ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Ανάλυση Στοιχείων Χρήσης Ψηφιακού Αποθετηρίου

Ψηφιακό Αποθετήριο ΚΑΛΛΙΠΟΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΔΟΝΤΗ Α. ΕΙΡΗΝΗΣ

Επιβλέπων

Μήτρου Νικόλαος

Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2024

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Ανάλυση Στοιχείων Χρήσης Ψηφιακού Αποθετηρίου

Ψηφιακό Αποθετήριο ΚΑΛΛΙΠΟΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΔΟΝΤΗ Α. ΕΙΡΗΝΗΣ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 9η Οκτωβρίου 2024

.....	.....	.....
Μήτρου Νικόλαος	Παπαβασιλείου Συμεών	Συκάς Ευστάθιος
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2024

.....

Δόντη Α. Ειρήνη

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η ραγδαία αύξηση δεδομένων εντείνει την ανάγκη για τη δημιουργία ψηφιακών αποθετηρίων, τα οποία προσφέρουν στους χρήστες περιεχόμενο σε οργανωμένη μορφή. Η αποτελεσματική οργάνωση και διαχείριση των δεδομένων που περιέχει ένα ψηφιακό αποθετήριο αποτελεί σημαντική προϋπόθεση για την εξασφάλιση της βέλτιστης λειτουργικότητας και αξιοποίησής του. Ένας τρόπος για να επιτευχθεί αυτό, πέρα από τη διαμόρφωση της ιεραρχικής δομής, είναι η συνεχής παρακολούθηση και ανάλυση των στοιχείων χρήσης του αποθετηρίου.

Στην παρούσα Διπλωματική Εργασία, πραγματοποιείται ανάλυση στοιχείων χρήσης και δεδομένων, ώστε να ανιχνευτούν συμπεριφορές των χρηστών, καθώς και τυχόν ελλείψεις στο περιεχόμενο ενός ψηφιακού αποθετηρίου. Επίσης, παρουσιάζονται διαδικασίες πρόβλεψης για τον αριθμό των νέων χρηστών που εισέρχονται στο αποθετήριο. Συγκεκριμένα, εφαρμόζονται μέθοδοι πρόβλεψης χρονοσειρών, με αξιοποίηση τεχνικών της Στατιστικής και της Μηχανικής Μάθησης, για να εκτιμηθούν μελλοντικές μεταβολές στην κίνηση του αποθετηρίου. Στη συνέχεια, τα μοντέλα που προκύπτουν αξιολογούνται με κατάλληλες μετρικές και συγκρίνονται, ώστε να επιλεχθεί το πιο αποδοτικό.

Λέξεις-Κλειδιά: Ψηφιακά Αποθετήρια, Ανάλυση Δεδομένων, Πρόβλεψη Δεδομένων, Στατιστική, Μηχανική Μάθηση, Μοντέλα

Abstract

The rapid growth of data intensifies the need to create digital repositories that enable users to discover and use content in organized formats. The effective organization and management of the data in a digital repository is an important condition for ensuring its optimal functionality and utilization. Apart from setting up the hierarchical structure, one way to achieve this is to monitor and analyze repository usage data continuously.

In this Thesis, an analysis of usage and data elements is carried out, to detect user behaviors in addition to deficiencies in the content of the digital repository. Also, forecasting procedures for the number of new users entering the repository are presented. Specifically, time series forecasting methods are applied, utilizing Statistical and Machine Learning techniques, to analyze future changes in the use of the repository. The resulting models are then evaluated with appropriate metrics and compared to select the most efficient one.

Keywords: Digital Repositories, Data Analysis, Data Prediction, Statistics, Machine Learning, Models

Ευχαριστίες

Στο σημείο αυτό, θέλω να απονείμω θερμές ευχαριστίες στον επιβλέποντα καθηγητή κ. Νικόλαο Μήτρου, ο οποίος με εμπιστεύτηκε και με καθοδήγησε καθ' όλη τη διάρκεια εκπόνησης της Διπλωματικής Εργασίας.

Επιπλέον, θέλω να ευχαριστήσω από καρδιάς τον Ιωάννη Μήτρου και τον Βασίλειο Ξηρό για τη διαρκή βοήθεια και τις εύστοχες παρεμβάσεις τους, οι οποίες συνέβαλαν στη βελτίωση της εργασίας μου.

Θα ήθελα, επίσης, να ευχαριστήσω την κα. Σταματίνα Κουτσιλέου για τη γλωσσική επιμέλεια της Διπλωματικής Εργασίας. Η συμβολή της στη βελτίωση του κειμένου υπήρξε πολύτιμη για το τελικό αποτέλεσμα.

Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στους γονείς μου, Αντώνη και Μαρία, καθώς και στην αδερφή μου Ελένη–Μαρία, για την πίστη τους στις δυνατότητές μου και τη στήριξή τους σε κάθε επιλογή μου.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

1 Εισαγωγή

1.1 Κίνητρο

Η τεχνολογική ανάπτυξη των τελευταίων ετών, κατά κύριο λόγο, έχει βελτιώσει τις συνθήκες και τις δυνατότητες σε όλους τους τομείς της ανθρώπινης δραστηριότητας. Για παράδειγμα, η μετάβαση από την εποχή των παραδοσιακών συγγραμμάτων στα ψηφιακά συγγράμματα αποτελεί αδιαμφισβήτητα ένα επίτευγμα της τεχνολογικής εξέλιξης.

Ένα από τα κύρια πλεονεκτήματα των ψηφιακών συγγραμμάτων είναι η άμεση πρόσβαση σε αυτά μέσω διαδικτύου, χωρίς χωρικούς και χρονικούς περιορισμούς, αλλά και χωρίς οικονομική επιβάρυνση των χρηστών στις περιπτώσεις που αυτά διατίθενται με ελεύθερες άδειες χρήσης. Επίσης, πολλά ψηφιακά συγγράμματα παρέχουν εμπλουτισμένο και διαδραστικό περιεχόμενο που ενισχύει την κατανόησή του σε σχέση με τα συμβατικά έντυπα βιβλία. Τα παραπάνω αποτελούν λόγους ώστε να προσελκύονται όλο και περισσότεροι χρήστες σε ψηφιακά αποθετήρια συγγραμμάτων. Η αυξανόμενη ζήτηση για τα περιεχόμενα των ψηφιακών συγγραμμάτων προκαλεί την ανάγκη για σωστή οργάνωση ενός αποθετηρίου. Η καλή διάρθρωση ενός αποθετηρίου θα διευκολύνει τους χρήστες να εντοπίζουν τα περιεχόμενα που τους ενδιαφέρουν πιο γρήγορα. Θα προσφέρεται, με αυτόν τον τρόπο, μία πιο ευχάριστη εμπειρία στους αναγνώστες, βελτιώνοντας την εμπιστοσύνη και την προτίμηση στην πλατφόρμα.

Στην περίπτωση που ένα αποθετήριο είναι οργανωμένο σύμφωνα με τις προτιμήσεις των χρηστών αυξάνεται η πιθανότητα να επιστρέψει ένας χρήστης μελλοντικά ή να το προτείνει σε άλλους. Με την ανάλυση των ήδη δημοσιευμένων στοιχείων ενός αποθετηρίου, εξερευνώνται τυχόν ελλείψεις σε συγγράμματα συγκεκριμένων κατηγοριών. Επίσης, η ανάλυση στοιχείων χρήσης ενός αποθετηρίου μπορεί να βοηθήσει στην αξιολόγηση του πώς τα δεδομένα χρηστών επηρεάζουν και επηρεάζονται από το περιεχόμενο της σελίδας. Αποτέλεσμα αυτού είναι η κατανόηση της δραστηριότητας των χρηστών και η εξαγωγή μοτίβων συμπεριφοράς.

Η ανάλυση και η πρόβλεψη χρονοσειρών πάνω σε στοιχεία του αποθετηρίου μπορούν να ενισχύσουν την κατανόηση της εποχικής δραστηριότητας χρηστών, καθώς ανιχνεύονται

στοιχεία τάσης και οι ανάγκες των χρηστών συναρτήσει του χρόνου, επιτρέποντας να προσαρμοστούν καλύτερα το περιεχόμενο και οι υπηρεσίες της πλατφόρμας στη ζητούμενη χρήση τους. Μία εξελιγμένη μέθοδος για πρόβλεψη χρονοσειρών αποτελεί η Μηχανική Μάθηση, η οποία μπορεί να προβλέψει μελλοντικές τιμές μίας μεταβλητής λαμβάνοντας υπόψιν παρελθοντικές τιμές.

Πρακτικά, υπάρχουν προκλήσεις σχετικά με το πόσο αντιπροσωπευτικές είναι οι αναλύσεις. Συγκεκριμένα, μπορεί να μην είναι επαρκές το δείγμα ή να μην αντιπροσωπεύει το σύνολο των χρηστών και συνεπώς είναι εύκολο να προκύψουν λανθασμένα συμπεράσματα. Επίσης, υπάρχουν δεδομένα χρηστών που απουσιάζουν, όπως η ηλικία, το επάγγελμα, και συνεπώς να μην είναι γνωστό πώς αντιπροσωπεύονται τα αποτελέσματα ανάλυσης σε σχέση με τις ομάδες χρηστών. Ωστόσο, η συμπεριφορά των χρηστών μπορεί να εξελιχθεί με τον χρόνο, λόγω αλλαγών προτιμήσεων ή λόγω της αλλαγής κοινωνικών τάσεων, γεγονός που μπορεί να οδηγήσει σε διαρκή αναθεώρηση των συμπερασμάτων ανάλυσης.

Μία λύση είναι η χρήση ευέλικτων μεθόδων ανάλυσης και η συχνή παρακολούθηση των δεδομένων χρήσης του αποθετηρίου. Στην περίπτωση της Μηχανικής Μάθησης, η πρόκληση είναι να αναπτυχθεί ένα περιγραφικό μοντέλο, με σχετικά ακριβείς προβλέψεις και να εξεταστεί με κατάλληλες μετρικές αξιολόγησης.

1.2 Αντικείμενο της Διπλωματικής Εργασίας

Η παρούσα Διπλωματική Εργασία έχει ως αντικείμενο τη μελέτη των στοιχείων χρήσης ενός αποθετηρίου με απώτερο στόχο τη βελτίωση των υποδομών και των υπηρεσιών του.

Στο πρώτο στάδιο της ερευνητικής μελέτης, πραγματοποιείται η ανάλυση στοιχείων η οποία στηρίζεται σε παραδοσιακές τεχνικές στατιστικής. Η ανάλυση στοιχείων συγκεντρώνει το μεγαλύτερο ενδιαφέρον στην παρούσα διπλωματική εργασία, αφού ανιχνεύονται συμπεριφορές των χρηστών κατά την περιήγησή τους στο αποθετήριο, καθώς και ελλείψεις στο περιεχόμενό του. Η σωστή ανάλυση και επεξεργασία στοιχείων καθορίζουν την πορεία της έρευνας, καθώς αναδεικνύονται μοτίβα και αντιφάσεις δεδομένων. Επίσης, αναδεικνύονται συσχετίσεις που μπορεί να υπάρχουν μεταξύ

διαφορετικών μετρικών, οι οποίες είναι πολύ χρήσιμες για τη συνέχεια της πειραματικής διαδικασίας.

Στο δεύτερο στάδιο της ερευνητικής μελέτης, πραγματοποιείται η διαδικασία πρόβλεψης χρονοσειρών σχετικά με τη μεταβλητή η οποία είναι καθοριστική για την επιτυχία ενός αποθετηρίου, δηλαδή το σύνολο νέων χρηστών οι οποίοι εισέρχονται στο αποθετήριο. Η εύρεση μελλοντικών προβλέψεων παρέχει πολύτιμες πληροφορίες για το πώς θα εξελιχθεί η επισκεψιμότητα του αποθετηρίου στο μέλλον. Στη συγκεκριμένη εργασία, καταγράφονται προβλήματα και ανάγκες που ανακύπτουν από την πρόβλεψη, αξιολογώντας την ακρίβεια της διαδικασίας μέσω σχετικών χρονοδιαγραμμάτων.

Στο τρίτο στάδιο της ερευνητικής μελέτης, χρησιμοποιείται η μέθοδος Μηχανικής Μάθησης για πρόβλεψη χρονοσειρών στο σύνολο νέων χρηστών με σκοπό τη σύγκριση της αποδοτικότητας των προβλέψεων και την αξιολόγησή τους. Η διαδικασία περιλαμβάνει την εκπαίδευση του μοντέλου Μηχανικής Μάθησης, χρησιμοποιώντας παρελθοντικά δεδομένα και έπειτα αξιολογείται η απόδοσή του, βάσει δεδομένων που δεν έχουν χρησιμοποιηθεί στη διαδικασία εκπαίδευσης. Επίσης, γίνεται σύγκριση αποδοτικότητας με τα αποτελέσματα που προέκυψαν από το δεύτερο στάδιο της ερευνητικής διαδικασίας.

Συνοψίζοντας, η παρούσα Διπλωματική Εργασία περιλαμβάνει τεχνικές ανάλυσης και πρόβλεψης δεδομένων χρήσης ενός συγκεκριμένου αποθετηρίου (ΚΑΛΛΙΠΟΣ) με τεχνικές στατιστικής, αλλά και με προηγμένες τεχνικές Μηχανικής Μάθησης. Αποτελεί μία έρευνα σχετικά με το πώς οι χρήστες διαχειρίζονται την ποιότητα και την ποσότητα του περιεχομένου και επισημαίνονται ζητήματα που αφορούν την παρουσίαση και την οργάνωση των δεδομένων του συγκεκριμένου αποθετηρίου με σκοπό τη βελτίωση της εμπειρίας χρήστη και την προσαρμογή του περιεχομένου βάσει των προτιμήσεων χρηστών.

1.3 Οργάνωση Τόμου

Η παρούσα Διπλωματική Εργασία αποτελείται από έξι κεφάλαια.

Συγκεκριμένα, στο πρώτο κεφάλαιο πραγματοποιείται μία εισαγωγή σχετικά με το αντικείμενο, το κίνητρο και τους στόχους της εργασίας.

Το δεύτερο κεφάλαιο αποτελεί ο ορισμός και η αναφορά σε παραδείγματα ψηφιακών αποθετηρίων. Επίσης, αναλύονται οι μέθοδοι συλλογής δεδομένων χρήσης ενός ψηφιακού αποθετηρίου, αφού παρουσιαστούν οι πιθανές πληροφορίες που συλλέγονται σχετικά με το πώς οι χρήστες αλληλεπιδρούν με το περιεχόμενο ενός ψηφιακού αποθετηρίου.

Στο τρίτο κεφάλαιο αναπτύσσεται το θέμα της δημοτικότητας δημοσιευμένων τεκμηρίων, τονίζοντας τη σημασία και τις προκλήσεις πρόβλεψης του συγκεκριμένου θέματος. Ακόμη, εξετάζονται οι παράγοντες που επηρεάζουν τη δημοτικότητα των ηλεκτρονικών συγγραμμάτων, οι μεθοδολογίες, καθώς και οι τρόποι αξιολόγησης της πρόβλεψής της.

Στο τέταρτο κεφάλαιο παρουσιάζονται και εξηγούνται δεδομένα χρήσης του αποθετηρίου ΚΑΛΛΙΠΟΣ. Επίσης, αναλύονται οι παράγοντες που επηρεάζουν τη χρήση εκπαιδευτικών ψηφιακών συγγραμμάτων και πραγματοποιείται η επεξεργασία των στοιχείων χρήσης του αποθετηρίου ΚΑΛΛΙΠΟΣ με τη βοήθεια της στατιστικής ανάλυσης. Παράλληλα, τίθενται υπό επεξεργασία δεδομένα των συγγραμμάτων για τυχόν ελλείψεις στο περιεχόμενο του συγκεκριμένου αποθετηρίου και εξετάζεται αν μπορούν να δικαιολογηθούν μοτίβα συμπεριφοράς χρηστών.

Στο πέμπτο κεφάλαιο, πραγματοποιείται η επεξεργασία των στοιχείων χρήσης του αποθετηρίου ΚΑΛΛΙΠΟΣ με τη βοήθεια πρόβλεψης χρονοσειρών. Εφαρμόζονται διάφορες μέθοδοι στατιστικής πρόβλεψης χρονοσειρών και μία μέθοδος πρόβλεψης χρονοσειρών με χρήση Μηχανικής Μάθησης. Γίνεται προσπάθεια βελτίωσης του μοντέλου Μηχανικής Μάθησης που αναπτύσσεται και συγκρίνεται με το αρχικό μοντέλο. Τα συμπεράσματα που προκύπτουν αναλύονται, εξετάζοντας την εκάστοτε διαδικασία με μετρικές αξιολόγησης.

Το έκτο κεφάλαιο αποτελεί τον επίλογο της παρούσας Διπλωματικής Εργασίας. Συγκεκριμένα, συνοψίζονται τα αποτελέσματα και τα συμπεράσματα της εργασίας, προτείνοντας ιδέες, σχετικές με το αντικείμενο εργασίας, για μελλοντική έρευνα.

2 Εισαγωγή στα Ψηφιακά Αποθετήρια και τη Χρήση τους

2.1 Ορισμός και Παραδείγματα Ψηφιακών Αποθετηρίων

Τα ψηφιακά αποθετήρια αποτελούν σημαντικά εργαλεία για την οργάνωση και την αποθήκευση ψηφιακού περιεχομένου. Στο παρόν κεφάλαιο, αναφέρονται ο ορισμός των ψηφιακών αποθετηρίων και οι κύριες κατηγορίες τους. Επίσης, παρουσιάζονται συγκεκριμένα παραδείγματα ψηφιακών αποθετηρίων, εστιάζοντας στον ακαδημαϊκό και τον πολιτιστικό τομέα.

2.1.1 Ορισμός

Ψηφιακό αποθετήριο ονομάζεται η εφαρμογή ή το σύστημα, το οποίο χρησιμεύει στην ηλεκτρονική απόθεση και διάθεση ψηφιακού περιεχομένου, σε μία ή περισσότερες μορφές: κειμένου, ήχου, εικόνας ή και βίντεο. Προσφέρει, μεταξύ άλλων, υπηρεσίες αναζήτησης, πλοήγησης και διαφύλαξης ψηφιακών δεδομένων, αντιστοιχίζοντας μεταδεδομένα και πνευματικά δικαιώματα σύμφωνα με διεθνή πρότυπα [1].

Τα βασικά στοιχεία που διαφοροποιούν ένα ψηφιακό αποθετήριο από άλλες ψηφιακές δομές, σύμφωνα με τους Heery και Anderson [2] είναι τα εξής:

- Το περιεχόμενο κατατίθεται σε ένα αποθετήριο, είτε από τον δημιουργό, είτε από τον ιδιοκτήτη, είτε από τρίτους.
- Το περιεχόμενο και τα μεταδεδομένα είναι διαχειρίσιμα από την αρχιτεκτονική του εκάστοτε αποθετηρίου.
- Το αποθετήριο παρέχει το βασικό σύνολο υπηρεσιών, όπως αναζήτηση, ανάκτηση, προσθήκη και έλεγχο πρόσβασης.
- Το αποθετήριο είναι καλά οργανωμένο και το περιεχόμενο που προσφέρει είναι ασφαλές και έγκυρο.

Υπάρχουν δύο κύριες κατηγορίες ψηφιακών αποθετηρίων, οι οποίες είναι τα θεματικά και τα ιδρυματικά αποθετήρια. Στα θεματικά αποθετήρια, περιέχεται περιεχόμενο μίας συγκεκριμένης κατηγορίας, ενώ στα ιδρυματικά αποθετήρια περιέχονται δεδομένα από διάφορες κατηγορίες, υποστηριζόμενα από κάποια ακαδημαϊκή κοινότητα ή ερευνητικό φορέα [3].

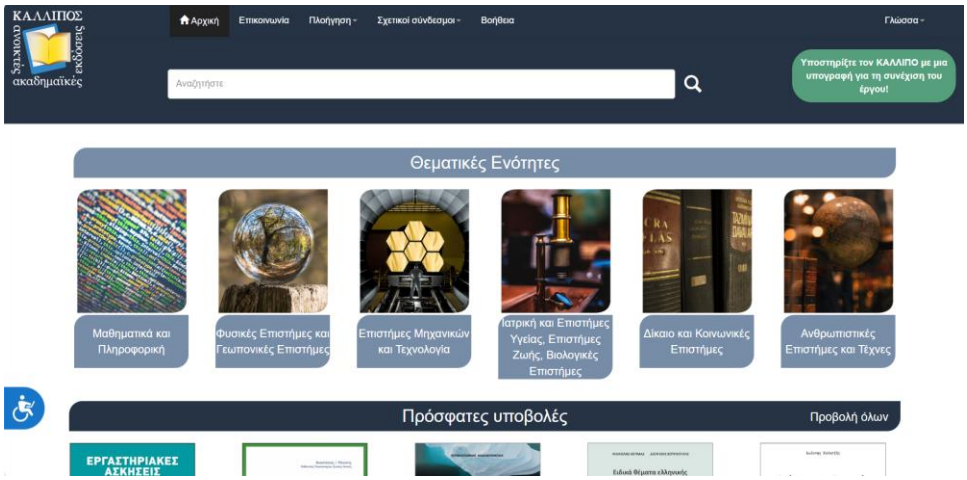
Τα θεματικά αποθετήρια είναι πιο εξειδικευμένα και προσφέρουν καλύτερη εικόνα σε συγκεκριμένους τομείς, καθώς συγκεντρώνουν το περιεχόμενο με γνώμονα το θέμα. Αντίθετα, τα ιδρυματικά αποθετήρια καλύπτουν μεγαλύτερη ποικιλία δεδομένων από διάφορες πηγές και στοχεύουν στη διάδοση ερευνητικών αποτελεσμάτων.

2.1.2 Παραδείγματα Ψηφιακών Αποθετηρίων

Παρακάτω, παρουσιάζονται παραδείγματα κάποιων γνωστών ψηφιακών αποθετηρίων τα οποία καλύπτουν μία ευρεία ποικιλία υλικού από διάφορους τομείς. Παρέχουν πρόσβαση σε πληθώρα υλικού, εκπαιδευτικού ή από επιστημονικές έρευνες, κώδικα λογισμικού έως εικόνες και ήχο.

2.1.2.1 ΚΑΛΛΙΠΟΣ [4]

Το αποθετήριο «ΚΑΛΛΙΠΟΣ» περιέχει συγγράμματα και εκπαιδευτικά βοηθήματα, τα οποία, είτε έχουν δημιουργηθεί από μέλη της ακαδημαϊκής και ερευνητικής κοινότητας στο πλαίσιο Δράσης «Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα και Βοηθήματα», είτε διατεθεί μέσω της ανοικτής πρόσκλησης παραχώρησης επιστημονικού περιεχομένου. Η Δράση ΚΑΛΛΙΠΟΣ έχει στόχο τη δημιουργία ακαδημαϊκών συγγραμμάτων, όπως προπτυχιακών και μεταπτυχιακών εγχειριδίων, μονογραφιών και βιβλιογραφικών οδηγών και τη διάθεσή τους μέσω του ομώνυμου ψηφιακού αποθετηρίου.



Σχήμα 2.1: Κεντρική Σελίδα του Ψηφιακού Αποθετηρίου ΚΑΛΛΙΠΟΣ [4]

Οι θεματικές ενότητες των συγγραμμάτων που περιέχονται στο αποθετήριο ΚΑΛΛΙΠΟΣ είναι οι παρακάτω:

Μαθηματικά και Πληροφορική

Περιέχονται συγγράμματα σχετικά με την επιστήμη των υπολογιστών, καθώς και τα θεωρητικά και εφαρμοσμένα μαθηματικά. Κάποια παραδείγματα περιεχομένου τέτοιων συγγραμμάτων είναι η άλγεβρα, η μαθηματική ανάλυση, ο προγραμματισμός, η τεχνητή νοημοσύνη κ.ά.

Φυσικές Επιστήμες και Γεωπονικές επιστήμες

Περιέχονται συγγράμματα σχετικά με τον κλάδο της φυσικής, της χημείας, της γεωλογίας και της γεωπονίας. Κάποια παραδείγματα περιεχομένου τέτοιων συγγραμμάτων είναι η μετεωρολογία, η αστροφυσική, η οργανομεταλλική-καταλυτική χημεία, η ανόργανη χημεία, η οργανική χημεία, η αρχαιογεωμορφολογία, η ιζηματολογία κ.ά.

Επιστήμες Μηχανικών και Τεχνολογία

Περιέχονται συγγράμματα σχετικά με την επιστήμη των μηχανικών όλων των κλάδων και της τεχνολογίας. Παραδείγματα περιεχομένου τέτοιων συγγραμμάτων είναι η

ηλεκτρολογία, η ηλεκτρονική, η μηχανολογία, η γεωργική μηχανική, η νευτώνεια μηχανική, η μηχανική των ρευστών, η οικιακή τεχνολογία κ.ά.

Ιατρική και Επιστήμες Υγείας, Επιστήμες Ζωής, Βιολογικές Επιστήμες

Περιέχονται συγγράμματα σχετικά με την ιατρική, τη βιολογία, τις επιστήμες υγείας και ζωής. Παραδείγματα περιεχομένου τέτοιων συγγραμμάτων είναι η φυσικοθεραπεία, η πνευμονολογία, η νευροφυσιολογία, η χειρουργική κ.ά.

Δίκαιο και Κοινωνικές Επιστήμες

Περιέχονται συγγράμματα σχετικά με την επιστήμη της νομικής και τις κοινωνικές επιστήμες. Κάποια παραδείγματα περιεχομένου τέτοιων συγγραμμάτων είναι το διεθνές δίκαιο, το τραπεζικό δίκαιο, ο φεμινισμός, η βιοηθική κ.ά.

Ανθρωπιστικές Επιστήμες και Τέχνες

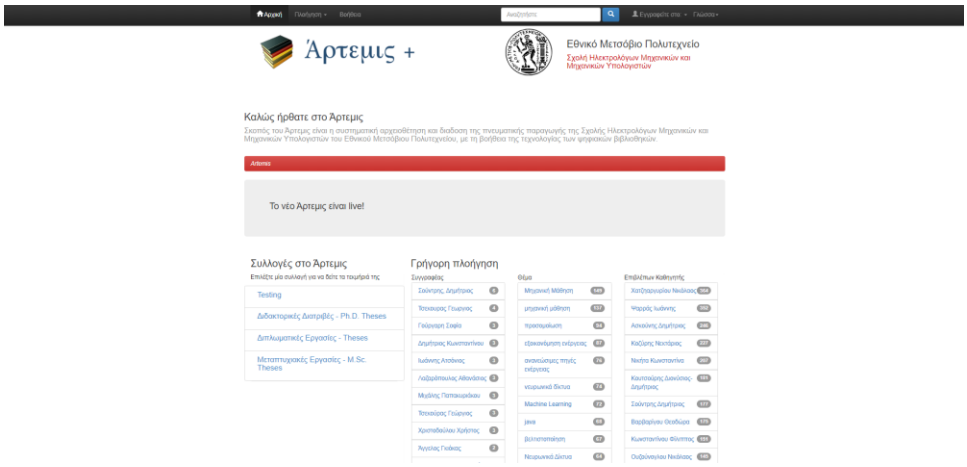
Περιέχονται συγγράμματα σχετικά με τις ανθρωπιστικές επιστήμες και τις τέχνες. Κάποια παραδείγματα περιεχομένου τέτοιων συγγραμμάτων είναι η γλωσσολογία, η φιλολογία, η λογοτεχνία και η μουσική.

Σε παρακάτω ενότητα, θα πραγματοποιηθεί η μελέτη των θεματικών ενοτήτων και του περιεχομένου του αποθετηρίου ΚΑΛΛΙΠΟΣ, καθώς θα γίνει λεπτομερής ανάλυση των στοιχείων χρήσης του.

2.1.2.2 Άρτεμις NTUA [5]

Το αποθετήριο «Άρτεμις» περιέχει συλλογές από Διδακτορικές Διατριβές, Διπλωματικές Εργασίες και Μεταπτυχιακές Εργασίες σπουδαστών του Εθνικού Μετσόβιου Πολυτεχνείου. Ο στόχος του είναι να οργανώσει την ακαδημαϊκή έρευνα της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου, αξιοποιώντας τις δυνατότητες των ψηφιακών βιβλιοθηκών.

Ο ιστότοπος «Άρτεμις» προσφέρει γρήγορη πλοήγηση ανάλογα με τον συγγραφέα, το θέμα και τον επιβλέποντα καθηγητή μέσα από την οποία μπορεί κάποιος χρήστης να αναζητήσει όποιο έγγραφο τον ενδιαφέρει.



Σχήμα 2.2: Κεντρική Σελίδα του Ψηφιακού Αποθετηρίου Άρτεμις NTUA [5]

Κάθε ακαδημαϊκό ίδρυμα παρέχει δικό του αποθετήριο στο οποίο φιλοξενούνται Διδακτορικές Διατριβές και Προπτυχιακές-Μεταπτυχιακές Εργασίες που εκπονούν σπουδαστές του συγκεκριμένου ιδρύματος. Για παράδειγμα, το Πανεπιστήμιο Πειραιώς παρέχει το αποθετήριο «Διώνη», ενώ το Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών παρέχει το αποθετήριο «Πέργαμος». Με αυτόν το τρόπο, κάθε σπουδαστής και ερευνητής, οποιουδήποτε ιδρύματος, έχει τη δυνατότητα να μοιράζεται, να αποθηκεύει και να αναζητεί επίσημα ακαδημαϊκά έγγραφα.

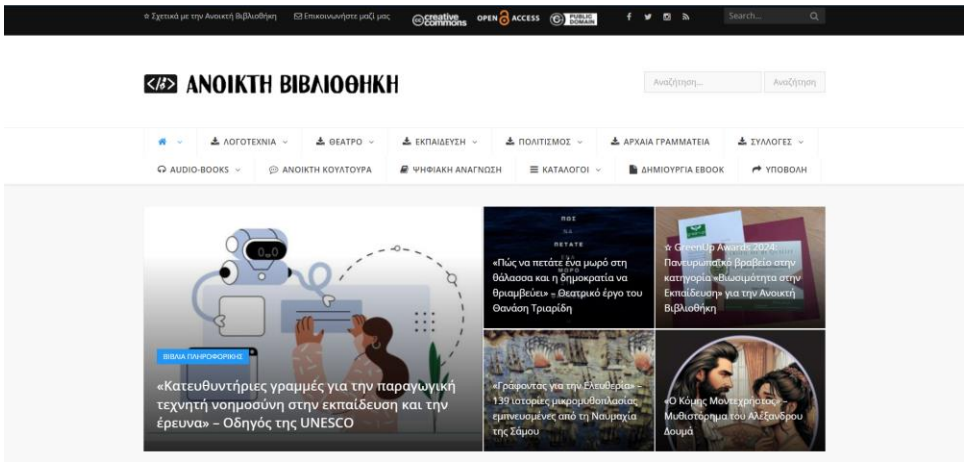
2.1.2.3 DSpace@NTUA [6]

Το αποθετήριο «DSpace@NTUA» λειτουργεί ως Ιδρυματικό αποθετήριο του Εθνικού Μετσόβιου Πολυτεχνείου στο οποίο δημοσιοποιείται το ερευνητικό έργο των μελών της Ακαδημαϊκής Κοινότητας. Συγκεκριμένα, αποθηκεύονται Διπλωματικές και Μεταπτυχιακές Εργασίες και Διδακτορικές Διατριβές των σπουδαστών της κοινότητας. Επίσης, παρατίθενται δημοσιεύσεις σε περιοδικά και πρακτικά συνεδριών των μελών ΔΕΠ, καθώς και ψηφιοποιημένο υλικό από τη Βιβλιοθήκη του Εθνικού Μετσόβιου Πολυτεχνείου.

Σχήμα 2.3: Κεντρική Σελίδα του Ψηφιακού Αποθετηρίου DSpace@NTUA [6]

2.1.2.4 Ανοικτή Βιβλιοθήκη [7]

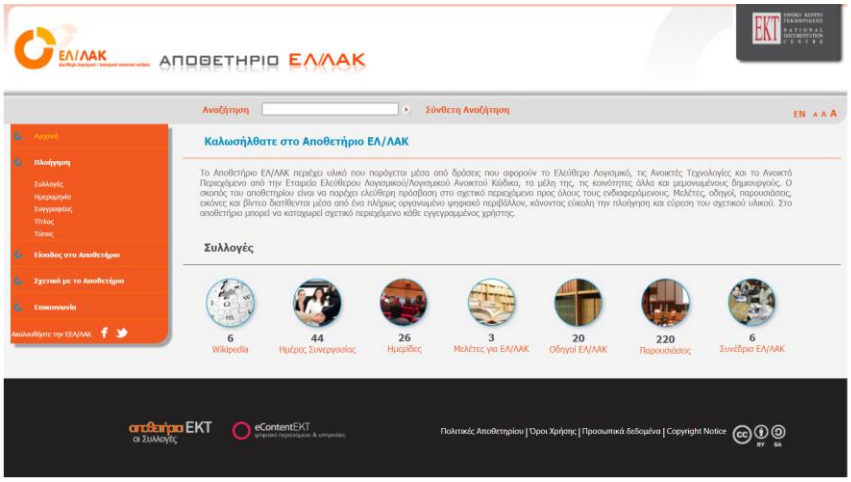
Το αποθετήριο «Ανοικτή Βιβλιοθήκη» ιδρύθηκε το 2010 και διαθέτει χιλιάδες ψηφιακά βιβλία, τα οποία δεν δεσμεύονται από πνευματικά δικαιώματα ή διανέμονται ελεύθερα και νόμιμα στο Διαδίκτυο από τους δημιουργούς. Περιέχει βιβλία από πολλές θεματικές κατηγορίες, προσφέροντας, με αυτόν τον τρόπο, ένα ευρύ φάσμα πληροφοριών και γνώσεων. Συγκεκριμένα, περιλαμβάνει βιβλία από πολλές κατηγορίες, όπως λογοτεχνία, θέατρο, εκπαίδευση, πολιτισμό κ.ά. Επίσης, παρέχει μία ποικιλία από ηχητικά βιβλία, επεκτείνοντας την προσβασιμότητα σε άτομα που χρειάζονται ή προτιμούν την ακρόαση υλικού παρά το διάβασμα. Επίσης, περιλαμβάνονται ανοικτές συζητήσεις, γεγονός που εμπλουτίζει το αποθετήριο με ιδέες και μπορεί να οδηγήσει σε ενδιαφέρουσες ανταλλαγές απόψεων.



Σχήμα 2.4: Κεντρική Σελίδα του Ψηφιακού Αποθετηρίου Ανοικτή Βιβλιοθήκη [7]

2.1.2.5 Αποθετήριο Ελεύθερου Λογισμικού/Λογισμικού Ανοικτού Κώδικα (ΕΛ/ΛΑΚ) [8]

Ο όρος «Ελεύθερο Λογισμικό/Λογισμικό Ανοικτού Κώδικα» περιγράφει το λογισμικό που δημοσιοποιείται με ειδικές άδειες, προκειμένου να επιτρέπεται στους χρήστες να το μελετήσουν και να το βελτιώσουν. Ο όρος «Ελεύθερο Λογισμικό» στηρίζεται στις ελευθερίες που προσφέρονται στον χρήστη μέσω ειδικών αδειών, ενώ ο όρος «Λογισμικό Ανοικτού Κώδικα» στηρίζεται στη δυνατότητα συνεργατικής ανάπτυξης κώδικα.



Σχήμα 2.5: Κεντρική Σελίδα του Ψηφιακού Αποθετηρίου ΕΛ/ΛΑΚ [8]

Το αποθετήριο ΕΛ/ΛΑΚ διαθέτει πηγαίο κώδικα ο οποίος παράγεται μέσα από κοινότητες και ανεξάρτητους δημιουργούς. Στόχος του αποθετηρίου αυτού είναι η ανοικτή πρόσβαση λογισμικού προς όλους τους χρήστες. Η καταχώρηση υλικού μπορεί να πραγματοποιηθεί στην περίπτωση που ο χρήστης είναι εγγεγραμμένος και μπορεί να δημοσιεύσει σχετικές μελέτες, εικόνες και βίντεο που επεξηγούν το υλικό. Το αποθετήριο ΕΛ/ΛΑΚ αναπτύσσεται από το Εθνικό Κέντρο Τεκμηρίωσης (ΕΚΤ) και από τον Οργανισμό Ανοικτών Τεχνολογιών (ΕΕΛΛΑΚ) στο πλαίσιο της δράσης «Εθνικό Πληροφοριακό Σύστημα Έρευνας και Τεχνολογίας/Κοινωνικά Δίκτυα – Περιεχόμενο Παραγόμενο από Χρήστες».

Η αναζήτηση υλικού μπορεί να γίνει βάσει της συλλογής που παρέχει το αποθετήριο, την ημερομηνία, τον συγγραφέα, τον τίτλο και τον τύπο υλικού.

Υπάρχουν αμέτρητα άλλα αποθετήρια, τα οποία παρέχουν μεγάλη ποικιλία περιεχομένου επιστημονικού και ερευνητικού χαρακτήρα. Κατά την περιήγησή τους σε αυτά, οι χρήστες μπορούν να βρουν αυτό που χρειάζονται για την έρευνά τους, σύμφωνα με τις απαιτήσεις και τις προτιμήσεις τους.

2.2 Ανάλυση και Συλλογή Δεδομένων Χρήσης Ψηφιακού Αποθετηρίου

Στο παρόν κεφάλαιο, παρουσιάζονται και αναλύονται κάποια από τα δεδομένα χρήσης, τα οποία συλλέγονται από ένα ψηφιακό αποθετήριο. Στη συνέχεια, θα εξεταστούν οι μέθοδοι συλλογής αυτών των δεδομένων, με σκοπό την κατανόηση των τεχνικών που χρησιμοποιούνται για την παρακολούθηση της αλληλεπίδρασης των χρηστών με το αποθετήριο.

2.2.1 Δεδομένα Χρήσης Ψηφιακού Αποθετηρίου

Τα δεδομένα χρήσης αφορούν πληροφορίες που συλλέγονται από δραστηριότητες χρήστη, με στόχο να γίνει αντιληπτός ο τρόπος που περιηγούνται σε κάποιο ψηφιακό αποθετήριο. Η ανάλυση των δεδομένων αυτών αναδεικνύει στοιχεία τάσης, μοτίβα συμπεριφοράς και

προτιμήσεις χρηστών. Μπορούν να χρησιμοποιηθούν για τη βελτίωση της εμπειρίας χρηστών και την καλύτερη διαχείριση του περιεχομένου μέσω της ανάπτυξης αποτελεσματικότερων υπηρεσιών και της λήψης αποφάσεων.

Παρακάτω, παρουσιάζονται οι βασικότερες πληροφορίες που συλλέγονται σχετικά με το πώς οι χρήστες αλληλεπιδρούν με το περιεχόμενο ενός ψηφιακού αποθετηρίου.

2.2.1.1 Περιοχή Προέλευσης Χρηστών

Η πληροφορία για την περιοχή προέλευσης των χρηστών που επισκέπτονται ένα αποθετήριο είναι σημαντική, καθώς γίνεται κατανοητή η γεωγραφική κατανομή των επισκεπτών. Επίσης, γίνονται κατανοητές οι πολιτισμικές προτιμήσεις των χρηστών και με αυτόν τον τρόπο μπορεί να προσαρμοστεί το περιεχόμενο με σκοπό την ανάπτυξη στρατηγικής και κατ' επέκταση την αύξηση της προσέλευσης χρηστών από περιοχές που δεν εμφανίζουν μεγάλη ζήτηση για το αποθετήριο.

2.2.1.2 Νέοι και Ενεργοί Χρήστες (New & Active Users) [9]

Οι *νέοι χρήστες* είναι εκείνοι που συνδέονται πρώτη φορά στο αποθετήριο. Συνεπώς, η ανάλυση των δραστηριοτήτων τους μπορεί να βοηθήσει στην κατανόηση της πρώτης αντίδρασης στα περιεχόμενα του αποθετηρίου. Η πρώτη αντίδραση από τους χρήστες είναι αρκετά σημαντική στην ανάλυση δεδομένων, καθώς γίνονται αντιληπτά τα αρχικά ενδιαφέροντα και οι ανάγκες τους.

Οι *ενεργοί χρήστες* είναι εκείνοι που παραμένουν στην ιστοσελίδα πάνω από κάποιο συγκεκριμένο χρονικό διάστημα. Το ελάχιστο χρονικό διάστημα ορίζεται από τον τρόπο με τον οποίο συλλέγονται τα στοιχεία χρήσης του αποθετηρίου. Η ανάλυση των ενεργών χρηστών ανά ημερομηνία είναι καταλυτική για τον εντοπισμό της διαδραστικότητας μεταξύ αποθετηρίου και χρηστών. Με άλλα λόγια, μπορούν να εντοπιστούν παράγοντες που μεταβάλλουν την αλληλεπίδραση των χρηστών με το αποθετήριο, όπως οι αλλαγές στο περιεχόμενο, η προώθηση και οι αλλαγές στην οργάνωση του αποθετηρίου.

2.2.1.3 Αφοσίωση Χρήστη (User Engagement) [9]

Η *αφοσίωση χρηστών* αποτελεί το χρονικό διάστημα κατά το οποίο ο χρήστης ξεκινά μία περίοδο σύνδεσης έως ότου αυτή ολοκληρωθεί. Η ολοκλήρωση μίας περιόδου σύνδεσης εξαρτάται από τον τρόπο που καταγράφονται τα δεδομένα χρήσης. Συγκεκριμένα, η περίοδος σύνδεσης τερματίζεται σε περιπτώσεις που ο χρήστης απομακρύνεται από την ιστοσελίδα, δηλαδή στην περίπτωση που κλείσει την καρτέλα, μετακινήσει τη σελίδα στο παρασκήνιο ακόμη και στην περίπτωση που εμφανιστούν σφάλματα στον ιστότοπο. Η αυξημένη αφοσίωση, πιθανότατα, σημαίνει ότι το περιεχόμενο είναι σημαντικό και χρήσιμο για τους επισκέπτες. Συνήθως, τα συστήματα καταγραφής καταγράφουν το ποσοστό αφοσίωσης χρηστών, το οποίο εκφράζεται ως το ποσοστό των περιόδων σύνδεσης αφοσίωσης στον ιστότοπο.

2.2.1.4 Απόκτηση και Επιστροφή Χρήστη (User Acquisition & Returning User) [9]

Η *απόκτηση χρηστών* αναφέρεται στη δράση αύξησης των επισκεπτών που εισέρχονται στο αποθετήριο. Αυτό γίνεται με πολλούς τρόπους, οι οποίοι καταγράφονται στο εκάστοτε σύστημα καταγραφής δεδομένων χρήσης. Αρχικά, ένας πολύ συχνός τρόπος απόκτησης χρηστών είναι η διαφήμιση στο Διαδίκτυο ή στα μέσα κοινωνικής δικτύωσης, καθώς και σε άλλες δραστηριότητες μάρκετινγκ. Η αποτελεσματική διαφήμιση ενός αποθετηρίου μπορεί, σε συνδυασμό με τη μελέτη της γεωγραφικής κατάτμησης, να βοηθήσει στη στόχευση συγκεκριμένου κοινού που ενδιαφέρεται για το περιεχόμενο της πλατφόρμας. Επίσης, άλλος ένας τρόπος απόκτησης χρηστών είναι η *οργανική αναζήτηση*, η οποία εξαρτάται από την προώθηση της μηχανής αναζήτησης που χρησιμοποιείται κάθε φορά. Οι τρόποι απόκτησης χρηστών πρέπει να μελετώνται συχνά και με στρατηγική σκέψη, καθώς αντικατοπτρίζουν τις διαφορετικές ροές χρηστών που εισέρχονται στην πλατφόρμα.

Η *επιστροφή χρηστών* αναφέρεται στην περίπτωση που παρελθοντικοί χρήστες επισκέπτονται εκ νέου στο αποθετήριο, για να περιηγηθούν στο περιεχόμενό του. Από την ανάλυση αυτή, μπορεί να προκύψει πόσο ικανοποιημένοι είναι οι χρήστες από την προηγούμενη εμπειρία τους με το υλικό του εκάστοτε αποθετηρίου. Όταν οι χρήστες επιστρέφουν αρκετές φορές στην ιστοσελίδα, σημαίνει ότι η πλατφόρμα προσφέρει

ικανοποιητική εμπειρία και παρέχει υλικό το οποίο αντικατοπτρίζει τα ενδιαφέροντα των επισκεπτών.

Μία αναλυτική περιγραφή των φαινομένων απόκτησης και επιστροφής χρηστών είναι σημαντική για την ανάπτυξη στρατηγικών, με σκοπό τη βελτίωση της ποιότητας που προσφέρει το εκάστοτε αποθετήριο και συνεπώς της συνολικής απόδοσής του.

2.2.1.5 Επιλογή Πρόσβασης σε Υλικό Αποθετηρίου

Η ανάλυση της επιλογής πρόσβασης στο υλικό ενός αποθετηρίου μπορεί να αποκαλύψει πολλά για τη συμπεριφορά των χρηστών. Συνήθως, τα αποθετήρια παρέχουν το υλικό τους δωρεάν στο κοινό, όμως υπάρχουν και κάποια στα οποία χρειάζεται πληρωμή ενός ποσού, είτε για την αγορά συγκεκριμένου περιεχομένου, είτε για συνδρομή εγγραφής. Επίσης, κάποια αποθετήρια παρέχουν το περιεχόμενό τους μέσω αδειοδότησης σε εξουσιοδοτημένους χρήστες, οι οποίοι λαμβάνουν προσκλήσεις για αποκλειστικό περιεχόμενο. Η επιλογή πρόσβασης σε υλικό αποθετηρίου εξαρτάται από την πλατφόρμα και τον σκοπό της. Πέρα από τις προϋποθέσεις πρόσβασης, ορισμένα αποθετήρια έχουν τη δυνατότητα επιλογής προβολής ή λήψης περιεχομένου. Η προβολή περιεχομένου επιτρέπει στους χρήστες να έχουν άμεση πρόσβαση στο περιεχόμενο, ενώ η λήψη περιεχομένου χρειάζεται πόρους, επιτρέποντας την πρόσβαση σε περιεχόμενο εκτός σύνδεσης. Οι παραπάνω προσεγγίσεις παρέχουν την ευελιξία και την ελευθερία επιλογής αποθήκευσης περιεχομένου στον εκάστοτε επισκέπτη του αποθετηρίου.

Με την ανάλυση του συνόλου χρηστών που επιλέγουν τον εκάστοτε τρόπο πρόσβασης στο αποθετήριο, γίνεται αντιληπτός ο τρόπος αξιοποίησης και η σημαντικότητα του περιεχομένου, ανάλογα με τις ανάγκες και τις προτιμήσεις των επισκεπτών.

2.2.2 Μέθοδοι Συλλογής Δεδομένων Χρήσης Ψηφιακού Αποθετηρίου

Σε αυτό το κεφάλαιο, εξετάζονται δύο βασικές μέθοδοι συλλογής δεδομένων χρήσης ενός ψηφιακού αποθετηρίου. Αρχικά, περιγράφεται η αυτόματη συλλογή δεδομένων χρήσης και συγκεκριμένα μέσω της πλατφόρμας Google Analytics. Στη συνέχεια, εξετάζεται η μη αυτόματη συλλογή δεδομένων χρήσης, η οποία περιλαμβάνει τη χρήση ερωτηματολογίων

και συνεντεύξεων, με στόχο την απόκτηση πληροφοριών απευθείας από τους χρήστες του αποθετηρίου.

2.2.2.1 Αυτόματη Συλλογή Δεδομένων Χρήσης

Η πιο συχνή μέθοδος για την αυτόματη συλλογή δεδομένων είναι η χρήση Διεπαφής Προγραμματισμού Εφαρμογών (Application Programming Interface ή API), καθώς υπάρχουν πολλά στοιχεία που χρειάζεται να καταγραφούν με συνέπεια και λεπτομέρεια. Επίσης, άλλος ένας λόγος που είναι αρκετά διαδεδομένη η χρήση των APIs, στις εφαρμογές συλλογής δεδομένων, είναι ότι έτσι γίνονται επεκτάσιμες, δηλαδή μπορούν να συνδεθούν με άλλες εφαρμογές για την καλύτερη ανάλυση και επεξεργασία υλικού.

Παρακάτω, παρουσιάζεται, μία από τις πιο διαδεδομένες διεπαφές προγραμματισμού εφαρμογών συλλογής στοιχείων ενός αποθετηρίου, το Google Analytics.

Google Analytics

Η πλατφόρμα Google Analytics προσφέρει στους προγραμματιστές τη δυνατότητα να αποκτήσουν πρόσβαση σε δεδομένα χρήσης εφαρμογών ή ιστοσελίδων. Παρέχει μία ευρεία ποικιλία δεδομένων σχετικά με την απόδοση και τη συμπεριφορά των χρηστών σε έναν ιστότοπο. Συγκεκριμένα, κάθε ιστοσελίδα, έχει τον προσωπικό λογαριασμό (account) και κάθε λογαριασμός έχει ένα ή περισσότερα προφίλ (profile). Ο προσωπικός λογαριασμός (account) είναι ένα αποθετήριο πληροφοριών και περιέχει ένα μοναδικό αναγνωριστικό, ώστε να επαληθεύεται ότι τα δεδομένα στέλνονται στο σωστό μέρος. Τα προφίλ (profiles) παραθέτουν τα δεδομένα στους διαχειριστές και παρέχουν εργαλεία που τροποποιούν τον τρόπο παρουσίασης των δεδομένων. Τα στοιχεία που παρέχονται, είναι αρκετά και, κυρίως, εστιάζουν στην ανάλυση της συμπεριφοράς των χρηστών κατά την περιήγησή τους στην ιστοσελίδα [10]. Το Google Analytics παρέχει μία ποικιλία μετρικών με σκοπό τη μέτρηση επιτυχίας της ιστοσελίδας.

Δύο σημαντικές μετρικές που παρέχει το Google Analytics, για τη μέτρηση της συνολικής κίνησης μίας ιστοσελίδας, είναι οι χρήστες (Users), οι οποίοι αποτελούν τους επισκέπτες της ιστοσελίδας και οι περίοδοι σύνδεσης (Sessions), οι οποίες καταγράφονται όταν ένας χρήστης ανοίγει την εφαρμογή στο προσκήνιο ή βλέπει τη σελίδα εντός ενός δεδομένου

χρονικού πλαισίου στο οποίο δεν υπάρχει ενεργή περίοδος σύνδεσης. Από προεπιλογή, μία συνεδρία ολοκληρώνεται μετά από 30 λεπτά αδράνειας χρήστη. Μερικές φορές, είναι πιο σημαντικές οι αντίστοιχες μετρικές των νέων χρηστών (New Users) ή των ενεργών χρηστών (Active Users), οι οποίες αφορούν τους χρήστες που εισέρχονται πρώτη φορά στην ιστοσελίδα ή τους χρήστες που εισέρχονται σε συγκεκριμένο χρονικό διάστημα αντίστοιχα. Επίσης, σημαντικός είναι και ο αριθμός των περιόδων σύνδεσης ανά χρήστη (Number of Sessions Per User), ο οποίος αποτελεί τις συνολικές περιόδους σύνδεσης διαιρεμένες με τους συνολικούς χρήστες. Με αυτόν τον τρόπο, καταμετρούνται οι χρήστες που επισκέπτονται την ιστοσελίδα πάνω από μία φορά. Οι τρόποι με τους οποίους οι χρήστες ανακαλύπτουν την ιστοσελίδα καταγράφονται στην αντίστοιχη αναφορά Απόκτησης Χρηστών (User acquisition) και αναλύονται διεξοδικά σε παρακάτω ενότητα.

Επίσης, το Google Analytics παρέχει μετρικές που καταμετρούν τις κινήσεις των χρηστών από τη χρονική στιγμή που επισκέπτονται την ιστοσελίδα. Αρχικά, μία μετρική αποτελεί το *ποσοστό εγκατάλειψης* (Bounce Rate), το οποίο είναι το ποσοστό των μη αφοσιωμένων επισκέψεων σε μία ιστοσελίδα (αυτών δηλαδή που δεν θα έχουν συνέχεια σε επόμενη σελίδα). Στις περιπτώσεις που η ιστοσελίδα αποτελείται από μία σελίδα η μετρική αυτή είναι περιττή, καθώς ο χρήστης δεν μπορεί εκ των πραγμάτων να συνεχίσει σε άλλη σελίδα. Επίσης, η μετρική αυτή είναι περιττή στην περίπτωση που ο χρήστης καθοδηγείται αυτόματα από τη ροή της ιστοσελίδας σε μία άλλη σελίδα, καθώς μεταφέρεται υποχρεωτικά σε μία συγκεκριμένη και δεν είναι η επιλογή του. Μία ακόμη βασική μετρική, που παρακολουθεί την απόδοση της ιστοσελίδας, είναι ο *αριθμός προβολών σελίδας* (Pageviews) και αποτελεί τον αριθμό των φορών που φορτώνεται η κάθε σελίδα από τους χρήστες. Μία παραλλαγή αποτελεί ο *αριθμός των μοναδικών προβολών* (Unique Pageviews), ο οποίος μετρά τον αριθμό των περιπτώσεων που η σελίδα φορτώνεται μία φορά σε κάθε περίοδο λειτουργίας. Μία αντίστοιχη μετρική είναι η *καταμέτρηση σελίδων ανά περίοδο σύνδεσης* (Pages Per Session) και μπορεί να δείξει πόσο αφοσιωμένοι είναι οι χρήστες στα περιεχόμενα της ιστοσελίδας. Τα στοιχεία αφοσίωσης των χρηστών παρουσιάζονται στη *σύνοψη αφοσίωσης* (Engagement overview).

Το Google Analytics αναλύει τα στοιχεία που προκύπτουν από τις περιόδους σύνδεσης, ώστε να συμπληρώσει μετρικές αφοσίωσης, όπως η αφοσίωση χρήστη (User engagement)

και ο μέσος χρόνος αφοσίωσης (Average engagement time). Ορισμένες αναφορές, όπως οι *Σελίδες και Οθόνες* (Pages and screens) περιλαμβάνουν μετρήσεις αφοσίωσης. Η αναφορά *Σελίδες και οθόνες* είναι μία προκατασκευασμένη αναφορά που εμφανίζει δεδομένα σχετικά με τις σελίδες που επισκέφτηκαν οι χρήστες στον ιστότοπο. Οι «σελίδες» αναφέρονται στους μεμονωμένους υπερσυνδέσμους ενός ιστοτόπου, ενώ οι «οθόνες» αναφέρονται στις μεμονωμένες οθόνες μίας εφαρμογής. Μία άλλη σημαντική αναφορά που παρέχει το Google Analytics, είναι η *αναφορά των εκδηλώσεων* (Events), η οποία παρουσιάζει το πόσες φορές ενεργοποιείται κάθε συμβάν και πόσοι χρήστες ενεργοποιούν κάθε συμβάν στον ιστότοπο [9]. Σε επόμενη ενότητα, θα αναλυθούν διεξοδικά τα συμβάντα, με σκοπό τη βελτίωση εμπειρίας χρήστη και περιεχομένου του αποθετηρίου ΚΑΛΛΙΠΟΣ.

Τα παραπάνω δεδομένα μπορούν να χρησιμοποιηθούν για να γίνει κατανοητό πώς αλληλεπιδρούν οι χρήστες με το περιεχόμενο, τις διαφημίσεις και τις λειτουργίες μίας ιστοσελίδας.

2.2.2.2 Μη Αυτόματη Συλλογή Δεδομένων Χρήσης

Η μη αυτόματη συλλογή δεδομένων χρήσης αναφέρεται στη διαδικασία κατά την οποία τα δεδομένα συλλέγονται, κυρίως, με τη βοήθεια του ανθρώπινου παράγοντα. Παραμένει απαραίτητη μέθοδος, καθώς μερικές πληροφορίες δεν μπορούν να ληφθούν με αυτόματο τρόπο και μπορεί να αποκαλύψει ελλείψεις ή ανάγκες ενός αποθετηρίου που επηρεάζουν τη συνολική εμπειρία χρήσης.

Στη συνέχεια, παρουσιάζονται οι πιο διαδεδομένες μέθοδοι μη αυτόματης συλλογής δεδομένων χρήσης.

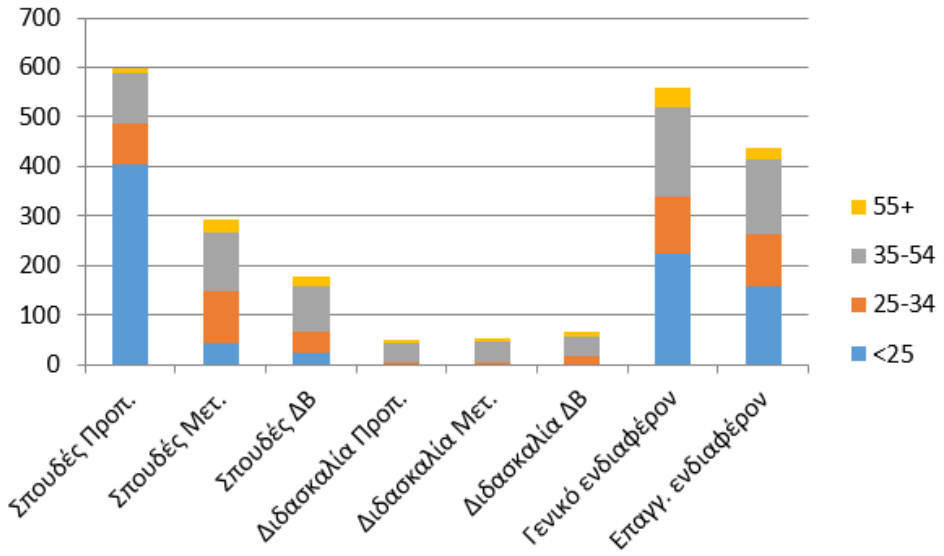
Ερωτηματολόγια

Τα ερωτηματολόγια αποτελούν μία αρκετά χρήσιμη μέθοδο για τη συλλογή δεδομένων. Συνήθως, οι ερωτήσεις που περιλαμβάνονται σε αυτά, για τη λήψη ποιοτικών δεδομένων, είναι ανοικτού τύπου με μορφή ελεύθερου κειμένου ή πολλαπλής επιλογής. Οι ερωτήσεις πρέπει να είναι σαφείς και συγκεκριμένες, ώστε να δίνονται στοχευμένες απαντήσεις πάνω σε συγκεκριμένα ζητήματα. Οι έρευνες με χρήση ερωτηματολογίων είναι έγκυρες και ιδανικές για την καταγραφή πεποιθήσεων και αντιλήψεων, με την προϋπόθεση ότι αφορούν σαφές και προκαθορισμένο δείγμα ατόμων [11].

Πέρα από τις βασικές πληροφορίες που μπορούν να παρέχουν τα συστήματα αυτόματης συλλογής στοιχείων χρήσης, υπάρχουν και άλλες που χρειάζονται περαιτέρω διευκρίνιση και καταγραφή. Ορισμένα αποθετήρια, όπως ο ΚΑΛΛΙΠΟΣ, δεν παρέχουν συγκεκριμένες πληροφορίες για τα χαρακτηριστικά των χρηστών. Αυτό συμβαίνει, γιατί η περιήγηση στο αποθετήριο δεν απαιτεί την εγγραφή των χρηστών. Οπότε, οι πληροφορίες σχετικά με το φύλο, την ηλικία, την εκπαίδευση και κάποια πιο εξειδικευμένα χαρακτηριστικά ή ενδιαφέροντα χρηστών μπορούν να συλλεχθούν μέσω της χρήσης ερωτηματολογίων σε ένα δείγμα χρηστών. Με αυτόν τον τρόπο, γίνονται πιο συγκεκριμένες οι έρευνες που αφορούν τη χρήση και τη συμπεριφορά των χρηστών στο αποθετήριο.

Σύμφωνα με συγκεκριμένη έρευνα, που αφορά το Αποθετήριο ΚΑΛΛΙΠΟΣ, συλλέχθηκαν στοιχεία τα οποία αφορούν το φύλο, την ηλικία, την εργασία, τις σπουδές, τη διδασκαλία και τα ενδιαφέροντα των χρηστών του Αποθετηρίου. Η έρευνα αυτή πραγματοποιήθηκε, σε συγκεκριμένο χρονικό διάστημα, με τη βοήθεια μίας φόρμας ερωτηματολογίου, η οποία εμφανιζόταν κατά την πρόσβαση στην κεντρική ιστοσελίδα του Αποθετηρίου. Στη συγκεκριμένη φόρμα, οι χρήστες του Αποθετηρίου χωρίζονται σε ηλικιακές ομάδες, οι οποίες είναι οι εξής: έως 25 ετών, από 25 έως 34 ετών, από 35 έως 54 ετών και από 55 ετών και άνω. Οι σπουδές και η διδασκαλία των χρηστών αφορούν το προπτυχιακό επίπεδο, το μεταπτυχιακό επίπεδο και τη δια βίου μάθηση, ενώ το ενδιαφέρον για το Αποθετήριο μπορεί να είναι γενικό ή επαγγελματικό.

Παρακάτω, παρουσιάζεται ένα διάγραμμα με τα αποτελέσματα της έρευνας:



Σχήμα 2.6: Αποτελέσματα Ερωτηματολογίου για το Αποθετήριο ΚΑΛΛΙΠΟΣ
(Περίοδος Αναφοράς: 31/01/2019 έως 04/02/2019)

Συγκεκριμένα, αναλύεται ο συνολικός αριθμός χρηστών ορισμένων ηλικιακών ομάδων με γνώμονα τις σπουδές, τη διδασκαλία και το ενδιαφέρον τους. Είναι φανερό ότι, οι χρήστες του αποθετηρίου είναι, κυρίως, προπτυχιακού επιπέδου και έχουν ηλικίες έως 25 ετών. Το περιεχόμενο του Αποθετηρίου προσελκύει περισσότερο τους χρήστες 35 έως 54 ετών που διδάσκουν σε κέντρα δια βίου μάθησης. Επιπλέον, οι χρήστες γενικού ενδιαφέροντος επιλέγουν περισσότερο το Αποθετήριο και ανήκουν, κυρίως, στην ηλικιακή ομάδα έως 25 ετών.

Συνεντεύξεις

Οι συνεντεύξεις χρησιμοποιούνται για τη συλλογή δεδομένων, με μία σειρά προκαθορισμένων ερωτήσεων. Μερικές φορές οι συνεντεύξεις καταγράφονται για περαιτέρω επεξεργασία. Είναι δυνατό οι συνεντεύξεις να είναι ή να μην είναι δομημένες, καθώς μπορεί να ακολουθήσουν μία αυστηρή γραμμή ερωτήσεων ή να είναι αποτέλεσμα έμπνευσης από ένα χαλαρό διάλογο με σκοπό οι συνεντευξιαζόμενοι να εκφράζονται ελεύθερα. Οι συνεντευξιαζόμενοι χρειάζεται να ακούν, να ρωτούν και να διερευνούν ανάλογα με την εξέλιξη της συνέντευξης. Οι συνεντεύξεις είναι ιδανικές για την καταγραφή πεποιθήσεων και αντιλήψεων σχετικά με ορισμένες καταστάσεις και φαινόμενα. Τα δεδομένα των συνεντεύξεων είναι σημαντικά για τη δημιουργία θεμάτων, μοντέλων και την ανακάλυψη των ενδιαφερόντων του κοινού. Πολλά ερευνητικά ερωτήματα μπορούν να απαντηθούν με έρευνες και με συνεντεύξεις. Η κύρια διαφορά των δύο τρόπων επιλογής είναι ότι οι συνεντεύξεις μπορούν να προσφέρουν περισσότερα δεδομένα για τον χαρακτήρα του συνεντευξιαζόμενου και συνεπώς να αποδώσουν πιο στοχευμένες απαντήσεις. Όμως, οι συνεντεύξεις απαιτούν περισσότερους πόρους, χρόνο προετοιμασίας και ανάλυσης, καθώς είναι δύσκολο να προσαρμόζονται οι ερωτήσεις, ξεχωριστά, για τον κάθε συμμετέχοντα [11].

Οι συνεντεύξεις μπορούν να χρησιμοποιηθούν, με σκοπό να καταγραφούν στοιχεία χρήσης ενός αποθετηρίου. Οι χρήστες, με αυτόν τον τρόπο, μπορούν να εκφράσουν ελεύθερα τις προτιμήσεις τους και τις ανάγκες τους σχετικά με το περιεχόμενο του εκάστοτε αποθετηρίου. Δίνεται η δυνατότητα να αποκαλύψουν προβλήματα που μπορεί να αντιμετωπίζουν κατά τη χρήση του αποθετηρίου, καθώς και τις αντιδράσεις τους σε διάφορες λειτουργίες και υπηρεσίες. Συνεπώς, οι συνεντεύξεις παρέχουν στους ερευνητές πολύτιμα σχόλια και προτάσεις για τη βελτίωση του αποθετηρίου και της εμπειρίας χρήστη.

3 Πρόβλεψη Δημοτικότητας Τεκμηρίων Ψηφιακού Αποθετηρίου

3.1 Εισαγωγή

Η πρόβλεψη δημοτικότητας δημοσιευμένων τεκμηρίων αποτελεί σήμερα ένα ενδιαφέρον αντικείμενο για τους επιστήμονες, καθώς υποβοηθά στη βελτίωση της εμπειρίας των χρηστών. Μέσω αυτής της ανάλυσης, παρέχονται πιο έγκαιρες και ακριβείς προτάσεις περιεχομένου ή πτυχές της δραστηριότητας των χρηστών. Τα παραπάνω, επιτρέπουν την ανάπτυξη συστημάτων που εξειδικεύονται στα προσωπικά ενδιαφέροντα των χρηστών, βελτιώνοντας την ποιότητα των παροχών ενός αποθετηρίου.

Παρακάτω, θα εξεταστεί η σημασία της πρόβλεψης δημοτικότητας των τεκμηρίων σε ένα ψηφιακό αποθετήριο, καθώς και οι προκλήσεις που μπορεί να αντιμετωπίζει αυτή η διαδικασία. Επίσης, θα αναλυθούν οι κύριοι παράγοντες που επηρεάζουν τη δημοτικότητα των ψηφιακών τεκμηρίων, καθώς και κάποιες μεθοδολογίες πρόβλεψης σχετικά με αυτό το ζήτημα.

3.2 Προκλήσεις στην Πρόβλεψη Δημοτικότητας Ηλεκτρονικών Βιβλίων

Η γνώση της δημοτικότητας ενός ψηφιακού βιβλίου επηρεάζει τον τρόπο με τον οποίο ένας δημιουργός ιστοσελίδας παρουσιάζει τα συγγράμματα στο εκάστοτε αποθετήριο. Οι δημιουργοί μίας ιστοσελίδας θα εστιάσουν να προωθήσουν, πιθανώς, επιτυχημένα συγγράμματα μεγάλης επισκεψιμότητας με σκοπό την ακόμη μεγαλύτερη απόκτηση επισκεπτών στην ιστοσελίδα.

Οι τεχνικές για την πρόβλεψη δημοτικότητας δημοσιευμένων τεκμηρίων βοηθούν στη βελτιστοποίηση των πόρων δικτύου εφαρμόζοντας στρατηγικές προσωπικής αποθήκευσης και αναπαραγωγής. Οι ακριβείς προβλέψεις επιτρέπουν την καλύτερη τοποθέτηση διαφημίσεων μεγιστοποιώντας τα έσοδα για τις εταιρείες που επενδύουν στην ψηφιακή διαφήμιση.

Όσον αφορά την πρόβλεψη δημοτικότητας ψηφιακού περιεχομένου, είναι κοινώς αποδεκτό ότι, αποτελεί μία πρόκληση σήμερα. Αρχικά, είναι πολύπλοκη και ασταθής η μέτρηση που είναι γνωστό ότι επηρεάζουν τη δημοτικότητα περιεχομένου, όπως η ποιότητα περιεχομένου, το θέμα, ο συγγραφέας και η συνάφεια του περιεχομένου με τους χρήστες. Ακόμη, διάφοροι παράγοντες όπως η σχέση μεταξύ γεγονότων στον φυσικό κόσμο και το περιεχόμενο που είναι δύσκολο να αποτυπωθούν και να συμπεριληφθούν σε ένα μοντέλο πρόβλεψης. Επιπλέον, η εξέλιξη της δημοτικότητας περιεχομένου μπορεί να περιγραφεί από πολύπλοκες διαδικτυακές αλληλεπιδράσεις οι οποίες είναι δύσκολο να προβλεφθούν [12].

3.3 Παράγοντες που Επηρεάζουν τη Δημοτικότητα των Ηλεκτρονικών Βιβλίων

Δεν υπάρχει ακριβής εξήγηση για το πώς μπορεί να γίνει δημοφιλές το οποιοδήποτε διαδικτυακό περιεχόμενο, αλλά υπάρχουν κάποιοι γνωστοί και κοινά αποδεκτοί παράγοντες. Ο εντοπισμός των παραγόντων που επηρεάζουν τη δημοτικότητα του περιεχομένου είναι σημαντικοί για τη δημιουργία πιο ακριβών μοντέλων πρόβλεψης, κατανοώντας ποιες είναι οι σωστές μέθοδοι που πρέπει να χρησιμοποιηθούν [12]. Οι παράγοντες που επηρεάζουν τη δημοτικότητα των ηλεκτρονικών βιβλίων είναι παρόμοιοι με εκείνους που αφορούν την προσέλευση χρηστών στην ιστοσελίδα, αλλά πιο συγκεκριμένα για το εκάστοτε σύγγραμμα. Συγκεκριμένα, κάποιοι από τους παράγοντες που επηρεάζουν τη δημοτικότητα των περιεχομένων μίας ιστοσελίδας και κατ' επέκταση των ηλεκτρονικών συγγραμμάτων παρουσιάζονται παρακάτω.

3.3.1 Ποιότητα Περιεχομένου

Η ποιότητα περιεχομένου που παρέχεται στο εκάστοτε σύγγραμμα δηλαδή η γραφή, η έρευνα, ο τίτλος και το επιστημονικό πεδίο στο οποίο επικεντρώνεται, επηρεάζει την ανταπόκριση του κοινού. Οι αναγνώστες προτιμούν περισσότερο ευανάγνωστα και καλογραμμένα βιβλία που θα τους προσφέρουν νέες γνώσεις και εμπειρίες. Από την άλλη μεριά, υπάρχουν στοιχεία που έχουν αρνητικό αντίκτυπο στη δημοτικότητα του

περιεχομένου. Ένα από αυτά είναι η παρουσία πολλαπλών εκδόσεων του ίδιου περιεχομένου που τείνει να περιορίσει τη δημοτικότητα κάθε μεμονωμένου αντιγράφου, καθώς διαμοιράζεται η προβολή του συγγράμματος.

3.3.2 Συστάσεις Συγγραμμάτων

Οι συστάσεις από συγγενείς, φίλους ή/και καθηγητές επηρεάζουν την επιλογή των αναγνωστών, καθώς, με την προτροπή αυτή, εξετάζουν περισσότερο προσεκτικά τα συγγράμματα. Επίσης, γνωστοί συγγραφείς έχουν, αδιαμφισβήτητα, μεγάλη επίδραση, αφού έχουν αποδείξει τη συγγραφική αξία τους και συνεπώς οι αναγνώστες τους εμπιστεύονται με περισσότερη ευκολία. Έχει παρατηρηθεί, με άλλα λόγια, ότι στα πρώτα στάδια μετά τη δημοσίευση ενός περιεχομένου ιστού, όσο μεγαλύτερη είναι η κοινωνική επίδραση του εκδότη ή του συγγραφέα, τόσο μεγαλύτερη είναι η αύξηση της δημοτικότητας του περιεχομένου.

3.3.3 Προώθηση και Μάρκετινγκ

Η προώθηση και το Μάρκετινγκ αποτελούν καθοριστικούς παράγοντες που επηρεάζουν τη δημοτικότητα των ηλεκτρονικών βιβλίων. Η διαφήμιση στα μέσα μαζικής ενημέρωσης, όπως ιστοσελίδες, τηλεόραση και κοινωνικά δίκτυα μπορεί να αυξήσει την επισκεψιμότητα ενός αποθετηρίου και να προσελκύσει νέους αναγνώστες παρουσιάζοντας το ως ελκυστική επιλογή ανάγνωσης. Επιπλέον, δημοφιλείς υπηρεσίες διαδικτύου, όπως εργαλεία αναζήτησης και εφαρμογές κοινής χρήσης κοινωνικών δικτύων, μπορούν να επεκτείνουν την ορατότητα του περιεχομένου μιας ιστοσελίδας, καθιστώντας τη σελίδα πιο εύκολα προσβάσιμη. Οι κριτικές και τα σχόλια, ιδίως, από αναγνώστες μπορεί να επηρεάσουν την απόφαση κάποιου αναγνώστη για κάποιο σύγγραμμα. Επίσης, η ποιότητα των υπηρεσιών διαδικτύου και εξυπηρετητών (servers) παίζει σημαντικό ρόλο στο πόσο δημοφιλές θα είναι το παρεχόμενο ψηφιακό υλικό. Η παράθεση παρόμοιων αποτελεσμάτων βάσει της περιγραφής με κάποιο σύνολο λέξεων-κλειδιών μπορεί να οδηγήσει, με μεγάλη πιθανότητα, στη δημοτικότητα συγγραμμάτων παρόμοιου περιεχομένου.

3.3.4 Προσωπικές Εμπειρίες και Προτιμήσεις

Το συναίσθημα αποτελεί σημαντικό παράγοντα προσέλευσης κοινού και κατ' επέκταση αύξησης της δημοτικότητας του περιεχομένου. Όταν ένας αναγνώστης έχει προσωπική σύνδεση με κάποιο σύγγραμμα το οποίο του προκαλεί συναισθήματα, δηλαδή το περιεχόμενο αντικατοπτρίζει τα ενδιαφέροντα και τις κλίσεις του, τότε υπάρχει μεγάλη πιθανότητα να το συστήσει σε άλλους και συνεπώς να αυξηθεί το συνολικό ενδιαφέρον για το εκάστοτε ψηφιακό σύγγραμμα. Οι αναγνώστες, συχνά, αναζητούν βιβλία που αντικατοπτρίζουν τις καθημερινές τους εμπειρίες, είτε για θέματα εργασίας, είτε για ψυχαγωγία.

3.4 Μεθοδολογίες Πρόβλεψης Δημοτικότητας Ηλεκτρονικών Συγγραμμάτων

Η πρόβλεψη της δημοτικότητας των ηλεκτρονικών συγγραμμάτων μπορεί να πραγματοποιηθεί με διάφορες μεθοδολογίες και τεχνικές. Μπορούν να προκύψουν πολλές προσεγγίσεις, ανάλογα με τα δεδομένα που είναι διαθέσιμα και τον στόχο της εκάστοτε πρόβλεψης δημοτικότητας. Παρακάτω, αναλύονται κάποιες από τις πιο σύγχρονες μεθοδολογίες πρόβλεψης δημοτικότητας των ηλεκτρονικών συγγραμμάτων.

3.4.1 Ανάλυση και Πρόβλεψη Χρονοσειρών με Στατιστικές Τεχνικές

Η ανάλυση χρονοσειρών (Time Series Analysis) είναι ένας συγκεκριμένος τρόπος ανάλυσης μίας ακολουθίας δεδομένων που συλλέγονται σε ένα χρονικό διάστημα. Στην ανάλυση χρονοσειρών, οι αναλυτές καταγράφουν τα σημεία δεδομένων σε σταθερά διαστήματα μίας καθορισμένης χρονικής περιόδου αντί να καταγράφουν απλώς τα σημεία δεδομένων κατά διαστήματα ή τυχαία. Αυτή η μεθοδολογία δεν συνίσταται απλώς στην πράξη συλλογής δεδομένων με την πάροδο του χρόνου. Η ανάλυση μπορεί να δείξει πώς αλλάζουν κρίσιμες μεταβλητές σε συγκεκριμένα χρονικά διαστήματα. Ο χρόνος αποτελεί μία κρίσιμη παράμετρο, επειδή τα δεδομένα προσαρμόζονται κατά τη διάρκεια των σημείων δεδομένων με διαφορετικό τρόπο σε διαφορετικά χρονικά διαστήματα. Παρέχει μία πρόσθετη πηγή πληροφοριών και μία καθορισμένη σειρά εξαρτήσεων μεταξύ των

δεδομένων. Η ανάλυση χρονοσειρών απαιτεί συνήθως μεγάλο αριθμό σημείων δεδομένων για να διασφαλιστεί η συνέπεια και η αξιοπιστία. Ένα εκτεταμένο σύνολο δεδομένων διασφαλίζει ότι υπάρχει αντιπροσωπευτικό μέγεθος δείγματος και ότι η ανάλυση μπορεί να περιορίσει τα θορυβώδη δεδομένα. Εξασφαλίζει, επίσης, ότι τυχόν τάσεις ή μοτίβα που ανακαλύφθηκαν μπορούν να ευθύνονται για εποχιακή διακύμανση. Επιπλέον, τα δεδομένα χρονοσειρών μπορούν να χρησιμοποιηθούν για πρόβλεψη μελλοντικών δεδομένων με βάση παρελθοντικά δεδομένα. Η ανάλυση χρονοσειρών χρησιμοποιείται για μη στάσιμα δεδομένα, δηλαδή για δεδομένα που κυμαίνονται συνεχώς με την πάροδο του χρόνου.

Ένα παράδειγμα ανάλυσης χρονοσειρών αποτελεί η ανάλυση χρηματιστηρίου με χρήση αυτοματοποιημένων αλγορίθμων συναλλαγών. Αντίστοιχα, η πρόβλεψη χρονοσειρών (Time Series Forecasting) μπορεί να χρησιμοποιηθεί σε προβλέψεις καιρικών συνθηκών, η οποία βοηθά τους μετεωρολόγους να προβλέψουν τα πάντα από την αυριανή πρόγνωση καιρού ως τα μελλοντικά χρόνια της κλιματικής αλλαγής [13]. Επίσης, οι προβλέψεις χρονοσειρών μπορούν να φανούν χρήσιμες και στην περίπτωση μελέτης δημοτικότητας ψηφιακών συγγραμμάτων σε αποθετήρια, καθώς τα αποτελέσματα λαμβάνονται υπόψιν για στρατηγικές αναβάθμισης της πλατφόρμας του αποθετηρίου.

Μερικές μέθοδοι πρόβλεψης δημοτικότητας ψηφιακών συγγραμμάτων, με τη βοήθεια πρόβλεψης χρονοσειρών, παρατίθενται στη συνέχεια.

3.4.1.1 Ανάλυση και Πρόβλεψη Χρονοσειρών με Μοντέλα AR, MA, ARIMA, ARIMAX και SARIMAX

Αυτοπαλινδρομικό Μοντέλο (Autoregressive Model-AR Model)

Ένα αυτοπαλινδρομικό μοντέλο προβλέπει μία τιμή σε μία χρονοσειρά χρησιμοποιώντας έναν γραμμικό συνδυασμό προηγούμενων τιμών της σειράς. Ο όρος υποδηλώνει ότι αποτελεί παλινδρόμηση της μεταβλητής έναντι του εαυτού της.

Το μοντέλο AR(p) μπορεί να οριστεί με την παρακάτω εξίσωση [14]:

$$Y_t = \varphi(L)y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} \quad (3.1)$$

όπου Y_t είναι η τιμή της σειράς τη χρονική στιγμή t, p είναι η σειρά του AR μοντέλου υποδεικνύοντας πόσο υστερεί σε προηγούμενες τιμές που έχουν χρησιμοποιηθεί και φ_1, φ_2, ..., φ_p είναι οι παράμετροι του μοντέλου.

Το μοντέλο AR συσχετίζει την τρέχουσα τιμή της σειράς με τις προηγούμενες τιμές της και υποθέτει ότι οι προηγούμενες τιμές έχουν γραμμική σχέση με την τρέχουσα τιμή.

Μοντέλο Κινούμενου Μέσου Όρου (Moving Average Model-MA Model)

Το μοντέλο κινούμενου μέσου όρου στο πλαίσιο χρονοσειρών χρησιμοποιεί προηγούμενους όρους σφαλμάτων, για να προβλέψει τη σειρά.

Το μοντέλο MA(q) μπορεί να οριστεί με την παρακάτω εξίσωση [14]:

$$y_t = \theta(L)\varepsilon_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (3.2)$$

όπου y_t είναι η τιμή της σειράς τη χρονική στιγμή t, q είναι η σειρά του MA μοντέλου υποδεικνύοντας πόσο υστερεί σε προηγούμενες τιμές σφάλματος που έχουν χρησιμοποιηθεί, θ_1, θ_2, ..., θ_q είναι οι παράμετροι του μοντέλου και ε_t είναι το σφάλμα τη χρονική στιγμή t.

Το μοντέλο MA συσχετίζει την τρέχουσα τιμή της σειράς με προηγούμενους όρους σφάλματος και αποτυπώνει τα απροσδόκητα γεγονότα στο παρελθόν που εξακολουθούν να επηρεάζουν τη σειρά.

Μοντέλο Αυτοπαλινδρομικού Ολοκληρωμένου Κινούμενου Μέσου Όρου (Autoregressive Integrated Moving Average Model-ARIMA Model)

Αποτελεί έναν συνδυασμό της μεθόδου αυτοπαλινδρόμησης και του κινούμενου μέσου όρου, προσθέτοντας την πτυχή της ολοκλήρωσης (I). Με τη μέθοδο της ολοκλήρωσης (I) διαφοροποιούνται δεδομένα, εξαλείφονται τάσεις και σταθεροποιείται, με αυτό τον τρόπο, ο μέσος όρος των χρονοσειρών.

Συνδυάζοντας τα τρία παραπάνω στοιχεία, το μοντέλο ARIMA έχει τη δυνατότητα να συλλάβει τυχόν μοτίβα και εξαρτήσεις δεδομένων χρονοσειρών, επιτρέποντας να γίνονται ακριβείς προβλέψεις [15].

Οπότε, ο τύπος που προκύπτει για το συγκεκριμένο μοντέλο είναι [14]:

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (3.3)$$

Το μοντέλο ARIMA χαρακτηρίζεται από τις παραμέτρους (p, d, q), όπου p είναι η σειρά του AR μοντέλου, q είναι η σειρά του MA μοντέλου και d είναι ο αριθμός των φορών που οι ακριβείς παρατηρήσεις χρειάζεται να διαφοροποιηθούν έτσι ώστε να γίνουν γραμμικά στάσιμες.

Αυτοπαλινδρομικός Ολοκληρωμένος Κινούμενος Μέσος Όρος με Εξωγενείς Παράγοντες (Autoregressive Integrated Moving Average with Exogenous Variables Model-ARIMAX Model)

Το μοντέλο πρόβλεψης ARIMAX αποτελεί μία βελτίωση του μοντέλου ARIMA, καθώς χρησιμοποιεί δεδομένα εξωγενών μεταβλητών, τα οποία επηρεάζουν τις παρατηρούμενες τιμές χρονοσειρών. Ο τύπος που εκφράζει το συγκεκριμένο μοντέλο είναι [14]:

$$Y_t = \beta X_t + \varphi(L) y_t + \theta(L) \varepsilon_t \quad (3.4)$$

όπου X_t είναι η εξωγενής μεταβλητή τη χρονική στιγμή t και β είναι ο συντελεστής της εξωγενούς μεταβλητής.

Εποχιακός Αυτοπαλινδρομικός Ολοκληρωμένος Κινούμενος Μέσος Όρος με Εξωγενείς Παράγοντες (Seasonal Autoregressive Integrated Moving Average with Exogenous Variables Model-SARIMAX Model)

Το μοντέλο SARIMAX αποτελεί μία προέκταση του μοντέλου ARIMAX, λαμβάνοντας υπόψιν εποχιακή παράμετρο που υποδηλώνει την περίοδο εμφάνισης μοτίβων εποχικότητας. Χαρακτηρίζεται από τις παραμέτρους (p, d, q)(P, D, Q, s), όπου p είναι η σειρά του AR μοντέλου, q είναι η σειρά του MA μοντέλου και d είναι ο αριθμός των φορών που οι ακριβείς παρατηρήσεις χρειάζεται να διαφοροποιηθούν έτσι ώστε να γίνουν γραμμικά ακίνητοι. Οι παράμετροι P, Q και D αποτελούν τις εποχικές παραμέτρους σε αντιστοιχία με τις παραμέτρους p, q και d. Η παράμετρος s εκφράζει την περίοδο της εποχικότητας, η οποία είναι ο αριθμός των χρονικών βημάτων μεταξύ των επαναλαμβανόμενων εποχών [16].

Ο τύπος που εκφράζει το συγκεκριμένο μοντέλο είναι [17]:

$$Y_t = \beta_0 + \sum_{i=1}^{k} \beta_i X_{i,t} + \omega_t \tag{3.5}$$

όπου β_0 σταθερή παράμετρος, β_1, β_2, …, β_k είναι οι παράμετροι του συντελεστή παλινδρόμησης των εξωγενών μεταβλητών, $X_{1,t}$, $X_{2,t}$, …, $X_{k,t}$ είναι οι ανεξάρτητες μεταβλητές, k είναι το σύνολο των εξωγενών μεταβλητών και ω_t μία στοχαστική υπολειπόμενη σειρά η οποία είναι ανεξάρτητη από τη σειρά εισόδου:

$$\omega_t = \frac{\theta_q(B)\Theta_Q(B^s)}{\varphi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D}\varepsilon_t \tag{3.6}$$

3.4.1.2 Ανάλυση και Πρόβλεψη Χρονοσειρών με τη Μέθοδο της Εκθετικής Εξομάλυνσης (Exponential Smoothing)

Η πρόβλεψη μελλοντικών τιμών χρονοσειρών είναι μία κρίσιμη διεργασία, την οποία χειρίζεται αποτελεσματικά η εκθετική εξομάλυνση. Η μέθοδος αυτή χρησιμοποιεί προηγούμενες παρατηρήσεις μίας χρονοσειράς για να προβλέψει τις μελλοντικές τιμές. Ονομάζεται εκθετική, επειδή εκχωρεί εκθετικά μειούμενα βάρη σε προηγούμενες παρατηρήσεις με τις πιο πρόσφατες παρατηρήσεις να έχουν υψηλότερα βάρη σε σχέση με

τις παλαιότερες. Με άλλα λόγια, προσαρμόζει τα βάρη με βάση τα δεδομένα, καταγράφοντας τυχόν μοτίβα και τάσεις. Είναι απλή στην εφαρμογή και αποδοτική, καθιστώντας το κατάλληλο για μεγάλα σύνολα δεδομένων και εφαρμογές πρόβλεψης σε πραγματικό χρόνο. Επίσης είναι προσαρμοστική μέθοδος, καθώς ενημερώνει τις προβλέψεις με βάση τα πιο πρόσφατα δεδομένα και συνεπώς ανταποκρίνεται στις αλλαγές που συμβαίνουν με την πάροδο του χρόνου. Δεν απαιτεί μεγάλο αριθμό σημείων δεδομένων, καθώς αποδίδει υψηλότερα βάρη σε πιο πρόσφατες παρατηρήσεις, καθιστώντας το κατάλληλο σε περιπτώσεις που η διαθεσιμότητα δεδομένων είναι περιορισμένη.

Υπάρχουν αρκετοί τύποι εκθετικής εξομάλυνσης, κάποιοι από τους οποίους παρατίθενται παρακάτω [18]:

Απλή Εκθετική Εξομάλυνση (Simple Exponential Smoothing)

Αποτελεί την πιο απλή μορφή εκθετικής εξομάλυνσης, καθώς χρησιμοποιεί την τρέχουσα παρατήρηση και την πρόβλεψη από την προηγούμενη περίοδο για την επόμενη πρόβλεψη. Ο τύπος για την Απλή Εκθετική Εξομάλυνση είναι:

$$F(t+1) = a \cdot Y(t) + (1-a) \cdot F(t) \qquad (3.7)$$

όπου $F(t+1)$ είναι η πρόβλεψη για την επόμενη περίοδο, Y(t) είναι η πραγματική παρατήρηση τη στιγμή t, $F(t)$ είναι η πρόβλεψη για την τρέχουσα περίοδο και α (αε(0,1)) είναι η παράμετρος εξομάλυνσης που ελέγχει το βάρος που έχει εκχωρηθεί στην τρέχουσα παρατήρηση έναντι του βάρους που έχει εκχωρηθεί στην πρόβλεψη από την προηγούμενη περίοδο. Μικρή τιμή της παραμέτρου α, δίνει μεγαλύτερη βαρύτητα στις προηγούμενες παρατηρήσεις, ενώ μεγαλύτερη τιμή του α δίνει μεγαλύτερη βαρύτητα στην τρέχουσα παρατήρηση.

Διπλή Εκθετική Εξομάλυνση (Double Exponential Smoothing)

Αποτελεί επέκταση της Απλής Εκθετικής Εξομάλυνσης ενσωματώνοντας ένα στοιχείο τάσης και επιπέδου. Είναι κατάλληλη για δεδομένα χρονοσειρών που παρουσιάζουν μία τάση, δηλαδή μία εναλλαγή στις τιμές των χρονοσειρών με την πάροδο του χρόνου.

Ο τύπος που περιγράφει τη Διπλή Εκθετική Εξομάλυνση είναι:

$$F(t+1) = L(t) + T(t) \tag{3.8}$$

L(t) είναι η συνιστώσα επιπέδου τη στιγμή t και T(t) είναι η συνιστώσα τάσης τη στιγμή t, δηλαδή:

$$L(t) = a \cdot Y(t) + (1-a) \cdot \big(L(t-1) + T(t-1)\big) \tag{3.9}$$

$$T(t) = \beta \cdot \big(L(t) - L(t-1)\big) + (1-\beta) \cdot T(t-1) \tag{3.10}$$

όπου β (βϵ(0,1)) είναι η παράμετρος εξομάλυνσης για την τάση. Η Διπλή Εκθετική Εξομάλυνση χρησιμοποιεί δύο παραμέτρους εξομάλυνσης, α για τον επίπεδο και β για την τάση, επιτρέποντας μεγαλύτερη ευελιξία στην καταγραφή διαφορετικών μοτίβων στα δεδομένα.

Τριπλή Εκθετική Εξομάλυνση (Triple Exponential Smoothing)

Αποτελεί επέκταση της Διπλής Εκθετικής Εξομάλυνσης, ενσωματώνοντας ένα εποχιακό στοιχείο πέρα από τα στοιχεία επιπέδου και τάσης. Είναι κατάλληλη για δεδομένα χρονοσειρών που παρουσιάζουν εποχικότητα, δηλαδή παρουσιάζουν ένα επαναλαμβανόμενο μοτίβο στις τιμές χρονοσειρών με την πάροδο του χρόνου. Ο τύπος για την Τριπλή Εκθετική Εξομάλυνση είναι:

$$F(t+1) = \big(L(t) + T(t)\big) \cdot S(t-m+1) \tag{3.11}$$

L(t) είναι η συνιστώσα επιπέδου τη στιγμή t, T(t) είναι η συνιστώσα τάσης τη στιγμή t και S(t) είναι η εποχική συνιστώσα τη στιγμή t, δηλαδή:

$$L(t) = a \cdot \frac{Y(t)}{S(t-m)} + (1-a) \cdot \big(L(t-1) + T(t-1)\big) \tag{3.12}$$

$$T(t) = \beta \cdot \big(L(t) - L(t-1)\big) + (1-\beta) \cdot T(t-1) \tag{3.13}$$

$$S(t) = \gamma \cdot \frac{Y(t)}{L(t)} + (1-\gamma) \cdot S(t-m) \tag{3.14}$$

όπου β (βϵ(0,1)) είναι η παράμετρος εξομάλυνσης για την τάση, m είναι η διάρκεια εποχικής περιόδου και γ (γϵ(0,1)) είναι η παράμετρος εξομάλυνσης για την εποχική συνιστώσα. Η Τριπλή Εκθετική Εξομάλυνση χρησιμοποιεί τρεις παραμέτρους εξομάλυνσης, α για τον επίπεδο, β για την τάση και γ για την εποχή. Με αυτόν τον τρόπο,

επιτρέπει την καταγραφή εποχιακών μοτίβων στα δεδομένα, καθιστώντας την κατάλληλη μέθοδο πρόβλεψης χρονοσειρών λαμβάνοντας υπόψιν, τόσο την τάση όσο και την εποχικότητα.

3.4.2 Μηχανική Μάθηση (Machine Learning)

Η χρήση μεθόδων μηχανικής μάθησης, για την πρόβλεψη της δημοτικότητας περιεχομένου ιστού, είναι εξαιρετικά διαδεδομένη, εξαιτίας της δυνατότητας επεξεργασίας μεγάλου όγκου δεδομένων και ανίχνευσης πολύπλοκων μοτίβων. Επίσης, η ανάπτυξη νέων αλγορίθμων και μεθόδων, στον τομέα της μηχανικής μάθησης, μπορεί να επιτρέψει την ανάπτυξη πιο ακριβών και αποδοτικών μοντέλων. Είναι σημαντικό να ληφθούν υπόψη πολλοί παράγοντες, όπως ο τύπος, οι πηγές και ο τρόπος επεξεργασίας των δεδομένων. Η χρήση μηχανικής μάθησης αποτελεί χρήσιμο εργαλείο στην περίπτωση μελέτης δημοτικότητας ψηφιακών συγγραμμάτων, καθώς χρειάζεται αποτελεσματική διαχείριση του μεγάλου όγκου δεδομένων που παρέχει ένα αποθετήριο.

Μερικές μέθοδοι πρόβλεψης δημοτικότητας ψηφιακών συγγραμμάτων, με τη βοήθεια της μηχανικής μάθησης, παρατίθενται παρακάτω.

3.4.2.1 Δέντρα Αποφάσεων (Decision Trees)

Πρόσφατες έρευνες προβλέψεων έδειξαν ότι τα δέντρα αποφάσεων παρέχουν εντυπωσιακή ακρίβεια στις πωλήσεις και σε άλλες εφαρμογές προβλέψεων. Τα δέντρα αποφάσεων χρησιμοποιούν επεξηγηματικές μεταβλητές, οι οποίες λέγονται *χαρακτηριστικά* (features) μίας εξαρτημένης μεταβλητής (*στόχος*). Οι προβλέψεις διαμορφώνονται από ένα σύνολο αποφάσεων που καθορίζουν πώς μπορούν να είναι κατηγοριοποιημένα τα δεδομένα. Η ρίζα είναι ο κόμβος που ξεκινά το γράφημα και είναι συνήθως η μεταβλητή που διαχωρίζει καλύτερα, βάσει του μέτρου ακρίβειας, τα δεδομένα. Στο δυαδικό δέντρο αποφάσεων, η ρίζα χωρίζεται σε δύο κλάδους με βάση μία συνθήκη που καθορίζεται από τον αλγόριθμο δημιουργίας δέντρων. Τα δεδομένα που πληρούν τις συνθήκες εκχωρούνται σε έναν από τους κλάδους, ενώ τα υπόλοιπα εκχωρούνται στον άλλον κλάδο. Κάθε αρχικός κλάδος μπορεί να χωριστεί περαιτέρω ή μπορεί να σταματήσει

σε κάποιον τελικό κόμβο, ο οποίος δεν μπορεί να χωριστεί παραπάνω. Η παραπάνω διαδικασία επαναλαμβάνεται μέχρι κάποια συνθήκη τερματισμού.

Μόνο οι μεταβλητές εισόδου που σχετίζονται με τη μεταβλητή στόχου χρησιμοποιούνται για τον διαχωρισμό των γονικών κόμβων σε ξεκάθαρους θυγατρικούς κόμβους της μεταβλητής στόχου. Μπορούν να χρησιμοποιηθούν τόσο διακριτές μεταβλητές εισόδου όσο και μεταβλητές συνεχούς εισόδου. Ο βαθμός «καθαρότητας» των θυγατρικών κόμβων μπορεί να υπολογιστεί μέσω της εντροπίας, του δείκτη Gini, το σφάλμα κατάταξης, το κέρδος πληροφορίας, τον λόγο κέρδους κ.λπ. Ένα πολύ περίπλοκο μοντέλο δέντρου αποφάσεων θα ήταν υπερβολικά προσαρμοσμένο στις υπάρχουσες παρατηρήσεις και θα είχε λίγες εγγραφές σε κάθε φύλλο, οπότε δε θα μπορούσε να προβλέψει αξιόπιστα τις μελλοντικές περιπτώσεις και συνεπώς θα είχε κακή γενίκευση. Για τον παραπάνω λόγο επιλέγονται, με βάση τον στόχο της ανάλυσης και τα χαρακτηριστικά της βάσης δεδομένων, κανόνες διακοπής κατά την κατασκευή ενός δέντρου αποφάσεων για να αποτραπεί η αυξημένη πολυπλοκότητα του μοντέλου. Σε κάποιες περιπτώσεις, οι κανόνες διακοπής δεν λειτουργούν καλά. Ένας εναλλακτικός τρόπος για τη δημιουργία ενός μοντέλου δέντρου είναι να αναπτυχθεί πρώτα ένα μεγάλο δέντρο και μετά να κλαδευτεί στο βέλτιστο μέγεθος, αφαιρώντας τους κόμβους που παρέχουν λιγότερες πρόσθετες πληροφορίες [19].

3.4.2.2 Μηχανές Ενίσχυσης Κλίσης (Gradient Boosting Machines)

Μηχανή Ενίσχυσης Κλίσης (Gradient Boosting Machine)

Οι *Μηχανές Ενίσχυσης Κλίσης* ανήκουν στην κατηγορία της μηχανικής μάθησης και είναι προσαρμόσιμες στις ιδιαίτερες ανάγκες της εκάστοτε εφαρμογής. Η γενικότερη ιδέα των μηχανών ενίσχυσης κλίσης είναι ότι προστίθενται νέα μοντέλα, διαδοχικά, στο σύνολο. Σε κάθε επανάληψη, εκπαιδεύεται ένα νέο αδύναμο μοντέλο βασικής εκπαίδευσης σε σχέση με το σφάλμα του γενικού συνόλου που έχει εκπαιδευτεί μέχρι εκείνη τη στιγμή.

Πιο συγκεκριμένα, ένας απλός αλγόριθμος μηχανής ενίσχυσης κλίσης ξεκινά με την αρχικοποίηση του μοντέλου με μία σταθερή τιμή, όπως τη μέση τιμή της μεταβλητής στόχου. Έπειτα, με επαναληπτικό τρόπο, δημιουργούνται κάποιες δομές, όπως είναι τα

δέντρα αποφάσεων και, κάθε ένα από αυτά, διορθώνουν τα λάθη που δημιουργούνται από τα προηγούμενα. Σε κάθε επανάληψη, υπολογίζεται η υπολειπόμενη τιμή, δηλαδή η διαφορά μεταξύ της προβλεπόμενης και της πραγματικής τιμής του τρέχοντος μοντέλου. Κάθε φορά που δημιουργείται κάποια δομή, προσαρμόζεται με την υπολειπόμενη τιμή και εκπαιδεύεται κατάλληλα, με σκοπό την πρόβλεψη των υπολειπόμενων τιμών. Οι προβλέψεις αυτές προστίθενται στις προηγούμενες προβλέψεις των μοντέλων. Η διαδικασία επαναλαμβάνεται έως ότου λάβει χώρα κάποιο κριτήριο τερματισμού. Το μοντέλο που προκύπτει, αποτελεί το άθροισμα όλων των προβλέψεων [20].

Μηχανή Ενίσχυσης Ελαφριάς Κλίσης (Light Gradient Boosting Machine)

Μία δημοφιλής παραλλαγή του απλού αλγορίθμου μηχανής ενίσχυσης κλίσης αποτελεί η *Μηχανή Ενίσχυσης Ελαφριάς Κλίσης* (Light Gradient Boosting Machine). Είναι σχεδιασμένη, για να χειρίζεται μεγάλα σύνολα δεδομένων και να αποδίδει ταχύτερα από άλλα πλαίσια ενίσχυσης κλίσης. Χρησιμοποιεί τη μέθοδο μονόπλευρης δειγματοληψίας για να χωρίσει δέντρα, με σκοπό τη μείωση της χρήσης μνήμης και τη βελτίωση της ακρίβειας. Επίσης, χρησιμοποιεί φυλλομετρική ανάπτυξη αντί για ανάπτυξη από επίπεδο σε επίπεδο, γεγονός που την καθιστά ταχύτερη από τις υπόλοιπες μεθόδους ενίσχυσης κλίσης. Η Μηχανή Ενίσχυσης Ελαφριάς Κλίσης μπορεί να χειριστεί διάφορους τύπους δεδομένων, συμπεριλαμβανομένων κατηγορικών και αριθμητικών δεδομένων. Περιλαμβάνει ενσωματωμένες λειτουργίες για προεπεξεργασία δεδομένων, όπως διασταυρωμένη επικύρωση και συντονισμό υπερπαραμέτρων, διευκολύνοντας τους χρήστες να βελτιστοποιήσουν τα μοντέλα τους [21].

Στον παρακάτω Πίνακα 3.1, παρουσιάζεται μία συνοπτική σύγκριση της μεθόδου Μηχανής Ενίσχυσης Ελαφριάς Κλίσης με τη μέθοδο Μηχανής Ενίσχυσης Κλίσης:

Χαρακτηριστικό	**Μηχανή Ενίσχυσης Κλίσης (GBM)**	**Μηχανή Ενίσχυσης Ελαφριάς Κλίσης (LightGBM)**
Διαχείριση Μεγάλου Συνόλου Δεδομένων	Υποφέρει από θέματα κλιμάκωσης για μεγάλα σύνολα δεδομένων.	Μπορεί να διαχειριστεί μεγάλα σύνολα δεδομένων, χάρη στον αποδοτικό και διανεμημένο αλγόριθμο εκπαίδευσης.
Ταχύτητα	Μπορεί να είναι αργό για μεγάλα και πολύπλοκα σύνολα δεδομένων.	Γρήγορη και κλιμακούμενη εκπαίδευση για πρόβλεψη. Μερικές φορές ξεπερνά, σε ταχύτητα, άλλα πλαίσια μηχανών ενίσχυσης.
Χρήση Μνήμης	Μεγάλη χρήση μνήμης, λόγω της αποθήκευσης όλων των δέντρων.	Μειωμένη χρήση μνήμης, λόγω του διαχωρισμού των φύλλων με τη μεγαλύτερη απώλεια.
Υπερπροσαρμογή	Επιρρεπής στην υπερπροσαρμογή.	Λιγότερο επιρρεπής στην υπερπροσαρμογή, λόγω των προηγμένων τεχνικών κανονικοποίησης.
Βελτίωση	Οι παράμετροι βελτιστοποίησης μπορεί να είναι χρονοβόρες και δύσκολες στην υλοποίηση.	Παρέχει ενσωματωμένες μεθόδους για αυτόματη βελτίωση υπερπαραμέτρων, κάνοντας ευκολότερη τη βελτιστοποίηση των μοντέλων.
Ακρίβεια	Μπορεί να επιτευχθεί μεγάλη ακρίβεια με κατάλληλη βελτίωση.	Συχνά ξεπερνά άλλες μεθόδους μηχανών ενίσχυσης κλίσης.
Υλοποίηση	Διαθέσιμη σε περισσότερες βιβλιοθήκες μηχανικής μάθησης.	Διαθέσιμη σε πλαίσια και σε αρκετές βιβλιοθήκες μηχανικής μάθησης.

Πίνακας 3.1: Σύγκριση Μεθόδων GBM και LightGBM [21]

3.5 Αξιολόγηση Πρόβλεψης Δημοτικότητας Ηλεκτρονικών Συγγραμμάτων

Η αξιολόγηση ενός μοντέλου πρόβλεψης δημοτικότητας ηλεκτρονικών συγγραμμάτων περιλαμβάνει την εκτίμηση της ακρίβειας των προκείμενων προβλέψεών του σε σχέση με τα πραγματικά δεδομένα. Αποτελεί ένα σημαντικό στάδιο στην ανάλυση, καθώς, με αυτόν τον τρόπο, επαληθεύεται όλη η διαδικασία επιλογής και επεξεργασίας δεδομένων. Οι πιο δημοφιλείς μετρικές αξιολόγησης παρουσιάζονται παρακάτω [22].

3.5.1 Μέσο Απόλυτο Σφάλμα (MAE)

Το μέσο απόλυτο σφάλμα (Mean Absolute Error) είναι ο υπολογισμός των μέσων απόλυτων διαφορών μεταξύ των αναμενόμενων $\hat{y}_i$ και των πραγματικών τιμών y_i. Ο τύπος που εκφράζει το μέσο απόλυτο σφάλμα είναι:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.15}$$

Αποτελεί μία σημαντική μέτρηση για την αξιολόγηση παλινδρομήσεων και των κατανομών συχνοτήτων, ενώ δεν επηρεάζεται από ακραίες τιμές. Μπορεί να εφαρμοστεί, για παράδειγμα, στην αξιολόγηση ακρίβειας των μοντέλων πρόβλεψης πωλήσεων μίας εταιρείας. Όσο μικρότερη είναι η τιμή του μέσου απόλυτου σφάλματος, τόσο πιο ακριβές είναι το μοντέλο.

3.5.2 Μέσο Τετραγωνικό Σφάλμα (MSE)

Το μέσο τετραγωνικό σφάλμα (Mean Squared Error) είναι ο υπολογισμός των μέσων τετραγωνικών διαφορών μεταξύ των αναμενόμενων $\hat{y}_i$ και των πραγματικών τιμών y_i. Ο τύπος που εκφράζει το μέσο απόλυτο σφάλμα είναι:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.16}$$

Το μέσο τετραγωνικό σφάλμα εφαρμόζεται για τον εντοπισμό ακραίων τιμών και ανωμαλιών σε ένα σύνολο δεδομένων. Όσο μικρότερη είναι η τιμή του μέσου

τετραγωνικού σφάλματος, τόσο πιο ακριβές είναι το μοντέλο. Ο τετραγωνισμός των διαφορών μεταξύ των αναμενόμενων $\hat{y}_i$ και των πραγματικών τιμών y_i υποδηλώνει ότι οι τιμές του μέσου τετραγωνικού σφάλματος είναι πάντα μη αρνητικές και ότι είναι ευαίσθητο σε μεγάλα σφάλματα, καθώς τετραγωνίζονται οι διαφορές. Μπορεί να εφαρμοστεί, για παράδειγμα, στην αξιολόγηση προβλέψεων μετοχών και τιμών χρηματοοικονομικών περιουσιακών στοιχείων με σκοπό τη βελτίωση των επενδυτικών στρατηγικών.

3.5.3 Τετραγωνική Ρίζα Μέσου Τετραγωνικού Σφάλματος (RMSE)

Η τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος (Root Mean Squared Error) είναι ο υπολογισμός της τετραγωνικής ρίζας των μέσων τετραγωνικών διαφορών μεταξύ των αναμενόμενων $\hat{y}_i$ και των πραγματικών τιμών y_i. Ο τύπος που εκφράζει την τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος είναι:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.17}$$

Το RMSE είναι, ουσιαστικά, η τετραγωνική ρίζα του MSE και υπολογίζει την τυπική απόκλιση του σφάλματος. Όμοια με το MSE, το RMSE χρησιμοποιείται σε παλινδρομήσεις και εκτιμήσεις μοντέλων που αφορούν αριθμητικές προβλέψεις. Όσο μικρότερη είναι η τιμή της τετραγωνικής ρίζας του μέσου τετραγωνικού σφάλματος, τόσο πιο ακριβές είναι το μοντέλο.

3.5.4 Μέσο Ποσοστιαίο Απόλυτο Σφάλμα (MAPE)

Το μέσο ποσοστιαίο απόλυτο σφάλμα (Mean Absolute Percentage Error) είναι ο υπολογισμός των μέσων απόλυτων διαφορών μεταξύ των αναμενόμενων $\hat{y}_i$ και των πραγματικών τιμών y_i. Ο τύπος που εκφράζει το μέσο απόλυτο σφάλμα είναι:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3.18}$$

Το MAPE τείνει να ευνοεί περισσότερο τα μοντέλα που υποτιμούν παρά που υπερεκτιμούν, καθώς, στην περίπτωση που η προβλεπόμενη τιμή υπερβαίνει την πραγματική τιμή, δεν μπορεί να υπάρχει άνω όριο στις τιμές. Στο επιχειρηματικό πλαίσιο, το MAPE μπορεί να αξιοποιηθεί στην ακρίβεια πρόβλεψης πωλήσεων στο Διαδίκτυο και στην πρόβλεψη ταμειακών ροών.

4 Ανάλυση Δεδομένων Χρήσης και Τεκμηρίων στο Αποθετήριο ΚΑΛΛΙΠΟΣ

4.1 Εισαγωγή

Στο παρόν κεφάλαιο, θα πραγματοποιηθεί η παρουσίαση κάποιων δεδομένων που παρέχει το Αποθετήριο ΚΑΛΛΙΠΟΣ, με σκοπό την κατανόηση της λειτουργίας και της αποδοτικότητάς του. Η παρουσίαση περιλαμβάνει την αναφορά για τα είδη τεκμηρίων, τα διαθέσιμα φίλτρα, τις μορφές αρχείων και τα μεταδεδομένα που περιέχει το Αποθετήριο.

Οι χρήστες έχουν τη δυνατότητα να αναζητήσουν το τεκμήριο που τους ενδιαφέρει, χρησιμοποιώντας τα διαθέσιμα φίλτρα για εξατομικευμένη αναζήτηση. Τα φίλτρα που είναι διαθέσιμα στο Αποθετήριο αφορούν το είδος τεκμηρίου, την κατάσταση, τον τύπο, τη γλώσσα, το έτος δημοσίευσης, τις θεματικές ενότητες και τον συγγραφέα του εκάστοτε τεκμηρίου. Με αυτόν τον τρόπο, η αναζήτηση των τεκμηρίων γίνεται πιο εύκολη και άμεση για τον επισκέπτη. Επίσης, σε οποιοδήποτε τεκμήριο, ο επισκέπτης του Αποθετηρίου μπορεί να υποβάλει σχόλια και αναφορές σφαλμάτων για την καλύτερη οργάνωση του περιεχομένου.

Το Αποθετήριο ΚΑΛΛΙΠΟΣ παρέχει τρία είδη τεκμηρίων: Συγγράμματα, κεφάλαια και μαθησιακά αντικείμενα, τα οποία δημοσιεύονται οριστικά ή προσωρινά. Πιο συγκεκριμένα, τα μαθησιακά αντικείμενα μπορεί να είναι βίντεο, ήχος, εικόνα, άσκηση, διαφάνειες, διαδραστικό αντικείμενο, προσομοίωση, αλγόριθμος και πίνακας. Από την άλλη μεριά, κάποιο σύγγραμμα μπορεί να είναι μονογραφία, λεξικό, εργαστηριακός και βιβλιογραφικός οδηγός, ιατρικός άτλας, προπτυχιακό ή μεταπτυχιακό εγχειρίδιο. Τα τεκμήρια που παρέχονται, είναι, κατά κύριο λόγο, σε ελληνική γλώσσα, αλλά υπάρχουν και κάποια που είναι σε αγγλική, γερμανική και ιταλική γλώσσα.

Τα κεφάλαια που παρέχονται στο Αποθετήριο βρίσκονται σε μορφή .pdf και είναι διαθέσιμα για ανάγνωση ή/και για τοπική αποθήκευση. Το ίδιο ισχύει και για τα συγγράμματα, με τη διαφορά ότι ο αναγνώστης έχει τη δυνατότητα να κάνει προεπισκόπηση ή/και να αποθηκεύσει τον πίνακα περιεχομένων ενός συγκεκριμένου συγγράμματος. Επίσης, ένας χρήστης μπορεί να κατεβάσει τα συνοπτικά στοιχεία ενός

συγγράμματος, στα οποία καταγράφονται η περίληψη και τα μεταδεδομένα. Τα μεταδεδομένα που παρέχονται είναι: Ο τίτλος, ο υπότιτλος, η γλώσσα, οι συγγραφείς, ο διεθνής πρότυπος αριθμός βιβλίου (ISBN), οι θεματικές κατηγορίες και οι λέξεις-κλειδιά. Στην περίπτωση των μαθησιακών αντικειμένων, η μορφή των αρχείων εξαρτάται από τον τύπο τους. Συγκεκριμένα, στην περίπτωση βίντεο τα αρχεία έχουν, κυρίως, τη μορφή .mp4, στην περίπτωση ήχου .mp3, στην περίπτωση εικόνας .jpeg, στην περίπτωση άσκησης .zip, στην περίπτωση διαδραστικού αντικειμένου .html, στην περίπτωση διαφανειών .pdf, στην περίπτωση προσομοίωσης .gif, στην περίπτωση πινάκων .xml και στην περίπτωση αλγορίθμων η μορφή αρχείων εξαρτάται από τη γλώσσα προγραμματισμού που χρησιμοποιείται, π.χ. .C, .m κ.λπ. Από τα παραπάνω, αποδεικνύεται ότι υπάρχει μεγάλη ποικιλία αρχείων που μπορούν να βελτιώσουν την εκπαιδευτική διαδικασία και να την κάνουν πιο προσιτή στον αναγνώστη.

Στη συνέχεια, θα εξεταστούν διάφοροι παράγοντες που επηρεάζουν τη χρήση των εκπαιδευτικών ψηφιακών συγγραμμάτων. Αυτοί οι παράγοντες περιλαμβάνουν: Το ακαδημαϊκό επίπεδο του χρήστη, την ευχρηστία των εργαλείων, την ποιότητα και την πολυγλωσσικότητα του περιεχομένου. Επιπλέον, θα περιγραφεί η διαδικασία συλλογής των δεδομένων χρήσης και συγγραμμάτων του Αποθετηρίου. Τα δεδομένα χρήσης του Αποθετηρίου ΚΑΛΛΙΠΟΣ συλλέγονται μέσω της εφαρμογής Google Analytics, η οποία καταγράφει τις ενέργειες των χρηστών και άλλες σημαντικές μετρήσεις, προσφέροντας χρήσιμες πληροφορίες για την αλληλεπίδραση με το περιεχόμενο. Από την άλλη μεριά, τα δεδομένα συγγραμμάτων συλλέγονται μέσω κατάλληλης Διασύνδεσης Προγραμματισμού Εφαρμογών (API), η οποία προσφέρει πληροφορίες για τα αντίστοιχα μεταδεδομένα. Τα δεδομένα χρήσης παρουσιάζονται εκτενώς με αντίστοιχες γραφικές παραστάσεις για την κατανόηση της συμπεριφοράς των χρηστών σε σχέση με το περιεχόμενο του Αποθετηρίου. Τα συγγράμματα, αντίστοιχα, αναλύονται με βάση την ημερομηνία έκδοσής τους και συσχετίζονται ανάλογα με τις θεματικές κατηγορίες στις οποίες ανήκουν.

4.2 Παράγοντες Χρήσης Ψηφιακών Συγγραμμάτων του Αποθετηρίου ΚΑΛΛΙΠΟΣ

Η χρήση ψηφιακών συγγραμμάτων σε αποθετήρια, όπως το αποθετήριο ΚΑΛΛΙΠΟΣ, επηρεάζεται από πολλούς παράγοντες. Ο εντοπισμός αυτών των παραμέτρων είναι καθοριστικός για την κατανόηση των προτιμήσεων των χρηστών. Παρακάτω, αναλύονται οι βασικότεροι παράγοντες που επηρεάζουν τη χρήση των ψηφιακών συγγραμμάτων. Από αυτή την ανάλυση, προκύπτουν πληροφορίες για την αποτελεσματικότητα και τη χρηστικότητα του Αποθετηρίου.

4.2.1 Κατηγορία Βιβλίου

Η κατηγορία του εκπαιδευτικού βιβλίου, που ενδιαφέρει κάποιον χρήστη, διαφέρει ανάλογα με το επίπεδο της γνώσης και της εκπαίδευσής του. Το αποθετήριο ΚΑΛΛΙΠΟΣ παρέχει, κυρίως, συγγράμματα και μαθησιακά αντικείμενα ακαδημαϊκού επιπέδου. Οι βασικότεροι τύποι ψηφιακού συγγράμματος είναι το προπτυχιακό εγχειρίδιο, το μεταπτυχιακό εγχειρίδιο και η μονογραφία. Τα βιβλία προπτυχιακών σπουδών είναι προσανατολισμένα για την εισαγωγή στα βασικά θέματα ενός τομέα της Επιστήμης, ενώ τα μεταπτυχιακά είναι εξειδικευμένα σε πιο προηγμένες γνώσεις. Η μονογραφία, από την άλλη, αποτελεί την εξειδικευμένη επιστημονική μελέτη ενός συγκεκριμένου θέματος από κάποιον επιστήμονα.

4.2.2 Πρόταση Διαβάσματος Ψηφιακού Συγγράμματος και Εξάμηνο Διδασκαλίας

Ορισμένα ψηφιακά συγγράμματα του αποθετηρίου ΚΑΛΛΙΠΟΣ είναι ειδικά σχεδιασμένα για συγκεκριμένα μαθήματα σε αντίστοιχα επίπεδα εκπαίδευσης. Ο εκάστοτε ακαδημαϊκός καθηγητής μπορεί να επιλέγει ψηφιακά συγγράμματα που αντιστοιχούν στις εκπαιδευτικές ανάγκες και στο επίπεδο μαθήματος, βελτιώνοντας την αποτελεσματικότητα της μάθησης και την ενσωμάτωση στοιχείων στο εκπαιδευτικό υλικό. Τα συγγράμματα αυτά τα περιλαμβάνει στην προτεινόμενη, για το μάθημά του, βιβλιογραφία. Με αυτόν τον τρόπο, οι φοιτητές καθοδηγούνται στη μελέτη συγκεκριμένων συγγραμμάτων που παρέχονται στο

Αποθετήριο, καθώς αυτά περιλαμβάνουν μέρος ή και το σύνολο της εξεταστέας ύλης μαθημάτων τους.

4.2.3 Αλληλεπίδραση και Εμπειρία Χρήστη

Η ενσωμάτωση πολυμεσικού περιεχομένου και αλληλεπιδραστικών στοιχείων στα εκπαιδευτικά ψηφιακά συγγράμματα του Αποθετηρίου ΚΑΛΛΙΠΟΣ κάνει την εκμάθηση πιο ενδιαφέρουσα και αποτελεσματική. Η διαθεσιμότητα συνοδευτικού υλικού μπορεί να κάνει το σύγγραμμα πιο δημοφιλές και προσιτό στους χρήστες του Αποθετηρίου. Όπως αναφέρθηκε σε προηγούμενο κεφάλαιο, το Αποθετήριο ΚΑΛΛΙΠΟΣ παρέχει πρόσβαση σε μαθησιακά αντικείμενα, τα οποία περιλαμβάνουν βίντεο, ήχο, εικόνα, άσκηση, διαφάνειες, διαδραστικό αντικείμενο, προσομοίωση, αλγόριθμο και πίνακα. Με αυτόν τον τρόπο, οι χρήστες αλληλεπιδρούν με το περιεχόμενο σε βαθύτερο επίπεδο, καθώς αξιοποιούν εργαλεία που ενισχύουν την κατανόηση και διευκολύνουν την εκπαιδευτική διαδικασία.

Παράλληλα, η ποιότητα του περιεχομένου που προσφέρεται μπορεί να είναι σημαντική, καθώς ένα εμπλουτισμένο υλικό μπορεί να βοηθήσει τον εκάστοτε φοιτητή να κατανοήσει έννοιες οπτικά ή/και ακουστικά που, σε άλλη περίπτωση, μπορεί να μην κατανοούσε με ευκολία. Η χρήση ποιοτικού περιεχομένου αναβαθμίζει την εμπειρία χρήστη, καθώς βελτιώνει την ευχρηστία και την αποτελεσματικότητα της διαδικασίας μάθησης. Το αποθετήριο ΚΑΛΛΙΠΟΣ παρέχει ενημερωμένο υλικό, το οποίο αξιολογείται πριν δημοσιευθεί, διασφαλίζοντας την αξιοπιστία και την ακρίβεια των πληροφοριών που προσφέρει.

4.2.4 Επίκαιρο και Ανανεωμένο Περιεχόμενο

Ειδικοί, πιθανώς, παρέχουν εισηγήσεις για το πώς μπορεί να ανανεωθεί το υλικό του εκάστοτε συγγράμματος. Στην περίπτωση των ψηφιακών συγγραμμάτων του Αποθετηρίου, η ενσωμάτωση νέων πληροφοριών γίνεται άμεσα σε σχέση με ένα έντυπο σύγγραμμα, ενισχύοντας, με αυτόν τον τρόπο, την αξία και την αξιοπιστία του συγγράμματος. Επίσης, η αναβάθμιση των εικόνων και των επεξηγηματικών

διαγραμμάτων με τα πιο πρόσφατα στοιχεία καθιστά το σύγγραμμα πιο ελκυστικό και ενδιαφέρον στον χρήστη.

4.2.5 Πολυγλωσσικότητα και Ένταξη σε Διεθνή Ευρετήρια

Η μετάφραση συγγραμμάτων σε πολλές γλώσσες μπορεί να ενισχύσει την αξιοποίηση τους, καθώς επιτρέπει σε περισσότερους αναγνώστες να έχουν πρόσβαση στο περιεχόμενό τους. Επίσης, η ένταξη συγγραμμάτων σε διεθνή ευρετήρια, τα καθιστά προσβάσιμα σε ανθρώπους από διάφορες χώρες και κουλτούρες.

Το Αποθετήριο ΚΑΛΛΙΠΟΣ παρέχει πάνω από 900 συγγράμματα και είναι, στο μεγαλύτερο ποσοστό, ελληνόγλωσσα. Τα μεταδεδομένα, ωστόσο, παρέχονται και στην αγγλική γλώσσα. Πρόσφατα, το Αποθετήριο ΚΑΛΛΙΠΟΣ έχει ενταχθεί στο διεθνές Ευρετήριο OERSI. Το ακρωνύμιο OERSI είναι η συντόμευση των λέξεων «Ευρετήριο Αναζήτησης Ανοικτών Εκπαιδευτικών Πόρων» (Open Educational Resources Search Index). Το OERSI αποτελεί μία μηχανή αναζήτησης με δωρεάν εκπαιδευτικό υλικό στην τριτοβάθμια εκπαίδευση. Συνδέει αποθετήρια ανοικτών εκπαιδευτικών πόρων (OER) με κρατική πρωτοβουλία, θεσμικά αποθετήρια πανεπιστημίων και βιβλιοθηκών, όπως και εξειδικευμένα θεματικά αποθετήρια ανοικτών εκπαιδευτικών πόρων (OER). Στο ευρετήριο OERSI δεν αποθηκεύεται περιεχόμενο, αλλά συγχωνεύονται μεταδεδομένα, ώστε να γίνεται ευκολότερη η αναζήτηση. Σε αντίθεση με τα αποθετήρια, το υλικό διατίθενται, αποκλειστικά, μέσω συνδεδεμένων πηγών [23].

Με αυτόν τον τρόπο, το περιεχόμενο του Αποθετηρίου ΚΑΛΛΙΠΟΣ γίνεται πιο δημοφιλές στο ευρύτερο κοινό, καθώς η ένταξη σε διεθνή ευρετήρια το καθιστά πιο αξιόπιστο. Επιπλέον, η μετάφραση τεκμηρίων του Αποθετηρίου σε πολλές γλώσσες μπορεί να αυξήσει την προσβασιμότητα, επιτρέποντας σε ανθρώπους από διαφορετικές χώρες και πολιτισμούς να αξιοποιήσουν και να κατανοήσουν το υλικό που παρέχεται.

4.3 Συλλογή Δεδομένων Χρήσης και Συγγραμμάτων του Αποθετηρίου ΚΑΛΛΙΠΟΣ

Τα περισσότερα δεδομένα που θα χρησιμοποιηθούν για την ανάλυση στοιχείων χρήσης του Αποθετηρίου προέρχονται από μετρήσεις της πλατφόρμας Google Analytics. Οι γραφικές παραστάσεις, οι οποίες αναλύουν τα στοιχεία που αφορούν βασικά στοιχεία των χρηστών, δημιουργούνται με τη βοήθεια του εργαλείου Looker Studio, στο οποίο συνδέονται τα δεδομένα από το Google Analytics. Το Looker Studio [24] είναι ένα εργαλείο, το οποίο μετατρέπει δεδομένα σε προσαρμόσιμους πίνακες εργαλείων και αναφορές, προσφέροντας πολλές δυνατότητες για την οπτικοποίηση και κοινή χρήση δεδομένων. Υπάρχει δυνατότητα σύνδεσης με πολλές πηγές δεδομένων, όπως το Google Analytics, Google Sheets, BigQuery κ.ά. Οι αναφορές γίνονται πιο διαδραστικές με τη βοήθεια φίλτρων προβολής και στοιχείων ελέγχου εύρους ημερομηνιών, επιτρέποντας στους διαχειριστές να εξερευνήσουν τα δεδομένα τους αποτελεσματικά. Οι υπόλοιπες γραφικές παραστάσεις δημιουργούνται με τη βοήθεια της γλώσσας προγραμματισμού Python και συγκεκριμένα με τη βοήθεια της βιβλιοθήκης Matplotlib [25].

Από την πλατφόρμα Google Analytics, αξιοποιούνται τα στοιχεία που παρέχονται για την απόκτηση χρηστών (User acquisition), τα οποία παρουσιάζουν τον τρόπο με τον οποίο οι χρήστες προσήλθαν στο αποθετήριο ανάλογα με την ημερομηνία. Στη συνέχεια, θα αναλυθούν διεξοδικά οι τρόποι με τους οποίους οι χρήστες ανακαλύπτουν το αποθετήριο. Επίσης, αξιοποιούνται στοιχεία που αφορούν την αφοσίωση χρηστών (Engagement), όπως το ποσοστό της αφοσίωσης (Engagement rate) και τα δεδομένα ενεργών χρηστών (Active users). Χρησιμοποιούνται, επιπλέον, μετρήσεις για γεγονότα που συμβαίνουν κατά την περιήγηση του αποθετηρίου (Event count), όπως οι λήψεις (book_downloads) και οι προβολές (book_full_view) συγγραμμάτων. Παράλληλα, μελετώνται στοιχεία για κάθε μία σελίδα του αποθετηρίου (Pages and screens), λαμβάνοντας δεδομένα χρηστών που αφορούν κάθε σύγγραμμα ξεχωριστά. Τα μονοπάτια σελίδων που αφορούν τα συγγράμματα είναι της μορφής /handle/book_handle, όπου book_handle το αναγνωριστικό κάθε συγγράμματος που έχει δημοσιευτεί στο Αποθετήριο.

Τα υπόλοιπα στοιχεία που αφορούν δεδομένα των συγγραμμάτων τα οποία έχουν δημοσιευθεί στο αποθετήριο ΚΑΛΛΙΠΟΣ λαμβάνονται, με τη βοήθεια προγράμματος σε

γλώσσα προγραμματισμού Python, από το API που αποτελείται από σελίδες της μορφής https://repository.kallipos.gr/book-info?handle={book_handle}, όπου book_handle το αναγνωριστικό κάθε συγγράμματος. Επίσης, η σελίδα https://repository.kallipos.gr/books, παρέχει το σύνολο των αναγνωριστικών των συγγραμμάτων που έχουν δημοσιευτεί μέχρι στιγμής στο Αποθετήριο. Με αυτόν τον τρόπο, καταγράφονται στοιχεία που αφορούν κάθε σύγγραμμα ξεχωριστά, δηλαδή τα κεφάλαια που περιέχει, τις θεματικές ενότητες στις οποίες ανήκει, το όνομα του συγγράμματος, το αναγνωριστικό (handle) και την ημερομηνία έκδοσης.

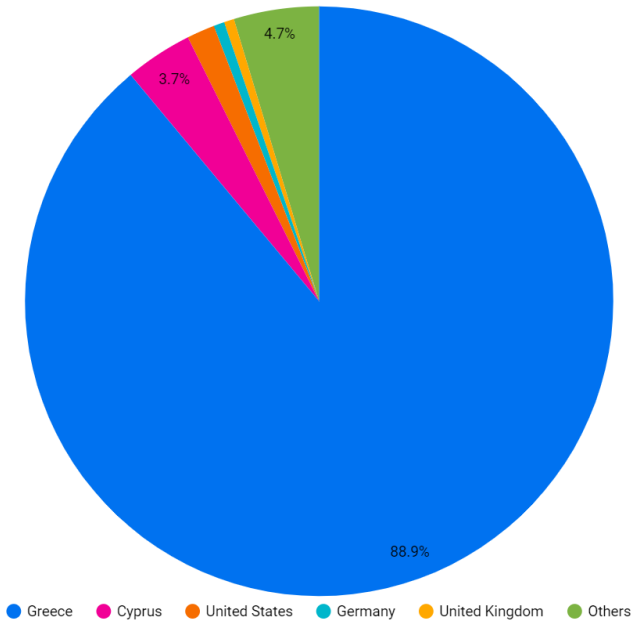
Σε παρακάτω ενότητα, θα συνδυαστούν όλα τα προαναφερθέντα δεδομένα με σκοπό να αναλυθούν και να εντοπιστούν μοτίβα συμπεριφοράς χρηστών και ενδιαφέροντα στοιχεία από τα δημοσιευμένα συγγράμματα του Αποθετηρίου.

4.4 Ανάλυση Δραστηριότητας Χρηστών του Αποθετηρίου ΚΑΛΛΙΠΟΣ

Η ανάλυση της δραστηριότητας χρηστών αποτελεί καθοριστικό παράγοντα για τον εντοπισμό μοτίβων συμπεριφοράς των επισκεπτών ενός αποθετηρίου. Σε αυτό το κεφάλαιο, εξετάζεται η αλληλεπίδραση των χρηστών με το Αποθετήριο ΚΑΛΛΙΠΟΣ, ώστε να γίνουν γνωστά τα στοιχεία που επηρεάζουν τη χρήση του και να δοθούν προτάσεις για τη βελτίωση της λειτουργικότητάς του. Με αυτή τη διαδικασία, το αποθετήριο θα μπορέσει να ανταποκριθεί καλύτερα στις ανάγκες των χρηστών και να είναι πιο χρηστικό.

4.4.1 Γεωγραφική Κατανομή Ενεργών Χρηστών

Μία σημαντική πληροφορία για τη γεωγραφική κατανομή επισκεπτών αποτελεί το σύνολο των ενεργών χρηστών ανά χώρα, δηλαδή τον αριθμό των μοναδικών χρηστών που επισκέφθηκαν τον ιστότοπο από χώρες του κόσμου, για χρονικό διάστημα από 1η Ιουλίου έως 31η Δεκεμβρίου του έτους 2023. Παρακάτω, παρουσιάζεται το διάγραμμα πίττας των ενεργών χρηστών ανά χώρα:

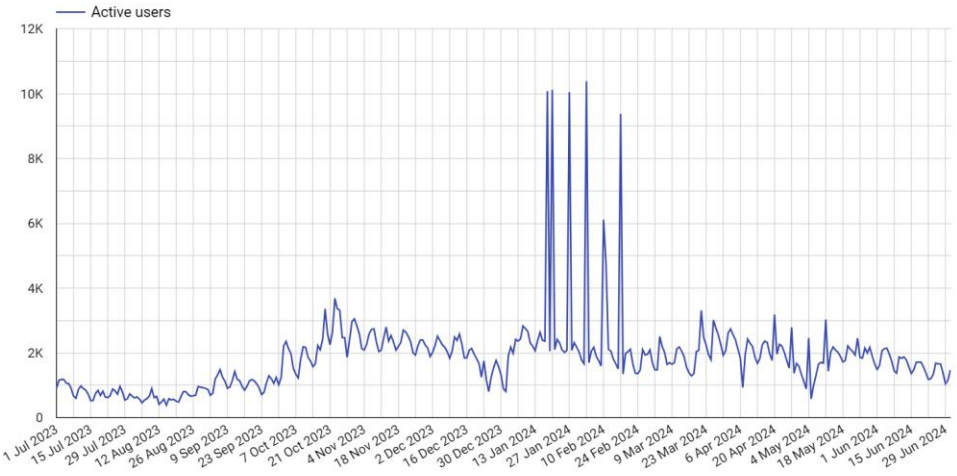


Σχήμα 4.1: Γεωγραφική Κατανομή Ενεργών Χρηστών του Αποθετηρίου ΚΑΛΛΙΠΟΣ

Παρατηρούμε ότι υπάρχουν ενεργοί χρήστες, οι οποίοι επισκέπτονται την ιστοσελίδα από πολλά μέρη του κόσμου. Οι περισσότεροι ενεργοί χρήστες, για το συγκεκριμένο χρονικό διάστημα, προέρχονται βεβαίως από την Ελλάδα με ποσοστό 88.9% και κατά δεύτερο ποσοστό από την Κύπρο (3.7%). Τα παραπάνω συμπεράσματα είναι λογικά, καθώς το Αποθετήριο παρέχει τεκμήρια τα οποία είναι γραμμένα, κυρίως, στην ελληνική γλώσσα. Για αυτόν τον λόγο, αποκλείονται χρήστες οι οποίοι δεν κατανοούν την ελληνική γλώσσα και συνεπώς το Αποθετήριο δεν έχει τη δυνατότητα να αποκτήσει διεθνή εμβέλεια.

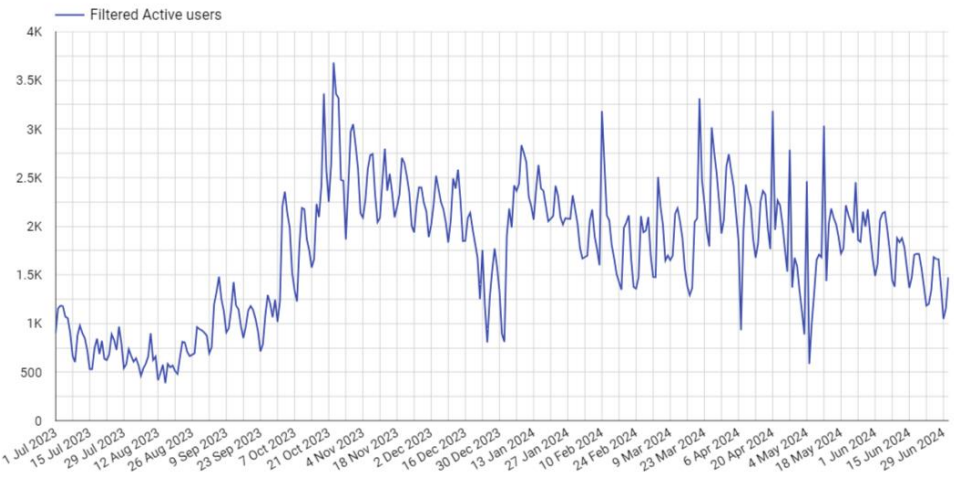
4.4.2 Ημερήσια Κατανομή Ενεργών Χρηστών

Η ανάλυση του αριθμού των ενεργών χρηστών μπορεί να βοηθήσει στον εντοπισμό συγκεκριμένων χρονικών περιόδων κατά τις οποίες το Αποθετήριο εμφανίζει, είτε μειωμένη, είτε αυξημένη κίνηση. Με αυτόν τον τρόπο, αξιολογείται η ζήτηση του περιεχομένου που διαθέτει το Αποθετήριο σε συγκεκριμένο χρονικό διάστημα. Παρακάτω, απεικονίζεται το διάγραμμα του συνόλου των ενεργών χρηστών (Active Users), δηλαδή του αριθμού των μοναδικών χρηστών που επισκέφθηκαν τον ιστότοπο σε συγκεκριμένη περίοδο σύνδεσης για χρονικό διάστημα από 1η Ιουλίου του έτους 2023 έως 1η Ιουλίου του έτους 2024:



Σχήμα 4.2: Ημερήσια Κατανομή Ενεργών Χρηστών του Αποθετηρίου ΚΑΛΛΙΠΟΣ

Οι απότομες αυξήσεις τιμών στο διάγραμμα οφείλονται στη συλλογή των μεταδεδομένων νέων βιβλίων για το διεθνές ευρετήριο OERSI, το οποίο αναφέρθηκε σε προηγούμενο κεφάλαιο. Για την καλύτερη ευκρίνεια και ανάλυση, παρουσιάζεται το διάγραμμα του συνόλου των ενεργών χρηστών, το οποίο έχει φιλτραριστεί στα σημεία με τις απότομες αυξήσεις τιμών:



Σχήμα 4.3: Ημερήσια Κατανομή Ενεργών Χρηστών του Αποθετηρίου ΚΑΛΛΙΠΟΣ (Φιλτραρισμένα Δεδομένα)

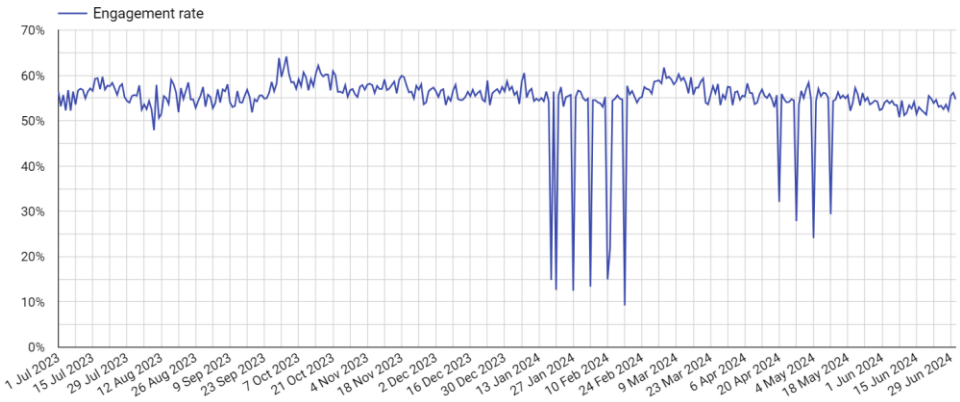
Είναι άξιο παρατήρησης, ότι οι ενεργοί χρήστες του αποθετηρίου μειώνονται αισθητά κατά τους καλοκαιρινούς μήνες, ενώ αυξάνονται κατά τους φθινοπωρινούς και ανοιξιάτικους μήνες. Αυτό είναι λογικό, καθώς το αποθετήριο περιλαμβάνει, κυρίως, συγγράμματα ακαδημαϊκών ιδρυμάτων, τα οποία είναι απαραίτητα σε περιόδους που οι Σχολές είναι ανοικτές και ιδίως σε εξεταστικές περιόδους και στην αρχή των εξαμήνων. Η μέγιστη τιμή των ενεργών χρηστών (3683) παρατηρείται στις 23 Οκτωβρίου του έτους 2023. Η συγκεκριμένη ημερομηνία περιλαμβάνεται στην περίοδο εισαγωγής σε διαλέξεις και προετοιμασίας για εργασίες, γι'αυτό υπάρχει αυξημένη κίνηση στο Αποθετήριο. Από την άλλη μεριά, η ελάχιστη τιμή των ενεργών χρηστών (388) παρατηρείται στις 15 Αυγούστου του έτους 2023. Η συγκεκριμένη ημερομηνία περιλαμβάνεται στην περίοδο καλοκαιρινών διακοπών και αποτελεί Εθνική Αργία, οπότε είναι λογική η ελάχιστη ζήτηση περιεχομένου από το Αποθετήριο.

4.4.3 Ημερήσιο Ποσοστό Αφοσίωσης Χρηστών

Άλλη μία σημαντική παράμετρος, η οποία είναι άξια μελέτης, είναι το ποσοστό αφοσίωσης (Engagement rate). Υπολογίζει το ποσοστό των επισκεπτών που αλληλεπιδρά με ένα τμήμα περιεχομένου, όπως να κάνουν κλικ στον σύνδεσμο ή να πραγματοποιούν περιήγηση σε σημαντικό χρόνο. Με άλλα λόγια, το ποσοστό αφοσίωσης (ΠΑ) ανιχνεύει πόσο ενεργά συμμετέχει ο εκάστοτε χρήστης στο περιεχόμενο της ιστοσελίδας. Υπολογίζεται με τον παρακάτω τύπο [9]:

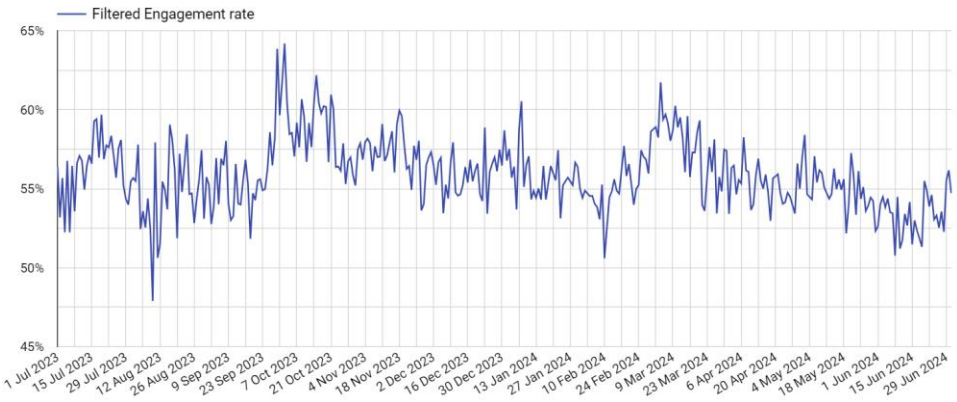
$$ΠΑ = \frac{Αφοσιωμένοι\ Περίοδοι\ Σύνδεσης}{Περίοδοι\ Σύνδεσης\ σε\ Καθορισμένη\ Χρονική\ Περίοδο} \cdot 100\% \qquad (4.1)$$

Παρακάτω, απεικονίζεται το διάγραμμα του ποσοστού αφοσίωσης χρηστών για χρονικό διάστημα από 1η Ιουλίου του έτους 2023 έως 1η Ιουλίου του έτους 2024:



Σχήμα 4.4: Ημερήσιο Ποσοστό Αφοσίωσης Χρηστών του Αποθετηρίου ΚΑΛΛΙΠΟΣ

Όπως και πριν, οι απότομες μειώσεις ποσοστών στο διάγραμμα οφείλονται στη συλλογή των μεταδεδομένων νέων βιβλίων για το διεθνές ευρετήριο OERSI. Για την καλύτερη ευκρίνεια και ανάλυση, παρουσιάζεται το διάγραμμα του ποσοστού αφοσίωσης χρηστών, το οποίο έχει φιλτραριστεί στα σημεία με τις απότομες μειώσεις τιμών:



Σχήμα 4.5: Ημερήσιο Ποσοστό Αφοσίωσης Χρηστών του Αποθετηρίου ΚΑΛΛΙΠΟΣ (Φιλτραρισμένα Δεδομένα)

Από τις αρχές Ιουλίου έως τις αρχές Αυγούστου, το ποσοστό αφοσίωσης των χρηστών κυμαίνεται μεταξύ 50% και 60%. Έπειτα, παρουσιάζεται μία κατακόρυφη πτώση του ποσοστού αφοσίωσης με το ελάχιστο ποσοστό να εμφανίζεται στις 9 Αυγούστου του έτους 2023 (47,89%). Στη συνέχεια, οι τιμές του ποσοστού αφοσίωσης αυξομειώνονται γύρω από το ποσοστό 55%, ενώ αυξάνονται στις αρχές Οκτωβρίου. Από το παραπάνω διάγραμμα είναι εμφανές, ότι το υψηλότερο ποσοστό αφοσίωσης (Engagement Rate) παρουσιάζεται αρχές Οκτωβρίου, πιθανώς, γιατί οι Σχολές ξεκινούν μαθήματα εκείνη την περίοδο και αρκετοί καθηγητές προτείνουν στους φοιτητές να μελετούν από βιβλία που περιέχονται στο Αποθετήριο. Η μέγιστη τιμή του ποσοστού αφοσίωσης παρατηρείται στις 2 Οκτωβρίου του έτους 2023 (64.19%). Κατά τη διάρκεια του Νοεμβρίου έως και τον Ιανουάριο, το ποσοστό αφοσίωσης κυμαίνεται γύρω από τις τιμές 53% με 60%. Το ποσοστό αφοσίωσης μειώνεται στα μέσα του Φεβρουαρίου, ενώ αυξάνεται στις αρχές του Μαρτίου, φτάνοντας το ποσοστό 62% στις 4 Μαρτίου του έτους 2024. Από τα μέσα του Μαρτίου έως και το τέλος του Ιουνίου το ποσοστό κυμαίνεται από 53% έως 60%, παρουσιάζοντας μία σταθερά καθοδική τάση.

4.4.4 Ημερήσια Απόκτηση Χρηστών ανά Πηγή Επισκεψιμότητας

Ενδιαφέρον παρουσιάζει η μελέτη του αριθμού των χρηστών που επισκέπτονται τον ιστότοπο για πρώτη φορά, κατηγοριοποιημένων ανά ομάδα καναλιών. Κανάλια ονομάζονται οι αφετηρίες επισκεψιμότητας του Αποθετηρίου και επιτρέπουν την παρακολούθηση απόδοσης όλων των καναλιών που δημιουργούν κίνηση χρηστών στον ιστότοπο.

Παρακάτω, περιγράφονται τα προκαθορισμένα κανάλια, δηλαδή οι αφετηρίες επισκεψιμότητας ιστοσελίδας [9]:

Οργανική Αναζήτηση (Organic Search)

Η απόκτηση χρηστών με οργανική αναζήτηση αναφέρεται στους χρήστες που φτάνουν στον ιστότοπο μέσω απλήρωτων αποτελεσμάτων μηχανών αναζήτησης, περιλαμβάνοντας κίνηση από μηχανές αναζήτησης όπως Google, Bing, Yahoo κ.λπ.

Απευθείας (Direct)

Η απόκτηση χρηστών με απευθείας σύνδεση αναφέρεται στους χρήστες που φτάνουν στον ιστότοπο, μέσω ενός αποθηκευμένου συνδέσμου ή εισάγοντας τη διεύθυνση URL του ιστότοπου.

Παραπομπή (Referral)

Η απόκτηση χρηστών με παραπομπή αναφέρεται στους χρήστες που φτάνουν στον ιστότοπο μέσω ενός συνδέσμου από τρίτους ιστοτόπους, όπως ιστολόγια και ιστότοπους ειδήσεων.

Οργανική Κοινωνική (Organic Social)

Η απόκτηση χρηστών με οργανική κοινωνική αναζήτηση αναφέρεται στους χρήστες που φτάνουν στον ιστότοπο μέσω συνδέσμων, χωρίς διαφημίσεις, από ιστότοπους κοινωνικής δικτύωσης, όπως Facebook και Twitter.

Οργανικό Βίντεο (Organic Video)

Η απόκτηση χρηστών με οργανικό βίντεο αναφέρεται στους χρήστες που φτάνουν στον ιστότοπο μέσω συνδέσμων χωρίς διαφημίσεις από ιστοσελίδες που παρέχουν βίντεο, όπως Youtube, Tik Tok και Vimeo.

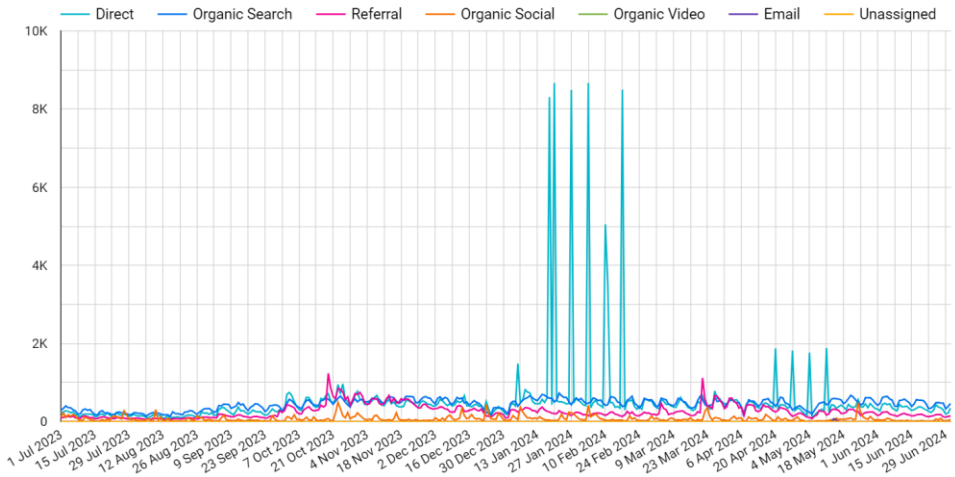
Ηλεκτρονική Διεύθυνση (Email)

Η απόκτηση χρηστών με ηλεκτρονική διεύθυνση αναφέρεται στους χρήστες που φτάνουν στον ιστότοπο μέσω συνδέσμων σε email.

Χωρίς Ανάθεση (Unassigned)

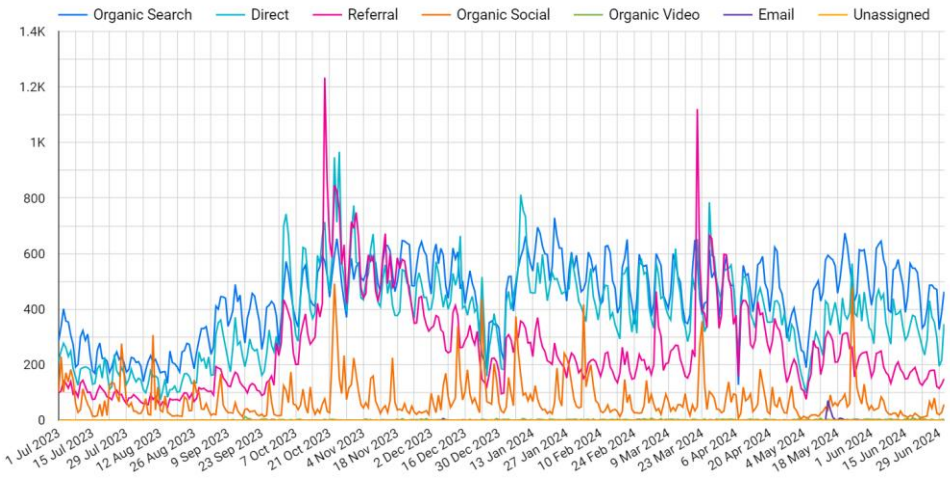
Η απόκτηση χρηστών χωρίς ανάθεση αναφέρεται στους χρήστες για τους οποίους δεν υπάρχουν άλλοι κανόνες καναλιού που αντιστοιχούν σε δεδομένα συμβάντος.

Παρακάτω, απεικονίζεται το διάγραμμα απόκτησης χρηστών στο Αποθετήριο ΚΑΛΛΙΠΟΣ για χρονικό διάστημα από 1η Ιουλίου του έτους 2023 έως 1η Ιουλίου του έτους 2024:



Σχήμα 4.6: Ημερήσια Απόκτηση Χρηστών του Αποθετηρίου ΚΑΛΛΙΠΟΣ

Όπως και πριν, οι απότομες αυξήσεις τιμών στο διάγραμμα απόκτησης χρηστών μέσω απευθείας σύνδεσης οφείλονται στη συλλογή των μεταδεδομένων νέων βιβλίων για το διεθνές ευρετήριο OERSI. Για την καλύτερη ευκρίνεια και ανάλυση, παρουσιάζεται το διάγραμμα απόκτησης χρηστών, το οποίο έχει φιλτραριστεί στα σημεία με τις απότομες αυξήσεις τιμών:



Σχήμα 4.7: Ημερήσια Απόκτηση Χρηστών του Αποθετηρίου ΚΑΛΛΙΠΟΣ (Φιλτραρισμένα Δεδομένα)

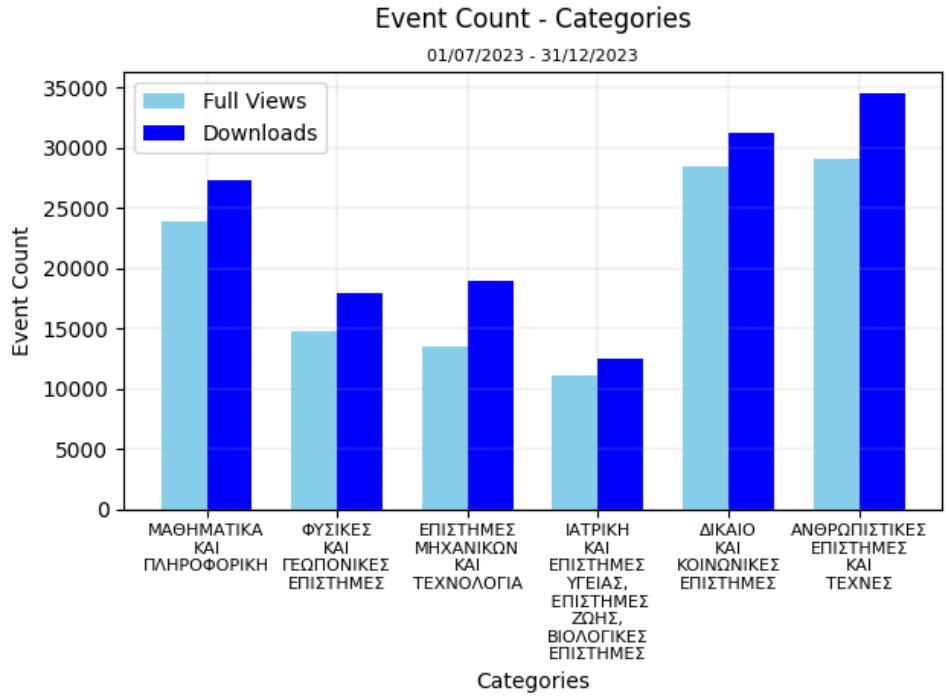
Όπως φαίνεται στο παραπάνω διάγραμμα, το πλήθος νέων χρηστών, κατά κύριο λόγο, επισκέφθηκαν το Αποθετήριο απευθείας, μέσω παραπομπής, οργανικής και οργανικής κοινωνικής αναζήτησης. Στις περιπτώσεις που ο νέος χρήστης προήλθε από οργανική αναζήτηση είτε απευθείας, είτε μέσω παραπομπής, είναι φανερό ότι το αντίστοιχο σύνολο νέων χρηστών εμφανίζει τοπικό ελάχιστο στα μέσα του Αυγούστου, στο τέλος του Δεκεμβρίου και στις αρχές του Μαΐου. Από την άλλη, παρατηρείται τοπικό μέγιστο στα μέσα του Οκτωβρίου, στις αρχές του Ιανουαρίου και στο τέλος του Μαΐου. Η προηγούμενη παρατήρηση είναι λογική, καθώς η αναζήτηση του Αποθετηρίου είναι λιγότερο δημοφιλής σε περιόδους διακοπών, όπως το καλοκαίρι, τα Χριστούγεννα και το Πάσχα, καθώς δεν είναι ανοικτές οι ακαδημαϊκές Σχολές. Αντίθετα, η αναζήτηση είναι περισσότερο δημοφιλής στην αρχή κάθε εξαμήνου και πριν τις εξεταστικές περιόδους, όταν οι ακαδημαϊκοί καθηγητές προτείνουν στους φοιτητές κάποιο ψηφιακό σύγγραμμα προς μελέτη. Στην περίπτωση της οργανικής κοινωνικής αναζήτησης, ο μέσος όρος του συνόλου

χρηστών είναι σχεδόν ίδιος, ενώ οι τιμές αυτές είναι σχεδόν πάντα λιγότερες από τις άλλες μεθόδους απόκτησης χρηστών και δεν ξεπερνούν τους 500 νέους χρήστες για μία συγκεκριμένη ημερομηνία. Τα παραπάνω αποδεικνύουν ότι η αναζήτηση του Αποθετηρίου, μέσω κοινωνικών σελίδων, είναι η λιγότερο διαδεδομένη. Ένας τρόπος για την αύξηση επίσκεψης στο Αποθετήριο μέσω της οργανικής κοινωνικής αναζήτησης, είναι η αύξηση των αναρτήσεων στα κοινωνικά δίκτυα με απώτερο σκοπό την ανάδειξη του περιεχομένου και την αύξηση της δημοτικότητας του Αποθετηρίου.

4.4.5 Προβολές και Λήψεις Ψηφιακών Συγγραμμάτων ανά Θεματική Ενότητα

Η ανάλυση των ενεργειών που πραγματοποιούν οι χρήστες στο Αποθετήριο, όπως η αποθήκευση ή οι προβολές ψηφιακών συγγραμμάτων, υποδεικνύει τις προθέσεις των χρηστών και τη μέθοδο περιήγησής τους στην ιστοσελίδα. Αποτελεί τρόπο για την αναβάθμιση της εμπειρίας των χρηστών και τη βελτίωση του περιεχομένου του αποθετηρίου.

Παρακάτω, παρατίθεται ένα ραβδόγραμμα κατατοπιστικό για τον τρόπο που επιλέγουν οι χρήστες να προσπελάσουν ένα ψηφιακό σύγγραμμα του Αποθετηρίου ανά θεματική ενότητα για χρονικό διάστημα από 1η Ιουλίου έως 31η Δεκεμβρίου του έτους 2023:



Σχήμα 4.8: Προβολές/Λήψεις ανά Θεματική Ενότητα του Αποθετηρίου ΚΑΛΛΙΠΟΣ

Υπάρχουν δύο τρόποι προσπέλασης ψηφιακών συγγραμμάτων εντός του Αποθετηρίου, τα οποία αποτυπώνονται στην παραπάνω γραφική παράσταση: Η προβολή πλήρους περιεχομένου συγγράμματος και η προβολή μέσω λήψης και αποθήκευσης συγγράμματος.

Στην περίπτωση προβολής πλήρους περιεχομένου συγγράμματος, καταγράφεται ο συνολικός αριθμός προβολής ηλεκτρονικών βιβλίων, ενώ στην περίπτωση προβολής μέσω αποθήκευσης συγγράμματος καταγράφεται ο συνολικός αριθμός λήψεων και αποθήκευσης του περιεχομένου, τοπικά σε συσκευή.

Αρχικά, από το διάγραμμα που προκύπτει για το συγκεκριμένο χρονικό διάστημα, είναι φανερό ότι η ποσοστιαία κατανομή των αποθηκεύσεων και των προβολών συγγραμμάτων είναι υψηλότερη στην κατηγορία «Ανθρωπιστικές Επιστήμες και Τέχνες» (ποσοστό αποθηκεύσεων: 24.236%, ποσοστό προβολών: 24.068%), έπειτα στην κατηγορία «Δίκαιο και Κοινωνικές Επιστήμες» (ποσοστό αποθηκεύσεων: 21.900%, ποσοστό προβολών: 23.609%) και μετά ακολουθούν οι υπόλοιπες κατηγορίες.

Σε όλες τις θεματικές ενότητες, η λήψη και αποθήκευση περιεχομένου είναι πιο συχνή από την προβολή του πλήρους περιεχομένου για το συγκεκριμένο χρονικό διάστημα, καθώς κάποιοι χρήστες μπορεί να προτιμούν την αποθήκευση του βιβλίου για μελλοντική ανάγνωση χωρίς να χρειάζεται πρόσβαση στο αποθετήριο για τη μετέπειτα προβολή. Από την άλλη μεριά, η προβολή πλήρους περιεχομένου μπορεί να προτιμάται, καθώς κάποιοι χρήστες θέλουν άμεση προσπέλαση του περιεχομένου χωρίς να καταναλώσουν πόρους για αποθήκευση. Επίσης, κάποιοι προτιμούν να προβάλουν το βιβλίο αρχικά, για να δουν αν τους καλύπτει το περιεχόμενο και μετά, αν όντως είναι στις προτιμήσεις τους, να το αποθηκεύουν τοπικά.

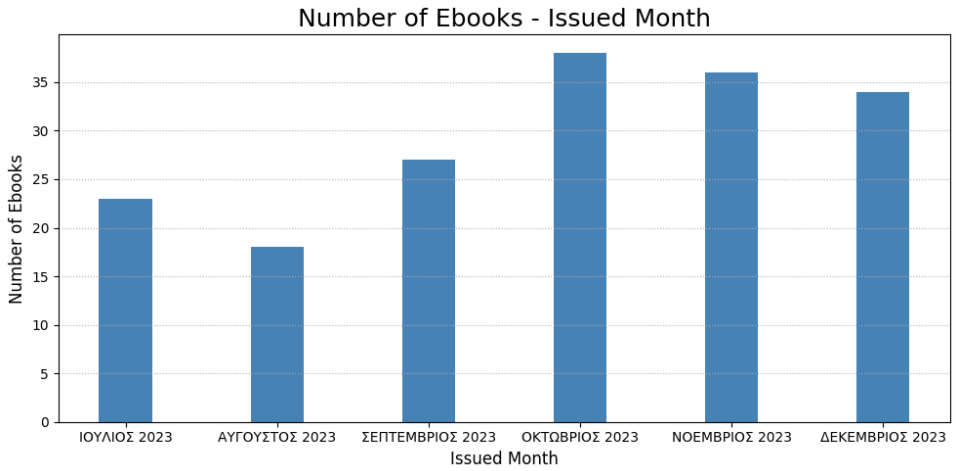
Σε γενικές γραμμές, δεν υπάρχουν κατηγορίες στις οποίες ο αριθμός προβολών να είναι χαμηλότερος από τον αριθμό αποθηκεύσεων, για το συγκεκριμένο χρονικό διάστημα.

4.5 Ανάλυση Δεδομένων Συγγραμμάτων του Αποθετηρίου ΚΑΛΛΙΠΟΣ

Η ανάλυση των δεδομένων συγγραμμάτων αναφέρεται στον αριθμό των τεκμηρίων (συνολικά και ανά κατηγορία) τα οποία είναι δημοσιευμένα στο Αποθετήριο, είναι δε καθοριστική για την οργάνωση, την ενημέρωση και τη διανομή πληροφοριών σχετικά με το περιεχόμενο που δημοσιοποιείται. Επειδή η Δράση ΚΑΛΛΙΠΟΣ είναι σε εξέλιξη, παρατηρείται μία συνεχής αυξητική τάση στο περιεχόμενο του Αποθετηρίου.

4.5.1 Νέα Συγγράμματα ανά Μήνα Δημοσίευσης

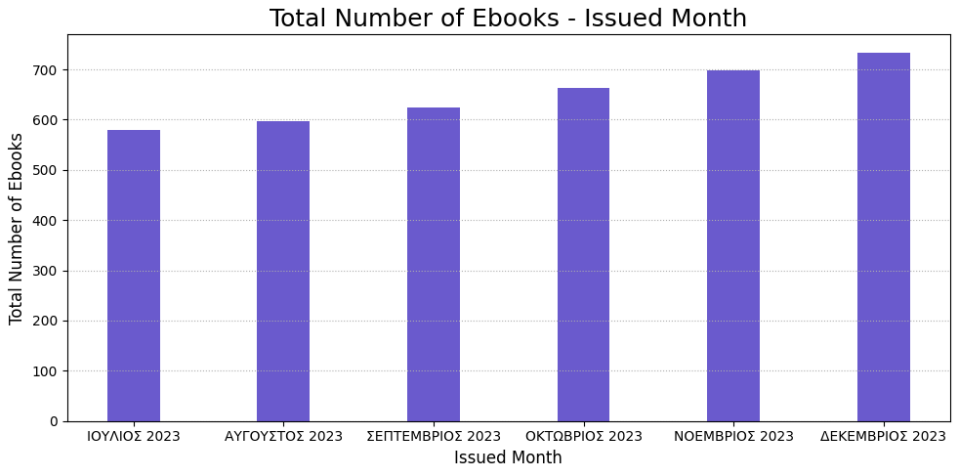
Το παρακάτω ραβδόγραμμα παρουσιάζει τον αριθμό νέων συγγραμμάτων που δημοσιεύθηκαν, ανά μήνα, από τον Ιούλιο έως τον Δεκέμβριο του έτους 2023:



Σχήμα 4.9: Συγγράμματα ανά Μήνα Δημοσίευσης του Αποθετηρίου ΚΑΛΛΙΠΟΣ

Το παραπάνω διάγραμμα ράβδων είναι κατατοπιστικό ως προς την ποσότητα των νέων τεκμηρίων που δημοσιοποιούνται στο αποθετήριο. Τα περισσότερα ψηφιακά συγγράμματα δημοσιεύθηκαν τον Οκτώβριο του έτους 2023 (δημοσιεύθηκαν 38 συγγράμματα), ενώ τα λιγότερα τον Αύγουστο του ίδιου έτους (δημοσιεύθηκαν 18 συγγράμματα). Γενικά, δεν υπάρχει σταθερός αριθμός ψηφιακών συγγραμμάτων που δημοσιεύονται ανά μήνα, επειδή η δεύτερη φάση του Έργου ΚΑΛΛΙΠΟΣ βρίσκεται σε εξέλιξη και υπάρχει διακύμανση στο νέο υλικό που ετοιμάζεται για ανάρτηση.

Μία παραλλαγή του παραπάνω διαγράμματος είναι το ραβδόγραμμα, το οποίο αφορά τον συνολικό αριθμό των ενεργών συγγραμμάτων που έχουν δημοσιευτεί μέχρι το τέλος του εκάστοτε μήνα από τον Ιούλιο έως τον Δεκέμβριο του έτους 2023 και παρουσιάζεται παρακάτω:



Σχήμα 4.10: Συνολικά Συγγράμματα ανά Μήνα Δημοσίευσης του Αποθετηρίου ΚΑΛΛΙΠΟΣ

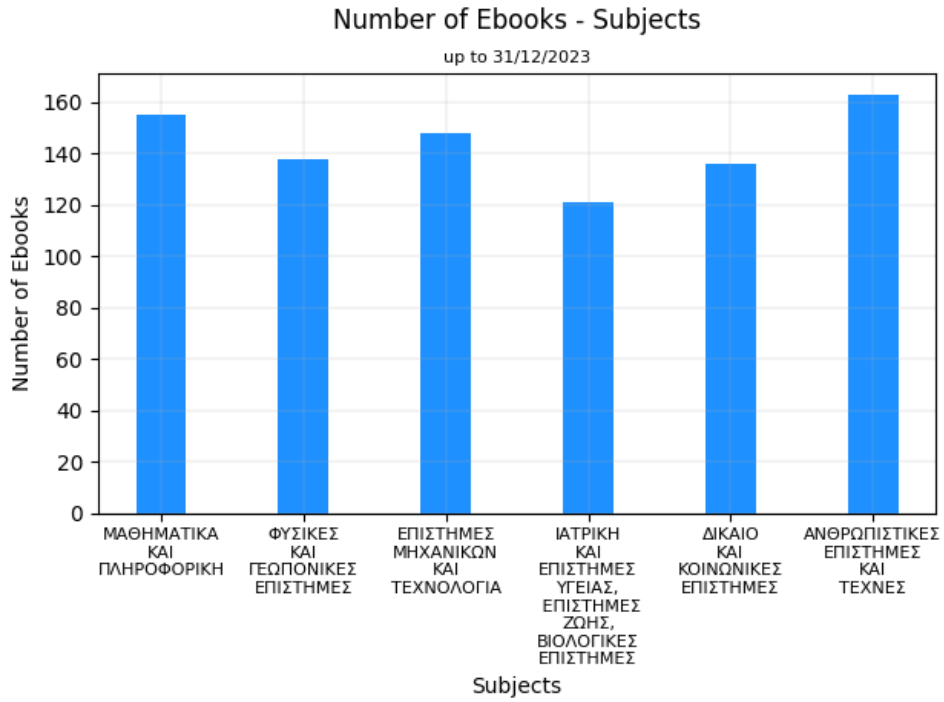
Η συνεχής προσθήκη νέων συγγραμμάτων στο Αποθετήριο δίνει τη δυνατότητα στους χρήστες να έχουν περισσότερες επιλογές περιεχομένου. Με αυτόν τον τρόπο, βελτιώνεται η εμπειρία χρήστη και η ποσότητα περιεχομένου, οπότε αποτελεί έναν λόγο που μπορεί να οδηγήσει σε αυξημένη αναζήτηση εντός του Αποθετηρίου.

Το συμπέρασμα αυτό αποδεικνύεται, αν συγκριθεί το παραπάνω ραβδόγραμμα με το διάγραμμα αφοσίωσης χρηστών που αναλύθηκε σε παραπάνω ενότητα. Στο τέλος του έτους 2023, το αποθετήριο παρουσιάζει μεγάλη ποικιλία περιεχομένου, καθώς περιλαμβάνει τα περισσότερα δυνατά συγγράμματα σε σχέση με τους υπόλοιπους μήνες του έτους. Στο διάγραμμα της αφοσίωσης χρηστών, το οποίο αναλύθηκε σε προηγούμενη ενότητα, παρατηρούμε ότι υπάρχει μία αυξανόμενη πορεία του ποσοστού αφοσίωσης στην ιστοσελίδα προς το τέλος του έτους 2023, πράγμα που σημαίνει ότι, ο εκάστοτε χρήστης, συμμετέχει πιο ενεργά στο περιεχόμενο της ιστοσελίδας εκείνη τη χρονική περίοδο. Ένας

λόγος, λοιπόν, που μπορεί να συμβαίνει αυτό, είναι η αύξηση της ποικιλίας συγγραμμάτων και η συνακόλουθη αύξηση ζήτησης και χρήσης τους.

4.5.2 Αριθμός Ψηφιακών Συγγραμμάτων ανά Κύρια Θεματική Κατηγορία

Παρακάτω παρουσιάζεται το ραβδόγραμμα με τον συνολικό αριθμό των ψηφιακών συγγραμμάτων που παρουσιάζονται στο Αποθετήριο, ανά κύρια θεματική κατηγορία, έως την τελευταία μέρα του έτους 2023:



Σχήμα 4.11: Συγγράμματα ανά Κύρια Θεματική Κατηγορία του Αποθετηρίου ΚΑΛΛΙΠΟΣ

Από το παραπάνω διάγραμμα, προκύπτει ότι οι κύριες κατηγορίες «Ανθρωπιστικές Επιστήμες και Τέχνες» (αριθμός συγγραμμάτων: 163) και «Μαθηματικά και Πληροφορική» (αριθμός συγγραμμάτων: 155) εκπροσωπούνται περισσότερο στα συγγράμματα του Αποθετηρίου. Αυτό είναι λογικό, καθώς οι συγκεκριμένες κατηγορίες

αποτελούν πεδία που καλύπτουν τα περισσότερα ακαδημαϊκά και ερευνητικά Ιδρύματα της χώρας.

Αντίθετα, οι κατηγορίες «Ιατρική και Επιστήμες Υγείας, Επιστήμες Ζωής, Βιολογικές Επιστήμες» (αριθμός συγγραμμάτων: 121) και «Δίκαιο και Κοινωνικές Επιστήμες» (αριθμός συγγραμμάτων: 136) εκπροσωπούνται λιγότερο στα συγγράμματα του Αποθετηρίου.

Σε γενικά πλαίσια, υπάρχει μία ισορροπημένη κατανομή των ψηφιακών συγγραμμάτων μεταξύ των διαφόρων επιστημονικών πεδίων, γεγονός που αποδεικνύει τη σχετικά ομοιόμορφη παροχή εκπαιδευτικού υλικού και την πρόσβαση σε ποικιλία περιεχομένου.

5 Πρόβλεψη Νέων Χρηστών του Αποθετηρίου ΚΑΛΛΙΠΟΣ

5.1 Εισαγωγή

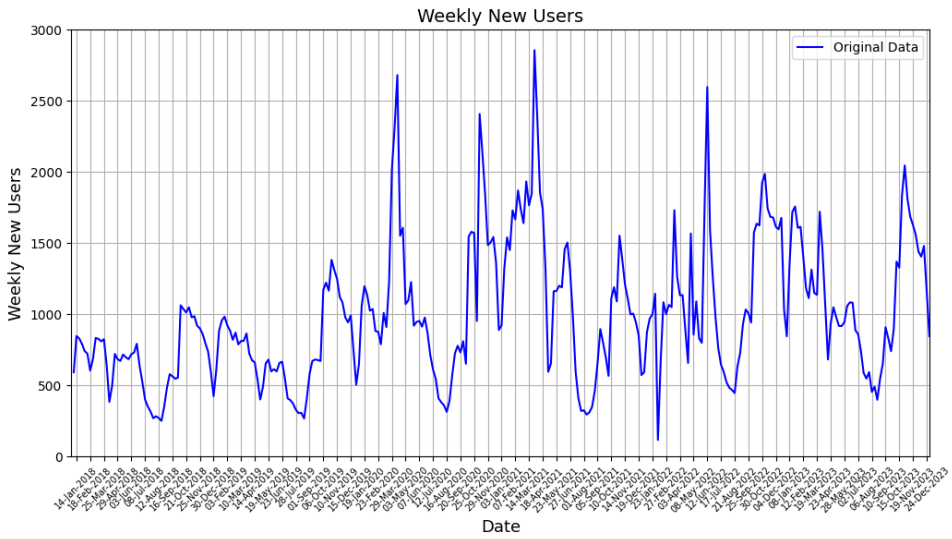
Η πρόβλεψη χρονοσειρών (Time Series Forecasting) αποτελεί μία σημαντική διαδικασία, στην οποία προβλέπονται στοιχεία σε συγκεκριμένα χρονικά διαστήματα. Συγκεκριμένα, η πρόβλεψη επισκεψιμότητας μπορεί να βοηθήσει στην καλύτερη διαχείριση μίας πλατφόρμας, καθώς επιτρέπει την πρόβλεψη μελλοντικών χρονικών περιόδων με αυξημένη κίνηση. Με αυτόν τον τρόπο, οι διαχειριστές της πλατφόρμας μπορούν να προγραμματίζουν τη συντήρησή της σε χρονικές περιόδους που δεν αναμένεται μεγάλη προσέλευση. Μία ενδιαφέρουσα επέκταση αποτελεί η πρόβλεψη επισκεψιμότητας σε συγκεκριμένες θεματολογίες της πλατφόρμας, καθώς έτσι θα γίνουν αντιληπτές οι οποιεσδήποτε προτιμήσεις χρηστών ως προς το περιεχόμενό της.

Η στατιστική ανάλυση παρέχει εργαλεία τα οποία επιτρέπουν την πρόβλεψη μελλοντικών τιμών μίας χρονοσειράς βάσει παρελθοντικών δεδομένων. Σε αυτή την ενότητα, θα εξεταστούν οι μέθοδοι του Εποχιακού Αυτοπαλινδρομικού Ολοκληρωμένου Κινούμενου Μέσου Όρου με Εξωγενείς Παράγοντες (SARIMAX), της Απλής Εκθετικής Εξομάλυνσης (Simple Exponential Smoothing), της Διπλής Εκθετικής Εξομάλυνσης (Double Exponential Smoothing) και της Τριπλής Εκθετικής Εξομάλυνσης (Triple Exponential Smoothing). Αυτές οι μέθοδοι είναι ευρέως γνωστές για την πρόβλεψη και την κατανόηση της συμπεριφοράς δεδομένων με την πάροδο του χρόνου. Στην προκειμένη περίπτωση, θα εφαρμοστούν οι παραπάνω μέθοδοι για την πρόβλεψη του μέσου όρου των νέων χρηστών του αποθετηρίου ΚΑΛΛΙΠΟΣ.

Επίσης, η Μηχανική Μάθηση αποτελεί έναν κλάδο, ο οποίος μπορεί να προσφέρει αποδοτικούς αλγόριθμους για την πρόβλεψη μελλοντικών τιμών μίας μεταβλητής βάσει παρελθοντικών δεδομένων. Η χρήση μεθόδων Μηχανικής Μάθησης προϋποθέτει τη δημιουργία και την εκπαίδευση ενός πολύπλοκου μοντέλου, το οποίο μπορεί να διαχειριστεί σύνθετα σύνολα δεδομένων. Σε αυτή την ενότητα, θα περιγραφεί όλη η διαδικασία προεπεξεργασίας και πρόβλεψης του μέσου όρου των νέων χρηστών του

αποθετηρίου ΚΑΛΛΙΠΟΣ με τη βοήθεια της Μηχανικής Μάθησης και συγκεκριμένα της μεθόδου Μηχανής Ενίσχυσης Ελαφριάς Κλίσης (Light Gradient Boosting Machine).

Τα δεδομένα που θα χρησιμοποιηθούν αποτελούν τον μέσο όρο των νέων χρηστών που επισκέπτονται το αποθετήριο ΚΑΛΛΙΠΟΣ, για κάθε εβδομάδα, από το έτος 2018 έως το 2023, όπως φαίνεται στην παρακάτω γραφική παράσταση:



Σχήμα 5.1: Εβδομαδιαίος Μέσος Όρος Νέων Χρηστών του Αποθετηρίου ΚΑΛΛΙΠΟΣ

Είναι φανερό, ότι υπάρχουν κάποια επαναλαμβανόμενα μοτίβα στις τιμές του μέσου όρου των νέων χρηστών ανά έτος. Για παράδειγμα, οι νέοι χρήστες που επισκέπτονται το αποθετήριο μειώνονται ραγδαία τις εβδομάδες του καλοκαιριού, καθώς τότε δεν υπάρχει ενεργή ακαδημαϊκή περίοδος. Συγκεκριμένα, τη δεύτερη εβδομάδα του Αυγούστου, κάθε έτους, υπάρχει η πιο μειωμένη προσέλευση νέων χρηστών με τη μέση τιμή τους να κυμαίνεται κάτω από 500. Αντίθετα, οι περισσότεροι χρήστες προτιμούν να επισκέπτονται το αποθετήριο τις φθινοπωρινές και ανοιξιάτικες εβδομάδες και, κυρίως, πριν ή κατά τη διάρκεια των εξεταστικών περιόδων. Είναι άξιο παρατήρησης ότι τον Οκτώβριο και τον Μάρτιο, κάθε έτους, υπάρχει η τάση να αυξάνονται οι νέοι χρήστες στο Αποθετήριο. Αυτό μπορεί να συμβαίνει, επειδή τότε ξεκινούν οι αντίστοιχες ακαδημαϊκές διαδικασίες του χειμερινού και του εαρινού εξαμήνου. Η μεγαλύτερη προσέλευση σημειώθηκε την τελευταία εβδομάδα του Μαρτίου του έτους 2021 με 2856 νέους χρήστες κατά μέσο όρο.

Στη συνέχεια, θα προβλεφθούν οι τιμές του μέσου όρου των νέων χρηστών για όλες τις εβδομάδες του έτους 2023. Από την ανάλυση, οι τιμές πρόβλεψης των νέων χρηστών αναμένεται να πλησιάζουν τα πραγματικά δεδομένα και να επιβεβαιωθούν τα παραπάνω μοτίβα για τα δεδομένα που προβλέπονται. Η αποδοτικότητα του μοντέλου, όμως, αναμένεται να έχει περιθώρια βελτίωσης με την εισαγωγή περισσότερων παραμέτρων και δεδομένων.

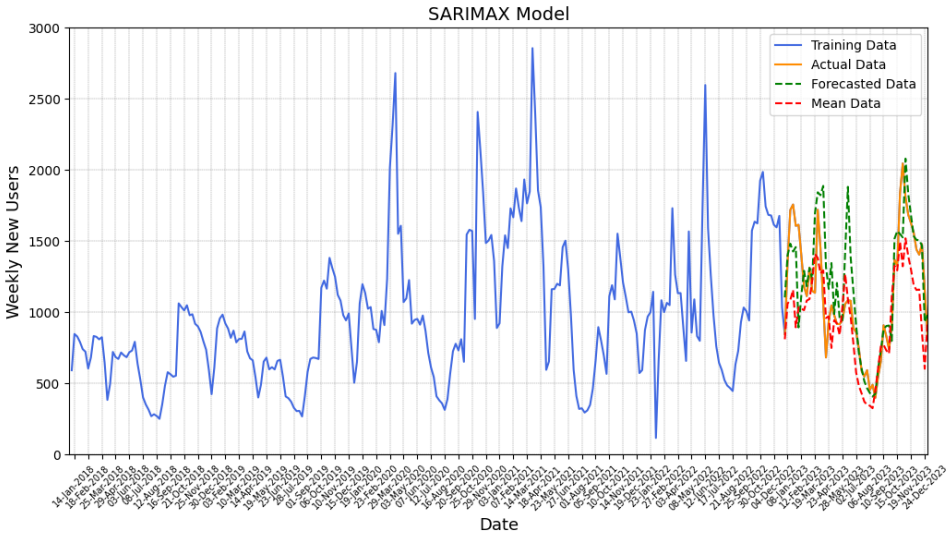
5.2 Πρόβλεψη Νέων Χρηστών με τη Μέθοδο SARIMAX

Μία ευέλικτη μέθοδος πρόβλεψης χρονοσειρών είναι η μέθοδος SARIMAX, η οποία είναι ιδανική για δεδομένα με εποχικές παρατηρήσεις. Επίσης, έχει τη δυνατότητα να ενσωματώνει εξωγενείς μεταβλητές στη μοντελοποίηση, ώστε να ληφθούν υπόψη εξωτερικοί παράγοντες και τάσεις που μπορούν να επηρεάζουν τα δεδομένα.

Στην προκειμένη περίπτωση, χρησιμοποιείται η γλώσσα προγραμματισμού Python και με τη βοήθεια της συνάρτησης auto_arima() της βιβλιοθήκης pmdarima [26], υπολογίζεται, μετά από δοκιμές, ποιο μοντέλο είναι το βέλτιστο. Έπειτα, χρησιμοποιείται η συνάρτηση SARIMAX() της βιβλιοθήκης statsmodels [27], για να αναπτυχθεί το βέλτιστο μοντέλο και να εκπαιδευτεί.

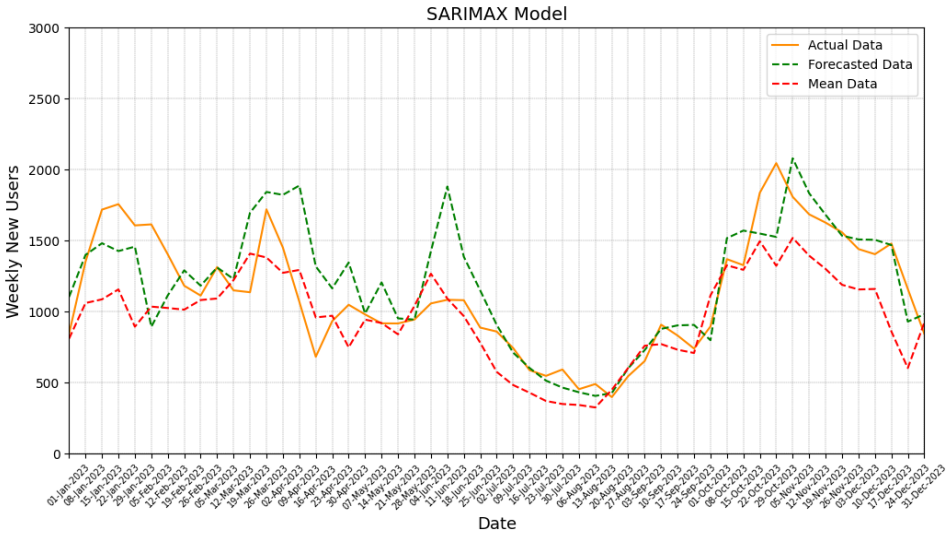
Αναπτύσσεται, με αυτόν τον τρόπο, ένα μοντέλο SARIMAX με δεδομένα ελέγχου, τον μέσο όρο των νέων χρηστών που επισκέφθηκαν το Αποθετήριο την πρώτη εβδομάδα του Ιανουαρίου έως την τελευταία εβδομάδα του Δεκεμβρίου του έτους 2023.

Το διάγραμμα στο οποίο καταγράφονται τα αρχικά δεδομένα, τα χρησιμοποιούμενα στην εκπαίδευση, τα δεδομένα πρόβλεψης και η μέση τιμή του μέσου όρου των εβδομαδιαίων νέων χρηστών για τα έτη 2018 έως 2022 παρουσιάζεται παρακάτω:



Σχήμα 5.2: Πρόβλεψη με το Μοντέλο SARIMAX

Παρακάτω, παρουσιάζεται η μεγέθυνση του παραπάνω διαγράμματος στο χρονικό διάστημα πρόβλεψης:



Σχήμα 5.3: Χρονικό Διάστημα Πρόβλεψης για το Μοντέλο SARIMAX

Το ακριβές μοντέλο που χρησιμοποιείται για το παραπάνω διάγραμμα, έπειτα από τις δοκιμές, είναι το SARIMAX(1, 0, 1)(2, 1, 0, 52). Συγκεκριμένα, οι παράμετροι που χρησιμοποιούνται για την κατασκευή του παραπάνω μοντέλου είναι p = 1 (αυτοπαλινδρομική σειρά), d = 0 (σειρά διαφοροποίησης), q = 1 (σειρά κινούμενου μέσου όρου), P = 2 (εποχιακή αυτοπαλινδρομική σειρά), D = 1 (εποχιακή σειρά διαφοροποίησης), Q = 0 (εποχιακή σειρά κινούμενου μέσου όρου) και s = 52 (διάρκεια του εποχιακού κύκλου). Η διάρκεια του εποχιακού κύκλου είναι s = 52, καθώς τα δεδομένα είναι εβδομαδιαία και παρουσιάζουν ετήσια περιοδικότητα.

Η πρόβλεψη του μέσου όρου των νέων χρηστών σε συνάρτηση με τον χρόνο (πράσινη διακεκομμένη γραμμή) πλησιάζει κάποιες από τις πραγματικές τιμές (πορτοκαλί γραμμή) και αυτό οφείλεται στη χρήση της συνάρτησης auto_arima() η οποία υπολογίζει ένα κατάλληλο μοντέλο SARIMAX για τα δεδομένα νέων χρηστών. Οι αποκλίσεις στα δεδομένα μπορεί να οφείλονται σε δεδομένα που περιέχουν θορύβους και από εξωγενείς παράγοντες.

Είναι άξιο παρατήρησης ότι το μοντέλο αποτυπώνει τη ραγδαία αύξηση του μέσου όρου των τιμών νέων χρηστών τις χειμερινές και φθινοπωρινές εβδομάδες, καθώς και τη μείωση των τιμών κατά τις καλοκαιρινές εβδομάδες του έτους 2023. Στις περισσότερες περιπτώσεις, οι τοπικά μέγιστες τιμές, που προβλέπονται, έχουν πιο υψηλές τιμές από τα πραγματικά δεδομένα, καθώς σε όλα τα προηγούμενα χρόνια υπήρχε μοτίβο αυξημένων τιμών.

Η αξιολόγηση του μοντέλου πραγματοποιείται με τη μέτρηση του μέσου απόλυτου σφάλματος ανάμεσα στα σύνολα ελέγχου και προβλέψεων. Στην προκειμένη περίπτωση, η τιμή του μέσου απόλυτου σφάλματος είναι MAE = 198.516.

Επιπλέον, για τον έλεγχο της αποδοτικότητας του μοντέλου, παρατίθεται η γραφική παράσταση του μέσου όρου των δεδομένων των ετών 2018 έως 2022. Η μορφή της γραφικής παράστασης του μέσου όρου είναι παρόμοιας μορφής με την αντίστοιχη των προβλέψεων.

Η αξιολόγηση του μέσου όρου πραγματοποιείται με τη μέτρηση του μέσου απόλυτου σφάλματος ανάμεσα στα σύνολα ελέγχου και μέσου όρου. Στην προκειμένη περίπτωση, η τιμή του μέσου απόλυτου σφάλματος είναι MAE = 232.035.

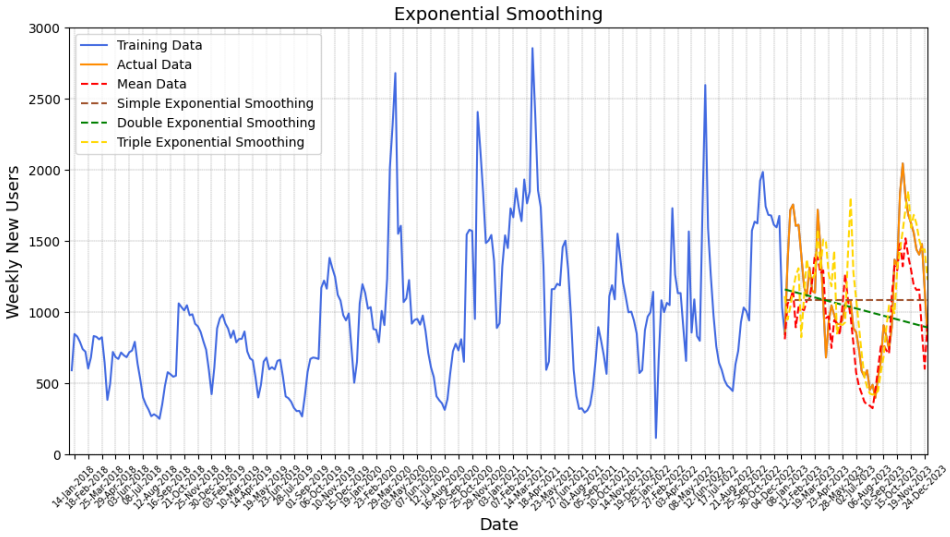
Η τιμή του μέσου απόλυτου σφάλματος, μεταξύ του μοντέλου πρόβλεψης και των πραγματικών δεδομένων, είναι μικρότερη σε σχέση με εκείνο του μέσου όρου και των πραγματικών δεδομένων. Αυτό σημαίνει ότι, το μοντέλο πρόβλεψης είναι πιο αποδοτικό από εκείνο του μέσου όρου των δεδομένων. Οπτικά, το παραπάνω συμπέρασμα μπορεί να επιβεβαιωθεί από το γεγονός ότι το μοντέλο πρόβλεψης προσαρμόζεται περισσότερο στα στοιχεία τάσης σε σχέση με τον μέσο όρο, κυρίως τις εβδομάδες κατά τις οποίες παρουσιάζουν ανοδική πορεία οι τιμές των νέων χρηστών. Όμως, το μοντέλο θα έπρεπε να είναι ακόμη πιο αποδοτικό, ώστε να προβλέπει με περισσότερη ακρίβεια τις τιμές των νέων χρηστών. Ένας λόγος που μπορεί να μη συμβαίνει αυτό, είναι ότι το χρονικό διάστημα και η ποιότητα των δεδομένων δεν είναι τόσο επαρκή για πιο αποδοτική πρόβλεψη. Περισσότερα δεδομένα νέων χρηστών και η εισαγωγή πιο πολύπλοκων παραμέτρων για πρόβλεψη, θα καθιστούσε το μοντέλο καταλληλότερο για έγκυρες προβλέψεις.

5.3 Πρόβλεψη Νέων Χρηστών με τη Μέθοδο Exponential Smoothing

Γενικά, η Εκθετική Εξομάλυνση (Exponential Smoothing) αποτελεί, όπως αναφέρθηκε σε παραπάνω ενότητα, μέθοδο πρόβλεψης χρονοσειρών, η οποία στηρίζεται σε αναθέσεις εκθετικά μειούμενων βαρών για προηγούμενες παρατηρήσεις. Μεγαλύτερα βάρη αποδίδονται σε πιο πρόσφατες παρατηρήσεις, ενώ, για πιο παλιές παρατηρήσεις, εκχωρούνται μειούμενα βάρη. Προϋποθέτει ότι το μέλλον θα είναι παρόμοιο με το πρόσφατο παρελθόν και, επομένως, παρέχει προβλέψεις δεδομένων χρονοσειρών με βάση προηγούμενες υποθέσεις, όπως η εποχικότητα και οι τάσεις.

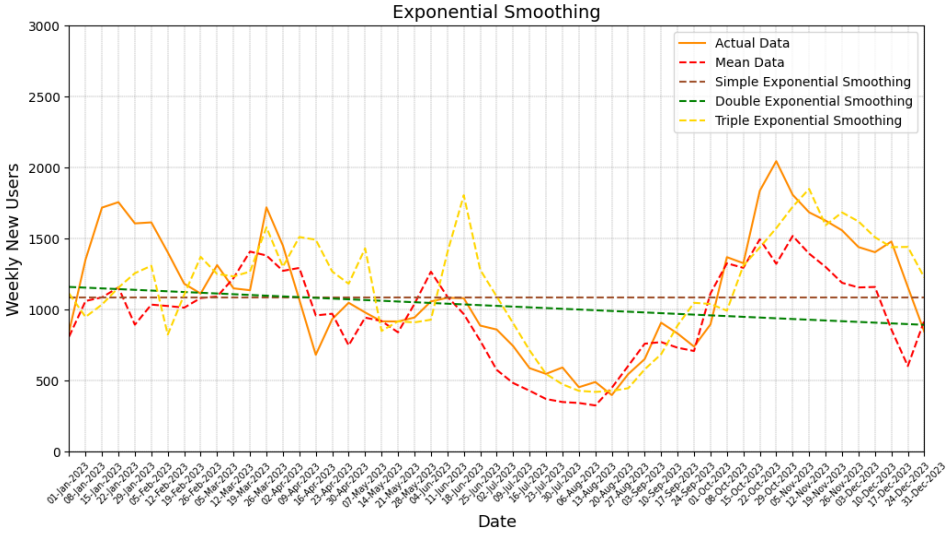
Στην προκειμένη περίπτωση, για την πρόβλεψη, θα χρησιμοποιηθούν οι μέθοδοι Απλής Εκθετικής Εξομάλυνσης (Simple Exponential Smoothing), Διπλής Εκθετικής Εξομάλυνσης (Double Exponential Smoothing) και Τριπλής Εκθετικής Εξομάλυνσης (Triple Exponential Smoothing). Για αυτόν τον σκοπό, χρησιμοποιείται η γλώσσα προγραμματισμού Python και συγκεκριμένα οι συναρτήσεις ExponentialSmoothing() και SimpleExpSmoothing() της βιβλιοθήκης statsmodels.

Παρακάτω, παρουσιάζεται το διάγραμμα εφαρμογής της Απλής, της Διπλής και της Τριπλής Εκθετικής Εξομάλυνσης στον μέσο όρο των νέων χρηστών για την πρώτη εβδομάδα του Ιανουαρίου έως την τελευταία εβδομάδα του Δεκεμβρίου του έτους 2023:



Σχήμα 5.4: Πρόβλεψη με τα Μοντέλα Exponential Smoothing

Παρακάτω, παρουσιάζεται η μεγέθυνση του παραπάνω διαγράμματος στο χρονικό διάστημα πρόβλεψης:



Σχήμα 5.5: Χρονικό Διάστημα Πρόβλεψης για τα Μοντέλα Exponential Smoothing

Στη συνέχεια, παρουσιάζεται ένας πίνακας με τις πιο βέλτιστες σταθερές α, β και γ, οι οποίες προέκυψαν από αναζήτηση πλέγματος (Grid Search) με τη βοήθεια της γλώσσας προγραμματισμού Python και συγκεκριμένα με χρήση της βιβλιοθήκης sklearn [28] που περιέχει τη συνάρτηση ParameterGrid(), καθώς και την τιμή του μέσου απόλυτου σφάλματος για το εκάστοτε μοντέλο:

Παράμετροι Εξομάλυνσης	**Απλή Εκθετική Εξομάλυνση**	**Διπλή Εκθετική Εξομάλυνση**	**Τριπλή Εκθετική Εξομάλυνση**
α (alpha)	0.919	0.798	0.091
β (beta)	-	0.050	0.131
γ (gamma)	-	-	0.515
MAE	339.122	337.186	231.304

Πίνακας 5.1: Τιμές Παραμέτρων Εξομάλυνσης για τα Μοντέλα Exponential Smoothing

Στην προκειμένη περίπτωση, η Απλή Εκθετική Εξομάλυνση εμφανίζει τη μεγαλύτερη τιμή του μέσου απόλυτου σφάλματος (MAE = 339.122) και έπειτα η Διπλή Εκθετική Εξομάλυνση με τιμή μέσου απόλυτου σφάλματος (MAE = 337.186). Τη μικρότερη τιμή του μέσου τετραγωνικού σφάλματος εμφανίζει η μέθοδος της Τριπλής Εκθετικής Εξομάλυνσης (MAE = 231.304), το οποίο σημαίνει ότι η μέθοδος πλησιάζει περισσότερο τις τιμές των πραγματικών δεδομένων σε σχέση με τις άλλες δύο μεθόδους.

Η Απλή Εκθετική Εξομάλυνση λαμβάνει υπόψη τις πιο πρόσφατες παρατηρήσεις, ενώ η Διπλή Εκθετική Εξομάλυνση λαμβάνει υπόψη τα στοιχεία τάσης των δεδομένων. Από την άλλη, η Τριπλή Εκθετική Εξομάλυνση προσαρμόζεται στις τάσεις, την εποχικότητα και τις διακυμάνσεις των δεδομένων. Στην περίπτωση της Απλής Εκθετικής Εξομάλυνσης, η πρόβλεψη είναι σταθερή, αποτυγχάνοντας να αποτυπώσει την εποχικότητα και τα μοτίβα τάσης που υπάρχουν στα πραγματικά στοιχεία. Αντίθετα, στην περίπτωση της Διπλής Εκθετικής Εξομάλυνσης, αποτυπώνεται μία ελαφρά καθοδική τάση ως μία προσπάθεια περιγραφής των μοτίβων τάσεων από τα πραγματικά δεδομένα, ενώ στην περίπτωση της Τριπλής Εκθετικής Εξομάλυνσης καταγράφονται ορισμένα εποχιακά μοτίβα που εντοπίζονται στα πραγματικά δεδομένα. Το μοντέλο της Τριπλής Εκθετικής Εξομάλυνσης παρέχει τις πιο ακριβείς προβλέψεις συγκριτικά με τις υπόλοιπες μεθόδους. Παρ' όλα αυτά,

υπάρχει περιθώριο βελτίωσης των προβλέψεων, ενσωματώνοντας πρόσθετα δεδομένα για την καλύτερη αποτύπωση των πολύπλοκων μοτίβων. Το περιθώριο βελτίωσης αποδεικνύεται από το γεγονός ότι η τιμή του μέσου απόλυτου σφάλματος ανάμεσα στο σύνολο ελέγχου και μέσου όρου (MAE = 232.035) είναι μικρότερο σε τιμή από τις τιμές του μέσου απόλυτου σφάλματος για τις μεθόδους της Απλής και Διπλής Εκθετικής Εξομάλυνσης.

5.4 Ανάλυση και Πρόβλεψη Νέων Χρηστών με τη Μέθοδο LightGBM

Παρακάτω, εφαρμόζεται η μέθοδος Μηχανής Ενίσχυσης Ελαφριάς Κλίσης (Light Gradient Boosting Machine), η οποία αποτελεί μία από τις πιο σύγχρονες και αποδοτικές τεχνικές ενίσχυσης κλίσης για προβλήματα παλινδρόμησης ή ταξινόμησης και αναλύθηκε διεξοδικά σε παραπάνω ενότητα.

Η διαδικασία ανάλυσης πραγματοποιήθηκε με τη βοήθεια της γλώσσας προγραμματισμού Python και συγκεκριμένα με χρήση της βιβλιοθήκης lightgbm [29] που περιέχει τη συνάρτηση LGBMRegressor() με την οποία κατασκευάζεται ένα μοντέλο ενίσχυσης ελαφριάς κλίσης για προβλήματα παλινδρόμησης, έπειτα από τη διαδικασία εκπαίδευσης που θα περιγραφεί σε παρακάτω ενότητα.

Πριν από αυτή την ανάλυση, θα χρειαστεί να γίνει μια πιο λεπτομερής επεξήγηση των στοιχείων που θα τύχουν επεξεργασίας, για να εξεταστούν τα στατιστικά χαρακτηριστικά της χρονοσειράς.

5.4.1 Εποχιακή Αποσύνθεση στο Σύνολο Νέων Χρηστών

Η Εποχιακή Αποσύνθεση (Seasonal Decomposition) αποτελεί μία στατιστική τεχνική που χρησιμοποιείται για τη διάσπαση μίας χρονοσειράς σε συνιστώσες. Η διαδικασία αυτή αποτελεί χρήσιμο εργαλείο για χρονοσειρές στις οποίες υπάρχουν μοτίβα που επαναλαμβάνονται ανά τακτά χρονικά διαστήματα.

Στην προκειμένη περίπτωση, έγινε διάσπαση των αρχικών στοιχείων σε στοιχεία τάσεων, σε εποχιακά στοιχεία και σε υπολειπόμενα στοιχεία, όπως αναλύεται παρακάτω [30]:

Στοιχεία Τάσης (Trend)

Αποτελούν στοιχεία που υποδεικνύουν την κατεύθυνση των δεδομένων με την πάροδο του χρόνου, δηλαδή την τάση των δεδομένων να αυξάνονται, να μειώνονται ή να παραμένουν σχετικά σταθερά.

Εποχιακά Στοιχεία (Seasonality)

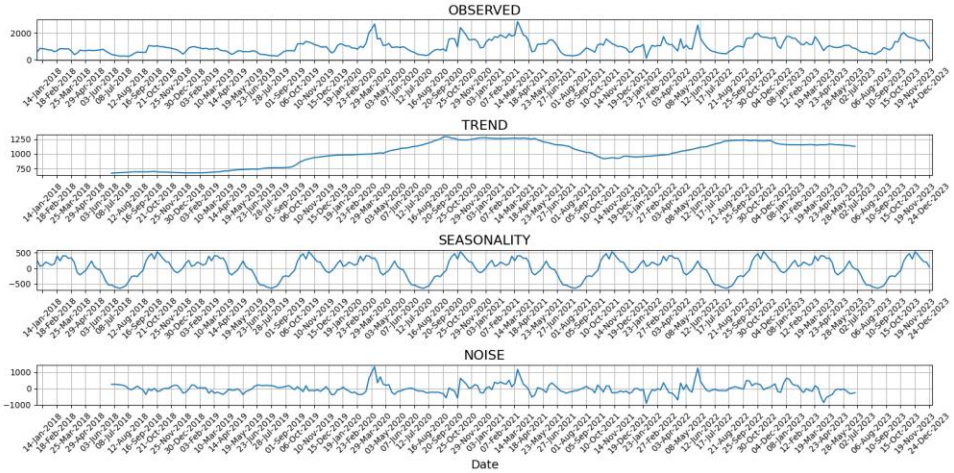
Αποτελούν στοιχεία που υποδεικνύουν τα επαναλαμβανόμενα μοτίβα ανά σταθερά χρονικά διαστήματα, δηλαδή τις τακτικές διακυμάνσεις στα δεδομένα που συμβαίνουν λόγω εποχιακών παραγόντων.

Υπολειπόμενα ή Θορυβώδη Στοιχεία (Residual/Noise)

Αποτελούν τα στοιχεία που υποδεικνύουν τα σφάλματα, δηλαδή τα δεδομένα τα οποία δεν μπορούν να εξηγηθούν από την τάση ή τα εποχιακά μοτίβα. Αντιπροσωπεύει τυχαίες διακυμάνσεις ή τον θόρυβο σε δεδομένα που δεν μπορούν να υπολογιστούν από άλλα στοιχεία. Η εξέταση των υπολειμμάτων μίας χρονοσειράς βοηθά στον εντοπισμό ανωμαλιών και συνεπώς γίνονται κατανοητά απροσδόκητα γεγονότα και ακραία στοιχεία που μπορούν να επηρεάσουν την ανάλυση.

Η διάσπαση πραγματοποιήθηκε με τη βοήθεια της γλώσσας προγραμματισμού Python και συγκεκριμένα της βιβλιοθήκης statsmodels, η οποία περιέχει τη συνάρτηση seasonal_decompose().

Παρακάτω, παρουσιάζονται τα αντίστοιχα διαγράμματα παρατηρούμενων στοιχείων, των στοιχείων τάσεων, των εποχιακών στοιχείων και των υπολειπόμενων στοιχείων του μέσου όρου των νέων χρηστών την πρώτη εβδομάδα του Ιανουαρίου έως την τελευταία εβδομάδα του Δεκεμβρίου του έτους 2023:



Σχήμα 5.6: Εποχιακή Αποσύνθεση του Εβδομαδιαίου Μέσου Όρου Νέων Χρηστών

Από τα παραπάνω διαγράμματα μπορούν να προκύψουν συμπεράσματα σχετικά με τα χαρακτηριστικά και τα μοτίβα που παρουσιάζουν οι νέοι χρήστες συναρτήσει του χρόνου. Το διάγραμμα της τάσης υποδεικνύει τη γενική κατεύθυνση των δεδομένων. Από το διάγραμμα, φαίνεται ότι υπάρχει μία αυξητική τάση μέχρι τους πρώτους μήνες του έτους 2021, με τη μέγιστη τιμή της (1295) να εμφανίζεται την τελευταία εβδομάδα του Σεπτεμβρίου του έτους 2020. Έπειτα, υπάρχει μία καθοδική πορεία της τάσης μέχρι το τέλος του έτους 2021, γεγονός που μπορεί να οφείλεται στην πανδημία Covid-19, εξαιτίας της οποίας δε δόθηκαν έγκαιρα τα έντυπα συγγράμματα της Υπηρεσίας Διαχείρισης Διδακτικών Συγγραμμάτων «Εύδοξος». Στη συνέχεια, οι τιμές της τάσης αυξάνονται, πιθανώς, λόγω της προσθήκης περιεχομένου στο Αποθετήριο μέσω της δεύτερης φάσης του Έργου ΚΑΛΛΙΠΟΣ. Η προσθήκη υλικού στο Αποθετήριο αυξάνει τις πιθανότητες για την προσέλευση νέων χρηστών, καθώς ενθαρρύνεται η αναζήτηση για νέα και ενημερωμένα δεδομένα.

Στο διάγραμμα στοιχείων εποχικότητας κάθε επαναλαμβανόμενο μοτίβο που εμφανίζεται, αντιστοιχεί σε μία εποχική περίοδο. Στην προκειμένη περίπτωση, βάσει της ανάλυσης, η εποχικότητα επαναλαμβάνεται κάθε χρόνο με το ίδιο μοτίβο. Υπάρχουν πολλές αυξομειώσεις τιμών κατά τη διάρκεια του έτους. Η πιο απότομη μείωση τιμών εποχικότητας πραγματοποιείται κατά τις καλοκαιρινές εβδομάδες, καθώς το Αποθετήριο παρουσιάζει τη μεγαλύτερη μείωση σε αριθμό νέων χρηστών. Όπως ήδη σημειώθηκε, περίπου από την πρώτη εβδομάδα του Σεπτεμβρίου κάθε έτους υπάρχει μία ανοδική πορεία των τιμών νέων χρηστών, γεγονός το οποίο σχετίζεται με την έναρξη των ακαδημαϊκών περιόδων. Η αύξηση νέων χρηστών, εκτός ενεργής ακαδημαϊκής περιόδου, πιθανόν, να είναι αποτέλεσμα της προσθήκης ποικιλίας συγγραμμάτων και συνεπώς της αύξησης αναζήτησής τους.

Το διάγραμμα των υπολειπόμενων στοιχείων υποδεικνύει τις τυχαίες διακυμάνσεις στα δεδομένα που δεν μπορούν να αναλυθούν από τα στοιχεία τάσης και εποχικότητας. Στην προκειμένη περίπτωση, το διάγραμμα παρουσιάζει κάποιες αιχμές που αντιστοιχούν στα αρχικά δεδομένα και μπορεί να οφείλονται σε εξωγενείς παράγοντες.

5.4.2 Αυτοσυσχέτιση στο Σύνολο Νέων Χρηστών

Πέρα από τα παραπάνω διαγράμματα, είναι σημαντικό να εξεταστεί και μία άλλη περάμετρος, η αυτοσυσχέτιση, προκειμένου να διασφαλιστεί η ορθότητα και η αξιοπιστία του μοντέλου. Η ανάλυση χρονοσειράς μέσω της συνάρτησης αυτοσυσχέτισης (ACF) υπολογίζεται ως η συνδιακύμανση της χρονοσειράς με καθυστερημένο αντίγραφο του εαυτού της. Έστω ότι $\{X_t\}$ αποτελεί μία σταθερή χρονοσειρά μεγέθους T και $\{X_{t-h}\}$ η χρονοσειρά με καθυστέρηση κατά h περιόδους. Η αυτοσυσχέτιση της $\{X_t\}$ υπολογίζεται σύμφωνα με τον παρακάτω τύπο [31]:

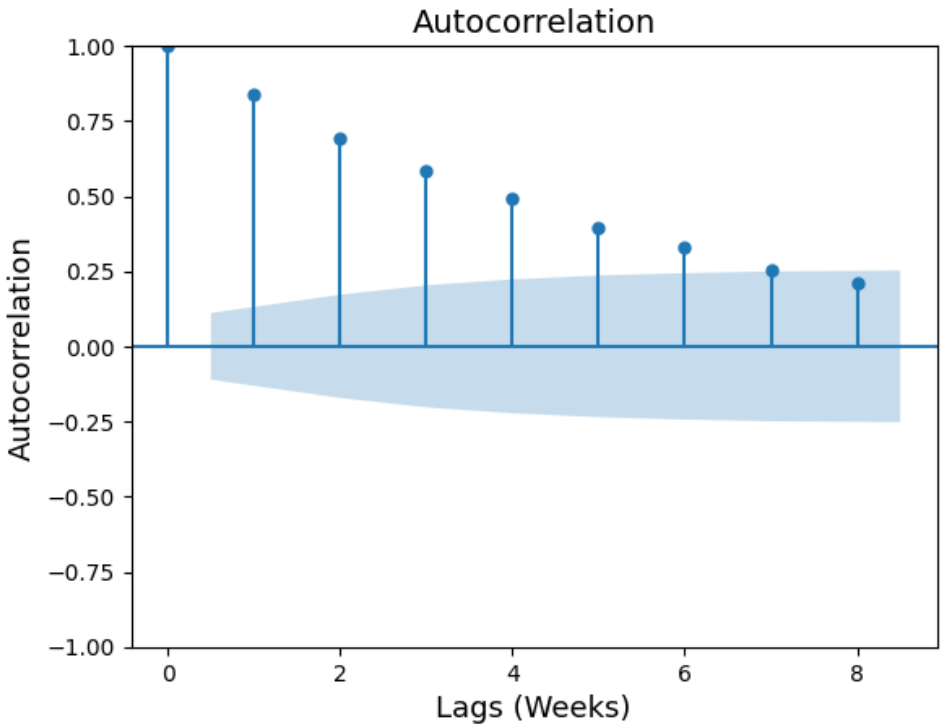
$$ACF_X(h) = \mathrm{Cov}(X_t, X_{t-h}) = E[(X_t - \mu_X)(X_{t-h} - \mu_X)] \tag{5.1}$$

ενώ σε κανονικοποιημένη μορφή δίνεται από τον τύπο:

$$\mathrm{Cor}(X_t, X_{t-h}) = \frac{\mathrm{Cov}(X_t, X_{t-h})}{\mathrm{Cov}(X_t, X_t)} = \frac{E[(X_t - \mu_X)(X_{t-h} - \mu_X)]}{E[(X_t - \mu_X)^2]} \tag{5.2}$$

όπου μ_X η αναμενόμενη τιμή της χρονοσειράς $\{X_t\}$.

Σε αυτή την περίπτωση, χρησιμοποιείται η γλώσσα προγραμματισμού Python και με τη βοήθεια της συνάρτησης plot_acf() της βιβλιοθήκης statsmodels, απεικονίζεται το διάγραμμα αυτοσυσχέτισης. Η εκάστοτε γραμμή στο διάγραμμα υποδεικνύει το επίπεδο αυτοσυσχέτισης για κάθε χρονική καθυστέρηση (lag). Παρακάτω, παρουσιάζεται το διάγραμμα αυτοσυσχέτισης στο σύνολο νέων χρηστών για χρονικά βήματα έως 8 εβδομάδες:



Σχήμα 5.7: Αυτοσυσχέτιση του Εβδομαδιαίου Μέσου Όρου Νέων Χρηστών

Είναι φανερό ότι η αυτοσυσχέτιση μειώνεται σταδιακά, καθώς αυξάνεται η χρονική καθυστέρηση (lag), γεγονός το οποίο οφείλεται στην εξάρτηση μεταξύ των τιμών της χρονοσειράς. Επίσης, φαίνεται ότι τα δεδομένα είναι αυτοπαλινδρομικά, δηλαδή ότι επηρεάζονται από προηγούμενες τιμές και συνεπώς η απόδοση του μοντέλου μπορεί να βελτιωθεί χρησιμοποιώντας καθυστερήσεις.

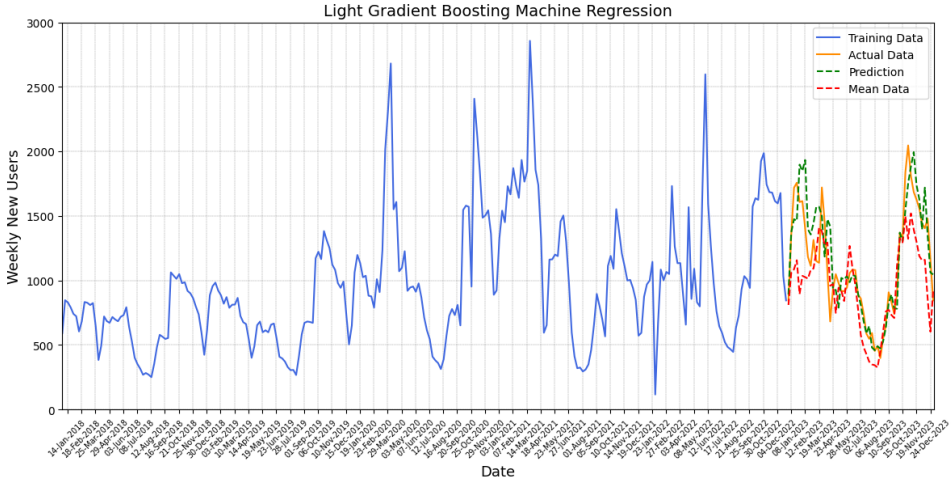
Η πτώση τιμών της παραπάνω γραφικής παράστασης υποδεικνύει ότι δεν υπάρχει ισχυρή μακροπρόθεσμη περιοδικότητα και ότι οι πρόσφατες παρατηρήσεις επηρεάζονται λιγότερο από παρελθοντικές τιμές. Το συμπέρασμα αυτό είναι λογικό, καθώς, όσο πιο χρονικά

απομακρυσμένες είναι οι παρατηρήσεις, τόσο πιο μικρή είναι η πιθανότητα να υπάρχει εξάρτηση μεταξύ τους.

5.4.3 Πρόβλεψη Νέων Χρηστών με τη Μέθοδο LightGBM

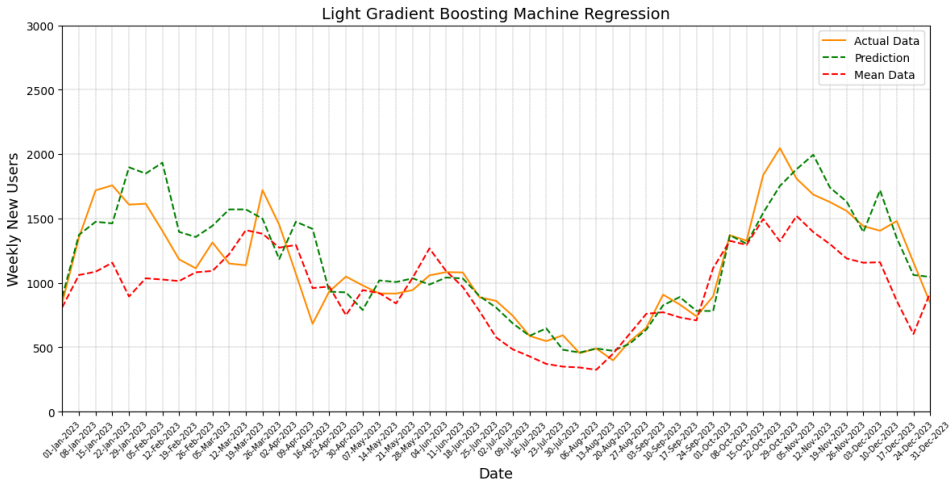
Η πρόβλεψη θα γίνει για τον μέσο όρο των νέων χρηστών την πρώτη εβδομάδα του Ιανουαρίου έως την τελευταία εβδομάδα του Δεκεμβρίου του έτους 2023. Η μεταβλητή (στόχος) στην οποία θα γίνει η πρόβλεψη είναι το σύνολο νέων χρηστών. Οι υπόλοιπες μεταβλητές (χαρακτηριστικά) που θα χρησιμοποιηθούν στην εκπαίδευση του μοντέλου είναι η ημέρα, ο μήνας και τα έτη κατά αύξοντα αριθμό (το έτος 2018 καταγράφεται ως 1, το έτος 2019 καταγράφεται ως 2 κ.λπ.). Επίσης, χρησιμοποιείται η μεταβλητή που έχει τις εβδομάδες κατά αύξοντα αριθμό και η μεταβλητή η οποία διατηρεί τον κανονικοποιημένο μέσο όρο των νέων χρηστών των προηγούμενων 4 εβδομάδων.

Παρακάτω, απεικονίζεται το διάγραμμα προβλέψεων του μέσου όρου των νέων χρηστών για το παραπάνω διάστημα, αφότου γίνει εκπαίδευση του μοντέλου Μηχανής Ενίσχυσης Ελαφριάς Κλίσης:



Σχήμα 5.8: Πρόβλεψη με το Μοντέλο LightGBM Regressor

Παρακάτω, παρουσιάζεται η μεγέθυνση του παραπάνω διαγράμματος στο χρονικό διάστημα πρόβλεψης:



Σχήμα 5.9: Χρονικό Διάστημα Πρόβλεψης για το Μοντέλο LightGBM Regressor

Τα πραγματικά δεδομένα απεικονίζονται με την πορτοκαλί γραμμή, τα αποτελέσματα των προβλέψεων απεικονίζονται με την πράσινη διακεκομμένη γραμμή, ενώ ο μέσος όρος των νέων χρηστών, ανά εβδομάδα, για τα έτη 2018 έως 2022, απεικονίζεται με την κόκκινη διακεκομμένη γραμμή.

Η τιμή του μέσου απόλυτου σφάλματος μεταξύ των πραγματικών δεδομένων και των προβλέψεων είναι MAE = 154.677.

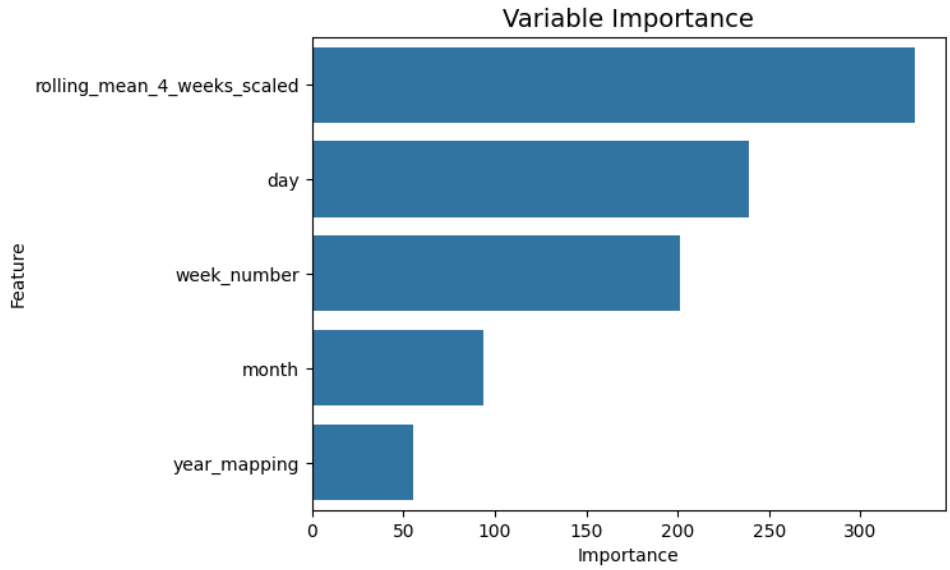
Αντίστοιχα, η τιμή του μέσου απόλυτου σφάλματος μεταξύ των πραγματικών δεδομένων και του μέσου όρου των νέων χρηστών είναι MAE = 232.035.

Οι προβλέψεις και ο μέσος όρος των νέων χρηστών προσεγγίζουν τις πραγματικές τιμές, αν και υπάρχουν περιπτώσεις με αποκλίσεις. Η μέση τιμή δεν εμφανίζει τόσο έντονες κορυφές, δηλαδή αποτελεί μία εξομαλυμένη γραφική παράσταση σε σχέση με εκείνη των προβλέψεων. Οι εποχιακές τάσεις αποτυπώνονται στα δεδομένα των πραγματικών τιμών και των προβλέψεων, όπως η αύξηση των νέων χρηστών κατά τις ανοιξιάτικες και φθινοπωρινές εβδομάδες.

Οι προβλέψεις υπολογίζουν τις εποχιακές τάσεις, αλλά όχι πάντοτε με ακρίβεια. Ένας τρόπος, για να μειωθεί η τιμή του μέσου απόλυτου σφάλματος μεταξύ των πραγματικών

δεδομένων και των προβλέψεων, είναι η παροχή περισσότερων δεδομένων των νέων χρηστών, παλαιότερων ημερομηνιών, στο σύνολο εκπαίδευσης και ελέγχου. Με αυτόν τον τρόπο, βελτιώνονται τα αποτελέσματα των προβλέψεων και τείνουν να πλησιάσουν τις πραγματικές τιμές του συνόλου των νέων χρηστών.

Παρακάτω, παρουσιάζεται ένα διάγραμμα, το οποίο παρουσιάζει τη σημασία των μεταβλητών σύμφωνα με το παραγόμενο μοντέλο:

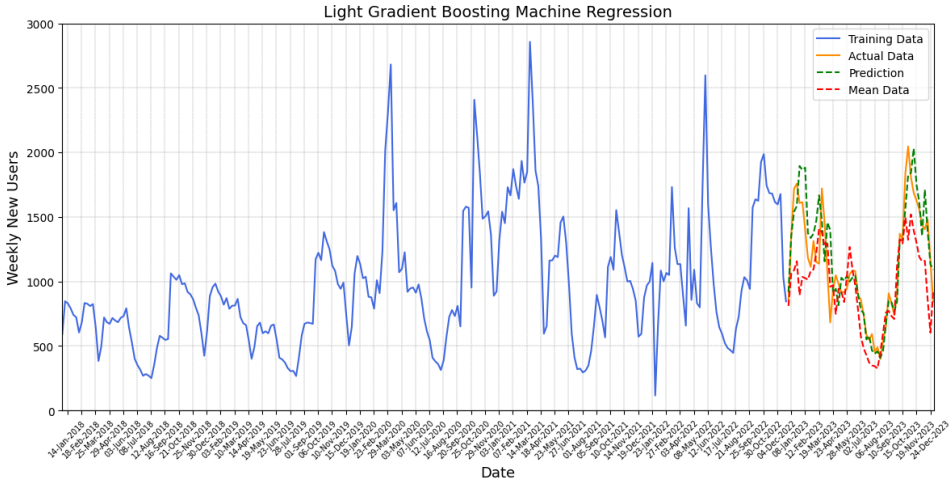


Σχήμα 5.10: Σημασία Μεταβλητών για το Μοντέλο LightGBM Regressor

Από το παραπάνω διάγραμμα, είναι φανερό ότι η μεταβλητή του κανονικοποιημένου μέσου όρου των νέων χρηστών των προηγούμενων 4 εβδομάδων παρέχει σημαντικές πληροφορίες για την πρόβλεψη της τάσης. Οι επόμενες τρεις πιο σημαντικές μεταβλητές είναι η ημέρα, η εβδομάδα κατά αύξουσα τιμή και ο μήνας, γεγονός το οποίο σημαίνει ότι η ημερομηνία επηρεάζει τον αριθμό των νέων χρηστών, πιθανώς, λόγω εποχιακών τάσεων. Έπειτα, η επόμενη σημαντικότερη μεταβλητή, δηλαδή η μεταβλητή που περιλαμβάνει τον αύξοντα αριθμό του έτους, δεν έχει τόση επιρροή στη διαδικασία της πρόβλεψης σε σχέση με τις υπόλοιπες μεταβλητές.

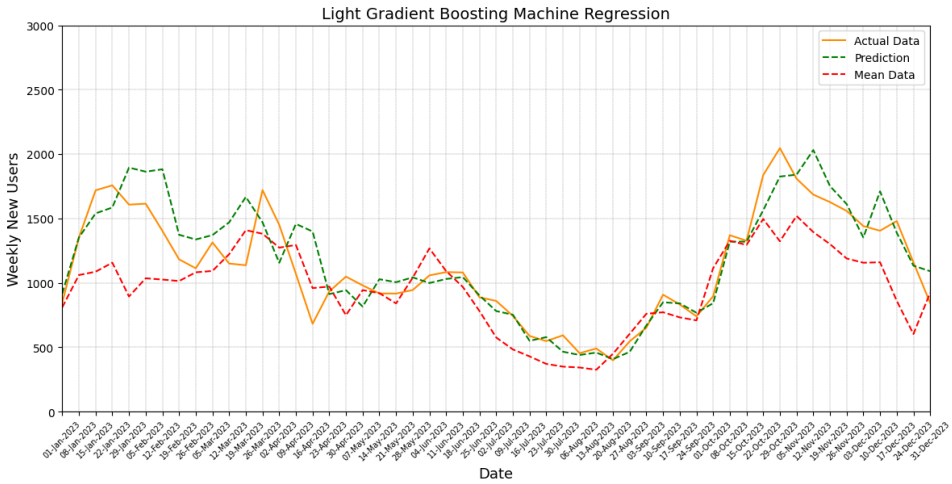
5.4.4 Πρόβλεψη Νέων Χρηστών με τη Μέθοδο LightGBM και Καθυστερήσεις

Όπως αποδείχθηκε προηγουμένως, τα δεδομένα φαίνεται να είναι αυτοσυσχετισμένα, οπότε θα γίνει προσπάθεια βελτίωσης της απόδοσης μοντέλου μέσω της προσθήκης καθυστερήσεων. Για αυτόν τον λόγο δημιουργείται μία νέα μεταβλητή, το χαρακτηριστικό με καθυστέρηση 28 εβδομάδων (count_prev_weeks) πάνω στα δεδομένα των νέων χρηστών. Παρακάτω, απεικονίζεται το διάγραμμα προβλέψεων του μέσου όρου των νέων χρηστών την πρώτη εβδομάδα του Ιανουαρίου έως την τελευταία εβδομάδα του Δεκεμβρίου του έτους 2023, αφότου γίνει εκπαίδευση του μοντέλου Μηχανής Ενίσχυσης Ελαφριάς Κλίσης και προσθήκη καθυστέρησης κατά 28 εβδομάδες:



Σχήμα 5.11: Πρόβλεψη με το Μοντέλο LightGBM Regressor και Καθυστερήσεις

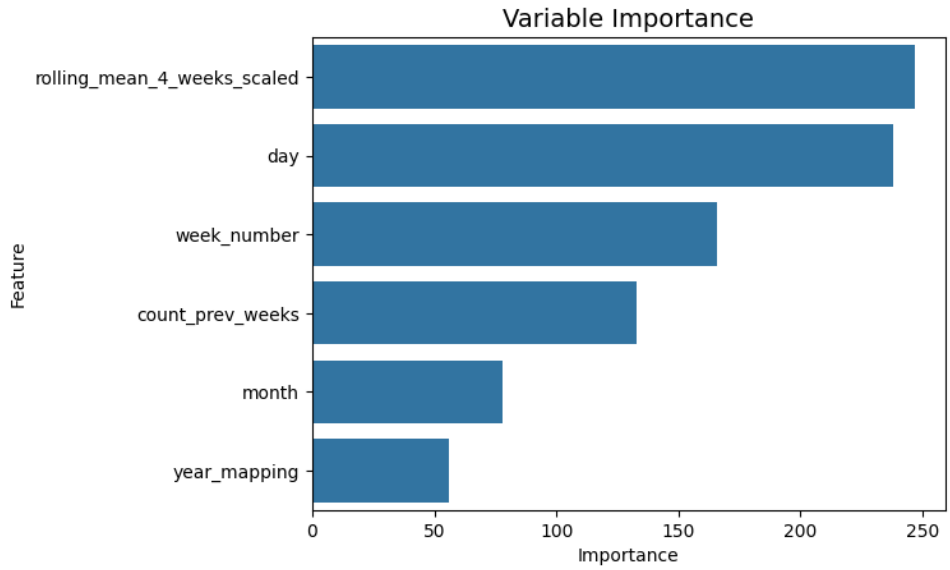
Παρακάτω, παρουσιάζεται η μεγέθυνση του παραπάνω διαγράμματος στο χρονικό διάστημα πρόβλεψης:



Σχήμα 5.12: Χρονικό Διάστημα Πρόβλεψης για το Μοντέλο LightGBM Regressor και Καθυστερήσεις

Από το παραπάνω διάγραμμα, είναι φανερό ότι οι προβλέψεις πλησιάζουν περισσότερο τις πραγματικές τιμές του συνόλου των νέων χρηστών ανά ημερομηνία σε σχέση με εκείνο χωρίς την εισαγωγή καθυστερήσεων. Η τιμή του μέσου απόλυτου σφάλματος είναι ίση με MAE = 143.555, το οποίο είναι μικρότερο από την αντίστοιχη τιμή του μέσου απόλυτου σφάλματος στην πρώτη περίπτωση. Οπότε, το μοντέλο, με την εισαγωγή καθυστερήσεων, είναι πιο αποδοτικό κατά $\frac{154.677-143.555}{154.677} \cdot 100\% \approx 7.190\%$.

Παρακάτω, παρουσιάζεται ένα διάγραμμα το οποίο παρουσιάζει τη σημασία των μεταβλητών σύμφωνα με το παραγόμενο μοντέλο, αφότου προστέθηκε η μεταβλητή με τις καθυστερήσεις:



Σχήμα 5.13: Σημασία Μεταβλητών για το Μοντέλο LightGBM Regressor και Καθυστερήσεις

Από το παραπάνω διάγραμμα, είναι άξιο παρατήρησης το γεγονός ότι το χαρακτηριστικό με καθυστέρηση (count_prev_weeks) αποτελεί σχετικά χρήσιμη μεταβλητή για την πρόβλεψη της μεταβλητής στόχου.

5.4.5 Επιλογή Αποδοτικών Παραμέτρων για το Μοντέλο LightGBM

Ένας τρόπος για τη βελτίωση της απόδοσης του μοντέλου της μηχανικής μάθησης, όπως αναλύθηκε προηγουμένως, είναι η επιλογή κατάλληλων παραμέτρων. Αυτή η διαδικασία περιλαμβάνει τη βελτιστοποίηση των υπερπαραμέτρων του μοντέλου και επηρεάζει άμεσα την ικανότητα του να προσαρμόζεται στα δεδομένα εκπαίδευσης.

Στην προκειμένη περίπτωση, εκτελούνται δοκιμές στις εξής υπερπαραμέτρους του μοντέλου LightGBM Regressor:

n_estimators: Ο αριθμός των ενισχυμένων δέντρων για προσαρμογή.

learning_rate: Η τιμή του ρυθμού εκμάθησης.

num_leaves: Ο αριθμός των μέγιστων φύλλων των βασικών δέντρων.

max_depth: Η τιμή του μέγιστου βάθους των βασικών δέντρων.

Τα χαρακτηριστικά που θα χρησιμοποιηθούν στην εκπαίδευση του εκάστοτε μοντέλου είναι η ημέρα (day), ο μήνας (month) και τα έτη κατά αύξοντα αριθμό (year_mapping). Επίσης, χρησιμοποιείται η μεταβλητή με τον αύξοντα αριθμό των εβδομάδων (week_number) και η μεταβλητή η οποία διατηρεί τον κανονικοποιημένο μέσο όρο των νέων χρηστών των προηγούμενων 4 εβδομάδων (rolling_mean_4_weeks_scaled).

Παρακάτω, παρουσιάζεται ένας πίνακας με τις δοκιμές των παραμέτρων που αναφέρθηκαν προηγουμένως και τις αντίστοιχες τιμές του μέσου απόλυτου σφάλματος:

Algorithm	**LGBMRegressor**				
Model	**n_estimators**	**learning_rate**	**num_leaves**	**max_depth**	**MAE**
Model A	Default (100)	Default (0.1)	Default (31)	Default (-1)	154.677
Model B	200	0.05	50	10	154.738
Model C	300	0.025	60	12	162.244
Model D	350	0.02	62	14	163.698
Model E	370	0.015	64	16	166.550
Train Dataset Attributes	day, month, year_mapping, week_number, rolling_mean_4_weeks_scaled				

Πίνακας 5.2: 1η Δοκιμή Παραμέτρων για το Μοντέλο LightGBM Regressor

Το πιο αποδοτικό μοντέλο που περιλαμβάνει τα χαρακτηριστικά day, month, year_mapping, week_number, rolling_mean_4_weeks_scaled είναι το **Model A**, καθώς η τιμή του μέσου απόλυτου σφάλματος έχει μικρότερη τιμή από τις υπόλοιπες δοκιμές (MAE = 154.677).

Για να γίνει προφανής η επιρροή του χαρακτηριστικού που έχει ως περιεχόμενο τον αύξοντα αριθμό των εβδομάδων (week_number), θα αφαιρεθεί από το σύνολο χαρακτηριστικών. Οπότε, επαναλαμβάνοντας την αρχική διαδικασία, προκύπτει ο παρακάτω πίνακας:

Algorithm	**LGBMRegressor**				
Model	**n_estimators**	**learning_rate**	**num_leaves**	**max_depth**	**MAE**
Model A	Default (100)	Default (0.1)	Default (31)	Default (-1)	175.932
Model B	200	0.05	50	10	173.788
Model C	300	0.025	60	12	175.119
Model D	350	0.02	62	14	176.598
Model E	370	0.015	64	16	174.815
Train Dataset Attributes	day, month, year_mapping, rolling_mean_4_weeks_scaled				

Πίνακας 5.3: 2η Δοκιμή Παραμέτρων για το Μοντέλο LightGBM Regressor

Η αφαίρεση του χαρακτηριστικού week_number αυξάνει τις τιμές του μέσου απόλυτου σφάλματος σε όλα τα μοντέλα. Το πιο αποδοτικό μοντέλο που περιλαμβάνει τα χαρακτηριστικά day, month, year_mapping, rolling_mean_4_weeks_scaled είναι το **Model B**, καθώς η τιμή του μέσου απόλυτου σφάλματος έχει μικρότερη τιμή από τις υπόλοιπες δοκιμές (MAE = 173.788).

Αντίστοιχα, για να γίνει προφανής η επιρροή του χαρακτηριστικού που εκφράζει τον κανονικοποιημένο μέσο όρο των νέων χρηστών των προηγούμενων 4 εβδομάδων (rolling_mean_4_weeks_scaled), θα αφαιρεθεί από το σύνολο χαρακτηριστικών. Οπότε, επαναλαμβάνοντας την αρχική διαδικασία, προκύπτει ο παρακάτω πίνακας:

Algorithm	**LGBMRegressor**				
Model	**n_estimators**	**learning_rate**	**num_leaves**	**max_depth**	**MAE**
Model A	Default (100)	Default (0.1)	Default (31)	Default (-1)	235.310
Model B	200	0.05	50	10	234.675
Model C	300	0.025	60	12	232.961
Model D	350	0.02	62	14	228.749
Model E	370	0.015	64	16	219.759
Train Dataset Attributes	day, month, year_mapping, week_number				

Πίνακας 5.4: 3η Δοκιμή Παραμέτρων για το Μοντέλο LightGBM Regressor

Η αφαίρεση του χαρακτηριστικού rolling_mean_4_weeks_scaled αυξάνει σημαντικά τις τιμές του μέσου απόλυτου σφάλματος σε όλα τα μοντέλα. Το πιο αποδοτικό μοντέλο που περιέχει τα χαρακτηριστικά day, month, year_mapping, week_number είναι το **Model E**, καθώς η τιμή του μέσου απόλυτου σφάλματος έχει μικρότερη τιμή από τις υπόλοιπες δοκιμές (MAE = 219.759).

Αντίστοιχα, για να γίνει προφανής η επιρροή του χαρακτηριστικού που εκφράζει τον κανονικοποιημένο μέσο όρο των νέων χρηστών των προηγούμενων 4 εβδομάδων (rolling_mean_4_weeks_scaled) και του χαρακτηριστικού που έχει ως περιεχόμενο τον αύξοντα αριθμό των εβδομάδων (week_number), θα αφαιρεθούν από το σύνολο χαρακτηριστικών. Οπότε, επαναλαμβάνοντας την αρχική διαδικασία, προκύπτει ο παρακάτω πίνακας:

Algorithm	**LGBMRegressor**				
Model	**n_estimators**	**learning_rate**	**num_leaves**	**max_depth**	**MAE**
Model A	Default (100)	Default (0.1)	Default (31)	Default (-1)	238.186
Model B	200	0.05	50	10	232.779
Model C	300	0.025	60	12	222.897
Model D	350	0.02	62	14	224.124
Model E	370	0.015	64	16	227.515
Train Dataset Attributes	day, month, year_mapping				

Πίνακας 5.5: 4η Δοκιμή Παραμέτρων για το Μοντέλο LightGBM Regressor

Η αφαίρεση του χαρακτηριστικού rolling_mean_4_weeks_scaled και week_number αυξάνει σημαντικά τις τιμές του μέσου απόλυτου σφάλματος σε όλα τα μοντέλα. Το πιο αποδοτικό μοντέλο που περιλαμβάνει τα χαρακτηριστικά day, month, year_mapping είναι το **Model C**, καθώς η τιμή του μέσου απόλυτου σφάλματος έχει μικρότερη τιμή από τις υπόλοιπες δοκιμές (MAE = 222.897).

Από τα παραπάνω, το χαρακτηριστικό rolling_mean_4_weeks_scaled αποδεικνύεται σημαντικό, καθώς συμβάλει στη μείωση του μέσου απόλυτου σφάλματος MAE. Επίσης, το χαρακτηριστικό week_number συμβάλλει στη βελτίωση της απόδοσης του μοντέλου. Η αφαίρεση και των δύο χαρακτηριστικών μειώνει την αποδοτικότητα του μοντέλου και συνεπώς είναι σημαντικό να συμπεριληφθούν στα χαρακτηριστικά της εκπαίδευσης.

Ένα μοντέλο είναι αποδοτικό, όταν οι τιμές του πλησιάζουν, όσο περισσότερο γίνεται, τις πραγματικές τιμές των δεδομένων. Με άλλα λόγια, ένα μοντέλο είναι πιο αποδοτικό από κάποιο άλλο, όταν η τιμή του μέσου απόλυτου σφάλματος έχει τη μικρότερη τιμή. Στην προκειμένη περίπτωση, το πιο αποδοτικό μοντέλο είναι το LGBMRegressor(n_estimators=100, learning_rate=0.1, num_leaves=31, max_depth=-1) διατηρώντας τα χαρακτηριστικά day, month, year_mapping, week_number, rolling_mean_4_weeks_scaled, αφού η τιμή του μέσου απόλυτου σφάλματος έχει μικρότερη τιμή από τις υπόλοιπες δοκιμές (MAE = 154.677).

5.5 Σύγκριση και Αξιολόγηση Μεθόδων Πρόβλεψης των Νέων Χρηστών

Παρακάτω, συνοψίζονται οι μέθοδοι πρόβλεψης, που αναλύθηκαν στα παραπάνω κεφάλαια, με τις αντίστοιχες τιμές του μέσου απόλυτου σφάλματος και της τετραγωνικής ρίζας του μέσου τετραγωνικού σφάλματος:

Μέθοδοι Πρόβλεψης	**MAE**	**RMSE**
SARIMAX (1, 0, 1)(2, 1, 0, 52)	198.516	284.277
Απλή Εκθετική Εξομάλυνση	339.122	410.568
Διπλή Εκθετική Εξομάλυνση	337.186	426.693
Τριπλή Εκθετική Εξομάλυνση	231.304	303.793
Μηχανή Ενίσχυσης Ελαφριάς Κλίσης	154.677	216.642

Πίνακας 5.6: Σύγκριση Μεθόδων Πρόβλεψης με Μετρικές Αξιολόγησης

Η μέθοδος Μηχανής Ενίσχυσης Ελαφριάς Κλίσης παρουσιάζει τη μικρότερη τιμή του μέσου απόλυτου σφάλματος MAE (154.677) και τιμή τετραγωνικής ρίζας του μέσου τετραγωνικού σφάλματος RMSE (216.642), γεγονός που την καθιστά την πιο αποδοτική μέθοδο πρόβλεψης σε σχέση με εκείνες που εξετάστηκαν. Το παραπάνω συμπέρασμα είναι λογικό, καθώς η μέθοδος Μηχανής Ενίσχυσης Ελαφριάς Κλίσης είναι μία σύγχρονη

μέθοδος που μπορεί να προσαρμοστεί σε πολύπλοκα μοτίβα των δεδομένων που προβλέπονται.

Η μέθοδος SARIMAX (1, 0, 1)(2, 1, 0, 52) ακολουθεί με τιμή του μέσου απόλυτου σφάλματος MAE = 198.516 και τιμή τετραγωνικής ρίζας του μέσου τετραγωνικού σφάλματος RMSE = 284.277, δηλαδή είναι λιγότερο ακριβής από τη μέθοδο Μηχανής Ενίσχυσης Ελαφριάς Κλίσης. Είναι κατάλληλη για χρονοσειρές με εποχικότητα και υψηλή αυτοσυσχέτιση, αλλά δεν αποδίδει καλά σε πολύπλοκα μοτίβα σε σχέση με τη μέθοδο Μηχανής Ενίσχυσης Ελαφριάς Κλίσης, η οποία στηρίζεται στην κατασκευή δέντρων απόφασης.

Οι μέθοδοι εκθετικής εξομάλυνσης εμφανίζουν σημαντικά υψηλότερες τιμές μέσου απόλυτου σφάλματος MAE και τιμές τετραγωνικής ρίζας του μέσου τετραγωνικού σφάλματος RMSE σε σχέση με τις υπόλοιπες μεθόδους. Συγκεκριμένα, η Απλή Εκθετική Εξομάλυνση εμφανίζει τιμή μέσου απόλυτου σφάλματος MAE = 339.122 και τιμή τετραγωνικής ρίζας του μέσου τετραγωνικού σφάλματος RMSE = 410.568, ενώ η Διπλή Εκθετική Εξομάλυνση εμφανίζει τιμή μέσου απόλυτου σφάλματος MAE = 337.186 και τιμή τετραγωνικής ρίζας του μέσου τετραγωνικού σφάλματος RMSE = 426.693.

Από την άλλη μεριά, η Τριπλή Εκθετική Εξομάλυνση εμφανίζει τιμή μέσου απόλυτου σφάλματος MAE = 231.304 και τιμή τετραγωνικής ρίζας του μέσου τετραγωνικού σφάλματος RMSE = 303.793. Η Απλή Εκθετική Εξομάλυνση είναι κατάλληλη για δεδομένα χωρίς τάσεις, ενώ η Διπλή Εκθετική Εξομάλυνση λαμβάνει υπόψιν τα μοτίβα τάσης. Η Τριπλή Εκθετική Εξομάλυνση λαμβάνει υπόψιν τα δεδομένα τάσης και εποχικότητας. Στην προκειμένη περίπτωση, τα αποτελέσματα της εκθετικής εξομάλυνσης δεν είναι τόσο αντιπροσωπευτικά, λόγω των μεγάλων τιμών του μέσου απόλυτου σφάλματος MAE και της τετραγωνικής ρίζας του μέσου τετραγωνικού σφάλματος RMSE.

6 Συμπεράσματα και Μελλοντικές Επεκτάσεις

6.1 Συμπεράσματα

Η παρούσα Διπλωματική Εργασία αποτελεί την αρχή για την πρόβλεψη της δημοτικότητας των τεκμηρίων και τη βελτίωση του περιεχομένου που παρέχει το αποθετήριο ΚΑΛΛΙΠΟΣ με βάση τις προτιμήσεις των χρηστών που το επισκέπτονται. Η συγκεκριμένη έρευνα είναι ενδιαφέρουσα και σημαντική, καθώς ελάχιστες εργασίες έχουν εστιάσει στη βελτίωση των υποδομών ενός αποθετηρίου.

Συνοψίζοντας, τα αποτελέσματα που προέκυψαν μέσω της ανάλυσης των δεδομένων χρήσης επιτρέπουν την κατανόηση των μοτίβων και των τάσεων συμπεριφοράς χρηστών, παρέχοντας πολύτιμες πληροφορίες που αξιοποιούνται για την εξέλιξη του Αποθετηρίου. Συγκεκριμένα, το πιο βασικό μοτίβο συμπεριφοράς χρηστών είναι η αύξηση προσέλευσης και αφοσίωσης χρηστών, η οποία πραγματοποιείται, κυρίως, όταν ξεκινούν τα ακαδημαϊκά εξάμηνα και η αντίστοιχη μείωση στις περιόδους διακοπών. Επίσης, παρατηρείται ότι ο λιγότερο διαδεδομένος τρόπος προσέλευσης χρηστών είναι εκείνος από τα μέσα κοινωνικής δικτύωσης και γι' αυτό χρειάζεται να διαδοθεί περισσότερο το έργο του αποθετηρίου ΚΑΛΛΙΠΟΣ στα κοινωνικά δίκτυα, μέσω διαφημίσεων και κοινοποιήσεων μεταξύ χρηστών. Άλλη μία ανάγκη που προέκυψε, μελετώντας τον πληθυσμιακό καταμερισμό, είναι εκείνη για την πολυγλωσσικότητα περιεχομένου του Αποθετηρίου, καθώς οι περισσότεροι ενεργοί επισκέπτες βρίσκονται στην Ελλάδα. Η πολυγλωσσικότητα περιεχομένου, η οποία πραγματοποιείται με τη μετάφραση τεκμηρίων σε πολλές γλώσσες, μπορεί να διευρύνει την προσβασιμότητα σε χρήστες από διαφορετικές χώρες και συνεπώς να αυξήσει την απήχηση του Αποθετηρίου. Επίσης, είναι φανερό ότι προτιμώνται οι αποθηκεύσεις συγγραμμάτων από τη θεματική κατηγορία «Δίκαιο και Κοινωνικές Επιστήμες» και έπειτα από τις «Ανθρωπιστικές Επιστήμες και Τέχνες».

Η μελέτη εμπλουτισμού του περιεχομένου του Αποθετηρίου μπορεί να στηριχθεί στις παραπάνω παρατηρήσεις και να προστεθεί περιεχόμενο που επιλέγεται συχνότερα, αφότου αναλυθεί η ζήτηση προβολών ή αποθήκευσης περιεχομένου σε κάθε θεματική ενότητα για επαρκή χρονικά διαστήματα. Το χρονικό διάστημα στο οποίο πραγματοποιήθηκε η

ανάλυση στοιχείων χρήσης του Αποθετηρίου είναι από 1η Ιουλίου του έτους 2023 έως την 1η Ιουλίου του έτους 2024, καθώς το Google Analytics άλλαξε σε νεότερη έκδοση από την 1η Ιουλίου του έτους 2023. Οι διαφορετικές εκδόσεις του Google Analytics δεν περιέχουν τις ίδιες μετρικές και δυνατότητες, γι' αυτό ήταν αναγκαία η χρήση δεδομένων από την πιο προηγμένη έκδοση.

Η μέθοδος Light Gradient Boosting Machine (LightGBM), που εφαρμόστηκε για την πρόβλεψη του μέσου όρου των εβδομαδιαίων νέων χρηστών για το έτος 2023, αποδείχθηκε η πιο αποδοτική μέθοδος. Επιπλέον, η ενσωμάτωση καθυστερήσεων στο μοντέλο συνέβαλε στη βελτίωση των παραγόμενων προβλέψεων. Αντίθετα, η μέθοδος SARIMAX αποδεικνύεται λιγότερο αποδοτική σε σύγκριση με τη LightGBM.

Οι λιγότερο ακριβείς προβλέψεις, σε σχέση με τα υπόλοιπα μοντέλα, παρουσιάστηκαν με τη μέθοδο Exponential Smoothing. Συγκεκριμένα, η μέθοδος Triple Exponential Smoothing ήταν η πιο αποδοτική από τις υπόλοιπες μεθόδους Exponential Smoothing, καθώς αποτυπώθηκαν κάποια σημεία τάσης και εποχικότητας στις προβλέψεις. Ακολουθεί η μέθοδος Double Exponential Smoothing, με την οποία έγινε προσπάθεια να αποτυπωθούν τα σημεία τάσης. Τέλος, με τη μέθοδο Simple Exponential Smoothing λήφθηκαν υπόψη μόνο τα πιο πρόσφατα στοιχεία για την πρόβλεψη.

Από τις τιμές των μετρικών αξιολόγησης των μοντέλων, παρατηρείται το περιθώριο για τη βελτίωση των μοντέλων, καθώς τα στοιχεία που χρησιμοποιήθηκαν δεν διαθέτουν την απαραίτητη πολυπλοκότητα για την επίτευξη πιο αξιόπιστων προβλέψεων. Η αρχική προσέγγιση ήταν η πρόβλεψη των νέων χρηστών, ημερησίως, για το έτος 2023, αλλά η συνάρτηση auto_arima() της βιβλιοθήκης pmdarima δεν υπολογίζει το πιο αποδοτικό μοντέλο SARIMAX για ένα μεγάλο σύνολο δεδομένων με ετήσια εποχικότητα. Παρ' όλα αυτά, με την πρόβλεψη του μέσου όρου των εβδομαδιαίων νέων χρηστών προκύπτουν ικανοποιητικά αποτελέσματα.

6.2 Μελλοντικές Επεκτάσεις

Όπως αναφέρθηκε προηγουμένως, η παρούσα εργασία αποτελεί το πρώτο βήμα στην ανάλυση και πρόβλεψη στοιχείων χρήσης του αποθετηρίου ΚΑΛΛΙΠΟΣ. Παρακάτω, θα παρουσιαστούν κάποιες ενδιαφέρουσες επεκτάσεις, καθώς και μελλοντικά πλάνα βελτίωσης της παρούσας εργασίας.

Η συνεχής παρακολούθηση των στοιχείων χρήσης θα βοηθήσει στη διαρκή προσαρμογή του περιεχομένου και συνεπώς στην ελκυστικότητα του Αποθετηρίου. Επίσης, η μελέτη μεγαλύτερων διαστημάτων, θα επιτρέψει τη συλλογή περισσότερων δεδομένων και συνεπώς την καλύτερη παρακολούθηση των τάσεων και μοτίβων συμπεριφορών των χρηστών. Ταυτόχρονα, η διαρκής συλλογή δεδομένων σχετικά με δημογραφικά στοιχεία, όπως η ηλικία, το φύλο και το επίπεδο σπουδών, μπορεί να βοηθήσει στην εύρεση εξιδεικευμένων αναγκών των χρηστών. Ο συνδυασμός δημογραφικών στοιχείων και στοιχείων χρήσης του Αποθετηρίου μπορεί να βοηθήσει στην ανάπτυξη μοντέλων τα οποία αναλύουν τις αλληλεπιδράσεις των χρηστών με διαφορετικές ενότητες.

Παράλληλα, η μεθοδολογία που αναπτύχθηκε για την πρόβλεψη των νέων χρηστών μπορεί να βελτιωθεί και να διευρυνθεί για τις θεματικές ενότητες που περιέχει το Αποθετήριο. Με αυτόν τον τρόπο, θα γίνει γνωστή η προτίμηση των χρηστών για τις θεματικές ενότητες που ενδιαφέρουν περισσότερο τους χρήστες. Επιπλέον, η ανάπτυξη συστημάτων που προτείνουν δημοφιλή τεκμήρια του Αποθετηρίου σε νέους χρήστες, βάσει των παρατηρούμενων τάσεων, μπορεί να αυξήσει την προσέλευση νέων χρηστών στο Αποθετήριο, χάρη στην εξατομικευμένη εμπειρία χρήστη.

Βιβλιογραφία

[1] Κ. Διαμαντής και Κ. Μπίκος, «Ανοικτή πρόσβαση στη γνώση – Ψηφιακά αποθετήρια,» σε *Σενάρια Διδασκαλίας με την υποστήριξη Ψηφιακών Μέσων*, Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις, 2022, p. 159. https://repository.kallipos.gr/handle/11419/8768.

[2] R. Heery και S. Anderson, «Digital repositories review,» Joint Information Systems Committee, 2005.

[3] «Εθνικό Κέντρο Τεκμηρίωσης,» [Ηλεκτρονικό]. Available: https://www.ekt.gr/el/faq/15913.

[4] «Αποθετήριο Κάλλιπος: Αρχική,» [Ηλεκτρονικό]. Available: https://repository.kallipos.gr/.

[5] «Artemis: Home,» [Ηλεκτρονικό]. Available: http://artemis.cslab.ece.ntua.gr:8080/jspui/.

[6] «DSpace Home,» [Ηλεκτρονικό]. Available: https://dspace.lib.ntua.gr/xmlui/.

[7] «Ανοικτή Βιβλιοθήκη — Ελεύθερα ψηφιακά βιβλία,» [Ηλεκτρονικό]. Available: https://www.openbook.gr/.

[8] «Αποθετήριο ΕΛ/ΛΑΚ: Αρχική,» [Ηλεκτρονικό]. Available: https://repository.ellak.gr/ellak/.

[9] Google, «Google Help,» [Ηλεκτρονικό]. Available: https://support.google.com/.

[10] M. Beasley, Practical Web Analytics for User Experience, 1η Έκδοση επιμ., Elsevier, 2013, pp. 29-30.

[11] E. Paradis, B. O. Brien, L. Nimmon, G. Bandiera και M. A. Martimianakis, «Design: Selection of Data Collection Methods,» *Journal of Graduate Medical Education,* τόμ. 8, αρ. 2, pp. 263-264, 2016.

[12] A. Tatar, M. D. de Amorim, S. Fdida και P. Antoniadis, «A survey on predicting the popularity of web content,» *Journal of Internet Services and Applications,* τόμ. 5, αρ. 1, p. 13, 13 Αύγουστος 2014.

[13] «Tableau,» 2023. [Ηλεκτρονικό]. Available: https://www.tableau.com/learn/articles.

[14] C. Ifeanyichukwu Ugoh, C. Alice Uzuke και D. Obioma Ugoh, «Application of ARIMAX Model on Forecasting Nigeria's GDP,» *American Journal of Theoretical and Applied Statistics,* τόμ. 10, αρ. 5, pp. 217-218, 29 Οκτώβριος 2021.

[15] M. D. Isa, «Understanding Time Series Forecasting with ARIMA,» 6 Ιούλιος 2023. [Ηλεκτρονικό]. Available: https://medium.com/.

[16] T. Andrianajaina, D. T. Razafimahefa, R. Rakotoarijaina και C. G. Haba, «Grid Search for SARIMAX Parameters for Photovoltaic Time Series Modeling,» *Global Journal of Energy Technology Research,* τόμ. 9, pp. 88-89, 23 Δεκέμβριος 2022.

[17] C. Nontapa, C. Kesamoon, N. Kaewhawong και P. Intrapaiboon, «A New Hybrid Forecasting Using Decomposition Method with SARIMAX Model and Artificial Neural Network,» *International Journal of Mathematics and Computer Science,* τόμ. 16, αρ. 4, pp. 1343-1344, 16 Μάρτιος 2021.

[18] N. Malkari, «Exponential Smoothing: A Method for Time Series Forecasting,» 19 Απρίλιος 2023. [Ηλεκτρονικό]. Available: https://medium.com/.

[19] E. Spiliotis, «Decision Trees for Time-Series Forecasting,» *ResearchGate,* pp. 31-33, Ιανουάριος 2022.

[20] İ. Kılıç, «Gradient Boosting Machines (GBM) with Python Example,» 23 Σεπτέμβριος 2023. [Ηλεκτρονικό]. Available: https://medium.com/.

[21] E. Saman, «Light Gradient Boosting Machine,» 25 Μάρτιος 2023. [Ηλεκτρονικό]. Available: https://medium.com/.

[22] «Metrics Evaluation: MSE, RMSE, MAE and MAPE,» 26 Φεβρουάριος 2024. [Ηλεκτρονικό]. Available: https://medium.com/.

[23] «About OERSI,» [Ηλεκτρονικό].
Available: https://oersi.org/resources/pages/en/about/.

[24] Google, «Looker Studio,» [Ηλεκτρονικό].
Available: https://lookerstudio.google.com.

[25] «Matplotlib - Visualization with Python,» [Ηλεκτρονικό].
Available: https://matplotlib.org/.

[26] «pmdarima,» [Ηλεκτρονικό]. Available: https://pypi.org/project/pmdarima/.

[27] «statsmodels,» [Ηλεκτρονικό]. Available: https://pypi.org/project/statsmodels/.

[28] «sklearn,» [Ηλεκτρονικό]. Available: https://scikit-learn.org/stable/.

[29] «LightGBM,» [Ηλεκτρονικό]. Available: https://lightgbm.readthedocs.io/en/stable/.

[30] N. Malkari, «Seasonal Decomposition,» 16 Απρίλιος 2023. [Ηλεκτρονικό].
Available: https://medium.com/.

[31] «Autocorrelation Function,» [Ηλεκτρονικό].
Available: https://www.sciencedirect.com.

Συντομογραφίες-Αρκτικόλεξα-Ακρωνύμια

ACF	Autocorrelation Function
API	Application Programming Interface
AR	Autoregressive
I	Integrated
ISBN	International Standard Book Number
(Light)GBM	(Light) Gradient Boosting Machine
MAPE	Mean Absolute Percentage Error
MA	Moving Average
MAE	Mean Absolute Error
OER(SI)	Open Educational Resources (Search Index)
(R)MSE	(Root) Mean Squared Error
(S)ARIMA(X)	(Seasonal) Autoregressive Integrated Moving Average (with Exogenous Variables Model)
URL	Uniform Resource Locator
ΔΕΠ	Διδακτικό Ερευνητικό Προσωπικό
ΕΕΛΛΑΚ	Οργανισμός Ανοικτών Τεχνολογιών
ΕΚΤ	Εθνικό Κέντρο Τεκμηρίωσης
ΕΛ/ΛΑΚ	Ελεύθερο Λογισμικό/Λογισμικό Ανοικτού Κώδικα
κ.ά.	και άλλα
κ.λπ.	και λοιπά
π.χ.	παραδείγματος χάρη
ΠΑ	Ποσοστό Αφοσίωσης

Απόδοση Ξενόγλωσσων Όρων

Απόδοση	Ξενόγλωσσος Όρος
Αναζήτηση Πλέγματος	Grid Search
Ανάλυση Χρονοσειρών	Time Series Analysis
Απευθείας	Direct
(Απλή/Διπλή/Τριπλή) Εκθετική Εξομάλυνση	(Simple/Double/Triple) Exponential Smoothing
Απόκτηση Χρηστών	User Acquisition
Αριθμός (Μοναδικών) Προβολών Σελίδας	(Unique) Pageviews
Αριθμός Περιόδων ανά Χρήστη	Number of Sessions Per User
Αυτοπαλινδρομικό Μοντέλο	Autoregressive Model
Αφοσίωση Χρήστη	User Engagement
Δέντρα Αποφάσεων	Decision Trees
Διεπαφή Προγραμματισμού Εφαρμογών	Application Programming Interface
Εκδηλώσεις	Events
Ενεργοί Χρήστες	Active Users
Εξυπηρετητές	Servers
Επιστροφή Χρήστη	Returning User
Εποχιακή Αποσύνθεση	Seasonal Decomposition
(Εποχιακός) Αυτοπαλινδρομικός Ολοκληρωμένος Κινούμενος Μέσος Όρος (με Εξωγενείς Παράγοντες)	(Seasonal) Autoregressive Integrated Moving Average (with Exogenous Variables Model)
Εποχικότητα	Seasonality
Ευρετήριο Αναζήτησης Ανοικτών Εκπαιδευτικών Πόρων	Open Educational Resources Search Index
Ηλεκτρονική Διεύθυνση	Email
Θόρυβος	Noise

Λογαριασμός	Account
Μέσο Απόλυτο Σφάλμα	Mean Absolute Error
Μέσο Ποσοστιαίο Απόλυτο Σφάλμα	Mean Absolute Percentage Error
Μέσο Τετραγωνικό Σφάλμα	Mean Squared Error
Μέσος Χρόνος Αφοσίωσης	Average Engagement Time
Μηχανή Ενίσχυσης (Ελαφριάς) Κλίσης	(Light) Gradient Boosting Machine
Μηχανική Μάθηση	Machine Learning
Μοντέλο Κινούμενου Μέσου Όρου	Moving Average Model
Νέοι Χρήστες	New Users
Οργανική Αναζήτηση	Organic Search
Οργανική Κοινωνική	Organic Social
Οργανικό Βίντεο	Organic Video
Παραπομπή	Referral
Περίοδοι Σύνδεσης	Sessions
Ποσοστό Αφοσίωσης	Engagement Rate
Ποσοστό Εγκατάλειψης	Bounce Rate
Πρόβλεψη Χρονοσειρών	Time Series Forecasting
Προφίλ	Profile
Σελίδες ανά Περίοδο Σύνδεσης	Pages Per Session
Σελίδες και Οθόνες	Pages and Screens
Σύνοψη Αφοσίωσης	Engagement Overview
Τάση	Trend
Τετραγωνική Ρίζα Μέσου Τετραγωνικού Σφάλματος	Root Mean Squared Error
Υπολειπόμενο	Residual
Χαρακτηριστικά	Features
Χρήστες	Users

Χρονική Καθυστέρηση	Lag
Χωρίς Ανάθεση	Unassigned