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 |

Contents

[1. Εισαγωγή 4](#_Toc192750157)

[1.1 Περιγραφή του Συνόλου Δεδομένων 4](#_Toc192750158)

[2. Προεπεξεργασία Δεδομένων (Data Preprocessing) 5](#_Toc192750159)

[2.1 Φόρτωση Δεδομένων 5](#_Toc192750160)

[2.2 Εξερεύνηση και Επεξεργασία Δεδομένων 5](#_Toc192750161)

[2.3 Μετασχηματισμοί 6](#_Toc192750162)

[2.4 Ανάλυση Δεδομένων 6](#_Toc192750163)

[3. Συσταδοποίηση (Clustering) 11](#_Toc192750164)

[3.1 Σκοπός & Σχεδιασμός 11](#_Toc192750165)

[3.2 Επιλογή χαρακτηριστικών και επεξεργασία 11](#_Toc192750166)

[3.3 K-Means 12](#_Toc192750167)

[3.3.1 Υπολογισμός μετρικών για πιθανά Κ 12](#_Toc192750168)

[3.3.2 Διαγράμματα 12](#_Toc192750169)

[3.4 DBSCAN 13](#_Toc192750170)

[3.5 Ερμηνεία και Σύγκριση Μεθόδων 14](#_Toc192750171)

[4. Ταξινόμηση (Classification) 20](#_Toc192750172)

[4.1 Ορισμός Στόχου (Label) 20](#_Toc192750173)

[4.2 Προετοιμασία συνόλου δεδομένων 20](#_Toc192750174)

[4.3 Δημιουργία και Εκπαίδευση Μοντέλων 21](#_Toc192750175)

[4.3.1 Support Vector Machine (SVM) 21](#_Toc192750176)

[4.3.2 Neural Network (Multi-Layer Perceptron) 21](#_Toc192750177)

[4.4 Αξιολόγηση Μεθόδων 22](#_Toc192750178)

[4.4.1 Αξιολόγηση SVM, MLP 22](#_Toc192750179)

[4.4.2 Σύγκριση SVM, MLP 22](#_Toc192750180)

[5. Συμπεράσματα 23](#_Toc192750181)

# Εισαγωγή

Η βιομηχανία του κινηματογράφου και του streaming ενσωματώνει όλο και περισσότερο τεχνικές Αναλυτικής Δεδομένων και Μηχανικής Μάθησης για τη λήψη αποφάσεων: από το να εκτιμήσει κανείς αν μια ταινία πρόκειται να γίνει εμπορική επιτυχία, μέχρι να προτείνει στοχευμένες ταινίες σε χρήστες με βάση τις προτιμήσεις τους.

Δημιουργείται λοιπόν το εύλογο ερώτημα, μπορεί ένα σύστημα μηχανικής μάθησης να προβλέψει αν μια νέα ταινία θα γίνει δημοφιλής προτού καν κυκλοφορήσει; Με όγκο δεδομένων όπως βαθμολογίες χρηστών, είδη (genres), ημερομηνίες κυκλοφορίας και μεταδεδομένα παραγωγής, υπάρχει η δυνατότητα να φτιάξουμε μοντέλα που να προσεγγίζουν μια τέτοια πρόβλεψη.

Στην παρούσα εργασία, αξιοποιούμε το MovieLens 100K Dataset, ένα σύνολο με περίπου 100.000 βαθμολογίες σε ταινίες, για να δείξουμε πώς μπορεί να υλοποιηθεί μια τέτοια ανάλυση και μοντελοποίηση.

Η δουλειά αυτή δεν είναι απλώς ακαδημαϊκή, αλλά βρίσκει εφαρμογή σε κινηματογραφικές παραγωγές, αλλά και πλατφόρμες streaming, όπως το Netflix, οι οποίες επιθυμούν να γνωρίζουν προκαταβολικά αν θα επενδύσουν σε μια ταινία ή να προωθήσουν συγκεκριμένους τίτλους. Μελετώντας επομένως τα μοτίβα (clusters) και το μοντέλο πρόβλεψης “δημοφιλίας”, μπορούμε να δούμε πώς η Αναλυτική Δεδομένων υποστηρίζει τη λήψη αποφάσεων στην κινηματογραφική βιομηχανία, από το marketing μέχρι τον σχεδιασμό ενός recommendation system.

## Περιγραφή του Συνόλου Δεδομένων

Το MovieLens dataset προέρχεται από το GroupLens Research του University of Minnesota, είναι ένα από τα πλέον χρησιμοποιούμενα σε ακαδημαϊκές δημοσιεύσεις για συστήματα συστάσεων.

Το **MovieLens 100K** αποτελείται από:

* **ratings.csv**: πίνακας βαθμολογιών ταινιών (userId, movieId, rating, timestamp).
* **movies.csv**: πίνακας με τα μεταδεδομένα των ταινιών (movieId, title, genres).
* **tags.csv:** πίνακας με χαρακτηρισμούς ταινιών (userId, movieId, tag, timestamp)
* **links.csv:** πίνακας με τα links των ταινιών στα imdb, themoviedb (movieId, imdbId, tmdbId)

**Βασικά Στατιστικά**

* Αριθμός εγγραφών: περίπου 100.000.
* Χρήστες: περίπου 943.
* Ταινίες: περίπου 1.682.

# 2. Προεπεξεργασία Δεδομένων (Data Preprocessing)

## 2.1 Φόρτωση Δεδομένων

Διαβάζουμε τα 4 csv αρχεία και τα περνάμε σε ξεχωριστά pandas data frames. Το MovieLens dataset είναι σπασμένο σε πολλαπλά αρχεία, οπότε πρέπει πρώτα να τα φορτώσουμε για να τα επεξεργαστούμε.

## 2.2 Εξερεύνηση και Επεξεργασία Δεδομένων

* Έλεγχος Κενών Τιμών (Οι ελλιπείς τιμές μπορούν να προκαλέσουν σφάλματα ή στρεβλώσεις στα μοντέλα): Διαπιστώθηκαν 8 μη διαθέσιμες τιμές στο links dataset, οι οποίες συμπληρώθηκαν με -1 για να δείξουμε ότι εδώ δεν υπάρχει πραγματική τιμή.
* Έλεγχος για Ακραίες Τιμές (Outliers): Δεν υπήρχε ουσιαστικό ζήτημα outliers στις βαθμολογίες (που κυμαίνονται σε [1,5]).
* Έλεγχος διπλότυπων εγγραφών (Οι διπλές εγγραφές ενδέχεται να αλλοιώσουν τα αποτελέσματα και να φουσκώσουν τεχνητά τις μετρήσεις): Δεν εντοπίστηκαν διπλότυπες εγγραφές.
* Αφαίρεση δεδομένων που δεν είναι χρήσιμα: Αφαιρέθηκαν οι στήλες timestamp , title καθώς ο χρόνος που δόθηκε η κριτική, αλλά και ο τίτλος μιας ταινίας δεν αποτελούν σημαντικά δεδομένα για την δική μας ανάλυση.

## 2.3 Μετασχηματισμοί

* **One-Hot Encoding**: Μετατρέπουμε τη στήλη genres (που περιέχει string “Action|Comedy|Drama” κ.λπ.) σε λίστα μεμονωμένων ειδών (π.χ. [“Action”, “Comedy”, “Drama”]). Exploding έτσι ώστε κάθε γραμμή να περιέχει ένα μόνο genre. Εφαρμόζουμε one-hot encoding ώστε να μετατρέψουμε κάθε είδος σε ξεχωριστή στήλη που θα έχει τιμή 0 ή 1. Ομαδοποιούμε πίσω (groupby) και επανενώνουμε (join), ώστε ο κάθε movieId να έχει στήλες 0 ή 1 για κάθε genre. Με αυτόν τον τρόπο έχουμε για κάθε είδος στο οποίο ανήκει η ταινία σε διακριτή, αριθμητική μορφή ώστε να χρησιμοποιηθούν σε μοντέλα μηχανικής μάθησης (clustering, classification).
* **Κανονικοποίηση βαθμολογιών**: Εφαρμόζεται StandardScaler στις βαθμολογίες ταινιών και έτσι μετατρέπονται σε κατανομή με μέσο όρο=0 και τυπική απόκλιση=1. Με αυτόν τον τρόπο εξισορροπείται η βαρύτητα των βαθμολογιών, γεγονός που βοηθάει και την αποτύπωση αλγορίθμων που χρησιμοποιούν αποστάσεις.
* **Ενοποίηση δεδομένων:** Διαμορφώνεται ένα ενιαίο Data frame το οποίο περιέχει την βαθμολογία (normalized) και τις κατηγορίες (one-hot-encoded).

## 2.4 Ανάλυση Δεδομένων

Προχωράμε σε παραγωγή διαγραμμάτων για να δούμε στατιστικά και συσχετίσεις των δεδομένων.

* + 1. Histogram των κανονικοποιημένων βαθμολογιών: Παρατηρούνται πως κατανέμονται οι βαθμολογίες μαζί με την πυκνότητά τους.
    2. Box plot των κανονικοποιημένων βαθμολογιών:

Παρατηρείται η κατανομή των τιμών των βαθμολογίων και οι ακραίες τιμές.

* + 1. Bar plot των ταινιών ανά είδος: Παρατηρείται ο αριθμός των ταινιών ανά κατηγορία.
    2. Correlation Heatmap μεταξύ των στηλών του ενοποιημένου data frame Παρατηρούνται οι συσχετίσεις ανάμεσα στα είδη των ταινιών.

A graph with blue lines and a blue line

AI-generated content may be incorrect.

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

Σύμφωνα με το διάγραμμα:

Οι περισσότερες βαθμολογίες συγκεντρώνονται γύρω από 0 έως 1, δηλαδή λίγο πάνω από τον μέσο όρο.

Υπάρχει επίσης αρκετός όγκος σε αρνητικές τιμές, που σημαίνει αρκετές βαθμολογίες κάτω του μέσου όρου.

Η διασπορά δείχνει πόσο σκόρπιες είναι οι βαθμολογίες γύρω από τον μέσο μετά την κανονικοποίηση.

A blue rectangular box with black lines

AI-generated content may be incorrect.

Εικόνα 2: Box Plot of normalized ratings

Σύμφωνα με το διάγραμμα:

Οι περισσότερες κανονικοποιημένες τιμές βρίσκονται ανάμεσα στο -1 και το +1.

Η διάμεσος είναι πάνω από το 0, άρα οι ταινίες παίρνουν ελαφρώς υψηλότερες βαθμολογίες από τον «ουδέτερο» μέσο όρο.

Υπάρχουν λίγες ακραίες κανονικοποιημένες τιμές αρκετά χαμηλά (π.χ. -3), που αντιπροσωπεύουν κάποιες ταινίες οι οποίες έχουν λάβει πολύ χαμηλές βαθμολογίες συγκριτικά με το γενικό σύνολο.

**A graph of a number of movies

AI-generated content may be incorrect.**

Εικόνα 3: Distribution of movie genres

Σύμφωνα με το διάγραμμα:

Το Drama είναι το πιο συνηθισμένο είδος (περίπου 4.500 ταινίες).

Ακολουθεί το Comedy με περίπου 3.500 ταινίες.

Σπανιότερα είδη είναι τα Musical, Western, IMAX, Film-Noir και οι ταινίες χωρίς συγκεκριμένη κατηγορία.

A graph with numbers and a red line

AI-generated content may be incorrect.

Εικόνα 4: Feature Correlation Heatmap

Σύμφωνα με το διάγραμμα:

Δεν υπάρχει έντονη εξάρτηση μεταξύ κάποιας μεμονωμένης στήλης (π.χ. rating) και των genres.

Κάποια είδη ταινιών (Animation–Children, Action–Adventure, Musical–Comedy) συνυπάρχουν μετρίως συχνά, οπότε έχουν θετική αλλά όχι υπερβολικά υψηλή συσχέτιση.

Η πιο δυνατή συσχέτιση είναι το Animation–Children στο 0.49.

Η έλλειψη ισχυρών συσχετίσεων υποδηλώνει ότι το dataset έχει αρκετή ποικιλία.

# 3. Συσταδοποίηση (Clustering)

## 3.1 Σκοπός & Σχεδιασμός

Αφού ολοκληρώσαμε την προεπεξεργασία των δεδομένων και αποκτήσαμε εικόνα για τη δομή του MovieLens 100K Dataset (διανομές βαθμολογιών, δημοφιλή είδη κ.λπ.), το επόμενο βήμα είναι η συσταδοποίηση (clustering). Στόχος μας είναι να ανακαλύψουμε μοτίβα ή ομάδες (clusters) εντός των δεδομένων που δεν είναι άμεσα ορατά με την απλή στατιστική ανάλυση. Ειδικότερα:

* **Επιλογή Χαρακτηριστικών**: Χρησιμοποιούμε τις κανονικοποιημένες βαθμολογίες (rating) και τα one-hot genres, παραλείποντας πεδία που δεν συνεισφέρουν (π.χ. movieId, userId).
* **Χρήση Κλασικών Αλγορίθμων**: Θα εφαρμόσουμε K-Means και DBSCAN, δύο διαφορετικές φιλοσοφίες (κεντροειδή vs. πυκνότητας).
* **Αξιολόγηση Συστάδων**: Υπολογίζουμε μετρικές ποιότητας (Inertia, Silhouette, Davies-Bouldin, Calinski-Harabasz) για να συγκρίνουμε τις λύσεις που προκύπτουν από κάθε αλγόριθμο και να επιλέξουμε την καλύτερη.

## 3.2 Επιλογή χαρακτηριστικών και επεξεργασία

Αφαιρούμε τους αναγνωριστικούς αριθμούς (movieId, userId), επειδή δεν συνεισφέρουν στην ομαδοποίηση. Εφαρμόζουμε StandardScaler για να φέρουμε όλα τα χαρακτηριστικά σε κλίμακα μέσου όρου 0 και τυπικής απόκλισης 1. Περιορίζουμε τον χώρο των χαρακτηριστικών σε 10 κύριες συνιστώσες (Principal Components). Σε υψηλές διαστάσεις, οι ευκλείδειες αποστάσεις χάνουν τη σημασία τους (“curse of dimensionality”). Η εφαρμογή PCA βελτιώνει την απόδοση των αλγορίθμων συσταδοποίησης (πιο γρήγορη εκτέλεση, λιγότερο noise).

Έτσι επιταχύνουμε το Clustering και βελτιώνουμε την ποιότητά του.

## 3.3 K-Means

### 3.3.1 Υπολογισμός μετρικών για πιθανά Κ

Για κάθε k από 2 έως 10, εκπαιδεύουμε ένα μοντέλο K-Means (n\_clusters=k) στα δεδομένα features\_reduced.

Μαζεύουμε διάφορες μετρικές ποιότητας συσταδοποίησης:

**Inertia** (άθροισμα ενδο-αποστάσεων):Όσο μικρότερη τόσο πιο συμπαγείς συστάδες έχουμε.

**Silhouette Score**: Όσο μεγαλύτερο τόσο καλύτερος διαχωρισμός.

**Davies-Bouldin Index** (clustering spread):Όσο μικρότερο τόσο το καλύτερο.

**Calinski-Harabasz Index**(ευδιάκριτα clusters) Όσο μεγαλύτερο τόσο καλύτερος διαχωρισμός .

Αποθηκεύουμε επίσης τη κατανομή των σημείων (cluster\_counts) σε κάθε συστάδα, και οπτικοποιούμε αυτές τις τιμές με bar plot για να δούμε αν υπάρχει ισορροπία μεγέθους μεταξύ των συστάδων.

Θα προχωρήσουμε σε υπολογισμό όλων των παραπάνω μετρικών και θα δημιουργήσουμε διαγράμματα για να πετύχουμε το καλύτερο δυνατό clustering.

### 3.3.2 Διαγράμματα

A graph with a line

AI-generated content may be incorrect.

Εικόνα 4: Inertia plot

A graph with red dots

AI-generated content may be incorrect.

Εικόνα 5 : Silhouette score for different K values

Η επιλογή του K είναι κρίσιμη για την ποιότητα του K-Means. Οι διάφορες μετρικές προσφέρουν συμπληρωματική εικόνα για το πώς τα δεδομένα ομαδοποιούνται. Συγκρίνοντας όλες τις παραπάνω μετρικές για όλα τα πιθανά K και σύμφωνα και με τα παραπάνω διαγράμματα καταλήγουμε σε κ=3.

## 3.4 DBSCAN

Δοκιμάζουμε διάφορους συνδυασμούς παραμέτρων eps και min\_samples για τον DBSCAN.

Αν ο αλγόριθμος βρει τουλάχιστον 2 clusters (δηλ. αν set(labels) έχει περισσότερα από 1 στοιχεία), μετράμε τα silhouette\_score, davies\_bouldin\_score, calinski\_harabasz\_score, καθώς και το πόσα σημεία χαρακτηρίστηκαν ως θόρυβος (label = -1).

Ενημερώνουμε τους «βέλτιστους» υπερπαραμέτρους eps και min\_samples όταν βρίσκουμε υψηλότερο Silhouette.

Ανακοινώνουμε τις βέλτιστες παραμέτρους (best\_eps, best\_min\_samples) που βρήκαμε.

«Κλειδώνουμε» τον DBSCAN σε αυτά τα βέλτιστα.

Βγάζουμε scatter plot των δεδομένων (μετά από 2D-PCA), χρωματίζοντας κάθε σημείο με βάση το cluster ID που βρήκε το DBSCAN. Το -1 (εάν υπάρχει) συνήθως αντιστοιχεί σε θόρυβο (outliers).

## 3.5 Ερμηνεία και Σύγκριση Μεθόδων

A graph of a number of green squares

AI-generated content may be incorrect.

Εικόνα 6: Average ratings per cluster

1. Το γράφημα δείχνει ότι τα τρία clusters διαφέρουν ελαφρώς στην κανονικοποιημένη (normalized) μέση βαθμολογία.
   * Cluster 0: Εμφανίζει μέση βαθμολογία λίγο πάνω από το
   * Cluster 1: Έχει ακόμα πιο θετική βαθμολογία, φτάνοντας 0.02
   * Cluster 2: Φαίνεται να συγκεντρώνει ταινίες με χαμηλότερο μέσο rating

Αυτό υποδεικνύει ότι τα clusters δεν διαχωρίζονται μόνο βάσει genre, αλλά και με βάση τη συνολική «δημοφιλία» ή προτίμηση των χρηστών.

A graph of a bar chart

AI-generated content may be incorrect.

Εικόνα 7: average presence of genre Action.

Παρατηρείται ότι στο Cluster 1 έχουμε πολύ υψηλό ποσοστό ταινιών Action (πάνω από 0.5 ως μέση τιμή, δηλ. περίπου το 50% των ταινιών εκεί έχουν Action = 1).

Αντίθετα, στο Cluster 0 η τιμή είναι σχεδόν 0.1, ενώ στο Cluster 2 γύρω στο 0.2-0.3.

Έτσι, το Cluster 1 είναι Action-oriented.

A graph of a graph

AI-generated content may be incorrect.

Εικόνα 8: average presence of genre Adventure

**Cluster 2** φαίνεται να έχει τη μεγαλύτερη μέση τιμή (~0.55), άρα πολλές ταινίες σε αυτό το cluster είναι Adventure.

**Cluster 1** έχει και αυτό αρκετό Adventure (~0.3), ενώ το **Cluster 0** είναι σχεδόν ανύπαρκτο σε Adventure.

Πάντως παρατηρείται ότι το είδος adventure, χαρακτηρίζει τη πλειοψηφία των ταινιών.

A graph with a bar

AI-generated content may be incorrect.

Εικόνα 9: average presence of genre animation

Το **Cluster 2** δεσπόζει με ~0.35, δηλαδή περίπου 35% των ταινιών εκεί είναι Animation.

Τα άλλα 2 clusters έχουν αμελητέο Animation (~0 ή 0.01).

Επομένως, το **Cluster 2** θα μπορούσε να ερμηνευτεί ως “Animation/Adventure” (και πιθανώς πιο παιδικό) σε μεγάλο ποσοστό.

A diagram of a cluster of dots

AI-generated content may be incorrect.

Εικόνα 10: K-means clusters

**Τι σημαίνουν τα 3 Clusters:**

* **Cluster 0** (purple): Τείνει να έχει μέση βαθμολογία ~0, δηλαδή μέτρια/ελαφρώς θετική βαθμολογία, χαμηλά ποσοστά Action, Adventure, Animation. Μπορεί να αφορά ταινίες **Drama/Comedy** που δεν είναι έντονα «περιπετειώδεις» ή «κινούμενα σχέδια».
* **Cluster 1** (teal): Πολύ Action, αρκετό Adventure, αλλά όχι Animation. Μάλλον ταινίες δράσης/περιπέτειας που παίρνουν και υψηλότερες βαθμολογίες (positive rating).
* **Cluster 2** (yellow): Έχει σημαντική παρουσία σε **Adventure** και **Animation**, αλλά η βαθμολογία είναι ελαφρώς κάτω του μέσου όρου. Ίσως περιέχει ταινίες πιο «παιδικές» ή «Animation/Adventure» που δεν έτυχαν τόσο καλής υποδοχής.

Σύγκριση με DBSCAN:

A diagram of red and blue dots

AI-generated content may be incorrect.

Εικόνα 11:DBSCAN clusters

Στην απεικόνιση DBSCAN (2D PCA) παρατηρείτε ότι εμφανίζονται πολλές διαφορετικές ετικέτες (0, 400, 800, 1200 κ.λπ.), υποδηλώνοντας ότι το DBSCAN παράγει πολυάριθμα μικρά clusters Αυτό:

* Δυσχεραίνει την ερμηνεία, διότι δεν έχουμε μόνο 3-5 ομάδες, αλλά δεκάδες/εκατοντάδες μικρές ομάδες.
* Η υψηλή τιμή Silhouette που μπορεί να προκύψει στο DBSCAN δείχνει «ισχυρή πυκνότητα» σε μικρά κομμάτια, όχι απαραίτητα χρήσιμα ως μεγάλες, διακριτές κατηγορίες.

Η εφαρμογή K-Means και DBSCAN στο MovieLens dataset ανέδειξε δύο διαφορετικές εικόνες:

O K-Means με μικρό αριθμό clusters (π.χ. 3–5) προσφέρει λίγες, ευδιάκριτες ομάδες ταινιών με ξεκάθαρη ερμηνεία (διαφορά στα genres, μέση βαθμολογία κ.λπ.)

O DBSCAN – ειδικά με τις τρέχουσες παραμέτρους – οδηγεί σε πολυάριθμες μικροσυστάδες που δυσχεραίνουν τη συνολική ανάγνωση. Παρά την ενδεχομένως υψηλή τιμή Silhouette στο DBSCAN, οι εκατοντάδες clusters που προκύπτουν δεν μπορούν να αξιοποιηθούν εύκολα σε πρακτικό πλαίσιο.

Συνεπώς ο K-Means είναι πιο κατάλληλη επιλογή από τον DBSCAN, για την κατηγοριοποίηση των ταινιών σε λιγότερες και , “ποιοτικές” ομάδες.

# 4. Ταξινόμηση (Classification)

## 4.1 Ορισμός Στόχου (Label)

Για τη δημιουργία μοντέλου ταξινόμησης, ορίσαμε τη δυαδική ετικέτα **is\_popular** για τις ταινίες, με σκοπό να προβλέψουμε τη αν η ταινία είναι δημοφιλής.

Κριτήριο: Αν ο αριθμός βαθμολογιών μίας ταινίας ξεπερνά τη διάμεσο του συνολικού δείγματος, τότε is\_popular = 1, αλλιώς 0.

Για κάθε ταινία (movieId), υπολογίζουμε πόσες βαθμολογίες (ratings) έχει λάβει συνολικά. Στη συνέχεια, θεωρούμε «δημοφιλή» (is\_popular=1) όποια ταινία έχει αριθμό βαθμολογιών πάνω από τον διάμεσο (median). Αλλιώς την κατατάσσουμε ως «μη δημοφιλή» (is\_popular=0).

## 4.2 Προετοιμασία συνόλου δεδομένων

Ενώνουμε τα μετα-δεδομένα των ταινιών που έχουν ήδη μετατραπεί σε one-hot encoding με την καινούρια στήλη is\_popular σε ένα ενιαίο data frame movies\_with\_features. Στη συνέχεια, διαχωρίζουμε Το X που περιλαμβάνει όλα τα χαρακτηριστικά (columns) εκτός από το movieId και το is\_popular. Το y είναι το label is\_popular (0 ή 1).

Χωρίζουμε το dataset σε 80% εκπαίδευση (train) και 20% έλεγχο (test).

Χρησιμοποιούμε stratify=y ώστε να διατηρηθεί η αναλογία (proportion) δημοφιλών/μη-δημοφιλών και στα δύο υποσύνολα.

Εφαρμόζουμε κλιμάκωση (StandardScaler) στα χαρακτηριστικά, για βέλτιστη συμπεριφορά αλγορίθμων όπως το SVM ή τα Νευρωνικά.

Το Train/Test split επιτρέπει αξιόπιστη εκτίμηση της απόδοσης σε αχαρτογράφητα δεδομένα.

Η κλιμάκωση βοηθά στον ομαλό υπολογισμό αποστάσεων και στη γρηγορότερη σύγκλιση των αλγορίθμων (ειδικά στο Neural Network).

## 4.3 Δημιουργία και Εκπαίδευση Μοντέλων

Θα επιλέξουμε τους αλγορίθμους Support Vector Machine (SVM) και Multi-Layer Perceptron (MLP). Το SVM είναι ένας αλγόριθμος ταξινόμησης που προσπαθεί να βρει το βέλτιστο υπερεπίπεδο (hyperplane) για να διαχωρίσει τις κλάσεις με το μεγαλύτερο περιθώριο (margin). Το MLP είναι ένα feed-forward νευρωνικό δίκτυο, αποτελούμενο από διαδοχικά κρυφά επίπεδα (hidden layers).

### 4.3.1 Support Vector Machine (SVM)

Ορίζουμε έναν ταξινομητή SVM με υπολογισμό πιθανοτήτων (probability=True) ώστε να μπορούμε να σχεδιάσουμε καμπύλες ROC/PR.

Προσθέτουμε class\_weight="balanced" ώστε να λαμβάνεται υπόψη τυχόν ανισορροπία στις κλάσεις.

Εκπαιδεύουμε το μοντέλο με τα κλιμακωμένα δεδομένα εκπαίδευσης.

Αποθηκεύουμε τις προβλέψεις ταξινόμησης (y\_pred\_svm) και τις πιθανότητες (svm\_probs).

### 4.3.2 Neural Network (Multi-Layer Perceptron)

Ορίζουμε ένα νευρωνικό δίκτυο με 2 κρυφά επίπεδα των 10 νευρώνων (hidden\_layer\_sizes=(10,10,)).

Το εκπαιδεύουμε ως ταξινομητή (MLPClassifier) για έως 500 επαναλήψεις.

Λαμβάνουμε τόσο τις «σκληρές» προβλέψεις (y\_pred\_nn = 0/1) όσο και τις προβλεπόμενες πιθανότητες (nn\_probs).

Ένα νευρωνικό δίκτυο μπορεί να μάθει μη γραμμικά μοτίβα και συχνά αποδίδει ικανοποιητικά σε δεδομένα με αρκετές μεταβλητές.

## 4.4 Αξιολόγηση Μεθόδων

### 4.4.1 Αξιολόγηση SVM, MLP

Accuracy & Classification Report: Υπολογίζουμε ακρίβεια (accuracy) και δείχνουμε λεπτομερώς precision, recall, F1-score για κάθε κλάση.

**Η ακρίβεια (accuracy)** δίνει μια συνολική εντύπωση,

**Precision:** Ποσοστό από τις προβλέψεις “θετικό” (1) που είναι πραγματικά σωστές (TP / (TP+FP)).

**Recall:** Ποσοστό από τα πραγματικά θετικά που αναγνωρίζονται (TP / (TP+FN)).

**F1:** Αρμονικός μέσος precision, recall.

**Confusion Matrix**:Προβάλλουμε σε ένα heatmap τον πίνακα σύγχυσης (TP, TN, FP, FN).

**ROC Curve & AUC:** Υπολογίζουμε τις τιμές FPR (False Positive Rate) και TPR (True Positive Rate) για διάφορα κατώφλια (thresholds).

Το εμβαδόν κάτω από την καμπύλη (AUC) δείχνει πόσο καλά διαχωρίζει το μοντέλο τις δύο κλάσεις ανεξαρτήτως threshold.

**Precision-Recall Curve & AUC**: Σχηματίζουμε την καμπύλη Precision-Recall και υπολογίζουμε το εμβαδόν (PR AUC).

Είναι ιδιαίτερα χρήσιμη όταν η θετική κλάση (δημοφιλής ταινία) είναι πιο σπάνια.

Μας δείχνει πώς μεταβάλλεται η precision (ακρίβεια εντοπισμού) σε σχέση με τη recall (τι ποσοστό των «δημοφιλών» εντοπίζουμε).

### 4.4.2 Σύγκριση SVM, MLP

Επισκόπηση Αποτελεσμάτων:

**SVM**

Accuracy: 0.766

Class 0 (μη δημοφιλής):

Precision: 0.97, Recall: 0.69, F1: 0.80

Από όλες τις ταινίες που προβλέπονται ως «μη δημοφιλείς», το 97% είναι πραγματικά μη δημοφιλείς (υψηλή precision). Ωστόσο, το μοντέλο εντοπίζει μόνο το 69% των πραγματικών 0 (medium recall).

Class 1 (δημοφιλής):

Precision: 0.57, Recall: 0.95, F1: 0.71

Το μοντέλο είναι πολύ επιθετικό στο να βαφτίζει μια ταινία «δημοφιλή» (high recall 0.95). Όμως η precision είναι μόλις 0.57, πράγμα που σημαίνει αρκετά false positives (πολλές ταινίες που δεν είναι δημοφιλείς προβλέπονται ως «1»).

Συνολικά για το SVM, βλέπουμε:

Εξαιρετικά υψηλό precision για την κλάση 0 και παράλληλα πολύ υψηλό recall για την κλάση 1. Άρα το μοντέλο θυσιάζει την ακρίβεια (precision) στην κλάση 1 προκειμένου να μην χάνει θετικές περιπτώσεις (1).

**Neural Network (MLP)**

Accuracy: 0.786

Class 0 (μη δημοφιλής):

Precision: 0.87, Recall: 0.81, F1: 0.84

Το δίκτυο έχει πιο ισορροπημένες τιμές precision–recall για την κλάση 0: Ούτε υπερβολικά επιθετικό, ούτε υπερβολικά συντηρητικό.

Class 1 (δημοφιλής):

Precision: 0.63, Recall: 0.73, F1: 0.67

Πρόκειται για μια μέτρια αρμονία μεταξύ precision και recall. Προβλέπει με αρκετή επιτυχία τις δημοφιλείς ταινίες, αλλά δεν είναι όσο «επιθετικό» όσο το SVM στον εντοπισμό τους.

Συνολικά για το MLP, βλέπουμε:

Ίσως πιο ισορροπημένη συμπεριφορά: δεν θυσιάζει ιδιαίτερα το precision της κλάσης 1 (0.63 > 0.57), αλλά δεν φτάνει σε τόσο υψηλό recall (0.73 < 0.95).

Σύγκριση & Ερμηνεία

Ακρίβεια (Accuracy):

MLP ~ 0.786, SVM ~ 0.766

Το MLP υπερέχει ελαφρώς σε απλό Accuracy.

Συμπεριφορά στις Κλάσεις:

SVM:

Πολύ υψηλό recall στη θετική κλάση (1), σημαίνει ότι σπάνια χάνει μια πραγματικά δημοφιλή ταινία (μόλις 5% FN).

Όμως, η precision για κλάση 1 είναι 0.57 ⇒ αρκετά FP. Ουσιαστικά, λέει «είναι δημοφιλής» και για αρκετές που δεν είναι, γι’ αυτό και πέφτει η ακρίβεια εκεί.

MLP:

Πιο «συμμετρικό» μοντέλο: δεν χάνει πάρα πολλές δημοφιλείς, αλλά ούτε κάνει τόσες πολλές λάθος προβλέψεις «1».

Recall(1)=0.73 ⇒ χάνει το 27% των δημοφιλών· έχει όμως καλύτερη precision(1)=0.63 σε σχέση με το SVM.

Τελικό Συμπέρασμα:

Το MLP υπερέχει ελαφρώς σε overall accuracy (0.786 vs 0.766) και διατηρεί πιο ισορροπημένη συμπεριφορά ανάμεσα σε precision/recall στην κλάση 1.

Το SVM κυνηγάει υψηλό recall στην κλάση 1 (0.95!), αλλά θυσιάζει την precision της κλάσης 1. Αυτό οδηγεί σε αρκετά false positives.

Αποφασίστε ποια είναι η προτεραιότητα: να μην χάνετε δημοφιλείς ταινίες (προτιμήστε SVM ή ρυθμίστε threshold στο MLP για υψηλό recall) ή να έχετε λιγότερες λανθασμένες προβλέψεις δημοφιλίας (προτιμήστε MLP ή ρυθμίστε threshold για καλύτερη precision).

A blue squares with white text

AI-generated content may be incorrect.

Εικόνα 12: Confusion Matrix SVM

A blue squares with white text

AI-generated content may be incorrect.

Εικόνα 13: Confusion Matrix MLP

# 5. Συμπεράσματα

Πάμε να δούμε τις εφαρμογές του μοντέλου μας.

1. Εφαρμογή στον Κύκλο Ζωής μιας Ταινίας

Στάδιο Παραγωγής

Πριν καν ξεκινήσει η παραγωγή ή ολοκληρωθεί το σενάριο, οι παραγωγοί μπορούν να «τεστάρουν» την ιδέα τους στο μοντέλο.

Αν ο αλγόριθμος (βασισμένος σε παρόμοιες ταινίες και χαρακτηριστικά) δείξει πιθανότητα υψηλής δημοφιλίας, αυτό ενθαρρύνει επιπλέον επενδύσεις σε σκηνικά, εφέ, προώθηση κ.λπ.

Στάδιο Προώθησης

Λίγο πριν βγει η ταινία στο κοινό, το μοντέλο δίνει δείκτες για το πόσο «εμπορική» θα είναι.

Η εταιρεία μπορεί να αποφασίσει πόσο budget θα διαθέσει για διαφήμιση, σε ποιες χώρες να στοχεύσει, ποια κανάλια (TV, social media) να εκμεταλλευτεί περισσότερο.

Μετά την Κυκλοφορία

Το σύστημα συνεχίζει να συλλέγει feedback (βαθμολογίες, κριτικές) και ενημερώνει ξανά το μοντέλο.

Έτσι, γίνεται συνεχής βελτίωση (re-training) του μοντέλου, ανάλογα με την αντίδραση του κοινού σε πραγματικό χρόνο.

Επιπλέον, αυτό προσφέρει feedback loops για επόμενες παραγωγές.

2. Πρακτική Χρησιμότητα για Διάφορους Παίκτες

Streaming Platforms

Το recommendation system βελτιώνεται αν ξέρει εκ των προτέρων ποιες ταινίες έχουν καλές πιθανότητες να «τραβήξουν» χρήστες.

Προτείνει πιο «ποιοτικό» περιεχόμενο, διατηρώντας ικανοποιημένους συνδρομητές.

Κινηματογραφικές Εταιρείες

Η προώθηση μιας ταινίας συνήθως κοστίζει πολλά χρήματα (μερικές φορές συγκρίσιμα με τον προϋπολογισμό γυρισμάτων).

Αν η προβλεπόμενη απήχηση είναι μικρή, ίσως στρέψουν τους πόρους τους σε άλλο project.

3. Συνολική Αξία του «Μοντέλου Πρόβλεψης Επιτυχίας»

Αποφυγή Ρίσκου: Βοηθά εταιρείες να επενδύουν χρήματα όπου υπάρχει υψηλότερη πιθανότητα επιτυχίας.

Στοχευμένες Καμπάνιες: Μειώνει κόστος διαφήμισης, αυξάνει την αποτελεσματικότητα των πωλήσεων.

Καινοτομία: Ενσωματώνει τεχνικές Machine Learning & Data Analytics σε παραδοσιακά «δημιουργικούς» κλάδους όπως ο κινηματογράφος.

Έτσι, ολόκληρη η αλυσίδα αξίας (από τη σύλληψη μιας ιδέας μέχρι τη διανομή της σε αίθουσες και streaming) γίνεται πιο δεδομενο-κεντρική (data-driven) και αποδοτική.