal{F}_t$ and $\mathcal{D}_t$ classifiers previously trained.

$T_{t+1} \leftarrow$ transactions at day $t + 1$.

for each transaction $x \in T_{t+1}$ **do**

compute $\mathcal{P}_{\mathcal{F}_t}(+, x)$

compute $\mathcal{P}_{\mathcal{D}_t}(+, x)$

compute $\mathcal{P}_{\mathcal{A}_t}(+, x)$ as in (9)

rank T_{t+1} according to $\mathcal{P}_{\mathcal{A}_t}(+, \cdot)$,

generate alerts A_t .

if update the classifier **then**

$F_{t+1} \leftarrow$ feedbacks from cards alerted in A_t .

$\mathcal{F}_{t+1} \leftarrow \text{TRAIN}(\{F_{t+1}, \dots, F_{t-Q}\})$

$D_{t+1-\delta} \leftarrow$ transactions authorized at $t + 1 - \delta$

$\mathcal{D}_{t+1} \leftarrow \text{TRAIN}(\{D_{t+1-\delta}, \dots, D_{t-(\delta+M)}\})$

return $\mathcal{F}_t, \mathcal{D}_t$ and $\mathcal{A}_t$ defined as in (9).

Figure 8. Proposed Learning Strategy

Κλείνει λέγοντας πως τα πειράματά που πραγματοποιήθηκαν σε δύο τεράστια σύνολα δεδομένων συναλλαγών πραγματικού κόσμου δείχνουν ότι, για να λαμβάνουμε ακριβείς ειδοποιήσεις, είναι υποχρεωτικό να δίνουμε μεγαλύτερη σημασία στα σχόλια κατά τη διάρκεια του μαθησιακού προβλήματος. Τα πειράματά δείχνουν επίσης ότι οι λύσεις που μειώνουν την επίδραση των σχολίων στη διαδικασία μάθησης συχνά επιστρέφουν λιγότερο ακριβείς ειδοποιήσεις.

Στο άρθρο [15] προτείνεται ένα μέτρο αξιολόγησης που αντιπροσωπεύει ρεαλιστικά, τα νομισματικά κέρδη και τις ζημιές εξαιτίας των απατών που συμβαίνουν καθώς και του εντοπισμού της. Επιπλέον, παρουσιάζει έναν ταξινομητή Bayes Minimum Risk, συμπεριλαμβανομένου του πραγματικού οικονομικού κόστους της ανίχνευσης απάτης με πιστωτικές κάρτες, προκειμένου να είναι διαθέσιμο ένα σύστημα ανίχνευσης το οποίο θα είναι ευαίσθητο στο κόστος. Η προτεινόμενη ευαισθησία στο κόστος μειώνει σημαντικά το κόστος, σύμφωνα με το άρθρο, σε σύγκριση με άλλες τεχνικές τελευταίας τεχνολογίας. Έπειτα μιλάει λίγο πιο συγκεκριμένα για το τι είναι ο Bayes Minimum Risk και μας παρουσιάζει επίσης πληροφορίες σχετικά με το σετ δεδομένων. Ο Bayes Minimum Risk είναι ένα μοντέλο απόφασης που βασίζεται στον ποσοτικό προσδιορισμό των αντισταθμίσεων μεταξύ διαφόρων αποφάσεων χρησιμοποιώντας τις πιθανότητες και το κόστος που συνοδεύει τέτοιες αποφάσεις. Στην περίπτωση εντοπισμού απάτης με πιστωτική κάρτα, υπάρχουν δύο αποφάσεις που λαμβάνει, είτε προβλέπει μια συναλλαγή ως απάτη είτε ως νόμιμη. Όσον αφορά το σετ δεδομένων: Χρησιμοποιείται ένα σετ δεδομένων το οποίο παρέχεται από μια μεγάλη Ευρωπαϊκή εταιρεία κάτοχο πιστωτικών καρτών. Περιέχει fraudulent και legitimate συναλλαγές που έγιναν με χρεωστικές και πιστωτικές κάρτες το 2012 .

Συνολικά περιέχει 80.000.000 ξεχωριστές συναλλαγές εκ των οποίων η κάθε μία έχει 27 γνωρίσματα. Από τα αρχικά γνωρίσματα επιλέγονται αυτά που περιέχουν χρήσιμες πληροφορίες με την βοήθεια της risk team της εταιρίας. Χρησιμοποιώντας τα αρχικά γνωρίσματα έχουν παραχθεί ακόμα 260 γνωρίσματα. Η ιδέα πίσω από την παραγωγή αυτών των γνωρισμάτων είναι η χρήση στρατηγική συγκέντρωσης συναλλαγών, προκειμένου να συλληφθεί τη συμπεριφορά των καταναλωτικών δαπανών στο πρόσφατο παρελθόν. Στο σέτ δεδομένων μόνο 20.000 συναλλαγές είναι μαρκαρισμένες ως *frauded*, γεγονός που οδηγεί σε *frauded ratio* 0.025%. Για το πείραμα επιλέγεται ένα μικρότερο υποσύνολο δεδομένων με μεγαλύτερο *frauded ratio*, που απαντά σε συγκεκριμένο γκρουπ συναλλαγών. Αυτή η βάση δεδομένων περιέχει 750.000 συναλλαγές με *frauded ratio* 0.467%. Για αυτό το σέτ δεδομένων οι πρώτοι 8 μήνες χρησιμοποιήθηκαν ως σέτ εκπαίδευσης και οι τελευταίοι 2 μήνες ως σέτ ελέγχου. Δημιουργήθηκαν 5 διαφορετικές βάσεις δεδομένων S1, S5, S10, S20 και S50, καθεμία από τις οποίες έχει διαφορετικό ποσοστό απάτης 1%, 5%, 10%, 20% και 50%, αντίστοιχα.

Database	Transactions	Frauds	Fraud Ratio	Fraud Amount
Total	750,000	3,500	0.467%	866,410
Train	625,000	2,900	0.464%	721,349
Test	125,000	600	0.480%	148,562
S1	290,000	2,900	1%	721,349
S5	58,000	2,900	5%	721,349
S10	29,000	2,900	10%	721,349
S20	17,500	2,900	20%	721,349
S50	5,800	2,900	50%	721,349

TABLE 24. Description of Databases: One Training Database and Different Under-sampled Databases Varying in Percentage of Frauds

Ο παραπάνω πίνακας 24 μας δείχνει τον τρόπο που έχει χωριστεί το αρχικό σέτ δεδομένων τόσο με τον διαχωρισμό σέτ εκπαίδευσης και ελέγχου αλλά και τον διαχωρισμό με την χρήση υπο δειγματοληψίας

Το άρθρο αξιολογεί τους αλγόριθμους LR, C4.5 και RF, στην πλήρη βάση δεδομένων και στις διαφορετικές βάσεις δεδομένων που δεν έχουν δειγματοληψία, προκειμένου να κάνει την σύγκριση. Καταλήγει στο γεγονός ότι οι παραδοσιακοί αλγόριθμοι αποδίδουν μόνο όταν λειτουργούν με μια μέθοδο υπο δειγματοληψίας.

Οι παράμετροι που χρησιμοποιούνται για σύγκριση διαφόρων συστημάτων ανίχνευσης απάτης είναι η ακρίβεια, το ποσοστό ανίχνευσης απάτης ως προς το True Positive και false positive, το κόστος και η εκπαίδευση απαιτείται, η εποπτευόμενη μάθηση. Τελειώνει λέγοντας πως όλες οι τεχνικές εντοπισμού απάτης με πιστωτικές κάρτες που συζητούνται σε αυτό το έγγραφο έρευνας έχουν τα δικά τους πλεονεκτήματα και αδυναμίες. Μια τέτοια έρευνα θα επιτρέψει την οικοδόμηση μιας υβριδικής προσέγγισης για τον εντοπισμό δόλιων συναλλαγών με πιστωτικές κάρτες.

Το άρθρο [17] ασχολείται με την έρευνα απατών σε πιστωτικές κάρτες χρησιμοποιώντας τον αλγόριθμο KNN και μία μέθοδο ανίχνευσης outlier. Αρχικά μιλάει για το κομμάτι της ταξινόμησης και μας λέει ότι αποτελείται από την ανάθεση επισημάνσεων κλάσης σε ένα σύνολο με μη κατηγοριοποιημένες περιπτώσεις. Συνεχίζει πάνω στο ίδιο κομμάτι και αναφέρει πως τα συστήματα ταξινόμησης εκτελούνται σε δύο προσεγγίσεις statistical and structural. Επιπλέον μας λέει ότι η ανίχνευση απατών είναι μία εξαιρετικά πολύπλοκη λειτουργία για να εκτελεστεί, καθώς και ότι δεν υπάρχει κάποιο σύστημα το οποίο μπορεί να εγγυηθεί 100% επιτυχία. Όλες οι υπάρχουσες μέθοδοι ενδέχεται να προβλέψουν μια απατηλή συναλλαγή και ταυτόχρονα να μην διαβεβαιώσουν για τα αποτελέσματα. Σύμφωνα με το άρθρο, προκειμένου να δουλέψουν οι τεχνικές ανίχνευσης απάτης βασισμένες στον KNN χρειάζονται δύο πράγματα : το μέτρο της απόστασης ή της ομοιότητας ανάμεσα σε δύο περιπτώσεις δεδομένων. Προσθέτει ότι η προσέγγιση outlier είναι αρκετά διαφορετική από άλλες παραδοσιακές μεθόδους παρατήρησης. Χρησιμοποιείται ώστε να ανιχνευθεί μία απρόσμενη συμπεριφορά ενός συστήματος χρησιμοποιώντας διαφορετικό μηχανισμό. Η ανίχνευση outlier έχει δύο κύριους τύπους ανάλογα με τις επισημάνσεις που δίνονται σε κάθε γκρουπ.

A. Στην εποπτευόμενη μέθοδο ανίχνευσης outlier, οι ειδικοί του τομέα μοντελοποιούν το σύστημα για να μάθει και να ταξινομήσει το outlier χρησιμοποιώντας ένα εκπαιδευτικό σύνολο δεδομένων.

B. Στην μη εποπτευόμενη μέθοδος ανίχνευσης outlier, αντικείμενα συγκεντρώνονται σε πολλαπλές ομάδες ανάλογα με τα χαρακτηριστικά τους. Αυτά τα αντικείμενα που απέχουν πολύ από οποιαδήποτε ομάδα επισημαίνονται ως ακραία σημεία.

Έπειτα επισημαίνει πως σε μία μέθοδο ανίχνευσης outlier η μη εποπτευόμενη μάθηση προτιμάται για την ανίχνευση απατών καθώς , μπορεί να οδηγήσει σε νέες εξηγήσεις και αναπαραστάσεις των δεδομένων που παρατηρούνται . Ένα ακόμα πλεονέκτημα της χρήσης μη εποπτευόμενης μάθησης είναι ότι δεν απαιτεί προηγούμενη επισημάνση δεδομένων ή γνώσεων σχετικά με δόλιες μεθόδους ή συναλλαγές . Έτσι δεν χρειάζεται να εκπαιδευτεί προκειμένου να κάνει τον διαχωρισμό ανάμεσα σε νόμιμες και μή νόμιμες συναλλαγές, αλλά ακολουθεί την φυσιολογική συμπεριφοριστική ακολουθία ως ασυνήθιστη ενέργειά ή απατηλή ενέργεια. Τέλος εξηγεί πως χρησιμοποιώντας υπέρ δειγματοληψία και εξάγοντας κύρια κατεύθυνση των δεδομένων μπορεί να χρησιμοποιηθεί η μέθοδος KNN του άρθρου ώστε να προσδιορίσει την ανωμαλία του παραδείγματος-στόχου. Έτσι η μέθοδος KNN μπορεί να χρησιμοποιηθεί για την ανίχνευσης απατών με τον περιορισμό της μνήμης.

Παράλληλα ο μηχανισμός ανίχνευσης outlier βοηθάει στην ανίχνευση απατών σε πιστωτικές κάρτες χρησιμοποιώντας λιγότερη μνήμη και υπολογιστικούς πόρους. Ιδιαίτερα η ανίχνευση outlier λειτουργεί γρήγορα και καλά σε μεγάλα σύνολα δεδομένων που βρίσκονται στο διαδίκτυο. Αλλά σε σύγκριση με τις μεθόδους ισχύος και άλλες γνωστές μεθόδους ανίχνευσης ανωμαλιών, τα πειραματικά αποτελέσματα αποδεικνύουν ότι η μέθοδος KNN είναι ακριβής και αποτελεσματική.

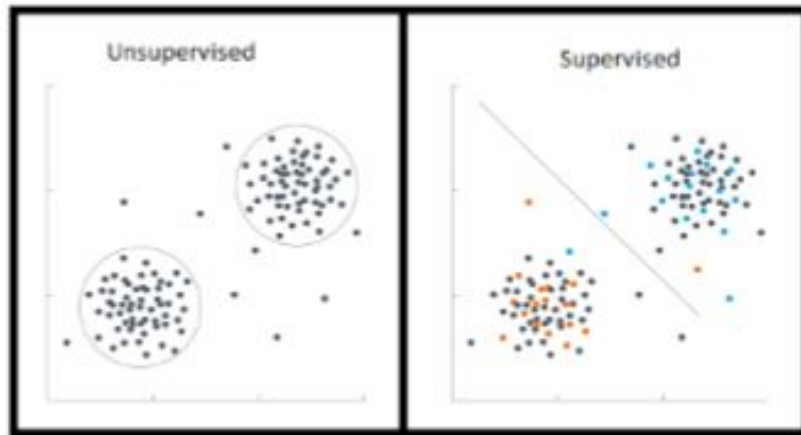


Figure 10. Outlier Detection Using Unsupervised and Supervised methods

Το παραπάνω σχήμα μας δείχνει την ανίχνευση outlier χρησιμοποιώντας εποπτευόμενες και μη εποπτευόμενες μεθόδους

Ο στόχος του άρθρου [18] είναι να αξιολογήσει τις διαφορετικές τοπολογίες πάνω στο κομμάτι Deep Learning όσον αφορά την αποτελεσματικότητά τους στην ανίχνευση απατών σε πιστωτικές κάρτες. Αναλύονται αρκετοί παράμετροι οι οποίο χρησιμοποιούνται για να κατασκευάσουν το μοντέλο έτσι ώστε να βρεθεί ο βέλτιστος συνδυασμός παραμέτρων για την ανίχνευση απειλών ενεργειών/δράσεων. Το σύνολο δεδομένων που χρησιμοποιήθηκε σε αυτήν τη μελέτη παρέχεται από ένα χρηματοπιστωτικό ίδρυμα που ασχολείται με τη λιανική τραπεζική. Περιέχει πληροφορίες για περίπου 80 εκκ. ανώνυμες συναλλαγές μέσα σε ένα χρονικό διάστημα 8 μηνών. Οι πληροφορίες που παρέχονται παρέχουν τόσο δεδομένα συναλλαγών όσο και λογαριασμών. Η μεταβλητή απόκρισης έχει προ-επισημανθεί ως 1 για δόλια συναλλαγή και 0 για μη δόλια συναλλαγή. Υπάρχει μία σημαντική ανισορροπία κλάσης. Μόνο το 0.14% είναι απειληκές συναλλαγές συγκριτικά με το 99.86% οι οποίες ταξινομούνται ως νόμιμες. Υπήρχαν εκτεταμένα δεδομένα που λείπουν σχετικά με την ώρα και τις κατηγορηματικές μεταβλητές, Συναλλαγές με ελλιπή δεδομένα δεν αφαιρέθηκαν από το σετ δεδομένων που χρησιμοποιήθηκε στην ανάλυση. Επίσης μα λέει ότι προστέθηκαν μεταβλητές dummy ώστε να επισημάνουν τα χαμένα δεδομένα για categorical and time variables που είχαν χαμένα δεδομένα. Στην έρευνα που πραγματοποιήθηκε, ταυτοποιήθηκαν και δημιουργήθηκαν μερικοί σημαντικοί και βιομηχανικοί παράγοντες όπως:

1) Frequency of transaction per month,

- 2)Missing data dummy variables
- 3)Dummy variables to indicate when a purchase occurs at a specific merchant
- 4)Account history variables such as number of transactions from the account in the 8-month period covered in the dataset, maximum and mean authorization amount for the account
- 5)A dummy variable that indicate if a transaction authorization amount made at a given merchant is greater than 10% of the standard deviation of the mean of the non-fraudulent transactions for the merchant and the transaction deviation from the non fraud mean is in the direction of the fraudulent mean.

Η ανάλυση που πραγματοποιείται εμπεριέχει τέσσερις Deep Learning τοπολογίες :

- 1)Artificial Neural Networks (ANNs)
- 2)Recurrent Neural Networks (RNNs)
- 3)Long Short-term Memory (LSTMs)
- 4) Gated Recurrent Units (GRUs).

Όπως αναφέρεται στο άρθρο, χρειάστηκε να σιγουρευθεί ότι κάθε λογαριασμός που δοκιμάστηκε είχε όλες του τις συναλλαγές παρών στο σετ εκπαίδευσης. Αυτό εμπεριέχει τον διαχωρισμό του σετ δεδομένων σε απατηλά και μη απατηλά και την εισαγωγή μοναδικών αριθμών λογαριασμού για το απατηλό σετ δεδομένων. Στη συνέχεια, δείχνονται τυχαία μη δόλιες συναλλαγές από το μη δόλιο σύνολο δεδομένων για τους αριθμούς λογαριασμού που εξήχθησαν. Ο δείκτης δειγματοληψίας που επιλέχθηκε ήταν 10: 1 ο οποίος έχει αποδειχθεί ιδανικός για τον εντοπισμό απάτης με πιστωτικές κάρτες. Τέλος, χρησιμοποιήθηκε κωδικοποίηση one-hot για να αντιπροσωπεύουν κατάλληλα τις κατηγορηματικές μεταβλητές.

Τελειώνει αναφέροντας τα αποτελέσματα για κάθε μία από τις τοπολογίες που αναφέρθηκαν παραπάνω , πιο συγκεκριμένα:

- 1)Συνολικά, η αρχιτεκτονική GRU παρήγαγε το μοναδικό μοντέλο με την καλύτερη απόδοση με 10-fold cross validated accuracy 0,916. Το μοντέλο περιείχε 6 κρυμμένα στρώματα το καθένα με 150 κόμβους.
- 2)Το μοντέλο LSTM με την καλύτερη απόδοση περιείχε 6 κρυμμένα στρώματα το καθένα με 150 κόμβους. Το καθένα είχε βαθμολογία ακρίβειας 0,912.
- 3)Το μοντέλο RNN με την καλύτερη απόδοση περιείχε 6 κρυμμένα στρώματα το καθένα με 150 κόμβους και είχε βαθμολογία ακρίβειας 0,90433.
- 4)Το ANN παρήγαγε τα χειρότερα αποτελέσματα μεταξύ των 4 τοπολογιών με βαθμολογία ακρίβειας 0,889. Το μοντέλο ANN με την καλύτερη απόδοση περιείχε 6 κρυμμένα στρώματα το καθένα με 150 κόμβους.

Σύνοψη της όλης εργασίας

Ο σκοπός της εργασίας είναι η σύγκριση διαφόρων αλγορίθμων μηχανικής μάθησης ο οποίοι χρησιμοποιούνται για την ανίχνευση απατών στις πιστωτικές κάρτες. Πρώτο βήμα είναι η παρουσίαση του μεγέθους του προβλήματος των απατών στις πιστωτικές κάρτες, κάτι που μας προβληματίζει αρκετά, διότι έχει σοβαρές επιπτώσεις στην πορεία των ηλεκτρονικών επιχειρήσεων και μιας και ζούμε σε μία εποχή που χαρακτηρίζεται από την ηλεκτρονική διασύνδεση και διακυβέρνηση , το καθιστά εξαιρετικά σημαντικό πρόβλημα. Στην συνέχεια παρουσιάζονται επιστημονικά άρθρα τα οποία σχετίζονται με το θέμα της ανίχνευσης απατών στις πιστωτικές κάρτες. Αυτό το βήμα είναι πολύ σημαντικό, καθώς μας φέρνει κοντά με διάφορους τρόπους σκέψης αλλά και έρευνες, γεγονός που μας βοηθάει να έχουμε μία ποιά συγκεκριμένη εικόνα σχετικά με το θέμα αλλά και με τους αλγορίθμους που χρησιμοποιούνται, με τεχνικές προεπεξεργασίας δεδομένων και πολλές άλλες τεχνικές που χρησιμοποιούνται με σκοπό να αντιμετωπιστεί αυτό το πρόβλημα των απατών. Μας επιτρέπει επίσης να συγκρίνουμε τα αποτελέσματα της εργασίας αυτής, με αποτελέσματα άλλων εργασιών ώστε να δούμε κατα πόσο είναι αποτελεσματική. Έπειτα παρουσιάζουμε και αναλύουμε το σετ δεδομένων που θα χρησιμοποιηθεί για την εξαγωγή αποτελεσμάτων, την μεθοδολογία που θα ακολουθηθεί για την εξαγωγή των αποτελεσμάτων που θα μας δώσουν το συμπέρασμα και βέβαια μια εικόνα για τους αλγορίθμους που θα χρησιμοποιηθούν και πως λειτουργεί ο καθένας από αυτούς. Επιπλέον , θα πραγματοποιηθεί αξιολόγηση των αποτελεσμάτων για να δούμε πόσο καλά απαντούν στις ανάγκες της εργασίας. Και στο τέλος, θα προταθούν τα συμπεράσματα του όλου εγχειρήματος , ποιές θα είναι οι μελλοντικές κινήσεις καθώς και προτάσεις γι αβελτίωση.

Κεφάλαιο 3ο

Περιγραφή Data Set

Το σετ δεδομένων που χρησιμοποιήθηκε στην εργασία αυτή , είναι από την ιστοσελίδα [kaggle.com](https://www.kaggle.com). Περιέχει συναλλαγές που πραγματοποιήθηκαν από ευρωπαίους κατόχους πιστωτικών καρτών τον μήνα Σεπτέμβριο το 2013. Το σετ δεδομένων αυτό, παρουσιάζει συναλλαγές που πραγματοποιήθηκαν μέσα σε δύο μέρες. Ποιά συγκεκριμένα έχουμε 492 απατηλές συναλλαγές από τις 284.807. Απο τα νούμερα αυτά καταλαβαίνουμε ότι έχουμε να αντιμετωπίσουμε ένα αρκετά ανισόρροπο σετ δεδομένων μια και οι συναλλαγές που είναι θετικές στο θέμα της απάτης αποτελούν το 0.172% του συνολικού σετ δεδομένων. Περιέχει

μόνο αριθμητικές τιμές, γεγονός που οφείλεται στην επεξεργασία PCA που έχει υποστεί. Δυστυχώς για λόγους εμπιστευτικότητας και προστασίας προσωπικών δεδομένων, είναι αδύνατο να έχουμε τα πραγματικά χαρακτηριστικά. Γι' αυτό τον λόγο τα χαρακτηριστικά που συναντάμε έχουμε την μορφή V1,V2,V3...,V28. Στο σημείο αυτό να σημειωθεί ότι τα μόνο από τα αρχικά χαρακτηριστικά τα οποία δεν έχουν υποστεί επεξεργασία PCA είναι τα χαρακτηριστικά "Time" και "Amount". Το χαρακτηριστικό "Time" περιέχει το χρονικό διάστημα ανάμεσα σε κάθε συναλλαγή με την πρώτη συναλλαγή. Το χαρακτηριστικό "Amount" περιέχει το συνολικό ποσό της κάθε συναλλαγής. Τέλος το χαρακτηριστικό "Class" παίρνει τιμές 0 ή 1 ανάλογα αν μια συναλλαγή είναι απατηλή(τότε παίρνει την τιμή 1) ή όχι (συνεπώς παίρνει την τιμή 0)

Περιγραφή Μεθοδολογίας

Προκειμένου να προχωρήσουμε με την όλη διαδικασία απαιτούνται ορισμένες ενέργειες έτσι ώστε να επιτευχθεί ο στόχος μας. Παρατηρείται ότι το σετ δεδομένων που έχουμε είναι αρκετά ανισόρροπο, δηλαδή έχουμε σημαντικά περισσότερα δείγματα από λογαριασμούς οι οποίοι δεν έχουν υποστεί απάτη συγκριτικά με αυτούς που έχουν υποστεί. Για να γίνει αντιληπτή αυτή η διαφορά, αρκεί να πούμε πως μόνο το 0.172% του συνόλου των λογαριασμών έχουν βρεθεί θύματα απάτης. Έπειτα από αρκετές δοκιμές, ο τρόπος προσέγγισης που καταλήξαμε ήταν ο εξής: Πρώτο βήμα είναι ο διαχωρισμός του συνόλου δεδομένων σε 10 μικρότερα σετ τα οποία είχαν τον ίδιο αριθμό λογαριασμών που έχουν υποστεί απάτη, με τους λογαριασμούς οι οποίοι είναι νόμιμοι, δηλαδή δεν είναι θύματα απάτης. Να σημειωθεί ότι ο αριθμός των λογαριασμών που ήταν θύματα απάτης κρατήθηκε ανέπαφος κατά την όλη διαδικασία διαχωρισμού. Αυτός ο διαχωρισμός επιτεύχθηκε με την χρήση της μεθόδου RESAMPLE στην καρτέλα Preprocess του εργαλείου WEKA. Στην συνέχεια, τρέξαμε τους αλγόριθμους :1)Random Forest (RF), 2) K-nearest Neighbours (KNN),3) Naive Bayes(NB),4)Logistic , 5)Support Vector Machine(SVM), δύο φορές για το κάθε υπο-σετ δεδομένων από τα 10 που παρήχθησαν από τον αρχικό διαχωρισμό, μία χρησιμοποιώντας cross validation x10 folds χωρίς feature selection και μία χρησιμοποιώντας cross validation x10 folds με feature selection και WrapperSubset Evaluator και μέθοδο Best First . Έπειτα, βγάλαμε έναν μέσο όρο για τις μετρήσεις precision,recall,f-measure για κάθε αλγόριθμο που χρησιμοποιήθηκε και για τις δύο περιπτώσεις, δηλαδή με και χωρίς feature selection και πάνω σε αυτά τα αποτελέσματα πραγματοποιούμε την αξιολόγηση.

Περιγραφή Αλγορίθμων

1) Αλγόριθμος Random Forest:

- Ο αλγόριθμος Random Forest είναι ένας αλγόριθμος ο οποίος σε γενικές γραμμές δημιουργεί ένα “δάσος” από ένα σύνολο “τυχαίων δέντρων”
- Ποιό συγκεκριμένα, ο RF έχει δύο στάδια, το πρώτο είναι η δημιουργία του τυχαίου “δάσους” και το δεύτερο στάδιο είναι η παραγωγή προβλέψεων με βάση το “δάσος” που δημιουργήθηκε στο πρώτο βήμα.
- Ο τρόπος με τον οποίο λειτουργεί είναι:
 - Επιλέγει τυχαία χ χαρακτηριστικά από ένα σύνολο Ψ χαρακτηριστικών όπου χ πολύ μικρότερο του $\Psi (\chi \ll \Psi)$
 - Από τα χ χαρακτηριστικά, υπολογίζει τον κόμβο κ χρησιμοποιώντας το καλύτερο σημείο διαχωρισμού
 - Χωρίζει τον κόμβο σε “κόμβους παιδιά” χρησιμοποιώντας τον βέλτιστο διαχωρισμό
 - Επαναλαμβάνονται τα πρώτα 3 στάδια έως ότου να έχει επιτευχθεί ένας αριθμός i κόμβων
 - Επαναλαμβάνονται τα πρώτα 4 στάδια n φορές έτσι ώστε να δημιουργηθεί n αριθμός δέντρων
- Ο τρόπος με τον οποίο πραγματοποιούνται οι προβλέψεις είναι :
 - Λαμβάνει τα δοκιμαστικά χαρακτηριστικά και χρησιμοποιεί τους κανόνες κάθε τυχαίου δέντρου απόφασης έτσι ώστε να προβλέψει το αποτέλεσμα το οποίο και αποθηκεύει
 - Υπολογίζει τις “ψήφους” για κάθε προβλεπόμενο στόχο
 - Εξετάζει τον στόχο με τους περισσότερους “ψήφους” ως την τελική πρόβλεψη του αλγορίθμου
- Ο λόγος που επιλέγεται ο RF είναι επειδή λειτουργεί πολύ καλά και αυτό οφείλεται στο γεγονός ότι τα μη συσχετισμένα μοντέλα μπορούν να παράγουν σύνολα προβλέψεων που είναι πιο ακριβείς από οποιαδήποτε από τις μεμονωμένες προβλέψεις. Ο λόγος για αυτό το αποτέλεσμα είναι ότι τα δέντρα μεριμνούν το ένα για άλλο σχετικά με τα ατομικά τους λάθη (αρκεί να μην έχουν πάντα λάθος στην ίδια κατεύθυνση). Ενώ ορισμένα δέντρα μπορεί να είναι λάθος, πολλά άλλα δέντρα θα είναι σωστά, έτσι ως ομάδα τα δέντρα μπορούν να κινηθούν προς τη σωστή κατεύθυνση

- Η παρακάτω εικόνα αποτελεί μια γραφική απεικόνιση του αλγορίθμου:

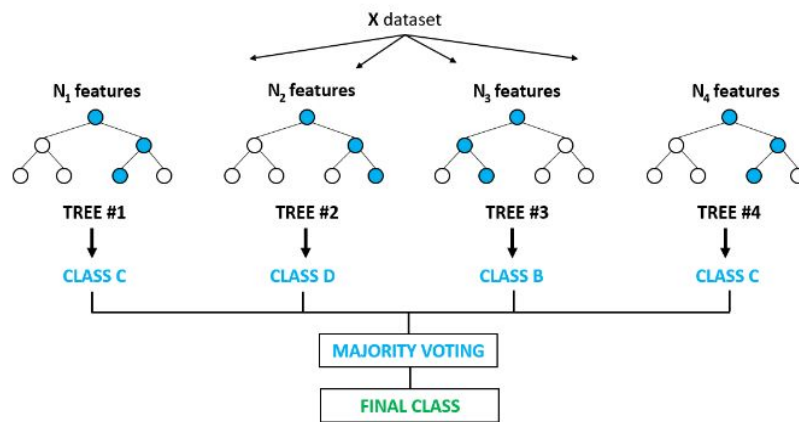


Figure 11. Graphic representation of Random Forest algorithm

2) Αλγόριθμος K-nearest Neighbours:

- Ο KNN αλγόριθμος θεωρεί ότι παρόμοια πράγματα υπάρχουν πολύ κοντά το ένα με το άλλο.
- Ο KNN συλλαμβάνει την ιδέα της ομοιότητας με κάποια μαθηματικά που μπορεί να μάθαμε στην παιδική μας ηλικία - υπολογίζοντας την απόσταση μεταξύ των σημείων σε ένα γράφημα.
- Το KNN λειτουργεί εντοπίζοντας τις αποστάσεις μεταξύ ενός ερωτήματος και όλων των παραδειγμάτων στα δεδομένα, επιλέγοντας τα καθορισμένα παραδείγματα αριθμών (K) που είναι πλησιέστερα στο ερώτημα και, στη συνέχεια, ψηφίζει για την πιο συχνή ετικέτα (στην περίπτωση της ταξινόμησης) ή υπολογίζει κατά μέσο όρο τις ετικέτες (σε η περίπτωση της παλινδρόμησης).
- Χρησιμοποιώ τον αλγόριθμο αυτό διότι:
 - Ο αλγόριθμος είναι απλός και εύκολο να εφαρμοστεί.
 - Δεν χρειάζεται να δημιουργηθεί ένα μοντέλο, να συντονιστούν αρκετές παράμετροι ή να γίνουν επιπλέον παραδοχές
 - Ο αλγόριθμος είναι ευέλικτος. Μπορεί να χρησιμοποιηθεί για ταξινόμηση, παλινδρόμηση και αναζήτηση
 - Παρακάτω βλέπουμε μια γραφική απεικόνιση της λειτουργία του αλγορίθμου :

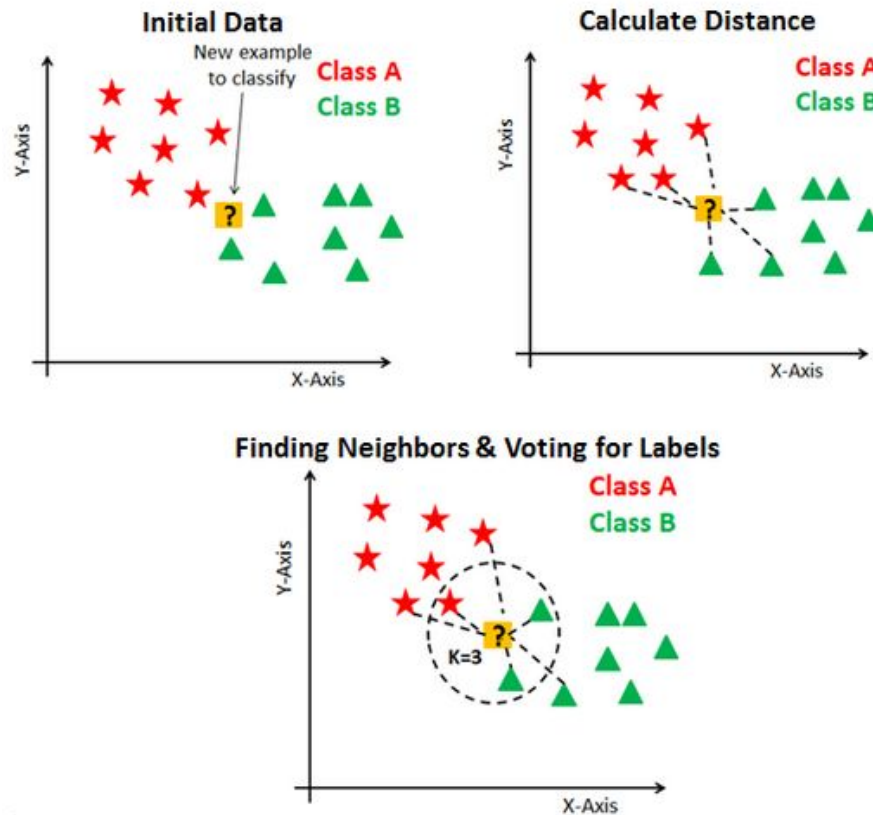


Figure 12. Graphic representation of KNN algorithm

3) Αλγόριθμος Naive Bayes:

- Είναι ένας πιθανοτικός αλγόριθμος μηχανικής μάθησης
- Ο όρος Naive χρησιμοποιείται διότι υποθέτει ότι τα χαρακτηριστικά που πηγαίνουν στο μοντέλο είναι ανεξάρτητα το ένα από το άλλο. Αυτό αλλάζει την τιμή μιας δυνατότητας, δεν επηρεάζει άμεσα ή αλλάζει την τιμή οποιασδήποτε από τις άλλες δυνατότητες που χρησιμοποιούνται στον αλγόριθμο.
- Επιλέγεται ο Naive Bayes διότι στηρίζεται στην θεωρία Bayes, ότι η πιθανότητα να συμβεί ένα γεγονός **A** είναι υψηλή εάν έχει συμβεί ένα άλλο γεγονός **B** κάτι που σχετίζεται αρκετά με το θέμα
- Τα πλεονεκτήματα της χρήσης του NB είναι τα εξής:
 - Απλότητα του μοντέλου
 - Ταχύτητα
 - Είναι κλιμακούμενος
 - Απαιτεί ελάχιστα δεδομένα

- Παρακάτω βλέπουμε μια εικονική αναπαράσταση του αλγορίθμου

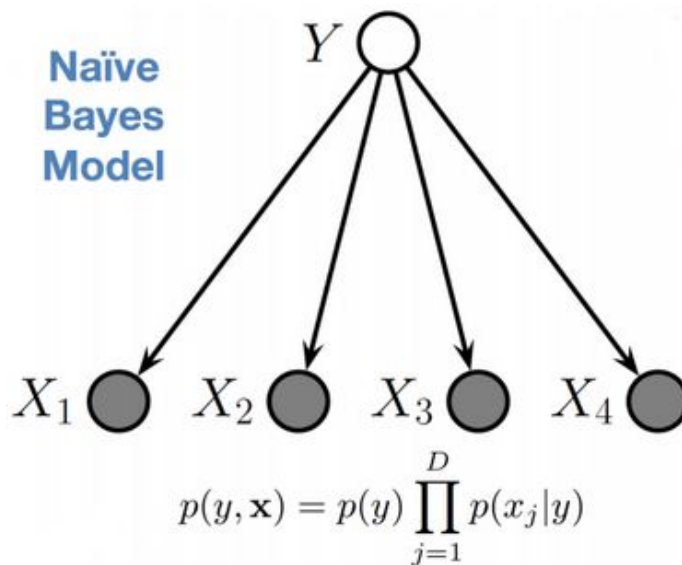


Figure 13. Graphic representation of Naive Bayes algorithm

4) Αλγόριθμος Logistic :

- Ο αλγόριθμος Logistic Regression είναι μια γρήγορη και απλή τεχνική, αλλά μπορεί να είναι πολύ αποτελεσματική σε ορισμένα προβλήματα.
- Ο αλγόριθμος μαθαίνει έναν συντελεστή για κάθε τιμή εισόδου, οι οποίοι συνδυάζονται γραμμικά σε μια συνάρτηση παλινδρόμησης και μετασχηματίζονται χρησιμοποιώντας μια συνάρτηση logistic
- Ο logistic regression υποστηρίζει μόνο δυαδικά προβλήματα ταξινόμησης, αν και η εφαρμογή Weka έχει προσαρμοστεί για να υποστηρίζει προβλήματα ταξινόμησης πολλαπλών κλάσεων.
- Παρακάτω βλέπουμε μια γραφική αναπαράσταση του αλγορίθμου :

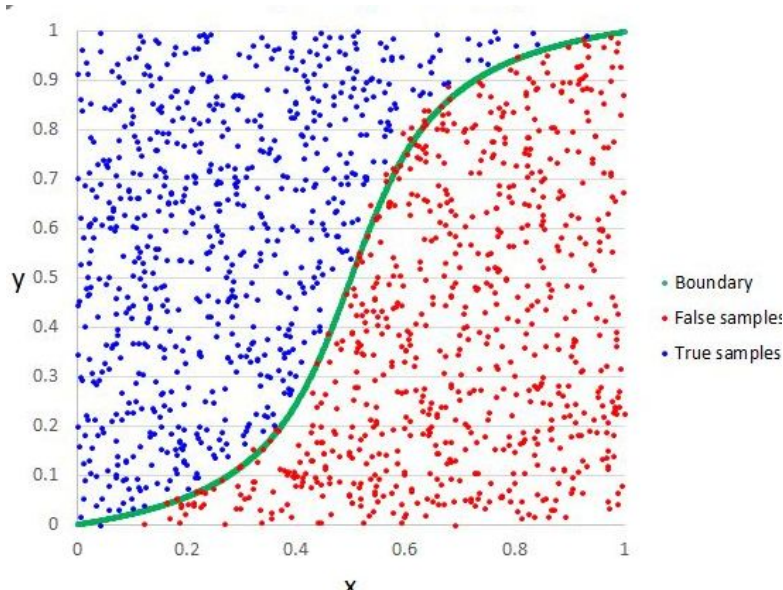


Figure 14. Graphic representation of Logistic Regression algorithm

5) Support Vector Machine:

- Το SVM αναπτύχθηκε για αριθμητικές μεταβλητές εισόδου, αν και θα μετατρέψει αυτόματα ονομαστικές τιμές σε αριθμητικές τιμές. Τα δεδομένα εισόδου ομαλοποιούνται επίσης πριν χρησιμοποιηθούν.
- Το SVM λειτουργεί βρίσκοντας μια γραμμή που διαχωρίζει καλύτερα τα δεδομένα στις δύο ομάδες. Αυτό γίνεται χρησιμοποιώντας μια διαδικασία βελτιστοποίησης που λαμβάνει υπόψη μόνο τις παρουσίες δεδομένων στο σύνολο δεδομένων εκπαίδευσης που είναι πιο κοντά στη γραμμή που διαχωρίζει καλύτερα τις κλάσεις. Οι περιπτώσεις ονομάζονται διανύσματα υποστήριξης, εξ ου και το όνομα της τεχνικής.
- Σε σχεδόν όλα τα προβλήματα ενδιαφέροντος, δεν μπορεί να σχεδιαστεί μια γραμμή για να διαχωριστούν τακτικά οι τάξεις, επομένως προστίθεται ένα περιθώριο γύρω από τη γραμμή για να χαλαρώσει ο περιορισμός, επιτρέποντας σε ορισμένες περιπτώσεις να ταξινομηθούν εσφαλμένα, αλλά επιτρέποντας ένα καλύτερο αποτέλεσμα συνολικά.
- Λίγα σύνολα δεδομένων μπορούν να διαχωριστούν με μια ευθεία γραμμή. Μερικές φορές πρέπει να επισημανθεί μια γραμμή με καμπύλες ή ακόμη και πολυγωνικές περιοχές. Αυτό επιτυγχάνεται με το SVM προβάλλοντας τα δεδομένα σε χώρο υψηλότερης διαστάσεων για τον σχεδιασμό των γραμμών και την πραγματοποίηση των προβλέψεων. Μπορούν να χρησιμοποιηθούν διαφορετικοί πυρήνες για τον έλεγχο της προβολής και του βαθμού ευελιξίας στο διαχωρισμό των τάξεων
- Παρακάτω βλέπουμε μια γραφική αναπαράσταση του SVM:

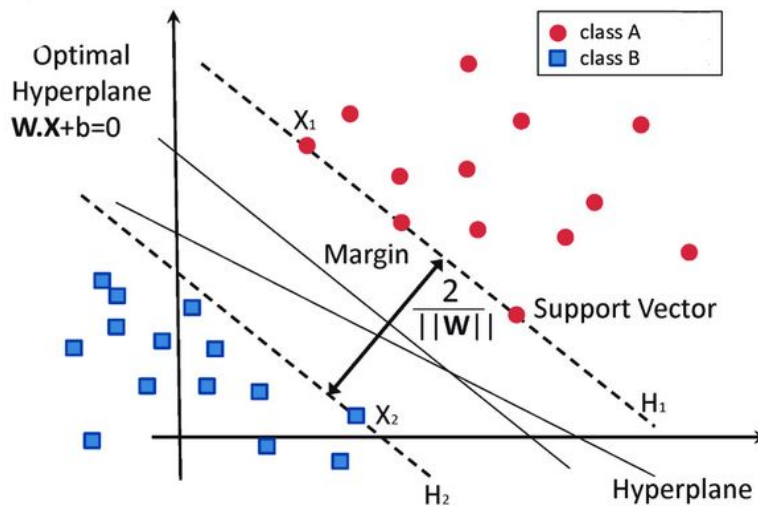


Figure 15. Graphic representation of SVM algorithm

Κεφάλαιο 4ο

Αποτελέσματα

Η σύγκριση των μέσων όρων των αλγορίθμων γίνεται με τα ποσοστά τους σε Precision, Recall και F-Measure. Το precision είναι είναι μια μέτρηση που ποσοτικοποιεί τον αριθμό των σωστών θετικών προβλέψεων, συνεπώς υπολογίζει την ακρίβεια για την minority class. Υπολογίζεται ως ο λόγος των σωστά προβλεπόμενων θετικών παραδειγμάτων δια του συνολικού αριθμού των θετικών παραδειγμάτων που είχαν προβλεφθεί. Το recall είναι μια μέτρηση που ποσοτικοποιεί τον αριθμό των σωστών θετικών προβλέψεων που έγιναν από όλες τις θετικές προβλέψεις που θα μπορούσαν να είχαν γίνει. Παρέχει επίσης μια ένδειξη χαμένων θετικών προβλέψεων. Και τέλος το f-measure παρέχει έναν τρόπο συνδυασμού τόσο του precision όσο και του recall σε ένα μόνο μέτρο που συλλαμβάνει και τις δύο ιδιότητες. Από μόνα τους ούτε το precision ούτε το recall μας δίνουν ολοκληρωμένη εικόνα. Μπορούμε να έχουμε εξαιρετικό precision και αρκετά καλό recall, ή εναλλακτικά, αρκετά καλό precision με εξαιρετικό recall. Το f-measure παρέχει έναν τρόπο να εκφράσουμε και τις δύο ανησυχίες με ένα μόνο αποτέλεσμα.

1ος πίνακας αποτελεσμάτων

Αλγόριθμοι	Precision	Recall	F-Measure
IBK	0,784	0,9632	0,9161
	0,8663	0,8624	0,9064
Naive Bayes	0,8654	0,9854	0,9227
	0,8863	0,843	0,9099
Random Forest	0,8911	0,9734	0,9302
	0,8797	0,8811	0,9238
Logistic	0,7572	0,8045	0,7801
	0,7913	0,7415	0,7457
SMO	0,9022	0,9226	0,9121
	0,92	0,8998	0,9098

Στον παραπάνω πίνακα απεικονίζεται ουσιαστικά ο μέσος όρος για το precision, recall καθώς και f-measure του κάθε αλγορίθμου χωρίς την χρήση feature selection, ο οποίος προκύπτει όπως έχω ήδη αναφέρει εφαρμόζοντας 10-fold cross validation σε κάθε αλγόριθμο σε κάθε ένα από τα 10 υπο-σετ που δημιουργήθηκαν από το αρχικό. Παρατηρούμε ότι με εξαίρεση τον Logistic Regression η μέτρηση F-measure του μέσου όρου των αλγορίθμων είναι άνω του 0,9. Η μεγαλύτερη ακρίβεια συναντάται στον SMO ενώ το καλύτερο Recall στους IBk και Random Forest.

2ος πίνακας αποτελεσμάτων

Αλγόριθμοι	Precision	Recall	F-Measure
IBK	0,8949	0,9621	0,9272
	0,9594	0,8866	0,9218
Naive Bayes	0,8973	0,986	0,9395
	0,9843	0,887	0,933
Random Forest	0,8951	0,9859	0,9381
	0,9842	0,8842	0,9315
Logistic	0,8776	0,9422	0,9087
	0,9376	0,8686	0,9017
SMO	0,8901	0,972	0,9293
	0,9692	0,8798	0,9225

Στον δεύτερο πίνακα έχει εφαρμοστεί η ίδια διαδικασία με πριν, μόνο που αυτή την φορά έχουμε χρησιμοποιήσει feature selection και πιο συγκεκριμένα Attribute Selection Classifier με WrapperSubset Evaluator και μέθοδο Best First. Τα attributes για κάθε αλγόριθμο είναι :

IBk : V4,V12,V14,V17,V20

RandomForest(RF): V1,V4,V6,V7,V8,V12,V13,V14,V17,V18,V19,V23,V24,V27,V28

NaiveBayes(NB): V4,V14,V17,V22,V23,V28

Logistic : V4,V12,V14

SMO : V14,V17

Βασιζόμενοι πάντα στα αποτελέσματα που παρουσιάζονται στο παραπάνω πίνακες βλέπουμε ότι κάποιο αλγόριθμοι υπερτερούν έναντι άλλων και αυτοί είναι οι : Random

Forest και Naive Bayes από τον πίνακα με του αλγορίθμους στους οποίους χρησιμοποιήθηκε feature selection

Naive Bayes	0,8973	0,986	0,9395
	0,9843	0,887	0,933
Random Forest	0,8951	0,9859	0,9381
	0,9842	0,8842	0,9315

Από αυτούς τους δύο βλέπουμε ότι αυτός που μας δίνει τα καλύτερα αποτελέσματα όσον αφορά το Precision, το Recall και το F-Measure είναι ο Naive Bayes ενώ ο Random Forest μας δίνει και αυτός πολύ κοντινά αποτελέσματα , για τους λόγους του πειράματος θα συνεχίσουμε με τον Naive Bayes

Κεφάλαιο 5ο

Πίνακας Σύγκρισης Εργασιών

Σε αυτό το σημείο θα πραγματοποιηθεί μία σύγκριση ανάμεσα στην εργασία μας και σε ορισμένες εργασίες από το state of the art κομμάτι. Οι εργασίες που επιλέχθηκαν για σύγκριση είναι οι: [1],[4],[7],[9],[10],[11],[12],[15]. Έχουμε συνολικά δηλαδή 9 εργασίες μαζί με την δική μας για σύγκριση. Ο παρακάτω πίνακας μας δίνει μια πιο λεπτομερή εικόνα για την κάθε εργασία. Πιο συγκεκριμένα έχουμε:

- Τον αριθμό του εκάστοτε άρθρου στην βιβλιογραφία
- Την μεθοδολογία που χρησιμοποιείται στην κάθε εργασία
- Το μέγεθος του σετ δεδομένων
- Τον τρόπο με τον οποίο γίνεται η αξιολόγηση της κάθε εργασίας
- Τις μετρικές αξιολόγησης της κάθε εργασίας

Άρθρο	Μεθοδολογία	Σετ Δεδομένων	Τρόπος Αξιολόγησης	Μετρικές Αξιολόγησης και Αποτελέσματα Βέλτιστου Αλγορίθμου
[1]	-Πειράματα στον αλγόριθμο Random Forest -Random Forest I	-30,000,000 συναλλαγές -62 χαρακτηριστικά	-Πείραμα 1:Χωρισμός Σετ σε 70% training και 30% testing	-accuracy -precision -recall -f measure

	-Random Forest II		-Πείραμα 2: Εξισορρόπηση του σετ δεδομένων απο 1:1 μέχρι 1:10 , χωρισμός και έλεγχος του σετ για κάθε εξισορρόπηση 70% training, 30% testing -Πείραμα 3: Επιλογή 5:1 εξισορρόπησης και random sampling	-Intervention Βέλτιστος Αλγόριθμος: Random Forest II Αποτελέσματα Αλγορίθμου: -accuracy: 96.77% -precision: 89.46% -recall: 95.27% -f measure: 0.9601
[4]	-C5.0 -Support Vector Machine -Artificial Neural Network -Naive Bayes Bayesian Belief Networks -Logistic Regression -Artificial Immune Systems	-10.000.000 συναλλαγές -8 χαρακτηριστικά	-Στην πρώτη φάση ελέγχονται οι αλγόριθμοι χωρίς να εφαρμοστεί κάποια μέθοδος -Στην δεύτερη φάση εφαρμόζεται random oversampling, one-class classification και cost sensitive	-accuracy -sensitivity -Area within the precision-recall Curve Βέλτιστοι αλγόριθμοι: C5.0, SVM, ANN Αποτελέσματα αλγορίθμων: C5.0: 96%,43%,0.6 SVM: 96%,39%,0.63 ANN: 96%,47%,0.62
[7]	-Support Vector Machine -Naive Bayes -K-Nearest Neighbor -Logistic Regression	Συνδυασμός δύο σετ δεδομένων -1)917.781 συναλλαγές με 10 χαρακτηριστικά -2)200 απατηλές συναλλαγές με 8 χαρακτηριστικά	-Χωρισμός του σετ δεδομένων σε 4 υπο σετ : 1)Συναλλαγές με επικίνδυνο Merchant Category Code 2)Συναλλαγές μεγαλύτερες των 100\$ 3)Συναλλαγές με επικίνδυνο ISO Response code 4)Συναλλαγές με άγνωστες διευθύνσεις διαδικτύου -Χρήση των 4 αυτών σετ με δύο διαφορετικούς τρόπους : α)Μετατροπή των raw δεδομένων σε numerical μορφή β)Επεξεργασία των raw δεδομένων χωρίς καμία	-accuracy -precision -recall -true positive rate -false positive rate -f1-measure -ROC -MCC Στο άρθρο αναφέρονται μόνο τα αποτελέσματα του accuracy τα οποία είναι: KNN : 74% NB: 83% LR: 72% SVM 91%

			μετατροπή -Data cleaning -Data Integration -Data Transformation -Data Reduction -Resampling με τις τεχνικές SMOTE και CNN,RUS	
[9]	-Decision Trees -Logistic Regression -Random Forest -Bayes Minimum Risk	-120.000.000 συναλλαγές -27 χαρακτηριστικά	-Έχει δημιουργηθεί ένα cost matrix με βάση το οποίο βλέπουν απο ποιον αλγόριθμο χάνουν τα λιγότερα χρήματα	-Cost Matrix Βέλτιστος αλγόριθμος: cost-sensitive decision tree με >60% savings
[10]	-Neural Network	-284.807 συναλλαγές -30 χαρακτηριστικά	-Χωρισμός σετ δεδομένων σε 70% training, 15%validation και 15% testing	-accuracy -precision -recall -F1 -Matthew Coefficient Αποτελέσματα αλγορίθμου: -accuracy:94.2%, -precision:91.6% -recall:97.3% -F1:94.40% -Matthew Coefficient: 88.62%.
[11]	-Random Forest -Neural Network	-Δεν αναγράφεται	-Ακολουθείται μια διαδικασία απο modules η οποία στο τέλος δίνει ένα confusion matrix , το οποίο αξιολογείται μέσα απο μία γραφική αναπαράσταση διότι έχουμε έτσι καλύτερη ακρίβεια	-Confusion Matrix Βέλτιστος Αλγόριθμος: Random Forest Η μόνη αναφορά σε αποτελέσματα είναι στο accuracy όπου αναφέρει ότι δίνει από 90 μέχρι και 95% accuracy
[12]	-Naive Bayes -Decision Trees -Random Forest -Neural Network -Linear Regression	-Γίνεται χρήση δύο σετ δεδομένων 1)284.807 συναλλαγές με 28 χαρακτηριστικά	-Σε κάθε σετ δεδομένων χρησιμοποιούνται οι ίδιοι αλγόριθμοι εφαρμόζοντας τόσο	-accuracy -sensitivity -Matthew Coefficient score

	-Logistic Regression -Support Vector Machine -ADABOOST	2)287.224 συναλλαγές με 11 χαρακτηριστικά -Το δεύτερο σετ δεδομένων είναι δεδομένα από τον πραγματικό κόσμο	το ADABOOST όσο και το Majority Voting και τέλος ελέγχονται τα αποτελέσματα του MCC(Matthew correlation coefficient)για κάθε αλγόριθμο	Βέλτιστος Αλγόριθμος: Random Forest και Naive Bayes Η καλύτερη μέτρηση του MCC για το πρώτο σετ δεδομένων προκύπτει από το majority voting και είναι 0.823, ενώ για το δεύτερο σετ δεδομένων με θόρυβο 30% η καλύτερη μέτρηση MCC είναι 0,942
[15]	-Bayes Minimum Risk	-80.000.000 συναλλαγές με 27 χαρακτηριστικά	-Δημιουργούνται πέντε διαφορετικά υπο-σετ το καθένα με διαφορετικό ποσοστό απατών -Ελέγχονται πρώτα παραδοσιακοί αλγόριθμοι που χρησιμοποιούνται στο CCFD -Έπειτα εφαρμόζεται ο BMR -Επιλέγεται το υπο σετ με τα καλύτερα αποτελέσματα και εφαρμόζεται σε αυτό threshold optimization -Εφαρμόζεται έπειτα ο BMR με προσαρμοσμένες πιθανότητες -Τέλος συγκρίνονται όλοι οι αλγόριθμοι σε ένα εύρος από 1 μέχρι 5 ευρώ	-misclassification -recall -precision -f measure Βέλτιστος Αλγόριθμος: Random Forest Minimum Risk Δίνει μόνο τα αποτελέσματα για τις τιμές Ca =1 , Ca=2,5 και Ca = 5 στις οποίες ο αλγόριθμος έχει τα καλύτερα αποτελέσματα και είναι: RF-MR Ca = 1 :20,929 Ca = 2,5:36,634 Ca = 5: 52,003

Σύγκριση Εργασιών

Στην εργασία [1] βλέπουμε ότι δίνεται ιδιαίτερη έμφαση στον αλγόριθμο Random Forest και πιο συγκεκριμένα σε δύο μορφές του αλγορίθμου αυτού οι οποίες είναι οι: Random tree based Random Forest και Classification and Regression tree based Random Forest. Πάνω σε αυτές τις δύο μορφές πραγματοποιούνται τρία πειράματα. Πρώτα χρησιμοποιούν ένα subset του αρχικού σετ δεδομένων, το οποίο αποτελείται από όλες τις απατηλές συναλλαγές του αρχικού σετ δεδομένων και απο 150.000 νόμιμες συναλλαγές. Έπειτα χωρίζουν το subset αυτό σε δύο μέρη, εκ των οποίων το πρώτο αποτελείται απο το 70% του subset και χρησιμοποιείται για την εκπαίδευση των αλγορίθμων και το δεύτερο αποτελείται από το υπόλοιπο 30% και χρησιμοποιείται για το κομμάτι του testing. Τέλος συγκρίνουν τα αποτελέσματα των δύο αλγορίθμων για να ελέγξουν πιο από τους δύο είναι καταλληλότερος για αυτό το subset. Στο δεύτερο πείραμα χρησιμοποιείται η τεχνική random undersampling στις νόμιμες συναλλαγές με σκοπό να αντιμετωπιστεί το πρόβλημα της ισορροπίας που υπάρχει στο αρχικό σετ δεδομένων. Έπειτα εξάγονται δέκα subsets τα οποία έχουν διαφορετική αναλογία απατηλών και νόμιμων συναλλαγών από 1:1 έως 1:10. Για κάθε ένα από αυτά τα δέκα subsets χρησιμοποιείται ο ίδιος διαχωρισμός που χρησιμοποιήθηκε και στο πρώτο πείραμα, δηλαδή χωρίζονται σε σετ εκπαίδευσης, το οποίο αποτελείται απο το 70% του εκάστοτε subset και σε testing σετ το οποίο αποτελείται απο το 30% του εκάστοτε subset. Στο τρίτο πείραμα χρησιμοποιείται μόνο το δεύτερο μοντέλο του Random Forest με ένα subset του οποίου η αναλογία απατηλών και νόμιμων συναλλαγών 1:5, καθώς με βάση τα αποτελέσματα από το δεύτερο πείραμα, αυτό το subset δίνει την καλύτερη μέτρηση όσον αφορά το f measure. Το testing σετ που χρησιμοποιείται στο πείραμα 3 αποτελείται από 4,7 εκατομμύρια νόμιμες συναλλαγές και απο 87 χιλιάδες απατηλές. Οι μετρικές που χρησιμοποιούνται και στα τρία πειράματα είναι οι: accuracy, precision, recall, f measure και intervention.

Συγκρίνοντας τις δύο εργασίες παρατηρούνται τόσο ομοιότητες όσο και διαφορές. Στην δικιά μας εργασία, χωρίσαμε το αρχικό σετ δεδομένων σε 10 μικρότερα και στην συνέχεια υπολογίζουμε τον μέσο όρο των μετρικών αξιολόγησης για τον κάθε αλγόριθμο που χρησιμοποιήσαμε, ενώ σε αυτή την εργασία έχουμε διαδοχικά πειράματα τα οποία επικεντρώνονται στην συμπεριφορά ενός αλγορίθμου ανάλογα με το subset και την αναλογία που αυτό έχει. Βλέπουμε ότι και στις δύο εργασίες οι μετρικές απόδοσης των αλγορίθμων είναι σχεδόν οι ίδιες με εξαίρεση το accuracy και το intervention που χρησιμοποιούνται επιπλέον στην εργασία [1]. Τέλος, το μοντέλο που ξεχωρίζει στην εργασία [1] είναι ο Classification and Regression tree based Random Forest ενώ στην δική μου εργασία είναι ο Naive Bayes με Wrapper Subset Evaluator και μέθοδο Best First.

Η εργασία [4] χρησιμοποιεί αρκετούς αλγορίθμους προκειμένου να μας παρουσιάσει ένα πιο ολοκληρωμένο αποτέλεσμα. Το σετ δεδομένων που χρησιμοποιεί αποτελείται από 10.000.000 συναλλαγές. Προκειμένου να αξιολογήσει τους αλγόριθμους η εργασία [4] χρησιμοποιεί μια διαδικασία που αποτελείται δύο φάσεις. Στην πρώτη φάση πραγματοποιείται μια απλή σύγκριση των αλγορίθμων με βάση τις μετρικές :accuracy, recall και Area within the precision-recall Curve. Στην δεύτερη φάση οι αλγόριθμοι που έχουν επιλεχθεί χρησιμοποιούνται με σκοπό να συγκριθούν οι εξής τρόποι προσέγγισης ανισόρροπης ταξινόμησης :Random Oversampling, One-Class Classification and και Cost Sensitive

Models. Τέλος συγκρίνονται τα αποτελέσματα και επιλέγονται οι ιδανικότεροι αλγόριθμοι. Στην προκειμένη περίπτωση είναι οι αλγόριθμοι: Linear Regression, C 5.0, Decision Tree Algorithm, Support Vector Machine και ο Artificial Neural Network.

Συγκρίνοντας την εργασία [4] με την προτεινόμενη υπάρχουν αρκετές ομοιότητες και διαφορές. Αρχικά η διαδικασία προσέγγισης είναι αρκετά διαφορετική από την προτεινόμενη καθώς η εργασία [4] κοιτάει την συμπεριφορά των αλγορίθμων όταν έρχονται αντιμέτωποι με ένα ανισόρροπο πολύ σετ δεδομένων. Επιπλέον χρησιμοποιείται πληθώρα αλγορίθμων, και μεγαλύτερο σετ δεδομένων, γεγονός που απαιτεί διαφορετική προσέγγιση. Βέβαια, συναντάμε και κοινά χαρακτηριστικά όπως είναι ορισμένες μετρικές που χρησιμοποιούνται για την αξιολόγηση των αλγορίθμων καθώς και ο συνδυαστικός τρόπος αξιολόγησης της δεύτερης φάσης.

Η εργασία [7] ακολουθεί μία διαφορετική προσέγγιση όσον αφορά την μεθοδολογία που χρησιμοποιεί. Αρχικά χωρίζει το σετ δεδομένων σε τέσσερα διαφορετικά υπο σετ, τα οποία είναι: Συναλλαγές με επικίνδυνο Merchant Category Code, Συναλλαγές μεγαλύτερες των 100\$, Συναλλαγές με επικίνδυνο ISO Response code, Συναλλαγές με άγνωστε διευθύνσεις διαδικτύου. Έπειτα, τα σετ αυτά χρησιμοποιούνται με δύο διαφορετικούς τρόπους. Τα πρώτα τρία με την μέθοδο μετατροπής των raw δεδομένων σε numerical μορφή και το τελευταίο με επεξεργασία των raw δεδομένων χωρίς καμία μετατροπή. Στην συνέχεια, λόγω του υψηλού ποσοστού ανισορροπίας που υπάρχει, εφαρμόζονται τεχνικές over sampling και under sampling προκειμένου να εξισορροπηθεί. Τέλος, εφαρμόζονται οι επιλεγμένοι αλγόριθμοι σε κάθε υπο σετ δεδομένων μετά την επεξεργασία και βάση των αποτελεσμάτων τους, τα οποία κρίνονται από τις μετρικές που χρησιμοποιήθηκαν στο άρθρο και είναι οι :accuracy, precision, recall, true positive rate, false positive rate, f1-measure, Receiver Operating Curve (ROC), Matthews Correlation Coefficient (MCC), επιλέγονται οι καταλληλότεροι. Τέλος, δεν επιλέγει τον καλύτερο αλγόριθμο, αλλά ξεχωρίζει τέσσερις και τονίζει την σημασία που έχουν οι τεχνικές resampling στην απόκτηση υψηλής απόδοσης από τον ταξινόμητή.

Συγκρίνοντας τις δύο εργασίες παρατηρούμε ότι λόγω της ανισορροπίας του αρχικού σετ δεδομένων καλούμαστε να δώσουμε λύσεις. Η λύση που δόθηκε στην προκειμένη περίπτωση, είναι η εφαρμογή μεθόδων όπως της μεθόδου over sampling SMOTE και της μεθόδου CNN και RUS, προκειμένου να προκύψουν ισορροπημένα σετ δεδομένων. Επίσης χρησιμοποιήθηκαν αρκετές μετρικές αξιολόγησης, μερικές από τις οποίες χρησιμοποιούμε και εμείς στην προτεινόμενη εργασία. Τέλος όσον αφορά τα αποτελέσματα βλέπουμε ότι και στις δύο εργασίες επικρατούν αποτελέσματα τα οποία προέρχονται από ισορροπημένα σετ δεδομένων, κάτι που τονίζει και το άρθρο [7] στο κομμάτι των συμπερασμάτων.

Το άρθρο [9] δίνει περισσότερη σημασία στα ποσά της κάθε συναλλαγής. Πιο συγκεκριμένα, προκειμένου να καθορίσει αν ένας αλγόριθμος είναι αποτελεσματικός ή όχι δημιουργείται ένας πίνακας σύγκρισης βασισμένος στα ποσά που είτε χάνονται λόγω των απατών ή είναι απλές νόμιμες συναλλαγές. Ο πίνακας αυτός όταν εξεταστούν όλες οι συναλλαγές, θεωρεί η εργασία [9], ότι αντιπροσωπεύει καλύτερα τα πραγματικά κόστη διότι εξάγει αποτελέσματα με βάση το συνολικό ποσό των χρημάτων που σώθηκαν από τις απάτες. Τέλος, αναφέρει ότι

οι αλγόριθμοι που συντέλεσαν στην διάσωση μεγαλύτερου ποσού χρημάτων είναι οι cost-sensitive decision tree και cost sensitive logistic regression.

Συγκριτικά με την προτεινόμενη εργασία, το άρθρο [9] έχει μία διαφορετική προσέγγιση όσον αφορά τον τρόπο αξιολόγησης των αλγορίθμων. Ο τρόπος με τον οποίο αξιολογούνται οι αλγόριθμοι του άρθρου δεν είναι με κάποιες από τις μετρικές αξιολόγησης που χρησιμοποιούμε στην εργασία μας όπως: το recall, precision και f-measure, αλλά μέσα από έναν πίνακα σύγκρισης, μέσα από τον οποίο προκύπτει ποιος αλγόριθμος είναι ικανότερος να “σώσει” περισσότερα χρήματα και με βάση αυτού επιλέγεται ο καταλληλότερος αλγόριθμος ο οποίος στην προκειμένη είναι ο cost-sensitive decision tree, ενώ στην προτεινόμενη εργασία είναι ο Naive Bayes Wrapper Subset Evaluator και μέθοδο Best First. Το άρθρο [10] χρησιμοποιεί το ίδιο σετ δεδομένων με αυτό της προτεινόμενης εργασίας. Στο άρθρο αυτό, μετά την επεξεργασία του σετ δεδομένων με τεχνικές under sampling, δημιουργείται ένα μοντέλο νευρωνικού δικτύου. Το μοντέλο αυτό, το οποίο αποτελείται από 29 νευρώνες εισαγωγής, ένα κρυφό τμήμα με 13 νευρώνες και ένα τμήμα εξόδου με 2 νευρώνες, είναι και αυτό το οποίο αξιολογείται. Τα αποτελέσματα του πάνω στις μετρικές που χρησιμοποιούνται για αυτό είναι: accuracy 94.2%, precision 91.6%, recall 97.3%, F1 94.40% και το Matthew Coefficient 88.62%. Με ισορροπημένα δεδομένα η ακρίβεια του μοντέλου είναι πολύ κοντά στην μέτρηση F1.

Τόσο το άρθρο [10] όσο και η εργασία μας χρησιμοποιούν το ίδιο σετ δεδομένων και παρατηρούμε ότι παρουσιάζουν κοινά προβλήματα και παρόμοιες λύσεις όσον αφορά το κομμάτι της επεξεργασίας του σετ δεδομένων. Βέβαια, το άρθρο [10] δεν εξετάζει αλγορίθμους και στο τέλος του συγκρίνει, με την όποια επεξεργασία ενδιάμεσα, αλλά αντιθέτως δημιουργεί ένα δικό του μοντέλο νευρωνικού δικτύου το οποίο και αξιολογεί με σημαντικά θετικές μετρήσεις.

Περνώντας στην επόμενη σύγκριση, πρέπει να πούμε ότι δεν αναφέρεται το μέγεθος του σετ δεδομένων που χρησιμοποιείται, παρα μόνο από που προήλθε. Το άρθρο [11], όπως και το άρθρο [1], επικεντρώνεται κυρίως στην χρήση του αλγορίθμου Random Forest. Πιο συγκεκριμένα, η μεθοδολογία που ακολουθεί το άρθρο προκειμένου να αξιολογήσει τους αλγορίθμους είναι, πρώτα ο διαχωρισμός του αρχικού σετ δεδομένων σε δύο εκ των οποίων το ένα θα χρησιμοποιηθεί για training και testing. Έπειτα, αφού έχει διαχωριστεί το σετ δεδομένων εφαρμόζεται ο αλγόριθμος Random Forest έτσι ώστε να υπάρξει καλύτερη ακρίβεια σχετικά με το θέμα των απατών σε πιστωτικές κάρτες. Με την εφαρμογή του Random Forest το σετ δεδομένων θα κατηγοριοποιηθεί σε τέσσερις κατηγορίες η οποίες θα παρουσιαστούν με την μορφή ενός πίνακα σύγκρισης. Με βάση αυτό τον πίνακα σύγκρισης θα πραγματοποιηθεί η ανάλυση και τέλος μετά την ανάλυση θα αποκτηθεί η ακρίβεια των απατηλών συναλλαγών στις πιστωτικές κάρτες του σετ δεδομένων η οποία θα εξαχθεί με την μορφή μιας γραφικής αναπαράστασης. Τέλος, μας λέει ότι η ακρίβεια του Random Forest σε αυτό το άρθρο κυμαίνεται από 90% μέχρι και 95%.

Η κύρια διαφορά που παρατηρείτε ανάμεσα στην εργασία [11] και στην προτεινόμενη είναι η επιλογή των αλγορίθμων της κάθε εργασίας. Το άρθρο [11] έχει επιλέξει να επικεντρωθεί σχεδόν αποκλειστικά στον αλγόριθμο Random Forest προκειμένου να δοθεί μία λύση ενώ η προτεινόμενη καλύπτει ένα μεγαλύτερο σύνολο αλγορίθμων. Επιπλέον, στο

άρθρο αναφέρεται μόνο τα αποτελέσματα που έχει ο αλγόριθμος όσον αφορά την ακρίβεια, παραλείποντας άλλες μετρικές. Τέλος, υπάρχουν και κοινά στοιχεία όπως για παράδειγμα η διαδικασία επεξεργασίας του σετ δεδομένων προκειμένου να το φέρουμε σε μια πιο διαχειρίσιμη κατάσταση.

Το άρθρο [12] χρησιμοποιεί δύο σετ δεδομένων, το πρώτο είναι το ίδιο σετ δεδομένων που χρησιμοποιούμε κι εμείς στην προτεινόμενη εργασία και το δεύτερο είναι ένα σετ δεδομένων που περιέχει πραγματικά δεδομένα συναλλαγών. Επιπλέον χρησιμοποιεί δύο τεχνικές προκειμένου να έχει καλύτερα αποτελέσματα, την Adaptive Boost ή AdaBoost και το Majority Voting. Αρχικά, δοκιμάζουν τον αλγόριθμους στο πρώτο σετ δεδομένων χωρίς να το επεξεργαστούν, πράγμα που οδηγεί σε αποτελέσματα με υψηλό δίκτη ακρίβειας. Έπειτα, στα ίδια μοντέλα χρησιμοποιείται χρησιμοποιείται η πρώτη από τις δύο τεχνικές, η AdaBoost προκειμένου να βελτιωθούν τα αποτελέσματα. Στην συνέχεια, σε ξεχωριστό πίνακα επιλέγει τα καλύτερα μοντέλα προτού εφαρμοστεί σε αυτά το AdaBoost και εφαρμόζει την δεύτερη τεχνική. Έπειτα συγκρίνει όλα τα αποτελέσματα με τα αποτελέσματα ενός από τα άρθρα που έχει στην βιβλιογραφία διότι χρησιμοποιούν το ίδιο σετ δεδομένων, με σκοπό να ελέγξει την απόδοσή τους. Μέσα από την σύγκριση αυτή επικρατεί ο αλγόριθμος Random Forest με 95% accuracy και 91% sensitivity. Περνάει μετά στο δεύτερο σετ δεδομένων στο οποίο, όπως και στο πρώτο δημιουργεί δύο πίνακες αποτελεσμάτων, έναν με την χρήση του AdaBoost και έναν με την χρήση του Majority Voting, μόνο που αυτή την φορά δεν συγκρίνει τα αποτελέσματα με τα αποτελέσματα ενός άλλου άρθρου, αλλά συγκρίνει τους δύο πίνακες μεταξύ τους, γεγονός που οδηγεί στο συμπέρασμα ότι τα αποτελέσματα που δίνονται από την χρήση του majority voting είναι καλύτερα. Τέλος, εφαρμόζει τρία διαφορετικά ποσοστά θορύβου σε κάθε δεδομένα, τα οποία ποσοστά είναι 10%, 20% και 30%, προκειμένου να ελέγξει την αντοχή του κάθε αλγορίθμου. Η καλύτερη μέτρηση του MCC για το πρώτο σετ δεδομένων προκύπτει από το majority voting και είναι 0.823, ενώ για το δεύτερο σετ δεδομένων με θόρυβο 30% η καλύτερη μέτρηση MCC είναι 0,942, πράγμα που δείχνει ότι η τεχνική majority voting έχει σταθερή απόδοση στην παρουσία θορύβου.

Συγκριτικά, το άρθρο [12] με την προτεινόμενη εργασία έχει αρκετά κοινά στοιχεία. Αρχικά, χρησιμοποιούν και οι δύο το ίδιο σετ δεδομένων, πολλοί από του αλγορίθμους που χρησιμοποιούνται στην προτεινόμενη εργασία, συναντιούνται και στο άρθρο [12]. Για το σετ δεδομένων που έχουν κοινό, το άρθρο [12] καταλήγει στον Random Forest και Naive Bayes ως επικρατέστερους αλγορίθμους, ενώ στην προτεινόμενη ο Random Forest έρχεται δεύτερος πίσω από τον Naive Bayes με Wrapper Subset Evaluator και μέθοδο Best First. Βέβαια συναντάμε και διαφορές όπως η μέθοδος αξιολόγησης των αλγορίθμων καθώς και η παρουσία δύο διαφορετικών σετ δεδομένων έναντι του ενός που χρησιμοποιείται στην προτεινόμενη εργασία.

Το άρθρο [15], έχει ένα πολύ μεγάλο σετ δεδομένων και η μεθοδολογία η οποία χρησιμοποιεί είναι αρκετά ιδιαίτερη. Αρχικά, εφαρμόζει τεχνικές under-sampling στις νόμιμες συναλλαγές προκειμένου να εξισορροπήσει το σετ δεδομένων. Δημιουργεί πέντε διαφορετικές βάσεις δεδομένων οι οποίες έχουν αντίστοιχα 1%, 5%, 10%, 20% και 50% απατηλές συναλλαγές έτσι ώστε να ελέγξει την συμπεριφορά των αλγορίθμων με διαφορετικές κατανομές απατηλών δεδομένων. Χρησιμοποιεί την τεχνική βελτίωσης κατωφλιού στους αλγορίθμους που

εκπαιδεύονται με την βάση που έχει 50% απατηλές συναλλαγές μιας και τα καλύτερα αποτελέσματα αλγορίθμων βρίσκονται σε αυτή την βάση δεδομένων. Εφαρμόζει τον Bayes Minimum Risk ως ταξινομητή και συγκρίνει τα αποτελέσματα των αλγορίθμων, τα οποία δείχνουν ότι δεν αποδίδουν καλύτερα. Βέβαια όπως αναφέρει το άρθρο, αυτό μπορεί να οφείλεται στο γεγονός ότι οι εκτιμήσεις πιθανότητας καταρτίστηκαν με βάση τις βάσεις δεδομένων, έτσι οι αλγόριθμοι υπερεκτιμούν τις πραγματικές πιθανότητες απάτης. Στην συνέχεια προσαρμόζονται αυτές οι εκτιμώμενες πιθανότητες και συγκρίνονται τα αποτελέσματα όλων των αλγορίθμων μεταβάλλοντας την παράμετρο από 1,00 ευρώ στα 5,00 ευρώ. Καταλήγει ότι ο βέλτιστος αλγόριθμος είναι ο Random Forest Minimum Risk

Το άρθρο [15] αποτελεί και την τελευταία σύγκριση που πραγματοποιείται με την προτεινόμενη εργασία. Η μέθοδος με την οποία θα αξιολογηθούν οι αλγόριθμοι στο άρθρο [15], δεν αποτελείται μόνο από τις μετρικές αξιολόγησης αλλά και από τα οικονομικά κόστη που θα διασφαλίσουν. Σε αντιδιαστολή με την προτεινόμενη εργασία, το άρθρο [15] επικεντρώνεται στην χρήση και προσαρμογή του Bayes Minimum Risk ως ταξινομητή. Η διαδικασία αξιολόγησης είναι αρκετά πολύπλοκη και έχει αρκετά στάδια και προσαρμογές μέχρι να καταλήξει στον επικρατέστερο αλγόριθμο. Επίσης, το σετ δεδομένων που χρησιμοποιείται στο άρθρο [15] είναι σημαντικά μεγαλύτερο από αυτό που χρησιμοποιείται στην προτεινόμενη. Βέβαια υπάρχουν και ομοιότητες ανάμεσα στις δύο εργασίες, όπως είναι οι αλγόριθμοι που χρησιμοποιήθηκαν και από τις δύο εργασίες καθώς και η λογική με την οποία χώρισαν το αρχικό σετ δεδομένων έτσι ώστε να προκύψουν καλύτερα αποτελέσματα από μικρότερα υπο σετ.

Συμπεράσματα

Στόχος της παρούσας πτυχιακής εργασίας είναι η σύγκριση κλασικών αλγορίθμων που χρησιμοποιούνται στην ανίχνευση απατών στις πιστωτικές κάρτες, πάνω σε ένα αρκετά ανισόρροπο σετ δεδομένων, με σκοπό να παρατηρήσουμε την συμπεριφορά τους και μέσω μιας διαφορετικής μεθοδολογίας, να δημιουργηθεί ένα μοντέλο το οποίο ανταποκρίνεται καλύτερα σε τέτοια σετ δεδομένων, τα οποία χαρακτηρίζονται από υψηλά ποσοστά ανισορροπίας. Μέσω της μεθοδολογίας την οποία ακολουθήσαμε για την επιλογή του καλύτερου μοντέλου, παρατηρείται ότι τα καλύτερα αποτελέσματα σε μία τέτοια διαδικασία προέρχονται από ένα ισορροπημένο σετ δεδομένων και πιο συγκεκριμένα με feature selection και evaluator. Ο αλγόριθμος που καταλήξαμε, όπως έχει αναφερθεί πολλαπλές φορές κατά την διάρκεια της εργασίας είναι ο Naive Bayes με Wrapper Subset Evaluator και μέθοδο Best First. Στο πρώτο κεφάλαιο της εργασίας, πραγματοποιείται μια εισαγωγή στο πρόβλημα που προκύπτει από τις απάτες στις πιστωτικές κάρτες, την ανάγκη που υπάρχει για την αντιμετώπιση του και τέλος εξηγούμε τι εννοούμε με τον όρο credit card fraud detection και ποια η ανάγκη για την ύπαρξη συστημάτων υποστήριξης απόφασης σε αυτό το πρόβλημα. Στο επόμενο κεφάλαιο πραγματοποιούμε μια βιβλιογραφική ανασκόπηση, έτσι

ώστε να αποκτηθεί μια καλύτερη εικόνα του προβλήματος, αλλά και να δουμε διαφορετικές προσεγγίσεις και τρόπους αντιμετώπισης του προβλήματος, όπως τις προτείνουν διάφοροι επιστήμονες του κλάδου μας. Στην συνέχεια παρουσιάζεται η σύνοψη της πτυχιακής μας εργασία, η οποία αποτελεί μια προοικονομία σχετικά με το τι πρόκειται να ακολουθήσει στο πειραματικό κομμάτι. Το κεφάλαιο τρία μπορούμε να πούμε ότι είναι ο πυρήνας της έρευνας που πραγματοποιήσαμε. Σε αυτό το κεφάλαιο αναλύεται το σετ δεδομένων καθώς και οι ιδιαιτερότητες αυτού. Επίσης αναλύουμε τον κάθε αλγόριθμο ξεχωριστά και εξηγούμε πως λειτουργεί, τόσο με λόγια όσο και με γραφικές αναπαραστάσεις προκειμένου να γίνουν κατανοητοί. Στο κεφάλαιο αυτό αναλύεται και η μεθοδολογία που ακολουθήσαμε προκειμένου να φτάσουμε στα αποτελέσματα μας. Προκειμένου να πάρει την τελική της μορφή η όλη διαδικασία που μόλις εξήγησα, από την επεξεργασία του σετ δεδομένων, στους αλγόριθμους και στο feature selection και evaluator, πραγματοποιήθηκαν πολλά πειράματα, μετατροπές, διαφορετικές μέθοδοι και προσεγγίσεις, με στόχο να παραχθούν ικανοποιητικά αποτελέσματα. Συνεχίζουμε με το κεφάλαιο τέσσερα, μέσα στο οποίο έχουμε τους πίνακες με τους αλγορίθμους. Αυτό είναι και το κεφάλαιο, στο οποίο πραγματοποιούμε την σύγκριση των αποτελεσμάτων που έχουμε, έτσι ώστε να καταλήξουμε στο καταλληλότερο αλγόριθμο για την έρευνά μας. Το κεφάλαιο πέντε περιέχει μια επιλογή από εργασίες, οι οποίες κρίθηκαν λίγο πιο σημαντικές από τις υπόλοιπες που χρησιμοποιήθηκαν στο κομμάτι της βιβλιογραφικής ανασκόπησης. Αυτές τις εργασίες τις βάζουμε με την σειρά σε έναν πίνακα με λίγα λόγια για την κάθε εργασία. Έπειτα, ακολουθεί η σύγκριση των εργασιών αυτών με την δική μας πτυχιακή εργασία, πράγμα που αποσκοπεί στην παρουσίαση ομοιοτήτων και διαφορών σχετικά με την προσέγγιση του θέματος, την μεθοδολογία καθώς και τον τρόπο αξιολόγησης των αλγορίθμων ανάμεσα σε αυτές τις εργασίες με την δικιά μας. Εν κατακλείδι, παρατηρούμε ότι υπάρχουν πολλές διαφορετικές προσεγγίσεις του θέματος που ακούει στο όνομα credit card fraud detection, καθώς και χώρος για βελτίωση παλαιών και παρουσίαση νέων προσεγγύσεων έτσι ώστε να μειωθεί περετέρο αυτή η πιθανότητα απάτης.

Κεφάλαιο 6ο

Βιβλιογραφία

[1]Xuan S., Liu G., Li Z., Zheng L., Wang S., & Jiang C.,S.,*Random forest for credit card fraud detection*, 15th International Conference on Networking, Sensing and Control (ICNSC), IEEE Xplore, 2018

- [2]Mittal S., & Tyagi S. , *Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection*, 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE Xplore, 2019
- [3]Shah V., Shah P., Shetty H., & Mistry Kamal, *Review of Credit Card Fraud Detection Techniques* , International Conference on System, Computation, Automation and Networking (ICSCAN), IEEE Xplore, 2019
- [4]Makki S., Assaghir Z., Taher Y., Haque R., Hacidb M.-S., & Zeineddine H. , *An Experimental Study with Imbalanced Classification Approaches for Credit Card Fraud Detection* , IEEE Xplore, IEEE Access, 2019
- [5]Jiang C., Song J., Liu G., Zheng L., & Luan W., *Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism* ,IEEE Internet of Things Journal, IEEE Xplore, 2018
- [6]Awoyemi J. O. Adetunmbi A. O. & Oluwadare S. A, *Credit card fraud detection using machine learning techniques: A comparative analysis* , International Conference on Computing Networking and Informatics (ICCNI) , IEEE Xplore , 2017
- [7]Thennakoon A., Bhagyan C., Premadasa S., Mihiranga S., & Kuruwitaarachchi N., *Real-time Credit Card Fraud Detection Using Machine Learning.* , 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) , IEEE Xplore , 2019
- [8]Carcillo F., Le Borgne Y.-A., Caelen O., Kessaci Y., Oblé F., & Bontempi G. ,*Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection*,Information Sciences, ScienceDirect,2019
- [9]Correa Bahnsen A., Aouada D., Stojanovic A., & Ottersten B.,*Feature engineering strategies for credit card fraud detection* , Expert Systems With Applications, ScienceDirect, ,2016
- [10]Georgieva S., Markova M., & Pavlov V.,*Using neural network for credit card fraud detection* , RENEWABLE ENERGY SOURCES AND TECHNOLOGIES , AIP Conference Proceedings, Scitation , 2019
- [11] M.Suresh Kumar, V.Soundarya ,S.Kavitha , E.S.Keerthika , E.Aswini, *CREDIT CARD FRAUD DETECTION USING RANDOM FOREST ALGORITHM*,3rd International Conference on Computing And Communication Technology ICCCT, 2019

- [12]Randhawa K., Loo C. K., Seera M., Lim C. P. & Nandi A. K.,*Credit Card Fraud Detection Using AdaBoost and Majority Voting*, IEEE Xplore ,IEEE Access, 2019
- [13]Kasa N., Dahbura A., Ravoori C., & Adams S. , *Improving Credit Card Fraud Detection by Profiling and Clustering Accounts*, Systems and Information Engineering Design Symposium (SIEDS), IEEE Xplore, 2019
- [14]Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi,Fellow IEEE and Gianluca Bontempi,Senior Member IEEE ,*Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy*,IEEE Transactions on Neural Networks and Learning Systems, IEEE Xplore , 2018
- [15]Bahnsen A. C., Stojanovic A., Aouada D., & Ottersten, B., *Cost Sensitive Credit Card Fraud Detection Using Bayes Minimum Risk*, 12th International Conference on Machine Learning and Applications , IEEE Xplore , 2013
- [16]Benson Edwin Raj, S., & Annie Portia, A., *Analysis on credit card fraud detection methods*, International Conference on Computer, Communication and Electrical Technology (ICCCET) , IEEE Xplore, 2013
- [17]Malini N., & Pushpa M.,*Analysis on credit card fraud identification techniques based on KNN and outlier detection*, Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB) , IEEE Xplore, 2017
- [18]Roy A., Sun J., Mahoney R., Alonzi L., Adams S., & Beling P., *Deep learning detecting fraud in credit card transactions*, Systems and Information Engineering Design Symposium (SIEDS), IEEE Xplore , 2018

Εργαλεία και Πηγές

- Το κατεξοχήν εργαλείο που χρησιμοποιήθηκε για την επεξεργασία του σετ δεδομένων και την εξαγωγή αποτελεσμάτων είναι το WEKA: <https://www.cs.waikato.ac.nz/ml/weka/>
- Το site στο οποίο βρέθηκε το data set είναι το Kaggle :<https://www.kaggle.com/>

- Η εύρεση των άρθρων που χρησιμοποιήθηκαν έγινε μέσα απο το Google Scholar και κατ επέκταση απο τα site IEEE Xplore, ScienceDirect, Scitation :
- A. Google Scholar : <https://scholar.google.com/>
 - B. IEEE Xplore: <https://ieeexplore.ieee.org/Xplore/home.jsp>
 - C. ScienceDirect: <https://www.sciencedirect.com/>
 - D. Scitation: <https://www.scitation.org/>
- Διαφάνειες και σημειώσεις μαθήματος “Αποθήκες & Εξόρυξη δεδομένων” Ιονίου πανεπιστημίου τμήματος πληροφορικής: <https://e-class.ionio.gr/courses/DCS172/>
- Σειρά μαθημάτων σχετικά με την εφαρμογή WEKA και το data mining :
- A. https://www.youtube.com/playlist?list=PLm4W7_iX_v4NqPUjceOGd-OKNV_O4c_cPD
 - B. https://www.youtube.com/playlist?list=PLm4W7_iX_v4OMSgc8xowC2h70s-unJKCp