ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΜΣ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ

ΚΑΤΕΥΘΥΝΣΗ: ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ

Ανίχνευση και πρόβλεψη απάτης στο λιανικό εμπόριο μέσω της χρήσης αλγορίθμων μηχανικής και βαθιάς μάθησης με στόχο τη βελτίωση της εμπειρίας του πελάτη.

Κούγιας Κωνσταντίνος – Νεκτάριος

ΜΕ2018

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ : Δ. Κυριαζής

ΠΕΙΡΑΙΑΣ, ΦΕΒΡΟΥΑΡΙΟΣ 2022

UNIVERSITY OF PIRAEUS

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

DEPARTEMNT OF DIGITAL SYSTEMS

POSTGRADUATE POGRAMME: "INFORMATION SYSTEMS & SERVICES"

DIRECTION: BIG DATA AND ANALYTICS

Fraud detection in retail transactions through the use of machine and deep learning algorithms in order to improve the customer experience.

Kougias Konstantinos – Nektarios

ME2018

SUPERVISOR PROFESSOR : D. Kyriazis

PIRAEUS, FEBRUARY 2022

Περίληψη

Τα τελευταία χρόνια εξαιτίας της ραγδαίας εξέλιξης της σύγχρονης τεχνολογίας και των προηγμένων μεθόδων επικοινωνίας, τα περιστατικά απάτης αυξάνονται διαρκώς σε ανησυχητικό βαθμό επιφέροντας τεράστιες οικονομικές απώλειες σε όλο τον κόσμο. Η απάτη έχει γίνει παγκόσμιο φαινόμενο που προκαλεί ανησυχία στον επιχειρηματικό κλάδο και επηρεάζει όλους του τύπους των επιχειρήσεων ανεξαρτήτως μεγέθους, κερδοφορίας ή κλάδου δραστηριότητας. Γι' αυτόν το λόγο οι σύγχρονες επιχειρήσεις καλούνται να ανιχνεύσουν και να προβλέψουν τον κίνδυνο απάτης σε κάθε μορφή με στόχο τη διασφάλιση των εσόδων τους και τη διατήρηση της αξιοπιστίας τους. Η ανίχνευση απάτης είναι μία διαδικασία εντοπισμού και ανακάλυψης κακόβουλων ενεργειών και πρακτικών.

Η Μηχανική Μάθηση και οι τεχνικές Εξόρυξης Δεδομένων συμβάλλουν καθοριστικά στον εντοπισμό και στην πρόβλεψη απάτης και έχουν εφαρμοστεί με επιτυχία στην ανίχνευση παράνομων δραστηριοτήτων όπως σε συναλλαγές με πιστωτικές κάρτες, στο διαδικτυακό εμπόριο, στο ξέπλυμα χρήματος, αλλά και σε περιπτώσεις απάτης κλάδων όπως τον ασφαλιστικό, τον κλάδο τηλεπικοινωνιών καθώς και τον ιατρικό και επιστημονικό κλάδο. Ειδικότερα, η ανίχνευση απάτης σε συναλλαγές αποτελεί μια διαδικασία ανάλυσης και επεξεργασίας μεγάλου όγκου δεδομένων καθώς και δημιουργίας προβλεπτικών μοντέλων, μέσω εφαρμογής αλγορίθμων Μηχανικής Μάθησης, που οδηγεί στον εντοπισμό ύποπτων συναλλαγών. Αυτή η τεχνολογία βαθιάς εκμάθησης αναγνωρίζει και μαθαίνει από πολύπλοκα μοτίβα και συνδυάζοντας σημαντικά στοιχεία συναλλαγών των χρηστών, από διάφορα κανάλια πωλήσεων, έχει την ικανότητα να ταξινομήσει εάν μια συναλλαγή είναι παράνομη ή νόμιμη.

Στόχος της παρούσας διπλωματικής εργασίας είναι η ανίχνευση και η πρόβλεψη ενδεχόμενης απάτης σε δεδομένα συναλλαγών που προέρχονται από τον κλάδο του λιανικού εμπορίου, μέσω τεχνικών εξόρυξης δεδομένων και χρήσης αλγορίθμων μηχανικής και βαθιάς μάθησης.

Λέξεις Κλειδιά: Ανίχνευση – Πρόβλεψη Απάτης, Λιανικό Εμπόριο, Εξόρυξη Δεδομένων, Αλγόριθμοι Μηχανικής Μάθησης, Κατηγοριοποίηση.

Abstract

In recent years, due to the rapid development of modern technology and advanced communication methods, fraud cases are constantly increasing at an alarming rate, resulting in huge financial losses around the world. Fraud has become a global phenomenon of concern to the business sector and affects all types of businesses regardless of size, profitability or industry. That's why modern businesses are called upon to detect and anticipate the risk of fraud in any form in order to secure their revenue and maintain their credibility. Detection of fraud is a process of detecting malicious actions and practices.

Machine Learning and Data Mining techniques are crucial in detecting and anticipating fraud and have been successfully applied to detect illegal activities such as credit card transactions, online commerce, money laundering, and fraudulent activities in industries such as insurance sector, the telecommunications sector as well as the medical and scientific sector. In particular, transaction fraud detection is a process of analysis and processing of large volumes of data as well as the creation of predictive models, through the application of Machine Learning algorithms, which leads to the detection of suspicious transactions. This deep learning technology recognizes and learns from complex patterns and combining important user transaction data from different sales channels, has the ability to classify whether a transaction is illegal or legal.

The aim of this master thesis is to detect and predict possible fraud in transaction data originating from the retail sector, through data mining techniques and the use of machine and deep learning algorithms.

Keywords: Fraud Detection - Prediction, Retail, Data Mining, Machine Learning Algorithms, Classification.

Ευχαριστίες

Με την ολοκλήρωση της μεταπτυχιακής διπλωματικής μου εργασίας θα ήθελα να ευχαριστήσω θερμά τους ανθρώπους που συνέβαλαν στην εκπόνησή της.

Αρχικά θα ήθελα να ευχαριστήσω τους υπεύθυνους της εταιρείας Qivos κα Φανή Χαρμπή και κ. Δημήτρη Σκόκα καθώς και τους συνεργάτες αυτών, οι οποίοι μεσολάβησαν για να αποκτήσω το dataset που χρησιμοποιήθηκε σε αυτήν την εργασία. Επιπλέον εξαιρετικά σημαντικές ήταν οι προτάσεις τους και οι διευκρινιστικές επισημάνσεις τους σχετικά με την πορεία της εργασίας.

Ιδιαίτερα θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Δημοσθένη Κυριαζή για την εμπιστοσύνη που μου έδειξε με την ανάθεση της διπλωματικής εργασίας αλλά και για το άρτιο κλίμα συνεργασίας που καλλιέργησε καθ' όλη τη διάρκεια εκπόνησής της.

Τέλος θα ήθελα να ευχαριστήσω την οικογένειά μου για τη συμπαράσταση και την κατανόηση που μου πρόσφερε σε όλη τη διάρκεια της ακαδημαϊκής μου πορείας.

Περιεχόμενα

Κατάλογος Εικόνων/Διαγραμμάτων

Κατάλογος Πινάκων

ΚΕΦΑΛΑΙΟ 1: Εισαγωγή

1.1 Δομή Μεταπτυχιακής Διπλωματικής Εργασίας

Η παρούσα διπλωματική εργασία διαρθρώνεται σε έξι (6) κεφάλαια ως εξής:

Στο πρώτο κεφάλαιο ορίζεται αρχικά η έννοια της απάτης και επισημαίνεται η συμβολή της Τεχνητής Νοημοσύνης και ειδικότερα της Μηχανικής Μάθησης στον εντοπισμό και στην πρόβλεψή της. Στη συνέχεια γίνεται αναφορά στη διαδικτυακή απάτη και την επιρροή της στον κλάδο λιανικού εμπορίου. Περιγράφονται επίσης οι συνηθέστερες μορφές διαδικτυακής απάτης και τονίζεται η ανάγκη του ψηφιακού μετασχηματισμού των επιχειρήσεων του κλάδου λιανεμπορίου, αναλύοντας τους λόγους που οδηγούν σε αυτήν την αναγκαιότητα.

Στο δεύτερο κεφάλαιο αναλύεται η έννοια των Μεγάλων Δεδομένων (Big Data) και παρουσιάζονται τα χαρακτηριστικά και οι τύποι δεδομένων καθώς και σημαντικοί τομείς εφαρμογής και αξιοποίησης αυτών. Επιπλέον το κεφάλαιο αυτό ασχολείται με την έννοια της Επιχειρηματικής Ευφυΐας και αναδεικνύονται τα οφέλη χρήσης της και τα διαφορετικά πεδία εφαρμογής της στις σύγχρονες επιχειρήσεις.

Το τρίτο κεφάλαιο αναφέρεται στην Εξόρυξη Δεδομένων και στα στάδια διαδικασίας ανακάλυψης γνώσης καθώς και στη Μηχανική Μάθηση όπου παρουσιάζονται τα βασικά είδη της και οι μέθοδοί της.

Στο τέταρτο κεφάλαιο γίνεται μια θεωρητική προσέγγιση αλγορίθμων μηχανικής μάθησης, αναλύεται η έννοια της Βαθιάς Μάθησης και περιγράφονται τα Τεχνητά Νευρωνικά Δίκτυα. Επιπλέον, το κεφάλαιο αυτό ασχολείται με τη Συλλογική Μάθηση και την ανάλυση της μεθόδου ενδυνάμωσης Boosting.

Στο πέμπτο κεφάλαιο παρουσιάζεται αναλυτικά η μεθοδολογία ανίχνευσης και πρόβλεψης απάτης που ακολουθήθηκε σε αυτήν την εργασία. Αρχικά παρουσιάζονται σχετικές μελέτες οι οποίες είχαν ασχοληθεί με το πρόβλημα της απάτης στο παρελθόν και στη συνέχεια ακολουθούν τα βήματα υλοποίησης, η περιγραφή των δεδομένων που χρησιμοποιήθηκαν, η διαδικασία της προεπεξεργασίας τους, τα αποτελέσματα που προκύπτουν από την υλοποίηση των αλγορίθμων που επιλέχθηκαν και τέλος η ερμηνεία της παραγόμενης γνώσης.

Τέλος στο έκτο κεφάλαιο γίνεται μια συγκριτική αξιολόγηση των αποτελεσμάτων που προκύπτουν από την εκτέλεση των αλγορίθμων, εξάγονται τα τελικά συμπεράσματα και παρατίθενται προτάσεις για μελλοντική επέκταση της εργασίας.

1.2 Ορισμός απάτης – Τεχνητή Νοημοσύνη & Ανίχνευση Απάτης

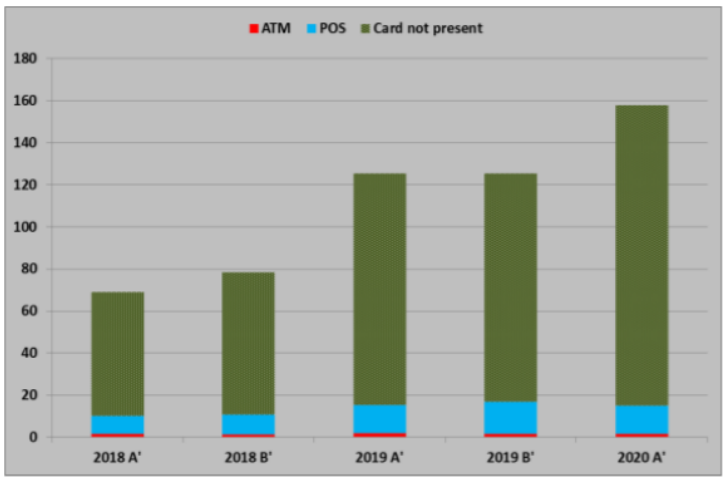
Σύμφωνα με το Διεθνή φορέα Institute of Internal Auditors (IIA) – Ινστιτούτο Εσωτερικών Ελεγκτών, ως απάτη (fraud) ορίζονται οποιεσδήποτε παράνομες πράξεις που χαρακτηρίζονται από δόλο, παραπλάνηση, απόκρυψη ή κατάχρηση εμπιστοσύνης. Αυτές οι πράξεις δεν εξαρτώνται από χρήση απειλής, βίας ή φυσικής δύναμης. Απάτες διαπράττονται από πρόσωπα και οργανισμούς προκειμένου να αποκτήσουν χρήματα, περιουσιακά στοιχεία ή υπηρεσίες για να αποφύγουν πληρωμές ή ζημιές ή να διασφαλίσουν ατομικό ή επιχειρηματικό συμφέρον [1]. Ως μία ευρύτερη

έννοια απάτη είναι μια πράξη εξαπάτησης που γίνεται για προσωπικό όφελος. Το φαινόμενο της απάτης συναντάται σε πάρα πολλές μορφές και απλώνεται σε όλα τα επίπεδα της επιχειρηματικής δραστηριότητας.

Η έξαρση της τεχνολογίας και της ψηφιοποίησης συμβάλλουν σημαντικά στην πρόληψη και την ανίχνευση απάτης. Συγκεκριμένα, η Τεχνητή Νοημοσύνη (Artificial Intelligence - AI) διαθέτει πολυάριθμες εφαρμογές που παρέχουν τη δυνατότητα εντοπισμού περιστατικών απάτης σε διάφορους τομείς και ιδίως όπου υπάρχουν προσωπικές και απόρρητες πληροφορίες, όπως σε πιστωτικές κάρτες. Η τεχνολογία Artificial Intelligence – AI χρησιμοποιείται κυρίως για αυτοματοποίηση εργασιών, ανάλυση συμπεριφοράς, εξυπηρέτηση πελατών και ανίχνευση απάτης. Με την πάροδο του χρόνου σημειώνονται όλο και περισσότερες επενδύσεις σε εταιρείες που ειδικεύονται στην τεχνολογία Artificial Intelligence [2]. Εξαιτίας του τεράστιου όγκου πληροφοριών και δεδομένων που αφορούν επιχειρηματικές συναλλαγές και που συνεχώς εξελίσσονται, ο εντοπισμός της απάτης είναι ιδιαίτερα περίπλοκος. Η Μηχανική Μάθηση (Machine Learning) ως κλάδος της Τεχνητής Νοημοσύνης, βασίζεται σε αλγορίθμους που μπορούν να εκπαιδεύονται από σύνολα δεδομένων και έχουν την ικανότητα να εξάγουν πρότυπα και να κατασκευάζουν μοντέλα ικανά για να κάνουν χρήσιμες προβλέψεις.

1.3 Διαδικτυακή Απάτη και Λιανικό Εμπόριο

Η ευρεία εξάπλωση και χρήση του διαδικτύου και η ψηφιακή επεξεργασία της κάθε πληροφορίας, έχουν επιφέρει επαναστατικές αλλαγές στην καθημερινή ζωή, στην παραγωγική διαδικασία και τις συναλλαγές, οδηγώντας σε ένα νέο περιβάλλον τόσο για τους καταναλωτές όσο και για τις επιχειρήσεις. Παρόλο που η πρόσβαση στο διαδίκτυο συνέβαλε στην άνθηση του επιχειρηματικού τομέα του ηλεκτρονικού εμπορίου, μετατρέποντας ολόκληρο τον κόσμο σε μια ενιαία παγκόσμια αγορά, παράλληλα μεγιστοποίησε τις δυνατότητες για διάπραξη νέων μορφών απάτης, προσφέροντας περισσότερες ευκαιρίες σε κακόβουλους δράστες να διαπράξουν ενέργειες απάτης μέσω διακαναλικών αγορών. Ιδιαίτερα κατά την τελευταία κρίση της πανδημίας Covid-19, οι περιορισμοί της φυσικής παρουσίας στα καταστήματα λιανικού εμπορίου, οδήγησαν σε απότομο άλμα του ηλεκτρονικού εμπορίου μεταξύ των χρηστών του διαδικτύου αλλά και ταυτόχρονη έξαρση της ηλεκτρονικής απάτης. Όπως αποκαλύπτει η Έκθεση Χρηματοπιστωτικής Σταθερότητας της Τράπεζας της Ελλάδος, έκρηξη της απάτης στις συναλλαγές με κάρτες πληρωμών καταγράφηκε το 2020. Η Τράπεζα της Ελλάδος αναλύοντας την απάτη ανά τύπο συναλλαγών με κάρτες πληρωμών όπως είναι οι συναλλαγές σε τερματικά ATM, οι πληρωμές σε τερματικά POS και οι εξ αποστάσεως συναλλαγές χωρίς τη φυσική παρουσία της κάρτας (card not present - CNP), διαπιστώνει ότι η πλειονότητα των περιστατικών απάτης εκδηλώνεται στις εξ αποστάσεως συναλλαγές και ειδικότερα στις συναλλαγές μέσω διαδικτύου [3].



Εικόνα 1.1 Διαχρονική εξέλιξη όγκου συναλλαγών απάτης ανά τύπο συναλλαγής.

Πηγή: Τράπεζα της Ελλάδος

Επιπλέον, εξαιτίας της πανδημίας Covid-19, αυξήθηκαν οι καταγγελίες και αναφορές στο Συνήγορο του Καταναλωτή σύμφωνα με την ετήσια έκθεση της Ανεξάρτητης Αρχής για το 2020. Ως επιταχυντής φαίνεται ότι λειτούργησε η πανδημία, η οποία έφερε τους Έλληνες πιο κοντά στις ηλεκτρονικές συναλλαγές αλλά και στις ηλεκτρονικές απάτες. Σύμφωνα με το Συνήγορο του Καταναλωτή, ο κλάδος των καταναλωτικών αγαθών λιανικού εμπορίου σημείωσε αύξηση του αριθμού των αναφορών που υπεβλήθησαν το 2020 σε σχέση με το 2019 κατά 75%. Η αύξηση των αναφορών οφείλεται στην εντατικοποίηση των ηλεκτρονικών αγορών, λόγω της επιβολής πρωτόγνωρων νομοθετικών περιορισμών στη φυσική λειτουργία των καταστημάτων κατά την περίοδο της πανδημίας [4] . Η απάτη στο ηλεκτρονικό εμπόριο γίνεται ολοένα και πιο ανεξέλεγκτη με εμπόρους και αγοραστές σε όλο τον κόσμο να πέφτουν θύματα. Το ραγδαία μεταβαλλόμενο και ανταγωνιστικό περιβάλλον που διαμορφώνεται μετά την κρίση της πανδημίας, αναμένεται να ωθήσει το λιανικό εμπόριο, που αποτελεί έναν από τους μεγαλύτερους κλάδους της οικονομίας παγκοσμίως, σε μεγάλες αλλαγές, με το ηλεκτρονικό εμπόριο να είναι η πιο σημαντική τάση που θα επηρεάσει το συγκεκριμένο κλάδο. Σύμφωνα με πρόσφατη έρευνα της εταιρείας McKinsey (“Retail speaks: Seven imperatives for the industry”, March 2021), μετά την τρέχουσα κρίση οι καταναλωτές προτίθενται να συνεχίσουν με μεγαλύτερη συχνότητα τις διαδικτυακές αγορές από ό,τι πριν την πανδημία, ενώ το ηλεκτρονικό εμπόριο προβλέπεται να αυξήσει σημαντικά το μερίδιό του επί των συνολικών λιανικών πωλήσεων [5]. Όπως επίσης αποκαλύπτει έρευνα της Ευρωπαϊκής Επιτροπής (“E-commerce statistics for individuals”, January 2021), οι καταναλωτές αξιολογούν ως σημαντική την ευελιξία που προσφέρει το ηλεκτρονικό εμπόριο για αγορές οποτεδήποτε και οπουδήποτε, έχοντας πρόσβαση σε ένα ευρύτερο φάσμα προϊόντων και υπηρεσιών και καθιστώντας τη σύγκριση τιμών πιο εύκολη [6].

Σε ένα πανκαναλικό περιβάλλον λιανικής, οι επιχειρήσεις είναι ευάλωτες, ανεξαρτήτως του μεγέθους τους. Συμβάντα απάτης ενδέχεται να οδηγήσουν σε απώλεια εσόδων, δυσφήμιση των επιχειρήσεων και ως εκ τούτου μείωση της εμπιστοσύνης των πελατών. Αν και είναι πολύ σημαντικό για τις σημερινές συνθήκες του λιανεμπορίου να αναπτυχθεί μία στρατηγική

πωλήσεων με όλα τα κανάλια εξυπηρέτησης, είναι εξίσου σημαντικό για τις επιχειρήσεις να επικεντρωθούν στον εντοπισμό και την πρόβλεψη της απάτης.

1.4 Μορφές Διαδικτυακής Απάτης

Το νέο ψηφιακό περιβάλλον που διαμορφώνεται στην εποχή μας, διευκολύνει την εξ αποστάσεως σύναψη συναλλαγών εκ μέρους των πολιτών - καταναλωτών με οικονομία χρόνου και χρήματος. Παράλληλα όμως η ανωνυμία του διαδικτύου καλλιεργεί συνθήκες ανάπτυξης νέων μορφών απάτης. Ιδιαίτερα κατά την περίοδο της πανδημίας και των αυστηρών περιοριστικών μέτρων εμφανίστηκαν νέοι τύποι απάτης στις ηλεκτρονικές συναλλαγές. Οι συνηθέστερες μορφές διαδικτυακής απάτης είναι οι ακόλουθες:

- **Απάτη για αγορές και πωλήσεις προϊόντων**
  Οι απατεώνες αναρτούν ψευδείς αγγελίες στο διαδίκτυο για πωλήσεις προϊόντων και εξαπατούν τους υποψήφιους αγοραστές, οι οποίοι αποστέλλουν προκαταβολές ή ολόκληρο το ποσό, χωρίς όμως ποτέ να λαμβάνουν αυτό που αγόρασαν. Δημιουργώντας επίσης μια πλασματική ιστοσελίδα, καταφέρνουν να υποκλέπτουν προσωπικά και οικονομικά στοιχεία από χρήστες του διαδικτύου, οι οποίοι έχοντας εξαπατηθεί, νομίζουν ότι πρόκειται για κάποιο διαδικτυακό κατάστημα και πραγματοποιούν τις αγορές τους. Ένα τέτοιο παράδειγμα κατά την περίοδο της πανδημίας αποτέλεσαν τα ψεύτικα site για αγορά μασκών και απολυμαντικών όπου οι απατεώνες παρότρυναν τους πολίτες να παραγγείλουν. Επιπλέον, αρκετές είναι οι περιπτώσεις όπου επιτήδειοι καταφέρνουν να αποκτούν φυσική πρόσβαση στα στοιχεία πιστωτικών καρτών πολιτών, τα οποία στη συνέχεια χρησιμοποιούν σε διαδικτυακές αγορές καθώς για τις αγορές αυτές δεν είναι απαραίτητη η φυσική κατοχή της πιστωτικής κάρτας, πάρα μόνο τα στοιχεία αυτής.
- **Απατηλά μηνύματα ηλεκτρονικού ταχυδρομείου – Ηλεκτρονικό ψάρεμα (phishing)**
  Στην περίπτωση αυτή οι πολίτες δίνουν οι ίδιοι, άθελά τους, τα στοιχεία σε κακόβουλους χρήστες του διαδικτύου (phishing). Ειδικότερα, ο ανυποψίαστος πολίτης λαμβάνει μήνυμα ηλεκτρονικού ταχυδρομείου από πιστωτικό ίδρυμα, στο οποίο τηρεί λογαριασμό με το οποίο του ζητείται να συμπληρώσει τα στοιχεία του (ονοματεπώνυμο, αριθμό λογαριασμού και πιστωτικής κάρτας, κωδικούς του e-banking κλπ.) για λόγους δήθεν ενημέρωσης των αρχείων της τράπεζας. Το μήνυμα, μέσω υπερσυνδέσμου, τους οδηγεί σε μια πλασματική ιστοσελίδα της τράπεζας, με αποτέλεσμα ο πολίτης να πείθεται και να χορηγεί τα επίμαχα στοιχεία. Οι απατεώνες αξιοποιούν άμεσα τα πολύτιμα αυτά στοιχεία που υπέκλεψαν για να μεταφέρουν τις καταθέσεις του ανύποπτου πολίτη σε λογαριασμούς τους σε εξωτικές χώρες από όπου είναι πρακτικά αδύνατο να ανακτηθούν.
- **Απατηλά μηνύματα SMS (Smishing)**
  Πρόκειται για απόπειρα επιτήδειων να αποκτήσουν προσωπικές και οικονομικές πληροφορίες ή κωδικούς ασφαλείας μέσω μηνυμάτων SMS, διαδικασία που είναι γνωστή ως «smishing» (από τις λέξεις «SMS» και «phishing»). Το μήνυμα κειμένου συνήθως ζητάει από το θύμα να κάνει κλικ σε έναν ηλεκτρονικό σύνδεσμο (link) ή να καλέσει έναν αριθμό τηλεφώνου, προκειμένου να επαληθεύσει, ενημερώσει ή επανεργοποιήσει το λογαριασμό του.

- **Απάτη με πρόφαση την επιδιόρθωση Η/Υ**
  Πρόκειται για απόπειρα παραπλάνησης χρηστών του διαδικτύου από άτομα που προσποιούνται τους τεχνικούς οι οποίοι εργάζονται σε κέντρα υποστήριξης μεγάλων εταιρειών λογισμικού. Μέσω κλήσεων από τηλέφωνα τρίτων χωρών και μιλώντας αγγλικά ενημερώνουν τους χρήστες ότι ο υπολογιστής τους είναι δήθεν μολυσμένος με ιούς. Από τα θύματα ζητούν να εγκαταστήσουν στον υπολογιστή τους λογισμικό απομακρυσμένης πρόσβασης, για τη δήθεν επιδιόρθωση του προβλήματος. Με αυτόν τον τρόπο αποκτούν πλήρη έλεγχο και έχουν πρόσβαση σε αποθηκευμένους κωδικούς, τους οποίους στη συνέχεια έχουν τη δυνατότητα να χρησιμοποιήσουν για παράνομες ενέργειες. Αυτή η μορφή απάτης βρήκε πρόσφορο έδαφος κατά την περίοδο της πανδημίας Covid-19, λόγω της τηλεργασίας, της τηλεκπαίδευσης και γενικά της αυξημένης χρήσης υπολογιστών.
- **Απάτη σχετιζόμενη με επενδύσεις**
  Άγνωστοι προσεγγίζουν πολίτες, είτε μέσω τηλεφώνου είτε μέσω email, υποδυόμενοι ότι εργάζονται σε δήθεν μεγάλες επενδυτικές εταιρείες του εξωτερικού και τους υπόσχονται υψηλές αποδόσεις επενδύοντας χρήματα μέσω εικονικών επενδυτικών πλατφορμών στο διαδίκτυο. Αυτού του είδους οι απάτες, σχετιζόμενες με επενδύσεις, μπορεί να περιλαμβάνουν επικερδείς επενδυτικές ευκαιρίες όπως μετοχές, ομόλογα, κρυπτονομίσματα, πολύτιμους λίθους, υπεράκτιες επενδύσεις σε ακίνητη περιουσία και εναλλακτικές πηγές ενέργειας. Για να είναι μάλιστα πειστικοί οι απατεώνες, τις περισσότερες φορές μιλούν πολύ καλά ελληνικά. Δυστυχώς αυτοί που τους εμπιστεύονται επενδύουν τα χρήματά τους τα οποία βλέπουν εικονικά να αυξάνονται, αλλά όταν ζητήσουν να τους αποδοθούν τα κέρδη διαπιστώνουν ότι αυτό δεν είναι εφικτό και ότι έχουν πέσει θύματα εξαπάτησης.
- **Απάτη με ψευδείς διαγωνισμούς**
  Πολύ συχνά ο καταναλωτής λαμβάνει προσκλήσεις για συμμετοχή σε διαγωνισμούς, για να στοιχηματίσει σε ηλεκτρονικά καζίνο, δέχεται δήθεν αναγγελίες κέρδους χρηματικών ποσών σε λοταρίες, όπου του ζητείται η καταβολή των εξόδων μεταφοράς των χρημάτων τα οποία φυσικά ποτέ δεν λαμβάνει [7].

Οι απατεώνες ανακαλύπτουν συνεχώς νέες μορφές απάτης και καθώς η φαντασία τους είναι απεριόριστη, εφευρίσκουν νέους τρόπους κάθε φορά για να εξαπατήσουν τα θύματά τους. Στη μάχη κατά της ηλεκτρονικής απάτης έχει μπει η τεχνολογία η οποία με τη χρήση ειδικών αλγορίθμων «μαθαίνει» να ανιχνεύει και να προβλέπει ύποπτες κινήσεις.

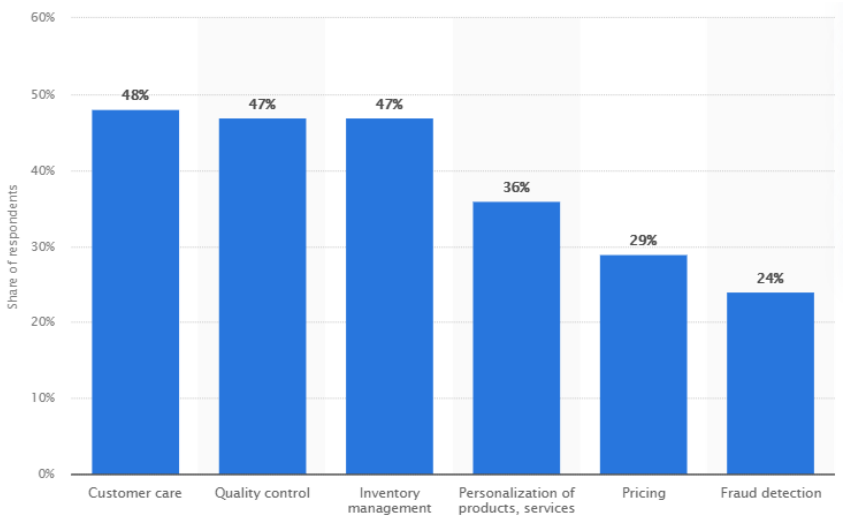
1.5 Λιανικό εμπόριο και ψηφιακός μετασχηματισμός

Η πανδημική κρίση στάθηκε η αφορμή για την ενίσχυση του ηλεκτρονικού εμπορίου αλλά και του ψηφιακού μετασχηματισμού των επιχειρήσεων του κλάδου λιανεμπορίου. Ο ψηφιακός μετασχηματισμός των επιχειρήσεων αποτελεί έναν από τους άξονες των στρατηγικών κατευθύνσεων του Εθνικού Σχεδίου Ανάκαμψης και Ανθεκτικότητας (Νοέμβριος 2020) και προβλέπει την υιοθέτηση ψηφιακών τεχνολογιών από τις επιχειρήσεις που αποτελεί πρόκληση ιδιαίτερα για τις ελληνικές μικρομεσαίες επιχειρήσεις, μειώνοντας το ψηφιακό χάσμα μεταξύ αυτών και του ευρωπαϊκού μέσου όρου, σημειώνει η Alpha Bank στο εβδομαδιαίο δελτίο οικονομικών εξελίξεων [8]. Το λιανικό εμπόριο αποτελεί από τους πιο αξιοσημείωτους τομείς παγκοσμίως που χρησιμοποιούν την τεχνητή νοημοσύνη (Artificial Intelligence - AI) και τη

δύναμη της μάθησης των μηχανών (Machine Learning) προσθέτοντας αξία στις επιχειρηματικές τους δραστηριότητες. Η τεχνητή νοημοσύνη αποτελεί μία προσομοίωση των διαδικασιών ανθρώπινης νοημοσύνης από μηχανές και ειδικότερα από συστήματα υπολογιστών. Τα μοντέλα προβλέψεων Machine Learning ανήκουν σε έξυπνες λύσεις ασφάλειας στο διαδίκτυο. Η ανίχνευση και η πρόβλεψη απάτης με μηχανική εκμάθηση είναι μια διαδικασία διερεύνησης δεδομένων και ανάπτυξης ενός μοντέλου που θα παρέχει τα καλύτερα αποτελέσματα στην αναγνώριση και πρόβλεψη παράνομων συναλλαγών συγκεντρώνοντας όλες τις σημαντικές πληροφορίες των συναλλαγών των χρηστών. Οι αλγόριθμοι Μηχανικής Μάθησης συνδυάζοντας δεδομένα γύρω από το προφίλ του κάθε πελάτη, έχουν την ικανότητα να αναλύουν την αγοραστική συμπεριφορά των πελατών και να οδηγούν σε σημαντικά συμπεράσματα για τις καταναλωτικές τους συνήθειες και τις προτιμήσεις τους, αυξάνοντας για τις επιχειρήσεις τη δυνατότητα παροχής εξατομικευμένων προϊόντων προς τους καταναλωτές. Η δημιουργία ομάδων πελατών (τμηματοποίηση πελατών) προσφέρει τη δυνατότητα στις επιχειρήσεις στοχευμένου μάρκετινγκ με προσωποποιημένες προτάσεις που εστιάζουν στον εκάστοτε πελάτη. Καθώς οι επιχειρήσεις καλούνται να διαχειριστούν έναν τεράστιο όγκο δεδομένων που παράγονται από τους πελάτες και τις συναλλαγές που πραγματοποιούνται, η επεξεργασία και η ανάλυση των δεδομένων αυτών δίνει τη δυνατότητα στις επιχειρήσεις λιανικού εμπορίου να παρακολουθούν πώς κινούνται καταναλωτικά οι αγοραστές και τι επιλέγουν, αλλά και να κατανοούν τα συναισθήματα, τις αντιδράσεις και τα πρότυπα των υποψηφίων πελατών τους, εντοπίζοντας παράλληλα δόλιες ή ύποπτες συμπεριφορές και αποτρέποντας περιστατικά απάτης. Η αξιοποίηση των πληροφοριών και η πολύτιμη γνώση που εξάγεται μέσα από την πληθώρα των δεδομένων, συντελούν στη βελτιστοποίηση της δραστηριότητας των επιχειρήσεων και κυρίως στην κατανόηση και στην προσαρμογή στις προτιμήσεις των πελατών τους, ικανοποιώντας έτσι τις απαιτήσεις τους και παρέχοντας προϊόντα προσαρμοσμένα αποκλειστικά στις ανάγκες τους. Η μοναδική ικανότητα της τεχνητής νοημοσύνης να ανακαλύπτει συμπεριφορικά πρότυπα, μέσω της μηχανικής μάθησης, επιτρέπει στις επιχειρήσεις να μαθαίνουν το προφίλ των πελατών τους και να σχεδιάζουν στρατηγικές που όχι μόνο βοηθούν στην προσέλκυση περισσότερων πελατών αλλά συμβάλλουν και στη διατήρηση των ήδη υφιστάμενων πελατών ενισχύοντας την πιστότητά τους [9]. Παράλληλα η ανίχνευση και πρόβλεψη απάτης με τη συμβολή της μηχανικής μάθησης και την ανάπτυξη αναλυτικών μοντέλων πρόβλεψης, βοηθά τις επιχειρήσεις στην προστασία των εσόδων τους, στη διατήρηση της φήμης τους και στην ενίσχυση της εμπιστοσύνης και της καλής πίστης των πελατών. Η διάπραξη απάτης και ο μη έγκαιρος εντοπισμός αυτής αποτελεί έναν εγγενή κίνδυνο με τον οποίο έρχεται αντιμέτωπη κάθε επιχείρηση οποιουδήποτε τομέα δραστηριότητος και ανεξαρτήτως μεγέθους.

Καθώς η λιανική βιομηχανία αλλάζει με ταχύτατους ρυθμούς και συνεχώς εξελίσσεται, κρίνεται πολύ σημαντική η επένδυση των επιχειρήσεων στον ψηφιακό μετασχηματισμό τους και την ενσωμάτωση ψηφιακών τεχνολογιών ώστε να μπορούν να ανταπεξέλθουν στον ανταγωνισμό που γίνεται ολοένα και πιο έντονος στον τομέα των λιανικών πωλήσεων. Δεδομένου ότι οι καταναλωτές σήμερα έχουν γίνει πιο απαιτητικοί με αυξημένες προσδοκίες για προϊόντα και υπηρεσίες και επιζητούν άμεση ικανοποίηση των επιθυμιών τους, οι επιχειρήσεις λιανικού εμπορίου κατευθύνονται ολοένα και περισσότερο στην αξιοποίηση της τεχνητής νοημοσύνης και της μηχανικής μάθησης, αναβαθμίζοντας τις υπηρεσίες τους, διευρύνοντας τις δυνατότητες του καταναλωτή και προσφέροντας καλύτερη εξυπηρέτηση στον πελάτη. Σύμφωνα με έρευνα που

πραγματοποιήθηκε για τις περιπτώσεις χρήσης της τεχνητής νοημοσύνης στη λιανική βιομηχανία παγκοσμίως, το 48% των ανταποκριθέντων στην έρευνα αναφέρει ότι η αξιοποίηση και ανάπτυξη της τεχνητής νοημοσύνης μπορεί να βοηθήσει στην εξυπηρέτηση των πελατών, το 47% στον ποιοτικό έλεγχο, το 47% στη διαχείριση αποθεμάτων, το 36% στην παροχή εξατομικευμένων προϊόντων και υπηρεσιών, το 29% στην πολιτική τιμολόγησης, ενώ το 24% στην ανίχνευση απάτης [10].



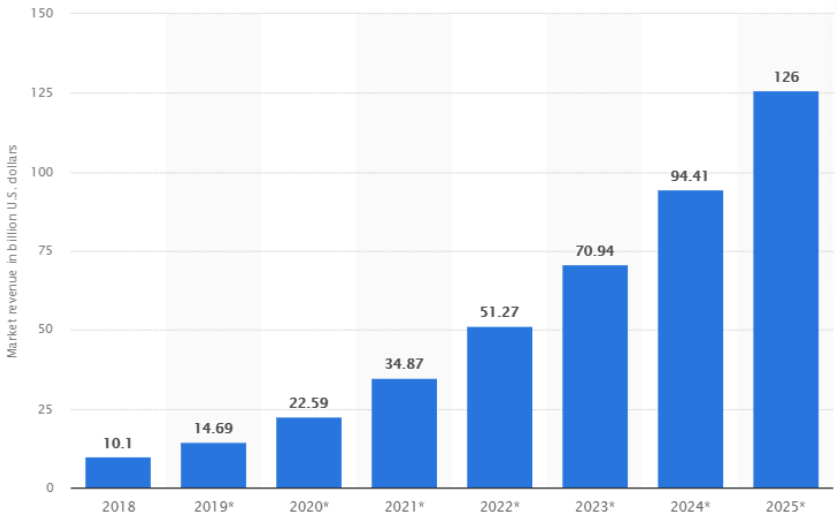
Εικόνα 1.2 Περιπτώσεις χρήσης AI σε καταναλωτικά αγαθά και λιανική βιομηχανία παγκοσμίως, 2020

Πηγή: Statista.com

Μεγάλες εταιρείες τεχνολογίας παγκοσμίου φήμης όπως η Google, η Microsoft, η Amazon, η IBM εξέλιξαν την τεχνητή νοημοσύνη στις λιανικές πωλήσεις παρέχοντας προηγμένου επιπέδου υπηρεσίες μέσω καινοτόμων τεχνολογικών λύσεων. Σήμερα μεγάλοι λιανοπωλητές έχουν στραφεί στις δυνατότητες του Machine Learning και των Μεγάλων Δεδομένων (Big Data) και μέσω της επεξεργασίας και ανάλυσης συμπεριφορικών δεδομένων, έχουν τη δυνατότητα παροχής στοχευμένων – εξατομικευμένων προτάσεων για προϊόντα τα οποία ανταποκρίνονται καλύτερα στις ανάγκες των καταναλωτών, καθιστώντας την εμπειρία με τα προϊόντα αυτά όσο το δυνατόν πιο μοναδική. Με τη χρήση αλγορίθμων μηχανικής μάθησης και τη δημιουργία προβλεπτικών μοντέλων για την ανίχνευση απάτης, απομακρύνεται ο κίνδυνος απώλειας εσόδων για τις επιχειρήσεις καθώς οι πωλήσεις δεν επηρεάζονται από δόλιες δραστηριότητες. Αποτρέποντας φαινόμενα απάτης επιτυγχάνεται επιπλέον η προστασία της αξιοπιστίας των επιχειρήσεων, η ενδυνάμωση των σχέσεων με τους πελάτες τους και γενικότερα η διασφάλιση των συναλλαγών.

Η μετάβαση προς την ψηφιακή τεχνολογία αποτελεί τη λύση για την επιτυχία των επιχειρήσεων, την αύξηση των πωλήσεων και της κερδοφορίας, παρέχοντας ταυτόχρονα ανταγωνιστικό πλεονέκτημα έναντι των αντιπάλων. Λύσεις τεχνητής νοημοσύνης και μηχανικής μάθησης με πολλές καινοτόμες εφαρμογές παρουσιάζονται σε διάφορους κλάδους της οικονομίας όπως στην υγεία, τις κατασκευές, το χρηματοοικονομικό κλάδο, την εκπαίδευση, την αυτοκινητοβιομηχανία και άλλους τομείς αποδίδοντας μεγάλη επιτυχία σε κάθε κλάδο και αλλάζοντας ριζικά τον κόσμο σε παγκόσμιο επίπεδο. Όπως προβλέπεται, τα έσοδα που θα παρουσιάσει η αγορά λογισμικού τεχνητής νοημοσύνης (AI) θα αυξηθούν ταχύτατα τα επόμενα χρόνια φθάνοντας τα 126

δισεκατομμύρια δολάρια το 2025 έναντι των περίπου 10 δισεκατομμυρίων δολαρίων που παρουσίαζε το 2018 [11].



Εικόνα 1.3 Έσοδα από την αγορά λογισμικού τεχνητής νοημοσύνης (AI) παγκοσμίως από το 2018 έως το 2025

Πηγή: Statista.com

Ειδικότερα για τον κλάδο λιανικού εμπορίου, η τεχνητή νοημοσύνη έχει διεισδύσει σημαντικά στις σημερινές επιχειρήσεις και μετασχηματίζοντας την παραδοσιακή εμπειρία λιανικής πώλησης, αλλάζει τον τρόπο με τον οποίο οι επιχειρήσεις αλληλεπιδρούν με τους πελάτες τους και οδηγεί στην αναβάθμιση των επιχειρηματικών δραστηριοτήτων με αυτοματοποίηση διαδικασιών, εξατομίκευση και βελτιωμένη απόδοση για τους καταναλωτές. Με την αξιοποίηση της μηχανικής μάθησης που αποτελεί μέρος της τεχνητής νοημοσύνης (AI) και την απόκτηση πολύτιμης γνώσης μέσω της ανάλυσης μεγάλων δεδομένων (Big data), προστίθεται μεγαλύτερη αξία και αποδοτικότητα και ικανοποιούνται διάφορες απαιτήσεις της λιανικής βιομηχανίας.

ΚΕΦΑΛΑΙΟ 2: Μεγάλα Δεδομένα – Επιχειρηματική Ευφυΐα

2.1 Μεγάλα Δεδομένα (Big Data)

Η παγκόσμια έκρηξη παραγωγής δεδομένων των τελευταίων δεκαετιών έχει οδηγήσει στο χαρακτηρισμό της σύγχρονης εποχής ως εποχή των «Μεγάλων Δεδομένων». Στη ραγδαία αύξηση των δεδομένων συνεισφέρουν, μεταξύ άλλων, πηγές όπως ο παγκόσμιος ιστός, τα κοινωνικά δίκτυα, το Διαδίκτυο των πραγμάτων (Internet of Things - IoT), καθώς και η τάση για ανοικτά δεδομένα. Σύμφωνα με την Ευρωπαϊκή Επιτροπή, τα προβλεπόμενα αριθμητικά στοιχεία για το 2025 παρουσιάζουν αύξηση κατά 530% του παγκόσμιου όγκου δεδομένων, από 33 zettabytes το 2018 σε 175 zettabytes. Παράλληλα η αξία της οικονομίας των δεδομένων στην Ευρωπαϊκή Ένωση προβλέπεται ίση με 829 δισ. € από 301 δισ. € (2,4% του ΑΕΠ της ΕΕ) το 2018. Οι επαγγελματίες του τομέα των δεδομένων στην Ευρωπαϊκή Ένωση προβλέπεται να ανέλθουν σε 10,9 εκατομμύρια από 5,7 εκατομμύρια το 2018 ενώ ποσοστό 65% του πληθυσμού της Ευρωπαϊκής Ένωσης προβλέπεται να διαθέτουν βασικές ψηφιακές δεξιότητες, από το 57% το 2018 [12].

Εικόνα 2.1 Προβλεπόμενα αριθμητικά στοιχεία για το 2025

Πηγή: ec.europa.eu

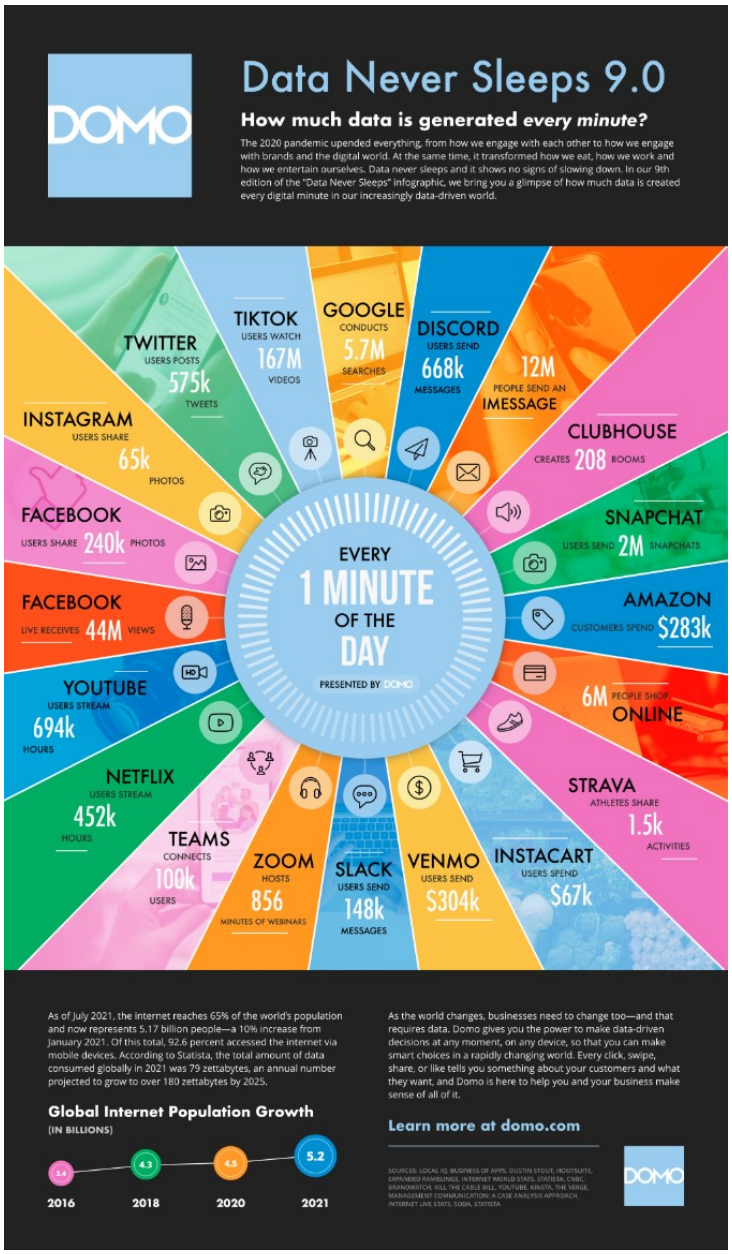
Συγκεκριμένα για το 2020, ο όγκος των δεδομένων που δημιουργήθηκαν ήταν υψηλότερος από ό,τι αναμενόταν προηγουμένως, εξαιτίας της αυξημένης χρήσης του Διαδικτύου λόγω της πανδημίας Covid-19, καθώς μεγάλο μέρος του πληθυσμού εργαζόταν και εκπαιδευόταν από το σπίτι και χρησιμοποιούσε πιο συχνά επιλογές οικιακής ψυχαγωγίας και διαδικτυακές αγορές προϊόντων.

Τα δεδομένα βρίσκονται στον πυρήνα του ψηφιακού μετασχηματισμού. Διαμορφώνουν τον τρόπο με τον οποίο παράγουμε, καταναλώνουμε και ζούμε. Η πρόσβαση σε ένα συνεχώς αυξανόμενο όγκο δεδομένων και η ικανότητα χρησιμοποίησής του έχει καθοριστική σημασία για την καινοτομία και την ανάπτυξη. Η ανάγκη για ψηφιακό μετασχηματισμό και καινοτομία είναι από τους βασικούς μοχλούς για επενδύσεις σε τεχνητή νοημοσύνη (AI) και τεχνολογίες μεγάλων δεδομένων. Οι ανάγκες της εποχής και η στροφή ολοένα και περισσοτέρων επιχειρήσεων στο ηλεκτρονικό εμπόριο, ενισχύει ακόμα περισσότερο την παραγωγή δεδομένων. Οι μεγαλύτερες εταιρείες e-commerce δέχονται μόλις σε λίγα λεπτά χιλιάδες συναλλαγές από πελάτες.

Οι σύγχρονες επιχειρήσεις καλούνται σήμερα να εστιάσουν στην αξιοποίηση μεγάλων δεδομένων ώστε να οδηγηθούν στη λήψη σωστών αποφάσεων, προσφέροντας τη βέλτιστη αξία στις επιχειρήσεις τους. Προκειμένου να αξιοποιήσουν τα μεγάλα δεδομένα οι επιχειρήσεις βασίζονται στην αποθηκευτική και επεξεργαστική ισχύ καθώς και σε ισχυρές ικανότητες και δεξιότητες ανάλυσης μέσω προηγμένων τεχνικών και εργαλείων ανάλυσης δεδομένων. Η υιοθέτηση μιας ολοκληρωμένης στρατηγικής γύρω από τα Big Data, θα δώσει σημαντικό πλεονέκτημα στις επιχειρήσεις αφού θα μπορούν να διαχειριστούν τα δεδομένα τους με αποτελεσματικό τρόπο, εξάγοντας σημαντικά συμπεράσματα και αποκομίζοντας πολύτιμη γνώση που οδηγεί στη βελτιστοποίηση των δραστηριοτήτων τους και την ενίσχυση του ανταγωνιστικού τους πλεονεκτήματος.

2.1.1 Big Data & Covid-19

Σε έναν ολοένα και πιο ψηφιακό κόσμο που εντείνεται από το ξέσπασμα της πανδημίας Covid-19, έχουμε εισέλθει σε μια νέα εποχή όπου η τεχνολογία και τα δεδομένα έχουν αποκτήσει πιο σημαντικούς ρόλους στην καθημερινή μας ζωή. Η πανδημία του 2020, ανέτρεψε τα πάντα από το πώς αλληλεπιδρούμε μεταξύ μας μέχρι το πώς αλληλεπιδρούμε με τον ψηφιακό κόσμο. Ταυτόχρονα άλλαξε τον τρόπο με τον οποίο τρώμε, εργαζόμαστε και διασκεδάζουμε. Η παγκόσμια πανδημία έχει αυξήσει την ψηφιοποίηση της καθημερινής ζωής. Η εργασία εξ αποστάσεως, η αύξηση των διαδικτυακών αγορών, η γνωριμία μας με ψηφιακούς χώρους συσκέψεων και διαδικτυακών σεμιναρίων, η διδασκαλία από απόσταση, η εικονική ψυχαγωγία είναι μερικά παραδείγματα της νέας ψηφιακής πραγματικότητας και των νέων ψηφιακών τάσεων που εμφανίστηκαν, αποδεικνύοντας τον ολοένα και πιο προσανατολισμένο στα δεδομένα κόσμο γύρω μας. Ο Covid-19 αποτέλεσε τη σπίθα για αυτήν την εκσυγχρονιστική τάση και συνέβαλε στην επιτάχυνση του μετασχηματισμού των επιχειρηματικών πρακτικών και διαδικασιών. Οι επιχειρήσεις επιτάχυναν τις ψηφιακές πρωτοβουλίες προκειμένου να ενισχύσουν την ευελιξία τους και να διατηρήσουν την ανταγωνιστικότητά τους. Επιπλέον, η πανδημία συνέβαλε σε μια σαρωτική, ανοδική τάση στη χρήση του διαδικτύου. Σύμφωνα με τα στοιχεία της εταιρείας Domo, από τον Ιανουάριο έως τον Ιούλιο του 2021 ο αριθμός των ατόμων που είχαν πρόσβαση στο διαδίκτυο αυξήθηκε κατά 10%, με 5,17 δισεκατομμύρια άτομα, ή το 65% του πληθυσμού, να συνδέονται. Αυτό έχει ως αποτέλεσμα την παραγωγή τεράστιου όγκου δεδομένων τα οποία διανέμονται μέσω του διαδικτύου και συνεχώς εξελίσσονται [13]. Η ανάλυση αυτών των δεδομένων μπορεί να βοηθήσει στην καλύτερη κατανόηση ενός κόσμου που κινείται με αυξανόμενες ταχύτητες. Η παρακάτω εικόνα παρουσιάζει πόσα δεδομένα δημιουργήθηκαν κάθε λεπτό σε πλατφόρμες και εφαρμογές υψηλής επισκεψιμότητας το 2020.



Εικόνα 2.2 Παραγόμενη ποσότητα δεδομένων κάθε 1 λεπτό στο διαδίκτυο

Πηγή: domo.com

2.1.2 Χαρακτηριστικά Μεγάλων Δεδομένων

Με την πάροδο των ετών και με την εξέλιξη της εφαρμογής των Big Data έχουν προστεθεί πολλά χαρακτηριστικά που τα προσδιορίζουν. Μεταξύ αυτών των χαρακτηριστικών, τρία είναι τα πιο δημοφιλή και έχουν ευρέως αναφερθεί και υιοθετηθεί τα οποία είναι γνωστά και ως τα 3Vs και είναι τα εξής:

3Vs – Douglas Laney (2001) – Βασικό μοντέλο [14]

- **Volume (Όγκος)** που σημαίνει την εισερχόμενη ροή δεδομένων και το σωρευτικό όγκο δεδομένων.
- **Velocity (Ταχύτητα)** που αναφέρεται στην υψηλή ταχύτητα με την οποία τα δεδομένα εισέρχονται στις βάσεις δεδομένων και στη συνέχεια επεξεργάζονται.
- **Variety (Ποικιλία)** που υποδηλώνει την ποικιλία των δεδομένων. Τα δεδομένα προέρχονται από διάφορες πηγές και παρέχονται σε διαφορετικές μορφές.

IBM – 4Vs

Η IBM πρόσθεσε επιπλέον το χαρακτηριστικό,

- **Veracity (Ακρίβεια - Εγκυρότητα)** που αναφέρεται στην ακρίβεια και την αξιοπιστία των δεδομένων. Η πρόσθετη διάσταση «V» αποτελεί την απάντηση στα προβλήματα ποιότητας που οι πελάτες της IBM αντιμετώπισαν ξεκινώντας δραστηριότητες με τα μεγάλα δεδομένα.

Yuri Demchenko – 5Vs

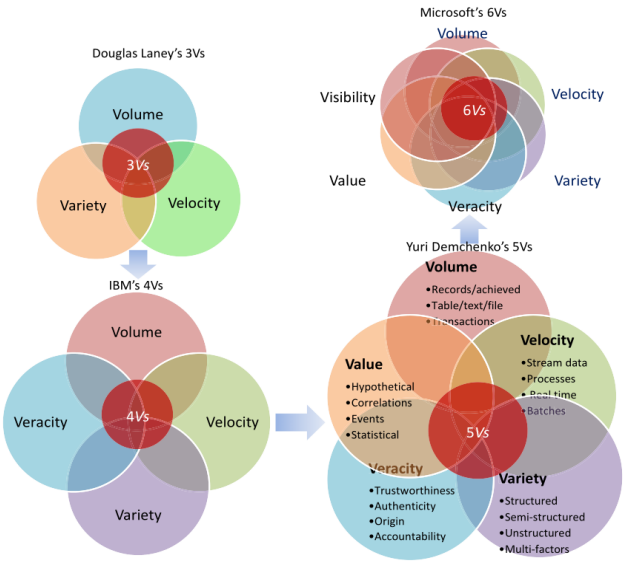
Ο Yuri Demchenko πρόσθεσε το 2013 το χαρακτηριστικό,

- **Value (Αξία)** που αναφέρεται στην αξία και στην εξαγωγή πολύτιμης γνώσης που προσδίδουν τα δεδομένα ύστερα από την ορθή επεξεργασία τους.

Microsoft – 6Vs

Για τη μεγιστοποίηση της επιχειρηματικής αξίας, η Microsoft επέκτεινε τα χαρακτηριστικά σε 6 Vs, προσθέτοντας το χαρακτηριστικό,

- **Visibility (Ορατότητα)** που υπογραμμίζει ότι πρέπει να υπάρχει μια πλήρη εικόνα των δεδομένων για να παρθεί μια σωστή απόφαση [15].



Εικόνα 2.3 Επέκταση βασικού μοντέλου 3Vs σε 6Vs

Πηγή: Buyya, Calheiros, Dastjerdi 2016

Πέραν των έξι προαναφερθέντων χαρακτηριστικών των μεγάλων δεδομένων, το μοντέλο έχει εξελιχθεί και έχει συμπεριλάβει περισσότερα χαρακτηριστικά.

2.1.3 Τύποι Δεδομένων

Οι τρεις βασικοί τύποι δεδομένων είναι οι εξής:

- **Δομημένα (structured)**

Ονομάζονται δομημένα διότι όταν εισάγονται σε μια σχεσιακή βάση δεδομένων, είναι οργανωμένα και αποθηκευμένα με σταθερή μορφή. Τα δεδομένα αυτού του τύπου χρησιμοποιούν κυρίως SQL, η οποία είναι μια γλώσσα προγραμματισμού που έχει δημιουργηθεί για σχεσιακές βάσεις δεδομένων. Τα δομημένα δεδομένα παρουσιάζουν σχετική ευκολία στην εισαγωγή, στην αποθήκευση και στην ανάλυσή τους. Οι συναλλαγές αποτελούν παράδειγμα δεδομένων δομημένης μορφής, τα οποία περιγράφουν με λεπτομέρεια τα στοιχεία συναλλαγών των πελατών.

- **Ημι – δομημένα (semi – structured)**

Τα ημι-δομημένα δεδομένα μπορεί να έχουν κάποιου είδους μορφή δομής αλλά δεν ακολουθούν μια συμβατική βάση δεδομένων. Η καταγραφή τέτοιου τύπου δεδομένων για ανάλυση είναι διαφορετική από τη λήψη μιας σταθερής μορφής αρχείου. Οι πίνακες σε υπολογιστικά φύλλα αποτελούν παράδειγμα ημι-δομημένων δεδομένων.

- **Αδόμητα (unstructured)**

Τα δεδομένα αυτού του τύπου δεν διαθέτουν καθορισμένη μορφή ή δομή και παρουσιάζουν ιδιαίτερη πολυπλοκότητα στην επεξεργασία και ανάλυσή τους καθώς επίσης είναι δύσκολο να αποθηκευτούν χρησιμοποιώντας παραδοσιακές σχεσιακές βάσεις δεδομένων. Παραδείγματα

αδόμητων δεδομένων είναι τα μηνύματα ηλεκτρονικού ταχυδρομείου, οι εικόνες, τα βίντεο, δεδομένα κινητής τηλεφωνίας κλπ. Τα κοινωνικά δίκτυα (Facebook, Twitter κλπ.) αποτελούν τεράστια πηγή τέτοιων δεδομένων. Εξαιτίας της εκτεταμένης χρήσης των μέσων κοινωνικής δικτύωσης και ως εκ τούτου της συνεχούς αύξησης δεδομένων αυτής της μορφής, η κατανόηση και διαχείρισή τους κρίνεται εξαιρετικά σημαντική.

2.1.4 Τομείς Εφαρμογής Μεγάλων Δεδομένων

Τα Μεγάλα Δεδομένα αξιοποιούνται σήμερα σε σημαντικούς τομείς της κοινωνικής και οικονομικής δραστηριότητας. Στόχος της εφαρμογής των Big Data είναι η βελτίωση της λειτουργίας των επιχειρήσεων, η βελτίωση της εμπειρίας των πελατών και η ανάλυση και εξαγωγή συμπερασμάτων που οδηγούν στη λήψη βέλτιστων αποφάσεων και στρατηγικών επιχειρηματικών κινήσεων. Ορισμένοι από τους κυριότερους κλάδους που αξιοποιούν τα μεγάλα δεδομένα είναι οι εξής:

- **Υγεία**

Ο όγκος των δεδομένων που παράγονται στο χώρο της υγείας είναι ιδιαίτερα σημαντικός. Η επεξεργασία και ανάλυση των δεδομένων επιτρέπει τη βελτίωση της ιατρικής περίθαλψης και συμβάλλει στην πρόβλεψη μελλοντικών ασθενειών, στην έγκαιρη και ακριβή διάγνωση, στη βελτίωση θεραπειών αλλά και στην αποδοτικότερη λειτουργία κλινικών μονάδων. Η ηλεκτρονική απεικόνιση ιστορικού υγείας και το ψηφιακό αρχείο με κλινικά δεδομένα ασθενών, παρέχουν καλύτερη εξατομικευμένη θεραπεία εστιασμένη σε κάθε περίπτωση ασθενούς. Επιπλέον οι φαρμακευτικές εταιρείες αξιοποιούν την ανάλυση των Big Data για την παρακολούθηση της αποτελεσματικότητας των φαρμάκων καθώς και την ανάπτυξη νέων φαρμάκων, παρακολουθώντας συγχρόνως το κόστος παρασκευής και πώλησης αλλά και την αποτελεσματικότητα κόστους/θεραπείας. Κατά την πρόσφατη κρίση της πανδημίας Covid-19 που συνεχίζει να πλήττει τον κόσμο παγκοσμίως, η αξιοποίηση μεγάλης κλίμακας δεδομένων που προέρχονται από κάθε γωνιά της γης, έχει δημιουργήσει μια ζωτική πηγή πληροφοριών και γνώσης βοηθώντας την ιατρική κοινότητα στην εξαγωγή σημαντικών συμπερασμάτων σχετικά με τη δράση του ιού. Η ανάλυση μεγάλων δεδομένων λειτουργεί ως μέσο παρακολούθησης, ελέγχου, έρευνας και πρόληψης του Covid-19 ως πανδημίας ενώ παράλληλα βοηθά στον περιορισμό εξάπλωσής της αλλά και στη λήψη κατάλληλων και έγκαιρων αποφάσεων. Επιπλέον τα δεδομένα που λαμβάνονται μπορούν να εκπαιδευτούν ξανά για την ανάπτυξη μελλοντικών μεθόδων πρόληψης [16].

- **Χρηματοπιστωτικός κλάδος**

Τα χρηματοπιστωτικά ιδρύματα αξιοποιούν τη δύναμη των Big Data δίνοντάς τους εξέχουσα σημασία. Οι τεχνολογίες Big Data επιτρέπουν την καλύτερη χαρτογράφηση των πελατών τους, διατηρώντας αναλυτικά στοιχεία τους αλλά και δεδομένα για όλες τις συναλλαγές που πραγματοποιούνται. Αναλύοντας τα στοιχεία των πελατών τους, δημιουργούν το προφίλ τους και διεξάγουν αναλύσεις τμηματοποίησης της αγοράς. Μαθαίνοντας τη συμπεριφορά των πελατών παρέχεται η δυνατότητα καλύτερης εξυπηρέτησης, προσφέροντας προσωποποιημένα χρηματοοικονομικά προϊόντα που ανταποκρίνονται στις ιδιαίτερες ανάγκες των πελατών, επιτυγχάνοντας ταυτόχρονα καλύτερη εμπειρία πελάτη και μεγαλύτερη ικανοποίηση. Επιπλέον

επιτυγχάνεται προσέλκυση και απόκτηση νέων πελατών με στοχευμένες καμπάνιες μάρκετινγκ, βάσει συγκεκριμένων προτιμήσεών τους. Πολύ σημαντικές εφαρμογές που προσφέρουν τα Big Data και η ανάλυσή τους στο χρηματοπιστωτικό κλάδο αποτελούν η διαχείριση του κινδύνου, η πρόληψη της απάτης αλλά και η πρόβλεψη απώλειας πελατών. Με τη σωστή εκτίμηση και διαχείριση για τον περιορισμό του κινδύνου, την έγκαιρη ανίχνευση ενδεχόμενης απάτης αλλά και την αποφυγή απώλειας πελατών διαφυλάσσεται η ομαλή λειτουργία των χρηματοπιστωτικών οργανισμών ενώ παράλληλα ενισχύεται η ανταγωνιστικότητά τους στην αγορά.

- **Πωλήσεις & Μάρκετινγκ**

Η χρήση των μεγάλων δεδομένων επηρεάζει καθοριστικά τον τομέα των πωλήσεων και το Μάρκετινγκ. Στις πωλήσεις χρησιμοποιούνται για την αναγνώριση της καταναλωτικής συμπεριφοράς των πελατών και την τμηματοποίηση της αγοράς (market segmentation). Η τμηματοποίηση της αγοράς συνίσταται στον εντοπισμό του καταναλωτικού κοινού σε ομάδες με ομοειδή χαρακτηριστικά. Τα μέλη μιας ομάδας παρουσιάζουν ομοιότητες σε σχέση με χαρακτηριστικά όπως η οικογενειακή και οικονομική κατάσταση, το μορφωτικό επίπεδο, η ηλικία, το φύλο, ο τόπος κατοικίας και κυρίως το ιστορικό αγορών και η καταναλωτική συμπεριφορά. Γνωρίζοντας τη σύνδεση του καταναλωτικού κοινού και τα χαρακτηριστικά της κάθε ομάδας, οι επιχειρήσεις μπορούν να υλοποιήσουν μια στρατηγική σχεδιασμού και διάθεσης προϊόντων και υπηρεσιών η οποία εστιάζει σε μια πιο στοχευμένη προσέγγιση πελατών που θα εξυπηρετεί τις ιδιαίτερες ανάγκες τους, επιτυγχάνοντας έτσι την καλύτερη εξυπηρέτηση των πελατών αλλά και την ενδυνάμωση των σχέσεών τους με τους πελάτες. Η Ανάλυση Συναισθήματος (Sentiment Analysis) επίσης, αποτελεί χρήσιμη τεχνική για να γνωρίσει η κάθε επιχείρηση τις απόψεις του καταναλωτικού κοινού σχετικά με τα προϊόντα της και τις υπηρεσίες της, οδηγώντας με αυτόν τον τρόπο τις επιχειρήσεις στη βελτίωση λειτουργιών και δραστηριοτήτων και στη δημιουργία αποτελεσματικών λύσεων με καινοτόμα προϊόντα που δίνουν ανταγωνιστικό πλεονέκτημα και πλήρη ικανοποίηση των πελατών. Επιπλέον, τα Big Data προσφέρουν τη δυνατότητα για στοχευμένο μάρκετινγκ που βασίζεται πάνω στις ανάγκες των ανθρώπων, με βάση την ανάλυση και το συνδυασμό δεδομένων από το προφίλ του κάθε πελάτη, που σχετίζονται με τις προτιμήσεις του, τις οικονομικές του δυνατότητες κλπ. Οι επιχειρήσεις αξιοποιώντας τα μεγάλα δεδομένα δημιουργούν εξατομικευμένες καμπάνιες μάρκετινγκ για διατήρηση, αλλά και προσέλκυση νέων πελατών επιφέροντας άνοδο των πωλήσεων και αύξηση των κερδών τους. Άλλες περιπτώσεις χρήσης των Big Data στο συγκεκριμένο κλάδο είναι η πρόβλεψη πωλήσεων, ο ποιοτικός έλεγχος, η πρόβλεψη απώλειας πελάτη, η ανίχνευση απάτης, ο καθορισμός πολιτικής τιμολόγησης κλπ.

- **Ασφαλιστικός κλάδος**

Ο ασφαλιστικός κλάδος αποτελεί έναν από τους σημαντικότερους πυλώνες της οικονομίας. Η αξιοποίηση των Big Data παρέχει τη δυνατότητα στις ασφαλιστικές εταιρείες να αντλήσουν πολύτιμες πληροφορίες σχετικά με τις αγοραστικές συνήθειες και προτιμήσεις των πελατών τους, προσφέροντάς τους υπηρεσίες προσανατολισμένες στις ανάγκες τους συμβάλλοντας έτσι στη διατήρηση των υφιστάμενων πελατών τους και κερδίζοντας την εμπιστοσύνη και αφοσίωσή τους. Σε συνδυασμό με τις κατάλληλες τεχνικές ανάλυσης, οι ασφαλιστικές εταιρείες μπορούν να μετατρέψουν τα big data σε πηγή χρήσιμων πληροφοριών για στοχευμένες προωθητικές ενέργειες,

με σκοπό την απόκτηση νέων πελατών αλλά και την ανάπτυξη νέων καινοτόμων υπηρεσιών, επιφέροντας άνοδο των πωλήσεων και ως εκ τούτου αύξηση της κερδοφορίας τους. Η διαχείριση του κινδύνου είναι εξαιρετικά σημαντική για τον ασφαλιστικό κλάδο. Η ανάλυση μεγάλων δεδομένων με την παράλληλη χρήση κατάλληλων εργαλείων ανάλυσης και τη δημιουργία προγνωστικών μοντέλων προσφέρει τη δυνατότητα στις επιχειρήσεις του κλάδου να προβλέψουν με μεγάλη επιτυχία το βαθμό του κινδύνου, επιτρέποντας στις ασφαλιστικές εταιρείες να καθορίσουν με μεγαλύτερη ακρίβεια το ύψος των ασφαλίστρων και να επιτύχουν έτσι ανταγωνιστικό πλεονέκτημα. Επιπλέον, η αξιοποίηση των μεγάλων δεδομένων κρίνεται ιδιαίτερα επωφελής για την αποτελεσματικότερη διαχείριση των αποζημιώσεων με απλοποίηση διαδικασιών, αποτρέποντας χρονοβόρες ενέργειες και μειώνοντας δραστικά το χρόνο διακανονισμού των ζημιών αλλά και για τη μείωση του κόστους αποζημίωσης καθώς πλέον υπάρχει η δυνατότητα να καθοριστεί το ακριβές ποσό που απαιτείται για την αποκατάσταση μιας ζημιάς. Σημαντική χρήση των Big Data στον κλάδο της Ασφάλισης είναι η ανίχνευση και πρόληψη της απάτης. Οι επιχειρήσεις του κλάδου χρησιμοποιούν την ανάλυση δεδομένων με παράλληλη εφαρμογή τεχνικών και προηγμένων αλγορίθμων, προκειμένου να εντοπίσουν με επιτυχία ενδεχόμενη απάτη. Τα big data με τη συμβολή αλγορίθμων Μηχανικής Μάθησης βοηθούν τις εταιρείες να εντοπίζουν μοτίβα και να ανιχνεύουν έγκαιρα την ασφαλιστική απάτη, προστατεύοντας με αυτόν τον τρόπο τα κέρδη τους, μειώνοντας το κόστος και διαφυλάττοντας τη φήμη και την πελατεία τους. Γενικότερα, οι ασφαλιστικές εταιρείες μετασχηματίζουν το επιχειρηματικό τους μοντέλο και βελτιστοποιούν τις επιχειρηματικές τους λειτουργίες, μέσω χρήσης τεχνολογιών Big Data και ψηφιοποίησης διαδικασιών, μεγιστοποιώντας την αποδοτικότητα και την αποτελεσματικότητά τους [17].

- **Εκπαίδευση**

Ο τομέας της Εκπαίδευσης αποτελεί σημαντική πηγή μεγάλων ποσοτήτων δεδομένων τα οποία προέρχονται από βάσεις δεδομένων εκπαιδευτικών συστημάτων. Η αξιοποίηση των Μεγάλων Δεδομένων συνεισφέρει στην πρόβλεψη των επιδόσεων των μαθητών και αναδεικνύει τους παράγοντες που επηρεάζουν τις επιδόσεις αυτές, αναλύοντας συμπεριφορές των ίδιων των μαθητών. Με την ανάλυση των εκπαιδευτικών δεδομένων λαμβάνονται χρήσιμες πληροφορίες που αφορούν το προφίλ του κάθε μαθητή – σπουδαστή σχετικά με τη μαθησιακή συμπεριφορά του, τις δεξιότητές του, τη νοοτροπία του, τη γνωστική του πρόοδο, τα ισχυρά του σημεία, τις αδυναμίες του, τις επιθυμίες του κλπ. Στόχος είναι η κατανόηση των αναγκών του κάθε εκπαιδευόμενου και η προσφορά προσωποποιημένης διδασκαλίας η οποία προσαρμόζεται στις ιδιαίτερες ανάγκες του. Αυτό βοηθά στην ανάπτυξη νέων πιο αποτελεσματικών μοντέλων διδασκαλίας και οδηγεί στη βελτίωση του εκπαιδευτικού συστήματος. Μέσω της ανάλυσης των εκπαιδευτικών δεδομένων αναδεικνύονται επίσης οι τάσεις της ηλεκτρονικής μάθησης με διαδικτυακή εκπαίδευση, αντικαθιστώντας τον παραδοσιακό τρόπο εκπαίδευσης. Η μέθοδος αυτής της εκπαίδευσης εφαρμόστηκε στις μέρες μας κατά την περίοδο της πανδημικής κρίσης. Πλήθος μελετών και αναλύσεων πραγματοποιήθηκαν οι οποίες σχετίζονται με την εξ αποστάσεως διαδικτυακή διδασκαλία αλλά και με τις δυνατότητες βελτίωσης του συγκεκριμένου τρόπου μάθησης. Επιπλέον η γνώση που προέρχεται από τη διαχείριση των Μεγάλων Δεδομένων συμβάλλει στην αξιολόγηση της ακαδημαϊκής προόδου των φοιτητών και στην πρόβλεψη μελλοντικής τους απόδοσης επιτυγχάνοντας με αυτόν τον τρόπο την αξιολόγηση της

Πανεπιστημιακής διαδικασίας και τη βελτιστοποίηση και ανανέωση των προγραμμάτων σπουδών, αυξάνοντας το ποσοστό αποφοίτησης. Γενικότερα, η γνώση που προέρχεται από εκπαιδευτικά δεδομένα και η ερμηνεία της γνώσης αυτής συντελεί στη λήψη σημαντικών αποφάσεων, στην πολυ-επίπεδη βελτίωση των εκπαιδευτικών δομών καθώς και στο μετασχηματισμό της μαθησιακής και παιδαγωγικής προσέγγισης.

2.2 Επιχειρηματική Ευφυΐα (Business Intelligence)

2.2.1 Τι είναι Επιχειρηματική Ευφυΐα

Η Επιχειρηματική Ευφυΐα ορίζεται ως ένα σύνολο από μεθόδους ανάλυσης, τεχνολογίες, ικανότητες και στρατηγικές, οι οποίες στόχο έχουν την επεξεργασία των διαθέσιμων δεδομένων και την εξαγωγή χρήσιμης πληροφορίας από αυτά, για την υποστήριξη της διαδικασίας λήψης επιχειρηματικών αποφάσεων. Ένας άλλος συγγενής όρος, ο οποίος γνωρίζει ιδιαίτερη διάδοση τον τελευταίο καιρό είναι η «Αναλυτική των Επιχειρήσεων» (Business Analytics). Η Επιχειρηματική Ευφυΐα επιτρέπει σε μια επιχείρηση ή έναν οργανισμό να μαθαίνει, να αντιλαμβάνεται καταστάσεις και συμβάντα, να σκέφτεται αφαιρετικά, να προβλέπει τάσεις και μελλοντικά συμβάντα, να σχεδιάζει και να καινοτομεί. Η παραγόμενη πληροφορία, μετουσιώνεται σε γνώση που αξιοποιείται από τα διοικητικά στελέχη, ώστε να δρομολογηθούν κατάλληλες δράσεις, που θα οδηγήσουν στον καθορισμό και την επίτευξη επιχειρηματικών στόχων, με τρόπο αποτελεσματικό και αποδοτικό [18].

Στη σημερινή εποχή η Επιχειρηματική Ευφυΐα είναι μείζονος σημασίας, για τη σωστή λειτουργία και την επιτυχή δραστηριότητα μιας επιχείρησης ή ενός οργανισμού, γνωρίζει πολλά πεδία εφαρμογής στη σύγχρονη επιχείρηση και βρίσκεται στο επίκεντρο του ενδιαφέροντος του επιχειρηματικού κόσμου.

Εικόνα 2.4 Business Intelligence

Πηγή: http://qph.fs.quoracdn.net/main-qimg-a067001e88db7bbd18b5ac6902927654

2.2.2 Συστήματα Επιχειρηματικής Ευφυΐας

Σε ένα συνεχώς μεταβαλλόμενο κόσμο, οι επιχειρήσεις θα πρέπει να βασίζονται στα συστήματα Επιχειρηματικής Ευφυΐας ώστε να έχουν τη δυνατότητα να ανταποκρίνονται καλύτερα στις

σύγχρονες τάσεις και τις μελλοντικές εξελίξεις. Οι σύγχρονες επιχειρήσεις έχουν στη διάθεσή τους τεράστιους όγκους δεδομένων που προέρχονται τόσο από εσωτερικές όσο και από εξωτερικές πηγές. Τα δεδομένα αυτά μπορεί να είναι αδόμητα, ποικιλόμορφα, να προέρχονται από απομακρυσμένα συστήματα της επιχείρησης ή και να περιέχουν ελλιπή στοιχεία. Παράλληλα όμως περιέχουν σημαντικές και πολύτιμες πληροφορίες για τις επιχειρήσεις. Τα συστήματα Επιχειρηματικής Ευφυΐας συγκεντρώνουν και επεξεργάζονται δεδομένα από διάφορες πηγές, παρέχοντας χρήσιμες πληροφορίες προς αξιοποίηση. Έτσι οι επιχειρήσεις τροφοδοτούνται έγκαιρα με γνώση, την οποία χρησιμοποιούν για τη λήψη αποφάσεων στρατηγικού προσανατολισμού. Η παροχή κατάλληλης πληροφόρησης αποτελεί καθοριστικό παράγοντα για τη λήψη επιτυχημένων αποφάσεων και τα συστήματα Επιχειρηματικής Ευφυΐας συμβάλλουν σε αυτήν την κατεύθυνση, προσφέροντας ποιοτική πληροφόρηση και μειώνοντας το βαθμό αβεβαιότητας κατά τη διαδικασία λήψης αποφάσεων [19]. Παράλληλα, η υιοθέτηση συστημάτων Επιχειρηματικής Ευφυΐας από τις επιχειρήσεις συμβάλλει στη βελτίωση των επιδόσεών τους αλλά και στην εξασφάλιση του ανταγωνιστικού τους πλεονεκτήματος. Η πορεία προς την ανακάλυψη γνώσης επιτυγχάνεται με τη χρήση εξελιγμένων τεχνικών και εργαλείων, εκτελώντας υψηλού επιπέδου ανάλυση των δεδομένων, με μεθόδους που προέρχονται από την Τεχνητή Νοημοσύνη και τη Μηχανική Μάθηση. Επιπλέον οι μεθοδολογίες που προσφέρει η Εξόρυξη Δεδομένων είναι ιδιαίτερα ικανές να δίνουν ολοκληρωμένες λύσεις Επιχειρηματικής Ευφυΐας, συμβάλλοντας με αυτόν τον τρόπο στη βελτιστοποίηση της απόδοσης και στην ενδυνάμωση της αποτελεσματικότητας των σύγχρονων επιχειρήσεων.

2.2.3 Οφέλη χρήσης της Επιχειρηματικής Ευφυΐας

Τα βασικά οφέλη που προσφέρονται για μια επιχείρηση από την αξιοποίηση συστημάτων επιχειρηματικής ευφυΐας είναι τα ακόλουθα:

- Παροχή υψηλής ποιότητας, ομοιογενών και συνεκτικών δεδομένων .
- Ανάδειξη της ουσιαστικής πληροφορίας με ταυτόχρονο και βασικό μέλημα την έγκαιρη πληροφόρηση.
- Βελτίωση της ποιότητας των αποφάσεων. Η αναβαθμισμένη και έγκαιρη πληροφόρηση συμβάλλει στη λήψη βελτιωμένων αποφάσεων.
- Συμβολή στη διαμόρφωση των στρατηγικών στόχων. Τα συστήματα επιχειρηματικής ευφυΐας απευθύνονται κυρίως στα υψηλά ή και κορυφαία στελέχη των επιχειρήσεων. Στο επίπεδο αυτό λαμβάνονται οι στρατηγικές αποφάσεις. Αξιοποιώντας την ποιοτική πληροφόρηση που αντλείται από τα συστήματα αυτά, καθορίζονται οι στρατηγικοί στόχοι.
- Απόκτηση ανταγωνιστικού πλεονεκτήματος. Η βελτίωση των αποφάσεων και μέσω αυτού η αύξηση της αποτελεσματικότητας και αποδοτικότητας της επιχείρησης, καθώς και ο καθορισμός σωστών στρατηγικών στόχων, μπορούν να οδηγήσουν σε αυξημένη ανταγωνιστικότητα, η οποία αποτελεί μόνιμη επιδίωξη κάθε επιχείρησης.
- Αύξηση της παραγωγικότητας και δυνατότητα μείωσης του κόστους. Γενικότερα, επιτυχημένα συστήματα επιχειρηματικής ευφυΐας συμβάλλουν στην αύξηση των πωλήσεων και της κερδοφορίας.
- Καλύτερη κατανόηση πελατών, αγορών, ανταγωνιστών, προμηθειών και πόρων.

- Άμεση ανταπόκριση στις αλλαγές της αγοράς, την πρόβλεψη και τον αντίκτυπό τους στην επιχείρηση.
- Παρακολούθηση των αποτελεσμάτων, τονίζοντας τις τάσεις, τις απειλές και τους ενδεχόμενους κινδύνους.
- Αύξηση της πιθανότητας πρόβλεψης συμβάντων και επιχειρηματικών ευκαιριών. Η βαθύτερη κατανόηση της αγοράς επιτρέπει τον εντοπισμό επιχειρηματικών ευκαιριών. Επιπλέον, οι μέθοδοι προγνωστικής ανάλυσης, επεξεργάζονται δεδομένα και επιτρέπουν τη διατύπωση προβλέψεων.

2.2.4 Εφαρμογές Επιχειρηματικής Ευφυΐας

Η Επιχειρηματική Ευφυΐα καλύπτει ένα ευρύ φάσμα εφαρμογών στις σύγχρονες επιχειρήσεις. Στο σημείο αυτό θα προχωρήσουμε σε μια συνοπτική παρουσίαση των συνηθεστέρων πεδίων εφαρμογής της σε μια επιχείρηση.

- **Διαχείριση Εφοδιαστικής Αλυσίδας**

Τα συστήματα επιχειρηματικής ευφυΐας παρέχουν κατάλληλη πληροφόρηση σχετικά με τα επίπεδα των αποθεμάτων, σε συνδυασμό με τις ανάγκες σε υλικά απαραίτητα για την παραγωγή προϊόντων. Εντοπίζονται έγκαιρα και αντιμετωπίζονται ελλείψεις και καθυστερήσεις σε παραγγελίες, ώστε να μην επιβραδύνεται η παραγωγή. Με τον τρόπο αυτό γίνεται καλύτερος έλεγχος της ροής των προϊόντων, αυξάνεται η ικανοποίηση του πελάτη με την έγκαιρη παράδοση και μειώνονται οι ακυρώσεις και οι επιστροφές. Επιπλέον, η επιχειρηματική ευφυΐα βρίσκει εφαρμογή στην επιλογή προμηθευτών, επεξεργάζοντας και αναλύοντας στοιχεία σχετικά με την ποιότητα των προϊόντων και υπηρεσιών τους, τη συνέπειά τους, τους χρόνους παράδοσης, τις τιμολογιακές πολιτικές τους, τις εκπτώσεις και προσφορές τους, τη χρηματοοικονομική τους κατάσταση κλπ.

- **Χρηματοοικονομική Ανάλυση και Διαχείριση**

Μέσω των συστημάτων επιχειρηματικής ευφυΐας καθίσταται δυνατή η παρακολούθηση των χρηματοοικονομικών ροών. Ειδικότερα ελέγχεται η πορεία των εσόδων και εξόδων της επιχείρησης, αναλύονται τα χρηματοοικονομικά μεγέθη και συντάσσονται χρηματοοικονομικές καταστάσεις για την εκτίμηση της επίδοσης της επιχείρησης. Παράλληλα γίνεται σύγκριση με τα μεγέθη του προϋπολογισμού ώστε, αν διαπιστωθούν αποκλίσεις, να ληφθούν οι αναγκαίες ενέργειες. Γενικότερα παρακολουθείται όλη η οικονομική πορεία της επιχείρησης και ελέγχεται η κερδοφορία της ώστε να παρέχεται μια πλήρη εικόνα για τις χρηματοοικονομικές της επιδόσεις.

- **Διαχείριση Ανθρώπινων Πόρων**

Ένα εξίσου σημαντικό πεδίο εφαρμογής της επιχειρηματικής ευφυΐας στις σύγχρονες επιχειρήσεις είναι ο τομέας του ανθρώπινου δυναμικού. Τα συστήματα επιχειρηματικής ευφυΐας έχουν την ικανότητα να διαχειριστούν θέματα που σχετίζονται με τη στελέχωση των επιχειρήσεων με ανθρώπινο δυναμικό, καθώς επίσης και θέματα αμοιβών και παραγωγικότητας. Επιτυγχάνεται ευκολότερη διαχείριση της μισθοδοσίας με πιο ακριβή έλεγχο και υπολογισμό των αμοιβών, φόρων, ασφαλιστικών εισφορών, υπερωριών κλπ. Παράλληλα διευκολύνεται ο έλεγχος της παραγωγικότητας, ελέγχονται οι χρόνοι προσέλευσης και αποχώρησης των εργαζομένων,

εντοπίζονται τα ταλέντα και οι πιο παραγωγικοί εργαζόμενοι και ταυτόχρονα σχεδιάζονται δράσεις για την ανάπτυξη και εξέλιξη των ταλαντούχων εργαζομένων. Επιπλέον καθίσταται πιο εύκολος και αποτελεσματικός ο σχεδιασμός για τις προσλήψεις και την κάλυψη των αναγκών σε εργατικό δυναμικό. Μέσω της επεξεργασίας και ανάλυσης δεδομένων, τα συστήματα επιχειρηματικής ευφυΐας επιτρέπουν επίσης την πρόβλεψη των αναγκών για ανθρώπινο δυναμικό προκειμένου να προχωρήσουν σε προσλήψεις.

- **Πωλήσεις**

Η αξιοποίηση των συστημάτων επιχειρηματικής ευφυΐας από τις σύγχρονες επιχειρήσεις, συμβάλλει στην παρακολούθηση και στον έλεγχο του σημαντικού τομέα των πωλήσεων. Αναλύοντας τα στοιχεία των πωλήσεων, οι επιχειρήσεις έχουν τη δυνατότητα να εξάγουν συμπεράσματα για την πορεία των πωλήσεων και συγκρίνοντας τα στοιχεία αυτά με τους στόχους που έχει θέσει η επιχείρηση, να λαμβάνονται κατάλληλα μέτρα σε περίπτωση απόκλισης. Επιπλέον μέσω της ανάλυσης των στοιχείων αναδεικνύονται νέες ευκαιρίες, διευκολύνεται η επιλογή βέλτιστων πρακτικών για την αντιμετώπιση ενδεχόμενων προβλημάτων και καθίσταται δυνατή η πρόβλεψη των μελλοντικών πωλήσεων.

- **Marketing**

Η επεξεργασία και η ανάλυση δεδομένων που αφορούν τους πελάτες, καθώς και η άντληση πολύτιμης σχετικής πληροφορίας είναι από τα σημαντικότερα και αποδοτικότερα πεδία εφαρμογής της Επιχειρηματικής Ευφυΐας. Βασικός στόχος είναι η κατανόηση της αγοραστικής συμπεριφοράς των καταναλωτών και η αναγνώριση των αναγκών και των προτιμήσεών τους. Οι πληροφορίες αυτές επιτρέπουν την προώθηση των πωλήσεων και την αξιοποίηση νέων ευκαιριών. Η επιχειρηματική ευφυΐα μπορεί να βελτιώσει την εμπειρία των πελατών και να ανταποκριθεί έγκαιρα και αποτελεσματικά στις μεταβαλλόμενες απαιτήσεις τους. Παράλληλα προσφέρει τη δυνατότητα στις επιχειρήσεις να συγκεντρώνουν πληροφορίες σχετικές με τις τάσεις στην αγορά και να καταλήγουν σε καινοτόμα προϊόντα ή υπηρεσίες. Επιπλέον, με τη χρήση των τεχνικών επιχειρηματικής ευφυΐας μπορεί να γίνει πολύ επιτυχημένη ανάλυση τμηματοποίησης της αγοράς, εντοπισμός δηλαδή συνόλων πελατών με ομοειδή χαρακτηριστικά και καταναλωτική συμπεριφορά. Αυτή η πληροφορία αξιοποιείται με τη διοργάνωση στοχευμένων διαφημιστικών εκστρατειών. Η αξιολόγηση των αποτελεσμάτων διαφημιστικών εκστρατειών είναι ένας ακόμα τομέας που διευκολύνεται με τη χρήση της επιχειρηματικής ευφυΐας. Επιλεγμένες διαφημιστικές δράσεις αποτιμώνται σε σχέση με το κόστος τους και τα οφέλη που απέφεραν, και γίνεται σύγκριση των πραγματικών αποτελεσμάτων με τα προϋπολογισμένα μεγέθη. Με τον τρόπο αυτό επιτυγχάνεται βελτιστοποίηση των διαφημιστικών πρακτικών [20].

- **Διαχείριση & Έλεγχος Απάτης**

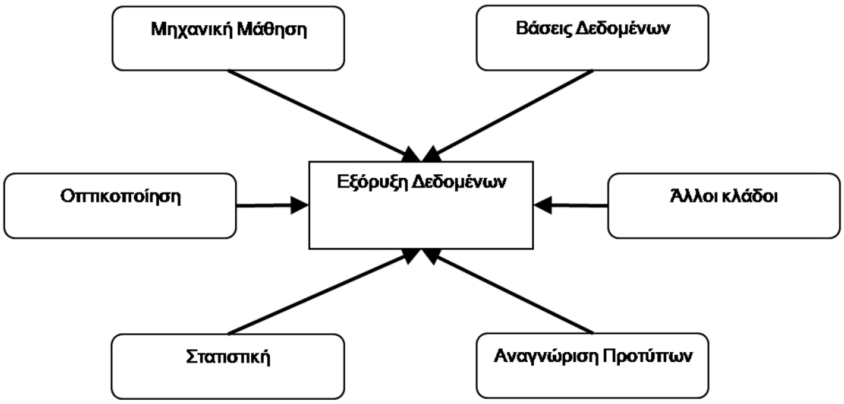
Μία πολύ σημαντική χρήση της επιχειρηματικής ευφυΐας στις σύγχρονες επιχειρήσεις είναι ο εντοπισμός και η ανακάλυψη παράνομων ενεργειών. Ένα σύστημα ανίχνευσης απάτης είναι ικανό, μέσω της επεξεργασίας και της ανάλυσης εξαιρετικά μεγάλου όγκου δεδομένων να αναγνωρίσει κακόβουλες πράξεις και πρακτικές. Οι επιχειρήσεις σήμερα συνειδητοποιούν όλο και περισσότερο την αναγκαιότητα ύπαρξης εξελιγμένων συστημάτων επιχειρηματικής ευφυΐας

για τον έγκαιρο εντοπισμό και την πρόβλεψη απάτης με σκοπό να διασφαλίσουν τα έσοδά τους, να αποτρέψουν αρνητικές συνέπειες από φαινόμενα απάτης και να προστατεύσουν τη φήμη τους.

ΚΕΦΑΛΑΙΟ 3: Εξόρυξη Δεδομένων – Μηχανική Μάθηση

3.1 Εξόρυξη Δεδομένων (Data Mining)

Καθώς ο παραγόμενος όγκος των δεδομένων, προερχόμενος από διάφορες πηγές, αυξάνεται συνεχώς με ταχείς ρυθμούς, κρίνεται αναγκαία η διαχείριση και η αξιοποίηση των δεδομένων αυτών με το βέλτιστο τρόπο ώστε να παραχθούν πολύτιμες και αξιόπιστες πληροφορίες. Η Εξόρυξη Δεδομένων αποτελεί τεχνική επεξεργασίας και ανάλυσης μεγάλης κλίμακας δεδομένων, τα οποία είναι αποθηκευμένα σε ογκώδεις βάσεις δεδομένων, που οδηγεί στην ανακάλυψη χρήσιμης γνώσης, μετατρέποντας τα δεδομένα σε αξία και πηγή πλούτου πληροφοριών, αναδεικνύοντας έτσι τη σημαντικότητα και τη συνεισφορά τους στη λήψη βέλτιστων αποφάσεων. Οι τεχνικές της εξόρυξης γνώσης οδηγούν στην ανάδειξη προτύπων και μοντέλων με ευρεία εμβέλεια σε διάφορα ερευνητικά πεδία. Έτσι, η εξόρυξη γνώσης από δεδομένα ορίζεται ως η διαδικασία εξαγωγής υπονοούμενης γνώσης, άγνωστης αλλά ενδεχομένως χρήσιμης γνώσης υπό τη μορφή συσχετίσεων, προτύπων και τάσεων, μέσω διαδικασίας εξέτασης, ανάλυσης και επεξεργασίας βάσεων δεδομένων, αντλώντας μεθοδολογίες από τη Στατιστική, τη Μηχανική Μάθηση, την Αναγνώριση Προτύπων, την τεχνολογία Βάσεων Δεδομένων, την Οπτικοποίηση και άλλων κλάδων (π.χ. τεχνητή νοημοσύνη) (Larose, 2004) [21].



Εικόνα 3.1 Η Εξόρυξη Δεδομένων ως αποτέλεσμα συμβολής άλλων κλάδων

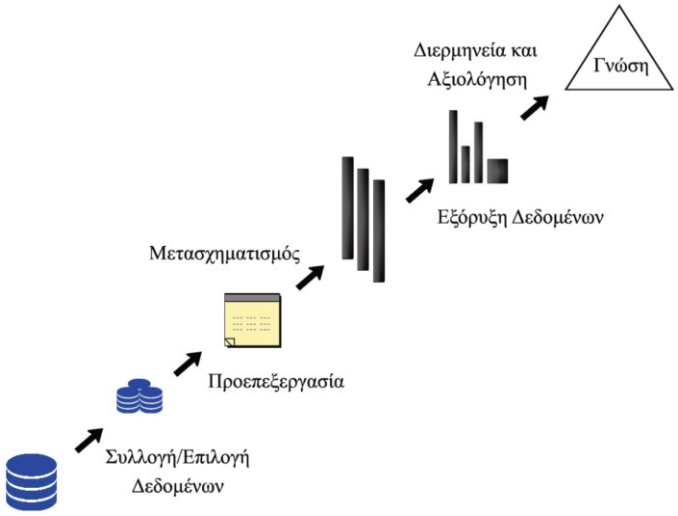
Πηγή: Κύρκος Ε., 2015

3.1.1 Στάδια της Διαδικασίας Ανακάλυψης Γνώσης

Η Ανακάλυψη Γνώσης από Βάσεις Δεδομένων αποτελείται από συγκεκριμένα στάδια. Πρόκειται για την αποκάλυψη ή παραγωγή λειτουργικής γνώσης μέσα από την ανάλυση των δεδομένων. Τα βασικά στάδια είναι:

1. Συλλογή Δεδομένων (Data Collection)
2. Προεπεξεργασία Δεδομένων (Preprocessing)
3. Μετασχηματισμός Δεδομένων (Transformation)

4. Εξόρυξη Δεδομένων (Data Mining)
5. Διερμηνεία και Αξιολόγηση (Interpretation/Evaluation)



Εικόνα 3.2 Βασικά στάδια Ανακάλυψης Γνώσης

Πηγή: Βερύκιος, Καγκλής, Σταυρόπουλος, 2015

- **Συλλογή/Επιλογή Δεδομένων**

Το πρώτο στάδιο είναι η συλλογή δεδομένων που σχετίζονται με το στόχο της ανάλυσης. Τα δεδομένα προέρχονται από πολλές και ετερογενείς πηγές. Η επιλογή των κατάλληλων δεδομένων είναι πολύ σημαντική για να διεξαχθεί επιτυχώς η Εξόρυξη Δεδομένων. Θα πρέπει να επιλεχθούν δεδομένα τα οποία θεωρούνται ότι περιέχουν ουσιαστική πληροφορία που σχετίζεται με τη συγκεκριμένη ανάλυση. Ενδεχόμενα προβλήματα που μπορεί να προκύψουν κατά τη συλλογή δεδομένων, αναλαμβάνει να τα αντιμετωπίσει το επόμενο στάδιο.

- **Προεπεξεργασία Δεδομένων**

Το δεύτερο στάδιο της προεπεξεργασίας δεδομένων αποτελεί μια πολύ σημαντική διαδικασία καθώς τα αρχικά δεδομένα πάσχουν από διαφόρων ειδών προβλήματα όπως πρόβλημα χαμένων τιμών, πρόβλημα θορύβου, πρόβλημα γενικότερων ασυνεπειών κλπ. Γενικά απαιτείται καθαρισμός και απομάκρυνση προβληματικών σημείων στο σύνολο των δεδομένων. Προβληματικά δεδομένα τα οποία περιέχουν λανθασμένες, ακραίες ή χαμένες τιμές μπορεί να αποπροσανατολίσουν τους αλγορίθμους εξόρυξης και να οδηγήσουν στην εξαγωγή άκυρων και λανθασμένων προτύπων.

- **Μετασχηματισμός Δεδομένων**

Ο μετασχηματισμός των δεδομένων αποτελεί το τρίτο στάδιο της ανακάλυψης γνώσης από βάσεις δεδομένων. Πρόκειται ουσιαστικά για τη μετατροπή των δεδομένων κάτω από ένα κοινό πλαίσιο για επεξεργασία. Για παράδειγμα μπορεί να γίνει μετατροπή αριθμητικών τιμών σε άλλες πιο «κατάλληλες» αριθμητικές τιμές (κανονικοποίηση) ή μετασχηματισμός συνεχών τιμών σε ονομαστικές τιμές (διακριτοποίηση).

- **Εξόρυξη Δεδομένων**

Σε αυτό το στάδιο εφαρμόζεται κάποιος αλγόριθμος για την παραγωγή ενός μοντέλου. Έχοντας καθαρίσει και μετασχηματίσει τα δεδομένα, είναι έτοιμα να χρησιμοποιηθούν από κάποιον αλγόριθμο ώστε να δημιουργηθεί ένα μοντέλο συνήθως κατηγοριοποίησης ή πρόβλεψης. Στόχος είναι το μοντέλο που θα δημιουργηθεί με βάση κάποια γνωστά δεδομένα, να μπορεί να δώσει απάντηση για την τιμή ενός χαρακτηριστικού – μεταβλητής στόχου, για νέα, άγνωστα δεδομένα.

- **Διερμηνεία και Αξιολόγηση**

Στο τελευταίο στάδιο γίνεται η διερμηνεία και η αξιολόγηση των αποτελεσμάτων που παρήχθηκαν από την όλη διαδικασία. Με την ολοκλήρωση αυτού του σταδίου εξάγονται τα συμπεράσματα και ακολουθεί η λήψη αποφάσεων και η ανάληψη δράσης.

3.1.1.1 Τύποι Μοντέλων

Τα μοντέλα που παράγονται από το στάδιο της Εξόρυξης Δεδομένων διακρίνονται σε δύο βασικούς τύπους: Τα μοντέλα πρόβλεψης (predictive) και τα περιγραφικά μοντέλα (descriptive).

- **Μοντέλα πρόβλεψης (predictive)**

Στόχος ενός μοντέλου πρόβλεψης είναι να προβλέψει τιμές για ένα συγκεκριμένο χαρακτηριστικό που παρουσιάζει ενδιαφέρον και που πιθανώς βασίζεται στη συμπεριφορά άλλων χαρακτηριστικών. Η πρόβλεψη δηλαδή αφορά στη χρήση κάποιων μεταβλητών μέσω των τιμών των οποίων μπορεί να εκτιμηθεί η άγνωστη ή μελλοντική τιμή ενός άλλου γνωρίσματος. Το χαρακτηριστικό πάνω στο οποίο πρόκειται να γίνει πρόβλεψη χαρακτηρίζεται συνήθως ως η στοχευμένη ή εξαρτώμενη μεταβλητή (μεταβλητή - στόχος) ενώ τα υπόλοιπα χαρακτηριστικά (attributes) που χρησιμοποιούνται για την πρόβλεψη χαρακτηρίζονται ως επεξηγηματικές ή ανεξάρτητες μεταβλητές.

- **Μοντέλα περιγραφικά (descriptive)**

Ένα περιγραφικό μοντέλο βρίσκει πρότυπα (patterns) ή σχέσεις (relations) που ενυπάρχουν στα δεδομένα και μελετά τις ιδιότητές τους ώστε να δοθεί μια αιτιολόγηση της συμπεριφοράς τους [22].

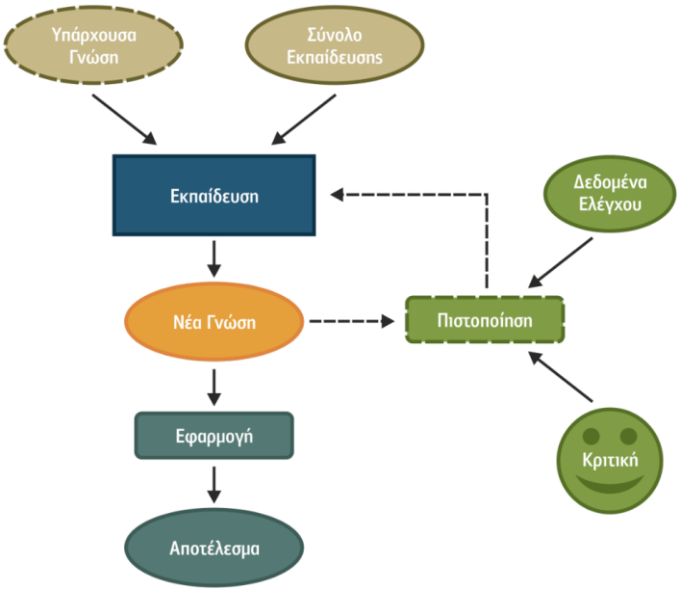
3.2 Μηχανική Μάθηση (Machine Learning)

Η Μάθηση (Learning) είναι μία από τις θεμελιώδεις ιδιότητες της νοήμονος συμπεριφοράς του ανθρώπου. Οι επιστήμονες του χώρου της Τεχνητής Νοημοσύνης προσπάθησαν να δημιουργήσουν υπολογιστικά συστήματα ικανά να μάθουν, να επιτύχουν δηλαδή, τη λεγόμενη Μηχανική Μάθηση (Machine Learning). Η Μηχανική Μάθηση έχει ως σκοπό τη δημιουργία μηχανών ικανών να μαθαίνουν, να βελτιώνουν δηλαδή την απόδοσή τους σε κάποιους τομείς μέσω της αξιοποίησης προηγούμενης γνώσης και εμπειρίας. Σύμφωνα με τον Mitchell (1997), «ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από την εμπειρία E ως προς κάποια κλάση εργασιών T και μέτρο απόδοσης P, αν η απόδοσή του σε εργασίες από το T, όπως μετριέται από το P, βελτιώνεται μέσω της εμπειρίας E» [23]. Στην Επαγωγική Μάθηση (Inductive Learning), με τη διαδικασία της επαγωγής (induction) ο άνθρωπος μαθαίνει κατανοώντας το περιβάλλον του μέσω παρατηρήσεων και δημιουργεί μια απλοποιημένη (αφαιρετική) εκδοχή του που ονομάζεται

νοητικό μοντέλο (mental model). Επιπλέον, ο άνθρωπος έχει τη δυνατότητα να οργανώνει και να συσχετίζει τις εμπειρίες και τις παρατηρήσεις του, δημιουργώντας νέες δομές που ονομάζονται νοητικά πρότυπα (mental patterns), με αξιοποίηση και του επαγωγικού (από το ειδικό στο γενικό – από τα εμπειρικά δεδομένα στη συναγωγή γενικής πρότασης) και του απαγωγικού συλλογισμού (από το γενικό στο ειδικό).

Στη μηχανική μάθηση μια μηχανή έχει την ικανότητα να δημιουργεί νέα μοντέλα και πρότυπα μάθησης από συγκεκριμένα παραδείγματα. Επομένως Μηχανική Μάθηση ονομάζεται η ικανότητα ενός υπολογιστικού συστήματος να δημιουργεί μοντέλα ή πρότυπα από ένα σύνολο δεδομένων. Ως ένας τομέας της επιστήμης των υπολογιστών, άρρηκτα συνδεδεμένος με την τεχνητή νοημοσύνη, η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων οι οποίοι έχουν την ικανότητα να εξάγουν πρότυπα και να δημιουργούν μοντέλα ικανά για να κάνουν χρήσιμες προβλέψεις, αποκτώντας και ενσωματώνοντας γνώση από σύνολα δεδομένων.

Η βασικότερη φάση κάθε αλγορίθμου είναι η εκπαίδευση, όπου ο αλγόριθμος χρησιμοποιεί ως είσοδο ένα σύνολο δεδομένων εκπαίδευσης (training set) προς επίτευξη του σκοπού του, τη δημιουργία νέας γνώσης. Επιπλέον, μπορεί είτε να χρησιμοποιήσει λιγότερο ή περισσότερο την υπάρχουσα γνώση είτε να μην τη χρησιμοποιήσει καθόλου. Την εκπαίδευση ακολουθεί η φάση της πιστοποίησης της παραγόμενης νέας γνώσης. Η πιστοποίηση πραγματοποιείται αρχικά με τη βοήθεια δεδομένων ελέγχου (test data) και στη συνέχεια μέσω κριτικής που κάνει ο ίδιος ο χρήστης βάσει των γνώσεων που διαθέτει για το πρόβλημα που επιχειρεί να λύσει ο αλγόριθμος. Τέλος, η νέα γνώση δίνεται προς χρήση σε εφαρμογές στις οποίες είναι απαραίτητη για να λυθούν πραγματικά προβλήματα [24].



Εικόνα 3.3 Φάσεις Μηχανικής Μάθησης

Πηγή: Γεωργούλη Α., 2015

3.2.1 Μέθοδοι Μηχανικής Μάθησης

Μία σημαντική διάκριση των μεθόδων μηχανικής μάθησης είναι από τον τρόπο με τον οποίο γίνεται η διαδικασία της μάθησης, δηλαδή με επίβλεψη, χωρίς επίβλεψη κλπ. Στο πεδίο της Μηχανικής Μάθησης υπάρχει ένας κατάλληλος τρόπος μάθησης προκειμένου να επιλυθεί ένα ενδεχόμενο πρόβλημα και για κάθε μέθοδο μάθησης υπάρχει τουλάχιστον ένας κατάλληλος αλγόριθμος που μπορεί να χρησιμοποιηθεί. Κάθε αλγόριθμος έχει μια διαφορετική προσέγγιση σχετικά με τον τρόπο υπολογισμού του μοντέλου πρόβλεψης και η καταλληλότητα του αλγορίθμου εξαρτάται από το είδος του προβλήματος που πρέπει να επιλυθεί, τη φύση των δεδομένων και το πλήθος των χαρακτηριστικών και περιπτώσεων. Οι δύο βασικές μέθοδοι Μηχανικής Μάθησης είναι οι ακόλουθες:

- **Επιβλεπόμενη μάθηση (Supervised learning)**

Στόχος της επιβλεπόμενης μάθησης είναι να δημιουργήσει μία συνάρτηση με την οποία να προσεγγίσει καλύτερα τη σχέση μεταξύ εισόδου και εξόδου που παρατηρείται στο σύνολο δεδομένων που της δίνεται. Πιο συγκεκριμένα, ο αλγόριθμος μηχανικής μάθησης εκπαιδεύεται ώστε να κατασκευάσει μία συνάρτηση η οποία αντιστοιχίζει τις εισόδους ενός συνόλου δεδομένων με τις γνωστές επιθυμητές εξόδους. Στη συνέχεια μπορεί να προβεί σε πρόβλεψη τιμών λαμβάνοντας νέες εισόδους για τις οποίες δεν είναι γνωστές οι τιμές εξόδων τους. Οι νέες είσοδοι θα ταξινομηθούν βασισμένες σε προηγούμενα δεδομένα εκπαίδευσης. Στην επιβλεπόμενη μάθηση διακρίνονται οι δύο παρακάτω κατηγορίες προβλημάτων:

1. Προβλήματα ταξινόμησης/κατηγοριοποίησης
   Η ταξινόμηση αφορά στη δημιουργία μοντέλων πρόβλεψης διακριτών τιμών (κλάσεις/κατηγορίες).
2. Προβλήματα παλινδρόμησης.
   Η παλινδρόμηση αφορά στη δημιουργία μοντέλων πρόβλεψης αριθμητικών – συνεχών τιμών.

- **Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)**

Στη μη επιβλεπόμενη μάθηση το σύστημα προσπαθεί να ανακαλύψει συσχετίσεις και ομάδες από τα δεδομένα βασιζόμενο μόνο στις ιδιότητές τους. Σαν αποτέλεσμα προκύπτουν πρότυπα (περιγραφές) κάθε ένα από τα οποία περιγράφει ένα μέρος από τα δεδομένα. Οι τιμές εξόδου δεν είναι γνωστές και ο αλγόριθμος χρησιμοποιεί μόνο τα πρότυπα εισόδου χωρίς να γνωρίζει σε ποια κλάση ανήκει το καθένα. Χρησιμοποιείται κυρίως σε προβλήματα Ανάλυσης Συσχετισμών και Συσταδοποίησης.

3.2.1.1 Κατηγοριοποίηση (Classification)

Πρόκειται για επιβλεπόμενη μέθοδο μάθησης και αποτελεί μία από τις βασικότερες τεχνικές της μηχανικής μάθησης με πλήθος εφαρμογών σε διάφορους τομείς. Η διαδικασία της κατηγοριοποίησης ή ταξινόμησης περιλαμβάνει την οργάνωση ενός συνόλου παρατηρήσεων που περιγράφονται από ένα σύνολο ιδιοτήτων (attributes) ή χαρακτηριστικών (features), σε μια σειρά από προκαθορισμένες κλάσεις. Οι τεχνικές της κατηγοριοποίησης χρησιμοποιούν ένα σύνολο εκπαίδευσης (training set), όπου όλες οι παρατηρήσεις είναι ήδη συνδεδεμένες με γνωστές

κλάσεις. Ο αλγόριθμος ταξινόμησης εκπαιδεύεται από αυτό το σύνολο ώστε να κατασκευάσει ένα μοντέλο – κατηγοριοποιητή (classifier) το οποίο στη συνέχεια ταξινομεί νέες εισόδους στις κατάλληλες κλάσεις [25]. Στην περίπτωση της δυαδικής κατηγοριοποίησης (binary classification) αναφερόμαστε σε δύο κλάσεις ενώ εάν οι κλάσεις είναι περισσότερες από δύο, η μέθοδος καλείται multiclass classification.

3.2.1.2 Παλινδρόμηση (Regression)

Η παλινδρόμηση είναι μία εργασία επιβλεπόμενης μάθησης που μοιάζει πολύ με την κατηγοριοποίηση. Υπάρχει ένα γνώρισμα - στόχος, οι τιμές του οποίου υπολογίζονται από τα υπόλοιπα γνωρίσματα. Οι αλγόριθμοι παλινδρόμησης εξετάζουν τις σχέσεις μεταξύ του γνωρίσματος – στόχου και των υπολοίπων γνωρισμάτων και κατασκευάζουν ένα μηχανισμό υπολογισμού. Η διαφορά με την κατηγοριοποίηση είναι ότι στην περίπτωση της παλινδρόμησης υπολογίζονται αριθμητικές τιμές. Στις επιχειρήσεις, τεχνικές παλινδρόμησης χρησιμοποιούνται για την πρόβλεψη αριθμητικών τιμών όπως το ύψος των πωλήσεων, το ύψος των κερδών κλπ.

3.2.1.3 Συσταδοποίηση (Clustering)

Η συσταδοποίηση ή αλλιώς ομαδοποίηση ανήκει στην κατηγορία εργασιών μη επιβλεπόμενης μάθησης. Έχοντας ένα σύνολο δεδομένων, στόχος της συσταδοποίησης είναι η δημιουργία συστάδων (clusters) δηλαδή ομάδων, οι οποίες θα περιέχουν όμοια δείγματα. Η διαφορά με την κατηγοριοποίηση έγκειται στο γεγονός ότι δεν υπάρχουν κατηγορίες οι οποίες είναι εκ των προτέρων γνωστές, δεν υπάρχει δηλαδή ένα γνώρισμα στο οποίο καταγράφεται η κατηγορία των παρατηρήσεων. Στόχος των αλγορίθμων της Ανάλυσης Συστάδων είναι να μεγιστοποιήσουν την ομοιότητα εντός των ομάδων και την ανομοιότητα μεταξύ των ομάδων. Αυτό σημαίνει ότι θα πρέπει να δημιουργήσουν διακριτές ομάδες με βάση ξεκάθαρα χαρακτηριστικά που περιγράφουν την κάθε ομάδα και την κάνουν να ξεχωρίζει από τις υπόλοιπες. Αφού σχηματιστούν οι ομάδες, μπορούν να θεωρηθούν ως κατηγορίες και να δημιουργηθούν κανόνες που να τις περιγράφουν. Ένα πολύ συνηθισμένο παράδειγμα εφαρμογής της συσταδοποίησης είναι ο επιμερισμός των πελατών σε ομοειδείς ομάδες. Η εφαρμογή αυτή ονομάζεται τμηματοποίηση αγοράς και είναι κεφαλαιώδους σημασίας γιατί επιτρέπει τη διεξαγωγή στοχευμένου μάρκετινγκ. Εάν η επιχείρηση γνωρίζει τα κοινά χαρακτηριστικά μεγάλων ομάδων πελατών, μπορεί να σχεδιάσει διαφημιστικές καμπάνιες ειδικά προσαρμοσμένες στα χαρακτηριστικά και τις απαιτήσεις αυτών των ομάδων με τελικό αποτέλεσμα την αύξηση της ικανοποίησης των πελατών.

3.2.1.4 Ανάλυση Συσχετίσεων

Η εξαγωγή κανόνων συσχέτισης (Mining Association Rules) θεωρείται σημαντική διεργασία της μη επιβλεπόμενης μάθησης. Οι κανόνες συσχέτισης ανακαλύπτουν κρυμμένες συσχετίσεις, μεταξύ των γνωρισμάτων ενός συνόλου των δεδομένων. Δεδομένου ενός συνόλου από δεδομένα, ένας κανόνας συσχέτισης A→B προβλέπει την εμφάνιση των χαρακτηριστικών του συνόλου B δεδομένης της εμφάνισης των χαρακτηριστικών του συνόλου A. Για παράδειγμα, εξορύσσονται κανόνες που αναφέρουν ότι όταν αγοράζεται το προϊόν A, τότε αγοράζεται ταυτόχρονα και το προϊόν B. Η Ανάλυση κανόνων Συσχέτισης εφαρμόζεται κυρίως στον τομέα των πωλήσεων καθώς επιτρέπει τον εντοπισμό καταναλωτικών προτύπων και την καλύτερη κατανόηση των πραγματικών αναγκών των πελατών. Οι πληροφορίες αυτές χρησιμοποιούνται για προσωποποιημένη προώθηση προϊόντων στους πελάτες ενώ παράλληλα βοηθά τις επιχειρήσεις

να τοποθετήσουν κατάλληλα τα προϊόντα στα ράφια, να σχεδιάσουν την κατάλληλη καμπάνια προώθησης των προϊόντων τους και να διαχειριστούν πιο αποδοτικά τα αποθεματικά τους [26].

ΚΕΦΑΛΑΙΟ 4: Θεωρία Αλγορίθμων Μηχανικής Μάθησης

4.1 Αλγόριθμοι Μηχανικής Μάθησης/ Βαθιά Μάθηση – Συλλογική Μάθηση

Σε αυτό το κεφάλαιο θα προβούμε αρχικά στη θεωρητική παρουσίαση των εξής αλγορίθμων:

- Δέντρα Απόφασης (Decision Trees)
- Λογιστική Παλινδρόμηση (Logistic Regression)
- Naïve Bayes Classifier
- k – πλησιέστερος γείτονας – KNN

Στη συνέχεια αναφερόμαστε στη Βαθιά Μάθηση και τα Τεχνητά Νευρωνικά Δίκτυα και περιγράφουμε το Multilayer Perceptron (MLP) το οποίο αποτελεί ένα ευρέως χρησιμοποιούμενο νευρωνικό δίκτυο πολλαπλών επιπέδων. Ακολουθεί αναφορά στη Συλλογική Μάθηση και περιγραφή των αλγορίθμων AdaBoost και Gradient Boost.

4.1.1 Δέντρα Απόφασης (Decision Trees)

Ο αλγόριθμος δέντρων αποφάσεων θεωρείται από τους πιο γνωστούς αλγορίθμους επιβλεπόμενης μάθησης ο οποίος δημιουργεί μοντέλα κατηγοριοποίησης (ή παλινδρόμησης) με μορφή τη δομή δέντρου. Τα δέντρα εκφράζουν μια ιεραρχία, δηλαδή εμπεριέχουν την έννοια του ανώτερου και του κατώτερου και έχουν ορισμένο βάθος. Ένα δέντρο απόφασης αποτελείται από κόμβους που αντιστοιχούν σε κάποιο χαρακτηριστικό – μεταβλητή του συνόλου εκπαίδευσης ο καθένας και διακρίνονται στους εξής:

- **Ρίζα:** Αρχικός κόμβος ο οποίος βρίσκεται στην κορυφή του δέντρου και χωρίζει το σύνολο εκπαίδευσης σε δύο ή περισσότερα υποσύνολα.
- **Εσωτερικοί κόμβοι:** Ενδιάμεσοι κόμβοι οι οποίοι με τη σειρά τους χωρίζουν το κάθε υποσύνολο του υποδέντρου σε μικρότερα υποσύνολα/υποομάδες.
- **Φύλλα (τελικοί κόμβοι):** Κάθε φύλλο αντιπροσωπεύει μια κλάση από το διακριτό σύνολο τάξεων του συνόλου εκπαίδευσης που εκφράζεται ως η τιμή του χαρακτηριστικού – στόχου (σύμφωνα με το οποίο θέλουμε να κατατάξουμε τον πληθυσμό). Οι τελικοί κόμβοι ή φύλλα του δέντρου αποτελούν την κατάληξη μιας σειράς διακλαδώσεων και δε διασπώνται περαιτέρω.

Όλοι οι κόμβοι, εκτός από τα φύλλα, έχουν εξερχόμενες ακμές, οι οποίες αντιστοιχούν σε μια συνθήκη βάσει της οποίας γίνεται η διάσπαση των δεδομένων (συνθήκη διάσπασης). Η ανάπτυξη του δέντρου πραγματοποιείται με επαναληπτικούς διαμερισμούς του συνόλου δεδομένων εκπαίδευσης και ο αλγόριθμος βρίσκει το μοναδικό εκείνο χαρακτηριστικό που διαχωρίζει καλύτερα τα δεδομένα σε κλάσεις. Η φάση ανάπτυξης συνεχίζεται μέχρις ότου δεν υπάρξει πλέον επαρκής διάσπαση ή ικανοποιηθεί κάποιο κριτήριο τερματισμού. Συνήθως το κριτήριο τερματισμού ικανοποιείται όταν όλες οι περιπτώσεις στο σύνολο εκπαίδευσης ανήκουν σε μια διακριτή κλάση. Παράλληλα, εφαρμόζεται μια διαδικασία «κλαδέματος» κατά την οποία τα

φύλλα που δεν προσθέτουν στην ικανότητα κατάταξης του δέντρου αφαιρούνται. Με το κλάδεμα αποφεύγεται η λεγόμενη υπερπροσαρμογή (overfitting) του δέντρου, δηλαδή η εκμάθηση μεμονωμένων περιπτώσεων και επιτυγχάνεται η δημιουργία ενός δέντρου που μαθαίνει γενικεύοντας με καλύτερη ικανότητα πρόβλεψης σε άγνωστα δεδομένα.

Κριτήρια διαχωρισμού

Γνωστά κριτήρια διαχωρισμού είναι:

Κέρδος πληροφορίας (information gain): Επιλέγεται το χαρακτηριστικό που οδηγεί στη διάσπαση με τη μικρότερη εντροπία (η εντροπία είναι μέτρο της «αταξίας» των δεδομένων). Η κατασκευή ενός δέντρου αποφάσεων αφορά την εύρεση ενός χαρακτηριστικού που επιστρέφει το υψηλότερο κέρδος πληροφορίας (δηλαδή τους πιο ομοιογενείς κλάδους). Το κριτήριο αυτό έχει την τάση να ευνοεί τα χαρακτηριστικά που παίρνουν πολλές διαφορετικές τιμές.

Λόγος κέρδους πληροφορίας (gain ratio): Είναι παραλλαγή του κέρδους πληροφορίας που κανονικοποιεί την πληροφορία ανά χαρακτηριστικό ώστε να είναι αντιπροσωπευτικότερη η σύγκριση.

Gini index: Είναι ένας δείκτης της μη καθαρότητας των δεδομένων (ακαθαρσίας δείκτης) και οδηγεί στην επιλογή του χαρακτηριστικού που επιτυγχάνει τη μεγαλύτερη βελτίωση στην καθαρότητα.

Απόσταση x^2 (Chi-square): Είναι το κριτήριο που εφαρμόζεται στα δέντρα τύπου CHAID και είναι ιδιαίτερα κατάλληλο σε ονομαστικά δεδομένα.

Ακρίβεια (accuracy): Επιλέγεται το χαρακτηριστικό που βελτιώνει περισσότερο τη συνολική ακρίβεια του δέντρου.

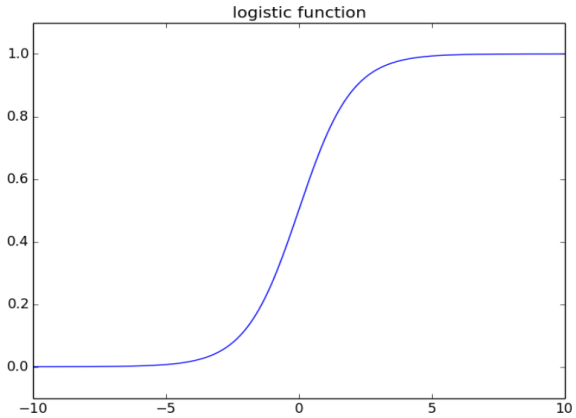
Για την επιτυχή κατασκευή ενός αποτελεσματικού δέντρου θα πρέπει να γίνουν επιπλέον διάφορες λειτουργίες όπως η επιλογή του κατάλληλου τύπου δέντρου, η επιλογή του μέγιστου βάθους, η επιλογή της ενεργοποίησης ή όχι του κλαδέματος, η ρύθμιση του ελαχίστου μεγέθους κόμβου προς διάσπαση κλπ. Σημαντικό πλεονέκτημα των δέντρων αποφάσεων ως μοντέλα κατάταξης/πρόβλεψης είναι ότι αποτελούν ιδιαίτερα παραστατικές και εύληπτες απεικονίσεις που παρέχουν άμεσα κατανοητή από τον άνθρωπο ερμηνεία της εξαχθείσας γνώσης. Υπάρχουν πολλοί αλγόριθμοι κατασκευής δέντρων αποφάσεων με διαφορετικά χαρακτηριστικά και πλεονεκτήματα ο καθένας. Αναφέρονται ως πιο διαδεδομένοι οι ID3, C4.5, CART, CHAID [27].

Αλγόριθμος C4.5

Ο αλγόριθμος C4.5 ή J48 ανήκει στη μεγάλη κατηγορία των αλγορίθμων ταξινόμησης οι οποίοι δημιουργούν δεντρικά μοντέλα ταξινομητών (δέντρα απόφασης). Είναι μια μετεξέλιξη του ID3 που παρουσιάζεται από τον ίδιο συγγραφέα (Quinlan 1993) [28]. Χρησιμοποιεί το λόγο κέρδους πληροφορίας (gain ratio) ως κριτήριο διαμερισμού. Ο διαμερισμός διακόπτεται όταν ο αριθμός των περιπτώσεων που πρόκειται να χωριστούν είναι κάτω από ένα ορισμένο όριο. Μετά τη φάση της ανάπτυξης του δέντρου εκτελείται κλάδεμα που βασίζεται στο σφάλμα. Ο C4.5 μπορεί να διαχειριστεί αριθμητικά δεδομένα. Επίσης μπορεί να διαχειριστεί από ένα σετ εκπαίδευσης τις ελλείπουσες/χαμένες τιμές χρησιμοποιώντας τους διορθωμένους λόγους κέρδους.

4.1.2 Λογιστική Παλινδρόμηση (Logistic Regression)

Η λογιστική παλινδρόμηση χρησιμοποιεί μια λογική λειτουργία για να μοντελοποιήσει μια δυαδική εξαρτώμενη μεταβλητή και αποτελεί ουσιαστικά μοντέλο ταξινόμησης των τιμών μιας μεταβλητής εξόδου Y με βάση τη θεωρία των πιθανοτήτων. Στο μοντέλο αυτό, όπου η μεταβλητή Y λαμβάνει δύο τιμές (δυαδικός χαρακτήρας), στόχος είναι η πρόβλεψη του αποτελέσματός της από το πλήθος ανεξάρτητων μεταβλητών. Εφαρμόζεται δηλαδή όταν θέλουμε να μοντελοποιήσουμε τις πιθανότητες μιας μεταβλητής απόκρισης ως συνάρτηση ορισμένων επεξηγηματικών μεταβλητών. Ο αλγόριθμος Logistic Regression είναι ένας από τους πιο συνηθισμένους αλγορίθμους Machine Learning για ταξινόμηση δύο κατηγοριών και ενδείκνυται σε περιπτώσεις όπου οι ανεξάρτητες μεταβλητές είναι κατηγορικές ή είναι συνεχείς και κατηγορικές. Σημαντική διαφορά από τη γραμμική παλινδρόμηση είναι ότι η τιμή της μεταβλητής εξόδου μπορεί να είναι κατηγορική ενώ στη γραμμική παλινδρόμηση είναι αποκλειστικά ποσοτική. Ονομάζεται λογιστική παλινδρόμηση λόγω της συνάρτησης που χρησιμοποιείται στον πυρήνα της μεθόδου, τη λογιστική συνάρτηση η οποία ονομάζεται επίσης και σιγμοειδής συνάρτηση. Είναι μια καμπύλη σχήματος S, που μπορεί να πάρει οποιοδήποτε πραγματικό αριθμό και να το χαρτογραφήσει σε τιμή μεταξύ 0 και 1 [29].



Εικόνα 4.1 Σιγμοειδής Συνάρτηση

Πηγή: J. Brownlee, 2020

4.1.3 Bayesian Classifiers – Naïve Bayes

Οι αλγόριθμοι ταξινόμησης Naïve Bayes αποτελούν μια οικογένεια απλών πιθανοτικών κατηγοριοποιητών που βασίζονται στο θεμελιωμένο θεώρημα του Bayes με ισχυρές υποθέσεις ανεξαρτησίας μεταξύ των χαρακτηριστικών. Ανεξαρτησία μεταξύ χαρακτηριστικών σημαίνει πως η ύπαρξη του ενός χαρακτηριστικού δεν επηρεάζεται ή συσχετίζεται με την ύπαρξη του άλλου. Ο κατηγοριοποιητής Bayes προβλέπει την πιθανότητα μιας δεδομένης εγγραφής/παρατήρησης να ανήκει σε μία από τις προκαθορισμένες κλάσεις. Οι Naïve Bayes αποτελούν μια απλουστευμένη εκδοχή των Bayesian δικτύων και έχουν προσαρμοστεί για τη διαδικασία κατηγοριοποίησης δεδομένων σε βαθμό που τους καθιστά ανταγωνιστικούς έναντι εξελιγμένων μεθόδων κατηγοριοποίησης όπως των Δέντρων Αποφάσεων και των Νευρωνικών Δικτύων.

Θεώρημα Bayes

Το θεώρημα Bayes στη θεωρία πιθανοτήτων και στη στατιστική έχει τη δυνατότητα να υπολογίζει την υπό συνθήκη πιθανότητα δύο γεγονότων A και B P(A/B) η οποία ορίζεται ως η πιθανότητα να συμβεί το γεγονός A δεδομένου ότι ισχύει ή έχει συμβεί το γεγονός B (δηλαδή ότι το B είναι αληθές). Η πιθανότητα αυτή εκφράζεται μαθηματικά από τον τύπο: $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$ όπου P(A) και P(B) η πιθανότητα των γεγονότων A και B αντίστοιχα και P(B/A) η πιθανότητα του γεγονότος B δεδομένου του A να είναι αληθές. Ο Naïve Bayes κατηγοριοποιητής αποτελεί άμεση εφαρμογή του θεωρήματος του Bayes όπου υποθέτουμε ότι B αποτελεί μια παρατήρηση του συνόλου δεδομένων και A είναι η υπόθεση ότι η παρατήρηση αυτή ανήκει στην κλάση C_i. Έτσι εάν B είναι ένα διάνυσμα n τιμών X=$(X_1, X_2, \dots, X_n)$ και υποθέτοντας ότι υπάρχουν m κλάσεις $c_1, c_2, \dots, c_m$ τότε σύμφωνα με το θεώρημα του Bayes, η πιθανότητα να ανήκει η παρατήρηση B στην κλάση C_i υπολογίζεται από τη σχέση: $P(C_i|B) = \frac{P(B)P(B|C_i)}{P(C_i)}$. Αφού γίνει ο υπολογισμός της πιθανότητας για κάθε κλάση C_i, τότε το μοντέλο κατηγοριοποιεί την παρατήρηση στην κλάση με τη μεγαλύτερη πιθανότητα.

Πλεονεκτήματα αλγορίθμου Naïve Bayes

- Είναι απλός, γρήγορος και εξαιρετικά κλιμακωτός αλγόριθμος
- Μπορεί να χρησιμοποιηθεί για ταξινόμηση τόσο σε δυαδικές όσο και σε περισσότερες κλάσεις.
- Παρέχει διαφορετικούς τύπους αλγορίθμων όπως Gaussian NB, Multinomial NB, Bernoulli NB.
- Μπορεί να εκπαιδευτεί εύκολα και σε μικρό σύνολο δεδομένων.
- Παρέχει μεγάλη επιλογή για προβλήματα ταξινόμησης κειμένου. Είναι δημοφιλής στην ταξινόμηση μηνυμάτων spam.

Μειονεκτήματα Naïve Bayes

Θεωρεί ότι όλα τα χαρακτηριστικά είναι ασυσχέτιστα οπότε δεν μπορεί να μάθει τη συσχέτιση μεταξύ τους. Μπορεί να μάθει μεμονωμένα χαρακτηριστικά γνωρίσματα χωρίς όμως να μπορεί να καθορίσει τη μεταξύ τους σχέση [30].

4.1.4 K – πλησιέστερος γείτονας (K – Nearest Neighbor - KNN)

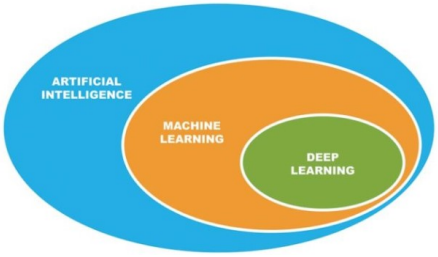
Ο αλγόριθμος K-Nearest Neighbor, γνωστός και ως KNN, αποτελεί έναν από τους πιο γνωστούς κατηγοριοποιητές στο πεδίο του Machine Learning. Βασίζεται στην έννοια της εγγύτητας και ταξινομεί ένα νέο σημείο/δείγμα σύμφωνα με την πλειοψηφία των γειτόνων του, δηλαδή το σημείο κατατάσσεται στην κλάση με τους περισσότερους k πλησιέστερους γείτονες με βάση κάποιο μέτρο απόστασης. Η τιμή του k, η οποία αντιστοιχεί στο πλήθος των πλησιέστερων γειτόνων, αποτελεί σημαντική παράμετρο για την αποτελεσματικότητα του αλγορίθμου και καθορίζεται από το χρήστη. Επίσης η μετρική της απόστασης παίζει πολύ σημαντικό ρόλο στην απόδοση του αλγορίθμου. Ως μετρική απόστασης συνήθως χρησιμοποιείται η Ευκλείδεια απόσταση όμως υπάρχουν και άλλες επιλογές όπως η απόσταση Minkowski, η απόσταση Manhattan, η απόσταση Mahalanobis και η απόσταση Hamming. Για τον προσδιορισμό του k

πλησιέστερου γείτονα κάποιου σημείου του συνόλου δεδομένων, πρέπει να υπολογιστεί ένα από τα παραπάνω μέτρα απόστασης. Ο αλγόριθμος KNN αποκαλείται «τεμπέλης μαθητής» (lazy learner) επειδή δε μαθαίνει αμέσως από το σύνολο εκπαίδευσης αλλά αποθηκεύει το σύνολο δεδομένων κατά τη διάρκεια του χρόνου εκπαίδευσης και δεν εκτελεί κανέναν υπολογισμό. Η διαδικασία εκμάθησης αναβάλλεται σε μια στιγμή που ζητείται πρόβλεψη για νέα δεδομένα. Δεν δημιουργεί δηλαδή εκ των προτέρων ένα μοντέλο και όλα τα σημεία δεδομένων χρησιμοποιούνται τη στιγμή της πρόβλεψης. Ο συγκεκριμένος αλγόριθμος είναι εύκολος στην κατανόηση και απλός στην εφαρμογή και μπορεί να χρησιμοποιηθεί τόσο για προβλήματα ταξινόμησης (συχνότερα) όσο και για προβλήματα παλινδρόμησης. Σημαντικό μειονέκτημά του είναι το υψηλό κόστος υπολογισμού καθώς αποθηκεύει όλα τα δεδομένα εκπαίδευσης και απαιτείται υπολογισμός της απόστασης μεταξύ των σημείων δεδομένων για όλα τα δείγματα εκπαίδευσης. Επιπλέον απαιτεί υψηλή αποθήκευση μνήμης, παρουσιάζει δυσκολία στον προσδιορισμό της κατάλληλης τιμής του k και η πρόβλεψη συνήθως είναι αργή [31].

4.2 Βαθιά Μάθηση

Η Τεχνητή Νοημοσύνη αναφέρεται στην ικανότητα των υπολογιστικών συστημάτων να αναπαράγουν τις γνωστικές λειτουργίες ενός ανθρώπου, επιτρέποντας στους υπολογιστές να μιμούνται την ανθρώπινη συμπεριφορά. Η Μηχανική Μάθηση αποτελεί μέρος της Τεχνητής Νοημοσύνης και αναφέρεται στην ικανότητα των υπολογιστικών συστημάτων να μαθαίνουν, ώστε να εξελίσσονται και να βελτιώνονται αυτόματα κατά την εκτέλεση μιας συγκεκριμένης εργασίας, μέσω της εμπειρίας, χωρίς να υπάρχει ανάγκη να προγραμματιστούν. Η διαδικασία μάθησης ξεκινά με δεδομένα προκειμένου να αναζητηθούν μοτίβα μέσα σε σύνολα δεδομένων και μέσω της χρήσης αλγορίθμων να κατασκευαστούν μοντέλα ώστε να μπορούν να προβούν σε σχετικές προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.

Η Βαθιά Μάθηση (Deep Learning) είναι ένα υποείδος της μηχανικής μάθησης, η οποία εμπνεύστηκε από τη δομή του ανθρώπινου εγκεφάλου και χρησιμοποιεί Τεχνητά Νευρωνικά Δίκτυα πολλαπλών επιπέδων [32]. Συνδυάζει την πρόοδο της υπολογιστικής ισχύος και των ειδικών τύπων νευρωνικών δικτύων ώστε να μαθαίνει περίπλοκα πρότυπα σε μεγάλες ποσότητες δεδομένων. Ο όρος "βαθύ" (deep) είναι ένας τεχνικός όρος που αναφέρεται στον αριθμό των κρυφών επιπέδων (hidden layers) που περιλαμβάνει ένα δίκτυο. Εάν υπάρχουν περισσότερα από τρία κρυμμένα επίπεδα, τότε ταξινομείται ως δίκτυο βαθιάς μάθησης και ονομάζεται βαθύ νευρωνικό δίκτυο (Deep Neural Network). Η βαθιά μάθηση χρησιμοποιώντας τα νευρωνικά δίκτυα διδάσκει στις μηχανές να αυτοματοποιούν τις εργασίες που εκτελούνται από τους ανθρώπους. Τα βαθιά νευρωνικά δίκτυα πολλαπλών κρυφών επιπέδων, επεξεργάζονται τα δεδομένα εισόδου με περισσότερες μαθηματικές λειτουργίες και εξαιτίας αυτού έχουν μεγαλύτερη υπολογιστική πολυπλοκότητα. Τόσο η μηχανική όσο και η βαθιά μάθηση αποτελούν υποσύνολα της τεχνητής νοημοσύνης.



Εικόνα 4.2 Η Μηχανική Μάθηση και η Βαθιά Μάθηση ως υποσύνολα της Τεχνητής Νοημοσύνης

Πηγή: nowmag.gr/machine-learning

4.2.1 Τεχνητά Νευρωνικά Δίκτυα

Τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks - ANNs) είναι υπολογιστικά μοντέλα που εμπνέονται από βιολογικές διαδικασίες μάθησης και βασίζονται στη λειτουργία του ανθρώπινου εγκεφάλου. Απαρτίζονται από απλούς υπολογιστικούς κόμβους που καλούνται νευρώνες (neurons), οι οποίοι συνδέονται μεταξύ τους δημιουργώντας ένα δίκτυο. Κάθε διασύνδεση νευρώνων στο δίκτυο χαρακτηρίζεται από κάποιο βάρος w_i που προσαρμόζεται κατά τη φάση εκπαίδευσης του δικτύου. Η εκπαίδευση συνολικά συνιστά τη διαδικασία προσαρμογής όλων των βαρών του δικτύου. Οι νευρώνες οργανώνονται σε τρία επίπεδα.

- **Επίπεδο εισόδου (input layer):** Είναι το πρώτο επίπεδο που αποτελείται από νευρώνες οι οποίοι θα χρησιμοποιηθούν για την εισαγωγή των δεδομένων στο νευρωνικό δίκτυο τα οποία θα υποστούν επεξεργασία. Οι νευρώνες εισόδου ανταποκρίνονται στις ανεξάρτητες ή προβλεπτικές μεταβλητές οι οποίες χρησιμεύουν για την πρόβλεψη των εξαρτημένων μεταβλητών.
- **Ενδιάμεσο επίπεδο (hidden layer):** Είναι το επίπεδο που συνδέεται με το προηγούμενο και το επόμενο από αυτό επίπεδο και ονομάζεται και κρυφό επίπεδο. Οι κρυφοί νευρώνες επεξεργάζονται την πληροφορία που λαμβάνουν από τους νευρώνες εισόδου και στη συνέχεια μεταφέρουν την επεξεργασμένη πληροφορία στο επόμενο επίπεδο. Μπορεί να υπάρχουν περισσότερα του ενός κρυφά επίπεδα υπολογιστικών νευρώνων.
- **Επίπεδο εξόδου (output layer):** Είναι το τελευταίο επίπεδο του δικτύου που από αυτό προκύπτουν τα αποτελέσματα μετά την εκπαίδευσή του [33].

Αρχικά οι νευρώνες πολλαπλασιάζουν κάθε είσοδό τους με το αντίστοιχο βάρος υπολογίζοντας το ολικό άθροισμα. Στη συνέχεια το άθροισμα τροφοδοτεί τη συνάρτηση ενεργοποίησης (activation function) για την παραγωγή της εξόδου. Οι τιμές που μπορούν να πάρουν τα σήματα εξόδου είναι ανάλογες με τη συνάρτηση ενεργοποίησης που εφαρμόζεται. Για παράδειγμα όταν εφαρμόζεται σιγμοειδής συνάρτηση, η έξοδος μπορεί να είναι μεταξύ του 0 και του 1. Αν και υπάρχουν αρκετές πιθανές επιλογές για τις συναρτήσεις, η σιγμοειδής συνάρτηση χρησιμοποιείται πιο συχνά καθώς είναι απλή, μη γραμμική και έχει παρόμοια συμπεριφορά με τους πραγματικούς νευρώνες. Τα νευρωνικά δίκτυα κατηγοριοποιούνται ως εξής:

- **Εμπρόσθια τροφοδότηση (feed forward):** Η ροή της πληροφορίας είναι μονοκατευθυνόμενη από την είσοδο στο κρυφό επίπεδο και στη συνέχεια στο επίπεδο εξόδου χωρίς να υπάρχει κάποια ανατροφοδότηση από την έξοδο. Αυτά τα δίκτυα

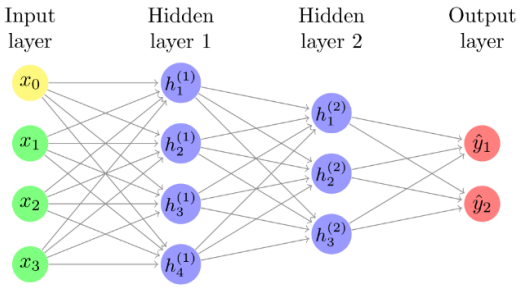
διακρίνονται σε μονοεπίπεδα (perceptron) ή πολυεπίπεδα (Multilayer Perceptron - MLP) και περιλαμβάνουν ένα ή περισσότερα κρυφά επίπεδα.

- **Ανατροφοδότηση (Feed Back):** Είναι τα δίκτυα που περιέχουν τουλάχιστον έναν κόμβο ανατροφοδότησης έτσι ώστε η πληροφορία στην έξοδο να τροφοδοτείται στις εισόδους των κόμβων των προηγούμενων επιπέδων.

Τα νευρωνικά δίκτυα είναι μη γραμμικά μοντέλα και αποτελούν πολύτιμα εργαλεία για τη μοντελοποίηση πολύπλοκων προβλημάτων επιδεικνύοντας υψηλές επιδόσεις σε διάφορες εφαρμογές, σε τομείς όπως η Ιατρική, η οικονομία, η βιομηχανία, αλλά και στο πεδίο των επιστημών γενικότερα.

4.2.1.1 Multilayer Perceptron (MLP)

Το Multilayer perceptron model είναι ένα από τα δημοφιλέστερα νευρωνικά δίκτυα που χρησιμοποιείται συχνά και η βασική του μονάδα είναι το perceptron (αισθητήρας). Αυτή η κατηγορία νευρωνικών δικτύων περιλαμβάνει περισσότερα από ένα ενδιάμεσα – κρυφά επίπεδα υπολογιστικών νευρώνων. Ο αλγόριθμος μπορεί να αντιμετωπίσει πιο απαιτητικά προβλήματα και να οδηγηθεί στην πραγματοποίηση πιο σύνθετων υπολογισμών, εξαιτίας της ύπαρξης περισσοτέρων κρυφών επιπέδων, διότι υπάρχει αύξηση του αριθμού των νευρώνων στο δίκτυο και άρα και της πληροφορίας των βαρών. Ο MLP ανήκει στην κατηγορία νευρωνικών δικτύων πρόσθιας τροφοδότησης (feed forward) πολλαπλών επιπέδων, στα οποία η ροή των πληροφοριών μετακινείται μόνο προς μια κατεύθυνση από την είσοδο, στο κρυφό επίπεδο και στη συνέχεια στο επίπεδο εξόδου και όχι προς την αντίθετη φορά (δεν υπάρχει ανατροφοδότηση από την έξοδο). Ο κάθε νευρώνας στα κρυφά επίπεδα είναι στην ουσία ένας αισθητήρας (perceptron) ο οποίος δέχεται ως input τα χαρακτηριστικά με τα βάρη τους, τα πολλαπλασιάζει και τα προσθέτει. Στο άθροισμα προστίθεται και η σταθερά bias. Στη συνέχεια το άθροισμα περνάει από τη συνάρτηση ενεργοποίησης και τελικά αντιστοιχίζονται τα αποτελέσματα σε ένα νευρώνα εξόδου. Αυτά τα δίκτυα καλούνται πολυεπίπεδοι αισθητήρες (Multilayer Perceptrons) [34].



Εικόνα 4.3 Αναπαράσταση MLP με 2 κρυφά επίπεδα

Πηγή: tex.stackexchange.com

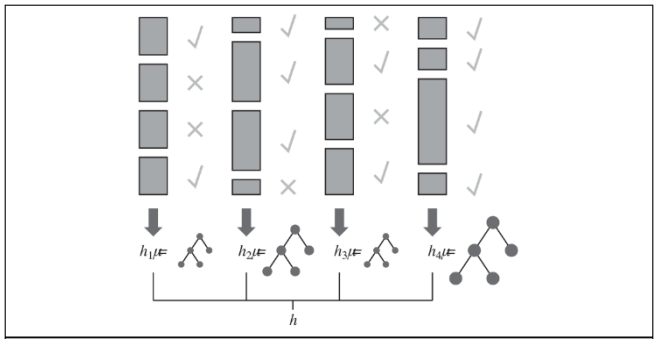
4.3 Συλλογική Μάθηση (Ensemble Learning)

Με τον όρο συλλογική μάθηση (ensemble learning) εννοούμε το συνδυασμό πολλαπλών μοντέλων μηχανικής μάθησης με σκοπό τη δημιουργία ισχυρότερων μοντέλων. Μία δημοφιλής και ευρέως χρησιμοποιούμενη μέθοδος δημιουργίας συλλογικών ταξινομητών είναι η ενδυνάμωση (boosting). Η μέθοδος Boosting αποτελεί μια τεχνική βελτίωσης της απόδοσης ενός

ασθενούς συστήματος μάθησης (weak learner). Συνδυάζοντας μοντέλα του ιδίου τύπου, όπως για παράδειγμα δέντρα αποφάσεων, είναι μια διαδοχική, επαναληπτική διαδικασία όπου κάθε νέο μοντέλο επηρεάζεται από την απόδοση όλων όσων δημιουργήθηκαν πριν από αυτό και προσπαθεί να διορθώσει τα σφάλματα του προηγούμενου. Ο επιδιωκόμενος στόχος είναι η μείωση του σφάλματος ταξινόμησης. Το τελικό μοντέλο είναι ο σταθμισμένος μέσος όρος όλων των «αδύναμων» μοντέλων. Κατά αυτόν τον τρόπο ένας ισχυρός συνδυασμένος ταξινομητής επιτυγχάνεται να κατασκευαστεί από «αδύναμους» μαθητές (weak learners) και η boosting μέθοδος μετατρέπει τους αδύναμους μαθητές σε δυνατούς. Το αποτέλεσμα είναι η δημιουργία μοντέλων με καλύτερη απόδοση καθώς βελτιώνεται η ακρίβεια της ταξινόμησης. Ωστόσο η μέθοδος boosting μερικές φορές μπορεί να οδηγήσει σε υπερπροσαρμογή των δεδομένων. Ένας μεγάλος αριθμός επαναλήψεων μπορεί να έχει ως αποτέλεσμα έναν υπερ-πολύπλοκο , συνθετικό ταξινομητή, ο οποίος είναι σημαντικά λιγότερο ακριβής από τον απλό ταξινομητή. Ένας πιθανός τρόπος αποφυγής του προβλήματος της υπερπροσαρμογής είναι η πραγματοποίηση όσο το δυνατόν λιγότερων επαναλήψεων.

Με τη μέθοδο συλλογικής μάθησης επιλέγεται μια συλλογή υποθέσεων ώστε να συνδυαστούν οι προβλέψεις τους με στόχο την ελαχιστοποίηση της πιθανότητας λανθασμένης ταξινόμησης. Η συλλογική μάθηση μειώνει την πιθανότητα λάθους δραστικά. Τα βήματα που ακολουθούνται κατά τη διαδικασία της ενδυνάμωσης είναι τα εξής:

1. Χρησιμοποιείται η έννοια του σταθμισμένου συνόλου εκπαίδευσης (weighted training set). Σε ένα τέτοιο σύνολο εκπαίδευσης, κάθε δείγμα φέρει ένα βάρος $w \geq 0$. Όσο μεγαλύτερο είναι το βάρος ενός δείγματος τόσο μεγαλύτερη είναι η σημασία του προς τη μάθηση μιας υπόθεσης. Αρχικά, όλα τα δείγματα εκπαίδευσης έχουν το ίδιο βάρος.
2. Δημιουργία πρώτης υπόθεσης όπου θα γίνει η κατηγοριοποίηση των δειγμάτων εκπαίδευσης σωστά ή λανθασμένα.
3. Πριν από κάθε επανάληψη, τα βάρη των δειγμάτων που κατέταξε λάθος η τελευταία υπόθεση, αυξάνονται και αυτών που κατέταξε σωστά μειώνονται. Δίνεται έμφαση στα δείγματα που στην τελευταία ταξινόμηση ταξινομήθηκαν λάθος. Στόχος είναι η επόμενη υπόθεση να έχει καλύτερο αποτέλεσμα σε σχέση με την πρώτη.
4. Η παραπάνω διεργασία επαναλαμβάνεται μέχρι τη δημιουργία M υποθέσεων, όπου M είναι η είσοδος στον αλγόριθμο boosting.



Εικόνα 4.4 Τρόπος λειτουργίας της μεθόδου boosting

Κάθε σκιασμένο ορθογώνιο αντιστοιχεί σε ένα δείγμα. Το ύψος του ορθογωνίου αντιστοιχεί στη βαρύτητά του. Τα σημάδια αποδοχής και άρνησης δείχνουν αν το δείγμα ταξινομήθηκε σωστά από την τρέχουσα υπόθεση. Το μέγεθος του δέντρου αποφάσεων υποδηλώνει τη βαρύτητα της υπόθεσης στην τελική συλλογή.

Πηγή: S. Russell and P. Norvig, 2010

Η τελική συλλογική υπόθεση είναι ένας σταθμισμένος πλειοψηφικός συνδυασμός όλων των υποθέσεων M, κάθε μία από τις οποίες είναι σταθμισμένη σύμφωνα με το πόσο καλά απέδωσε το σύνολο εκπαίδευσης [35].

4.3.1 AdaBoost (Adaptive Boosting)

Ο αλγόριθμος AdaBoost είναι ένας δημοφιλής αλγόριθμος κατασκευής συλλογικών ταξινομητών ο οποίος βελτιώνει τους ασθενείς ταξινομητές μέσω μιας επαναληπτικής διαδικασίας. Η βασική ιδέα είναι να δοθεί περισσότερο προσοχή σε δείγματα από το σύνολο εκπαίδευσης που είναι δυσκολότερο να ταξινομηθούν. Ο βαθμός της προσοχής που δίνεται ποσοτικοποιείται με την αντιστοίχιση βαρών σε κάθε δείγμα του συνόλου εκπαίδευσης.

Αρχικά το ίδιο βάρος αντιστοιχίζεται σε όλα τα δείγματα. Σε κάθε επανάληψη τα βάρη από όλα τα δείγματα που δεν ταξινομήθηκαν σωστά αυξάνονται και παράλληλα μειώνονται τα βάρη από τα δείγματα που ταξινομήθηκαν σωστά. Αυτό έχει ως αποτέλεσμα ο ασθενής ταξινομητής να επικεντρώνεται στα δείγματα από το σύνολο εκπαίδευσης που είναι δύσκολο να ταξινομηθούν, με το να γίνονται επιπλέον επαναλήψεις και να δημιουργούνται περισσότεροι ταξινομητές. Επίσης, σε κάθε επιμέρους ταξινομητή αντιστοιχίζεται ένα βάρος, το οποίο μετράει τη συνολική ακρίβεια του ταξινομητή και είναι συνάρτηση του συνολικού βάρους των σωστά ταξινομημένων δειγμάτων. Επομένως υψηλότερα βάρη δίνονται στους περισσότερο ακριβείς ταξινομητές. Τα βάρη αυτά χρησιμοποιούνται κατά την ταξινόμηση των νέων δειγμάτων. Η επαναλητική αυτή διαδικασία παρέχει μια ομάδα ταξινομητών που αλληλοσυμπληρώνονται. Ο AdaBoost ως αλγόριθμος ενδυνάμωσης – boosting βελτιώνει την απόδοση των ασθενών ταξινομητών αυξάνοντας την ακρίβεια της ταξινόμησης [36].

4.3.2 Gradient Boost

Ο αλγόριθμος Gradient Boost ο οποίος χρησιμοποιείται για προβλήματα κατηγοριοποίησης και παλινδρόμησης χτίζει διαδοχικά ένα μοντέλο πρόβλεψης με τη μορφή ενός συνόλου αδύναμων μοντέλων πρόβλεψης, συνήθως δέντρων αποφάσεων. Το κάθε δέντρο προσπαθεί να διορθώσει τα λάθη του προηγούμενου. Η μέθοδος Gradient Boosting αποτελεί μια επέκταση μιας μεθόδου ενδυνάμωσης ως εξής:

Gradient Boosting = Gradient Descent + Boosting

Δηλαδή αυτή η μέθοδος χρησιμοποιεί τον αλγόριθμο «gradient descent». Η διαδικασία εκμάθησης στη συγκεκριμένη μέθοδο προσαρμόζεται διαδοχικά στα μοντέλα με σκοπό να παρέχει μια ακριβέστερη εκτίμηση της τιμής της μεταβλητής εξόδου. Η κύρια διαφορά μεταξύ των αλγορίθμων ενδυνάμωσης, AdaBoost και Gradient Descent, είναι ο τρόπος με τον οποίο εντοπίζουν τις αδυναμίες των «weak learners», καθώς ο αλγόριθμος AdaBoost επιβαρύνει τα δείγματα του συνόλου εκπαίδευσης με υψηλό βάρος ενώ ο αλγόριθμος Gradient Descent χρησιμοποιεί τις κλίσεις (gradients) συναρτήσεων κόστους. Ο αλγόριθμος Gradient Boost περιλαμβάνει τα κάτωθι στοιχεία:

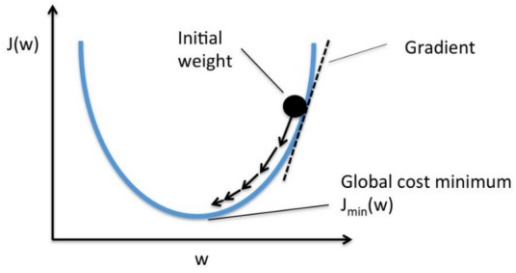
1. Βελτιστοποίηση συνάρτησης κόστους: το μοντέλο δημιουργείται σταδιακά, ελαχιστοποιώντας κάθε φορά μια διαφοροποιήσιμη συνάρτηση κόστους. Η επιλογή συνάρτησης κόστους εξαρτάται από τον τύπο προβλήματος που πρέπει να επιλυθεί.
2. Έναν «weak learner» που πρέπει να κάνει προβλέψεις: Συνήθως τα δέντρα απόφασης χρησιμοποιούνται ως «weak learners», όπως συμβαίνει και στον AdaBoost.

3. Ένα πρόσθετο μοντέλο που προσθέτει τους "weak learners" με στόχο να ελαχιστοποιηθεί η τιμή της συνάρτησης κόστους c. Στον Gradient Descent, ελαχιστοποιείται η τιμή της συνάρτησης κόστους, κατά την προσθήκη δέντρων [37].

Σχετικά με τον αλγόριθμο Gradient Descent παρακάτω παρουσιάζεται με συνοπτικό τρόπο η διαδικασία που ακολουθεί:

Πρόκειται για έναν αλγόριθμο βελτιστοποίησης που στόχος του είναι η ελαχιστοποίηση της συνάρτησης κόστους J(w) η οποία είναι παραμετροποιημένη από ένα μοντέλο παραμέτρων w.

- Αρχικά επιλέγονται βάρη w τυχαία.
- Υπολογίζεται η κλίση G της συνάρτησης κόστους. Η τιμή της εξαρτάται από τις τιμές των παραμέτρων του μοντέλου και τη συνάρτηση κόστους.
- Τα βάρη μεταβάλλονται με συνεχείς επαναλήψεις για την ελαχιστοποίηση του σφάλματος. Μία παράμετρος που επηρεάζει την ταχύτητα εκμάθησης και το τελικό ποσοστό σφάλματος είναι ο ρυθμός μάθησης a (learning rate). Όταν η τιμή του ρυθμού μάθησης είναι μικρή, η σύγκλιση θα χρειαστεί περισσότερες επαναλήψεις πάνω στο dataset με αποτέλεσμα η διαδικασία εκμάθησης να αργεί ενώ όταν η τιμή του είναι μεγάλη τότε η μεταβολή των βαρών είναι μεγάλη και αυξάνεται η πιθανότητα παρέκκλισης από το ελάχιστο.
- Η διαδικασία επαναλαμβάνεται μέχρι να σταματήσει να μειώνεται η τιμή της συνάρτησης κόστους ή να τερματίσει κάποιο κριτήριο [38].



Εικόνα 4.5 Gradient Descent

Πηγή: https://www.youtube.com/watch?v=b4Vyma9wPHo

ΚΕΦΑΛΑΙΟ 5: Μεθοδολογία Ανίχνευσης και Πρόβλεψης Απάτης

5.1 Σχετικές μελέτες

Η απάτη αποτελεί ένα πρόβλημα που αφορά και επηρεάζει όλους τους τομείς επιχειρηματικής δραστηριότητας και έχει γίνει αντικείμενο μελέτης σε αρκετές έρευνες. Πολλοί επιστήμονες και ερευνητές έχουν ασχοληθεί και εξακολουθούν να ασχολούνται με το κρίσιμο αυτό ζήτημα. Παρακάτω θα αναφερθούμε σε μερικές από αυτές τις μελέτες που αφορούν το συγκεκριμένο πρόβλημα.

Αρχικά στη μελέτη με τίτλο "Machine Learning for Prediction of Imbalanced Data: Credit Fraud Detection" [39] εξαιτίας της ανισορροπίας των κλάσεων, χρησιμοποιούνται δύο τεχνικές εξισορρόπησής τους και συγκεκριμένα οι τεχνικές SMOTE και ADASYN. Στη συνέχεια

υλοποιούνται οι αλγόριθμοι μηχανικής μάθησης random forest, k nearest neighbors, decision tree και logistic regression. Επιτυγχάνεται ακρίβεια μεγαλύτερη από 99% για τις περισσότερες περιπτώσεις και με τις δύο τεχνικές εξισορρόπησης.

Επίσης στη μελέτη με τίτλο "Pattern Analysis for Transaction Fraud Detection" [40] εφαρμόζονται οι αλγόριθμοι μηχανικής μάθησης: decision tree, logistic regression και linear regression. Πραγματοποιείται ισομοιρασμός των κλάσεων και χρησιμοποιείται η τεχνική της συσχέτισης ώστε να βρεθούν ποιες μεταβλητές σχετίζονται περισσότερο με την απάτη. Σύμφωνα με τα αποτελέσματα από την εκτέλεση των αλγορίθμων, ο αλγόριθμος decision tree ταξινόμησε σωστά ως απάτη το 83.7% όλων των περιπτώσεων ενώ πρόβλεψε λανθασμένα ως απάτη μόλις το 0.6% των περιπτώσεων. Ο αλγόριθμος logistic regression εμφάνισε ποσοστό 87% σωστά ταξινομημένων περιπτώσεων και ποσοστό 1,7% λανθασμένα ταξινομημένων περιπτώσεων ως απάτη. Ο logistic regression παρουσίασε καλύτερα αποτελέσματα σωστά ταξινομημένων περιπτώσεων σε σύγκριση με τον Decision Tree αλλά ταυτόχρονα εμφάνισε μεγαλύτερο ποσοστό λανθασμένα ταξινομημένων περιπτώσεων (ταξινόμησε λανθασμένα ως απάτη 3 φορές περισσότερο). Τέλος, το μοντέλο linear regression ταξινόμησε σωστά το 81,5% των περιπτώσεων ενώ ταξινόμησε λανθασμένα το 0,7%.

Στην έρευνα [41] δίνεται έμφαση στη δημιουργία ενός συστήματος web-based υπηρεσιών (SOAP και REST) που μπορεί να εντοπίσει απάτες σε οικονομικές συναλλαγές με χρήση πιστωτικών καρτών. Χρησιμοποιεί πέντε αλγορίθμους μηχανικής μάθησης και συγκεκριμένα τους αλγορίθμους SVM, MLP, Random Forest Regression, Autoencoder και Isolation Forest. Επιτυγχάνει ακρίβεια μεγαλύτερη του 99% για τον αλγόριθμο Random Forest Regression και μεγαλύτερη του 91% για τους υπόλοιπους αλγορίθμους και με τις δύο web-based υπηρεσίες (SOAP και REST).

Σύμφωνα με τη μελέτη [42], πολλοί τρόποι έχουν προταθεί για τον εντοπισμό της απάτης όπως το Anomaly Detection και οι τεχνικές Μηχανικής Μάθησης καθώς και μοντέλα βαθιάς Μηχανικής Μάθησης. Όπως επισημαίνεται στη συγκεκριμένη μελέτη, ο εντοπισμός της απάτης είναι ιδιαίτερα δύσκολος, καθώς η απάτη εμφανίζεται με ποικίλες μορφές και δεν υπάρχει τέλειος κανόνας διαχωρισμού της απάτης από τις κανονικές/νόμιμες περιπτώσεις. Στη συγκεκριμένη έρευνα χρησιμοποιούνται οι αλγόριθμοι Naïve Bayes, Logistic Regression, Decision Trees, Gradient Boosted Trees, Random Forest, Neural Network καθώς και οι unsupervised μέθοδοι Autoencoder και Isolation Forest για τον καθορισμό των κλάσεων που ανήκουν οι παρατηρήσεις, καθώς υπάρχουν περιπτώσεις όπου δεν έχουμε προκαθορισμένες κλάσεις. Επιτυγχάνεται Precision 89.1% για τον αλγόριθμο Logistic Regression, 88% για τον αλγόριθμο Gradient Boosting, 84% για τον αλγόριθμο Random Forest, 79.5% για τον αλγόριθμο Neural Network, 76.2% για τον αλγόριθμο Decision Tree και 54.3% για τον αλγόριθμο Naïve Bayes. Από τις unsupervised τεχνικές επιτυγχάνεται precision 94.4% για τον αλγόριθμο Autoencoder και 72.3% για τον αλγόριθμο Isolation Forest.

Επίσης στην έρευνα [43] δίνεται έμφαση σε μεθόδους βαθιάς μηχανικής μάθησης για τον εντοπισμό της απάτης. Αν και αναφέρεται ότι αλγόριθμοι όπως ο SVM, ο Decision Tree και ο Logistic Regression προτείνονται πολύ συχνά για την ανίχνευση απάτης, ακολουθώντας

παραδοσιακούς τρόπους μηχανικής μάθησης, η συγκεκριμένη έρευνα επικεντρώνεται στην εφαρμογή των συνελικτικών δικτύων CNN και LSTM προκειμένου να εντοπιστεί η απάτη.

Τέλος στη μελέτη με τίτλο "Various Methods for Fraud Transaction Detection in Credit Cards" [44] χρησιμοποιήθηκαν τα κρυμμένα μοντέλα Markov (Hidden Markov Model) καθώς και τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Network) και συγκεκριμένα ο αλγόριθμος Convolutional Neural Network – CNN και τέλος ο Logistic Regression προκειμένου να ανιχνευθεί η απάτη. Η συγκεκριμένη μελέτη χρησιμοποιεί για πρώτη φορά τα συνελικτικά δίκτυα (CNN) για τον εντοπισμό της απάτης.

5.2 Βήματα Υλοποίησης

Η ανίχνευση απάτης αποτελεί μία από τις πλέον σημαντικές χρήσεις της Μηχανικής Μάθησης. Οι τεχνικές Μηχανικής Μάθησης έχουν την ικανότητα να ανακαλύπτουν απατηλά μοτίβα και να προβλέπουν δόλιες συναλλαγές.

Στην περίπτωσή μας έχουμε να διαχειριστούμε πέντε αρχεία που περιλαμβάνουν πραγματικά δεδομένα με καταγεγραμμένες συναλλαγές, τα οποία προέρχονται από το χώρο του λιανικού εμπορίου. Από αυτά τα αρχεία επιλέγουμε το αρχείο "Invoice data" καθώς περιλαμβάνει σημαντικές πληροφορίες για τις συναλλαγές που πραγματοποιούνται όπως το συνολικό ποσό πληρωμής, το ποσό έκπτωσης, την τιμή μονάδας προϊόντος, την ποσότητα κλπ., καθώς και το αρχείο "Coupons data", το οποίο περιλαμβάνει πληροφορίες που σχετίζονται με τη χρήση εκπτωτικών κουπονιών, τα οποία ενώνουμε μέσω την κοινής μεταβλητής invoiceid (αναγνωριστικό απόδειξης). Επίσης πραγματοποιούμε ένωση των δύο προηγούμενων αρχείων με το αρχείο "Member data", μέσω της κοινής μεταβλητής loyaltymemberid (αναγνωριστικό μέλους), επιλέγοντας να λάβουμε υπόψη μας μόνο τις μεταβλητές οι οποίες σχετίζονται με στοιχεία των πελατών τα οποία θεωρούνται σημαντικά και είναι απαραίτητα για τη μελέτη μας. Τα υπόλοιπα αρχεία αγνοήθηκαν καθώς περιείχαν μη χρήσιμες πληροφορίες που δεν σχετίζονταν με την επίλυση του προβλήματος που καλούμαστε να διαχειριστούμε.

Στη συνέχεια το πρόβλημα που ανακύπτει είναι ότι έχουμε δεδομένα χωρίς προκαθορισμένες κλάσεις, δηλαδή τα δεδομένα μας δεν έχουν ετικέτες με την κλάση στην οποία ανήκει κάθε εγγραφή. Με άλλα λόγια δεν έχουμε δείγματα των οποίων οι διαθέσιμες εγγραφές έχουν χαρακτηριστεί ως ύποπτες ή νόμιμες. Στο σημείο αυτό πρέπει να επισημανθεί ότι είναι ιδιαίτερα δύσκολο και περίπλοκο να εντοπιστούν περιπτώσεις απάτης σε σύνολα δεδομένων που περιέχουν συναλλαγές πελατών, λόγω του τεράστιου όγκου δεδομένων τέτοιου τύπου που συνεχώς ανανεώνονται και εξελίσσονται, καθώς και εξαιτίας των διαφορετικών μορφών διάπραξης απάτης που μπορεί να υπάρχουν. Το γεγονός ότι είναι δυνατόν να υπάρχουν ποικίλοι τρόποι με τους οποίους μπορεί να διαπραχθεί απάτη αλλά και διαφορετικά σενάρια με τα οποία μπορεί να εμφανιστεί, κάνει τη διαδικασία ανίχνευσης απάτης ιδιαίτερα δαπανηρή καθώς απαιτείται μεγάλο κόστος προκειμένου να πραγματοποιηθεί μία λεπτομερής έρευνα για κάθε μια περίπτωση συναλλαγής ξεχωριστά. Επιπλέον, πάντα υπάρχει πιθανότητα σχετικού σφάλματος και λανθασμένης εκτίμησης καθώς είναι πιθανό ορισμένες εγγραφές να έχουν χαρακτηριστεί ως περιπτώσεις απάτης ενώ στην πραγματικότητα δεν είναι, ή κάποιες περιπτώσεις απάτης να έχουν καταγραφεί λανθασμένα ως νόμιμες. Ακόμη ενδέχεται να μην έχουν εντοπιστεί και αποκαλυφθεί ορισμένες περιπτώσεις απάτης. Για αυτόν τον λόγο στην περίπτωσή μας προσπαθούμε να

αναζητήσουμε συγκεκριμένα μοτίβα αγορών τα οποία θεωρούνται ύποπτα και ενδεχομένως αποτελούν ένδειξη απάτης, εστιάζοντας σε συγκεκριμένες μόνο περιπτώσεις.

Πιο συγκεκριμένα:

Στην πρώτη περίπτωση εντοπίζονται αποδείξεις συναλλαγών στις οποίες, αν αθροιστούν τα ποσά εκπτώσεων για κάθε συναλλαγή, διαπιστώνεται ότι υπάρχουν περιπτώσεις στις οποίες το ποσό της έκπτωσης είναι μεγαλύτερο από τη συνολική αξία του προϊόντος. Παράλληλα εντοπίζονται περιπτώσεις συναλλαγών για τις οποίες έχει χρησιμοποιηθεί εκπτωτικό κουπόνι και το συνολικό ποσό πληρωμής τους είναι αρνητικό ή μηδενικό.

Παρακάτω παραθέτουμε στιγμιότυπο του κώδικα και των αποτελεσμάτων που προκύπτουν από την πρώτη περίπτωση.

```
# Detecting fraud 1st way
fraud_df = pd.DataFrame(final_df.groupby('invoiceid', as_index=False)[['unitprice', 'discountamount', 'lineamount']]
                        .agg('sum'))

# Comparing floating point doesn't work well in pandas so we will use numpy
unitprice = np.array(fraud_df['unitprice'].tolist()).round(2)
discount = np.array(fraud_df['discountamount'].tolist()).round(2)
invoiceids = fraud_df['invoiceid'].tolist()
keepids = []

bool_compare = np.greater(discount, unitprice)

# keepids contains all invoiceids that total(sum) of discount > unitprice
for i in range(0, len(bool_compare)):
    if bool_compare[i]:
        keepids.append(invoiceids[i])
print("Length of keepids: ", len(keepids)) #899

# Keep from Coupons dataset all those invoiceids that exist in keepids list
coupon_disc = Coupons_df[Coupons_df['invoiceid'].isin(keepids)][['invoiceid', 'couponset_discount_amount']]

# InvoiceID 541932 (40,10), 642229 (40, 10), 346523 (20, 30) have 2 coupons. 50eur in coupons total.
coupon_disc = pd.DataFrame(coupon_disc.groupby('invoiceid', as_index=False)['couponset_discount_amount'].agg('sum'))

# Merge column couponset_discount_amount with fraud_df on invoiceid
fraud_df = pd.merge(coupon_disc, fraud_df, on='invoiceid', how='left')

# Take non zero coupon
fraud_df = fraud_df[fraud_df['couponset_discount_amount'] != 0]
fraud_df = fraud_df[fraud_df['lineamount'] <= 0.0]
```

Εικόνα 5.1 Fraud Detection 1st way

Στη δεύτερη περίπτωση εντοπίζονται αποδείξεις συναλλαγών στις οποίες παρόλο που αναγράφεται η συνολική ποσότητα του αγορασθέντος προϊόντος, η συνολική αξία πληρωμής του προϊόντος είναι αρνητική. Επίσης σημειώνουμε ότι λαμβάνουμε υπόψη μας τις συναλλαγές στις οποίες έχει χρησιμοποιηθεί έκπτωση με κουπόνι.

Ακολουθεί στιγμιότυπο του κώδικα και των αποτελεσμάτων που προκύπτουν από τη δεύτερη περίπτωση.

```
# Detecting fraud 2nd way
fraud_df2 = pd.DataFrame(final_df.groupby('invoiceid', as_index=False)[['quantity', 'unitprice', 'discountamount']].
                         agg('sum'))
fraud_df2 = fraud_df2[fraud_df2['unitprice'] <= 0.0]
fraud_df2 = fraud_df2[fraud_df2['quantity'] >= 0.0]

invoiceids = fraud_df2['invoiceid'].tolist()
coupon_disc = Coupons_df[Coupons_df['invoiceid'].isin(invoiceids)][['invoiceid', 'couponset_discount_amount']]

# Merge column couponset_discount_amount with fraud_df on invoiceid
fraud_df2 = pd.merge(coupon_disc, fraud_df2, on='invoiceid', how='left')
fraud_df2 = fraud_df2[fraud_df2['couponset_discount_amount'].notna()]
```

Εικόνα 5.2 Fraud Detection 2nd way

Ως τρίτη περίπτωση και αφού πρώτα ομαδοποιήσουμε τις παρατηρήσεις μας ως προς τον κωδικό του κάθε πελάτη, επιλέγουμε ως ύποπτες για απάτη τις παρατηρήσεις που έχουν αθροιστικά για κάθε πελάτη, ποσό πληρωμής (lineamount) μικρότερο του μηδενός. Επίσης ως ύποπτες παρατηρούνται περιπτώσεις όπου το συνολικό ποσό πληρωμής του πελάτη δε συμφωνεί με το ποσό που θα έπρεπε να ισχύει αφού αφαιρεθεί η έκπτωση, όπως φαίνεται και από το στιγμιότυπο του κώδικα που ακολουθεί.

```
# Detecting fraud 3rd way
fraud_df3 = pd.DataFrame(final_df.groupby('customerid', as_index=False)[['unitprice', 'discountamount', 'lineamount',
                                                                          'quantity']].agg('sum'))
fraud_df3 = fraud_df3[fraud_df3['lineamount'] < 0.0]
print("Fraud df 3 shape: ", fraud_df3.shape)

customerid = fraud_df3['customerid'].tolist()
unitprice = np.array(fraud_df3['unitprice'].tolist()).round(2)
lineamount = np.array(fraud_df3['lineamount'].tolist()).round(2)
discountamount = np.array(fraud_df3['discountamount'].tolist()).round(2)
cust_ids = []

# cust_ids will contain all customerids that total(sum) of lineamount != unitprice - discount
for i in range(0, len(customerid)):
    if not np.isclose([abs(unitprice[i] - discountamount[i])], [abs(lineamount[i])]):
        cust_ids.append(customerid[i])
```

Εικόνα 5.3 Fraud Detection 3rd way

Λαμβάνοντας υπόψη μας τις παραπάνω περιπτώσεις παρατηρούμε ότι από το συνολικό μέγεθος των 1.657.396 συναλλαγών μόνο οι 9.119 αποτελούν υποψία απάτης που αντιστοιχούν σε ποσοστό 0.55% επί των συνολικών συναλλαγών. Στη συνέχεια, ονομάζουμε την εξαρτημένη μεταβλητή – μεταβλητή στόχο "isfraud" στην οποία δίνουμε δύο τιμές, την τιμή "1" σε περίπτωση απάτης και την τιμή "0" σε περίπτωση νόμιμης συναλλαγής (no fraud). Έχοντας πλέον ορίσει τις κλάσεις στις οποίες ανήκουν οι εγγραφές, το πρόβλημά μας μετατρέπεται σε πρόβλημα δυαδικής ταξινόμησης για το αν μια συναλλαγή είναι προϊόν απάτης ή όχι (πρόβλεψη απάτης). Για την επίλυσή του θα χρησιμοποιήσουμε τη μέθοδο της κατηγοριοποίησης (classification) η οποία αποτελεί μέθοδο επιβλεπόμενης μάθησης, ακολουθώντας τα παρακάτω βήματα:

- Συλλογή/Επιλογή δεδομένων
- Προεπεξεργασία δεδομένων
- Διαδικασία επιλογής ταξινομητών – Εκπαίδευση
- Αποτελέσματα – Ερμηνεία Παραγόμενης γνώσης

Στην υλοποίηση της εκπαίδευσης και αξιολόγησης των ταξινομητών χρησιμοποιήθηκε η γλώσσα Python. Για το μέρος της μηχανικής μάθησης χρησιμοποιήθηκε η βιβλιοθήκη scikit-learn καθώς επίσης έγινε και χρήση των βιβλιοθηκών NumPy και Pandas. Πιο αναλυτικά:

Η python είναι μια αντικειμενοστραφής γενικού σκοπού και υψηλού επιπέδου γλώσσα προγραμματισμού, ευρέως χρησιμοποιούμενη η οποία δημιουργήθηκε από τον Ολλανδό Guido van Rossum και κυκλοφόρησε δημοσίως το 1991. Χαρακτηρίζεται από την εύκολη αναγνωσιμότητα του κώδικά της και την ευκολία χρήσης της σε διάφορες εφαρμογές. Η απλή και ευανάγνωστη σύνταξή της επιτρέπει στους προγραμματιστές να γράψουν γρήγορα και αποτελεσματικά τον κώδικά τους, δαπανώντας λιγότερο χρόνο για την επιδιόρθωση σφαλμάτων και αντιμετώπιση προβλημάτων. Πρόκειται για γλώσσα ανοιχτού κώδικα η οποία είναι φιλική σε οποιοδήποτε λειτουργικό σύστημα. Αποτελεί ένα πολύ ισχυρό εργαλείο για εφαρμογές στον τομέα της επιστήμης δεδομένων (data science) και χρησιμοποιείται από επιστήμονες δεδομένων για την ανάπτυξη αλγορίθμων και γενικότερα τη διαχείριση των δεδομένων. Διαθέτει πολλές ολοκληρωμένες βιβλιοθήκες διευκολύνοντας ιδιαίτερα την εκτέλεση πολύπλοκων εργασιών. Οι πιο δημοφιλείς βιβλιοθήκες για την επιστήμη δεδομένων, οι οποίες καλύπτουν κάθε στάδιο της ανάλυσης των δεδομένων είναι οι εξής:

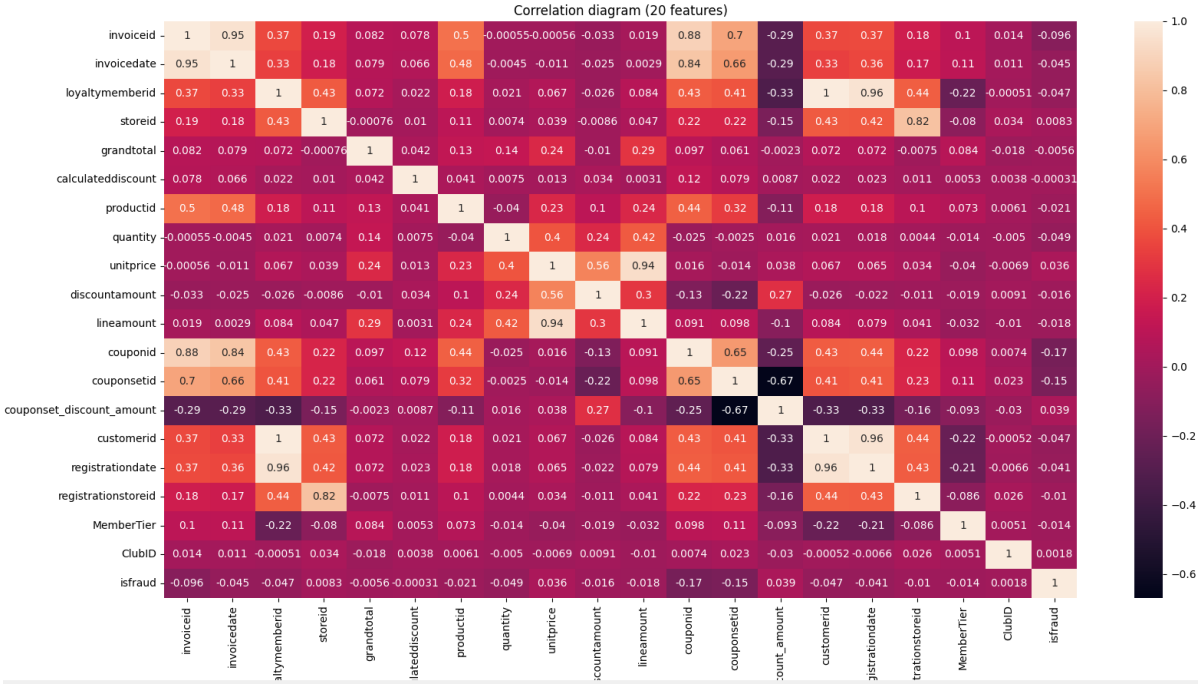
- Pandas: Πολύ χρήσιμη και διαδεδομένη βιβλιοθήκη που παρέχει υψηλής απόδοσης δομές δεδομένων (data structures) για το χειρισμό, τον καθαρισμό και την προετοιμασία δεδομένων ώστε να αποτελέσουν input για τη διαδικασία της μηχανικής μάθησης. Το pandas επιτρέπει την αποτελεσματική μορφοποίηση των δεδομένων στην κατάλληλη μορφή ώστε να είναι δυνατή η πραγματοποίηση σωστών προβλέψεων στο machine learning.
- Scikit-learn: βιβλιοθήκη μηχανικής μάθησης για δημιουργία, εκπαίδευση και υλοποίηση μοντέλων. Παρέχει ένα μεγάλο αριθμό αλγορίθμων για κατηγοριοποίηση, παλινδρόμηση και συσταδοποίηση. Παρέχει επίσης εργαλεία για την επικύρωση των αποτελεσμάτων και τη διασφάλιση της βέλτιστης επιλογής των μοντέλων. Συνδυάζεται εύκολα με τις υπόλοιπες βασικές βιβλιοθήκες της Python διευκολύνοντας την πραγματοποίηση διαφόρων εργασιών μηχανικής μάθησης.
- NumPy – Numerical Python: σημαντική βιβλιοθήκη για αριθμητικούς υπολογισμούς υψηλής απόδοσης. Η βιβλιοθήκη numpy προσθέτει υποστήριξη για μεγάλους πολυδιάστατους πίνακες και γραφήματα μαζί με μια μεγάλη συλλογή μαθηματικών συναρτήσεων που εφαρμόζονται σε αυτούς. Την πιο βασική λειτουργία της βιβλιοθήκης αποτελεί η δομή δεδομένων για πολυδιάστατους πίνακες [45].
- Imb.learn (imbalanced learn): βιβλιοθήκη η οποία προσφέρει διάφορες τεχνικές over-sampling και under-sampling που χρησιμοποιούνται συχνά σε ανισορροπία του συνόλου δεδομένων. Παράλληλα είναι συμβατή με τη βιβλιοθήκη scikit-learn.
- Matplotlib: βιβλιοθήκη η οποία παρέχει υποστήριξη για τη δημιουργία γραφικών παραστάσεων στην python υψηλής ποιότητας.

5.3 Συλλογή / Επιλογή Δεδομένων

Κατόπιν συνένωσης των αρχείων όπως προαναφέραμε και αφού πλέον είναι γνωστές οι κλάσεις των παρατηρήσεων/εγγραφών, προκύπτει το σύνολο των δεδομένων μας το οποίο αποτελείται αρχικά από 1.666.515 γραμμές και 18 στήλες που αφορούν το σύνολο των χαρακτηριστικών – μεταβλητών που λαμβάνονται υπόψη. Όπως παρατηρούμε το σύνολο δεδομένων περιέχει πολλές κενές τιμές οι οποίες δημιουργούν πρόβλημα στους κατηγοριοποιητές. Για αυτόν το λόγο

αποφασίστηκε να αφαιρεθούν οι κενές τιμές με τον ακόλουθο τρόπο. Αρχικά αντικαθιστούμε τις κενές τιμές των μεταβλητών ClubID, registrationstoreid, registrationdate και MemberTier με την πιο συχνά εμφανιζόμενη τιμή της κάθε μεταβλητής. Στη συνέχεια διαγράφουμε όλες τις γραμμές που έχουν κενές τιμές στις μεταβλητές storeid και couponset_discount_amount. Επίσης αφαιρέθηκαν και όλες οι διπλά εγγεγραμμένες παρατηρήσεις.

Στη συνέχεια εξετάζουμε την αλληλεπίδραση των χαρακτηριστικών μεταξύ τους, χρησιμοποιώντας τη μήτρα συσχέτισης (correlation matrix) όπου πρόκειται για έναν πίνακα συσχετισμού που εμφανίζει συντελεστές συσχέτισης μεταξύ διαφόρων χαρακτηριστικών.



Εικόνα 5.4 Μήτρα Συσχέτισης

Όπως παρατηρούμε υπάρχουν δύο χαρακτηριστικά με υψηλό βαθμό συσχέτισης. Οπότε μπορεί να εξαιρεθεί από το μοντέλο πρόβλεψης ένα από τα δύο με στόχο να επιλεχθούν τα καλύτερα χαρακτηριστικά. Οι μεταβλητές customerid και loyalitymemberid είχαν συσχέτιση 1 οπότε αποφασίζουμε να αφαιρεθεί μια από τις δύο. Συγκεκριμένα αφαιρέθηκε η μεταβλητή customerid. Επίσης αφαιρούμε και τη μεταβλητή isused που είχε σχεδόν παντού τιμή 1.

Έτσι προκύπτει το τελικό σύνολο δεδομένων (dataset) το οποίο αποτελείται από 196867 γραμμές και 19 μεταβλητές μαζί με τη μεταβλητή στόχο "isfraud".

Πίνακας 5.1 Παρουσίαση χαρακτηριστικών συνόλου δεδομένων

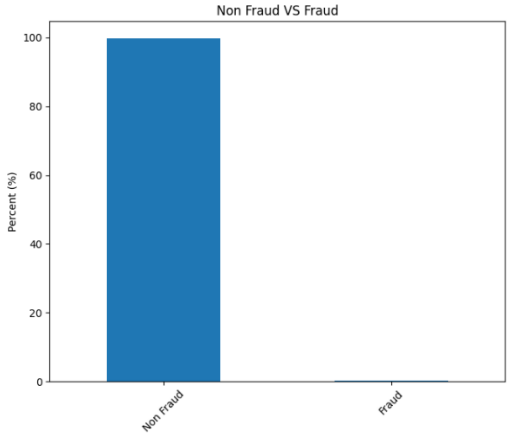
Μεταβλητή	Επεξήγηση/ Περιγραφή	Τύπος Μεταβλητής
Invoiceid	Αναγνωριστικό απόδειξης	Κλειδί, ακέραιος
invoicedate	Ημερομηνία όπου η απόδειξη εκδόθηκε	Ημερομηνία YYYY-MM-DD hh:mm:ss
loyaltymemberid	Αναγνωριστικό μέλους	Κλειδί, ακέραιος
Storeid	Αναγνωριστικό καταστήματος	Κλειδί, ακέραιος
grandtotal	Ποσό μετά τις εκπτώσεις	Δεκαδικός
Calculateddiscount	Έκπτωση	Δεκαδικός
Productid	Αναγνωριστικό προϊόντος	Κλειδί, ακέραιος
Quantity	Ποσότητα του κάθε προϊόντος στην απόδειξη	Δεκαδικός
Unitprice	Τιμή του κάθε προϊόντος στην απόδειξη	Δεκαδικός
Discountamount	Ποσό έκπτωσης για κάθε προϊόν στην απόδειξη	Δεκαδικός
Lineamount	Ποσό όπου το μέλος πλήρωσε για κάθε προϊόν στην απόδειξη	Δεκαδικός
Couponid	Αναγνωριστικό κουπονιού	Κλειδί, ακέραιος
Couponsetid	Αναγνωριστικό σετ κουπονιού	Κλειδί, ακέραιος
couponset_discount_amount	Ποσό έκπτωσης σετ κουπονιού	Ακέραιος
Registrationdate	Ημερομηνία όπου γράφτηκε το μέλος	Ημερομηνία YYYY-MM-DD hh:mm:ss
Registrationstoreid	Κατάστημα όπου το μέλος γράφτηκε	Κλειδί, ακέραιος
MemberTier	Επίπεδο όπου ανήκει το μέλος	Ακέραιος
ClubID	Club όπου ανήκει το μέλος	Ακέραιος
isfraud	Η μεταβλητή εξόδου – στόχος παίρνει την τιμή "1" σε περίπτωση απάτης και "0" σε περίπτωση νόμιμης συναλλαγής.	Δυαδικό 0,1

Όπως παρατηρούμε το σύνολο των μεταβλητών μας αποτελείται από ποσοτικές μεταβλητές δηλαδή μεταβλητές numeric – αριθμητικές.

5.4 Προεπεξεργασία Δεδομένων (Data Preprocessing)

Η προεπεξεργασία δεδομένων είναι μια πολύ σημαντική διαδικασία καθώς τα αρχικά δεδομένα πάσχουν από διαφόρων ειδών προβλήματα όπως πρόβλημα χαμένων τιμών, πρόβλημα θορύβου (σφάλματα στις τιμές των χαρακτηριστικών) που αποπροσανατολίζουν τον αλγόριθμο, προβλήματα γενικότερων ασυνεπειών και ανωμαλιών κλπ. Γενικά απαιτείται καθαρισμός των «ακάθαρτων δεδομένων» και απομάκρυνση των προβληματικών σημείων στο σύνολο δεδομένων, πριν τη δημιουργία μοντέλων μηχανικής μάθησης, ώστε να είναι κατάλληλα για χρήση. Στην περίπτωσή μας και κατόπιν της συνένωσης των αρχείων μας, παρατηρήσαμε ότι προέκυπταν κενές τιμές σε κάποιες μεταβλητές τις οποίες έπρεπε να διαχειριστούμε κατάλληλα ώστε να προχωρήσουμε σωστά στην κατηγοριοποίηση. Έτσι αποφασίστηκε να αφαιρεθούν αυτές οι κενές τιμές με τους τρόπους που προαναφέραμε, καθώς επίσης αφαιρέθηκαν και οι διπλά εγγεγραμμένες παρατηρήσεις ώστε να ενισχυθεί η απόδοση των μοντέλων και η αποτελεσματικότητα των κατηγοριοποιητών.

Στη συνέχεια, προβάλλουμε εικόνα η οποία δείχνει τη μεγάλη ανισορροπία που προκύπτει μεταξύ των κλάσεων "1" (απάτη - fraud) και "0" (νόμιμη συναλλαγή – no fraud). Συγκεκριμένα οι παρατηρήσεις της κλάσης "0" ανέρχονται σε ποσοστό 99.73% έναντι της κλάσης "1" οι οποίες ανέρχονται σε ποσοστό 0.27% που αντιστοιχεί σε μόλις 522 συναλλαγές από το συνολικό μέγεθος των 196.867.



Εικόνα 5.5 Κατανομή συναλλαγών ανά κλάση

Απάτη (fraud) – νόμιμη (no fraud)

Αυτή η ανισορροπία των κλάσεων θα οδηγούσε τα μοντέλα εκπαίδευσης σε λανθασμένη κατηγοριοποίηση καθώς παρατηρήσεις της κλάσης που υπολείπεται θα τις κατηγοριοποιούσαν ως παρατηρήσεις της κλάσης που υπερισχύει. Για αυτόν το λόγο θα πρέπει να εξισορροπήσουμε τη μεγάλη διαφορά στις παρατηρήσεις μεταξύ της κλάσης "0" και "1" ώστε να εκπαιδευτούν καλύτερα οι ταξινομητές και να βελτιώσουμε τα αποτελέσματα. Για το σκοπό αυτό χρησιμοποιούνται κυρίως οι τεχνικές over-sampling και under-sampling. Σε αυτές τις τεχνικές είτε

προσθέτουμε παρατηρήσεις στην κλάση που υπολείπεται είτε αφαιρούμε παρατηρήσεις από την κλάση που υπερισχύει. Αποφασίστηκε να εφαρμοστεί μια over-sampling τεχνική, μέσω της βιβλιοθήκης imblearn της Python, κατά την οποία προστίθενται παρατηρήσεις στη μειονεκτική κλάση. Με αυτόν τον τρόπο ο αριθμός των δόλιων συναλλαγών (απάτη) αυξάνεται από 522 σε 196.345 όσο δηλαδή και το πλήθος των νόμιμων συναλλαγών. Το τελικό σύνολο εκπαίδευσης ανέρχεται σε 392.690 εγγραφές με την αναλογία των fraud – no fraud να είναι 1:1 όπως φαίνεται στην παρακάτω εικόνα.

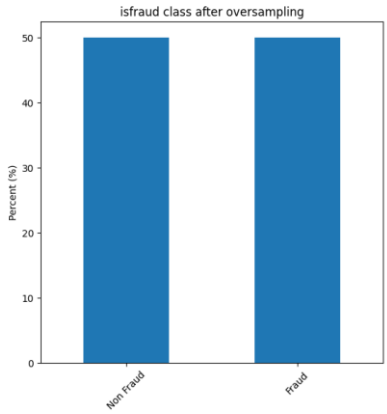
```
Both classes should have equal amount of observations:
 0    196345
1    196345
Name: isfraud, dtype: int64
```

Εικόνα 5.6 Απάτη (fraud) - Νόμιμες (No fraud)

Οι κλάσεις εξισορροπήθηκαν ως εξής:

Κλάση 0 – νόμιμη συναλλαγή (no fraud): 50%

Κλάση 1 – απάτη (fraud): 50% όπως αποτυπώνεται στο παρακάτω διάγραμμα



Εικόνα 5.7 Αναλογία μεταβλητής isfraud

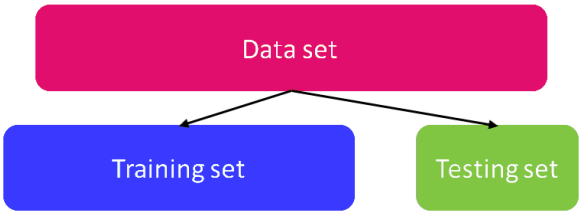
5.5 Διαδικασία Επιλογής Ταξινομητών – Εκπαίδευση

Έχοντας τελειώσει με τη διαδικασία της προεπεξεργασίας, τα δεδομένα είναι έτοιμα να χρησιμοποιηθούν από τους κατάλληλους αλγορίθμους για τη δημιουργία μοντέλων πρόβλεψης ώστε να ολοκληρωθεί η διαδικασία της κατηγοριοποίησης. Προκειμένου να επιλύσουμε το πρόβλημα δυαδικής ταξινόμησης (απάτη – μη απάτη) που καλούμαστε να αντιμετωπίσουμε, επιλέγουμε τους παρακάτω ταξινομητές: Δέντρα Απόφασης, Λογιστική Παλινδρόμηση, Naïve Bayes, K – πλησιέστερος γείτονας, Multilayer Perceptron - MLP, AdaBoost, Gradient Boost.

Για την εκπαίδευση των αλγορίθμων το σύνολο των δεδομένων διαχωρίστηκε με δύο διαφορετικές τεχνικές ως εξής:

Στην πρώτη τεχνική το σύνολο δεδομένων διασπάται σε δύο υποσύνολα όπου το πρώτο χρησιμοποιείται ως σύνολο εκπαίδευσης και το δεύτερο ως σύνολο επικύρωσης ή δοκιμής.

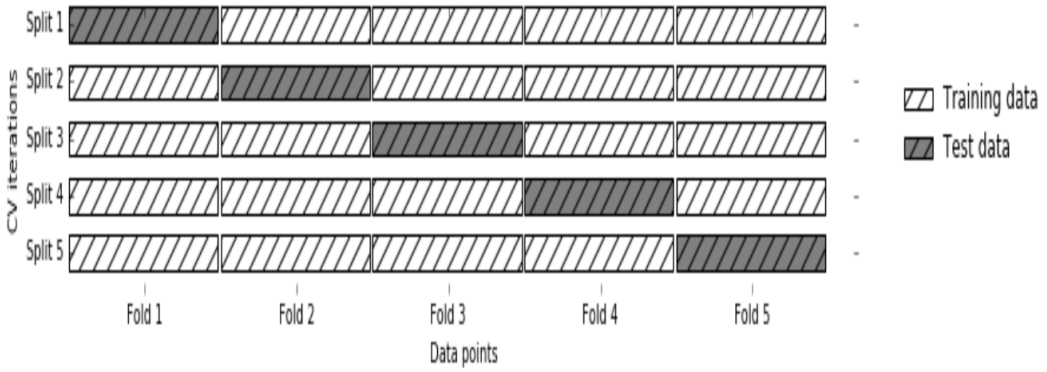
- Σύνολο εκπαίδευσης (training set): Αποτελεί το σύνολο δεδομένων πάνω στο οποίο βασίζεται η κατασκευή των μοντέλων πρόβλεψης. Πάνω σε αυτό το σύνολο δεδομένων εκπαιδεύεται ένας αλγόριθμος για τη δημιουργία προβλεπτικού μοντέλου.
- Σύνολο επικύρωσης ή δοκιμής (test set): Αποτελεί το σύνολο δεδομένων πάνω στο οποίο εξετάζεται η απόδοση των μοντέλων πρόβλεψης. Με αυτό το σύνολο δεδομένων ελέγχουμε την ακρίβεια του μοντέλου στα υπόλοιπα άγνωστα δεδομένα, εκτιμώντας με αυτόν τον τρόπο την ικανότητά του να προβλέπει σωστά την κλάση νέων – άγνωστων παρατηρήσεων.



Εικόνα 5.8 Διαχωρισμός συνόλου δεδομένων

Στη συγκεκριμένη περίπτωση το σύνολο εκπαίδευσης αποτελείται από το 80% των παρατηρήσεων και το σύνολο επικύρωσης/δοκιμής από το υπόλοιπο 20%. Ο διαχωρισμός του συνόλου των δεδομένων σε δύο υποσύνολα γίνεται με τυχαίο τρόπο.

Η δεύτερη τεχνική ονομάζεται k-fold cross validation ή διασταυρούμενη επικύρωση k φορών. Το σύνολο δεδομένων χωρίζεται σε k υποσύνολα, στη συγκεκριμένη περίπτωση σε πέντε (5), τα οποία είναι ίσου μεγέθους και ονομάζονται "folds". Χρησιμοποιούμε το ένα υποσύνολο από αυτά ως σετ δοκιμής/επικύρωσης και τα υπόλοιπα υποσύνολα ως σετ εκπαίδευσης. Η συγκεκριμένη διαδικασία επαναλαμβάνεται 5 φορές, όπου κάθε φορά σε κάθε επανάληψη, το ένα υποσύνολο διατηρείται ως σύνολο δοκιμής ενώ τα υπόλοιπα χρησιμοποιούνται για να εκπαιδεύσουν τον κατηγοριοποιητή. Η συνολική ακρίβεια πρόβλεψης του συστήματος αξιολογείται ως μέσος όρος των 5 δοκιμών. Το ποσοστό σφάλματος των διαφόρων επαναλήψεων αποτελεί το μέσο όρο του ολικού ποσοστού σφάλματος. Η εκτίμηση σφάλματος υπολογίζεται δηλαδή κατά μέσο όρο σε όλες τις δοκιμές k για να επιτευχθεί η συνολική αποτελεσματικότητα του μοντέλου.



Εικόνα 5.9 Τεχνική Cross Validation με k = 5

Κατά την εκτέλεση των αλγορίθμων είναι πολύ σημαντικό να πειραματιστούμε, μεταβάλλοντας τις τιμές των παραμέτρων των ταξινομητών, με σκοπό να βελτιώσουν την εκπαίδευσή τους και τα

αποτελέσματά τους. Στόχος είναι να επιλεχθούν οι παράμετροι με τα πιο ακριβή αποτελέσματα ανά κατηγοριοποιητή για να αξιολογηθούν.

5.5.1 Αξιολόγηση μοντέλων

Αφού δημιουργηθεί το μοντέλο, αξιολογούμε την απόδοσή του, αναδεικνύοντας με αυτόν τον τρόπο την αξιοπιστία του αλγορίθμου. Τα βασικότερα μέτρα τα οποία χρησιμοποιούνται για την αξιολόγηση της αποτελεσματικότητας ενός συστήματος είναι τα ακόλουθα:

- **Ορθότητα (Accuracy):** Εκφράζει το ποσοστό των σωστά κατηγοριοποιημένων παρατηρήσεων (θετικών και αρνητικών) προς το συνολικό αριθμό όλων των παρατηρήσεων. Αποτελεί βασικό μέτρο αξιολόγησης, αλλά σε σύνολα δεδομένων που παρουσιάζουν ισχυρή ανισορροπία κλάσεων, το Accuracy δεν μπορεί να χαρακτηριστεί ως σωστό κριτήριο για την αξιολόγηση της απόδοσης ενός αλγορίθμου καθώς ο κατηγοριοποιητής προβλέπει με υψηλά ποσοστά επιτυχίας την κλάση που υπερέχει σε σχέση με την κλάση που υπολείπεται. Σε αυτήν την περίπτωση, των μη ισορροπημένων δεδομένων, τα μέτρα της ανάκλησης (recall) και ακρίβειας (Precision) αποδίδουν καλύτερα. **Accuracy = (TP+TN)/(TP+FP+TN+FN)**
- **Ακρίβεια (Precision):** Ορίζεται ως μέτρο αποτελεσματικότητας όταν ο στόχος μας είναι ο περιορισμός των ψευδών θετικών κατηγοριοποιημένων παρατηρήσεων. Για κάθε τάξη, ορίζεται ως η αναλογία των πραγματικών θετικών προς το άθροισμα ενός αληθινού θετικού και ψευδώς θετικού (ακρίβεια θετικών προβλέψεων). **Precision = (TP/TP+FP)**
- **Ανάκληση (Recall):** Εκφράζει το ποσοστό των αληθώς θετικών κατηγοριοποιημένων παρατηρήσεων που σωστά κατηγοριοποιήθηκαν ως θετικά (True positive - TP). Για κάθε τάξη ορίζεται ως ο λόγος των πραγματικών θετικών προς το άθροισμα των πραγματικών θετικών και των ψευδών αρνητικών. **Recall = TP/TP+FN**

Ο πίνακας σύγχυσης (Confusion Matrix) αποτελεί μια διάταξη πίνακα που επιτρέπει την οπτικοποίηση της απόδοσης ενός μοντέλου χρησιμοποιώντας στοιχεία τα οποία συμβολίζονται ως εξής:

- **TP:** Πλήθος θετικών παρατηρήσεων που έχουν προβλεφθεί σωστά από το μοντέλο – αληθή θετικά
- **FN:** Πλήθος θετικών παρατηρήσεων που έχουν προβλεφθεί λανθασμένα ως αρνητικά – ψευδή αρνητικά
- **FP:** Πλήθος αρνητικών παρατηρήσεων που έχουν προβλεφθεί λανθασμένα ως θετικά – ψευδή θετικά
- **TN:** Πλήθος αρνητικών παρατηρήσεων που έχουν προβλεφθεί σωστά από το μοντέλο – αληθή αρνητικά.
- **F1 Score:** Είναι ένας σταθμισμένος αρμονικός μέσος που υπολογίζεται με βάση τα αποτελέσματα της ακρίβειας και της ανάκλησης, συνδυάζοντας τα δύο μέτρα και δίνοντας καλύτερα αποτελέσματα. Η υψηλότερη τιμή του δείχνει ότι τα δύο μέτρα (precision και recall) είναι ικανοποιητικά.
- **Καμπύλη ROC:** Ένα ισχυρό μέτρο για την εκτίμηση της ανά κλάση ακρίβειας και απόδοσης του κατηγοριοποιητή είναι η καμπύλη ROC (Receiver Operating

Characteristic) ή καμπύλη χαρακτηριστικών λειτουργίας δέκτη η οποία είναι χρήσιμο οπτικό εργαλείο που υπολογίζει τις ψευδείς θετικές (άξονας x) και τις πραγματικές θετικές τιμές (άξονας y). Για τη σύγκριση των κατηγοριοποιητών χρησιμοποιούμε το μέτρο σύγκρισης που είναι η περιοχή κάτω από την καμπύλη ROC που είναι γνωστή ως Area Under Curve (AUC). Τα μοντέλα τα οποία είναι κάτω από τη διαγώνιο της καμπύλης ROC θεωρούνται χειρότερα σε σχέση με αυτά που είναι πάνω από αυτή. Όσο ένας κατηγοριοποιητής βρίσκεται πιο κοντά στην αριστερή πάνω γωνία της καμπύλης, τόσο καλύτερη είναι αφού έχει πραγματική θετική τιμή 1 και ψευδής θετική τιμή 0. Αντίθετα, όσο πιο κοντά βρίσκεται το μοντέλο στη διαγώνιο τόσο λιγότερο ακριβές είναι, ενώ αν πέσει ακριβώς πάνω στη διαγώνιο είναι εντελώς τυχαίο.

5.6 Αποτελέσματα – Ερμηνεία Παραγόμενης Γνώσης

Ολοκληρώνοντας το έργο της εκπαίδευσης προχωρούμε στην αξιολόγηση των αποτελεσμάτων από την εκτέλεση των αλγορίθμων. Οι τιμές των παραμέτρων των ταξινομητών αποφασίστηκαν κατόπιν επαναλαμβανομένων δοκιμών και επιλέχθηκαν αυτές που συμβάλλουν στη βέλτιστη απόδοση των μοντέλων. Κατόπιν συνεχών δοκιμών, παρατηρήθηκε ότι οι ταξινομητές Logistic Regression και Multilayer Perceptron - MLP παρουσίαζαν ιδιαίτερα χαμηλές επιδόσεις ταξινόμησης. Οπότε αποφασίστηκε να χρησιμοποιηθεί η τεχνική της τυποποίησης (Standardization) με το StandardScaler της βιβλιοθήκης Scikit-learn, σύμφωνα με την οποία τα δεδομένα προσαρμόζονται σε ένα άλλο πιο κατάλληλο εύρος τιμών, ώστε να βελτιωθεί η εκπαίδευσή τους και η αποτελεσματικότητά τους. Οι ταξινομητές αξιολογήθηκαν ως προς την ορθότητα (accuracy) ακρίβεια (precision), ανάκληση (recall) και το μέτρο F-Measure (αρμονικός μέσος). Εκτελώντας τον αλγόριθμο Δέντρα Απόφασης προκύπτουν τα παρακάτω αποτελέσματα.

Πίνακας 5.2 Εξαγωγή προβλεπτικού μοντέλου με Δέντρα Απόφασης

<table>
<tr><td rowspan="2">Precision</td><td>Train 80%</td><td>92.7%</td><td colspan="2">Confusion Matrix (Train 80%)</td></tr>
<tr><td>Cross Validation 5</td><td>92.8%</td><td>36308</td><td>2843</td></tr>
<tr><td rowspan="2">Recall</td><td>Train 80%</td><td>85.8%</td><td>6013</td><td>33374</td></tr>
<tr><td>Cross Validation 5</td><td>85.8%</td><td colspan="2">Confusion Matrix (Cross Validation 5)</td></tr>
<tr><td rowspan="2">F1-Score</td><td>Train 80%</td><td>89.1%</td><td>182285</td><td>14060</td></tr>
<tr><td>Cross Validation 5</td><td>89.2%</td><td>30278</td><td>166067</td></tr>
<tr><td rowspan="2">Accuracy</td><td>Train 80%</td><td>88.72%</td><td>True Positive</td><td>False Positive</td></tr>
<tr><td>Cross Validation 5</td><td>88.71%</td><td>False Negative</td><td>True Negative</td></tr>
</table>

Παρατηρούμε ότι η προβλεπτική ικανότητα του μοντέλου αγγίζει το 88.72%. Το μοντέλο σύμφωνα με την πρώτη τεχνική διάσπασης – διαχωρισμού των δεδομένων και συγκεκριμένα του training και test set εμφανίζει 69.682 σωστά κατανεμημένες περιπτώσεις στο δείγμα και 8.856 λανθασμένα κατανεμημένες από τις 78.538 περιπτώσεις σύμφωνα με το Confusion Matrix, ενώ σύμφωνα με τη δεύτερη τεχνική του k fold - cross validation εμφανίζει 348.352 σωστά κατανεμημένες περιπτώσεις στο δείγμα και 44.338 λανθασμένα κατανεμημένες από τις 392.690 περιπτώσεις. Το Recall (Ανάκληση) ή Ευαισθησία εκφράζεται από το ποσοστό TP Rate και ανέρχεται σε ποσοστό 85.8% που αποτελεί το καλύτερο ποσοστό και από τις δύο τεχνικές. Επίσης το Precision ανέρχεται σε ποσοστό 92.8% και το F1-Score σε 89.2% που αποτελούν τα καλύτερα ποσοστά και από τις δύο τεχνικές διάσπασης.

Από την εκτέλεση του αλγορίθμου Λογιστικής Παλινδρόμησης προκύπτουν τα κάτωθι αποτελέσματα.

Πίνακας 5.3 Εξαγωγή προβλεπτικού μοντέλου με Λογιστική Παλινδρόμηση

<table>
<tr><td rowspan="2">Precision</td><td>Train 80%</td><td>84.1%</td><td colspan="2">Confusion Matrix (Train 80%)</td></tr>
<tr><td>Cross Validation 5</td><td>83.4%</td><td>32908</td><td>6243</td></tr>
<tr><td rowspan="2">Recall</td><td>Train 80%</td><td>91.0%</td><td>3269</td><td>36118</td></tr>
<tr><td>Cross Validation 5</td><td>90.9%</td><td colspan="2">Confusion Matrix (Cross Validation 5)</td></tr>
<tr><td rowspan="2">F1-Score</td><td>Train 80%</td><td>87.4%</td><td>163722</td><td>32623</td></tr>
<tr><td>Cross Validation 5</td><td>87.0%</td><td>16454</td><td>179891</td></tr>
<tr><td rowspan="2">Accuracy</td><td>Train 80%</td><td>87.89%</td><td>True Positive</td><td>False Positive</td></tr>
<tr><td>Cross Validation 5</td><td>87.50%</td><td>False Negative</td><td>True Negative</td></tr>
</table>

Παρατηρούμε ότι η προβλεπτική ικανότητα του μοντέλου αγγίζει το 87.89%. Το μοντέλο σύμφωνα με την πρώτη τεχνική διάσπασης – διαχωρισμού των δεδομένων και συγκεκριμένα του training και test set εμφανίζει 69.026 σωστά κατανεμημένες περιπτώσεις στο δείγμα και 9.512 λανθασμένα κατανεμημένες από τις 78.538 περιπτώσεις σύμφωνα με το Confusion Matrix, ενώ σύμφωνα με τη δεύτερη τεχνική του k fold - cross validation εμφανίζει 343.613 σωστά

κατανεμημένες περιπτώσεις στο δείγμα και 49.077 λανθασμένα κατανεμημένες από τις 392.690 περιπτώσεις. Το Recall (Ανάκληση) ή Ευαισθησία εκφράζεται από το ποσοστό TP Rate και ανέρχεται σε ποσοστό 91% που αποτελεί το καλύτερο ποσοστό και από τις δύο τεχνικές. Επίσης το Precision ανέρχεται σε ποσοστό 84.1% και το F1-Score σε 87.4% που αποτελούν τα καλύτερα ποσοστά και από τις δύο τεχνικές διάσπασης.

Ο αλγόριθμος Naïve Bayes αφού εκπαιδεύτηκε από το σύνολο δεδομένων εκπαίδευσης δημιούργησε ένα μοντέλο Naïve Bayes Model από το οποίο προκύπτουν τα παραπάνω αποτελέσματα.

Πίνακας 5.4 Εξαγωγή προβλεπτικού μοντέλου Naïve Bayes.

<table>
<tr><td rowspan="2">Precision</td><td>Train 80%</td><td>70.1%</td><td colspan="2">Confusion Matrix (Train 80%)</td></tr>
<tr><td>Cross Validation 5</td><td>50.8%</td><td>27458</td><td>11693</td></tr>
<tr><td rowspan="2">Recall</td><td>Train 80%</td><td>52.4%</td><td>24962</td><td>14425</td></tr>
<tr><td>Cross Validation 5</td><td>48.3%</td><td colspan="2">Confusion Matrix (Cross Validation 5)</td></tr>
<tr><td rowspan="2">F1-Score</td><td>Train 80%</td><td>60.0%</td><td>99715</td><td>96630</td></tr>
<tr><td>Cross Validation 5</td><td>49.5%</td><td>106656</td><td>89689</td></tr>
<tr><td rowspan="2">Accuracy</td><td>Train 80%</td><td>53.33%</td><td>True Positive</td><td>False Positive</td></tr>
<tr><td>Cross Validation 5</td><td>48.23%</td><td>False Negative</td><td>True Negative</td></tr>
</table>

Παρατηρούμε ότι η προβλεπτική ικανότητα του μοντέλου αγγίζει το 55.33%. Το μοντέλο σύμφωνα με την πρώτη τεχνική διάσπασης – διαχωρισμού των δεδομένων και συγκεκριμένα του training και test set εμφανίζει 41.883 σωστά κατανεμημένες περιπτώσεις στο δείγμα και 36.655 λανθασμένα κατανεμημένες από τις 78.538 περιπτώσεις σύμφωνα με το Confusion Matrix, ενώ σύμφωνα με τη δεύτερη τεχνική του k fold - cross validation εμφανίζει 189.404 σωστά κατανεμημένες περιπτώσεις στο δείγμα και 203.286 λανθασμένα κατανεμημένες από τις 392.690 περιπτώσεις. Το Recall (Ανάκληση) ή Ευαισθησία εκφράζεται από το ποσοστό TP Rate και ανέρχεται σε ποσοστό 52.4% που αποτελεί το καλύτερο ποσοστό και από τις δύο τεχνικές. Επίσης το Precision ανέρχεται σε ποσοστό 70.1% και το F1-Score σε 60.0% που αποτελούν τα καλύτερα ποσοστά και από τις δύο τεχνικές διάσπασης.

Στη συνέχεια παρουσιάζονται τα αποτελέσματα από την εκτέλεση του αλγορίθμου Κ – πλησιέστερος γείτονας (KNN)

Πίνακας 5.5 Εξαγωγή προβλεπτικού μοντέλου KNN

<table>
<tr><td rowspan="2">Precision</td><td>Train 80%</td><td>99.9%</td><td colspan="2">Confusion Matrix (Train 80%)</td></tr>
<tr><td>Cross Validation 5</td><td>99.7%</td><td>39093</td><td>58</td></tr>
<tr><td rowspan="2">Recall</td><td>Train 80%</td><td>96.9%</td><td>1265</td><td>38122</td></tr>
<tr><td>Cross Validation 5</td><td>88.6%</td><td colspan="2">Confusion Matrix (Cross Validation 5)</td></tr>
<tr><td rowspan="2">F1-Score</td><td>Train 80%</td><td>98.3%</td><td>195706</td><td>639</td></tr>
<tr><td>Cross Validation 5</td><td>93.8%</td><td>25102</td><td>171243</td></tr>
<tr><td rowspan="2">Accuracy</td><td>Train 80%</td><td>98.32%</td><td>True Positive</td><td>False Positive</td></tr>
<tr><td>Cross Validation 5</td><td>93.44%</td><td>False Negative</td><td>True Negative</td></tr>
</table>

Παρατηρούμε ότι η προβλεπτική ικανότητα του μοντέλου αγγίζει το 98.32%. Το μοντέλο σύμφωνα με την πρώτη τεχνική διάσπασης – διαχωρισμού των δεδομένων και συγκεκριμένα του training και test set εμφανίζει 77.215 σωστά κατανεμημένες περιπτώσεις στο δείγμα και 1.323 λανθασμένα κατανεμημένες από τις 78.538 περιπτώσεις σύμφωνα με το Confusion Matrix, ενώ σύμφωνα με τη δεύτερη τεχνική του k fold - cross validation εμφανίζει 366.949 σωστά κατανεμημένες περιπτώσεις στο δείγμα και 25.741 λανθασμένα κατανεμημένες από τις 392.690 περιπτώσεις. Το Recall (Ανάκληση) ή Ευαισθησία εκφράζεται από το ποσοστό TP Rate και ανέρχεται σε ποσοστό 96.9% που αποτελεί το καλύτερο ποσοστό και από τις δύο τεχνικές. Επίσης το Precision ανέρχεται σε ποσοστό 99.9% και το F1-Score σε 98.3% που αποτελούν τα καλύτερα ποσοστά και από τις δύο τεχνικές διάσπασης.

Προχωρώντας με βαθιά μάθηση (deep learning), χρησιμοποιήσαμε Τεχνητά Νευρωνικά Δίκτυα και συγκεκριμένα το Multilayer Perceptron model. Από την εφαρμογή του αλγορίθμου MLP προκύπτουν τα εξής αποτελέσματα όπως παρουσιάζονται στην κάτωθι εικόνα.

Πίνακας 5.6 Εξαγωγή προβλεπτικού μοντέλου MLP

<table>
<tr><td rowspan="2">Precision</td><td>Train 80%</td><td>99.4%</td><td colspan="2">Confusion Matrix (Train 80%)</td></tr>
<tr><td>Cross Validation 5</td><td>99.6%</td><td>38904</td><td>247</td></tr>
<tr><td rowspan="2">Recall</td><td>Train 80%</td><td>99.8%</td><td>67</td><td>39320</td></tr>
<tr><td>Cross Validation 5</td><td>100.0%</td><td colspan="2">Confusion Matrix (Cross Validation 5)</td></tr>
<tr><td rowspan="2">F1-Score</td><td>Train 80%</td><td>99.6%</td><td>195609</td><td>736</td></tr>
<tr><td>Cross Validation 5</td><td>99.8%</td><td>0</td><td>196345</td></tr>
<tr><td rowspan="2">Accuracy</td><td>Train 80%</td><td>99.60%</td><td>True Positive</td><td>False Positive</td></tr>
<tr><td>Cross Validation 5</td><td>99.81%</td><td>False Negative</td><td>True Negative</td></tr>
</table>

Παρατηρούμε ότι η προβλεπτική ικανότητα του μοντέλου αγγίζει το 99.81%. Το μοντέλο σύμφωνα με την πρώτη τεχνική διάσπασης – διαχωρισμού των δεδομένων και συγκεκριμένα του training και test set εμφανίζει 78.224 σωστά κατανεμημένες περιπτώσεις στο δείγμα και 314 λανθασμένα κατανεμημένες από τις 78.538 περιπτώσεις σύμφωνα με το Confusion Matrix, ενώ σύμφωνα με τη δεύτερη τεχνική του k fold - cross validation εμφανίζει 391.954 σωστά κατανεμημένες περιπτώσεις στο δείγμα και 736 λανθασμένα κατανεμημένες από τις 392.690 περιπτώσεις. Το Recall (Ανάκληση) ή Ευαισθησία εκφράζεται από το ποσοστό TP Rate και ανέρχεται σε ποσοστό 100.0% που αποτελεί το καλύτερο ποσοστό και από τις δύο τεχνικές. Επίσης το Precision ανέρχεται σε ποσοστό 99.6% και το F1-Score σε 99.8% που αποτελούν τα καλύτερα ποσοστά και από τις δύο τεχνικές διάσπασης.

Στη συνέχεια προχωρήσαμε στην εκτέλεση αλγορίθμων ενδυνάμωσης (boosting) – Συλλογική Μάθηση. Εκτελώντας τον αλγόριθμο AdaBoost προκύπτουν τα κάτωθι αποτελέσματα.

Πίνακας 5.7 Εξαγωγή προβλεπτικού μοντέλου AdaBoost

<table>
<tr><td rowspan="2">Precision</td><td>Train 80%</td><td>95.4%</td><td colspan="2">Confusion Matrix (Train 80%)</td></tr>
<tr><td>Cross Validation 5</td><td>94.9%</td><td>37353</td><td>1798</td></tr>
<tr><td rowspan="2">Recall</td><td>Train 80%</td><td>96.4%</td><td>1406</td><td>37981</td></tr>
<tr><td>Cross Validation 5</td><td>96.3%</td><td colspan="2">Confusion Matrix (Cross Validation 5)</td></tr>
<tr><td rowspan="2">F1-Score</td><td>Train 80%</td><td>95.9%</td><td>186321</td><td>10024</td></tr>
<tr><td>Cross Validation 5</td><td>95.6%</td><td>7170</td><td>189175</td></tr>
<tr><td rowspan="2">Accuracy</td><td>Train 80%</td><td>95.92%</td><td>True Positive</td><td>False Positive</td></tr>
<tr><td>Cross Validation 5</td><td>95.62%</td><td>False Negative</td><td>True Negative</td></tr>
</table>

Παρατηρούμε ότι η προβλεπτική ικανότητα του μοντέλου αγγίζει το 95.92%. Το μοντέλο σύμφωνα με την πρώτη τεχνική διάσπασης – διαχωρισμού των δεδομένων και συγκεκριμένα του training και test set εμφανίζει 75.334 σωστά κατανεμημένες περιπτώσεις στο δείγμα και 3.204 λανθασμένα κατανεμημένες από τις 78.538 περιπτώσεις σύμφωνα με το Confusion Matrix, ενώ σύμφωνα με τη δεύτερη τεχνική του k fold - cross validation εμφανίζει 375.496 σωστά κατανεμημένες περιπτώσεις στο δείγμα και 17.194 λανθασμένα κατανεμημένες από τις 392.690 περιπτώσεις. Το Recall (Ανάκληση) ή Ευαισθησία εκφράζεται από το ποσοστό TP Rate και ανέρχεται σε ποσοστό 96.4% που αποτελεί το καλύτερο ποσοστό και από τις δύο τεχνικές. Επίσης το Precision ανέρχεται σε ποσοστό 95.4% και το F1-Score σε 95.9% που αποτελούν τα καλύτερα ποσοστά και από τις δύο τεχνικές διάσπασης.

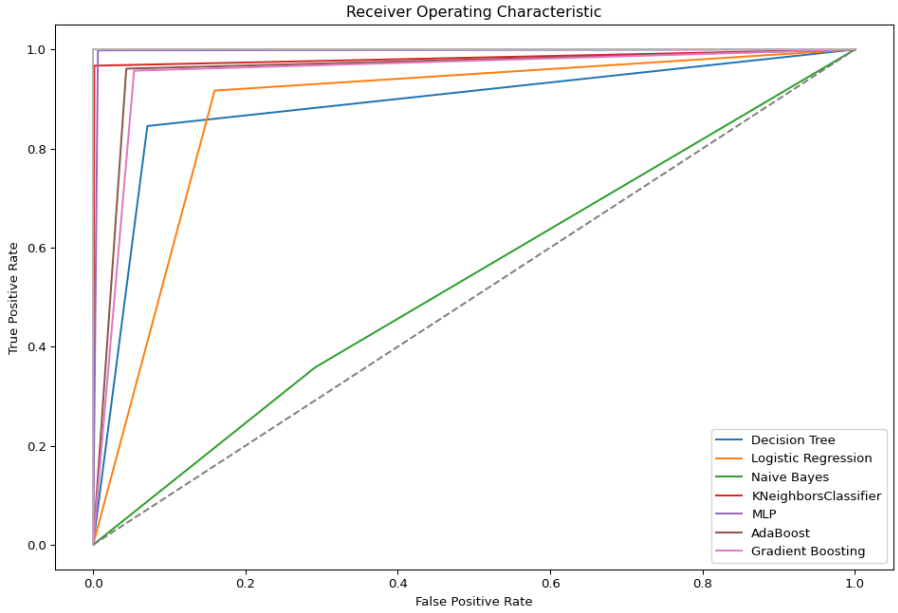
Τέλος εφαρμόζουμε έναν άλλο αλγόριθμο ενδυνάμωσης και συγκεκριμένα τον Gradient Boost, από την εκτέλεση του οποίου προκύπτουν τα παρακάτω αποτελέσματα.

Πίνακας 5.8 Εξαγωγή προβλεπτικού μοντέλου με Gradient Boost

<table>
<tr><td rowspan="2">Precision</td><td>Train 80%</td><td>94.8%</td><td colspan="2">Confusion Matrix (Train 80%)</td></tr>
<tr><td>Cross Validation 5</td><td>94.2%</td><td>37111</td><td>2040</td></tr>
<tr><td rowspan="2">Recall</td><td>Train 80%</td><td>94.7%</td><td>2063</td><td>37324</td></tr>
<tr><td>Cross Validation 5</td><td>95.1%</td><td colspan="2">Confusion Matrix (Cross Validation 5)</td></tr>
<tr><td rowspan="2">F1-Score</td><td>Train 80%</td><td>94.8%</td><td>184932</td><td>11413</td></tr>
<tr><td>Cross Validation 5</td><td>94.6%</td><td>9605</td><td>186740</td></tr>
<tr><td rowspan="2">Accuracy</td><td>Train 80%</td><td>94.78%</td><td>True Positive</td><td>False Positive</td></tr>
<tr><td>Cross Validation 5</td><td>94.65%</td><td>False Negative</td><td>True Negative</td></tr>
</table>

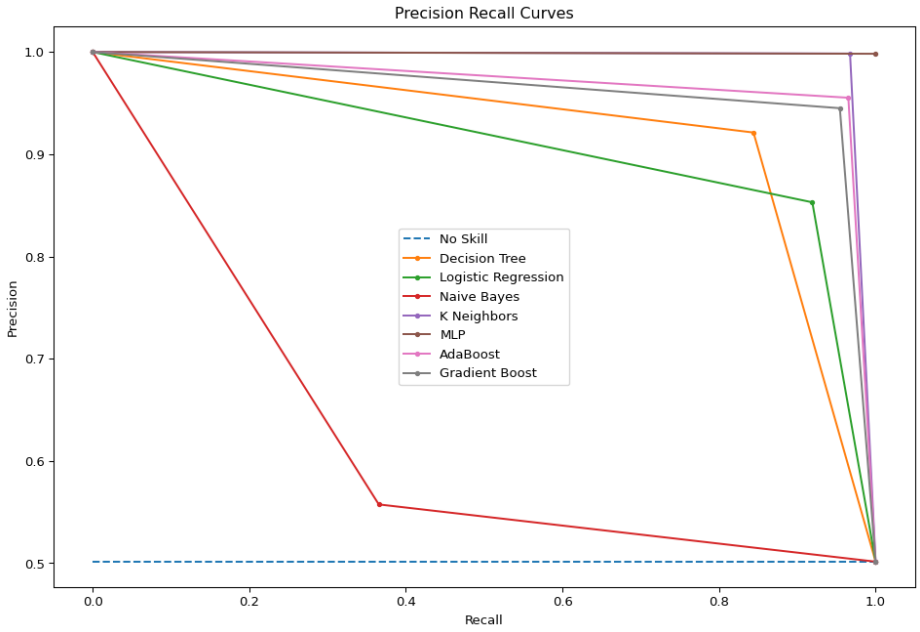
Παρατηρούμε ότι η προβλεπτική ικανότητα του μοντέλου αγγίζει το 94.78%. Το μοντέλο σύμφωνα με την πρώτη τεχνική διάσπασης – διαχωρισμού των δεδομένων και συγκεκριμένα του training και test set εμφανίζει 74.435 σωστά κατανεμημένες περιπτώσεις στο δείγμα και 4.103 λανθασμένα κατανεμημένες από τις 78.538 περιπτώσεις σύμφωνα με το Confusion Matrix, ενώ σύμφωνα με τη δεύτερη τεχνική του k fold - cross validation εμφανίζει 371.672 σωστά κατανεμημένες περιπτώσεις στο δείγμα και 21.018 λανθασμένα κατανεμημένες από τις 392.690 περιπτώσεις. Το Recall (Ανάκληση) ή Ευαισθησία εκφράζεται από το ποσοστό TP Rate και ανέρχεται σε ποσοστό 95.1% που αποτελεί το καλύτερο ποσοστό και από τις δύο τεχνικές. Επίσης το Precision ανέρχεται σε ποσοστό 94.8% και το F1-Score σε 94.8% που αποτελούν τα καλύτερα ποσοστά και από τις δύο τεχνικές διάσπασης.

Παρακάτω προβάλλουμε τις καμπύλες ROC των ταξινομητών που χρησιμοποιήθηκαν.



Εικόνα 5.10 Καμπύλες ROC

Ένα άλλο μέτρο αξιολόγησης ενός ταξινομητή είναι το εμβαδόν κάτω από την καμπύλη ROC (Area Under the Curve - AUC). Παρακάτω προβάλλουμε τις καμπύλες AUC Ακρίβειας-Ευαισθησίας (Precision - Recall) των επτά (7) ταξινομητών.



Εικόνα 5.11 Καμπύλες AUC Precision -Recall

Από τα παραπάνω, προκύπτει ότι ο MLP παρουσιάζει τα καλύτερα αποτελέσματα και ακολουθούν οι αλγόριθμοι KNN, AdaBoost, Gradient Boost, Decision Tree και Logistic Regression. Τη χειρότερη επίδοση παρουσιάζει ο αλγόριθμος Naïve Bayes.

ΚΕΦΑΛΑΙΟ 6: Συγκριτική Αξιολόγηση Αλγορίθμων - Συμπεράσματα – Μελλοντικές Επεκτάσεις

6.1 Συγκριτική Αξιολόγηση - Συμπεράσματα

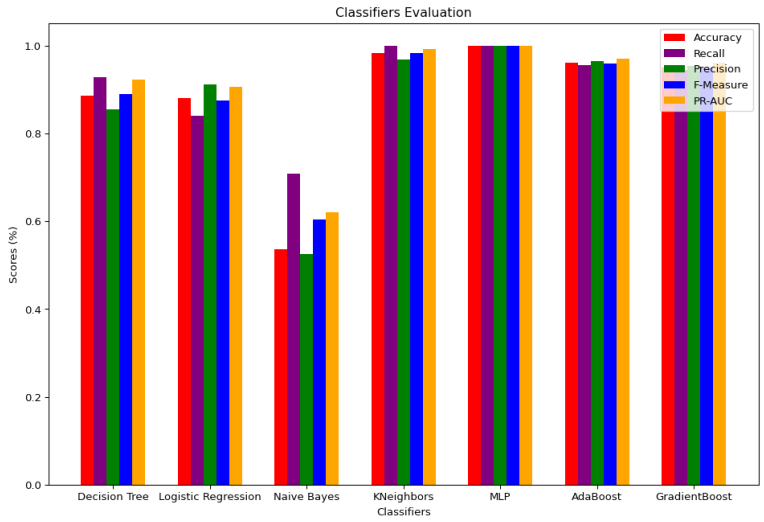
Στην παρούσα διπλωματική εργασία στόχος μας ήταν η ανίχνευση και πρόβλεψη απάτης σε πραγματικά δεδομένα συναλλαγών, προερχόμενα από τον κλάδο του λιανικού εμπορίου. Για τον εντοπισμό της απάτης και μη γνωρίζοντας την κλάση που ανήκουν οι παρατηρήσεις, εστιάσαμε σε συγκεκριμένα μοτίβα αγορών τα οποία θεωρήθηκαν ύποπτα και λαμβάνοντας υπόψη μας τις περιπτώσεις αυτές τις χαρακτηρίσαμε ως απάτη (υποψία απάτης). Στη συνέχεια αφού δημιουργήσαμε την κλάση που ανήκουν οι εγγραφές και κατόπιν κατάλληλης προεπεξεργασίας των δεδομένων καθώς και τεχνικών εξισορρόπησης των κλάσεων, χρησιμοποιήσαμε διάφορους αλγορίθμους μηχανικής και βαθιάς μάθησης με σκοπό τη δημιουργία προβλεπτικών μοντέλων προκειμένου να επιλύσουμε πλέον ένα πρόβλημα δυαδικής ταξινόμησης που αφορά στην πρόβλεψη για το αν μια συναλλαγή αποτελεί προϊόν απάτης ή όχι, εφαρμόζοντας τη μέθοδο της κατηγοριοποίησης. Αξιολογώντας συγκριτικά τα αποτελέσματα από την εκτέλεση των ταξινομητών, παρουσιάζουμε τον παρακάτω πίνακα.

Πίνακας 6.1 Συγκριτικός πίνακας Αξιολόγησης

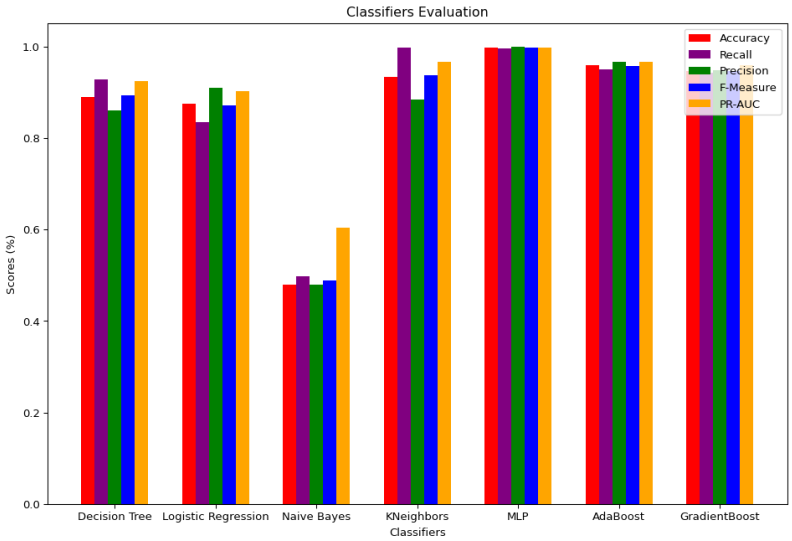
Ταξινομητές	Τρόπος διάσπασης δεδομένων	Accuracy	Recall	Precision	F-Measure	PR AUC
Decision Tree	Train 80%	88.72%	85.8%	92.7%	89.1%	92.2%
	K - Cross Validation=5	88.71%	85.8%	92.8%	89.2%	92.3%
Logistic Regression	Train 80%	87.89%	91.0%	84.1%	87.4%	90.6%
	K - Cross Validation=5	87.50%	90.9%	83.4%	87.0%	90.2%
Naïve Bayes	Train 80%	53.33%	52.4%	70.1%	60.0%	61.7%
	K - Cross Validation=5	48.23%	48.3%	50.8%	49.5%	60.6%
KNN	Train 80%	98.32%	96.9%	**99.9%**	98.3%	99.1%
	K - Cross Validation=5	93.44%	88.6%	99.7%	93.8%	96.6%
MLP	Train 80%	99.60%	99.8%	99.4%	99.6%	99.6%
	K - Cross Validation=5	**99.81%**	**100%**	99.6%	**99.8%**	**99.8%**
AdaBoost	Train 80%	95.92%	96.4%	95.4%	95.9%	96.8%
	K - Cross Validation=5	95.62%	96.3%	94.9%	95.6%	96.7%
Gradient Boost	Train 80%	94.78%	94.7%	94.8%	94.8%	95.9%
	K - Cross Validation=5	94.65%	95.1%	94.2%	94.6%	95.9%

Όπως παρατηρούμε τα καλύτερα αποτελέσματα παρουσιάζει ο αλγόριθμος MLP με πολύ υψηλές επιδόσεις σε όλα τα μέτρα αποτελεσματικότητας. Ο αλγόριθμος KNN παρουσιάζει επίσης πολύ υψηλές επιδόσεις, ιδιαίτερα με την πρώτη τεχνική διαχωρισμού των δεδομένων του Train 80% - Test 20%. Οι ταξινομητές ενδυνάμωσης AdaBoost και Gradient Boost εμφανίζουν ως καλύτερο ποσοστό Accuracy 95.92% και 94.78% αντίστοιχα, αλλά και υψηλά ποσοστά σε Recall, Precision, F-Measure και PR AUC. Πολύ ικανοποιητικά αποτελέσματα επιτυγχάνουν οι αλγόριθμοι Δέντρα Απόφασης και Λογιστική Παλινδρόμηση με καλύτερα ποσοστά Accuracy 88.72% και 87.89% αντίστοιχα, αλλά και πολύ καλά αποτελέσματα στα υπόλοιπα μέτρα αποτελεσματικότητας. Τις χαμηλότερες επιδόσεις εμφάνισε ο αλγόριθμος Naïve Bayes παρόλο που ήταν εξαιρετικά γρήγορος.

Παρακάτω ακολουθούν τα συγκριτικά διαγράμματα της απόδοσης των ταξινομητών σύμφωνα και με τις δύο μεθόδους διαχωρισμού των δεδομένων.



Εικόνα 6.1 Συγκριτικό διάγραμμα απόδοσης ταξινομητών με Train 80% - Test 20%



Εικόνα 6.2 Συγκριτικό διάγραμμα απόδοσης ταξινομητών με Cross Validation k=5

Συμπερασματικά, η απάτη αποτελεί ένα πολύ σημαντικό πρόβλημα που απειλεί τις επιχειρήσεις όλων των κλάδων. Ειδικότερα, στον κλάδο του λιανικού εμπορίου, φαινόμενα απάτης αποτελούν κίνδυνο τόσο για την οικονομική ευημερία των επιχειρήσεων όσο και για τη φήμη τους, καθώς κλονίζονται οι σχέσεις εμπιστοσύνης με τους πελάτες τους. Η χρήση μεθοδολογιών και μοντέλων πρόβλεψης απάτης είναι μείζονος σημασίας για τις επιχειρήσεις, ιδιαίτερα στη σημερινή εποχή με την έκρηξη του ηλεκτρονικού εμπορίου και την ευρεία εξάπλωση των διαδικτυακών συναλλαγών. Οι τεχνικές της Μηχανικής Μάθησης αποδεικνύονται ιδιαίτερα αποτελεσματικές στην ανίχνευση απάτης και την πρόβλεψη δόλιων συναλλαγών, διασφαλίζοντας και ενδυναμώνοντας με αυτόν τον τρόπο την αξιοπιστία, τη φερεγγυότητα και το κύρος των επιχειρήσεων, καλλιεργώντας ταυτόχρονα ένα αίσθημα ασφάλειας και σιγουριάς μεταξύ πελάτη και επιχείρησης που οδηγεί στη βελτίωση της εμπειρίας του πελάτη.

6.2 Μελλοντικές Επεκτάσεις

Σε μια μελλοντική εργασία, ένα σύστημα ανίχνευσης θα μπορούσε ενδεχομένως να εντοπίζει αλλαγές στα μοτίβα συναλλαγών που σχετίζονται με την αλλαγή συμπεριφοράς των καταναλωτών συγκρίνοντας για παράδειγμα τη συμπεριφορά ενός πελάτη με μια αναμενόμενη συμπεριφορά, κατόπιν κατάλληλης ομαδοποίησης των καταναλωτών. Επιπλέον, για τον εντοπισμό της απάτης θα μπορούσε να χρησιμοποιηθεί η τεχνική Ανίχνευσης Ανωμαλιών (Anomaly Detection) με βάση το προφίλ των καταναλωτών και το μοντέλο της αγοραστικής τους συμπεριφοράς, ανακαλύπτοντας συμπεριφορές που αποκλίνουν από το φυσιολογικό, δηλαδή ασυνήθιστες συμπεριφορές που αποτελούν ένδειξη απάτης. Ενδιαφέρον επίσης θα παρουσίαζε η ανάπτυξη νέων πιο αποτελεσματικών τεχνικών εύρεσης απάτης σε δεδομένα συναλλαγών όπως για παράδειγμα μια περίπτωση προσέγγισης με τη χρήση Συνελικτικών Νευρωνικών Δικτύων (Convolutional Neural Networks - CNN). Στόχος πάντα της κάθε τεχνικής είναι η αποτελεσματικότητά της ως προς τον εντοπισμό και τον περιορισμό φαινομένων απάτης. Γενικότερα, καθώς οι στρατηγικές απάτης συνεχώς ανανεώνονται, θα πρέπει οι τεχνικές και τα συστήματα ανίχνευσης να ενημερώνονται και να εξελίσσονται διαρκώς.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] The Institute of Internal Auditors Global - IIA Διεθνή Πρότυπα, [Ηλεκτρονικό]. Available: https://global.theiia.org/translations/PublicDocuments/IPPF-Standards-2017-Greek.pdf.

[2] «Οι top 7 Ευρωπαϊκές Εταιρείες ανίχνευσης απάτης για το 2020,» secnews, 11 02 2020. [Ηλεκτρονικό]. Available: https://www.secnews.gr/211110/oi-top-7-evropaikes-etaireies-anichnefsis-apatis-gia-to-2020/.

[3] Εκθεση Χρηματοπιστωτικής Σταθερότητας Τράπεζα της Ελλάδος, 01 2021. [Ηλεκτρονικό]. Available: https://www.bankofgreece.gr/Publications/ΕΚΘΕΣΗ_ΧΡΗΜΑΤΟΠΙΣΤΩΤΙΚΗΣ_ΣΤΑΘΕΡΟΤΗΤΑΣ_ΙΑΝΟΥΑΡΙΟΣ_2021.pdf.

[4] Συνήγορος του Καταναλωτή Ανεξάρτητη Αρχή, Ετήσια έκθεση 2020. [Ηλεκτρονικό]. Available: http://www.synigoroskatanaloti.gr/docs/StK-Annual-Report-2020-Summary.pdf.

[5] McKinsey, «Seven imperatives for rethinking retail,» 03 2021. [Ηλεκτρονικό]. Available: https://www.mckinsey.com/~/media/McKinsey/Industries/Retail/Our%20Insights/retail%20speaks%20seven%20imperatives%20for%20the%20industry/retail-speaks-full-report.pdf.

[6] «E-commerce statistics for individuals,» 01 2021. [Ηλεκτρονικό]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=E-commerce_statistics_for_individuals.

[7] Συνήγορος του Καταναλωτή Ανεξάρτητη Αρχή, [Ηλεκτρονικό]. Available: http://www.synigoroskatanaloti.gr/docs/info/info-Hlektroniko-Egklima.pdf.

[8] www.alpha.gr Εβδομαδιαίο Δελτίο Οικονομικών Εξελίξεων, 30 03 2021. [Ηλεκτρονικό]. Available: https://www.alpha.gr/-/media/alphagr/files/group/agores/weekly-economic-report/2021/weekly_30032021.pdf.

[9] Gary Fowler, «Using AI To Increase Retail Sales,» 2019. [Ηλεκτρονικό]. Available: https://www.forbes.com/sites/forbesbusinessdevelopmentcouncil/2019/07/30/using-ai-to-increase-retail-sales/#71c99b8652a9.

[10] Statista, 02 2020. [Ηλεκτρονικό]. Available: https://www.statista.com/statistics/1197958/ai-use-cases-consumer-goods-retail-global/.

[11] Statista, 03 2020. [Ηλεκτρονικό]. Available: https://www.statista.com/statistics/607716/worldwide-artificial-intelligence-market-revenues/.

[12] «Ευρωπαϊκή στρατηγική για τα δεδομένα,» European Commission, [Ηλεκτρονικό]. Available: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_el.

[13] J. Josh, «What ‘Data Never Sleeps 9.0’ Proves About the Pandemic,» 4 October 2021. [Ηλεκτρονικό]. Available: https://www.domo.com/blog/what-data-never-sleeps-9-0-proves-about-the-pandemic/.

[14] D. Laney, 3D Data Management: Controlling Data Volume, Velocity, and Variety, META Group, 2001.

[15] R. Buyya, R. N. Calheiros και A. V. Dastjerdi, Big data : principles and paradigms, Cambridge Morgan Kaufmann, 2016.

[16] Indian J Orthop, «Significant Applications of Big Data in COVID-19 Pandemic,» τόμ. 54(4): 526–528, αρ. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7204193/, Jul 2020.

[17] C. Henry, «How Big data is changing the insurance industry,» March 2020. [Ηλεκτρονικό]. Available: https://chrishtopher-henry-38679.medium.com/how-big-data-is-changing-the-insurance-industry-293bb243a820.

[18] «Τι είναι Επιχειρηματική Ευφυΐα - Business Intelligence (BI), softcon,» [Ηλεκτρονικό]. Available: http://softcon.gr/10-arthrografua/62-bi-what-is.

[19] A. Ferrari, Business Intelligence Systems, Uncertainty in Decision-Making and Effectiveness of Organizational Coordination, 2011.

[20] Ε. Κύρκος, Επιχειρηματική Ευφυΐα & Εξόρυξη Δεδομένων, Αθήνα: ΣΕΑΒ,, 2015.

[21] D. T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, New York: Wiley, 2004.

[22] Β. Σ. Βερύκιος, Β. Καγκλής και Ή. Κ. Σταυρόπουλος, Η επιστήμη των δεδομένων μέσα από τη γλώσσα R κεφ.1, Εισαγωγή στην Εξόρυξη Δεδομένων, Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015.

[23] T. Mitchell, Machine Learning, McGraw Hill International Edition, 1997.

[24] Α. Γεωργούλη, Τεχνητή νοημοσύνη κεφ.4, Μηχανική Μάθηση, Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015.

[25] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2012.

[26] Κύρκος Γ. Ε., Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων κεφ. 6, Εξόρυξη Γνώσης Από Δεδομένα, Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών ΒΙβλιοθηκών, 2015.

[27] Oded Maimon and Lior Rokach, Data Mining and Knowledge Discovery Handbook, Chapter 9, 2nd ed., 2010.

[28] Quinlan J.R., C4.5: Programs for Machine Learning, Los Altos: Morgan Kaufmann, 1993.

[29] J. Brownlee, «Logistic Regression for Machine Learning,» 15 August 2020. [Ηλεκτρονικό]. Available: https://machinelearningmastery.com/logistic-regression-for-machine-learning/.

[30] R. Saxena, «Dataaspirant,» 06 February 2017. [Ηλεκτρονικό]. Available: https://dataaspirant.com/naive-bayes-classifier-machine-learning/.

[31] A. Joby, «What Is K-Nearest Neighbor? An ML Algorithm to Classify Data, learn.g2,» 19 July 2021. [Ηλεκτρονικό]. Available: https://learn.g2.com/k-nearest-neighbor.

[32] Pang-Ning Tan, Michael Steinbach και Vipin Kumar, Introduction to Data Mining, Pearson, 2006.

[33] Judith Hurwitz και Daniel Kirsch, Machine Learning for Dummies, John Wiley & Sons, Inc., 2018.

[34] Marius-Constantin Popescu, Valentina E. Balas, Liliana Perescu-Popescu και Nikos Mastorakis, Multilayer Perceptron and Neural Networks, WSEAS Trans. Cir. and Sys. 8, 7, 2009.

[35] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, Third ed., Pearson, 2010.

[36] L. Rokach, «Ensemble-based classifiers, Artificial Intelligence Review,» Springer, 2010.

[37] J. Brownlee, «A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning,» September 9, 2016.

[38] K. Dhandhania, «How to understand Gradient Descent, the most popular ML algorithm,» 18 June 2018. [Ηλεκτρονικό]. Available: https://freecodecamp.org/news/understanding-gradient-descent-the-most-popular-ml-algorithm-a66c0d97307f/.

[39] Thanh Cong Tran και Tran Khanh Dang, «Machine Learning for Prediction of Imbalanced Data: Credit Fraud Detection,» 2021.

[40] N. Cochrane et al., «Pattern Analysis for Transaction Fraud Detection,» 2021.

[41] D. Prusti, A. Kumar, I. S. Purusottam και S. Kumar Rath, «A design methodology for web-based services to detect fraudulent transactions in credit card. In 14th Innovations in Software Engineering Conference (formerly known as India Software Engineering Conference) (ISEC 2021),» 2021.

[42] I. Psychoula, A. Gutmann, P. Mainali, S. H. Lee, P. Dunphy και F. Petitcolas, «Explainable Machine Learning for Fraud Detection,» Oct 2021.

[43] T. T. Nguyen, H. Tahir, M. Abdelrazek και A. Babar, «Deep Learning Methods for Credit Card Fraud Detection,» 2020.

[44] H. Manek, N. Kataria, S. Jain και C. Bhole, «Various Methods for Fraud Transaction Detection in Credit Cards,» November 2018.

[45] GoldenGatePro, «Οι 3 σημαντικότερες Python βιβλιοθήκες για την ανάλυση δεδομένων,» 01 Feb 2019. [Ηλεκτρονικό]. Available: https://medium.com/@GoldenGatePro/python-libraries-data-science-bbc98c1bb148.