ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«vaseis-app2: Επέκταση της εφαρμογής vaseis-app με μηχανισμούς αυτόματης εισαγωγής δεδομένων και διευρυμένες λειτουργίες ανάλυσης και παρουσίασης δεδομένων»

Του φοιτητή
Λαπούσι Γκενάρντο
Αρ. Μητρώου: 2020076

Επιβλέπων
Ουγιάρογλου Στέφανος
Επίκουρος Καθηγητής

25 Μαίου 2026

Τίτλος Δ.Ε. . vaseis-app2: Επέκταση της εφαρμογής vaseis-app με μηχανισμούς αυτόματης εισαγωγής δεδομένων και διευρυμένες λειτουργίες ανάλυσης και παρουσίασης δεδομένων
Κωδικός Δ.Ε. 26107
Ονοματεπώνυμο φοιτητή: Λαπούσι Γκενάρντο
Ονοματεπώνυμο εισηγητή: Ουγιάρογλου Στέφανος
Ημερομηνία ανάληψης Δ.Ε. 17-01-2026
Ημερομηνία περάτωσης Δ.Ε. 31-05-2026

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

«Στην οικογένειά μου, που πίστεψε σε εμένα περισσότερο από όλους και μου έδωσε τη δύναμη να φτάσω μέχρι εδώ.»

Πρόλογος

Το βασικό κίνητρο για την επιλογή του συγκεκριμένου θέματος ήταν η επιθυμία μου να εργαστώ πάνω σε ένα υφιστάμενο λογισμικό, αναλαμβάνοντας την πρόκληση να το αναβαθμίσω και να το επεκτείνω. Θεωρώ πως η διαδικασία κατανόησης, διόρθωσης και εξέλιξης ενός ήδη υπάρχοντος κώδικα προσομοιώνει άριστα τις πραγματικές απαιτήσεις του σύγχρονου εργασιακού περιβάλλοντος στο οποίο επιθυμώ να ενταχθώ. Η πορεία αυτή περιλάμβανε αρκετές τεχνικές προκλήσεις και δοκιμές, ωστόσο, η διαδικασία επίλυσης των σφαλμάτων αποτέλεσε μια εξαιρετικά πολύτιμη μαθησιακή εμπειρία, η οποία οδήγησε στην επιτυχή βελτίωση της εφαρμογής.

Περίληψη

Η παρούσα διπλωματική εργασία εστιάζει στην τεχνολογική και λειτουργική αναβάθμιση της διαδικτυακής εφαρμογής "vaseisapp", ενός πληροφοριακού συστήματος που παρέχει ιστορικά δεδομένα και στατιστικά για τις βάσεις εισαγωγής των Πανελλαδικών Εξετάσεων. Βασικό κίνητρο αποτέλεσε η αντιμετώπιση περιορισμών όπως η χειροκίνητη και χρονοβόρα συλλογή δεδομένων, η απουσία προηγμένων αναλυτικών εργαλείων και οι ελλείψεις στη χρηστικότητα της διεπαφής.

Για την επίλυση αυτών, αναπτύχθηκε αρχικά ένας πλήρως αυτοματοποιημένος μηχανισμός άντλησης δεδομένων (web scraping) από τον ιστότοπο του Υπουργείου Παιδείας, ο οποίος εξαλείφει την ανθρώπινη παρέμβαση και εξασφαλίζει τη βιωσιμότητα του συστήματος. Στη συνέχεια, υλοποιήθηκε μια εκτενής πειραματική μελέτη αξιοποιώντας αλγορίθμους Μηχανικής Μάθησης. Συγκεκριμένα, εφαρμόστηκαν τεχνικές μη επιβλεπόμενης μάθησης (k-Means, Ιεραρχική Συσταδοποίηση) για την ομαδοποίηση των πανεπιστημιακών τμημάτων με βάση τη διαχρονική συμπεριφορά των μορίων τους, καθώς και μοντέλα επιβλεπόμενης μάθησης (Random Forest, XGBoost, k-NN) για την πρόβλεψη μελλοντικών βάσεων. Παράλληλα, εκσυγχρονίστηκε πλήρως η διεπαφή χρήστη (frontend) με την προσθήκη διαδραστικών πινάκων, δυναμικών φίλτρων και εργαλείων οπτικοποίησης, καθιστώντας τα πολύπλοκα αποτελέσματα άμεσα κατανοητά στο ευρύ κοινό.

Τα αποτελέσματα έδειξαν ότι τα μοντέλα αποδίδουν ικανοποιητικές προβλέψεις και αναδεικνύουν κρυφά μοτίβα τάσεων, αν και η ακρίβειά τους περιορίζεται αναπόφευκτα από εξωγενείς παράγοντες που δεν αποτυπώνονται στα διαθέσιμα δεδομένα. Συνολικά, η συνδυαστική χρήση αυτοματοποίησης, αναλυτικών μοντέλων και μοντέρνου σχεδιασμού μεταμόρφωσε το vaseisapp σε ένα ολοκληρωμένο εργαλείο ανάλυσης, θέτοντας παράλληλα ισχυρά θεμέλια για μελλοντικές επεκτάσεις, όπως η ενσωμάτωση νευρωνικών δικτύων (LSTM) και ο εμπλουτισμός των μοντέλων με επιπρόσθετα χαρακτηριστικά των τμημάτων.

«vaseis-app2: Extension of the vaseis-app application with automatic data entry mechanisms and expanded data analysis and presentation functions»

LLapushi Genardo

Abstract

This thesis focuses on the technological and functional upgrade of the "vaseisapp" web application, an information system that provides historical data and statistics on the admission thresholds for the Panhellenic Examinations. The primary motivation was to address limitations such as manual and time-consuming data collection, the lack of advanced analytical tools, and shortcomings in the interface's usability.

To address these issues, a fully automated web scraping mechanism was initially developed to extract data from the Ministry of Education's website, which eliminates human intervention and ensures the system's sustainability. Subsequently, an extensive experimental study was conducted using machine learning algorithms. Specifically, unsupervised learning techniques (k-Means, Hierarchical Clustering) were applied to cluster university departments based on the temporal behavior of their enrollment figures, as well as supervised learning models (Random Forest, XGBoost, k-NN) to predict future admission thresholds. At the same time, the user interface (frontend) was fully modernized with the addition of interactive tables, dynamic filters, and visualization tools making complex results immediately understandable to the general public.

The results showed that the models provide satisfactory predictions and reveal hidden trend patterns, although their accuracy is inevitably limited by exogenous factors not captured in the available data. Overall, the combined use of automation, analytical models, and modern design has transformed vaseisapp into a comprehensive analysis tool, while laying a strong foundation for future expansions, such as the integration of neural networks (LSTM) and the enrichment of models with additional features of the segments.

Ευχαριστίες

Ένα μεγάλο ευχαριστώ οφείλω στην οικογένειά μου για την αδιάκοπη στήριξη, την υπομονή και την ενθάρρυνση που μου προσέφεραν όλα αυτά τα χρόνια των σπουδών μου, δίνοντάς μου τη δύναμη να φτάσω στο σημείο που βρίσκομαι σήμερα. Επίσης θα ήθελα να ευχαριστήσω τον Επίκουρο Καθηγητή κ. Στέφανο Ουγιαρόγλου για την πολύτιμη βοήθεια και την εμπιστοσύνη του κατά την εκπόνηση της διπλωματικής μου εργασίας.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1ο: Εισαγωγή

1.1 Οι Πανελλαδικές εξετάσεις και η σημασία των βάσεων

Οι Πανελλαδικές εξετάσεις αποτελούν τον βασικό θεσμό επιλογής των υποψηφίων για την τριτοβάθμια εκπαίδευση στην Ελλάδα. Διεξάγονται ετησίως στο τέλος της σχολικής χρονιάς και βασίζονται στην αξιολόγηση των μαθητών σε τέσσερα συγκεκριμένα μαθήματα ανάλογα με το επιστημονικό πεδίο που έχουν επιλέξει. Το αποτέλεσμά τους σε συνδυασμό με τη συμπλήρωση του μηχανογραφικού καθορίζει την εισαγωγή σε πανεπιστημιακά τμήματα μέσω των βάσεων εισαγωγής. Οι βάσεις εισαγωγής αποτελούν το κατώτατο όριο μορίων που καθορίζει τον τελευταίο εισακτέο σε κάθε πανεπιστημιακό τμήμα και διαμορφώνονται κάθε χρόνο από τον συνδυασμό τριών μεταβλητών: τον βαθμό δυσκολίας των θεμάτων, τον αριθμό των διαθέσιμων θέσεων και τις προτιμήσεις των υποψηφίων στο μηχανογραφικό τους. Η διαμόρφωσή τους δεν είναι στατική και προκύπτει δυναμικά συνδυάζοντας τις επιδόσεις των υποψηφίων, τον βαθμό δυσκολίας των θεμάτων και τα δεδομένα προσφοράς και ζήτησης θέσεων ανά επιστημονικό πεδίο. Επιπλέον με την καθιέρωση της ελάχιστης βάσης εισαγωγής (ΕΒΕ), η είσοδος σε μια σχολή απαιτεί όχι μόνο τη συγκέντρωση των απαραίτητων μορίων αλλά και την επίτευξη ενός ελάχιστου μέσου όρου βαθμολογίας. Συνεπώς οι εξετάσεις αυτές προσδιορίζονται ως ο προθάλαμος των υποψηφίων αλλά και η ευκαιρία τους να ανταγωνιστούν ισότιμα για μια θέση στην ανώτατη εκπαίδευση της χώρας.

1.2 Η εφαρμογή ιστού vaseisapp

Η εφαρμογή vaseis-app αποτελεί πληροφοριακό σύστημα για την ενημέρωση των υποψηφίων σχετικά με τις Πανελλαδικές Εξετάσεις, το οποίο αναπτύχθηκε στα πλαίσια προηγούμενης διπλωματικής εργασίας. Η ιστοσελίδα λειτουργεί ως ένα σημείο πληροφόρησης όπου συγκεντρώνονται οι βάσεις εισαγωγής και τα στατιστικά στοιχεία των σχολών της τριτοβάθμιας εκπαίδευσης. Στην τρέχουσα μορφή της η εφαρμογή παρέχει τη δυνατότητα αναζήτησης και προβολής των βάσεων εισαγωγής ανά έτος. Η παρουσίαση των δεδομένων γίνεται κυρίως μέσω στατικών πινάκων. Το σημαντικότερο ζήτημα της υφιστάμενης κατάστασης εντοπίζεται στη διαδικασία συντήρησης και ενημέρωσης της βάσης δεδομένων διότι το Υπουργείο Παιδείας δημοσιεύει κάθε χρόνο μεγάλο αριθμό διάσπαρτων αρχείων για βάσεις εισαγωγής και στατιστικά προτιμήσεων. Για την επίλυση αυτού του προβλήματος συγκεντρώθηκαν τα δεδομένα πολλών ετών, σχεδιάστηκε σχεσιακή βάση δεδομένων και αναπτύχθηκε ένα δημόσιο διαδικτυακό API για εύκολη ανάκτηση και αξιοποίηση των πληροφοριών. Επίσης δημιουργήθηκε μια διαδικτυακή εφαρμογή που επιτρέπει στους χρήστες να αναζητούν, να συγκρίνουν και να οπτικοποιούν τις βάσεις και τα στατιστικά των τμημάτων. Η παρούσα διπλωματική εργασία προσφέρει μια ολοκληρωμένη λύση που απλοποιεί την πρόσβαση στα δεδομένα και διευκολύνει τόσο τους μαθητές όσο και τους υπόλοιπους ενδιαφερόμενους στην αναζήτηση βάσεων των πανελλαδικών.

1.3 Κίνητρο

Η παρούσα ερευνητική εργασία βασίζεται στην ανάγκη επίλυσης συγκεκριμένων μεθοδολογικών και επιχειρησιακών προκλήσεων που εντοπίζονται στη διαχείριση, την ανάλυση και την οπτικοποίηση των δεδομένων που αφορούν τις βάσεις εισαγωγής στην τριτοβάθμια εκπαίδευση. Τα κίνητρα για την υλοποίηση του προτεινόμενου συστήματος συνοψίζονται στους ακόλουθους τρεις άξονες: την αυτοματοποίηση της ροής συλλογής και επεξεργασίας δεδομένων, την εφαρμογή μοντέλων μηχανικής μάθησης για την εξαγωγή γνώσης και τον εκσυγχρονισμό/βελτιστοποίηση της διεπαφής χρήστη.

Μέχρι σήμερα, η συγκέντρωση των αποτελεσμάτων και των μορίων εισαγωγής από την επίσημη ιστοσελίδα του Υπουργείου Παιδείας πραγματοποιείται μέσω αποσπασματικών, χειροκίνητων ενεργειών. Η απουσία μιας τυποποιημένης και αυτόνομης υποδομής υποβάλλει τον εκάστοτε

διαχειριστή σε μια χρονοβόρα και επαναλαμβανόμενη διαδικασία λήψης, καθαρισμού και μορφοποίησης των αρχείων σε ετήσια βάση. Το γεγονός αυτό αυξάνει κατακόρυφα την πιθανότητα ανθρώπινου λάθους και καθιστά το σύστημα μη βιώσιμο σε βάθος χρόνου. Συνεπώς, βασική επιδίωξη αποτελεί η ανάπτυξη ενός αυτοματοποιημένου μηχανισμού άντλησης δεδομένων από τον ιστό (web scraping), ο οποίος θα εκτελεί τη μεταφόρτορωση, τον μετασχηματισμό και την ασφαλή αποθήκευση των πληροφοριών στη βάση δεδομένων, εξαλείφοντας την ανάγκη για ανθρώπινη παρέμβαση.

Οι παραδοσιακές μέθοδοι στατιστικής επισκόπησης αδυνατούν να αποτυπώσουν τις σύνθετες, μη γραμμικές συσχετίσεις και τις διαχρονικές τάσεις που διέπουν τις διακυμάνσεις των βάσεων εισαγωγής. Ένα σημαντικό κίνητρο της παρούσας μελέτης είναι η μετάβαση από την απλή περιγραφική παρουσίαση σε προηγμένες τεχνικές Μηχανικής Μάθησης, με στόχο τόσο την πρόβλεψη όσο και την εύρεση κρυφών μοτίβων. Στο πλαίσιο αυτό, αξιοποιούνται αλγόριθμοι επίβλεψης (Supervised Learning) για τη διεκπεραίωση εργασιών παλινδρόμησης (regression), όπως τα μοντέλα Random Forest, XGBoost και k-Nearest Neighbors (k-NN), επιτρέποντας την ακριβή εκτίμηση των μελλοντικών τάσεων. Παράλληλα, επιστρατεύονται τεχνικές μη επιβλεπόμενης μάθησης (Unsupervised Learning) για τη συσταδοποίηση (clustering) των σχολών, εφαρμόζοντας τους αλγορίθμους k-Means και Agglomerative Hierarchical Clustering, ώστε να εντοπιστούν ομοιογενείς ομάδες τμημάτων με βάση την κοινή συμπεριφορά των μορίων τους.

Η χρηστικότητα της υφιστάμενης εφαρμογής περιόριζε την αποτελεσματική αλληλεπίδραση του τελικού χρήστη με τον όγκο των παραγόμενων πληροφοριών. Η ανάγκη για μια πιο διαισθητική, δυναμική και ανθρωποκεντρική εμπειρία έδωσε το τρίτο κίνητρο της εργασίας. Οι πραγματοποιηθείσες βελτιώσεις στη διεπαφή (UI/UX) στοχεύουν στην άμεση και ομαλή πλοήγηση, προσφέροντας παράλληλα προηγμένες δυνατότητες φιλτραρίσματος. Με τον τρόπο αυτό, η πολυπλοκότητα των αποτελεσμάτων της μηχανικής μάθησης και των στατιστικών γραφημάτων μετατρέπεται σε κατανοητή οπτική πληροφορία, προσβάσιμη τόσο από εξειδικευμένους αναλυτές όσο και από το ευρύ κοινό (π.χ. υποψηφίους και εκπαιδευτικούς).

1.4 Συνεισφορά

Η αξία της συγκεκριμένης εργασίας δεν περιορίζεται στην ανάδειξη των τεχνικών δυσλειτουργιών του παρελθόντος, αλλά εστιάζει στη δημιουργία και διάθεση λειτουργικών ψηφιακών προϊόντων.

Στο πλαίσιο του έργου σχεδιάστηκε και τέθηκε σε εφαρμογή ένα αυτόνομο σύστημα web scraping, το οποίο καταργεί πλήρως την ανάγκη για ανθρώπινη εποπτεία. Το σύστημα βασίζεται σε μια ευέλικτη αρχιτεκτονική που προσαρμόζεται αυτόματα σε ενδεχόμενες τροποποιήσεις του ιστότοπου του Υπουργείου, εγγυώμενο τη σταθερή ροή και τη σωστή αποθήκευση των δεδομένων.

Η μελέτη εισάγει ένα ολοκληρωμένο υπολογιστικό πλαίσιο για τη δοκιμή αλγορίθμων supervised learning (όπως k-NN, XGBoost και Random Forest), προσαρμοσμένο αποκλειστικά στα δεδομένα των ελληνικών εισαγωγικών εξετάσεων. Στόχος είναι ο εντοπισμός και η μαθηματική τεκμηρίωση του καταλληλότερου μοντέλου πρόβλεψης με βάση τη μείωση των δεικτών σφάλματος MAE και MSE.

Υλοποιήθηκε μια εναλλακτική προσέγγιση ταξινόμησης των τμημάτων μέσω αλγορίθμων ιεραρχικής συσταδοποίησης (Agglomerative) και k-Means. Η καινοτομία έγκειται στην οργάνωση των σχολών σε διακριτές ομάδες με βάση το ιστορικό διακύμανσης των μορίων τους, αφού προηγουμένως εφαρμόστηκαν προηγμένες τεχνικές κανονικοποίησης.

Η έρευνα ολοκληρώθηκε με τη δημιουργία μιας λειτουργικής διαδικτυακής πλατφόρμας. Η εφαρμογή ενσωματώνει στο backend της τις διεργασίες μηχανικής μάθησης, επιτρέποντας την εκτέλεση των αλγορίθμων σε πραγματικό χρόνο ανάλογα με τις επιλογές του χρήστη.

Παραδόθηκε ένα πλήρως ανανεωμένο frontend, το οποίο βελτιστοποιεί την εμπειρία πλοήγησης. Μέσα από διαδραστικούς πίνακες και φίλτρα, οι χρήστες (γονείς, μαθητές και αναλυτές) μπορούν να επεξεργαστούν τα αποτελέσματα των αλγορίθμων με την ίδια ευκολία τόσο από ηλεκτρονικούς υπολογιστές όσο και από φορητές συσκευές.

1.5 Δομή εργασίας

Η παρούσα διατριβή αναπτύσσεται δομικά σε εννέα διακριτά κεφάλαια που καθοδηγούν μεθοδολογικά τον αναγνώστη από τη θεωρία στην πράξη. Αρχικά, τίθεται το γενικότερο πλαίσιο των Πανελλαδικών εξετάσεων, παρουσιάζεται η εφαρμογή vaseisapp και τεκμηριώνονται τα κίνητρα και η συνεισφορά της έρευνας. Στη συνέχεια, διενεργείται μια κριτική επισκόπηση της προγενέστερης έκδοσης του συστήματος ως προς την αρχιτεκτονική και τη βάση δεδομένων του, ενώ παράλληλα παρουσιάζεται το νέο τεχνολογικό οικοσύστημα, τα frameworks και οι βιβλιοθήκες που επιλέχθηκαν για το backend και το frontend.

Η τεχνική υλοποίηση ξεκινά με την ανάλυση του αυτοματοποιημένου μηχανισμού άντλησης δεδομένων (web scraping) από τον ιστότοπο του Υπουργείου Παιδείας και συνεχίζει με την περιγραφή του ανασχεδιασμένου frontend, όπου ενσωματώθηκαν responsive πίνακες και δυναμικά φίλτρα για τη βελτιστοποίηση της χρηστικότητας. Ακολούθως, αναπτύσσεται το θεωρητικό υπόβαθρο των χρονοσειρών, των μεθόδων συσταδοποίησης, των τεχνικών οπτικοποίησης και των αλγορίθμων παλινδρόμησης. Το πειραματικό σκέλος της εργασίας ξεδιπλώνεται με την ομαδοποίηση των τμημάτων βάσει της διαχρονικής συμπεριφοράς των μορίων τους, καθώς και με την παρουσίαση των αποτελεσμάτων των μοντέλων πρόβλεψης και κατηγοριοποίησης. Η μελέτη ολοκληρώνεται με τη σύνοψη των τελικών συμπερασμάτων και την πρόταση κατευθύνσεων για μελλοντικές τεχνικές βελτιώσεις του συστήματος.

Κεφάλαιο 2ο: Μελέτη της Υφιστάμενης Εφαρμογής

2.1 Αρχιτεκτονική

Η αρχιτεκτονική προσέγγιση που διέπει την αρχική έκδοση του πληροφοριακού συστήματος vaseisapp εδράζεται στις αρχές της κατανεμημένης σχεδίασης και των υπηρεσιών ιστού (Service-Oriented Architecture). Ο θεμελιώδης δομικός κανόνας που ακολουθήθηκε είναι ο πλήρης λειτουργικός διαχωρισμός του επιπέδου διαχείρισης και αποθήκευσης των δεδομένων από το περιβάλλον παρουσίασης και αλληλεπίδρασης με τον τελικό χρήστη.

Στον πυρήνα του backend αναπτύχθηκε ένα stateless διαδικτυακό RESTful API, το οποίο υλοποιήθηκε εξολοκλήρου στη γλώσσα προγραμματισμού PHP. Το εν λόγω API αναλαμβάνει τον ρόλο του ενδιάμεσου λογισμικού (middleware), διαχειριζόμενο αποκλειστικά ασύγχρονα αιτήματα μέσω του πρωτοκόλλου HTTP, με κύρια χρήση της ασφαλούς μεθόδου GET για την ανάκτηση πληροφοριών . Η απόκριση του συστήματος προς τα έξω πραγματοποιείται μέσω της δομημένης και ελαφριάς μορφοποίησης JSON, γεγονός που καθιστά το API εύκολα προσπελάσιμο και αξιοποιήσιμο από τρίτους ερευνητές ή εξωτερικές εφαρμογές.

Στην αντίπερα όχθη, το frontend συνιστά μια αυτόνομη διαδικτυακή εφαρμογή (web application) η οποία "καταναλώνει" τους πόρους του API μέσω του Fetch API της JavaScript, εξασφαλίζοντας δυναμική ανανέωση του περιεχομένου χωρίς την ανάγκη ολικής επαναφόρτωσης της ιστοσελίδας. Η αρχιτεκτονική αυτή επιλογή προσφέρει υψηλή επεκτασιμότητα και συντηρησιμότητα, καθώς επιτρέπει την ανεξάρτητη αναβάθμιση του backend ή του frontend χωρίς να επηρεάζεται η συνολική σταθερότητα του πληροφοριακού συστήματος.

2.2 Βάση

Η ανάγκη για συστηματική, ασφαλή και αποδοτική διαχείριση του σύνθετου και πολυδιάστατου όγκου δεδομένων που αφορούν τις Πανελλαδικές εξετάσεις οδήγησε στην επιλογή του σχεσιακού συστήματος διαχείρισης βάσεων δεδομένων (RDBMS) MariaDB. Το σχήμα της βάσης δεδομένων σχεδιάστηκε με γνώμονα την ελαχιστοποίηση του πλεονασμού πληροφορίας και τη διατήρηση της ακεραιότητας των οντοτήτων, αποτελούμενο από έξι διακριτούς πίνακες που διασυνδέονται μέσω σχέσεων ένα-προς-πολλά (1:N):

- **university**: Αποθηκεύει τα τριτοβάθμια εκπαιδευτικά ιδρύματα της χώρας, περιλαμβάνοντας το αυτόματα παραγόμενο αναγνωριστικό (id), τον επίσημο σύντομο τίτλο (title) και την πλήρη, χειροκίνητα επιμελημένη ονομασία τους (full_title).

- **dept**: Καταγράφει τα ακαδημαϊκά τμήματα χρησιμοποιώντας ως πρωτεύον κλειδί τον μοναδικό κωδικό του Υπουργείου Παιδείας (code) και συνδέεται ιεραρχικά με τον πίνακα των πανεπιστημίων μέσω του ξένου κλειδιού uni_id.

- **examtype**: Κωδικοποιεί τους διαφορετικούς τύπους και κατηγορίες εξετάσεων (π.χ. ΓΕΛ, ΕΠΑΛ, ημερήσια, εσπερινά λύκεια, καθώς και τα συστήματα 90% και 10%).

- **specialcat**: Διαχειρίζεται τις ειδικές κοινωνικές κατηγορίες υποψηφίων (π.χ. πολύτεκνοι, τρίτεκνοι) που ίσχυαν κατά τη διάρκεια των ετών.

- **base**: Συνιστά έναν εκ των κεντρικών πινάκων, αποθηκεύοντας τα ιστορικά μόρια εισαγωγής του πρώτου και του τελευταίου επιτυχόντα, τις διαθέσιμες θέσεις, το επιστημονικό πεδίο και το έτος εξέτασης.

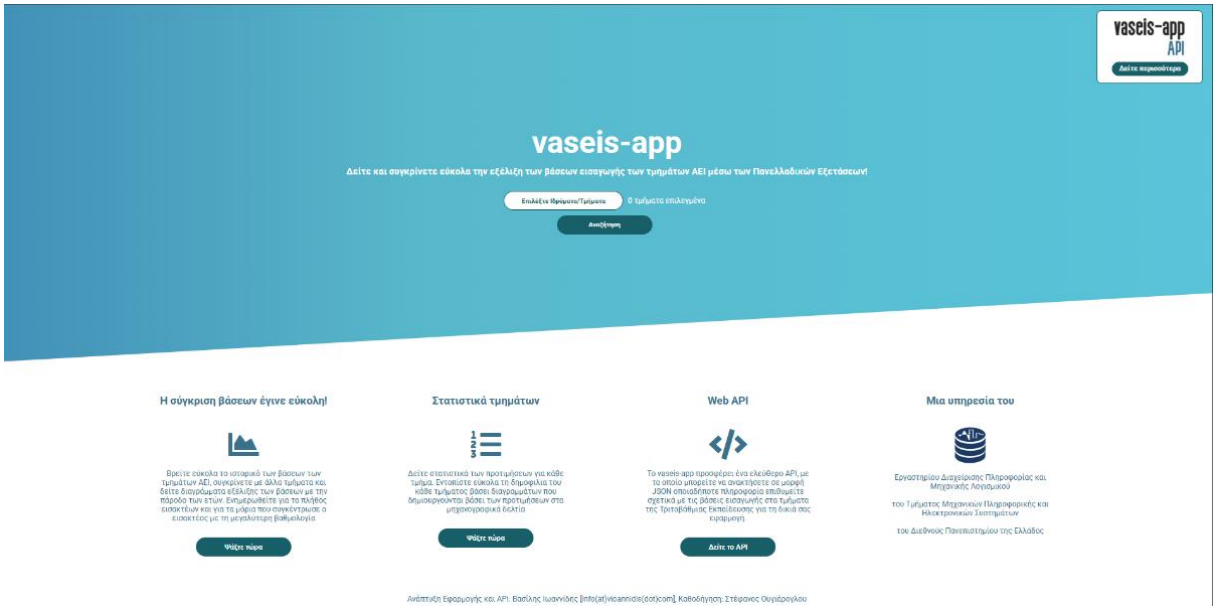
- **statistics**: Αποτυπώνει τα στατιστικά στοιχεία των προτιμήσεων των υποψηφίων, όπως αυτά δηλώθηκαν με σειρά προτεραιότητας στα μηχανογραφικά τους δελτία.

Η τροφοδότηση της βάσης δεδομένων με νέες πληροφορίες πραγματοποιείται μέσω ενός ειδικού περιβάλλοντος backend. Ο διαχειριστής οφείλει να μορφοποιήσει εκ των προτέρων τα ακατέργαστα αρχεία του Υπουργείου σε δύο συγκεκριμένα πρότυπα αρχεία CSV ανά έτος (ένα για τις βάσεις και ένα για τα στατιστικά). Ο ενσωματωμένος αλγόριθμος επεξεργασίας αναλαμβάνει την ανάγνωση αυτών των αρχείων και, μέσω ελεγχόμενων SQL transactions, εκτελεί τις απαραίτητες εγγραφές, διασφαλίζοντας ότι τυχόν σφάλματα κατά τη μεταφόρτωση δεν θα αλλοιώσουν την υφιστάμενη δομή της βάσης.

2.3 Παρουσίαση

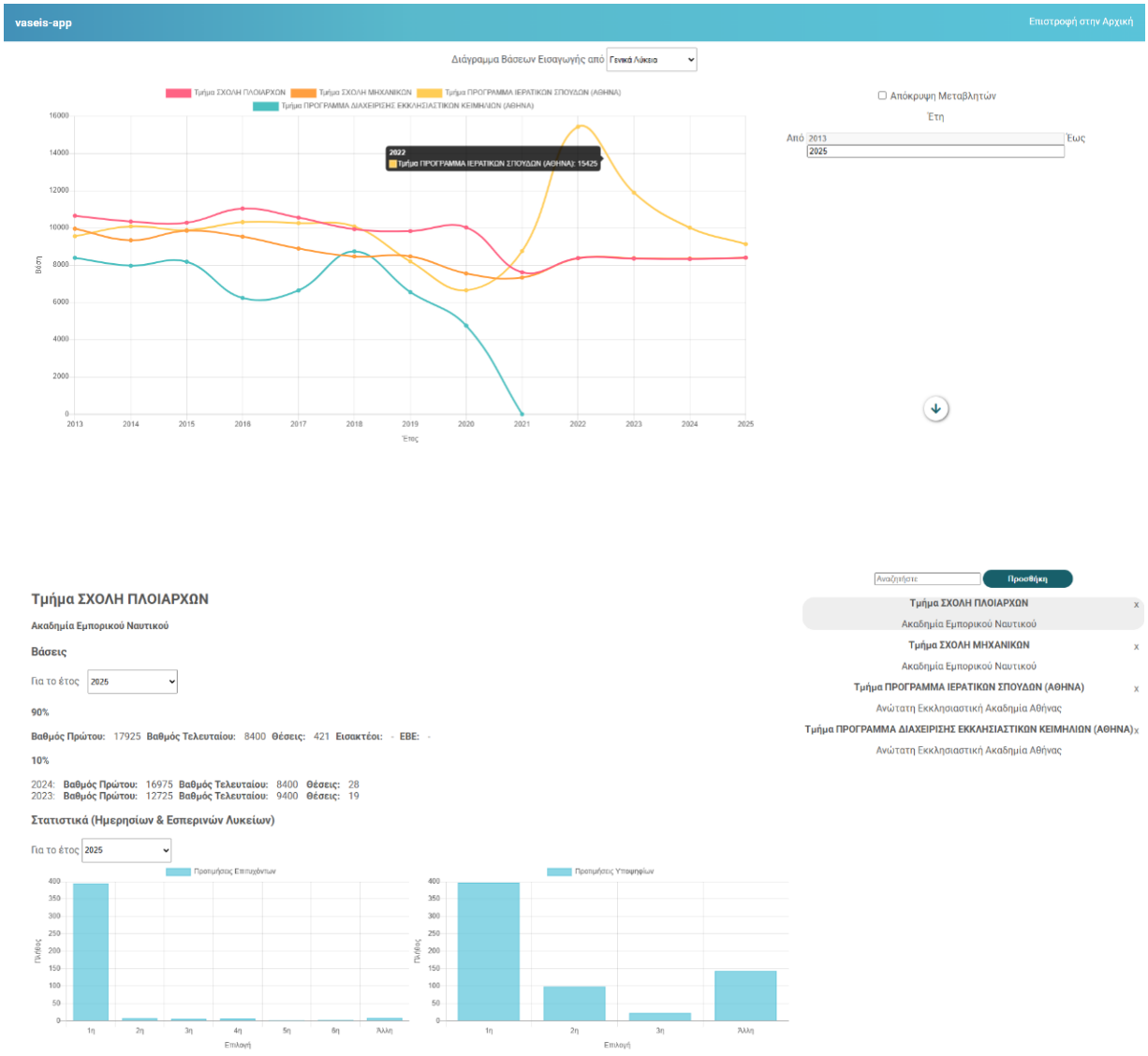
Η διεπαφή χρήστη (User Interface) της υφιστάμενης εφαρμογής σχεδιάστηκε με απόλυτο προσανατολισμό στη χρηστικότητα (Usability) και την απλοποίηση της εμπειρίας πλοήγησης (User Experience). Στόχος ήταν η μετατροπή των δυσνόητων και διάσπαρτων αρχείων του Υπουργείου Παιδείας σε μια ενιαία, κατανοητή οπτική πληροφορία, προσβάσιμη από χρήστες χωρίς εξειδικευμένες τεχνικές γνώσεις. Η εφαρμογή διαρθρώνεται λειτουργικά σε δύο βασικά επίπεδα:

- **Περιβάλλον Αναζήτησης και Επιλογής (Αρχική Σελίδα)**: Παρέχει στον χρήστη ένα φιλικό περιβάλλον με δυναμικά κριτήρια και εργαλεία αναζήτησης, μέσω των οποίων μπορεί να εντοπίσει και να προσθέσει σε μια προσωπική λίστα μελέτης τα ακαδημαϊκά τμήματα που τον ενδιαφέρουν.



Εικόνα 1.Αρχική Σελίδα VaseisApp

- **Περιβάλλον Προβολής Δεδομένων (Σελίδα Δεδομένων)**: Αναλαμβάνει την παρουσίαση του ιστορικού των μορίων και των στατιστικών προτιμήσεων για τα επιλεγμένα τμήματα. Η οπτικοποίηση των δεδομένων επιτυγχάνεται με την ενσωμάτωση της βιβλιοθήκης Chart.js, η οποία παράγει διαδραστικά γραφήματα γραμμής (line charts) που επιτρέπουν την άμεση διαχρονική σύγκριση των βάσεων εισαγωγής.



Εικόνα 2.Σελίδα Προβολής Δεδομένων

Για την υποστήριξη της προσβασιμότητας από ένα ευρύ φάσμα συσκευών, η σχεδίαση ενσωματώνει τεχνικές αποκριτικής συμπεριφοράς (Responsive Web Design) μέσω εγγενών CSS media queries. Το frontend προσαρμόζει αυτόματα τη διάταξη των πινάκων, των μενού και των γραφημάτων, διασφαλίζοντας ότι η ροή της πληροφορίας παραμένει εξίσου λειτουργική και ευανάγνωστη τόσο στην επιφάνεια εργασίας ενός ηλεκτρονικού υπολογιστή όσο και στις περιορισμένες διαστάσεις της οθόνης μιας φορητής συσκευής.

Κεφάλαιο 3ο: Γλώσσες και Τεχνολογίες

3.1 Η έννοια του Web Scraping

Το web scraping αποτελεί μια αυτοματοποιημένη διαδικασία εξαγωγής δεδομένων από ιστοσελίδες, με στόχο τη μετατροπή του μη δομημένου περιεχομένου του διαδικτύου σε οργανωμένη και αξιοποιήσιμη πληροφορία. Καθώς ο παγκόσμιος ιστός περιλαμβάνει τεράστιο όγκο δεδομένων σε μορφές όπως HTML, κείμενο, εικόνες ή PDF, η χειροκίνητη συλλογή τους είναι χρονοβόρα και επιρρεπής σε λάθη [1].

Το web scraping βασίζεται στη χρήση bots (spiders), τα οποία αποστέλλουν αιτήματα σε web servers, λαμβάνουν το περιεχόμενο των σελίδων και στη συνέχεια το αναλύουν ώστε να εντοπίσουν και να εξάγουν τα επιθυμητά στοιχεία [2].Τα παραπάνω στοιχεία που περιέχουν την επιθυμητή πληροφορία εξάγονται σε μορφές όπως CSV, βάσεις δεδομένων ή υπολογιστικά φύλλα. Αν και η χρήση των API αποτελεί πιο αξιόπιστη μέθοδο πρόσβασης σε δομημένα δεδομένα, το web scraping είναι απαραίτητο όταν τέτοια API δεν υπάρχουν.

Η Python αποτελεί την πιο δημοφιλή γλώσσα για web scraping, χάρη στη λιτή σύνταξή της και στις ισχυρές βιβλιοθήκες της, όπως οι BeautifulSoup, Requests και Pandas, που διευκολύνουν την αποστολή HTTP αιτημάτων, την ανάλυση HTML και την επεξεργασία των εξαγόμενων δεδομένων [3].

Παρά τα πλεονεκτήματά του, το web scraping παρουσιάζει προκλήσεις, όπως οι συχνές αλλαγές στη δομή των ιστοσελίδων, οι μηχανισμοί προστασίας (CAPTCHAs, IP blocking) και οι νομικοί/ηθικοί περιορισμοί που σχετίζονται με τους όρους χρήσης και την προστασία δεδομένων [4]. Συνολικά το web scraping αποτελεί κρίσιμη τεχνολογία για την αξιοποίηση των δεδομένων του διαδικτύου, προσφέροντας ταχύτητα, ακρίβεια και δυνατότητα ενσωμάτωσης σε προηγμένα συστήματα ανάλυσης και αυτοματοποίησης.

3.2 Γλώσσα προγραμματισμού Python

Η Python αποτελεί μία από τις σημαντικότερες γλώσσες προγραμματισμού της σύγχρονης εποχής, καθώς συνδυάζει απλότητα, ευελιξία και υψηλή λειτουργικότητα. Ως γλώσσα υψηλού επιπέδου και διερμηνευόμενη, επιτρέπει στον χρήστη να επικεντρώνεται στη λογική επίλυσης προβλημάτων, χωρίς να επιβαρύνεται από περίπλοκες συντακτικές δομές. Η ευκολία εκμάθησής της έχει συμβάλει στην αξιοποίηση της από αρχάριους προγραμματιστές, ενώ ταυτόχρονα η υποστήριξη πολλαπλών παραδειγμάτων προγραμματισμού διαδικαστικού, αντικειμενοστραφούς και λειτουργικού την καθιστά κατάλληλη για απαιτητικές εφαρμογές της βιομηχανίας και της έρευνας [5] .

Η Python έχει εξελιχθεί σε βασικό εργαλείο για την ανάλυση δεδομένων, χάρη στο πλούσιο οικοσύστημα βιβλιοθηκών της, όπως οι NumPy, Pandas, Matplotlib και Seaborn και Scikit-learn, οι οποίες καλύπτουν αριθμητικούς υπολογισμούς, διαχείριση δεδομένων, οπτικοποίηση και βασικές τεχνικές μηχανικής μάθησης. Η διαδικασία ανάλυσης δεδομένων περιλαμβάνει διάφορα στάδια όπως η συλλογή, η επεξεργασία, ο καθαρισμός και η μοντελοποίηση τα οποία υλοποιούνται αποτελεσματικά μέσω των εργαλείων της Python [6].

Αξίζει να σημειωθεί ότι η Python έχει ανακυρηχθεί ως γλώσσα πρώτης επιλογής για νέους αλλά και έμπειρους προγραμματιστές διότι προσφέρει φορητότητα. Είναι επίσης γλώσσα ανοιχτού κώδικα και διαθέτει μεγάλο αριθμό περιβαλλόντων ανάπτυξης (IDE) που διευκολύνουν την παραγωγικότητα. Ο μεγάλος αριθμός μελών στην κοινότητά της συμβάλλει στη συνεχή βελτίωση και επέκταση των δυνατοτήτων της, ενώ η χρήση της σε τομείς όπως η τεχνητή νοημοσύνη, το web development, η ανάλυση δεδομένων και η δημιουργία αυτοματοποιημένων διαδικασιών ενισχύει ακόμη περισσότερο τη σημασία της στον χώρο της πληροφορικής και του προγραμματισμού .

Συνολικά, η Python συνδυάζει χαρακτηριστικά που την καθιστούν ιδανική τόσο για εκπαιδευτικούς σκοπούς όσο και για επαγγελματικές εφαρμογές υψηλής πολυπλοκότητας. Η ευκολία στην χρήση της, η ισχυρή υποστήριξη βιβλιοθηκών και η δυνατότητα εφαρμογής της σε πλήθος επιστημονικών και τεχνολογικών πεδίων επιβεβαιώνουν την κυρίαρχη θέση της ως μία από τις πιο σημαντικές γλώσσες προγραμματισμού.

3.3 Βιβλιοθήκες

3.3.1 Βιβλιοθήκες Επικοινωνίας και Συλλογής (Web Scraping)

Η **BeautifulSoup** είναι μια από τις πιο σημαντικές και ευέλικτες βιβλιοθήκες της Python για web scraping. Η ικανότητά της να αναλύει HTML και XML έγγραφα και να μετατρέπει το αδόμητο περιεχόμενο σε μια οργανωμένη δομή μορφής δένδρου διευκολύνοντας έτσι την πλοήγηση και εξαγωγή των χρήσιμων δεδομένων [7]. Χαρακτηρίζεται ως εύκολη στη χρήση, ειδικά σε σύγκριση με πιο σύνθετες μεθόδους όπως τα regular expressions. Χρησιμοποιείται στο στάδιο της ανάλυσης δεδομένων (Parsing Data), δηλαδή όταν ο crawler έχει ήδη κατεβάσει το HTML και πρέπει να εντοπίσει Tags, attributes, links και κείμενο [8].

Η βιβλιοθήκη **Requests** είναι ένα από τα βασικά εργαλεία στη διαδικασία web scraping, καθώς χρησιμοποιείται για την αποστολή HTTP αιτημάτων και την ανάκτηση του HTML περιεχομένου των ιστοσελίδων. Η Requests αποτελεί το αρχικό βήμα της ροής εργασίας, ο crawler «κατεβάζει» τη σελίδα μέσω ενός αιτήματος και στη συνέχεια το περιεχόμενο περνά στο στάδιο της ανάλυσης (parsing) με BeautifulSoup. Η βιβλιοθήκη λειτουργεί ως ο μηχανισμός που επιτρέπει την πρόσβαση στα δεδομένα, παρέχοντας σταθερότητα και απλότητα στη διαδικασία λήψης HTML το οποίο είναι κρίσιμο για την εξαγωγή δεδομένων όπου απαιτείται συνεχής έλεγχος για νέες ή ενημερωμένες πληροφορίες [9]. Στο συγκεκριμένο άρθρο τονίζεται ότι η Requests είναι απαραίτητη για την αυτοματοποίηση της διαδικασίας, καθώς επιτρέπει την επαναλαμβανόμενη και αξιόπιστη ανάκτηση περιεχομένου, πάνω στην οποία χτίζονται όλα τα επόμενα στάδια του μοντέλου scraping.

Στο πλαίσιο του web scraping απαιτείται μεγάλη προσοχή στη σωστή διαχείριση των URLs κάτι που αποτελεί βασική λειτουργία της βιβλιοθήκης **urllib.parse**. Η συγκεκριμένη βιβλιοθήκη είναι υπομονάδα της Python που επιτρέπει την ανάλυση, σύνθεση και τροποποίηση των URLs [8]. Διευκολύνει με αυτόν τον τρόπο την πλοήγηση σε πολυσέλιδες ιστοσελίδες, την κατασκευή δυναμικών αιτημάτων και την εξαγωγή κρίσιμων στοιχείων όπως το path, τα query parameters και τα fragments. Η σωστή κατανόηση της δομής ενός URL είναι απαραίτητη για τεχνικές όπως η αναγνώριση συνδέσμων και η στοχευμένη εξαγωγή δεδομένων από συγκεκριμένες ενότητες μιας ιστοσελίδας. Η urllib.parse λειτουργεί συμπληρωματικά με εργαλεία όπως Requests και Beautiful Soup, παρέχοντας τον μηχανισμό που επιτρέπει στον scraper να «καταλαβαίνει» και να χειρίζεται τη δομή των διευθύνσεων του ιστού, κάτι που αποτελεί θεμέλιο για πιο σύνθετες διαδικασίες scraping και ανάλυσης HTML [7].

3.3.2 Βιβλιοθήκες Επεξεργασίας Δεδομένων και Μηχανικής Μάθησης

Η **pandas** παρουσιάζεται ως μία από τις πιο σημαντικές και ευρέως χρησιμοποιούμενες βιβλιοθήκες της Python για την αυτοματοποίηση της ανάλυσης δεδομένων. Αυτό επιτυγχάνεται χάρη στις ισχυρές δυνατότητές της στην διαχείριση, τον καθαρισμό και τον μετασχηματισμό των δεδομένων. Προσφέρει ευέλικτες δομές δεδομένων όπως τα DataFrames που επιτρέπουν αποτελεσματική οργάνωση, φιλτράρισμα, συγχώνευση και μετατροπή μεγάλων συνόλων δεδομένων καθιστώντας την ιδανική για το στάδιο της επεξεργασίας δεδομέων (data preprocessing) [10]. Παραμένει η πιο δημοφιλής βιβλιοθήκη διότι βελτιώνει δραστικά την ταχύτητα και την ακρίβεια της ανάλυσης, μειώνοντας τον χρόνο επεξεργασίας από ώρες ή ημέρες σε λίγα δευτερόλεπτα, και καθιστώντας την ιδανική για καθαρισμό και προετοιμασία δεδομένων σε πλήθος εφαρμογών [11].

Η βιβλιοθήκη **re (Regular Expressions)** αποτελεί ισχυρό εργαλείο της Python για την επεξεργασία και τη διαχείριση αλφαριθμητικών (strings). Βασίζεται στην έννοια των κανονικών εκφράσεων οι οποίες είναι ειδικές συμβολαιοσειρές που περιγράφουν ένα μοτίβο αναζήτησης μέσα σε ένα κείμενο. Οι κανονικές εκφράσεις παρουσιάζονται ως ένα από τα πιο αποδοτικά εργαλεία για τον εντοπισμό μοτίβων μέσα σε τεράστιους όγκους πηγαίου κώδικα, επιτρέποντας την ταχεία αναζήτηση, το φιλτράρισμα και την κατηγοριοποίηση χωρίς την ανάγκη πλήρους συντακτικής ανάλυσης [12]. Επίσης η συγκεκριμένη βιβλιοθήκη δεν χρησιμοποιείται μόνο για την εξαγωγή δεδομένων, αλλά και στην λειτουργία της μηχανής αναζήτησης ώστε οι χρήστες να μπορούν να βρίσκουν αποτελέσματα ακόμη και όταν δεν γνωρίζουν την ακριβή λέξη-κλειδί [13].

Η βιβλιοθήκη **scikit-learn** αποτελεί ένα από τα πιο ισχυρά και ευρέως διαδεδομένα εργαλεία της Python για την εφαρμογή αλγορίθμων Μηχανικής Μάθησης (Machine Learning). Η ικανότητά της να παρέχει ένα σταθερό, αποδοτικό και φιλικό προς τον χρήστη περιβάλλον (API) για την ανάλυση και μοντελοποίηση δεδομένων, την καθιστά ιδανική επιλογή τόσο για επιβλεπόμενη (supervised) όσο και για μη επιβλεπόμενη μάθηση (unsupervised learning) [14]. Χαρακτηρίζεται ως εξαιρετικά ευέλικτη, καθώς ενσωματώνεται άψογα με άλλες βιβλιοθήκες επεξεργασίας, όπως η pandas και η NumPy, διευκολύνοντας την προετοιμασία των δεδομένων πριν από την εκπαίδευση των μοντέλων. Στο πλαίσιο της παρούσας εργασίας, η scikit-learn χρησιμοποιείται στο στάδιο της εξαγωγής προτύπων (pattern extraction) και ανάλυσης τάσεων [14]. Συγκεκριμένα, αξιοποιείται για την εκτέλεση αλγορίθμων συσταδοποίησης (clustering), όπως ο k-Means και η Ιεραρχική Συσταδοποίηση (Agglomerative Hierarchical Clustering), με σκοπό την ταξινόμηση και ομαδοποίηση των πανεπιστημιακών τμημάτων βάσει της διαχρονικής συμπεριφοράς των βάσεών τους. Η βιβλιοθήκη παρέχει τους απαραίτητους μηχανισμούς που επιτρέπουν την εύκολη παραμετροποίηση των αλγορίθμων, όπως τον καθορισμό του πλήθους των συστάδων, προσφέροντας παράλληλα υψηλή υπολογιστική ταχύτητα και αξιοπιστία .

3.3.3 Βιβλιοθήκες Διαχείρισης Αρχείων και Συστήματος

Η βιβλιοθήκη **os (Operating System)** επιτρέπει στον προγραμματιστή να αλληλεπιδρά άμεσα και με ασφάλεια με το λειτουργικό σύστημα, παρέχοντας ένα σύνολο εργαλείων για τη διαχείριση αρχείων, φακέλων και διαδρομών. Λειτουργεί ως ενδιάμεσο επίπεδο ανάμεσα στον κώδικα και το σύστημα αρχείων, αναλαμβάνοντας εργασίες όπως ο δυναμικός εντοπισμός διαδρομών (π.χ. ο φάκελος της επιφάνειας εργασίας), η δημιουργία νέων καταλόγων για την οργάνωση δεδομένων αλλά και ο έλεγχος ύπαρξης αρχείων πριν από την επεξεργασία τους [15]. Με αυτόν τον τρόπο, ο προγραμματιστής αποφεύγει σφάλματα που σχετίζονται με διαφορετικές δομές φακέλων ή λειτουργικά συστήματα. Η χρήση της βιβλιοθήκης os ενισχύει σημαντικά την φορητότητα της εφαρμογής καθώς επιτρέπει στον κώδικα να προσαρμόζεται αυτόματα στο περιβάλλον όπου εκτελείται εξασφαλίζοντας σταθερή και προβλέψιμη λειτουργία σε κάθε υπολογιστή.

Η βιβλιοθήκη **glob** αποτελεί χρήσιμο μέσο για τη διαχείριση του συστήματος αρχείων, παρέχοντας τη δυνατότητα εντοπισμού αρχείων και καταλόγων βάσει συγκεκριμένων προτύπων. Η λειτουργία της βασίζεται στον ισχυρό τρόπο αναζήτησης και επιλογής αρχείων με βάση μοτίβα (patterns), χρησιμοποιώντας γνωστικούς χαρακτήρες όπως *, ? και ** [16] . Συγκεκριμένα χρησιμοποιείται για τη μαζική ανάκτηση λιστών αρχείων με συγκεκριμένες επεκτάσεις (π.χ. .xlsx, .zip) από τους τοπικούς φακέλους αποθήκευσης. Σε αντίθεση με απλές μεθόδους παράθεσης αρχείων, η glob επιτρέπει το βέλτιστο φιλτράρισμα των περιεχομένων ενός φακέλου απευθείας κατά την ανάγνωση, διευκολύνοντας την αυτοματοποιημένη ροή εργασιών χωρίς την ανάγκη για επαναληπτικούς ελέγχους. Η απλότητα, η φορητότητα και η άμεση ενσωμάτωσή της με βιβλιοθήκες όπως η pandas την καθιστούν βασικό εργαλείο για κάθε υπολογιστικό έργο που απαιτεί συστηματική πρόσβαση σε αρχεία [16].

Η βιβλιοθήκη **zipfile** αποτελεί απαραίτητο εφόδιο για τον χειρισμό αρχείων συμπιεσμένης μορφής τύπου ZIP. Αποτελεί μια σταθερή και ευρέως χρησιμοποιούμενη λύση για τη διαχείριση συμπιεσμένων δεδομένων προσφέροντας αξιοπιστία και ευελιξία σε διάφορες εφαρμογές. Επιπλέον παρέχει ένα ολοκληρωμένο σύνολο εργαλείων για την ανάγνωση, την εγγραφή, τη δημιουργία και την εξαγωγή περιεχομένων από αρχεία υποστηρίζοντας παράλληλα διάφορες μεθόδους συμπίεσης [17]. Η

ενσωμάτωση της zipfile κρίθηκε απαραίτητη για τη διαχείριση των δεδομένων που ανακτώνται από την ιστοσελίδα του Υπουργείου Παιδείας. Τα δεδομένα των Βάσεων και των Στατιστικών συχνά διατίθονται σε συμπιεσμένη μορφή για λόγους περιορισμού του όγκου μεταφοράς. Η συγκεκριμένη βιβλιοθήκη αναλαμβάνει την αποσυμπίεση των αρχείων στον τοπικό δίσκο ώστε να επεξεργαστούν κατάλληλα σύμφωνα με την αρχική δομή τους.

3.3.4 Βιβλιοθήκες Βάσης Δεδομένων και Διεπαφής

Η **SQLAlchemy** αποτελεί Object-Relational Mapping (ORM) για τη γλώσσα προγραμματισμού Python, προσφέροντας επικοινωνία με σχεσιακές βάσεις δεδομένων. Είναι σχεδιασμένη ώστε να γεφυρώνει το χάσμα μεταξύ του αντικειμενοστρεφούς προγραμματισμού και των σχεσιακών βάσεων δεδομένων και επιτρέπει την αλληλεπίδραση με τη βάση χρησιμοποιώντας κώδικα Python αντί για SQL. Με αυτό τον τρόπο αυξάνεται η παραγωγικότητα αλλά και η αναγνωσιμότητα του κώδικα. Η βιβλιοθήκη αυτή φημίζεται για την ευελιξία της και διευκολύνει σημαντικά τη διαδικασία δοκιμών, καθώς επιτρέπει την εναλλαγή μεταξύ διαφορετικών συστημάτων βάσεων δεδομένων (όπως PostgreSQL και SQLite) χωρίς αλλαγές στον κώδικα [18]. Αυτό καθιστά δυνατή την αναπαραγωγή πολύπλοκων ροών δεδομένων σε περιβάλλοντα δοκιμών με ελάχιστη προσπάθεια. Στην παρούσα υλοποίηση, η επιλογή της SQLAlchemy έγινε κυρίως για να διευκολυνθεί η μαζική εισαγωγή των δεδομένων από τα αρχεία CSV στη MySQL, αποφεύγοντας τη συγγραφή πολύπλοκων και επαναλαμβανόμενων εντολών SQL.

Το **Tkinter** εξυπηρετεί στην δημιουργία γραφικών διεπαφών (GUI) και ξεχωρίζει επειδή είναι ενσωματωμένο στην Python χωρίς να απαιτείται εξωτερική εγκατάσταση. Προσφέρει ένα πλούσιο σύνολο από γραφικά στοιχεία (widgets) όπως κουμπιά, πλαίσια, ετικέτες και παράθυρα τα οποία επιτρέπουν την δημιουργία διαδραστικών εφαρμογών. Χρησιμοποιεί event-driven αρχιτεκτονική, πράγμα που σημαίνει ότι η εφαρμογή ανταποκρίνεται σε ενέργειες του χρήστη (π.χ. κλικ, επιλογές, μετακινήσεις). Η ευκολία στον συνδυασμό του Tkinter με άλλες βιβλιοθήκες το κάνει ιδανικό για εφαρμογές που χρειάζονται οπτικοποίηση και επεξεργασία δεδομένων ή αυτοματοποίηση διαδικασιών [19]. Συνολικά, αποτελεί μια πρακτική λύση για τη δημιουργία καθαρών, λειτουργικών και φιλικών εφαρμογών προς τον χρήστη.

3.4 Συστήματα Διαχείρισης Βάσεων Δεδομένων

Τα Συστήματα Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ) αποτελούν το θεμέλιο για την ασφαλή αποθήκευση, ανάκτηση και διαχείριση της πληροφορίας. Στο πλαίσιο της παρούσας εργασίας, επιλέχθηκε η χρήση ενός Σχεσιακού Συστήματος Διαχείρισης Βάσεων Δεδομένων (RDBMS), καθώς προσφέρει το απαραίτητο πλαίσιο για τη διαχείριση δομημένων δεδομένων με σύνθετες αλληλεπιδράσεις. Το σχεσιακό μοντέλο βασίζεται στην οργάνωση των δεδομένων σε πίνακες (σχέσεις), όπου κάθε γραμμή αντιπροσωπεύει μια οντότητα και κάθε στήλη ένα χαρακτηριστικό της. Τα κύρια χαρακτηριστικά που καθιστούν τα RDBMS ιδανικά για το *vaseis-app2* είναι:

- **Αναφορική Ακεραιότητα (Referential Integrity):** Μέσω της χρήσης Ξένων Κλειδιών (Foreign Keys), διασφαλίζεται ότι δεν μπορούν να υπάρξουν "ορφανές" εγγραφές. Για παράδειγμα, μια εγγραφή στον πίνακα base δεν μπορεί να αναφέρεται σε έναν κωδικό τμήματος που δεν υφίσταται στον πίνακα dept.

- **Γλώσσα SQL (Structured Query Language):** Η SQL παρέχει μια ισχυρή, τυποποιημένη γλώσσα για την εκτέλεση σύνθετων ερωτημάτων, όπως η συσχέτιση δεδομένων από πολλαπλούς πίνακες (π.χ. συνδυασμός βάσεων εισαγωγής με τα στατιστικά προτιμήσεων), η οποία θα ήταν εξαιρετικά δυσχερής σε μη σχεσιακές δομές.

Στο πλαίσιο της υλοποίησης του vaseis-app2, επιλέχθηκε η MySQL ως μέσο υλοποίησης του παραπάνω μοντέλου, λόγω της συνδυαστικής της ικανότητας να προσφέρει υψηλή απόδοση και ευκολία

ενσωμάτωσης. Αποτελεί ένα ανοιχτού κώδικα Σύστημα Διαχείρισης Βάσεων Δεδομένων (DBMS) και ειδικότερα ένα Σχεσιακό Σύστημα Διαχείρισης Βάσεων Δεδομένων (RDBMS). Η λειτουργία της βασίζεται στη χρήση της γλώσσας SQL και στην οργάνωση των δεδομένων σε πίνακες, γεγονός που επιτρέπει την αποδοτική αποθήκευση, ανάκτηση και επεξεργασία πληροφοριών [20]. Χρησιμοποιείται ευρέως σε διαδικτυακά περιβάλλοντα λόγω της ταχύτητας, της αξιοπιστίας και της ευκολίας με την οποία επιτρέπει την αναζήτηση, ταξινόμηση και τροποποίηση δεδομένων [21]. Η MySQL διακρίνεται για την υψηλή της απόδοση σε απλά ερωτήματα, την επεκτασιμότητά της σε βάσεις δεδομένων πολύ μεγάλου μεγέθους και τα ισχυρά επίπεδα ασφάλειας που προσφέρει μέσω μηχανισμών ελέγχου πρόσβασης και κρυπτογράφησης [20]. Επιπλέον, η χρήση του SQLAlchemy ORM (Object-Relational Mapping) γεφυρώνει το χάσμα μεταξύ του σχεσιακού μοντέλου της βάσης και του αντικειμενοστραφούς μοντέλου της Python. Το ORM επιτρέπει στον προγραμματιστή να χειρίζεται τους πίνακες της MySQL ως κλάσεις Python, ενισχύοντας τη συντηρησιμότητα του κώδικα και παρέχοντας έμφυτη προστασία από επιθέσεις τύπου SQL Injection.

3.5 Τεχνολογίες Front-End (HTML, CSS, JavaScript)

Η αποτελεσματική παρουσίαση των αποτελεσμάτων των Πανελλαδικών Εξετάσεων απαιτεί ένα σύγχρονο και αποκρινόμενο (responsive) περιβάλλον εργασίας. Για την επίτευξη αυτού του στόχου, η διεπαφή του *vaseis-app2* βασίστηκε στον συνδυασμό των θεμελιωδών τεχνολογιών του παγκόσμιου ιστού (HTML, CSS, JavaScript), διασφαλίζοντας τη βέλτιστη αλληλεπίδραση του χρήστη με το σύστημα.

3.5.1 HTML

Ως θεμέλιο κάθε ψηφιακής παρουσίας στο διαδίκτυο η HTML (HyperText Markup Language) λειτουργεί ως το πρωτόκολλο οργάνωσης και δόμησης πληροφοριών. Δεν πρόκειται για μια συμβατική γλώσσα προγραμματισμού, αλλά για ένα σύστημα αρχιτεκτονικής σήμανσης που επιτρέπει στους περιηγητές (browsers) να ερμηνεύουν και να προβάλλουν το περιεχόμενο με συγκεκριμένη ιεραρχία, όπως είναι οι επικεφαλίδες, οι λίστες και τα σώματα κειμένου [22] .

Η συντακτική μορφολογία της γλώσσας βασίζεται στη χρήση στοιχείων (elements), τα οποία οριοθετούνται από ειδικά σύμβολα ετικετών (tags). Κάθε δομικό στοιχείο περικλείεται από μια ετικέτα έναρξης (<tag>) και μια ετικέτα τερματισμού (</tag>), ενώ μπορεί να φέρει επιπρόσθετα ορίσματα (attributes). Τα ορίσματα αυτά παρέχουν συμπληρωματικές οδηγίες στο σύστημα, καθορίζοντας συμπεριφορές όπως η μέθοδος επικοινωνίας μιας φόρμας ή η πηγή μιας εικόνας.

Από την πρώτη της εμφάνιση το 1993, η γλώσσα εξελίχθηκε ραγδαία, με την HTML5 να αποτελεί την πλέον σύγχρονη και ισχυρή έκδοση. Η HTML5 επέφερε μια επανάσταση στη σημασιολογική ανάλυση του ιστού, εισάγοντας εξειδικευμένες ετικέτες που αντικατέστησαν τις γενικές δομές, προσδίδοντας νόημα στο περιεχόμενο (semantic web) [23]. Παράλληλα, κατέστησε περιττή τη χρήση εξωτερικών πρόσθετων (όπως το Flash) για την αναπαραγωγή πολυμέσων, ενσωματώνοντας εγγενή υποστήριξη για ήχο και βίντεο [24].

Επιπλέον, η HTML5 διεύρυνε τους ορίζοντες της διαδραστικότητας μέσω προηγμένων API. Εργαλεία όπως το Canvas επιτρέπουν τη δυναμική δημιουργία γραφικών και οπτικοποιήσεων σε πραγματικό χρόνο, κάτι που αξιοποιήθηκε στην παρούσα εργασία για την απεικόνιση των στατιστικών δεδομένων. Συμπερασματικά, η HTML5 πέρα από τον βασικό πυλώνα της εφαρμογής *vaseis-app2*, αποτελεί και το μέσο που μετατρέπει τα ακατέργαστα δεδομένα σε μια συγκροτημένη και λειτουργική διεπαφή χρήστη.

3.5.2 CSS

Η τεχνολογία των CSS (Cascading Style Sheets - Κλιμακωτά Φύλλα Στυλ) αποτελεί το απαραίτητο συμπλήρωμα της HTML, αναλαμβάνοντας αποκλειστικά την οπτική παρουσίαση και τη μορφοποίηση των δομημένων εγγράφων. Η θεμελιώδης φιλοσοφία της CSS βασίζεται στον πλήρη διαχωρισμό του περιεχομένου (content) από την εμφάνιση (presentation). Αυτή η αρχή επιτρέπει στους προγραμματιστές να τροποποιούν την αισθητική μιας ολόκληρης εφαρμογής παρεμβαίνοντας σε ένα κεντρικό αρχείο στυλ, χωρίς να επηρεάζεται η δομική ακεραιότητα του κώδικα HTML.

Η λειτουργία της CSS βασίζεται σε ένα σύστημα κανόνων. Κάθε κανόνας αποτελείται από έναν επιλογέα (selector), ο οποίος στοχεύει ένα συγκεκριμένο στοιχείο της HTML, και ένα μπλοκ δηλώσεων (declaration block) που ορίζει τις οπτικές παραμέτρους. Οι παράμετροι αυτές περιλαμβάνουν ένα ευρύ φάσμα ιδιοτήτων, όπως η τυπογραφία, τα χρωματικά μοντέλα, τα περιθώρια (margins) και οι αποστάσεις (padding). Η ιεραρχική φύση της γλώσσας ("Cascading") επιτρέπει τον καθορισμό προτεραιοτήτων στους κανόνες, διασφαλίζοντας ότι οι μορφοποιήσεις εφαρμόζονται με συνέπεια σε όλη την έκταση της ιστοσελίδας [25].

Η εμφάνιση του προτύπου CSS3 έφερε ριζικές αλλαγές στον τρόπο σχεδίασης, εισάγοντας δυνατότητες που παλαιότερα απαιτούσαν σύνθετο κώδικα ή χρήση εικόνων. Χαρακτηριστικά όπως οι σκιές (box-shadows), οι στρογγυλεμένες γωνίες (border-radius), οι διαβαθμίσεις (gradients) και τα εφέ μετάβασης (transitions) επέτρεψαν τη δημιουργία πλούσιων και διαδραστικών διεπαφών.

Ωστόσο, η σημαντικότερη συνεισφορά της CSS3 στη σύγχρονη εποχή είναι η υποστήριξη του Responsive Web Design (Αποκρινόμενος Σχεδιασμός) [23]. Μέσω των Media Queries, η CSS3 επιτρέπει στην εφαρμογή να αντιλαμβάνεται τα τεχνικά χαρακτηριστικά της συσκευής του χρήστη (όπως το πλάτος της οθόνης ή την ανάλυση) και να αναδιατάσσει δυναμικά το περιεχόμενο [23]. Στην περίπτωση του *vaseis-app2*, η τεχνολογία αυτή είναι ζωτικής σημασίας, καθώς διασφαλίζει ότι οι πολυδιάστατοι πίνακες των βάσεων εισαγωγής παραμένουν ευανάγνωστοι και λειτουργικοί τόσο σε μεγάλες οθόνες σταθερών υπολογιστών όσο και σε περιορισμένες επιφάνειες φορητών συσκευών.

Συνοψίζοντας, η CSS3 δεν αποτελεί απλώς ένα εργαλείο διακόσμησης, αλλά έναν κρίσιμο μηχανισμό που βελτιώνει την εμπειρία του χρήστη (User Experience - UX). Με την οργάνωση της πληροφορίας σε καθαρά οπτικά πλαίσια και τη χρήση προηγμένων συστημάτων διάταξης (όπως το *Flexbox* και το *Grid*), η CSS3 καθιστά τη διεπαφή της εφαρμογής ελκυστική, προσβάσιμη και επαγγελματική, ενισχύοντας την αξιοπιστία του πληροφοριακού συστήματος.

3.5.3 Java Script

Η JavaScript αποτελεί την κυρίαρχη γλώσσα προγραμματισμού για την ανάπτυξη δυναμικών διεπαφών στην πλευρά του πελάτη (client-side scripting). Ενώ η HTML καθορίζει τη δομή και η CSS την αισθητική, η JavaScript προσδίδει "νοημοσύνη" και συμπεριφορά στο περιβάλλον εργασίας, επιτρέποντας την εκτέλεση των σύνθετων λειτουργιών και την αλληλεπίδραση με τον χρήστη σε πραγματικό χρόνο. Πρόκειται για μια διερμηνευμένη γλώσσα υψηλού επιπέδου, η οποία εκτελείται απευθείας από τον περιηγητή, καθιστώντας την απαραίτητη για τη δημιουργία σύγχρονων διαδικτυακών εφαρμογών.

Ένα από τα πλέον κρίσιμα χαρακτηριστικά της JavaScript είναι η ικανότητά της να διαχειρίζεται το DOM (Document Object Model), δηλαδή την ιεραρχική αναπαράσταση της ιστοσελίδας, επιτρέποντας τη δυναμική τροποποίηση του περιεχομένου χωρίς την ανάγκη ολικής ανανέωσης (page refresh). Ιδιαίτερη σημασία για το *vaseis-app2* έχει η χρήση των ασύγχρονων αιτημάτων (Asynchronous JavaScript and XML - AJAX) μέσω του Fetch API. Η τεχνολογία αυτή επιτρέπει στην εφαρμογή να επικοινωνεί με το back-end στο παρασκήνιο, να ανακτά δεδομένα από τη βάση (σε μορφή JSON) και να ενημερώνει τους πίνακες ή τα γραφήματα άμεσα. Η πρακτική αυτή προσφέρει μια ομαλή και

ταχύτατη αλληλεπίδραση, επιβεβαιώνοντας σύγχρονες μελέτες που καταδεικνύουν ότι η ορθή χρήση της JavaScript βελτιώνει κατακόρυφα τη συνολική απόδοση (performance) και την εμπειρία του χρήστη (User Experience - UX) [26].

Πέρα από τη διαχείριση δεδομένων, η JavaScript αποτελεί τον κινητήριο μοχλό για τη διαδραστικότητα της εφαρμογής. Μέσω αυτής υλοποιούνται:

- Φιλτράρισμα Δεδομένων: Η δυνατότητα του χρήστη να αναζητά σχολές, να επιλέγει έτη ή επιστημονικά πεδία και να βλέπει τα αποτελέσματα να αλλάζουν ακαριαία.
- Έλεγχος Εγκυρότητας (Client-side Validation): Η επαλήθευση των στοιχείων που εισάγει ο χρήστης στις φόρμες πριν αυτά αποσταλούν στον εξυπηρετητή.
- Δυναμική Γραφική Απεικόνιση: Η τροφοδοσία βιβλιοθηκών οπτικοποίησης που μετατρέπουν τις στατικές αριθμητικές τιμές σε διαδραστικές χρονοσειρές και διαγράμματα σύγκρισης μορίων.

Η εξέλιξη της JavaScript μέσω των προτύπων ECMAScript την έχει μετατρέψει από μια απλή γλώσσα προσθήκης εφέ σε ένα πανίσχυρο εργαλείο ανάπτυξης λογισμικού. Στο πλαίσιο της παρούσας διπλωματικής, η JavaScript λειτουργεί ως ο "ενορχηστρωτής" που συνδέει τη στατική δομή της HTML με τα δεδομένα της MySQL, δημιουργώντας ένα ζωντανό και αποκρινόμενο πληροφοριακό σύστημα που ανταποκρίνεται στις σύγχρονες απαιτήσεις χρηστικότητας.

Κεφάλαιο 4ο: Υλοποίηση Μηχανισμών Αυτόματης Εισαγωγής

4.1 Ανάλυση της πηγής δεδομένων (Ιστοσελίδα Υπουργείου Παιδείας)

Η κύρια πηγή άντλησης των δεδομένων είναι η επίσημη ιστοσελίδα του Υπουργείου Παιδείας, Θρησκευμάτων και Αθλητισμού (https://www.minedu.gov.gr/baseis-an). Η διαδικασία ανάρτησης και η δομή των πληροφοριών παρουσιάζουν τα εξής χαρακτηριστικά:

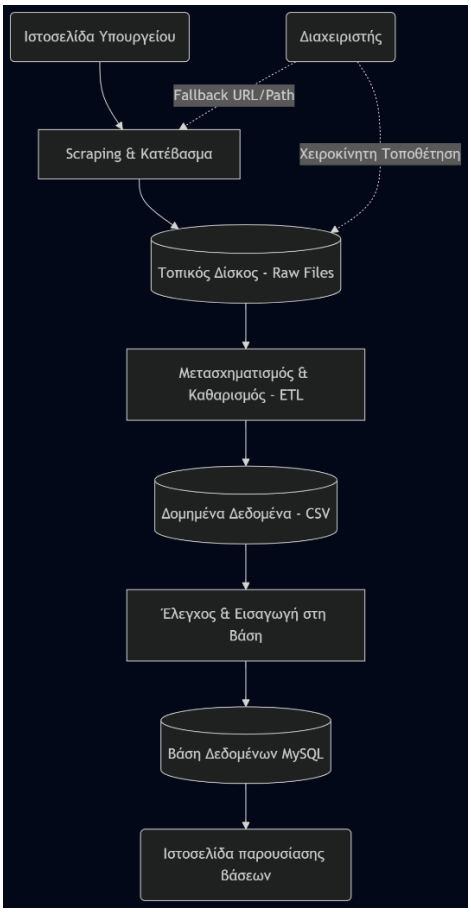
- **Τρόπος Διάθεσης:** Τα δεδομένα ανακοινώνονται μια φορά τον χρόνο, συνήθως τέλη Ιουλίου. Η διάθεσή τους γίνεται μέσω δελτίων τύπου που περιλαμβάνουν συνδέσμους (links) για την λήψη των αρχείων.

- **Τύπος Δεδομένων:** Τα δεδομένα παρέχονται σε συμπιεσμένη μορφή (.zip). Η επιλογή αυτή υποχρεώνει το σύστημα για μια διαδικασία αποσυμπίεσης πριν από οποιαδήποτε επεξεργασία.

- **Δομή Περιεχομένου:** Μετά την αποσυμπίεση τα δεδομένα είναι οργανωμένα σε αρχεία υπολογιστικών φύλλων (Excel). Η πληροφορία είναι διαχωρισμένη σε διαφορετικά αρχεία ανάλογα με την κατηγορία των υποψηφίων, όπως: ΓΕΛ (Ημερήσια/Εσπερινά),ΕΠΑΛ (Ημερήσια/Εσπερινά),Κατηγορίες 10% .

- **Περιεχόμενο Αρχείων:** Κάθε αρχείο Excel περιλαμβάνει σημαντικά πεδία, όπως ο κωδικός της σχολής, το όνομα του ιδρύματος, το όνομα του τμήματος, το είδος θέσης,το επιστημονικό πεδίο, καθώς και τη βάση εισαγωγής (μόρια) του τελευταίου εισακτέου.

Η απουσία ενός δομημένου API (Application Programming Interface) καθιστά απαραίτητη την υλοποίηση ενός μηχανισμού Web Scraping. Ο μηχανισμός αυτός πρέπει να βρίσκεται σε θέση να εντοπίζει δυναμικά τον σωστό σύνδεσμο μέσα στην σελίδα, να πραγματοποιεί την λήψη των αρχείων Excel και να τα διαχειρίζεται, μετατρέποντάς τα σε μορφή συμβατή με τη βάση δεδομένων της εφαρμογής. Για την αντιμετώπιση των παραπάνω περιορισμών και την πλήρη αυτοματοποίηση της διαδικασίας, σχεδιάστηκε και υλοποιήθηκε μια ολοκληρωμένη ροή δεδομένων (data pipeline). Η συνολική αρχιτεκτονική της διαδικασίας αυτής, η οποία ξεκινά από τον εντοπισμό της πηγής και εκτείνεται μέχρι την τελική παρουσίαση των δεδομένων, αποτυπώνεται σχηματικά στην Εικόνα 3. Η αρχιτεκτονική αυτή αναλύεται βήμα βήμα στις υποενότητες που ακολουθούν, ξεκινώντας από την ανάπτυξη του Crawler , προχωρώντας στον καθαρισμό των δεδομένων και καταλήγοντας στον συγχρονισμό τους με τη βάση δεδομένων.



Εικόνα 3.Διάγραμμα αρχιτεκτονικής και ροής δεδομένων

4.2 Ανάπτυξη του Crawler/Scraper

Ο σκοπός του μηχανισμού είναι η αυτοματοποιημένη πλοήγηση στην ιστοσελίδα του Υπουργείου Παιδείας, ο εντοπισμός των ανακοινώσεων για τις βάσεις εισαγωγής και η λήψη των αντίστοιχων αρχείων δεδομένων.Για την υλοποίηση αξιοποιήθηκαν οι παρακάτω βιβλιοθήκες:

- **Requests:** Χρησιμοποιήθηκε για την αποστολή HTTP αιτημάτων (GET requests) προς τους εξυπηρετητές του Υπουργείου. Μέσω αυτής ανακτάται ο πηγαίος κώδικας (HTML) των σελίδων και υλοποιείται η λήψη (download) των αρχείων.

- **BeautifulSoup (bs4):** Αποτέλεσε το βασικό εργαλείο ανάλυσης (parsing) του εγγράφου HTML. Χρησιμοποιήθηκε για την αναζήτηση και εξαγωγή συγκεκριμένων δομικών στοιχείων, όπως οι σύνδεσμοι (<a> tags) που περιέχουν τα επιθυμητά αρχεία.

- **Urllib.parse:** Απαραίτητη για τη διαχείριση των διαδικτυακών διευθύνσεων. Αξιοποιήθηκε η συνάρτηση urljoin για τη δημιουργία πλήρων, έγκυρων URL από σχετικά (relative) paths, καθώς και η unquote για την αποκωδικοποίηση χαρακτήρων, εξασφαλίζοντας τη σωστή ανάγνωση των ελληνικών ονομάτων στα αρχεία.

Ο αλγόριθμος απόκτησης των βάσεων ακολουθεί τα παρακάτω βήματα αυτοματοποίησης:

- **Αναζήτηση Ανακοίνωσης:** Το σύστημα πραγματοποιεί αίτημα στην κεντρική σελίδα ανακοινώσεων του Υπουργείου. Έπειτα σαρώνει όλους τους συνδέσμους αναζητώντας τίτλους που να περιέχουν λέξεις-κλειδιά (π.χ Βάσεις ή Στατιστικά ή «Σχετικά με τα αποτελέσματα εισαγωγής») σε συνδυασμό με το δηλωμένο έτος εξέτασης (π.χ 2024 ή 2025).

- **Φιλτράρισμα και Λήψη Αρχείων:** Μόλις εντοπιστεί η σωστή ανακοίνωση που πληροί τα κριτήρια αναζήτησης, ο αλγόριθμος μεταβαίνει στη σελίδα των αποτελεσμάτων. Εκεί επιλέγει τους συνδέσμους που καταλήγουν σε αναγνωρίσιμους τύπους αρχείων (.zip, .xls, .xlsx) και εφαρμόζει λεκτικούς ελέγχους (string matching) για να διατηρήσει μόνο τα αρχεία που αφορούν τα ΓΕΛ και τα ΕΠΑΛ.

- **Αποσυμπίεση και Οργάνωση:** Μετά τη λήψη των αρχείων στον τοπικό δίσκο, αξιοποιείται η βιβλιοθήκη zipfile για την αποσυμπίεση των αρχείων. Μέσω ελέγχου της ονομασίας τους, το σύστημα διαχωρίζει αυτόματα τα εξαγόμενα έγγραφα σε υποφακέλους ("Vaseis" και "Statistics"), προετοιμάζοντάς τα για το επόμενο στάδιο της ανάλυσης.

4.3 Καθαρισμός και Μετασχηματισμός Δεδομένων

Μετά την επιτυχή λήψη των αρχείων, ακολουθεί το κρίσιμο στάδιο της προετοιμασίας των δεδομένων. Επειδή τα αρχεία που παρέχονται από το Υπουργείο Παιδείας (σε μορφή Excel) είναι μορφοποιημένα για ανθρώπινη ανάγνωση περιέχουν κενά κελιά και ανομοιογενείς στοιχεία. Για την επίλυση αυτού του προβλήματος χρησιμοποιήθηκε η βιβλιοθήκη Pandas, διαχωρίζοντας τη διαδικασία σε δύο διακριτές ροές επεξεργασίας (pipelines), ανάλογα με τον τύπο των αρχείων.

Ροή Επεξεργασίας Αρχείων Βάσεων Εισαγωγής . Τα αρχεία αυτά περιέχουν τα μόρια εισαγωγής και τις διαθέσιμες θέσεις ανά τμήμα. Ο αλγόριθμος εφαρμόζει τα εξής βήματα:

- **Τυποποίηση Σχήματος :** Η ανάγνωση του αρχείου παρακάμπτει την πρώτη περιγραφική γραμμή (header=1). Έπειτα μέσω ενός λεξικού αντιστοίχισης (rename mapping), οι ελληνικές επικεφαλίδες (όπως "ΚΩΔΙΚΟΣ ΣΧΟΛΗΣ", "ΕΙΔΟΣ ΘΕΣΗΣ", "ΒΑΘΜΟΣ ΠΡΩΤΟΥ") μετονομάζονται στα προκαθορισμένα αγγλικά πεδία συμβατά με τη Σχεσιακή Βάση (code, cat_title, vasiprotou αντίστοιχα).

- **Αυστηροποίηση Τύπων Δεδομένων:** Η στήλη του κωδικού της σχολής (code) αποτελεί το βασικό κλειδί ταυτοποίησης. Για τον λόγο αυτό, μετατρέπεται υποχρεωτικά σε αριθμητική μορφή. Οι γραμμές χωρίς έγκυρο κωδικό απορρίπτονται εντελώς μέσω της μεθόδου dropna(), και αυτές οι τιμές που απομένουν ορίζονται αυστηρά ως ακέραιοι (integers), διασφαλίζοντας την ακεραιότητα των δεδομένων.

- **Διαχείριση Ελλειπουσών Στηλών:** Το σύστημα ελέγχει δυναμικά αν στο πρωτογενές αρχείο υφίστανται οι στήλες για το Επιστημονικό Πεδίο (field), τους εισακτέους (admissions) ή την Ελάχιστη Βάση Εισαγωγής (minimumBase). Εάν αυτές απουσιάζουν καθώς η μορφοποίηση του Υπουργείου συχνά αλλάζει από έτος σε έτος δημιουργούνται αυτόματα και αρχικοποιούνται με την τιμή (NULL).

- **Ανάλυση Επικεφαλίδας και Εξαγωγή Τίτλου:** Καθώς ο τύπος της εξέτασης δεν βρίσκεται σε ξεχωριστή στήλη, η πρώτη, ακατέργαστη γραμμή του Excel διαβάζεται μεμονωμένα. Ο αλγόριθμος αναζητά τη λέξη-κλειδί "ΕΠΙΛΟΓΗ", απομονώνει το σχετικό τμήμα του κειμένου (αφαιρώντας περιττούς χαρακτήρες όπως το σύμβολο "--") και, με χρήση κανονικών

εκφράσεων (re.sub), εντοπίζει το τετραψήφιο έτος ώστε να προσθέσει αυτόματα τον όρο "-ΗΜΕΡΗΣΙΑ". Η παραγόμενη συμβολοσειρά αποθηκεύεται στη στήλη title.

- **Φιλτράρισμα και Συγχώνευση:** Το τελικό DataFrame περιορίζεται αποκλειστικά στις επιθυμητές στήλες. Αφού ολοκληρωθεί ο καθαρισμός όλων των αρχείων, τα επιμέρους σύνολα δεδομένων ενοποιούνται και εξάγονται ως ένα προσωρινό ενιαίο αρχείο (vaseis.csv). Το τελικό αυτό αρχείο είναι που θα χρησιμοποιηθεί για την εισαγωγή των δεδομένων στην βάση.

Ροή Επεξεργασίας Αρχείων Στατιστικών . Τα αρχεία στατιστικών περιέχουν τον αριθμό των προτιμήσεων των υποψηφίων ανά σχολή και απαιτούν πιο σύνθετους μετασχηματισμούς:

- **Τυποποίηση Σχήματος και Μετονομασία:** Το σύστημα αγνοεί την πρώτη γραμμή του Excel (header=1). Στη συνέχεια, εφαρμόζεται ένα προκαθορισμένο λεξικό αντιστοίχισης (rename mapping) που μετατρέπει τις ελληνικές επικεφαλίδες (όπως "ΣΧΟΛΗ", "ΠΡΟΤΙΜΗΣΗ 1η") στα τυποποιημένα αγγλικά πεδία της βάσης (name, protimisi1 έως protimisi7).

- **Εξαγωγή δεδομένων από την Ονομασία Αρχείων:** Ο αλγόριθμος αντλεί πληροφορία από το ίδιο το όνομα του αρχείου στο σύστημα αρχείων. Εάν το όνομα περιέχει τη λέξη «επιτυχόντων», στο DataFrame καταχωρείται η τιμή 1 στη νέα στήλη category, ενώ αν περιέχει τη λέξη «υποψηφίων», καταχωρείται η τιμή 0.

- **Ανάλυση Επικεφαλίδας με Κανονικές Εκφράσεις:** Για την εξαγωγή του ακριβούς τύπου εξέτασης, ο κώδικας διαβάζει μεμονωμένα την πρώτη γραμμή του πρωτογενούς αρχείου. Μέσω κανονικών εκφράσεων (π.χ. r"(?:10%\s+)?(ΓΕΛ|ΕΠΑΛ)...") , εντοπίζει αν πρόκειται για εξετάσεις ΓΕΛ ή ΕΠΑΛ και μετασχηματίζει τη συμβολοσειρά, προσθέτοντας αυτόματα τον όρο «ΗΜΕΡΗΣΙΑ & ΕΣΠΕΡΙΝΑ». Το παραγόμενο αποτέλεσμα αποθηκεύεται στη στήλη αναγνώρισης id.

- **Διαχείριση Ελλειπουσών Στηλών (Data Imputation):** Γίνεται δυναμικός έλεγχος για την ύπαρξη όλων των στηλών προτίμησης. Εάν κάποιες σειρές (όπως η 4η, 5η ή 6η προτίμηση) απουσιάζουν από το αρχικό Excel, το σύστημα τις προσθέτει αυτόματα, αρχικοποιώντας τες με μηδενικές τιμές (0), ώστε να εξασφαλιστεί η συνοχή και η συνέπεια πριν τη συγχώνευση.

- **Συγχώνευση (Concatenation):** Αφού ολοκληρωθεί ο καθαρισμός κάθε μεμονωμένου αρχείου, το σύνολο των DataFrames συγχωνεύεται σε ένα ενιαίο σύνολο δεδομένων, το οποίο εξάγεται ως αρχείο CSV (stats.csv). Το τελικό αυτό αρχείο είναι που θα χρησιμοποιηθεί για την εισαγωγή των δεδομένων στην βάση.

4.4 Διαδικασία Συγχρονισμού με τη Βάση Δεδομένων

Η διαδικασία φόρτωσης των βάσεων εισαγωγής από το αρχείο vaseis.csv στον πίνακα base της MySQL, αποτελεί το πιο δύσκολο τμήμα της ροής ETL (Extract, Transform, Load). Λόγω της αυστηρής δομής της βάσης δεδομένων, ο αλγόριθμος ο οποίος σχεδιάστηκε εξασφαλίζει την ακεραιότητα των αναφορών πριν από οποιαδήποτε εισαγωγή.Τα βήματα που εκτελούνται είναι τα εξής:

- **Προετοιμασία και Μετασχηματισμός Τύπων.** Αρχικά, το σύστημα φορτώνει τα δεδομένα στη μνήμη και διαγράφει αχρείαστες πληροφορίες, όπως τη στήλη του ιδρύματος (institution) ώστε να έχει το csv τις ίδες στήλες με τον πίνακα base. Στην συνέχεια εφαρμόζονται αυστηροί κανόνες τύπων δεδομένων, ο κωδικός της σχολής (code) μετατρέπεται υποχρεωτικά σε ακέραιο

αριθμό, αφαιρώντας τις γραμμές με μη έγκυρους κωδικούς. Για τα αριθμητικά πεδία (Βάση Πρώτου, Βάση Τελευταίου, Εισακτέοι, Ελάχιστη Βάση Εισαγωγής), οι κενές τιμές αντικαθίστανται με την τιμή μηδέν (0) και μετατρέπονται και αυτοί σε ακέραιοι. Έπειτα για τα υπόλοιπα πεδία, οι ελλείπουσες τιμές (NaN) μετατρέπονται στον τύπο None της Python ώστε να καταγραφούν ως NULL στη MySQL.

- **Έλεγχος και Ενημέρωση Πινάκων.** Για να μην αποτύχει η εισαγωγή των δεδομένων λόγω παραβίασης Ξένων Κλειδιών το πρόγραμμα αντλεί (μέσω της βιβλιοθήκης SQLAlchemy) τα υπάρχοντα δεδομένα από τους πίνακες της βάσης, examtype (Τύπος Εξέτασης), dept (Τμήματα) και specialcat (Ειδικές Κατηγορίες). Μετά απο σύγκριση των νέων δεδομένων με αυτά που ήδη υπάρχουν,εντοπίζει τις νέες εγγραφές που λείπουν από τη βάση. Στην συνέχεια οι ελλείπουσες εγγραφές εξάγονται τοπικά σε βοηθητικά αρχεία καταγραφής (missing_dept.csv, missing_examtype.csv, missing_specialcat.csv). Έπειτα εισάγονται αυτόματα στους αντίστοιχους πίνακες της βάσης με χρήση ερωτημάτων INSERT IGNORE, ενημερώνοντας πλήρως το σχεσιακό μοντέλο.

- **Ασφαλής Τελική Φόρτωση.** Εφόσον διασφαλιστεί ότι όλοι οι συσχετιζόμενοι πίνακες έχουν ενημερωθεί, πραγματοποιείται η τελική φόρτωση στον βασικό πίνακα base. Η εισαγωγή γίνεται γραμμή προς γραμμή εκτελώντας το ερώτημα INSERT IGNORE. Αυτό το είδος εισαγωγής καθιστά τη διαδικασία ασφαλή απέναντι σε πολλαπλές εκτελέσεις , καθώς το σύστημα της βάσης δεδομένων θα αγνοήσει σιωπηλά τις εγγραφές που προκαλούν διπλοτυπίες, προστατεύοντας την ακεραιότητα των ιστορικών δεδομένων. Τέλος το Γραφικό Περιβάλλον (GUI) ενημερώνεται δυναμικά, ειδοποιώντας τον χρήστη για την επιτυχή ολοκλήρωση της προσθήκης των δεδομένων στην βάση.

```python
# Εισαγωγή vaseis στην βάση

df = pd.read_csv(os.path.join(foldervas, "vaseis.2025.csv"))
# διαγραφή της στήλης institution αν υπάρχει
if "institution" in df.columns:
    df = df.drop(columns=["institution"])
# στήλη code πρεπει να είναι αριθμητική
df['code'] = pd.to_numeric(df['code'], errors='coerce')
df = df.dropna(subset=['code'])
df['code'] = df['code'].astype(int)
# αλλαγή NaN με 0 σε vasiprotou και vasitel
df['vasiprotou'] = df['vasiprotou'].fillna(0).astype(int)
df['vasitel'] = df['vasitel'].fillna(0).astype(int)
# όλα τα NaN σε None για να γίνουν NULL στη MySQL
df = df.where(pd.notnull(df), None)
# παιρνω τα δεδομένα από τους πίνακες
exam_titles = pd.read_sql("SELECT title FROM examtype", engine)
dept_codes = pd.read_sql("SELECT code FROM dept", engine)
specialcat_titles = pd.read_sql("SELECT title FROM specialcat", engine)
# ποιες τιμές λείπουν
missing_titles = df[~df['title'].isin(exam_titles['title'])]['title'].dropna().unique()
missing_codes = df[~df['code'].isin(dept_codes['code'])]['code']
missing_cat_titles = df[~df['cat_title'].isin(specialcat_titles['title'])]['cat_title'].dropna().unique()
print("Missing from examtype:", len(missing_titles))
print("Missing from dept:", len(missing_codes))
print("Missing from specialcat:", len(missing_cat_titles))
# αποθήκευση των missing values
pd.DataFrame(missing_titles, columns=["title"]).to_csv("missing_examtype.csv", index=False)
pd.DataFrame(missing_codes, columns=["code"]).to_csv("missing_dept.csv", index=False)
pd.DataFrame(missing_cat_titles, columns=["cat_title"]).to_csv("missing_specialcat.csv", index=False)
print("The following files were created: missing_examtype.csv, missing_dept.csv, missing_specialcat.csv")
# εισαγωγή των missing values
with engine.begin() as conn:
    for t in missing_titles:
        conn.execute(text("INSERT IGNORE INTO examtype (title) VALUES (:title)"), {"title": t})
    for c in missing_codes:
        conn.execute(text("INSERT IGNORE INTO dept (code) VALUES (:code)"), {"code": int(c)})
    for ct in missing_cat_titles:
        conn.execute(text("INSERT IGNORE INTO specialcat (title) VALUES (:cat_title)"), {"cat_title": ct})

print("The tables examtype, dept, specialcat were updated with the new values")
```

```python
# όλα τα NaN σε None για να γίνουν NULL στη MySQL
df = df.where(pd.notnull(df), None)
# καποιες στήλες να γίνουν 0 αντί για NULL
df['vasiprotou'] = df['vasiprotou'].fillna(0).astype(int)
df['vasitel'] = df['vasitel'].fillna(0).astype(int)
df['admissions'] = df['admissions'].fillna(0).astype(int)
df['minimumBase'] = df['minimumBase'].fillna(0).astype(int)
# εισαγωγή με IGNORE για να αγνοεί διπλότυπα
with engine.begin() as conn:
    for _, row in df.iterrows():
        conn.execute(text("""
            INSERT IGNORE INTO base
            (code, title, cat_title, positions, admissions, field, year, vasiprotou, vasitel, minimumBase)
            VALUES (:code, :title, :cat_title, :positions, :admissions, :field, :year, :vasiprotou, :vasitel, :minimumBase)
        """), row.to_dict())

print("The data was inserted into the base table")


label.config(text="Search Complete! \n Statistics and Vaseis imported to Database")
```

Εικόνα 4.Κώδικας εισαγωγής Βάσεων στην Βάση Δεδομένων

Η εισαγωγή των στατιστικών προτιμήσεων αποτελεί μια ξεχωριστή ροή (pipeline), καθώς απαιτεί ριζικό ανασχηματισμό των δεδομένων πριν την αποθήκευσή τους στο σχεσιακό μοντέλο. Η διαδικασία υλοποιείται μέσω των παρακάτω σταδίων:

- **Αναδιάρθρωση Δεδομένων.** Το αρχείο CSV των στατιστικών διαβάζεται στη μνήμη παρουσιάζοντας τα δεδομένα σε ευρεία μορφή (wide format), όπου κάθε προτίμηση αποτελεί ξεχωριστή στήλη. Χρησιμοποιώντας τη συνάρτηση melt της βιβλιοθήκης Pandas, εκτελείται αποσυγκέντρωση των δεδομένων. Ως μεταβλητές ταυτοποίησης παραμένουν οι στήλες code, id, category και year, ενώ οι στήλες προτιμήσεων συγχωνεύονται σε δύο νέες στήλες, την protimisi (που δηλώνει τη σειρά προτίμησης) και το plithos (που δηλώνει τον αριθμό των υποψηφίων).

- **Κανονικοποίηση και Μετατροπή Τύπων.** Μετά τον ανασχηματισμό, οι τιμές της νέας πλέον στήλης protimisi (π.χ. "protimisi1", "protimisi_other") καθαρίζονται. Μέσω συναρτήσεων αντικατάστασης συμβολοσειρών (str.replace) αφαιρείται το πρόθεμα "protimisi" και η τιμή "other" αντιστοιχίζεται στον αριθμό 7. Το τελικό αποτέλεσμα μετατρέπεται σε ακέραιο τύπο (integer).

- **Βελτιστοποίηση Χώρου.** Για τη μείωση του όγκου των δεδομένων και την αποφυγή περιττών εγγραφών στη βάση, εφαρμόζεται ένας μηχανισμός φιλτραρίσματος Συγκεκριμένα αν η εγγραφή αφορά υποψηφίους (category = 0), εξετάζεται για τις προτιμήσεις 4, 5 και 6. Στην περίπτωση που το πλήθος αυτών των προτιμήσεων είναι μηδενικό (plithos = 0),τότε η εγγραφή απορρίπτεται δυναμικά και δεν προωθείται για εισαγωγή.Με αυτό τον τρόπο βελτιστοποιείται η απόδοση των μελλοντικών ερωτημάτων (queries).

- **Ασφαλής Μαζική Εισαγωγή.** Τέλος, το παραγόμενο σύνολο δεδομένων df_long διατρέχεται γραμμή προς γραμμή. Η διαχείριση της σύνδεσης με τη MySQL γίνεται μέσω του engine.begin(), το οποίο διασφαλίζει τη σωστή διαχείριση των συναλλαγών. Η εισαγωγή εκτελείται με χρήση της εντολής INSERT IGNORE σε συνδυασμό με παραμετροποιημένα ερωτήματα, επιτρέποντας στο σενάριο να επανεκτελεστεί με ασφάλεια χωρίς να υπάρχει ο κίνδυνος δημιουργίας διπλότυπων εγγραφών.

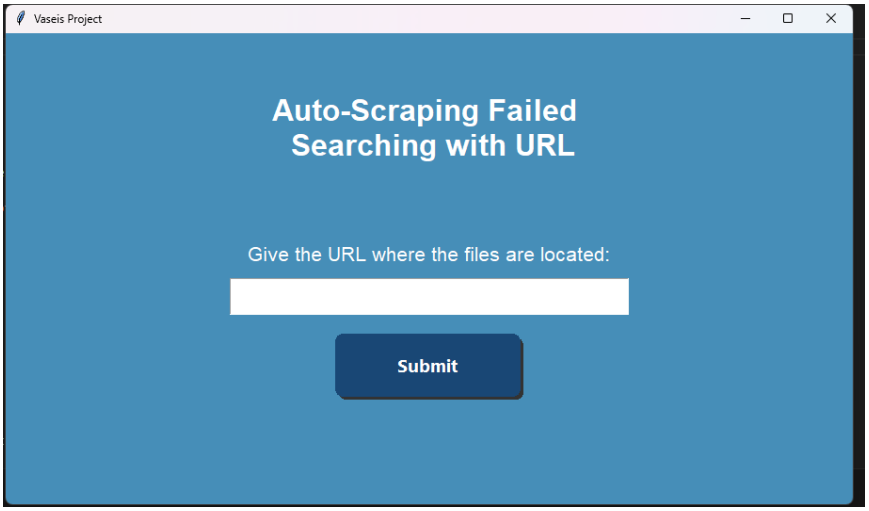
```python
# Εισαγωγή statistics στην βάση

df = pd.read_csv(os.path.join(folderstat, "stats.2025.csv"))


# Μετατροπή από wide σε long format
protimisi_cols = [c for c in df.columns if c.startswith("protimisi")]
df_long = df.melt(
    id_vars=["code","id","category","year"],
    value_vars=protimisi_cols,
    var_name="protimisi",
    value_name="plithos"
)

# καθαρισμός: protimisi1 -> 1, protimisi2 -> 2, ..., protimisi_other -> 7
df_long["protimisi"] = df_long["protimisi"].str.replace("protimisi","")
df_long["protimisi"] = df_long["protimisi"].replace({"other":"7"})
df_long["protimisi"] = df_long["protimisi"].astype(int)

# αν category=0 και protimisi είναι 4,5,6 τότε κράτα μόνο όσες έχουν plithos > 0
mask = ~((df_long["category"] == 0) & (df_long["protimisi"].isin([4,5,6])) & (df_long["plithos"] == 0))
df_long = df_long[mask]


# εισαγωγή στον πίνακα statistics
with engine.begin() as conn:
    for _, row in df_long.iterrows():   # <-- Χρησιμοποιούμε df_long
        conn.execute(text("""
            INSERT IGNORE INTO statistics
            (code, id, category, protimisi, plithos, year)
            VALUES (:code, :id, :category, :protimisi, :plithos, :year)
        """), row.to_dict())

print("The data was inserted into the statistics table")
```

Εικόνα 5.Κώδικας εισαγωγής Στατιστικών στην Βάση Δεδομένων

4.5 Μηχανισμοί Ανθεκτικότητας

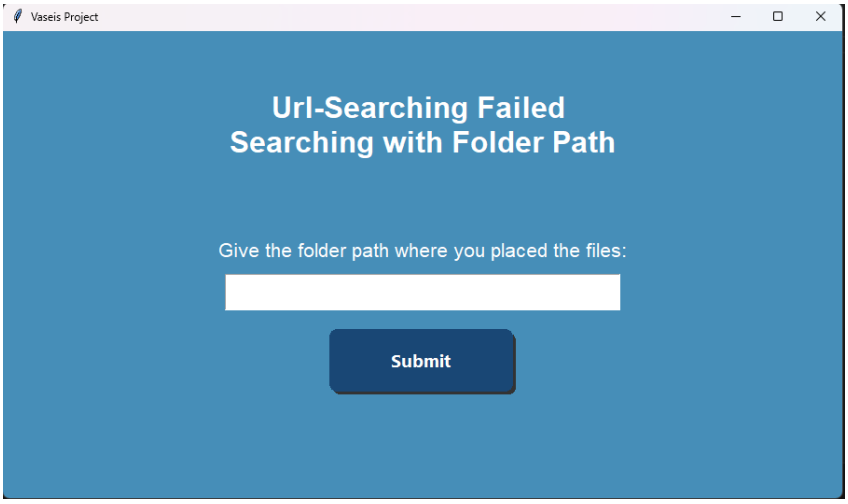
Ένα κρίσιμο χαρακτηριστικό του συστήματος είναι η ανοχή του σε σφάλματα, η οποία υλοποιήθηκε μέσω του Γραφικού Περιβάλλοντος (GUI) που αναπτύχθηκε με τη βιβλιοθήκη tkinter. Σε περίπτωση που η αυτόματη εύρεση της ανακοινώσης αποτύχει για οποιονδήποτε λόγο όπως λόγω αναδιάρθρωσης της ιστοσελίδας του Υπουργείου το σύστημα δεν καταρρέει. Αντιθέτως, ενεργοποιεί ένα εναλλακτικό σενάριο ροής:

- Ειδοποιεί τον χρήστη για την αποτυχία σύνδεσης και του ζητά να εισάγει απευθείας το URL της σελίδας με τα αποτελέσματα (URL Search).



Εικόνα 6.Παράθυρο tkinter αποτυχία Web Scraping

- Εφόσον ούτε το URL λειτουργήσει ή αν δεν υπάρχουν αρχεία στο δίκτυο, το σύστημα δίνει την τελική επιλογή στον χρήστη να τοποθετήσει τα αρχεία χειροκίνητα σε έναν τοπικό φάκελο, παρέχοντας το αντίστοιχο path στην εφαρμογή για να συνεχιστεί η διαδικασία της επεξεργασίας.



Εικόνα 7.Παράθυρο tkinter αποτυχία Url Search

Ο μηχανισμός αυτός μετατρέπει το σύστημα από μια εύθραυστη αυτοματοποιημένη διαδικασία σε μια ανθεκτική υπηρεσία διασφαλίζοντας ότι η ακεραιότητα και η ενημέρωση της βάσης δεδομένων

παραμένουν αδιαπραγμάτευτες, ανεξάρτητα από τις εξωτερικές αλλαγές στο περιβάλλον αντλήσεως δεδομένων.

Κεφάλαιο 5ο: Διευρυμένες Λειτουργίες Παρουσίασης

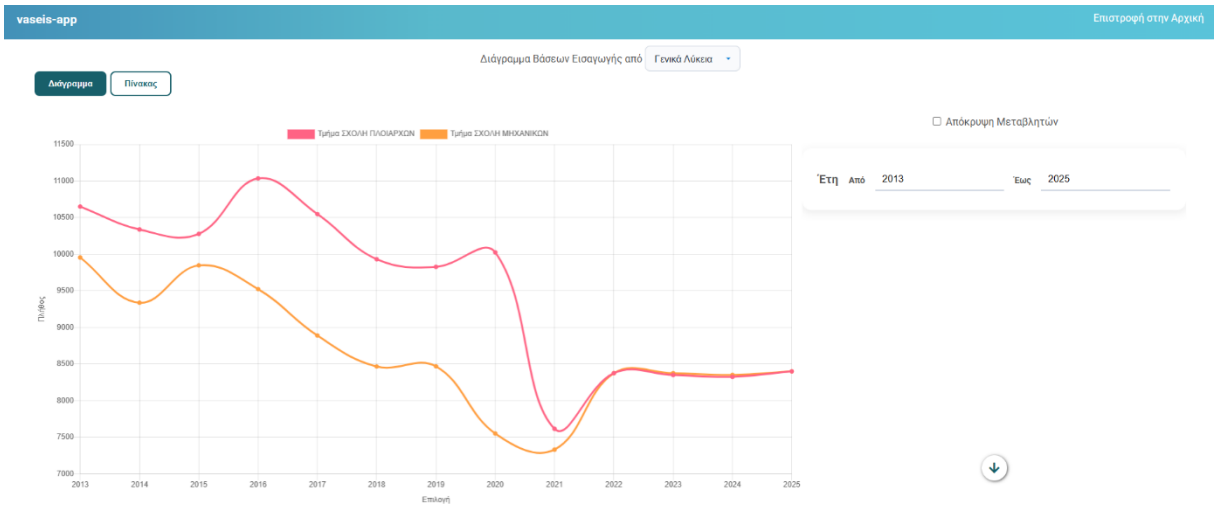
5.1 Ανάπτυξη λειτουργιών ανάλυσης

Η ανάπτυξη των λειτουργιών ανάλυσης επικεντρώθηκε στη μετατροπή των πρωτογενών δεδομένων σε δυναμική πληροφόρηση, επιτρέποντας στον χρήστη να εξάγει συμπεράσματα πέρα από την απλή ανάγνωση των βάσεων εισαγωγής. Παρόλο που η βασική δομή επεξεργασίας των δεδομένων παρέμεινε πιστή στον αρχικό σχεδιασμό της εφαρμογής, προστέθηκαν νέες λειτουργίες που διευρύνουν το πεδίο έρευνας. Ο πυρήνας αυτής της διαδικασίας βασίστηκε στην ανάλυση χρονοσειρών, όπου μέσω εξειδικευμένων αλγορίθμων JavaScript επιτεύχθηκε η αυτόματη ευθυγράμμιση και σύγκριση δεδομένων από το 2013 έως το 2025. Το σύστημα διαχειρίζεται αποτελεσματικά τις ασυνέχειες των δεδομένων, εξασφαλίζοντας ότι η μαθηματική απεικόνιση στις γραφικές παραστάσεις παραμένει ακριβής. Παράλληλα, ενσωματώθηκε η δυνατότητα πολυπαραμετρικής ανάλυσης και στατιστικής επεξεργασίας των προτιμήσεων των υποψηφίων, αναδεικνύοντας τη δυναμική και την ελκυστικότητα κάθε σχολής μέσω της ανάλυσης της σειράς προτίμησης των επιτυχόντων (1η, 2η επιλογή κ.ο.κ.), όπως αυτή αποτυπώνεται στα νέα ραβδογράμματα στατιστικών. Η κορύφωση των αναλυτικών δυνατοτήτων του συστήματος επιτυγχάνεται με την ενσωμάτωση αλγορίθμων συσταδοποίησης (cluster analysis) για την αυτόματη ομαδοποίηση των τμημάτων. Συγκεκριμένα, υλοποιήθηκε ο αλγόριθμος K-Means καθώς και τεχνικές Ιεραρχικής Συσταδοποίησης (Hierarchical Clustering), υποστηρίζοντας διαφορετικά κριτήρια διασύνδεσης όπως οι μέθοδοι Ward, Complete, Average και Single linkage. Οι αλγόριθμοι αυτοί δίνουν τη δυνατότητα στον χρήστη να ομαδοποιήσει τις σχολές είτε στατικά, με κριτήριο το απόλυτο βαθμολογικό τους επίπεδο, είτε δυναμικά, βάσει της κοινής τους συμπεριφοράς και των ετήσιων μεταβολών τους στο χρόνο.

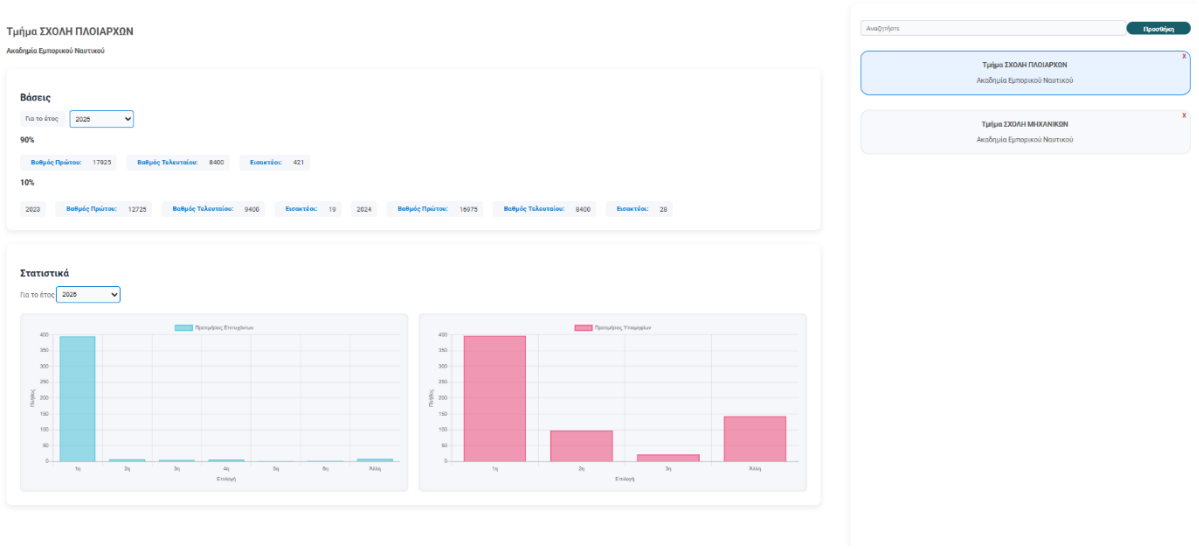
Για την υποστήριξη και την πλήρη ερμηνεία αυτών των μοντέλων μη επιβλεπόμενης μάθησης, η εφαρμογή εξοπλίστηκε με εξειδικευμένα διαδραστικά εργαλεία οπτικοποίησης. Αυτά περιλαμβάνουν διαγράμματα Ανάλυσης Κυρίων Συνιστωσών (PCA) για τη γεωμετρική αποτύπωση των συστάδων, Parallel Coordinates plots για την παρακολούθηση των χρονικών τροχιών των βάσεων, καθώς και Χάρτες Θερμότητας (Heatmaps) για την ανάδειξη των γραμμικών συσχετίσεων μεταξύ των εξεταστικών ετών.

5.2 Σχεδιασμός και Υλοποίηση του User Interface (UI)

Ο σχεδιασμός της διεπαφής χρήστη (UI) υλοποιήθηκε με γνώμονα τις αρχές του σύγχρονου Dashboard Design, αναβαθμίζοντας ριζικά την οπτική ταυτότητα της εφαρμογής χωρίς να αλλοιωθεί η γνώριμη δομή πλοήγησης. Η σημαντικότερη αλλαγή αφορά τη χρήση του Card Layout, όπου κάθε ενότητα πληροφορίας (Βάσεις, Στατιστικά, Διάγραμμα) περικλείεται σε αυτόνομα λευκά πλαίσια με στρογγυλεμένες γωνίες και διακριτικές σκιές (drop shadows), προσδίδοντας βάθος και καθαρότητα στην οθόνη. Για την οπτικοποίηση χρησιμοποιήθηκε η βιβλιοθήκη Chart.js 2.x, η οποία παραμετροποιήθηκε ώστε να προσφέρει ομαλές καμπύλες Bézier και διαδραστικά tooltips. Σημαντική βελτίωση παρατηρείται στο Sidebar (Δεξιά Στήλη), όπου η λίστα των επιλεγμένων τμημάτων ανασχεδιάστηκε με σύγχρονα "pills" που περιλαμβάνουν έντονο περίγραμμα και καθαρό εικονίδιο διαγραφής (X). Η χρήση του συστήματος Flexbox επέτρεψε μια απόκριση διάταξη, ενώ η αισθητική αναβάθμιση ολοκληρώθηκε με τη χρήση ημιδιαφανών χρωμάτων, soft-shadows και μια πιο αέρινη τυπογραφία, προσφέροντας μια επαγγελματική εμπειρία χρήσης που θυμίζει σύγχρονες αναλυτικές πλατφόρμες.

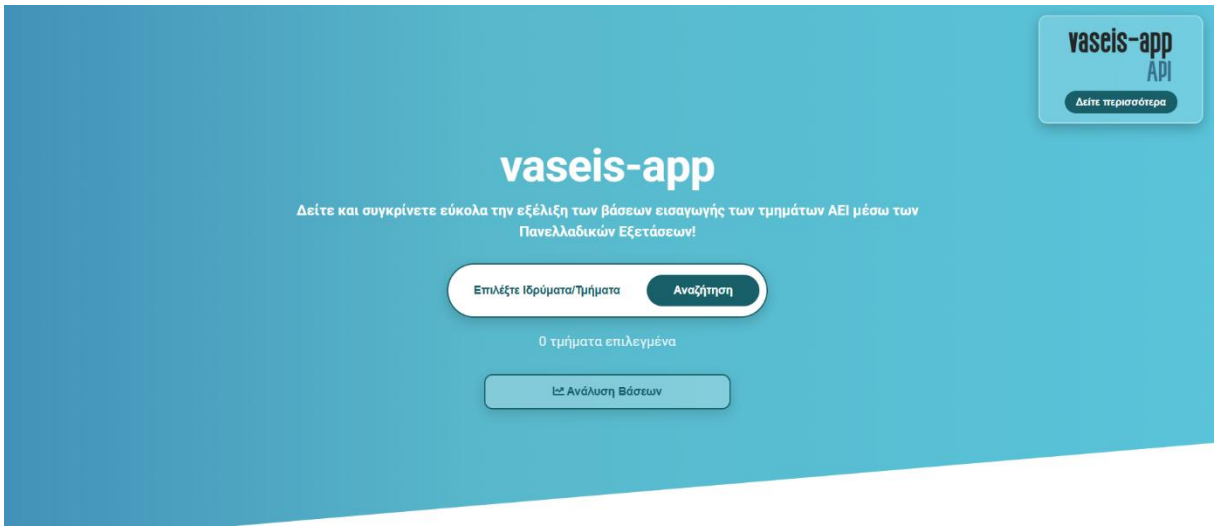


Εικόνα 8.Νέο UI Σελίδας Προβολής Δεδομένων



Εικόνα 9.Νέο UI Σελίδας Προβολής Δεδομένων

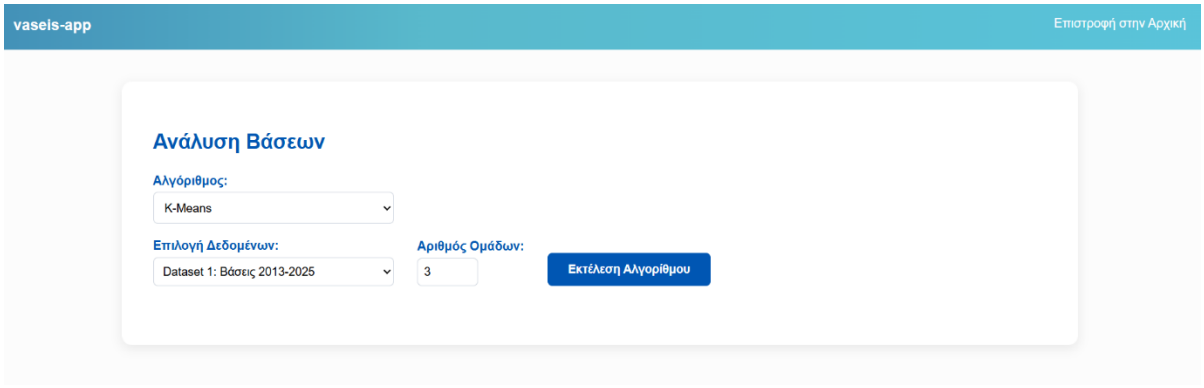
Προκειμένου να ενσωματωθούν ομαλά οι νέες προηγμένες δυνατότητες συσταδοποίησης (Clustering) χωρίς να υπερφορτωθεί η βασική οθόνη αναζήτησης, ανασχεδιάστηκε η εμπειρία πλοήγησης από την αρχική σελίδα της εφαρμογής. Πιο συγκεκριμένα, κάτω από την κεντρική μπάρα αναζήτησης τμημάτων, προστέθηκε ένα νέο, διακριτό κομβίο (button) με την ένδειξη «Ανάλυση Βάσεων».



Εικόνα 10.Η αρχική σελίδα της εφαρμογής με δυνατότητα για «Ανάλυση Βάσεων»

Μέσω αυτού του κομβίου, ο χρήστης δρομολογείται σε ένα εντελώς νέο και εξειδικευμένο περιβάλλον (view), αποκλειστικά αφιερωμένο στην ανάλυση δεδομένων (Data Analysis). Στο ανώτερο τμήμα αυτής της νέας σελίδας κυριαρχεί ένας δυναμικός πίνακας ελέγχου (Control Panel), ο οποίος σχεδιάστηκε ακολουθώντας την ίδια καθαρή φιλοσοφία του Card Layout. Μέσα από αυτή την κάρτα, ο χρήστης καλείται να παραμετροποιήσει το πείραμα της μηχανικής μάθησης με απλό και κατανοητό τρόπο.

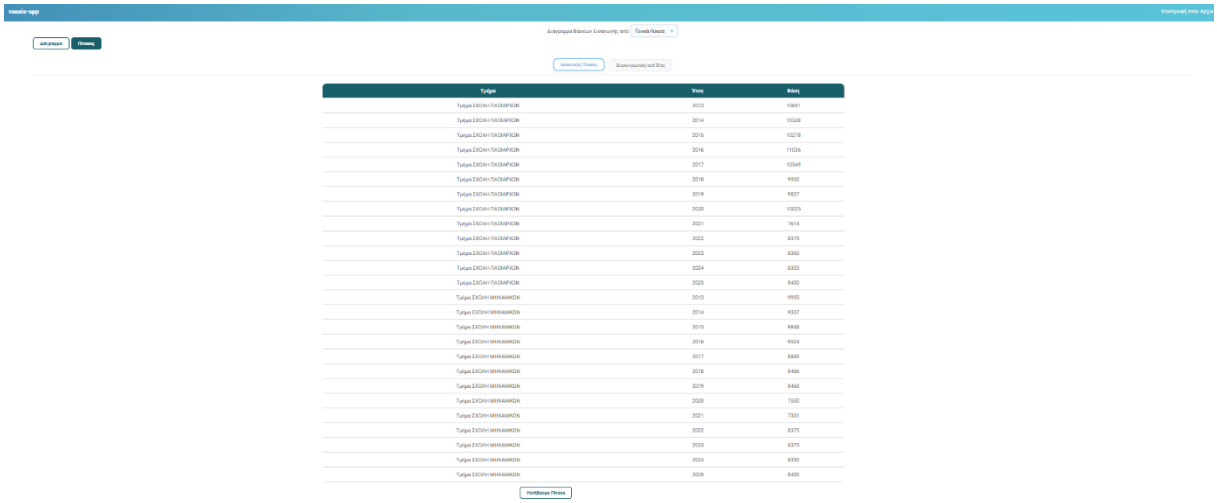
Το περιβάλλον ελέγχου παρέχει δύο πτυσσόμενα μενού (dropdowns) για την επιλογή του επιθυμητού αλγορίθμου (όπως K-Means) και την επιλογή του συνόλου δεδομένων (Dataset). Παράλληλα, υπάρχει ένα πεδίο ελεύθερης εισαγωγής (input field) όπου ο χρήστης μπορεί να ορίσει τον αριθμό των ομάδων (k) που επιθυμεί να δημιουργήσει ο αλγόριθμος. Η διαδικασία ολοκληρώνεται με το κεντρικό κομβίο «Εκτέλεση Αλγορίθμου», το οποίο πυροδοτεί τους υπολογισμούς στο παρασκήνιο (backend) και επιστρέφει ομαλά τα παραγόμενα διαγράμματα και τις στατιστικές μετρικές (Silhouette Score, SSE) ακριβώς από κάτω, δημιουργώντας μια διαδραστική ροή εργασίας (workflow).



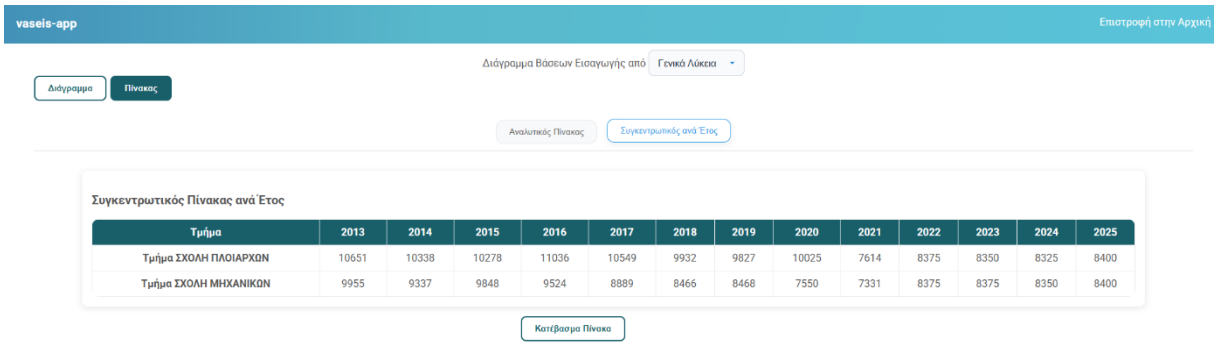
Εικόνα 11. Δυναμικός πίνακας ελέγχου (Control Panel) για την παραμετροποίηση και εκτέλεση της ανάλυσης δεδομένων.

5.3 Παρουσίαση αποτελεσμάτων μέσω πίνακα

Η παρουσίαση των αποτελεσμάτων σε μορφή πίνακα οργανώθηκε σε δύο διακριτά επίπεδα, διατηρώντας τον υφιστάμενο αναλυτικό πίνακα και προσθέτοντας μια νέα, σύνθετη μορφή προβολής. Ο αναλυτικός πίνακας συνεχίζει να λειτουργεί ως λεπτομερής βάση δεδομένων, ενώ ο νέος συγκεντρωτικός πίνακας σχεδιάστηκε ως μια δισδιάστατη μήτρα δεδομένων (matrix view). Σε αυτή τη νέα προσθήκη, τα τμήματα τοποθετούνται στον κατακόρυφο άξονα και τα έτη από το 2013 έως το 2025 στον οριζόντιο, διευκολύνοντας την ιστορική παρακολούθηση μιας σχολής με μια μόνο ματιά. Η υλοποίηση ενισχύθηκε με λειτουργίες sticky headers για τη διατήρηση της αναφοράς των ετών κατά την κύλιση και zebra-striping για βελτιωμένη αναγνωσιμότητα. Επιπλέον, προστέθηκαν νέα controls εναλλαγής (Buttons) μεταξύ των προβολών, τα οποία εμφανίζονται δυναμικά μόνο όταν ο χρήστης επιλέξει την ενότητα του πίνακα. Τέλος, αναπτύχθηκε ένα έξυπνο σύστημα εξαγωγής δεδομένων σε μορφή Excel μέσω της βιβλιοθήκης XLSX, το οποίο προσαρμόζεται αυτόματα στην ενεργή προβολή της οθόνης, προσφέροντας είτε την κλασική αναλυτική λίστα είτε την πλήρη στατιστική μήτρα ετών.



Εικόνα 12. Εμφάνιση αποτελεσμάτων με μορφή αναλυτικού πίνακα



Εικόνα 13. Εμφάνιση αποτελεσμάτων με μορφή συγκεντρωτικού πίνακα.

Κεφάλαιο 6ο: Αλγόριθμοι Μηχανικής Μάθησης και Χρονοσειρές

6.1 Αλγόριθμοι συσταδοποίησης

6.1.1 Εισαγωγή

Οι αλγόριθμοι συσταδοποίησης αποτελούν ένα από τα σημαντικότερα εργαλεία της μη επιβλεπόμενης μάθησης, καθώς επιτρέπουν την ομαδοποίηση δεδομένων χωρίς να απαιτείται προηγούμενη γνώση. Η βασική τους λειτουργία είναι να εντοπίζουν υποκείμενες δομές μέσα σε μεγάλα και συχνά ετερογενή σύνολα δεδομένων, δημιουργώντας συστάδες στις οποίες τα στοιχεία παρουσιάζουν υψηλή ομοιότητα μεταξύ τους και ταυτόχρονα διαφοροποιούνται από στοιχεία άλλων ομάδων. Με αυτόν τον τρόπο η συσταδοποίηση βοηθά στην αποκάλυψη προτύπων που δεν είναι άμεσα εμφανή, επιτρέποντας μια πιο ουσιαστική κατανόηση της συμπεριφοράς των δεδομένων. Στην πράξη, οι αλγόριθμοι αυτοί χρησιμοποιούνται σε ένα ευρύ φάσμα εφαρμογών από την τμηματοποίηση χρηστών και την ανάλυση προτιμήσεων, μέχρι την ανίχνευση ανωμαλιών και την ομαδοποίηση χαρακτηριστικών σε πολυδιάστατα σύνολα δεδομένων. Ενδεικτικά, μέθοδοι όπως ο K-Means, που βασίζεται στη βελτιστοποίηση κεντροειδών, και ο Agglomerative Clustering, που ακολουθεί ιεραρχική προσέγγιση, προσφέρουν διαφορετικές οπτικές και πλεονεκτήματα ανάλογα με τη φύση του προβλήματος. Η χρήση τους καθιστά δυνατή την εξαγωγή συμπερασμάτων που δεν θα μπορούσαν να προκύψουν με απλή παρατήρηση, ενώ παράλληλα λειτουργούν ως θεμέλιο για πιο σύνθετες διαδικασίες ανάλυσης και οπτικοποίησης δεδομένων.

6.1.2 Χρονοσειρές

Οι χρονοσειρές αποτελούν μια ιδιαίτερη κατηγορία δεδομένων όπου η χρονική διάσταση έχει καθοριστικό ρόλο, καθώς κάθε παρατήρηση συνδέεται άμεσα με τις προηγούμενες και επηρεάζει τις επόμενες. Πρόκειται ουσιαστικά για ακολουθίες τιμών που καταγράφονται σε διαδοχικές χρονικές στιγμές, συνήθως με σταθερό χρονικό βήμα, και χρησιμοποιούνται για την ανάλυση της εξέλιξης ενός φαινομένου μέσα στον χρόνο. Η μελέτη χρονοσειρών επιτρέπει την αναγνώριση τάσεων, εποχικών μοτίβων, κυκλικών συμπεριφορών και απότομων μεταβολών. Σε αντίθεση με τα στατικά δεδομένα, οι χρονοσειρές απαιτούν μεθοδολογίες που λαμβάνουν υπόψη τη χρονική τους εξέλιξη [27]. Η ανάλυσή τους μπορεί να γίνει τόσο μέσω τεχνικών μη επιβλεπόμενης μάθησης (για την ομαδοποίηση παρόμοιων τάσεων) όσο και μέσω επιβλεπόμενης μάθησης (για την πρόβλεψη μελλοντικών τιμών). Η χρήση τους είναι ιδιαίτερα σημαντική σε εφαρμογές όπου η χρονική εξέλιξη παίζει κρίσιμο ρόλο, όπως οικονομικά δεδομένα, μετεωρολογικές μετρήσεις και, στο πλαίσιο της παρούσας εργασίας, η μελέτη της διακύμανσης των βάσεων εισαγωγής ανά έτος.

Η μαθηματική προσέγγιση των χρονοσειρών προϋποθέτει, κατά κανόνα, την αποσύνθεσή τους στα βασικά δομικά τους στοιχεία. Μια τυπική χρονοσειρά απαρτίζεται από την τάση (trend), η οποία υποδηλώνει τη μακροπρόθεσμη κατεύθυνση των δεδομένων (όπως η σταδιακή άνοδος ή πτώση της ζήτησης για συγκεκριμένα επιστημονικά πεδία), και την εποχικότητα (seasonality), που αφορά επαναλαμβανόμενα μοτίβα σε σταθερά χρονικά διαστήματα. Επιπλέον, περιλαμβάνει την κυκλικότητα (cyclicity), η οποία περιγράφει πολυετείς διακυμάνσεις που συχνά οφείλονται σε ευρύτερους εξωγενείς παράγοντες (π.χ. μεταρρυθμίσεις στο εκπαιδευτικό σύστημα ή αλλαγές στον τρόπο υπολογισμού των μορίων), καθώς και τον τυχαίο θόρυβο (noise/irregularity), που αντιπροσωπεύει απρόβλεπτες και μη συστηματικές μεταβολές, όπως η απότομη αυξομείωση των βάσεων λόγω του αυξημένου βαθμού δυσκολίας των θεμάτων σε μια συγκεκριμένη εξεταστική χρονιά. Η κατανόηση αυτών των συνιστωσών είναι ζωτικής σημασίας, καθώς διαφοροποιεί ριζικά την ανάλυση χρονοσειρών από την επεξεργασία στατικών δεδομένων, επιβάλλοντας τον αυστηρό σεβασμό στη χρονική αλληλουχία (order of observations) κατά την εφαρμογή οποιουδήποτε αλγορίθμου μηχανικής μάθησης [27].

6.1.3 Συσταδοποίηση χρονοσειρών

Η συσταδοποίηση χρονοσειρών αποτελεί μια εξειδικευμένη μορφή μη επιβλεπόμενης μάθησης, η οποία στοχεύει στην ομαδοποίηση χρονοεξαρτώμενων δεδομένων με βάση τη μορφή, τη δυναμική και τα κοινά μοτίβα που παρουσιάζουν μέσα στον χρόνο. Σε αντίθεση με την κλασική συσταδοποίηση, εδώ βασικό ρόλο παίζει η χρονική αλληλουχία των παρατηρήσεων [28]. Για την εύρεση παρόμοιων χρονοσειρών χρησιμοποιούνται μετρικές ομοιότητας, όπως η Ευκλείδεια απόσταση, η οποία εφαρμόζεται αποτελεσματικά όταν τα δεδομένα μοιράζονται το ίδιο χρονικό πλαίσιο και την ίδια διάρκεια. Στόχος είναι να εντοπιστούν χρονοσειρές που παρουσιάζουν παρόμοια εξελικτική συμπεριφορά. Η συγκεκριμένη διαδικασία είναι ιδιαίτερα χρήσιμη σε εφαρμογές όπου η συνολική τάση ενός φαινομένου έχει μεγαλύτερη σημασία από μεμονωμένες τιμές. Στο πλαίσιο της παρούσας εργασίας, η συσταδοποίηση επιτρέπει την ομαδοποίηση τμημάτων με βάση την ιστορική εξέλιξη των βάσεων εισαγωγής τους, αναδεικνύοντας κοινά μοτίβα ,όπως ομάδες σχολών με σταθερά ανοδική ή καθοδική πορεία, προσφέροντας μια πιο βαθιά κατανόηση της δυναμικής τους.

Θεμελιώδες ζήτημα κατά την εφαρμογή αλγορίθμων συσταδοποίησης σε χρονοσειρές αποτελεί η επιλογή της κατάλληλης μετρικής απόστασης (distance metric), καθώς αυτή καθορίζει τον τρόπο με τον οποίο ορίζεται η ομοιότητα μεταξύ δύο ακολουθιών. Στην περίπτωση που τα δεδομένα είναι απόλυτα ευθυγραμμισμένα χρονικά και διαθέτουν το ίδιο ακριβώς μήκος ,όπως συμβαίνει με τις ετήσιες βάσεις εισαγωγής των τμημάτων, η Ευκλείδεια απόσταση (Euclidean Distance) αποτελεί την πλέον ενδεδειγμένη και υπολογιστικά αποδοτική επιλογή. Η συγκεκριμένη μετρική υπολογίζει τη διαφορά των τιμών ,σημείο προς σημείο, για κάθε αντίστοιχο έτος, επιτρέποντας στους αλγόριθμους να ομαδοποιήσουν απευθείας τα πρωτογενή δεδομένα [28]. Η προσέγγιση αυτή εξασφαλίζει ότι στην ίδια συστάδα (cluster) θα τοποθετηθούν τμήματα των οποίων οι βάσεις όχι μόνο κυμαίνονται σε παρόμοια επίπεδα μορίων, αλλά ακολουθούν και παράλληλες αυξομειώσεις μέσα στα χρόνια. Με τον τρόπο αυτό, η συσταδοποίηση υπερβαίνει την απλή στατική ταξινόμηση και αποκαλύπτει βαθύτερες, δυναμικές σχέσεις προσφοράς και ζήτησης στον ακαδημαϊκό χάρτη.

6.1.4 Ο αλγόριθμος k-means / elbow

Ο αλγόριθμος **K-Means** αποτελεί μία από τις πιο διαδεδομένες τεχνικές συσταδοποίησης και χρησιμοποιείται για την ομαδοποίηση δεδομένων σε K συστάδες με βάση την ομοιότητά τους. Η βασική του λειτουργία στηρίζεται στην έννοια του κεντροειδούς, δηλαδή του σημείου που αντιπροσωπεύει το κέντρο κάθε συστάδας. Αρχικά επιλέγονται τυχαία K κεντροειδή και στη συνέχεια κάθε παρατήρηση ανατίθεται στη συστάδα με το κοντινότερο κέντρο. Τα κεντροειδή ενημερώνονται επαναληπτικά μέχρι να σταθεροποιηθούν, ελαχιστοποιώντας το συνολικό άθροισμα των τετραγωνικών αποστάσεων των σημείων από το κέντρο της συστάδας τους, SSE (Sum of Squared Errors) [29]. Ο K-Means είναι ιδιαίτερα αποτελεσματικός σε μεγάλα σύνολα δεδομένων και προσφέρει γρήγορη σύγκλιση, ωστόσο απαιτεί να οριστεί εκ των προτέρων ο αριθμός των συστάδων, κάτι που δεν είναι πάντα προφανές [29].

Για τον προσδιορισμό του κατάλληλου αριθμού συστάδων χρησιμοποιείται συχνά η **μέθοδος Elbow**, η οποία βασίζεται στη μελέτη της μεταβολής του SSE για διαφορετικές τιμές του K [29]. Καθώς αυξάνεται ο αριθμός των συστάδων, το SSE μειώνεται, όμως μετά από ένα σημείο η μείωση γίνεται οριακή. Το σημείο στο οποίο η καμπύλη παρουσιάζει απότομη αλλαγή κλίσης σχηματίζοντας ένα «αγκώνα» θεωρείται ως η βέλτιστη τιμή του K. Η λογική πίσω από τη μέθοδο είναι ότι πριν από τον «αγκώνα» η προσθήκη νέων συστάδων βελτιώνει ουσιαστικά την ομαδοποίηση, ενώ μετά από αυτό το σημείο η βελτίωση είναι μικρή και δεν δικαιολογεί την αύξηση της πολυπλοκότητας του μοντέλου. Στο πλαίσιο της παρούσας εργασίας, ο συνδυασμός K-Means και Elbow επιτρέπει την ορθολογική επιλογή του αριθμού συστάδων και την αποτελεσματική ομαδοποίηση των δεδομένων, προσφέροντας μια ξεκάθαρη εικόνα των υποκείμενων μοτίβων που εμφανίζονται στις βάσεις εισαγωγής.

6.1.5 Ιεραρχική συσταδοποίηση (ward, complete, single, average)

Η **ιεραρχική συσταδοποίηση** αποτελεί μια από τις πιο ευέλικτες και κατανοητές μεθόδους ομαδοποίησης, καθώς δημιουργεί μια δενδροειδή δομή (dendrogram) που απεικονίζει τη διαδικασία συγχώνευσης ή διάσπασης συστάδων σε διαφορετικά επίπεδα. Σε αντίθεση με αλγορίθμους όπως ο **K-Means**, η ιεραρχική συσταδοποίηση δεν απαιτεί τον προκαθορισμό του αριθμού συστάδων. Αντίθετα, επιτρέπει στον αναλυτή να επιλέξει το κατάλληλο επίπεδο αποκοπής του δενδρογράμματος, ανάλογα με τη δομή των δεδομένων και τον βαθμό ομοιότητας που επιθυμεί να διατηρηθεί.

Στην **συγχωνευτική (agglomerative)** εκδοχή της, η οποία χρησιμοποιείται συχνότερα, κάθε παρατήρηση ξεκινά ως ξεχωριστή συστάδα και σε κάθε βήμα συγχωνεύονται οι δύο πιο «κοντινές» συστάδες σύμφωνα με ένα συγκεκριμένο κριτήριο σύνδεσης [30]. Τα πιο διαδεδομένα κριτήρια είναι τα εξής:

- **Single Linkage**: ορίζει την απόσταση μεταξύ δύο συστάδων ως τη μικρότερη απόσταση μεταξύ οποιουδήποτε ζεύγους σημείων. Τείνει να δημιουργεί «μακριές» και επιμήκεις συστάδες (φαινόμενο chaining), κάτι που μπορεί να οδηγήσει σε λιγότερο συμπαγείς ομάδες [30].
- **Complete Linkage**: χρησιμοποιεί τη μεγαλύτερη απόσταση μεταξύ σημείων δύο συστάδων. Παράγει πιο συμπαγείς και σφικτές συστάδες, αλλά μπορεί να είναι ευαίσθητο σε ακραίες τιμές [30].
- **Average Linkage**: υπολογίζει τη μέση απόσταση μεταξύ όλων των ζευγών σημείων δύο συστάδων [30]. Αποτελεί μια ισορροπημένη προσέγγιση μεταξύ single και complete linkage, προσφέροντας σταθερότερα αποτελέσματα.
- **Ward Linkage**: συγχωνεύει τις συστάδες που προκαλούν τη μικρότερη αύξηση στη συνολική ενδοσυσταδιακή διακύμανση [30]. Παράγει ιδιαίτερα συμπαγείς και καλά διαχωρισμένες συστάδες και θεωρείται μία από τις πιο αποτελεσματικές μεθόδους για αριθμητικά δεδομένα.

Η επιλογή του κατάλληλου κριτηρίου σύνδεσης εξαρτάται από τη φύση των δεδομένων και τη μορφή των συστάδων που επιδιώκεται να εντοπιστούν. Στο πλαίσιο της παρούσας εργασίας, η ιεραρχική συσταδοποίηση χρησιμοποιείται για την ανάλυση της ομοιότητας μεταξύ τμημάτων με βάση τα χαρακτηριστικά τους, επιτρέποντας την οπτική διερεύνηση της δομής των δεδομένων μέσω του δενδρογράμματος και την αναγνώριση ομάδων με κοινά μοτίβα συμπεριφοράς.

6.1.6 Οπτικοποίηση των cluster assignments με διάγραμμα parallel coordinates

Η οπτικοποίηση των αποτελεσμάτων συσταδοποίησης με διάγραμμα parallel coordinates αποτελεί μια ιδιαίτερα χρήσιμη τεχνική για την κατανόηση της δομής των συστάδων σε πολυδιάστατα δεδομένα. Το διάγραμμα αυτό απεικονίζει κάθε χαρακτηριστικό ως έναν κατακόρυφο άξονα και κάθε παρατήρηση ως μια πολυγωνική γραμμή που διατρέχει όλους τους άξονες. Με αυτόν τον τρόπο, οι παρατηρήσεις που ανήκουν στην ίδια συστάδα εμφανίζουν παρόμοια μορφή γραμμής, επιτρέποντας την εύκολη αναγνώριση κοινών μοτίβων και διαφορών μεταξύ των clusters. Η χρωματική κωδικοποίηση των γραμμών ανά συστάδα ενισχύει την οπτική διάκριση και βοηθά στον εντοπισμό χαρακτηριστικών που συμβάλλουν περισσότερο στη διαφοροποίηση των ομάδων. Η μέθοδος parallel coordinates είναι ιδιαίτερα αποτελεσματική όταν ο αριθμός των διαστάσεων είναι μεγάλος, καθώς επιτρέπει την ταυτόχρονη απεικόνιση όλων των χαρακτηριστικών χωρίς απώλεια πληροφορίας. Στο πλαίσιο της παρούσας εργασίας, η συγκεκριμένη οπτικοποίηση χρησιμοποιείται για να αναδειχθούν οι διαφορές μεταξύ των τμημάτων ως προς τα επιλεγμένα χαρακτηριστικά, προσφέροντας μια καθαρή εικόνα της εσωτερικής συνοχής και της μεταξύ τους διαφοροποίησης.

6.1.7 Οπτικοποίηση των cluster assignments με μείωση διαστάσεων (PCA)

Η μείωση διαστάσεων μέσω της μεθόδου Principal Component Analysis (PCA) αποτελεί μια από τις πιο διαδεδομένες τεχνικές για την οπτικοποίηση πολυδιάστατων δεδομένων σε δύο ή τρεις διαστάσεις. Η PCA μετασχηματίζει τα αρχικά χαρακτηριστικά σε ένα νέο σύνολο μη συσχετισμένων συνιστωσών, οι οποίες διατηρούν το μεγαλύτερο δυνατό ποσοστό της συνολικής διακύμανσης των δεδομένων. Με αυτόν τον τρόπο, καθίσταται δυνατή η απεικόνιση των συστάδων σε χαμηλότερη διάσταση, επιτρέποντας την οπτική αξιολόγηση της ποιότητας της συσταδοποίησης. Στο διάγραμμα PCA, τα σημεία που ανήκουν στην ίδια συστάδα τείνουν να συγκεντρώνονται σε κοντινές περιοχές, ενώ οι διαφορετικές συστάδες εμφανίζουν διακριτές χωρικές κατανομές. Η μέθοδος αυτή είναι ιδιαίτερα χρήσιμη για την επιβεβαίωση της διαχωρισιμότητας των clusters, καθώς και για την κατανόηση της συμβολής των χαρακτηριστικών στη διαμόρφωση των κύριων συνιστωσών. Στην παρούσα εργασία, η PCA χρησιμοποιείται ως συμπληρωματικό εργαλείο οπτικοποίησης, επιτρέποντας την αποτύπωση της δομής των συστάδων σε δισδιάστατο χώρο και διευκολύνοντας την ερμηνεία των αποτελεσμάτων της συσταδοποίησης.

6.2 Ανάλυση Παλινδρόμησης (Regression Analysis)

6.2.1 Εισαγωγή

Η ανάλυση παλινδρόμησης (Regression Analysis) αποτελεί μία από τις πιο θεμελιώδεις τεχνικές της στατιστικής και της μηχανικής μάθησης, καθώς επιτρέπει τη διερεύνηση της σχέσης μεταξύ μιας εξαρτημένης μεταβλητής και μίας ή περισσότερων ανεξάρτητων μεταβλητών. Στόχος της είναι να περιγράψει, να ποσοτικοποιήσει και να προβλέψει πώς μεταβάλλεται η εξαρτημένη μεταβλητή όταν αλλάζουν οι ανεξάρτητες. Μέσω της παλινδρόμησης, μπορούμε να εντοπίσουμε τάσεις, να αξιολογήσουμε τη σημασία των παραγόντων που επηρεάζουν ένα φαινόμενο και να δημιουργήσουμε μοντέλα πρόβλεψης με πρακτική εφαρμογή σε πλήθος επιστημονικών πεδίων. Οι πιο συνηθισμένες μορφές είναι η γραμμική παλινδρόμηση, όπου η σχέση μεταξύ των μεταβλητών θεωρείται γραμμική, και η πολυωνυμική ή μη γραμμική παλινδρόμηση, που επιτρέπει πιο σύνθετες συσχετίσεις. Στο πλαίσιο της παρούσας εργασίας, η ανάλυση παλινδρόμησης χρησιμοποιείται για να διερευνηθεί η επίδραση συγκεκριμένων χαρακτηριστικών στις βάσεις εισαγωγής και να εκτιμηθεί η δυνατότητα πρόβλεψης μελλοντικών τιμών, συμβάλλοντας στην καλύτερη κατανόηση των παραγόντων που διαμορφώνουν την εξέλιξή τους.

6.2.2 Random Forest (decision trees)

Τα **δέντρα αποφάσεων (Decision Trees)** αποτελούν μία από τις πιο απλές και κατανοητές τεχνικές μηχανικής μάθησης, καθώς μιμούνται τον τρόπο με τον οποίο ένας άνθρωπος παίρνει αποφάσεις μέσα από μια σειρά διαδοχικών ερωτήσεων. Κάθε κόμβος του δέντρου αντιστοιχεί σε έναν έλεγχο πάνω σε κάποιο χαρακτηριστικό, ενώ κάθε κλαδί οδηγεί σε διαφορετικό αποτέλεσμα ανάλογα με την τιμή του χαρακτηριστικού [31]. Η διαδικασία εκπαίδευσης ενός δέντρου βασίζεται στη διάσπαση των δεδομένων σε όλο και πιο ομοιογενή υποσύνολα, χρησιμοποιώντας μετρικές όπως η εντροπία ή ο δείκτης Gini για να επιλεγεί το χαρακτηριστικό που προσφέρει τον καλύτερο διαχωρισμό. Τα δέντρα αποφάσεων έχουν το πλεονέκτημα ότι είναι εύκολα στην ερμηνεία και μπορούν να αποτυπώσουν μη γραμμικές σχέσεις, όμως συχνά παρουσιάζουν υψηλή ευαισθησία στο θόρυβο και τείνουν να υπερπροσαρμόζονται όταν γίνονται πολύ βαθιά ή πολύ λεπτομερή. Παρά τα μειονεκτήματά τους, αποτελούν τη βάση για πιο σύνθετες και ισχυρές μεθόδους.

Μία από αυτές τις μεθόδους είναι ο αλγόριθμος **Random Forest**, ο οποίος αξιοποιεί τα decision trees ως δομικά στοιχεία και τα συνδυάζει σε ένα ισχυρό σύνολο . Η βασική ιδέα του Random Forest είναι ότι πολλά δέντρα αποφάσεων εκπαιδεύονται ανεξάρτητα το ένα από το άλλο, χρησιμοποιώντας διαφορετικά τυχαία υποσύνολα των δεδομένων και των χαρακτηριστικών. Αυτή η διαδικασία, γνωστή

ως bootstrap sampling και random feature selection, εισάγει ποικιλία μεταξύ των δέντρων και μειώνει σημαντικά την πιθανότητα υπερπροσαρμογής το οποίο αποτελεί κύριο μειονέκτημα των μεμονωμένων decision trees [31] [31]. Η τελική πρόβλεψη προκύπτει από τον συνδυασμό των προβλέψεων όλων των δέντρων: στην ταξινόμηση μέσω πλειοψηφικής ψήφου και στην παλινδρόμηση μέσω του μέσου όρου. Αυτός ο μηχανισμός προσφέρει μεγάλη σταθερότητα, καθώς τα λάθη ενός δέντρου εξισορροπούνται από τα υπόλοιπα.

Ένα ακόμη σημαντικό πλεονέκτημα του Random Forest είναι η δυνατότητά του να παρέχει μετρήσεις σημασίας χαρακτηριστικών, επιτρέποντας στον αναλυτή να εντοπίσει ποιοι παράγοντες επηρεάζουν περισσότερο την τελική πρόβλεψη. Παράλληλα, ο αλγόριθμος απαιτεί ελάχιστη προεπεξεργασία δεδομένων, δεν επηρεάζεται από διαφορετικές κλίμακες τιμών και μπορεί να χειριστεί εύκολα τόσο αριθμητικά όσο και κατηγορικά χαρακτηριστικά [31]. Η ανθεκτικότητά του στον θόρυβο και η ικανότητά του να συλλαμβάνει πολύπλοκες, μη γραμμικές σχέσεις τον καθιστούν ιδανικό για εφαρμογές όπου τα δεδομένα παρουσιάζουν έντονες διακυμάνσεις ή ασυνέπειες.

6.2.3 XGBoost

Ο αλγόριθμος **XGBoost** (Extreme Gradient Boosting) αποτελεί μία από τις πιο ισχυρές και αποδοτικές τεχνικές μηχανικής μάθησης, ειδικά σε προβλήματα παλινδρόμησης και ταξινόμησης όπου απαιτείται υψηλή ακρίβεια και ικανότητα γενίκευσης. Βασίζεται στην οικογένεια αλγορίθμων **boosting**, όπου πολλά «αδύναμα» μοντέλα συνήθως μικρά δέντρα αποφάσεων εκπαιδεύονται διαδοχικά. Κάθε νέο δέντρο προσπαθεί να διορθώσει τα λάθη των προηγούμενων, μαθαίνοντας από τα υπολείμματα (residuals) της πρόβλεψης. Με αυτόν τον τρόπο, το τελικό μοντέλο δεν προκύπτει από ένα μόνο δέντρο, αλλά από τον συνδυασμό πολλών μικρών βημάτων βελτίωσης, κάτι που επιτρέπει στο XGBoost να εντοπίζει πολύπλοκα μοτίβα και μη γραμμικές σχέσεις στα δεδομένα [32].

Ένα από τα βασικά πλεονεκτήματα του XGBoost είναι ότι ενσωματώνει τεχνικές **regularization** (L1 και L2), οι οποίες περιορίζουν την πολυπλοκότητα των δέντρων και μειώνουν την πιθανότητα υπερπροσαρμογής. Παράλληλα, χρησιμοποιεί βελτιστοποιημένους αλγορίθμους για την εύρεση των καλύτερων διασπάσεων, καθώς και έξυπνες τεχνικές διαχείρισης μνήμης και υπολογισμών, που το καθιστούν σημαντικά ταχύτερο από άλλες υλοποιήσεις boosting. Η δυνατότητα για παράλληλη επεξεργασία, η υποστήριξη για sparse δεδομένα, καθώς και η χρήση shrinkage (learning rate) και tree pruning, συμβάλλουν στη σταθερότητα και την υψηλή απόδοση του μοντέλου [32].

Επιπλέον, το XGBoost παρέχει εργαλεία για την αξιολόγηση της **σημασίας των χαρακτηριστικών**, επιτρέποντας στον αναλυτή να εντοπίσει ποιοι παράγοντες επηρεάζουν περισσότερο την τελική πρόβλεψη [32]. Αυτό είναι ιδιαίτερα χρήσιμο σε εφαρμογές όπου η ερμηνευσιμότητα παίζει σημαντικό ρόλο, καθώς βοηθά στην κατανόηση της συμπεριφοράς του μοντέλου και στην εξαγωγή ουσιαστικών συμπερασμάτων από τα δεδομένα. Στο πλαίσιο της παρούσας εργασίας, το XGBoost μπορεί να αξιοποιηθεί για την πρόβλεψη των βάσεων εισαγωγής, καθώς έχει την ικανότητα να συλλαμβάνει λεπτές διαφορές και αλληλεπιδράσεις μεταξύ των χαρακτηριστικών, προσφέροντας υψηλή ακρίβεια και σταθερότητα στα αποτελέσματα. Η ικανότητά του να αποδίδει καλά ακόμη και σε δεδομένα με θόρυβο ή ελλείψεις το καθιστά ιδανική επιλογή για πολύπλοκα εκπαιδευτικά δεδομένα όπως αυτά των βάσεων.

6.2.4 KNN

Ο αλγόριθμος **K-Nearest Neighbors (KNN)** αποτελεί μία από τις πιο απλές αλλά ταυτόχρονα αποτελεσματικές μεθόδους μηχανικής μάθησης, ιδιαίτερα σε προβλήματα ταξινόμησης και παλινδρόμησης. Η βασική του ιδέα είναι ότι μια νέα παρατήρηση μπορεί να προβλεφθεί εξετάζοντας

τις κοντινότερες υπάρχουσες παρατηρήσεις στο σύνολο δεδομένων. Αντί να εκπαιδεύει ένα μοντέλο με την κλασική έννοια, ο KNN αποθηκεύει τα δεδομένα και στη φάση της πρόβλεψης αναζητά τα **K πιο κοντινά σημεία** με βάση ένα μέτρο απόστασης, όπως η Ευκλείδεια απόσταση [33]. Στην ταξινόμηση, η τελική κατηγορία προκύπτει από την πλειοψηφία των γειτόνων, ενώ στην παλινδρόμηση από τον μέσο όρο των τιμών τους. Η απλότητα της μεθόδου την καθιστά ιδιαίτερα εύκολη στην κατανόηση και στην εφαρμογή, ενώ ταυτόχρονα μπορεί να αποδώσει πολύ καλά σε δεδομένα με σαφή τοπικά μοτίβα.

Παρά την απλότητά του, ο KNN έχει ορισμένα σημαντικά χαρακτηριστικά που επηρεάζουν την απόδοσή του. Η επιλογή του αριθμού **K** είναι κρίσιμη: μικρές τιμές οδηγούν σε ευαισθησία στον θόρυβο, ενώ πολύ μεγάλες τιμές μπορεί να «θολώσουν» τα όρια μεταξύ των κατηγοριών. Επίσης, ο αλγόριθμος επηρεάζεται έντονα από την κλίμακα των χαρακτηριστικών, καθώς οι αποστάσεις υπολογίζονται απευθείας πάνω στις τιμές των δεδομένων. Για τον λόγο αυτό, συχνά απαιτείται κανονικοποίηση ή τυποποίηση των χαρακτηριστικών πριν την εφαρμογή του. Ένα ακόμη μειονέκτημα είναι ότι ο KNN μπορεί να γίνει υπολογιστικά ακριβός σε μεγάλα σύνολα δεδομένων, καθώς η πρόβλεψη απαιτεί τον υπολογισμό αποστάσεων από όλα τα σημεία του dataset [33]. Παρ' όλα αυτά, η μέθοδος παραμένει ιδιαίτερα χρήσιμη σε περιπτώσεις όπου η δομή των δεδομένων είναι τοπική και η σχέση μεταξύ των παρατηρήσεων δεν μπορεί να περιγραφεί εύκολα με ένα παγκόσμιο μοντέλο.

Ο KNN μπορεί να αξιοποιηθεί για την ανάλυση και πρόβλεψη των βάσεων εισαγωγής, καθώς επιτρέπει την αναγνώριση τμημάτων με παρόμοια χαρακτηριστικά και την εξαγωγή προβλέψεων βασισμένων σε τοπικές ομοιότητες [33]. Η μέθοδος είναι ιδιαίτερα χρήσιμη όταν οι βάσεις παρουσιάζουν μοτίβα που δεν είναι απαραίτητα γραμμικά ή όταν η συμπεριφορά ενός τμήματος μοιάζει με αυτήν άλλων τμημάτων με παρόμοια ιστορικά δεδομένα. Παράλληλα, η απλότητα και η διαφάνειά του καθιστούν τον KNN μια καλή συμπληρωματική μέθοδο σε σχέση με πιο σύνθετους αλγορίθμους όπως το Random Forest και το XGBoost, προσφέροντας μια διαφορετική οπτική στη δομή των δεδομένων και στη διαδικασία πρόβλεψης.

Κεφάλαιο 7ο: Cluster Analysis στις Χρονοσειρές των Βάσεων Εισαγωγής

7.1 Σύνολα Δεδομένων

Για την εξαγωγή αξιόπιστων και ουσιαστικών συμπερασμάτων από την εφαρμογή των αλγορίθμων συσταδοποίησης, κρίθηκε απαραίτητο να οργανωθούν τα διαθέσιμα δεδομένα σε στοχευμένα και καλοδομημένα σύνολα. Η επιλογή των δεδομένων δεν έγινε τυχαία, αλλά βασίστηκε σε δύο βασικούς άξονες: το χρονικό εύρος και τον τύπο της μεταβλητής που θέλουμε να αναλύσουμε. Με αυτόν τον τρόπο εξασφαλίστηκε ότι κάθε σύνολο δεδομένων θα αναδεικνύει διαφορετικές πτυχές της συμπεριφοράς των τμημάτων.

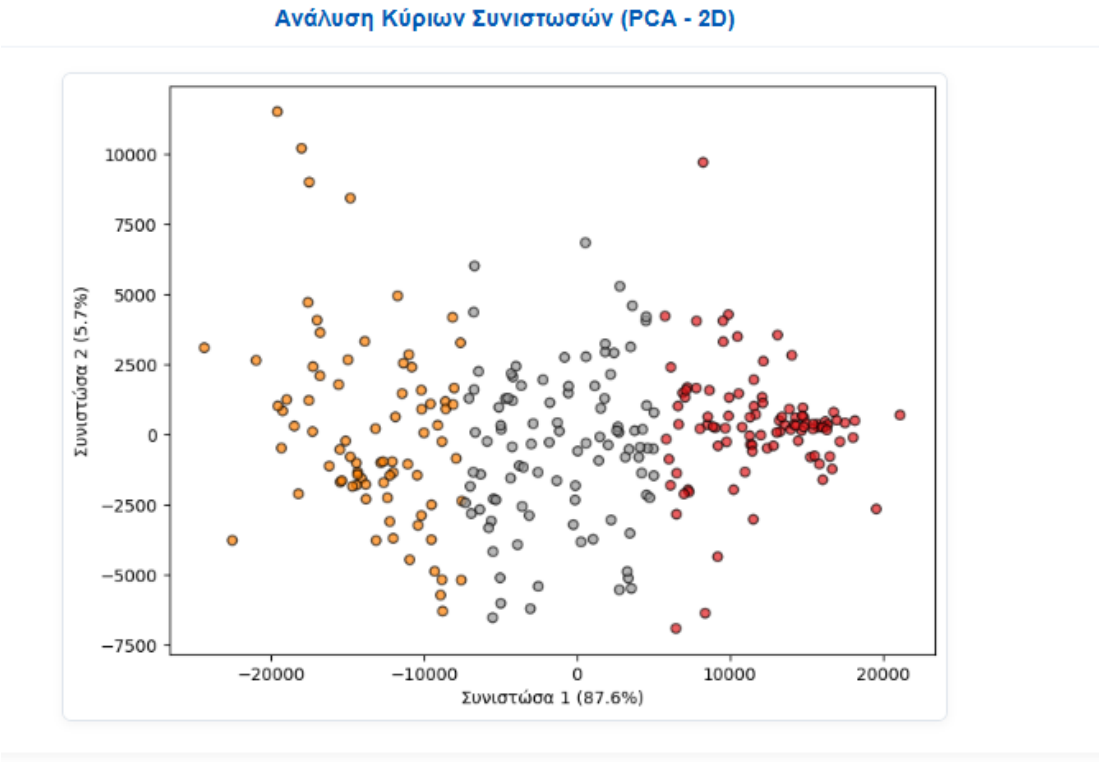
Συγκεκριμένα, δημιουργήθηκαν τέσσερα διακριτά σύνολα δεδομένων (datasets). Τα δύο πρώτα περιλαμβάνουν τις απόλυτες βάσεις εισαγωγής ανά έτος (από το 2013 μέχρι το 2025 και από το 2019 μέχρι το 2025 αντίστοιχα), δηλαδή τα πραγματικά μόρια που απαιτήθηκαν για την εισαγωγή σε κάθε πανεπιστημιακό τμήμα. Αυτά τα σύνολα μας επιτρέπουν να εξετάσουμε το γενικό επίπεδο δυσκολίας και τη σχετική «βαθμολογική θέση» κάθε σχολής στο σύστημα. Τα άλλα δύο datasets περιέχουν τις ετήσιες μεταβολές των βάσεων (από το 2013 μέχρι το 2025 και από το 2019 μέχρι το 2025 αντίστοιχα), υπολογίζοντας τη διαφορά από χρονιά σε χρονιά ($\text{Βάση}_t - \text{Βάση}_{t-1}$), ώστε να αποτυπωθεί η δυναμική συμπεριφορά των τμημάτων. Με αυτόν τον τρόπο μπορούμε να εντοπίσουμε κοινές τάσεις, αντιδράσεις σε αλλαγές του συστήματος και μοτίβα μεταβλητότητας. Ο συνδυασμός των δύο προσεγγίσεων, στατικής και δυναμικής, προσφέρει μια ολοκληρωμένη εικόνα της εξέλιξης των βάσεων εισαγωγής.

7.2 Συσταδοποίηση Τμημάτων με βάση εισαγωγής στο διάστημα 2013-2025 (k-means, hierarchical)

7.2.1 K-means

Κατά την πρώτη πειραματική διαδικασία, ο αλγόριθμος K-Means εφαρμόστηκε στις απόλυτες τιμές των βάσεων εισαγωγής για τη δωδεκαετία 2013-2025, με τον αριθμό των συστάδων να ορίζεται σε k=3. Η ποιότητα του προκύπτοντος διαχωρισμού αξιολογήθηκε μέσω δύο βασικών μετρικών, **Silhouette Score:** 0.419 και **Sum of Squared Errors (SSE):** 8.526.003.302. Η τιμή του δείκτη Silhouette κρίνεται ικανοποιητική για δεδομένα τέτοιου είδους, υποδηλώνοντας ότι οι συστάδες διαθέτουν επαρκή βαθμό συνοχής και είναι διακριτές μεταξύ τους. Ακολουθεί λεπτομερής ανάλυση και ερμηνεία των αποτελεσμάτων μέσα από τα επιμέρους γραφήματα της μελέτης.

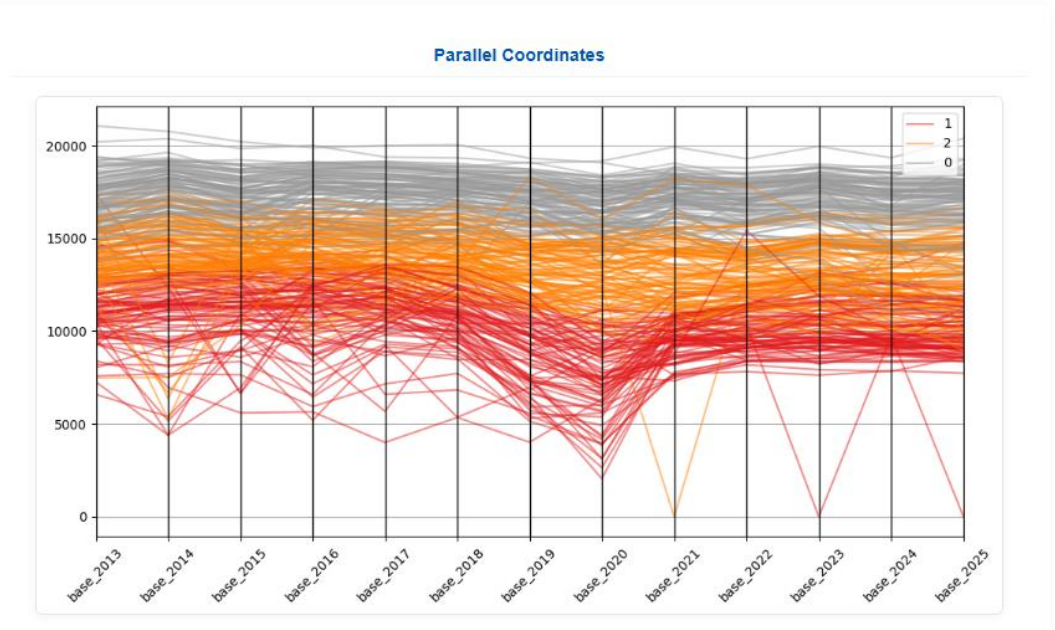
Για τον έλεγχο της ομοιογένειας και της οπτικής επαλήθευσης των παραγόμενων ομάδων, πραγματοποιήθηκε προβολή των πολυδιάστατων δεδομένων στο δισδιάστατο επίπεδο με τη χρήση της τεχνικής PCA.



Εικόνα 14.Διάγραμμα διασποράς των συστάδων του K-Means στο επίπεδο των δύο πρώτων κύριων συνιστωσών (PCA) για την περίοδο 2013-2025.

Όπως προκύπτει από την Εικόνα 14, οι δύο πρώτες κύριες συνιστώσες κατορθώνουν να συγκρατήσουν το 93.3% της συνολικής διακύμανσης του δείγματος. Στο διάγραμμα διασποράς παρατηρείται ένας σαφής οριζόντιος διαχωρισμός των τριών ομάδων κατά μήκος του άξονα της πρώτης συνιστώσας. Στα αριστερά εντοπίζεται το Cluster 2 (πορτοκαλί χρώμα), στο κέντρο αναπτύσσεται το Cluster 0 (γκρι χρώμα) και στα δεξιά διαμορφώνεται το Cluster 1 (κόκκινο χρώμα). Η απουσία έντονων επικαλύψεων μεταξύ των σημείων επιβεβαιώνει γεωμετρικά ότι ο αλγόριθμος εντόπισε τρία διακριτά βαθμολογικά προφίλ σχολών.

Η μελέτη της διαχρονικής πορείας των βάσεων εισαγωγής για κάθε τμήμα ανά ομάδα πραγματοποιήθηκε με τη χρήση διαγράμματος παράλληλων συντεταγμένων, όπου κάθε έτος αποτελεί έναν διακριτό κάθετο άξονα.



Εικόνα 15.Διάγραμμα παράλληλων συντεταγμένων των βάσεων εισαγωγής ανά συστάδα για την περίοδο 2013-2025.

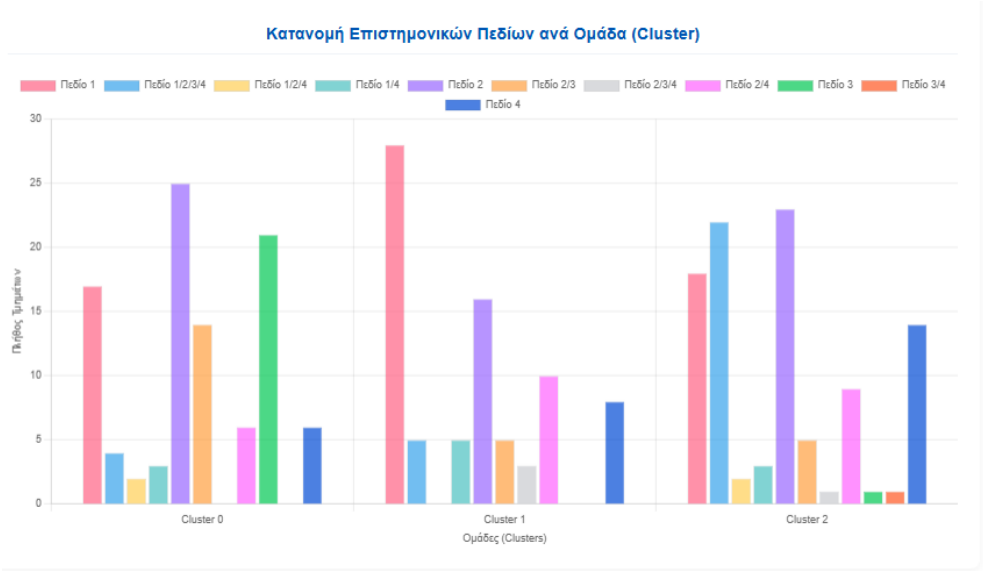
Σύμφωνα με την Εικόνα 15, οι τροχιές των τμημάτων οργανώνονται σε τρία επίπεδα, τα οποία αντιστοιχούν στο βαθμολογικό μέγεθος των σχολών:

- **Cluster 0 (Γκρι γραμμές):** Αντιπροσωπεύει την υψηλόβαθμη «ελίτ» των σχολών, με τις βάσεις να κινούνται σταθερά καθ' όλη τη διάρκεια της δωδεκαετίας πάνω από το όριο των 15.000 μορίων, αγγίζοντας σε πολλές περιπτώσεις τα 20.000 μόρια.

- **Cluster 2 (Πορτοκαλί γραμμές):** Περιλαμβάνει τα τμήματα μεσαίας βαθμολογικής ζώνης, τα οποία συγκεντρώνονται κυρίως μεταξύ 10.000 και 15.000 μορίων, επιδεικνύοντας μια σχετικά σταθερή και παράλληλη πορεία.

- **Cluster 1 (Κόκκινες γραμμές):** Απεικονίζει τα χαμηλόβαθμα τμήματα, οι βάσεις των οποίων βρίσκονται κάτω από τα 10.000 μόρια. Η ομάδα αυτή παρουσιάζει τη μεγαλύτερη εσωτερική διασπορά, καθώς και μια έντονα πτωτική τάση που κορυφώνεται γύρω στο έτος 2020, γεγονός που αποτυπώνει την επίδραση των αυξομειώσεων στη δυσκολία των θεμάτων στις σχολές χαμηλής ζήτησης. Σημειώνεται, επίσης, η ύπαρξη ελάχιστων τμημάτων-outliers που καταγράφουν μηδενικές τιμές σε συγκεκριμένα έτη, κάτι που σχετίζεται πιθανώς με αναστολές λειτουργίας ή μηδενική κάλυψη θέσεων.

Για την κατανόηση της επιστημονικής ταυτότητας των συστάδων, αναλύθηκε η συχνότητα εμφάνισης των επιστημονικών πεδίων στο εσωτερικό της κάθε ομάδας.



Εικόνα 16. Ραβδόγραμμα κατανομής των Επιστημονικών Πεδίων στο εσωτερικό των τριών συστάδων.

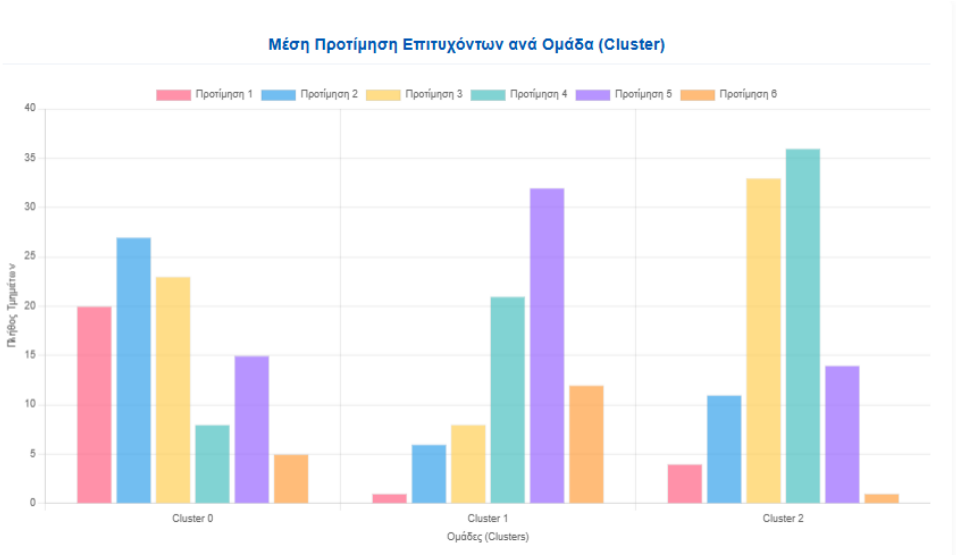
Από το ραβδόγραμμα της Εικόνας 16, προκύπτει μια ξεκάθαρη διαφοροποίηση στη σύνθεση των ομάδων:

- Στο **Cluster 0** (υψηλόβαθμα τμήματα) παρατηρείται κυριαρχία του Πεδίου 2 (Θετικές και Τεχνολογικές Επιστήμες) και του Πεδίου 3 (Επιστήμες Υγείας και Ζωής), επιβεβαιώνοντας ότι οι παραδοσιακά υψηλόβαθμες ιατρικές και πολυτεχνικές σχολές συγκεντρώνονται εδώ.

- Στο **Cluster 1** (χαμηλόβαθμα τμήματα) παρουσιάζεται έντονη συγκέντρωση των ανθρωπιστικών επιστημών (Πεδίο 1), γεγονός που συνάδει με τη γενικευμένη πτώση των βάσεων που παρατηρείται στα τμήματα αυτά τα τελευταία χρόνια, καθώς και σχολών του Πεδίου 2 με περιορισμένη γεωγραφική ζήτηση.

- Στο **Cluster 2** (μεσαίας βαθμολογίας τμήματα) καταγράφεται μια πιο ισορροπημένη κατανομή, με ισχυρή όμως παρουσία του Πεδίου 4 (Επιστήμες Οικονομίας και Πληροφορικής) και του Πεδίου 2, αποτυπώνοντας τα τμήματα της μεσαίας ζώνης των κεντρικών και περιφερειακών ιδρυμάτων.

Η δυναμική των συστάδων συμπληρώνεται από την ανάλυση της μέσης σειράς προτίμησης των επιτυχόντων, η οποία αναδεικνύει τον βαθμό ελκυστικότητας των σχολών.

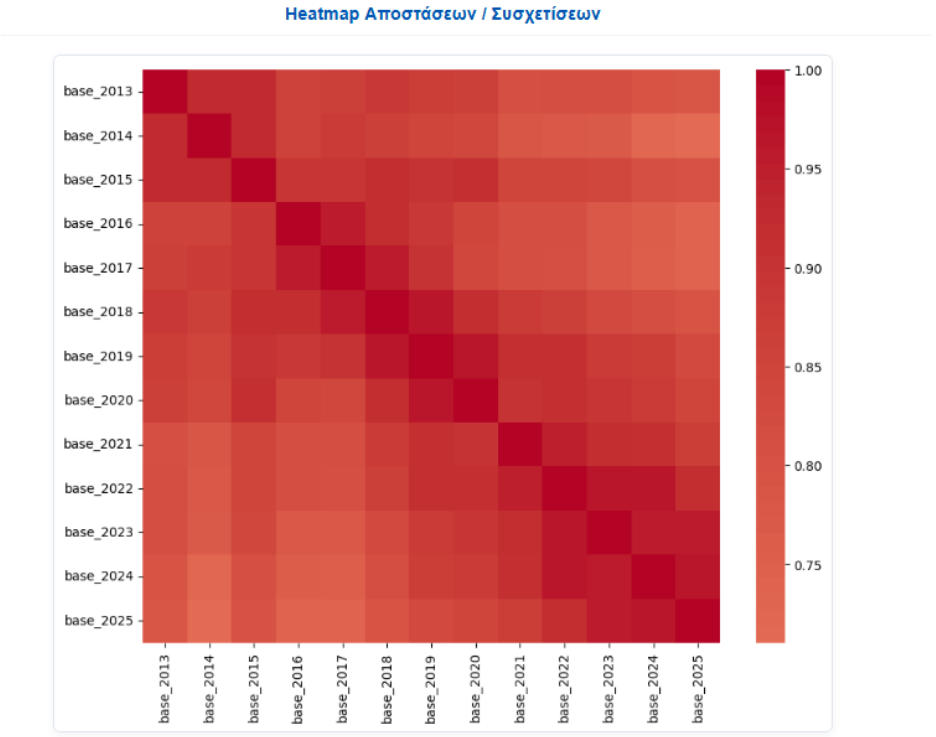


Εικόνα 17. Ραβδόγραμμα συχνότητας της σειράς προτίμησης των επιτυχόντων ανά συστάδα.

Η ανάλυση του ραβδογράμματος της Εικόνας 17, φέρνει στην επιφάνεια πολύ ενδιαφέροντα στοιχεία για την ψυχολογία των υποψηφίων:

- Στο **Cluster 0** (υψηλόβαθμα τμήματα), η πλειονότητα των επιτυχόντων εισάγεται έχοντας τη σχολή στη 2η ή στην 3η προτίμησή τους, ενώ αξιοσημείωτο είναι και το ποσοστό της 1ης προτίμησης. Το στοιχείο αυτό δείχνει ότι τα τμήματα αυτά αποτελούν πρωταρχικούς στόχους.
- Στο **Cluster 1** (χαμηλόβαθμα τμήματα), η κατανομή μετατοπίζεται ξεκάθαρα προς τα δεξιά. Οι περισσότεροι υποψήφιοι εισάγονται σε αυτά τα τμήματα έχοντάς τα ως 5η ή 6η επιλογή στο μηχανογραφικό τους, επιβεβαιώνοντας ότι πρόκειται για σχολές «δεύτερης ευκαιρίας» ή λύσεις ανάγκης.
- Στο **Cluster 2** (μεσαία τμήματα), η συντριπτική πλειονότητα των εισακτέων είχε τα τμήματα αυτά ως 3η ή 4η επιλογή, αποτυπώνοντας τον ρόλο των μεσαίων σχολών ως ενδιάμεσων επιλογών ασφαλείας.

Η στατική εικόνα των βάσεων ολοκληρώνεται με τη μελέτη της γραμμικής συσχέτισης των βάσεων εισαγωγής μεταξύ των ετών, η οποία αποτυπώνεται σε έναν χάρτη θερμότητας (Heatmap).



Εικόνα 18.Heatmap kmeans των βάσεων εισαγωγής μεταξύ των ετών 2013-2025.

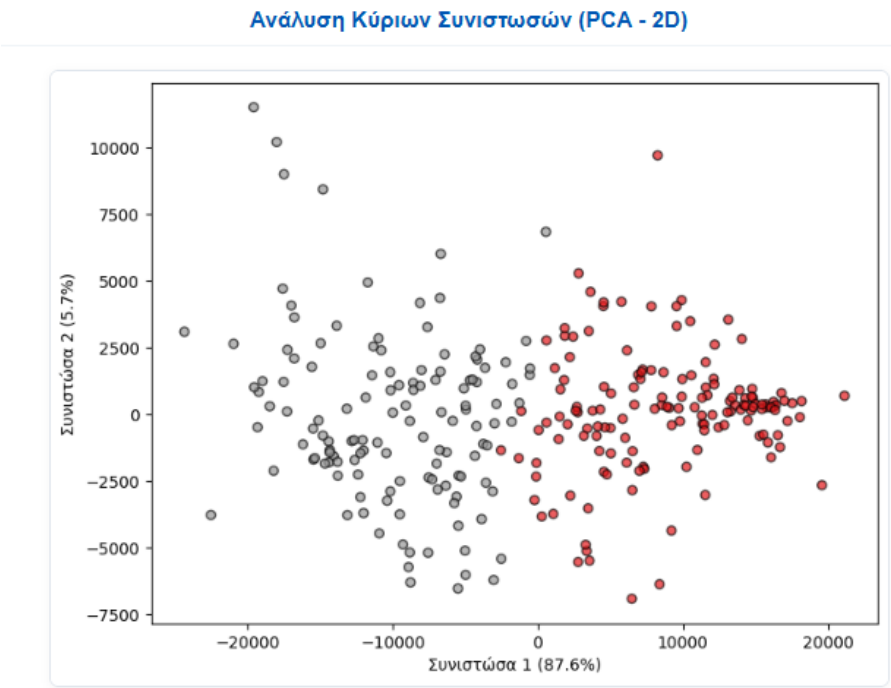
Ο χάρτης θερμότητας της Εικόνας 18, εμφανίζει μια εξαιρετικά συμπαγή δομή. Όλες οι τιμές των συντελεστών συσχέτισης κινούνται σε πολύ υψηλά επίπεδα (άνω του 0.70), γεγονός που υποδηλώνει ότι η γενική ιεραρχική δομή των βάσεων εισαγωγής στην Ελλάδα παραμένει εξαιρετικά σταθερή στο χρόνο (οι υψηλόβαθμες σχολές παραμένουν υψηλόβαθμες και οι χαμηλόβαθμες αντίστοιχα χαμηλόβαθμες).

Η συσχέτιση είναι σχεδόν απόλυτη (κοντά στο 0.95 με 1.0) κατά μήκος της κυρίας διαγωνίου, δηλαδή μεταξύ διαδοχικών ετών (π.χ. 2013 με 2014). Παρατηρείται, ωστόσο, μια πολύ ομαλή και σταδιακή εξασθένηση των χρωμάτων (μείωση του συντελεστή προς το 0.75) καθώς απομακρυνόμαστε από τη διαγώνιο, με τις χαμηλότερες συσχετίσεις να εντοπίζονται στη σύγκριση των ετών 2013-2014 με τα έτη 2024-2025. Η προοδευτική αυτή μείωση αποτυπώνει τη σταδιακή, μακροπρόθεσμη αλλαγή των προτιμήσεων της κοινωνίας και των υποψηφίων σε ορίζοντα δεκαετίας.

7.2.2 Hierarchical

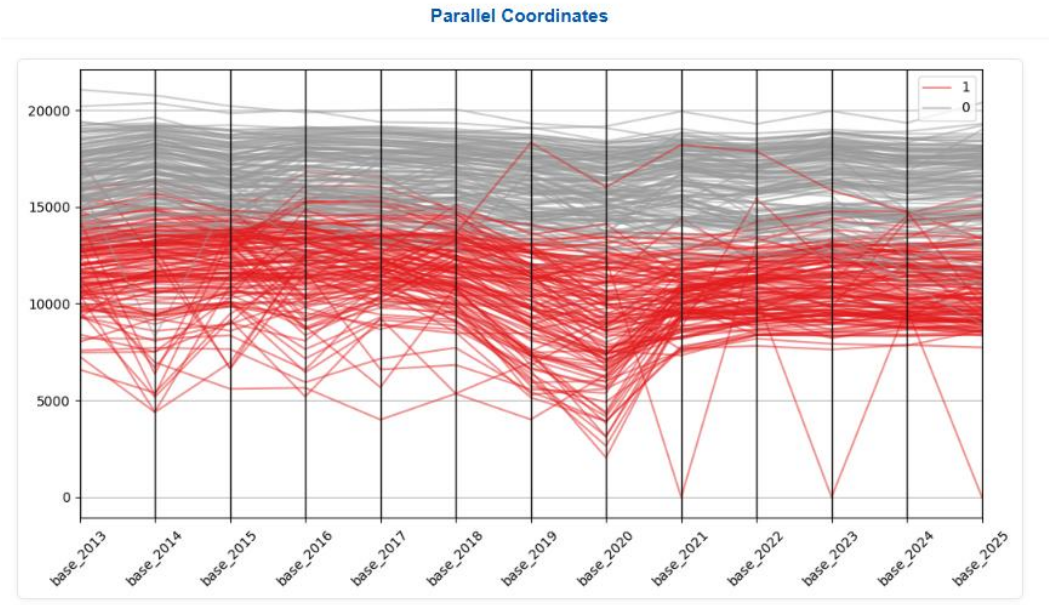
Για την επαλήθευση της σταθερότητας των παραπάνω αποτελεσμάτων, το ίδιο σύνολο δεδομένων (απόλυτες τιμές 2013-2025) υποβλήθηκε σε Ιεραρχική Συσταδοποίηση, εξετάζοντας τέσσερις διαφορετικές μεθόδους διασύνδεσης (Ward, Complete, Average, Single). Από τη σύγκριση των ιεραρχικών αλγορίθμων φάνηκε ότι η μέθοδος **Ward** είναι η πιο αξιόπιστη για τις απόλυτες βάσεις της περιόδου 2013-2025 (με Silhouette Score 0.528 και SSE 12.831.884.288). Ο αλγόριθμος χώρισε τα δεδομένα με τον καλύτερο τρόπο σε 2 μεγάλες ομάδες (k=2). Αντί να σπάσει τις σχολές σε πολλά μικρά επίπεδα, τις χώρισε σε δύο βασικούς «πόλους» με ξεκάθαρα χαρακτηριστικά.

Για να κατανοήσουμε καλύτερα αυτές τις δύο ομάδες, ακολουθεί η ανάλυση των γραφημάτων τους. Ο διαχωρισμός των δύο ομάδων φαίνεται καθαρά στο διάγραμμα της ανάλυσης κυρίων συνιστωσών (PCA).



Εικόνα 19.Διάγραμμα διασποράς των 2 συστάδων της μεθόδου Ward (PCA - 2D) για την περίοδο 2013-2025.

Στην Εικόνα 19, βλέπουμε ότι οι δύο ομάδες ξεχωρίζουν απόλυτα. Στην αριστερή πλευρά βρίσκεται το Cluster 0 (Γκρι χρώμα) και στη δεξιά πλευρά το Cluster 1 (Κόκκινο χρώμα). Το γεγονός ότι οι κουκκίδες των δύο ομάδων σχεδόν δεν ανακατεύονται μεταξύ τους στο κέντρο, αποδεικνύει ότι τα δύο αυτά σύνολα έχουν τελείως διαφορετικές βαθμολογίες. Η πορεία των βάσεων για αυτές τις δύο ομάδες μέσα στα χρόνια φαίνεται στο διάγραμμα παράλληλων συντεταγμένων.

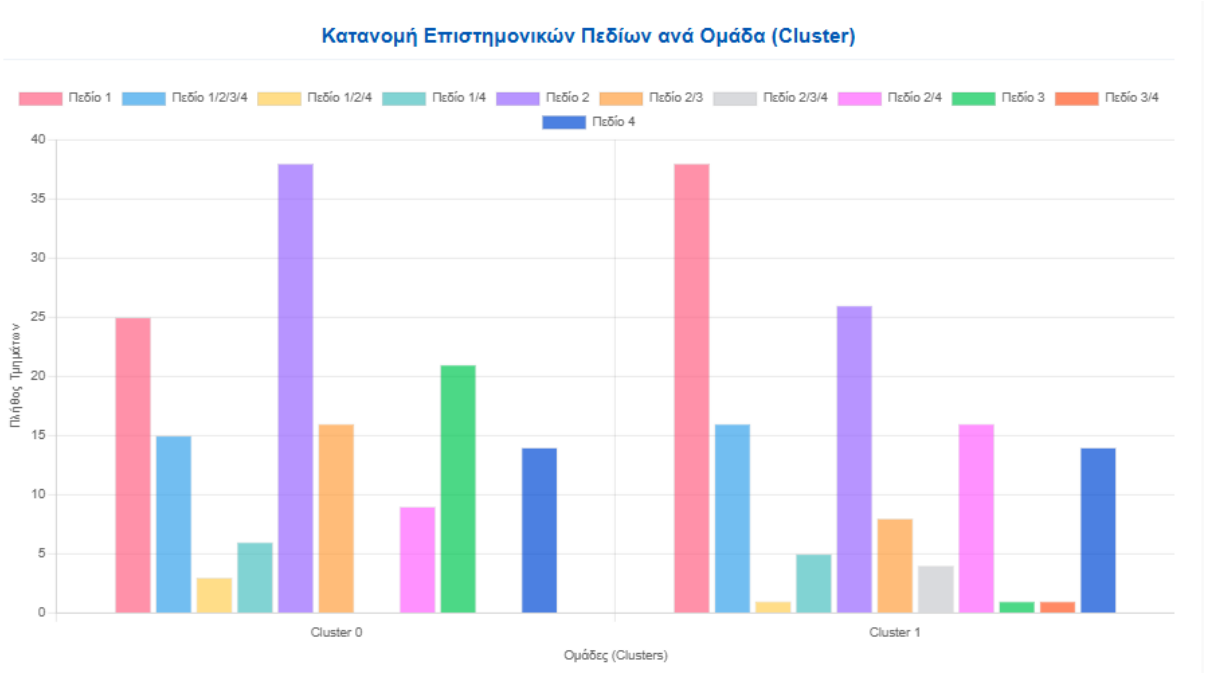


Εικόνα 20.Διάγραμμα παράλληλων συντεταγμένων των βάσεων εισαγωγής (Ward, k=2) για την περίοδο 2013-2025.

Η Εικόνα 20, δείχνει ξεκάθαρα τις διαφορές τους:

- **Cluster 0 (Γκρι γραμμές - Υψηλόβαθμες Σχολές):** Περιλαμβάνει τα τμήματα που έχουν σταθερά πάνω από 13.000 με 14.000 μόρια. Η πορεία τους είναι σχεδόν ευθεία, δείχνοντας ότι οι βάσεις τους δεν επηρεάζονται εύκολα από το πόσο δύσκολα είναι τα θέματα των Πανελλαδικών.
- **Cluster 1 (Κόκκινες γραμμές - Μεσαίες/Χαμηλόβαθμες Σχολές):** Περιλαμβάνει όλες τις σχολές κάτω από τα 14.000 μόρια. Εδώ υπάρχει μεγάλη αυξομείωση στις βάσεις. Φαίνονται καθαρά οι μεγάλες πτώσεις (όπως το 2020), καθώς αυτά τα τμήματα επηρεάζονται πολύ έντονα από τις αλλαγές στο σύστημα και τις επιδόσεις των μαθητών.

Για να δούμε ποιες σχολές ανήκουν σε κάθε ομάδα, αναλύουμε την κατανομή των επιστημονικών πεδίων.

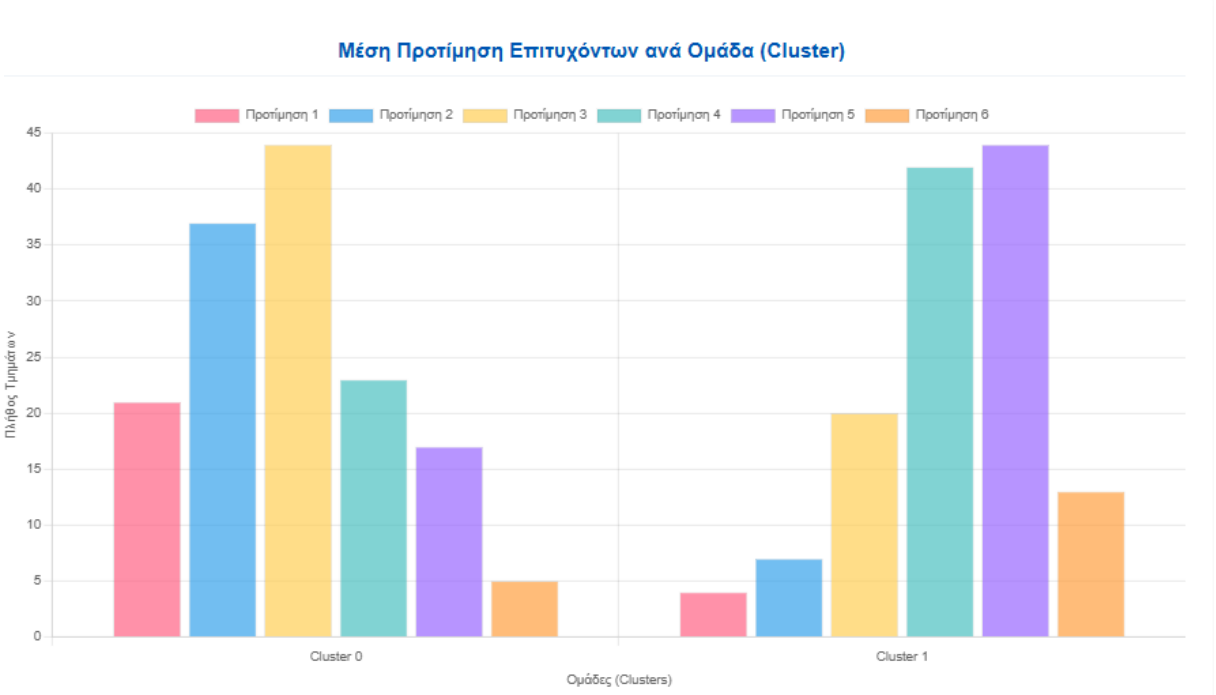


Εικόνα 21.Ραβδόγραμμα κατανομής των Επιστημονικών Πεδίων στις 2 συστάδες της μεθόδου Ward.

Από το ραβδόγραμμα της Εικόνας 21, προκύπτουν τα εξής:

- Στο **Cluster 0** (υψηλόβαθμες) κυριαρχούν το 2ο Πεδίο (Θετικές Επιστήμες) και το 3ο Πεδίο (Επιστήμες Υγείας). Εδώ βρίσκονται συνήθως οι ιατρικές και οι κεντρικές πολυτεχνικές σχολές.
- Στο **Cluster 1** (μεσαίες/χαμηλόβαθμες) κυριαρχεί το 1ο Πεδίο (Ανθρωπιστικές Σπουδές), κάτι που επιβεβαιώνει τη γενική πτώση των βάσεων στις θεωρητικές σχολές. Επίσης, υπάρχουν πολλά τμήματα του 4ου Πεδίου (Οικονομία/Πληροφορική) και περιφερειακά τμήματα του 2ου Πεδίου.

Ο τρόπος που επιλέγουν οι μαθητές αυτές τις σχολές μάς δείχνει τη δυναμική της κάθε ομάδας στο επόμενο ραβδόγραμμα.



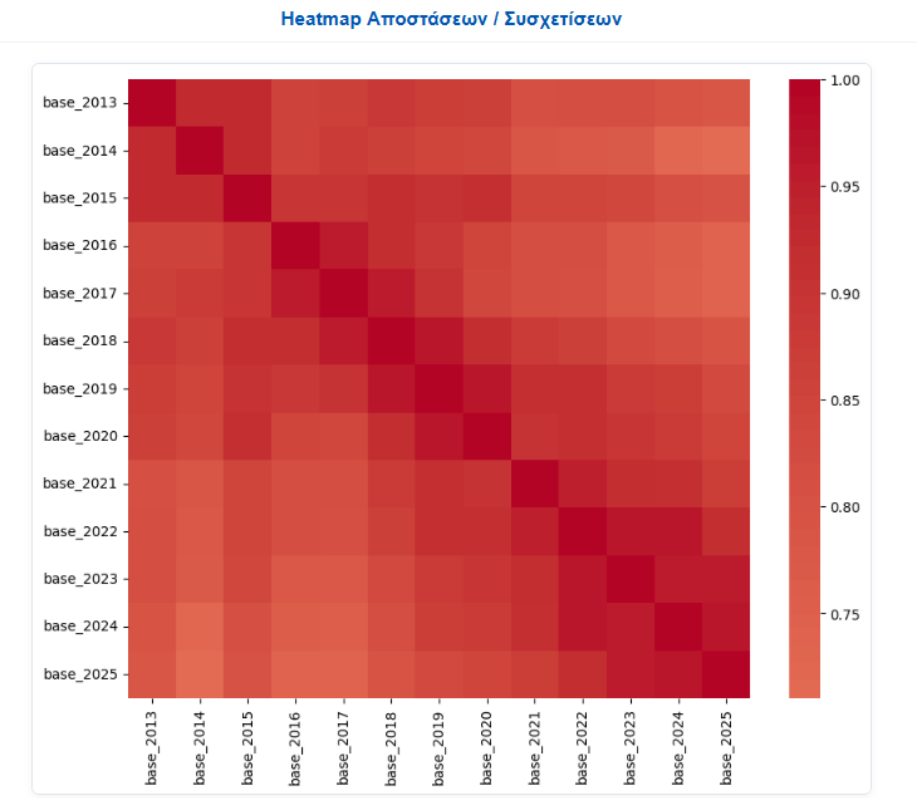
Εικόνα 22.Ραβδόγραμμα συχνότητας της σειράς προτίμησης των επιτυχόντων ανά συστάδα.

Η Εικόνα 22, δείχνει δύο διαφορετικούς τρόπους συμπλήρωσης του μηχανογραφικού:

- Στο **Cluster 0**, οι περισσότεροι μαθητές μπαίνουν έχοντας τη σχολή ως 1η, 2η ή 3η επιλογή. Αυτό σημαίνει ότι αυτές οι σχολές αποτελούν τον βασικό τους στόχο.

- Αντίθετα, στο **Cluster 1**, οι υψηλές προτιμήσεις είναι ελάχιστες. Οι περισσότεροι υποψήφιοι περνούν σε αυτές τις σχολές έχοντάς τες ως 4η, 5η όπου είναι και η κορυφή ή 6η επιλογή. Αυτό δείχνει ότι τα συγκεκριμένα τμήματα επιλέγονται συνήθως ως εναλλακτικές λύσεις δηλαδή λύσεις ασφαλείας για να εξασφαλίσουν την είσοδό τους στο πανεπιστήμιο.

Η ανάλυση ολοκληρώνεται με τη μελέτη των συσχετίσεων μεταξύ των ετών, που μας βοηθά να καταλάβουμε τη σταθερότητα του συστήματος.



Εικόνα 23.Heatmap hierarchical των βάσεων εισαγωγής μεταξύ των ετών 2013-2025.

Όπως παρατηρούμε στην Εικόνα 23, στο διάγραμμα κυριαρχεί το έντονο κόκκινο χρώμα. Οι συντελεστές συσχέτισης είναι εξαιρετικά υψηλοί σε όλη τη δωδεκαετία, καθώς κινούνται σε ένα πολύ στενό εύρος από **0.75 έως 1.00**. Αυτή η καθολική συσχέτιση αποδεικνύει ότι η ιεραρχία των σχολών στην Ελλάδα παραμένει σχεδόν αναλλοίωτη στο χρόνο. Οι σχολές που ήταν υψηλόβαθμες το 2013 (π.χ. Ιατρικές, Πολυτεχνικές) διατηρούν την ίδια ακριβώς πλεονεκτική θέση και το 2025, παρά τις όποιες αλλαγές έγιναν στα εξεταστικά συστήματα. Αυτή η ισχυρή μακροχρόνια σταθερότητα των δεδομένων είναι ο λόγος που επέτρεψε στη μέθοδο Ward να χαράξει ένα τόσο καθαρό γεωμετρικό όριο και να χωρίσει με απόλυτη επιτυχία τις σχολές σε δύο σταθερούς πόλους.

Συμπερασματικά για την ιεραρχική ανάλυση των απόλυτων βάσεων της δωδεκαετίας, η μέθοδος **Ward** επιλέχθηκε ως η βέλτιστη και πιο αξιόπιστη προσέγγιση. Σε αντίθεση με τις μεθόδους Complete, Average και κυρίως τη Single Linkage (η οποία εμφάνισε έντονα το φαινόμενο της αλυσιδωτής συγχώνευσης, εκτινάσσοντας το SSE στα 36.239.955.015), η μέθοδος Ward ελαχιστοποιεί την εσωτερική διακύμανση των ομάδων. Αυτός ο μηχανισμός επέτρεψε στον αλγόριθμο να φιλτράρει τον θόρυβο και να χαράξει ένα καθαρό γεωμετρικό όριο (Silhouette 0.528, SSE 12.831.884.288), χωρίζοντας με απόλυτη ερμηνευτική σαφήνεια τις σχολές στους δύο βασικούς βαθμολογικούς πόλους της χώρας.

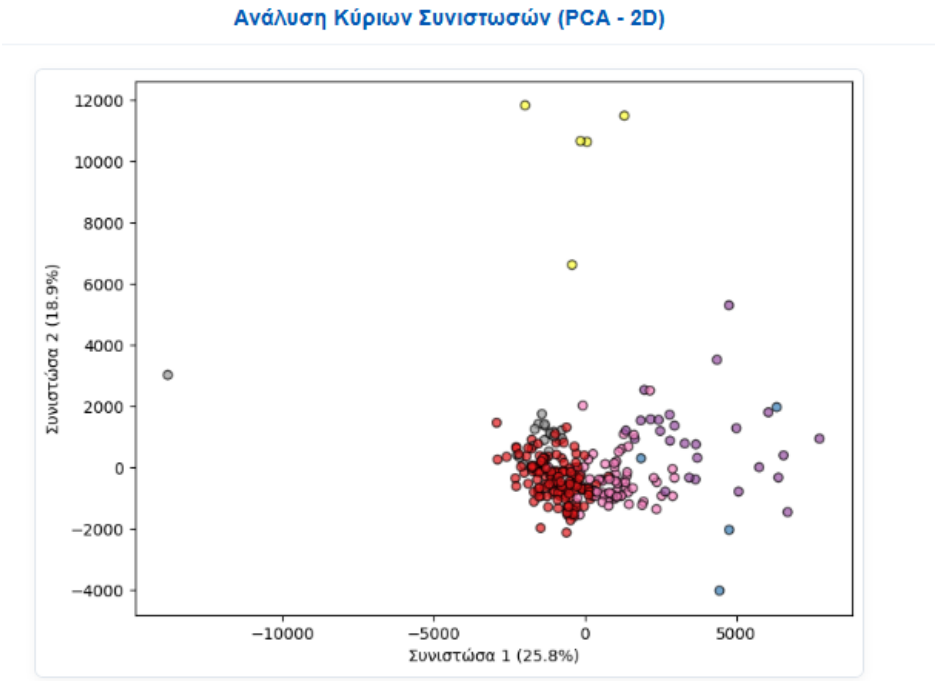
7.3 Συσταδοποίηση Τμημάτων βάσει κοινής συμπεριφοράς στο διάστημα 2013-2025 (k-means, hierarchical)

7.3.1 K-means

Σε αυτή τη φάση της μελέτης, το ενδιαφέρον μετατοπίζεται από τα απόλυτα βαθμολογικά επίπεδα στη δυναμική συμπεριφορά των τμημάτων. Τα δεδομένα μετασχηματίστηκαν κατάλληλα ώστε να αποτυπώνουν τις ετήσιες διαφορές των βάσεων εισαγωγής ($Βάση_{τ}$ – $Βάση_{τ-1}$). Σκοπός είναι ο εντοπισμός κοινών τάσεων και ομαδοποιήσεων με κριτήριο το πώς αντιδρούν οι σχολές στο χρόνο (π.χ. ταυτόχρονη άνοδος ή πτώση).

Για το σκοπό αυτό, ο αλγόριθμος K-Means εφαρμόστηκε με επιλογή k=6 συστάδων, αποδίδοντας τις εξής μετρικές αξιολόγησης, **Silhouette Score:** 0.237 και **Sum of Squared Errors (SSE):** 2.281.012.214. Η εμφανής μείωση του δείκτη Silhouette σε σχέση με την ανάλυση των απόλυτων τιμών κρίνεται μαθηματικά αναμενόμενη. Οι ετήσιες μεταβολές χαρακτηρίζονται από έντονο θόρυβο, καθώς επηρεάζονται άμεσα από εξωτερικούς και αστάθμητους παράγοντες, όπως η αυξομειούμενη δυσκολία των θεμάτων στις Πανελλαδικές εξετάσεις, με αποτέλεσμα οι συστάδες να παρουσιάζουν μεγαλύτερη γεωμετρική διασπορά.

Η γεωμετρική δομή των συστάδων των ετήσιων μεταβολών αποτυπώνεται στο δισδιάστατο επίπεδο με τη χρήση της τεχνικής PCA.

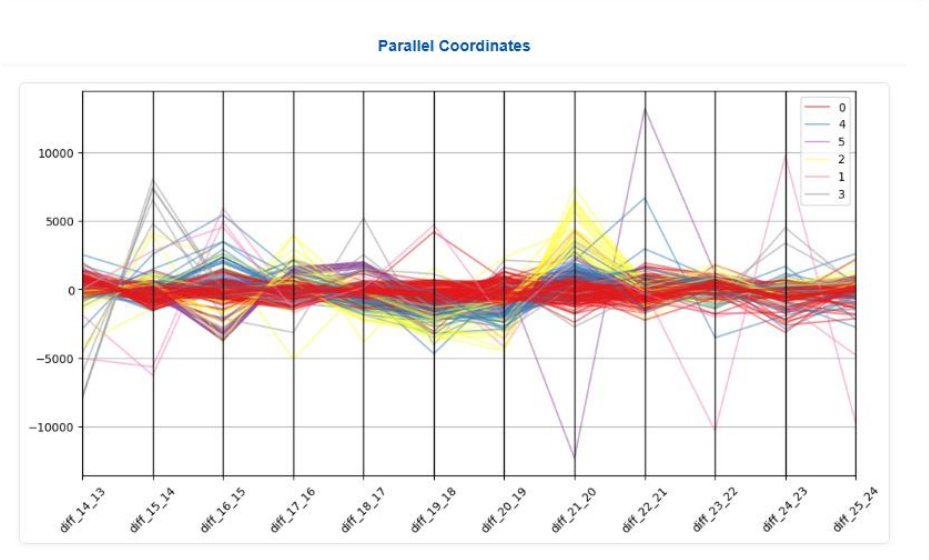


Εικόνα 24.Διάγραμμα διασποράς των συστάδων του K-Means στο επίπεδο των δύο πρώτων κύριων συνιστωσών (PCA) για τις ετήσιες μεταβολές (2013-2025).

Από την Εικόνα 24, παρατηρείται ότι οι δύο πρώτες συνιστώσες ερμηνεύουν συνδυαστικά το 47.3% της συνολικής διακύμανσης. Η χαμηλότερη αυτή τιμή (σε σχέση με το 93.3% των απόλυτων τιμών) αναδεικνύει την πολυδιάστατη και σύνθετη φύση των χρονικών μεταβολών. Υπάρχει μια εξαιρετικά πυκνή, κεντρική μάζα σημείων (κόκκινο και έντονο ροζ χρώμα) που συγκεντρώνεται γύρω από το μηδέν, αναδεικνύοντας ότι η συντριπτική πλειονότητα των σχολών ακολουθεί μια σταθερή, μέση

συμπεριφορά. Περιφερειακά αυτής της μάζας, ο αλγόριθμος δημιουργεί μικρά, αραιά και εξειδικευμένα clusters. Για παράδειγμα, στο ανώτερο τμήμα του άξονα της δεύτερης συνιστώσας απομονώνονται ελάχιστα κίτρινα σημεία, ενώ στα δεξιά αναπτύσσονται μωβ και γαλάζιες συστάδες. Η δομή αυτή επιβεβαιώνει ότι ο K-Means με k=6 λειτούργησε αποτελεσματικά ως μηχανισμός εντοπισμού και απομόνωσης ακραίων συμπεριφορών (outliers).

Η χρονική εξέλιξη των ετήσιων αυξομειώσεων για κάθε μία από τις 6 ομάδες αποτυπώνεται στο διάγραμμα παράλληλων συντεταγμένων.

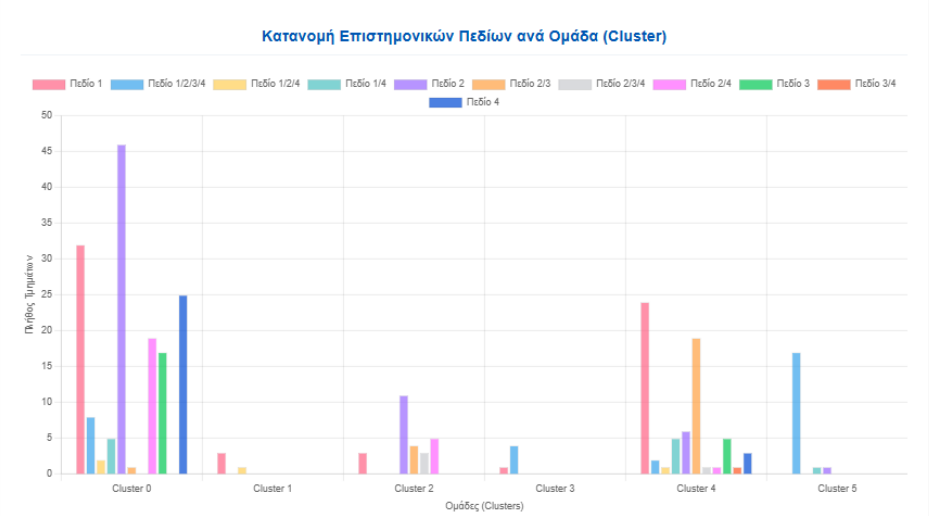


Εικόνα 25.Διάγραμμα παράλληλων συντεταγμένων των ετήσιων μεταβολών ανά συστάδα για την περίοδο 2013-2025.

Το διάγραμμα της Εικόνας 25, επιτρέπει την ταυτοποίηση της χρονικής συμπεριφοράς των ομάδων:

- **Cluster 0 (Κόκκινες γραμμές):** Αποτελεί τον βασικό πυρήνα του δείγματος. Οι γραμμές είναι σχεδόν απόλυτα συγκεντρωμένες γύρω από τον άξονα του μηδενός, υποδηλώνοντας τμήματα με μηδενικές ή εξαιρετικά ήπιες ετήσιες μεταβολές.
- **Υπόλοιπα Clusters (1, 2, 3, 4, 5):** Αντιπροσωπεύουν σχολές με έντονη μεταβλητότητα και μεγάλες αποκλίσεις. Παρατηρείται ότι κάθε ομάδα απομονώνει συγκεκριμένα χρονικά γεγονότα. Για παράδειγμα, η μωβ γραμμή (Cluster 5) καταγράφει μια κατακόρυφη άνοδο άνω των 10.000 μορίων στο διάστημα diff_22_21 και μια αντίστοιχη πτώση στο diff_21_20. Αντίστοιχα, οι ροζ γραμμές εμφανίζουν ακραίες βυθίσεις στο diff_23_22 και στο diff_25_24. Οι συμπεριφορές αυτές αποτυπώνουν τμήματα που αντέδρασαν με ακραίο τρόπο σε συστημικές αλλαγές, όπως η εισαγωγή της ΕΒΕ ή η ριζική αναδιάρθρωση των συντελεστών βαρύτητας.

Προκειμένου να διαπιστωθεί αν οι 6 συμπεριφορές συνδέονται με συγκεκριμένα γνωστικά αντικείμενα, εξετάστηκε η κατανομή των επιστημονικών πεδίων στο εσωτερικό τους.

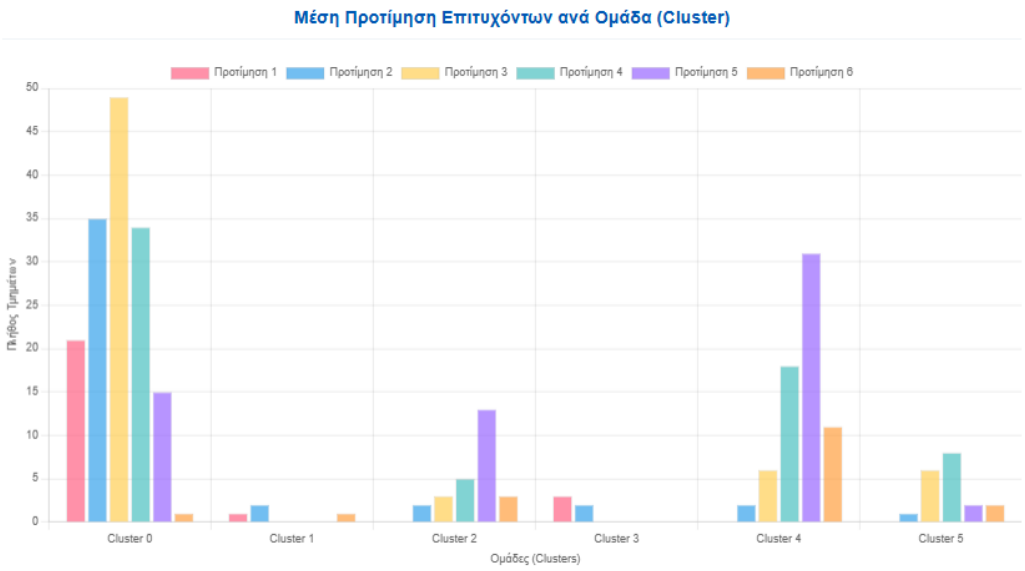


Εικόνα 26.Ραβδόγραμμα κατανομής των Επιστημονικών Πεδίων στο εσωτερικό των 6 συστάδων των μεταβολών.

Η ανάλυση της Εικόνας 26, προσφέρει τα εξής σημαντικά συμπεράσματα:

- Στο **Cluster 0** (ήπιες μεταβολές) εντάσσεται ο μεγάλος όγκος των σχολών από όλα τα επιστημονικά πεδία, με κυρίαρχα το Πεδίο 1 (Ανθρωπιστικές), το Πεδίο 2 (Θετικές) και το Πεδίο 4 (Οικονομίας/Πληροφορικής). Αυτό δείχνει ότι η πλειονότητα των σχολών, ανεξαρτήτως αντικειμένου, διατηρεί μια σταθερή χρονική συμπεριφορά.
- Στα **Clusters 2, 4 και 5** παρατηρείται επιλεκτική συγκέντρωση σχολών. Στο Cluster 4, για παράδειγμα, εμφανίζεται έντονη παρουσία του Πεδίου 1 (Ανθρωπιστικές) και του Πεδίου 2/3. Στο Cluster 2 εντοπίζονται κυρίως σχολές του Πεδίου 2. Τα ραβδογράμματα αυτά αποδεικνύουν ότι οι ακραίες χρονικές αυξομειώσεις δεν είναι τυχαίες, αλλά πλήττουν συγκεκριμένες κατηγορίες σχολών (κυρίως περιφερειακά τμήματα των θετικών επιστημών και των ανθρωπιστικών σπουδών).

Η εξέταση της μέσης σειράς προτίμησης των επιτυχόντων αναδεικνύει τη σύνδεση μεταξύ της ελκυστικότητας μιας σχολής και της σταθερότητας της βάσης της.

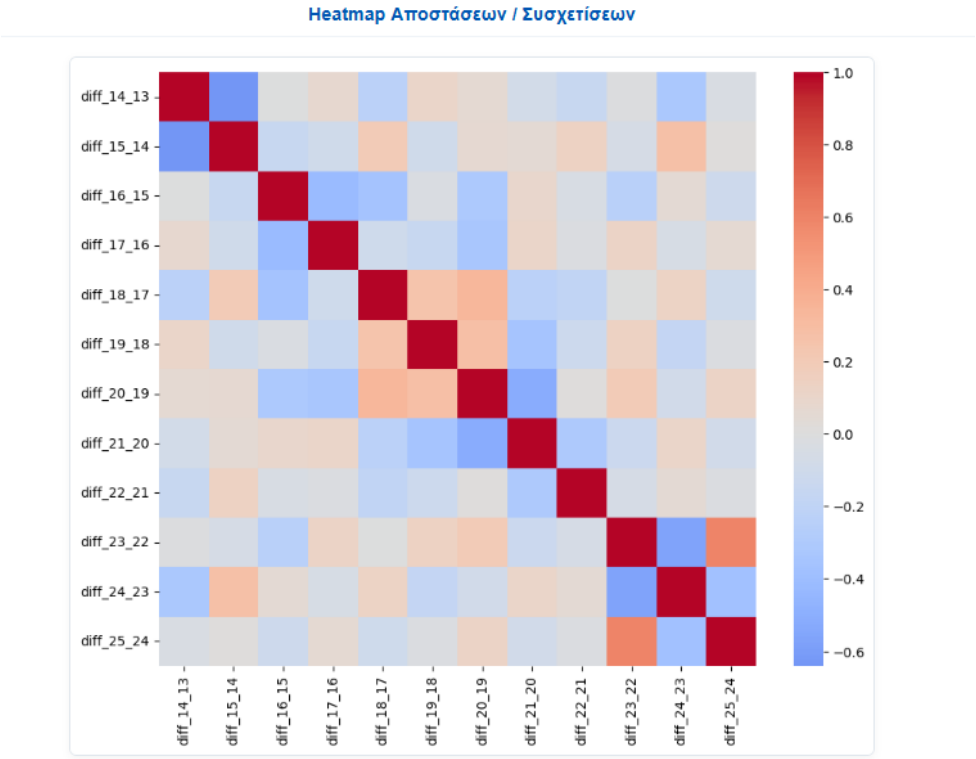


Εικόνα 27.Ραβδόγραμμα συχνότητας της σειράς προτίμησης των επιτυχόντων ανά συστάδα μεταβολών.

Σύμφωνα με την Εικόνα 27, καταγράφεται μια σαφέστατη μεθοδολογική διαφοροποίηση:

- Στο **Cluster 0** (που περιλαμβάνει τη μάζα των σταθερών σχολών), η κατανομή των προτιτήσεων είναι ομαλή και κορυφώνεται στην 3η προτίμηση, με ισχυρή παρουσία της 1ης και 2ης επιλογής. Πρόκειται για σχολές με παγιοποιημένη ζήτηση.
- Αντίθετα, στα **Clusters 2, 4 και 5** (υψηλή μεταβλητότητα), η κατανομή μετατοπίζεται δραματικά προς τα δεξιά. Στο Cluster 4, για παράδειγμα, η συντριπτική πλειονότητα των εισακτέων είχε τα τμήματα αυτά ως 5η επιλογή στο μηχανογραφικό τους. Το στοιχείο αυτό επιβεβαιώνει ότι οι σχολές που εμφανίζουν τις πιο ασταθείς, απρόβλεπτες και ακραίες ετήσιες διακυμάνσεις είναι εκείνες που λειτουργούν ως επιλογές ανάγκης ή χαμηλής προτεραιότητας για τους υποψηφίους.

Η δυναμική ανάλυση ολοκληρώνεται με τον υπολογισμό των συντελεστών συσχέτισης μεταξύ των ετήσιων μεταβολών.



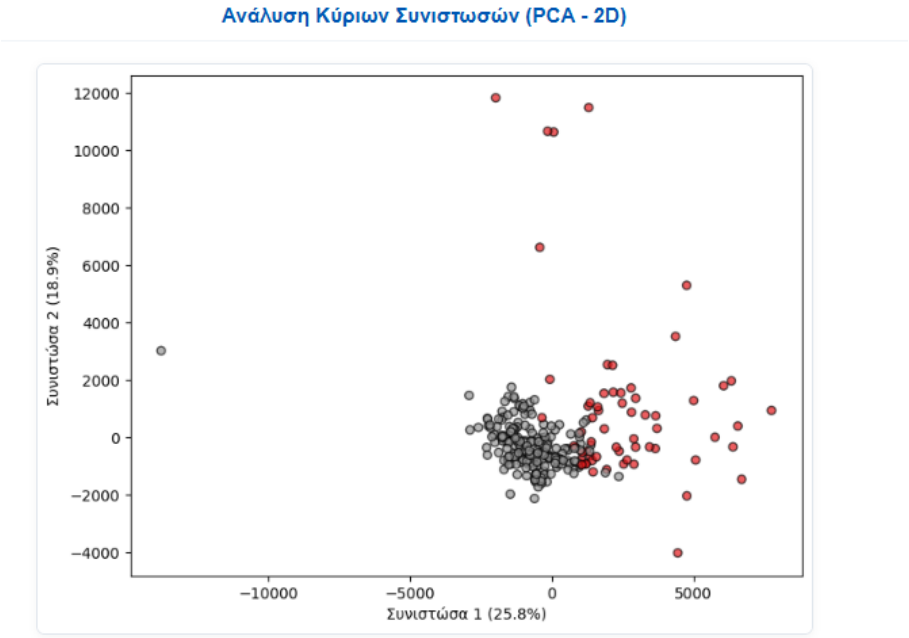
Εικόνα 28.Heatmap kmeans των ετήσιων μεταβολών μεταξύ των ετών 2013-2025.

Ο χάρτης θερμότητας της Εικόνας 28, παρουσιάζει μια ριζικά διαφορετική εικόνα σε σχέση με αυτόν των απόλυτων τιμών, καθώς κυριαρχούν οι χαμηλές θετικές, αλλά και οι αρνητικές τιμές συσχέτισης (οι συντελεστές κυμαίνονται κυρίως από -0.40 έως +0.20). Για παράδειγμα, η μεταβολή diff_15_14 εμφανίζει αρνητική συσχέτιση (μπλε χρώμα) με τη μεταβολή diff_14_13. Η γενικευμένη απουσία ισχυρής γραμμικής συσχέτισης αποδεικνύει ότι η μεταβολή των βάσεων μιας χρονιάς δεν αποτελεί οδηγό για την επόμενη. Η πορεία των ετήσιων μεταβολών είναι αυτόνομη και καθορίζεται σχεδόν αποκλειστικά από τις ετήσιες ιδιαιτερότητες (βαθμός δυσκολίας θεμάτων, αριθμός εισακτέων) της εκάστοτε εξεταστικής περιόδου, δικαιολογώντας την ανάγκη του K-Means με k=6 να επιστρατεύσει πολλές ομάδες για να ερμηνεύσει αυτόν τον θόρυβο.

7.3.2 Hierarchical

Για την επαλήθευση των τάσεων που ανέδειξε ο K-Means, το ίδιο dataset των ετήσιων μεταβολών υποβλήθηκε σε Ιεραρχική Συσταδοποίηση, εφαρμόζοντας τις τέσσερις βασικές μεθόδους διασύνδεσης (Ward, Complete, Average, Single). Η μέθοδος **Ward** με **k=2** ομάδες ήταν η μοναδική που απέφυγε την παγίδα των ακραίων τιμών (outliers). Ο αλγόριθμος (με Silhouette Score 0.374 και SSE 3.648.639.501) κατάφερε να χωρίσει το σύνολο των τμημάτων σε δύο ουσιαστικά προφίλ συμπεριφοράς, τη σταθερή πλειοψηφία και τη συστηματικά ευμετάβλητη μειοψηφία. Για να κατανοήσουμε αυτά τα δύο προφίλ, αναλύουμε τα παραγόμενα διαγράμματα.

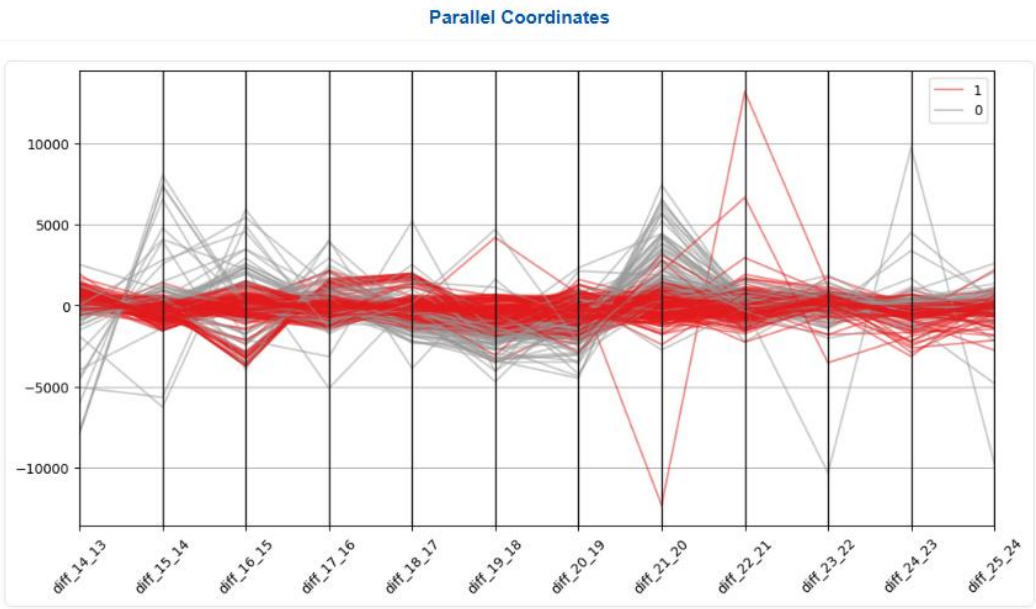
Η διαφορά στη συμπεριφορά των δύο ομάδων αποτυπώνεται στον χώρο μέσω της ανάλυσης κυρίων συνιστωσών.



Εικόνα 29.Διάγραμμα διασποράς των 2 συστάδων ετήσιων μεταβολών της μεθόδου Ward (PCA - 2D) για την περίοδο 2013-2025.

Στην Εικόνα 29, οι δύο πρώτες συνιστώσες ερμηνεύουν το 44.7% της συνολικής διακύμανσης. Παρατηρούμε ότι το Cluster 1 (Κόκκινο χρώμα) διαθέτει έναν εξαιρετικά πυκνό πυρήνα, αλλά και αρκετά σημεία που απλώνονται δεξιά και πάνω. Αντίθετα, το Cluster 0 (Γκρι χρώμα) συγκεντρώνεται στο αριστερό τμήμα του διαγράμματος. Ο αλγόριθμος ουσιαστικά διαχώρισε τη βασική μάζα του συστήματος (κόκκινο) από μια συγκεκριμένη ομάδα τμημάτων (γκρι) που παρουσιάζει μια διαφορετική, δική της δυναμική διακυμάνσεων.

Η πραγματική φύση αυτών των δύο ομάδων γίνεται απόλυτα κατανοητή όταν δούμε την πορεία των μεταβολών τους στον χρόνο.

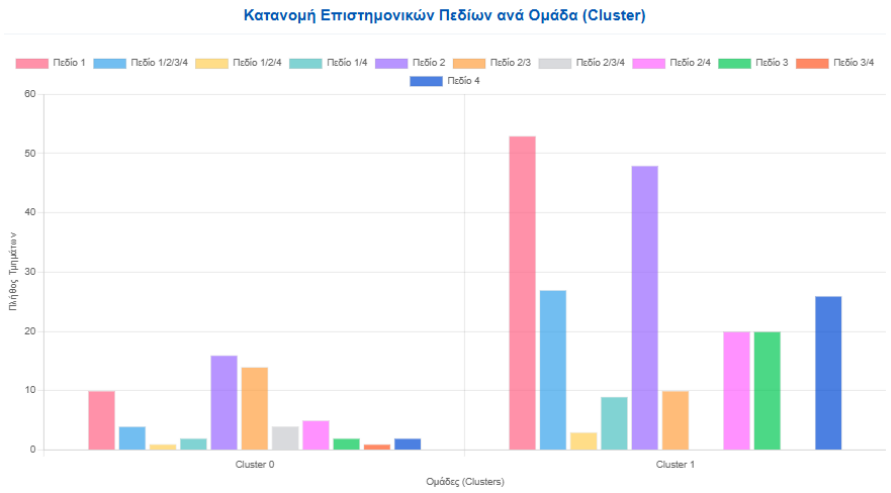


Εικόνα 30.Διάγραμμα παράλληλων συντεταγμένων των ετήσιων μεταβολών (Ward, k=2) για την περίοδο 2013-2025.

Το διάγραμμα της Εικόνας 30, αποκαλύπτει δύο ξεκάθαρους «χαρακτήρες» σχολών:

- **Cluster 1 (Κόκκινες γραμμές - Η Σταθερή Πλειοψηφία):** Αποτελεί τον βασικό κορμό της τριτοβάθμιας εκπαίδευσης. Η συντριπτική πλειονότητα αυτών των γραμμών κινείται σε ένα πολύ στενό εύρος γύρω από τον άξονα του μηδενός. Αυτό σημαίνει ότι αυτές οι σχολές απορροφούν ομαλά τις όποιες αλλαγές (π.χ. δυσκολία θεμάτων), χωρίς να εμφανίζουν ακραίες αυξομειώσεις στις βάσεις τους (με εξαίρεση ελάχιστα μεμονωμένα τμήματα που εντάχθηκαν εδώ αναγκαστικά λόγω της επιλογής k=2).
- **Cluster 0 (Γκρι γραμμές - Η Συστηματική Μεταβλητότητα):** Η μικρότερη αυτή ομάδα παρουσιάζει μια συνεχή «πριονωτή» εικόνα. Οι γραμμές κάνουν διαρκώς μεγάλες αυξήσεις τη μία χρονιά, βαθιές πτώσεις την επόμενη, όπως φαίνεται ξεκάθαρα στα έτη 2014-15 και 2021-23. Πρόκειται για σχολές εξαιρετικά ευάλωτες στις εξεταστικές συγκυρίες.

Η ταυτότητα αυτών των ευάλωτων και σταθερών σχολών επιβεβαιώνεται από την κατανομή των επιστημονικών πεδίων.

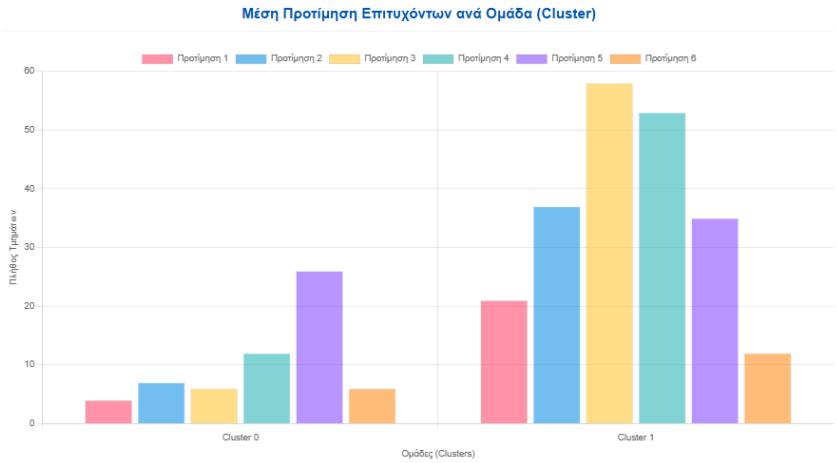


Εικόνα 31.Ραβδόγραμμα κατανομής των Επιστημονικών Πεδίων στις 2 συστάδες μεταβολών (Ward).

Από το ραβδόγραμμα της Εικόνας 31, διαπιστώνουμε τα εξής:

- Στο **Cluster 1** (σταθερή ομάδα) υπάρχει μια μαζική και απόλυτα ισορροπημένη εκπροσώπηση όλων των πεδίων (Πεδίο 1, 2, 3 και 4). Είναι ο καθρέφτης ολόκληρου του εκπαιδευτικού συστήματος.
- Αντίθετα, το **Cluster 0** (ευάλωτη ομάδα) έχει μια πολύ στοχευμένη σύσταση. Κυριαρχείται από το 2ο Πεδίο (Θετικές Επιστήμες - μωβ μπάρα) και τα κοινά τμήματα του Πεδίου 2/3 (πορτοκαλί μπάρα). Αυτό μας δείχνει ότι η έντονη αστάθεια δεν είναι τυχαία, αλλά πλήττει κυρίως συγκεκριμένες σχολές θετικών επιστημών (συνήθως περιφερειακά τμήματα μηχανικών, γεωπονίας ή φυσικομαθηματικών).

Η διαφορά στην ελκυστικότητα των δύο ομάδων αποτυπώνεται εντυπωσιακά στις επιλογές των μαθητών.

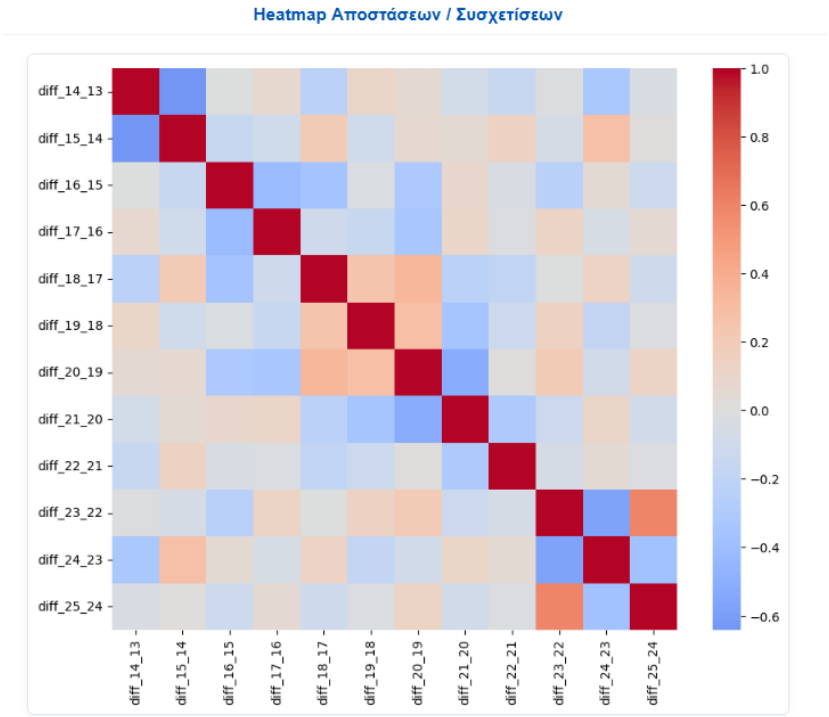


Εικόνα 32.Ραβδόγραμμα συχνότητας της σειράς προτίμησης των επιτυχόντων ανά συστάδα μεταβολών.

Το ραβδόγραμμα της Εικόνας 32, επιβεβαιώνει έναν βασικό κανόνα της εκπαιδευτικής ανάλυσης:

- Στο **Cluster 1** (σταθερές μεταβολές), η κατανομή είναι απολύτως φυσιολογική. Οι μαθητές περνούν σε αυτές τις σχολές κυρίως ως 3η και 4η επιλογή (κορύφωση του γραφήματος). Είναι οι βασικές, ρεαλιστικές επιλογές των υποψηφίων.
- Στο **Cluster 0** (ακραίες μεταβολές), η εικόνα αλλάζει δραματικά. Η συντριπτική πλειονότητα των εισακτέων περνάει σε αυτά τα τμήματα έχοντάς τα ως **5η επιλογή**. Η προτίμηση 1 και 2 είναι σχεδόν ανύπαρκτη. Αυτό αποδεικνύει περίτρανα ότι οι σχολές που εμφανίζουν τις μεγαλύτερες δακυμανσεις από χρονιά σε χρονιά, είναι ακριβώς εκείνες που δηλώνονται στο τέλος του μηχανογραφικού ως λύσεις έσχατης ανάγκης.

Τέλος, εξετάζουμε το πώς συσχετίζονται οι αυξομειώσεις των διαφόρων ετών μεταξύ τους.



Εικόνα 33. Heatmap hierarchical των ετήσιων μεταβολών 2013-2025.

Ο χάρτης θερμότητας της Εικόνας 33, παρουσιάζει μια εντελώς αντίθετη εικόνα από αυτή των απόλυτων βάσεων. Εδώ απουσιάζει το έντονο κόκκινο χρώμα. Αντιθέτως, κυριαρχούν οι χαμηλές, μηδενικές, αλλά και αρνητικές συσχετίσεις (οι μπλε αποχρώσεις). Μια αρνητική συσχέτιση (όπως για παράδειγμα μεταξύ του diff_21_20 και του diff_22_21) σημαίνει ότι λειτουργεί ένας μηχανισμός «διόρθωσης» στην αγορά: μια μεγάλη πτώση των βάσεων τη μία χρονιά (λόγω δυσκολίας θεμάτων), οδηγεί συχνά σε άνοδο την αμέσως επόμενη, επειδή οι υποψήφιοι της νέας χρονιάς «επιτίθενται» μαζικά στις σχολές που είδαν να έχουν χαμηλά μόρια.

Συμπερασματικά για τις ετήσιες μεταβολές της δωδεκαετίας, η μέθοδος **Ward** προκρίνεται ως η μοναδική κατάλληλη ιεραρχική τεχνική, καθώς απέφυγε μια σημαντική μεθοδολογική παγίδα. Παρόλο που οι μέθοδοι Complete, Average και Single κατέγραψαν πλασματικά υψηλό Silhouette Score (0.769), η γεωμετρική ανάλυση έδειξε ότι αυτό αποτελεί μαθηματικό τεχνούργημα (artifact), καθώς οι αλγόριθμοι αυτοί απομόνωσαν απλώς 1-2 ακραία outliers σε ένα cluster, αφήνοντας το 99% των σχολών σε ένα χαοτικό δεύτερο σύνολο. Η μέθοδος Ward, εστιάζοντας στην ελαχιστοποίηση της διακύμανσης,

πέτυχε το χαμηλότερο SSE (3.648.639.501) και αποκάλυψε δύο πραγματικά, ισορροπημένα προφίλ χρονικής συμπεριφοράς (σταθερότητα έναντι μεταβλητότητας).

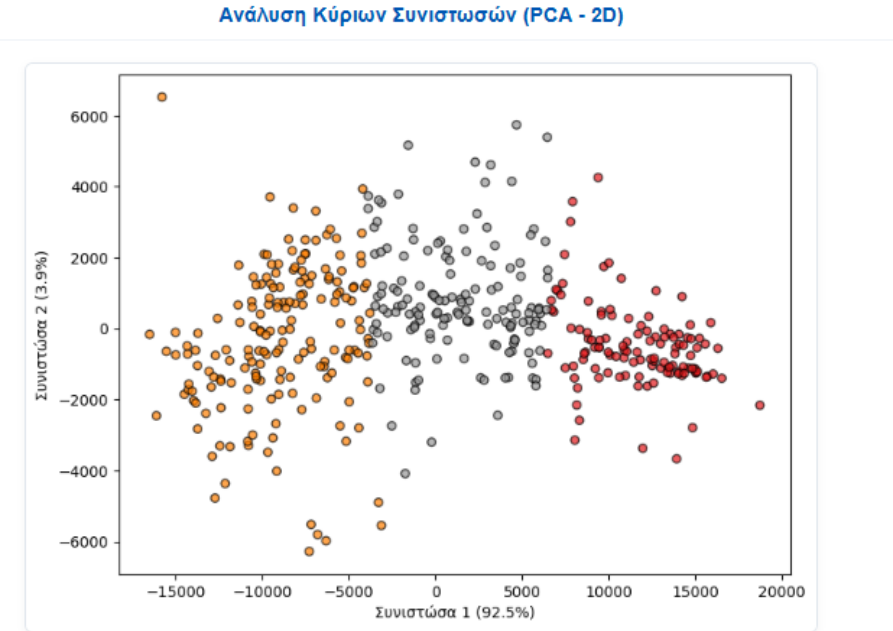
7.4 Συσταδοποίηση Τμημάτων με βάση εισαγωγής στο διάστημα 2019-2025 (k-means, hierarchical)

7.4.1 K-means

Σε αυτή την ενότητα, ο χρονικός ορίζοντας περιορίζεται στην πρόσφατη εξαετία 2019-2025. Η επιλογή αυτής της περιόδου είναι στρατηγικής σημασίας, καθώς περιλαμβάνει σημαντικές δομικές αλλαγές στο ελληνικό εκπαιδευτικό σύστημα, με κυριότερη τη θεσμοθέτηση και εφαρμογή της Ελάχιστης Βάσης Εισαγωγής (ΕΒΕ) από το 2021 αλλά και την δημιουργία νέων τμημάτων το 2019 όταν τα ΤΕΙ έγιναν ΑΕΙ. Η ανάλυση εστιάζει στις απόλυτες τιμές των βάσεων εισαγωγής, προκειμένου να αποτυπωθεί η σύγχρονη βαθμολογική διάρθρωση των πανεπιστημιακών τμημάτων. Για το σύνολο αυτό, ο αλγόριθμος K-Means εφαρμόστηκε με επιλογή k=3 συστάδων, καταγράφοντας τις εξής μετρικές, **Silhouette Score:** 0.476 και **Sum of Squared Errors (SSE):** 6.714.445.534

Το Silhouette Score (0.476) αποτελεί την υψηλότερη τιμή που καταγράφηκε σε όλες τις δοκιμές των απόλυτων τιμών. Το γεγονός αυτό αποδεικνύει μαθηματικά ότι ο περιορισμός του χρονικού εύρους εξομάλυνε τις μακροχρόνιες ασυνέχειες, επιτρέποντας στον αλγόριθμο να δημιουργήσει εξαιρετικά συμπαγείς, διακριτές και γεωμετρικά καθαρές ομάδες σχολών.

Η γεωμετρική καθαρότητα και ο βαθμός διαχωρισμού των τριών ομάδων στο σύγχρονο dataset αποτυπώνονται στο δισδιάστατο επίπεδο μέσω της τεχνικής PCA.

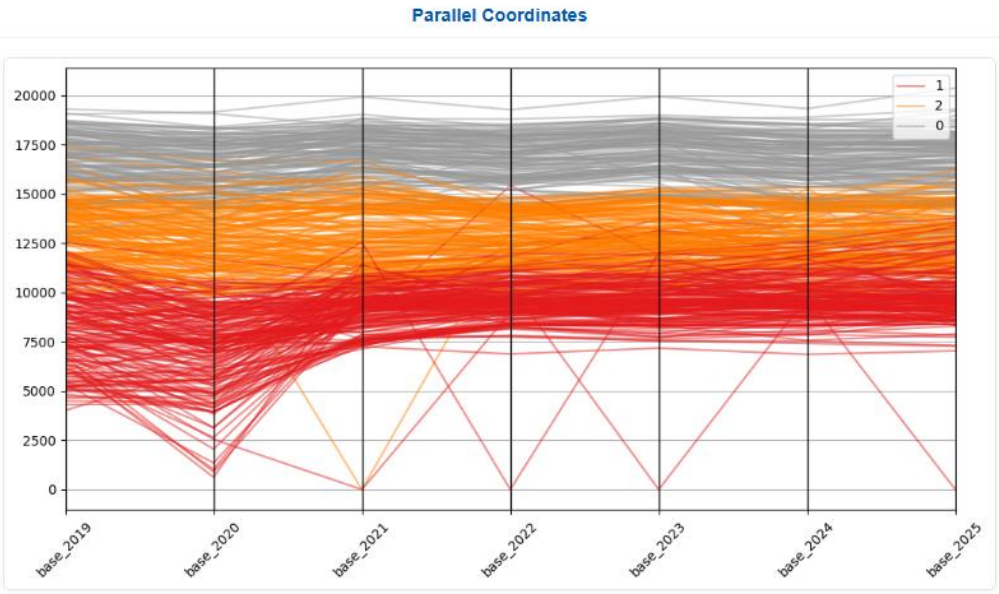


Εικόνα 34.Διάγραμμα διασποράς των συστάδων του K-Means στο επίπεδο των δύο πρώτων κύριων συνιστωσών (PCA) για την περίοδο 2019-2025.

Από την Εικόνα 34, προκύπτει ότι οι δύο πρώτες κύριες συνιστώσες ερμηνεύουν το 96.4% της συνολικής πληροφορίας και διακύμανσης των δεδομένων. Το διάγραμμα διασποράς αναδεικνύει έναν υποδειγματικό γραμμικό διαχωρισμό κατά μήκος του πρώτου άξονα. Στα αριστερά αναπτύσσεται το Cluster 2 (πορτοκαλί χρώμα), στο κέντρο το Cluster 0 (γκρι χρώμα) και στα δεξιά το Cluster 1 (κόκκινο

χρώμα). Η πλήρης απουσία επικαλύψεων και η σαφής απόσταση μεταξύ των ορίων των ομάδων πιστοποιούν την υψηλή στατιστική ποιότητα της συσταδοποίησης.

Η διαχρονική εξέλιξη των απόλυτων μορίων για κάθε τμήμα ανά ομάδα μελετάται μέσα από το διάγραμμα παράλληλων συντεταγμένων της πρόσφατης περιόδου.

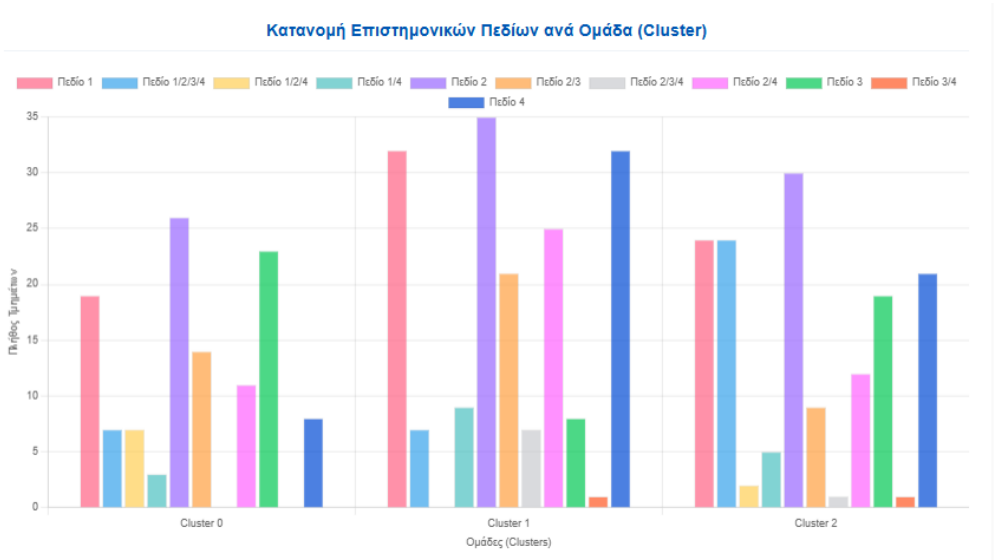


Εικόνα 35.Διάγραμμα παράλληλων συντεταγμένων των βάσεων εισαγωγής ανά συστάδα για την περίοδο 2019-2025.

Το διάγραμμα της Εικόνας 35, αποτυπώνει τη βαθμολογική διαστρωμάτωση, αναδεικνύοντας παράλληλα μια πολύ σημαντική συστημική επίδραση:

- **Cluster 0 (Γκρι γραμμές):** Απομονώνει τα διαχρονικά υψηλόβαθμα τμήματα της χώρας, με τις βάσεις τους να κινούνται σταθερά πάνω από το όριο των 15.000 μορίων, εμφανίζοντας μια εξαιρετικά ομαλή και παράλληλη πορεία, ανεπηρέαστη από εξωτερικές μεταβολές.
- **Cluster 2 (Πορτοκαλί γραμμές):** Συσσωρεύει τα τμήματα της μεσαίας βαθμολογικής κλίμακας, τα οποία καταλαμβάνουν τον χώρο μεταξύ 10.000 και 15.000 μορίων.
- **Cluster 1 (Κόκκινες γραμμές):** Περιλαμβάνει τα χαμηλόβαθμα τμήματα της χώρας (κάτω των 10.000 μορίων). Στο cluster αυτό παρατηρείται ένα εξαιρετικά ενδιαφέρον φαινόμενο: μετά το έτος 2021, οι γραμμές των βάσεων σταματούν να παρουσιάζουν τις ακραίες πτώσεις των προηγούμενων ετών και εμφανίζουν μια σαφή τάση σύγκλισης και τεχνητής σταθεροποίησης προς τα πάνω. Η συμπεριφορά αυτή αποτελεί την άμεση μαθηματική αποτύπωση της εφαρμογής της Ελάχιστης Βάσης Εισαγωγής (ΕΒΕ), η οποία απέκλεισε τις πολύ χαμηλές βαθμολογίες, «κόβοντας» ουσιαστικά την ουρά της κατανομής.

Προκειμένου να αναλυθεί η επιστημονική ταυτότητα των τριών ομάδων για το διάστημα 2019-2025, εξετάστηκε η κατανομή των επιστημονικών πεδίων στο εσωτερικό τους.

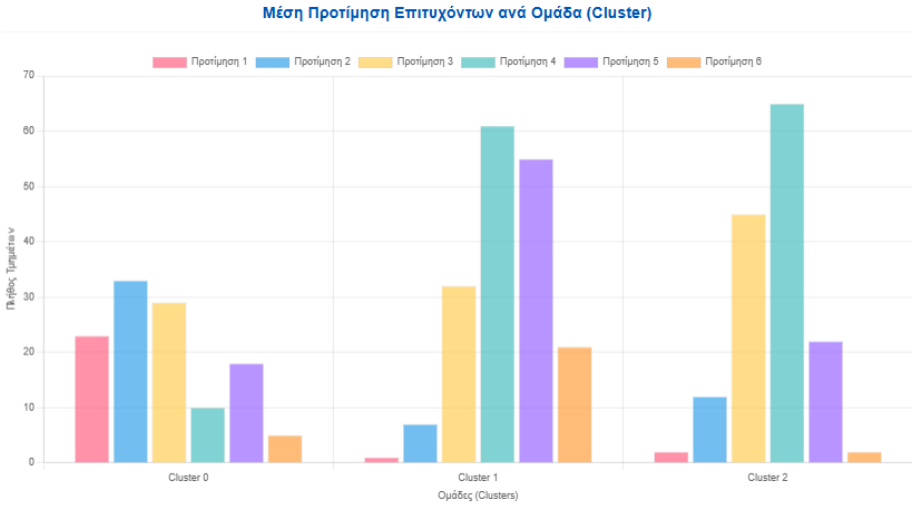


Εικόνα 36.Ραβδόγραμμα κατανομής των Επιστημονικών Πεδίων στο εσωτερικό των τριών συστάδων.

Η ανάλυση του ραβδογράμματος της Εικόνας 36, επιβεβαιώνει τις παραδοσιακές τάσεις της εκπαιδευτικής ζήτησης, προσαρμοσμένες στα σύγχρονα δεδομένα:

- Το **Cluster 0** (υψηλόβαθμα) κυριαρχείται από το Πεδίο 3 (Επιστήμες Υγείας) και το Πεδίο 2 (Θετικές/Τεχνολογικές Επιστήμες), στεγάζοντας τις ιατρικές, οδοντιατρικές και τις κεντρικές πολυτεχνικές σχολές.
- Το **Cluster 1** (χαμηλόβαθμα) εμφανίζει τη μεγαλύτερη συγκέντρωση τμημάτων από το Πεδίο 1 (Ανθρωπιστικές Σπουδές) και το Πεδίο 4 (Οικονομία και Πληροφορική), αντανακλώντας τη μειωμένη ζήτηση που καταγράφεται στα περιφερειακά τμήματα αυτών των κατευθύνσεων.
- Το **Cluster 2** (μεσαία) παρουσιάζει μια ισορροπημένη δομή, με αξιοσημείωτη παρουσία του Πεδίου 1, του Πεδίου 2 και του Πεδίου 4, επιβεβαιώνοντας ότι τα περισσότερα τμήματα αυτών των επιστημών αποτελούν τη βάση της μεσαίας βαθμολογικής ζώνης.

Η ελκυστικότητα των σύγχρονων συστάδων αποτυπώνεται ανάγλυφα μέσα από τη μέση σειρά προτίμησης που δήλωσαν οι επιτυχόντες στα μηχανογραφικά τους δελτία.

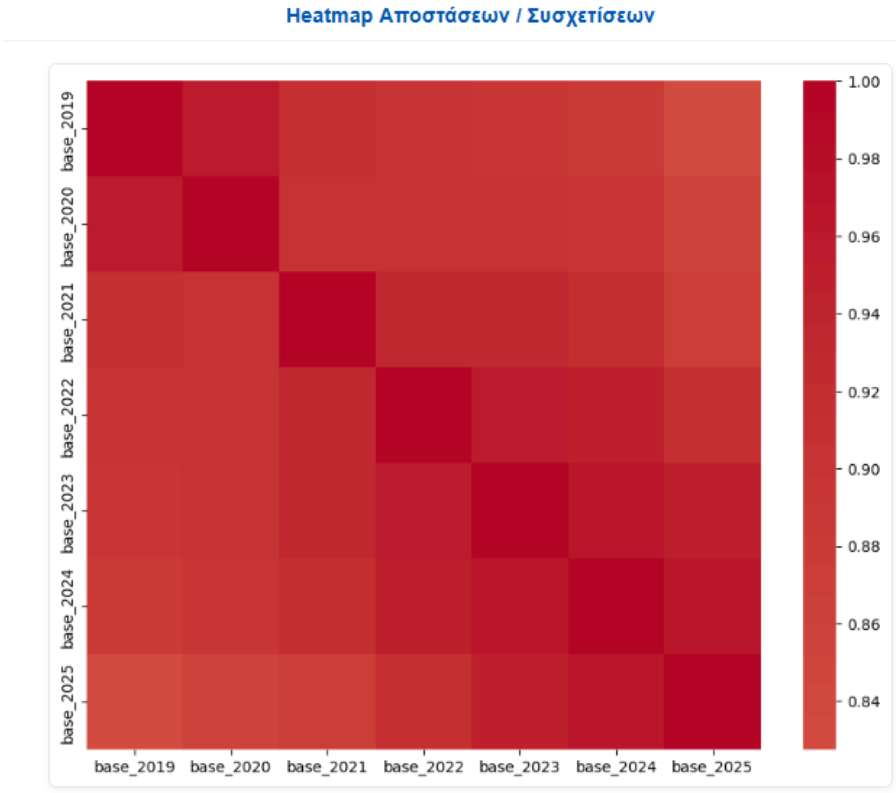


Εικόνα 37.Ραβδόγραμμα συχνότητας της σειράς προτίμησης των επιτυχόντων ανά συστάδα.

Όπως παρατηρείται στην Εικόνα 37, η κατανομή των προτιμήσεων ακολουθεί μια ξεκάθαρη φθίνουσα πορεία ανάλογα με το βαθμολογικό επίπεδο:

- Στο **Cluster 0** (υψηλόβαθμα), οι εισακτέοι πέτυχαν στη σχολή τους όντας κυρίως η 1η, η 2η ή η 3η επιλογή τους, γεγονός που αποδεικνύει ότι η ζήτηση για αυτά τα τμήματα παραμένει στοχευμένη και συνειδητή.
- Στο **Cluster 2** (μεσαία), η κατανομή παρουσιάζει κορύφωση στην 3η και 4η επιλογή, λειτουργώντας ως η χρυσή τομή μεταξύ επιθυμίας και βαθμολογικής ασφάλειας.
- Στο **Cluster 1** (χαμηλόβαθμα), η κατανομή μετατοπίζεται εμφανώς προσφέροντας υψηλά ποσοστά στις θέσεις 4 και 5, γεγονός που πιστοποιεί ότι τα τμήματα αυτά επιλέγονται κατά κανόνα προς το τέλος του μηχανογραφικού δελτίου, ως επιλογές ύστατης ανάγκης για την εξασφάλιση της εισαγωγής στην τριτοβάθμια εκπαίδευση.

Η ανάλυση του αλγορίθμου K-Means για την πρόσφατη εξαετία ολοκληρώνεται με τη μελέτη των χρονικών συσχετίσεων (κατά Pearson) μεταξύ των απόλυτων βάσεων των ετών.



Εικόνα 38.Heatmap Kmeans των βάσεων εισαγωγής μεταξύ των ετών 2019-2025.

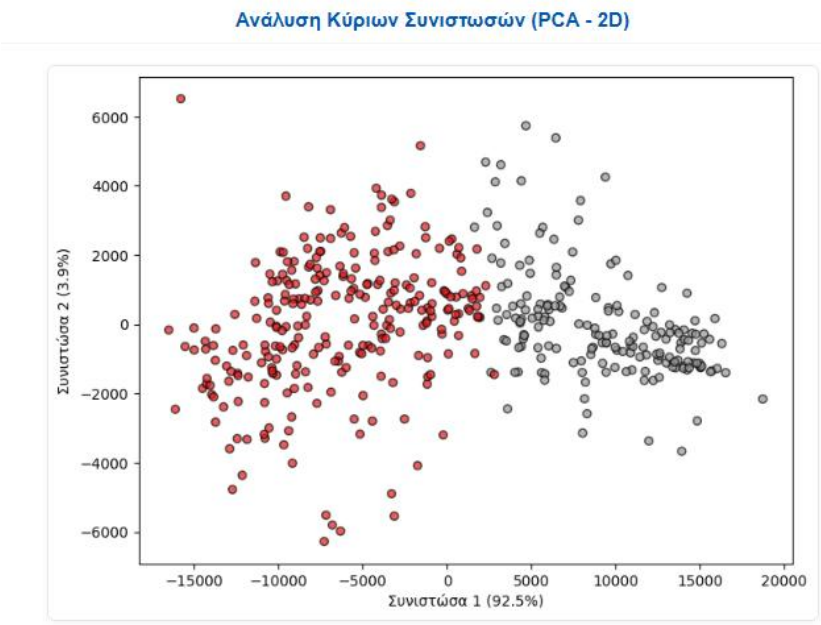
Ο χάρτης θερμότητας της Εικόνας 38, παρουσιάζει την πιο συμπαγή και «θερμή» εικόνα από όλους τους αντίστοιχους χάρτες της μελέτης. Όπως παρατηρείται στην κλίμακα στα δεξιά του γραφήματος, όλοι οι συντελεστές συσχέτισης είναι εξαιρετικά υψηλοί, κινούμενοι σε ένα πολύ στενό και υψηλό εύρος, από 0.84 έως 1.00.

Η σχεδόν απόλυτη αυτή γραμμική συσχέτιση αποδεικνύει ότι στην πρόσφατη εξαετία η δομή και η ιεραρχία των βάσεων στην Ελλάδα έχει αποκτήσει μια εντυπωσιακή σταθερότητα. Οι συστημικές αλλαγές (όπως η εισαγωγή της ΕΒΕ) δεν ανέτρεψαν τη σειρά των σχολών, αλλά αντίθετα «κλείδωσαν» τις βαθμολογικές αποστάσεις μεταξύ τους. Μια σχολή που ήταν υψηλόβαθμη το 2019, παραμένει σχεδόν νομοτελειακά υψηλόβαθμη και το 2025, καθιστώντας το εκπαιδευτικό σύστημα εξαιρετικά προβλέψιμο ως προς τη διαστρωμάτωσή του.

7.4.2 Hierarchical

Με σκοπό τη διασταύρωση και την επαλήθευση της ορθότητας των αποτελεσμάτων του K-Means, το dataset της περιόδου 2019-2025 υποβλήθηκε σε Ιεραρχική Συσταδοποίηση, αξιολογώντας τις τέσσερις κλασικές μεθόδους διασύνδεσης (Ward, Complete, Average, Single). Η μέθοδος **Ward** με **k=2** ομάδες κυριάρχησε απόλυτα (Silhouette Score 0.566 και SSE 11.433.204.040). Ο αλγόριθμος χώρισε τον ακαδημαϊκό χάρτη σε δύο μεγάλους, ξεκάθαρους πόλους, προσφέροντας την πιο καθαρή εικόνα για το πώς διαμορφώθηκαν οι βάσεις τα τελευταία χρόνια. Ακολουθεί η ανάλυση των γραφημάτων για την πλήρη κατανόηση αυτών των δύο ομάδων.

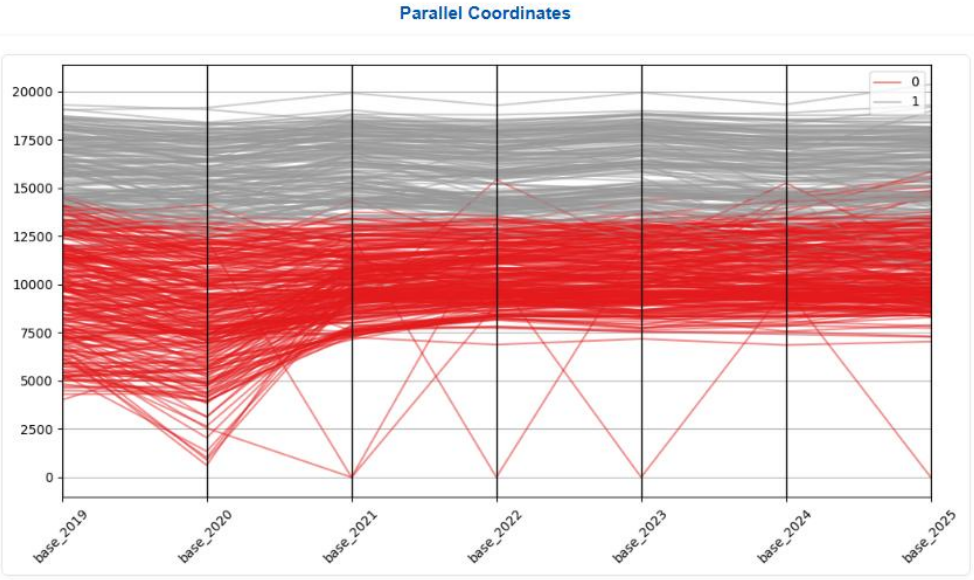
Ο διαχωρισμός των δύο ομάδων αποτυπώνεται ιδανικά στο διάγραμμα της ανάλυσης κυρίων συνιστωσών.



Εικόνα 39.Διάγραμμα διασποράς των 2 συστάδων της μεθόδου Ward (PCA - 2D) για τις βάσεις 2019-2025.

Στην Εικόνα 39, βλέπουμε έναν σχεδόν τέλειο γεωμετρικό διαχωρισμό. Η πρώτη κύρια συνιστώσα (που συγκεντρώνει το εντυπωσιακό 92.5% της πληροφορίας) λειτουργεί ως το απόλυτο σύνορο. Αριστερά απλώνεται το Cluster 0 (Κόκκινο χρώμα) και δεξιά το Cluster 1 (Γκρι χρώμα). Η απουσία ανάμειξης στο κέντρο δικαιολογεί το υψηλό Silhouette Score και αποδεικνύει ότι τα δύο αυτά σύνολα σχολών δεν έχουν καμία βαθμολογική σχέση μεταξύ τους.

Η εξέλιξη των βάσεων αυτών των δύο ομάδων δείχνει καθαρά και την επίδραση των πρόσφατων αλλαγών στο σύστημα (όπως η ΕΒΕ).

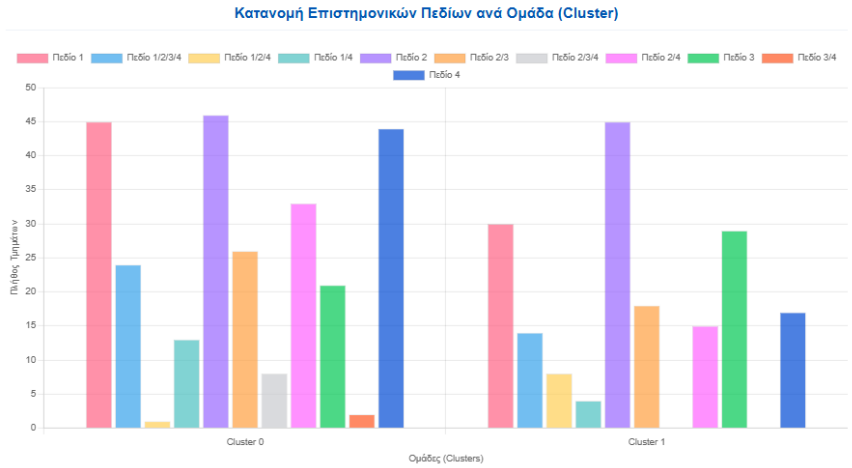


Εικόνα 40.Διάγραμμα παράλληλων συντεταγμένων των βάσεων εισαγωγής (Ward, k=2) για την περίοδο 2019-2025.

Η Εικόνα 40, είναι εξαιρετικά αποκαλυπτική:

- **Cluster 1 (Γκρι γραμμές - Υψηλόβαθμες Σχολές):** Είναι τα τμήματα που βρίσκονται σταθερά στο πάνω μέρος του γραφήματος (κυρίως πάνω από τα 13.000 μόρια). Οι γραμμές τους είναι απολύτως ευθείες και παράλληλες, δείχνοντας ότι αυτά τα τμήματα έμειναν ανεπηρέαστα από τις όποιες αλλαγές του συστήματος.
- **Cluster 0 (Κόκκινες γραμμές - Μεσαίες/Χαμηλόβαθμες Σχολές):** Περιλαμβάνει τα τμήματα του κάτω μισού. Εδώ παρατηρούμε καθαρά το "φαινόμενο της ΕΒΕ": Ενώ τα έτη 2019 και 2020 υπάρχουν βαθιές πτώσεις προς το μηδέν, από το 2021 και μετά (έτος εφαρμογής της ΕΒΕ) οι γραμμές σταματούν να πέφτουν και συγκλίνουν μαζικά προς τα πάνω, δημιουργώντας μια τεχνητή σταθεροποίηση στη βάση του γραφήματος.

Για να διαπιστώσουμε ποια τμήματα ανήκουν σε κάθε πόλο, εξετάζουμε την κατανομή των πεδίων.

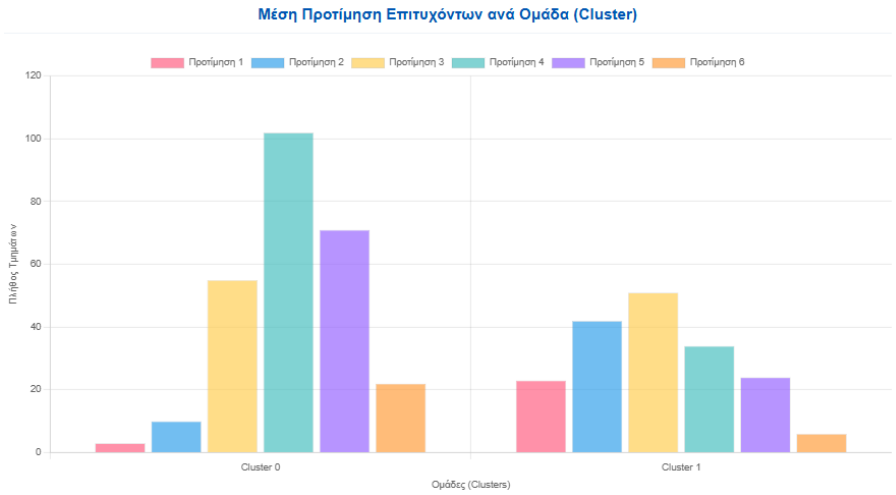


Εικόνα 41.Ραβδόγραμμα κατανομής των Επιστημονικών Πεδίων στις 2 συστάδες της μεθόδου Ward (2019-2025).

Από το ραβδόγραμμα της Εικόνας 41, προκύπτει ξεκάθαρα η ταυτότητα των ομάδων:

- Το **Cluster 1** (υψηλόβαθμα) κυριαρχείται από το 3ο Πεδίο (Επιστήμες Υγείας - πράσινη μπάρα) και το 2ο Πεδίο (Θετικές/Πολυτεχνικές - μωβ μπάρα).
- Το **Cluster 0** (μεσαία/χαμηλόβαθμα) έχει τελείως διαφορετική σύσταση. Εκεί υπερτερούν οι σχολές του 1ου Πεδίου (Ανθρωπιστικές - ροζ μπάρα) και του 4ου Πεδίου (Οικονομία/Πληροφορική - μπλε μπάρα), αποδεικνύοντας ότι τα συγκεκριμένα πεδία απαρτίζουν τον κύριο όγκο των σχολών μέσης και χαμηλής ζήτησης στη σύγχρονη εποχή.

Ο τρόπος με τον οποίο δηλώνουν οι υποψήφιοι αυτά τα τμήματα δείχνει τη δυναμική της κάθε ομάδας.

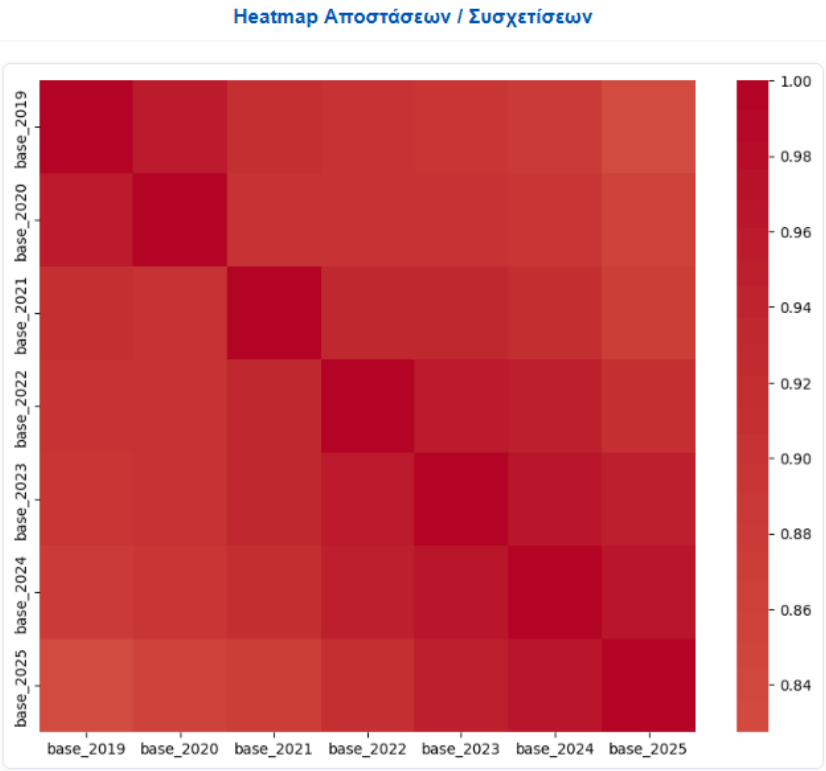


Εικόνα 42.Ραβδόγραμμα συχνότητας της σειράς προτίμησης των επιτυχόντων ανά συστάδα (2019-2025).

Στην Εικόνα 42, αποτυπώνεται καθαρά η αξία των σχολών στα μάτια των μαθητών:

- Στο **Cluster 1** (υψηλόβαθμα), οι περισσότεροι επιτυχόντες πέρασαν έχοντας τη σχολή ως 2η ή 3η επιλογή, ενώ υπάρχει και ισχυρό ποσοστό 1ης επιλογής. Πρόκειται για τμήματα που αποτελούν προτεραιότητα στα μηχανογραφικά.
- Στο **Cluster 0** (μεσαία/χαμηλόβαθμα), το γράφημα έχει μετατοπιστεί έντονα δεξιά. Η απόλυτη κορυφή βρίσκεται στην 4η και 5η επιλογή. Αυτό δείχνει ότι ο μεγάλος αυτός πόλος σχολών επιλέγεται κυρίως ως "δίχτυ ασφαλείας" από τους υποψηφίους προς το τέλος του μηχανογραφικού τους.

Τέλος, βλέπουμε πόσο στενά συνδέονται οι βάσεις αυτών των ετών μεταξύ τους.



Εικόνα 43.Heatmap hierarchical των βάσεων εισαγωγής 2019-2025.

Η Εικόνα 43, παρουσιάζει ίσως την πιο "θερμή" και συμπαγή εικόνα (σκούρο κόκκινο παντού). Οι συσχετίσεις κινούνται σε εξαιρετικά υψηλά επίπεδα (από 0.84 έως 1.00). Αυτό σημαίνει ότι στην πρόσφατη εξαετία, η ιεραρχία των σχολών "κλείδωσε". Όποιες αλλαγές κι αν έγιναν στο σύστημα, η βαθμολογική απόσταση μεταξύ μιας υψηλόβαθμης και μιας χαμηλόβαθμης σχολής έμεινε πρακτικά αναλλοίωτη, κάνοντας το σύστημα εξαιρετικά προβλέψιμο. Αυτή η απόλυτη σταθερότητα εξηγεί γιατί ο αλγόριθμος Ward μπόρεσε να χωρίσει τα τμήματα τόσο τέλεια σε 2 ομάδες.

Συμπερασματικά για την πρόσφατη περίοδο των απόλυτων βάσεων, η μέθοδος **Ward** αναδεικνύεται ως η κορυφαία ιεραρχική επιλογή, καταγράφοντας την υψηλότερη ποιότητα διαχωρισμού (Silhouette 0.566) και εξαιρετικά χαμηλό εσωτερικό σφάλμα (SSE 11.433.204.040). Ο variance-based μηχανισμός της αποδείχθηκε ιδανικός για το πιο εξομαλυσμένο περιβάλλον της εξαετίας 2019-2025. Κατάφερε να εξουδετερώσει πλήρως το Chaining Effect που αλλοίωσε τη Single Linkage (SSE 36.598.518.160) και, σε αντίθεση με τη μέθοδο Complete, προσέφερε έναν απόλυτα ομοιογενή δυαδικό διαχωρισμό, αποτυπώνοντας με ακρίβεια τη σύγχρονη βαθμολογική διάρθρωση των τμημάτων.

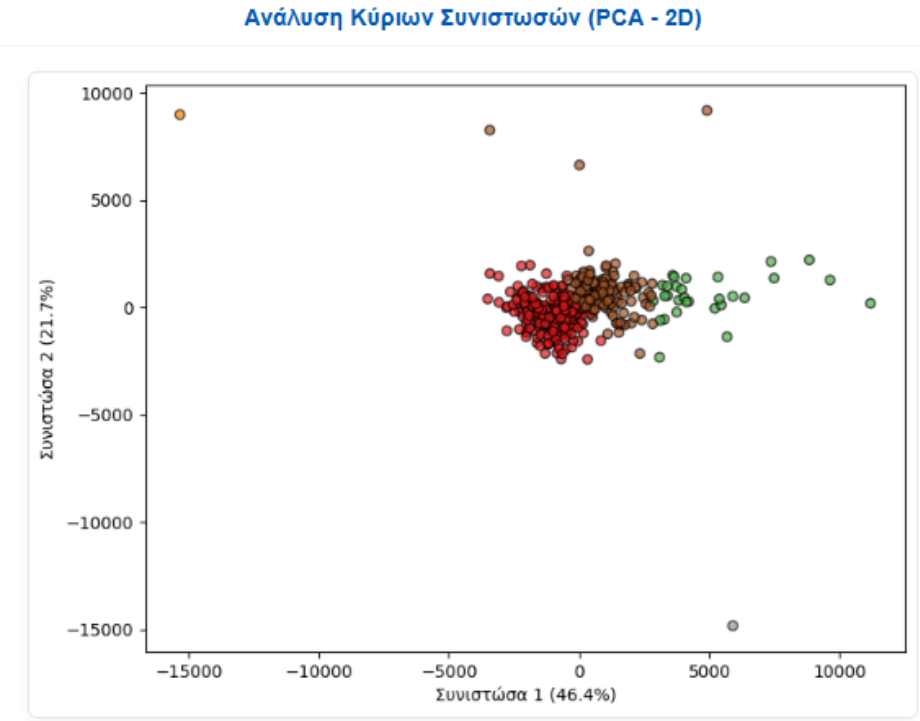
7.5 Συσταδοποίηση Τμημάτων βάσει κοινής συμπεριφοράς στο διάστημα 2019-2025 (k-means, hierarchical)

7.5.1 K-means

Στο τελευταίο στάδιο της ανάλυσης, μελετήθηκε η δυναμική συμπεριφορά των τμημάτων αποκλειστικά για την πρόσφατη εξαετία (2019-2025). Τα δεδομένα μετασχηματίστηκαν στις ετήσιες διαφορές των βάσεων εισαγωγής, με στόχο την αποτύπωση των σύγχρονων τάσεων και της αντίδρασης των σχολών στις πρόσφατες συστημικές αλλαγές (όπως η καθιέρωση της ΕΒΕ). Ο αλγόριθμος K-Means εφαρμόστηκε με επιλογή k=5 συστάδων καταγράφοντας τις εξής μετρικές, **Silhouette Score:** 0.277 και

Sum of Squared Errors (SSE): 1.886.280.888 . Η βελτίωση του Silhouette Score σε σύγκριση με την ανάλυση των μεταβολών ολόκληρης της περιόδου υποδηλώνει ότι οι αυξομειώσεις των τελευταίων ετών ακολουθούν πιο συγκροτημένα και αναγνωρίσιμα μοτίβα.

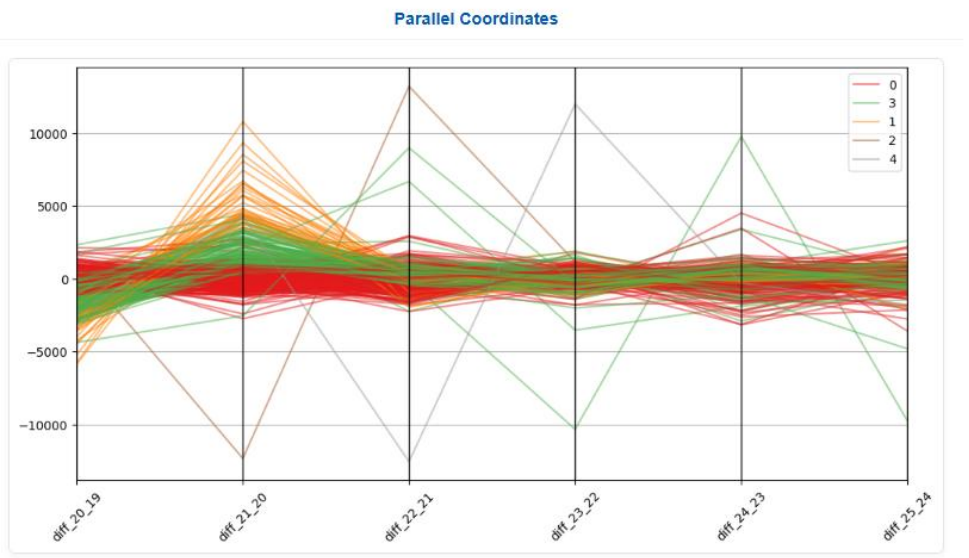
Η ικανότητα του αλγορίθμου να διαχωρίσει τις διαφορετικές χρονικές συμπεριφορές αποτυπώνεται γεωμετρικά μέσω της τεχνικής PCA.



Εικόνα 44.Διάγραμμα διασποράς των 5 συστάδων του K-Means στο επίπεδο των δύο πρώτων κύριων συνιστωσών (PCA) για τις ετήσιες μεταβολές.

Όπως παρατηρείται στην Εικόνα 44, οι δύο πρώτες κύριες συνιστώσες κατορθώνουν να συγκρατήσουν ένα υψηλό ποσοστό της συνολικής πληροφορίας 68.1%, η γεωμετρική κατανομή των συστάδων αποκαλύπτει τους εγγενείς περιορισμούς της ομαδοποίησης στο συγκεκριμένο dataset. Δεν υφίσταται ξεκάθαρος και φυσικός διαχωρισμός μεταξύ των βασικών ομάδων στο κέντρο του διαγράμματος. Το Κόκκινο cluster (0) και το Καφέ cluster (2) αλληλοεπικαλύπτονται σε μεγάλο βαθμό, δημιουργώντας ένα ενιαίο, αδιαχώριστο νέφος σημείων. Αυτό υποδηλώνει ότι οι ετήσιες μεταβολές στην πρόσφατη εξαετία δεν σχηματίζουν διακριτές ομάδες συμπεριφοράς, αλλά ένα συνεχές φάσμα διακυμάνσεων, πάνω στο οποίο ο αλγόριθμος K-Means αναγκάστηκε να επιβάλει «τεχνητά» γεωμετρικά όρια (αναγκαστική κατάτμηση). Ουσιαστικός και καθαρός διαχωρισμός επιτυγχάνεται μόνο στην περιφέρεια της κατανομής, όπου ο αλγόριθμος καταφέρνει να απομονώσει μεμονωμένα τμήματα με ακραίες αυξομειώσεις (όπως το γκρι σημείο στο κάτω μέρος ή τα πράσινα σημεία στα δεξιά). Συνεπώς, για τον κύριο όγκο των δεδομένων, η συσταδοποίηση σε αυτό το επίπεδο λειτούργησε περισσότερο ως μηχανισμός ανίχνευσης ακραίων τιμών (outlier detection) παρά ως δημιουργός ξεκάθαρων, ποιοτικά διαφορετικών ομάδων.

Η μορφή και το μέγεθος των αυξομειώσεων που χαρακτηρίζουν την κάθε ομάδα αναλύονται μέσω του διαγράμματος παράλληλων συντεταγμένων.

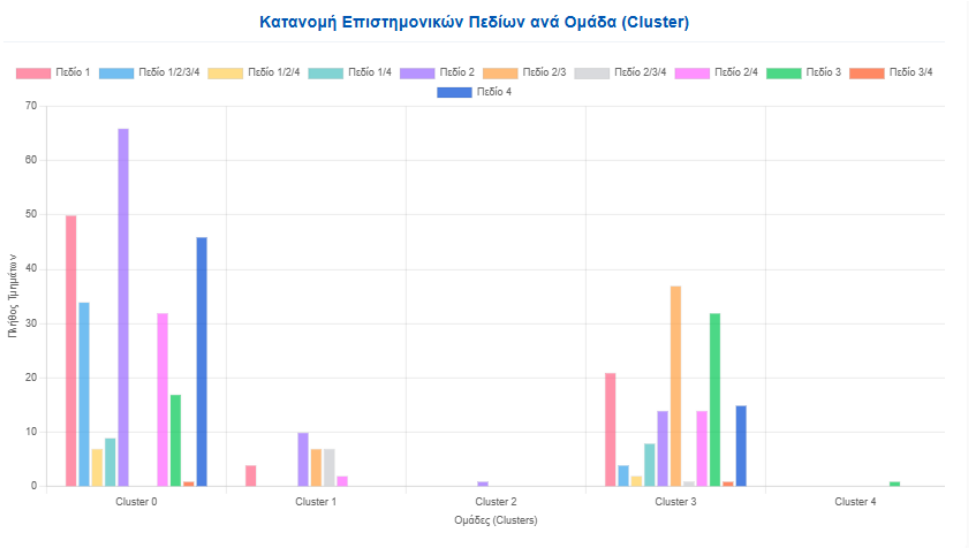


Εικόνα 45.Διάγραμμα παράλληλων συντεταγμένων των ετήσιων μεταβολών ανά συστάδα (2019-2025).

Το διάγραμμα της Εικόνας 45, είναι ιδιαίτερα αποκαλυπτικό όσον αφορά την επιρροή συγκεκριμένων χρονικών γεγονότων:

- **Cluster 0 (Κόκκινες γραμμές):** Αντιπροσωπεύει τη γενική ομαλότητα του συστήματος. Οι μεταβολές κυμαίνονται σε ένα στενό εύρος κοντά στο μηδέν, αποδεικνύοντας ότι η πλειοψηφία των τμημάτων απορρόφησε ομαλά τις αλλαγές του εξεταστικού συστήματος.
- **Υπόλοιπα Clusters (1, 2, 3, 4):** Καθένα από αυτά τα μικρά clusters απομονώνει μια συγκεκριμένη «σοκ»-αντίδραση. Για παράδειγμα, το Cluster 1 (Πορτοκαλί) καταγράφει μια τεράστια άνοδο άνω των 10.000 μορίων στο diff_21_20 (επίδραση της ΕΒΕ). Το Cluster 4 (Γκρι) εμφανίζει μια βίαιη πτώση στο diff_22_21 και άμεση επαναφορά στο diff_23_22. Η συμπεριφορά αυτών των γραμμών αντικατοπτρίζει τις απότομες ανακατατάξεις που υπέστησαν κυρίως τα χαμηλόβαθμα τμήματα τα τελευταία χρόνια.

Για να διαπιστωθεί εάν οι ακραίες αυτές μεταβολές είναι χαρακτηριστικό συγκεκριμένων επιστημών, αναλύθηκε η κατανομή των πεδίων.

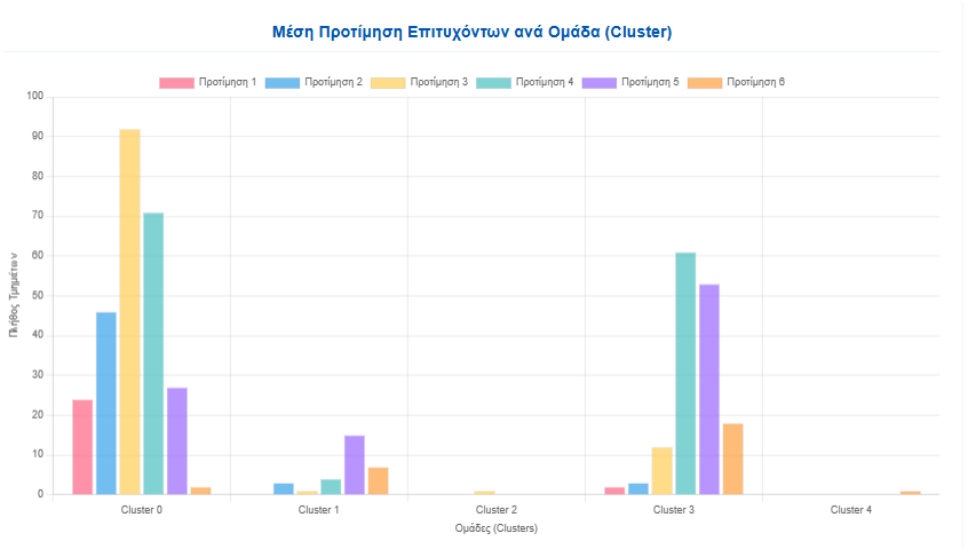


Εικόνα 46.Ραβδόγραμμα κατανομής των Επιστημονικών Πεδίων στο εσωτερικό των 5 συστάδων μεταβολών.

Το ραβδόγραμμα της Εικόνας 46, αποδεικνύει ότι οι έντονες αναταράξεις δεν έπληξαν όλα τα πεδία ομοιόμορφα:

- Στο **Cluster 0** (η ομάδα της σταθερότητας), παρατηρείται μαζική εκπροσώπηση όλων των πεδίων, με ιδιαίτερη δυναμική στο Πεδίο 2 (Θετικές), το Πεδίο 4 (Οικονομία) και το Πεδίο 1 (Ανθρωπιστικές).
- Στα μικρότερα clusters της υψηλής μεταβλητότητας (όπως το **Cluster 3**), υπάρχει επιλεκτική συγκέντρωση. Στο Cluster 3 παρατηρείται έντονη παρουσία του Πεδίου 2/3 (Πορτοκαλί μπάρα) και του Πεδίου 3 (Πράσινη μπάρα). Αυτό καταδεικνύει ότι ορισμένες αλλαγές (π.χ. στην ποσόστωση ή στους συντελεστές μαθημάτων) προκάλεσαν στοχευμένες αυξομειώσεις αποκλειστικά σε τμήματα των επιστημών υγείας και των θετικών επιστημών.

Η σχέση μεταξύ της μεταβλητότητας μιας σχολής και της προτεραιότητάς της στις επιλογές των υποψηφίων εξετάζεται μέσω της μέσης σειράς προτίμησης.

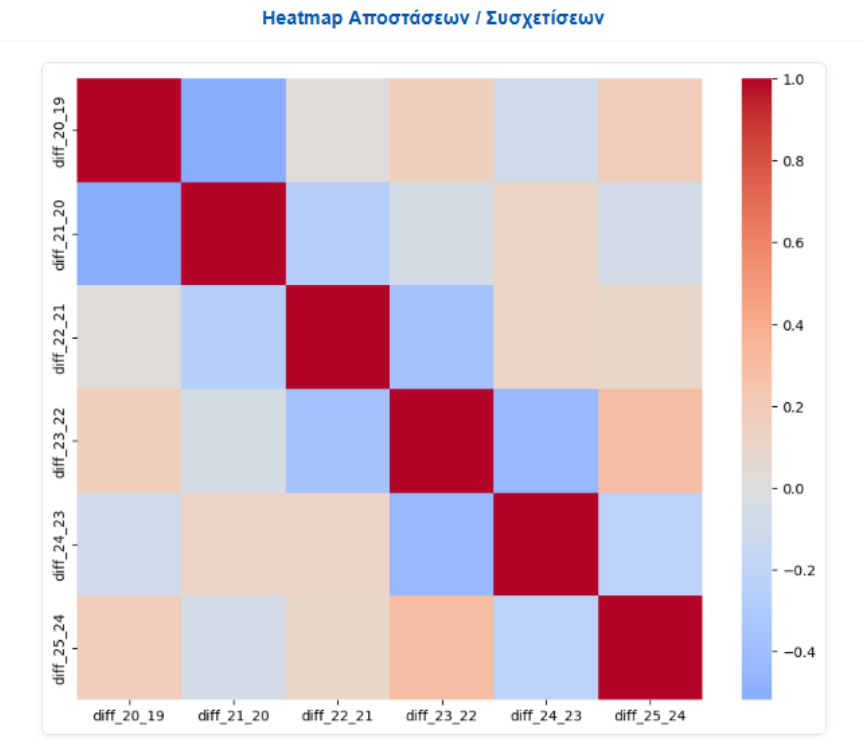


Εικόνα 47.Ραβδόγραμμα συχνότητας της σειράς προτίμησης των επιτυχόντων ανά συστάδα μεταβολών.

Η Εικόνα 47, έρχεται να επιβεβαιώσει ένα πάγιο μοτίβο συμπεριφοράς:

- Στο σταθερό **Cluster 0**, η κατανομή είναι φυσιολογική, με απόλυτη κορύφωση στην 3η προτίμηση και πολύ ισχυρή παρουσία της 2ης και της 4ης. Πρόκειται για σχολές που αποτελούν βασικούς κορμούς των μηχανογραφικών.
- Στα clusters της ακραίας μεταβλητότητας (όπως το **Cluster 1** και το **Cluster 3**), η εικόνα αλλάζει ριζικά. Η συντριπτική πλειονότητα των εισακτέων είχε τα τμήματα αυτά ως 4η ή 5η επιλογή (στο Cluster 3) ή ως 5η επιλογή (στο Cluster 1). Αποδεικνύεται περίτρανα ότι οι σχολές που υπέστησαν τις μεγαλύτερες βαθμολογικές αναταράξεις την τελευταία εξαετία ήταν εκείνες που εξ αρχής αποτελούσαν λύσεις ανάγκης και χαμηλής προτεραιότητας.

Η δυναμική ανάλυση ολοκληρώνεται με τον χάρτη θερμότητας των συντελεστών συσχέτισης Pearson για τις ετήσιες μεταβολές.



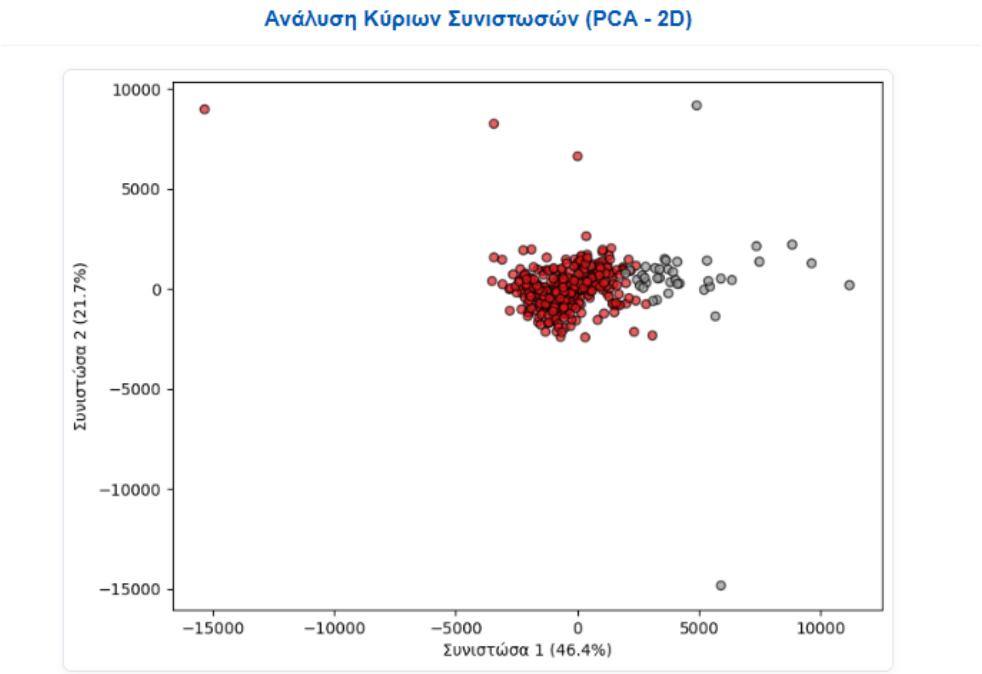
Εικόνα 48.Heatmap kmeans των ετήσιων μεταβολών 2019-2025.

Σε πλήρη αντίθεση με την απόλυτη σταθερότητα των βάσεων εισαγωγής, ο χάρτης της Εικόνας 48, αναδεικνύει την αυτονομία των ετήσιων μεταβολών. Παρατηρούνται ενδιαφέρουσες αρνητικές συσχετίσεις (μπλε αποχρώσεις), όπως μεταξύ της μεταβολής diff_21_20 και της diff_20_19. Η αρνητική αυτή συσχέτιση μεταφράζεται ως διόρθωση της αγοράς: μια απότομη πτώση τη μία χρονιά συχνά ακολουθείται από μια τεχνητή άνοδο την επόμενη, λόγω αλλαγής της συμπεριφοράς των υποψηφίων που "κυνηγούν" τις χαμηλές βάσεις της προηγούμενης χρονιάς. Γενικά, οι χαμηλές τιμές συσχέτισης επιβεβαιώνουν ότι οι ετήσιες αυξομειώσεις αποτελούν μεμονωμένα γεγονότα, εξαρτώμενα πλήρως από τη συγκυρία της εκάστοτε χρονιάς.

7.5.2 Hierarchical

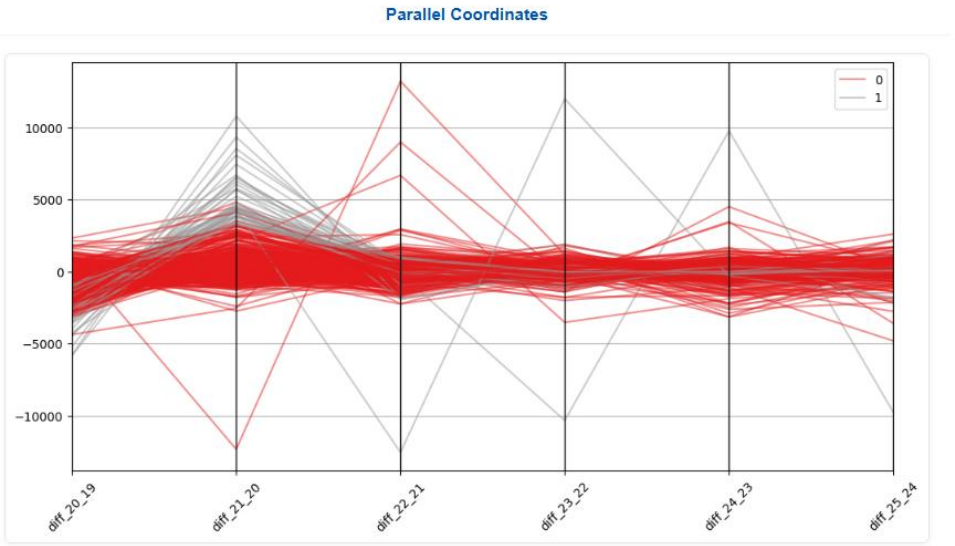
Η ανάλυση των σύγχρονων ετήσιων μεταβολών ολοκληρώθηκε με την εφαρμογή των αλγορίθμων ιεραρχικής συσταδοποίησης, αξιολογώντας τις τέσσερις κλασικές μεθόδους διασύνδεσης (Ward, Complete, Average, Single). Όπως διαπιστώθηκε από τη σύγκριση των μετρικών για την πρόσφατη εξαετία, η μέθοδος **Ward** με **k=2** ομάδες ήταν και πάλι η μόνη που προσέφερε ρεαλιστικά αποτελέσματα (Silhouette Score 0.529 και το χαμηλότερο SSE: 2.942.447.562). Αντίθετα με άλλες μεθόδους που ξεγελάστηκαν από ακραίες τιμές (outliers), ο αλγόριθμος Ward κατάφερε να διαχωρίσει με επιτυχία το σύνολο των τμημάτων σε δύο ξεκάθαρους χαρακτήρες, σε αυτούς που κράτησαν σταθερή πορεία και σε αυτούς που υπέστησαν τα μεγαλύτερες βαθμολογικές απώλειες. Ακολουθούν αναλυτικά αυτά τα δύο προφίλ μέσα από τα διαγράμματα.

Η γεωμετρική διαφορά στη συμπεριφορά των δύο ομάδων φαίνεται στο διάγραμμα της ανάλυσης κυρίων συνιστωσών.



Εικόνα 49.Διάγραμμα διασποράς των 2 συστάδων ετήσιων μεταβολών της μεθόδου Ward (PCA - 2D) για την περίοδο 2019-2025.

Στην Εικόνα 49, οι δύο πρώτες συνιστώσες ερμηνεύουν ένα ικανοποιητικό 68.1% της πληροφορίας. Βλέπουμε καθαρά ότι το Cluster 0 (Κόκκινο χρώμα) αποτελεί έναν πολύ πυκνό, κεντρικό πυρήνα. Αντίθετα, το Cluster 1 (Γκρι χρώμα) αποτελείται από πολύ λιγότερα τμήματα, τα οποία είναι διάσπαρτα προς τα δεξιά. Ο αλγόριθμος, δηλαδή, ξεχώρισε τη μάζα των φυσιολογικών σχολών από τις εξαιρέσεις (outliers) που είχαν ακραία συμπεριφορά. Το τι ακριβώς σημαίνει ακραία συμπεριφορά αποτυπώνεται στο πώς άλλαξαν οι βάσεις αυτών των σχολών από χρονιά σε χρονιά.

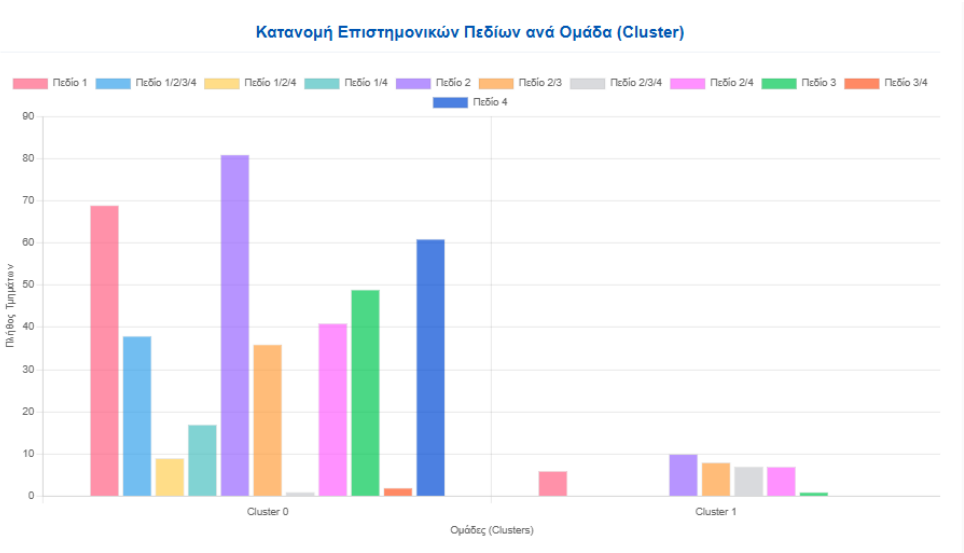


Εικόνα 50.Διάγραμμα παράλληλων συντεταγμένων των ετήσιων μεταβολών (Ward, k=2) για την περίοδο 2019-2025.

Η Εικόνα 50, είναι απόλυτα διαφωτιστική:

- **Cluster 0 (Κόκκινες γραμμές - Η Σταθερότητα):** Είναι ο κορμός του συστήματος. Οι γραμμές κινούνται ομαλά, πολύ κοντά στο μηδέν. Αυτά τα τμήματα απορρόφησαν τις αλλαγές (π.χ. καθιέρωση της ΕΒΕ) χωρίς να παρουσιάσουν χαοτικές αυξομειώσεις στα μόριά τους.
- **Cluster 1 (Γκρι γραμμές - Τα Βαθμολογικά Σοκ):** Εδώ συγκεντρώνονται τα τμήματα με τις ακραίες πτώσεις και εκτινάξεις (ζιγκ-ζαγκ). Βλέπουμε τεράστιες βυθίσεις στο diff_21_20 και στο diff_22_21 και απότομες ανόδους. Πρόκειται για τα τμήματα που ταρακουνήθηκαν βίαια από τις νέες συνθήκες εισαγωγής.

Από ποια επιστημονικά πεδία προέρχονται όμως αυτά τα τμήματα που υπέστησαν τα σοκ;

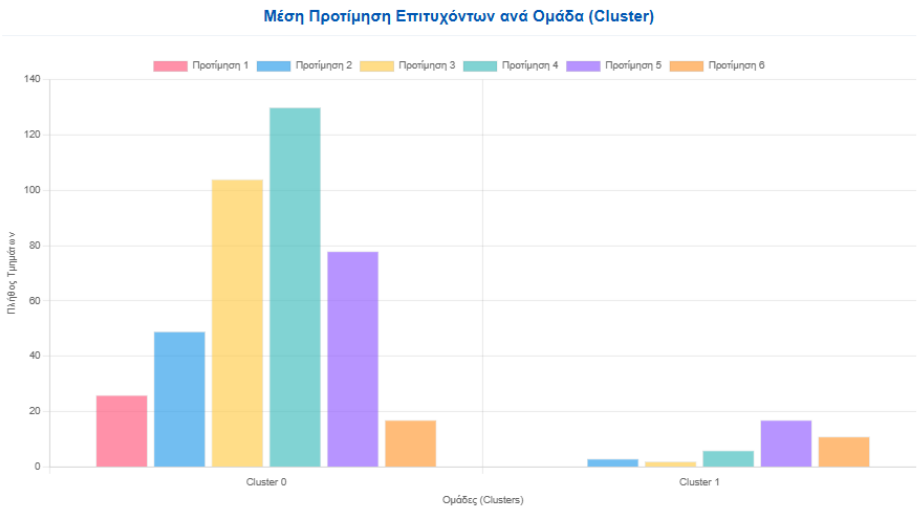


Εικόνα 51.Ραβδόγραμμα κατανομής των Επιστημονικών Πεδίων στις 2 συστάδες μεταβολών (Ward, 2019-2025).

Το ραβδόγραμμα της Εικόνας 51, μας δίνει την απάντηση:

- Το **Cluster 0** (η σταθερή ομάδα) περιλαμβάνει τη συντριπτική πλειοψηφία των σχολών από όλα τα πεδία (κυρίως Πεδίο 1 - ροζ, Πεδίο 2 - μωβ, Πεδίο 4 - μπλε).
- Αντίθετα, στο **Cluster 1** (η ασταθής ομάδα) υπάρχουν ελάχιστες σχολές, οι οποίες προέρχονται κυρίως από το 2ο Πεδίο (Θετικές Επιστήμες - μωβ) και τα κοινά τμήματα του 2ου/4ου και 2ου/3ου πεδίου. Η εφαρμογή της ΕΒΕ έπληξε (προκαλώντας αστάθεια) κυρίως τα περιφερειακά τμήματα των θετικών και τεχνολογικών επιστημών.

Πώς αντιμετωπίζουν οι υποψήφιοι αυτά τα ασταθή τμήματα στο μηχανογραφικό τους;

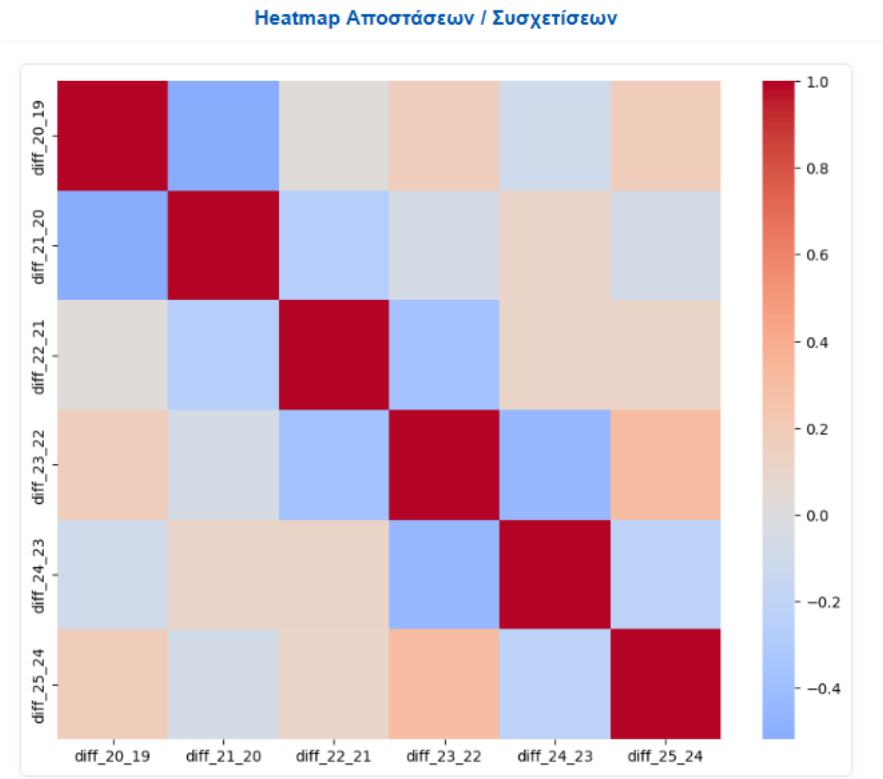


Εικόνα 52.Ραβδόγραμμα συχνότητας της σειράς προτίμησης των επιτυχόντων ανά συστάδα μεταβολών.

Το ραβδόγραμμα της Εικόνας 52, δείχνει τον κανόνα ότι η αστάθεια ισούται με την χαμηλή ζήτηση.

- Στο **Cluster 0** (σταθερά τμήματα), η κατανομή είναι φυσιολογική, με τις περισσότερες εισαγωγές να γίνονται στην 3η (κίτρινο) και 4η (τιρκουάζ) επιλογή.
- Στο **Cluster 1** (ασταθή τμήματα), οι πρώτες επιλογές απλώς δεν υπάρχουν. Η συντριπτική πλειονότητα των εισακτέων τα είχε δηλώσει ως **5η (μωβ) και 6η (πορτοκαλί) επιλογή**. Είναι ξεκάθαρα σχολές που λειτουργούν ως έσχατη λύση (backup) για τους μαθητές, γι' αυτό και οι βάσεις τους κατακρημνίζονται ή εκτινάσσονται ανάλογα με τη χρονιά.

Τέλος, βλέπουμε πώς σχετίζονται οι μεταβολές των διαφόρων ετών στην πρόσφατη εξαετία.



Εικόνα 53.Heatmap hierarchical των ετήσιων μεταβολών 2019-2025.

Ο χάρτης θερμότητας της Εικόνας 53, επιβεβαιώνει την έλλειψη γραμμικής συσχέτισης, καθώς κυριαρχούν τα ανοιχτά χρώματα (κοντά στο μηδέν) και το μπλε (αρνητική συσχέτιση). Το πιο χαρακτηριστικό παράδειγμα είναι η έντονη αρνητική συσχέτιση (μπλε χρώμα) ανάμεσα στο diff_21_20 και στο diff_20_19. Αυτό δείχνει τον μηχανισμό "αυτοδιόρθωσης" των βάσεων: μετά την απότομη πτώση που μπορεί να καταγραφεί μια χρονιά, οι υποψήφιοι της επόμενης χρονιάς στρέφονται μαζικά προς αυτές τις (φαινομενικά) "φθηνές" σε μόρια σχολές, προκαλώντας έτσι την τεχνητή άνοδό τους την επόμενη χρονιά.

Συμπερασματικά για τις σύγχρονες ετήσιες μεταβολές, η μέθοδος **Ward** επιλέχθηκε ως η βέλτιστη ιεραρχική προσέγγιση διότι αντιμετώπισε επιτυχώς τον έντονο θόρυβο της περιόδου. Οι υπόλοιποι αλγόριθμοι (Complete, Average, Single) παρήγαγαν παραπλανητικά υψηλούς δείκτες Silhouette (έως 0.831), επειδή περιορίστηκαν στο να απομονώσουν ελάχιστα τμήματα με ακραία σκαμπανεβάσματα λόγω της εισαγωγής της ΕΒΕ, στερούμενοι ερμηνευτικής αξίας. Η μέθοδος Ward, ελαχιστοποιώντας συστηματικά το σφάλμα (SSE 2.942.447.562), πέτυχε έναν ουσιαστικό διαχωρισμό, ομαδοποιώντας ορθολογικά τα τμήματα με βάση την πραγματική διαχρονική τους σταθερότητα και τη σειρά προτίμησης των υποψηφίων.

Κεφάλαιο 8ο: Πρόβλεψη Μελλοντικών Τιμών Βάσης Εισαγωγής

8.1 Σύνολο δεδομένων

Η ποιότητα, η αντιπροσωπευτικότητα και η δομή του συνόλου δεδομένων (dataset) αποτελούν τον θεμέλιο λίθο για την επιτυχή εκπαίδευση των μοντέλων μηχανικής μάθησης. Για τις ανάγκες της παρούσας έρευνας, πραγματοποιήθηκε συλλογή, ενοποίηση και συστηματική επεξεργασία ιστορικών στοιχείων που αφορούν τις Πανελλαδικές Εξετάσεις των Γενικών Λυκείων (ΓΕΛ) για ένα εκτενές χρονικό εύρος.

Το τελικό dataset οργανώνεται ανά κωδικό και όνομα τμήματος, πανεπιστημιακό ίδρυμα και έτος εξετάσεων. Κάθε διακριτή εγγραφή αντιπροσωπεύει τη συμπεριφορά και τα αποτελέσματα ενός συγκεκριμένου ακαδημαϊκού τμήματος σε ένα συγκεκριμένο έτος. Τα χαρακτηριστικά που απαρτίζουν το σύνολο δεδομένων και τροφοδοτήθηκαν στα μοντέλα κατηγοριοποιούνται σε τρεις βασικές ομάδες:

1. **Στοιχεία Ταυτότητας Τμήματος:** Περιλαμβάνουν το όνομα του πανεπιστημιακού τμήματος (tmhma), το ανώτατο εκπαιδευτικό ίδρυμα στο οποίο ανήκει (panepistimio), το έτος αναφοράς της εξέτασης (etos), καθώς και το επιστημονικό πεδίο ή τα πεδία στα οποία είναι ενταγμένο το τμήμα (epistimoniko_pedio), όπως το 1ο, 2ο, 3ο ή 4ο πεδίο.
2. **Δείκτες Ζήτησης και Προσφοράς:** Περιλαμβάνουν τον συνολικό αριθμό των εισακτέων που όρισε το Υπουργείο Παιδείας για το τμήμα τη δεδομένη χρονιά (positions), τον αριθμό των υποψηφίων που δήλωσαν το τμήμα ως πρώτη τους επιλογή στο μηχανογραφικό τους δελτίο (plithos_1on_protimiseon) καθώς και τη «Μέση Προτίμηση», έναν σύνθετο στατιστικό δείκτη που αποτυπώνει τη μέση σειρά προτίμησης του τμήματος στο σύνολο των υποβληθέντων μηχανογραφικών.
3. **Ιστορικά Βαθμολογικά Δεδομένα & Χρονικές Υστερήσεις (Time Lags):** Περιλαμβάνουν τη βάση εισαγωγής (τα μόρια του τελευταίου εισαχθέντος) του τμήματος κατά το τρέχον έτος (vasi), καθώς και τις βάσεις εισαγωγής των προηγούμενων ετών σε βάθος τριετίας (vasi_minus_1 για το έτος t-1, vasi_minus_2 για το έτος t-2, και vasi_minus_3 για το έτος t-3). Παράλληλα, ενσωματώθηκε ο βαθμός σε μόρια του πρώτου επιτυχόντος στο τμήμα (vathmos_protou), ο οποίος λειτουργεί ως δείκτης προσέλκυσης αριστούχων υποψηφίων.

Ανάλογα με τη φύση του πειράματος, ορίστηκαν δύο διαφορετικές μεταβλητές στόχοι (target variables). Για το πρόβλημα της παλινδρόμησης (regression), ως στόχος ορίστηκε αρχικά η μεταβλητή target_vasi_plus_1 (η πραγματική βάση εισαγωγής του επόμενου έτους t+1), η οποία σε μεταγενέστερο στάδιο βελτιστοποίησης μετασχηματίστηκε στη μεταβλητή target_diff, η οποία ισούται με τη διαφορική μεταβολή των μορίων (target_vasi_plus_1 - vasi). Για το πρόβλημα της κατηγοριοποίησης (classification), η μεταβλητή target_diff μετατράπηκε σε διακριτές κλάσεις που αντιπροσωπεύουν την τάση κίνησης της βάσης (π.χ. "Άνοδος", "Πτώση", "Σταθερότητα").

Πριν από την εισαγωγή των δεδομένων στους αλγορίθμους, εφαρμόστηκαν αυστηρές τεχνικές προεπεξεργασίας (data preprocessing). Επειδή ορισμένα τμήματα ήταν νέα ή είχαν υποστεί μετονομασίες, εμφάνιζαν κενές τιμές (NaN) στα ιστορικά τους χαρακτηριστικά. Πραγματοποιήθηκε απαλοιφή των εγγραφών αυτών (dropna), εξασφαλίζοντας ότι τα μοντέλα θα εκπαιδευτούν αποκλειστικά σε τμήματα με πλήρες ιστορικό βάθος τουλάχιστον τριών ετών. Επιπλέον, χαρακτηριστικά όπως το πλήθος των πρώτων προτιμήσεων και οι θέσεις εισακτέων μετατράπηκαν σε καθαρούς ακέραιους αριθμούς (int), ενώ εφαρμόστηκε κατάλληλη κωδικοποίηση χαρακτήρων για τη σωστή διατήρηση των ελληνικών ονομάτων των σχολών και των πανεπιστημίων.

8.1.1 Πειραματική Διαδικασία

Η πειραματική διαδικασία σχεδιάστηκε με βάση τη λογική του χρονικού διαχωρισμού (Time-based Train/Test Split), ώστε να αποφευχθεί η διαρροή δεδομένων (data leakage) και τα μοντέλα να

αξιολογηθούν σε πραγματικές συνθήκες πρόβλεψης μελλοντικών ετών. Το σύνολο δεδομένων διαχωρίστηκε ως εξής:

- **Σύνολο Εκπαίδευσης (Train Set):** Περιλαμβάνει όλα τα ιστορικά δεδομένα των τμημάτων έως και το έτος 2023.
- **Σύνολο Αξιολόγησης (Test Set):** Περιλαμβάνει τα δεδομένα του έτους 2024, με σκοπό την πρόβλεψη και επαλήθευση των πραγματικών βάσεων του 2025.

Στο πλαίσιο της έρευνας πραγματοποιήθηκαν δύο βασικά πειράματα:

1. **Πείραμα Παλινδρόμησης (Regression):** Αντικείμενο του συγκεκριμένου πειράματος αποτελεί ο προσδιορισμός της ακριβούς συνεχούς τιμής της βάσης εισαγωγής (ή εναλλακτικά της ετήσιας διαφορικής μεταβολής των μορίων) για κάθε ακαδημαϊκό τμήμα. Προς κατεύθυνση αυτή, επιστρατεύθηκαν οι αλγόριθμοι XGBoost Regressor, Random Forest Regressor και K-Neighbors Regressor (KNN). Η αξιολόγηση της προβλεπτικής ικανότητας και της ακρίβειας των μοντέλων βασίστηκε στη μετρική του Μέσου Απόλυτου Σφάλματος (Mean Absolute Error - MAE), η οποία ποσοτικοποιεί τη μέση απόκλιση της εκτιμώμενης τιμής από την πραγματική, εκπεφρασμένη σε μόρια.
2. **Πείραμα Κατηγοριοποίησης (Classification):** Στοχεύει στην πρόβλεψη της γενικής τάσης μεταβολής των βάσεων εισαγωγής. Στο πλαίσιο αυτό, το πρόβλημα μετασχηματίζεται από τον προσδιορισμό μιας ακριβούς τιμής στην ταξινόμηση των τμημάτων σε διακριτές κλάσεις (π.χ. «Άνοδος», «Πτώση» ή «Σταθερότητα»), με βάση προκαθορισμένα όρια διακύμανσης των μορίων.

8.2 Πειραματικά αποτελέσματα regression

Στο πείραμα της παλινδρόμησης αξιολογήθηκαν τρεις διαφορετικές προσεγγίσεις μηχανικής μάθησης με στόχο την απευθείας πρόβλεψη της απόλυτης τιμής της βάσης εισαγωγής για το έτος αξιολόγησης. Για την εκπαίδευση των μοντέλων χρησιμοποιήθηκαν τα ιστορικά χαρακτηριστικά των προηγούμενων ετών, ο βαθμός του πρώτου εισαχθέντος, καθώς και τα στοιχεία προσφοράς και ζήτησης (θέσεις και πρώτες προτιμήσεις υποψηφίων).

Ως μέτρο σύγκρισης (Baseline) για την αξιολόγηση της απόδοσης των αλγορίθμων ορίστηκε η **πραγματική μέση διαφορά των βάσεων μεταξύ των ετών 2024 και 2025**, η οποία ανήλθε στα **424.4 μόρια**. Η τιμή αυτή αντιπροσωπεύει τη φυσική στατιστική διακύμανση του συστήματος των εξετάσεων ανάμεσα στα δύο αυτά έτη.

Τα αποτελέσματα του Μέσου Απόλυτου Σφάλματος (MAE) σε μόρια για κάθε προσέγγιση αποτυπώνονται στον παρακάτω πίνακα, αναδεικνύοντας τις 4 χαρακτηριστικές τιμές της πειραματικής διαδικασίας:

Αλγόριθμος / Μέθοδος	**MAE (Μόρια)**
K-Neighbors Regressor (KNN)	535.1 μόρια
XGBoost Regressor	519.1 μόρια
Random Forest Regressor	508.5 μόρια
Πραγματική Διαφορά 2024–2025 (Baseline)	424.4 μόρια

Πίνακας 1.Σύγκριση Μέσου Απόλυτου Σφάλματος (MAE) Μοντέλων Regression

Από τη συγκριτική ανάλυση των αποτελεσμάτων προκύπτει ότι κατά την απευθείας πρόβλεψη του τελικού μεγέθους της βάσης, ο αλγόριθμος **Random Forest** παρουσίασε την καλύτερη προβλεπτική ικανότητα μεταξύ των μοντέλων, επιτυγχάνοντας το χαμηλότερο Μέσο Απόλυτο Σφάλμα με 508.5 μόρια. Ο αλγόριθμος **XGBoost** ακολούθησε με MAE 519.1 μόρια, ενώ ο **KNN** εμφάνισε τη χαμηλότερη ακρίβεια με σφάλμα 535.1 μόρια.

Ένα ιδιαίτερα αξιοσημείωτο εύρημα είναι ότι και τα τρία μοντέλα μηχανικής μάθησης παρουσίασαν σφάλμα πρόβλεψης (508.5 έως 535.1 μόρια) το οποίο είναι υψηλότερο από την πραγματική μέση διαφορά των δύο ετών (424.4 μόρια). Η συμπεριφορά αυτή υποδηλώνει τη μεγάλη δυσκολία που αντιμετωπίζουν οι αλγόριθμοι όταν καλούνται να μάθουν το ακριβές αριθμητικό μέγεθος της βάσης. Αυτό συμβαίνει διότι οι απόλυτες τιμές των μορίων παρουσιάζουν τεράστιο εύρος (από σχολές των 9.000 μορίων έως σχολές των 19.000 μορίων), με αποτέλεσμα ο «θόρυβος» των δεδομένων να δυσχεραίνει την εκπαίδευση.

Παρόλα αυτά, η υπεροχή των δενδρικών μοντέλων (Random Forest και XGBoost) έναντι του KNN επιβεβαιώνει ότι οι αρχιτεκτονικές που βασίζονται σε δέντρα απόφασης μπορούν να διαχειριστούν πιο αποτελεσματικά τις μη γραμμικές σχέσεις μεταξύ των θέσεων εισακτέων και των προτιμήσεων των υποψηφίων.

8.3 Πειραματικά αποτελέσματα κατηγοριοποίησης

Στο δεύτερο σκέλος της πειραματικής διαδικασίας, το πρόβλημα της πρόβλεψης των βάσεων εισαγωγής μετασχηματίστηκε από συνεχή παλινδρόμηση σε διακριτή ταξινόμηση (classification). Στόχος των μοντέλων είναι η πρόβλεψη της τάσης μεταβολής των μορίων ανά τμήμα. Για την ολοκληρωμένη διερεύνηση του προβλήματος, πραγματοποιήθηκαν δύο διαφορετικά σενάρια κατηγοριοποίησης:

1. **Ταξινόμηση 3 Κλάσεων (3-Class Setup):** Οι βάσεις κατηγοριοποιούνται με βάση τη μεταβολή τους σε *«Πτώση»*, *«Σταθερότητα»* και *«Άνοδος»*.
2. **Ταξινόμηση 5 Κλάσεων (5-Class Setup):** Για βαθύτερη ανάλυση της έντασης της μεταβολής, οι κλάσεις επεκτείνονται σε *«Μεγάλη Πτώση»*, *«Μικρή Πτώση»*, *«Σταθερότητα»*, *«Μικρή Άνοδος»* και *«Μεγάλη Άνοδος»*.

Για την αξιολόγηση των αλγορίθμων Random Forest, XGBoost και K-Nearest Neighbors (kNN) χρησιμοποιήθηκαν οι μετρικές της Συνολικής Ακρίβειας (Accuracy), της Πιστότητας (Precision), της Ανάκλησης (Recall) και του F1-Score (weighted average). Στο πρώτο πείραμα, τα μοντέλα κλήθηκαν να διαχωρίσουν τα τμήματα στις τρεις βασικές κατευθύνσεις. Τα συγκεντρωτικά αποτελέσματα από τα classification reports των μοντέλων αποτυπώνονται στον παρακάτω πίνακα.

Αλγόριθμος Ταξινόμησης	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest Classifier	61.42%	60.10%	61.42%	59.83%
XGBoost Classifier	59.39%	58.74%	59.39%	58.83%
K-Neighbors Classifier (kNN)	53.04%	52.88%	53.04%	52.92%

Πίνακας 2.Μετρικές Αξιολόγησης Μοντέλων για 3 Κλάσεις

Όπως προκύπτει από τα πειραματικά αποτελέσματα, ο αλγόριθμος Random Forest πέτυχε την κορυφαία επίδοση με συνολική ακρίβεια 61.42%, παρουσιάζοντας την καλύτερη ισορροπία μεταξύ Precision και Recall. Ο XGBoost ακολούθησε με μικρή διαφορά, επιτυγχάνοντας Accuracy 59.39%. Αντίθετα, ο αλγόριθμος kNN εμφάνισε σημαντική υστέρηση, με την ακρίβειά του να περιορίζεται στο 53.04%.

Η ανάλυση των επιμέρους κλάσεων (per-class metrics) έδειξε ότι και τα τρία μοντέλα παρουσιάζουν εξαιρετικά υψηλά ποσοστά ευστοχίας στην αναγνώριση των τμημάτων που θα σημειώσουν *«Άνοδο»* και *«Πτώση»*. Η μεγαλύτερη πρόκληση εντοπίστηκε στην κλάση της

«Σταθερότητας». Αυτό συμβαίνει διότι η σταθερότητα ορίζεται σε ένα πολύ στενό αριθμητικό εύρος μορίων, με αποτέλεσμα οι οριακές αυξομειώσεις των βάσεων να προκαλούν εσφαλμένες ταξινομήσεις προς τις γειτονικές κλάσεις, μειώνοντας έτσι τη συνολική επίδοση.

Προκειμένου να ελεγχθεί αν οι αλγόριθμοι μπορούν να διακρίνουν την ένταση της μεταβολής των βάσεων, το πρόβλημα επεκτάθηκε σε 5 κλάσεις. Τα αποτελέσματα της συγκεκριμένης διάταξης παρουσιάζονται στον επόμενο πίνακα.

Αλγόριθμος Ταξινόμησης	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest Classifier	44.41%	42.66%	44.41%	42.69%
XGBoost Classifier	43.14%	42.50%	43.14%	42.62%
K-Neighbors Classifier (kNN)	37.81%	37.91%	37.81%	37.82%

Πίνακας 3. Μετρικές Αξιολόγησης Μοντέλων για 5 Κλάσεις

Η εισαγωγή 5 κλάσεων αύξησε κατακόρυφα την πολυπλοκότητα του προβλήματος, γεγονός που αποτυπώνεται στην πτώση της συνολικής ορθότητας όλων των μοντέλων. Το Random Forest διατηρεί τα πρωτεία της ακρίβειας με ποσοστό 44.41%, με τον XGBoost να έπεται σε απόσταση αναπνοής (43.14%), ενώ ο kNN περιορίζεται στο 37.81%.

Η πτώση της ακρίβειας κατά περίπου 17% σε σχέση με το πείραμα των 3 κλάσεων κρίνεται αναμενόμενη και δικαιολογείται απόλυτα από τη φύση των Πανελλαδικών Εξετάσεων. Ο διαχωρισμός, για παράδειγμα, μιας «Μικρής Άνδου» από μια «Μεγάλη Άνοδο» εξαρτάται από εξαιρετικά ευαίσθητες ισορροπίες, όπως οι τοπικές προτιμήσεις των υποψηφίων ανά γεωγραφική περιφέρεια και ο ακριβής αριθμός των αριστούχων ανά επιστημονικό πεδίο.

Ωστόσο, η μελέτη των πινάκων σύγχυσης (confusion matrices) των μοντέλων αποκάλυψε μια πολύ θετική συμπεριφορά: τα σφάλματα των μοντέλων Random Forest και XGBoost δεν είναι χαοτικά. Στη συντριπτική τους πλειονότητα, όταν οι αλγόριθμοι αποτυγχάνουν, ταξινομούν την εγγραφή στην αμέσως διπλανή κλάση (π.χ. προβλέπουν «Μικρή Άνοδο» αντί για «Μεγάλη Άνοδο») και σχεδόν ποτέ δεν υποπίπτουν σε διαμετρικά αντίθετα σφάλματα (όπως η πρόβλεψη «Μεγάλης Πτώσης» σε τμήμα που τελικά σημείωσε «Μεγάλη Άνοδο»). Αυτό επιβεβαιώνει ότι τα δενδρικά μοντέλα έχουν καταφέρει να αποκτήσουν μια ισχυρή μαθηματική αντίληψη της υποκείμενης δομής των δεδομένων.

8.4 Συζήτηση

Η συνολική αποτίμηση των αποτελεσμάτων, τόσο στο επίπεδο της παλινδρόμησης όσο και στο επίπεδο της κατηγοριοποίησης, αναδεικνύει τη σημασία των δεδομένων συμπεριφοράς των υποψηφίων (όπως οι πρώτες προτιμήσεις των μηχανογραφικών δελτίων) έναντι των καθαρά ιστορικών βαθμολογικών στοιχείων. Η υπεροχή των δενδρικών μοντέλων (Random Forest και XGBoost) έναντι του αλγορίθμου KNN επιβεβαίωσε ότι οι αρχιτεκτονικές που βασίζονται σε δέντρα απόφασης μπορούν να διαχειριστούν πολύ πιο αποτελεσματικά τις μη γραμμικές σχέσεις και τις κλιμακωτές διαφορές μεταξύ των μορίων των σχολών.

Ωστόσο, εξετάζοντας τα αποτελέσματα υπό το πρίσμα της πρακτικής τους εφαρμογής σε ένα πραγματικό περιβάλλον, προκύπτει ένα κρίσιμο συμπέρασμα. Παρόλο που οι αλγόριθμοι επέδειξαν ικανοποιητική συμπεριφορά κατά την πειραματική διαδικασία, τα τελικά αποτελέσματα κρίθηκαν ανεπαρκή για την ασφαλή ενσωμάτωση του μοντέλου πρόβλεψης στην επίσημη ιστοσελίδα/εφαρμογή.

Η απόφαση να μην συμπεριληφθεί η λειτουργία αυτόματης πρόβλεψης στην παραγωγική σελίδα βασίζεται σε τρεις κεντρικούς άξονες:

1. **Το Μέγεθος του Σφάλματος στην Παλινδρόμηση:** Ένα Μέσο Απόλυτο Σφάλμα (MAE) της τάξης των 508 έως 535 μορίων στην απευθείας πρόβλεψη της βάσης θεωρείται ιδιαίτερα υψηλό για τα δεδομένα των Πανελλαδικών Εξετάσεων. Στην πράξη, μια απόκλιση 500 μορίων μπορεί να κρίνει την επιτυχία ή την αποτυχία ενός υποψηφίου σε μια σχολή. Η προβολή τέτοιων προβλέψεων στη σελίδα θα μπορούσε να αποβεί παραπλανητική για τους μαθητές και τους γονείς κατά τη συμπλήρωση του μηχανογραφικού τους δελτίου.

2. **Τα Όρια της Κατηγοριοποίησης 5 Κλάσεων:** Στο σενάριο όπου επιχειρήθηκε η λεπτομερής πρόβλεψη της τάσης (Μεγάλη/Μικρή Άνοδος ή Πτώση), η ακρίβεια του κορυφαίου μοντέλου περιορίστηκε στο 44.41%. Ένα ποσοστό ευστοχίας κάτω από 50% εισάγει υψηλό επίπεδο αβεβαιότητας και ρίσκου, το οποίο αντιτίθεται στα πρότυπα αξιοπιστίας που πρέπει να διέπουν μια εκπαιδευτική διαδικτυακή πλατφόρμα.

3. **Η Ύπαρξη Εξωτερικών, Μη Προβλέψιμων Παραγόντων:** Το σύστημα διαμόρφωσης των βάσεων στην Ελλάδα επηρεάζεται οριζόντια από αστάθμητους παράγοντες, όπως ο αιφνίδιος βαθμός δυσκολίας των θεμάτων κάθε έτους και οι αλλαγές στο θεσμικό πλαίσιο (π.χ. Ελάχιστη Βάση Εισαγωγής). Καθώς οι παράγοντες αυτοί δεν αποτυπώνονται στο ιστορικό σύνολο δεδομένων, κανένα μοντέλο μηχανικής μάθησης δεν μπορεί να τους προβλέψει εκ των προτέρων, με αποτέλεσμα οι προβλέψεις για το μέλλον (όπως το έτος 2025) να ενέχουν ρίσκο αστοχίας.

Συμπερασματικά, η παρούσα έρευνα πέτυχε τον επιστημονικό της στόχο, ο οποίος ήταν η χαρτογράφηση, η ανάλυση και η δοκιμή των ορίων της Μηχανικής Μάθησης πάνω στα δεδομένα των βάσεων εισαγωγής. Ωστόσο, η μεταφορά αυτών των αποτελεσμάτων σε μια δημόσια προσβάσιμη ιστοσελίδα απορρίφθηκε, καθώς προτεραιότητα δόθηκε στην εγκυρότητα της εφαρμογής. Η εργασία αυτή θέτει τις βάσεις για μελλοντική έρευνα, η οποία, με την πιθανή ενσωμάτωση δειγματοληπτικών δεδομένων από τις επιδόσεις των μαθητών σε πραγματικό χρόνο (π.χ. κατά τη διάρκεια της βαθμολόγησης), ίσως καταφέρει να γεφυρώσει το χάσμα και να επιτρέψει μια ασφαλή διαδικτυακή ενσωμάτωση στο μέλλον.

Κεφάλαιο 9ο: Συμπεράσματα και Μελλοντικές Επεκτάσεις

Η παρούσα εργασία είχε ως στόχο την αναβάθμιση και επέκταση της εφαρμογής vaseisapp, τόσο σε επίπεδο λειτουργικότητας όσο και σε επίπεδο ανάλυσης δεδομένων. Αρχικά, υλοποιήθηκε ένας πλήρως αυτοματοποιημένος μηχανισμός εισαγωγής δεδομένων, ο οποίος εξαλείφει την ανάγκη χειροκίνητης παρέμβασης και μειώνει σημαντικά τον χρόνο ενημέρωσης της βάσης. Η αυτοματοποίηση αυτή αποτελεί ουσιαστική βελτίωση σε σχέση με την προηγούμενη διαδικασία και συμβάλλει στη βιωσιμότητα και επεκτασιμότητα της εφαρμογής.

Παράλληλα, πραγματοποιήθηκε μια εκτενής πειραματική μελέτη πάνω σε αλγορίθμους Μηχανικής Μάθησης, με στόχο την κατανόηση των τάσεων των βάσεων εισαγωγής και την πρόβλεψη μελλοντικών τιμών. Τα αποτελέσματα της συσταδοποίησης ανέδειξαν ομάδες τμημάτων με κοινή συμπεριφορά διαχρονικά, προσφέροντας μια νέα οπτική στην ανάλυση των χρονοσειρών. Η χρήση τόσο του k-means όσο και των ιεραρχικών μεθόδων έδειξε ότι οι βάσεις εισαγωγής παρουσιάζουν δομές που μπορούν να αξιοποιηθούν για περαιτέρω μελέτη και ταξινόμηση. Αντίστοιχα, τα πειράματα παλινδρόμησης έδειξαν ότι μοντέλα όπως το Random Forest και το XGBoost μπορούν να αποδώσουν ικανοποιητικές προβλέψεις, αν και η μεταβλητότητα των βάσεων και οι εξωγενείς παράγοντες περιορίζουν την ακρίβεια των αποτελεσμάτων.

Επιπλέον, η ενσωμάτωση των νέων λειτουργιών στο frontend, όπως η βελτιωμένη παρουσίαση δεδομένων και η δυνατότητα οπτικοποίησης των clusters, ενισχύει σημαντικά την εμπειρία χρήσης και καθιστά την εφαρμογή πιο λειτουργική και κατανοητή για τον τελικό χρήστη. Η εργασία απέδειξε ότι ο συνδυασμός αυτοματοποιημένων διαδικασιών, σύγχρονων τεχνικών ανάλυσης και βελτιωμένης διεπαφής μπορεί να μεταμορφώσει μια εφαρμογή σε ένα ολοκληρωμένο πληροφοριακό εργαλείο.

Παρά τα θετικά αποτελέσματα, υπάρχουν ορισμένοι περιορισμοί που αξίζει να αναφερθούν. Η πρόβλεψη των βάσεων επηρεάζεται από παράγοντες που δεν αποτυπώνονται στα διαθέσιμα δεδομένα, όπως αλλαγές στον τρόπο εξέτασης, μεταβολές στις προτιμήσεις των υποψηφίων ή κοινωνικοοικονομικές συνθήκες. Επιπλέον, η συσταδοποίηση βασίζεται αποκλειστικά σε χρονοσειρές βάσεων και όχι σε πρόσθετα χαρακτηριστικά των τμημάτων, κάτι που θα μπορούσε να εμπλουτίσει τα αποτελέσματα.

Μελλοντικά, η εργασία μπορεί να επεκταθεί σε πολλαπλές κατευθύνσεις. Μια σημαντική προοπτική είναι η ενσωμάτωση πιο προηγμένων μοντέλων πρόβλεψης, όπως νευρωνικά δίκτυα για χρονοσειρές (LSTM, GRU). Επίσης, η αξιοποίηση επιπλέον χαρακτηριστικών (π.χ. αριθμός εισακτέων, γεωγραφική θέση, επαγγελματικά δικαιώματα) θα μπορούσε να οδηγήσει σε πιο ακριβή μοντέλα συσταδοποίησης και παλινδρόμησης. Τέλος, η περαιτέρω βελτίωση της διεπαφής, η προσθήκη διαδραστικών εργαλείων ανάλυσης και η παροχή εξατομικευμένων προβλέψεων για υποψηφίους αποτελούν ενδιαφέρουσες κατευθύνσεις για την εξέλιξη της εφαρμογής.

Συνολικά, η εργασία συνέβαλε ουσιαστικά τόσο στη λειτουργική αναβάθμιση του vaseisapp όσο και στην επιστημονική μελέτη των βάσεων εισαγωγής, προσφέροντας ένα ισχυρό υπόβαθρο για μελλοντική έρευνα και ανάπτυξη.

Βιβλιογραφία

[1] V. Singrodia, A. Mitra, and S. Paul, "A Review on Web Scrapping and its Applications," in *2019 International Conference on Computer Communication and Informatics (ICCCI)*, 2019, pp. 1–6. doi: 10.1109/ICCCI.2019.8821809.

[2] S. GOEL, M. BANSAL, A. K. SRIVASTAVA, and N. ARORA, "Web Crawling-based Search Engine using Python," in *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2019, pp. 436–438. doi: 10.1109/ICECA.2019.8821866.

[3] C. Bhatt, Gaitri, D. Kumar, R. Chauhan, A. Vishvakarma, and T. Singh, "Web Scraping: Huge Data Collection from Web," in *2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology, ICSEIET 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 375–378. doi: 10.1109/ICSEIET58677.2023.10303037.

[4] R. Chauhan, A. Negi, and M. Manchanda, "An Extensive Review on Web Scraping Technique using Python," in *Proceedings of the 2023 2nd International Conference on Augmented Intelligence and Sustainable Systems, ICAISS 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1134–1138. doi: 10.1109/ICAISS58487.2023.10250745.

[5] R. T. Narayanan, "Novice Programmer to New-Age Application Developer: What Makes Python their First Choice?"

[6] K. Nongthombam and D. Sharma, "Data Analysis using Python." [Online]. Available: www.ijert.org

[7] S. Pant, N. Yadav, Milan, M. Sharma, Y. Bedi, and A. Raturi, "Web Scraping Using Beautiful Soup," in *2024 International Conference on Knowledge Engineering and Communication Systems, ICKECS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICKECS61492.2024.10617017.

[8] L. Zhao and H. Sun, "Design and Implementation of Web Crawler System Based on Python Technology," in *Proceedings - 2023 3rd International Signal Processing, Communications and Engineering Management Conference, ISPCEM 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 390–395. doi: 10.1109/ISPCEM60569.2023.00076.

[9] V. Bisht, R. Choyal, A. S. Negi, and E. K. Singh, "Utilizing Python for Web Scraping and Incremental Data Extraction," in *2nd International Conference on Automation, Computing and Renewable Systems, ICACRS 2023 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1450–1455. doi: 10.1109/ICACRS58579.2023.10404702.

[10] P. Bhardwaj, C. Choudhury, and P. Batra, "Automating Data Analysis with Python: A Comparative Study of Popular Libraries and their Application," in *Proceedings - International Conference on Technological Advancements in Computational Sciences, ICTACS 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1243–1248. doi: 10.1109/ICTACS59847.2023.10390032.

[11] Z. Lu, P. Wan, H. Zhu, and Z. Chen, "Research on Network Virus Log Data Analysis and Processing Based on Pandas," in *Proceedings of 2024 IEEE 4th International Conference on Information Technology, Big Data and Artificial Intelligence, ICIBA 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 762–766. doi: 10.1109/ICIBA62489.2024.10868996.

[12] H. Ye, W. Chen, W. Dou, G. Wu, and J. Wei, “Knowledge-Based Environment Dependency Inference for Python Programs,” in *Proceedings - International Conference on Software Engineering*, IEEE Computer Society, Jul. 2022, pp. 1245–1256. doi: 10.1145/3510003.3510127.

[13] F. A. Riski, N. Selviandro, and M. Adrian, “Implementation of Web Scraping on Job Vacancy Sites Using Regular Expression Method,” in *2022 1st International Conference on Software Engineering and Information Technology, ICoSEIT 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 204–209. doi: 10.1109/ICoSEIT55604.2022.10029964.

[14] J. Pragathi and D. Sagar, “A Design and Analysis of Customer Characteristics-based Classification using Intelligent Algorithm,” in *2024 IEEE 9th International Conference for Convergence in Technology, I2CT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/I2CT61223.2024.10543884.

[15] *ASPLOS XVI : sixteenth International Conference on Architectural Support for Programming Languages and Operating Systems, March 5-11, 2011, Newport Beach, CA, USA*. ACM Press, 2011.

[16] S. Shapiro, “Pattern-Based File and Data Access with Python Glob: A Comprehensive Guide for Computational Research,” Sep. 2025, [Online]. Available: http://arxiv.org/abs/2509.08843

[17] C. Costa, P. K. Chrysanthis, M. Costa, E. Stavrakis, and N. Nicolaou, “Towards a Signature Based Compression Technique for Big Data Storage,” in *Proceedings - 2023 IEEE 39th International Conference on Data Engineering Workshops, ICDEW 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 100–104. doi: 10.1109/ICDEW58674.2023.00022.

[18] J. Borrow, P. La Plante, J. Aguirre, and P. K. G. Williams, “Making Research Data Flow With Python,” in *Proceedings of the 23rd Python in Science Conference*, SciPy, Jul. 2024, pp. 236–246. doi: 10.25080/hwga5253.

[19] S. Vasavi, P. D. L. Nikhita Sri, and P. V. Sai Krishna, “GUI-Enabled Boundary Regularization System for Urban Buildings Using the Tkinter,” in *Proceedings - 2nd IEEE International Conference on Device Intelligence, Computing and Communication Technologies, DICCT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 424–429. doi: 10.1109/DICCT61038.2024.10532910.

[20] J. Wahyudi, M. Asbari, I. Sasono, T. Pramono, and D. Novitasari, “Database Management in MYSQL,” 2022.

[21] S. Sotnik, V. Manakov, and V. Lyashenko, “Overview: PHP and MySQL Features for Creating Modern Web Projects,” 2023. [Online]. Available: www.ijeais.org/ijaisr

[22] R. Levering and M. Cutler, “The Portrait of a Common HTML Web Page,” 2006.

[23] R. Shankar Upadhayay and V. Kumar Barodiya, “A Comprehensive Review of HTML5 and CSS3: Advancements, Features, and Implications,” 2018.

[24] A. Kumar Ratha, S. Sahu, and P. Meher, “HTML5 in Web Development: A New Approach,” *International Research Journal of Engineering and Technology*, [Online]. Available: www.irjet.net

[25] U. Kaushal, G. Singh, and T. Parashar, “Responsive Webpage Using HTML CSS,” in *International Conference on Cyber Resilience, ICCR 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICCR56254.2022.9995922.

[26] R. Tulsyan, P. Shukla, A. Kumar, and T. Singh, “The Impact of JavaScript Frameworks on Website Performance and User Experience”, doi: 10.1109/ICBDML.2024.51.

[27] G. Dudek, “STD: A Seasonal-Trend-Dispersion Decomposition of Time Series,” *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 10, pp. 10339–10350, Oct. 2023, doi: 10.1109/TKDE.2023.3268125.

[28] B. Cai, G. Huang, N. Samadiani, G. Li, and C. H. Chi, “Efficient Time Series Clustering by Minimizing Dynamic Time Warping Utilization,” *IEEE Access*, vol. 9, pp. 46589–46599, 2021, doi: 10.1109/ACCESS.2021.3067833.

[29] E. Umargono, J. E. Suseno, and V. G. S. K., “K-Means Clustering Optimization using the Elbow Method and Early Centroid Determination Based-on Mean and Median,” Scitepress, Jul. 2020, pp. 234–240. doi: 10.5220/0009908402340240.

[30] O. Eric U. and O. Michael O., “Overview of Agglomerative Hierarchical Clustering Methods,” *British Journal of Computer, Networking and Information Technology*, vol. 7, no. 2, pp. 14–23, Jun. 2024, doi: 10.52589/bjcnit-cv9poogw.

[31] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, “Random Forests and Decision Trees,” 2012. [Online]. Available: www.IJCSI.org

[32] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[33] A. K. Albohali, A. Alshareef, and M. Shantal, “Machine Learning Approach to Predict Student Performance Based on Previous Records,” in *2024 International Conference on Computer and Applications, ICCA 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICCA62237.2024.10927802.