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 3](#_Toc192854728)

[1.1 Περιγραφή του Συνόλου Δεδομένων 3](#_Toc192854729)

[2. Προεπεξεργασία Δεδομένων (Data Preprocessing) 4](#_Toc192854730)

[2.1 Φόρτωση Δεδομένων 4](#_Toc192854731)

[2.2 Εξερεύνηση και Επεξεργασία Δεδομένων 4](#_Toc192854732)

[2.3 Μετασχηματισμοί 4](#_Toc192854733)

[2.4 Ανάλυση Δεδομένων 5](#_Toc192854734)

[3. Συσταδοποίηση (Clustering) 9](#_Toc192854735)

[3.1 Σκοπός & Σχεδιασμός 9](#_Toc192854736)

[3.2 Επιλογή χαρακτηριστικών και επεξεργασία 9](#_Toc192854737)

[3.3 K-Means 10](#_Toc192854738)

[3.3.1 Υπολογισμός μετρικών για πιθανά Κ 10](#_Toc192854739)

[3.3.2 Διαγράμματα 11](#_Toc192854740)

[3.4 DBSCAN 12](#_Toc192854741)

[3.5 Ερμηνεία και Σύγκριση Μεθόδων 13](#_Toc192854742)

[4. Ταξινόμηση (Classification) 17](#_Toc192854743)

[4.1 Ορισμός Στόχου (Label) 17](#_Toc192854744)

[4.2 Προετοιμασία συνόλου δεδομένων 18](#_Toc192854745)

[4.3 Δημιουργία και Εκπαίδευση Μοντέλων 18](#_Toc192854746)

[4.3.1 Support Vector Machine (SVM) 19](#_Toc192854747)

[4.3.2 Neural Network (Multi-Layer Perceptron) 19](#_Toc192854748)

[4.4 Αξιολόγηση Μεθόδων 20](#_Toc192854749)

[4.4.1 Αξιολόγηση SVM, MLP 20](#_Toc192854750)

[4.4.2 Σύγκριση SVM, MLP 21](#_Toc192854751)

[5. Συμπεράσματα 23](#_Toc192854752)

# Εισαγωγή

Η βιομηχανία του κινηματογράφου και του streaming αξιοποιεί ολοένα και περισσότερες τεχνικές Αναλυτικής Δεδομένων και Μηχανικής Μάθησης στη λήψη αποφάσεων.

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.

Ένα φυσικό ερώτημα που προκύπτει είναι: μπορεί ένα σύστημα Μηχανικής Μάθησης να προβλέψει αν μια νέα ταινία θα γίνει δημοφιλής προτού καν κυκλοφορήσει; Με τη διαθεσιμότητα μεγάλου όγκου δεδομένων (βαθμολογίες χρηστών, είδη, ημερομηνίες κυκλοφορίας, μεταδεδομένα παραγωγής κ.ά.), μπορούμε να επιχειρήσουμε την ανάπτυξη μοντέλων που προσεγγίζουν μια τέτοια πρόβλεψη.

Στην παρούσα εργασία χρησιμοποιείται το MovieLens 100K dataset (περίπου 100.000 βαθμολογίες ταινιών) για την διερεύνηση του παραπάνω ερωτήματος.

Η προσέγγισή δεν είναι μόνο ακαδημαϊκή, καθώς έχει και πρακτικές εφαρμογές σε κινηματογραφικά στούντιο και πλατφόρμες streaming (π.χ. Netflix) που θα ήθελαν να γνωρίζουν εκ των προτέρων ποιες ταινίες αξίζει να χρηματοδοτήσουν ή να προωθήσουν. Μελετώντας τα μοτίβα (clusters) στα δεδομένα και κατασκευάζοντας ένα μοντέλο πρόβλεψης “δημοφιλίας”, αναδεικνύεται πώς η Αναλυτική Δεδομένων μπορεί να υποστηρίξει τη λήψη αποφάσεων στην κινηματογραφική βιομηχανία – από το marketing μιας ταινίας μέχρι τον σχεδιασμό ενός συστήματος συστάσεων.

Οι επόμενες ενότητες καλύπτουν διαδοχικά την προεπεξεργασία και εξερεύνηση των δεδομένων, την εφαρμογή μεθόδων συσταδοποίησης (K-Means, DBSCAN) για την ανάδειξη ομάδων ταινιών, την ανάπτυξη και αξιολόγηση μοντέλων ταξινόμησης (SVM, MLP) για την πρόβλεψη της δημοφιλίας και, τέλος, τα συμπεράσματα όπου συζητώνται η αποτελεσματικότητα των προσεγγίσεων και οι πρακτικές τους προεκτάσεις.

## Περιγραφή του Συνόλου Δεδομένων

Το MovieLens 100K dataset προέρχεται από το GroupLens Research του University of Minnesota και αποτελεί ένα από τα πλέον χρησιμοποιούμενα σύνολα δεδομένων σε ακαδημαϊκές μελέτες για συστήματα συστάσεων. Περιλαμβάνει τέσσερα αρχεία CSV:

* ratings.csv – βαθμολογίες ταινιών (στήλες: userId, movieId, rating, timestamp)
* movies.csv – μεταδεδομένα ταινιών (στήλες: movieId, title, genres)
* tags.csv – ετικέτες που απέδωσαν οι χρήστες σε ταινίες (στήλες: userId, movieId, tag, timestamp)
* links.csv – σύνδεσμοι των ταινιών σε εξωτερικές βάσεις (στήλες: movieId, imdbId, tmdbId)

Βασικά στατιστικά του dataset:

* Συνολικές βαθμολογίες (εγγραφές): ~100.000
* Αριθμός χρηστών: 943
* Αριθμός ταινιών: 1.682

# Προεπεξεργασία Δεδομένων (Data Preprocessing)

## 2.1 Φόρτωση Δεδομένων

Τα τέσσερα αρχεία CSV του dataset φορτώθηκαν σε ξεχωριστά pandas Data Frames (ratings, movies, tags, links). Η διάθεση των δεδομένων σε πολλαπλά αρχεία απαιτεί την αρχική φόρτωσή τους προτού μπορέσουμε να τα συνδυάσουμε και να τα αναλύσουμε.

## 2.2 Εξερεύνηση και Επεξεργασία Δεδομένων

Πριν την ανάπτυξη μοντέλων, πραγματοποιήθηκαν βασικά βήματα καθαρισμού και προετοιμασίας:

* 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 (Outliers): Οι βαθμολογίες των ταινιών κυμαίνονται μόνο μεταξύ 1 και 5, 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 το title (τίτλος ταινίας). Τα πεδία αυτά δεν συνεισφέρουν στην πρόβλεψη ή στη συσταδοποίηση, επομένως αφαιρώντας τα απλοποιείται το dataset χωρίς απώλεια πληροφορίας σχετικής με το πρόβλημα.

## 2.3 Μετασχηματισμοί

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

* One-Hot Encoding (Κατηγοριοποίηση Ειδών): Το πεδίο των ειδών ταινίας (genres) που περιέχει συμβολοσειρές με πολλαπλά είδη (π.χ. "Action|Comedy|Drama") μετασχηματίστηκε σε διακριτά δυαδικά χαρακτηριστικά. Για κάθε ταινία, διαχωρίστηκαν τα είδη σε μεμονωμένες εγγραφές (διαδικασία *explode*) 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 movieId και συνενώθηκαν, ώστε κάθε ταινία να αντιπροσωπεύεται από μία γραμμή με πολλαπλά χαρακτηριστικά (στήλες) που υποδεικνύουν την παρουσία ή απουσία κάθε genre.
* Κανονικοποίηση Βαθμολογιών: Εφαρμόστηκε κλιμάκωση StandardScaler στις βαθμολογίες των ταινιών. Μετά τον μετασχηματισμό αυτό, οι βαθμολογίες έχουν μέσο όρο 0 και τυπική απόκλιση 1. Η κανονικοποίηση εξισορροπεί την κλίμακα των βαθμολογιών, έτσι ώστε καμία ταινία να μη “υπερτερεί” ή “υστερεί” λόγω μεγέθους τιμών, και διευκολύνει αλγορίθμους που βασίζονται σε αποστάσεις (π.χ. K-Means, SVM) να λειτουργούν πιο αξιόπιστα.
* Ενοποίηση Δεδομένων: Τα επεξεργασμένα δεδομένα βαθμολογιών (κανονικοποιημένες) και τα χαρακτηριστικά είδους (one-hot encoded genres) συνενώθηκαν σε ένα ενιαίο dataset. Κάθε ταινία πλέον αναπαρίσταται από ένα σύνολο αριθμητικών χαρακτηριστικών: τη normalized βαθμολογία της και πληθώρα δεικτών (0/1) για την κατηγορία/είδος της. Αυτό το τελικό DataFrame των χαρακτηριστικών θα χρησιμοποιηθεί τόσο στη συσταδοποίηση όσο και στην ταξινόμηση.

## 2.4 Ανάλυση Δεδομένων

Με το dataset καθαρισμένο και μετασχηματισμένο, ξεκινά η διερευνητική ανάλυση μέσω γραφημάτων για να κατανοήσουμε καλύτερα τις κατανομές και τυχόν συσχετίσεις των δεδομένων:

* Κατανομή βαθμολογιών (Histogram με KDE):
  + 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.
  + 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.

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

Διάγραμμα Box Plot (κανονικοποιημένες βαθμολογίες):

* Το μεγαλύτερο μέρος των κανονικοποιημένων βαθμολογιών βρίσκεται στο εύρος περίπου [-1, +1].
* Η διάμεσος (median) των βαθμολογιών είναι ελαφρώς πάνω από το 0. Αυτό υποδεικνύει ότι, συνολικά, οι ταινίες τείνουν να λαμβάνουν λίγο υψηλότερες αξιολογήσεις από τον «ουδέτερο» μέσο όρο.
* Παρατηρούνται ορισμένες ακραίες χαμηλές τιμές (outliers) γύρω στο -3. Αυτές αντιστοιχούν σε λίγες ταινίες που πήραν πολύ χαμηλές βαθμολογίες συγκριτικά με τις υπόλοιπες – δηλαδή ταινίες που το κοινό αξιολόγησε εξαιρετικά αρνητικά.

A blue rectangular box with black lines

AI-generated content may be incorrect.

Εικόνα 2: Box Plot of normalized ratings

* Κατανομή Ειδών Ταινίας (Bar Plot ανά Genre):
  + Το Drama είναι το είδος με τις περισσότερες ταινίες στο dataset (περίπου 4.500 εμφανίσεις ειδών σε ταινίες αν συμπεριλάβουμε τις πολλαπλές κατηγορίες).
  + Ακολουθεί η 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.

*A graph of a number of movies

AI-generated content may be incorrect.Εικόνα 3: Distribution of movie genres*

* Χάρτης Συνεργασίας Χαρακτηριστικών (Correlation Heatmap):
  + Δεν παρατηρείται κάποια ισχυρή συσχέτιση μεταξύ της βαθμολογίας μιας ταινίας και οποιουδήποτε συγκεκριμένου είδους. Δηλαδή, το αν μια ταινία ανήκει σε κάποιο genre δεν φαίνεται να σχετίζεται *άμεσα* με το ύψος της βαθμολογίας της.
  + 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–Adventure ή Musical–Comedy. Οι συσχετίσεις αυτές όμως δεν είναι εξαιρετικά υψηλές – απλώς δείχνουν ότι υπάρχουν ορισμένοι συνδυασμοί ειδών που είναι δημοφιλείς.
  + Η ισχυρότερη θετική συσχέτιση μεταξύ δύο συγκεκριμένων κατηγοριών είναι περίπου 0.49, ανάμεσα σε Animation και Children. Αυτό είναι αναμενόμενο, καθώς πολλές ταινίες κινουμένων σχεδίων απευθύνονται σε παιδικό/οικογενειακό κοινό.
  + Γενικά, η απουσία πολύ υψηλών συσχετίσεων (κοντά στο 1) υποδηλώνει ότι το dataset περιλαμβάνει μεγάλη ποικιλία ταινιών. Δεν υπάρχει δηλαδή μία μικρή ομάδα ειδών που να εμφανίζονται πάντα μαζί· τα είδη διαμοιράζονται σε πολλές διαφορετικές συνδυαστικές περιπτώσεις.

A graph with numbers and a red line

AI-generated content may be incorrect.Με βάση την παραπάνω εξερεύνηση, αποκτήσαμε μια βασική κατανόηση της δομής των δεδομένων – π.χ. ποια είδη είναι πιο κοινά, πώς διανέμονται οι αξιολογήσεις – γεγονός που μας βοηθάει να προσεγγίσουμε τα επόμενα στάδια ανάλυσης με καλύτερη εικόνα. Στη συνέχεια, περνάμε στη φάση της συσταδοποίησης (clustering), για να ανακαλύψουμε κρυμμένα μοτίβα ομάδων ταινιών.

Εικόνα 4: Feature Correlation Heatmap

# 3. Συσταδοποίηση (Clustering)

## 3.1 Σκοπός & Σχεδιασμός

Αφού ολοκληρώθηκε η προεπεξεργασία των δεδομένων και αποκτήσαμε εικόνα για τη διανομή τους (βαθμολογίες, δημοφιλή είδη κ.λπ.), το επόμενο βήμα είναι η συσταδοποίηση (clustering). Στόχος είναι να ανακαλύψουμε μοτίβα ή ομάδες ταινιών που δεν είναι άμεσα ορατές μέσω απλής στατιστικής ανάλυσης. Συνοπτικά, ο σχεδιασμός για το clustering έχει ως εξής:

* Επιλογή Χαρακτηριστικών: Ως χαρακτηριστικά για το clustering χρησιμοποιούμε τις κανονικοποιημένες βαθμολογίες (rating) των ταινιών και τις δυαδικές μεταβλητές των genres (one-hot encoded genres). Αγνοούμε πεδία όπως τα IDs (movieId, userId), καθώς είναι απλώς αναγνωριστικά και δεν περιέχουν πληροφορία περιεχομένου.
* Αλγόριθμοι: Θα εφαρμοστούν δύο διαφορετικοί αλγόριθμοι clustering – ο κλασικός K-Means και ο DBSCAN. Οι δύο αυτοί αλγόριθμοι 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). Η σύγκριση τους θα μας δώσει δύο οπτικές για τα δεδομένα.
* Αξιολόγηση Συστάδων: Για να εκτιμήσουμε την ποιότητα των clusters που θα παραχθούν, θα υπολογίσουμε διάφορες μετρικές αξιολόγησης συσταδοποίησης: Inertia, Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index. Αυτές οι μετρικές θα μας επιτρέψουν να συγκρίνουμε ποσοτικά τα αποτελέσματα του K-Means και του DBSCAN (και διαφορετικές παραμετροποιήσεις τους) ώστε να επιλέξουμε την καλύτερη ομαδοποίηση των ταινιών.

## 3.2 Επιλογή χαρακτηριστικών και επεξεργασία

Για να διασφαλίσουμε ότι οι αλγόριθμοι clustering θα λειτουργήσουν αποδοτικά, εφαρμόσαμε κάποια επιπλέον βήματα προετοιμασίας στα χαρακτηριστικά:

* Αφαιρέσαμε από το σύνολο χαρακτηριστικών τις στήλες αναγνωριστικών (movieId, userId), διότι δεν περιέχουν πληροφορία σχετική με το περιεχόμενο ή την ομαδοποίηση.
* Εφαρμόσαμε εκ νέου StandardScaler σε όλα τα αριθμητικά χαρακτηριστικά (περιλαμβάνει πλέον την normalized βαθμολογία και τις στήλες genres 0/1) ώστε να έχουν μέσο όρο 0 και τυπική απόκλιση 1. Αυτό είναι σημαντικό για τον K-Means, ο οποίος στηρίζεται σε αποστάσεις – η κλιμάκωση διασφαλίζει ότι καμία μεταβλητή δεν κυριαρχεί λόγω μονάδων μέτρησης.
* Μειώσαμε τη 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 (ειδικά των genres) 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 φιλτράρουμε επίσης πιθανό θόρυβο στα δεδομένα.
* Τα παραπάνω βήματα βελτιώνουν την αποδοτικότητα του clustering: αφενός επιταχύνουν τους υπολογισμούς (λιγότερες διαστάσεις προς επεξεργασία) και αφετέρου μπορούν να οδηγήσουν σε πιο «καθαρές» συστάδες.

## 3.3 K-Means

### 3.3.1 Υπολογισμός μετρικών για πιθανά Κ

Για τον αλγόριθμο K-Means, χρειάζεται να ορίσουμε τον αριθμό των clusters *K*. Αυτό είναι ένα κρίσιμο υπερ-παράμετρο, καθώς μπορεί να αλλάξει δραστικά την ομαδοποίηση. Προκειμένου να επιλέξουμε το βέλτιστο K, ακολουθήσαμε την εξής διαδικασία:

* Εκπαιδεύσαμε διαδοχικά μοντέλα K-Means για κάθε πιθανό αριθμό clusters από K = 2 έως Κ = 10 χρησιμοποιώντας το σύνολο χαρακτηριστικών που προέκυψε μετά το PCA.
* Για κάθε τιμή του K, υπολογίσαμε και καταγράψαμε διάφορες μετρικές ποιότητας clustering:
  + Inertia – άθροισμα των αποστάσεων όλων των σημείων από το πλησιέστερο κέντρο cluster (ενδο-ομαδική απόσταση). Χαμηλότερη τιμή υποδηλώνει πιο συμπαγείς συστάδες.
  + Silhouette Score – μέσο συντελεστή σιλουέτας για όλα τα σημεία, ο οποίος κυμαίνεται από -1 έως 1. Υψηλότερες τιμές υποδεικνύουν καλύτερο διαχωρισμό μεταξύ των clusters (οι αποστάσεις των σημείων από το cluster τους είναι πολύ μικρότερες από την απόσταση προς άλλα clusters).
  + Δείκτης Davies-Bouldin – μετράει την “διάχυση” των clusters λαμβάνοντας υπόψη την απόσταση κάθε cluster από τα γειτονικά του (χαμηλότερες τιμές σημαίνουν πιο διακριτά και συμπαγή clusters).
  + Δείκτης Calinski-Harabasz – λόγος της διασποράς μεταξύ των clusters προς τη διασπορά εντός των clusters (υψηλότερες τιμές σημαίνουν ότι τα clusters είναι καλά διαχωρισμένα και τα σημεία κάθε cluster βρίσκονται κοντά στο κέντρο του).
* Επιπλέον, για κάθε K εξετάσαμε το μέγεθος των clusters (πόσες ταινίες εμπίπτουν σε κάθε cluster). Καταγράψαμε την κατανομή μεγέθους συστάδων και τη σχεδιάσαμε σε γράφημα μπάρας, ώστε να ελέγξουμε αν για κάποια K προκύπτουν έντονα άνισες ομάδες (π.χ. ένα cluster με υπερβολικά πολλά ή ελάχιστα σημεία).
* Τα αποτελέσματα των παραπάνω αναλύσεων απεικονίστηκαν σε γραφήματα (π.χ. διάγραμμα Inertia vs K, διάγραμμα Silhouette vs K, κ.ά.) για να διευκολύνουν την οπτική εντοπίσει του βέλτιστου K.

### 3.3.2 Διαγράμματα

A graph with a line

AI-generated content may be incorrect.

Εικόνα 5: Inertia plot

A graph with red dots

AI-generated content may be incorrect.

Εικόνα 6 : Silhouette score for different K values

A diagram of a cluster of dots

AI-generated content may be incorrect.Συγκρίνοντας όλες τις μετρικές ποιότητας για K=2 έως 10, παρατηρήσαμε σύγκλιση ενδείξεων υπέρ ενός συγκεκριμένου αριθμού clusters. Συγκεκριμένα, διακρίναμε (elbow) στο διάγραμμα της Inertia και υψηλή τιμή Silhouette score γύρω από το K=3. Με βάση τον συνδυασμό των μετρικών και των διαγραμμάτων, επιλέξαμε ως βέλτιστο αριθμό συστάδων το K=3 για τον αλγόριθμο K-Means. Αυτό σημαίνει ότι οι ταινίες ομαδοποιούνται καλύτερα σε τρεις κύριες κατηγορίες σύμφωνα με τα χαρακτηριστικά μας.

Εικόνα 7: K-means clusters

## 3.4 DBSCAN

Σε αντίθεση με τον K-Means, ο DBSCAN δεν απαιτεί apriori ορισμό του αριθμού clusters, αλλά στηρίζεται σε δύο άλλες παραμέτρους: την ακτίνα γειτονιάς (επιτρεπόμενη απόσταση) που ορίζει πότε δύο σημεία θεωρούνται κοντά (*εps*) και τον ελάχιστο αριθμό σημείων που απαιτούνται για να σχηματίσουν μια συστάδα (*min\_samples*). Η εύρεση των κατάλληλων παραμέτρων έγινε ως εξής:

* Πραγματοποιήσαμε πειραματισμό δοκιμάζοντας πολλούς συνδυασμούς τιμών για τις παραμέτρους εps και min\_samples.
* Για κάθε συνδυασμό παραμέτρων, εφαρμόσαμε τον DBSCAN στο dataset. Εάν ο αλγόριθμος κατάφερνε να αναγνωρίσει τουλάχιστον 2 clusters (δηλαδή αν δεν χαρακτήριζε όλα τα σημεία ως θόρυβο ή όλα σε ένα cluster), τότε για εκείνο τον συνδυασμό υπολογίζαμε τις μετρικές ποιότητας των αποτελεσμάτων: Silhouette, Davies-Bouldin, Calinski-Harabasz. Παράλληλα, σημειώναμε και το ποσοστό των σημείων που θεωρήθηκαν θόρυβος (ετικέτα cluster = -1), καθώς ένα πολύ μεγάλο ποσοστό θορύβου σημαίνει ότι ο DBSCAN απορρίπτει πολλά δεδομένα ως outliers.
* Καθώς δοκιμάζαμε διαφορετικές τιμές, παρακολουθούσαμε ποιος συνδυασμός (εps, min\_samples) απέδιδε τον υψηλότερο Silhouette score – χρησιμοποιήσαμε το Silhouette ως κύριο κριτήριο για βέλτιστη συσταδοποίηση, μιας και συγκεντρώνει πληροφορία τόσο για την ενδο-συνοχή όσο και για τον μεταξύ τους διαχωρισμό των clusters.
* Τελικά, επιλέξαμε το ζεύγος παραμέτρων που μεγιστοποίησε το Silhouette ως το βέλτιστο Με αυτές τις τιμές εξισορροπήθηκε και το ποσοστό θορύβου σε αποδεκτά επίπεδα.
* Με το βέλτιστο μοντέλο DBSCAN, προχωρήσαμε σε οπτικοποίηση των αποτελεσμάτων. Επειδή τα clusters του DBSCAN μπορεί να είναι πολλά και δυσδιάστατα, χρησιμοποιήσαμε και πάλι PCA για να μειώσουμε το dataset σε 2 διαστάσεις και δημιουργήσαμε ένα scatter plot. Κάθε σημείο (ταινία) στο διάγραμμα χρωματίστηκε σύμφωνα με το cluster ID που του ανέθεσε ο DBSCAN, ενώ τα σημεία που χαρακτηρίστηκαν ως θόρυβος φέρουν ειδική ετικέτα (-1). Αυτό μας επέτρεψε να δούμε σχηματικά την κατανομή των ταινιών στους χώρους πυκνότητας που βρήκε ο αλγόριθμος.

A diagram of red and blue dots

AI-generated content may be incorrect.

Εικόνα 8:DBSCAN clusters

## 3.5 Ερμηνεία και Σύγκριση Μεθόδων

Η εφαρμογή του K-Means με K=3 παρήγαγε τρεις ομάδες ταινιών. Αναλύοντας τη σύσταση κάθε cluster ως προς τα genres και τη μέση βαθμολογία των ταινιών, μπορούμε να ερμηνεύσουμε τι χαρακτηρίζει καθεμία από αυτές:

|  |  |
| --- | --- |
| Cluster 0 | Περιλαμβάνει κυρίως ταινίες των οποίων η μέση κανονικοποιημένη βαθμολογία είναι γύρω στο 0 (δηλαδή πρόκειται για ταινίες με μέτριες προς ελαφρώς θετικές αξιολογήσεις). Στο cluster αυτό, η παρουσία ειδών δράσης ή φαντασίας είναι χαμηλή – τα ποσοστά των genres Action, Adventure και Animation είναι μικρά. Αντίθετα, πολλές από αυτές τις ταινίες ανήκουν πιθανώς σε είδη όπως το Drama ή/και το Comedy. Συνοπτικά, το Cluster 0 φαίνεται να αντιστοιχεί σε ταινίες χωρίς έντονα στοιχεία δράσης/περιπέτειας ή παιδικής φαντασίας, με μεσαίου επιπέδου αποδοχή από το κοινό. |
| Cluster 1 | Περιλαμβάνει ταινίες που κατά μέσο όρο έχουν ελαφρώς υψηλότερη βαθμολογία από το Cluster 0 (λίγο πάνω από το 0, π.χ. ~0.02 normalized). Ξεχωρίζει για το πολύ υψηλό ποσοστό ταινιών Action – περίπου οι μισές ταινίες σε αυτό το cluster είναι δράσης. Επίσης, έχει σημαντική παρουσία του είδους Adventure. Αντιθέτως, σχεδόν καμία ταινία εδώ δεν είναι Animation. Το Cluster 1, επομένως, μπορεί να χαρακτηριστεί ως το cluster των Action/Adventure ταινιών, οι οποίες μάλιστα τείνουν να έχουν σχετικά θετικές αξιολογήσεις (οι πιο “δημοφιλείς” ανάμεσα στα τρία clusters). |
| Cluster 2 | Περιλαμβάνει ταινίες με ελαφρώς χαμηλότερη του μέσου όρου βαθμολογία (αρνητική κανονικοποιημένη βαθμολογία – δηλαδή αυτές οι ταινίες γενικά δεν απέσπασαν πολύ καλές κριτικές από τους χρήστες). Το cluster αυτό έχει τη μεγαλύτερη παρουσία του είδους Adventure (κατά προσέγγιση στο ~55% των ταινιών του cluster) και αξιοσημείωτη παρουσία Animation (περίπου 35% των ταινιών είναι κινουμένων σχεδίων). Αντίθετα, οι ταινίες δράσης είναι λίγες. Το Cluster 2 φαίνεται να αντιστοιχεί σε ταινίες που είναι κυρίως περιπέτειες και κινούμενα σχέδια (πιθανώς οικογενειακές/παιδικές ταινίες) οι οποίες όμως δεν έτυχαν υψηλής αποδοχής – δηλαδή, παρά το φαντασμαγορικό είδος τους, οι αξιολογήσεις τους ήταν μέτριες προς χαμηλές. |

Γενικά, παρατηρούμε ότι τα clusters δεν διαχωρίζονται αποκλειστικά βάσει είδους, αλλά φαίνεται να σχετίζονται και με τη συνολική “δημοφιλία” ή αποδοχή των ταινιών (μέση βαθμολογία). Για παράδειγμα, το Cluster 1 συγκεντρώνει τις πιο καλά βαθμολογημένες ταινίες δράσης/περιπέτειας, ενώ το Cluster 2 συγκεντρώνει περιπέτειες/animation με σχετικά χαμηλότερες βαθμολογίες. Αυτό υποδηλώνει ότι στο dataset μας ορισμένα είδη ταινιών συνδυάζονται με συγκεκριμένα επίπεδα επιτυχίας.

Σύγκριση K-Means και DBSCAN: Η εφαρμογή των δύο μεθόδων clustering ανέδειξε διαφορετικές “οπτικές” των δεδομένων:

* Ο βέλτιστος K-Means (K=3) παρείχε λίγες, σαφείς ομάδες ταινιών. Μπορούμε να συνοψίσουμε τις τρεις αυτές ομάδες και να τις ερμηνεύσουμε (όπως έγινε παραπάνω), παρατηρώντας διαφορές στα genres και στη μέση βαθμολογία κάθε ομάδας.
* Ο βέλτιστος DBSCAN, αντίθετα, παρήγαγε έναν πολύ μεγαλύτερο αριθμό clusters. Στο διάγραμμα 2D PCA των αποτελεσμάτων του DBSCAN, εμφανίστηκαν εκατοντάδες διαφορετικές ετικέτες clusters (π.χ. cluster IDs 0, 1, 2, ..., 400, 800, …) κάτι που σημαίνει ότι ο DBSCAN διαμόρφωσε πολυάριθμες μικροσυστάδες.
* Οι πολλές μικρές συστάδες του DBSCAN δυσχεραίνουν την ερμηνεία των αποτελεσμάτων. Δεν είναι πρακτικό να εξαχθούν συμπεράσματα αν έχουμε δεκάδες ή εκατοντάδες ομάδες – δεν μπορούμε εύκολα να περιγράψουμε “τι είδους ταινίες” ανήκουν σε καθεμία από τόσο πολλές ομάδες. Αντίθετα, με τον K-Means και 3 clusters, μπορούμε να δώσουμε μια θεματική ερμηνεία σε κάθε ομάδα.
* Παρά την ενδεχομένως υψηλή τιμή Silhouette που πέτυχε ο DBSCAN (επειδή σχημάτισε πολύ “σφιχτές” μικρο-ομάδες σημείων), αυτό το αποτέλεσμα δεν μεταφράζεται σε χρήσιμα clusters για ανάλυση υψηλού επιπέδου. Ο DBSCAN ουσιαστικά εντόπισε πολλές μικρές περιοχές πυκνότητας, αλλά αυτές δεν αντιστοιχούν σε *γενικές κατηγορίες* ταινιών που θα μπορούσε να εκμεταλλευτεί κάποιος εύκολα.
* Συνοψίζοντας, στο συγκεκριμένο πρόβλημα ο K-Means αποδείχθηκε πιο κατάλληλος, καθώς προσφέρει μια συνοπτική και κατανοητή κατηγοριοποίηση των ταινιών σε λίγες ομάδες. Ο DBSCAN με τις τρέχουσες ρυθμίσεις απέδωσε πληθώρα clusters που δεν είναι εύκολο να αξιοποιηθούν πρακτικά. Επομένως, για την ομαδοποίηση των ταινιών του MovieLens dataset σε διακριτές κατηγορίες, προτιμούμε την προσέγγιση του K-Means έναντι του DBSCAN.

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

A graph of a bar chart

AI-generated content may be incorrect.Αυτό υποδεικνύει ότι τα clusters δεν διαχωρίζονται μόνο βάσει genre, αλλά και με βάση τη συνολική «δημοφιλία» ή προτίμηση των χρηστών.

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 with a bar

AI-generated content may be incorrect.A graph of a graph

AI-generated content may be incorrect.

Εικόνα 8: average presence of genre Adventure / Εικόνα 9: average presence of genre animation

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.

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.

# 4. Ταξινόμηση (Classification)

Αφού εξετάσαμε μια μη-εποπτευόμενη προσέγγιση (clustering) για την ανάλυση των δεδομένων, προχωράμε στην εποπτευόμενη μάθηση με στόχο την πρόβλεψη της “επιτυχίας” μιας ταινίας. Συγκεκριμένα, θα επιχειρήσουμε να εκπαιδεύσουμε μοντέλα ταξινόμησης που να προβλέπουν αν μια ταινία είναι δημοφιλής ή όχι με βάση τα χαρακτηριστικά της. Αυτό απαιτεί πρώτα να ορίσουμε ποια ταινία θεωρούμε “δημοφιλή” και στη συνέχεια να εκπαιδεύσουμε αλγορίθμους πάνω σε δεδομένα με γνωστές τέτοιες ετικέτες.

## 4.1 Ορισμός Στόχου (Label)

Για να προσεγγίσουμε το ζήτημα της πρόβλεψης δημοφιλίας, χρειάζεται να μετατρέψουμε το πρόβλημα σε ένα πρόβλημα δυαδικής ταξινόμησης. Ορίσαμε λοιπόν μια δυαδική ετικέτα στόχο, την οποία θα προβλέπουν τα μοντέλα μας:

* is\_popular: Μια ταινία χαρακτηρίζεται ως “δημοφιλής” (1) αν ο αριθμός αξιολογήσεων που έχει λάβει είναι πάνω από ένα ορισμένο όριο, αλλιώς χαρακτηρίζεται “μη δημοφιλής” (0).
* Κριτήριο: Χρησιμοποιήσαμε ως όριο τη διάμεσο του πλήθους αξιολογήσεων όλων των ταινιών. Δηλαδή, υπολογίσαμε πόσες βαθμολογίες έχει η “μεσαία” ταινία στο σύνολο και θεωρήσαμε δημοφιλείς όλες όσες έχουν περισσότερες αξιολογήσεις από αυτό το median. Αν μια ταινία έχει αριθμό αξιολογήσεων κάτω από τη διάμεσο, χαρακτηρίζεται ως μη δημοφιλής.

Με αυτόν τον τρόπο, περίπου το μισό των ταινιών λαμβάνει ετικέτα 1 (δημοφιλείς) και το άλλο μισό 0 (μη δημοφιλείς), πράγμα που καθιστά το dataset ισορροπημένο ως προς τις κλάσεις και το κριτήριο αρκετά αυστηρό αλλά και αντικειμενικό. (Εναλλακτικά κριτήρια θα μπορούσαν να είναι το μέσο rating ή ένας συνδυασμός rating και αριθμού ψήφων, αλλά επιλέξαμε το πλήθος αξιολογήσεων ως ένδειξη “δημοφιλίας” – π.χ. μια ταινία με πολλές βαθμολογίες θεωρούμε ότι συζητήθηκε/προβλήθηκε πολύ.)

## 4.2 Προετοιμασία συνόλου δεδομένων

Αφού ορίστηκε η ετικέτα is\_popular για κάθε ταινία, προχωρήσαμε στη δημιουργία του τελικού πίνακα χαρακτηριστικών για την ταξινόμηση και στον διαχωρισμό των δεδομένων σε training και test:

* Συγχωνεύσαμε τη στήλη is\_popular με τα ήδη επεξεργασμένα δεδομένα των ταινιών. Δημιουργήσαμε δηλαδή ένα νέο DataFrame το οποίο, για κάθε movieId, περιλαμβάνει όλα τα χαρακτηριστικά (π.χ. genres σε one-hot μορφή, normalized βαθμολογία) συν τη στήλη-στόχο is\_popular.
* Διαχωρίσαμε τα δεδομένα σε features (X) και label (y). Το X περιλαμβάνει όλες τις χαρακτηριστικές στήλες εκτός από το movieId και το is\_popular (το movieId δεν έχει νόημα ως χαρακτηριστικό, ενώ το is\_popular είναι αυτό που θέλουμε να προβλέψουμε). Το y είναι το αντίστοιχο διάνυσμα τιμών της μεταβλητής is\_popular (μήκους ίσου με τον αριθμό ταινιών).
* Έπειτα, χωρίσαμε το σύνολο των ταινιών σε δύο υποσύνολα: σύνολο εκπαίδευσης (training) και σύνολο δοκιμής (test). Χρησιμοποιήσαμε 80% των δεδομένων για εκπαίδευση και κρατήσαμε 20% ως test, ώστε να αξιολογήσουμε την απόδοση των μοντέλων σε “άγνωστα” δεδομένα. Επιπλέον, ο διαχωρισμός έγινε με stratify = y, δηλαδή διατηρήθηκε η ίδια αναλογία δημοφιλών/μη δημοφιλών ταινιών και στα δύο σύνολα (περίπου 50-50), για να μην υπάρξει μεροληψία.
* Στα χαρακτηριστικά του X εφαρμόστηκε εκ νέου StandardScaler μετά τον διαχωρισμό train-test. Συγκεκριμένα, προσαρμόσαμε τον scaler στα δεδομένα εκπαίδευσης και τον χρησιμοποιήσαμε για να κλιμακώσουμε τόσο τα train όσο και τα test features. Με αυτόν τον τρόπο, όλα τα χαρακτηριστικά (όπως οι 0/1 στήλες genres και η normalized βαθμολογία) βρίσκονται σε συγκρίσιμη κλίμακα. Η κλιμάκωση αυτή είναι σημαντική για αλγορίθμους όπως το SVM και τα νευρωνικά δίκτυα, καθώς βελτιώνει τη σταθερότητα των υπολογισμών και την ταχύτητα σύγκλισης του μοντέλου.

## 4.3 Δημιουργία και Εκπαίδευση Μοντέλων

Για την επίλυση του προβλήματος ταξινόμησης επιλέξαμε να εκπαιδεύσουμε δύο διαφορετικούς τύπους μοντέλων, ώστε να συγκρίνουμε τις επιδόσεις τους: έναν Support Vector Machine (SVM) και ένα νευρωνικό δίκτυο τύπου Multi-Layer Perceptron (MLP). Η επιλογή αυτών των αλγορίθμων δίνει δύο διαφορετικές προσεγγίσεις: το SVM αναζητά ένα γραμμικό (ή μη γραμμικό μέσω kernel) όριο απόφασης με μέγιστο περιθώριο μεταξύ των κλάσεων, ενώ το MLP είναι ένα πολυεπίπεδο νευρωνικό δίκτυο που μπορεί να μάθει πιο σύνθετες, μη γραμμικές σχέσεις στα δεδομένα.

### 4.3.1 Support Vector Machine (SVM)

Για το SVM, χρησιμοποιήσαμε την υλοποίηση από τη βιβλιοθήκη scikit-learn (SVC classifier) και ρυθμίσαμε τις εξής παραμέτρους/τεχνικές:

* Ενεργοποιήσαμε τον υπολογισμό πιθανοτήτων στο SVM (probability=True) έτσι ώστε μετά την εκπαίδευση να μπορούμε να εξάγουμε πιθανότητες πρόβλεψης για την κλάση “δημοφιλής”. Αυτό θα μας επιτρέψει να σχεδιάσουμε καμπύλες αξιολόγησης όπως ROC και Precision-Recall.
* Θέσαμε class\_weight="balanced" κατά την αρχικοποίηση του SVM. Με αυτόν τον τρόπο, το μοντέλο θα λαμβάνει υπόψη ότι οι κλάσεις 0 και 1 είναι (σχεδόν) ισοπληθείς, και αποφεύγεται τυχόν μεροληψία αν υπήρχε ανισορροπία. (Στο συγκεκριμένο dataset οι κλάσεις είναι ισορροπημένες λόγω του τρόπου που ορίσαμε το label, αλλά αυτό το βήμα είναι γενικά καλή πρακτική).
* Εκπαιδεύσαμε το SVM (fit) χρησιμοποιώντας τα δεδομένα εκπαίδευσης (μετά την κλιμάκωση).
* Μετά την εκπαίδευση, χρησιμοποιήσαμε το μοντέλο για να προβλέψουμε τις ετικέτες στο σύνολο δοκιμής. Αποθηκεύσαμε τόσο τις προβλεφθείσες κλάσεις (y\_pred\_svm, δηλαδή 0 ή 1 για κάθε ταινία του test set) όσο και τις προβλεφθείσες πιθανότητες (svm\_proba) που δίνει το μοντέλο για την κλάση 1 (δημοφιλής). Οι πιθανότητες αυτές θα χρησιμεύσουν για τις καμπύλες ROC/PR.

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

Για το νευρωνικό δίκτυο, χρησιμοποιήσαμε τον MLP Classifier της scikit-learn, με τις εξής ρυθμίσεις:

* Καθορίσαμε την αρχιτεκτονική του MLP με 2 κρυφά επίπεδα των 10 νευρώνων το καθένα (hidden\_layer\_sizes=(10, 10)). Ένα δίκτυο με 2 hidden layers μπορεί να μάθει αρκετά σύνθετες συναρτήσεις, ενώ το μέγεθος 10 νευρώνων θεωρήθηκε επαρκές για το μέγεθος του προβλήματος (δοκιμάστηκε και μεγαλύτερο χωρίς σημαντική διαφορά).
* Θέσαμε μέγιστο αριθμό *epochs* (iterations) = 100 για την εκπαίδευση. Αυτό δίνει στο δίκτυο αρκετές ευκαιρίες να συγκλίνει.
* Εκπαιδεύσαμε το MLP στα δεδομένα εκπαίδευσης. Κατά την εκπαίδευση, το δίκτυο προσαρμόζει τα βάρη του μέσα από backpropagation ώστε να ταξινομεί σωστά τις ταινίες ως δημοφιλείς ή όχι.
* Μετά την εκπαίδευση, εξαγάγαμε τις προβλέψεις κλάσης του δικτύου στο test set (y\_pred\_nn) καθώς και τις πιθανότητες πρόβλεψης για την κλάση 1 (nn\_proba). Όπως και στο SVM, αυτές οι πιθανότητες θα μας βοηθήσουν στην αποτίμηση πέρα από τη στεγνή ακρίβεια (π.χ. μέσω ROC AUC).
* Να σημειωθεί ότι τα νευρωνικά δίκτυα έχουν την ικανότητα να μοντελοποιούν μη-γραμμικές σχέσεις στα δεδομένα. Δεδομένου ότι τα χαρακτηριστικά μας περιλαμβάνουν αρκετές δυαδικές στήλες (genres) και μία συνεχή (rating), ένα MLP μπορεί θεωρητικά να συλλάβει πολύπλοκες συσχετίσεις μεταξύ τους, πιθανώς αποδοτικότερα από ένα γραμμικό μοντέλο.

## 4.4 Αξιολόγηση Μεθόδων

### 4.4.1 Αξιολόγηση SVM, MLP

Για την αξιολόγηση των μοντέλων SVM και MLP χρησιμοποιήσαμε διάφορες μετρικές, ώστε να έχουμε πληρέστερη εικόνα της απόδοσής τους:

|  |  |
| --- | --- |
| Accuracy  (Ακρίβεια) | Το ποσοστό των παραδειγμάτων που ταξινομήθηκαν σωστά συνολικά. Δίνει μια πρώτη γενική ένδειξη της επιτυχίας του μοντέλου. |
| Precision  (Ευστοχία) | Ο λόγος των προβλέψεων θετικής κλάσης (δημοφιλής) που ήταν πράγματι σωστές. Δηλαδή πόσο “ακριβείς” είμαστε όταν το μοντέλο προβλέπει ότι μια ταινία είναι δημοφιλής – συγκεκριμένα, $Precision = \frac{TP}{TP+FP}$ (True Positives / (True Positives + False Positives)). Υψηλό precision σημαίνει λίγα false positives. |
| Recall  (Ανάκληση) | Ο λόγος των πραγματικά θετικών δειγμάτων που το μοντέλο κατάφερε να εντοπίσει. Με άλλα λόγια, από όλες τις δημοφιλείς ταινίες, πόσες τις βρήκε το μοντέλο; Υπολογίζεται ως $Recall = \frac{TP}{TP+FN}$. Υψηλό recall σημαίνει λίγα false negatives (το μοντέλο δεν “χάνει” πολλές δημοφιλείς ταινίες). |
| F1-score | Ο αρμονικός μέσος του precision και του recall. Το $F1$ δίνει μια συνολική εικόνα συνυπολογίζοντας τόσο την ακρίβεια όσο και την ανάκληση. Είναι ιδιαίτερα χρήσιμο όταν θέλουμε μια μόνο τιμή για απόδοση και οι κλάσεις είναι ανισόρροπες ή όταν υπάρχει trade-off μεταξύ precision και recall. |
| Confusion Matrix | Ο πίνακας σύγχυσης, που παρουσιάζει τον αριθμό των True Positives (TP), True Negatives (TN), False Positives (FP) και False Negatives (FN) των προβλέψεων του μοντέλου. Τον απεικονίσαμε ως heatmap για να είναι πιο εύληπτος. Ο πίνακας σύγχυσης βοηθά στο να δούμε λεπτομερώς πού κάνει λάθος το μοντέλο (π.χ. προβλέπει πολλές μη δημοφιλείς ως δημοφιλείς ή το αντίστροφο;). |
| Καμπύλη ROC & AUC | Υπολογίσαμε τις συντεταγμένες για την καμπύλη ROC (Receiver Operating Characteristic) και το εμβαδόν κάτω από αυτήν (AUC). Η καμπύλη ROC απεικονίζει το trade-off μεταξύ True Positive Rate (TPR = Recall) και False Positive Rate (FPR) καθώς μεταβάλλεται το threshold ταξινόμησης του μοντέλου. Το AUC (Area Under Curve) συνοψίζει την απόδοση: τιμή 1 σημαίνει τέλειος διαχωρισμός, 0.5 σημαίνει τυχαία επιλογή. Ένα υψηλό AUC δείχνει ότι το μοντέλο διαχωρίζει καλά τις δημοφιλείς από τις μη δημοφιλείς ταινίες ανεξαρτήτως του επιλεγμένου threshold. |
| Καμπύλη Precision-Recall & AUC | Επειδή στη συγκεκριμένη περίπτωση η θετική κλάση (δημοφιλής ταινία) δεν είναι πολύ σπάνια (περίπου 50%), η καμπύλη Precision-Recall είναι συμπληρωματική της ROC για αξιολόγηση. Τη σχεδιάσαμε υπολογίζοντας Precision και Recall για διάφορα thresholds και επίσης υπολογίσαμε το PR AUC (εμβαδόν κάτω από την καμπύλη Precision-Recall). Αυτή η καμπύλη είναι ιδιαίτερα χρήσιμη όταν η θετική κλάση είναι σπάνια, καθώς εστιάζει περισσότερο στην ικανότητα του μοντέλου να εντοπίζει θετικά παραδείγματα χωρίς να παράγει πολλά false positives. Στο πρόβλημά μας, μια υψηλή PR AUC θα σημαίνει ότι το μοντέλο συνδυάζει καλά precision και recall στην κλάση των δημοφιλών ταινιών. |

### 4.4.2 Σύγκριση SVM, MLP

Αφού εκπαιδεύσαμε και αξιολογήσαμε τα δύο μοντέλα, συγκεντρώσαμε τα βασικά αποτελέσματα στο test set για άμεση σύγκριση:

Αποτελέσματα μοντέλων στο Test Set:

|  |  |  |
| --- | --- | --- |
| Μετρική | SVM | MLP |
| Accuracy | 0.766 | 0.786 |
| Precision (Class 0) | 0.97 | 0.87 |
| Recall (Class 0) | 0.69 | 0.81 |
| F1-score (Class 0) | 0.80 | 0.84 |
| Precision (Class 1) | 0.57 | 0.63 |
| Recall (Class 1) | 0.95 | 0.73 |
| F1-score (Class 1) | 0.71 | 0.67 |
| *(Σημείωση: Class 0 = μη δημοφιλής, Class 1 = δημοφιλής ταινία)* |  |  |

Από τον παραπάνω πίνακα και τα αναλυτικά αποτελέσματα (classification report, καμπύλες κ.λπ.), μπορούμε να συγκρίνουμε τα δύο μοντέλα:

|  |  |
| --- | --- |
| Συνολική Ακρίβεια | Το MLP πέτυχε ελαφρώς καλύτερο συνολικό accuracy (78.6%) σε σύγκριση με το SVM (76.6%). Η διαφορά δεν είναι πολύ μεγάλη, αλλά υποδηλώνει ότι συνολικά το νευρωνικό δικτύο ταξινόμησε σωστά λίγο περισσότερες ταινίες. |
| Συμπεριφορά στην Κλάση 1  (Δημοφιλείς ταινίες) | Εδώ βλέπουμε μια σημαντική διαφορά φιλοσοφίας των μοντέλων. Το SVM έχει πολύ υψηλό Recall για την κλάση 1 – περίπου 0.95, που σημαίνει ότι βρίσκει το 95% των πραγματικά δημοφιλών ταινιών (σχεδόν καμία δημοφιλής δεν διέλαθε την προσοχή του). Ωστόσο, αυτό έγινε εις βάρος του Precision της κλάσης 1, το οποίο είναι μόλις 0.57. Δηλαδή, το SVM έχει την τάση να χαρακτηρίζει πολλές ταινίες ως “δημοφιλείς”, με αποτέλεσμα αρκετές από αυτές να είναι στην πραγματικότητα μη δημοφιλείς (false positives). Αντίθετα, το MLP είναι πιο ισορροπημένο: έχει Recall ~0.73 για την κλάση 1 (χάνει μερικές δημοφιλείς ταινίες – περίπου 27% δεν τις εντόπισε), αλλά καλύτερο Precision ~0.63 (λιγότερα false positives από το SVM). Συνεπώς, το SVM είναι πιο “επιθετικό” στην πρόβλεψη της δημοφιλίας (προσπαθεί να μη χάσει καμία επιτυχία, ακόμα κι αν σημάνει μερικούς λάθος συναγερμούς), ενώ το MLP είναι πιο συγκρατημένο και κάνει λιγότερες λανθασμένες θετικές προβλέψεις. |
| Συμπεριφορά στην Κλάση 0  (Μη δημοφιλείς ταινίες) | Εφόσον η κλάση 0 είναι το αντίθετο της 1, οι παρατηρήσεις αντιστρέφονται. Το SVM έχει εξαιρετικά υψηλό Precision για την κλάση 0 (0.97), που σημαίνει ότι όταν προβλέπει ότι μια ταινία δεν θα είναι δημοφιλής, σχεδόν πάντα έχει δίκιο. Όμως, το Recall για την κλάση 0 είναι πιο μέτριο (0.69), δηλαδή το SVM χάνει περίπου το 31% των μη δημοφιλών ταινιών (μερικές τις προβλέπει λανθασμένα ως δημοφιλείς). Το MLP, από την άλλη, μείωσε ελαφρώς το precision στο 0.87 (κάποιες προβλέψεις “μη δημοφιλής” αποδείχθηκαν λάθος), αλλά ανέβασε το recall στο 0.81 (βρίσκει περισσότερες από τις πραγματικά μη δημοφιλείς). Με απλά λόγια, το MLP κάνει λίγο περισσότερα λάθη χαρακτηρίζοντας λίγες δημοφιλείς ταινίες ως μη δημοφιλείς (μειώνοντας το precision\_0), αλλά εντοπίζει περισσότερες μη δημοφιλείς σωστά σε σχέση με το SVM. |
| F1-score | Το F1 συνδυάζει τις προηγούμενες μετρικές. Για την κλάση 0, το MLP έχει καλύτερο F1 (0.84 vs 0.80), ενώ για την κλάση 1 το SVM έχει οριακά καλύτερο F1 (0.71 vs 0.67) εξαιτίας του πολύ υψηλού recall του. Ωστόσο, το F1 της κλάσης 1 για το MLP δεν πέφτει πολύ (0.67), δείχνοντας ότι συνολικά το MLP διατηρεί μια πιο συμμετρική απόδοση. |

Τελική σύγκριση: Το MLP υπερέχει ελαφρώς σε overall accuracy και παρουσιάζει μια πιο ισορροπημένη συμπεριφορά ανάμεσα στις δύο κλάσεις. Το SVM, αντιθέτως, φαίνεται να έχει ρυθμιστεί έτσι ώστε να μην χάνει σχεδόν καμία δημοφιλή ταινία (υψηλό recall\_1), κάτι που όμως οδηγεί σε περισσότερους ψευδώς θετικούς συναγερμούς (χαμηλότερο precision\_1). Η επιλογή μεταξύ των δύο μοντέλων μπορεί να εξαρτηθεί από το ποια λάθη θεωρούνται πιο “κοστοβόρα” στην πράξη:

* Αν προτεραιότητα είναι να μην χαθεί καμία ταινία που θα μπορούσε να είναι επιτυχία (δηλ. να εντοπιστούν *όλες* οι δημοφιλείς), τότε το SVM θα ήταν προτιμότερο ή θα μπορούσε κανείς να ρυθμίσει κατάλληλα το threshold του MLP ώστε να αυξήσει το recall (έναντι του precision).
* Αν προτεραιότητα είναι η αποφυγή ψευδών συναγερμών (δηλ. να μην θεωρηθούν κατά λάθος πολλές μέτριες ταινίες ως επιτυχίες), τότε το MLP είναι καλύτερη επιλογή από προεπιλογή, καθώς διατηρεί υψηλότερο precision, ή θα μπορούσε να ρυθμιστεί το SVM/threshold ώστε να αυξηθεί το precision θυσιάζοντας λίγο recall.

A blue squares with white text

AI-generated content may be incorrect.Σε κάθε περίπτωση, και τα δύο μοντέλα δίνουν χρήσιμα αποτελέσματα, με το MLP να έχει ένα ελαφρύ προβάδισμα στην ισορροπία μεταξύ των μετρικών. Οι καμπύλες ROC και Precision-Recall και για τα δύο μοντέλα έδειξαν AUC αρκετά πάνω από 0.8, υποδεικνύοντας ότι τόσο το SVM όσο και το MLP έχουν σημαντική προβλεπτική ικανότητα ως προς το διαχωρισμό των ταινιών σε δημοφιλείς και μη.

A blue squares with white text

AI-generated content may be incorrect.Εικόνα 12: Confusion Matrix SVM

Εικόνα 13: Confusion Matrix MLP

# 5. Συμπεράσματα

Εν κατακλείδι, τα αποτελέσματα της παρούσας ανάλυσης δείχνουν ότι ένα σύστημα Μηχανικής Μάθησης μπορεί όντως να συμβάλει στην πρόβλεψη της επιτυχίας μιας νέας ταινίας πριν καν αυτή κυκλοφορήσει. Αναλύοντας το MovieLens 100K dataset:

* Η μεθοδολογία K-Means clustering εντόπισε τρεις διακριτές ομάδες ταινιών, που χαρακτηρίζονται από διαφορετικούς συνδυασμούς ειδών και επιπέδων δημοφιλίας. Αυτό μας έδωσε μια εικόνα των κρυφών μοτίβων στο σύνολο δεδομένων (π.χ. μια ομάδα με action ταινίες υψηλής βαθμολογίας, μια με παιδικές περιπέτειες χαμηλότερης απήχησης, κ.ο.κ.).
* Η συγκριτική ανάλυση έδειξε ότι ο K-Means υπερτερεί έναντι του DBSCAN για τη συγκεκριμένη περίπτωση, παρέχοντας λίγα και ουσιαστικά clusters που μπορούν να ερμηνευθούν και να αξιοποιηθούν.
* Στο εποπτευόμενο σκέλος, το μοντέλο ταξινόμησης MLP κατάφερε να προβλέψει με ακρίβεια περίπου 78% ποιες ταινίες θα θεωρηθούν δημοφιλείς (με βάση τον αριθμό αξιολογήσεων). Το MLP επέδειξε πιο ισορροπημένη συμπεριφορά σε σύγκριση με το SVM, το οποίο όμως πέτυχε πολύ υψηλή ανάκληση για τις επιτυχίες. Και τα δύο μοντέλα πάντως πέτυχαν σημαντικά καλύτερη απόδοση από τυχαία πρόβλεψη, υπογραμμίζοντας ότι υπάρχουν αξιοποιήσιμα πρότυπα στα δεδομένα που σχετίζονται με τη “δημοφιλία” μιας ταινίας.

Πέρα από τις αριθμητικές αξιολογήσεις, είναι σημαντικό να δούμε πώς αυτά τα ευρήματα μπορούν να αξιοποιηθούν στην πράξη. Παρακάτω παρουσιάζονται ορισμένες πρακτικές εφαρμογές και ωφέλειες ενός τέτοιου μοντέλου πρόβλεψης στην κινηματογραφική βιομηχανία:

|  |  |
| --- | --- |
| 1. Εφαρμογή στον Κύκλο Ζωής μιας Ταινίας | o Στάδιο Παραγωγής: Πολύ πριν μια ταινία βγει στις αίθουσες – ακόμα και πριν ξεκινήσουν τα γυρίσματα ή ολοκληρωθεί το σενάριο – οι παραγωγοί μπορούν να “τεστάρουν” την ιδέα τους στο μοντέλο πρόβλεψης. Αν ο αλγόριθμος, με βάση παρόμοιες ταινίες και τα χαρακτηριστικά τους, δείξει υψηλή πιθανότητα επιτυχίας, αυτό μπορεί να ενθαρρύνει μεγαλύτερες επενδύσεις στην παραγωγή (σε προϋπολογισμό, σκηνικά, ειδικά εφέ, κάστινγκ γνωστών ηθοποιών κ.λπ.). Αν αντίθετα η πρόβλεψη είναι αρνητική, ίσως επανεκτιμήσουν στοιχεία του project (σενάριο, είδος) ή να είναι πιο συντηρητικοί στα κόστη.  o Στάδιο Προώθησης (Marketing): Λίγο πριν την κυκλοφορία μιας νέας ταινίας, το μοντέλο μπορεί να δώσει μια εκτίμηση του πόσο “εμπορική” αναμένεται να είναι. Με βάση αυτή την πληροφορία, η διανομέας εταιρεία αποφασίζει πόσο budget θα διαθέσει για διαφήμιση και προωθητικές ενέργειες. Για παράδειγμα, αν το μοντέλο προβλέπει χαμηλή απήχηση, ίσως επιλέξουν μια πιο στοχευμένη και περιορισμένη καμπάνια, εξοικονομώντας πόρους. Αν προβλέπεται μεγάλη επιτυχία, θα επενδύσουν περισσότερα στη διαφήμιση και ίσως επεκτείνουν την προβολή σε περισσότερες αγορές. Επίσης, μπορούν να αποφασίσουν σε ποιες χώρες ή σε ποια κανάλια (TV, social media, events) θα εστιάσουν ανάλογα με το κοινό-στόχο της ταινίας.  o Μετά την Κυκλοφορία: Μόλις η ταινία κυκλοφορήσει, το σύστημα συνεχίζει να συλλέγει δεδομένα – όπως πραγματικές βαθμολογίες, κριτικές κοινού, έσοδα box office – και μπορεί να ενημερώνει το μοντέλο (re-training) σε πραγματικό χρόνο. Έτσι, αν το κοινό αντιδρά διαφορετικά από το αναμενόμενο, το μοντέλο προσαρμόζεται. Αυτό αποτελεί ένα feedback loop που μπορεί να φανεί χρήσιμο και σε μελλοντικές παραγωγές: οι δημιουργοί μαθαίνουν από την απόδοση της ταινίας και τα χαρακτηριστικά της τί πήγε καλά και τί όχι, μέσω της “ματιάς” του μοντέλου. |
| 2. Πρακτική Χρησιμότητα για Διάφορους Παίκτες | o Streaming Platforms: Πλατφόρμες όπως το Netflix, Amazon Prime κ.ά. βασίζονται στις προτάσεις περιεχομένου για να κρατήσουν τους συνδρομητές ικανοποιημένους. Ένα μοντέλο που προβλέπει εκ των προτέρων ποιες νέες ταινίες έχουν μεγάλες πιθανότητες να “τραβήξουν” το ενδιαφέρον του κοινού, μπορεί να ενσωματωθεί στον σχεδιασμό περιεχομένου της πλατφόρμας. Για παράδειγμα, αν μια πλατφόρμα γνωρίζει ότι μια επερχόμενη ταινία (δική της παραγωγή ή εξωτερική) πιθανότατα θα είναι πολύ δημοφιλής, μπορεί να τη διαφημίσει έντονα στους χρήστες, να την τοποθετήσει σε περίοπτη θέση στην αρχική σελίδα, ή να εξασφαλίσει τα δικαιώματά της. Επιπλέον, οι αλγόριθμοι recommendation μπορούν να δώσουν έμφαση σε τέτοιες ταινίες, προτείνοντας τες ευρύτερα. Συνολικά, η γνώση προκαταβολικά της πιθανής επιτυχίας βοηθά στο να προταθεί πιο “ποιοτικό” και επιτυχημένο περιεχόμενο, διατηρώντας τους συνδρομητές ενεργούς και ικανοποιημένους.  o Κινηματογραφικές Εταιρείες & Διανομείς: Για τις παραδοσιακές κινηματογραφικές εταιρείες, το marketing μιας ταινίας είναι μια δαπανηρή επένδυση – συχνά δαπανώνται ποσά συγκρίσιμα με τον προϋπολογισμό παραγωγής σε προωθητικές ενέργειες. Ένα εργαλείο πρόβλεψης επιτυχίας μπορεί να λειτουργήσει ως “δείκτης εμπιστοσύνης” στο προϊόν. Αν η προβλεπόμενη απήχηση μιας ταινίας είναι μικρή, οι εταιρείες μπορεί να επιλέξουν να περιορίσουν τις δαπάνες διαφήμισης ή ακόμα και να ανακατανείμουν πόρους σε άλλα project με καλύτερες προοπτικές. Αν πάλι προβλέπεται blockbuster, θα φροντίσουν να μεγιστοποιήσουν την προβολή του. Επιπλέον, οι κινηματογραφικές αίθουσες (σινεμά) θα μπορούσαν να χρησιμοποιήσουν τέτοιες προβλέψεις για να κανονίσουν ανάλογα τον αριθμό αιθουσών ή προβολών που θα διαθέσουν σε μια νέα ταινία. |
| 3. Συνολική Αξία του “Μοντέλου Πρόβλεψης Επιτυχίας” | o Αποφυγή Ρίσκου: Σ’ έναν κλάδο με παραδοσιακά μεγάλο ρίσκο (πολλές ταινίες δεν αποδίδουν εμπορικά παρά τις επενδύσεις), η ύπαρξη ενός data-driven μοντέλου προσφέρει ένα πρόσθετο επίπεδο αντικειμενικής αξιολόγησης. Οι εταιρείες μπορούν να λαμβάνουν πιο τεκμηριωμένες αποφάσεις για το πού θα επενδύσουν τα χρήματά τους, μειώνοντας την πιθανότητα ακριβών αποτυχιών.  o Στοχευμένες Καμπάνιες: Με την πρόβλεψη της απήχησης, το μάρκετινγκ γίνεται πιο αποδοτικό. Οι πόροι κατανέμονται καλύτερα – π.χ. δεν ξοδεύονται υπέρογκα ποσά σε ταινίες που το μοντέλο “σημαίνει” ως προβληματικές, ενώ επενδύονται περισσότερα σε αυτές που έχουν δυναμική. Αυτό οδηγεί σε μειωμένο κόστος διαφήμισης και ενδεχομένως αυξημένη αποτελεσματικότητα των καμπανιών (αφού εστιάζουν εκεί που υπάρχει πρόσφορο έδαφος στο κοινό).  o Καινοτομία στον Κλάδο: Η εισαγωγή τεχνικών Machine Learning και Data Analytics φέρνει μια νότα καινοτομίας σε έναν παραδοσιακά “δημιουργικό” και εν μέρει υποκειμενικό χώρο όπως ο κινηματογράφος. Φυσικά, δεν αντικαθιστούν την ανθρώπινη δημιουργικότητα ή το καλλιτεχνικό ένστικτο, αλλά λειτουργούν συμπληρωματικά, παρέχοντας δεδομένα και προβλέψεις που μπορούν να υποστηρίξουν (ή να αμφισβητήσουν) τις καθιερωμένες πρακτικές.  o Δεδομενο-κεντρική Προσέγγιση: Συνολικά, η αλυσίδα αξίας μιας ταινίας – από τη σύλληψη της ιδέας, το στάδιο παραγωγής, τη διανομή στις αίθουσες μέχρι και το streaming – μπορεί να γίνει πιο δεδομενο-κεντρική (data-driven) και αποδοτική. Οι αποφάσεις που λαμβάνονται σε κάθε στάδιο μπορούν να υποβοηθούνται από δεδομένα και αναλύσεις, μειώνοντας την αβεβαιότητα. Αυτό δεν σημαίνει ότι εξαφανίζεται ο κίνδυνος ή η έκπληξη (πάντα θα υπάρχουν sleeper hits και ανέλπιστες αποτυχίες), αλλά το επιχειρησιακό κομμάτι του κινηματογράφου γίνεται πιο στοχευμένο και μετρήσιμο. |

Με τη βοήθεια λοιπόν της Αναλυτικής Δεδομένων και της Μηχανικής Μάθησης, η κινηματογραφική βιομηχανία αποκτά ένα ακόμη εργαλείο για να μετουσιώσει τα δεδομένα σε γνώση και να στηρίξει αποφάσεις που άλλοτε λαμβάνονταν μόνο εμπειρικά. Τα μοντέλα που αναπτύξαμε στο πλαίσιο αυτής της εργασίας αποτελούν ένα παράδειγμα του πώς μπορούμε να ποσοτικοποιήσουμε τη “δημοφιλία” και να την προβλέψουμε, συνδέοντας έτσι τον κόσμο της δημιουργικής παραγωγής με την επιστήμη των δεδομένων προς όφελος και των δύο.