Πανεπιστήμιο Θεσσαλίας
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Μοντέλο Πρόβλεψης Επιτυχίας Κινηματογραφικών Ταινιών με χρήση Αλγορίθμων Μηχανικής Μάθησης

Διπλωματική Εργασία

ΔΙΟΝΥΣΙΟΣ Σ. ΜΠΑΣΔΑΝΗΣ

Επιβλέπων
Μιχαήλ Βασιλακόπουλος
Αναπληρωτής Καθηγητής

Βόλος, Ιούνιος 2020

Πανεπιστήμιο Θεσσαλίας
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Μοντέλο Πρόβλεψης Επιτυχίας Κινηματογραφικών Ταινιών με χρήση Αλγορίθμων Μηχανικής Μάθησης

Διπλωματική Εργασία

ΔΙΟΝΥΣΙΟΣ Σ. ΜΠΑΣΔΑΝΗΣ

Επιτροπή επίβλεψης

Επιβλέπων	Συνεπιβλέπουσα	Συνεπιβλέπουσα
Μιχαήλ Βασιλακόπουλος	Ελένη Τουσίδου	Παναγιώτα Τσομπανοπούλου
Αναπληρωτής Καθηγητής	Μέλος ΕΔΙΠ	Αναπληρώτρια Καθηγήτρια

Βόλος, Ιούνιος 2020

Πανεπιστήμιο Θεσσαλίας
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Ο/Η συγγραφέας αυτής της εργασίας βεβαιώνει ότι κάθε βοήθεια την οποία είχε για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης βεβαιώνει ότι έχει αναφέρει τις όποιες πηγές από τις οποίες έκανε χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται επακριβώς, είτε παραφρασμένες.

University of Thessaly
Faculty of Engineering
Department of Electrical & Computer Engineering

Predictive Model of Motion Picture Success using Machine Learning Algorithms

Diploma Thesis

DIONISIS S. BASDANIS

Supervisor
Michael Vasilakopoulos
Associate Professor University of Thessaly

Volos, June 2020

Περίληψη

Ο μεγάλος όγκος των διαθέσιμων δεδομένων προκάλεσε το ενδιαφέρον των ερευνητών για μελέτη και ανάπτυξη αλγορίθμων για την αξιοποίηση αυτών των δεδομένων και, κατά συνέπεια, τη διεξαγωγή συμπερασμάτων. Τον ρόλο της ανάλυσης των δεδομένων και της εύρεσης μοτίβων έχει αναλάβει ο κλάδος της εξόρυξης δεδομένων και της μηχανικής μάθησης. Αντικείμενο της εργασίας είναι η ανάλυση μιας βάσης δεδομένων που αφορά τις κινηματογραφικές ταινίες με σκοπό την πρόβλεψη της επιτυχίας τους (πρόβλεψη της βαθμολογίας τους) με χρήση αλγορίθμων μηχανικής μάθησης. Για αυτήν την εργασία χρησιμοποιήθηκε μία βάση βαθμολογημένων ταινίων από την ιστοσελίδα του IMDb και μία βάση με εγγραφές σχετικές με τα βραβεία όσκαρ των ηθοποιών και των σκηνοθετών. Το πρόβλημα προσεγγίστηκε με τρείς τρόπους. Αρχικά, έγινε η ανάλυση των δεδομένων και στη συνέχεια, μέσα από τα πειράματα, προέκυψε το συμπέρασμα πως οι ηθοποιοί, οι σκηνοθέτες, το είδος της ταινίας και η ανάλυση της εικόνας είναι οι παράγοντες που επηρεάζουν στο αν μια ταινία είναι επιτυχημένη ή όχι. Στη συνέχεια, στη δεύτερη προσέγγιση , εφαρμόστηκε η μέθοδος της ανάλυσης των κύριων συνιστωσών όπου μέσω των νέων γνωρισμάτων που δημιουργήθηκαν φάνηκε πως βελτιώθηκε το ποσοστό της ορθής πρόβλεψης. Τέλος, υλοποιήθηκε μια τρίτη προσέγγιση όπου, χρησιμοποιώντας τη βάση δεδομένων των βραβείων Όσκαρ, έγινε η προσπάθεια για καλύτερη και ακριβέστερη ταξινόμηση των γνωρισμάτων ηθοποιοί, σκηνοθέτες και λέξεις πλοκής. Μέσα από πειράματα φάνηκε πως τα αποτελέσματα του τρίτου αυτού μοντέλου προσέγγισαν και τελικά ήταν παρόμοια με αυτά του δεύτερου.

Λέξεις Κλειδιά

εξόρυξη δεδομένων, μηχανική μάθηση, R, Weka, βάση δεδομένων IMDb, ανάλυση κύριων συνιστωσών (ΑΚΣ).

Abstract

The large volume of data has aroused interest in the study and development of algorithms that use them to draw conclusions. The role of data analysis and pattern finding has been taken by the data mining and machine learning scientific branches. The subject of diploma is a database of motion pictures analysis to predict their success (predicting their score) using machine learning algorithms. For this work a database of graded films from the IMDb website and a database related to the oscar winners for actors and directors were used. The problem was approached in three ways. Initially, the data were analyzed and through the graphs the conclusion was drawn that the actors, the directors, the type of film and the analysis of the image are the factors that influence whether a film is successful or not. Following the first approach, the method of analyzing the principal components was applied, through which new features were created, a combination of the initial ones, and it seemed to improve the percentage of the correct prediction. Finally, another approach was taken in which the actors, directors and the plot keywords were taken as features and in combination with the database of the Oscars, the effort was made to better and more accurately classify these features. Thus, the classification was made by combining some factors and studying the coefficients of these factors. The results of the third approach showed that they are close with those of the second.

Keywords

data mining, machine learning, R, Weka, IMDb database, principal component analysis (PCA).

Αφιερωμένο στην οικογένεια μου.

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω την Συνεπιβλέπουσα Δρ. Ελένη Τουσίδου για την πολύτιμη βοήθεια και καθοδήγηση της κατά τη διάρκεια αυτής της εργασίας. Οφείλω ευχαριστίες και στα άλλα μέλη της επιτροπής της διπλωματικής εργασίας μου, τον Επιβλέποντα Αναπληρωτή Καθηγήτη κ. Μιχαήλ Βασιλακόπουλο και την Συνεπιβλέπουσα Αναπληρώτρια Καθηγήτρια κ. Παναγιώτα Τσομπανοπούλου για τις συμβουλές τους και όλους τους καθηγητές που είχα καθ' όλη την διάρκεια της πορείας μου, για τις γνώσεις που μου πρόσφεραν. Επίσης, ευχαριστώ τους φίλους που ήταν κοντά μου και με υποστήριξαν. Πάνω από όλα, είμαι ευγνώμων στην οικογένεια μου για την ολόψυχη αγάπη τους και την στήριξη τους όλα αυτά τα χρόνια.

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Κεφάλαιο 1

Εισαγωγή

Τα τελευταία χρόνια με τη ραγδαία ανάπτυξη της τεχνολογίας το ίντερνετ και τα μέσα κοινωνικής δικτύωσης έχουν μπεί στην καθημερινή ζωή των ανθρώπων και αποτελούν αναπόσπαστο κομμάτι τους. Ο όγκος της πληροφορίας και των δεδομένων έχει ακολουθήσει ανάλογη πορεία. Με όλα αυτά τα δεδομένα ο ερευνητής έχει ένα επιπλέον εργαλείο για να αντιμετωπίσει τα άλυτα προβλήματα που τον αποσχολούν. Όμως όλο αυτό το πλήθος πληροφοριών δεν είναι καθόλου χρήσιμο από μόνο του. Χρειάζεται επεξεργασία προκειμένου να ανακαλυφθούν προηγουμένως άγνωστα μοτίβα και να διεξαχθεί χρήσιμη πληροφορία. Αυτό οδήγησε στην ανάπτυξη των κλάδων της επιστήμης που ασχολούνται με την αποδοτικότερη αξιοποίηση των δεδομένων. Ένας από αυτούς τους κλάδους είναι και η **εξόρυξη δεδομένων**. Ο όρος περιγράφει τη λεπτομερή διαδικασία εξέτασης μεγάλων βάσεων δεδομένων για αναζήτηση ενδιαφερόντων προτύπων και σχέσεων. Στην πράξη, η εξόρυξη δεδομένων παρέχει εργαλεία με τα οποία μπορούν να αναλυθούν αυτόματα μεγάλες ποσότητες δεδομένων [10].

Ένα μεγάλο τμήμα της ανάλυσης των δεδομένων έχει σαν σκοπό την δημιουργία μοντέλων που προβλέπουν κάποια συμπεριφορά ή που προτείνουν κάποια επιλογή. Τέτοιες δυνατότητες έχουν πλέον πάρα πολλά συστήματα. Παραδείγματα της καθημερινότητας όπου χρησιμοποιούνται αυτά τα συστήματα είναι η πρόβλεψη του καιρού, το χρηματιστήριο, η προσαρμογή των διαφημίσεων, οι αθλητικές αναλύσεις και επιδόσεις. Με τον σκοπό της πρόβλεψης αξιοποιούνται και στην παρούσα διπλωματική εργασία αυτά τα εργαλεία της εξόρυξης δεδομένων.

1.1 Αντικείμενο της διπλωματικής

Αυτή η διπλωματική εργασία έχει σαν σκοπό να προβλέψει την επιτυχία μίας κινηματογραφικής ταινίας με χρήση αλγορίθμων μηχανικής μάθησης. Με αυτή την πρόβλεψη ο θεατής γνωρίζει εκ των προτέρων πόσο “καλή” θα είναι η ταινία που επρόκειται να παρακολουθήσει, γεγονός που τον βοηθάει στην επιλογή της ταινίας. Συνακόλουθα, η πρόβλεψη αυτή μπορεί να χρησιμοποιηθεί από ένα παραγωγό ταινιών προκειμένου να γνωρίζει κατά πόσο μια ταινία έχει πιθανότητες να του αποφέρει κέρδη και κατ’ επέκταση μπορούν να επωφεληθούν όλοι οι παράγοντες που σχετίζονται με τη βιομηχανία του κινηματογράφου.

Έχοντας έναν μεγάλο όγκο δεδομένων που αφορούν κινηματογραφικές ταινίες γίνεται προ-

σπάθεια πρόβλεψης της βαθμολογίας που θα έχει μια ταινία όταν βγεί στην αγορά. Η βάση δεδομένων έχει δημιουργηθεί από πληροφορίες που έχουν παρθεί από τον διαδικτυακό ιστότοπο κριτικής ταινιών "IMDB"[17] οι οποίες είναι διαθέσιμες στον ιστότοπο της "kaggle"[18]. Στόχος είναι η επεξεργασία της βάσης και η άντληση όσο το δυνατόν περισσότερης πληροφορίας για ακριβέστερη πρόβλεψη. Σημαντικό κομμάτι της μελέτης είναι ο εντοπισμός των μοτίβων που ακολουθούν τα δεδομένα και τελικά η επιλογή των γνωρισμάτων που είναι πιο αντιπροσωπευτικά και καθορίζουν σε μεγαλύτερο βαθμό το αποτέλεσμα.

1.1.1 Συνεισφορά

Η συνεισφορά της διπλωματικής αυτής εργασίας συνοψίζεται ως εξής:

1. Μελετήθηκε το σύστημα του WEKA και της R όπου έγιναν και τα αντίστοιχα πειράματα πάνω στα δεδομένα.

2. Χρησιμοποιήθηκαν εννέα αλγόριθμοι για την πρόβλεψη της βαθμολογίας μίας ταινίας.

3. Αξιολογήθηκε η επίδοση των αλγορίθμων και βρέθηκε ότι μεγαλύτερη απόδοση είχε ο αλγόριθμος "Τυχαία Δάση" (Random Forest) καθώς και η μέθοδος της ανάλυσης των κύριων συνιστωσών (ΑΚΣ ή PCA-Principal Component Analysis).

4. Το πρόβλημα προσεγγίστηκε με τρεις διαφορετικούς τρόπους των οποίων η αποτελεσματικότητα μελετήθηκε σε βάθος και παρουσιάζεται σε επιμέρους κεφάλαια.

1.2 Οργάνωση του τόμου

Η διπλωματική είναι οργανωμένη σε ενότητες. Εργασίες σχετικές με το αντικείμενο της διπλωματικής παρουσιάζονται στο Κεφάλαιο 2. Στο Κεφάλαιο 3 γίνεται η θεωρητική περγραφή των μεθόδων και των τεχνικών που χρησιμοποιήθηκαν σε αυτήν την εργασία. Στο Κεφάλαιο 4 περιγράφεται η πρώτη προσέγγιση της λύσης του προβλήματος όπου, γίνεται ανάλυση της συμπεριφοράς του κάθε γνωρίσματος, οδηγώντας έτσι στην επιλογή των γνωρισμάτων τα οποία τελικά χρησιμοποιούνται στην εκτέλεση των αλγορίθμων. Στο Κεφάλαιο 5 παρουσιάζεται η χρήση της μεθόδου ΑΚΣ για την μείωση των γνωρισμάτων και την χρήση των κύριων συνιστωσών που προκύπτουν στους αλγορίθμους. Το Κεφάλαιο 6 αναφέρεται σε μια πιο συγκερκιμένη μέθοδο προσέγγισης που έχει προκύψει απο την παρατήρηση των δεδομένων, την εμπειρία ενός συνήθη θεατή και την δοκιμή των αποτελεσμάτων. Στο Κεφάλαιο 7 γίνεται η περιγραφή των εργαλείων καθώς και οι τεχνικές λεπτομέρειες της διπλωματικής. Τέλος στο κεφάλαιο 8 παρουσιάζονται τα συμπεράσματα που εξήχθησαν από αυτή τη μελέτη ενώ επίσης γίνεται αναφορά για μελλοντικές επεκτάσεις που είναι δυνατό να πραγματοποιηθούν για μεγαλύτερη κάλυψη επί του συγκεκριμένου θέματος και εφαρμογή των αποτελεσμάτων σε άλλες εργασίες.

Κεφάλαιο 2

Συγγενικές εργασίες

2.1 Εισαγωγή

Στο κεφάλαιο γίνεται αναφορά σε προηγούμενες μελέτες ενώ επίσης παρουσιάζονται επιγραμματικά οι τεχνικές που έχουν εφαρμοστεί σε αυτήν την εργασία. Τέλος επισημαίνονται οι ομοιότητες που υπάρχουν στην διπλωματική και τις μελέτες που αναφέρονται.

2.2 Μοντέλα πρόβλεψης και σύστασης.

Η βιομηχανία κινηματογραφικών ταινιών είναι μια μεγάλη επιχείρηση που αφορά μεγάλο μέρος του πλανήτη. Μια έκθεση από την εταιρεία παρακολούθησης της βιομηχανίας Nash Information Services δείχνει ότι οι πωλήσεις εισιτηρίων αυξάνονται σταθερά τα τελευταία 20 χρόνια. Αυτή η αύξηση συμπίπτει με μια αυξανόμενη ποσότητα δεδομένων σχετικών με τις ταινίες, στις οποίες έχουν στραφεί οι ερευνητές για να βρουν τρόπους να ανακαλύψουν γνωρίσματα που τις χαρακτηρίζουν ως επιτυχημένες ή αποτυχημένες [13] και να εξετάσουν την αλληλεπίδραση μεταξύ των αξιολογήσεων και των εσόδων μετά την κυκλοφορία μιας ταινίας [14].

Το άρθρο [11] επικεντρώνεται σε χαρακτηριστικά που σχετίζονται με τις αξιολογήσεις των ταινιών από τους χρήστες, την ανακάλυψη εάν οι ταινίες μεγάλου προϋπολογισμού είναι πιο δημοφιλείς από τις αντίστοιχες χαμηλού προϋπολογισμού, εάν υπάρχει σχέση μεταξύ ταινιών που παράγονται κατά τη διάρκεια της «χρυσής εποχής» (όπως Citizen Kane, It's A Wonderful Life, κ.λπ.) και αν οι ηθοποιοί είναι πιθανό να συμβάλλουν στην επιτυχία μιας ταινίας. Το άρθρο αναφέρει επίσης τις τεχνικές που χρησιμοποιούνται, δίνοντας την εφαρμογή και τη χρησιμότητά τους. Διαπιστώνει ότι από το IMDB είναι δύσκολο να πραγματοποιηθεί εξόρυξη δεδομένων, λόγω της μορφής των δεδομένων προέλευσης. Κάνει επίσης μερικές ενδιαφέρουσες παρατηρήσεις, όπως ότι ο προϋπολογισμός μιας ταινίας δεν δείχνει πόσο καλή θα είναι, ότι υπάρχει μια πτωτική τάση στην ποιότητα των ταινιών με την πάροδο του χρόνου, καθώς και ότι ο σκηνοθέτης και οι ηθοποιοί που εμπλέκονται σε μια ταινία είναι οι πιο σημαντικοί παράγοντες για την επιτυχία της. Στην παρούσα εργασία ακολουθούνται παρόμοιες τεχνικές στην πρώτη και την τρίτη προσέγγιση για την προεπεξεργασία των δεδομένων καθώς και για την διακριτοποίηση της κλάσης. Επίσης ανακαλύπτονται παρόμοια συμπεράσματα όσον αφορά τους ηθοποιούς και τους σκηνοθέτες, ότι δηλαδή είναι από

τα κύρια γνωρίσματα που καθορίζουν το αποτέλεσμα.

Στο άρθρο [4] παρουσιάζεται η χρήση νευρωνικών δικτύων. Τα νευρωνικά δίκτυα έχουν εφαρμοστεί επιτυχώς σε ένα ευρύ φάσμα εποπτευόμενων και μη εποπτευόμενων εφαρμογών μάθησης. Ωστόσο, οι μέθοδοι των νευρωνικών δικτύων δεν χρησιμοποιούνται συνήθως για εργασίες εξόρυξης δεδομένων, διότι συχνά παράγουν ακατανόητα μοντέλα και απαιτούν πολύ χρόνο εκπαίδευσης. Σε αυτό το άρθρο, περιγράφονται αλγόριθμοι εκμάθησης νευρωνικών δικτύων που είναι σε θέση να παράγουν κατανοητά μοντέλα και που δεν απαιτούν υπερβολικούς χρόνους εκπαίδευσης. Συγκεκριμένα, συζητιούνται δύο κατηγορίες προσεγγίσεων για εξόρυξη δεδομένων με νευρωνικά δίκτυα. Ο πρώτος τύπος προσέγγισης, που συχνά ονομάζεται εξαγωγή κανόνων, περιλαμβάνει την εξαγωγή συμβολικών μοντέλων από εκπαιδευμένα νευρωνικά δίκτυα. Η δεύτερη προσέγγιση είναι η κατανόηση των απλών, εύκολα κατανοητών δικτύων. Υποστηρίζεται ότι, δεδομένης της τρέχουσας τεχνολογίας, οι μέθοδοι νευρωνικών δικτύων αξίζουν μια θέση στις εργαλειοθήκες των ειδικών της εξόρυξης δεδομένων. Σε αυτό το άρθρο χρησιμοποιείται ο αλγόριθμος Back Propagation για την εκπαίδευση του νευρωνικού δικτύου, αλγόριθμος ο οποίος δοκιμάζεται και στην παρούσα εργασία.

Στο άρθρο [6], αναλύεται η επιτυχία της ταινίας από την άποψη τόσο των οικονομικών κερδών όσο και των κριτικών της ταινίας. Η βέλτιστη επιτυχία ορίζεται ως μια ταινία που είναι τόσο κερδοφόρα όσο και πολύ αναγνωρισμένη, ενώ το χειρότερο αποτέλεσμα της περιλαμβάνει οικονομική απώλεια και κακές κριτικές ταυτόχρονα. Τα χαρακτηριστικά γνωρίσματα που είναι εμφανή τόσο στα οικονομικά όσο και στα κρίσιμα αποτελέσματα αναγνωρίζονται σε μια προσπάθεια να αποκαλυφθεί τι κάνει μια «καλή» ταινία «καλή» και μια «κακή» ταινία «κακή», καθώς επίσης και να εξηγήσει ποσοτικά κοινά φαινόμενα στη βιομηχανία του κινηματογράφου. Για να διεξαχθούν τα παραπάνω συμπεράσματα χρησιμοποιείται η ανάλυση κύριων συνιστωσών για να μειωθούν τα γνωρίσματα της βάσης, ενώ ως κατηγοριοποιητές χρησιμοποιούνται οι μηχανές διανυσμάτων υποστήριξης (SVM). Επίσης το άρθρο [8] διερευνά τη χρήση του Principal Component Analysis (PCA) για τη μείωση δεδομένων υψηλής διάστασης και τη βελτίωση της προγνωστικής απόδοσης ορισμένων γνωστών μεθόδων μηχανικής μάθησης.

Κεφάλαιο 3

Θεωρητικό υπόβαθρο

3.1 Εισαγωγή

Στο κεφάλαιο αυτό περιγράφονται οι τεχνικές και οι αλγόριθμοι που χρησιμοποιήθηκαν στην παρούσα διπλωματική εργασία. Γίνεται μια εισαγωγή στην θεωρία της εξόρυξης δεδομένων και παρουσιάζονται συνοπτικά τα κύρια στοιχεία των αλγορίθμων.

3.2 Μέθοδοι εξόρυξης δεδομένων

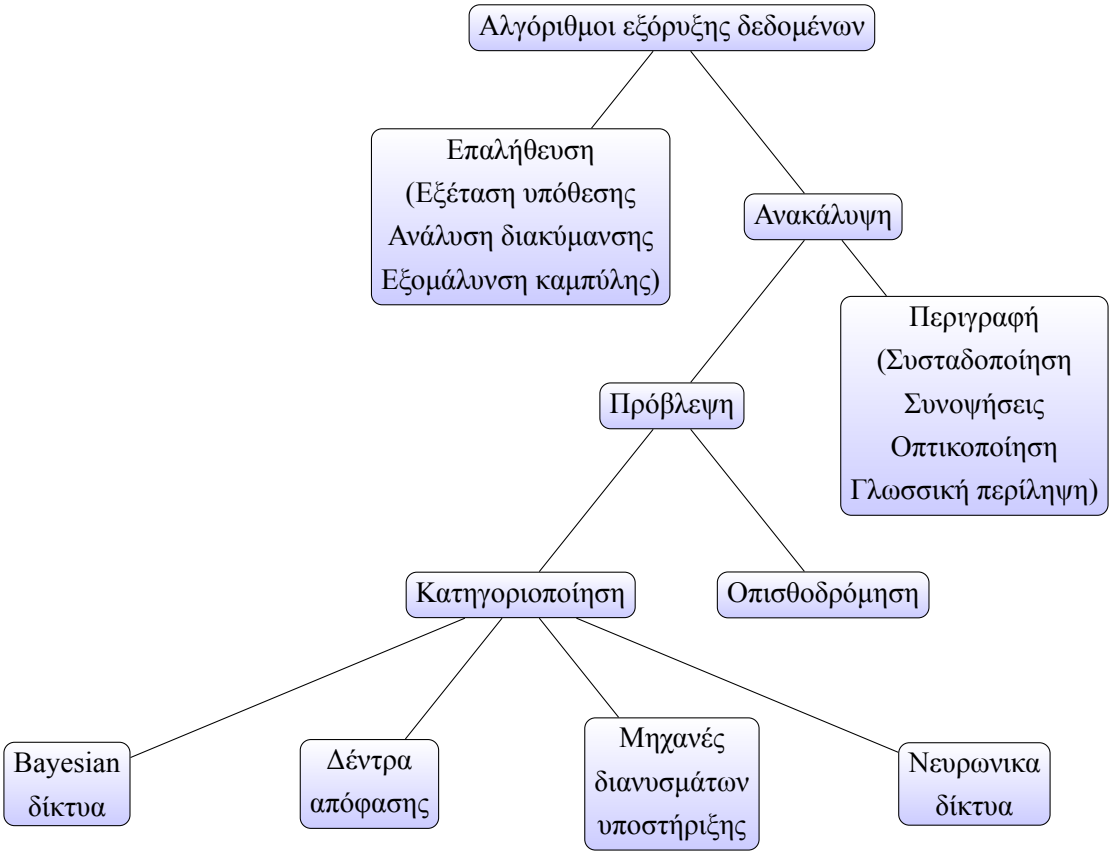
Είναι χρήσιμο να γίνει διάκριση μεταξύ δύο κύριων τύπων εξόρυξης δεδομένων: προσανατολισμένος στην επαλήθευση (verification-oriented) (το σύστημα επαληθεύει την υπόθεση του χρήστη) και προσανατολισμένος στην ανακάλυψη (discovery-oriented) (το σύστημα βρίσκει αυτόνομους νέους κανόνες και μοτίβα). Το σχήμα 3.1 απεικονίζει αυτήν την ταξινόμηση [10].

Κάθε τύπος έχει τη δική του μεθοδολογία. Οι μέθοδοι **περιγραφής**, οι οποίες προσδιορίζουν αυτόματα μοτίβα στα δεδομένα, περιλαμβάνουν μεθόδους πρόβλεψης και περιγραφής. Οι μέθοδοι περιγραφής επικεντρώνονται στην κατανόηση του τρόπου λειτουργίας των υποκείμενων δεδομένων, ενώ οι μέθοδοι με βάση την πρόβλεψη στοχεύουν στη δημιουργία ενός μοντέλου συμπεριφοράς για τη λήψη νέων και αόρατων δειγμάτων και για την πρόβλεψη τιμών μιας ή περισσότερων μεταβλητών που σχετίζονται με το δείγμα.

Οι περισσότερες από τις τεχνικές με γνώμονα την **ανακάλυψη** βασίζονται στην επαγωγική μάθηση, όπου ένα μοντέλο κατασκευάζεται ρητά ή έμμεσα με γενίκευση από επαρκή αριθμό παραδειγμάτων εκπαίδευσης. Η υποκείμενη υπόθεση της επαγωγικής προσέγγισης είναι ότι το εκπαιδευμένο μοντέλο εφαρμόζεται σε μελλοντικά αόρατα παραδείγματα.

Οι **μέθοδοι επαλήθευσης**, από την άλλη πλευρά, αξιολογούν μια υπόθεση που προτείνεται από μια εξωτερική πηγή (όπως ένας ειδικός κ.λπ.). Αυτές οι μέθοδοι περιλαμβάνουν τις πιο συνηθισμένες μεθόδους παραδοσιακών στατιστικών, όπως την εξομάλυνση καμπύλης και την ανάλυση διακύμανσης. Αυτές οι μέθοδοι συνδέονται λιγότερο με την εξόρυξη δεδομένων από τις αντίστοιχες με την ανακάλυψη, επειδή τα περισσότερα προβλήματα εξόρυξης δεδομένων αφορούν την επιλογή μιας υπόθεσης (από ένα σύνολο υποθέσεων) αντί να δοκιμάσουν μια γνωστή. Το επίκεντρο των παραδοσιακών στατιστικών μεθόδων είναι συνήθως στην εκτίμηση μοντέλου σε αντίθεση

με έναν από τους κύριους στόχους της εξόρυξης δεδομένων, την ταυτοποίηση μοντέλου.



Σχήμα 3.1: Ταξινόμηση των μεθόδων εξόρυξης δεδομένων.

3.3 Κατηγοριοποίηση

Η κατηγοριοποίηση, όπως περιγράφεται στο βιβλίο [12], η οποία είναι η εργασία εκχώρησης αντικειμένων σε μία από τις διάφορες προκαθορισμένες κατηγορίες, είναι ένα ευρέως γνωστό πρόβλημα που περιλαμβάνει πολλές και διαφορετικές εφαρμογές. Παραδείγματα αποτελούν η ανίχνευση ανεπιθύμητων ηλεκτρονικών μηνυμάτων (spam), βασιζόμενη στην επικεφαλίδα και το περιεχόμενο του μηνύματος, η κατηγοριοποίηση κυττάρων ως κακοήθη ή καλοήθη βασιζόμενη στα αποτελέσματα εξετάσεων MRI (μαγνητική τομογραφία), και η κατηγοριοποίηση των γαλαξιών με βάση το σχήμα τους.

Η κατηγοριοποίηση είναι η εργασία εκμάθησης μιας συνάρτησης-στόχου (target function) f, η οποία απεικονίζει κάθε σύνολο χαρακτηριστικών x σε μια από τις καθορισμένες ετικέτες κατηγορίας. Η συνάρτηση-στόχος, είναι γνωστή ανεπίσημα και ως μοντέλο κατηγοριοποίησης (classification model).

Μια τεχνική κατηγοριοποίησης (ή διαφορετικά ένας κατηγοριοποιητής), είναι μια συστηματική προσέγγιση για την δημιουργία μοντέλων κατηγοριοποίησης από ένα σύνολο δεδομένων εισόδου. Παραδείγματα αποτελούν οι κατηγοριοποιητές δένδρων απόφασης, οι βασισμένοι σε κανόνες κατηγοριοποιητές, τα νευρωνικά δίκτυα, οι μηχανές διανυσμάτων υποστήριξης, και οι

απλοϊκοί κατηγοριοποιητές Bayes. Κάθε τεχνική χρησιμοποιεί έναν αλγόριθμο μάθησης (learnig algorithm), για να εντοπίσει ένα μοντέλο που ταιριάζει καλύτερα στη σχέση μεταξύ του συνόλου χαρακτηριστικών και της ετικέτας κατηγορίας των δεδομένων εισόδου. Το μοντέλο που παράγεται από τον αλγόριθμο μαθησης πρέπει τόσο να ταιριάζει καλά στα δεδομένα εισόδου όσο και να προβλέπει σωστά τις ετικέτες κατηγορίας των εγγραφών, τις οποίες δεν γνωρίζει. Επομένως, ένας βασικός στόχος του αλγορίθμου μάθησης είναι να δημιουργήσει μοντέλα που έχουν την ικανότητα γενίκευσης, δηλαδή μοντέλα που προβλέπουν με ακρίβεια τις ετικέτες κατηγοριών για εγγραφές, οι οποίες προηγουμένως ήταν άγνωστες.

Πρώτον, πρέπει να δοθεί ένα σύνολο εκπαίδευσης (training set) που να αποτελείται από εγγραφές των οποίων οι ετικέτες κατηγορίας είναι γνωστές. Το σύνολο εκπαίδευσης χρησιμοποιείται για να κατασκευαστεί ένα μοντέλο κατηγοριοποίησης, το οποίο με τη σειρά του εφαρμόζεται σε ένα σύνολο ελέγχου (test set), που αποτελείται από εγγραφές με άγνωστες ετικέτες.

Η εκτίμηση της απόδοσης ενός μοντέλου κατηγοριοποίησης, βασίζεται στο πλήθος των εγγραφών ελέγχου που έχουν προβλεφθεί σωστά και λανθασμένα από το μοντέλο. Αυτές οι μετρήσεις τοποθετούνται σε έναν πίνακα που είναι γνωστός ως μήτρα σύγχυσης (confusion matrix). Ο Πίνακας 3.1 παριστάνει τη μήτρα σύγχυσης για ένα πρόβλημα δυαδικής κατηγοριοποίησης. Κάθε κατηγοριοποίηση f_{ij}, αυτού του πίνακα δηλώνει το πλήθος των εγγραφών της κατηγορίας i, που προβλέφθηκε ότι ανήκει στην κατηγορία j. Για παράδειγμα, η καταχώρηση f_{01}, είναι το πλήθος των εγγραφών από την κατηγορία 0 που λανθασμένα προβλέφθηκε ότι ανήκει στην κατηγορία 1. Βασιζόμενοι στις καταχωρήσεις της μήτρας σύγχυσης, το συνολικό πλήθος των σωστών προβλέψεων που κάνει το μοντέλο είναι $(f_{11}+f_{00})$ και το συνολικό πλήθος των λανθασμένων προβλέψεων είναι $(f_{10}+f_{01})$.

Παρά το γεγονός ότι η μήτρα σύγχυσης παρέχει τις πληροφορίες που απαιτούνται για να καθοριστεί το πόσο καλά λειτουργεί ένα μοντέλο κατηγοριοποίησης, η σύνοψη των πληροφοριών σε ένα απλό νούμερο κάνει πιο εύκολη τη σύγκρουση της απόδοσης διαφορετικών μοντέλων. Αυτό μπορεί να γίνει χρησιμοποιώντας ένα μέτρο απόδοσης όπως είναι η ακρίβεια (accuracy), η οποία ορίζεται ώς ακολούθως:

$$Accuracy = \frac{NC}{T} = \frac{f_{11}+f_{00}}{f_{11}+f_{10}+f_{01}+f_{00}} \tag{3.1}$$

όπου NC είναι το πλήθος των ορθών προβλέψεων και T το συνολικό πλήθος των προβλέψεων.

Ισοδύναμα, η απόδοση ενός μοντέλου που μπορεί να εκφραστεί με βάση το ρυθμό σφάλματος (error rate), ο οποίος δίνεται από την ακόλουθη εξίσωση:

$$Error\ rate = \frac{NW}{T} = \frac{f_{10}+f_{01}}{f_{11}+f_{10}+f_{01}+f_{00}} \tag{3.2}$$

όπου NW είναι το πλήθος των λανθασμένων προβλέψεων και T το συνολικό πλήθος των προβλέψεων.

Οι πιο πολλοί αλγόριθμοι κατηγοριοποίησης αναζητούν μοντέλα, τα οποία επιτυγχάνουν τη μεγαλύτερη ακρίβεια, ή ισοδύναμα, το μικρότερο ρυθμό σφάλματος, όταν εφαρμόζεται σε ένα συνολο ελέγχου

Πίνακας 3.1: Μήτρα σύγχυσης ενός δυαδικού προβλήματος

		Προβλεφθείσα	κατηγορία
		Κατηγορία=1	Κατηγορία=0
Πραγματική	Κατηγορία=1	f_{11}	f_{10}
Κατηγορία	Κατηγορία=0	f_{01}	f_{00}

3.4 Κατάρα των πολλών διαστάσεων

Αν η είσοδος αποτελείται από μεγάλο αριθμό διαστάσεων [10] (δηλαδή, ο αριθμός των γνωρισμάτων) τότε αυξάνεται το μέγεθος του χώρου αναζήτησης με εκθετικό τρόπο και έτσι αυξάνεται η πιθανότητα ο κατηγοριοποιητής να βρει λανθασμένα μοτίβα που δεν ισχύουν γενικά. Είναι ευρέως γνωστό ότι ο απαιτούμενος αριθμός δειγμάτων για ταξινόμηση αυξάνεται ως συνάρτηση των διαστάσεων. Έχει αποδειχθεί (Fukunaga (1990)) ότι ο απαιτούμενος αριθμός δειγμάτων για εκπαίδευση σχετίζεται γραμμικά με τη διάσταση ενός γραμμικού ταξινομητή και με το τετράγωνο των διαστάσεων για έναν τετραγωνικό ταξινομητή. Όσον αφορά τους μη παραμετρικούς ταξινομητές όπως τα δέντρα αποφάσεων, η κατάσταση είναι ακόμη πιο σοβαρή. Έχει εκτιμηθεί ότι καθώς ο αριθμός των διαστάσεων αυξάνεται, το μέγεθος του δείγματος πρέπει να αυξηθεί εκθετικά προκειμένου να έχει μια αποτελεσματική εκτίμηση.

3.5 Επιλογή γνωρισμάτων

Όταν η βάση δεδομένων αποτελείται από μεγάλο πλήθος γνωρισμάτων, δηλαδή από πολλές διαστάσεις, αποτελεί σοβαρό εμπόδιο στην αποτελεσματικότητα των περισσότερων επαγωγικών αλγορίθμων [10]. Η επιλογή γνωρισμάτων είναι ένας αποτελεσματικός τρόπος αντιμετώπισης του προβλήματος αυτού. Ο στόχος της επιλογής γνωρισμάτων είναι να προσδιοριστούν εκείνα τα χαρακτηριστικά στο σύνολο δεδομένων που είναι σημαντικά και να απορριφθούν άλλα που πιθανόν να είναι άσχετα και περιττά. Δεδομένου ότι η επιλογή γνωρισμάτων μειώνει τη διάσταση των δεδομένων, οι αλγόριθμοι εξόρυξης δεδομένων μπορούν να λειτουργούν γρηγορότερα και πιο αποτελεσματικά. Ο λόγος για τη βελτιωμένη απόδοση οφείλεται κυρίως σε μια πιο συμπαγή, εύκολα ερμηνευμένη αναπαράσταση της έννοιας του στόχου.

3.6 Κατηγοριοποιητές του Bayes

Οι Bayesian κατηγοριοποιητές [12] βασίζονται στο θεώρημα του Bayes (θεώρημα 3.1) [12]. Πρόκειται για στατιστικούς ταξινομητές οι οποίοι κατηγοριοποιούν με βάση την πιθανότητα μια δεδομένη εγγραφή να ανήκει σε μια συγκεκριμένη κλάση.

Θεώρημα 3.1 (Θεώρημα Bayes)**.** *Έστω ότι τα X και Y είναι ένα ζεύγος από τυχαίες μεταβλητές. Η από κοινού πιθανότητα, P(X = x, Y = y) αναφέρεται στην πιθανότητα η μεταβλητή X να λαβει την*

τιμή x και η μεταβλητή Y να λάβει την τιμή y. Η υπό συνθήκη πιθανότητα είναι η πιθανότητα μιας τυχαίας μεταβλητής να λάβει μια συγκεκριμένη τιμή δεδομένου ότι το αποτέλεσμα για μια άλλη τυχαία μεταβλητή είναι γνωστό. Για παράδειγμα, η υπό συνθήκη πιθανότητα P(Y = y | X = x) αναφέρεται στην πιθανότητα η μεταβλητή Y να λάβει την τιμή y, δοθέντος του γεγονότος ότι η μεταβλητή X έχει παρατηρηθεί ότι λαμβάνει την τιμή x. Οι από κοινού και υπό συνθήκη πιθανότητες των X και Y σχετίζονται με τον ακόλουθο τρόπο:

$$P(X, Y) = P(Y \mid X) \times P(X) = P(X \mid Y) \times P(Y) \tag{3.3}$$

Η αλλαγή στη διάταξη των δύο τελευταίων εκφράσεων της εξίσωσης 3.3 οδηγεί στην ακόλουθη σχέση, η οποία είναι γνωστή ως το θεώρημα του Bayes:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}. \tag{3.4}$$

Στην παρούσα εργασία έχει γίνει χρήση του αλγορίθμου Naive Bayes του περιβάλλοντος της Weka. Πιο αναλυτική περιγραφή του αλγορίθμου γίνεται στο άρθρο [9].

3.7 Δέντρα απόφασης

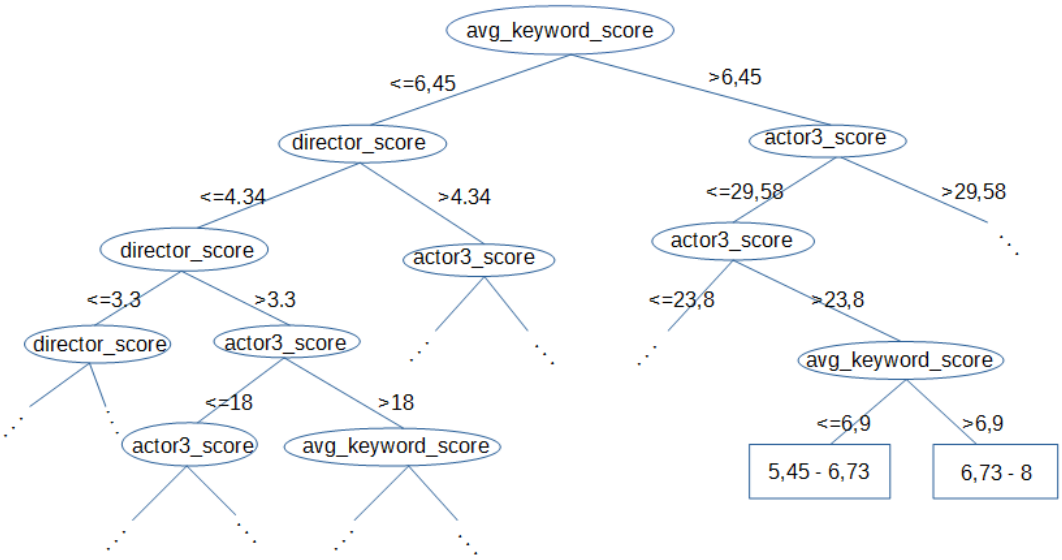
Τα δέντρα αποφάσεων είναι απλές αλλά επιτυχημένες τεχνικές για την πρόβλεψη και εξήγηση της σχέσης μεταξύ ορισμένων μετρήσεων για ένα στοιχείο και της τιμής στόχου του. Εκτός από τη χρήση τους στην εξόρυξη δεδομένων, τα δέντρα αποφάσεων, τα οποία αρχικά προέρχονταν από τη λογική, τη διαχείριση και τα στατιστικά στοιχεία, είναι σήμερα εξαιρετικά αποτελεσματικά εργαλεία σε άλλους τομείς όπως η εξόρυξη κειμένων, η εξαγωγή πληροφοριών, η μηχανική εκμάθηση και η αναγνώριση προτύπων. Τα δέντρα απόφασης προσφέρουν πολλά οφέλη:

- Ευελιξία για μια μεγάλη ποικιλία εργασιών εξόρυξης δεδομένων, όπως κατηγοριοποίηση (classification), οπισθοδρόμηση (regression), συσταδοποίηση (clustering) και επιλογή χαρακτηριστικών (feature selection).
- Ευελιξία στο χειρισμό μιας ποικιλίας δεδομένων εισόδου: ονομαστικά, αριθμητικά και κείμενο.
- Προσαρμοστικότητα στην επεξεργασία συνόλων δεδομένων που ενδέχεται να έχουν σφάλματα ή να λείπουν τιμές.
- Υψηλή προγνωστική απόδοση για μία σχετικά μικρή υπολογιστική προσπάθεια.
- Χρήσιμα για μεγάλα σύνολα δεδομένων.

Το δέντρο απόφασης [12] είναι μία ιεραρχική δομή που αποτελείται από κόμβους και κατευθυνόμενες ακμές. Οι τύποι κόμβων που συναντιούνται σε αυτή τη δομή είναι:

- ο **κόμβος ρίζα** ο οποίος δεν έχει εισερχόμενες ακμές αλλά μπορεί να έχει εξερχόμενες.

- ο **εσωτερικοί κόμβοι** οι οποίοι έχουν μία εισερχόμενη ακμή και δύο ή περισσότερες εξερχόμενες.
- τα **φύλλα** ή **τερματικοί κόμβοι** οι οποίοι έχουν μία εισερχόμενη ακμή και καμία εξερχόμενη.



Σχήμα 3.2: Παράδειγμα δέντρου απόφασης, αποτέλεσμα από τον J48 αλγόριθμο.

Κάθε εσωτερικός κόμβος αντιπροσωπεύει μια δοκιμή σε ένα χαρακτηριστικό, κάθε κόμβος φύλλου αντιπροσωπεύει μια ετικέτα κλάσης (η απόφαση λαμβάνεται αφού υπολογιστούν όλα τα χαρακτηριστικά) και τα κλαδιά αντιπροσωπεύουν συζεύξεις χαρακτηριστικών που οδηγούν σε αυτές τις ετικέτες κλάσης. Οι διαδρομές από ρίζα σε φύλλο αντιπροσωπεύουν κανόνες ταξινόμησης. Σε αυτή την εργασία έχουν χρησιμοποιηθεί οι αλγόριθμοι J48 και Τυχαία Δάση (Random Forest) που παρουσιάζονται περιληπτικά παρακάτω. Ένα παράδειγμα δέντρου απόφασης φαίνεται στο σχήμα 3.2.

3.7.1 J48 αλγόριθμος

Ο αλγόριθμος J48 χρησιμοποιείται για τη δημιουργία ενός κλαδευμένου ή μη κλαδευμένου δέντρου αποφάσεων C4.5. Στην παρούσα εργασία χρησιμοποιούνται κλαδεμένα δέντρα αποφάσεων. Το κλάδεμα βοηθάει στην αποκοπή ορισμένων κλαδιών του αρχικού δέντρου, με τρόπο που να βελτιώνει την ικανότητα γενίκευσης του δέντρου απόφασης. Επίσης με το κλάδεμα του δέντρου απογφεύγεται το φαινόμενο της υπερπροσαρμογής (overfitting).

Ο αλγόριθμος C4.5 [12], μια εξέλιξη του ID3, χρησιμοποιεί την αναλογία κέρδους (gain ratio) ως κριτήριο διαχωρισμού.

$$Gain\ ratio = \frac{\Delta_{info}}{Split\ Info} \tag{3.5}$$

όπου $Split\ Info = -\sum_{i=1}^{k} P(v_i) log_2 P(v_i)$ και k, είναι ο συνολικός αριθμός των διαχωρισμών. Αν κάθε τιμή χαρακτηριστικού έχει το ίδιο πλήθος εγγραφών, τότε $\forall i : P(u_i) = \frac{1}{k}$. Το κέρδος Δ,

δίνεται από την εξίσωση 3.6:

$$\Delta = I(parent) - \sum_{j=1}^{k} \frac{N(u_j)}{N} I(u_j) \tag{3.6}$$

όπου $I(.)$, είναι το μέτρο ανομοιογένειας ενός δεδομένου κόμβου, N είναι το συνολικό πλήθος των εγγραφών στον κόμβο-γονέα, k είναι το πλήθος των τιμών ενός χαρακτηριστικού και $N(u_j)$ είναι το πλήθος των εγγραφών που σχετίζονται με τον κόμβο-παιδί u_x

Ο διαχωρισμός σταματά όταν ο αριθμός των περιπτώσεων που θα διαχωριστούν είναι κάτω από ένα ορισμένο κατώφλι. Το κλάδεμα με βάση το σφάλμα πραγματοποιείται μετά τη φάση ανάπτυξης. Ο C4.5 μπορεί να χειριστεί αριθμητικά χαρακτηριστικά. Μπορεί επίσης να ανταπεξέλθει σε ένα σύνολο εκπαίδευσης που ενσωματώνει τιμές που λείπουν χρησιμοποιώντας διορθωμένα κριτήρια αναλογίας κέρδους.

3.7.2 Τυχαία δάση (random forest)

Ένα τυχαίο σύνολο δασών χρησιμοποιεί ένα μεγάλο αριθμό μεμονωμένων δέντρων αποφάσεων που δεν έχουν κλαδευτεί και που δημιουργούνται με τυχαιοποίηση του διαχωρισμού σε κάθε κόμβο του δέντρου αποφάσεων [10]. Κάθε δέντρο είναι πιθανό να είναι λιγότερο ακριβές από ένα δέντρο που έχει δημιουργηθεί με τον ακριβή διαχωρισμό. Όμως, συνδυάζοντας πολλά από αυτά τα «κατά προσέγγιση» δέντρα σε ένα σύνολο, μπορούμε να βελτιώσουμε την ακρίβεια, παράγοντας συχνά καλύτερα αποτελέσματα από ένα μεμονωμένο δέντρο με ακριβείς διαχωρισμούς. Ο ακριβής αλγόριθμος περιγράφεται στην πηγή [2].

3.8 Νευρωνικά δίκτυα

Η έρευνα σχετικά με τα τεχνητά νευρωνικά δίκτυα είναι εμπνευσμένη από τη δομή και τη λειτουργία του εγκεφάλου [20]. Βασικό δομικό στοιχείο του εγκεφάλου είναι οι νευρώνες, δηλαδή τα νευρικά κύτταρα τα οποία δημιουργούν ένα πυκνό δίκτυο επικοινωνίας μεταξύ τους. Κίνητρο για τη μελέτη του νευρώνα και των νευρωνικών δικτύων είναι η ελπίδα ανακαλύψης ενός νέου υπολογιστικού μοντέλου βασισμένου σε μια διακτυακή δομή παρόμοια με αυτή του εγκεφάλου. Αυτή η καινούργια υπολογιστική πλατφόρμα - γνωστη ως Connectionist Model- θα είναι πιο κατάλληλη για ανάπτυξη ευφυών αλγορίθμων και γενικότερα διαδικασιών σχετιζόμενων με τη νοημοσύνη, όπως η μάθηση, η μνήμη, η γενίκευση, η ομαδοποίηση προτύπων, κλπ.

Τα τεχνητά νευρωνικά δίκτυα χρησιμοποιούν πολύ απλοποιημένα μοντέλα νευρώνων τέτοια ώστε να διατηρούν μόνο τα πολύ αδρά χαρακτηριστικά των λεπτομερών μοντέλων που χρησιμοποιούνται στη νευρολογία. Θα έλεγε κανείς ότι τα συνήθη τεχνητά νευρωνικά μοντέλα έχουν ελάχιστη σχέση με τα βιολογικά νευρωνικά συστήματα. Ωστόσο πιστεύεται ότι οι λεπτομέρειες δεν έχουν ιδιαίτερη σημασία στην κατανόηση της ευφυούς συμπεριφοράς των βιολογικών νευρωνικών συστημάτων. Ακόμη και αυτά τα απλά μοντέλα νευρώνων μπορούν να δημιουργήσουν ιδιαιτέρως ενδιαφέροντα δίκτυα αρκεί να πληρούν δύο βασικά χαρακτηριστικά:

1. Οι νευρώνες να έχουν ρυθμιζόμενες παραμέτρους ώστε να διευκολύνεται η διαδικασία της μάθησης -ιδιότητα γνωστή ως **πλαστικότητα** των νευρώνων.

2. Το δίκτυο να αποτελείται από μεγάλο πλήθος νευρώνων ώστε να επιτυγχάνεται ο **παραλληλισμός** της επεξεργασίας και η **κατανομή** της πληροφορίας.

Η πρόκληση που αντιμετωπίζει η θεωρία των τεχνητών νευρωνικών δικτύων είναι η εύρεση κατάλληλων αλγορίθμων εκπαίδευσης των δικτύων και ανάκλησης της πληροφορίας που αυτά περιέχουν, έτσι ώστε να προσομοιάζονται ευφυείς διαδικασίες, όπως αυτές που αναφέρθηκαν παραπάνω. Για την επίτευξη αυτού του στόχου απαιτείται ο ορισμός του κατάλληλου περιβάλλοντος εκπαίδευσης, πχ. αν το δίκτυο θα εκπαιδεύεται με επίβλεψη, δηλαδή με τη χρήση κάποιων δεδομένων οδηγών-δασκάλων, ή αν το δίκτυο θα αφήνεται μόνο του να αυτο-οργανωθεί και με ποιο συγκεκριμένο κριτήριο και στόχο. Στην παρούσα εργασία χρησιμοποιείται ο αλγόριθμος Multilayer Perceptron που είναι υλοποιημένος στην Weka.

3.9 Κατηγοριοποιητής κανόνων

Ένας κατηγοριοποιητής κανόνων [12] είναι μια τεχνική για την κατηγοριοποίηση εγγραφών χρησιμοποιώντας μία συλλογή από κανόνες "if...then...else". Οι κανόνες του μοντέλου αναπαριστώνται σε μια διαζευκτική κανονική μορφή, $R = (r_1 \vee r_2 \vee ...r_k)$, όπου το R είναι γνωστό ως το σύνολο κανόνων (rule set) και τα r_i είναι οι κανόνες κατηγοριοποίησης ή διαζεύξεις. Κάθε κανόνας κατηγοριοποίησης μπορεί να εκφραστεί με τον ακόλουθο τύπο:

$$r_i : (Condition) \longrightarrow y_i. \tag{3.7}$$

Το αριστερό μέρος του κανόνα ονομάζεται προηγούμενο κανόνα (rule antecedent) ή συνθήκη εισόδου (precondition). Περιλαμβάνει μια σύζευξη των ελέγχων των χαρακτηριστικών:

$$Condition_i : (A_1 opv_1) \wedge (A_2 opv_2) \wedge ...(A_k opv_k), \tag{3.8}$$

όπου (A_j, v_j) είναι ένα ζεύγος χαρακτηριστικού-τιμής και op είναι ο λογικός τελεστής που επιλέγεται από το σύνολο $\{=, \neq, <, >, \geq, \leq\}$. Κάθε έλεγχος χαρακτηριστικού $(A_j opv_j)$ είναι γνωστός ως σύζευξη. Το δεξί μέρος του κανόνα ονομάζεται επακόλουθο κανόνα (rule consequent), και περιέχει την προβλεπόμενη κατηγορία y_i. Ένας κανόνας r καλύπτει μία εγγραφή x μόνον όταν η συνθήκη εισόδου του r ταιριάζει στα χαρακτηριστικά του x. Ο κανόνας r λέμε ότι πυροδοτείται ή ενεργοποιείται όταν καλύπτει μια δοθείσα εγγραφή.

3.10 Κατηγοριοποιητής πλησιέστερου γείτονα

Ένας κατηγοριοποιητής πλησιέστερου γείτονα [12] αναπαριστά κάθε δείγμα ως ένα σημείο δεδομένων σε ένα χώρο d διαστάσεων, όπου d είναι το πλήθος των χαρακτηριστικών. Δοθέντος ενός δείγματος ελέγχου, υπολογίζεται η εγγύτητά του σε σχέση με τα υπόλοιπα σημεία δεδομένων του συνόλου εκπαίδευσης, χρησιμοποιώντας ένα από τα μέτρα εγγύτητας. Οι k-πλησιέστεροι γείτονες ενός δοθέντος δείγματος z αναφέρονται στα k σημεία που είναι πλησιέστερα στο z.

3.11 Ανάλυση κύριων συνιστωσών (ΑΚΣ - PCA)

Ο στόχος της Ανάλυσης Κύριων Συνιστωσών είναι να βρεθεί ένα νέο σύνολο διαστάσεων (χαρακτηριστικών) που αντικατοπτρίζει καλύτερα τη μεταβλητότητα των δεδομένων [12]. Ειδικότερα, η πρώτη διάσταση επιλέγεται ώστε να λάβει όσο το δυνατόν μεγαλύτερη μεταβλητότητα. Η δεύτερη διάσταση είναι ορθογώνια ως προς την πρώτη, και υπό αυτόν τον περιορισμό, λαμβάνει όσο το δυνατόν μεγαλύτερο μέρος από την υπόλοιπη μεταβλητότητα και ούτε καθεξής.

Μαθηματικό υπόβαθρο

Οι στατιστικοί συνοψίζουν τη μεταβλητότητα ενός συνόλου από πολυμεταβλητά δεδομένα, δηλαδή δεδομένα που έχουν πολλά συνεχή χαρακτηριστικά, υπολογίζοντας τη μητρα συνδιακύμανσης S των δεδομένων.

Ορισμός

Δοθείσης μια μήτρας D διαστάσεων $m * n$, της οποίας οι m γραμμές είναι τα αντικείμενα δεδομένων και οι n στήλες είναι τα χαρακτηριστικά, η μήτρα συνδιακύμανσης της D είναι μια μήτρα S, της οποίας οι καταχωρήσεις s_{ij} ορίζονται ως ακολούθως:

$$s_{ij} = covariance(d_{*i}, d_{*j}). \tag{3.9}$$

Λεκτικά, το στοιχείο s_{ij} είναι η συνδιακύμανση των χαρακτηριστικών (στηλών) i και j των δεδομένων.

Ορισμός

Η συνδιακύμανση ή $C_{ov}(X, Y)$ είναι μέτρο του βαθμού συσχετίσεως δύο μεταβλητων X και Y.

$$C_{ov}(X, Y) = E[(X - E(X))(Y - E(Y))] \tag{3.10}$$

Αν $i = j$, δηλαδή τα χαρακτηριστικά είναι ίδια, τότε η συνδιακύμανση είναι η διακύμανση του χαρακτηριστικού. Αν η μήτρα δεδομένων D, έχει υποστεί προεπεξεργασία, ώστε ο μέσος κάθε χαρακτηριστικού να είναι 0, τότε $S = D^T D$. Ένας στόχος της ΑΚΣ είναι να παράγει ένα μετασχηματισμό των δεδομένων, ο οποίος να ικανοποιεί τις ακόλουθες ιδιότητες:

1. Κάθε ζεύγος νέων χαρακτηριστικών έχει συνδιακύμανση 0 (για ξεχωριστά χαρακτηριστικά).

2. Τα χαρακτηριστικά ταξινομούνται σε σχέση με το ποσοστό της διακύμανσης των δεδομένων που λαμβάνουν.

3. Το πρώτο χαρακτηριστικό λαμβάνει όσο το δυνατόν περισσότερη μεταβλητότητα των δεδομένων.

4. Κάθε διαδοχικό χαρακτηριστικό, υπό την απαίτηση ορθογωνικότητας, λαμβάνει όσο το δυνατόν περισσότερη από την υπολειπόμενη διακύμανση.

Ένας μετασχηματισμός των δεδομένων που διαθέτει αυτές τις ιδιότητες μπορεί να ληφθεί χρησιμοποιώντας την ανάλυση ιδιοτιμών της μήτρας συνδιακύμανσης. Έστω ότι $\lambda_1, ..., \lambda_n$ είναι οι ιδιοτιμές της μήτρας S. Όλες οι ιδιοτιμές είναι μη αρνητικές και μπορούν να ταξινομηθούν με τέτοιον τρόπο, ώστε $\lambda_1 \geq \lambda_2 \geq ...\lambda_{m-1} \geq \lambda_m$. (Οι μήτρες συνδιακύμανσης αποτελούν παράδειγμα

αυτών που ονομάζονται **θετικά ημιορισμένες μήτρες**, οι οποίες, μεταξύ των άλλων ιδιοτήτων, έχουν μη αρνητικές ιδιοτιμές.) Έστω ότι $U = [u_1, ..., u_n]$ είναι η μήτρα των ιδιοτιμών της S. Αυτά τα ιδιοδιανύσματα ταξινομούνται ώστε το *i*-οστό ιδιοδιάνυσμα να αντιστοιχεί στην *i*-οστή ιδιοτιμή. Τέλος, ας υποτεθεί ότι η μήτρα δεδομένων D έχει υποστεί προεπεξεργασία ώστε ο μέσος κάθε χαρακτηριστικού (στήλη) να είναι 0. Μπορούμε να διατυπώσουμε τις ακόλουθες προτάσεις.

- Η μήτρα δεδομένων $D^{'} = DU$ είναι το σύνολο των μετασχηματισμένων δεδομένων που ικανοποιεί τις συνθήκες που τέθηκαν παραπάνω.
- Κάθε νέο χαρακτηριστικό είναι ένας γραμμικός συνδυασμός των αρχικών χαρακτησριστικών. Ειδικότερα, τα βάρη του γραμμικού συνδυασμού για το *i*-οστό χαρακτηριστικό είναι οι συνιστώσες του *i*-οστού ιδιοδιανύσματος. Αυτό προκύπτει από το γεγονός ότι η *j*-οστή στήλη της μήτρας $D^{'}$ δίνεται από τη Du_j και τον ορισμό του πολλαπλασιασμού μήτρας-διανύσματος.
- Η διακύμανση του νέου χαρακτηριστικού *i* είναι λ_i.
- Το άθροισμα της διακύμανσης των αρχικών χαρακτηριστικών, είναι ίσο με το άθροισμα της διακύμανσης των νέων χαρακτηριστικών.
- Τα νέα χαρακτηριστικά ονομάζονται **κύριες συνιστώσες**, δηλαδή, το πρώτο χαρακτηριστικό είναι η πρώτη κύρια συνιστώσα, το δεύτερο νέο χαρακτηριστικό είναι η δεύτερη κύρια συνιστώσα και ούτε καθεξής.

Το ιδιοδιάνυσμα που σχετίζεται με τη μεγαλύτερη ιδιοτιμή δείχνει την κατεύθυνση στην οποία τα δεδομένα έχουν τη μεγαλύτερη διακύμανση. Με άλλα λόγια, αν όλα τα διανύσματα δεδομένων προβληθούν στη γραμμή που ορίζεται από αυτό το διάστημα, οι τιμές που θα προκύψουν θα έχουν τη μέγιστη διακύμανση σε σχέση με όλες τις άλλες πιθανές κατευθύνσεις. Το ιδιοδιάνυσμα που σχετίζεται με τη δεύτερη μεγαλύτερη ιδιοτιμή είναι η κατεύθυνση (ορθογώνια σε εκείνη του πρώτου ιδιοδιανύσματος), στην οποία τα δεδομένα έχουν τη μεγαλύτερη υπολοιπόμενη διακύμανση.

Τα ιδιοδιανύσματα της μήτρας S ορίζουν ένα νέο σύνολο από άξονες. Πράγματι, η ΑΚΣ μπορεί να θεωρηθεί ως μια περιστροφή των αρχικών συντεταγμένων αξόνων σε ένα νέο σύνολο από άξονες, οι οποίοι ευθυγραμμίζονται με τη μεταβλητότητα των δεδομένων. Η συνολική μεταβλητότητα των δεδομένων διατηρείται, αλλά τα νέα χαρακτηριστικά είναι τώρα μη συσχετιζόμενα.

3.12 Διαδοχική ελάχιστη βελτιστοποίηση (SMO)

Εφαρμόζεται ο αλγόριθμος της διαδοχικής ελάχιστης βελτιστοποίησης του John Platt για την εκπαίδευση ενός "support vector" κατηγοριοποιητή. Αυτή η υλοποίηση αντικαθιστά όλες τις τιμές που λείπουν και μετατρέπει τα ονομαστικά χαρακτηριστικά σε δυαδικά. Επίσης ομαλοποιεί όλα τα χαρκτηριστικά από προεπιλογή (οι συντελεστές στην έξοδο βασίζονται στα κανονικοποιημένα δεδομένα και όχι στα αρχικά δεδομένα - αυτό είναι σημαντικό για τον κατηγοριοποιητή). Το πρόβλημα των πολλών κλάσεων επιλύεται χρησιμοποιώντας κατάταξη κατα ζεύγη (1 έναντι 1). Πιο λεπτομερής περιγραφή για τον αλγόριθμο παρουσιάζεται στο άρθρο [15].

Ο αλγόριθμος στην πιο απλή εκδοχή του (γραμμικα support vector machines για τη διαχωρίσιμη περίπτωση) λειτουργεί ως ακολούθως [3]:

Έστω ότι έχουμε l συνολικα παρατηρήσεις. Κάθε παρατήρηση αποτελείται απο ένα ζευγάρι: το διάνυσμα $x_i \epsilon R^n, i = 1, ..., l$ και την συσχετιζόμενη "αλήθεια" y_i, δοσμένη σε μάς από μια αξιόπιστη πηγή. Στο πρόβλημα αναγνώρισης δέντρων για παράδειγμα, x_i μπορεί να είναι ένα διάνυσμα τιμών pixel και y_i θα ήταν 1 άν η εικόνα περιέχει ένα δέντρο ή -1 διαφορετικά. Υποθέτουμε σταθερό y για δεδομένο x. Ας υποθέσουμε τώρα ότι έχουμε μία μηχανή της οποίας το έργο είναι να μάθουμε την αντιστοίχηση $x \mapsto y$.

Η μηχανή θεωρείται ότι είναι ντετερμινιστική: για μια δεδομένη είσοδο x και επιλογή a, θα δίνει πάντα την ίδια έξοδο $f(x, a)$. Μια συγκεκριμένη επιλογή a δημιουργεί αυτό που ονομάζουμε "εκπαιδευμένο μηχάνημα". Έτσι, για παράδειγμα, ένα νευρωνικό δίκτυο με σταθερή αρχιτεκτονική, με a που αντιστοιχεί στα βάρη και τα "biases", είναι μια μηχανή εκμάθησης υπό αυτή την έννοια.

Άρα έχουμε τα δεδομένα εκπαίδευσης $x_i, y_i, i = 1, ..., l, y_i \epsilon \{-1, 1\}, x_i \epsilon R^d$. Ας υποθέσουμε ότι έχουμε κάποιο υπερεπίπεδο που διαχωρίζει τα θετικά από τα αρνητικά παραδείγματα. Τα σημεία x που βρίσκονται στο υπερεπίπεδο ικανοποιούν τη συνθήκη $w*x+b = 0$, όπου w κανονικό στο υπερεπίπεδο, $|b|/||w||$ είναι η κάθετη απόσταση απο την αρχή των αξόνων και $||w||$ είναι η ευκλείδεια νόρμα του w. Ακόμα $d_+(d_-)$ είναι η μικρότερη απόσταση από το διαχωρίστικό υπερεπίπεδο στο πλησιέστερο θετικό (αρνητικό) παράδειγμα. Ορίζουμε το "περιθώριο" του διαχωριστικού υπερεπιπέδου να είναι $d_+ + d_-$. Για την γραμμικά διαχωρίσιμη περίπτωση, ο αλγόριθμος support vector αναζητά το διαχωριστικό υπερεπίπεδο με το μεγαλύτερο περιθώριο. Αυτό μπορεί να πάρει την ακόλουθη φόρμα: ας υποθέσουμε ότι όλα τα δεδομένα εκπαίδευσης ικανοποιούν τους ακόλουθους περιορισμούς:

$$x_i * w + b \geq +1 \quad \text{για} \quad y_i = +1 \tag{3.11}$$

$$x_i * w + b \leq -1 \quad \text{για} \quad y_i = -1 \tag{3.12}$$

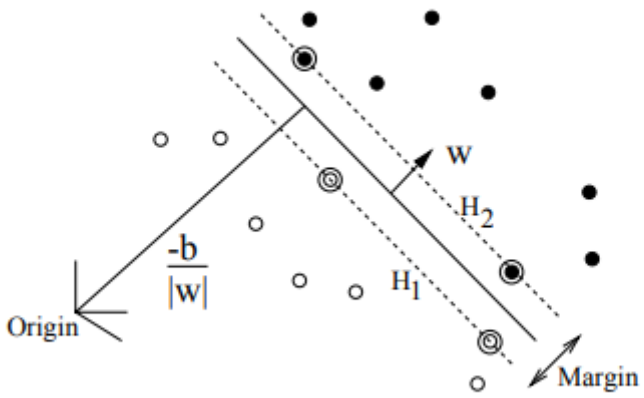
Αυτό μπορεί να συνδυαστεί σε ένα σύνολο ανισοτήτων:

$$y_i(x_i * w + b) - 1 \geq 0, \forall i \tag{3.13}$$

Ας δούμε τα σημεία για τα οποία ισχύει η ισότητα στην Εξ.(3.11) (αυτο το σημείο ισοδυναμεί με την επιλογή κλίμακας για τα w και b). Αυτά τα σημεία βρίσκονται στο υπερεπίπεδο $H1 : x_i * w + b = 1$ με κανονικό w και κάθετη απόσταση από την αρχή $|1 - b|/w$. Ομοίως, τα σημεία για τα οποία ισχύει η ισότητα στην Εξ.(3.12) βρίσκονται στο υπερεπίπεδο $H2 : x_i * w + b = -1$, με κανονικό πάλι w και κάθετη απόσταση απότην αρχή $|1 - b|/b$. Ως εκ τούτου $d_+ = d_- = 1/w$ και το περιθώριο είναι απλά $2/w$. Σημειώστε ότι τα $H1$ και $H2$ είναι παράλληλα και ότι δεν υπάρχουν σημεία προς εκπαίδευση ανάμεσα τους. Έτσι μπορούμε να βρούμε το ζεύγος των υπερεπιπέδων που δίνει το μέγιστο περιθώριο ελαχιστοποιώντας το $||w||^2$ υπό τους περιρισμούς της Εξ.(3.13).

Έτσι αναμένουμε ότι η λύση για μια τυπική διασδιάστατη περίπτωση θα έχει τη μορφη που φαίνεται στην εικόνα 1. Αυτά τα σημεία εκπαίδευσης για τα οποία ισχύει η ισότητα στην Εξ.(3.13) (δηλαδή εκείνα που καταλήγουν να βρίσκονται σε ένα από τα υπερεπίπεδα $H1, H2$) και των οποίων

η αφαίρεση θα άλλαζε τη λύση που βρέθηκε, ονομάζονται support vectors. Τα σημεία φαίνονται στην εικόνα 3.3 με τους επιπλέον κύκλους.



Σχήμα 3.3: Γραμμικά διαχωρίσιμα υπερεπίπεδα για την διαχωρίσιμη περίπτωση. Τα support vectors είναι κυκλωμένα.

Κεφάλαιο 4

Προεπεξεργασία και πρώτη προσέγγιση του προβλήματος

4.1 Εισαγωγή

Η παρούσα εργασία αντλεί δεδομένα από το IMDB (Internet Movie Database), μιας δωρεάν, διαχειριζόμενης από τον χρήστη, ιστοσελίδας πληροφοριών για περισσότερες από 390.000 ταινίες, τηλεοπτικές σειρές και βιντεοπαιχνίδια, όπου υπάρχουν πληροφορίες όπως ο τίτλος, το είδος της ταινίας, οι βαθμολογίες των χρηστών και πολλές άλλες [17]. Αποτελεί έγκυρη πηγή καθώς είναι η πιο δημοφιλής στο χώρο των κινηματογραφικών ταινιών και χρησιμοποιείται από εκατομμύρια χρήστες. Η βάση επιλέχθηκε καθώς περιέχει επαρκή αριθμό εγγραφών καθώς και αρκετά μεγάλη ποικιλία γνωρισμάτων. Αποτελείται απο 5043 ταινίες και εκτείνεται σε 66 χώρες σε χρονικό διάστημα 100 ετών. Περιέχει 28 μεταβλητές όπου το “imdb score” αποτελεί την ετικέτα κλάσης ενώ οι υπόλοιπες 27 αποτελούν πιθανά γνωρίσματα πρόβλεψης. Τα πεδία του αρχείου φαίνονται αναλυτικά στους πίνακες 4.1 και 4.2.

Όνομα μεταβλητής	**Περιγραφή**
movie_title	Ο τίτλος της ταινίας
duration	Η διάρκεια της ταινίας σε λεπτά
director_name	Το όνομα του σκηνοθέτη της ταινίας
director_facebook_likes	Ο αριθμός των facebook likes του σκηνοθέτη
actor_1_name	Πρωταρχικός ηθοποιός που πρωταγωνιστεί στην ταινία
actor_1_facebook_likes	Ο αριθμός των likes του Actor_1 στην σελίδα του Facebook

Πίνακας 4.1: Τα αρχικά γνωρίσματα της βάσης.

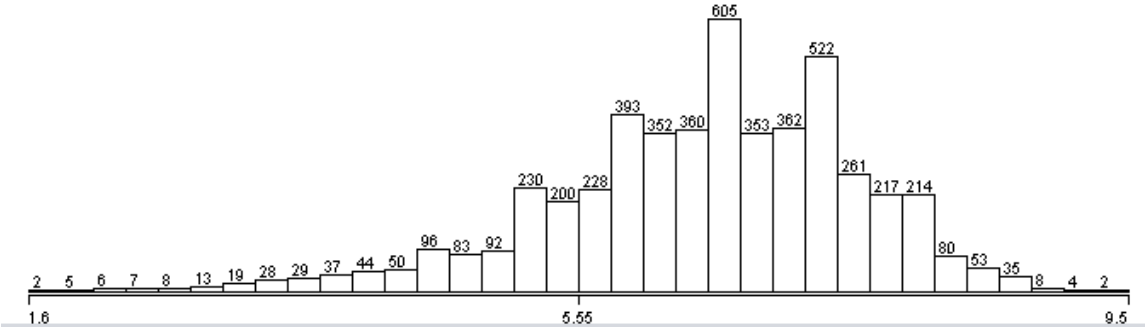
actor_2_name	Άλλος ηθοποιός που πρωταγωνιστεί στην ταινία
actor_2_facebook_likes	Ο αριθμός των likes του Actor_2 στην σελίδα του Facebook
actor_3_name	Άλλος ηθοποιός που πρωταγωνιστεί στην ταινία
actor_3_facebook_likes	Ο αριθμός των likes του Actor_3 στην σελίδα του Facebook
num_user_for_reviews	Αριθμός χρηστών που έδωσαν μια κριτική
num_critic_for_reviews	Ο αριθμός των κριτικών στο IMDb
num_voted_users	Αριθμός ατόμων που ψήφισαν για την ταινία
cast_total_facebook_likes	Συνολικός αριθμός likes στο facebook για ολόκληρο το καστ της ταινίας
movie_facebook_likes	Αριθμός των likes στο Facebook στην σελίδα της ταινίας
plot_keywords	Λέξεις-κλειδιά που περιγράφουν την πλοκή της ταινίας
facenumber_in_poster	Αριθμός των ηθοποιών που εμφανίζονται στην αφίσα της ταινίας
color	Χρωματισμός ταινίας. ‘Black and White’ ή ‘Color’
genres	Κατηγοριοποίηση ταινιών όπως ‘Animation’, ‘Comedy’, ‘Romance’, ‘Horror’, ‘Sci-Fi’, ‘Action’, ‘Family’
title_year	Το έτος κυκλοφορίας της ταινίας (1916: 2016)
language	Η γλώσσα της ταινίας
country	Χώρα όπου παράγεται η ταινία
content_rating	Βαθμολογία περιεχομένου της ταινίας
aspect_ratio	Αναλογία διαστάσεων στην οποία δημιουργήθηκε η ταινία
movie_imdb_link	IMDB link της ταινίας
gross	Μικτά κέρδη της ταινίας σε δολάρια
budget	Προϋπολογισμός της ταινίας σε δολάρια
imdb_score	Η βαθμολογία της ταινίας στο IMDB

Πίνακας 4.2: Τα αρχικά γνωρίσματα της βάσης.

4.2 Αρχικές παρατηρήσεις για τη βάση

Σκοπός της εργασίας είναι η πρόβλεψη της βαθμολογίας μιας ταινίας. Το γνώρισμα που δείχνει την βαθμολογία στην διαθέσιμη βάση δεδομένων είναι το imdb_score το οποίο είναι αριθμητικό και η κατανομή του φαίνεται στην εικόνα 4.1. Στο γράφημα, στον οριζόντιο άξονα, απεικονίζεται η τιμή του γνωρίσματος imdb_score και στον κατακόρυφο άξονα ο αριθμός των ταινιών. Επειδή

το γνώρισμα είναι αριθμητικό, θα πρέπει να κατηγοριοποιηθεί σε κλάσεις ώστε να μπορεί να γίνει μια αποδοτική πρόβλεψη.



Σχήμα 4.1: Γράφημα κατανομής δεδομένων στο γνώρισμα ”imdb_score”

Αρχικά το αρχείο της βάσης με τις 5043 εγγραφές εισάγεται στο πρόγραμμα της Weka και της R. Παρατηρείται ότι υπάρχουν 45 διπλότυπα τα οποία και αφαιρούνται οπότε η βάση αποτελείται πλέον από 4998 εγγραφές.

Χωρίς καμία περαιτέρω επεξεργασία διακριτοποιείται το γνώρισμα imdb_score με το κατάλληλο φίλτρο, μετατρέπεται σε ονομαστικό και δοκιμάζονται οι αλγορίθμοι που φαίνονται στον πίνακα 4.3 με γνώρισμα κλάσης το imdb_score. Πιο συγκεκριμένα ο πίνακας 4.3 δείχνει το ποσοστό (επί τοις εκατό) των εγγραφών που έχουν κατηγοριοποιηθεί σωστά με τους αντίστοιχους κάδους. Οι δοκιμές έχουν γίνει στο πρόγραμμα Weka. Συνακόλουθα τα πειράματα έχουν γίνει με την επιλογή Cross-validation folds = 10 (4.2.1). Όπου υπάρχει παύλα σημαίνει πως ο αλγόριθμος δεν μπόρεσε να τερματίσει ώστε να δώσει αποτελέσματα.

Πίνακας 4.3: Χωρίς επεξεργασία των δεδομένων

Αλγόριθμος	Κάδοι#3	Κάδοι#5	Κάδοι#10
NaiveBayes	39.22	28.55	18.76
Logistic	-	-	-
MultilayerPerceptron	-	-	-
ClassificationViaRegression	-	-	-
LogitBoost	70.93	63.20	39.44
DecisionTable	70.51	57.88	34.52
J48	64.39	52.98	30.04
RandomForest	-	-	-

Παρατηρείται πως οι αλγόριθμοι Logistic, MultilayerPerceptron, ClassificationViaRegression και ο RandomForest δεν καταφέρνουν να παράγουν κάποιο αποτέλεσμα. Αντίθετα, οι αλγόριθμοι NaiveBayes, LogitBoost, DecisionTable και J48 κατηγοριοποιούν τα δεδομένα με χειρότερη απόδοση αυτή του NaiveBayes και καλύτερη, με μικρή διαφορά σε σχέση με τον DecisionTable, αυτή του LogitBoost. Ακόμη είναι φανερό και αναμενόμενο πως όσο αυξάνεται ο αριθμός των κλάσεων

μειώνεται και η απόδοση των αλγορίθμων. Για την συνέχεια της εργασίας επιλέγεται να γίνουν τα πειράματα με έξι κλάσεις έχοντας έτσι ακρίβεια ως προς την πρόβλεψη και μία αξιοπρεπή απόδοση.

Στους πίνακες 4.4, 4.5 και 4.6 φαίνεται, στην δεύτερη στήλη, πως είναι διαμορφωμένες οι κλάσεις, ενώ στην τρίτη στήλη παρουσιάζεται ο αριθμός των εγγραφών που ανήκει σε κάθε κλάση. Αυτή η τιμή χρησιμοποιείται ως βάρος στους αλγορίθμους όπως φαίνεται και στην τέταρτη στήλη.

Πίνακας 4.4: Στατιστικά για τους τρείς κάδους στο γνώρισμα imdb_score

Αριθμός	Κλάση	Αριθμός	Βάρος
1	(0-4.23]	218	218.0
2	(4.23-6.87]	2849	2849.0
3	(6.87-10]	1931	1931.0

Πίνακας 4.5: Στατιστικά για τους πέντε κάδους στο γνώρισμα imdb_score

Αριθμός	Κλάση	Αριθμός	Βάρος
1	(0-3.18]	53	53.0
2	(3.18-4.76]	322	322.0
3	(4.76-6.34]	1722	1722.0
4	(6.34-7.92]	2577	2577.0
5	(7.92-10]	324	324.0

Πίνακας 4.6: Στατιστικά για τους δέκα κάδους στο γνώρισμα imdb_score

Αριθμός	Κλάση	Αριθμός	Βάρος
1	(0-2.39]	16	16.0
2	(2.39-3.18]	37	37.0
3	(3.18-3.97]	101	101.0
4	(3.97-4.76]	221	221.0
5	(4.76-5.55]	574	574.0
6	(5.55-6.34]	1148	1148.0
7	(6.34-7.13]	1505	1505.0
8	(7.13-7.92]	1072	1072.0
9	(7.92-8.71]	303	303.0
10	(8.71-10]	21	21.0

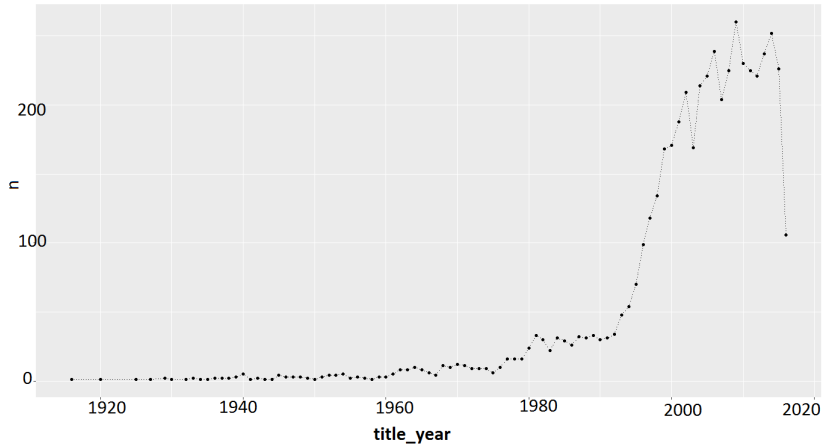
4.2.1 Διασταυρωμένη επικύρωση (cross-validation)

Η διασταυρωμένη επικύρωση (cross-validation) [5], μια τυποποιημένη τεχνική αξιολόγησης, είναι ένας συστηματικός τρόπος εκτέλεσης επαναλαμβανόμενων ποσοστών διαχωρισμού. Διαχωρίζεται ένα σύνολο δεδομένων σε 10 κομμάτια (”πτυχές”-”folds”), στη συνέχεια κρατιέται το κάθε κομμάτι με τη σειρά του (για έλεγχο) για να δοκιμάστουν και να εκπαιδευτούν τα υπόλοιπα 9 μαζί. Αυτό δίνει 10 αποτελέσματα αξιολόγησης, τα οποία είναι υπολογισμένα κατά μέσο όρο. Στην διασταυρωμνη επικύρωση, όταν γίνεται η αρχική διαίρεση, διασφαλίζεται ότι κάθε πτυχή περιέχει περίπου το σωστό ποσοστό των τιμών κλάσης. Έχοντας κάνει 10 φορές την επικύρωση και υπολογίσει τα αποτελέσματα της αξιολόγησης, το πρόγραμμα Weka επικαλείται τον αλγόριθμο εκμάθησης μια τελική (11η) φορά σε ολόκληρο το σύνολο δεδομένων για να αποκτήσει το μοντέλο που εκτυπώνει. Αυτή η προσέγγιση έχει το πλεονέκτημα ότι χρησιμοποιεί όσο το δυνατό περισσότερα δεδομένα για την εκπαίδευση. Επιπλέον, τα σύνολα ελέγχου είναι αμοιβαία αποκλειόμενα και ουσιαστικά καλύπτουν όλο το σύνολο δεδομένων. Το μειονέκτημα της μεθόδου, είναι ότι είναι υπολογιστικά ακριβή η επανάληψη N φορές [12].

4.3 Πρώτη προσέγγιση - Επιλογή γνωρισμάτων με οπτικοποίηση

Σε αυτή την προσέγγιση γίνεται μελέτη στα αρχικά δεδομένα και στη συνέχεια επιλέγονται τα γνωρίσματα που φαίνεται να καθορίζουν και να επηρεάζουν την βαθμολογία μια ταινίας [7].

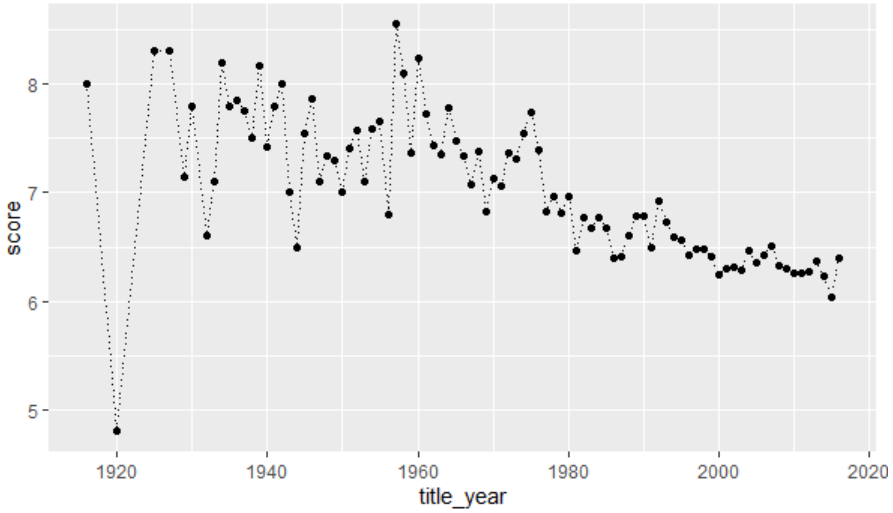
Στο γράφημα της εικόνας 4.2 φαίνεται η ετήσια παραγωγή ταινιών όπως αυτή έχει καταγραφεί στη βάση δεδομένων. Ο οριζόντιος άξονας αναπαριστά τις τιμές του γνωρίσματος title_year ενώ ο κάθετος άξονας τον αριθμό των ταινιών.



Σχήμα 4.2: Αριθμός ταινίων που έχουν βγει κάθε χρονια

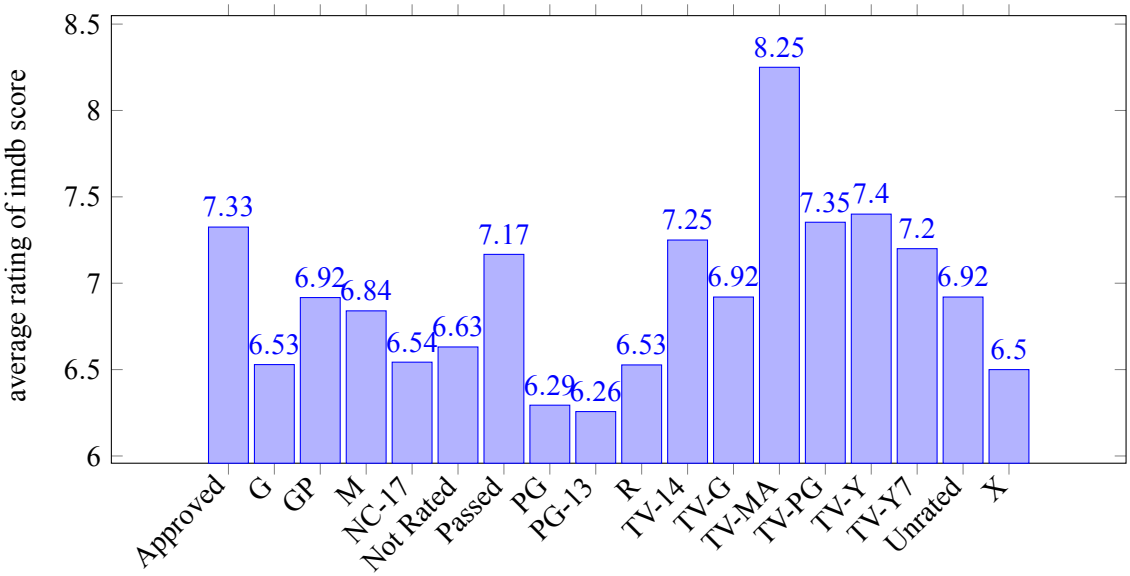
Παρατηρείται ότι στα δεδομένα υπάρχουν περισσότερες ταίνιες για τα τελευταία 30 χρόνια. Πρίν την χρονολογία του 1990 ο αριθμός των ταινιών είναι αρκετά μικρός και μάλλον όχι αντιπροσωπευτικός. Για το λόγο αυτό διατηρούνται στην βάση οι ταινίες που έχουν κυκλοφορήσει τα τελευταία τριάντα χρόνια.

Στο γράφημα 4.3 που ακολουθεί φαίνεται ο μέσος όρος του γνωρίσματος imdb_score ανά έτος. Στον οριζόντιο άξονα απεικονίζεται το γνώρισμα title_year και στον κάθετο άξονα η βαθμολογία των ταινιών (imdb_score). Παρατηρείται έντονη διακύμανση των μέσων όρων για τις χρονολογίες πριν το 1990 και το γεγονός αυτό δικαιολογείται καθώς δεν υπάρχουν αρκετές εγγραφές για το διάστημα αυτό. Για τις ταινίες τα τελευταία τριάντα χρόνια φαίνεται πως ο μέσος όρο της βαθμολογίας τους κυμαίνεται από 6 εως 7.



Σχήμα 4.3: Αριθμός ταινίων που έχουν βγει κάθε χρονια

Στην εικόνα 4.4 φαίνεται πως διακυμαίνεται ο μέσος όρος της βαθμολογίας μια ταινίας σε σχέση με τη βαθμολογία του περιεχομένου (content_rating). Στον άξονα των x αναπαριστούνται οι κατηγορίες που υπάρχουν στο γνώρισμα content_rating και στον άξονα των y αναπαρίσταται ο μέσος όρων των βαθμολογιών των ταινιών όπου υπάρχουν οι συγκεκριμένες κατηγορίες.



Σχήμα 4.4: Μέσος όρος των τιμών του γνωρίσματος content_rating.

Οι κατηγορίες που υπάρχουν στο γνώρισμα content_rating είναι:

Κατηγορία	**Περιγραφή**
Approved	Κατάλληλο για όλους
G (general)	Κατάλληλο για όλους
GP	Κατάλληλο για όλους, προτείνεται η γονική συναίνεση
M	Για ενήλικες
NC-17	Άνω τν 17
Not Rated	Δεν υπάρχει κάποια ταξινόμηση στις υπόλοιπες κατηγορίες (χρησιμοποιείται π.χ. σε κάποιο trailer)
Passed	Εγκρίνεται για ...
PG	Κατάλληλο με γονική συναίνεση
PG-13	Με γονική συναίνεση, άνω των 13 ετών
R	Κάτω των 17 ετώβ απαιτείται συνοδός γονέα
TV-14	Απαραίτητη η γονική συναίνεση, άνω των 14 ετών
TV-G	Κατάλληλο για όλους
TV-MA	Για ενήλικους
TV-PG	Κατάλληλο για όλους προτείνεται η γονική συναίνεση
TV-Y	Για παιδιά
TV-Y7	Για παιδιά άνω των 7 ετών
Unrated	Δεν ανήκει σε κάποια κατηγορία
X	ακατάλληλο για ανηλίκους

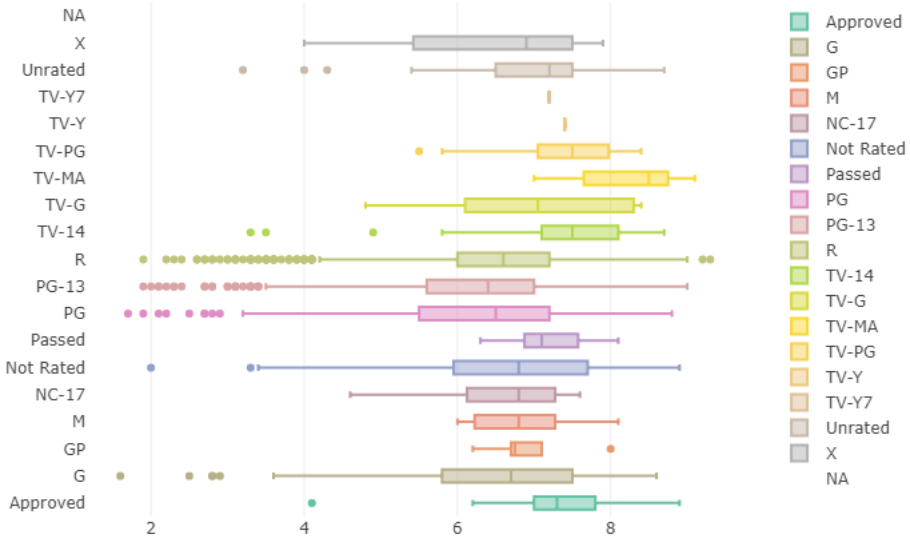
Πίνακας 4.7: Οι κατηγορίες του γνωρίσματος content_rating.

Με βάση το προηγούμενο γράφημα η υψηλότερη μέση βαθμολογία φαίνεται να είναι στην κατηγορία TV-MA, δηλαδή στην κατηγορία που επιτρέπεται να παρακολουθήσουν την ταινία οι ενήλικες.

Όταν ένα σύνολο δεδομένων έχει ακραίες τιμές, συνοψίζεται μια τυπική τιμή χρησιμοποιώντας τη διάμεσο σε αντίθεση με το μέσο όρο [16]. Όταν ένα σύνολο δεδομένων έχει ακραίες τιμές, η μεταβλητότητα συνοψίζεται συχνά από μια τιμή που ονομάζεται εύρος μεταξύ τεταρτημορίων, η οποία είναι η διαφορά μεταξύ του πρώτου και του τρίτου τεταρτημορίου. Το πρώτο τεταρτημόριο, που υποδηλώνεται Q1, είναι η τιμή στο σύνολο δεδομένων που κρατά το 25% των τιμών κάτω από αυτό. Το τρίτο τεταρτημόριο, που υποδηλώνεται Q3, είναι η τιμή στο σύνολο δεδομένων που κρατά το 25% των τιμών πάνω από αυτό. Τα τεταρτημόρια μπορούν να προσδιοριστούν ακολουθώντας την ίδια προσέγγιση που χρησιμοποιήσαμε για τον προσδιορισμό της διαμέσου, αλλά τώρα εξετάζουμε κάθε μισό του συνόλου δεδομένων ξεχωριστά. Το εύρος μεταξύ των τεταρτημορίων ορίζεται ως εξής:

$$Interquartile \quad range = Q3 - Q1$$

Για τον έλεγχο της διακύμανσης των τιμών δημιουργείται το γράφημα της εικόνας 4.5 όπου φαίνεται πως διαφέρουν οι βαθμολογίες σε κάθε κατηγορία του content rating.



Σχήμα 4.5: Κατανομή βαθμολογιών σε κάθε κατηγορία

Είναι φανερό ότι το IQR (interquartile range) κάθε κατανομής είναι πάνω από 5. Τα υψηλότερα imdb_scores τείνουν να είναι του TV-MA τύπου βαθμολογίας περιεχομένου. Η κατηγορία R έχει τον μεγαλύτερο αριθμό ακραίων τιμών που κυμαίνονται από σκορ 1.9 έως 4.2. Μετά την εξαγωγή των συμπερασμάτων για το γνώρισμα content_rating, θα δοκιμαστεί για το αν βοηθάει στην καλύτερη πρόβλεψη της βαθμολογίας.

Συνακόλουθα, για το γνώρισμα color, εφόσον οι 4185 εγγραφές έχουν την τιμή “Color” ενώ μόνο οι 209 έχουν τιμή “Black and White” και 19 έχουν κενή τιμή, κρατιούνται για την βάση μόνο οι έγχρωμες ταινίες.

Επίσης δημιουργείται ο πίνακας της εικόνας 4.6 που αναφέρεται στις διαφορετικές τιμές του γνωρίσματος aspect_ratio (δηλαδή την ανάλυση της εικόνας).

	1.33	1.37	1.44	1.5	1.66	1.75	1.77	1.78	1.85	2	2.2	2.24	2.35	2.39	2.4	2.55	2.76
1.33	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.37	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.44	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1.66	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0
1.75	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
1.77	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
1.78	0	0	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0
1.85	0	0	0	0	0	0	0	0	1300	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
2.2	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0
2.24	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
2.35	0	0	0	0	0	0	0	0	0	0	0	0	1732	0	0	0	0
2.39	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0
2.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
2.55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
2.76	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Σχήμα 4.6: Πίνακας για τις τιμές του γνωρίσματος aspect ratio.

Όπως παρατηρείται οι περισσότερες εγγραφές έχουν τιμές 1.85 και 2.35. Έτσι δημιουργούνται τρείς κατηγορίες a) 1.85, b)2.35 και c)Others ώστε να έχουμε στις κατηγορίες ισάξιο αριθμό εγγραφών. Στον πίνακα 4.8 φαίνονται οι κατηγορίες και αριθμός εγγραφών σε κάθε κατηγορία για το γνώρισμα aspect_ratio.

a) 1.85	b) 2.35	c) Others
1300	1732	102

Πίνακας 4.8: Κατηγοριοποίηση των εγγραφών του γνωρίσματος aspect_ratio.

Επιπλέον, στον πίνακα 4.9 φαίνονται οι τιμές που λείπουν σε κάθε γνώρισμα.

γνώρισμα	**τιμές που λείπουν**
duration	1
actor_3_facebook_likes	4
actor_2_name	1
actor_3_name	4
facenumber_in_poster	5
plot_keywords	5
language	1
content_rating	11
budget	141
actor_2_facebook_likes	1
aspect_ratio	31

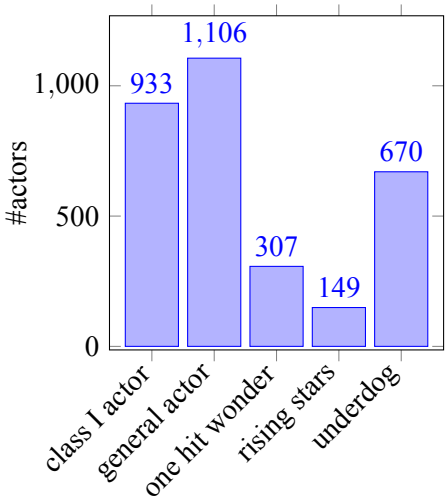
Πίνακας 4.9: Ο αριθμός των τιμών που λείπουν από τα γνωρίσματα.

Οι τιμές που λείπουν είναι πολύ λίγες σε σχέση με τον αριθμό των εγγραφών και άρα μπορούμε να χρησιμοποιήσουμε τα γνωρίσματα χωρίς να επηρεάζεται το αποτέλεσμα από αυτές τις τιμές.

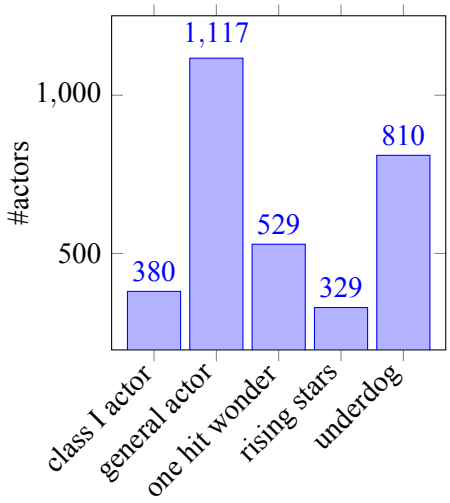
Σε αυτή την προσέγγιση η κατηγοριοποίηση των γνωρισμάτων των ηθοποιών και των σκηνοθετών μπαίνουν σε πέντε κλάσεις (στη μεταβλητή number_movie είναι καταγεγραμμένος ο αριθμός των ταινιών που παίζει ένας ηθοποιός):

1. **Class I actor** (Πρώτης κλάσης ηθοποιοί): average_rating >= 6.5 και number_movie >= 5.
2. **Rising Stars** (Ανερχόμενα αστέρια): average_rating >= 6.5 και number_movie >= 3.
3. **One hit wonder** (Ηθοποιοί με λίγες αλλά μεγάλες επιτυχίες): average_rating >= 7 και number_movie < 3.
4. **General actor** (Γενικοί ηθοποιοί): average_rating >= 6 και δεν ισχύουν τα παραπάνω.
5. **Underdog** (Όχι φαβορί): αν δεν ανήκει σε καμία από τις παραπάνω κατηγορίες.

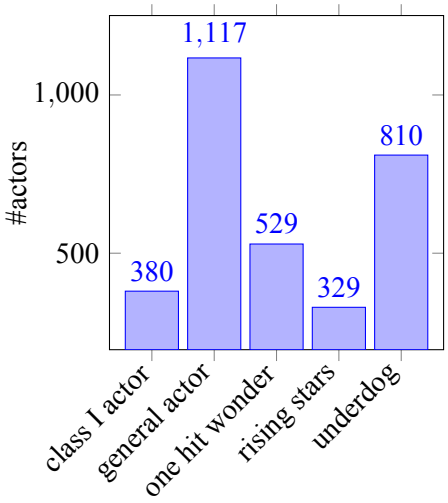
Παρακάτω φαίνονται τα γραφήματα για το πως είναι διαμορφωμένα πλέον τα γνωρίσματα actor1_rating, actor2_rating, actor3_rating, director_rating. Στον οριζόντιο άξονα απεικονίζονται οι κατηγορίες στις οποίες ανήκουν οι ηθοποιοί ενώ στον κάθετο άξονα είναι ο αριθμός των ηθοποιοών ή των σκηνοθετών αντίστοιχα.



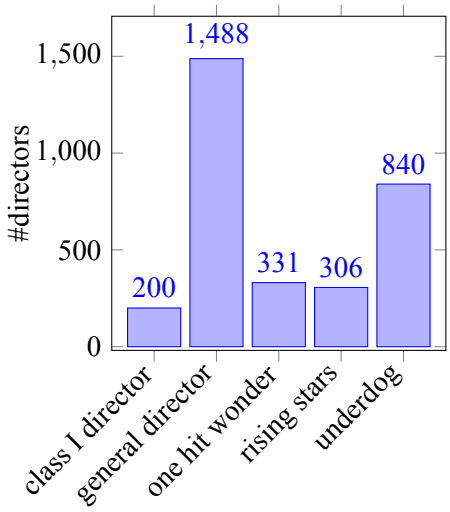
Σχήμα 4.7: Πρώτοι ηθοποιοί.



Σχήμα 4.8: Δεύτεροι ηθοποιοί.



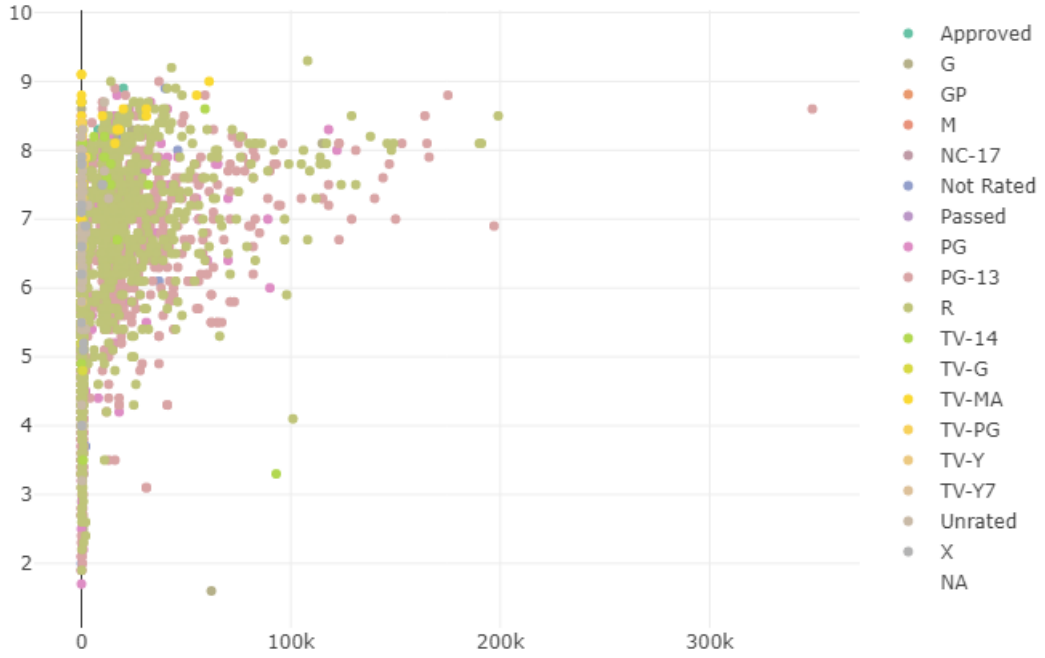
Σχήμα 4.9: Τρίτοι ηθοποιοί.



Σχήμα 4.10: Σκηνοθέτες.

Η τάση της ταξινόμησης των ηθοποιών και των σκηνοθετών στις πέντε κλάσεις στα τέσσερα παραπάνω γραφήματα φαίνεται να είναι παρόμοια. Οι περισσότεροι ηθοποιοί και σκηνοθέτες κατατάσσονται στην δεύτερη κλάση, general actor και general director αντίστοιχα. Ενώ η αμέσως επόμενη κλάση σε πλήθος είναι η κλάση “underdog”. Τέλος, παρατηρείται πως σαν πρώτοι ηθοποιοί συνήθως επιλέγονται ηθοποιοί απο την κλάση “class I actor”, ενώ λίγοι από τη συγκεκριμένη κλάση παίζουν σαν δεύτεροι ή τρίτοι ηθοποιοί.

Στη συνέχεια ελέγχεται αν το γνώρισμα facebook_likes επηρεάζει το γνώρισμα imdb_score. Στο γράφημα της εικόνας 4.11 στον οριζόντιο άξονα είναι ο αριθμός των facebook likes για την ταινία και στον κάθετο άξονα είναι η βαθμολογία της ταινίας (imdb_score). Ακόμη, με διαφορετικό χρώμα γίνονται φανερές οι κατηγορίες του γνωρίσματος content_rating. Στα δεξιά του γραφήματος παρουσιάζονται και οι κατηγορίες με το αντίστοιχο χρώμα.

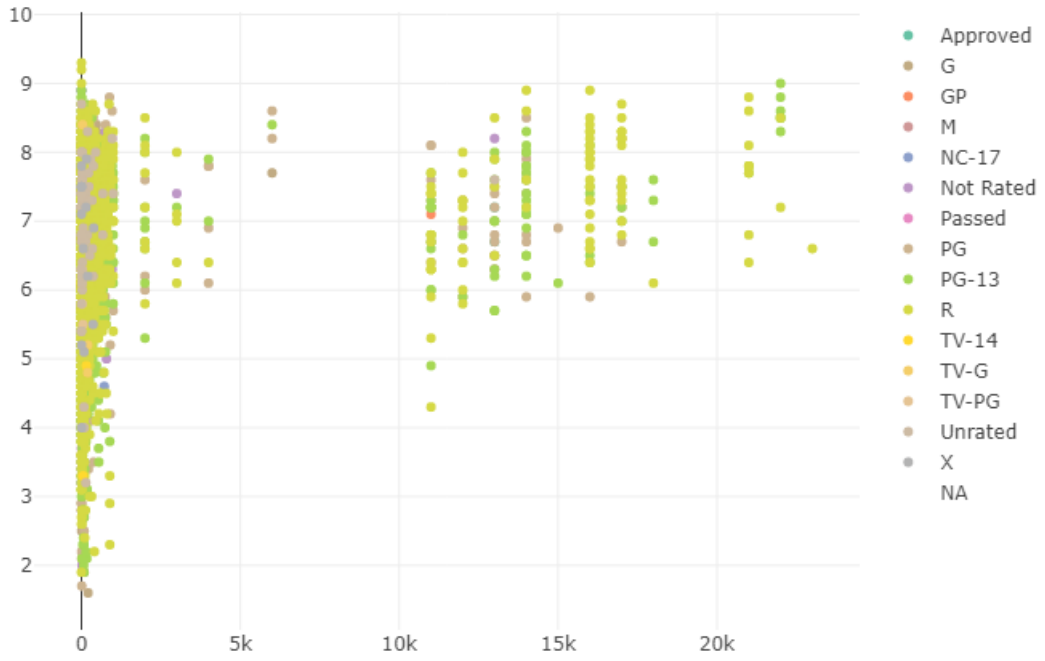


Σχήμα 4.11: Γράφημα για τις τιμές του γνωρίσματος facebook_likes ως προς το imdb_score.

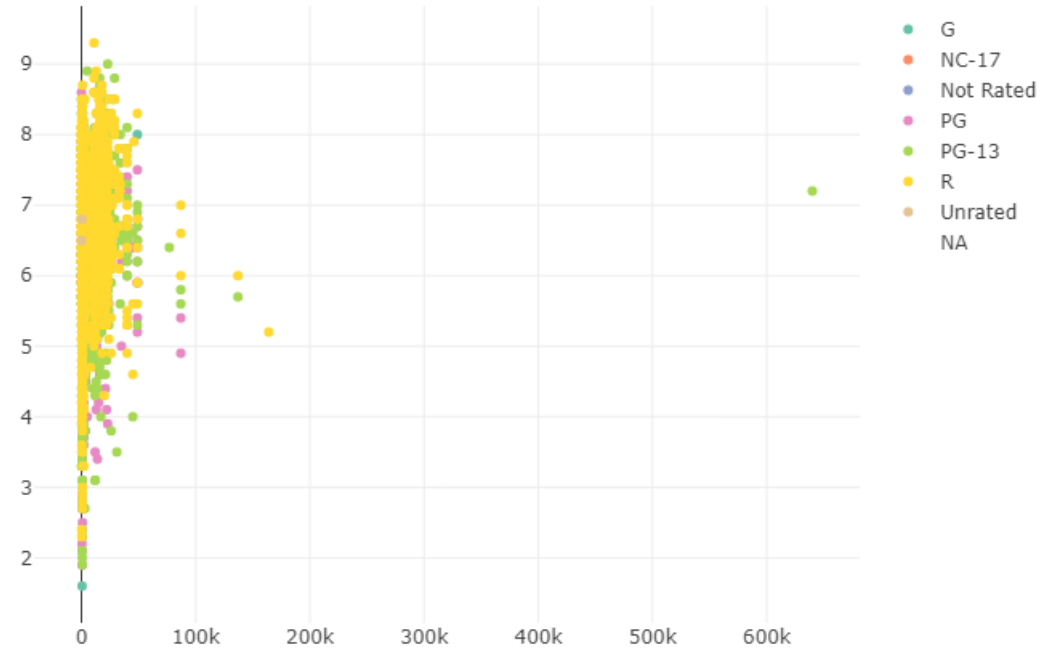
Στο συγκεκριμένο γνώρισμα δεν μπορεί να βρεθεί κάποια τάση. Φαίνεται να υπάρχουν ταινίες που έχουν υψηλές βαθμολογίες IMDB αλλά μικρό αριθμό από likes στο Facebook.

Επίπρόσθετα, και παρόμοια με το γράφημα 4.11, ελέγχεται αν επηρεάζει ο αριθμός των facebook likes των σκηνοθετών το imdb score στο γράφημα 4.12. Όμοια με τις ταινίες, δεν φαίνεται να βοηθάει το γνωρισμα director_facebook_likes στην διεξαγωγή κάποιου συμπεράσματος . Η κατανομή στο γράφημα δείχνει πως πολλοί σκηνοθέτες έχουν μικρό αριθμό από likes ανεξαρτήτως του πόσο καλοί είναι. Υπάρχουν σκηνοθέτες με υψηλές βαθμολογίες IMDB και με μικρό αριθμό likes στο Facebook, άρα το γνώρισμα δεν είναι κατάλληλο για ακριβή πρόβλεψη.

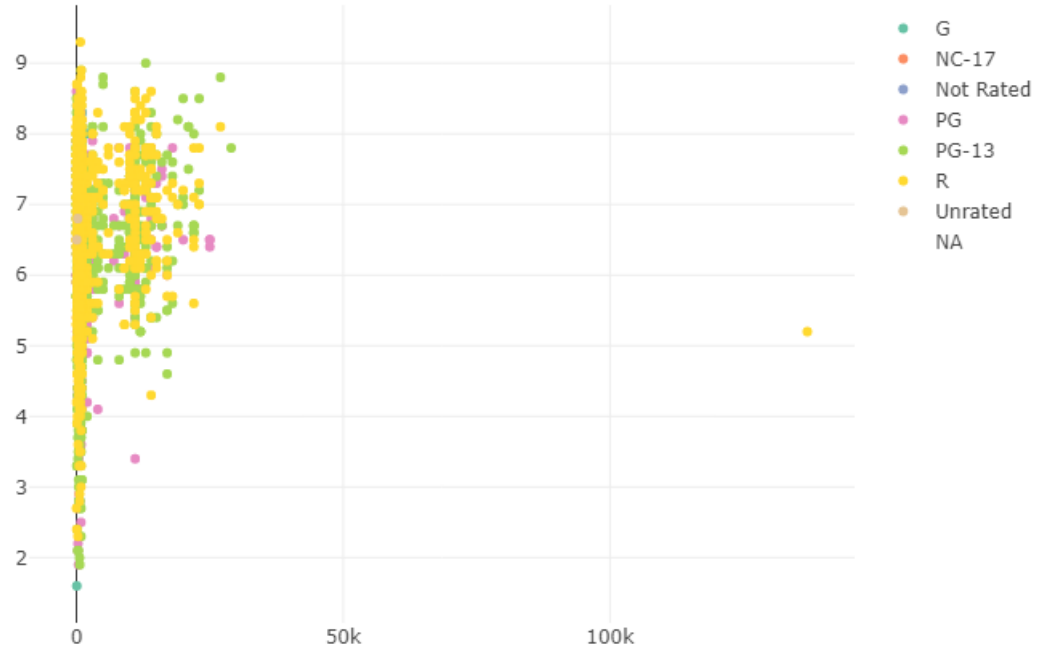
Παρεμφερή είναι και τα συμπεράσματα για τα facebook likes των ηθοποιών, όπως παρατηρείται και στα γραφήματα των ηθοποιών που φαίνονται στις εικόνες 4.13, 4.14 και 4.15. Στον οριζόντιο άξονα είναι τα facebook likes των ηθοποιών και στον κάθετο άξονα είναι η βαθμολογία της ταινίας στις οποίες συμμετέχουν οι συγκεκριμένοι ηθοποιοί. Μία παρατήρηση στα προηγούμενα γραφήματα είναι πως στα γραφήματα των ηθοποιών φαίνεται να κυριαρχεί το κίτρινο χρώμα. Αυτό δείχνει πως τα likes για τους ηθοποιούς γίνονται σε ταινίες που ανήκουν στην κατηγορία R του γνωρίσματος content rating.



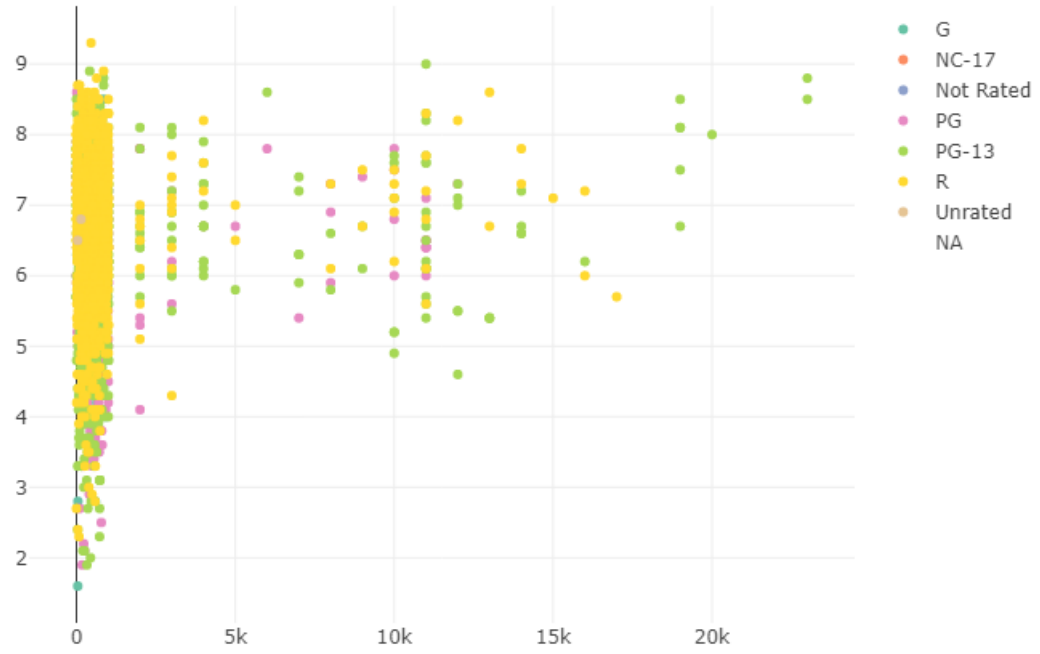
Σχήμα 4.12: Γράφημα για τις τιμές του γνωρίσματος director_facebook_likes ως προς το imdb_score.



Σχήμα 4.13: Γράφημα για τις τιμές του γνωρίσματος actor1_facebook_likes ως προς το imdb_score.



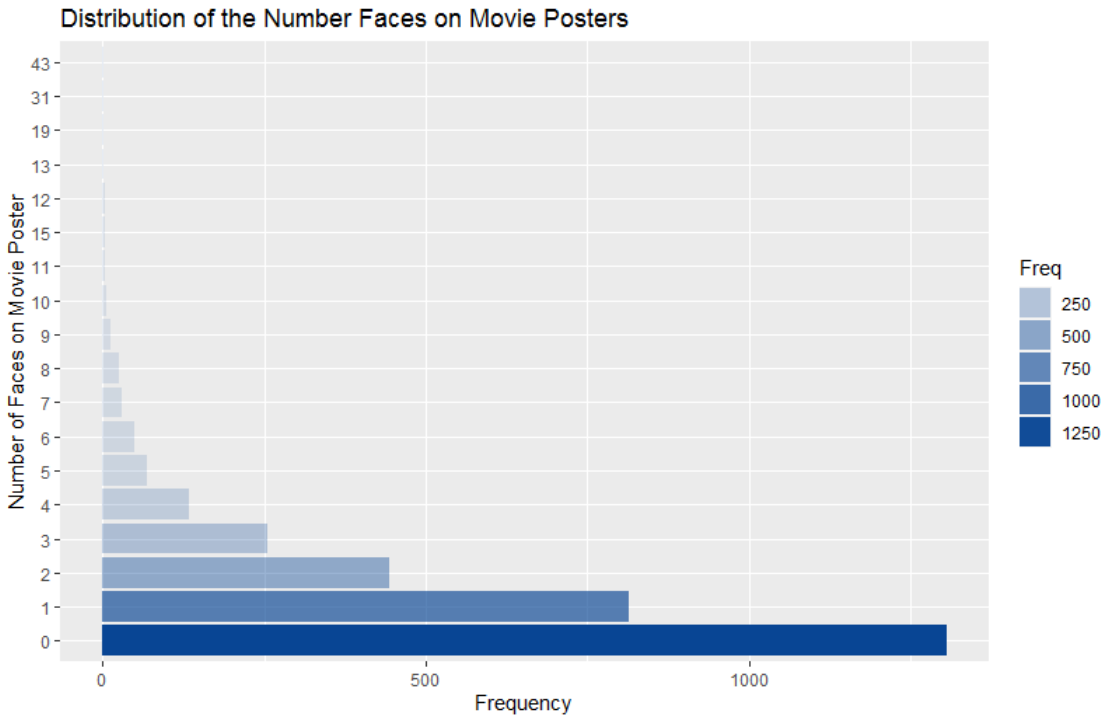
Σχήμα 4.14: Γράφημα για τις τιμές του γνωρίσματος actor2_facebook_likes ως προς το imdb_score.



Σχήμα 4.15: Γράφημα για τις τιμές του γνωρίσματος actor3_facebook_likes ως προς το imdb_score.

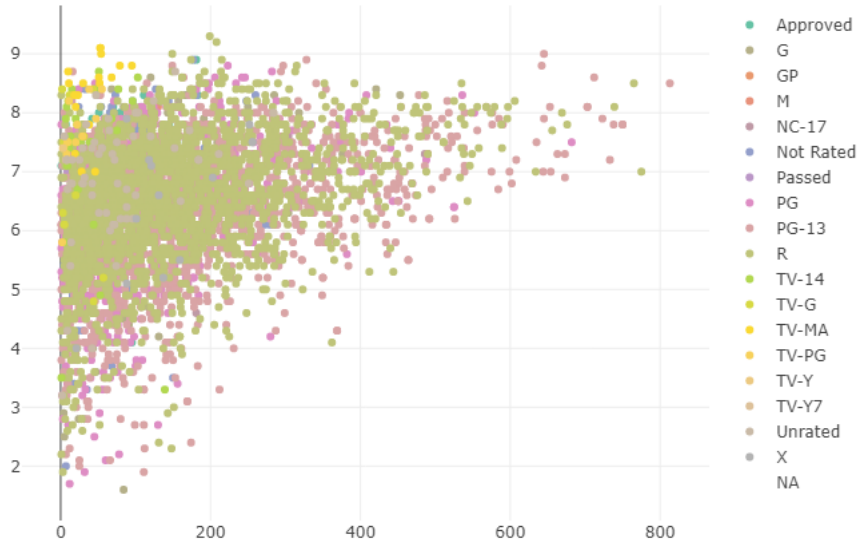
Τα συμπεράσματα για τα τρία παραπάνω γραφήματα που αφορούν τα likes των ηθοποιών είναι παρόμοια με αυτά που προέκυψαν για τους σκηνοθέτες. Δεν φανερώνεται κάποια τάση.

Ένα ακόμη γνώρισμα που μπορεί να μελετηθεί για το αν επηρεάζει την βαθμολογία μιας ταινίας είναι ο αριθμός των προσώπων στην αφίσα της ταινίας (facenumber_in_poster). Στο γράφημα του σχήματος 4.16 στον οριζόντιο άξονα φανερώνεται η συχνότητα των ταινιών και στον κάθετο άξονα είναι ο αριθμός των προσώπων της αφίσας. Επίσης στα δεξιά φαίνεται η κλίμακα του μπλέ χρώματος (σκούρο-ανοιχτό) που αντιστοιχεί στην συχνότητα με την οποία εμαφανίζεται η συγκεκριμένη αφίσα. Γίνεται ευδιάκριτο πως οι περισσότερες ταινίες έχουν στις αφίσες τους μέχρι δύο πρόσωπα. Καθώς δεν υπάρχει μεγάλη διαφοροποίηση στην κατηγοριοποίηση των εγγραφών ως προς αυτό το γνώρισμα, συνέπως ούτε σε αυτή την περίπτωση είναι δυνατό να βρεθεί κάποιο μοτίβο που να είναι εποικοδομητικό ως προς τη βελτίωση της πρόβλεψης.



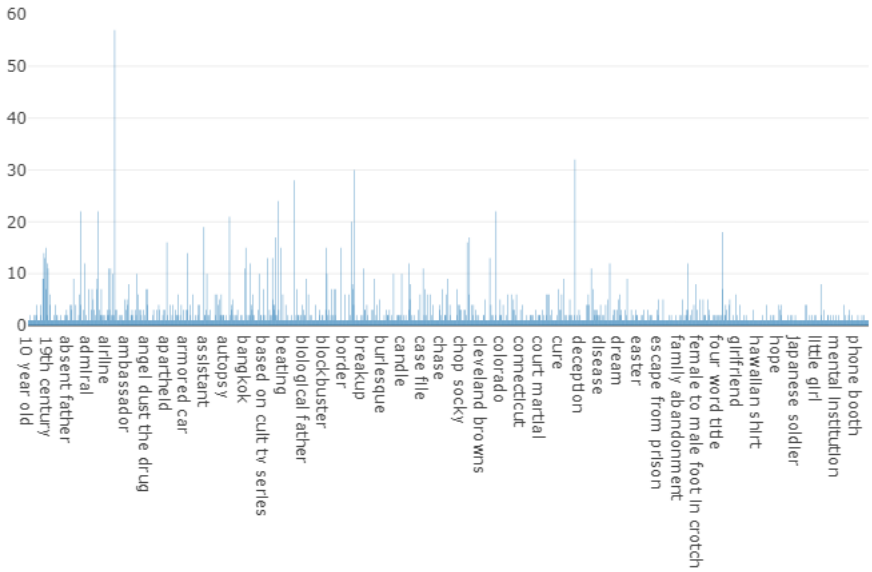
Σχήμα 4.16: Γράφημα για τις τιμές του γνωρίσματος facenumber_in_poster.

Στη συνέχεια εξετάζεται το γνώρισμα num_critic_for_reviewes. Στο γράφημα του σχήματος 4.17 ο οριζόντιος άξονας δείχνει την τιμή του γνωρίσματος num_critic_for_reviewes (αριθμός των κριτικών στο IMDB) και ο κάθετος άξονας δείχνει την βαθμολογία της ταινίας (imdb_score). Και σε αυτό το γράφημα φαίνεται να κυριαρχεί η R κατηγορία του content rating, αυτή τη φορά με πράσινο χρώμα. Σε αυτό το γνώρισμα φαίνεται να υπάρχει καλύτερη κατηγοριοποίηση των εγγραφών η οποία μπορεί να επηρεάσει την πρόβλεψη της βαθμολογίας. Το αρνητικό σε αυτή την περίπτωση είναι πως δεν υπάρχει η δυνατότητα να είναι γνωστός ο αριθμός των κριτικών πριν την κυκλοφορία μιας ταινία, άρα θα πρέπει να το παραβλέψουμε.



Σχήμα 4.17: Γράφημα για τις τιμές του γνωρίσματος num_critic_for_reviewes ως προς το imdb_score.

Τέλος γίνεται έλεγχος για το γνώρισμα plot_keywords. Δημιουργείται το γράφημα της εικόνας 4.18 στο οποίο απεικονίζονται οι διαφορετικές λέξεις κλειδιά στον άξονα των x και ο αριθμός εμφάνισης της κάθε λέξης στον άξονα των y. Όπως μπορεί να γίνει φανερό υπάρχουν πολλές διαφορετικές λέξεις που εμφανίζονται λίγες φορές η καθεμία. Για το λόγο αυτό ούτε το γνώρισμα plot_keywords μπορεί να βοηθήσει στην βελτίωση της πρόβλεψης.



Σχήμα 4.18: Μέρος του γραφήματος για το πόσες φορές εμφανίζεται η κάθε τιμη του plot_keywords.

Έχοντας κάνει όλη την προηγούμενη μελέτη του κεφαλαίου σχετικά με τα γνωρίσματα γίνονται τα πειράματα για να επαληθευτούν όσα έχουν προηγηθεί. Τα γνωρίσματα που τελικά επηρεάζουν περισσότερο την κλάση, και είναι δυνατό να είναι γνωστά πριν την κυκλοφορία της ταινίας, είναι τα genre, director_rating, actor1_rating, actor2_rating, actor3_rating και aspect_ratio.

Λαμβάνοντας υπόψη όλα τα παραπάνω, πραγματοποιήθηκαν πειράματα με μια σειρά γνωστών αλγορίθμων ταξινόμησης. Τα αποτελέσματα των πειραμάτων φαίνονται στον πίνακα 4.10. Στην δεύτερη στήλη ο αριθμός αντιπροσωπεύει το ποσοστό με το οποίο ο αλγόριθμος καταφέρνει να προβλέψει με επιτυχία την κλάση της ταινίας με την μέθοδο του cross validation. Όλοι οι αλγόριθμοι φαίνεται να έχουν παρόμοια απόδοση με μεγαλύτερη αυτή του Logistic με ποσοστό επιτυχίας 72.95.

Πίνακας 4.10: Αποτελέσματα από την πρώτη προσέγγιση

Αλγόριθμος	Αποτέλεσμα
Naive Bayes	72.92
Logistic	72.95
Multilayer Perceptron	69.73
SMO	71.23
ClassificationViaRegression	72.69
LogitBoost	72.65
DecisionTable	72.06
J48	72.16
RandomForest	70.04

Κεφάλαιο 5

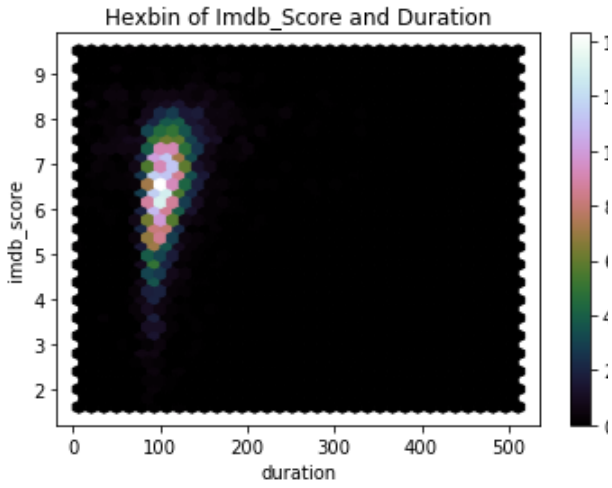
Ανάλυση κύριων συνιστωσών - Δεύτερη προσέγγιση του προβλήματος

5.1 Εισαγωγή

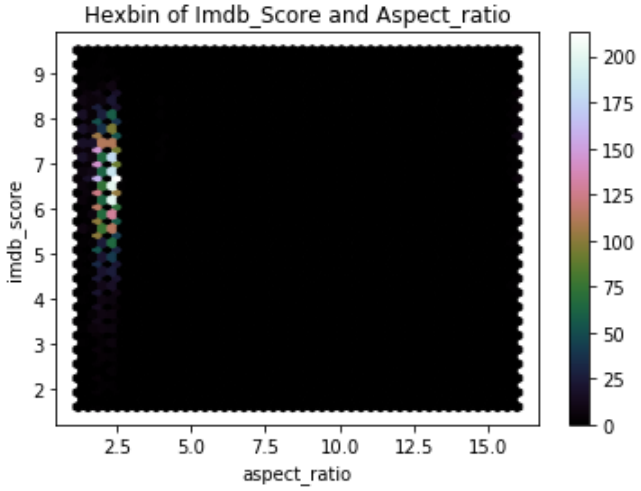
Σε αυτό το κεφάλαιο χρησιμοποιείται η ίδια βάση δεδομένων με τις ταινίες από το IMDB. Σε αυτή την προσέγγιση όμως δεν οπτικοποιούνται τα δεδομένα. Χρησιμοποιείται η ανάλυση κύριων συνιστωσών για να μειωθούν τα γνωρίσματα στην προσπάθεια να γίνει πιο αποδοτική η πρόβλεψη του αποτελέσματος. Για την μέθοδο της ανάλυσης των κύριων συνιστωσών χρησιμοποιήθηκε η υλοποίηση της στην γλώσσα python.

5.2 Ανάλυση κύριων συνιστωσών (ΑΚΣ)

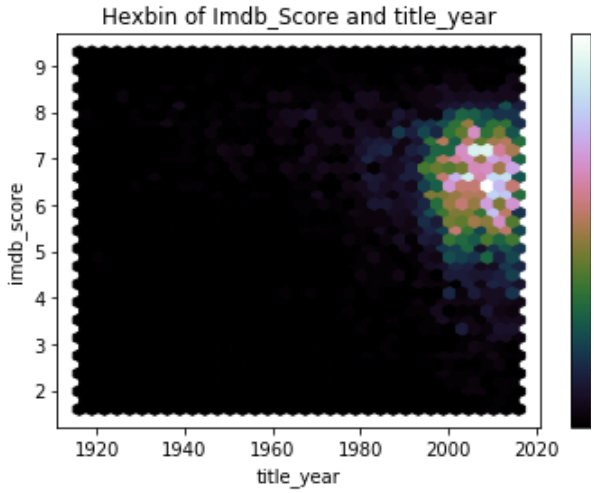
Αρχικά αναλύεται η βάση movie_metadata.csv η οποία περιλαμβάνει 28 γνωρίσματα και 5043 εγγραφές, αφαιρούνται τα γνωρίσματα που δεν έχουν αριθμητικές τιμές, οπότε απομένουν τα υπόλοιπα 16 [1]. Ο λόγος για τον οποίο χρειάζονται τα αριθμητικά γνωρίσματα και όχι τα υπόλοιπα, είναι πως για την χρήση της μεθόδου ΑΚΣ είναι απαραίτητη η κανονικοποίηση των γνωρισμάτων, η εύρεση ιδιοτιμών και ιδιοδιανυσμάτων τα οποία μπορούν να υπολογιστούν από αριθμητικά δεδομένα. Ακόμη τα γνωρίσματα gross, num_user_for_reviews, num_critic_for_reviews, num_voted_users αφαιρούνται από την βάση καθώς δεν είναι δυνατό να είναι γνωστές οι τιμές τους πριν την κυκλοφορία μιας ταινίας. Στη συνέχεια διαγράφονται τα στοιχεία των οποίων η τιμή είναι NA (δηλαδή λείπει η τιμή) και γίνεται κανονικοποίηση. Αφού γίνουν τα παραπάνω αναλύονται οι συσχετίσεις των γνωρισμάτων δημιουργώντας γραφήματα hexbin και Pearson. Τα γραφήματα φαίνονται στην εικόνα 5.5.



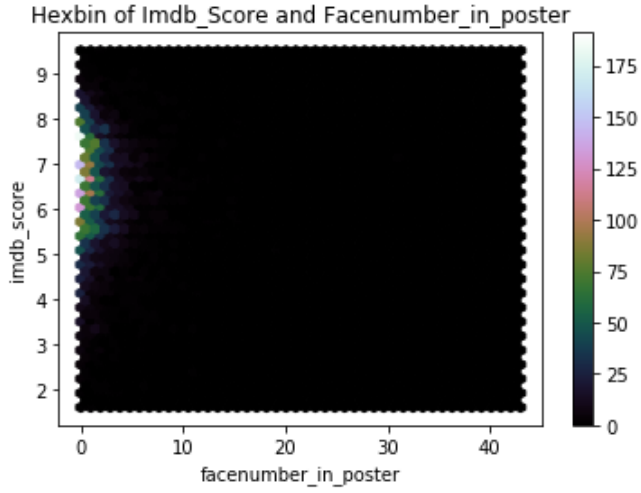
Σχήμα 5.1: Το γνώρισμα duration.



Σχήμα 5.2: Το γνώρισμα aspect_ratio.



Σχήμα 5.3: Το γνώρισμα title_year.

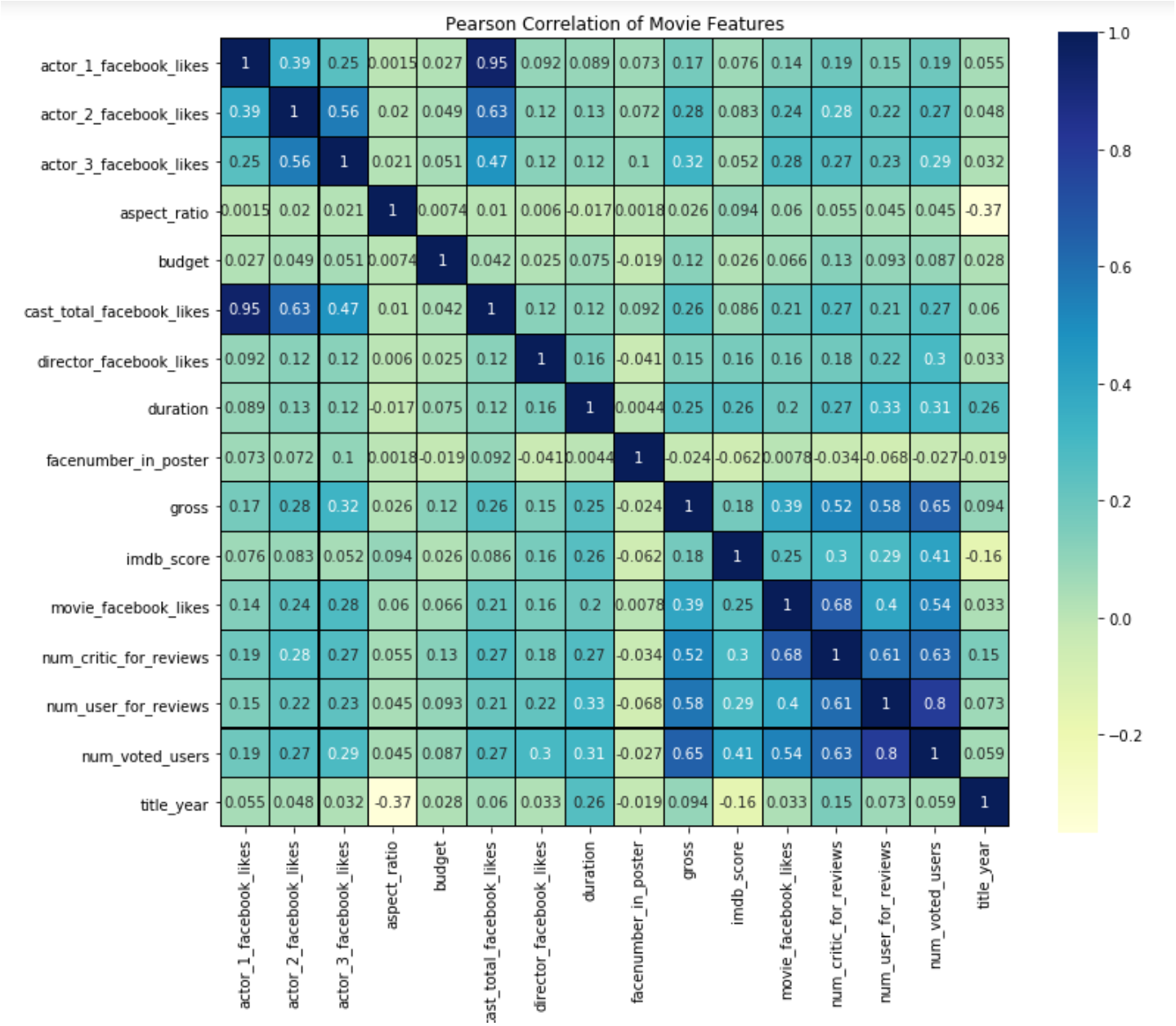


Σχήμα 5.4: Το γνώρισμα facenumber_in_poster.

Σχήμα 5.5: Γραφήματα hexbin για τα γνωρίσματα duration, aspect_ratio, title_year και facenumber_in_poster.

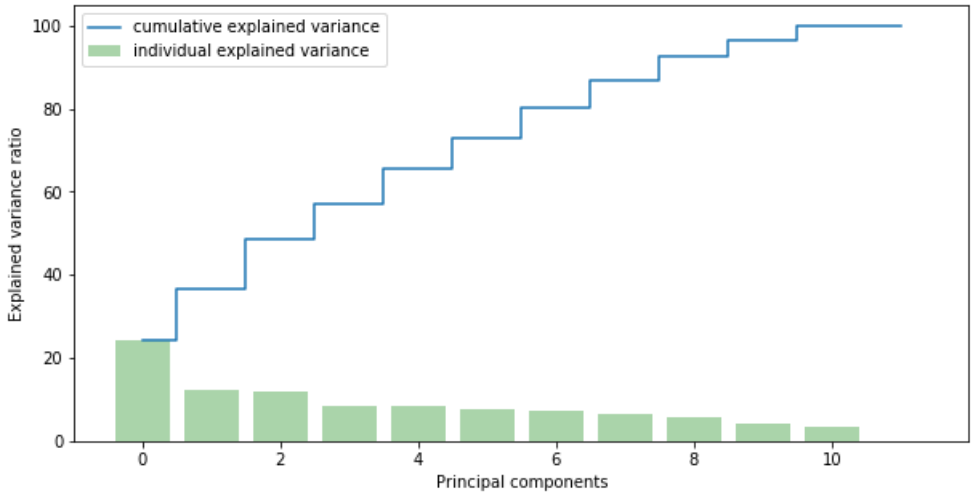
Τα γραφήματα hexbin δημιουργούνται για να παρατηρηθεί η κατανομή των εγγραφών κάποιων γνωρισμάτων για το αν μπορούν να επηρεάσουν το γνώρισμα imdb_score. Είναι φανερό ότι υπάρχουν φωτεινές περιοχές κοντά στις βαθμολογίες 6 και 7, το οποίο ερμηνεύεται πως υπάρχουν περισσότερες εγγραφές με αυτές τις τιμές.

Στο γράφημα συσχετισμού του Pearson είναι δυνατή η κατανόηση για το ποια γνωρίσματα συσχετίζονται μεταξύ τους. Στην εικόνα 5.6 τα τετράγωνα που έχουν πιο σκούρο χρώμα αντιστοιχούν σε γνωρίσματα με μεγαλύτερη συσχέτιση. Αν δύο γνωρίσματα έχουν μεγάλο βαθμό συσχέτισης, επηρεάζουν με παρόμοιο τρόπο το αποτέλεσμα και άρα δεν χρειάζεται να χρησιμοποιηθούν και τα δυο στο μοντέλο πρόβλεψης. Επειδή υπάρχουν γνωρίσματα με υψηλή μεταξύ τους συσχέτιση χρησιμοποιείται η μέθοδος της ανάλυσης κύριων συνιστωσών για να δημιουργηθούν γνωρίσματα τα οποία είναι κάθετα μεταξύ τους.



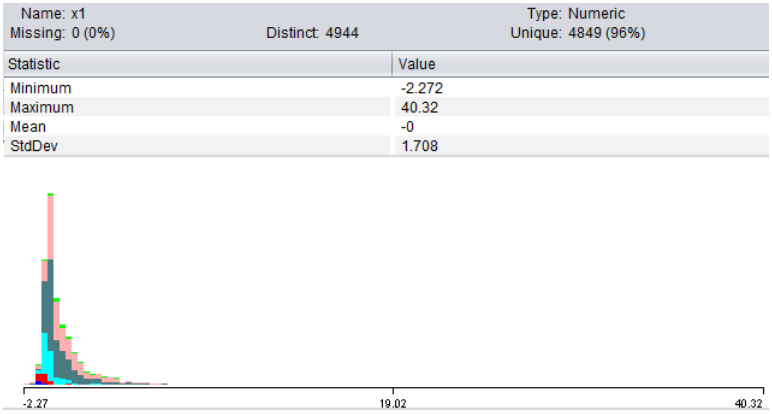
Σχήμα 5.6: Γράφημα συσχετισμού Pearson

Στην συνέχεια της διαδικασίας υπολογίζονται τα ιδιοδιανύσματα και οι ιδιοτιμές του πίνακα καθώς αυτά απαιτούνται για την ανάλυση μας (ΑΚΣ). Τα ιδιοδιανύσματα ερμηνεύουν τις κατευθύνσεις του νέου χώρου γνωρισμάτων, οι ιδιοτιμές το μέγεθος τους ενώ η διακύμανση δείχνει πόσα στοιχεία μπορούν να αποδοθούν σε καθένα από τα κύρια στοιχεία.

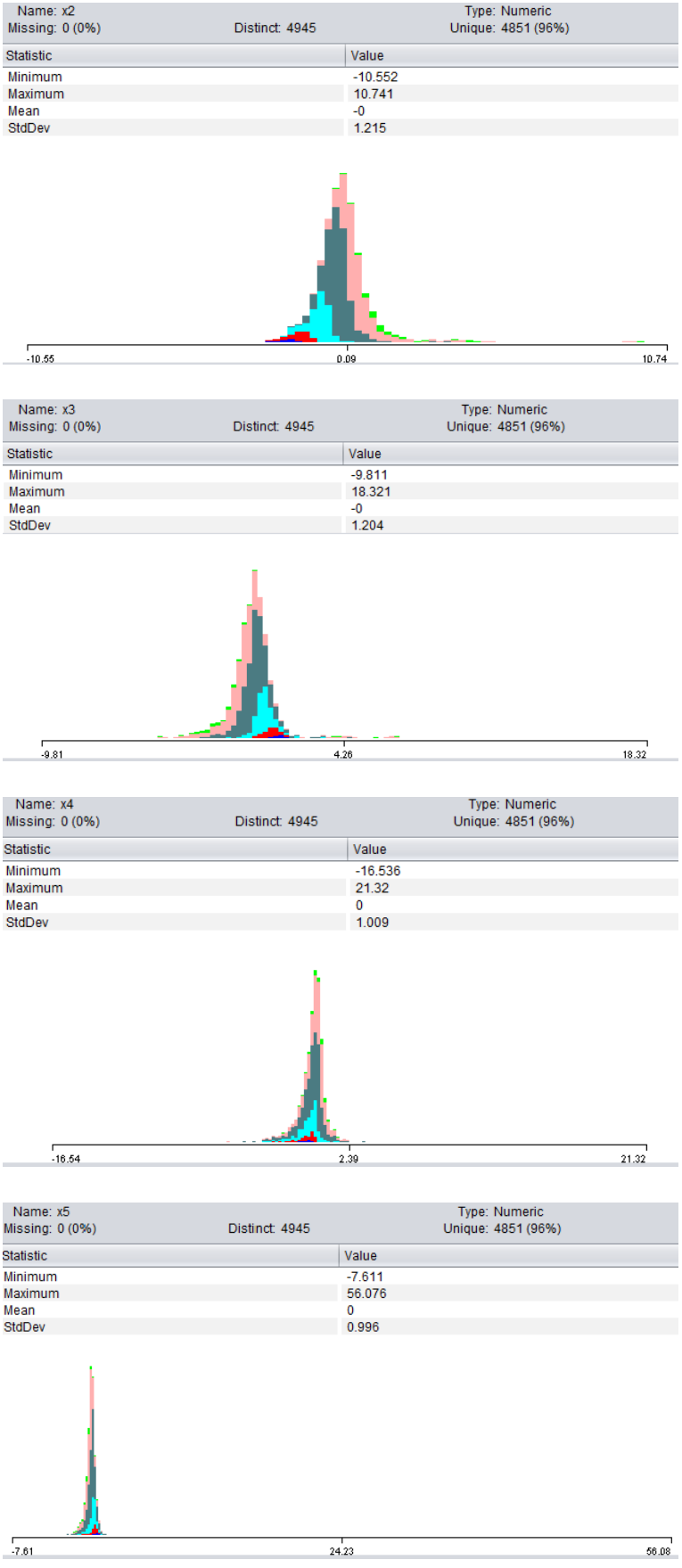


Σχήμα 5.7: Γράφημα που δείχνει τη διακύμανση των κύριων συνιστωσών

Στο γράφημα στην εικόνα 5.7 με πράσινο χρώμα παρουσιάζεται η διακύμανση της κάθε κύριας συνιστώσας που δημιουργείται από την ανάλυση κύριων συνιστωσών. Η μπλέ γραμμή απεικονίζει το άθροισμα (άξονας y) των x κύριων συνιστωσών (άξονας x). Από το γράφημα είναι φανερό ότι το 90% της διακύμανσης μπορεί να αντικατασταθεί από τις 8 κύριες συνιστώσες. Έτσι κρατώντας τις 8 κύριες συνιστώσες εφαρμόζουμε την Ανάλυση Κύριων Συνιστωσών. Προστίθεται σαν γνώρισμα η κλάση που θα γίνει η κατηγοριοποίηση (imdb_score) και διακριτοποιείται σε έξι κλάσεις. Στις εικόνες 5.8, 5.9 και 5.10 φαίνονται αναλυτικά οι οχτώ κύριες συνιστώσες που έχουν δημιουργηθεί και πώς είναι κατανεμημένες οι εγγραφές τους.



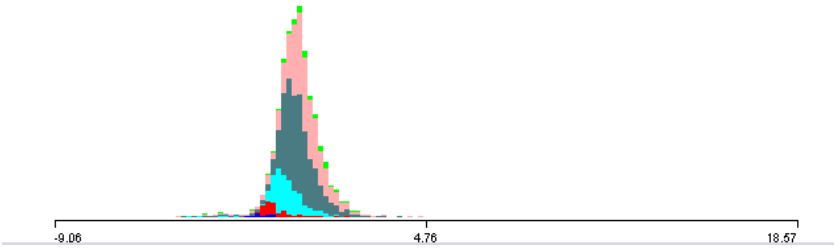
Σχήμα 5.8: Η πρώτη κύρια συνιστώσα από την ανάλυση PCA.



Σχήμα 5.9: Οι 2η, 3η, 4η και 5η κύριες συνιστώσες από την ανάλυση PCA.

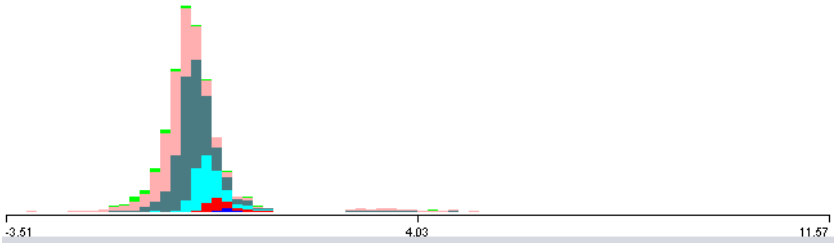
Name: x6
Missing: 0 (0%)
Distinct: 4945
Type: Numeric
Unique: 4851 (96%)

Statistic	Value
Minimum	-9.056
Maximum	18.569
Mean	0
StdDev	0.96



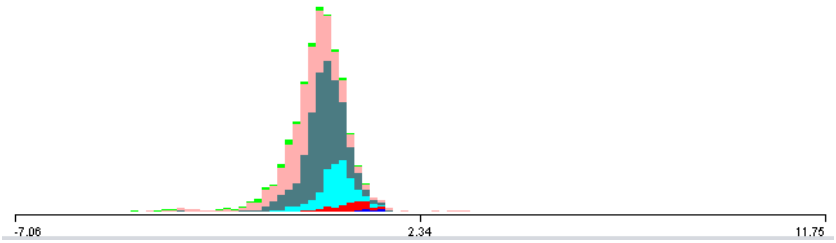
Name: x7
Missing: 0 (0%)
Distinct: 4945
Type: Numeric
Unique: 4851 (96%)

Statistic	Value
Minimum	-3.51
Maximum	11.572
Mean	0
StdDev	0.927



Name: x8
Missing: 0 (0%)
Distinct: 4944
Type: Numeric
Unique: 4849 (96%)

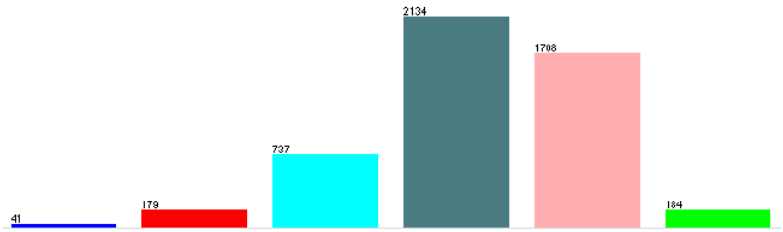
Statistic	Value
Minimum	-7.065
Maximum	11.746
Mean	0
StdDev	0.881



Σχήμα 5.10: Οι 6η, 7η και 8η κύριες συνιστώσες από την ανάλυση PCA.

Name: imdb_score Type: Nominal
Missing: 0 (0%) Distinct: 6 Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-2.916667]'	41	41.0
2	'(2.916667-4.233333]'	179	179.0
3	'(4.233333-5.55]'	737	737.0
4	'(5.55-6.866667]'	2134	2134.0
5	'(6.866667-8.183333]'	1768	1768.0
6	'(8.183333-inf)'	184	184.0



Σχήμα 5.11: Γραφήμα για το πως κατανέμονται οι εγγραφές της κλάσης imdb_score

Τέλος, γίνονται οι δοκιμές των αλγορίθμων και τα αποτελεματα φαίνονται στον πίνακα 5.1. Στην δεύτερη στήλη ο αριθμός αναπαριστά το ποσοστό επιτυχίας της πρόβλεψης με την μέθοδο cross validation. Παρατηρείται πως ο αλγόριθμος Naive Bayes έχει πολύ κακή απόδοση, χειρότερη από αυτή της προηγούμενης προσέγγισης. Επίσης ο SMO έχει και αυτός χαμηλό ποσοστό επιτυχημένης πρόβλεψης. Οι αλγόριθμοι LogitBoost και DecisionTable κυμαίνονται σε ποσοστά όμοια με αυτά της προηγούμενης προσέγγισης. Οι υπόλοιποι αλγόριθμοι τρέχουν με πολύ καλύτερα αποτελέσματα, ο καλύτερος από τους οποίους είναι ο Random Forest με ποσοστό επιτυχημένης πρόβλεψης 87.55.

Πίνακας 5.1: Αποτελέσματα από την PCA ανάλυση

Αλγόριθμος	Αποτέλεσμα
Naive Bayes	54.79
Logistic	84.49
Multilayer Perceptron	85.88
SMO	65.73
ClassificationViaRegression	84.81
LogitBoost	74.8
DecisionTable	75.45
J48	81.92
RandomForest	87.55

Κεφάλαιο 6

Μελέτη των συντελεστών - Τρίτη προσέγγιση του προβλήματος

6.1 Εισαγωγή

Σε αυτήν την τρίτη προσέγγιση γίνεται προσπάθεια εκμετάλλευσης των συμπερασμάτων που προέκυψαν από τα προηγούμενα πειράματα, αναφορικά με το ποια φαίνεται να είναι τα γνωρίσματα που επηρεάζουν/καθορίζουν το αποτέλεσμα. Το συμπέρασμα που προέκυψε από τις δύο προηγούμενες προσεγγίσεις είναι πως οι ηθοποιοί, οι σκηνοθέτες καθώς και οι λέξεις κλειδιά μιας ταινίας παίζουν κύριο ρόλο στο πόσο επιτυχημένη θα είναι μια ταινία. Συνεπώς κύριος στόχος αυτής της προσέγγισης είναι να γίνει καλύτερη και ακριβέστερη ταξινόμηση των γνωρισμάτων αυτών. Η προσέγγιση ολοκληρώνεται με πειράματα για να διαπιστωθεί κατα πόσο αυτή η υπόθεση είναι βάσιμη και κατα πόσο είναι αποδοτική στην τελική πρόβλεψη.

6.2 Προεπεξεργασία του αρχείου academy_awards.csv

Για την καλύτερη ταξινόμηση των ηθοποιών και των σκηνοθετών χρησιμοποιείται η βάση του αρχείου academy_awards.csv η οποία παρέχει πληροφορίες για τα όσκαρ και τις υποψηφιότητες τους. Αρχική σκέψη είναι να συνδυαστούν τα γνωρίσματα αυτά μάζι με τις άλλες παραμέτρους από την προηγούμενη βάση για να επιτευχθεί η ακριβέστερη κατάταξη και βαθμολόγηση των ηθοποιών-σκηνοθετών. Η κατάταξη αυτή θα χρησιμοποιηθεί στη συνέχεια για την πρόβλεψη της επιτυχίας της ταινίας. Στον πίνακα 6.1 φαίνονται τα γνωρίσματα των δεδομένων της βάσης academy_awards.csv.

Οι αρχικές εγγραφές του αρχείου είναι 9964. Η βάση περιέχει τις βραβεύσεις από το 1927. Εφόσον στη μελέτη δεν συμμετέχουν οι ηθοποιοί από αρκετά παλαιότερες χρονολογίες αφαιρούνται οι εγγραφές που αναφέρονται σε βραβεύσεις πριν το 1970. Για να δημιουργηθεί το αρχείο των υποψηφιοτήτων και των βραβείων όσκαρ των ηθοποιών, επιλέγονται οι γραμμές στις οποίες το γνώρισμα “Award” έχει τις τιμές : “Actor”, “Actress”, “Actor in a Leading Role”, “Actor in a Supporting Role”, “Actress in a Leading Role” και “Actress in a Supporting Role”. Έπειτα κρατιούνται μόνο τα γνωρίσματα “Winner” και “Name”, όπου οι τιμές του γνωρίσματος “Winner”

Όνομα μεταβλητής	Περιγραφή
Year	The year of the event
Ceremony	oscars
Award	Actor Actress
Winner	Did the actor/actress win?
Name	name of actor/actress
Film	name of film

Πίνακας 6.1: Τα γνωρίσματα της βάσης από το αρχείο academy_awards.csv.

είναι NA και 1 ανάλογα αν έχει νικήσει ή όχι. Αθροίζονται με βάση τα ονόματα και έτσι βγαίνουν σε μία λίστα τα όσκαρ που έχει κερδίσει ο καθένας ηθοποιός. Με παρόμοιο τρόπο υπολογίζονται και οι υποψηφιότητες του κάθε ηθοποιού με τη μόνη διαφορά ότι αντικαθίστανται στο γνώρισμα ”Winner” οι τιμές που έχουν τιμή NA με την τιμή 1. Τέλος συνενώνονται όλα τα στοιχεία σε ένα νέο αρχείο actors_nom_oscars.csv το οποίο έχει σαν γνωρίσματα τα πεδία Name, Nominations και Oscars. Επαναλαμβάνουμε την ίδια διαδικασία για τους σκηνοθέτες και προκύπτει το αρχείο director_oscars.csv που έχει γνωρίσματα τα πεδία Name, Nominations και Oscars.

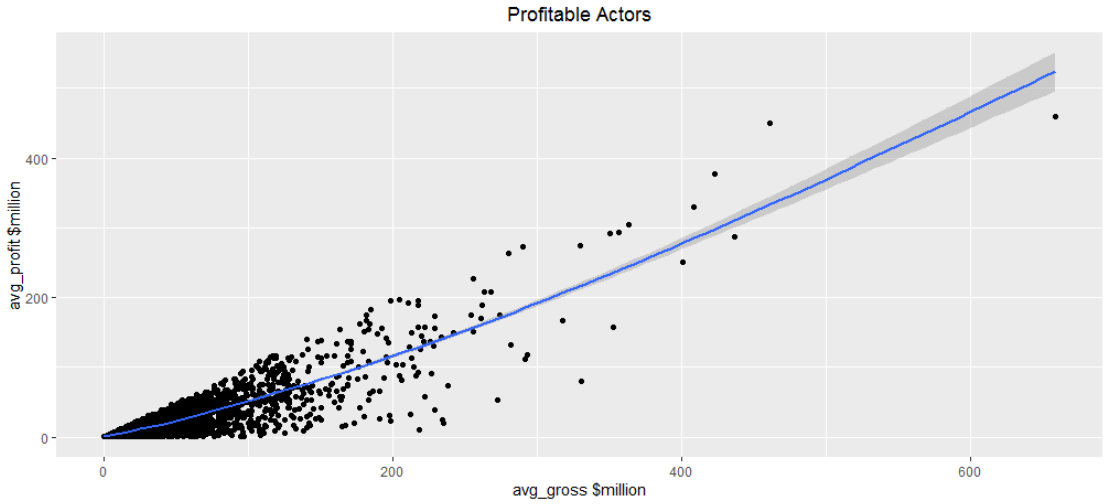
6.3 Προεπεξεργασία ηθοποιών από την αρχική βάση

Αρχικά σβήνονται τα διπλότυπα από όλο το αρχείο και διαγράφονται όλες οι γραμμές στις οποίες τα γνώρισματα gross και budget έχουν την τιμή NA. Έπειτα αποθηκεύονται σε μια δομή (data frame) actor1 τα γνωρίσματα actor_1_name, actor_1_facebook_likes, gross, budget, imdb_score και δημιουργείται ένα επιπλέον γνώρισμα profit = gross - budget. Το επόμενο βήμα είναι να διαφοροποιηθούν τα γνωρίσματα της δομής actor1. Για κάθε ξεχωριστό όνομα ηθοποιού που βρίσκεται στη δομή και για το σύνολο των ταινιών όπου συμμετέχει, υπολογίζονται τα γνωρίσματα:

- facebook_likes: ο μέγιστος αριθμός από likes με τον οποίο εμφανίζεται ο ηθοποιός
- total_gross: το άρθροισμα όλων των τιμών του γνωρίσματος gross
- avg_gross: ο μέσος όρος των τιμών του γνωρίσματος gross
- total_profit: το άθροισμα των τιμών των γνωρισμάτων profit
- avg_profit: ο μέσος όρος των τιμών των γνωρισμάτων profit
- avg_imdb_score: ο μέσος όρος των τιμών των γνωρισμάτων imdb_score
- top_imdb_score: το μέγιστο των τιμών για το γνώρισμα imdb_score

Η ίδια διαδικασία γίνεται αντίστοιχα και για τους actor2 και actor3. Στο τέλος συνενώνονται οι δομές των actor1, actor2 και actor3 σε μία και τα δεδομένα αποθηκεύονται στο αρχείο actors.csv.

Τα αρχεία actors_nom_oscars.csv και actors.csv συνενώνονται. Πλέον η βάση για την κατηγοριοποίηση των ηθοποιών έχει 4600 εγγραφές και τα εξής γνωρίσματα: actor_name, facebook_likes, total_gross, avg_gross, total_profit, avg_profit, avg_imdb_score, top_imdb_score, nominations και oscars.



Σχήμα 6.1: Γράφημα που δείχνει το γνώρισμα avg_profit σε σχέση με το ”avg_gross”

Στη συνέχεια, μελετώντας τις συσχετίσεις των γνωρισμάτων βγήκε το συμπέρασμα πως τα γνωρίσματα avg_gross και avg_profit έχουν γραμμική συσχέτιση όπως φαίνεται και στο σχήμα 6.1. Οπότε η παρουσία του ενός από τα δύο γνωρίσματα μάλλον είναι περιττή.

Για όλους τους ηθοποιοιούς γίνεται κανονικοποίηση στην κλίμακα του δέκα (δηλαδή διαιρούνται με το μέγιστο και πολλαπλασιάζονται επί δέκα) τα γνωρίσματα: total_gross, avg_gross, total_profit και avg_profit. Επειδή το μέγιστο σκορ ηθοποιού που εμφανίζεται είναι κοντά στο 6.5, προστίθενται επιπλέον σαν παραμέτροι τα: nominations, oscars και facebook_likes οπότε ο καλύτερος ηθοποιός φθάνει κοντά στο δέκα.

Οι ίδιοι παράγοντες χρησιμοποιούνται και για τους σκηνοθέτες. Κανονικοποιούνται στην κλίμακα του 10 τα γνωρίσματα: total_gross, avg_gross, total_profit και avg_profit. Παρόμοια με τους ηθοποιούς προστίθενται και οι παράμετροι nominations και oscars. Σε αυτή τη περίπτωση δεν προστίθεται η παράμετρος των facebook_likes καθώς οι σκηνοθέτες δεν εκτίθενται στη διασημότητα όπως οι ηθοποιοί.

Οι συντελεστές των γνωρισμάτων δεν είναι σταθεροί αλλά αναζητούνται μέσα από τα πειράματα οι καλύτεροι δυνατοί καθώς και τα γνωρίσματα που έχουν μεγαλύτερη επιρροή στο αποτέλεσμα. Το γράφημα του Pearson 5.6 βοηθάει στην κατανόηση για το ποια γνωρίσματα έχουν μεγάλη συσχέτιση ώστε να αξιοποιηθεί στα πειράματα.

Οι εξισώσεις από τις οποίες γίνεται η κατηγοριοποίηση των ηθοποιών και των σκηνοθετών είναι οι ακόλουθες:

$$actor\_score = top\_imdb\_score + avg\_imdb\_score + total\_gross + avg\_gross + total\_profit + avg\_profit + nominations + oscars + (0.5 \quad \text{αν} \quad facebool\_likes >= 1000) \quad (6.1)$$

$$director\_score = top\_imdb\_score + avg\_imdb\_score + total\_gross + avg\_gross + total\_profit + avg\_profit + nominations + oscars \quad (6.2)$$

Επίσης προστίθεται ως γνώρισμα οι λέξεις κλειδιά (plot_keywords). Η κατηγοριοποίηση της κάθε λέξης αυτού του γνωρίσματος έχει γίνει υπολογίζοντας τον μέσο όρο των imdb_score των ταινιών στις οποίες εμφανίζονται και έπειτα υπολογίζεται ο μέσος όρος των λέξεων στην ταινία. Για παράδειγμα, αν το γνώρισμα plot_keyword μιας ταινίας είναι το: avatar, future, marine, native, paraplegic και οι μέσοι όροι των λέξεων είναι 6.05, 6.92, 6.275, 7.9, 7.9 αντίστοιχα, τότε για την συγκεκριμένη ταινία το γνώρισμα plot_keywords_score θα έχει την τιμή (6.05 + 6.92 + 6.275 + 7.9 + 7.9) / 5 = 7. Άρα πλέον τα γνωρίσματα που χρησιμοποιούμε είναι actor_1_score, actor_2_score, actor_3_score, director_score και plot_keywords_score.

Στους παρακάτω πίνακες παρουσιάζονται τα πειράματα για την δοκιμή των συντελεστών των ηθοποιών και των σκηνοθετών. Χρησιμοποιείται και το plot_keywords_score αλλά δεν αλλάζει κάτι στους συντελεστές του για αυτό το λόγο δεν φαίνεται στον πίνακα. Το γνώρισμα imdb_score κατηγοριοποιείται σε 6 κλάσεις όπως και στις προηγούμενες προσεγγίσεις.

	Πείραμα 1ο	Πείραμα 1ο	Πείραμα 2ο	Πείραμα 2ο
	actor_score	director_score	actor_score	director_score
top_imdb_score	0.2	0.2	0	0
avg_imdb_score	0.6	0.6	0.8	0.8
total_gross	0.09	0.1125	0.09	0.1125
avg_gross	0.06	0.0375	0.06	0.0375
total_profit	0.03	0.0375	0.03	0.0375
avg_profit	0.02	0.0125	0.02	0.0125
nominations	0.2	0.2	0.2	0.2
oscars	0.1	0.1	0.1	0.1
NaiveBayes	74.14		74.97	
Logistic	75.51		75.48	
MultilayerPerceptron	76.2		76.4	
SMO	72.98		73.69	
ClassificationViaRegression	75.17		74.66	
LogitBoost	78.82		78.5	
DecisionTable	75.11		75.29	
J48	77.7		78.76	
RandomForest	80.13		80.59	

	Πείραμα 3ο	Πείραμα 3ο	Πείραμα 4ο	Πείραμα 4ο
	actor_score	director_score	actor_score	director_score
top_imdb_score	0.8	0.8	0.4	0.4
avg_imdb_score	0	0	0.4	0.4
total_gross	0.09	0.1125	0.09	0.1125
avg_gross	0.06	0.0375	0.06	0.0375
total_profit	0.03	0.0375	0.03	0.0375
avg_profit	0.02	0.0125	0.02	0.0125
nominations	0.2	0.2	0.2	0.2
oscars	0.1	0.1	0.1	0.1
NaiveBayes	72.95		74	
Logistic	73.92		75.2	
MultilayerPerceptron	75.94		76.23	
SMO	71.49		72.46	
ClassificationViaRegression	79.9		75.66	
LogitBoost	82.9		80.02	
DecisionTable	80.27		75.48	
J48	82.61		78.62	
RandomForest	84.98		81.1	

	Πείραμα 5ο	Πείραμα 5ο	Πείραμα 6ο	Πείραμα 6ο
	actor_score	director_score	actor_score	director_score
top_imdb_score	0.8	0.8	0.8	0.8
avg_imdb_score	0	0	0	0
total_gross	0	0	0.15	0.15
avg_gross	0.15	0.15	0	0
total_profit	0.03	0.0375	0.03	0.0375
avg_profit	0.02	0.0125	0.02	0.0125
nominations	0.2	0.2	0.2	0.2
oscars	0.1	0.1	0.1	0.1
NaiveBayes	73.69		72.92	
Logistic	73.97		73.77	
MultilayerPerceptron	75.94		76.25	
SMO	71.6		71.32	
ClassificationViaRegression	78.28		80.7	
LogitBoost	81.21		83.4	
DecisionTable	78.62		81.7	
J48	81.16		82.95	
RandomForest	83.32		85.83	

	Πείραμα 7ο	Πείραμα 7ο	Πείραμα 8ο	Πείραμα 8ο
	actor_score	director_score	actor_score	director_score
top_imdb_score	0.8	0.8	0.8	0.8
avg_imdb_score	0	0	0	0
total_gross	0.075	0.075	0.15	0.15
avg_gross	0.075	0.075	0	0
total_profit	0.03	0.0375	0.05	0.05
avg_profit	0.02	0.0125	0	0
nominations	0.2	0.2	0.2	0.2
oscars	0.1	0.1	0.1	0.1
NaiveBayes	73.18		73.11	
Logistic	73.88		73.83	
MultilayerPerceptron	76.08		76.48	
SMO	71.52		71.29	
ClassificationViaRegression	78.14		80.05	
LogitBoost	82.33		84.2	
DecisionTable	79.45		81.41	
J48	81.33		83.15	
RandomForest	84.94		86.15	

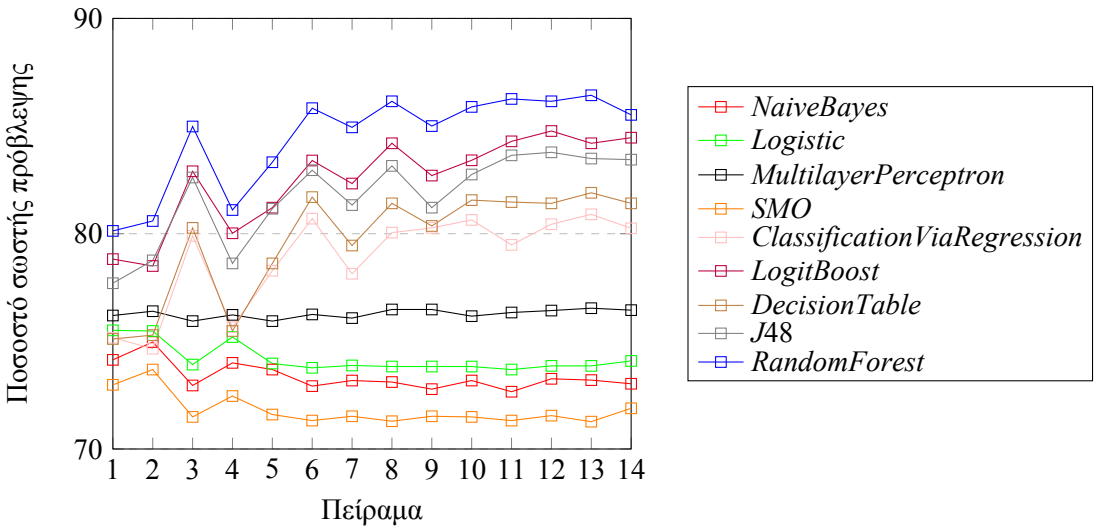
	Πείραμα 9ο	Πείραμα 9ο	Πείραμα 10ο	Πείραμα 10ο
	actor_score	director_score	actor_score	director_score
top_imdb_score	0.8	0.8	0.8	0.8
avg_imdb_score	0	0	0	0
total_gross	0.15	0.15	0.15	0.15
avg_gross	0	0	0	0
total_profit	0	0	0.025	0.025
avg_profit	0.05	0.05	0.025	0.025
nominations	0.2	0.2	0.2	0.2
oscars	0.1	0.1	0.1	0.1
NaiveBayes	72.78		73.18	
Logistic	73.83		73.83	
MultilayerPerceptron	76.48		76.17	
SMO	71.52		71.49	
ClassificationViaRegression	80.27		80.64	
LogitBoost	82.7		83.41	
DecisionTable	80.36		81.56	
J48	81.21		82.75	
RandomForest	85		85.89	

	Πείραμα 11ο	Πείραμα 11ο	Πείραμα 12ο	Πείραμα 12ο
	actor_score	director_score	actor_score	director_score
top_imdb_score	0.8	0.8	0.8	0.8
avg_imdb_score	0	0	0	0
total_gross	0.15	0.15	0.15	0.15
avg_gross	0	0	0	0
total_profit	0.05	0.05	0.05	0.05
avg_profit	0	0	0	0
nominations	0.3	0.3	0	0
oscars	0	0	0.3	0.3
NaiveBayes	72.66		73.26	
Logistic	73.69		73.86	
MultilayerPerceptron	76.34		76.43	
SMO	71.32		71.55	
ClassificationViaRegression	79.48		80.44	
LogitBoost	84.29		84.77	
DecisionTable	81.47		81.41	
J48	83.64		83.78	
RandomForest	86.26		86.15	

	Πείραμα 13ο	Πείραμα 13ο	Πείραμα 14ο	Πείραμα 14ο
	actor_score	director_score	actor_score	director_score
top_imdb_score	0.8	0.8	0.8	0.8
avg_imdb_score	0	0	0	0
total_gross	0.2	0.2	0	0
avg_gross	0	0	0	0
total_profit	0	0	0.2	0.2
avg_profit	0	0	0	0
nominations	0.15	0.15	0.15	0.15
oscars	0.15	0.15	0.15	0.15
NaiveBayes	73.2		73.03	
Logistic	73.86		74.09	
MultilayerPerceptron	76.54		76.45	
SMO	71.27		71.89	
ClassificationViaRegression	80.9		80.25	
LogitBoost	84.2		84.46	
DecisionTable	81.9		81.41	
J48	83.49		83.44	
RandomForest	86.43		85.52	

Στα τέσσερα πρώτα πειράματα αλλάζει μόνο ο συντελεστής των παραμέτρων top_imdb_score και avg_imdb_score ενώ οι υπόλοιποι συντελεστές είναι σταθεροί. Παρατηρείται πως το τρίτο πείραμα έχει καλύτερα αποτελέσματα σε σχέση με τα άλλα τρία, καθώς στους περισσότερους αλγορίθμους έχει καλύτερες αποδόσεις συνεπώς διατηρούνται οι συντελεστές αυτού του πειράματος για τα δύο αυτά γνωρίσματα στα επόμενα πειράματα. Στη συνέχεια, στα πειράματα πέντε, έξι και εφτά δοκιμάζονται οι συντελεστές των total_gross και avg_gross. Καλύτερα αποτελέσματα φαίνεται να έχουν οι αλγόριθμοι στο πείραμα έξι. Έχοντας τους πρώτους τέσσερις συντελεστές, στα πειράματα οχτώ, εννιά και δέκα αλλάζουν οι συντελεστές των total_profit και avg_profit. Από τα τρία αυτά πειράματα κρατιούνται οι συντελεστές του πειράματος οχτώ. Επίσης από τα πειράματα έντεκα και δώδεκα, που αλλάζουν οι συντελεστές για τα nominations και τα oscars, φαίνεται οι δύο παράμετροι να επηρεάζουν παρόμοια το αποτέλεσμα. Τέλος, είχαμε παρατηρήσει πως τα γνωρίσματα profit και gross έχουν γραμμική συσχέτιση στο σχήμα 6.1, και για αυτό γίνονται τα πειράματα δεκατρία και δεκατέσσερα μηδενίζοντας το gross (πείραμα 13) και το profit (πείραμα 14).

Στη συνέχεια, στο σχήμα 6.2 φαίνεται η συμπεριφορά των αλγορίθμων σε όλα τα πειράματα. Στον οριζόντιο άξονα απεικονίζεται ο αριθμός του πειράματος και στον κάθετο άξονα απεικονίζεται η απόδοση του αλγορίθμου. Στο παράρτημα στα δεξιά γίνεται φανερός ο αλγόριθμος με το αντίστοιχο χρώμα. Όπως μπορεί να γίνει εύκολα αντιληπτό ο αλγόριθμος RandomForest έχει τις καλύτερες αποδόσεις για τα συγκεκριμένα πειράματα με την μεγίστη απόδοση να φτάνει στο 86.43 στο πείραμα δεκατρία, η οποία πλησιάζει το αποτέλεσμα της δεύτερης προσέγγισης. Επίσης φαίνεται πως ο αλγόριθμος SMO έχει κατά μέσο όρο τις χειρότερες αποδόσεις. Συνολικά, στο πείραμα δεκατρία οι αλγόριθμοι έχουν κατα μέσο όρο την καλύτερη συμπεριφορά και αυτοί φαίνεται να είναι οι ιδανικότεροι συντελεστές για τις παραμέτρους. Οι παράμετροι που επηρεάζουν περισσότερο το αποτέλεσμα είναι οι: top_imdb_score, total_gross, nominations και τα oscars.



Σχήμα 6.2: Γραφική που δείχνει την συμπεριφορά των αλγωρίθμων

Κεφάλαιο 7

Τεχνικές λεπτομέρειες

Στο κεφάλαιο αυτό γίνεται μία περιγραφή για τα εργαλεία που χρησιμοποιήθηκαν στην διπλωματική εργασία. Αναφέρεται η διαδικασία χρήσης των εργαλειών για την προεπεξεργασία και την δοκιμή των πειραμάτων. Επίσης δίνονται κάποιες γενικές τεχνικές πληροφορίες για το πρόγραμμα WEKA, την γλώσσα R και την πλατφόρμα RStudio και το περιβάλλον του Jupyter και Jupyter Notebook ειδικότερα.

7.1 Λεπτομέρειες υλοποίησης

Για την πρώτη και την τρίτη προσέγγιση του προβλήματος, η προεπεξεργασία των δεδομένων έγινε στην προγραμματιστική πλατφόρμα RStudio. Αφού έγιναν οι διαδικασίες που περιγράφονται στα κεφάλαια 4 και 6, τα επεξεργασμένα δεδομένα γράφτηκαν σε αρχεία ”.csv” και ”.arff”. Στη συνέχεια, τα αρχεία αυτά διαβάστηκαν από την πλατφόρμα WEKA, χρησιμοποιήθηκε το φίλτρο της διακριτοποίησης για το γνώρισμα της κλάσης και δοκιμάστηκαν οι αλγόριθμοι που παρουσιάζονται στα προηγούμενα κεφάλαια. Για την δεύτερη προσέγγιση η διαδικασία της επεξεργασίας των δεδομένων και της εφαρμογής της μεθόδου των κύριων συνιστωσών πραγματοποιήθηκε στο εργαλείο Jupyter Notebook, από όπου προέκυψαν και τα αντίστοιχα γραφήματα.

7.2 Πλατφόρμες και προγραμματιστικά εργαλεία

7.2.1 RStudio

Η R είναι ένα περιβάλλον λογισμικού ανοιχτού κώδικα για στατιστικούς υπολογισμούς και γραφικά. Η R μεταγλωττίζεται και εκτελείται σε Windows, Mac OS X και σε πολλές πλατφόρμες UNIX (όπως Linux). Για τις περισσότερες πλατφόρμες, η R διανέμεται σε δυαδική μορφή για ευκολία εγκατάστασης. Το πρόγραμμα λογισμικού R ξεκίνησε για πρώτη φορά από τους Robert Gentleman και Ross Ihaka. Η γλώσσα επηρεάστηκε σε μεγάλο βαθμό από τη γλώσσα S, η οποία αρχικά αναπτύχθηκε στο Bell Laboratories από τον John Chambers και τους συναδέλφους του. Έκτοτε, με την καθοδήγηση και τα ταλέντα της βασικής ομάδας ανάπτυξης της R, η R έχει εξελιχθεί σε κοινή γλώσσα για στατιστικούς υπολογισμούς σε πολλούς κλάδους της ακαδημαϊκής

κοινότητας και σε διάφορες βιομηχανίες. Η R είναι κάτι πολύ περισσότερο από βασική γλώσσα. Διαθέτει ένα παγκόσμιο σύστημα αποθετηρίου, το Comprehensive R Archive Network (CRAN) —http: //cran.r-project.org— για πακέτα πρόσθετων χρηστών που έχουν συνεισφέρει για τη συμπλήρωση της βασικής διανομής. Από το 2011, υπήρχαν περισσότερα από 3.000 τέτοια πακέτα που φιλοξενούνται στο CRAN και πολλά άλλα σε άλλους ιστότοπους. Συνολικά, η R έχει επί του παρόντος λειτουργικότητα για να αντιμετωπίσει ένα τεράστιο φάσμα προβλημάτων και έχει ακόμη περιθώριο ανάπτυξης. Η R έχει σχεδιαστεί γύρω από τη βασική γλώσσα δέσμης ενεργειών, αλλά επιτρέπει επίσης την ενσωμάτωση με μεταγλωττισμένο κώδικα γραμμένο σε C, C ++, Fortran, Java κ.λπ., για υπολογιστικά εντατικές εργασίες ή για εργαλεία μόχλευσης που παρέχονται για άλλες γλώσσες [19].

Το RStudio είναι ένα ολοκληρωμένο περιβάλλον ανάπτυξης (IDE) για την R, μια γλώσσα προγραμματισμού για στατιστικούς υπολογιστές και γραφικά. Διατίθεται σε δύο μορφές: Το RStudio Desktop είναι μια κανονική εφαρμογή για επιτραπέζιους υπολογιστές, ενώ ο διακομιστής RStudio εκτελείται σε έναν απομακρυσμένο διακομιστή και επιτρέπει την πρόσβαση στο RStudio χρησιμοποιώντας ένα πρόγραμμα περιήγησης στο Web. Το RStudio IDE είναι διαθέσιμο με την GNU Affero General Public License έκδοση 3. Το AGPL v3 είναι μια άδεια ανοιχτού κώδικα που εγγυάται την ελευθερία κοινοποίησης του κώδικα.Το RStudio IDE είναι εν μέρει γραμμένο στη γλώσσα προγραμματισμού C ++ και χρησιμοποιεί το πλαίσιο Qt για τη γραφική διεπαφή χρήστη του. Το μεγαλύτερο ποσοστό του κώδικα είναι γραμμένο σε Java. Η JavaScript είναι επίσης μεταξύ των γλωσσών που χρησιμοποιούνται. Το RStudio IDE αναπτύχθηκε από την RStudio, PBC, μια εμπορική επιχείρηση που ιδρύθηκε από τον JJ Allaire, δημιουργό της γλώσσας προγραμματισμού ColdFusion. Το RStudio, PBC δεν έχει καμία επίσημη σύνδεση με το Ίδρυμα R, έναν μη κερδοσκοπικό οργανισμό που βρίσκεται στη Βιέννη της Αυστρίας, ο οποίος είναι υπεύθυνος για την εποπτεία της ανάπτυξης του περιβάλλοντος R για στατιστικούς υπολογισμούς.

7.2.2 Weka

Το Waikato Environment for Knowledge Analysis (Weka) είναι μια δημοφιλής σουίτα λογισμικού μηχανικής μάθησης γραμμένο σε Java, που αναπτύχθηκε στο Πανεπιστήμιο του Waikato της Νέας Ζηλανδίας. Είναι ελεύθερο λογισμικό υπό την άδεια GNU General Public License. Το Weka (προφέρεται για να κάνει ρίμα με τη Μέκκα) είναι μια σουίτα, η οποία περιέχει μια συλλογή από εργαλεία οπτικοποίησης και αλγορίθμους για την ανάλυση δεδομένων και την προγνωστική μοντελοποίηση, μαζί με γραφικές διεπαφές χρήστη για εύκολη πρόσβαση σε αυτές τις λειτουργίες. Η αρχική μη-Java έκδοση του Weka ήταν ένα Tcl/Tk front-end (ως επί το πλείστον τρίτων) για μοντελοποίηση αλγορίθμων που εφαρμόζονται σε άλλες γλώσσες προγραμματισμού, περιέχοντας δυνατότητες προεπεξεργασίας δεδομένων σε C, και ένα σύστημα βασισμένο σε Makefile για τη πραγματοποίηση πειραμάτων μηχανικής μάθησης. Αυτή η αρχική έκδοση είχε σχεδιαστεί ως ένα εργαλείο για την ανάλυση των δεδομένων από γεωργικούς τομείς, αλλά η πιο πρόσφατη πλήρης έκδοση βασισμένη σε Java(Weka 3), η ανάπτυξη της οποίας άρχισε το 1997, έχει πλέον πολλούς τομείς εφαρμογής,κυρίως εκπαιδευτικούς σκοπούς και έρευνες. Στα πλεονεκτήματα του Weka περιλαμβάνονται:

- Δωρεάν διαθεσιμότητα υπό την GNU Γενική Άδεια Δημόσιας χρήσης.
- Φορητότητα, δεδομένου ότι έχει υλοποιηθεί πλήρως στην γλώσσα προγραμματισμού Java και έτσι τρέχει σε σχεδόν κάθε σύγχρονη υπολογιστική πλατφόρμα.
- Μια ολοκληρωμένη συλλογή δεδομένων προεπεξεργασίας και τεχνικών μοντελοποίησης.
- Ευκολία στη χρήση λόγω των γραφικών διεπαφών χρήστη.

Το Weka είναι μια συλλογή αλγορίθμων μηχανικής μάθησης για καθήκοντα εξόρυξης δεδομένων. Περιέχει εργαλεία για την προετοιμασία των δεδομένων, την κατηγοριοποίηση, την παλινδρόμηση, την ομαδοποίηση, την εξόρυξη κανόνων συσχέτισης και την οπτικοποίηση. Όλες οι τεχνικές του Weka στηρίζονται στην υπόθεση ότι τα δεδομένα είναι διαθέσιμα ως ένα απλό αρχείο ή συσχέτιση, όπου κάθε σημείο δεδομένων περιγράφεται από ένα σταθερό αριθμό των χαρακτηριστικών (κανονικά, αριθμητικά ή ονομαστικά χαρακτηριστικά, αλλά και κάποιοι άλλοι τύποι χαρακτηριστικών υποστηρίζονται επίσης). Το Weka παρέχει πρόσβαση σε SQL βάσεις δεδομένων , χρησιμοποιώντας Java Database Connectivity και μπορεί να επεξεργαστεί το αποτέλεσμα που επιστρέφονται από ένα ερώτημα βάσης δεδομένων. Δεν είναι ικανό για εξόρυξη από πολυ-σχεσιακές βάσεις δεδομένων, αλλά υπάρχει ξεχωριστό λογισμικό για τη μετατροπή μιας συλλογής συνδεδεμένων πινάκων της βάσης δεδομένων σε έναν πίνακα που είναι κατάλληλος για επεξεργασία χρησιμοποιώντας το Weka. Άλλη μια σημαντική περιοχή που δεν καλύπτεται προς το παρόν από τους αλγορίθμους που περιλαμβάνονται στο Weka είναι η μοντελοποίηση αλληλουχιών.

7.2.3 Jupyter

Το Project Jupyter είναι ένας μη κερδοσκοπικός οργανισμός που δημιουργήθηκε για να "αναπτύξει λογισμικό ανοιχτού κώδικα, ανοιχτά πρότυπα και υπηρεσίες για διαδραστικούς υπολογιστές σε δεκάδες γλώσσες προγραμματισμού". Από το IPython το 2014 από τον Fernando Pérez, το Project Jupyter υποστηρίζει περιβάλλοντα εκτέλεσης σε αρκετές δεκάδες γλώσσες. Το όνομα του Project Jupyter είναι μια αναφορά στις τρεις βασικές γλώσσες προγραμματισμού που υποστηρίζονται από τον Jupyter, οι οποίες είναι οι Julia, Python και R, και επίσης ένα αφιέρωμα στα σημειωματάρια του Galileo που καταγράφουν την ανακάλυψη των φεγγαριών του Δία. Το Project Jupyter ανέπτυξε και υποστήριξε τα διαδραστικά υπολογιστικά προϊόντα Jupyter Notebook, JupyterHub και JupyterLab, την επόμενη γενιά του Jupyter Notebook.

Το Jupyter Notebook (πρώην IPython Notebooks) είναι ένα διαδραστικό υπολογιστικό περιβάλλον που βασίζεται στον ιστό για τη δημιουργία εγγράφων Jupyter notebook. Ο όρος "σημειωματάριο" μπορεί να αναφέρεται συστηματικά σε πολλές διαφορετικές οντότητες, κυρίως στην εφαρμογή Ιστού Jupyter, στον διακομιστή ιστού Jupyter Python ή στη μορφή εγγράφου Jupyter ανάλογα με το περιβάλλον. Ένα έγγραφο Jupyter Notebook είναι ένα έγγραφο JSON, το οποίο ακολουθώντας ένα διαμορφωμένο σχήμα περιέχει μια σειρά παραγγελίας κυψελών εισόδου / εξόδου που μπορούν να περιέχουν κώδικα, κείμενο (χρησιμοποιώντας Markdown), μαθηματικά, γραφικές παραστάσεις και εμπλουτισμένα μέσα, συνήθως τελειώνουν με το ".ipynb" επέκταση.

Κεφάλαιο 8

Επίλογος

Το κεφάλαιο αυτό συνοψίζει την μελέτη που εκπονήθηκε στην διπλωματική εργασία. Γίνεται ανακεφαλαίωση των προηγούμενων κεφαλαίων, παρουσιάζονται τα συμπεράσματα τα οποία έχουν εξαχθεί και παρατίθενται οι επεκτάσεις και οι μελλοντικές βελτιώσεις που μπορούν να βασιστούν στην παρούσα εργασία.

8.1 Σύνοψη και συμπεράσματα

Αντικείμενο μελέτης της παρούσας εργασίας αποτέλεσαν τα μοντέλα πρόβλεψης. Ειδικότερα, τα μοντέλα πρόβλεψης της επιτυχίας των κινηματογραφικών ταινιών μέσω της βαθμολόγησης τους στο δημοφιλέστερο σύστημα αξιολόγησης ταινιών. Πιο συγκεκριμένα, μεγάλο κομμάτι της υλοποίησης των μοντέλων αυτών στηρίχτηκε στην επεξεργασία των δεδομένων της βάσης προκειμένου να αυξηθεί η απόδοση των κατηγοριοποιητών. Για το σκοπό αυτό αρχικά έγινε επισκόπηση της βιβλιογραφίας, από την οποία αναζητήθηκε ο τρόπος ανάλυσης των δεδομένων,η λειτουργία και η χρησιμότητα του κάθε αλγορίθμου, καθώς και η εύρεση αντίστοιχων μελετών για τη διαχείρηση παρόμοιων προβλημάτων. Από την μελέτη αυτή προέκυψε το συμπέρασμα ότι τα γνωρίσματα των ηθοποιών και των σκηνοθετών παίζουν καθοριστικό ρόλο για την πρόβλεψη του αποτελέσματος. Επίσης έγινε αντιληπτό πως η μέθοδος της ανάλυσης των κύριων συνιστωσών είναι αυτή που μπορεί να κάνει τους αλγορίθμους της κατηγοριοποίησης αποδοτικότερους.

Αρχικά έγινε δοκιμή των αλγορίθμων κατηγοριοποίησης χωρίς καμία επεξεργασία των δεδομένων για να υπάρχει η δυνατότητα να συγκρίθούν τα αποτελέσματα που θα προέκυπταν από τις επόμενες προσεγγίσεις. Οι πρώτες δοκιμές έγιναν με την διακριτοποίηση του γνωρίσματός της κλάσης. Πραγματοποιήθηκαν πειράματα για διαφορετικούς αριθμούς κλάσεων με σκοπό την εύρεση της τομής μεταξύ ακρίβειας πρόβλεψης και απόδοσης αλγορίθμων. Έτσι αποφασίστηκε ο αριθμός των κλάσεων να είναι έξι. Να σημειωθεί ότι λόγω του γεγονότος πως τα δεδομένα δεν δέχτηκαν καμία επεξεργασία αρκετοί από τους αλγορίθμους, στο περιβάλλον της Weka, δεν κατάφεραν να βγάλουν κάποιο αποτέλεσμα.

Έπειτα, ξεκίνησε η πρώτη προσέγγιση του προβλήματος στην οποία έγινε ανάλυση των δεδομένων στο πρόγραμμα του RStudio. Κάθε γνώρισμα μελετήθηκε ξεχωριστά για να γίνει φανερός ο τρόπος με τον οποίο οι εγγραφές του ήταν κατανεμημένες και για το αν το γνώρισμα αυτό

μπορούσε να καθορίσει την κλάση. Κρίθηκε αναγκαίο κάποια γνωρίσματα που είχαν πολλές διαφορετικές εγγραφές να κατηγοριοποιηθούν σε κάποιες κλάσεις. Εξήχθει σαν συμπέρασμα πως τα γνωρίσματα που παίζουν το σημαντικότερο ρόλο για την κλάση ήταν το είδος τη ταινίας, οι σκηνοθέτες,ο πρώτος ηθοποιός, ο δεύτερος ηθοποιός και ο αριθμός των κριτικών. Με την ανάλυση αυτή το ποσοστό σωστής πρόβλεψης των αλγορίθμων ήταν κοντά στο 70% με μεγαλύτερη αυτή του LogitBoost.

Στη δεύτερη προσέγγιση αποφασίστηκε να εφαρμοστεί ως μέθοδος για τον καθορισμό των γνωρισμάτων η ανάλυση των κύριων συνιστωσών. Με την ανάλυση αυτή δημιουργούνται κάποια νέα γνωρίσματα που είναι συνδυασμός των αρχικών και είναι κάθετα μεταξύ τους. Επιλέχθηκαν οι οχτώ από τις δέκα συνιστώσες που είχε σαν έξοδο ο αλγόριθμος, καθώς το μεγαλύτερο ποσοστό της διακύμανσης των δεδομένων φάνηκε πως καθορίζεται από τις οχτώ πρώτες. Εφόσον εξήχθησαν αυτές οι συνιστώσες έγιναν τα πειράματα. Σε αυτή την περίπτωση τα αποτελέσματα ήταν πολύ καλύτερα, σε σχέση με την προηγούμενη προσέγγιση, για τους περισσότερους από τους αλγορίθμους. Την υψηλότερη απόδοση είχε ο αλγόριθμος Random Forest με απόδοση 87.55%. Επίσης επαληθεύεται και στην περίπτωση αυτής της εργασίας πως η ανάλυση των κύριων συνιστωσών βοηθούν στην βελτίωση της απόδοσης των περισσότερων αλγορίθμων κατηγοριοποίησης.

Μετά τις δύο πρώτες προσεγγίσεις, ακολούθησε μια τρίτη προσέγγιση. Αφού έγινε φανερό ότι οι ηθοποιοί και οι σκηνοθέτες παίζουν καθοριστικό ρόλο στο πόσο επιτυχημένη είναι μια ταινία επιλέχθηκε να ταξινομηθούν τα γνωρίσματα αυτά με διαφορετικό τρόπο, ώστε αυτή τους η ταξινόμηση να είναι πιο ακριβής. Έτσι εκτός από τα ήδη υπάρχοντα γνωρίσματα για τις ταινίες γενικά, προστέθηκαν, μέσω άλλης βάσης δεδομένων, τα βραβεία oscar καθώς και οι υποψηφιότητες για τα βραβεία αυτά εφόσον κρίθηκαν ως σημαντικός παράγοντας στην ανάδειξη του πόσο “καλός” είναι ένας ηθοποιός. Αφού πραγματοποίηθηκαν αυτές οι ενέργειες και πάρθηκαν κάποια στατιστικά δεδομένα από τις εγγραφές, τα γνωρίσματα βαθμολογήθηκαν στην κλίμακα του δέκα με βάση κάποιες παραμέτρους. Στην συνέχεια έγιναν πολλαπλά πειράματα προκειμένου να καθοριστεί ο βαθμός στον οποίο η κάθε παράμετρος επηρεάζει τη βαθμολογία ηθοποιών και σκηνοθετών, και να προκύψουν συντελεστές για κάθε μία παράμετρο. Συνακόλουθα δοκιμάστηκαν ως παράμετρος οι λέξεις της πλοκής μιας ταινίας, αφού και αυτές θα μπορούσαν να επηρεάσουν το αποτέλεσμα με μία διαφορετική κατηγοριοποίηση. Έτσι υπολογίστηκε ο μέσος όρος της βαθμολογίας της ταινίας στην οποία εμφανίζονταν η κάθε λέξη του γνωρίσματος ενώ παράλληλα υπολογίστηκε ως επιπλέον γνώρισμα ο μέσος όρος των βαθμολογιών των λέξεων πλοκής της κάθε ταινίας. Από τα πειράματα φάνηκε πως ο νέος αυτός υπολογισμός βελτίωσε το αποτέλεσμα. Την καλύτερη απόδοση φάνηκε να την έχει και πάλι ο αλγόριθμος Random Forest όπως φαίνεται στο δέκατο τρίτο πείραμα με τιμή ίση με 86.43%.

8.2 Μελλοντικές επεκτάσεις

Όπως μπορούμε να διαπιστώσουμε από την καθημερινότητα μας, μπορούμε να καταλάβουμε πως τα συστήματα πρόβλεψης έχουν αρχίσει να καταλαμβάνουν όλο και μεγαλύτερο μέρος και να μας επηρεάζουν. Ο κλάδος της εξόρυξης δεδομένων και της μηχανικής μάθησης έχουν γνωρίσει σημαντική άνθηση στις μέρες μας. Φαίνεται λοιπόν πως αυτά τα συστήματα θα έχουν όλο και

μεγαλύτερη χρήση στο μέλλον. Επομένως είναι καθοριστικό να μπορούν να κατανοήθούν με τον καλύτερο δυνατό τρόπο τα συστήματα αυτά και να βελτιωθούν.

Η εν λόγω εργασία μπορεί να αποτελέσει πυλώνα για μελλοντικές μελέτες ώστε να ανακαλυφθούν περισσότερα χαρακτηριστικά ταινιών που να βελτιώνουν τα παρόντα αποτελέσματα. Υπάρχουν περιθώρια εξέλιξης τόσο στην ακρίβεια πρόβλεψης, το ποσοστό επιτυχίας κατηγοριοποίησης, όσο και στον αριθμό των γνωρισμάτων. Η χρήση και η ανάλυση διαφορετικών βάσεων δεδομένων μπορεί να βοηθήσει στην εξαγωγή περισσότερων και ακριβέστερων συμπερασμάτων.

Όσο μεγαλύτερο το δείγμα τόση μεγαλύτερη και ακριβέστερη μπορεί να είναι και η ανάλυση των δεδομένων. Συνεπώς, καθως νέες ταινίες θα συνεχίζουν να κυκλοφορούν, νέα άτομα και νέες τάσεις στον χώρο του κινηματογράφου θα εμφανίζονται, η βάση δεδομένων κρίνεται αναγκαίο να ανταποκρίνεται στα νέα αυτά δεδομένα και να προσαρμόζεται. Σαν μελλοντικός στόχος είναι η υλοποίηση ενός τέτοιου συστήματος.

Τέλος, η υλοποίηση και η χρήση της μελέτης αυτής σε μία εφαρμογή που να είναι προσβάσιμη στο ευρύτερο κοινό θα ήταν επιθυμητή. Με τη χρήση μια τέτοιας εφαρμογής, ο χρήστης θα έχει τη δυνατότητα να επιλέγει με μεγαλύτερη ευκολία και περισσότερη σιγουριά τις ταινίες που επιθυμεί να παρακολοθήσει. Ακόμη βοηθιούνται και οι παραγωγοί ταινιών καθώς μπορούν να κρίνουν τα έσοδα που θα προκύψουν και ανάλογα να προβουν σε επένδυση μιας νέας ταινίας ή όχι. Κατ' επέκταση με την αποτελεσματικότερη πρόβλεψη επωφελούνται όλοι οι παράγοντες που έχουν σχέση με την βιομηχανία του κινηματογράφου.

Βιβλιογραφία

[1] amysfernweh. https://amysfernweh.wordpress.com/2017/09/08/pythonprincipal-component-analysis-and-k-means-clustering-with-imdb-movie-datasets/. Ημερομηνία πρόσβασης 26-4-2020.

[2] Leo Breiman. Random Forest. *Machine Learning*, 45:5 – 32, 2001.

[3] Christopher J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, 1998.

[4] Mark W. Craven και Jude W. Shavlik. Using neural networks for data mining. *Future Gener. Comput. Syst.*, 13(2-3):211–229, 1997.

[5] futurelearn. https://www.futurelearn.com/courses/data-mining-with-weka/0/steps/25384. Ημερομηνία πρόσβασης: 15-4-2020.

[6] Zheng Gao, Vincent Malic, Shutian Ma και Patrick C. Shih. How to Make a Successful Movie: Factor Analysis from both Financial and Critical Perspectives. Στο *Information in Contemporary Society - 14th International Conference, iConference 2019, Washington, DC, USA, March 31 - April 3, 2019, Proceedings*Natalie Greene Taylor, Caitlin Christian-Lamb, Michelle H. Martin και Bonnie A. Nardi, επιμελητές, τόμος 11420 στο *Lecture Notes in Computer Science*, σελίδες 669–678. Springer, 2019.

[7] github. https://github.com/abheek24/IMDB-MovieRevenue-Prediction. Ημερομηνία πρόσβασης: 23-4-2020.

[8] Tom Howley, Michael G. Madden, Marie-Louise O'Connell και Alan G. Ryder. The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. *Knowl. Based Syst.*, 19(5):363–370, 2006.

[9] George H. John και Pat Langley. Estimating Continuous Distributions in Bayesian Classifiers. Στο *UAI '95: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, August 18-20, 1995*Philippe Besnard και Steve Hanks, επιμελητές, σελίδες 338–345. Morgan Kaufmann, 1995.

[10] Oded Maimon Lior Rokach. *Data Mining With Decision Trees*. World Scientific Publishing Co. Pte. Ltd., 5 Toh Tuck Link, Singapore 596224, 2008.

[11] J. Eccleston M. Saraee, S. White. A data mining approach to analysis and prediction of movie ratings. 2004.

[12] Vipin Kumar Pang-Ning Tan, Michael Steinbach. *Εισαγωγή στην Εξόρυξη Δεδομένων*. Τζιολά, 2010.

[13] S.Abraham Ravid. Information, Blockbusters and Stars - A Study of the Film Industry. 1997.

[14] Dawn Iacobucci Sangkil Moon, Paul K. Bergey. Dynamic Effects Among Movie Ratings, Movie Revenues, and Viewer Satisfaction. *Journal of Marketing*, 74:108–121, 2010.

[15] C. Bhattacharywa K.R.K Murphy S.S. Keerthi, S.K. Shevade. Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Compulation. 2001.

[16] summarizing data. `http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_summarizingdata/bs704_summarizingdata7.html`. Ημερομηνία πρόσβασης: 5-6-2020.

[17] The International Movie Database (IMDb). `http://www.imdb.com`. Ημερομηνία πρόσβασης: 28-2-2020.

[18] The online community Kaggle. `http://www.kaggle.com`. Ημερομηνία πρόσβασης: 28-2-2020.

[19] John Verzani. *Getting Started with RStudio*. O'Reilly, 1005 Gravenstein Highway North, Sebastopol, 2011.

[20] Κ. Διαμαντάρας. *Τεχνητά Νευρωνικά Δίκτυα*. Κλειδάριθμος, Αθήνα, 1ή έκδοση, 2007.

Συντομογραφίες

βλπ	βλέπε
κ.λπ.	και λοιπά
κ.ο.κ	και ούτω καθεξής
PCA	Principal Component Analysis
ΑΚΣ	Ανάλυση Κύριων Συνιστωσών
SVM	Support Vector Machines
SMO	Sequential Minimal Optimization
IMDB	Internet Movie Database
Weka	Waikato Environment for Knowledge Analysis
IQR	interquartile range
PBC	Public-benefit corporations
NA	not valued

Ορολογία - Γλωσσάρι

Ελληνικός όρος	Αγγλικός όρος
εξόρυξη δεδομένων	data mining
μηχανική μάθηση	machine learning
διακύμανση	variance
τυχαία δάση	random forest
γνώρισμα	feature, attribute
διασταυρούμενη επικύρωση	cross validation