2026-01

þÿ š ± Ä ± ½ ± » É Ä ¹ º - Â Ä ¬ Ã µ ¹ Â Ã Ä ¹ Â

þÿ · » µ º Ä Á ¿ ½ ¹ º - Â ± ³ ¿ Á - Â ¼ - Ã É ± ¼

þÿ ´ µ ´ ¿ ¼ - ½ É ½ º ± ¹ ¼ · Ç ± ½ ¹ º ® Â ¼ –

þÿ ” µ ¼ ¯ Á · , £ ¿ Æ ¯ ±

þÿ œ µ Ä ± À Ä Å Ç ¹ ± º Ì Á Ì ³ Á ± ¼ ¼ ± Ã Ä · ½ ‘ ½ ¬ » Å Ã · ” µ ´ ¿ ¼ - ½ É ½ º ± ¹ § Á · ¼ ± Ä ¿ ¿ ¹ º ¿ ½ ¿
þÿ £ Ç ¿ » ® Ÿ ¹ º ¿ ½ ¿ ¼ ¹ º Î ½ • À ¹ Ã Ä · ¼ Î ½ º ± ¹ ” ¹ ¿ ¯ º · Ã · Â , ± ½ µ À ¹ Ã Ä ® ¼ ¹ ¿ • µ ¬ À ¿ » ¹ Â

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

Πανεπιστήμιο Νεάπολις Πάφος.

«ΚΑΤΑΝΑΛΩΤΙΚΕΣ ΤΑΣΕΙΣ ΣΤΙΣ ΗΛΕΚΤΡΟΝΙΚΕΣ ΑΓΟΡΕΣ ΜΕΣΩ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ»

Σοφία Δεμίρη

Ιανουάριος, 2026

Πανεπιστήμιο Νεάπολις Πάφου

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

Διπλωματική Εργασία η οποία υποβλήθηκε προς απόκτηση εξ αποστάσεως μεταπτυχιακού τίτλου σπουδών στην ανάλυση δεδομένων και χρηματοοικονομική τεχνολογία στο Πανεπιστήμιο Νεάπολις Πάφος.

«ΚΑΤΑΝΑΛΩΤΙΚΕΣ ΤΑΣΕΙΣ ΣΤΙΣ ΗΛΕΚΤΡΟΝΙΚΕΣ ΑΓΟΡΕΣ ΜΕΣΩ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ»

Σοφία Δεμίρη

Ιανουάριος, 2026

Πνευματικά δικαιώματα

Η ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ

Η Σοφία Δεμίρη, γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «καταναλωτικές τάσεις στις ηλεκτρονικές αγορές μέσω ανάλυσης δεδομένων και μηχανικής μάθησης», αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Η Δηλούσα

Σοφία Δεμίρη

Πίνακας περιεχομένων

Πίνακας Σχημάτων

Πίνακας Διαγραμμάτων

Ονοματεπώνυμο Φοιτητή: Σοφία Δεμίρη

Τίτλος Διπλωματικής Εργασίας: ΚΑΤΑΝΑΛΩΤΙΚΕΣ ΤΑΣΕΙΣ ΣΤΙΣ ΗΛΕΚΤΡΟΝΙΚΕΣ ΑΓΟΡΕΣ ΜΕΣΩ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Η παρούσα Διπλωματική Εργασία εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση εξ αποστάσεως μεταπτυχιακού τίτλου στο Πανεπιστήμιο Νεάπολις και εγκρίθηκε στις / /2026 από τα μέλη της Εξεταστικής Επιτροπής.

Εξεταστική Επιτροπή:

Πρώτος επιβλέπων: Κωνσταντίνος Παναγιωτάκης, Καθηγητής, Τμήμα ΔΕΤ, ΕΛΜΕΠΑ

Μέλος Εξεταστικής Επιτροπής: Χρήστος Λεμονάκης, Αναπληρωτής Καθηγητής, Τμήμα ΔΕΤ, ΕΛΜΕΠΑ

Μέλος Εξεταστικής Επιτροπής: Γεώργιος Μαστοράκης, Αναπληρωτής Καθηγητής, Τμήμα ΔΕΤ, ΕΛΜΕΠΑ

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα καθηγητή μου Κωνσταντίνο Παναγιωτάκη, για την πολύτιμη βοήθεια του και την καθοδήγηση του καθ' όλη την διάρκεια της διπλωματικής μου εργασίας. Η συμβολή του υπήρξε ιδιαίτερα σημαντική για την υλοποίηση της εργασίας παρέχοντας ουσιαστικές παρατηρήσεις, αξιόλογα εργαλεία και προτάσεις που αποδείχθηκαν καθοριστικές για την ολοκλήρωση της.

Τον ευχαριστώ για την εμπιστοσύνη και την ενθάρρυνση που μου έδειξε, καθώς συνέλαβαν ουσιαστικά στην πρόοδο μου και στην επιτυχή ολοκλήρωση της παρούσας διπλωματικής εργασίας.

Περίληψη

Η παρούσα διπλωματική εργασία εστιάζει στη μελέτη των καταναλωτικών τάσεων στις ηλεκτρονικές αγορές, αξιοποιώντας τεχνικές ανάλυσης δεδομένων και μεθόδους μηχανικής μάθησης, με έμφαση στην ανάλυση συσχετίσεων μεταξύ δημογραφικών και συμπεριφορικών χαρακτηριστικών των καταναλωτών. Σκοπός της έρευνας είναι η διερεύνηση και η κατανόηση της αγοραστικής συμπεριφοράς στο ψηφιακό περιβάλλον, καθώς και η ανάδειξη προτύπων που δύναται να συμβάλλουν στη βελτίωση της αποδοτικότητας και της στρατηγικής προσέγγισης των σύγχρονων επιχειρήσεων που δραστηριοποιούνται στο ηλεκτρονικό εμπόριο.

Η έρευνα βασίζεται σε ένα πραγματικό σύνολο δεδομένων σχετικά ηλεκτρονικού εμπορίου, το οποίο διατίθεται δημόσια μέσω της πλατφόρμας kaggle και αφορά την πλατφόρμα Olist. Το συγκεκριμένο σύνολο δεδομένων περιλαμβάνει πραγματικές συναλλαγές, χαρακτηριστικά παραγγελιών καθώς και πληροφορίες σχετικά με τη διαδικασία παράδοσης. Όλα τα παραπάνω, αξιοποιούνται για την ανάλυση καταναλωτικών τάσεων και την λήψη επιχειρηματικών αποφάσεων. Η μεθοδολογική προσέγγιση επικεντρώνεται στη διερευνητική ανάλυση δεδομένων και στον υπολογισμό συντελεστών συσχέτισης με στόχο την αποτύπωση των σχέσεων μεταξύ δημογραφικών χαρακτηριστικών, καταναλωτικών επιλογών και ικανοποίησης των καταναλωτών.

Τα αποτελέσματα της ανάλυσης αποδεικνύουν την ύπαρξη σημαντικών συσχετίσεων μεταξύ συγκεκριμένων χαρακτηριστικών και αγοραστικών προτύπων, καθώς και τη δυνατότητα διάκρισης ομάδων καταναλωτών με παρόμοια χαρακτηριστικά. Συνολικά, η εργασία αναδεικνύει τη χρησιμότητα της ανάλυσης δεδομένων και της μηχανικής μάθησης ως εργαλεία ερμηνείας της καταναλωτικής συμπεριφοράς στο πλαίσιο των ηλεκτρονικών αγορών, ενισχύοντας την κατανόηση των σύγχρονων καταναλωτικών τάσεων.

Λέξεις κλειδιά

Μηχανική μάθηση, Νευρωνικά δίκτυα, ηλεκτρονικές αγορές, ταξινόμηση, ανάλυση συσχετίσεων, αλγόριθμος Apriori

Abstract

This thesis focuses on the study of consumer trends in online shopping by employing data analysis techniques and machine learning methods, with particular emphasis on examining correlations between demographic and behavioral characteristics of consumers. The aim of the research is to investigate and understand purchasing behavior in the digital environment, as well as to identify patterns that may contribute to improving the efficiency and strategic approach of contemporary enterprises operating in the field of electronic commerce.

The research is based on a synthetic dataset, which includes variables such as age, income, gender, marital status, product category, customer satisfaction level, and payment method. The methodological approach centers on exploratory data analysis and the calculation of correlation measures, in order to capture the relationships between demographic characteristics, consumer choices, and levels of customer satisfaction.

The results of the analysis demonstrate the existence of significant correlations between specific characteristics and purchasing patterns, as well as the potential to distinguish groups of consumers with similar profiles. Overall, the study highlights the usefulness of data analysis and machine learning as tools for interpreting consumer behavior in the context of online shopping, thereby enhancing the understanding of contemporary consumer trends.

Keywords

machine learning, neural networks, online market, classification, correlation analysis, Apriori algorithm

Αφιερώνεται

«Στην οικογένειά μου, τους φίλους μου και τον σύντροφο μου για την αγάπη, τη στήριξη και την υπομονή τους σε όλη την διάρκεια αυτής της απαιτητικής πορείας, καθώς στάθηκαν δίπλα μου σε κάθε δυσκολία, προσφέροντας μου στήριξη, υπομονή και ενθάρρυνση. Επίσης, την αφιερώνω στον Πίκο, τον τετράποδο φίλο μου, που με τη σιωπηλή του παρουσία και την αφοσίωση του υπήρξε καθημερινή πηγή δύναμης και συντροφιάς μέχρι την ολοκλήρωση της παρούσας εργασίας, υπενθυμίζοντας μου καθημερινά την σημασία της επιμονής και της ισορροπίας.»

Εισαγωγή

Οι ηλεκτρονικές αγορές αποτελούν αναπόσπαστο μέρος της σύγχρονης οικονομικής δραστηριότητας, καθώς έχουν επηρεαστεί σημαντικά από την ανάπτυξη του ηλεκτρονικού εμπορίου και των ψηφιακών τεχνολογιών, διότι συνέλαβαν στη μεταβολή του τρόπου με τον οποίο οι καταναλωτές αξιολογούν και αποκτούν προϊόντα και υπηρεσίες. Η μετάβαση από τα παραδοσιακά κανάλια πώλησης φυσικής παρουσίας σε ψηφιακές πλατφόρμες έχει δημιουργήσει νέα δεδομένα στις σχέσεις μεταξύ επιχειρήσεων και καταναλωτών, ενισχύοντας τον ανταγωνισμό και αυξάνοντας την δυνατότητα εύρεσης πληροφοριών σχετικά με τιμές, αξιολογήσεις και χαρακτηριστικά προϊόντων (Laudon & Traver, 2024; Parker et al., 2016). Επιπλέον, οι ηλεκτρονικές αγορές έχουν περιορίσει τα γεωγραφικά προβλήματα και έχουν διευκολύνει την πρόσβαση των καταναλωτών σε ένα ευρύ φάσμα επιλογών, συμβάλλοντας στη διαμόρφωση ενός δυναμικού και συνεχώς εξελισσόμενου ψηφιακού περιβάλλοντος ((Turban et al., 2017) (*Digital Economy Report 2021 (Overview)*, 2021). Όπως παρατηρείται στο Σχήμα 1, τα έσοδα του ηλεκτρονικού εμπορίου παρουσιάζουν σταθερή αύξηση κατά την εξεταζόμενη περίοδο.

Σχήμα 1: Εξέλιξη των ετήσιων εσόδων του ηλεκτρονικού εμπορίου στις Ηνωμένες Πολιτείες για την περίοδο 2014–2024 (σε δισεκατομμύρια δολάρια). Πηγή: U.S. Census Bureau.

Η εκρηκτική ανάπτυξη των ηλεκτρονικών αγορών έχει οδηγήσει στη δημιουργία μεγάλου όγκου δεδομένων, τα οποία αποτυπώνουν σε πραγματικό χρόνο, την αγοραστική συμπεριφορά και τις αλληλεπιδράσεις των καταναλωτών. Η ανάλυση αυτών των δεδομένων αποτελεί αντικείμενο αυξανόμενου ερευνητικού ενδιαφέροντος, καθώς συμβάλλει στην διευρευνηση των καταναλωτικών τάσεων και την κατανόηση των παραγόντων που επηρεάζουν τις αγοραστικές αποφάσεις (Chen et al., 2012; Gandomi & Haider, 2015). Η μελέτη των συσχετίσεων μεταξύ δημογραφικών χαρακτηριστικών και καταναλωτικών επιλογών θεωρείται πολύ σημαντική, καθώς λειτουργεί ως βασικό εργαλείο για την ερμηνεία της αγοραστικής συμπεριφοράς και για την εξαγωγή συμπερασμάτων σχετικά με τη λειτουργία του ηλεκτρονικού εμπορίου (Wedel & Kannan, 2016).

Στο πλαίσιο αυτό, οι μέθοδοι μηχανικής μάθηση και η ανάλυση δεδομένων καθίστανται καθοριστικής σημασίας για τη μελέτη της καταναλωτικής συμπεριφοράς. Μέσω τεχνικών διερευνητικής ανάλυσης, ταξινόμησης και ανάλυσης συσχετίσεων, διαδραματίζεται η αποτύπωση σύνθετων σχέσεων μεταξύ διαφορετικών χαρακτηριστικών των καταναλωτών, καθώς και η διερεύνηση επαναλαμβανόμενων μοτίβων αγοραστικής συμπεριφοράς ((Han et al., 2011; Jordan & Mitchell, 2015). Οι προσεγγίσεις αυτές, προσφέρουν ένα αξιόπιστο υπόβαθρο για την ερμηνεία της δυναμικής του ηλεκτρονικού εμπορίου.

Η παρούσα διπλωματική εργασία δομείται σε πέντε βασικά κεφάλαια. Αρχικά παρουσιάζεται το θεωρητικό υπόβαθρο και στη συνέχεια αναλύεται η μεθοδολογία και τα δεδομένα της μελέτης. Τέλος, παρουσιάζονται τα βασικά ευρήματα και τα συμπεράσματα της εργασίας.

ΚΕΦΑΛΑΙΟ 1: ΣΤΟΧΟΙ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

1.1 Στόχοι

Η παρούσα εργασία έχει ως στόχο την κατανόηση της καταναλωτικής συμπεριφοράς στις ηλεκτρονικές αγορές μέσω της ανάλυσης δεδομένων και της αξιοποίησης μεθόδων μηχανικής μάθησης. Συγκεκριμένα, η έρευνα επικεντρώνεται στην μελέτη των σχέσεων μεταξύ δημογραφικών χαρακτηριστικών και αγοραστικών επιλογών, με σκοπό να αναγνωρίσει τυχόν συσχετίσεις και επαναλαμβανόμενα πρότυπα συμπεριφοράς στο ψηφιακό περιβάλλον (Lemon & Verhoef, 2016).

Ειδικότερα, η εργασία στοχεύει στην ανάλυση δεδομένων που εστιάζουν σε βασικά χαρακτηριστικά των καταναλωτών, όπως δημογραφικά στοιχεία, επίπεδο ικανοποίησης και καταναλωτικές επιλογές, καθώς και στη διερεύνηση του τρόπου με τον οποίο τα χαρακτηριστικά αυτά συνδέονται με τη συμπεριφορά των καταναλωτών στις ηλεκτρονικές αγορές. Μέσω της συνεχής μελέτης των δεδομένων, επιδιώκεται η εξαγωγή συμπερασμάτων για την κατανόηση των μηχανισμών που επηρεάζουν τις καταναλωτικές αποφάσεις στο πλαίσιο του ηλεκτρονικού εμπορίου.

1.2 Αποτελέσματα

Η παρούσα μελέτη ανέδειξε σαφείς συσχετίσεις μεταξύ δημογραφικών χαρακτηριστικών και αγοραστικών προτύπων στις ηλεκτρονικές αγορές. Η ανάλυση των δεδομένων αποκάλυψε ότι συγκεκριμένα προσωπικά και συμπεριφορικά χαρακτηριστικά σχετίζονται με διαφοροποιήσεις στην αγοραστική συμπεριφορά των καταναλωτών στις ηλεκτρονικές αγορές, δίνοντας την δυνατότητα να ερμηνευθούν πρότυπα αγοραστικής συμπεριφοράς στο ψηφιακό περιβάλλον (Kumar et al., 2016).

Παράλληλα, τα αποτελέσματα αποδεικνύουν ότι η αξιοποίηση τεχνικών ανάλυσης δεδομένων και μεθόδων μηχανικής μάθησης, διευκόλυνε την κατανόηση των ηλεκτρονικών αγορών. Ο εντοπισμός κοινών χαρακτηριστικών μεταξύ καταναλωτών βοήθησε στην καλύτερη κατανόηση των σύγχρονων καταναλωτικών τάσεων στις ηλεκτρονικές αγορές (Chen et al., 2011).

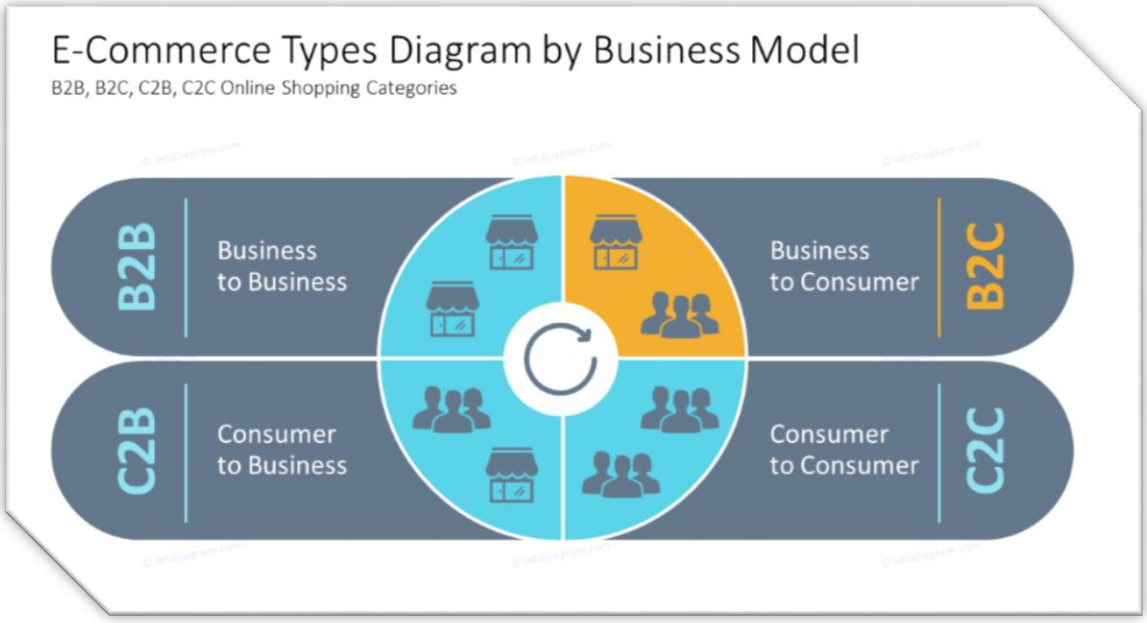
ΚΕΦΑΛΑΙΟ 2: ΗΛΕΚΤΡΟΝΙΚΟ ΕΜΠΟΡΙΟ & ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

2.1 Ηλεκτρονικό Εμπόριο

Το Ηλεκτρονικό εμπόριο (e-commerce) έχει αλλάξει δραστικά τις σχέσεις αγοραστή και πωλητή και πλέον αποτελεί ένα σύγχρονο τρόπο αγοράς αγαθών μέσω διαδικτύου (Laudon & Traver, 2024). Η τεχνολογική πρόοδος έχει βοηθήσει τις αγορές ώστε να πραγματοποιούνται με μεγαλύτερη ταχύτητα, πιο εύκολα και πολλές φορές χωρίς φυσική επαφή ((Turban et al., 2017; Vladimir, 1996); (Vladimir, 1996).

Η ραγδαία ανάπτυξη του ηλεκτρονικού εμπόριου έγινε περισσότερο εμφανής κατά τη διάρκεια της πανδημίας του covid-19, όπου τα περιοριστικά μέτρα για μείωση μετακινήσεων και κοινωνικής αποστασιοποίησης, ώθησαν τους καταναλωτές να στραφούν σε διαδικτυακές αγορές (OECD, 2020; Sheth, 2020). Εκ τοτε, το ηλεκτρονικό εμπόριο έχει εδραιωθεί στις βασικές επιλογές των καταναλωτών, παρέχοντας τους ευκολία, ταχύτητα και ποικιλία επιλογών σε προϊόντα και υπηρεσίες με το πάτημα ενός κουμπιού (*Global Retail E-Commerce Sales 2022-2028| Statista*, n.d.).

Οι κύριες μορφές ηλεκτρονικού εμπορίου αποτελούνται από συναλλαγές μεταξύ επιχειρήσεων και καταναλωτών (B2C), μεταξύ επιχειρήσεων (B2B) καθώς επίσης και μεταξύ καταναλωτών (C2C) και καταναλωτών προς επιχειρήσεις (C2B). Η ανάπτυξη αυτών των τριών μορφών, είναι ο λόγος δημιουργίας ηλεκτρονικών αγορών (marketplace) οι οποίοι εκπροσωπούν το ψηφιακό εμπόριο. Οι βασικές μορφές ηλεκτρονικού εμπορίου παρουσιάζονται συνοπτικά στο Σχήμα 2 (Rochet & Tirole, 2006).



Σχήμα 2: Βασικές μορφές ηλεκτρονικού εμπορίου βάσει του επιχειρηματικού μοντέλου συναλλαγών (B2B, B2C, C2B, C2C). Πηγή: InfoDiagram

2.2 Ηλεκτρονικές αγορές

Ο όρος ηλεκτρονικών αγορών πραγματεύεται την διαδικασία στην οποία οι εταιρείες προσφέρουν τα προϊόντα και τις υπηρεσίες τους μέσω μιας κοινής διαδικτυακής πλατφόρμας (Chaffey et al., 2019). Σε σχέση με τα παραδοσιακά καταστήματα, όπου οι συναλλαγές πραγματοοποιούνται απευθείας μεταξύ πελατών και μιας επιχείρησης, οι ηλεκτρονικές αγορές λειτουργούν ως ο ενδιάμεσος παράγοντας που φέρνει σε επαφή τα καταστήματα με τους καταναλωτές (Hagiu & Wright, 2015). Η αγοραστική συμπεριφορά των καταναλωτών αποτυπώνεται ως μια διαδοχική διαδικασία που περιλαμβάνει διακριτά στάδια, ξεκινώντας από την αναγνώριση της ανάγκης, συνεχίζοντας με την αναζήτηση πληροφοριών και την αξιολόγηση εναλλακτικών και καταλήγοντας στη μετα-αγοραστική αξιολόγηση, όπως παρουσιάζεται στο Σχήμα 3 (Edelman & Singer, n.d.).

Σχήμα 3: Στάδια της καταναλωτικής απόφασης στο ψηφιακό περιβάλλον.
Πηγή: Προσαρμογή από τη σχετική βιβλιογραφία

Μια γνωστή ελληνική πλατφόρμα ηλεκτρονικών αγορών είναι το Skroutz, το οποίο θεωρείται κόμβος αγορών που παρέχει εκκατομύρια προϊόντα από χιλιάδες ελληνικές εταιρείες. Οι καταναλωτές έχουν την δυνατότητα να συγκρίνουν τιμές και να διαβάζουν σχόλια και αξιολογήσεις άλλων καταναλωτών πριν προχωρήσουν σε αγορά (*October 2003 THE DIGITIZATION OF WORD-OF-MOUTH*, n.d.). Αντίστοιχα και η ηλεκτρονική κινέζικη πλατφόρμα, Alibaba, συνδέει εκκατομύρια πωλητές των οποίων οι πωλήσεις επηρεάζονται από την φήμη και τις αξιολογήσεις των αγοραστών (Li et al., 2019).

Οι πλατφόρμες ηλεκτρονικών αγορών δημιουργούν πηγές δεδομένων και καταγράφουν τα στάδια της αγοραστικής συμπεριφοράς, από την αναζήτηση έως την αγορά των προϊόντων καθώς και τη μεταγενέστερη αξιολόγηση από τον καταναλωτή. Τέτοιου είδους δεδομένα αξιοποιούνται σε προγνωστικά μοντέλα και στοχευμένες προωθητικές ενέργειες, με στόχο την ενίσχυση της αποδοτικότητας των επιχειρήσεων (Kaur & Kang, 2016).

Παράλληλα, οι ηλεκτρονικές αγορές αποτελούν βασικό πυλώνα για την διαμόρφωση καταναλωτικών τάσεων, καθώς επιτρέπουν την ανάλυση μεγάλου όγκου δεδομένων σε πραγματικό χρόνο. Με τον τρόπο αυτό, καθίσταται δυνατή η ανάπτυξη διάφορων μοντέλων μηχανικής μάθησης που σχετίζονται με την αγοραστική συμπεριφορά, τη δημιουργία συστημάτων προτάσεων και την κατηγοροιοποίηση πελατών βάσει των προσωπικών τους προτιμήσεων (Kolodin et al., 2020).

Στο πλαίσιο αυτό, η μηχανική μάθηση αποτελεί βασικό εργαλείο για την ανάλυση και την αξιοποίηση των δεδομένων ηλεκτρονικού εμπορίου, γεγονός που εξετάζεται αναλυτικά στο επόμενο κεφάλαιο.

2.3 Συμπεριφορά Καταναλωτή στο Ψηφιακό Περιβάλλον

Η ανάπτυξη του ηλεκτρονικού εμπορίου έχει επηρεάσει σημαντικά τη συμπεριφορά των καταναλωτών στο ψηφιακό περιβάλλον. Πλέον οι ψηφιακές πληροφορίες, οι εμπειρίες και οι αλληλεπιδράσεις διαδραματίζουν καθοριστικό ρόλο στη διαδικασία λήψης αγοραστικών αποφάσεων (Kannan & Li, 2017; Lemon & Verhoef, 2016). Μερικοί από τους κύριους παράγοντες που επηρεάζουν την αγοραστική συμπεριφορά σε σημαντικό βαθμό, είναι η κοινωνική επιρροή, οι διαδικτυακές κριτικές και η αυξημένη προσβασιμότητα του καταναλωτή σε πλήθος πληροφοριών (Chen et al., 2012; Gefen & Pavlou, 2011).

Τα βασικά χαρακτηριστικά που αντιπροσωπεύουν τους ψηφιακούς καταναλωτές είναι η υψηλή ενημερότητα και η προσβασιμότητα τους σε πλήθος πληροφοριών για σύγκριση προϊόντων σε πραγματικό χρόνο. Μέσω ψηφιακών πλατφορμών, όπως το Skroutz, καθώς και των μέσων κοινωνικής δικτύωσης, οι χρήστες – καταναλωτές δημιουργούν προσδοκίες βασισμένες στις εμπειρίες άλλων χρηστών και σε προτάσεις προϊόντων που δημιουργούνται από αλγορίθμους συστάσεων (Filieri, 2015). Οι διαδικτυακές αξιολογήσεις θεωρούνται αξιόπιστες, καθώς ενισχύουν την εμπιστοσύνη των καταναλωτών, δεδομένου ότι επηρεάζουν την αντίληψη της ποιότητας και μειώνουν την αντιλαμβανόμενη αβεβαιότητα (Mudambi & Schuff, 2010; Rose et al., 2012).

Επιπλέον, ένα ακόμη σημαντικό στοιχείο της ψηφιακής καταναλωτικής συμπεριφοράς είναι η εξατομίκευση (personalization). Πρόκειται για συστήματα μηχανικής μάθησης τα οποία εστιάζουν στην ανάλυση προτιμήσεων, το ιστορικό πλοήγησης και αγορών και στοχεύουν στη δημιουργία εξατομικευμένων προτάσεων προϊοντων με στόχο την αύξηση της πιθανότητας πραγματοποίησης αγορών. (*Personalized Online Advertising Effectiveness: The Interplay of What, When, and Where | Marketing Science*, n.d.; *Putting One-to-One Marketing to Work: Personalization, Customization, and Choice | Marketing Letters | Springer Nature Link*, n.d.; *The Netflix Recommender System: Algorithms, Business Value, and Innovation: ACM Transactions on Management Information Systems: Vol 6, No 4*, n.d.). Επίσης, η εμπιστοσύνη, ο κίνδυνος και η αντιλαμβανόμενη χρησιμότητα είναι παράγοντες που επηρεάζουν την αγοραστική συμπεριφορά. Σύμφωνα με μελέτες, η εμπιστοσύνη στην πλατφόρμα, στο πωλητή και στην ασφάλεια πληρωμών είναι βασικοί παράγοντες για την ολοκλήρωση μίας αγοράς (Gefen et al., 2003; Kim et al., 2008). Ωστόσο, η κοινωνική επιρροή ενισχύει τη συμπεριφορά του χρήστη ώστε να προχωρήσει σε αγορά ή ακόμη και να απορρίψει το προϊόν (*The Effect of Word of Mouth on Sales: Online Book Reviews - Judith A. Chevalier, Dina Mayzlin, 2006*, n.d.; Verhoef et al., 2015).

2.4 Ο Ρόλος της Ανάλυσης Δεδομένων και της Μηχανικής Μάθησης στο Ηλεκτρονικό Εμπόριο

Με την εξέλιξη της τεχνολογίας, η μηχανική μάθηση αποτελεί έναν από τους σημαντικότερους παράγοντες του ηλεκτρονικού εμπορίου. Στόχος της είναι να μπορέσουν οι επιχειρήσεις να διαχειριστούν τον όγκο πληροφοριών που έχουν αντλήσει από τις διαδικτυακές αγορές, με σκοπό τη βελτιστοποίηση και την ενίσχυση της εμπειρίας του πελάτη. Το ηλεκτρονικό εμπόριο παρέχει τεράστια ποικιλία δεδομένων, όπως τα *clicks*, το ιστορικό αγορών, τα δημογραφικά στοιχεία των πελατών καθώς και το πλήθος προβολών προϊόντων, τα οποία μπορούν να αξιοποιηθούν για την εξαγωγή επιχειρησιακής γνώσης μέσω τεχνικών ανάλυσης δεδομένων (Chen et al., 2012; Günther et al., 2017; Wamba et al., 2017). Η διαδικασία ανάλυσης δεδομένων για τη μελέτη των καταναλωτικών τάσεων στις ηλεκτρονικές αγορές ακολουθεί μια δομημένη σειρά σταδίων, όπως αποτυπώνεται συνοπτικά στο Σχήμα 4.

Σχήμα 4: Διαδικασία ανάλυσης δεδομένων για τη μελέτη των καταναλωτικών τάσεων στις ηλεκτρονικές αγορές
Πηγή: Ίδια επεξεργασία

Ειδικότερα, η μηχανική μάθηση χρησιμοποιείται ως εργαλείο για την κατανόηση και την πρόβλεψη της συμπεριφοράς του καταναλωτή, δημιουργώντας συστήματα εξατομίκευσης (personalization) τα οποία προσαρμόζουν την εμπειρία του καταναλωτή από προηγούμενες προτιμήσεις και ενέργειες. Χαρακτηριστικό παράδειγμα αποτελούν τα συστήματα συστάσεων τα οποία χρησιμοποιούν αλγορίθμους όπως collaborative filtering και matrix factorization, οι οποίοι έχουν την τάση να προτείνουν προϊόντα που είναι πολύ πιθανόν να προσελκύσουν το ενδιαφέρον των πελατών (Jannach & Adomavicius, 2016; Ricci et al., 2022). Σημαντικά παραδείγματα αποτελεσματικότητας τέτοιων αλγορίθμων είναι η εμφάνιση τους σε πλατφόρμες όπως η Amazon και το Netflix. Οι συγκεκριμένες εταιρείες,

χρησιμοποιούν μοντέλα μηχανικής μάθησης για να προσαρμόζουν και να εξατομικεύουν την εμπειρία των χρηστών η οποία επιτυγχάνει μεγαλύτερα ποσοστά ικανοποίησης τους και παράλληλα αυξάνει τις πωλήσεις (Smith & Linden, 2017; *The Netflix Recommender System: Algorithms, Business Value, and Innovation: ACM Transactions on Management Information Systems: Vol 6, No 4*, n.d.).

Ακόμη, η ανάλυση δεδομένων αποτελεί βασικό εργαλείο για τη διαχείριση τιμών και δυναμικών στρατηγικών τιμολόγησης (dynamic pricing), όπου οι τιμές διαμορφώνονται σε πραγματικό χρόνο βάσει της ζήτησης και της διαθεσιμότητας προϊόντων. Σύμφωνα με μελέτες, οι εταιρείες που χρησιμοποιούν στρατηγικές δυναμικής τιμολόγησης πετυχαίνουν μεγαλύτερα περιθώρια κέρδους και ανταγωνιστικό πλεονέκτημα (*Dynamic Pricing and Learning*, 2013);(Elmaghraby & Keskinocak, 2003). Παράλληλα, οι τεχνικές πρόβλεψης ζήτησης, οι οποίες είναι βασισμένες σε μοντέλα μηχανικής μάθησης, ωθούν τις επιχειρήσεις να λάβουν αποφάσεις που επικεντρώνονται στη διαχείριση αποθεμάτων, διαφοροποίηση προϊόντων και προγραμματισμό προμηθειών.

Η ανάλυση μεγάλων δεδομένων πραγματεύεται την κατανόηση καταναλωτικών τάσεων και μοτίβων της αγοράς, χρησιμοποιώντας διάφορες τεχνικές, που θα αναλύσουμε στη συνέχεια, όπως η τμηματοποίηση πελατών (customer segmentation) που στηρίζεται σε δημογραφικά και συμπεριφορικά στοιχεία και έχει ως στόχο να βελτιστοποιήσει διάφορες στρατηγικές μάρκετινγκ και να δημιουργήσει πιο στοχευμένες καμπάνιες (Ngai et al., 2009). Γενικότερα, τα μεγάλα δεδομένα (big data) σχετίζονται με την λήψη στρατηγικών αποφάσεων, ενώ οι επιχειρήσεις που τα χρησιμοποιούν τείνουν να έχουν υψηλότερα επίπεδα παραγωγικότητας και καινοτομίας (Brynjolfsson et al., 2011).

Επιπλέον, οι επιχειρήσεις μπορούν να εκμεταλλευτούν δεδομένα του ηλεκτρονικού εμπορίου για τον εντοπισμό απάτης (fraud detection). Συγκεκριμένα, χρησιμοποιούνται αλγόριθμοι ταξινόμησης όπως Random Forest, Support Vector Machines ή Neural Networks οι οποίοι έχουν την δυνατότητα να εντοπίζουν ύποπτες συναλλαγές, μειώνοντας τον οικονομικό κίνδυνο και ενισχύοντας την ασφάλεια των πλατφορμών (Bhatnagar & Ghose, 2004) ; (Dal Pozzolo et al., 2015).

Συνολικά, η ανάλυση δεδομένων και η μηχανική μάθηση αποτελούν βασικά εργαλεία για τα ηλεκτρονικά καταστήματα, καθώς βοηθούν στην πρόβλεψη των καταναλωτικών συμπεριφορών, στην βελτιστοποίηση των λειτουργιών καθώς και στη δημιουργία ανταγωνιστικού πλεονεκτήματος.

ΚΕΦΑΛΑΙΟ 3: Κατηγορίες και Τύποι Μοντέλων Μηχανικής Μάθησης

3.1 Εισαγωγή στη Μηχανική Μάθηση

Τα τελευταία χρόνια, η ραγδαία εξέλιξη της τεχνολογίας έχει αναπτύξει διάφορους τομείς όπως η τεχνητή νοημοσύνη (*Artificial Inteligence - AI),* η οποία περιλαμβάνει πολλούς επιμέρους κλάδους. Ένας από τους σημαντικότερους είναι η μηχανική μάθηση (*Machine Learnirng - ML*), η οποία επικεντρώνεται στην ικανότητα του υπολογιστικού συστήματος να «μαθαίνει» από δεδομένα και να λειτουργεί ανεξάρτητα και αυτόνομα, χωρίς να χρειάζεται η ανθρώπινη παρέμβαση για προγραμματισμό (Mitchell, 1997).

Η λειτουργία της μηχανικής μάθησης επικεντρώνεται στην άντληση γνώσεων από πεδία όπως η στατιστική, η πληροφορική και τα μαθηματικά. Ουσιαστικά, η διαδικασία έχει ως στόχο την ανάπτυξη αλγορίθμων και μοντέλων, τα οποία θα ερευνούν πρότυπα (*patterns*), θα κάνουν προβλέψεις και θα λαμβάνουν αποφάσεις οι οποίες θα βασίζονται σε εμπειρικά δεδομένα (*Pattern Recognition and Machine Learning | Springer Nature Link,* n.d.). Ειδικότερα, πρόκειται για αλγορίθμους μηχανικής μάθησης οι οποίοι εκπαιδεύονται σε σύνολα δεδομένων και, μέσω επαναλαμβανόμενης επεξεργασίας, βελτιώνουν σταδιακά την απόδοσή τους (Kim, 2016).

Η μηχανική μάθηση περιλαμβάνει πολλά κοινά στοιχεία με την στατιστική, ωστόσο η πρώτη εμφανίζει σημαντικές διαφορές ως προς το στόχο και την φιλοσοφία της. Συγκεκριμένα, ο τομέας της στατιστικής επικεντρώνεται στην εξήγηση των σχέσεων μεταξύ των μεταβλητών και στον έλεγχο υποθέσεων, με την προϋπόθεση ότι βασίζονται σε θεωρητικά μοντέλα και κατανομές (Bzdok et al., 2018; Shmueli, 2010). Αντιθέτως, η μηχανική μάθηση εστιάζει στην πρόβλεψη και στην ικανότητα ενός μοντέλου να μπορεί να ανταπεξέλθει, με όσο το δυνατόν μεγαλύτερη ακρίβεια, σε νέα και άγνωστα δεδομένα (Breiman, 2001; Murphy, 2012). Παρά τις διαφορές τους, είναι σημαντικό να σημειώσουμε ότι η μηχανική μάθηση έχει εξελιχθεί με βάση την στατιστική, καθώς τα μοντέλα της χρησιμοποιούν στατιστικές αρχές. Ενώ η στατιστική επικεντρώνεται στην ερμηνεία των σχέσεων μεταξύ μεταβλητών και στην εξαγωγή συμπερασμάτων, η μηχανική μάθηση δίνει έμφαση στην πρόβλεψη και στη γενίκευση, όπως για παράδειγμα στην εκτίμηση της πιθανότητας ένας χρήστης να πραγματοποιήσει αγορά (Chong et al., 2015; Sharda et al., 2018).

Η ουσία της μηχανικής μάθησης έγκειται στην δημιουργία συστημάτων, τα οποία βελτιώνουν την αποτελεσματικότητα τους μέσω της έκθεσης τους σε νέες πληροφορίες και στην εμπειρία που έχουν αναπτύξει. Στην πραγματικότητα, ο αλγόριθμος εξελίσσεται συνεχώς καθώς τροφοδοτείται από όλο και περισσότερα δεδομένα τα οποία τον καθιστούν ικανό να λαμβάνει αποφάσεις καθώς και να κάνει προβλέψεις με υψηλή ακρίβεια.

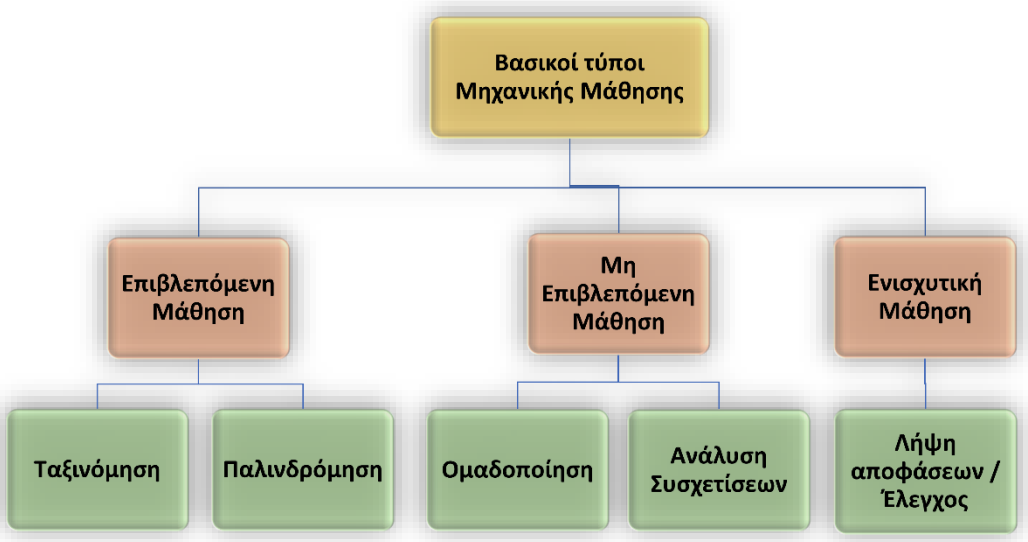
Η εφαρμογή της μηχανικής μάθησης έχει γίνει απαραίτητη σε πολλούς τομείς. Συγκεκριμένα , έχει χρησιμοποιηθεί στην ιατρική, για την έγκαιρη διάγνωση ασθενειών όπως ο καρκίνος οδηγώντας τους ιατρούς να προτείνουν ταχύτερα θεραπείες στους ασθενείς

τους (Esteva et al., 2017). Ακόμη, σημαντική είναι η εμφάνιση της στο χρηματοοικονομικό τομέα τόσο για την ανάλυση οικονομικών συναλλαγών για διαπίστωση ύποπτων κινήσεων, όσο και για τις προβλέψεις των τιμών των μετοχών (‘(PDF) Machine Learning in Accounting & Finance’, 2025). Παράλληλα, χρησιμοποιείται από μεγάλες εταιρείες όπως η Google η οποία την έχει ενσωματώσει στην εφαρμογή των χαρτών της, ώστε να αναλύει και να προβλέπει την κίνηση στους δρόμους (Vishwakarma & Verma, 2025).

Ωστόσο, ένας από τους σημαντικότερους τομείς που έχουν επηρεαστεί από τη χρήση της μηχανικής μάθησης είναι το ηλεκτρονικό εμπόριο. Συγκεκριμένα, η μηχανική μάθηση συμβάλλει στην κατανόηση των προτιμήσεων των καταναλωτών, αξιοποιώντας δημογραφικά χαρακτηριστικά και δεδομένα που προκύπτουν από το ιστορικό προηγούμενων αγορών (Nguyen et al., 2014).

Μέσω της ανάλυσης της αγοραστικής συμπεριφοράς, όπως των προϊόντων που προστίθενται στο ηλεκτρονικό καλάθι, τα συστήματα μηχανικής μάθησης είναι σε θέση να δημιουργούν εξατομικευμένες προτάσεις προϊόντων, προσαρμοσμένες στις ανάγκες και τα ενδιαφέροντα κάθε χρήστη. Επομένως, η μηχανική μάθηση ενισχύει το ψηφιακό marketing των επιχειρήσεων, και υποστηρίζει τη βελτιστοποίηση της διαχείρισης αποθεμάτων βάσει των χαρακτηριστικών και των συνηθειών του πελατολογίου τους (Han et al., 2023).

Στο πλαίσιο αυτό, η μηχανική μάθηση διακρίνεται σε βασικές κατηγορίες, οι οποίες παρουσιάζονται συνοπτικά στο Σχήμα 5.



Σχήμα 5: Βασικοί τύποι μηχανικής μάθησης
Πηγή: Ίδια επεξεργασία

3.2 Κατηγορίες Μηχανικής Μάθησης

Ο τομέας της Μηχανικής Μάθησης διακρίνεται σε τρεις κύριες κατηγορίες ανάλογα με το είδος και τη μορφή των δεδομένων που καλείται να αντιμετωπίσει ο αλγόριθμος κατά την εκπαίδευσή του. Πρόκειται για την κατηγοριοποίηση της σε Επιβλεπόμενη Μάθηση (supervised learning), Μη Επιβλεπόμενη Μάθηση (unsupervised learning) και Ενισχυτική Μάθηση (reinforcement learning).

Κάθε κατηγορία εφαρμόζεται σε διαφορετικού τύπου προβλήματα, με την βοήθεια κατάλληλων τεχνικών και μεθόδων. Οι προσεγγίσεις αυτές, χρησιμοποιούνται σε ποικίλους τομείς, όπως η υγεία, το ηλεκτρονικό εμόριο, οι χρηματοοικονομικές συναλλαγές τονίζοντας τη σημασία της μηχανικής μάθησης στις σύγχρονες εφαρμογές.

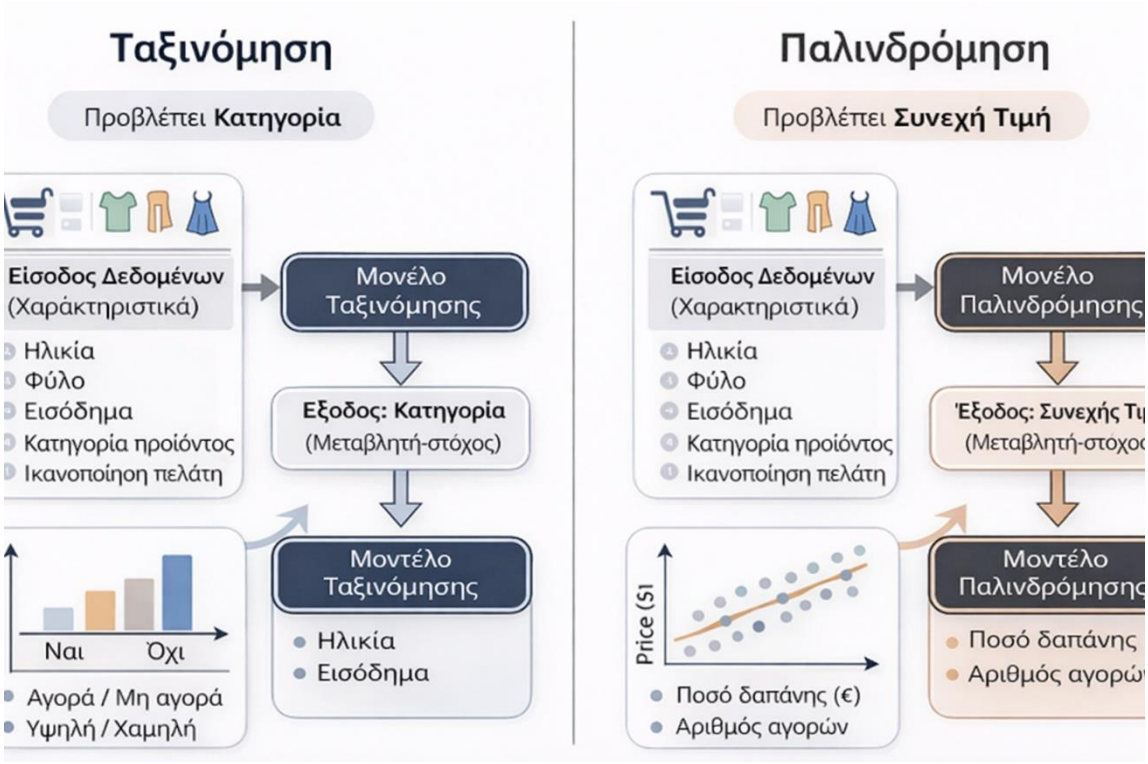
3.2.1 Επιβλεπόμενη Μάθηση

Η επιβλεπόμενη μάθηση (supervised learning) αποτελεί την πιο διαδεδομένη και συνηθέστερη μορφή της μηχανικής μάθησης. Το πλαίσιο εκπαίδευσης της βασίζεται σε ένα σύνολο δεδομένων (dataset) το οποίο περιλαμβάνει τόσο τις εισόδους (features) όσο και τις αντίστοιχες εξόδους (labels). Σκοπός της επιβλεπόμενης μάθησης είναι να εκπαιδεύσει τον αλγόριθμο ώστε να καταφέρει να κατανοήσει τη σχέση που συνδέει τις εισόδους με τις αντίστοιχες εξόδους, προκειμένου στη συνέχεια να προβλέψει με ακρίβεια τις άγνωστες εξόδους. (Nasteski, 2017)

Η διαδικασία αυτή, θα μπορούσε να παρομοιαστεί με την εκπαίδευση ενός μαθητή από ένα δάσκαλο, ο μαθητής – μοντέλο εκπαιδεύεται σε γνωστά δεδομένα, λαμβάνοντας διορθώσεις για τα σφάλματά του από τον καθηγητή – αλγόριθμο για τα λάθη του, προκειμένου να βελτιωθεί σταδιακά (Kim, 2016). Κατά την διαδικασία αυτή, ο αλγόριθμος υπολογίζει τη διαφορά που προκύπτει από τις προβλεπόμενες και τις πραγματικές τιμές, γνωστή και ως σφάλμα (error) και προσαρμόζει τις παραμέτρους του ώστε να βελτιώσει την ακρίβεια πρόβλεψης (Shalev-Shwartz & Ben-David, 2014).

Η βασική έννοια της επιβλεπόμενης μάθησης είναι η ικανότητα γενίκευσης (generalization), κατά την οποία το μοντέλο είναι αποδοτικότερο στα νέα, άγνωστα δεδομένα (Vapnik, 1999). Για τον λόγο αυτό, εφαρμόζονται τεχνικές όπως η κανονικοποίηση (regularization), η διασταυρούμενη επικύρωση (cross-validation), και η προσεκτική επιλογή χαρακτηριστικών (Guyon & Elisseeff, n.d.).

Η επιβλεπόμενη μάθηση συναντάνται συχνά στο ηλεκτρονικό εμπόριο, ιδιαίτερα σε συστήματα πρόβλεψης αγοραστικής συμπεριφοράς, σε εφαρμογές για κατηγοριοποίηση και τμηματοποίηση πελατών καθώς και σε συστήματα που χρησιμοποιούν εξατομικευμένες προτάσεις προϊόντων στους πελάτες.



Σχήμα 6: σύγκριση ταξινόμησης και παλινδρόμησης στο πλαίσιο της επιβλεπόμενης μάθησης
Πηγή: Ίδια επεξεργασία

Στην παρούσα εργασία δίνεται έμφαση σε προβλήματα ταξινόμησης, καθώς ο στόχος είναι η πρόβλεψη κατηγοριών καταναλωτικής συμπεριφοράς.

3.2.1.1 Ταξινόμηση (classification)

Η ταξινόμηση αποτελεί έναν από τους κύριους τύπους επιβλεπόμενης μάθησης, όπου οι αλγόριθμοι αναλύουν δεδομένα, τα ταξινομούν σε κατηγορίες- κλάσεις και προβλέπουν την κατηγορία στην οποία ανήκουν τα νέα δεδομένα (Kotsiantis, n.d.) Οι δύο βασικότερες μορφές ταξινόμησης είναι η δυαδική ταξινόμηση (binary classification) και η πολυκατηγορική ταξινόμηση (multiclass classification). Στην πρώτη περίπτωση, η έξοδος περιέχει δύο κατηγορίες (π.χ. στο ηλεκτρονικό εμπόριο οι κατηγορίες είναι αγορά και μη αγορά προϊόντος), ενώ στη δεύτερη οι κατηγορίες είναι περισσότερες από δύο όπως διάφορα επίπεδα ικανοποίησης πελατών (π.χ. υψηλή, μέτρια, χαμηλή).

Μεταξύ των πιο γνωστών αλγορίθμων ταξινόμησης περιλαμβάνονται η Λογιστική Παλινδρόμηση (Logistic Regression), τα Δέντρα Αποφάσεων (Decision Trees), τα Τυχαία Δάση (Random Forests), ο Naive Bayes, οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines – SVM) και τα Νευρωνικά Δίκτυα (Neural Networks). Για να αξιολογηθούν οι συγκεκριμένοι αλγόριθμοι, συνήθως χρησιμοποιούνται μετρικές όπως η ακρίβεια (accuracy), η ευαισθησία (recall), η ακρίβεια πρόβλεψης (precision), ο δείκτης F1

(F1-score), καθώς και ο πίνακας σύγχυσης (confusion matrix) (Han et al., 2011; MacQueen, 1967).

Όσον αφορά το ηλεκτρονικό εμπόριο, η ταξινόμηση χρησιμοποιείται για την πρόβλεψη της πιθανότητας αγοράς ενός προϊόντος από έναν καταναλωτή. Ειδικότερα, λαμβάνοντας υπόψη δημογραφικά δεδομένα όπως η ηλικία, το εισόδημα, ο τόπος κατοικίας καθώς και ιστορικό από προηγούμενες αγορές, ο αλγόριθμος μπορεί να ταξινομήσει τους καταναλωτές σε κατηγορίες – κλάσεις ανάλογα με την πιθανότητα να πραγματοποιήσουν μελλοντική αγορά (π.χ. υψηλή, μέτρια ή χαμηλή πιθανότητα).

Ακόμη, στις σύγχρονες ψηφιακές πλατφόρμες και στα social media, όλο και περισσότεροι χρήστες σχολιάζουν και αξιολογούν τα προϊόντα και τις επιχειρήσεις. Στο πλαίσιο αυτό λαμβάνει χώρα η ταξινόμηση συναισθημάτων (sentiment classification) μέσω της οποίας αναλύονται τα σχόλια των χρηστών και ταξινομούνται σε θετικά, ουδέτερα και αρνητικά συναισθήματα.

Γενικότερα, η ταξινόμηση είναι μία σημαντική τεχνική που χρησιμοποιείται από πολλές επιχειρήσεις, καθώς η ανάλυση προφίλ και συναισθημάτων των καταναλωτών, βελτιστοποιούν τις στρατηγικές marketing και την προώθηση προϊόντων, διεκδικώντας την πιστότητα των πελατών τους (Gajowniczek & Ząbkowski, 2014).

3.2.1.2 Παλινδρόμηση (Regression)

Ένας ακόμη βασικός τύπος επιβλεπόμενης μάθησης είναι η παλινδρόμηση, η οποία χρησιμοποιείται για προβλέψεις που βασίζονται σε συνεχείς, αριθμητικές μεταβλητές (James et al., 2021). Σε αντίθεση με τη μέθοδο της ταξινόμησης, η οποία κατατάσσει τα δεδομένα της σε διακριτές κατηγορίες, η παλινδρόμηση έχει ως στόχο την πρόβλεψη μίας συνεχούς τιμής, βασισμένης σε προηγούμενα δεδομένα.

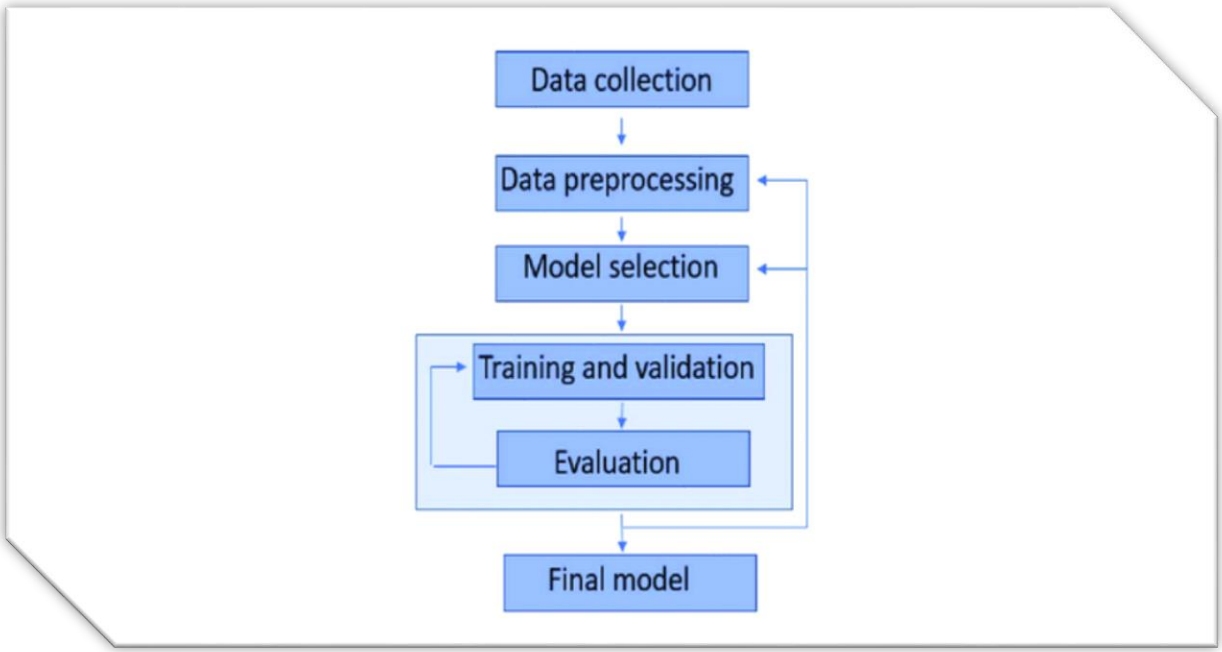
Υπάρχουν πολλοί αλγόριθμοι παλινδρόμησης, μερικοί από τους πιο διαδεδομένους είναι η Γραμμική Παλινδρόμηση (Linear Regression), η Πολυωνυμική Παλινδρόμηση (Polynomial Regression), και τα Νευρωνικά Δίκτυα Neural Networks.

Για να αξιολογηθούν τα μοντέλα της παλινδρόμησης, χρησιμοποιούνται μετρικές όπως το Μέσο Απόλυτο Σφάλμα (Mean Absolute Error – MAE), το Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error – MSE) και η Τετραγωνική Ρίζα του Μέσου Τετραγωνικού Σφάλματος (Root Mean Squared Error – RMSE), οι οποίες εντοπίζουν τη διαφορά μεταξύ πραγματικών και προβλεπόμενων τιμών. Για να αποφευχθεί η υπερπροσαρμογή (overfitting), δηλαδή η διαδικασία όπου ο αλγόριθμος μαθαίνει υπερβολικά στα δεδομένα εκπαίδευσης με αποτέλεσμα να αδυνατεί να γενικεύσει σε νέα δεδομένα, χρησιμοποιούνται τεχνικές κανονικοποίησης όπως L1 (Lasso) και L2 (Ridge) regularization (Hastie et al., 2009).

Η παλινδρόμηση είναι μία μέθοδος που χρησιμοποιείται από πολλές επιχειρήσεις στο πλαίσιο του ηλεκτρονικού εμπορίου, καθώς αναλύονται δεδομένα όπως οι διαφημιστικές δαπάνες, οι τιμές προϊόντων, δημογραφικά χαρακτηριστικά και το ιστορικό αγορών των καταναλωτών. Όλες αυτές οι μεταβλητές θεωρούνται ανεξάρτητες και στόχος της παλινδρόμησης είναι να ανακαλύψει τη σχέση τους με την εξαρτημένη μεταβλητή η οποία συνήθως αντιπροσωπεύει τις πωλήσεις.

Επιπλέον, η παλινδρόμηση εφαρμόζεται συχνά για την εκτίμηση μελλοντικών τιμών αγοράς προϊόντων, λαμβάνοντας υπόψη τη συμπεριφορά των καταναλωτών καθώς και εποχικούς παράγοντες, όπως η αυξημένη ζήτηση κατά την περίοδο των Χριστουγέννων. Χαρακτηριστικό παράδειγμα αποτελεί η Amazon η οποία χρησιμοποιεί μεθόδους προγνωστικής ανάλυσης και πολλαπλής γραμμικής παλινδρόμησης για την πρόβλεψη της ζήτησης των προϊόντων της και τη βελτιστοποίηση της διαχείρισης αποθεμάτων. Βασιζόμενη στην ανάλυση ιστορικών δεδομένων αγορών, δημογραφικών στοιχείων και εποχικότητας, η εταιρεία είναι σε θέση να προβλέπει τις μελλοντικές πωλήσεις ανά περιοχή και προϊόν, γεγονός που οδηγεί σε μείωση του λειτουργικού κόστους, βελτίωση των διαδικασιών logistics και διατήρηση υψηλών επιπέδων ικανοποίησης των πελατών όσον αφορά τους χρόνους παράδοσης των προϊόντων.

Τέλος, η παλινδρόμηση μετατρέπει τα δεδομένα σε ποσοτικές προβλέψεις και υποστηρίζει τη λήψη στρατηγικών αποφάσεων από τις επιχειρήσεις, τόσο σε επενδυτικό επίπεδο όσο και σε επίπεδο διαφημιστικών καμπανιών, με στόχο τη μείωση του επιχειρηματικού κινδύνου και την αύξηση της κερδοφορίας (James et al., 2021). Τα βασικά στάδια ανάπτυξης, εκπαίδευσης και αξιολόγησης ενός μοντέλου επιβλεπόμενης μάθησης παρουσιάζονται συνοπτικά στο Σχήμα 7.



Σχήμα 7: Βασικά στάδια ανάπτυξης, εκπαίδευσης και αξιολόγησης μοντέλου επιβλεπόμενης μάθησης
Πηγή: Ίδια επεξεργασία

3.2.2 Μη Επιβλεπόμενη Μάθηση

Η μη επιβλεπόμενη μάθηση είναι μία από τις τρεις βασικές μορφές της μηχανικής μάθησης. Στη συγκεκριμένη κατηγορία, ο αλγόριθμος εντοπίζει πρότυπα και σχέσεις σε σύνολα δεδομένων χωρίς την παρουσία προκαθορισμένων κατηγοριών ή ετικετών . Ο αλγόριθμος προσπαθεί να εντοπίσει την «λογική» που κρύβεται πίσω από τα δεδομένα, ανιχνεύοντας συσχετίσεις, ομοιότητες ή διαφορές, χωρίς να γνωρίζει εκ των προτέρων τις σωστές απαντήσεις, σε σύγκριση με την επιβλεπόμενη μάθηση, η οποία εκπαιδεύει τον αλγόριθμο με τα δεδομένα που έχει από τις εισόδους και τις αντίστοιχες εξόδους τους .

Η μη επιβλεπόμενη μάθηση έχει αυξηθεί με ταχύτατους ρυθμούς τα τελευταία χρόνια καθώς η συνεχής αύξηση του όγκου των δεδομένων που διαθέτουν οι επιχειρήσεις, εντείνει την ανάγκη για καινοτόμες μεθόδους ανάλυσης, χωρίς ανθρώπινη παρέμβαση. Συνεπώς, η συγκεκριμένη μορφή μάθησης είναι σημαντική για εταιρείες που διαθέτουν εκτενή σύνολα δεδομένων, χωρίς προηγούμενη γνώση των κρίσιμων χαρακτηριστικών τους, και προσπαθούν να εντοπίσουν άγνωστες συσχετίσεις και μοτίβα.

Οι συνηθέστερες τεχνικές που εφαρμόζονται στη μη επιβλεπόμενη μάθηση είναι η ομαδοποίηση (clustering), η ανάλυση συσχετίσεων (association analysis) και η μείωση διαστάσεων (dimensionality reduction). Η ομαδοποίηση αποτελεί την πιο γνωστή τεχνική, η οποία χρησιμοποιείται για να εντοπίσει ομάδες δεδομένων με παρόμοια χαρακτηριστικά, όπως πελάτες με κοινές συνήθειες και προτιμήσεις. Η ανάλυση συσχετίσεων στοχεύει στον εντοπισμό σχέσεων που παρουσιάζουν μερικά προϊόντα μεταξύ τους, όπως για παράδειγμα οι καταναλωτές που αγοράζουν φορητό ηλεκτρονικό υπολογιστή έχουν αυξημένη πιθανότητα να αγοράσουν και θήκη (Kaur & Kang, 2016).

Μία ακόμη από τις σημαντικότερες εφαρμογές της μη επιβλεπόμενης μάθησης είναι η ανίχνευση απάτης (fraud detection) και η κυβερνοασφάλεια, όπου αναλύονται μοτίβα συναλλαγών με σκοπό τον εντοπισμό ύποπτων συμπεριφορών, χωρίς να έχουν οριστεί εκ των προτέρων κανόνες ή παραδείγματα ύποπτων συναλλαγών. Όσον αφορά το ηλεκτρονικό εμπόριο, η μη επιβλεπόμενη μάθηση συμβάλλει στην κατανόηση της καταναλωτικής συμπεριφοράς, επιτρέποντας της να πραγματοποιεί τμηματοποίηση πελατών (customer segmentation) και να ενισχύει τις στρατηγικές μάρκετινγκ της.

Συνοψίζοντας, η μη επιβλεπόμενη μάθηση, μέσω των τεχνικών της, δίνει την δυνατότητα στις επιχειρήσεις να εντοπίζουν τάσεις και μοτίβα στα δεδομένα τους, υποστηρίζοντας την κατανόηση της καταναλωτικής συμπεριφοράς και την ανάπτυξη στοχευμένων στρατηγικών που μπορούν να τους προσφέρουν ανταγωνιστικό πλεονέκτημα στην αγορά.

3.2.2.1 Ομαδοποίηση (Clustering)

Η τεχνική της ομαδοποίησης αποτελεί μία από τις κυριότερες τεχνικές της μη επιβλεπόμενης μάθησης και έχει ως στόχο τον εντοπισμό φυσικών ομάδων μέσα σε πλήθος πληροφοριών. Ειδικότερα, χρησιμοποιείται για να διαχωρίσει τα δεδομένα σε ομάδες, γνωστές ως clusters, έτσι ώστε τα δεδομένα που βρίσκονται μέσα σε ίδια clusters να κατέχουν υψηλό βαθμό ομοιότητας ενώ τα υπόλοιπα δεδομένα να παρουσιάζουν αξιόλογες διαφορές μεταξύ τους. Οι ομάδες αυτές δημιουργούνται σύμφωνα με κοινά χαρακτηριστικά τα οποία μπορεί να προκύπτουν από δημογραφικά στοιχεία, αγοραστικά ή συμπεριφορικά χαρακτηριστικά.

Η ομαδοποίηση χρησιμοποιείται κυρίως σε περιπτώσεις όπου δεν υπάρχουν προκαθορισμένες κατηγορίες και οι οργανισμοί έχουν την ανάγκη να ανακαλύψουν πρότυπα και ομοιότητες στα δεδομένα τους. Επομένως, μέσω αυτής της τεχνικής οι επιχειρήσεις αποκτούν τη δυνατότητα να λάμβάνουν καλύτερες στρατηγικές αποφάσεις, καθώς μπορούν να κατανοήσουν σε βάθος το προφίλ και το πλαίσιο που κυμένονται το πελατολόγιο τους ώστε να εντοπίζουν ομάδες καταναλωτών με παρόμοιες προτιμήσεις (Xu & Wunsch, 2005).

Σημαντική είναι η εμφάνιση της στο ηλεκτρονικό εμπόριο όπου η ομαδοποίηση πελατών χρησιμοποείται για την τμηματοποίηση πελατών (*customer segmentation*), αξιοποιώντας πληροφορίες όπως οι προτιμήσεις σε προϊόντα, ο τόπος κατοικίας, το ιστορικό αγορών καθώς και πολλά άλλα χαρακτηριστικά που παίζουν σημαντικό ρόλο στην διαμόρφωση του καταναλωτικού το προφίλ (Tan et al., 2016; Wedel & Kamakura, 2000). Για παράδειγμα, ο αλγόριθμος ομαδοποίησης έχει την δυνατότητα να εντοπίζει ηλικιακές συσχετίσεις, όπως οι νεαρές ηλικίες 18-30 ετών συνηθίζουν να αγοράζουν ηλεκτρονικά προϊόντα (υπολογιστές, κινητά τηλέφωνα και ηχεία) ενώ καταναλωτές μεγαλύτερης ηλικίακής κλίμακας 40-65 ετών προτιμούν οικιακές συσκευές όπως σκεύη κουζίνας ή ηλεκτρικές σκούπες. Τα παραπάνω δεδομένα, μπορούν να αξιοποιηθούν από τις εταιρείες για τη διαμόρφωση προσαρμοσμένων διαφημιστικών καμπανιών, ανάλογα με το κοινό στο οποίο απευθύνονται.

Ένα άλλο στοχευμένο παράδειγμα αξιοποίησης τεχνικών ομαδοποίησης αποτελεί η χρήση αλγορίθμων clustering στην διαδικτυακή πλατφόρμα Netflix με τους οποίους ομαδοποιούνται οι τηλεθεατές βάσει των προτιμήσεων τους. Τα χαρακτηριστικά που αναλύονται οι μεταβλητές που αναλύονται αφορούν, μεταξύ άλλων, το είδος ταινίας (π.χ. κωμωδία, δράση), τις αξιολογήσεις, τη συχνότητα προβολής και το ιστορικό τηλεθέασης. Συνεπώς, το Netflix έχει την δυνατότητα να προσαρμόζει ποιο είδος ταινίας ή σειράς θα προτείνει σε κάθε ομάδα ώστε να είναι αντίστοιχο βάσει του προφίλ της. Σύμφωνα με μελέτες, περίπου το 80% του περιεχομένου που παρακολουθούν οι τηλεθεατές στην συγκεκριμένη πλατφόρμα, βασίζεται σε τέτοιου είδους εξατομικευμένες προτάσεις (*The Netflix Recommender System: Algorithms, Business Value, and Innovation: ACM Transactions on Management Information Systems: Vol 6, No 4,* n.d.).

Για την επιτυχημένη εφαρμογή της ομαδοποιήσης εφαρμόζονται διάφοροι αλγόριθμοι. Ένας από τους πιο διαδεδομένους είναι ο K-Means, και ακολουθούν οι Hierarchical Clustering και το DBSCAN. Ο K-Means διακρίνεται για την απλότητα και την υπολογιστική του αποδοτικότητα, δεδομένο που τον καθιστά κατάλληλο για μεγάλα σύνολα δεδομένων. Η

βασική αρχή λειτουργίας του στηρίζεται στην διάσπαση ενός συνόλου δεδομένων σε ομάδες k, γνωστές ως clusters, έτσι ώστε τα στοιχεία της ίδιας ομάδας να παρουσιάζουν μεγαλύτερη ομοιότητα και σαφή διαφοροποίηση από τα στοιχεία των άλλων ομάδων (Jain, 2010).

3.2.2.2 Ανάλυση συσχετίσεων (Association Analysis)

Μία ακόμη σημαντική τεχνική της μη επιβλεπόμενης μάθησης είναι η ανάλυση συσχετίσεων (Association Analysis) η οποία χρησιμοποιείται σε περιπτώσεις όπου οι οργανισμοί επιθυμούν να εντοπίσουν συνδυασμούς στοιχείων που εμφανίζονται συχνά μαζί στα δεδομένα τους. Αξιοποιώντας την συγκεκριμένη τεχνική, αναζητούνται πρότυπα και σχέσεις μεταξύ μεταβλητλων σε πλήθος δεδομένων, με στόχο την κατανόηση της ταυτόχρονης ύπαρξης συγκεκριμένων αντικειμένων ή χαρακτηριστικών. Συγκεκριμένα, η ανάλυση συσχετίσεων αποσκοπεί στον εντοπισμό προτύπων και σχέσεων μεταξύ μεταβλητών, εξάγοντας κανόνες που περιγράφουν τον βαθμό στον οποίο η παρουσία ενός στοιχείου σχετίζεται με την εμφάνιση ενός άλλου (Agrawal et al., 1993)

Οι κανόνες αυτοί ονομάζονται κανόνες συσχέτισης (association rules), αντιπροσωπεύονται συνήθως από τις μεταβλητές X και Y, ως σύνολα αντικειμένων και εκφράζονται με την μορφή «αν X, τότε Y» τα οποία περιγράφουν την συσχέτιση μεταξύ των μεταβλητών, δηλαδή κατά πόσο η εμφάνιση του αντικειμένου X αυξάνει την πιθανότητα της εμφάνισης του αντικειμένου Y.

Η ανάλυση καλαθιού αγορών (Market Basket Analysis), αποτελεί μία από τις πιο γνωστές εφαρμογές της ανάλυσησς συσχετίσεων. Η συγκεκριμένη μορφή, επιτρέπει στους οργανισμούς να εντοπίζουν επαναλαμβανόμενα πρότυπα συναλλαγών τα οποία υπάρχουν στα δεδομένα αγορών. Χρησιμοποιείται ευρέως τόσο στο ηλεκτρονικό εμπόριο όσο και στα καταστήματα λιανικής βοηθώντας τις επιχειρησεις να προωθούν προϊόντα με στρατηγικές διασταυρούμενης προώθησης (cross-selling) ή με μεθόδους ομαδοποιημένων προσφορών (bundle offers) με στόχο την αύξηση αξίας κάθε συναλλαγής (Tan et al., 2016). Για παράδειγμα, έχει παρατηρηθεί ότι οι καταναλωτές που αγοράζουν κινητό τηλέφωνο, συνήθως αγοράζουν και ακουστικά. Η γνώση αυτή μπορεί να αξιοποιηθεί από τις επιχειρήσεις για την εφαρμογή συνδυαστικών προσφορών, προσφέροντας έκπτωση σε συμπληρωματικά προϊόντα όταν αυτά αγοράζονται μαζί με ένα βασικό προϊόν (Brijs et al., 2004).

Χαρακτηριστικό παράδειγμα χρήσης της ανάλυσης συσχετίσεων αποτελεί η αλυσίδα σούπερ μάρκετ Tesco, η οποία αντλεί δεδομένα από τις κάρτες πιστότητας των πελατών της. Κάνοντας χρήση του αλγορίθμου Apriori, ανακάλυψε ότι οι καταναλωτές που αγόραζαν πάνες είχαν την τάση να αγόραζαν και ενεργειακά ποτά. Η συσχέτιση μεταξύ αυτών των δύο προϊόντων οδήγησε την Tesco να αναπροσαρμόσει την διάταξη των προϊόντων της στα ράφια και να εφαρμόσει στοχευμένες προσφορές, γεγονός που βοήθησε στην αύξηση των πωλήσεων και κατ’επεκταση, της κερδοφορίας της (Brijs et al., 2004; Larose & Larose, 2014).

Όσον αφορά την υλοποιήση της, η ανάλυση συσχετίσεων στηρίζεται κυρίως σε τρεις δείκτες: την υποστήριξη (support), την εμπιστοσύνη (confidence) και την ανύψωση (lift), οι οποίοι υπολογίζονται συνήθως με την χρήση του αλγορίθμου Apriori. Η υποστήριξη αναφέρεται στη συχνότητα εμφάνισης ενός συνδυασμού προϊόντων στο σύνολο των

συναλλαγων, η εμπιστοσύνη μετρά την πιθανότητα να αγοραστεί το προίον Y όταν έχει αγοραστεί το προϊόν X και η ανύψωση αξιολογεί την δύναμη της σχέσης μεταξύ των προϊόντων X και Y σε σχέση με την πιθανότητα να εμφανιστούν τυχαία σε συναλλαγές.

Συνοψίζοντας, η ανάλυση συσχετίσεων αποτελεί ένα πολύτιμο εργαλείο για τους οργανισμούς καθώς επιτρέπει την αναγνώριση καταναλωτικών τάσεων και πρότυπων στις ηλεκτρονικές αγορές. Οι οργανισμοί αξιοποιώντας τις πληροφορίες που προκύπτουν από την ανάλυση αγοραστικών επιλογών των καταναλωτών, έχουν την δυνατότητα να ενισχύουν τη θέση τους στην αγορά, να βελτιώνουν την εμπειρια του πελάτη και να εφαρμόζουν πιο αποτελεσματικές στρατηγικές προώθησης.

3.2.3 Ενισχυτική Μάθηση

Η ενισχυτική μάθηση (Reinforcement Learning – RL) αποτελεί τον τρίτο τύπο Μηχανικής μάθησης ο οποίος διαφοροποιείται σημαντικά από τους δύο προηγούμενους τύπους που έχουμε αναφέρει. Στην ενισχυτική μάθηση, η διαδικασία εκμάθησης στηρίζεται στην αλληλεπίδραση ενός πράκτορα (agent) με ένα περιβάλλον (environment). Ειδικότερα, ο πράκτορας βρίσκεται σε μία συγκεκριμένη κατάσταση του περιβάλλοντος και εκτελεί μία συγκεκριμένη ενέργεια, ενώ παράλληλα το περιβάλλον αλληλεπιδρά σε αυτή την ενέργεια ενημερώνοντας την νέα του κατάσταση και επιστρέφοντας στον πράκτορα τόσο την νέα κατάσταση όσο και τις αντίστοιχες ανταμοιβές ή ποινές ανάλογα με την έκβαση της δράσης του (Sutton & Barto, n.d.).

Γενικότερα, το σύστημα εκπαιδεύεται από τις δοκιμές και τα σφάλματα από τα οποία λαμβάνει ανταμοιβές ή ποινές ανάλογα με τις ενέργειες που εκτελεί. Ο πράκτορας εκπαιδεύεται και μαθαίνει ποιες ενέργειες οδηγούν σε αρνητικά και ποιες σε θετικά αποτελέσματα. Στόχος του είναι να αναπτύξει μία στρατηγική γνωστή ως πολιτική (policy), η οποία θα τον οδηγήσει στην επιλογή της ενέργειας που θα του μεγιστοποιήση την μελλοντική συνολική ανταμοιβή του (Kaelbling et al., n.d.).

Η ενισχυτική μάθηση επικεντρώνεται στην λήψη αποφάσεων σε περιβάλλοντα αβεβαιότητας και βασίζεται σε μαθηματικά μοντέλα, όπως οι Μαρκοβιακές διαδικασίες απόφασης (Markov Decision Processes – MDPs). Στο πλαίσιο αυτό, η μελλοντική κατάσταση του συστήματος εξαρτάται μόνο από την παρούσα κατάσταση και τις ενέργεις που εκτελεί ο πράκτορας, γεγονός που καθιστά εφικτή τη μοντελοποίηση σύνθετων προβλημάτων λήψης αποφάσεων που επικεντρώνεται σε δυναμικά περιβάλλοντα (Arulkumaran et al., 2017). Η ενισχυτική μάθηση χρησιμοποιείται σε σύγχρονες εφαρμογές, όπως τα συστήματα αυτόνομης λήψης αποφάσεων και οι αλγόριθμοι παιχνιδιών, όπου το σύστημα εκπαιδεύεται μέσω της συνεχούς αλληλεπίδρασης με το περιβάλλον (Brunner, n.d.).

3.3 Βαθιά μάθηση και Νευρωνικά δίκτυα (Deep Learning & Neural Networks)

Η βαθιά μάθηση αποτελεί έναν εκρηκτικά αναπτυσσόμενο κλάδο της μηχανικής μάθησης ο οποίος στηρίζεται στα τεχνητά νευρωνικα δίκτυα για την επεξεργασία δεδομένων και κατ'επέκταση την εξαγωγή προτύπων. Η ονομασία τους δεν είναι τυχαία, καθώς προκύπτει από τους νευρώνες του ανθρώπινου εγκεφάλου οι οποίοι αλληλεπιδρούν μεταξύ τους και μεταδίδουν πληροφορίες με την χρήση σημάτων. Παράλληλα στα τεχνητά νευρωνικά δίκτυα, οι νευρώνες οργανώνονται σε διακριτά στρώματα (layers) και επεξεργάζονται δεδομένα τα οποία προσαρμόζονται κατά την διαδικασία εκπαίδευσης (Kim, 2016).

Στην βαθιά μάθηση τα μοντέλα μαθαίνουν να λειτουργούν με ιεραρχικες αναπαραστάσεις δεδομένων μέσω της χρήσης κρυφών στρωμάτων (hidden layers). Ειδικότερα, σε εφαρμογές όπως η αναγνώριση μία εικόνας, τα αρχικά στρώματα εστιάζουν σε απλές δομές όπως γραμμές και άκρα ενώ τα βαθύτερα στρώματα προχωρούν σε αναγνώριση πιο σύνθετων μοτίβων, όπως προσώπων και αντικειμένων. Επομένως, η ικανότητα αυτή καθιστά την βαθιά μάθηση περισσότερο αποτελεσματική σε προβλήματα που εμφανίζουν υψηλή πολυπλοκότητα και μεγάλη διάσταση δεδομένων (LeCun et al., 2015; *Schmidhuber - 2015 - Deep Learning in Neural Networks An Overview | PDF | Deep Learning | Artificial Neural Network*, n.d.).

Τα νευρωνικά δίκτυα στηρίζονται στην διαδικασία οπισθοδιάδοσης του σφάλματος (backpropagation), κατά την οποία το μοντέλο ενημερώνει τα δεδομένα του και υπολογίζει τη διαφορά μεταξύ προβλεπόμενης και πραγματικής τιμής και προσαρμόζει τα βάρη του, ώστε να μειώσει το συνολικό σφάλμα πρόβλεψης. Για την αποτελεσματικότερη διαδικασία επκπαίδευσης χρησιμοποιούνται αλγόριθμοι όπως οι Stochastic Gradient Descent (SGD) και οι παραλλαγές του, Adam ή RMSProp (Ruder, 2017). Ωστόσο, η βαθιά μάθηση χρειάζεται μεγάλη υπολογιστικική ισχύ καθώς έχει να αντιμετωπίσει μεγάλο όγκο δεδομένων.

Όσον αφορά το ηλεκτρονικό εμπόριο, η βαθιά μάθηση έχει κάνει την εμφάνιση της σε διάφορους τομείς. Μία από τις σημαντικότερες εμφανίσεις της είναι τα συστήματα συστάσεων όπου τα μοντέλα της αναλύουν αγοραστικές συμπεριφορές και προτιμήσεις γα να δημιουργουν εξατομικευμένες προτάσεις (Zhang et al., 2019). Επιπλέον, εφαρμόζεται στη δυναμική τιμολόγηση κατά την οποια πραγματοποιείται η πρόβλεψη ζήτησης προϊοντων σε πραγματικό χρόνο με σκοπό την προσαρμογή τιμών από τις επιχειρήσεις, βάσει των συνθηκών της εκάστοτε αγοράς.

Η βαθιά μάθηση αποτελεί βασικό εργαλείο για την σύγχρονη ψηφιακή οικονομία βοηθώντας τις επειχηρήσεις να αναλύουν και να επεξεργάζονται πολύπλοκα δεδομένα με απότερο σκοπό να λαμβάνουν αποφάσεις και να επιτυγχάνουν μεγαλύτερη ακρίβεια και προβλέψεις.

ΚΕΦΑΛΑΙΟ 4 Μεθοδολογία και Δεδομένα

4.1 Επιλογή και περιγραφή του dataset (Kaggle)

Το πρακτικό μέρος της διπλωματικής εργασίας βασίζεται σε ένα σύνολο δεδομένων, το "Brazilian E-Commerce Public Dataset by Olist", το οποίο έχει δημοσιευτεί στην διαδικτυακή πλατφόρμα Kaggle από την βραζιλιάνικη εταιρεία Olist. Η Olist αποτελεί μία εταιρεία που δραστηριοποιείται στις ηλεκτρονικές αγορές (marketplace), συγκεκριμένα είναι μία ενδιάμεση πλατφόρμα ανάμεσα σε καταναλωτές και επιχειρήσεις. Πρόκειται για μία πλατφόρμα που είναι αντίστοιχη με το Skroutz (*Brazilian E-Commerce Public Dataset by Olist*, n.d.).

Το συγκεκριμένο dataset περιλαμβάνει πραγματικά δεδομένα που προκύπτουν από περίπου 100.000 παραγγελίες οι οποίες πραγματοποιήθηκαν μέσω της πλατφόρμας Olist, σε χρονικό διάστημα περίπου δύο ετών. Η χρήση πραγματικών συναλλαγών συμβάλλει στην αξιοπιστία των δεδομένων, το οποίο καθίσταται κατάλληλο για αναλύσεις που βασίζονται σε καταναλωτική συμπεριφορά και την εφαρμογή τεχνικών μηχανικής μάθησης στο ηλεκτρονικό εμπόριο ('(PDF) An Innovation Resistance Theory Perspective on Mobile Payment Solutions', 2025). Επιπρόσθετα, το συγκεκριμένο dataset έχει χρησιμοποιηθεί σε πλήθος ερευνητικών εργασιών και εκπαιδευτικών παραδειγμάτων, γεγονός που αναδεικνύει τη συμβολή του στην επιστημονική ανάλυση (*Kaggle*, n.d.).

Σε αντίθεση με άλλα dataset που αποτυπώνουν τα δεδομένα τους σε έναν ενιαίο πίνακα, το συγκεκριμένο δομείται σε πολλούς επιμέρους πίνακες οι οποίοι αντιπροσωπεύουν διαφορετικές πτυχές της αγοραστικής διαδικασίας. Ειδικότερα, μερικοί από τους πίνακες αποτελούν αρχεία που αφορούν τις παραγγελίες (orders), τα προϊόντα (products), τα στοιχεία των πελατών (customers), τις πληρωμές (payments), τις αξιολογήσεις (reviews), τους πωλητές (sellers) και τα γεωγραφικά στοιχεία (geolocation). Οι πίνακες συνδέονται μεταξύ τους μέσω μοναδικών αναγνωριστικών όπως order_id, customer_id, product_id και seller_id, τα οποία καθιστούν εφικτή την ενοποίηση των δεδομένων και την δημιουργία ενός ενιαίου πίνακα για αναλυτικούς σκοπούς (Kotsiantis et al., 2006).

Ο βασικός πίνακας είναι το αρχείο olist_orders_dataset.csv που περιλαμβάνει στοιχεία όπως το μοναδικό αναγνωριστικό της παραγγελίας (order_id), το αναγνωριστικό του πελάτη (customer_id), την ημερομηνία καταχώρισης, έγκρισης και παράδοσης της παραγγελίας, καθώς και πληροφορίες που αφορούν την κατάσταση της παραγγελίας (π.χ. delivered, canceled). Το αρχείο olist_order_items_dataset.csv περιλαμβάνει δεδομένα σχετικά με τα επιμέρους προϊόντα κάθε παραγγελίας (order items), τις αντίστοιχες ποσότητες, τιμές και τον πωλητή (seller_id). Το αρχείο olist_order_payments_dataset.csv περιέχει πληροφορίες σχετικά με την πληρωμή των παραγγελιών, τον αριθμό των δόσεων καθώς και το τελικό ποσό πληρωμής.

Ένα σημαντικά χρήσιμο αρχείο είναι το olist_order_reviews_dataset.csv, το οποίο επικεντρώνεται στις αξιολογήσεις των καταναλωτών σχετικά με τις παραγγελίες. Το συγκεκριμένο περιέχει μεταβλητές όπως η βαθμολογία (review_score), η ημερομηνία υποβολής αξιολογήσης και σε ορισμένες περιπτώσεις περιλαμβάνονται σχόλια. Ωστόσο το αρχείο olist_products_dataset.csv περιέχει λεπτομέρειες σχετικά με τα προϊόντα, τις διαστάσεις, και το βάρος ενώ το olist_customers_dataset.csv καθώς και ανωνυμοποιημένα

στοιχεία των πελατών, όπως πόλη και ταχυδρομικός κώδικας. Τέλος, το olist_geolocation_dataset.csv περιλαμβάνει γεωγραφικά χαρακτηριστικά όπως το γεωγραφικό πλάτος και μήκος.

Η ύπαρξη πολλών διαφορετικών συσχετιζόμενων πινάκων καθιστά το συγκεκριμένο dataset κατάλληλο για την συγκεκριμένη διπλωματική εργασία καθώς έπειτα απο την ενοποιήση τους, μέσω της διαδικασίας συγχώνευσης (merge), δίνεται η δυνατότητα ανάλυσης πολύπλοκων σχέσεων σχετικά με τις καταναλωτικές τάσεις στις ηλεκτρονικές αγορές. Ο συνδυασμός πληροφοριών που σχετίζονται με τα προϊόντα, τις πληρωμές, τους χρόνους παράδοσης και τις αξιολογήσεις των καταναλωτών παρέχει την δυνατότητα ανάπτυξης αναλυτικών μοντέλων για την αξιολόγηση της αγοραστικής συμπεριφοράς, την τμηματοποίηση των πελατών καθώς και την πρόβλεψη κρίσιμων δεικτών απόδοσης στο ηλεκτρονικό εμπόριο (Wamba et al., 2017).

4.2 Προετοιμασία και ενοποίηση δεδομένων

Πριν την πραγματοποίηση της περιγραφικής ανάλυσης και των τεχνικών μηχανικής μάθησης, προηγείται η προετοιμασία και η ενοποίηση των δεδομένων. Αυτό αποτελεί το πρώτο κρίσιμο βήμα της μεθοδολογίας ανάλυσης δεδομένων, διότι είναι γνωστό ότι η ποιότητα και η δομή τους είναι υπεύθυνες για την εγκυρότητα και την αξιοπιστία των αποτελεσμάτων που προκύπτουν από τη μελέτη (García et al., 2015; Pyle, 1999).

Στην παρούσα διπλωματική εργασία, για την επεξεργασία των δεδομένων χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python, στο περιβάλλον Anaconda, καθώς προσφέρει ένα πλήρες πλαίσιο ανάλυσης δεδομένων, με την χρήση εξιδεικευμένων βιβλιοθηκών. Ειδικότερα, χρησιμοποιήθηκε η βιβλιοθήκη pandas, η οποία προσφέρει αποδοτική διαχείριση δομημένων δεδομένων σε μορφή πινάκων (DataFrames) και αποτελεί κύριο εργαλείο για τέτοιου είδους διαδικασίες προετοιμασίας δεδομένων και ενοποιήσης (McKinney, 2012). Μέσω της συγκεκριμένης βιβλιοθήκης πραγματοποιήθηκε η εισαγωγή των αρχείων CSV τα οποία συνθέτουν το σύνολο δεδομένων της Olist.

```
# Εισαγωγή της βιβλιοθήκης pandas για ανάλυση και διαχείριση δεδομένων
import pandas as pd

# Φόρτωση του dataset παραγγελιών (orders),το οποίο περιλαμβάνει πληροφορίες
# σχετικά με την κατάσταση και τα χρονικά σημεία εξέλιξης κάθε παραγγελίας
orders = pd.read_csv("olist_orders_dataset.csv")

# Εμφάνιση των πρώτων εγγραφών του συνόλου δεδομένων για προκαταρκτική διερεύνηση
orders.head()
```

	order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	or
0	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-10-04 19:55:00	
1	53cdb2fc8bc7dce0b6741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	2018-07-24 20:41:37	2018-07-26 03:24:27	2018-07-26 14:31:00	
2	47770eb9100c2d0c44946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	2018-08-08 08:38:49	2018-08-08 08:55:23	2018-08-08 13:50:00	
3	949d5b44dbf5de918fe9c16f97b45f8a	f88197465ea7920adcdbec7375364d82	delivered	2017-11-18 19:28:06	2017-11-18 19:45:59	2017-11-22 13:39:59	
4	ad21c59c0840e6cb83a9ceb5573f8159	8ab97904e6daea8866dbdbc4fb7aad2c	delivered	2018-02-13 21:18:39	2018-02-13 22:20:29	2018-02-14 19:46:34	

Αρχικά, έγινε εισαγωγή των επιμέρους συνόλων δεδομένων που εστιάζουν στις παραγγελίες, τους πελάτες και τις συναλλαγές με σκοπό τον έλεγχο της δομής των δεδομένων μέσω βασικών διερευνητικών εντολών, όπως η info(). Στη συνέχεια, ελέγχθηκε ο αριθμός των εγγραφών, οι διαθέσιμες μεταβλητές και οι τύποι δεδομένων ώστε να εξακριβωθεί εάν οι τιμές καταχωρήθηκνα ορθός. Συγκεκριμένα, εξετάστηκε εάν οι μεταβλητές που δηλώνουν ημερομηνία έχουν όντως διατυπωθεί σε χρονικό ορίζοντα ως χρονικά δεδομένα και όχι ως απλό κείμενο, έτσι ώστε να εξασφαλιστεί η δυνατότητα αξιόπιστης χρήσης στην ανάλυση. Αυτή η διαδικασία, εκτός από την κατανόηση των δεδομένων, συνέβαλε και στον εντοπισμό πιθανών ασυνεπειών ή προβλημάτων ποιότητας.

```
# Έλεγχος της δομής του dataset παραγγελιών,
# Παρουσίαση τύπων δεδομένων και πλήθους μη κενών τιμών για κάθε μεταβλητή
orders.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99441 entries, 0 to 99440
Data columns (total 8 columns):
 #   Column                         Non-Null Count  Dtype 
---  ------                         --------------  ----- 
 0   order_id                       99441 non-null  object
 1   customer_id                    99441 non-null  object
 2   order_status                   99441 non-null  object
 3   order_purchase_timestamp       99441 non-null  object
 4   order_approved_at              99281 non-null  object
 5   order_delivered_carrier_date   97658 non-null  object
 6   order_delivered_customer_date  96476 non-null  object
 7   order_estimated_delivery_date  99441 non-null  object
dtypes: object(8)
memory usage: 6.1+ MB
```

```
# Περιγραφική σύνοψη των κατηγορικών (μη αριθμητικών) μεταβλητών του dataset παραγγελιών
# Παρουσιάζονται το πλήθος εγγραφών, οι μοναδικές τιμές και η συχνότερη εμφάνιση κάθε μεταβλητής
orders.describe()
```

	order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date
count	99441	99441	99441	99441	99281	97658
unique	99441	99441	8	98875	90733	81018
top	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2018-04-11 10:48:14	2018-02-27 04:31:10	2018-05-09 15:48:00
freq	1	1	96478	3	9	47

Έπειτα, εφόσον δόθηκε η εντολή read () στην python, τα δεδομένα αναγνώστηκαν και αποθηκεύτηκαν στην μνήμη. Ακολούθησε η εργασία της ενοποίησης των επιμέρους πινάκων σε έναν γενικό πίνακα, μέσω της διαδικασίας συγχώνευσης merge. Η μέθοδος της συγχώνευσης πραγματοποιήθηκε σχετικά με κοινά αναγνωριστικά πεδία όπως το *order_id* και *customer_id*, δημιουργώντας συσχέτιση πληροφοριών όσον αφορά τις παραγγελίες, τα χαρακτηριστικά πελατών και τα στοιχεία συναλλαγών. Από την διαδικασία αυτή, προέκυψε ένα ενιαίο σύνολο δεδομένων κατάλληλο για την ανάλυση της αγοραστικής συμπεριφοράς και την εφαρμογή τεχνικών μηχανικής μάθησης (*(PDF) Data Cleaning*, n.d.).

```python
# Φόρτωση του dataset προϊόντων ανά παραγγελία (τιμές προϊόντων και κόστος μεταφοράς)
items = pd.read_csv("olist_order_items_dataset.csv")
# Αρχική επισκόπηση των δεδομένων
items.head()
# Έλεγχος δομής, τύπων δεδομένων και πληρότητας των μεταβλητών
items.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 112650 entries, 0 to 112649
Data columns (total 7 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   order_id             112650 non-null  object
 1   order_item_id        112650 non-null  int64
 2   product_id           112650 non-null  object
 3   seller_id            112650 non-null  object
 4   shipping_limit_date  112650 non-null  object
 5   price                112650 non-null  float64
 6   freight_value        112650 non-null  float64
dtypes: float64(2), int64(1), object(4)
memory usage: 6.0+ MB
```

```python
# Φόρτωση του dataset προϊόντων, το οποίο περιλαμβάνει πληροφορίες
# σχετικά με την κατηγορία και τα βασικά χαρακτηριστικά κάθε προϊόντος
products = pd.read_csv("olist_products_dataset.csv")

products.head()
```

	product_id	product_category_name	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_lengt
0	1e9e8ef04dbcff4541ed26657ea517e5	perfumaria	40.0	287.0	1.0	225.0	
1	3aa071139cb16b67ca9e5dea641aaa2f	artes	44.0	276.0	1.0	1000.0	
2	96bd76ec8810374ed1b65e29197717f	esporte_lazer	46.0	250.0	1.0	154.0	
3	cef67bcfe19066a932b7673e239eb23d	bebes	27.0	261.0	1.0	371.0	
4	9dc1a7de274444849c219cff195d0b71	utilidades_domesticas	37.0	402.0	4.0	625.0	

```python
# Ενοποίηση δεδομένων παραγγελιών και πελατών
# Συγχώνευση μέσω μοναδικού αναγνωριστικού customer_id
# Επιλέγεται εσωτερική σύζευξη (inner join) ώστε να διατηρηθούν
# μόνο οι παραγγελίες που αντιστοιχούν σε έγκυρους πελάτες
orders_customers = pd.merge(
    orders,
    customers,
    on="customer_id",
    how="inner"
)
```

```python
# Ενοποίηση δεδομένων παραγγελιών, πελατών και προϊόντων με βάση το order_id
# Δημιουργία του τελικού συνόλου δεδομένων για ανάλυση
full_data = pd.merge(
    orders_customers,
    items,
    on="order_id",
    how="inner"
)

# Προβολή των πρώτων γραμμών του dataset
full_data.head()
```

order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_
7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-10-04 19:55:00	2017-10-10 21:
741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	2018-07-24 20:41:37	2018-07-26 03:24:27	2018-07-26 14:31:00	2018-08-07 15:
1946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	2018-08-08 08:38:49	2018-08-08 08:55:23	2018-08-08 13:50:00	2018-08-17 18:
1fe9c16f97b45f8a	f88197465ea7920adcdbec7375364d82	delivered	2017-11-18 19:28:06	2017-11-18 19:45:59	2017-11-22 13:39:59	2017-12-02 00:
3a9ceb5573f8159	8ab97904e6daea8866dbdbc4fb7aad2c	delivered	2018-02-13 21:18:39	2018-02-13 22:20:29	2018-02-14 19:46:34	2018-02-16 18:

Μετά την ολοκλήρωση της διαδικασίας συγχώνευσης των δεδομένων, το επόμενο βήμα επικεντρώνεται στον έλεγχο ύπαρξης ελλιπών τιμών στις μεταβλητές του ενοποιημένου συνόλου δεδομένων, μέσω της εντολής isnull(), καθώς επίσης και τον υπολογισμό του ποσοστού εμφάνισης τους. Ο έλεγχος αυτός είναι σημαντικός διότι αναλύει τα δεδομένα και εντοπίζει, εάν υπάρχουν, κενά που προκύπτουν από τυχόν ακυρώσεις παραγγελιών ή από καθυστερήσεις παραδόσεων ή ακόμα και ελλιπών καταγραφών κατά την συλλογή δεδομένων. Στο συγκεκριμένο σύνολο τα αποτελέσματα έδειξαν ότι το μεγαλύτερο ποσοστό

των δεδομένων είναι πλήρες ενώ ελλιπείς τιμές εντοπίζονται σε μεταβλητές που σχετίζονται με το χρόνο παράδοσης.

```
# Έλεγχος αριθμού ελλιπών τιμών ανά μεταβλητή
full_data.isnull().sum()
```

```
order_id                            0
customer_id                         0
order_status                        0
order_purchase_timestamp            0
order_approved_at                  15
order_delivered_carrier_date     1194
order_delivered_customer_date    2454
order_estimated_delivery_date       0
customer_unique_id                  0
customer_zip_code_prefix            0
customer_city                       0
customer_state                      0
order_item_id                       0
product_id                          0
seller_id                           0
shipping_limit_date                 0
price                               0
freight_value                       0
dtype: int64
```

```
# Υπολογισμός ποσοστού ελλιπών τιμών ανά μεταβλητή
(full_data.isnull().sum() / len(full_data)) * 100
```

```
order_id                         0.000000
customer_id                      0.000000
order_status                     0.000000
order_purchase_timestamp         0.000000
order_approved_at                0.013316
order_delivered_carrier_date     1.059920
order_delivered_customer_date    2.178429
order_estimated_delivery_date    0.000000
customer_unique_id               0.000000
customer_zip_code_prefix         0.000000
customer_city                    0.000000
customer_state                   0.000000
order_item_id                    0.000000
product_id                       0.000000
seller_id                        0.000000
shipping_limit_date              0.000000
price                            0.000000
freight_value                    0.000000
dtype: float64
```

Τέλος, μετά τον έλεγχο ποιότητας των δεδομένων, δημιουργήθηκε ένα καθαρό και ενοποιημένο σύνολο δεδομένων, κατάλληλο για περαιτέρω ανάλυση. Το σύνολο αυτό είναι το υπόβαθρο για την πραγματοποίηση της περιγραφικής ανάλυσης και την εφαρμογή τεχνικών μηχανικής μάθησης που ακολουθούν.

4.3 Περιγραφική Ανάλυση Δεδομένων

4.3.1 Ρόλος της περιγραφικής ανάλυσης στο ηλεκτρονικό εμπόριο

Η εκρηκτική ανάπτυξη του ηλεκτρονικού εμπορίου έχει οδηγήσει τις επιχειρήσεις στην συλλογή και την αποθήκευση τεράστιου όγκου δεδομένων με στόχο την ανάλυση των καταναλωτών σχετικά με την καταναλωτική τους συμπεριφορά, τις συναλλαγές τους καθώς και ότι σχετίζεται με τις διαδικασίες logistics. Ωστόσο, η απλή συλλογή των δεδομένων, χωρίς περαιτέρω ανάλυση, δεν αρκεί για τη δημιουργία επιχειρησιακής αξίας καθώς απαιτείται η συστηματική ανάλυσή τους μέσω της περιγραφικής ανάλυσης δεδομένων (descriptive data analysis), η οποία αφορά το πρώτο βήμα κάθε αναλυτικής διαδικασίας ((PDF) Data Science for Business, n.d.).

Η περιγραφική ανάλυση εστιάζει στην κατανόηση των κύριων πληροφοριών ενός συνόλου δεδομένων χρησιμοποιώντας στατιστικά μετρα, πίνακες και γραφικές απεικονίσεις. Στο πλαίσιο του ηλεκτρονικού εμπορίου, βοηθά τις επιχειρήσεις να κατανοήσουν ερωτήματα σχετικά με το πλήθος των παραγγελιών, τη γεωγραφική κατανομή της ζήτησης, την συχνότητα καθώς και την μέση αξία των συναλλαγών, συμβάλλοντας στη λήψη τεκμηριωμένων επιχειρηματικών αποφάσεων (Moro et al., 2016).

Η περιγραφική ανάλυση προηγείται της μηχανικής μάθησης και της προγνωστικής ανάλυσης. Σύμφωνα με τους James et al. (2021), πρέπει να κατανοηθεί η δομή των δεδομένων καθώς και τα βασικά μοτίβα ώστε τα μοντέλα που θα αναπτυχθούν στη συνέχεια να είναι αξιόπιστα. Σε περίπτωση που παραλειφθεί το συγκεκριμένο στάδιο, υπάρχει κίνδυνος για έκδοση λανθασμένων συμπερασμάτων ή εσφαλμένης ερμηνείας των αποτελεσμάτων (‘Multivariate Data Analysis’ by Joseph F. Hair, n.d.; (PDF) Data Science for Business, n.d.).

Στο περιβάλλον του ηλεκτρονικού εμπορίου, η περιγραφική ανάλυση συμβάλλει στην κατανόηση των δεδομένων σχετικά με την αγοραστική συμπεριφορά των καταναλωτών. Αναλύοντας μεταβλητές όπως η κατάσταση της παραγγελίας, η γεωγραφική τοποθεσία των πελατών και η αξία των αγορών οι επιχειρήσεις έχουν την δυνατότητα να εντοπίσουν τάσεις και πρότυπα. Για παράδειγμα, η ανάλυση της κατάστασης παραγγελίας (delivered, canceled, shipped) μέσω της όποιας παρέχονται σημαντικές πληροφορίες για την αποδοτικότητα των διαδικασιών logistics και την εμπειρία του πελάτη, μπορούν να επηρεάσουν άμεσα την ικανοποιήση, την πιστότητα και την διατήρηση των πελατών (Chaffey & Ellis-Chadwick, 2019).

Επιπλέον, η γεωγραφική ανάλυση των παραγγελίων συμβάλλει στην κατανόηση της κατανομής και της ζήτησης των προϊόντων ανά περιοχή ώστε να γίνεται πιο στοχευμένος στρατηγικός σχεδιασμός που θα οδηγεί σε βελτιστοποιημένα δίκτυα διανομής. Ωστόσο, η περιγραφική ανάλυση στην αξία της παραγγελίας, πραγματεύεται τη μέση καταναλωτική δαπάνη και βοηθά στη δημιουργία στρατηγικών τιμολόγησης και προώθησης προϊόντων βάσει δεδομένων (Wedel & Kannan, 2016).

Τέλος, σημαντικό πλεονέκτημα της περιγραφικής ανάλυσης είναι η συμβολή της στην βελτίωση της ποιότητας των δεδομένων. Ειδικότερα, ο έλεγχος των κενών τιμών, των ακραίων παρατηρήσεων και ασυνεπειών θέτει τα θεμέλια για τα επόμενα στάδια της ανάλυσης και εξασφαλίζει ότι τα δεδομένα που θα χρησιμοποιηθούν είναι αξιόπιστα για επεξεργασία (Batini, 2009). Σύμφωνα με τους (Shmueli et al., 2017), η ποιότητα των δεδομένων αποτελεί καθοριστικό παράγοντα για την επιτυχία οποιουδήποτε μοντέλου μηχανικής μάθησης. Στο πλαίσιο της παρούσας διπλωματικής εργασίας, η περιγραφική ανάλυση αποτελεί την γέφυρα μεταξύ της θεωρητικής προσέγγισης και της πρακτικής εφαρμογής των μεθόδων μηχανικής μάθησης σχετικά με τα δεδομένα πραγματικών συναλλαγών στο ηλεκτρονικό εμπόριο.

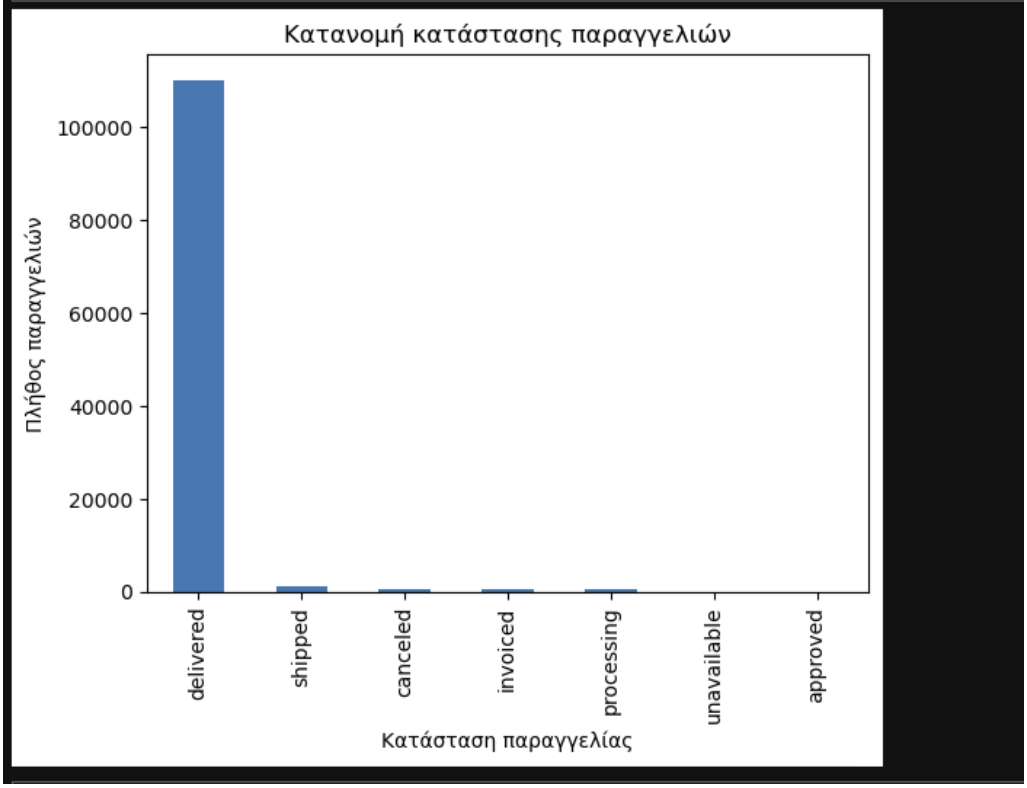
4.3.2 Περιγραφική ανάλυση του συνόλου δεδομένων Olist

Με στόχο την κατανόηση των δεδομένων και την ανάδειξη βασικών προτύπων που σχετίζονται με τις ηλεκτρονικές συναλλαγές, πραγματοποιήθηκε περιγραφική ανάλυση στο σύνολο των δεδομένων Olist, που δημιουργήθηκε έπειτα από το πλαίσιο της προετοιμασίας και ενοποίησης των δεδομένων. Η ανάλυση αυτή επιτρέπει την αρχική καταγραφή της συμπεριφοράς των παραγγελιών και χρησιμοποιείται ως βασικό εργαλείο για την διερεύνηση των δεδομένων.

Αρχικά, εξετάστηκε η μεταβλητή order_status, που παρουσιάζει την κατάσταση κάθε παραγγελίας. Σύμφωνα με το διάγραμμα 1, η πλειοψηφία των παραγγελιών ολοκληρώνεται με επιτυχία και χαρακτηρίζεται ως delivered. Αυτό παρουσιάζει υψηλό επίπεδο

αποτελεσματικότητας σχετικά με την εκτέλεση και παράδοση των παραγγελιών, το οποίο είναι πολύ σημαντικό για την εμπειρία του πελάτη στο ηλεκτρονικό εμπόριο.

```
# Εισαγωγή της βιβλιοθήκης matplotlib για οπτικοποίηση δεδομένων
import matplotlib.pyplot as plt

# Οπτικοποίηση της κατανομής των παραγγελιών ως προς την κατάσταση ολοκλήρωσης
status_counts = full_data["order_status"].value_counts()

status_counts.plot(kind="bar")
plt.title("Κατανομή κατάστασης παραγγελιών")
plt.xlabel("Κατάσταση παραγγελίας")
plt.ylabel("Πλήθος παραγγελιών")
plt.show()
```



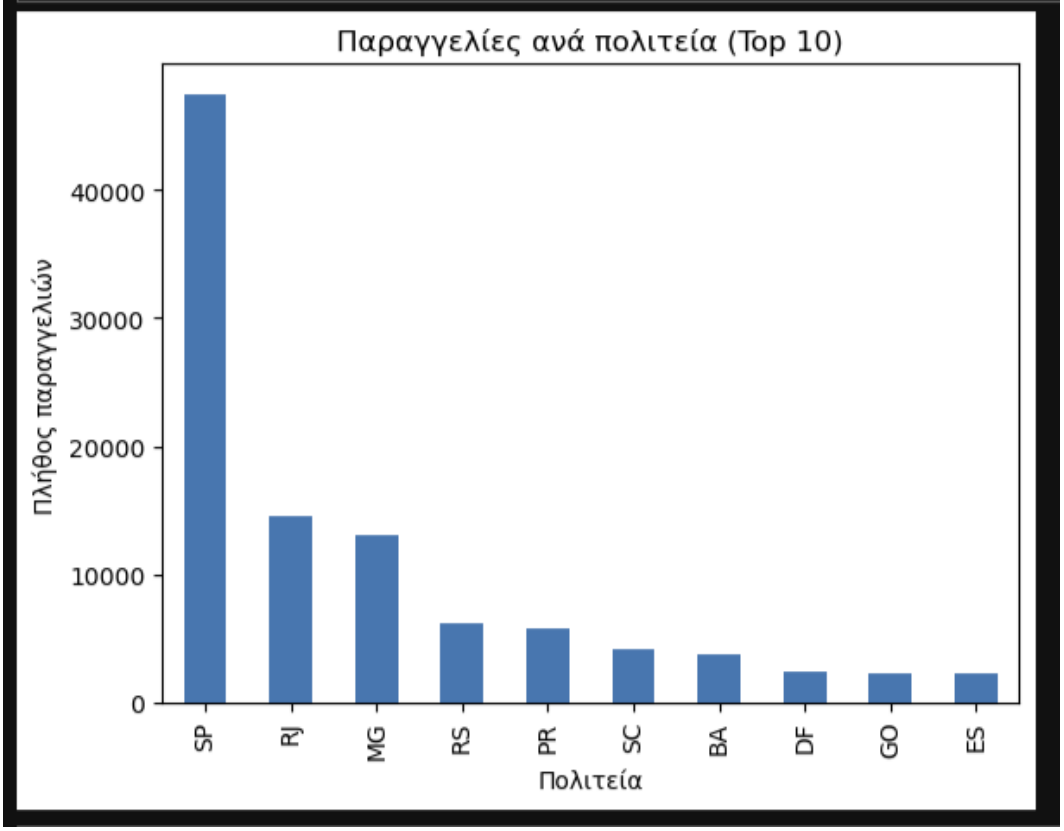
Διάγραμμα 1 Κατάσταση Παραγγελιών

Ωστόσο, παρατηρείται μικρό ποσοστό παραγγελιών που εμφανίζεται στις κατηγορίες canceled ή unavailable, γεγονός που θεωρείται αναμενόμενο σε πραγματικές συναλλαγές και μπορεί να εξαρτάται από διαθεσιμότητα προϊόντων, καθυστερήσεις παραδόσεων ή ακυρώσεις από τους καταναλωτές. Η ανάλυση της κατάστασης παραγγελίας είναι ιδιαίτερα σημαντική καθώς μπορεί να αξιολογήσει την αποδοτικότητα των διαδικασιών logistics, να εντοπίσει σημεία που χρειάζονται βελτίωση καθώς και να εστιάσει στην ικανοποίηση και την διατήρηση των πελατών ((Hübner et al., 2016).

Στη συνέχεια, εξετάστηκε η μεταβλητή state_status, η οποία αναφέρεται στην γεωγραφική κατανομή των παραγγελιών ανάλογα με την πολιτεία κατοικίας του πελατολογίου. Το διάγραμμα 2 πραγματεύεται τις δέκα πολιτείες που εμφανίζουν το μεγαλύτερο πλήθος

παραγγελιών. Το συγκεκριμένο σχεδιάγραμμα αποκαλύπτει ότι η αγοραστική δραστηριότητα επικεντρώνεται σε συγκεκριμένες πολιτείες το οποίο μπορεί να αποδοθεί σε πληθυσμιακούς, οικονομικούς και κοινωνικούς παράγοντες. Ενδεικτικά, η πολιτεία SP (São Paulo), παρουσιάζει τον μεγαλύτερο αριθμό παραγγελιών. Το εύρημα αυτό, πιθανόν να οφείλεται στο γεγονός ότι η πολιτεία αποτελεί το μεγαλύτερο οικονομικό και εμπορικό κέντρο της Βραζιλίας, λόγω της υψηλής πληθυσμιακής συγκέντρωσης με ανεπτυγμένη αγορά. Αντιθέτως, οι πολιτείας GO (Goiás) και ES (Espírito Santo) παρουσιάζουν μικρότερο αριθμό παραγγελιών εξαιτίας του μικρότερου πληθυσμού. Το στοιχείο ζήτησης γεωγραφικής κατανομής είναι βασικό εργαλείο για τις επιχειρήσεις ηλεκτρονικού εμπορίου, καθώς μπορεί να χρησιμοποιηθεί στην ανάπτυξη στρατηγικού σχεδιασμού δικτύων διανομής και στη βελτιστοποίηση των διαδικασιών logistics, αλλά και στον σχεδιασμό στοχευμένων προωθητικών ενεργειών ((Offline Showrooms in Omnichannel Retail: Demand and Operational Benefits | Management Science, n.d.).

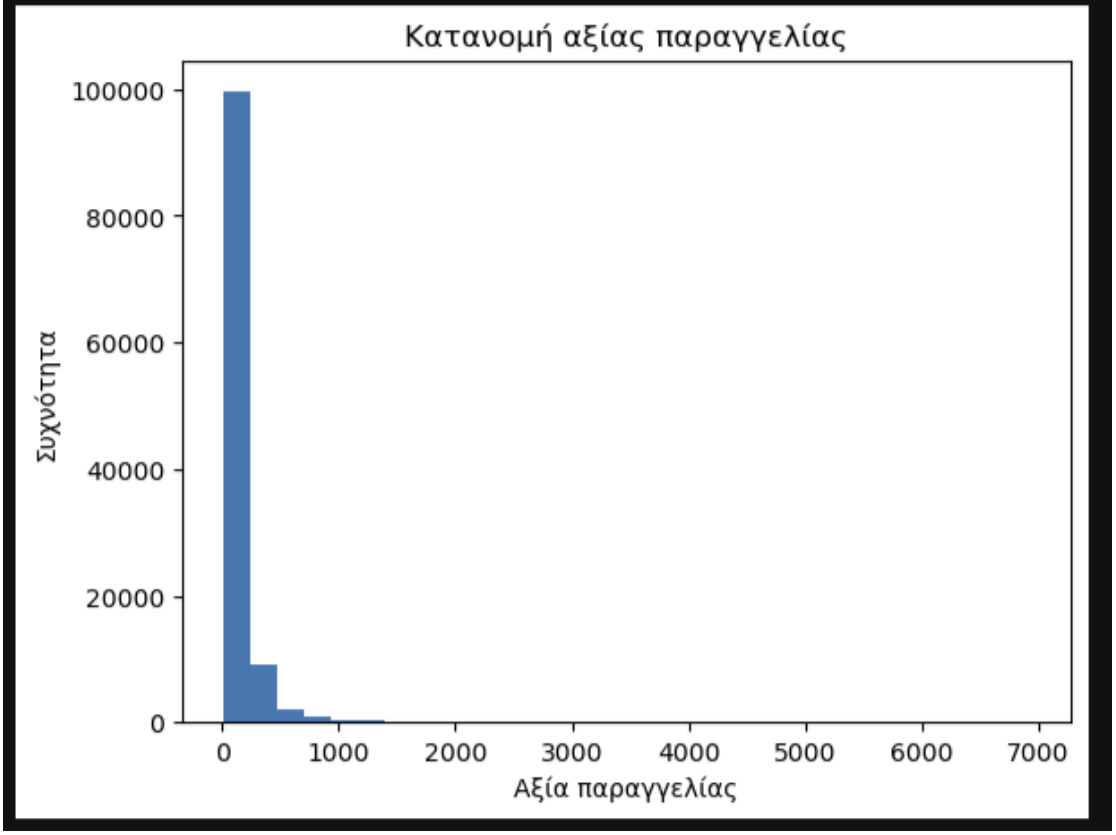
```python
# Οπτικοποίηση των 10 πολιτειών με το μεγαλύτερο πλήθος παραγγελιών
state_counts = full_data["customer_state"].value_counts().head(10)

state_counts.plot(kind="bar")
plt.title("Παραγγελίες ανά πολιτεία (Top 10)")
plt.xlabel("Πολιτεία")
plt.ylabel("Πλήθος παραγγελιών")
plt.show()
```



Διάγραμμα 2 Παραγγελίες ανά πολιτεία

Το τρίτο διάγραμμα, παρουσιάζει την κατανομή αξίας παραγγελίας η οποία προκύπτει από το άθροισμα τιμής του προϊόντος και του κόστους μεταφοράς. Συγκεκριμένα, γίνεται αντιληπτό ότι η πλειονότητα των συναλλαγών επικεντρώνεται σε χαμηλή έως μεσαία αξία προϊόντων, όπως προϊόντα καθημερινή χρήσης (είδη οικιακής κατανάλωσης ή αξεσουάρ). Αντίθετα, παραγγελίες μεγάλης αξίας (ηλεκτρονικές συσκευές ή έπιπλα) εμφανίζονται σπανιότερα, γεγονός που αποδίδει την δεξία ασυμμετρία της κατανομής. Το στοιχείο αυτό συνδέεται με την ανάλυση του προήγουμενου σχήματος καθώς εικάζεται πως το γεγονός αυτό εξαρτάται από οικονομικούς και κοινωνικούς παράγοντες που διαφοροποιούνται ανά γεωγραφική περιοχή.

```
# Ιστόγραμμα κατανομής αξίας παραγγελίας
plt.hist(full_data["order_value"], bins=30)
plt.title("Κατανομή αξίας παραγγελίας")
plt.xlabel("Αξία παραγγελίας")
plt.ylabel("Συχνότητα")
plt.show()
```



Διάγραμμα 3 Κατανομή αξίας παραγγελίας

Συνολικά, τα παραπάνω σχεδιαγράμματα συνέβαλαν στην κατανόηση της αγοραστικής συμπεριφοράς των καταναλωτών και αποτέλεσαν υποστηρικτική βάση για τα επόμενα στάδια της μελέτης, όπως είναι η προεπεξεργασία δεδομένων και η εφαρμογή τεχνικών μηχανικής μάθησης.

4.4 Προεπεξεργασία Δεδομένων και Feature Engineering

4.4.1 Επιλογή και καθαρισμός μεταβλητών

Το στάδιο της προεπεξεργασίας δεδομένων αποτελεί την επιλογή και τον καθαρισμό των μεταβλητών που κρίθηκαν απαραίτητες για την εφαρμογή τεχνικών μηχανικής μάθησης. Τα αρχικά σύνολα δεδομένων περιλαμβάνουν μεγάλο αριθμό μεταβλητών. Δεν είναι όλες οι μεταβλητές χρήσιμες ή σχετικές με το αντικείμενο της ανάλυσης. Για το λόγο αυτό, γίνεται διαχωρισμός και επιλέγονται οι μεταβλητές που συμβάλουν στη μοντελοποίηση και την εξαγωγή συμπερασμάτων.

Εφόσον, εξετάστηκαν οι διαθέσιμες μεταβλητές του τελικού ενοποιημένου συνόλου, επιλέχθηκαν εκείνες οι οποίες αναφέρονται στην διαδικασία ηλεκτρονικής συναλλαγής και συμπεριφοράς καταναλωτή. Κάποιες μεταβλητές, όπως οι μοναδικοί κωδικοί παραγγελιών ή πελατών, δεν χρησιμοποιήθηκαν ως είσοδοι στα μοντέλα μάθησης, διότι δεν αποτελούν πληροφορία πρόβλεψης και δεν εξυπηρετούν στην ερμηνεία της αγοραστικής συμπεριφοράς (Kuhn & Johnson, 2019).

Στη συνέχεια, πραγματοποιήθηκε ο έλεγχος ποιότητας των επιλεγμένων μεταβλητών. Ειδικότερα, οι μεταβλητές που εμφάνιζαν προβλήματα, όπως ελλιπείς τιμές, εξετάστηκαν μέσω της εντολής full_data.isnull().sum(), ώστε να εξεταστεί η χρησιμότητα τους και να αξιολογηθεί ο βαθμός στον οποίο η απουσία τιμών θα μπορούσε να επηρεάσει τα αποτελέσματα της ανάλυσης. Σε περίπτωση που οι ελλιπείς τιμές αφορούν μικρό αριθμό εγγραφών και δεν επηρεάζουν το σύνολο των δεδομένων, οι εγγραφές αυτές διατηρούνται. Ωστόσο, οι μεταβλητές που επηρεάζουν μεγάλο ποσοστό των δεδομένων ή η πληροφορία που προσφέρουν είναι περιορισμένη, δεν θα συμπεριλαμβάνονται στο σύνολο της ανάλυσης (García et al., 2015).

Επιπλέον, πραγματοποιήθηκε έλεγχος σχετικά με την ύπαρξη ακραίων ή μη έγκυρων τιμών οι οποίες θα μπορούσαν να επηρεάσουν αρνητικά την απόδοση των μοντέλων μηχανικής μάθησης και να εξάγουν λανθασμένα συμπεράσματα, καθώς τα μοντέλα θα εκπαιδεύονται με λάθος παραδείγματα. Παραδείγματα μη έγκυρων ή ακραίων τιμών αποτελούν προϊόντα με τιμή αξίας 0 € ή παραγγελία πολύ υψηλής αξίας 1.000.000 €, όταν οι υπόλοιπες παραγγελίες κυμαίνονται μεταξύ 20 και 200 €. Η συγκεκριμένη διαδικασία πραγματοποιήθηκε για να προσφέρει ένα αξιόπιστο σύνολο δεδομένων και να αποφύγει τους θορύβους, δίνοντας την δυνατότητα στον αλγόριθμο να εντοπίσει ουσιαστικά πρότυπα.

Όλα τα παραπάνω σχετικά με τον καθαρισμό των μεταβλητών, διασφάλισαν την καταλληλόλητα των δεδομένων, βελτίωσαν την κατανόηση τους και θεωρήθηκαν αντιπροσωπευτικά για την αγοραστική συμπεριφορά του ηλεκτρονικού εμπορίου.

4.4.2 Δημιουργία νέων χαρακτηριστικών (Feature Engineering)

Για να βελτιωθεί η πληροφορία που παρέχεται στα μοντέλα μηχανικής μάθησης, εφαρμόστηκε η διαδικασία δημιουργίας νέων χαρακτηριστικών (feature engineering), η οποία βασίστηκε στον μετασχηματισμό υπαρκτών μεταβλητών με σκοπό την ανάπτυξη νέων χαρακτηριστικών με αυξημένη ερμηνευτική και προβλεπτική αξία. Η διαδικασία αυτή αποτελεί καθοριστικό ρόλο για την απόδοση και την αξιοπιστία των μοντέλων μηχανικής μάθησης, διότι η ποιότητα των χαρακτηριστικών προσφέρει στον αλγόριθμο την ικανότητα να εντοπίζει πρότυπα στα δεδομένα (Zheng & Casari, 2018).

Αρχικά, δημιουργήθηκε η μεταβλητή order_value από το άθροισμα της τιμής των προϊόντων και του κόστους μεταφοράς. Η συγκεκριμένη μεταβλητή συνεισφέρει στην άμεση πληροφόρηση για την χρηματική αξία των συναλλαγών και βοηθά στην ανάλυση της αγοραστικής συμπεριφοράς, ενώ παράλληλα επιτρέπει την ποσοτική αποτύπωση της καταναλωτικής δαπάνης.

```
# Υπολογισμός συνολικής αξίας παραγγελίας ως άθροισμα
# της τιμής προϊόντος και του κόστους μεταφοράς
full_data["order_value"] = full_data["price"] + full_data["freight_value"]
```

Στη συνέχεια, μετασχηματίστηκαν οι χρονικές πληροφορίες των παραγγελιών για να υπολογιστεί η απόκλιση της πραγματικής ημερομηνίας παράδοσης από την εκτιμώμενη. Η μεταβλητή αυτή ονομάστηκε delivery_delay_days και προέκυψε αφαιρώντας την εκτιμώμενη ημερομηνία παράδοσης από την παραγματική ημερομηνία. Οι ημερομηνίες μετατράπηκαν σε αριθμητική μορφή, ώστε να είναι κατάλληλες για τους αλγορίθμους της μηχανικής μάθησης και να παρέχεται η δυνατότητα ανάλυσης της απόδοσης των διαδικασιών logistics σε ποσοτικό επίπεδο. Σε περίπτωση θετικής τιμής της μεταβλητής, τα προϊόντα παραδόθηκαν με χρονοκαθυστέρηση, ενώ σε περίπτωση αρνητικού αποτελέσματος υποδηλώνεται πρόωρη παράδοση. Όπως αποτυπώνεται και στο πίνακα παρακάτω, από τις πρώτες ενδεικτικές εγγραφές που χρησιμοποιήθηκαν για ανάλυση και μοντελοποίηση, η στήλη με την μεταβλητή delivery_delay_days αποδικνύει πως όλες οι παραγγελίες παραδόθηκαν στην ώρα τους (αρνητική τιμή).

```
# Μετατροπή των χρονικών μεταβλητών σε τύπο datetime
full_data["order_delivered_customer_date"] = pd.to_datetime(
    full_data["order_delivered_customer_date"]
)
full_data["order_estimated_delivery_date"] = pd.to_datetime(
    full_data["order_estimated_delivery_date"]
)

# Υπολογισμός ημερών καθυστέρησης παράδοσης ως η διαφορά
# μεταξύ πραγματικής και εκτιμώμενης ημερομηνίας παράδοσης
full_data["delivery_delay_days"] = (
    full_data["order_delivered_customer_date"]
    - full_data["order_estimated_delivery_date"]
).dt.days
```

```
# Επιλογή βασικών μεταβλητών για περαιτέρω ανάλυση και μοντελοποίηση
analysis_data = full_data[
    [
        "customer_state",
        "order_value",
        "delivered_binary",
        "delivery_delay_days"
    ]
]
analysis_data.head()
```

	customer_state	order_value	delivered_binary	delivery_delay_days
0	SP	38.71	1	-8.0
1	BA	141.46	1	-6.0
2	GO	179.12	1	-18.0
3	RN	72.20	1	-13.0
4	SP	28.62	1	-10.0

Τέλος, δημιουργήθηκε η δυαδική μεταβλητή delivered_binary η οποία λαμβάνει τιμές 0 ή 1 και αποτυπώνει την ολοκλήρωση της παραγγελίες. Συγκεκριμένα, η τιμή 1 αντιστοιχεί στις παραγγελίες που ολοκληρώθηκαν και παραδόθηκαν επιτυχώς στον πελάτη, ενώ η τιμή 0 αντιστοιχεί στις παραγγελίες που δεν παραδόθηκαν. Η μεταβλητή αυτή, διαχωρίζει τις παραγγελίες σε επιτυχείς και μη επιτυχείς και μπορεί να χρησιμοποιηθεί ως μεταβλητή-στόχος σε μοντέλα επιβλεπόμενης μάθησης (Guyon & Elisseeff, n.d.).

```python
# Δημιουργία δυαδικής μεταβλητής για την κατάσταση παράδοσης
# 1: παραγγελία παραδόθηκε, 0: διαφορετική κατάσταση
full_data["delivered_binary"] = full_data["order_status"].apply(
    lambda x: 1 if x == "delivered" else 0
)

# Κατανομή παραγγελιών ως προς την κατάσταση παράδοσης
full_data["delivered_binary"].value_counts()
```

```
delivered_binary
1    110197
0      2453
```

4.5.1 Ορισμός μεταβλητής-στόχου και χαρακτηριστικών εισόδου

Στην επιβλεπόμενη μάθηση βασικό εργαλείο της διαδικασίας αποτελεί η μεταβλητή στόχος (*target variable*) και τα χαρακτηριστικά εισόδου (*input features*) που χρησιμοποιούνται για την εκπαίδευση των μοντέλων. Ειδικότερα, η μεταβλητή στόχος επικεντρωνεται στο φαινόμενο που θέλουμε να προβλέψουμε, για παράδειγμα εάν μία παραγγελία θα παραδοθεί, ενώ τα χαρακτηριστικά εισόδου εστιάζουν στα δεδομένα που έχουμε στη διάθεση μας σχετικά με την παραγγελία, όπως τι κόστισε ή πόσες μέρες έκανε για να παραδοθεί (Hastie et al., 2009).

Στην παρούσα ανάλυση, ως μεταβλητή στόχο έχουμε ορίσει την δυαδική μεταβλητή delivered_binary η οποία παρουσιάζει την ολοκλήρωση της παραγγελίας. Η συγκεκριμένη μεταβλητή επιτρέπει την ανάπτυξη ενός προβλήματος δυαδικής ταξινόμησης το οποίο έχει ως στόχο να προβλέψει εάν η παραγγελία ήταν επιτυχής λαμβάνοντας την τιμή 1 ή δεν ολοκληρώθηκε λαμβάνοντα την τιμή 0.

Επιπλέον, τα χαρακτηριστικά εισόδου εκπροσωπούνται από τις μεταβλητές που περιγράφουν την οικονομική και την λειτουργική διάσταση του ηλεκτρονικού εμπορίου. Ειδικότερα, οι μεταβλητές που χρησιμοποιήθηκαν είναι η συνολική αξία παραγγελίας (order_value), ο χρόνος παράδοσης σε ημέρες (delivery_delay_days), καθώς και γεωγραφικά και λειτουργικά χαρακτηριστικά τα οποία σχετίζονται με τον πελάτη και τη διαδικασία εκτέλεσης της παραγγελίας. Οι συγκεκριμένες μεταβλητές κρίθηκαν απαραίτητες, καθώς εξυπηρετούν την απόδοση των διαδικασιών logistics και την εμπειρία του πελάτη.

Ο διαχωρισμός σε μεταβλητή στόχο και χαρακτηριστικά εισόδου είναι απαραίτητος διότι θα συμβάλει στη σωστή εκπαίδευση και στην αξιολόγηση των μοντέλων ταξινόμησης που θα αναπτυχθούν σε επόμενες ενότητες εξασφαλίζοντας αξιόπιστα και συγκρίσιμα αποτελέσματα (Müller & Guido, 2016).

4.5.2 Περιγραφή μοντέλου ταξινόμησης - Logistic Regression

Για την αντιμετώπιση του προβλήματος δυαδικής ταξινόμησης που παρουσιάστηκε σε προηγούμενες ενότητες, χρησιμοποιήθηκε προσέγγιση επιβλεπόμενης μάθησης. Ο σκοπός της ανάλυσης είναι η πρόβλεψη της μεταβλητής delivered_binary η οποία καταγράφει εάν μία παραγγελία ολοκληρώθηκε ή όχι, βασιζόμενη σε χαρακτηριστικά εισόδου διαφορετικών πτυχών του ηλεκτρονικού εμπορίου (Alpaydin, 2020).

Στην παρούσα διπλωματική εργασία, επιλέχθηκε το μοντέλο Logistic Regression, το οποίο αποτελεί μία από τις πιο διαδεδομένες και αξιόπιστες μεθόδους για πρόβλήματα δυαδικής ταξινόμησης. Ειδικότερα, το μοντέλο εκτιμά ότι το παρατηρούμενο γεγονός ανήκει σε μία από τις δύο κατηγορίες, το οποίο το καθιστά ιδιαίτερα κατάλληλο σε αναλύσεις όπου η μεταβλητή λαμβάνει δυαδικές τιμές.

Η συγκεκριμένη μέθοδος επιλέχθηκε για διάφορους λόγους. Αρχικά, χαρακτηρίζεται από απλότητα και υψηλή ερμηνευσιμότητα, διότι πραγματεύεται κατανόηση της επίδράσης των επιμέρους χαρακτηριστιών εισόδου στην τελική πρόβλεψη. Επιπλέον, δεν χρειάζεται εκτενή προεπεξεργασία των δεδομένων και παρουσιάζει υψηλά ποσοστά απόδοσης σε σύνολα δεδομένων μεσαίου μεγέθους, όπως το σύνολο δεδομένων Olist (Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow - Aurélien Géron - Βιβλία Google, n.d.).

Η Logistic Regression, χρησιμοποιείται ευρέως σε προβλήματα πρόβλεψης που σχετίζονται με την αγοραστική συμπεριφορά των καταναλωτών και την αξιολόγηση επιχειρησιακών διαδικασιών στο ηλεκτρονικό εμπόριο. Στο πλαίσιο της εργασίας, η συμβολή της κρίνεται ιδιαίτερα σημαντική καθώς επιτρέπει την εξαγωγή συμπερασμάτων σχετικά με τους παράγοντες που επηρεάζουν την ολοκλήρωση ή μη των παραγγελιών. Παράλληλα, αποτελεί αξιόπιστο σημείο αναφοράς για την σύγκριση μελλοντικών ή σύνθετων προσεγγίσεων μηχανικής μάθησης (Jr et al., 2013).

4.5.3 Εκπαίδευση του μοντέλου και αξιολόγηση απόδοσης

Στο στάδιο αυτό πραγματοποιήθηκε η εκπαίδευση του μοντέλου Logistic Regression, με σκοπό την πρόβλεψη της μεταβλητής-στόχου. Αρχικά ορίστηκαν τα χαρακτηριστικά εισόδου (featutes) και η μεταβλητή- στόχος (target variable). Συγκεκριμένα ως χαρακτηριστικά εισόδου χρησιμοποιήθηκαν οι μεταβλητές order_value και delivery_delay_days, οι οποίες χαρακτηρίστηκαν με την μεταβλητή x, ενώ για την μεταβλητή - στόχο χρησιμοποιήθηκε η δυαδική μεταβλητή delivered_binary η οποία χαρακτηρίστηκε ως y.

Για την καλύτερη απόδοση του μοντέλου, το σύνολο δεδομένων διαμοιράστηκε σε σύνολο εκπαίδευσης και σύνολο ελέγχου (training set και test set αντίστοιχα) με την χρήση της συνάρτησης train_test_split. Ο διαμοιρασμός αυτός επιτρέπει την εκτίμηση απόδοσης του μοντέλου σε δεδομένα που δεν χρησιμοποιήθηκαν κατά την διαδιακασία της εκπαίδευσης, δηλαδή κάνει πρόβλεψη σε παραγγελίες που δεν έχει ξανά δει, προσομοιώνοντας του την ικανότητα να γενικεύει σε νέες παρατηρήσεις (Kohavi, 2001).

Συγκεκριμένα, στην παρούσα ανάλυση το σύνολο δεδομένων διαχωρίζεται σε ποσοστό 80% για εκπαίδευση και 20% για έλεγχο. Η παράμετρος random_state=42 δηλώνει ότι ο

διαχωρισμός των δεδομένων είναι αναπαραγώγιμος, καθώς η διαδικασία εκπαίδευσης και αξιολόγησης βασίζεται σε σταθερές παραμέτρους και επιτρέπει την επανάληψη της διαδικασίας με τα ίδια αποτελέσματα.

Ωστόσο, το μοντέλο Logistic Regression, εκπαιδεύτηκε με τη μέθοδο fit. Στην εκπαίδευση, έμαθε τη σχέση μεταξύ των χαρακτηριστικών εισόδου (*X_train*) και των πραγματικών τιμών της μεταβλητής-στόχου (y_train), ώστε να είναι σε θέση μελλοντικά να παράγει προβλέψεις για νέες παρατηρήσεις. Εφόσον ολοκληρώθηκε και η εκπαίδευση fit, το μοντέλο εκπαιδεύτηκε στο σύνολο εκπαιδειδευσης (X_train, y_train) και στη συνέχεια εφαρμόστηκε στο σύνολο ελέγχου (X_test), δημιουργώντας τις προβλέψεις (y_pred), οι οποίες χρησιμοποιούνται για να αξιολογηθεί η απόδοση του. Η διαδικασία αυτή, θεωρείται βασικό στάδιο στην ανάπτυξη μοντέλων, διότι επιτρέπει τον έλεγχο της ακρίβειας και της γενίκευσης του μοντέλου σε άγνωστα δεδομένα (Brownlee, 2022).

```
analysis_data = full_data[
    ["order_value", "delivery_delay_days", "delivered_binary"]
].dropna()

# Επιλογή χαρακτηριστικών (features) και μεταβλητής στόχου
X = analysis_data[["order_value", "delivery_delay_days"]]
y = analysis_data["delivered_binary"]

# Διαχωρισμός δεδομένων σε σύνολα εκπαίδευσης και ελέγχου
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Εκπαίδευση μοντέλου λογιστικής παλινδρόμησης
model = LogisticRegression()
model.fit(X_train, y_train)
```

4.5.4 Αξιολόγηση της απόδοσης του μοντέλου Logistic Regression

Στην παρούσα ενότητα αξιολογείται η απόδοση του μοντέλου Logistic Regression, το οποίο εκπαιδεύτηκε στο σύνολο εκπαίδευσης (X_train, y_train) με στόχο την πρόβλεψη της μεταβλητής στόχο delivered_binary σε δεδομένα που δε χρησιμοποιήθηκαν κατά την εκπαίδευση, και σχετίζονται με το ποια παραγγελία ολοκληρώθηκε επιτυχώς (τιμή 1) ή όχι (τιμή 0).

Αρχικά, το εκπαιδευμένο μοντέλο χρησιμοποιήθηκε στο σύνολο ελέγχου (X_test) και δημιουργήσε τις προβλεπόμενες κλάσεις y_pred, το οποίο υποδηλώνει εάν οι παραγγελίες προβλέπονται ως παραδομένες ή όχι.

```
y_pred = model.predict(X_test)
```

Στη συνέχεια, για την ποσοτική αξιολόγηση της απόδοσης του μοντέλου χρησιμοποιήθηκε η συνάρτηση classification_report, η οποία υπολογίζει τις μετρικές αξιολόγησης precision, recall και F1-score για κάθε κλάση, καθώς και τη συνολική ακρίβεια (accuracy) του μοντέλου (‘(PDF) Evaluation’, 2025).

Τα αποτελέσματα έδειξαν υψηλό ποσοστό συνολικής ακρίβειας (accuracy περίπου 98%). Αυτό οφείλεται στην επιτυχή πρόβλεψη της πλειονότητας των παραγγελιών οι οποίες ολοκληρώθηκαν επιτυχώς (κλάση 1 - delivered), για τις οποίες οι δεικτες precision και recall είναι ιδιαίτερα υψηλοί. Ωστόσο, στην κλάση 0 (not delivered) οι μετρικές precision και recall είναι μηδενικές. Αυτό σημαίνει ότι το μοντέλο δεν κατάφερε να προβλέψει καμία παραγγελία ως μη παραδοθείσα.

Το γεγονός αυτό σχετίζεται άμεσα με την έντονη ανισορροπία των κλάσεων στο σύνολο των δεδομένων, όπου ο αριθμός των παραγγελιών που δεν ολοκληρώθηκαν είναι πολύ μικρότερος από αυτό που παραδόθηκαν. Σε αυτές τις περιπτώσεις, τα μοντέλα ταξινόμησης έχουν την τάση να ευνοούν την πλειοψηφική κλάση και να οδηγούνται σε παραπλανητικές υψηλές τιμές ακρίβειας (He & Garcia, 2009).

```
# Αξιολόγηση απόδοσης μοντέλου
from sklearn.metrics import accuracy_score, classification_report
y_pred = model.predict(X_test)
# Αναλυτική αξιολόγηση της απόδοσης του μοντέλου ανά κατηγορία
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       492
           1       0.98      1.00      0.99     22038

    accuracy                           0.98     22530
   macro avg       0.49      0.50      0.49     22530
weighted avg       0.96      0.98      0.97     22530
```

Για την καλύτερη ερμηνεία των αποτελεσμάτων, υπολογίστηκε και ο πίνακας σύγχυσης (*confusion matrix*), ο οποίος απεικονίζει τη σχέση μεταξύ πραγματικών και προβλεπόμενων κλάσεων. Ο πίνακας επιβεβαιώνει ότι όλες οι παρατηρήσεις ταξινομήθηκαν ως delivered, ενώ οι περιπτώσεις not delivered δεν εντοπίστηκαν. Από το πόρισμα αυτό οδηγούμαστε στο συμπέρασμα ότι παρ'όλο που η συνολική ακρίβεια είναι υψηλή, το μοντέλο δεν είναι αποτελεσματικό στην μειοψηφική κλάση.

```
# Υπολογισμός της μήτρας σύγχυσης (confusion matrix) για το μοντέλο ταξινόμησης
# Το αποτέλεσμα δείχνει ότι το μοντέλο ταξινομεί όλες τις παρατηρήσεις
# ως "delivered" (κλάση 1), χωρίς να προβλέπει καμία μη παραδομένη παραγγελία (κλάση 0),
# λόγω της έντονης ανισορροπίας μεταξύ των κλάσεων στο σύνολο δεδομένων
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, y_pred)
```

```
array([[    0,   492],
       [    0, 22038]])
```

Τέλος, παρότι το accuracy είναι υψηλό, περίπου 0.98 , θα πρέπει να ερμηνευτεί με προσοχή διότι η ανισορροπία στις τιμές των παραδομένων και μη παραδομένων παραγγελιών μπορεί να οδηγήσει σε υπερεκτίμηση απόδοσης, ενώ το μοντέλο δεν μπορεί να εντοπίσει την μειοψηφική και κρίσιμη κλάση των μη παραδοθέντων. Στην πραγματικότητα, το μοντέλο εφόσον εκπαιδεύτηκε σε δεδομένα όπου περίπου 22.000 παραγγελίες παραδόθηκαν και περίπου 500 δεν παραδόθηκαν, έχει κάνει καλή πρόβεψη καθώς είναι πιθανότερο οι παραγγελίες να παραδοθούν. Παρόλα αυτά, το μοντέλο δεν έχει λάβει υπόψη τις μη παραδομένες παραγγελίες και, στη πράξη, αυτό μπορεί να δημιουργήσει προβλήματα, περιόριζοντας τη χρησιμότητα του για την υποστήριξη αποφάσεων και την ανάπτυξη στρατηγικών logistics (Fawcett & Provost, 1997).

4.5.5 Αντιμετώπιση της ανισορροπίας κλάσεων και βελτίωση του μοντέλου

Στην προηγούμενη ενότητα διαπιστώθηκε ότι, παρά το γεγονός της υψηλής συνολικής ακρίβειας του μοντέλου Logistic Regression, η απόδοση του ως προς την μειοψηφική κλάση, στην προκειμένη οι παραγγελίες που δεν ολοκληρώθηκαν, δεν ήταν ικανοποιητική. Το φαινόμενο αυτό σχετίζεται με την ανισορροπία κλάσεων (class imbalance) που χαρακτηρίζει το σύνολο των δεδομένων, καθώς η πλειονότητα των δεδομένων αφορά τις παραγγελίες που παραδόθηκαν επιτυχώς.

Η ανισορρόπια κλάσεων συναντάται συχνά σε εφαρμογές του ηλεκτρονικού εμπορίου όπου οι «αρνητικές» παρατηρήσεις εμφανίζονται σπανιότερα από τις «θετικές». Σε αυτές τις περιπτώσεις τα μοντέλα ταξινόμησης ευνοούν την επικράτουσα κλάση, στην προκειμένη την παράδοση της παραγγελίας, επιτυγχάνοντας υψηλή ακρίβεια αλλά ταυτόχρονα χαμηλή ικανότητα εντοπισμού των κρίσιμων παρατηρήσεων (Fiore et al., 2019; Krawczyk, 2016).

Μία απλή και αποτελεσματική προσέγγιση για τέτοιου είδους προβλήματα είναι η χρήση των σταθμισμένων κλάσεων (class weighting) κατά την διαδικασία εκπαίδευσης του μοντέλου. Η μέθοδος αυτή έχει την ικανότητα να δίνει μεγαλύτερη «βαρύτητα» στα σφάλματα που δημιουργούνται σχετικά με την μειοψηφική κλάση, «ενθαρύνοντας» το μοντέλο να την λάβει υπόψη και να την αντιμετωπίσει καλύτερα. Η παράμετρος που χρησιμοποιείται για αυτή τη μέθοδο στη Logistic Regression είναι η class_weight = 'balanced' (Buda et al., 2018).

```python
# Εκπαίδευση λογιστικής παλινδρόμησης με εξισορρόπηση κλάσεων (class_weight='balanced')
# Στόχος είναι η βελτίωση της πρόβλεψης της μειοψηφικής κατηγορίας (μη παραδομένες παραγγελίες),
# ακόμη και αν αυτό οδηγήσει σε μείωση της συνολικής ακρίβειας
from sklearn.linear_model import LogisticRegression

model_balanced = LogisticRegression(class_weight='balanced')
model_balanced.fit(X_train, y_train)

y_pred_balanced = model_balanced.predict(X_test)
```

Μετά την ολοκλήρωση του συγκεκριμένου βήματος, το μοντέλο αξιολογήθηκε για την απόδοση του μέσω της συνάρτησης classification_report και των ίδιων μετρικών όπως στην προηγούμενη ενότητα προκειμένου να εκτιμηθεί η επίδραση της παραμέτρου class_weight = 'balanced' στην απόδοση του μοντέλου. Τα αποτελέσματα εμφάνισαν μεγάλη βελτίωση σχετικά με την ικανότητα εντοπισμού των μη παραδομένων παραγγελιών με το recall να λαμβάνει τιμή 0,28 σε αντίθεση με την προηγούμενη εφαρμογή που ήταν μηδέν. Το γεγονός αυτό αποδικνύει ότι το μοντέλο είναι σε θέση πλέον να αναγνωρίσει μέρος των προβληματικών περιπτώσεων.

Επιπλέον, η συνολική τιμή της ακρίβειας μειώθηκε σε 0,73 από 0,98 που είχε καταγραφεί προηγουμένως, εύρημα που είναι αναμενόμενο σε περιπτώσεις εξισορρόπησης κλάσεων.

Όσον αφορά την κλάση 1 (delivered), το μοντέλο συνεχίζει να διατηρεί υψηλή ακρίβεια 0,98, ωστόσο η μετρική τιμή recall μειώθηκε σε 0,74 γεγονός που αποδυκνύει ότι κάποιες από τις παραγγελίες ταξινομούνται πλέον λανθασμενα ως μη παραδομένες.

```
# Αξιολόγηση του εξισορροπημένου μοντέλου ταξινόμησης
# Η χρήση εξισορρόπησης κλάσεων οδηγεί σε καλύτερη ανίχνευση μη παραδομένων παραγγελιών,
# ενώ η συνολική ακρίβεια μειώνεται, λόγω της αυξημένης έμφασης στη μειοψηφική κατηγορία
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred_balanced))
```

```
              precision    recall  f1-score   support

           0       0.02      0.28      0.04       492
           1       0.98      0.74      0.84     22038

    accuracy                           0.73     22530
   macro avg       0.50      0.51      0.44     22530
weighted avg       0.96      0.73      0.83     22530
```

Το γενικό συμπέρασμα είναι ότι η χρήση σταθμισμένων κλάσεων οδηγεί συγκριτικά σε μία πιο ισορροπημένη και επιχειρησιακά χρήσιμη συμπεριφορά του μοντέλου. Παρά το γεγονός ότι η συνολική ακρίβεια μειώνεται, το μοντέλο έχει πλέον τη δυνατότητα να εντοπίζει προβληματικές παραγγελίες, οι οποίες είναι επιχειρησιακά πιο κρίσιμες για τη βελτίωση των διαδικασιών logistics.

4.6 Ανάλυση Συσχετίσεων Αγορών (Market Basket Analysis)

4.6.1 Στόχος και σκοπός της Ανάλυσης Συσχετίσεων

Μετά την ολοκλήρωση της ανάλυσης ταξινόμησης η οποία εστίαζε στην πρόβλεψη της επιτυχούς ολοκλήρωσης παραγγελιών μέσω της δυαδικής μεταβλητής, η παρούσα ενότητα επικεντρώνεται στην ανάλυση συσχετίσεων αγορών (Association Rule Mining). Πρόκειται για μια τεχνική της μη επιβλεπόμενης μάθησης η οποία στοχεύει στον εντοπισμό επαναλαμβανόμενων αγοραστικών συνδυασμών προϊόντων οι οποίοι εμφανίζονται στις ίδιες συναλλαγές (Agrawal et al., 1993).

Ο στόχος της συγκεκριμένης μεθόδους είναι να απαντάει σε ερωτήματα του τύπου «Ποια προϊόντα τείνουν να αγοράζονται μαζί από τους καταναλωτές;». Τα αποτελέσματα συνήθως έχουν την μορφή κανόνων συσχέτισης όπως «Αν ένας πελάτης αγοράσει το προϊόν Α, τότε είναι πιθανό να αγοράσει και το προϊόν Β», προσφέροντας χρήσιμες πληροφορίες σχετικά με την καταναλωτική συμπεριφορά .

Η ανάλυση συσχετίσεων είναι σημαντικό εργαλείο στο πλαίσιο του ηλεκτρονικού εμπορίου καθώς χρησιμοποιεί στρατηγικές όπως διασταυρούμενων πωλήσεων (cross-selling), προτάσεων προϊόντων (recommendation systems) και ομαδοποιημένων προσφορών (bundle offers). Μέσω των αποτελεσμάτων της οι επιχειρήσεις μπορούν να αυξήσουν την αξία καλαθιού των καταναλωτών, προτείνοντας προϊόντα που σχετίζονται με τις αγοραστικές τους επιλογές (MobasherBamshad et al., 2000).

Σε αντίθεση με τις τεχνικές που χρησιμοποιεί η επιβλεπόμενη μάθηση, η ανάλυση συσχετίσεων δεν χρειάζεται μεταβλητή στόχο. Ωστόσο, επικεντρώνεται στην δομή των συναλλαγών και στην παρατήρηση των αγοραστικών προτύπων που προκύπτουν από τα δεδομένα. Αυτός είναι και ο λόγος που η συγκεκριμένη τεχνική δεν αντικαθιστά την ταξινόμηση, αντιθέτως θεωρείται συμπληρωματική καθώς προσφέρει πληροφορίες σχετικά με τα αγοραστικά πρότυπα των καταναλωτών, πέρα από την πρόβλεψη ολοκλήρωσης παραγγελίας.

Στην παρούσα διπλωματική εργασία, η ανάλυση συσχετίσεων χρησιμοποιείται στο σύνολο δεδομένων Olist, με στόχο την ανακάλυψη προτύπων συνδυασμών προϊόντων και την εξαγωγή κανόνων συσχέτισης που μπορούν να χρησιμοποιηθούν ως εργαλείο για την υποστήριξη αποφάσεων marketing και προτάσεων προϊόντων.

4.6.2 Προετοιμασία δεδομένων συναλλαγών

Για την εφαρμογή της ανάλυσης συσχετίσεων, τα δεδομένα οργανώθηκαν σε συναλλαγές ώστε κάθε συναλλαγή να αντιπροσωπεύει το καλάθι αγορών το οποίο περιλαμβάνει τα προϊόντα που αγοράστηκαν μαζί. Ειδικότερα, στην παρούσα εργασία, ως συναλλαγή θα ορίζεται η κάθε παραγγελία (order_id), ενώ τα προϊόντα που περιλαμβάνονται σε αυτή θα αντιπροσωπεύουν το «καλάθι αγορών». Αυτή η αναπαράσταση εξυπηρετεί στον εντοπισμό επαναλαμβανόμενων συνδυασμών προϊόντων που αγοράζονται μαζί (Agrawal et al., 1993).

Για το γεγονός αυτό, χρησιμοποιήθηκαν οι πίνακες orders και order_items του συνόλου δεδομένων Olist. Όσον αφορά τον πρώτο πίνακα αναφέρεται στην ταυτότητα κάθε παραγγελίας ενώ ο δεύτερος περιλαμβάνει τις πληροφορίες των προϊόντων που βρίσκονται σε κάθε παραγγελία. Ο συνδυασμός των δύο πινάκων επιτρέπει τη σύνδεση των προϊόντων με κάθε παραγγελία.

Αρχικά, φορτώθηκε το σύνολο δεδομένων olist_order_items_dataset το οποίο έχει πληροφορίες για τα προϊόντα που περιέχονται σε κάθε παραγγελία. Έπειτα, επιλέχθηκαν οι στήλες *order_id* και *product_id*, οι οποίες είναι απαραίτητες για τον σχηματισμό των συναλλαγών στο πλαίσιο της ανάλυσης συσχετίσεων. Στη συνέχεια οι πίνακες αυτοί ενώθηκαν σύμφωνα με το κοινό πεδίο order_id, με στόχο την δημιουργία ενός ενιαίου συνόλου δεδομένων που αντικατροπτίζει ποια προϊόντα αντιστοιχούν σε κάθε παραγγελία.

```
# Επιλογή των απαραίτητων μεταβλητών για ανάλυση συσχετίσεων προϊόντων (Market Basket Analysis)
# Κάθε παραγγελία αντιστοιχεί σε ένα σύνολο προϊόντων
transactions = items[['order_id', 'product_id']]
```

Έπειτα, πραγματοποιήθηκε ομαδοποίηση ανά παραγγελία με την δημιουργία πριόντων για κάθε order_id. Με αυτό τον τρόπο κάθε παραγγελία εμφανίζεται ως το σύνολο των προϊόντων που αγοράστηκαν μαζί, δηλαδή ως το καλάθι αγορών κάθε συναλλαγής.

```
# Ομαδοποίηση προϊόντων ανά παραγγελία
# Δημιουργία "καλαθιού αγορών" για κάθε παραγγελία
basket = transactions.groupby('order_id')['product_id'].apply(list)
```

Στη συνέχεια εξετάστηκε το πλήθος των προϊόντων τα οποία περιλαμβάνονται σε κάθε καλάθι αγορών. Βέβαια, αποδείχθηκε πως η πλειονότητα των παραγγελιών περιείχε μόνο ένα προϊόν (όπως φαίνεται στο product_id – 86843 καλάθια με ένα προϊόν) , κάτι το οποίο περιορίζει τη μέθοδο των συσχετίσεων. Για αυτό το λόγο, περιόριστηκε το σύνολο δεδομένων σε καλάθια που περιείχαν τουλάχστον δύο διαφορετικά προϊόντα. Έπειτα από το φιλτάρισμα που πραγματοποιήθηκε το σύνολο των συναλλαγών κυμαίνεται πλέον στα 3236 καλάθια και αυτό εξασφαλίζει ότι τα δεδομένα είναι κατάλληλα για την εφαρμογή τεχνικών ανάλυσης συσχετίσεων. Ακόμη, στην παρακάτω εικόνα, δίνεται η ποσότητα των ID των προϊόντων που εμφανίζονται πάνω από μία φορά. (Tan et al., 2016).

```
# Εμφάνιση των συχνότερων καλαθιών αγορών (ίδιων συνδυασμών προϊόντων)
# Κάθε εγγραφή αντιστοιχεί σε ένα μοναδικό σύνολο προϊόντων ανά παραγγελία
basket.value_counts().head(10)
```

```
product_id
[99a4788cb24856965c36a24e339b6058]    397
[aca2eb7d00ea1a7b8ebd4e68314663af]    346
[d1c427060a0f73f6b889a5c7c61f2ac4]    291
[53b36df67ebb7c41585e8d54d6772e08]    290
[154e7e31ebfa092203795c972e5804a6]    256
[2b4609f8948be18874494203496bc318]    254
[422879e10f46682990de24d770e7f83d]    249
[3dd2a17168ec895c781a9191c1e95ad7]    240
[389d119b48cf3043d311335e499d9c6b]    225
[7c1bd920dbdf22470b68bde975dd3ccf]    216
Name: count, dtype: int64
```

```
# Κατανομή του πλήθους προϊόντων ανά καλάθι αγορών
# Παρουσιάζεται πόσες παραγγελίες περιέχουν 1, 2, 3 κ.ο.κ. προϊόντα
basket.apply(len).value_counts().sort_index().head(10)
```

```
product_id
1     88863
2      7516
3      1322
4       505
5       204
6       198
7        22
8         8
9         3
10        8
Name: count, dtype: int64
```

```
# Φιλτράρισμα καλαθιών που περιλαμβάνουν τουλάχιστον δύο διαφορετικά προϊόντα
# Το βήμα αυτό είναι απαραίτητο για την εξόρυξη κανόνων συσχέτισης,
# καθώς τα καλάθια με ένα μόνο προϊόν δεν παράγουν συσχετίσεις
basket_2plus = basket[basket.apply(lambda x: len(set(x)) >= 2)]

print("Σύνολο καλαθιών:", len(basket))
print("Καλάθια με ≥2 προϊόντα:", len(basket_2plus))
```

```
Σύνολο καλαθιών: 98666
Καλάθια με ≥2 προϊόντα: 3236
```

4.6.3 Κωδικοποίηση καλαθιών αγορών (Transaction Encoding)

Η παρούσα ενότητα ασχολείται με την κωδικοποιήση των καλαθιών σε δυαδική μορφή, ώστε στη συνέχεια να μπορεί να πραγματοποιήθει η εφαρμογή του αλγορίθμου Apriori. Ειδικότερα, κάθε γραμμή αντιστοιχεί σε μία συναλλαγή του καλαθιου αγορών ενώ κάθε στήλη αντιπροσωπεύει ένα μοναδικό προϊόν. Η τιμή 1 (true) δηλώνει ότι το προϊόν βρίσκεται στο καλάθι ενώ η 0 (false) ότι δεν περιλαμβάνεται.

Για τον λόγο αυτό χρησιμοποιήθηκε ο TransactionEncoder της βιβλιοθήκης mlxtend με τον οποίο η λίστα προϊόντων κάθε συναλλαγής μετατρέπεται σε δυαδικό πίνακα παρουσίας (true) / απουσίας (false). Η διαδικασία αυτή είναι απαραίτητη για την αναπαράσταση των συναλλαγών σε κατάλληλη μορφή για την εξαγωγή συχνών συνδυασμών προϊόντων (Raschka, 2018).

Το τελικό dataset που δημιουργείται ονομάζεται basket_encoded και πρόκειται για ένα δυαδικό πίνακα που περιλαμβάνει μεγάλο αριθμό στηλών το οποίο προκύπτει από το πλήθος διαφορετικών προϊόντων στο σύνολο των δεδομένων. Από οσο φαίνεται και στο κώδικα, οι περισσότερες τιμές είναι false κάτι το οποίο είναι απολύτως φυσιολογικό εφόσον το κάθε καλάθι περιλαμβάνει περιορισμένο αριθμό προϊόντων σε σχέση με το σύνολο των διαθέσιμων προϊόντων (Hahsler et al., 2005).

```python
from mlxtend.preprocessing import TransactionEncoder
import pandas as pd
```

```python
# Επανακωδικοποίηση των φιλτραρισμένων καλαθιών (≥2 προϊόντα) σε δυαδική μορφή,
# ώστε να χρησιμοποιηθούν ως είσοδος στον αλγόριθμο Apriori
te = TransactionEncoder()
te_array = te.fit(basket_2plus).transform(basket_2plus)

basket_encoded = pd.DataFrame(te_array, columns=te.columns_)
```

```python
# Προβολή του δυαδικά κωδικοποιημένου πίνακα συναλλαγών
# μετά το φιλτράρισμα καλαθιών με τουλάχιστον δύο προϊόντα
basket_encoded.head()
```

	0011c512eb256aa0dbbb544d8dffcf6e	001b72dfd63e9833e8c02742adf472e3	0042f1a9a7e0edd1400c6cd0fda065f8	005030ef108f58b46b78116f754d8d38	0060b415594
0	False	False	False	False	
1	False	False	False	False	
2	False	False	False	False	
3	False	False	False	False	
4	False	False	False	False	

5 rows × 4885 columns

4.6.4 Εξόρυξη συσχετίσεων με Apriori και κανόνες συσχέτισης

Στο πλαίσιο αυτής της ενότητας αναλύεται η συσχέτιση των προϊόντων που τείνουν να εμφανίζονται μαζί. Η ανάλυση αυτή προσφέρει πληροφορίες σχετικά με την αγοραστική συμπεριφορά των καταναλωτών και οι επιχειρήσεις μπορούν να την εκμεταλλευτούν ώστε να αναπτύξουν στρατηγικές διασταυρούμενων πωλήσεων (cross-selling).

Εφόσον δημιουργήθηκε το δυαδικά κωδικοποιημένο σύνολο basket_encoded το επόμενο βήμα είναι να εφαρμοστεί ο αλγόριθμος Apriori για την εξαγωγή συχνών συνόλων προϊόντων (frequent itemsets). Εξαιτίας του μεγάλου αριθμού διαφορετικών προϊόντων, ορίστηκε το min_support ίσο με 0,005 δηλαδή με 0,5% για να περιοριστεί η υπολογιστικη πολυπλοκότητα και να παραμείνουν οι συνδυασμοί προϊόντων που εμφανίζονται με ικανοποιητική συχνότητα στο σύνολο των συναλλαγών.

Από τον πίνακα των αποτελεσμάτων διακρίνουμε ότι τα πιο συχνά itemsets αποτελούνται κυρίως από μεμονωμένα προϊόντα, με τιμές support να κυμαίνονται από 0,83% εως 1,6%. Αυτό είναι αναμενόμενο, διότι το μεγαλύτερο ποσοστό των παραγγελιών περιέχει ένα μόνο προϊόν, όπως προέκυψε και από την ανάλυση των καλαθιών αγοράς.

```python
# Εξόρυξη συχνών συνόλων προϊόντων με τον αλγόριθμο Apriori,
# εφαρμόζοντας κατώφλι υποστήριξης ίσο με 0.5%
from mlxtend.frequent_patterns import apriori, association_rules

# Συχνά σύνολα προϊόντων (frequent itemsets)
frequent_itemsets = apriori(
    basket_encoded,
    min_support=0.005,      # κατώφλι 0.005 (0.5%)
    use_colnames=True
)

frequent_itemsets = frequent_itemsets.sort_values("support", ascending=False)

frequent_itemsets.head(10)
```

	support	itemsets
9	0.016069	(99a4788cb24856965c36a24e339b6058)
4	0.014833	(36f60d45225e60c7da4558b070ce4b60)
10	0.011743	(e53e557d5a159f5aa2c5e995dfdf244b)
2	0.011125	(35afc973633aaeb6b877ff57b2793310)
13	0.010507	(36f60d45225e60c7da4558b070ce4b60, e53e557d5a1...
6	0.009889	(422879e10f46682990de24d770e7f83d)
8	0.008962	(53759a2ecddad2bb87a079a1f1519f73)
12	0.008962	(35afc973633aaeb6b877ff57b2793310, 99a4788cb24...
5	0.008653	(389d119b48cf3043d311335e499d9c6b)
3	0.008344	(368c6c730842d78016ad823897a372db)

Στη συνέχεια από τα συχνά εμφανιζόμενα σύνολα προϊόντων εξήχθησαν κανόνες συσχέτισης με χρήση της μετρικής lift, η οποία εκφράζει τη δύναμη συσχέτισης μεταξύ των

προϊόντων και τέθηκε ελάχιστο όριο lift ≥ 1.0 για να διατηρηθούν μόνο οι κανόνες που παρουσιάζουν θετική συσχέτιση.

```
# Δημιουργία κανόνων συσχέτισης από τα συχνά σύνολα προϊόντων
# με χρήση του δείκτη lift, ώστε να εντοπιστούν μη τυχαίες συσχετίσεις
rules = association_rules(
    frequent_itemsets,
    metric="lift",
    min_threshold=1.0
)

# Τα αποτελέσματα ταξινομούνται βάσει lift και confidence
# για ευκολότερη ερμηνεία των ισχυρότερων κανόνων
rules = rules.sort_values(["lift", "confidence"], ascending=False)
rules.head(10)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	representativity	leverage	co
4	(f4f67ccaece962d013a4e1d7dc3a61f7)	(4fcb3d9a5f4871e8362dfedbdb02b064)	0.005562	0.005871	0.005253	0.944444	160.853801	1.0	0.005221	1
5	(4fcb3d9a5f4871e8362dfedbdb02b064)	(f4f67ccaece962d013a4e1d7dc3a61f7)	0.005871	0.005562	0.005253	0.894737	160.853801	1.0	0.005221	
1	(e53e557d5a159f5aa2c5e995dfdf244b)	(36f60d45225e60c7da4558b070ce4b60)	0.011743	0.014833	0.010507	0.894737	60.320175	1.0	0.010333	
0	(36f60d45225e60c7da4558b070ce4b60)	(e53e557d5a159f5aa2c5e995dfdf244b)	0.014833	0.011743	0.010507	0.708333	60.320175	1.0	0.010333	
3	(99a4788cb24856965c36a24e339b6058)	(35afc973633aaeb6b877ff57b2793310)	0.016069	0.011125	0.008962	0.557692	50.130342	1.0	0.008783	
2	(35afc973633aaeb6b877ff57b2793310)	(99a4788cb24856965c36a24e339b6058)	0.011125	0.016069	0.008962	0.805556	50.130342	1.0	0.008783	

Για λόγους ευκολότερης ερμηνείας και εστίασης στους πιο αξιόπιστους κανόνες παρέμειναν μόνο οι βασικές μετρικές support, confidence και lift. Επίσης, πραγμτοποιήθηκε ακόμη ένα φιλτράρισμα με confidence ≥ 20% και lift ≥ 1.20. στο πλαίσιο αυτό, η ανάλυση έχει επικεντρωθεί στις πιο ισχυρές και επιχειρησιακά χρήσιμες συσχετίσεις προϊόντων.

```
# Φιλτράρισμα των κανόνων συσχέτισης βάσει κατωφλίων confidence και lift,
# με στόχο τη διατήρηση μόνο των πιο αξιόπιστων και ουσιαστικών συσχετίσεων
rules_filtered = rules[
    (rules["confidence"] >= 0.20) &   # π.χ. τουλάχιστον 20%
    (rules["lift"] >= 1.20)           # π.χ. lift αρκετά πάνω από 1
].sort_values(["lift","confidence"], ascending=False)

rules_filtered.head(10)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	representativity	leverage	co
4	(f4f67ccaece962d013a4e1d7dc3a61f7)	(4fcb3d9a5f4871e8362dfedbdb02b064)	0.005562	0.005871	0.005253	0.944444	160.853801	1.0	0.005221	1
5	(4fcb3d9a5f4871e8362dfedbdb02b064)	(f4f67ccaece962d013a4e1d7dc3a61f7)	0.005871	0.005562	0.005253	0.894737	160.853801	1.0	0.005221	
1	(e53e557d5a159f5aa2c5e995dfdf244b)	(36f60d45225e60c7da4558b070ce4b60)	0.011743	0.014833	0.010507	0.894737	60.320175	1.0	0.010333	
0	(36f60d45225e60c7da4558b070ce4b60)	(e53e557d5a159f5aa2c5e995dfdf244b)	0.014833	0.011743	0.010507	0.708333	60.320175	1.0	0.010333	
3	(99a4788cb24856965c36a24e339b6058)	(35afc973633aaeb6b877ff57b2793310)	0.016069	0.011125	0.008962	0.557692	50.130342	1.0	0.008783	
2	(35afc973633aaeb6b877ff57b2793310)	(99a4788cb24856965c36a24e339b6058)	0.011125	0.016069	0.008962	0.805556	50.130342	1.0	0.008783	

Τα αποτελέσματα έδειξαν κανόνες με εξαιρτετικα υψηλές τιμές lift που σε ορισμένες περιπτώσεις ξεπερνούν τις 150, το οποίο αποδικνύει πολύ ισχυρή συσχέτιση μεταξύ συγκεκριμένων προϊόντων. Για παράδειγμα σε μερικούς κανόνες η πιθανότητα αγοράς ενός προϊόντος Α αυξάνεται δραματικά όταν έχει αγοραστεί το προϊόν Β.

Επιπλέον, οι τιμές confidence σε αρκετούς κανόνες ξεπερνούν το ποσοστό των 80 % το οποίο αποδικνύει ότι όταν ένα προίον Α υπάρχει στην παραγγελία, υπάρχει μεγάλη πιθανότητα να περιλαμβάνεται και το συσχετιζόμενο προίον Β. Αντιθέτως, οι σχετικά χαμηλές τιμές support δείχνουν ότι οι συγκεκριμένοι συνδυασμοί, αν και αποδικνύονται πολύ ισχυροί, εμφανίζονται σε περιορισμένο ποσοστό του συνόλου. Αυτό σχετίζεται με την δομή του dataset, όπου η πλειονότητα των παραγγελιών περιλαμβάνει ένα προϊόν στο καλάθι.

ΚΕΦΑΛΑΙΟ 5. Συμπεράσματα

5.1 Σύνοψη μελέτης

Ο σκοπός της παρούσας διπλωματικής εργασίας ήταν η διερεύνηση της αγοραστικής συμπεριφοράς των καταναλωτών με την χρήση διάφορων τεχνικών μηχανικής μάθησης και ανάλυσης δεδομένων. Ιδιαίτερα έγινε χρήση του συνόλου δεδομένων Olist, το οποίο περιλαμβάνει πραγματικές συναλλαγές ηλεκτρονικού εμπορίου σε χρονική περίοδο δύο ετών, με το οποίο παρέχεται η δυνατότητα ανάλυσης τόσο της λειτουργικής πλευράς των παραγγελιών όσο και προτύπων αγορών των καταναλωτών.

Κατά την διαδικασία της μελέτης χρησιμοποιήθηκαν δύο διαφορετικές αλλά συμπληρωματικές προσεγγίσεις. Η πρώτη ανήκει στην επιβλεπόμενη μάθηση και βασίζεται στο μοντέλο Logistic Regression το οποίο εστίασε στην πρόβλεψη της επιτυχούς ολοκλήρωσης των παραγγελίων. Η δεύτερη αποτελεί προσέγγιση της μη επιβλεπόμενης μάθησης και αφορά την ανάλυση συσχετίσεων με την χρήση του αλγορίθμου Apriori για τον εντοπισμό συνδυασμών προϊόντων που αγοράζονται μαζί.

Οι δύο προσεγγίσεις αποσκοπούν σε διαφορετικούς σκοπούς και προσφέρουν διαφορετικού τύπου αποτελέσματα. Το γεγονός αυτό προσφέρει μια πολυδιάστατη κατανόση της συμπεριφορας των καταναλωτών και των διαδικασιών του ηλεκτρονικού εμπορίου.

5.2 Συμπεράσματα από την επιβλεπόμενη μάθηση (Logistic Regression)

Η πρώτη προσέγγιση Logistic Regression που εφαρμόστηκε στο πλαίσιο της διπλωματικής είχε ως στόχο να προβλέψει κατά πόσο οι παραγγελίες ολοκληρώνονται επιτυχώς ή όχι. Σύμφωνα με τα αποτελέσματα, το μοντέλο λάμβανε υψηλή συνολική ακρίβεια. Το γεγονός αυτό υποδηλώνει, ότι το μοντέλο μπορούσε να αναγνωρίσει με μεγάλη αποτελεσματικότητα τα πλειοψηφοί πρότυπα δεδομένων.

Ωστόσο, η ανάλυση ανέδειξε κάποιους περιορισμούς. Συγκεκριμένα, διαπιστώθηκε έντονη ανισορροπία μεταξύ της πλειοψηφικής και της μειοψηφικής κλάσης (delivered έναντι non-delivered παραγγελιών), η οποία ήταν βασικό εμπόδιο στο να πραγματοποιηθεί αποτελεσματική πρόβλεψη για την μειοψηφική κλάση των μη παραδομένων παραγγελιών. Συνεπώς, παρόλο που εμφανίστηκε υψηλή ακρίβεια, η χρησιμότητά του ως εργαλείο έγκαιρης ανίχνευσης προβληματικών παραγγελιών ήταν περιορισμένη.

Επομένως, η Logistic Regression είναι περισσότερο κατάλληλη για χρήση ως εργαλείο γενικής εκτίμησης κινδύνου και ανάλυσης τάσεων, παρά ως μηχανισμός αυτόματης λήψης αποφάσεων σε πραγματικό χρόνο. Υπάρχει δυνατότητα βελτίωσης εφόσον συνδυαστεί με τεχνικές αντιμετώπισης της ανισορροπίας κλάσεων ή με πιο σύνθετες μεθόδους μηχανικής μάθησης.

5. 3 Συμπεράσματα από την ανάλυση συσχετίσεων (Apriori)

Η δεύτερη προσέγγιση της παρούσας διπλωματικής εργασίας, που στηρίχτηκε στην ανάλυση συσχετίσεω δεν εστίαζε στην πρόβλεψη αλλά αντίθετα είχε ως στόχο την ανακάλυψη προτύπων αγοραστικής συμπεριφοράς. Βασικό εργαλείο της ήταν ο αλγόριθμος Apriori μέσω του οποίου εντοπίστηκαν σύνολα προϊόντων και κανόνες συσχέτισης που αποκάλυψαν τα προϊόντα που έχουν την τάση να αγοράζονται μαζί.

Γενικότερα, διαπιστώθηκε ότι το μεγαλύτερο μέρος των παραγγελιών αποτελούνταν από μόνο ένα προϊόν στο καλάθι, παρόλα αυτά υπήρξαν μερικοί συνδυασμοί προϊόντων με ιδιαίτερα υψηλές τιμές κανόνων lift και confidence. Οι συγκεκριμένοι υποδηλώνουν ότι τα προϊόντα μεταξύ τους έχουν ισχυρές και μη τυχαίες σχέσεις, όπου μπορούν να χρησιμοποιηθούν σε στρατηγικές διασταυρούμενων πωλήσεων cross-selling, προτάσεις προϊόντων και σχεδιασμό προωθητικών ενεργειών.

Σε αντίθεση με τη Logistic Regression, η ανάλυση συσχετίσεων δεν επηρεάζεται από ανισορροπία δεδομένων αλλά μπορεί να προσφέρει άμεσα ερμηνεύσιμα αποτελέσματα. Το γεγονός αυτό την καθιστά χρήσιμο εργαλείο για επιχειρησιακές εφαρμογές στο ηλεκτρονικό εμπόριο, συμβάλλοντας σε μια πιο ολοκληρωμένη ανάλυση σε συνδυασμό με την επιβλεπόμενη μάθηση.

5. 4 Σύγκριση και συμπληρωματικότητα των μεθόδων

Όσον αφορά την σύγκριση των δύο προσεγγίσεων καταδικνύεται ότι καμία από τις δύο δεν υπερτερεί από την άλλη, διότι εξυπηρετούν διαφορετικούς σκοπούς. Η Logistic Regression εστιάζει στην πρόβλεψη της έκβασης μιας παραγγελίας και βασίζεται σε χαρακτηριστικά που σχετίζονται με την ολοκλήρωση των παραγγελιών. Ωστόσο, η ανάλυση συσχετίσεων επικεντρώνεται στη συμπεριφορά των καταναλωτών και στη δημιουργία σχέσεων μεταξύ των προϊόντων.

Ο συνδυασμός των δύο αυτών προσεγγίσεων προσφέρει μία πιο ολοκληρωμένη ανάλυση του ηλεκτρονικού εμπορίου. Η επιβλεπόμενη μάθηση επικεντρώνεται στις λειτουργικές αποφάσεις και την διαχείριση κινδύνου δημιουργώντας την δυνατότητα για βελτίωση στρατηγικών logistics ενώ η μη επιβλεπόμενη μάθηση ετσιάζει σε στρατηγικές marketing και πωλήσεων δίνοντας την δυνατότητα ανάπτυξης στρατηγικών προώθησης προϊόντων.

Συνεπώς, η συνδυαστική εφαρμογή των δύο προσεγγίσεων προσφέρει την δυνατότητα στις επιχειρήσεις να λαμβάνουν τόσο προληπτικά μέτρα όσο και τις κατάλληλες στρατηγικές αποφάσεις, ενισχύοντας την επιχειρησιακή τους απόδοση.

Βιβλιογραφία

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, SIGMOD '93*, 207–216. https://doi.org/10.1145/170035.170072

Alpaydin, E. (2020). *Introduction to Machine Learning, fourth edition*. MIT Press.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). A Brief Survey of Deep Reinforcement Learning. *IEEE Signal Processing Magazine*, *34*(6), 26–38. https://doi.org/10.1109/MSP.2017.2743240

Batini, C. (2009). Data quality assessment. In *Encyclopedia of Database Systems* (pp. 608–612). Springer.

Bhatnagar, A., & Ghose, S. (2004). Segmenting consumers based on the benefits and risks of Internet shopping. *Journal of Business Research, Mobility and Markets: Emerging Outlines of M-Commerce*, *57*(12), 1352–1360. https://doi.org/10.1016/S0148-2963(03)00067-5

Brazilian E-Commerce Public Dataset by Olist. (n.d.). Retrieved 27 January 2026, from https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brijs, T., Swinnen, G., Vanhoof, K., & Wets, G. (2004). Building an Association Rules Framework to Improve Product Assortment Decisions. *Data Mining and Knowledge Discovery*, *8*(1), 7–23. https://doi.org/10.1023/B:DAMI.0000005256.79013.69

Brownlee, J. (2022). *Machine learning mastery*. Machine Learning Mastery.

Brunner, F. (n.d.). *Mastering the game of Go with deep neural networks and tree search (Silver et al., 2016)*.

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). *Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?* (SSRN Scholarly Paper No. 1819486). Social Science Research Network. https://doi.org/10.2139/ssrn.1819486

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233–234. https://doi.org/10.1038/nmeth.4642

Chaffey, D., & Ellis-Chadwick, F. (2019). *Digital Marketing*. Pearson UK.

Chaffey, D., Hemphill, T., & Edmundson-Bird, D. (2019). *Digital Business and E-commerce Management*. Pearson UK.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165–1188. https://doi.org/10.2307/41703503

Chen, Y., Fay, S., & Wang, Q. (2011). The Role of Marketing in Social Media: How Online Consumer Reviews Evolve. *SSRN Electronic Journal*, *25*. https://doi.org/10.2139/ssrn.1710357

Chong, A., Ngai, E., Ch'ng, E., Li, B., & Lee, F. (2015). Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach. *International Journal of Operations & Production Management*.

Dal Pozzolo, A., Caelen, O., Johnson, R., & Bontempi, G. (2015, December 9). *Calibrating Probability with Undersampling for Unbalanced Classification*. https://doi.org/10.1109/SSCI.2015.33

Digital Economy Report 2021 (Overview). (2021). https://unctad.org/system/files/official-document/der2021_overview_en_0.pdf

Dynamic pricing and learning. (2013). s.n.].

Edelman, D. C., & Singer, M. (n.d.). *Competing on Customer Journeys*.

Elmaghraby, W., & Keskinocak, P. (2003). Dynamic Pricing in the Presence of Inventory Considerations: Research Overview, Current Practices, and Future Directions. *Management Science*, *49*, 1287–1309. https://doi.org/10.1287/mnsc.49.10.1287.17315

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Correction: Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *546*(7660), 686–686. https://doi.org/10.1038/nature22985

Fawcett, T., & Provost, F. (1997). *Combining Data Mining and Machine Learning for Effective Fraud Detection*.

Filieri, R. (2015). What makes online reviews helpful? A diagnosticity-adoption framework to explain informational and normative influences in e-WOM. *Journal of Business Research*, *68*(6), 1261–1270. https://doi.org/10.1016/j.jbusres.2014.11.006

Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, *479*, 448–455. https://doi.org/10.1016/j.ins.2017.12.030

Gajowniczek, K., & Ząbkowski, T. (2014). Short Term Electricity Forecasting Using Individual Smart Meter Data. *Procedia Computer Science*, *35*, 589–597. https://doi.org/10.1016/j.procs.2014.08.140

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining* (Vol. 72). Springer International Publishing. https://doi.org/10.1007/978-3-319-10247-4

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in Online Shopping: An Integrated Model. *MIS Quarterly*, *27*(1), 51–90. https://doi.org/10.2307/30036519

Gefen, D., & Pavlou, P. (2011). The Boundaries of Trust and Risk: The Quadratic Moderating Role of Institutional Structures. *Information Systems Research*, *Articles in Advance*, 1–20. https://doi.org/10.2307/23274654

Global retail e-commerce sales 2022-2028| Statista. (n.d.). Retrieved 27 January 2026, from https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/

Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems*, *26*(3), 191–209. https://doi.org/10.1016/j.jsis.2017.07.003

Guyon, I., & Elisseeff, A. (n.d.). *An Introduction to Variable and Feature Selection*.

Hagiu, A., & Wright, J. (2015). Marketplace or Reseller? *Management Science*, *61*(1), 184–203.

Hahsler, M., Grün, B., & Hornik, K. (2005). arules—A Computational Environment for Mining Association Rules and Frequent Item Sets. *Journal of Statistical Software*, *14*, 1–25. https://doi.org/10.18637/jss.v014.i15

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd edn). Morgan Kaufmann. https://www.perlego.com/book/1836800/data-mining-concepts-and-techniques-pdf

Han, J., Pei, J., & Tong, H. (2023). *Data mining: Concepts and techniques* (Fourth edition). Morgan Kaufmann is an imprint of Elsevier.

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow—Aurélien Géron—Βιβλία Google. (n.d.). Retrieved 27 January 2026, from https://books.google.gr/books?hl=el&lr=&id=X5ySEAAAQBAJ&oi=fnd&pg=PT17&dq=Hands-on+Machine+Learning+with+Scikit-Learn+Keras+and+TensorFlow+G%C3%A9ron&ots=yC3zti05uJ&sig=HpvTX7rcuYYBWLlEcPHaBy18ang&redir_esc=y#v=onepage&q&f=false

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hübner, A., Kuhn, H., & Wollenburg, J. (2016). Last mile fulfilment and distribution in omni-channel grocery retailing: A strategic planning framework. *International Journal of Retail & Distribution Management*, *44*, 228–247. https://doi.org/10.1108/IJRDM-11-2014-0154

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, Award Winning Papers from the 19th International Conference on Pattern Recognition (ICPR)*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Jannach, D., & Adomavicius, G. (2016). Recommendations with a Purpose. *Proceedings of the 10th ACM Conference on Recommender Systems, RecSys '16*, 7–10. https://doi.org/10.1145/2959100.2959186

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Jr, D. W. H., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons.

Kaelbling, L. P., Littman, M. L., Moore, A. W., & Hall, S. (n.d.). Reinforcement Learning: A Survey. *Reinforcement Learning*.

Kaggle: Your Machine Learning and Data Science Community. (n.d.). Retrieved 27 January 2026, from https://www.kaggle.com/

Kannan, P. K., & Li, H. "Alice". (2017). Digital marketing: A framework, review and research agenda. *International Journal of Research in Marketing*, *34*(1), 22–45. https://doi.org/10.1016/j.ijresmar.2016.11.006

Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. *Procedia Computer Science*, *85*, 78–85. https://doi.org/10.1016/j.procs.2016.05.180

Kim, D. J., Ferrin, D. L., & Rao, H. R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision Support Systems*, *44*(2), 544–564. https://doi.org/10.1016/j.dss.2007.07.001

Kim, K. G. (2016). Book Review: Deep Learning. *Healthcare Informatics Research*, *22*(4), 351. https://doi.org/10.4258/hir.2016.22.4.351

Kohavi, R. (2001). *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. *14*.

Kolodin, D., Telychko, O., Rekun, V., Tkalych, M., & Yamkovyi, V. (2020). *Artificial Intelligence in E-Commerce: Legal Aspects*. 96–102. https://doi.org/10.2991/aebmr.k.200318.012

Kotsiantis, S. B. (n.d.). *Supervised Machine Learning: A Review of Classification Techniques*.

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). *Data Preprocessing for Supervised Leaning*. *1*(1).

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

Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315108230

Kumar, V., Dixit, A., Javalgi, R. (Raj) G., & Dass, M. (2016). Research framework, strategies, and applications of intelligent agent technologies (IATs) in marketing. *Journal of the Academy of Marketing Science*, *44*(1), 24–45. https://doi.org/10.1007/s11747-015-0426-9

Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons.

Laudon, K. C., & Traver, C. G. (2024). *E-commerce: Business, technology, society* (Eighteenth edition, global edition). Pearson.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lemon, K. N., & Verhoef, P. C. (2016). Understanding Customer Experience Throughout the Customer Journey. *Journal of Marketing*, *80*(6), 69–96. https://doi.org/10.1509/jm.15.0420

Li, X., Wu, C., & Mai, F. (2019). The effect of online reviews on product sales: A joint sentiment-topic analysis. *Information & Management, Social Commerce and Social Media: Behaviors in the New Service Economy*, *56*(2), 172–184. https://doi.org/10.1016/j.im.2018.04.007

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics: 5.1* (pp. 281–298). University of California Press. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992

McKinney, W. (2012). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Inc.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

MobasherBamshad, CooleyRobert, & SrivastavaJaideep. (2000). Automatic personalization based on Web usage mining. *Communications of the ACM*. (New York, NY, USA). https://doi.org/10.1145/345124.345169

Moro, S., Rita, P., & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. *Journal of Business Research*, *69*(9), 3341–3351. https://doi.org/10.1016/j.jbusres.2016.02.010

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful review? A study of customer reviews on Amazon. com (SSRN Scholarly Paper No. ID 2175066). *Social Science Research Network, Rochester, NY*.

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, Inc.

'Multivariate Data Analysis' by Joseph F. Hair. (n.d.). Retrieved 27 January 2026, from https://digitalcommons.kennesaw.edu/facpubs/2925/

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *HORIZONS.B*, *4*, 51–62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2, Part 2), 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). Exploring the filter bubble: The effect of using recommender systems on content diversity. *Proceedings of the 23rd International Conference on World Wide Web, WWW '14*, 677–686. https://doi.org/10.1145/2566486.2568012

October 2003 THE DIGITIZATION OF WORD-OF-MOUTH: PROMISE AND CHALLENGES OF ONLINE. (n.d.). Studylib.Net. Retrieved 27 January 2026, from https://studylib.net/doc/13713445/october-2003-the-digitization-of-word-of-mouth---promise-...

OECD. (2020). E-commerce in the time of COVID-19. *OECD Policy Responses to Coronavirus (COVID-19)*. https://doi.org/10.1787/3a2b78e8-en

Offline Showrooms in Omnichannel Retail: Demand and Operational Benefits | Management Science. (n.d.). Retrieved 26 January 2026, from https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2016.2684

Parker, G. G., Alstyne, M. W. V., & Choudary, S. P. (2016). *Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You*. W. W. Norton & Company.

Pattern Recognition and Machine Learning | Springer Nature Link. (n.d.). Retrieved 26 January 2026, from https://link.springer.com/9780387310732

(PDF) An innovation resistance theory perspective on mobile payment solutions. (2025). *ResearchGate*. https://doi.org/10.1016/j.jretconser.2020.102059

(PDF) Data Cleaning: Problems and Current Approaches. (n.d.). ResearchGate. Retrieved 27 January 2026, from https://www.researchgate.net/publication/220282831_Data_Cleaning_Problems_and_Current_Approaches

(PDF) Data Science for Business. (n.d.). ResearchGate. Retrieved 27 January 2026, from https://www.researchgate.net/publication/256438799_Data_Science_for_Business

(PDF) Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. (2025). *ResearchGate*.

https://www.researchgate.net/publication/276412348_Evaluation_From_precision_recall_and_F-measure_to_ROC_informedness_markedness_correlation

(PDF) Machine Learning in Accounting & Finance: Architecture, Scope & Challenges. (2025). *ResearchGate*. https://doi.org/10.5539/ijbm.v17n5p13

Personalized Online Advertising Effectiveness: The Interplay of What, When, and Where | Marketing Science. (n.d.). Retrieved 26 January 2026, from https://pubsonline.informs.org/doi/abs/10.1287/mksc.2015.0930

Putting one-to-one marketing to work: Personalization, customization, and choice | Marketing Letters | Springer Nature Link. (n.d.). Retrieved 27 January 2026, from https://link.springer.com/article/10.1007/s11002-008-9056-z

Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann.

Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *Journal of Open Source Software*, *3*(24), 638. https://doi.org/10.21105/joss.00638

Ricci, F., Rokach, L., & Shapira, B. (2022). Recommender Systems: Techniques, Applications, and Challenges. In F. Ricci, L. Rokach, & B. Shapira (Eds), *Recommender Systems Handbook* (pp. 1–35). Springer US. https://doi.org/10.1007/978-1-0716-2197-4_1

Rochet, J.-C., & Tirole, J. (2006). Two-sided markets: A progress report. *The RAND Journal of Economics*, *37*(3), 645–667. https://doi.org/10.1111/j.1756-2171.2006.tb00036.x

Rose, S., Clark, M., Samouel, P., & Hair, N. (2012). Online Customer Experience in e-Retailing: An empirical model of Antecedents and Outcomes. *Journal of Retailing*, *88*(2), 308–322. https://doi.org/10.1016/j.jretai.2012.03.001

Ruder, S. (2017). *An overview of gradient descent optimization algorithms* (arXiv:1609.04747). arXiv. https://doi.org/10.48550/arXiv.1609.04747

Schmidhuber—2015—Deep Learning in Neural Networks An Overview | PDF | Deep Learning | Artificial Neural Network. (n.d.). Scribd. Retrieved 27 January 2026, from https://www.scribd.com/document/353968817/SCHMIDHUBER-2015-Deep-Learning-in-Neural-Networks-an-Overview

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Sharda, R., Delen, D., & Turban, E. (2018). *Business intelligence, analytics, and data science: A managerial perspective* (Fourth edition). Pearson.

Sheth, J. (2020). Impact of Covid-19 on consumer behavior: Will the old habits return or die? *Journal of Business Research*, *117*, 280–283. https://doi.org/10.1016/j.jbusres.2020.05.059

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Jr, K. C. L. (2017). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. John Wiley & Sons.

Smith, B., & Linden, G. (2017). Two Decades of Recommender Systems at Amazon.com. *IEEE Internet Computing*, *21*(3), 12–18. https://doi.org/10.1109/MIC.2017.72

Sutton, R. S., & Barto, A. G. (n.d.). *Reinforcement Learning: An Introduction*.

Tan, P.-N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education India.

The Effect of Word of Mouth on Sales: Online Book Reviews—Judith A. Chevalier, Dina Mayzlin, 2006. (n.d.). Retrieved 26 January 2026, from https://journals.sagepub.com/doi/abs/10.1509/jmkr.43.3.345

The Netflix Recommender System: Algorithms, Business Value, and Innovation: ACM Transactions on Management Information Systems: Vol 6, No 4. (n.d.). Retrieved 27 January 2026, from https://dl.acm.org/doi/abs/10.1145/2843948

Turban, E., Whiteside, J., King, D., & Outland, J. (2017). *Introduction to Electronic Commerce and Social Commerce*. Springer International Publishing. https://doi.org/10.1007/978-3-319-50091-1

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999. https://doi.org/10.1109/72.788640

Verhoef, P. C., Kannan, P. K., & Inman, J. J. (2015). From Multi-Channel Retailing to Omni-Channel Retailing: Introduction to the Special Issue on Multi-Channel Retailing. *Journal of Retailing, Multi-Channel Retailing*, *91*(2), 174–181. https://doi.org/10.1016/j.jretai.2015.02.005

Vishwakarma, A., & Verma, Y. (2025). The Brain Behind the Map: AI and Traffic Prediction in Google Maps. *International Journal of Scientific Research*, *11*.

Vladimir, Z. (1996). Electronic Commerce: Structures and Issues. *International Journal of Electronic Commerce*, *1*(1), 3–23. https://doi.org/10.1080/10864415.1996.11518273

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, *70*, 356–365. https://doi.org/10.1016/j.jbusres.2016.08.009

Wedel, M., & Kamakura, W. A. (2000). *Market Segmentation* (Vol. 8). Springer US. https://doi.org/10.1007/978-1-4615-4651-1

Wedel, M., & Kannan, P. K. (2016). Marketing Analytics for Data-Rich Environments. *Journal of Marketing*, *80*(6), 97–121. https://doi.org/10.1509/jm.15.0413

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678. https://doi.org/10.1109/TNN.2005.845141

Zhang, Q., Qiu, L., Wu, H., Wang, J., & Luo, H. (2019). Deep Learning Based Dynamic Pricing Model for Hotel Revenue Management. *2019 International Conference on Data Mining Workshops (ICDMW)*, 370–375. https://doi.org/10.1109/ICDMW.2019.00061

Zheng, A., & Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, Inc.

Παράρτημα Α – Κώδικας Ανάλυσης

Στο παράρτημα Α, παρατίθεται πλήρης κώδικας ο οποίος αναπτύχθηκε σε γλώσσα προγραμματισμού Python για την προεπεξεργασία των δεδομένων, την εκπαίδευση και αξιολόγηση των μοντέλων, καθώς και την εξαγωγή κανόνων συσχέτισης. Ο κώδικας είναι διαθέσιμος μέσω του ακόλουθου δημόσιου διαδικτυακού συνδέσμου GitHub:

https://github.com/sofiadem181-byte/ecommerce-logistics-ml/blob/main/demirisofia_ecommerce_ml.py

Παράρτημα Β - Ενδεικτικά Στιγμιότυπα Κώδικα και Αποτελεσμάτων

Ο παρών κώδικας ανάλυσης δεδομένων αναπτύχθηκε στο πλαίσιο της παρούσας διπλωματικής εργασίας και χρησιμοποιήθηκε για την προεπεξεργασία και ανάλυση των δεδομένων. Παρακάτω φαίνονται ενδεικτικά κάποιες από τις εντολές που χρησιμοποιήθηκαν μέσω Python στην πλατφόρμα Anacoda. Τα αποτελέσματα του κώδικα καταγράφονται στο κεφάλαιο 4 της παρούσας εργασίας.

Φόρτωση και αρχική διερεύνηση δεδομένων:

```
# Εισαγωγή της βιβλιοθήκης pandas για ανάλυση και διαχείριση δεδομένων
import pandas as pd

# Φόρτωση του dataset παραγγελιών (orders),το οποίο περιλαμβάνει πληροφορίες
# σχετικά με την κατάσταση και τα χρονικά σημεία εξέλιξης κάθε παραγγελίας
orders = pd.read_csv("olist_orders_dataset.csv")

# Εμφάνιση των πρώτων εγγραφών του συνόλου δεδομένων για προκαταρκτική διερεύνηση
orders.head()
```

```
# Έλεγχος της δομής του dataset παραγγελιών,
# Παρουσίαση τύπων δεδομένων και πλήθους μη κενών τιμών για κάθε μεταβλητή
orders.info()
```

```
# Περιγραφική σύνοψη των κατηγορικών (μη αριθμητικών) μεταβλητών του dataset παραγγελιών
# Παρουσιάζονται το πλήθος εγγραφών, οι μοναδικές τιμές και η συχνότερη εμφάνιση κάθε μεταβλητής
orders.describe()
```

```
# Φόρτωση του dataset πελατών, το οποίο περιλαμβάνει δημογραφικές και γεωγραφικές πληροφορίες
customers = pd.read_csv("olist_customers_dataset.csv")

# Εμφάνιση των πρώτων εγγραφών για αρχική διερεύνηση των δεδομένων
customers.head()
```

```
# Έλεγχος της δομής του dataset πελατών
# Παρουσίαση τύπων δεδομένων και ύπαρξης κενών τιμών στις μεταβλητές
customers.info()
```

```
# Φόρτωση του dataset προϊόντων ανά παραγγελία (τιμές προϊόντων και κόστος μεταφοράς)
items = pd.read_csv("olist_order_items_dataset.csv")
# Αρχική επισκόπηση των δεδομένων
items.head()
# Έλεγχος δομής, τύπων δεδομένων και πληρότητας των μεταβλητών
items.info()
```

```
# Φόρτωση του dataset προϊόντων, το οποίο περιλαμβάνει πληροφορίες
# σχετικά με την κατηγορία και τα βασικά χαρακτηριστικά κάθε προϊόντος
products = pd.read_csv("olist_products_dataset.csv")

products.head()
```

Συγχώνευση πινάκων:

```
# Ενοποίηση δεδομένων παραγγελιών και πελατών
# Συγχώνευση μέσω μοναδικού αναγνωριστικού customer_id
# Επιλέγεται εσωτερική σύζευξη (inner join) ώστε να διατηρηθούν
# μόνο οι παραγγελίες που αντιστοιχούν σε έγκυρους πελάτες
orders_customers = pd.merge(
    orders,
    customers,
    on="customer_id",
    how="inner"
)
```

```
# Ενοποίηση δεδομένων παραγγελιών, πελατών και προϊόντων με βάση το order_id
# Δημιουργία του τελικού συνόλου δεδομένων για ανάλυση
full_data = pd.merge(
    orders_customers,
    items,
    on="order_id",
    how="inner"
)

# Προβολή των πρώτων γραμμών του dataset
full_data.head()
```

```
# Έλεγχος αριθμού ελλιπών τιμών ανά μεταβλητή
full_data.isnull().sum()
```

```
# Υπολογισμός ποσοστού ελλιπών τιμών ανά μεταβλητή
(full_data.isnull().sum() / len(full_data)) * 100
```

Δημιουργία μεταβλητών:

```
# Δημιουργία δυαδικής μεταβλητής για την κατάσταση παράδοσης
# 1: παραγγελία παραδόθηκε, 0: διαφορετική κατάσταση
full_data["delivered_binary"] = full_data["order_status"].apply(
    lambda x: 1 if x == "delivered" else 0
)

# Κατανομή παραγγελιών ως προς την κατάσταση παράδοσης
full_data["delivered_binary"].value_counts()
```

```
# Μετατροπή των χρονικών μεταβλητών σε τύπο datetime
full_data["order_delivered_customer_date"] = pd.to_datetime(
    full_data["order_delivered_customer_date"]
)
full_data["order_estimated_delivery_date"] = pd.to_datetime(
    full_data["order_estimated_delivery_date"]
)

# Υπολογισμός ημερών καθυστέρησης παράδοσης ως η διαφορά
# μεταξύ πραγματικής και εκτιμώμενης ημερομηνίας παράδοσης
full_data["delivery_delay_days"] = (
    full_data["order_delivered_customer_date"]
    - full_data["order_estimated_delivery_date"]
).dt.days
```

```
# Επιλογή βασικών μεταβλητών για περαιτέρω ανάλυση και μοντελοποίηση
analysis_data = full_data[
    [
        "customer_state",
        "order_value",
        "delivered_binary",
        "delivery_delay_days"
    ]
]
analysis_data.head()
```

Εκπαίδευση μοντέλου ταξινόμησης:

```
analysis_data = full_data[
    ["order_value", "delivery_delay_days", "delivered_binary"]
].dropna()

# Επιλογή χαρακτηριστικών (features) και μεταβλητής στόχου
X = analysis_data[["order_value", "delivery_delay_days"]]
y = analysis_data["delivered_binary"]

# Διαχωρισμός δεδομένων σε σύνολα εκπαίδευσης και ελέγχου
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Εκπαίδευση μοντέλου λογιστικής παλινδρόμησης
model = LogisticRegression()
model.fit(X_train, y_train)
```

```
# Αξιολόγηση απόδοσης μοντέλου
from sklearn.metrics import accuracy_score, classification_report
y_pred = model.predict(X_test)
# Αναλυτική αξιολόγηση της απόδοσης του μοντέλου ανά κατηγορία
print(classification_report(y_test, y_pred))
```

```
# Υπολογισμός της μήτρας σύγχυσης (confusion matrix) για το μοντέλο ταξινόμησης
# Το αποτέλεσμα δείχνει ότι το μοντέλο ταξινομεί όλες τις παρατηρήσεις
# ως "delivered" (κλάση 1), χωρίς να προβλέπει καμία μη παραδομένη παραγγελία (κλάση 0),
# λόγω της έντονης ανισορροπίας μεταξύ των κλάσεων στο σύνολο δεδομένων
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, y_pred)
```

```
# Αναλυτική αξιολόγηση της απόδοσης του μοντέλου ανά κατηγορία
# Παρατηρείται υψηλή ακρίβεια για την πλειοψηφική κατηγορία (delivered),
# ενώ το μοντέλο αποτυγχάνει πλήρως να αναγνωρίσει μη παραδομένες παραγγελίες,
# γεγονός που επιβεβαιώνει το πρόβλημα της ανισορροπίας κλάσεων (class imbalance)
from sklearn.metrics import classification_report

print(classification_report(y_test, y_pred))
```

```
# Έλεγχος ενδεικτικών προβλέψεων του μοντέλου
# Διαπιστώνεται ότι όλες οι προβλέψεις ανήκουν στην κλάση "delivered" (1),
# επιβεβαιώνοντας τη μονομερή συμπεριφορά του μοντέλου
y_pred[:10]
```

```
# Μήτρα σύγχυσης του αρχικού (μη εξισορροπημένου) μοντέλου
# Το μοντέλο ταξινομεί όλες τις παρατηρήσεις στην πλειοψηφική κατηγορία ("delivered"),
# χωρίς σωστές προβλέψεις για μη παραδομένες παραγγελίες
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, y_pred)
```

```
# Υπολογισμός συνολικής ακρίβειας του αρχικού μοντέλου
# Η υψηλή τιμή της ακρίβειας είναι παραπλανητική,
# καθώς οφείλεται στην επικράτηση της πλειοψηφικής κλάσης
from sklearn.metrics import accuracy_score
accuracy_score(y_test, y_pred)
```

```
# Εκπαίδευση λογιστικής παλινδρόμησης με εξισορρόπηση κλάσεων (class_weight='balanced')
# Στόχος είναι η βελτίωση της πρόβλεψης της μειοψηφικής κατηγορίας (μη παραδομένες παραγγελίες),
# ακόμη και αν αυτό οδηγήσει σε μείωση της συνολικής ακρίβειας
from sklearn.linear_model import LogisticRegression

model_balanced = LogisticRegression(class_weight='balanced')
model_balanced.fit(X_train, y_train)

y_pred_balanced = model_balanced.predict(X_test)
```

```
# Αξιολόγηση του εξισορροπημένου μοντέλου ταξινόμησης
# Η χρήση εξισορρόπησης κλάσεων οδηγεί σε καλύτερη ανίχνευση μη παραδομένων παραγγελιών,
# ενώ η συνολική ακρίβεια μειώνεται, λόγω της αυξημένης έμφασης στη μειοψηφική κατηγορία
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred_balanced))
```

```
# Επιλογή των απαραίτητων μεταβλητών για ανάλυση συσχετίσεων προϊόντων (Market Basket Analysis)
# Κάθε παραγγελία αντιστοιχεί σε ένα σύνολο προϊόντων
transactions = items[['order_id', 'product_id']]
```

```
# Ομαδοποίηση προϊόντων ανά παραγγελία
# Δημιουργία "καλαθιού αγορών" για κάθε παραγγελία
basket = transactions.groupby('order_id')['product_id'].apply(list)
```

```
# Μετασχηματισμός των συναλλαγών σε δυαδικό πίνακα παρουσίας/απουσίας προϊόντων,
# προκειμένου να καταστεί δυνατή η εφαρμογή αλγορίθμων εξόρυξης συσχετίσεων,
# όπως ο αλγόριθμος Apriori
from mlxtend.preprocessing import TransactionEncoder

te = TransactionEncoder()
te_array = te.fit(basket).transform(basket)

basket_encoded = pd.DataFrame(te_array, columns=te.columns_)
```

Ανάλυση συσχετίσεων (Apriori):

```
# Εξόρυξη συχνών συνόλων προϊόντων με τον αλγόριθμο Apriori,
# εφαρμόζοντας κατώφλι υποστήριξης ίσο με 0.5%
from mlxtend.frequent_patterns import apriori, association_rules

# Συχνά σύνολα προϊόντων (frequent itemsets)
frequent_itemsets = apriori(
    basket_encoded,
    min_support=0.005,      # κατώφλι 0.005 (0.5%)
    use_colnames=True
)

frequent_itemsets = frequent_itemsets.sort_values("support", ascending=False)

frequent_itemsets.head(10)
```

```
# Δημιουργία κανόνων συσχέτισης από τα συχνά σύνολα προϊόντων
# με χρήση του δείκτη lift, ώστε να εντοπιστούν μη τυχαίες συσχετίσεις
rules = association_rules(
    frequent_itemsets,
    metric="lift",
    min_threshold=1.0
)

# Τα αποτελέσματα ταξινομούνται βάσει lift και confidence
# για ευκολότερη ερμηνεία των ισχυρότερων κανόνων
rules = rules.sort_values(["lift", "confidence"], ascending=False)
rules.head(10)
```

```
# Δημιουργία απλοποιημένου πίνακα κανόνων συσχέτισης
# και μετατροπή των συνόλων προϊόντων σε λίστες,
# ώστε να διευκολυνθεί η ερμηνεία και παρουσίαση των αποτελεσμάτων
rules_small = rules[["antecedents","consequents","support","confidence","lift"]].copy()

rules_small["antecedents"] = rules_small["antecedents"].apply(lambda x: list(x))
rules_small["consequents"] = rules_small["consequents"].apply(lambda x: list(x))

rules_small.head(10)
```

```
# Φιλτράρισμα των κανόνων συσχέτισης βάσει κατωφλίων confidence και lift,
# με στόχο τη διατήρηση μόνο των πιο αξιόπιστων και ουσιαστικών συσχετίσεων
rules_filtered = rules[
    (rules["confidence"] >= 0.20) &   # π.χ. τουλάχιστον 20%
    (rules["lift"] >= 1.20)           # π.χ. lift αρκετά πάνω από 1
].sort_values(["lift","confidence"], ascending=False)

rules_filtered.head(10)
```