

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΕΡΓΑΣΤΗΡΙΟ ΕΠΕΞΕΡΓΑΣΙΑΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΥΠΟΛΟΓΙΣΜΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**Δημιουργία γενικευμένου συστήματος προτάσεων με εφαρμογή σε σύνολα αλληλεξαρτώμενων δεδομένων**

Εκπόνηση: Επιβλέποντες:

**Ιωάννης-Παναγιώτης Μπουντουρίδης**  Καθ. **Ανδρέας Συμεωνίδης**

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

ΑΕΜ: 8872 Ερευν. **Σωτήρης Τσαρούχης**

ΘΕΣΣΑΛΟΝΙΚΗ 2021

# **Ευχαριστίες**

Καταρχάς θα ήθελα να ευχαριστήσω τον καθηγητή κ. Ανδρέα Συμεωνίδη 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**

**Διπλωματική Εργασία**

**Τίτλος: Δημιουργία γενικευμένου συστήματος προτάσεων με εφαρμογή σε σύνολα αλληλεξαρτώμενων δεδομένων**

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 αυτών των συστημάτων εξυπηρετεί την ερμηνεία της δραστηριότητας των χρηστών ώστε να προτείνει προϊόντα ή υπηρεσίες που είναι όσο το δυνατόν πιο κοντά στα ενδιαφέροντα τους. H βιομηχανία της ένδυσης, καθώς και η βιομηχανία του κινηματογράφου αποτελούν αναμφίβολα πεδία που βασίζουν την επιτυχία τους στα παραπάνω συστήματα.

Σκοπός της παρούσας διπλωματικής είναι η δημιουργία ενός γενικευμένου συστήματος προτάσεων που θα εφαρμόζει τεχνικές προεπεξεργασίας δεδομένων και πρόβλεψης προτιμήσεων. Στα πλαίσια αυτής, η βιομηχανία της ένδυσης και η βιομηχανία του συνδρομητικού κινηματογράφου χρησιμοποιήθηκαν ως πεδία δοκιμής. Στην περίπτωση της βιομηχανίας της ένδυσης, ο τελικός στόχος είναι η δημιουργία ενός συστήματος που θα υποβοηθά σχεδιαστές παρέχοντας προτάσεις για τον σχεδιασμό ρούχων. Από την άλλη πλευρά, στη βιομηχανία του κινηματογράφου ο 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 με βάση χρήσιμες μετρικές με σκοπό την ερμηνεία των προτάσεων του συστήματος, δηλαδή του πόσο κοντά είναι οι προτάσεις των συστημάτων στις προτιμήσεις των χρηστών. Τα αποτελέσματα των προτάσεων που παρουσιάζονται τελικά, παράγονται από το σύστημα που είχε την καλύτερη επίδοση στην δημιουργία πετυχημένων προτάσεων.

# **Abstract**

**Diploma Thesis**

**Title: Creation of a generalized recommendation system with application to interdependent data sets**

Nowadays, big software and technology giants owe a significant part of their success to the time they have invested in building reliable recommendation systems. The role of these systems is to interpret users' activity and suggest products or services that are as close to their interests as possible. The clothing industry and the film streaming industry are undoubtedly fields that base their success on the systems mentioned above.

This diploma thesis aims to create a generalized recommendation system that will apply data preprocessing and recommendation techniques. In this context, the implemented system will be used in the clothing and the streaming cinema industry. For the clothing industry, the goal is to create a system that will assist designers by providing clothing design suggestions. On the other hand, the purpose of the streaming cinema industry is to create a system that will provide film suggestions to users as close to their preferences as possible.

For the clothing recommendation experiments, two datasets with clothes from the company ASOS and their user ratings are included: the first has a rating rate of 30% % of total ratings and the second 60%. For the streaming recommendation experiments, movieLens dataset is utilized; it is a set that consists of movies with ratings published by the GroupLens team.

The experiments performed on the above systems are evaluated through useful metrics to interpret the recommendations and how close they are to their preferences. The final recommendations presented are produced from the system with the best performance in creating successful proposals.

# **Πίνακας περιεχομένων**

[**Ευχαριστίες i**](#_Toc65342186)

[**Περίληψη ii**](#_Toc65342187)

[**Abstract iii**](#_Toc65342188)

[**Πίνακας περιεχομένων iv**](#_Toc65342189)

[**Λίστα εικόνων viii**](#_Toc65342190)

[**Λίστα πινάκων xii**](#_Toc65342191)

[**1 Εισαγωγή 1**](#_Toc65342192)

[1.1 Γενικά 1](#_Toc65342193)

[1.2 Ορισμός προβλήματος 2](#_Toc65342194)

[1.3 Στόχος της Διπλωματικής 2](#_Toc65342195)

[1.4 Μεθοδολογία διπλωματικής 3](#_Toc65342196)

[1.5 Οργάνωση Κεφαλαίων 4](#_Toc65342197)

[**2** **Θεωρητικό Υπόβαθρο** 5](#_Toc65342198)

[2.1 Κατηγορίες συστημάτων προτάσεων 5](#_Toc65342199)

[2.1.1 Συστήματα Προτάσεων Συνεργατικού Φιλτραρίσματος (Collaborative filtering) 5](#_Toc65342200)

[2.1.1.1 Συστήματα ανάλυσης πίνακα σε ιδιάζουσες τιμές (Singular Value Decomposition) [5] 8](#_Toc65342201)

[2.1.1.2 Νευρωνικά Συνεργατικά Συστήματα (Neural Collaborative Filtering) 9](#_Toc65342202)

[2.1.2 Συστήματα Προτάσεων Φιλτραρίσματος Περιεχομένου (Content-based Filtering) 12](#_Toc65342203)

[2.1.2.1 Τυχαία Δάση (Random Forest) 13](#_Toc65342204)

[2.1.2.2 Βαθιά Νευρωνικά Δίκτυα Περιεχομένου (Deep Content Neural Networks) 14](#_Toc65342205)

[2.1.3 Υβριδικά Συστήματα Προτάσεων (Hybrid) 15](#_Toc65342206)

[2.1.3.1 Σταθμισμένα Υβριδικά Συστήματα 17](#_Toc65342207)

[2.2 Βιβλιογραφική ανασκόπηση 18](#_Toc65342208)

[2.2.1 Τεχνικές Παραγοντοποίησης Πίνακα για Συστήματα προτάσεων (Matrix Factorization Techniques for Recommender Systems) 18](#_Toc65342209)

[2.2.2 Συνεργατικά Νευρωνικά Δίκτυα (Neural Collaborative Filtering 2017) 19](#_Toc65342210)

[2.2.3 Συγκεντρωτικά Συστήματα Προτάσεων με Τυχαία Δάση (Aggregated Recommendation through Random Forests 2014) 19](#_Toc65342211)

[2.3 Μετρικές Αξιολόγησης Συστημάτων Προτάσεων 20](#_Toc65342212)

[2.3.1 Ομοιότητα δεδομένων 20](#_Toc65342213)

[2.3.2 Ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 21](#_Toc65342214)

[2.3.3 Μέσο απόλυτο σφάλμα (MAE) 21](#_Toc65342215)

[2.3.4 Κάλυψη (Coverage) 22](#_Toc65342216)

[2.3.5 Ποικιλομορφία (diversity) 22](#_Toc65342217)

[2.3.6 Καινοτομία (novelty) 22](#_Toc65342218)

[**3** **Υλοποίηση συστημάτων προτάσεων** 24](#_Toc65342219)

[3.1 Προ-επεξεργασία Δεδομένων 24](#_Toc65342220)

[3.1.1 Σετ δεδομένων Asos - Βήματα 24](#_Toc65342221)

[3.1.1.1 Συλλογή και οργάνωση δεδομένων 24](#_Toc65342222)

[3.1.1.2 Χαρακτηριστικά Ρούχων (features) 26](#_Toc65342223)

[3.1.1.3 Βαθμολογίες Ρούχων 29](#_Toc65342224)

[3.1.2 Σετ δεδομένων Movielens-100k 29](#_Toc65342225)

[3.1.2.1 Χαρακτηριστικά ταινιών (features) 29](#_Toc65342226)

[3.1.2.2 Βαθμολογίες Ταινιών 30](#_Toc65342227)

[3.2 Συνεργατικό φιλτράρισμα με επίκεντρο τον χρήστη 30](#_Toc65342228)

[3.2.1 Δεδομένα εισόδου (input data) 30](#_Toc65342229)

[3.2.2 Βήματα υλοποίησης 31](#_Toc65342230)

[3.2.2.1 Εύρεση όμοιων χρηστών 31](#_Toc65342231)

[3.2.2.2 Στάθμιση σημασίας ομοιότητας (significance weighting) 31](#_Toc65342232)

[3.2.2.3 Κεντράρισμα μέση τιμής δεδομένων (mean-center data) 31](#_Toc65342233)

[3.2.2.4 Συνάρτηση πρόβλεψης αξιολόγησης 32](#_Toc65342234)

[3.3 Συνεργατικό φιλτράρισμα με ανάλυση πίνακα σε ιδιάζουσες τιμές 33](#_Toc65342235)

[3.3.1 Δεδομένα Εισόδου (input data) 33](#_Toc65342236)

[3.3.2 Βήματα υλοποίησης 33](#_Toc65342237)

[3.3.2.1 Επιλογή βιβλιοθήκης μοντέλου 33](#_Toc65342238)

[3.3.2.2 Επιλογή παραμέτρων 33](#_Toc65342239)

[3.4 Συνεργατικό φιλτράρισμα μέσω νευρωνικών δικτύων 35](#_Toc65342240)

[3.4.1 Δεδομένα εισόδου (input data) 35](#_Toc65342241)

[3.4.2 Βήματα υλοποίησης 35](#_Toc65342242)

[3.4.2.1 Κωδικοποίηση Χρηστών-Προϊόντων (Ρούχων/Ταινιών) 35](#_Toc65342243)

[3.4.2.2 Επίπεδο Ενσωμάτωσης 35](#_Toc65342244)

[3.4.2.3 Αρχιτεκτονική πολύ-επίπεδο νευρωνικού δικτύου 36](#_Toc65342245)

[3.5 Φιλτράρισμα Περιεχομένου με Τυχαία Δάση 38](#_Toc65342246)

[3.5.1 Δεδομένα εισόδου 38](#_Toc65342247)

[3.5.2 Βήματα υλοποίησης 38](#_Toc65342248)

[3.5.2.1 Επιλογή βιβλιοθήκης μοντέλου 38](#_Toc65342249)

[3.5.2.2 Επιλογή παραμέτρων 38](#_Toc65342250)

[3.6 Φιλτράρισμα Περιεχομένου με Νευρωνικά Δίκτυα 40](#_Toc65342251)

[3.6.1 Δεδομένα Εισόδου 40](#_Toc65342252)

[3.6.2 Βήματα υλοποίησης 40](#_Toc65342253)

[3.6.2.1 Επιλογή βιβλιοθήκης μοντέλου 40](#_Toc65342254)

[3.6.2.2 Αρχιτεκτονική Νευρωνικού Δικτύου 40](#_Toc65342255)

[**4** **Πειράματα και αποτελέσματα** 43](#_Toc65342256)

[4.1 Διαχωρισμός δεδομένων 43](#_Toc65342257)

[4.2 Σύνολα δεδομένων 44](#_Toc65342258)

[4.3 Μετρικές Αξιολόγησης 45](#_Toc65342259)

[4.3.1 Accuracy, Recall, Precision και F1 Score 45](#_Toc65342260)

[4.3.2 RMSE και MAE 46](#_Toc65342261)

[4.4 Υλοποίηση συνεργατικού φιλτραρίσματος με επίκεντρο τον χρήστη (User-Based) 46](#_Toc65342262)

[4.4.1 Accuracy, recall, precision και f1\_score 47](#_Toc65342263)

[4.4.2 Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 48](#_Toc65342264)

[4.4.3 Κάλυψη και Ποικιλία 49](#_Toc65342265)

[4.5 Υλοποίηση συνεργατικού φιλτραρίσματος με ανάλυση πίνακα ιδιαζουσών τιμών (SVD) 50](#_Toc65342266)

[4.5.1 Accuracy, recall, precision και f1\_score 50](#_Toc65342267)

[4.5.2 Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 52](#_Toc65342268)

[4.5.3 Κάλυψη και Ποικιλία 53](#_Toc65342269)

[4.6 Υλοποίηση Συνεργατικού Νευρωνικού Φιλτραρίσματος 55](#_Toc65342270)

[4.6.1 Accuracy, recall, precision και f1\_score 55](#_Toc65342271)

[4.6.2 Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 57](#_Toc65342272)

[4.6.3 Κάλυψη και Ποικιλία 58](#_Toc65342273)

[4.7 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Τυχαία Δάση 60](#_Toc65342274)

[4.7.1 Accuracy, recall, precision και f1\_score 60](#_Toc65342275)

[4.7.2 Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 61](#_Toc65342276)

[4.7.3 Κάλυψη και Ποικιλία 62](#_Toc65342277)

[4.8 Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα 64](#_Toc65342278)

[4.8.1 Accuracy, recall, precision και f1\_score 64](#_Toc65342279)

[4.8.2 Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 66](#_Toc65342280)

[4.8.3 Κάλυψη και Ποικιλία 67](#_Toc65342281)

[4.9 Υλοποίηση Υβριδικού Συστήματος 69](#_Toc65342282)

[4.9.1 Accuracy, recall, precision και f1\_score 69](#_Toc65342283)

[4.9.2 Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 71](#_Toc65342284)

[4.9.3 Κάλυψη και Ποικιλία 72](#_Toc65342285)

[4.10 Συνοπτική αναπαράσταση όλων των υλοποιήσεων 75](#_Toc65342286)

[4.10.1 Accuracy, recall, precision και f1\_score 75](#_Toc65342287)

[4.10.2 Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 76](#_Toc65342288)

[4.11 Αποτελέσματα αναζητήσεων 78](#_Toc65342289)

[4.11.1 Κατηγορία: Μάλλινα 79](#_Toc65342290)

[4.11.2 Κατηγορία: Παλτό 80](#_Toc65342291)

[4.11.3 Κατηγορία: Φόρεμα 81](#_Toc65342292)

[4.11.4 Κατηγορία: Σορτσάκια 82](#_Toc65342293)

[4.11.5 Κατηγορία: Γυναικεία Πουκάμισα 83](#_Toc65342294)

[4.11.6 Κατηγορία: Αντρικά Πουκάμισα 84](#_Toc65342295)

[4.11.7 Κατηγορία: Γυναικεία Μαγιό 85](#_Toc65342296)

[**5** **Συμπεράσματα και Μελλοντικές Επεκτάσεις** 86](#_Toc65342297)

[5.1 Γενικά Συμπεράσματα 86](#_Toc65342298)

[5.2 Μελλοντικές επεκτάσεις 87](#_Toc65342299)

[**Βιβλιογραφία** 88](#_Toc65342300)

# **Λίστα εικόνων**

[**Εικόνα 1** - Αναπαράσταση ανάλυσης πίνακα σε ιδιάζουσες τιμές (SVD) [6] 8](#_Toc65342301)

[**Εικόνα 2** - Αρχιτεκτονική Νευρωνικών Συνεργατικών Συστημάτων [9] 10](#_Toc65342302)

[**Εικόνα 3** - Οπτικοποίηση μοντέλου: Τυχαία Δάση [17] 13](#_Toc65342303)

[**Εικόνα 4 -** Αρχιτεκτονική Νευρωνικών Δικτύων 14](#_Toc65342304)

[**Εικόνα 5** - Ταξινομία Υβριδικών Συστημάτων [19] 16](#_Toc65342305)

[**Εικόνα 6** - Διάγραμμα ρούχων ανά κατηγορία αρχικού σετ δεδομένων 25](#_Toc65342306)

[**Εικόνα 7** - Αναπαράσταση περίπτωσης επικάλυψης ρούχων ανά κατηγορία 25](#_Toc65342307)

[**Εικόνα 8** - Διάγραμμα ρούχων ανά κατηγορία τελικού σετ δεδομένων 26](#_Toc65342308)

[**Εικόνα 9** - Δημιουργία Πίνακα Συγκεντρωτικών Χαρακτηριστικών 28](#_Toc65342309)

[**Εικόνα 10** - Οπτική αναπαράσταση βημάτων μοντέλου ανάλυσης πίνακα σε ιδιάζουσες τιμές 34](#_Toc65342310)

[**Εικόνα 11** - Οπτική αναπαράσταση βημάτων μοντέλου συνεργατικού φιλτραρίσματος νευρωνικών δικτύων 37](#_Toc65342311)

[**Εικόνα 12** - Οπτική αναπαράσταση βημάτων μοντέλου Τυχαίων Δασών 39](#_Toc65342312)

[**Εικόνα 13** - Οπτική αναπαράσταση βημάτων μοντέλου Νευρωνικών Περιεχομένου 42](#_Toc65342313)

[**Εικόνα 14** - Αναπαράσταση Διαχωρισμού Δεδομένων - 1 43](#_Toc65342314)

[**Εικόνα 15** - Αναπαράσταση Διαχωρισμού Δεδομένων - 2 44](#_Toc65342315)

[**Εικόνα 16** - Διαχωρισμός αξιολογήσεων [0,10] 45](#_Toc65342316)

[**Εικόνα 17** - Διαχωρισμός αξιολογήσεων [0,5] 46](file:///C:\Users\Mpountouridis\Desktop\giannis\BML-Recommendation-Engine\διπλωματική_διορθώσεις.docx#_Toc65342317)

[**Εικόνα 18** - Υλοποίηση User-Based - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score ανά γείτονα – data30 47](#_Toc65342318)

[**Εικόνα 19** - Υλοποίηση User-Based - Διάγραμμα μετρικών αccuracy, recall, precision και f1\_score ανά γείτονα – data60 47](#_Toc65342319)

[**Εικόνα 20** - Υλοποίηση User-Based - Διάγραμμα μετρικών rmse και mae ανά γείτονα – data30 48](#_Toc65342320)

[**Εικόνα 21** - Υλοποίηση User-Based - Διάγραμμα μετρικών rmse και mae ανά γείτονα - data60 48](#_Toc65342321)

[**Εικόνα 22 -** Υλοποίηση User-Based - Διάγραμμα Κάλυψης και Ποικιλίας – data30 49](#_Toc65342322)

[**Εικόνα 23** - Υλοποίηση User-Based - Διάγραμμα Κάλυψης και Ποικιλίας – data60 50](#_Toc65342323)

[**Εικόνα 24** - Υλοποίηση SVD - Διάγραμμα μετρικών αccuracy, recall, precision και f1\_score ανά τιμή χαρακτηριστικών – data30 51](#_Toc65342324)

[**Εικόνα 25** - Υλοποίηση SVD - Διάγραμμα μετρικών αccuracy, recall, precision και f1\_score ανά τιμή χαρακτηριστικών – data60 51](#_Toc65342325)

[**Εικόνα 26** - Υλοποίηση SVD - Διάγραμμα μετρικών αccuracy, recall, precision και f1\_score ανά τιμή χαρακτηριστικών – movielens 100K 51](#_Toc65342326)

[**Εικόνα 27** - Υλοποίηση SVD - Διάγραμμα μετρικών rmse και mae ανά τιμή χαρακτηριστικών – data30 52](#_Toc65342327)

[**Εικόνα 28** - Υλοποίηση SVD - Διάγραμμα μετρικών rmse και mae ανά τιμή χαρακτηριστικών – data30 52](#_Toc65342328)

[**Εικόνα 29** - Υλοποίηση SVD - Διάγραμμα μετρικών rmse και mae ανά τιμή χαρακτηριστικών – movielens 100K 53](#_Toc65342329)

[**Εικόνα 30** - Υλοποίηση SVD - Διάγραμμα Κάλυψης και Ποικιλίας – data30 54](#_Toc65342330)

[**Εικόνα 31** - Υλοποίηση SVD - Διάγραμμα Κάλυψης και Ποικιλίας – data60 54](#_Toc65342331)

[**Εικόνα 32** - Υλοποίηση SVD - Διάγραμμα Κάλυψης και Ποικιλίας – movielens-100k 55](#_Toc65342332)

[**Εικόνα 33 -** Υλοποίηση NCF - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 56](#_Toc65342333)

[**Εικόνα 34** - Υλοποίηση NCF - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data60 56](#_Toc65342334)

[**Εικόνα 35** - Υλοποίηση NCF - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K 56](#_Toc65342335)

[**Εικόνα 36** - Υλοποίηση NCF - Διάγραμμα μετρικών rmse και mae – data30 57](#_Toc65342336)

[**Εικόνα 37** - Υλοποίηση NCF - Διάγραμμα μετρικών rmse και mae – data60 57](#_Toc65342337)

[**Εικόνα 38** - Υλοποίηση NCF - Διάγραμμα μετρικών rmse και mae – movielens 100K 57](#_Toc65342338)

[**Εικόνα 39** - Υλοποίηση NCF - Διάγραμμα Κάλυψης και Ποικιλίας – data30 58](#_Toc65342339)

[**Εικόνα 40** - Υλοποίηση NCF - Διάγραμμα Κάλυψης και Ποικιλίας – data60 59](#_Toc65342340)

[**Εικόνα 41** - Υλοποίηση NCF - Διάγραμμα Κάλυψης και Ποικιλίας – movielens-100k 59](#_Toc65342341)

[**Εικόνα 42** - Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 60](#_Toc65342342)

[**Εικόνα 43** - Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data60 60](#_Toc65342343)

[**Εικόνα 44 -** Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K 61](#_Toc65342344)

[**Εικόνα 45** - Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών rmse και mae – data30 61](#_Toc65342345)

[**Εικόνα 46** - Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών rmse και mae – data60 61](#_Toc65342346)

[**Εικόνα 47** - Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών rmse και mae – movielens 100K 62](#_Toc65342347)

[**Εικόνα 48** - Υλοποίηση Τυχαίων Δασών - Διάγραμμα Κάλυψης και Ποικιλίας – data30 63](#_Toc65342348)

[**Εικόνα 49** - Υλοποίηση Τυχαίων Δασών - Διάγραμμα Κάλυψης και Ποικιλίας – data60 63](#_Toc65342349)

[**Εικόνα 50** - Υλοποίηση Τυχαίων Δασών - Διάγραμμα Κάλυψης και Ποικιλίας – movielens-100k 64](#_Toc65342350)

[**Εικόνα 51** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 65](#_Toc65342351)

[**Εικόνα 52** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 65](#_Toc65342352)

[**Εικόνα 53** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K 65](#_Toc65342353)

[**Εικόνα 54** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών rmse και mae – data30 66](#_Toc65342354)

[**Εικόνα 55** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών rmse και mae – data60 66](#_Toc65342355)

[**Εικόνα 56** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών rmse και mae – movielens 100K 66](#_Toc65342356)

[**Εικόνα 57** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα Κάλυψης και Ποικιλίας – data30 67](#_Toc65342357)

[**Εικόνα 58** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα Κάλυψης και Ποικιλίας – data60 68](#_Toc65342358)

[**Εικόνα 59** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα Κάλυψης και Ποικιλίας – movielens-100k 68](#_Toc65342359)

[***Εικόνα* 60 *-*** Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 70](#_Toc65342360)

[**Εικόνα 61** - Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data60 70](#_Toc65342361)

[**Εικόνα 62** - Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K 70](#_Toc65342362)

[**Εικόνα 63** - Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – data30 71](#_Toc65342363)

[**Εικόνα 64** - Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – data30 72](#_Toc65342364)

[**Εικόνα 65** - Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – movielens 100K 72](#_Toc65342365)

[**Εικόνα 66** - Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα Κάλυψης και Ποικιλίας – data30 73](#_Toc65342366)

[**Εικόνα 67** - Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα Κάλυψης και Ποικιλίας – data60 74](#_Toc65342367)

[**Εικόνα 68** - Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα Κάλυψης και Ποικιλίας – movielens-100k 74](#_Toc65342368)

[**Εικόνα 69** - Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 76](#_Toc65342369)

[**Εικόνα 70** - Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data60 76](#_Toc65342370)

[**Εικόνα 71** - Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K 76](#_Toc65342371)

[**Εικόνα 72** - Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – data30 77](#_Toc65342372)

[**Εικόνα 73** - Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – data30 77](#_Toc65342373)

[**Εικόνα 74** - Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – movielens 100K 78](#_Toc65342374)

[**Εικόνα 75 -** Παρουσίαση αποτελεσμάτων ρούχων. Αναζήτηση: Γυναικεία Μάλλινα Μπλουζάκια 79](#_Toc65342375)

[**Εικόνα 76** - Παρουσίαση αποτελεσμάτων ρούχων. Αναζήτηση: Γυναικεία Παλτό 80](#_Toc65342376)

[**Εικόνα 77** - Παρουσίαση αποτελεσμάτων ρούχων. Αναζήτηση: Γυναικεία Φορέματα 81](#_Toc65342377)

[**Εικόνα 78** - Παρουσίαση αποτελεσμάτων ρούχων. Αναζήτηση: Αντρικά Σορτσ 82](#_Toc65342378)

[**Εικόνα 79** - Παρουσίαση αποτελεσμάτων ρούχων. Αναζήτηση: Γυναικεία Πουκάμισα 83](#_Toc65342379)

[**Εικόνα 80** - Παρουσίαση αποτελεσμάτων ρούχων. Αναζήτηση: Ανδρικά Πουκάμισα 84](#_Toc65342380)

[**Εικόνα 81 -** Παρουσίαση αποτελεσμάτων ρούχων. Αναζήτηση: Γυναικεία Μαγιό 85](#_Toc65342381)

# **Λίστα πινάκων**

[**Πίνακας 1** - Αναπαράσταση δεδομένων και εύρεση ομοιότητας 20](#_Toc65342382)

[**Πίνακας 2** – Αναπαράσταση διαφορετικών μεγεθών κάθε ρούχου 27](#_Toc65342383)

[**Πίνακας 3** – Αναπαράσταση διαφορετικής φόρμας κάθε ρούχου 27](#_Toc65342384)

[**Πίνακας 4** – Αναπαράσταση σχεδιασμού γιακά κάθε ρούχου 27](#_Toc65342385)

[**Πίνακας 5** – Αναπαράσταση διαφορετικής κατηγορίας κάθε ρούχου 27](#_Toc65342386)

[**Πίνακας 6** – Αναπαράσταση λέξεις κλειδιών κάθε ρούχου 28](#_Toc65342387)

[**Πίνακας 7** - Αναπαράσταση καθαρισμού περιττών καταχωρήσεων 28](#_Toc65342388)

[**Πίνακας 8** Αναπαράσταση κάθε κατηγορίας των ταινιών 29](#_Toc65342389)

[**Πίνακας 9** Αναπαράσταση βαθμολογίας των ταινιών από κάθε χρήστη 30](#_Toc65342390)

[**Πίνακας 10** - Αναπαράσταση του πίνακα χρηστών-αξιολογήσεων 30](#_Toc65342391)

[**Πίνακας 11** - Αναπαράσταση του πίνακα χρηστών-αξιολογήσεων - Υπολογισμός ομοιότητας 31](#_Toc65342392)

[**Πίνακας 12** - Αναπαράσταση πίνακα με κεντραρισμένη μέση τιμή αξιολογήσεων 32](#_Toc65342393)

[**Πίνακας 13** - Αναπαράσταση του πίνακα χρηστών-αξιολογήσεων 33](#_Toc65342394)

[**Πίνακας 14** – Αναπαράσταση δομής πολύ-επίπεδου νευρωνικού 36](#_Toc65342395)

[**Πίνακας 15** – Αναπαράσταση Δομής πολύ-επίπεδου νευρωνικού περιεχομένου 41](#_Toc65342396)

[**Πίνακας 16** - Συνοπτική αναπαράσταση για κάθε σύνολο δεδομένων 45](#_Toc65342397)

**Κεφάλαιο**

# **Εισαγωγή**

## **Γενικά**

Στην σημερινή εποχή ένας ευρέως αναπτυσσόμενος κλάδος είναι αυτός της ένδυσης και της μόδας, ο οποίος αποτελεί μία κερδοφόρα και ταχέως 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 κοινωνικών, επαγγελματικών και προσωπικών αναγκών.

Σε αυτό (υπέρ-)καταναλωτικό κλίμα, μεγάλοι οίκοι μόδας καλούνται να βρουν τρόπους για να ικανοποιήσουν όσο το δυνατόν περισσότερους πελάτες, ενώ ταυτόχρονα να μεγιστοποιήσουν το κέρδος που έχουν τα προϊόντα τους. Προς αυτή την κατεύθυνση, προσπαθούν να δημιουργούν κολεξιόν που θα ακολουθούν τις συνεχώς εξελισσόμενες τάσεις της μόδας και θα οδηγήσουν τους καταναλωτές να τις προτιμήσουν. Για αυτό το λόγο, οι μεγάλοι οίκοι μόδας πρέπει να δημιουργούν σε κάθε νέα καμπάνια τους τις προϋποθέσεις ώστε να κερδίζουν τις εντυπώσεις, να μειώνουν την αβεβαιότητα των πωλήσεων τους και να ακολουθούν πιστά τις τάσεις της μόδας ώστε να οδηγηθούν τελικά στην επίτευξη του στόχου τους.

Κάτι αντίστοιχο φαίνεται συμβαίνει και με τις μεγάλες εταιρίες συνδρομητικού κινηματογράφου (Netflix, Amazon Prime), οι οποίες παρέχουν στους χρήστες τους χιλιάδες ταινίες για να περάσουν ευχάριστα τον χρόνο τους. Με το πέρασμα του χρόνου, όλο και περισσότερες ταινίες προστίθενται στην τεράστια λίστα των υπάρχοντών ταινιών, με αποτέλεσμα η αναζήτηση μιας ταινίας με βάση τις προτιμήσεις του χρήστη να γίνεται διαδικασία επίπονη και χρονοβόρα. Αυτό το γεγονός εύλογα προβληματίζει τις εταιρίες διανομής, και δημιουργεί ένα κλίμα αβεβαιότητας που χρειάζεται να ξεπεραστεί.

## **Ορισμός προβλήματος**

Όπως αναφέρθηκε, η βιομηχανία της ένδυσης και η βιομηχανία του συνδρομητικού κινηματογράφου αποτελούν ευρέως αναπτυσσόμενους κλάδους. Μία εταιρεία μόδας για να κατανοήσει τις ανάγκες των πελατών της και να δημιουργήσει νέες τάσεις, χρειάζεται να προσανατολισθεί σε νέα τεχνολογικά εργαλεία. Το ίδιο ισχύει και για τις εταιρείες συνδρομητικού κινηματογράφου, οι οποίες είναι απαραίτητο να προτείνουν στους χρήστες ταινίες της προτίμησης τους. Με αυτόν τον τρόπο διατηρούν το ενδιαφέρον των χρηστών και να τους απαλλάσσουν από τον κόπο αναζήτησης σε μία τεράστια βάση δεδομένων.

Συνεπώς, τα δεδομένα χιλιάδων ρούχων που χρειάζεται να διαχειριστεί μια εταιρία ένδυσης ή τα δεδομένα χιλιάδων ταινιών μιας εταιρίας συνδρομητικού κινηματογράφου φέρνουν στην επιφάνεια την ανάγκη ενός σύγχρονου συστήματος προτάσεων. Στόχος είναι η διαχείριση αυτού του τεράστιου όγκου δεδομένων με σκοπό να προτείνονται προϊόντα ή υπηρεσίες κοντά στις προτιμήσεις των εταιρειών ή/και των χρηστών τους.

Μ’ αυτά τα συστήματα, οι οίκοι ένδυσης θα μπορούν να αναπροσαρμόζονται συνεχώς στις μεταβαλλόμενες τάσεις της μόδας. Αντίστοιχα, οι εταιρείες συνδρομητικού κινηματογράφου θα μπορούν να προσαρμόζουν τις προτάσεις τους σε επίκαιρες σειρές και ταινίες, κοντά στις προτιμήσεις των χρηστών. Η αξιολόγηση κάθε πρότασης είναι ύψιστης σημασίας, καθώς αποτελεί μέρος της διαδικασίας βελτίωσης του συστήματος. Για το λόγο αυτό, ο μεγάλος όγκος των προτιμήσεων των χρηστών για τις δύο περιπτώσεις που αναφέρθηκαν, είναι αναγκαίος για να δημιουργηθεί το σωστό πλαίσιο λειτουργίας του συστήματος στο μεταβαλλόμενο περιβάλλον της αγοράς.

## **Στόχος της 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 νέων προϊόντων. Πιο συγκεκριμένα, οι σχεδιαστές θα έχουν στην βάση δεδομένων τους μια λίστα από προϊόντα ένδυσης, ένα μικρό ποσοστό των οποίων θα έχουν αξιολογηθεί στο παρελθόν. Αντίστοιχα, θα υπάρχουν αξιολογήσεις εθελοντών ή πελατών που θα έχουν εκφράσει την αρέσκεια ή την δυσαρέσκειά τους για κάποια από τα προϊόντα της βάσης δεδομένων που έχει πρόσβαση ο σχεδιαστής. Αυτά τα δεδομένα θα αξιοποιούνται, ώστε η καθοδήγηση που θα προβάλλεται στο σχεδιαστή να είναι προσαρμοσμένη στην δική του αισθητική, αλλά και στις προτιμήσεις των πελατών-εθελοντών.

Για την βιομηχανία του κινηματογράφου, τελικός στόχος είναι η δημιουργία ενός συστήματος που θα παρέχει προτάσεις ταινιών στους χρήστες με βάση τις προτιμήσεις τους. Πιο συγκεκριμένα, το σύστημα θα προσπαθήσει να χαρτογραφήσει τις προτιμήσεις κάθε χρήστη και στη συνέχεια θα αναζητήσει μέσα σε μία τεράστια βάση δεδομένων από ταινίες και σειρές, σύμφωνα με αυτές τις προτιμήσεις. Με αυτό τον τρόπο μία συνδρομητική εταιρία κινηματογράφου μπορεί να διατηρήσει το ενδιαφέρον των συνδρομητών της και να ελαχιστοποιήσει τις απώλειες απογραφών από το σύστημα της.

## **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 αυτών των 2 είναι ότι το πρώτο έχει το 30% του συστήματος σε αξιολογήσεις από το σύνολο των ρούχων που μπορούν να γίνουν ενώ το δεύτερο 60%. Τα δύο αυτά σύνολα έχουν 100 χρήστες και 1700 ρούχα που ανήκουν σε 10 διαφορετικές κατηγορίες. To τρίτο σύνολο περιλαμβάνει 100 χιλιάδες αξιολογήσεις σε 9000 ταινίες, από 600 χρήστες.

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

1. Συγκέντρωση των δεδομένων,
2. Προεπεξεργασία των δεδομένων των προϊόντων ένδυσης,
3. Καθορισμός της/των κατηγορίας/ιών κάθε προϊόντος,
4. Διαχωρισμός των δεδομένων σε δεδομένα εκπαίδευσης και ελέγχου,
5. Δημιουργία των συστημάτων προτάσεων,
6. Συνεργασία των συστημάτων προτάσεων που δημιουργήθηκαν μέσω ενός υβριδικού συστήματος, και
7. Αξιολόγηση και επιστημονική τεκμηρίωση της ορθής λειτουργίας του συστήματος με χρήση μετρικών.

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

1. Συγκέντρωση των δεδομένων,
2. Προεπεξεργασία των δεδομένων των ταινιών,
3. Καθορισμός των κατηγοριών κάθε ταινίας,
4. Διαχωρισμός των δεδομένων σε δεδομένα εκπαίδευσης και ελέγχου,
5. Δημιουργία των συστημάτων προτάσεων,
6. Συνεργασία των συστημάτων προτάσεων που δημιουργήθηκαν μέσω ενός υβριδικού συστήματος, και
7. Αξιολόγηση και επιστημονική τεκμηρίωση της ορθής λειτουργίας του συστήματος με χρήση μετρικών.

## **Οργάνωση Κεφαλαίων**

Η παρούσα διπλωματικής αποτελείται από 5 κεφάλαια και οργανώνεται ως εξής:

* **Κεφάλαιο 2:** Περιγράφεται το βασικό θεωρητικό υπόβαθρο, στο οποίο στηρίχθηκε η διπλωματική, και η αρχιτεκτονική των εργαλείων που χρησιμοποιήθηκαν για την υλοποίηση του συστήματος.
* **Κεφάλαιο 3:** Γίνεται διεξοδική ανάλυση των βημάτων κάθε προτεινόμενης υλοποίησης, καθώς και αναφορά των εργαλείων που χρησιμοποιήθηκαν.
* **Κεφάλαιο 4**: Γίνεται η παρουσίαση των πειραμάτων, η αξιολόγηση της απόδοσης που επιτυγχάνει το σύστημα στην εξαγωγή προτάσεων και η οπτική αναπαράσταση των αποτελεσμάτων.
* **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 ορισμένες κατηγορίες συστημάτων που, όπως θα δούμε στα επόμενα κεφάλαια, θα αποτελέσουν τα θεμέλια για την δημιουργία του προτεινόμενου συστήματος.

### **Συστήματα Προτάσεων Συνεργατικού Φιλτραρίσματος (Collaborative filtering)**

Τα συστήματα προτάσεων που ανήκουν στην κατηγορία του συνεργατικού φιλτραρίσματος χρησιμοποιούν αποκλειστικά τις βαθμολογίες χρηστών. 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 το σύνολο των βαθμολογιών των χρηστών του συστήματος και παράγουν προβλέψεις σχετικά με το τι χρειάζεται ένας χρήστης, έχοντας ως βάση τους πιο κοντινούς (ως προς τις προτιμήσεις) σε αυτόν χρήστες [1].

Η κυρία πρόκληση στην σχεδίαση ενός μοντέλου συνεργατικό φιλτραρίσματος είναι η σωστή διαχείριση των *πινάκων αξιολογήσεων* (*user-item rating matrix*) που περιλαμβάνουν τις αξιολογήσεις κάθε χρήστη για όλα τα προϊόντα της βάσης *(Πίνακας 13).* Οι πίνακες αυτοί είναι *αραιοί* (*sparse*), καθώς πολλές θέσεις αξιολογήσεων παραμένουν κενές. Πιο συγκεκριμένα, σε μία βάση δεδομένων με χιλιάδες προϊόντα, οι περισσότεροι χρήστες αξιολογούν ένα μικρό κομμάτι του συνόλου όλων των προϊόντων, με αποτέλεσμα οι περισσότερες αξιολογήσεις να παραμένουν απροσδιόριστες.

Για κάθε χρήστη η βασική ιδέα του συνεργατικού φιλτραρίσματος είναι να καλύψει αυτές στις απροσδιόριστες αξιολογήσεις λαμβάνοντας υπόψη τους υψηλά συσχετιζόμενους χρήστες που έχουν παρόμοιες προτιμήσεις.

Υπάρχουν δύο είδη μεθόδων που χρησιμοποιούνται κατά κόρον στις υλοποιήσεις συνεργατικού φιλτραρίσματος: οι (α) *μέθοδοι βασισμένοι στη μνήμη (memory-based methods),* και οι (β) *μέθοδοι βασισμένοι στα μοντέλα (model-based methods)* [2].

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

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 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 τον 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 κυμαίνεται 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.

1. *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. Πιο συγκεκριμένα, προσαρμόζουν ένα παραμετροποιήσιμο σύστημα στον πίνακα αξιολογήσεων, που χρησιμοποιείται για την πρόβλεψη απροσδιόριστων βαθμολογιών, και κατ’ επέκταση για τη δημιουργία προτάσεων προς τους χρήστες.

Σε αυτή την υποκατηγορία συστημάτων ανήκουν:

* 1. οι *ταξινομητές κοντινότερου γείτονα (knn classifiers).*
  2. οι *Bayesian ταξινομητές (Bayesian classifiers).*
  3. 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: (1) το *Singular Value Decomposition (SVD),* (2) *το regularized SVD,* (3) *το Non-negative Matrix Factorization (NMF),* (4) *το Probabilistic Matrix Factorization (PMF),* (5) *το Bayesian PMF,* (6) *το Non-linear PMF,* (7) *το Maximum Margin Matrix Factorization (MMMF),* και (8) *Nonlinear Principal Component Analysis (NPCA)*.

Τα μοντέλα που αναφέρθηκαν τελευταία αποκαλούνται *λανθάνοντα παράγοντα (Latent factor models)* [3]. 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. [4]

#### **Συστήματα ανάλυσης πίνακα σε ιδιάζουσες τιμές (Singular Value Decomposition)** [5]

Τα συστήματα *ανάλυσης πίνακα σε ιδιάζουσες τιμές (SVD)* συνιστούν μια τεχνική παραγοντοποίησης πινάκων με βάση την οποία ένας *πίνακας Α* διαστάσεων παραγοντοποιείται σε 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:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

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 φθίνουσα σειρά. Με την βοήθεια των πινάκων και , οι χρήστες και τα προϊόντα (ρούχα/ταινίες) 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. Αντίστοιχα, είναι εφικτή η διατήρηση των πρώτων στηλών των πινάκων και και η αποβολή των υπόλοιπων στηλών (Εικόνα 1). 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) [6]

Πρώτον, είναι σημαντική η ανάλυση των λανθανουσών σχέσεων που υπάρχουν ανάμεσα στους χρήστες και τα προϊόντα (ρούχα/ταινίες). Κάθε ιδιάζουσα τιμή αναφέρεται σε έναν άγνωστο παράγοντα, που χαρακτηρίζει με σημαντικό τρόπο 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 τη *μείωση των 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 των συστάσεων, αλλά και στη απόδοσή του συστήματος ως προς τις απαιτήσεις σε αποθηκευτικό χώρο.

Ένα μειονεκτήματα της προσέγγισης αφορά τον χρόνο εκτέλεσής της, ο οποίος είναι συνήθως μεγαλύτερος από τη προσέγγιση συνεργατικού φιλτραρίσματος με βάση τη μνήμη. 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)* ώστε να άρουν αυτό τον περιορισμό [7]. Τέλος η προσέγγιση απαιτεί συνήθως απαλοιφή τον μηδενικών τιμών του αρχικού πίνακα μέσω εξειδικευμένης προεπεξεργασίας.

#### **Νευρωνικά Συνεργατικά Συστήματα (Neural Collaborative Filtering)**

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 [8].

Το συνεργατικό φιλτράρισμα μέσω νευρωνικών δικτύων άρει αυτό τον περιορισμό.



**Εικόνα 2** - Αρχιτεκτονική Νευρωνικών Συνεργατικών Συστημάτων [9]

Η αρχιτεκτονική των Νευρωνικών Συνεργατικών Συστημάτων όπως φαίνεται στην *(Εικόνα 2)* 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 εξαρτημένης από το συνεργατικό φιλτράρισμα, 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 και βοηθά στην αντιμετώπιση προβλημάτων εξαγωγής προτάσεων σε συστήματα με λίγα διαθέσιμα δεδομένα αξιολογήσεων από τους χρήστες.

Πάνω από το επίπεδο εισόδου βρίσκεται το επίπεδο ενσωμάτωσης, ένα πλήρως διασυνδεδεμένο επίπεδο, στο οποίο προβάλλεται η αραιή αναπαράσταση των χρηστών/αντικειμένων σε διανύσματα χαμηλότερων διαστάσεων. Το ενσωματωμένο επίπεδο για τον χρήστη, όπως και το ενσωματωμένο επίπεδο για το προϊόν αντίστοιχα, περιλαμβάνει τα ίδια λανθάνοντα διανύσματα με αυτά των *συστημάτων ανάλυσης ιδιάζουσας τιμής (SVD).*

Τα ενσωματωμένα επίπεδα συνενώνονται και τροφοδοτούν ένα πολύ-επίπεδο νευρωνικό δίκτυο, τα οποία χαρακτηρίζονται και ως 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 ή/και 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)* [10] ή *βάση περιθωρίου απώλειας (margin-based loss)* [11]. 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 πολύ-επίπεδο νευρωνικό σύστημα. Αυτός ο σχεδιασμός έχει υιοθετηθεί ευρέως σε πολυτροπικές εργασίες *βαθιάς μηχανικής μάθησης (multimodal deep learning work)* [12].

Μια απλή συνένωση των χαρακτηριστικών δε λαμβάνει υπόψη καμία αλληλεπίδραση των λανθάνων χαρακτηριστικών μεταξύ χρηστών και προϊόντων 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. Αυτό δεν είναι εφικτό να πραγματοποιηθεί για μοντέλα ανάλυσης πίνακα ιδιαζουσών τιμών.

Πιο συγκεκριμένα ορίζεται το σύστημα:

|  |  |  |
| --- | --- | --- |
|  |  | (2) |
|  |  | (3) |

|  |  |  |
| --- | --- | --- |
|  | *,* | (4) |
|  |  | (5) |

όπου ο πίνακας με βάρη, το bias του νευρωνικού δικτύου, και η συνάρτηση ενεργοποίησης για το επίπεδο. Αναλύοντας κάθε διαθέσιμη συνάρτηση ενεργοποίησης παρατηρούμε ότι:

1. *Η συνάρτηση σιγμοειδούς περιορίζει κάθε νευρώνα να παίρνει τιμές στο διάστημα (0,1), κάτι που μπορεί να έχει αρνητικές επιπτώσεις στην 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.* [13]
2. *Η συνάρτηση υπερβολικής εφαπτόμενης είναι καλύτερη επιλογή από αυτή της σιγμοειδούς και έχει υιοθετηθεί ευρέως* [14]*, αφού διορθώνει σε μεγάλο βαθμό το θέμα του κορεσμού.*
3. *Η συνάρτηση ReLU είναι μια εξίσου καλή λύση, αφού έχει αποδειχθεί μια εύλογη επιλογή που λύνει τα προβλήματα κορεσμού* [15]*, ενθαρρύνει αραιές ενεργοποιήσεις και είναι κατάλληλη για αραιά δεδομένα μειώνοντας τις πιθανότητες υπέρ-εκπαίδευσης.*

Τα αποτελέσματα των εργασιών πάνω στο αντικείμενο δείχνουν ότι η συνάρτηση ReLU έχει ελαφρώς καλύτερη απόδοση από την συνάρτηση υπερβολικής εφαπτόμενης, η οποία με την σειρά της είναι αποδεδειγμένα καλύτερη από τη σιγμοειδή.

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. Με την προϋπόθεση της χρήσης ενός μικρού αριθμού κρυφών μονάδων στα υψηλότερα επίπεδα, δημιουργούνται συνθήκες μάθησης πιο αφαιρετικών χαρακτηριστικών των δεδομένων [16]. Εμπειρικά, συνήθως 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 περιλαμβάνει αναλυτικά τα χαρακτηριστικά κάθε προϊόντος. Τα δεδομένα αυτά υφίστανται την απαραίτητη επεξεργασία, ώστε να εξαχθούν τα πιο σημαντικά χαρακτηριστικά και να αποβληθεί η περιττή πληροφορία. Για παράδειγμα, στην βιομηχανία της μόδας, 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 σε νέους χρήστες του συστήματος, εξαιτίας της έλλειψης δεδομένων που περιγράφουν προηγούμενες αγορές ή προτιμήσεις τους. Αυτό συμβαίνει γιατί τα συστήματα προτάσεων φιλτραρίσματος περιεχομένου βλέπουν τα χαρακτηριστικά κάθε προϊόντος που ο χρήστης δηλώνει ότι του αρέσει ή όχι και εξάγουν 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).*



**Εικόνα 3** - Οπτικοποίηση μοντέλου: Τυχαία Δάση [17]

Για την δημιουργία ενός δέντρου απόφασης (decision tree), αρχικά ανατίθενται στη ρίζα του τα δείγματα εκπαίδευσης. Κάθε κόμβος περιλαμβάνει ένα υποσύνολο των δειγμάτων, το οποίο μέσω της εφαρμογής ενός κατάλληλου ελέγχου διαχωρίζεται σε δυο ή περισσότερα 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 το *Gini index*, η εντροπία, το *misclassification error κ.α*.

Κάθε δένδρο του δάσους μπορεί να αναπτυχθεί ως το μέγιστο βάθος του ή μέχρι ορισμένο όριο βάθους που τίθεται πριν την διαδικασία εκπαίδευσης. Η μέθοδος Bagging, με ταξινομητές τα δέντρα απόφασης, αποτελεί μια ειδική κατηγορία των Random Forests. Σ’ αυτή την περίπτωση, η τυχαιότητα ενσωματώνεται στο μοντέλο 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. Στην συνέχεια, λαμβάνονται υπόψη οι ψήφοι των ταξινομητών, σύμφωνα με ένα πλειοψηφικό σύστημα ψηφοφορίας (majority voting scheme) [18]. 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. Αναθέτει τις λειτουργίες των νευρώνων σε ένα απλό στοιχείο ικανό να αθροίσει την είσοδο του και να κανονικοποιεί την έξοδο του. Οι νευρώνες είναι συνδεδεμένοι σε αυθαίρετα σύνθετα τεχνητά νευρωνικά δίκτυα. Για ταξινόμηση χρησιμοποιούνται συνήθως νευρωνικά δίκτυα εμπρόσθιας τροφοδότησης πολλών επιπέδων (multi-layer perceptions).



**Εικόνα 4** **-** Αρχιτεκτονική Νευρωνικών Δικτύων

Οι νευρώνες οργανώνονται σε επίπεδα *(Εικόνα 4)*:

* 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, 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, εφαρμόζονται οι τιμές των χαρακτηριστικών 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. Κάθε σύστημα προτάσεων φαίνεται να δρα περιοριστικά όταν δρα μεμονωμένα σε προβλήματα που υπάρχουν πολλές πηγές δεδομένων και το σύστημα αξιοποιεί μόνο ένα μέρος αυτών. 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 διερεύνηση 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 *(Εικόνα 5)*:

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 [19]

Τα υβριδικά συστήματα προτάσεων μπορούν να ταξινομηθούν στις ακόλουθες κατηγορίες [19]:

1. Σταθμισμένα συστήματα (weighted): Σε αυτήν την περίπτωση, οι βαθμολογίες πολλών συστημάτων ενσωματώνονται 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 του στο επόμενο σύστημα, 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 συστήματα, η έξοδος χρησιμοποιείται ως είσοδος σε διαδοχικό σύστημα. Ο συνηθισμένος συνδυασμός που χρησιμοποιείται είναι αυτός του συνεργατικού φιλτραρίσματος με το φιλτράρισμα περιεχομένου. Το σύστημα τροποποιείται ώστε οι τεχνικές συνεργατικού φιλτραρίσματος να περιέχουν χαρακτηριστικά περιεχομένου προϊόντων. Με αυτό τον τρόπο το σύστημα θεωρείται ότι δρα ως ένα ενιαίο μοντέλο, για 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, συνδυάζονται οι βαθμολογίες διαφορετικών συστημάτων προτάσεων 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 ως:

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

H προαναφερθείσα εξίσωση μπορεί να γραφεί σε όρους μεμονωμένων καταχωρίσεων του πίνακα:

|  |  |  |
| --- | --- | --- |
|  |  | (7) |

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

Υπάρχουν διαφορετικές μέθοδοι για τον υπολογισμό των βαρών. Στην πιο απλή περίπτωση, τα βάρη παίρνουν την τιμή της ίσης ψήφου, που στην συγκεκριμένη περίπτωση αυτή είναι . Ωστόσο, αυτή η περίπτωση είναι αρκετά ιδανική, καθώς τα συστήματα προτάσεων ποικίλουν.

Για την επιλογή των βέλτιστων βαρών είναι απαραίτητη η αξιολόγηση της επίδοσης κάθε συνδυασμού Με τη βοήθεια οποιαδήποτε αλγορίθμου παλινδρόμησης μπορεί να κρατηθεί ένα μικρό κλάσμα (25%) από γνωστές καταχωρήσεις του πίνακα βαθμολογίας διαστάσεων και να εκπαιδευτεί το μοντέλο δημιουργώντας τα βέλτιστα βάρη. Στη συνέχεια με βάση τα βάρη αυτά μπορεί εφαρμοστεί πρόβλεψη στο υπόλοιπο 75% των καταχωρήσεων. 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:

|  |  |  |
| --- | --- | --- |
|  |  | (8) |

|  |  |  |
| --- | --- | --- |
|  |  | (9) |

Οι παραπάνω μετρικές παρέχουν την αξιολόγηση ενός συγκεκριμένου συνδυασμού βαρών . Μια προσέγγιση που επικρατεί στην βιβλιογραφία και μπορεί να προσδιορίσει τις βέλτιστες τιμές του για την ελαχιστοποίηση αυτών των μετρικών, είναι η χρήση *γραμμικής παλινδρόμησης (linear regression).* [20]

## **Βιβλιογραφική ανασκόπηση**

### **Τεχνικές Παραγοντοποίησης Πίνακα για Συστήματα προτάσεων (Matrix Factorization Techniques for Recommender Systems)**

Η εργασία [21] διερευνά την χρήση τεχνικών παραγοντοποίησης ενός πίνακα τον οποίο μπορούν να αξιοποιήσουν τα συστήματα προτάσεων. Τα μοντέλα παραγοντοποίησης πίνακα αντιστοιχούν τόσο τους χρήστες όσο και τα προϊόντα σε έναν κοινό λανθάνοντα παράγοντα διαστάσεων k, έτσι ώστε οι αλληλεπιδράσεις χρηστών-προϊόντων να μοντελοποιούνται ως εσωτερικά προϊόντα σε αυτόν τον χώρο. Κατά συνέπεια, κάθε προϊόν i όπως και κάθε χρήστης j σχετίζονται με τα αντίστοιχα διανύσματα και . Τα στοιχεία αυτών των διανυσμάτων περιλαμβάνουν τιμές που δείχνουν τον τρόπο με τον οποίο ανταποκρίνονται οι χρήστες και τα προϊόντα (θετικά ή αρνητικά) σε αυτούς τους παράγοντες (latent factors). Το εσωτερικό γινόμενο αυτών των διανυσμάτων αντικατοπτρίζει την αλληλεπίδραση του χρήστη j στο προϊόν i. Κατά την διάρκεια της διαδικασίας παραγοντοποίησης πινάκων, η ανάλυση της συμπεριφοράς των χρηστών στις αλληλεπιδράσεις που κρύβονται πίσω απ’ τις αξιολογήσεις τους δίνουν την δυνατότητα ανίχνευσης έμμεσης πληροφορίας. Στην εργασία τονίζεται ο τρόπος με τον οποίον η σωστή ερμηνεία αυτών των αλληλεπιδράσεων προβλέπει και την σωστή βαθμολογία του χρήστη για ένα προϊόν, ώστε να βελτιώσει τις προτάσεις του συστήματος.

### **Συνεργατικά Νευρωνικά Δίκτυα (Neural Collaborative Filtering 2017)**

Στην εργασία [22] διερευνάται η χρήση των βαθιών νευρωνικών δικτύων στα συστήματα προτάσεων. Η προσέγγιση των νευρικών δικτύων έχει επίκεντρο το συνεργατικό φιλτράρισμα, μια τεχνική που προσπαθεί να ερμηνεύσει τις αλληλεπιδράσεις μεταξύ των χρηστών προς την δημιουργία επιτυχημένων προτάσεων. Πιο αναλυτικά η εργασία παρουσιάζει:

* Την αρχιτεκτονική ενός νευρωνικού δικτύου για τη μοντελοποίηση και την αναγνώριση των αλληλεπιδράσεων μεταξύ χρηστών και προϊόντων,
* Την μέθοδο ανάλυσης πίνακα ιδιαζουσών τιμών, ως ειδική περίπτωση του συνεργατικού νευρωνικού φιλτραρίσματος, και
* Την σπουδαιότητα των νευρωνικών δικτύων στην ανίχνευση των μη γραμμικών σχέσεων μεταξύ χρηστών και προϊόντων που μέθοδοι ανάλυσης ιδιαζουσών τιμών αδυνατούν να εντοπίσουν.

Η εργασία διερευνά διαφορετικές αρχιτεκτονικές νευρωνικών δικτύων για συνεργατικό φιλτράρισμα, χωρίς να αποκλείει όμως και άλλους συνδυασμούς μοντέλων ανίχνευσης αλληλεπιδράσεων χρήστη-προϊόντων.

### **Συγκεντρωτικά Συστήματα Προτάσεων με Τυχαία Δάση (Aggregated Recommendation through Random Forests 2014)**

Στην εργασία [23] προτείνονται τα Τυχαία δάση για τη δημιουργία προτάσεων, χρησιμοποιώντας τα δημογραφικά χαρακτηριστικά των χρηστών, τις πληροφορίες του προφίλ και τις βαθμολογίες που έχουν κάνει στα προϊόντα της αρεσκείας τους. Η προσέγγισή που ακολουθείται περιλαμβάνει τρία στάδια. Αρχικά, στο στάδιο της προεπεξεργασίας, οι πίνακες χρηστών, προσωπικών στοιχείων και βαθμολογίας του αρχικού συνόλου δεδομένων συγκεντρώνονται σε έναν ενιαίο πίνακα. Στη συνέχεια, κατασκευάζεται ένα σετ εκπαίδευσης και ένα σετ ελέγχου μέσω της πολλαπλής επικύρωσης (k-fold cross-validation). Ακολουθεί το στάδιο εκπαίδευσης και ελέγχου όπου οι δημογραφικές και προσωπικές πληροφορίες, καθώς και οι πληροφορίες περιεχομένου των χρηστών τροφοδοτούνται στο μοντέλο. Τέλος, υιοθετούνται τέσσερις προσεγγίσεις πρόβλεψης: (1) *τυπική ψηφοφορία (standard voting)*, (2) *σταθμισμένος μέσος όρος (weighted average),* (3) *ψηφοφορία βάσει συγκέντρωσης διανομής (aggregation based voting)* και (4) *μέσος όρος συγκέντρωσης διανομής (aggregation based average),* για τον υπολογισμό των τιμών αξιολόγησης του Τυχαίου Δάσους.

## **Μετρικές Αξιολόγησης Συστημάτων Προτάσεων**

### **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. Η ομοιότητα των συνημίτονων είναι ένα μέτρο ομοιότητας που μπορεί να χρησιμοποιηθεί για τη σύγκριση δύο διανυσμάτων. Το μέτρο υπολογίζει το συνημίτονο της γωνίας μεταξύ των διανυσμάτων (έστω **x** και **y**). Μια τιμή συνημίτονου 0 σημαίνει ότι τα δύο διανύσματα είναι 90 μοίρες το ένα στο άλλο. Όσο πλησιέστερη είναι η τιμή συνημίτονου στο 1, τόσο μικρότερη είναι η γωνία και τόσο μεγαλύτερη είναι η αντιστοιχία μεταξύ διανυσμάτων [24]. Στο παράδειγμα που ακολουθεί έχουμε ένα πίνακα αξιολογήσεων χρηστών v και u και αντίστοιχα τους τύπους για τον υπολογισμό της ομοιότητας των δύο χρηστών *(Πίνακας 1).*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *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

|  |  |  |
| --- | --- | --- |
|  |  | (10) |

|  |  |  |
| --- | --- | --- |
|  |  | (11) |

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

|  |  |  |
| --- | --- | --- |
|  |  | (12) |

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. [25]

Αυτό σημαίνει ότι για μοντέλα που τα μεγάλα σφάλματα είναι ανεπιθύμητα ή ακόμη και απαγορευτικά, η μετρική αυτή θα πρέπει να λαμβάνεται σοβαρά υπόψη. Πιο αναλυτικά, η μετρική δίνεται από τον τύπο:

|  |  |  |
| --- | --- | --- |
|  |  | (13) |

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:

|  |  |  |
| --- | --- | --- |
|  |  | (14) |

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 υπόψη τα μεγάλα και μικρά σφάλματα εκτίμησης με το ίδιο βάρος [25]. Ταυτόχρονα, δεν λαμβάνει υπόψη την κατεύθυνση του σφάλματος, αφού περιλαμβάνει την μέση τιμή στον υπολογισμό του τύπου.

### **Κάλυψη (Coverage)**

Σύμφωνα με την βιβλιογραφία [26], 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, και
* με το ποσοστό των διαθέσιμων προϊόντων που πράγματι συνιστώνται σε χρήστες

Στην παρούσα διπλωματική εργασία υιοθετείται ο πρώτος ορισμός, δηλαδή ως κάλυψη προβλέψεων από το σύνολο του καταλόγου. [27]

|  |  |  |
| --- | --- | --- |
|  |  | (15) |

### **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. [28] Πιο συγκεκριμένα θεωρούμε την ποικιλομορφία ως τον μέσο όρο της ανομοιότητας μεταξύ όλων των πιθανών ζευγών που μπορούν να προκύψουν από μια πρόταση. Ο τύπος περιγράφεται από την σχέση:

|  |  |  |
| --- | --- | --- |
|  |  | (16) |

Είναι σαφές ότι η ποιότητα της μέτρησης εξαρτάται σε μεγάλο βαθμό από την συνάρτηση ομοιότητας που θα εφαρμοστεί στο σύνολο των δεδομένων. [29]

### **Καινοτομία (novelty)**

Η έννοια της καινοτομίας στα συστήματα προτάσεων, 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 προϊόντων με τα οποία ο χρήστης δεν είχε καμία αλληλεπίδραση στο παρελθόν, δηλαδή του είναι άγνωστα [30]. Αναλυτικά ο τύπος περιγράφεται από την σχέση:

|  |  |  |
| --- | --- | --- |
|  |  | (17) |

όπου το 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 που θεωρεί τα λιγότερο δημοφιλή προϊόντα ως καινοτόμες προτάσεις στους χρήστες [31].

**Κεφάλαιο**

# **Υλοποίηση συστημάτων προτάσεων**

Το παρόν κεφάλαιο περιλαμβάνει:

* την αναλυτική περιγραφή της διαδικασίας προ-επεξεργασίας των δεδομένων,
* τις διαφορετικές μορφές εισόδου δεδομένων για κάθε υλοποίηση, και
* τη λειτουργία κάθε υλοποίησης, με ανάλυση κάθε βήματος.

## **Προ-επεξεργασία Δεδομένων**

### **Σετ δεδομένων Asos - Βήματα**

#### **Συλλογή και οργάνωση δεδομένων**

Το αρχικό σετ δεδομένων δημιουργήθηκε μέσω ενός συστήματος συλλογής δεδομένων από το site της εταιρίας μόδας Asos. Οι κατηγορίες του αρχικού σετ έχουν ανομοιογενή δεδομένα, δηλαδή δεν ακολουθούν ομοιόμορφη κατανομή. Αυτό είναι αναμενόμενο, καθώς κάθε κατηγορία προϊόντων έχει διαφορετικό αριθμό προϊόντων λόγω εποχής και ζήτησης. Ωστόσο, ο αριθμός των κατηγοριών φάνηκε να είναι αρκετά υψηλός (*Εικόνα 6*), με αποτέλεσμα πολλές κατηγορίες να είναι επικαλυπτόμενες.



**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*).



**Εικόνα 7** - Αναπαράσταση περίπτωσης επικάλυψης ρούχων ανά κατηγορία

Με βάση το αρχικό δείγμα δεδομένων, έγινε μια προσπάθεια καθαρισμού από παραπλανητικές καταχωρήσεις. Σε κάποιες κατηγορίες υπήρχαν προϊόντα που είχαν τοποθετηθεί λανθασμένα. Για παράδειγμα, υπήρχαν καταχωρημένα μαγιό (swimming suits) στα κουστούμια (suits). Περιπτώσεις σαν την παραπάνω φιλτραρίστηκαν, ώστε το τελικό σετ να περιλάμβανε κατά το μέγιστο δυνατόν ορθές καταχωρήσεις.

Οι 20 κατηγορίες που υπήρχαν αρχικά θεωρούσαν τα σετ σαν ξεχωριστές κατηγορίες, για παράδειγμα οι πιτζάμες και τα σετ πιτζάμες, τα παντελόνια και τα σετ παντελόνια θεωρούνταν ξεχωριστές κατηγορίες. Αυτοί οι διαχωρισμοί παραλείφθηκαν και ο αριθμός των κατηγοριών μειώθηκε σημαντικά.

Το τελικό, διαμορφωμένο σετ δεδομένων περιλαμβάνει δέκα κατηγορίες ρούχων, όσο κατά το δυνατόν ορθά καταχωρημένες, μη επικαλυπτόμενες, αλλά όχι απαραίτητα ομοιόμορφα κατανεμημένες (*Εικόνα 8*).



**Εικόνα 8** - Διάγραμμα ρούχων ανά κατηγορία τελικού σετ δεδομένων

#### **Χαρακτηριστικά Ρούχων (features)**

Κάθε προϊόν ένδυσης 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, κ.α.), η φόρμα του ρούχου (κολλητό, χαλαρό κ.α.) και το είδος γιακά του. Σκοπός του βήματος της προ-επεξεργασίας είναι η δημιουργία ενός διανύσματος χαρακτηριστικών που θα εκφράζει κάθε ρούχο με τα χαρακτηριστικά που το διέπουν.

Τα χαρακτηριστικά τα οποία εκφράζουν ένα ρούχο είναι διανύσματα τα οποία έχουν την τιμή 1 όταν εκφράζονται από αυτό το χαρακτηριστικό και 0 για όσα χαρακτηριστικά δεν το εκφράζουν:

* Ο πίνακας που ακολουθεί δείχνει τα χαρακτηριστικά του νούμερο για το ρούχο. Στην συγκεκριμένη καταχώρηση το ρούχο είναι μεγέθους LONG

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SHORT | MEDIUM | KNEE | LONG | CAPRI | ¾ LONG |
| 0 | 0 | 0 | 1 | 0 | 0 |

**Πίνακας 2** – Αναπαράσταση διαφορετικών μεγεθών κάθε ρούχου

* Ο πίνακας που ακολουθεί δείχνει τα χαρακτηριστικά της φόρμας του ρούχου. Στην συγκεκριμένη καταχώρηση το ρούχο είναι μεγέθους φόρμας RELAXED FIT

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| RELAXED FIT | REGULAR FIT | SLIM FIT | CARGO FIT | BIKER FIT |
| 0 | 1 | 0 | 0 | 0 |

**Πίνακας 3** – Αναπαράσταση διαφορετικής φόρμας κάθε ρούχου

* Ο πίνακας που ακολουθεί δείχνει τα χαρακτηριστικά του γιακά του ρούχου. Στην συγκεκριμένη καταχώρηση το ρούχο έχει σχέδιο γιακά V

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ROUND NECK | COLLAR | V NECK | HOODED | OFF SHOULDER | HALTERNECK | TURTLENECK |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 |

**Πίνακας 4** – Αναπαράσταση σχεδιασμού γιακά κάθε ρούχου

* Ο πίνακας που ακολουθεί δείχνει την κατηγορία του ρούχου. Στην συγκεκριμένη καταχώρηση το ρούχο είναι κατηγορίας πιτζάμα

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PYJAMAS | SKIRT | TROUSERS | … | CARDIGAN | SHIRT |
| 1 | 0 | 0 |  | 0 | 0 |

**Πίνακας 5** – Αναπαράσταση διαφορετικής κατηγορίας κάθε ρούχου

* Ο πίνακας που ακολουθεί δείχνει τα χαρακτηριστικά για τις λέξεις κλειδιά του ρούχου. Στην συγκεκριμένη καταχώρηση το ρούχο έχει λέξη κλειδί VNECKPYJAMA

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SHORTBLOUSES | SHORTTROUSERS | … | KNEESET | COLLARDRESS | VNECKPYJAMAS |
| 0 | 0 |  | 0 | 0 | 1 |

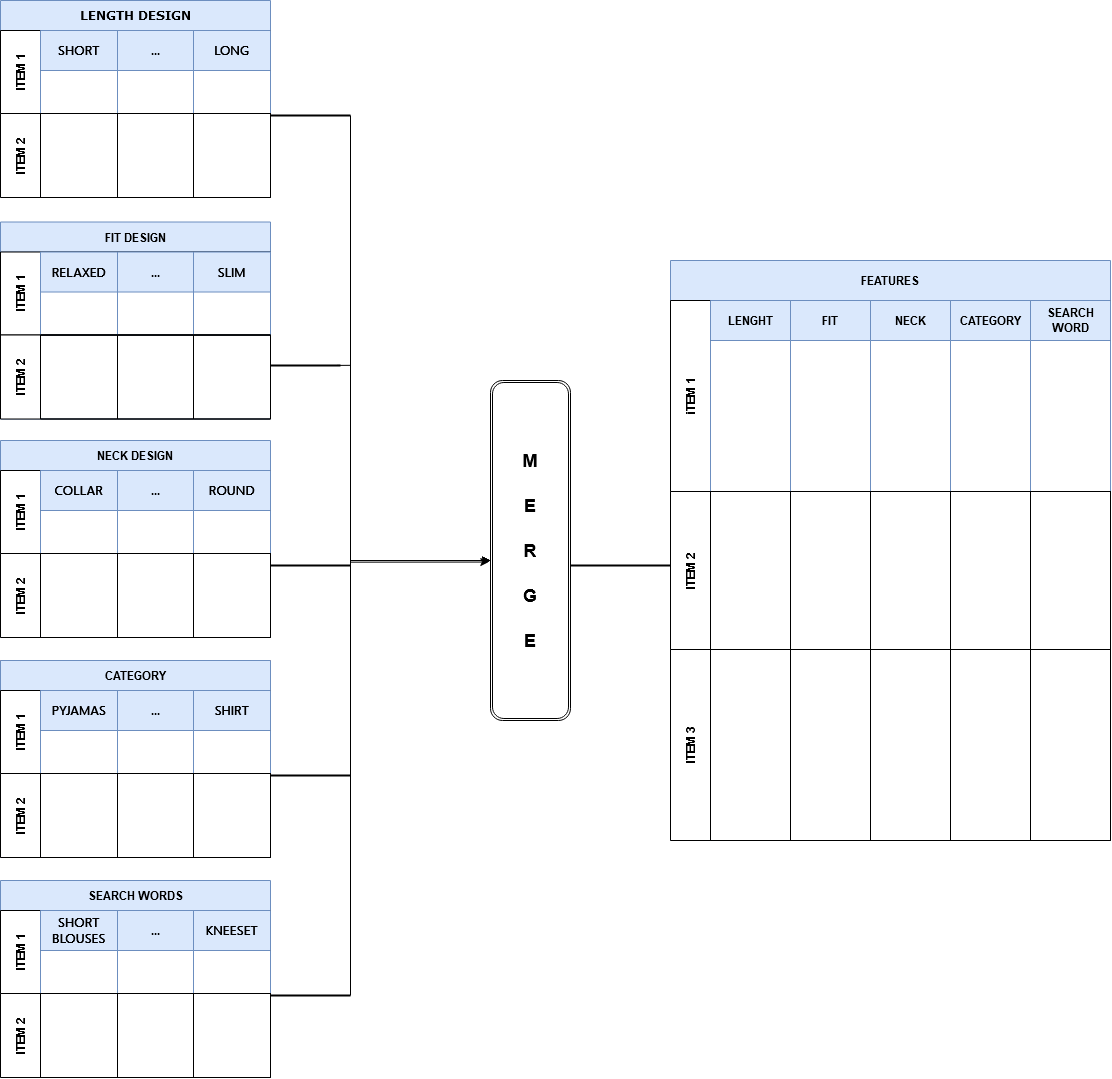
**Πίνακας 6** – Αναπαράσταση λέξεις κλειδιών κάθε ρούχου

Να σημειωθεί ότι σε κάθε ένα από τους παραπάνω πίνακες έγινε ένας καθαρισμός τον περιττών καταχωρήσεων. Το παρακάτω παράδειγμα δείχνει μια περίπτωση περιττής καταχώρησης. Το αποτέλεσμα αυτής της διαδικασίας είναι η μείωση των διαστάσεων κάθε χαρακτηριστικού χωρίς να μειώνεται η πληροφορία.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PYJAMAS,SET | SET | PYJAMAS |  | SET | PYJAMAS |
| 1 | 0 | 0 |  | 1 | 1 |

**Πίνακας 7** - Αναπαράσταση καθαρισμού περιττών καταχωρήσεων

Μετά την εξαγωγή των σημαντικότερων χαρακτηριστικών που περιγράφουν κάθε προϊόν, ακολουθεί η ένωση όλων αυτών των χαρακτηριστικών, ώστε να δημιουργηθεί ένα ενιαίο σύνολο (*Εικόνα 9*), 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 Ρούχων**

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 και με παρεμφερή τρόπο έχοντας μικρές διακυμάνσεις,
* οι χρήστες δεν βαθμολογούν απαραίτητα τα ίδια ρούχα, και
* οι βαθμολογίες κάθε χρήστη δεν είναι ομοιόμορφα κατανεμημένες, φαινόμενο που συναντάται στα σενάρια του πραγματικού κόσμου.

Έχοντας ως καθοδήγηση τις παραπάνω προϋποθέσεις δημιουργούνται δύο σετ δεδομένων αξιολογήσεων. To πρώτο περιλαμβάνει αξιολογήσεις σε ποσοστό 30% επί του συνόλου του δείγματος, ενώ το δεύτερο 60%.

### **Σετ δεδομένων Movielens-100k**

Για το σετ δεδομένων movielens-100k δε χρειάστηκε κάποια προεπεξεργασία, παρά μόνο μια μικρή οργάνωση. Πρόκειται για δεδομένα δημοσιευμένα από την GroupLens που περιλαμβάνουν 100 χιλιάδες αξιολογήσεις από 600 χρήστες σε 9000 ταινίες. Για κάθε ταινία αντιστοιχεί ένα διάνυσμα κατηγοριών, όπου κάθε τιμή της θέσης του διανύσματος υποδεικνύει αν η ταινία ανήκει στην συγκεκριμένη κατηγορία.

#### **Χαρακτηριστικά ταινιών (features)**

Κάθε ταινία του συνόλου δεδομένων περιγράφεται από την κατηγορία στην οποία ανήκει. Το σύνολο δεδομένων περιλαμβάνει ταινίες που μπορούν να ανήκουν σε μία ή περισσότερες κατηγορίες.

Οι κατηγορίες που εκφράζονται οι ταινίες είναι διανύσματα τα οποία έχουν την τιμή 1 όταν εκφράζονται από αυτή την κατηγορία και 0 όταν δεν εκφράζονται από μια κατηγορία:

* Ο πίνακας που ακολουθεί δείχνει τις κατηγορίες που εκφράζουν μια ταινία. Στην συγκεκριμένη καταχώρηση η ταινία είναι κωμωδία και μυστηρίου

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Drama | Comedy | Thriler | … | Sci-Fi | Mystery | Documentary |
| 0 | 1 | 0 |  | 0 | 1 | 0 |

**Πίνακας 8** Αναπαράσταση κάθε κατηγορίας των ταινιών

#### **Βαθμολογίες Ταινιών**

Οι βαθμολογία κάθε ταινίας από τους χρήστες είναι οργανωμένη σε έναν πίνακα βαθμολογιών της μορφής που ακολουθεί παρακάτω. Το εύρος κάθε βαθμολογίας κυμαίνεται [0 έως 5].

|  |  |  |
| --- | --- | --- |
| USER | MOVIE | RATING |
| 0 | 1 | 4.5 |
| 13 | 6 | 3.0 |
| 56 | 44 | 1.0 |

**Πίνακας 9** Αναπαράσταση βαθμολογίας των ταινιών από κάθε χρήστη

## **Συνεργατικό φιλτράρισμα με επίκεντρο τον χρήστη**

### **Δεδομένα εισόδου (input data)**

Στην παρούσα υλοποίηση, η είσοδος δεδομένων που απαιτείται για την εξαγωγή προτάσεων είναι *ο πίνακας χρηστών-αξιολογήσεων* *(Πίνακας 10)* που περιλαμβάνει τις αξιολογήσεις κάθε χρήστη για όλα τα ρούχα/ταινίες που περιέχονται στη βάση δεδομένων. Παρακάτω φαίνεται μια τυπική αναπαράσταση του πίνακα χρηστών-αξιολογήσεων:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Ρούχο / Ταινία 1 | … | Ρούχο / Ταινία m |
| χρήστης 1 | 5 | … | ? |
| χρήστης 2 | 4 | … | 2 |
| …. | ….. | … | … |
| χρήστης n | 7 | … | ? |

**Πίνακας 10** - Αναπαράσταση του πίνακα χρηστών-αξιολογήσεων

### **Βήματα υλοποίησης**

#### **Εύρεση όμοιων χρηστών**

Το αμέσως επόμενο βήμα αυτής της υλοποίησης είναι η εύρεση των χρηστών με την μεγαλύτερη ομοιότητα *(Πίνακας 11)* 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 11** - 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) [32] δεδομένου των κοινών αξιολογήσεων ορίζεται ένα threshold b τέτοιο ώστε:

|  |  |  |
| --- | --- | --- |
|  |  | (18) |

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

Εφόσον ολοκληρωθεί η εύρεση των όμοιων χρηστών μέσα από την στάθμιση της ομοιότητας, είναι αναγκαίο το κεντράρισμα της μέσης τιμής των δεδομένων κάθε 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. [33]

Η νέα αξιολόγηση κάθε χρήστη προκύπτει από τον τύπο

|  |  |  |
| --- | --- | --- |
|  |  | (19) |

όπου η αξιολόγηση του χρήστη u στο ρούχο j και η μέση τιμή των αξιολογήσεων (15).

Παρακάτω ακολουθεί σχηματικά η αναπαράσταση του νέου *πίνακα χρηστών-αξιολογήσεων* (*Πίνακας 12*) με κεντραρισμένη μέση τιμή.

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

**Πίνακας 12** - 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 τη μέση βαθμολογία του χρήστη που είναι στο επίκεντρο. Παρακάτω, παρουσιάζεται ο τύπος της πρόβλεψης των απροσδιόριστων αξιολογήσεων.

|  |  |  |
| --- | --- | --- |
|  |  | (20) |

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.

## **Συνεργατικό φιλτράρισμα με ανάλυση πίνακα σε ιδιάζουσες τιμές**

### **Δεδομένα Εισόδου (input data)**

Όμοια με την προηγούμενη υλοποίηση, η είσοδος δεδομένων που απαιτείται για την εξαγωγή προτάσεων είναι ο *πίνακας χρηστών-αξιολογήσεων* (*Πίνακας 13*) 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 13** - Αναπαράσταση του πίνακα χρηστών-αξιολογήσεων

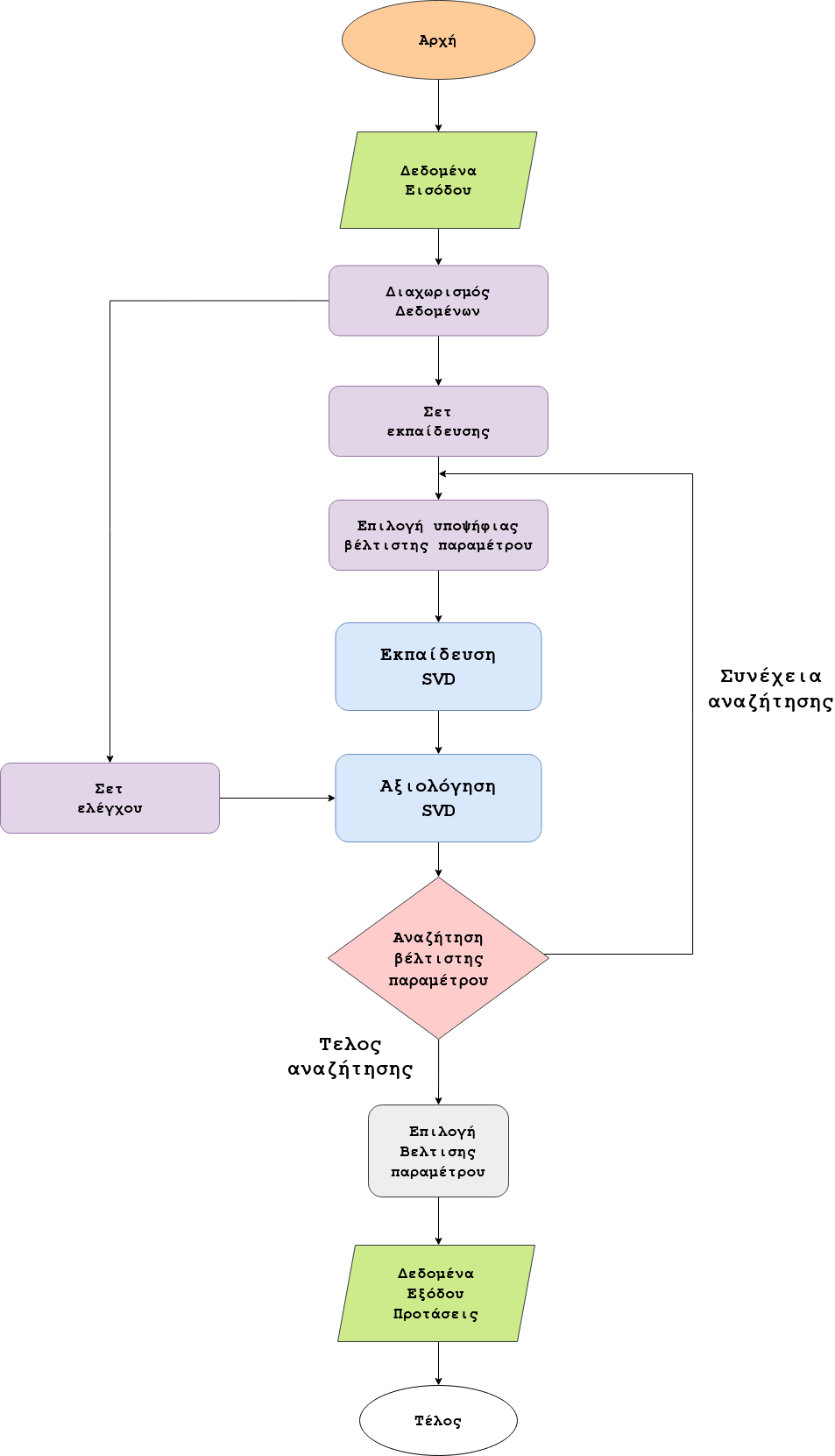
### **Βήματα υλοποίησης**

#### **Επιλογή βιβλιοθήκης μοντέλου**

Το μοντέλο ανάλυσης πίνακα σε ιδιάζουσες τιμές είναι απαραίτητο να εκπαιδευτεί με τις αξιολογήσεις των χρηστών, και να αξιολογηθεί, βάσει ορισμένων χρήσιμων μετρικών. Η υλοποίηση τρέχει κάνοντας χρήση της βιβλιοθήκης Numpy με την εντολή np.linalg.svd. Η οποία παίρνει σαν είσοδο τον πίνακα προς παραγοντοποίηση.

#### **Επιλογή παραμέτρων**

Η χρησιμότητα της ανάλυσης πίνακα σε ιδιάζουσες τιμές έγκειται στη δυνατότητα διατήρησης των k μεγαλύτερων ιδιαζουσών τιμών. Ένα μοντέλο ανάλυσης πίνακα σε ιδιάζουσες τιμές μπορεί να εκπαιδευτεί έχοντας του τεθεί διαφορετικές τιμές στον αριθμό k. Συνεπώς κατά την εκπαίδευση του μοντέλου γίνονται διάφορες δοκιμές τιμών k που μέσω ενός πακέτου ελέγχου επιλέγεται η βέλτιστή τιμή της παραμέτρου.



**Εικόνα 10** - Οπτική αναπαράσταση βημάτων μοντέλου ανάλυσης πίνακα σε ιδιάζουσες τιμές

Το μοντέλο αρχικά δέχεται τα δεδομένα εισόδου τα οποία και διαχωρίζει σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου. Στην συνέχεια γίνεται κάποια επιλογή παραμέτρων συγκεκριμένων τιμών που αποτελούν υποψήφιες βέλτιστες τιμές. Το σύστημα εκπαιδεύεται με βάση αυτές και κάνει επικύρωση στο σετ ελέγχου. Τελικό στάδιο είναι η επιλογή της βέλτιστης παραμέτρου και η εξαγωγή προτάσεων βάση αυτής (*Εικόνα 10).*

## **Συνεργατικό φιλτράρισμα μέσω νευρωνικών δικτύων**

### **Δεδομένα εισόδου (input data)**

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-Προϊόντων (Ρούχων/Ταινιών)**

Τα πεδία χρηστών / προϊόντων που υπάρχουν στο σύνολο δεδομένων με τις αξιολογήσεις είναι μη διαδοχικοί ακέραιοι αριθμοί (π.χ. clothID: 3, 12, 15, 21) 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 (π.χ. clothID: 0, 1, 2, 3). Για την επίτευξη αυτής της ανάγκης γίνεται η χρήση της *κλάσης LabelEncoder* *του scikit-learn.* Ο λόγος που συμβαίνει αυτή η κωδικοποίηση είναι διότι η βιβλιοθήκη keras χρειάζεται αυτά τα δεδομένα να είναι κωδικοποιημένα σε διαδοχικούς ακέραιους ώστε να μπορεί να τρέξει η υλοποίηση.

#### **Επίπεδο Ενσωμάτωσης**

Εφόσον ολοκληρωθεί η κωδικοποίηση δεδομένων του αρχικού σετ, και ο διαχωρισμός των δεδομένων σε 80% σετ εκπαίδευσης και 20% σετ ελέγχου από κάθε χρήστη, ακολουθεί η τροφοδότηση του επιπέδου ενσωμάτωσης, ένα πλήρως διασυνδεδεμένο 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 λανθάνοντα 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 το επίπεδο ενσωμάτωσης προϊόντων.

#### **Αρχιτεκτονική πολύ-επίπεδο νευρωνικού δικτύου**

To ενσωματωμένο επίπεδο εξάγει την αναπαράσταση για κάθε χρήστη και προϊόν σε πυκνά διανύσματα. Τα διανύσματα αυτά περιλαμβάνουν τις 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.

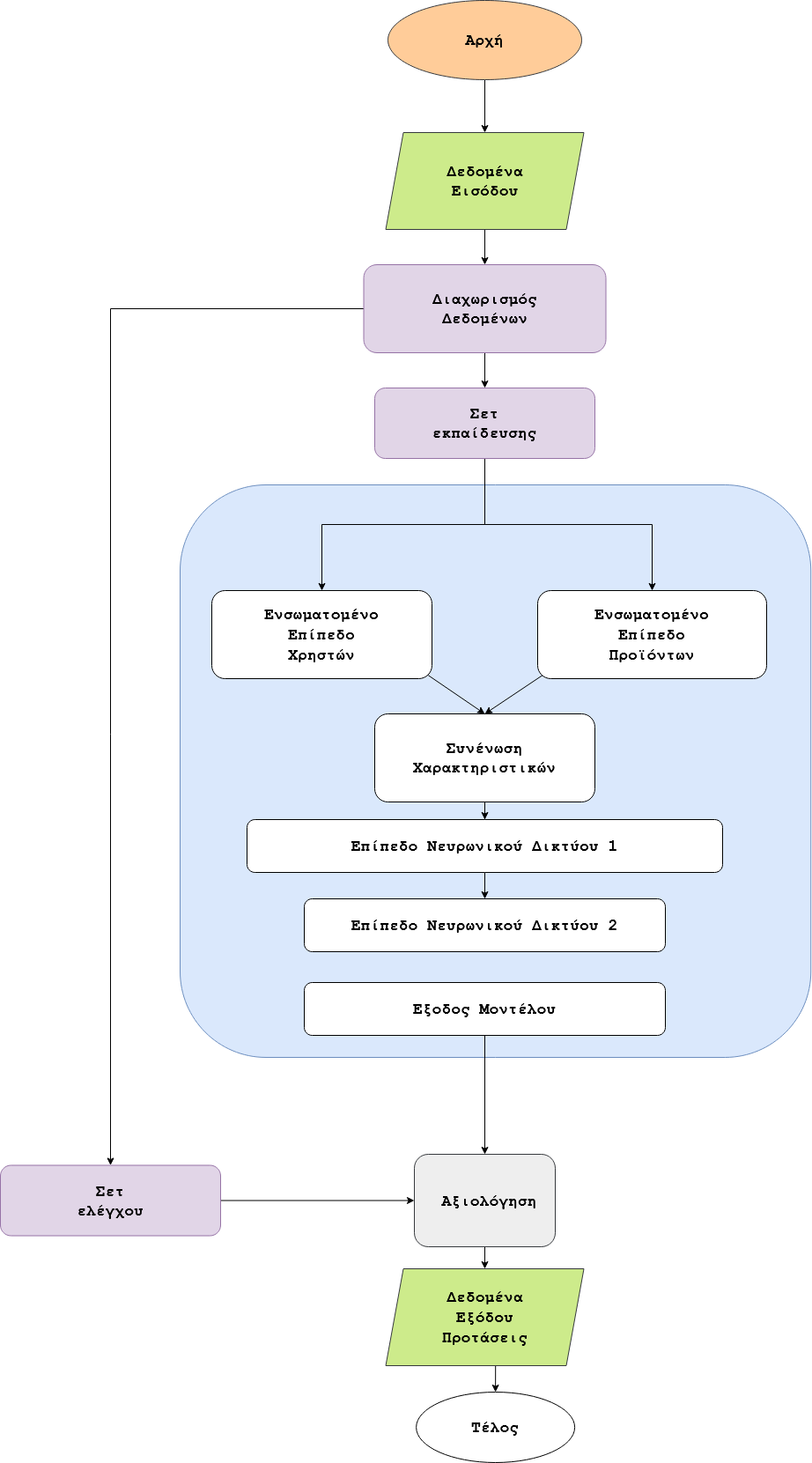
Για την διευθέτηση αυτού του ζητήματος γίνεται χρήση του νευρωνικού δικτύου, ώστε να γίνει η εκπαίδευση για την αναγνώριση αυτών των αλληλεπιδράσεων και την αναγνώριση μη γραμμικών σχέσεων, μια λειτουργία αδύνατη για τα μοντέλα ανάλυσης πίνακα ιδιαζουσών τιμών. Αυτά τα χαρακτηριστικά εισέρχονται σε ένα πολύ-επίπεδο νευρωνικό δίκτυο, του οποίου η αρχιτεκτονική μοιάζει με πύργος, μια αρχιτεκτονική ευρέως χρησιμοποιούμενη που επιτρέπει στο δίκτυο να αναγνωρίζει στα υψηλά στρώματα χαρακτηριστικά με μεγαλύτερο επίπεδο αφαίρεσης [16].

Τα επίπεδα δημιουργούνται με την χρήση της κλάσης keras.layers.Dense και οι νευρώνες 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 *(Πίνακας 14).* [34]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Δομή Αρχιτεκτονικής Πολύ-επίπεδου Νευρωνικού | | | | |
| Επίπεδα | **Νευρώνες** | **Συνάρτηση Ενεργοποίησης** | **Εγκατάλειψη**  **(Dropout)** | **Αρχικοποίηση βαρών** |
| Επίπεδο 1 | 4096 | ReLu | 50 % | he\_normal |
| Επίπεδο 2 | 2048 | ReLu | 50 % | he\_normal |
| Επίπεδο 3 | 1024 | ReLu | 30 % | he\_normal |
| Επίπεδο 4 | 512 | ReLu | 30 % | he\_normal |
| Επίπεδο 5 | 256 | ReLu | 30 % | he\_normal |
| Επίπεδο 6 | 128 | ReLu | 30 % | he\_normal |
| Επίπεδο 7 | 1 | Sigmoid | 0 % | Glorot Normal |

**Πίνακας 14** – Αναπαράσταση δομής πολύ-επίπεδου νευρωνικού

Το μοντέλο αρχικά δέχεται τα δεδομένα εισόδου, τα οποία διαχωρίζει σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου. Στην συνέχεια, τα δεδομένα εκπαίδευσης εισάγονται στα επίπεδα ενσωμάτωσης, όπου εξάγονται οι αναπαραστάσεις κάθε χρήστη και προϊόντος σε πυκνά διανύσματα. Αυτά τα χαρακτηριστικά τροφοδοτούν ένα πολύ-επίπεδο νευρωνικό το οποίο εκπαιδεύεται ώστε να τα αναγνωρίζει. Τελικό στάδιο είναι η εξαγωγή των τελικών προτάσεων *(Εικόνα 11).*



**Εικόνα 11** - Οπτική αναπαράσταση βημάτων μοντέλου συνεργατικού φιλτραρίσματος νευρωνικών δικτύων

## **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, καθώς και ο πίνακας χρηστών-αξιολογήσεων που περιλαμβάνει τις αξιολογήσεις κάθε χρήστη για όλα τα ρούχα που περιέχονται στη βάση δεδομένων.

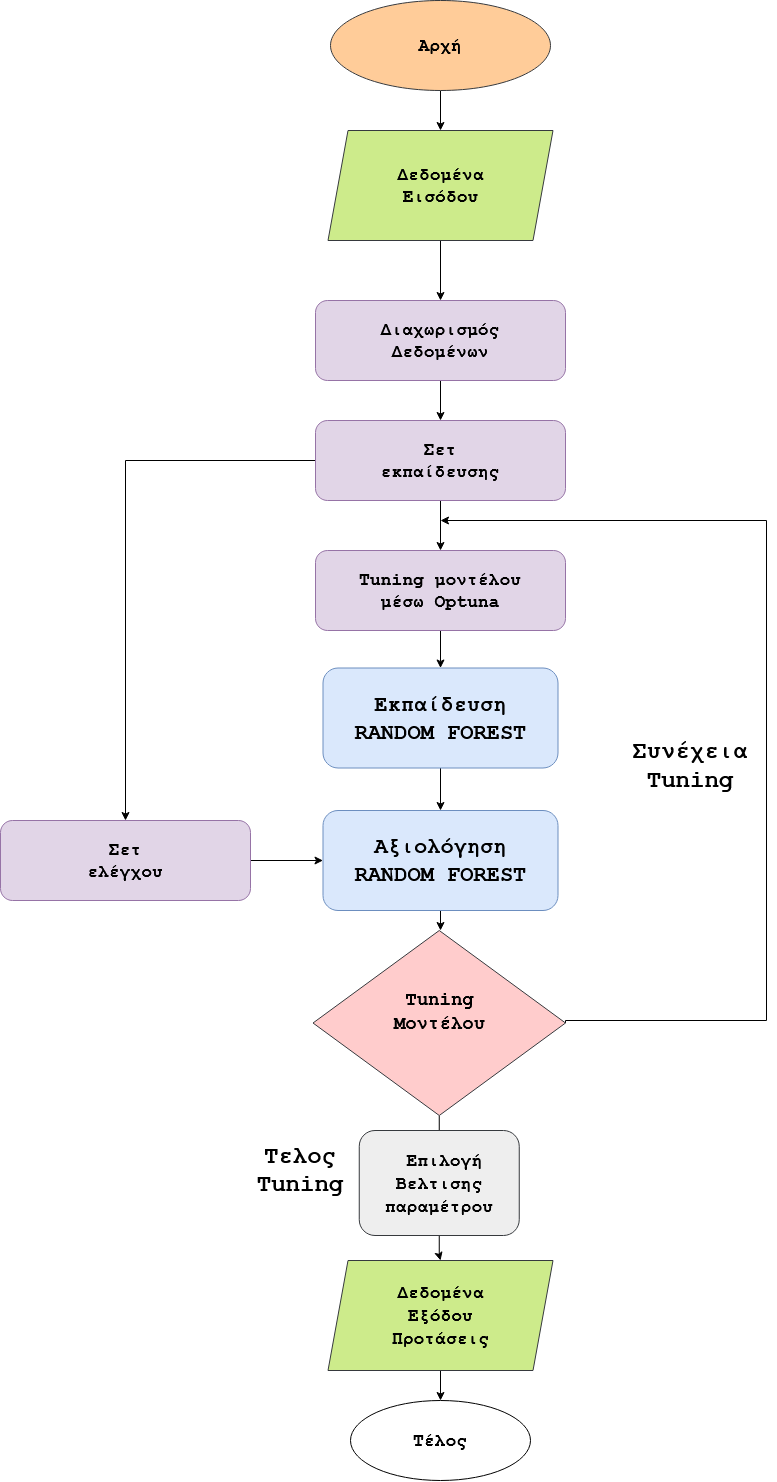
### **Βήματα υλοποίησης**

#### **Επιλογή βιβλιοθήκης μοντέλου**

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

#### **Επιλογή παραμέτρων**

Ένα τυχαίο δάσος μπορεί να εκπαιδευτεί για διάφορες τιμές των διαθέσιμων 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 [35], η οποία αναζητά σε ένα χώρο περιορισμένου εύρους τις βέλτιστες παραμέτρους που δίνουν επιθυμητές επιδόσεις τα μοντέλα Τυχαίων Δασών.



**Εικόνα 12** - Οπτική αναπαράσταση βημάτων μοντέλου Τυχαίων Δασών

Το μοντέλο αρχικά δέχεται τα δεδομένα εισόδου, τα οποία διαχωρίζει σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου. Στην συνέχεια, το μοντέλο εκπαιδεύεται μεταβάλλοντας κάποιες παραμέτρους ώστε να επιτευχθούν οι βέλτιστες επιδόσεις. Μόλις η διαδικασία δοκιμών ολοκληρωθεί το τελικό στάδιο είναι η εξαγωγή των τελικών προτάσεων μέσα από την επιλογή των καλύτερων παραμέτρων *(Εικόνα 12).*

## **Φιλτράρισμα Περιεχομένου με Νευρωνικά Δίκτυα**

### **Δεδομένα Εισόδου**

Για τις υλοποιήσεις περιεχομένου, τα δεδομένα εισόδου που απαιτούνται 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, 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 η λήψη των απαραίτητων δεδομένων για κάθε χρήστη, 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.

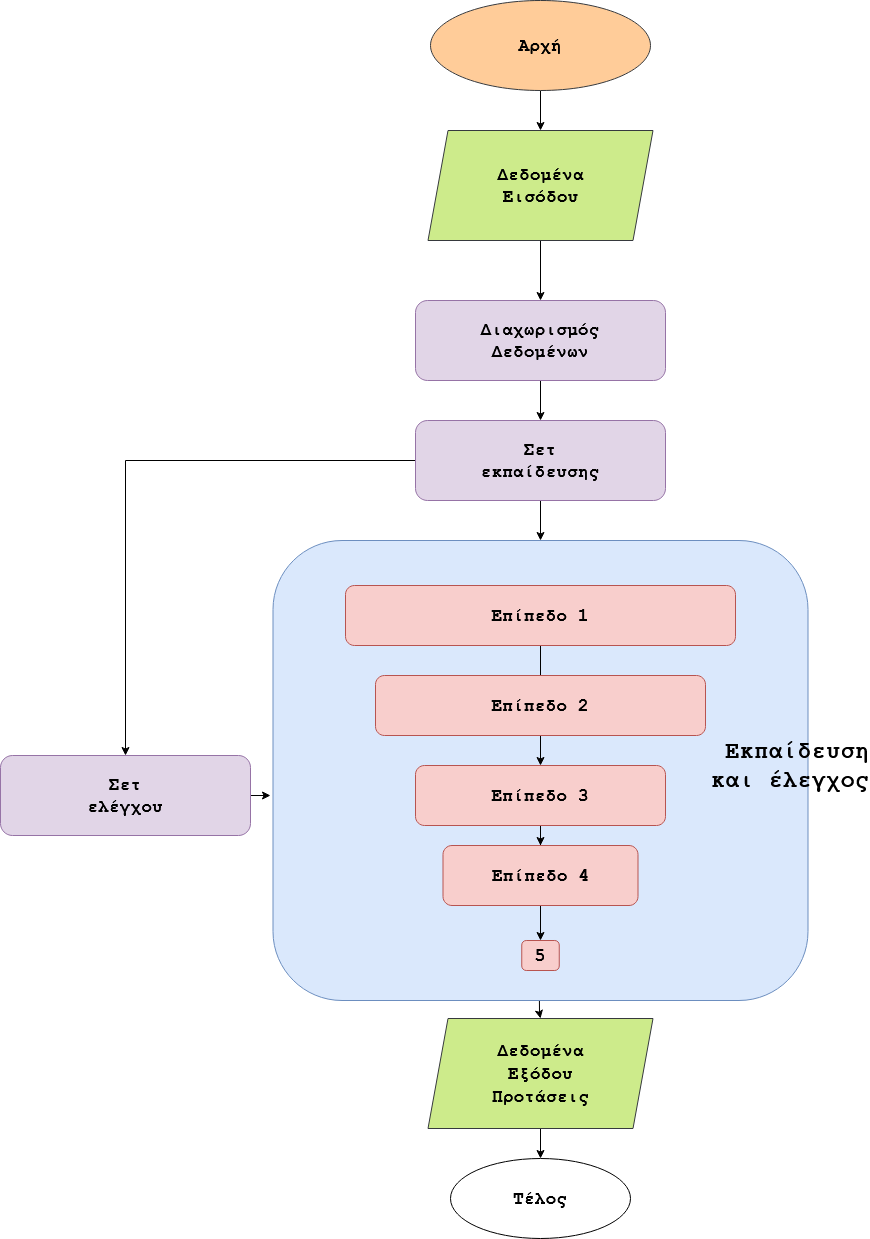
Τα χαρακτηριστικά αυτά εισέρχονται σε ένα πολύ-επίπεδο νευρωνικό δίκτυο του οποίου η αρχιτεκτονική είναι ίδια με της υλοποίησης του συνεργατικού νευρωνικού δικτύου. Όπως αναφέρθηκε ήδη, η συγκεκριμένη αρχιτεκτονική είναι ευρέως χρησιμοποιούμενη που επιτρέπει στο δίκτυο να αναγνωρίζει στα υψηλά στρώματα 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 νευρώνων, 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 *(Πίνακας 15).* Η αρχικοποίηση βαρών είναι ακριβώς αντίστοιχη του Συνεργατικού Νευρωνικού Δικτύου [26].

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Δομή Αρχιτεκτονικής Πολύ-επίπεδου Νευρωνικού | | | | |
| Επίπεδα | **Νευρώνες** | **Συνάρτηση Ενεργοποίησης** | **Εγκατάλειψη**  **(Dropout)** | **Αρχικοποίηση βαρών** |
| Επίπεδο 1 | 2048 | ReLu | 20 % | he\_normal |
| Επίπεδο 2 | 1024 | ReLu | 20 % | he\_normal |
| Επίπεδο 3 | 512 | ReLu | 10 % | he\_normal |
| Επίπεδο 4 | 256 | ReLu | 10 % | he\_normal |
| Επίπεδο 5 | 1 | Sigmoid | 0 % | Glorot normal |

**Πίνακας 15** – Αναπαράσταση Δομής πολύ-επίπεδου νευρωνικού περιεχομένου

Η διαδικασία που ακολουθείται είναι η εξής: αρχικά, το μοντέλο δέχεται τα δεδομένα εισόδου και τα διαχωρίζει σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου. Στην συνέχεια, εκπαιδεύεται τροφοδοτώντας το νευρωνικό δίκτυο με τα χαρακτηριστικά των ρούχων και τις βαθμολογίες που τους αντιστοιχούν. Μόλις η διαδικασία εκπαίδευσης ολοκληρωθεί το τελικό στάδιο είναι η εξαγωγή προτάσεων *(Εικόνα 13).* Ένα πλεονέκτημα αυτής της υλοποίησης είναι ότι τα νευρωνικά δίκτυα εκπαιδεύουν τα βάρη τους αυτόματα, ώστε τελικά να επιλέγονται οι βέλτιστη παράμετροι χωρίς να χρειάζεται κάποιο tuning κατά την διάρκεια της εκπαίδευσης. Αυτό δεν γίνεται στην περίπτωση της υλοποίησης με τα τυχαία δάση.



**Εικόνα 13** - Οπτική αναπαράσταση βημάτων μοντέλου Νευρωνικών Περιεχομένου

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. Με δεδομένο ότι ο μοναδικός τρόπος αξιολόγησης των προτάσεων ήταν η χρήση των ήδη γνωστών βαθμολογιών, τα δεδομένα χωρίστηκαν με τέτοιο τρόπο 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 14).*



**Εικόνα 14** - 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 15*).



**Εικόνα 15** - Αναπαράσταση Διαχωρισμού Δεδομένων - 2

## **Σύνολα δεδομένων**

Ο παραπάνω διαχωρισμός που περιεγράφηκε εφαρμόζεται σε 3 σύνολα δεδομένων, όπως αναφέρθηκε και στο Κεφάλαιο 4.1.

Τα δύο πρώτα αφορούν προϊόντα ένδυσης της εταιρίας ASOS και περιλαμβάνουν αξιολογήσεις από 100 χρήστες σε 1700 ρούχα. Η διαφορά μεταξύ αυτών των δύο συνόλων είναι ότι στο πρώτο περιλαμβάνει τις αξιολογήσεις του 30% του συνόλου των προϊόντων που μπορεί να αξιολογήσει ο χρήστης, ενώ το δεύτερο το 60% των αξιολογήσεων σε σχέση 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 η βελτίωσή τους, καθώς το σύστημα στη δεύτερη περίπτωση διαθέτει περισσότερα δεδομένα εκπαίδευσης απ’ ότι στην πρώτη.

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.

Οι αξιολογήσεις στα πρώτα δύο σύνολα δεδομένων της εταιρίας Asos κυμαίνονται στην κλίμακα [0 έως 10] ενώ οι αξιολογήσεις του movielens με τις ταινίες στην κλίμακα [0 έως 5]. Το γεγονός ότι δύο πρώτα σύνολα δεδομένων αποτελούνται από τεχνητές αξιολογήσεις, ενώ το τρίτο σύνολο δεδομένων από πραγματικές αξιολογήσεις είναι μια καλή αφορμή για να σχολιαστεί η επίδοση του γενικευμένου συστήματος προτάσεων σε τεχνητά και πραγματικά δεδομένα, ως τεκμήριο της αξιοπιστία του.

Στον *Πίνακας 16* παρουσιάζονται συνοπτικά τα προαναφερθέντα σύνολα δεδομένων:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Όνομα Σετ Δεδομένων | Αντικείμενα | Χρήστες | Αξιολογήσεις | Ποσοστό αξιολογήσεων (%) | Κλίμακα | Κατηγορίες |
| Asos30 | 1700 | 100 | 50k | 30 % | [0,10] | 10 |
| Asos60 | 1700 | 100 | 100k | 60 % | [0,10] | 10 |
| Movielens-100k | 9000 | 600 | 100k | 2 % | [0,5] | 11 |

**Πίνακας 16** - Συνοπτική αναπαράσταση για κάθε σύνολο δεδομένων

## **Μετρικές Αξιολόγησης**

### **Accuracy, Recall, Precision και F1 Score**

Ένα σύστημα προτάσεων έχει ως τελικό στόχο να προβάλει στον χρήστη προϊόντα ή υπηρεσίες όσο πιο κοντά στις προτιμήσεις του είναι εφικτό. Γι’ αυτόν τον λόγο, είναι απαραίτητο να αξιολογήσουμε πόσο πετυχημένες είναι οι προτάσεις αυτών των συστημάτων, με μετρικές όπως το accuracy, το precision, και το recall. Οι προτάσεις θεωρούνται ένα πρόβλημα δυαδικής ταξινόμησης, όπου ένα σύστημα με χαμηλό precision και recall θα υστερεί αφού θα αδυνατεί να αναγνωρίσει τις πραγματικές προτιμήσεις του χρήστη.

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 *(Εικόνα 16)* 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 (*Εικόνα 17*).



**Εικόνα 16** - Διαχωρισμός αξιολογήσεων [0,10]



**Εικόνα 17** - Διαχωρισμός αξιολογήσεων [0,5]

### **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 πάντα αρκετά υψηλά ή χαμηλά, τα αποτελέσματα των 4 μετρικών δεν καταφέρνουν να αποτυπώσουν με σωστό τρόπο την καλή ή την όχι τόσο καλή απόδοση του συστήματος.

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. Αυτό τον περιορισμό έρχονται να άρουν οι μετρικές RSME και MAE που αναφέρθηκαν στις ενότητες 2.3.2 και 2.3.3. Η μετρική RMSE, με την οποία υπολογίζουμε την ρίζα του μέσου τετραγωνικού σφάλματος, 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 το αντίστροφο) η μετρική αυτή να αυξάνεται περισσότερο. 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**

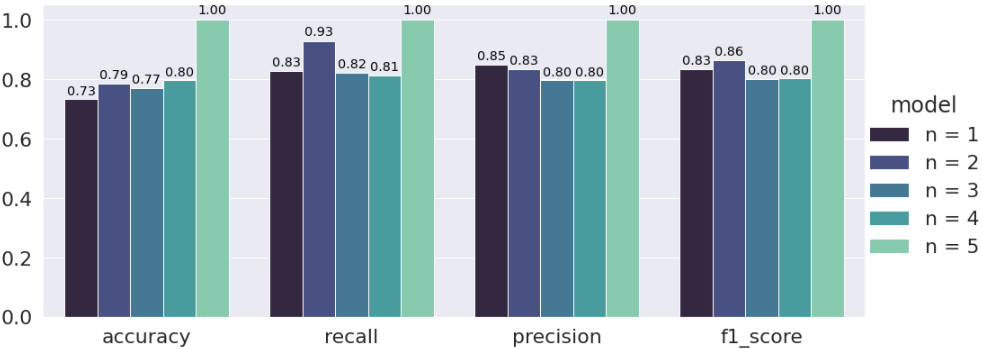
Το διάγραμμα των τεσσάρων μετρικών για τα σύνολα data30, data60 δείχνει πως η ικανότητα πρόβλεψης του μοντέλου αλλάζει συναρτήσει του αριθμούγειτόνων.

Για το σύνολο δεδομένων data30 το accuracy βελτιώνεται αισθητά από 73% με 1 γείτονα σε 0.8% με 4 γείτονες, ενώ με 5 γείτονες φτάνουν και οι 4 μετρικές στο 100% επιτυχημένης ταξινόμησης. To recall δίνει καλή βελτίωση από 0.83 με ένα γείτονα σε 0.93 με δύο γείτονες ενώ τα precision και f1\_score κυμαίνονται περίπου στα ίδια για έως 4 γείτονες με μια ελαφρά μείωση *(Εικόνα 18).*

Για το σύνολο δεδομένων data60 το accuracy να βελτιώνεται αισθητά από 72% με 1 γείτονα σε 0.80% με 4 γείτονες όπως όμοια και με 5 γείτονες. To recall δίνει καλή βελτίωση από 0.82% με ένα γείτονα σε 0.93% με δύο γείτονες και 0.96% με 4 γείτονες ενώ τα precision και f1\_score κυμαίνονται περίπου στα ίδια για έως 4 γείτονες με μια ελαφρά μείωση *(Εικόνα 19).*

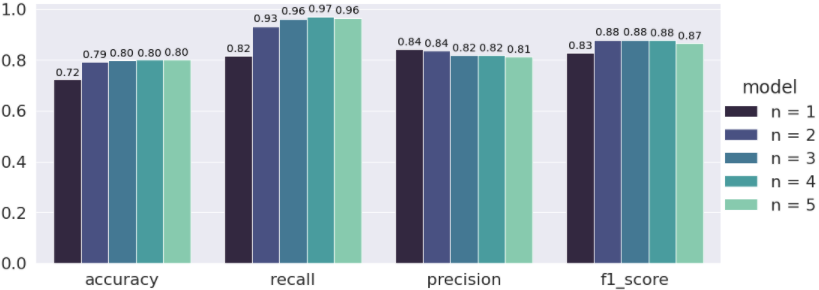
Γενικότερα, η αύξηση του αριθμού των γειτόνων είναι εφικτό να δώσει καλύτερα αποτελέσματα στην ταξινόμηση των προτάσεων, ωστόσο όσο αυξάνει αυτός ο αριθμός τόσο λιγότερες προτάσεις δημιουργούνται.

Σετ δεδομένων: data30



**Εικόνα 18** - Υλοποίηση User-Based - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score ανά γείτονα – data30

Σετ δεδομένων: data60



**Εικόνα 19** - 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 δείχνουν πως η αύξηση των γειτόνων στο μοντέλομειώνει αισθητά τις δύο μετρικές σφάλματος.

Για το σύνολο data30, η ρίζα μέσου τετραγωνικού σφάλματος μειώνεται αισθητά αφού από 3.44 με 1 γείτονα φτάνει στο 2.11 με 5 γείτονες. To ίδιο συμβαίνει και με το μέσο απόλυτο σφάλμα, από 2.77 με 1 γείτονα φτάνει στο 2.07 με 5 γείτονες *(Εικόνα 20).*

Για το σύνολο data60 παρουσιάζονται παρόμοια αποτελέσματα. καθώς η ρίζα μέσου τετραγωνικού σφάλματος μειώνεται από το 3.50 με 1 γείτονα στο 2.43 με 5 γείτονες και το μέσο απόλυτο σφάλμα από 2.77 με 1 γείτονα στο 2.14 με 5 γείτονες *(Εικόνα 21).*

Θα περίμενε κανείς να δει βελτιωμένα αποτελέσματα στο σύνολο data60 σε σχέση με το data30, λόγω του γεγονότος ότι υπάρχουν περισσότερα δεδομένα για να εκπαιδευτούν οι προτάσεις. Παρόλα αυτά του το σύστημα ωστόσο αυτό δε συμβαίνει, πράγμα που σημαίνει ότι η αύξηση τον δεδομένων δεν επηρεάζει την αξιοπιστία της υλοποίησης.

|  |  |
| --- | --- |
| Σετ δεδομένων: data30 | Σετ δεδομένων: data60 |
|  |  |
| **Εικόνα 20** - Υλοποίηση User-Based - Διάγραμμα μετρικών rmse και mae ανά γείτονα – data30 | **Εικόνα 21** - Υλοποίηση User-Based - Διάγραμμα μετρικών rmse και mae ανά γείτονα - data60 |

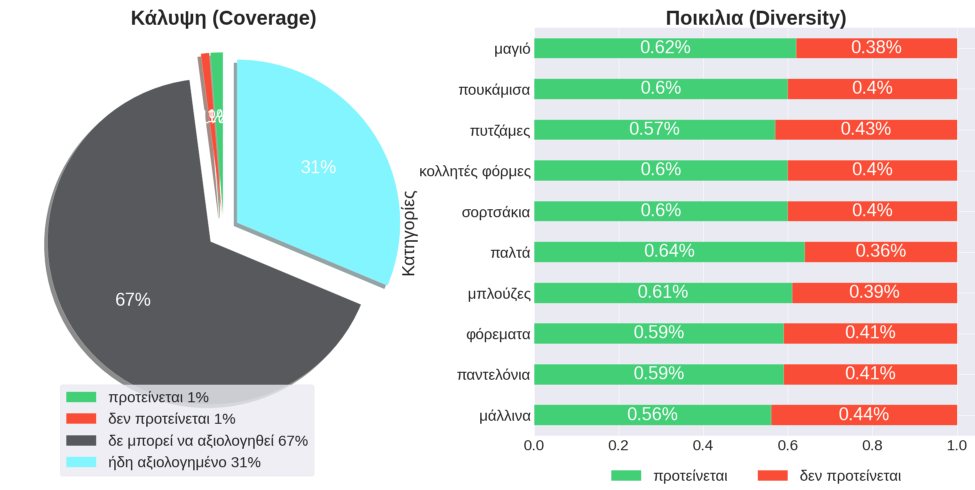
### **Κάλυψη και Ποικιλία**

Για το σετ δεδομένων data30 φαίνεται ότι από τα μη αξιολογημένα ρούχα βαθμολογήθηκαν θετικά το 2% και αρνητικά μόλις το 1% συνολικά για όλους τους χρήστες στο σύστημα, ενώ το μεγαλύτερο ποσοστό (67%) έμεινε μη αξιολογημένο *(*Εικόνα 22*).* Στο σετ δεδομένων data60 βαθμολογήθηκαν θετικά το 5% και αρνητικά το 4% ενώ το μεγαλύτερο ποσοστό (30%) έμεινε μη αξιολογημένο *(*Εικόνα 23*).*

Οι κατηγορίες των ρούχων στο σύνολο δεδομένων data30 δείχνουν μια ισορροπία στις προτάσεις, δηλαδή δεν απορρίπτουν σε μεγάλο ποσοστό πολλά ρούχα από μια συγκεκριμένη κατηγορία. Το ίδιο ακριβώς ισχύει και για το σύνολο data60.

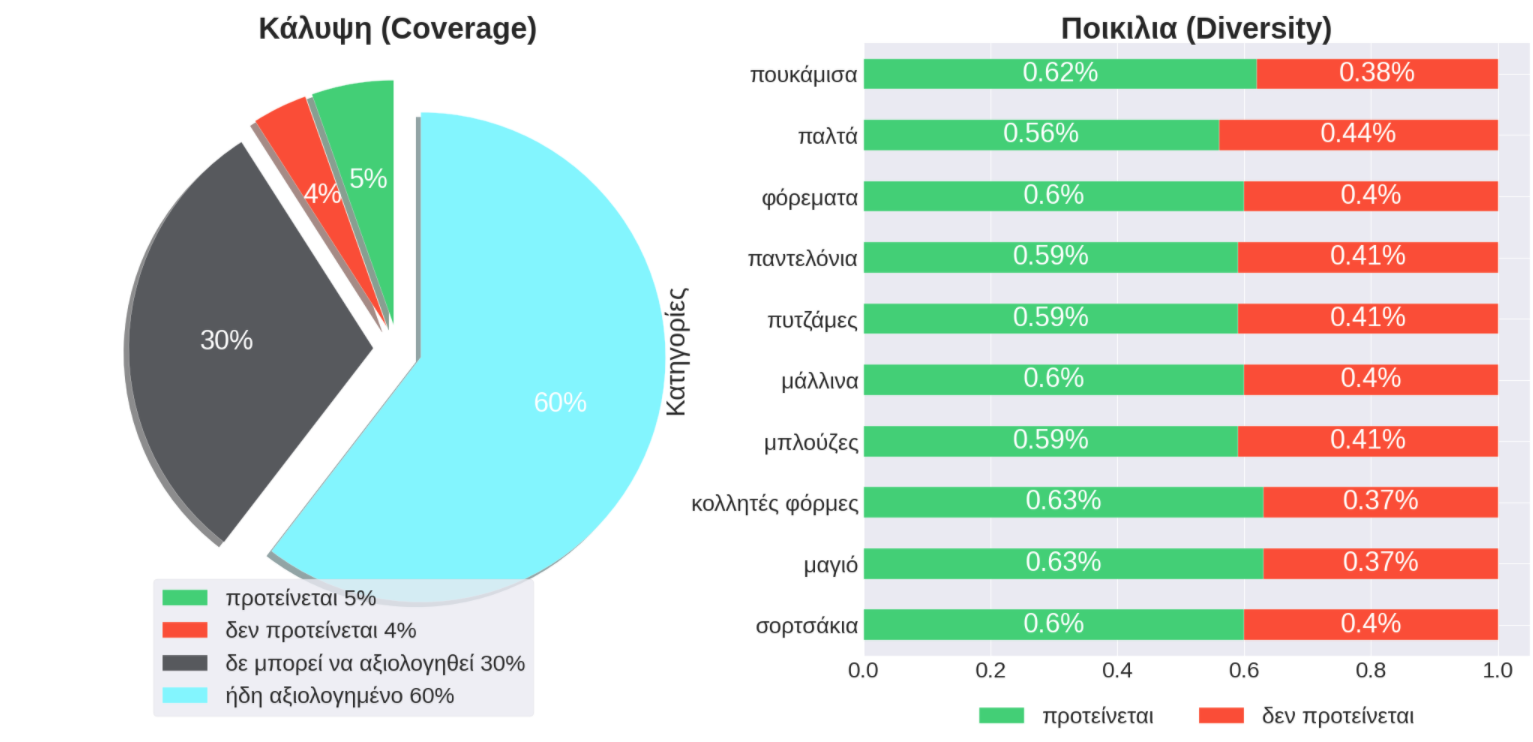
Στην συνέχεια, τα διαγράμματα Κάλυψης και Ποικιλίας δείχνουν πως η παρούσα υλοποίηση του συνεργατικού φιλτραρίσματος αποτυγχάνει να ανακαλύψει ικανοποιητικά τα προϊόντα της βάσης, ωστόσο παρέχει μια ισορροπημένη ποικιλία ανά κατηγορία ρούχου. Ένας λόγος είναι ότι η φύση αυτής της υλοποίησης έχει ισχυρή εξάρτηση μεταξύ των αξιολογήσεων από χρήστη σε χρήστη.

Σετ δεδομένων: data30



**Εικόνα 22 -** Υλοποίηση User-Based - Διάγραμμα Κάλυψης και Ποικιλίας – data30

Σετ δεδομένων: data60



**Εικόνα 23** - 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 και data60, ενώ αλλάζει ελάχιστα για το σύνολο movielens-100k.

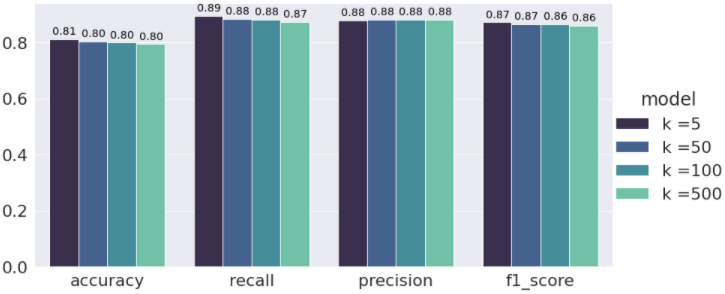
Βλέπουμε για το σύνολο δεδομένων data30 ότι για διαφορετικές τιμές k το accuracy δε μεταβάλλεται από το 77%, το recall κυμαίνεται 99% με 100%, όπως το precision 77% με 78%, ενώ το f1\_score παραμένει σταθερό στο 87% *(Εικόνα 24).* Το σύνολο δεδομένων data60 για διαφορετικές τιμές k το accuracy και το recall δε μεταβάλλονται από το 82% και 100% αντίστοιχα, το ίδιο ισχύει για τα precision και f1\_score *(Εικόνα 25)*. Τέλος, το σύνολο movielens 100K για διαφορετικές τιμές του k έχει accuracy που κυμαίνεται από 80% με 81% recall από 87% με 89%, precision σταθερό στα 88% και f1\_score από 86% με 87% *(Εικόνα 26).*

Ένα συμπέρασμα που προκύπτει είναι ότι εφόσον υπάρχει ελάχιστη ή καθόλου επίδραση στην ακρίβεια του μοντέλου με τη μεταβολή της τιμής των χαρακτηριστικών, αυτό κάνει το σύστημα λιγότερο ευαίσθητο στην ρύθμιση του και περισσότερο αξιόπιστο. Αποτελεί, μεν, μια μετρική που πρέπει να ρυθμιστεί σωστά, ωστόσο μια λανθασμένη απόφαση δε θα δημιουργούσε μεγάλο πρόβλημα.

Δεύτερο συμπέρασμα που προκύπτει από τις επιδόσεις των συνόλων data30 και data60 είναι ότι με την αύξηση 30% των κριτικών προκαλούνται θετικέςεπιδράσεις στα ζητήματα ταξινόμησης αφού έχουμε αύξηση του accuracy και precision από 77% σε 82% ενώ του f1\_score από 87% σε 90%.

|  |  |
| --- | --- |
| Σετ δεδομένων: data30 | Σετ δεδομένων: data60 |
|  |  |
| **Εικόνα 24** - Υλοποίηση SVD - Διάγραμμα μετρικών αccuracy, recall, precision και f1\_score ανά τιμή χαρακτηριστικών – data30 | **Εικόνα 25** - Υλοποίηση SVD - Διάγραμμα μετρικών αccuracy, recall, precision και f1\_score ανά τιμή χαρακτηριστικών – data60 |

Σετ δεδομένων: movielens 100K



**Εικόνα 26** - 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)**

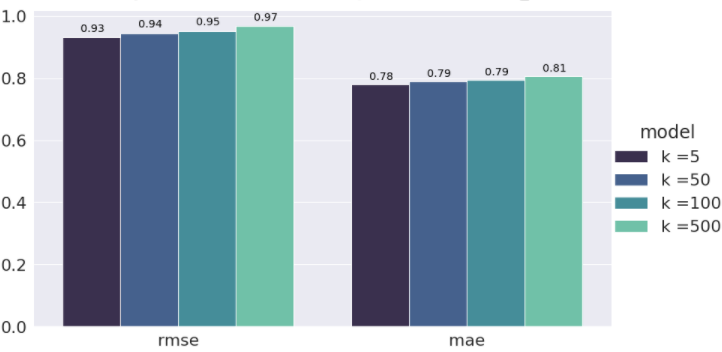
Τα διαγράμματα των ΜΑΕ και RMSE ανά τιμή χαρακτηριστικών δείχνουν πως η ικανότητα πρόβλεψης του μοντέλου αλλάζει ελάχιστα για όλα τα σύνολα, ως προς το σφάλμα πρόβλεψης.

Για το σύνολο δεδομένων data30 παρατηρείται ότι η ρίζα μέσου τετραγωνικού σφάλματος κυμαίνεται από 2.73 έως 2.77 και το μέσο απόλυτο σφάλμα από 2.34 έως 2.40 για διαφορετικές τιμές του k *(Εικόνα 27).* Αντίστοιχα, στο σύνολο δεδομένων data60 η ρίζα μέσου τετραγωνικού σφάλματος κυμαίνεται από 2.55 έως 2.70 και το μέσο απόλυτο σφάλμα από 2.17 έως 2.33 για διαφορετικές τιμές του k *(Εικόνα 28).* Τέλος, στο σύνολο movielens 100K η ρίζα μέσου τετραγωνικού σφάλματος κυμαίνεται από 0.93 έως 0.97 και το μέσο απόλυτο σφάλμα από 0.78 έως 0.81 για διαφορετικές τιμές του k *(Εικόνα 29).*

Τα συμπεράσματα που προκύπτουν είναι όμοια με αυτά των τεσσάρων μετρικών, δηλαδή ότι εφόσον υπάρχει ελάχιστη ή καθόλου επίδραση στην ακρίβεια του μοντέλου με τη μεταβολή της τιμής των χαρακτηριστικών, αυτό κάνει το σύστημα λιγότερο ευαίσθητο στην ρύθμιση του και περισσότερο αξιόπιστο. Ομοίως, από τις επιδόσεις των συνόλων data30 και data60 προκύπτει ότι με την αύξηση 30% των κριτικών προκαλούνται θετικέςεπιδράσεις στα ζητήματα ταξινόμησης, καθώς έχουμε βελτίωση του ελάχιστου τετραγωνικού σφάλματος από 2.73 σε 2.55 και του μέσου απόλυτου σφάλματος από 2.40 σε 2.17.

|  |  |
| --- | --- |
| Σετ δεδομένων: data30 | Σετ δεδομένων: data60 |
|  |  |
| **Εικόνα 27** - Υλοποίηση SVD - Διάγραμμα μετρικών rmse και mae ανά τιμή χαρακτηριστικών – data30 | **Εικόνα 28** - Υλοποίηση SVD - Διάγραμμα μετρικών rmse και mae ανά τιμή χαρακτηριστικών – data30 |

Σετ δεδομένων: movielens 100K



**Εικόνα 29** - Υλοποίηση SVD - Διάγραμμα μετρικών rmse και mae ανά τιμή χαρακτηριστικών – movielens 100K

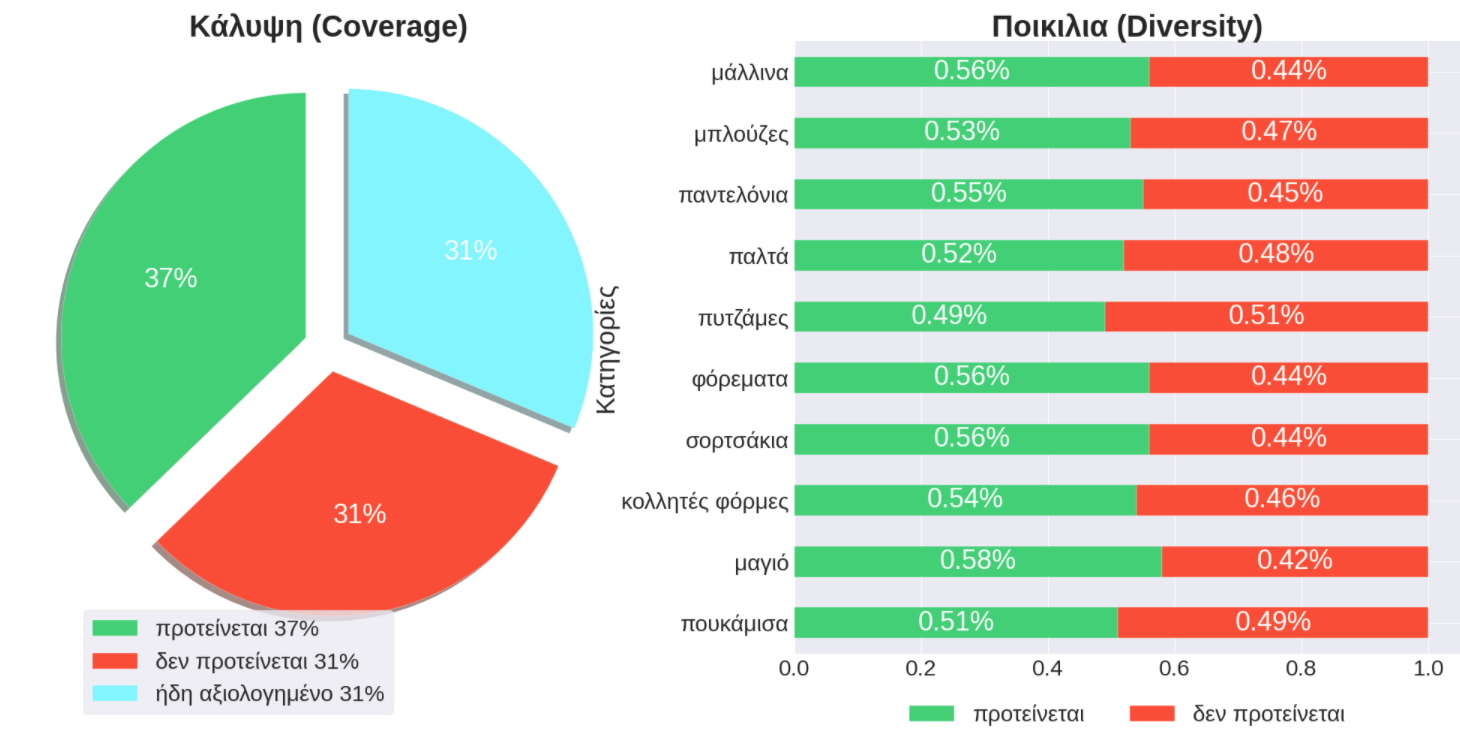
### **Κάλυψη και Ποικιλία**

Για το σετ δεδομένων data30 φαίνεται ότι από τα μη αξιολογημένα ρούχα βαθμολογήθηκαν θετικά το 37% και αρνητικά μόλις το 31% συνολικά για όλους τους χρήστες στο σύστημα *(Εικόνα 30).* Για το σετ δεδομένων data60 βαθμολογήθηκαν θετικά το 21% και αρνητικά το 18% *(Εικόνα 31).* Τέλος, για το σετ δεδομένων movielens-100k βαθμολογήθηκαν θετικά και αρνητικά 49% συνολικά *(Εικόνα 32).*

Οι κατηγορίες των ρούχων στο σύνολο δεδομένων data30 δείχνουν μια ισορροπία στις προτάσεις, δηλαδή δεν απορρίπτουν σε μεγάλο ποσοστό πολλά ρούχα από μια συγκεκριμένη κατηγορία. Όλες οι κατηγορίες φαίνονται να προτείνονται σε ποσοστά μεγαλύτερα του 49%. Τα ίδια ακριβώς ισχύουν και για το σύνολο data60. Τέλος, για το movielens-100k υπάρχει πάλι ισορροπία στις προτάσεις, σε μικρότερο βαθμό, ωστόσο κάθε κατηγορία προτείνεται με ποσοστά μεγαλύτερα του 47%.

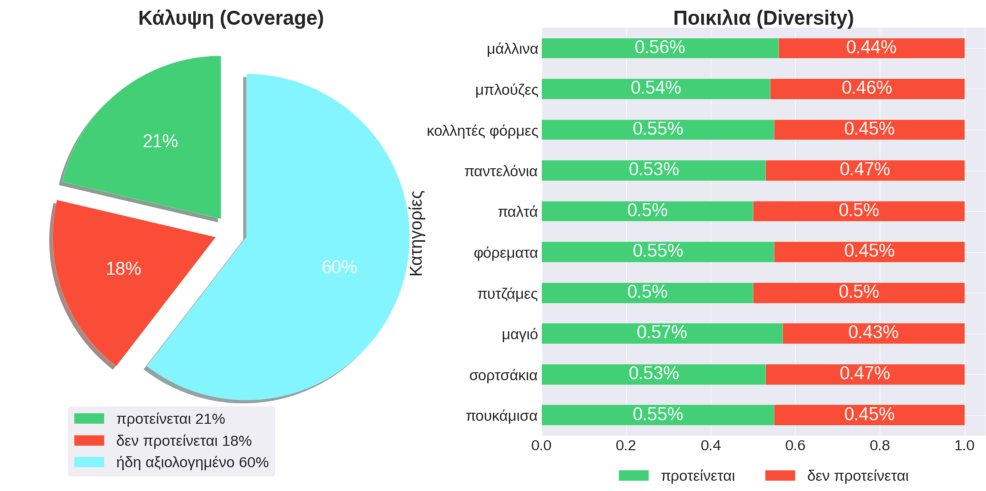
Συνοψίζοντας, τα διαγράμματα Κάλυψης και Ποικιλίας δείχνουν πως η ικανότητα ανακάλυψης των προϊόντων της βάσης είναι ικανοποιητική για όλα τα σύνολα. Επιπλέον, υπάρχει ισορροπία στις προτάσεις ανά κατηγορία, δηλαδή το σύστημα δεν απορρίπτει μεγάλο εύρος ρούχων/ταινιών κάποιας κατηγορίας. Αυτό θα είχε ως αποτέλεσμα να μην γινόταν σωστά η ανακάλυψη αυτής της κατηγορίας.

Σετ δεδομένων: data30



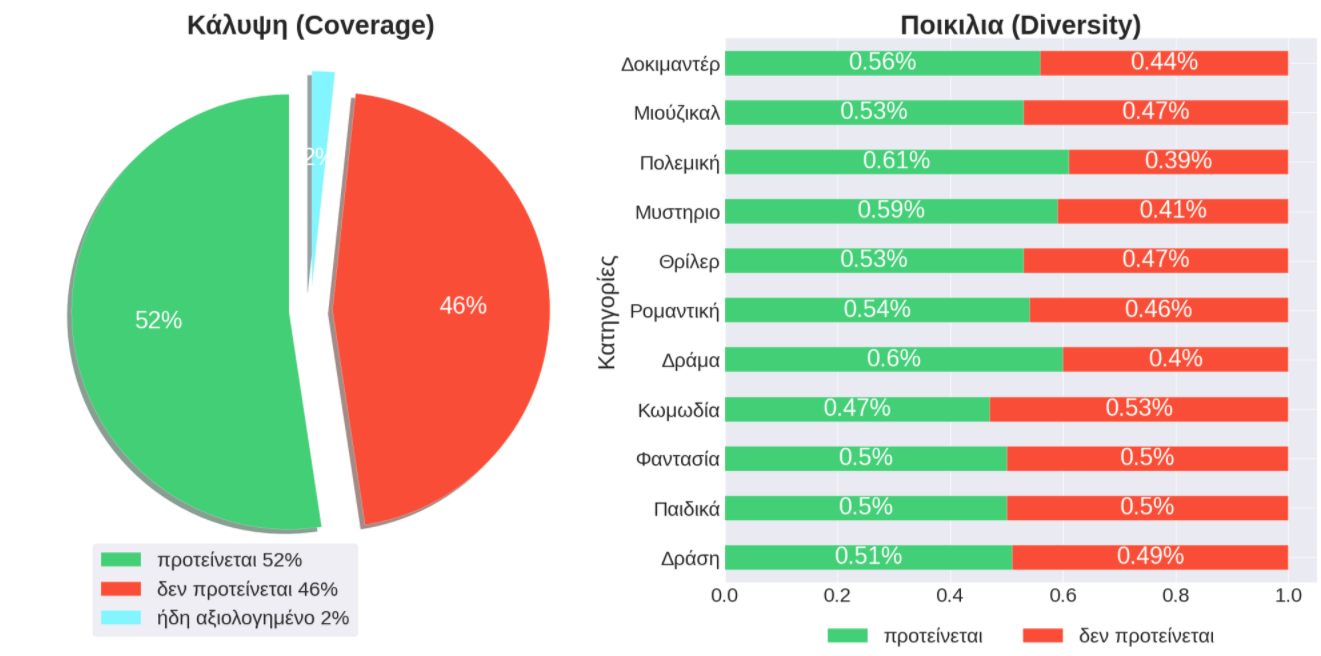
**Εικόνα 30** - Υλοποίηση SVD - Διάγραμμα Κάλυψης και Ποικιλίας – data30

Σετ δεδομένων: data60



**Εικόνα 31** - Υλοποίηση SVD - Διάγραμμα Κάλυψης και Ποικιλίας – data60

Σετ δεδομένων: movielens-100k



**Εικόνα 32** - 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 καλύτερη εκπαίδευση του συνεργατικού νευρωνικού δικτύου. Στη συνέχεια 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**

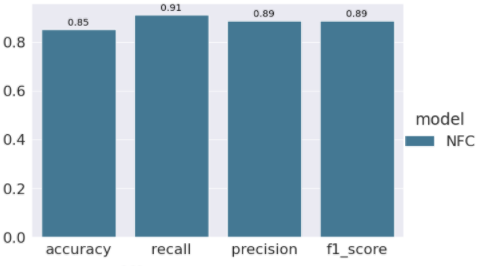
Το διάγραμμα των τεσσάρων μετρικών για όλα τα σύνολα δείχνει ικανοποιητικά ποσοστά accuracy, recall και precision. Ένα πλεονέκτημα αυτού του συστήματος είναι ότι προσαρμόζει με αυτοματοποιημένο τρόπο τις παραμέτρους του μοντέλου, διαδικασία που είναι κομμάτι των νευρωνικών δικτύων.

Βλέπουμε για το σύνολο δεδομένων data30 ότι accuracy-precision φτάνουν μόλις το 78%, recall στο 100%, ενώ το f1\_score 87% *(Εικόνα 33).* Το σύνολο δεδομένων data60 έχει accuracy και precision φτάνει μόλις στο 83%, recall 100%, και το f1\_score 90% *(Εικόνα 34).* Τέλος το σύνολο movielens 100K έχει accuracy 85%, precision και f1\_score 89% ενώ το recall 91% *(Εικόνα 35).*

Το συμπέρασμα που προκύπτει είναι ότι για τα σύνολα δεδομένων data30 και data60 (με την αύξηση 30% των κριτικών τους) έχουν 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 και data60 πετυχαίνει καλύτερα ποσοστά accuracy και precision το οποίο σημαίνει ότι έχει περισσότερη ακρίβεια στις προτάσεις του, έστω κι αν δεν τις βρίσκει όλες.

|  |  |
| --- | --- |
| Σετ δεδομένων: data30 | Σετ δεδομένων: data60 |
|  |  |
| **Εικόνα 33 -** Υλοποίηση NCF - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 | **Εικόνα 34** - Υλοποίηση NCF - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data60 |

Σετ δεδομένων: movielens 100K



**Εικόνα 35** - Υλοποίηση NCF - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K

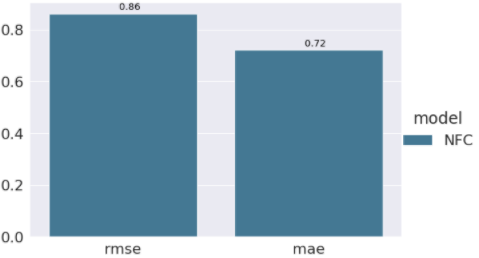
### **Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE)**

Για το σύνολο δεδομένων data30, η ρίζα μέσου τετραγωνικού σφάλματος είναι 2.75 και το μέσο απόλυτο σφάλμα 2.41 (Εικόνα 36). Το σύνολο δεδομένων data60 έχει ρίζα μέσου τετραγωνικού σφάλματος κυμαίνεται 2.47 και το μέσο απόλυτο σφάλμα από 2.04 *(Εικόνα 37).* Τέλος, το σύνολο movielens 100K έχει ρίζα μέσου τετραγωνικού σφάλματος 0.86 και μέσο απόλυτο σφάλμα 0.72 *(Εικόνα 38).*

Τα συμπέρασμα που προκύπτει από τις επιδόσεις των συνόλων data30 και data60 είναι ότι με την αύξηση 30% των κριτικών προκαλούνται θετικέςεπιδράσεις στα ζητήματα σφάλματος, αφού έχουμε βελτίωση του ελάχιστου τετραγωνικού σφάλματος από 2.75 σε 2.47 και του μέσου απόλυτου σφάλματος από 2.41 σε 2.04.

|  |  |
| --- | --- |
| Σετ δεδομένων: data30 | Σετ δεδομένων: data60 |
|  |  |
| **Εικόνα 36** - Υλοποίηση NCF - Διάγραμμα μετρικών rmse και mae – data30 | **Εικόνα 37** - Υλοποίηση NCF - Διάγραμμα μετρικών rmse και mae – data60 |

Σετ δεδομένων: movielens 100K



**Εικόνα 38** - Υλοποίηση NCF - Διάγραμμα μετρικών rmse και mae – movielens 100K

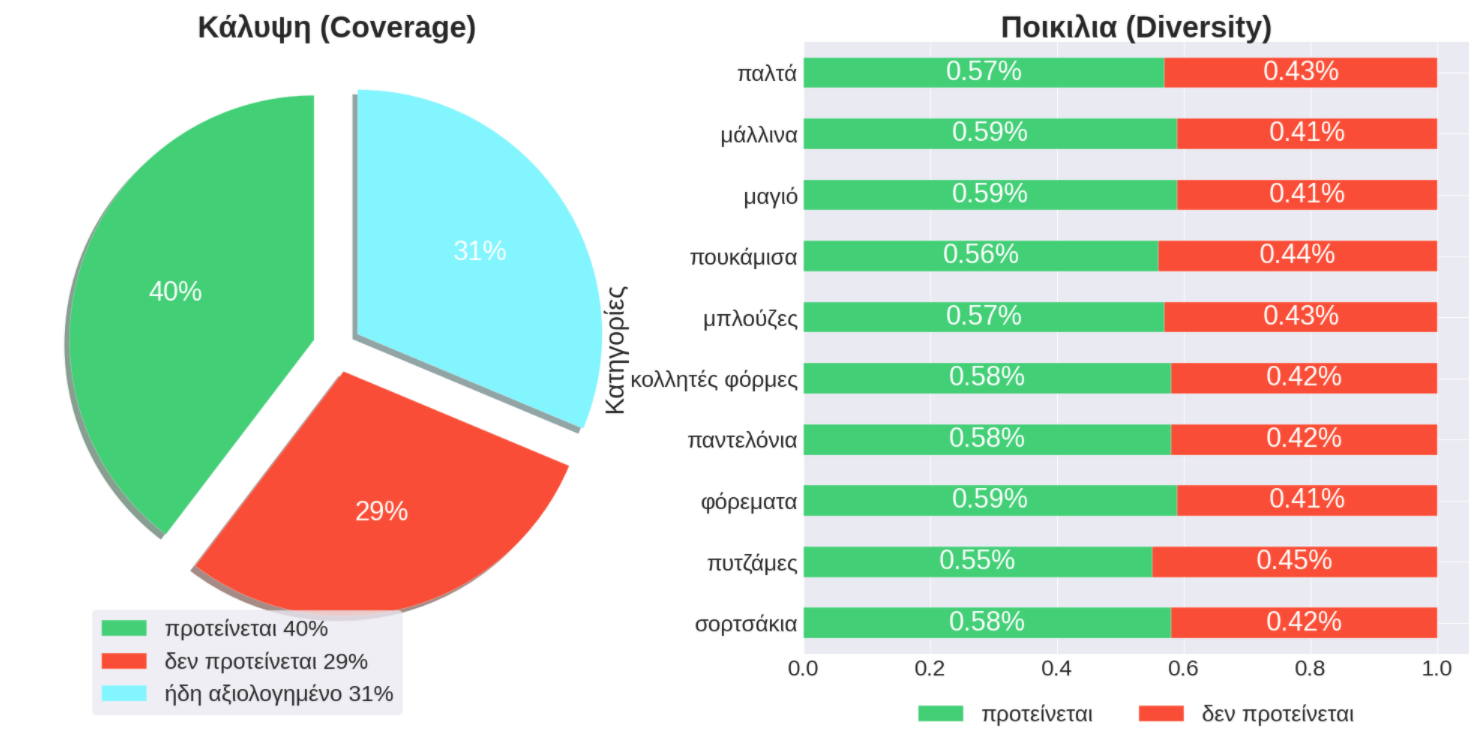
### **Κάλυψη και Ποικιλία**

Για το σετ δεδομένων data30 φαίνεται ότι από τα μη αξιολογημένα ρούχα βαθμολογήθηκαν θετικά το 40% και αρνητικά μόλις το 29% συνολικά για όλους τους χρήστες στο σύστημα *(Εικόνα 39).* Για το σετ δεδομένων data60 βαθμολογήθηκαν θετικά το 21% και αρνητικά το 19% *(Εικόνα 40).* Τέλος για το σετ δεδομένων movielens-100k βαθμολογήθηκαν θετικά 44% και αρνητικά 56% *(Εικόνα 41).*

Οι κατηγορίες των ρούχων στο σύνολο δεδομένων data30 δείχνουν μια ισορροπία στις προτάσεις, δηλαδή δεν απορρίπτουν σε μεγάλο ποσοστό πολλά ρούχα από μια συγκεκριμένη κατηγορία. Όλες οι κατηγορίες φαίνονται να προτείνονται σε ποσοστά μεγαλύτερα του 55%. Τα ίδια ακριβώς ισχύουν και για το σύνολο data60 σε ποσοστό λίγο χαμηλότερο, κοντά στο 50%. Τέλος, για το movielens-100k υπάρχει πάλι μια ισορροπία στις προτάσεις, αλλά όχι στον ίδιο βαθμό με τα δύο προηγούμενα σετ, ωστόσο κάθε κατηγορία προτείνεται με ποσοστά μεγαλύτερα του 39%.

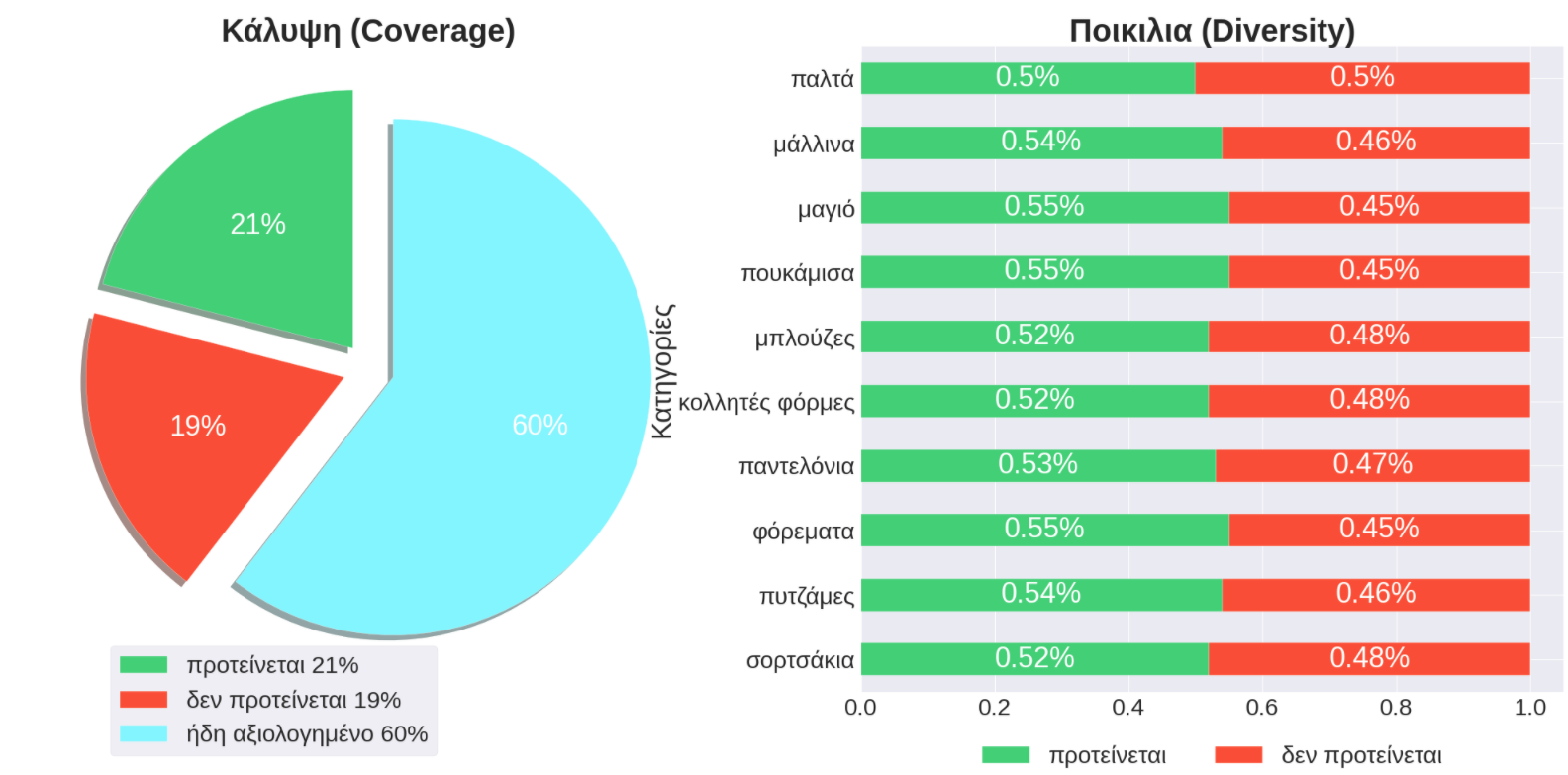
Τα διαγράμματα Κάλυψης και Ποικιλίας δείχνουν πως η ικανότητα ανακάλυψης των προϊόντων της βάσης είναι ικανοποιητική για όλα τα σύνολα. Επίσης υπάρχει ισορροπία στις προτάσεις που υπάρχουν ανά κατηγορία, δηλαδή το σύστημα δεν απορρίπτει μεγάλο εύρος ρούχων/ταινιών κάποιας κατηγορίας. Αυτό θα είχες ως αποτέλεσμα να μην γινόταν σωστά η ανακάλυψη αυτής της κατηγορίας.

Σετ δεδομένων: data30



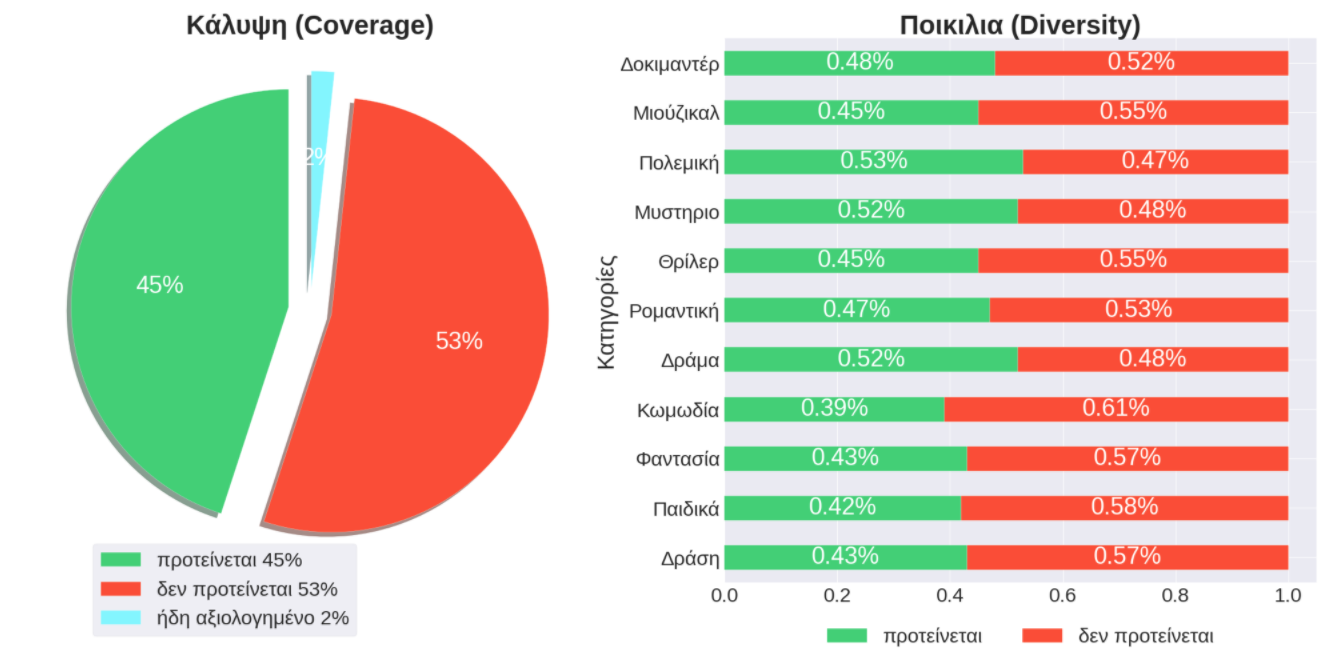
**Εικόνα 39** - Υλοποίηση NCF - Διάγραμμα Κάλυψης και Ποικιλίας – data30

Σετ δεδομένων: data60



**Εικόνα 40** - Υλοποίηση NCF - Διάγραμμα Κάλυψης και Ποικιλίας – data60

Σετ δεδομένων: movielens-100k



**Εικόνα 41** - 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 καλύτερη εκπαίδευση του Τυχαίου Δάσους μεταβάλλοντας τον αριθμό των εκτιμητών και το μέγιστο βάθος κάθε δένδρου ταξινόμησης. Στην συνέχεια, το εκπαιδευμένο μοντέλο εφαρμόζεται στο σετ ελέγχου. Τέλος, δημιουργούνται προτάσεις, οι οποίες συγκρίνονται με τις πραγματικές που υπάρχουν στο σετ ελέγχου, και αξιολογείται η ικανότητα εξαγωγής προτάσεων του μοντέλου.

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

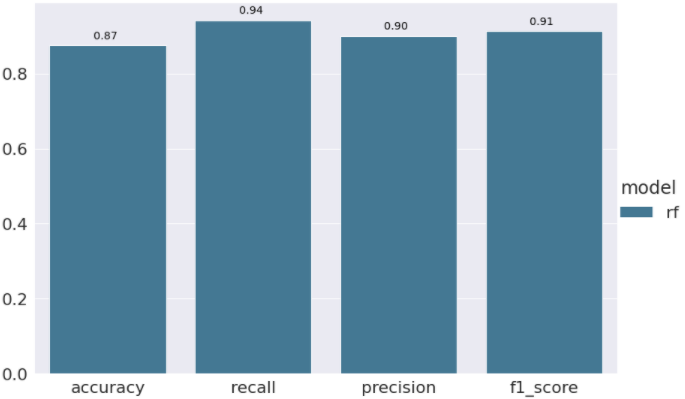
Το διάγραμμα των τεσσάρων μετρικών για όλα τα σύνολα δείχνει ικανοποιητικά ποσοστά accuracy, recall και precision. Ένα πλεονέκτημα της υλοποίησης είναι ότι προσαρμόζει με αυτοματοποιημένο τις παραμέτρους των Τυχαίων Δασών με τη χρήση της βιβλιοθήκης Optuna.

Βλέπουμε για το σύνολο δεδομένων data30 ότι accuracy-precision φτάνουν μόλις το 78%, recall στο 96%, ενώ το f1\_score 86% *(Εικόνα 42).* Το σύνολο δεδομένων data60 έχει accuracy και precision φτάνει μόλις στο 80%, recall 97%, και το f1\_score 87% *(Εικόνα 43).* Τέλος, το σύνολο movielens 100K έχει accuracy 87%, precision και f1\_score 90% με 91%, ενώ το recall 94% *(Εικόνα 44).*

Ένα συμπέρασμα που προκύπτει εκ του αποτελέσματος είναι ότι τα σύνολα δεδομένων data30 και data60 έχουν ανεπαίσθητες επιδράσεις στα ζητήματα ταξινόμησης, αφού το accuracy απ’ το 77% στο 80%, το recall από το 96% στο 97% αλλαγές πολύ μικρές. Από την άλλη, το σύνολο δεδομένων movielens-100k παρόλο που δεν έχει το ίδιο υψηλό recall σε σχέση με τα data30 και data60 πετυχαίνει καλύτερα ποσοστά accuracy και precision. Αυτό σημαίνει ότι έχει περισσότερη ακρίβεια στις προτάσεις αρεσκείας των χρηστών, ακόμα κι αν δεν τις βρίσκει όλες.

|  |  |
| --- | --- |
| Σετ δεδομένων: data30 | Σετ δεδομένων: data60 |
|  |  |
| **Εικόνα 42** - Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 | **Εικόνα 43** - Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data60 |

Σετ δεδομένων: movielens 100K



**Εικόνα 44 -** Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K

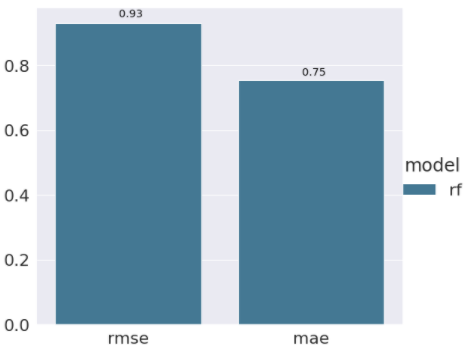
### **Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE)**

Για το σύνολο δεδομένων data30 η ρίζα μέσου τετραγωνικού σφάλματος είναι 2.35 και το μέσο απόλυτο σφάλμα 1.85 (*Εικόνα 45)*. Το σύνολο δεδομένων data60 έχει ρίζα μέσου τετραγωνικού σφάλματος κυμαίνεται 2.13 και το μέσο απόλυτο σφάλμα από 1.64 *(Εικόνα 46).* Τέλος, το σύνολο movielens 100K έχει ρίζα μέσου τετραγωνικού σφάλματος 0.93 και μέσο απόλυτο σφάλμα 0.75 *(Εικόνα 47).*

Τα συμπέρασμα που προκύπτει από τις επιδόσεις των συνόλων data30 και data60 είναι ότι με την αύξηση 30% των κριτικών προκαλούνται θετικέςεπιδράσεις στα ζητήματα σφάλματος αφού έχουμε βελτίωση του ελάχιστου τετραγωνικού σφάλματος από 2.35 σε 2.14 και του μέσου απόλυτου σφάλματος από 1.85 σε 1.65.

|  |  |
| --- | --- |
| Σετ δεδομένων: data30 | Σετ δεδομένων: data60 |
|  |  |
| **Εικόνα 45** - Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών rmse και mae – data30 | **Εικόνα 46** - Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών rmse και mae – data60 |

Σετ δεδομένων: movielens 100K



**Εικόνα 47** - Υλοποίηση Τυχαία Δάση - Διάγραμμα μετρικών rmse και mae – movielens 100K

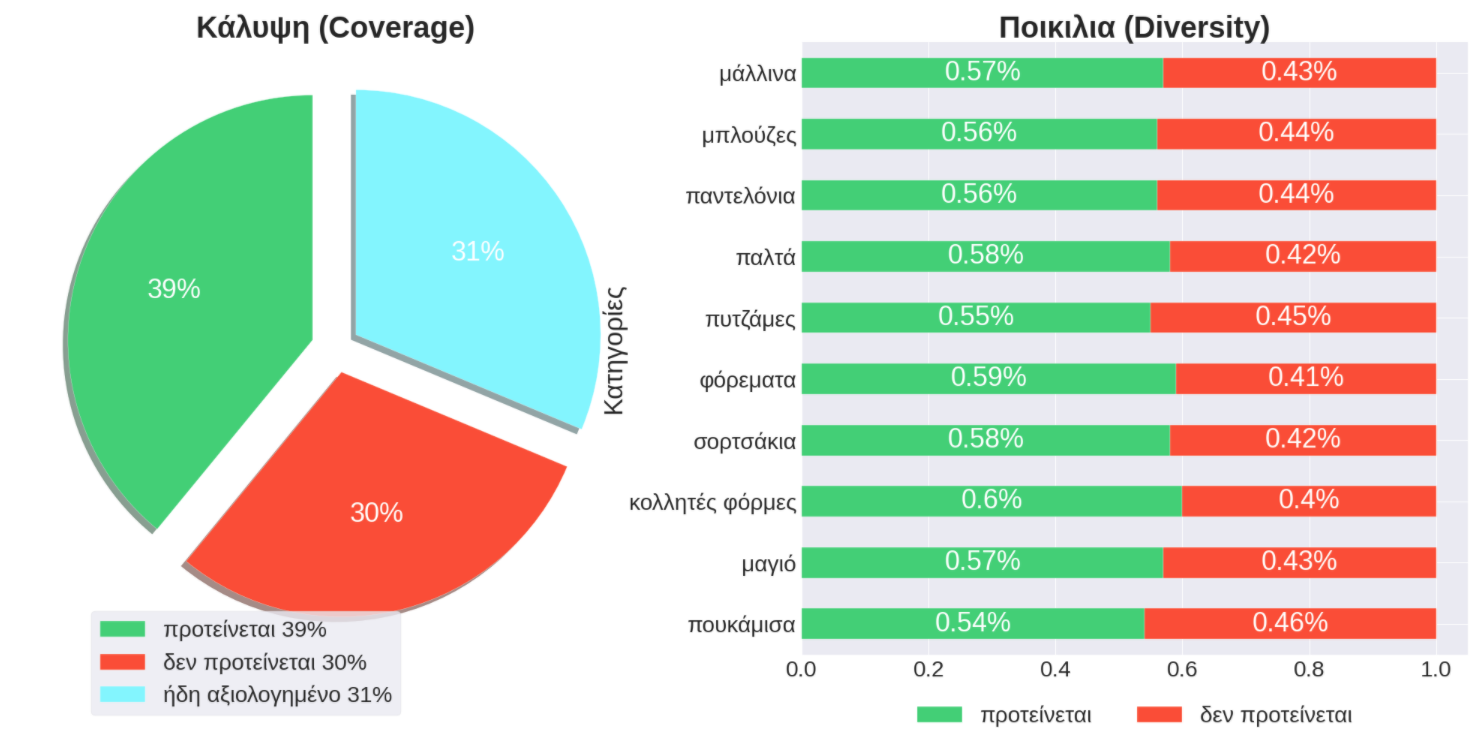
### **Κάλυψη και Ποικιλία**

Για το σετ δεδομένων data30 φαίνεται ότι από τα μη αξιολογημένα ρούχα βαθμολογήθηκαν θετικά το 39% και αρνητικά μόλις το 30% συνολικά για όλους τους χρήστες στο σύστημα *(Εικόνα 48).* Για το σετ δεδομένων data60 βαθμολογήθηκαν θετικά το 22% και αρνητικά το 18% *(Εικόνα 49).* Τέλος για το σετ δεδομένων movielens-100k βαθμολογήθηκαν θετικά 47% και αρνητικά 51% *(Εικόνα 50).*

Οι κατηγορίες των ρούχων στο σύνολο δεδομένων data30 δείχνουν μια ισορροπία στις προτάσεις, δηλαδή δεν απορρίπτουν σε μεγάλο ποσοστό πολλά ρούχα από μια συγκεκριμένη κατηγορία. Όλες οι κατηγορίες φαίνονται να προτείνονται σε ποσοστά μεγαλύτερα του 54%. Τα ίδια ακριβώς ισχύουν και για το σύνολο data60 σε ποσοστό πάλι 54%. Τέλος, για το movielens-100k υπάρχει πάλι μια ισορροπία στις προτάσεις, αλλά όχι στον ίδιο βαθμό με τα δύο προηγούμενα σετ. Ωστόσο, κάθε κατηγορία προτείνεται με ποσοστά μεγαλύτερα του 35% με 40%.

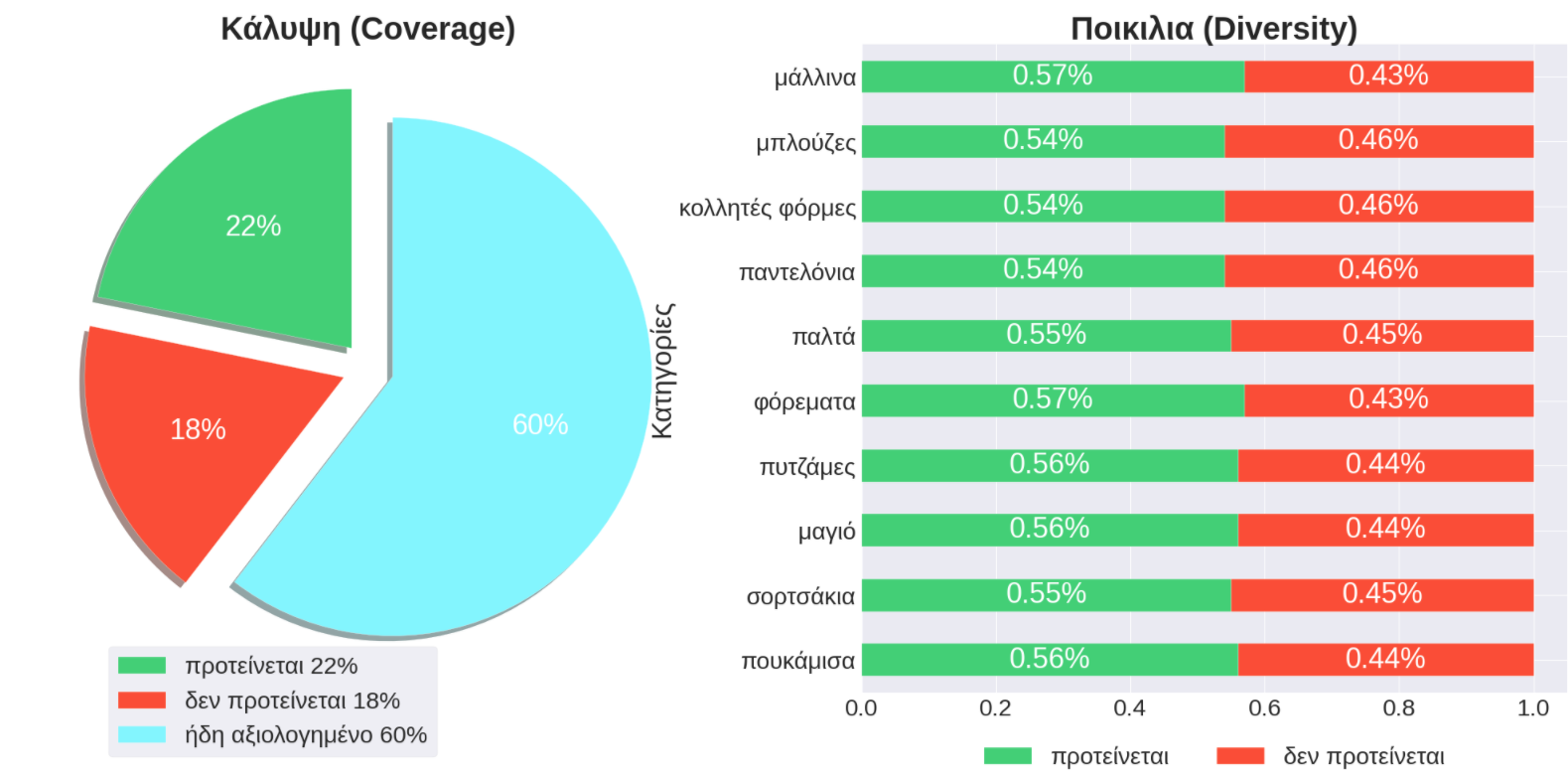
Τα διαγράμματα Κάλυψης και Ποικιλίας δείχνουν πως η ικανότητα ανακάλυψης των προϊόντων της βάσης είναι ικανοποιητική για όλα τα σύνολα. Επίσης, υπάρχει ισορροπία στις προτάσεις που υπάρχουν ανά κατηγορία, δηλαδή το σύστημα δεν απορρίπτει μεγάλο εύρος ρούχων/ταινιών κάποιας κατηγορίας. Αυτό θα είχε ως αποτέλεσμα να μην γινόταν σωστά η ανακάλυψη αυτής της κατηγορίας.

Σετ δεδομένων: data30



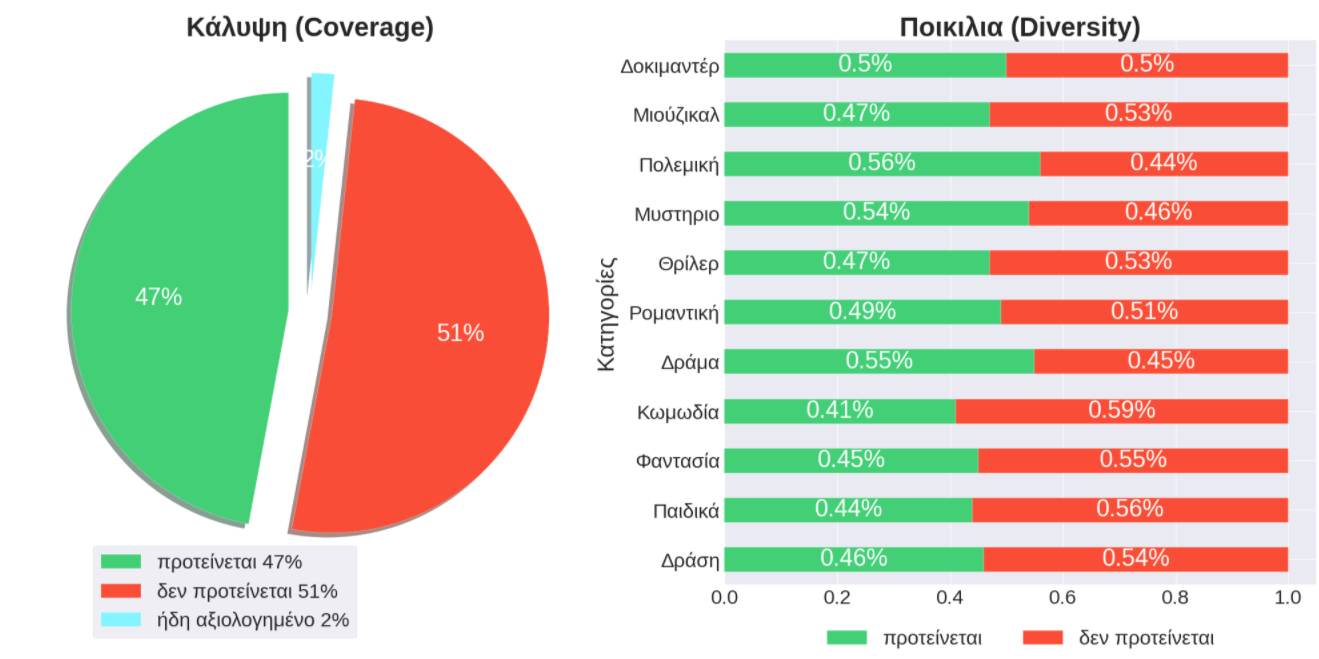
**Εικόνα 48** - Υλοποίηση Τυχαίων Δασών - Διάγραμμα Κάλυψης και Ποικιλίας – data30

Σετ δεδομένων: data60



**Εικόνα 49** - Υλοποίηση Τυχαίων Δασών - Διάγραμμα Κάλυψης και Ποικιλίας – data60

Σετ δεδομένων: movielens-100k



**Εικόνα 50** - 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 καλύτερη εκπαίδευση του νευρωνικού φιλτραρίσματος περιεχομένου. Στην συνέχεια δημιουργούνται προτάσεις, οι οποίες συγκρίνονται με τις πραγματικές τιμές που υπάρχουν στο σετ ελέγχου, και αξιολογείται η ικανότητα εξαγωγής προτάσεων του μοντέλου.

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

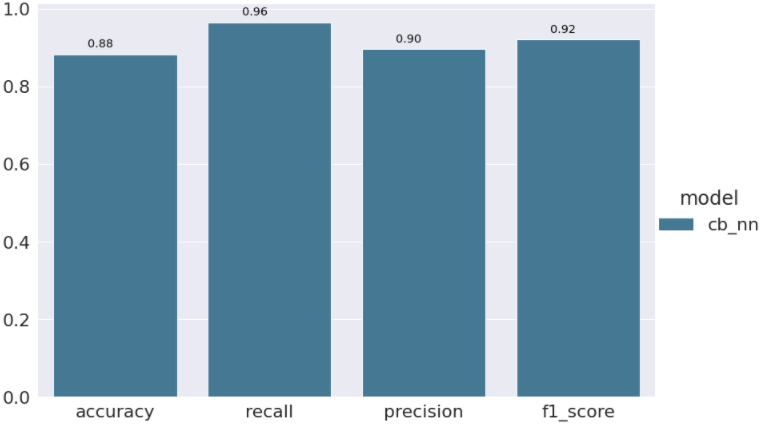
Το διάγραμμα των τεσσάρων μετρικών για όλα τα σύνολα δείχνει ικανοποιητικά ποσοστά accuracy, recall και precision. Ένα πλεονέκτημα αυτής της υλοποίησης είναι ότι προσαρμόζει με αυτοματοποιημένο τρόπο τα βάρη κάθε Νευρωνικού Δικτύου Περιεχομένου.

Βλέπουμε για το σύνολο δεδομένων data30 ότι accuracy-precision φτάνουν μόλις τα 76% κ 81%, recall στο 89%, ενώ το f1\_score 84% *(Εικόνα 51).* Το σύνολο δεδομένων data60 έχει accuracy στο 81% precision στο 84%, recall 91%, και το f1\_score 87% *(Εικόνα 52).* Τέλος το σύνολο movielens 100K έχει accuracy 88%, precision 90%, f1\_score 92% και recall 96% *(Εικόνα 53).*

Ένα συμπέρασμα που προκύπτει από τα σύνολα δεδομένων data30 και data60 είναι ότι με την αύξηση 30% των κριτικών τους έχουμε μικρές θετικές επιδράσεις στα ζητήματα ταξινόμησης, αφού το accuracy απ’ το 76% στο 80%, το recall από το 89% στο 90%. Από την άλλη το σύνολο δεδομένων movielens-100k έχει υψηλό recall σε σχέση με data30, data60 και πετυχαίνει καλύτερα ποσοστά accuracy και precision.

|  |  |
| --- | --- |
| Σετ δεδομένων: data30 | Σετ δεδομένων: data60 |
|  |  |
| **Εικόνα 51** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 | **Εικόνα 52** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30 |

Σετ δεδομένων: movielens 100K



**Εικόνα 53** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K

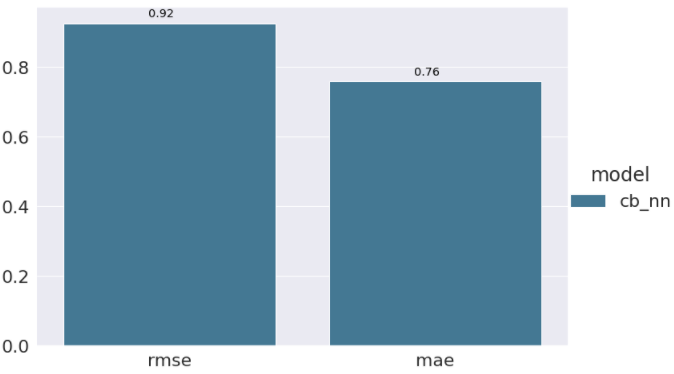
### **Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE)**

Για το σύνολο δεδομένων data30 η ρίζα μέσου τετραγωνικού σφάλματος είναι 2.49 και το μέσο απόλυτο σφάλμα 1.88 *(Εικόνα 54).* Το σύνολο δεδομένων data60 έχει ρίζα μέσου τετραγωνικού σφάλματος είναι 2.21 και το μέσο απόλυτο σφάλμα από 1.58 (Εικόνα 55). Τέλος το σύνολο movielens 100K έχει ρίζα μέσου τετραγωνικού σφάλματος 0.92 και μέσο απόλυτο σφάλμα 0.76 *(Εικόνα 56).*

Τα συμπέρασμα που προκύπτει από τις επιδόσεις των συνόλων data30 και data60 είναι ότι με την αύξηση 30% των κριτικών προκαλούνται θετικέςεπιδράσεις στα ζητήματα σφάλματος, αφού έχουμε βελτίωση του ελάχιστου τετραγωνικού σφάλματος από 2.49 σε 2.21 και του μέσου απόλυτου σφάλματος από 1.88 σε 1.58.

|  |  |
| --- | --- |
| Σετ δεδομένων: data30 | Σετ δεδομένων: data60 |
|  |  |
| **Εικόνα 54** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών rmse και mae – data30 | **Εικόνα 55** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών rmse και mae – data60 |

Σετ δεδομένων: movielens 100K



**Εικόνα 56** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα μετρικών rmse και mae – movielens 100K

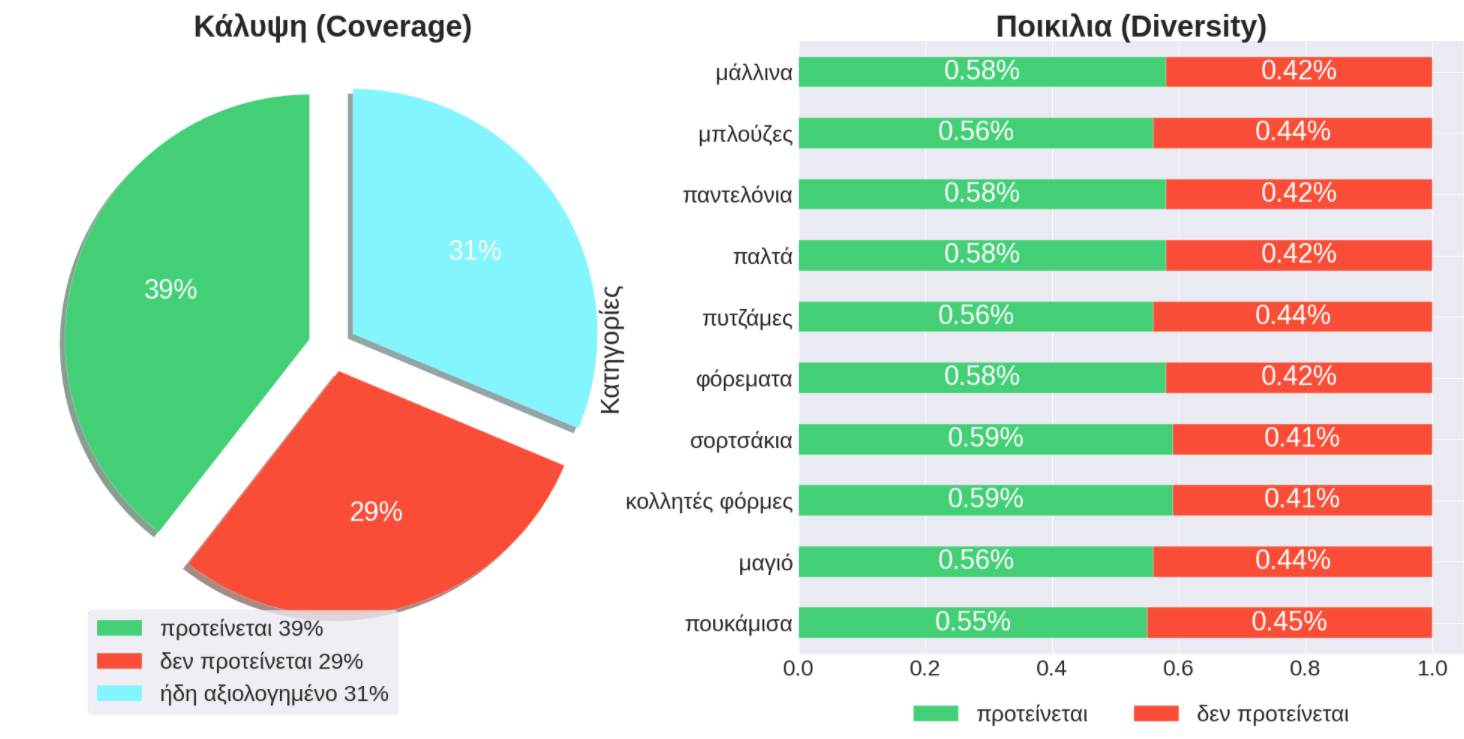
### **Κάλυψη και Ποικιλία**

Για το σετ δεδομένων data30 φαίνεται ότι από τα μη αξιολογημένα ρούχα βαθμολογήθηκαν θετικά το 40% και αρνητικά μόλις το 30% συνολικά για όλους τους χρήστες στο σύστημα *(Εικόνα 57).* Για το σετ δεδομένων data60 βαθμολογήθηκαν θετικά το 23% και αρνητικά το 17% *(Εικόνα 58).* Τέλος για το σετ δεδομένων movielens-100k βαθμολογήθηκαν θετικά 56% και αρνητικά 42% *(Εικόνα 59).*

Οι κατηγορίες των ρούχων στο σύνολο δεδομένων data30 δείχνουν μια ισορροπία στις προτάσεις, δηλαδή δεν απορρίπτουν σε μεγάλο ποσοστό πολλά ρούχα από μια συγκεκριμένη κατηγορία. Όλες οι κατηγορίες φαίνονται να προτείνονται σε ποσοστά μεγαλύτερα του 55%. Τα ίδια ακριβώς ισχύουν και για το σύνολο data60 σε ποσοστό 57%. Τέλος, για το movielens-100k υπάρχει πάλι μια ισορροπία στις προτάσεις αλλά όχι στον ίδιο βαθμό με τα δύο προηγούμενα σετ, πάλι κι εκεί ωστόσο κάθε κατηγορία προτείνεται με ποσοστά μεγαλύτερα του 48%.

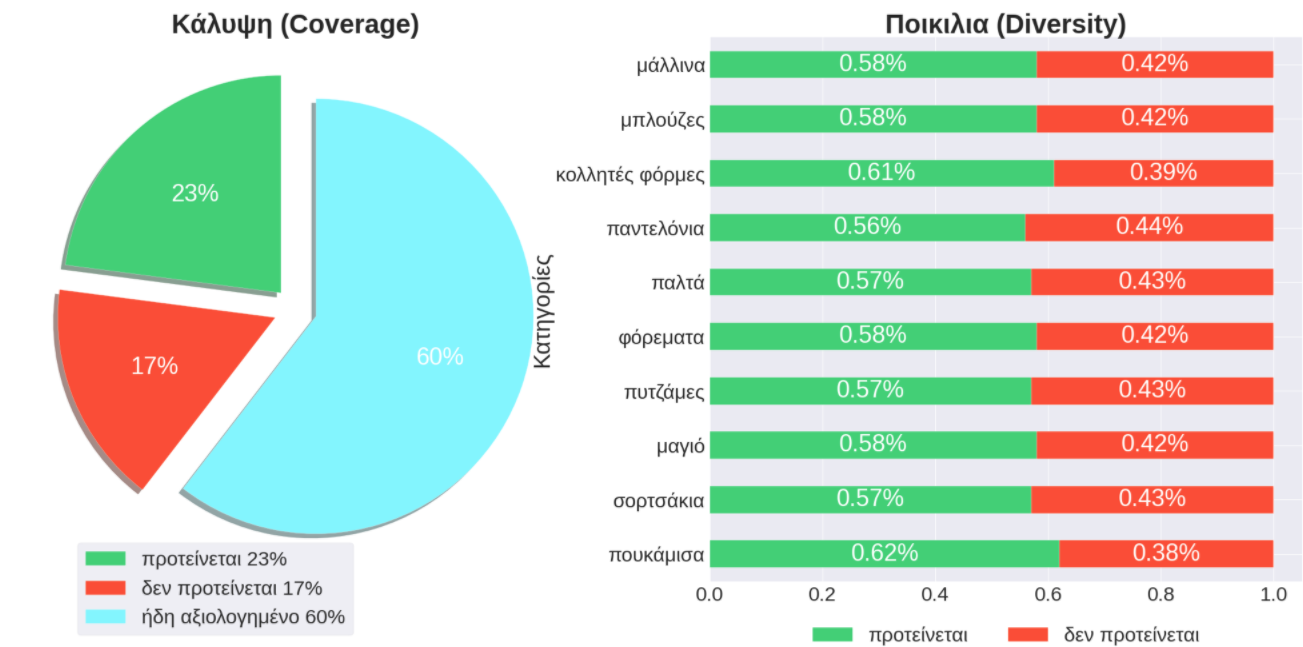
Συνολικά, τα διαγράμματα Κάλυψης και Ποικιλίας δείχνουν πως η ικανότητα ανακάλυψης των προϊόντων της βάσης είναι ικανοποιητική για όλα τα σύνολα. Επιπλέον, υπάρχει ισορροπία στις προτάσεις που υπάρχουν ανά κατηγορία, δηλαδή το σύστημα δεν απορρίπτει μεγάλο εύρος ρούχων/ταινιών κάποιας κατηγορίας. Αυτό θα είχε ως αποτέλεσμα να μην γινόταν σωστά η ανακάλυψη αυτής της κατηγορίας.

Σετ δεδομένων: data30



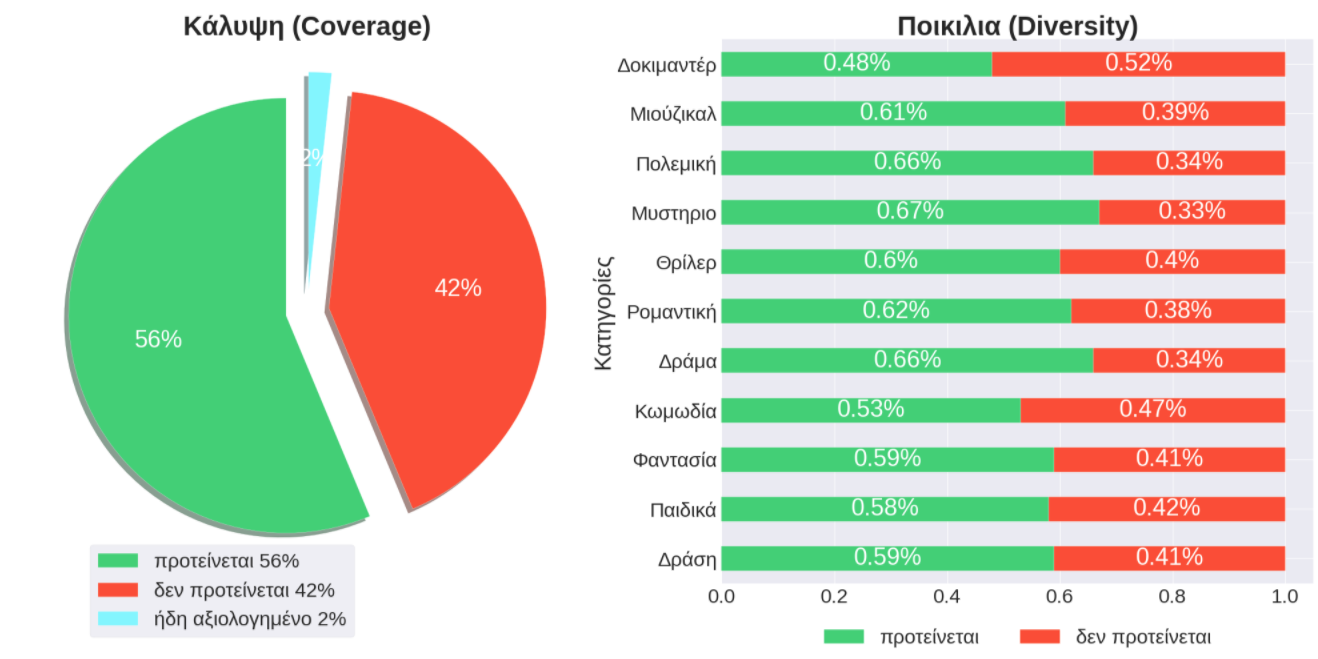
**Εικόνα 57** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα Κάλυψης και Ποικιλίας – data30

Σετ δεδομένων: data60



**Εικόνα 58** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα Κάλυψης και Ποικιλίας – data60

Σετ δεδομένων: movielens-100k



**Εικόνα 59** - Υλοποίηση Φιλτραρίσματος Περιεχομένου με Νευρωνικά Δίκτυα - Διάγραμμα Κάλυψης και Ποικιλίας – movielens-100k

## **Υλοποίηση Υβριδικού Συστήματος**

Η υλοποίηση του υβριδικού συστήματος χρησιμοποιεί όλα τα παραπάνω μοντέλα που παρουσιάστηκαν. Αφού εξαχθούν οι προτάσεις από κάθε μοντέλο, το υβριδικό σύστημα δημιουργεί δύο πακέτα προτάσεων. Το πρώτο πακέτο 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 πακέτο αφορά την εξαγωγή προτάσεων μέσω της χρήσης ενός μοντέλου γραμμικής παλινδρόμησης. Το 30% των αξιολογήσεων χρησιμοποιούνται για την εκπαίδευση του μοντέλου το οποίο με την σειρά του δημιουργεί προτάσεις για το υπόλοιπο 70%.

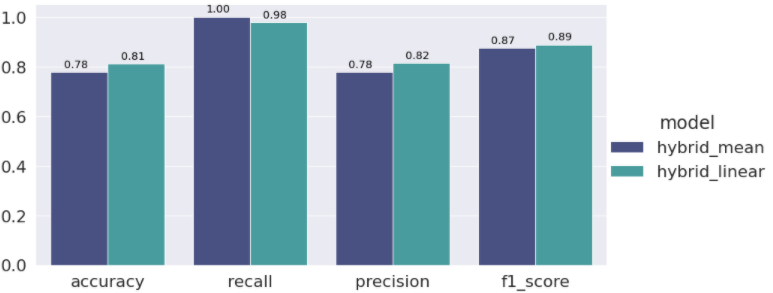
### **Accuracy, recall, precision και f1\_score**

Τα διαγράμματα των τεσσάρων μετρικών για όλα τα σύνολα δείχνει ικανοποιητικά ποσοστά accuracy, recall και precision.

Βλέπουμε για το σύνολο δεδομένων data30 το μοντέλο ίσης ψήφου έχει accuracy και precision φτάνουν μόλις στο 78%, recall στο 100%, και f1\_score 87%. Το μοντέλο γραμμικής παλινδρόμησης στο ίδιο σετ φαίνεται να τα πάει καλύτερα αφού έχει accuracy και precision στα 81% και 82% αντίστοιχα, recall στο 98% (λίγο χαμηλότερο από αυτό της ίσης ψήφου αλλά όχι σημαντικά), και f1\_score 89%. *(Εικόνα 60).* Παρόμοια ποσοστά προκύπτουν και για το σύνολο δεδομένων data60, το οποίο για το μοντέλο ίσης ψήφου έχει accuracy στο 80%, precision στο 79%, recall 100%, και το f1\_score 88%. Παρόλα αυτά, το μοντέλο γραμμικής παλινδρόμησης και στο δεύτερο σετ τα πάει καλύτερα από αυτό της ίσης ψήφου με ποσοστά accuracy και precision στο 84%, recall 97%, και το f1\_score 90%. *(Εικόνα 61).* Τέλος, στο σύνολο movielens 100K το μοντέλο ίσης ψήφου έχει accuracy στο 86%, precision στο 88%, recall 92%, και το f1\_score 89% με παραπλήσια ποσοστά ακολουθεί το μοντέλο γραμμικής παλινδρόμησης *(Εικόνα 62) .*

Ένα συμπέρασμα που προκύπτει είναι ότι για όλα τα σύνολα δεδομένων οι προτάσεις του υβριδικού μέσω γραμμικής παλινδρόμησης έχουν καλύτερη επίδοση από το υβριδικό σύστημα ίσης ψήφου. Ένα δεύτερο συμπέρασμα όσον αφορά τα σύνολα δεδομένων data30 και data60 (αναφορικά στο μοντέλο γραμμικής παλινδρόμησης), προκύπτει ότι η αύξηση 30% των κριτικών προκαλεί θετικές επιδράσεις στα ζητήματα ταξινόμησης αφού το accuracy αυξάνεται απ’ το 81% στο 84%, το precision από το 82% στο 84% και το f1\_score απ’ το 87% στο 89%, με μια ελαφρά μείωση στο recall. Από την άλλη πλευρά, το σύνολο δεδομένων movielens-100k κρατά και τις 4 μετρικές σε ποσοστά άνω του 90%

Σετ δεδομένων: data30



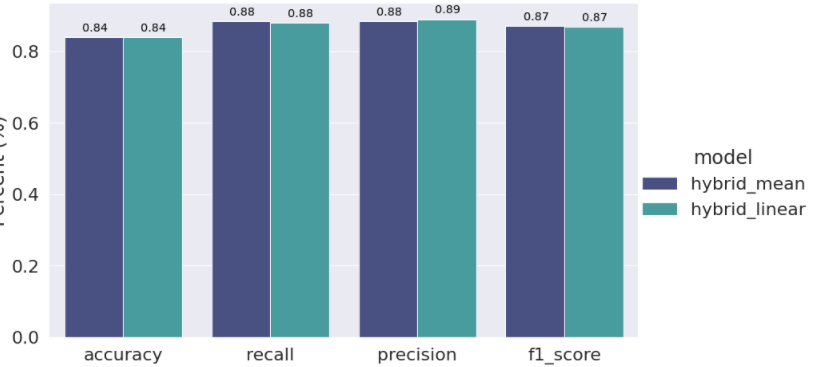
***Εικόνα 60 -*** Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30

Σετ δεδομένων: data60



**Εικόνα 61** - Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data60

Σετ δεδομένων: movielens 100K



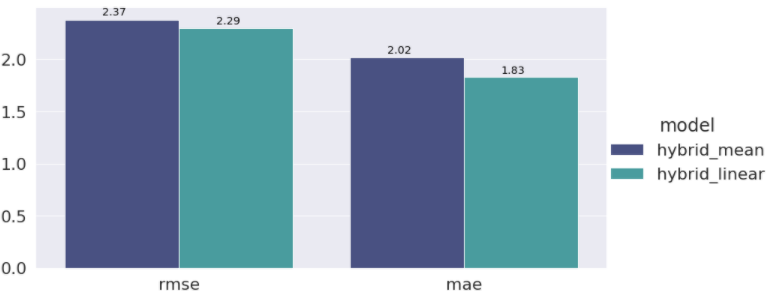
**Εικόνα 62** - Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K

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

Όσον αφορά το υβριδικό μοντέλο προτάσεων ίσης ψήφου για το σύνολο δεδομένων data30, η ρίζα μέσου τετραγωνικού σφάλματος είναι 2.37 και το μέσο απόλυτο σφάλμα 2.02, ενώ το υβριδικό μοντέλο γραμμικής παλινδρόμησης πετυχαίνει ρίζα μέσου τετραγωνικού σφάλματος 2.29 και μέσο απόλυτο σφάλμα 1.83 *(Εικόνα 63).* Στο σύνολο δεδομένων data60 το υβριδικό μοντέλο ίσης ψήφου πετυχαίνει ρίζα μέσου τετραγωνικού σφάλματος είναι 2.10 και μέσο απόλυτο σφάλμα από 1.74, ενώ το υβριδικό μοντέλο γραμμικής παλινδρόμησης πετυχαίνει ρίζα μέσου τετραγωνικού σφάλματος είναι 1.97 και μέσο απόλυτο σφάλμα από 1.52 *(Εικόνα 64).* Τέλος, όσον αφορά το σύνολο movielens 100K, το υβριδικό μοντέλο προτάσεων ίσης ψήφου έχει ρίζα μέσου τετραγωνικού σφάλματος 0.85 και μέσο απόλυτο σφάλμα 0.70, ενώ το υβριδικό μοντέλο γραμμικής παλινδρόμησης πετυχαίνει ρίζα μέσου τετραγωνικού σφάλματος είναι 0.82 και μέσο απόλυτο σφάλμα από 0.67 *(Εικόνα 65).*

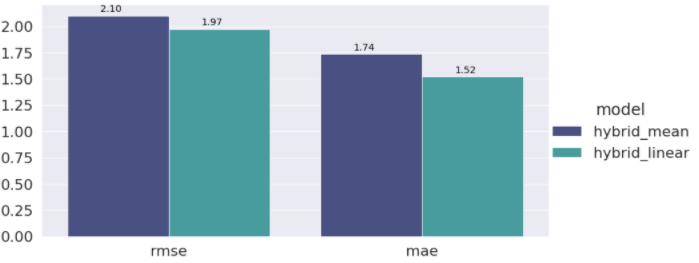
Ένα συμπέρασμα που προκύπτει παρατηρώντας το σφάλμα για όλα τα σύνολα δεδομένων είναι πως τελικά το υβριδικό σύστημα μέσω γραμμικής παλινδρόμησης πετυχαίνει μικρότερό σφάλμα του υβριδικό σύστημα ίσης ψήφου. Επιπλέον, από τις επιδόσεις των συνόλων data30 και data60 (αναφορικά με το υβριδικό μοντέλο γραμμικής παλινδρόμησης) φαίνεται ότι με την αύξηση 30% των κριτικών προκαλούνται θετικέςεπιδράσεις στα ζητήματα σφάλματος αφού έχουμε βελτίωση του ελάχιστου τετραγωνικού σφάλματος από 2.37 σε 1.99 και του μέσου απόλυτου σφάλματος από 1.83 σε 1.54.

Σετ δεδομένων: data30



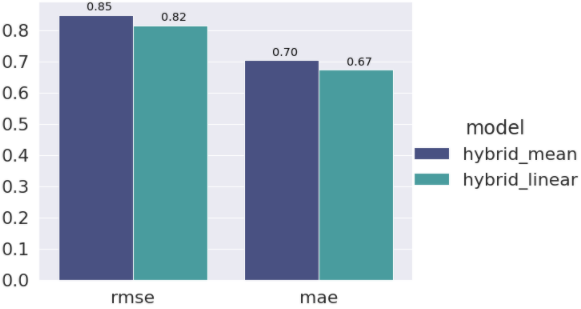
**Εικόνα 63** - Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – data30

Σετ δεδομένων: data60



**Εικόνα 64** - Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – data30

Σετ δεδομένων: movielens 100K



**Εικόνα 65** - Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – movielens 100K

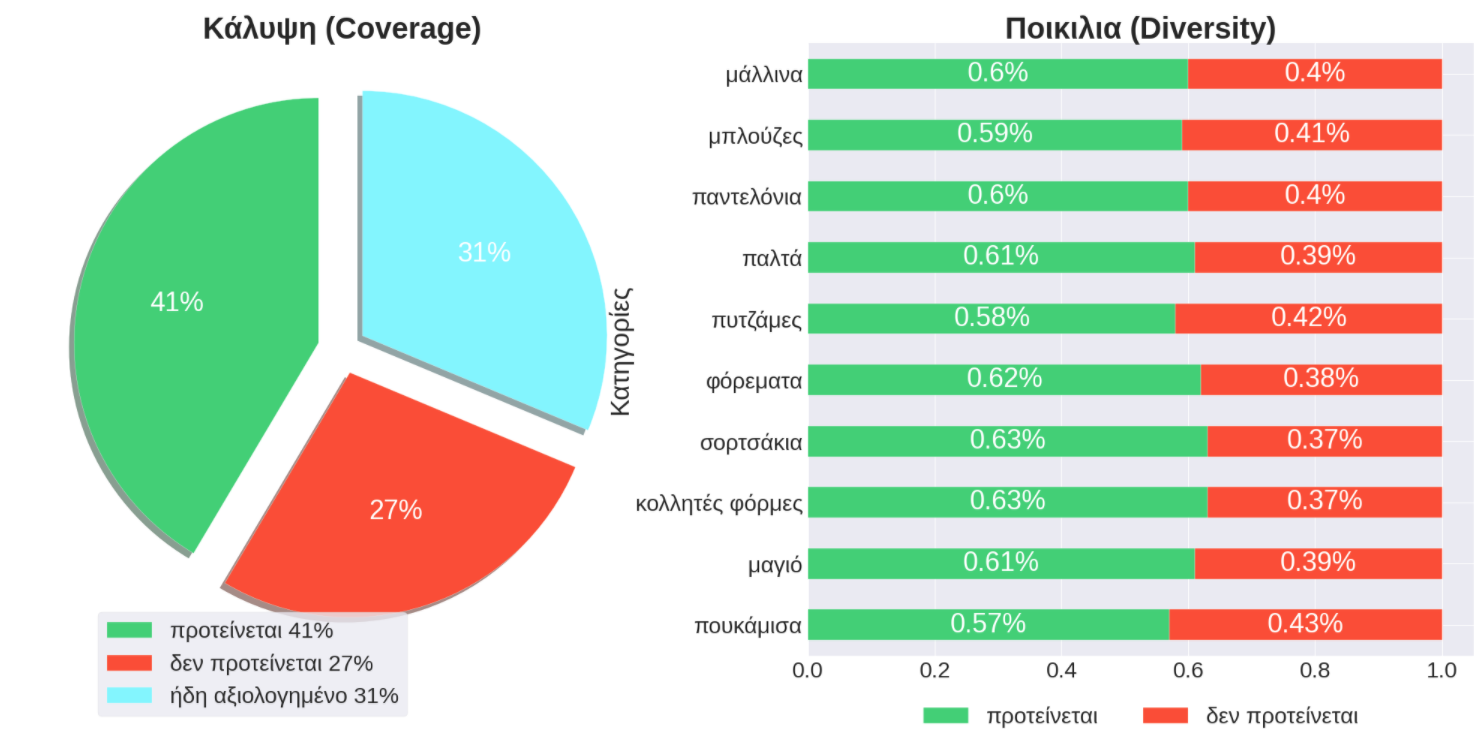
### **Κάλυψη και Ποικιλία**

Για το σετ δεδομένων data30 φαίνεται ότι από τα μη αξιολογημένα ρούχα βαθμολογήθηκαν θετικά το 41% και αρνητικά μόλις το 27% συνολικά για όλους τους χρήστες στο σύστημα *(Εικόνα 66).* Για το σετ δεδομένων data60 βαθμολογήθηκαν θετικά το 23% και αρνητικά το 17% *(Εικόνα 67).* Τέλος, για το σετ δεδομένων movielens-100k βαθμολογήθηκαν θετικά 47% και αρνητικά 51% *(Εικόνα 68).*

Οι κατηγορίες των ρούχων στο σύνολο δεδομένων data30 δείχνουν μια ισορροπία στις προτάσεις, δηλαδή δεν απορρίπτουν σε μεγάλο ποσοστό πολλά ρούχα από μια συγκεκριμένη κατηγορία. Όλες οι κατηγορίες φαίνονται να προτείνονται σε ποσοστά μεγαλύτερα του 57%. Τα ίδια ακριβώς ισχύουν και για το σύνολο data60 σε ποσοστό 56%. Τέλος, για το movielens-100k υπάρχει πάλι μια ισορροπία στις προτάσεις, αλλά όχι στον ίδιο βαθμό με τα δύο προηγούμενα σετ, ωστόσο κάθε κατηγορία προτείνεται με ποσοστά μεγαλύτερα του 45% με 50%.

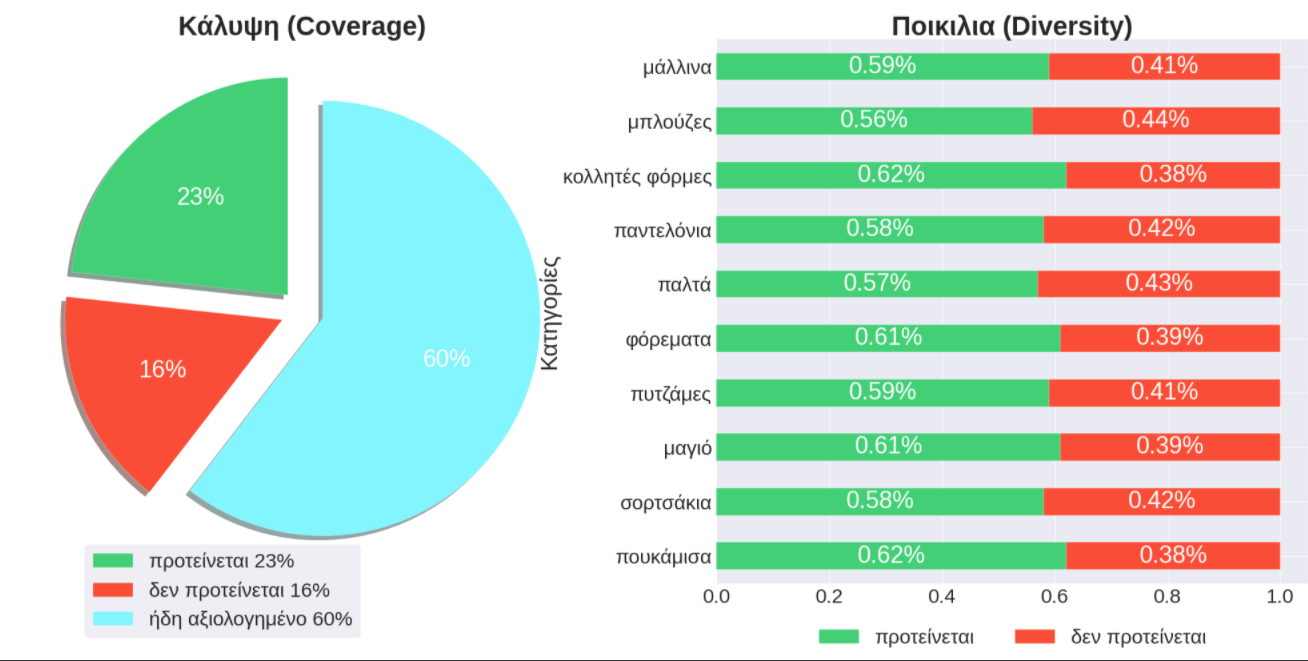
Συνοψίζοντας, τα διαγράμματα Κάλυψης και Ποικιλίας για την υλοποίηση του υβριδικού συστήματος μέσω γραμμικής παλινδρόμησης δείχνουν πως η ικανότητα ανακάλυψης των προϊόντων της βάσης είναι ικανοποιητική για όλα τα σύνολα. Επίσης υπάρχει ισορροπία στις προτάσεις που υπάρχουν ανά κατηγορία, δηλαδή το σύστημα δεν απορρίπτει μεγάλο εύρος ρούχων/ταινιών κάποιας κατηγορίας. Αυτό θα είχε ως αποτέλεσμα να μην γινόταν σωστά η ανακάλυψη αυτής της κατηγορίας.

Σετ δεδομένων: data30



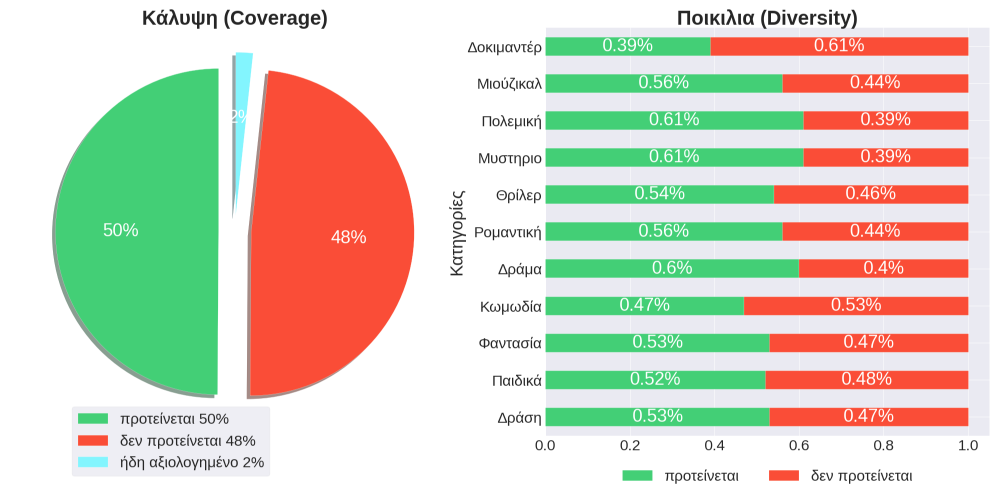
**Εικόνα 66** - Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα Κάλυψης και Ποικιλίας – data30

Σετ δεδομένων: data60



**Εικόνα 67** - Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα Κάλυψης και Ποικιλίας – data60

Σετ δεδομένων: movielens-100k



**Εικόνα 68** - Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα Κάλυψης και Ποικιλίας – movielens-100k

## **Συνοπτική αναπαράσταση όλων των υλοποιήσεων**

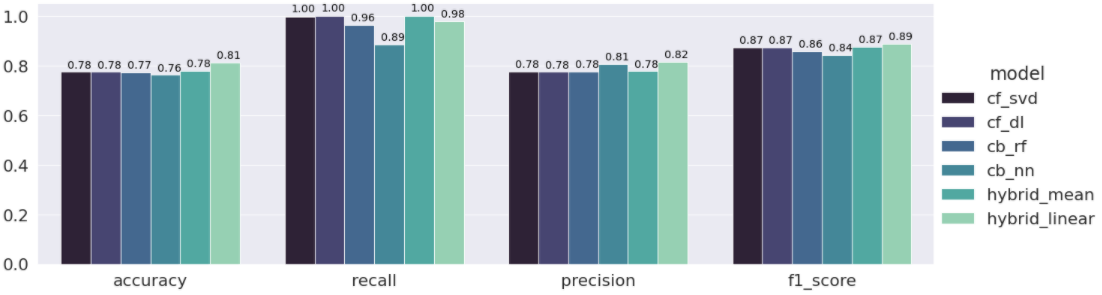
### **Accuracy, recall, precision και f1\_score**

Για το σύνολο δεδομένων data30, από τα αποτελέσματα φαίνεται ότι οι 3 πρώτες υλοποιήσεις (Ανάλυση πίνακα σε ιδιάζουσες τιμές, Συνεργατικά Νευρωνικά Δίκτυα και Τυχαία Δάση) φέρουν παρόμοια αποτελέσματα σε θέματα ταξινόμησης, αφού το accuracy και precision τους κυμαίνεται στο 77% με 78%, το recall στο 96% με 100% και το f1\_score στο 86% με 87%. Το χαμηλότερο recall φαίνεται να το έχει η υλοποίηση Νευρωνικών Δικτύων Περιεχομένου, χωρίς ωστόσο να υστερεί στις υπόλοιπες 3 μετρικές. Την καλύτερη επίδοση φαίνεται να έχει το υβριδικό μοντέλο γραμμικής παλινδρόμησης, αφού ξεπερνά σε όλες της μετρικές τα υπόλοιπα μοντέλα με εξαίρεση το recall, όπου αγγίζει το 98% *(Εικόνα 69).*

Ομοίως, για το σύνολο δεδομένων data60, οι 3 πρώτες υλοποιήσεις (Ανάλυση πίνακα σε ιδιάζουσες τιμές, Συνεργατικά Νευρωνικά Δίκτυα και Τυχαία Δάση) φέρουν παρόμοια αποτελέσματα σε θέματα ταξινόμησης, αφού το accuracy και precision τους κυμαίνεται στο 80% με 83%, το recall στο 97% με 100% και το f1\_score στο 87% με 91%. Κι εδώ το χαμηλότερο recall φαίνεται να παράγει η υλοποίηση Νευρωνικών Δικτύων Περιεχομένου, χωρίς ωστόσο να υστερεί στις υπόλοιπες 3 μετρικές. Την καλύτερη επίδοση την έχει πάλι το υβριδικό μοντέλο γραμμικής παλινδρόμησης, που ξεπερνά σε όλες της μετρικές τα υπόλοιπα μοντέλα με εξαίρεση το recall όπου αγγίζει το 97% *(Εικόνα 70).*

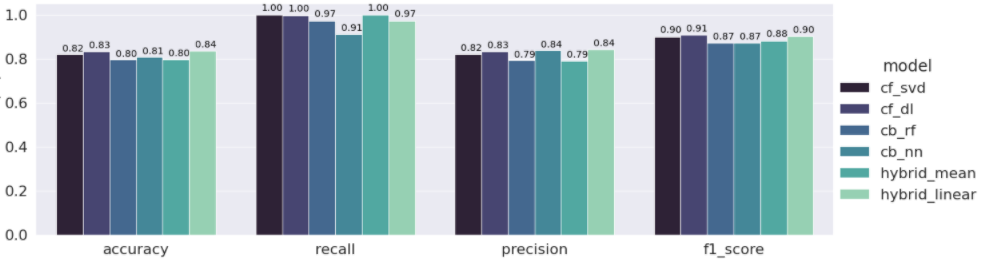
Τέλος, για το σύνολο δεδομένων movielens-100k καλύτερη υλοποίηση σε θέματα ταξινόμησης φαίνεται να είναι τα Νευρωνικά δίκτυα Περιεχομένου με επιδόσεις: 88% accuracy 96% recall, 90%, precision και 92% f1\_score. Δεύτερη καλύτερη υλοποίηση είναι αυτή που περιλαμβάνει τυχαία δάση με: 88% accuracy, 94% recall, 90% precision, και 92% f1\_score. Τρίτη στη σειρά βρίσκεται η υβριδική υλοποίηση γραμμικής παλινδρόμησης, ενώ στο τέλος βρίσκονται οι επιδόσεις των υλοποιήσεων Ανάλυση Πίνακα Ιδιάζουσας τιμής και του Συνεργατικού Νευρωνικού δικτύου *(Εικόνα 71).*

Σετ δεδομένων: data30



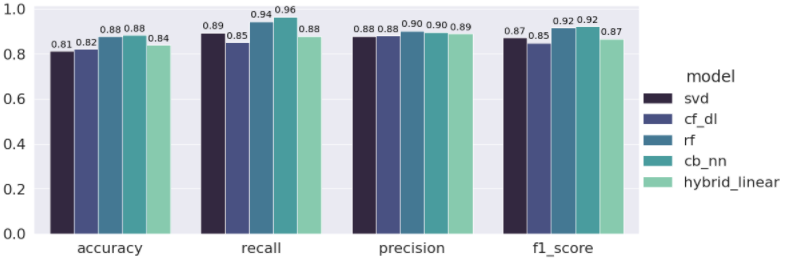
**Εικόνα 69** - Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data30

Σετ δεδομένων: data60



**Εικόνα 70** - Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – data60

Σετ δεδομένων: movielens 100K



**Εικόνα 71** - Υλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών accuracy, recall, precision και f1\_score – movielens 100K

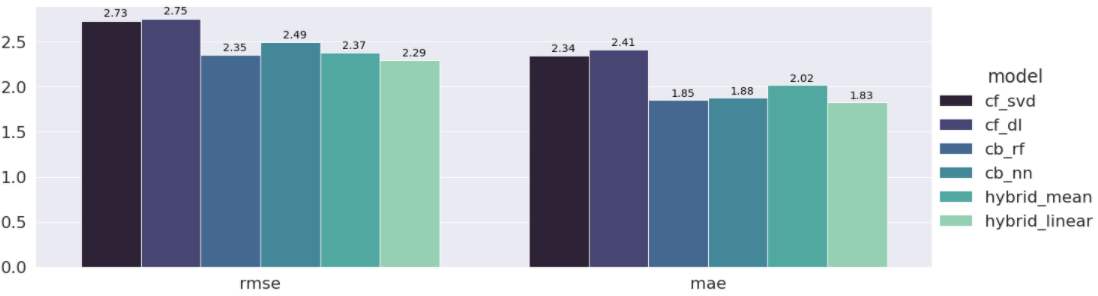
### **Μέσο απόλυτο σφάλμα (MAE) και ρίζα μέσου τετραγωνικού σφάλματος (RMSE)**

Για το σύνολο δεδομένων data30 βλέπουμε ότι οι 2 πρώτες υλοποιήσεις (Ανάλυση πίνακα σε ιδιάζουσες τιμές και Συνεργατικά Νευρωνικά Δίκτυα) φέρουν παρόμοιες τιμές σφάλματος, αφού έχουν ρίζα μέσου τετραγωνικού σφάλματος που κυμαίνεται από 2.73 έως 2.75 και παρόμοιο μέσο απόλυτο σφάλμα που κυμαίνεται από 2.34 έως 2.41. Αντίστοιχα, οι επόμενες δύο υλοποιήσεις (Τυχαία Δάση και Νευρωνικά Δίκτυα Περιεχομένου) φέρουν κι αυτές παρόμοιες τιμές σφάλματος αφού έχουν ρίζα μέσου τετραγωνικού σφάλματος που κυμαίνεται από 2.35 έως 2.49 και μέσο απόλυτο σφάλμα που κυμαίνεται από 1.85 έως 1.88. Την καλύτερη επίδοση φαίνεται να έχει το υβριδικό μοντέλο γραμμικής παλινδρόμησης, αφού έχει την ελάχιστη ρίζα μέσου τετραγωνικού σφάλματος 2.29 και το ελάχιστο μέσο απόλυτο σφάλμα 1.83 *(Εικόνα 72).*

Σε παρόμοια πλαίσια, για το σύνολο δεδομένων data60 βλέπουμε ότι οι 2 πρώτες υλοποιήσεις (Ανάλυση πίνακα σε ιδιάζουσες τιμές, Συνεργατικά Νευρωνικά Δίκτυα) φέρουν παρόμοιες τιμές σφάλματος, αφού έχουν ρίζα μέσου τετραγωνικού σφάλματος που κυμαίνεται από 2.47 έως 2.55 και παρόμοιο μέσο απόλυτο σφάλμα που κυμαίνεται από 2.04 έως 2.17. Αντίστοιχα, οι επόμενες δύο υλοποιήσεις (Τυχαία Δάση και Νευρωνικά Δίκτυα Περιεχομένου) φέρουν κι αυτές παρόμοιες τιμές σφάλματος, αφού έχουν ρίζα μέσου τετραγωνικού σφάλματος που κυμαίνεται από 2.14 έως 2.21 και μέσο απόλυτο σφάλμα που κυμαίνεται από 1.58 έως 1.65. Την καλύτερη επίδοση φαίνεται να έχει το υβριδικό μοντέλο γραμμικής παλινδρόμησης με ελάχιστη ρίζα μέσου τετραγωνικού σφάλματος 1.97 και ελάχιστο μέσο απόλυτο σφάλμα 1.52 *(Εικόνα 73).*

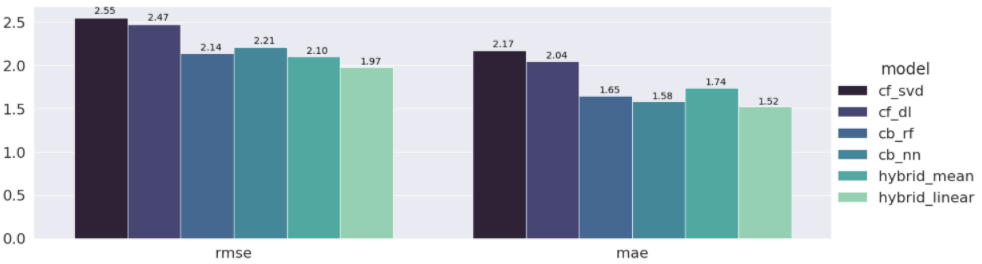
Τέλος, για το σύνολο δεδομένων movielens-100k παρόμοια αποτελέσματα φέρουν οι υλοποιήσεις Ανάλυση Πίνακα Ιδιάζουσων τιμών, Τυχαία Δάση και Νευρωνικά Δίκτυα περιεχομένου με τιμές ρίζας μέσου τετραγωνικού σφάλματος να κυμαίνονται από 0.91 έως 0.93 και μέσο απόλυτο σφάλμα από 0.75 έως 0.78. Ακολουθούν στην καλύτερη επίδοση η υλοποίηση του Νευρωνικού Συνεργατικού Δικτύου, ενώ την καλύτερη επίδοση την είχε εκ νέου το υβριδικό μοντέλο γραμμικής παλινδρόμησης με την ελάχιστη ρίζα μέσου τετραγωνικού σφάλματος 0.82 και το ελάχιστο μέσο απόλυτο σφάλμα 0.67. *(Εικόνα 74).*

Σετ δεδομένων: data30



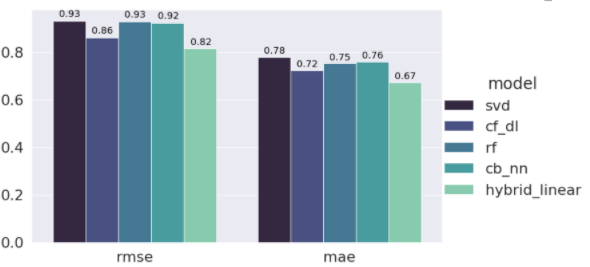
**Εικόνα 72** - Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – data30

Σετ δεδομένων: data60



**Εικόνα 73** - Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – data30

Σετ δεδομένων: movielens 100K



**Εικόνα 74** - Yλοποίηση Υβριδικού Συστήματος - Διάγραμμα μετρικών rmse και mae – movielens 100K

## **Αποτελέσματα αναζητήσεων**

Παρακάτω παρουσιάζονται μερικές από τις προτάσεις του συστήματος. Το σύστημα που επιλέχθηκε για την εξαγωγή των προτάσεων είναι το υβριδικό σύστημα γραμμικής παλινδρόμησης, καθώς παρουσίασε τις καλύτερες επιδόσεις στα πειραματικά αποτελέσματα.

Μετά την ολοκλήρωση την αναζήτησης του ρούχου παρουσιάζονται:

* Συνολικά 5 φωτογραφίες με ρούχα που αρέσουν ήδη στον χρήστη
* Το σύνολο των ρούχων που αρέσουν για αυτή την κατηγορία στο χρήστη
* Συνολικά 5 φωτογραφίες με ρούχα που προτείνονται από το σύστημα
* Το σύνολο των ρούχων που μπορούν να προταθούν από το σύστημα για αυτή την κατηγορία στο χρήστη

### **Κατηγορία: Μάλλινα**

*Παρατηρήσεις: (*Εικόνα 75*)*

1. Παρατηρούμε ότι η τέταρτη μάλλινη μπλούζα της πρώτης σειράς που αρέσει στον σχεδιαστή είναι αρκετά παρόμοια με την τέταρτη και Πέμπτη μαύρη μάλλινη μπλούζα της δεύτερης σειράς που προτείνει το σύστημα.
2. Παρατηρούμε ότι η τρίτη και πέμπτη μάλλινη γυναικεία μπλούζα της πρώτης σειράς που αρέσει στον σχεδιαστή είναι αρκετά παρόμοια (διαφορετικού χρώματος) με την πρώτη και δεύτερη της δεύτερης σειράς που προτείνει το σύστημα (μάκρος, γιακάς, μανίκια).
3. Οι υπόλοιπες προτάσεις που υπάρχουν προτάσεις δεν μπορούν να εξηγηθούν από την παρούσα φωτογραφία. Πολύ πιθανόν να αποτελούν αποτέλεσμα προτάσεων από αλληλεπιδράσεις άλλων χρηστών



**Εικόνα 75 -** Παρουσίαση αποτελεσμάτων ρούχων. Αναζήτηση: Γυναικεία Μάλλινα Μπλουζάκια

### **Κατηγορία: Παλτό**

*Παρατηρήσεις: (Εικόνα 76):*

1. Το τρίτο γκρι γυναικείο παλτό της πρώτης σειράς που αρέσει στον σχεδιαστή είναι αρκετά παρόμοιο με το τρίτο γκρι γυναικείο παλτό της δεύτερης σειράς που προτείνεται από το σύστημα.
2. Το τρίτο γκρι γυναικείο παλτό της πρώτης σειράς που αρέσει στον σχεδιαστή είναι αρκετά παρόμοιο με το δεύτερο σομόν γυναικείο παλτό της δεύτερης σειράς που προτείνεται από το σύστημα.
3. Το πέμπτο πράσινο γυναικείο παλτό με την γούνα στο γιακά πρώτης σειράς που αρέσει στον σχεδιαστή είναι αρκετά παρόμοιο με το τέταρτο γκρι γυναικείο παλτό της δεύτερης σειράς που προτείνεται από το σύστημα.
4. Οι υπόλοιπες προτάσεις που υπάρχουν προτάσεις δεν μπορούν να εξηγηθούν από την παρούσα φωτογραφία και πολύ πιθανόν να αποτελούν αποτέλεσμα προτάσεων από αλληλεπιδράσεις άλλων χρηστών



**Εικόνα 76** - Παρουσίαση αποτελεσμάτων ρούχων. Αναζήτηση: Γυναικεία Παλτό

### **Κατηγορία: Φόρεμα**

*Παρατηρήσεις: (Εικόνα 77)*

1. Tο δεύτερο και το τέταρτο γυναικείο φόρεμα της πρώτης σειράς που αρέσει στο σχεδιαστή μοιάζει αρκετά με το τρίτο και τέταρτο γυναικείο φόρεμα ως προς τα μανίκια το μάκρος που προτείνει το σύστημα.
2. Το τρίτο φόρεμα της πρώτης σειράς που αρέσει στον σχεδιαστή έχει σχέδια τύπου επαναλαμβανόμενα (pattern design) παρόμοια με το δεύτερο φόρεμα της δεύτερης σειράς που προτείνει το σύστημα.
3. Το γυναικείο φόρεμα της πρώτης που αρέσει στον σχεδιαστή είναι παρόμοιου χρώματος με το τρίτο και τέταρτο φόρεμα της δεύτερης σειράς που προτείνει το σύστημα.
4. Οι υπόλοιπες προτάσεις που υπάρχουν προτάσεις δεν μπορούν να εξηγηθούν από την παρούσα φωτογραφία και πολύ πιθανόν να αποτελούν αποτέλεσμα προτάσεων από αλληλεπιδράσεις άλλων χρηστών.



**Εικόνα 77** - Παρουσίαση αποτελεσμάτων ρούχων. Αναζήτηση: Γυναικεία Φορέματα

΅

### **Κατηγορία: Σορτσάκια**

*Παρατηρήσεις: (*Εικόνα 78*):*

1. Το πρώτο μαύρο σορτσάκι με την Ρίγα στη μέση που αρέσει στον σχεδιαστή μοιάζει αρκετά από άποψη σχεδίου με το πρώτο σορτσάκι της δεύτερης σειράς πάλι με τη ρίγα στη μέση που προτείνει το σύστημα.
2. Το δεύτερο τζιν σορτσάκι σετ που αρέσει το σχεδιαστή είναι ίδιο με το δεύτερο τζιν σορτσάκι σετ που προτείνει το σύστημα.
3. Το τέταρτο και πέμπτο σορτσάκι της πρώτης σειράς που αρέσει το σχεδιαστή είναι μονόχρωμα όπως αντίστοιχα με το τέταρτο και πέμπτο σορτσάκι της δεύτερης σειράς που προτείνει το σύστημα.
4. Οι υπόλοιπες προτάσεις που υπάρχουν προτάσεις δεν μπορούν να εξηγηθούν από την παρούσα φωτογραφία και πολύ πιθανόν να αποτελούν αποτέλεσμα προτάσεων από αλληλεπιδράσεις άλλων χρηστών



**Εικόνα 78** - Παρουσίαση αποτελεσμάτων ρούχων. Αναζήτηση: Αντρικά Σορτσ

### **Κατηγορία: Γυναικεία Πουκάμισα**

*Παρατηρήσεις: (Εικόνα 79):*

1. Το τρίτο γυναικείο λεοπάρ πουκάμισο της πρώτης σειράς φορέσει το σχεδιαστή μοιάζει αρκετά με το τέταρτο γυναικείο πουκάμισο της δεύτερης σειράς που προτείνει το σύστημα.
2. Το πρώτο, τέταρτο και το πέμπτο είναι άσπρα γυναικεία πουκάμισα που αρέσουν στον σχεδιαστή και μοιάζουν με το πέμπτο πουκάμισο της δεύτερης σειράς που προτείνει το σύστημα.
3. Το δεύτερο γυναικείο πουκάμισο που αρέσει σχεδιαστή έχει παρόμοιο ύφασμα με το δεύτερο πουκάμισο που προτείνει το σύστημα.
4. Οι υπόλοιπες προτάσεις που υπάρχουν προτάσεις δεν μπορούν να εξηγηθούν από την παρούσα φωτογραφία και πολύ πιθανόν να αποτελούν αποτέλεσμα προτάσεων από αλληλεπιδράσεις άλλων χρηστών



**Εικόνα 79** - Παρουσίαση αποτελεσμάτων ρούχων. Αναζήτηση: Γυναικεία Πουκάμισα

### **Κατηγορία: Αντρικά Πουκάμισα**

*Παρατηρήσεις: (Εικόνα 80):*

1. Το πρώτο πουκάμισο της πρώτης σειράς με τα ριγέ σχέδια που αρέσει στον σχεδιαστή μοιάζει αρκετά με το πέμπτο πουκάμισο της δεύτερης σειράς που προτείνει το σύστημα
2. Το τρίτο αντρικό πουκάμισο της πρώτης σειράς που αρέσει τον σχεδιαστή μοιάζει αρκετά με το τρίτο πουκάμισο της δεύτερης σειράς που προτείνει το σύστημα.
3. Το δεύτερο αντρικό πουκάμισο της πρώτης σειράς που είναι σκούρο μπλε μοιάζει αρκετά με το τέταρτο πουκάμισο της δεύτερης σειράς σου προτείνει το σύστημα.
4. Οι υπόλοιπες προτάσεις που υπάρχουν προτάσεις δεν μπορούν να εξηγηθούν από την παρούσα φωτογραφία και πολύ πιθανόν να αποτελούν αποτέλεσμα προτάσεων από αλληλεπιδράσεις άλλων χρηστών



**Εικόνα 80** - Παρουσίαση αποτελεσμάτων ρούχων. Αναζήτηση: Ανδρικά Πουκάμισα

### **Κατηγορία: Γυναικεία Μαγιό**

*Παρατηρήσεις: (*Εικόνα 81*):*

1. Το δεύτερο μαγιό της πρώτης σειράς που αρέσει στον σχεδιαστή μοιάζει αρκετά με το δεύτερο μαγιό της δεύτερης σειράς που προτείνει το σύστημα. Τα δύο αυτά μαγιό είναι ασπρόμαυρα με παρόμοια σχέδια και διαστάσεις
2. Το τέταρτο μαγιό της πρώτης σειράς που αρέσει στον σχεδιαστή είναι μαύρο ολόσωμο και μοιάζει αρκετά με το μαγιό στην δεύτερη σειρά της τέταρτης θέσης που προτείνει το σύστημα.
3. Το μαγιό στην πέμπτη θέση της πρώτης σειράς που ο σχεδιαστής έχει δηλώσει όπως του αρέσει είναι ένα λεοπάρ μαγιό και υπάρχει αντίστοιχα πρόταση παρόμοιου λεοπάρ μαγιό στην δεύτερη σειρά της πέμπτης θέσης
4. Οι υπόλοιπες προτάσεις που υπάρχουν προτάσεις δεν μπορούν να εξηγηθούν από την παρούσα φωτογραφία και πολύ πιθανόν να αποτελούν αποτέλεσμα προτάσεων από αλληλεπιδράσεις άλλων χρηστών



**Εικόνα 81 -** Παρουσίαση αποτελεσμάτων ρούχων. Αναζήτηση: Γυναικεία Μαγιό

# **Συμπεράσματα και Μελλοντικές Επεκτάσεις**

Στο τελευταίο κεφάλαιο παρουσιάζονται τα συμπεράσματα που προέκυψαν από την πειραματική διαδικασία, αλλά και οι μελλοντικές επεκτάσεις είναι εφικτό να γίνουν ώστε να βελτιωθεί το γενικευμένο σύστημα προτάσεων.

## **Γενικά Συμπεράσματα**

Η παρούσα διπλωματική εστιάζει κυρίως στην ανάπτυξη ενός συστήματος γενικευμένων προτάσεων όπου τα δεδομένα προσαρμόζονται στο σύστημα προ-επεξεργάζονται και εξάγουν χρήσιμα συμπεράσματα. Τα πειράματα που ακολούθησαν στο γενικευμένο σύστημα αφορούν δεδομένα εισόδου από κριτικές ρούχων για τη δημιουργία προτάσεων σε σχεδιαστές και δεδομένα εισόδου από κριτικές ταινιών για την παροχή προτάσεων, ικανών, να κερδίσουν το ενδιαφέρον κάθε χρήστη.

Για την δημιουργία των προτάσεων χρησιμοποιούνται 4 μοντέλα τα οποία μέσω της χρήσης ενός υβριδικού συστήματος αξιοποιούνται με διαφορετική βαρύτητα για την εξαγωγή των τελικών προτάσεων. Κάθε ένα από τα 4 μοντέλα έχουν τις δικές τους παραμέτρους και ρυθμίζονται κατάλληλα με όσο κατά τον δυνατόν αυτοματοποιημένο τρόπο. Κατά την διάρκεια των πειραμάτων, η διαδικασία ρύθμισης αυτών τον παραμέτρων φάνηκε αρκετά σημαντική στην επίδραση των επιδόσεων ακρίβειας, ταξινόμησης και σφάλματος.

Όπως φάνηκε από τα πειράματα το μοντέλο συνεργατικού φιλτραρίσματος ως προς τον χρήστη μπόρεσε να δώσει προτάσεις με υψηλή ακρίβεια χωρίς ωστόσο να ανακαλύψει πετυχημένα ολόκληρη την βάση αφού οι προτάσεις αυτές που δημιουργήθηκαν ήταν για μικρό ποσοστό των ρούχων. Από την άλλη η υλοποίηση της Ανάλυσης Πίνακας σε Ιδιάζουσες τιμές και του Συνεργατικού Νευρωνικού Φιλτραρίσματος έδωσαν προτάσεις με καλή ακρίβεια και ποικιλία αφού λάμβαναν υπόψη τις αλληλεπιδράσεις άλλων χρηστών. Τέλος η υλοποίηση των Τυχαίων Δασών και η υλοποίηση του Νευρωνικού Δικτύου Περιεχομένου φέραν και αυτές τις προτάσεις με υψηλή ακρίβεια και κάποια εξειδίκευση αφού αφορούσαν προτάσεις που δεν λάμβαναν υπόψη τους υπόλοιπους χρήστης παρά μόνο τον ίδιο. Το υβριδικό σύστημα απέδειξε ότι μπορεί να αξιοποιήσει όλα τα θετικά στοιχεία μιας υλοποίησης και να φέρει το μέγιστο δυνατό αποτέλεσμα. Σε όλα τα σύνολα δεδομένων το υβριδικό σύστημα γραμμικής παλινδρόμησης είχε τα καλύτερα ποσοστά ταξινόμησης και το ελάχιστο σφάλμα εκτίμησης.

Απ’ όλη την πορεία της παρούσας διπλωματικής φαίνεται τελικά πως τα συστήματα προτάσεων μπορούν να λύσουν σύγχρονα προβλήματα πολλών επιχειρήσεων διαφορετικού τύπου. Στο κομμάτι της ένδυσης, μπορεί να επιτευχθεί η βελτιστοποίηση της παραγωγής και η δημιουργία προϊόντων κοντά στις προτιμήσεις των καταναλωτών, που αποτελούν και τα συστατικά της επιτυχίας στην εξέλιξη μιας επιχείρησης οίκου μόδας. Όμοια επιχειρήσεις συνδρομητικού κινηματογράφου έχουν ανάγκη την σωστή παροχή προτάσεων ώστε να μην χάσουν τους μηνιαίους συνδρομητές τους.

## **Μελλοντικές επεκτάσεις**

Μια πρώτη μελλοντική επέκταση για την καλύτερη λειτουργία του συστήματος θα μπορούσε να είναι η δημιουργία ενός καλύτερου συνόλου δεδομένων, με περισσότερα ρούχα και κατηγορίες. Αυτό για να συμβεί χρειάζεται οι εταιρίες ένδυσης να οργανώσουν καλά τα δεδομένα των ρούχων τους και μέσω ενός εθελοντικού ή μισθωτού προσωπικού να λαμβάνουν καθημερινά αξιολογήσεις. Τα δεδομένα αυτά που, εν τέλη, θα συλλεχθούν, θα προσφέρουν αναμφίβολα περισσότερη ακρίβεια στις προβλέψεις του μοντέλου και κατ’ επέκταση καλύτερες προτάσεις εμπορικής αξίας στον σχεδιαστή.

Ένας ακόμη μελλοντικός στόχος θα μπορούσε να είναι η αξιολόγηση των προτάσεων του συστήματος από τον ίδιο τον σχεδιαστή. Αυτό μπορεί να συμβεί με την εμφάνιση ενός παραθύρου στο σύστημα το οποίο θα ρωτά τον σχεδιαστή αν η πρόταση του ήταν αρεστή ή όχι. Με αυτό τον τρόπο το σύστημα θα μπορεί να αναθεωρήσει και τελικώς να διορθώσει τυχόν ατέλειες στις προτάσεις.

Τέλος, το σύστημα θα μπορούσε να αναβαθμιστεί λαμβάνοντας όχι μόνο υπόψη τις αξιολογήσεις των προϊόντων αλλά και τις πωλήσεις ώστε με αυτό τον τρόπο να ικανοποιούνται και εμπορικές ανάγκες που θα αποτρέψουν ζημιογόνες επενδύσεις σε νέες κολεξιόν.

# **Βιβλιογραφία**

|  |  |
| --- | --- |
| [1] | «Wikipedia,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Collaborative\_filtering. |
| [2] | D. Jannach , M. Zanker , A. Felfernig και G. Friedrich, Recommender Systems: An Introduction, 2010, pp. 13-22. |
| [3] | Y. Koren, Factorization Meets the Neighborhood: a Multifaceted Collabora-tive Filtering Model, 2008. |
| [4] | Υ. Koren, R. Bell και C. Volinsky,, Matrix factorization Techniques forRecommender Systems, 2009. |
| [5] | S. Berkovsky , I. Cantador και D. Tikk, Collaborative Recommendations: Algorithms, Practical Challenges And Applications, 2018, pp. 32-40. |
| [6] | A. Tulloch, «Facebook Research,» [Ηλεκτρονικό]. Available: https://research.fb.com/blog/2014/09/fast-randomized-svd/. |
| [7] | E. Jane και T. Bailey , Folding-up: A Hybrid Method for Updating the Partial Singular Value Decomposition in Latent Semantic Indexing, 2006. |
| [8] | S. Rendle, Factorization machines, 2010, p. 995–1000. |
| [9] | «Towards Science,» [Ηλεκτρονικό]. Available: https://towardsdatascience.com/neural-collaborative-filtering-96cef1009401. |
| [10] | S. Rendle, . C. Freudenthaler, Z. Gantner και L. Schmidt-Thieme, Bayesian personalized ranking from implicit feedback, 2009, p. 452–461. |
| [11] | R. Socher, . D. Chen, C. D. Manning και . A. N. Reasoning, Reasoning with neural tensor networks for knowledge base completion, 2013, p. 926–934. |
| [12] | N. Srivastava και R. R. Salakhutdinov, Multimodal learning with deep boltzmann machines, 2012, p. 2222–2230. |
| [13] | B. Xu, R. Huang και M. Li, Revise Saturated Activation Functions, 2015. |
| [14] | 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. |
| [15] | X. Glorot, A. Bordes και . Y. Bengio, Deep sparse rectifier neural networks., 2011. |
| [16] | K. He, X. Zhang, . S. Ren και J. Sun, Deep residual learning for image recognition, 2016. |
| [17] | «WikiPedia,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Random\_forest#/media/File:Random\_forest\_diagram\_complete.png. |
| [18] | «Wikipedia,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Random\_forest. |
| [19] | R. Burke, Hybrid recommender systems: Survey and experiments. User Modeling and User-adapted Interaction, 2002, p. 331–370. |
| [20] | M. Jahrer, A. Toscher και R. Legenstein, Combining predictions for accurate recommender, ACM KDD Conference, 2010., p. 693–702. |
| [21] | Y. Koren, R. Bell και C. Volinsky, Matrix Factorization Techniques For Recommender Systems. |
| [22] | H. Xiangnan , L. Liao και H. Zhang, Neural Collaborative Filtering, 2017. |
| [23] | H.-R. Zhang, F. Min και X. He, Aggregated Recommendation through Random Forests, 2014. |
| [24] | J. Han, M. Kamber και J. Pei, Data Mining: Concepts and Techniques, 2012. |
| [25] | T. Chai και . R. Draxler, «Root mean square error (RMSE) or mean absolute error (MAE)?,» p. 2, 2014. |
| [26] | J. Herlocker , . J. Konstan, L. Terveen και J. Riedl , Evaluating collaborative filtering recommender systems., ACM Transactions on Information Systems, 2004, pp. 5-53. |
| [27] | D. Jannach, M. Ge και C. Delgado-Battenfeld, Beyond accuracy: Evaluating recommender systems by coverage and serendipity, Barcelona, 2010. |
| [28] | K. Bradley και B. Smyth, Improving recommendation diversity, Maynooth: AICS-01, 2001, pp. 75-84. |
| [29] | M. Kunavera και T. Požrlb, Diversity in recommender systems – A survey, Ljubljana, 2017. |
| [30] | L. Zhang, The Definitionof Novelty in Recommendation System, Guiyang: School of management Xiamen university Xiamen, 2013. |
| [31] | L. Lü, M. Medo, H. Yeung και . Z. Yi-Cheng, Recommender systems, Physics Reports, 2012. |
| [32] | G. Karypis και C. Desrosiers, A comprehensive survey of neighborhood-basedrecommendation methods, Springer, 2011, pp. 21-22. |
| [33] | X. Su και T. M. Khoshgoftaar, A Survey of Collaborative Filtering Techniques, USA: Hindawi Publishing Corporation, 2009, pp. 6-7. |
| [34] | [Ηλεκτρονικό]. Available: https://www.deeplearningwizard.com/deep\_learning/boosting\_models\_pytorch/weight\_initialization\_activation\_functions/. |
| [35] | «Optuna,» [Ηλεκτρονικό]. Available: https://optuna.org/. |
| [36] | C. C. Aggarwal, Recommender Systems - The Textbook, Springer, 2006. |
| [37] | B. Sarwar, G. Karypis, J. Konstan και J. Riedl, Item-based Collaborative Filtering RecommendationAlgorithms, Hong Kong, 2001, pp. 6-7. |