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

1.1 Πρόλογος

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

1.2 Περιγραφή του συνόλου δεδομένων

Το MovieLens 100K dataset προέρχεται από το group Lens Research και αποτελεί ένα από τα πλέον χρησιμοποιούμενα σύνολα δεδομένων σε ακαδημαϊκές μελέτες για συστήματα συστάσεων.

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

2. Προ επεξεργασία δεδομένων (Data Preprocessing)

2.1 Εξερεύνηση και Επεξεργασία δεδομένων

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

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

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

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

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

2.2 Μετασχηματισμοί

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

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

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

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

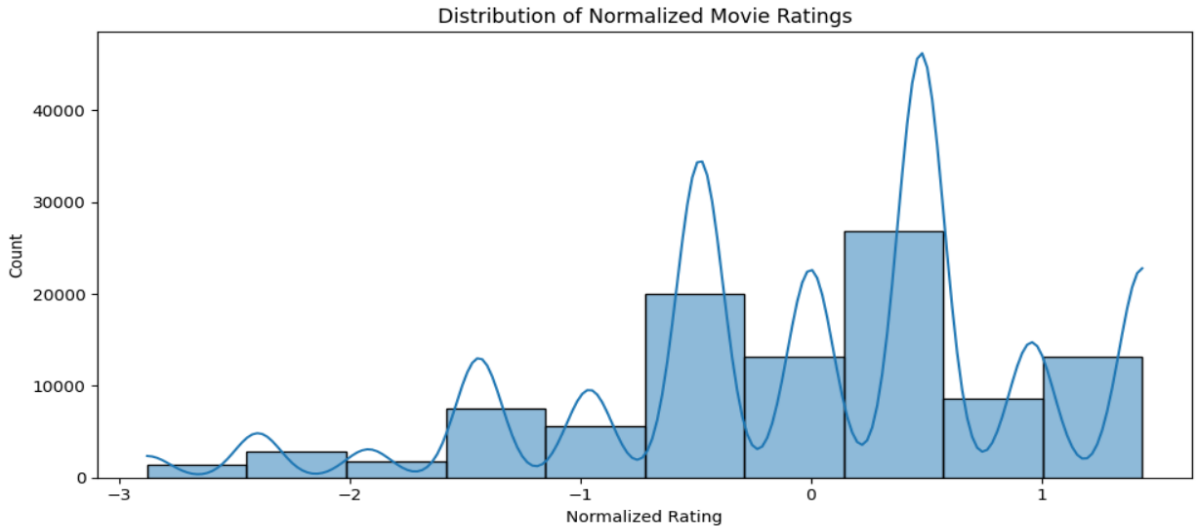
2.3 Ανάλυση δεδομένων

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

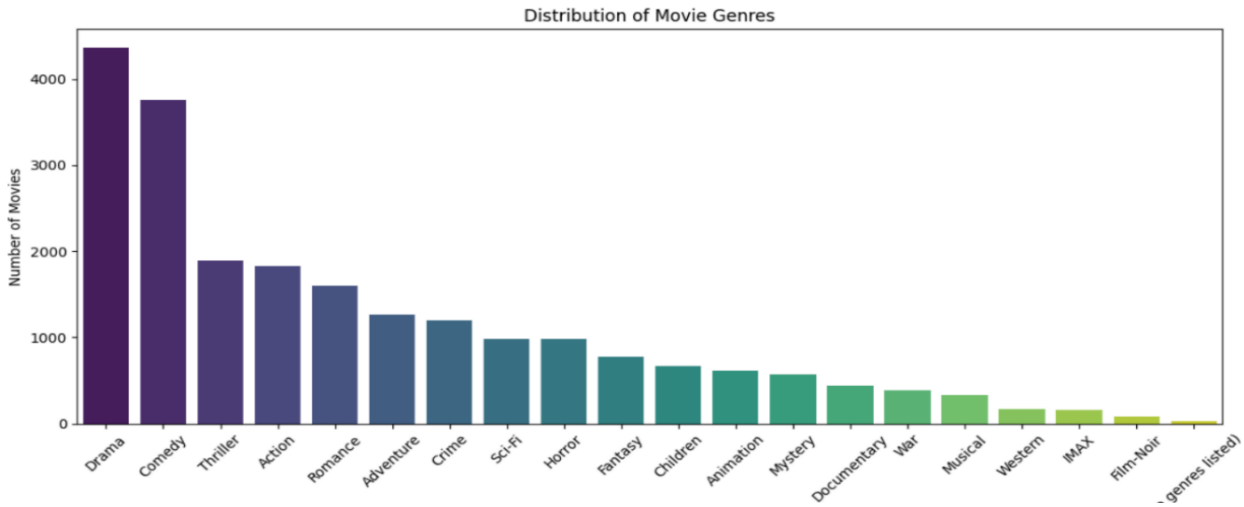


Εικόνα 1: Histogram with kde for the distribution of normalized ratings

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

• Ακολουθεί το Comedy με περίπου 3.500 εμφανίσεις.

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



Εικόνα 2: Bar Plot for the distribution of movie genres

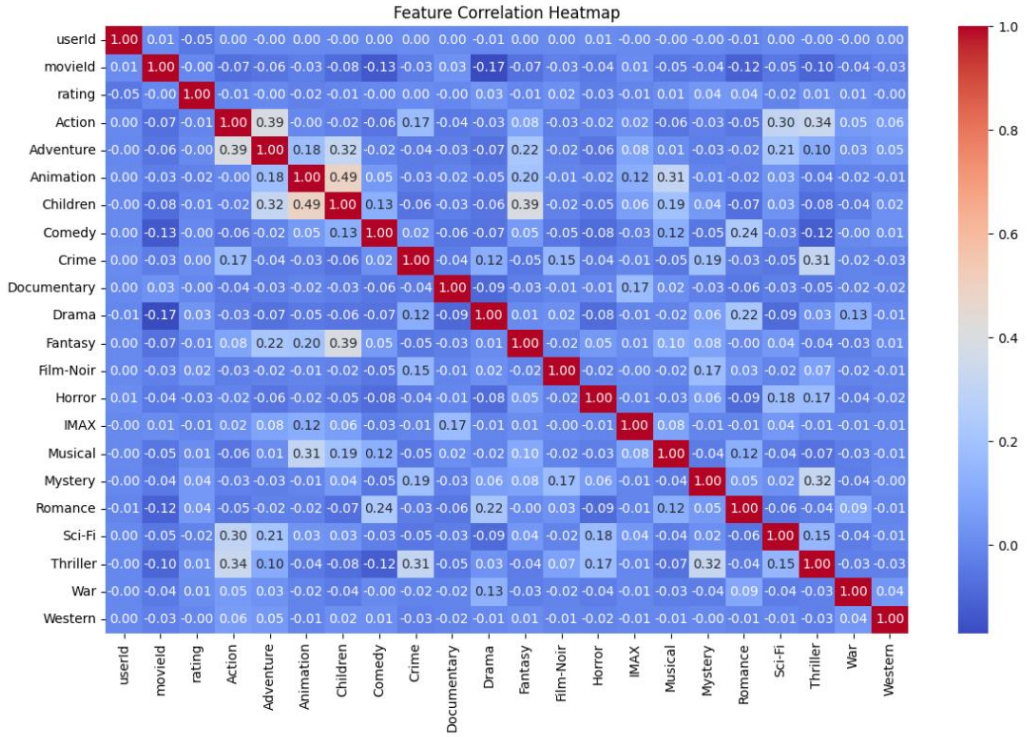
Συσχέτιση χαρακτηριστικών:

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

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

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

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



Εικόνα 3: Feature Correlation Heatmap

3. Συσταδιοποίηση (Clustering):

3.1 Σχεδιασμός και σκοπός

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

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

• Αξιολόγηση Συστάδων: Για να εκτιμηθεί η ποιότητα των clusters που προκύπτουν από τους παραπάνω αλγορίθμους, θα χρησιμοποιηθούν διαγράμματα και μετρικές αξιολόγησης των συστάδων.
Πιο συγκεκριμένα θα υπολογιστούν: Inertia, Silhouette Score, Davies-Bouldin Index. Οι παραπάνω μετρικές θα αξιοποιηθούν με σκοπό να συγκριθεί η αποτελεσματικότητα των αλγορίθμων clustering.

3.2 Επιλογή χαρακτηριστικών και επεξεργασία

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

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

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

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

3.3 K-means

3.3.1 Υπολογισμός μετρικών για εύρεση βέλτιστου K

Ο συγκεκριμένος αλγόριθμος χρειάζεται τον αριθμό των clusters K που θα δημιουργήσει.

Προκειμένου να βρεθεί το βέλτιστο K ακολουθήθηκε η εξής διαδικασία:

Εκπαιδεύτηκαν μοντέλα K-MEANS για K από 2 έως 10. Για κάθε τιμή του K υπολογίστηκαν και κατεγράφησαν οι μετρικές ποιότητας όπως:

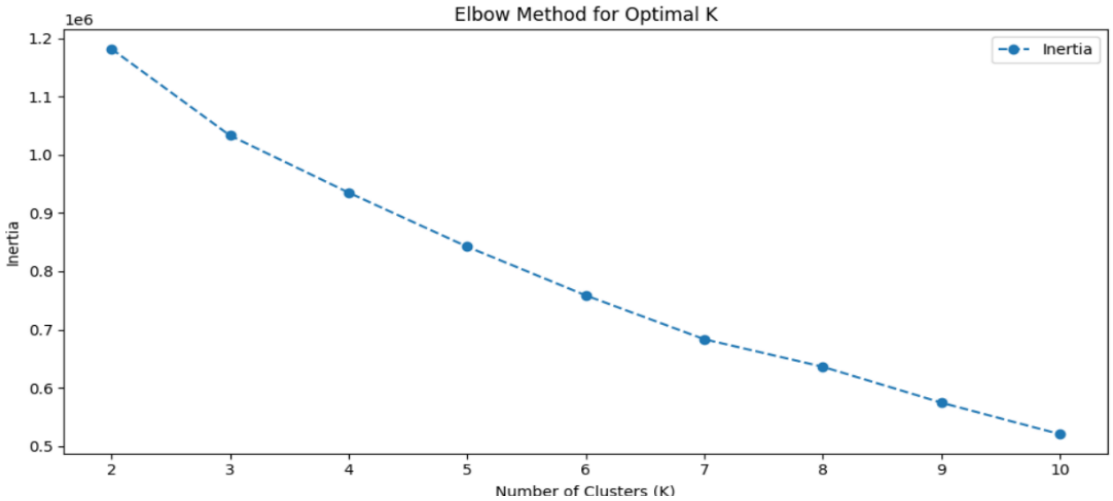
silhouette score: για τον διαχωρισμό των clusters

inertia: δείχνει πόσο συμπαγή είναι τα clusters

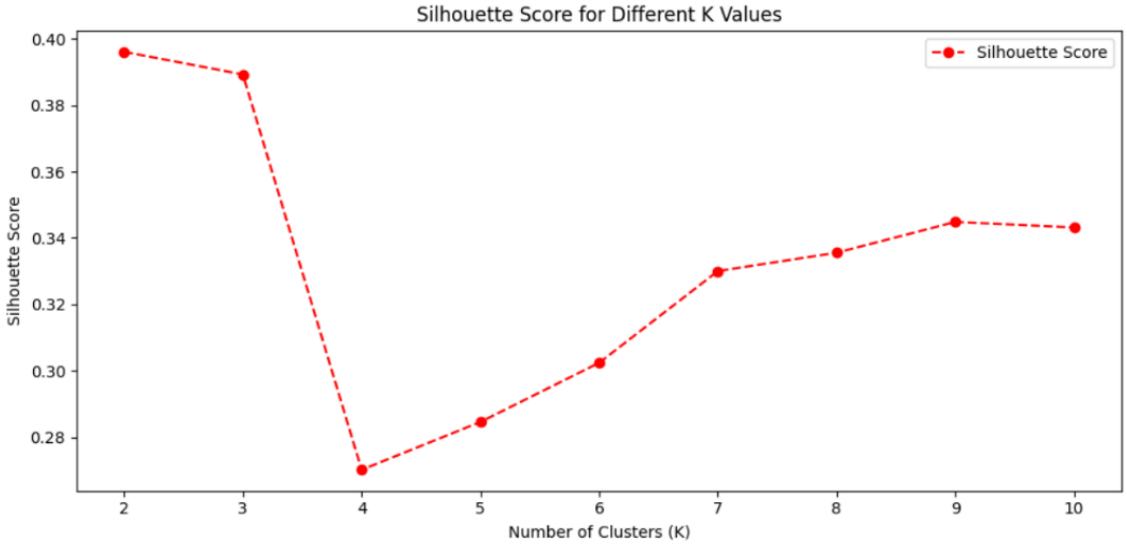
Davies Bouldin index: δείχνει την ανεξαρτησία των clusters

Σε συνδυασμό με τον υπολογισμό των παραπάνω μετρικών, εξετάστηκε και το μέγεθος των clusters για κάθε K με σκοπό να ελεγχθεί το αν προκύπτουν υπερβολικά άνισα clusters.

3.3.2 Διαγραμματική απεικόνιση και επιλογή βέλτιστου K

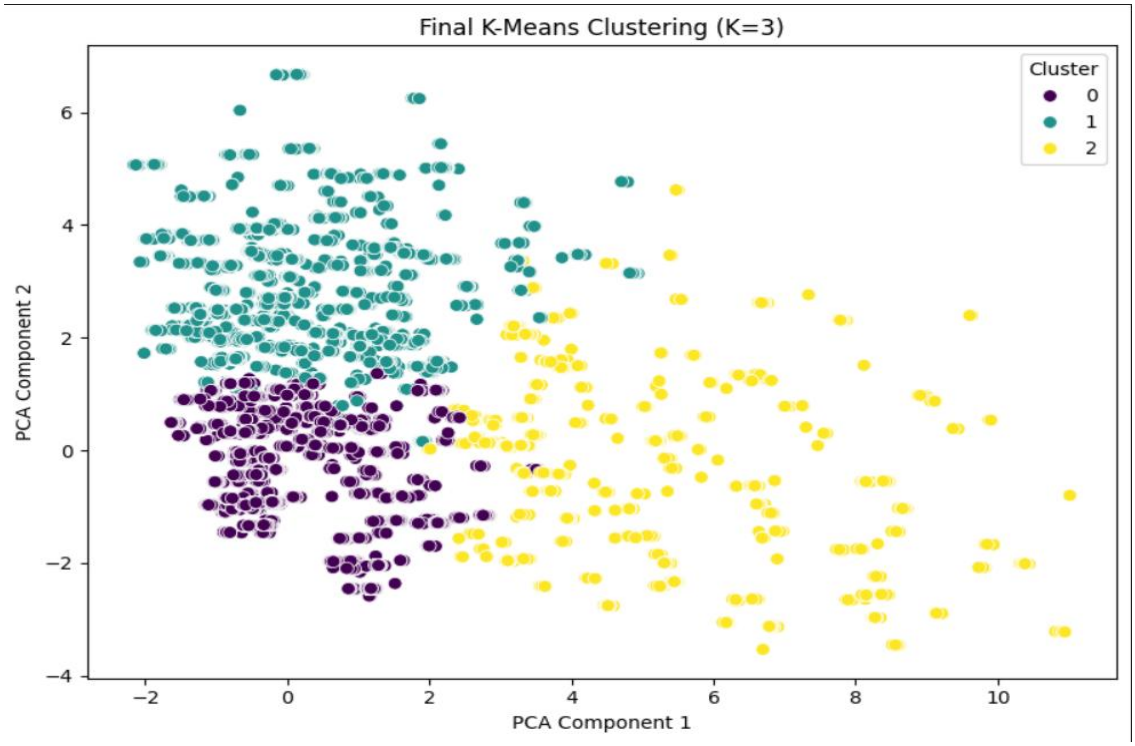


Εικόνα 4: Inertia Plot



Εικόνα 5: Silhouette score plot

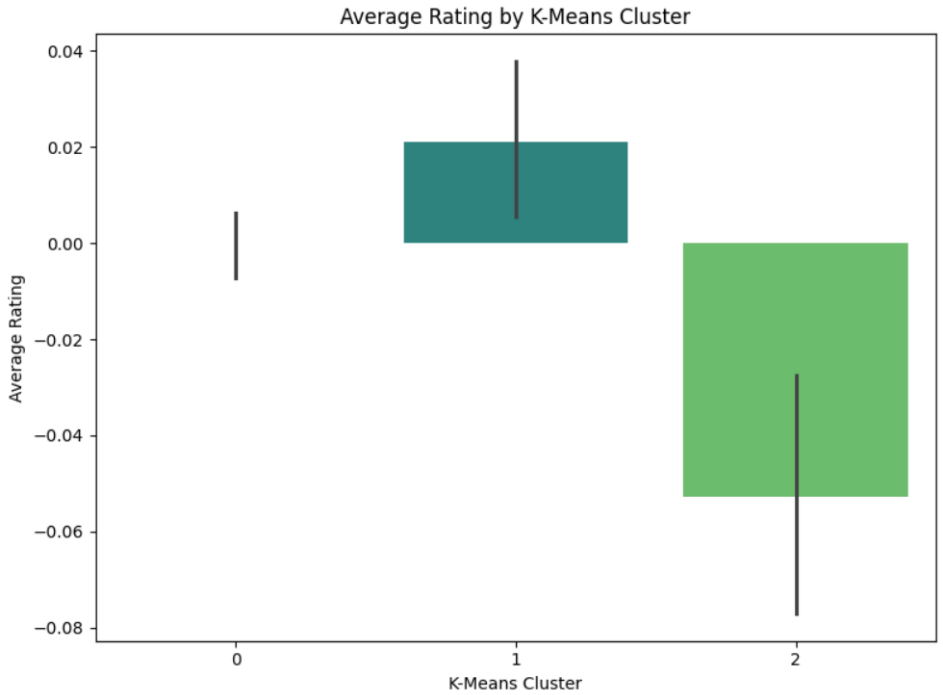
Συγκρίνοντας τις προαναφερθείσες μετρικές ποιότητας για όλα τα πιθανά K, παρατηρήθηκε σύγκλιση ενδείξεων υπέρ ενός συγκεκριμένου αριθμού clusters. Με τη βοήθεια των διαγραμμάτων παρατηρείται στο διάγραμμα Inertia elbow στο K=3, καθώς και στο διάγραμμα silhouette score το K=3 έχει υψηλό σκορ.



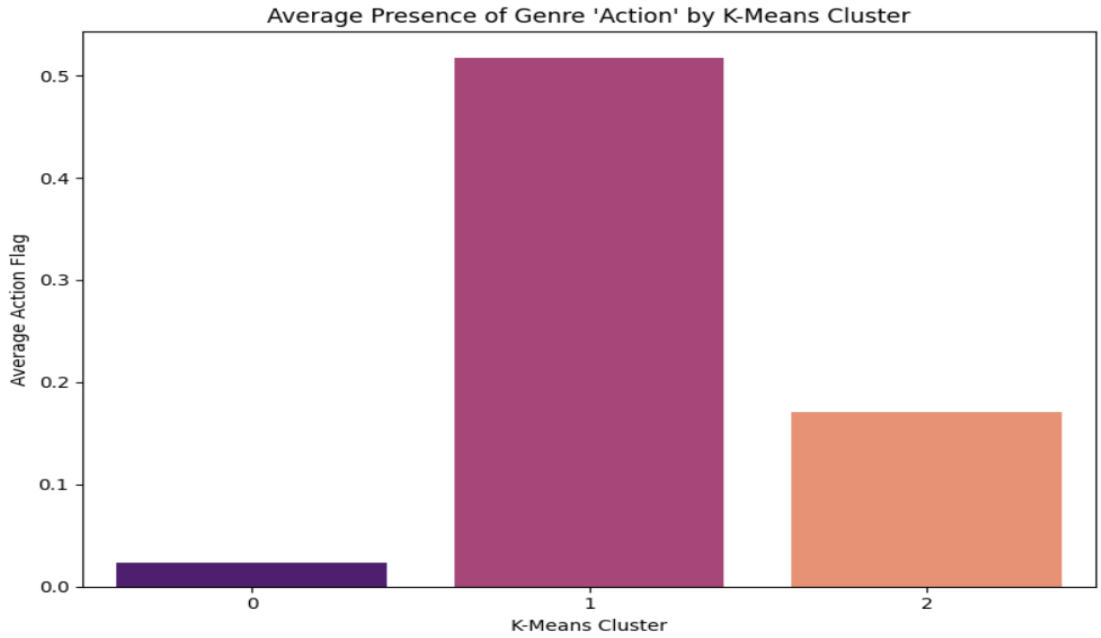
Εικόνα 6: K-means clusters scatter plot

Καταλήγοντας στο παραπάνω διάγραμμα, έχει επιτευχθεί καλός διαχωρισμός με ελάχιστες επικαλύψεις και αρκετά ισχυρά clusters.

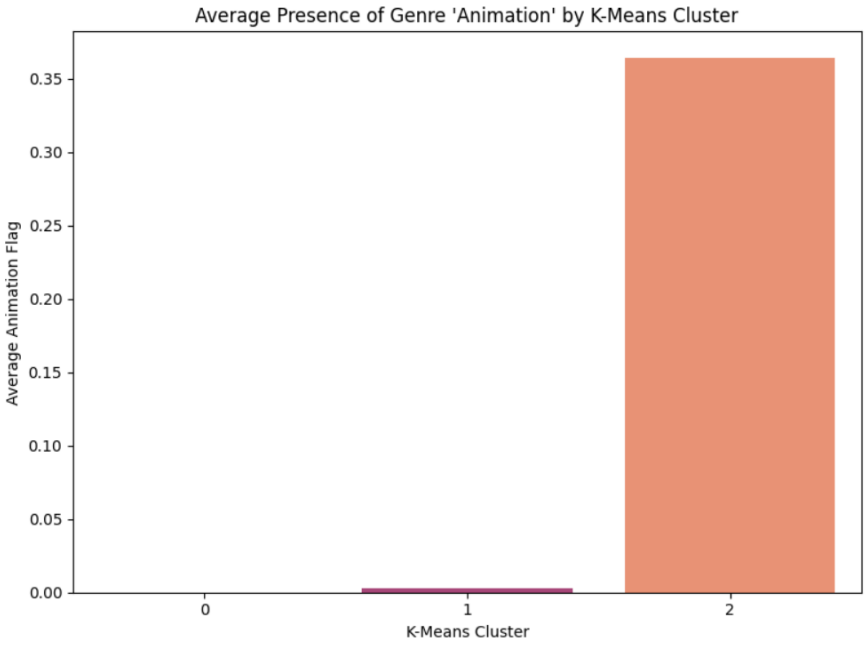
Αναλύοντας κάθε cluster ξεχωριστά και εξάγοντας στατιστικά και διαγράμματα, διακρίνονται τα παρακάτω συμπεράσματα:



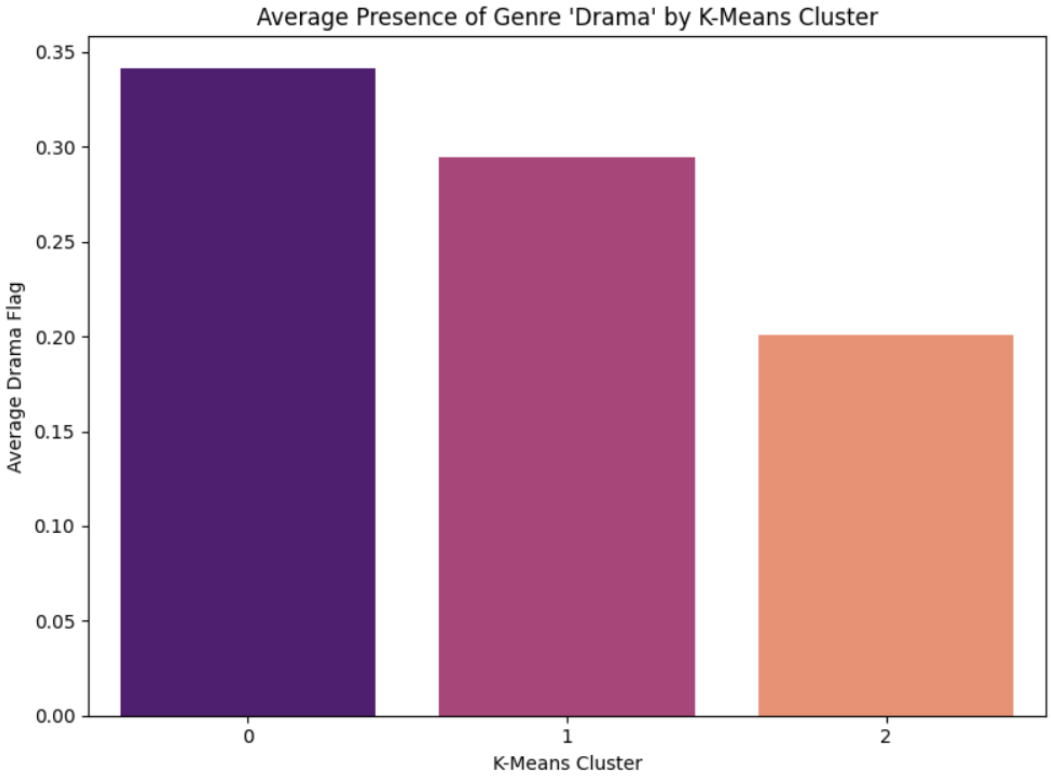
Εικόνα 7: Average rating per cluster



Εικόνα 8: Action movies per cluster



Εικόνα 9 Animation movies per cluster



Εικόνα 10: Drama movies per cluster

Cluster0: Περιλαμβάνει κυρίως πολλά είδη ταινιών με μέτριες προς ελαφρώς θετικές βαθμολογίες.

Cluster1: Αποτελείται κυρίως από Action, Adventure ταινίες με σχετικά θετικές βαθμολογίες.

Cluster2: Συνδυάζει Animation, Adventure με μέτριες προς χαμηλές βαθμολογίες.

Το drama genre περιέχεται σχεδόν σε όλες τις ταινίες, οπότε είναι ένας περισσότερο γενικός χαρακτηρισμός μιας ταινίας παρά μια ειδική κατηγορία.

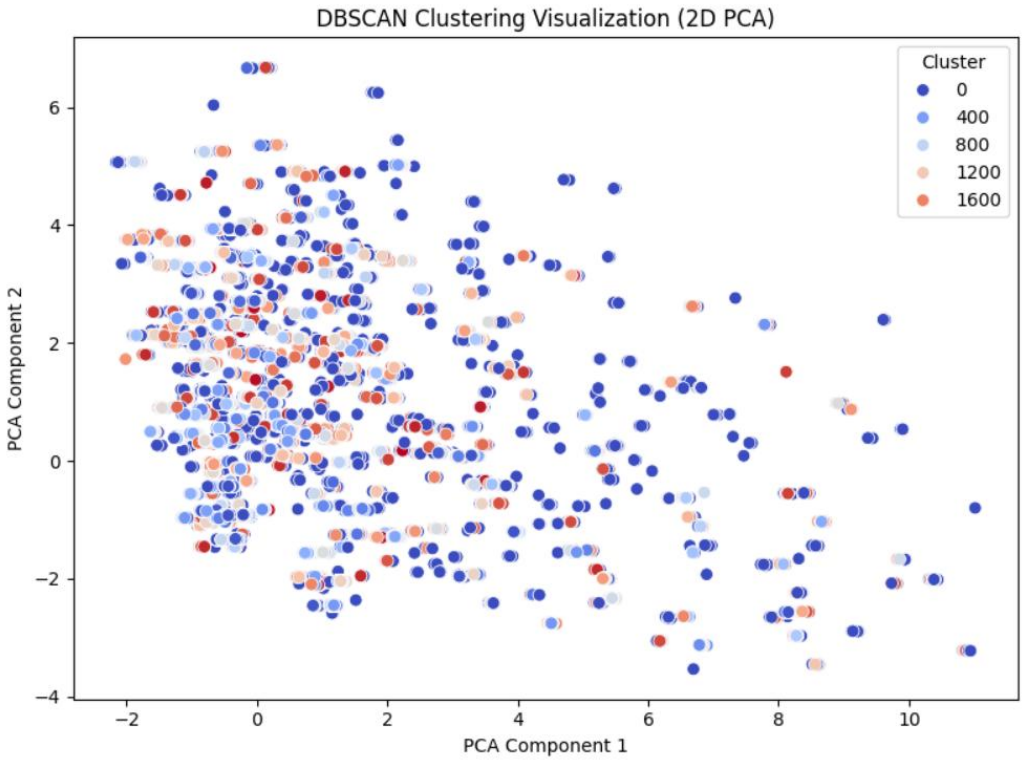
3.4 DBSCAN

Σε αντίθεση με τον K-means, ο αλγόριθμος DBSCAN δεν απαιτεί προκαθορισμό του αριθμού των clusters και στηρίζεται στις εξής παραμέτρους:

Eps: η επιτρεπόμενη απόσταση που κάνει 2 σημεία στο χώρο να θεωρούνται κοντά

Minimum Samples: ο ελάχιστος αριθμός σημείων που απαιτούνται για να σχηματίσουν ένα cluster

Προκειμένου να βρεθούν οι βέλτιστες παράμετροι για τον DBSCAN, πραγματοποιήθηκαν πολλοί συνδυασμοί τιμών των παραπάνω παραμέτρων. Για τους συνδυασμούς τιμών που δημιουργούσαν τουλάχιστον 2 clusters, υπολογίστηκαν οι μετρικές ποιότητας silhouette και davies-bouldin. Επειδή τα clusters του DBSCAN είναι πολλά και πολυδιάστατα, χρησιμοποιήθηκε PCA για να μειωθούν το data frame σε 2 διαστάσεις.



Εικόνα 11: DBSCAN clusters scatter plot

Παρατηρείται πως ο DBSCAN διαμόρφωσε πάρα πολλά clusters, χωρίς καθαρό διαχωρισμό, γεγονός που αποδεικνύει πως τα δεδομένα δεν παρουσιάζουν αρκετά πυκνές ομάδες που να μπορεί να αναγνωρίσει ο συγκεκριμένος αλγόριθμος.

3.5 Σύγκριση μεθόδων

Η σύγκριση των δύο παραπάνω αλγορίθμων clustering έδειξε διαφορετικές προσεγγίσεις στην ομαδοποίηση των ταινιών. Πιο συγκεκριμένα:

Ο K-means παρήγαγε ευδιάκριτα clusters που μπορούν να ερμηνευτούν θεματικά, καθώς εμφανίζουν διαφοροποιήσεις ως προς τα είδη και τη βαθμολογία τους.

Αντιθέτως ο DBSCAN δημιούργησε πολύ μεγάλο αριθμό από μικρά clusters, που δυσκολεύει την εξαγωγή χρήσιμων συμπερασμάτων.

Σύμφωνα με τις παραπάνω παρατηρήσεις, ο K-means αποτελεί την πιο κατάλληλη μέθοδο clustering πάνω σε αυτά τα δεδομένα, δίνοντας πιο ξεκάθαρα και ερμηνεύσιμα clusters, σε αντίθεση με τον DBSCAN που έδωσε τεράστιο αριθμό μη-ερμηνεύσιμων clusters.

4. Ταξινόμηση (Classification)

Μετά την ολοκλήρωση των διαδικασιών μη-εποπτευόμενης μάθησης, επόμενο βήμα αποτελεί η δημιουργία μοντέλων ταξινόμησης εποπτευόμενης μάθησης για την πρόβλεψη της επιτυχίας μιας ταινίας.

4.1 Ορισμός ετικέτας στόχου

Η ανάλυση προσεγγίζεται ως πρόβλημα δυαδικής ταξινόμησης, ορίζοντας την ετικέτα στόχο is_popular. Με αυτόν τον τρόπο μια ταινία θεωρείται δημοφιλής αν ο αριθμός των αξιολογήσεών της είναι πάνω από ένα καθορισμένο όριο, αλλιώς ανήκει στις μη δημοφιλείς. Το όριο που καθορίστηκε είναι η διάμεσος του αριθμού αξιολογήσεων. Με αυτόν τον τρόπο το dataset παραμένει ισορροπημένο μεταξύ 2 κατηγοριών, καθιστώντας την ταξινόμηση αντικειμενική και αυστηρή, χωρίς προκαταλήψεις ανάμεσα σε προτιμήσεις και ήδη.

4.2 Προετοιμασία του συνόλου δεδομένων

Η ετικέτα συγχωνεύτηκε με τα επεξεργασμένα δεδομένα των ταινιών, δημιουργώντας το τελικό dataset για ταξινόμηση. Διαχωρίστηκε σε features τα οποία περιλαμβάνουν όλα τα χαρακτηριστικά των ταινιών (εκτός από το αναγνωριστικό ταινίας και την ετικέτα) και label που είναι η ετικέτα. Χρησιμοποιήθηκε το 80% των δεδομένων για εκπαίδευση και 20% για δοκιμή, διατηρώντας την αναλογία των κλάσεων ισορροπημένη. Τέλος, κλιμακώνονται τα χαρακτηριστικά πριν την εκπαίδευση τον μοντέλων, για να αποφευχθούν τυχόν ανισορροπίες τιμών.

4.3 Δημιουργία και Εκπαίδευση μοντέλων

Για την ταξινόμηση, επιλέχθηκαν δύο μοντέλα με διαφορετική προσέγγιση:

Support Vector Machine (SVM): Αναζητά βέλτιστο όριο απόφασης μεταξύ των κλάσεων

Multi-Layer Perceptron (MLP): Ένα νευρωνικό δίκτυο που μπορεί να μάθει πιο σύνθετα μοτίβα στα δεδομένα.

4.3.1 Support Vector Machine

Το SVM εκπαιδεύτηκε χρησιμοποιώντας τον svc classifier της scikit-learn με συγκεκριμένες ρυθμίσεις, όπως balanced weight class για την ισορροπία των κλάσεων και probability true με σκοπό την πρόβλεψη πιθανοτήτων και σχεδιασμό καμπυλών.

4.3.2 Multi-layer Perceptron

Το νευρωνικό δίκτυο εκπαιδεύτηκε με τον MLPClassifier της scikit-learn με συγκεκριμένες ρυθμίσεις, όπως early stopping αν δεν βελτιώνεται η απόδοση και verbosity για να παρακολουθείται η πρόοδος του μοντέλου κατά την εκπαίδευση.

Μετά την εκπαίδευση και των δύο μοντέλων, εξάχθηκαν οι προβλέψεις κλάσης του δικτύου στο test set καθώς και οι πιθανότητες πρόβλεψης για την κλάση 1, που θα χρησιμεύσουν στην οπτικοποίηση των μετρικών απόδοσής τους.

4.4 Αξιολόγηση μεθόδων

Η αξιολόγηση των μοντέλων έγινε με τις ακόλουθες μετρικές:

Accuracy (Ακρίβεια): Το ποσοστό των σωστά ταξινομημένων ταινιών.

Precision (Ευστοχία): Πόσες από τις ταινίες που προβλέφθηκαν ως δημοφιλείς ήταν όντως επιτυχίες.

Recall (Ανάκληση): Από τις πραγματικά δημοφιλείς ταινίες, πόσες εντοπίστηκαν σωστά.

F1-score: Ο αρμονικός μέσος Precision & Recall, χρήσιμο όταν υπάρχει trade-off μεταξύ των δύο.

Confusion Matrix: Παρουσιάζει τα λάθη του μοντέλου (false positives & false negatives).

ROC & AUC: Δείχνει την ικανότητα του μοντέλου να διαχωρίζει τις κλάσεις σε διαφορετικά thresholds.

Precision-Recall Curve & AUC: Ιδιαίτερα χρήσιμη όταν η θετική κλάση είναι λιγότερο συχνή.

4.4.1 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
Class 0 = μη δημοφιλής, Class 1 = δημοφιλής		

Από τον παραπάνω πίνακα και τα αναλυτικά αποτελέσματα μπορούν να συγκριθούν τα δύο μοντέλα:

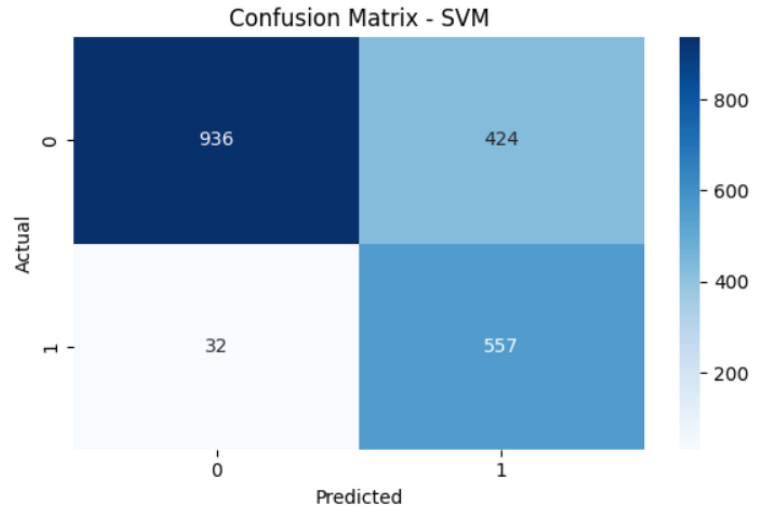
Το MLP έχει ελαφρώς καλύτερο accuracy, δείχνοντας πιο ισορροπημένη ταξινόμηση.

Στην κλάση 1, το SVM έχει πολύ καλό recall εντοπίζοντας σχεδόν όλες τις δημοφιλείς ταινίες, αλλά χαμηλό precision που σημαίνει ότι κάνει αρκετά false positives. Το MLP είναι πιο ισορροπημένο και στα δύο.

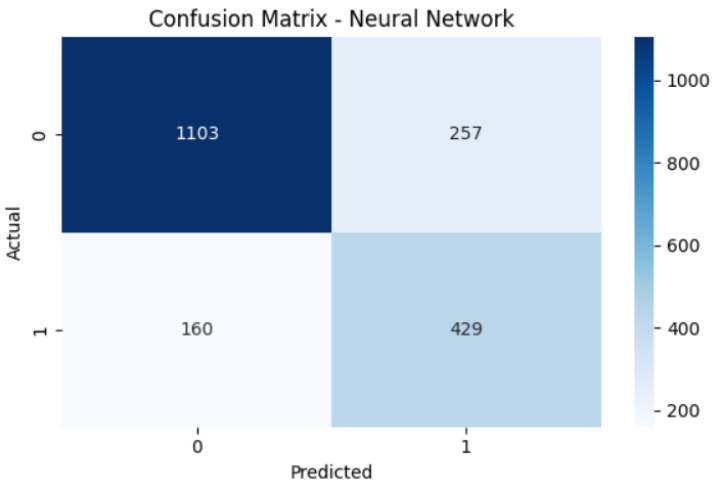
Στην κλάση 0, το SVM έχει πολύ ψηλό precision αλλά μέτριο προς χαμηλό recall, ενώ το MLP είναι πάλι πιο ισορροπημένο με υψηλότερες μετρικές σε αυτήν την κλάση.

4.4.2 Διαγραμματική απεικόνιση

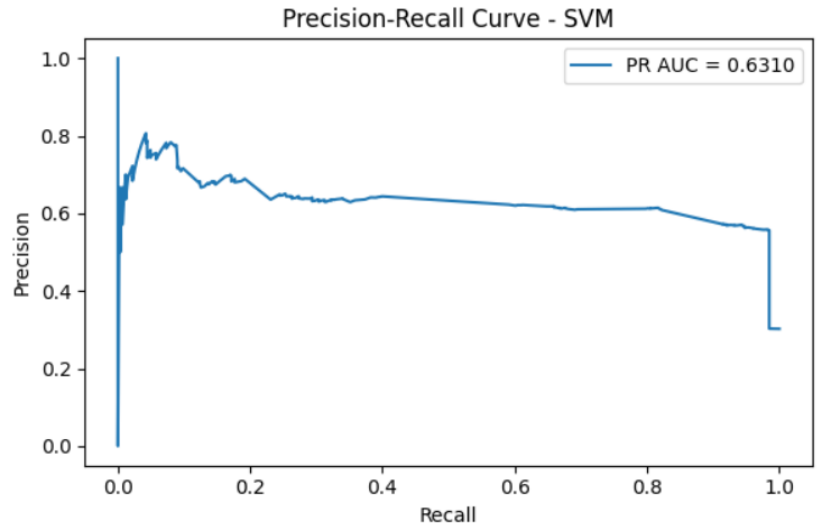
Η αριθμητική σύγκριση ενισχύεται με τα παρακάτω διαγράμματα:



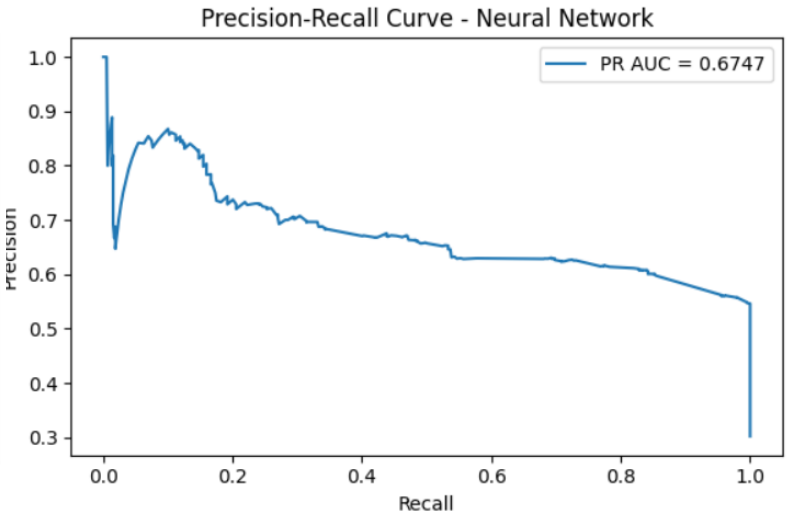
Εικόνα 12: SVM Confusion Matrix



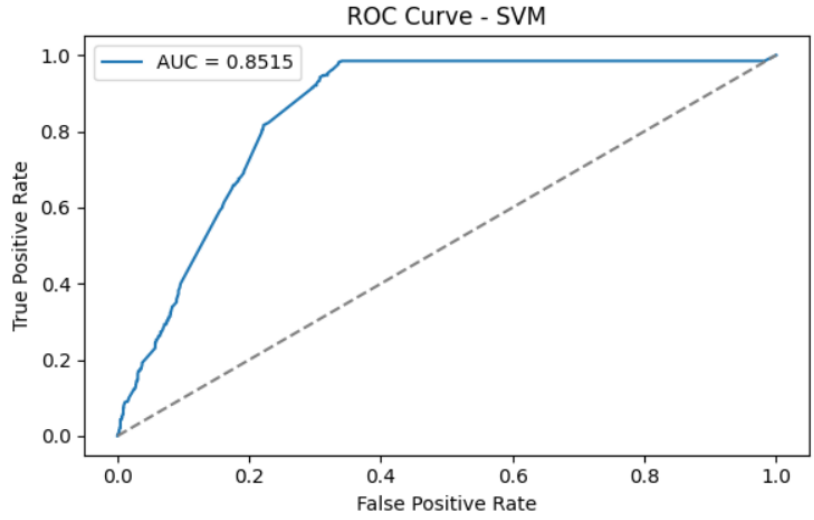
Εικόνα 13: MLP Confusion Matrix



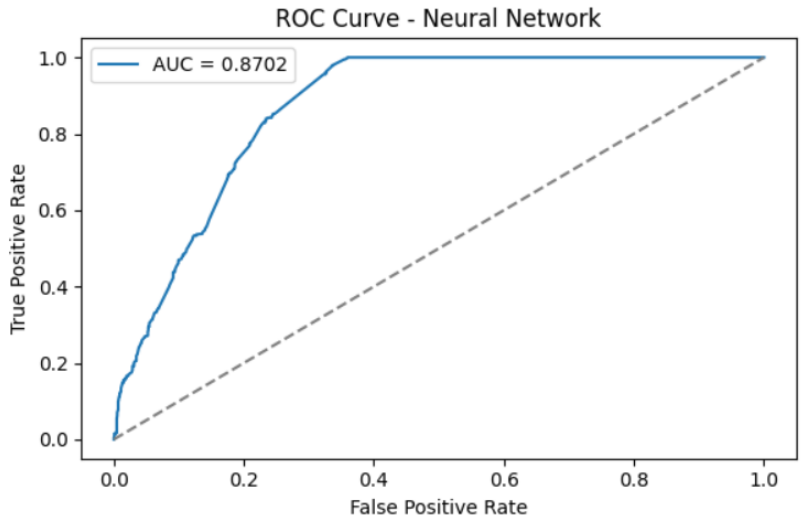
Εικόνα 14: SVM PR curve



Εικόνα 15: MLP PR curve



Εικόνα 16: SVM ROC curve



Εικόνα 17: MLP ROC curve

Παρατηρείται και μέσα από τα διαγράμματα, πως το SVM βρίσκει περισσότερες δημοφιλείς ταινίες αλλά κάνει και περισσότερα λάθη σύμφωνα με τα confusion matrices, ενώ από τις καμπύλες παρατηρείται πως το MLP είναι πιο ισορροπημένο μοντέλο και υπερτερεί του SVM τόσο στο trade-off μεταξύ precision-recall όσο και μεταξύ true positive rate και false positive rate.

4.4.3 Τελική Αποτίμηση

Η παραπάνω ανάλυση σε συνδυασμό με την διαγραμματική σύγκριση των δύο μεθόδων ταξινόμησης, έδειξε πως και τα δύο μοντέλα είναι ικανά να προβλέψουν αν μια ταινία θα είναι επιτυχημένη, το καθένα με διαφορετικά χαρακτηριστικά. Αν στόχος είναι να βρεθούν όλες οι πιθανώς επιτυχίες φαίνεται να είναι καλύτερη η επιλογή του SVM εφόσον έχει καλύτερες πιθανότητες να την βρει, εις βάρος όμως της ακρίβειας κάνοντας αρκετές λανθασμένες θετικές προβλέψεις. Αν στόχος είναι μια πιο ισορροπημένη προσέγγιση με λιγότερες λανθασμένες προβλέψεις η επιλογή του MLP είναι προτιμότερη. Τα δύο μοντέλα αποτελούν χρήσιμα εργαλεία για την πρόβλεψη επιτυχίας μιας ταινίας, με την επιλογή να εξαρτάται από τις ανάγκες και προτεραιότητες της εκάστοτε εφαρμογής.

5. Συμπεράσματα

Η παρούσα ανάλυση έδειξε πως η χρήση πρακτικών της Αναλυτικής Δεδομένων και της Μηχανικής Μάθησης μπορεί να συμβάλει στην πρόβλεψη επιτυχίας μιας ταινίας πριν αυτή κυκλοφορήσει.

Βασικά τεχνικά ευρήματα:

Ανακαλύφθηκαν συσχετίσεις και μοτίβα των δεδομένων μέσω στατιστικής ανάλυσης και μεθόδων clustering, με τον K-means να αποδίδει πιο διακριτά αποτελέσματα από τον DBSCAN. Στις μεθόδους εποπτευόμενης μάθησης, και τα δύο μοντέλα είχαν μεγάλη ικανότητα πρόβλεψης με το MLP να είναι πιο ισορροπημένο από το πιο επιθετικό, αλλά λιγότερο ακριβές SVM.

Εφαρμογές του μοντέλου στην πραγματική ζωή:

Τέτοια μοντέλα μπορούν να αξιοποιηθούν στην βιομηχανία του κινηματογράφου όσο και στις βιομηχανίες που παρέχουν υπηρεσίες streaming. Με τη βοήθεια αυτού του μοντέλου θα μπορούσε να καθοριστεί το budget και το casting μιας ταινίας, αλλά και ολόκληρο το marketing plan της ταινίας. Αντίστοιχα, θα μπορούσε να χρησιμοποιηθεί σε recommendation systems από πλατφόρμες streaming.

Συνολική αξία του μοντέλου:

Επιτυγχάνεται μείωση ρίσκου στις κινηματογραφικές επενδύσεις.

Ενισχύει διαφημιστικές καμπάνιες για αποδοτικότερη προώθηση και μείωση κόστους.

Βοηθάει στη λήψη στρατηγικών αποφάσεων με βάση τα δεδομένα.

Εν κατακλείδι, η ανάλυση δεδομένων και η μηχανική μάθηση μπορούν να αποτελέσουν πολύτιμο εργαλείο στρατηγικού σχεδιασμού βοηθώντας τεράστιες βιομηχανίες, όπως κινηματογραφική.