

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Δημιουργία γενικευμένου συστήματος προτάσεων με εφαρμογή σε σύνολα αλληλεξαρτώμενων δεδομένων

Εκπόνηση:                                          Επιβλέποντες:

Ιωάννης-Παναγιώτης Μπουντουρίδης                        Καθ. Περικλής Α. Μήτκας

giannis.[boudou@gmail.com](mailto:boudou@gmail.com)                                       Δρ. Αντώνης Χρυσόπουλος

ΑΕΜ: 8872                                          Ερευν. Σωτήρης Τσαρούχης

ΘΕΣΣΑΛΟΝΙΚΗ 2020

# Ευχαριστίες

Καταρχάς θα ήθελα να ευχαριστήσω τον καθηγητή κ. Περικλή Μήτκα την εμπιστοσύνη που μου έδειξε αναθέτοντάς αυτήν την διπλωματική εργασία.

Επίσης θα ήθελα να ευχαριστήσω τον μεταδιδακτορικό ερευνητή του εργαστηρίου Δρ. Αντώνη Χρυσόπουλο για την εξαιρετική συνεργασία κατά τη διάρκεια εκπόνησης της διπλωματικής εργασίας.

Επιπλέον θα ήθελα να ευχαριστήσω τον ερευνητή του εργαστηρίου Σωτήρη Τσαρούχη για τις συμβουλές του, την καθοδήγηση και τη συνεχή επικοινωνία που είχαμε.

Τέλος θα ήθελα να ευχαριστήσω τους ανθρώπους που με έχουν στηρίξει όσο κανένας άλλος δηλαδή τους γονείς μου Μπάμπη και Στέλλα και τον αδερφό μου Τάσο που αποτελούν τα σημαντικότερα πρόσωπα στην πορεία της ζωής μου.

# Περίληψη

Στην σημερινή εποχή μεγάλοι κολοσσοί λογισμικού και τεχνολογίας οφείλουν ένα σημαντικό κομμάτι της επιτυχίας τους στον χρόνο που επένδυσαν με στόχο να δημιουργήσουν αξιόπιστα συστήματα προτάσεων. Ο ρόλος των συστημάτων προτάσεων εξυπηρετεί την ερμηνεία της δραστηριότητας των χρηστών και προτείνει προϊόντα ή υπηρεσίες που είναι όσο κατά το δυνατόν πιο πιθανά να είναι κοντά στα ενδιαφέροντα των χρηστών. H βιομηχανία της ένδυσης καθώς και η βιομηχανία του κινηματογράφου αποτελούν αναμφίβολα πεδία που βασίζονται στα παραπάνω συστήματα.

Σκοπός της παρούσας διπλωματικής είναι η δημιουργία ενός γενικευμένου συστήματος προτάσεων που θα εφαρμόζει τεχνικές προεπεξεργασίας δεδομένων και πρόβλεψης προτιμήσεων. Τα πεδία εφαρμογής στο παραπάνω γενικευμένο σύστημα είναι η βιομηχανία της ένδυσης και η βιομηχανία του κινηματογράφου. Για την βιομηχανία της ένδυσης τελικός στόχος είναι η δημιουργία ενός συστήματος που θα υποβοηθά σχεδιαστές παρέχοντας προτάσεις για τον σχεδιασμό ρούχων ενώ για την βιομηχανία του κινηματογράφου τελικός στόχος είναι η δημιουργία ενός συστήματος που θα παρέχει προτάσεις ταινιών στους χρήστες όσο γίνεται πιο κοντά στις προτιμήσεις τους.

Το κομμάτι της ένδυσης περιλαμβάνει δύο σετ δεδομένων με ρούχα από την εταιρία ASOS και αξιολογήσεις χρηστών, το ένα set έχει ποσοστό αξιολογήσεων 30% και το άλλο 60%. Για το κομμάτι του κινηματογράφου γίνεται η χρήση του σετ δεδομένων movieLens, ένα σετ που περιλαμβάνει ταινίες με αξιολογήσεις δημοσιευμένο από την ομάδα GroupLens.

Τα πειράματα που διεξάχθηκαν στα παραπάνω συστήματα αξιολογούνται μέσα από κάποιες χρήσιμες μετρικές με σκοπό ερμηνεία των προτάσεων του συστήματος δηλαδή του πόσο κοντά είναι οι προτάσεις των συστημάτων στις προτιμήσεις των χρηστών. Τα αποτελέσματα των προτάσεων που παρουσιάζονται στο τέλος γίνονται με το σύστημα που είχε την καλύτερη επίδοση στην δημιουργία πετυχημένων προτάσεων.

# Abstract

Να γίνει η μετάφραση της περίληψης μετά την ολοκλήρωση της

# Λίστα εικόνων

[Εικόνα 1 - Αναπαράσταση ανάλυσης πίνακα σε ιδιάζουσες τιμές (SVD) [3] 21](#_Toc62231494)

[Εικόνα 2 Αρχιτεκτονική Νευρωνικών Συνεργατικών Συστημάτων [5] 23](#_Toc62231495)

[Εικόνα 3 Οπτικοποίηση μοντέλου: Τυχαία Δάση [13] 27](#_Toc62231496)

[Εικόνα 4 Αρχιτεκτονική Νευρωνικών Δικτύων 28](#_Toc62231497)

[Εικόνα 5 Ταξινομία Υβριδικών Συστημάτων [14] 30](#_Toc62231498)

[Εικόνα 6 Διάγραμμα ρούχων ανά κατηγορία αρχικού σετ δεδομένων 37](#_Toc62231499)

[Εικόνα 7 Αναπαράσταση περίπτωσης επικάλυψης ρούχων ανά κατηγορία 38](#_Toc62231500)

[Εικόνα 8 Διάγραμμα ρούχων ανά κατηγορία τελικού σετ δεδομένων 38](#_Toc62231501)

[Εικόνα 9 Δημιουργία Πίνακα Συγκεντρωτικών Χαρακτηριστικών 40](#_Toc62231502)

[Εικόνα 10 Αναπαράσταση Διαχωρισμού Δεδομένων - 1 51](#_Toc62231503)

[Εικόνα 11 Αναπαράσταση Διαχωρισμού Δεδομένων - 2 52](#_Toc62231504)

[Εικόνα 12 Διαχωρισμός αξιολογήσεων 52](#_Toc62231505)

[Εικόνα 13 Υλοποίηση User-Based - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score ανά γείτονα – data30 53](#_Toc62231506)

[Εικόνα 14 Υλοποίηση User-Based - Διάγραμμα μετρικών αccuracy, recall, precision και f1\_score ανά γείτονα – data60 54](#_Toc62231507)

[Εικόνα 15 Υλοποίηση User-Based - Διάγραμμα μετρικών rmse και mae ανά γείτονα – data30 55](#_Toc62231508)

[Εικόνα 16 Υλοποίηση User-Based - Διάγραμμα μετρικών rmse και mae ανά γείτονα - data60 55](#_Toc62231509)

[Εικόνα 17 Υλοποίηση User-Based - Διάγραμμα Κάλυψης και Ποικιλίας – data30 56](#_Toc62231510)

[Εικόνα 18 Υλοποίηση User-Based - Διάγραμμα Κάλυψης και Ποικιλίας – data60 56](#_Toc62231511)

[Εικόνα 19 Υλοποίηση SVD - Διάγραμμα μετρικών αccuracy, recall, precision και f1\_score ανά τιμή χαρακτηριστικών – data30 57](#_Toc62231512)

[Εικόνα 20 Υλοποίηση SVD - Διάγραμμα μετρικών αccuracy, recall, precision και f1\_score ανά τιμή χαρακτηριστικών – data60 58](#_Toc62231513)

[Εικόνα 21 Υλοποίηση SVD - Διάγραμμα μετρικών αccuracy, recall, precision και f1\_score ανά τιμή χαρακτηριστικών – movielens 100K 58](#_Toc62231514)

[Εικόνα 22 Υλοποίηση SVD - Διάγραμμα μετρικών rmse και mae ανά τιμή χαρακτηριστικών – data30 59](#_Toc62231515)

[Εικόνα 23 Υλοποίηση SVD - Διάγραμμα μετρικών rmse και mae ανά τιμή χαρακτηριστικών – data60 60](#_Toc62231516)

[Εικόνα 24 Υλοποίηση SVD - Διάγραμμα μετρικών rmse και mae ανά τιμή χαρακτηριστικών – movielens 100K 60](#_Toc62231517)

[Εικόνα 25 Υλοποίηση SVD - Διάγραμμα Κάλυψης και Ποικιλίας – data30 61](#_Toc62231518)

[Εικόνα 26 Υλοποίηση SVD - Διάγραμμα Κάλυψης και Ποικιλίας – data60 61](#_Toc62231519)

[Εικόνα 27 Υλοποίηση NCF - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 62](#_Toc62231520)

[Εικόνα 28 Υλοποίηση NCF - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data60 63](#_Toc62231521)

[Εικόνα 29 Υλοποίηση NCF - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K 63](#_Toc62231522)

[Εικόνα 30 Υλοποίηση NCF - Διάγραμμα μετρικών rmse και mae – data30 64](#_Toc62231523)

[Εικόνα 31 Υλοποίηση NCF - Διάγραμμα μετρικών rmse και mae – data60 64](#_Toc62231524)

[Εικόνα 32 Υλοποίηση NCF - Διάγραμμα μετρικών rmse και mae – movielens 100K 65](#_Toc62231525)

[Εικόνα 33 Υλοποίηση NCF - Διάγραμμα Κάλυψης και Ποικιλίας – data30 65](#_Toc62231526)

[Εικόνα 34 Υλοποίηση NCF - Διάγραμμα Κάλυψης και Ποικιλίας – data60 66](#_Toc62231527)

[Εικόνα 35 Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 67](#_Toc62231528)

[Εικόνα 36 Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data60 67](#_Toc62231529)

[Εικόνα 37 Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K 68](#_Toc62231530)

[Εικόνα 38 Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών rmse και mae – data30 68](#_Toc62231531)

[Εικόνα 39 Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών rmse και mae – data60 69](#_Toc62231532)

[Εικόνα 40 Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών rmse και mae – movielens 100K 69](#_Toc62231533)

[Εικόνα 41 Υλοποίηση Τυχαίων Δασών - Διάγραμμα Κάλυψης και Ποικιλίας – data30 70](#_Toc62231534)

[Εικόνα 42 Υλοποίηση Τυχαίων Δασών - Διάγραμμα Κάλυψης και Ποικιλίας – data60 70](#_Toc62231535)

[Εικόνα 43 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 71](#_Toc62231536)

[Εικόνα 44 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 72](#_Toc62231537)

[Εικόνα 45 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K 72](#_Toc62231538)

[Εικόνα 46 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών rmse και mae – data30 73](#_Toc62231539)

[Εικόνα 47 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών rmse και mae – data60 73](#_Toc62231540)

[Εικόνα 48 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών rmse και mae – movielens 100K 74](#_Toc62231541)

[Εικόνα 49 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα Κάλυψης και Ποικιλίας – data30 74](#_Toc62231542)

[Εικόνα 50 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα Κάλυψης και Ποικιλίας – data60 75](#_Toc62231543)

[Εικόνα 51 Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 76](#_Toc62231544)

[Εικόνα 52 Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data60 76](#_Toc62231545)

[Εικόνα 54 Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K 77](#_Toc62231546)

[Εικόνα 55 Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – data30 77](#_Toc62231547)

[Εικόνα 56 Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – movielens 100K 78](#_Toc62231548)

[Εικόνα 57 Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα Κάλυψης και Ποικιλίας – data30 79](#_Toc62231549)

[Εικόνα 58 Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα Κάλυψης και Ποικιλίας – data60 79](#_Toc62231550)

[Εικόνα 59 Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 80](#_Toc62231551)

[Εικόνα 60 Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K 80](#_Toc62231552)

[Εικόνα 61 Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – data30 81](#_Toc62231553)

[Εικόνα 62 Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – movielens 100K 81](#_Toc62231554)

Πίνακας περιεχομένων

[Ευχαριστίες 2](#_Toc62231713)

[Περίληψη 3](#_Toc62231714)

[Abstract 4](#_Toc62231715)

[Λίστα εικόνων 5](#_Toc62231716)

[1 Εισαγωγή 12](#_Toc62231717)

[1.1 Γενικά 12](#_Toc62231718)

[1.2 Ορισμός προβλήματος 13](#_Toc62231719)

[1.3 Στόχοι διπλωματικής 13](#_Toc62231720)

[1.4 Μεθοδολογία διπλωματικής 14](#_Toc62231721)

[1.5 Οργάνωση Κεφαλαίων 14](#_Toc62231722)

[2 Θεωρητικό Υπόβαθρο 16](#_Toc62231723)

[2.1 Κατηγορίες συστημάτων προτάσεων 16](#_Toc62231724)

[2.1.1 Συστήματα Προτάσεων Συνεργατικού Φιλτραρίσματος (Collaborative filtering) 16](#_Toc62231725)

[2.1.1.1 Συστήματα ανάλυσης πίνακα σε ιδιάζουσες τιμές (SVD) 19](#_Toc62231726)

[2.1.1.2 Νευρωνικά Συνεργατικά Συστήματα (NCF) 21](#_Toc62231727)

[2.1.2 Συστήματα Προτάσεων Φιλτραρίσματος Περιεχομένου (Content-based Filtering) 24](#_Toc62231728)

[2.1.2.1 Τυχαία Δάση (Random Forest) 25](#_Toc62231729)

[2.1.2.2 Βαθιά Νευρωνικά Δίκτυα Περιεχομένου (Deep Content Neural Networks) 26](#_Toc62231730)

[2.1.3 Υβριδικά Συστήματα Προτάσεων (Hybrid) 27](#_Toc62231731)

[2.1.3.1 Σταθμισμένα Υβριδικά Συστήματα 29](#_Toc62231732)

[2.2 Μετρικές αξιολόγησης (Χρήσιμες μετρικές) 30](#_Toc62231733)

[2.2.1 Ομοιότητα δεδομένων 30](#_Toc62231734)

[2.2.2 Ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 31](#_Toc62231735)

[2.2.3 Μέσο απόλυτο σφάλμα (MAE) 32](#_Toc62231736)

[2.2.4 Κάλυψη (Coverage) 32](#_Toc62231737)

[2.2.5 Ποικιλομορφία (diversity) 33](#_Toc62231738)

[2.2.6 Καινοτομία (novelty) 33](#_Toc62231739)

[3 Υλοποιήσεις συστημάτων προτάσεων 35](#_Toc62231740)

[3.1 Προ-επεξεργασία Δεδομένων 35](#_Toc62231741)

[3.1.1 Συλλογή και οργάνωση δεδομένων 35](#_Toc62231742)

[3.1.2 Χαρακτηριστικά Προϊόντων (features) 37](#_Toc62231743)

[3.1.3 Βαθμολογίες Προϊόντων 38](#_Toc62231744)

[3.2 Συνεργατικό φιλτράρισμα με επίκεντρο τον χρήστη 39](#_Toc62231745)

[3.2.1 Είσοδος δεδομένων (data – input) 39](#_Toc62231746)

[3.2.2 Βήματα υλοποίησης 39](#_Toc62231747)

[3.2.2.1 Εύρεση όμοιων χρηστών 39](#_Toc62231748)

[3.2.2.2 Στάθμιση σημασίας ομοιότητας (significance weighting) 40](#_Toc62231749)

[3.2.2.3 Κεντράρισμα μέση τιμής δεδομένων (mean-center data) 40](#_Toc62231750)

[3.2.2.4 Συνάρτηση πρόβλεψης αξιολόγησης 41](#_Toc62231751)

[3.3 Συνεργατικό φιλτράρισμα με επίκεντρο τον προϊόν 42](#_Toc62231752)

[3.3.1 Προ-επεξεργασία δεδομένων (pre-processing) 42](#_Toc62231753)

[3.3.2 Κεντράρισμα μέση τιμής δεδομένων (mean-center data) 42](#_Toc62231754)

[3.3.3 Εύρεση όμοιων προϊόντων (ρούχων) 42](#_Toc62231755)

[3.3.4 Συνάρτηση πρόβλεψης αξιολόγησης 43](#_Toc62231756)

[3.4 Συνεργατικό φιλτράρισμα με ανάλυση πίνακα σε ιδιάζουσες τιμές 43](#_Toc62231757)

[3.4.1 Είσοδος Δεδομένων (data – input) 43](#_Toc62231758)

[3.4.2 Βήματα υλοποίησης 44](#_Toc62231759)

[3.4.2.1 Επιλογή βιβλιοθήκης μοντέλου 44](#_Toc62231760)

[3.5 Συνεργατικό φιλτράρισμα μέσω νευρωνικών δικτύων 44](#_Toc62231761)

[3.5.1 Δεδομένα εισόδου 44](#_Toc62231762)

[3.5.2 Βήματα υλοποίησης 44](#_Toc62231763)

[3.5.2.1 Κωδικοποίηση Χρηστών-Ρούχων 44](#_Toc62231764)

[3.5.2.2 Επίπεδο Ενσωμάτωσης 45](#_Toc62231765)

[3.5.2.3 Αρχιτεκτονική πολύ-επίπεδο νευρωνικού δικτύου 45](#_Toc62231766)

[3.6 Φιλτράρισμα Περιεχομένου με Τυχαία Δάση 46](#_Toc62231767)

[3.6.1 Δεδομένα εισόδου 46](#_Toc62231768)

[3.6.2 Βήματα υλοποίησης 46](#_Toc62231769)

[3.6.2.1 Επιλογή βιβλιοθήκης μοντέλου 46](#_Toc62231770)

[3.6.2.2 Επιλογή παραμέτρων 46](#_Toc62231771)

[3.7 Φιλτράρισμα Περιεχομένου με Νευρωνικά Δίκτυα 47](#_Toc62231772)

[3.7.1 Δεδομένα Εισόδου 47](#_Toc62231773)

[3.7.2 Βήματα υλοποίησης 47](#_Toc62231774)

[3.7.2.1 Επιλογή βιβλιοθήκης μοντέλου 47](#_Toc62231775)

[3.7.2.2 Αρχιτεκτονική Νευρωνικού Δικτύου 47](#_Toc62231776)

[4 Πειράματα και αποτελέσματα 49](#_Toc62231777)

[4.1 Διαχωρισμός δεδομένων 49](#_Toc62231778)

[4.2 Διαχωρισμός αξιολογήσεων 50](#_Toc62231779)

[4.3 Υλοποίηση συνεργατικού φιλτραρίσματος με επίκεντρο τον χρήστη (User-Based) 51](#_Toc62231780)

[4.3.1 Accuracy, recall, precision και f1\_score 51](#_Toc62231781)

[4.3.2 Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 52](#_Toc62231782)

[4.3.3 Κάλυψη και Ποικιλία 53](#_Toc62231783)

[4.4 Υλοποίηση συνεργατικού φιλτραρίσματος με ανάλυση πίνακα ιδιαζουσών τιμών (SVD) 55](#_Toc62231784)

[4.4.1 Accuracy, recall, precision και f1\_score 55](#_Toc62231785)

[4.4.2 Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 57](#_Toc62231786)

[4.4.3 Κάλυψη και Ποικιλία 58](#_Toc62231787)

[4.5 Υλοποίηση Συνεργατικού Νευρωνικού Φιλτραρίσματος 59](#_Toc62231788)

[4.5.1 Accuracy, recall, precision και f1\_score 60](#_Toc62231789)

[4.5.2 Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 61](#_Toc62231790)

[4.5.3 Κάλυψη και Ποικιλία 63](#_Toc62231791)

[4.6 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Τυχαία Δάση 64](#_Toc62231792)

[4.6.1 Accuracy, recall, precision και f1\_score 64](#_Toc62231793)

[4.6.2 Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 66](#_Toc62231794)

[4.6.3 Κάλυψη και Ποικιλία 67](#_Toc62231795)

[4.7 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα 69](#_Toc62231796)

[4.7.1 Accuracy, recall, precision και f1\_score 69](#_Toc62231797)

[4.7.2 Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 70](#_Toc62231798)

[4.7.3 Κάλυψη και Ποικιλία 72](#_Toc62231799)

[4.8 Yλοποίηση Υβριδικού Συστήματος 73](#_Toc62231800)

[4.8.1 Accuracy, recall, precision και f1\_score 73](#_Toc62231801)

[4.8.2 Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 75](#_Toc62231802)

[4.8.3 Κάλυψη και Ποικιλία 76](#_Toc62231803)

[4.9 Συνοπτική αναπαράσταση όλων των υλοποιήσεων 78](#_Toc62231804)

[4.9.1 Accuracy, recall, precision και f1\_score 78](#_Toc62231805)

[4.9.2 Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 79](#_Toc62231806)

[4.10 Αποτελέσματα αναζητήσεων 80](#_Toc62231807)

[4.10.1 Κατηγορία: Μάλλινα 80](#_Toc62231808)

[4.10.2 Κατηγορία: Παλτό 80](#_Toc62231809)

[4.10.3 Κατηγορία: Φόρεμα 81](#_Toc62231810)

[4.10.4 Κατηγορία: Σορτσάκια 81](#_Toc62231811)

[4.10.5 Κατηγορία: Πουκάμισα 82](#_Toc62231812)

[4.10.6 Κατηγορία: Μαγιο 83](#_Toc62231813)

[Βιβλιογραφία 84](#_Toc62231814)

Κεφάλαιο

# Εισαγωγή

## Γενικά

Στην σημερινή εποχή ένας ευρέως αναπτυσσόμενος κλάδος είναι αυτός της ένδυσης και της μόδας που αποτελεί μία κερδοφόρα και αρκετά εξελισσόμενη βιομηχανία. Σε πολλές χώρες του κόσμου η έννοια της ένδυσης παύει να αποτελεί απλά μία βασική ανάγκη. Πολλοί άνθρωποι οδηγούνται στην αγορά προϊόντων ένδυσης για την ικανοποίηση ευρύτερων αναγκών όπως κοινωνικών, επαγγελματικών ή και προσωπικών .

Σε αυτό καταναλωτικό κλίμα μεγάλοι οίκοι μόδας καλούνται να λύσουν προβλήματα που θα μεγιστοποιήσουν την απήχηση και το κέρδος που έχουν τα ρούχα τους. Σκοπός για την επίτευξη αυτού του στόχου είναι η δημιουργία κολεξιόν που θα ακολουθούν τις συνεχώς εξελισσόμενες τάσεις της μόδας που θα κάνουν τους καταναλωτές να ενδώσουν. Η έλλειψη πωλήσεων και η δημιουργία υψηλών αποθεμάτων ρούχων παρωχημένων είναι προβλήματα που πάση θυσία μεγάλοι κολοσσοί προσπαθούν να αποφύγουν.

Είναι σημαντικό, μεγάλοι οίκοι μόδας σε συνεργασία με άλλες εταιρίες προϊόντων παραγωγής ένδυσης σε κάθε καμπάνια τους να δημιουργούν τις προϋποθέσεις ώστε να κερδίζουν τις εντυπώσεις, να μειώνουν την αβεβαιότητα των πωλήσεων τους και να ακολουθούν πιστά τις τάσεις ώστε τελικά οδηγηθούν στην παραπάνω επίτευξη του στόχου τους.

Κάτι αντίστοιχο φαίνεται συμβαίνει και με τις μεγάλες εταιρίες συνδρομητικού κινηματογράφου οι οποίες παρέχουν υπηρεσίες στους χρήστες τους χιλιάδες ταινίες να δουν και να περάσουν ευχάριστα τον χρόνο τους. Καθώς όμως τα χρόνια περνούν όλο και περισσότερες ταινίες προστίθενται στην ίδια τεράστια λίστα με αποτέλεσμα η αναζήτηση μιας ταινίας κοντά στις προτιμήσεις του χρήστη να γίνεται διαδικασία δύσκολη και χρονοβόρα γεγονός που εύλογα προβληματίζει τις εταιρίες και δημιουργεί ένα κλίμα αβεβαιότητας που χρειάζεται να ξεπεραστεί.

## Ορισμός προβλήματος

Αν αναλογιστεί κανείς τον όγκο δεδομένων που χρειάζεται να διαχειριστεί μια εταιρία ρούχων η οποία εργάζεται δεκάδες χρόνια στο χώρο ή μια εταιρία συνδρομητικού κινηματογράφου η οποία δεν έπαψε ποτέ να προσθέτει νέες ταινίες και σειρές στην ήδη τεράστια βάση δεδομένων της τότε μπορεί κάλλιστα να αντιληφθεί την χρησιμότητα ενός γενικευμένου συστήματος προτάσεων και τον ρόλο του.

Ο ορισμός του προβλήματος είναι η διαχείριση αυτού του τεράστιο όγκο που μεγάλες εταιρίες καλούνται να επεξεργαστούν με σκοπό να βρουν προϊόντα ή υπηρεσίες κοντά στις προτιμήσεις τους, για της ίδιες ή τους χρήστες τους.

## Στόχοι διπλωματικής

Στόχος της παρούσας διπλωματικής είναι η δημιουργία ενός γενικευμένου συστήματος προτάσεων που θα εφαρμόζει τεχνικές προεπεξεργασίας δεδομένων και πρόβλεψης προτιμήσεων. Τα πεδία εφαρμογής στο παραπάνω γενικευμένο σύστημα είναι η βιομηχανία της ένδυσης και η βιομηχανία του συνδρομητικού κινηματογράφου. Για την βιομηχανία της ένδυσης τελικός στόχος είναι η δημιουργία ενός συστήματος που θα υποβοηθά σχεδιαστές παρέχοντας προτάσεις για τον σχεδιασμό ρούχων ενώ για την βιομηχανία του κινηματογράφου τελικός στόχος είναι η δημιουργία ενός συστήματος που θα παρέχει προτάσεις ταινιών στους χρήστες όσο γίνεται πιο κοντά στις προτιμήσεις τους.

## Μεθοδολογία διπλωματικής

Για την πραγματοποίηση της διπλωματικής εργασίας χρησιμοποιήθηκαν 3 σετ δεδομένων:

* asos30 dataset
* asos60 dataset
* movielens-100k dataset

Tα πρώτα δύο σετ δεδομένων περιλαμβάνουν ρούχα από την εταιρία ASOS μαζί με τα χαρακτηριστικά τους. Η διαφορά τους είναι ότι το πρώτο σύνολο έχει το 30% του συστήματος σε αξιολογήσεις από το σύνολο των ρούχων ενώ το δεύτερο 60%. Τα δύο αυτά σύνολα έχουν 100 χρήστες και 1700 ρούχα 10 διαφορετικών κατηγοριών. To τρίτο σύνολο έχει 600 χρήστες με συνολικά 100 χιλιάδες αξιολογήσεις σε 9000 ταινίες.

Τα στάδια για την δημιουργία του τελικού συστήματος προτάσεων ένδυσης ήταν:

1. Η συγκέντρωση δεδομένων
2. Η προεπεξεργασία δεδομένων των ρούχων
3. Ο καθορισμός των κατηγοριών κάθε ρούχου
4. Ο διαχωρισμός των δεδομένων σε δεδομένα εκπαίδευσης και ελέγχου
5. Η δημιουργία συστημάτων προτάσεων
6. Η συνεργασία των συστημάτων προτάσεων που δημιουργήθηκαν μέσω ενός υβριδικού συστήματος
7. Η χρήστη μετρικών για την αξιολόγηση και επιστημονική τεκμηρίωση της ορθής λειτουργίας του συστήματος

Τα στάδια για την δημιουργία του τελικού συστήματος προτάσεων συνδρομητικού κινηματογράφου ήταν:

1. Η συγκέντρωση δεδομένων
2. Η προεπεξεργασία δεδομένων των ταινιών
3. Ο καθορισμός των κατηγοριών κάθε ταινίας
4. Ο διαχωρισμός των δεδομένων σε δεδομένα εκπαίδευσης και ελέγχου
5. Η δημιουργία συστημάτων προτάσεων
6. Η συνεργασία των συστημάτων προτάσεων που δημιουργήθηκαν μέσω ενός υβριδικού συστήματος
7. Η χρήστη μετρικών για την αξιολόγηση και επιστημονική τεκμηρίωση της ορθής λειτουργίας του συστήματος

## Οργάνωση Κεφαλαίων

Η οργάνωση της παρούσας διπλωματικής αποτελείται από 5 κεφάλαια και είναι ως εξής:

* Κεφάλαιο 2: Περιγράφεται το βασικό θεωρητικό υπόβαθρο στο οποίο στηρίχθηκε η διπλωματική και η αρχιτεκτονική των εργαλείων που χρησιμοποιήθηκαν στην υλοποίηση.
* Κεφάλαιο 3: Γίνεται διεξοδική ανάλυση των βημάτων κάθε υλοποίησης καθώς και η αναφορά των εργαλείων που χρησιμοποιήθηκαν
* Κεφάλαιο 4: Γίνεται η παρουσίαση της , η αξιολόγηση του πόσο καλά επιτυγχάνει τελικά το σύστημα να εξάγει προτάσεις και η οπτική αναπαράσταση τους.
* Κεφάλαιο 5: Παρουσιάζονται τα τελικά συμπεράσματα και οι πιθανές μελλοντικές επεκτάσεις

Κεφάλαιο

# Θεωρητικό Υπόβαθρο

## Κατηγορίες συστημάτων προτάσεων

Μετά από μια εισαγωγική ανασκόπηση για την έννοια, τον ρόλο και την συμβολή των συστημάτων προτάσεων είναι αναγκαίο να αναλυθούν κάποιες συγκεκριμένες κατηγορίες συστημάτων που όπως θα δούμε στα επόμενα κεφάλαια θα αποτελέσουν θεμέλια για την δημιουργία του τελικού συστήματος.

### Συστήματα Προτάσεων Συνεργατικού Φιλτραρίσματος (Collaborative filtering)

Τα συστήματα προτάσεων που ανήκουν στην κατηγορία του συνεργατικού φιλτραρίσματος (collaborative filtering), σε αντίθεση με άλλα συστήματα, δε χρησιμοποιούν πληροφορίες περιεχομένου των προϊόντων, παρά μόνο τον τις βαθμολογίες χρηστών δηλαδή τον πίνακα χρήστη-αντικειμένου βαθμολογιών (user-item, rating matrix). H βασική ιδέα σε αυτή την κατηγορία αλγόριθμων είναι η εύρεση χρηστών με αρκετά παρόμοιες αξιολογήσεις. Η αποτελεσματικότητα των αλγορίθμων στηρίζεται στην παρατήρηση ότι οι χρήστες που έχουν παρόμοιες προτιμήσεις, αξιολογούν με παρόμοιο τρόπο. Οι τεχνικές αυτές συνήθως λαμβάνουν ένα σύνολο από τις βαθμολογίες των χρηστών του συστήματος και παράγουν προβλέψεις σχετικά με το τι χρειάζεται ένας χρήστης, βασιζόμενες στους πιο κοντινούς (ως προς τις προτιμήσεις) σε αυτόν χρήστες. Tα μοντέλα συνεργατικού φιλτραρίσματος χρησιμοποιούν αξιολογήσεις παρεχόμενες από όμοιους χρήστες με σκοπό τη δημιουργία προτάσεων. Η κυρία πρόκληση στην σχεδίαση ενός μοντέλου συνεργατικό φιλτραρίσματος είναι η σωστή διαχείριση των πινάκων αξιολογήσεων (user-item rating matrix). Οι πίνακες αυτοί, καθώς πολλές θέσεις αξιολογήσεων παραμένουν κενές έχουν πυκνότητα που είναι αρκετά αραιή (sparse). Πιο συγκεκριμένα σε μία βάση δεδομένων με χιλιάδες προϊόντα οι περισσότεροι χρήστες αξιολογούν ένα μικρό κομμάτι του συνόλου όλων των προϊόντων ως αποτέλεσμα οι περισσότερες αξιολογήσεις να παραμένουν απροσδιόριστες. Για κάθε χρήστη η κυρία ιδέα του συνεργατικού φιλτραρίσματος είναι να καλύψει αυτές στις απροσδιόριστες αξιολογήσεις λαμβάνοντας υπ’ όψη τους υψηλά συσχετιζόμενους χρήστες που έχουν παρόμοιες προτιμήσεις. Υπάρχουν δύο είδη μεθόδων που χρησιμοποιούνται κατά κόρον στις υλοποιήσεις συνεργατικού φιλτραρίσματος γνωστοί και ως *μέθοδοι βασισμένοι στη μνήμη (memory-based methods)* και *μέθοδοι βασισμένοι στα μοντέλα (model-based methods)*.

* *Μέθοδοι βασισμένοι στη μνήμη (memory-based methods):* Αποτελούν τα θεμέλια του συνεργατικό φιλτραρίσματος καθώς αποτέλεσαν τις πρώτες επιδιώξεις δημιουργίας προτάσεων που βασίστηκαν σε μια παρατήρηση ότι όμοιες αξιολογήσεις υποδηλώνουν και όμοιες προτιμήσεις. Για να εξερευνήσει κανείς την προτίμηση των χρηστών με αυτή την κατηγορία μεθόδων μπορεί να το κάνει με δύο τεχνικές: η μια τεχνική έχει επίκεντρο τις βαθμολογίες του χρήστη ενώ η άλλη τεχνική έχει επίκεντρο τις βαθμολογίες των προϊόντων.

1. *Τεχνική με επίκεντρο τον χρήστη (user-based):* Σε αυτή την περίπτωση αν υποθέσουμε ότι ο χρήστης Α είναι το επίκεντρο που θέλουμε να κάνουμε προτάσεις, τότε χρήστες όμοιοι με τις αξιολογήσεις του A δηλαδή χρήστες με όμοιες προτιμήσεις αξιοποιούνται για την εκτίμηση των αξιολογήσεων του χρήστη Α σε προϊόντα που δεν έχει εξερευνήσει. Το σύνολο των αξιολογήσεων του χρήστη Α συγκρίνεται με τα σύνολα αξιολογήσεων των υπόλοιπων χρηστών ώστε να βρεθούν ποιοι χρήστες είναι πιο όμοιοι. Η ομοιότητα κάθε χρήστη, με του χρήστη Α, βασίζεται μόνο σε αξιολογήσεις που έχουν κάνει οι χρήστες σε κοινά προϊόντα. Συναρτήσεις ομοιότητας με είσοδο τις κοινές αξιολογήσεις των συγκρινόμενων χρηστών επιστρέφουν μια τιμή που κυμμένεται από το -1 έως 1. Όσο πιο μεγάλη τιμή επιστρέφει η συνάρτηση τόσο πιο όμοιες θεωρούνται οι αξιολογήσεις των δύο συγκρινόμενων χρηστών. Ένα πρόβλημα που προκύπτει είναι ότι κάποιοι χρήστες τείνουν να βαθμολογούν αρκετά ψηλά ενώ άλλοι χαμηλά. Η λύση σε αυτό το πρόβλημα είναι η αφαίρεση της μέσης τιμής των αξιολογήσεων του χρήστη από κάθε αξιολόγηση προϊόντος ξεχωριστά ώστε να δημιουργηθούν mean-centered δεδομένα.
2. *Τεχνική με επίκεντρο το προϊόν (item-based):* Σε αυτή την περίπτωση αν πάλι υποθέσουμε ότι ο χρήστης Α είναι το επίκεντρο που θέλουμε να κάνουμε προτάσεις, τότε για ένα προϊόν Β ψάχνουμε ένα σύνολο προϊόντων S όμοιο με το προϊόν Β και χρησιμοποιούμε τις αξιολογήσεις του συνόλου S για την πρόβλεψη της αξιολόγησης του Β. Είναι προφανές λοιπόν πως σε αυτή την μέθοδο για την πρόβλεψη των απροσδιόριστων αξιολογήσεων του χρήστη Α γίνεται αναζήτηση των πιο όμοιων προϊόντων και όχι των πιο όμοιων χρηστών. Συνεπώς τα Κ πιο όμοια προϊόντα του συνόλου S έχουν αξιολογήσεις που θα καθορίσουν το αποτέλεσμα της πρόβλεψης για την τιμή αξιολόγησης που θα πάρει το προϊόν Β.

* *Μέθοδοι βασισμένοι στα μοντέλα (model-based methods):* Σε αντίθεση με τους αλγόριθμους που βασίζονται στη μνήμη, οι αλγόριθμοι βασιζόμενοι στα μοντέλα χρησιμοποιούν τις βαθμολογίες των χρηστών προκειμένου να εκπαιδεύσουν ένα μοντέλο. Οι μέθοδοι βασισμένοι στα μοντέλα προσαρμόζουν ένα παραμετροποιημένο σύστημα στον πίνακα αξιολογήσεων το οποίο στη συνέχεια χρησιμοποιείται για την πρόβλεψη απροσδιόριστων βαθμολογιών και κατ’ επέκταση για τη δημιουργία προτάσεων προς τους χρήστες. Σε αυτή την υποκατηγορία συστημάτων ανήκουν: οι ταξινομητές κοντινότερου γείτονα (knn classifiers), οι Bayesian ταξινομητές (Bayesian classifiers) αλλά και οι μέθοδοι βασιζόμενοι σε παλινδρόμηση (regression-based methods). Το ευρύτερο φάσμα αυτών των τεχνικών είναι: η μηχανική μάθηση (machine learning) και η εξόρυξη δεδομένων (data mining). Μία αποτελεσματική κατηγορία μοντέλων που βασίζεται στην χαμηλού βαθμού παραγοντοποίηση μήτρας (low-rank matrix factorization) με τεχνικές όπως το Singular Value Decomposition (SVD), το regularized SVD, το Non-negative Matrix Factorization (NMF), το Probabilistic Matrix Factorization (PMF), το Bayesian PMF, το Non-linear PMF, το Maximum Margin Matrix Factorization (MMMF) και Nonlinear Principal Component Analysis (NPCA) . Τα μοντέλα που αναφέρθηκαν τελευταία αποκαλούνται λανθάνοντα παράγοντα (Latent factor models) [1]. Αυτές οι μέθοδοι έχουν γίνει δημοφιλείς τα τελευταία χρόνια επειδή συνδυάζουν καλό βαθμό κλιμάκωσης και ευστοχία προβλέψεων. Επιπρόσθετα, προσφέρουν ευελιξία στην μοντελοποίηση διαφόρων καταστάσεων από την πραγματική ζωή. Ωστόσο, επισημαίνεται ότι όλων των ειδών οι model-based προσεγγίσεις είναι συχνά χρονοβόρες στην υλοποίηση και την ενημέρωσή τους. [2]

#### Συστήματα ανάλυσης πίνακα σε ιδιάζουσες τιμές (SVD)

Τα συστήματα ανάλυσης πίνακα σε ιδιάζουσες τιμές (Singular Value Decomposition) συνιστούν μια τεχνική παραγοντοποίησης πινάκων με βάση την οποία ένας πίνακας R διαστάσεων παραγοντοποιείται στο γινόμενο πινάκων. Τα συστήματα αυτού του τύπου ανήκουν στην κατηγορία συστημάτων συνεργατικού φιλτραρίσματος με μεθόδους βασισμένες στην παραμετροποίηση και εκπαίδευση ενός μοντέλου. Η ανάλυση πίνακα αυτών των μοντέλων είναι της μορφής:

όπου:

* ένας ορθογώνιος πίνακας διαστάσεων
* ο ανάστροφος πίνακας διαστάσεων
* ένας διαγώνιος πίνακας διαστάσεων

Τα στοιχεία του πίνακα είναι οι ιδιάζουσες τιμές του πίνακα και είναι ταξινομημένα σε σειρά φθίνουσα. Με την βοήθεια των πινάκων και οι χρήστες και τα προϊόντα (ρούχα) μπορούν να αναπαρασταθούν σε έναν χώρο διαστάσεων. Η χρησιμότητα της ανάλυσης πίνακα σε ιδιάζουσες τιμές έγκειται στη δυνατότητα διατήρησης των k μεγαλύτερων ιδιαζουσών τιμών του διαγώνιου πίνακα και η αποβολή των υπόλοιπων τιμών όπως αντίστοιχα η διατήρηση των πρώτων στηλών των πινάκων και και η αποβολή των υπόλοιπων στηλών. Τα πλεονεκτήματα που προκύπτουν από τη χρήση της ανάλυσης πίνακα σε ιδιάζουσες τιμές συμβάλλουν σημαντικά στην απόδοση και ακρίβεια του συστήματος συστάσεων.



Εικόνα 1 - Αναπαράσταση ανάλυσης πίνακα σε ιδιάζουσες τιμές (SVD) [3]

Πρώτον, σημαντική είναι η ανάλυση των λανθανουσών σχέσεων που υπάρχουν ανάμεσα στους χρήστες και τα προϊόντα (ρούχα). Κάθε ιδιάζουσα τιμή αναφέρεται σε έναν άγνωστο παράγοντα, που χαρακτηρίζει σημαντικά τη σχέση ανάμεσα στους χρήστες και τα προϊόντα. Διατηρώντας τις μεγαλύτερες ιδιάζουσες τιμές, εξασφαλίζεται η διατήρηση των πιο σημαντικών παραγόντων, άσχετα εάν η ερμηνεία τους είναι κατά πάσα πιθανότητα αδύνατη. Επιπλέον, η τεχνική αυτή εξασφαλίζει μείωση διαστάσεων (dimensionality reduction) του αρχικού προβλήματος. Εξαλείφοντας τις μικρότερες ιδιάζουσες τιμές αρκεί να αποθηκεύσουμε μόνο τους τρεις παραγόμενους πίνακες μειώνοντας το σύνολο των αποθηκευμένων στοιχείων από σε .Συνεπώς η επιλογή της τιμής έχει άμεσες συνέπειες στην ακρίβεια συστάσεων αλλά και στη απόδοσή του συστήματος ως προς τις απαιτήσεις σε αποθηκευτικό χώρο. Τα μειονεκτήματα της προσέγγισης αφορούν τον χρόνο εκτέλεσής της, ο οποίος είναι συνήθως μεγαλύτερος από τη προσέγγιση συνεργατικού φιλτραρίσματος με βάση τη μνήμη. Ωστόσο τα αποτελέσματα της ανάλυσης του αρχικού πίνακα μπορούν να αποθηκευτούν και να επαναχρησιμοποιηθούν για την παραγωγή προβλέψεων. Μάλιστα, εξαιτίας της μείωσης των διαστάσεων, παρουσιάζονται πιο σύντομοι χρόνοι εκτέλεσης κατά το στάδιο δημιουργίας συστάσεων σε σύγκριση με τα αντίστοιχα στάδια των υπόλοιπων προσεγγίσεων. Δεύτερο μειονέκτημα αποτελεί η ανάγκη για εκ νέου εκτέλεση της διαδικασίας ανάλυσης εξαιτίας αλλαγών στον αρχικό πίνακα δεδομένων (προσθήκη νέων χρηστών ή αντικειμένων). Στη βιβλιογραφία όμως παρατηρούνται προσπάθειες για τη δημιουργία τεχνικών ανανέωσης και επέκτασης των τριών παραγόμενων πινάκων (SVD updating & folding up). Τέλος η προσέγγιση απαιτεί συνήθως απαλοιφή τον μηδενικών τιμών του αρχικού πίνακα μέσω εξειδικευμένης προεπεξεργασίας. [4]

#### Νευρωνικά Συνεργατικά Συστήματα (NCF)

Όπως αναφέρθηκε ήδη στα συστήματα προτάσεων ανάλυσης πίνακα σε ιδιάζουσες τιμές (SVD) η εκτίμηση των προτάσεων κάθε χρήστη προκύπτει από το εσωτερικό γινόμενο των διανυσμάτων κάθε σειράς του πίνακα χρηστών με κάθε στήλη του πίνακα προϊόντων. Στην σύγχρονη βιβλιογραφία των τελευταίων ετών υποστηρίζεται ότι ο υπολογισμός του εσωτερικού γινομένου περιορίζει την εκφραστικότητα των λανθάνων διανυσμάτων αφού γίνεται ο εντοπισμός μόνο των γραμμικών και όχι των μη γραμμικών σχέσεων. Το συνεργατικό φιλτράρισμα μέσω νευρωνικών δικτύων άρει αυτό τον περιορισμό.



Εικόνα 2 Αρχιτεκτονική Νευρωνικών Συνεργατικών Συστημάτων [5]

Η αρχιτεκτονική των Νευρωνικών Συνεργατικών Συστημάτων περιλαμβάνει:

* Το επίπεδο εισόδου (input layer)
* Το ενσωματωμένο επίπεδο (embedding layer)
* Τα επίπεδα του νευρωνικού συνεργατικού φιλτραρίσματος (Neural Collaborative layers)
* To τελικό επίπεδο εξόδου (output layer) που επιστρέφει την εκτίμηση των αξιολογήσεων ελαχιστοποιώντας το σφάλμα πρόβλεψης

Το επίπεδο εισόδου αποτελείται από δύο αραιά διανύσματα χαρακτηριστικών (sparse feature vectors) για την αναγνώριση κάθε χρήστη και προϊόντος αντίστοιχα. Δεδομένου ότι το σύστημα επικεντρώνεται στην δημιουργία μιας πρότασης καθαρά εξαρτώμενης του συνεργατικού φιλτραρίσματος η είσοδος απαιτεί τον καθορισμό του χρήστη και του προϊόντος προς εξέταση για την δημιουργία μιας πρότασης. Αυτός ο τρόπος γενικής αναπαράστασης χρηστών και προϊόντων επιτρέπει την εύκολη προσαρμογή του συστήματος να αντιμετωπίζει προβλήματα εξαγωγής προτάσεων σε συστήματα με λίγα διαθέσιμα δεδομένα αξιολογήσεων από τους χρήστες. Πάνω από το επίπεδο εισόδου βρίσκεται το επίπεδο ενσωμάτωσης, ένα πλήρες συνδεδεμένο επίπεδο στο οποίο προβάλλεται η αραιή αναπαράσταση των χρηστών ή αντικειμένων σε διανύσματα χαμηλότερων διαστάσεων. Το ενσωματωμένο επίπεδο για τον χρήστη όπως και το ενσωματωμένο επίπεδο για το προϊόν αντίστοιχα μπορεί να θεωρηθεί ότι περιλαμβάνει τα λανθάνον διανύσματα όμοια των συστημάτων ανάλυσης ιδιάζουσας τιμής (SVD). Τα ενσωματωμένα επίπεδα συνενώνονται και τροφοδοτούν ένα πολύ-επίπεδο νευρωνικό δίκτυο τα οποία χαρακτηρίζονται και επίπεδα συνεργατικού φιλτραρίσματος. Κάθε επίπεδο έχει ικανότητα να χαρτογραφήσει τα λανθάνοντα διανύσματα στις κατάλληλες βαθμολογίες πρόβλεψης. Κάθε στρώμα μπορεί να προσαρμοστεί και να ανακαλύψει ορισμένες αλληλεπιδράσεις χρηστών – προϊόντων. Το μέγεθος του στρώματος επηρεάζει αυτή την ικανότητα του μοντέλου. Το τελικό επίπεδο της εξόδου είναι η προβλεπόμενη βαθμολογία και η εκπαίδευση πραγματοποιείται ελαχιστοποιώντας την απώλεια της προβλεπόμενης με την πραγματική τιμή . Σημειώνεται ότι ένας άλλος τρόπος εκπαίδευσης του μοντέλου πραγματοποιείται με χρήση ζεύγους μάθησης όπως αυτή της Bayesian εξατομικευμένης κατάταξης (Bayesian Personalized Ranking) [6] ή βάση περιθωρίου απώλειας (margin-based loss) [7]. Δεδομένου ότι τα συστήματα μοντελοποιούν τα αραιά διανύσματα χρηστών και προϊόντων σε πυκνά χαμηλότερων διαστάσεων, είναι διαισθητικό να συνδυάζονται τα χαρακτηριστικά μέσω της συνένωσης και να τροφοδοτούν ένα πολυεπίπεδο νευρωνικό σύστημα. Αυτός ο σχεδιασμός έχει υιοθετηθεί ευρέως σε πολυτροπικές εργασίες βαθιάς μηχανικής μάθησης (multimodal deep learning work) [8]. Μια απλή συνένωση των χαρακτηριστικών δε λαμβάνει υπόψη καμία αλληλεπίδραση μεταξύ χρηστών και προϊόντων των λανθάνων χαρακτηριστικών που είναι απαραίτητη για την εφαρμογή του συνεργατικού φιλτραρίσματος. Για την διευθέτηση αυτού του ζητήματος γίνεται η χρήση του νευρωνικού δικτύου ώστε να γίνει η εκπαίδευση για την αναγνώριση αυτών των αλληλεπιδράσεων και την αναγνώριση μη γραμμικών σχέσεων λειτουργία αδύνατη για μοντέλα ανάλυσης πίνακα ιδιάζουσων τιμών. Πιο συγκεκριμένα ορίζεται το σύστημα:

,

*,*

*,*

όπου ο πίνακας από βάρυ το bias και η συνάρτηση ενεργοποίησης για το επίπεδο. Αναλύοντας κάθε διαθέσιμη συνάρτηση ενεργοποίησης παρατηρούμε ότι:

1. *Η συνάρτηση σιγμοειδούς περιορίζει κάθε νευρώνα να είναι σε διάστημα (0,1) κάτι που μπορεί να έχει αρνητικές επιπτώσεις την απόδοση του μοντέλου. Άλλωστε είναι γνωστό ότι ένα από τα μειονεκτήματα της συνάρτησης αυτής είναι ότι υποφέρει από κορεσμό. Οι νευρώνες σταματούν να μαθαίνουν όταν η παράγωγος είναι κοντά στο μηδέν ή ένα.* [9]
2. *Η συνάρτηση υπερβολικής εφαπτόμενης είναι καλύτερη επιλογή από αυτή της σιγμοειδούς και έχει υιοθετηθεί ευρέως* [10] *αφού ανακουφίζει τα θέματα σε μεγάλο βαθμό τα θέματα του κορεσμού.*
3. *Η συνάρτηση ReLU είναι ακόμη μια καλή λύση αφού έχει αποδειχθεί μια εύλογη επιλογή που λύνει τα προβλήματα κορεσμού* [11] *ενθαρρύνει αραιές ενεργοποιήσεις και είναι κατάλληλη για αραιά δεδομένα μειώνοντας τις πιθανότητες υπερεκπαίδευσης.*

Τα αποτελέσματα δείχνουν ότι η συνάρτηση ReLU δίνει ελαφρώς καλύτερη απόδοση από την συνάρτηση υπερβολικής εφαπτόμενης η οποία με την σειρά της είναι αποδεδειγμένα καλύτερη από τη σιγμοειδές.

Όσον αφορά το σχεδιασμό του δικτύου μια τυπική λύση είναι ο σχεδιασμός πύργου, όπου το πρώτο στρώμα είναι το ευρύτερο και κάθε διαδοχικό έχει μικρότερο αριθμό νευρώνων. H παραπάνω λύση είναι μια λύση που υιοθετείται ευρέως. Με την προϋπόθεση της χρήσης ενός μικρού αριθμού κρυφών μονάδων στα υψηλότερα επίπεδα δημιουργούνται συνθήκες μάθησης πιο αφαιρετικών χαρακτηριστικών των δεδομένων [12]. Εμπειρικά εφαρμόζεται η δομή του πύργου μειώνοντας κατά το ήμισυ το μέγεθος του στρώματος κάθε διαδοχικού ανώτερου επιπέδου.

### Συστήματα Προτάσεων Φιλτραρίσματος Περιεχομένου (Content-based Filtering)

Τα συστήματα προτάσεων φιλτραρίσματος περιεχομένου έχουν κύριο στόχο την ανάλυση των βασικών χαρακτηριστικών και ιδιοτήτων των προϊόντων. Αναγκαία προϋπόθεση είναι η ύπαρξη ενός συνόλου δεδομένων που θα περιλαμβάνει αναλυτικά τα χαρακτηριστικά κάθε προϊόντος. Τα δεδομένα αυτά υφίστανται απαραίτητη επεξεργασία ώστε να εξαχθούν τα πιο σημαντικά χαρακτηριστικά και να αποβληθεί η περιττή πληροφορία. Τα χαρακτηριστικά κάθε ρούχου ορίζονται από την βιομηχανία παραγωγής τους. Το μέγεθος, η μάρκα, το μάκρος των μανικιών και άλλα χαρακτηριστικά είναι λίγο πολύ τετριμμένα δηλαδή αποτελούν έννοιες που χρησιμοποιούνται ευρέως σε όλο τον κόσμο. Δυστυχώς, τα συστήματα βασισμένα στο περιεχόμενο υποφέρουν από σημαντικούς περιορισμούς που τα καθιστούν δύσχρηστα και συνεπώς τις περισσότερες φορές συνδυάζονται με άλλες τεχνικές και μεθόδους. Ένα σημαντικό πρόβλημα είναι το πρόβλημα της εξειδίκευσης. Τα προϊόντα που προτείνονται από το σύστημα περιορίζονται κυρίως στα χαρακτηριστικά των προϊόντων που παρουσίασε αρχικά ο χρήστης. Με αυτό τον τρόπο, είναι πρακτικά αδύνατο να παρουσιαστούν νέα προϊόντα στον χρήστη που να καλύπτουν διαφορετικές ανάγκες του. Τέλος ιδιαίτερο πρόβλημα αποτελεί η αδυναμία δημιουργίας προφίλ και παροχής συστάσεων στους νέους χρήστες του συστήματος εξαιτίας της έλλειψης σε δεδομένα που περιγράφουν προηγούμενες αγορές ή προτιμήσεις τους. Tα συστήματα προτάσεων φιλτραρίσματος περιεχομένου βλέπουν τα χαρακτηριστικά κάθε προϊόντος που ο χρήστης δηλώνει ότι του αρέσει ή όχι και εξάγει κάποια συμπεράσματα για τις προτιμήσεις του. Υπάρχουν διάφορα συστήματα προτάσεων φιλτραρίσματος περιεχομένου, παρακάτω θα αναλυθούν αυτά που υλοποιήθηκαν στο κομμάτι αυτής της διπλωματικής εργασίας.

#### Τυχαία Δάση (Random Forest)

Τα τυχαία δάση είναι μοντέλα ιδανικά για την ανάλυση των βασικών χαρακτηριστικών και ιδιοτήτων των προϊόντων αφού αποτελούν μια ειδική κατηγορία συνδυαστικών μεθόδων ταξινόμησης που χρησιμοποιεί ως ταξινομητές δένδρα απόφασης.



Εικόνα 3 Οπτικοποίηση μοντέλου: Τυχαία Δάση [13]

Για την δημιουργία ενός δέντρου απόφασης (decision tree) ανατίθενται αρχικά στη ρίζα του τα δείγματα εκπαίδευσης. Κάθε κόμβος περιλαμβάνει ένα υποσύνολο των δειγμάτων το οποίο μέσω της εφαρμογής ενός κατάλληλου ελέγχου το διαχωρίζει σε δυο ή περισσότερα μικρότερα υποσύνολα σε κάθε διαδοχικό επίπεδο. Ο έλεγχος συνήθως αφορά ένα υποσύνολο των χαρακτηριστικών των δειγμάτων εκπαίδευσης. Η επιλογή του καλύτερου διαχωρισμού γίνεται σύμφωνα με ένα κατάλληλο μέτρο όπως π.χ. Gini index, εντροπία, misclassification error. Κάθε δένδρο του δάσους αναπτύσσει το μέγιστο βάθος που μπορεί να αναπτύξει ή μέχρι ορισμένο όριο βάθους που έχει τεθεί πριν την διαδικασία εκπαίδευσης. Η μέθοδος Bagging χρησιμοποιώντας για ταξινομητές δέντρα απόφασης αποτελεί μια ειδική κατηγορία των Random Forests. Σ ’ αυτή την περίπτωση η τυχαιότητα ενσωματώνεται στο μοντέλο και μέσω της τυχαίας επιλογής Ν παραδειγμάτων εκπαίδευσης, με επανατοποθέτηση, από το αρχικό σύνολο εκπαίδευσης. Η διαδικασία ταξινόμησης «άγνωστων» παραδειγμάτων πραγματοποιείται μέσω της διάσχισης των δέντρων του δάσους ξεκινώντας από τη ρίζα και καταλήγοντας σε ένα από τα φύλλα του δέντρου και στη συνέχεια συνδυάζοντας τις προβλέψεις των ταξινομητών σύμφωνα με ένα πλειοψηφικό σύστημα ψηφοφορίας (majority voting scheme). Κάθε παράδειγμα ανατίθεται στην κατηγορία με τη μεγαλύτερη συχνότητα.

#### Βαθιά Νευρωνικά Δίκτυα Περιεχομένου (Deep Content Neural Networks)

Η ευρύτερη κατηγορία νευρωνικών δικτύων μιμείται σε μεγάλο βαθμό τα βιολογικά νευρωνικά δίκτυα αναθέτοντας τις λειτουργίες των νευρώνων σε ένα απλό στοιχείο ικανό να αθροίσει την είσοδο του και να κανονικοποιεί την έξοδο. Οι νευρώνες είναι συνδεδεμένοι σε αυθαίρετα σύνθετα τεχνητά νευρωνικά δίκτυα. Για ταξινόμηση συνήθως χρησιμοποιούνται νευρωνικά δίκτυα εμπρόσθιας τροφοδότησης πολλών επιπέδων (multi-layer perceptions).



Εικόνα 4 Αρχιτεκτονική Νευρωνικών Δικτύων

Οι νευρώνες οργανώνονται σε επίπεδα:

* ένα επίπεδο εισόδου (input layer) το οποίο αντιστοιχεί στα χαρακτηριστικά
* ένα ή περισσότερα κρυμμένα επίπεδα (hidden layers)
* ένα επίπεδο εξόδου (output layer) το οποίο αντιστοιχεί στις κατηγορίες ή τιμές εξόδου ανάλογα τον τύπο του νευρωνικού.

Στόχος του αλγορίθμου μάθησης είναι να καθορίσει τα βάρη των συνδέσεων μεταξύ των νευρώνων (τα οποία χρησιμοποιούνται για να υπολογίσουμε σταθμισμένα αθροίσματα σε κάθε νευρώνα) με στόχο να μειώσει το ποσοστό σφάλματος ταξινόμησης ή το μέσο τετραγωνικό σφάλμα παλινδρόμησης, ανάλογα με τις προδιαγραφές που έχουν τεθεί. Για την ταξινόμηση ή παλινδρόμηση ενός νέου στοιχείου εισόδου οι τιμές των χαρακτηριστικών εφαρμόζονται στις εισόδου στους νευρώνες εισόδου του νευρωνικού δικτύου. Αυτές οι τιμές σταθμίζονται σύμφωνα με τις συνδέσεις μεταξύ των νευρώνων και τα σταθμισμένα αθροίσματά τους υπολογίζονται σε κάθε νευρώνα του επόμενου επιπέδου νευρώνων. Τα αποτελέσματα δίνονται στους νευρώνες εξόδου και αποτελούν την τελική εκτίμηση του νευρωνικού δικτύου.

### Υβριδικά Συστήματα Προτάσεων (Hybrid)

Η ανάλυση που ακολούθησε στις προηγούμενες ενότητες αναδεικνύει ότι το φάσμα των συστημάτων προτάσεων ποικίλει, με κάθε τεχνική να έχει τα δικά της δυνατά και αδύναμα σημεία. Βέβαια κάθε σύστημα προτάσεων φαίνεται να δρα περιοριστικά όταν δρα μεμονωμένα σε προβλήματα που υπάρχουν πολλές πηγές δεδομένων και το σύστημα αξιοποιεί μόνο ένα μέρος αυτής της μερίδας. Τα Υβριδικά Συστήματα Προτάσεων προσπαθούν να άρουν αυτό ακριβώς τον περιορισμό. Σε γενικές γραμμές είναι επιθυμητό να γίνει αξιοποίηση όλων των διαθέσιμων δεδομένων από διαφορετικές πηγές για τη δημιουργία ισχυρών συμπερασμάτων. Σε αυτή την ενότητα θα γίνει η εξερεύνηση αυτών των δυνατοτήτων με συστήματα που συνδυάζουν/εναλλάσσουν δύο ή περισσότερες τεχνικές για να επιτύχουν καλύτερη απόδοση και να περιορίσουν όσο περισσότερο γίνεται τα μειονεκτήματα οποιουδήποτε μεμονωμένου συστήματος πρότασης. Φυσικά ο τρόπος με τον οποίο συνδυάζονται τα συστήματα ποικίλει. Σύμφωνα με την βιβλιογραφία υπάρχουν τρεις τρόποι σχεδιασμού υβριδικών συστημάτων:

1. Σχεδιασμός ενσωμάτωσης (ensemble): Σε αυτό τον σχεδιασμό υβριδικών συστημάτων δύο ή περισσότεροι είσοδοι συστημάτων προτάσεων συνδυάζονται και μετασχηματίζονται σε μια έξοδο. Ένα παράδειγμα τέτοιων συστημάτων θα μπορούσε να είναι ο συνδυασμός ενός συστήματος προτάσεων ανάλυσης περιεχομένου με ένα σύστημα προτάσεων συνεργατικού φιλτραρίσματος.
2. Μονολιθικός Σχεδιασμός (monolithic): Σε αυτή την περίπτωση δημιουργείται ένας ολοκληρωμένος αλγόριθμος προτάσεων κάνοντας χρήση διαφορετικούς τύπους δεδομένων. Πολλές φορές σε αυτό τον σχεδιασμό συστημάτων δεν υπάρχει σαφής διάκριση μεταξύ των αλγορίθμων επομένως προσεγγίσεις τέτοιου τύπου τείνουν να ενσωματώσουν διάφορες πηγές δεδομένων.
3. Μικτός σχεδιασμός (mixed): Όπως και στα συστήματα ενσωμάτωσης αυτά τα συστήματα ενσωματώνουν πολλαπλούς αλγορίθμους ως μαύρα κουτιά (black boxes) ωστόσο παρουσιάζονται όλες οι προτάσεις από κάθε σύστημα η μια δίπλα στην άλλη.



Εικόνα 5 Ταξινομία Υβριδικών Συστημάτων [14]

Τα υβριδικά συστήματα προτάσεων μπορούν να ταξινομηθούν στις ακόλουθες κατηγορίες [14]:

1. Σταθμισμένα συστήματα (weighted): Σε αυτήν την περίπτωση, οι βαθμολογίες πολλών συστημάτων συνίστανται σε μια ενιαία βαθμολογία, υπολογίζοντας τα σταθμισμένα βάρη των βαθμολογιών που αντιστοιχούν σε κάθε μοντέλο. Η μεθοδολογία στάθμισης μπορεί να είναι ευριστική, ή μπορεί να χρησιμοποιεί επίσημα στατιστικά μοντέλα.
2. Συστήματα εναλλαγής (switching): Σε αυτά τα συστήματα ένας αλγόριθμος εναλλάσσει τα μοντέλα ανάλογα τις τρέχουσες ανάγκες.
3. Συστήματα Cascade: Σε αυτή την περίπτωση των συστημάτων πολλά συστήματα προτάσεων βρίσκονται σε σειρά έτσι ώστε ένα σύστημα προτάσεων να εξάγει προτάσεις που για το επόμενο σύστημα αποτελούν είσοδο βελτιώνοντας διαδοχικά την τελική πρόταση.
4. Συστήματα Feature Augmentation: Όπως και στα συστήματα Cascade η έξοδος ενός συστήματος προτάσεων χρησιμοποιείται στην είσοδο για το επόμενο. Ενώ το σύστημα cascade βελτιώνει διαδοχικά τις προτάσεις του προηγούμενου συστήματος, το σύστημα feature augmentation αντιμετωπίζει ως χαρακτηριστικά την είσοδο και την έξοδο για το επόμενο σύστημα.
5. Συστήματα Feature Combination: Σε αυτά τα συστήματα χαρακτηριστικά διαφορετικών πηγών δεδομένων συνδυάζονται και χρησιμοποιούνται ως ένα μοναδικό σύστημα προτάσεων.
6. Συστήματα Meta-Level: Όπως και στα cascade η έξοδος χρησιμοποιείται ως είσοδος σε διαδοχικό σύστημα. Ο σύνηθες συνδυασμός που χρησιμοποιείται είναι αυτός του συνεργατικού φιλτραρίσματος με του φιλτραρίσματος περιεχομένου. Το σύστημα τροποποιείται ώστε τεχνικές συνεργατικού φιλτραρίσματος να περιέχουν χαρακτηριστικά περιεχομένου προϊόντων. Με αυτό τον τρόπο το σύστημα θεωρείται ότι δρα ως ένα ενιαίο μοντέλο γι αυτό και κατατάσσεται στην κατηγορία των μονολιθικών συστημάτων.
7. Μεικτά Συστήματα: Όπως και στα συστήματα ενσωμάτωσης αυτά τα συστήματα ενσωματώνουν πολλαπλούς αλγορίθμους ως μαύρα κουτιά (black boxes) ωστόσο παρουσιάζονται όλες οι προτάσεις από κάθε σύστημα η μια δίπλα στην άλλη.

#### Σταθμισμένα Υβριδικά Συστήματα

Στα σταθμισμένα υβριδικά συστήματα, βαθμολογίες διαφορετικών συστημάτων προτάσεων συνδυάζονται για τον υπολογισμό της εξόδου του συστήματος. Οι αξιολογήσεις προϊόντων από χρήστες περιέχονται στον πίνακα αξιολογήσεων *διαστάσεων* . Ένα υβριδικό σύστημα που περιλαμβάνει συστήματα προτάσεων παράγει νέους πίνακες αξιολογήσεων . Οι έξοδοι των συστημάτων συνδυάζονται με την χρήση ενός συνόλου βαρών για την εκτίμηση του τελικού αποτελέσματος. Για ένα σύνολο βαρών η συνάρτηση εξόδου του σταθμισμένου υβριδικού συστήματος ορίζεται ώς:

Στην πιο απλή περίπτωση οι τιμές για τα βάρη παίρνουν την τιμή της ίσης ψήφου, στην συγκεκριμένη περίπτωση αυτή είναι . Ωστόσο αυτή η περίπτωση είναι αρκετά ιδανική καθώς τα συστήματα προτάσεων ποικίλουν. Για να επιλεχθούν τα βέλτιστα βάρη είναι απαραίτητο να αξιολογηθεί η επίδοση της αποτελεσματικότητας του συνδυασμού των βαρών Υπάρχουν διαφορετικές μέθοδοι για τον υπολογισμό των βαρών. H προαναφερθείσα εξίσωση μπορεί να γραφεί σε όρους μεμονωμένων καταχωρίσεων του πίνακα:

Όπου είναι η πρόβλεψη του συστήματος πρότασης που αφορά τον χρήστη για το προϊόν ενώ    αφορά την τελική πρόβλεψη.

Για να προσδιοριστούν τα βέλτιστα βάρη, είναι απαραίτητο να αξιολογηθεί η αποτελεσματικότητα ενός συγκεκριμένου συνδυασμού βαρών ., Μια απλή προσέγγιση είναι να κρατηθεί ένα μικρό κλάσμα (25%) από γνωστές καταχωρήσεις του πίνακα βαθμολογίας διαστάσεων και να δημιουργηθούν σύμφωνα με τους πίνακες πρόβλεψης εφαρμόζοντας διαφορετικούς συνδυασμούς βαρών στο υπόλοιπο 75% των καταχωρήσεων στο R. Για ένα δεδομένο συνδυασμό βαρών = (), η αποτελεσματικότητα μπορεί να αξιολογηθεί χρησιμοποιώντας είτε το μέσο τετράγωνο σφάλμα (MSE) είτε το μέσο απόλυτο σφάλμα (MAE) του προβλεπόμενου πίνακα:

Οι παραπάνω μετρικές παρέχουν μια αξιολόγηση ενός συγκεκριμένου συνδυασμού βαρών . Μια απλή προσέγγιση που μπορεί να προσδιορίσει τις βέλτιστες τιμές του για την ελαχιστοποίηση αυτών των μετρικών, είναι η χρήση γραμμικής παλινδρόμησης (linear regression). [15]

## Μετρικές αξιολόγησης (Χρήσιμες μετρικές)

### Ομοιότητα δεδομένων

Υπάρχουν διαφορές μετρικές για τον υπολογισμό της ομοιότητας των δεδομένων, οι επικρατέστερες είναι η μετρική Pearson (Pearson correlation similarity) και η μετρική του συνημίτονου (Cosine similarity). Κάθε μέθοδος επιστρέφει έναν αριθμό από το -1 ως το 1. Όσο μεγαλύτερος είναι ο αριθμός αυτός τόσο μεγαλύτερη είναι θεωρείται πως είναι η ομοιότητα των δεδομένων. Στο παράδειγμα που ακολουθεί έχουμε ένα πίνακα αξιολογήσεων χρηστών v και u και αντίστοιχα τους τύπους για τον υπολογισμό της ομοιότητας των δύο χρηστών.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *cloth 1* | *cloth 2* | *cloth 3* | *cloth 4* | *cloth 5* |
| *User: v* | 7 |  | 3 | 8 | 2 |
| *User: u* | 9 | 6 | 2 |  | 2 |

Πίνακας 1 - Αναπαράσταση δεδομένων και εύρεση ομοιότητας

Εξίσωση 1 - Μετρική ομοιότητας pearson

Εξίσωση 2 - Μετρική ομοιότητας συνημίτονου

Όπου οι θέσεις που βαθμολόγησε ο χρήστης *u*, οι θέσεις που βαθμολόγησε ο χρήστης *v* και οι θέσεις που βαθμολόγησαν και οι δύο χρήστες *u* και *v.*

Αντίστοιχα και είναι οι βαθμολογίες των χρηστών u και v στο προϊόν k ενώ και είναι η μέσες βαθμολογίες των χρηστών u και v.

Από το πιο πάνω παράδειγμα προκύπτει ότι

Συνεπώς οι ομοιότητες των χρηστών θα είναι ως προς pearson και για την ομοιότητα συνημίτονου.

### Ρίζα μέσου τετραγωνικού σφάλματος (RMSE)

Η ρίζα μέσου τετραγωνικού σφάλματος είναι μια μετρική αξιολόγησης της εκτίμησης ενός μοντέλου πρόβλεψης. Δεδομένου ότι το σφάλμα πρόβλεψης τετραγωνίζεται η μετρική RMSE δίνει βάρος περισσότερο στα σφάλματα με μεγάλη απόκλιση και λιγότερο σε αυτά με μικρή. [16] Αυτό σημαίνει ότι για μοντέλα που τα μεγάλα σφάλματα είναι ανεπιθύμητα ή ακόμη και απαγορευτικά, η μετρική αυτή θα πρέπει να λαμβάνεται σοβαρά υπ’ όψη. Πιο αναλυτικά η μετρική δίνεται από τον τύπο:

όπου είναι οι τιμές εκτίμησης, είναι οι πραγματικές τιμές και είναι ο αριθμός των πραγματικών τιμών.

### Μέσο απόλυτο σφάλμα (MAE)

To μέσο απόλυτο σφάλμα είναι μια μετρική αξιολόγησης που μετρά πόσο κοντά εκτιμήθηκαν οι προβλέψεις του συστήματος σε σχέση με τις πραγματικές βαθμολογίες. Ο τύπος περιγράφεται από την σχέση:

όπου είναι οι τιμές εκτίμησης, είναι οι πραγματικές τιμές και είναι ο αριθμός των πραγματικών τιμών.

Το μέσο απόλυτο σφάλμα σε αντίθεση με την ρίζα του μέσου τετραγωνικού σφάλματος λαμβάνει υπ’ όψη τα μεγάλα και μικρά σφάλματα εκτίμησης με το ίδιο βάρος [16] ενώ αγνοεί την κατεύθυνση του σφάλματος αφού περιλαμβάνει την μέση τιμή στον υπολογισμό του τύπου.

### Κάλυψη (Coverage)

Σύμφωνα με την βιβλιογραφία [17] η κάλυψη (coverage) ενός συστήματος προτάσεων είναι το ποσοστό των στοιχείων από το σύνολο του συστήματος για τα οποία το ίδιο το σύστημα μπορεί να κάνει προτάσεις. Ο όρος κάλυψη συσχετίστηκε κυρίως με δύο έννοιες:

* Με το ποσοστό των προϊόντων για τα οποία το σύστημα είναι σε θέση να δημιουργήσει μια πρόταση
* Με το ποσοστό των διαθέσιμων προϊόντων που πράγματι συνιστώνται ποτέ σε χρήστες

Παρόλο που διαφορετικοί συγγραφείς διαφέρουν ως προς την ορολογία, στην παρούσα διπλωματική εργασία υιοθετείται ο πρώτος ορισμός δηλαδή ως κάλυψη προβλέψεων από το σύνολο του καταλόγου. [18]

### Ποικιλομορφία (diversity)

Υπάρχουν πολλοί ορισμοί που περιγράφουν την έννοια της ποικιλομορφίας ως μετρική, η απλούστερη είναι αυτή του Bradley και Smyth που ορίζουν την ποικιλομορφία ως το αντίθετο της ομοιότητας η οποία και υιοθετείται στην παρούσα διπλωματική. [19] Πιο συγκεκριμένα θεωρούμε την ποικιλομορφία ώς τον μέσο όρο της ανομοιότητας μεταξύ όλων των πιθανών ζευγών που μπορούν να προκύψουν από μια πρόταση. Ο τύπος περιγράφεται από την σχέση:

Είναι σαφές ότι η ποιότητα της μέτρησης εξαρτάται σε μεγάλο βαθμό από την συνάρτηση ομοιότητας που θα εφαρμοστεί στο σύνολο των δεδομένων. [20]

### Καινοτομία (novelty)

Η έννοια της καινοτομίας (novelty) στα συστήματα προτάσεων σύμφωνα με τη βιβλιογραφία για ένα δεδομένο χρήστη, καθορίζεται από την αναλογία των άγνωστων αντικειμένων στη λίστα προτάσεων προς το σύνολο. Δηλαδή είναι ένα μέγεθος που εκφράζει το ποσοστό των προτάσεων που τα προϊόντα δεν είχαν στο παρελθόν κάποια αλληλεπίδραση απ’ τον δεδομένο χρήστη δηλαδή του ήταν άγνωστα [21]. Αναλυτικά ο τύπος περιγράφεται από την σχέση:

Όπου το R είναι το σύνολο των προτεινόμενων προϊόντων και είναι μια δυαδική συνάρτηση που επιστρέφει 1 εάν ο δεδομένος χρήστης γνωρίζει ήδη το προϊόν διαφορετικά επιστρέφει 0. Καθώς τα συστήματα προτάσεων αδυνατούν με μεγάλη σιγουριά να κρίνουν εάν ο χρήστης δεν αλληλοεπίδρασε ποτέ με προϊόν, υπάρχει μια προσέγγιση θεωρεί τα λιγότερο δημοφιλή προϊόντα ως καινοτόμες προτάσεις στους χρήστες [22].

Κεφάλαιο

# Υλοποιήσεις συστημάτων προτάσεων

Το παρόν κεφάλαιο περιλαμβάνει την περιγραφή προ-επεξεργασίας των δεδομένων που απαιτήθηκε καθώς και την λειτουργία κάθε υλοποίησης συμπεριλαμβανομένου τα δεδομένα εισόδου της, τα βήματα των τεχνικών της και τις μεθόδους αξιολόγησης των επιδόσεων της.

## Προ-επεξεργασία Δεδομένων

### Συλλογή και οργάνωση δεδομένων

Το αρχικό σετ δεδομένων δημιουργήθηκε μέσω ενός συστήματος συλλογής δεδομένων από τον μεγάλο κολοσσό οίκου μόδας Asos. Οι κατηγορίες του αρχικού σετ έχουν ανομοιογενή δεδομένα δηλαδή δεν ακολουθούν ομοιόμορφη κατανομή γεγονός λογικό αφού στην αγορά κάθε κατηγορία προϊόντων δεν έχει τον ίδιο αριθμό λόγω ζήτησης. Παρόλα αυτά ο αριθμός των κατηγοριών φάνηκε να είναι αρκετά υψηλός με αποτέλεσμα πολλές κατηγορίες να είναι επικαλυπτόμενες δηλαδή να σχετίζονται σε μεγάλο βαθμό η κατηγορία που εκπροσωπούν.



Εικόνα 6 Διάγραμμα ρούχων ανά κατηγορία αρχικού σετ δεδομένων

Η παρακάτω εικόνα δείχνει μια περίπτωση επικάλυψης όπου πολλά παντελόνια είναι καταχωρημένα και σαν φόρμες, πολλά σορτσάκια είναι καταχωρημένα επίσης σαν φόρμες ενώ τέλος κάποια παντελόνια είναι καταχωρημένα ταυτόχρονα ως σορτσάκια και φόρμες.



Εικόνα 7 Αναπαράσταση περίπτωσης επικάλυψης ρούχων ανά κατηγορία

Η διπλωματική εργασία περιλαμβάνει δέκα κατηγορίες ρούχων όσο κατά το δυνατόν μη επικαλυπτόμενες και όχι απαραίτητα ομοιόμορφα κατανεμημένες.



Εικόνα 8 Διάγραμμα ρούχων ανά κατηγορία τελικού σετ δεδομένων

### Χαρακτηριστικά Προϊόντων (features)

Κάθε προϊόν μπορεί να περιγράφεται από χαρακτηριστικά τα οποία υποδηλώνουν άμεσα ή έμμεσα την κατηγορία ή την χρησιμότητά του προϊόντος, δηλαδή πληροφορίες απαραίτητες για τα συστήματα προτάσεων. Το σύνολο δεδομένων περιλαμβάνει ρούχα τα οποία αποτελούνται από χαρακτηριστικά και μετρικές ευρέως χρησιμοποιούμενες στις εταιρίες ρούχων όπως για παράδειγμα το μέγεθος (small, medium, large κτλ.) το είδος του υφάσματος(μάλλινο, βαμβακερό), η φόρμα του ρούχου (κολλητό, χαλαρό) και το είδος γιακά του. Σκοπός αυτού του βήματος της προ-επεξεργασίας είναι η δημιουργία ενός διανύσματος χαρακτηριστικών που θα εκφράζει κάθε ρούχο με τα χαρακτηριστικά που το διέπουν.

Στην παρούσα διπλωματική χαρακτηριστικά τα οποία εκφράζουν ένα ρούχο θεωρήθηκαν:

* Το νούμερο του ρούχου:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SHORT | MEDIUM | KNEE | LONG | CAPRI | ¾ LONG |
|  |  |  |  |  |  |

* Η φόρμα του ρούχου:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| RELAXED FIT | REGULAR FIT | SLIM FIT | CARGO FIT | BIKER FIT |
|  |  |  |  |  |

* Ο σχεδιασμός γιακά του:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ROUND NECK | COLLAR | V NECK | HOODED | OFF SHOULDER | HALTERNECK | TURTLENECK |
|  |  |  |  |  |  |  |

* Η κατηγορία ρούχου:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PYJAMAS | SKIRT | TROUSERS | … | CARDIGAN | SHIRT |
|  |  |  |  |  |  |

* Οι λέξεις κλειδιά που το χαρακτηρίζουν:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SHORTBLOUSES | SHORTTROUSERS | … | KNEESET | COLLARDRESS | VNECKPYJAMAS |
|  |  |  |  |  |  |

Μετά την εξαγωγή τον πιο σημαντικών χαρακτηριστικών που εκφράζουν κάθε προϊόν ακολουθεί η ένωση όλων αυτών των χαρακτηριστικών ώστε να δημιουργηθεί ένα ενιαίο σύνολο όπου κάθε προϊόν θα εκφράζεται ως ένα διάνυσμα πολλών διαστάσεων με κάθε διάσταση να αποτελεί και μια ξεχωριστή ιδιότητα.



Εικόνα 9 Δημιουργία Πίνακα Συγκεντρωτικών Χαρακτηριστικών

### Βαθμολογίες Προϊόντων

Για τα δεδομένα που συλλέχθηκαν από την ιστοσελίδα της εταιρίας Asos δημιουργήθηκαν δύο σετ δεδομένων με τεχνητές αξιολογήσεις χρηστών οι οποίες προσεγγίζουν την ανθρώπινη συμπεριφορά με τρόπο ώστε:

* Οι χρήστες βαθμολογούν σε προϊόντα αρκετά παρόμοια με παρόμοιο τρόπο έχοντας μικρές διακυμάνσεις
* Οι χρήστες δεν βαθμολογούν απαραίτητα τα ίδια ρούχα
* Οι βαθμολογίες κάθε χρήστη δεν είναι ομοιόμορφα κατανεμημένες φαινόμενο που συναντάται σε πραγματικά σετ δεδομένων

Έχοντας ως καθοδήγηση τις παραπάνω προϋποθέσεις δημιουργούνται δύο σετ δεδομένων αξιολογήσεων. To πρώτο περιλαμβάνει αξιολογήσεις συνολικού ύψους 30% του συστήματος ενώ το δεύτερο ύψους 60%.

## Συνεργατικό φιλτράρισμα με επίκεντρο τον χρήστη

### Είσοδος δεδομένων (data – input)

Στην παρούσα υλοποίηση η είσοδος δεδομένων που απαιτείται για την εξαγωγή προτάσεων είναι ο πίνακας χρηστών-αξιολογήσεων που περιλαμβάνει τις αξιολογήσεις κάθε χρήστη για όλα τα ρούχα που περιέχονται στη βάση δεδομένων. Παρακάτω φαίνεται μια τυπική αναπαράσταση του πίνακα χρηστών-αξιολογήσεων:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ρούχο 1 | ρούχο 2 | … | ρούχο m-1 | ρούχο m |
| χρήστης 1 | 5 | 5 | … | 6 | ? |
| χρήστης 2 | 4 | 9 | … | 3 | 2 |
| …. | ….. | … | … | ? | … |
| χρήστης n | 7 | 9 | … | 8 | ? |

Πίνακας 2 - Αναπαράσταση του πίνακα χρηστών-αξιολογήσεων

### Βήματα υλοποίησης

#### Εύρεση όμοιων χρηστών

Το αμέσως επόμενο βήμα αυτής της υλοποίησης είναι η εύρεση των πιο όμοιων χρηστών ως προς τον χρήστη που είναι το επίκεντρο. Για το παράδειγμα που ακολουθεί θεωρείται επίκεντρο ο χρήστης 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ρούχο 1 | … | ρούχο m-1 | ρούχο m | ομοιότητα |
| χρήστης 1 | 5 | … | 6 |  | 1 |
| χρήστης 2 | 4 | … | 3 | 2 | 0.67 |
| …. | ….. | … | … | … |  |
| χρήστης n | 7 | … | 8 |  | 0.789 |

Πίνακας 3 - Αναπαράσταση του πίνακα χρηστών-αξιολογήσεων - Υπολογισμός ομοιότητας

#### Στάθμιση σημασίας ομοιότητας (significance weighting)

Η αξιοπιστία του αποτελέσματος της συνάρτησης ομοιότητας πολλές φορές εξαρτάται από τον αριθμό των κοινών αξιολογήσεων για τους συγκρινόμενους χρήστες. Όταν δύο χρήστες έχουν μικρό αριθμό κοινών αξιολογήσεων τότε το αποτέλεσμα της ομοιότητας θα πρέπει να μειώνεται από έναν παράγοντα που θα τονίζει την σημασία των κοινών αξιολογήσεων. Αυτή η μέθοδος γνωστή και ως στάθμιση της ομοιότητας (significance weighting) [23] δεδομένου των κοινών αξιολογήσεων ορίζεται ένα threshold b τέτοιο ώστε:

Εξίσωση 3 - Αναπαράσταση τύπου στάθμισης σημασίας ομοιότητας

#### Κεντράρισμα μέση τιμής δεδομένων (mean-center data)

Εφόσον ολοκληρωθεί η εύρεση των πιο όμοιων χρηστών μέσα από την στάθμιση της ομοιότητας είναι αναγκαίο το κεντράρισμα της μέσης τιμής των δεδομένων κάθε όμοιου χρήστη για την προσαρμογή του bias των αξιολογήσεων.

To bias των αξιολογήσεων προκαλείται όταν οι χρήστες τείνουν να δίνουν πάντα υψηλές ή πάντα χαμηλές βαθμολογίες σε όλα τα στοιχεία. [24]

Η νέα αξιολόγηση κάθε χρήστη προκύπτει από τον τύπο

Όπου η αξιολόγηση του χρήστη u στο ρούχο j και η μέση τιμή των αξιολογήσεων.

Παρακάτω ακολουθεί σχηματικά η αναπαράσταση του νεου πίνακα χρηστών-αξιολογήσεων με κεντραρισμένη μέση τιμή

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Ρούχο 1 | … | Ρούχο m | Μέση τιμή | Σταθμισμένη ομοιότητα |
| Χρήστης 1 |  | … |  |  | …. |
| …. | …. | … | …. | … | …. |
| Χρήστης n |  |  |  |  | …. |

Πίνακας 4 - Αναπαράσταση πίνακα με κεντραρισμένη μέση τιμή αξιολογήσεων

#### Συνάρτηση πρόβλεψης αξιολόγησης

Οι προβλέψεις των απροσδιόριστων αξιολογήσεων υπολογίζονται ως ο σταθμισμένος μέσος όρος των αποκλίσεων από τον μέσο όρο του κάθε χρήστη προσθέτοντάς τον στη μέση βαθμολογία του χρήστη που είναι στο επίκεντρο.

Εξίσωση 4 - Τύπος της πρόβλεψης των απροσδιόριστων αξιολογήσεων

Παραπάνω παρουσιάζεται ο τύπος της πρόβλεψης των απροσδιόριστων αξιολογήσεων.

Αν u είναι ο χρήστης που έχουμε στο επίκεντρο και j το ρούχο που θέλουμε να εκτιμήσουμε τότε   είναι η εκτιμώμενη αξιολόγηση. Όπου η μέση τιμή των αξιολογήσεων του χρήστη u, το σύνολο των k πιο όμοιων χρηστών, η ομοιότητα του χρήστη u με τον χρήστη v και

## Συνεργατικό φιλτράρισμα με επίκεντρο τον προϊόν

### Προ-επεξεργασία δεδομένων (pre-processing)

Στην παρούσα υλοποίηση σκοπός της προεπεξεργασίας δεδομένων είναι παρόμοιος με τις προηγούμενης δηλαδή η δημιουργία ενός πίνακα χρηστών-αξιολογήσεων που θα περιλαμβάνει τις αξιολογήσεις κάθε χρήστη για όλα τα ρούχα που περιέχονται στη βάση δεδομένων.

### Κεντράρισμα μέση τιμής δεδομένων (mean-center data)

Αμέσως μετά την δημιουργία του πίνακα χρηστών-αξιολογήσεων είναι αναγκαίο το κεντράρισμα της μέσης τιμής των δεδομένων του κάθε χρήστη για την προσαρμογή του bias των αξιολογήσεων.

Αντίθετα με την προηγούμενη εφαρμογή το κεντράρισμα της μέσης τιμής των αξιολογήσεων γίνεται πριν την εύρεση των πιο όμοιων ρούχων.

### Εύρεση όμοιων προϊόντων (ρούχων)

Έχοντας κεντράρει την μέση τιμή των αξιολογήσεων κάθε χρήστη στο μηδέν το επόμενο βήμα αυτής της υλοποίησης είναι η εύρεση των πιο όμοιων ρούχων. Ο υπολογισμός της ομοιότητας με τη χρήση του βασικού μέτρου συνημίτονο (regular cosine similarity) σε περίπτωση που βασίζεται σε προϊόντα (item-based) έχει ένα σημαντικό μειονέκτημα - η διαφορά στην κλίμακα βαθμολογίας μεταξύ διαφορετικών χρηστών δεν λαμβάνεται υπόψη. Η προσαρμοσμένη ομοιότητα συνημίτονου αντισταθμίζει αυτό το μειονέκτημα αφαιρώντας τον αντίστοιχο μέσος όρο των χρηστών από κάθε προσωπική του βαθμολογία. [25]

Θεωρούμε τις και της κεντραρισμένες ως προς την μέση τιμή αξιολογήσεις ενώ αντίστοιχα και τις θέσεις που έχουν υποβάλει οι χρήστες και σε προϊόντα.

Εξίσωση 5 - Αναπαράσταση τύπου προσαρμοσμένης ομοιότητας

### Συνάρτηση πρόβλεψης αξιολόγησης

## Συνεργατικό φιλτράρισμα με ανάλυση πίνακα σε ιδιάζουσες τιμές

### Είσοδος Δεδομένων (data – input)

Όμοια με την προηγούμενη υλοποίηση, η είσοδος δεδομένων που απαιτείται για την εξαγωγή προτάσεων είναι ο πίνακας χρηστών-αξιολογήσεων που περιλαμβάνει τις αξιολογήσεις κάθε χρήστη για όλα τα ρούχα που περιέχονται στη βάση δεδομένων. Παρακάτω φαίνεται μια τυπική αναπαράσταση του πίνακα χρηστών-αξιολογήσεων:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ρούχο 1 | ρούχο 2 | … | ρούχο m-1 | ρούχο m |
| χρήστης 1 | 5 | 5 | … | 6 | ? |
| χρήστης 2 | 4 | 9 | … | 3 | 2 |
| …. | ….. | … | … | ? | … |
| χρήστης n | 7 | 9 | … | 8 | ? |

Πίνακας 5 Αναπαράσταση του πίνακα χρηστών-αξιολογήσεων

### Βήματα υλοποίησης

#### Επιλογή βιβλιοθήκης μοντέλου

## Συνεργατικό φιλτράρισμα μέσω νευρωνικών δικτύων

### Δεδομένα εισόδου

Όμοια με τις δύο προηγούμενες υλοποιήσεις, η είσοδος δεδομένων που απαιτείται για την εξαγωγή προτάσεων είναι ο πίνακας χρηστών-αξιολογήσεων που περιλαμβάνει τις αξιολογήσεις κάθε χρήστη για όλα τα ρούχα που περιέχονται στη βάση δεδομένων.

### Βήματα υλοποίησης

#### Κωδικοποίηση Χρηστών-Ρούχων

Τα πεδία χρηστών / ρούχων που υπάρχουν στο σύνολο δεδομένων με τις αξιολογήσεις είναι μη διαδοχικοί ακέραιοι αριθμοί που αντιπροσωπεύουν κάποιο μοναδικό αναγνωριστικό. Για την λειτουργία αυτής της υλοποίησης χρειάζεται αυτοί οι αριθμοί να είναι διαδοχικοί ξεκινώντας από το μηδέν έως το τελικό εύρος. Για την επίτευξη αυτής της ανάγκης γίνεται η χρήση της κλάσης LabelEncoder του scikit-learn.

#### Επίπεδο Ενσωμάτωσης

Εφόσον ολοκληρωθεί η κωδικοποίηση δεδομένων του αρχικού σετ και ο διαχωρισμός των δεδομένων σε σετ εκπαίδευσης και ελέγχου ακολουθεί η τροφοδότηση του επιπέδου ενσωμάτωσης, ένα πλήρες συνδεδεμένο επίπεδο το οποίο μαθαίνει να προβάλλει την αναπαράσταση χρηστών ή προϊόντων σε πυκνά διανύσματα. Το ενσωματωμένο επίπεδο για τον χρήστη όπως και το ενσωματωμένο επίπεδο για το προϊόν αντίστοιχα μπορεί να θεωρηθεί ότι περιλαμβάνει τα λανθάνον διανύσματα όμοια των συστημάτων ανάλυσης ιδιάζουσας τιμής (SVD). Η κλάση που χρησιμοποιείται είναι η *tf.keras.layers.Embedding* μέσω της οποίας δημιουργείται το επίπεδο ενσωμάτωσης χρηστών και επίπεδο ενσωμάτωσης προϊόντων.

#### Αρχιτεκτονική πολύ-επίπεδο νευρωνικού δικτύου

To ενσωματωμένο επίπεδο εξάγει την αναπαράσταση κάθε χρήστη και προϊόν σε πυκνά διανύσματα, χαρακτηριστικά τα οποία περιλαμβάνουν αλληλεπιδράσεις μεταξύ των χρηστών και είναι απαραίτητο να αναλυθούν για να δημιουργηθούν προτάσεις. Μια απλή συνένωση των χαρακτηριστικών δε λαμβάνει υπόψη καμία αλληλεπίδραση μεταξύ χρηστών και προϊόντων των λανθάνων χαρακτηριστικών που είναι απαραίτητη για την εφαρμογή του συνεργατικού φιλτραρίσματος. Για την διευθέτηση αυτού του ζητήματος γίνεται η χρήση του νευρωνικού δικτύου ώστε να γίνει η εκπαίδευση για την αναγνώριση αυτών των αλληλεπιδράσεων και την αναγνώριση μη γραμμικών σχέσεων λειτουργία αδύνατη για μοντέλα ανάλυσης πίνακα ιδιαζουσών τιμών. Αυτά τα χαρακτηριστικά λοιπόν εισέρχονται σε ένα πολύ-επίπεδο νευρωνικό δίκτυο του οποίου η αρχιτεκτονική μοιάζει με πύργου, μια αρχιτεκτονική ευρέως χρησιμοποιούμενη που επιτρέπει στο δίκτυο να αναγνωρίζει στα υψηλά στρώματα χαρακτηριστικά πιο αφαιρετικά [12]. Τα επίπεδα δημιουργούνται με την χρήση της κλάσης keras.layers.Dense και τα βάρη κάθε διαδοχικού επιπέδου από το αρχικό μειώνονται κατά το ήμισυ. Η συνάρτηση ενεργοποίησης που χρησιμοποιείται είναι η ReLU σε όλα τα επίπεδα εκτός του τελικού που γίνεται η χρήση της sigmoid. Η αρχικοποίηση των βαρών κάθε επιπέδου γίνεται μέσω της κανονικής κατανομής he\_normal ενώ η αρχικοποίηση του τελευταίου επιπέδου μέσω της κανονικής Glorot κατανομής. [26]

## Φιλτράρισμα Περιεχομένου με Τυχαία Δάση

### Δεδομένα εισόδου

Για τις υλοποιήσεις περιεχομένου η είσοδος δεδομένων που απαιτείται για την εξαγωγή προτάσεων είναι ο πίνακας χαρακτηριστικών κάθε ρούχου και ο πίνακας χρηστών-αξιολογήσεων που περιλαμβάνει τις αξιολογήσεις κάθε χρήστη για όλα τα ρούχα που περιέχονται στη βάση δεδομένων.

### Βήματα υλοποίησης

#### Επιλογή βιβλιοθήκης μοντέλου

Για την εξαγωγή προτάσεων κάθε Τυχαίο Δάσος χρειάζεται να εκπαιδευτεί με της αξιολογήσεις του χρήστη και να αναλύσει τα χαρακτηριστικά κάθε ρούχου ώστε να αντιληφθεί τις προτιμήσεις του. Πρώτο βήμα αυτής της υλοποίησης είναι για κάθε χρήστη να λαμβάνει τα απαραίτητα δεδομένα δηλαδή κάθε αξιολόγηση που αντιστοιχεί στα χαρακτηριστικά ενός ρούχου. Η υλοποίηση τρέχει κάθε Τυχαίο Δάσος κάνοντας χρήση της βιβλιοθήκης RandomForestRegressor της sklearn, μια βιβλιοθήκη που δοθέντος ενός αριθμού δέντρων αποφάσεων ταξινόμησης σε διάφορα υπο-δείγματα του συνόλου δεδομένων χρησιμοποιεί τον μέσο όρο για τη βελτίωση της ακρίβειας πρόγνωσης των προτάσεων.

#### Επιλογή παραμέτρων

Ένα τυχαίο δάσος μπορεί να εκπαιδευτεί έχοντας του τεθεί διαφορετικές ρυθμίσεις παραμέτρων. Ο αριθμός εκτιμητών και το μέγιστο βάθος κάθε δένδρου απόφασης των τυχαίων δασών είναι οι βασικότεροι παράμετροι που επηρεάζουν την επίδοση των μοντέλων. Γι’ αυτό το σκοπό έγινε χρήση της βιβλιοθήκης ανοιχτού κώδικα python: Optuna η οποία αναζητά σε ένα χώρο περιορισμένου εύρους τις βέλτιστες παραμέτρους που δίνουν επιθυμητές επιδόσεις τα μοντέλα Τυχαίων Δασών

## Φιλτράρισμα Περιεχομένου με Νευρωνικά Δίκτυα

### Δεδομένα Εισόδου

Για τις υλοποιήσεις περιεχομένου η είσοδος δεδομένων που απαιτείται για την εξαγωγή προτάσεων είναι ο πίνακας χαρακτηριστικών κάθε ρούχου και ο πίνακας χρηστών-αξιολογήσεων που περιλαμβάνει τις αξιολογήσεις κάθε χρήστη για όλα τα ρούχα που περιέχονται στη βάση δεδομένων.

### Βήματα υλοποίησης

#### Επιλογή βιβλιοθήκης μοντέλου

Παρόμοια με τα Τυχαία Δάση, για εξαγωγή προτάσεων κάθε Νευρωνικό Δίκτυο χρειάζεται να εκπαιδευτεί με τις αξιολογήσεις του χρήστη και να αναλύσει τα χαρακτηριστικά κάθε προϊόντος ώστε να αντιληφθεί τις προτιμήσεις του. Πρώτο βήμα της υλοποίησης είναι για κάθε χρήστη να λαμβάνει τα απαραίτητα δεδομένα δηλαδή κάθε αξιολόγηση που αντιστοιχεί στα χαρακτηριστικά κάθε ρούχου. Η υλοποίηση τρέχει το Νευρωνικό Δίκτυο κάνοντας χρήση την κλάση keras.models.Sequential της βιβλιοθήκης Keras. Τα επίπεδα του δικτύου δημιουργούνται με κλάση keras.layers.Dense συνοδευόμενη από επίπεδα tf.keras.layers.Dropout που απενεργοποιούν τυχαία ένα ποσοστό βαρών του στρώματος και αποτρέπουν την υπέρ-εκπαίδευση του μοντέλου.

#### Αρχιτεκτονική Νευρωνικού Δικτύου

Κάθε χρήστης έχει αξιολογήσει μια σειρά ρούχων με διαφορετικές βαθμολογίες και κάθε κριτική εκφράζει με πολύπλοκο τρόπο τις προτιμήσεις του. Για την αναγνώριση αυτών των προτιμήσεων και την εξαγωγή ισχυρών συμπερασμάτων γίνεται η χρήση του Νευρωνικού Δικτύου. Τα χαρακτηριστικά αυτά εισέρχονται σε ένα πολύ-επίπεδο νευρωνικό δίκτυο του οποίου η αρχιτεκτονική είναι ίδια με της υλοποίησης του συνεργατικού νευρωνικού δικτύου που μοιάζει με πύργου. Όπως αναφέρθηκε ήδη η συγκεκριμένη αρχιτεκτονική είναι ευρέως χρησιμοποιούμενη που επιτρέπει στο δίκτυο να αναγνωρίζει στα υψηλά στρώματα χαρακτηριστικά πιο αφαιρετικά [12]. Κάθε επίπεδο του δικτύου δημιουργείται με την κλάση Dense συνοδευόμενη με επίπεδα Dropout για την τυχαία απενεργοποίηση μικρού ποσοστού από τα βάρη. Το αρχικό επίπεδο έχει το μεγαλύτερο αριθμό βαρών ενώ κάθε διαδοχικό μειώνεται κατά το ήμισυ. Η συνάρτηση ενεργοποίησης που χρησιμοποιείται σε όλα τα επίπεδα είναι η ReLU εκτός του τελικού που είναι η sigmoid. Η αρχικοποίηση βαρών είναι ακριβώς αντίστοιχη του Συνεργατικού Νευρωνικού Δικτύου [26].

Κεφάλαιο

# Πειράματα και αποτελέσματα

## Διαχωρισμός δεδομένων

Τα σύνολα δεδομένων για την διεξαγωγή των πειραμάτων αποτελούνται από χρήστες που έχουν αξιολογήσει προϊόντα με αρνητικές, ενδιάμεσες και θετικές βαθμολογίες. Καθώς ο μοναδικός τρόπος αξιολόγησης των προτάσεων ήταν η χρήση των ήδη γνωστών βαθμολογιών, τα δεδομένα χωρίστηκαν ώστε: το 80% των αξιολογήσεων κάθε χρήστη να αφορούν τα δεδομένα εκπαίδευσης, το 10% των αξιολογήσεων κάθε χρήστη να αφορούν τα δεδομένα επικύρωσης και το υπόλοιπο 10% τα δεδομένα ελέγχου.



Εικόνα 10 Αναπαράσταση Διαχωρισμού Δεδομένων - 1

Από αυτό τον διαχωρισμό δεδομένων εξαιρούνται υλοποιήσεις που κατά την εκπαίδευση τους δε χρησιμοποιούν δεδομένα επικύρωσης. Αυτές οι υλοποιήσεις θα έχουν το ίδιο σετ ελέγχου και ελαφρός αυξημένο σετ εκπαίδευσης.



Εικόνα 11 Αναπαράσταση Διαχωρισμού Δεδομένων - 2

## Σύνολα δεδομένων

Ο παραπάνω διαχωρισμός που περιγράφηκε εφαρμόζεται σε 3 σύνολα δεδομένων. Τα δύο πρώτα αφορούν ρούχα της εταιρίας ASOS και περιλαμβάνουν αξιολογήσεις από 100 χρήστες σε 1700 ρούχα. Η διαφορά μεταξύ αυτών των δύο συνόλων είναι ότι το πρώτο περιλαμβάνει το 30% των αξιολογήσεων σε σχέση με το σύνολο των αξιολογήσεων που μπορεί να κάνει ο χρήστης ενώ το δεύτερο το 60% των αξιολογήσεων σε σχέση πάλι με το σύνολο. Μέσα από την εξαγωγή πειραμάτων σε αυτά τα δύο σύνολα δεδομένων μπορεί να γίνει αντιληπτή η επιρροή των αξιολογήσεων στις προτάσεις και να σχολιαστεί κατά πόσο τις βελτιώνει αφού το σύστημα στην μια περίπτωση διαθέτει περισσότερα δεδομένα να εκπαιδευτεί απ’ ότι στην άλλη. Το τρίτο σύνολο δεδομένων είναι το movielens-100k περιλαμβάνει 9000 ταινίες για τις οποίες υπάρχουν 100 χιλιάδες αξιολογήσεις από 600 χρήστες και έχει δημοσιευθεί από την ομάδα GroupLens. Το γεγονός ότι δύο πρώτα σύνολα δεδομένων αποτελούνται από τεχνητές αξιολογήσεις ενώ το τρίτο σύνολο δεδομένων από πραγματικές αξιολογήσεις είναι μια καλή αφορμή να σχολιαστεί η επίδοση του γενικευμένου συστήματος προτάσεων σε τεχνητά και πραγματικά δεδομένα ως τεκμήριο της αξιοπιστία του. Οι αξιολογήσεις στα πρώτα δύο σύνολα δεδομένων της εταιρίας asos κυμαίνονται στην κλίμακα [0 εως 10] ενώ οι αξιολογήσεις του movielens με τις ταινίες στην κλίμακα [0 εως 5]. Παρακάτω παρουσιάζονται συνοπτικά σε ένα πίνακα τα σύνολα δεδομένων:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Σύνολο Δεδομένων | Ρούχα / Ταινίες | Χρήστες | Αξιολογήσεις | Κλίμακα |
| Asos30 | 1700 | 100 | 50k | [0,10] |
| Asos60 | 1700 | 100 | 100k | [0,10] |
| Movielens-100k | 1700 | 600 | 100k | [0,5] |

Πίνακας 6 Συνοπτική αναπαράσταση για κάθε σύνολο δεδομένων

## Μετρικές Αξιολόγησης

### Accuracy, Recall, Precision και f1\_score

Ένα σύστημα προτάσεων έχει σαν τελικό στόχο να προβάλει στον χρήστη προϊόντα ή υπηρεσίες όσο κατά το δυνατόν πιο κοντά στις προτιμήσεις του. Γι’ αυτό ακριβώς τον λόγο είναι απαραίτητο να αξιολογήσουμε πόσο πετυχημένα υλοποιούν το καθήκον τους αυτά τα συστήματα, με μετρικές όπως το accuracy, το precision, και το recall θεωρώντας τις προτάσεις σαν ένα πρόβλημα δυαδικής ταξινόμησης όπου ένα σύστημα με χαμηλό precision και recall θα υστερεί αφού θα αδυνατεί να αναγνωρίσει τις πραγματικές προτιμήσεις του χρήστη.



Εικόνα 12 Διαχωρισμός αξιολογήσεων [0,10]



Εικόνα 13 Διαχωρισμός αξιολογήσεων [0,5]

O διαχωρισμός των αξιολογήσεων και στα 3 σύνολα δεδομένων έγινε στην μέση του εύρους. Για τα 2 σύνολα δεδομένων Asos με κλίμακα [0 εως 10] θεωρούμε τις αξιολογήσεις μεγαλύτερες του 5 ως πιθανές αξιολογήσεις κοντά στις προτιμήσεις του χρήστη ενώ για το σύνολο δεδομένων movielens με κλίμακα [0 εως 5] θεωρούμε τις αξιολογήσεις μεγαλύτερες του 2.5 ως πιθανές αξιολογήσεις κοντά στις προτιμήσεις του χρήστη.

### RMSE και MAE

Η χρήση των παραπάνω 4 μετρικών δίνει ικανοποιητικά αποτελέσματα για την αξιολόγηση της ικανότητας των προτάσεων του συστήματος ωστόσο από μόνη της δεν αρκεί. Σε περιπτώσεις που οι χρήστες τείνουν να βαθμολογούν συχνά αρκετά υψηλά ή χαμηλά τα αποτελέσματα των 4 μετρικών ίσως παραπλανητικά παρουσιάσουν κάποια επιτυχία ή αποτυχία του συστήματος η οποία δεν αντικατοπτρίζεται στην πραγματικότητα. Αυτό τον περιορισμό έρχονται να άρουν οι μετρικές RSME και MAE. Δεδομένου ότι το σύστημα εκτιμά την αξιολόγηση σε ένα προϊόν και στην συνέχεια συγκρίνει αυτή την εκτίμηση με την πραγματική τιμή, είναι εμφανές πως υπάρχει ανάγκη για τη χρήση κάποιας μετρικής η οποία δείχνει πόσο κοντά είναι η εκτίμηση του συστήματος στην αληθινή προτίμηση του χρήστη. Μια μετρική που ικανοποιεί αυτή την ανάγκη είναι η RMSE με την οποία υπολογίζουμε την ρίζα του μέσου τετραγωνικού σφάλματος. Ο υπολογισμός αυτής της μετρικής τιμωρεί περισσότερο τα μεγάλα σφάλματα σε σχέση με τα μικρότερα λόγω του τετραγώνου με αποτέλεσμα όταν το σύστημα εκτιμά υψηλές τιμές αξιολογήσεων σε προϊόντα που έχουν αξιολογηθεί χαμηλά (και αντίστροφα) η μετρική αυτή αυξάνει δραματικά. Από την άλλη η μετρική MAE τιμωρεί εξίσου το σφάλμα εκτίμησης μικρών και μεγάλων αποκλίσεων δηλαδή αντικατοπτρίζει την γενική περίπτωση κάθε εκτίμησης του συστήματος σε σχέση με την πραγματική τιμή.

## Υλοποίηση συνεργατικού φιλτραρίσματος με επίκεντρο τον χρήστη (User-Based)

Η παρούσα υλοποίηση χρησιμοποιεί τον δεύτερο διαχωρισμό δεδομένων. Τα δεδομένα εκπαίδευσης χρησιμοποιούνται και δημιουργούν προτάσεις για το σετ ελέγχου. Οι προτάσεις αυτές συγκρίνονται με τις πραγματικές τιμές που υπάρχουν στο σετ ελέγχου και έτσι αξιολογείται η ικανότητα εξαγωγής προτάσεων του μοντέλου.

### Accuracy, recall, precision και f1\_score

Το διάγραμμα των τεσσάρων μετρικών **ανά γείτονα** για τα σύνολα data30, data60 δείχνει πως η ικανότητα πρόβλεψης του μοντέλου **αλλάζει συναρτήσει του αριθμού γειτόνων**. Για μεγάλες τιμές αριθμού γειτόνων έχουμε αύξηση του recall συνεπώς και του ποσοστού προτάσεων που πράγματι αρέσουν στους χρήστες και εντοπίστηκαν από το μοντέλο χωρίς ιδιαίτερη μείωση του precision δηλαδή του ποσοστού επιτυχίας των προϊόντων που θεωρήθηκαν αρεστά από το μοντέλο και όντως ήταν.

Σετ δεδομένων: data30



Εικόνα 14 Υλοποίηση User-Based - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score ανά γείτονα – data30

Σετ δεδομένων: data60

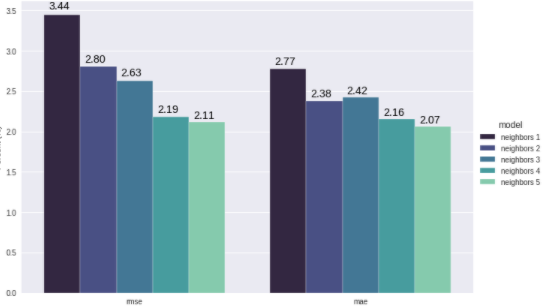


Εικόνα 15 Υλοποίηση User-Based - Διάγραμμα μετρικών αccuracy, recall, precision και f1\_score ανά γείτονα – data60

### Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE)

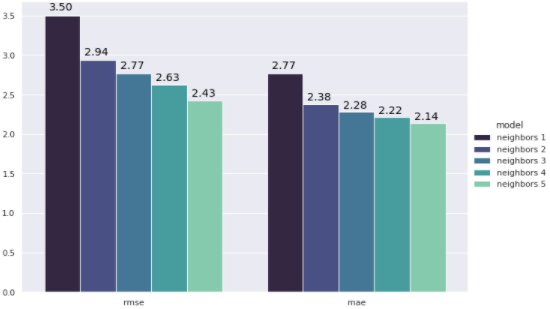
Τα διαγράμματα των ΜΑΕ και RMSE ανά γείτονα δείχνουν πως η αύξηση των γειτόνων στο μοντέλο **μειώνει αισθητά τις δύο μετρικές σφάλματος.**

**Σετ δεδομένων: data30**



Εικόνα 16 Υλοποίηση User-Based - Διάγραμμα μετρικών rmse και mae ανά γείτονα – data30

**Σετ δεδομένων: data60**



Εικόνα 17 Υλοποίηση User-Based - Διάγραμμα μετρικών rmse και mae ανά γείτονα - data60

### Κάλυψη και Ποικιλία

Τα διαγράμματα Κάλυψης και Ποικιλίας δείχνουν πως η παρούσα υλοποίηση του συνεργατικού φιλτραρίσματος **αποτυγχάνει να ανακαλύψει ικανοποιητικά τα προϊόντα της βάσης** ωστόσο παρέχει μια ισορροποιημένη ποικιλία ανά κατηγορία ρούχου. Ένας λόγος είναι ότι η φύση αυτής της υλοποίησης έχει ισχυρή εξάρτηση μεταξύ των αξιολογήσεων από χρήστη σε χρήστη. Όπως θα φανεί οι επόμενες υλοποιήσεις ξεπερνούν αυτό τον περιορισμό.

**Σετ δεδομένων: data30**

Εικόνα 18 Υλοποίηση User-Based - Διάγραμμα Κάλυψης και Ποικιλίας – data30

**Σετ δεδομένων: data60**



Εικόνα 19 Υλοποίηση User-Based - Διάγραμμα Κάλυψης και Ποικιλίας – data60

## Υλοποίηση συνεργατικού φιλτραρίσματος με ανάλυση πίνακα ιδιαζουσών τιμών (SVD)

Η παρούσα υλοποίηση χρησιμοποιεί τον πρώτο διαχωρισμό δεδομένων. Τα δεδομένα επαλήθευσης χρησιμοποιούνται για την βέλτιστη επιλογή της τιμής k των χαρακτηριστικών και στην συνέχεια εφαρμόζεται στο σετ ελέγχου. Τέλος δημιουργούνται προτάσεις, συγκρίνονται με τις πραγματικές τιμές που υπάρχουν στο σετ ελέγχου και έτσι αξιολογείται η ικανότητα εξαγωγής προτάσεων του μοντέλου.

### Accuracy, recall, precision και f1\_score

Τα διαγράμματα των τεσσάρων μετρικών ανά τιμή χαρακτηριστικών δείχνουν πως η ικανότητα πρόβλεψης του μοντέλου **δεν αλλάζει** για τα σύνολα data30, data60 ενώ **αλλάζει ελάχιστα** για το σύνολο movielens-100k. Συνεπώς, το γεγονός ότι υπάρχει καθόλου ή ελάχιστη επίδραση στην ακρίβεια του μοντέλου συναρτήσει των χαρακτηριστικών κάνει το σύστημα λιγότερο ευαίσθητο στην ρύθμιση του και περισσότερο αξιόπιστο.

Συγκρίνοντας τις επιδόσεις των συνόλων data30 με data60 βλέπουμε ότι η αύξηση 30% των κριτικών **έχει θετικές** επιδράσεις στα ζητήματα ταξινόμησης.

|  |  |
| --- | --- |
| Σετ δεδομένων: data30 | Σετ δεδομένων: data60 |
|  |  |
| Εικόνα 20 Υλοποίηση SVD - Διάγραμμα μετρικών αccuracy, recall, precision και f1\_score ανά τιμή χαρακτηριστικών – data30 | Εικόνα 21 Υλοποίηση SVD - Διάγραμμα μετρικών αccuracy, recall, precision και f1\_score ανά τιμή χαρακτηριστικών – data60 |

Σετ δεδομένων: movielens 100K



Εικόνα 22 Υλοποίηση SVD - Διάγραμμα μετρικών αccuracy, recall, precision και f1\_score ανά τιμή χαρακτηριστικών – movielens 100K

### Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE)

Τα διαγράμματα των ΜΑΕ και RMSE ανά τιμή χαρακτηριστικών δείχνουν πως η ικανότητα πρόβλεψης του μοντέλου **αλλάζει ελάχιστα** για όλα τα σύνολα ως προς το σφάλμα πρόβλεψης.

Συγκρίνοντας τις επιδόσεις των συνόλων data30 με data60 βλέπουμε ότι η αύξηση 30% των κριτικών **έχει θετικές** επιδράσεις στα ζητήματα ταξινόμησης.

|  |  |
| --- | --- |
| Σετ δεδομένων: data30 | Σετ δεδομένων: data60 |
|  |  |
| Εικόνα 23 Υλοποίηση SVD - Διάγραμμα μετρικών rmse και mae ανά τιμή χαρακτηριστικών – data30 | Εικόνα 24 Υλοποίηση SVD - Διάγραμμα μετρικών rmse και mae ανά τιμή χαρακτηριστικών – data30 |

Σετ δεδομένων: movielens 100K



Εικόνα 25 Υλοποίηση SVD - Διάγραμμα μετρικών rmse και mae ανά τιμή χαρακτηριστικών – movielens 100K

### Κάλυψη και Ποικιλία

Τα διαγράμματα Κάλυψης και Ποικιλίας δείχνουν πως η ικανότητα ανακάλυψης των προϊόντων της βάσης είναι ικανοποιητική για όλα τα σύνολα. Επίσης **υπάρχει ισορροπία στις προτάσεις που υπάρχουν ανα κατηγορία** δηλαδή το σύστημα δεν απορρίπτει μεγάλο εύρος ρούχων/ταινιών κάποιας κατηγορίας κάτι που θα α ήταν προβληματικό καθώς θα εμποδίζονταν η ανακάλυψη αυτής της κατηγορίας.

Σετ δεδομένων: data30



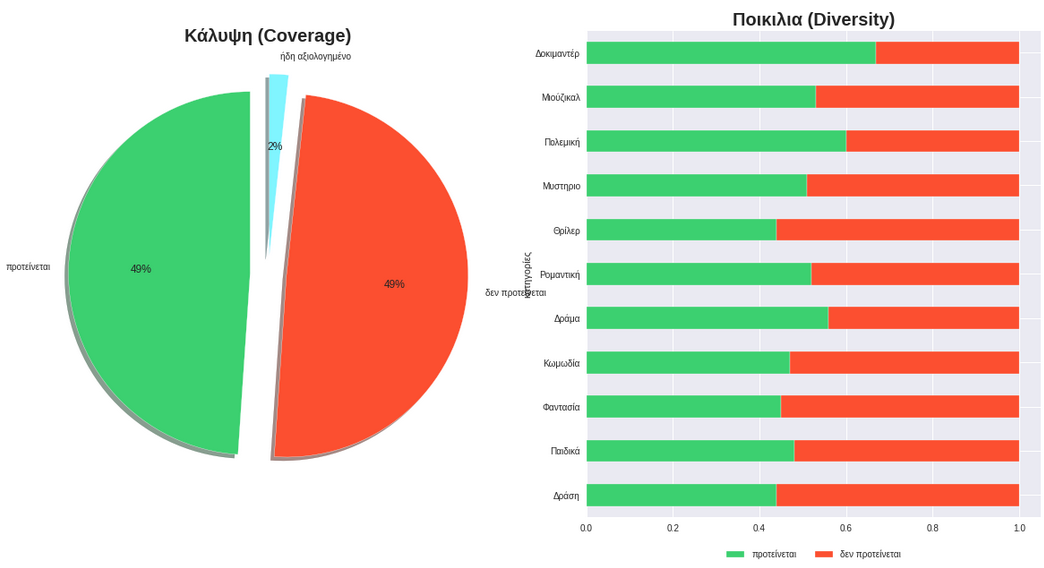
Εικόνα 26 Υλοποίηση SVD - Διάγραμμα Κάλυψης και Ποικιλίας – data30

Σετ δεδομένων: data60



Εικόνα 27 Υλοποίηση SVD - Διάγραμμα Κάλυψης και Ποικιλίας – data60

Σετ δεδομένων: movielens-100k



Εικόνα 28 Υλοποίηση SVD - Διάγραμμα Κάλυψης και Ποικιλίας – movielens-100k

## Υλοποίηση Συνεργατικού Νευρωνικού Φιλτραρίσματος

Η παρούσα υλοποίηση χρησιμοποιεί τον πρώτο διαχωρισμό δεδομένων. Τα δεδομένα επαλήθευσης χρησιμοποιούνται για την βέλτιστη εκπαίδευση του συνεργατικού νευρωνικού δικτύου. Αμέσως μετά δημιουργούνται προτάσεις, συγκρίνονται με τις πραγματικές τιμές που υπάρχουν στο σετ ελέγχου και έτσι αξιολογείται η ικανότητα εξαγωγής προτάσεων του μοντέλου.

### Accuracy, recall, precision και f1\_score

Το διάγραμμα των τεσσάρων μετρικών για όλα τα σύνολα **δείχνει ικανοποιητικά ποσοστά accuracy, recall και precision.** Ένα πλεονέκτημα αυτού του συστήματος είναι ότι προσαρμόζει με αυτοματοποιημένο τρόπο τις παραμέτρους του μοντέλου που είναι στην φύση των νευρωνικών δικτύων. Βλέπουμε για τα σύνολα δεδομένων data30, data60 ότι η αύξηση 30% των κριτικών **έχει ανεπαίσθητες επιδράσεις στα ζητήματα ταξινόμησης**. Από την άλλη το σύνολο δεδομένων movielens-100k παρόλο που δεν έχει το ίδιο υψηλό recall σε σχέση με data30, data60 **πετυχαίνει καλύτερα ποσοστά accuracy και precision** πραγμα που σημαίνει ότι έχει περισσότερη ακρίβεια στις προτάσεις αρεσκίας των χρηστών ακόμα κι ας μην της βρίσκει όλες.

|  |  |
| --- | --- |
| Σετ δεδομένων: data30 | Σετ δεδομένων: data60 |
|  |  |
| Εικόνα 29 Υλοποίηση NCF - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 | Εικόνα 30 Υλοποίηση NCF - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data60 |

Σετ δεδομένων: movielens 100K

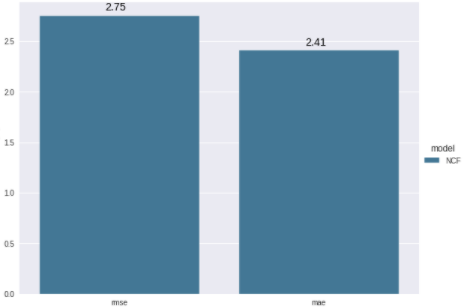


Εικόνα 31 Υλοποίηση NCF - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K

### Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE)

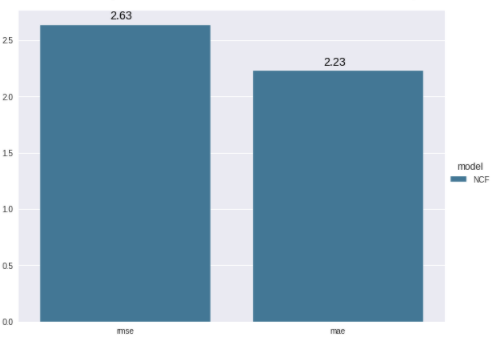
Τα διαγράμματα των ΜΑΕ και RMSE για τα σύνολα data30, data60 δείχνουν ότι η αύξηση 30% των κριτικών **έχει ουσιώδεις επιδράσεις** στο ζήτημα του σφάλματος αφου μειώνει το RMSE από 2,75 σε 2,63 και το MAE από 2,41 σε 2,23 αντίστοιχα. Τα διαγράμματα των ΜΑΕ και RMSE για το σύνολο movielens-100k έχει κι αυτό **ικανοποιητικά αποτελέσματα** αφού η ρίζα του μέσου τετραγωνικού σφάλματος παραμένει χαμηλότερη της μονάδας.

**Σετ δεδομένων: data30**



Εικόνα 32 Υλοποίηση NCF - Διάγραμμα μετρικών rmse και mae – data30

**Σετ δεδομένων: data60**



Εικόνα 33 Υλοποίηση NCF - Διάγραμμα μετρικών rmse και mae – data60

**Σετ δεδομένων: movielens 100K**



Εικόνα 34 Υλοποίηση NCF - Διάγραμμα μετρικών rmse και mae – movielens 100K

### Κάλυψη και Ποικιλία

Τα διαγράμματα Κάλυψης και Ποικιλίας για τα σύνολα data30, data60 και movielens-100k δείχνουν πως η παρούσα υλοποίηση (NCF) **ανακαλύπτει ικανοποιητικά τα προϊόντα της βάσης** στα σύνολα δεδομένων. Πιο συγκεκριμένα στα σύνολα δεδομένων data30, data60 **υπάρχει ισορροπία στις προτάσεις που υπάρχουν ανα κατηγορία** δηλαδή το σύστημα δεν απορρίπτει μεγάλο εύρος ρούχων κάποιας κατηγορίας κάτι που θα α ήταν προβληματικό καθώς δε θα εμπόδιζε του σχεδιαστές να ανακαλύψουν ρούχα αυτής της κατηγορίας. Το ίδιο ισχύει και για το σύνολο δεδομένων movieLens-100k το οποίο παρά τον χαμηλό αριθμό κρητικών καλύπτει και τις 9000 ταινίες του συστήματος απορρίπτοντας όχι σημαντικά μεγάλα ποσοστά ταινιών ανά κατηγορία.

**Σετ δεδομένων: data30**



Εικόνα 35 Υλοποίηση NCF - Διάγραμμα Κάλυψης και Ποικιλίας – data30

**Σετ δεδομένων: data60**



Εικόνα 36 Υλοποίηση NCF - Διάγραμμα Κάλυψης και Ποικιλίας – data60

**Σετ δεδομένων: movielens-100k**



Εικόνα 37 Υλοποίηση NCF - Διάγραμμα Κάλυψης και Ποικιλίας – movielens-100k

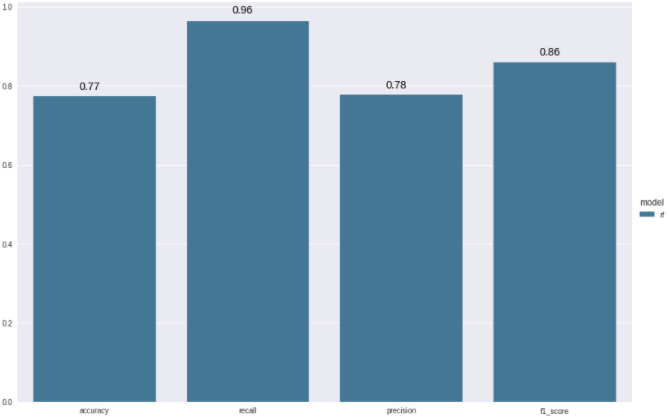
## Υλοποίηση Φιλτραρίσματος Περιεχομένου με Τυχαία Δάση

Η παρούσα υλοποίηση χρησιμοποιεί τον πρώτο διαχωρισμό δεδομένων. Τα δεδομένα επαλήθευσης χρησιμοποιούνται για την βέλτιστη εκπαίδευση του Τυχαίου Δάσους μεταβάλλοντας τον αριθμό εκτιμητών και το μέγιστο βάθος κάθε δένδρου ταξινόμησης. Στην συνέχεια το εκπαιδευμένο μοντέλο εφαρμόζεται στο σετ ελέγχου. Τέλος δημιουργούνται προτάσεις, συγκρίνονται με τις πραγματικές που υπάρχουν στο σετ ελέγχου και έτσι αξιολογείται η ικανότητα εξαγωγής προτάσεων του μοντέλου.

### Accuracy, recall, precision και f1\_score

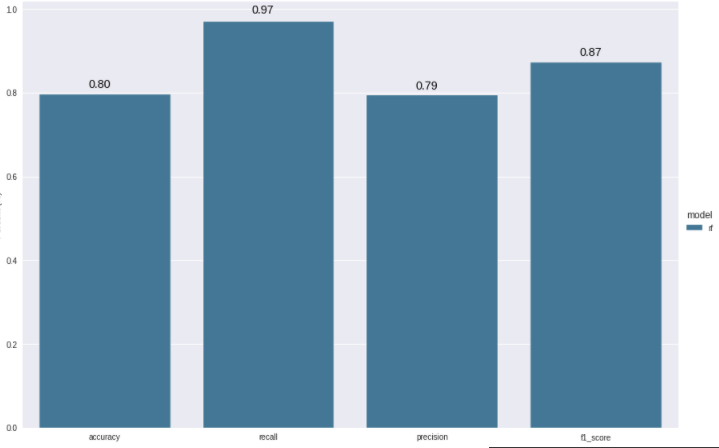
Το διάγραμμα των τεσσάρων μετρικών για τα σύνολα data30, data60 και movielens-100k **δείχνει ικανοποιητικά ποσοστά accuracy,recall και precision**. Ένα πλεονέκτημα αυτού του συστήματος είναι ότι προσαρμόζει με αυτοματοποιημένο τρόπο τις παραμέτρους του μοντέλου χρησιμοποιώντας την βιβλιοθήκη επιλογής βέλτιστων παραμέτρων optuna. Βλέπουμε για τα σύνολα δεδομένων data30, data60 ότι η αύξηση 30% των κριτικών **έχει ελαφρώς θετικές επιδράσεις στα ζητήματα ταξινόμησης**. Από την άλλη το σύνολο δεδομένων movielens-100k παρόλο που δεν έχει το ίδιο υψηλό recall σε σχέση με data30, data60 πετυχαίνει καλύτερα ποσοστά accuracy και precision πραγμα που σημαίνει ότι έχει περισσότερη ακρίβεια στις προτάσεις αρεσκίας των χρηστών ακόμα κι ας μην της βρίσκει όλες.

**Σετ δεδομένων: data30**



Εικόνα 38 Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30

**Σετ δεδομένων: data60**



Εικόνα 39 Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data60

**Σετ δεδομένων: movielens 100K**

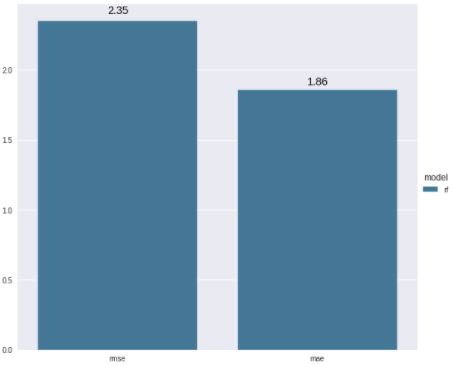


Εικόνα 40 Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K

### Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE)

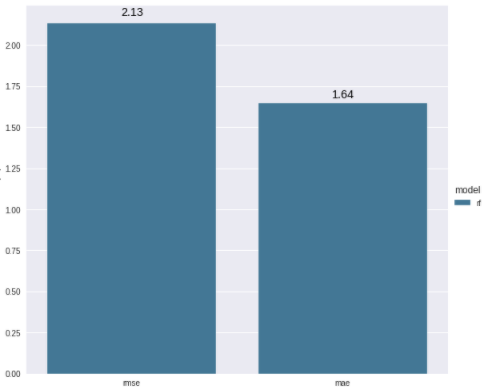
Τα διαγράμματα των ΜΑΕ και RMSE για τα σύνολα data30, data60 δείχνουν ότι η αύξηση 30% των κριτικών **έχει ουσιώδεις επιδράσεις** στο ζήτημα του σφάλματος αφου μειώνει το RMSE από 2,35 σε 2,13 και το MAE από 1,86 σε 1,64 αντίστοιχα. Τα διαγράμματα των ΜΑΕ και RMSE για το σύνολο movielens-100k έχει κι αυτό ικανοποιητικά αποτελέσματα αφού η ρίζα του μέσου τετραγωνικού σφάλματος παραμένει χαμηλότερη της μονάδας.

**Σετ δεδομένων: data30**



Εικόνα 41 Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών rmse και mae – data30

**Σετ δεδομένων: data60**



Εικόνα 42 Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών rmse και mae – data60

**Σετ δεδομένων: movielens 100K**



Εικόνα 43 Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών rmse και mae – movielens 100K

### Κάλυψη και Ποικιλία

Τα διαγράμματα Κάλυψης και Ποικιλίας για τα σύνολα data30, data60 και movielens-100k δείχνουν πως η παρούσα υλοποίηση (Τυχαία Δάση) ανακαλύπτει ικανοποιητικά τα προϊόντα της βάση στα σύνολα δεδομένων. Στα σύνολα δεδομένων data30, data60 υπάρχει ισορροπία στις προτάσεις που υπάρχουν ανα κατηγορία δηλαδή το σύστημα δεν απορρίπτει μεγάλο εύρος ρούχων κάποιας κατηγορίας. Το ίδιο ισχύει και για το σύνολο δεδομένων movieLens-100k το οποίο παρά τον χαμηλό αριθμό κρητικών καλύπτει και τις 9000 ταινίες του συστήματος απορρίπτοντας όχι σημαντικά μεγάλα ποσοστά ταινιών ανά κατηγορία.

**Σετ δεδομένων: data30**



Εικόνα 44 Υλοποίηση Τυχαίων Δασών - Διάγραμμα Κάλυψης και Ποικιλίας – data30

**Σετ δεδομένων: data60**



Εικόνα 45 Υλοποίηση Τυχαίων Δασών - Διάγραμμα Κάλυψης και Ποικιλίας – data60



Εικόνα 46 Υλοποίηση Τυχαίων Δασών - Διάγραμμα Κάλυψης και Ποικιλίας – movielens-100k

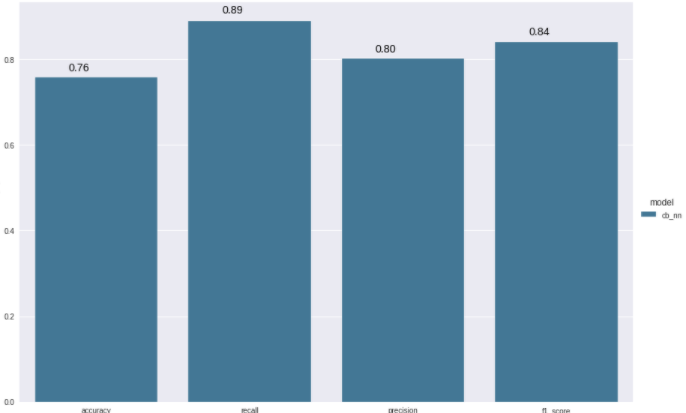
## Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα

Η παρούσα υλοποίηση χρησιμοποιεί τον πρώτο διαχωρισμό δεδομένων. Τα δεδομένα επαλήθευσης χρησιμοποιούνται για την βέλτιστη εκπαίδευση του νευρωνικού δικτύου φιλτραρίσματος περιεχομένου. Αμέσως μετά δημιουργούνται προτάσεις, συγκρίνονται με τις πραγματικές τιμές που υπάρχουν στο σετ ελέγχου και έτσι αξιολογείται η ικανότητα εξαγωγής προτάσεων του μοντέλου.

### Accuracy, recall, precision και f1\_score

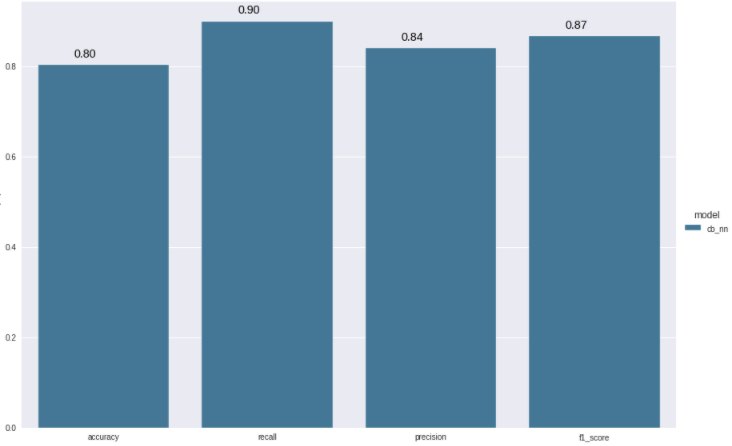
Το διάγραμμα των τεσσάρων μετρικών **για τα σύνολα data30, data60 και movielens-100k** δείχνει ικανοποιητικά ποσοστά accuracy,recall και precision. Ένα πλεονέκτημα αυτού του συστήματος είναι ότι **προσαρμόζει με αυτοματοποιημένο** αφού άλλωστε είναι στην φύση των νευρωνικών να προσαρμόζουν αυτόματα τις παραμέτρους τους. Βλέπουμε για τα σύνολα δεδομένων data30, data60 ότι η αύξηση 30% των κριτικών **έχει ελαφρώς θετικές** επιδράσεις στα ζητήματα ταξινόμησης χωρίς ωστόσο αυτή την φορά να έχουν recall όμοιο με τις προηγούμενες υλοποιήσεις. Από την άλλη το σύνολο δεδομένων movielens-100k **έχει αρκετά υψηλό recall** σε σχέση με τα σύνολα data30, data60 αλλά και **καλύτερα ποσοστά accuracy και precision.**

**Σετ δεδομένων: data30**



Εικόνα 47 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30

**Σετ δεδομένων: data60**



Εικόνα 48 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30

**Σετ δεδομένων: movielens 100K**

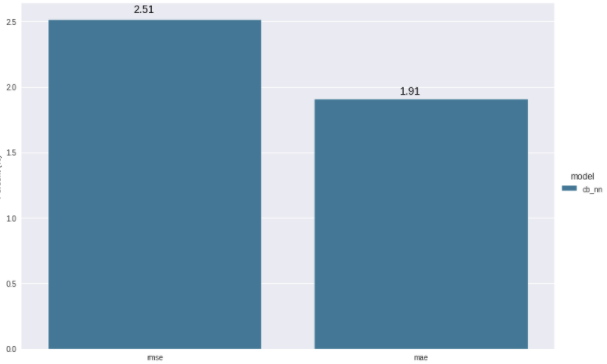


Εικόνα 49 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K

### Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE)

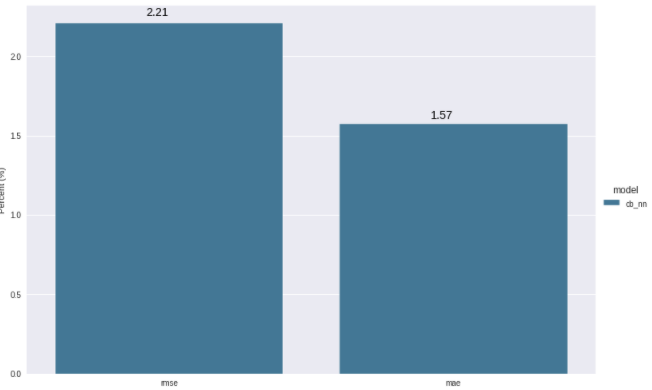
Τα διαγράμματα των ΜΑΕ και RMSE για τα σύνολα data30, data60 δείχνουν ότι η αύξηση 30% των κριτικών **έχει ουσιώδεις επιδράσεις** στο ζήτημα του σφάλματος αφου μειώνει το RMSE από 2,51 σε 2,21 και το MAE από 1,91 σε 1,57 αντίστοιχα. Τα διαγράμματα των ΜΑΕ και RMSE για το σύνολο movielens-100k έχει κι αυτό ικανοποιητικά αποτελέσματα αφού η ρίζα του μέσου τετραγωνικού σφάλματος παραμένει χαμηλότερη της μονάδας.

**Σετ δεδομένων: data30**



Εικόνα 50 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών rmse και mae – data30

**Σετ δεδομένων: data60**



Εικόνα 51 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών rmse και mae – data60

**Σετ δεδομένων: movielens 100K**



Εικόνα 52 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών rmse και mae – movielens 100K

### Κάλυψη και Ποικιλία

Τα διαγράμματα Κάλυψης και Ποικιλίας για τα σύνολα data30, data60 και movielens-100k δείχνουν πως η παρούσα υλοποίηση (Φιλτραρίσμα Περιεχομένου με Νευρωνικά Δίκτυα) **ανακαλύπτει ικανοποιητικά τα προϊόντα της βάση στα σύνολα δεδομένων**. Στα σύνολα δεδομένων data30, data60 **υπάρχει ισορροπία στις προτάσεις που υπάρχουν ανα κατηγορία** δηλαδή το σύστημα δεν απορρίπτει μεγάλο εύρος ρούχων κάποιας κατηγορίας. Το ίδιο ισχύει και για το σύνολο δεδομένων movieLens-100k το οποίο παρά τον χαμηλό αριθμό κρητικών καλύπτει και τις 9000 ταινίες του συστήματος απορρίπτοντας όχι σημαντικά μεγάλα ποσοστά ταινιών ανά κατηγορία.

**Σετ δεδομένων: data30**



Εικόνα 53 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα Κάλυψης και Ποικιλίας – data30

**Σετ δεδομένων: data60**



Εικόνα 54 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα Κάλυψης και Ποικιλίας – data60

**Σετ δεδομένων: movielens-100k**



Εικόνα 55 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα Κάλυψης και Ποικιλίας – movielens-100k

## Υλοποίηση Υβριδικού Συστήματος

Η παρούσα υλοποίηση χρησιμοποιεί τα παραπάνω μοντέλα που παρουσιάστηκαν. Αφού εξαχθούν οι προτάσεις από κάθε μοντέλο το υβριδικό σύστημα δημιουργεί δύο είδη προτάσεων. Το πρώτο είδος αφορά την εξαγωγή προτάσεων μέσω της λογικής ίσης ψήφου με αποτέλεσμα η τελική βαθμολογία του προϊόντος να είναι η μέση τιμή όλων των αξιολογήσεων από κάθε σύστημα. Το δεύτερο είδος αφορά την εξαγωγή προτάσεων μέσω της χρήσης ενός μοντέλου γραμμικής παλινδρόμησης. Έτσι το 30% των αξιολογήσεων χρησιμοποιούνται για την εκπαίδευση του μοντέλου το οποίο με την σειρά του δημιουργεί προτάσεις για το υπόλοιπο 70%.

### Accuracy, recall, precision και f1\_score

Το διάγραμμα των τεσσάρων μετρικών **για τα σύνολα data30, data60 και movielens-100k** δείχνει ικανοποιητικά ποσοστά accuracy,recall και precision στην γενική τους περίπτωση. Συγκεκριμένα και για τα 3 σύνολα δεδομένων βλέπουμε μια ελαφρά μείωση του recall και αύξηση των accuracy,precision και f1\_score πράγμα που δείχνει πως οι προτάσεις του υβριδικού μέσω γραμμικής παλινδρόμησεις δίνουν προτάσεις πετυχημένες με λιγότερα προϊόντα που τελικά δεν άρεσαν στον χρήστη χάνοντας ένα ποσοστό του recall διόλου σημαντικό.

**Σετ δεδομένων: data30**



Εικόνα 56 Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30

**Σετ δεδομένων: data60**



Εικόνα 57 Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data60

**Σετ δεδομένων: movielens 100K**

Εικόνα 58 Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K

### Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE)

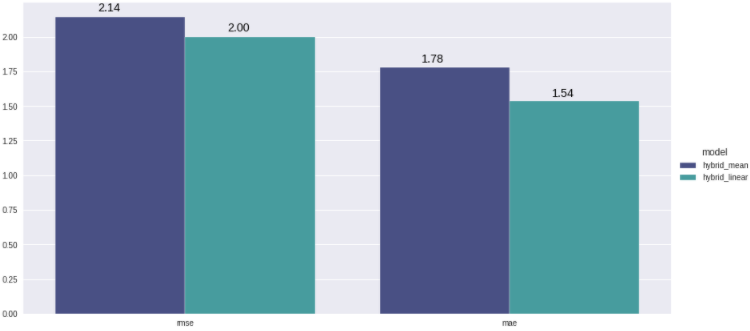
Τα διαγράμματα των ΜΑΕ και RMSE για τα σύνολα data30, data60 δείχνουν ότι η αύξηση 30% των κριτικών **έχει ουσιώδεις επιδράσεις** στο ζήτημα του σφάλματος. Το ίδιο συμβαίνει και με τα αποτελέσματα του υβριδικού συστήματος ίσης ψήφου (hybrid\_mean) με το υβριδικό σύστημα γραμμικής παλινδρόμησής (hybrid\_linear). Φαίνεται πως η γραμμική παλινδρόμηση μειώνει αισθητά το σφάλμα και για τα 3 σύνολα δεδομένων αφού δίνει περισσότερη βαρύτητα στα μοντέλα που εξάγουν καλές προτάσεις και λιγότερη στα μοντέλα που κάνουν λάθη.

**Σετ δεδομένων: data30**



Εικόνα 59 Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – data30

**Σετ δεδομένων: data60**



**Σετ δεδομένων: movielens 100K**



Εικόνα 60 Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – movielens 100K

### Κάλυψη και Ποικιλία

Τα διαγράμματα Κάλυψης και Ποικιλίας για τα σύνολα data30, data60 και movielens-100k δείχνουν πως το υβριδικό σύστημα **ανακαλύπτει ικανοποιητικά τα προϊόντα της βάση στα σύνολα δεδομένων**. Τα σύνολα δεδομένων data30, data60 **υπάρχει ισορροπία στις προτάσεις που υπάρχουν ανα κατηγορία** δηλαδή το σύστημα δεν απορρίπτει μεγάλο εύρος ρούχων κάποιας κατηγορίας. Το ίδιο ισχύει και για το σύνολο δεδομένων movieLens-100k το οποίο παρά τον χαμηλό αριθμό κρητικών καλύπτει και τις 9000 ταινίες του συστήματος απορρίπτοντας όχι σημαντικά μεγάλα ποσοστά ταινιών ανά κατηγορία.

**Σετ δεδομένων: data30**



Εικόνα 61 Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα Κάλυψης και Ποικιλίας – data30

**Σετ δεδομένων: data60**



Εικόνα 62 Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα Κάλυψης και Ποικιλίας – data60

## Συνοπτική αναπαράσταση όλων των υλοποιήσεων

### Accuracy, recall, precision και f1\_score

**Σετ δεδομένων: data30**

Εικόνα 63 Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30

**Σετ δεδομένων: movielens 100K**

Εικόνα 64 Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K

### Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE)

**Σετ δεδομένων: data30**

Εικόνα 65 Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – data30

**Σετ δεδομένων: movielens 100K** 

Εικόνα 66 Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – movielens 100K

## Αποτελέσματα αναζητήσεων

### Κατηγορία: Μάλλινα



### Κατηγορία: Παλτό



### Κατηγορία: Φόρεμα



### Κατηγορία: Σορτσάκια



### Κατηγορία: Πουκάμισα





### Κατηγορία: Μαγιο



# Βιβλιογραφία

|  |  |
| --- | --- |
| [1] | Y. Koren, Factorization Meets the Neighborhood: a Multifaceted Collabora-tive Filtering Model, 2008. |
| [2] | Υ. Koren, R. Bell και C. Volinsky,, Matrix factorization Techniques forRecommender Systems, 2009. |
| [3] | A. Tulloch, «Facebook Research,» [Ηλεκτρονικό]. Available: https://research.fb.com/blog/2014/09/fast-randomized-svd/. |
| [4] | E. Jane και T. Bailey , Folding-up: A Hybrid Method for Updating the Partial Singular Value Decomposition in Latent Semantic Indexing, 2006. |
| [5] |  |
| [6] | S. Rendle, . C. Freudenthaler, Z. Gantner και L. Schmidt-Thieme, Bayesian personalized ranking from implicit feedback, 2009, p. 452–461. |
| [7] | R. Socher, . D. Chen, C. D. Manning και . A. N. Reasoning, Reasoning with neural tensor networks for knowledge base completion, 2013, p. 926–934. |
| [8] | N. Srivastava και R. R. Salakhutdinov, Multimodal learning with deep boltzmann machines, 2012, p. 2222–2230. |
| [9] | B. Xu, R. Huang και M. Li, Revise Saturated Activation Functions, 2015. |
| [10] | A. M. Elkahky, Y. Song και . X. He, A multi-view deep learning approach for cross domain user modeling in recommendation systems, 2015, p. 278–288. |
| [11] | X. Glorot, A. Bordes και . Y. Bengio, Deep sparse rectifier neural networks., 2011. |
| [12] | K. He, X. Zhang, . S. Ren και J. Sun, Deep residual learning for image recognition, 2016. |
| [13] | «WikiPedia,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Random\_forest#/media/File:Random\_forest\_diagram\_complete.png. |
| [14] | R. Burke, Hybrid recommender systems: Survey and experiments. User Modeling and User-adapted Interaction, 2002, p. 331–370. |
| [15] | M. Jahrer, A. Toscher και R. Legenstein, Combining predictions for accurate recommender, ACM KDD Conference, 2010., p. 693–702. |
| [16] | T. Chai και . R. Draxler, «Root mean square error (RMSE) or mean absolute error (MAE)?,» p. 2, 2014. |
| [17] | J. Herlocker , . J. Konstan, L. Terveen και J. Riedl , Evaluating collaborative filtering recommender systems., ACM Transactions on Information Systems, 2004, pp. 5-53. |
| [18] | D. Jannach, M. Ge και C. Delgado-Battenfeld, Beyond accuracy: Evaluating recommender systems by coverage and serendipity, Barcelona, 2010. |
| [19] | K. Bradley και B. Smyth, Improving recommendation diversity, Maynooth: AICS-01, 2001, pp. 75-84. |
| [20] | M. Kunavera και T. Požrlb, Diversity in recommender systems – A survey, Ljubljana, 2017. |
| [21] | L. Zhang, The Definitionof Novelty in Recommendation System, Guiyang: School of management Xiamen university Xiamen, 2013. |
| [22] | L. Lü, M. Medo, H. Yeung και . Z. Yi-Cheng, Recommender systems, Physics Reports, 2012. |
| [23] | G. Karypis και C. Desrosiers, A comprehensive survey of neighborhood-basedrecommendation methods, Springer, 2011, pp. 21-22. |
| [24] | X. Su και T. M. Khoshgoftaar, A Survey of Collaborative Filtering Techniques, USA: Hindawi Publishing Corporation, 2009, pp. 6-7. |
| [25] | B. Sarwar, G. Karypis, J. Konstan και J. Riedl, Item-based Collaborative Filtering RecommendationAlgorithms, Hong Kong, 2001, pp. 6-7. |
| [26] | C. C. Aggarwal, Recommender Systems - The Textbook, Springer, 2006. |
| [27] | H. Xiangnan , L. Liao και H. Zhang, Neural Collaborative Filtering, 2017. |