****

*Εργασία στο μάθημα: « Επιχειρηματική Αναλυτική και Τεχνολογίες Εξατομίκευσης – 7o εξάμηνο» - Β’ Ομαδική Εργασία*

**ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΞΑΓΩΓΗ ΣΥΜΠΕΡΑΣΜΑΤΩΝ ΓΙΑ ΤΗΝ ΕΤΑΙΡΙΑ   
ΑΒ ΒΑΣΙΛΟΠΟΥΛΟΣ**

**ΦΟΙΤΗΤΕΣ**

Σίντι Βρουσάι: 8210017   
Ιάσονας Νεραντζάκης: 8210103   
Κατερίνα Ρούση: 8210130

**Επιβλέπουσα: Α. Πραματάρη**

**Αθήνα – Ιανουάριος 2025**

# **Εισαγωγή**

## **Στόχος**

Στόχος της παρούσας εργασίας είναι η ανάλυση των δεδομένων που μας δόθηκαν από την εταιρία supermarket ΑΒ Βασιλόπουλος Α.Ε.. Μετά από μελέτη των δεδομένων και εφαρμογή των κατάλληλων τεχνικών και αλγορίθμων ανάλυσης, θα εξαχθούν πολύτιμα συμπεράσματα σχετικά με τη χρησιμότητα του κάθε πεδίου δεδομένων, το είδος των πληροφοριών που μπορούν να αντληθούν από αυτά καθώς και τον βέλτιστο τρόπο αξιοποίησής τους, με σκοπό πάντα την συμβολή στον εμπλουτισμό της πολιτικής της εταιρίας και στην αύξηση του ανταγωνιστικού της πλεονεκτήματος.

## **Περιγραφή δεδομένων**

Το σύνολο των δεδομένων περιλαμβάνει πολλαπλά πεδία τα οποία έχουν συλλεχθεί από τα καταστήματα της αλυσίδας supermarket σε πανελλαδική εμβέλεια. Επομένως, κάθε εγγραφή αντιπροσωπεύει ένα κατάστημα, για το οποίο μας παρέχονται οι εξής πληροφορίες:

* Δεδομένα συναλλαγών όπως μέσα έσοδα ανά μήνα, μέσος όγκος πωλήσεων ανά μήνα, μέσα έσοδα από προσφορές και μέσο μερίδιο shopping trips.
* Δεδομένα τμήματος, τα οποία περιλαμβάνουν το ποσοστό των εσόδων στο οποίο συνεισφέρει κάθε τμήμα (με άθροισμα στο 1).
* Χαρακτηριστικά καταστήματος, όπως αριθμός ορόφων, αριθμός ταμείων, ύπαρξη χώρου στάθμευσης και μέγεθος σε τετραγωνικά μέτρα.
* Χαρακτηριστικά περιοχής, όπως είδος τοποθεσίας και wealth index περιοχής.

## **Μεθοδολογία ανάλυσης**

Η ανάλυση που πραγματοποιήθηκε περιλαμβάνει τον καθαρισμό και κανονικοποίηση των δεδομένων που λάβαμε και την εφαρμογή αλγορίθμων συσχετισμού (correlation) και παλινδρόμησης (regression) με σκοπό τον εντοπισμό και την αξιολόγηση του είδους των σχέσεων που έχουν τα διάφορα πεδία μεταξύ τους, καθώς και ποια διαδραματίζουν σημαντικότερο ρόλο στην πρόβλεψη και αξιολόγηση των εσόδων του κάθε καταστήματος.

Στη συνέχεια, και ως βασικό κομμάτι της ανάλυσης, πραγματοποιήθηκε συσταδοποίηση (clustering) των δεδομένων με χρήση αλγορίθμων που κρίθηκαν κατάλληλοι, ώστε να εντοπιστούν κανόνες στη συμπεριφορά των καταναλωτών και να τμηματοποιηθεί το σύνολο των δεδομένων. Απώτερος σκοπός της συσταδοποίησης είναι φυσικά η επίτευξη του διαχωρισμού των καταστημάτων σε κατηγορίες. Έτσι, θα γίνει ευκολότερη η εξατομίκευση της στρατηγικής στο καθένα από αυτά, με στόχο την βελτιστοποίηση των εσόδων και την κατηγοριοποίησή τους με χρήση ουσιαστικών μετρικών, σε αντίθεση με την παρούσα πολιτική, η οποία τα διαχωρίζει με βάση την τοποθεσία τους και μόνο.

Η ανάλυση πραγματοποιήθηκε με χρήση της βιβλιοθήκης pandas της Python,

# **Ανάλυση δεδομένων**

## **Καθαρισμός και Κανονικοποίηση**

Πριν ξεκινήσει η ανάλυση των δεδομένων, κρίθηκε απαραίτητο αυτά να καθαριστούν και να κανονικοποιηθούν ώστε τα αποτελέσματα που θα εξαχθούν να είναι αξιόπιστα.

Ως πρώτο βήμα, έγινε αντικατάσταση των πεδίων που είχαν ένδειξη έλλειψης πληροφορίας (π.χ. 0, n/a κ.α.) με τον μέσο όρο της εκάστοτε στήλης. Ο λόγος που έγινε αντικατάσταση των κενών στοιχείων με ενδεικτικές τιμές και δεν αφαιρέθηκε ολόκληρη η εγγραφή, είναι ο μικρός όγκος των δεδομένων που ήταν διαθέσιμα (319 γραμμές – καταστήματα), γεγονός το οποίο καθιστά κάθε γραμμή εξαιρετικά χρήσιμη.

Συγκριτικά, η στήλη με τα φρέσκα ψάρια και τα θαλασσινά είχε αρκετά κενά στοιχεία, ακριβώς εξήντα, το οποίο ενδεχομένως να ορίζει τα αποτελέσματα για την στήλη αυτή ως μη αντιπροσωπευτικά.

Τέλος, πραγματοποιήθηκε κανονικοποίηση των δεδομένων, διασφαλίζοντας ότι οι πληροφορίες είναι κατάλληλα και σωστά διαμορφωμένες, ώστε να υποστηρίζουν την υλοποίηση μοντέλων, όπως η ανάλυση συσχέτισης.

## **PCA, συσχετίσεις και παλινδρόμηση**

Με σκοπό να αποκαλυφθούν οι σημαντικότερες συσχετίσεις μεταξύ των δεδομένων, εφαρμόστηκε το στατιστικό μοντέλο Pearsons Pairwise Correlation. Ωστόσο, επειδή υπήρχε υψηλή συσχέτιση (multicollinearity) μεταξύ κάποιων χαρακτηριστικών, δημιουργήθηκε το πρόβλημα της υπεραντιπροσώπευσης πληροφορίας, με αποτέλεσμα την ασυνέπεια του στατιστικού μοντέλου και την απώλεια ουσιαστικής πληροφορίας και αξιόπιστου αποτελέσματος. Για να λυθεί το πρόβλημα αυτό, εφαρμόστηκε Principal Component Analysis (PCA). Η διαδικασία PCA μετασχηματίζει τις συσχετισμένες μεταβλητές σε νέες μεταβλητές (κύριες συνιστώσες) που είναι ανεξάρτητες μεταξύ τους. Η λύση μέσω PCA επιτρέπει τη μείωση διαστάσεων και τη δημιουργία γραμμικά μη συσχετισμένων χαρακτηριστικών, αντιμετωπίζοντας τη βασική αιτία του προβλήματος.

Η PCA εφαρμόστηκε στις μεταβλητές που σχετίζονται με τα έσοδα των καταστημάτων (avg\_monthly\_revenue, avg\_monthly\_volume, avg\_monthly\_promo\_revenue, avg\_monthly\_transactions και avg\_basket\_spend).  
 Προέκυψε μία καινούρια βασική συνιστώσα, η revenue\_PCA, η οποία αντιπροσώπευσε το μεγαλύτερο μέρος της διακύμανσης των δεδομένων, δίνοντας pca explained variance ratio ίσο με 0,85. Για τις επόμενες μετρήσεις, όλες οι μεταβλητές εσόδων θα αντιπροσωπεύονται από αυτή.

Στη συνέχεια, υπολογίστηκαν οι συσχετίσεις (correlations) μεταξύ των εσόδων και κάθε μίας από τις μεταβλητές μεγέθους και περιοχής του κάθε καταστήματος. Όπως προαναφέρθηκε, ο λόγος που πραγματοποιήθηκε αυτή η ανάλυση, ήταν για να εντοπιστούν οι μεταβλητές που η μεταβολή τους επηρεάζει περισσότερο τα έσοδα, ώστε αυτές να χρησιμοποιηθούν στη συνέχεια για τη συσταδοποίηση των καταστημάτων. Τα αποτελέσματα παρουσιάζονται στο παρακάτω heatmap.

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά, ορθογώνιο παραλληλόγραμμο

Περιγραφή που δημιουργήθηκε αυτόματα

Όπως φαίνεται, υπάρχει ισχυρή θετική συσχέτιση των εσόδων με τον αριθμό των ταμείων που είναι διαθέσιμα και το μέγεθος του εκάστοτε καταστήματος. Οι μεταβλητές που αντιστοιχούν στο Parking και στο WI (Wealth Index) εμφανίζουν μέτρια συσχέτιση, ενώ η τοποθεσία και ο αριθμός ορόφων σχεδόν μηδενική.

Για διασταύρωση των αποτελεσμάτων, επιλέχθηκε η εφαρμογή επίσης ενός μοντέλου Random Forest Regression, με χρήση των παραπάνω μεταβλητών και την αξιολόγηση του βαθμού στον οποίο επηρεάζουν το revenue\_PCA. Το αποτέλεσμα μετά από την παλινδρόμηση έδειξε πως η πιο σημαντική από τις μεταβλητές είναι αυτή του αριθμού των ταμείων (importance = 0.45), ακολουθεί το μέγεθος του καταστήματος (0.35) και το WI (0,15) και τελευταίες η τοποθεσία και η ύπαρξη πάρκινγκ.

Ε

## **Συσταδοποίηση**

# **Συμπεράσματα και προτάσεις**

## **Συμπεράσματα**

## **Προτάσεις**