ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΕΛΛΑΔΟΣ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Real-Time Air Quality Forecasting & Monitoring

«Εικόνα»

Φοιτητής:
Κωνσταντίνος Κατσαούνης
Αριθμός Μητρώου: it 185435

Επιβλέπων:
Χαράλαμπος Μπράτσας

07 Φεβρουαρίου 2026

Real-Time Air Quality Forecasting & Monitoring
Student's full name Konstantinos Katsaounis
Supervisor's full name Charalampos Bratsas
Date of undertaking 15-10-2025
Date of completion 07-02-2026

Περίληψη

Στη περιοχή της Θεσσαλονίκης όπως και σε πολλές περιοχές της Ελλάδας, η ρύπανση της ατμόσφαιρας συνιστά μία από τις σημαντικότερες απειλές για τη δημόσια υγεία, καθώς τα μορφολογικά χαρακτηριστικά των λεγόμενων «αστικών φαραγγιών» και η τοπογραφική διαμόρφωση της περιοχής δημιουργούν συνθήκες που δημιουργούν τη συγκέντρωση επιβλαβών ρυπαντών. Η διπλωματική αυτή εργασία αναπτύσσει ένα εξελιγμένο σύστημα προγνωστικής αξιολόγησης της ατμοσφαιρικής ποιότητας, το οποίο προκύπτει ως αναβαθμισμένη συνέχεια της υφιστάμενης πλατφόρμας «AirQ». Σε αντίθεση με την πρωτότυπη εκδοχή που χρησιμοποιούσε το εργαλείο «Prophet» για τη πρόβλεψη χρονοσειρών, η τρέχουσα μελέτη υιοθετεί μια αισθητά πιο περίπλοκη και αποτελεσματική αρχιτεκτονική. Συγκεκριμένα, χρησιμοποιείται ο αλγόριθμος XGBoost και μια πλήρως αυτοματοποιημένη διαδικασία MLOps, αντικαθιστώντας τις παλαιότερες προσεγγίσεις.

Η συγκεκριμένη μεθοδολογία ενσωματώνει δεδομένα τηλεπισκόπησης που συλλέγονται από το προηγμένο όργανο TROPOMI του δορυφόρου Sentinel-5P από το πρόγραμμα Copernicus, τα οποία συνδυάζονται με μετεωρολογικές πληροφορίες προερχόμενες από την πλατφόρμα Open Meteo. Το σύστημα λειτουργεί μέσω αυτοματοποιημένων διαδικασιών ETL, οι οποίες διαχειρίζονται τη λήψη και την γεωχωρική κατεργασία αρχείων μορφής NetCDF. Επιπρόσθετα, αξιοποιείται το περιβάλλον Optuna για την βελτιστοποίηση των υπερπαραμέτρων μέσω τεχνικών Μπεϋζιανής αναζήτησης. Η διαχείριση των μοντέλων και η καταγραφή των πειραματικών δοκιμών επιτυγχάνεται με τη χρήση του εργαλείου MLflow, εξασφαλίζοντας την λειτουργικότητα και την αναπαραγωγιμότητα όλου εγχειρήματος.

Η τεχνική υλοποίηση βασίζεται στην αρχιτεκτονική containers/Docker, ενσωματώνοντας όλο το βασικό πυρήνα του κώδικα στο backend και μία διαδραστική διεπαφή χρήστη στον κώδικα του frontend που αναπτύχθηκε με το πλαίσιο Streamlit.Από την αξιολόγηση του συστήματος επιβεβαιώνουν ότι η χρήση του XGBoost, σε συνδυασμό με τη χρήση lag features, επιτυγχάνει σημαντικά υψηλότερη προγνωστική ακρίβεια συγκριτικά με τις προγενέστερες προσπάθειες. Το σύστημα που προέκυψε διαθέτει πλήρη αυτονομία λειτουργίας, παρέχοντας αξιόπιστες προβλέψεις και μετασχηματίζοντας πολύπλοκα δορυφορικά δεδομένα σε άμεσα χρησιμοποιήσιμη πληροφορία, συμβάλλοντας ουσιαστικά στην προστασία της υγείας των κατοίκων.

Λέξεις-κλειδιά: Ατμοσφαιρική ρύπανση, Sentinel-5P, XGBoost, MLOps, Θεσσαλονίκη, Δεδομένα, Πρόγνωση, Μοντελοποίηση.

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Λίστα ακρωνυμίων

Ελληνικός όρος	**Αγγλικός όρος**	**Ακρ.**
Παγκόσμιος Οργανισμός Υγείας	World Health Organization	Π.Ο.Υ.
Διοξείδιο του Αζώτου	Nitrogen Dioxide	NO_2
Όζον	Ozone	O_3
Μονοξείδιο του Άνθρακα	Carbon Monoxide	CO
Διοξείδιο του Θείου	Sulfur Dioxide	SO_2
Φορμαλδεΰδη	Methanal	$HCHO$
Αιωρούμενα σωματίδια	Aerosols	-
Ευρωπαϊκής Ένωσης	-	Ε.Ε.
Ευρωπαϊκός Οργανισμός Διαστήματος	-	Ε.Ο.Δ.
Μηχανές διανυσμάτων στήριξης	Support Vector Machines	SVM
-	Extract, Transform, Load	ETL
-	Automated Machine Learning	AutoML
Μέσο απόλυτο σφάλμα	Mean Absolute Error	MAE
Μέσο τετραγωνικό σφάλμα	Mean Squared Error	MSE
Ρίζα Μέσου Τετραγωνικού Σφάλματος	Root Mean Squared Error	RMSE
-	Global Monitoring for Environment and Security	GMES
-	Gradient Boosted Decision Trees	GBDT
-	Virtual Machine	VM
-	Supervised Learning	SL

Κεφάλαιο 1

Εισαγωγή

1.1 Εισαγωγή και περιγραφή του προβλήματος

1.1.1 Παγκόσμιο πλαίσιο και οι επιπτώσεις στην υγεία

Σε παγκόσμια κλίμακα, η ρύπανση της ατμόσφαιρας συγκαταλέγεται μεταξύ των κυριότερων περιβαλλοντικών απειλών που πλήττουν τη δημόσια υγεία. Αναγνωρίζοντας τη σοβαρότητα του ζητήματος, ο Π.Ο.Υ. το 2021 σε αναθεώρηση των κατευθυντήριων οδηγιών αναφορικά με την ατμοσφαιρική ποιότητα, επιβάλλοντας πιο αυστηρά όρια για τους κρίσιμους ρύπους. Η επιστημονική βιβλιογραφία τεκμηριώνει ότι η έκθεση σε ρυπαντές, συμπεριλαμβανομένων του NO_2, του O_3, του CO και του SO_2, συνεπάγεται σε άμεσους κινδύνους για την ανθρώπινη υγεία. Οι επιπτώσεις ευθύνονται από παθήσεις του αναπνευστικού και του καρδιαγγειακού συστήματος έως διαταραχές της νευρολογικής λειτουργίας και ακόμη την εμφάνιση καρκίνου.

Αναλυτικότερα:

NO_2: Η κυρίαρχη πηγή προέλευσής του εντοπίζεται στην κυκλοφοριακή συμφόρηση, ενώ επιδημιολογικά δεδομένα τον συσχετίζονται με αυξημένα ποσοστά νοσηλειών λόγω άσθματος και χρόνιας αποφρακτικής πνευμονοπάθειας.

O_3: Πρόκειται για δευτερογενή ρύπο που προκύπτει μέσω φωτοχημικών αντιδράσεων, επιφέροντας δυσμενείς συνέπειες στην πνευμονική λειτουργία.

SO_2 & CO: Η προέλευσή τους συνδέεται με διεργασίες καύσης (θέρμανση κτιρίων, βιομηχανική δραστηριότητα εργοστασίων), προκαλώντας οξέα αναπνευστικά συμπτώματα και αυξάνοντας τον καρδιοπνευμονικό κίνδυνο.

HCHO (Φορμαλδεΰδη): Παρότι συχνά εξετάζεται ως ρύπος κλειστών χώρων, η παρουσία του σε αστικά περιβάλλοντα υπό τη μορφή πτητικής οργανικής ένωσης και η δορυφορική ανίχνευσή του αποκτούν κρίσιμη σημασία της καρκινογόνου ουσίας.

Επιπλέον, πρόσφατα ερευνητικά ευρήματα έχουν καταδείξει θετική συσχέτιση μεταξύ της παρατεταμένης έκθεσης σε αυτούς τους ρύπους και της αυξημένης θνησιμότητας που παρατηρήθηκε κατά την πανδημία COVID-19. Το γεγονός αυτό υπογραμμίζει την αυξημένη ευπάθεια του ανοσοποιητικού μηχανισμού σε γεωγραφικές περιοχές με υψηλά επίπεδα ατμοσφαιρικής ρύπανσης.

1.1.2 Η κατάσταση στην Θεσσαλονίκη

Η Θεσσαλονίκη αποτελεί ιδιαίτερα ενδιαφέρουσα περίπτωση μελέτης, καθώς συνδυάζει υψηλά επίπεδα κυκλοφοριακής συμφόρησης με συγκεκριμένα τοπογραφικά χαρακτηριστικά που επιβαρύνουν το πρόβλημα της ρύπανσης. Αν και οι εθνικοί μέσοι όροι ενδέχεται να κυμαίνονται εντός των επιτρεπτών ορίων, τα ερευνητικά δεδομένα αποκαλύπτουν ότι οι σταθμοί παρακολούθησης ρύπανσης στο αστικό κέντρο έχουν καταγράψει ιστορικά υπερβάσεις των ετήσιων ορίων για το NO_2, με τιμές που αγγίζουν τα $45\ \mu g/m^3$, όταν το όριο ορίζεται στα $40\ \mu g/m^3$.

Η άμεση εξάρτηση των επιπέδων ρύπανσης από το τεράστιο όγκο οχημάτων επιβεβαιώθηκε κατά τη διάρκεια των περιοριστικών μέτρων lockdown που επιβλήθηκαν λόγω της πανδημίας COVID-19. Συγκεκριμένα, στον ελληνικό χώρο παρατηρήθηκαν μειώσεις στις συγκεντρώσεις NO_2 που κυμάνθηκαν από 25% έως 65%. Επιπρόσθετα, η πυκνή αστική δομή και τα γεωμορφολογικά χαρακτηριστικά της Θεσσαλονίκης διαμορφώνουν συνθήκες «αστικών φαραγγιών», τα οποία λειτουργούν ως παγίδες για τους ρύπους. Αυτό το φαινόμενο οδηγεί στον εγκλωβισμό ρύπων μεταξύ των οποίων τα Aerosols και η HCHO στα κατώτερα ατμοσφαιρικά στρώματα, με αποτέλεσμα την περαιτέρω υποβάθμιση της ποιότητας του αέρα που εισπνέουν οι κάτοικοι.

Συνεπώς, καθίσταται επιτακτική η ανάπτυξη ενός συστήματος που δεν θα περιορίζεται στην απλή καταγραφή των δεδομένων, αλλά θα παρέχει προγνωστικές δυνατότητες σε πραγματικό χρόνο, συμβάλλοντας ουσιαστικά στην προστασία της δημόσιας υγείας στην ευρύτερη περιοχή.

1.2 Αντικείμενο και στόχοι της πλατφόρμας

1.2.1 Αυτοματοποιημένο ETL και λήψη αποφάσεων

Η δημιουργία ολοκληρωμένων αυτοματοποιημένων αγωγών ETL για σκοπούς περιβαλλοντικής παρακολούθησης προσφέρει σημαντική επιστημονική και λειτουργική προστιθέμενη αξία, καθώς γεφυρώνει το κενό ανάμεσα στη συλλογή πρωτογενών δεδομένων και την πρακτική υποστήριξη διαδικασιών λήψης αποφάσεων. Τα αυτοματοποιημένα συστήματα εξασφαλίζουν την αξιοπιστία των συχνών ροών δεδομένων μέσω της άμεσης διαδικασίας καθαρισμού και της συμπλήρωσης ελλειπόντων τιμών, αντιμετωπίζοντας αποδοτικά τα κενά που συχνά χαρακτηρίζουν τα δίκτυα αισθητήρων [1].

Οι εν λόγω αγωγοί διευκολύνουν την ομαλή ενοποίηση διαφορετικών συνόλων ανοικτών δεδομένων όπως εκείνα που διατίθενται από τα προγράμματα Copernicus Sentinel και Open-Meteo επιτρέποντας την ευρείας κλίμακας εφαρμογή προηγμένων μεθόδων μηχανικής μάθησης και τεχνι-

κών σύντηξης δεδομένων [2]. Τέτοιες αρχιτεκτονικές προάγουν την επεκτασιμότητα, καθιστώντας εφικτή την παράλληλη επεξεργασία πολλαπλών ροών δεδομένων με υψηλή χωρική και χρονική ανάλυση.

Επιπρόσθετα, η ενσωμάτωση αυτών των αγωγών σε διαδικτυακά συστήματα υποστήριξης λήψης αποφάσεων παρέχει στους ενδιαφερόμενους φορείς πρόσβαση σε εξειδικευμένες γνώσεις για κρίσιμες λειτουργίες, συμπεριλαμβανομένης της αξιολόγησης περιβαλλοντικών κινδύνων και της πρόγνωσης συγκεντρώσεων ατμοσφαιρικών ρύπων. Εν τέλει, η αυτοματοποίηση αυτή περιορίζει την ανθρώπινη παρέμβαση, μειώνει τον υπολογιστικό φόρτο και καθιστά εφικτή την προληπτική περιβαλλοντική διαχείριση ως αντίδραση στις μεταβαλλόμενες οικολογικές συνθήκες.

1.2.2 Επεξεργασία σε πραγματικό χρόνο

Η υιοθέτηση ολοκληρωμένων αγωγών ETL από άκρο σε άκρο (*end-to-end*) για την επεξεργασία μετεωρολογικών δεδομένων, σε αντίθεση με τα αυτόνομα μοντέλα, επιφέρει σημαντικές βελτιώσεις τόσο στην ακεραιότητα των δεδομένων όσο και στην επιχειρησιακή απόδοση. Σύμφωνα με τους Zhang και Thorburn [1], τα ολοκληρωμένα συστήματα που βασίζονται σε υποδομές υπολογιστικού νέφους ενισχύουν την αξιοπιστία των δεδομένων μέσω της εκτέλεσης καθαρισμού σε πραγματικό χρόνο και της αυτοματοποιημένης συμπλήρωσης ελλειπουσών τιμών, μειώνοντας έτσι τη στατιστική μεροληψία που χαρακτηρίζει τα αποσπασματικά σύνολα δεδομένων.

Σε σύγκριση με τις μεμονωμένες μεθοδολογίες, οι αγωγοί αυτοί εξασφαλίζουν ανώτερη επεκτασιμότητα, αξιοποιώντας αρχιτεκτονικές που επιτρέπουν την ταυτόχρονη διαχείριση εκατοντάδων διαφορετικών ροών δεδομένων. Περαιτέρω, όπως τεκμηριώνουν οι Ben Bouallègue κ.ά. [3] και Himeur κ.ά. [2], οι ροές εργασίας που βασίζονται στο deep learning από άκρο σε άκρο καθιστούν εφικτή την εξαγωγή συμπερασμάτων σε πραγματικό χρόνο με ταχύτητες που υπερβαίνουν κατά πολλαπλάσιες τάξεις μεγέθους εκείνες της συμβατικής αριθμητικής ολοκλήρωσης.

Η υπολογιστική αυτή ικανότητα αποδεικνύεται κρίσιμη για την ενίσχυση των διαδικασιών λήψης αποφάσεων, καθώς επιτρέπει στις αρμόδιες αρχές να προβλέπουν περιβαλλοντικές απειλές, όπως επεισόδια έντονης ρύπανσης ή ακραία μετεωρολογικά φαινόμενα, σύμφωνα με τους Méndez κ.ά. [4]. Μέσω της ενοποίησης της συλλογής δεδομένων με τη προγνωστική μοντελοποίηση, οι αυτοματοποιημένοι αγωγοί εγγυώνται τη δομική συνοχή και θεμελιώνουν μια ισχυρή υποδομή για την προληπτική περιβαλλοντική διαχείριση και τη χάραξη πολιτικών βασισμένων σε τεκμηριωμένα δεδομένα.

1.2.3 Διαδικτυακή οπτικοποίηση και επικοινωνία δεδομένων

Οι διαδικτυακοί πίνακες ελέγχου και τα εργαλεία οπτικοποίησης έχουν καταστεί κρίσιμοι μηχανισμοί για τη μετατροπή περίπλοκων περιβαλλοντικών συνόλων δεδομένων σε πρακτικά χρησιμοποιήσιμη γνώση προς το ευρύ κοινό. Οι εφαρμογές αυτές λειτουργούν ως «επιμελημένος φακός», καθιστώντας εφικτή την παρακολούθηση των δυναμικών οικολογικών μεταβολών από μη ειδικευμένα κοινά με μια απλή ματιά, μέσω δομημένων σχεδιαστικών προτύπων και οπτικών αφηρημένων

αναπαραστάσεων [5].

Μέσω της ενσωμάτωσης τεχνολογιών IoT και ανάλυσης μαζικών δεδομένων, αυτές οι πλατφόρμες ενισχύουν ουσιαστικά την προσβασιμότητα των δεδομένων, παρέχοντας ροές πληροφοριών υψηλής ταχύτητας σχετικά με την ατμοσφαιρική ποιότητα, τη διαχείριση υδάτινων πόρων και την έξυπνη γεωργία [6]. Επιπροσθέτως, η λειτουργία σε πραγματικό χρόνο που χαρακτηρίζει αυτά τα εργαλεία προωθεί τη συλλογική λήψη αποφάσεων και την ενεργό συμμετοχή των πολιτών, ενισχύοντας ένα κοινό όραμα περιβαλλοντικής βιωσιμότητας και λογοδοσίας [7].

Μέσω διαδραστικών δυνατοτήτων, όπως η εξερεύνηση δεδομένων και η παροχή λεπτομερειών κατόπιν αιτήματος, οι πίνακες ελέγχου καθιστούν εφικτή τη βαθύτερη αντίληψη των περιβαλλοντικών τάσεων. Ως εκ τούτου, τα εργαλεία αυτά ενδυναμώνουν τους εμπλεκόμενους φορείς να αντιδρούν προληπτικά στις περιβαλλοντικές προκλήσεις, ενώ παράλληλα μειώνουν το γνωστικό φορτίο που συνδέεται με την ερμηνεία δεδομένων μεγάλης κλίμακας.

1.3 Δομή της εργασίας

Η παρούσα διπλωματική εργασία οργανώνεται σε έξι κεφάλαια, τα οποία οδηγούν τον αναγνώστη μέσω μιας συνεπούς πορείας από τη θεωρητική βάση προς την τεχνική πραγματοποίηση και, εν τέλει, προς την κριτική αξιολόγηση της προτεινόμενης προσέγγισης. Αρχικά, το δεύτερο κεφάλαιο καθορίζει το θεωρητικό πλαίσιο της μελέτης, εξετάζοντας τις παραμέτρους της ατμοσφαιρικής ρύπανσης και τις δυνατότητες δορυφορικής παρακολούθησης της Γης μέσω του προγράμματος Copernicus και του αισθητήρα TROPOMI του δορυφόρου Sentinel-5P. Παράλληλα, διερευνώνται οι τεχνολογίες μηχανικής μάθησης, εστιάζοντας στον αλγόριθμο XGBoost και στο πλαίσιο βελτιστοποίησης υπερπαραμέτρων Optuna, όπως επίσης και στα εργαλεία ανάπτυξης λογισμικού που υιοθετήθηκαν. Προχωρώντας στον σχεδιασμό της λύσης, το τρίτο κεφάλαιο διερευνά τις λειτουργικές και μη λειτουργικές προδιαγραφές του συστήματος, παρουσιάζοντας την αρχιτεκτονική μικροϋπηρεσιών, το σχήμα της βάσης δεδομένων και τις διεπαφές προγραμματισμού εφαρμογών (APIs) που αξιοποιήθηκαν για την απόκτηση των πρωτογενών δεδομένων. Κατόπιν, το τέταρτο κεφάλαιο επικεντρώνεται στην πραγματοποίηση του συστήματος, περιγράφοντας λεπτομερώς τον αγωγό ETL για τη συλλογή και κατεργασία των γεωχωρικών δεδομένων, τη διαδικασία εκπαίδευσης και ρύθμισης των μοντέλων, καθώς και την κατασκευή της αναπαραγώγιμης υποδομής μέσω της πλατφόρμας Docker. Στη συνέχεια, το πέμπτο κεφάλαιο παρουσιάζει την τελική εφαρμογή και τη διεπαφή χρήστη που δημιουργήθηκε με το εργαλείο Streamlit. Στο ίδιο κεφάλαιο διενεργείται η αξιολόγηση της επίδοσης των μοντέλων μέσω στατιστικών δεικτών, εξετάζεται η σπουδαιότητα των χαρακτηριστικών εισόδου και αναλύεται μια μελέτη περίπτωσης που αφορά τη γεωγραφική περιοχή της Θεσσαλονίκης. Η εργασία ολοκληρώνεται με το έκτο κεφάλαιο, όπου ανακεφαλαιώνονται τα τελικά συμπεράσματα, τονίζονται οι περιορισμοί που αναδύθηκαν και υποδεικνύονται κατευθύνσεις για μελλοντικές επεκτάσεις του προγνωστικού μηχανισμού. Μέσω αυτής της διάρθρωσης, επιδιώκεται η ανάπτυξη ενός ολοκληρωμένου συστήματος που συνδέει τη δορυφορική τηλεπισκόπηση με την προηγμένη υπολογιστική ανάλυση για την παροχή ακριβών προβλέψεων ρύπανσης, προσφέροντας τη βάση για τη διαμόρφωση τεκμηριωμένων αποφάσεων. Η μελέτη εκκινεί στο επόμενο κεφάλαιο με τη διεξοδική βιβλιογραφική επισκόπηση των επιστημονικών εννοιών

και των τεχνολογιών αιχμής που τεκμηριώνουν τη θεωρητική θεμελίωση της έρευνας.

Κεφάλαιο 2

Θεωρητικό υπόβαθρο & τεχνολογίες

2.1 Εισαγωγή

Η αραιή χωρική κάλυψη και η άνιση γεωγραφική κατανομή αποτελούν θεμελιώδεις περιορισμούς των συμβατικών επίγειων δικτύων μέτρησης, με αποτέλεσμα τη δημιουργία κατακερματισμένων συνόλων δεδομένων κατά την εκτίμηση της ποιότητας του αέρα σε περιφερειακή κλίμακα [4]. Συνεχείς παρατηρήσεις υψηλής χωρικής ανάλυσης για κρίσιμους ατμοσφαιρικούς ρύπους παρέχονται σε παγκόσμια κλίμακα μέσω της δορυφορικής αποστολής του Sentinel-5P, η οποία εφαρμόζει το όργανο παρακολούθησης της τροπόσφαιρας TROPOMI, αντιμετωπίζοντας έτσι τα κρίσιμα αυτά κενά [8]. Ετσι, η υψηλή διαστασιμότητα και η πολυπλοκότητα που χαρακτηρίζουν τα δορυφορικά δεδομένα απαιτούν την υιοθέτηση προηγμένων αναλυτικών μεθοδολογιών. Η ενσωμάτωση αλγορίθμων Μηχανικής Μάθησης καθίσταται απαραίτητη, δεδομένου ότι αυτοί επιλύουν αποτελεσματικά τις περίπλοκες μη γραμμικές χωροχρονικές συσχετίσεις δηλαδή τις σχέσεις που τα συμβατικά μοντέλα χημείας-μεταφοράς CTMs και οι παραδοσιακές στατιστικές προσεγγίσεις συχνά αποτυγχάνουν να αναπαραστήσουν επαρκώς [9].

Μέσω της αυτοματοποιημένης εξαγωγής γνώσης από ογκώδεις ροές περιβαλλοντικών δεδομένων, τα συστήματα Μηχανικής Μάθησης βελτιώνουν ουσιαστικά τόσο την ακρίβεια όσο και την ταχύτητα των προγνωστικών μοντέλων. Συνεπώς, η σύζευξη της δορυφορικής τηλεπισκόπησης με την υπολογιστική νοημοσύνη συνιστά μετασχηματιστική αλλαγή παραδείγματος στον τομέα των περιβαλλοντικών επιστημών, μετατρέποντας την παρακολούθηση της ποιότητας του αέρα σε ολιστική, δεδομενοκεντρική επιστήμη με δυνατότητα υποστήριξης πρωτοβουλιών δημόσιας υγείας και διαχείρισης αστικού χώρου σε πραγματικό χρόνο.

2.2 Ατμοσφαιρική ρύπανση

2.2.1 Βασικοί ρύποι και οι επιπτώσεις τους στον άνθρωπο

Ένα πολύπλοκο μείγμα πρωτογενών και δευτερογενών ρύπων διαμορφώνει την ποιότητα του ατμοσφαιρικού αέρα στα αστικά κέντρα, επιφέροντας σοβαρές συνέπειες στη δημόσια υγεία. Ανθρωπογενείς δραστηριότητες αποτελούν την κύρια πηγή προέλευσης των σημαντικότερων πρωτογενών ρύπων, συμπεριλαμβανομένων του NO_2, του CO και του SO_2 [10]. Η καύση ορυκτών καυσίμων στις οδικές μεταφορές και στην οικιακή θέρμανση συνιστά, σύμφωνα με τη διεθνή βιβλιογραφία, τις κυριότερες πηγές εκπομπής. Γι'αυτόν τον λόγω, το τροποσφαιρικό O_3 παρουσιάζει διαφορετική συμπεριφορά, καθώς δεν έλκεται άμεσα στην ατμόσφαιρα αλλά παράγεται μέσω φωτοχημικών μετασχηματισμών πρόδρομων ρυπαντών υπό την επίδραση της ηλιακής ακτινοβολίας [11]. Το αναπνευστικό και καρδιαγγειακό σύστημα αποτελούν τους πρωταρχικούς στόχους των αρνητικών επιπτώσεων αυτών των ρύπων. Συγκεκριμένα, η έκθεση σε σωματιδιακό υλικό σχετίζεται άμεσα με την ανάπτυξη άσθματος, τη χρόνια αποφρακτική πνευμονοπάθεια, καθώς και με αυξημένη πιθανότητα εμφάνισης ισχαιμικών επεισοδίων και εγκεφαλικών περιστατικών που εμποδίζει τη φυσιολογική μεταφορά οξυγόνου, οδηγώντας σε καρδιολογικές και νευρολογικές επιπλοκές.

Ακόμα,αξίζει να αναφερθεί ότι η $HCHO$, η οποία πέρα από τις έντονες ερεθιστικές ιδιότητές της στους βλεννογόνους, έχει κατηγοριοποιηθεί ως καρκινογόνος παράγοντας και συνδέεται με νεοπλάσματα της ρινοφαρυγγικής περιοχής [12]. Συνολικά, η αλληλεπίδραση των προαναφερθέντων ρύπων επιδεινώνει την ποιότητα ζωής και συμβάλλει στην αύξηση των δεικτών πρόωρης θνησιμότητας μεταξύ του αστικού πληθυσμού.

2.2.2 Σημαντικά χαρακτηριστικά της ρύπανσης στη Θεσσαλονίκη

Η ιδιαίτερη τοπογραφική διαμόρφωση και οι μετεωρολογικές συνθήκες της περιοχής ασκούν καθοριστική επίδραση στην ποιότητα του αέρα εντός του πολεοδομικού συγκροτήματος της Θεσσαλονίκης. Η συσσώρευση τοξικών ουσιών στην ατμόσφαιρα ευνοείται από τον περιορισμένο φυσικό αερισμό, τα αυξημένα ποσοστά σχετικής υγρασίας και τις συχνές αναστροφές της θερμοκρασίας, οι οποίες δυσχεραίνουν σημαντικά τη διασπορά των ρύπων.

Έντονη εποχική μεταβλητότητα είναι αυτή που χαρακτηρίζει τόσο τις πηγές όσο και τις συγκεντρώσεις των ρύπων. Την περίοδο του χειμώνα, η ευρεία χρήση καύσης βιομάζας ξύλου για οικιακή θέρμανση καθιστά τα Aerosols και το CO τους κύριους ρύπους, προκαλώντας συχνά φαινόμενα αιθαλομίχλης. Οι εκπομπές από την οδική κυκλοφορία, αντιθέτως, συνιστούν σταθερή πηγή διαχρονικά, με ιδιαίτερη επιβάρυνση στο κεντρικό αστικό ιστό.

Κατά την καλοκαιρινή περίοδο, η έντονη ηλιακή ακτινοβολία ενεργοποιεί σημαντική φωτοχημική δραστηριότητα, με αποτέλεσμα την παραγωγή υψηλών συγκεντρώσεων δευτερογενών ρύπων, συμπεριλαμβανομένων του όζοντος και της φορμαλδεΰδης. Η συστηματική παρακολούθηση υψηλής χρονικής ανάλυσης καθίσταται επιτακτική λόγω της πολυπλοκότητας των χημικών μετασχηματισμών και της ταχείας διακύμανσης των επιπέδων ρύπανσης σε σχέση με τις τοπικές πηγές εκπο-

μπών. Η λεπτομερής καταγραφή των ρυπαντών αποτελεί απαραίτητη προϋπόθεση για την αναγνώριση των πηγών έκθεσης και την ανάπτυξη αποτελεσματικών πολιτικών προστασίας της δημόσιας υγείας στην πόλη.

2.3 Δορυφορική παρατήρηση

2.3.1 Copernicus και τα ανοικτά δεδομένα

Η εμβληματική πρωτοβουλία της Ε.Ε. στον τομέα της άντλησης δεδομένων από δορυφόρους, αποτελεί το πρόγραμμα Copernicus, το οποίο προηγουμένως ήταν γνωστό ως GMES. Χαρακτηρίζεται συχνά ως «τα μάτια της Ευρώπης στη Γη», παρέχοντας ολοκληρωμένη απεικόνιση της πλανητικής κατάστασης μέσω της συλλογής συνεχών και υψηλής ακρίβειας δεδομένων.

Τρεις πυλώνες αποτελούν τη βάση της αρχιτεκτονικής του προγράμματος: το διαστημικό σκέλος, το σκέλος επίγειων μετρήσεων και οι υπηρεσίες [13]. Ο Ε.Ο.Δ. διαδραματίζει καθοριστικό ρόλο, συντονίζοντας το διαστημικό σκέλος και αναλαμβάνοντας τον σχεδιασμό, την ανάπτυξη και τη λειτουργική διαχείριση των εξειδικευμένων δορυφορικών αποστολών Sentinel.

Η πολιτική «πλήρους, ελεύθερης και ανοικτής» πρόσβασης στα δεδομένα (*full, free, and open data policy*) είναι πλέον από τα σημαντικότερα που χαρακτηριζει το Copernicus. Μέσω αυτής της στρατηγικής, έχει επιτευχθεί ο εκδημοκρατισμός της πρόσβασης σε πληροφορίες υψηλής ανάλυσης, με το κόστος για ερευνητικούς φορείς και δημόσιες υπηρεσίες να μηδενίζεται πλήρως [14]. Η εξάλειψη των οικονομικών φραγμών καθιστά το πρόγραμμα καταλυτικό παράγοντα για την επιστημονική καινοτομία στους τομείς της περιβαλλοντικής παρακολούθησης και της κλιματικής αλλαγής. Η ελεύθερη διαθεσιμότητα των δεδομένων Sentinel επιτρέπει τη δημιουργία καινοτόμων υπηρεσιών και εφαρμογών προστιθέμενης αξίας, καθιστώντας το Copernicus θεμελιώδες εργαλείο για τη χάραξη πολιτικών σχετικά με τη βιώσιμη ανάπτυξη και την προστασία των οικοσυστημάτων σε παγκόσμιο επίπεδο.

2.4 Μηχανική Μάθηση

2.4.1 Μετασχηματισμός χρονοσειρών σε προβλήματα επιβλεπόμενης μάθησης

Η ανάπτυξη προγνωστικών μοντέλων βασίζεται θεμελιωδώς στην εποπτευόμενη μάθηση γνωστή και ως SL, όπου επιδιώκεται η εκμάθηση μιας συνάρτησης χαρτογράφησης που συνδέει ένα σύνολο εισόδων (X) με μια μεταβλητή εξόδου (Y), αξιοποιώντας ιστορικά δεδομένα που περιέχουν τις πραγματικές τιμές-στόχους [15]. Στο πλαίσιο της παρούσας εργασίας, η πρόβλεψη των συγκεντρώσεων αέριων ρύπων προσεγγίζεται ως πρόβλημα παλινδρόμησης. Η φύση των δεδομένων υπαγορεύει αυτήν την επιλογή, δεδομένου ότι οι συγκεντρώσεις ρύπων όπως NO_2 ή O_3 εκφράζονται

σε συνεχείς αριθμητικές τιμές εντός συγκεκριμένου εύρους, σε αντιδιαστολή με τα προβλήματα ταξινόμησης που αφορούν διακριτές κατηγορίες [16].

Σημαντική θεωρητική πρόκληση παρουσιάζεται κατά την εφαρμογή κλασικών αλγορίθμων μηχανικής μάθησης, όπως το XGBoost, το Random Forest ή οι SVM, σε χρονοσειρές [17]. Οι αλγόριθμοι αυτοί έχουν σχεδιαστεί για την επεξεργασία πινακοειδών δεδομένων, όπου κάθε γραμμή αντιμετωπίζεται ως ανεξάρτητη από τις υπόλοιπες. Ετσι, οι χρονοσειρές διακρίνονται από εγγενή χρονική εξάρτηση, με τη σειρά των δεδομένων να διαδραματίζει καθοριστικό ρόλο στην εξαγωγή συμπερασμάτων. Η επίτευξη συμβατότητας απαιτεί τον μετασχηματισμό της μονοδιάστατης χρονοσειράς σε δομημένο σχήμα εποπτευόμενης μάθησης μέσω της τεχνικής του «συρόμενου παραθύρου», η οποία είναι επίσης γνωστή ως «ενσωμάτωση χρονικής υστέρησης» [18]. Η δημιουργία χαρακτηριστικών υστέρησης αποτελεί τη βάση αυτού του μετασχηματισμού. Ειδικότερα, η τιμή της χρονοσειράς σε χρονική στιγμή t που συνιστά την εξαρτημένη μεταβλητή (y_t) μοντελοποιείται ως συνάρτηση των προγενέστερων παρατηρήσεων. Οι τιμές των προηγούμενων χρονικών βημάτων μετατρέπονται σε ανεξάρτητες μεταβλητές εισόδου. Με αυτόν τον τρόπο, το μοντέλο καθίσταται ικανό να «εκπαιδευτεί» αναγνωρίζοντας μοτίβα και τάσεις από ιστορικά δεδομένα προκειμένου να προβλέψει τη μελλοντική συμπεριφορά του φαινομένου. Η προσέγγιση αυτή καθιστά δυνατή τη χρήση ισχυρών αλγορίθμων που, παρότι δεν σχεδιάστηκαν εξαρχής για ακολουθίες, είναι ικανοί να αποκαλύψουν σύνθετες μη γραμμικές σχέσεις στα δεδομένα ρύπανσης [19].

2.4.2 Αρχιτεκτονική και μαθηματική προσέγγιση του XGBoost

Η τεχνική του XGBoost αποτελεί μια εξελιγμένη προσέγγιση στην υλοποίηση των δέντρων απόφασης με ενίσχυση κλίσης GBDT, επιτυγχάνοντας εξαιρετική βελτιστοποίηση. Διαδοχικά δέντρα ενσωματώνονται στο μοντέλο μέσω της προσθετικής εκπαίδευσης additive training, με κάθε νέο δέντρο να στοχεύει στη μείωση των αποκλίσεων που παρουσιάζουν τα προγενέστερα [20].

Η βελτιστοποίηση μιας πολύπλοκης αντικειμενικής συνάρτησης objective function συνιστά τη μαθηματική βάση της υπεροχής του αλγορίθμου. Στο χρονικό βήμα t, για δεδομένα που περιλαμβάνουν n παραδείγματα και m χαρακτηριστικά, η αντικειμενική συνάρτηση διατυπώνεται ως εξής [21]:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{2.1}$$

Η συνάρτηση απώλειας l ποσοτικοποιεί την απόκλιση ανάμεσα στην πραγματική τιμή y_i και την προβλεπόμενη $\hat{y}_i$, ενώ ο όρος $\Omega(f_t)$ εκφράζει την κανονικοποίηση , η οποία περιορίζει την πολυπλοκότητα του μοντέλου προς αποτροπή της υπερπροσαρμογής. Αναλυτικότερα, η κανονικοποίηση διαμορφώνεται ως:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2.2}$$

Ο αριθμός των φύλλων του δέντρου συμβολίζεται από το T, ενώ το w αντιστοιχεί στο βάρος του κάθε φύλλου. Οι παράμετροι γ και λ αποτελούν συντελεστές ποινής: η πρώτη ρυθμίζει τη διαδικασία δημιουργίας φύλλων, ενώ η δεύτερη εξομαλύνει τα βάρη που αντιστοιχούν στις προβλέψεις [21], [22].

Για την ταχεία και ακριβή βελτιστοποίηση της αντικειμενικής συνάρτησης, εφαρμόζεται ανάπτυγμα Taylor δεύτερης τάξης που προσεγγίζει τη συνάρτηση απώλειας. Σε αντίθεση ως προς τους συμβατικούς αλγορίθμους ενίσχυσης κλίσης που αξιοποιούν αποκλειστικά την παράγωγο πρώτης τάξης, η συγκεκριμένη μέθοδος ενσωματώνει και τη δεύτερη παράγωγο, διευκολύνοντας έτσι μια εξελιγμένη προσαρμογή [21]:

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{2.3}$$

Η πρώτη παράγωγος g_i και η δεύτερη παράγωγος h_i της συνάρτησης απώλειας υπολογίζονται ως προς την πρόβλεψη του αμέσως προηγούμενου σταδίου:

$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}) \tag{2.4}$$

$$h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}) \tag{2.5}$$

Σύνθετες μη γραμμικές σχέσεις διαχειρίζονται αποτελεσματικά μέσω αυτής της προσέγγισης, γεγονός που καθιστά τον αλγόριθμο ιδιαίτερα αποδοτικό για πινακοειδή και δομημένα δεδομένα. Σύμφωνα με την επιστημονική βιβλιογραφία, τα δενδρικά μοντέλα παρουσιάζουν εξαιρετική ικανότητα επεξεργασίας ετερογενών χαρακτηριστικών που χαρακτηρίζονται από διαφορετικές κλίμακες και μη τυποποιημένα πρότυπα [23]. Συμπληρωματικά, η μέθοδος επιδεικνύει σημαντική ανθεκτικότητα απέναντι σε ακραίες τιμές, ενώ παράλληλα διαθέτει ενσωματωμένες λειτουργίες για την αυτοματοποιημένη αντιμετώπιση ελλιπών δεδομένων, επιλέγοντας αυτόνομα την κατεύθυνση διακλάδωσης που ελαχιστοποιεί την απώλεια κατά την εκπαίδευση [22]. Συνεπώς, η ταυτόχρονη εφαρμογή κανονικοποίησης $L1$ (Lasso) και $L2$ (Ridge) περιορίζει σημαντικά την πιθανότητα υπερπροσαρμογής—παράγοντας καθοριστικός όταν αναλύονται σύνολα δεδομένων περιορισμένης έκτασης, όπως συμβαίνει τυπικά σε έρευνες ατμοσφαιρικής ρύπανσης [24].

2.4.3 Θεωρητικό υπόβαθρο των υπερπαραμέτρων Optuna

Οι υπερπαράμετροι συνιστούν θεμελιώδη στοιχεία που καθορίζουν τη δομή και τη λειτουργική συμπεριφορά των μοντέλων μηχανικής μάθησης, ορίζονται προ της έναρξης της εκπαιδευτικής διαδικασίας και δεν εκτιμώνται άμεσα από τα δεδομένα [25]. Ως η πλέον κρίσιμη παράμετρος του βελτιστοποιητή, ο ρυθμός μάθησης ρυθμίζει το μέγεθος του βήματος κατά κάθε επανάληψη προς την επίτευξη σύγκλισης [26]. Υψηλές τιμές ενδέχεται να προκαλέσουν ταλαντώσεις ή αποτυχία σύγκλισης, ενώ εξαιρετικά χαμηλές τιμές συνεπάγονται με εξαιρετικά αργή εκπαίδευση με αυξημένο

υπολογιστικό κόστος [27]. Παράλληλα, το μέγιστο βάθος ελέγχει την πολυπλοκότητα των δενδρικών αρχιτεκτονικών, επηρεάζοντας άμεσα την ικανότητά τους να εντοπίζουν σύνθετα πρότυπα ωστόσο, εξαιρετικά βαθιά δέντρα αυξάνουν τον κίνδυνο υπερπροσαρμογής [28]. Η αυτοματοποίηση της αναζήτησης βέλτιστων διαμορφώσεων επιτυγχάνεται μέσω του πλαισίου Optuna, το οποίο εφαρμόζει Μπεϋζιανή βελτιστοποίηση.

Η στρατηγική αυτή χρησιμοποιεί ένα πιθανοτικό μοντέλο για την αναπαράσταση της αντικειμενικής συνάρτησης και μια συνάρτηση απόκτησης για την επιλογή των επόμενων σημείων αξιολόγησης [26]. Μέσω αυτής της επαναληπτικής προσέγγισης, το Optuna επιτυγχάνει ισορροπία μεταξύ της εξερεύνησης άγνωστων περιοχών και της εκμετάλλευσης περιοχών με υψηλή απόδοση [25]. Κατά αυτόν τον τρόπο, το σύστημα περιορίζει τον αριθμό των απαιτούμενων δοκιμών, διασφαλίζοντας ταυτόχρονα υψηλή ακρίβεια και αποδοτικότητα στον εντοπισμό της βέλτιστης αρχιτεκτονικής μοντέλου [27]. Μετά τον καθορισμό της βέλτιστης αρχιτεκτονικής μέσω του Optuna, η αξιολόγηση της ικανότητας του μοντέλου να γενικεύει σε νέα δεδομένα πραγματοποιείται μέσω συγκεκριμένων στατιστικών δεικτών. Για την κατανόηση της ακρίβειας και της αξιοπιστίας του μοντέλου XGBoost στην πρόβλεψη ατμοσφαιρικών ρύπων, καθίσταται αναγκαία η αξιολόγηση της προγνωστικής του απόδοσης. Τρεις θεμελιώδεις μετρικές παλινδρόμησης υιοθετούνται για την ποσοτική αποτίμηση της απόκλισης ανάμεσα στις πραγματικές συγκεντρώσεις ρύπων και τις τιμές που προβλέπει το μοντέλο.

MAE

Η μετρική MAE ποσοτικοποιεί τον μέσο όρο των απόλυτων αποκλίσεων που παρατηρούνται μεταξύ προβλεπόμενων και πραγματικών τιμών, ανεξάρτητα από την κατεύθυνση του σφάλματος. Μαθηματικά εκφράζεται ως:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.6}$$

Η επιστημονική βιβλιογραφία χαρακτηρίζει το MAE ως γραμμική μετρική όπου κάθε μεμονωμένο σφάλμα συνεισφέρει ισοδύναμα στον τελικό μέσο όρο [29]. Συγκριτικά με μετρικές που βασίζονται σε τετραγωνική συνάρτηση σφάλματος, το MAE παρουσιάζει μειωμένη ευαισθησία απέναντι σε ακραίες τιμές, γεγονός που αποτελεί σημαντικό πλεονέκτημα [30]. Στο πλαίσιο της ατμοσφαιρικής ρύπανσης, μια διαυγής απεικόνιση της μέσης απόκλισης του μοντέλου από τις πραγματικές συγκεντρώσεις ρυπαντών παρέχεται μέσω του MAE.

RMSE

Ως τετραγωνική μετρική αξιολόγησης, η RMSE διατυπώνεται μαθηματικά ως εξής:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.7}$$

Η υπολογιστική διαδικασία της μετρικής επιφέρει ενισχυμένη έμφαση σε σημαντικές αποκλίσεις, αφού οι διαφορές τετραγωνίζονται προτού εξαχθεί η τετραγωνική τους ρίζα [29]. Συνεπώς, το RMSE αποκτά καθοριστική σημασία στην πρόβλεψη ατμοσφαιρικής ρύπανσης. Η αποφυγή εκτεταμένων σφαλμάτων κατά την πρόγνωση κρίσιμων επεισοδίων υψηλής ρύπανσης είναι ζωτικής σημασίας για το συγκεκριμένο πεδίο, δεδομένου ότι εσφαλμένες προβλέψεις τέτοιου είδους ενδέχεται να προκαλέσουν σοβαρές συνέπειες στη δημόσια υγεία και να επηρεάσουν την υιοθέτηση προληπτικών στρατηγικών.

Συντελεστής προσδιορισμού (R^2)

Η ποιότητα προσαρμογής του μοντέλου στα διαθέσιμα δεδομένα αξιολογείται μέσω του (R^2), ο οποίος ορίζεται ως:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.8}$$

Το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής που ερμηνεύεται από τις ανεξάρτητες μεταβλητές του XGBoost αντανακλάται στη συγκεκριμένη μετρική. Όταν η τιμή R^2 προσεγγίζει τη μονάδα, αποδεικνύεται ότι το μοντέλο επιτυγχάνει αποτελεσματική ερμηνεία της μεταβλητότητας των ρύπων, ενσωματώνοντας παράγοντες όπως μετεωρολογικά χαρακτηριστικά και χρονικές μεταβλητές.

2.5 Τεχνολογίες ανάπτυξης λογισμικού

2.5.1 Εικονικοποίηση μέσω Docker

Ένα σύγχρονο παράδειγμα εικονικοποίησης που αναδιαμορφώνει τον τρόπο ανάπτυξης και εκτέλεσης λογισμικού συνιστά η τεχνολογία Docker, προσφέροντας σημαντικά πλεονεκτήματα σε σχέση με τις παραδοσιακές εικονικές μηχανές VM. Οι εικονικές μηχανές στηρίζονται σε υπερεπόπτη για την προσομοίωση υλικού και επιβάλλουν την ύπαρξη πλήρους φιλοξενούμενου λειτουργικού συστήματος για κάθε εκτελούμενη περίπτωση γεγονός που συνεπάγεται σημαντική επιβάρυνση πόρων και χρονικές καθυστερήσεις κατά την εκκίνηση. Αντιθέτως, τα containers του Docker υιοθετούν εικονικοποίηση σε επίπεδο λειτουργικού συστήματος [32].

Ο διαμοιρασμός του πυρήνα του συστήματος φιλοξενίας καθιστά τα containers εξαιρετικά ελαφριά, επιτυγχάνοντας ταχύτατους χρόνους εκκίνησης και διασφαλίζοντας βέλτιστη αξιοποίηση των

υπολογιστικών πόρων, με δυνατότητα εκτέλεσης πολλαπλών απομονωμένων υπηρεσιών στο ίδιο υλικό [33].

Πέραν της αρχιτεκτονικής υπεροχής, το Docker εξελίσσεται σε καθοριστικό εργαλείο για την αντιμετώπιση της «κρίσης αναπαραγωγιμότητας» στις επιστήμες δεδομένων και την πληροφορική [34]. Η δυνατότητά του να δημιουργεί κώδικα με τις απαραίτητες βιβλιοθήκες, εξαρτήσεις και αρχεία ρυθμίσεων σε ένα ενιαίο, αμετάβλητο είδωλο εγγυάται την πανομοιότυπη εκτέλεση του λογισμικού σε οποιαδήποτε υποδομή, ανεξάρτητα από το υποκείμενο περιβάλλον. Κατά αυτόν τον τρόπο, επιλύεται αποτελεσματικά το πρόβλημα ασυμβατότητας μεταξύ διαφορετικών συστημάτων—γνωστό στην πρακτική των προγραμματιστών ως "*it works on my machine*" διασφαλίζοντας την αξιοπιστία, τη μεταφερσιμότητα και τη διαφάνεια της ερευνητικής διαδικασίας σε όλα τα στάδια ανάλυσης [35].

2.5.2 Ταχεία ανάπτυξη εφαρμογών με το Streamlit

Το Streamlit αποτελεί ένα πρωτοποριακό πλαίσιο εργασίας ανοικτού κώδικα, ειδικά σχεδιασμένο για την ταχεία κάλυψη των αναγκών της Μηχανικής Μάθησης και της Επιστήμης Δεδομένων. Η φιλοσοφία του εστιάζει στην αποδοτικότητα, καθιστώντας δυνατή τη μετατροπή σεναρίων Python σε πλήρως διαδραστικές διαδικτυακές εφαρμογές, χωρίς να απαιτείται από τον χρήστη γνώση τεχνολογιών διεπαφής όπως HTML, CSS ή JavaScript. Αυτή η δυνατότητα καθιστά την ανάπτυξη προσβάσιμη σε αναλυτές δεδομένων, επιταχύνοντας σημαντικά τη διαδικασία ταχείας ανάπτυξης εφαρμογών (Rapid Application Development) [36]. Η απλοποίηση της δομής του προγράμματος επιτυγχάνεται μέσω της αρχιτεκτονικής του Streamlit, η οποία αντιμετωπίζει τα διαδραστικά στοιχεία ως απλές μεταβλητές. Ο κώδικας διατηρεί γραμμική και ευανάγνωστη δομή, γεγονός που μειώνει δραστικά τον χρόνο μεταξύ της σύλληψης μιας προγνωστικής ιδέας και της τελικής της υλοποίησης. Σε σύγκριση με εναλλακτικές λύσεις, όπως το Plotly Dash —το οποίο παρέχει μεγαλύτερη ευελιξία στον σχεδιασμό διεπαφής αλλά απαιτεί βαθύτερη κατανόηση της αρχιτεκτονικής ιστού— το Streamlit υπερτερεί στην ταχύτητα και την ευκολία χρήσης για μη εξειδικευμένους προγραμματιστές [37]. Για τους σκοπούς της παρούσας εργασίας, η υιοθέτησή του επιτρέπει την απόλυτη εστίαση στην ανάπτυξη των προγνωστικών μοντέλων και την ανάλυση των περιβαλλοντικών δεδομένων, παρακάμπτοντας τις χρονοβόρες πολυπλοκότητες της μηχανικής διεπαφών και διασφαλίζοντας την άμεση οπτικοποίηση των αποτελεσμάτων.

Η υιοθέση του Streamlit συνιστά το τελικό στοιχείο που ενοποιεί το φάσμα των διαφορετικών τεχνολογικών εργαλείων που παρουσιάστηκαν προηγουμένως, διαμορφώνοντας έτσι ένα ολοκληρωμένο και συνεκτικό περιβάλλον ανάπτυξης. Τα δορυφορικά δεδομένα παρατήρησης της Γης που προέρχονται από το πρόγραμμα Copernicus συγκεκριμένα τον δορυφόρο Sentinel-5P υφίστανται μετατροπή σε προβλεπτική πληροφορία μέσω της αξιοποίησης του XGBoost, ενώ πλέον καθίσταται εφικτή η άμεση διασύνδεση αυτής της επεξεργασμένης γνώσης με τους τελικούς αποδέκτες. Επιπροσθέτως, το γεγονός ότι ολόκληρη η υπολογιστική διαδικασία ενθυλακώνεται εντός ενός Docker container εξασφαλίζει την επαναληψιμότητα και τη φορητότητα της λύσης, επιτρέποντας τη μετάβαση από το θεωρητικό πλαίσιο στην πρακτική εφαρμογή με συνέπεια και σταθερότητα. Συνεπώς, το σύστημα αποκτά την ικανότητα να ανταποκριθεί αποτελεσματικά στα πολύπλοκα ζητήματα που σχετίζονται με τη ρύπανση του αέρα σε μητροπολιτικές περιοχές, με την Θεσσαλονίκη να απο-

τελεί χαρακτηριστικό παράδειγμα. Έχοντας πλέον θέσει τα εννοιολογικά και τεχνικά θεμέλια, η έρευνα μεταβαίνει στη φάση της καταγραφής λειτουργικών προδιαγραφών και του σχεδιασμού της αρχιτεκτονικής του αυτοματοποιημένου μηχανισμού πρόβλεψης.

Κεφάλαιο 3

Ανάλυση απαιτήσεων και σχεδιασμός

3.1 Εισαγωγή

Αφού στο προηγούμενο κεφάλαιο τέθηκαν τα θεωρητικά θεμέλια σχετικά με την ατμοσφαιρική ρύπανση, τις τεχνικές τηλεπισκόπησης και τις μεθόδους μηχανικής μάθησης, στο παρόν κεφάλαιο επικεντρώνεται η προσοχή στο λειτουργικό πλαίσιο και τον τεχνικό σχεδιασμό της προτεινόμενης λύσης. Επιδίωξη του κεφαλαίου αποτελεί η καταγραφή των αρχιτεκτονικών επιλογών μέσω των οποίων οι θεωρητικές αρχές μετατρέπονται σε ένα αυτοματοποιημένο, κλιμακώσιμο και πλήρως λειτουργικό σύστημα για την παρακολούθηση και πρόγνωση ρύπων.

Αρχικά, προσδιορίζονται οι απαιτήσεις λειτουργικού και μη λειτουργικού χαρακτήρα, οι οποίες καθορίζουν τις προδιαγραφές αξιοπιστίας και φορητότητας που πρέπει να πληροί το σύστημα. Ακολούθως, παρουσιάζεται η αρχιτεκτονική των μικροϋπηρεσιών που έχει υιοθετηθεί, η οποία στηρίζεται στην τεχνολογία Docker και επιτυγχάνει τον διαχωρισμό των κρίσιμων λειτουργιών σε επίπεδα Backend και Frontend. Ειδική προσοχή αποδίδεται στη δομή του σχήματος της βάσης δεδομένων SQLite, που λειτουργεί ως επιχειρησιακό αποθετήριο, αλλά και στην τεχνική ενσωμάτωσης των εξωτερικών διεπαφών προγραμματισμού APIs. Συνεπώς, μέσω της συνδυαστικής αξιοποίησης δεδομένων που προέρχονται από το πρόγραμμα Copernicus του δορυφόρου Sentinel-5P και την πλατφόρμα Open-Meteo, διαμορφώνεται μια ολοκληρωμένη ροή ETL, η οποία αποτελεί τον κεντρικό άξονα της προγνωστικής μοντελοποίησης που θα εξεταστεί στα επόμενα κεφάλαια.

3.2 Απαιτήσεις συστήματος

3.2.1 Επισκόπηση ροής εργασιών συστήματος

Η λειτουργία του συστήματος που προτείνεται στηρίζεται σε μια ολοκληρωμένη αλυσίδα επεξεργασίας (*pipeline*), η οποία συγκροτείται από τρία ξεχωριστά υποσυστήματα υλοποιημένα σε Python, με κάθε ένα να αντιμετωπίζει συγκεκριμένες τεχνικές απαιτήσεις για την αποτελεσματική παρακο-

λούθηση της ατμοσφαιρικής ρύπανσης [38].

Το κεντρικό αρχείο της εφαρμογής είναι το *ingestion_data.py*, το οποίο εγκαινιάζει τη διαδικασία αναλαμβάνοντας τη συλλογή και τον καθαρισμό των πληροφοριών. Η διαδικασία αυτή εκτελείται αυτόματα μέσω της υπηρεσίας Task Scheduler σε καθημερινή βάση, εξασφαλίζοντας την αδιάλειπτη ενημέρωση της βάσης δεδομένων. Σε αυτό το στάδιο αντιμετωπίζεται μια κρίσιμη τεχνική δυσκολία που αφορά την καθυστέρηση των πληροφοριών: επειδή οι δορυφορικές μετρήσεις του Sentinel-5P απαιτούν χρόνο επεξεργασίας, το σύστημα ανακτά δεδομένα με χρονική υστέρηση μίας ημέρας ($t-1$). Για την εξασφάλιση της σταθερότητας, ενσωματώνονται ισχυροί μηχανισμοί χειρισμού σφαλμάτων που συμπεριλαμβάνουν συστηματική καταγραφή σφαλμάτων και όλης της διαδικασίας και ελέγχους εγκυρότητας των απαντήσεων από τα API, εμποδίζοντας την καταχώρηση ελλιπών ή αλλοιωμένων δεδομένων στη βάση.

Στην καρδιά της εφαρμογής, το αρχείο *training.py* υλοποιεί την εκπαίδευση των μοντέλων, αυτοματοποιώντας τον κύκλο ζωής της μηχανικής μάθησης. Το πλαίσιο MLflow ενσωματώνεται στην αρχιτεκτονική για την παρακολούθηση των πειραμάτων, ενώ μια βάση δεδομένων SQLite χρησιμοποιείται για την αποθήκευση των μετρικών απόδοσης και των υπερπαραμέτρων. Υψηλή ανοχή σφαλμάτων επιδεικνύεται από το σύστημα, καθώς οι δομές *try/except* που χρησιμοποιούνται επιτρέπουν την πρόοδο της εκπαίδευσης για τους υπόλοιπους ρύπους ακόμη κι αν συμβεί αποτυχία σε συγκεκριμένο ρύπο. Υψηλοί υπολογιστικοί πόροι απαιτούνται από τη διαδικασία, οι οποίοι καλύπτονται μέσω της παραλληλοποίησης των υπολογισμών (*n_jobs=-1*), επιταχύνοντας σημαντικά τη βελτιστοποίηση του αλγορίθμου XGBoost.

Τελικά, την παραγωγή των προγνώσεων διαχειρίζεται το τρίτο μέρος της εφαρμογής στο αρχείο *predict.py*, το οποίο υιοθετεί μια αναδρομική προσέγγιση [4]. Με αυτή τη μέθοδο, οι προβλεπόμενες τιμές κάθε χρονικού βήματος αξιοποιούνται ως είσοδοι για το επόμενο στάδιο. Παρόλο που παρέχει ευελιξία, η προσέγγιση αυτή εμπεριέχει τον κίνδυνο της συσσώρευσης σφάλματος, όπου οι αρχικές μικρές αποκλίσεις μπορούν να ενισχυθούν κατά τη διάρκεια του 24ωρου προγνωστικού ορίζοντα. Παράλληλα, η ποιότητα των εξωτερικών μετεωρολογικών προγνώσεων από την υπηρεσία Open-Meteo επηρεάζει άμεσα την ακρίβεια των αποτελεσμάτων. Επιπρόσθετα, το φαινόμενο της «ψυχρής εκκίνησης» αποτελεί πρόκληση σε περιπτώσεις διακοπής λειτουργίας, δεδομένου ότι για την αρχικοποίηση των χαρακτηριστικών υστέρησης απαιτούνται πρόσφατα ιστορικά δεδομένα. Μια κλιμακώσιμη και στιβαρή λύση για την αυτοματοποιημένη πρόβλεψη της ποιότητας αέρα στο αστικό περιβάλλον προσφέρεται μέσω της συνεργασίας των τριών αυτών υποσυστημάτων.

3.2.2 Μη λειτουργικές απαιτήσεις

Φορητότητα και αναπαραγωγιμότητα

Η πλήρης φορητότητα του προτεινόμενου συστήματος και η πιστή αναπαραγωγή του περιβάλλοντος εκτέλεσης σε διαφορετικές υπολογιστικές υποδομές κρίνονται απαραίτητες. Γι' αυτόν το λόγο, η τεχνολογία containerization υιοθετείται μέσω της πλατφόρμας Docker.

Η χρήση της βασικής εικόνας *python:3.9-slim* που βασίζεται στη διανομή Debian αντί της ελαφρύ-

τερης Alpine Linux αποτέλεσε στρατηγική επιλογή. Η απόφαση αυτή κρίθηκε αναγκαία για την εξασφάλιση της συμβατότητας των βιβλιοθηκών επιστήμης δεδομένων, όπως η NumPy, η Pandas και η XGBoost, οι οποίες εξαρτώνται από συγκεκριμένες εξαρτήσεις συστήματος που συχνά αποτυγχάνουν στο περιβάλλον Alpine. Επιπροσθέτως, η αρχή Don't Repeat Yourself εφαρμόζεται για τη βελτιστοποίηση της δομής του λογισμικού στο *Dockerfile* του Frontend, το οποίο εισάγει απευθείας από το Backend την απαραίτητη βοηθητική λογική, εμποδίζοντας τον επαναλαμβανόμενο κώδικα και διευκολύνοντας τη συντήρηση.

Αξιοπιστία, ανθεκτικότητα και ακεραιότητα

Η υποστήριξη υψηλής ανοχής σφαλμάτων και η ανθεκτικότητα πρέπει να χαρακτηρίζουν την αρχιτεκτονική του συστήματος, ειδικά στις κρίσιμες διεργασίες πρόσβασης δεδομένων και ταυτοποίησης. Το πρωτόκολλο OAuth2 Password Grant υλοποιείται από την υπηρεσία αυθεντικοποίησης στο αρχείο *auth.py*, διασφαλίζοντας ένα ασφαλές πλαίσιο για τη διαχείριση χρηστών. Επίσης, τα αυστηρά χρονικά όρια της τάξης των 10 δευτερολέπτων που έχουν οριστεί, περιορίζουν όλες τις κλήσεις προς εξωτερικές διεπαφές προγραμματισμού APIs, αποτρέποντας φαινόμενα εμπλοκής του συστήματος. Ταυτόχρονα, ένας εκτενής μηχανισμός καταγραφής έχει ενσωματωθεί, ο οποίος διατηρεί τις πλήρεις JSON αποκρίσεις των αιτημάτων, παρέχοντας αποτελεσματική ιχνηλασιμότητα σφαλμάτων και λεπτομερή αποσφαλμάτωση όταν παρουσιάζεται απρόβλεπτη συμπεριφορά. Ετσι, σε δεύτερο χρόνο μπορείς να ανατρέξεις στο αρχείο καταγραφής σφαλμάτων και να έχεις πλήρη εικόνα τι πήγε ή δεν πήγε καλά.

Απόδοση

Για τη διατήρηση της υψηλής απόκρισης της διεπαφής χρήστη UI, προηγμένοι μηχανισμοί προσωρινής αποθήκευσης ενσωματώνονται στην εφαρμογή Streamlit. Ειδικότερα, ο διακοσμητής *decorator @st.cache_data* χρησιμοποιείται για τη διαχείριση του φόρτου των ιστορικών πληροφοριών. Με αυτή την προσέγγιση εξασφαλίζεται ότι οι υπολογιστικά απαιτητικές διεργασίες ανάκτησης και επεξεργασίας δεδομένων δεν επαναλαμβάνονται χωρίς λόγο σε κάθε αλληλεπίδραση με τον χρήστη. Έτσι, το σύστημα διατηρεί την ταχύτητα και αποδοτικότητά του, εξασφαλίζοντας ομαλή πλοήγηση ακόμη και όταν ο όγκος των δεδομένων αυξάνεται σημαντικά με την πάροδο του χρόνου.

3.3 Αρχιτεκτονική υλοποίηση και διαχείριση δεδομένων

Μια δεδομενοκεντρική προσέγγιση ακολουθείται από την αρχιτεκτονική του συστήματος που προτείνεται, η οποία στηρίζεται στον αυστηρό διαχωρισμό αρμοδιοτήτων μέσω δύο διακριτών βάσεων δεδομένων SQLite. Με αυτήν την επιλογή επιτυγχάνεται η πλήρης αποσύνδεση των επιχειρησιακών πληροφοριών από τις διεργασίες παρακολούθησης πειραμάτων μηχανικής μάθησης.Για την αντιμετώπιση της πολυπλοκότητας που χαρακτηρίζει την επιχειρησιακή λειτουργία προϊόντων μηχανικής μάθησης, η παρούσα μελέτη υιοθετεί μια αρχιτεκτονική που βασίζεται σε περιέκτες μέσω των τεχνολογιών Docker και Docker Compose. Σύμφωνα με τους Nüst κ.ά. [39], η τεχνολογία

containerization προσφέρει μηχανικά αναγνώσιμες οδηγίες που περικλείουν το σύνολο του υπολογιστικού περιβάλλοντος, εξασφαλίζοντας την υπολογιστική αναπαραγωγιμότητα—μια θεμελιώδη προϋπόθεση για την επιστημονική ακεραιότητα. Ακολουθώντας τις αρχές των λειτουργιών μηχανικής μάθησης MLOps, η υποδομή αυτή προάγει την επεκτασιμότητα και τη συστημική συνοχή, αποσυνδέοντας τα λειτουργικά στοιχεία από την υποκείμενη υλική αρχιτεκτονική [40].

Ως κύρια γλώσσα ανάπτυξης επιλέχθηκε η Python 3.9, λόγω της καθιερωμένης θέσης της ως βασικού εργαλείου για τις ροές εργασίας επιστήμης δεδομένων και της ομαλής ενσωμάτωσής της με αυτοματοποιημένους αγωγούς CI/CD [40]. Τηρώντας την αναγκαιότητα της καθήλωσης εκδόσεων, η επιλογή της Python 3.9 αποτρέπει την αστάθεια που συνδέεται με τους «κινούμενους στόχους» στο πλαίσιο των εξαρτήσεων λογισμικού [39]. Η ρητή αυτή διαχείριση εκδόσεων εγγυάται την ιχνηλασιμότητα και την ικανότητα παραγωγής πανομοιότυπων πειραματικών αποτελεσμάτων σε όλο το φάσμα του κύκλου ζωής του προϊόντος, γεφυρώνοντας αποτελεσματικά το χάσμα ανάμεσα στην απόδειξη της ιδέας και στην παραγωγική εφαρμογή.

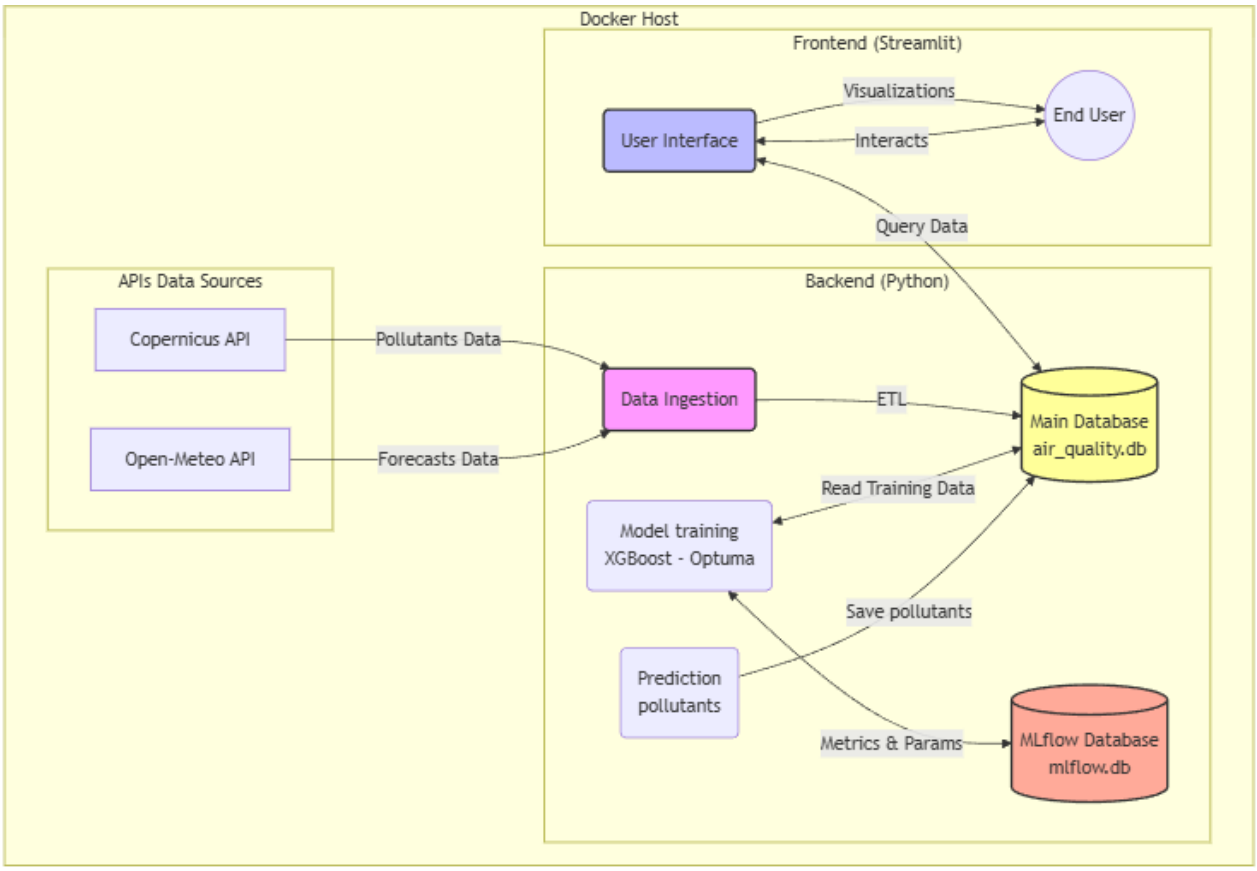
3.3.1 Διάγραμμα ροής δεδομένων

Όπως αποτυπώνεται στο Διάγραμμα 3.1, η ροή πληροφοριών εκκινεί με την άντληση πρωτογενών δεδομένων από εξωτερικές διεπαφές προγραμματισμού εφαρμογών APIs. Η διαδικασία ETL εξελίσσεται ως ακολούθως:

Εισαγωγή (*Ingestion*): Η συλλογή και αποθήκευση των ακατέργαστων δεδομένων πραγματοποιείται στην κύρια βάση Δεδομένων *air_quality.db*, η οποία λειτουργεί ως επιχειρησιακό αποθετήριο του συστήματος και περιλαμβάνει ιστορικά μετεωρολογικά δεδομένα, ιστορικές μετρήσεις ρύπων και τις τελικές προγνώσεις των καιρικών συνθηκών για τις επόμενες 16 ημέρες.

Εκπαίδευση (*Training*): Από την κύρια βάση δεδομένων ανακτώνται τα απαραίτητα δεδομένα για τη μονάδα εκπαίδευσης, ενώ η δεύτερη βάση δεδομένων MLflow *mlflow.db* χρησιμοποιείται για την τήρηση της ιστορικότητας των πειραμάτων. Συγκεκριμένα, οι μετρικές απόδοσης, οι εκδόσεις των μοντέλων και οι υπερπαράμετροι καταγράφονται σε αυτήν, εξασφαλίζοντας την ιχνηλασιμότητα της ερευνητικής διαδικασίας.

Οπτικοποίηση: Για την οπτικοποίηση, τα δεδομένα αντλούνται αποκλειστικά από την κύρια βάση και έτσι η διεπαφή παρουσιάζει στον τελικό χρήστη τα επεξεργασμένα αποτελέσματα.



Σχήμα 3.1: Διάγραμμα ροής δεδομένων συστήματος

3.3.2 Μικροϋπηρεσίες

Δύο αυτόνομες μικροϋπηρεσίες συγκροτούν το σύστημα, με κάθε μία να εκτελείται σε απομονωμένο περιβάλλον:

Υπηρεσία Backend

Ο υπολογιστικός πυρήνας αποτελείται από αυτήν την υπηρεσία και λειτουργεί στο παρασκήνιο. Η μονάδα εισαγωγής δεδομένων μέσω προγραμματισμένων εργασιών, ο αγωγός εκπαίδευσης του αλγορίθμου XGBoost και η μονάδα πρόβλεψης που τροφοδοτεί την κύρια βάση με νέες εκτιμήσεις περιλαμβάνονται στη λειτουργία της. Όλες αυτές οι υπηρεσίες περιλαμβάνονται στον φάκελο *backend* του προγράμματος και είναι διαχωρισμένες από τον κώδικα του UI.

Υπηρεσία Frontend

Μέσω του πλαισίου Streamlit υλοποιείται αυτή η υπηρεσία και είναι απόλυτα απομονωμένη από τον υπολογιστικό φόρτο του backend. Η σύνδεσή της γίνεται αποκλειστικά με την κύρια βάση δεδομένων *air_quality.db*, εξασφαλίζοντας μια διαδραστική και ταχεία εμπειρία χρήστη που παραμένει αδιατάρακτη από τις βαριές διεργασίες εκπαίδευσης μοντέλων. Τέλος, ο κώδικας του frontend ανήκει στον φάκελο *frontend* της εφαρμογής, τα οποία έχουν αναπτυχθεί στο περιβάλλον VSCode.

3.4 Σχεδιασμός βάσης δεδομένων

Το σύστημα SQLite επιλέχθηκε για τη διαχείριση των δεδομένων του συστήματος, λόγω της απλότητας και της χαμηλής κατανάλωσης πόρων. Ο «*serverless*» και αρχειοκεντρικός χαρακτήρας του αποτέλεσε τη βάση της απόφασης αυτής, καθώς απλοποιεί εντυπωσιακά τη διαδικασία ανάπτυξης με την εφαρμογή του Docker, αφού δεν χρειάζεται η εγκατάσταση και παραμετροποίηση αυτόνομου διακομιστή βάσης δεδομένων.

3.4.1 Σχήμα βάσης δεδομένων

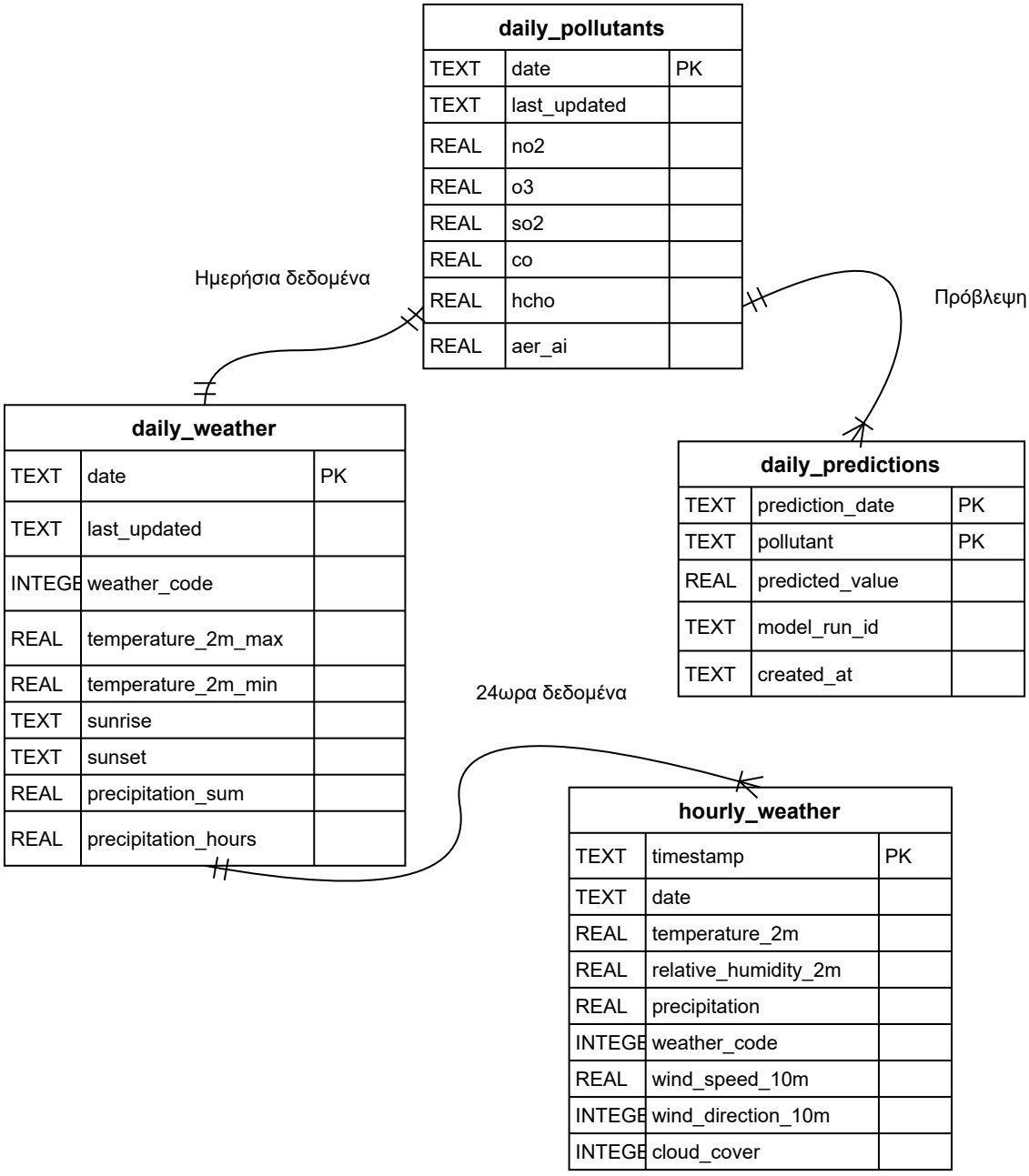
Με γνώμονα τη βελτιστοποίηση της ταχύτητας ανάκτησης πληροφοριών και τη διασφάλιση της ακεραιότητας των δεδομένων έχει σχεδιαστεί το σχήμα της βάσης. Τέσσερις κύριοι πίνακες απαρτίζουν τη δομή του *air_quality.db*:

***daily_pollutants*:** Το αποθετήριο των δεδομένων τηλεπισκόπησης από τον Sentinel-5P συνιστά αυτός ο πίνακας. Η στήλη *date* χρησιμοποιείται ως πρωτεύον κλειδί, εξασφαλίζοντας μία μοναδική εγγραφή ανά ημέρα. Ως αριθμοί κινητής υποδιαστολής *REAL* αποθηκεύονται οι συγκεντρώσεις των ατμοσφαιρικών ρύπων (*no2, o3, so2, co, hcho, aer_ai*).

***daily_weather*:** Συγκεντρωτικά μετεωρολογικά ημερήσια δεδομένα από την υπηρεσία Open-Meteo περιλαμβάνονται σε αυτόν, όπως το συνολικό ύψος βροχόπτωσης (*precipitation_sum*) και οι ακραίες θερμοκρασίες (*temperature_2m_max*, *temperature_2m_min*).

***hourly_weather*:** Μετεωρολογικές μεταβλητές σε υψηλή χρονική ανάλυση ωριαία αποθηκεύονται εδώ, συμπεριλαμβανομένων της ταχύτητας και διεύθυνσης του ανέμου (*wind_speed_10m*, *wind_direction_10m*). Ο δείκτης *idx_hourly_date* έχει υλοποιηθεί στη στήλη *date* για τη βελτιστοποίηση των ερωτημάτων σύνδεσης *joins* με τους ημερήσιους πίνακες, μειώνοντας σημαντικά τον χρόνο απόκρισης κατά την επεξεργασία μεγάλου όγκου δεδομένων.

***daily_predictions*:** Τα αποτελέσματα της προγνωστικής μοντελοποίησης φιλοξενούνται σε αυτόν τον πίνακα. Ένα σύνθετο πρωτεύον κλειδί στις στήλες *prediction_date* και *pollutant* επιβάλλεται από τη σχεδίαση. Καθοριστικός είναι αυτός ο περιορισμός, διότι εγγυάται ότι για κάθε ημερομηνία υπάρχουν μοναδικές προβλέψεις ανά ρύπο, διατηρώντας την ποιότητα των δεδομένων που τροφοδοτούν τη διεπαφή χρήστη.



Σχήμα 3.2: Διάγραμμα βάσης δεδομένων

Συνολικά, η εφαρμογή περιορισμών όπως το *NOT NULL* και τα σαφώς καθορισμένα πρωτεύοντα κλειδιά εξασφαλίζουν την ανθεκτικότητα του συστήματος έναντι εσφαλμένων εισαγωγών δεδομένων από τους αυτοματοποιημένους αγωγούς ETL.

3.5 Πηγές άντλησης δεδομένων APIs

Στην αυτοματοποιημένη επικοινωνία με δύο πρωτογενείς πηγές δεδομένων μέσω σύγχρονων διεπαφών προγραμματισμού βασίζεται η τροφοδοσία του συστήματος με μετεωρολογικές και περιβαλλοντικές πληροφορίες.

3.5.1 Ενσωμάτωση ατμοσφαιρικών και μετεωρολογικών δεδομένων μέσω Sentinel-5P και Open-Meteo API

Η αποστολή Sentinel-5P αποτελεί θεμελιώδες στοιχείο για την τηλεπισκόπηση της ατμόσφαιρας, κυρίως μέσω του οργάνου παρακολούθησης της τροπόσφαιρας TROPOMI. Σύμφωνα με τους Zhao κ.ά. [41], οι αποστολές Sentinel προσφέρουν υψηλή χωρική διακριτική ικανότητα και διαρκή παγκόσμια κάλυψη, καθιστώντας εφικτή την ακριβή ανίχνευση ιχνοαερίων, όπως το NO_2 και το O_3. Για την επιτυχή ερμηνεία αυτών των συνόλων δεδομένων, η ενσωμάτωση με API μετεωρολογικής επανανάλυσης καθίσταται απαραίτητη. Οι Ramadan κ.ά. [42] υπογραμμίζουν ότι μετεωρολογικές παράμετροι συμπεριλαμβανομένων της ανεμολογικής δυναμικής, της θερμοκρασίας και της υγρασίας επηρεάζουν σημαντικά την ατμοσφαιρική ποιότητα.

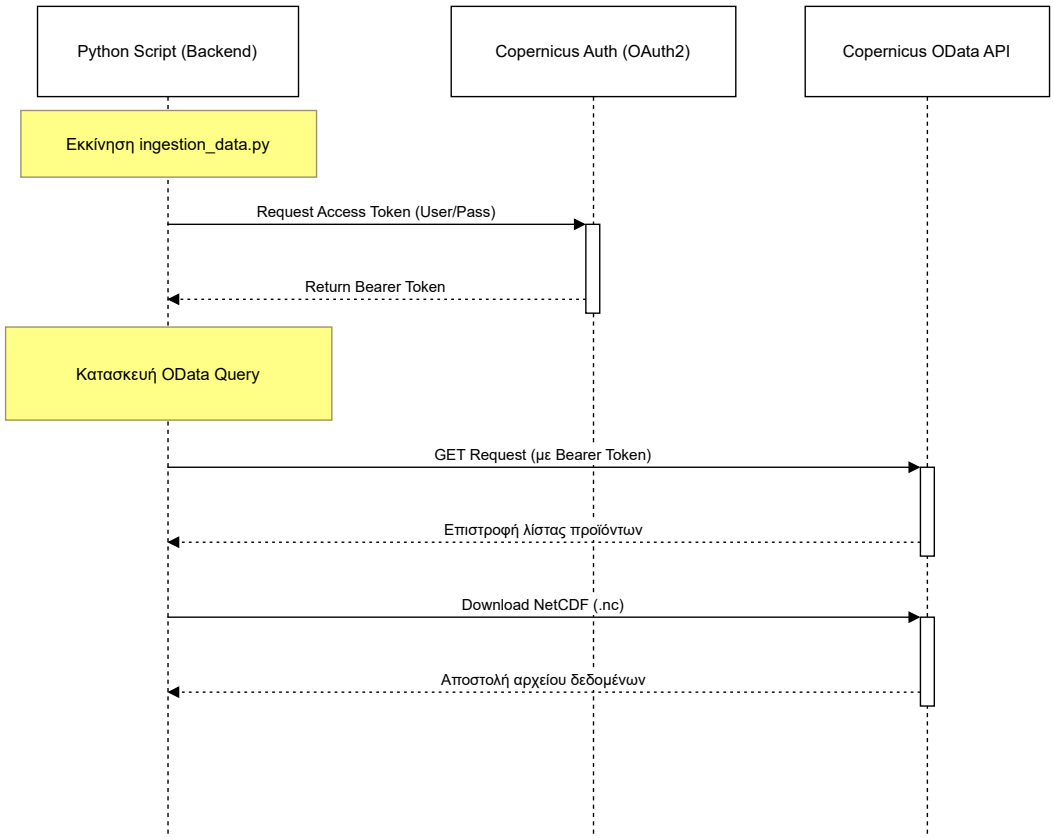
Η πρόσβαση σε αυτές τις μεταβλητές μέσω API, όπως το Open-Meteo, καθιστά δυνατή την αποτελεσματική μοντελοποίηση της διασποράς ρύπων και της χημικής μεταφοράς μεγάλων αποστάσεων από τους ερευνητές. Ενώ τα δορυφορικά δεδομένα προσφέρουν στιγμιαία αποτυπώματα των αεριακών συγκεντρώσεων, η συνδυαστική αξιοποίησή τους με τοπικά μετεωρολογικά δεδομένα διαμορφώνει ένα ολιστικό πλαίσιο παρακολούθησης. Συνεπώς, η ολοκληρωμένη αυτή μεθοδολογία επιτυγχάνει ανώτερη ακρίβεια και δομική συνοχή σε σύγκριση με την ανάλυση διακριτών συνόλων δεδομένων, γεφυρώνοντας αποτελεσματικά το χάσμα ανάμεσα στις παγκόσμιες παρατηρήσεις και την τοπική περιβαλλοντική δυναμική.

3.5.2 Copernicus Dataspace Ecosystem Sentinel-5P

Μέσω του Copernicus Dataspace Ecosystem πραγματοποιείται η ανάκτηση των δορυφορικών παρατηρήσεων και συγκεκριμένα από τον Sentinel-5P, χρησιμοποιώντας τα πρωτόκολλα OData και STAC. Ωστόσο, επιλέχθηκε τελικά το OData λόγω της απλουστευμένης διαδικασίας λήψης δεδομένων μετά τις αναβαθμίσεις του Copernicus. Το πρότυπο OAuth2 Password Grant υλοποιεί τη διαδικασία ταυτοποίησης, με την αξιοποίηση της κλάσης *LegacyApplicationClient* από τη βιβλιοθήκη *oauthlib*. Με αυτήν την προσέγγιση το σύστημα αποκτά ένα διακριτικό πρόσβασης χωρίς την ανάγκη ανθρώπινης παρέμβασης, εξασφαλίζοντας την απρόσκοπτη λειτουργία των υπηρεσιών στο παρασκήνιο.

Στο αρχείο γραμμένο με Python το *ingestion_data.py* εκτελείται η στρατηγική υποβολής ερωτημάτων, κατασκευάζοντας δυναμικά αιτήματα OData. Με τη χρήση εξειδικευμένων φίλτρων, συγκεκριμένα της συνάρτησης *OData.CSC.Intersects*, η λήψη δεδομένων περιορίζεται αποκλειστικά

εντός του γεωγραφικού πλαισίου της Θεσσαλονίκης, όπως αυτό καθορίζεται στις παραμέτρους παραμετροποίησης.



Σχήμα 3.3: Sequence Diagram της διαδικασίας ανάκτησης δεδομένων από το Copernicus API.

Την επεξεργασία των πρωτογενών αρχείων αφορά ένα κρίσιμο στάδιο της διαδικασίας ETL, το οποίο υλοποιείται από το αρχείο *processing.py* στον φάκελο *backend*. Σε μορφή NetCDF (.nc) μεταφορτώνονται τα δεδομένα, τα οποία περιέχουν πολυδιάστατους πίνακες με τις τιμές των εικονοστοιχείων για το καθορισμένο Bounding Box. Με τις βιβλιοθήκες *xarray* και *numpy*, το σύστημα διαβάζει τα αρχεία, εφαρμόζει χωρική μάσκα βάσει των γεωγραφικών ορίων και υπολογίζει τον χωρικό μέσο όρο των έγκυρων παρατηρήσεων *np.nanmean*. Η διαδικασία αυτή είναι θεμελιώδης, διότι μετασχηματίζει ογκώδεις ακατέργαστους δορυφορικούς όγκους δεδομένων σε μεμονωμένες αριθμητικές τιμές *float* ανά ημέρα και ανά ρύπο, καθιστώντας εφικτή και αποδοτική την αποθήκευσή τους στη βάση δεδομένων SQLite.

3.5.3 Open-Meteo API (Weather Data)

Η διεπαφή Open-Meteo REST API αξιοποιείται από το σύστημα για την άντληση μετεωρολογικών δεδομένων, αποτελώντας μια υπηρεσία ανοικτής πρόσβασης που λειτουργεί με βάση τον προσδιο-

ρισμό γεωγραφικών συντεταγμένων. Δύο διακριτά σημεία πρόσβασης χρησιμοποιούνται από την αρχιτεκτονική της λύσης: το Archive για την ανάκτηση ιστορικών δεδομένων που είναι απαραίτητα για την εκπαίδευση των μοντέλων, και το Forecast για τη λήψη προγνώσεων με 16ήμερο ορίζοντα. Το σύστημα αιτείται ημερήσιες συγκεντρωτικές τιμές, όπως το συνολικό ύψος βροχόπτωσης, η μέγιστη θερμοκρασία και ο κωδικός καιρού, καθώς και ωριαία χαρακτηριστικά, συμπεριλαμβανομένων της σχετικής υγρασίας και της ταχύτητας και διεύθυνσης του ανέμου. Τα δεδομένα λαμβάνονται σε μορφή JSON, η οποία αναλύεται και εισάγεται απευθείας στις αντίστοιχες δομές της βάσης δεδομένων. Η συνδυαστική χρήση των ανωτέρω πηγών επιτρέπει τη δημιουργία ενός πλούσιου συνόλου χαρακτηριστικών, το οποίο συνιστά τη βάση για την προγνωστική μοντελοποίηση με τον αλγόριθμο XGBoost.

Πίνακας 3.1: Μετεωρολογικές παράμετροι που ανακτώνται από το Open-Meteo API.

Παράμετρος	**Τύπος**	**Περιγραφή**
Precipitation Sum	Daily	Συνολικό ημερήσιο ύψος βροχόπτωσης (mm)
Temperature Max	Daily	Μέγιστη ημερήσια θερμοκρασία (2m)
Weather Code	Daily	Κωδικός καιρού (WMO code)
Relative Humidity	Hourly	Σχετική υγρασία αέρα (2m)
Wind Speed	Hourly	Ταχύτητα ανέμου (10m)
Wind Direction	Hourly	Διεύθυνση ανέμου (10m)

Η ολοκλήρωση του σχεδιαστικού πλαισίου επιτυγχάνεται μέσω της σύζευξης των διεπαφών Open-Meteo και Copernicus, οι οποίες αποτελούν τον κρίσιμο σύνδεσμο ανάμεσα στις πρωτογενείς πηγές δεδομένων και την αρχιτεκτονική βασισμένη σε μικροϋπηρεσίες που προτείνεται. Το σύστημα, ξεκινώντας από την αρχική διατύπωση των λειτουργικών προδιαγραφών και εκτεινόμενο στον σχεδιασμό της δομής των αποθηκευτικών χώρων —όπου επιχειρησιακά δεδομένα και μετρικές παρακολούθησης διαχωρίζονται μέσω της πλατφόρμας MLflow— έως την εξασφάλιση της επαναληψιμότητας που παρέχει η τεχνολογία Docker, συγκροτεί πλέον μια ενιαία και αποτελεσματική υπολογιστική δομή. Συνεπώς, έχοντας εδραιωθεί το σχεδιαστικό υπόβαθρο, η έρευνα προσανατολίζεται στην πρακτική διάσταση της ανάπτυξης, αφήνοντας πίσω το θεωρητικό στάδιο της ανάλυσης απαιτήσεων και της αρχιτεκτονικής σύλληψης. Το κεφάλαιο που ακολουθεί επικεντρώνεται στην τεχνική υλοποίηση του συνόλου, αναλύοντας λεπτομερώς την οργάνωση του πηγαίου κώδικα, την κατασκευή του αυτοματοποιημένου αγωγού εξαγωγής-μετασχηματισμού-φόρτωσης δεδομένων, καθώς και τις παραμετρικές ρυθμίσεις που καθιστούν το μοντέλο πρόβλεψης λειτουργικά διαθέσιμο για τις ανάγκες του αστικού ιστού της Θεσσαλονίκης.

Κεφάλαιο 4

Υλοποίηση συστήματος Backend & ML

4.1 Εισαγωγή

Στο παρόν κεφάλαιο επιχειρείται η λεπτομερής τεχνική τεκμηρίωση της προτεινόμενης λύσης, μεταθέτοντας το ενδιαφέρον από την εννοιολογική σχεδίαση και την αρχιτεκτονική ανάλυση του Κεφαλαίου 3 προς την πραγματική ανάπτυξη λογισμικού και την υλοποίηση κώδικα.

Η συγκεκριμένη ενότητα επιδιώκει να καταγράψει με λεπτομέρεια τις τεχνολογικές επιλογές, τα προγραμματιστικά πρότυπα και τα εργαλεία που αξιοποιήθηκαν προκειμένου να μετατραπεί το θεωρητικό πλαίσιο σε μία πλήρως λειτουργική και αυτοματοποιημένη πλατφόρμα για την πρόβλεψη της ατμοσφαιρικής ρύπανσης. Οι αρχές της σπονδυλωτής αρχιτεκτονικής διέπουν την υλοποίηση, εξασφαλίζοντας την αυτόνομη λειτουργία κάθε υποσυστήματος, ενώ παράλληλα εγγυώνται την πλήρη συνεργασία μεταξύ των διαφόρων μονάδων. Το κεφάλαιο οργανώνεται ακολουθώντας τη φυσική διαδρομή των δεδομένων μέσα στο σύστημα, ξεκινώντας από τις πρωτογενείς πηγές πληροφορίας και φτάνοντας στη πραγματική φάση των προγνωστικών αποτελεσμάτων. Συγκεκριμένα, η Ενότητα 4.2 εξετάζει τον αγωγό ETL, υπεύθυνο για την αυτοματοποιημένη άντληση δεδομένων από την πλατφόρμα Copernicus και την υπηρεσία Open-Meteo, καθώς επίσης και για τον εξειδικευμένο γεωχωρικό μετασχηματισμό αρχείων NetCDF. Στην συνέχεια, η Ενότητα 4.3 επικεντρώνεται στον αγωγό Μηχανικής Μάθησης, καλύπτοντας τη διαδικασία εκπαίδευσης μέσω του αλγορίθμου XGBoost, τη βελτιστοποίηση υπερπαραμέτρων με το πλαίσιο Optuna, καθώς και τη διαχείριση του κύκλου ζωής μοντέλων μέσω του MLflow. Στην Ενότητα 4.4 παρουσιάζεται η μηχανή πρόβλεψης και εφαρμόζεται η αναδρομική μεθοδολογία για την παραγωγή προγνώσεων σε πολλαπλά χρονικά βήματα. Τέλος, η Ενότητα 4.5 αποτυπώνει την υποδομή του συστήματος, η οποία στηρίζεται στην τεχνολογία Docker, διασφαλίζοντας τη φορητότητα, την απομόνωση περιβαλλόντων και την κλιμακωσιμότητα της οριστικής υλοποίησης.

4.2 Διαδικασία συλλογής ακατέργαστων δεδομένων ETL Pipeline

4.2.1 Αυθεντικοποίηση και λήψη δορυφορικών δεδομένων

Το αρχικό και κρισιμότερο βήμα του αγωγού ETL συνίσταται στην άντληση δεδομένων από την πλατφόρμα Copernicus Dataspace Ecosystem. Η είσοδος στις υπηρεσίες τηλεπισκόπησης πραγματοποιείται μέσω του πρωτοκόλλου OAuth2 Password Grant. Στο επίπεδο τεχνικής υλοποίησης, το αρχείο *auth.py* χρησιμοποιεί τη βιβλιοθήκη *requests_oauthlib*, και συγκεκριμένα την κλάση *LegacyApplicationClient*, για τη διεκπεραίωση της ανταλλαγής διακριτικών πρόσβασης τα tokens.

```python
from urllib3.util.retry import Retry                    # auth.py
from requests.adapters import HTTPAdapter
from requests_oauthlib import OAuth2Session
from oauthlib.oauth2 import LegacyApplicationClient

def get_oauth_session(client_id, username, password, token_url):
# Ορισμός συνολικών επαναπροσπάθειων
retry_strategy = Retry(
total=3,                                    # Μέγιστο πλήθος προσπαθειών
backoff_factor=1,                           # Συντελεστής οπισθοχώρησης
status_forcelist=[500, 502, 503, 504],      # Σφάλματα που ενεργοποιούν retry
allowed_methods=["POST", "GET"]
)

adapter = HTTPAdapter(max_retries=retry_strategy)
try:
    client = LegacyApplicationClient(client_id=client_id)
    oauth = OAuth2Session(client=client)

    # Κάθε αίτημα σε http/https θα περνάει από τον έλεγχο retry
    oauth.mount('https://', adapter)
    oauth.mount('http://', adapter)

    # Ανάκτηση Token
    oauth.fetch_token(
    token_url=token_url,
    username=username,
    password=password,
    client_id=client_id
    )
    return oauth
except ..
```

Κώδικας 4.1: Υλοποίηση μηχανισμού επαναπροσπάθειας και αυθεντικοποίησης OAuth2

Η ανθεκτικότητα της διαδικασίας ταυτοποίησης αποτέλεσε προτεραιότητα κατά την ανάπτυξη. Μέσω των κλάσεων *HTTPAdapter* και *Retry*, η συνάρτηση *get_oauth_session* ενσωματώνει εξελιγμένη στρατηγική επαναπροσπάθειας. Συγκεκριμένα, το σύστημα παραμετροποιήθηκε για την αυτόματη εκτέλεση έως τριών προσπαθειών (*total=3*) όταν εμφανίζονται παροδικά σφάλματα δικτύου ή αστοχίες διακομιστή (κωδικοί κατάστασης 500, 502, 503, 504). Επιπλέον, ένας συντελεστής προ-

οδευτικής οπισθοχώρησης *backoff_factor* εξασφαλίζει ότι ο αγωγός δεν διακόπτεται απροσδόκητα από προσωρινές δυσλειτουργίες των API, αναμένοντας αντίθετα τη σταθεροποίηση της σύνδεσης.

Το βασικό αρχείο *ingestion_data.py* αναλαμβάνει την καθημερινή επιχειρησιακή λειτουργία του συστήματος. Η σχεδίαση του υποσυστήματος επιτρέπει την πλήρως αυτοματοποιημένη λειτουργία μέσω προγραμματισμένων εργασιών. Η βασική λογική στηρίζεται στον δυναμικό προσδιορισμό της ημερομηνίας, ορίζοντάς την προεπιλεγμένη ως την προηγούμενη ημέρα από την τρέχουσα με την εντολή *datetime.now() - timedelta(days=1)*. Αυτή η προσέγγιση είναι απαραίτητη εξαιτίας της εγγενούς καθυστέρησης στη δημοσίευση των δορυφορικών προϊόντων Sentinel-5P. Ως κεντρικός ενορχηστρωτής, το αρχείο ingestion_data.py συντονίζει διαδοχικά την ταυτοποίηση, τη λήψη δεδομένων ατμοσφαιρικών ρύπων, την ανάκτηση μετεωρολογικών δεδομένων και την αποθήκευσή όλων αυτών στην βάση δεδομένων air_quality.db.

4.2.2 Διαδικασία αυτοματοποίησης συλλογής

Το αρχείο *ingestion_data.py* αναλαμβάνει την καθημερινή επιχειρησιακή λειτουργία του συστήματος. Η σχεδίαση του υποσυστήματος επιτρέπει την πλήρως αυτοματοποιημένη εκτέλεση μέσω προγραμματισμένων εργασιών λειτουργικού συστήματος. Η βασική λογική στηρίζεται στον δυναμικό προσδιορισμό της ημερομηνίας, ορίζοντάς την προεπιλεγμένη ως την προηγούμενη ημέρα από την τρέχουσα με την εντολή *datetime.now() - timedelta(days=1)*. Αυτή η προσέγγιση είναι απαραίτητη εξαιτίας της εγγενούς καθυστέρησης στη δημοσίευση των δορυφορικών προϊόντων Sentinel-5P.

Ως κεντρικός ενορχηστρωτής, το αρχείο *ingestion_data.py* συντονίζει διαδοχικά την ταυτοποίηση, τη λήψη δεδομένων ατμοσφαιρικών ρύπων, την ανάκτηση μετεωρολογικών δεδομένων και την αποθήκευσή όλων αυτών στην βάση δεδομένων *air_quality.db*, όπως αποτυπώνεται στον Κώδικα 4.2.

```
from datetime import datetime, timedelta

def main():                          # ingestion_data.py
init_db()

yesterday_dt = datetime.now() - timedelta(days=1) # (T-1 ημέρα)

date_str = yesterday_dt.strftime('%Y-%m-%d')
start_time = yesterday_dt.strftime('%Y-%m-%dT00:00:00Z')
end_time = yesterday_dt.strftime('%Y-%m-%dT23:59:59Z')
time_interval_list = [start_time, end_time]

# Εκτέλεση συναρτήσεων APIs για την άντληση δεδομένων
ingest_historical_pollutants(date_str, time_interval_list)
ingest_weather(date_str)

```

Κώδικας 4.2: Κύρια κλήση της συνάρτηση και υπολογισμού προηγούμενης ημέρας

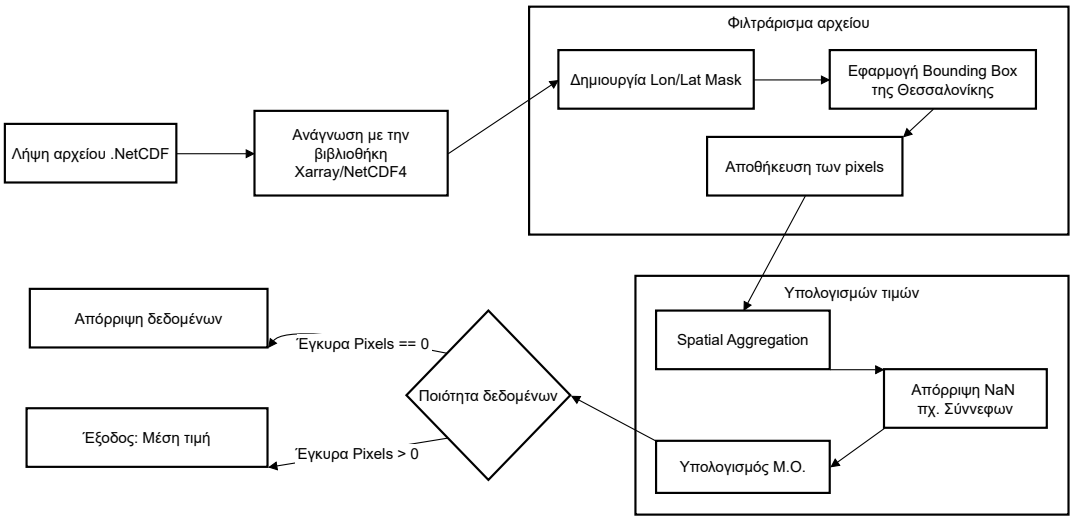
Η αρχική πληρότητα του συστήματος με ιστορικά δεδομένα, αναγκαία για την εκπαίδευση των μοντέλων, επιτυγχάνεται μέσω του αρχείου *backfill_data.py*. Αντίθετα με τον ημερήσιο αγωγό,

αυτό το εργαλείο λειτουργεί ως μηχανισμός αρχικοποίησης. Με τη βιβλιοθήκη pandas και τη συνάρτηση *pandas.date_range*, εκτελείται βρόχος επανάληψης σε χειροκίνητα καθορισμένο χρονικό εύρος *START_DATE* έως *END_DATE*. Κεντρικό χαρακτηριστικό της αρχιτεκτονικής αποτελεί η επαναχρησιμοποίηση της συνάρτησης *ingest_historical_pollutants* από το κύριο σενάριο εισαγωγής. Έτσι εξασφαλίζεται ότι τα ιστορικά δεδομένα υφίστανται επεξεργασία πανομοιότυπη με αυτή των δεδομένων πραγματικής λειτουργίας, διατηρώντας τη στατιστική συνέπεια και την ακεραιότητα των εισόδων προς τον αλγόριθμο μηχανικής μάθησης.

4.2.3 Επεξεργασία των δεδομένων μετά από λήψη του αρχείου NetCDF

Το στάδιο επεξεργασίας των ακατέργαστων δορυφορικών αρχείων συνιστά κρίσιμη φάση μετασχηματισμού, κατά την οποία η πληροφορία μετατρέπεται από γεωχωρικά σύνολα σε δομημένες αριθμητικές τιμές. Για τον χειρισμό της πολυπλοκότητας του ιεραρχικού μορφότυπου NetCDF, εφαρμόζεται η βιβλιοθήκη xarray με τη μηχανή netcdf4. Η επιλογή αυτή καθιστά δυνατό τον χειρισμό πολυδιάστατων δορυφορικών δεδομένων ως επισημασμένων πινάκων, απλοποιώντας σημαντικά τη συσχέτιση μεταβλητών ρύπανσης με γεωγραφικές συντεταγμένες.

Η τεχνική Boolean Masking χρησιμοποιείται για την απομόνωση της υπό μελέτη περιοχής. Η συνάρτηση *process_s5p_file* ανακτά από το αρχείο τους πίνακες γεωγραφικού μήκους και πλάτους, κατασκευάζοντας στη συνέχεια λογική μάσκα με βάση το προκαθορισμένο *Bounding Box* των γεωγραφικών ορίων της Θεσσαλονίκης. Η πράξη *combined_mask = lon_mask & lat_mask* εξασφαλίζει την αυστηρή επιλογή εικονοστοιχείων που εμπίπτουν αποκλειστικά στον αστικό ιστό, αποκλείοντας περιττές πληροφορίες από παρακείμενες περιοχές.



Σχήμα 4.1: Διάγραμμα απεικόνισης λειτουργίας κατά την λήψη του αρχείου .NetCDF.

Ακολούθως, η μείωση διαστασιμότητας των δεδομένων επιτυγχάνεται μέσω της διαδικασίας Spatial Aggregation. Επιδιώκεται ο μετασχηματισμός του δισδιάστατου πλέγματος επιλεγμένων εικονο-

στοιχείων σε μία ενιαία, αντιπροσωπευτική τιμή συγκέντρωσης ανά ημέρα. Η συνάρτηση *numpy.nanmean* διαδραματίζει θεμελιώδη ρόλο σε αυτό το στάδιο, δεδομένου ότι οι δορυφορικές παρατηρήσεις συχνά εμπεριέχουν τιμές *NaN* (*Not a Number*) λόγω υψηλής νεφοκάλυψης ή σφαλμάτων ανάκτησης. Η *nanmean* υπολογίζει τον μέσο όρο βασιζόμενη αποκλειστικά σε έγκυρες παρατηρήσεις, απορρίπτοντας αυτόματα συννεφιασμένα ή ελαττωματικά pixels.

Ως τελικό βήμα, εφαρμόζεται κρίσιμη δικλείδα ελέγχου ποιότητας: όταν το σύνολο των έγκυρων εικονοστοιχείων εντός των ορίων είναι μηδενικό *valid_pixels == 0*, το αρχείο απορρίπτεται χωρίς η τιμή να εισαχθεί στη βάση δεδομένων. Η πρακτική αυτή αποτρέπει την εισαγωγή θορύβου και κενών καταχωρήσεων στην επιχειρησιακή βάση, εγγυώμενη την ακεραιότητα των δεδομένων που θα τροφοδοτήσουν τον αλγόριθμο XGBoost.

4.2.4 Μοντελοποίηση με XGBoost και βελτιστοποίηση με την χρήση Optuna

Η σύζευξη του XGBoost με το πλαίσιο βελτιστοποίησης Optuna συνιστά σημαντική εξέλιξη στο πεδίο της προγνωστικής μοντελοποίησης περιβαλλοντικών και μετεωρολογικών χρονοσειρών.

Το XGBoost επιδεικνύει εξαιρετική αποτελεσματικότητα σε πολυδιάστατα σύνολα δεδομένων, χάρη στην ικανότητά του να αναγνωρίζει σύνθετα μη γραμμικά μοτίβα και στην εγγενή του δυνατότητα διαχείρισης ελλειπουσών τιμών μέσω του διαχωρισμού δέντρων με επίγνωση αραιότητας [43]. Ως μέθοδος συνόλου που βασίζεται σε δενδρικές δομές, αξιοποιεί έναν στόχο μάθησης δεύτερης τάξης και όρους κανονικοποίησης προκειμένου να περιορίσει την υπερπροσαρμογή, εξασφαλίζοντας ισχυρή ικανότητα γενίκευσης στις ποικίλες χωροχρονικές κλίμακες που διακρίνουν τα δεδομένα ατμοσφαιρικής ποιότητας [44].

Η απόδοση του XGBoost ενισχύεται ουσιαστικά μέσω του συνδυασμού με το Optuna, μια αυτοματοποιημένη μεθοδολογία βελτιστοποίησης υπερπαραμέτρων. Σε αντίθεση με τις συμβατικές τεχνικές *Grid* ή *Random Search*, οι οποίες απαιτούν εξαντλητικούς υπολογιστικούς πόρους και δεν διαθέτουν επαναληπτική μαθησιακή ικανότητα, το Optuna χρησιμοποιεί τον Tree-structured Parzen Estimator [45]. Η Μπεϋζιανή αυτή προσέγγιση καθιστά εφικτή την ταχύτερη σύγκλιση, κατασκευάζοντας δυναμικά χώρους αναζήτησης και ιεραρχώντας συνδυασμούς παραμέτρων που ιστορικά έχουν επιτύχει ανώτερα αποτελέσματα επικύρωσης.

Εμπειρικά ευρήματα καταδεικνύουν ότι η εν λόγω συνέργεια βελτιώνει συνεχώς τους βασικούς δείκτες επίδοσης, μειώνοντας πρωτίστως το ριζικό μέσο τετραγωνικό σφάλμα RMSE και το μέσο απόλυτο σφάλμα MAE, ενώ παράλληλα μεγιστοποιεί τον συντελεστή προσδιορισμού (R^2) [46]. Επομένως, ο συνδυασμός XGBoost-Optuna αποτελεί μια εξαιρετικά επεκτάσιμη και αποδοτική λύση για την περιβαλλοντική παρακολούθηση και την πρόβλεψη σε πραγματικό χρόνο.

4.3 Διαδικασία εκπαίδευσης μοντέλου

Μέσω του αρχείου *train.py* υλοποιείται η εκπαίδευση των μοντέλων παλινδρόμησης XGBoost, ακολουθώντας δομημένη ροή που ενσωματώνει την επεξεργασία δεδομένων, τη βελτιστοποίηση

υπερπαραμέτρων και τη διακυβέρνηση πειραμάτων.

4.3.1 Προετοιμασία δεδομένων και διαμόρφωση παραμέτρων

Η αρχική φάση εκπαίδευσης στηρίζεται στη συνάρτηση *load_and_prepare_data()*, η οποία αναλαμβάνει την ανάκτηση πρωτογενών δεδομένων από τη βάση SQLite. Η συγχώνευση με SQL queries της εντολής inner join μεταξύ ιστορικών επιπέδων ρύπων και μετεωρολογικών δεδομένων πραγματοποιείται από το σύστημα. Στην απόδοση του μοντέλου, η διαδικασία που εφαρμόζεται, διαδραματίζει καθοριστικό ρόλο εστιάζοντας στις ακόλουθες κατηγορίες:

- **Κυκλικά χαρακτηριστικά (*sesonality*):** Η συνάρτηση *create_date_features()* μετασχηματίζει τον χρόνο σε τριγωνομετρικές συναρτήσεις *month_sin* και *month_cos*. Έτσι, το μοντέλο αντιλαμβάνεται την περιοδικότητα και εποχικότητα των ρύπων (π.χ. τα αυξημένα επίπεδα NO_2 κατά τη χειμερινή περίοδο λόγω θέρμανσης), διατηρώντας παράλληλα τη συνέχεια μεταξύ Δεκεμβρίου και Ιανουαρίου.
- **Μεταβλητή προηγούμενης μέρας (*Lag Features*):** Δημιουργείται η μεταβλητή *target_pollutant_lag_1*, αντιπροσωπεύοντας την τιμή του ρύπου της προηγούμενης ημέρας ($t-1$). Σε χρονοσειρές, αυτή αποτελεί τον ισχυρότερο προγνωστικό παράγοντα, δεδομένου ότι η παρούσα ατμοσφαιρική κατάσταση εξαρτάται άμεσα από την πρόσφατη ιστορία.
- **Συγκέντρωση δεδομένων (*Aggregation*):** Δεδομένης της ημερήσιας ανάλυσης των δορυφορικών δεδομένων έναντι της ωριαίας των μετεωρολογικών, εφαρμόζεται η μέθοδος *groupby-('date').agg()* για τον υπολογισμό στατιστικών μεγεθών (*mean*, *min*, *max*, *std*). Χαρακτηριστικά όπως θερμοκρασία και ταχύτητα ανέμου ενσωματώνονται στο μοντέλο με τρόπο που αναδεικνύει τις ακραίες τιμές και την ενδοημερήσια διακύμανση.

4.3.2 Βελτιστοποίηση υπερπαραμέτρων με το Optuna

Για την επίτευξη μέγιστης ακρίβειας, το πλαίσιο AutoML Optuna χρησιμοποιείται στη διαδικασία βελτιστοποίησης. Η συνάρτηση *objective()* ορίζει τη διαδικασία, επιδιώκοντας την ελαχιστοποίηση του MAE σε σύνολο επικύρωσης.

Οι κρίσιμες υπερπαράμετροι του αλγορίθμου XGBoost περιλαμβάνονται στο διάστημα αναζήτησης:

- *n_estimators*: Αριθμός δέντρων απόφασης (εύρος 300 έως 1000).
- *learning_rate*: Ρυθμός εκμάθησης για την αποφυγή υπερπροσαρμογής.
- *max_depth*: Μέγιστο βάθος δέντρων για την καταγραφή σύνθετων αλληλεπιδράσεων.
- *subsample* και *colsample_bytree*: Ποσοστά δειγματοληψίας δεδομένων και χαρακτηριστικών.

Πίνακας 4.1: Χώρος αναζήτησης υπερπαραμέτρων (Search Space) για τον αλγόριθμο XGBoost.

Υπερπαράμετρος	**Εύρος / Τιμές**	**Περιγραφή**
n_estimators	$300 - 1000$	Πλήθος δέντρων απόφασης (int)
learning_rate	$0.05 - 0.30$	Ρυθμός εκμάθησης (float)
max_depth	$5 - 15$	Μέγιστο βάθος δέντρου
subsample	$0.7 - 1.0$	Ποσοστό δειγματοληψίας δεδομένων
colsample_bytree	$0.7 - 1.0$	Ποσοστό δειγματοληψίας χαρακτηριστικών
min_child_weight	$1 - 5$	Ελάχιστο άθροισμα βαρών σε κόμβο

Η βελτιστοποίηση διεξάγεται για 50 δοκιμές *OPTUNA_N_TRIALS*, εξασφαλίζοντας την ανεύρεση βέλτιστης διαμόρφωσης που εξισορροπεί την πολυπλοκότητα με τη γενικευτική ικανότητα του μοντέλου.

4.3.3 Παρακολούθηση πειραμάτων με το MLflow

Για την οργάνωση και διαχείριση του κύκλου ζωής των αλγορίθμων μηχανικής μάθησης, αξιοποιείται η πλατφόρμα MLflow, η οποία συγκεντρώνει το σύνολο των εκτελέσεων στο πλαίσιο του πειράματος *«Pollutant_Forecasting»*. Κάθε επαναληπτική διαδικασία εκπαίδευσης αποτελεί ένα αυτόνομο «Run», γεγονός που εξασφαλίζει την πλήρη ιχνηλασιμότητα και την ικανότητα αναπαραγωγής των πειραματικών ευρημάτων.

Το αρχείο *train.py* ενσωματώνει προηγμένες λειτουργίες καταγραφής, οι οποίες αναλύονται στα ακόλουθα επίπεδα:

- **Μεταδεδομένα και ταξινόμηση:** Η χρήση της συνάρτησης *set_tag* επιτρέπει την επισήμανση κάθε εκτέλεσης με βασικά χαρακτηριστικά, όπως ο συγκεκριμένος ρύπος (*pollutant*) και η αρχιτεκτονική που εφαρμόζεται *model_type*. Η προσέγγιση αυτή διευκολύνει την ταχεία αναζήτηση και φιλτράρισμα μέσω του γραφικού περιβάλλοντος του MLflow.
- **Παράμετροι και μετρικές:** Αποθηκεύονται οι βέλτιστες τιμές υπερπαραμέτρων που εξήχθησαν από τη διαδικασία βελτιστοποίησης Optuna, ο κατάλογος των εισερχόμενων χαρακτηριστικών (*features*), αλλά και ο συντελεστής κανονικοποίησης. Επιπλέον, καταχωρείται η τελική επίδοση του μοντέλου όσον αφορά στην τιμή MAE.
- **Validation και serialization:** Κατά τη φάση αποθήκευσης *log_model*, εφαρμόζονται οι μηχανισμοί *infer_signature* και *input_example*, οι οποίοι καθορίζουν με ακρίβεια τη δομή των αναμενόμενων δεδομένων εισόδου. Μέσω αυτής της επικύρωσης σχήματος, αποτρέπονται πιθανά προβλήματα συμβατότητας κατά την επακόλουθη αξιοποίηση του μοντέλου.

Πίνακας 4.2: Συγκεντρωτικός πίνακας στοιχείων παρακολούθησης στο MLflow.

Κατηγορία	**Στοιχείο Καταγραφής**	**Τεχνικό Κλειδί / Λειτουργία**
Experiment	Όνομα πειράματος	*Pollutant_Forecasting_XGB*
Tags	Ρύπος-Στόχος	*pollutant* (π.χ. no2, o3)
	Τύπος αλγορίθμου	*model_type* («XGBoost»)
	Κατάσταση εκτέλεσης	*status* («success», «error»)
Params	Υπερπαράμετροι	*learning_rate*, *max_depth*
	Χαρακτηριστικά εισόδου	*features_list*
	Συντελεστής κλιμάκωσης	*scale_factor*
Metrics	Σφάλμα πρόβλεψης	*best_mae*
Artifacts	Αποθήκευση μοντέλου	*mlflow.xgboost.log_model*
	Επικύρωση σχήματος	*infer_signature*, *input_example*
	Model Registry	*registered_model_name*

- **Model registry:** Η ολοκλήρωση της διαδικασίας συνοδεύεται από την αυτοματοποιημένη καταχώρηση του μοντέλου στο κεντρικό αποθετήριο, γεγονός που απλοποιεί τη διαχείριση πολλαπλών εκδόσεων και επιτρέπει την άμεση πρόσβαση στην καταλληλότερη έκδοση για ανάπτυξη σε παραγωγικό περιβάλλον.

Η συγκεκριμένη αρχιτεκτονική εξασφαλίζει την ομαλή μετάβαση από το στάδιο του πειραματισμού στην επιχειρησιακή ωριμότητα, παρέχοντας ενοποιημένο σημείο ελέγχου για την ολότητα της ροής εργασίας MLOps.

4.4 Διαδικασία πρόβλεψης

Στο αρχείο *predict.py* υλοποιείται η διαδικασία πρόβλεψης, αναλαμβάνοντας τη δημιουργία εκτιμήσεων για τις επόμενες 16 ημέρες με τη χρήση μετεωρολογικών προγνώσεων ως εισόδου.

4.4.1 Μεθοδολογία αναδρομικής διαδικασίας για την πρόβλεψη

Λόγω της εγγενούς χρονικής αλληλουχίας που διέπει το πρόβλημα, όπου η εκτίμηση των ρύπων εξαρτάται άμεσα από χρονοκαθυστερημένα χαρακτηριστικά, κρίθηκε απαραίτητη η εφαρμογή μιας αναδρομικής στρατηγικής για την παραγωγή των προγνώσεων. Η αρχιτεκτονική του αρχείου *predict.py* διαρθρώνεται σε πέντε διακριτά λειτουργικά στάδια, τα οποία εξασφαλίζουν την αξιοπιστία και τη συνοχή των αποτελεσμάτων:

- **Προεπεξεργασία δεδομένων και εποχικότητα:** Η συνάρτηση *prepare_future_data* δεν αρκείται στην απλή ανάγνωση των μετεωρολογικών προγνώσεων, αλλά εφαρμόζει τεχνικές

στατιστικής συγχώνευσης για να επιτύχει δομική αντιστοιχία με τα δεδομένα εκπαίδευσης. Υπολογίζονται παράμετροι όπως η μέση τιμή και η τυπική απόκλιση, ενώ οι χρονικές μεταβλητές υποβάλλονται σε κυκλική κωδικοποίηση χρήση *month_sin*, *day_year_cos*, επιτρέποντας στο μοντέλο να αντιλαμβάνεται τη χρονική περιοδικότητα.

- **Μοντέλο (MLflow):** Η φόρτωση των αλγορίθμων πραγματοποιείται δυναμικά μέσω του *models:/[model_name]/latest*. Αυτή η πρακτική διασφαλίζει ότι το σύστημα χρησιμοποιεί πάντα την πιο πρόσφατη εγκεκριμένη έκδοση από το Model Registry, εξαλείφοντας την ανάγκη για χειροκίνητη παραμετροποίηση.

- **Αναδρομικός βρόχος πρόβλεψης:** Η διαδικασία εκτελείται επαναληπτικά, όπου η εκτίμηση της χρονικής στιγμής t ενσωματώνεται άμεσα ως χαρακτηριστικό εισόδου lag-1 feature για τον υπολογισμό της επόμενης χρονικής στιγμής $t + 1$.

- **Αποκανονικοποίηση:** Για ρύπους που έχουν υποστεί κλιμάκωση κατά την εκπαίδευση (π.χ. NO_2, HCHO), εφαρμόζεται αντίστροφος μετασχηματισμός. Εάν η απόλυτη τιμή της πρόβλεψης υπερβαίνει τη μονάδα, διαιρείται με τον συντελεστή κλιμάκωσης 10^6, επαναφέροντας τις τιμές στις φυσικές τους μονάδες μέτρησης.

- **Διασφάλιση ακεραιότητας δεδομένων:** Η εγγραφή στη βάση SQLite θωρακίζεται με αυστηρά πρωτόκολλα. Πριν την εντολή *to_sql*, εκτελείται διαδικασία διαγραφής *DELETE* για τις μελλοντικές εγγραφές του εκάστοτε ρύπου. Έτσι, αποτρέπονται παραβιάσεις μοναδικότητας και διασφαλίζεται ότι η βάση φιλοξενεί αποκλειστικά τις πλέον επικαιροποιημένες προγνώσεις.

διατήρηση παρωχημένων δεδομένων.

4.5 Υποδομή και παραμετροποίηση συστήματος Docker

Η τεχνολογία containerization αποτελεί τη βάση της αρχιτεκτονικής της υποδομής, εξασφαλίζοντας τη φορητότητα και την ομοιομορφία μεταξύ των περιβαλλόντων ανάπτυξης και παραγωγής.

```
/Bachelors
├── docker-compose.yml
├── backend/
│   ├── Dockerfile
│   └── requirements.txt
└── app_ui/
    ├── Dockerfile
    └── requirements.txt
```

Σχήμα 4.2: Ιεραρχική δομή αρχείων Docker και εξαρτήσεων του συστήματος.

4.5.1 Διαχείριση και οργάνωση αρχείων Dockerfiles

Για τη δόμηση των υπηρεσιών Backend και Frontend επιλέγεται η εικόνα *python:3.9-slim* βασισμένη σε Debian. Η συγκεκριμένη επιλογή παρέχει τη βέλτιστη ισορροπία ανάμεσα στο περιορισμένο μέγεθος ειδώλου και την πλήρη συμβατότητα με εξειδικευμένες γεωχωρικές και επιστημονικές βιβλιοθήκες που απαιτούνται.

Στο επίπεδο του Frontend, η αρχή DRY εφαρμόζεται μέσω του στρατηγικού ορισμού στη ρίζα του έργου. Επομένως, το *Dockerfile* του Frontend ενσωματώνει αυτούσια σενάρια κοινής λογικής του Backend όπως *config.py* και *database_utils.py*, εξαλείφοντας τον πλεονασμό κώδικα και διασφαλίζοντας ότι οι δύο υπηρεσίες χρησιμοποιούν πάντα ταυτόσημες ρυθμίσεις και μεθόδους πρόσβασης δεδομένων.

4.5.2 Docker Compose

Ο ρόλος του Docker Compose συνίσταται στην ενορχήστρωση και τη διαχείριση του κύκλου λειτουργίας των επιμέρους υπηρεσιών, ορίζοντας ταυτόχρονα τις διασυνδέσεις δικτύου και τις εξαρτήσεις που συνδέουν τα διαφορετικά στοιχεία μεταξύ τους. Μέσω της τεχνικής των κοινόχρηστων τόμων Shared Volumes, επιτυγχάνεται η διαρκής αποθήκευση πληροφορίας, καθώς συγκεκριμένοι φάκελοι του κεντρικού συστήματος αντιστοιχίζονται σε καταλόγους εντός των containers. Η προσέγγιση αυτή αποδεικνύεται ιδιαίτερα σημαντική για τη διατήρηση τόσο της βάσης SQLite air_quality.db όσο και των αρχείων MLflow mlflow.db, εξασφαλίζοντας τη δυνατότητα άμεσης πρόσβασης σε προβλέψεις και μεταδεδομένα που παράγονται συνεχώς από το Backend κατά την αναδρομική επεξεργασία. Παράλληλα, η προστασία ευαίσθητων πληροφοριών και διαπιστευτηρίων πραγματοποιείται με τη χρήση αρχείου περιβάλλοντος .env, το οποίο αποσυνδέει τις παραμέτρους ρύθμισης από τον εκτελέσιμο κώδικα. Η χρήση του Docker Compose επιτρέπει την ολιστική ενσωμάτωση όλων των επιμέρους συστημάτων που παρουσιάστηκαν, συνενώνοντας τον αυτοματισμό της αλυσίδας ETL, τη φάση εκπαίδευσης του μοντέλου και τη διαδικασία παραγωγής προβλέψεων σε μία ενοποιημένη αρχιτεκτονική. Η εκκίνηση της υποδομής μέσω του αρχείου διαμόρφωσης καθιστά το Backend τυποποιημένο και έτοιμο προς παραγωγική χρήση και δίνει την δυνατότητα υποστήριξης της επιχειρησιακής εκτέλεσης του προγνωστικού αλγορίθμου. Έχοντας ολοκληρώσει την τεχνική ανάπτυξη του Backend, η μελέτη προχωρά στην απεικόνιση των αποτελεσμάτων μέσω της διαδραστικής διεπαφής Streamlit και στην εκτενή αξιολόγηση της προγνωστικής απόδοσης του συνολικού συστήματος. Το κεφάλαιο που ακολουθεί περιλαμβάνει την παρουσίαση των πειραματικών ευρημάτων για το αστικό περιβάλλον της Θεσσαλονίκης, συμπληρωμένα από την ποσοτική αξιολόγηση των μετρικών σφάλματος και την επίδειξη της λειτουργίας της οριστικής εφαρμογής.

Κεφάλαιο 5

Παρουσίαση εφαρμογής & αξιολόγηση

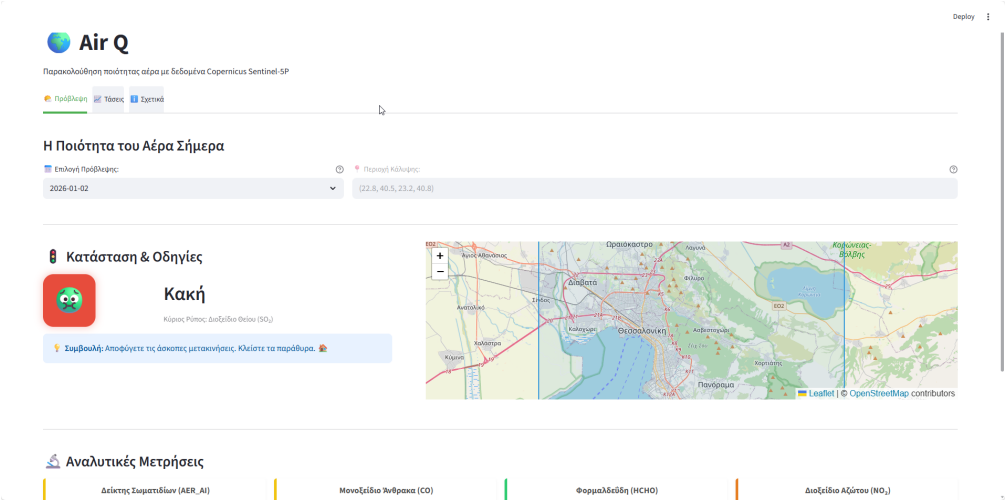
5.1 Εισαγωγή

Το παρόν κεφάλαιο επικεντρώνεται στην επαλήθευση και την αξιολόγηση της προτεινόμενης λύσης, αφού προηγουμένως, στο Κεφάλαιο 4, αναλύθηκε εκτενώς η αρχιτεκτονική και η τεχνική υλοποίηση του συστήματος. Στόχο αποτελεί η απόδειξη της επιχειρησιακής ωριμότητας της πλατφόρμας, καθώς και η διαπίστωση της ικανότητάς της να παράγει ακριβείς προβλέψεις υπό πραγματικές λειτουργικές συνθήκες. Η διαδικασία αξιολόγησης πραγματοποιείται σε πολλαπλά στρώματα, περιλαμβάνοντας τόσο την τεχνική αρτιότητα των αλγορίθμων μηχανικής μάθησης όσο και τη χρηστικότητα που προσφέρει η εφαρμογή στους τελικούς χρήστες.

Η οργάνωση του κεφαλαίου ακολουθεί κλιμακωτή δομή. Στην Ενότητα 5.2, περιγράφεται αρχικά το περιβάλλον διεπαφής χρήστη *User Interface*, το οποίο αναπτύχθηκε με τη χρήση του πλαισίου Streamlit, αναδεικνύοντας τα διαδραστικά χαρακτηριστικά και τις μεθόδους οπτικοποίησης που καθιστούν προσιτά τα δεδομένα. Ακολούθως, η Ενότητα 5.3 καταγράφει τα πειραματικά ευρήματα από τη διαδικασία εκπαίδευσης, αξιοποιώντας ποσοτικά μέτρα όπως το MAE και το MSE για την εκτίμηση της ακρίβειας των μοντέλων. Τελικά, η Ενότητα 5.4 περιλαμβάνει λεπτομερή μελέτη περίπτωσης με επίκεντρο την πόλη της Θεσσαλονίκης, εξετάζοντας πώς το σύστημα ανταποκρίνεται δυναμικά σε συγκεκριμένα ατμοσφαιρικά επεισόδια και επιβεβαιώνοντας την ικανότητά του να ερμηνεύει με ακρίβεια φαινόμενα συσσώρευσης και έκπλυσης ρύπων.

5.2 User Interface Streamlit

Η καρτέλα «Πρόβλεψη» συγκεντρώνει την κεντρική λειτουργικότητα της εφαρμογής, παρέχοντας στους χρήστες μια ολοκληρωμένη εικόνα των αναμενόμενων επιπέδων ατμοσφαιρικής ρύπανσης. Η δημιουργία της διεπαφής βασίστηκε στην αρχή της χρηστικότητας, εξασφαλίζοντας άμεση πρόσβαση σε κρίσιμες πληροφορίες μέσω διαδραστικών και οπτικών μέσων.



Σχήμα 5.1: Κεντρική σελίδα UI (Frontend).

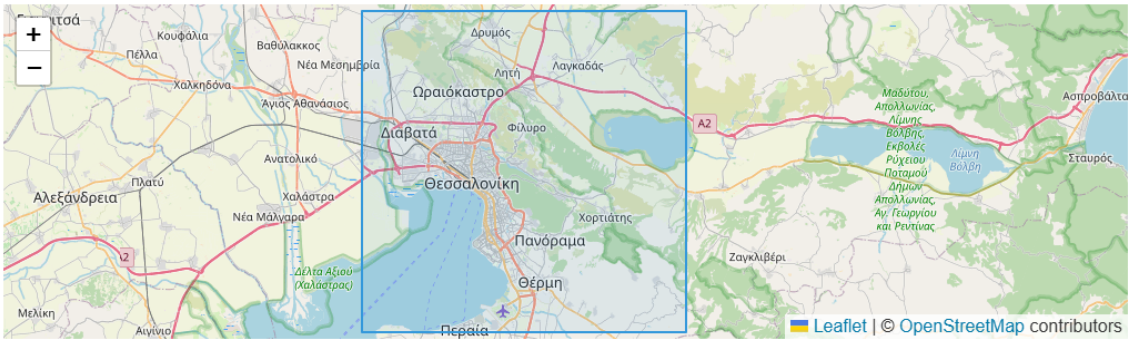
Διάταξη και αρχικής σελίδας

Στην κορυφή της σελίδας υπάρχουν ειδικά *widgets* από τη βιβλιοθήκη Streamlit, τα οποία επιτρέπουν στους χρήστες να παραμετροποιήσουν την προβολή. Μέσω της συνάρτησης *st.selectbox*, ο χρήστης μπορεί να επιλέξει την επιθυμητή ημερομηνία από το διαθέσιμο χρονικό ορίζοντα προβλέψεων, ενώ ταυτόχρονα παρουσιάζονται πληροφορίες για το γεωγραφικό περίγραμμα *Bounding Box* της περιοχής της Θεσσαλονίκης. Το backend σύστημα τροφοδοτείται δυναμικά από αυτήν την επιλογή, ανακτώντας τα αντίστοιχα δεδομένα από τη βάση για την ενημέρωση των οπτικοποιήσεων.

Διαδραστική χαρτογράφηση

Με τη χρήση της βιβλιοθήκης Folium και την ενσωμάτωσή της μέσω του *st_folium*, υλοποιήθηκε ο διαδραστικός χάρτης Εικόνα 5.2, ο οποίος αποτελεί κεντρικό στοιχείο της οπτικοποίησης. Οι συντεταγμένες της Θεσσαλονίκης, όπως ορίζονται στο αρχείο *config.py*, καθορίζουν το αυτόματο επίκεντρο του χάρτη.

Ένα γεωμετρικό παραλληλόγραμμο *folium.Rectangle* χρησιμοποιείται για την οριοθέτηση της περιοχής μελέτης, παρέχοντας το απαραίτητο χωρικό πλαίσιο. Η δομή *STATUS_CONFIG* παρέχει τις κατηγοριοποιήσεις στις οποίες βασίζεται η χρωματική κωδικοποίηση για την απόδοση της ποιότητας του αέρα. Χρώματα όπως το πράσινο για «Εξαιρετική» κατάσταση και το κόκκινο για «Κακή» εξασφαλίζουν άμεση οπτική πληροφόρηση στον χρήστη σχετικά με τη σοβαρότητα της ρύπανσης.

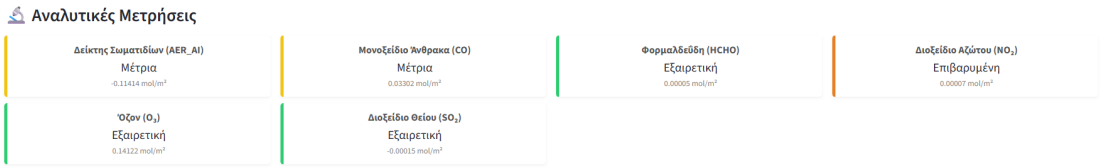


Σχήμα 5.2: Διαδραστικός χάρτης πρόβλεψης στο Bounding Box

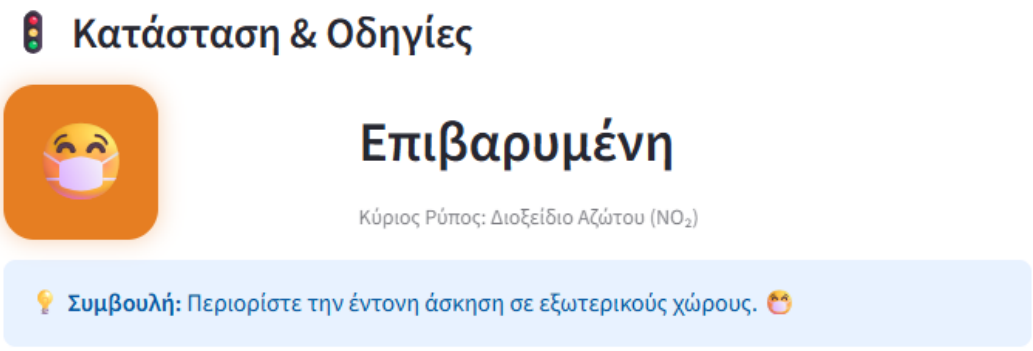
Κάρτες κατάστασης και ειδικές συμβουλές

Συνοπτικές κάρτες πληροφοριών χρησιμοποιούνται στο τμήμα των αποτελεσμάτων Εικόνα 5.3 και Εικόνα 5.4 για την παρουσίαση της ποιοτικής αξιολόγησης της ημέρας από το σύστημα. Κάθε κάρτα περιλαμβάνει την αριθμητική τιμή της πρόβλεψης, τον χαρακτηρισμό της κατάστασης (π.χ. «Μέτρια», «Εξαιρετική») και το αντίστοιχο εικονίδιο.

Ιδιαίτερη σημασία δίνεται στην παροχή πρακτικών οδηγιών, όπως προτροπές για περιορισμό της σωματικής άσκησης σε εξωτερικούς χώρους, οι οποίες παράγονται αυτόματα από τη λογική του συστήματος βάσει των προκαθορισμένων ορίων ασφαλείας. Με αυτόν τον τρόπο, η εφαρμογή μετατρέπει τα ακατέργαστα δεδομένα σε χρήσιμη πληροφορία για την προστασία της δημόσιας υγείας.



Σχήμα 5.3: Κάρτες πληροφοριών με τις τιμές τους σε mol/m^2



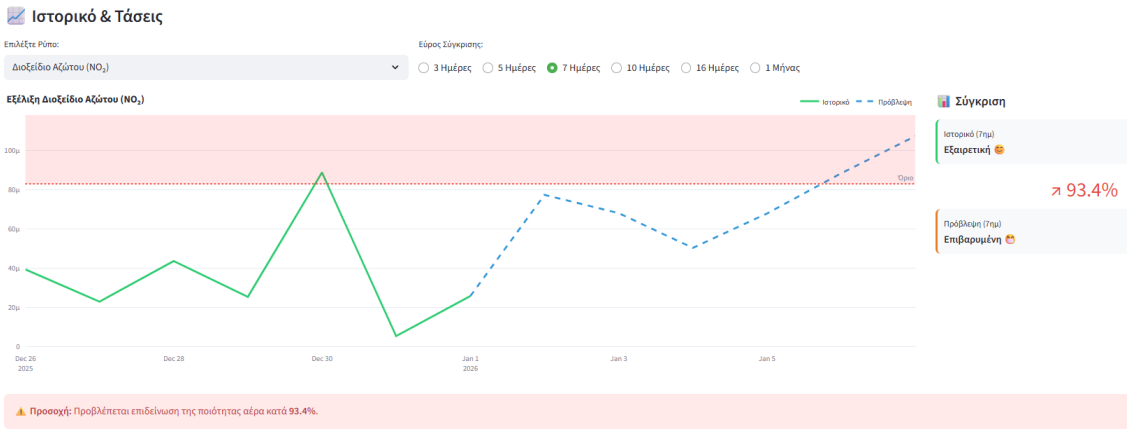
Σχήμα 5.4: Συμβουλή και υπόδειξη ρύπου με την μεγαλύτερη βαρύτητα

5.3 Καρτέλα ανάλυσης τάσεων και στατιστικών

Το υποσύστημα χρονικής ανάλυσης της εφαρμογής υλοποιείται μέσω της καρτέλας «Τάσεις» (Εικόνα 5.5), παρέχοντας στον χρήστη τη δυνατότητα συγκριτικής αξιολόγησης μεταξύ ιστορικών μετρήσεων και μελλοντικών εκτιμήσεων. Διαδραστικοί έλεγχοι διατίθενται στη διεπαφή μέσω των στοιχείων *st.selectbox* και *st.radio*, επιτρέποντας την επιλογή του επιθυμητού ρύπου και του χρονικού παραθύρου σύγκρισης, το οποίο μπορεί να κυμαίνεται από τρεις ημέρες έως έναν μήνα, αλλά με την ιδιαιτερότητα ότι στις μελλοντικές προβλέψεις η ανώτερη πρόβλεψη ανέρχεται έως 16 ημέρες.

Δυναμικά γραφήματα χρονοσειρών παράγονται με τη βιβλιοθήκη Plotly, η οποία χρησιμοποιείται για την οπτικοποίηση των δεδομένων ανά ρύπο ξεχωριστά. Στο γράφημα υφίσταται σαφής οπτικός διαχωρισμός: συνεχής πράσινη γραμμή αποδίδει την ιστορική πορεία, ενώ διακεκομμένη μπλε γραμμή υποδηλώνει την πρόβλεψη, τονίζοντας τη στατιστική φύση της εκτίμησης. Επιπροσθέτως, ενσωματώνεται αυτοματοποιημένος μηχανισμός ασφαλείας που σχεδιάζει κόκκινη οριζόντια γραμμή ορίου (*add_hline*) και σκιασμένη περιοχή κινδύνου (*add_hrect*) όταν προβλέπεται ότι τα επίπεδα ρύπανσης θα ξεπεράσουν τα όρια υγείας.

Στο δεξιό τμήμα της καρτέλας πραγματοποιείται αλγοριθμικός υπολογισμός της ποσοστιαίας διαφοράς μεταξύ του ιστορικού και του προβλεπόμενου μέσου όρου. Δυναμικοί δείκτες κατεύθυνσης (βέλη ↗/↘) και χρωματική κωδικοποίηση (κόκκινο για επιδείνωση, πράσινο για βελτίωση) χρησιμοποιούνται από το σύστημα για να μετατρέψει τα στατιστικά ευρήματα σε άμεσα κατανοητή πληροφορία. Επομένως, αποκτάται από τον χρήστη μια σαφής εικόνα της αναμενόμενης τάσης της ποιότητας του αέρα στην υπό εξέταση περιοχή.



Σχήμα 5.5: Σύγκριση ιστορικών δεδομένων και δεδομένων πρόβλεψης

5.4 Ειδική περίπτωση στην Θεσσαλονίκη

Η δυναμική των ρύπων επηρεάζεται καθοριστικά από το ιδιαίτερο μικροκλίμα και τη γεωμορφολογία της Θεσσαλονίκης, η οποία περικλείεται από τον Θερμαϊκό Κόλπο και την περιβάλλουσα λοφοσειρά. Φαινόμενα θερμοκρασιακής αναστροφής εκδηλώνονται συχνά στην πόλη κατά τη χειμερινή περίοδο, με αποτέλεσμα τον εγκλωβισμό των ρύπων στο οριακό στρώμα. Αντιθέτως, η διασπορά ή η έκπλυση συμβάλλουν στη δραστική μείωση των συγκεντρώσεων όταν εκδηλώνονται έντονα μετεωρολογικά φαινόμενα.

Σενάριο Α: Επεισόδιο συσσώρευσης ρύπων στις 14 Φεβρουαρίου 2025

Ψυχρές αέριες μάζες επηρέαζαν την πόλη κατά τη διάρκεια της 14ης Φεβρουαρίου 2025, με την ελάχιστη θερμοκρασία να φτάνει τους 5,3°C και τη βροχόπτωση να παραμένει σχεδόν μηδενική (0,2 mm). Η συσσώρευση NO_2 ευνοήθηκε από αυτές τις συνθήκες σε συνδυασμό με την αυξημένη χρήση συστημάτων θέρμανσης. Από τον δορυφόρο Sentinel-5P καταγράφηκε πραγματική τιμή $0,000138\,\mathrm{mol/m^2}$, επίπεδο που χαρακτηρίζεται ιδιαίτερα υψηλό για τα δεδομένα της περιοχής. Πρόβλεψη ύψους $0,000132\,\mathrm{mol/m^2}$ παρήχθη από το προτεινόμενο μοντέλο XGBoost, επιτυγχάνοντας εξαιρετική σύγκλιση με την πραγματικότητα.

Το αποτέλεσμα της απόφασης του μοντέλου όπως αναλύθηκε, αποκαλύπτει ότι τα χαρακτηριστικά της χαμηλής θερμοκρασίας και της ατμοσφαιρικής σταθερότητας αξιοποιήθηκαν ορθά από το σύστημα. Η ανοδική τάση και ο κίνδυνος υποβάθμισης της ποιότητας του αέρα αναγνωρίστηκαν με ακρίβεια από το μοντέλο, παρά την ελαφρά υποεκτίμηση που είναι αναμενόμενη σε ακραία γεγονότα αιχμής.

Σενάριο Β: Γεγονός έκπλυσης στις (31 Ιουλίου 2025)

Μια τυπική καλοκαιρινή καταιγίδα στις 31 Ιουλίου 2025 αποτελεί το δεύτερο σενάριο, η οποία συνοδεύτηκε από έντονη βροχόπτωση ύψους 47,8 mm. Ως φυσικός «καθαριστής» *scrubber* λειτουργεί η βροχή σε αυτές τις περιπτώσεις, απομακρύνοντας τα αιωρούμενα σωματίδια και τους αέριους ρύπους μέσω της διαδικασίας της υγρής εναπόθεσης. Σε εξαιρετικά χαμηλά επίπεδα υποχώρησε η πραγματική τιμή του ρύπου ($1,18 \times 10^{-5}\,\mathrm{mol/m^2}$), τάξη μεγέθους σημαντικά μικρότερη σε σύγκριση με τις χειμερινές μετρήσεις. Τιμή $1,50 \times 10^{-5}\,\mathrm{mol/m^2}$ προβλέφθηκε από το μοντέλο, καταγράφοντας με επιτυχία τη δραστική πτώση.

Ως το κυρίαρχο χαρακτηριστικό στη λογική του μοντέλου αναδείχθηκε η βροχόπτωση, αποδεικνύοντας ότι η φυσική σχέση μεταξύ έντονου υετού και ατμοσφαιρικής κάθαρσης έχει ενσωματωθεί σωστά στον αλγόριθμο.

Πίνακας 5.1: Συγκριτικός Πίνακας Σεναρίων Αξιολόγησης (NO_2)

Παράμετρος	**Σενάριο Α (Χειμώνας)**	**Σενάριο Β (Καλοκαίρι)**
Ημερομηνία	14/02/2025	31/07/2025
Θερμοκρασία (Min)	5,3°C	22,1°C
Βροχόπτωση	0,2 mm	47,8 mm
Πραγματική Τιμή (mol/m^2)	**0,000138**	**0,000012**
Πρόβλεψη Μοντέλου (mol/m^2)	**0,000132**	**0,000015**
Συμπέρασμα	Επιτυχής ανίχνευση συσσώρευσης	Επιτυχής ανίχνευση έκπλυσης

Η επιτυχής αντιμετώπιση αυτών των δύο πολικά διαφορετικών περιπτώσεων —αφενός της συγκέντρωσης ρύπων που διερευνήθηκε στο πρώτο σενάριο και αφετέρου της ατμοσφαιρικής διασποράς που εξετάστηκε στο δεύτερο— καταδεικνύει την αξιοπιστία και την προσαρμοστικότητα της προτεινόμενης μεθοδολογίας. Το γεγονός ότι ο αλγόριθμος XGBoost αναγνωρίζει και ποσοτικοποιεί αποτελεσματικά τη μετάβαση από φάσεις υψηλής ατμοσφαιρικής ρύπανσης σε περιόδους βελτίωσης της ποιότητας του αέρα στη Θεσσαλονίκη, αξιοποιώντας μετεωρολογικά δεδομένα, αναβαθμίζει την εφαρμογή από απλή θεωρητική κατασκευή σε λειτουργικό μέσο περιβαλλοντικής επιτήρησης. Επιπλέον, η συνέπεια μεταξύ των υπολογισμένων τιμών και των πραγματικών δορυφορικών παρατηρήσεων, ακόμη και σε περιπτώσεις απότομων ατμοσφαιρικών διακυμάνσεων, υπογραμμίζει την επιχειρησιακή εγκυρότητα της ανάπτυξης και προετοιμάζει το έδαφος για την ολοκληρωμένη αποτίμηση της ερευνητικής προσπάθειας. Με βάση τα στοιχεία αυτά, η μελέτη μεταβαίνει στο τελικό της στάδιο, όπου συντίθενται τα κύρια ερευνητικά πορίσματα, αναλύονται οι αδυναμίες και οι περιορισμοί που εντοπίστηκαν κατά τη διαδικασία, και προτείνονται στοχευμένες προοπτικές για την περαιτέρω βελτίωση και επέκταση του προγνωστικού συστήματος.

Κεφάλαιο 6

Συμπεράσματα και μελλοντικές επεκτάσεις

6.1 Σύνοψη εργασίας

Ο σχεδιασμός και η υλοποίηση ενός ολοκληρωμένου, αυτοματοποιημένου συστήματος πρόβλεψης ατμοσφαιρικής ρύπανσης για το πολεοδομικό συγκρότημα της Θεσσαλονίκης έως 16 ημέρες, αποτέλεσε το επίκεντρο της παρούσας πτυχιακής εργασίας. Από τη λεπτομερή ανάλυση του προβλήματος της αστικής ρύπανσης και την ανάγκη για αξιόπιστα δεδομένα υψηλής χωρικής κάλυψης ξεκίνησε η πορεία της έρευνας.

Δεδομένα τηλεπισκόπησης από τον δορυφόρο Sentinel-5P του προγράμματος *Copernicus* αξιοποιήθηκαν προς αυτή την κατεύθυνση και συνδυάστηκαν με μετεωρολογικές παραμέτρους από το *Open-Meteo*. Οι αρχές των *cloud-native* εφαρμογών αποτέλεσαν τη βάση της αρχιτεκτονικής του συστήματος, η οποία ενσωμάτωσε τον κατάλληλο αγωγό ETL για την άντληση δεδομένων, μια προηγμένη διαδικασία εκπαίδευσης μηχανικής μάθησης με τον αλγόριθμο XGBoost και τη χρήση εργαλείων MLOps Optuna, MLflow για τη βελτιστοποίηση και διακυβέρνηση των μοντέλων. Η σταθερότητα και η κλιμακωσιμότητα της λύσης εξασφαλίστηκαν από την τελική υλοποίηση σε περιβάλλον Docker.

6.2 Συμπεράσματα

Ορισμένα θεμελιώδη συμπεράσματα προκύπτουν από την ολοκλήρωση και την αξιολόγηση του συστήματος:

- **Πλήρης αυτοματοποίηση:** Η επίτευξη πλήρους αυτονομίας αποτελεί το σημαντικότερο επίτευγμα της εργασίας. Σε καθημερινή βάση λειτουργεί το σύστημα χωρίς την ανάγκη ανθρώπινης παρέμβασης, εκτελώντας διαδοχικά τη συλλογή, την επεξεργασία, την εκπαίδευση και την παραγωγή προγνώσεων. Υψηλή είναι η επιχειρησιακή ετοιμότητα του συστήματος, καθώς μόνο το αρχείο *backfill_data.py* παραμένει χειροκίνητο τμήμα, το οποίο απαιτείται

αποκλειστικά για την αρχικοποίηση των ιστορικών δεδομένων ή την άντληση παλιότερων ιστορικών δεδομένων ατμοσφαιρικής ρύπανσης από τα τρέχων για καλύτερη πρόβλεψη του μοντέλου.

- **Αλγοριθμική αποτελεσματικότητα:** Ιδιαίτερα αποδοτικός για τον χειρισμό των δομημένων δεδομένων χρονοσειρών αποδείχθηκε ο αλγόριθμος XGBoost. Ταχύτητα στην εκπαίδευση και υψηλή ερμηνευσιμότητα προσέφερε ο XGBoost σε σύγκριση με πιο σύνθετα μοντέλα βαθιάς μάθησης, καταγράφοντας με ακρίβεια τις μη γραμμικές σχέσεις μεταξύ μετεωρολογικών συνθηκών και συγκέντρωσης ρύπων.
- **Φορητότητα:** Η πλήρη αναπαραγωγιμότητα του συστήματος διασφαλίστηκε από τη χρήση της τεχνολογίας Docker. Η μεταφορά της εφαρμογής από το περιβάλλον ανάπτυξης στην παραγωγή χωρίς δυσλειτουργίες επετράπη από την απομόνωση των εξαρτήσεων, επικυρώνοντας τη σημασία των σύγχρονων πρακτικών DevOps στην επιστήμη δεδομένων.

6.3 Περιορισμοί και δυσκολίες

Συγκεκριμένες προκλήσεις και περιορισμοί αντιμετωπίστηκαν από την έρευνα, παρά την επιτυχή υλοποίηση:

- **Χρονοβόρα άντληση δεδομένων:** Η υπολογιστική υστέρηση κατά τη συλλογή ιστορικών δεδομένων αποτελεί τον σημαντικότερο περιορισμό. Εξαιρετικά χρονοβόρα αποδείχθηκε η διαδικασία ανάκτησης μέσω του *backfill_data.py*, καθώς περίπου 5 ώρες επεξεργασίας απαιτούσε η λήψη δεδομένων για ένα διάστημα μόλις 10 ημερών. Στον τεράστιο όγκο των ακατέργαστων αρχείων *.zip* του *Copernicus Open Access Hub* και στο σημαντικό δικτυακό και υπολογιστικό κόστος οφείλεται αυτό, το οποίο απαιτείται για την αποσυμπίεση και την εκχύλιση συγκεκριμένων pixels για την περιοχή της Θεσσαλονίκης αλλά και γενικότερα. Μια πρόκληση που απαιτεί αυξημένους πόρους καθίσταται η εκπαίδευση σε δεδομένα πολλών μηνών λόγω του περιορισμού αυτού.
- **Δορυφορικοί Περιορισμοί:** Περιορισμοί που σχετίζονται με τη νεφοκάλυψη συνεπάγεται η εξάρτηση από τον Sentinel-5P, καθώς σε απώλεια δεδομένων οδηγούν οι ημέρες με υψηλή συννεφιά. Επιπροσθέτως, για την καταγραφή των επιπέδων υποβάθρου είναι επαρκής η χωρική ανάλυση του δορυφόρου ($3.5 \times 5.5\,\mathrm{km}$), αλλά ο εντοπισμός τοπικών εστιών ρύπανσης σε επίπεδο δρόμου δεν επιτρέπεται.

6.4 Μελλοντικές επεκτάσεις

Οι βάσεις για μια σειρά από μελλοντικές βελτιώσεις και επεκτάσεις τίθενται από την παρούσα εργασία:

- **Ενοποίηση με την εφαρμογή «AirQ»:** Η ενσωμάτωση του backend πρόβλεψης με την υπάρχουσα εφαρμογή «AirQ» αποτελεί φυσική εξέλιξη του συστήματος, καθώς θα αντικαταστήσει το παρών μοντέλο πρόβλεψης. Η διάχυση της πληροφορίας σε ένα ευρύτερο κοινό θα επιτραπεί από αυτήν τη σύνδεση.

- **Προσωποποιημένες ειδοποιήσεις:** Προσωποποιημένες ειδοποιήσεις υγείας θα μπορεί να παρέχει το σύστημα μέσω της παραπάνω ενοποίησης. Η δυνατότητα να δημιουργούν προφίλ και να λαμβάνουν στοχευμένες προειδοποιήσεις βάσει της ευαισθησίας τους (π.χ.αθλητές, ηλικιωμένοι, καρδιακοί), αντί για γενικές συστάσεις.

- **Γεωγραφική κλιμάκωση:** Άμεσα υλοποιήσιμη με την απλή προσθήκη των γεωγραφικών συντεταγμένων είναι η επέκταση του συστήματος σε άλλες μεγάλες Ελληνικές πόλεις, όπως η Αθήνα, ο Βόλος και η Πάτρα, λόγω της του τρόπου που είναι φτιαγμένος ο κώδικας της εφαμοργής.

- **Υβριδικά μοντέλα δεδομένων:** Ο συνδυασμός των δορυφορικών δεδομένων με επίγειους αισθητήρες IoT. Τους περιορισμούς της δορυφορικής ανάλυσης θα μπορούσε να αντισταθμίσει μια τέτοια υβριδική προσέγγιση, προσφέροντας υψηλότερη χρονική και χωρική ακρίβεια στις προγνώσεις.

Συνοψίζοντας, αυτή η αυτοματοποιημένη εφαρμογή κατέδειξε τον καίριο ρόλο που διαδραματίζει η διασύνδεση τηλεπισκοπικών δορυφορικών τεχνολογιών με αλγορίθμους μηχανικής μάθησης στην αντιμετώπιση επιτακτικών περιβαλλοντικών ζητημάτων. Η ανάπτυξη ενός πλήρως αυτοματοποιημένου και εύχρηστου προγνωστικού μηχανισμού αποτυπώνει σημαντική επιστημονική προσφορά, έστω και υπό το βάρος δομικών δυσχερειών όπως οι χρονικοί περιορισμοί στην επεξεργασία δεδομένων και τα εμπόδια που προκύπτουν από τη νεφική κάλυψη. Συνεπώς, το παρόν έργο υπερβαίνει τα όρια μιας αμιγώς τεχνολογικής εφαρμογής, καθώς αποτελεί ουσιώδη συμβολή στην προστασία της κοινής υγείας και την ενίσχυση βιώσιμων αναπτυξιακών πρακτικών στο αστικό περιβάλλον. Τέλος, η προοπτική ενσωμάτωσης του συστήματος σε διευρυμένες χωρικές κλίμακες και πολυεπίπεδες εφαρμογές προοιωνίζει τη διαμόρφωση ενός ψηφιακού περιβαλλοντικού οικοσυστήματος, το οποίο δύναται να βελτιώσει ουσιαστικά τις συνθήκες διαβίωσης στους σύγχρονους αστικούς ιστούς.

Βιβλιογραφία

[1] Y. Zhang και P. J. Thorburn, «Handling missing data in near real-time environmental monitoring: A system and a review of selected methods», *Future Generation Computer Systems*, τόμ. 128, σσ. 63–72, 2022. DOI: 10.1016/j.future.2021.09.033

[2] Y. Himeur, B. Rimal, A. Tiwary και A. Amira, «Using artificial intelligence and data fusion for environmental monitoring: A review and future perspectives», *Information Fusion*, τόμ. 86-87, σσ. 44–75, 2022. DOI: 10.1016/j.inffus.2022.06.003

[3] Z. B. Bouallègue κ.ά., «The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context», *Bulletin of the American Meteorological Society*, τόμ. 105, αρθμ. 6, E864–E883, 2024. DOI: 10.1175/BAMS-D-23-0162.1

[4] M. Méndez, M. G. Merayo και M. Núñez, «Machine learning algorithms to forecast air quality: a survey», *Artificial Intelligence Review*, τόμ. 56, αρθμ. 9, σσ. 10 031–10 066, 2023. DOI: 10.1007/s10462-023-10424-4

[5] B. Bach κ.ά., «Dashboard Design Patterns», *IEEE Transactions on Visualization and Computer Graphics*, τόμ. 29, αρθμ. 1, σσ. 342–352, 2023. DOI: 10.1109/TVCG.2022.3209448

[6] Y. Hajjaji, W. Boulila, I. R. Farah, I. Romdhani και A. Hussain, «Big data and IoT-based applications in smart environments: A systematic review», *Computer Science Review*, τόμ. 39, σ. 100 318, 2021. DOI: 10.1016/j.cosrev.2020.100318

[7] M. G. Schultz κ.ά., «Can deep learning beat numerical weather prediction?», *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, τόμ. 379, αρθμ. 2194, σ. 20 200 097, 2021. DOI: 10.1098/rsta.2020.0097

[8] Q. Liao, M. Zhu, L. Wu, X. Pan, X. Tang και Z. Wang, «Deep Learning for Air Quality Forecasts: a Review», *Current Pollution Reports*, τόμ. 6, αρθμ. 4, σσ. 399–409, 2020. DOI: 10.1007/s40726-020-00159-z

[9] W. Mao, W. Wang, L. Jiao, S. Zhao και A. Liu, «Modeling air quality prediction using a deep learning approach: Method optimization and evaluation», *Sustainable Cities and Society*, τόμ. 65, σ. 102 567, 2021. DOI: 10.1016/j.scs.2020.102567

[10] F. Gholami, M. Tomas, Z. Gholami και M. Vakili, «Technologies for the nitrogen oxides reduction from flue gas: A review», *Science of The Total Environment*, τόμ. 714, σ. 136 712, 2020. DOI: 10.1016/j.scitotenv.2020.136712

[11] P. Han κ.ά., «Calibrations of Low-Cost Air Pollution Monitoring Sensors for CO, NO2, O3, and SO2», *Sensors*, τόμ. 21, αρθμ. 1, σ. 256, 2021. DOI: 10.3390/s21010256

[12] S. Dimitroulopoulou κ.ά., «Indoor air quality guidelines from across the world: An appraisal considering energy saving, health, productivity, and comfort», *Environment International*, τόμ. 178, σ. 108 127, 2023. DOI: 10.1016/j.envint.2023.108127

[13] Copernicus Sentinel Online. «Sentinel-5P Mission Overview - SentiWiki», επίσκεψη 28 Δεκ. 2025. διεύθν.: https://sentiwiki.copernicus.eu/web/s5p-mission

[14] Royal Netherlands Meteorological Institute (KNMI). «TROPOMI: Tropospheric Monitoring Instrument», επίσκεψη 28 Δεκ. 2025. διεύθν.: https://www.tropomi.eu

[15] K. Benidis κ.ά., «Deep Learning for Time Series Forecasting: Tutorial and Literature Survey», *ACM Comput. Surv.*, τόμ. 55, αρθμ. 6, 121:1–121:36, 2022. DOI: 10.1145/3533382

[16] T. S. Talagala, R. J. Hyndman και G. Athanasopoulos, «Meta-learning how to forecast time series», *Journal of Forecasting*, τόμ. 42, αρθμ. 6, σσ. 1476–1501, 2023. DOI: 10.1002/for.2963

[17] S.-X. Lv, L. Peng, H. Hu και L. Wang, «Effective machine learning model combination based on selective ensemble strategy for time series forecasting», *Information Sciences*, τόμ. 612, σσ. 994–1023, 2022. DOI: 10.1016/j.ins.2022.09.002

[18] Z. Shen, Y. Zhang, J. Lu, J. Xu και G. Xiao, «A novel time series forecasting model with deep learning», *Neurocomputing*, τόμ. 396, σσ. 302–313, 2020. DOI: 10.1016/j.neucom.2018.12.084

[19] P. Mancuso, V. Piccialli και A. M. Sudoso, «A machine learning approach for forecasting hierarchical time series», *Expert Systems with Applications*, τόμ. 182, σ. 115 102, 2021. DOI: 10.1016/j.eswa.2021.115102

[20] J. Dong, Y. Chen, B. Yao, X. Zhang και N. Zeng, «A neural network boosting regression model based on XGBoost», *Applied Soft Computing*, τόμ. 125, σ. 109 067, 2022. DOI: 10.1016/j.asoc.2022.109067

[21] T. Chen και C. Guestrin, «XGBoost: A Scalable Tree Boosting System», στο *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, σσ. 785–794. DOI: 10.1145/2939672.2939785

[22] O. Sagi και L. Rokach, «Approximating XGBoost with an interpretable decision tree», *Information Sciences*, τόμ. 572, σσ. 522–542, 2021. DOI: 10.1016/j.ins.2021.05.055

[23] R. Shwartz-Ziv και A. Armon, «Tabular data: Deep learning is not all you need», *Information Fusion*, τόμ. 81, σσ. 84–90, 2022. DOI: 10.1016/j.inffus.2021.11.011

[24] V. Borisov, T. Leemann κ.ά., «Deep Neural Networks and Tabular Data: A Survey», *IEEE Transactions on Neural Networks and Learning Systems*, τόμ. 35, αρθμ. 6, σσ. 7499–7519, 2024. DOI: 10.1109/TNNLS.2022.3229161

[25] B. Bischl κ.ά., «Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges», *WIREs Data Mining and Knowledge Discovery*, τόμ. 13, e1484, 2023, ISSN: 1942-4795.

[26] L. Yang και A. Shami, «On hyperparameter optimization of machine learning algorithms: Theory and practice», *Neurocomputing*, τόμ. 415, σσ. 295–316, 2020, ISSN: 0925-2312.

[27] J. Wu, S. Chen και X. Liu, «Efficient hyperparameter optimization through model-based reinforcement learning», *Neurocomputing*, τόμ. 409, σσ. 381–393, 2020, ISSN: 0925-2312.

[28] Z. K. Maseer κ.ά., «Benchmarking of Machine Learning for Anomaly Based Intrusion Detection Systems in the CICIDS2017 Dataset», *IEEE Access*, τόμ. 9, σσ. 22 351–22 370, 2021, ISSN: 2169-3536.

[29] N. Ahmad, Y. Ghadi, M. Adnan και M. Ali, «Load Forecasting Techniques for Power System: Research Challenges and Survey», *IEEE Access*, τόμ. 10, σσ. 71 054–71 090, 2022. DOI: 10.1109/ACCESS.2022.3187839

[30] J. Qi, J. Du, S. M. Siniscalchi, X. Ma και C.-H. Lee, «On Mean Absolute Error for Deep Neural Network Based Vector-to-Vector Regression», *IEEE Signal Processing Letters*, τόμ. 27, σσ. 1485–1489, 2020. DOI: 10.1109/LSP.2020.3016837

[31] S. Demir και E. K. Sahin, «An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost», *Neural Computing and Applications*, τόμ. 35, αρθμ. 4, σσ. 3173–3190, 2023. DOI: 10.1007/s00521-022-07856-4

[32] A. M. Potdar, N. D g, S. Kengond και M. M. Mulla, «Performance Evaluation of Docker Container and Virtual Machine», *Procedia Computer Science*, τόμ. 171, σσ. 1419–1428, 2020. DOI: 10.1016/j.procs.2020.04.152

[33] A. K. Verma, R. Patel, P. Rai και S. Patel, «Performance Comparison Between Docker Container and Virtual Machine Microservices», στο *2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*, 2024, σσ. 718–723. DOI: 10.1109/CSNT60213.2024.10546015

[34] C. Boettiger, «An introduction to Docker for reproducible research», *SIGOPS Oper. Syst. Rev.*, τόμ. 49, αρθμ. 1, σσ. 71–79, 2015. DOI: 10.1145/2723872.2723882

[35] R. S. Canon, «The Role of Containers in Reproducibility», στο *2020 2nd International Workshop on Containers and New Orchestration Paradigms for Isolated Environments in HPC (CANOPIE-HPC)*, 2020, σσ. 19–25. DOI: 10.1109/CANOPIEHPC51917.2020.00008

[36] Streamlit Inc., *Streamlit Documentation*, 2025. επίσκεψη 28 Δεκ. 2025. διεύθν.: https://docs.streamlit.io/

[37] Streamlit Inc., *Building Data Apps with Streamlit (Official Site)*, 2025. επίσκεψη 28 Δεκ. 2025. διεύθν.: https://streamlit.io/

[38] Z. Idrees και L. Zheng, «Low cost air pollution monitoring systems: A review of protocols and enabling technologies», *Journal of Industrial Information Integration*, τόμ. 17, σ. 100 123, 2020. DOI: 10.1016/j.jii.2019.100123

[39] D. Nüst κ.ά., «Ten simple rules for writing Dockerfiles for reproducible data science», *PLOS Computational Biology*, τόμ. 16, αρθμ. 11, e1008316, 2020. DOI: 10.1371/journal.pcbi.1008316

[40] S. Benzidia, N. Makaoui και O. Bentahar, «The impact of big data analytics and artificial intelligence on green supply chain process integration and hospital environmental performance», *Technological Forecasting and Social Change*, τόμ. 165, σ. 120 557, 2021. DOI: 10.1016/j.techfore.2020.120557

[41] Q. Zhao κ.ά., «An Overview of the Applications of Earth Observation Satellite Data: Impacts and Future Trends», *Remote Sensing*, τόμ. 14, αρθμ. 8, σ. 1863, 2022. DOI: 10.3390/rs14081863

[42] M. N. A. Ramadan, M. A. H. Ali, S. Y. Khoo, M. Alkhedher και M. Alherbawi, «Real-time IoT-powered AI system for monitoring and forecasting of air pollution in industrial environment», *Ecotoxicology and Environmental Safety*, τόμ. 283, σ. 116 856, 2024. DOI: 10.1016/j.ecoenv.2024.116856

[43] Z. Li, «Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost», *Computers, Environment and Urban Systems*, τόμ. 96, σ. 101 845, 2022. DOI: 10.1016/j.compenvurbsys.2022.101845

[44] A. Barnwal, H. Cho και T. Hocking, «Survival Regression with Accelerated Failure Time Model in XGBoost», *Journal of Computational and Graphical Statistics*, τόμ. 31, αρθμ. 4, σσ. 1292–1302, 2022. DOI: 10.1080/10618600.2022.2067548

[45] P. Srinivas και R. Katarya, «hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost», *Biomedical Signal Processing and Control*, τόμ. 73, σ. 103 456, 2022. DOI: 10.1016/j.bspc.2021.103456

[46] J. M. Ahn, J. Kim και K. Kim, «Ensemble Machine Learning of Gradient Boosting (XGBoost, LightGBM, CatBoost) and Attention-Based CNN-LSTM for Harmful Algal Blooms Forecasting», *Toxins*, τόμ. 15, αρθμ. 10, σ. 608, 2023. DOI: 10.3390/toxins15100608