ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Τμήμα Πληροφορικής

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

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 όπως βαθμολογίες χρηστών, είδη, ημερομηνίες κυκλοφορίας και πολλά άλλα, μπορούμε να επιχειρήσουμε την ανάπτυξη μοντέλων που προσεγγίζουν μια τέτοια πρόβλεψη.

Στην παρούσα εργασία χρησιμοποιείται το Movie Lens 100K dataset για την διερεύνηση του παραπάνω ερωτήματος.

Η προσέγγισή που ακολουθείται δεν είναι αμιγώς ακαδημαϊκή, καθώς έχει και πρακτικές εφαρμογές σε κινηματογραφικά στούντιο και πλατφόρμες streaming που θα ήθελαν να γνωρίζουν εκ των προτέρων ποιες ταινίες αξίζει να χρηματοδοτήσουν ή να προωθήσουν. Μελετώντας τα μοτίβα στα δεδομένα και κατασκευάζοντας ένα μοντέλο πρόβλεψης επιτυχίας, αναδεικνύεται πώς η Αναλυτική Δεδομένων μπορεί να υποστηρίξει τη λήψη αποφάσεων στην κινηματογραφική βιομηχανία.

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 επιτυχίας και τέλος, τα συμπεράσματα όπου συζητώνται η αποτελεσματικότητα των προσεγγίσεων και οι πρακτικές τους προεκτάσεις.

1.1 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 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)

2. Προ επεξεργασία Δεδομένων (Data Preprocessing)

2.1 Φόρτωση Δεδομένων

Τα τέσσερα αρχεία CSV του dataset φορτώθηκαν σε ξεχωριστά pandas Data Frames (ratings, movies, tags, links).

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 : 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 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.

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 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 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 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 μην υπερτερεί λόγω μεγέθους τιμών.

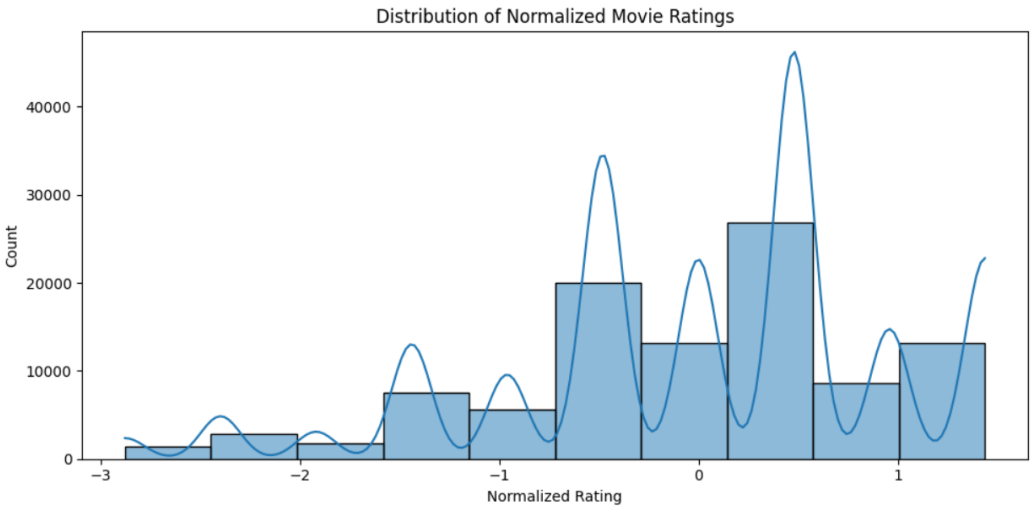
- Ενοποίηση Δεδομένων: Τα επεξεργασμένα δεδομένα βαθμολογιών και τα χαρακτηριστικά είδους (one-hot encoded genres) συνενώθηκαν σε ένα ενιαίο dataset. Κάθε ταινία πλέον αναπαρίσταται από ένα σύνολο αριθμητικών χαρακτηριστικών: τη κανονικοποιημένη βαθμολογία της και πληθώρα δεικτών (0/1) για την κατηγορία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

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:

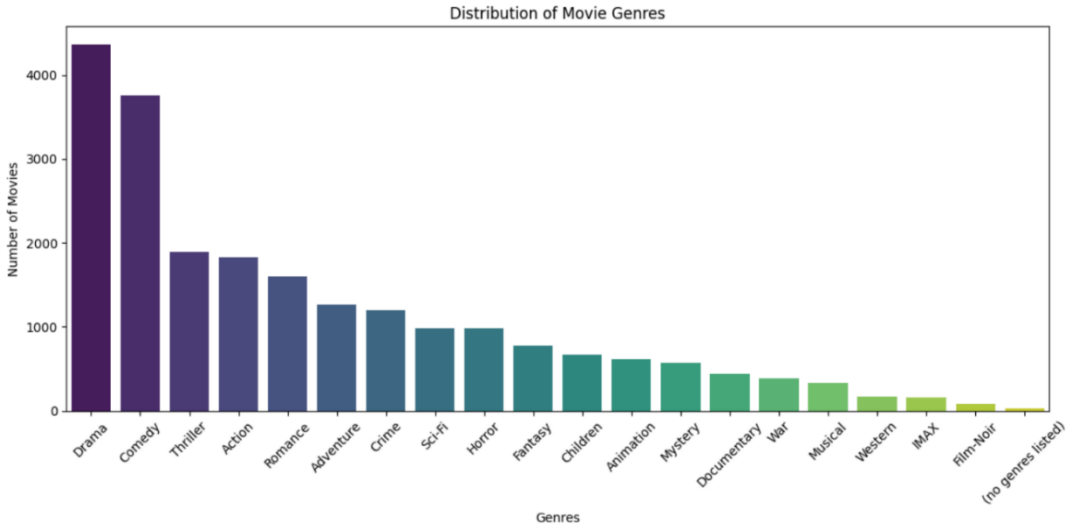
- Οι περισσότερες ταινίες έχουν μέση βαθμολογία συγκεντρωμένη μεταξύ 0 και 1 (ελαφρώς πάνω από τον μέσο όρο του dataset).
- Υπάρχει όμως και σημαντικός όγκος ταινιών με αρνητικές κανονικοποιημένες βαθμολογίες, δηλαδή αρκετές ταινίες έλαβαν βαθμολογίες κάτω του συνολικού μέσου όρου.
- Η διασπορά των τιμών δείχνει ότι οι αξιολογήσεις παρουσιάζουν μια σχετική διακύμανση γύρω από το μηδέν, αλλά οι περισσότερες δεν απέχουν δραματικά από τον μέσο όρο.



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

Κατανομή Ειδών Ταινίας (Bar Plot ανά Genre):

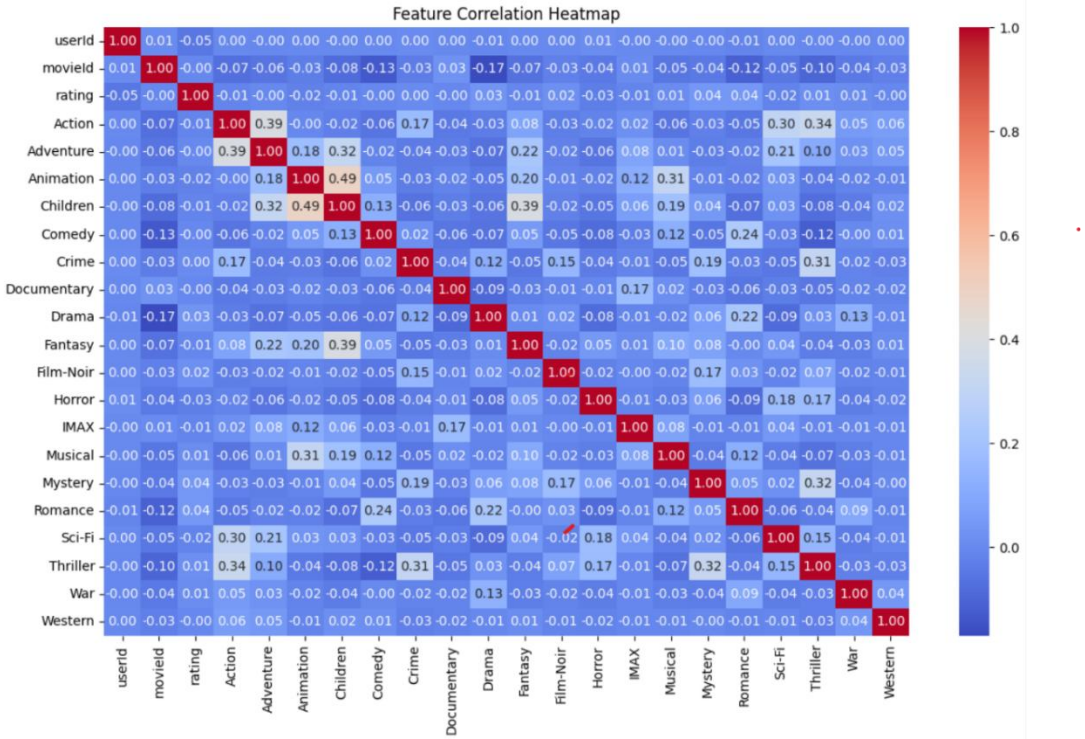
- Το Drama είναι το είδος με τις περισσότερες ταινίες στο dataset (περίπου 4.500 εμφανίσεις ειδών σε ταινίες αν συμπεριλάβουμε τις πολλαπλές κατηγορίες).
- Ακολουθεί η Comedy με περίπου 3.500 εμφανίσεις.
- Αντίθετα, σπανιότερα εμφανίζονται είδη όπως το Musical, το Western, το Film-Noir και το IMAX, καθώς και περιπτώσεις ταινιών που δεν έχουν δηλωμένο είδος. Αυτές οι κατηγορίες είναι ελάχιστες στο σύνολο.



Εικόνα 2: Distribution of movie genres

Χάρτης Συνεργασίας Χαρακτηριστικών (Correlation Heatmap):

- 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. Οι συσχετίσεις αυτές όμως δεν είναι εξαιρετικά υψηλές.
- 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.



Εικόνα 3: Feature Correlation Heatmap

3. Συσταδοποίηση (Clustering)

3.1 Σκοπός & Σχεδιασμός

Αφού ολοκληρώθηκε η προεπεξεργασία των δεδομένων και αποκτήσαμε εικόνα για τη διανομή τους το επόμενο βήμα είναι η συσταδοποίηση. 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:

- Αλγόριθμοι: Θα εφαρμοστούν δύο διαφορετικοί αλγόριθμοι clustering, ο 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 ώστε να επιλέξουμε την καλύτερη ομαδοποίηση των ταινιών.

3.2 Επιλογή χαρακτηριστικών και επεξεργασία

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

- Επιλογή Χαρακτηριστικών: Ως χαρακτηριστικά για το clustering χρησιμοποιούμε τις κανονικοποιημένες βαθμολογίες 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.
- 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, αυτή η κλιμάκωση διασφαλίζει ότι καμία μεταβλητή δεν κυριαρχεί λόγω μονάδων μέτρησης.
- Μειώσαμε τη διάσταση του διανύσματος χαρακτηριστικών εφαρμόζοντας Principal Component Analysis 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 φιλτράρουμε επίσης πιθανό θόρυβο στα δεδομένα.

3.3 K-Means

3.3.1 Υπολογισμός μετρικών για πιθανά K

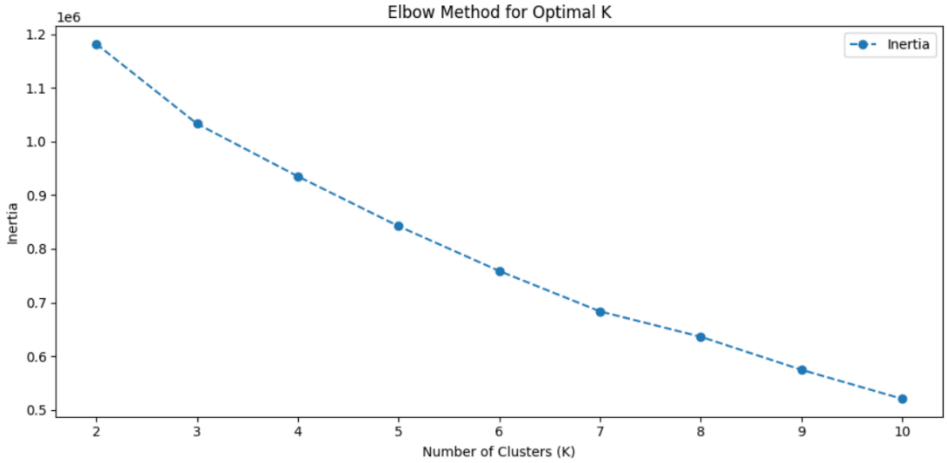
Για τον αλγόριθμο K-Means χρειάζεται να ορίσουμε τον αριθμό των clusters.

Προκειμένου να επιλέξουμε το βέλτιστο K, ακολουθήσαμε την εξής διαδικασία:

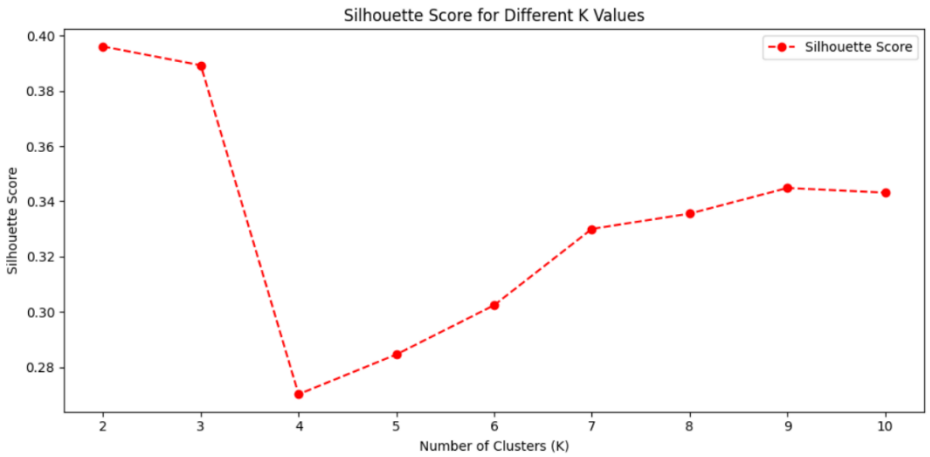
- Εκπαιδεύσαμε διαδοχικά μοντέλα K-Means για κάθε πιθανό αριθμό clusters για K από 2 έως 10 χρησιμοποιώντας το σύνολο χαρακτηριστικών που προέκυψε μετά το PCA.
- Για κάθε τιμή του K, υπολογίσαμε και καταγράψαμε διάφορες μετρικές ποιότητας clustering:
    - Inertia: άθροισμα των αποστάσεων όλων των σημείων από το πλησιέστερο κέντρο cluster. 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.
    - Δείκτης Davies-Bouldin: μετράει την ανεξαρτησία 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. 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).

3.3.2 Διαγράμματα K-means

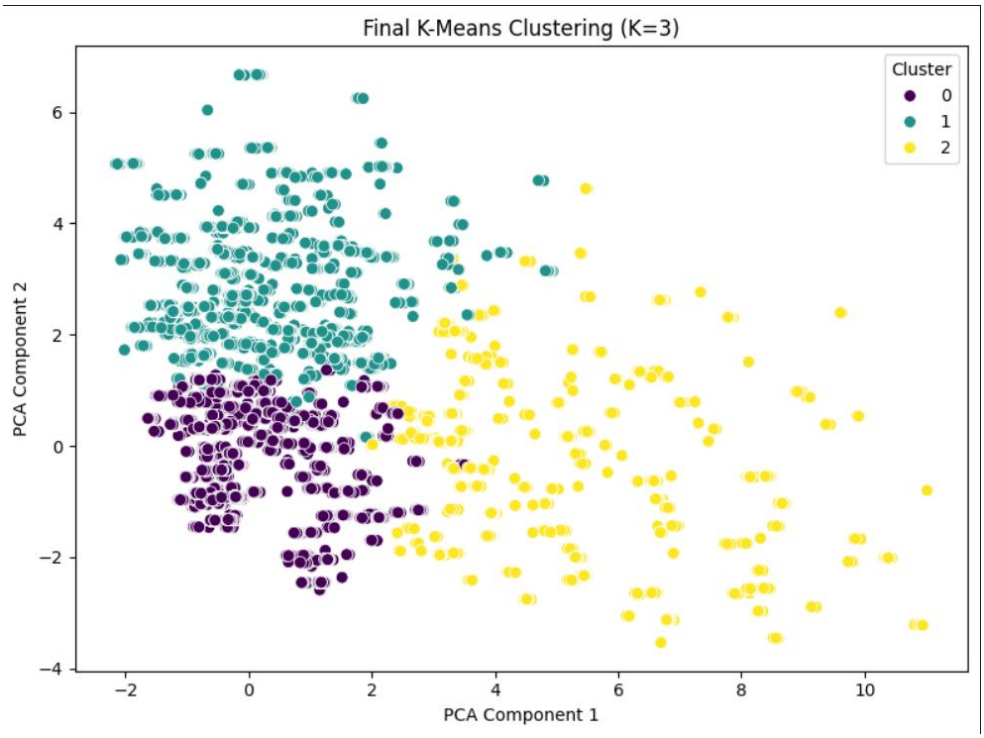


Εικόνα 4: Inertia plot



Εικόνα 5 : Silhouette score for different K values

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 6: K-means clusters

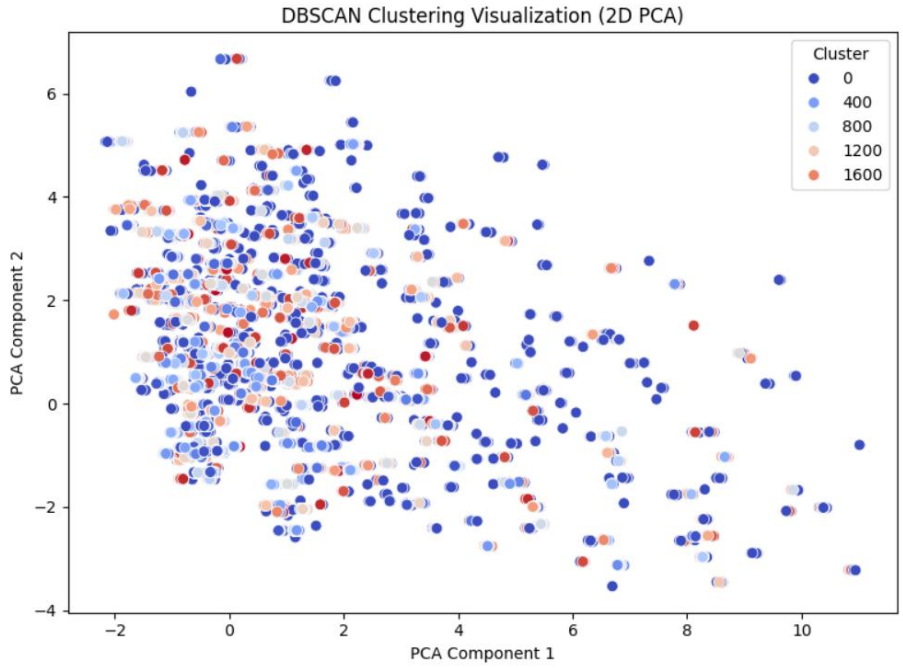
3.4 DBSCAN

Σε αντίθεση με τον K-Means, ο DBSCAN δεν απαιτεί προκαθορισμό του αριθμού clusters, αλλά στηρίζεται σε δύο άλλες παραμέτρους: την ακτίνα γειτονιάς (επιτρεπόμενη απόσταση) που ορίζει πότε δύο σημεία θεωρούνται κοντά (*eps*) 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 eps 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.

- Το Silhouette score χρησιμοποιήθηκε ως κύριο κριτήριο για βέλτιστη συσταδοποίηση, μιας και συγκεντρώνει πληροφορία τόσο για την ενδο-συνοχή όσο και για τον μεταξύ τους διαχωρισμό των clusters.
- 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.



Εικόνα 7: 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. Αντίθετα, πολλές από αυτές τις ταινίες ανήκουν σε είδη όπως το Drama και το Comedy. Συνοπτικά, το Cluster 0 φαίνεται να αντιστοιχεί σε ταινίες χωρίς έντονα στοιχεία δράσης ή animation, με μεσαίου επιπέδου αποδοχή από το κοινό.
Cluster 1	Περιλαμβάνει ταινίες που κατά μέσο όρο έχουν ελαφρώς υψηλότερη βαθμολογία από το Cluster 0. 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. Το Cluster 1, επομένως, μπορεί να χαρακτηριστεί ως το cluster των Action και Adventure ταινιών, οι οποίες μάλιστα τείνουν να έχουν σχετικά θετικές αξιολογήσεις.
Cluster 2	Περιλαμβάνει ταινίες με ελαφρώς χαμηλότερη του μέσου όρου βαθμολογία. 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). Το Cluster 2 φαίνεται να αντιστοιχεί σε ταινίες που είναι κυρίως περιπέτειες και κινούμενα σχέδια οι οποίες όμως δεν έτυχαν υψηλής αποδοχής καθώς 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 αποδοχή των ταινιών. Για παράδειγμα, το Cluster 1 συγκεντρώνει τις πιο καλά βαθμολογημένες ταινίες δράσης, 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, 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.

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 Movie Lens dataset σε διακριτές κατηγορίες, προτιμούμε την προσέγγιση του K-Means έναντι του DBSCAN.

4. Ταξινόμηση (Classification)

Αφού εξετάστηκε μια μη-εποπτευόμενη προσέγγιση (clustering) για την ανάλυση των δεδομένων, επόμενο βήμα αποτελεί η εποπτευόμενη μάθηση με στόχο την πρόβλεψη της επιτυχίας μιας ταινίας. Συγκεκριμένα, θα εκπαιδευτούν μοντέλα ταξινόμησης που προβλέπουν αν μια ταινία είναι δημοφιλής ή όχι με βάση τα χαρακτηριστικά της. Αυτό απαιτεί πρώτα να οριστεί ποια ταινία θεωρείται δημοφιλή και στη συνέχεια να εκπαιδευτούν αλγορίθμους πάνω σε δεδομένα με γνωστές τέτοιες ετικέτες.

4.1 Ορισμός Στόχου (Label)

Για να προσεγγιστεί το ζήτημα της πρόβλεψης επιτυχίας , χρειάζεται η μετατροπή του προβλήματος σε πρόβλημα δυαδικής ταξινόμησης. Ορίστηκε λοιπόν μια δυαδική ετικέτα στόχος, την οποία θα προβλέπουν τα μοντέλα που θα δημιουργηθούν:

- is_popular: Μια ταινία χαρακτηρίζεται ως δημοφιλής αν ο αριθμός αξιολογήσεων που έχει λάβει είναι πάνω από ένα ορισμένο όριο, αλλιώς χαρακτηρίζεται μη δημοφιλής.
- Κριτήριο: Χρησιμοποιήσαμε ως όριο τη διάμεσο του πλήθους αξιολογήσεων όλων των ταινιών. Δηλαδή, υπολογίστηκαν πόσες βαθμολογίες έχει η μέση ταινία στο σύνολο και θεωρήθηκαν ως δημοφιλείς όλες όσες έχουν περισσότερες αξιολογήσεις από αυτή την τιμή. 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 που σημαίνει ότι είναι δημοφιλής η ταινία και το άλλο μισό με ετικέτα 0, που σημαίνει πως η ταινία είναι μη δημοφιλείς, πράγμα που καθιστά το dataset ισορροπημένο ως προς τις κλάσεις και το κριτήριο αρκετά αυστηρό αλλά και αντικειμενικό.

4.2 Προετοιμασία συνόλου δεδομένων

Αφού ορίστηκε η ετικέτα is_popular για κάθε ταινία, επόμενο βήμα αποτελεί η δημιουργία του τελικού πίνακα χαρακτηριστικών για την ταξινόμηση και στον διαχωρισμό των δεδομένων σε training και test:

- Συγχωνεύτηκε η στήλη is_popular με τα ήδη επεξεργασμένα δεδομένα των ταινιών. Δημιουργήθηκε δηλαδή ένα νέο Data Frame το οποίο, για κάθε movieId, περιλαμβάνει όλα τα χαρακτηριστικά συν τη στήλη-στόχο is_popular.

- Διαχωρίστηκαν 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 αποτελεί την πρόβλεψη). Το y είναι το αντίστοιχο διάνυσμα τιμών της μεταβλητής is_popular (μήκους ίσου με τον αριθμό ταινιών).
- Έπειτα, χωρίστηκε το σύνολο των ταινιών σε δύο υποσύνολα: σύνολο εκπαίδευσης (training) και σύνολο δοκιμής (test). Χρησιμοποιήσαμε 80% των δεδομένων για training και κρατήσαμε 20% ως test, ώστε να αξιολογηθεί η απόδοση των μοντέλων σε άγνωστα δεδομένα. Επιπλέον, ο διαχωρισμός έγινε με stratify = y, δηλαδή διατηρήθηκε η ίδια αναλογία δημοφιλών και 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. 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 επιλέχθηκαν να εκπαιδευτούν δύο διαφορετικοί τύποι μοντέλων, ώστε να συγκριθούν οι 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 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, χρησιμοποιήθηκε ο SVC classifier από τη βιβλιοθήκη scikit-learn και ρυθμίστηκαν οι εξής τεχνικές:

- Ενεργοποιήθηκε ο υπολογισμός πιθανοτήτων στο SVM (probability=True) έτσι ώστε μετά την εκπαίδευση να μπορούν να εξαχθούν πιθανότητες πρόβλεψης για την κλάση δημοφιλής. Αυτό επιτρέπει τον σχεδιασμό καμπυλών αξιολόγησης όπως ROC και Precision-Recall.
- Θέτεται class_weight="balanced" κατά την αρχικοποίηση του SVM. Με αυτόν τον τρόπο, το μοντέλο θα λαμβάνει υπόψη ότι οι κλάσεις 0 και 1 είναι ισοπληθείς, και αποφεύγεται τυχόν μεροληψία αν υπήρχε ανισορροπία.
- Εκπαιδεύτηκε το SVM χρησιμοποιώντας τα δεδομένα εκπαίδευσης μετά την κλιμάκωση.

- Μετά την εκπαίδευση, χρησιμοποιήθηκε το μοντέλο για την πρόβλεψη της ετικέτας στο σύνολο δοκιμής. Αποθηκεύτηκαν τόσο οι προβλεφθείσες κλάσεις όσο και οι προβλεφθείσες πιθανότητες που δίνει το μοντέλο για την κλάση 1, που αφορά τις δημοφιλείς ταινίες. Οι πιθανότητες αυτές θα χρησιμεύσουν για τις καμπύλες ROC/PR.

4.3.2 Neural Network (Multi-Layer Perceptron)

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

- Καθορίστηκε η 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).
- Ορίστηκε μέγιστος αριθμό εποχών (iterations) = 100 για την εκπαίδευση. Αυτό δίνει στο δίκτυο αρκετές ευκαιρίες να συγκλίνει.
- Εκπαιδεύτηκε 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, εξάχθηκαν οι προβλέψεις κλάσης του δικτύου στο test set καθώς και οι πιθανότητες πρόβλεψης για την κλάση 1. Όπως και στο SVM, αυτές οι πιθανότητες θα μας βοηθήσουν στην αποτίμηση πέρα από τη στεγνή ακρίβεια (π.χ. μέσω ROC AUC).

4.4 Αξιολόγηση Μεθόδων

4.4.1 Αξιολόγηση SVM, MLP

Για την αξιολόγηση των μοντέλων SVM και MLP χρησιμοποιήθηκαν διάφορες μετρικές, ώστε να υπάρχει 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 ταινία) που ήταν πράγματι σωστές. Δηλαδή πόσο ακριβείς είμαστε όταν το μοντέλο προβλέπει ότι μια ταινία είναι δημοφιλής 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. Υψηλό recall σημαίνει λίγα false negatives (το μοντέλο δεν “χάνει” πολλές δημοφιλείς ταινίες).

F1-score	Ο αρμονικός μέσος του precision και του recall. 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. 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	Υπολογίστηκαν οι 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. Σχεδιάστηκε υπολογίζοντας Precision και Recall για διάφορα thresholds και 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.

4.4.2 Σύγκριση SVM, MLP

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 μπορούν να συγκριθούν τα δύο μοντέλα:

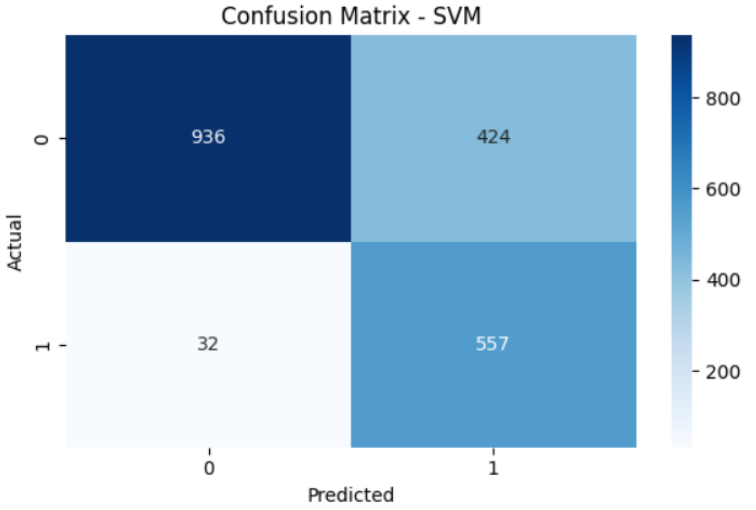
Συνολική Ακρίβεια	Το MLP πέτυχε ελαφρώς καλύτερο συνολικό accuracy (78.6%) σε σύγκριση με το SVM (76.6%). Η διαφορά δεν είναι πολύ μεγάλη, αλλά υποδηλώνει ότι συνολικά το νευρωνικό δικτύο ταξινόμησε σωστά λίγο περισσότερες ταινίες.

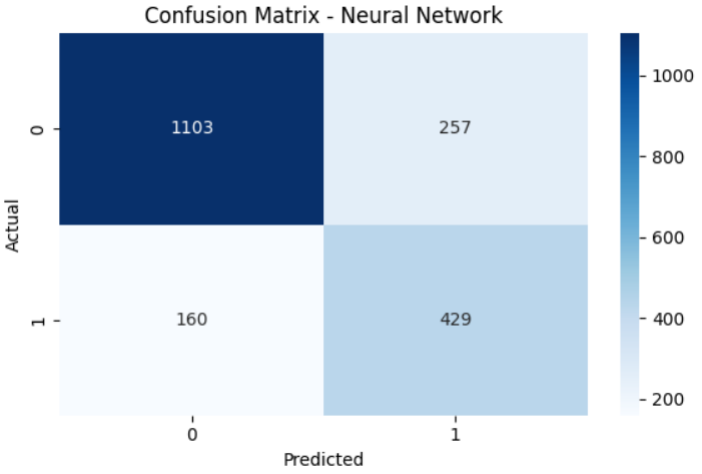
Συμπεριφορά στην Κλάση 1 (Δημοφιλείς ταινίες)	Παρατηρείται μια σημαντική διαφορά φιλοσοφίας των μοντέλων. Το SVM έχει πολύ υψηλό Recall για την κλάση 1 , περίπου 0.95, που σημαίνει ότι βρίσκει το 95% των πραγματικά δημοφιλών ταινιών. Ωστόσο, αυτό έγινε εις βάρος του Precision της κλάσης 1, το οποίο είναι 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, αλλά καλύτερο Precision 0.63 (λιγότερα false positives από το SVM). Συνεπώς, το SVM είναι πιο επιθετικό στην πρόβλεψη της επιτυχίας (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, 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.

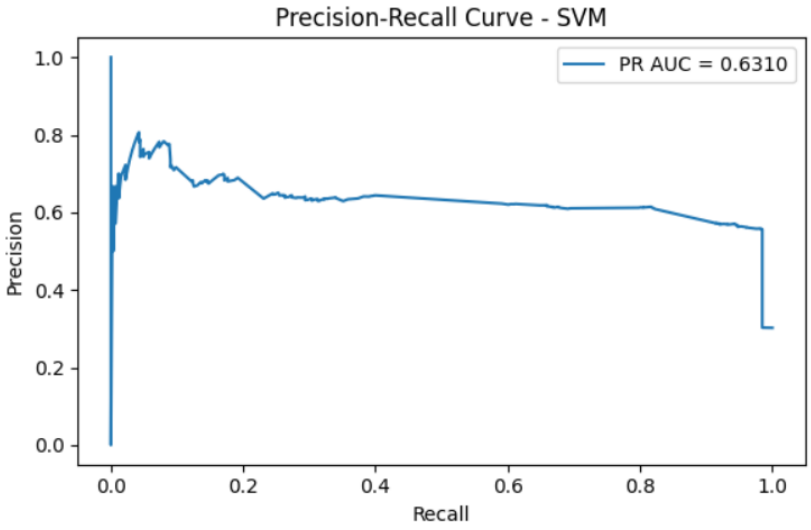
Τελική σύγκριση: Το MLP υπερέχει ελαφρώς σε overall accuracy και παρουσιάζει μια πιο ισορροπημένη συμπεριφορά ανάμεσα στις δύο κλάσεις. Το SVM, αντιθέτως, φαίνεται να έχει ρυθμιστεί έτσι ώστε να μην χάνει σχεδόν καμία δημοφιλή ταινία, κάτι που όμως οδηγεί σε περισσότερους ψευδώς θετικούς χαρακτηρισμούς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

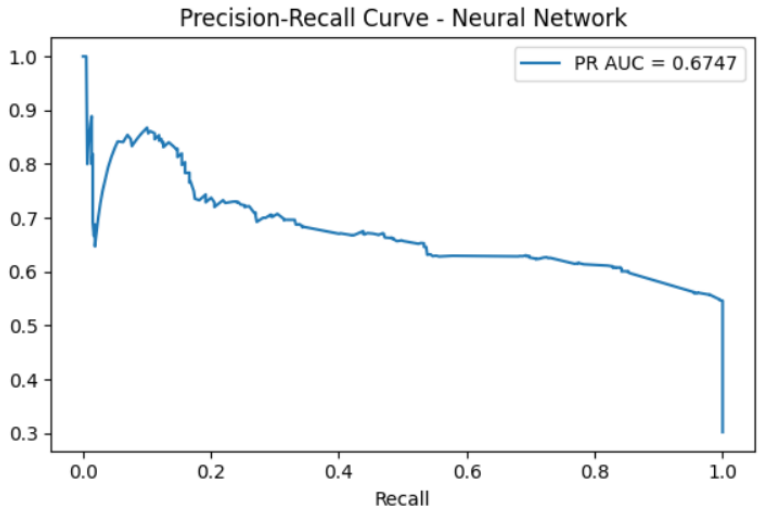
- Αν προτεραιότητα είναι να μην χαθεί καμία ταινία που θα μπορούσε να είναι επιτυχία, τότε το SVM θα ήταν προτιμότερο.
- Αν προτεραιότητα είναι η αποφυγή ψευδών συναγερμών (δηλαδή να μην θεωρηθούν κατά λάθος πολλές μέτριες ταινίες ως επιτυχίες), τότε το MLP είναι καλύτερη επιλογή.
- Σε κάθε περίπτωση, 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.

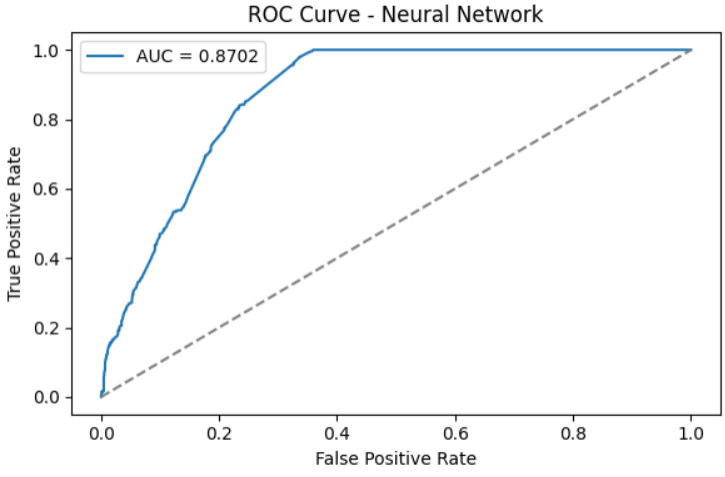
Διαγραμματική απεικόνιση:

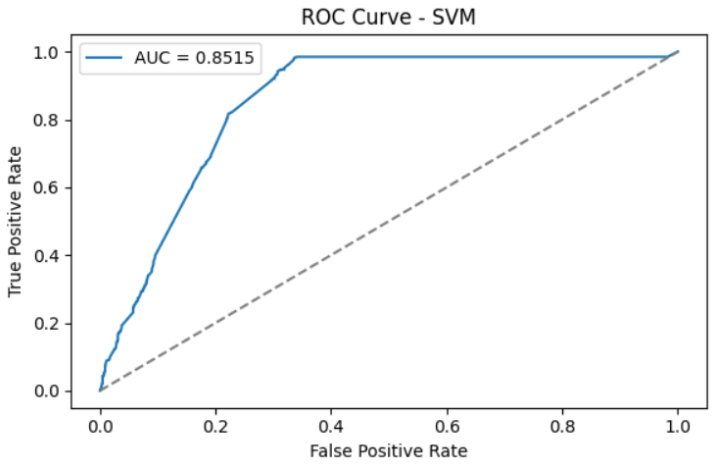












5. Συμπεράσματα

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 την επιτυχία 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:

Εφαρμογή στον Κύκλο Ζωής μιας Ταινίας	Στάδιο Παραγωγής: Πολύ πριν μια ταινία βγει στις αίθουσες 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 της ταινίας (σενάριο, είδος) ή να είναι πιο συντηρητικοί στα κόστη. Στάδιο Προώθησης: 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. Αφού κυκλοφορήσει: Μόλις η ταινία κυκλοφορήσει, το σύστημα συνεχίζει να συλλέγει δεδομένα 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 που μπορεί να φανεί χρήσιμο και σε μελλοντικές παραγωγές.

Πρακτική Χρησιμότητα για Διάφορους Παίκτες	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 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. Κινηματογραφικές Εταιρείες & Διανομείς: Για τις παραδοσιακές κινηματογραφικές εταιρείες, το marketing μιας ταινίας είναι μια δαπανηρή επένδυση. 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 θα μπορούσαν να χρησιμοποιήσουν τέτοιες προβλέψεις για να κανονίσουν ανάλογα τον αριθμό αιθουσών ή προβολών που θα διαθέσουν σε μια νέα ταινία.

Συνολική Αξία του Μοντέλου Πρόβλεψης Επιτυχίας:

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.

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 προβλέπει ως προβληματικές, ενώ επενδύονται περισσότερα σε αυτές που έχουν δυναμική. Αυτό οδηγεί σε μειωμένο κόστος διαφήμισης και ενδεχομένως αυξημένη αποτελεσματικότητα των ενεργειών.

Καινοτομία στον Κλάδο: Η εισαγωγή τεχνικών Τεχνητής Νοημοσύνης 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 και να αμφισβητήσουν τις καθιερωμένες πρακτικές.

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 την επιτυχία και να την προβλέψουμε, συνδέοντας έτσι τον κόσμο της δημιουργικής παραγωγής με την επιστήμη των δεδομένων προς όφελος και των δύο.