ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Τμήμα Πληροφορικής

Εικόνα που περιέχει ορθογώνιο παραλληλόγραμμο, σχεδίαση

Το περιεχόμενο που δημιουργείται από τεχνολογία AI ενδέχεται να είναι εσφαλμένο.

Εργασία Μαθήματος *Αναλυτική Δεδομένων και Μηχανική Μάθηση*

|  |  |
| --- | --- |
| Αρ. Άσκησης - Τίτλος Άσκησης | *Αναλυτική Δεδομένων και Μηχανική Μάθηση* |
| Όνομα φοιτητή - Αρ. Μητρώου | Ραυτόπουλος Μάριος – ΜΠΚΕΔ24034 |
| Ημερομηνία παράδοσης | 15/03/25 |

# Εισαγωγή

## 1.1 Πρόλογος

Στη σύγχρονη εποχή, η κινηματογραφική βιομηχανία και οι πλατφόρμες αναπαραγωγής ταινιών αξιοποιούν ολοένα και περισσότερο τις τεχνικές της αναλυτικής δεδομένων και της μηχανικής μάθησης. Οι εφαρμογές τους εκτείνονται από την εκτίμηση του αν μια ταινία θα γίνει εμπορική επιτυχία, μέχρι την παροχή στοχευμένων προτάσεων περιεχομένου σε χρήστες βάσει των προτιμήσεών τους. Προκύπτει λοιπόν, ένα εύλογο ερώτημα: Γίνεται να προβλεφθεί αν μια ταινία θα γίνει δημοφιλής προτού αυτή κυκλοφορήσει; Στην παρούσα εργασία χρησιμοποιείται το Movie Lens 100K dataset για την διερεύνηση του παραπάνω ερωτήματος. Η προσέγγιση που ακολουθείται δεν εξυπηρετεί μόνο ακαδημαϊκούς σκοπούς, καθώς βρίσκει άμεσες εφαρμογές σε εταιρείες που συμμετέχουν στην κινηματογραφική παραγωγή, είτε για δημιουργία ταινίας, είτε για διαφήμιση μιας ταινίας, είτε για σύσταση σε πλατφόρμες streaming, οι οποίες δείχνουν και την σοβαρότητα της θέσης της στη λήψη αποφάσεων γύρω από τον συγκεκριμένο κλάδο. Οι επόμενες ενότητες καλύπτουν διαδοχικά όλες τις ενέργειες που θα συμβάλουν στην δημιουργία του μοντέλου πρόβλεψης επιτυχίας της ταινίας.

## 1.2 Περιγραφή του συνόλου δεδομένων

Το Movie Lens 100K dataset προέρχεται από το GROUP Lens Research του University of Minnesota και αποτελεί ένα από τα πλέον χρησιμοποιούμενα σύνολα δεδομένων σε ακαδημαϊκές μελέτες για συστήματα συστάσεων.

Περιλαμβάνει δεδομένα ταινιών όπως βαθμολογίες ταινιών, μεταδεδομένα ταινιών, χαρακτηρισμούς και ετικέτες που αποδόθηκαν από χρήστες, καθώς και συνδέσμους σε εξωτερικές βάσεις όπως το IMDB.

# Προ επεξεργασία δεδομένων ( Data Preprocessing)

## Εξερεύνηση και Επεξεργασία δεδομένων

Τα τέσσερα αρχεία CSV (movies, ratings, tags, links) του dataset φορτώθηκαν σε ξεχωριστά dataframes. Πριν την ανάπτυξη μοντέλων, πραγματοποιήθηκαν βασικά βήματα καθαρισμού και προετοιμασίας.

Πιο συγκεκριμένα πραγματοποιήθηκε:

• Έλεγχος Κενών Τιμών: Διαπιστώθηκαν 8 ελλιπείς τιμές στο αρχείο links.csv, οι οποίες συμπληρώθηκαν με -1, ώστε να υποδηλώνεται η απουσία δεδομένου (οι κενές τιμές θα μπορούσαν να προκαλέσουν σφάλματα ή στρεβλώσεις στα μοντέλα).

• Έλεγχος Ακραίων Τιμών : Οι βαθμολογίες των ταινιών κυμαίνονται μεταξύ 1 και 5 με ελάχιστα outliers, επομένως δεν παρουσιάζεται ζήτημα ακραίων τιμών σε αυτό το πεδίο.

• Έλεγχος Διπλότυπων Εγγραφών: Ελέγχθηκε αν υπάρχουν διπλότυπα (που θα μπορούσαν να αλλοιώσουν τα αποτελέσματα ή να μετρήσουν διπλά ορισμένες πληροφορίες) και δεν εντοπίστηκαν επαναλαμβανόμενες εγγραφές στα δεδομένα.

• Αφαίρεση Μη Χρήσιμων Πεδίων: Αφαιρέθηκαν ορισμένες στήλες που δεν θεωρούνται χρήσιμες για την ανάλυση, συγκεκριμένα το timestamp (η χρονική σήμανση της αξιολόγησης) και ο τίτλος της κάθε ταινίας. Τα πεδία αυτά δεν συνεισφέρουν στην πρόβλεψη ή στη συσταδοποίηση, επομένως αφαιρώντας τα απλοποιείται το dataset χωρίς σημαντική απώλεια πληροφορίας.

## Μετασχηματισμοί

Μετά τον καθαρισμό των δεδομένων, πραγματοποιήθηκαν επιπλέον μετασχηματισμοί στα δεδομένα για να καταστούν κατάλληλα για ανάλυση από αλγορίθμους μηχανικής μάθησης.

Ειδικότερα:

• One-Hot Encoding (Κατηγοριοποίηση Ειδών): Το πεδίο των ειδών ταινίας που περιέχει συμβολοσειρές με είδη (π.χ. action, animation, comedy) μετασχηματίστηκε σε διακριτά δυαδικά χαρακτηριστικά. Για κάθε ταινία, διαχωρίστηκαν τα είδη σε μεμονωμένες εγγραφές και στη συνέχεια εφαρμόστηκε one-hot encoding: δημιουργήθηκε μία στήλη για κάθε είδος με τιμή 1 ή 0 ανάλογα με το αν η ταινία ανήκει ή όχι σε αυτό το είδος. Τέλος, τα δεδομένα ομαδοποιήθηκαν ξανά ανά αναγνωριστικό ταινίας και συνενώθηκαν, ώστε κάθε ταινία να αντιπροσωπεύεται από μία γραμμή με πολλαπλά χαρακτηριστικά που υποδεικνύουν την παρουσία ή απουσία κάθε είδους.

• Κανονικοποίηση Βαθμολογιών: Εφαρμόστηκε κλιμάκωση StandardScaler στις βαθμολογίες των ταινιών. Μετά τον μετασχηματισμό αυτό, οι βαθμολογίες έχουν μέσο όρο 0 και τυπική απόκλιση 1. Η διαδικασία αυτή, εξισορροπεί την κλίμακα των βαθμολογιών, έτσι ώστε καμία ταινία να μην υπερτερεί λόγω μεγέθους τιμών.

• Ενοποίηση Δεδομένων: Τα επεξεργασμένα δεδομένα βαθμολογιών και τα χαρακτηριστικά είδους (one-hot encoded genres) συνενώθηκαν σε ένα ενιαίο dataset. Κάθε ταινία πλέον αναπαρίσταται από ένα σύνολο αριθμητικών χαρακτηριστικών: τη κανονικοποιημένη βαθμολογία της και πληθώρα δεικτών (0/1) για την κατηγορία της. Αυτό το τελικό DataFrame των χαρακτηριστικών θα χρησιμοποιηθεί τόσο στη συσταδοποίηση όσο και στην ταξινόμηση.

## Ανάλυση δεδομένων

Έχοντας λοιπόν το τελικό DataFrame ξεκινάει η διερευνητική ανάλυση δεδομένων και η χρήση διαγραμμάτων για να κατανοηθούν καλύτερα οι κατανομές και οι συσχετίσεις των δεδομένών.

Κατανομή βαθμολογιών:

• Οι περισσότερες ταινίες έχουν μέση βαθμολογία συγκεντρωμένη μεταξύ 0 και 1 (ελαφρώς πάνω από τον μέσο όρο του dataset).

• Υπάρχει όμως και σημαντικός όγκος ταινιών με αρνητικές κανονικοποιημένες βαθμολογίες, δηλαδή αρκετές ταινίες έλαβαν βαθμολογίες κάτω του συνολικού μέσου όρου.

A graph with blue lines and a blue line

AI-generated content may be incorrect.• Η διασπορά των τιμών δείχνει ότι οι αξιολογήσεις παρουσιάζουν μια σχετική διακύμανση γύρω από το μηδέν, αλλά οι περισσότερες δεν απέχουν δραματικά από τον μέσο όρο.

Εικόνα 1: Histogram with kde for the distribution of normalized ratings

Κατανομή ειδών ταινίας:

• Το Drama είναι το είδος με τις περισσότερες ταινίες στο dataset (περίπου 4.500 εμφανίσεις ειδών σε ταινίες αν συμπεριλάβουμε τις πολλαπλές κατηγορίες).

• Ακολουθεί το Comedy με περίπου 3.500 εμφανίσεις.

• Αντίθετα, σπανιότερα εμφανίζονται είδη όπως το Musical, το Western, το Film-Noir και το IMAX, καθώς και περιπτώσεις ταινιών που δεν έχουν δηλωμένο είδος. Αυτές οι κατηγορίες είναι ελάχιστες στο σύνολο.

Εικόνα : Bar Plot for the distribution of movie genres

*A graph of a number of movies

AI-generated content may be incorrect.*

Συσχέτιση χαρακτηριστικών:

• Δεν παρατηρείται κάποια ισχυρή συσχέτιση μεταξύ της βαθμολογίας μιας ταινίας και οποιουδήποτε συγκεκριμένου είδους.

• Κάποια genres εμφανίζουν μεταξύ τους μετρίου βαθμού συσχέτιση, υποδηλώνοντας ότι συχνά συνυπάρχουν στις ίδιες ταινίες. Για παράδειγμα, τα είδη Animation και Children εμφανίζονται μαζί αρκετά συχνά, όπως και τα Action με το Adventure.

• Η ισχυρότερη συσχέτιση μεταξύ δύο συγκεκριμένων κατηγοριών είναι 0.49, ανάμεσα σε Animation και Children. Αυτό είναι αναμενόμενο, καθώς πολλές ταινίες κινουμένων σχεδίων απευθύνονται σε παιδιά.

• Η απουσία πολύ υψηλών συσχετίσεων, τιμές κοντά στο 1, υποδηλώνει ότι το dataset περιλαμβάνει μεγάλη ποικιλία ταινιών.

A graph with numbers and a red line

AI-generated content may be incorrect.

Εικόνα : Feature Correlation Heatmap

# Συσταδιοποίηση (Clustering):

## 3.1 Σχεδιασμός και σκοπός

Με την ολοκλήρωση της προ επεξεργασίας και ανάλυσης δεδομένων αποκτήθηκε εικόνα επί του συνόλου των δεδομένων, με το επόμενο βήμα να αποτελεί η συσταδιοποίηση. Κύριος στόχος της συγκεκριμένης διαδικασίας αποτελεί η ανακάλυψη μοτίβων και συσχετίσεων των δεδομένων που δεν είναι ορατές μέσω απλής στατιστικής ανάλυσης.

Συνοπτικά, ο σχεδιασμός για το clustering έχει ως εξής:

• Αλγόριθμοι: Θα εφαρμοστούν δύο διαφορετικοί αλγόριθμοι, ο K-Means και ο DBSCAN. Οι αλγόριθμοι αυτοί αντιπροσωπεύουν διαφορετικές προσεγγίσεις: ο K-Means βασίζεται στην εύρεση κεντροειδών (centroid-based clustering), ενώ ο DBSCAN βασίζεται στην έννοια της πυκνότητας των σημείων (density-based clustering).

• Αξιολόγηση Συστάδων: Για να εκτιμηθεί η ποιότητα των clusters που προκύπτουν από τους παραπάνω αλγορίθμους, θα χρησιμοποιηθούν διαγράμματα και μετρικές αξιολόγησης των συστάδων.

Πιο συγκεκριμένα θα υπολογιστούν: Inertia, Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index. Οι παραπάνω μετρικές θα αναλυθούν και θα αξιοποιηθούν με σκοπό να συγκριθεί η αποτελεσματικότητα των αλγορίθμων clustering.

## 3.2 Επιλογή χαρακτηριστικών και επεξεργασία

Για την διασφάλιση ότι οι αλγόριθμοι clustering θα λειτουργήσουν αποδοτικά, εφαρμόστηκαν κάποια επιπλέον βήματα προετοιμασίας στα χαρακτηριστικά:

• Επιλογή Χαρακτηριστικών: Ως χαρακτηριστικά για το clustering χρησιμοποιήθηκαν οι κανονικοποιημένες βαθμολογίες των ταινιών και οι δυαδικές μεταβλητές των genres (one-hot encoded genres). Αγνοούνται τα πεδία IDs (movieId, userId), καθώς είναι απλώς αναγνωριστικά και δεν περιέχουν πληροφορία περιεχομένου.

• Εφαρμόστηκε εκ νέου StandardScaler σε όλα τα αριθμητικά χαρακτηριστικά (περιλαμβάνει πλέον την normalized βαθμολογία και τις στήλες genres 0/1). Αυτό είναι σημαντικό για τον K-Means, ο οποίος στηρίζεται σε αποστάσεις. Αυτή η κλιμάκωση διασφαλίζει ότι καμία μεταβλητή δεν κυριαρχεί λόγω μονάδων μέτρησης.

• Μειώθηκε η διάσταση του διανύσματος χαρακτηριστικών εφαρμόζοντας Principal Component Analysis (PCA) και κρατώντας τις 10 κυριότερες συνιστώσες. Δεδομένου ότι μετά το one-hot encoding το πλήθος χαρακτηριστικών είναι αρκετά μεγάλο, η μείωση διαστάσεων βοηθά να μετριαστεί το φαινόμενο “curse of dimensionality” (στις πολύ υψηλές διαστάσεις οι αποστάσεις χάνουν την ερμηνεία τους). Με το PCA φιλτράρεται επίσης πιθανός θόρυβος στα δεδομένα.