



ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ:

«Εντοπισμός δόλιας συναλλαγής πιστωτικής κάρτας με χρήση μηχανικής μάθησης»

Υποψήφιος: Σιώζος Θεοχάρης Παναγιώτης (Π2018093)

Επιβλέπων: Θεμιστοκλής Έξαρχος (Επίκουρος Καθηγητής)

ΚΕΡΚΥΡΑ, 2023

Περίληψη

Η πτυχιακή εργασία ασχολείται με το σημαντικό πρόβλημα του εντοπισμού δόλιων συναλλαγών πιστωτικής κάρτας με χρήση μηχανικής μάθησης. Ο σκοπός της εργασίας είναι να αναπτυχθεί ένα σύστημα που μπορεί να ανιχνεύει αυτές τις απάτες με χρήση αλγορίθμων μηχανικής μάθησης. Η εργασία προσεγγίζει το πρόβλημα αυτό με τη συλλογή και την ανάλυση μεγάλου όγκου δεδομένων συναλλαγών πιστωτικής κάρτας. Στη συνέχεια, χρησιμοποιεί διάφορες τεχνικές μηχανικής μάθησης, όπως αναγνώριση προτύπων, ταξινόμηση, και ανίχνευση ανωμαλιών, προκειμένου να εκπαιδεύσει ένα μοντέλο που μπορεί να προβλέψει τις δόλιες συναλλαγές με υψηλή ακρίβεια. Η εργασία αναλύει επίσης τα πλεονεκτήματα και τους περιορισμούς της χρήσης μηχανικής μάθησης σε αυτό το πεδίο και προτείνει πιθανές βελτιώσεις και επεκτάσεις για μελλοντικές έρευνες. Με αυτόν τον τρόπο, η εργασία αναδεικνύει τη σημασία της μηχανικής μάθησης στον εντοπισμό δόλιων συναλλαγών πιστωτικής κάρτας και τη συνεισφορά της στην ασφάλεια των χρηστών και των χρηματοπιστωτικών θεσμών.

Abstract

The thesis focuses on addressing the crucial issue of detecting fraudulent credit card transactions using machine learning techniques. The primary objective of this work is to develop a system capable of identifying such fraudulent activities by harnessing the power of machine learning algorithms. The study approaches this problem by collecting and analyzing a large volume of credit card transaction data. Subsequently, various machine learning techniques, such as pattern recognition, classification, and anomaly detection, are employed to train a model capable of predicting fraudulent transactions with high accuracy. The thesis also discusses the advantages and limitations of utilizing machine learning in this field and suggests potential enhancements and extensions for future research. In doing so, it underscores the significance of machine learning in the detection of credit card fraud and its contribution to the security of users and financial institutions.

Πίνακας περιεχομένων

Κεφάλαιο 1^ο - Πιστωτικές κάρτες & Πρόβλημα

1.Εισαγωγή.....	7
1.1 Μέθοδοι διαπραξης απατης.....	8
1.2 Για ποιον λογο η χρήση ML στο συγκεκριμένο πρόβλημα.....	15
1.3 Αναφορά και επισκόπηση της βιβλιογραφίας.....	16
1.4 .Επισκόπηση δεδομένων.....	23

Κεφάλαιο 2^ο - Επεξεργασία στο Σύνολο Δεδομένων

2. Προεπεξεργασία δεδομένων.....	25
2.1. Κατάργηση περιττού χαρακτηριστικού.....	25
2.2. Ελλειπής/διπλός χειρισμός.....	25
2.3. Μετατροπή σε ονομαστική.....	25
2.4. Αφαίρεση ακραίων τιμών.....	26
2.5. Αριθμητικός μετασχηματισμός.....	26
2.6. Χειρισμός ανισορροπιών.....	27
2.7. Τυποποίηση.....	28
2.8.Μηχανική χαρακτηριστικών.....	28

Κεφάλαιο 3^ο - Χρήση αλγορίθμων και Μεθόδων

3. Χτισσιμο Μοντελου.....	29
3.1. Διασταυρούμενη επικύρωση.....	29
3.2. Αξιολόγηση επιδόσεων.....	29
3.3. Κ Πλησιέστερος γείτονας.....	31
3.4. Διάγραμμα αποφάσεων.....	33
3.5. Τυχαίο δάσος.....	35
3.5 Υποστήριξη Vector Machine.....	37
3.6 Bagging.....	40
3.6 AdaBoostM1.....	42

Κεφάλαιο 4^ο - Συμπεράσματα, Αποτελέσματα & Σχολιασμός

4.Ανάλυση αποτελεσμάτων.....	45
------------------------------	----

4.1.Συμπέρασμα:.....	50
----------------------	----

Κεφάλαιο 5ο - Η Εφαρμογή

5. το Application της Πτυχιακής.....	52
5.1 “Choose one of the values below”.....	52
5.2 "History of past payment".....	54
5.3 "Write your answer in the fields below".....	55
5.4 "Amount of previous payments".....	56
5.5 "Amount of bill statement".....	57
5.6 “Submit και Show Results”.....	57
5.7 Ανάλυση της Εφαρμογής.....	58

Κεφάλαιο 6ο - Βιβλιογραφία

6. Πηγές - Αναφορές.....	61
--------------------------	----

Κατάλογος Σχημάτων

Εικόνα 1: Κατανομές των χαρακτηριστικών.....	25
Σχήμα 2: Εφέ φίλτρου NumericToNominal πριν (αριστερά), μετά (δεξιά) επίδραση του φίλτρου.....	27
Σχήμα 3: Επιδράσεις λογαριθμικού μετασχηματισμού πριν(αριστερά), μετά(δεξιά).....	28
Εικόνα 4: Επιδράσεις μετασχηματισμού ρίζας κύβου, πριν (πάνω), μετά (κάτω).....	34
Σχήμα 5: K Επιδόσεις πλησιέστερου γείτονα σε σύνολο δεδομένων IGAE.....	35
Σχήμα 6: Επιδόσεις δέντρου αποφάσεων σε σύνολο δεδομένων IGAE.....	37
Διάγραμμα 7: Τυχαίες δασικές επιδόσεις σε σύνολο δεδομένων IGAE.....	40
Σχήμα 8: Υποστήριξη απόδοσης διανυσματικής μηχανής συνόλου δεδομένων IGAE.....	41
Σχήμα 9: Επιδόσεις σε σάκους (RepTree) σε σχέση με το σύνολο δεδομένων IGAE.....	43
Σχήμα 10: AdaBoostM1 (Decision Stump) Performance over IGAE dataset.....	45

Λίστα πινάκων

Πίνακας 1: Περιγραφική ανάλυση των αριθμητικών χαρακτηριστικών.....	25
Πίνακας 2: Διαδικασία ρύθμισης παραμέτρων.....	27
Πίνακας 3: Απόδοση του αλγορίθμου του πλησιέστερου γείτονα K.....	28

Πίνακας 4: Ρύθμιση υπερπαραμέτρων του αλγορίθμου του δέντρου αποφάσεων.....	36
Πίνακας 5: Απόδοση του αλγορίθμου του διαγράμματος αποφάσεων.....	37
Πίνακας 6: Ρύθμιση υπερπαραμέτρων του αλγορίθμου τυχαίου δάσους.....	38
Πίνακας 7: Απόδοση αλγορίθμου τυχαίου δάσους.....	39
Πίνακας 8: Ρύθμιση υπερπαραμέτρων αλγορίθμου διανυσματικής μηχανής υποστήριξης...	40
Πίνακας 9: Απόδοση αλγορίθμου διανυσματικής μηχανής υποστήριξης.....	41
Πίνακας 10: Ρύθμιση υπερπαραμέτρων του ταξινομητή συσκευασίας.....	41
Πίνακας 11: Απόδοση της συσκευασίας με χρήση του RepTree Classifier.....	42
Πίνακας 12: Ρύθμιση υπερπαραμέτρων του ταξινομητή AdaBoostM1.....	44
Πίνακας 13: Απόδοση του AdaBoostM1 με χρήση του DecisionStump Classifier.....	45
Πίνακας 14: σταθμισμένος μέσος όρος αποτελεσμάτων από τους αλγορίθμους.....	46
Πίνακας 15: Αποτελέσματα πρόβλεψης κατηγορίας βάθους των αλγορίθμων.....	47

Κεφάλαιο 1^ο - Πιστωτικές κάρτες & Πρόβλημα

1. Εισαγωγή

Στην σημερινή εποχή χάρη στην ραγδαία τεχνολογική άνοδο των τελευταίων ετών η ανθρωπότητα έχει καταφέρει να εντάξει στις οικονομικές της συναλλαγές και την μέθοδο μέσο των πιστωτικών καρτών. Αναλυτικότερα πιστωτική κάρτα είναι μία κάρτα την οποία εκδίδουν οι τράπεζες, ώστε να μπορεί ο κόσμος να αγοράζει προϊόντα και υπηρεσίες. Είναι ένα είδος συναλλαγής που εξυπηρετεί τον καταναλωτή και τον διευκολύνει απο το να κατέχει μετρητά χρήματα ρισκάροντας να πέσει θύμα κλοπής. Είναι ομως ο συγκεκριμένος τρόπος συναλλαγής απόλυτα ασφαλής για τον πολίτη? Ενδελεχούν κίνδυνοι οι οποίοι μπορούν να διαταράξουν αυτο το χρηματοοικονομικό σύστημα? Εαν ναί μπορούν να αντιμετωπισθούν και πως? Ολα αυτα είναι εύλογα ερωτήματα τα οποία θα απαντηθούν μέσα απο την παρούσα εργασία οπως επίσης και θα αναλυθεί εις βάθος πως ο κλάδος της πληροφορικής και ειδικότερα της **μηχανικής μάθησης** έχει την δυνατότητα να συνεισφέρει στην καταπολέμηση των κινδύνων αυτου του φαινομένου.

Ιστορική αναδρομή στην αντιμετώπιση του εντοπισμού απάτης των πιστωτικών καρτών:

Το ιστορικό της ανίχνευσης απάτης μέσω πιστωτικών καρτών μπορεί να ανιχνευθεί στις πρώτες μέρες των ίδιων των πιστωτικών καρτών. Αρχικά, ο εντοπισμός απάτης μέσω πιστωτικών καρτών γινόταν κυρίως χειροκίνητα, με τις εταιρείες να εξετάζουν τις συναλλαγές για ύποπτη δραστηριότητα. Καθώς η τεχνολογία προχωρούσε, οι εταιρείες άρχισαν να χρησιμοποιούν αυτοματοποιημένα συστήματα για να επισημαίνουν πιθανές δόλιες συναλλαγές.

Στη δεκαετία του 1980, οι τεχνικές μηχανικής μάθησης εφαρμόστηκαν για πρώτη φορά στον εντοπισμό απάτης με πιστωτικές κάρτες. Αυτά τα πρώιμα συστήματα βασίστηκαν σε αλγόριθμους βασισμένους σε κανόνες, οι οποίοι ήταν σε θέση να προσδιορίσουν ορισμένα μοτίβα στις συναλλαγές που ήταν ενδεικτικά της απάτης.

Στις δεκαετίες του 1990 και του 2000, τα νευρωνικά δίκτυα και άλλες τεχνικές μηχανικής εκμάθησης χρησιμοποιούνταν όλο και περισσότερο για τον εντοπισμό απάτης με πιστωτικές κάρτες. Αυτά τα συστήματα ήταν σε θέση να μάθουν από ιστορικά δεδομένα και να προσδιορίσουν μοτίβα που δεν ήταν εύκολα αναγνωρίσιμα από συστήματα βασισμένα σε κανόνες.

Σήμερα, η ανίχνευση απάτης μέσω πιστωτικών καρτών είναι ένα εξαιρετικά εξελιγμένο πεδίο, με εταιρείες που χρησιμοποιούν έναν συνδυασμό τεχνικών όπως η μηχανική μάθηση, η τεχνητή νοημοσύνη και η ανάλυση μεγάλων δεδομένων για τον εντοπισμό και την πρόληψη της απάτης. Αυτό περιλαμβάνει όχι μόνο δεδομένα συναλλαγών αλλά και δεδομένα συμπεριφοράς, γεωγραφική τοποθεσία και δακτυλικά αποτυπώματα συσκευής.

Γενικότερα η πιθανότητα εντοπισμού απάτης μέσω πιστωτικής κάρτας εξαρτάται από διάφορους παράγοντες, συμπεριλαμβανομένου του είδους της απάτης, των μεθόδων που χρησιμοποιούνται για τη διάπραξη της απάτης και των συστημάτων εντοπισμού που υπάρχουν. Ορισμένοι τύποι απάτης με πιστωτικές κάρτες, όπως το skimming, είναι σχετικά εύκολο να εντοπιστούν επειδή περιλαμβάνουν τη φυσική τροποποίηση συσκευών ανάγνωσης καρτών. Άλλοι τύποι απάτης, όπως οι εξαγορές λογαριασμών, μπορεί να είναι πιο δύσκολο να εντοπιστούν επειδή περιλαμβάνουν τη μη εξουσιοδοτημένη χρήση ενός υπάρχοντος λογαριασμού και όχι τη δημιουργία νέου.

Οι μέθοδοι που χρησιμοποιούνται για τη διάπραξη απάτης με πιστωτικές κάρτες μπορούν επίσης να επηρεάσουν την πιθανότητα εντοπισμού της. Για παράδειγμα, οι απατεώνες που χρησιμοποιούν πολλαπλές ταυτότητες και διευθύνσεις για να ανοίξουν πολλούς λογαριασμούς μπορεί να είναι πιο πιθανό να εντοπιστούν από εκείνους που χρησιμοποιούν μία ταυτότητα. Τα ισχύοντα συστήματα ανίχνευσης διαδραματίζουν επίσης σημαντικό ρόλο στην πιθανότητα εντοπισμού απάτης με πιστωτικές κάρτες.

Τα προηγμένα συστήματα που χρησιμοποιούν μηχανική μάθηση, τεχνητή νοημοσύνη και αναλυτικά δεδομένα μεγάλων δεδομένων μπορούν να εντοπίσουν μοτίβα και ανωμαλίες που μπορεί να παραλείψουν τα απλούστερα συστήματα, καθιστώντας τα πιο αποτελεσματικά στον εντοπισμό απάτης. Συνολικά, είναι δύσκολο να πούμε ποιος τύπος απάτης πιστωτικών καρτών είναι πιο πιθανό να εντοπιστεί, καθώς εξαρτάται από τις συγκεκριμένες συνθήκες και τα ισχύοντα συστήματα ανίχνευσης. Αλλά με την πρόοδο της τεχνολογίας, τα συστήματα ανίχνευσης βελτιώνονται στον εντοπισμό και την πρόληψη της απάτης, καθιστώντας πιο δύσκολο για τους απατεώνες να περνούν απαρατήρητοι.

1.1 Μέθοδοι διάπραξης απάτης με πιστωτικές κάρτες:

Υπάρχουν πολλές διαφορετικές μέθοδοι που μπορούν να χρησιμοποιηθούν για τη διάπραξη απάτης με πιστωτικές κάρτες. Μερικές από τις πιο κοινές μεθόδους περιλαμβάνουν:

- **Skimming:** Περιλαμβάνει τη χρήση μιας μικρής συσκευής που τοποθετείται πάνω από μια συσκευή ανάγνωσης καρτών, όπως ένα ATM ή μια αντλία βενζίνης, για τη λήψη πληροφοριών πιστωτικής κάρτας όταν εισάγεται μια κάρτα.

- **Phishing:** Αυτή είναι μια μέθοδος απόκτησης πληροφοριών πιστωτικής κάρτας μέσω της χρήσης email ή ιστότοπων που φαίνονται νόμιμοι, αλλά στην πραγματικότητα έχουν σχεδιαστεί για την κλοπή προσωπικών πληροφοριών.

- **Εξαγορά λογαριασμού (Account takeover):** Αυτό συμβαίνει όταν ένας απατεώνας αποκτά πρόσβαση στον λογαριασμό πιστωτικής κάρτας ενός ατόμου και τον χρησιμοποιεί για να κάνει μη εξουσιοδοτημένες αγορές.

- **Απάτη χωρίς κάρτα (Card-not-present fraud):** Αυτός ο τύπος απάτης συμβαίνει όταν ένα άτομο κάνει μια αγορά στο διαδίκτυο ή μέσω τηλεφώνου χρησιμοποιώντας έναν κλεμμένο αριθμό πιστωτικής κάρτας.

- **Απάτη εφαρμογών (Application fraud):** Αυτό συμβαίνει όταν ένα άτομο χρησιμοποιεί ψευδή ταυτότητα ή άλλες ψευδείς πληροφορίες για να αποκτήσει πιστωτική κάρτα.

- **Πλαστή κάρτα (Counterfeit card):** Όταν ένας απατεώνας δημιουργεί μια ψεύτικη πιστωτική κάρτα χρησιμοποιώντας κλεμμένες ή κατασκευασμένες πληροφορίες.

- **Απάτη αλληλογραφίας (Mail fraud):** Δηλαδή όταν ένα άτομο παρακολουθεί εφαρμογές πιστωτικών καρτών ή δηλώσεις πιστωτικών καρτών στο ταχυδρομείο και χρησιμοποιεί τις πληροφορίες για να κάνει δόλιες αγορές.

- **Κλοπή ταυτότητας (Identity theft):** Είναι όταν ένα άτομο χρησιμοποιεί τα προσωπικά στοιχεία ενός άλλου ατόμου, συμπεριλαμβανομένων των στοιχείων πιστωτικής κάρτας, για να πραγματοποιήσει δόλιες αγορές ή να ανοίξει νέους λογαριασμούς.

Απο την όψη του θύματος:

Οι πιστωτικές και οι χρεωστικές κάρτες οφελούν πολύ τον χρήστη να μπορέσει να προβεί επιτυχημένα σε αγορές που θέλει να υλοποιήσει, σε χρηματικές του συναλλαγές όπως είναι η μεταφορά χρημάτων ή ακόμη και στην εξόφληση λογαριασμών. Επομένως συμπερνούμε πως ο εξής τρόπος έχει εγκαθεδρευθεί στις ζωές των ανθρώπων και μάλιστα λόγω όλης αυτής της

συχνής χρήσης έχει στοχοποιηθεί και απο τους δράστες απάτης των πιστωτικών καρτών. Άλλωστε με στατιστικά στοιχεία έχει αποδεχθεί πως οι απάτες πιστωτικών καρτών αποτελούν την πιο κοινή μορφή δόλιας απάτης στον εικονικό κόσμο. Σχεδόν πάντα ο σκοπός της απάτης είναι η άντληση χρημάτων απο το θύμα και η επιβεβαίωση της οικονομικής του κατάστασης. Υπάρχουν πολλές μεθοδολογίες που μπορεί να χρησιμοποιήσει ο δράστης για να θέσει σε εφαρμογή το σχέδιο του. Τηλεφωνικές υπηρεσίες, Internet, e-banking και πολλές άλλες συσκευές και λειτουργίες είναι υπεύθυνες για τις διαδικασίες πιστωτικής κάρτας του χρήστη.

Εκεί ακριβώς στοχεύουν και οι απατεώνες, στην απρόσεκτη συμπεριφορά του θύματος και ως συνέπεια την υλοποίηση των εγκληματικών ενεργειών. Οι πιο κοινές επιθέσεις με τις οποίες μπορούν να δράσουν οι απατεώνες στον κυβερνοχώρο είναι το **Phishing, Vishing, Mishng** και το **Smishing** όπου παρέχεται η δυνατότητα να συλλεχθούν λεπτομερείς πληροφορίες για την πιστωτική απάτη. Τις περισσότερες φορές το θύμα δεν έχει καν επίγνωση πως έχει πέσει θύμα απάτης εως ότου έρθει η στιγμή που θα αντικρίσει των τραπεζικό του λογαριασμό.

Απο την όψη των εταιρειών:

Αρχικά αξίζει να σημειωθεί πως την σημερινή κοινωνία απαρτίζει ένας αυξανόμενος αριθμός απο καινούριες εταιρείες οι οποίες ξεπροβάλλουν καθημερινώς και όλες έχουν ως κύριο σκοπό τους να προσφέρουν την καλύτερη ποιότητα προϊόντων και υπηρεσιών τους καταναλωτές. Για να το πετύχουν αυτο οι εταιρείες είναι απαραίτητο να **διασφαλίζουν τα δεδομένα** τα οποία με τα οποία λειτουργεί ο οργανισμός. Αυτα τα δεδομένα μπορεί να συναντηθούν σε διάφορες μορφές και αποθηκεύονται απο την εκάστοτε εταιρεία με σκοπό να επεξεργαστούν και μετέπειτα να διατηρηθεί ότι κριθεί σημαντικό. Σε ολή αυτην την διαδικασία δεν μπορεί να αγνωηθεί ενα μεγάλο ποσοστό ρίσκου που είναι η περίπτωση να κλαπούν αυτα τα δεδομένα (παραδείγματος χάριν οικονομικές πληροφορίες) και να ζημιωθει ολόκληρη η εταιρεία ή μεμονωμένα ο καταναλωτής. Η κλοπή και γενικότερα η απάτη παρουσιάζεται με πολλές μορφές και μεθοδολογίες. Μερικές εξ'αυτών είναι η πλαστογραφία επιταγών, η πώληση πλαστών αντικειμένων είτε μέσο διαδυκτίου είτε στον φυσικό κόσμο, η κλοπή ταυτότητας, οι ασφαλιστικές απάτες και φυσικά οι απάτες με πιστωτικές κάρτες. Στην περίπτωση που θα αναλυθεί παρακάτω δηλαδή την περίπτωση απάτης με πιστωτική κάρτα ουσιαστικά αναφέρεται σε μια κατάσταση που ο απατεώνας κάνει χρήση της πιστωτικής κάρτας εν αγνοία του αληθινού κατόχου σπαταλώντας τους οικονομικούς πόρους και προκαλώντας του οικονομική ζημία.

Οι συναλλαγές που πραγματοποιούνται με χρήση πιστωτικών καρτών που αποκτήθηκαν παγκοσμίως ανήλθαν σε 1,8 δισεκατομμύρια ευρώ το 2016 το κακό με την άνοδο του συγκεκριμένου τρόπου συναλλαγής είναι πως η απάτη είναι ανάλογη με το μέγεθος των συναλλαγών. Δηλαδή όσο υπάρχει τεράστια αύξηση του όγκου των συναλλαγών με πιστωτικές κάρτες τόσο θα αυξάνονται και οι πιστωτικές απάτες. Παρότι έχουν βρεθεί εξελιγμένα συστήματα ανίχνευσης της απάτης ωστόσο οι απατεώνες και υποχωρούν αν' αυτού βρίσκουν νέους τρόπους για να πετύχουν τους δόλιους σκοπούς τους. Το 2017, σημειώθηκαν 1.579 παραβιάσεις δεδομένων και σχεδόν 179 εκατομμυρίων αρχείων, μεταξύ των οποίων ήταν και απάτες με πιστωτικές κάρτες η πιο κοινή μορφή με 133.015 αναφορές, στη συνέχεια απάτες που σχετίζονται με την απασχόληση ή φορολογικές απάτες με 82.051 αναφορές, τηλεφωνικές

απάτες με 55.045 αναφορές ακολουθούμενες από τραπεζικές απάτες με 50.517 αναφορές από τα στατιστικά στοιχεία που δημοσίευσε η **FTC**.

Το εμπόριο πιστωτικών καρτών στο Dark Web

Το εμπόριο πληροφοριών πιστωτικών καρτών στον σκοτεινό ιστό (**dark web**) αναφέρεται στην αγορά και πώληση κλεμμένων πληροφοριών πιστωτικών καρτών σε underground online marketplaces. Αυτές οι αγορές είναι συνήθως προσβάσιμες μόνο μέσω της χρήσης ειδικού λογισμικού, όπως το Tor, το οποίο επιτρέπει στους χρήστες να έχουν πρόσβαση στο διαδίκτυο ανώνυμα.

Οι πληροφορίες της πιστωτικής κάρτας που πωλούνται στο σκοτεινό ιστό λαμβάνονται συχνά με διάφορα μέσα, όπως απάτες ηλεκτρονικού "ψαρέματος" (phishing), skimming και παραβιάσεις δεδομένων. Οι πληροφορίες μπορεί να περιλαμβάνουν τον αριθμό της πιστωτικής κάρτας, την ημερομηνία λήξης και τον κωδικό CVV, καθώς και άλλα προσωπικά στοιχεία, όπως το όνομα και τη διεύθυνση του κατόχου της κάρτας.

Το dark web είναι μια σημαντική πηγή πληροφοριών κλεμμένων πιστωτικών καρτών για απατεώνες, καθώς παρέχει έναν βολικό και ανώνυμο τρόπο για να αγοράσουν και να πουλήσουν τις πληροφορίες. Αυτές οι πληροφορίες μπορούν να χρησιμοποιηθούν για την πραγματοποίηση δόλιων αγορών στο διαδίκτυο ή στο κατάστημα ή για τη δημιουργία νέων λογαριασμών χρησιμοποιώντας κλεμμένες ταυτότητες.

Για την καταπολέμηση αυτού του ζητήματος, οι υπηρεσίες επιβολής του νόμου και οι ιδιωτικές εταιρείες εργάζονται για να εντοπίσουν και να κλείσουν αυτές τις αγορές του dark web και να εντοπίσουν τα άτομα που βρίσκονται πίσω από αυτές. Επιπλέον, οι εταιρείες πιστωτικών καρτών και τα χρηματοπιστωτικά ιδρύματα έχουν εφαρμόσει προηγμένα μέτρα ασφαλείας όπως **EMV chips** και **tokenization**, για να κάνουν πιο δύσκολο για τους απατεώνες να χρησιμοποιούν κλεμμένες πληροφορίες πιστωτικών καρτών.

Συνολικά, το εμπόριο πληροφοριών πιστωτικών καρτών στον σκοτεινό ιστό είναι ένα σοβαρό πρόβλημα που μπορεί να έχει σοβαρές συνέπειες τόσο για τα άτομα όσο και για τις επιχειρήσεις. Είναι σημαντικό να λαμβάνετε μέτρα για την προστασία των προσωπικών δεδομένων και οι χρήστες να είναι σε επαγρύπνηση έναντι ύποπτων δραστηριοτήτων.

Μηχανική μάθηση και η συνεισφορά της στον εντοπισμό της απάτης:

Η ανάλυση μηχανικής μάθησης αναφέρεται στη διαδικασία χρήσης αλγορίθμων και στατιστικών μοντέλων για την αυτόματη αναγνώριση προτύπων, τάσεων και σχέσεων στα δεδομένα. Αυτό μπορεί να χρησιμοποιηθεί σε ένα ευρύ φάσμα εφαρμογών, συμπεριλαμβανομένης της ανίχνευσης απάτης, της αναγνώρισης εικόνων, της επεξεργασίας φυσικής γλώσσας και της μοντελοποίησης πρόβλεψης.

Οι συγκεκριμένες τεχνικές που χρησιμοποιούνται στην ανάλυση μηχανικής μάθησης μπορεί να ποικίλλουν ανάλογα με τον τύπο των δεδομένων που αναλύονται και τους στόχους της ανάλυσης. Ορισμένες κοινές τεχνικές περιλαμβάνουν την εποπτευόμενη μάθηση (**supervised**

learning), την μάθηση χωρίς επίβλεψη (**unsupervised learning**) και την ενισχυτική μάθηση (**reinforcement learning**).

- Η εποπτευόμενη μάθηση (supervised learning) είναι ένας τύπος μηχανικής μάθησης όπου ένας αλγόριθμος εκπαιδεύεται σε ένα επισημασμένο σύνολο δεδομένων, πράγμα που σημαίνει ότι η σωστή έξοδος είναι ήδη γνωστή. Αυτή η διαδικασία επιτρέπει στον αλγόριθμο να μάθει από τα δεδομένα και να κάνει προβλέψεις για νέα, αόρατα δεδομένα.

- Η μάθηση χωρίς επίβλεψη (unsupervised learning) είναι ένας τύπος μηχανικής μάθησης που εντοπίζει υποκείμενα μοτίβα ή ομαδοποιήσεις σε δεδομένα χωρίς τη χρήση δεδομένων με ετικέτα. Οι κοινές τεχνικές περιλαμβάνουν τη ομαδοποίηση και τη μείωση διαστάσεων.

- Η ενισχυτική μάθηση (reinforcement learning) είναι ένας τύπος μηχανικής μάθησης όπου μαθαίνει να λαμβάνει αποφάσεις αλληλεπιδρώντας με ένα περιβάλλον και λαμβάνοντας ανταμοιβές ή ποινές για ορισμένες ενέργειες.

Τα μοντέλα μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για την ανάλυση μεγάλων ποσοτήτων δεδομένων γρήγορα και με ακρίβεια, καθιστώντας τα ένα ισχυρό εργαλείο για τον εντοπισμό προτύπων και τον εντοπισμό ανωμαλιών που μπορεί να υποδηλώνουν απάτη.

Η μηχανική εκμάθηση μπορεί να βοηθήσει στον εντοπισμό απάτης με πιστωτικές κάρτες αναλύοντας μοτίβα στα δεδομένα συναλλαγών. Αυτό μπορεί να περιλαμβάνει τον εντοπισμό ασυνήθιστων ή ύποπτων προτύπων συμπεριφοράς, όπως μεγάλους αριθμούς μικρών συναλλαγών ή συναλλαγών που πραγματοποιούνται εκτός των κανονικών συνηθειών δαπανών του κατόχου της κάρτας. Επιπλέον, τα μοντέλα μηχανικής εκμάθησης μπορούν να εκπαιδευτούν σε ιστορικά δεδομένα για τον εντοπισμό μοτίβων που είναι ενδεικτικά της απάτης και, στη συνέχεια, να χρησιμοποιηθούν για τον αυτόματα εντοπισμό και επισήμανση πιθανών δόλιων συναλλαγών σε πραγματικό χρόνο. Αυτό μπορεί να βοηθήσει τα χρηματοπιστωτικά ιδρύματα να εντοπίζουν γρήγορα και να ανταποκρίνονται σε δόλιες δραστηριότητες, ελαχιστοποιώντας τις πιθανές απώλειες που σχετίζονται με την απάτη με πιστωτικές κάρτες.

Αναλυτικότερα περί της Μάστιγας της απάτης:

Στην παρακάτω παράγραφο θα αναλύσουμε την αποφυγή της πιστωτικής και χρεωστικής απάτης. Η αποφυγή της απάτης χωρίζεται σε δύο κατηγορίες στην **πρόληψη** και στην **ανίχνευση**. Η πρόληψη ουσιαστικά είναι η αποφυγή τυχόν επιθέσεων από απατεώνες λειτουργώντας ως ένα στρώμα προστασίας ενώ αντίθετα η ανίχνευση συμβαίνει όταν η πρόληψη έχει ήδη αποτύχει. Δηλαδή η ανίχνευση βοηθά στον εντοπισμό και την ειδοποίηση αφού έχει υλοποιηθεί η δόλια συναλλαγή.

Η απάτη με πιστωτική κάρτα κατηγοριοποιείται στις εξής δύο περιπτώσεις:

Στην **απάτη με κάρτα χωρίς Παρουσία (CNP)** και στην **απάτη με παρουσία της κάρτας (CP)**. Η μία είναι με την κλοπή της φυσικής κάρτας και η άλλη είναι εικονικά. Στη πρώτη περίπτωση ο δράστης κλέβει την πιστωτική κάρτα του θύματος ενώ στην δεύτερη κλέβει ευαίσθητες πληροφορίες όπως τον αριθμό της κάρτας, τον κωδικό cvv, τον τύπο της κάρτας κ.α. Με την κλοπή των στοιχείων της πιστωτικής κάρτας ένας απατεώνας μπορεί να αποσπάσει μεγάλο χρηματικό ποσό ή να το αγοράσει αγαθά χρεώνοντας τον κάτοχο της κάρτας. Οι απάτες με πιστωτικές κάρτες είναι ένας εύκολος στόχος, χωρίς κινδύνους και ένα σημαντικό χρηματικό ποσό μπορεί να αφαιρεθεί εν αγνοία του ιδιοκτήτη σε σύντομο χρονικό διάστημα. Οι απατεώνες πάντα προσπαθούν να κάνουν νόμιμη κάθε δόλια συναλλαγή η οποία και καθιστά τον εντοπισμό της απάτης ακόμα πιο δύσκολη.

Επιπλέον το σύνολο των δεδομένων των πιστωτικών καρτών είναι **εξαιρετικά ανισορροπημένο** αυτό συμβαίνει διότι υπάρχουν περισσότερες νόμιμες συναλλαγές συγκριτικά με τις δόλιες και έτσι συμπερασματικά είναι ακόμη πιο δύσκολο να εντοπιστεί η επικίνδυνη συναλλαγή. Οι τεχνικές machine learning εμπίπτουν στις κατηγορίες μάθησης χωρίς επίβλεψη και στις κατηγορίες μάθησης με επίβλεψη το ποιά θα χρησιμοποιηθεί κρίνεται αναλόγως το data set. Στην επίβλεψη με μάθηση κύριο χαρακτηριστικό είναι η ταξινόμηση των ανωμαλιών. Οι περισσότεροι machine learning algorithms παίρνουν πολύ χρόνο για training σκοπό παρά για predicting (πρόβλεψη). Ακόμη ένα βασικό ζήτημα που επηρεάζει τον εντοπισμό της οικονομικής απάτης είναι το **future selection** δηλαδή το χαρακτηριστικό της επιλογής. Στόχος του είναι να φιλτράρει τα χαρακτηριστικά που τα περισσότερα περιγράφουν τις πτυχές του εντοπισμού και των χαρακτήρων. Κατά την εξέταση ενός συστήματος το κόστος της δόλιας συμπεριφοράς και το κόστος πρόληψης λαμβάνονται υπόψη.

Η ραγδαία ανάπτυξη που έχει ο κόσμος στη σημερινή εποχή και η εξελισσόμενη χρηματοπιστωτική βιομηχανία έχουν οδηγήσει στη χαλάρωση του καταναλωτή. Ειδικότερα την περίοδο της παγκόσμιας επιδημικής κρίσης covid 19 κατά την οποία πολλές χρηματοπιστωτικές συναλλαγές τελέστηκαν σε διαδικτυακές πλατφόρμες. Ως αποτέλεσμα έχει την έξαρση σε πολλά οικονομικά εγκλήματα όπως είναι η απάτη πιστωτικής και χρεωστικής κάρτας που είχαν σημαντική άνοδο αυτό το διάστημα. Το πρόβλημα αυτό έχει ανησυχήσει πολλές κυβερνήσεις και χρηματοπιστωτικά ιδρύματα όχι λόγω των νομισματικών απωλειών που θα φαινομενικά αποτελεί την πιο ζημειωγόνα συνέπεια αλλά διότι η ύπαρξη αυτού του φαινομένου βλάπτει σοβαρά τη φήμη του έθνους ή των αντίστοιχων οργανισμών και επιπλέον μπορεί να έχει ως συνέπεια την ενθάρρυνση άλλων δραστών να προβούν σε αντίστοιχες ενέργειες.

Οι περισσότερες τράπεζες και χρηματοπιστωτικά ιδρύματα χρησιμοποιούν βάσει κανονισμών συστήματα τα οποία κάποιος εξειδικευμένος χρησιμοποιεί παλαιότερα δεδομένα απάτης με σκοπό να δημιουργήσει ένα σύνολο κανόνων προβλέψεις στο σύστημα το οποίο έπειτα με την σειρά του θα σημάνει συναγερμό εάν μία συναλλαγή ταιριάζει με αυτούς τους κανόνες. Δηλαδή ουσιαστικά μέσω των παλαιότερων παραβιάσεων απομνημονεύονται τα χαρακτηριστικά και εάν τα χαρακτηριστικά εντοπιστούν ξανά ενεργοποιούνται τα **red flags**. Οι Κανόνες αυτοί όμως έχουν το ελάττωμα πώς εξαιτίας της χειροκίνητης διαδικασίας που τελείται για την ανάπτυξη ενός τέτοιου συστήματος είναι αντιδραστικοί, στερούνται ευδεξίας και συνέπειας καθώς όλη διαδικασία υλοποίησης είναι εξαιρετικά χρονοβόρα. Απο την άλλη μεριά η μηχανική μάθηση (machine learning) επιτρέπει στους υπολογιστές να μαθαίνουν χωρίς

να είναι ρητά προγραμματισμένοι για τον εντοπισμό στη συγκεκριμένη περίπτωση απάτης πιστωτικών καρτών. Επομένως η μηχανική μάθηση χρησιμοποιείται για τον καλύτερο και τάχυστο εντοπισμό στις δόλιες συναλλαγές. Ωστόσο όμως η ανίχνευση μπορεί να αποτελέσει πρόκληση για τη μηχανική μάθηση εξαιτίας διάφορων λόγων όπως είναι η διανομή των δεδομένων που είναι **εξαιρετικά ανισορροπή**. Ο ρυθμός των δόλιων συναλλαγών είναι πολύ μικρός σε σχέση με των νόμιμων συναλλαγών και τα δεδομένα εξελίσσονται συνεχώς με την πάροδο του χρόνου. Επίσης η έλλειψη δεδομένων είναι μεγάλη διότι εμπλέκεται και το απόρρητο της διαδικασίας. Σαν απο μηχανής θεός οι λύσεις που παρουσιάστηκαν για να ξεπεραστούν αυτές οι προκλήσεις ήταν η ύπαρξη υβριδικών μοντέλων λύση για την οποία έχουν σχηματιστεί πολλές έρευνες ώστε να χρησιμοποιηθούν υβριδικά μοντέλα μηχανικής μάθησης για τον εντοπισμό της πιστωτικής απάτης και την αντιμετώπιση του φαινομένου.

Παρουσίαση σχετικών μοντέλων μηχανικής μάθησης για τον εντοπισμό απάτης με πιστωτική κάρτα:

Υπάρχουν πολλά μοντέλα μηχανικής εκμάθησης που μπορούν να χρησιμοποιηθούν για τον εντοπισμό απάτης με πιστωτικές κάρτες, το καθένα με τα δικά του δυνατά και αδύνατα σημεία. Μερικά δημοφιλή μοντέλα περιλαμβάνουν:

- **Ανίχνευση ανωμαλιών (Anomaly Detection):** Αυτό το μοντέλο χρησιμοποιεί τεχνικές εκμάθησης χωρίς επίβλεψη για να εντοπίσει ασυνήθιστα μοτίβα στα δεδομένα συναλλαγών που μπορεί να υποδηλώνουν απάτη. Αυτό μπορεί να περιλαμβάνει τον εντοπισμό συναλλαγών που πραγματοποιούνται εκτός των συνηθισμένων συνηθειών δαπανών του κατόχου της κάρτας ή που έχουν ασυνήθιστα ποσά ή συχνότητες.

- **Random Forest:** Τα μοντέλα Random Forest είναι ένα σύνολο δέντρων αποφάσεων. Αυτά τα δέντρα μπορούν να χρησιμοποιηθούν για την ταξινόμηση των δεδομένων συναλλαγών ως δόλιων ή μη. Μπορούν επίσης να χρησιμοποιηθούν για τη δημιουργία γραφήματος σπουδαιότητας χαρακτηριστικών, το οποίο μπορεί να δείξει ποια χαρακτηριστικά είναι πιο σημαντικά για τον εντοπισμό απάτης.

- **Νευρωνικά δίκτυα (Neural Networks):** Τα νευρωνικά δίκτυα είναι ένας τύπος μοντέλου βαθιάς μάθησης που μπορεί να χρησιμοποιηθεί για τον εντοπισμό πολύπλοκων μοτίβων στα δεδομένα συναλλαγών. Μπορούν να χρησιμοποιηθούν για την ταξινόμηση των συναλλαγών ως δόλιες ή μη, και μπορούν επίσης να χρησιμοποιηθούν για τον προσδιορισμό του είδους της απάτης που συμβαίνει.

- **Λογιστική παλινδρόμηση (Logistic Regression):** Τα μοντέλα λογιστικής παλινδρόμησης είναι ένας τύπος εποπτευόμενου μοντέλου μάθησης που μπορεί να χρησιμοποιηθεί για την πρόβλεψη της πιθανότητας μιας συναλλαγής να είναι δόλια. Μπορούν να χρησιμοποιηθούν για την επισήμανση συναλλαγών που έχουν μεγάλη πιθανότητα απάτης.

- **Υποστήριξη διανυσματικών μηχανών (SVM-Support Vector Machines):** Τα SVM είναι ένας τύπος εποπτευόμενου μοντέλου εκμάθησης που έχει την δυνατότητα να χρησιμοποιηθεί για την ταξινόμηση των δεδομένων συναλλαγών ως δόλιων ή μη. Λειτουργούν βρίσκοντας ένα όριο ή «υπερπλάνο» που διαχωρίζει τις δόλιες συναλλαγές από τις μη δόλιες συναλλαγές.

Αξίζει να σημειωθεί ότι η επιλογή του μοντέλου και το σύνολο δεδομένων που χρησιμοποιείται είναι πολύ σημαντικά για την απόδοση του μοντέλου. Επίσης, τα μοντέλα πρέπει να βελτιωθούν για να αποδώσουν βέλτιστα και επίσης πρέπει να ενημερώνονται τακτικά για να προσαρμόζονται σε νέους τύπους απάτης.

1.2 Γιατί να χρησιμοποιήσω μηχανική μάθηση ως λύση στο συγκεκριμένο πρόβλημα?

Ο εντοπισμός δόλιας συναλλαγής πιστωτικών καρτών αποτελεί ένα σημαντικό πρόβλημα για τον τραπεζικό και τον χρηματοπιστωτικό τομέα, καθώς και για τους κατόχους πιστωτικών καρτών. Η χρήση μηχανικής μάθησης για την ανίχνευση δόλιων συναλλαγών έχει αναδειχθεί ως μια αποτελεσματική και αξιόπιστη προσέγγιση, και υπάρχουν πολλοί λόγοι που καθιστούν αυτήν την μέθοδο προτιμητέα έναντι άλλων προσεγγίσεων.

Προβλεψιμότητα και Ακρίβεια

Η μηχανική μάθηση χρησιμοποιεί αλγόριθμους που μπορούν να αναγνωρίσουν αρνητικά πρότυπα και τάσεις που είναι δύσκολο να ανιχνευθούν από τον ανθρώπινο εγκέφαλο. Αυτό σημαίνει ότι η μηχανική μάθηση μπορεί να εντοπίσει ανομαλίες και ανοιχτές συναλλαγές με υψηλό βαθμό ακρίβειας.

Εξατομίκευση και Προσαρμογή

Κάθε χρήστης πιστωτικής κάρτας έχει μοναδικές συνήθειες και συμπεριφορές στις συναλλαγές του. Η μηχανική μάθηση επιτρέπει την προσαρμογή του μοντέλου ανίχνευσης δόλιων συναλλαγών στον κάθε χρήστη. Αυτό σημαίνει ότι το μοντέλο μπορεί να λαμβάνει υπόψη τις μοναδικές συνήθειες και να προσαρμόζει τις αποφάσεις του στη συγκεκριμένη κατάσταση του χρήστη.

Ταχύτητα και Αυτοματοποίηση

Η μηχανική μάθηση επιτρέπει την ταχεία επεξεργασία και ανάλυση μεγάλων όγκων δεδομένων, κάτι που είναι αδύνατο να επιτευχθεί με τον παραδοσιακό ανθρώπινο τρόπο. Αυτό επιτρέπει την αυτοματοποίηση της διαδικασίας εντοπισμού δόλιων συναλλαγών, μειώνοντας τον χρόνο που απαιτείται για την αξιολόγηση κάθε συναλλαγής.

Αντιμετώπιση Αναδράσεων

Το μοντέλο μηχανικής μάθησης μπορεί να βελτιώσει την απόδοση του με την πάροδο του χρόνου, μαθαίνοντας από τις νέες δεδομένες. Αυτό του επιτρέπει να αντιμετωπίζει αποτελεσματικά τις αναδράσεις και να προσαρμόζεται στην εξέλιξη των τεχνικών που χρησιμοποιούν οι απατεώνες.

Συνεχής Εξέλιξη των Τεχνικών Απάτης

Οι απατεώνες συνεχώς εξελίσσουν τις τεχνικές τους για να αποφύγουν τον εντοπισμό. Η μηχανική μάθηση επιτρέπει την αναγνώριση νέων μοτίβων απάτης με πολύ υψηλή ευαισθησία.

Αντιμετώπιση Μεγάλων Ογκων Δεδομένων

Οι τράπεζες και οι χρηματοπιστωτικές εταιρείες διαχειρίζονται τεράστιους όγκους συναλλαγών καθημερινά. Η μηχανική μάθηση είναι η μόνη αποτελεσματική λύση για την ανάλυση και αξιολόγηση αυτών των μεγάλων συνόλων δεδομένων.

Ανίχνευση Κρυφών Σχέσεων

Η μηχανική μάθηση μπορεί να ανιχνεύσει κρυφές σχέσεις μεταξύ δεδομένων που δεν είναι εμφανείς στον ανθρώπινο εξεταστή. Αυτό μπορεί να αποτελέσει κλειδί για τον εντοπισμό της απάτης.

Συνεχής Εκπαίδευση

Τα μοντέλα μηχανικής μάθησης μπορούν να εκπαιδεύονται συνεχώς με νέα δεδομένα και να βελτιώνονται με τον χρόνο. Αυτό κάνει την προσέγγιση ανθεκτική στις αλλαγές και την εξέλιξη των τεχνικών απάτης.

Συνολικά, η χρήση μηχανικής μάθησης για τον εντοπισμό δόλιων συναλλαγών πιστωτικών καρτών προσφέρει ανωτερότητα σε θέματα απόδοσης, ακρίβειας, και αντιμετώπισης αναδράσεων, καθιστώντας τη μία από τις πιο αποτελεσματικές λύσεις για το πρόβλημα της απάτης στις πιστωτικές κάρτες.

1.3 Αναφορά και επισκόπηση της βιβλιογραφίας.

[1] Εισαγωγή:

Στην έρευνα της βιβλιογραφίας με τίτλο "Credit Card Fraud Detection - Machine Learning methods" που συντάχθηκε από τους Dejan Varmedja, Mirjana Karanovic, Srdjan Sladojevic,

Marko Arsenovic, Andras Anderla συζητά τη χρήση μεθόδων μηχανικής μάθησης για τον εντοπισμό απάτης με πιστωτικές κάρτες.

Η απάτη με πιστωτικές κάρτες είναι ένα σημαντικό πρόβλημα για τον χρηματοπιστωτικό κλάδο, προκαλώντας ζημίες δισεκατομμυρίων δολαρίων κάθε χρόνο. Οι μέθοδοι μηχανικής μάθησης γίνονται ολοένα και πιο δημοφιλείς για τον εντοπισμό απάτης με πιστωτικές κάρτες λόγω της ικανότητάς τους να αναλύουν μεγάλο όγκο δεδομένων και να εντοπίζουν μοτίβα που μπορεί να είναι δύσκολο να εντοπιστούν χρησιμοποιώντας παραδοσιακές μεθόδους. Αυτό το άρθρο παρουσιάζει μια επισκόπηση των διαφορετικών τεχνικών μηχανικής εκμάθησης που χρησιμοποιούνται στον εντοπισμό απάτης με πιστωτικές κάρτες.

Μέθοδοι:

Η εργασία εξετάζει διάφορες τεχνικές μηχανικής μάθησης που χρησιμοποιούνται για τον εντοπισμό απάτης με πιστωτικές κάρτες, συμπεριλαμβανομένων των δέντρων αποφάσεων, της λογιστικής παλινδρόμησης, των μηχανών διανυσμάτων υποστήριξης, των τεχνητών νευρωνικών δικτύων και των μεθόδων συνόλου. Οι συγγραφείς επισημαίνουν τα πλεονεκτήματα και τους περιορισμούς κάθε μεθόδου και παρέχουν παραδείγματα των εφαρμογών τους στον εντοπισμό απάτης με πιστωτικές κάρτες. Οι συγγραφείς συζητούν επίσης τη σημασία της επιλογής χαρακτηριστικών και των τεχνικών προεπεξεργασίας για τη βελτίωση της ακρίβειας των μοντέλων ανίχνευσης απάτης. Η επιλογή χαρακτηριστικών βοηθά στον εντοπισμό σχετικών χαρακτηριστικών που συμβάλλουν στον εντοπισμό δόλιων συναλλαγών, ενώ τεχνικές προεπεξεργασίας όπως η κανονικοποίηση και η κλιμάκωση βοηθούν στη μείωση του αντίκτυπου των ακραίων τιμών και στη βελτίωση της απόδοσης των μοντέλων μηχανικής εκμάθησης.

Αποτελέσματα:

Η εργασία παρουσιάζει τα αποτελέσματα αρκετών μελετών που καταδεικνύουν την αποτελεσματικότητα των μεθόδων μηχανικής εκμάθησης στον εντοπισμό απάτης με πιστωτικές κάρτες. Οι συγγραφείς συγκρίνουν την απόδοση διαφορετικών αλγορίθμων μηχανικής μάθησης σε ένα σύνολο δεδομένων συναλλαγών με πιστωτικές κάρτες και αποδεικνύουν ότι οι μέθοδοι συνόλου όπως το Random Forest και η Gradient Boosting αποδίδουν καλύτερα από άλλες μεθόδους. Οι συγγραφείς συζητούν επίσης τις προκλήσεις που σχετίζονται με τον εντοπισμό απάτης με πιστωτικές κάρτες και υπογραμμίζουν την ανάγκη για συνεχή βελτίωση και βελτίωση των μοντέλων μηχανικής μάθησης για να συμβαδίζουν με τα εξελισσόμενα μοτίβα απάτης.

Συμπέρασμα:

Συμπερασματικά, η εργασία παρέχει μια ολοκληρωμένη επισκόπηση των τεχνικών μηχανικής εκμάθησης που χρησιμοποιούνται για τον εντοπισμό απάτης μέσω πιστωτικών καρτών. Οι συγγραφείς καταδεικνύουν την αποτελεσματικότητα διαφόρων μεθόδων και υπογραμμίζουν

τη σημασία των τεχνικών επιλογής χαρακτηριστικών και προεπεξεργασίας για τη βελτίωση της ακρίβειας των μοντέλων ανίχνευσης απάτης.

Το έγγραφο υπογραμμίζει την ανάγκη για συνεχή βελτίωση και τελειοποίηση των μοντέλων μηχανικής μάθησης για να συμβαδίζουν με τα εξελισσόμενα μοτίβα απάτης. Οι πληροφορίες που παρέχονται από αυτό το έγγραφο μπορεί να είναι χρήσιμες για χρηματοπιστωτικά ιδρύματα και άλλους οργανισμούς που εμπλέκονται στον εντοπισμό και την πρόληψη απάτης για την ανάπτυξη αποτελεσματικών μοντέλων ανίχνευσης απάτης χρησιμοποιώντας τεχνικές μηχανικής μάθησης. Οι συγγραφείς χρησιμοποιούν ένα σύνολο δεδομένων από το Kaggle, το οποίο περιέχει συναλλαγές με πιστωτική κάρτα που χαρακτηρίζονται ως δόλιες ή μη. Οι συγγραφείς συγκρίνουν διάφορους αλγόριθμους μηχανικής μάθησης, συμπεριλαμβανομένης της λογιστικής παλινδρόμησης, των δέντρων αποφάσεων και των νευρωνικών δικτύων, για να προσδιορίσουν την πιο αποτελεσματική μέθοδο για τον εντοπισμό απάτης. Διαπιστώνουν ότι οι αλγόριθμοι Random Forest και gradient boosting έχουν την καλύτερη απόδοση, με ποσοστό ακρίβειας άνω του 99%.

Αυτή η ερευνητική εργασία από τους Anuruddha Thennakoon, Chee Bhagyan, Sasitha Premadasa, Shalitha Mihiranga, Nuwan Kuruwitaarachchi περιγράφει ένα σύστημα ανίχνευσης απάτης με πιστωτικές κάρτες σε πραγματικό χρόνο (Real Time) χρησιμοποιώντας αλγόριθμους μηχανικής μάθησης.

[2] Εισαγωγή:

Η εργασία των Thennakoon, A., Bhagyan, C., Premadasa, S., Mihiranga, S., & Kuruwitaarachchi, N. (2019). Real-time Credit Card Fraud Detection Using Machine Learning. Για το 9th International Conference on Cloud Computing, Data Science & Engineering παρουσιάζει ένα σύστημα ανίχνευσης απάτης σε πραγματικό χρόνο που βασίζεται σε μηχανική μάθηση για συναλλαγές με πιστωτικές κάρτες.

Μέθοδοι:

Το έγγραφο προτείνει ένα σύστημα ανίχνευσης απάτης σε πραγματικό χρόνο που χρησιμοποιεί μια προσέγγιση που βασίζεται στη μηχανική μάθηση για τον εντοπισμό δόλιων συναλλαγών. Οι ερευνητές χρησιμοποίησαν ένα σύνολο δεδομένων συναλλαγών με πιστωτικές κάρτες για να εκπαιδεύσουν και να αξιολογήσουν το μοντέλο τους. Χρησιμοποίησαν διαφορετικούς αλγόριθμους μηχανικής μάθησης, συμπεριλαμβανομένων των δέντρων αποφάσεων, του k-πλησιέστερου γείτονα και της λογιστικής παλινδρόμησης, για να αξιολογήσουν την απόδοση του συστήματός τους.

Οι ερευνητές πρότειναν επίσης μια τεχνική επιλογής χαρακτηριστικών και προεπεξεργασίας για τη μείωση της διάστασης του συνόλου δεδομένων και τη βελτίωση της απόδοσης του μοντέλου. Χρησιμοποίησαν ένα τεστ χ-τετράγωνο για να προσδιορίσουν τα πιο σημαντικά χαρακτηριστικά και εφάρμοσαν κανονικοποίηση για να κλιμακώσουν τα δεδομένα.

Αποτελέσματα:

Οι ερευνητές απέδειξαν την αποτελεσματικότητα του συστήματος ανίχνευσης απάτης σε πραγματικό χρόνο αξιολογώντας την απόδοσή του σε ένα πραγματικό σύνολο δεδομένων συναλλαγών με πιστωτικές κάρτες. Το σύστημα πέτυχε ακρίβεια 98,7% και ψευδώς θετικό ποσοστό 0,6%. Οι συγγραφείς συνέκριναν επίσης την απόδοση του συστήματός τους με τα παραδοσιακά συστήματα ανίχνευσης απάτης που βασίζονται σε κανόνες και έδειξαν ότι το σύστημά τους που βασίζεται σε μηχανική μάθηση υπερτερεί των συστημάτων που βασίζονται σε κανόνες.

Οι ερευνητές συζήτησαν επίσης τη φύση του συστήματός τους σε πραγματικό χρόνο και την ικανότητά του να ανιχνεύει απάτη σε πραγματικό χρόνο, μειώνοντας τον κίνδυνο δόλιων συναλλαγών.

Συμπέρασμα:

Συμπερασματικά, η εργασία "Real-time Credit Card Fraud Detection Using Machine Learning" προτείνει ένα σύστημα ανίχνευσης απάτης σε πραγματικό χρόνο που χρησιμοποιεί αλγόριθμους μηχανικής μάθησης για τον εντοπισμό δόλιων συναλλαγών με πιστωτικές κάρτες. Οι συγγραφείς απέδειξαν την αποτελεσματικότητα του συστήματός τους επιτυγχάνοντας υψηλή ακρίβεια και χαμηλά ποσοστά ψευδώς θετικών. Το έγγραφο υπογραμμίζει τη σημασία των τεχνικών επιλογής χαρακτηριστικών και προεπεξεργασίας για τη βελτίωση της απόδοσης των συστημάτων ανίχνευσης απάτης που βασίζονται στη μηχανική μάθηση.

Η φύση του συστήματος σε πραγματικό χρόνο μειώνει επίσης τον κίνδυνο δόλιας συναλλαγής. Οι πληροφορίες που παρέχονται από αυτό το έγγραφο μπορεί να είναι χρήσιμες για χρηματοπιστωτικά ιδρύματα και άλλους οργανισμούς που εμπλέκονται στον εντοπισμό και την πρόληψη απάτης για την ανάπτυξη αποτελεσματικών συστημάτων ανίχνευσης απάτης σε πραγματικό χρόνο χρησιμοποιώντας τεχνικές μηχανικής μάθησης. Η εργασία θέτει τα θεμέλια για μελλοντική έρευνα στον τομέα του εντοπισμού και της πρόληψης απάτης με πιστωτικές κάρτες.

Οι ερευνητές προτείνουν ένα σύστημα που χρησιμοποιεί έναν αλγόριθμο ανίχνευσης ανωμαλιών για τον εντοπισμό δόλιων συναλλαγών με βάση διάφορα χαρακτηριστικά όπως το ποσό, η τοποθεσία και η ώρα της συναλλαγής. Οι ερευνητές συγκρίνουν διάφορους αλγόριθμους μηχανικής μάθησης όπως για παράδειγμα των δέντρων αποφάσεων, των k-πλησιέστερων γειτόνων και των μηχανών υποστήριξης διανυσμάτων (support vector machines) με σκοπό να βρουν την πιο αποτελεσματική μέθοδο για τον εντοπισμό απάτης με πιστωτική κάρτα. Καταλήγουν ότι ο αλγόριθμος του δέντρου αποφάσεων έχει την καλύτερη απόδοση, με ποσοστό ακρίβειας άνω του 98%.

[3] Εισαγωγή:

Αυτή η ερευνητική εργασία που συντάχθηκε από τους Vaishnavi Nath Dornadula και Geetha S δίνει μια προσέγγιση μηχανικής μάθησης όπως και όλες οι προηγούμενες για τον εντοπισμό απάτης με πιστωτικές κάρτες. Οι συγγραφείς χρησιμοποιούν ένα σύνολο δεδομένων συναλλαγών με πιστωτικές κάρτες που χαρακτηρίζονται ως δόλιες ή μη και συγκρίνουν διάφορους αλγόριθμους μηχανικής μάθησης, συμπεριλαμβανομένης της λογιστικής

παλινδρόμησης, των δέντρων αποφάσεων και των k-πλησιέστερων γειτόνων, για να προσδιορίσουν την πιο αποτελεσματική μέθοδο για τον εντοπισμό απάτης.

Μέθοδοι:

Η εργασία προτείνει μια προσέγγιση βασισμένη στη μηχανική μάθηση για τον εντοπισμό απάτης με πιστωτικές κάρτες χρησιμοποιώντας τέσσερις αλγόριθμους: δέντρα αποφάσεων, λογιστική παλινδρόμηση, k-πλησιέστερο γείτονα και νευρωνικά δίκτυα. Οι συγγραφείς χρησιμοποίησαν ένα δημοσίως διαθέσιμο σύνολο δεδομένων συναλλαγών με πιστωτικές κάρτες για να εκπαιδεύσουν και να αξιολογήσουν τα μοντέλα τους. Χρησιμοποίησαν τεχνικές επιλογής χαρακτηριστικών και προεπεξεργασίας για να βελτιώσουν την ακρίβεια των μοντέλων τους. Οι συγγραφείς συνέκριναν την απόδοση των μοντέλων μηχανικής εκμάθησης τους με το παραδοσιακό σύστημα ανίχνευσης απάτης που βασίζεται σε κανόνες και έδειξαν ότι η προσέγγισή τους που βασίζεται στη μηχανική μάθηση υπερέρχει του συστήματος που βασίζεται σε κανόνες. Αξιολόγησαν επίσης την απόδοση των μοντέλων τους όσον αφορά την ακρίβεια, την ακρίβεια, την ανάκληση και τη βαθμολογία F1.

Αποτελέσματα:

Οι συγγραφείς απέδειξαν την αποτελεσματικότητα της προσέγγισής τους που βασίζεται στη μηχανική μάθηση για τον εντοπισμό απάτης με πιστωτικές κάρτες, επιτυγχάνοντας υψηλή ακρίβεια και βαθμολογία F1. Το μοντέλο με τις καλύτερες επιδόσεις ήταν το μοντέλο νευρωνικών δικτύων, το οποίο πέτυχε ακρίβεια 99,6% και βαθμολογία F1 0,883. Οι συγγραφείς έδειξαν επίσης ότι η προσέγγισή τους που βασίζεται στη μηχανική μάθηση έχει χαμηλότερο ποσοστό ψευδώς θετικών σε σύγκριση με το παραδοσιακό σύστημα που βασίζεται σε κανόνες. Αυτό είναι σημαντικό επειδή ένα χαμηλότερο ποσοστό ψευδώς θετικών συναλλαγών σημαίνει ότι λιγότερες νόμιμες συναλλαγές επισημαίνονται ως δόλιες, μειώνοντας την ταλαιπωρία για τους πελάτες.

Συμπέρασμα:

Συμπερασματικά, η εργασία προτείνει μια προσέγγιση βασισμένη στη μηχανική μάθηση για τον εντοπισμό απάτης μέσω πιστωτικών καρτών. Οι συγγραφείς απέδειξαν την αποτελεσματικότητα της προσέγγισής τους επιτυγχάνοντας υψηλή ακρίβεια και βαθμολογία F1 διατηρώντας παράλληλα ένα χαμηλό ποσοστό ψευδώς θετικών. Αυτή η ερευνητική εργασία παρέχει μια ανασκόπηση διαφόρων προσεγγίσεων μηχανικής μάθησης που χρησιμοποιούνται για τον εντοπισμό απάτης με πιστωτικές κάρτες. Συζητά για διάφορους αλγόριθμους μηχανικής μάθησης συμπεριλαμβανομένης της λογιστικής παλινδρόμησης, των δέντρων αποφάσεων, των τεχνητών νευρωνικών δικτύων και των μηχανών υποστήριξης διανυσμάτων και την αποτελεσματικότητά τους στην ανίχνευση της απάτης. Επίσης διάφορα σύνολα δεδομένων που χρησιμοποιήθηκαν σε προηγούμενες έρευνες και τις προκλήσεις στον εντοπισμό απάτης με πιστωτικές κάρτες όπως η ανισορροπία τάξης (class imbalance) και η υψηλή διάσταση (high dimensionality). Η εργασία ολοκληρώνεται με μια συζήτηση μελλοντικών κατευθύνσεων έρευνας στον εντοπισμό απάτης με πιστωτικές κάρτες.

[4] Εισαγωγή:

Η ασφάλεια των συναλλαγών με πιστωτικές κάρτες είναι ένας σημαντικός παράγοντας για τη διασφάλιση της οικονομικής ασφάλειας και την πρόληψη της απάτης. Στην προσπάθεια να βελτιωθεί η απάτη, η μηχανική μάθηση έχει αναδειχθεί ως ένα ισχυρό εργαλείο. Σε αυτήν την αναφορά, εξετάζεται η χρήση του Γενετικού Αλγορίθμου (GA) για την επιλογή των βέλτιστων χαρακτηριστικών σε ένα μοντέλο μηχανικής μάθησης που αποσκοπεί στην ανίχνευση απάτης πιστωτικών καρτών.

Μέθοδος: Στην αναφορά αυτή, περιγράφεται η μεθοδολογία που χρησιμοποιήθηκε για την υλοποίηση του μοντέλου ανίχνευσης απάτης. Αναλύονται λεπτομερώς οι διαδικασίες προεπεξεργασίας των δεδομένων, η επιλογή των αλγορίθμων μηχανικής μάθησης που χρησιμοποιήθηκαν και η εφαρμογή του Γενετικού Αλγορίθμου για την βελτιστοποίηση των χαρακτηριστικών.

Αποτελέσματα: Παρουσιάζονται τα αποτελέσματα της μελέτης, συμπεριλαμβανομένης της απόδοσης του μοντέλου στον αντιγραφέα απάτης πιστωτικών καρτών. Επίσης, συζητούνται τα οφέλη της χρήσης του Γενετικού Αλγορίθμου για την επιλογή των χαρακτηριστικών σε σχέση με άλλες προσεγγίσεις.

Συμπεράσματα: Η αναφορά καταλήγει σε συμπεράσματα σχετικά με την αποτελεσματικότητα της προτεινόμενης μεθόδου και τον ρόλο του Γενετικού Αλγορίθμου στην επιτυχή ανίχνευση απάτης πιστωτικών καρτών.

Συνοψίζοντας: Η αναφορά αυτή αναλύει μια προσέγγιση για την ανίχνευση απάτης πιστωτικών καρτών βασισμένη στη μηχανική μάθηση και τον Γενετικό Αλγορίθμο για την επιλογή των βέλτιστων χαρακτηριστικών. Οι προτεινόμενες μέθοδοι και τα αποτελέσματα της μελέτης παρέχουν σημαντική εισαγωγή στον τομέα της ανίχνευσης απάτης πιστωτικών καρτών με χρήση της μηχανικής μάθησης και του Γενετικού Αλγορίθμου.

[5] Εισαγωγή:

Η απάτη στον τομέα των οικονομικών έχει γίνει ένα σοβαρό πρόβλημα που επηρεάζει τις επιχειρήσεις και την κοινωνία γενικότερα. Αυτή η αναφορά αποτελεί μια συστηματική ανασκόπηση της υπάρχουσας βιβλιογραφίας που ασχολείται με τη χρήση της μηχανικής μάθησης για την ανίχνευση χρηματοοικονομικής απάτης.

Συμπεράσματα: Η ανασκόπηση της βιβλιογραφίας αποκάλυψε τα εξής σημαντικά ευρήματα:

- **Πληθώρα Μεθόδων Μηχανικής Μάθησης:** Η βιβλιογραφία αναφέρει πολλές μεθόδους μηχανικής μάθησης που χρησιμοποιούνται για την ανίχνευση χρηματοοικονομικής απάτης. Ανάμεσά τους, οι αλγόριθμοι SVM (Μηχανή Διανυσματικής Στήριξης) και NN (Νευρωνικά Δίκτυα) ξεχωρίζουν ως δημοφιλείς.
- **Τύποι Απάτης:** Η ανασκόπηση διακρίνει διάφορους τύπους απάτης, με την απάτη σε πιστωτικές κάρτες να είναι ο πιο διαδεδομένος.
- **Ανάγκη για Μελλοντικές Έρευνες:** Παρατηρήθηκε ότι υπάρχει χώρος για μελλοντικές έρευνες στον τομέα της ανίχνευσης απάτης. Συγκεκριμένα, η χρήση της συνάθροισης πολλαπλών αλγορίθμων μηχανικής μάθησης και η εξέταση των ανεπεξέργαστων πρακτικών όπως η ανίχνευση ανωμαλιών θεωρούνται υποσχόμενοι τομείς.

Περιορισμοί: Οι περιορισμοί αυτής της ανασκόπησης περιλαμβάνουν τον περιορισμό στα αγγλικά κείμενα και τον περιορισμό σε δημοσιεύσεις σε συνέδρια και περιοδικά. Επίσης, μπορεί να υπάρχουν πιθανές αναξιόπιστες βάσεις δεδομένων που δεν λήφθηκαν υπόψη.

Συμπεράσματα: Η ανασκόπηση αυτή αποτελεί μια χρήσιμη πηγή γνώσης για την ανίχνευση χρηματοοικονομικής απάτης με χρήση μηχανικής μάθησης. Τα ευρήματα και οι προτάσεις για μελλοντικές έρευνες που περιέχονται στο άρθρο αυτό αποτελούν σημαντική συνεισφορά στον τομέα της ανίχνευσης χρηματοοικονομικής απάτης.

Σύνοψη Ερευνών

Η απάτη με πιστωτικές κάρτες είναι ένα σοβαρό πρόβλημα στον τραπεζικό και χρηματοπιστωτικό κλάδο. Τα τελευταία χρόνια, οι μέθοδοι μηχανικής μάθησης έχουν αποκτήσει δημοτικότητα στον εντοπισμό δόλιων συναλλαγών λόγω της ικανότητάς τους να επεξεργάζονται μεγάλες ποσότητες δεδομένων σε πραγματικό χρόνο. Αυτό το άρθρο εξετάζει επτά ερευνητικές έρευνες σχετικά με τις μεθόδους μηχανικής εκμάθησης για τον εντοπισμό απάτης με πιστωτικές κάρτες.

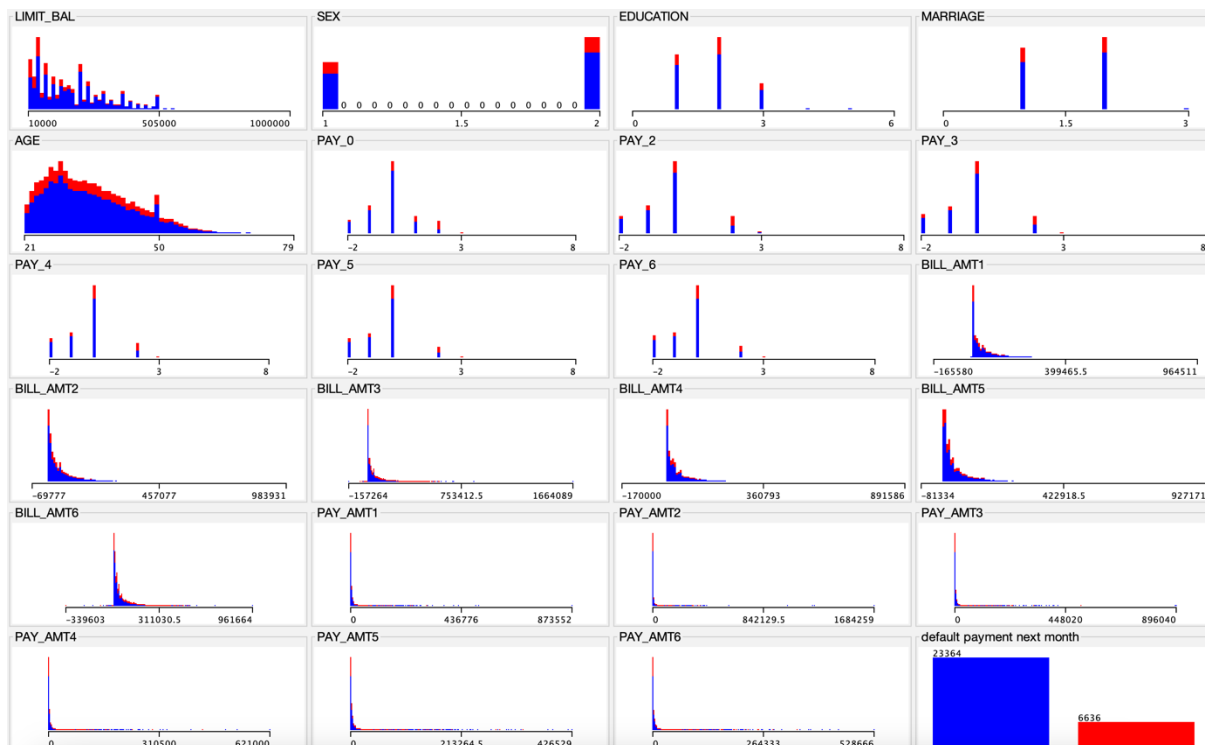
Τα άρθρα καλύπτουν διάφορες τεχνικές, όπως μάθηση με επίβλεψη και χωρίς επίβλεψη, βαθιά μάθηση και μεθόδους συνόλου. Η ανασκόπηση επισημαίνει τα πλεονεκτήματα και τους περιορισμούς κάθε μεθόδου και συγκρίνει την απόδοσή της χρησιμοποιώντας διαφορετικές μετρήσεις όπως η ακρίβεια, η ακρίβεια, η ανάκληση και η βαθμολογία F1. Τα αποτελέσματα δείχνουν ότι οι μέθοδοι μηχανικής μάθησης έχουν βελτιώσει την ακρίβεια και την αποτελεσματικότητα των συστημάτων ανίχνευσης απάτης. Η ανασκόπηση εξετάζει επίσης τις προκλήσεις και τις μελλοντικές ερευνητικές κατευθύνσεις σε αυτόν τον τομέα.

Συνολικά, η ανασκόπηση παρέχει πολύτιμες πληροφορίες για ερευνητές και

επαγγελματίες που ενδιαφέρονται να αναπτύξουν αποτελεσματικά και ισχυρά συστήματα ανίχνευσης απάτης για συναλλαγές με πιστωτικές κάρτες

1.4 Επισκόπηση δεδομένων (Data Overview)

Το σύνολο δεδομένων περιλαμβάνει 25 χαρακτηριστικά και 30.000 περιπτώσεις και το χαρακτηριστικό στόχος είναι η " default payment next month ". Για την εμφάνιση της κατανομής στο Weka, το χαρακτηριστικό προορισμού έχει μετατραπεί σε ονομαστικό χαρακτηριστικό χρησιμοποιώντας το φίλτρο "NumericToNominal". Το φίλτρο numericaToNominal λαμβάνει την παράμετρο attributeIndices, η οποία θα πρέπει να οριστεί στον συγκεκριμένο αριθμό χαρακτηριστικού που πρέπει να αλλάξει. Στο χαρακτηριστικό προορισμού, υπάρχουν 23364 παρουσίες χωρίς σφάλμα (0) και παρουσίες σφάλματος '6636' (1). Η κατανομή των χαρακτηριστικών φαίνεται στο σχήμα 1.



Σχήμα 1: Κατανομές των χαρακτηριστικών

Από την κατανομή στην εικόνα 1, υπάρχουν ορατά προβλήματα δεξιάς ασυμμετρίας και ορισμένα από τα χαρακτηριστικά δείχνουν κανονική κατανομή του αρχείου καταγραφής. Ο πίνακας 1 δείχνει την περιγραφική ανάλυση του συνόλου δεδομένων:

Πίνακας 1: Περιγραφική ανάλυση των αριθμητικών χαρακτηριστικών

Attributes	mean	std	min	50%	max
LIMIT_BAL	167484.32	129747.66	10000.0	140000.0	1000000.0
AGE	35.49	9.22	21.0	34.0	79.0
BILL_AMT1	51223.33	73635.86	- 165580.0	22381.5	964511.0
BILL_AMT2	49179.08	71173.77	-69777.0	21200.0	983931.0
BILL_AMT3	47013.15	69349.39	- 157264.0	20088.5	1664089.0
BILL_AMT4	43262.95	64332.86	- 170000.0	19052.0	891586.0
BILL_AMT5	40311.4	60797.16	-81334.0	18104.5	927171.0
BILL_AMT6	38871.76	59554.11	- 339603.0	17071.0	961664.0
PAY_AMT1	5663.58	16563.28	0.0	2100.0	873552.0
PAY_AMT2	5921.16	23040.87	0.0	2009.0	1684259.0
PAY_AMT3	5225.68	17606.96	0.0	1800.0	896040.0
PAY_AMT4	4826.08	15666.16	0.0	1500.0	621000.0
PAY_AMT5	4799.39	15278.31	0.0	1500.0	426529.0
PAY_AMT6	5215.5	17777.47	0.0	1500.0	528666.0

Από την περιγραφική ανάλυση, τα εύρη είναι διαφορετικά για κάθε χαρακτηριστικό και τα εύρη είναι αρκετά υψηλά. Έτσι, πρέπει να ομαλοποιηθεί το σύνολο δεδομένων, να προεπεξεργαστούν τα χαρακτηριστικά, να εφαρμοθούν κάποια από τη διαδικασία μηχανικής χαρακτηριστικών για να τα τροφοδοτηθεί τελικά στο μοντέλο μηχανικής μάθησης.

Κεφάλαιο 2^ο - Επεξεργασία στο Σύνολο Δεδομένων

2. Προεπεξεργασία δεδομένων (Data Preprocessing)

Data preprocessing αναφέρεται στην επεξεργασία του συνόλου δεδομένων πριν από την εφαρμογή του μοντέλου μηχανικής μάθησης. Η προεπεξεργασία είναι απαραίτητη για την κάλυψη της περίπτωσης χρήσης της δήλωσης προβλήματος και για την επίτευξη του καλύτερου αποτελέσματος από το μοντέλο. Τα ακόλουθα βήματα λαμβάνονται για την προεπεξεργασία του συνόλου δεδομένων για την εφαρμογή των μοντέλων μηχανικής μάθησης:

2.1. Κατάργηση περιττού χαρακτηριστικού (Remove Unnecessary Attribute)

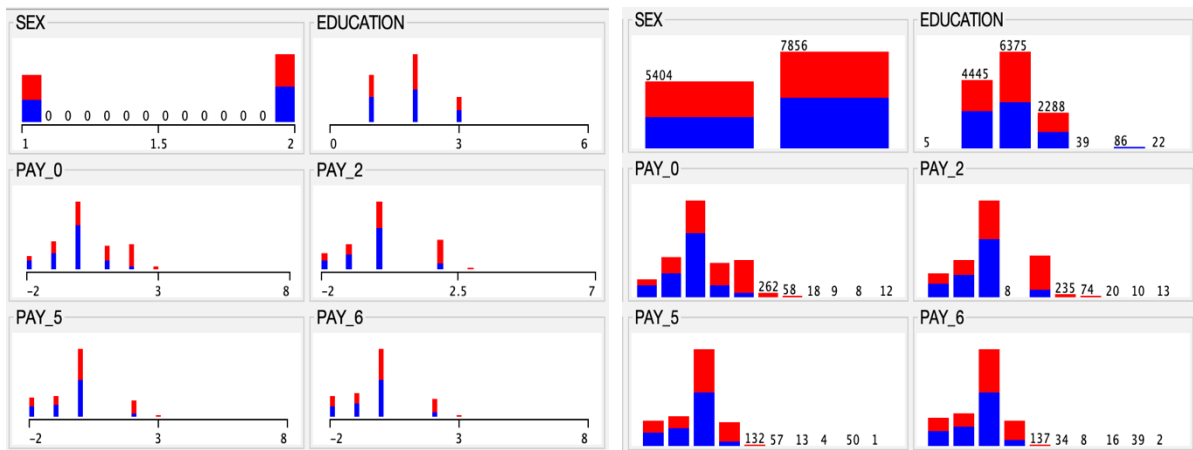
Στην αρχή, το χαρακτηριστικό "ID" καταργήθηκε επειδή το "ID" δεν είναι απαραίτητο για την πρόβλεψη.

2.2. Ελλιπής/διπλός χειρισμός (Missing/Duplicates Handling)

Το σύνολο δεδομένων δεν έχει δείγματα που λείπουν. Ωστόσο, υπάρχουν λιγότερες διπλότυπες περιπτώσεις. Για την κατάργηση των διπλότυπων δειγμάτων, έχει χρησιμοποιηθεί το φίλτρο "RemoveDuplicate". 35 διπλότυπα δείγματα εντοπίζονται και καταργούνται εφαρμόζοντας το φίλτρο.

2.3. Μετατροπή σε ονομαστική (Convert into Nominal)

Weka μπορεί να χειριστεί τα ονομαστικά χαρακτηριστικά ενώ κάνει την ταξινόμηση. Ορισμένα από τα χαρακτηριστικά στο σύνολο δεδομένων είναι ονομαστικά, τα οποία κωδικοποιούνται σε αριθμητικές τιμές. Η μετατροπή γίνεται χρησιμοποιώντας το φίλτρο 'NumericToNominal'. Στην παράμετρο attributeIndices πρέπει να οριστούν οι αριθμοί χαρακτηριστικών. Η εικόνα 4 δείχνει τα αποτελέσματα μετά την εφαρμογή του φίλτρου:



Σχήμα 2: Εφέ φίλτρου NumericToNominal πριν (αριστερά), μετά (δεξιά) επίδραση του φίλτρου

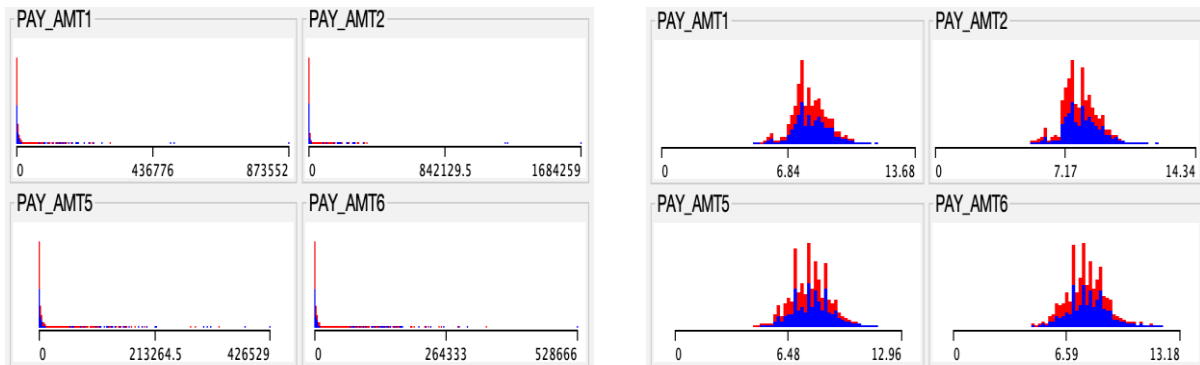
2.4. Αφαίρεση ακραίων τιμών (Remove Outliers)

Από την κατανομή των χαρακτηριστικών, ελέγχοντας την ασυμμετρία, μπορεί να υποτεθεί ότι ενδέχεται να υπάρχουν κάποιες ακραίες τιμές στο σύνολο δεδομένων. Επομένως, για τον προσδιορισμό των ακραίων τιμών στο σύνολο δεδομένων, χρησιμοποιείται το φίλτρο «InterquartileRange» με τις προεπιλεγμένες τιμές παραμέτρων. 4238 περιπτώσεις εντοπίζονται ως ακραίες τιμές. Μετά τον προσδιορισμό των ακραίων τιμών, η αφαίρεση των ακραίων τιμών γίνεται χρησιμοποιώντας το φίλτρο "RemoveWithValues". Ο αριθμός χαρακτηριστικού outlier πρέπει να οριστεί στην παράμετρο `attributeIndex` και το `nominalIndices` πρέπει να ορίσει την «τελευταία» τιμή που υποδεικνύει την τιμή outliers.

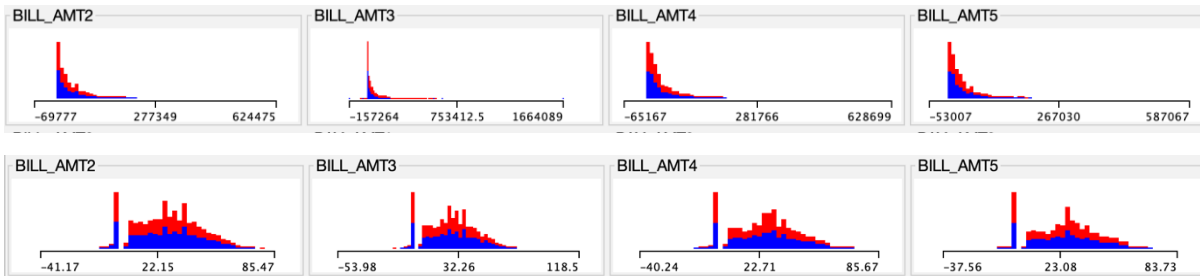
2.5. Αριθμητικός μετασχηματισμός (Numeric Transform)

Από την κατανομή χαρακτηριστικών στην εικόνα 1, μερικά χαρακτηριστικά δείχνουν κανονική κατανομή log και μερικά δείχνουν μέτρια προβλήματα δεξιάς ασυμμετρίας. Τα χαρακτηριστικά που εμφανίζουν κανονική κατανομή log (PAY_AMT_1, PAY_AMT_2, PAY_AMT_3, PAY_AMT_4, PAY_AMT_5, PAY_AMT_6) έχουν μετασχηματιστεί log χρησιμοποιώντας το φίλτρο 'NumericTransform' ('log' στην παράμετρο `methodName`). Τα άλλα χαρακτηριστικά που εμφανίζουν μέτρια δεξιά ασυμμετρία (LIMIT_BAL, BILL_AMT_1, BILL_AMT_2, BILL_AMT_3, BILL_AMT_4, BILL_AMT_5, BILL_AMT_6) έχουν μετασχηματιστεί εφαρμόζοντας τη συνάρτηση ρίζας κύβου ('cbrt' στην παράμετρο

methodName) χρησιμοποιώντας το φίλτρο 'NumericTransform'. Τα σχήματα 3 και 4 δείχνουν τη διάταξη μετασχηματισμού «log» και «cbrt» του φίλτρου:



σχήμα 3: Επιδράσεις μετασχηματισμού κορμών πριν(αριστερά), μετά(δεξιά)



Εικόνα 4: Επιδράσεις μετασχηματισμού ρίζας κύβου, πριν (πάνω), μετά (κάτω)

2.6. Χειρισμός ανισορροπιών (Imbalance Handling)

Το χαρακτηριστικό class δεν είναι ισορροπημένο και φαίνεται στο σχήμα 1. Το χαρακτηριστικό πρέπει να είναι ισορροπημένο για να αποκτήσετε ένα γενικευμένο μοντέλο. Η χαμηλότερη τάξη έχει 6027 περιπτώσεις, κάτι που είναι αρκετά καλό για την εκπαίδευση ενός μοντέλου. Έτσι, η υποδειγματοληψία είναι μια καλή επιλογή σε αυτή την περίπτωση, καθώς η υπερδειγματοληψία παράγει εικονικά δεδομένα που μπορεί να μειώσουν την αποτελεσματικότητα του μοντέλου ενώ θα εισαχθούν στα πραγματικά δεδομένα. Επομένως, η υποδειγματοληψία γίνεται χρησιμοποιώντας το φίλτρο «SpreadSubsample». Η παράμετρος 'distributionSpread' έχει οριστεί σε -1 για να γίνει η υποδειγματοληψία. Η πλειοψηφική κατηγορία μειώθηκε σε 6027 περιπτώσεις μετά την εφαρμογή της τεχνικής υποδειγματοληψίας. Η υποδειγματοληψία δεν επηρεάζει την ποιότητα των δεδομένων και είναι ο πιο χρήσιμος όρος για τον χειρισμό των ζητημάτων ανισορροπίας, ενώ υπάρχουν αρκετές παρουσίες δεδομένων.

2.7. Τυποποίηση (Standardization)

Τέλος, τυποποιούμε το σύνολο δεδομένων χρησιμοποιώντας το φίλτρο 'Standardize'. Η τυποποίηση, γνωστή και ως κλιμάκωση χαρακτηριστικών ή κανονικοποίηση, είναι μια τεχνική προεπεξεργασίας δεδομένων που χρησιμοποιείται στη μηχανική μάθηση για τη μετατροπή των χαρακτηριστικών ενός συνόλου δεδομένων σε κοινή κλίμακα. Η διαδικασία περιλαμβάνει την αναπροσαρμογή των τιμών κάθε χαρακτηριστικού έτσι ώστε να έχουν μέσο όρο 0 και τυπική απόκλιση 1. Αυτός ο μετασχηματισμός είναι απαραίτητος όταν τα χαρακτηριστικά έχουν διαφορετικές κλίμακες ή μονάδες μέτρησης, καθώς βοηθά να φέρει όλα τα χαρακτηριστικά σε συγκρίσιμο επίπεδο, εμποδίζοντας ορισμένα χαρακτηριστικά να κυριαρχήσουν στη μαθησιακή διαδικασία. Οι προεπιλεγμένες τιμές παραμέτρων χρησιμοποιούνται για την τυποποίηση του συνόλου δεδομένων.

2.8 Μηχανική χαρακτηριστικών (Feature Engineering)

Feature engineering αποτελεί μέρος της τροποποίησης των χαρακτηριστικών με τρόπο που μπορεί να βοηθήσει στη βελτίωση της απόδοσης των αλγορίθμων. Το σύνολο δεδομένων έχει 24 χαρακτηριστικά μετά την κατάργηση του χαρακτηριστικού ID. Έτσι, η επιλογή χαρακτηριστικών θα μπορούσε να διαδραματίσει σημαντικό ρόλο εδώ πριν από την εφαρμογή του μοντέλου μηχανικής μάθησης. Η μέθοδος επιλογής Info Gain Attribute χρησιμοποιείται χρησιμοποιώντας το φίλτρο 'AttributeSelection'. Η τιμή κατωφλίου ορίστηκε σε 0 για να καταργηθεί το χαρακτηριστικό που βαθμολογήθηκε λιγότερο ή ίσο με 0. Η μέθοδος επιλογής χαρακτηριστικών αφαίρεσε 5 χαρακτηριστικά (BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6) από το σύνολο δεδομένων που είναι λιγότερο σημαντικά για την πρόβλεψη.

Κεφάλαιο 3^ο - Χρήση αλγορίθμων και Μεθόδων

3. Χτίσιμο μοντέλου (Model Building)

Τώρα, είναι η σειρά να δημιουργήσουμε το καλύτερο μοντέλο για αυτό το συγκεκριμένο σύνολο δεδομένων. Το σύνολο δεδομένων έχει υποβληθεί σε προεπεξεργασία για να έχει καλύτερο αποτέλεσμα για αυτό το μη ισορροπημένο σύνολο δεδομένων. Έτσι, εφαρμόσαμε αλγόριθμους στο ακατέργαστο σύνολο δεδομένων, στο προεπεξεργασμένο σύνολο δεδομένων και στο σύνολο δεδομένων Information Gain Feature Evaluation (IGAE) και συγκρίναμε το αποτέλεσμά τους για να δούμε τις βελτιώσεις. 6 διαφορετικοί αλγόριθμοι: Random Forest (RF), K Nearest Neighbors (KNN) και Decision Tree (DT), Support Vector Machine (SVM), Bagging, Adaptive Boosting with Multiclass Classification (AdaBoostM1).

3.1. Διασταυρούμενη επικύρωση (Cross Validation)

- i. Για την αξιολόγηση της απόδοσης του αλγορίθμου, χρησιμοποιήθηκε 10-fold cross validation. Η μέθοδος cross validation μειώνει τα προβλήματα υπερφόρτωσης και βελτιώνει την απόδοση γενίκευσης του μοντέλου. Επομένως, το cross validation είναι πιο αξιόπιστη από τη μέθοδο διαχωρισμού ποσοστού και το 10 φορές είναι ένας προτεινόμενος αριθμός από το μεγαλύτερο μέρος του ερευνητή [1]. Το cross validation έχει τα ακόλουθα πλεονεκτήματα:
- ii. Ο εκτιμητής cross validation έχει μικρότερη διακύμανση από έναν μεμονωμένο εκτιμητή συνόλου αναμονής
- iii. Τα αποτελέσματα cross validation είναι πιο αξιόπιστα από ένα μεμονωμένο σύνολο αποτελεσμάτων
- iv. Μειώνει επίσης τις πιθανότητες υπερβολικής τοποθέτησης

3.2. Αξιολόγηση επιδόσεων (Performance Evaluation)

Για να ελεγχθεί η αποτελεσματικότητα του μοντέλου στο σύνολο δεδομένων και να συγκριθούν τα αποτελέσματά τους, χρησιμοποιούνται διάφορες μετρήσεις αξιολόγησης. Εδώ, έχουν ληφθεί υπόψη οι μετρήσεις αξιολόγησης Accuracy, AUROC και F-Measure.

- **Accuracy:** Η ακρίβεια αναφέρεται στην ορθότητα του μοντέλου, η οποία υποδεικνύει πόσο ακριβής είναι η πρόβλεψη και μπορεί να υπολογιστεί χρησιμοποιώντας τον παρακάτω τύπο:

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

- **Precision:** Η ακρίβεια απαντά στην ερώτηση: "Από όλες τις περιπτώσεις που προβλέφθηκαν ως θετικές, πόσες ήταν πραγματικά θετικές;" Η ακρίβεια είναι ιδιαίτερα χρήσιμη σε καταστάσεις όπου τα ψευδώς θετικά (εσφαλμένα προβλεπόμενα θετικά) είναι δαπανηρά ή ανεπιθύμητα. Βοηθά στην αξιολόγηση της ικανότητας του μοντέλου να αποφεύγει ψευδώς θετικά σφάλματα. Υπολογίζεται χρησιμοποιώντας τον παρακάτω τύπο:

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

- **Recall:** Το Recall απαντά στην ερώτηση: "Από όλες τις θετικές περιπτώσεις στο σύνολο δεδομένων, πόσες αναγνώρισε σωστά το μοντέλο;" Η ανάκληση είναι ιδιαίτερα χρήσιμη σε καταστάσεις όπου τα ψευδώς αρνητικά (εσφαλμένα προβλεπόμενα αρνητικά) είναι δαπανηρά ή ανεπιθύμητα. Βοηθά στην αξιολόγηση της ικανότητας του μοντέλου να αποφεύγει την απώλεια θετικών περιπτώσεων.

$$Precision = \frac{TP}{(TP + FN)}$$

- **F-Measure:** Η ακρίβεια είναι το ποσοστό των ανακτημένων παρουσιών που είναι σχετικές. Τα ποσοστά των σχετικών περιπτώσεων που ανακτήθηκαν είναι αυτά που συνθέτουν τη θετική προγνωστική αξία και ανάκληση ή ευαισθησία. Τέλος, η ανάκληση και η ακρίβεια είναι οι αναπαραστάσεις του F-Measure.

$$F - Measure = \frac{2 * Precision * Recall}{(Precision + Recall)}$$

Το F-Measure χρησιμοποιείται εδώ στο τμήμα ρύθμισης παραμέτρων επειδή το σύνολο δεδομένων ήταν λίγο μη ισορροπημένο και η ακρίβεια δεν είναι καλή επιλογή για το μη ισορροπημένο σύνολο δεδομένων.

3.3. K Πλησιέστερος γείτονας (K Nearest Neighbor)

Ο K Nearest Neighbor, γνωστός και ως IBk, είναι ένας από τους διάσημους αλγόριθμους που βασίζονται στην απόσταση στον κόσμο της μηχανικής μάθησης [2]. Είναι ένας μη παραμετρικός και τεμπέλης αλγόριθμος μάθησης, που σημαίνει ότι δεν κάνει υποθέσεις σχετικά με την υποκείμενη κατανομή δεδομένων και δεν δημιουργεί ένα ρητό μοντέλο κατά τη διάρκεια της φάσης εκπαίδευσης. Η κύρια ιδέα πίσω από το KNN είναι να ταξινομήσει ή να προβλέψει ένα νέο σημείο δεδομένων λαμβάνοντας υπόψη τα πλησιέστερα σημεία δεδομένων K από το σύνολο δεδομένων εκπαίδευσης. Η κλάση ή η τιμή του νέου σημείου δεδομένων καθορίζεται από την πλειοψηφία ή τον μέσο όρο των τιμών των πλησιέστερων γειτόνων K. Με άλλα λόγια, το KNN εκχωρεί την κλάση ή την τιμή που είναι πιο κοινή μεταξύ των πλησιέστερων γειτόνων K.

Το KNN είναι σχετικά εύκολο να εφαρμοστεί και να κατανοηθεί, καθιστώντας το ένα καλό σημείο εκκίνησης για πολλές εργασίες ταξινόμησης και παλινδρόμησης. Ωστόσο, μπορεί να είναι υπολογιστικά ακριβό, ειδικά με μεγάλα σύνολα δεδομένων, καθώς απαιτεί υπολογισμό αποστάσεων για κάθε πρόβλεψη. Επιπλέον, η επιλογή μιας κατάλληλης τιμής για το K και η επιλογή της σωστής μέτρησης απόστασης είναι ζωτικής σημασίας για την επίτευξη των καλύτερων αποτελεσμάτων.

Εδώ, το IBk χρησιμοποιείται χρησιμοποιώντας διάφορες τιμές K από 1-15 για να ανακαλύψει το καλύτερο αποτέλεσμα από το σύνολο δεδομένων. Ο συντονισμός παραμέτρων παρατηρείται στο επιλεγμένο σύνολο δεδομένων και εφαρμόζεται και στα 3 σύνολα δεδομένων που συζητήθηκαν προηγουμένως. Το μέτρο F χρησιμοποιήθηκε για την αξιολόγηση των επιδόσεων. Η διαδικασία και τα αποτελέσματα παρουσιάζονται στον ακόλουθο πίνακα 2.

Πίνακας 2: Διαδικασία ρύθμισης παραμέτρων

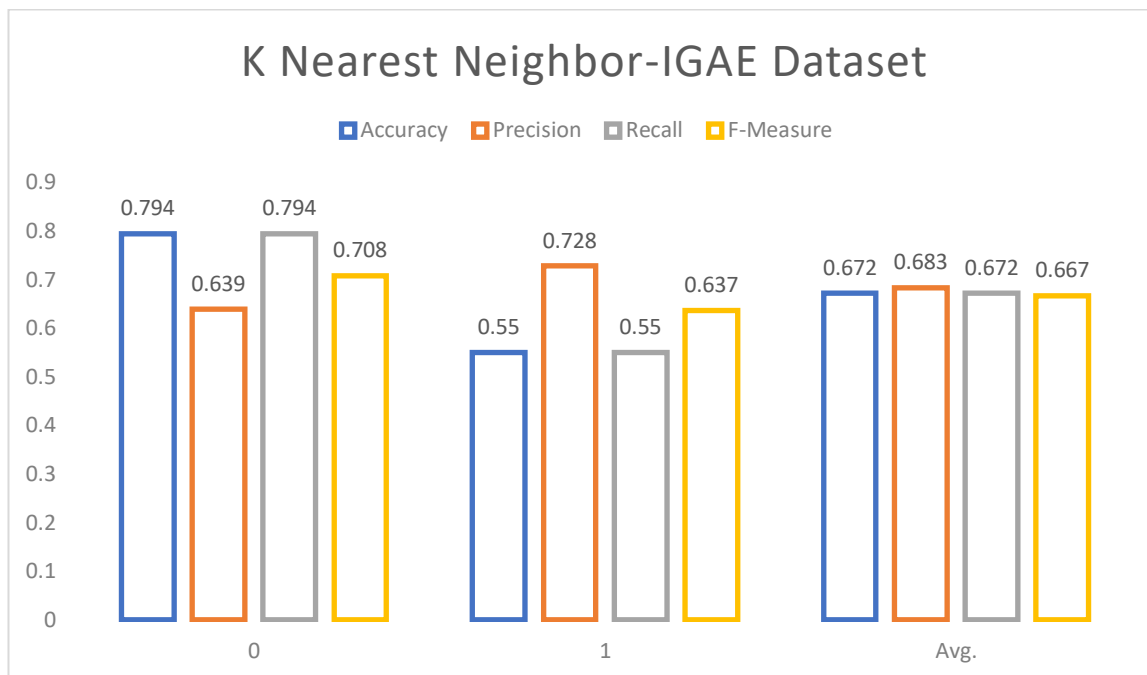
Parameter	F-Measure
K=1	0.613
K=3	0.636
K=5	0.655
K=7	0.664
K=9	0.672
K=11	0.672
K=13	0.682
K=13	0.672

Για το $K = 13$ είναι η επίτευξη του υψηλότερου F-Μέτρου. Ο πίνακας 3 δείχνει την απόδοση των αλγορίθμων χρησιμοποιώντας το καλύτερο σύνολο παραμέτρων:

Πίνακας 3: Απόδοση αλγορίθμου πλησιέστερου γείτονα K

Dataset	Accuracy (weighted avg.)	Precision (weighted avg.)	Recall (weighted avg.)	F-Measure (weighted avg.)
Raw	0.809	0.788	0.809	0.786
Raw Processed	0.672	0.683	0.672	0.667
IGAE	0.672	0.683	0.672	0.667

Το σχήμα 5 δείχνει την απόδοση του KNN στο σύνολο δεδομένων IGAE:



3.4. Δέντρο αποφάσεων (Decision Tree)

Ένα δέντρο αποφάσεων είναι ένας δημοφιλής αλγόριθμος εποπτευόμενης μηχανικής μάθησης που χρησιμοποιείται τόσο για εργασίες ταξινόμησης όσο και παλινδρόμησης [3]. Δημιουργεί ένα μοντέλο που μοιάζει με δέντρο όπου κάθε εσωτερικός κόμβος αντιπροσωπεύει μια απόφαση που βασίζεται σε ένα χαρακτηριστικό, κάθε κλάδος αντιστοιχεί στο αποτέλεσμα της απόφασης και κάθε κόμβος φύλλου αντιπροσωπεύει μια ετικέτα κλάσης ή μια προβλεπόμενη τιμή. Τα δέντρα αποφάσεων είναι ιδιαίτερα ερμηνεύσιμα, καθώς η διαδικασία λήψης αποφάσεων αναπαρίσταται σε μια απλή οπτική μορφή. Μπορούν να χειριστούν τόσο αριθμητικά όσο και κατηγορικά δεδομένα και είναι σχετικά μη ευαίσθητα σε ακραίες τιμές. Ωστόσο, τα δέντρα αποφάσεων είναι επιρρεπή σε υπερβολική προσαρμογή, πράγμα που σημαίνει ότι μπορεί να απομνημονεύσουν τα δεδομένα προπόνησης και να έχουν κακή απόδοση σε άγνωστα δεδομένα. Τεχνικές όπως το κλάδεμα, ο καθορισμός ελάχιστου αριθμού δειγμάτων ανά φύλλο ή η χρήση μεθόδων συνόλου όπως το Random Forest ή το Gradient Boosting μπορούν να βοηθήσουν στον μετριασμό της υπερβολικής τοποθέτησης και στη βελτίωση της απόδοσης.

Εδώ, χρησιμοποιείται το δέντρο αποφάσεων και ελέγχονται διάφορες τιμές παραμέτρων, συμπεριλαμβανομένων των τιμών minNumObj από 2-7, των τιμών κλαδέματος (True, False) και των δυαδικών διαιρέσεων (True, False) και της εμπιστοσύνης (0,15-0,35) για να μάθετε το καλύτερο αποτέλεσμα από το σύνολο δεδομένων. Ο συντονισμός παραμέτρων παρατηρείται στο επιλεγμένο σύνολο δεδομένων και εφαρμόζεται και στα 3 σύνολα

δεδομένων που συζητήθηκαν προηγουμένως. Το μέτρο F χρησιμοποιήθηκε για την αξιολόγηση των επιδόσεων. Η διαδικασία και τα αποτελέσματα παρουσιάζονται στον ακόλουθο πίνακα 4:

Πίνακας 4: Ρύθμιση υπερπαραμέτρων του αλγορίθμου του δέντρου αποφάσεων

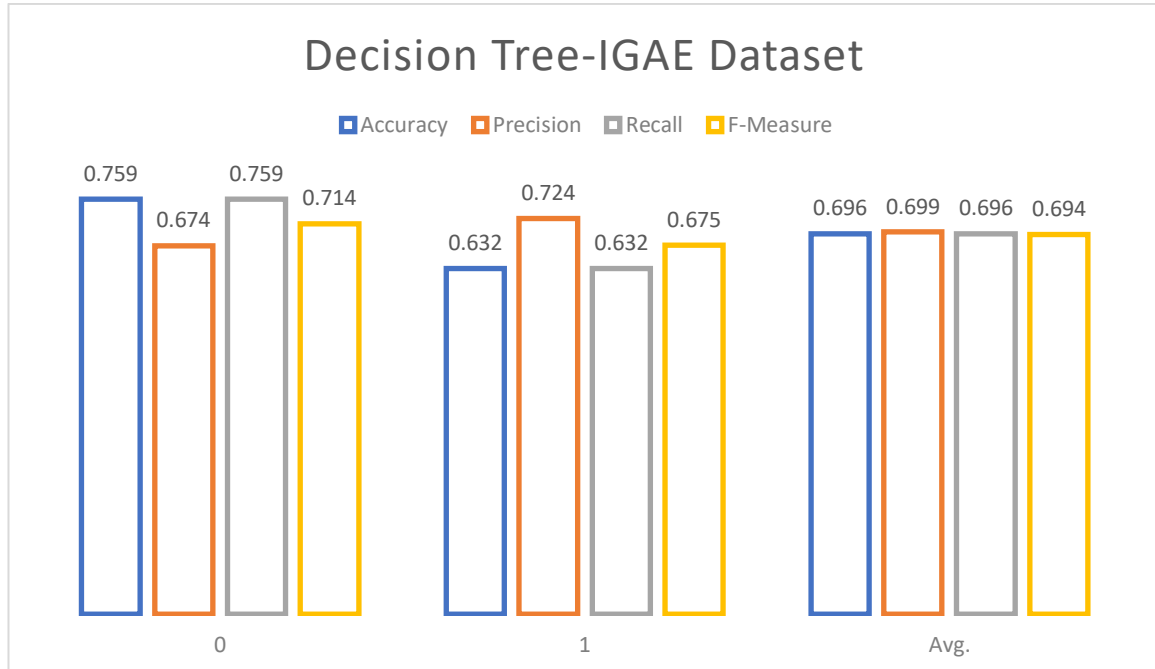
Parameter	F-Measure
default	0.685
Unpruned= True	0.656
binarySplits=True	0.686
binarySplits=True, Confidence=0.15	0.699
binarySplits=True, Confidence=0.20	0.685
binarySplits=True, Confidence=0.30	0.682
binarySplits=True, Confidence=0.35	0.678
binarySplits=True, Confidence=0.15, minNumObj= 4	0.709
binarySplits=True, Confidence=0.15, minNumObj= 5	0.706
binarySplits=True, Confidence=0.15, minNumObj= 6	0.710
binarySplits=True, Confidence=0.15, minNumObj= 7	0.708

Το καλύτερο σύνολο παραμέτρων που βρέθηκε είναι binarySplits=True, confidence=0.15, minNumObj=6. Ο πίνακας 5 δείχνει την απόδοση των αλγορίθμων χρησιμοποιώντας τις καλύτερες παραμέτρους:

Πίνακας 5: Απόδοση αλγορίθμου δέντρου αποφάσεων

Dataset	Accuracy (weighted avg.)	Precision (weighted avg.)	Recall (weighted avg.)	F-Measure (weighted avg.)
Raw	0.817	0.799	0.817	0.795

Raw Processed	0.694	0.696	0.694	0.693
IGAE	0.696	0.699	0.696	0.694



Σχήμα 6: Απόδοση δέντρου αποφάσεων σε σύνολο δεδομένων IGAE

3.5. Τυχαίο δάσος (Random Forest)

Το Random Forest είναι μια μέθοδος εκμάθησης συνόλου που βασίζεται σε δέντρα αποφάσεων [4]. Είναι ένας από τους πιο ισχυρούς και ευρέως χρησιμοποιούμενους αλγόριθμους μηχανικής μάθησης για εργασίες ταξινόμησης και παλινδρόμησης.

Τα βασικά πλεονεκτήματα του Random Forest είναι η ικανότητά του να χειρίζεται μεγάλα σύνολα δεδομένων με υψηλή διάσταση, η αντοχή του στην υπερβολική προσαρμογή και η ανθεκτικότητά του έναντι ακραίων στοιχείων (outliers) και θορυβωδών δεδομένων (noisy data). Επιπλέον, το Random Forest παρέχει βαθμολογίες σπουδαιότητας χαρακτηριστικών, υποδεικνύοντας ποια χαρακτηριστικά είναι πιο σχετικά με την πραγματοποίηση προβλέψεων.

Χρησιμοποιείται το Random Forest και έχει ελεγχθεί με διάφορες τιμές παραμέτρων, συμπεριλαμβανομένων των τιμών minNumObj που κυμαίνονται από 2-7, τιμές κλαδέματος (True, False) και δυαδικά splits (True, False) και εμπιστοσύνη (0,15-0,35) καλύτερο αποτέλεσμα από το σύνολο δεδομένων. Ο συντονισμός παραμέτρων παρατηρείται στο

επιλεγμένο σύνολο δεδομένων χαρακτηριστικών και εφαρμόζεται και στα 3 σύνολα δεδομένων που συζητήθηκαν προηγουμένως. Το F-measure έχει χρησιμοποιηθεί για την αξιολόγηση της απόδοσης. Η διαδικασία και τα αποτελέσματα φαίνονται στον παρακάτω πίνακα 6:

Πίνακας 6: Υπερπαραμετρικός συντονισμός του αλγορίθμου Random Forest

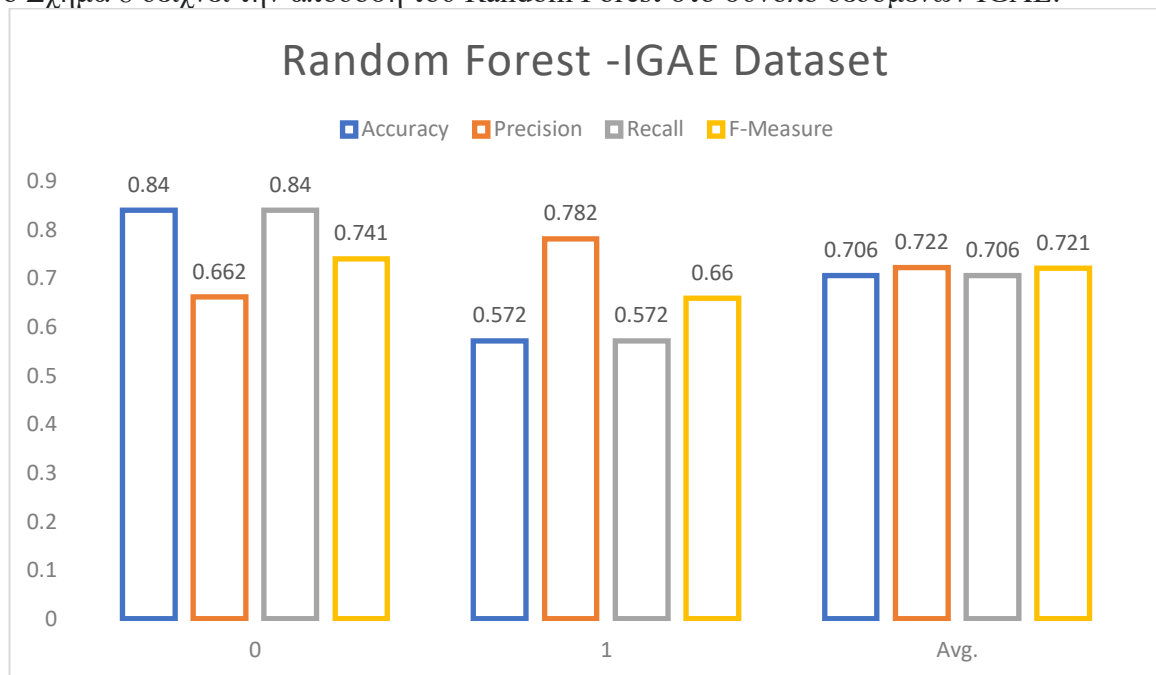
Parameter	F-Measure
default	0.704
max_depth =2	0.715
max_depth =3	0.717
max_depth =4	0.717
max_depth =5	0.716
max_depth =6	0.713
max_depth =3, numFeatures = 2	0.716
max_depth =3, numFeatures = 3	0.721
max_depth =3, numFeatures = 4	0.721
max_depth =3, numFeatures = 5	0.722
max_depth =3, numFeatures = 6	0.721
max_depth =3, numFeatures = 7	0.722
max_depth =3, numFeatures = 8	0.715
max_depth =3, numFeatures = 9	0.715

Το καλύτερο σύνολο παραμέτρων που βρέθηκε είναι max_depth = 3, numFeatures = 7. Ο Πίνακας 7 δείχνει την απόδοση των αλγορίθμων που χρησιμοποιούν τις καλύτερες παραμέτρους:

Πίνακας 7: Απόδοση αλγόριθμου *random forest*

Dataset	Accuracy (weighted avg.)	Precision (weighted avg.)	Recall (weighted avg.)	F-Measure (weighted avg.)
Raw	0.816	0.798	0.816	0.787
Raw Processed	0.709	0.721	0.709	0.725
IGAE	0.706	0.722	0.706	0.721

Το Σχήμα 8 δείχνει την απόδοση του Random Forest στο σύνολο δεδομένων IGAE:



Εικόνα 7: Random Forest σε σύνολο δεδομένων IGAE

3.5 Support Vector Machine

Το Support Vector Machine (SVM) που υλοποιείται στο Weka (LibSVM) είναι ένας ισχυρός και ευρέως χρησιμοποιούμενος εποπτευόμενος αλγόριθμος μηχανικής μάθησης που έχει αποδειχθεί αποτελεσματικός σε διάφορους τομείς, όπως η αναγνώριση προτύπων, η ανάλυση εικόνας και η επεξεργασία φυσικής γλώσσας [5]. Ένα από τα βασικά πλεονεκτήματα του SVM είναι η ικανότητά του να χειρίζεται σύνολα δεδομένων με μεγάλο αριθμό χαρακτηριστικών ή διαστάσεων. Σε ερευνητικά σενάρια όπου ο αριθμός των χαρακτηριστικών υπερβαίνει τον αριθμό των περιπτώσεων, το SVM παραμένει αποτελεσματικό, καθιστώντας το ένα πολύτιμο εργαλείο για την αντιμετώπιση πολύπλοκων και μεγάλης κλίμακας συνόλων δεδομένων.

Επιπλέον, το SVM είναι γνωστό για την στιβαρότητά του έναντι της υπερβολικής προσαρμογής, μια κοινή ανησυχία στη μηχανική μάθηση. Επιδιώκοντας να μεγιστοποιήσει το περιθώριο μεταξύ των τάξεων, το SVM μειώνει τον κίνδυνο απομνημόνευσης θορύβου στα

δεδομένα εκπαίδευσης, διασφαλίζοντας ότι μπορεί να γενικεύσει καλά σε μη ορατά δεδομένα. Αυτή η αξιοπιστία είναι μεγάλης σημασίας σε εφαρμογές πραγματικού κόσμου, όπου η απόδοση του μοντέλου σε νέα και αόρατα δεδομένα είναι απαραίτητη για πρακτική χρήση.

Χρησιμοποιείται Support Vector Machine και έχει ελεγχθεί με διάφορες τιμές παραμέτρων, όπως KernelType, το C κυμαίνεται από 0,1 έως το 1 και το εύρος gamma από 0,01 έως 0,5 για να βρεθεί το καλύτερο αποτέλεσμα από το σύνολο δεδομένων. Ο συντονισμός παραμέτρων παρατηρείται στο επιλεγμένο σύνολο δεδομένων χαρακτηριστικών και εφαρμόζεται και στα 3 σύνολα δεδομένων που συζητήθηκαν προηγουμένως. Το F-measure έχει χρησιμοποιηθεί για την αξιολόγηση της απόδοσης. Η διαδικασία και τα αποτελέσματα φαίνονται στον παρακάτω πίνακα 8:

Πίνακας 8: Συντονισμός υπερπαραμέτρων του αλγόριθμου *Support Vector Machine*

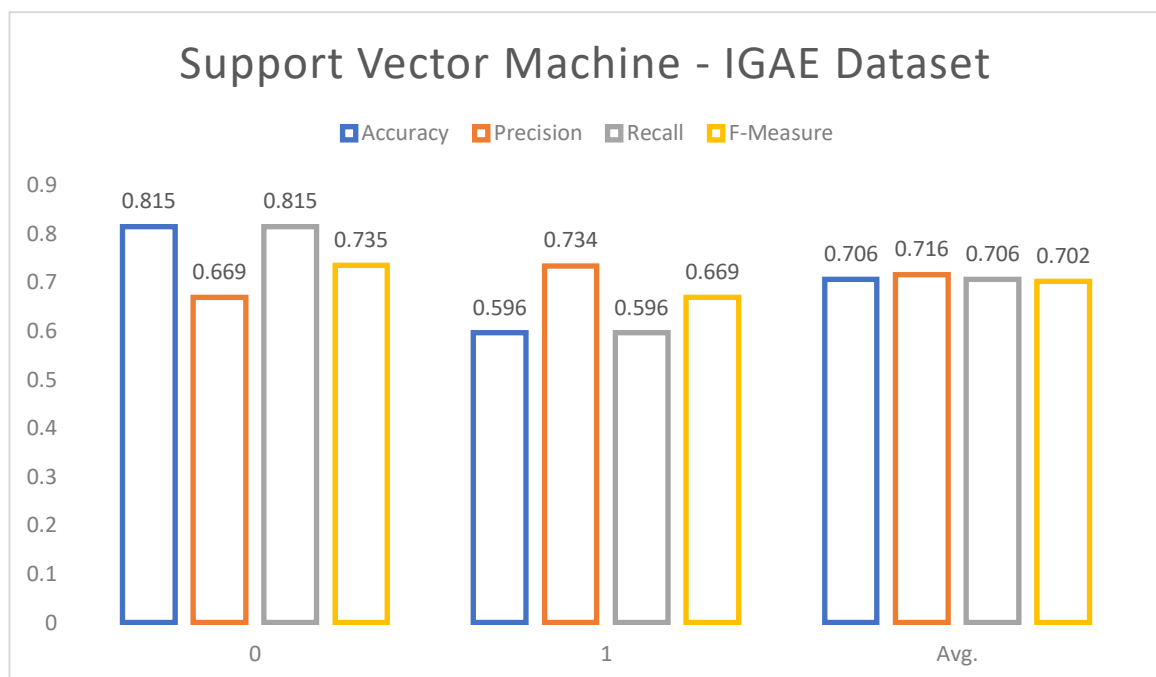
Parameter	F-Measure
default	0.703
kernelType = polynomial	0.566
kernelType = sigmoid	0.689
kernelType = rbf, C = 0.1	0.691
kernelType = rbf, C = 0.5	0.701
kernelType = rbf, C = 5	0.714
kernelType = rbf, C = 10	0.715
kernelType = rbf, C = 10, gamma = 0.01	0.715
kernelType = rbf, C = 10, gamma = 0.05	0.701
kernelType = rbf, C = 10, gamma = 0.1	0.675
kernelType = rbf, C = 10, gamma = 0.5	0.656

Το καλύτερο σύνολο παραμέτρων που βρέθηκε είναι $\text{kernelType} = \text{'rbf'}$, $C = 10$, $\text{gamma} = 0,01$. Ο Πίνακας 9 δείχνει την απόδοση των αλγορίθμων που χρησιμοποιούν τις καλύτερες παραμέτρους:

Πίνακας 9: Απόδοση Αλγόριθμου *Support Vector Machine*

Dataset	Accuracy (weighted avg.)	Precision (weighted avg.)	Recall (weighted avg.)	F-Measure (weighted avg.)
Raw	0.780	0.729	0.780	0.692
Raw Processed	0.709	0.718	0.709	0.706
IGAE	0.706	0.716	0.706	0.702

Το σχήμα 8 δείχνει την απόδοση του Support Vector Machine πάνω από το σύνολο δεδομένων IGAE:



3.6 Bagging

Το Bagging, συντομογραφία του Bootstrap Aggregating, είναι μια τεχνική εκμάθησης συνόλου που στοχεύει στη βελτίωση της απόδοσης και της ευρωστίας των μοντέλων μηχανικής εκμάθησης [6]. Περιλαμβάνει τη δημιουργία πολλαπλών αντιγράφων του ίδιου αλγορίθμου, το καθένα εκπαιδευμένο σε διαφορετικά υποσύνολα των αρχικών δεδομένων εκπαίδευσης. Αυτά τα υποσύνολα δημιουργούνται μέσω του bootstrapping, μιας διαδικασίας τυχαίας δειγματοληψίας περιπτώσεων με αντικατάσταση από το σετ εκπαίδευσης.

Το βασικό πλεονέκτημα του bagging έγκειται στην ικανότητά του να μειώνει τη διακύμανση και να αποτρέπει την υπερβολική τοποθέτηση. Εκπαιδεύοντας πολλαπλά μοντέλα σε διαφορετικά υποσύνολα δεδομένων και συγκεντρώνοντας τις προβλέψεις τους μέσω ψηφοφορίας ή υπολογισμού του μέσου όρου, το bagging βοηθά στη δημιουργία ενός πιο γενικευμένου και σταθερού μοντέλου συνόλου. Αυτή η προσέγγιση συνόλου οδηγεί συχνά σε καλύτερη γενίκευση σε μη ορατά δεδομένα, καθιστώντας την ιδιαίτερα χρήσιμη σε σενάρια όπου μεμονωμένα μοντέλα ενδέχεται να είναι επιρρεπή σε διαφορετικά σφάλματα.

Ο ταξινομητής bagging χρησιμοποιείται χρησιμοποιώντας τον ταξινομητή RepTree και έχει ελεγχθεί με διάφορες τιμές παραμέτρων, συμπεριλαμβανομένων των τιμών bagSizePercent από 20 έως 100, επανάληψης κυμαίνεται από 5 έως 20. Ο συντονισμός του ταξινομητή Reptree έχει επίσης εξετάσει τη χρήση του minNum (2-5) και του NoPruning (True , False) παραμέτρους για να βρεί το καλύτερο αποτέλεσμα από το σύνολο δεδομένων. Ο συντονισμός παραμέτρων παρατηρείται στο επιλεγμένο σύνολο δεδομένων χαρακτηριστικών και εφαρμόζεται και στα 3 σύνολα δεδομένων που συζητήθηκαν προηγουμένως. Το F-measure έχει χρησιμοποιηθεί για την αξιολόγηση της απόδοσης. Η διαδικασία και τα αποτελέσματα φαίνονται στον παρακάτω πίνακα 10:

Πίνακας 10: Υπερπαραμετρικός συντονισμός του ταξινομητή Bagging

Parameter	F-Measure
default	0.695
bagSizePercent = 20	0.706

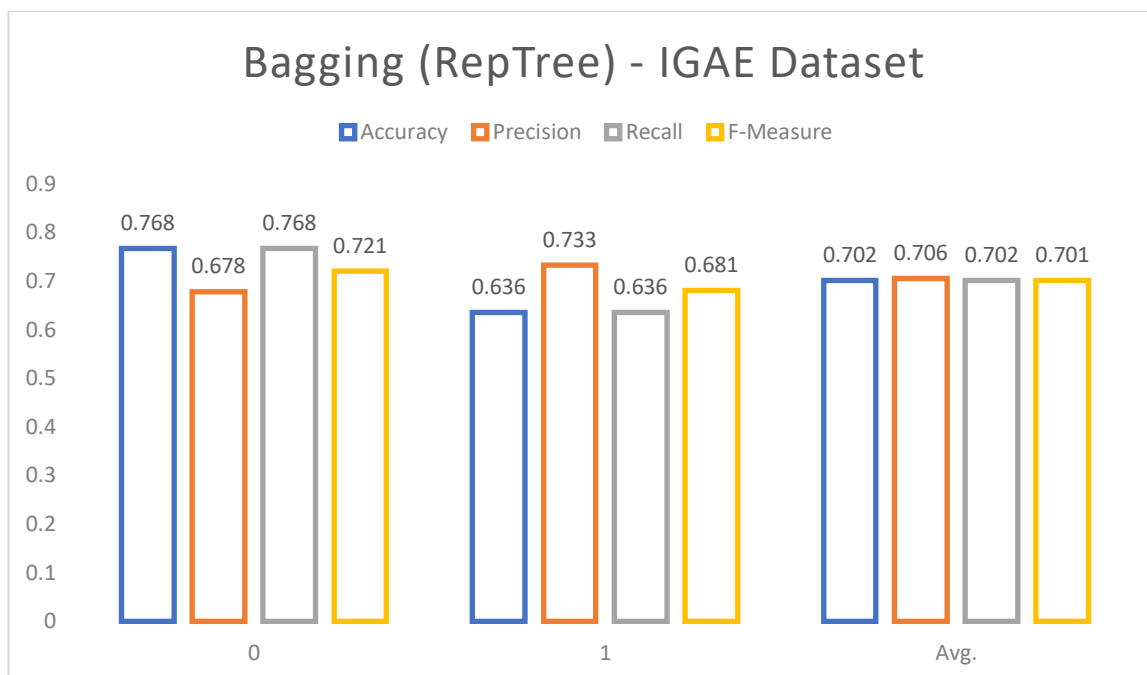
bagSizePercent = 40	0.706
bagSizePercent = 50	0.712
bagSizePercent = 60	0.701
bagSizePercent = 70	0.700
bagSizePercent = 90	0.701
bagSizePercent = 50, Iteration = 5	0.707
bagSizePercent = 50, Iteration = 10	0.712
bagSizePercent = 50, Iteration = 20	0.711
bagSizePercent = 50, Iteration = 10, RepTree(minNum = 3)	0.714
bagSizePercent = 50, Iteration = 10, RepTree(minNum = 5)	0.712
bagSizePercent = 50, Iteration = 10, RepTree(minNum = 5, noPruning = True)	0.697

Το καλύτερο σύνολο παραμέτρων που βρέθηκε είναι bagSizePercent = 50, Iteration = 10, RepTree (minNum = 3). Ο Πίνακας 11 δείχνει την απόδοση των αλγορίθμων που χρησιμοποιούν τις καλύτερες παραμέτρους:

Πίνακας 11: Απόδοση του Bagging με χρήση του RepTree Classifier

Dataset	Accuracy (weighted avg.)	Precision (weighted avg.)	Recall (weighted avg.)	F-Measure (weighted avg.)
Raw	0.820	0.803	0.820	0.798
Raw Processed	0.701	0.704	0.701	0.700
IGAE	0.702	0.706	0.702	0.701

Το σχήμα 9 δείχνει την απόδοση Bagging (RepTree) στο σύνολο δεδομένων IGAE:



Εικόνα 9: Απόδοση Bagging (RepTree) σε σύνολο δεδομένων IGAE

3.6 AdaBoostM1

Το AdaBoostM1, συντομογραφία της Adaptive Boosting with Multiclass classification, είναι ένας δημοφιλής αλγόριθμος εκμάθησης συνόλου που βελτιώνει την απόδοση των μοντέλων μηχανικής μάθησης [7]. Λειτουργεί εκπαιδεύοντας διαδοχικά πολλαπλούς αδύναμους ταξινομητές, δίνοντας υψηλότερα βάρη σε εσφαλμένα ταξινομημένα στιγμιότυπα σε κάθε επανάληψη. Το τελικό μοντέλο είναι μια συνάθροιση αυτών των αδύναμων ταξινομητών, σταθμισμένα με βάση την ατομική τους απόδοση.

Το πρωταρχικό πλεονέκτημα του AdaBoostM1 έγκειται στην ικανότητά του να εστιάζει σε περιπτώσεις που είναι δύσκολο να ταξινομηθούν, βελτιώνοντας αποτελεσματικά τη συνολική ακρίβεια του μοντέλου συνόλου. Προσαρμόζοντας επαναληπτικά τα βάρη των εσφαλμένων ταξινομημένων παρουσιών, το AdaBoostM1 μπορεί να δώσει προτεραιότητα στη μάθηση από δύσκολα σημεία δεδομένων, οδηγώντας σε ένα πιο ισχυρό και ακριβές μοντέλο.

Το AdaBoostM1 χρησιμοποιείται και ελέγχεται με διάφορες παραμέτρους, όπως useResampling (True, False), weight Threshold (60 - 100), Iteration (5 - 50) και Classifier

(Random Tree, Decision Stump, Hoeffding Tree, RepTree). Ο συντονισμός παραμέτρων παρατηρείται στο επιλεγμένο σύνολο δεδομένων χαρακτηριστικών και εφαρμόζεται και στα 3 σύνολα δεδομένων που συζητήθηκαν προηγουμένως. Η διαδικασία και τα αποτελέσματα φαίνονται στον παρακάτω πίνακα 12:

Πίνακας 12: Συντονισμός υπερπαραμέτρων του ταξινομητή AdaBoostM1

Parameter	F-Measure
default	0.683
useResampling = True	0.687
useResampling = True, weightThreshold = 80	0.655
useResampling = True, weightThreshold = 60	0.655
useResampling = True, Iteration = 5	0.677
useResampling = True, Iteration = 20	0.713
useResampling = True, Iteration = 50	0.713
useResampling = True, Iteration = 50, Classifier = Hoeffding Tree	0.704
useResampling = True, Iteration = 50, Classifier = Random Tree	0.662

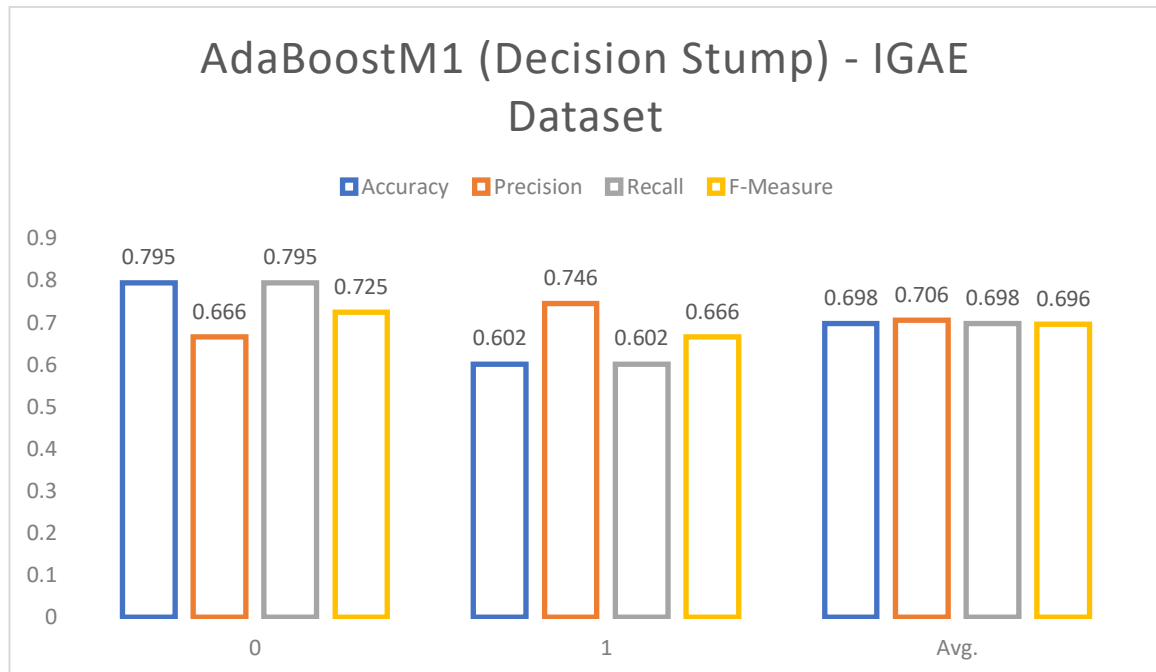
Το καλύτερο σύνολο παραμέτρων που βρέθηκε είναι useResampling = True, Iteration = 20. Ο Πίνακας 13 δείχνει την απόδοση των αλγορίθμων που χρησιμοποιούν τις καλύτερες παραμέτρους:

Πίνακας 13: Απόδοση του AdaBoostM1 χρησιμοποιώντας τον ταξινομητή DecisionStump

Dataset	Accuracy (weighted avg.)	Precision (weighted avg.)	Recall (weighted avg.)	F-Measure (weighted avg.)
Raw	0.817	0.799	0.817	0.791

Raw Processed	0.698	0.704	0.698	0.695
IGAE	0.698	0.706	0.698	0.696

Το σχήμα 10 δείχνει την απόδοση του AdaBoostM1 στο σύνολο δεδομένων IGAE:



Κεφάλαιο 4ο - Συμπεράσματα, Αποτελέσματα & Σχολιασμός

4. Ανάλυση αποτελεσμάτων (Result Analysis)

Οι 6 διαφορετικοί αλγόριθμοι χρησιμοποιήθηκαν πάνω από τα σύνολα δεδομένων για να συγκριθεί η απόδοση των αλγορίθμων και να βρεθεί ο καλύτερος από αυτούς. Ο Πίνακας 14 δείχνει τη σταθμισμένη μέση απόδοση (weighted average performance) όλων των ταξινομητών στα διαφορετικά σύνολα δεδομένων:

Πίνακας 14: weighted average αποτελέσματα από τους αλγόριθμους

Algorithm	Accuracy (weighted avg.)	Precision (weighted avg.)	Recall (weighted avg.)	F-Measure (weighted avg.)
Raw Dataset				
Random Forest	0.816	0.798	0.816	0.787
Decision Tree	0.817	0.799	0.817	0.795
KNN	0.809	0.788	0.809	0.786
SVM	0.780	0.729	0.780	0.692
Bagging	0.820	0.803	0.820	0.798
AdaBoostM1	0.817	0.799	0.817	0.791
Raw Preprocessed Dataset				
Random Forest	0.709	0.721	0.709	0.725
Decision Tree	0.694	0.696	0.694	0.693
KNN	0.672	0.683	0.672	0.667
SVM	0.709	0.718	0.709	0.706
Bagging	0.701	0.704	0.701	0.700
AdaBoostM1	0.698	0.704	0.698	0.695
IGAE Dataset				
Random Forest	0.706	0.722	0.706	0.721
Decision Tree	0.696	0.699	0.696	0.694
KNN	0.672	0.683	0.672	0.667
SVM	0.706	0.716	0.706	0.702

Bagging	0.702	0.706	0.702	0.701
AdaBoostM1	0.698	0.706	0.698	0.696

Από τον πίνακα 8 μπορεί να παρατηρηθεί ότι όλοι οι αλγόριθμοι που εμφανίζουν υψηλότερη βαθμολογία στο ακατέργαστο σύνολο δεδομένων (raw dataset) και το Random Forest εμφανίζουν το καλύτερο αποτέλεσμα μεταξύ τους. Όμως, καθώς η τάξη είναι ανισορροπημένη, δεν είναι δυνατή η βάση του σταθμισμένου μέσου όρου βαθμολογίας (weighted average score) και τίποτα δεν μπορεί να αποφασιστεί από αυτά τα αποτελέσματα. Βλεποντας την απόδοση της τάξης σε βάθος για κάθε τάξη για να δω αν υπάρχουν διαφορές. Ο Πίνακας 15 δείχνει τη σε βάθος απόδοση κατηγορίας των ταξινομητών.

Πίνακας 15: Αποτελέσματα πρόβλεψης σε βάθος κλάσης των αλγορίθμων

Algorithm	Class	Accuracy	Precision	Recall	F-Measure
Raw Dataset					
Random Forest	0	0.960	0.830	0.960	0.890
	1	0.308	0.685	0.308	0.425
	Avg.	0.816	0.798	0.816	0.787
Decision Tree	0	0.947	0.838	0.947	0.890
	1	0.357	0.659	0.357	0.463
	Avg.	0.817	0.799	0.817	0.795
KNN	0	0.944	0.833	0.944	0.885
	1	0.334	0.631	0.334	0.437
	Avg.	0.809	0.788	0.809	0.786
SVM	0	0.994	0.782	0.994	0.875
	1	0.025	0.542	0.024	0.027
	Avg.	0.780	0.729	0.780	0.692
Bagging	0	0.951	0.840	0.951	0.892
	1	0.361	0.674	0.361	0.470
	Avg.	0.820	0.803	0.820	0.789
AdaBoostM1	0	0.955	0.834	0.955	0.890
	1	0.330	0.676	0.330	0.443
	Avg.	0.817	0.799	0.817	0.791
Raw Preprocessed Dataset					
	0	0.828	0.669	0.828	0.740

Random Forest	1	0.590	0.774	0.559	0.669
	Avg.	0.709	0.721	0.709	0.725
Decision Tree	0	0.747	0.675	0.747	0.709
	1	0.640	0.717	0.640	0.676
	Avg.	0.694	0.696	0.694	0.693
KNN	0	0.794	0.638	0.794	0.708
	1	0.550	0.728	0.550	0.627
	Avg.	0.672	0.683	0.672	0.667
SVM	0	0.809	0.674	0.809	0.736
	1	0.609	0.762	0.609	0.677
	Avg.	0.709	0.718	0.709	0.706
Bagging	0	0.763	0.679	0.763	0.718
	1	0.639	0.729	0.639	0.681
	Avg.	0.701	0.704	0.701	0.700
AdaBoostM1	0	0.787	0.668	0.787	0.723
	1	0.608	0.741	0.608	0.668
	Avg.	0.698	0.704	0.698	0.695
IGAE Dataset					
Random Forest	0	0.840	0.662	0.840	0.741
	1	0.572	0.782	0.572	0.660
	Avg.	0.706	0.722	0.706	0.721
Decision Tree	0	0.759	0.674	0.759	0.714
	1	0.632	0.724	0.632	0.675
	Avg.	0.696	0.699	0.696	0.694
KNN	0	0.794	0.639	0.794	0.708
	1	0.550	0.728	0.550	0.637
	Avg.	0.672	0.683	0.672	0.667
SVM	0	0.815	0.669	0.815	0.735
	1	0.596	0.734	0.596	0.669
	Avg.	0.706	0.716	0.706	0.702
Bagging	0	0.768	0.678	0.768	0.721
	1	0.636	0.733	0.636	0.681

	Avg.	0.702	0.706	0.702	0.701
AdaBoostM1	0	0.795	0.666	0.795	0.725
	1	0.602	0.746	0.602	0.666
	Avg.	0.698	0.706	0.698	0.696

Έτσι, τα αποτελέσματα υποδεικνύουν ότι ο αλγόριθμος Random Forest και SVM έχει γενικά καλύτερη απόδοση από άλλους αλγόριθμους σε όλα τα σύνολα δεδομένων. Επιτυγχάνουν υψηλότερη average accuracy, precision, recall και F-measure, υποδεικνύοντας την αποτελεσματικότητά τους στον χειρισμό της εργασίας ταξινόμησης. Άλλοι αλγόριθμοι έχουν επίσης αρκετά καλή απόδοση, παρουσιάζοντας ανταγωνιστικά αποτελέσματα με το Random Forest και το SVM σε ορισμένες περιπτώσεις. Ωστόσο, το KNN τείνει να έχει ελαφρώς χαμηλότερη απόδοση σε σύγκριση με τους άλλους αλγόριθμους στις περισσότερες περιπτώσεις.

Σε μια γενική άποψη, οι αλγόριθμοι έχουν εξαιρετική απόδοση στο ακατέργαστο σύνολο δεδομένων, αλλά κάθε φορά που προσπαθούσαμε να δούμε την απόδοση του επιπέδου τάξης, υπήρχε κάτι ενδιαφέρον. Στο ακατέργαστο σύνολο δεδομένων, όλοι οι ταξινομητές προβλέπουν με ακρίβεια άνω του 90% για την κλάση μη σφάλματος -«0» αλλά μόλις 35% για την κλάση σφάλματος «1», γεγονός που υποδεικνύει μια τεράστια προκατάληψη του μοντέλου. Ο σκοπός αυτής της εργασίας είναι ο εντοπισμός των πιστωτικών σφαλμάτων, επομένως ο εντοπισμός των ελαττωματικών συναλλαγών είναι πιο σημαντικός από τις συναλλαγές χωρίς σφάλμα. Επιπλέον, το πρώτο και κύριο χαρακτηριστικό ενός καλού μοντέλου είναι η γενίκευση. Επομένως, είναι σημαντικό να έχουμε ένα γενικευμένο μοντέλο και όχι ένα μοντέλο υψηλής απόδοσης. Έτσι, η πρόβλεψη για το ακατέργαστο σύνολο δεδομένων δεν φαίνεται καλή.

Για το προεπεξεργασμένο και το σύνολο δεδομένων IGAE, τα αποτελέσματα είναι αυστηρά εδώ, η διαφορά απόδοσης των κλάσεων έχει μειωθεί πολύ. Το σφάλμα: κλάση «1» εμφανίζει το f-measure μεταξύ 55-64% με αυτά τα σύνολα δεδομένων που υποδηλώνει ότι τα προβλήματα ανισορροπίας έχουν επιλυθεί και τώρα τα μοντέλα είναι λιγότερο προκατειλημμένα και γενικευμένα και έτοιμα για χρήση. Κάποιοι από τους αλγόριθμους είχαν

καλύτερη απόδοση σε ακατέργαστα προεπεξεργασμένα και κάποιοι καλύτερα στο σύνολο δεδομένων IGAE. Σε αυτή τη μελέτη, το σύνολο δεδομένων IGAE επιλέχθηκε έναντι των άλλων λόγω της σχετικά μικρής διαφοράς απόδοσης περίπου 0,1% σε σύγκριση με τα σύνολα δεδομένων Raw Pre-επεξεργασμένα. Επιπλέον, το σύνολο δεδομένων IGAE προσέφερε το πρόσθετο πλεονέκτημα της μείωσης του αριθμού των χαρακτηριστικών κατά 24%, καθιστώντας το μια ευνοϊκή επιλογή για την επίτευξη συγκρίσιμων αποτελεσμάτων με ένα πιο συνοπτικό σύνολο χαρακτηριστικών.

Μεταξύ των αλγορίθμων που δοκιμάστηκαν, το Random Forest και το SVM πέτυχαν την υψηλότερη συνολική ακρίβεια και F-Measure. Ωστόσο, και τα δύο μοντέλα αντιμετώπισαν προκλήσεις στη γενίκευση, ειδικά όταν ασχολούνταν με την Κλάση 1 (δόλιες συναλλαγές). Το F-Measure του Random Forest για την Κλάση 1 ήταν μόνο 66,0%, και το F-Measure του SVM για την Κλάση 1 ήταν μόνο 66,9%, υποδεικνύοντας σημαντικές δυσκολίες στον ακριβή εντοπισμό δόλιων δραστηριοτήτων. Από την άλλη πλευρά, το Decision Tree είχε καλή απόδοση όσον αφορά τη γενίκευση. Ωστόσο, η απόδοσή του για την Κλάση 0 ήταν συγκριτικά χαμηλότερη. Ωστόσο, το Δέντρο Αποφάσεων έδειξε ανταγωνιστική γενικευμένη απόδοση και πέτυχε καλή βαθμολογία για περιπτώσεις απάτης. Το KNN έδειξε τη χαμηλότερη απόδοση για την Κλάση 1, υποδεικνύοντας τους περιορισμούς τους στον σωστό προσδιορισμό των νόμιμων συναλλαγών.

Είναι ενδιαφέρον ότι ο ταξινομητής **Bagging with RepTree** αναδείχθηκε ως το μοντέλο με τις καλύτερες επιδόσεις όσον αφορά το F-Measure, πετυχαίνοντας ένα εντυπωσιακό F-Measure 72,1% για **non fraud** και 68,1% για περιπτώσεις απάτης. Ξεπέρασε τους άλλους αλγόριθμους, συμπεριλαμβανομένων των Random Forest και SVM, όσον αφορά την επίτευξη ισορροπίας μεταξύ ακρίβειας και ανάκλησης και για τις δύο κατηγορίες. Επιπλέον, ο αλγόριθμος AdaBoostM1 και Decision Tree έδειξε επίσης πολλά υποσχόμενα γενικευμένα αποτελέσματα. Αν και δεν ξεπέρασε τις επιδόσεις του Bagging με τον ταξινομητή RepTree, η περαιτέρω βελτίωση και βελτιστοποίηση θα μπορούσαν ενδεχομένως να βελτιώσουν την απόδοσή τους. Λαμβάνοντας υπόψη την κρίσιμη φύση του εντοπισμού απάτης με πιστωτικές κάρτες, η εστίαση στη μέτρηση F-Measure είναι απαραίτητη για την αξιολόγηση της ικανότητας των

αλγορίθμων να χειρίζονται τις ανισορροπίες κλάσεων και να ελαχιστοποιούν τα ψευδώς αρνητικά και τα ψευδώς θετικά.

Συμπερασματικά, μεταξύ των έξι αλγορίθμων που δοκιμάστηκαν, ο ταξινομητής **Bagging with RepTree** αποδείχθηκε το πιο ισχυρό και ισορροπημένο μοντέλο για τον εντοπισμό απάτης με πιστωτικές κάρτες, ακολουθούμενο από το **Decision Tree** και το **AdaBoostM1**. Ενώ το Random Forest και το SVM πέτυχαν την υψηλότερη συνολική ακρίβεια και το F-Measure, αντιμετώπισαν προκλήσεις στη γενίκευση, ιδιαίτερα για την Κλάση 1. Το Decision Tree είχε καλή απόδοση όσον αφορά τη γενίκευση, παρά τη συγκριτικά χαμηλότερη ακρίβειά του για την Κλάση 0.

4.1. Συμπέρασμα (Conclusion)

Η μελέτη διερεύνησε την απόδοση έξι αλγορίθμων μηχανικής μάθησης (Random Forest, Decision Tree, KNN, SVM, Bagging, AdaBoostM1) για το κρίσιμο έργο του εντοπισμού απάτης με πιστωτικές κάρτες. Η μελέτη έδειξε τρεις διαφορετικές παραλλαγές του συνόλου δεδομένων, συμπεριλαμβανομένου του Raw Dataset, Raw Preprocessed Dataset και Information Gain Attribute Selection Dataset, για την αξιολόγηση της ευρωστίας των μοντέλων υπό διάφορες συνθήκες. Τα ευρήματά αποκάλυψαν ότι ο αλγόριθμος Bagging αναδείχθηκε ως η πιο πολλά υποσχόμενη επιλογή για τον εντοπισμό απάτης με πιστωτικές κάρτες. Πέτυχε με συνέπεια το υψηλότερο F-Measure σε όλες τις κατηγορίες απάτης, διαχειριζόταν επίσης ισορροπημένη και ακριβή απόδοση για τον εντοπισμό τόσο νόμιμων όσο και δόλιων συναλλαγών. Επιπλέον, ο ταξινομητής Bagging with RepTree έδειξε καλύτερη γενίκευση σε σύγκριση με άλλους αλγόριθμους, καθιστώντας τον κατάλληλο μοντέλο για αυτό το πρόβλημα ταξινόμησης. Ωστόσο, παρατηρείται επίσης ότι οι διαφορές απόδοσης μεταξύ των έξι αλγορίθμων ήταν σχετικά μικρές, υποδηλώνοντας ότι κάθε μοντέλο είναι ικανό να επιτύχει ανταγωνιστικά αποτελέσματα στον εντοπισμό απάτης πιστωτικών καρτών. Επομένως, ανάλογα με τις συγκεκριμένες απαιτήσεις και τους διαθέσιμους πόρους, οι ερευνητές και οι επαγγελματίες μπορούν να επιλέξουν το καταλληλότερο μοντέλο με βάση άλλες εκτιμήσεις,

όπως η υπολογιστική αποτελεσματικότητα και η ερμηνευτικότητα.

Παρά τα πολλά υποσχόμενα αποτελέσματα, υπάρχουν αρκετοί τρόποι για μελλοντική εργασία στον εντοπισμό απάτης με πιστωτικές κάρτες. Πρώτον, η διερεύνηση πρόσθετων τεχνικών προεπεξεργασίας δεδομένων θα μπορούσε ενδεχομένως να βελτιώσει περαιτέρω την απόδοση του μοντέλου. Επιπλέον, η ενσωμάτωση μεθόδων συνόλου, όπως η στοίβαξη, θα μπορούσε να είναι ευεργετική για την ενίσχυση της συνολικής προγνωστικής ισχύος. Ο συνδυασμός των δυνατοτήτων πολλών μοντέλων μπορεί να οδηγήσει σε ακόμη πιο ισχυρές δυνατότητες ανίχνευσης απάτης. Επιπλέον, καθώς τα μοτίβα απάτης με πιστωτικές κάρτες εξελίσσονται συνεχώς, είναι σημαντικό να διατηρηθεί η απόδοση του μοντέλου με την πάροδο του χρόνου. Η εφαρμογή μιας προσαρμοστικής προσέγγισης μάθησης ή η ανάπτυξη συστημάτων παρακολούθησης σε πραγματικό χρόνο μπορεί να βοηθήσει το μοντέλο να προσαρμοστεί σε νέα πρότυπα απάτης και να διασφαλίσει τη συνεχή ακρίβεια.

Κεφάλαιο 5^ο - Η Εφαρμογή

5. To Application της Πτυχιακής

Credit Card Fraud Detector

a university thesis

Choose one of the values below:

Sex:

Education:

Marriage:

Εντοπισμός δόλιας συναλλαγής πιστωτικής κάρτας η εφαρμογή

Η εφαρμογή "Credit Card Fraud Detector" έχει σχεδιαστεί για να βοηθήσει να ανιχνεύσετε εάν μια συναλλαγή με πιστωτική κάρτα είναι δολιά ή όχι. Ο χρήστης ακολουθεί τις παρακάτω οδηγίες για να χρησιμοποιήσει την εφαρμογή και να καταλάβει εάν οι συναλλαγές της πιστωτικής του κάρτας είναι δόλιες.

5.1 Κατηγορία Εφαρμογής: “Choose one of the values below”

- Ο χρήστης ανοίγει την εφαρμογή "Credit Card Fraud Detector" στη συσκευή του.
- Στην αρχική οθόνη, θα δεί πρώτα απο όλα τρία πεδία για την εισαγωγή δεδομένων της συναλλαγής: "Sex" (φύλο), "Education" (εκπαίδευση), και "Marriage" (γάμος).

Choose one of the values below:

Sex:

male

Education:

University

Marriage:

yes

yes

no

Συμπλήρωση των σημείων με τα δεδομένα της συναλλαγής

- Για το πεδίο "Sex" (φύλο), επιλέξτε αν είστε "male" (άνδρας) ή "female" (γυναίκα).
- Για το πεδίο "Education" (εκπαίδευση), επιλέξτε το επίπεδο εκπαίδευσής σας: "university" (πανεπιστήμιο), "high school" (λύκειο), ή "Elementary school" (δημοτικό).
- Για το πεδίο "Marriage" (γάμος), επιλέξτε "yes" (ναι) εάν είστε παντρεμένοι, αλλιώς επιλέξτε "no" (όχι).

5.2 Κατηγορία Εφαρμογής: "History of past payment"

- Για κάθε έναν από τους μήνες September, August, July, June, May, April, μπορείτε να εισαγάγετε μια τιμή από το -2 έως το 8. Αυτές οι τιμές αντιπροσωπεύουν την κατάσταση των προηγούμενων πληρωμών σας για τον αντίστοιχο μήνα.

The screenshot shows a form titled "History of past payment" in green text. Below the title, there are six rows, each for a month. Each row consists of a label (e.g., "September:") and a text input field. The values entered in the input fields are: 5 for September, 2 for August, 5 for July, 6 for June, 8 for May, and 3 for April.

Month	Value
September:	5
August:	2
July:	5
June:	6
May:	8
April:	3

- Η τιμή 0 σημαίνει ότι δεν υπήρχαν προβλήματα με τις πληρωμές σας για αυτόν τον μήνα.
- Η τιμή -2 μπορεί να χρησιμοποιηθεί εάν υπήρξαν δολιες πληρωμές για αυτόν τον μήνα.
- Η τιμή 8 μπορεί να χρησιμοποιηθεί εάν οι πληρωμές ήταν εξαιρετικά καλές για αυτόν τον μήνα.
- Συμπληρώστε τις τιμές για κάθε μήνα με βάση την ιστορία των πληρωμών σας και την εμπειρία σας

5.3 Κατηγορία Εφαρμογής: "Write your answer in the fields below"

Write your answer in the fields below:

age:

amount of the given credit:

Επειτα ο χρήστης συμπληρώνει τα δύο πεδία ως εξής:

- Στο πεδίο "**age**" (ηλικία), εισάγετε την ηλικία σας. Αυτή η πληροφορία βοηθά στον έλεγχο της αντισταθμιστικής ικανότητας.
- Στο πεδίο "**amount of the given credit**" (ποσό του δανείου που παρέχεται), εισάγετε το ποσό του δανείου που σχετίζεται με την εκάστοτε συναλλαγή.

5.4 Κατηγορία Εφαρμογής: "Amount of previous payments"

amount of previous payments

September:

August:

July:

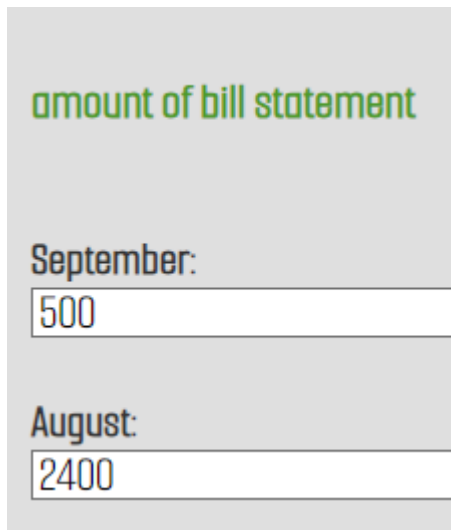
June:

May:

April:

- Για κάθε έναν από τους μήνες September, August, July, June, May, και April, ο χρήστης μπορεί να εισάγει το αντίστοιχο ποσό που πληρώσε για τον εκάστοτε μήνα.
- Μπορεί να χρησιμοποιήσει αριθμούς για να αντιπροσωπεύσει τα ποσά. Ο χρήστης χρειάζεται να συμπληρώσει τα πεδία με προσοχή και ακρίβεια.

5.5 Κατηγορία Εφαρμογής: "Amount of bill statement"



amount of bill statement

September:

500

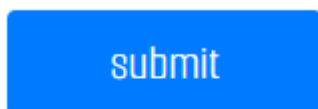
August:

2400

- Σε αυτή την κατηγορία, πρέπει να απαντήσει ο χρήστης σχετικά με τα ποσά των ανακοινώσεών σας για τους μήνες September και August.
- Για κάθε έναν από τους μήνες September και August, πρέπει ο χρήστης να εισαγει το ποσό της ανακοίνωσης πληρωμής που ελαβε για τον εκάστοτε μήνα.
- Ο χρήστης μπορεί να χρησιμοποιήσει αριθμούς για να αναπαραστήσει τα ποσά. Ο χρήστης οφείλει να συμπληρώσει τα πεδία με προσοχή και ακρίβεια.

5.6 "Submit και Show Results"

- Αφού ο χρήστης έχει συμπληρώσει όλα τα απαραίτητα πεδία σχετικά με τη συναλλαγή πιστωτικής κάρτας, πατάει το κουμπί "Submit" (Υποβολή) για να εκτελέσει το μοντέλο.
- Αφού πατήσει το κουμπί "Submit," το μοντέλο θα τρέξει και θα αξιολογήσει τη συναλλαγή πιστωτικής κάρτας.



submit

Success!

- Μετά τον υπολογισμό, θα εμφανιστεί ένα μήνυμα που θα δηλώνει εάν η διαδικασία εκτέλεσης του μοντέλου ήταν επιτυχής ("Success") ή όχι.

- Αν η διαδικασία είναι επιτυχής, ο χρήστης πρέπει να πατήσει το κουμπί "Show Results" (Εμφάνιση Αποτελεσμάτων) για να δει τα αποτελέσματα της ανίχνευσης δόλου.
- Στην οθόνη θα εμφανίζεται το αποτέλεσμα της ανίχνευσης, π.χ., "Η συναλλαγή σας είναι δόλια" ή "Η συναλλαγή σας δεν είναι δόλια με ποσοστό 80%."

show results

Ο αλγόριθμος με τις τιμές που εισήγαγε ο χρήστης έχει ως αποτέλεσμα ότι η συναλλαγή δεν είναι δόλια κατά 0.725%

- Εάν εμφανίζεται ένα ποσοστό, αυτό αντιπροσωπεύει το επίπεδο πιθανότητας της συναλλαγής να μην είναι δόλια. ο χρήστης μπορεί να χρησιμοποιήσει αυτή την πληροφορία για να λάβει αποφάσεις.
- Αν η διαδικασία δεν είναι επιτυχής, θα πρέπει να εξετάσει τα στοιχεία που καταχώρησε και να προσπαθήσει να τα διορθώσει.

5.7 Ανάλυση της Εφαρμογής "Credit Card Fraud Detector"

Η εφαρμογή "Credit Card Fraud Detector" αποτελεί ένα ισχυρό εργαλείο για τον χρήστη που τον βοηθά στον εντοπισμό δόλιων συναλλαγών χρησιμοποιώντας τεχνικές μηχανικής μάθησης.

Εξατομίκευση της Ανίχνευσης

Η εφαρμογή επιτρέπει στον χρήστη να εισαγάγει διάφορα στοιχεία που αφορούν τη συναλλαγή του, όπως η ηλικία, το ποσό του δανείου, τα ποσά των προηγούμενων πληρωμών και τα ποσά

των ανακοινώσεων. Αυτή η εξατομίκευση επιτρέπει στο μοντέλο μηχανικής μάθησης να προσαρμόσει την ανίχνευση δόλιων συναλλαγών σύμφωνα με τη συγκεκριμένη κατάσταση του χρήστη.

Αυξημένη Ακρίβεια

Η χρήση της μηχανικής μάθησης επιτρέπει στην εφαρμογή να αναλύσει πολλαπλά χαρακτηριστικά της συναλλαγής και να προβλέψει την πιθανότητα δόλιας συναλλαγής με υψηλή ακρίβεια.

Γρήγορη Αξιολόγηση

Ο χρήστης μπορεί να λάβει γρήγορη αξιολόγηση για τη συναλλαγή του μετά την υποβολή των στοιχείων. Αυτό εξοικονομεί χρόνο και επιτρέπει τη λήψη άμεσων αποφάσεων.

Προοπτικές Χρήσης

Ανίχνευση Δόλιων Συναλλαγών

Η κύρια χρήση της εφαρμογής είναι η ανίχνευση δόλιων συναλλαγών σε πιστωτικές κάρτες. Οι χρήστες μπορούν να εισαγάγουν τα στοιχεία της συναλλαγής τους και να λάβουν γρήγορη αξιολόγηση για την εγκυρότητά της.

Βελτιστοποίηση Οικονομικών Επιλογών

Οι χρήστες μπορούν να χρησιμοποιήσουν την εφαρμογή για να αξιολογήσουν τις οικονομικές τους επιλογές και να λάβουν αποφάσεις βασισμένες σε αναλύσεις από το μοντέλο μηχανικής μάθησης.

Εκπαίδευση στην Οικονομική Ευφυΐα

Η εφαρμογή μπορεί να χρησιμοποιηθεί ως εκπαιδευτικό εργαλείο για να βοηθήσει τους χρήστες να κατανοήσουν τους παράγοντες που επηρεάζουν τις χρηματιστηριακές αποφάσεις και την ανίχνευση δόλιων συναλλαγών.

Συνοψίζοντας, η εφαρμογή "Credit Card Fraud Detector" παρέχει στους χρήστες μια ισχυρή εργαλειοθήκη για την ανίχνευση δόλιων συναλλαγών και την βελτιστοποίηση των οικονομικών τους επιλογών. Αυτή η εφαρμογή μπορεί να βοηθήσει στην προστασία των

οικονομικών και προσωπικών πληροφοριών του χρήστη, καθιστώντας την ιδιαίτερα χρήσιμη για την ασφάλεια και την οικονομική του ευεξία.

Κεφάλαιο 6^ο - Βιβλιογραφία

6. Πηγές - Αναφορές

- [1]. Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019). Credit Card Fraud Detection - Machine Learning methods. *18th International Symposium INFOTEH-JAHORINA*. <https://doi.org/10.1109/infoteh.2019.8717766>
- [2]. Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S., & Kuruwitaarachchi, N. (2019). Real-time Credit Card Fraud Detection Using Machine Learning. 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). <https://doi.org/10.1109/confluence.2019.8776942>
- [3]. Dornadula, V. N., & Geetha, S. (2019). Credit Card Fraud Detection using Machine Learning Algorithms. *Procedia Computer Science*, 165, 631–641. <https://doi.org/10.1016/j.procs.2020.01.057>
- [4]. Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the GA algorithm for feature selection. *Journal of Big Data*, 9(1). <https://doi.org/10.1186/s40537-022-00573-8>
- [5]. Ali, A., Razak, S. A., Othman, S. H., Eisa, T. a. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial Fraud Detection Based on Machine Learning: A Systematic Literature review. *Applied Sciences*, 12(19), 9637. <https://doi.org/10.3390/app12199637>
- [6]. Berrar, D. (2019). Cross-Validation. In *Elsevier eBooks* (pp. 542–545). <https://doi.org/10.1016/b978-0-12-809633-8.20349-x>

[7]. Kalhotra, S. K., Dongare, S. V., Kasthuri, A., & Kaur, D. (2022). Data mining and machine learning techniques for credit card fraud detection. *ECS Transactions*, 107(1), 4977–4985. <https://doi.org/10.1149/10701.4977ecst>

[8] Shmatko, O., Fedorchenko, V., & Prochukhan, D. (2021). DETECTING CREDIT CARD FRAUD USING MACHINE LEARNING ALGORITHMS. *InterConf*, 393–403. <https://doi.org/10.51582/interconf.19-20.08.2021.037>

[9] Almuteer, A. H., Aloufi, A. A., Alrashidi, W. O., Alshobaili, J. F., & Ibrahim, D. M. (2021). Detecting Credit Card Fraud using Machine Learning. *International Journal of Interactive Mobile Technologies*, 15(24), 108–122. <https://doi.org/10.3991/ijim.v15i24.27355>