ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών «Κυβερνοασφάλεια και Επιστήμη Δεδομένων»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	Τμηματοποίηση Πελατών και Ανάλυση Καλαθιού Αγορών σε Περιβάλλον Ηλεκτρονικού Εμπορίου Customer Segmentation and Market Basket Analysis for eCommerce
Ονοματεπώνυμο Φοιτητή	Αντώνιος Μωραΐτης
Πατρώνυμο	Παναγιώτης
Αριθμός Μητρώου	ΜΠΚΕΔ21002
Επιβλέπων	Δημήτριος Αποστόλου, Καθηγητής

Ημερομηνία Παράδοσης **Απρίλιος 2026**

Στην ολοκλήρωση της παρούσας μεταπτυχιακής διατριβής, ιδιαίτερα σημαντική ήταν η συμβολή του Διδάσκοντα του ΠΜΣ κ. Ανδρέα Ζάρα, που προσέφερε επιστημονική και συμβουλευτική καθοδήγηση σε όλα τα στάδια εκπόνησής της.

Τριμελής Εξεταστική Επιτροπή

Παναγιώτης Κοτζανικολάου	Δημήτριος Αποστόλου	Διονύσιος Σωτηρόπουλος
Καθηγητής	Καθηγητής	Αναπληρωτής Καθηγητής

Ευχαριστίες

Στο πλαίσιο ολοκλήρωσης της παρούσας μεταπτυχιακής διατριβής, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες προς τον κ. Ανδρέα Ζάρα για την πολύτιμη επιστημονική καθοδήγηση, τις ουσιαστικές υποδείξεις και τη συνεχή υποστήριξή του σε όλα τα στάδια της εκπόνησης της εργασίας. Η συμβολή του υπήρξε καθοριστική για τη διαμόρφωση της μεθοδολογικής προσέγγισης και την ολοκλήρωση της παρούσας ερευνητικής προσπάθειας. Επιπροσθέτως, θα ήθελα να εκφράσω την ειλικρινή μου ευγνωμοσύνη προς την κοπέλα μου για την αμέριστη στήριξη, την κατανόηση και την υπομονή που επέδειξε καθ' όλη τη διάρκεια της εκπόνησης της διατριβής. Η συμβολή της, σε προσωπικό επίπεδο, υπήρξε ιδιαίτερα σημαντική για την επιτυχή ολοκλήρωση της προσπάθειας αυτής. Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου για τη διαρκή ενθάρρυνση και τη στήριξή της. Η παρουσία και η εμπιστοσύνη τους συνέβαλαν ουσιαστικά στην επίτευξη του στόχου αυτού.

Περίληψη

Η μηχανική μάθηση προσφέρει στην βιομηχανία λιανικού εμπορίου την υποστήριξη για την πρόβλεψη επιτυχών πωλήσεων σε εγγεγραμμένους πελάτες. Το Market Basket Analysis (MBA) είναι μια εκ των κορυφαίων εφαρμογών του machine learning στο λιανικό εμπόριο. Βοηθά τις επιχειρήσεις να γνωρίζουν τι προϊόντα αγοράζει ο εγγεγραμμένος πελάτης, έτσι ώστε ο ιστότοπος του ηλεκτρονικού καταστήματος να είναι σχεδιασμένος με ανάλογο τρόπο. Αυτό που μελετάται κυρίως είναι η προηγούμενη αγοραστική δραστηριότητα. Οι εταιρείες επίσης το αξιοποιούν για σταυροειδείς πωλήσεις (cross-selling), προτείνοντας συμπληρωματικά προϊόντα ή υπηρεσίες σε έναν πελάτη. Έτσι αυξάνεται η αξία της παραγγελίας, αλλά και η ικανοποίηση και η διευκόλυνση του πελάτη. Το Market Basket Analysis (MBA) δημιουργεί εξατομίκευση της αγοραστικής εμπειρίας που οδηγεί τόσο στην αύξηση των εσόδων βραχυπρόθεσμα, όσο και στην διαχρονική αφοσίωση του πελάτη.Η ψηφιακή πλατφόρμα παράδοσης φαγητού, ειδών σούπερ μάρκετ, μικροαγορών κ.α. που μελετάται στην παρούσα διατριβή είναι ένα εξαιρετικό παράδειγμα αξιοποίησης του Market Basket Analysis (MBA) με σκοπό τις σταυροειδείς πωλήσεις. Μαζί με την αγορά συγκεκριμένου προϊόντος, προτείνεται στον πελάτη μια λίστα άλλων προϊόντων που πιθανόν να ενδιαφέρουν τον πελάτη. Το ποια προϊόντα εμφανίζονται σε αυτή την λίστα βασίζεται στην προηγούμενη αγοραστική δραστηριότητα του πελάτη, το ιστορικό αναζήτησης του, τι αγοράζουν άλλοι πελάτες συνδυαστικά με ένα συγκεκριμένο προϊόν, αλλά και άλλους παράγοντες.

Abstract

Machine learning offers the retail industry the support to predict successful sales to registered customers. Market Basket Analysis (MBA) is one of the leading applications of machine learning in retail. It helps businesses know what products a customer buys, so that the online store website is designed accordingly. What is mainly studied is the previous purchasing activity of the people. Businesses also use it for cross-selling, suggesting complementary products or services to a customer. This increases the value of the order, as well as customer satisfaction and convenience. Market Basket Analysis (MBA) creates a personalized shopping experience that leads to both short-term revenue growth and long-term customer loyalty.The online food delivery platform studied in this thesis is an excellent example of leveraging Market Basket Analysis (MBA) for the purpose of cross-selling. Along with the purchase of a specific product, the customer is suggested a list of other products that may be of interest to the customer. Which products appear on this list is based on the customer's previous purchasing activity, their search history, what other customers buy in combination with a specific product, and other factors.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Κεφάλαιο 1

ΕΙΣΑΓΩΓΗ

Η εποχή των Μεγάλων Δεδομένων (Big Data) είναι ένα νέο ψηφιακό οικοσύστημα που διαμορφώθηκε αφενός χάρη στην ραγδαία ανάπτυξη τεχνολογιών πληροφορικής και αφετέρου λόγω της αύξησης του όγκου των διαθέσιμων δεδομένων, τα οποία συνιστούν τους κυριότερους πόρους του επιχειρηματικού περιβάλλοντος για την εξαγωγή πολύτιμης πληροφορίας και γνώσης με στόχο την κατανόηση σύνθετων φαινομένων και την λήψη αποφάσεων για την διαχείριση τους. Παράδειγμα αποτελεί η ανάλυση δεδομένων των συναλλαγών στον τομέα του ηλεκτρονικού εμπορίου και των υπηρεσιών διανομής προϊόντων που βοηθά στην κατανόηση των μοτίβων συμπεριφοράς των καταναλωτών και στην πρόβλεψη των προτιμήσεων τους. Με την εξόρυξη δεδομένων μέσω τεχνικών μηχανικής μάθησης καταδεικνύονται πρότυπα αγοραστικής συμπεριφοράς του καταναλωτικού κοινού βοηθώντας στον σχεδιασμό στοχευμένων στρατηγικών marketing από την εκάστοτε επιχείρηση.(Chen et al., 2012, p. 1166· Davenport, 2014, p. 23)

Η παρούσα εργασία αποσκοπεί στην συλλογή και την ανάλυση δεδομένων επώνυμης ηλεκτρονικής πλατφόρμας παραγγελιών προϊόντων ευρείας κατανάλωσης για την εύρεση και κατανόηση της καταναλωτικής συμπεριφοράς των χρηστών της και την εξαγωγή συμπερασμάτων με στόχο την αξιοποίηση τους για τη βελτίωση των επιχειρηματικών διαδικασιών και στρατηγικών μάρκετινγκ. Απαραίτητες τεχνικές για την επίτευξη αυτού του εγχειρήματος είναι η ανάλυση RFM και η MArket Basket Analysis.(Provost & Fawcett, 2013, p. 27· Han et al., 2012, p. 8).

Η ανάλυση RFM (Recency, Frequency, Monetary) επιτρέπει την κατηγοριοποίηση των πελατών με βάση την συχνότητα των αγορών τους (Frequency), το πόσο πρόσφατα αγόρασαν (Recency) και το οικονομικό αποτύπωμα τους (Monetary). Η Market Basket Analysis από την άλλη είναι μια τεχνική εξόρυξης δεδομένων (data mining) και εντοπισμού προϊόντων που τείνουν να αγοράζονται μαζί βοηθώντας στην δημιουργία κανόνων συσχέτισης και αξιοποίησης τους για τη βελτιστοποίηση των πωλήσεων. Η τεχνική αυτή είναι διαδεδομένη σε περιβάλλοντα ηλεκτρονικού εμπορίου και λιανικής πώλησης προσφέροντας προτάσεις στρατηγικών όπως το cross-selling.(Wedel & Kannan, 2016, p. 98).

Ο συνδυασμός των παραπάνω τεχνικών προσφέρει μεγάλο πλεονέκτημα και σημαντικά οφέλη στους επιχειρηματικούς τομείς και ιδιαίτερα στις επιχειρήσεις ηλεκτρονικής διανομής προϊόντων. Η προώθηση προσωποποιημένων προτάσεων συμβάλλει στην βελτίωση της καταναλωτικής εμπειρίας, δημιουργώντας ένα περιβάλλον εμπιστοσύνης, ενισχύοντας την πιστότητα και την ικανοποίηση του καταναλωτή. Παράλληλα, δημιουργείται ανταγωνιστικό πλεονέκτημα και βελτιστοποιείται το κέρδος της επιχείρησης.(Davenport, 2014, p. 45· Chen et al., 2012, p. 1170).

1.1 ΣΚΟΠΟΣ

Σκοπός της εν λόγω διατριβής είναι η ανάλυση της αγοραστικής συμπεριφοράς των πελατών μέσω της αξιοποίησης δεδομένων συναλλαγών από ηλεκτρονική πλατφόρμα παραγγελιών φαγητού.Πιο συγκεκριμένα, η εργασία αποσκοπεί:

- Στην κατανόηση των προτύπων αγοραστικής συμπεριφοράς των πελατών.
- Στην κατηγοριοποίηση των πελατών με βάση τη μεθοδολογία RFM (Recency, Frequency, Monetary).
- Στην εφαρμογή τεχνικών ομαδοποίησης (clustering) για τον εντοπισμό τμημάτων πελατών με παρόμοια χαρακτηριστικά.
- Στην εξαγωγή κανόνων συσχέτισης προϊόντων μέσω της Market Basket Analysis.

- Στην αξιολόγηση των αποτελεσμάτων και τη διατύπωση προτάσεων για τη βελτίωση των στρατηγικών marketing.

Για την επίτευξη των παραπάνω απαιτείται η τμηματοποίηση των πελατών με βάση την καταναλωτική τους συμπεριφορά και το μοτίβο αλληλεπίδρασης με την υπηρεσία, η ανάλυση του προφίλ και των χαρακτηριστικών κάθε τμήματος, η μελέτη των δεδομένων κάθε τμήματος μεμονωμένα καθώς και η κατάταξη των προϊόντων και των υπηρεσιών ανάλογα με τους καταναλωτές κάθε τμήματος. Σε ένα δεύτερο στάδιο θα πραγματοποιηθεί ανάλυση καλαθιού αγορών για περαιτέρω διερεύνηση και τελικά ερμηνεία των αποτελεσμάτων.

Η ανάλυση των παραπάνω στοχεύει στην καλύτερη κατανόηση της πελατειακής βάσης και στην πρόταση πιο αποδοτικών στρατηγικών για την λήψη αποδοτικότερων επιχειρηματικών αποφάσεων.

Στα κεφάλαια που ακολουθούν, αναλύονται βασικές έννοιες που σχετίζονται με τα Μεγάλα Δεδομένα (Big Data), τα χαρακτηριστικά τους και οι εφαρμογές τους σε οργανισμούς, πραγματοποιείται η κύρια ανάλυση των δεδομένων και τελικά παρουσιάζονται τα βασικά συμπεράσματα της μελέτης, καθώς και προτάσεις για περαιτέρω έρευνα και εφαρμογή των αποτελεσμάτων σε πραγματικά επιχειρηματικά περιβάλλοντα.

Κεφάλαιο 2

2.1 ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ (BIG DATA)

Η έννοια των Μεγάλων Δεδομένων (Big Data) δημιουργήθηκε σαν αποτέλεσμα της ραγδαίας ανάπτυξης των τεχνολογιών πληροφορικής και της συνεχώς αυξανόμενης παραγωγής δεδομένων. Από τις πρώτες βάσεις δεδομένων έως τα σύγχρονα κατανεμημένα συστήματα, η ανάγκη για αποθήκευση και επεξεργασία μεγάλου όγκου δεδομένων έχει εξελιχθεί σημαντικά. Κατά τις τελευταίες δεκαετίες, η εξάπλωση του διαδικτύου, των κινητών συσκευών και των κοινωνικών δικτύων οδήγησε σε εκρηκτική αύξηση των δεδομένων. Επιπλέον, η ανάπτυξη τεχνολογιών όπως το Διαδίκτυο των Πραγμάτων (IoT) και οι αισθητήρες συνέβαλε στην παραγωγή δεδομένων σε πραγματικό χρόνο.Η εξέλιξη των Big Data συνοδεύτηκε από την ανάπτυξη νέων τεχνολογιών, όπως τα κατανεμημένα συστήματα αποθήκευσης και επεξεργασίας (π.χ. Hadoop και Spark), τα οποία επιτρέπουν την αποδοτική διαχείριση μεγάλων ποσοτήτων δεδομένων. Πλέον τα Big Data αποτελούν βασικό εργαλείο για την εξαγωγή γνώσης και τη λήψη αποφάσεων σε πολλούς τομείς.

Η εποχή των μεγάλων δεδομένων έχει έρθει για να μείνει. Ο κόσμος των μεγάλων δεδομένων αναφέρεται στον τρόπο με τον οποίο επεξεργαζόμαστε, αναλύουμε και αξιοποιούμε τεράστιες ποσότητες δεδομένων από διάφορες πηγές. Από τις αγορές και την οικονομία, μέχρι την υγεία και την προσωπική μας ζωή, τα μεγάλα δεδομένα είναι πανταχού παρόντα. Ο κόσμος αυτός έχει αλλάξει τον τρόπο με τον οποίο επικοινωνούμε, εργαζόμαστε, ψυχαγωγούμαστε και λαμβάνουμε αποφάσεις. Η συλλογή, ανάλυση και χρήση των μεγάλων δεδομένων έχει καταστεί ουσιώδης για πολλές επιχειρήσεις και οργανισμούς, καθώς τους παρέχει τη δυνατότητα να λαμβάνουν ακριβείς και αποτελεσματικές αποφάσεις, βασισμένες στα δεδομένα και όχι στην υπόθεση. Τα τελευταία χρόνια, η αύξηση των δεδομένων έχει εκρηκτικό ρυθμό και επηρεάζει την οικονομία, την πολιτική, την κοινωνία και την καθημερινότητά μας. Τονίζεται ότι η σημασία των δεδομένων δεν βρίσκεται απλά στον αριθμό τους, αλλά και στην ικανότητά μας να αντλούμε πληροφορίες από αυτά και να τα μετατρέπουμε σε γνώση. Η επεξεργασία των δεδομένων μας επιτρέπει να ανακαλύπτουμε μοτίβα και σχέσεις μεταξύ αυτών των δεδομένων και να εξάγουμε συμπεράσματα για τον κόσμο που μας περιβάλλει.

Οι τεχνολογικές εξελίξεις και η αύξηση των συνδεδεμένων συσκευών έχουν οδηγήσει στην παραγωγή και συλλογή μεγάλων δεδομένων σε κάθε τομέα της ζωής μας. Οι εταιρείες και οργανισμοί συλλέγουν δεδομένα για να κατανοήσουν τις προτιμήσεις και τις συνήθειες των καταναλωτών, ενώ η κυβέρνηση και οι υπηρεσίες παρέχουν υπηρεσίες βασισμένες σε

δεδομένα για να βελτιώσουν την ασφάλεια και την ποιότητα ζωής των πολιτών. Ωστόσο, η συλλογή και ανάλυση τόσων μεγάλων δεδομένων δημιουργεί πολλά ζητήματα που αφορούν την ιδιωτικότητα και την ασφάλεια των δεδομένων. Επιπλέον, η ανάλυση μεγάλων δεδομένων μπορεί να οδηγήσει σε στερεότυπα και αποκλεισμούς, καθώς και σε αποφάσεις που λαμβάνονται από αλγορίθμους χωρίς ανθρώπινη παρέμβαση. Αυτή η κατάσταση οδηγεί σε μια σύγχυση ανάμεσα στο πραγματικό και το εικονικό κόσμο, καθιστώντας τον κόσμο μας πιο σύνθετο και δύσκολο να τον κατανοήσουμε. Τα μεγάλα δεδομένα απαιτούν από εμάς να είμαστε σε θέση να τα αξιοποιούμε με σωστό τρόπο και να αντιμετωπίζουμε τις προκλήσεις που προκύπτουν από αυτά. Η ανάλυση και η εξόρυξη δεδομένων από αυτές τις πηγές μας επιτρέπουν να εξάγουμε πληροφορίες και συμπεράσματα που μπορούν να βοηθήσουν στη λήψη αποφάσεων και την επίλυση προβλημάτων. Ωστόσο, η σωστή επεξεργασία και ανάλυση των δεδομένων απαιτεί εξειδικευμένες γνώσεις και εργαλεία, καθώς και τη συμμετοχή και την υποστήριξη επαγγελματικών ομάδων και οργανισμών. Με τη σωστή χρήση και αξιοποίηση των μεγάλων δεδομένων μπορούμε να βελτιώσουμε την καθημερινότητά μας και να αντιμετωπίσουμε καλύτερα τις προκλήσεις του σύγχρονου κόσμου.

Η αύξηση των δεδομένων τα τελευταία χρόνια έχει γίνει αντιληπτή σε όλους τους τομείς της ζωής μας. Ο αριθμός των δεδομένων που συλλέγονται σε παγκόσμιο επίπεδο αυξάνεται εκθετικά κάθε χρόνο, μερικές φορές ακόμη και κάθε μήνα ή και πιο συχνά. Το 2020, για παράδειγμα, δημιουργήθηκαν περίπου 65 ζεταμπάιτ από δεδομένα, ενώ αναμένεται να δημιουργηθούν περίπου 175 ζεταμπάιτ ανά ημέρα μέχρι το 2025, σύμφωνα με μια έκθεση της IDC. Το 2020, ο όγκος των δεδομένων που δημιουργήθηκαν και αναπαράχθηκαν έφτασε σε νέο υψηλό. Η ανάπτυξη ήταν υψηλότερη από ό,τι αναμενόταν προηγουμένως λόγω της αυξημένης ζήτησης λόγω της πανδημίας COVID-19, καθώς περισσότεροι άνθρωποι εργάζονταν και μάθαιναν από το σπίτι και χρησιμοποιούσαν πιο συχνά επιλογές οικιακής ψυχαγωγίας.

Η αυξημένη χρήση των κινητών συσκευών, των αισθητήρων και των συστημάτων παρακολούθησης, καθώς και η υπερβολική παραγωγή και κατανάλωση ψηφιακού περιεχομένου, όπως βίντεο, εικόνες και ήχοι, έχουν οδηγήσει σε μια αύξηση του όγκου των δεδομένων. Αυτό όμως έχει αναδείξει σοβαρά ζητήματα σχετικά με την ιδιωτικότητα και την ασφάλεια των δεδομένων. Υπάρχουν ανησυχίες ότι η συλλογή και η ανάλυση των δεδομένων μπορεί να οδηγήσει σε παραβίαση της ιδιωτικότητας των ατόμων, ενώ επίσης μπορεί να δημιουργήσει ανισότητες στην πρόσβαση στις υπηρεσίες και την επιρροή που ασκούν στους ανθρώπους. Επιπλέον, η αποθήκευση και η επεξεργασία των δεδομένων μπορεί να έχει σοβαρές συνέπειες στην ασφάλεια και την ακεραιότητα τους, με τον κίνδυνο της κλοπής και της κακόβουλης χρήσης τους. Όλα αυτά αποδεικνύουν ότι η διαχείριση των μεγάλων δεδομένων είναι ένα ζήτημα που απαιτεί στρατηγική σκέψη και αντίληψη για τις συνέπειες που μπορεί να έχει στην κοινωνία.

2.2 ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ BIG DATA (5VS)

Τα 5 V των μεγάλων δεδομένων αποτελούν ένα σύνολο χαρακτηριστικών που περιγράφουν τη φύση των μεγάλων δεδομένων και πώς αυτά πρέπει να διαχειρίζονται. Τα 5 V είναι:

1. Όγκος (Volume): Αναφέρεται στο μέγεθος των δεδομένων. Η αύξηση του όγκου των δεδομένων έχει οδηγήσει στην ανάγκη για νέες τεχνολογίες αποθήκευσης και επεξεργασίας δεδομένων.
2. Ταχύτητα (Velocity): Αναφέρεται στην ταχύτητα με την οποία παράγονται, συλλέγονται και αναλύονται τα δεδομένα. Η ταχύτητα είναι ένα σημαντικό χαρακτηριστικό για τις επιχειρήσεις που χρειάζονται άμεση ανταπόκριση στα δεδομένα.
3. Ποικιλία (Variety): Αναφέρεται στο εύρος των δεδομένων και τις διαφορετικές πηγές από τις οποίες προέρχονται. Η ποικιλία των δεδομένων μπορεί να είναι δομημένη ή μη δομημένη και μπορεί να προέρχεται από διάφορες πηγές, όπως κοινωνικά δίκτυα, αισθητήρες και κινητά τηλέφωνα.

4. Ακρίβεια (Veracity): Αναφέρεται στην ακρίβεια των δεδομένων, δηλαδή στο πόσο αξιόπιστα και ακριβή είναι τα δεδομένα που συλλέγονται και πώς μπορούν να χρησιμοποιηθούν με ασφάλεια.
5. Αξία (Value): Αναφέρεται στην αξία των δεδομένων, δηλαδή στο πόσο χρήσιμα είναι τα δεδομένα για την επίλυση προβλημάτων, την πρόβλεψη τάσεων και την λήψη αποφάσεων.

Η ανάπτυξη των μεγάλων δεδομένων δεν φαίνεται να έχει τερματιστεί, καθώς η τεχνολογική εξέλιξη συνεχίζεται με φρενήρεις ρυθμούς. Αναμένεται να συλλέγονται ακόμα περισσότερα δεδομένα από διαφορετικές πηγές, όπως οι συσκευές IoT, οι κοινωνικές δικτυακές πλατφόρμες, οι αισθητήρες και άλλες συσκευές που συνδέονται στο διαδίκτυο. Αυτή η συλλογή δεδομένων αναμένεται να έχει τεράστιες επιπτώσεις στην κοινωνία μας, καθώς οι εταιρείες και οι οργανισμοί θα έχουν ακόμα περισσότερη πρόσβαση σε προσωπικά δεδομένα και θα μπορούν να αναλύουν τη συμπεριφορά των ανθρώπων με ακόμα μεγαλύτερη ακρίβεια. Παράλληλα, η ανάπτυξη της τεχνητής νοημοσύνης και του machine learning αναμένεται να οδηγήσει σε ακόμα πιο αποτελεσματικές μεθόδους επεξεργασίας και ανάλυσης των δεδομένων.

Η βιομηχανική επανάσταση 4.0 είναι μια εξέλιξη της βιομηχανίας που χαρακτηρίζεται από τη χρήση τεχνολογιών όπως το Διαδίκτυο των Πραγμάτων (IoT), την Τεχνητή Νοημοσύνη (AI) και τα Μεγάλα Δεδομένα (Big Data) για την αναβάθμιση των βιομηχανικών διεργασιών και τη βελτίωση της παραγωγικότητας. Τα Μεγάλα Δεδομένα είναι ένα σημαντικό στοιχείο της βιομηχανικής επανάστασης 4.0 και χρησιμοποιούνται για τη συλλογή και ανάλυση των δεδομένων που παράγονται από τα συστήματα παραγωγής και τα προϊόντα. Αυτή η ανάλυση των δεδομένων μπορεί να οδηγήσει σε βελτιώσεις στην παραγωγή, στην ποιότητα των προϊόντων και στη μείωση των δαπανών. Επιπλέον, τα Μεγάλα Δεδομένα μπορούν να χρησιμοποιηθούν για τη βελτίωση της συντήρησης των μηχανημάτων και των εξοπλισμών, μέσω της παρακολούθησης της λειτουργίας τους και της αναγνώρισης προβλημάτων που χρειάζονται προληπτική συντήρηση. Τα Μεγάλα Δεδομένα μπορούν επίσης να χρησιμοποιηθούν για τη βελτίωση της επικοινωνίας μεταξύ των συστημάτων παραγωγής και των πελατών, μέσω της συλλογής και ανάλυσης δεδομένων για την κατανόηση της συμπεριφοράς των πελατών και την πρόβλεψη των αναγκών τους. Ωστόσο, η χρήση των Μεγάλων Δεδομένων δεν είναι απλώς μια τεχνική πτυχή της βιομηχανικής επανάστασης 4.0, αλλά απαιτεί και την αλλαγή του τρόπου σκέψης των επιχειρήσεων και την προσαρμογή τους σε νέες διεργασίες και τεχνολογίες. Αυτό απαιτεί συχνά την εκπαίδευση και την επαγγελματική ανάπτυξη των εργαζομένων σε νέες δεξιότητες και ικανότητες. Συνολικά, τα Μεγάλα Δεδομένα είναι ένα σημαντικό στοιχείο της βιομηχανικής επανάστασης 4.0 και μπορούν να βοηθήσουν τις επιχειρήσεις να βελτιώσουν την απόδοσή τους και να προσαρμοστούν σε μια καινοτόμα και ανταγωνιστική αγορά.

2.3 ΕΦΑΡΜΟΓΕΣ BIG DATA ΣΕ ΟΡΓΑΝΙΣΜΟΥΣ

Η γνώση που παρέχεται από την συλλογή και την επεξεργασία των Μαζικών Δεδομένων μπορεί να προσφέρει οφέλη σε σχεδόν οποιαδήποτε επιχείρηση ή βιομηχανία. Παρόλα' αυτά, οι πιο μεγάλοι οργανισμοί με πολυπλοκότερες επιχειρησιακές αρμοδιότητες είναι συχνά σε θέση να κάνουν την πιο ουσιαστική χρήση των Μεγάλων Δεδομένων και ωφελούνται περισσότερο.

Οικονομικά: Μια μελέτη του 2020 (Journal of Big Data) επισημαίνει ότι τα Μεγάλα Δεδομένα «παίζουν σημαντικότατο ρόλο στην αλλαγή του τομέα των χρηματοπιστωτικών υπηρεσιών, και κυρίως στο εμπόριο και τις επενδύσεις, τη φορολογική μεταρρύθμιση, τον εντοπισμό και τη διερεύνηση της απάτης, την ανάλυση κινδύνων και την αυτοματοποίηση». Τα Big Data βοήθησαν επίσης στον μετασχηματισμό της χρηματοοικονομικής βιομηχανίας αναλύοντας τα δεδομένα πελατών και τα σχόλια για να αποκτήσουν τις πολύτιμες πληροφορίες που απαιτούνται για τη βελτίωση της ικανοποίησης και της εμπειρίας των πελατών. Τα itemsets συναλλαγών είναι μερικά από τα μεγαλύτερα και ταχύτερα κινούμενα στον κόσμο. Η στροφή σε όλο και πιο προηγμένες λύσεις διαχείρισης μεγάλων δεδομένων θα συνδράμει ώστε οι τράπεζες και τα χρηματοπιστωτικά ιδρύματα να προστατεύσουν αυτά τα δεδομένα και να τα

χρησιμοποιήσουν με τρόπους που ωφελούν και προστατεύουν τόσο τον πελάτη όσο και την επιχείρηση.

Υγειονομική Περίθαλψη: Η ανάλυση μεγάλων δεδομένων βοηθά τους επαγγελματίες υγείας να κάνουν περισσότερο ακριβείς και πιο τεκμηριωμένες διαγνώσεις. Επιπλέον, τα Big Data βοηθούν τους διαχειριστές νοσοκομείων να εντοπίσουν τις τάσεις, να διαχειριστούν τους κινδύνους και να ελαχιστοποιήσουν τις περιττές δαπάνες – οδηγώντας τους υψηλότερους δυνατούς προϋπολογισμούς σε τομείς της φροντίδας και της έρευνας των ασθενών.

Μεταφορά και Εφοδιαστική: Το Φαινόμενο Amazon είναι ένας όρος που περιγράφει τον τρόπο με τον οποίο η Amazon έχει θέσει τη γραμμή για τις προσδοκίες παράδοσης της επόμενης ημέρας, όπου οι πελάτες απαιτούν πλέον αυτού του είδους την ταχύτητα αποστολής για οτιδήποτε παραγγείλουν ηλεκτρονικά. Το περιοδικό Business επισημαίνει ότι ως άμεσο αποτέλεσμα του Φαινομένου Amazon, “ο αγώνας εφοδιαστικής του τελευταίου μιλίου θα αυξηθεί περισσότερο ανταγωνιστικός”. Οι εταιρείες εφοδιαστικής βασίζονται όλο και περισσότερο στα Big Data analytics για να βελτιστοποιήσουν τον προγραμματισμό διαδρομών, την ενοποίηση φορτίων και τα μέτρα απόδοσης καυσίμου.

Εκπαίδευση Κατά τη διάρκεια της πανδημίας, τα εκπαιδευτικά ιδρύματα σε όλο τον κόσμο χρειάστηκε να επανεφεύρουν τα προγράμματα σπουδών τους και τις μεθόδους διδασκαλίας τους για να υποστηρίξουν την απομακρυσμένη μάθηση. Μια σημαντική πρόκληση σε αυτή τη διαδικασία είναι η εξεύρεση αξιόπιστων τρόπων ανάλυσης και αξιολόγησης της απόδοσης των μαθητών και η συνολική αποτελεσματικότητα των μεθόδων διδασκαλίας στο διαδίκτυο. Ένα άρθρο του 2020 σχετικά με τον αντίκτυπο των Μεγάλων Δεδομένων στην εκπαίδευση και την ηλεκτρονική μάθηση κάνει μια παρατήρηση σχετικά με τους εκπαιδευτικούς: «Τα μεγάλα δεδομένα τους κάνουν να αισθάνονται πολύ πιο σίγουροι για την εξατομίκευση της εκπαίδευσης, την ανάπτυξη μικτής μάθησης, τη μετατροπή των συστημάτων αξιολόγησης και την προώθηση της δια βίου μάθησης».

Ενέργεια και Επιχειρήσεις Κοινής Ωφελείας Σύμφωνα με τις ΗΠΑ. Γραφείο Στατιστικών Εργασίας, εταιρείες κοινής ωφέλειας δαπανούν πάνω από 1,4 δισεκατομμύρια δολάρια ηπα σε αναγνώστες μετρητών και συνήθως βασίζονται σε αναλογικούς μετρητές και σπάνιες χειροκίνητες αναγνώσεις. Οι έξυπνοι αναγνώστες μετρητών παραδίδουν τα ψηφιακά στοιχεία πολλές φορές την ημέρα και, με το όφελος της ανάλυσης μεγάλων δεδομένων, αυτό μπορεί να ενημερώσει την αποδοτικότερη ενεργειακή χρήση και την ακριβέστερη τιμολόγηση και πρόβλεψη. Επιπλέον, όταν οι εργαζόμενοι στον τομέα απελευθερώνονται από την καταμέτρηση, η καταγραφή και η ανάλυση δεδομένων μπορούν να βοηθήσουν πιο γρήγορα στην ανακατανομή τους εκεί όπου χρειάζονται επειγόντως επισκευές και αναβαθμίσεις.

Τα Big Data έχουν σημαντικές εφαρμογές σε ένα ευρύ φάσμα οργανισμών και επιχειρήσεων. Στον τομέα του λιανικού εμπορίου, η ανάλυση δεδομένων επιτρέπει την κατανόηση της συμπεριφοράς των πελατών και τη βελτιστοποίηση των στρατηγικών marketing. Μέσω της αξιοποίησης δεδομένων συναλλαγών, οι επιχειρήσεις μπορούν να εντοπίσουν πρότυπα αγοραστικής συμπεριφοράς, να προβλέψουν τη ζήτηση προϊόντων και να δημιουργήσουν εξατομικευμένες προτάσεις. Επιπλέον, η ανάλυση δεδομένων δρα επικουρικά στη βελτίωση της διαχείρισης αποθεμάτων και την αλυσίδα ανεφοδιασμού. Στον χρηματοοικονομικό τομέα, τα Big Data χρησιμοποιούνται για την ανίχνευση απάτης, την αξιολόγηση κινδύνου και την ανάλυση επενδυτικών τάσεων. Στον τομέα της υγείας, συμβάλλουν στη διάγνωση ασθενειών και στη βελτίωση της παροχής υπηρεσιών υγείας. Η αξιοποίηση των Big Data επιτρέπει στους οργανισμούς να λαμβάνουν αποφάσεις βασισμένες σε δεδομένα (data-driven decision making), ενισχύοντας την αποδοτικότητα και την ανταγωνιστικότητά τους.

2.4 ΕΠΑΓΓΕΛΜΑΤΑ ΣΧΕΤΙΖΟΜΕΝΑ ΜΕ BIG DATA ΚΑΙ ΠΟΙΟΣ Ο ΡΟΛΟΣ ΤΟΥ DATA SCIENTIST

Η εξέλιξη του κλάδου των Big Data έχει οδηγήσει στην ανάγκη για δημιουργία νέων επαγγελματικών ρόλων, οι οποίοι επικεντρώνονται στη τόσο στη συλλογή,όσο και στην

επεξεργασία και την ανάλυση δεδομένων. Τέτοιοι είναι ο Data Scientist, ο Data Analyst και ο Machine learning engineer.

Ο ρόλος του Data Scientist είναι να ασχολείται με τη συλλογή, την οργάνωση και την ανάλυση δεδομένων, με τελικό σκοπό να εντοπίζει τάσεις ή χρήσιμες πληροφορίες που δεν είναι ευθέως αναγνωρίσιμες. Συνήθως αρχίζει με τη συγκέντρωση όλων των δεδομένων που χρειάζεται και συνεχίζει με την προετοιμασία και τον “καθαρισμό” τους, ώστε να είναι κατάλληλα για ανάλυση. Συνεχίζει εφαρμόζοντας διάφορες τεχνικές και μοντέλα, με στόχο να ανιχνεύσει μοτίβα ή να κάνει προβλέψεις, που μπορούν να βοηθήσουν την επιχείρηση να κατανοήσει καλύτερα τους πελάτες της, τις ανάγκες της αγοράς ή την πορεία ενός προϊόντος.

Με την διαχείριση και επεξεργασία δεδομένων ασχολείται και ο Data Analyst, ωστόσο οι δύο ρόλοι δεν πρέπει να συγχέονται. Ο τρόπος με τον οποίο προσεγγίζουν την πληροφορία και το είδος των αποφάσεων που μπορούν να υποστηρίξουν είναι αρκετά διαφορετικός. Για να γίνει πιο ξεκάθαρο το πλαίσιο, θα πρέπει να εξετάσουμε τι είναι ένας Data Analyst, αφού ο εν λόγω επαγγελματίας εστιάζει κυρίως στη μελέτη των δεδομένων που υπάρχουν ήδη, στην εξαγωγή βασικών συμπερασμάτων και στη δημιουργία αναφορών που περιγράφουν την τρέχουσα εικόνα μιας επιχείρησης. Ο Data Scientist όμως προχωρά ένα βήμα παραπέρα. Χτίζει μοντέλα που μπορούν να προβλέψουν το μέλλον και να υποστηρίξουν στρατηγικές αποφάσεις. Με πιο απλά λόγια ο Data Analyst εξηγεί το παρελθόν και το παρόν ενώ ο Data Scientist προσπαθεί να δείξει τι μπορεί να συμβεί στο μέλλον.

Οι προγραμματιστές μηχανικής μάθησης αναπτύσσουν στρατηγικές βασισμένοι στην ανάλυση δεδομένων με MBA, βοηθώντας τις επιχειρήσεις να αυξήσουν τα έσοδά τους με τους ακόλουθους τρόπους:

- Με την προσφορά έκπτωσης σε ένα από τα σχετικά προϊόντα
- Τα σχετικά προϊόντα τοποθετούνται το ένα κοντά στο άλλο έτσι ώστε ότα ο πελάτης αγοράσει το ένα να επηρεαστεί και να αγοράσει και το άλλο

Κεφάλαιο 3

3.1 CUSTOMER SEGMENTATION

Η τμηματοποίηση πελατών αποτελεί μία θεμελιώδη διαδικασία στο συγχρονο μαρκετινγκ κατά την οποία οι πελάτες μπορούν ομαδοποιούνται σε επιμέρους κατηγορίες με βάση τα χαρακτηριστικά που παρουσιάζουν ομοιότητες. Η βασικη αρχη στηρίζεται στο οτι οι αγορες που μελετάμε δεν είναι ομοιογενείς και διαφέρουν σημαντικά μεταξύ τους. Η διαδικασία επιτρέπει στις επιχειρήσεις να εντοπίσουν πιο αποτελεσματικά ομάδες ανθρώπων οι οποίοι κατέχουν κοινές συμπεριφορές και να επενδύσουν περισσότερους πόρους σε αυτούς.

Η τμηματοποίηση των πελατών έγινε μέσω της διαδικασίας STP (Segmentation, Targeting, Positioning) που στηρίζεται σε ποικίλους παράγοντες όπως δημογραφικά , γεωγραφικά, ψυχογραφικά και συμπεριφορικά δεδομένα. Συγκεκριμένα, η δημογραφική τμηματοποίηση λαμβάνει υπόψη μεταβλητές όπως η ηλικία, το φύλο και το εισόδημα. Η γεωγραφική στηρίζεται στην τοποθεσία δράσης των καταναλωτών και τα χωρικά όρια που αυτοί έγκενται. Η ψυχογραφική μελετά δεδομένα κοινωνικού υπόβαθρου όπως οι αξίες , ο τρόπος ζωής και τα ενδιαφέροντα του καταναλωτή ενώ η συμπεριφορική τμηματοποίηση επικεντρώνεται σε καταναλωτικά πρότυπα, διακυμάνσεις πιστότητας και γενικότερα στη στάση του καταναλωτή απέναντι σε ένα προϊόν. (Wedel, M., & Kamakura, W. (2000). Market Segmentation: Conceptual and Methodological Foundations. Springer.)

Η τμηματοποίηση των πελατών έχει εξελιχθεί σημαντικά στις μέρες μας λόγω του μεγάλου όγκου δεδομένων που παρέχονται από σύγχρονες αναλύσεις και της προόδου των τεχνικών των αναλύσεων αυτών. Η χρήση δεδομένων big data, οι αλγόριθμοι και οι εφαρμογές τεχνητής νοημοσύνης έχουν υποβοηθήσει σημαντικά τις επιχειρήσεις , προσφέροντας μια πιο καθαρή εικόνα των σύνθετων προτύπων καταναλωτικής συμπεριφοράς , δίνοντας την

δυνατότητα εξέλιξης δυναμικών και προσαρμοσμένων τμημάτων αγοράς.(Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly.) που προσφέρουν εξατομικευμένη εμπειρία κατανάλωσης στοχευοντας τελικά στην ενίσχυση της σχέσης επιχείρησης- καταναλωτή. Συνιστά επίσης σημαντικό εργαλείο για το marketing και την ανάπτυξη στρατηγικών με μακροπρόθεσμα αποτελέσματα στην μείωση του ανταγωνισμού και την σύνδεση του προϊόντος με τους πελάτες. Συνδέεται έτσι άμεσα με την διαχείριση των σχέσεων πελατών (Customer Relationship Management – CRM), διευκολύνοντας την αναγνώριση των πλέον πολύτιμων και πιστών ακολούθων του προϊόντος ,προτείνοντας πρωτοβουλίες που ενδυναμώνουν τη διατήρηση και την αφοσίωσή τους.(Payne, A., & Frow, P. (2005). A Strategic Framework for Customer Relationship Management. Journal of Marketing.)

Για την ορθή τμηματοποίηση των καταναλωτών προτιμώνται μεθοδολογικές προσεγγίσεις όπως αλγόριθμοι συσχέτισης, ταξινόμησης, παλινδρόμησης και ομαδοποίησης. Η ανάλυση RFM συνιστά την πιο απλή αλλά αποδοτική μέθοδο προσέγγισης ειδικά σε περιβάλλοντα με επαρκή δεδομένα καταναλωτικών συναλλαγών ενώ οι τελευταίες είναι οι πλέον αποτελεσματικές καθώς προσφέρουν τη δυνατότητα ανακάλυψης φυσικών ομάδων μέσα στο δεδομένα χωρίς να έχει προηγηθεί καθορισμένη κατηγοριοποίηση. Στην παρούσα εργασία , επιλέχθηκε για την τμηματοποίηση των πελατών η μεθοδολογία RFM καθώς παρέχει σαφώς ερμηνεύσιμα και επιχειρησιακά αξιοποιήσιμα αποτελέσματα και είναι κατάλληλη για το υπό εξέταση σύνολο δεδομένων δίνοντας μια πληρέστερη εικόνα όπως θα φανεί στο επόμενο κεφάλαιο.

3.1.1 RFM Analysis

Η ανάλυση RFM (Recency, Frequency, Monetary) θεωρείται η πλέον καθιερωμένη και αποτελεσματική μεθοδολογία για την ανάλυση της καταναλωτικής συμπεριφοράς αξιοποιώντας δεδομένα από το ιστορικών των συναλλαγών για τον υπολογισμό του οικονομικού αποτυπώματος και της αξίας κάθε πελάτη.Ο όρος Recency είναι το χρονικό διάστημα που έχει μεσολαβήσει από την τελευταία αγορά ενός καταναλωτή, καταδεικνύοντας το επίπεδο και το χρονικό πλαίσιο της πρόσφατης δραστηριότητάς του.(Bult, J. R., & Wansbeek, T. (1995). Optimal selection for direct mail. Marketing Science.) Ο όρος Frequency αναφέρεται στην συχνότητα, δηλαδή στον αριθμό των αγορών που έχει πραγματοποιήσει ένας πελάτης σε συγκεκριμένο χρονικό πλαισιο, αντανακλώντας τη συνέπεια του συγκεκριμένου πελάτη ως προς το προιον. Τέλος, ο όρος Monetary εκφράζει τη συνολική χρηματική αξία των συναλλαγών του εκάστοτε αγοραστή αποκαλύπτοντας τη συνεισφορά του στα συνολικά έσοδα της επιχείρησης.

Πλέον, η μέθοδος RFM έχει καθιερωθεί ως το κύριο εργαλείο του Customer Relationship Management (CRM), καθώς συνιστά σημαντική πηγή δεδομένων , παρέχοντας δυνατότητες όπως ο εντοπισμός πελατών υψηλής αξίας, η αναγνώριση πιστών και συχνά ενεργών πελατών αλλά και αντίθετα αυτών που παρουσιάζουν μειωμένη δραστηριότητα και ενδέχεται να απαιτούν στοχευμένες ενέργειες επαναδραστηριοποίησης. Με τον τρόπο αυτό οι επιχειρήσεις μπορούν να μεταβάλλουν τη στρατηγική marketing τους από μια γενική προσέγγιση σε πιο εξατομικευμένη πελατοκεντρική με στόχο την αύξηση πόρων.

Η εφαρμογή αυτής της ανάλυσης στηρίζεται στην μετατροπή των τριών αυτών βασικών μεταβλητών σε βαθμολογικές κατηγορίες (stores). Οι κατηγορίες αυτές εν συνεχεία συνδυάζονται κατάλληλα για τη δημιουργία διακριτών τμημάτων πελατών που φέρουν κοινά αγοραστικά χαρακτηριστικά, διευκολύνοντας έτσι την αναγνώριση προτύπων καταναλωτικής συμπεριφοράς.

Η ανάλυση αυτή σε συνδυασμό με άλλες αναλυτικές τεχνικές συνιστά το ιδανικό εργαλείο σε περιβάλλοντα ηλεκτρονικού εμπορίου λόγω της απλότητας της, της πρακτικής εφαρμογής της καθώς και της σαφήνειας των αποτελεσμάτων της. Σε τέτοια πλαίσια, όπου τα δεδομένα συναλλαγών είναι άμεσα διαθέσιμα η εφαρμογή της συμβάλλει ουσιαστικά στη βελτιστοποίηση της εμπειρίας του πελάτη, στη διαμόρφωση προσωποποιημένων προσφορών και στη μεγιστοποίηση της αξίας διάρκειας ζωής πελάτη (Customer Lifetime Value – CLV).

Στην παρούσα εργασία η τεχνική αυτή ανταποκρίνεται πλήρως στα χαρακτηριστικά της διαθέσιμης δεξαμενής δεδομένων γι αυτό και επιλέχθηκε ως εργαλείο τμηματοποίησης των πελατών.

3.1.2 Clustering Algorithms

Οι αλγόριθμοι ομαδοποίησης (clustering) είναι ένα εργαλείο ανάλυσης δεδομένων μη επιβλεπόμενης μηχανικής μάθησης που χρησιμοποιούνται τον διαχωρισμό ενός συνόλου δεδομένων σε ομάδες με βάση την ομοιότητα των χαρακτηριστικών τους . Τα δεδομένα της ίδιας ομάδας θα πρέπει να είναι όσο το δυνατόν πιο όμοια μεταξύ τους ενώ οι διαφορές μεταξύ διαφορετικών ομάδων θα πρέπει να είναι φανερές (Jain, 2010).

Στην τμηματοποίηση πελατών, τα καταναλωτικά πρότυπα όπως οι αγοραστικές συνήθειες, τα επίπεδα δαπάνης και η συχνότητα αλληλεπίδρασης με την επιχείρηση χρησιμοποιούνται για τον διαχωρισμό των ομάδων πελατών. (Wedel & Kamakura, 2000). Η επιλογή του κατάλληλου αλγορίθμου clustering εξαρτάται από τη φύση των δεδομένων και τους στόχους της ανάλυσης. Ο αλγόριθμος *K-means* αποτελεί μία από τις πιο διαδεδομένες μεθόδους, καθώς έχει υψηλή υπολογιστική απόδοση και θεωρείται κατάλληλος για μεγάλα σύνολα αριθμητικών δεδομένων, επιδιώκοντας τον σχηματισμό συμπαγών και διακριτών ομάδων (MacQueen, 1967). Άλλοι αλγόριθμοι από την άλλη, όπως οι ιεραρχικές μέθοδοι (*hierarchical clustering*) επιτρέπουν τη σταδιακή δημιουργία ομάδων προσφέροντας οπτική αναπαράσταση μέσω δενδροδιαγραμμάτων. Διευκολύνουν έτσι, την ερμηνεία της δομής των δεδομένων (Aggarwal & Reddy, 2014). Αλγόριθμοι όπως το *DBSCAN* χρησιμοποιούνται για τον εντοπισμό ομάδων αυθαίρετου σχήματος σε πιο σύνθετα δεδομένα (Ester et al., 1996). Η επιτυχία της ομαδοποίησης εξαρτάται σε μεγάλο βαθμό από τα στάδια προεπεξεργασίας των δεδομένων, όπως η κανονικοποίηση των μεταβλητών και η επιλογή κατάλληλων χαρακτηριστικών, τα οποία επηρεάζουν άμεσα την απόδοση των αλγορίθμων (Han, Kamber & Pei, 2011). Τα αποτελέσματα αξιολογούνται με τη χρήση δεικτών όπως ο *Silhouette Score*, που αποτιμά την ποιότητα του διαχωρισμού μεταξύ διαφορετικών ομάδων.

Στην επόμενη ενότητα παρουσιάζονται αναλυτικά βασικοί αλγόριθμοι clustering που εφαρμόστηκαν στην παρούσα εργασία, καθώς και τα κριτήρια επιλογής τους με βάση τα χαρακτηριστικά του εξεταζόμενου συνόλου δεδομένων.

3.1.2.1 K-Means Algorithm

Ο αλγόριθμος *K-Means* χρησιμοποιείται στην ανάλυση δεδομένων. Συνιστά μια μέθοδο βασισμένη στον διαχωρισμό των παρατηρήσεων σε έναν προκαθορισμένο αριθμό ομάδων (*clusters*).Ο στόχος είναι να μειωθεί όσο το δυνατόν περισσότερο η διακύμανση μέσα σε κάθε ομάδα, δηλαδή το άθροισμα των αποστάσεων των σημείων από το κέντρο (centroid) της ομάδας στην οποία ανήκουν. Με αυτόν τον τρόπο προκύπτουν συστάδες που είναι πιο συνεκτικές και ομοιογενείς (James MacQueen, 1967).

Η λειτουργία του αλγορίθμου K-Means περιλαμβάνει δύο φάσεις. Η πρώτη συνιστά την αρχικοποίηση και την επαναληπτική βελτιστοποίηση κατά την οποία επιλέγονται τυχαία K σημεία από το σύνολο δεδομένων, τα οποία χρησιμοποιούνται ως αρχικά κεντροειδή (centroid). Ο αριθμός των συστάδων K οφείλει να έχει καθοριστεί εκ των προτέρων καθώς επηρεάζει σημαντικά την ποιότητα των αποτελεσμάτων.Εξίσου σημαντική είναι και η αρχική επιλογή των κεντροειδών. Διαφορετικές αρχικοποιήσεις ενδέχεται να οδηγήσουν σε διαφορετικές τελικές λύσεις. (Arthur & Vassilvitskii, 2007). Στη δεύτερη φάση, κάθε παρατήρηση αντιστοιχίζεται στο πλησιέστερο κεντροειδές, συνήθως με βάση την ευκλείδεια απόσταση. Στη συνέχεια, τα κεντροειδή επαναϋπολογίζονται ως ο μέσος όρος των σημείων που ανήκουν σε κάθε συστάδα και η διαδικασία επαναλαμβάνεται διαδοχικά, εναλλάσσοντας τα βήματα ανάθεσης και ενημέρωσης, έως ότου επιτευχθεί σύγκλιση, μια φάση όπου δεν παρατηρούνται σημαντικές αλλαγές στις αναθέσεις ή στις θέσεις των κεντροειδών.Η υπολογιστική αποδοτικότητα και η ευκολία υλοποίησης είναι βασικά πλεονεκτήματα του K-Means και το συνιστούν κατάλληλο για μεγάλα σύνολα δεδομένων..

3.1.2.2 DBSCAN Algorithm

Ο DBSCAN είναι ένας μη παραμετρικός αλγόριθμος ομαδοποίησης βασισμένος στην πυκνότητα. Εντοπίζει ομάδεσ δεδομένων σε περιοχές υψηλής συγκέντρωσης στοιχείων και χαρακτηρίζει τα υπόλοιπα σημεία ως θόρυβο (*noise*) (Ester et al., 1996). Μπορεί να εντοπίσει συστάδεσ ακαθόριστου σχήματος και δεν απαιτεί προκαθορισμένο αριθμό ομάδων σε αντίθεση με τον K-means.

Η λειτουργία του αλγορίθμου βασίζεται σε δύο βασικές παραμέτρους, στην ακτίνα γειτονιάς και τον ελάχιστο αριθμό σημείων (*MinPts*) που απαιτούνται για τον σχηματισμό μιας πυκνής περιοχής. Με βάση αυτές τις παραμέτρους, τα σημεία ταξινομούνται σε τρεις κατηγορίες: σημεία πυρήνα (*core points*), με επαρκή αριθμό γειτόνων εντός της ακτίνας , συνοριακά σημεία (*border points*), εντοπισμένα κοντά σε πυκνές περιοχές χωρίς να πληρούν το κριτήριο πυρήνα, και σημεία θορύβου (*noise*), που δεν ανήκουν σε καμία συστάδα.

Η ομαδοποίηση ξεκινά επιλέγοντας ένα τυχαίου σημείου και στην συνέχεια εξετάζοντασ την γειτονιά του. Εάν το σημείο αποτελεί σημείο πυρήνα, δημιουργείται μία νέα συστάδα και επεκτείνεται προσθέτοντας όλα τα σημεία που είναι στενά συνδεδεμένα με αυτό. Η διαδικασία επαναλαμβάνεται διαδοχικά μέχρι το σημείο κορεσμού όπου δεν μπορούν να προστεθούν νέα σημεία στη συστάδα. Ο αλγόριθμος εξετάζει τα εναπομείναντα μη ταξινομημένα σημεία, επαναλαμβάνοντας τη διαδικασία μέχρι την ομαδοποίηση όλων των δεδομένων.

Βασικό πλεονέκτημα του DBSCAN είναι η ικανότητά εντοπισμού συστάδων αυθαιρέτου σχήματος, και η αποτελεσματική διαχείριση δεδομένων που περιέχουν θόρυβο ή ακραίες τιμές. Επιπλέον, δεν επηρεάζεται από την αρχικοποίηση, σε αντίθεση με τον K-Means, γεγονός που οδηγεί σε πιο σταθερά αποτελέσματα.

Συνολικά, ο DBSCAN είναι ένας ισχυρός και ευέλικτος αλγόριθμος ομαδοποίησης, κατάλληλος για εφαρμογές όπου οι συστάδες δεν έχουν κανονικό σχήμα και τα δεδομένα περιλαμβάνουν θόρυβο. Για τον λόγο αυτό, χρησιμοποιείται ευρέως σε πεδία ανάλυσης χωρικών δεδομένων, ανίχνευσης ανωμαλιών και τμηματοποίησης πελατών με σύνθετα καταναλωτικά πρότυπα συμπεριφοράς.

3.1.2.4 Gaussian Mixture Model (GMM)

Το *Gaussian Mixture Model (GMM)* είναι μία πιθανοκρατική μέθοδος μέγιστης σημασίας για την ομαδοποίηση και την αναγνώριση καταναλωτικών προτύπων. Απαραίτητη προυπόθεση η οποία προυποθέτει ότι τα δεδομένα προέρχονται από συνδυασμό πολλαπλών κανονικών κατανομών, καθεμία εκ των οποίων αντιστοιχεί σε μία συστάδα (Bishop, 2006). Για την εκτίμηση παραμέτρων ενός μείγματος μπορουν να χρησιμοποιηθούν προσεγγίσεις όπως γραφικοί μέθοδοι, αντιστοίχιση ροπών, μπεϋζιανές προσεγγίσεις και η μέθοδος MLE. Οι παράμετροι του εν λόγω μοντέλου εκτιμώνται μέσω του αλγορίθμου *Expectation-Maximization (EM)*, ο οποίος ξεκινώντας από ένα αρχικό μοντέλο μίγματος επαναληπτικά υπολογίζει τις πιθανότητες συμμετοχής των δεδομένων (Expectation) και ενημερώνει τις παραμέτρους των κατανομών (Maximization), έως ότου επιτευχθεί σύγκλιση (Dempster et al., 1977).

Ο GMM έχει ως κύριο πλεονέκτημα την ευελιξία. Μπορεί να μοντελοποιήσει συστάδες με διαφορετικά σχήματα και διακυμάνσεις, δίνοντας μια πιο ρεαλιστική αποτύπωση σύνθετων δεδομένων (McLachlan & Peel, 2000). Ωστόσο, προκειμένου να επιτευχθεί η σύγκλιση του αλγόριθμου αυτού απαιτείται προκαθορισμός του αριθμού των συστάδων.

3.2 DATA MANAGEMENT

Στο κεφάλαιο αυτό θα ασχοληθούμε με την διαχείριση των δεδομένων (Data Management), το κεφάλαιο αυτο αποτελεί κρίσιμο μέρος της διαδικασίας, καθώς μας επιτρεπει να διασφαλίσουμε

την ποιότητα και την αξιοπιστία των δεδομένων που χρησιμοποιούμε. Η σωστή προετοιμασία των δεδομένων μας είναι μια διαδικασία η οποία είναι απαραίτητη για την εξαγωγή έγκυρων αποτελεσμάτων και ευρημάτων από τα τα δεδομένα μας.

Στην μελέτη που θα ακολουθήσει χρησιμοποιήθηκαν δεδομένα από ηλεκτρονική πλατφόρμα παραγγελιών προϊόντων σούπερ μάρκετ. Τα δεδομένα μας περιέχουν στοιχεία συναλλαγών των παραγγελιών ,τους πελάτες που συνδέονται με αυτές τις συναλλαγές και χρονικά δεδομένα εγγραφής του πελάτη στην πλατφόρμα και των ημερομηνιών που έχουν πραγματοποιηθεί οι παραγγελίες. Η διαδικασία που θα ακολουθήσει είναι ο καθαρισμός των δεδομένων, παρουσίαση και ανάλυση των δεδομένων μας, τμηματοποίηση της πελατειακής βάσης σε cluster και market basket analysis σε δυο κατηγοριες πελατών.

Για την ανάλυση των δεδομένων χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python.Η επεξεργασία και ανάλυση πραγματοποιήθηκε με τη χρήση βιβλιοθηκών όπως Pandas και NumPy για τη διαχείριση δεδομένων, η Scikit-learn για την εφαρμογή αλγορίθμων μηχανικής μάθησης.

3.2.1 Data cleaning and preparation

Το στάδιο της προεπεξεργασίας δεδομένων αποτελεί κρίσιμο σημείο για την αξιοπιστία της ανάλυσης. Στο πλαίσιο της εργασίας των σύνολο δεδομένων που έχουμε προέρχεται από ηλεκτρονική πλατφόρμα διανομής προιοντων σουπερ μαρκετ. Το σύνολο των δεδομένων που χρησιμοποιήθηκε αποτελειται 772.266 εγγραφές, 43.617 διακριτούς πελάτες και 4.025 διαφορετικά προϊόντα. Ο μεγάλος όγκος των δεδομένων μας και η ποικιλία αυτών καθηστούν απαιραιτιτο των καθαρισμό και φιλατρρισμός των δεδομένων ώστε να διασφαλιστεί η ποιότητα της αναλυσης και των αποτελεσμάτων μας

Πίνακας 1 -Παρουσιασή βασικών μεταβλητών των δεδομένων

Μεταβλητή	**Περιγραφή**	**Τύπος Δεδομένων**
order_id	Μοναδικό αναγνωριστικό παραγγελίας	Κατηγορικό (String)
customer_id	Μοναδικό αναγνωριστικό πελάτη	Κατηγορικό (String)
order_value	Συνολική αξία παραγγελίας	Αριθμητικό (Float)
order_timestamp	Ημερομηνία και ώρα παραγγελίας	Ημερομηνία/Ώρα (Datetime - UTC)
product_name	Όνομα προϊόντος	Κατηγορικό (String)
product_category	Κατηγορία προϊόντος	Κατηγορικό (String)
product_price	Τιμή προϊόντος	Αριθμητικό (Float)
items_sold	Ποσότητα προϊόντος	Αριθμητικό (Float)
geographical_region_of_user	Περιοχή πελάτη	Κατηγορικό (String)
geographical_region_of_shop	Περιοχή καταστήματος	Κατηγορικό (String)

registered_at	Ημερομηνία εγγραφής πελάτη	Ημερομηνία/Ώρα (Datetime - UTC)

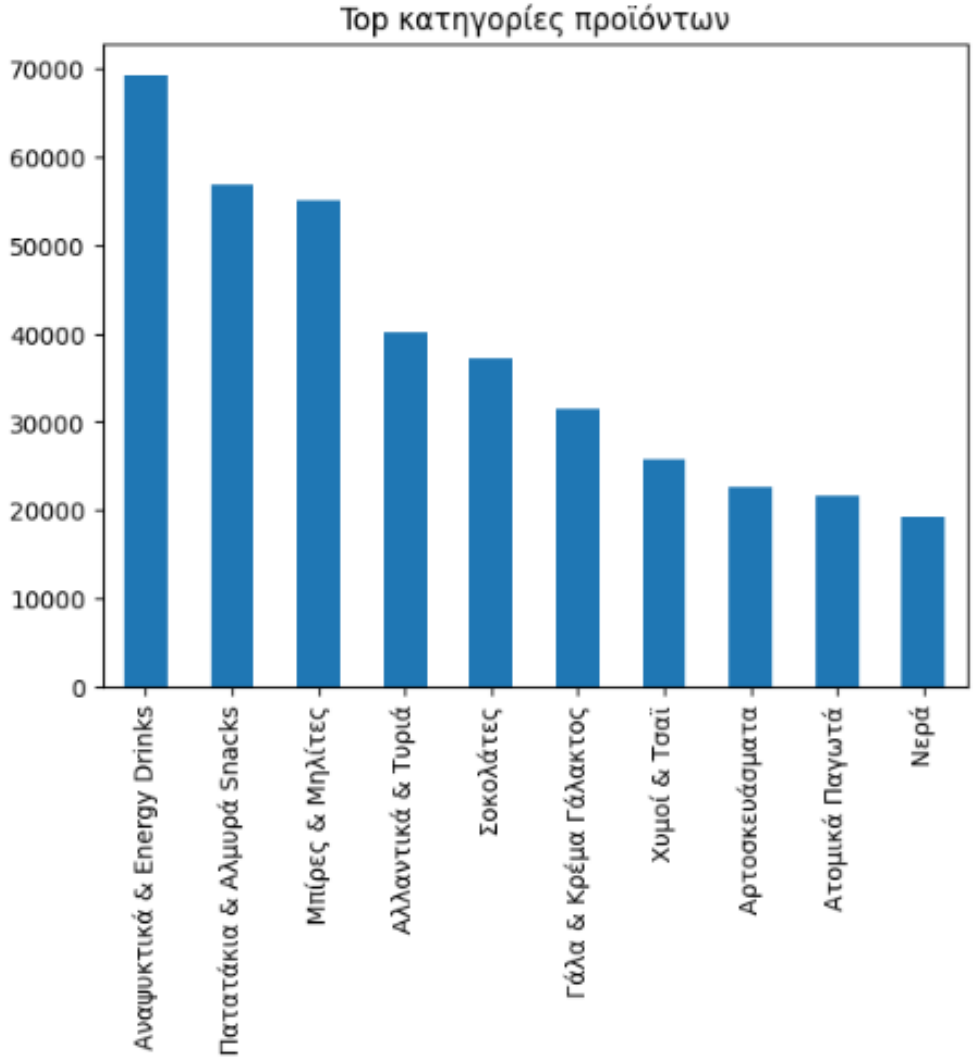
Στον Πίνακα παρουσιάζονται οι μεταβλητές του συνόλου δεδομένων μας, της περιγραφής τους καθως και των τύπων δεδομένων τους.

Τέλος, για την βελτίωση της ποιότητας των δεδομένων και την αποφυγή στρεβλώσεων στην ανάλυση, πραγματοποιήθηκε αφαίρεση ακραίων τιμών από το αρχικό σύνολο δεδομένων. Η διαδικασία βασίστηκε στη μέθοδο του ενδοτεταρτημοριακού εύρους (Interquartile Range - IQR), όπου υπολογίστηκαν το πρώτο (Q1) και το τρίτο τεταρτημόριο (Q3) της μεταβλητής αξίας παραγγελίας.Οι τιμές που βρίσκονται εκτός του διαστήματος, θεωρήθηκαν ακραίες και αφαιρέθηκαν από το dataset. Με τον τρόπο αυτό εξασφαλίσαμε οτι η αναλυση βασίζεται σε πιο αντιπροσωπευτικά δεδομένα.

3.2.2 Data exploration and visualization

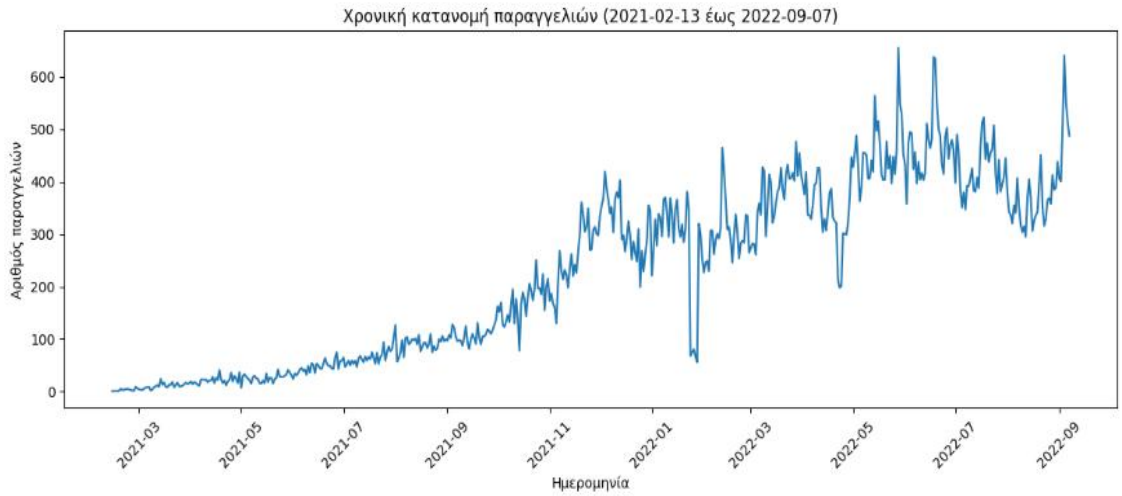
Στην συνέχεια θα παρουσιαστούν κάποια διαγράμματα τα οποία δημιουργήθηκαν μέσω της Python, με στόχο την εις βάθος κατανόηση των χαρακτηριστικών του συνόλου δεδομένων μας.

Πίνακας 2-Top Κατηγορίες Προϊόντων



Στο παραπάνω διάγραμμα απεικονίζονται οι μεγαλύτερες σε ζήτηση κατηγορίες προϊόντων σε συνάρτηση με τον αριθμό παραγγελιών που εμφανίζονται. Τα αναψυκτικά και τα ενεργειακά ποτά αποτελούν τη δημοφιλέστερη κατηγορία προϊόντων καθώς εμφανίζονται σε περίπου 70.000 παραγγελίες. Ακολουθούν με μικρή διαφορά τα πατατάκια και τα αλμυρά σνακς καθώς και οι μπύρες και οι μηλίτες που εμφανίζονται σε περίπου 60.000 παραγγελίες γεγονός που δηλώνει τη μεγάλη ζήτηση αυτών των προϊόντων. Ακολουθούν με ένα μεσαίο αριθμό κατανάλωσης τα αλλαντικά και τα τυριά και οι σοκολάτες ενώ σε μικρότερο αριθμό παραγγελιών εντοπίστηκαν τα αρτοσκευάσματα και τα ατομικά παγωτά, πράγμα που υποδηλώνει τη χαμηλή ζήτηση και κατανάλωση των προϊόντων αυτών. Από το διάγραμμα συμπεραίνουμε πως τα προϊόντα άμεσης κατανάλωσης εντοπίστηκαν στο μεγαλύτερο αριθμό παραγγελιών ενώ τα βασικά προϊόντα όπως τα αρτοσκευάσματα είχαν χαμηλή ζήτηση καθώς εντοπίστηκαν σε αρκετά μικρό αριθμό παραγγελιών.

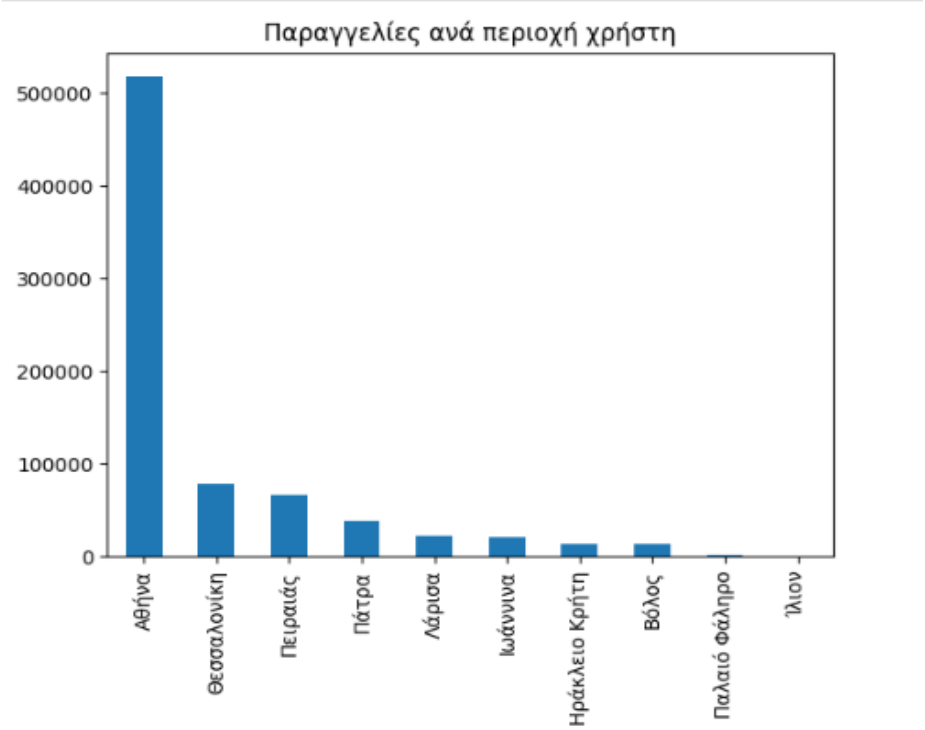
Πίνακας 3-Χρονική Κατανομή Παραγγελιών



Στο παραπάνω διάγραμμα απεικονίζεται η χρονική κατανομή των παραγγελιών στην πλατφόρμα από τον Φεβρουάριο του 2021 έως τον Σεπτέμβριο του 2022. Συγκεκριμένα έχουν ληφθεί δεδομένα για τη χρονική περίοδο 13 Φεβρουαρίου 2021 έως τις 7 Σεπτεμβριου 2022. Παρατηρούμε μια σταθερή άνοδο του αριθμού των παραγγελιών με την πάροδο του χρόνου με μικρές πτώσεις τον Φεβρουάριο του 2022 και τον Μάρτιο του ίδιου έτος.

Η συνεχώς αυξανόμενη τάση του αριθμού των παραγγελιών υποδηλώνει την συνεχώς αυξανόμενη χρήση της πλατφόρμας με την πάροδο του χρόνου. Μάλιστα οι μεγαλύτερες τιμες σημειώνονται τα τελευταία έτη αποτυπώνοντας τη δυναμική της εξέλιξη, την αυξανόμενη ζήτηση από τους καταναλωτές καθώς και την εντονότερα δραστηριότητα εντός της πλατφόρμας κατά την εξεταζόμενη περίοδο.

Πίνακας 4 - Παραγγελίες ανα περιοχή



Το παραπάνω διάγραμμα παρουσιάζει τον αριθμό των παραγγελιών ανά γεωγραφική περιοχή.Η Αθήνα συγκεντρώνει το μεγαλύτερο πλήθος παραγγελιών, γεγονός που υποδηλώνει ότι αποτελεί την κύρια αγορά της πλατφόρμας και το βασικό κέντρο δραστηριότητας. Ακολουθούν η Θεσσαλονίκη και ο Πειραιάς, οι οποίες εμφανίζουν σημαντικά μικρότερο αλλά αξιοσημείωτο όγκο παραγγελιών.

3.2.3 Customer behavior based on Recency, Frequency, and Monetary metrics

Για την ανάλυση της συμπεριφοράς των πελατών χρησιμοποιήθηκε η μεθοδολογία RFM (Recency, Frequency, Monetary), ένα από τα αποτελεσματικότερα εργαλεία για την αξιολόγηση της δαπάνης του αγοραστικού κοινού. Με αυτή την προσέγγιση γίνεται ευκολότερα κατανοητή η αλληλεπίδραση των πελατών με την πλατφόρμα, καθώς και ο εντοπισμός διαφορετικών επιπέδων αξίας και δραστηριότητας.

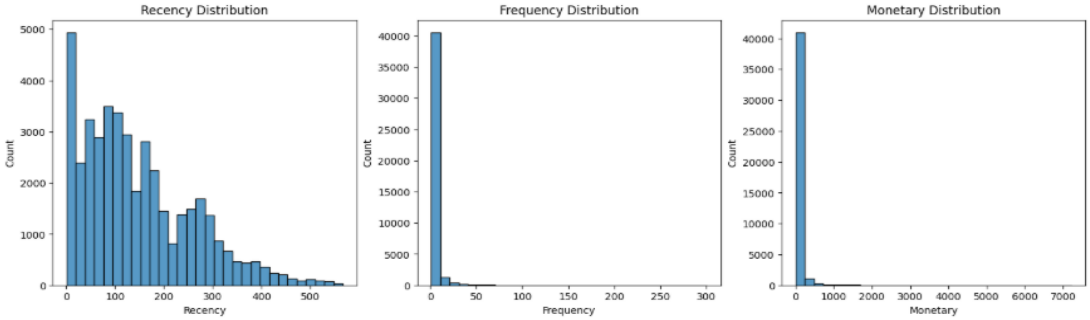
Η μεταβλητή Recency (R) εκφράζει το χρονικό διάστημα που έχει μεσολαβήσει από την τελευταία παραγγελία του εκάστοτε πελάτη μέχρι την ημερομηνία αναφοράς της ανάλυσης. Οι χαμηλές τιμές Recency υποδηλώνουν πρόσφατη δραστηριότητα και έντονη αλληλεπίδραση με την πλατφόρμα, ενώ υψηλές τιμές πιθανώς να υποδεικνύουν πελάτες με μειωμένη ή καθόλου δραστηριότητα.

Η μεταβλητή Frequency (F) αντιπροσωπεύει τον αριθμό των παραγγελιών που έχει πραγματοποιήσει κάθε πελάτης. Δεδομένου ότι το αρχικό σύνολο δεδομένων περιλαμβάνει εγγραφές σε επίπεδο προϊόντος, πραγματοποιήθηκε ομαδοποίηση των δεδομένων σε επίπεδο παραγγελίας (order_id), ώστε να υπολογιστεί με ακρίβεια ο αριθμός των διακριτών συναλλαγών ανά πελάτη.

Η μεταβλητή Monetary (M) εκφράζει τη συνολική χρηματική αξία των συναλλαγών κάθε πελάτη και υπολογίστηκε ως το άθροισμα της αξίας των παραγγελιών (order value). Η μεταβλητή αυτή αποτελεί βασικό δείκτη της οικονομικής συνεισφοράς κάθε πελάτη στην πλατφόρμα

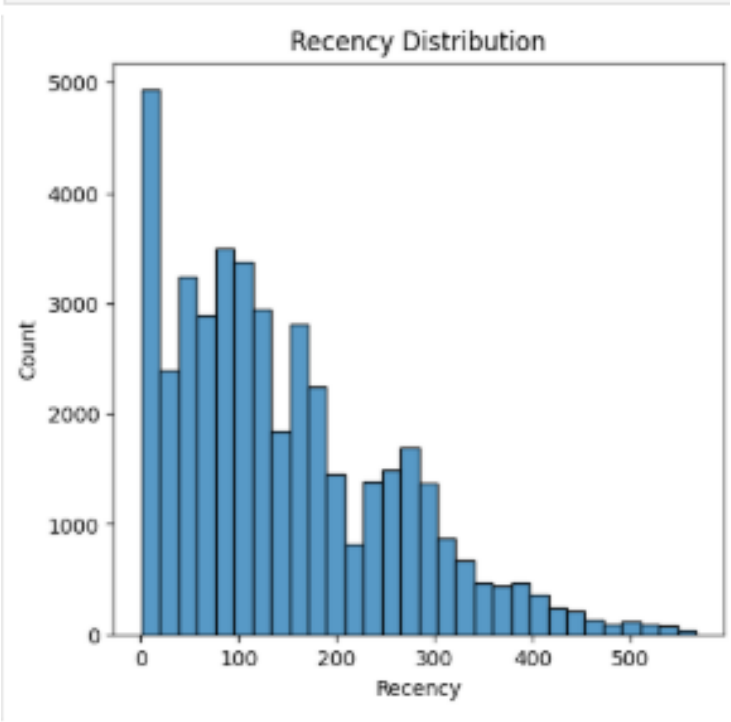
.

Πίνακας 5- Κατανομές των μεταβλητών Recency, Frequency και Monetary



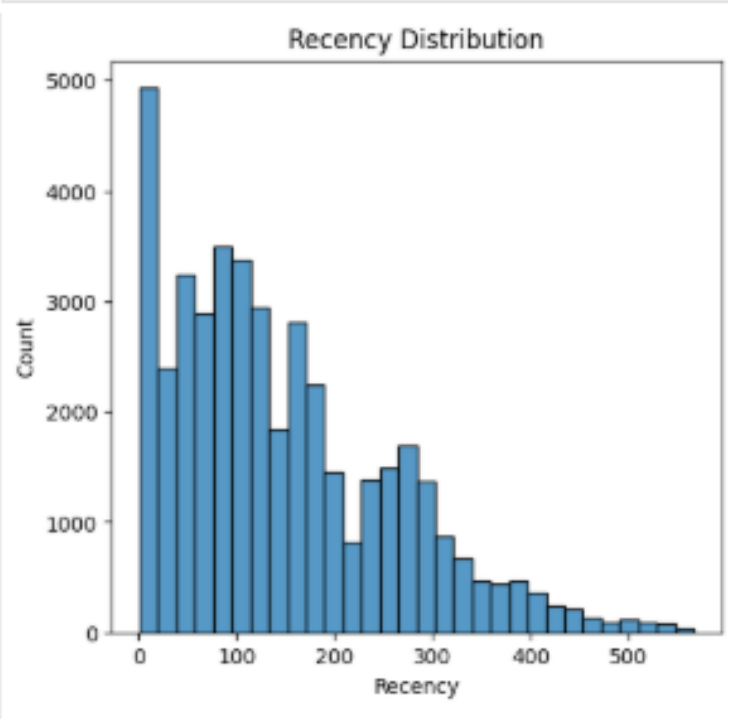
Η ανάλυση των κατανομών των μεταβλητών Recency, Frequency και Monetary έδειξε σημαντικά χαρακτηριστικά της πελατειακής συμπεριφοράς. Όπως παρουσιάζεται στον Πίνακα 5, οι κατανομές των μεταβλητών έχουν έντονη δεξιά ασυμμετρία (right-skewed), κάτι που υποδηλώνει ότι το μεγαλύτερο μέρος των πελατών εμφανίζει χαμηλά επίπεδα δραστηριότητας και αγοραστικής δραστηριότητας, ενώ η μειονότητα παρουσιάζει σημαντικά υψηλότερες τιμές.

Πίνακας 6-Recency



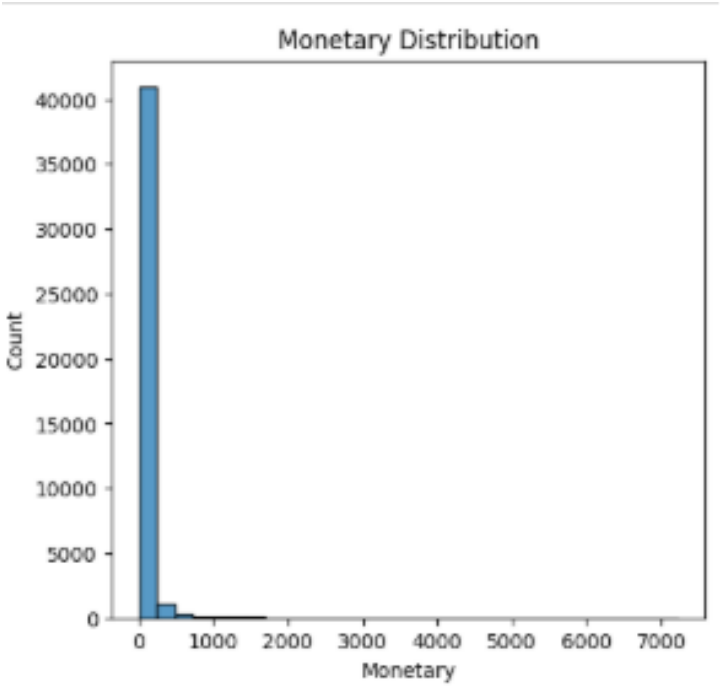
Συγκεκριμένα, για τη μεταβλητή Recency παρατηρείται μεγάλη συγκέντρωση πελατών σε χαμηλές τιμές, γεγονός που υποδηλώνει ότι ένα σημαντικό ποσοστό χρηστών έχει πραγματοποιήσει πρόσφατες αγορές. Παράλληλα, η ύπαρξη ουράς προς υψηλές τιμές υποδεικνύει την παρουσία πελατών που έχουν μεγάλο χρονικό διάστημα να πραγματοποιήσουν συναλλαγή, γεγονός που μπορεί να συνδέεται με πιθανή αποχώρηση (customer churn).

Πίνακας 7- Frequency



Για τη μεταβλητή Frequency, το παραπάνω ιστόγραμμα δείχνει ότι η πλειονότητα των πελατών πραγματοποιεί περιορισμένο αριθμό παραγγελιών, ενώ ένα μικρό ποσοστό εμφανίζει ιδιαίτερα υψηλή συχνότητα αγορών και είναι εκείνοι που χαρακτηρίζονται από υψηλό επίπεδο αφοσίωσης προς την πλατφόρμα.

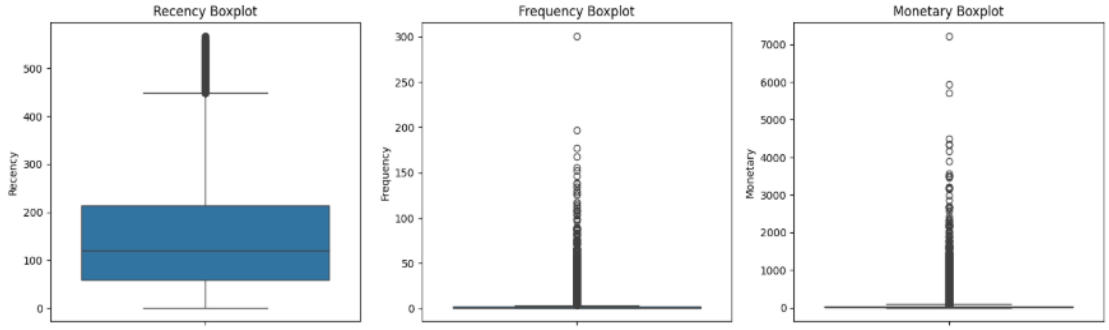
Πίνακας 8 - Monetary



Αντίστοιχα, η μεταβλητή Monetary παρουσιάζει έντονη συγκέντρωση σε χαμηλές τιμές, με λίγους πελάτες να εμφανίζουν ιδιαίτερα υψηλή συνολική δαπάνη. Το εύρημα αυτό

υποδηλώνει ότι η συνεισφορά στα συνολικά έσοδα δεν κατανέμεται ομοιόμορφα, αλλά συγκεντρώνεται σε ένα μικρό ποσοστό πελατών υψηλής αξίας.

Πίνακας 9 - Boxplots των μεταβλητών Recency, Frequency και Monetary



Η περαιτέρω ανάλυση μέσω boxplots επιβεβαιώνει τα παραπάνω ευρήματα. Όπως παρουσιάζεται στην Πίνακας 9 - Boxplots των μεταβλητών Recency, Frequency και Monetary, και οι τρεις μεταβλητές εμφανίζουν σημαντική διασπορά και παρουσία ακραίων τιμών (outliers).

Πιο συγκεκριμένα, στον πίνακα για τη μεταβλητή Recency εντοπίζονται πελάτες με ιδιαίτερα υψηλές τιμές, οι οποίοι έχουν μεγάλο χρονικό διάστημα να πραγματοποιήσουν δαπάνη και ενδεχομένως να βρίσκονται σε κίνδυνο αποχώρησης. Στον πίνακα για τη μεταβλητή Frequency παρατηρούνται πελάτες με εξαιρετικά υψηλό αριθμό παραγγελιών, που μπορούν να χαρακτηριστούν ως πιστοί πελάτες με υψηλή αγοραστική δραστηριότητα. Αντίστοιχα, στον πίνακα για την μεταβλητή Monetary εμφανίζονται πελάτες με εξαιρετικά υψηλές δαπάνες, οι οποίοι συνεισφέρουν δυσανάλογα μεγάλο μέρος των συνολικών εσόδων.

Στο σύνολό της, η ανάλυση RFM αναδεικνύει έντονη ανισοκατανομή που υπάρχει στην πελατειακή βάση της πλατφόρμας, όπου το μεγαλύτερο μέρος των πελατών χαρακτηρίζεται από χαμηλή δραστηριότητα και δαπάνη, ενώ ένα μικρό ποσοστό πελατών υψηλής αξίας φαίνεται να παίζει καθοριστικό ρόλο στη εισροή εσόδων. Τα αποτελέσματα της ανάλυσης συμβάλλουν στην περαιτέρω κατηγοριοποίηση των πελατών και την ανάπτυξη στοχευμένων στρατηγικών μάρκετινγκ

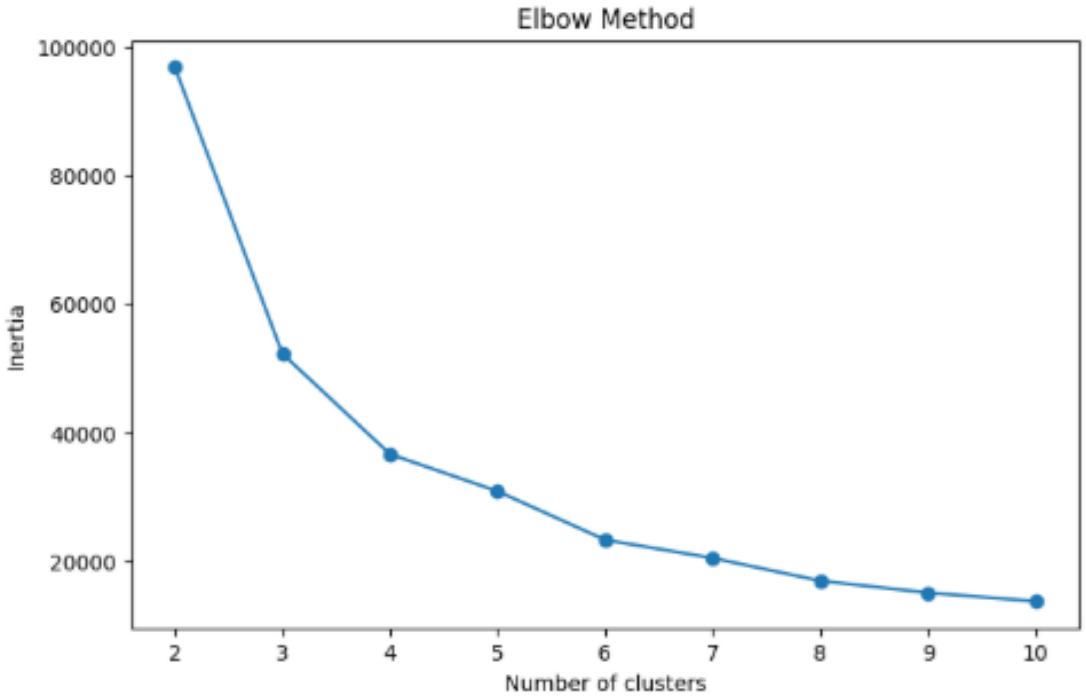
3.2.4 Customer categorization analysis

Η κατηγοριοποίηση των πελατών πραγματοποιήθηκε με τη χρήση της μεθόδου K-Means clustering, με σκοπό τον διαχωρισμό των πελατών σε ομοιογενείς ομάδες με παρόμοια αγοραστική δραστηριότητα. Η ανάλυση βασίστηκε στις μεταβλητές Recency, Frequency και Monetary, πο προέκυψαν από τη διαδικασία RFM

.Αρχικά, εφαρμόστηκε κανονικοποίηση (standardization) των δεδομένων, ώστε να εξαλειφθούν διαφορές κλίμακας μεταξύ των μεταβλητών και να διασφαλιστεί η ορθή λειτουργία του αλγορίθμου. Στη συνέχεια, χρησιμοποιήθηκε η μέθοδος Elbow για τον προσδιορισμό του κατάλληλου αριθμού clusters.

Όπως παρουσιάζεται στον Πίνακας 10 - Elbow Method για τον προσδιορισμό του αριθμού clusters, η τιμή της αδράνειας (inertia) μειώνεται σημαντικά έως το σημείο k=4, ενώ μετά το k=5 η μείωση γίνεται πιο ομαλή. Το σημείο καμπής εντοπίζεται μεταξύ των τιμών k=4 και k=5, και για τον λόγο αυτό επιλέχθηκε η τιμή k=5, η οποία επιτρέπει έναν πιο λεπτομερή και ουσιαστικό διαχωρισμό των πελατών.

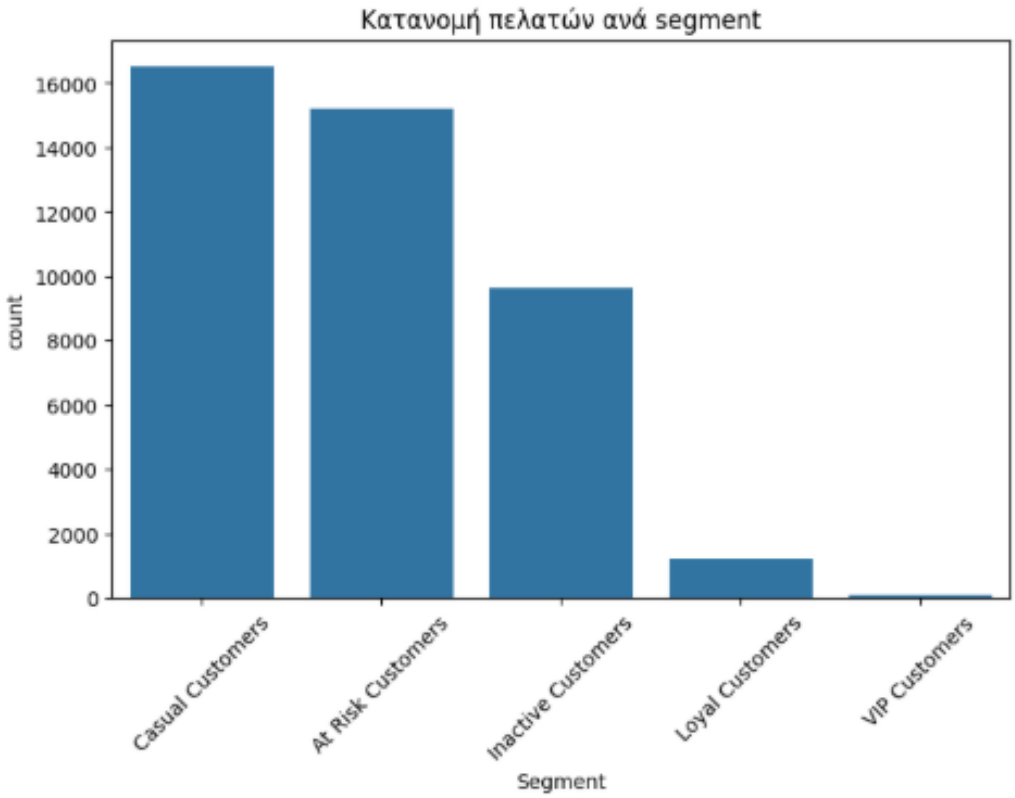
Πίνακας 10 - Elbow Method για τον προσδιορισμό του αριθμού clusters



Έπειτα από την εφαρμογή του αλγορίθμου K-Means, έγινε κατανομή των πελατών σε πέντε διακριτές ομάδες (segments), οι οποίες χαρακτηρίστηκαν με βάση τις μέσες τιμές των μεταβλητών Recency, Frequency και Monetary.

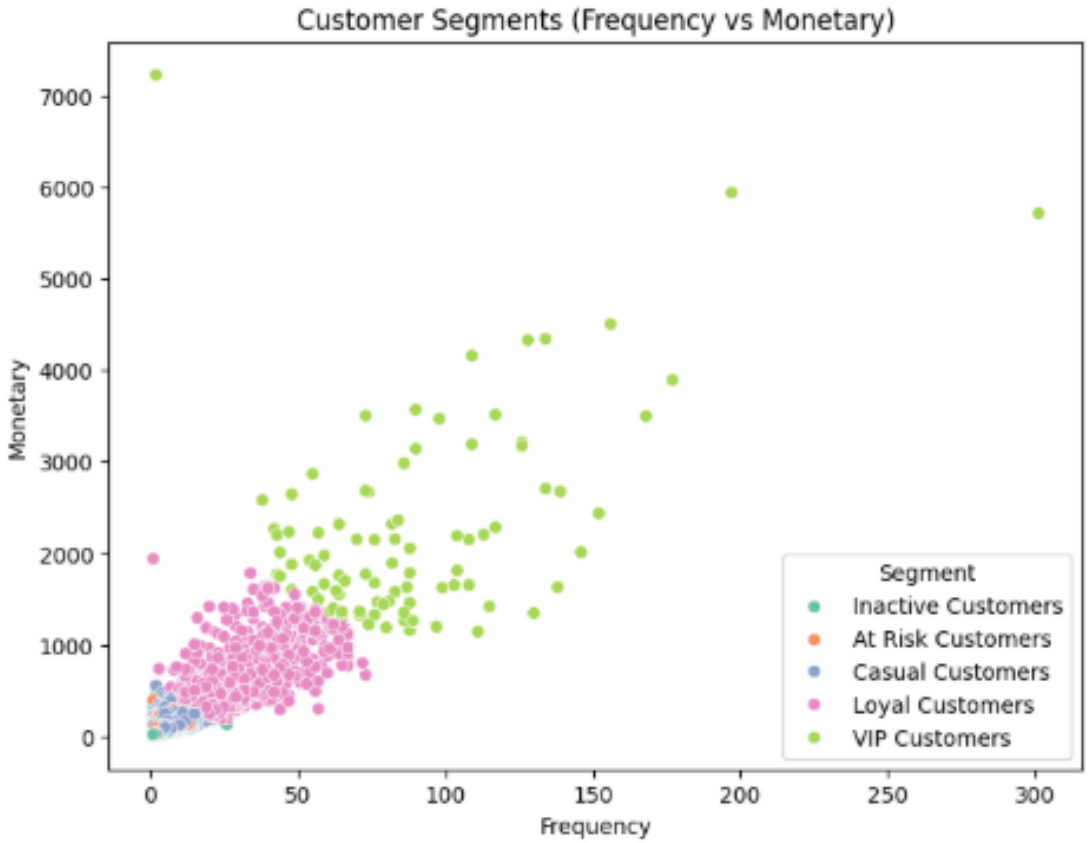
Οι πελάτες κατανεμήθηκαν ανά segment όπως παρουσιάζεται στον Πίνακα 11, όπου παρατηρείται ότι η πλειονότητα των πελατών εντάσσεται στις κατηγορίες χαμηλής και μέτριας αξίας, ενώ ένα πολύ μικρό ποσοστό βρίσκεται στην κατηγορία υψηλής αξίας.

Πίνακας 11- Κατανομή πελατών ανά segment



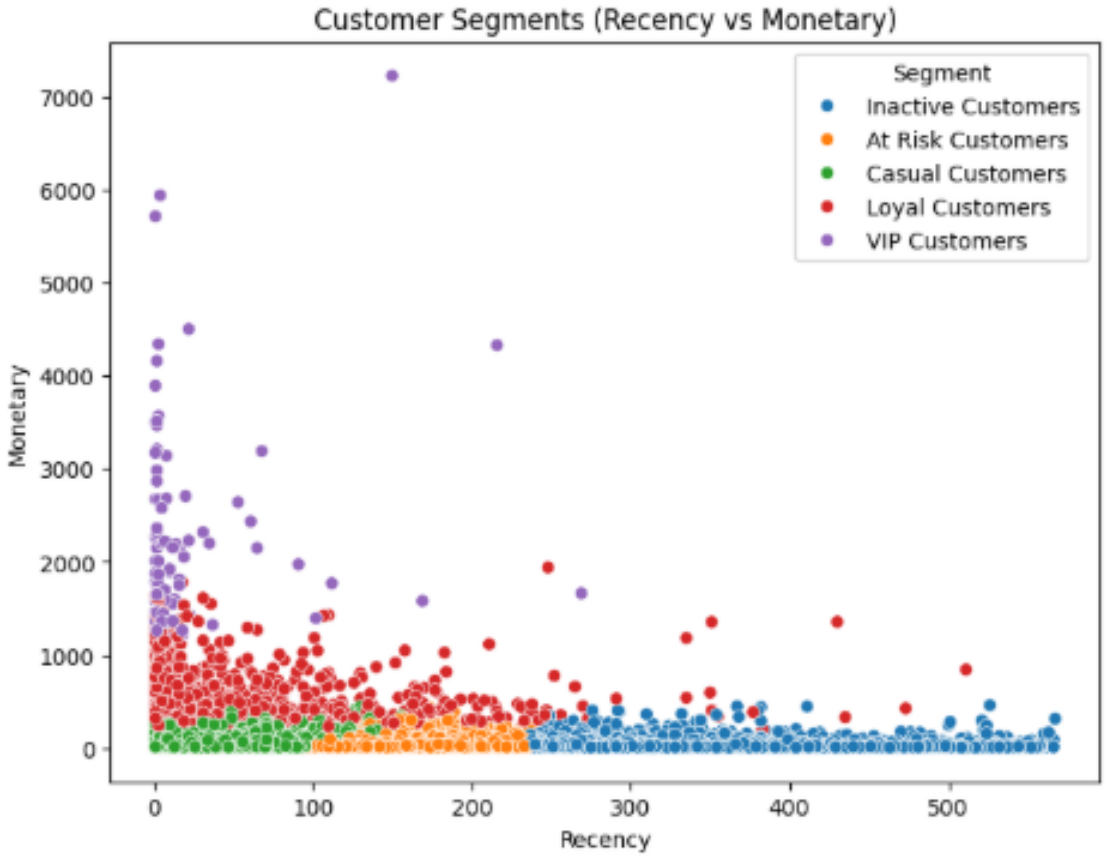
Η οπτικοποίηση των clusters σύμφωνα με τις μεταβλητές Frequency και Monetary παρουσιάζεται στον Πίνακα 12. Όπως διακρίνεται οι πελάτες υψηλής αξίας (VIP) διαχωρίζονται εμφανώς από τους υπόλοιπους, καθώς παρουσιάζουν σημαντικά υψηλότερες τιμές τόσο στη συχνότητα όσο και στη συνολική δαπάνη.

Πίνακας 12 - Οπτικοποίηση των segments (Frequency vs Monetary)



Επιπλέον, η σχέση μεταξύ Recency και Monetary απεικονίζεται στον Πίνακας 13, όπου διακρίνεται ότι οι πελάτες υψηλής αξίας χαρακτηρίζονται από χαμηλές τιμές Recency (πρόσφατη δραστηριότητα) και υψηλή δαπάνη, ενώ οι ανενεργοί πελάτες εμφανίζουν υψηλές τιμές Recency και χαμηλή συνολική αξία συναλλαγών.

Πίνακας 13 - Οπτικοποίηση των segments (Recency vs Monetary)



Με βάση τα χαρακτηριστικά των clusters, προέκυψαν οι ακόλουθες κατηγορίες πελατών:

- VIP Customers: Οι πελάτες αυτοί παρουσιάζουν πολύ υψηλή συχνότητα αγορών και υψηλή συνολική δαπάνη, ενώ έχουν πραγματοποιήσει πρόσφατες συναλλαγές. Αποτελούν την πιο πολύτιμη κατηγορία πελατών και συμβάλλουν σημαντικά στα συνολικά έσοδα της πλατφόρμας.
- Loyal Customers: Πρόκειται για πελάτες με σταθερή αγοραστική συμπεριφορά, σχετικά υψηλή συχνότητα αγορών και μέτρια έως υψηλή δαπάνη. Διατηρούν ενεργή σχέση με την πλατφόρμα και αποτελούν σημαντική ομάδα για στρατηγικές διατήρησης.
- Casual Customers: Οι πελάτες αυτοί πραγματοποιούν περιορισμένο αριθμό αγορών και εμφανίζουν χαμηλή συνολική δαπάνη. Αντιπροσωπεύουν μεγάλο ποσοστό της πελατειακής βάσης, αλλά με χαμηλή συνεισφορά στα έσοδα.
- At Risk Customers: Οι πελάτες αυτοί εμφανίζουν μειωμένη πρόσφατη δραστηριότητα (υψηλότερη Recency) και χαμηλή συχνότητα αγορών, γεγονός που υποδηλώνει πιθανή απομάκρυνση από την πλατφόρμα. Απαιτούν στοχευμένες ενέργειες επαναδραστηριοποίησης.
- Inactive Customers: Πρόκειται για πελάτες με πολύ υψηλές τιμές Recency και πολύ χαμηλή δραστηριότητα, οι οποίοι έχουν ουσιαστικά σταματήσει να χρησιμοποιούν την πλατφόρμα. Η επαναπροσέγγισή τους αποτελεί πρόκληση για την επιχείρηση.

Η ανάλυση των segments τονίζειι την έντονη ανισοκατανομή της πελατειακής βάσης, όπου ένα μικρό ποσοστό πελατών υψηλής αξίας (VIP) συνεισφέρει δυσανάλογα μεγάλο μέρος των εσόδων. Αντίθετα, το μεγαλύτερο μέρος των πελατών ανήκει σε κατηγορίες χαμηλής δραστηριότητας και δαπάνης.

Τα ευρήματα αυτά υπερτονίζουν την ανάγκη για ανάπτυξη διαφοροποιημένων στρατηγικών μάρκετινγκ, όπως είναι η επιβράβευση των VIP πελατών, η ενίσχυση της αφοσίωσης των loyal customers και η επαναδραστηριοποίηση των at risk και inactive πελατών

3.3 MAKET BASKET ANALYSIS

Η ανάλυση καλαθιού αγορών (Market Basket Analysis) είναι μία σημαντική τεχνική στον τομέα της εξόρυξης δεδομένων, που βοηθά στον εντοπισμό σχέσεων μεταξύ προϊόντων που αγοράζονται μαζί. Μέσω αυτής της μεθόδου κατανοούμε καλύτερα τη συμπεριφορά των πελατών και μπορούμε να πάρουμε πιο σωστές αποφάσεις βασισμένες σε δεδομένα.

Η βασική αρχή είναι η ανάλυση των ιστορικών συναλλαγών, όπου κάθε συναλλαγή αναπαρίσταται ως ένα σύνολο καλαθιού. Μέσω της επεξεργασίας των ιστορικών δεδομένων εξάγονται κανόνες συσχέτισης (association rules), οι οποίοι εκφράζουν σχέσεις της μορφής: $X \rightarrow Y$, όπου η παρουσία ενός συνόλου προϊόντων X συνεπάγεται αυξημένη πιθανότητα εμφάνισης ενός άλλου συνόλου προϊόντων Y. (Agrawal & Srikant, 1994)

3.3.1 Αλγόριθμος Apriori

Ο αλγόριθμος Apriori είναι μια από τις πιο κλασικές προσεγγίσεις εξόρυξης συχνών συνόλων στοιχείων (frequent itemsets). Βασίζεται στην ιδιότητα της καθοδικής μονοτονίας (downward closure property), σύμφωνα με την οποία εάν ένα σύνολο προϊόντων είναι συχνό, τότε όλα τα υποσύνολά του είναι επίσης συχνά (Agrawal & Srikant, 1994).

Η λειτουργία του αλγορίθμου είναι επαναληπτική και περιλαμβάνει:

- Τη δημιουργία υποψήφιων συνόλων (candidate itemsets)
- Τον υπολογισμό του δείκτη support
- Την απόρριψη των μη συχνών συνόλων

Παρά τη σαφήνεια και την ερμηνευσιμότητά του, ο αλγόριθμος Apriori χρειάζεται αρκετή υπολογιστική δύναμη, καθώς απαιτεί πολλαπλές σαρώσεις του συνόλου δεδομένων και δημιουργεί μεγάλο αριθμό υποψηφίων συνδυασμών και έτσι περιορίζεται η αποδοτικότητα του σε μεγάλα σύνολα δεδομένων.

3.3.2 Αλγόριθμος FP-Growth

Ο αλγόριθμος FP-Growth αποτελεί έναν αποτελεσματικό τρόπο για τον εντοπισμό των πιο συχνών συνόλων στοιχείων σε ένα σύνολο δεδομένων. Αυτός ο αλγόριθμος είναι μια εξέλιξη του αλγορίθμου Apriori, αντιμετωπίζοντας τις αδυναμίες του Apriori και προσφέροντας μια αποτελεσματικότερη λύση για τον εντοπισμό συχνών συνόλων. Ο αλγόριθμος αυτός ξεκινά συμπιέζοντας τη βάση δεδομένων εισόδου, αναπτύσσοντας έτσι μια παρουσία ενός συχνού δέντρου μοτίβου. Στη συνέχεια, η συμπιεσμένη βάση δεδομένων διαιρείται σε μερικές υπό όρους βάσεις δεδομένων, όπου κάθε βάση δεδομένων αντιπροσωπεύει ένα μοναδικό συχνό μοτίβο. Τέλος, πραγματοποιείται εξόρυξη κάθε βάσης δεδομένων ξεχωριστά. Ως εκ τούτου, το κόστος αναζήτησης είναι σημαντικά μειωμένο, προσφέροντας καλή εκλεκτικότητα. Τα βήματα του αλγορίθμου FP-Growth είναι τα εξής:

1. Δημιουργία FP-Tree: Διαβάζουμε τα δεδομένα και χτίζουμε το FP-Tree. Κάθε αντικείμενο προστίθεται στο δέντρο με βάση τη συχνότητά του στα δεδομένα. Τα αντικείμενα ταξινομούνται βάσει της συχνότητας εμφάνισής τους.
2. Κατασκευή Conditional FP-Tree: Δημιουργούμε ένα Conditional FP-Tree για κάθε συνδυασμό στοιχείων. Αφαιρούμε τα λιγότερο συχνά στοιχεία από τα δεδομένα και επαναλαμβάνουμε τη διαδικασία για κάθε νέο Conditional FP-Tree.

3. Αναδρομική εφαρμογή του αλγορίθμου στα Conditional FP-Trees: Επαναλαμβάνουμε τη διαδικασία κατασκευής FP-Tree και Conditional FP-Tree για κάθε νέο σύνολο δεδομένων. Επαναλαμβάνουμε αυτό το βήμα μέχρι να μην υπάρχουν πλέον στοιχεία για εξέταση.
4. Εξαγωγή συχνών συνόλων στοιχείων: Εξάγουμε τα συχνά σύνολα στοιχείων από το FP Tree, λαμβάνοντας υπόψη τα Conditional FP-Trees που έχουν δημιουργηθεί. Κάθε σύνολο στοιχείων που εξάγεται αποτελεί ένα συχνό σύνολο.

3.3.3 Αλγόριθμος Eclat

Οι αλγόριθμοι Apriori και FP-Growth αποτελούν παραδείγματα μεθόδων κανόνων συσχέτισης που χρησιμοποιούν οριζόντιες μορφές δεδομένων. Μια μέθοδος κανόνων συσχέτισης που αξιοποιεί μια κάθετη μορφή δεδομένων είναι η ECLAT, ή αλλιώς ο μετασχηματισμός κλάσης ισοδυναμίας. Ο κύριος σκοπός του αλγορίθμου μετασχηματισμού κλάσης ισοδυναμίας είναι να εντοπίζει συχνά στοιχεία σε ένα σύνολο συναλλαγών και λειτουργεί αποκλειστικά σε μια βάση δεδομένων κάθετης διάταξης. Ο αλγόριθμος ECLAT σαρώνει τη βάση δεδομένων μόνο μία φορά για να βρει το συχνό σύνολο στοιχείων, σε αντίθεση με τον αλγόριθμο Apriori, ο οποίος χρειάζεται περισσότερο χρόνο για να εντοπίσει συχνά σύνολα στοιχείων, καθώς απαιτείται η επανειλημμένη σάρωση της βάσης δεδομένων, μια διαδικασία που απαιτεί περισσότερο χρόνο. Ο αλγόριθμος ECLAT εντοπίζει στοιχεία από κάτω προς τα πάνω, ακολουθώντας μια πρώτη αναζήτηση κατά βάθος. Αυτός ο αλγόριθμος υπολογίζει μόνο την τιμή υποστήριξης, ενώ η τιμή εμπιστοσύνης δεν υπολογίζεται σε αυτόν τον αλγόριθμο. Τα βήματα του αλγορίθμου ECLAT είναι τα εξής:

1. Δημιουργία Ισοδύναμων Κλάσεων (Equivalence Classes): Αρχικά, τα δεδομένα συναλλαγών ομαδοποιούνται σε ισοδύναμες κλάσεις με βάση τα στοιχεία που περιέχουν. Κάθε στοιχείο αποτελεί την ισοδύναμη κλάση του εαυτού του.
2. Εύρεση Συχνών Συνόλων Στοιχείων εντός των Κλάσεων: Στη συνέχεια, αναζητούμε συχνά σύνολα στοιχείων μέσα σε κάθε ισοδύναμη κλάση, δηλαδή τις συνδυασμένες συχνότητες των στοιχείων σε κάθε κλάση.
3. Συνένωση Συχνών Συνόλων Στοιχείων: Τέλος, συνδυάζουμε τα συχνά σύνολα στοιχείων που έχουν εντοπιστεί σε κάθε ισοδύναμη κλάση για να δημιουργήσουμε τα συχνά σύνολα στοιχείων για ολόκληρο το σύνολο των δεδομένων.

3.3.4 Μεθοδολογία

Για την εξαγωγή των συχνών συνόλων αντικειμένων χρησιμοποιήθηκε ο αλγόριθμος FP-Growth (Frequent Pattern Growth), ο οποίος αποτελεί μία αποδοτική προσέγγιση σε σύγκριση με παραδοσιακές μεθόδους όπως ο Apriori.

Ο FP-Growth βασίζεται στη δημιουργία μιας δομής δεδομένων, γνωστής ως FP-Tree, η οποία επιτρέπει τη συμπίεση των συναλλαγών και την αποδοτική αποθήκευση των επαναλαμβανόμενων προτύπων. Σε αντίθεση με άλλες μεθόδους, δεν απαιτείται η δημιουργία υποψήφιων συνόλων, γεγονός που μειώνει σημαντικά τον υπολογιστικό φόρτο.

Η διαδικασία περιλαμβάνει δύο βασικά στάδια:

- Κατασκευή του FP-Tree από τα δεδομένα των συναλλαγών
- Αναδρομική εξόρυξη των συχνών συνόλων μέσω της ανάλυσης του δέντρου

Μετά την εξαγωγή των συχνών συνόλων, δημιουργήθηκαν κανόνες συσχέτισης, οι οποίοι αξιολογήθηκαν με βάση τρεις βασικούς δείκτες:

- Support: ποσοστό εμφάνισης ενός συνδυασμού προϊόντων στο σύνολο των συναλλαγών
- Confidence: πιθανότητα αγοράς ενός προϊόντος δεδομένης της αγοράς ενός άλλου
- Lift: μέτρο ισχύος της συσχέτισης μεταξύ προϊόντων

3.3.5 Αποτελέσματα Ανάλυσης

Στην παρούσα ενότητα παρουσιάζονται τα αποτελέσματα της ανάλυσης καλαθιού αγορών για δύο κατηγορίες πελατών: τους Casual και τους Loyal. Οι κανόνες συσχέτισης που προέκυψαν αξιολογήθηκαν βάσει των δεικτών support, confidence και lift, και επιλέχθηκαν οι σημαντικότεροι εξ αυτών.

Πίνακας 14 - Αποτελέσματα MBA Πελατών Casual

RULE	Support	Confidence	Lift
Τροφή για γάτες μους Gourmet Gold με Ψάρια Ωκεανού (85gr) → Τροφή για γάτες μους Gourmet Gold με Βοδινό (85gr)	0,002	0,487	144,478
Τροφή για γάτες μους Gourmet Gold με Βοδινό (85gr) → Τροφή για γάτες μους Gourmet Gold με Ψάρια Ωκεανού (85gr)	0,002	0,469	144,478
Τροφή για γάτες μους Gourmet Gold με Βοδινό (85gr) → Τροφή για γάτες μους Gourmet Gold Με Κοτόπουλο (85gr)	0,001	0,363	135,484
Τροφή για γάτες μους Gourmet Gold Με Κοτόπουλο (85gr) → Τροφή για γάτες μους Gourmet Gold με Βοδινό (85gr)	0,001	0,457	135,484
Τροφή για γάτες μους Gourmet Gold Με Κοτόπουλο (85gr) → Τροφή για γάτες μους Gourmet Gold με Ψάρια Ωκεανού (85gr)	0,001	0,402	123,774
Τροφή για γάτες μους Gourmet Gold με Ψάρια Ωκεανού (85gr) → Τροφή για γάτες μους Gourmet Gold Με Κοτόπουλο (85gr)	0,001	0,331	123,774
Μηλίτης Somersby apple φιάλη (330ml) → Μηλίτης Somersby mango lime φιάλη (330ml)	0,004	0,731	101,503
Μηλίτης Somersby mango lime φιάλη (330ml) → Μηλίτης Somersby apple φιάλη (330ml)	0,004	0,573	101,503
Oriental Express Noodles Pot Μοσχαράκι & Μαύρο Πιπέρι (85gr) → Oriental Express Noodles Pot Κοτόπουλο & Πράσινο Κρεμμύδι (85gr)	0,001	0,397	99,22
Oriental Express Noodles Pot Κοτόπουλο & Πράσινο Κρεμμύδι (85gr) → Oriental Express Noodles Pot Μοσχαράκι & Μαύρο Πιπέρι (85gr)	0,001	0,295	99,22
Μηλίτης Somersby watermelon φιάλη (330ml) → Μηλίτης Somersby mango lime φιάλη (330ml)	0,004	0,713	98,924
Μηλίτης Somersby mango lime φιάλη (330ml) → Μηλίτης Somersby watermelon φιάλη (330ml)	0,004	0,588	98,924
Μηλίτης Somersby watermelon φιάλη (330ml), Μηλίτης Somersby apple φιάλη (330ml) → Μηλίτης Somersby mango lime φιάλη (330ml)	0,002	0,709	98,457
Μηλίτης Strongbow Gold Apple (330ml) → Μηλίτης Strongbow Red Berries (330ml)	0,001	0,436	94,121
Μηλίτης Strongbow Red Berries (330ml) → Μηλίτης Strongbow Gold Apple (330ml)	0,001	0,295	94,121

RULE	Support	Confidence	Lift
Τσάι FuzeTea λεμόνι & λουίζα (500ml) → Τσάι FuzeTea Peach & Hibiscus (500ml)	0,001	0,45	86,828
Τσάι FuzeTea Peach & Hibiscus (500ml) → Τσάι FuzeTea λεμόνι & λουίζα (500ml)	0,001	0,22	86,828
Παγωτό B&J Chocolate Fudge Brownie (100ml), Παγωτό B&J Cookie Dough (100ml) → Παγωτό ξυλάκι B&J Peacepop (80ml)	0,001	0,316	77,427
Oriental Express Noodles 3` Ψητές Γαρίδες (87gr) → Oriental Express Noodles 3` Λαχανικά Oriental (87gr)	0,001	0,324	74,187

Ανάλυση Κανόνων – Casual Πελάτες

Κανόνας 1: Gourmet Gold (Ψάρια → Βοδινό)

Support = 0.002, Confidence = 48.7%, Lift = 144.478

Ο συγκεκριμένος κανόνας παρουσιάζει εξαιρετικά υψηλή τιμή lift, γεγονός που υποδηλώνει ισχυρή συσχέτιση μεταξύ των δύο προϊόντων. Η πιθανότητα συνδυαστικής αγοράς είναι σημαντικά μεγαλύτερη από την τυχαία, γεγονός που δείχνει ότι οι καταναλωτές προτιμούν να αγοράζουν διαφορετικές γεύσεις της ίδιας σειράς προϊόντων. Η συμπεριφορά αυτή συνδέεται με την ανάγκη για ποικιλία.

Κανόνας 2: Somersby apple → mango lime

Support = 0.004, Confidence = 73.1%, Lift = 101.503

Ο κανόνας αυτός εμφανίζει ιδιαίτερα υψηλό confidence, υποδηλώνοντας ότι όταν αγοράζεται η γεύση apple, υπάρχει μεγάλη πιθανότητα αγοράς και της mango lime. Το υψηλό lift επιβεβαιώνει ότι η σχέση αυτή είναι ισχυρή και μη τυχαία. Η συμπεριφορά αυτή υποδεικνύει τάση για κατανάλωση πολλαπλών γεύσεων, πιθανόν στο πλαίσιο κοινωνικής κατανάλωσης.

Κανόνας 3: Noodles (Μοσχάρι → Κοτόπουλο

Support = 0.001, Confidence = 39.7%, Lift = 99.220

Παρατηρείται σημαντική συσχέτιση μεταξύ διαφορετικών γεύσεων του ίδιου προϊόντος. Παρά το χαμηλό support, το υψηλό lift δείχνει ότι η ταυτόχρονη αγορά είναι πολύ πιο πιθανή από το αναμενόμενο. Αυτό υποδηλώνει ότι οι καταναλωτές επιλέγουν εναλλακτικές επιλογές του ίδιου προϊόντος.

Πίνακας 15 - Αποτελέσματα MBA Πελατών Loyal

RULE	Support	Confidence	Lift
Τροφή για γάτες μους Gourmet Gold με Ψάρια Ωκεανού (85gr) → Τροφή για γάτες μους Gourmet Gold με Βοδινό (85gr)	0,004	0,597	86,838
Τροφή για γάτες μους Gourmet Gold με Βοδινό (85gr) → Τροφή για γάτες μους Gourmet Gold με Ψάρια Ωκεανού (85gr)	0,004	0,561	86,838

RULE	Support	Confidence	Lift
Τροφή για γάτες μους Gourmet Gold Με Κοτόπουλο (85gr) → Τροφή για γάτες μους Gourmet Gold με Βοδινό (85gr)	0,002	0,503	73,202
Τροφή για γάτες μους Gourmet Gold με Βοδινό (85gr) → Τροφή για γάτες μους Gourmet Gold Με Κοτόπουλο (85gr)	0,002	0,36	73,202
Τροφή για γάτες μους Gourmet Gold Με Κοτόπουλο (85gr) → Τροφή για γάτες μους Gourmet Gold με Ψάρια Ωκεανού (85gr)	0,002	0,464	71,863
Τροφή για γάτες μους Gourmet Gold με Ψάρια Ωκεανού (85gr) → Τροφή για γάτες μους Gourmet Gold Με Κοτόπουλο (85gr)	0,002	0,353	71,863
Schweppes Pink Grapefruit (330ml), Schweppes Πορτοκάλι με γεύση Άνθος Πορτοκαλιού (330ml) → Schweppes Λεμόνι με γεύση Περγαμόντο - Ιβίσκος (330ml)	0,002	0,594	46,789
Παγωτό πύραυλος Corneto Hazelnut (120ml) → Παγωτό πύραυλος Cornetto Classico (120ml)	0,002	0,459	44,799
Παγωτό πύραυλος Cornetto Classico (120ml) → Παγωτό πύραυλος Corneto Hazelnut (120ml)	0,002	0,229	44,799
Schweppes Pink Grapefruit (330ml), Schweppes Λεμόνι με γεύση Περγαμόντο - Ιβίσκος (330ml) → Schweppes Πορτοκάλι με γεύση Άνθος Πορτοκαλιού (330ml)	0,002	0,717	44,193
Schweppes Λεμόνι με γεύση Περγαμόντο - Ιβίσκος (330ml) → Schweppes Πορτοκάλι με γεύση Άνθος Πορτοκαλιού (330ml)	0,008	0,6	36,983
Schweppes Πορτοκάλι με γεύση Άνθος Πορτοκαλιού (330ml) → Schweppes Λεμόνι με γεύση Περγαμόντο - Ιβίσκος (330ml)	0,008	0,469	36,983
Παγωτό πύραυλος Cornetto Chocolate (120ml) → Παγωτό πύραυλος Cornetto Classico (120ml)	0,003	0,345	33,634
Παγωτό πύραυλος Cornetto Classico (120ml) → Παγωτό πύραυλος Cornetto Chocolate (120ml)	0,003	0,285	33,634
Monster Ultra Red (500ml) → Monster Energy Zero Ultra (500ml)	0,002	0,272	24,151
Monster Energy Zero Ultra (500ml) → Monster Ultra Red (500ml)	0,002	0,182	24,151
Χαρτάκια Rizla Micron (Πακέτο) → Φιλτράκια Rizla Ultra Slim (Πακέτο)	0,002	0,168	23,918
Φιλτράκια Rizla Ultra Slim (Πακέτο) → Χαρτάκια Rizla Micron (Πακέτο)	0,002	0,303	23,918
Φιλτράκια Swan Extra Slim 54’ (Πακέτο) → Χαρτάκια Rizla γαλάζια (Τεμ)	0,003	0,406	22,879

RULE	Support	Confidence	Lift
Χαρτάκια Rizla γαλάζια (Τεμ) → Φιλτράκια Swan Extra Slim 54’ (Πακέτο)	0,003	0,179	22,879
Schweppes Πορτοκάλι με γεύση Άνθος Πορτοκαλιού (330ml), Schweppes Λεμόνι με γεύση Περγαμόντο - Ιβίσκος (330ml) → Schweppes Pink Grapefruit (330ml)	0,002	0,321	21,652
Αραβική πίτα με γαλοπούλα & ένταμ (200gr) → Αραβική πίτα με κοτόπουλο & σως μουστάρδας (200gr)	0,002	0,208	18,517

Ανάλυση Κανόνων – Loyal Πελάτες

Κανόνας 1: Gourmet Gold (Ψάρια → Βοδινό)

Support = 0.004, Confidence = 59.7%, Lift = 86.838

Σε αντίθεση με τους casual πελάτες, εδώ παρατηρείται υψηλότερο support και confidence, γεγονός που υποδηλώνει πιο συχνή και σταθερή επανάληψη της ίδιας αγοραστικής συμπεριφοράς. Αν και το lift είναι χαμηλότερο, η σχέση παραμένει ισχυρή.

Κανόνας 2: Schweppes (συνδυασμός προϊόντων)

Support = 0.002, Confidence = 59.4%, Lift = 46.789

Ο κανόνας αυτός περιλαμβάνει συνδυασμό περισσότερων του ενός προϊόντων που οδηγούν σε αγορά ενός τρίτου. Η υψηλή τιμή confidence δείχνει ότι οι loyal πελάτες έχουν σαφή προτίμηση σε συγκεκριμένες μάρκες και τείνουν να αγοράζουν προϊόντα της ίδιας κατηγορίας συνδυαστικά.

Κανόνας 3: Cornetto Hazelnut → Classico

Support = 0.002, Confidence = 45.9%, Lift = 44.799

Ο συγκεκριμένος κανόνας υποδηλώνει ισχυρή συσχέτιση μεταξύ διαφορετικών εκδοχών του ίδιου προϊόντος. Οι καταναλωτές εμφανίζουν προτίμηση στη μάρκα και επιλέγουν περισσότερες από μία γεύσεις, γεγονός που υποδεικνύει υψηλό επίπεδο πιστότητας (brand loyalty).

Το γενικό συμπέρασμα των ευρημάτων ανέδειξε σημαντικές διαφορές ανάμεσα στις δύο ομάδες πελατών. Οι casual πελάτες παρουσιάζουν πιο υψηλές τιμές lift, κάτι που υποδεικνύει ισχυρούς αλλά λιγότερο τακτικούς συσχετισμούς. Οι loyal πελάτες εμφανίζουν πιο υψηλά επίπεδα support και confidence, κάτι που δείχνει μια πιο σταθερή και προβλέψιμη αγοραστική συμπεριφορά. Κοινό στοιχείο και των δύο ομάδων είναι η προτίμηση σε διαφορετικές γεύσεις του ίδιου προϊόντος, η μεγάλη παρουσία brand loyalty και η ύπαρξη ευκαιριών για στοχευμένες στρατηγικές cross-selling και bundling. Κατά συνέπεια, τα ευρήματα της ανάλυσης μπορούν να χρησιμοποιηθούν για τη βελτίωση των προωθητικών ενεργειών και την ενίσχυση της εμπορικής απόδοσης.

Κεφάλαιο 4

4.1 ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παρούσα μεταπτυχιακή εργασία εστιάστηκε στη μελέτη και την ανάλυση της συμπεριφοράς των χρηστών μιας ψηφιακής πλατφόρμας online παραγγελιών προϊόντων σούπερ μάρκετ, χρησιμοποιώντας τεχνικές data analysis και μεθόδους machine learning. Στους βασικούς στόχους της έρευνας περιλαμβάνονται η κατανόηση των προτύπων καταναλωτικής συμπεριφοράς, η κατηγοριοποίηση των πελατών και η εξαγωγή χρήσιμων συμπερασμάτων που μπορούν να βοηθήσουν στη στρατηγική λήψη αποφάσεων.

Στην αρχή, εξετάστηκε το θεωρητικό υπόβαθρο των Big Data και η σημασία τους στις σύγχρονες επιχειρηματικές δραστηριότητες. Η συνεχής αύξηση τόσο της ποσότητας, όσο και της ποικιλίας των δεδομένων, δημιουργεί την επιτακτική ανάγκη για εξέλιξη προηγμένων τεχνικών ανάλυσης, που θα δίνουν τη δυνατότητα στις επιχειρήσεις να μετατρέπουν τα δεδομένα σε πολύτιμη γνώση. Σε αυτό το πλαίσιο, η ανάλυση δεδομένων πελατών είναι ένα ιδιαίτερα σημαντικό εργαλείο, καθώς βοηθά στην κατανόηση των αναγκών, των προτιμήσεων και της συμπεριφοράς των καταναλωτών.

Έπειτα, έγινε μια εκτίμηση του dataset της πλατφόρμας, που περιελάμβανε δεδομένα για συναλλαγές, προϊόντα και πελάτες. Αυτό το κομμάτι της μελέτης φαίνεται να είναι σημαντικό, γιατί έδωσε τη δυνατότητα για πιο αποτελεσματική επεξεργασία των δεδομένων. Έγιναν και κάποιες διαδικασίες προετοιμασίας, όπως ο καθαρισμός και η οπτικοποίηση των δεδομένων.

Η εφαρμογή της ανάλυσης RFM (Recency, Frequency, Monetary) αποτέλεσε το κύριο εργαλείο για την κατηγοριοποίηση των πελατών. Με την συγκεκριμένη μεθοδολογία, κατέστη δυνατός ο διαχωρισμός των πελατών με βάση τον χρόνο της τελευταίας τους αγοράς, τη συχνότητα συναλλαγών και τη συνολική οικονομική τους αξία. Τα αποτελέσματα ανέδειξαν ότι η πελατειακή βάση δεν είναι ομοιογενής, αλλά διακρίνεται σε ομάδες με διαφορετικά χαρακτηριστικά. Συνολικά, εντοπίστηκαν πέντε βασικές κατηγορίες πελατών.

Περαιτέρω ανάλυση με τη χρήση αλγορίθμων ομαδοποίησης (clustering) οδήγησε στην αναγνώριση φυσικών ομάδων πελατών με κοινά χαρακτηριστικά. Οι αλγόριθμοι αυτοί αποκάλυψαν κρυφές συσχετίσεις στα δεδομένα, διευκολύνοντας τον εντοπισμό τμημάτων της αγοράς που δεν θα μπορούσαν να αναγνωριστούν με παραδοσιακές στατιστικές μεθόδους. Η διαδικασία αυτή ενίσχυσε σημαντικά την κατανόηση της συμπεριφοράς των πελατών και δημιούργησε τις βάσεις για την ανάπτυξη στοχευμένων στρατηγικών marketing.

Ταυτόχρονα, μέσω των δεικτών support, confidence και lift, έγινε αξιολόγηση της συχνότητας και της ισχύος των συνδυαστικών αγορών. Τα ευρήματα έδειξαν πως ορισμένα προϊόντα εμφανίζονται συστηματικά μαζί στις αγορές των καταναλωτών, υποδεικνύοντας την ύπαρξη ισχυρών προτύπων καταναλωτικής συμπεριφοράς. Υψηλές τιμές του δείκτη lift υποδηλώνουν έντονη συσχέτιση μεταξύ προϊόντων, γεγονός που μπορεί να αξιοποιηθεί για την ανάπτυξη στρατηγικών cross-selling και up-selling, ενισχύοντας τη συνολική αξία κάθε συναλλαγής.

Συνολικά, τα αποτελέσματα της έρευνας επιβεβαιώνουν ότι η αξιοποίηση των πελατειακών δεδομένων αποτελεί κρίσιμο παράγοντα επιτυχίας για τις σύγχρονες επιχειρήσεις. Η δυνατότητα ανάλυσης της συμπεριφοράς των πελατών και πρόβλεψης των αναγκών τους επιτρέπει την ανάπτυξη πιο αποτελεσματικών και στοχευμένων στρατηγικών, οδηγώντας σε αυξημένη ικανοποίηση και αφοσίωση των πελατών.

Με βάση τα παραπάνω ευρήματα, διατυπώνονται ορισμένες κρίσιμες προτάσεις για την αξιοποίηση των αποτελεσμάτων στο επίπεδο της επιχείρησης. Για αρχή, η βελτίωση της εμπειρίας χρήστη στην πλατφόρμα αποτελεί καθοριστικό παράγοντα για την ενίσχυση της ικανοποίησης των πελατών. Η συμπεριφορά των καταναλωτών επηρεάζεται άμεσα από παράγοντες όπως η ευχρηστία, η ταχύτητα εξυπηρέτησης και η αξιοπιστία. Η προσθήκη εναλλακτικών στον τρόπο πληρωμής και παράδοσης ενδεχομένως να ενισχύσει την ευελιξία και να αυξήσει τη χρήση της πλατφόρμας.

Επιπρόσθετα, τα αποτελέσματα της ανάλυσης RFM μπορούν να αξιοποιηθούν για την ανάπτυξη προσωποποιημένων στρατηγικών marketing. Οι επιχειρήσεις μπορούν να στοχεύσουν διαφορετικά τμήματα πελατών μέσω εξειδικευμένων ενεργειών, όπως για παράδειγμα προγράμματα επιβράβευσης για πιστούς πελάτες ή ειδικές προσφορές για την επανενεργοποίηση λιγότερο ενεργών χρηστών. Με τον τρόπο αυτό επιτυγχάνεται αποτελεσματικότερη κατανομή πόρων και βελτιστοποίηση των ενεργειών marketing.

Η αξιοποίηση της Ανάλυσης Καλαθιού Αγορών (Market Basket Analysis) μπορεί επίσης να συμβάλλει σημαντικά στην αύξηση των πωλήσεων. Η προβολή συναφών προϊόντων, η δημιουργία πακέτων προσφορών και η βελτιστοποίηση της παρουσίασης των προϊόντων στην πλατφόρμα, ενισχύουν τις συνδυαστικές αγορές, αυξάνοντας τη μέση αξία παραγγελίας και προσφέροντας πιο εξατομικευμένες εμπειρίες στους πελάτες.

Τέλος, η ενσωμάτωση μεθόδων μηχανικής μάθησης στις επιχειρησιακές διαδικασίες μπορεί να αποφέρει σημαντικά οφέλη. Η ανάπτυξη μοντέλων προβλέψεων επιτρέπει την εκτίμηση της μελλοντικής συμπεριφοράς των πελατών, τη βελτίωση της διαχείρισης αποθεμάτων και την έγκαιρη ανίχνευση πιθανών κινδύνων, όπως η αποχώρηση πελατών. Ιδιαίτερη σημασία πρέπει να δοθεί στη διασφάλιση των προσωπικών δεδομένων των πελατών. Η συμμόρφωση με τους κανονισμούς προστασίας δεδομένων και η εφαρμογή κατάλληλων πρακτικών ασφαλείας αποτελούν βασικές προϋποθέσεις τόσο για τη διατήρηση της εμπιστοσύνης των καταναλωτών, όσο και για τη βιωσιμότητα της επιχείρησης.

Συμπερασματικά, η παρούσα μελέτη υποδεικνύει ότι η συνδυαστική αξιοποίηση τεχνικών ανάλυσης δεδομένων, όπως η ανάλυση RFM, οι αλγόριθμοι ομαδοποίησης και η Ανάλυση Καλαθιού Αγορών, προσφέρει ουσιαστική γνώση σχετικά με τη συμπεριφορά των πελατών. Η αξιοποίηση αυτής της γνώσης οδηγεί στη διαμόρφωση αποτελεσματικότερων στρατηγικών, στην αύξηση των εσόδων και στη βελτίωση της συνολικής εμπειρίας των πελατών, προσδίδοντας σημαντικό ανταγωνιστικό πλεονέκτημα στις επιχειρήσεις.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS Quarterly, 36(4), pp. 1165–1188

Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.

Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. Journal of Marketing, 80(6), pp. 97–121.

Davenport, T. H. (2014). Big data at work: Dispelling the myths, uncovering the opportunities. Harvard Business Review Press.

Wedel, M., & Kamakura, W. (2000). Market Segmentation: Conceptual and Methodological Foundations. Springer. pp. 1-10, pp 8-15

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly pp. 1166-1175

Payne, A., & Frow, P. (2005). A Strategic Framework for Customer Relationship Management. Journal of Marketing.) pp.168-178

Bult, J. R., & Wansbeek, T. (1995). Optimal selection for direct mail. Marketing Science. Pp 55-60

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), pp. 651–666.

Wedel, M., & Kamakura, W. A. (2000). Market Segmentation: Conceptual and Methodological Foundations. Springer.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium

Aggarwal, C. C., & Reddy, C. K. (2014). Data Clustering: Algorithms and Applications. CRC Press.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. KDD

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probabilit pp. 281-297

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer. pp 423-450

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B, 39(1), pp. 1–22

McLachlan, G., & Peel, D. (2000). Finite Mixture Models. Wiley pp.20-35

Agrawal, R., & Srikant, R. (1994). Mining association rules between sets of items in large databases. Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), 487–499

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data pp. 1–12

Zaki, M. J. (2000). Fast discovery of association rules. In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pp. 457–460