Σελίδα

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΙΩΑΝΝΙΝΩΝ

ΣΧΟΛΗ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΜΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΔΙΚΤΥΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΑΝΑΛΥΣΗ ΚΑΙ ΣΧΕΔΙΑΣΜΟΣ ΣΥΣΤΗΜΑΤΩΝ ΣΥΣΤΑΣΕΩΝ

Παχούλας Γεώργιος

Επιβλέπων: Στύλιος Χρυσόστομος

Καθηγητής

Άρτα, Σεπτέμβριος, 2023

ANALYSIS AND DESIGN OF RECOMMENDATION SYSTEMS

Εγκρίθηκε από τριμελή εξεταστική επιτροπή

Άρτα, 29/09/2023

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Επιβλέπων καθηγητής

   Στύλιος Χρυσόστομος,

   Καθηγητής

2. Μέλος επιτροπής

   Καρβέλης Πέτρος

   Επίκουρος Καθηγητής

3. Μέλος επιτροπής

   Λιάγκου Βασιλική

   Επίκουρη Καθηγήτρια

Δήλωση μη λογοκλοπής

Δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα μεταπτυχιακή εργασία είναι εκ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία.

Υπογραφή

Παχούλας, Γεώργιος

ΕΥΧΑΡΙΣΤΙΕΣ

Ευχαριστώ θερμά τον επιβλέποντα καθηγητή της διπλωματικής εργασίας Καθηγητή Χρυσόστομο Στύλιο για την εμπιστοσύνη και την πολύτιμη καθοδήγηση που μου έδειξε. Επιπλέον δεν θα μπορούσα να μην ευχαριστήσω το Δρ. κ. Jeries Besharat για την στήριξη που μου παρείχε καθόλη την διάρκεια της μελέτης.

Επιπλέον θα ήθελα να απευθύνω ευχαριστίες στους συναδέλφους μου στο Εργαστήριο Γνώσης και Ευφυούς Πληροφορικής οι οποίοι με στήριξαν και μου έδωσαν κουράγιο ώστε να μπορέσω να ολοκληρώσω αυτή τη δύσκολη προσπάθεια.

Τέλος θέλω από τα βάθη της καρδιάς μου να ευχαριστήσω την οικογένειά μου για την για την ανεκτίμητη στήριξή τους, τόσο για την παρούσα εργασία, όσο και για την ολοκλήρωση των σπουδών μου.

ΠΕΡΙΛΗΨΗ

Τα συστήματα συστάσεων αποτελούν ένα σημαντικό κομμάτι του σύγχρονου ψηφιακού κόσμου, διαπιστώνοντας αυτοματοποιημένα τις τάσεις και τις προτιμήσεις κάθε χρήστη μέσα από τον τρόπο καθημερινής αλληλεπίδρασης και προσφέροντας προσαρμοσμένες συστάσεις και πληροφορίες σε κάθε χρήστη. Στο επίκεντρο των συστημάτων συστάσεων βρίσκεται η ιδέα της εξατομικευμένης συμβουλής και εμπειρίας, δίνοντας έμφαση στη δημιουργία αξίας για τον κάθε χρήστη. Τα συστήματα συστάσεων βασίζονται στην ιδέα ότι μπορούν να παρέχουν προσαρμοσμένες συστάσεις στους χρήστες, λαμβάνοντας υπόψη τα προφίλ τους, τις προτιμήσεις τους και την δραστηριότητα τους στο διαδίκτυο και την αλληλεπίδρασή τους με ένα πληροφοριακό σύστημα. Σήμερα πραγματοποιείται συνεχή παρακολούθηση, συλλογή και καταγραφή όλων των δραστηριοτήτων και αλληλεπιδράσεων μας με τα πληροφοριακά συστήματα οπότε η κατανόηση και η αξιοποίηση των συστημάτων συστάσεων αποτελούν ζωτικό στοιχείο για την ανάδειξη και την ανταγωνιστικότητα των επιχειρήσεων και την βελτίωση της ψηφιακής μας εμπειρίας.

Σκοπός της παρούσας διπλωματικής εργασίας ήταν η μελέτη, ο σχεδιασμός και η ανάπτυξη ενός συστήματος συστάσεων και η ενσωμάτωση του σε εφαρμογή τουριστικού περιεχομένου με σκοπό την δημιουργία εξατομικευμένων συστάσεων στους χρήστες. Τα δεδομένα που εξετάζονται, συλλέγονται ανώνυμα από την εφαρμογή και παράγονται από την αλληλεπίδραση των χρηστών με την εφαρμογή όπως είναι η πλοήγηση στις σελίδες του δικτυακού τόπου, η αλληλεπίδραση του χρήστη μέσα στις διάφορες διεπαφές, η αναζήτηση πληροφοριών και αντικειμένων κλπ. Το σύστημα συστάσεων που κατασκευάστηκε βασίζεται στο περιεχόμενο των αντικειμένων προς σύσταση (Content – Based Filtering) και έχει ως σκοπό την πρόταση παρόμοιων αντικειμένων με ίδια χαρακτηριστικά που να αντιστοιχούν στις προτιμήσεις των χρηστών. Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στα πλαίσια του ΠΜΣ με τίτλο «Πληροφορική και Δίκτυα» του τμήματος Πληροφορικής και Τηλεπικοινωνιών του Πανεπιστημίου Ιωαννίνων. Για την δημιουργία του συστήματος συστάσεων χρησιμοποιήθηκαν προηγμένες τεχνικές σχεδιασμού και ανάπτυξης λογισμικού.

Λέξεις-κλειδιά: Διαδικτυακός προγραμματισμός, συστήματα συστάσεων, σύσταση βάση περιεχομένου, συνεργατικό φιλτράρισμα, υβριδικά συστήματα συστάσεων, προφίλ χρηστών.

ABSTRACT

Recommendation systems are an important part of the digital world, offering personalized recommendations and information to each user. At the heart of recommender systems is the idea of a personalised experience, emphasising the importance of each user. Recommendation systems are based on the idea that they can provide personalized recommendations to users, taking into account their profiles, preferences and historical activity. Today, with the increasing collection and availability of information, understanding and leveraging recommendation systems is a vital element in making businesses stand out and competitive and improving our digital experience.

The purpose of this thesis was the study, design and development of a recommendation system and its integration in a tourist content application in order to create personalized recommendations to users. The data considered are collected anonymously by the application and are generated by the interaction of users with the application, such as for example navigating the pages - interfaces, searching for objects, etc. The recommendation system built is based on the content of the objects to be recommended (Content - Based Filtering) and aims to suggest similar objects with user preferences. This thesis was carried out within the framework of the MSc in "Informatics and Networks" of the Department of Informatics and Telecommunications of the University of Ioannina. Advanced software design and development techniques were used to create the system.

Keywords: Web programming, recommendation system, content – based recommendation, collaborative filtering, hybrid recommendation systems, user profiles.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΕΥΧΑΡΙΣΤΙΕΣ....7
ΠΕΡΙΛΗΨΗ....8
ABSTRACT....10
ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ....11
ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ....16
ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ/ΕΙΚΟΝΩΝ....17
ΕΙΣΑΓΩΓΗ....20
1 Συστήματα συστάσεων....21
1.1 Τι είναι τα συστήματα συστάσεων....21
1.2 Ιστορική αναδρομή....23
1.3 Δομικά μέρη ενός συστήματος συστάσεων....25
1.3.1 Αντικείμενα (Items)....25
1.3.2 Χρήστες (Users)....26
1.3.3 Συναλλαγές (Transactions)....26
1.4 Χαρακτηριστικά και οφέλη συστημάτων συστάσεων....27
1.4.1 Οφέλη ως προς τους παρόχους υπηρεσιών....28
1.4.2 Οφέλη ως προς τους χρήστες....29
1.5 Κατηγορίες συστημάτων συστάσεων....29
1.5.1 Συστήματα βασισμένα στο περιεχόμενο (Content – based filtering)....31
1.5.2 Συστήματα βασισμένα στην συνεργασία (Collaborative filtering)....34
1.5.3 Συστήματα με βάση την γνώση (Knowledge – based systems)....38
1.5.4 Υβριδικά συστήματα συστάσεων (Hybrid Recommendation Systems)....40
1.6 Προβλήματα – περιορισμοί στα συστήματα συστάσεων....44

1.6.1 Πρόβλημα ψυχρής εκκίνησης (Cold – start problem) 44
1.6.2 Πρόβλημα της συνωνυμίας (Synonymy problem) 45
1.6.3 Πρόβλημα της ιδιωτικότητας (Privacy problem) 45
1.6.4 Πρόβλημα γκρίζου προβάτου (Grey sheep problem) 46
1.6.5 Πρόβλημα αραιών και διάσπαρτων δεδομένων (Data Sparity) 46
1.6.6 Πρόβλημα επιθέσεων Shilling (Shilling attacks problem) 47
1.6.7 Πρόβλημα περιορισμένης ανάλυσης περιεχομένου και υπερεξειδίκευσης (Limited Content Analysis and Overspecialization problem) 47
1.6.8 Πρόβλημα καθυστέρησης (Latency problem) 48
1.7 Αξιολόγηση συστημάτων συστάσεων 48
1.7.1 Offline μέθοδος 50
1.7.2 Online μέθοδος 54
1.8 Παραδείγματα συστημάτων συστάσεων 55
1.8.1 Netflix 55
1.8.2 Amazon 56
1.8.3 YouTube 57
2 Μηχανική μάθηση και εξόρυξη δεδομένων 58
2.1 Εξόρυξη δεδομένων 58
2.1.1 Τεχνικές εξόρυξης δεδομένων 59
2.1.2 Περιγραφή μεθοδολογίας εξόρυξης δεδομένων 67
2.2 Μηχανική μάθηση (Machine Learning) 71
2.2.1 Επιβλεπόμενη μάθηση (Supervised Learning) 72
2.2.2 Μη επιβλεπόμενη μάθηση (Unsupervised Learning) 73
2.2.3 Ημι – επιβλεπόμενη μάθηση (Semi – supervised learning) 74
2.2.4 Ενισχυτική μάθηση (Reinforcement learning) 74
3 Δεδομένα σε τουριστικές εφαρμογές 76
3.1 Δεδομένα UGC 76

3.1.1 Διαδικτυακά δεδομένα κινητού 77
3.1.2 Δεδομένα φωτογραφιών 77
3.2 Δεδομένα συναλλαγών ανά ενέργεια χρήστη 77
3.2.1 Δεδομένα αναζήτησης από το διαδίκτυο (Search Data) 78
3.2.2 Δεδομένα επισκεψιμότητας ιστοσελίδας 78
3.3 Μεθοδολογία συλλογής δεδομένων 78
3.3.1 Παρατήρηση 79
3.3.2 Συνέντευξη 79
3.3.3 Ερωτηματολόγια 80
3.3.4 Καταγραφή δραστηριοτήτων χρήστη (Clickstreams) 82
4 Προφίλ χρηστών 84
4.1 Τύποι προφίλ χρηστών 84
4.1.1 Στατικό προφίλ 85
4.1.2 Δυναμικό προφίλ 85
4.2 Τύποι δεδομένων σε προφίλ χρηστών 86
4.3 Αναπαράσταση προφίλ χρηστών 88
4.3.1 Βασισμένο σε όρους 88
4.3.2 Βασισμένο σε σημασιολογικά δίκτυα 90
4.3.3 Με βάση τις έννοιες (Concept – based profiles) 92
4.4 Διαδικασία δημιουργίας προφίλ χρήστη 93
4.4.1 Συλλογή δεδομένων χρήστη 93
4.4.2 Δημιουργία προφίλ χρήστη 96
4.4.3 Ενημέρωση προφίλ χρήστη 97
5 Αλγόριθμοι και δεδομένα υλοποίησης 98
5.1 Αλγόριθμοι υλοποίησης 98
5.1.1 Δέντρα αποφάσεων (Decision Trees) 98
5.1.2 Random Forest 108

5.1.3 Σύγκριση Decision Tree & Random Forest....115
5.1.4 Επιλογή του καταλληλότερου αλγορίθμου....117
5.2 Δεδομένα ερωτηματολογίων στην εφαρμογή....118
5.2.1 Ερωτηματολόγιο Hotel (Hotel Survey)....119
5.2.2 Ερωτηματολόγιο Restaurant (Restaurant Survey)....120
5.2.3 Ερωτηματολόγιο Recommendation (Recommendation Survey)....121
5.3 Δεδομένα καταγραφής δραστηριοτήτων στην εφαρμογή (clickstreams)....121
5.3.1 Παραδείγματα χρήσης clickstreams στην υλοποίηση....127
6 Τεχνολογίες υλοποίησης....131
6.1 Τεχνολογίες από την πλευρά του διακομιστή (Backend)....131
6.1.1 Node.js....131
6.1.2 REST API....134
6.1.3 Python....136
6.1.4 MongoDB....138
6.2 Τεχνολογίες από την πλευρά του πελάτη (Frontend)....141
6.2.1 Angular....142
6.2.2 HTML....145
6.3 Επιπλέον πακέτα και βιβλιοθήκες....146
6.3.1 Pandas....147
6.3.2 Scikit – learn....149
7 Σχεδιασμός και ανάπτυξη συστήματος συστάσεων....152
7.1 Περιγραφή του συστήματος συστάσεων....153
7.2 Απαιτήσεις συστήματος....154
7.2.1 Λειτουργικές απαιτήσεις....154
7.2.2 Μη λειτουργικές απαιτήσεις....155
7.3 Επιμέρους μονάδες του συστήματος....156
7.3.1 Μονάδα συλλογής δεδομένων....157

7.3.2 Μονάδα διαχείρισης και ανάλυσης δεδομένων 158
7.3.3 Μονάδα εξαγωγής κατηγοριών και υποκατηγοριών 159
7.3.4 Μονάδα δημιουργίας προφίλ χρηστών 161
7.3.5 Μονάδα δημιουργίας συστάσεων 163
7.3.6 Μονάδα παρουσίασης συστάσεων 166
8 Συμπεράσματα και μελλοντικές προσθήκες 169
ΠΑΡΑΡΤΗΜΑ Α - ΕΡΩΤΗΜΑΤΟΛΟΓΙΟ Recommendation Survey 171
ΒΙΒΛΙΟΓΡΑΦΙΑ 175

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 1.1 Συγκριτικός πίνακας των μεθόδων αξιολόγησης (Πηγή: Najmani et al., 2022) 50
Πίνακας 5.1 Συγκριτικός πίνακας Decision Tree & Random Forest 116
Πίνακας 7.1 Πίνακας προτιμήσεων του χρήστη 162
Πίνακας 7.2 Πίνακας χαρακτηριστικών περιεχομένου 162
Πίνακας 7.3 Πίνακας σταθμισμένων χαρακτηριστικών 163
Πίνακας 7.4 Προφίλ χρήστη 163
Πίνακας 7.5 Πίνακας χαρακτηριστικών υποψήφιων παραλιών 164
Πίνακας 7.6 Σταθμισμένος πίνακας υποψήφιων παραλιών 164
Πίνακας 7.7 Πίνακας πιθανού ενδιαφέροντος 165
Πίνακας 7.8 Η συνάρτηση classify χρησιμοποιείται για την κατηγοριοποίηση των δεδομένων κάνοντας χρήση τον αλγόριθμο Random Forest 166

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ/ΕΙΚΟΝΩΝ

Εικόνα 1.1 Κατηγοριοποίηση των συστημάτων συστάσεων 30
Εικόνα 1.2 Παράδειγμα σταθμισμένου υβριδικού συστήματος.....41
Εικόνα 1.3 Παράδειγμα εναλλασσόμενου υβριδικού συστήματος.....41
Εικόνα 1.4 Παράδειγμα μεικτού υβριδικού συστήματος.....42
Εικόνα 1.5 Παράδειγμα υβριδικού συστήματος βασιζόμενο στα χαρακτηριστικά 42
Εικόνα 1.6 Παράδειγμα cascade υβριδικού συστήματος.....43
Εικόνα 1.7 Παράδειγμα υβριδικού συστήματος επαύξησης χαρακτηριστικών.....43
Εικόνα 1.8 Offline μέθοδος αξιολόγησης (Πηγή: Najmani et al., 2022) 52
Εικόνα 1.9 Online μέθοδος αξιολόγησης (Πηγή: Najmani et al., 2022).....55
Εικόνα 1.10 Εξατομικευμένες συστάσεις στην πλατφόρμα του Netflix.....56
Εικόνα 1.11 Παράδειγμα συστάσεων από την πλατφόρμα του Amazon.....57
Εικόνα 1.12 Παράδειγμα συστάσεων στην πλατφόρμα του YouTube 57
Εικόνα 2.1 Διαδικασία εξόρυξης δεδομένων. 59
Εικόνα 2.2 Κλάδοι σύνδεσης της εξόρυξης δεδομένων.....59
Εικόνα 2.3 Ταξινομητές ανεπιθύμητων μηνυμάτων αλληλογραφίας. 60
Εικόνα 2.4 Συσταδοποίηση με μοντέλο πυκνότητας 63
Εικόνα 2.5 Συσταδοποίηση με κεντροειδή μοντέλο 64
Εικόνα 2.6 Στάδια εξόρυξης δεδομένων 68
Εικόνα 2.7 Παράδειγμα Επιβλεπόμενης Μάθησης. Κατηγοριοποίηση email. (Heidenreich, 2018).....73
Εικόνα 2.8 Παράδειγμα συσταδοποίησης. (Heidenreich, 2018).....74
Εικόνα 3.1 Διαχωρισμός μεγάλων δεδομένων για τουριστικές εφαρμογές.....76
Εικόνα 4.1 Διαχωρισμός των προφίλ χρηστών σε στατικά και δυναμικά προφίλ.....85
Εικόνα 4.2 Παράδειγμα χρήσης του PEA. 89
Εικόνα 4.3 Παράδειγμα εφαρμογής του InfoWeb.....91
Εικόνα 4.4 Ιεράρχηση εννοιών.....93
Εικόνα 5.1 Παράδειγμα δομής ενός δέντρου αποφάσεων 99
Εικόνα 5.2 Δέντρο απόφασης με τα δομικά του στοιχεία.....101
Εικόνα 5.4 Παράδειγμα λειτουργίας ενός Random Forest.....109
Εικόνα 5.5 Αναπαράσταση της τεχνικής Boosting 110
Εικόνα 5.6 Παράδειγμα λειτουργίας ενός Random Forest.....111
Εικόνα 5.7 Αποτελέσματα σύγκρισης των αλγορίμων Decision Tree & Random Forest 117

Εικόνα 5.8 Παράδειγμα του ερωτηματολογίου που αφοράνε την αξιολόγηση του ξενοδοχείου....119
Εικόνα 5.9 Παράδειγμα του ερωτηματολογίου που αφοράνε την αξιολόγηση των εστιατορίων του ξενοδοχείου. 120
Εικόνα 5.10 Παράδειγμα του ερωτηματολογίου που αφοράνε τις προτιμήσεις των πελατών του ξενοδοχείου. 121
Εικόνα 5.11 Κώδικας αποστολής δεδομένων καταγραφής στον διακομιστή. 123
Εικόνα 5.12 Παράδειγμα ενός αντικειμένου clickstream που δημιουργείτε κατά την αλληλεπίδραση του χρήστη με την εφαρμογή 125
Εικόνα 5.13 Συνάρτηση για την δημιουργία των αντικειμένων clickstream και αποθήκευση των δεδομένων στο local storage 126
Εικόνα 5.14 Παράδειγμα ενέργειας του χρήστη και συγκεκριμένα εφαρμογής ενός φίλτρου αναζήτησης στην λίστα των παραλιών (αριστερό τμήμα) και δημιουργία του αντίστοιχου αντικειμένου clickstream (δεξιό τμήμα) 127
Εικόνα 5.15 Παράδειγμα δημιουργίας αντικειμένων clickstream στην σελίδα σημείων ενδιαφερόντων της εφαρμογής....128
Εικόνα 5.16 Παράδειγμα δημιουργίας αντικειμένων clickstream μέσω της χρήσης των συνδέσμων σε υπηρεσίες κοινωνικής δικτύωσης....130
Εικόνα 6.1 Παράδειγμα διαχωρισμού του τμήματος back – end από το front - end 131
Εικόνα 6.2 Το λογότυπο του Node.js 133
Εικόνα 6.3 Παράδειγμα βασικών οντοτήτων που σχετίζονται με το REST API. 135
Εικόνα 6.4 Το logo της γλώσσας προγραμματισμού Python 137
Εικόνα 6.5 Παράδεγμα δομής ενός εγγράφου στην MongoDB 141
Εικόνα 6.6 Το logo της βιβλιοθήκης Pandas. 147
Εικόνα 6.7 Παράδειγμα χρήσης της δομής δεδομένων Data Frame. 148
Εικόνα 6.8 Το logo της βιβλιοθήκης Scikit – Learn. 149
Εικόνα 6.9....151
Εικόνα 7.1 Στιγμιότυπο από τις βασικές λίστες περιεχομένου της εφαρμογής 153
Εικόνα 7.2 Βασική ροή της εφαρμογής....156
Εικόνα 7.3 Περιγραφή της αρχιτεκτονικής του συστήματος συστάσεων....157
Εικόνα 7.4 Η συνάρτηση retrieveDBdata χρησιμοποιείται για την διασύνδεση και λήψη δεδομένων από την βάση δεδομένων. 158
Εικόνα 7.5 Παράδειγμα κατηγορίας και υποκατηγορίας 160

Εικόνα 7.6 Η συνάρτηση update_subcatIds χρησιμοποιείται για την διασύνδεση με την βάση και την λήψη πληροφοριών με βάση το id των υποκατηγοριών.....161
Εικόνα 7.7 Η κατάλληλη μορφή δεδομένων για την αποθήκευση τους στην βάση δεδομένων.....167
Εικόνα 7.8 Παράδειγμα σύστασης στην κατηγορία φαγητών. 168
Εικόνα 7.9 Παράδειγμα σύστασης στην κατηγορία παραλιών 168
Εικόνα 7.10 Παράδειγμα σύστασης στην κατηγορία δραστηριοτήτων. 168

ΕΙΣΑΓΩΓΗ

Στις μέρες μας βιώνουμε μια νέα επανάσταση στην πληροφορική, η οποία έχει οδηγήσει στην δημιουργία του σύγχρονου ψηφιακού κόσμου, ο οποίος χαρακτηρίζεται από την εκρηκτική αύξηση της διαθέσιμης πληροφορίας και των ψηφιακών πόρων που είναι διαθέσιμοι για τους χρήστες. Το τεράστιο μέγεθος και είδος των διαθέσιμων πληροφοριών σημαίνει αυξημένη πολυπλοκότητα του περιβάλλοντος πληροφοριών, η οποία κατά συνέπεια δημιουργεί την ανάγκη εύρεσης και εφαρμογής αυτοματοποιημένων μεθόδων που θα βοηθήσουν τους χρήστες να επιλέγουν ευαίσθητα σε περιεχόμενο προϊόντα και υπηρεσίες που ταιριάζουν στις προτιμήσεις και τις ανάγκες τους. Σε αυτό το πλαίσιο, τα συστήματα συστάσεων αναδύονται ως κρίσιμη τεχνολογία που ανταποκρίνεται σε αυτήν την πρόκληση.

Τα συστήματα συστάσεων είναι πληροφοριακά συστήματα που έχουν ως στόχο να δημιουργήσουν αυτοματοποιημένα προφίλ χρηστών, να συμπεράνουν τις προτιμήσεις τους και να εξάγουν χρήσιμες πληροφορίες ώστε να προτείνουν στους χρήστες τα κατάλληλα προϊόντα ή υπηρεσίες που τους ταιριάζουν. Τα συστήματα συστάσεων εφαρμόζονται ευρέως σε κάθε τομέα ψηφιακής δραστηριότητας από το ηλεκτρονικό εμπόριο και την ψυχαγωγία έως την υγεία και την εκπαίδευση. Συμβάλλουν σημαντικά στη βελτίωση της εμπειρίας των χρηστών, την αύξηση της προσωποποίησης και την αύξηση της αποτελεσματικότητας των επιχειρήσεων.

Στόχος της παρούσας διπλωματικής εργασίας είναι η λεπτομερής μελέτη και ανάλυση των συστημάτων συστάσεων, συμπεριλαμβανομένων των αλγορίθμων, των μεθόδων αξιολόγησης και των προκλήσεων που αντιμετωπίζουν. Επιπλέον θα μελετηθεί η διαδικασία δημιουργίας προφίλ χρηστών συμπεριλαμβανομένων των διαφορετικών τύπων προφίλ ανάλογα με τους υπάρχοντες τρόπους αναπαράστασης καθώς και την διαδικασία δημιουργίας ενός προφίλ χρήστη.

1 Συστήματα συστάσεων

1.1 Τι είναι τα συστήματα συστάσεων

Τα συστήματα συστάσεων στοχεύουν να συμπεράνουν τα ενδιαφέροντα των χρηστών και να τους προτείνουν προϊόντα και υπηρεσίες που είναι πολύ πιθανό να είναι ενδιαφέροντα για αυτούς. Η ανάπτυξη αυτών των συστημάτων προέκυψε από το γεγονός ότι οι άνθρωποι στην καθημερινότητα τους, είτε ενσυναίσθητα είτε ασυναίσθητα λαμβάνουν υπόψιν διάφορες προτάσεις από το περίγυρό τους προκειμένου να προβούν σε ορισμένες αποφάσεις. Για παράδειγμα πολλές φορές, είτε ένας πωλητής που μας γνωρίζει είτε ο οικογενειακός και επαγγελματικός περίγυρος που γνωρίζει τις προτιμήσεις μας, έρχεται και μας προτείνει ποιο βιβλίο να διαβάσουμε, ποιο τουριστικό προορισμό να επισκεφθούμε καθώς και ποια τηλεοπτική σειρά θα πρέπει να επιλέξουμε και η οποία ταιριάζει με τις επιθυμίες και το χαρακτήρα μας. Είναι προφανές ότι τέτοιου είδους συστήματα είναι ιδιαίτερα ωφέλιμα για τους χρήστες, στους οποίους συστήνονται οι κατάλληλες αποφάσεις αλλά και για όσους αξιοποιούν τέτοια συστήματα καθώς ικανοποιούν τόσο τους πελάτες τους όσο και οργανώνουν καλύτερα τις παρεχόμενες υπηρεσίες.. Τα τελευταία χρόνια, η ανάπτυξη του διαδικτύου έχει αυξήσει σημαντικά τον όγκο και την πολυπλοκότητα των διαθέσιμων δεδομένων που συλλέγονται και καταγράφονται και επομένως των διαθέσιμων προς επεξεργασία πληροφοριών. Ο όγκος, η ποσότητα και τα χαρακτηριστικά των δεδομένων που συλλέγονται απαιτούν ιδιαίτερη προσπάθεια για την επεξεργασία τους, την ανάλυσή τους ώστε να επιλεγούν οι πιο σημαντικές και απαραίτητες πληροφορίες με βάση τις οποίες μπορούν να εξαχθούν τα βέλτιστα συμπεράσματα που θα αξιοποιηθούν στην διαδικασία λήψης αποφάσεων.

Τα συστήματα συστάσεων έρχονται να υποστηρίξουν την διαδικασία επίλυσης αυτού του προβλήματος βελτιώνοντας την ποιότητα και την διαδικασία λήψης αποφάσεων, εντοπίζοντας τα κατάλληλα προϊόντα ή υπηρεσίες για κάθε χρήστη μέσα από ένα μεγάλο σύνολο διαθέσιμων επιλογών. Στην σημερινή εποχή τα συστήματα αυτά έχουν ενσωματωθεί σε πολλές γνωστές διαδικτυακές υπηρεσίες και επιχειρήσεις πως η Amazon, YouTube, Yahoo, Netflix, TripAdvisor, Booking, Spotify κ.α.. Έχει αποδειχθεί ότι η χρήση των συστημάτων συστάσεων έχει οδηγήσει αυτές τις υπηρεσίες και επιχειρήσεις σε εκθετική αύξηση των επισκέψεων, των πωλήσεων καθώς και σε μεγαλύτερη ικανοποίηση των πελατών τους. Κύριος στόχος των συστημάτων

συστάσεων είναι να υποστηρίζουν όλους τους χρήστες, είτε διαθέτουν είτε όχι σημαντική προσωπική εμπειρία ή πληροφόρηση για να αξιολογήσουν τη πληθώρα των εναλλακτικών προσεγγίσεων, αποφάσεων, επιλογών και αντικειμένων που είναι στην διάθεσή τους.

Κάθε σύστημα συστάσεων συνήθως προτείνει στον χρήστη την επιλογή κάποιων αντικειμένων (ως αντικείμενο αποκαλούμε το αποτέλεσμα που εξάγει το σύστημα συστάσεων), το οποίο μπορεί να είναι μία ταινία ή τηλεοπτική σειρά, ένα ξενοδοχείο, ένας τουριστικός προορισμός ή και ένα τραγούδι. Συνήθως τα αποτελέσματα τα οποία εξάγονται από το σύστημα είναι εξατομικευμένες λίστες αντικειμένων οι οποίες είναι διαφορετικές για κάθε χρήστη ή για κάθε ομάδα χρηστών. Βέβαια υπάρχουν και οι μη εξατομικευμένες λίστες συστάσεων οι οποίες είναι πιο εύκολο να εξαχθούν, αλλά δεν λαμβάνουν υπόψη τις ιδιαίτερες προτιμήσεις και χαρακτηριστικά του κάθε χρήστη. Για να εξαχθούν οι εξατομικευμένες λίστες συστάσεων απαιτείται να δημιουργηθούν προφίλ χρηστών τα οποία βασίζονται στα ιδιαίτερα χαρακτηριστικά, την προσωπικότητα, τις επιλογές και τις ανάγκες των χρηστών.

Με την εκθετική επέκταση του διαδικτύου, το πλήθος, το μέγεθος και η διαφορετικότητα των δεδομένων και πληροφοριών που συλλέγονται, καταγράφονται και είναι διαθέσιμες στο ευρύ είναι τεράστιο οπότε είναι ιδιαίτερα δύσκολο εάν όχι αδύνατο για τον χρήστη να τις επεξεργαστεί και να τις χρησιμοποιήσει για τις ανάγκες τους. Οπότε έχουν προταθεί αυτοματοποιημένες μέθοδοι κατηγοριοποίησης και διαχωρισμού των περιττών από των χρήσιμων πληροφοριών. Οι μέθοδοι αυτοί έχουν την δυνατότητα με την εφαρμογή αυτοματοποιημένων τεχνικών να εξάγουν και να παρέχουν στον χρήστη ακριβώς αυτό που επιθυμεί και όχι άχρηστες ή άσχετες πληροφορίες.

Συνοψίζοντας τα Συστήματα Συστάσεων (ΣΣ) είναι συστήματα τα οποία αντιμετωπίζουν το πρόβλημα της υπερπληροφόρησης, φιλτράροντας τις ζωτικής σημασίας πληροφορίες, οι οποίες προέρχονται από μεγάλο όγκο παραγόμενων πληροφοριών αναφορικά με τις προτιμήσεις, τα ενδιαφέροντα ή ακόμη και την διαδικτυακή συμπεριφορά κάθε χρήστη σχετικά με κάποιο αντικείμενο που πιθανόν να τον ενδιαφέρει να επιλέξει.

1.2 Ιστορική αναδρομή

Το διαδίκτυο (Internet) με την μορφή που το γνωρίζουμε σήμερα, ξεκίνησε την ανάπτυξη του στην δεκαετία του 1950, σε διάφορα εργαστήρια πληροφορικής στις Ηνωμένες Πολιτείες της Αμερικής, το Ηνωμένο Βασίλειο και την Γαλλία (Kim, 2005). Το 1969 στάλθηκε το πρώτο μήνυμα μέσω του δικτύου ARPANET, από το Πανεπιστήμιο της Καλιφόρνια, Los Angeles (UCLA) προς το Εκπαιδευτικό Ίδρυμα του Stanford (SRI). Κατά την ανάπτυξη του διαδικτύου μπορούμε να ορίσουμε 3 ορόσημα. Αρχικά έχουμε το Web1.0, που περιλαμβάνει την πλειονότητα των ιστοσελίδων που δημιουργήθηκαν το διάστημα 1994 έως 2004, όπου παρείχαν στους χρήστες τη δυνατότητα ανάγνωσης πληροφοριών (στατικές ιστοσελίδες), χωρίς όμως να τους παρέχουν την δυνατότητα αλληλεπίδρασης και σχολιασμού (Jacksi and Abass, 2019).

Αυτό άλλαξε περίπου το 2004, με την εμφάνιση του Web 2.0, όπου οι χρήστες από παθητικοί αναγνώστες μετατράπηκαν σε ενεργούς θεατές με δυνατότητα διαμοιρασμού, αλληλεπίδρασης και σχολιασμού της παρατιθέμενης πληροφορίας, με χαρακτηριστικό εκπρόσωπο τις σελίδες κοινωνικής δικτύωσης. Χαρακτηριστική είναι η περιγραφή που έχει δοθεί από τον Τέρι Flew : «[Η] μετάβαση από τις προσωπικές ιστοσελίδες στα ιστολόγια και στις συλλογές τους, από τη δημοσίευση στη συμμετοχή, από το περιεχόμενο του ιστού ως αποτέλεσμα μεγάλης προκαταβεβλημένης επένδυσης σε μία τρέχουσα διαδραστική διαδικασία, και από τα συστήματα διαχείρισης περιεχομένου στις συνδέσεις με βάση ετικέτες επισήμανσης (tag)» (Flew, 2008).

Το Διαδίκτυο όπως εξελίσσεται σε συνδυασμό με την ολοένα αναπτυσσόμενη ψηφιακή τεχνολογία έχει δημιουργήσει μια τεράστια αποθηκευμένη βάση γνώσεων και πληροφοριών που ιδιαίτερα δύσκολο για χρήστες να προσδιορίσουν το στόχο της έρευνάς τους. Αυτή η αξιοποίηση του διαδικτύου σε ολοένα και περισσότερες ανθρώπινες δραστηριότητες οδήγησε στην εξέλιξη του παραδοσιακού Web σε αυτό που αποκαλούμε Web 2.0. Ο όρος Web 2.0 χρησιμοποιείται για να περιγράψει τη νέα γενιά του Παγκόσμιου Ιστού, όπου ο αρχικός παθητικός τρόπος παρουσίασης των πληροφοριών μεταλλάσσεται. Έννοιες όπως διαδραστικότητα, δυναμικό περιεχόμενο, συνεργασία, συνεισφορά και κοινότητα διαδραματίζουν πλέον πρωταγωνιστικό ρόλο.

Τα κύρια χαρακτηριστικά του Web 2.0 είναι τα εξής:

- Πλούσιες και διαδραστικές διασυνδέσεις (interfaces) χρηστών
- Συνεχής και άμεση ανανέωση των δεδομένων και του λογισμικού
- Προώθηση του δημοκρατικού χαρακτήρα του διαδικτύου με τους χρήστες να έχουν τον πρωταγωνιστικό ρόλο
- Δυνατότητα κατηγοριοποίησης του περιεχομένου από το χρήστη με σημασιολογικές έννοιες για ευκολότερη αναζήτηση της πληροφορίας.
- Δυνατότητα για ανοιχτή επικοινωνία, ανάδραση, διάχυση πληροφοριών, άμεση συγκέντρωση και εκμετάλλευση της γνώσης των χρηστών για διάφορα ζητήματα.
- Αμφίδρομη επικοινωνία του χρήστη με επιχειρήσεις ή οργανισμούς που μπορεί να έχει σαν αποτέλεσμα την επίδρασή του στη υιοθέτηση κατευθύνσεων και λήψη αποφάσεων.

Με την ραγδαία ανάπτυξη του Web 2.0 εμφανίστηκαν δύο νέες έννοιες που έχουν ως στόχο την καλύτερη οργάνωση και πρόβλεψη των αναγκών των χρηστών. Αυτές είναι η εξατομίκευση (Personalization) και η παραγωγή συστάσεων (Recommendation).

Η ιδέα χρήσης των υπολογιστών για την σύσταση των καλύτερων στοιχείων για τον χρήστη υπήρχε από την αρχή της πληροφορικής. Η πρώτη εφαρμογή της ιδέας των συστημάτων συστάσεων (ΣΣ) εμφανίστηκε το 1979 με την δημιουργία ενός συστήματος που ονομαζόταν Grundy έναν «ψηφιακό» βιβλιοθηκάριο ο οποίος πρότεινε στον χρήστη τι βιβλία να διαβάσει. Στις αρχές της δεκαετίας του 1990 κυκλοφόρησε το Tapestry το πρώτο ΣΣ προς εμπορική χρήση. Το Tapestry αποτελούσε μία εφαρμογή η οποία βοηθούσε τους ανθρώπους να βρουν τα αγαπημένα τους άρθρα. Ακολούθησαν και άλλες παρόμοιες εφαρμογές με το Tapestry όπως το Ringo, Bellcore, Jester, GroupLens. Το αποκορύφωμα των ΣΣ σημειώθηκε στα τέλη του 1990 από την Amazon χρησιμοποιώντας την μέθοδο Collaborative Filtering (CF) μία από τις πιο γνωστές μεθόδους στον τομέα των ΣΣ, η οποία χρησιμοποιήθηκε για να προτείνει στους χρήστες συγκεκριμένα προϊόντα και είχε ως συνέπεια την κατακόρυφη αύξηση των εσόδων της εταιρείας. Η Amazon έχει κατοχυρώσει την μέθοδο CF με δίπλωμα ευρεσιτεχνίας. Η επιτυχία της Amazon οδήγησε και άλλες εταιρείες στην χρήση ΣΣ. Μετά την επιτυχημένη εποχή στα τέλη της δεκαετίας του 1990, η βιομηχανία πρόσφερε γενναιόδωρη χρηματοδότηση για την υλοποίηση περαιτέρω έρευνας των ΣΣ. Ο πιο

δημοφιλής διαγωνισμός για τα ΣΣ πραγματοποιήθηκε από το Netflix, έναν πάροχο ροής βίντεο και ταινιών μέσω διαδικτύου. Ο διαγωνισμός διεξάχθηκε το 2006 και το έπαθλο ήταν 1 εκατομμύριο δολάρια σε όποιον μπορούσε να βελτιώσει τουλάχιστον κατά 10% την μηχανή συστάσεων της εταιρείας. Το 2010 το YouTube υλοποίησε και πρόσθεσε στον ιστότοπό του ένα ΣΣ.

1.3 Δομικά μέρη ενός συστήματος συστάσεων

1.3.1 Αντικείμενα (Items)

Ως αντικείμενα χαρακτηρίζουμε όσα προτείνει ένα σύστημα συστάσεων. Τα αντικείμενα μπορούν να χαρακτηρίζονται από την πολυπλοκότητα τους, την αξία ή τη χρησιμότητάς τους. Η αξία ενός αντικειμένου μπορεί να είναι θετική εάν είναι χρήσιμο για τον χρήστη ή αρνητική εάν το αντικείμενο δεν είναι κατάλληλο και ο χρήστης έλαβε λανθασμένη απόφαση κατά την επιλογή του. Ένας χρήστης όταν επιλέγει ένα αντικείμενο θα επιβαρύνεται πάντα με ένα κόστος το οποίο αποτελείται από το πραγματικό χρηματικό κόστος που έχει καταβάλει για το αντικείμενο και το γνωστικό κόστος της αναζήτησης του αντικειμένου.

Για παράδειγμα σε ένα ειδησεογραφικό σύστημα συστάσεων ο διαχειριστής του, θα πρέπει να λάβει υπόψη του την πολυπλοκότητα ενός ειδησεογραφικού στοιχείου, δηλαδή τη δομή του, την αναπαράσταση του κειμένου, και τη χρονικά εξαρτώμενη σημασία κάθε ειδησεογραφικού στοιχείου. Παράλληλα θα πρέπει να κατανοήσει ότι ακόμη και αν ο χρήστης δεν πληρώσει για την ανάγνωση των ειδήσεων, υπάρχει πάντα ένα γνωστικό κόστος που συνδέεται με την αναζήτηση και την ανάγνωση ειδήσεων. Εάν ένα επιλεγμένο στοιχείο είναι σχετικό με τον χρήστη, το κόστος αυτό κυριαρχείται από το όφελος της απόκτησης χρήσιμων πληροφορίων. Στην περίπτωση που το συγκεκριμένο στοιχείο δεν είναι σχετικό, η καθαρή αξία του για τον χρήστη και η σύσταση του είναι αρνητική. Για παράδειγμα στον τομέα χρηματοοικονομικών επενδύσεων και στην πώληση αυτοκινήτων, το πραγματικό χρηματικό κόστος των αντικειμένων αποτελεί σημαντικό στοιχείο που πρέπει να λαμβάνεται υπόψιν κατά την επιλογή της καταλληλότερης προσέγγισης για μια σύσταση.

Οι ειδήσεις, ιστοσελίδες, βιβλία, ταινίες αποτελούν αντικείμενα με χαμηλή πολυπλοκότητα και αξία. Από την άλλη κινητά τηλέφωνα, ηλεκτρονικοί υπολογιστές κλπ.. είναι αντικείμενα με μεγαλύτερη πολυπλοκότητα και αξία. Τα πιο πολύπλοκα

αντικείμενα στο πλαίσιο τις ερευνητικής προσπάθειας για δημιουργία Συστημάτων Συστάσεων, που έχουν εξεταστεί είναι οι χρηματοοικονομικές επενδύσεις, οι θέσεις εργασίας, τα ταξίδια και τα ασφαλιστήρια συμβόλαια.

1.3.2 Χρήστες (Users)

Οι χρήστες κάθε συστήματος συστάσεων διαθέτουν διαφορετικούς στόχους και χαρακτηριστικά. Προκειμένου να εξατομικεύσουν τις παραγόμενες συστάσεις και την αλληλεπίδραση μεταξύ των χρηστών και του υπολογιστή τα συστήματα συστάσεων χρησιμοποιούν μια σειρά πληροφοριών για κάθε χρήστη. Η δομή των καταγραφόμενων πληροφοριών γίνεται με διάφορους τρόπους και η επιλογή των πληροφοριών προς χρήση εξαρτάται από την τεχνική που θα χρησιμοποιηθεί για την δημιουργία των συστάσεων. Τα δεδομένα των χρηστών οδηγούν στην δημιουργία ενός μοντέλου για κάθε χρήστη. Το μοντέλο χρήστη σκιαγραφεί το προφίλ του χρήστη, δηλαδή κωδικοποιεί τις προτιμήσεις και τις ανάγκες του. Το μοντέλο χρήστη παίζει πάντα κεντρικό ρόλο σε ένα σύστημα συστάσεων καθώς χωρίς αυτό δεν θα ήταν δυνατή η εξατομίκευση των διαθέσιμων επιλογών που θα προκύψουν.

Για παράδειγμα στην περίπτωση του συνεργατικού φιλτραρίσματος οι χρήστες αναπαρίστανται ως μια λίστα που περιέχει τις αξιολογήσεις που έδωσε ο κάθε χρήστης για ορισμένα αντικείμενα. Σε ένα δημογραφικό σύστημα συστάσεων χρησιμοποιούνται κοινωνικοδημογραφικά χαρακτηριστικά όπως η ηλικία, το φύλο, το επάγγελμα και η εκπαίδευση.

Οι χρήστες μπορούν αξιοποιώντας τα δεδομένα να περιγράφουν πρότυπα συμπεριφοράς τους για παράδειγμα πληροφορίες αναφορικά με τις προτιμήσεις βασιζόμενοι στην πλοήγηση που κάνουν σε κάθε ιστοτόπο. Επιπλέον τα δεδομένα χρηστών μπορεί να περιλαμβάνουν σχέσεις μεταξύ των χρηστών, όπως το επίπεδο εμπιστοσύνης αυτών των σχέσεων μεταξύ των χρηστών. Το σύστημα συστάσεων μπορεί να χρησιμοποιήσει αυτές τις πληροφορίες προκειμένου να συστήσει προτιμήσεις σε χρήστες που προτιμήθηκαν από παρόμοιους ή έμπιστους χρήστες.

1.3.3 Συναλλαγές (Transactions)

Ως συναλλαγή αναφέρεται η καταγεγραμμένη αλληλεπίδραση του χρήστη και του συστήματος συστάσεων. Οι συναλλαγές είναι δεδομένα τα οποία περιέχουν σημαντικές πληροφορίες που παράγονται κατά τη διάρκεια της αλληλεπίδρασης ανθρώπου –

υπολογιστή και τα οποία είναι χρήσιμα για τον αλγόριθμο παραγωγής συστάσεων που χρησιμοποιεί το σύστημα. Για παράδειγμα ένα αρχείο καταγραφής συναλλαγών μπορεί να περιέχει μια αναφορά στο αντικείμενο που επιλέχθηκε από τον χρήστη και μια περιγραφή του αντικειμένου για τη συγκεκριμένη σύσταση. Σε ορισμένες περιπτώσεις η συναλλαγή μπορεί να περιλαμβάνει ρητή ανατροφοδότηση που ο χρήστης έχει παράσχει όπως για παράδειγμα η βαθμολογία για το επιλεγμένο αντικείμενο.

Οι αξιολογήσεις αποτελούν την πιο δημοφιλή μορφή δεδομένων συναλλαγών που συλλέγει ένα σύστημα συστάσεων. Οι αξιολογήσεις μπορεί να συλλέγονται ρητά η σιωπηρά. Στη ρητή συλλογή αξιολογήσεων, οι χρήστες καλούνται να εκφράσουν τη γνώμη τους για ένα στοιχείο σε μία κλίμακα αξιολόγησης.

Οι αξιολογήσεις διαθέτουν διάφορες μορφές όπως:

- Δυαδικές αξιολογήσεις που μοντελοποιούν επιλογές στις οποίες ο χρήστης καλείται να αποφασίσει αν ένα συγκεκριμένο στοιχείο είναι καλό ή κακό.
- Αριθμητικές αξιολογήσεις όπως για παράδειγμα η κλίμακα των αστεριών από 1 έως το 5.
- Τακτικές αξιολογήσεις όπως “συμφωνώ απόλυτα”, “διαφωνώ”, “ουδέτερο”, κλπ..

Στις συναλλαγές που συλλέγουν σιωπηρές αξιολογήσεις, το σύστημα έχει ως στόχο να συμπεράνει τη γνώμη του χρήστη με βάση τις ενέργειές του. Για παράδειγμα, εάν ένας χρήστης εισάγει τη λέξη-κλειδί "γιόγκα" στο Amazon.com, θα του παρασχεθεί ένας μακρύς κατάλογος βιβλίων, τα οποία περιέχουν ή συσχετίζονται με την συγκεκριμένη λέξη. Σε αντάλλαγμα, ο χρήστης μπορεί να κάνει κλικ σε ένα συγκεκριμένο βιβλίο της λίστας για να λάβει πρόσθετες πληροφορίες.

1.4 Χαρακτηριστικά και οφέλη συστημάτων συστάσεων

Βασικοί τομείς οι οποίοι έχουν συνεισφέρει στην δημιουργία των συστημάτων συστάσεων είναι:

- **Ανάκτηση πληροφοριών**: Τα συστήματα ανάκτησης πληροφοριών είναι τα συστήματα τα οποία έχουν την δυνατότητα να αναζητήσουν, να ταξινομήσουν και να ανακτήσουν πολλά δεδομένα τα οποία εκτός από κείμενο μπορούν να περιλαμβάνουν και άλλες μορφές όπως εικόνα, βίντεο, ήχος κ.α.. Επιπλέον τα

συστήματα αυτά δίνουν την δυνατότητα στον χρήστη να βελτιστοποιήσει την λίστα των αποτελεσμάτων.

- **Η εξατομίκευση**: Η εξατομίκευση είναι μία μορφή μάρκετινγκ η οποία προσπαθεί να δημιουργήσει προϊόντα τα οποία είναι ιδανικά για τον κάθε χρήστη ξεχωριστά. Αυτό συμβαίνει έπειτα από αλληλεπίδραση με τον κάθε χρήστη και από τον έλεγχο του ιστορικού των ενεργειών του χρήστη.
- **Διαχείριση εμπιστοσύνης**: Στο διαδίκτυο υπάρχουν πολλές πληροφορίες οι οποίες προέρχονται από άτομα τα οποία μεροληπτούν πάνω σε ορισμένα θέματα. Έτσι είναι σημαντικό να είναι γνωστή η πηγή της πληροφορίας έτσι ώστε να μπορέσει να κρίνει ο ενδιαφερόμενος αν οι πληροφορίες είναι έγκυρες ή όχι. Ο όγκος των δεδομένων στο διαδίκτυο είναι τόσο μεγάλος που είναι πολύ δύσκολο να αξιολογηθούν όλες οι πηγές που υπάρχουν. Έτσι μια τεχνική που χρησιμοποιείται για να εξακριβωθεί αν μια πηγή είναι ορθή είναι η εμπιστοσύνη σε άτομα τα οποία είναι γνωστά.

1.4.1 Οφέλη ως προς τους παρόχους υπηρεσιών

Η ωφελιμότητα χρήσης τέτοιων συστημάτων από παρόχους υπηρεσιών έγκειται σε:

- **Αύξηση του αριθμού των πωληθέντων αντικειμένων**. Αυτή είναι πιθανώς η πιο σημαντική λειτουργία για ένα εμπορικό Σύστημα Συστάσεων, δηλ. οι πάροχοι υπηρεσιών να είναι σε θέση να πουλήσουν ένα πρόσθετο σύνολο αντικειμένων σε σύγκριση με εκείνα που συνήθως πωλούνται χωρίς καμία σύσταση. Αυτό επιτυγχάνεται επειδή τα στοιχεία για τα οποία γίνεται σύσταση είναι πιθανό να ταιριάζουν στις ανάγκες και τις επιθυμίες του κάθε χρήστη
- **Πώληση διαφοροποιημένων ειδών**. Μια άλλη σημαντική λειτουργία ενός Συστήματος Συστάσεων είναι να επιτρέψει στο χρήστη να επιλέξει στοιχεία που μπορεί να είναι δύσκολο να βρεθούν χωρίς συγκεκριμένες συστάσεις.
- **Αύξηση της ικανοποίησης των χρηστών**. Ένα καλά σχεδιασμένο Σύστημα Συστάσεων έχει ως κύριο σκοπό να βελτιώσει την εμπειρία του χρήστη στη χρήση ενός ιστοτόπου ή μιας εφαρμογής.
- **Αύξηση της αφοσίωσης των χρηστών.** Ένας χρήστης πρέπει να είναι πιστός σε κάποιον ιστότοπο και όταν τον επισκέπτεται ξανά , αυτός να τον αναγνωρίζει ως παλιό πελάτη και τον αντιμετωπίζει ως πολύτιμο επισκέπτη.

- **Καλύτερη κατανόηση των επιθυμιών του χρήστη.** Μια άλλη σημαντική λειτουργία ενός Συστήματος Συστάσεων, η οποία μπορεί να χρησιμοποιηθεί σε πολλές άλλες εφαρμογές, είναι η περιγραφή των προτιμήσεων του χρήστη, η οποία είτε συλλέγεται ρητά είτε προβλέπεται από το σύστημα

1.4.2 Οφέλη ως προς τους χρήστες

Όσο αναφορά τους χρήστες, η χρήση Συστήματος Συστάσεων είναι επιθυμητή και ωφέλιμη λόγω της αποτελεσματικότητας που έχουν στην υποστήριξη των στόχων του κάθε χρήστη:

- **Εύρεση καλών αντικειμένων.** Δηλαδή, να προτείνονται στον χρήστη αντικείμενα ταξινομημένα σε λίστα μαζί με τις προβλέψεις τους, και τα οποία να ανταποκρίνονται σε τι πραγματικά τα ήθελε.
- **Εύρεση όλων των καλών αντικειμένων.** Συστήνει όλα τα στοιχεία που μπορούν να ικανοποιήσουν τις ανάγκες των χρηστών.
- **Σχολιασμός που συνοδεύει τις προτάσεις**. Λαμβάνοντας υπόψη κάποιες προϋπάρχουσες προτάσεις, να τονίζονται ορισμένες από αυτές ανάλογα με τις μακροπρόθεσμες προτιμήσεις του χρήστη.
- **Σύσταση μιας ακολουθίας.** Αντί το Σύστημα Συστάσεων να δημιουργήσει μια μόνο σύσταση, να προτείνει μια σειρά στοιχείων που ικανοποιούν το χρήστη στο σύνολό τους.
- **Σύσταση συνδυασμού.** Δηλαδή, να προτείνεται στο χρήστη μια σειρά αντικειμένων που μπορούν να ταιριάξουν μεταξύ τους.

Κατά συνέπεια, ένα Σύστημα Συστάσεων πρέπει να ισορροπήσει τις ανάγκες ενός παρόχου αλλά και του χρήστη και να προσφέρει μια υπηρεσία πολύτιμη και για τους δύο.

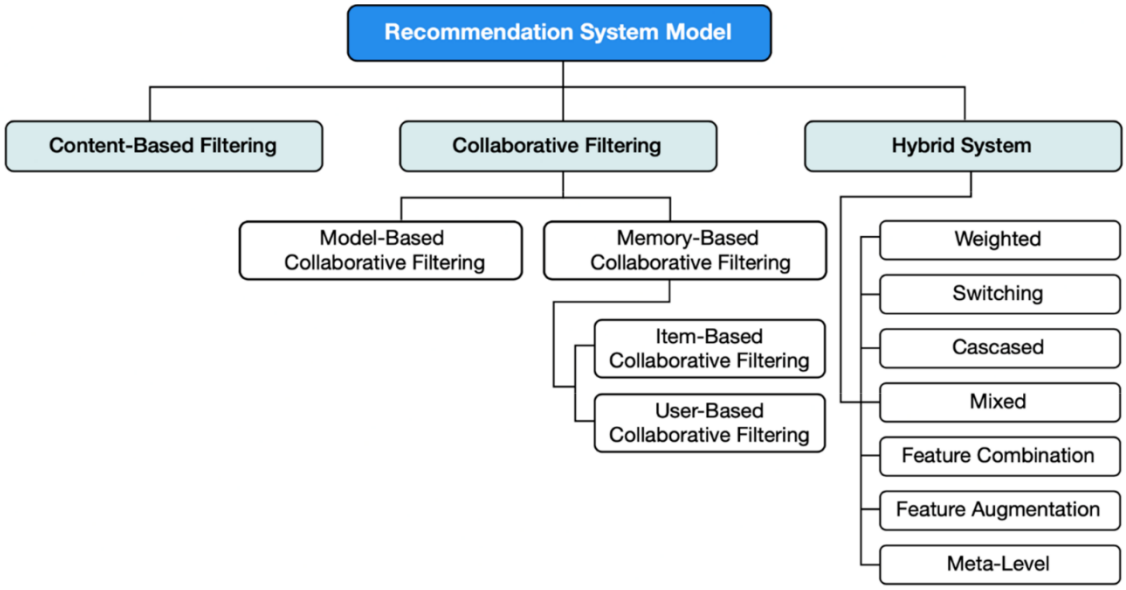
1.5 Κατηγορίες συστημάτων συστάσεων

Τα Συστήματα Συστάσεων διακρίνονται σε κατηγορίες, ανάλογα με τον τρόπο που εξάγονται οι συστάσεις που θα προταθούν στους χρήστες (Patel et al., 2017).

- **Συστήματα Συστάσεων βασιζόμενα στο περιεχόμενο (Content Based)**: Η τεχνική αυτή εστιάζει στα χαρακτηριστικά των αντικειμένων και βασίζεται σε πληροφορίες που συλλέγονται από το περιεχόμενο ιστοσελίδων, εκδόσεων, ειδήσεων σχετικά με αυτά. Με βάσει τα συλλεγόμενα στοιχεία δομείται το

προφίλ του χρήστη το οποίο στη συνέχεια χρησιμοποιείται από το σύστημα συστάσεων για να προβλέψει και να προτείνει αντικείμενα στο χρήστη.

- **Συστήματα Συστάσεων βασιζόμενα στη συνεργασία (Collaborative Systems)**: Αυτά τα συστήματα αξιοποιούν την συμπεριφορά του παρελθόντος που είχε ο χρήστης προκειμένου να εξάγουν αποφάσεις σχετικά με το εάν ένας χρήστης θα βρει χρήσιμες τις πληροφορίες ή όχι. Δηλαδή, συγκρίνουν αν οι χρήστες βρήκαν χρήσιμες τις πληροφορίες σε παρόμοιες αναζητήσεις στο παρελθόν και λαμβάνουν ανάλογες αποφάσεις σχετικά με τη χρησιμότητα ή μη.
- **Συστήματα Συστάσεων βασιζόμενα στη γνώση (Knowledge - Based Systems):** Τα Συστήματα Συστάσεων βασίζονται στη γνώση που έχουν αποκτήσει για τους χρήστες, για τη συμπεριφορά τους, για τις προτιμήσεις τους, για τα αντικείμενα, και για τους διαθέσιμους τρόπους που υπάρχουν για να καλύψουν τις ανάγκες των χρηστών και ανάλογα παρέχουν συστάσεις στους χρήστες.
- **Υβριδικά συστήματα συνεργασίας (Hybrid Systems):** Τα υβριδικά Συστήματα Συστάσεων βασίζονται στον συνδυασμό των προηγούμενων κατηγοριών σε μια προσπάθεια για βελτίωση του τελικού αποτελέσματος.



Εικόνα 1.1 Κατηγοριοποίηση των συστημάτων συστάσεων

1.5.1 Συστήματα βασισμένα στο περιεχόμενο (Content – based filtering)

Τα συστήματα συστάσεων που βασίζονται στο περιεχόμενο είναι συστήματα που χρησιμοποιούν φιλτράρισμα βάσει περιεχομένου, δηλαδή αναλύουν ένα σύνολο πληροφοριών και περιγραφών σχετικά με τα χαρακτηριστικά των αντικειμένων τα οποία έχουν προηγουμένως βαθμολογηθεί από τους χρήστες. Στη συνέχεια δημιουργούν ένα μοντέλο ή προφίλ των ενδιαφερόντων του χρήστη με βάση αυτά τα χαρακτηριστικά. Αυτό το προφίλ συνιστά μια δομημένη παρουσίαση των ενδιαφερόντων του χρήστη και έχει τη δυνατότητα να προσαρμόζεται και να βελτιώνεται όταν ο χρήστης αλλάζει ή επεκτείνει τα ενδιαφέροντά του, όπως για παράδειγμα στην περίπτωση ύπαρξης νέων αντικειμένων που μπορεί να τον ενδιαφέρουν. Η διαδικασία σύστασης συνίσταται βασικά στην αντιστοίχιση των χαρακτηριστικών του προφίλ των χρηστών σε σχέση με τα χαρακτηριστικά ενός αντικειμένου. Το αποτέλεσμα είναι μία σχετική απόφαση για το επίπεδο του ενδιαφέροντος του κάθε χρήστη για το αντικείμενο αυτό. Με βάση την συγκεκριμένη τεχνική προτείνονται στο χρήστη αντικείμενα όμοια με αυτά που είχε προτιμήσει στο παρελθόν. Οι αξιολογήσεις του συστήματος είναι ατομικές από κάθε χρήστη και δεν έχει καμία εξάρτηση με όλους τους υπόλοιπους. Το κύριο σημείο εδώ είναι το ίδιο το αντικείμενο. Δηλαδή τα χαρακτηριστικά που το εκφράζουν και το πώς το ένα αντικείμενο διαφέρει από τα υπόλοιπα. Κάθε αντικείμενο είναι διαφορετικό έχοντας διαφορετικά γνωρίσματα που το περιγράφουν.

Τα συστήματα φιλτραρίσματος που βασίζονται στο περιεχόμενο επιλέγουν αντικείμενα με βάση τη συσχέτιση μεταξύ του περιεχομένου των αντικειμένων και των προτιμήσεων του χρήστη, ενώ τα συνεργατικά συστήματα φιλτραρίσματος επιλέγουν αντικείμενα με βάση τη συσχέτιση μεταξύ των χρηστών.

Προκειμένου να καταγραφεί τι προτιμά κάποιος χρήστης χρησιμοποιούνται άμεσες ή έμμεσες πληροφορίες - ανατροφοδότηση. Η άμεση ανατροφοδότηση αποτελεί την ρητή βαθμολόγηση από τον χρήστη ενός αντικειμένου χρησιμοποιώντας μια κλίμακα, ενώ η την λειτουργία της έμμεσης ανατροφοδότησης βασίζεται στις προτιμήσεις του χρήστη που συνάγονται από την παρατήρηση των πράξεων του.

Μια βελτιωμένη μέθοδο ανάκτησης της πληροφορίας,βασίζεται στο προφίλ των χρηστών. Το προφίλ των χρηστών συνίσταται στις προτιμήσεις και τις ανάγκες των χρηστών. Για τη δημιουργία των προφίλ των χρηστών εφαρμόζονται μέθοδοι που περιλαμβάνουν (συλλογή πληροφοριών για προτιμήσεις, ανάγκες, κλπ. Στη

συνέχεια εφαρμόζονται αλγόριθμοι μηχανικής μάθησης (machine -learning algorithms) με τη βοήθεια των οποίων το σύστημα μαθαίνει το προφίλ του χρήστη. Το κομβικό σημείο αυτής της προσέγγισης είναι η επεξεργασία του περιεχομένου που περιγράφει τα αντικείμενα που πρόκειται να προταθούν. Τα αντικείμενα ενδέχεται να είναι διαφορετικά μεταξύ τους, και αυτό είναι σε συνάρτηση με το πλήθος τόσο των ίδιων των αντικειμένων όσο και των χαρακτηριστικών τους που μπορεί να διαφέρουν στο είδος και στην ποικιλία (Javed et al., 2021).

Συγκρίνοντας την τεχνικής φιλτραρίσματος βασιζόμενη στο περιεχόμενο με την τεχνική του συνεργατικού φιλτραρίσματος διαπιστώνεται ότι η πρώτη τεχνική πλεονεκτεί στα ακόλουθα σημεία:

- **Ανεξαρτησία από τους υπόλοιπους χρήστες:** Σε αντίθεση με το συνεργατικό φιλτράρισμα δεν απαιτούνται δεδομένα από άλλους χρήστες του συστήματος πέρα από τον ενεργό – τρέχον χρήστη. Μόλις ο τρέχον χρήστης περιηγηθεί ή αναζητήσει κάποια αντικείμενα, το σύστημα συστάσεων θα είναι σε θέσει να ξεκινήσει τις αντίστοιχες συστάσεις. Αυτό το πλεονέκτημα είναι ιδανικό για επιχειρήσεις που δεν διαθέτουν μεγάλο αριθμό χρηστών ή επιχειρήσεις που έχουν μεγάλο αριθμό πελατών αλλά μικρό αριθμό αλληλοεπιδράσεων των χρηστών σε ορισμένες κατηγορίες.
- **Οι συστάσεις είναι σε μεγάλο βαθμό σχετικές με τον χρήστη:** Οι συστάσεις που πραγματοποιούνται μπορούν να προσαρμοστούν σε μεγάλο βαθμό στα ενδιαφέροντα του χρήστη, συμπεριλαμβανομένων των συστάσεων εξειδικευμένων αντικειμένων καθώς αυτή η μεθοδολογία βασίζεται στην αντιστοίχιση των χαρακτηριστικών ή των ιδιοτήτων του αντικειμένου με το προφίλ του χρήστη.
- **Δεν επηρεάζεται από το πρόβλημα της ψυχρής εκκίνησης:** Παρόλο που και το φιλτράρισμα βάσει περιεχομένου χρειάζεται κάποιες αρχικές εισροές από τους χρήστες προκειμένου να ξεκινήσουν οι συστάσεις, η ποιότητα αυτών των αρχικών – πρώιμων συστάσεων είναι γενικά καλύτερη από ένα σύστημα συνεργατικού φιλτραρίσματος το οποίο απαιτεί την προσθήκη και συσχέτιση μεγάλου βαθμού δεδομένων προτού βελτιστοποιηθεί.
- **Διαφάνεια προς τους χρήστες:** Μέσω του φιλτραρίσματος βάσει περιεχομένου παρέχεται μια αίσθηση ανοιχτότητας προς τον χρήστη, ενισχύοντας έτσι το επίπεδο

εμπιστοσύνης του στις προσφερόμενες συστάσεις. Εισάγοντας τα χαρακτηριστικά ή τις περιγραφές ενός αντικειμένου, είναι δυνατόν να εξηγηθεί πώς λειτουργεί το σύστημα συστάσεων και για ποιο λόγο τοποθετεί το συγκεκριμένο αντικείμενο στη λίστα προτάσεων. Αυτά τα χαρακτηριστικά είναι δείκτες που μπορεί κανείς να συμβουλευτεί για να προσδιορίσει το ποσοστό «εμπιστοσύνης» της σύστασης που γίνεται.

- **Ευκολία στη υλοποίηση τους:** Οι τεχνικές των δεδομένων που χρησιμοποιούνται στην τεχνική φιλτραρίσματος βάσει περιεχομένου είναι σχετικά απλές σε σχέση με συστήματα συνεργατικού φιλτραρίσματος που βασίζονται στην ομοιότητα των χρηστών. Ουσιαστικά η πραγματική και απαιτητική διαδικασία ενός συστήματος βάσει περιεχομένου είναι η ανάθεση των χαρακτηριστικών.

Αντίστοιχα, τα συστήματα φιλτραρίσματος περιεχομένου μειονεκτούν σε σχέση με τα συστήματα συνεργατικού φιλτραρίσματος ως προς τα ακόλουθα:

- **Πρόβλημα σε περίπτωση επεκτασιμότητας:** Κάθε φορά που ένα νέο προϊόν ή υπηρεσία προστίθεται στο σύστημα τα χαρακτηριστικά του θα πρέπει να επισημαίνονται, το οποίο είναι μια επίπονη διαδικασία καθιστώντας την επεκτασιμότητα δύσκολη και χρονοβόρα.
- **Λανθασμένα ή ασυνεπή χαρακτηριστικά:** Τα συστήματα που βασίζονται στο περιεχόμενο βασίζονται σε χαρακτηριστικά τα οποία επισημαίνονται από ειδικούς πάνω σε αυτά. Ενδεχομένως εκατομμύρια στοιχεία μπορεί να χρειάζονται χαρακτηριστικά και, δεδομένου ότι τα χαρακτηριστικά μπορεί να είναι υποκειμενικά, πολλά μπορεί να έχουν επισημανθεί εσφαλμένα. Μια διαδικασία που διασφαλίζει ότι τα χαρακτηριστικά εφαρμόζονται με συνέπεια και ακρίβεια είναι υψίστης σημασίας. Διαφορετικά, ένα σύστημα συστάσεων βασισμένο στο περιεχόμενο δεν θα λειτουργήσει όπως προβλέπεται.
- **Μεγάλος βαθμός εξειδίκευσης:** Ένα σύστημα που βασίζεται στο περιεχόμενο δεν διαθέτει κάποια εξειδικευμένη τεχνική προκειμένου να μπορέσει να διαχειριστεί το απροσδόκητο. Το σύστημα προτείνει στοιχεία που έχουν υψηλή βαθμολογία σε σύγκριση με το προφίλ του χρήστη και, επομένως, ο χρήστης κατηγοριοποιείται με βάση αντικείμενα παρόμοια με αυτά που έχουν ήδη αξιολογηθεί.

- **Περιορισμένη ανάλυση δεδομένων:** Οι τεχνολογίες που βασίζονται στο περιεχόμενο έχουν φυσικούς περιορισμούς ως προς τον αριθμό και τους τύπους χαρακτηριστικών που σχετίζονται με τα αντικείμενα που προτείνουν. Συχνά είναι απαραίτητο ένα πεδίο γνώσης, π.χ., για τις συστάσεις ταινιών, το σύστημα πρέπει να γνωρίζει τους ηθοποιούς και σκηνοθέτες.

1.5.2 Συστήματα βασισμένα στην συνεργασία (Collaborative filtering)

Τα Συστήματα Συστάσεων που βασίζονται στη συνεργασία χρησιμοποιούν τη μέθοδο του συνεργατικού φιλτραρίσματος (Collaborative Filtering). Το συνεργατικό φιλτράρισμα είναι ένας δημοφιλής αλγόριθμος που βασίζει τις προβλέψεις και τις συστάσεις του στη συμπεριφορά και τη βαθμολόγηση αντικειμένων από άλλους χρήστες του συστήματος. Η βασική υπόθεση της μεθόδου αυτής είναι πως οι γνώμες των άλλων χρηστών μπορούν να συγκεντρωθούν και να οργανωθούν με τρόπο που θα εξάγεται μία λογική υπόθεση για την προτίμηση του ενεργού χρήστη. Οι τεχνικές που ανήκουν σε αυτή τη κατηγορία χρησιμοποιούν τις βαθμολογίες των χρηστών για να καθορίσουν τη σχέση μεταξύ τους και να εξάγουν προβλέψεις για βαθμολογίες που θα έδινε κάθε χρήστης σε κάποιο νέο αντικείμενο. Οι τεχνικές συνεργατικού φιλτραρίσματος χρησιμοποιούνται σε πολλές εφαρμογές συστημάτων συστάσεων. Δημοφιλείς εφρμογές που χρησιμοποιούν τέτοιες τεχνικές είναι οι LinkedIn, Facebook, Twitter και Amazon.

Η πλειονότητα των μεθόδων συνεργατικού φιλτραρίσματος λειτουργούν δημιουργώντας προβλέψεις για την προτίμηση του χρήστη και έπειτα παράγουν συστάσεις βαθμολογώντας υποψήφια αντικείμενα από τις εκτιμώμενες προτιμήσεις. Αυτή η μέθοδος ανήκει στην κατηγορία συνεργατικού φιλτραρίσματος με βάση το χρήστη. Χαρακτηριστικά, για κάθε χρήστη βρίσκεται ένα σύνολο «πλησιέστερων χρηστών / γειτόνων» με τους οποίους, με βάση τις μέχρι τώρα εκτιμήσεις υπάρχει ο ισχυρότερος συσχετισμός. Τα αποτελέσματα για τα άγνωστα στοιχεία προβλέπονται με βάση συνδυασμό αποτελεσμάτων που είναι γνωστά από τους «πλησιέστερους γείτονες». Το σύστημα μπορεί να προτείνει αντικείμενα (όπως βιβλία, μουσική κ.λπ.) στους χρήστες βασισμένο στις εκτιμήσεις των στοιχείων, αντί του περιεχομένου των στοιχείων, γεγονός που μπορεί να βελτιώσει την ποιότητα των συστάσεων. Εναλλακτικά, υπάρχει και το φιλτράρισμα με βάση το αντικείμενο, π.χ. ο χρήστης που αγόρασε το χ προϊόν είναι πιθανό να αγοράσει και το ψ. Με τη μέθοδο αυτή δημιουργείται ένας πίνακας αντικείμενο-αντικείμενο και καθορίζει σχέσεις μεταξύ των αντικειμένων δημιουργώντας

ζευγάρια από αυτά. Στη συνέχεια, συνάγει τις προτιμήσεις του ενεργού χρήστη εξετάζοντας τον πίνακα και ταιριάζοντας τις πληροφορίες για το χρήστη.

Στις τεχνικές αυτής της προσέγγισης υπάρχει μια λίστα χρηστών u {u1, u2 .. uN}, μια λίστα με αντικείμενα i {i1, i2 .. iN} και μια λίστα με αντικείμενα lui τα οποία έχει αξιολογήσει ο χρήστης. Οι αξιολογήσεις του χρήστη μπορούν να είναι ρητές, π.χ. ο χρήστης αξιολογεί ένα στοιχείο σε μια κλίμακα, είτε σιωπηρές όπως αναζητήσεις ή αγορές/ επιλογές που έκανε ο χρήστης (Xiaoyuan & Taghi, 2009).

Το συνεργατικό φιλτράρισμα χρησιμοποιεί τρεις κύριες τεχνικές:

- Τεχνική που βασίζεται στη μνήμη (memory-based)
- Τεχνική που βασίζεται στο μοντέλο (model-based)
- Την υβριδική τεχνική.

Τεχνική που βασίζεται στη μνήμη (Memory Based)

Η τεχνική αυτή χρησιμοποιεί τις πληροφορίες βαθμολόγησης των χρηστών για να υπολογίσει την ομοιότητα μεταξύ χρηστών ή αντικειμένων και να δημιουργήσει τις συστάσεις της. Τα κυριότερα πλεονεκτήματα της προσέγγισης που βασίζεται στη μνήμη είναι η επεξηγηματικότατα των αποτελεσμάτων, η οποία είναι μία πολύ σημαντική πλευρά των συστημάτων συστάσεων, η ευκολία στη χρήση και στην εφαρμογή της καθώς και η ανεξαρτησία της από το περιεχόμενο του αντικειμένου που προτείνεται.

Αντίθετα, το κυριότερο μειονέκτημα της είναι η μειωμένη επίδοση όταν τα δεδομένα που έχει στη διάθεσή της είναι λιγοστά. Εάν ένα νέο στοιχείο εμφανιστεί στη βάση δεδομένων, δεν υπάρχει κανένας τρόπος να συσταθεί σε έναν χρήστη έως ότου να ληφθούν περισσότερες πληροφορίες για αυτό μέσω μιας άλλης εκτίμησης χρήστη είτε διευκρινίζοντας ποια άλλα στοιχεία είναι παρόμοια με αυτό. Το πρόβλημα είναι γνωστό με την ονομασία «ψυχρή εκκίνηση» (cold-start) ή πρόβλημα της «πρώτης εκτίμησης» (first-rater), καθώς συστάσεις απαιτούνται για τα στοιχεία που κανένας χρήστης δεν έχει εκτιμήσει ακόμα.

Τεχνική που βασίζεται στο μοντέλο (Model Based)

Η τεχνική που βασίζεται στο μοντέλο χρησιμοποιεί τις αξιολογήσεις των χρηστών ως σύνολο εκπαίδευσης αλγόριθμων μηχανικής εκμάθησης για τη δημιουργία ενός μοντέλου πρόβλεψης. Στη συνέχεια, το μοντέλο χρησιμοποιείται για την πρόβλεψη της βαθμολογίας για ένα αντικείμενο. Υπάρχουν πολλοί τρόποι για πρόβλεψη πραγματικών δεδομένων, οι σημαντικότεροι από τους οποίους είναι η συσταδοποίηση και οι πιθανοτικοί αλγόριθμοι. Η τεχνική που βασίζεται σε μοντέλα αντιμετωπίζει το πρόβλημα των λιγοστών δεδομένων καλύτερα από τη memory based τεχνική, λύνει το πρόβλημα της κλιμάκωσης σε περίπτωση μεγάλων συνόλων δεδομένων και βελτιώνει την απόδοση της πρόβλεψης. Στα μειονεκτήματά της συγκαταλέγονται η δυσκολία του χτισίματος του μοντέλου και της εξήγησης των προβλέψεων. Όπως σε όλες τις μεθόδους απαιτείται να βρεθεί η βέλτιστη ισορροπία μεταξύ της απόδοσης της πρόβλεψης και της κλιμακοσιμότητας.

Ακετές εφαρμογές χρησιμοποιούν ένα συνδυασμό των τεχνικών που βασίζονται στη μνήμη και των τεχνικών που βασίζονται στο μοντέλο, που καλείται υβριδική τεχνική. Ένα υβριδικό σύστημα που συνδυάζει τις τεχνικές Α και Β επιχειρεί να χρησιμοποιήσει τα πλεονεκτήματα της Α για να διορθώσει τα μειονεκτήματα της Β. Επομένως, προβλήματα όπως αυτό της έλλειψης πληροφοριών μπορούν να ξεπεραστούν και να βελτιωθεί έτσι η απόδοση του συστήματος, όμως αυξάνεται και η πολυπλοκότητα του. Συνήθως τα περισσότερα εμπορικά συστήματα σύστασης είναι υβριδικά όπως αυτό που χρησιμοποιεί η Google.

Σύμφωνα με τον Koren (Y.Koren, 2008), δυο από τις πιο επιτυχημένες προσεγγίσεις συνεργατικού φιλτραρίσματος είναι:

1. Τα μοντέλα λανθάνοντος παράγοντα (latent factor models). Σ' αυτά η βασική υπόθεση είναι ότι υπάρχει μια ακαθόριστη αναπαράσταση χρηστών και προϊόντων με λίγες διαστάσεις, όπου η αντιστοίχιση χρηστών-προϊόντων μπορεί να μοντελοποιηθεί με ακρίβεια. Αυτά τα μοντέλα δημιουργούν προφίλ και στα προϊόντα και στους χρήστες.
2. Τα μοντέλα γειτνίασης (neighborhood models) που αναλύουν ομοιότητες μεταξύ των προϊόντων ή των χρηστών. Οι μέθοδοι γειτνίασης βασίζονται στον υπολογισμό των σχέσεων μεταξύ προϊόντων ή εναλλακτικά μεταξύ χρηστών. Μια προσέγγιση

προσανατολισμένη στα προϊόντα αξιολογεί την προτίμηση ενός χρήστη για ένα προϊόν με βάση την κατάταξη που διενεργεί ο ίδιος χρήστης για παρόμοια προϊόντα. Κατά μια έννοια, αυτές οι μέθοδοι καθιστούν ικανούς τους χρήστες να αξιολογούν μια πληθώρα προϊόντων και να προτιμούν αυτά που είναι πιο κοντά στα ενδιαφέροντά τους. Έτσι, δεν υπάρχει πλέον ανάγκη να συγκρίνουμε χρήστες με προϊόντα αλλά απευθείας συσχετίζουμε προϊόντα με προϊόντα.

Τα μοντέλα λανθάνοντος παράγοντα (latent factor models) όπως είναι τα μοντέλα singular value decomposition (SVD), περιλαμβάνουν μια εναλλακτική προσέγγιση και μετασχηματίζουν και τα προϊόντα και τους χρήστες στον ίδιο χώρο καθιστώντας και τα δυο μέρη απευθείας συγκρίσιμα. Ο λανθάνων χώρος (latent space) προσπαθεί να εξηγήσει τις αξιολογήσεις χαρακτηρίζοντας και τα προϊόντα και τους χρήστες, σύμφωνα με τους παράγοντες που προκύπτουν αυτόματα από την ανατροφοδότηση του χρήστη. Π.χ. όταν τα προϊόντα είναι ταινίες, οι παράγοντες μπορεί να μετρούν προφανείς διαστάσεις, όπως κωμωδία ή δράμα, ποσότητα δράσης, προσανατολισμός στα παιδιά ή λιγότερο καθορισμένες διαστάσεις, όπως το βάθος της ανάπτυξης του χαρακτήρα ή εντελώς άρρητες διαστάσεις.

Τα μοντέλα γειτνίασης (neighborhood models) είναι πιο αποτελεσματικά στην ανίχνευση τοπικών σχέσεων. Βασίζονται σε λίγες σημαντικές σχέσεις γειτνίασης, αγνοώντας συχνά την πλειοψηφία των αξιολογήσεων από ένα χρήστη. Συνεπώς αυτές οι μέθοδοι δεν είναι ικανές να αποτυπώσουν το σύνολο των αδύναμων σημείων που βρίσκονται σε όλες τις αξιολογήσεις ενός χρήστη.

Αντίθετα, τα μοντέλα λανθάνοντος παράγοντα είναι αποτελεσματικά στην εκτίμηση της συνολικής δομής που συσχετίζει ταυτόχρονα τα περισσότερα ή όλα τα προϊόντα μεταξύ τους. Όμως αυτά τα μοντέλα έχουν χαμηλή απόδοση στην ανίχνευση ισχυρών συνδέσεων των προϊόντων μέσα από ένα μικρό σύνολο στενά συσχετισμένων προϊόντων, ενώ τα μοντέλα γειτνίασης σε αυτή την περίπτωση τα καταφέρνουν καλύτερα

Ο πιο ευρέως χρησιμοποιούμενος αλγόριθμος για συνεργατικό φιλτράρισμα είναι ο αλγόριθμος συστάσεων πλησιέστερων γειτόνων - k nearest neighbors recommendation algorithm (kNN) (Bobadilla et al., 2013). Εννοιολογικά είναι πολύ κοντά με την προσέγγιση του συνεργατικού φιλτραρίσματος, η εύρεση παρόμοιων χρηστών είναι ισοδύναμη με την εύρεση γειτόνων για έναν συγκεκριμένο χρήστη (Ricci, Rokach, Shapira, & Kantor, 2011). Οι αλγόριθμοι User-Based Top-N Recommendation

προσδιορίζουν τους k παρόμοιους χρήστες (γείτονες) με τον ενεργό χρήστη u (Xiaoyuan & Taghi, 2009). Η επιλογή του N θα πρέπει να γίνεται προσεκτικά διότι εάν το N είναι πολύ μεγάλο, θα χρειαστεί υπερβολική ποσότητα μνήμης για την αποθήκευση των λιστών γειτονιάς και η πρόβλεψη θα είναι αργή (Ricci, Rokach, Shapira, & Kantor, 2011, p. 130). Ενώ από την άλλη η επιλογή μιας υπερβολικά μικρής τιμής για το N μπορεί να οδηγήσει σε μη ακριβείς συστάσεις.

1.5.3 Συστήματα με βάση την γνώση (Knowledge – based systems)

Τα συστήματα που βασίζονται στη γνώση προτείνουν αντικείμενα που βασίζονται σε συγκεκριμένο πεδίο γνώσης σχετικά με το πώς ορισμένα χαρακτηριστικά ενός στοιχείου ανταποκρίνονται στις ανάγκες και τις προτιμήσεις των χρηστών και, τελικά, πως τα αντίστοιχα χαρακτηριστικά συσχετίζονται με τον χρήστη. Τα παραδοσιακά συστήματα σύστασης (με βάση το περιεχόμενο φιλτράρισμα και συνεργατικά) είναι κατάλληλα για τη σύσταση καθημερινών προϊόντων ή υπηρεσιών όπως βιβλία, ταινίες, ή ειδήσεις. Ωστόσο, ειδικά στο πλαίσιο των προϊόντων όπως αυτοκίνητα, υπολογιστές, διαμερίσματα, ή οι χρηματοπιστωτικές υπηρεσίες αυτές οι προσεγγίσεις δεν είναι η καλύτερη επιλογή. Για παράδειγμα, τα διαμερίσματα δεν αγοράζονται πολύ συχνά, γεγονός που καθιστά μάλλον ανέφικτη τη συλλογή πολλών αξιολογήσεων για ένα συγκεκριμένο αντικείμενο (όπως είναι οι βαθμολογίες που απαιτούνται από τους αλγορίθμους συνεργατικού φιλτραρίσματος). Επιπλέον, δεν θα είναι ικανοποιημένοι οι χρήστες των εφαρμογών με συστάσεις ετών για τις προτιμήσεις αντικειμένων (όπως θα ήταν οι συστάσεις ενός αλγόριθμου με βάση το περιεχόμενο).

Οι μέθοδοι συστάσεων που βασίζονται στη γνώση βοηθούν στην αντιμετώπιση αυτών των προκλήσεων με την αξιοποίηση ρητών απαιτήσεων των χρηστών και βαθιάς γνώσης για τον τομέα ενδιαφέροντος των συγκεκριμένων προϊόντων για τον υπολογισμό των συστάσεων. Τα συστήματα αυτά σε μεγάλο βαθμό επικεντρώνονται σε πηγές γνώσης που δεν αξιοποιούνται από το συνεργατικό και με βάση το περιεχόμενο φιλτράρισμα. Σε σύγκριση με το συνεργατικό φιλτράρισμα και το φιλτράρισμα με βάση το περιεχόμενο, τα συστήματα αυτά δεν χρειάζεται να αντιμετωπίσουν το πρόβλημα της «ψυχρής εκκίνησης», αλλά η υλοποίησή τους είναι δυσκολότερη.

Αξιοσημείωτα εισηγητικά συστήματα που βασίζονται στη γνώση είναι τα βασισμένα σε περιπτώσεις (case based). Σε αυτά τα συστήματα η συνάρτηση ομοιότητας υπολογίζει πόσο οι ανάγκες των χρηστών (περιγραφή του προβλήματος) ταιριάζουν με τις συστάσεις (λύσεις του προβλήματος). Εδώ ο βαθμός ομοιότητας μπορεί να ερμηνεύεται άμεσα ως η χρησιμότητα της σύστασης για τον χρήστη. Τα συστήματα που βασίζονται σε περιορισμούς (constraint based) είναι ένα άλλο είδος των συστημάτων σύστασης που βασίζονται στη γνώση. Όσον αφορά τη γνώση που χρησιμοποιούν, τα δύο συστήματα είναι παρόμοια: συλλέγονται οι απαιτήσεις των χρηστών, σε περιπτώσεις ασυνεπών απαιτήσεων που δεν μπορούν να βρεθούν λύσεις, προτείνονται αυτομάτως επιδιορθώσεις και τα αποτελέσματα συστάσεων μπορούν να εξηγηθούν. Η σημαντική διαφορά έγκειται στον τρόπο που υπολογίζονται οι λύσεις.

Τα συστήματα με βάση τις περιπτώσεις καθορίζουν τις συστάσεις τους βάσει μετρικών ομοιότητας ενώ τα συστήματα με βάση τους περιορισμούς εκμεταλλεύονται κυρίως προκαθορισμένες βάσεις γνώσης που περιέχουν σαφείς κανόνες για το πώς να συσχετίζουν τις απαιτήσεις του πελάτη με τα χαρακτηριστικά στοιχείου. Τα συστήματα που βασίζονται στη γνώση τείνουν να λειτουργούν καλύτερα από τα άλλα κατά την έναρξη της εφαρμογής τους, αλλά αν δεν είναι εξοπλισμένα με εργαλεία μάθησης δεν υπερτερούν σε σχέση με άλλες μεθόδους που μπορούν να εκμεταλλευτούν τα αρχεία καταγραφής της αλληλεπίδρασης ανθρώπου / υπολογιστή.

Η μέθοδος σύστασης που βασίζεται σε περιπτώσεις (case based) είναι μία από τις πιο επιτυχημένες μεθόδους μηχανικής μάθησης. Η τεχνική αυτή είναι μία κυκλική και ολοκληρωμένη διαδικασία επίλυσης προβλημάτων που στηρίζεται στη μάθηση από την εμπειρία και έχει τέσσερα κύρια στάδια: την ανάκτηση, επαναχρησιμοποίηση, προσαρμογή και διατήρηση. Με την εμφάνιση ενός νέου προβλήματος αναζητά ένα παρόμοιο πρόβλημα του παρελθόντος (όπου να έχει ήδη επιλύσει μία παρόμοια υπόθεση), και στη συνέχεια επαναχρησιμοποιεί την υπόθεση εκείνη για την επίλυση του σημερινού προβλήματος. Σε αυτές τις προσεγγίσεις μια υπόθεση και ένα προϊόν ουσιαστικά θεωρούνται πανομοιότυπα αντικείμενα .

Το πρόβλημα της σύστασης συνήθως αντιπροσωπεύεται από ένα σύνολο χαρακτηριστικών του προϊόντος, εκείνες που καθορίζονται από το χρήστη, και η λύση της υπόθεσης είναι το ίδιο το προϊόν . Στο βασικό σενάριο χρήσης, ο πελάτης ψάχνει να αγοράσει κάποιο προϊόν και καθιστά σαφείς ορισμένες απαιτήσεις σχετικά με το προϊόν

αυτό. Το σύστημα αναζητά την βασική υπόθεση για τα προϊόντα που ταιριάζουν με τις απαιτήσεις του χρήστη. Η διαδικασία ανάκτησης οδηγείται από ένα μετρητή ομοιότητας που υπολογίζει την ομοιότητα της περιγραφής του προβλήματος, δηλαδή, τις σημερινές απαιτήσεις των χρήστη με τα προϊόντα στη βάση δεδομένων της υπόθεσης. Μια σειρά από προϊόντα στη συνέχεια ανακτώνται από τη βασική υπόθεση και τα προϊόντα αυτά συνιστώνται στο χρήστη. Αν ο χρήστης δεν είναι ικανοποιημένος με τις υποδείξεις μπορεί να τροποποιήσει τις απαιτήσεις του και ένας νέος κύκλος σύστασης ξεκινάει.

1.5.4 Υβριδικά συστήματα συστάσεων (Hybrid Recommendation Systems)

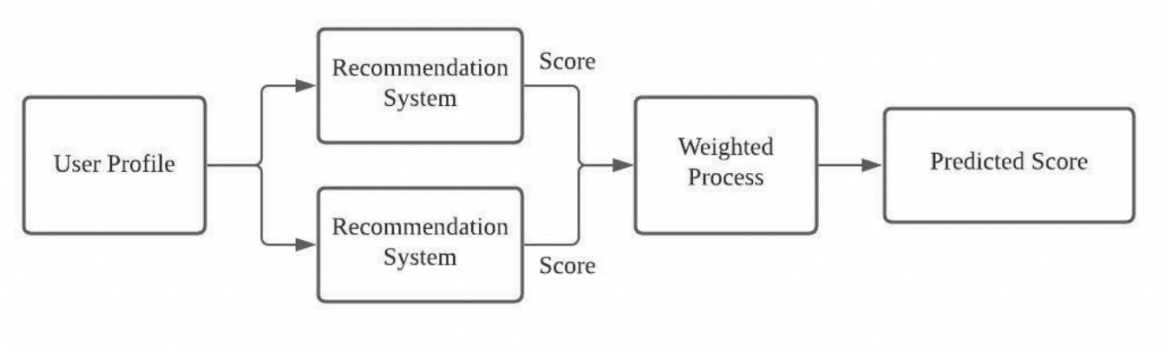
Ένα υβριδικό σύστημα συστάσεων είναι ένας ειδικός τύπος συστάσεων που μπορεί να θεωρηθεί ως ο συνδυασμός της μεθόδου περιεχομένου και του συνεργατικού φιλτραρίσματος. Ο συνδυασμός των δύο μεθόδων μπορεί να βοηθήσει στην αντιμετώπιση των αδυναμιών που μπορεί να προκύψουν κατά τη χρήση των μεθόδων ξεχωριστά καθώς επίσης μπορεί να είναι πιο αποτελεσματική σε ορισμένες περιπτώσεις. Υπάρχουν διάφοροι τρόποι υλοποίησης υβριδικών συστημάτων, όπως κάνοντας αρχικά χρήση των συμβατικών μεθόδων ξεχωριστά και στη συνέχεια γίνεται συνδυασμός των προβλέψεων ή γίνεται χρήση μίας από τις συμβατικές μεθόδους στην οποία προσθέτουμε δυνατότητες τις άλλης και το αντίστροφο. Σύμφωνα με μελέτες που συγκρίνουν την απόδοση των συμβατικών προσεγγίσεων με των υβριδικών μεθόδων η χρήση υβριδικών συστημάτων συμβάλει στην δημιουργία πιο αξιόπιστων συστάσεων (Çano and Morisio, 2017).

Όπως αναφέρθηκε και πιο πάνω υπάρχουν διάφοροι τρόποι υλοποίησης ενός υβριδικού συστήματος οι οποίοι είναι (Seth and Sharaff, 2022):

- **Σταθμισμένο σύστημα (Weighted System)**

Σε αυτού του τύπου συστήματα ορίζονται αρχικά τα μοντέλα τα οποία θα συμβάλουν στην ερμηνεία του συνόλου δεδομένων. Το σύστημα συστάσεων θα λάβει τις εξόδους από κάθε ένα από τα μοντέλα και θα συνδυάσει το αποτέλεσμα σε στατικές σταθμίσεις οι οποίες παραμένουν σταθερές στο σύνολο εκπαίδευσης και δοκιμής. Για παράδειγμα γίνεται συνδυασμός ενός μοντέλου βασισμένο στο περιεχόμενο και ένα μοντέλο συνεργατικού φιλτραρίσματος και το καθένα από τα μοντέλα λαμβάνει βάρος 50% προς

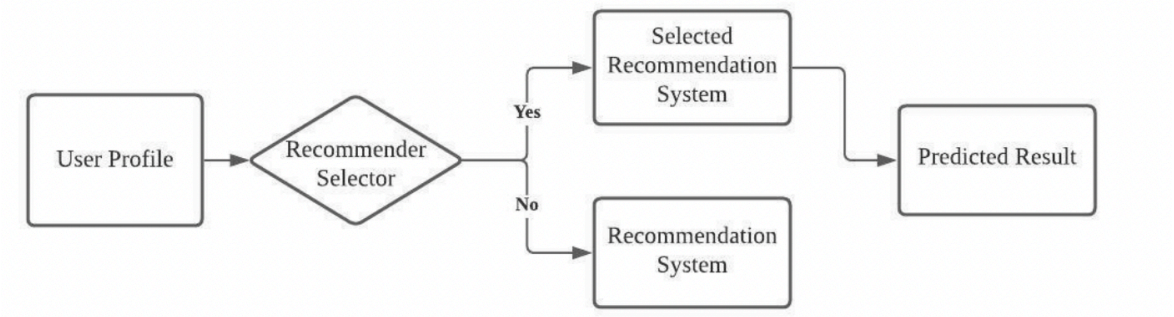
την τελική πρόβλεψη. Το πλεονέκτημα της χρήσης του σταθμισμένου υβριδικού είναι ότι ενσωματώνουμε πολλαπλά μοντέλα για την υποστήριξη του συνόλου δεδομένων στη διαδικασία σύστασης με γραμμικό τρόπο.



Εικόνα 1.2 Παράδειγμα σταθμισμένου υβριδικού συστήματος

- **Εναλλασσόμενο σύστημα (Switching system)**

Το υβριδικό σύστημα εναλλαγής επιλέγει το κατάλληλο μοντέλο κάθε φορά με βάση την κατάσταση. Αρχικά θα πρέπει να γίνει ο ορισμός των κριτηρίων επιλογής με βάση το προφίλ του χρήστη ή άλλων χαρακτηριστικών. Το σύστημα αυτό εισάγει άλλο ένα πρόσθετο επίπεδο πάνω στο μοντέλο σύστασης το οποίο επιλέγει το κατάλληλο μοντέλο για χρήση.

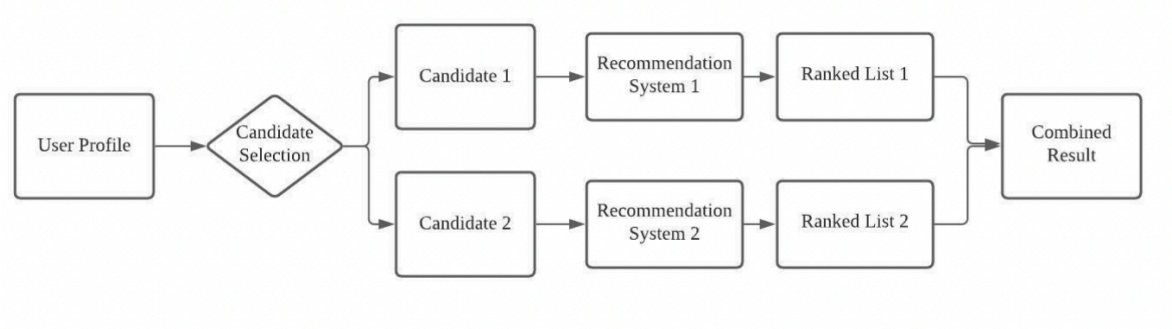


Εικόνα 1.3 Παράδειγμα εναλλασσόμενου υβριδικού συστήματος

- **Μικτό σύστημα (Mixed System)**

Το σύστημα αυτό λαμβάνει πρώτα το προφίλ του χρήστη και τα χαρακτηριστικά του για να δημιουργήσει διαφορετικά σύνολα υποψήφιων συνόλων δεδομένων. Τα σύνολα αυτά εισάγονται στο μοντέλο συστάσεων το οποίο συνδυάζει την πρόβλεψη για να παράγει το αποτέλεσμα της σύστασης. Το μικτό υβριδικό σύστημα συστάσεων είναι σε θέση να

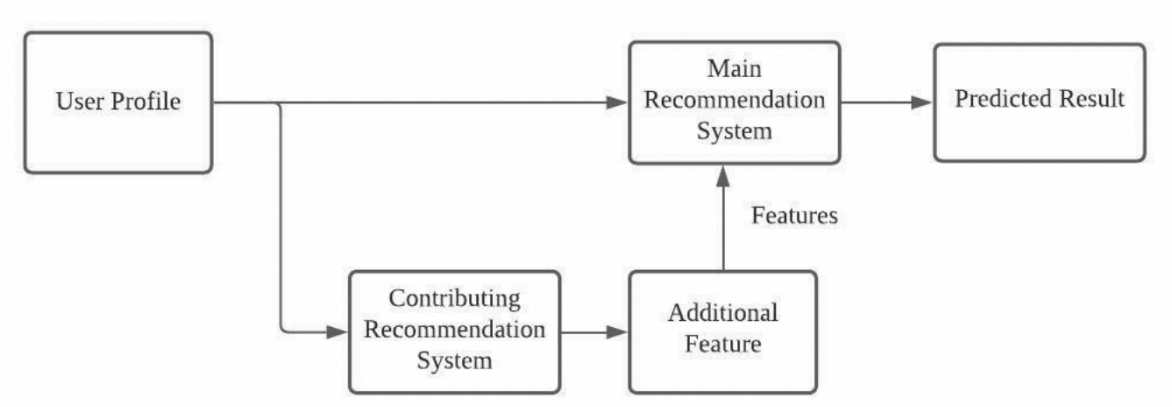
κάνει ταυτόχρονα μεγάλο αριθμό συστάσεων και να προσαρμόζει το μερικό σύνολο δεδομένων στο κατάλληλο μοντέλο, με σκοπό την επίτευξη καλύτερης απόδοσης.



Εικόνα 1.4 Παράδειγμα μεικτού υβριδικού συστήματος

- **Συστήματα βασιζόμενα στο συνδυασμό χαρακτηριστικών (Feature combination Systems)**

Στο σύστημα αυτό γίνεται προσθήκη ενός εικονικού μοντέλου συστάσεων το οποίο μηχανή δημιουργίας χαρακτηριστικών με βάση το αρχικό σύνολο προφίλ χρήστη. Για παράδειγμα, μπορούμε να εισάγουμε χαρακτηριστικά ενός συνεργατικού μοντέλου συστάσεων σε ένα μοντέλο συστάσεων που βασίζεται σε περιεχόμενο. Το υβριδικό μοντέλο είναι σε θέση να λάβει υπόψη του τα συνεργατικά δεδομένα από το υποσύστημα, βασιζόμενο αποκλειστικά σε ένα μοντέλο.

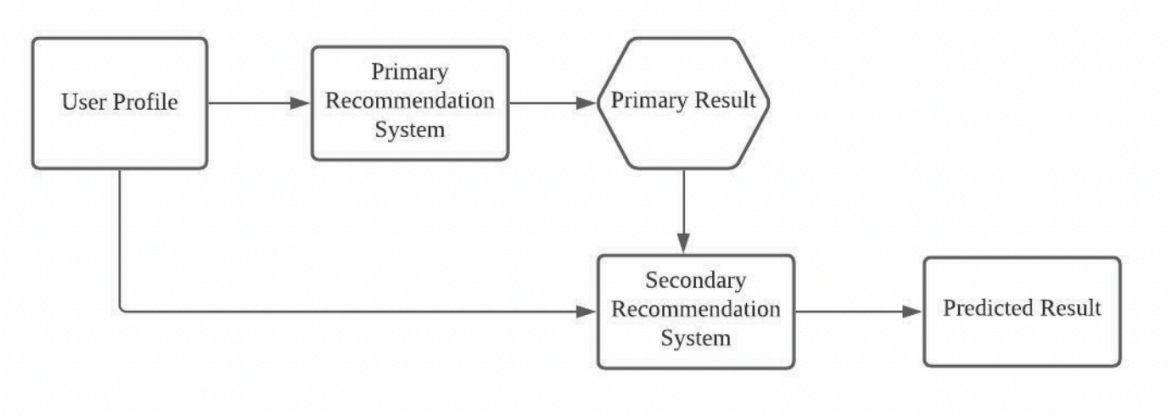


Εικόνα 1.5 Παράδειγμα υβριδικού συστήματος βασιζόμενο στα χαρακτηριστικά

- **Cascade System**

Το σύστημα Cascade ορίζει ένα σύστημα συστάσεων αυστηρά ιεραρχικής δομής, έτσι ώστε το κύριο σύστημα συστάσεων να παράγει το πρωταρχικό αποτέλεσμα και στη

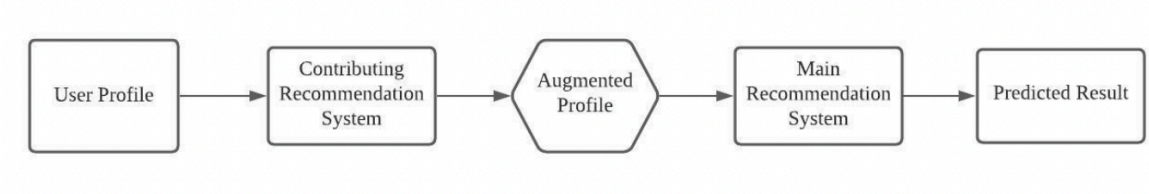
συνέχεια μέσω της χρήσης του δευτερεύοντος μοντέλου να επιλύονται ζητήματα του πρωταρχικού αποτελέσματος. Συνήθως στην πράξη τα περισσότερα σύνολα δεδομένων είναι αραιά, το δευτερεύον μοντέλο σύστασης μπορεί να είναι αποτελεσματικό κατά του προβλήματος της ίσης βαθμολόγησης ή του προβλήματος ελλιπών δεδομένων.



Εικόνα 1.6 Παράδειγμα cascade υβριδικού συστήματος

- **Σύστημα επαύξησης χαρακτηριστικών (Feature Augmentation)**

  Το σύστημα αυτό αρχικά βασίζει την λειτουργία του ένα μοντέλο συστάσεων το οποίο χρησιμοποιείται για τη δημιουργία μιας αξιολόγησης ή ταξινόμησης του προφίλ χρήστη η οποία χρησιμοποιείται περαιτέρω για την παραγωγή του τελικού προβλεπόμενου αποτελέσματος. Το υβριδικό σύστημα επαύξησης χαρακτηριστικών είναι σε θέση να βελτιώσει την απόδοση του βασικού συστήματος χωρίς να αλλάξει το κύριο μοντέλο σύστασης. Για παράδειγμα, χρησιμοποιώντας τον κανόνα συσχέτισης, είναι σε θέση να ενισχύσει το σύνολο δεδομένων του προφίλ χρήστη. Με το επαυξημένο σύνολο δεδομένων, θα βελτιωθεί η απόδοση του μοντέλου συστάσεων βάσει περιεχομένου.



Εικόνα 1.7 Παράδειγμα υβριδικού συστήματος επαύξησης χαρακτηριστικών

- **Σύστημα μετα – επιπέδου (Meta – level system)**

Το σύστημα αυτό είναι παρόμοιο με το σύστημα επαύξησης χαρακτηριστικών κατά το οποίο το επαυξημένο – ενημερωμένο σύνολο δεδομένων παρέχεται στο κύριο μοντέλο συστάσεων. Σε σχέση με το μοντέλο επαύξησης χαρακτηριστικών το συγκεκριμένο μοντέλο αντικαθιστά το αρχικό σύνολο δεδομένων με ένα μοντέλο το οποίο λειτουργεί ως είσοδο στο κύριο μοντέλο συστάσεων.

1.6 Προβλήματα – περιορισμοί στα συστήματα συστάσεων

Σε αυτή την ενότητα παρουσιάζονται διάφορα προβλήματα και προκλήσεις που συναντώνται στα συστήματα συστάσεων. Όπως και ο αλγόριθμος PVR, ο αλγόριθμος Top N video-ranker συνδυάζει μετρικές εξατομίκευσης και δημοτικότητας κατά τη δημιουργία των συστάσεών του.

1.6.1 Πρόβλημα ψυχρής εκκίνησης (Cold – start problem)

Το πρόβλημα της ψυχρής εκκίνησης αφορά την εισαγωγή νέων προϊόντων τα οποία δεν μπορούν να βαθμολογηθούν ή να αγοραστούν από τους χρήστες, με συνέπεια να οδηγούν σε λιγότερο ακριβείς συστάσεις.

Το πρόβλημα της ψυχρής εκκίνησης μπορεί να λυθεί με πολλούς τρόπους, όπως:

- Ζητώντας από τον χρήστη να βαθμολογήσει τα νέα στοιχεία που εισάγονται.
- Ζητώντας από τους χρήστες να δηλώσουν ρητά τις προτιμήσεις τους συνολικά.
- Να συνάγει προτάσεις στοιχείων στον χρήστη με βάση τα δημογραφικά στοιχεία που συλλέχθηκαν.

Οι δημογραφικές πληροφορίες χρήστη μπορούν να χρησιμοποιηθούν για να γίνουν γνωστά στοιχεία σχετικά με την τοποθεσία, τον ταχυδρομικό κώδικα καθώς και τις αλληλεπιδράσεις του νέου χρήστη με το σύστημα, προκειμένου να προτείνονται στοιχεία με βάση τις αξιολογήσεις που παρέχονται από άλλους παρόμοιους χρήστες με παρόμοια δημογραφικά στοιχεία. Ένα ζήτημα που μπορεί να εμφανιστεί αφορά συγκεκριμένους τομείς, όπου ενδέχεται να υπάρχουν "sleepers" (στοιχεία εν υπνώσει) που είναι στοιχεία τα οποία είναι σημαντικά αλλά δεν έχουν αξιολογηθεί. Ο χειρισμός των sleepers μπορεί να γίνει χρησιμοποιώντας μεταδεδομένα ή μεθόδους που βασίζονται σε περιεχόμενο, χρησιμοποιώντας τη δημοτικότητα στοιχείων, την εντροπία των στοιχείων και την εξατομίκευση του χρήστη, καθώς και χρήση Συνδεδεμένων Ανοικτών Δεδομένων (Linked Open Data-LOD) εξάγοντας δεδομένα σχετικά με στοιχεία που χρησιμοποιούν

πηγές δεδομένων LOD, χωρίς να περιμένουν από τους χρήστες να παρέχουν ρητά αξιολογήσεις. (Khusro, Ali, Ullah, 2016)

1.6.2 Πρόβλημα της συνωνυμίας (Synonymy problem)

Η συνωνυμία προκύπτει όταν ένα αντικείμενο αναπαρίσταται με δύο ή περισσότερα διαφορετικά ονόματα ή καταχωρήσεις που έχουν παρόμοια σημασία. Σε τέτοιες περιπτώσεις, ο χρήστης δεν μπορεί να προσδιορίσει εάν οι όροι αντιπροσωπεύουν διαφορετικά στοιχεία ή το ίδιο αντικείμενο. Η διακύμανση στη χρήση περιγραφικών όρων είναι μεγαλύτερη από ό,τι συνήθως και η υπερβολική χρήση συνωνύμων λέξεων μειώνει την απόδοση των συστάσεων. Δεδομένου ότι τα περιεχόμενα των αντικειμένων αγνοούνται επιμελώς, το σύστημα συστάσεων δεν λαμβάνει υπόψιν την λανθάνουσα συσχέτιση μεταξύ των στοιχείων. Αυτός είναι ο λόγος για τον οποίο δεν συνίστανται νέα στοιχεία, εφόσον αυτά αξιολογούνται από τους χρήστες. Για την άμβλυνση των προβλημάτων της συνωνυμίας, θα μπορούσαν να χρησιμοποιηθούν διαφορετικές τεχνικές, συμπεριλαμβανομένων των οντολογιών, των τεχνικών SVD και Latent Semantic Indexing (LSI). (Khusro, Ali, Ullah, 2016)

1.6.3 Πρόβλημα της ιδιωτικότητας (Privacy problem)

Η παροχή προσωπικών πληροφοριών στα συστήματα συστάσεων έχει ως αποτέλεσμα καλύτερες υπηρεσίες συστάσεων, αλλά απαιτείται να λαμβάνονται υπόψιν τα ζητήματα απορρήτου και ασφάλειας δεδομένων. Οι χρήστες διστάζουν να τροφοδοτήσουν δεδομένα σε συστήματα συστάσεων που χαρακτηρίζονται από ζητήματα απορρήτου δεδομένων. Ένα σύστημα συστάσεων θα πρέπει να εμπνέει εμπιστοσύνη στους χρήστες του. Στην τεχνική συνεργατικού φιλτραρίσματος για παράδειγμα, τα δεδομένα χρήστη, συμπεριλαμβανομένων των αξιολογήσεων, αποθηκεύονται σε ένα κεντρικό αποθετήριο το οποίο μπορεί να παραβιαστεί με αποτέλεσμα την κακή χρήση δεδομένων. Για το σκοπό αυτό, μπορούν να χρησιμοποιηθούν κρυπτογραφικοί μηχανισμοί παρέχοντας εξατομικευμένες συστάσεις χωρίς τη συμμετοχή τρίτων και ομότιμων χρηστών. Άλλες τεχνικές περιλαμβάνουν τη χρήση τεχνικών τυχαιοποιημένης διαταραχής, που επιτρέπει στους χρήστες να δημοσιεύουν τα προσωπικά τους δεδομένα χωρίς να εκθέτουν την ταυτότητά τους, και τη χρήση τεχνολογιών Σημασιολογικού Ιστού, ειδικά οντολογιών σε συνδυασμό με τεχνικές NLP για τον μετριασμό της ανεπιθύμητης έκθεσης πληροφοριών. (Khusro, Ali, Ullah, 2016)

1.6.4 Πρόβλημα γκρίζου προβάτου (Grey sheep problem)

Το γκρίζο πρόβατο εμφανίζεται σε αμιγώς συστήματα συνεργατικού φιλτραρίσματος, όπου οι απόψεις ενός χρήστη δεν ταιριάζουν με καμία ομάδα και επομένως δεν μπορεί να επωφεληθεί από τις συστάσεις. Το καθαρό φιλτράρισμα που βασίζεται στο περιεχόμενο μπορεί να επιλύσει αυτό το ζήτημα, όπου προτείνονται αντικείμενα με εκμετάλλευση του προσωπικού προφίλ χρήστη και των περιεχομένων των αντικειμένων που προτείνονται. Ομοίως, η αραιή βαθμολογία και ο πρώτος βαθμολογητής στο συνεργατικό φιλτράρισμα μπορούν επίσης να επιλυθούν με την τεχνική φιλτραρίσματος περιεχομένου. Οι χρήστες γκρίζου προβάτου μπορούν να εντοπιστούν και να διαχωριστούν από άλλους χρήστες εφαρμόζοντας τεχνικές ομαδοποίησης εκτός σύνδεσης, συμπεριλαμβανομένης της ομαδοποίησης k-mean. Με αυτόν τον τρόπο η απόδοση βελτιώνεται και το σφάλμα σύστασης είναι ελάχιστο. (Khusro, Ali, Ullah, 2016)

1.6.5 Πρόβλημα αραιών και διάσπαρτων δεδομένων (Data Sparity)

Η διαθεσιμότητα τεράστιου μεγέθους δεδομένων και η απροθυμία των χρηστών να βαθμολογήσουν αντικείμενα δημιουργούν ένα σύνολο διαθέσιμων δεδομένων με αρκετές ελλείψεις, δηλαδή τα δεδομένα να είναι αραιά και διάσπαρτα με αποτέλεσμα να επηρεάζεται η ευστοχία και η αποδοχή των προτάσεων και, συνεπώς, η αποτελεσματικότητα του συστήματος συστάσεων. Η αραιή βαθμολογία στα συστήματα συνεργατικού φιλτραρίσματος (Collaborative Filtering - CF) καθιστά δύσκολη την πραγματοποίηση ακριβών προβλέψεων σχετικά με τα στοιχεία. Το CF χρησιμοποιεί τους πλησιέστερους γείτονες για να προτείνει στοιχεία και οι λιγότερες βαθμολογίες καθιστούν υπολογιστικά δύσκολο τον υπολογισμό γειτόνων. Το πρόβλημα μπορεί να είναι σοβαρό σε διάφορα συστήματα συστάσεων, καθώς χρησιμοποιούν πολυδιάστατα διανύσματα, όπου καθίσταται πολύ δύσκολο να παρέχονται ακριβείς συστάσεις για πολύ λίγα αξιολογημένα στοιχεία. Για την αντιμετώπιση αυτής της κατάστασης, μπορούν να χρησιμοποιηθούν διάφορες προσεγγίσεις, όπως το πολυδιάστατο μοντέλο συστάσεων, οι τεχνικές Single Value Decomposition (SVD), το δημογραφικό φιλτράρισμα και η χρήση αλγόριθμου CF ενισχυμένου περιεχομένου. (Khusro, Ali, Ullah, 2016)

1.6.6 Πρόβλημα επιθέσεων Shilling (Shilling attacks problem)

Υπάρχει η πιθανότητα ένας κακόβουλος χρήστης ή ανταγωνιστής να εισέλθει σε ένα σύστημα και να αρχίσει να δίνει ψευδείς αξιολογήσεις για ορισμένα στοιχεία, προκειμένου, είτε να αυξήσει τη δημοτικότητα του προϊόντος, είτε να τη μειώσει. Τέτοιου είδους επιθέσεις πραγματοποιούνται με τη δημιουργία ψευδών προφίλ και την ανάπτυξη πολλών λογαριασμών χρηστών, αλλά προϋποθέτουν, επίσης, τη γνώση σχετικά με κάποια στοιχεία, όπως για παράδειγμα τη μέση τιμή αξιολόγησης (rating) ενός χαρακτηριστικού. Τέτοιες επιθέσεις μπορούν να κλονίσουν την εμπιστοσύνη στο σύστημα συστάσεων, καθώς και να μειώσουν την απόδοση και την ποιότητα των συστάσεων. Αυτή η απειλή προκαλεί μεγαλύτερη ανησυχία στις τεχνικές CF αλλά μικρότερη απειλή για την τεχνική CF που βασίζεται σε είδη. Υπάρχουν διαφορετικά μοντέλα επίθεσης όπως bandwagon, random, average και reversed bandwagon. Οι επιθέσεις μπορούν να ανιχνευθούν μέσω διαφορετικών προσεγγίσεων όπως γενικά χαρακτηριστικά και συγκεκριμένα χαρακτηριστικά μοντέλων, μετατόπιση πρόβλεψης και αναλογία επιτυχιών. Αυτοί οι τύποι επιθέσεων μπορούν να κατηγοριοποιηθούν με διαστάσεις όπως η πρόθεση επίθεσης, το μέγεθος της επίθεσης και η απαιτούμενη γνώση για την έναρξη της επίθεσης. (Khusro, Ali, Ullah, 2016), (Stalidis, Kardaras,2015)

1.6.7 Πρόβλημα περιορισμένης ανάλυσης περιεχομένου και υπερεξειδίκευσης (Limited Content Analysis and Overspecialization problem)

Τα συστήματα συστάσεων που βασίζονται σε περιεχόμενο (Content-Based-CB), εξαρτώνται από στοιχεία και χρήστες που πρόκειται να υποβληθούν σε επεξεργασία με τεχνικές ανάκτησης πληροφοριών. Η περιορισμένη διαθεσιμότητα περιεχομένου οδηγεί σε προβλήματα, συμπεριλαμβανομένης της υπερεξειδίκευσης. Εδώ, τα στοιχεία αντιπροσωπεύονται από τα υποκειμενικά τους χαρακτηριστικά. Οι δυνατότητες που αντιπροσωπεύουν τις προτιμήσεις των χρηστών με καλύτερο τρόπο, δεν λαμβάνονται υπόψη. Για πολλούς τομείς, το περιεχόμενο είτε είναι σπάνιο, όπως βιβλία, είτε είναι δύσκολο να αποκτηθεί και να αναπαρασταθεί, όπως ταινίες. Σε τέτοιες περιπτώσεις, δεν μπορούν να προταθούν σχετικά στοιχεία εκτός εάν το αναλυόμενο περιεχόμενο περιέχει αρκετές πληροφορίες για να χρησιμοποιηθούν για τη διάκριση στοιχείων που αρέσουν/δεν αρέσουν στον χρήστη. Αυτό οδηγεί επίσης στην αναπαράσταση δύο διαφορετικών στοιχείων με το ίδιο σύνολο χαρακτηριστικών, όπου, π.χ., τα καλογραμμένα ερευνητικά άρθρα μπορεί να είναι δύσκολο να διακριθούν από τα κακά

εάν και τα δύο αντιπροσωπεύονται με το ίδιο σύνολο λέξεων-κλειδιών. Η περιορισμένη ανάλυση περιεχομένου οδηγεί σε υπερβολική εξειδίκευση στην οποία τα συστήματα προτείνουν στοιχεία που σχετίζονται στενά με το προφίλ χρήστη και δεν προτείνουν νέα στοιχεία. Προκειμένου να προτείνονται νέα και ασυνήθιστα αντικείμενα μαζί με γνωστά αντικείμενα, πρέπει να εισαχθούν πρόσθετα hacks και σημειώσεις τυχαία, που μπορούν να επιτευχθούν με τη χρήση γενετικών αλγορίθμων που φέρνουν ποικιλομορφία στις προτάσεις που γίνονται. (Khusro, Ali, Ullah, 2016)

1.6.8 Πρόβλημα καθυστέρησης (Latency problem)

Τα συστήματα που βασίζονται στο συνεργατικό φιλτράρισμα αντιμετωπίζουν το πρόβλημα του λανθάνοντος χρόνου όταν νέα στοιχεία προστίθενται πιο συχνά στη βάση δεδομένων, όπου το σύστημα προτείνει μόνο τα ήδη βαθμολογημένα στοιχεία καθώς τα νέα στοιχεία που προστέθηκαν δεν έχουν ακόμη βαθμολογηθεί. Η χρήση του φιλτραρίσματος βάσει περιεχομένου μπορεί να μειώσει τους χρόνους αναμονής, αλλά μπορεί να εισάγει υπερεξειδίκευση. Για να αντιμετωπιστεί αυτή η κατάσταση, μπορεί να χρησιμοποιηθεί η προσέγγιση βάσει κατηγορίας σε συνδυασμό με το στερεότυπο του χρήστη. Για περαιτέρω αύξηση της απόδοσης, μπορούν να εφαρμοστούν πολλές τεχνικές ομαδοποίησης και οι υπολογισμοί μπορούν να γίνουν εκτός σύνδεσης. Οι προσεγγίσεις CF που βασίζονται σε μοντέλα μπορούν επίσης να βελτιώσουν την επεκτασιμότητα και την απόδοση. (Khusro, Ali, Ullah, 2016)

1.7 Αξιολόγηση συστημάτων συστάσεων

Ο σωστός σχεδιασμός του συστήματος αξιολόγησης είναι ζωτικής σημασίας για την κατανόηση της αποτελεσματικότητας των διαφόρων αλγορίθμων συστάσεων. Η αξιολόγηση των συστημάτων συστάσεων είναι συχνά πολύπλευρη και ένα μόνο κριτήριο δεν μπορεί να είναι αντικειμενική. Ένας λανθασμένος σχεδιασμός της πειραματικής αξιολόγησης μπορεί να οδηγήσει είτε σε κατάφωρη υποτίμηση είτε σε υπερεκτίμηση της πραγματικής ακρίβειας ενός συγκεκριμένου αλγορίθμου ή μοντέλου.

Τα συστήματα συστάσεων μπορούν να αξιολογηθούν μέσω:

- Offline μεθόδων
- Online μεθόδων

Τα ακόλουθα ζητήματα είναι σημαντικά από την άποψη του σχεδιασμού μεθόδων αξιολόγησης για συστήματα συστάσεων:

- **Στόχοι αξιολόγησης:** Αν και είναι δελεαστικό να χρησιμοποιηθούν μετρήσεις ακρίβειας για την αξιολόγηση συστημάτων συστάσεων, μια τέτοια προσέγγιση μπορεί συχνά να παρέχει μια ελλιπή εικόνα της εμπειρίας του χρήστη. Αν και οι μετρήσεις ακρίβειας είναι αναμφισβήτητα τα πιο σημαντικά στοιχεία του, πολλοί δευτερεύοντες στόχοι, όπως η καινοτομία, η εμπιστοσύνη και η κάλυψη είναι σημαντικοί για την εμπειρία του χρήστη. Αυτό συμβαίνει επειδή αυτές οι μετρήσεις έχουν σημαντικές βραχυπρόθεσμες και μακροπρόθεσμες επιπτώσεις στα ποσοστά μετατροπών. Ωστόσο, ο πραγματικός ποσοτικός προσδιορισμός ορισμένων από αυτούς τους παράγοντες είναι συχνά αρκετά υποκειμενικός και συχνά δεν υπάρχουν σκληρά μέτρα για την παροχή μιας αριθμητικής μέτρησης.
- **Ζητήματα πειραματικού σχεδιασμού:** Ακόμη και όταν χρησιμοποιείται η ακρίβεια ως μέτρηση, είναι σημαντικό να σχεδιάζονται τα πειράματα έτσι ώστε η ακρίβεια να μην υπερεκτιμάται ή υποτιμάται. Για παράδειγμα, εάν το ίδιο σύνολο καθορισμένων χαρακτηρισμών χρησιμοποιείται τόσο για την κατασκευή του μοντέλου όσο και για την αξιολόγηση της ακρίβειας, τότε η ακρίβεια θα υπερεκτιμηθεί κατάφωρα. Σε αυτό το πλαίσιο, ο προσεκτικός πειραματικός σχεδιασμός είναι σημαντικός.
- **Μετρήσεις ακρίβειας:** Παρά τη σημασία άλλων δευτερευόντων μέτρων, οι μετρήσεις ακρίβειας εξακολουθούν να είναι το μοναδικό πιο σημαντικό στοιχείο στην αξιολόγηση. Τα συστήματα συστάσεων μπορούν να αξιολογηθούν είτε ως προς την ακρίβεια πρόβλεψης μιας βαθμολογίας είτε ως προς την ακρίβεια κατάταξης των στοιχείων. Επομένως, ένας αριθμός κοινών μετρήσεων όπως το μέσο απόλυτο σφάλμα και το μέσο τετραγωνικό σφάλμα χρησιμοποιούνται συχνά. Η αξιολόγηση των ταξινομήσεων μπορεί να πραγματοποιηθεί με τη χρήση διαφόρων μεθόδων, όπως υπολογισμούς με βάση τη χρησιμότητα, με συντελεστές συσχέτισης κατάταξης και με τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη. (Aggarwal, 2016)

	Reproducibility	Reliability of results	Cost of the preparation	Evaluation	Stability
Offline Evaluation	In offline evaluation there is a fixed dataset and possible fixed user interactions with it, so comparing to online evaluations, the results of this type of evaluation is also reproducible in an easier way.	We use a fixed dataset, so it's difficult to have a good picture and reliable results.	In offline evaluation, no preparation cost is needed when they use the open dataset.	Evaluation is difficult, because in offline evaluation we have just information about the evaluation of user to the items.	The evaluation of users in offline evaluation is stable.
Online Evaluation	It's difficult to reproduce the same results, because in online evaluation, we measure the actions of users in the real time.	In opposite to offline evaluations, the online experiment has the possibility to collect real time user interaction that performs real tasks with the RS, which can help us to have more reliable results.	In online evaluation, the preparation cost is very high because of the settings of users (metrics, tasks, etc.)	In online evaluation, we can ask users about their evaluations not like offline evaluation, so the evaluation is very easy.	The evaluation of users in online evaluation is not stable because users give their evaluations after using the system.
	Possibility of extensibility	**Scalability**	**Time**	**Deep analysis**	**Sparsity**
Offline Evaluation	In offline evaluation, we use a statistic dataset, so we cannot add new metrics.	It's not difficult to have many users and items in offline evaluation.	Since we cannot use the system in real time, it's difficult to analyze with the passed time in offline evaluation.	We have a static data so we cannot make further analysis.	The sparsity of ratings datasets can limit the items that can be evaluated, so we cannot evaluate the relevance of a recommended item to a user if we don't have the evaluation of this user to this item.
Online Evaluation	In online evaluation, we can always interact with users in contrast to offline evaluation, so it's easy to add new metrics.	It's difficult to have many users and items in the online evaluation because users must use the systems in several contexts.	In online evaluation, it's not difficult to analyze with time passed because they can use the system in real time.	Since we have a real time data instead of statistic data, we can give a deep analysis in several contexts.	We don't have this problem in online evaluation, because we have real time user, we can ask users about the evaluation.

Πίνακας 1.1 Συγκριτικός πίνακας των μεθόδων αξιολόγησης (Πηγή: Najmani et al., 2022)

1.7.1 Offline μέθοδος

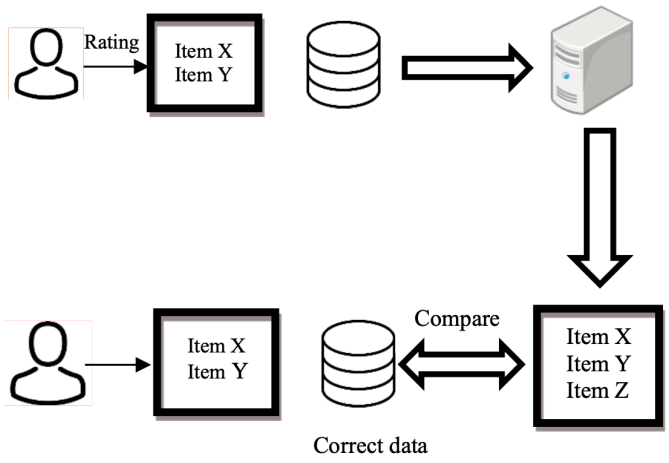
Οι μέθοδοι αξιολόγησης εκτός σύνδεσης διεξάγονται με την χρήση ενός προ – συλλεγμένου συνόλου δεδομένων χρηστών που επιλέγουν ή βαθμολογούν αντικείμενα. Μέσω της χρήσης αυτού του συνόλου δεδομένων γίνεται προσπάθεια προσομοίωσης της συμπεριφοράς των χρηστών που αλληλοεπιδρούν με το σύστημα συστάσεων. Με αυτόν τον τρόπο γίνεται μια υπόθεση ότι η συμπεριφορά των χρηστών κατά την διάρκεια

ανάπτυξης του συστήματος θα είναι σε μεγάλο βαθμό παρόμοια με τη συμπεριφορά των χρηστών όταν αναπτυχθεί το σύστημα.

Τα πειράματα εκτός σύνδεσης χρησιμοποιούνται σε μεγάλο βαθμό καθώς δεν απαιτούν αλληλεπίδραση με πραγματικούς χρήστες επιτρέποντας την σύγκριση ενός μεγάλου φάσματος υποψήφιων αλγορίθμων με χαμηλό κόστος. Το μειονέκτημα αυτής της μεθόδου αξιολόγησης είναι ότι μπορούν να απαντήσουν μόνο σε περιορισμένες ερωτήσεις σχετικά με την ικανότητα πρόβλεψης ενός αλγορίθμου. Ο στόχος των offline μεθόδων αξιολόγησης είναι να φιλτράρουν τις ακατάλληλες προσεγγίσεις, περιορίζοντας έτσι τον αριθμό των υποψήφιων αλγορίθμων που θα ελεγχθούν.

Προκειμένου να γινεί αξιολόγηση αλγορίθμων με την offline μέθοδο, είναι απαραίτητο να γίνει προσομοίωση της διαδικτυακής διαδικασίας στην οποία η τελική μορφή του συστήματος θα χρησιμοποιηθεί από τους χρήστες και θα προσφέρει προβλέψεις ή συστάσεις και τις οποίες θα διορθώνουν οι χρήστες. Η διαδικασία αυτή γίνεται με την χρήση ιστορικών δεδομένων χρηστών στα οποία γίνεται απόκρυψη ορισμένων εγγραφών – αλληλεπιδράσεων προκειμένου να προσομοιωθεί η γνώση του τρόπου με την οποία ένας χρήστης θα βαθμολογήσει ένα στοιχείο ή σε ποιες συστάσεις θα βασιστεί ο χρήστης. Υπάρχουν διάφοροι τρόποι με του οποίος γίνεται η επιλογή των στοιχείων προς απόκρυψη αλλά η επιλογή αυτή είναι προτιμότερο να γίνεται με τέτοιο τρόπο ο οποίος προσομοιώνει όσο το δυνατόν καλύτερα την εφαρμογή. Στην περίπτωση αξιολόγησης πειραμάτων σε μεγάλα σύνολα δεδομένων υπάρχει περιορισμός λόγο του υπολογιστικού κόστους.

Τα δεδομένα που χρησιμοποιούνται σε αυτή την μεθοδολογία θα πρέπει να ταιριάζουν όσο το δυνατόν περισσότερο με τα δεδομένα τα οποία αναμένεται να αντιμετωπίσει το σύστημα όταν αναπτυχθεί πλήρως. Είναι σημαντικό ότι στα δεδομένα αυτά θα πρέπει να διασφαλιστεί ότι δεν υπάρχει μεροληψία στις κατανομές των χρηστών, των στοιχείων και των αξιολογήσεων που επιλέγονται. Πριν την χρήση των δεδομένων θα πρέπει να γίνει ένα φιλτράρισμα εκ των προτέρων προκειμένου να αποκλειστούν στοιχεία ή χρήστες με χαμηλές μετρήσεις με σκοπό την μείωση του κόστους του πειράματος. Η ενέργεια αυτή συνεπάγεται την εισαγωγή μιας συστηματικής μεροληψίας στα δεδομένα.



Εικόνα 1.8 Offline μέθοδος αξιολόγησης (Πηγή: Najmani et al., 2022)

Η μέθοδος αυτή ακολουθεί μια διαδικασία αξιολόγησης εκπαίδευσης – δοκιμής κατά την οποία:

- Αρχικά τα δεδομένα χρηστών χωρίζονται σε σύνολο εκπαίδευσης (train set) και στο σύνολο δοκιμής - ελέγχου (test set).
- Εκπαίδευση των αλγορίθμων συστάσεων στο σύνολο εκπαίδευσης.
- Δημιουργία ενός καταλόγου συστάσεων από σύνολο υποψήφιων χαρακτηριστικών (συνήθως αντικείμενα τα οποία δεν έχει αξιολογήσει ο χρήστης στο σύνολο εκπαίδευσης) για κάθε χρήστη.
- Δοκιμή της ακρίβειας των προβλέψεων ή της κατάταξης των συστάσεων κάνοντας χρήση του συνόλου δοκιμής – ελέγχου (test set).
- Υπολογισμός του μέσου όρου των βαθμολογιών των μετρικών αποτελεσμάτων για όλους τους χρήστες από το σύνολο δοκιμών ελέγχου (test set).

Στις offline μεθόδους, τα μέτρα ακρίβειας μπορούν συχνά να παρέχουν μια ελλιπή εικόνα του πραγματικού ποσοστού μετατροπής ενός συστήματος συστάσεων. Αρκετά άλλα δευτερεύοντα μέτρα παίζουν επίσης ρόλο. Ως εκ τούτου, είναι σημαντικό να σχεδιαστεί προσεκτικά το σύστημα αξιολόγησης, έτσι ώστε οι μετρήσεις να αντικατοπτρίζουν πραγματικά την αποτελεσματικότητα του συστήματος από την πλευρά του χρήστη. Για λόγους ολοκλήρωσης παρουσιάζονται στη συνέχεια τα βασικά μέτρα που χρησιμοποιούνται.

1.7.1.1 Mean Absolue Error (MAE)

Το Mean Absolute Error ή το Μέσο Απόλυτο Σφάλμα αποτελεί τη διαφορά μεταξύ της πραγματικής τιμής (βαθμολογίας) και της προβλεπόμενης τιμής. Το MAE είναι το πιο δημοφιλές και ευρέως χρησιμοποιούμενο. Πρόκειται για ένα μέτρο απόκλισης της σύστασης από τη συγκεκριμένη τιμή του χρήστη. Όσο χαμηλότερη είναι η τιμή MAE, τόσο καλύτερο θα είναι το μοντέλο.

$$MAE = \frac{1}{N}\sum_{u,i}|p_{u,i} - r_{u,i}|$$

όπου $p_{u,i}$ είναι η προβλεπόμενη βαθμολογία για το χρήστη u στο στοιχείο i, $r_{u,i}$ είναι η πραγματική βαθμολογία και N είναι ο συνολικός αριθμός αξιολογήσεων στο σύνολο στοιχείων.

1.7.1.2 Root Mean Squared Error (RMSE)

Το RMSE ή το Ριζικό Μέσο Τετραγωνικό Σφάλμα είναι παρόμοιο με το MAE, με μόνη διαφορά ότι η απόλυτη τιμή του υπολείμματος τετραγωνίζεται και λαμβάνεται η τετραγωνική ρίζα ολόκληρου του όρου για σύγκριση. Το πλεονέκτημα της χρήσης του RMSE έναντι του MAE είναι ότι «τιμωρεί» περισσότερο τον όρο όταν το σφάλμα είναι υψηλό. Το RMSE δίνει μεγαλύτερη έμφαση στο μεγαλύτερο απόλυτο σφάλμα και όσο χαμηλότερο είναι το RMSE, τόσο καλύτερη είναι η ακρίβεια της σύστασης.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(p_{u,i} - r_{u,i})^2}{N}}$$

1.7.1.3 Hit Rate

Το ποσοστό επιτυχίας είναι μια καλύτερη εναλλακτική λύση του MAE ή του RMSE. Για να μετρηθεί το ποσοστό επιτυχίας, δημιουργούμε πρώτα τις κορυφαίες N συστάσεις για όλους τους χρήστες στο σύνολο δεδομένων δοκιμής μας. Κατά τον υπολογισμό του ποσοστού επιτυχίας, δίνεται προτεραιότητα στη λίστα με τις κορυφαίες N συστάσεις και όχι στον χρήστη. Όσο μεγαλύτερο είναι το ποσοστό επιτυχίας, τόσο καλύτερο θα είναι το σύστημα συστάσεων. Το ποσοστό επιτυχίας είναι εύκολο να κατανοηθεί, αλλά η μέτρησή του είναι δύσκολη.

$$HIT\ RATE = \frac{Hits\ in\ test}{Number\ of\ users}$$

Ο καλύτερος τρόπος για τη μέτρησή του είναι με τη χρήση της μεθόδου Leave One Out Cross-Validation. Στην μέθοδο αυτή αρχικά υπολογίζετε η κορυφαία Ν λίστα συστάσεων για κάθε χρήστη στα δεδομένα εκπαίδευσης και γίνεται σκόπιμη αφαίρεση ενός από αυτά τα στοιχεία από τα δεδομένα εκπαίδευσης του χρήστη. Στη συνέχεια, γίνεται δοκιμή της ικανότητας του Συστήματος Συστάσεων να συστήνει "αυτό" το στοιχείο που αφαιρέθηκε σκόπιμα στη φάση δοκιμής.

Ορισμένες προσκλήσεις – περιορισμοί της μεθόδου:

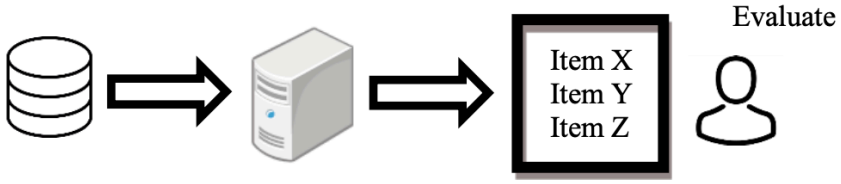
- Είναι δύσκολο να επιτευχθεί ένα συγκεκριμένο στοιχείο σωστά από το να πετύχουμε μία από τις Ν συστάσεις. Ειδικά όταν υπάρχει ένας τεράστιος αριθμός αντικειμένων που πρέπει να αντιμετωπιστεί.
- Το ποσοστό επιτυχίας στο Leave One Out Cross Validation είναι πολύ μικρό και δύσκολο να το μετρηθεί, εκτός αν το dataset που διατίθεται είναι πολύ μεγάλο.

1.7.2 Online μέθοδος

Οι online μέθοδοι αξιολόγησης ονομάζονται επίσης «διαδικτυακό πείραμα» ή «πείραμα ζωντανού χρήστη». Πραγματοποιούνται σε απευθείας σύνδεση με πραγματικούς χρήστες που αλληλοεπιδρούν με ένα σύστημα συστάσεων και τους αξιολογούν συλλέγοντας σχετικές μετρήσεις με τη συμπεριφορά των χρηστών σε πραγματικό χρόνο. Η online μέθοδος αξιολόγησης εξαρτάται από πολλούς παράγοντες όπως για παράδειγμα ο βαθμός εξοικείωσης του χρήστη με τα στοιχεία και τη διεπαφή του συστήματος συστάσεων που το παρουσιάζονται, την πρόθεση του χρήστη να γνωρίζει τις πληροφορίες που χρειάζεται κτλπ.

Σε αντίθεση με τις offline μεθόδους αξιολόγησεις, το online πείραμα έχει την δυνατότητα να συλλέξει σε πραγματικό χρόνο την αλληλεπίδραση των χρηστών που εκτελούν πραγματικές εργασίες με το σύστημα συστάσεων όπως προτιμήσεις, απόψεις, αλληλεπιδράσεις (κλικ) κλπ. γεγονός που μπορεί να βοηθήσει στο σχηματισμό μιας καλής εικόνας για την απόδοση του συστήματος συστάσεων καθώς θα υπάρχουν πιο αξιόπιστα αποτελέσματα.

Τα δυναμικά δεδομένα πραγματικού χρόνου έχουν επίσης και ορισμένα μειονεκτήματα στην αξιολόγηση. Προκειμένου να διατηρηθεί και να αναπτυχθεί ένα σύστημα απαιτούνται ένας μεγάλος αριθμός πόρων όπως κατάλληλη υποστήριξη χρηστών και αρκετή υπολογιστική ικανότητα, όποτε σε αυτή την περίπτωση το σύστημα θα ήταν αρκετά ακριβό για να υλοποιηθεί. Επιπλέον για να δημιουργηθεί και να προετοιμαστεί το περιβάλλον για την δοκιμή – αξιολόγηση του συστήματος συστάσεων θα πρέπει να δαπανηθεί τεράστιος χρόνος. Επίσης, οι χρήστες με την πάροδο του χρόνου γνωρίζουν το σύστημα και τυχόν αλλαγές θα γίνουν αντιληπτές, οπότε θα επηρεάσουν το αποτέλεσμα της αξιολόγησης. (Najmani et al., 2022).



Εικόνα 1.9 Online μέθοδος αξιολόγησης (Πηγή: Najmani et al., 2022)

1.8 Παραδείγματα συστημάτων συστάσεων

Παρακάτω παρουσιάζονται ορισμένα αξιοσημείωτα παραδείγματα συστημάτων συστάσεων στον πραγματικό κόσμο.

1.8.1 Netflix

Η μηχανή συστάσεων του Netflix αποτελεί ίσως το πιο γνωστό και ευρέως χρησιμοποιούμενο σύστημα συστάσεων. Ο αλγόριθμός του βασίζεται στο ιστορικό προβολών, τη βαθμολογία και τη συμπεριφορά αναζήτησης των χρηστών με σκοπό την πρόταση ταινιών και τηλεοπτικών εκπομπών που πιθανόν να ταιριάζει με τις προτιμήσεις του χρήστη.

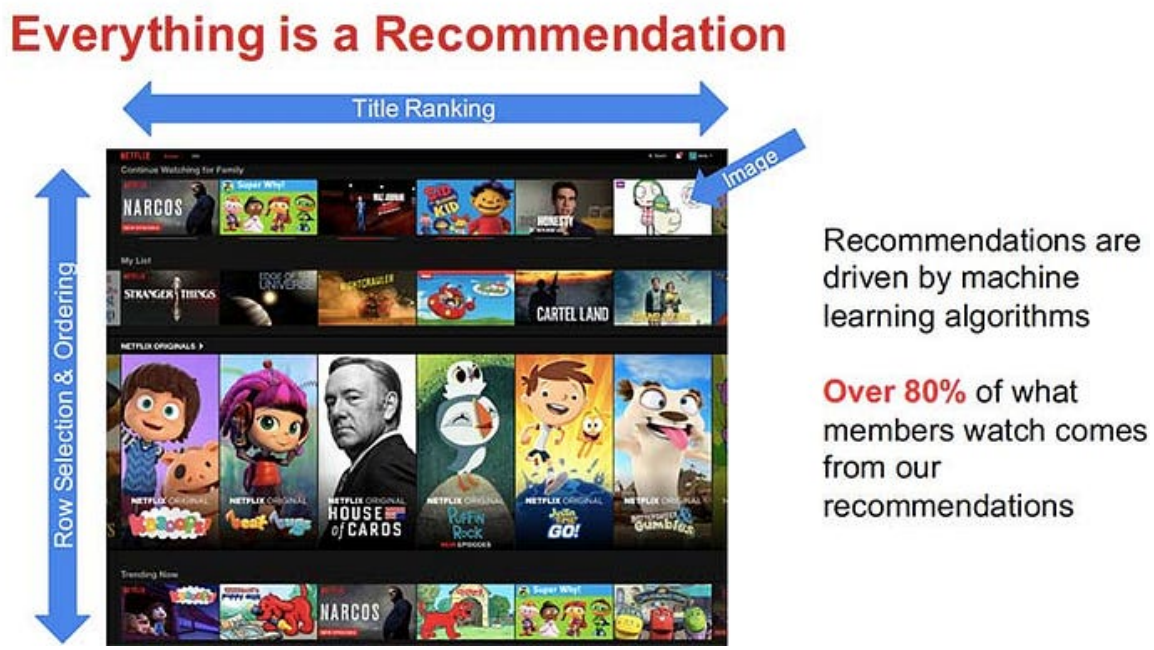
Ορισμένοι από τους αλγορίθμους που χρησιμοποιεί το Netflix είναι:

Εξατομικευμένη κατάταξη βίντεο (Personalized Video Ranking – PVR)

Ο αλγόριθμος αυτός συνήθως φιλτράρει τον κατάλογο των ταινιών με βάση κάποια κριτήρια περιεχομένου (π.χ. αμερικάνικες τηλεοπτικές εκπομπές, κωμωδίες κ.α.) σε συνδυασμό με τα χαρακτηριστικά του χρήστη.

Top – N video Ranker

Έχει σχεδιαστεί για να τον εντοπισμό ενός περιορισμένου αριθμού συστάσεων αλλά ελέγχοντας ολόκληρο τον κατάλογο περιεχομένου. Ενώ ο αλγόριθμος PVR αξιολογείται και βελτιστοποιείται χρησιμοποιώντας μετρικές και αλγορίθμους που εστιάζουν στην κατάταξη που παράγεται για ολόκληρο τον κατάλογο, ο αλγόριθμος Top N video-ranker βασίζεται σε μετρικές και αλγορίθμους που εστιάζουν στα κορυφαία εκατοστημόρια της κατάταξης του καταλόγου (Ko et al., 2022) (Steck et al., 2021).

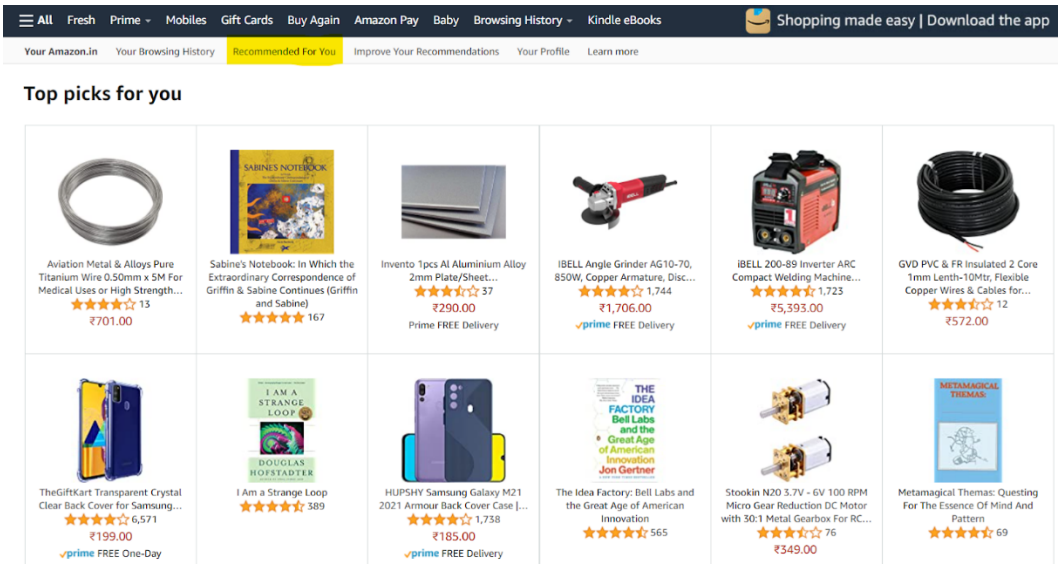


Εικόνα 1.10 Εξατομικευμένες συστάσεις στην πλατφόρμα του Netflix

(Πηγή: https://mikescogs20.medium.com/netflix-recommendation-system-inside-the-algorithm-55edc1712748)

1.8.2 Amazon

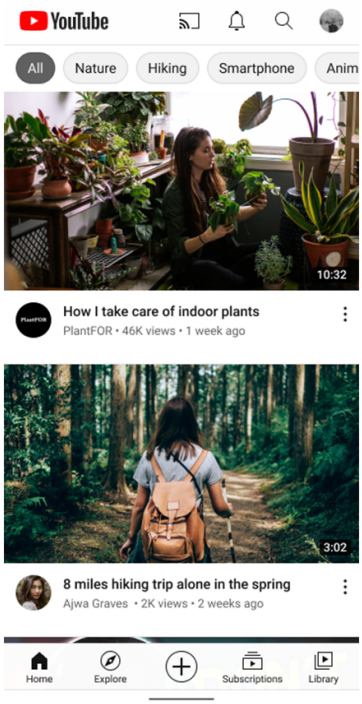
Η μηχανή συστάσεων της Amazon προτείνει προϊόντα με βάση το ιστορικό αγορών, το ιστορικό αναζήτησης και τη συμπεριφορά περιήγησης ενός χρήστη. Πραγματοποιεί εξατομικευμένες συστάσεις βασιζόμενες σε προηγούμενες αγορές του χρήστη, τα προϊόντα που έχει δει και τα αντικείμενα που έχουν προστεθεί στο καλάθι αγορών του (Ko et al., 2022).



Εικόνα 1.11 Παράδειγμα συστάσεων από την πλατφόρμα του Amazon

1.8.3 YouTube

Η μηχανή συστάσεων του YouTube βασίζεται στο ιστορικό προβολής των χρηστών δηλαδή των βίντεο που του άρεσαν καθώς και στο ιστορικό αναζήτησης. Επιπλέον παράγοντες που λαμβάνονται υπόψιν είναι η διάρκεια παρακολούθησης ενός βίντεο, τα αγαπημένα κανάλια του χρήστη και άλλες συνήθειες προβολής με σκοπό την δημιουργία εξατομικευμένων συστάσεων (Ko et al., 2022) (Davidson et al., 2010).



Εικόνα 1.12 Παράδειγμα συστάσεων στην πλατφόρμα του YouTube

2 Μηχανική μάθηση και εξόρυξη δεδομένων

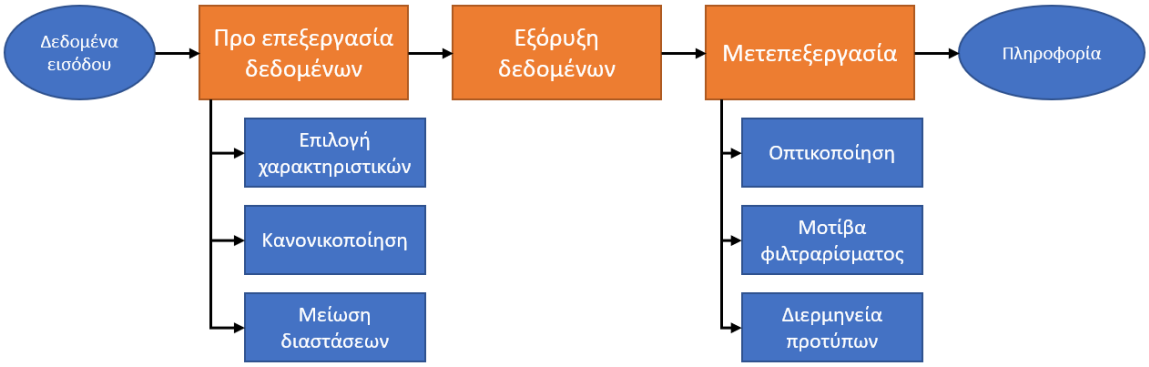
Η μηχανική μάθηση και η εξόρυξη δεδομένων είναι δύο στενά συνδεδεμένοι τομείς στον κόσμο της πληροφορικής, και η συνεργασία τους έχει φέρει ανακαλύψεις και εφαρμογές που έχουν αλλάξει τον τρόπο που αντιλαμβανόμαστε και αναλύουμε τα δεδομένα.

2.1 Εξόρυξη δεδομένων

Η εξόρυξη δεδομένων είναι μια διαδικασία ανάλυσης μεγάλων ποσοτήτων δεδομένων με ενδελεχή και αποτελεσματικό τρόπο, προκειμένου να εξαχθούν χρήσιμα και πιθανώς απροσδόκητα πρότυπα από τα δεδομένα. Πιο συγκεκριμένα πρόκειται για μια διαδικασία η οποία λαμβάνει σαν είσοδο πρωτογενή δεδομένα και τα μετατρέπει σε χρήσιμη πληροφορία. Τα δεδομένα που εισάγονται αποθηκεύονται σε διάφορες μορφές, είτε σε ένα κεντρικό αποθετήριο είτε κατανέμονται σε διάφορες περιοχές. Τα δεδομένα που χρησιμοποιούνται σε αυτή τη διαδικασία συναντάται σε διάφορους τύπους δεδομένων όπως πίνακες, κείμενο, χρονοσειρές, εικόνες, γραφήματα κ.λπ. ή μπορεί να έχουν και μία χωρική ή χρονική διάσταση. Αυτός είναι και ο λόγος για τον οποίο σε επόμενα στάδια υποβάλλονται σε επεξεργασία για να μετατραπούν σε κατάλληλη μορφή για επεξεργασία και εξαγωγής γνώσης (Han et al., 2016). Στην εικόνα 2.1 παρουσιάζεται η διαδικασία της εξόρυξης δεδομένων όπως περιγράφεται και πιο πάνω.

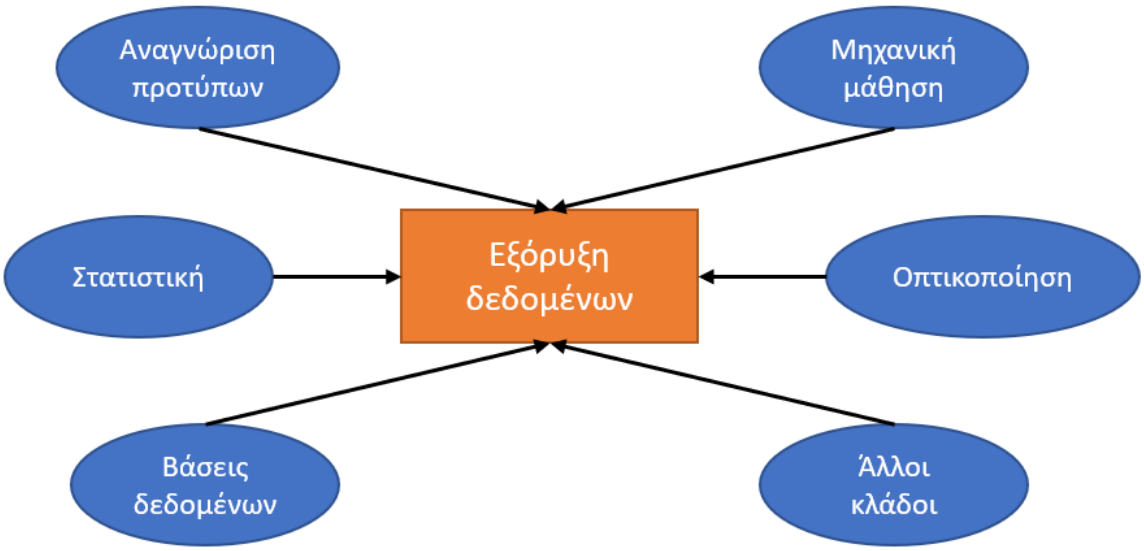
Η εξόρυξη δεδομένων οφείλει:

- Εξόρυξη γνώσης από διαφορετικές πηγές.
- Την διαχείριση τεράστιου μεγέθους δεδομένων.
- Το χειρισμό διαφορετικών τύπων δεδομένων.
- Την απαίτηση για απόδοση και εξέλιξη των αλγορίθμων εξόρυξης δεδομένων.
- Την διαχείριση της ανομοιογένειας και της κατανεμημένης φύσης των δεδομένων.



Εικόνα 2.1 Διαδικασία εξόρυξης δεδομένων.

Η εξόρυξη δεδομένων συνεργάζεται στενά και με άλλους επιστημονικούς κλάδους όπως η στατιστική (statistics), η τεχνητή νοημοσύνη (artificial intelligence), η μηχανική μάθηση (machine learning), τα συστήματα υποστήριξης αποφάσεων (decision support systems), τα συστήματα άμεσης ανάλυσης δεδομένων (OLAP) και το ταίριασμα προτύπων (pattern matching).



Εικόνα 2.2 Κλάδοι σύνδεσης της εξόρυξης δεδομένων

2.1.1 Τεχνικές εξόρυξης δεδομένων

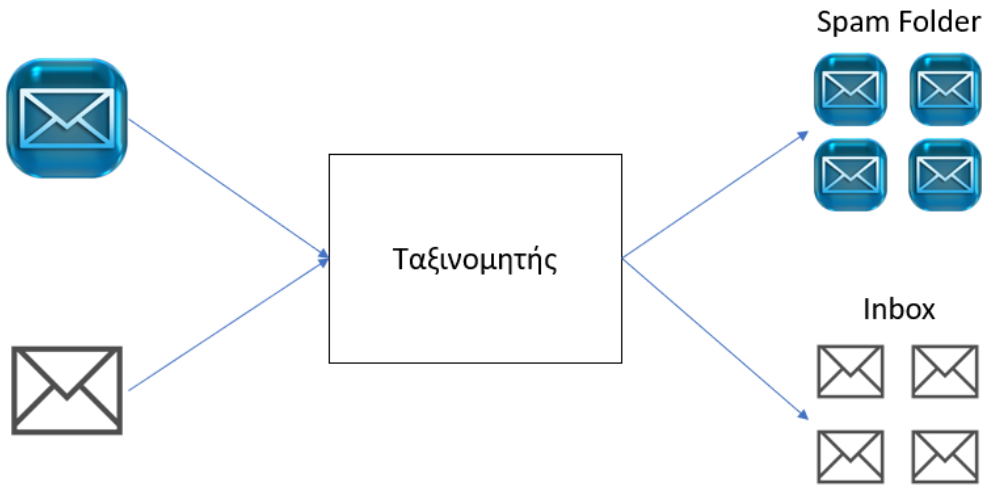
Υπάρχουν πολλές διαφορετικές τεχνικές εξόρυξης δεδομένων διαθέσιμες. Ανάλογα με το είδος της πληροφορίας και τη μέθοδο που χρησιμοποιείται για την εξαγωγή της, ταξινομείται σε διαφορετική κατηγορία. Μερικές βασικές μέθοδοι εξόρυξης δεδομένων περιγράφονται παρακάτω. Είναι σημαντικό να σημειωθεί ότι, στην Επιστήμη Δεδομένων, η μεγαλύτερη διαφορά έγκειται στον τρόπο με τον οποίο χρησιμοποιούνται

τα δεδομένα για εκπαίδευση. Αντί να χρησιμοποιείται ένας δάσκαλος ή κάποιος ειδικός για τη μεταφορά γνώσης από τον έναν στον άλλον, στην Επιστήμη των Δεδομένων η εκπαίδευση πραγματοποιείται με τη χρήση δεδομένων (Verykios, 2015).

Η ποιότητα των πληροφοριών σχετίζεται κυρίως με την ακρίβεια και της ποσότητα της πληροφορίας. Επομένως, σε πολλές εφαρμογές είναι άσκοπο να έχουμε υψηλή απόδοση αποτελεσμάτων με υπολογισμούς που απαιτούν απροσδιόριστα πολύ χρόνο ή να έχουμε πολύ γρήγορα αποτελέσματα πολύ χαμηλής ποιότητας. Η αντιστάθμιση που εμφανίζεται συχνά σε πολλούς αλγόριθμους εξόρυξης δεδομένων είναι η ανάγκη επιλογής μεταξύ ακρίβειας δεδομένων και ταχύτητας επεξεργασίας.

2.1.1.1 Ταξινόμηση (Classification)

Η ταξινόμηση είναι μια θεμελιώδης τεχνική στην εξόρυξη δεδομένων και μπορεί να εφαρμοστεί σχεδόν σε κάθε κλάδο. Πρόκειται για μια διαδικασία κατά την οποία τα σημεία δεδομένων από μεγάλα σύνολα δεδομένων αντιστοιχίζονται σε κατηγορίες με βάση τον τρόπο χρήσης τους. Η τεχνική της ταξινόμησης πρόκειται για μια προγνωστική μέθοδο. Στόχος της είναι η δημιουργία ενός μοντέλου – ταξινομητή (classifier) με βάση τα υπάρχοντα δεδομένα. Ουσιαστικά, είναι η εκμάθηση μιας συνάρτησης που αντιστοιχίζει ένα αντικείμενο στις τιμές μιας κατηγορικής μεταβλητής, γνωστής και ως κλάση (ή κατηγορία). Η έννοια της μάθησης, είναι η συμπεριφορά των ευφυών συστημάτων, που μελετάται από πεδία όπως η μηχανική μάθηση ή η τεχνητή νοημοσύνη. (Verykios, 2015).



Εικόνα 2.3 Ταξινομητές ανεπιθύμητων μηνυμάτων αλληλογραφίας.

Στην εξόρυξη δεδομένων, η ταξινόμηση θεωρείται μια μορφή ομαδοποίησης - δηλαδή, είναι χρήσιμη για την εξαγωγή συγκρίσιμων σημείων δεδομένων για συγκριτική ανάλυση. Η ταξινόμηση χρησιμοποιείται επίσης για τον προσδιορισμό ευρέων ομάδων στο πλαίσιο ενός δημογραφικού στοιχείου, ενός κοινού-στόχου ή μιας βάσης χρηστών, μέσω των οποίων οι επιχειρήσεις μπορούν να αποκτήσουν ισχυρότερες γνώσεις. Υπάρχουν διάφοροι μέθοδοι για την ταξινόμηση εξόρυξης δεδομένων όπως:

- **Λογιστική παλινδρόμηση:** Αυτός ο αλγόριθμος προσπαθεί να δείξει την πιθανότητα ενός συγκεκριμένου αποτελέσματος εντός δύο πιθανών αποτελεσμάτων. Για παράδειγμα, μια υπηρεσία ηλεκτρονικού ταχυδρομείου μπορεί να χρησιμοποιήσει τη λογιστική παλινδρόμηση για να προβλέψει αν ένα μήνυμα ηλεκτρονικού ταχυδρομείου είναι ανεπιθύμητο ή όχι.
- **Δέντρα αποφάσεων:** Αφού ταξινομηθούν τα δεδομένα, μπορούν να τεθούν ερωτήσεις παρακολούθησης και τα αποτελέσματα να διαμορφωθούν σε ένα διάγραμμα που ονομάζεται δέντρο αποφάσεων. Για παράδειγμα, αν μια εταιρεία υπολογιστών θέλει να προβλέψει την πιθανότητα αγοράς φορητών υπολογιστών, μπορεί να ρωτήσει: Είναι ο δυνητικός αγοραστής φοιτητής; Τα δεδομένα ταξινομούνται σε δέντρα αποφάσεων "Ναι" και "Όχι", ενώ άλλες ερωτήσεις μπορούν να τεθούν στη συνέχεια με παρόμοιο τρόπο.
- **Κ – κοντινότεροι γείτονες (KNN):** Αφού ταξινομηθούν τα δεδομένα, μπορούν να τεθούν ερωτήσεις παρακολούθησης και τα αποτελέσματα να διαμορφωθούν σε ένα διάγραμμα που ονομάζεται δέντρο αποφάσεων. Για παράδειγμα, αν μια εταιρεία υπολογιστών θέλει να προβλέψει την πιθανότητα αγοράς φορητών υπολογιστών, μπορεί να ρωτήσει: Είναι ο δυνητικός αγοραστής φοιτητής; Τα δεδομένα ταξινομούνται σε δέντρα αποφάσεων "Ναι" και "Όχι", ενώ άλλες ερωτήσεις μπορούν να τεθούν στη συνέχεια με παρόμοιο τρόπο.
- **Naive Bayes**: Βασισμένος στο θεώρημα των πιθανοτήτων Bayes, αυτός ο αλγόριθμος χρησιμοποιεί ιστορικά δεδομένα για να προβλέψει αν θα συμβούν παρόμοια γεγονότα με βάση ένα διαφορετικό σύνολο δεδομένων.
- **Μηχανή διανυσμάτων υποστήριξης (SVM)**: Αυτός ο αλγόριθμος μηχανικής μάθησης χρησιμοποιείται συχνά για τον καθορισμό της γραμμής που χωρίζει καλύτερα ένα σύνολο δεδομένων σε δύο κλάσεις. Ένα SVM μπορεί να βοηθήσει στην ταξινόμηση εικόνων και χρησιμοποιείται σε λογισμικό αναγνώρισης προσώπου και γραφής.

Παραδείγματα ταξινόμησης στις επιχειρήσεις αποτελούν τα χρηματοπιστωτικά ιδρύματα όπου ταξινομούν τους καταναλωτές με βάση πολλές μεταβλητές για να προωθήσουν νέα δάνεια ή να προβλέψουν τους κινδύνους πιστωτικών καρτών. Εν τω μεταξύ, οι εφαρμογές καιρού ταξινομούν τα δεδομένα για να προβλέψουν τα σύνολα χιονόπτωσης και άλλα παρόμοια στοιχεία. Τα παντοπωλεία χρησιμοποιούν επίσης την ταξινόμηση για να ομαδοποιήσουν τα προϊόντα με βάση τους καταναλωτές που τα αγοράζουν, βοηθώντας στην πρόβλεψη των αγοραστικών προτύπων. (Neha and Reddy, 2020)

2.1.1.2 Συσταδοποίηση (Clustering)

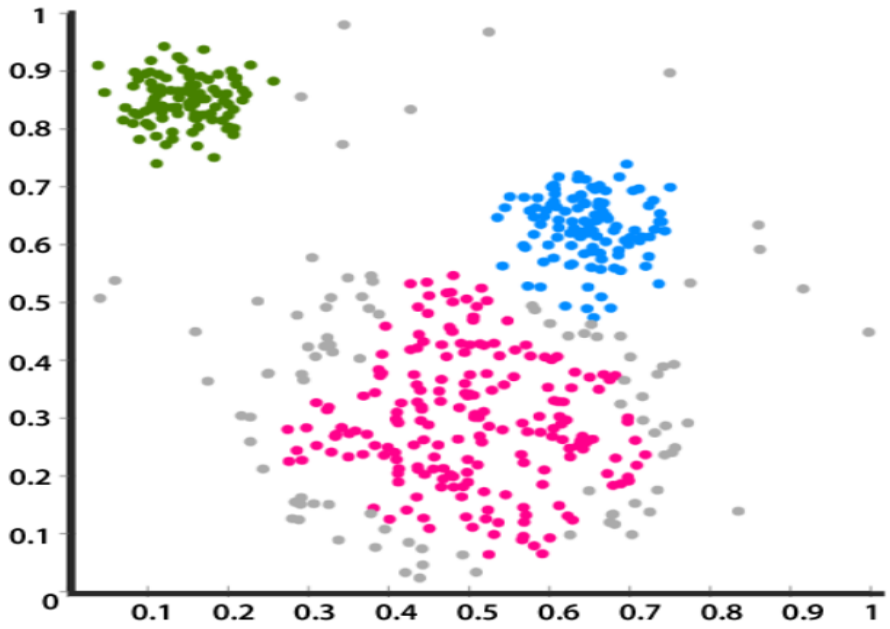
Η συσταδοποίηση είναι μια τεχνική που χρησιμοποιείται για την οπτική αναπαράσταση δεδομένων - όπως σε γραφήματα που δείχνουν τάσεις αγοράς ή δημογραφικά στοιχεία πωλήσεων για ένα συγκεκριμένο προϊόν. Η συσταδοποίηση αναφέρεται στη διαδικασία ομαδοποίησης μιας σειράς διαφορετικών σημείων δεδομένων με βάση τα χαρακτηριστικά τους. Με τον τρόπο αυτό, οι data miners μπορούν να χωρίσουν απρόσκοπτα τα δεδομένα σε υποσύνολα, επιτρέποντας πιο τεκμηριωμένες αποφάσεις όσον αφορά ευρύτερα δημογραφικά στοιχεία (όπως καταναλωτές ή χρήστες) και τις αντίστοιχες συμπεριφορές τους.

Οι μέθοδοι που χρησιμοποιούνται συνήθως για τη συσταδοποίηση των δεδομένων είναι:

- **Μέθοδος διαχωρισμού**: Αυτή περιλαμβάνει τη διαίρεση ενός συνόλου δεδομένων σε μια ομάδα συγκεκριμένων συστάδων για αξιολόγηση με βάση τα κριτήρια κάθε μεμονωμένης συστάδας. Σε αυτή τη μέθοδο, τα σημεία δεδομένων ανήκουν σε μία μόνο ομάδα ή συστάδα.
- **Ιεραρχική μέθοδος**: Με την ιεραρχική μέθοδο, τα σημεία δεδομένων αποτελούν μια ενιαία συστάδα, τα οποία ομαδοποιούνται με βάση τις ομοιότητες. Αυτές οι νεοδημιουργηθείσες συστάδες μπορούν στη συνέχεια να αναλυθούν ξεχωριστά η μία από την άλλη.
- **Μέθοδος με βάση την πυκνότητα**: Μια μέθοδος μηχανικής μάθησης όπου τα σημεία δεδομένων που απεικονίζονται μαζί αναλύονται περαιτέρω, αλλά τα σημεία δεδομένων από μόνα τους χαρακτηρίζονται ως "θόρυβος" και απορρίπτονται.
- **Μέθοδος βασισμένη σε πλέγμα**: Αυτή περιλαμβάνει τη διαίρεση των δεδομένων σε κελιά σε ένα πλέγμα, τα οποία στη συνέχεια μπορούν να ομαδοποιηθούν με βάση

μεμονωμένα κελιά και όχι με βάση ολόκληρη τη βάση δεδομένων. Ως αποτέλεσμα, η συσταδοποίηση με βάση το πλέγμα έχει γρήγορο χρόνο επεξεργασίας.

- **Μέθοδος με βάση το μοντέλο**: Σε αυτή τη μέθοδο, δημιουργούνται μοντέλα για κάθε συστάδα δεδομένων για τον εντοπισμό των καλύτερων δεδομένων που ταιριάζουν στο συγκεκριμένο μοντέλο.

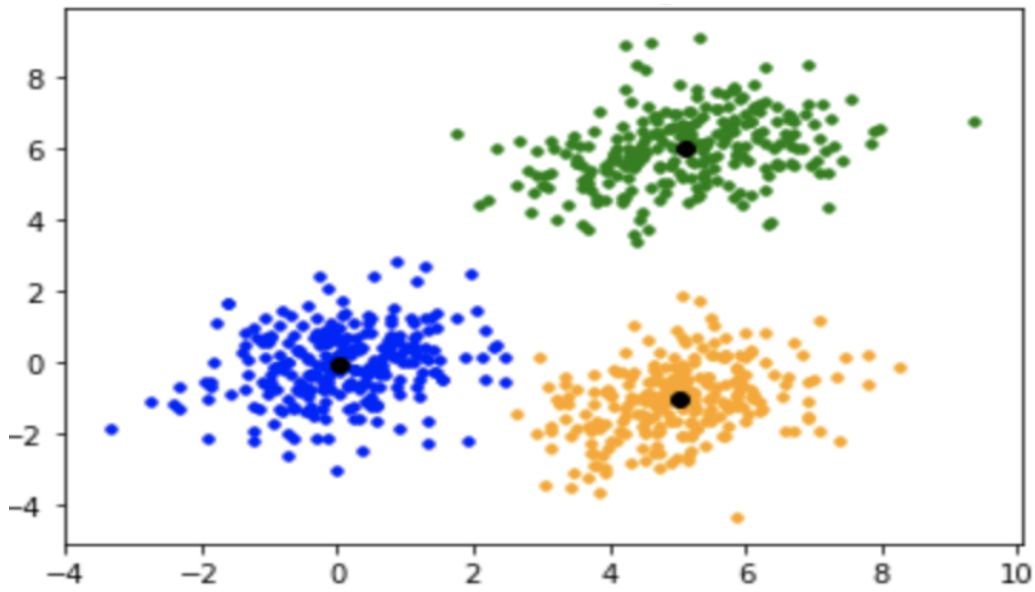


Εικόνα 2.4 Συσταδοποίηση με μοντέλο πυκνότητας

Η συσταδοποίηση βοηθά τις επιχειρήσεις να διαχειρίζονται αποτελεσματικότερα τα δεδομένα τους. Για παράδειγμα, οι λιανοπωλητές μπορούν να χρησιμοποιούν μοντέλα συσταδοποίησης για να καθορίσουν ποιοι πελάτες αγοράζουν συγκεκριμένα προϊόντα, σε ποιες ημέρες και με ποια συχνότητα. Αυτό μπορεί να βοηθήσει τους λιανοπωλητές να στοχεύσουν προϊόντα και υπηρεσίες σε πελάτες που ανήκουν σε μια συγκεκριμένη δημογραφική ομάδα ή περιοχή.

Η συσταδοποίηση μπορεί να βοηθήσει τα παντοπωλεία να ομαδοποιήσουν τα προϊόντα με βάση διάφορα χαρακτηριστικά (μάρκα, μέγεθος, κόστος, γεύση κ.λπ.) και να κατανοήσουν καλύτερα τις τάσεις των πωλήσεών τους. Μπορεί επίσης να βοηθήσει τις εταιρείες ασφάλισης αυτοκινήτων που θέλουν να προσδιορίσουν ένα σύνολο πελατών που έχουν συνήθως υψηλές ετήσιες αποζημιώσεις, προκειμένου να τιμολογούν αποτελεσματικότερα τα συμβόλαια. Επιπλέον, οι τράπεζες και τα χρηματοπιστωτικά ιδρύματα θα μπορούσαν να χρησιμοποιήσουν την ομαδοποίηση για να κατανοήσουν καλύτερα τον τρόπο με τον οποίο οι πελάτες χρησιμοποιούν τις προσωπικές έναντι των

εικονικών υπηρεσιών, ώστε να σχεδιάσουν καλύτερα τις ώρες λειτουργίας των καταστημάτων και το προσωπικό. (Neha and Reddy, 2020)



Εικόνα 2.5 Συσταδοποίηση με κεντροειδή μοντέλο

2.1.1.3 Κανόνες συσχέτισης (Association Rules)

Οι κανόνες συσχέτισης χρησιμοποιούνται για την εύρεση συσχετίσεων μεταξύ σημείων σε ένα σύνολο δεδομένων. Οι ειδικοί με την εξόρυξη δεδομένων χρησιμοποιούν τη συσχέτιση για να ανακαλύψουν μοναδικές ή ενδιαφέρουσες σχέσεις μεταξύ μεταβλητών σε βάσεις δεδομένων. Η συσχέτιση χρησιμοποιείται συχνά για να βοηθήσει τις εταιρείες να καθορίσουν την έρευνα και τη στρατηγική μάρκετινγκ. Υπάρχουν δύο κύριες προσεγγίσεις που χρησιμοποιούν τη συσχέτιση στην εξόρυξη δεδομένων είναι η μονοδιάστατη και η πολυδιάστατη μέθοδος.

- **Μονοδιάστατη συσχέτιση**: Αυτό περιλαμβάνει την αναζήτηση μιας επαναλαμβανόμενης περίπτωσης ενός σημείου δεδομένων ή ενός χαρακτηριστικού. Για παράδειγμα, ένας έμπορος λιανικής πώλησης μπορεί να αναζητήσει στη βάση δεδομένων του τις περιπτώσεις αγοράς ενός συγκεκριμένου προϊόντος.
- **Πολυδιάστατη συσχέτιση**: Αυτό περιλαμβάνει την αναζήτηση περισσότερων από ένα σημείων δεδομένων σε ένα σύνολο δεδομένων. Ο ίδιος λιανοπωλητής μπορεί να θέλει να γνωρίζει περισσότερες πληροφορίες από το τι αγόρασε ένας πελάτης - όπως η ηλικία του, η μέθοδος αγοράς (μετρητά ή πιστωτική κάρτα) ή η ηλικία του.

Η ανάλυση της αυτοσχέδιας αγοραστικής συμπεριφοράς είναι ένα παράδειγμα συσχέτισης - δηλαδή, οι λιανοπωλητές παρατηρούν σε μελέτες δεδομένων ότι οι γονείς που ψωνίζουν για προμήθειες παιδικής φροντίδας είναι πιο πιθανό να αγοράσουν ειδικά τρόφιμα ή ποτά για τον εαυτό τους κατά τη διάρκεια του ίδιου ταξιδιού. Αυτές οι αγορές μπορούν να αναλυθούν μέσω στατιστικής συσχέτισης.

Η ανάλυση συσχέτισης έχει πολλές άλλες χρήσεις στις επιχειρήσεις. Για τους λιανοπωλητές, είναι ιδιαίτερα χρήσιμη για τη διατύπωση προτάσεων αγοράς. Για παράδειγμα, εάν ένας πελάτης αγοράσει ένα smartphone, ένα tablet ή μια συσκευή βιντεοπαιχνιδιών, η ανάλυση συσχέτισης μπορεί να προτείνει σχετικά αντικείμενα όπως καλώδια, εφαρμόσιμο λογισμικό και προστατευτικές θήκες. Επιπλέον, η συσχέτιση χρησιμοποιείται από την κυβέρνηση για τη χρησιμοποίηση δεδομένων απογραφής και το σχεδιασμό δημόσιων υπηρεσιών- χρησιμοποιείται επίσης από τους γιατρούς για την αποτελεσματικότερη διάγνωση διαφόρων ασθενειών και καταστάσεων (Neha and Reddy, 2020).

2.1.1.4 Παλινδρόμηση (Regression)

Η παλινδρόμηση είναι μια διαδικασία που χρησιμοποιείται για την πρόβλεψη αριθμητικών δεδομένων που λείπουν ή δεν είναι διαθέσιμα και όχι ξεχωριστή ονομασία κλάσεων όπως στην ταξινόμηση. Αυτή είναι μια άλλη μέθοδος για την πρόβλεψη των αποτελεσμάτων. Ο σκοπός της διαδικασίας είναι να αναπτύξει μια συνάρτηση αντιστοίχισης που παίρνει ένα αντικείμενο και το εκχωρεί σε μια πραγματική μεταβλητή. Ορισμένοι παράγοντες που μπορούν να επηρεάσουν τις τιμές μιας εξαρτημένης μεταβλητής χρησιμοποιούνται για την πρόβλεψή τους.

Η παλινδρόμηση επιτρέπει τις προβλέψεις από δεδομένα, μαθαίνοντας τη σχέση μεταξύ χαρακτηριστικών των δεδομένων και μερικών παρατηρήσεων με συνεχείς τιμές. Κοινοί αλγόριθμοι παλινδρόμησης είναι:

- Γραμμική παλινδρόμηση
- Μη – γραμμική παλινδρόμηση
- Γενικευμένα γραμμικά μοντέλα
- Δέντρα απόφασης
- Νευρωνικά δίκτυα

Η απλούστερη και παλαιότερη μορφή παλινδρόμησης είναι η γραμμική παλινδρόμηση που χρησιμοποιείται για την εκτίμηση μιας σχέσης μεταξύ δύο μεταβλητών. Η τεχνική αυτή χρησιμοποιεί τον μαθηματικό τύπο της ευθείας γραμμής ($y = mx + b$). Με απλά λόγια, αυτό σημαίνει απλώς ότι, δεδομένης μιας γραφικής παράστασης με άξονα Y και X, η σχέση μεταξύ X και Y είναι μια ευθεία γραμμή με λίγες ακραίες τιμές. Άλλες τεχνικές παλινδρόμησης είναι οι εξής (Chapple, 2022) :

- **Τυπική πολλαπλή παλινδρόμηση** εξετάζει ταυτόχρονα όλες τις μεταβλητές πρόβλεψης. Για παράδειγμα, 1) ποια είναι η σχέση μεταξύ του εισοδήματος και της εκπαίδευσης (προβλεπτικές μεταβλητές) και της επιλογής της γειτονιάς (προβλεπόμενη μεταβλητή)- και 2) σε ποιο βαθμό συμβάλλει κάθε μία από τις επιμέρους προβλεπτικές μεταβλητές σε αυτή τη σχέση;
- **Βηματική πολλαπλή παλινδρόμηση** απαντά σε ένα εντελώς διαφορετικό ερώτημα. Ένας αλγόριθμος βηματικής παλινδρόμησης θα αναλύσει ποιοι προγνωστικοί παράγοντες χρησιμοποιούνται καλύτερα για την πρόβλεψη της επιλογής γειτονιάς - δηλαδή το βηματικό μοντέλο αξιολογεί τη σειρά σπουδαιότητας των μεταβλητών πρόβλεψης και στη συνέχεια επιλέγει ένα σχετικό υποσύνολο. Αυτός ο τύπος προβλήματος παλινδρόμησης χρησιμοποιεί "βήματα" για την ανάπτυξη της εξίσωσης παλινδρόμησης. Δεδομένου αυτού του τύπου παλινδρόμησης, όλοι οι προγνωστικοί παράγοντες ενδέχεται να μην εμφανίζονται καν στην τελική εξίσωση παλινδρόμησης.
- **Ιεραρχική παλινδρόμηση**, όπως και η βηματική, είναι μια διαδοχική διαδικασία, αλλά οι μεταβλητές πρόβλεψης εισάγονται στο μοντέλο με μια προκαθορισμένη σειρά που καθορίζεται εκ των προτέρων, δηλαδή ο αλγόριθμος δεν περιέχει ένα ενσωματωμένο σύνολο εξισώσεων για τον καθορισμό της σειράς εισαγωγής των προβλεπτικών παραγόντων. Αυτό χρησιμοποιείται συχνότερα όταν το άτομο που δημιουργεί την εξίσωση παλινδρόμησης έχει εξειδικευμένη γνώση του τομέα.
- **Παλινδρόμηση κατά ομάδες** είναι επίσης παρόμοια με τη βηματική παλινδρόμηση, αλλά αναλύει σύνολα μεταβλητών και όχι μεμονωμένες μεταβλητές.

2.1.2 Περιγραφή μεθοδολογίας εξόρυξης δεδομένων

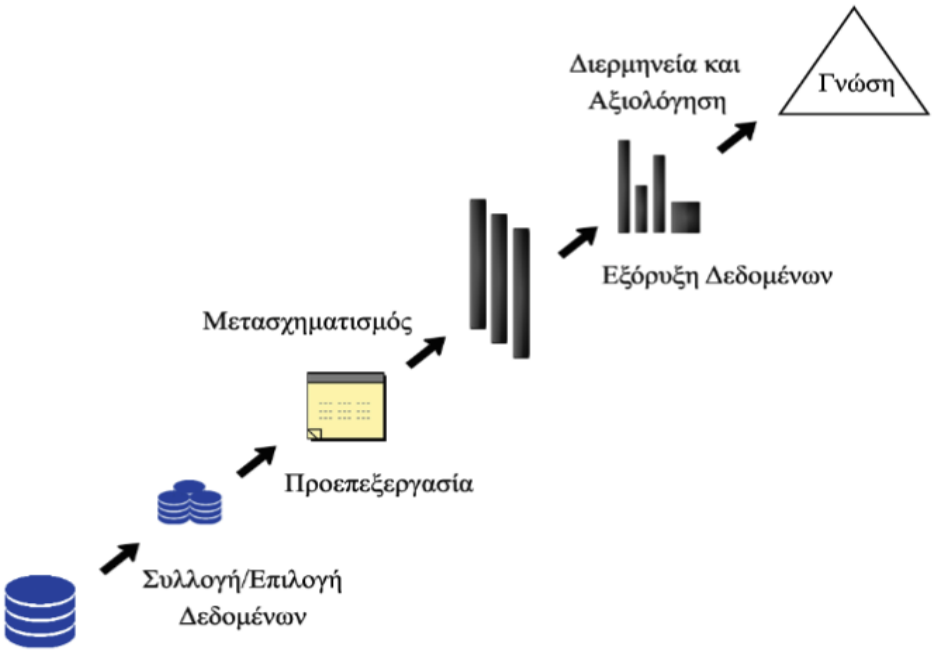
Η διαδικασία εξαγωγής προτύπων από δεδομένα συμβαίνει εδώ και αρκετά χρόνια. Οι πρώτες μέθοδοι για τον εντοπισμό προτύπων ήταν εκείνες που βασίστηκαν στην θεωρία Bayes και στην ανάλυσης της παλινδρόμησης. Η αυξανόμενη χρήση της τεχνολογίας των υπολογιστών έχει οδηγήσει σε αύξηση του όγκου των δεδομένων, γεγονός που έχει δημιουργήσει την ανάγκη για συστήματα που μπορούν να διαχειριστούν αποτελεσματικά αυτές τις πληροφορίες. Η αυτοματοποιημένη επεξεργασία δεδομένων έχει αντικαταστήσει τη χειροκίνητη ανάλυση δεδομένων ως την τυπική μέθοδο για την ανάλυση δεδομένων. Αυτό οφείλεται στον αυξανόμενο όγκο και την πολυπλοκότητα των συλλογών δεδομένων. Σε αυτό συνέβαλαν άλλες ανακαλύψεις της επιστήμης των υπολογιστών, όπως τα νευρωνικά δίκτυα, η ομαδοποίηση, οι γενετικοί αλγόριθμοι, τα δέντρα αποφάσεων και οι μηχανές διανυσμάτων υποστήριξης.

Η εξόρυξη δεδομένων είναι η διαδικασία χρήσης αυτών των μεθόδων για την εξερεύνηση δεδομένων με σκοπό την εύρεση κρυφών μοτίβων. Αυτό γεφυρώνει το χάσμα μεταξύ εφαρμοσμένων στατιστικών και τεχνητής νοημοσύνης χρησιμοποιώντας τεχνικές διαχείρισης βάσεων δεδομένων. Για να κάνουμε τη θεωρία και τους διαθέσιμους αλγόριθμους να λειτουργούν πιο γρήγορα και αποτελεσματικά, πρέπει να τους κάνουμε πιο αποτελεσματικούς όταν εφαρμόζονται σε μεγάλα σύνολα δεδομένων.

Η εξόρυξη δεδομένων συχνά θεωρείται συνώνυμη με ένα άλλο ευρέως χρησιμοποιούμενο όρο, την ανακάλυψη γνώσης από δεδομένα. Άλλοι απλώς βλέπουν την εξόρυξη δεδομένων ως ένα ουσιαστικό βήμα στη διαδικασία ανακάλυψης γνώσης. Η ανακάλυψη γνώσης είναι μια διαδικασία που αποτελείται από τα παρακάτω βήματα:

1. **Καθαρισμός δεδομένων:** Αφορά την αφαίρεση του θορύβου, πληροφοριών και εσφαλμένων δεδομένων που είναι περιττές.
2. **Ενσωμάτωση δεδομένων:** Συνδυασμός πολλαπλών πηγών δεδομένων.
3. **Συλλογή και επιλογή δεδομένων:** Αφορά την διαδικασία ανάλυσης των δεδομένων που ανακτώνται από τη βάση δεδομένων.
4. **Μετασχηματισμός δεδομένων:** Στη διαδικασία αυτή τα δεδομένα μετασχηματίζονται ή ενοποιούνται σε μορφές κατάλληλες για εξόρυξη, εκτελώντας λειτουργίες σύνοψης ή συνάθροισης για παράδειγμα.
5. **Εξόρυξη δεδομένων:** Χρήση έξυπνων τεχνικών και διαδικασιών για να ανακαλυφθούν πρότυπα.

6. **Ερμηνεία / Αξιολόγηση προτύπων:** Κατά την διάρκεια αυτή της διαδικασίας αναγνωρίζονται τα πραγματικά ενδιαφέροντα πρότυπα που οδηγούν στην δημιουργία γνώσης. Η αξιολόγηση γίνεται με βάση ορισμένους δείκτες αποτελεσματικότητας.
7. **Παρουσίαση της γνώσης:** Αφορά την χρήση τεχνικών παρουσίασης με σκοπό την οπτικοποίηση των δεδομένων. Στις περισσότερες φορές τα αποτελέσματα παρουσιάζονται με τη μορφή διαγραμμάτων.



Εικόνα 2.6 Στάδια εξόρυξης δεδομένων

Μερικά από τα στάδια της διαδικασίας ανακάλυψης γνώσης μπορούν να επαναληφθούν πολλές φορές. Τα βήματα επιλογής, της προ επεξεργασίας και του μετασχηματισμού αναφέρονται συχνά ως βήματα προετοιμασίας δεδομένων. Η προσπάθεια που απαιτείται σε κάθε στάδιο της διαδικασίας δεν είναι η ίδια.

2.1.2.1 Ολοκλήρωση και καθαρισμός δεδομένων

Η αφετηρία για την ανάλυση των δεδομένων είναι τα πηγαία δεδομένα. Γενικά τα δεδομένα αποθηκεύονται σε διάφορες πηγές, όπως συστήματα παρακολούθησης συναλλαγών, ανεξάρτητες βάσεις δεδομένων, ανεξάρτητα αρχεία, εξωτερικές πηγές κ.λπ. Αυτά τα δεδομένα συνήθως έχουν διάφορα προβλήματα, όπως σφάλματα, αντιφάσεις και τιμές που λείπουν. Τα δεδομένα πρέπει να συλλέγονται από διάφορες

πηγές να ομογενοποιηθούν και να καθαριστούν πριν χρησιμοποιηθούν. Δεδομένα με ανακριβείς τιμές ή τιμές που λείπουν μπορούν να οδηγήσουν τους αλγόριθμους εξόρυξης να παράγουν μη έγκυρα και ανακριβή αποτελέσματα. Ορισμένοι αλγόριθμοι εξόρυξης δεδομένων είναι σε θέση να αντιμετωπίσουν προβλήματα δεδομένων μόνοι τους, αλλά αυτό δεν συμβαίνει πάντα. Ο καλύτερος τρόπος αντιμετώπισης προβλημάτων δεδομένων μπορεί αν μην είναι πάντα η πιο αποτελεσματική προσέγγιση. Είναι γενικά προτιμότερο ο αναλυτής να καθαρίζει τα δεδομένα μόνος του παρά από κάποια αυτοματοποιημένη διαδικασία. Με αυτόν τον τρόπο ο αναλυτής έχει περισσότερο έλεγχο στη διαδικασία και μπορεί να διασφαλίσει ότι τα δεδομένα καθαρίζονται σωστά. Συνήθως τα δεδομένα αφού απαλλαγούν από τα προβλήματα αποθηκεύονται σε μια βάση δεδομένων.

2.1.2.2 Συλλογή δεδομένων (Selection)

Κατά τη φάση επιλογής δεδομένων, ορίζεται το σύνολο δεδομένων στο οποίο θα αναζητηθεί το μοτίβο. Η επιλογή των δεδομένων είναι πολύ σημαντική γιατί καθορίζει τα αποτελέσματα που θα έχει το μοντέλο. Συχνά ο τρόπος αποθήκευσης των δεδομένων δεν ταιριάζει με τους αλγορίθμους ανακάλυψης που απαιτούνται για την εξαγωγή των δεδομένων και την οργάνωσή τους σε δομές που θα είναι προσβάσιμες από αυτούς.

2.1.2.3 Προ επεξεργασία (Preprocessing)

Προτού τα δεδομένα είναι έτοιμα να χρησιμοποιηθούν στο επόμενο στάδιο της διαδικασίας ανακάλυψης γνώσης απαιτείται πολλή δουλεία και χρόνος. Στο στάδιο αυτό σκοπός είναι η διατήρηση των δεδομένων που είναι χρήσιμα για την εξαγωγή συμπερασμάτων και την αντιμετώπιση περιπτώσεων όπου τα δεδομένα είναι ελλιπή, παραπλανητικά ή θορυβώδη. Αυτή η φάση είναι επίσης γνωστή ως η φάση καθαρισμού δεδομένων.

2.1.2.4 Μετασχηματισμός (Transformation)

Η προ επεξεργασία δεδομένων είναι ένα σημαντικό μέρος της εξόρυξης δεδομένων και είναι απαραίτητη πριν από την εφαρμογή των αλγορίθμων. Τα δεδομένα συλλέγονται με μεθόδους που δεν επιτρέπουν την εξαγωγή γνώσης από αυτά, με αποτέλεσμα να δημιουργείται θόρυβος και να επηρεάζεται αρνητικά η ποιότητά τους. Υπάρχουν πολλοί διαφορετικοί τρόποι συλλογής δεδομένων, συμπεριλαμβανομένων των απαντήσεων σε ερωτηματολόγια, των συσκευών μέτρησης, των συνεντεύξεων, των παρατηρήσεων ή του

Διαδικτύου. Ο θόρυβος και οι ασυνεπείς τιμές μπορεί να είναι αποτέλεσμα ανθρώπινου λάθους κατά την εισαγωγή τιμών ή από προβλήματα της συσκευής.

Οι μέθοδοι εξόρυξης δεδομένων βασίζονται έντονα σε δεδομένα. Τα δεδομένα θα αποτελέσουν τη βάση για τυχόν συμπεράσματα. Η επιλογή των σωστών δεδομένων είναι ζωτικής σημασίας για τη λήψη ακριβών αποφάσεων. Αρχικά η επιλογή δεδομένων αναφέρεται στην επιλογή ενός ή πολλών χαρακτηριστικών. Η επιλογή των χαρακτηριστικών σχετίζεται άμεσα με την εργασία που εκτελεί ο αναλυτής. Κάποια χαρακτηριστικά μπορεί να είναι κατάλληλα για μια εργασία παρά για μια άλλη. Ο αναλυτής αρχικά επιλέγει χαρακτηριστικά που πιστεύει ότι περιέχουν σημαντικές πληροφορίες σχετικές με την ανάλυσή του.

Η αρχική επιλογή των χαρακτηριστικών δεν είναι αρκετή. Στην αρχική φάση της ανάλυσης, ο αναλυτής αποκλείει χαρακτηριστικά που φαίνονται άσχετα με την ανάλυσή του. Στη συνέχεια όμως, η επιλογή μπορεί να μην είναι προφανής, καθώς το ίδιο μέγεθος μπορεί να καταγράφεται με διαφορετικούς τρόπους.

Στο στάδιο επιλογής χαρακτηριστικών, πραγματοποιείται και ο μετασχηματισμός των δεδομένων. Για παράδειγμα, οι αριθμοί μπορούν να μετατραπούν σε άλλους αριθμούς ή οι αριθμοί μπορούν να μετατραπούν σε ονομαστικές τιμές. Κάθε φορά που χρησιμοποιούνται μέθοδοι ανάλυσης που απαιτούν διαφορετικά δεδομένα από τα αυτά που χρησιμοποιούνται αρχικά, μπορεί να χρειαστεί να γίνει προσαρμογή των δεδομένων. Ορισμένες μέθοδοι ταξινόμησης είναι πιο αποτελεσματικές όταν εργάζονται με μεγάλες τιμές και λιγότερο αποτελεσματικές όταν εργάζονται με μικρές τιμές. Σε τέτοιες περιπτώσεις, τα μεγέθη των τιμών στα διάφορα πεδία πρέπει να είναι συγκρίσιμα. Το τελικό αποτέλεσμα αυτής της διαδικασίας είναι ένα σύνολο δεδομένων που θα χρησιμοποιηθεί για την αναγνώριση προτύπων.

2.1.2.5 Εκπαίδευση (Train / Testing)

Στο στάδιο αυτό γίνεται επιλογή τις τεχνικής που θα ακολουθηθεί δηλαδή ποιος αλγόριθμος θα εφαρμοστεί. Αυτό εξαρτάται από το είδος της γνώσης που θα αναζητηθεί. Υπάρχουν δύο είδη γνώσης που μπορούν να αποκτηθούν τα πρότυπα πληροφοριών και τα πρότυπα πρόβλεψης. Η εξόρυξη δεδομένων και η ανακάλυψη γνώσης συχνά υποκαθιστούν ο ένας τον άλλον. Σύμφωνα με την παραπάνω διαδικασία, η φάση εξόρυξης δεδομένων είναι ένα μόνο βήμα στη συνολική διαδικασία εξόρυξης γνώσης.

2.1.2.6 Ερμηνεία αποτελεσμάτων (Interpretation – Evaluation)

Μόλις δημιουργηθεί το μοντέλο, πρέπει να αξιολογηθεί για τα αποτελέσματά του. Αυτή η διαδικασία αξιολόγησης πρέπει στη συνέχεια να χρησιμοποιηθεί για να προσδιοριστεί η σημασία του μοντέλου. Σε προβλήματα κατηγοριοποίησης, ο πίνακας σύγχυσης (confusion matrix) αποτελεί ένα χρήσιμο εργαλείο για την κατανόηση των αποτελεσμάτων. Είναι σημαντικό να γίνεται χρήση διαφόρων γραφικών αναπαραστάσεων για την κατανόηση των αποτελεσμάτων της μελέτης.

2.2 Μηχανική μάθηση (Machine Learning)

Η Μηχανική Μάθηση είναι πεδίο της επιστήμης υπολογιστών που ανήκει στον τομέα της Τεχνητής Νοημοσύνης και ασχολείται με τη δημιουργία αλγορίθμων που έχουν ως στόχο να «μαθαίνουν» από δεδομένα, δηλαδή να αποκτούν την ικανότητα στην επιπλέον πρόσκτηση γνώσης μέσω της αλληλεπίδρασης με το περιβάλλον στο οποίο δραστηριοποιούνται και την ικανότητα να βελτιώνουν μέσω της επανάληψης τον τρόπο τον οποίο εκτελούν την ενέργεια αυτή. Σύμφωνα με τον Άρθουρ Σάμουελ (Samuel, 1959), η μηχανική μάθηση «*δίνει τη δυνατότητα στους υπολογιστές να μαθαίνουν χωρίς να έχουν ρητά προγραμματιστεί*».

Ο πιο επίσημος ορισμός δόθηκε από τον Τομ Μ. Μίτσελ (Mitchell, 1997), σύμφωνα με τον οποίο: «*Ένα πρόγραμμα υπολογιστή θεωρείται ότι μαθαίνει από μία εμπειρία Ε, σε σχέση με μία σειρά από έργα Τ και απόδοση μετρημένη με Ρ, αν η απόδοση του στα έργα Τ, μετρημένη με Ρ, βελτιώνεται με την εμπειρία Ε*».

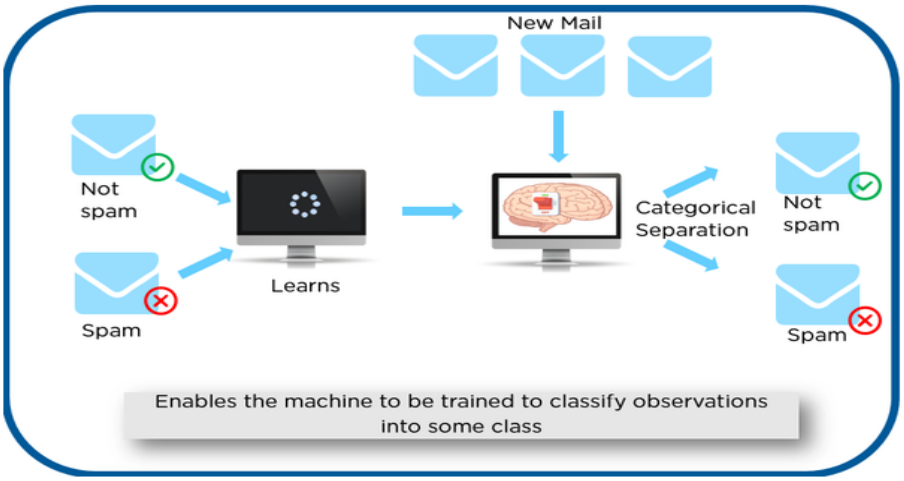
Εφαρμόζεται σε μία πληθώρα εφαρμογών, στις οποίες ο ρητός σχεδιασμός και προγραμματισμός δεν είναι εφικτός, όπως για παράδειγμα στις μηχανές αναζήτησης, στην οπτική αναγνώριση χαρακτήρων, στα φίλτρα ανεπιθύμητων μηνυμάτων, κ.λπ. Η ΕΔ χρησιμοποιεί τεχνικές πρόβλεψης ή κατηγοριοποίησης της μηχανικής μάθησης. Με τη μηχανική μάθηση σε ένα σύστημα, είναι εφικτή η πρόβλεψη, που μέσω της ανατροφοδότησης (feedback), μπορεί να οδηγήσει στη μάθηση. Η μάθηση βασίζεται στα παραδείγματα, την αποθηκευμένη γνώση, και την ανατροφοδότηση. Όταν μελλοντικά συμβεί ανάλογη περίπτωση, η ανατροφοδότηση χρησιμοποιείται για να κάνει την ίδια πρόβλεψη ή για να κάνει μια εντελώς διαφορετική πρόβλεψη. Η στατιστική είναι πολύ σημαντική σε προγράμματα μηχανικής μάθησης γιατί τα αποτελέσματα των προβλέψεων πρέπει να είναι στατιστικά σημαντικά.

Τα είδη της μηχανικής μάθησης χωρίζονται βάσει του αποτελέσματος και του τρόπου με τον οποίο λειτουργεί το σύστημα εκμάθησης και ανατροφοδοτείται. Υπάρχουν τέσσερις κύριες κατηγορίες:

- Επιβλεπόμενη μάθηση
- Μη-επιβλεπόμενη μάθηση
- Ημι-επιβλεπόμενη μάθηση
- Ενισχυτική μάθηση

2.2.1 Επιβλεπόμενη μάθηση (Supervised Learning)

Ο όρος supervised learning αναφέρεται στη διαδικασία του να τροφοδοτήσεις έναν αλγόριθμο με εγγραφές στις οποίες μια μεταβλητή απόκρισης (output variable) που μας ενδιαφέρει είναι γνωστή και ο αλγόριθμος προσπαθεί να «μάθει» πώς να προβλέψει την τιμή με νέες εγγραφές όπου το αποτέλεσμα είναι άγνωστο. Δηλαδή, μοντελοποιούν μια μεταβλητή απόκρισης βασιζόμενοι σε μία ή περισσότερες επεξηγηματικές μεταβλητές (input variable). Το σύνολο δεδομένων είναι μια συλλογή από ζευγάρια εισόδων και εξόδου της μορφής (x_i , y_i), i = 1,2,…,n. Το x_i ονομάζεται διάνυσμα χαρακτηριστικών, όπου κάθε τιμή του διανύσματος περιγράφει το i-οστό στοιχείο του συνόλου δεδομένων (εγγραφή). Το y_i είναι η “ετικέτα” και είναι συνήθως ένα στοιχείο από ένα σύνολο κατηγοριών ή ένας πραγματικός αριθμός, αλλά μπορεί να είναι και μια πιο σύνθετη δομή (διάνυσμα, γράφος, κ.ο.κ). Το x_i είναι το σύνολο των ανεξάρτητων μεταβλητών και το y_i είναι η εξαρτημένη μεταβλητή. Πρακτικά, το μοντέλο με ένα σύνολο δεδομένων εκπαίδευσης βρίσκει σχέσεις ανάμεσα στις τιμές του διανύσματος χαρακτηριστικών, αφού γνωρίζει ποια είναι η επιθυμητή έξοδος. Έπειτα από την διαδικασία εκπαίδευσης και αφού θα έχει ικανοποιητικά αποτελέσματα, το μοντέλο θα πρέπει να είναι σε θέση να προβλέπει την κατηγορία (ή ετικέτα) μιας άγνωστης εγγραφής, με μεγάλη πιθανότητα, σωστά. Αλγόριθμοι που βασίζονται σε αυτήν την μάθηση, είναι αλγόριθμοι κατηγοριοποίησης και παλινδρόμησης. (Burkov, 2019; Heidenreich, 2018)



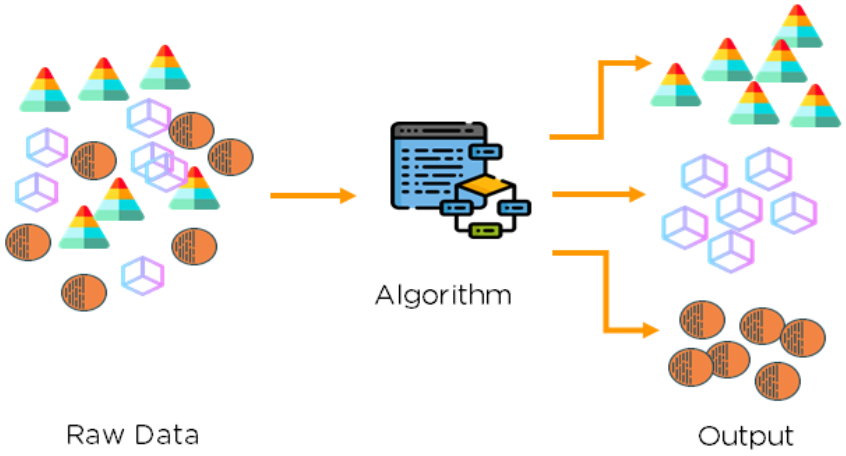
Εικόνα 2.7 Παράδειγμα Επιβλεπόμενης Μάθησης. Κατηγοριοποίηση email. (Heidenreich, 2018)

Γνωστοί αλγόριθμοι/τεχνικές επιβλεπόμενης μάθησης είναι:

- Δέντρα Απόφασης
- Support Vector Machines (SVM)
- Naïve Bayes
- K-nearest neighbors (kNN) κ.ά.

2.2.2 Μη επιβλεπόμενη μάθηση (Unsupervised Learning)

Η μη-επιβλεπόμενη μάθηση χρησιμοποιείται όταν δεν υπάρχει μεταβλητή απόκρισης να προβλεφθεί ή να ταξινομηθεί. Πιο συγκεκριμένα, ο όρος unsupervised learning αναφέρεται στην ανάλυση που κάποιος επιχειρεί, για να μάθει κάτι άλλο για τα δεδομένα πέρα από την πρόβλεψη της τιμής μίας μεταβλητής που τον ενδιαφέρει, όπως για παράδειγμα, το αν ανήκει σε κάποιο cluster. Οι τεχνικές μάθησης χωρίς επίβλεψη χρησιμοποιούνται όταν δεν υπάρχει κάποιο πεδίο να προβλεφθεί αλλά, διερευνώνται οι σχέσεις μεταξύ των δεδομένων ώστε να ανακαλυφθεί η γενική δομή τους.



Εικόνα 2.8 Παράδειγμα συσταδοποίησης. (Heidenreich, 2018)

Με άλλα λόγια, πρόκειται για αυτοματοποιημένη παραγωγή νέας γνώσης, όπου δεν είναι γνωστές οι επιθυμητές έξοδοι για το σύνολο εκπαίδευσης και το μοντέλο κατασκευάζεται γενικά για κάποιο σύνολο εισόδων με σκοπό να βρει τη δομή του συνόλου αυτού. Ενδεικτικοί αλγόριθμοι αυτής της κατηγορίας μάθησης είναι:

- Ο αλγόριθμος Apriori,
- k-means,
- Density-based spatial clustering of applications with noise (DBSCAN),
- Autoencoders
- Local Outlier Factor κ.ά.

2.2.3 Ημι – επιβλεπόμενη μάθηση (Semi – supervised learning)

Στην ημι-επιβλεπόμενη μάθηση, τα δεδομένα περιέχουν και δεδομένα με ετικέτα και χωρίς ετικέτα. Ο αριθμός των δεδομένων με ετικέτα είναι συνήθως πολύ υψηλότερος. Σκοπός της ημι-επιβλεπόμενης μάθησης είναι ίδιος με την επιβλεπόμενη μάθηση, με την διαφορά ότι η χρήση πολλών δεδομένων χωρίς ετικέτα οδηγούν στην δημιουργία ενός καλύτερου μοντέλου, για κάποιες περιπτώσεις (Burkov, 2019; Heidenreich, 2018).

2.2.4 Ενισχυτική μάθηση (Reinforcement learning)

Ενισχυτική μάθηση είναι μια τεχνική, όπου το μοντέλο επηρεάζεται από το περιβάλλον. Το μοντέλο αντιλαμβάνεται την κατάσταση του περιβάλλοντος, ως ένα διάνυσμα χαρακτηριστικών. Το μοντέλο είναι σε θέση να δρα. Η κάθε δράση του μοντέλου ανταμείβεται ή να τιμωρείται, και μπορεί να αλλάξει την κατάσταση του περιβάλλοντος.

Σκοπός του μοντέλου είναι να εκτελεί την καλύτερη δράση, σε κάθε κατάσταση του περιβάλλοντος. Η ενισχυτική μάθηση χρησιμοποιείται σε ηλεκτρονικά παιχνίδια, στην ρομποτική, σε προσομοιώσεις και σε συστήματα διαχείρισης πόρων (Burkov, 2019; Heidenreich, 2018).

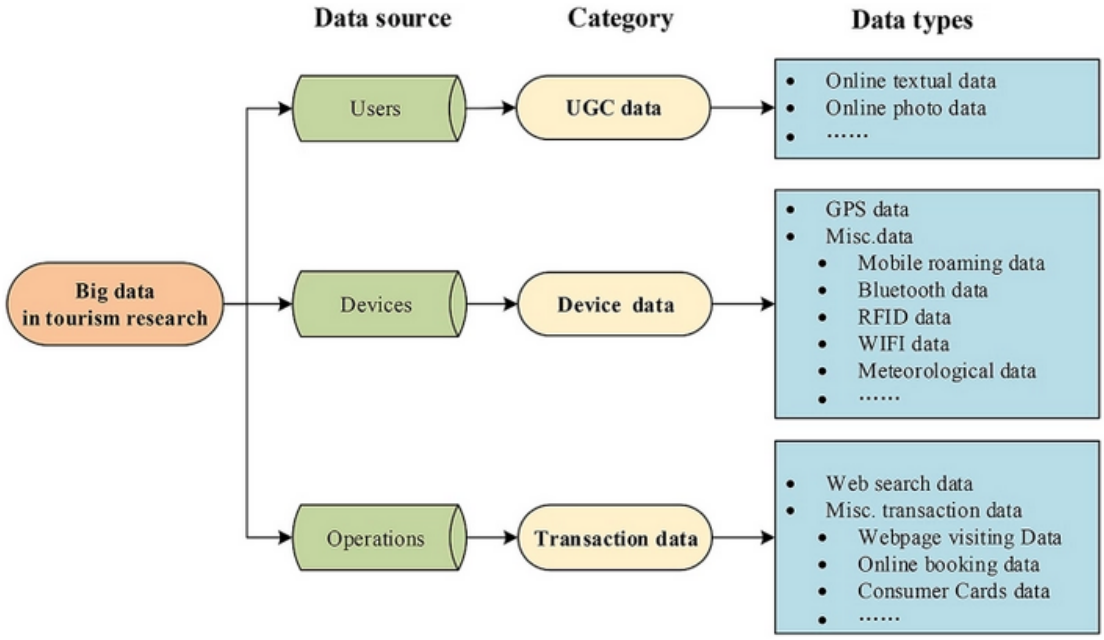
Ενδεικτικοί αλγόριθμοι ενισχυτικής μάθησης είναι οι:

- Monte Carlo
- Q-Learning
- SARSA
- DQN

3 Δεδομένα σε τουριστικές εφαρμογές

Τα μεγάλα δεδομένα που σχετίζονται με τον τουρισμό διακρίνονται σε τρεις κύριες κατηγορίες:

- UGC (User Generated Content) δεδομένα (που παράγονται από τους χρήστες)
- Δεδομένα συσκευών (που προέρχονται από συσκευές)
- Δεδομένα συναλλαγών (που προέρχονται από ενέργειες)



Εικόνα 3.1 Διαχωρισμός μεγάλων δεδομένων για τουριστικές εφαρμογές

3.1 Δεδομένα UGC

Η συντομογραφία UGC αναφέρεται στην έκφραση “User Generated Data” δηλαδή δεδομένα που παράγονται από χρήστες. Στην ψηφιακή εποχή, η έκρηξη του διαδικτύου και των κοινωνικών μέσων ενημέρωσης έχει αλλάξει δραματικά τον τρόπο λειτουργίας του τουρισμού, παρέχοντας πλέον μια ευρύχωρη πλατφόρμα για την ανταλλαγή δεδομένων UGC. Αποτελούν μια σημαντική κατηγορία μεγάλων δεδομένων τα οποία μπορούν να αποκτηθούν με ιδιαίτερη ευκολία και χρησιμοποιούνται για την προώθηση της τουριστικής έρευνας μέσω δύο τύπων δεδομένων οι οποίοι είναι (Li et al., 2018):

- Διαδικτυακά δεδομένα κειμένου, όπως κριτικές προϊόντων και ιστολόγια που δημοσιεύονται στα μέσα κοινωνικής δικτύωσης και ερωτηματολόγια.

- Δεδομένα φωτογραφιών που δημοσιεύονται σε ιστότοπους κοινής χρήσης φωτογραφιών.

3.1.1 Διαδικτυακά δεδομένα κινητού

Τα μέσα κοινωνικής δικτύωσης τα οποία ευδοκιμούν με την έντονη ανάπτυξη του Διαδικτύου προσφέρουν πλατφόρμες στις οποίες οι τουρίστες μπορούν να διαδώσουν πλήθος πληροφοριών σχετικές με τον τουρισμό όπως ταξιδιωτικές κριτικές και εμπειρίες. Για παράδειγμα, οι ταξιδιώτες μπορούν να εκφράσουν την ικανοποίηση και τη δυσαρέσκειά τους προς τα προϊόντα ή της υπηρεσίες μίας τουριστικής επιχείρησης δημιουργώντας μία πλούσια συλλογή διαδικτυακών κριτικών. Οι τουρίστες μπορούν επίσης να διαμοιράσουν τις ταξιδιωτικές τους εμπειρίες σε ιστολόγια όπως το Twitter και το Facebook παρέχοντας πληροφορίες προς άλλους τουρίστες. Το σύνολο των διαδικτυακών κριτικών, τα δεδομένα των ιστολογίων και άλλα συναφή δεδομένα με την μορφή κειμένου συνιστούν ένα ειδικό τύπο μεγάλων δεδομένων κειμένου που μεταφέρουν συναισθήματα, αισθήματα και διαθέσεις των τουριστών.

3.1.2 Δεδομένα φωτογραφιών

Με την έντονη εξέλιξη του IoT, ποικίλες συσκευές ή αισθητήρες έχουν αναπτυχθεί και χρησιμοποιούνται για την παρακολούθηση της κίνησης των τουριστών παρέχοντας μαζικά δεδομένα υψηλής ποιότητας για την τουριστική διαχείριση. Η απόκτηση αυτών των δεδομένων είναι αρκετά δαπανηρή καθώς το κόστος τους εξαρτάται από τις περιοχές που καλύπτονται και την περίοδο που διατίθεται για τη μελέτη. Τα δεδομένα αυτά αποτελούνται από δεδομένα GPS, δεδομένα περιαγωγής κινητής τηλεφωνίας, Bluetooth δεδομένα, δεδομένα RFID και δεδομένα Wi-Fi. Επιπλέον οι αυτόματοι αισθητήρες μετεωρολογικών σταθμών συλλέγουν ένα μεγάλο σύνολο μετεωρολογικών δεδομένων για την εξυπηρέτηση της λήψης ταξιδιωτικών αποφάσεων, δεδομένου ότι ο καιρός αποτελεί σημαντικό παράγοντα για τον τουρισμό. Όλα τα παραπάνω δεδομένα, δομημένα και αδόμητα έχουν ήδη εφαρμοστεί στην τουριστική έρευνα και όπως διαπιστώνεται επιφέρουν ιδιαίτερα οφέλη (Li et al., 2018).

3.2 Δεδομένα συναλλαγών ανά ενέργεια χρήστη

Αποτελεί άλλον ένα σημαντικό τύπο μεγάλων δεδομένων για την τουριστική έρευνα ο οποίος καταγράφει λειτουργίες που σχετίζονται με τον τουρισμό (ή συναλλαγές και δραστηριότητες στον τουριστικό τομέα) όπως ή αναζήτηση στο διαδίκτυο, η

επισκεψιμότητα των ιστοσελίδων, ηλεκτρονικές κρατήσεις και αγορές. Τα δεδομένα αυτά χρησιμοποιούνται ήδη για την προώθηση τουριστικών υπηρεσιών, βελτιστοποίηση μηχανών αναζήτησης (SEO), την τουριστική συμπεριφορά και το τουριστικό μάρκετινγκ. Συνήθως για τη λήψη αυτών των δεδομένων χρησιμοποιούνται προηγμένες υπηρεσίες ιστού, όπως το Google Analytics.

3.2.1 Δεδομένα αναζήτησης από το διαδίκτυο (Search Data)

Οι μηχανές αναζήτησης αποτελούν μια αναδυόμενη πηγή μεγάλων δεδομένων για την τουριστική έρευνα επιτρέποντας και καταγράφοντας τις λειτουργίες αναζήτησης στο διαδίκτυο για περιεχόμενα που σχετίζονται με τον τουρισμό. Ειδικότερα οι τουρίστες μπορούν να αναζητήσουν ταξιδιωτικές πληροφορίες μέσω μιας μηχανής αναζήτησης, αφήνοντας τα αντίστοιχα ίχνη αναζήτησης. Τα ίχνη αυτά καταγράφονται και επεξεργάζονται με σκοπό τον σχηματισμό ενός μεγαλύτερου είδους δεδομένων τα οποία ονομάζονται δεδομένα αναζήτησης στον ιστό και αντικατοπτρίζουν άμεσα την προσοχή του κοινού προς ένα τουριστικό αντικείμενο, συμβάλλοντας θετικά στην τουριστική αγορά.

3.2.2 Δεδομένα επισκεψιμότητας ιστοσελίδας

Βοηθούν στην κατανόηση της συμπεριφοράς των επισκεπτών κατά την περιήγηση στο διαδίκτυο όπως για παράδειγμα πως βρίσκουν τον ιστότοπο και πως αλληλοεπιδρούν με αυτόν βελτιώνοντας έτσι το διαδικτυακό μάρκετινγκ όσον αφορά την προσαρμογή του περιεχομένου και τον σχεδιασμό του ιστότοπου. Έρευνα έχει δείξει ότι με βάση τα μοντέλα παλινδρόμησης οι νέες επισκέψεις και οι επαναλαμβανόμενες επισκέψεις σε τουριστικούς ιστότοπους επηρεάζουν τις σελίδες ανά επίσκεψη (Li et al., 2018).

3.3 Μεθοδολογία συλλογής δεδομένων

Η μεθοδολογία που ακολουθείται για τη συλλογή δεδομένων βασίζεται σε ορισμένες αρχές, οι οποίες παρουσιάζονται παρακάτω:

- Πριν την έναρξη της έρευνας, θα πρέπει να καθοριστεί η ανάλυση των δεδομένων. Η μεθοδολογία της έρευνας που θα ακολουθηθεί, θα πρέπει να προσαρμοστεί σε αυτό που αναζητάτε και όχι το αντίστροφο.

- Κατά τη διάρκεια της έρευνας διατυπώνονται, συχνά, ο προβληματισμός και η διαμόρφωση της ανάλυσης των δεδομένων. Ωστόσο, είναι προτιμότερο να διατυπώνονται εξ αρχής.
- Είναι σημαντικό να διευκρινιστούν διεξοδικά οι ερευνητικές υποθέσεις, καθώς με βάση αυτών θα καθοριστούν τα ερευνητικά εργαλεία που θα χρησιμοποιηθούν και η ανάλυση που θα πραγματοποιηθεί.
- Τα ερευνητικά εργαλεία που θα χρησιμοποιηθούν μπορεί να είναι διαφορετικά, καθώς είναι δυνατό να μην παρέχουν όλα τα ερευνητικά εργαλεία τα ίδια ικανοποιητικά αποτελέσματα.
- Το θεωρητικό πλαίσιο της έρευνας, ο προβληματισμός και οι ερευνητικές υποθέσεις συνδέονται άμεσα με τη μεθοδολογία της συλλογής δεδομένων.

Για τη διεξαγωγή μιας έρευνας, υπάρχουν τρία είδη πηγών πληροφοριών: ο λόγος (συνέντευξη, ερωτηματολόγιο), τα γεγονότα (παρατήρηση) και, τέλος, τα «ίχνη» (γραπτά, στατιστικές).

Οι τρεις από τις πιο γνωστές μεθόδους συλλογής υλικού είναι:

- Η παρατήρηση
- Η συνέντευξη
- Το ερωτηματολόγιο
- Καταγραφή δραστηριοτήτων χρήστη

3.3.1 Παρατήρηση

Η παρατήρηση αποτελεί τη διαδικασία στην οποία κάποιο φαινόμενο ή συμπεριφορά παρατηρείται προγραμματισμένα, οργανωμένα, συστηματοποιημένα, από άτομα ειδικευμένα ή εκπαιδευμένα για τον ρόλο αυτό. Στην παρατήρηση τα γεγονότα καταγράφονται και στη συνέχεια επαληθεύονται. Η παρατήρηση διακρίνεται σε άμεση, συμμετοχική και έμμεση ή προκαλούμενη ή πειραματική παρατήρηση.

3.3.2 Συνέντευξη

Στη συνέντευξη, η οποία αποτελεί μια από τις πιο γνωστές μεθόδους συλλογής υλικού, υποβάλλονται ερωτήσεις από τον ερευνητή στον ερωτώμενο. Αυτό στο οποίο αποσκοπεί ο ερευνητής είναι να μάθει τί σκέφτεται ο ερωτώμενος σχετικά με κάποιο ζήτημα και στη συνέχεια, να συγκρίνει τις γνώμες και τις απόψεις όλων των ερωτώμενων. Ύστερα

από αυτό, γίνεται η ομαδοποίηση των απόψεων των ερωτώμενων. Η διαδικασία της συνέντευξης διακρίνεται σε κατευθυνόμενη ή δομημένη, ημι-κατευθυνόμενη και ελεύθερη συνέντευξη.

Στο ερωτηματολόγιο, το οποίο αποτελεί ένα έντυπο, περιλαμβάνεται μια σειρά από δομημένες ερωτήσεις, οι οποίες θα πρέπει να απαντηθούν γραπτά και με μία συγκεκριμένη σειρά από τον ερωτώμενο. Η συλλογή δεδομένων στα ερωτηματολόγια γίνεται ζητώντας να απαντηθούν οι ίδιες ακριβώς ερωτήσεις (Lagoumintzis, Vlachopoulos and Koutsogiannis, 2015).

3.3.3 Ερωτηματολόγια

Τα ερωτηματολόγια αποτελούν ερευνητικά εργαλεία που χρησιμοποιούν ένα σύνολο ερωτήσεων για τη συλλογή δεδομένων από τους ερωτώμενους. Τα μέσα αυτά περιλαμβάνουν είτε γραπτές είτε προφορικές ερωτήσεις. Τα ερωτηματολόγια μπορεί να είναι ποιοτικά ή ποσοτικά και μπορούν να διεξάγονται διαδικτυακά, τηλεφωνικά, σε χαρτί ή πρόσωπο με πρόσωπο, ενώ οι ερωτήσεις δεν είναι απαραίτητο να χορηγούνται με την παρουσία ερευνητή.

Τα ερωτηματολόγια διαθέτουν είτε ανοικτές είτε κλειστές ερωτήσεις και μερικές φορές χρησιμοποιούν ένα μείγμα και των δύο. Οι ερωτήσεις ανοικτού τύπου επιτρέπουν στους ερωτώμενους να απαντήσουν με δικά τους λόγια με όσες λεπτομέρειες επιθυμούν. Οι κλειστές ερωτήσεις παρέχουν στους ερωτώμενους μια σειρά προκαθορισμένων απαντήσεων από τις οποίες μπορούν να επιλέξουν .

Αν και η σημασία των ερωτηματολογίων στην έρευνα είναι σαφής, υπάρχουν πλεονεκτήματα και μειονεκτήματα στη χρήση αυτών των μέσων για τη συλλογή πληροφοριών (Flores, 2021).

Πλεονεκτήματα των ερωτηματολογίων

Ορισμένα από τα πλεονεκτήματα της χρήσης ερωτηματολογίων ως ερευνητικά εργαλεία περιλαμβάνουν:

- **Πρακτικότητα:** Τα ερωτηματολόγια επιτρέπουν στους ερευνητές να διαχειρίζονται στρατηγικά το κοινό – στόχο, τις ερωτήσεις και τη μορφή τους, ενώ παράλληλα συγκεντρώνουν μεγάλες ποσότητες δεδομένων για οποιοδήποτε θέμα.

- **Ταχύτητα:** Τα δεδομένα της έρευνας συγκεντρώνονται γρήγορα και αβίαστα χρησιμοποιώντας φορητά «εργαλεία».
- **Οικονομική αποδοτικότητα:** Μέσω των ερωτηματολογίων δεν χρειάζεται η πρόσληψη ερευνητών για τον διαμοιρασμό των ερωτήσεων, αντίθετα μπορεί να τις τοποθετήσετε σε κάποιο ιστότοπο, ή να αποσταλθούν μέσω ηλεκτρονικού ταχυδρομείου στους ερωτηθέντες με μικρό έως καθόλου κόστος.
- **Ευκολία ανάλυσης:** Τα ερωτηματολόγια διαθέτουν ορισμένα ενσωματωμένα εργαλεία που αυτοματοποιούν τις αναλύσεις, καθιστώντας την ερμηνεία των αποτελεσμάτων τις ανάλυσης γρηγορότερη και ευκολότερη.
- **Συγκρισιμότητα:** Οι ερευνητές μπορούν να χρησιμοποιούν το ίδιο ερωτηματολόγιο κάθε χρόνο προκειμένου να συγκρίνουν και να αντιπαραβάλλουν τα αποτελέσματα των ερευνών για να αποκτήσουν πολύτιμες πληροφορίες.
- **Άνεση των ερωτηθέντων:** Κατά τη συμπλήρωση ενός ερωτηματολογίου, οι ερωτηθέντες είναι εντελώς ανώνυμοι και δεν υπόκεινται σε αγχωτικούς χρονικούς περιορισμούς, βοηθώντας τους να αισθάνονται χαλαροί και ενθαρρύνοντάς τους να παρέχουν ειλικρινείς απαντήσεις.
- **Επεκτασιμότητα:** Τα ερωτηματολόγια είναι εξαιρετικά επεκτάσιμα, επιτρέποντας στους ερευνητές να τα διανέμουν σε δημογραφικές ομάδες οπουδήποτε στον κόσμο.

Μειονεκτήματα των ερωτηματολογίων

Τα ερωτηματολόγια έχουν επίσης και ορισμένα μειονεκτήματα όπως:

- **Παράλειψη ερωτήσεων:** Οι ερευνητές θα πρέπει να βεβαιωθούν ότι απαιτούν από τους ερωτώμενους να απαντήσουν όλες τις ερωτήσεις της έρευνας. Διαφορετικά μπορεί να υπάρχει ο κίνδυνος οι ερωτηθέντες να αφήσουν τις ερωτήσεις αναπάντητες.
- **Κόπωση από την έρευνα:** Οι ερωτηθέντες μπορεί να βιώσουν κόπωση από την έρευνα, εάν λάβουν ένα ερωτηματολόγιο πολύ μεγάλο.
- **Δυσκολίες ερμηνείας:** Εάν μια ερώτηση δεν είναι αρκετά απλή, οι ερωτηθέντες μπορεί να δυσκολευτούν να την ερμηνεύσουν με ακρίβεια. Για αυτό είναι σημαντικό οι ερωτήσεις να διατυπώνονται με σαφήνεια και συντομία καθώς και με επεξηγήσεις όταν είναι απαραίτητο.
- **Προσκλήσεις ανάλυσης:** Αν και οι κλειστές ερωτήσεις είναι εύκολο να αναλυθούν, οι ανοικτές ερωτήσεις προϋποθέτουν τον ανθρώπινο παράγοντα για την

εξέταση και ερμηνεία τους. Οι ερευνητές θα πρέπει να περιορίσουν τις ανοιχτές ερωτήσεις για να αποκτήσουν περισσότερα ποσοτικοποιημένα δεδομένα τα οποία μπορούν να αξιολογήσουν και να αξιοποιήσουν πιο γρήγορα.

3.3.4 Καταγραφή δραστηριοτήτων χρήστη (Clickstreams)

Τα δεδομένα ροής κλικ και η ανάλυση ροής κλικ αποτελούν διαδικασίες που εμπλέκονται στη συλλογή, ανάλυση και αναφορά συγκεντρωτικών δεδομένων σχετικά με το ποιες σελίδες επισκέπτεται ένας χρήστης ενός ιστοτόπου και με ποια σειρά. Η διαδρομή που ακολουθεί ο χρήστης σε έναν ιστότοπο ονομάζεται ροή κλικ. Η ροή κλικ αφορά τη σύνδεση των ενεργειών που έχει πραγματοποιήσει ένας χρήστης σε μία μόνο συνεδρία. Αυτό σημαίνει ότι προσδιορίζεται πού πραγματοποιήθηκε αναζήτηση, κλικ ή αγορά εντός μιας ενιαίας συνεδρίας. Με απλούς όρους, παρακολουθεί τη συμπεριφορά του χρήστη καταγράφοντας σε ποιο σημείο της οθόνης του υπολογιστή κάνει κλικ.

Υπάρχουν δύο επίπεδα ανάλυσης ροών κλικ – clickstreams (Leach et al., 2005):

- **Ανάλυση κυκλοφορίας (Traffic Analytics):** Λειτουργεί στο επίπεδο διακομιστή (server).

Συλλέγει και αναλύει τα ακόλουθα δεδομένα:

  - Πόση ώρα χρειάζεται κάθε σελίδα για να φορτώσει.
  - Πόσες σελίδες προβάλλονται σε έναν χρήστη.
  - Πόσο συχνά ο χρήστης πατάει το κουμπί επιστροφής του προγράμματος περιήγησης.
  - Πόσα δεδομένα μεταδίδονται προτού ο χρήστης μεταβεί σε διαφορετική σελίδα.

- **Ανάλυση στοιχείων ηλεκτρονικού εμπορίου (E-commerce analytics):** Αυτή η ανάλυση χρησιμοποιεί δεδομένα clickstream για να προσδιορίσει την αποτελεσματικότητα ενός ιστοτόπου όσον αφορά τις μετατροπές τις συναλλαγές.

Ασχολείται με τα ακόλουθα δεδομένα:

  - Σε ποιες σελίδες παραμένει ο αγοραστής.
  - Τι τοποθετεί ή αφαιρεί ο αγοραστής από το καλάθι αγορών.
  - Ποια αντικείμενα αγοράζει ο αγοραστής.
  - Αν ο αγοραστής χρησιμοποιεί κωδικό κουπονιού.
  - Τον προτιμώμενο τρόπο πληρωμής του αγοραστή.

Υπάρχουν πολλά οφέλη που μπορούν να αποκομίσουν οι οργανισμοί από τα δεδομένα ροής κλικ και την ανάλυση ροής κλικ.

Μερικά από αυτά είναι τα εξής:

- **Πληροφορίες σχετικά με τους χρήστες:** Τα δεδομένα τα οποία συλλέγονται μπορούν να περιλαμβάνουν τους όρους αναζήτησης που χρησιμοποιούνται, τις σελίδες στις οποίες εισέρχονται, τις λειτουργίες της ιστοσελίδας που χρησιμοποιούνται και την προσθήκη ή την αφαίρεση στοιχείων από ένα καλάθι, τα οποία μπορούν να οδηγήσουν σε πιο αξιότιμες πληροφορίες.
- **Διαδρομές χρηστών:** Οι επιχειρήσεις μπορούν να χρησιμοποιήσουν τα δεδομένα που προέρχονται από τα clickstream για να δουν τις διαφορετικές διαδρομές που ακολουθούν οι διαδικτυακοί επισκέπτες ή οι πελάτες τους για να φτάσουν σε μια σελίδα ή για να πραγματοποιήσουν μια αγορά.
- **Τάσεις σχετικά με τους πελάτες:** Η συλλογή και ανάλυση των clickstreams ενός μεγάλου αριθμού επισκεπτών επιτρέπει σε έναν οργανισμό να εντοπίσει τάσεις σε διάφορους τομείς όπως πόση ώρα παραμένουν σε μια σελίδα, τι κάνουν μόλις βρεθούν εκεί, τον αριθμό των μοναδικών και επαναλαμβανόμενων επισκεπτών κλπ..
- **Ψηφιακό marketing:** Τα δεδομένα clickstream μπορούν να χρησιμοποιηθούν για τον προσδιορισμό της ποσότητας της επισκεψιμότητας που προέρχεται από διαφημιστικά banners και καμπάνιες. Τα δεδομένα αυτά παρέχουν πληροφορίες σχετικά με το ποιες διαφημίσεις είναι πιο αποτελεσματικές και οδηγούν στη βελτιστοποίηση του ρυθμού επισκεψιμότητας των χρηστών. Η ανάλυση των clickstream μπορεί επίσης να εξάγει ποιες ώρες της ημέρας, του μήνα ή του έτους μια στρατηγική marketing είναι πιο αποτελεσματική.
- **Διορθώσεις σε θέματα του UX:** Εάν κάποιο ποσοστό των χρηστών εγκαταλείπει γρήγορα μια σελίδα ή ένα ιστότοπο, αυτό μπορεί να είναι μια ένδειξη ότι η συγκεκριμένη σελίδα είναι ανεπαρκώς βελτιστοποιημένη ή δεν περιέχει αρκετές πληροφορίες αξίας. Μέσω της ανάλυσης των δεδομένων clickstream δίνεται η δυνατότητα σε μια επιχείρηση να αναγνωρίσει.

4 Προφίλ χρηστών

Το user profiling, γνωστό και ως προφίλ χρήστη, αναφέρεται στη διαδικασία συλλογής, αποθήκευσης, ανάλυσης και καταγραφής πληροφοριών σχετικά με τη συμπεριφορά, τις προτιμήσεις και τα χαρακτηριστικά ενός χρήστη ή καταναλωτή. Αποτελεί μια διαδικασία που χρησιμοποιείται κυρίως στον τομέα της τεχνολογίας, ιδίως στον κόσμο του διαδικτύου και των ψηφιακών εφαρμογών, για να κατανοήσει και να προσαρμόσει την εμπειρία του χρήστη.

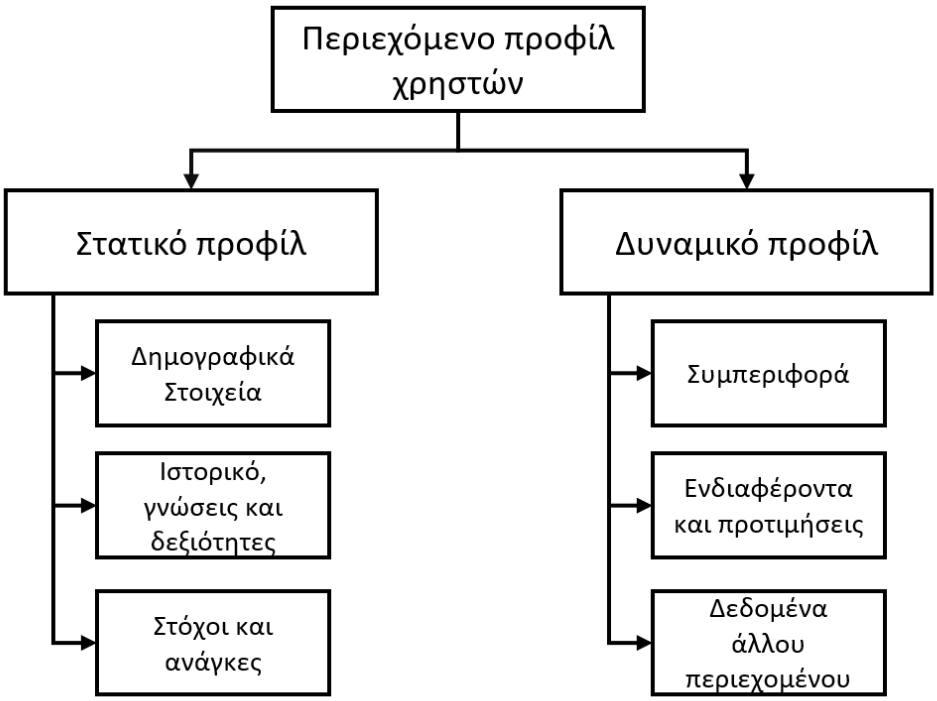
Αφού οι πληροφορίες συλλεγούν, αναλύονται και αξιοποιούνται από προηγμένους αλγορίθμους μηχανικής μάθησης και τεχνητής νοημοσύνης, παρέχονται εξατομικευμένες προτάσεις και πληροφορίες στον χρήστη. Στον κόσμο του τουριστικού ενδιαφέροντος, το user profiling χρησιμοποιείται για να προτείνει προορισμούς, ξενοδοχεία, εστιατόρια, δραστηριότητες και άλλες εμπειρίες που ανταποκρίνονται στις προτιμήσεις και τις ανάγκες του κάθε ταξιδιώτη. Εντέλει, το user profiling βοηθά στην κατανόηση του χρήστη και στην παροχή εξατομικευμένων εμπειριών, δημιουργώντας θετικές και ικανοποιητικές αλληλεπιδράσεις με τις ψηφιακές πλατφόρμες και υπηρεσίες.

4.1 Τύποι προφίλ χρηστών

Τα προφίλ χρηστών συνήθως αναπαρίστανται με τη μορφή πλούσιας σημασιολογικής δομής και με ένα σύνολο σταθμισμένων λέξεων – κλειδιών. Παρόλα αυτά, η αναπαράσταση με σταθμισμένες λέξεις – κλειδιά χρησιμοποιείται συνήθως επειδή η εξαγωγή του προφίλ με αυτή την μεθοδολογία από ένα έγγραφο ή από άλλες πηγές πραγματοποιείται αυτόματα.

Το προφίλ χρήστη μπορεί να ομαδοποιηθεί σε δύο βασικές κατηγορίες:

- Στατικό προφίλ
- Δυναμικό προφίλ



Εικόνα 4.1 Διαχωρισμός των προφίλ χρηστών σε στατικά και δυναμικά προφίλ

4.1.1 Στατικό προφίλ

Η εργασία σκιαγράφησης προφίλ χρήστη θεωρείται ως μία προσέγγιση μάθησης με επίβλεψη. Η προσέγγιση στατικού προφίλ αποτελεί μία διαδικασία ανάλυσης των προβλέψιμων και στατικών χαρακτηριστικών του χρήστη. Οι πληροφορίες που παρέχει ο χρήστης μέσω του στατικού προφίλ χρησιμοποιούνται στον προσδιορισμό του είδους των πληροφοριών για τις οποίες ο χρήστης εμφανίζει ενδιαφέρον. Ένα στατικό προφίλ είναι ένας τύπος προφίλ που διατηρεί πληροφορίες του χρήστη για μεγάλο χρονικό διάστημα δηλαδή δεν υφίστανται καμία αλλαγή ή τροποποίηση όπως για παράδειγμα, η ηλικία και το φύλο του χρήστη. Ως εκ τούτου, τα χαρακτηριστικά του χρήστη και τα διαθέσιμα περιεχόμενα είναι στατικής φύσης, δηλαδή παραμένουν αμετάβλητα εντός μίας περιόδου. Το πρόβλημα με αυτή τη μορφή σκιαγράφησης προφίλ είναι ότι οι χρήστες σπάνια παρέχουν όλες τις πληροφορίες τους με ακρίβεια, καθώς θεωρούν ότι οι προσωπικές πληροφορίες που ζητούνται είναι ιδιαίτερα σημαντικές για να τις «αποκαλύψουν» με αποτέλεσμα το στατικό προφίλ να θεωρείται αναξιόπιστο.

4.1.2 Δυναμικό προφίλ

Σε αντίθεση με το στατικό προφίλ στο δυναμικό προφίλ η σκιαγράφηση του προφίλ γίνεται αυτοματοποιημένα όποτε κατά συνέπεια ενημερώνονται αυτόματα καθώς οι

ιδιότητες και τα περιεχόμενα του χρήστη μεταβάλλονται με την πάροδο του χρόνου. Στο δυναμικό προφίλ, οι πληροφορίες του χρήστη σχετικά με τη συμπεριφορά του χρήστη επιδιώκουν τον προσδιορισμό μελλοντικών πληροφοριών του χρήστη περισσότερο από ότι οι παρούσες πληροφορίες. Χαρακτηρίζεται και ως συμπεριφορικό ή προσαρμοστικό προφίλ. Το δυναμικό προφίλ είναι πάντα ακριβές σε καταστάσεις όπου υπάρχει μεγάλη ροή δεδομένων. Επιπλέον, χρησιμοποιείται η υπάρχουσα οντολογία χρηστών για να κατευθύνει την εξαγωγή του προφίλ, να καθορίσει το σύνολο των σχέσεων και για την παροχή του λεξικού οντοτήτων. Το δυναμικό προφίλ που λαμβάνει υπόψη τον χρόνο μπορεί να διακρίνει μεταξύ μακροπρόθεσμων και βραχυπρόθεσμων μακροπρόθεσμα ενδιαφέροντα. Ενώ το βραχυπρόθεσμο προφίλ απεικονίζει το τρέχον ενδιαφέρον του χρήστη, το μακροπρόθεσμο αντιπροσωπεύει το ενδιαφέρον που δεν αλλάζει πάντα.

4.2 Τύποι δεδομένων σε προφίλ χρηστών

Υπάρχουν διάφοροι τύποι δεδομένων που μπορούν να συνυπάρχουν σε ένα προφίλ χρήστη όπως προσωπικά δεδομένα του χρήστη, γνώσεις και δεξιότητες του χρήστη, ανάγκες και στόχοι του χρήστη τα οποία είναι στατικά δεδομένα. Από την άλλη, υπάρχουν και άλλα δεδομένα τα οποία είναι δυναμικά όπως στοιχεία συμπεριφοράς του χρήστη, στοιχεία ενδιαφέροντος και προτιμήσεων του χρήστη. Οι παραπάνω κατηγορίες διαδέχονται η μία την άλλη με βάση τη δυναμική των δεδομένων από τα δεδομένα χαμηλού επιπέδου έως τα δυναμικά δεδομένα υψηλού επιπέδου.

Παρακάτω περιγράφονται ορισμένοι τύποι δεδομένων που χρησιμοποιούνται στα προφίλ χρηστών:

- **Δημογραφικά στοιχεία:** Τα δημογραφικά στοιχεία ενός χρήστη είναι τα πολύ βασικά χαρακτηριστικά, όπως το όνομα, η χώρα, το φύλο, η ηλικία, η μητρική γλώσσα, εκπαίδευση, μέλη της οικογένειας κ.λπ.
- **Ιστορικό, γνώσεις και δεξιότητες:** Αποτελούν ιδιαίτερα σημαντικό χαρακτηριστικό για την μοντελοποίηση μαθητών σε συστήματα διδασκαλία και προσαρμοστικά εκπαιδευτικά συστήματα. Οι γνώσεις τις οποίες διαθέτει ένας μαθητής είναι το κύριο χαρακτηριστικό που καθορίζει ένα προσαρμοστικό σύστημα και οι γνώσεις του σχετικά με το διδασκόμενο αντικείμενο είναι σημαντικές για την παροχή κατάλληλης βοήθειας τόσο στον ίδιο τον μαθητή όσο και στην προσαρμογή του περιεχομένου των μαθημάτων. Σε ορισμένα συστήματα συστάσεων οι

δεξιότητες που διαθέτει ένας χρήστης ή υπάλληλος καθώς και ο ρόλος και η απόδοση του μέσα στην επιχείρηση είναι σημαντικά για να παρέχουν τις κατάλληλες θέσεις εργασίας για κάποιον υποψήφιο ώστε να προτείνουν τα κατάλληλα άτομα σε έναν υπεύθυνο προσλήψεων.

- **Στόχοι και ανάγκες:** Οι στόχοι ενός χρήστη αποτελούν σημαντικό στοιχείο για την ανίχνευση του σκοπού από την εφαρμογή που χρησιμοποιεί. Στόχος είναι η απόκτηση σχετικών πληροφοριών από την περιήγηση του χρήστη στον Παγκόσμιο Ιστό.
- **Ενδιαφέροντα και προτιμήσεις:** Τα ενδιαφέροντα και οι προτιμήσεις ενός χρήστη αποτελούν κύριο μέρος της εξατομίκευσης. Τα ενδιαφέροντα μπορεί να αφοράνε θέματα εργασίας, θέματα κοινωνικών δικτύων, θέματα ιστοτόπων, θέματα εγγράφων. Αυτά τα θέματα μπορούν να προέρχονται από διάφορες πηγές, όπως το ιστορικό αγορών ή ιστορικό περιήγησης. Μερικές φορές τα ενδιαφέροντα των χρηστών ταξινομούνται σε βραχυπρόθεσμα ή μακροπρόθεσμα ενδιαφέροντα. Τα μακροπρόθεσμα προφίλ κατασκευάζονται από τις Google Directory χρησιμοποιώντας τα θέματα των αποτελεσμάτων των ροών με κλικ, και τα βραχυπρόθεσμα μακροπρόθεσμα μοντέλα χρησιμοποιώντας μια προσωρινή μνήμη cache των αποτελεσμάτων που έγιναν πρόσφατα κλικ.
- **Συμπεριφορά:** Η συμπεριφορά αποτελεί ένα είδος πληροφοριών που συλλέγονται σιωπηρά. Είναι σημαντικό η συμπεριφορά να είναι επαναλαμβανόμενη προκειμένου να εντοπιστούν μοτίβα που χρησιμοποιούνται από προσαρμοστικά συστήματα ή ευφυείς πράκτορες για την υποβοήθηση της συμπεριφοράς του χρήστη. Σε ορισμένες έρευνες παρατηρούνται τρεις τύποι συμπεριφοράς του χρήστη:
    - Χρόνος προσοχής
    - Επιλογές του χρήστη μέσω κλικ ποντικιού
    - Κίνηση του ποντικιού
- **Δεδομένα άλλου περιεχομένου:** Τα δεδομένα αυτά προσδιορίζονται ως οποιαδήποτε πληροφορία, η οποία χρησιμοποιείται για να χαρακτηρίσει μια κατάσταση οποιουδήποτε στοιχείου. Ως πληροφορίες οι οποίες μπορούν να προσδιορίσουν το προφίλ ενός χρήστη θεωρούνται οποιεσδήποτε προτιμήσεις του, αντιπάθειες του ακόμη και οι συναισθηματικές του καταστάσεις. Τα συναισθήματα ενός χρήστη έχουν άμεσο αντίκτυπο στον τρόπο με τον οποίο χρησιμοποιεί την εφαρμογή λογισμικού. Για παράδειγμα η ταχύτητα πληκτρολόγησης και τα κλικ του

ποντικιού είναι οι κυριότεροι παράμετροι για την ανίχνευση και την αναγνώριση των βασικών ανθρώπινων συναισθημάτων.

4.3 Αναπαράσταση προφίλ χρηστών

Τα προφίλ των χρηστών συνήθως αναπαρίστανται γενικά ως σύνολα σταθμισμένων λέξεων – κλειδιών, σημασιολογικών δικτύων ή σταθμισμένων εννοιών. Τα προφίλ λέξεων – κλειδιών είναι τα απλούστερα στην κατασκευή τους, αλλά επειδή απαιτείται η συλλογή και η αναπαράσταση όλων ή των περισσότερων λέξεων με τις οποίες συνδέονται τα ενδιαφέροντα του χρήστη που μπορούν να συναντηθούν σε μελλοντικά έγγραφα, απαιτούν μεγάλο όγκο ανατροφοδότησης του χρήστη προκειμένου να μάθουν την ορολογία με την οποία μπορεί να συζητηθεί ένα θέμα. Αυτό το πρόβλημα μοιράζονται επίσης τα περισσότερα προφίλ που βασίζονται σε σημασιολογικά δίκτυα - πρέπει να μαθαίνουν την ορολογία με την οποία συναντιούνται οι έννοιες. Τα προφίλ εννοιών, αντίθετα, εκπαιδεύονται σε παραδείγματα για κάθε έννοια εκ των προτέρων, και έτσι ξεκινούν με μια υπάρχουσα χαρτογράφηση μεταξύ λεξιλογίου και εννοιών. Έτσι, μπορούν να δημιουργήσουν προφίλ που είναι ανθεκτικά στις παραλλαγές της ορολογίας με λιγότερη ανατροφοδότηση από τους χρήστες .

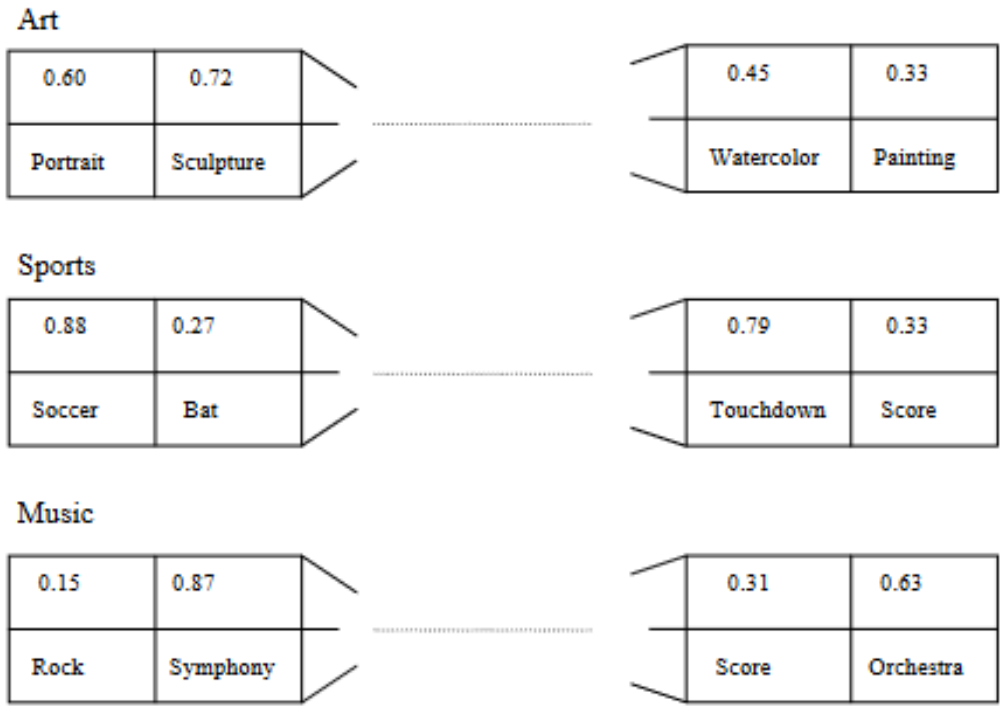
4.3.1 Βασισμένο σε όρους

Το μοντέλο χρήστη που βασίζεται σε διανύσματα ή σε όρους (λέξεις – κλειδιά) αποτελεί την πιο κοινή μέθοδο για την αναπαράσταση των ενδιαφερόντων ενός χρήστη. Τα δεδομένα αυτά μπορούν να εξαχθούν αυτόματα από έγγραφα του Ιστού κατά την περιήγησή του χρήστη είτε να δοθούν απευθείας από τον ίδιο τον χρήστη. Συνήθως αποτελούνται από μία συλλογή δεδομένων τα περιέχουν ένα σύνολο όρων (ή διανυσματικός χώρος όρων) σε συνδυασμό με σταθμισμένα (weighted) διανύσματα λέξεων – κλειδιών τα οποία μπορούν να χρησιμοποιηθούν για την επέκταση των ερωτημάτων των χρηστών. Τα βάρη, τα οποία συνήθως συνδέονται με τις λέξεις – κλειδιά, είναι αριθμητικές αναπαραστάσεις των ενδιαφερόντων του χρήστη. Κάθε λέξη – κλειδί μπορεί να αντιπροσωπεύει ένα θέμα ενδιαφέροντος ή οι λέξεις – κλειδιά μπορεί να ομαδοποιηθούν σε κατηγορίες για να αντικατοπτρίζουν μια πιο τυποποιημένη αναπαράσταση των ενδιαφερόντων του χρήστη.

Οι λέξεις – κλειδιά που αναφέρονται στα χαρακτηριστικά του χρήστη εξάγονται από τα έγγραφα που επισκέπτονται κατά την περιήγησή τους από τους σελιδοδείκτες που

αποθηκεύονται από τον χρήστη, ή οι λέξεις – κλειδιά που παρέχονται ρητά από τον χρήστη. Κάθε λέξη – κλειδί συνδέεται συνήθως με ένα αριθμητικό βάρος που αντιπροσωπεύει τη σημασία του στο προφίλ. Τόσο τα προφίλ των χρηστών όσο και τα έγγραφα που ανακτώνται από το σύστημα ως αποτέλεσμα μια αναζήτησης αναπαρίσταται με τη μορφή ενός σταθμισμένου διανύσματος λέξεων – κλειδιών. Αυτά τα διανύσματα συγκρίνονται στη συνέχεια με το προφίλ χρησιμοποιώντας τον τύπο του συνημίτονού και μόνο τα αντίστοιχα έγγραφα για εκείνα τα διανύσματα που είναι πιο κοντά στο προφίλ διαβιβάζονται στον χρήστη. Ένα από τα κυριότερα μειονεκτήματα των προφίλ με βάση τις λέξεις – κλειδιά είναι ότι πολλές λέξεις έχουν παραπάνω από μία εννοιολογική σημασία. Εξαιτίας αυτού του γεγονότος πολλές λέξεις – κλειδιά στο προφίλ χρήστη είναι διφορούμενες, καθιστώντας το προφίλ ανακριβές.

Παράδειγμα εφαρμογής αποτελεί το PEA ένας εξατομικευμένος βοηθός αναζήτησης στον Παγκόσμιο Ιστό που δημιουργεί προφίλ χρηστών με βάση λέξεις – κλειδιά χρησιμοποιώντας όρους που εξάγονται από τους σελιδοδείκτες (αποθηκευμένες σελίδες) του χρήστη. Η εφαρμογή αυτή διαφέρει από άλλες προσεγγίσεις καθώς αντί να δημιουργεί ένα ενιαίο προφίλ για τον χρήστη, ο χρήστης αναπαρίσταται ως ένα σύνολο διανυσμάτων λέξεων – κλειδιών και σταθμισμένων διανυσμάτων ανά σελιδοδείκτη.

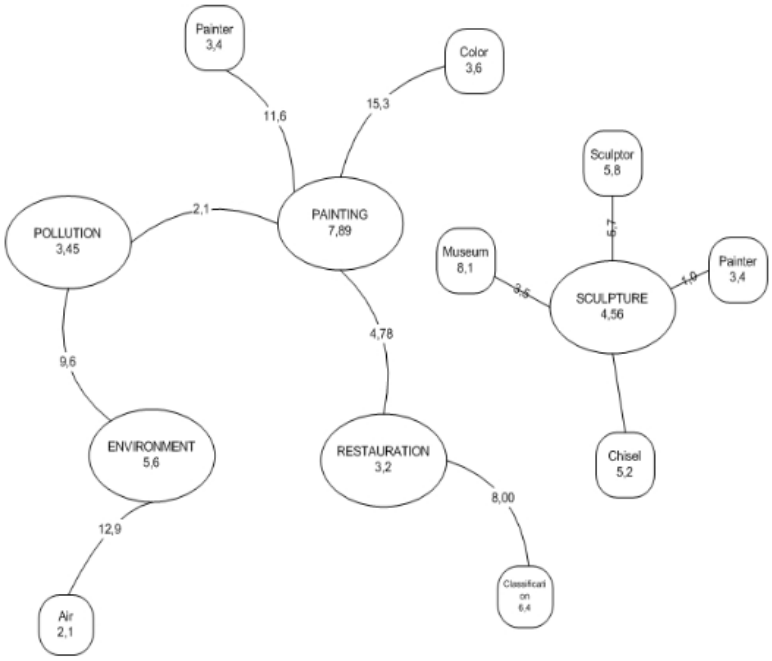


Εικόνα 4.2 Παράδειγμα χρήσης του PEA.

Η λογική πίσω από αυτή την μεθοδολογία είναι ότι αν ένας χρήστης ενδιαφέρεται για δύο θέματα, ο συνδυασμός των λέξεων – κλειδιών και από τα δύο θέματα σε ένα ενιαίο διάνυσμα δεν παρέχει τόσο ακριβή δεδομένα. Αντίθετα, η αναπαράσταση κάθε περιοχής ενδιαφέροντος δηλαδή κάθε σελιδοδείκτη ως ξεχωριστό διάνυσμα είναι πιθανό να παρέχει ένα πιο ακριβές προφίλ. Καθώς ο χρήστης περιηγείται, επιπλέον σελίδες συνιστώνται στο χρήστη όταν το διάνυσμα της καινούργιας σελίδας είναι παρόμοιο με το διάνυσμα ενός υπάρχοντος σελιδοδείκτη (Gauch et al., 2007).

4.3.2 Βασισμένο σε σημασιολογικά δίκτυα

Με σκοπό την αντιμετώπιση της πολυσημίας η οποία αποτελεί ένα εγγενές πρόβλημα στην μεθοδολογία αναπαράστασης των προφίλ χρηστών με βάση λέξεις – κλειδιά δημιουργήθηκε μία νέα μεθοδολογία κατά την οποία τα προφίλ αναπαρίστανται με ένα σταθμισμένο σημασιολογικό δίκτυο στο οποίο κάθε κόμβος αντιπροσωπεύει μια έννοια. Κάθε κόμβος περιέχει μια συγκεκριμένη λέξη που βρίσκεται στο έγγραφο και οι οποίες συνδέονται μεταξύ τους ανάλογα με την συνύπαρξη των δύο λέξεων. Ωστόσο, η απεικόνιση μεμονωμένων λέξεων ως κόμβους στο σημασιολογικό δίκτυο δεν ήταν ακριβής για να διακρίνει τις έννοιες των λέξεων οπότε εφαρμόστηκε μια ομαδοποίηση των σχετικών λέξεων με βάση την έννοια τους που ονομάζονται σύνολα συνωνύμων ή synsets. Οπότε πλέον η αναπαράσταση ενός προφίλ χρήστη με χρήση σημασιολογικών δικτύων στο οποίο οι κόμβοι είναι σύνολα, τα τόξα – συνδέσεις είναι οι συν – εμφανίσεις των μελών των συνόλων σε ένα έγγραφο που ενδιαφέρει τον χρήστη και τα βάρη των κόμβων και των τόξων αντιπροσωπεύουν το επίπεδο ενδιαφέροντος του χρήστη.



Εικόνα 4.3 Παράδειγμα εφαρμογής του InfoWeb.

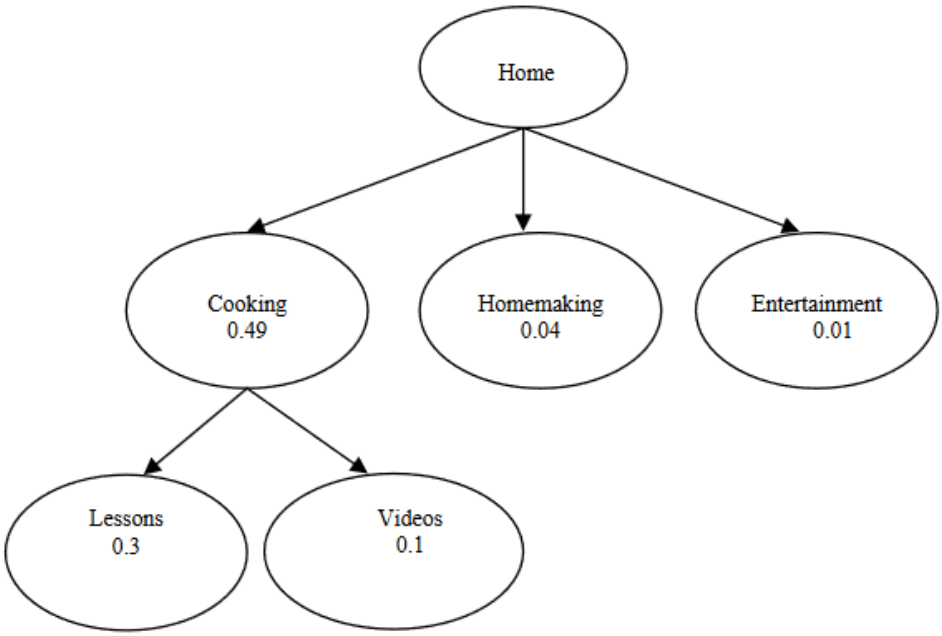
Παράδειγμα εφαρμογής είναι το InfoWeb, ένα σύστημα φιλτραρίσματος για ηλεκτρονικά έγγραφα ψηφιακών βιβλιοθηκών που κατασκευάζουν προφίλ βασισμένα σε σημασιολογικά δίκτυα που αντιπροσωπεύουν τα μακροπρόθεσμα ενδιαφέροντα των χρηστών. Κάθε χρήστης προφίλ αναπαρίσταται ως σημασιολογικό δίκτυο εννοιών. Αρχικά, κάθε σημασιολογικό δίκτυο περιέχει μια συλλογή από μη συνδεδεμένους κόμβους στους οποίους κάθε κόμβος αντιπροσωπεύει μια έννοια. Οι κόμβοι εννοιών που ονομάζονται «πλανήτες» περιέχουν ένα και μόνο αντιπροσωπευτικό σταθμισμένο όρο για την συγκεκριμένη έννοια. Καθώς συγκεντρώνονται περισσότερες πληροφορίες για τον χρήστη, το προφίλ εμπλουτίζεται ώστε να περιλαμβάνει πρόσθετες σταθμισμένες λέξεις – κλειδιά που σχετίζονται με τις έννοιες. Αυτές οι λέξεις – κλειδιά αποθηκεύονται σε δευτερεύοντες κόμβους, που ονομάζονται δορυφόροι, οι οποίοι συνδέονται με τους αντίστοιχους κόμβους – έννοιες «πλανήτες». Επιπλέον προστίθενται σύνδεσμοι μεταξύ των πλανητών που αντιπροσωπεύουν συσχετίσεις μεταξύ των εννοιών (Gauch et al., 2007).

4.3.3 Με βάση τις έννοιες (Concept – based profiles)

Τα προφίλ χρηστών που βασίζονται σε έννοιες είναι παρόμοια με τα προφίλ που βασίζονται σε σημασιολογικά δίκτυα υπό την έννοια ότι και τα δύο αντιπροσωπεύονται από εννοιολογικούς κόμβους και σχέσεις μεταξύ αυτών των κόμβων. Παρόλα αυτά, στα εννοιολογικά προφίλ, οι κόμβοι αντιπροσωπεύουν αφηρημένα θέματα που θεωρούνται ενδιαφέροντα για τον χρήστη, και όχι συγκεκριμένες λέξεις ή σύνολα σχετικών λέξεων. Αυτή η μεθοδολογία μπορεί να θεωρηθεί παρόμοια με τα προφίλ λέξεων – κλειδιών στο γεγονός ότι συχνά αναπαρίστανται ως διανύσματα σταθμισμένων χαρακτηριστικών, αλλά τα χαρακτηριστικά αντιπροσωπεύουν έννοιες και όχι λέξεις ή σύνολα λέξεων. Εφαρμόζονται διάφοροι μηχανισμοί για να εκφράσουν πόσο ο χρήστης ενδιαφέρεται για κάθε θέμα. Η απλούστερη τεχνική είναι με την χρήση μιας αριθμητικής τιμής που ονομάζεται βάρος και η οποία συνδέεται με κάθε θεματική ενότητα. Αργότερα προτάθηκε η χρήση ιεραρχικών εννοιών αντί για ένα επίπεδο σύνολο εννοιών καθώς αυτό θα επέτρεπε στο σύστημα να κάνει γενικεύσεις. Κατά την ιεράρχηση των εννοιών τα επίπεδα μπορεί να είναι σταθερά, ή να αλλάζουν δυναμικά ανάλογα με τα ενδιαφέροντα του χρήστη. Οι ιεραρχίες εννοιών χρησιμοποιήθηκαν αρχικά για την αναπαράσταση του περιεχομένου των ιστοσελίδων, αλλά πιο πρόσφατα χρησιμοποιήθηκαν για την αναπαράσταση προφίλ χρηστών. Όσο περισσότερα επίπεδα χρησιμοποιούνται τόσο πιο συγκεκριμένη μπορεί να γίνει η αναπαράσταση του προφίλ χρήστη.

Καθώς η διαδικασία δημιουργίας μιας ευρείας και βαθιάς ιεραρχίας εννοιών αποτελεί μια δαπανηρή, και κυρίως χειροκίνητη διαδικασία, τα προφίλ συνήθως βασίζονται σε υποσύνολα υφιστάμενων ιεραρχιών εννοιών. Όταν γίνεται χρήση ενός υπάρχοντος καταλόγου ως πηγή εννοιών, ορισμένοι μετασχηματισμοί πρέπει να πραγματοποιηθούν για να μετατραπούν τα περιεχόμενα του καταλόγου σε ιεραρχία εννοιών. Επειδή ο κατάλογος έχει σχεδιαστεί για να επιτρέπει την περιήγηση του τελικού χρήστη, δεν αποτελούν όλοι οι σύνδεσμοι γονέα – παιδιού (ιεραρχία). Ορισμένα θέματα χωρίζονται αλφαβητικά άλλα χωρίζονται γεωγραφικά ενώ άλλα μπορεί να έχουν ελάχιστα έως και εκατοντάδες παιδιά. Το έργο σκιαγράφησης – αναπαράστασης του προφίλ θα πρέπει να λάβει υπόψη όλα αυτά τα ζητήματα και να αποφασίσει ποια από τα θέματα του καταλόγου να συμπεριληφθούν στην ιεραρχία εννοιών.

Ένα από τα πρώτα έργα για τη δημιουργία προφίλ χρηστών βασισμένα σε έννοιες ήταν το OBIWAN. Αρχικά το σύστημα χρησιμοποιούσε μια ιεραρχία εννοιών αναφοράς που περιείχε 4.417 θέματα από τα τέσσερα ανώτερα επίπεδα του ιστότοπου Magellan. Αφού ο ιστότοπος Magellan διέκοψε την λειτουργία του η ομάδα πειραματίστηκε με ιεραρχίες θεμάτων από την ιστοσελίδα Yahoo και Lycos, επιλέγοντας τελικά το Open Directory Project (ODP) ως αντικαταστάτη κυρίως επειδή ο κατάλογος τους είναι ανοικτού κώδικα. Αρχικά αναπαριστούσαν προφίλ με την χρήση 1.869 εννοιών από τα τρία ανώτερα επίπεδα της ιεραρχίας εννοιών του ODP (Gauch et al., 2007).



Εικόνα 4.4 Ιεράρχηση εννοιών

4.4 Διαδικασία δημιουργίας προφίλ χρήστη

Παρακάτω περιγράφεται η διαδικασία δημιουργίας προφίλ χρήστη. Παρέχει το σύνολο των διαδικασιών που εμπλέκονται στο προφίλ χρήστη, όπως η κατασκευή προφίλ, η συλλογή πληροφοριών χρήστη και η ενημέρωση του προφίλ χρήστη.

4.4.1 Συλλογή δεδομένων χρήστη

Η συλλογή πληροφοριών για έναν συγκεκριμένο χρήστη αποτελεί την έναρξη της δημιουργίας του προφίλ χρήστη. Το σύστημα συστάσεων θα πρέπει να περιέχει πληροφορίες για το χρήστη έτσι ώστε να μπορεί να «καταλάβει» ποιες είναι οι ανάγκες

του χρήστη και να κάνει τις ανάλογες προτάσεις. Επομένως, για τη δημιουργία εξατομικευμένων συστάσεων, η απόκτηση δεδομένων χρήστη, τα οποία αντιπροσωπεύουν τις ανάγκες και τις προτιμήσεις του, είναι μια κρίσιμη διαδικασία. Η αλληλεπίδραση του χρήστη με διάφορα στοιχεία είναι υπεύθυνη για την αναγνώριση των προτιμήσεων του χρήστη από το σύστημα. (Eke et al., 2019)

Οι μέθοδοι με τις οποίες συλλέγονται τα δεδομένα είναι τρεις (Farid et al., 2018):

- Η ρητή μέθοδος (χειροκίνητα)
- Η έμμεση (ημιαυτόματα)
- Η υβριδική (αυτόματα)

4.4.1.1 Ρητή συλλογή πληροφοριών χρήστη

Η απόκτηση πληροφοριών χρήστη με τη ρητή μέθοδο περιλαμβάνει την καταγραφή των προσωπικών πληροφοριών με την άμεση παρέμβαση του χρήστη. Αυτό σημαίνει ότι οι χρήστες είναι αυτοί οι οποίοι παρέχουν πληροφορίες που τους αφορούν, μέσα από μια διαδικασία έρευνας και εγγραφής ή με τη συμπλήρωση φόρμας και ερωτηματολογίων, τα οποία μπορεί να περιλαμβάνουν τη γνώμη του χρήστη για κάποιο συγκεκριμένο θέμα. Πολλοί ιστότοποι συλλέγουν τις προτιμήσεις των χρηστών παρέχοντας εξατομικευμένες υπηρεσίες στους χρήστες και στη συνέχεια ζητώντας τους απευθείας να δώσουν προσωπικές πληροφορίες για να δημιουργήσουν ένα προφίλ. Οι πληροφορίες που λαμβάνονται μπορεί να αποτελούνται από δημογραφικά χαρακτηριστικά όπως το όνομα του χρήστη, η διεύθυνσή του, ο αριθμός τηλεφώνου του, η κατάσταση γάμου, η κατάσταση εργασίας, τα γενέθλιά του, τα προσωπικά του ενδιαφέροντα και τα χόμπι. Άλλα στοιχεία του χρήστη, που μπορούν να αποτελέσουν ρητές πληροφορίες, είναι πληροφορίες όπως η διαδικτυακή του συναλλαγή ή η δραστηριότητά του στον ιστό. Για παράδειγμα, τέτοια στοιχεία μπορεί να είναι οι κατηγορίες του προϊόντος που αγοράστηκε περισσότερο και το πόσο συχνά ο χρήστης επισκέπτεται τον ιστό (Eke et al., 2019).

Παρόλο που η απόκτηση δεδομένων του χρήστη με τη ρητή μέθοδο είναι σχετικά εύκολη διαδικασία και έχει λιγότερο θόρυβο, συνοδεύεται από κάποια μειονεκτήματα. Πρώτον, υπάρχει άμεση παρέμβαση του χρήστη με αποτέλεσμα ο χρήστης να επιβαρύνεται από τη διαδικασία, καθώς επενδύει χρόνο και προσπάθεια συμπληρώνοντας τα στοιχεία του και εκφράζοντας τις προτιμήσεις και τα ενδιαφέροντά του. Επίσης, σε περίπτωση που οι

χρήστες δεν είναι πρόθυμοι να μοιραστούν προσωπικές πληροφορίες, η δημιουργία του προφίλ χρήστη δεν μπορεί να πραγματοποιηθεί. Ένα άλλο πρόβλημα που προκύπτει αφορά το θέμα της ιδιωτικότητας. Οι χρήστες μπορεί όχι μόνο να μην δέχονται να παρέχουν τα προσωπικά τους στοιχεία, αλλά και να δίνουν ανακριβείς πληροφορίες στο σύστημα. Επιπλέον, οι προτιμήσεις του χρήστη μπορεί να αλλάζουν με την πάροδο του χρόνου, αλλά το προφίλ παραμένει στατικό, με αποτέλεσμα να δημιουργείται ένα όλο και πιο ανακριβές προφίλ. Τα μειονεκτήματα αυτά προκαλούν προβλήματα στην απόδοση των συστημάτων συστάσεων και δυσκολεύουν τη δημιουργία ενός ακριβούς προφίλ των χρηστών (Nadee, 2016; Eke et al. 2019).

4.4.1.2 Έμμεση συλλογή πληροφοριών χρήστη

Οι πληροφορίες που συλλέγονται με τη ρητή μέθοδο δεν είναι πάντα διαθέσιμες και δεν περιέχουν πάντα αρκετές λεπτομέρειες για τη δημιουργία ενός επαρκούς προφίλ χρήστη. Για το λόγο αυτό, μπορεί να επιλεγεί η έμμεση συλλογή πληροφοριών χρήστη. Οι περισσότερες πληροφορίες χρήστη που συλλέγονται έμμεσα πραγματοποιούνται με έξυπνους πράκτορες λογισμικού ή τεχνικές εξόρυξης δεδομένων και βασίζονται στη συμπεριφορά και τη δραστηριότητα των χρηστών. Ένας τρόπος συλλογής δεδομένων με έμμεσο τρόπο περιλαμβάνει αρχεία καταγραφής χρήσης ιστού, ροών κλικ, ιστορικών περιήγησης, εγγραφών αγορών και περιεχομένου ή δομικών πληροφοριών από ιστοσελίδες που επισκέφθηκαν. Τα ιστορικά περιήγησης αποτελούν μια κοινή πηγή πληροφοριών χρήστη που αποκτώνται έμμεσα. Με τη συλλογή πληροφοριών έμμεσα, το κυριότερο πλεονέκτημα είναι ότι δεν υπάρχει παρέμβαση του χρήστη και δεν απαιτείται καμία παραπάνω προσπάθειά του κατά τη δημιουργία του προφίλ. Επιπλέον, η πρόσβαση στα δεδομένα είναι εύκολη και συνεχής και τα δεδομένα ενημερώνονται αυτόματα κάθε φορά που οι χρήστες αλληλοεπιδρούν με το σύστημα.

Παρόλα αυτά, η μετατροπή της συμπεριφοράς του χρήστη σε προτιμήσεις και πληροφορίες είναι ιδιαίτερα δύσκολη. Το πόσο ακριβή είναι τα δεδομένα που αντλούνται εξαρτάται άμεσα από τη σωστή ερμηνεία της συμπεριφοράς του χρήστη. Για παράδειγμα, υπάρχει η περίπτωση ο χρήστης να προχωρήσει σε κάποια αγορά ενός προϊόντος, το οποίο όμως δεν προορίζεται για αυτόν. Με αυτόν τον τρόπο, θα συλλεχθούν λανθασμένα δεδομένα για το συγκεκριμένο χρήστη, καθώς το προϊόν που αγόρασε ανήκει σε άλλον και μπορεί να μην είναι της αρεσκείας του. Ένα άλλο μειονέκτημα αυτής της μεθόδου συλλογής πληροφοριών είναι ότι απαιτεί την ανάπτυξη

υψηλής ποιότητας εφαρμογών ή προσθηκών, οι οποίες θα πρέπει στη συνέχεια να εγκατασταθούν στον υπολογιστή του χρήστη (Nadee, 2016).

4.4.1.3 Υβριδική συλλογή πληροφοριών χρήστη

Τα υβριδικά προφίλ χρηστών δημιουργούνται με ημιαυτόματες τεχνικές με περιορισμένη συμμετοχή χρηστών. Το σύστημα συλλέγει πληροφορίες χρήστη και χρήσης με μια μικτή προσέγγιση έμμεσων, ως μη αυτόματων, και ρητών μεθόδων. Αυτή η προσέγγιση βοηθά στη δημιουργία ενός πιο αποτελεσματικού προφίλ χρήστη και, επιπλέον, διατηρεί την ακρίβεια των χρονικών πληροφοριών, καθώς οι πληροφορίες ενημερώνονται συχνά (Farid et al., 2018).

4.4.2 Δημιουργία προφίλ χρήστη

Κατά τη δημιουργία προφίλ χρήστη, χρησιμοποιούνται διάφοροι αλγόριθμοι μάθησης και συστήματα ανάκτησης πληροφοριών με βάση την επιλογή της αναπαράστασης. Η κατασκευή προφίλ μπορεί να κατηγοριοποιηθεί στο σημασιολογικό δίκτυο, τη λέξη-κλειδί και το προφίλ έννοιας. Υπάρχει η δυνατότητα δημιουργίας προφίλ χρήστη χειροκίνητα από τους χρήστες ή τους ειδικούς. Ωστόσο, αυτή η διαδικασία είναι παρεμβατική, χρονοβόρα και δύσκολη για πολλούς χρήστες, γεγονός που εμποδίζει την επέκταση της υιοθέτησης εξατομικευμένων υπηρεσιών. Για το λόγο αυτό, πιο δημοφιλής είναι η τεχνική που χρησιμοποιεί αυτόματα τα σχόλια των χρηστών για τη δημιουργία. Άλλες προσεγγίσεις, όπως νευρωνικά δίκτυα και γενετικοί αλγόριθμοι, που βασίζονται σε πιθανότητες ή μοντέλα διανυσματικού χώρου, χρησιμοποιούνται γενικότερα και έχει παρατηρηθεί ότι είναι πιο αποτελεσματικές σε αρκετούς τομείς.

Παρά το γεγονός ότι η δημιουργία προφίλ χρήστη βασίζεται στα ενδιαφέροντα και τις προτιμήσεις του χρήστη, διάφορες μελέτες που έχουν διεξαχθεί εξετάζουν τη δημιουργία προφίλ με βάση θέματα και τομείς που μπορεί να μην ενδιαφέρουν το χρήστη. Λαμβάνοντας υπόψη και αυτόν τον τρόπο, υπάρχουν πλέον δύο διαθέσιμες μέθοδοι, οι οποίες μπορούν να εφαρμοστούν στο σύστημα ώστε να εντοπίσουν τα κρίσιμα στοιχεία (που αφορούν τις προτιμήσεις του χρήστη) και, ταυτόχρονα, να εξαλείψουν τα μη σχετικά (Eke et al., 2019).

4.4.3 Ενημέρωση προφίλ χρήστη

Η ενημέρωση των δεδομένων, τα οποία συλλέχθηκαν για τη δημιουργία του προφίλ χρήστη, συνήθως πραγματοποιείται μετά την επιτυχή κατασκευή του. Η ενημέρωση σημαίνει την υποβολή ενός συγκεκριμένου ερωτήματος στο σύστημα. Αντίστοιχα, το σύστημα ανακτά το συγκεκριμένο στοιχείο, το οποίο αναζητήθηκε, και τις λέξεις-κλειδιά του ερωτήματος και στη συνέχεια επαληθεύει την εμφάνιση του στόχου στο προφίλ. Το ερώτημα ενισχύεται με βάση τα κριτήρια επιλογής του χρήστη, εάν η επαλήθευση είναι θετική. Επιπλέον, το σύστημα παρέχει χρήσιμες υπηρεσίες στον χρήστη με βάση την υβριδική λύση ή τις αντιστοιχίσεις περιεχομένου χρήστη και την εκτίμηση (Eke et al., 2019).

5 Αλγόριθμοι και δεδομένα υλοποίησης

Τα δεδομένα που θα χρησιμοποιηθούν στο σύστημα συστάσεων προέρχονται από μια εφαρμογή κινητών συσκευών η οποία αναπτύχθηκε για την ανάδειξη των προϊόντων και υπηρεσιών μιας ξενοδοχειακής μονάδας. Σκοπός της εφαρμογής είναι η συλλογή πλήθος δεδομένων από τους χρήστες – πελάτες με σκοπό τόσο την καλύτερη εμπειρία των πελατών μέσω εξατομικευμένων συστάσεων βάση των προτιμήσεων τους όσο και την παρακολούθηση και βελτίωση των υπηρεσιών που παρέχει η επιχείρηση.

Στην εφαρμογή τα δεδομένα συλλέγονται κατά κύριο λόγο με δύο τρόπους:

- Μέσω ερωτηματολογίων,
- Μέσω καταγραφής της δραστηριότητας των χρηστών (clickstreams)

5.1 Αλγόριθμοι υλοποίησης

Για τις ανάγκες της υλοποίησης του έργου επιλέχθηκαν δύο πολύ γνωστοί αλγόριθμοι: τα Δέντρα αποφάσεων και το Random Forest (RF). Διερευνήθηκε η απόδοσή τους και τα αποτελέσματα τους με σκοπό την χρήση ενός από αυτούς. Ακολουθεί η αναλυτική παρουσίασή τους.

5.1.1 Δέντρα αποφάσεων (Decision Trees)

Τα δέντρα αποφάσεων αποτελούν έναν δημοφιλή αλγόριθμο μηχανικής μάθησης. Είναι εύκολα στην κατανόηση, στην ερμηνεία και στην εφαρμογή τους, γεγονός που τους καθιστά ιδανική επιλογή για αρχάριους στον τομέα της μηχανικής μάθησης.

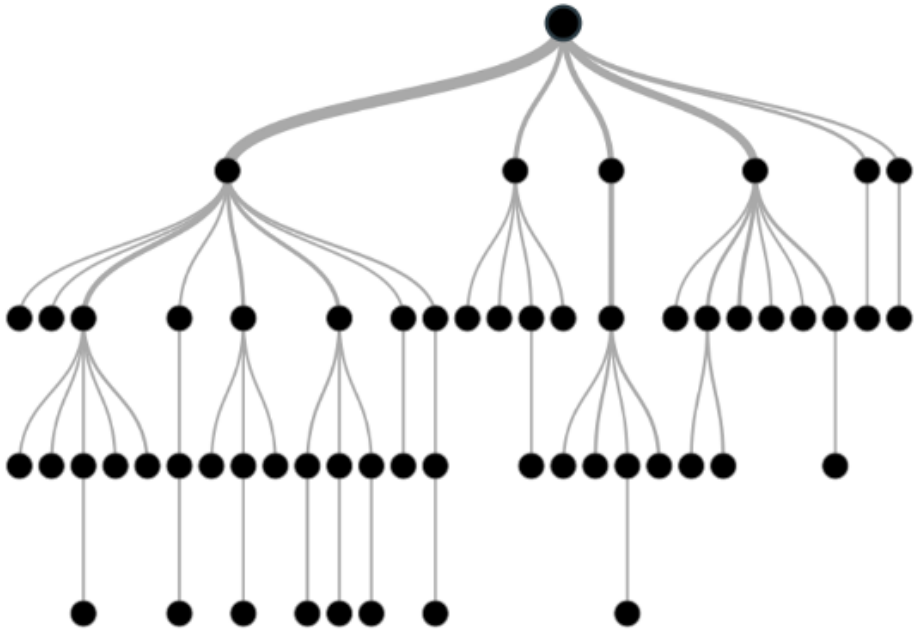
Ένα δέντρο αποφάσεων είναι ένα μοντέλο πρόβλεψης που χρησιμοποιεί μια δομή η οποία μοιάζει με διάγραμμα ροής για να λαμβάνει αποφάσεις με βάση τα δεδομένα εισόδου. Διαχωρίζει τα δεδομένα σε κλάδους και αναθέτει αποτελέσματα σε κόμβους φύλλων. Τα δέντρα αποφάσεων χρησιμοποιούνται για εργασίες ταξινόμησης και παλινδρόμησης, παρέχοντας εύκολα κατανοητά μοντέλα.

Αποτελούν ένα ιεραρχικό μοντέλο που χρησιμοποιείται στην υποστήριξη αποφάσεων και απεικονίζει τις αποφάσεις και τα πιθανά αποτελέσματα τους, ενσωματώνοντας τυχαία γεγονότα, δαπάνες πόρων και χρησιμότητα. Αυτό το αλγοριθμικό μοντέλο χρησιμοποιεί δηλώσεις ελέγχου υπό όρους και είναι μη παραμετρική, επιβλεπόμενη

μάθηση. Η δενδρική δομή αποτελείται από έναν κόμβο ρίζας, κλαδιά εσωτερικούς κόμβους και κόμβους φύλλων, σχηματίζοντας μια ιεραρχική, δενδρική δομή.

Ο στόχος της χρήσης ενός δέντρου αποφάσεων είναι η δημιουργία ενός μοντέλου εκπαίδευσης που μπορεί να χρησιμοποιηθεί για την πρόβλεψη της κλάσης ή της τιμής της μεταβλητής – στόχου με τη μάθηση απλών κανόνων απόφασης που προκύπτουν από προηγούμενα δεδομένα (δεδομένα εκπαίδευσης).

Πρόκειται για ένα εργαλείο που έχει εφαρμογές που εκτείνονται σε διάφορους τομείς. Τα δέντρα αποφάσεων μπορούν να χρησιμοποιηθούν τόσο για προβλήματα ταξινόμησης όσο και για προβλήματα παλινδρόμησης. Το ίδιο το όνομα υποδηλώνει ότι χρησιμοποιεί ένα διάγραμμα ροής σαν δομή δέντρου για να δείξει τις προβλέψεις που προκύπτουν από μια σειρά διαχωρισμών με βάση τα χαρακτηριστικά. Για την πρόβλεψη μιας ετικέτας κλάσης για μια εγγραφή η διαδικασία ξεκινάει από τη ρίζα του δέντρου. Γίνεται σύγκριση τις τιμής του χαρακτηριστικού της ρίζας με το χαρακτηριστικό της εγγραφής. Βάσει της σύγκρισης, ακολουθείτε ο κλάδος που αντιστοιχεί σε αυτή την τιμή και μεταβαίνει στον επόμενο κόμβο (Saini, 2021).



Εικόνα 5.1 Παράδειγμα δομής ενός δέντρου αποφάσεων

5.1.1.1 Δομικά στοιχεία ενός δέντρου αποφάσεων

Τα δέντρα αποφάσεων διαθέτουν ορισμένα δομικά στοιχεία καθώς και κάποιες λειτουργίες στις οποίες βασίζονται για την δημιουργία των τελικών προβλέψεων αποφάσεων.

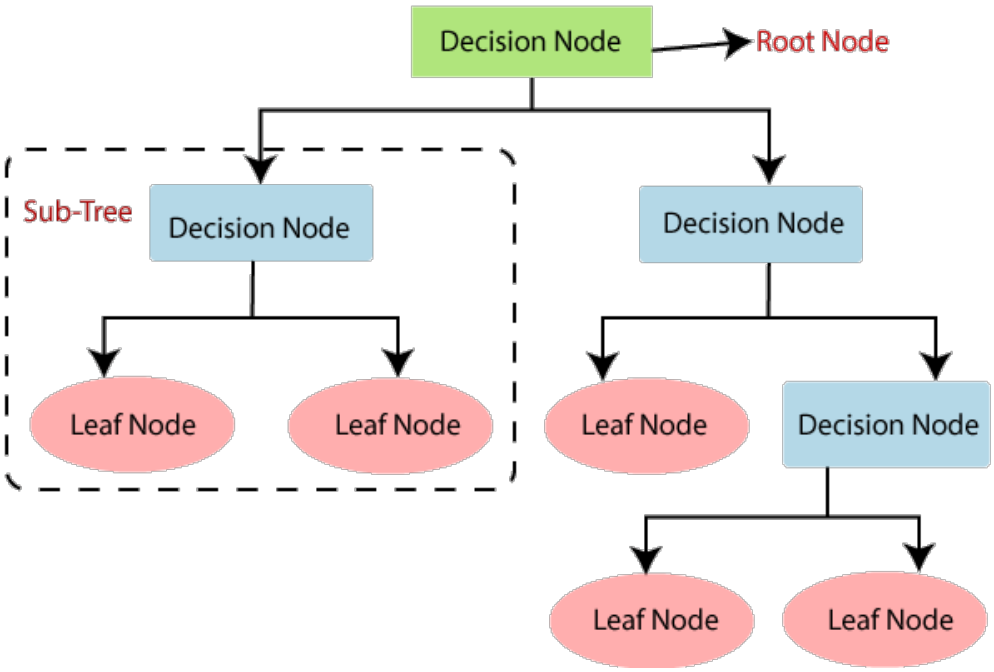
Αυτά τα δομικά στοιχεία είναι:

- **Κόμβος ρίζας (Root node):** Είναι ο κόμβος που υπάρχει στην αρχή ενός δέντρου αποφάσεων (βάση του δέντρου). Αντιπροσωπεύει ολόκληρο τον πληθυσμό ή το δείγμα και αυτό διαιρείται περαιτέρω σε δύο ή περισσότερα ομοιογενή σύνολα.
- **Κόμβοι απόφασης / εσωτερικός (Decision / Internal node):** Οι κόμβοι που προκύπτουν μετά τη διάσπαση των κόμβων ρίζας. Συμβολίζουν μια επιλογή σχετικά με ένα χαρακτηριστικό εισόδου.
- **Κόμβος φύλλων / Τερματικό κόμβος (Leaf / Terminal node):** Ένας κόμβος που δεν είναι δυνατή ή περαιτέρω διάσπαση του (χωρίς κόμβους – παιδιά) και υποδεικνύει μια ετικέτα κλάσης ή μια αριθμητική τιμή.
- **Υποδέντρο / Κλάδος (Sub – tree / Branch):** Ένα υποτμήμα του δέντρου αποφάσεων που ξεκινά από έναν εσωτερικό κόμβο και καταλήγει στους κόμβους φύλλων.

Πιο κάτω παρουσιάζονται ορισμένες ζωτικές διαδικασίες που πραγματοποιούνται σε ένα δέντρο αποφάσεων:

- **Διάσπαση (Splitting):** Διάσπαση είναι η διαδικασία διαίρεσης του κόμβου απόφασης / κόμβου ρίζας σε υποκόμβους σύμφωνα με τις δεδομένες συνθήκες.
- **Κλάδεμα (Pruning):** Κλάδεμα είναι η διαδικασία αφαίρεσης των ανεπιθύμητων κλάδων από το δέντρο. Ως ανεπιθύμητος κλάδος θεωρείται ένας που δεν παρέχει πρόσθετες πληροφορίες ή οδηγούν σε υπερπροσαρμογή (overfitting).

Κάθε κόμβος στο δέντρο λειτουργεί ως περίπτωση δοκιμής για κάποιο χαρακτηριστικό, και κάθε ακμή που κατεβαίνει από τον κόμβο αντιστοιχεί στις πιθανές απαντήσεις στην περίπτωση δοκιμής. Η διαδικασία αυτή είναι αναδρομική στη φύση της και επαναλαμβάνεται για κάθε υποδέντρο που έχει τη ρίζα του στο νέο κόμβο.



Εικόνα 5.2 Δέντρο απόφασης με τα δομικά του στοιχεία

(Πηγή: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm)

5.1.1.2 Τρόπος λειτουργίας των δέντρων αποφάσεων

Σε ένα δέντρο απόφασης, για την πρόβλεψη της κλάσης του δεδομένου συνόλου δεδομένων, ο αλγόριθμος ξεκινά από τον κόμβο-ρίζα του δέντρου. Αυτός ο αλγόριθμος συγκρίνει τις τιμές του χαρακτηριστικού της ρίζας με το χαρακτηριστικό της εγγραφής (πραγματικό σύνολο δεδομένων) και, με βάση τη σύγκριση, ακολουθεί τον κλάδο και μεταπηδά στον επόμενο κόμβο.

Για τον επόμενο κόμβο, ο αλγόριθμος συγκρίνει και πάλι την τιμή του χαρακτηριστικού με τους άλλους υποκόμβους και προχωράει περαιτέρω. Συνεχίζει τη διαδικασία μέχρι να φτάσει στον κόμβο φύλλο του δέντρου.

Η πλήρης διαδικασία μπορεί να γίνει καλύτερα κατανοητή χρησιμοποιώντας τον παρακάτω αλγόριθμο:

- **Βήμα 1:** Δημιουργία του δέντρου από τον κόμβο ρίζας S, ο οποίος περιέχει το πλήρες σύνολο δεδομένων.
- **Βήμα 2:** Εύρεση του καλύτερου χαρακτηριστικού στο σύνολο δεδομένων χρησιμοποιώντας το μέτρο επιλογής χαρακτηριστικών (ASM).

- **Βήμα 3:** Διαχωρισμός του S σε υποσύνολα που περιέχουν πιθανές τιμές για τα καλύτερα χαρακτηριστικά.
- **Βήμα 4:** Δημιουργία του κόμβου απόφασης ο οποίος περιέχει το καλύτερο χαρακτηριστικό.
- **Βήμα 5:** Δημιουργία αναδρομικά νέων δέντρων απόφασης χρησιμοποιώντας τα υποσύνολα του συνόλου δεδομένων που δημιουργήθηκαν στο βήμα 3. Η διαδικασία επαναλαμβάνεται μέχρι να μην μπορούν να ταξινομηθούν περαιτέρω οι κόμβοι ονομάζοντας τους τελικούς κόμβους ως κόμβους φύλλα.

Κατά την εφαρμογή ενός δέντρου αποφάσεων, το κύριο ζήτημα που προκύπτει είναι πώς θα γίνει η επιλογή του καλύτερου χαρακτηριστικού για τον κόμβο-ρίζα και για τους εσωτερικούς κόμβους. Έτσι, για την επίλυση τέτοιων προβλημάτων υπάρχει μια τεχνική που ονομάζεται μέτρο επιλογής χαρακτηριστικών ή ASM (Attribute selection measure). Με αυτό το μέτρο, γίνεται πιο εύκολα η διαδικασία επιλογής του καλύτερου χαρακτηριστικού για τους κόμβους του δέντρου.

Υπάρχουν δύο δημοφιλείς τεχνικές για το ASM οι οποίες είναι οι εξής:

- Κέρδος πληροφορίας (Information Gain)
- Δείκτης Gini (Gini Index)
- Gain ratio (Αναλογία κέρδους)

Κέρδος πληροφορίας (Information Gain)

Το κέρδος πληροφορίας είναι μια στατιστική ιδιότητα που μετράει τη μείωση της αβεβαιότητας δεδομένου κάποιου χαρακτηριστικού και είναι επίσης αποφασιστικός παράγοντας για το ποιο χαρακτηριστικό θα πρέπει να επιλεγεί ως κόμβος απόφασης ή ριζικός κόμβος. Η κατασκευή ενός δέντρου απόφασης έχει να κάνει με την εύρεση ενός χαρακτηριστικού που αποδίδει το μεγαλύτερο κέρδος πληροφορίας και τη μικρότερη εντροπία.

Υπολογίζει τη διαφορά μεταξύ της εντροπίας πριν από τη διάσπαση του συνόλου δεδομένων με βάση τις δεδομένες τιμές χαρακτηριστικών. Για παράδειγμα ο αλγόριθμος ID3 χρησιμοποιεί κέρδος πληροφορίας.

Μαθηματικά, η IG αναπαρίσταται ως εξής:

$$Information = Entropy\,(before) - \sum_{j=1}^{K} Entropy(j, after)$$

Όπου before είναι το σύνολο δεδομένων πριν από τη διάσπαση, K είναι ο αριθμός των υποσυνόλων που δημιουργούνται από τη διάσπαση και (j, after) είναι το υποσύνολο j μετά τη διάσπαση.

Εντροπία

Η εντροπία είναι ένα μέτρο της τυχαιότητας των πληροφοριών που υποβάλλονται σε επεξεργασία. Όσο υψηλότερη είναι η εντροπία, τόσο πιο δύσκολο είναι να εξαχθούν συμπεράσματα από τις πληροφορίες αυτές. Προσδιορίζει την τυχαιότητα στα δεδομένα.

Η εντροπία μπορεί να υπολογιστεί ως εξής:

$$Entropy(s) = -P(yes)\log 2 * P(yes) - P(no)\log 2 * P(no)$$

Δείκτης Gini (Gini Index)

Ο δείκτης Gini θεωρείται ως μία συνάρτηση κόστους που χρησιμοποιείται για την αξιολόγηση των διαχωρισμών στο σύνολο δεδομένων. Ένα χαρακτηριστικό με χαμηλό δείκτη Gini θα πρέπει να προτιμάται σε σύγκριση με τον υψηλό δείκτη Gini.

Ο δείκτης Gini μπορεί να υπολογιστεί ως εξής:

$$Gini = 1 - \sum_{i=1}^{c} (pi)^2$$

Αναλογία κέρδους (Gain ratio)

Το κέρδος πληροφορίας μεροληπτεί προς την επιλογή χαρακτηριστικών με μεγάλο αριθμό τιμών ως κόμβους ρίζας. Αυτό σημαίνει ότι προτιμά το χαρακτηριστικό με μεγάλο αριθμό διακριτών τιμών.

Ο αλγόριθμος C4.5, μια βελτίωση του ID3, χρησιμοποιεί την αναλογία κέρδους, η οποία είναι μια τροποποίηση του κέρδους πληροφορίας η οποία μειώνει την προκατάληψή του και είναι συνήθως η καλύτερη επιλογή. Ο Gain ratio ξεπερνά το πρόβλημα με το κέρδος πληροφορίας λαμβάνοντας υπόψη τον αριθμό των διακλαδώσεων που θα προκύψουν

πριν από τη διάσπαση. Διορθώνει το κέρδος πληροφορίας λαμβάνοντας υπόψη την εγγενή πληροφορία μιας διάσπασης.

Ο δείκτης Gain ration μπορεί να υπολογιστεί ως εξής:

$$Gain\ Ratio = \frac{Information\ Gain}{Split\ Info} = \frac{Entropy(before) - \sum_{j=1}^{K} Entropy(j, after)}{\sum_{j=1}^{K} w_j \log_2 w_j}$$

Συνήθως, τα σύνολα δεδομένων του πραγματικού κόσμου έχουν μεγάλο αριθμό χαρακτηριστικών, γεγονός που οδηγεί σε μεγάλο αριθμό διαχωρισμών, ο οποίος με τη σειρά του δίνει ένα τεράστιο δέντρο. Η κατασκευή τέτοιων δέντρων απαιτεί χρόνο και μπορεί να οδηγήσει σε υπερπροσαρμογή. Αυτό σημαίνει ότι το δέντρο θα δώσει πολύ καλή ακρίβεια στο σύνολο δεδομένων εκπαίδευσης, αλλά θα δώσει κακή ακρίβεια στα δεδομένα δοκιμής.

5.1.1.3 Τύποι δέντρων αποφάσεων

Υπάρχουν δύο βασικοί τύποι δέντρων αποφάσεων κατηγορικά και συνεχή. Ο διαχωρισμός βασίζεται στον τύπο των μεταβλητών που χρησιμοποιούνται ως αποτέλεσμα.

Δέντρα απόφασης κατηγορικών μεταβλητών

Σε ένα δέντρο απόφασης κατηγορικών μεταβλητών, η απάντηση εντάσσεται με σαφήνεια στη μία ή την άλλη κατηγορία. Για παράδειγμα σε μια ρίψη νομίσματος το αποτέλεσμα είναι κορώνα ή γράμματα. Σε αυτόν τον τύπο δέντρου αποφάσεων, τα δεδομένα τοποθετούνται σε μία μόνο κατηγορία με βάση τις αποφάσεις στους κόμβους σε όλο το δέντρο (Charbuty and Abdulazeez, 2021).

Δέντρα απόφασης για συνεχείς μεταβλητές

Ένα δέντρο αποφάσεων συνεχούς μεταβλητής είναι ένα δέντρο όπου δεν υπάρχει μια απλή απάντηση με ναι ή όχι. Είναι επίσης γνωστό ως δέντρο παλινδρόμησης επειδή η μεταβλητή απόφασης ή αποτελέσματος εξαρτάται από άλλες αποφάσεις που βρίσκονται πιο ψηλά στο δέντρο ή από τον τύπο της επιλογής που εμπλέκεται στην απόφαση.

Βασικό πλεονέκτημα ενός δέντρου αποφάσεων με συνεχείς μεταβλητές είναι ότι το αποτέλεσμα μπορεί να προβλεφθεί με βάση πολλαπλές μεταβλητές και όχι με βάση μία

μόνο μεταβλητή όπως σε ένα δέντρο αποφάσεων με κατηγορικές μεταβλητές. Τα δέντρα αποφάσεων συνεχών μεταβλητών χρησιμοποιούνται για τη δημιουργία προβλέψεων. Το σύστημα μπορεί να χρησιμοποιηθεί τόσο για γραμμικές όσο και για μη γραμμικές σχέσεις, εάν επιλεγεί ο σωστός αλγόριθμος (Charbuty and Abdulazeez, 2021).

Μερικοί βασικοί αλγόριθμοι των δέντρων αποφάσεων είναι:

- **CART:** Ο όρος CART αποτελεί συντομογραφία για το Classification And Regression Trees) δηλαδή δέντρα ταξινόμησης και παλινδρόμησης και εισήχθη από τον Leo Breiman. Αυτός ο αλγόριθμος χρησιμοποιεί συνήθως την τιμή Gini για τον προσδιορισμό του ιδανικού χαρακτηριστικού για διαχωρισμό. Η τιμή Gini μετράει πόσο συχνά ένα τυχαία επιλεγμένο χαρακτηριστικό ταξινομείται εσφαλμένα. Κατά την αξιολόγηση με χρήση της μεταβλητής Gini, η χαμηλότερη τιμή είναι πιο ιδανική.
- **ID3:** Ο όρος ID3 αποτελεί συντομογραφία για το Iterative Dichotomiser 3 δηλαδή επαναληπτικός διχοτομητής 3. Αυτός ο αλγόριθμος αξιοποιεί την εντροπία και το κέρδος πληροφορίας ως μετρικές για την αξιολόγηση των υποψήφιων διαχωρισμών.
- **C4.5:** Αυτός ο αλγόριθμος θεωρείται μεταγενέστερη επανάληψη του ID3. Μπορεί να χρησιμοποιήσει το κέρδος πληροφορίας ή τους λόγους κέρδους για την αξιολόγηση των σημείων διαχωρισμού εντός των δέντρων απόφασης.

5.1.1.4 Πλεονεκτήματα και Μειονεκτήματα των δέντρων αποφάσεων

Τα δέντρα αποφάσεων είναι ιδιαίτερα χρήσιμα για εργασίες εξόρυξης δεδομένων και ανακάλυψης γνώσης. Παρόλα αυτά διαθέτουν και ορισμένες προκλήσεις από την χρήση τους. Πιο κάτω παρουσιάζονται τα βασικά οφέλη και οι προκλήσεις των δέντρων αποφάσεων (Charbuty and Abdulazeez, 2021).

Πλεονεκτήματα

- **Εύκολη ερμηνεία:** Η λογική Boolean (True / False) σε κάθε κόμβο απόφασης και οι οπτικές αναπαραστάσεις των δέντρων αποφάσεων τα καθιστούν ευκολότερα και κατανοητά ως προς τους χρήστες. Η ιεραρχική φύση ενός δέντρου αποφάσεων καθιστά επίσης εύκολο την διάκριση των χαρακτηριστικών που είναι πιο σημαντικά, κάτι που δεν είναι σαφές με άλλους αλγορίθμους όπως τα νευρωνικά δίκτυα.
- **Ελαχίστη έως καθόλου προετοιμασία δεδομένων:** Τα δέντρα αποφάσεων έχουν ορισμένα χαρακτηριστικά, τα οποία τα καθιστούν πιο ευέλικτα από άλλους

ταξινομητές. Μπορούν να χειριστούν διάφορους τύπους δεδομένων δηλαδή διακριτές ή συνεχείς τιμές, και οι συνεχείς τιμές μπορούν να μετατραπούν σε κατηγορικές τιμές μέσω της χρήσης κατωφλιών. Επιπλέον, μπορούν να χειριστούν τιμές με ελλείπουσες τιμές, οι οποίες μπορεί να είναι προβληματικές για άλλους ταξινομητές όπως ο Naïve Bayes.

- **Πιο ευέλικτο:** Το δένδρο αποφάσεων μπορεί να χρησιμοποιηθεί τόσο για εργασίες ταξινόμησης όσο και για εργασίες παλινδρόμησης, καθιστώντας το πιο ευέλικτο από ορισμένους άλλους αλγορίθμους. Αυτό σημαίνει ότι εάν δύο μεταβλητές συσχετίζονται σε μεγάλο βαθμό, ο αλγόριθμος θα επιλέξει μόνο ένα από τα χαρακτηριστικά για διαχωρισμό.
- **Επεκτασιμότητα:** Τα δέντρα αποφάσεων μπορούν να χειριστούν μεγάλα σύνολα δεδομένων και μπορούν εύκολα να παραλληλιστούν για να βελτιώσουν τον χρόνο επεξεργασίας.
- **Χειρισμός μη γραμμικών σχέσεων:** Τα δέντρα αποφάσεων μπορούν να χειριστούν μη γραμμικές σχέσεις μεταξύ μεταβλητών, καθιστώντας τα κατάλληλη επιλογή για σύνθετα σύνολα δεδομένων.
- **Ανοχή στις ελλείπουσες τιμές:** Τα δέντρα αποφάσεων είναι σε θέση να χειρίζονται τις ελλιπείς τιμές στα δεδομένα, καθιστώντας τα κατάλληλη επιλογή για σύνολα δεδομένων με ελλιπή δεδομένα.
- **Δυνατότητα χειρισμού ανισόρροπων δεδομένων:** Τα δέντρα αποφάσεων μπορούν να χειριστούν ανισοβαρή σύνολα δεδομένων, όπου μια κλάση αντιπροσωπεύεται σε μεγάλο βαθμό σε σύγκριση με τις άλλες, σταθμίζοντας τη σημασία των μεμονωμένων κόμβων με βάση την κατανομή των κλάσεων.

Μειονεκτήματα

- **Επιρρεπείς σε υπερπροσαρμογή:** Τείνουν να υπερπροσαρμόζονται και δεν γενικεύονται καλά σε νέα δεδομένα. Αυτό το σενάριο μπορεί να αποφευχθεί μέσω των διαδικασιών προ-κλάδευσης ή μετα-κλάδευσης. Το προ-κλάδεμα σταματά την ανάπτυξη του δέντρου όταν υπάρχουν ανεπαρκή δεδομένα, ενώ το μετα-κλάδεμα αφαιρεί τα υποδέντρα με ανεπαρκή δεδομένα μετά την κατασκευή του δέντρου.
- **Εκτιμητές υψηλής διακύμανσης:** Μικρές παραλλαγές εντός των δεδομένων μπορεί να παράγουν ένα πολύ διαφορετικό δέντρο αποφάσεων. Το bagging, ή ο μέσος όρος των εκτιμήσεων, μπορεί να είναι μια μέθοδος μείωσης της διακύμανσης

των δέντρων απόφασης. Ωστόσο, αυτή η προσέγγιση είναι περιορισμένη, καθώς μπορεί να οδηγήσει σε υψηλά συσχετιζόμενους προγνωστικούς παράγοντες.

- **Πιο δαπανηρό:** Δεδομένου ότι τα δέντρα αποφάσεων ακολουθούν μια προσέγγιση άπληστης αναζήτησης κατά την κατασκευή τους, η εκπαίδευσή τους μπορεί να είναι πιο δαπανηρή σε σύγκριση με άλλους αλγορίθμους.
- **Δεν υποστηρίζεται πλήρως από το scikit-learn:** Το Scikit-learn είναι μια δημοφιλής βιβλιοθήκη μηχανικής μάθησης που βασίζεται στην Python. Ενώ αυτή η βιβλιοθήκη διαθέτει μια ενότητα δέντρων απόφασης (DecisionTreeClassifier), η τρέχουσα υλοποίηση δεν υποστηρίζει κατηγορικές μεταβλητές.

5.1.1.5 Εφαρμογές των Decision Tree

1) Διαχείριση πελατειακών σχέσεων

Μια συχνά χρησιμοποιούμενη προσέγγιση για τη διαχείριση των σχέσεων με τους πελάτες είναι η διερεύνηση του τρόπου με τον οποίο τα άτομα αποκτούν πρόσβαση σε διαδικτυακές υπηρεσίες. Μια τέτοια διερεύνηση πραγματοποιείται κυρίως με τη συλλογή και ανάλυση δεδομένων χρήσης των ατόμων και στη συνέχεια με την παροχή συστάσεων βάσει των εξαγόμενων πληροφοριών. Σε έρευνα έχουν εφαρμοστεί δέντρα αποφάσεων για να διερευνήσουν τις σχέσεις μεταξύ των αναγκών και των προτιμήσεων των πελατών και της επιτυχίας των ηλεκτρονικών αγορών. Τα δέντρα αποφάσεων που δημιουργήθηκαν δείχνουν ότι η επιτυχία ενός ηλεκτρονικού καταστήματος εξαρτάται σε μεγάλο βαθμό από τη συχνότητα των αγορών των πελατών και την τιμή των προϊόντων. Τα ευρήματα που ανακαλύπτονται από τα δέντρα αποφάσεων είναι χρήσιμα για την κατανόηση των αναγκών και των προτιμήσεων των πελατών τους.

2) Διάγνωση σφαλμάτων

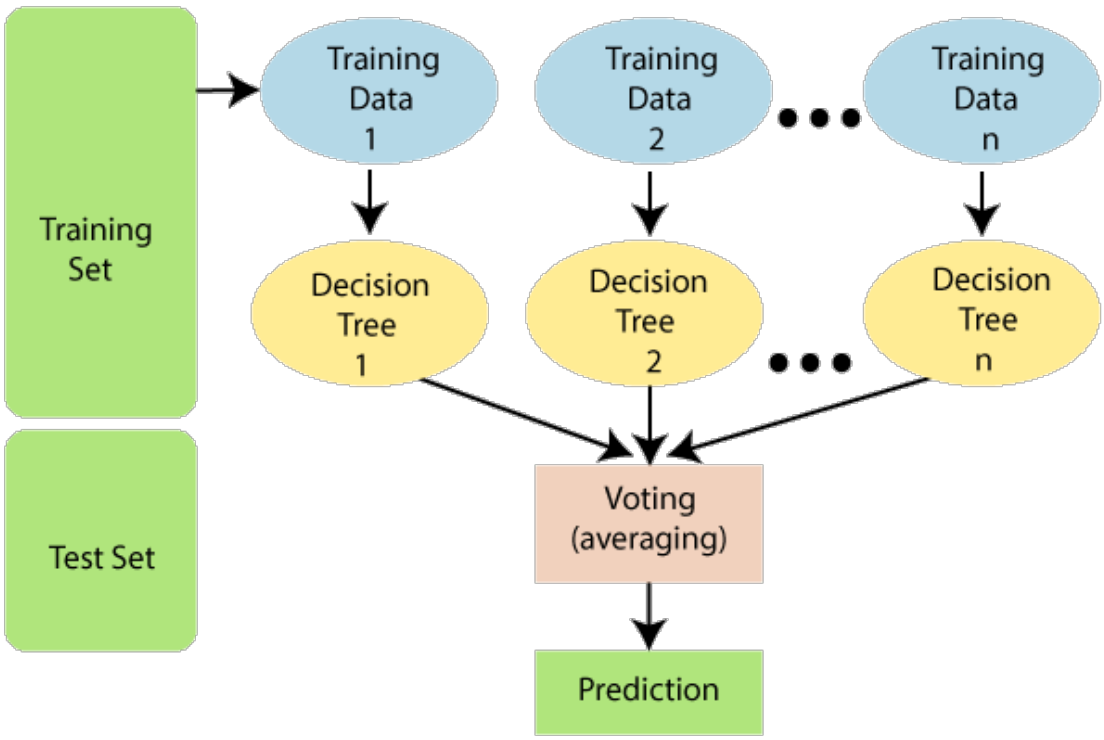
Μια ευρέως χρησιμοποιούμενη εφαρμογή στον τομέα της μηχανικής είναι η ανίχνευση σφαλμάτων, ειδικά στην αναγνώριση ενός ελαττωματικού ρουλεμάν σε περιστροφικές μηχανές. Αυτό οφείλεται πιθανώς στο γεγονός ότι το έδρανο είναι ένα από τα σημαντικότερα εξαρτήματα που επηρεάζει άμεσα τη λειτουργία μιας περιστροφικής μηχανής. Για την ανίχνευση της ύπαρξης ελαττωματικού ρουλεμάν, οι μηχανικοί τείνουν να μετρούν τα σήματα κραδασμών και ακουστικής εκπομπής (AE) που εκπέμπονται από την περιστροφική μηχανή. Ωστόσο, η μέτρηση περιλαμβάνει έναν αριθμό μεταβλητών, ορισμένες από τις οποίες μπορεί να είναι λιγότερο σημαντικές για τη διερεύνηση. Τα

δέντρα αποφάσεων είναι ένα πιθανό εργαλείο για την απομάκρυνση τέτοιων άσχετων μεταβλητών, καθώς μπορούν να χρησιμοποιηθούν για τους σκοπούς της επιλογής χαρακτηριστικών.

Οι Sugumaran και Ramachandran (2007) δημιουργούν ένα μοντέλο δέντρου αποφάσεων για τον εντοπισμό των χαρακτηριστικών που μπορεί να επηρεάσουν σημαντικά τη διερεύνηση ενός ελαττωματικού ρουλεμάν. Μέσω της επιλογής χαρακτηριστικών, επιλέχθηκαν τρία χαρακτηριστικά για τη διάκριση των ελαττωματικών συνθηκών ενός ρουλεμάν, δηλαδή η ελάχιστη τιμή του σήματος δόνησης, η τυπική απόκλιση του σήματος δόνησης και η κύρτωση. Τα επιλεγμένα χαρακτηριστικά, στη συνέχεια, χρησιμοποιήθηκαν για τη δημιουργία ενός άλλου μοντέλου δέντρου αποφάσεων. Οι αξιολογήσεις από αυτό το μοντέλο δείχνουν ότι πάνω από το 95% του συνόλου δεδομένων δοκιμής έχει ταξινομηθεί σωστά. Ένα τόσο υψηλό ποσοστό ακρίβειας υποδηλώνει ότι η απομάκρυνση των ασήμαντων χαρακτηριστικών εντός ενός συνόλου δεδομένων είναι μια άλλη συμβολή των δέντρων απόφασης.

5.1.2 Random Forest

Ο αλγόριθμος Random Forest - RF είναι ένας αλγόριθμος μηχανικής μάθησης που ανήκει στην τεχνική μάθησης με επίβλεψη (supervised learning). Μπορεί να χρησιμοποιηθεί τόσο για προβλήματα ταξινόμησης όσο και για προβλήματα παλινδρόμησης στην Μηχανική Μάθηση. Βασίζεται στην έννοια της μάθησης συνόλου, η οποία είναι μια διαδικασία συνδυασμού πολλαπλών ταξινομητών για την επίλυση ενός σύνθετου προβλήματος και τη βελτίωση της απόδοσης του μοντέλου.



Εικόνα 5.3 Παράδειγμα λειτουργίας ενός Random Forest

(Πηγή: https://www.javatpoint.com/machine-learning-random-forest-algorithm)

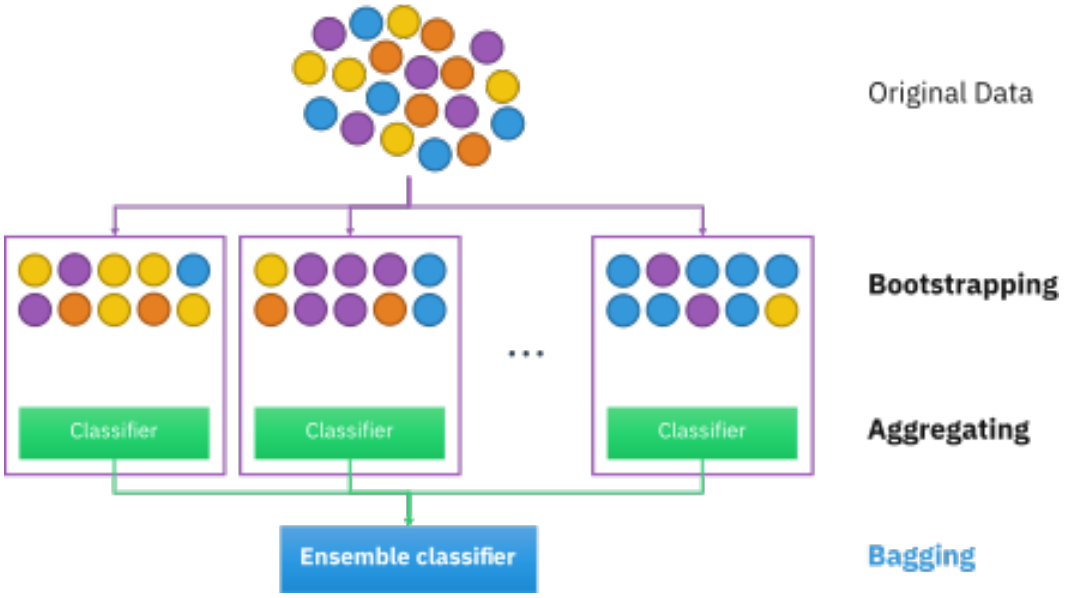
Όπως υποδηλώνει και το όνομα του αλγορίθμου ο Random Forest είναι ένας ταξινομητής ο οποίος περιέχει έναν αριθμό δέντρων απόφασης σε διάφορα υποσύνολα ενός συγκεκριμένου συνόλου δεδομένων και λαμβάνει το μέσο όρο για να βελτιώσει την ακρίβεια πρόβλεψης του εν λόγω συνόλου δεδομένων. Αντί να βασίζεται σε ένα δέντρο απόφασης, ο RF λαμβάνει την πρόβλεψη από κάθε δέντρο και με βάση τις πλειοψηφικές ψήφους των προβλέψεων δημιουργεί την τελική έξοδο – αποτέλεσμα. Όσο μεγαλύτερος ο αριθμός των δέντρων στο «δάσος» τόσο μεγαλύτερη η ακρίβεια του τελικού αποτελέσματος αλλά και τόσο μικρότερη η εμφάνιση του προβλήματος υπερπροσαρμογής (overfitting) (Parmar et al., 2019).

5.1.2.1 Τρόπος λειτουργίας ενός RF

Ο αλγόριθμος Random Forest βασίζεται στην τεχνική μάθησης συνόλου η οποία συνδυάζει πολλαπλά μοντέλα. Επομένως γίνεται χρήση μιας συλλογής δεδομένων για να γίνουν οι προβλέψεις και όχι ένα μεμονωμένο μοντέλο. Στον Random Forest χρησιμοποιούνται η μέθοδος **Bagging**.

Το bagging, επίσης γνωστό ως Bootstrap Aggregation είναι μια τεχνική που χρησιμοποιείται από τον αλγόριθμο Random Forest. Αρχικά επιλέγει ένα τυχαίο δείγμα

υποσύνολο από το σύνολο των δεδομένων. Ως εκ τούτου, κάθε μοντέλο δημιουργείται από τα δείγματα (Bootstrap Samples) που παρέχονται από τα αρχικά δεδομένα με αντικατάσταση, γνωστό και ως δειγματοληψία σειράς. Αυτό το βήμα της δειγματοληψίας σειράς με αντικατάσταση ονομάζεται bootstrap. Κάθε μοντέλο εκπαιδεύεται ανεξάρτητα το οποίο παράγει και ένα αποτέλεσμα. Το τελικό αποτέλεσμα του μοντέλου βασίζεται στην ψηφοφορία πλειοψηφίας μετά το συνδυασμό των αποτελεσμάτων όλων των μοντέλων. Αυτό το βήμα που περιλαμβάνει το συνδυασμό όλων των αποτελεσμάτων και τη δημιουργία εξόδου με βάση την ψηφοφορία πλειοψηφίας, είναι γνωστό ως συνάθροιση.



Εικόνα 5.4 Αναπαράσταση της τεχνικής Boosting

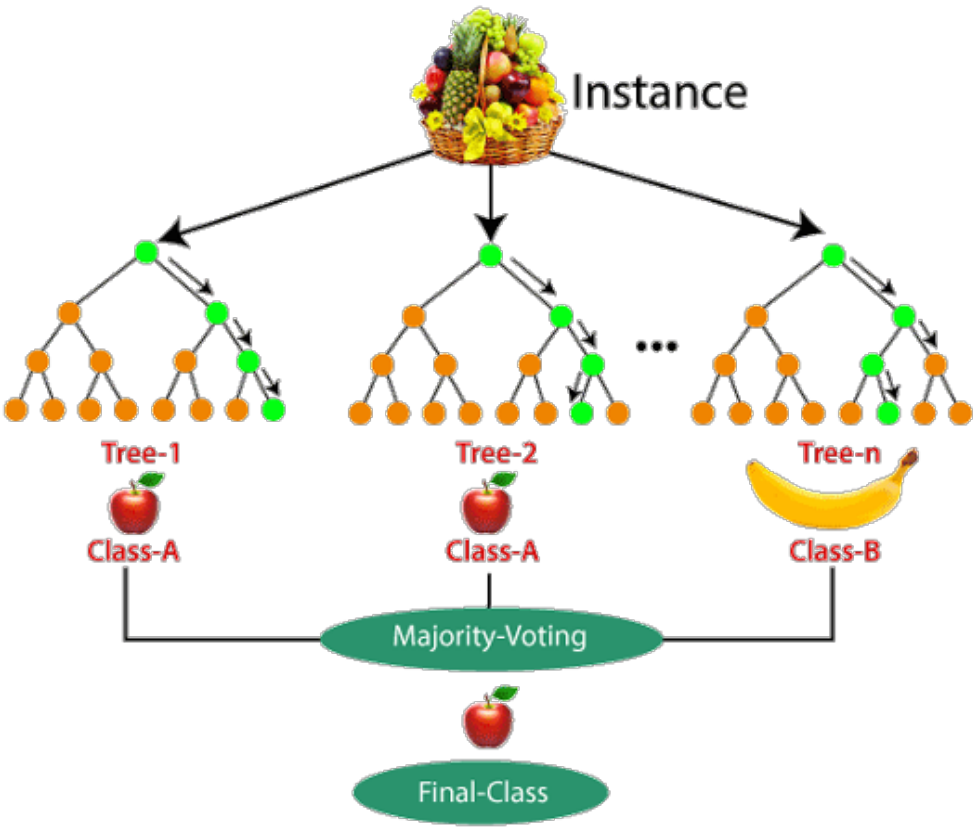
(Πηγή: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#Working_of_Random_Forest_Algorithm)

Ο αλγόριθμος Random Forest λειτουργεί σε δύο φάσεις. Στην πρώτη φάση γίνεται η δημιουργία ενός τυχαίου δάσους συνδυάζοντας N δέντρα αποφάσεων. Στην δεύτερη φάση γίνονται οι προβλέψεις για κάθε δέντρο που έχει δημιουργηθεί στην πρώτη φάση (Donges, 2023).

Η διαδικασία λειτουργίας ενός Random Forest είναι η εξής:

- **Βήμα 1:** Γίνεται η επιλογή ενός υποσυνόλου σημείων δεδομένων και ένα υποσυνόλου χαρακτηριστικών για την κατασκευή κάθε δέντρου απόφασης.
- **Βήμα 2:** Κατασκευάζονται μεμονωμένα δέντρα απόφασης που σχετίζονται με τα επιλεγμένα σημεία δεδομένων.

- **Βήμα 3:** Υπολογισμός της πρόβλεψης που παράγει κάθε δέντρο απόφασης ως έξοδος.
- **Βήμα 4:** Υπολογισμός της τελικής εξόδου με βάση την ψηφοφορία πλειοψηφίας ή τη μέση τιμή για την ταξινόμηση και την παλινδρόμηση.



Εικόνα 5.5 Παράδειγμα λειτουργίας ενός Random Forest

(Πηγή: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#Working_of_Random_Forest_Algorithm)

5.1.2.2 Βασικά χαρακτηριστικά και παράμετροι ενός RF

Βασικά χαρακτηριστικά του Random Forest

Ορισμένα από τα πιο σημαντικά χαρακτηριστικά ενός Random Forest είναι (Donges, 2023):

- **Ποικιλομορφία (Diversity):** Κάθε δέντρο έχει μοναδικές ιδιότητες, ποικιλία και χαρακτηριστικά που διαφέρουν από άλλα δέντρα.
- **Παραλληλισμός (Parallelization):** Κάθε δέντρο δημιουργείται ανεξάρτητα από διαφορετικά δεδομένα και χαρακτηριστικά. Αυτό σημαίνει ότι μπορεί να

χρησιμοποιηθεί πλήρως η CPU και οι πυρήνες της για την δημιουργία τυχαίων δασών.

- **Σταθερότητα (Stability):** Η σταθερότητα προκύπτει επειδή το τελικό αποτέλεσμα βασίζεται στο Bagging δηλαδή στην ψηφοφορία πλειοψηφίας ή στο μέσο όρο.
- **Ανοσία στην κατάρα της διαστατικότητας:** Δεδομένου ότι κάθε δέντρο δεν εξετάζει όλα τα χαρακτηριστικά, ο χώρος τω χαρακτηριστικών μειώνεται.

Υπερπαράμετροι (hyperparameters)

Οι παράμετροι ή αλλιώς υπερπαράμετροι (hyperparameters) στο τυχαίο δάσος χρησιμοποιούνται είτε για να αυξήσουν την προβλεπτική ικανότητα του μοντέλου είτε για να κάνουν το μοντέλο πιο γρήγορο.

Ορισμένες από αυτές τις παραμέτρους είναι:

- **N_estimators:** Αριθμός των δέντρων που κατασκευάζει ο αλγόριθμος πριν την μέση τιμή των προβλέψεων.
- **Max_features:** Μέγιστος αριθμός χαρακτηριστικών που χρησιμοποιεί το τυχαίο δάσος πριν εξετάσει τη διάσπαση ενός κόμβου.
- **Mini_sample_leaf:** Καθορίζει τον ελάχιστο αριθμό φύλλων που απαιτούνται για τη διάσπαση ενός εσωτερικού κόμβου.

Οι ακόλουθες υπερπαράμετροι χρησιμοποιούνται για την αύξηση της ταχύτητας του μοντέλου:

- **N_jobs:** Δηλώνει τον αριθμό των επεξεργαστών που θα χρησιμοποιηθούν. Εάν η τιμή είναι 1 μπορεί να χρησιμοποιηθεί μόνο ένας επεξεργαστής, αλλά εάν η τιμή είναι -1 δεν υπάρχει όριο στον αριθμό των επεξεργαστών.
- **Random_state:** Ελέγχει την τυχαιότητα του δείγματος. Το μοντέλο θα παράγει πάντα τα ίδια αποτελέσματα αν έχει μια συγκεκριμένη τιμή της τυχαίας κατάστασης και αν του έχουν δοθεί οι ίδιες υπερπαράμετροι και τα ίδια δεδομένα εκπαίδευσης.
- **Oob_score:** Η OOB (Out Of the Bag) είναι μια μέθοδος διασταυρούμενης επικύρωσης του τυχαίου δάσους. Σε αυτήν, το ένα τρίτο του δείγματος δεν χρησιμοποιείται για την εκπαίδευση των δεδομένων αλλά για την αξιολόγηση της απόδοσής του.

5.1.2.3 Πλεονεκτήματα και μειονεκτήματα των RF

Υπάρχουν ορισμένα βασικά πλεονεκτήματα καθώς και προκλήσεις που παρουσιάζουν οι αλγόριθμοι Random Forest κατά την χρήση τους σε προβλήματα ταξινόμησης ή παλινδρόμησης.

Πλεονεκτήματα

1. **Μειωμένος κίνδυνος υπερπροσαρμογής**

Τα δέντρα αποφάσεων διατρέχουν τον κίνδυνο υπερπροσαρμογής, καθώς τείνουν να προσαρμόζουν αυστηρά όλα τα δείγματα στα δεδομένα εκπαίδευσης. Ωστόσο, όταν υπάρχει ένας ισχυρός αριθμός δέντρων απόφασης σε ένα τυχαίο δάσος, ο ταξινομητής δεν θα υπερπροσαρμόσει το μοντέλο, καθώς ο μέσος όρος των ασυσχέτιστων δέντρων μειώνει τη συνολική διακύμανση και το σφάλμα πρόβλεψης.

2. **Παρέχει ευελιξία**

Δεδομένου ότι ο αλγόριθμος Random Forest μπορεί να χειριστεί τόσο εργασίες παλινδρόμησης όσο και ταξινόμησης με υψηλό βαθμό ακρίβειας, είναι μια δημοφιλής μέθοδος μεταξύ των επιστημόνων δεδομένων. Το feature bagging καθιστά επίσης τον ταξινομητή random forest ένα αποτελεσματικό εργαλείο για την εκτίμηση ελλιπών τιμών, καθώς διατηρεί την ακρίβεια όταν ένα μέρος των δεδομένων λείπει.

3. **Εύκολος προσδιορισμός της σημασίας των χαρακτηριστικών**

Ο Random Forest διευκολύνει την αξιολόγηση της σημασίας των μεταβλητών ή της συμβολής τους στο μοντέλο. Υπάρχουν μερικοί τρόποι αξιολόγησης της σημασίας των χαρακτηριστικών. Η σημασία Gini και η μέση μείωση της ακαθαρσίας (MDI) χρησιμοποιούνται συνήθως για να μετρήσουν πόσο μειώνεται η ακρίβεια του μοντέλου όταν αποκλείεται μια συγκεκριμένη μεταβλητή. Ωστόσο, η σημαντικότητα μετάθεσης, επίσης γνωστή ως μέση μείωση της ακρίβειας (MDA), είναι ένα άλλο μέτρο σημαντικότητας. Το MDA προσδιορίζει τη μέση μείωση της ακρίβειας με τυχαία αντιμετάθεση των τιμών των χαρακτηριστικών σε δείγματα oob.

Μειονεκτήματα

1. **Χρονοβόρα διαδικασία**

Σε σύγκριση με τα δέντρα αποφάσεων είναι πιο αργοί καθώς υπολογίζουν δεδομένα για κάθε μεμονωμένο δέντρο απόφασης.

2. Απαιτεί περισσότερους πόρους

Δεδομένου ότι τα τυχαία δάση επεξεργάζονται μεγαλύτερα σύνολα δεδομένων, απαιτούν περισσότερους πόρους για την αποθήκευση των δεδομένων αυτών.

3. Απαιτεί περισσότερους πόρους

Η πρόβλεψη ενός μεμονωμένου δέντρου απόφασης είναι ευκολότερο να ερμηνευτεί σε σύγκριση με ένα δάσος από αυτά.

5.1.2.4 Εφαρμογές των RF

Υπάρχουν πολλοί τομείς στους οποίους μπορεί να χρησιμοποιηθεί η ανάλυση τυχαίων δασών (Random Forest).

Πιο κάτω παρουσιάζονται ορισμένα παραδείγματα χρήσης των RF:

1) Τμηματοποίηση πελατών

Κατά την τμηματοποίηση των πελατών γίνεται ανάλυση δεδομένων σχετικά με τα ετήσια ποσά δαπανών διαφόρων πελατών διαφόρων κατηγοριών προϊόντων για την εσωτερική δομή. Αυτό έχει ως σκοπό την καλύτερη κατανόηση της ετερογένειας των διαφόρων τύπων πελατών με τους οποίους αλληλεπιδρά ένας έμπορος χονδρικής. Με τον τρόπο αυτό, ο διανομέας θα αποκτήσει εικόνα για το πώς το πρόγραμμα διανομής του μπορεί να διαμορφωθεί καλύτερα ώστε να ανταποκρίνεται στις ανάγκες κάθε πελάτη. Με βάση τα επίπεδα του εισοδήματός τους, ο αλγόριθμος Random Forest εφαρμόστηκε σε δεδομένα της απογραφής των ΗΠΑ για το τμήμα των πελατών. Ο αλγόριθμος εφαρμόστηκε σε 16.000+ γραμμές δεδομένων με 14 χαρακτηριστικά (συνδυασμός ποιοτικών και ποσοτικών παραγόντων). Η ακρίβεια πρόβλεψης του μοντέλου στο σύνολο δοκιμών (20% τυχαία επιλεγμένα δεδομένα από το σύνολο δεδομένων) είναι περίπου 85%.

2) Πρόβλεψη διαβήτη

Ο διαβήτης είναι ένας τύπος μεταβολικής νόσου που προκαλείται από έλλειψη ινσουλίνης λόγω δυσλειτουργίας του παγκρέατος. Ο διαβήτης μπορεί να ωθήσει ένα άτομο σε παθολογική καταστροφή των β-κυττάρων του παγκρέατος, κώμα,

καρδιαγγειακή δυσλειτουργία, νεφρική και αμφιβληστροειδική ανεπάρκεια, ανεπάρκεια των αρθρώσεων, σεξουαλική δυσλειτουργία, παθογόνες επιδράσεις στην ανοσία, απώλεια βάρους και περιφερικές αγγειακές παθήσεις. Ως εκ τούτου, προτάθηκε ένα στιβαρό πλαίσιο για την έγκαιρη ανίχνευση του διαβήτη, όπου συμπληρώθηκαν η απόρριψη ακραίων τιμών, οι ελλείπουσες τιμές, η τυποποίηση δεδομένων, η K-πλάσια επικύρωση και διάφοροι ταξινομητές μηχανικής μάθησης (ML). Το Random Forest αποδίδει καλύτερα σε αυτό, μεταξύ όλων των ταξινομητών.

3) Σύσταση προιόντος

Το Amazon έχει αποδείξει ότι οι κριτικές για τα αγαθά λειτουργούν. Το 35% των πωλήσεών της οφείλεται στη μηχανή συστάσεων. Αλλά ο εντοπισμός των σωστών τάσεων στις πωλήσεις προϊόντων και στη συμπεριφορά αγορών απαιτεί μεγάλη υπολογιστική ισχύ. Ο Random Forest έχει εφαρμοστεί για την κατασκευή ενός πλαισίου συστάσεων. Η βασική παραδοχή είναι ότι όσο υψηλότερη είναι η βαθμολογία κριτικής για έναν συγκεκριμένο καταναλωτή τόσο πιο πιθανό είναι να συστηθεί για τη θέση. Για την κατασκευή ενός πιο λογικού και αποτελεσματικού σχήματος με τη χρήση του MSE για τη βαθμολογία και του Confusion Matrix για τη σύσταση ή μη, αξιολογείται η απόδοση του μοντέλου.

5.1.3 Σύγκριση Decision Tree & Random Forest

Ο αλγόριθμος Decision Tree και ο Random Forest είναι δύο δημοφιλείς μέθοδοι που χρησιμοποιούνται στη μηχανική μάθηση. Και οι δύο μέθοδοι μπορούν να χρησιμοποιηθούν για εργασίες ταξινόμησης και παλινδρόμησης αλλά υπάρχουν ορισμένες βασικές διαφορές μεταξύ τους.

Decision Tree	Random Forest
Το δέντρο αποφάσεων είναι ένα δενδροειδές μοντέλο αποφάσεων μαζί με τα πιθανά αποτελέσματα σε ένα διάγραμμα.	Ένας αλγόριθμος ταξινόμησης που αποτελείται από πολλά δέντρα αποφάσεων που συνδυάζονται για να λάβουν ένα πιο ακριβές αποτέλεσμα σε σύγκριση με ένα μόνο δέντρο.
Υπάρχει πάντα ένα περιθώριο υπερπροσαρμογής (overfitting), λόγω της παρουσίας διακύμανσης.	Ο αλγόριθμος Random Forest αποφεύγει και αποτρέπει την υπερπροσαρμογή με τη χρήση πολλαπλών δέντρων.

Τα αποτελέσματα δεν είναι ακριβή.	Παρέχει ακριβή και έγκυρα αποτελέσματα.
Τα δέντρα αποφάσεων απαιτούν χαμηλό υπολογισμό, μειώνοντας έτσι τον χρόνο υλοποίησης και μεταφέροντας χαμηλή ακρίβεια.	Ένα τυχαίο δάσος απαιτεί περισσότερους υπολογισμούς. Η διαδικασία παραγωγής και ανάλυσης είναι χρονοβόρα.
Είναι εύκολο να οπτικοποιηθεί. Το μόνο πρόβλημα είναι η προσαρμογή του μοντέλου του δέντρου αποφάσεων.	Δύσκολο στην οπτικοποίηση, καθώς προσδιορίζει το μοτίβο πίσω από τα δεδομένα.

Πίνακας 5.1 Συγκριτικός πίνακας Decision Tree & Random Forest

Πολυπλοκότητα

Με βάση τους τύπους παλινδρόμησης και ταξινόμησης, ένα δέντρο αποφάσεων παράγει μια σειρά αποφάσεων που χρησιμοποιούνται για την εξαγωγή συγκεκριμένων αποτελεσμάτων. Ενώ είναι απλό και εύκολο στην ερμηνεία, η διαδικασία διαχωρισμού των δεδομένων και πρόβλεψης της εξόδου είναι γρήγορη. Από την άλλη πλευρά, στην περίπτωση του αλγορίθμου τυχαίου δάσους, υπάρχουν πολλαπλά στάδια ορισμού των δέντρων και άλλων κρίσιμων μεταβλητών που αυξάνουν άμεσα την πολυπλοκότητα του μοντέλου σε κάθε κόμβο.

Υπερπροσαρμογή

Κατά την εφαρμογή τους, και οι δύο αλγόριθμοι είναι εκτεθειμένοι στην υπερπροσαρμογή, δημιουργώντας μια κατάσταση συμπιεσμένης συμφόρησης κατά την εκπαίδευση των δεδομένων. Ο αντίκτυπος στο νέο μοντέλο δεδομένων υποδεικνύει αρνητική απόδοση όταν το σύνολο δεδομένων αποτυγχάνει στα κριτήρια επικύρωσης. Σε τέτοια σενάρια, ένα δέντρο απόφασης έχει περισσότερες πιθανότητες υπερπροσαρμογής. Αντ' αυτού, ο αλγόριθμος τυχαίου δάσους μπορεί να μειώσει την έκθεσή του με πολλαπλά δέντρα.

Επεξεργασία δεδομένων

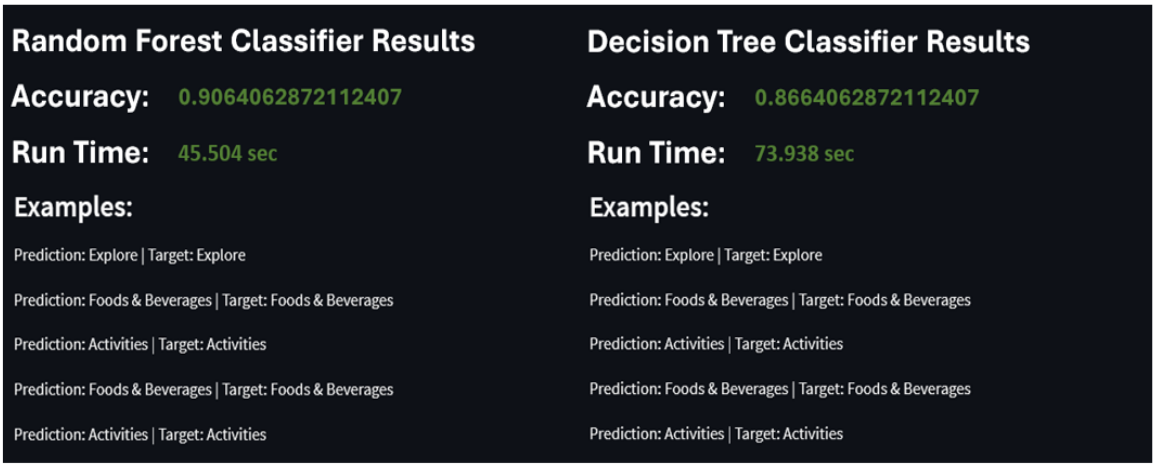
Σε ένα δέντρο αποφάσεων, η βασική αιτία κάθε δήλωσης προβλήματος συμβολίζεται ως κόμβος ρίζας. Φέρει μια σειρά από κόμβους απόφασης που αντιπροσωπεύουν διάφορες αποφάσεις. Από τους κόμβους απόφασης, οι κόμβοι φύλλων δείχνουν τον αντίκτυπο αυτών των αποφάσεων. Αυτοί οι κόμβοι διακλαδίζονται περαιτέρω για την απόκτηση

καλύτερων πληροφοριών και θα συνεχίσουν να το κάνουν μέχρι όλοι οι κόμβοι να έχουν παρόμοια συνεπή δεδομένα.

Ο αλγόριθμος τυχαίου δάσους λειτουργεί με βάση ένα συλλογικό αποτέλεσμα πολλαπλών δέντρων αποφάσεων. Ορισμένα μπορεί να μην δίνουν ένα σωστό απαιτούμενο αποτέλεσμα, αλλά με τη συγχώνευση όλων των δέντρων, ένα συλλογικό αποτέλεσμα μπορεί να είναι ακριβές και να χρησιμοποιηθεί για περαιτέρω στάδια.

5.1.4 Επιλογή του καταλληλότερου αλγορίθμου

Μελετώντας τα δεδομένα που συλλέγονται από την εφαρμογή και εφαρμόζοντας τους δύο παραπάνω αλγορίθμους στα δεδομένα αυτά, προκύπτει ότι ο αλγόριθμος Random Forest (RF) είναι ο καταλληλότερος αλγόριθμος για την διαδικασία εξόρυξης δεδομένων. Πιο συγκεκριμένα ο αλγόριθμος RF εμφανίζει ελαφρώς καλύτερο ποσοστό επιτυχίας από τον αλγόριθμο Decision Tree. Εκτός από το ποσοστό επιτυχίας λαμβάνεται υπόψιν και ο χρόνος εκτέλεσης των αλγορίθμων καθώς ο RF σημειώνει καλύτερους χρόνους εκτέλεσης από τον Decision Tree. Επομένως λαμβάνοντας υπόψιν όλα τα παραπάνω προτερήματα του RF καθώς και τα επιπλέον πλεονεκτήματα που προσφέρει όπως αποφυγή υπερπροσαρμογής (overfitting) των δεδομένων κλπ.. στην υλοποίηση μας θα χρησιμοποιηθεί ο αλγόριθμος RF.



Εικόνα 5.6 Αποτελέσματα σύγκρισης των αλγορίμων Decision Tree & Random Forest

5.2 Δεδομένα ερωτηματολογίων στην εφαρμογή

Σκοπός των ερωτηματολογίων είναι τόσο η παροχή μίας άμεσης ανατροφοδότησης της επιχείρησης για την ποιότητα και την ικανοποίηση των πελατών της από τις υπηρεσίες που παρέχει καθώς και την βελτίωση των υπηρεσιών αυτών.

Τα ερωτηματολόγια έχουν σχεδιαστεί με τέτοιο τρόπο ώστε να περιέχουν όσο το δυνατόν ελάχιστο αριθμό ερωτήσεων ώστε να μην προκαλούν δυσφορία στον χρήστη κατά την συμπλήρωση αλλά παράλληλα να είναι αρκετές ώστε να μπορεί να ληφθεί ανατροφοδότηση τόσο για τις υπηρεσίες που παρέχονται όσο και δεδομένα για τις προτιμήσεις του χρήστη και τις συστάσεις του.

Τα ερωτηματολόγια που περιέχονται στην εφαρμογή είναι τα εξής:

- Hotel Survey
- Restaurant Survey
- Improve your Experience (Recommendation Survey)

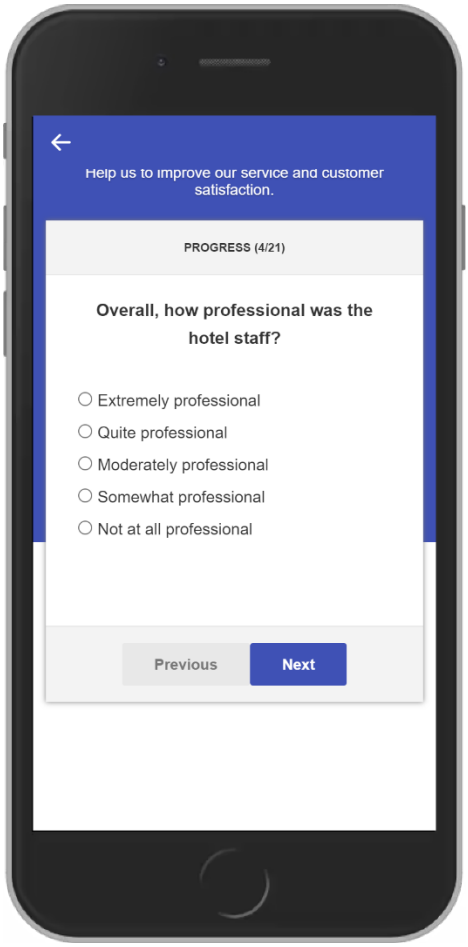
Όλα τα ερωτηματολόγια αποτελούν κομμάτι της εφαρμογής καθώς έχουν δημιουργηθεί προγραμματιστικά χωρίς κάποιο έτοιμο «εξωτερικό» εργαλείο ή υπηρεσία αλλά με την χρήση του framework Angular.

Τα ερωτηματολόγια είναι ανώνυμα καθώς δεν ζητάνε προσωπικές πληροφορίες από τον χρήστη. Η ταυτοποίηση και ο διαχωρισμός τους γίνεται μέσω του μοναδικού αναγνωριστικού κωδικού της συσκευής UUID.

Ένα UUID (συντομογραφία του Universal Unique IDentifier) είναι μία συμβολοσειρά 36 χαρακτήρων που μπορεί να χρησιμοποιηθεί για την αναγνώριση πληροφοριών. Τα UUID χρησιμοποιούνται ευρέως εν μέρει επειδή είναι πολύ πιθανό να είναι μοναδικά σε παγκόσμιο επίπεδο, πράγμα που σημαίνει ότι το UUID δεν είναι μόνο μοναδικό στον πίνακα της βάσης δεδομένων της εφαρμογής μας, αλλά είναι πιθανότατα η μοναδική γραμμή με αυτό το UUID σε οποιοδήποτε σύστημα οπουδήποτε. Τεχνικά δεν είναι αδύνατο το ίδιο UUID που παράγεται να χρησιμοποιηθεί κάπου αλλού, αλλά οι πιθανότητες να συμβεί είναι πολύ μικρές. (Montgomery et al., 2004)

5.2.1 Ερωτηματολόγιο Hotel (Hotel Survey)

Το ερωτηματολόγιο αυτό αφορά το πόσο ικανοποιημένος είναι ο πελάτης από την διαμονή του στο ξενοδοχείο. Αποτελείται από είκοσι ερωτήσεις τις οποίες ο χρήστης θα πρέπει να τις συμπληρώσει. Κάθε ερώτηση περιέχει από έξι επιλογές εκτός από την τελευταία ερώτηση η οποία είναι ανοικτού τύπου και επιτρέπει στον χρήστη να γράψει το δικό του κείμενο. Δίνεται η δυνατότητα στον χρήστη της σταδιακής συμπλήρωσης και αποθήκευσης των επιλογών του. Για παράδειγμα ο χρήστης συμπληρώνει το ερωτηματολόγιο μέχρι την ερώτηση πέντε και για κάποιο λόγο χρειάζεται να κλείσει την εφαρμογή, αν ξανά ανοίξει την εφαρμογή και επιλέξει να συμπληρώσει το ίδιο ερωτηματολόγιο τότε θα συνεχίσει από την ερώτηση που είχε μείνει χωρίς να χρειάζεται να ξανά συμπληρώσει τις προηγούμενες ερωτήσεις.

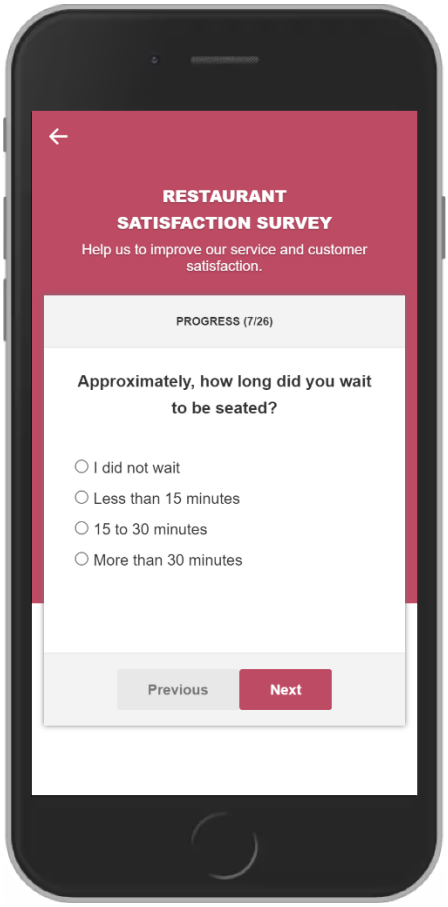


Εικόνα 5.7 Παράδειγμα του ερωτηματολογίου που αφοράνε την αξιολόγηση του ξενοδοχείου.

Το σύνολο των ερωτημάτων του ερωτηματολογίου «Hotel Survey» παρατίθεται στο Παράρτημα Α

5.2.2 Ερωτηματολόγιο Restaurant (Restaurant Survey)

Το ερωτηματολόγιο αυτό αφορά το πόσο ικανοποιημένος είναι ο πελάτης από τα τρία εστιατόρια που διαθέτει το ξενοδοχείο. Οι ερωτήσεις αφορούν τον χρόνο αναμονής για την παράδοση του φαγητού, την ποιότητα του φαγητού και των εγκαταστάσεων καθώς και τον τρόπο λειτουργίας των εργαζομένων. Αποτελείται από είκοσι πέντε ερωτήσεις τις οποίες ο χρήστης θα πρέπει να τις συμπληρώσει όλες. Κάθε ερώτηση περιέχει από έξι επιλογές εκτός από την τελευταία ερώτηση η οποία είναι ανοικτού τύπου και επιτρέπει στον χρήστη να γράψει το δικό του κείμενο. Επίσης και σε αυτό το ερωτηματολόγιο δίνεται η δυνατότητα στον χρήστη της σταδιακής συμπλήρωσης και αποθήκευσης των επιλογών του σε περίπτωση διακοπής της χρήσης της εφαρμογής και συνέχειας από την ερώτηση που είχε μείνει.

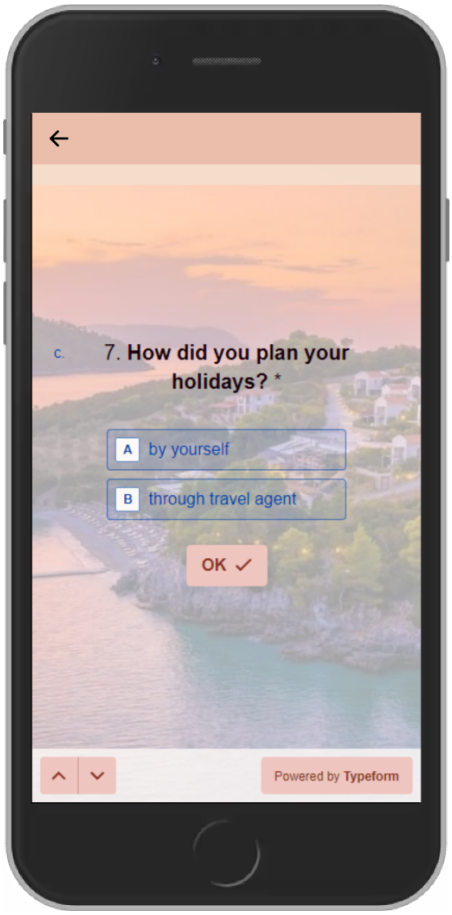


Εικόνα 5.8 Παράδειγμα του ερωτηματολογίου που αφοράνε την αξιολόγηση των εστιατορίων του ξενοδοχείου.

Το σύνολο των ερωτημάτων που αφορούν το restaurant του ξενοδοχείου παρατίθεται στο Παράρτημα Β.

5.2.3 Ερωτηματολόγιο Recommendation (Recommendation Survey)

Αυτό το ερωτηματολόγιο περιέχει ερωτήσεις που αφορούν τον χρήστη σε προσωπικό επίπεδο δηλαδή ηλικία, οικογενειακή κατάσταση καθώς και προτιμήσεις για τον τρόπο διασκέδασης, για τις δραστηριότητες που του αρέσουν. Όλες αυτές οι ερωτήσεις συμβάλουν στην δημιουργία ενός προφίλ του χρήστη με τις δραστηριότητες και τα ενδιαφέροντα του τα οποία στην συνέχεια θα ληφθούν για την δημιουργία των κατάλληλων συστάσεων προς τον χρήστη εκάστοτε χρήστη.



Εικόνα 5.9 Παράδειγμα του ερωτηματολογίου που αφοράνε τις προτιμήσεις των πελατών του ξενοδοχείου.

Το σύνολο των ερωτημάτων που αφορούν το Recommendation Survey του ξενοδοχείου παρατίθεται στο Παράρτημα Γ.

5.3 Δεδομένα καταγραφής δραστηριοτήτων στην εφαρμογή (clickstreams)

Η εφαρμογή εκτός από τα ερωτηματολόγια για την συλλογή δεδομένων διαθέτει και την δυνατότητα καταγραφής ενεργειών – δραστηριοτήτων του χρήστη. Το χαρακτηριστικό

αυτό συμβάλει στην συλλογή δεδομένων από τον χρήστη για την δημιουργία του αντιστοίχου προφίλ του. Μέσω των προφίλ και των δεδομένων δραστηριοτήτων μπορούμε να εξάγουμε πληροφορίες για τις προτιμήσεις του χρήστη, τις τάσεις που προτιμάνε οι περισσότεροι χρήστες κλπ. Αναλύοντας αυτά τα δεδομένα και τα μοτίβα που προκύπτουν μπορούμε να προβούμε στις αντίστοιχες συστάσεις ανάλογα με τα ενδιαφέροντα του κάθε χρήστη.

Η καταγραφή δραστηριοτήτων πραγματοποιείται σε κάθε βασική λειτουργία που κάνει ο χρήστης:

Ορισμένες από αυτές τις βασικές λειτουργίας είναι:

- Πλοήγηση στις σελίδες της εφαρμογής.
- Πραγματοποίηση αναζήτησης στις λίστες τις εφαρμογής.
- Επιλογή σημείου στο χάρτη.
- Σάρωση AR σημείου.
- Επιλογή ενεργειών όπως:
    - Επίσκεψης εξωτερικού διαδικτυακού συνδέσμου (π.χ. Facebook, Instagram, Twitter κλπ.).
    - Επικοινωνία μέσω email ή τηλεφώνου με το ξενοδοχείο και τους εξωτερικούς συνεργάτες.

Τα δεδομένα καταγραφής δημιουργούνται κάθε φορά που πραγματοποιούνται μία από τις παραπάνω ενέργειες και αποθηκεύονται στο local storage. Το local storage είναι μία ιδιότητα που επιτρέπει σε ιστοτόπους και εφαρμογές JavaScript να αποθηκεύουν ζεύγη κλειδιών – τιμών σε ένα πρόγραμμα περιήγησης ιστού χωρίς ημερομηνία λήξης. Αυτό σημαίνει ότι τα δεδομένα που αποθηκεύονται στο πρόγραμμα περιήγησης θα παραμείνουν ακόμη και μετά το κλείσιμο του παραθύρου του προγράμματος περιήγησης. Τα δεδομένα καταγραφής αποθηκεύονται σε έναν πίνακα στο local storage με την ονομασία clickstreams και αποστέλνονται στον διακομιστή (server) κάθε 10 λεπτά. Όλα τα δεδομένα που συλλέγονται στέλνονται στην διακομιστή ο οποίος τα αποθηκεύει στην βάση δεδομένων σε ένα collection με την ονομασία streams. Μόλις γίνει η αποστολή των δεδομένων στον διακομιστή γίνεται εκκαθάριση του local storage προκειμένου να συλλεχθούν νέα δεδομένα καταγραφής. Όλα τα δεδομένα καταγραφής δημιουργούνται από την συνάρτηση clickStreamPush Εικόνα 5.6. Η συνάρτηση αυτή καλείται σε κάθε

βασική ενέργεια του χρήστη και λαμβάνει βασικές πληροφορίες προκειμένου να δημιουργήσει το αντικείμενο και να το αποθηκεύσει στο local storage.

Εκτός από τις συστάσεις τα δεδομένα που προκύπτουν από την καταγραφή των κινήσεων μπορούν να χρησιμοποιηθούν για την εξαγωγή και άλλον χρήσιμων πληροφοριών που μπορεί να συμβάλλουν θετικά στην εξέλιξη των υπηρεσιών του ξενοδοχείου. Όπως για παράδειγμα οι χρόνοι παραμονής των χρηστών της εφαρμογής. Πιο συγκεκριμένα διατηρώντας τους χρόνους δημιουργίας της κάθε ενέργειας του χρήστη μπορούμε να υπολογίσουμε την διάρκεια παραμονής του στην κάθε σελίδα. Παρατηρώντας τους χρόνους παραμονής η επιχείρηση θα μπορεί να δει ποιες σελίδες παρουσιάζουν μεγαλύτερο ενδιαφέρον και ποιες όχι καθώς επίσης και την βελτίωση και αλλαγή των σελίδων αυτών.

Ένα δεδομένο καταγραφής αναπαρίσταται με την μορφή ενός αντικειμένου το οποίο περιέχει εννέα χαρακτηριστικά ή αλλιώς εννέα ζευγάρια κλειδιών – τιμών.

```
1 usage  George Pachoulas
clickStreamSend() : void  {
  if (localStorage.getItem( key: 'clickStreams')) {
    const data: any = {};
    data.stream = JSON.parse(String(localStorage.getItem( key: 'clickStreams')));
    if (this.connectionStatus === 'online') {
      if (data.stream.length > 0) {
        this.clickStreamService.sendClickStream(data)
          .pipe(takeUntil(this.destroyed))
          .subscribe( next: (res: any) : void  => {
            // console.log('response is: ', res);
            localStorage.removeItem( key: 'clickStreams');
          }, error: error => {
            console.log(error);
          });
      }
    }
  }
}
```

Εικόνα 5.10 Κώδικας αποστολής δεδομένων καταγραφής στον διακομιστή.

Τα χαρακτηριστικά αυτά είναι τα εξής:

- **Action:** Καταγράφει ενέργειας που αφοράνε τα σημεία του χάρτη όπως την επιλογή ενός σημείου ή την χρήση κάποιου συνδέσμου μέσα από τις πληροφορίες του σημείου. Αποτελεί ένα προαιρετικό χαρακτηριστικό και σε περίπτωση που δεν

καταγράφεται κάποιο δεδομένο αυτού του τύπου παραμένει ως ένα κενό αλφαριθμητικό (string).

- **Current:** Καταγράφει την τρέχουσα σελίδα στην οποία βρίσκεται ο χρήστης. Αποτελεί ένα απαραίτητο χαρακτηριστικό και δεν μπορεί να είναι κενό.
- **DetailsID:** Καταγράφεται κατά την είσοδο σε κάποιο από τις σελίδες λεπτομερειών (φαγητών, δραστηριοτήτων, παραλιών) ή στην επιλογή ενός σημείου στο χάρτη. Αποτελεί ένα προαιρετικό χαρακτηριστικό και σε περίπτωση που δεν καταγράφεται κάποιο δεδομένο αυτού του τύπου παραμένει ως ένα κενό αλφαριθμητικό (string).
- **FbLogin:** Καταγράφει τα δεδομένα από το λογαριασμό Facebook του χρήστη σε περίπτωση που συνδεθεί μέσω αυτής της πλατφόρμας. Αποτελεί ένα προαιρετικό χαρακτηριστικό και σε περίπτωση που δεν καταγράφεται κάποιο δεδομένο αυτού του τύπου παραμένει ως ένα κενό αλφαριθμητικό (string).
- **Filter:** Καταγράφετε κατά την χρήση των φίλτρων στις σελίδες των λιστών (φαγητών, δραστηριοτήτων, παραλιών) ή στην σελίδα του χάρτη. Αποτελεί ένα προαιρετικό χαρακτηριστικό και σε περίπτωση που δεν καταγράφεται κάποιο δεδομένο αυτού του τύπου παραμένει ως ένα κενό αλφαριθμητικό (string).
- **Previous:** Καταγράφει την σελίδα από την οποία έρχεται ο χρήστης (προηγούμενη σελίδα). Αποτελεί ένα απαραίτητο χαρακτηριστικό και δεν μπορεί να είναι κενό εκτός από την εκκίνηση της εφαρμογής κατά την οποία δεν υπάρχει προηγούμενη σελίδα.
- **Social:** Καταγράφει δεδομένα κατά την επίσκεψη εξωτερικών διαδικτυακών συνδέσμων και κατά την επικοινωνία (τηλέφωνο και email). Αποτελεί ένα προαιρετικό χαρακτηριστικό και σε περίπτωση που δεν καταγράφεται κάποιο δεδομένο αυτού του τύπου παραμένει ως ένα κενό αλφαριθμητικό (string).
- **Timestamp:** Καταγράφει την χρονική στιγμή που γίνεται η ενέργεια σε Unix time. Αποτελεί ένα απαραίτητο χαρακτηριστικό και δεν μπορεί να είναι κενό.
- **Uuid:** Περιέχει το αναγνωριστικό Uuid του χρήστη το οποίο δημιουργείται αυτόματα με την εκκίνηση της εφαρμογής σε περίπτωση που δεν υπάρχει. Αποτελεί ένα απαραίτητο χαρακτηριστικό και δεν μπορεί να είναι κενό.

Εικόνα 5.11 Παράδειγμα ενός αντικειμένου clickstream που δημιουργείτε κατά την αλληλεπίδραση του χρήστη με την εφαρμογή

```
no usages  George Pachoulas
clickStreamPush(social: any, filters: any, action: any) : void  {
  if (localStorage.getItem( key: 'uuid')) {
    this.uuid = localStorage.getItem( key: 'uuid')
    // console.log(this.uuid);
  }
  let details : string  = ''
  if (this.currentPage.includes('food-beverage-details')) {
    details = this.currentPage.split( separator: '/')[4]
  } else if (this.currentPage.includes('activities-details')) {
    details = this.currentPage.split( separator: '/')[3]
  } else if (this.currentPage.includes('beaches-details')) {
    details = this.currentPage.split( separator: '/')[3]
  }
  if (action.detailsID) {
    details = action.detailsID;
    delete action.detailsID;
  }
  let login: any = ''
  if (this.authService.currentSocialUser) {
    login = this.authService.currentSocialUser.response;
    login.authToken = this.authService.currentSocialUser.authToken;
  }
  const data : {uid: any, timestamp: number, ...  = {
    uid: this.uuid,
    timestamp: moment().unix(),
    previous: this.previousPage,
    current: this.currentPage,
    detailsID: details,
    social: social,
    filter: filters,
    action: action,
    fbLogin: login
  };
  if (data.uid !== null) {
    // console.log(data.uid);
    if (!localStorage.getItem( key: 'clickStreams')) {
      localStorage.setItem('clickStreams', JSON.stringify( value: [data]))
    } else {
      const streams = JSON.parse(String(localStorage.getItem( key: 'clickStreams')));
      streams.push(data);
      localStorage.setItem('clickStreams', JSON.stringify(streams))
    }
  }
}
```

Εικόνα 5.12 Συνάρτηση για την δημιουργία των αντικειμένων clickstream και αποθήκευση των δεδομένων στο local storage

5.3.1 Παραδείγματα χρήσης clickstreams στην υλοποίηση

5.3.1.1 Clickstream στα φίλτρα αναζήτησης

Όλα οι επιλογές που γίνονται στα φίλτρα τόσο στις σελίδες με τις λίστες φαγητών, δραστηριοτήτων και παραλιών καθώς επίσης και στον χάρτη καταγράφονται στο χαρακτηριστικό filter. Βασικός σκοπός συλλογής αυτών των δεδομένων είναι η δημιουργία του προφίλ χρήστη.

Το χαρακτηριστικό filter αποθηκεύει ένα αντικείμενο της παρακάτω μορφής:

- **Category:** Καταγράφει την κατηγορία που έχει επιλεχθεί από τα φίλτρα τόσο στις λίστες όσο και στον χάρτη.
- **Page:** Καταγράφει την σελίδα στην οποία έγινε η ενέργεια.
- **Type:** Καταγράφει τον τύπο που έχει επιλεχθεί από τα φίλτρα (σε όσα διαθέτουν).

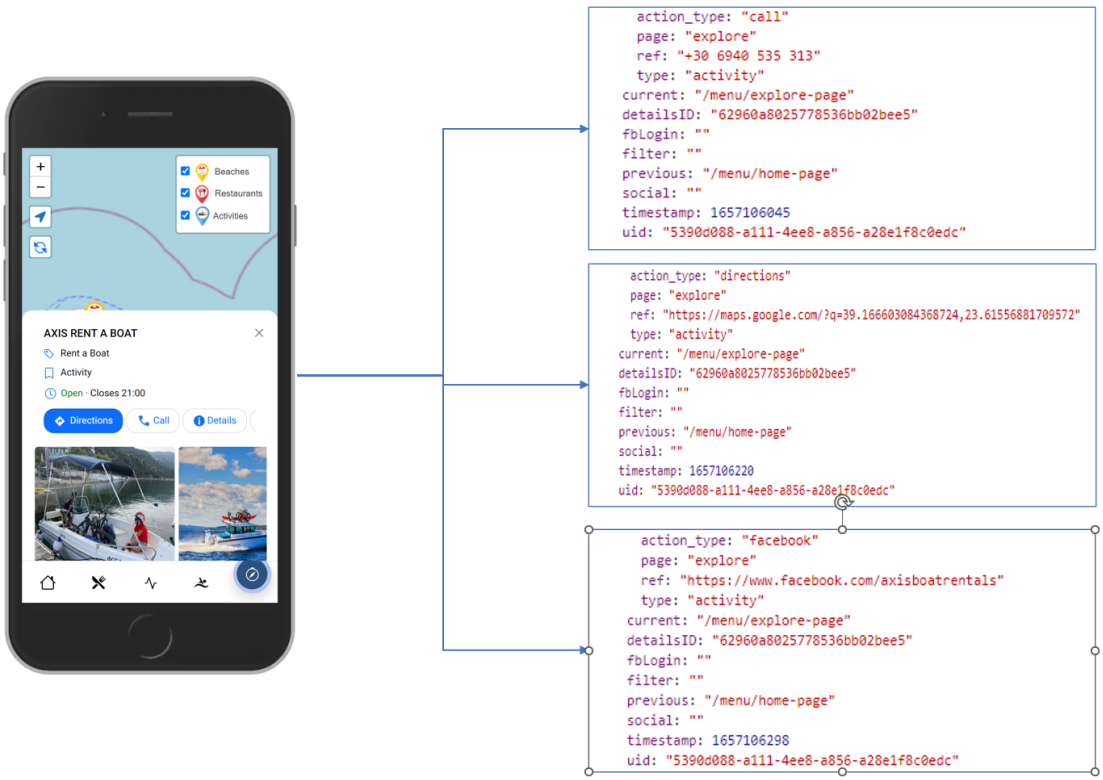
Εικόνα 5.13 Παράδειγμα ενέργειας του χρήστη και συγκεκριμένα εφαρμογής ενός φίλτρου αναζήτησης στην λίστα των παραλιών (αριστερό τμήμα) και δημιουργία του αντίστοιχου αντικειμένου clickstream (δεξιό τμήμα)

5.3.1.2 Clickstream στην σελίδα του χάρτη

Όλα οι κινήσεις και οι ενέργειες που καταγράφονται στην σελίδα Explore της εφαρμογής η οποία περιέχει και τον χάρτη με τα σημεία ενδιαφέροντος καταγράφονται στο χαρακτηριστικό action. Καταγράφονται τόσο η επιλογή ενός σημείου στο χάρτη όσο και η επιλογή κάποιες από τις διαθέσιμες ενέργειες του σημείου.

Το χαρακτηριστικό action αποθηκεύει ένα αντικείμενο της παρακάτω μορφής:

- **Action_type:** Καταγράφει τον τύπο της ενέργειας που έγινε με τα εξής:
  - **BottomSheet:** Στην περίπτωση της επιλογής και μόνο ενός σημείου στο χάρτη.
  - **Call:** Στην περίπτωση της επιλογής "Call".
  - **Facebook:** Στην περίπτωση της επιλογής "Facebook".
  - **Website:** Στην περίπτωση της επιλογής "Website".
- **Page:** Καταγράφει την σελίδα στην οποία έγινε η ενέργεια (χρήσιμο για την καταγραφή δραστηριοτήτων με την ανίχνευση AR).
- **Ref:** Περιέχει σε περίπτωση επιλογής κάποιας ενέργειας στο σημείο τον εξωτερικό σύνδεσμο, τηλέφωνο επικοινωνίας ή email.
- **Type:** Καταγράφει τον τύπο του σημείου ενδιαφέροντος που άνοιξε ή του αντικειμένου AR που ανιχνεύθηκε (π.χ. beach, activity, restaurant, wine, food, cocktail).



Εικόνα 5.14 Παράδειγμα δημιουργίας αντικειμένων clickstream στην σελίδα σημείων ενδιαφερόντων της εφαρμογής

5.3.1.3 Clickstream στα social της εφαρμογής

Όλες οι ενέργειες που γίνονται κατά την επιλογή των εξωτερικών συνδέσμων του ξενοδοχείου καταγράφονται στο χαρακτηριστικό social. Μέσα από την συλλογή αυτών των δεδομένων η επιχείρηση θα μπορέσει να έχει μια εικόνα της χρήσης των μέσων κοινωνικής δικτύωσης όπως ποια μέσα χρησιμοποιούν οι χρήστες ποια δεν έχουν μεγάλη απήχηση καθώς να βελτιώσουν τα μέσα αυτά ώστε να τραβήξουν το ενδιαφέρον των χρηστών.

Το χαρακτηριστικό social αποθηκεύει ένα αντικείμενο της παρακάτω μορφή:

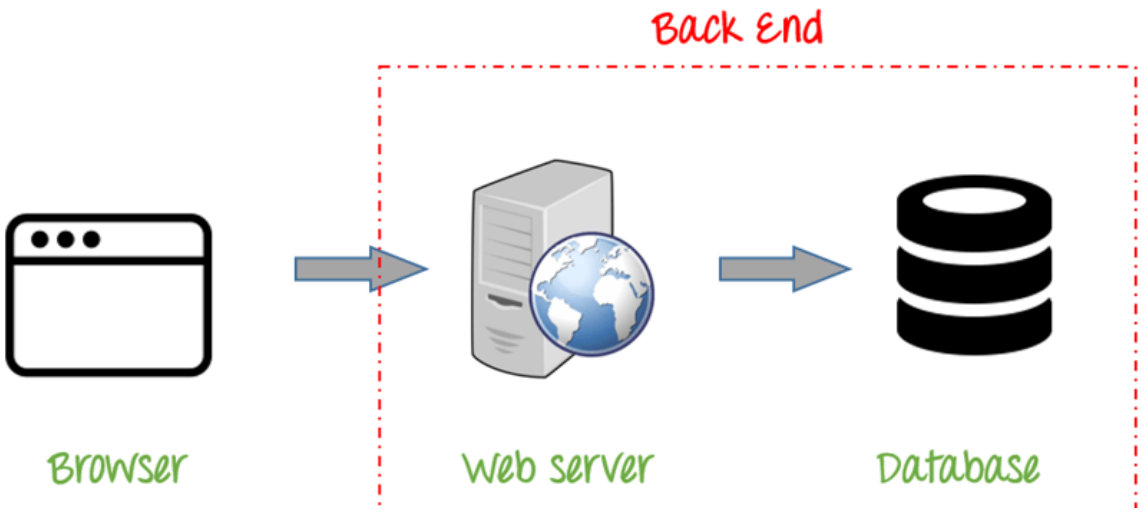
- **Type:** Καταγράφει τον τύπο του συνδέσμου που επιλέχθηκε από τα εξής:
  - Facebook
  - Twitter
  - Instagram
  - YouTube
  - TripAdvisor
  - LinkedIn
  - WhatsApp
  - Email
  - Telephone
- **URL:** Περιέχει τον σύνδεσμο ως προς την ιστοσελίδα (στην περίπτωση των email και telephone περιέχει την διεύθυνση ηλεκτρονικού ταχυδρομείου και τηλέφωνο αντίστοιχα).

Εικόνα 5.15 Παράδειγμα δημιουργίας αντικειμένων clickstream μέσω της χρήσης των συνδέσμων σε υπηρεσίες κοινωνικής δικτύωσης

6 Τεχνολογίες υλοποίησης

6.1 Τεχνολογίες από την πλευρά του διακομιστή (Backend)

Το back – end αναφέρεται στην ανάπτυξη από την πλευρά του εξυπηρετητή. Συμπεριλαμβάνει βάσεις δεδομένων, σενάρια (scripts), αρχιτεκτονική ιστοσελίδων. Περιλαμβάνει παρασκηνιακές δραστηριότητες που συμβαίνουν κατά την εκτέλεση οποιασδήποτε ενέργειας σε έναν ιστότοπο. Το backend επικοινωνεί με το frontend, στέλνοντας και λαμβάνοντας πληροφορίες που θα εμφανίζονται ως ιστοσελίδα. Ο κώδικας που γράφετε από προγραμματιστές back – end βοηθά τα προγράμματα περιήγησης να επικοινωνούν με πληροφορίες βάσης δεδομένων. Για παράδειγμα κάθε φορά που ο χρήστης συμπληρώνει μια φόρμα επικοινωνίας, ή πραγματοποιεί μια αγορά και γενικά οποιαδήποτε αλληλεπίδραση χρήστη από την πλευρά του πελάτη το πρόγραμμα περιήγησης στέλνει ένα αίτημα στον διακομιστή, το οποίο επιστρέφει πληροφορίες με τη μορφή κώδικα διεπαφής που το πρόγραμμα περιήγησης στην συνέχεια «μεταγλωττίζει» και τις εμφανίζει στην διεπαφή χρήστη.



Εικόνα 6.1 Παράδειγμα διαχωρισμού του τμήματος back – end από το front - end

6.1.1 Node.js

Το Node.js είναι ένα ανοιχτού κώδικα, περιβάλλον εκτέλεσης για τον προγραμματισμό εφαρμογών από την πλευρά του διακομιστή και είναι χτισμένο σε περιβάλλον Javascript. Το Node.js παρέχει μία αρχιτεκτονική βάσει συμβάντων και μία non-blocking I/O API σχεδιασμένη να βελτιστοποιεί την απόδοση και κλιμάκωση μιας εφαρμογής. Σε αντίθεση με τα περισσότερα σύγχρονα περιβάλλοντα ανάπτυξης

εφαρμογών δικτύων μία διεργασία node στηρίζεται σε ένα μοντέλο ασύγχρονης επικοινωνίας εισόδου/εξόδου και περιέχει μία ενσωματωμένη βιβλιοθήκη για να επιτρέπει στις εφαρμογές να λειτουργούν ως ένας αυτόνομος διαδικτυακός διακομιστής.Κάποιοι από τους χρήστες του λογισμικού NodeJS είναι το Linkedin, Microsoft, Netflix, Paypal και Yahoo.

Το Node.js αναπτύχθηκε το 2009 από τον Ryan Dahl και άλλους προγραμματιστές που εργάζονταν στη Joyent (The History of Node.js, 2020). Το Node.js δημιουργήθηκε και εκδόθηκε πρώτη φορά για χρήση των Linux το 2009. Η ανάπτυξη και διατήρησή του καθοδηγήθηκαν από τον Dahl και χρηματοδοτήθηκαν από την Joyent, την εταιρεία που εργαζόταν ο Dahl. Ο Dahl εμπνεύστηκε τη δημιουργία του node από την ανάγκη του να ενημερώσει τον χρήστη σε πραγματικό χρόνο για την κατάσταση ενός αρχείου που ανεβάζει στο διαδίκτυο.

Το Node.js επιτρέπει τη δημιουργία διαδικτυακών διακομιστών και εργαλείων δικτύωσης που χρησιμοποιούν την JavaScript και μία συλλογή από modules που χειρίζονται διάφορες βασικές λειτουργίες. Τα modules χειρίζονται το σύστημα αρχείων εισόδου/εξόδου, τη δικτύωση (HTTP, TCP, UDP, DNS, ή TLS/SSL), δυαδικά δεδομένα (ενδιάμεσες μνήμες), λειτουργίες κρυπτογράφησης, ροές δεδομένων και άλλες βασικές λειτουργίες. Τα modules του Node χρησιμοποιούν ένα API που είναι σχεδιασμένο να μειώνει την πολυπλοκότητα της συγγραφής εφαρμογών διακομιστών. Τα πλαίσια μπορούν να χρησιμοποιηθούν για να επιταχύνουν την ανάπτυξη εφαρμογών και κοινών πλαισίων όπως τα Express.js, Socket.IO και Connect.

Το Node.js αρχικά χρησιμοποιήθηκε για την κατασκευή προγραμμάτων δικτύου, όπως διαδικτυακών διακομιστών, κάτι που το καθιστούσε παρόμοιο με την PHP και την Python. Η μεγαλύτερη διαφορά μεταξύ της PHP και του Node.js είναι ότι η PHP είναι μια blocking γλώσσα, όπου οι εντολές εκτελούνται αφού πρώτα έχει ολοκληρωθεί η προηγούμενη εντολή, ενώ το Node.js είναι μια non-blocking γλώσσα όπου οι εντολές εκτελούνται παράλληλα, και χρησιμοποιεί ανακλήσεις για να σηματοδοτήσει κάποια ολοκλήρωση.

Εικόνα 6.2 Το λογότυπο του Node.js

Το 2011, ένας διαχειριστής πακέτων εισήχθη στη βιβλιοθήκη Node.js, με το όνομα npm. Ο διαχειριστής πακέτων επιτρέπει τη δημοσίευση και τον διαμοιρασμό των ανοιχτού κώδικα βιβλιοθηκών Node.js από την κοινότητα, και είναι σχεδιασμένος να απλοποιεί την εγκατάσταση, αναβάθμιση και απεγκατάσταση των βιβλιοθηκών. Το πρώτο Node.js που κατασκευάστηκε για να υποστηρίζει τα Windows κυκλοφόρησε τον Ιούλιο του 2011.

Χαρακτηριστικά του Node.js

Παρακάτω ακολουθούν μερικά από τα σημαντικά χαρακτηριστικά που καθιστούν το Node.js την πρώτη επιλογή για τις αρχιτεκτονικές διακομιστή (Effendy et al., 2021):

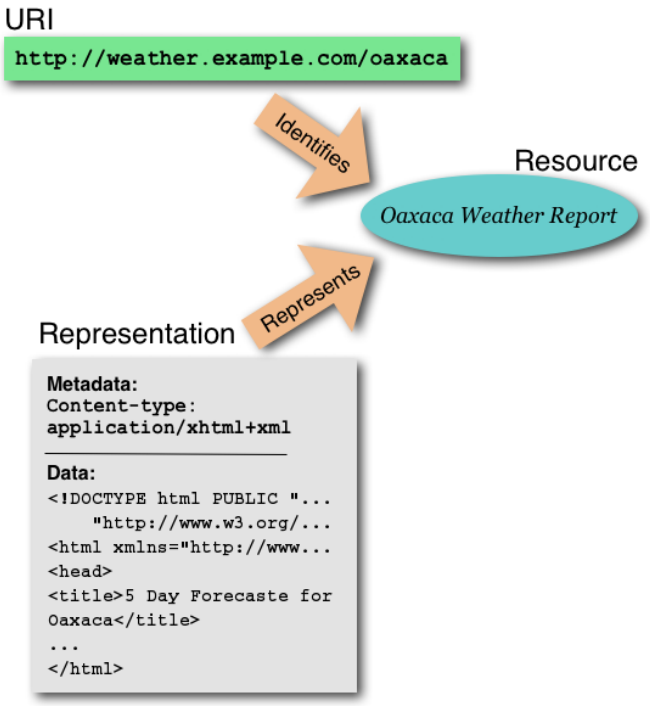
- **Ασύγχρονα και οδηγούμενα από συμβάντα:** Όλα τα API της βιβλιοθήκης Node.js είναι ασύγχρονα, δηλαδή ο διακομιστής δεν περιμένει το API να επιστρέψει δεδομένα αλλά μεταβαίνει στο επόμενο API που έχει κληθεί και μέσω ενός μηχανισμού ειδοποίησης συμβάντων λαμβάνει την απάντηση από την προηγούμενη κλήση API.
- **Γρήγορη χρόνοι απόκρισης:** Η βιβλιοθήκη Node.js βασίζεται στη μηχανή V8 JavaScript Engine του Google Chrome η οποία είναι πολύ γρήγορη στην εκτέλεση του κώδικα.
- **Χρήση ενός thread με δυνατότητες επεκτασιμότητας:** Το Node.js χρησιμοποιεί ένα thread από την επεξεργαστική ισχύ με βρόχο συμβάντων το οποίο του επιτρέπει την διαχείριση πολλαπλών συμβάντων κάτι που καθιστά τον διακομιστή εξαιρετικά επεκτάσιμο σε αντίθεση με τους παραδοσιακούς

διακομιστές που δημιουργούν περιορισμένα νήματα για τη διαχείριση των αιτημάτων.

6.1.2 REST API

Το πρωτόκολλο REST εμφανίστηκε τα τελευταία χρόνια και έχει καταφέρει να εδραιωθεί σαν ένα μοντέλο σχεδιασμού υπηρεσιών μέσω διαδικτύου λόγω του ότι είναι απλό στη χρήση του. Ορίζει ένα είδος αρχιτεκτονικής μέσω της οποία μπορούν να σχεδιαστούν Web Services. Οι υπηρεσίες αυτές εστιάζουν στους πόρους του συστήματος, αλλά και πιο συγκεκριμένα πως οι πόροι αυτοί χαρακτηρίζονται και μεταφέρονται διαμέσου του πρωτοκόλλου HTTP από έναν ευρύ αριθμό χρηστών. Οι χρήστες μπορεί με τη σειρά τους να είναι άλλα Web Services πιθανώς γραμμένα και σε διαφορετική γλώσσα. Αρχικά, το REST πρωτοεμφανίστηκε το 2000 από τον Roy Fielding στα πλαίσια της διδακτορικής του εργασίας, με τίτλο «Architectural Styles and the Design of Network-based Software Architectures». Η εργασία αυτή παρουσίαζε ένα σύνολο από αρχιτεκτονικές αρχές λογισμικού, οι οποίες χρησιμοποιούσαν τον Ιστό σαν πλατφόρμα των κατανεμημένων υπολογιστικών συστημάτων. Έδωσε απάντηση στην ανάγκη του IETF (Internet Engineering Task Force) για την ύπαρξη ενός μοντέλου για πως θα έπρεπε να λειτουργεί ο Ιστός (Κοροβέσης and Μανώλης, 2016).

Η βασική ιδέα σε κάθε REST API είναι οι πόροι, που αναγνωρίζονται χάρη στα URI και μπορεί να αναπαριστούν web pages. Ένας πόρος είναι ένα αντικείμενο που αποτελείται από έναν τύπο, συσχετιζόμενα δεδομένα, σχέσεις με άλλους πόρους, καθώς και μια σειρά από μεθόδους που λειτουργούν πάνω σε αυτό. Είναι παρόμοιο με ένα αντικείμενο σε μία αντικειμενοστραφή γλώσσα προγραμματισμού, με τη διαφορά ότι μόνο μερικές πρότυπες μέθοδοι ορίζονται για τον πόρο (πρότυπο HTTP : GET, POST, PUT και DELETE), ενώ ένα αντικείμενο συνήθως έχει πολλές μεθόδους HTTP (Κοροβέσης and Μανώλης, 2016).



Εικόνα 6.3 Παράδειγμα βασικών οντοτήτων που σχετίζονται με το REST API.

Οι πόροι μπορούν να ομαδοποιηθούν σε συλλογές - σορούς. Κάθε συλλογή είναι ομοιογενής, δεν έχει συγκεκριμένη σειρά και περιέχει μόνο έναν τύπο πόρων. Πόροι μπορεί επίσης να υπάρχουν εκτός της συλλογής. Όλα τα αντικείμενα και οι πόροι του συστήματος είναι προσβάσιμα μέσω ενός μοναδικού αναγνωριστικού. Στο REST αυτό επιτυγχάνεται χρησιμοποιώντας URI. Όταν πραγματοποιείται ένα αίτημα HTTP σε ένα πρόγραμμα περιήγησης, πρέπει να περιέχει το URI του αντικειμένου που ζητείται ή τοποθετείται στον διακομιστή (Κοροβέσης and Μανώλης, 2016).

Πλεονεκτήματα του REST API

Η αρχιτεκτονική REST παρουσιάζει τα ακόλουθα πλεονεκτήματα:

- Τα στοιχεία είναι απλοποιημένα και επεκτάσιμα επειδή δεν υπάρχει ανάγκη διατήρησης της κατάστασης της εφαρμογής..
- Οι αιτήσεις μπορούν να διεκπεραιωθούν παράλληλα.
- Οι αιτήσεις μπορούν να γίνουν κατανοητές αν απομονωθούν, με αποτέλεσμα μια απλοποιημένη οργάνωση και δυναμική αναδιοργάνωση των υπηρεσιών.
- Επιτρέπει την αποτελεσματική χρήση της προσωρινής μνήμης HTTP και του διακομιστή μεσολάβησης για τη διαχείριση υψηλότερων φόρτων εργασίας.

6.1.3 Python

Η Python είναι μια διερμηνευμένη, αντικειμενοστραφής υψηλού επιπέδου γλώσσα προγραμματισμού με δυναμική σημασιολογία, με κομψό συντακτικό που επιτρέπει στους προγραμματιστές να επικεντρώνονται περισσότερο στην επίλυση προβλημάτων παρά στα συντακτικά λάθη. Σχεδιασμένη για να είναι εύκολη αλλά και διασκεδαστική, το όνομα "Python" είναι μια αναφορά στη βρετανική κωμική ομάδα Monty Python. Η Python έχει τη φήμη μιας γλώσσας φιλικής προς τους αρχάριους, αντικαθιστώντας την Java ως την πιο ευρέως χρησιμοποιούμενη εισαγωγική γλώσσα, επειδή χειρίζεται μεγάλο μέρος της πολυπλοκότητας για τον χρήστη, επιτρέποντας στους αρχάριους να επικεντρωθούν στην πλήρη κατανόηση των εννοιών του προγραμματισμού και όχι στις μικρολεπτομέρειες.

Δημιουργήθηκε από τον Ολλανδό Γκίντο βαν Ρόσσουμ (Guido van Rossum) το 1989 στο ερευνητικό κέντρο Centrum Wiskunde & Informatica (CWI) και κυκλοφόρησε για πρώτη φορά το 1991. Βασική έμπνευση για τον Γκίντο βαν Ρόσσουμ ήταν η γλώσσα προγραμματισμού ABC. Η Python 2.0 κυκλοφόρησε το 2000 και το 2008 κυκλοφόρησε η έκδοση 3.0. Η Python 3.0 είναι η πρώτη γλώσσα προγραμματισμού που σπάει την προς τα πίσω συμβατότητα με προηγούμενες εκδόσεις, για να αντιμετωπιστούν κάποια λάθη που υπήρχαν σε παλαιότερες εκδόσεις και να γίνει ακόμα πιο απλός και σαφής η συγγραφή κώδικα σε αυτήν.

Η Python χρησιμοποιείται για την ανάπτυξη ιστοσελίδων από την πλευρά του διακομιστή, την ανάπτυξη λογισμικού, τα μαθηματικά, την δημιουργία πηγαίου κώδικα για το σύστημα και είναι δημοφιλής για την ταχεία ανάπτυξη εφαρμογών καθώς και για την δυνατότητα της να συνδέει υφιστάμενα στοιχεία λόγω των υψηλού επιπέδου ενσωματωμένων δομών δεδομένων, της δυναμικής τυποποίησης και της δυναμικής δέσμευσης. Το κόστος συντήρησης των προγραμμάτων μειώνεται με την Python λόγω της εύκολα εκμάθησης του συντακτικού και της έμφασης στην αναγνωσιμότητα. Επιπλέον, η υποστήριξη της Python για ενότητες και πακέτα διευκολύνει τα αρθρωτά προγράμματα και την επαναχρησιμοποίηση του κώδικα. Η Python είναι μια γλώσσα ανοικτού κώδικα της κοινότητας, οπότε πολλοί ανεξάρτητοι προγραμματιστές δημιουργούν συνεχώς βιβλιοθήκες και λειτουργίες για αυτήν.

Εικόνα 6.4 Το logo της γλώσσας προγραμματισμού Python

Χαρακτηριστικά της Python

Ορισμένα από τα χαρακτηριστικά της Python είναι (Python - Overview, 2022):

- **Εύκολη εκμάθηση:** Η Python έχει λίγες λέξεις – κλειδιά, απλή δομή και σαφώς καθορισμένο συντακτικό επιτρέποντας έτσι την γρήγορη κατανόησή της.
- **Εύκολη ανάγνωση:** Ο κώδικας σε Python είναι πιο σαφής και ορατός στα μάτια.
- **Εύκολη συντήρηση:** Ο πηγαίος κώδικας σε Python είναι αρκετά εύκολος στη συντήρηση.
- **Ευρεία τυποποιημένη βιβλιοθήκη:** Ο κύριος όγκος της βιβλιοθήκης της Python είναι συμβατός με πολλαπλές πλατφόρμες όπως Unix, Macintosh και Windows.
- **Δια δραστική λειτουργία:** Υποστηρίζει μια λειτουργία που επιτρέπει τη δια δραστική δοκιμή και αποσφαλμάτωση αποσπασμάτων κώδικα.
- **Βάσεις δεδομένων:** Παρέχει διεπαφές σε όλες τις μεγάλες εμπορικές βάσεις δεδομένων.
- **Προγραμματισμός GUI:** Η Python υποστηρίζει εφαρμογές GUI που μπορούν να δημιουργηθούν και να μεταφερθούν σε άλλα συστήματα και πλατφόρμες.
- **Επεκτάσιμη:** Υπάρχει δυνατότητα προσθήκης χαμηλού επιπέδου modules στον διερμηνέα της Python επιτρέποντας την προσαρμογή των εργαλείων της ώστε να γίνουν πιο αποδοτικά.
- **Ευρεία υποστήριξη δεδομένων:** Παρέχει δυναμικούς τύπους δεδομένων πολύ υψηλού επιπέδου και υποστηρίζει δυναμικό έλεγχο τύπων.
- **Ενσωμάτωση:** Μπορεί εύκολα να ενσωματωθεί με C, C++, COM, ActiveX, CORBA και Java.

Πλεονεκτήματα της Python

Ορισμένα από τα πλεονεκτήματα της Python είναι τα εξής:

- Η Python είναι δωρεάν και ανοιχτή, ώστε ο καθένας να μπορεί να τη κατεβάσει και να τη χρησιμοποιήσει αμέσως.
- Αποτελεί γλώσσα προγραμματισμού υψηλού επιπέδου με σύνταξη που είναι παρόμοια στην αγγλική γεγονός που την καθιστά εύκολη επιλογή για την κατανόηση και εκμάθηση από αρχάριους.
- Επειδή ο κώδικας είναι απλός, η παραγωγικότητα είναι συγκριτικά υψηλότερη από άλλες γλώσσες προγραμματισμού.
- Σε περίπτωση εμφάνισης σφάλματος, η Python σταματά την κωδικοποίηση μέχρι να επιλυθεί το σφάλμα συμβάλλοντας έτσι στη δημιουργία κώδικα χωρίς σφάλματα.
- Είναι ανεξάρτητη από το σύστημα, πράγμα που σημαίνει ότι δεν χρειάζεται να αλλάξει ο κώδικας όταν γίνεται χρήση σε διαφορετικές πλατφόρμες.
- Διαθέτει πολυάριθμα πακέτα στην βιβλιοθήκη της βοηθώντας τους χρήστες να εργάζονται σε διάφορες εφαρμογές με ευκολία.

Μειονεκτήματα της Python

Ορισμένα από τα μειονεκτήματα της Python είναι τα εξής:

- Η διαδικασία εκτέλεσης είναι σχετικά πιο αργή.
- Οι δομές της Python χρειάζονται πρόσθετη μνήμη.
- Μπορεί να οδηγήσει σε σφάλματα κατά τη διάρκεια εκτέλεσης.
- Δεν είναι η καλύτερη επιλογή όταν αλληλοεπιδρά με βάσεις δεδομένων.
- Η επεξεργαστική ισχύ της είναι αργή σε σύγκριση με άλλες γλώσσες.

6.1.4 MongoDB

Η MongoDB είναι μια από τις πιο δημοφιλείς βάσεις δεδομένων NoSQL ανοιχτού κώδικα γραμμένη σε C++. Από τον Φεβρουάριο του 2015, η MongoDB είναι το τέταρτο πιο δημοφιλές σύστημα διαχείρισης βάσεων δεδομένων. Η λέξη Mongo προέρχεται βασικά από το Humongous. Η MongoDB αναπτύχθηκε για πρώτη φορά από έναν οργανισμό που εδρεύει στη Νέα Υόρκη και ονομάζεται 10gen το έτος 2007. Αργότερα

η 10gen άλλαξε το όνομα και έγινε γνωστή ως MongoDB Inc από σήμερα. Στην αρχή, η MongoDB αναπτύχθηκε βασικά ως βάση δεδομένων PAAS (Platform as a Service). Το έτος 2009, παρουσιάστηκε ως βάση δεδομένων ανοικτού κώδικα με την ονομασία MongoDB 1.0. Η MongoDB 7.0 είναι η τρέχουσα σταθερή έκδοση που κυκλοφόρησε το 2023.

Η MongoDB είναι μια βάση δεδομένων εγγράφων στην οποία αποθηκεύονται δεδομένα σε μορφή αντικειμένων JSON εγγράφων με δυναμικό σχήμα. Αυτό σημαίνει ότι μπορείτε να γίνει αποθήκευση δεδομένων χωρίς να υπάρχει κάποια ανησυχία σχετικά με την δομή των δεδομένων, όπως ο αριθμός των πεδίων ή οι τύποι των πεδίων για την αποθήκευση τιμών (Matallah et al., 2021).

Χαρακτηριστικά της MongoDB

Ορισμένα από τα χαρακτηριστικά της MongoDB είναι:

- **Βάση δεδομένων χωρίς σχήμα**: Είναι το σπουδαίο χαρακτηριστικό που παρέχει η MongoDB. Μια βάση δεδομένων χωρίς σχήμα σημαίνει ότι μια συλλογή μπορεί να περιέχει διαφορετικούς τύπους εγγράφων ή με άλλα λόγια, στη βάση δεδομένων MongoDB, μια ενιαία συλλογή μπορεί να περιέχει πολλαπλά έγγραφα και αυτά τα έγγραφα μπορεί να αποτελούνται από διαφορετικό αριθμό πεδίων, περιεχόμενο και μέγεθος. Δεν είναι απαραίτητο το ένα έγγραφο να είναι παρόμοιο με ένα άλλο έγγραφο όπως στις σχεσιακές βάσεις δεδομένων. Λόγω αυτού του μοναδικού χαρακτηριστικού, η MongoDB παρέχει μεγάλη ευελιξία στις βάσεις δεδομένων.
- **Βάση δεδομένων εγγράφων**: Στην MongoDB, όλα τα δεδομένα αποθηκεύονται σε έγγραφα αντί για πίνακες όπως στα RDBMS. Σε αυτά τα έγγραφα, τα δεδομένα αποθηκεύονται σε πεδία (ζεύγος κλειδιών-τιμών) αντί για γραμμές και στήλες, γεγονός που καθιστά τα δεδομένα πολύ πιο ευέλικτα σε σύγκριση με τα RDBMS. Και κάθε έγγραφο περιέχει το μοναδικό του αναγνωριστικό αντικειμένου.
- **Ευρετηρίαση**: Αυτό διευκολύνει και απαιτεί λιγότερο χρόνο για την απόκτηση ή την αναζήτηση δεδομένων από τη δεξαμενή δεδομένων. Εάν τα δεδομένα δεν είναι ευρετηριασμένα, τότε η βάση δεδομένων αναζητά κάθε έγγραφο με το καθορισμένο ερώτημα το οποίο απαιτεί πολύ χρόνο και δεν είναι τόσο αποτελεσματικό.
- **Επεκτασιμότητα:** Η MongoDB παρέχει οριζόντια επεκτασιμότητα με τη βοήθεια του sharding. Sharding σημαίνει διανομή δεδομένων σε πολλούς διακομιστές. Με το

sharding ένας μεγάλος όγκος δεδομένων διαμερίζεται σε κομμάτια δεδομένων χρησιμοποιώντας το κλειδί και αυτά τα κομμάτια δεδομένων κατανέμονται ομοιόμορφα σε shards που βρίσκονται σε πολλούς φυσικούς διακομιστές.

- **Αντίγραφο:** Δημιουργεί πολλαπλά αντίγραφα των δεδομένων και στέλνει αυτά τα αντίγραφα σε διαφορετικό διακομιστή, έτσι ώστε εάν ένας διακομιστής αποτύχει, τότε τα δεδομένα ανακτώνται από έναν άλλο διακομιστή.
- **Συγκέντρωση:** Επιτρέπει την εκτέλεση πράξεων στα ομαδοποιημένα δεδομένα και την εξαγωγή ενός ενιαίου αποτελέσματος. Είναι παρόμοια με τη ρήτρα GROUPBY της SQL. Παρέχει τρεις διαφορετικές μεθόδους συνάθροισης, δηλ. αγωγό συνάθροισης, συνάρτηση map-reduce και μεθόδους συνάθροισης ενός σκοπού.
- **Υψηλή απόδοση:** Η απόδοση της MongoDB είναι και η ανθεκτικότητα των δεδομένων, σε σύγκριση με άλλη βάση δεδομένων, είναι πολύ υψηλή λόγω των χαρακτηριστικών της όπως η επεκτασιμότητα, η ευρετηρίαση, η αντιγραφή κ.λπ.

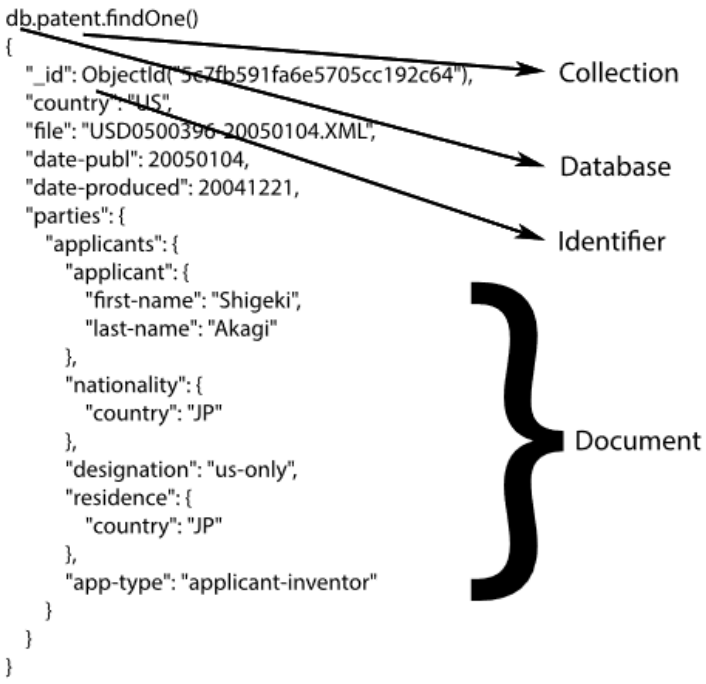
Δομικά στοιχεία της MongoDB

Παρακάτω παρουσιάζονται ορισμένοι βασικοί όροι που χρησιμοποιούνται στην MongoDB:

- **Συλλογή (collection):** Μια συλλογή στη MongoDB είναι μια ομάδα εγγράφων. Η συλλογή στη MongoDB είναι παρόμοια με τον πίνακα στη σχεσιακή βάση δεδομένων SQL και είναι ένα βασικό δομικό στοιχείο, το οποίο περιέχει την ίδια ομάδα εγγράφων. Υπάρχουν διάφοροι τύποι συλλογών στη MongoDB. Ο προεπιλεγμένος τύπος συλλογής είναι επεκτάσιμος σε μέγεθος. Η κάθε συλλογή στη MongoDB έχει ένα μοναδικό όνομα (Jha, 2022).
- **Έγγραφο (document):** Στη MongoDB, οι εγγραφές δεδομένων αποθηκεύονται ως έγγραφα BSON. Το έγγραφο δημιουργείται χρησιμοποιώντας ζεύγη πεδίου-τιμής ή ζεύγη κλειδιού-τιμής και η τιμή του πεδίου μπορεί να είναι οποιουδήποτε τύπου BSON. Σε αντίθεση με μια παραδοσιακή βάση δεδομένων όπου μια γραμμή είναι σταθερή, ένα έγγραφο στη MongoDB μπορεί να αποτελείται από οποιονδήποτε αριθμό κλειδιών και τιμών (MongoDB Document - Structure and Sample Documents, 2022).
- **Συσχετίσεις (associations):** Οι σχέσεις στη MongoDB αντιπροσωπεύουν τον τρόπο με τον οποίο τα διάφορα έγγραφα συνδέονται λογικά μεταξύ τους. Οι

σχέσεις μπορούν να μοντελοποιηθούν μέσω των προσεγγίσεων Embedded και Referenced. Οι σχέσεις αυτές μπορεί να είναι (MongoDB - Relationships, 2022):

- Ένα προς ένα
- Ένα προς πολλά
- Πολλά προς ένα
- Πολλά προς πολλά



Εικόνα 6.5 Παράδεγμα δομής ενός εγγράφου στην MongoDB

6.2 Τεχνολογίες από την πλευρά του πελάτη (Frontend)

Η ανάπτυξη ιστού front-end ή αλλιώς γνωστή ως ανάπτυξη από την πλευρά του πελάτη είναι το μέρος με το οποίο αλληλοεπιδρούν οι χρήστες. Το front–end αποτελείται από όλα όσα βλέπει ο χρήστης όταν περιηγείται στο Διαδίκτυο, από γραμματοσειρές και χρώματα έως αναπτυσσόμενα μενού και πεδία εισαγωγής κειμένου τα οποία είναι ένας συνδυασμός HTML, CSS και JavaScript τα οποία ελέγχονται από το πρόγραμμα περιήγησης του υπολογιστή.

Η κυριότερη πρόσκληση στην ανάπτυξη διεπαφών είναι ότι τα εργαλεία και οι τεχνικές που χρησιμοποιούνται εξελίσσονται συνεχώς και έτσι ο προγραμματιστής πρέπει να ενημερώνεται διαρκώς για το πως αναπτύσσεται το πεδίο. Ο στόχος του σχεδιασμού ενός ιστοτόπου είναι να διασφαλιστεί η ευκολία χρήσης του από τους χρήστες που θα το

επισκεφτούν. Αυτό περιπλέκεται περισσότερο από το γεγονός ότι οι χρήστες στην σημερινή εποχή χρησιμοποιούν μια μεγάλη ποικιλία συσκευών με διαφορετικά μεγέθη οθόνης και αναλύσεις με αποτέλεσμα να αναγκάζουν τον σχεδιαστή να λάβει υπόψη αυτές τις πτυχές κατά τον σχεδιασμό του ιστοτόπου, προκειμένου να διασφαλισθεί η σωστή λειτουργία σε διαφορετικά προγράμματα περιήγησης (cross-device) καθώς και σε διαφορετικά λειτουργικά συστήματα (cross-platform).

6.2.1 Angular

Το Angular είναι ένα ανοιχτού κώδικα πλαίσιο (framework) βασισμένο σε TypeScript για την δημιουργία δυναμικών εφαρμογών ιστού χρησιμοποιώντας HTML και TypeScript. Το πλαίσιο Angular λειτουργεί διαβάζοντας πρώτα τη σελίδα HTML η οποία διαθέτει ενσωματωμένα σε αυτήν πρόσθετα χαρακτηριστικά. Η Angular ερμηνεύει αυτά τα χαρακτηριστικά ως οδηγίες για τη σύνδεση τμημάτων εισόδου ή εξόδου της σελίδας σε ένα μοντέλο που αποτελείται από τυπικές μεταβλητές JavaScript. Η Angular βασίζεται στην πεποίθηση ότι ο δηλωτικός προγραμματισμός πρέπει να χρησιμοποιείται για τη δημιουργία διεπαφών χρήστη και τη σύνδεση στοιχείων του λογισμικού ενώ ο επιτακτικός προγραμματισμός είναι καταλληλότερος για τον ορισμό της επιχειρηματικής λογικής μιας εφαρμογής. Αποτελεί μια επέκταση για το παραδοσιακό HTML για να παρουσιάσει δυναμικό περιεχόμενο μέσω αμφίδρομης δέσμευσης δεδομένων που επιτρέπει τον αυτόματο συγχρονισμό μοντέλων και προβολών (Geetha et al., 2022).

Πλεονεκτήματα του Angular

Παρακάτω παρουσιάζονται ορισμένα πλεονεκτήματα από το πλαίσιο Angular:

- **Αρχιτεκτονική βασισμένη σε συστατικά που παρέχουν υψηλότερη ποιότητα κώδικα:** Τα συστατικά της Angular μπορούν να θεωρηθούν ως μικρά κομμάτια της διεπαφής χρήστη, όπως ένα τμήμα της εφαρμογής. Ενώ κάθε συστατικό ενθυλακώνεται με τη λειτουργικότητά του, υπάρχει μια αυστηρή ιεραρχία των συστατικών στο Angular.

  Κυριότερα πλεονεκτήματα της αρχιτεκτονικής που βασίζεται σε συστατικά:

  - **Συντηρησιμότητα:** Τα στοιχεία που αποσυνδέονται εύκολα μεταξύ τους μπορούν εύκολα να αντικατασταθούν με καλύτερες υλοποιήσεις. Βασικά, η ομάδα μηχανικών θα είναι πιο αποτελεσματική στη συντήρηση και την

ενημέρωση του κώδικα στο πλαίσιο της επαναληπτικής ροής εργασίας ανάπτυξης.

- **Επαναχρησιμότητα:** Οι προγραμματιστές μπορούν να επαναχρησιμοποιήσουν τα συστατικά σε διάφορα μέρη μιας εφαρμογής. Αυτό είναι ιδιαίτερα χρήσιμο σε εφαρμογές επιχειρηματικής εμβέλειας, όπου διαφορετικά συστήματα συγκλίνουν αλλά μπορεί να έχουν πολλά παρόμοια στοιχεία, όπως πλαίσια αναζήτησης, επιλογής ημερομηνιών, λίστες ταξινόμησης.
- **Αναγνωσιμότητα:** Η ενθυλάκωση διασφαλίζει επίσης ότι οι νέοι προγραμματιστές – οι οποίοι έχουν πρόσφατα ενταχθεί σε ένα έργο – μπορούν να διαβάζουν καλύτερα τον κώδικα και τελικά να φτάνουν γρηγορότερα στο επίπεδο της παραγωγικότητας τους.
- **Φιλικό προς το unit-testing:** Η ανεξάρτητη φύση των συστατικών απλοποιεί τις δομικές μονάδες διαδικασίας διασφάλισης ποιότητας που αποσκοπούν στην επαλήθευση της απόδοσης των μικρότερων τμημάτων της εφαρμογής, των μονάδων.

- **RxJS (αποδοτικός, ασύγχρονος προγραμματισμός):** Επιτρέπει τον ανεξάρτητο παράλληλο χειρισμό γεγονότων και τη συνέχιση της εκτέλεσης χωρίς να χρειαστεί ο χρήστης να περιμένει να συμβεί κάποιο γεγονός. Η βιβλιοθήκη RxJS λειτουργεί με Observables, ένα είδος σχεδιαγράμματος που περιγράφει τον τρόπο με τον οποίο συνδυάζονται οι ροές δεδομένων και τον τρόπο με τον οποίο η εφαρμογή αντιδρά στις μεταβλητές αυτών των ροών. Προφανώς, ο ασύγχρονος προγραμματισμός υπήρχε και πριν από το RxJS, αλλά αυτή η βιβλιοθήκη έχει κάνει πολλά πράγματα ευκολότερα.
- **Υψηλή απόδοση:** Υπάρχουν πολλοί παράγοντες που μπορούν να βοηθήσουν στην ταχύτερη υποβολή των αιτήσεων ορισμένοι από αυτούς είναι:
    - **Ιεραρχική έγχυση εξαρτημάτων:** Η τεχνική αυτή αποσυνδέει τα πραγματικά συστατικά από τις εξαρτήσεις τους εκτελώντας τα παράλληλα μεταξύ τους. Η Angular κατασκευάζει ένα ξεχωριστό δέντρο εγχύσεων εξάρτησης που μπορεί να τροποποιηθεί χωρίς επαναδιαμόρφωση των συστατικών. Έτσι οι κλάσεις δεν έχουν εξαρτήσεις από μόνες τους αλλά τις καταναλώνουν από την εξωτερική πηγή.
    - **Angular Universal:** Αποτελεί υπηρεσία που επιτρέπει την απόδοση της προβολής εφαρμογών σε έναν διακομιστή αντί για τα προγράμματα περιήγησης του πελάτη. Η Google παρέχει ένα σύνολο εργαλείων είτε για

την προκαταρκτική αναπαράσταση της εφαρμογής είτε για την εκ νέου αναπαράστασή της για κάθε αίτημα ενός χρήστη.

- **Διαφορικό φορτίο:** Το διαφορικό φορτίο προστέθηκε στην Angular 8 ως άλλη τεχνική βελτιστοποίησης. Η διαφορική φόρτωση είναι ένας τρόπος φόρτωσης περιεχομένου και βελτιστοποίησης του μεγέθους δέσμης. Αυτό που στην πραγματικότητα κάνει, είναι η δημιουργία δύο διαφορετικών πακέτων για παλαιά προγράμματα περιήγησης και νέα προγράμματα περιήγησης. Η Angular χρησιμοποιεί σύγχρονη σύνταξη και polyfills για τα νεότερα προγράμματα περιήγησης, ενώ θα δημιουργήσει ένα ξεχωριστό πακέτο με σταθερή σύνταξη για τα παλαιότερα προγράμματα περιήγησης.

- **Μακροπρόθεσμη υποστήριξη της Google:** Ορισμένοι μηχανικοί λογισμικού θεωρούν το γεγονός και μόνο ότι η Angular υποστηρίζεται από την Google ως σημαντικό πλεονέκτημα της τεχνολογίας. Η Google ανακοίνωσε τη μακροπρόθεσμη υποστήριξη (TLS) για την Angular. Αυτό ουσιαστικά σημαίνει ότι η Google σκοπεύει να παραμείνει στο οικοσύστημα της Angular και να την αναπτύξει περαιτέρω, προσπαθώντας να κρατήσει τις ηγετικές θέσεις μεταξύ των εργαλείων front-end engineering.

Μειονεκτήματα του Angular

Ορισμένα μειονεκτήματα του πλαισίου Angular είναι:

- **Πολυπλοκότητα:** Παρόλο που το σημαντικότερο πλεονέκτημα της Angular είναι τα συστατικά ο τρόπος διαχείρισης τους είναι πολύ περίπλοκος. Για παράδειγμα για ένα συστατικό μπορεί να χρειαστεί έως και πέντε αρχεία. Άλλα σημεία προβληματισμού είναι οι ειδικές για την Angular βιβλιοθήκες τρίτων κατασκευαστών και η σύνταξή της. Κατά συνέπεια, μεγάλο μέρος του χρόνου ανάπτυξης στην Angular αναλώνεται σε επαναλαμβανόμενα πράγματα.
- **Μεγάλη καμπύλη εκμάθησης:** Σε περίπτωση που κάποιος προγραμματιστής είναι εξοικειωμένος με την JavaScript θελήσει να χρησιμοποιήσει την Angular θα έχει δυσκολίες σε σύγκριση με παρόμοιες προσθήκες σε React ή Vue. Η γκάμα των θεμάτων και των πτυχών που πρέπει να καλυφθούν είναι μεγάλη: modules, dependency injection, components, services, templates κ.λπ. Ένα άλλο εμπόδιο είναι το RxJS, μια βιβλιοθήκη αντιδραστικού προγραμματισμού για ασύγχρονο

προγραμματισμό. Η εκμάθησή της, τουλάχιστον σε βασικό επίπεδο, είναι υποχρεωτική για τη χρήση του Angular.

- **Η μετάβαση παλαιών συστημάτων από AngularJS σε Angular απαιτεί χρόνο:** Υπάρχει μια μνημειώδης διαφορά μεταξύ AngularJS και Angular, και το ίδιο ισχύει και για την πορεία της μετάβασης από το παρελθόν στο μέλλον. Υπάρχουν διάφοροι τρόποι για να γίνει η μετάβαση, ένας από τους οποίους είναι να χρησιμοποιηθεί μια υβριδική προσέγγιση. Συνεπάγεται την ταυτόχρονη λειτουργία τόσο της παλιάς όσο και της νέας Angular, ενώ γίνεται ενημέρωση σταδιακά σε ολόκληρο το προϊόν. Όχι μόνο χρειάζεται χρόνος, αλλά θα πρέπει να γίνεται επανεξέταση σε πολλά εργαλεία, να γίνει μετάβαση σε μια νέα γλώσσα και να ασχοληθεί κάποιος με μια πιο βαριά εφαρμογή καθώς έχετε και τις δύο Angular σε λειτουργία. Μια άλλη τεχνική που μπορεί να εφαρμοστεί κατά τη διάρκεια της μετάβασης ονομάζεται lazy loading. Το Lazy loading είναι μια τεχνική βελτιστοποίησης, η οποία ουσιαστικά σημαίνει ότι φορτώνονται μόνο τα μέρη της εφαρμογής (ή το περιεχόμενό της) που κλήθηκαν από τον χρήστη. Όταν απαιτείται το στοιχείο ή η λειτουργία, η Angular θα αξιοποιήσει το τμήμα της εφαρμογής και θα το κάνει render. Με το lazy loading, τα μέρη της εφαρμογής AngularJS μπορούν να κληθούν και να αποδοθούν εντός της εφαρμογής Angular.

6.2.2 HTML

Η γλώσσα HTML (HyperText Markup Language) αποτελεί μια τυποποιημένη γλώσσα σήμανσης που χρησιμοποιείται για τη δημιουργία ιστοσελίδων. Επιτρέπει τη δημιουργία και τη δομή τμημάτων, παραγράφων και συνδέσμων με τη χρήση στοιχείων HTML (τα δομικά στοιχεία μιας ιστοσελίδας), όπως ετικέτες και χαρακτηριστικά. Η HTML δεν θεωρείται γλώσσα προγραμματισμού, καθώς δεν μπορεί να δημιουργήσει δυναμική λειτουργικότητα, αν και πλέον θεωρείται επίσημο πρότυπο το διαδικτύου (Roy et al., 2023).

Η HTML έχει πολλές περιπτώσεις χρήσης, συγκεκριμένα:

- **Ανάπτυξη ιστοσελίδων:** Οι προγραμματιστές χρησιμοποιούν κώδικα HTML για να σχεδιάσουν τον τρόπο με τον οποίο ένα πρόγραμμα περιήγησης εμφανίζει στοιχεία ιστοσελίδας, όπως κείμενο, υπερσυνδέσμους και αρχεία πολυμέσων.

- **Πλοήγηση στο Διαδίκτυο:** Οι χρήστες μπορούν εύκολα να προηγηθούνε και να εισάγουν συνδέσμους μεταξύ σχετικών σελίδων και ιστότοπων, καθώς η HTML χρησιμοποιείται σε μεγάλο βαθμό για την ενσωμάτωση υπερσυνδέσμων.

Πλεονεκτήματα της HTML

Ορισμένα από τα πλεονεκτήματα της HTML είναι :

- **Φιλική προς τους αρχάριους:** Η HTML έχει ένα καθαρό και συνεπής συντακτικό ενώ είναι ιδιαίτερα εύκολη στην εκμάθηση.
- **Υποστήριξη:** Η HTML χρησιμοποιείται ευρέως με πολλούς πόρους διαθέτοντας μια μεγάλη κοινότητα για την υποστήριξη της.
- **Ευέλικτη:** Ενσωματώνεται εύκολα με γλώσσες backend όπως PHP και Node.js.
- **Προσβάσιμη:** Είναι ανοιχτού κώδικα και τρέχει εγγενώς σε όλα τα προγράμματα περιήγησης ιστού.

Μειονεκτήματα της HTML

- **Ξεχωριστές σελίδες:** Οι χρήστες θα πρέπει να δημιουργούν ξεχωριστές ιστοσελίδες για την HTML, ακόμα και αν τα στοιχεία είναι τα ίδια.
- **Συμβατότητα με τα προγράμματα περιήγησης:** Ορισμένα προγράμματα περιήγησης υιοθετούν νέα χαρακτηριστικά με αργούς ρυθμούς. Αυτό έχει ως αποτέλεσμα ορισμένοι παλαιότερη περιηγητές να μην υποστηρίζουν πάντα τις νεότερες ετικέτες.
- **Υποστήριξη μόνο στατικών σελίδων:** Η HTML χρησιμοποιεί μόνο στατικές ιστοσελίδες. Για δυναμική λειτουργικότητα, ίσως χρειαστεί να χρησιμοποιηθεί JavaScript ή PHP.

6.3 Επιπλέον πακέτα και βιβλιοθήκες

Οι βιβλιοθήκες είναι ένα σύνολο χρήσιμων λειτουργιών που εξαλείφουν την ανάγκη συγγραφής κώδικα από την αρχή. Μία βιβλιοθήκη είναι ένα σύνολο από ήδη υλοποιημένους κώδικες που μπορούν να χρησιμοποιηθούν επαναληπτικά για να μειωθεί ο χρόνος που απαιτείται για την κωδικοποίηση. Το γεγονός αυτό είναι ιδιαίτερα

σημαντικό καθώς αυτά τα κομμάτια κώδικα μπορούν να χρησιμοποιηθούν σε αρκετά μέρη χωρίς να χρειάζεται να γραφτούν από την αρχή κάθε φορά. Αποτελεί παρόμοια φιλοσοφία με τις φυσικές βιβλιοθήκες, δηλαδή πρόκειται για μια συλλογή επαναχρησιμοποιήσιμων πόρων, πράγμα που σημαίνει ότι κάθε βιβλιοθήκη έχει μια βασική πηγή.

Ορισμένες βιβλιοθήκες – πακέτα που χρησιμοποιήθηκαν κατά την υλοποίηση του συστήματος συστάσεων είναι:

- Pandas
- NumPy
- Scikit – learn
- Matplotlib

6.3.1 Pandas

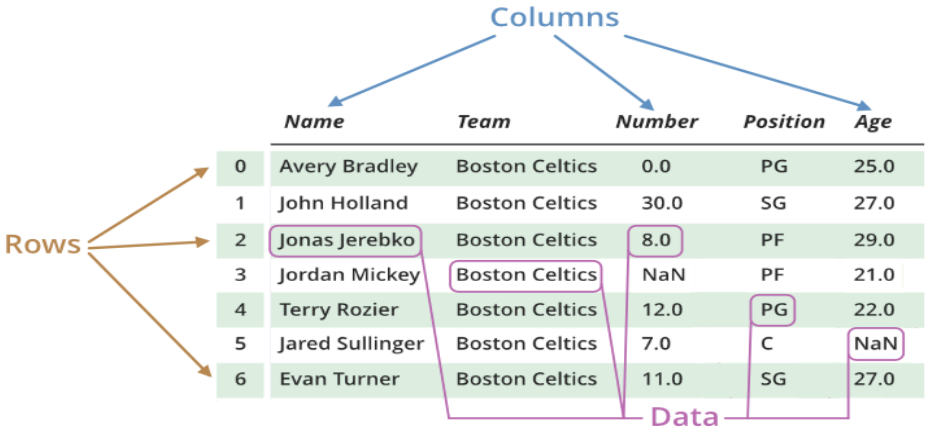
Το Pandas είναι ένα πακέτο – βιβλιοθήκη της Python ανοιχτού κώδικα που χρησιμοποιείται ευρέως για την επιστήμη και ανάλυση δεδομένων και εργασίες μηχανικής μάθησης. Η ονομασία της βιβλιοθήκης αυτής προέρχεται από το Pan (Panel) πίνακας και το Das (Data) δεδομένα. Βασίζεται πάνω σε ένα άλλο πακέτο που ονομάζεται Numpy, το οποίο παρέχει υποστήριξη για πολυδιάστατους πίνακες. Καθώς αποτελεί ένα από τα πιο δημοφιλή πακέτα επεξεργασίας δεδομένων, το Pandas συνεργάζεται καλά με πολλές άλλες ενότητες επιστήμης δεδομένων μέσα στο οικοσύστημα της Python.

Εικόνα 6.6 Το logo της βιβλιοθήκης Pandas.

(Πηγή: https://upload.wikimedia.org/wikipedia/commons/e/ed/Pandas_logo.svg)

Το Pandas βασίζει την λειτουργία του στην χρήση μιας δομής δεδομένων που ονομάζεται DataFrame. Το Pandas DataFrame είναι μια δισδιάστατη δομή δεδομένων με δυνατότητα

αλλαγής μεγέθους δυνητικά ετερογενής, με πίνακες δεδομένων και με επισημασμένους άξονες ευθυγραμμίζοντας τα δεδομένα σε γραμμές και στήλες. Τα DataFrame αποτελούνται από τρία κύρια συστατικά τα δεδομένα, τις γραμμές και τις στήλες.



Εικόνα 6.7 Παράδειγμα χρήσης της δομής δεδομένων Data Frame.

Οι κυριότεροι τομείς στους οποίους η βιβλιοθήκη Pandas είναι:

- Χρηματοοικονομικά
- Οικονομικά
- Ανάλυση

Το Pandas διευκολύνει την εκτέλεση πολλών από τις χρονοβόρες και επαναλαμβανόμενες εργασίες που σχετίζονται με την επεξεργασία και διαχείριση δεδομένων όπως:

- Καθαρισμός δεδομένων
- Συμπλήρωση δεδομένων
- Κανονικοποίηση δεδομένων
- Συγχωνεύσεις και ενώσεις
- Οπτικοποίηση δεδομένων
- Φιλτράρισμα των δεδομένων
- Εισαγωγή και διαγραφή στηλών δομών δεδομένων
- Αναδιαμόρφωση συνόλων δεδομένων
- Αντικείμενα DataFrame για χειρισμό δεδομένων με ενσωματωμένη ευρετηρίαση.

- Λειτουργίες χρονοσειρών όπως δημιουργία εύρους ημερομηνιών και μετατροπείς συχνοτήτων, στατιστικά στοιχεία κινούμενου παραθύρου, γραμμικές παλινδρομήσεις κινούμενου παραθύρου, μετατόπιση ημερομηνίας.
- Στατιστική ανάλυση
- Επιθεώρηση δεδομένων
- Φόρτωση και αποθήκευση δεδομένων

6.3.2 Scikit – learn

Η Scikit – learn αποτελεί την πιο χρήσιμη και ισχυρή βιβλιοθήκη για μηχανική μάθηση στην Python. Αυτή η βιβλιοθήκη, η οποία είναι σε μεγάλο βαθμό γραμμένη σε Python, βασίζεται στις NumPy, SciPy και Matplotlib. Παρέχει μια επιλογή αποτελεσματικών εργαλείων για μηχανική μάθηση και στατιστική μοντελοποίηση, συμπεριλαμβανομένης της ταξινόμησης, της παλινδρόμησης, της μείωσης της διαστατικότητας και της ομαδοποίησης μέσω μιας διεπαφής στην Python. Η βιβλιοθήκη αυτή είναι σε μεγάλο βαθμό γραμμένη σε Python και βασίζεται στις NumPy, SciPy και Matplotlib.

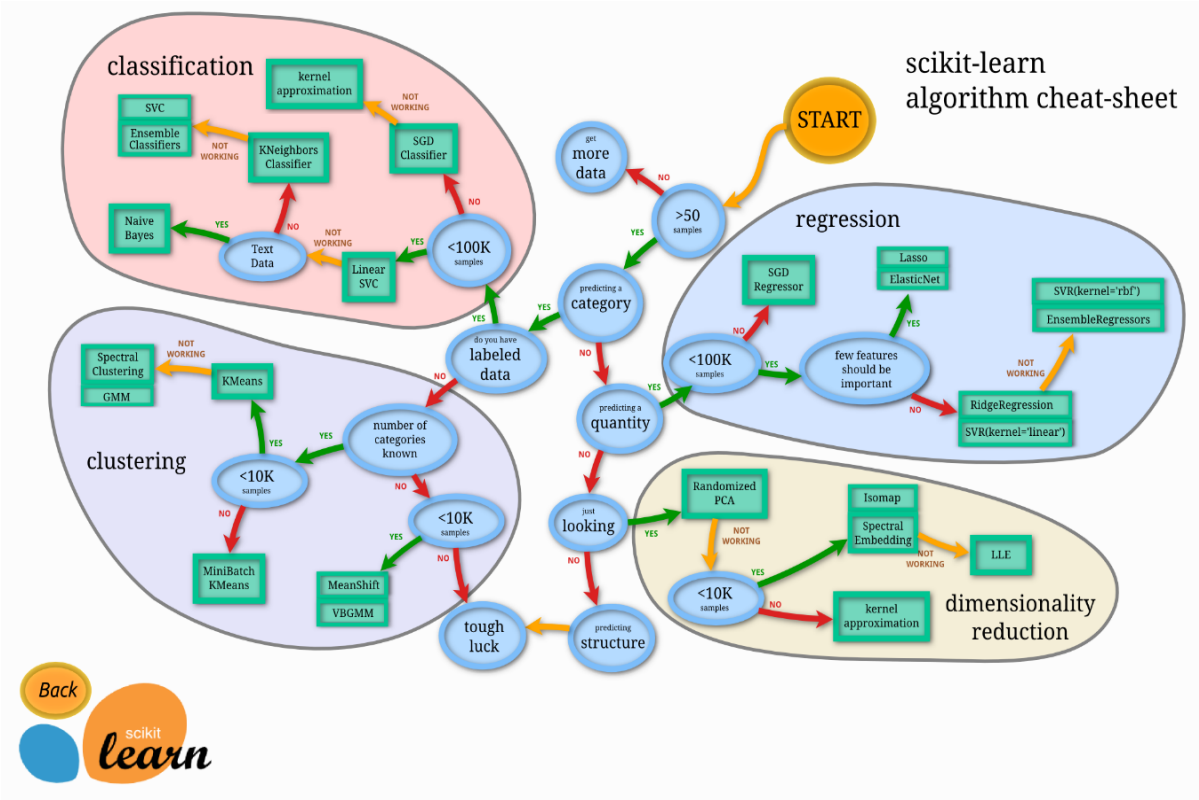
Εικόνα 6.8 Το logo της βιβλιοθήκης Scikit – Learn.

(Πηγή: : https://upload.wikimedia.org/wikipedia/commons/0/05/Scikit_learn_logo_small.svg)

Ορισμένα από τα χαρακτηριστικά της scikit – learn είναι (Scikit Learn - Modelling Process, 2022):

- **Υποστήριξη αλγορίθμων επιβλεπόμενης μάθησης:** Σχεδόν όλοι οι δημοφιλείς αλγόριθμοι επιβλεπόμενης μάθησης, όπως η γραμμική παλινδρόμηση, η μηχανή διανυσμάτων υποστήριξης (SVM), το δέντρο αποφάσεων κ.λπ. αποτελούν μέρος του scikit – learn.

- **Υποστήριξη μάθησης χωρίς επίβλεψη:** Διαθέτει επίσης όλους τους γνωστούς αλγορίθμους μάθησης χωρίς επίβλεψη, από ομαδοποίηση, ανάλυση παραγόντων, την ανάλυση κύριων παραγόντων (PCA) έως τα νευρωνικά δίκτυα χωρίς επίβλεψη.
- **Ομαδοποίηση (Clustering):** Το μοντέλο αυτό χρησιμοποιείται για την ομαδοποίηση δεδομένων χωρίς ετικέτες.
- **Μείωση διαστάσεων (Dimensionality Reduction):** Η μεθοδολογία αυτή χρησιμοποιείται για τη μείωση του αριθμού χαρακτηριστικών στα δεδομένα τα οποία θα χρησιμοποιηθούν στην περαιτέρω ανάλυση, οπτικοποίηση και επιλογή χαρακτηριστικών.
- **Διασταυρούμενη επικύρωση (Cross Validation):** Ελέγχει την ακρίβεια των εποπτευόμενων μοντέλων σε νέα δεδομένα.
- **Εξαγωγή χαρακτηριστικών (Feature extraction):** Χρησιμοποιείται για την εξαγωγή των χαρακτηριστικών από ένα σύνολο δεδομένων κειμένου ή εικόνας.
- **Επιλογή χαρακτηριστικών (Feature selection):** Χρησιμοποιείται κυρίως στα μοντέλα με επίβλεψη και συμβάλει στον προσδιορισμό των χρήσιμων χαρακτηριστικών για τη δημιουργία μοντέλων με επίβλεψη.
- **Μέθοδοι Ensemble:** Συμβάλει στον συνδυασμό των προβλέψεων πολλαπλών εποπτευόμενων μοντέλων.
- **Ανοιχτός κώδικας:** Αποτελεί βιβλιοθήκη ανοιχτού κώδικα και εμπορικά αξιοποιήσιμη με άδεια BSD.


scikit-learn
algorithm cheat-sheet
START
get more data
>50 samples
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
Naive Bayes
Text Data
Linear SVC
<100K samples
clustering
Spectral Clustering
GMM
KMeans
number of categories known
<10K samples
MiniBatch KMeans
MeanShift
VBGMM
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
few features should be important
RidgeRegression
SVR(kernel='linear')
dimensionality reduction
Randomized PCA
Isomap
Spectral Embedding
LLE
kernel approximation
Back
scikit learn

Εικόνα 6.9

7 Σχεδιασμός και ανάπτυξη συστήματος συστάσεων

Ο σχεδιασμός ενός συστήματος αποτελεί ένα κρίσιμο βήμα στην ανάπτυξη και υλοποίηση κάθε επιτυχημένου έργου. Είναι η φάση κατά την οποία προσδιορίζουμε και αναλύουμε λεπτομερώς τις απαιτήσεις του συστήματος που θα υλοποιηθεί, ώστε να επιτευχθούν οι στόχοι του με τον καλύτερο και πιο αποτελεσματικό τρόπο. Οι απαιτήσεις αποτελούν τον οδηγό για τη σχεδίαση του συστήματος, καθώς προσδιορίζουν τις λειτουργίες, τις δυνατότητες και τις προδιαγραφές που πρέπει να πληροί.

Η σημαντικότητα του σχεδιασμού ενός συστήματος αντανακλάται στην ικανότητα του να αντιμετωπίζει αποτελεσματικά προβλήματα και να παρέχει λύσεις που καλύπτουν τις ανάγκες και τις προσδοκίες των χρηστών. Είναι ο βασικός πυλώνας που υποστηρίζει την εξέλιξη και την καινοτομία καθώς τα απαιτούμενα χαρακτηριστικά και οι λειτουργίες πρέπει να προσαρμόζονται στις μεταβαλλόμενες ανάγκες της κοινωνίας και της τεχνολογίας. Ο σχεδιασμός ενός συστήματος είναι επίσης απαραίτητος για να εξασφαλιστεί η αποτελεσματική διαχείριση των πόρων και των διεργασιών, μειώνοντας τον κίνδυνο σφαλμάτων και ανεπιθύμητων αποτελεσμάτων. Ένας καλός σχεδιασμός μετριάζει τις πιθανές δυσλειτουργιών και αντιμετωπίζει τις προκλήσεις, διασφαλίζοντας την ομαλή λειτουργία του συστήματος.

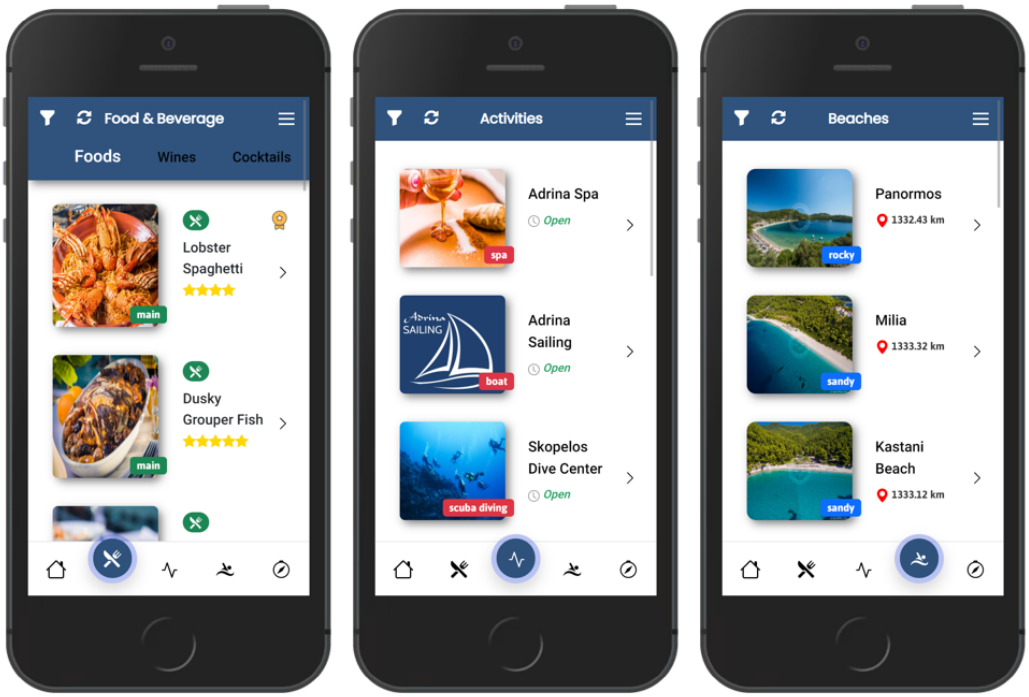
Ο σχεδιασμός του συστήματος απαιτεί προσεκτική ανάλυση των αναγκών των χρηστών και των ενδιαφερόμενων μερών, καθώς και την κατανόηση των συναφών προβλημάτων και περιορισμών. Κάθε απαίτηση πρέπει να είναι σαφής, μετρήσιμη και εφικτή. Στόχος είναι να προσδιοριστεί με ακρίβεια τι ακριβώς πρέπει να κάνει το σύστημα και πώς πρέπει να το κάνει, προσφέροντας μια ολοκληρωμένη εικόνα της λειτουργικότητας που αναμένεται να παρέχει.

Ο καλός σχεδιασμός εξυπηρετεί τον τελικό χρήστη. Με τον εστιασμένο, συνεπή σχεδιασμό, μπορούν να δημιουργηθούν υψηλής ποιότητας και ευχάριστες εμπειρίες για τους χρήστες. Αυτό μπορεί να οδηγήσει σε αυξημένη ικανοποίηση, πελατειακή πιστότητα και θετικές αντιδράσεις που μπορούν να βοηθήσουν το σύστημα να επιτύχει τους στόχους του και να ξεχωρίσει από τους ανταγωνιστές του.

7.1 Περιγραφή του συστήματος συστάσεων

Στα πλαίσια της διπλωματικής εργασίας δημιουργήθηκε ένα σύστημα συστάσεων το οποίο λειτουργεί συμπληρωματικά στο παρασκήνιο μιας εφαρμογής τουριστικού περιεχομένου και βασίζεται στα δεδομένα που συλλέγονται από την αλληλεπίδραση των χρηστών της εφαρμογής. Όλα τα δεδομένα που συλλέγονται είναι ανώνυμα και βασίζονται στον μοναδικό κωδικό της συσκευής του κάθε χρήστη (UUID) για την εξατομίκευση των συστάσεων. Σκοπός της εφαρμογής είναι να προβάλει και να αναδείξει στους χρήστες επιπλέον περιεχόμενα και επιλογές που ταιριάζουν με τα ενδιαφέροντα και τις επιλογές τους.

Το σύστημα βασίζει την λειτουργία του σε ένα σύνολο υπομονάδων – υπό συστημάτων (περιγράφονται σε επόμενη ενότητα) τα οποία συνεργάζονται αρμονικά μεταξύ τους και προσφέρουν τις ολοκληρωμένες συστάσεις στους χρήστες της εφαρμογής. Η εφαρμογή διαθέτει ένα μεγάλο σύνολο δεδομένων όπως δραστηριότητες, παραλίες και παραδοσιακά φαγητά τα οποία μπορούν οι τουρίστες να εξερευνήσουν μέσω της εφαρμογής. Σκοπός του συστήματος συστάσεων είναι να προτείνει στους χρήστες περιεχόμενο το οποίο δεν έχουν ανακαλύψει – προβάλει ακόμα το οποίο όμως είναι παρόμοιο με αυτά που έχουν δει ήδη.



Εικόνα 7.1 Στιγμιότυπο από τις βασικές λίστες περιεχομένου της εφαρμογής

7.2 Απαιτήσεις συστήματος

Οι απαιτήσεις αποτελούν την καρδιά και τον πυρήνα κάθε επιτυχημένου συστήματος. Αποτελούν τον κρίσιμο καθοριστικό παράγοντα που καθορίζει το τι πρέπει να πετύχει ένα σύστημα και πώς πρέπει να το επιτύχει. Ο σωστός ορισμός και καταγραφή των απαιτήσεων αποτελεί το θεμέλιο για την επιτυχή υλοποίηση, τη λειτουργία και την ικανοποίηση των αναγκών των χρηστών.

Οι απαιτήσεις καθορίζουν τις λειτουργίες και τις δυνατότητες που πρέπει να διαθέτει το σύστημα, καθώς και τους περιορισμούς και τις προδιαγραφές που πρέπει να τηρούνται. Αποτελούν τον κοινό γνώμονα μεταξύ των μελών της ομάδας ανάπτυξης, των πελατών και των ενδιαφερόμενων μερών. Μέσω των απαιτήσεων, οι στόχοι και οι προσδοκίες γίνονται σαφείς και περιγράφονται με ακρίβεια.

Οι απαιτήσεις χωρίζονται γενικά σε δύο τύπους:

- Λειτουργικές απαιτήσεις
- Μη λειτουργικές απαιτήσεις

7.2.1 Λειτουργικές απαιτήσεις

Οι λειτουργικές απαιτήσεις του τρέχοντος συστήματος συστάσεων είναι:

- **Προσωποποιημένες συστάσεις:** Το σύστημα πρέπει να παρέχει συστάσεις που είναι προσαρμοσμένες στις προτιμήσεις του κάθε χρήστη, λαμβάνοντας υπόψη τις δράσεις και τις πληροφορίες του.
- **Παροχή συστάσεων για φαγητά και ποτά:** Το σύστημα πρέπει να προτείνει φαγητά και ποτά που ανταποκρίνονται στις ανάγκες και τις προτιμήσεις των χρηστών.
- **Παροχή συστάσεων για δραστηριότητες:** Το σύστημα πρέπει να προτείνει δραστηριότητες που ανταποκρίνονται στις ανάγκες και τις προτιμήσεις των χρηστών.
- **Παροχή συστάσεων για παραλίες:** Το σύστημα πρέπει να προτείνει παραλίες που ανταποκρίνονται στις ανάγκες και τις προτιμήσεις των χρηστών.
- **Ακρίβεια και αξιοπιστία:** Οι συστάσεις πρέπει να είναι ακριβείς και αξιόπιστες, έτσι ώστε να ικανοποιούν τις ανάγκες των χρηστών και να αυξάνουν την πιθανότητα θετικής ανταπόκρισης από τους χρήστες.

- **Αναγνώριση προτιμήσεων:** Το σύστημα πρέπει να μπορεί να ανιχνεύει και να καταγράφει τις προτιμήσεις του χρήστη, ώστε να βελτιώνει συνεχώς τις συστάσεις του.
- **Ανάλυση συμπεριφοράς χρηστών:** Το σύστημα πρέπει να αναλύει τη συμπεριφορά των χρηστών και τις αλληλεπιδράσεις τους με το περιεχόμενο, προκειμένου να κατανοήσει καλύτερα τις προτιμήσεις τους.
- **Επεκτασιμότητα:** Το σύστημα πρέπει να είναι εύκολο να επεκταθεί για να υποστηρίξει μεγαλύτερο αριθμό χρηστών, περιεχομένου και δεδομένων.

7.2.2 Μη λειτουργικές απαιτήσεις

Οι μη λειτουργικές απαιτήσεις είναι ένα σύνολο προδιαγραφών που περιγράφουν τις δυνατότητες και τους περιορισμούς λειτουργίας του συστήματος και επιχειρούν να βελτιώσουν τη λειτουργικότητά του. Οι μη λειτουργικές απαιτήσεις είναι σημαντικές επειδή μπορούν να έχουν σημαντικό αντίκτυπο στη συνολική ποιότητα και επιτυχία ενός συστήματος λογισμικού. Συχνά συνδέονται στενά με την απόδοση, την ασφάλεια και τη χρηστικότητα του συστήματος. Βοηθούν επίσης να διασφαλιστεί ότι το σύστημα είναι συντηρήσιμο, φορητό και συμβατό με τους σχετικούς νόμους και κανονισμούς.

Κυρίως ασχολούνται με θέματα όπως:

- Φορητότητα
- Ασφάλεια
- Συντηρησιμότητα
- Αξιοπιστία
- Επεκτασιμότητα
- Απόδοση
- Ευελιξία

Ορισμένες μη λειτουργικές απαιτήσεις του συστήματος συστάσεων στην εφαρμογή είναι:

- **Ασφάλεια και απόρρητο:** Το σύστημα θα πρέπει να προστατεύει τα προσωπικά δεδομένων των χρηστών – πελατών και να διασφαλίζει την ασφάλεια των χρηστών.
- **Απόκριση σε πραγματικό χρόνο:** Οι συστάσεις πρέπει να παρέχονται άμεσα κατά τη διάρκεια της χρήσης της εφαρμογής.

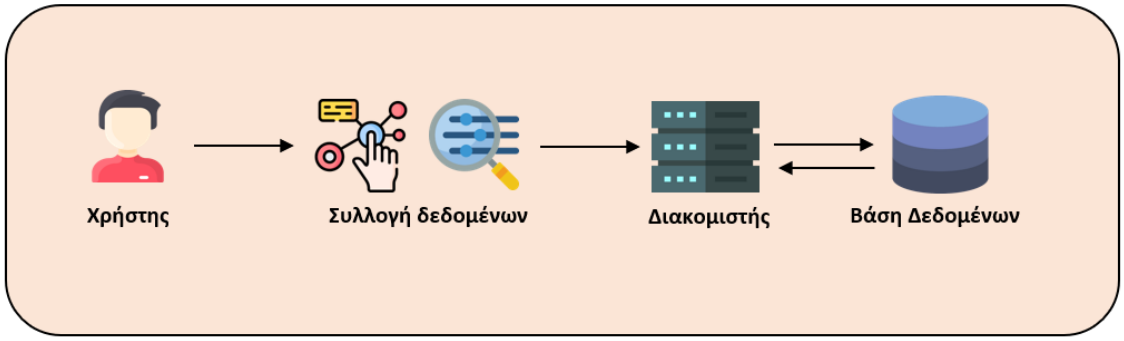
- **Διαθεσιμότητα:** Το σύστημα πρέπει να είναι πάντα διαθέσιμο και να μπορεί να ανταποκρίνεται σε μεγάλο αριθμό χρηστών.
- **Κλιμακωσιμότητα:** Το σύστημα πρέπει να μπορεί εύκολα να κλιμακωθεί για να υποστηρίξει αυξημένο αριθμό χρηστών και δεδομένων.
- **Χρηστικότητα και φιλική προς τον χρήστη διεπαφή:** Η διεπαφή πρέπει να είναι εύκολη στη χρήση και να προσφέρει απλό και ευχάριστο τρόπο χρήσης της εφαρμογής.

7.3 Επιμέρους μονάδες του συστήματος

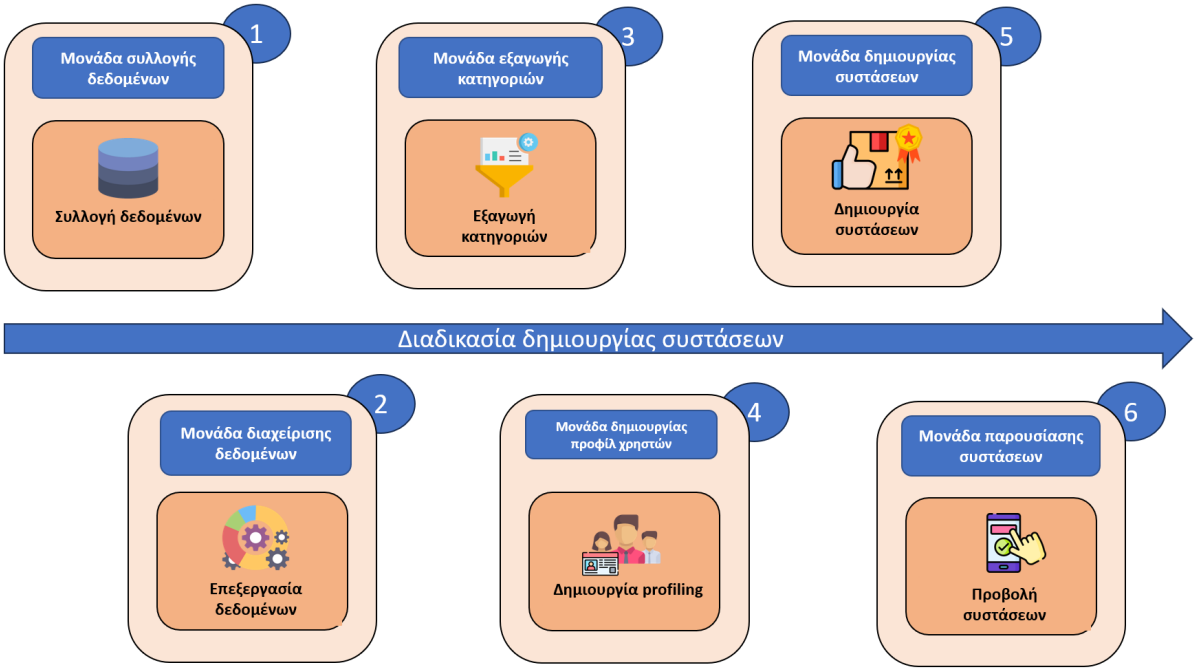
Ένα σύστημα συστάσεων είναι μια εφαρμογή που παρέχει προτάσεις σε χρήστες για αντικείμενα ή πληροφορίες που πιθανώς θα τους ενδιαφέρουν. Οι σύγχρονες τεχνικές συστάσεων βασίζονται συνήθως σε αλγόριθμους μηχανικής μάθησης και αξιοποιούν δεδομένα που συλλέγονται από τους χρήστες και τα αντικείμενα του συστήματος για να δημιουργήσουν εξατομικευμένες προτάσεις.

Οι βασικές υπομονάδες ενός συστήματος συστάσεων περιλαμβάνουν τα εξής:

- Μονάδα συλλογής δεδομένων
- Μονάδα διαχείρισης και ανάλυσης δεδομένων
- Μονάδα εξαγωγής κατηγοριών και υποκατηγοριών
- Μονάδα δημιουργίας προφίλ χρηστών
- Μονάδα δημιουργίας συστάσεων
- Μονάδα παρουσίασης συστάσεων



Εικόνα 7.2 Βασική ροή της εφαρμογής



Εικόνα 7.3 Περιγραφή της αρχιτεκτονικής του συστήματος συστάσεων

7.3.1 Μονάδα συλλογής δεδομένων

Το βασικότερο συστατικό στο σύστημα συστάσεων είναι τα δεδομένα των χρηστών τα οποία προέρχονται από την εφαρμογή. Τα δεδομένα αυτά περιγράφουν την αλληλεπίδραση των χρηστών με την εφαρμογή (clickstreams). Διαθέτουν στοιχεία όπως τις μεταβάσεις των χρηστών από κάποια σελίδα – οθόνη της εφαρμογής σε κάποια άλλη, αναζητήσεις και φιλτραρίσματα των χρηστών στο περιεχόμενο της εφαρμογής κλπ. Όλα αυτά αποθηκεύονται ανώνυμα με την χρήση του UUID της συσκευής του χρήστη στην βάση δεδομένων μας.

Προκειμένου να πραγματοποιηθεί η δημιουργία συστάσεων θα πρέπει να γίνεται η απαραίτητη σύνδεση με την βάση δεδομένων έτσι ώστε να γίνεται λήψη των νέων δεδομένων clickstreams που δημιουργούνται συνεχώς από τους χρήστες. Η βάση δεδομένων είναι μια MongoDB οι οποία βασίζεται στα έγγραφα (collections) στα οποία αποθηκεύονται τα δεδομένα. Για τα δεδομένα των χρηστών έχει δημιουργηθεί ένα αντίστοιχο collection με την ονομασία clickstream. Μόλις γίνει η λήψη των δεδομένων, δίνονται σαν είσοδο στο επόμενο στάδιο – μονάδα που είναι υπεύθυνη για την διαχείριση και την ανάλυση τους με σκοπό την προετοιμασία τους για την εξαγωγή πληροφοριών.

Η διασύνδεση με την βάση δεδομένων γίνεται μέσω της βιβλιοθήκης pymongo η οποία περιγράφεται σε επόμενη ενότητα. Ο λόγος για τον οποίο «τραβάμε» τα δεδομένα από

την βάση δεδομένων και όχι από κάποιο τοπικό αρχείο είναι γιατί τα δεδομένα αυτά των χρηστών ανανεώνονται συνέχεια με νέα οπότε το σύστημα συστάσεων θα πρέπει να τροφοδοτείται συνεχώς με νέα δεδομένα για να παρέχει ακριβής συστάσεις στους χρήστες.

```python
import pandas as pd
import pymongo

def retrieveDBdata():

    client = pymongo.MongoClient("mongodb://localhost:27017/")
    db = client['stageAdrina']

    databases = ["menus", "wines", "cocktails", "restaurants", "activities", "beaches"]

    id_names = {}
    for database in databases:
        collections = db[database]
        all_documents = collections.find()
        df = pd.DataFrame(all_documents)
        df['_id'] = df['_id'].astype(str)

        uid = df['_id']
        names = df['name']

        result_df = pd.concat([uid, names], axis=1)
        for index, values in result_df.iterrows():
            id_names[values['_id']] = values['name']
    json_str = json.dumps(id_names)

    with open("dataset/link-to-name.json", 'w') as file:
        file.write(json_str)
```

Εικόνα 7.4 Η συνάρτηση retrieveDBdata χρησιμοποιείται για την διασύνδεση και λήψη δεδομένων από την βάση δεδομένων.

7.3.2 Μονάδα διαχείρισης και ανάλυσης δεδομένων

Σε αυτή την μονάδα τα δεδομένα προετοιμάζονται ώστε να χρησιμοποιηθούν στις επιμέρους διεργασίες – μονάδες για την δημιουργία συστάσεων. Σε πρώτη φάση γίνεται απαλοιφή των χαρακτηριστικών που δεν λαμβάνονται υπόψη κατά την διαδικασία εξαγωγής των συστάσεων.

Μια εγγραφή αλληλεπίδρασης του χρήστη (clickstream) διαθέτει εννέα χαρακτηριστικά τα οποία είναι (περιγράφονται λεπτομερώς σε προηγούμενο παραδοτέο):

- Uid
- Timestamp
- Social

- Previous
- Current
- Filter
- FbLogin
- DetailsID
- Action

Από αυτά τα εννέα γίνεται χρήση μόνο των uid, current, timestamp καθώς τα υπόλοιπα χαρακτηριστικά δεν συμβάλλουν στην δημιουργία συστάσεων, και η διατήρησή τους θα επηρέαζε αρνητικά την απόδοση του μοντέλου.

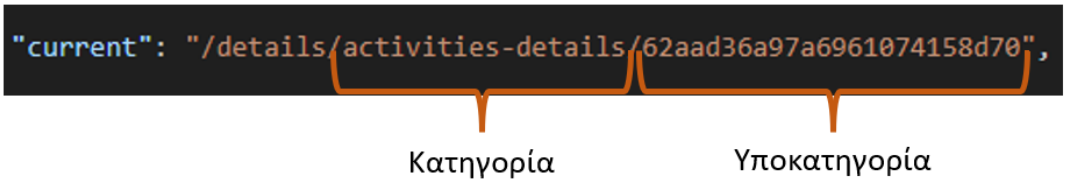
Επιπλέον πραγματοποιείται και καθαρισμός των δεδομένων με σκοπό την απαλοιφή του θορύβου από το σύνολο δεδομένων. Συνήθως γίνεται αφαίρεση δεδομένων που με κάποιο τρόπο λόγο κάποιας δυσλειτουργίας του συστήματος δεν έχει συγκρατηθεί το uuid του χρήστη ή χρονική στιγμή της αλληλεπίδρασης (timestamp).

Όλες αυτές οι προβληματικές περιπτώσεις θα πρέπει να αφαιρούνται από το σύνολο των δεδομένων καθώς ενδέχεται σε μεγάλο βαθμό να επηρεάσουν αρνητικά τα αποτελέσματα των επόμενων υπό μονάδων και γενικότερα του μοντέλου συστάσεων. Μόλις γίνει η απαλοιφή των μη χρήσιμων χαρακτηριστικών και ο καθαρισμός τα δεδομένα δίνονται ως είσοδο στην επόμενη μονάδα η οποία είναι υπεύθυνη για την εξαγωγή των κατηγοριών και υποκατηγοριών.

7.3.3 Μονάδα εξαγωγής κατηγοριών και υποκατηγοριών

Εφόσον έχει γίνει ο καθαρισμός των δεδομένων και η απαλοιφή των περιττών χαρακτηριστικών αναλαμβάνει η μονάδα εξαγωγής των κατηγοριών και υποκατηγοριών. Οι κατηγορίες που εξάγονται αφορούν τις πέντε μεγάλες κατηγορίες περιεχομένων που εμφανίζονται στην εφαρμογή οι οποίες είναι φαγητά (foods), κρασιά (wines), κοκτέιλ (cocktails), δραστηριότητες (activities), παραλίες (beaches). Από την άλλη, οι υποκατηγορίες περιγράφουν κάποια συγκεκριμένη εγγραφή μέσα από της βασικές κατηγορίες όπως για παράδειγμα ένα συγκεκριμένο φαγητό, ή μία συγκεκριμένη παραλία. Η εξαγωγή των κατηγοριών και υποκατηγοριών επιτυγχάνεται με την επεξεργασία της μεταβλητής – πεδίου current από τα δεδομένα του χρήστη (clickstreams) που δημιουργούνται κατά την αλληλεπίδραση του με την εφαρμογή. Το

πεδίο current περιέχει στοιχεία για την τρέχουσα σελίδα που βρίσκεται ο χρήστης. Μέσα από αυτό το πεδίο γίνεται απευθείας εξαγωγή της κατηγορίας και της υποκατηγορίας.



Εικόνα 7.5 Παράδειγμα κατηγορίας και υποκατηγορίας

Στην παραπάνω εικόνα παρουσιάζεται ένα παράδειγμα από μια εγγραφή χρήστη ο οποίος επισκέφθηκε μια συγκεκριμένη εγγραφή από την λίστα δραστηριοτήτων της εφαρμογής. Τα δεδομένα αλληλεπίδρασης δημιούργησαν την εξής εγγραφή που παρουσιάζεται στην Εικόνα 8.3. Δεδομένου του ότι η υποκατηγορία δίνεται με τον μοναδικό αναγνωριστικό _id γίνεται σύνδεση με την βάση δεδομένων για να ληφθούν η πληροφορίες του συγκεκριμένου _id. Μετά την λήψη των λεπτομερών δεδομένων του περιεχομένου εφαρμόζονται τεχνικές στατιστικής ανάλυσης οι οποίες λαμβάνουν τα δεδομένα του χρήστη και εξάγουν ορισμένα στατιστικά μέτρα που θα χρησιμοποιηθούν για την δημιουργία του προφίλ. Πιο συγκεκριμένα το σύστημα – profiler υπολογίζει τη συχνότητα και τη δημοσιότητα για τις αναζητήσεις του κάθε χρήστη ξεχωριστά. Για κάθε χρήστη δημιουργείται ένα αντικείμενο τύπου Profiler.

Το σύστημα profiler ξεκινάει την προετοιμασία των δεδομένων του κάθε χρήστη. Δέχεται ως είσοδο όλο το ιστορικό αναζητήσεων και την ταυτότητα των χρηστών με σκοπό να τοποθετηθούν οι αναζητήσεις που έκανες ένας συγκεκριμένος χρήστης σε ένα αντικείμενο. Ο κάθε χρήστης είναι μοναδικός και έχει το δικό του UID με σκοπό τον διαχωρισμό του από τους υπόλοιπους. Μέσω αυτών των πληροφοριών μπορούμε να εντοπίσουμε το ενδιαφέρον του κάθε χρήστη και να του συστήνουμε μια υποκατηγορία με βάση τη δραστηριότητά του.

```python
def update_subcatIds():
    databases = ["menus", "wines", "cocktails", "restaurants", "activities", "beaches"]
    foodsDetails = "food-beverage-details"
    beachesDetails = "beaches-details"
    activitiesDetails = "activities-details"
    exploreDetails = "explore-details"
    ARDetails = "augmented-reality-details"

    id_names = {}
    for database in databases:
        collections = db[database]
        all_documents = collections.find()
        df = pd.DataFrame(all_documents)
        df['_id'] = df['_id'].astype(str)

        uid = df['_id']
        names = df['name']

        result_df = pd.concat([uid, names], axis=1)
        for index, values in result_df.iterrows():
            id_names[values['_id']] = values['name']

    json_str = json.dumps(id_names)

    with open("dataset/link-to-name.json", 'w') as file:
        file.write(json_str)
```

Εικόνα 7.6 Η συνάρτηση update_subcatIds χρησιμοποιείται για την διασύνδεση με την βάση και την λήψη πληροφοριών με βάση το id των υποκατηγοριών.

7.3.4 Μονάδα δημιουργίας προφίλ χρηστών

Εφόσον έχει γίνει ο υπολογισμός συχνοτήτων για τα περιεχόμενα που έχει επισκεφθεί ο χρήστης τώρα θα πρέπει να δημιουργηθεί το αντίστοιχο προφίλ το οποίο θα συμβάλλει στην σύσταση νέου περιεχομένου. Ας δούμε για παράδειγμα πως δουλεύει το σύστημα συστάσεων για την περίπτωση των παραλιών. Αρχικά δημιουργείται ένας πίνακας ο οποίος αντιπροσωπεύει την συχνότητα προτίμησης του περιεχομένου από τον χρήστη. Έστω για παράδειγμα ότι στον Πίνακα 7.1 ο χρήστης έχει επισκεφθεί μέσω τις εφαρμογής τις παραλίες Πάνορμος, Μηλιά, Καστανή με συχνότητα επίσκεψης 2, 10 και 8 αντίστοιχα. Στην συνέχεια θα κατασκευαστεί ο πίνακας των χαρακτηριστικών των παραλιών. Στον πίνακα αυτό χρησιμοποιούμε τα σύνολα των χαρακτηριστικών των παραλιών ως στήλες και περιγράφουμε τις παραλίες σαν εγγραφές στον πίνακα. Όσες παραλίες διαθέτουν αυτό το χαρακτηριστικό λαμβάνουν την τιμή 1 ενώ σε αντίθετη περίπτωση λαμβάνουν το 0.

Στις παραλίες τα κυριότερα χαρακτηριστικά είναι:

- Είδος εδάφους (Αμμώδες ή Πετρώδες)
- Οργάνωση παραλίες
- Πρόσβαση στην παραλία με σκάφος ή χωρίς

Προτιμήσεις χρηστών

Παραλίες	Προτίμηση χρήστη
Πάνορμος	2
Μηλιά	10
Καστανή	8

Πίνακας 7.1 Πίνακας προτιμήσεων του χρήστη

Οπότε για τις παραλίες Πάνορμος, Μηλιά, Καστανή δημιουργείται ο πίνακας που παρουσιάζεται στον Πίνακα 7.2. Αφού δημιουργηθούν και οι δύο πίνακες γίνεται πολλαπλασιασμός των δύο πινάκων (προτιμήσεις χρήστη & πίνακας χαρακτηριστικών) και προκύπτει ο σταθμισμένος πίνακα συνόλου χαρακτηριστικών Πίνακας 7.3. Αυτό ο πίνακας ουσιαστικά βαθμολογεί – παρουσιάζει την βαρύτητα που έχει κάθε χαρακτηριστικό κατά την επιλογή περιεχομένου από τον χρήστη.

Πίνακας παραλιών

	Οργανωμένη	Πετρώδες	Αμμώδες	Πρόσβαση με βάρκα
Πάνορμος	0	1	1	0
Μηλιά	1	1	1	1
Καστανή	1	0	1	0

Πίνακας 7.2 Πίνακας χαρακτηριστικών περιεχομένου

Σταθμισμένος πίνακας παραλιών

	Οργανωμένη	Πετρώδες	Αμμώδες	Πρόσβαση με βάρκα
Πάνορμος	0	2	2	0
Μηλιά	10	10	10	10
Καστανή	8	0	8	0

Πίνακας 7.3 Πίνακας σταθμισμένων χαρακτηριστικών

Τώρα έχοντας δημιουργήσει τον πίνακα των σταθμισμένων χαρακτηριστικών μπορούμε να διαμορφώσουμε το προφίλ του ενεργού χρήστη. Ουσιαστικά προσθέτουμε τα σταθμισμένα είδη και στη συνέχεια τα κανονικοποιούμε για να βρούμε το προφίλ του χρήστη. Το αποτέλεσμα της ενέργειας αυτής είναι ο πίνακας στον Πίνακα 7.4. Βλέποντας τα αποτελέσματα του πίνακα βλέπουμε ότι ο χρήστης προτιμάει περισσότερο αμμώδες παραλίες και στην συνέχεια οργανωμένες.

Προφίλ χρήστη

	Οργανωμένη	Πετρώδες	Αμμώδες	Πρόσβαση με βάρκα
Χρήστης	0,3	0,2	0,33	0,16

Πίνακας 7.4 Προφίλ χρήστη

7.3.5 Μονάδα δημιουργίας συστάσεων

Η μονάδα δημιουργίας συστάσεων εκτελείται αφού όλες οι παραπάνω διαδικασίες έχουν ολοκληρωθεί με επιτυχία. Στην μονάδα αυτή δίνεται σαν είσοδος το προφίλ χρήστη που έχει διαμορφωθεί σε προηγούμενο στάδιο καθώς και το υπόλοιπο περιεχόμενο που δεν έχει επιλέξει ο χρήστης. Σκοπός αυτής της μονάδας είναι να βρει ποιο από αυτό το περιεχόμενο είναι καταλληλότερο για να συσταθεί στον χρήστη. Προκειμένου να ξεκινήσει αυτή η διαδικασία θα πρέπει να δημιουργηθεί ο πίνακας χαρακτηριστικών του υποψήφιου περιεχομένου. Στο παράδειγμα των παραλιών έχουμε τις παραλίες Χόβολο,

Έλιος, Βελανιό τις οποίες θα ελέγξουμε αν συμπίπτουν με τις προτιμήσεις του χρήστη. Τα χαρακτηριστικά των παραλιών αυτών παρουσιάζονται στον Πίνακα 7.5.

Πίνακας υποψήφιων παραλιών

	Οργανωμένη	Πετρώδες	Αμμώδες	Πρόσβαση με βάρκα
Χόβολο	1	1	0	0
Έλιος	1	0	1	0
Βελανιό	1	0	1	0

Πίνακας 7.5 Πίνακας χαρακτηριστικών υποψήφιων παραλιών

Στην συνέχεια πολλαπλασιάζουμε τον πίνακα προφίλ χρήστη που έχουμε δημιουργήσει στο προηγούμενο στάδιο με τον πίνακα χαρακτηριστικών υποψήφιων παραλιών. Από αυτή την διαδικασία προκύπτει ο πίνακας σταθμισμένων παραλιών Πίνακας 7.6 ο οποίος δείχνει την βαρύτητα κάθε είδους σε σχέση με το προφίλ χρήστη.

Σταθμισμένος πίνακας παραλιών

	Οργανωμένη	Πετρώδες	Αμμώδες	Πρόσβαση με βάρκα
Χόβολο	0,3	0,2	0	0
Έλιος	0	0	0,33	0
Βελανιό	0,3	0	0,33	0

Πίνακας 7.6 Σταθμισμένος πίνακας υποψήφιων παραλιών

Τέλος υπολογίζεται το άθροισμα χαρακτηριστικών κάθε παραλίας – γραμμής για να πάρουμε το πιθανό επίπεδο ενδιαφέροντος του ενεργού χρήστη για αυτές τις τρεις παραλίες. Ουσιαστικά πρόκειται για την λίστα συστάσεων την οποία γίνεται ταξινόμηση για να κατατάξουμε τις παραλίες και να τις προτείνουμε στον χρήστη. Ο Πίνακας 7.7 παρουσιάζει τα πιθανά επίπεδα ενδιαφέροντος. Στον παρακάτω πίνακα είναι προφανές

ότι η πιο πιθανή παραλία ενδιαφέροντος είναι το Βελανιό με ποσοστό ενδιαφέροντος 0,63 ενώ στην συνέχεια ακολουθεί η παραλία Χόβολο με 0,56.

Πιθανό επίπεδο ενδιαφέροντος

	Επίπεδο Ενδιαφέροντος
Χόβολο	0,56
Έλιος	0,33
Βελανιό	0,63

Πίνακας 7.7 Πίνακας πιθανού ενδιαφέροντος

Σε αυτή τη μονάδα χρησιμοποιείται ο αλγόριθμος Random Forest (RF) με σκοπό την εξαγωγή των συστάσεων για τους χρήστες. Δεδομένου ότι ο αλγόριθμος RF βασίζεται σε ορισμένες παραμέτρους που δίνονται σαν όρισμα στον αλγόριθμο και επηρεάζουν τον τρόπο εκτέλεσης του, πραγματοποιήθηκαν ένα σύνολο από πειράματα με σκοπό την εύρεση των κατάλληλων παραμέτρων του αλγορίθμου RF οι οποίες παρουσιάζονται στην συνέχεια.

Για την εκτέλεση του Random Forest έχουν ορισθεί οι παρακάτω παράμετροι εκτέλεσης:

- **N_estimatorsint (Συνολικός αριθμός δέντρων):** 100
- **Max_depthint (Συνολικό βάθος):** None
- **Min_samples_splitint (Ελάχιστο βάθος δέντρων):** 2

```python
def classify(data, st):
    date = datetime.now().strftime("%Y_%m_%d")

    date_encoded = pd.get_dummies(data['date'], prefix='date')

    data_encoded = pd.concat([data.drop('date', axis=1), date_encoded], axis=1)

    X = data_encoded.drop('class', axis=1)
    y = data_encoded['class']

    results = []

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    RandomForest = Results()
    RandomForest.model = "Random Forest Classifier"

    classifier = RandomForestClassifier()
    classifier.fit(X_train, y_train)

    y_pred = classifier.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    RandomForest.accuracy = accuracy
    for i in range(5):  # Print the first 5 predictions
        RandomForest.predictionResults += "Prediction: " + str(y_pred[i] + "\tTarget: " + str(y_test.iloc[i]) + "\n")
    dump(classifier, f'weights/RandomForestWeights.joblib')
    dump(classifier, f'backupFiles/BackupRandomForestWeights_{date}.joblib')

    st.write("## " + RandomForest.model + " Results\n")
    st.write("### Accuracy: ", RandomForest.accuracy)
    st.write("### Examples:\n")
    for i in range(5):  # Print the first 5 predictions
        st.write(f"Prediction:  {y_pred[i]} \t|\t Target: {y_test.iloc[i]} \n")
```

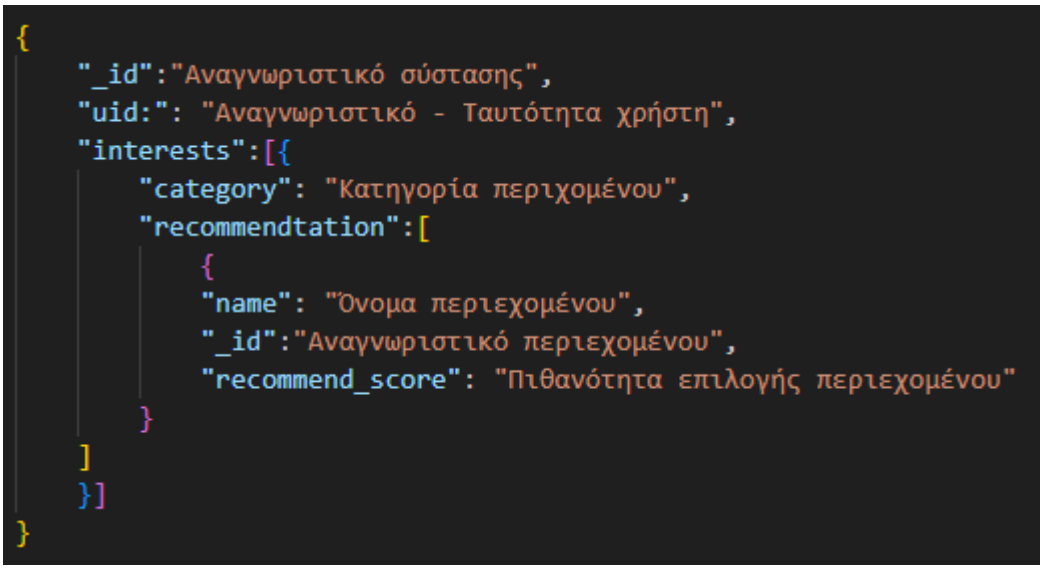
Πίνακας 7.8 Η συνάρτηση classify χρησιμοποιείται για την κατηγοριοποίηση των δεδομένων κάνοντας χρήση τον αλγόριθμο Random Forest

7.3.6 Μονάδα παρουσίασης συστάσεων

Το τελευταίο βήμα – μονάδα του συστήματος συστάσεων είναι η παρουσίαση των συστάσεων που έχουν προκύψει. Μέχρι τώρα όποιες συστάσεις έχουν πραγματοποιηθεί παρουσιάζονται μόνο στον διακομιστή και με κάποιο τρόπο θα πρέπει να είναι προσβάσιμες και ορατές από τους χρήστες. Για αυτό τον σκοπό γίνεται αποθήκευση των δεδομένων σύστασης στην βάση δεδομένων. Στην βάση MongoDB έχει δημιουργηθεί ένα collection με το όνομα predictions. Αρχικά τα δεδομένα διαμορφώνονται σε μία συγκεκριμένη μορφή κατάλληλη για την αποθήκευση τους στην βάση δεδομένων.

Το αντικείμενο που δημιουργείται για την αποθήκευση στην βάση δεδομένων διαθέτει τρία στοιχεία – κλειδιά. Το στοιχείο _id το οποίο περιέχει το μοναδικό αναγνωριστικό της συγκεκριμένης σύστασης. Το στοιχείο uid το οποίο είναι το αναγνωριστικό του χρήστη προέρχεται από την συσκευή του και είναι μοναδικό για κάθε χρήστη. Επιπλέον υπάρχει και το στοιχείο interests και περιέχει ένα πίνακα αντικειμένων τα οποία είναι πιθανά να επιλέξει ο χρήστης και τα οποία του προτείνονται. Κάθε αντικείμενο αναπαριστά μια σύσταση σε μία συγκεκριμένη κατηγορία της εφαρμογής όπως δραστηριότητες, παραλία ή φαγητό. Πιο συγκεκριμένα το αντικείμενο σύστασης περιέχει

το όνομα και τον αναγνωριστικό του περιεχομένου καθώς και το πιθανότητα πρότασης που προκύπτει. Στην Εικόνα 7.6 προβάλλεται η βασική δομή του αντικειμένου συστάσεων.



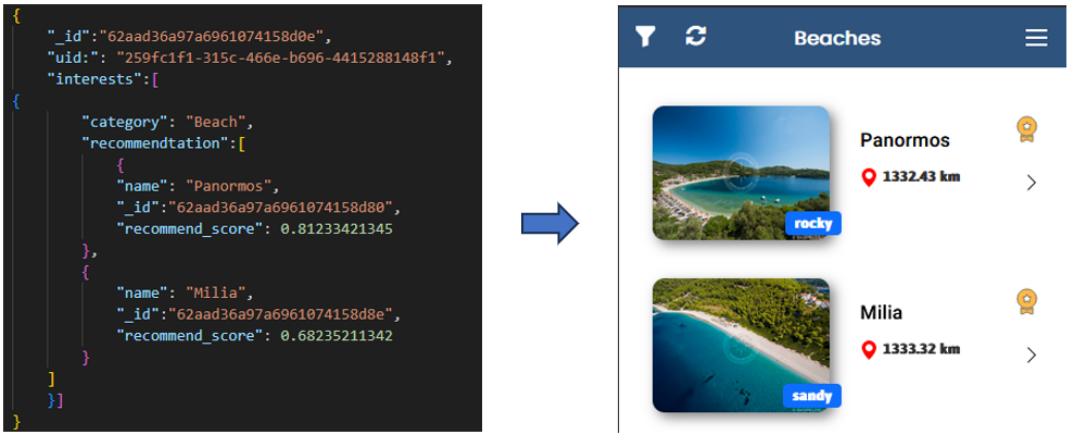
Εικόνα 7.7 Η κατάλληλη μορφή δεδομένων για την αποθήκευση τους στην βάση δεδομένων.

Αν ο χρήστης για τον οποίο δημιουργούνται οι συστάσεις αποτελεί νέα εγγραφή για το collection prediction τότε απλώς αποθηκεύεται η εγγραφή του στην βάση. Στην περίπτωση που ο χρήστης δεν είναι καινούργιος αλλά έχουν ξανά δημιουργηθεί συστάσεις για αυτόν γίνεται εύρεση του χρήστη και ανανέωση των υπαρχόντων δεδομένων με τα νέα δεδομένα συστάσεων.

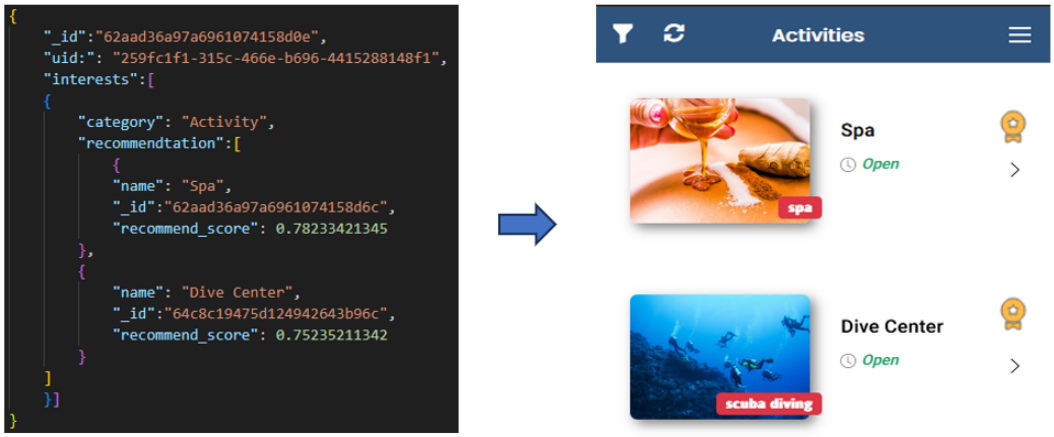
Τα δεδομένα αυτά στην συνέχεια προβάλλονται στην εφαρμογή. Ο χρήστης με το που ανοίξει την εφαρμογή κάνει ένα αίτημα στην βάση δεδομένων στέλνοντας το uuid του προκειμένου να λάβει τις διαθέσιμες συστάσεις (αν υπάρχουν) για αυτόν.

Μόλις η εφαρμογή λάβει αυτά τα δεδομένα γίνεται η προβολή τους στις λίστες περιεχομένων της εφαρμογής. Η προβολή των συστάσεων γίνεται στις κεντρικές λίστες της εφαρμογής δηλαδή στην λίστα φαγητών – κρασιών – κοκτέιλ, στην λίστα δραστηριοτήτων και στην λίστα παραλιών. Ο διαχωρισμός των απλών εγγραφών με αυτές των προτάσεων γίνεται με την χρήση ενός ευδιάκριτου συμβόλου.

Εικόνα 7.8 Παράδειγμα σύστασης στην κατηγορία φαγητών.



Εικόνα 7.9 Παράδειγμα σύστασης στην κατηγορία παραλιών



Εικόνα 7.10 Παράδειγμα σύστασης στην κατηγορία δραστηριοτήτων.

8 Συμπεράσματα και μελλοντικές προσθήκες

Η παρούσα εργασία είχε ως σκοπό την μελέτη των συστημάτων συστάσεων και πιο συγκεκριμένα την χρήση αυτών των συστημάτων στον τομέα του τουρισμού, μέσα από μια θεωρητική προσέγγιση καθώς επίσης και με την ανάπτυξη μιας εφαρμογής η οποία κάνει χρήση ενός συστήματος συστάσεων. Αρχικά έγινε αναφορά στα συστήματα συστάσεων στα δομικά μέρη τους καθώς και στα είδη των συστημάτων αυτών.

Στην εφαρμογή που δημιουργήθηκε στα πλαίσια της εργασίας χρησιμοποιήθηκαν ορισμένοι αλγόριθμοι και τεχνικές για την δημιουργία των προφίλ των χρηστών καθώς και των συστάσεων τους. Αρχικά έγινε δοκιμή ορισμένων αλγορίθμων κατηγοριοποίησης για την δημιουργία των προφίλ χρηστών ενώ έγιναν και δοκιμές μέσω στατιστικής ανάλυσης. Αφού δημιουργήθηκαν τα προφίλ και βρέθηκαν οι προτιμήσεις των χρηστών έγινε αντιστοιχία των χαρακτηριστικών των αντικειμένων με τις προτιμήσεις του χρήστη ώστε να προταθούν και άλλα παρόμοια αντικείμενα με τα ίδια χαρακτηριστικά. Αποτέλεσμα αυτών των ενεργειών είναι η δημιουργία ενός συστήματος συστάσεων βασιζόμενα στο περιεχόμενο των αντικειμένων (Content Based Filtering).

Όσο αφορά τα εργαλεία που χρησιμοποιήθηκαν από την πλευρά του frontend έγινε χρήση του framework Angular σε συνδυασμό με τις γλώσσες Typescript, HTML/CSS. Από την πλευρά του backend η εφαρμογή βασίστηκε στο πλαίσιο Node.js ενώ τα δεδομένα αποθηκεύονται σε μια βάση MongoDB. Για το σύστημα συστάσεων χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python σε συνδυασμό με ορισμένες βιβλιοθήκες όπως η Pandas και η Scikit – learn η οποίες είναι υπεύθυνες για την διαχείριση μεγάλων δεδομένων και εφαρμογή αλγορίθμων μηχανικής μάθησης αντίστοιχα. Μετά το πέρας της εφαρμογής έχουν γίνει κατανοητά σε ικανοποιητικό βαθμό πλέον η λειτουργία μιας διαδικτυακής εφαρμογής τόσο από την πλευρά του frontend όσο και από την πλευρά του backend τα οποία συνεργάζονται μεταξύ τους καθώς επίσης κατανοητό έγινε και η δημιουργία του συστήματος συστάσεων το οποίο τροφοδοτεί την εφαρμογή και παρέχει στους χρήστες εξατομικευμένες συστάσεις.

Οι χρήστες μέσω των κινητών τηλεφώνων τους έχουν την δυνατότητα να χρησιμοποιήσουν την εφαρμογή μέσω ενός απλού περιηγητή όπως ο Google Chrome, ο Mozilla, ο Safari και άλλοι. Επιπλέον δίνεται η δυνατότητα στους χρήστες μέσω του περιηγητή να κάνουν λήψη της εφαρμογής στις κινητές συσκευές τους.

Η επιστήμη της πληροφορικής είναι ένας κλάδος ο οποίος βρίσκει εφαρμογή σε πολλούς τομείς της καθημερινής ζωής. Ο κλάδος αυτός συνεχώς εξελίσσεται και βελτιώνεται με νέες μεθόδους και τεχνολογίες. Κατά αυτό τον τρόπο και η παρούσα εργασία και υλοποίηση υπάρχουν ορισμένες αλλαγές και προσθήκες που θα μπορούσαν να πραγματοποιηθούν μελλοντικά για την βελτίωση της.

Παρακάτω αναφέρονται συνοπτικά μερικές αλλαγές – βελτιώσεις:

- **Διασύνδεση με άλλες πλατφόρμες:** Μία σημαντική προσθήκη είναι η διασύνδεση με άλλες πλατφόρμες κοινωνικής δικτύωσης όπως το FaceBook μέσα από την οποία θα συλλέγονται δεδομένα όπως δημογραφικά στοιχεία και άλλα τα οποία θα χρησιμοποιούνται κατά την δημιουργία των προφίλ χρηστών και θα συμβάλλουν θετικά στην δημιουργία των συστάσεων.
- **Αξιοποίηση των ερωτηματολίων:** Προς το παρόν το σύστημα συστάσεων βασίζεται μόνο στην αλληλεπίδραση των χρηστών με την εφαρμογή (clickstreams). Μια πιθανή βελτίωση είναι η χρήση των ερωτηματολογίων και κυρίως του ερωτηματολογίου συστάσεων το οποίο θα βοηθήσει στην δημιουργία ενός αρχικού προφίλ για τους νέους χρήστες καταπολεμώντας έτσι προβλήματα που έχουν τα συστήματα συστάσεων όπως το πρόβλημα της ψυχρής εκκίνησης (cold start).
- **Προσθήκη και αξιοποίηση αξιολογήσεων:** Μια πιθανή βελτίωση είναι η προσθήκη της δυνατότητας αξιολογήσεων στους χρήστες. Τα δεδομένα των αξιολογήσεων θα μπορούσαν να συμβάλλουν και αυτά στον εμπλουτισμό του προφίλ ενός χρήστη.
- **Συνδυασμός των μεθόδων συστάσεων:** Άλλη μία πιθανή βελτίωση θα ήταν η δημιουργία και η εφαρμογή ενός υβριδικού συστήματος συστάσεων το οποίο θα αντιμετωπίζει τις αδυναμίες που μπορεί να προκύψουν κατά τη χρήση των μεθόδων ξεχωριστά καθώς επίσης μπορεί να είναι πιο αποτελεσματική.

ΠΑΡΑΡΤΗΜΑ Α - ΕΡΩΤΗΜΑΤΟΛΟΓΙΟ

Recommendation Survey

Παρακάτω παρατίθεται οι ερωτήσεις που περιέχονται στο Recommendation Survey καθώς και οι επιλογές τους:

1. Age?
   - < 25
   - 25 – 35
   - 36 – 45
   - 46 – 55
   - > 55
2. Gender?
   - Female
   - Male
   - Third gender
3. Country of origin?
   - List of countries
4. Marital status?
   - Single
   - Married
5. Type of tourism?
   - Recreational
   - Historical
   - Cultural
   - Other
6. Your budget level for your holidays (average)?
   - 1 (< 800 euro)
   - 2
   - 3 (3.000 euro)
   - 4
   - 5 (> 10.000 euro)
7. How did you plan your holidays?

- By yourself
- Through travel agent

8. All information was selected for the location via?
   - TripAdvisor
   - Booking
   - Trivago
   - Other
9. Types of museums?
   - Archaeological
   - Folklore
   - History
   - Other
10. Culture events?
    - Theater
    - Festival
    - Cultural Activities
    - Other
11. Boat trips and water sports?
    - Diving
    - Boat tours
    - Dolphin watching
    - Other
12. Shopping?
    - Souvenirs
    - Traditional products
    - Clothes / Shoes
    - Other
13. Types of nightlife?
    - Beach bar
    - Bar
    - Cocktail lounge
    - Wine bar
14. Types of food?
    - Local cuisine

- Tavern
- Street food
- Restaurant

15. Coffee & drink?
    - Café
    - Take away
    - Hotel lounge
    - Other
16. Do you like Skopelos island?
    - 1
    - 2
    - 3
    - 4
    - 5

17. Kindness of local people?
    - 1
    - 2
    - 3
    - 4
    - 5
18. Information received before your arrival in your destination?
    - 1
    - 2
    - 3
    - 4
    - 5
19. Accommodation?
    - 1
    - 2
    - 3
    - 4
    - 5
20. Infrastructure

- 1
- 2
- 3
- 4
- 5

21. Customer service
    - 1
    - 2
    - 3
    - 4
    - 5
22. Wi – Fi
    - 1
    - 2
    - 3
    - 4
    - 5
23. Activities
    - 1
    - 2
    - 3
    - 4
    - 5
24. Would you visit Skopelos island again?
    - 1
    - 2
    - 3
    - 4
    - 5
25. Please, note any suggestion or idea that will make your next visit better?
    - Free text

ΒΙΒΛΙΟΓΡΑΦΙΑ

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). "Recommender systems survey". Knowledge-Based Systems, pp. 109-132.

Burkov, A., 2019. The Hundred-Page Machine Learning Book by Andriy Burkov.

Çano, E., Morisio, M., 2017. Hybrid recommender systems: A systematic literature review. IDA 21, 1487–1524. https://doi.org/10.3233/IDA-163209

Charbuty, B., Abdulazeez, A., 2021. Classification Based on Decision Tree Algorithm for Machine Learning. JASTT 2, 20–28. https://doi.org/10.38094/jastt20165

Charu C. Aggarwal, 2016, Recommender Systems, The Textbook, Springer Nature Switzerland AG.

Comparative Analysis of Data Visualization Libraries Matplotlib and Seaborn in Python, 2021. . IJATCSE 10, 277–281. https://doi.org/10.30534/ijatcse/2021/391012021

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., Sampath, D., 2010. The YouTube video recommendation system, in: Proceedings of the Fourth ACM Conference on Recommender Systems. Presented at the RecSys '10: Fourth ACM Conference on Recommender Systems, ACM, Barcelona Spain, pp. 293–296. https://doi.org/10.1145/1864708.1864770

Durner, D., Leis, V., Neumann, T., 2021. JSON Tiles: Fast Analytics on Semi-Structured Data, in: Proceedings of the 2021 International Conference on Management of Data. Presented at the SIGMOD/PODS '21: International Conference on Management of Data, ACM, Virtual Event China, pp. 445–458. https://doi.org/10.1145/3448016.3452809

Effendy, F., Taufik, Adhilaksono, B., 2021. Performance Comparison of Web Backend and Database: A Case Study of Node.JS, Golang and MySQL,

Mongo DB. RACSC 14, 1955–1961. https://doi.org/10.2174/2666255813666191219104133

Eke, C., Norman, A., Shuib, L. and Nweke, H., 2019. A Survey of User Profiling: State-of-the-Art, Challenges, and Solutions. IEEE Access, 7, pp.144907-144924.

Flew, T. (2008). New Media: An Introduction (3rd edition). Melbourne: Oxford University Press, 19.

Gauch, S., Speretta, M., Chandramouli, A. and Micarelli, A., 2007. User Profiles for Personalized Information Access. The Adaptive Web, pp.54-89.

Geetha, G., Mittal, M., Prasad, K.M., Ponsam, J.G., 2022. Interpretation and Analysis of Angular Framework, in: 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS). Presented at the 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), IEEE, Chennai, India, pp. 1–6. https://doi.org/10.1109/ICPECTS56089.2022.10047474

Han, J., Kamber, M. and Pei, J. (2016) Data Mining: Concepts and Techniques. Waltham, Ma: Morgan Kaufmann.

Harris, C.R., Millman, K.J., Van Der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., Van Kerkwijk, M.H., Brett, M., Haldane, A., Del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Heidenreich, H. (2018) Machine learning for the average person: What are the types of machine, Hunter Heidenreich. Available at: http://hunterheidenreich.com/blog/breaking_down_ml_for_the_average_person/

Jacksi, K., Abass, S.M., 2019. Development History Of The World Wide Web 8.

Javed, U., Shaukat, K., A. Hameed, I., Iqbal, F., Mahboob Alam, T., Luo, S., 2021. A Review of Content-Based and Context-Based Recommendation Systems. Int. J. Emerg. Technol. Learn. 16, 274. https://doi.org/10.3991/ijet.v16i03.18851

Kim, B.-K. (2005). "Internationalising the Internet, the Co-evolution of Influence and Technology". Edward Elgar, 51–55.

Ko, H., Lee, S., Park, Y., Choi, A., 2022. A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields. Electronics 11, 141. https://doi.org/10.3390/electronics11010141

Koren, Y., 2008, August. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 426-434).

Lagoumintzis, G., Vlachopoulos, G. and Koutsogiannis, K., 2015. Μεθοδολογία Της Έρευνας Στις Επιστήμες Υγείας. Kallipos, Open Academic Editions, pp.41 - 51.

Leach, P., Mealling, M., Salz, R., 2005. A Universally Unique IDentifier (UUID) URN Namespace (No. RFC4122). RFC Editor. https://doi.org/10.17487/rfc4122

Li, J., Xu, L., Tang, L., Wang, S. and Li, L., 2018. Big data in tourism research: A literature review. Tourism Management, 68, pp.301-323.

Matallah, H., Belalem, G., Bouamrane, K., 2021. Comparative Study Between the MySQL Relational Database and the MongoDB NoSQL Database: International Journal of Software Science and Computational Intelligence 13, 38–63. https://doi.org/10.4018/IJSSCI.2021070104

Montgomery, A.L., Li, S., Srinivasan, K., Liechty, J.C., 2004. Modeling Online Browsing and Path Analysis Using Clickstream Data. Marketing Science 23, 579–595. https://doi.org/10.1287/mksc.1040.0073

Nadee, W., 2016. Modeling User Profiles for Recommender Systems. Doctor of Philosophy. Queensland University of Technology.

Najmani, K., Ajallouda, L., Benlahmar, E., Sael, N. and Zellou, A., 2022. Offline and Online Evaluation for Recommender Systems. *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)*,.

Neha, K., Reddy, M.Y., 2020. A Study On Applications Of Data Mining 9.

Parmar, A., Katariya, R., Patel, V., 2019. A Review on Random Forest: An Ensemble Classifier, in: Hemanth, J., Fernando, X., Lafata, P., Baig, Z. (Eds.), International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018, Lecture Notes on Data Engineering and Communications Technologies. Springer International Publishing, Cham, pp. 758–763. https://doi.org/10.1007/978-3-030-03146-6_86

Patel, B., Desai, P., Panchal, U., 2017. Methods of recommender system: A review, in: 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS). Presented at the 2017 4th International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), IEEE, Coimbatore, pp. 1–4. https://doi.org/10.1109/ICIIECS.2017.8275856

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (2011). Recommender Systems Handbook. New York: Springer.

Roy, S., Daniel, C., Agrawal, M., 2023. Chapter 12 Introduction to HTML and Website Development, in: Fundamentals of Information Technology. University of South Florida Library. https://doi.org/10.5038/TKOC5299

Seth, R., Sharaff, A., 2022. A Comparative Overview of Hybrid Recommender Systems: Review, Challenges, and Prospects, in: Raja, R., Nagwanshi, K.K., Kumar, S., Laxmi, K.R. (Eds.), Data Mining and Machine Learning Applications. Wiley, pp. 57–98. https://doi.org/10.1002/9781119792529.ch3

Shah Khusro, Zafar Ali and Irfan Ullah, 2016, Recommender Systems: Issues, Challenges, and Research Opportunities, Springer Science+Business Media Singapore.

Steck, H., Baltrunas, L., Elahi, E., Liang, D., Raimond, Y., Basilico, J., 2021. Deep learning for recommender systems: A Netflix case study. AI Magazine 42, 7–18. https://doi.org/10.1609/aimag.v42i3.18140

Verykios, V., Kagklis, V., & Stavropoulos, I. (2015). Εισαγωγή στην Εξόρυξη Δεδομένων [Chapter]. In Verykios, V., Kagklis, V., & Stavropoulos, E. 2015. Η επιστήμη των δεδομένων μέσα από τη γλώσσα R [Undergraduate textbook]. Kallipos, Open Academic Editions. chapter 1. http://hdl.handle.net/11419/2966

Xiaoyuan, S., & Taghi, K. M. (2009, January). A Survey of Collaborative Filtering Techniques. Advances in Artificial Intelligence, 2009, pp. 1-20.

[Οπισθόφυλλο. Κενή σελίδα]