«Σχολή Θετικών Επιστημών & Τεχνολογίας»

«Μεταπτυχιακή Εξειδίκευση στα Συστήματα Κινητού και Διάχυτου Υπολογισμού»

Διπλωματική Εργασία

«Χρήση μηχανικής μάθησης για την πρόβλεψη της ατμοσφαιρικής ρύπνασης σε αστικά περιβάλλοντα»

Δημήτριος Κάββουρας

Επιβλέπων καθηγητής: Θεόδωρος Παναγιωτακόπουλος

Πάτρα, Σεπτέμβριος 2023

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

«Χρήση μηχανικής μάθησης για την πρόβλεψη της ατμοσφαιρικής ρύπνασης σε αστικά περιβάλλοντα»

Δημήτριος Κάββουρας

Επιτροπή Επίβλεψης Διπλωματικής Εργασίας

Επιβλέπων Καθηγητής:
Θεόδωρος Παναγιωτακόπουλος
Επιστημονικός συνεργάτης, ΕΑΠ

Συν-Επιβλέπων Καθηγητής:
Αχιλλέας Καμέας
Καθηγητής ΕΑΠ

Πάτρα, Σεπτέμβριος 2023

«Θα ήθελα να ευχαριστήσω τον Κύριο Θεόδωρο Παναγωτακόπουλο, επιβλέποντα καθηγητή, για την βοήθεια την υποστήριξη, την κατανόηση, και την καθοδήγηση που έδειξε σε όλη την διάρκεια εκπόνησης της εργασίας, καθώς και τον συν-επιβλέπων Κύριο Αχιλλέα Καμέα. *»*

Περίληψη

Η πρόβλεψη της ποιότητας του αέρα είναι ένα δύσκολο πρόβλημα λόγω της δυναμικής φύσης, της αστάθειας και της μεγάλης μεταβλητότητας στο χρόνο και στο χώρο των ρύπων και των σωματιδίων. Ταυτόχρονα, η ικανότητα μοντελοποίησης, πρόβλεψης και παρακολούθησης της ποιότητας του αέρα αποκτά ολοένα και μεγαλύτερη σημασία, ιδίως στις αστικές περιοχές, λόγω των παρατηρούμενων κρίσιμων επιπτώσεων της ατμοσφαιρικής ρύπανσης στην υγεία των πολιτών στο περιβάλλον και την οικονομία, οπότε είναι σημαντική η πρόβλεψη της ποιότητας του αέρα. Για τον σκοπό αυτό, θα πρέπει να εφαρμοστούν νέες τεχνικές μηχανικής μάθησης. Στην εργασία παρουσιάζουμε διάφορους αλγορίθμους μηχανικής μάθησης που χρησιμοποιούνται για την πρόβλεψη της ποιότητας του αέρα. Με βάση την έρευνα στον τομέα, προτείνουμε στατιστικά(ARMA- FFT -Theta), μοντέλα παλινδρόμησης και βαθιάς μάθησης(LSTM-CNN-MLP), τα οποία μπορούν να χρησιμοποιηθούν για την πρόβλεψη της ατμοσφαιρικής ρύπανσης, και του δείκτη ποιότητας του αέρα . Αυτοί οι αλγόριθμοι έχουν δοκιμαστεί χρησιμοποιώντας χρονοσειρές για σωματίδια PM10, PM2.5 και AQI(δείκτης ποιότητας αέρα). Τα αποτελέσματα έδειξαν ότι οι αλγόριθμοι ARIMA και SVM έχουν την καλύτερη απόδοση στην πρόβλεψη των συγκεντρώσεων των ατμοσφαιρικών ρύπων που μελετήθηκαν (PM2.5, PM10), ενώ τα μοντέλα βαθιάς μάθησης(LSTM) παρουσιάζουν καλύτερη απόδοση για τον δείκτη ποιότητας του αέρα.

Λέξεις – Κλειδιά

Δείκτης Ποιότητας Αέρα, Παλυνδρόμηση Τυχαίο Δάσος, Μηχανές Διανυσμάτων Υποστήριξης, Βαθια Μάθηση, Νευρωνικά Δίκτυα, Arima.

Abstract

Predicting air quality is a difficult problem because of the dynamic nature, instability and high variability in time and space of pollutants and particles. At the same time, the ability to model, predict and monitor air quality is becoming increasingly important, especially in urban areas, due to the observed critical impacts of air pollution on citizens' health in the environment and the economy, so air quality prediction is important. To this end, new machine learning techniques should be applied. In this paper, we present several machine learning algorithms used for air quality prediction. Based on the research in the field, we propose statistical(ARMA- FFT -Theta), regression and deep learning models(LSTM-CNN-MLP), which can be used to predict air pollution, and air quality index . These algorithms have been tested using time series for PM10, PM2.5 and AQI (air quality index) particles. The results showed that Arima and SVM algorithms have the best performance in predicting the concentrations of the studied air pollutants (PM2.5, PM10), while the deep learning(LSTM) models show better performance for air quality index.

Keywords

Air quality Index, Machine learing, Random Forest, Support Vector Machine, Deep Learing, Neural Network. Arima.

Περιεχόμενα

Κατάλογος Εικόνων / Σχημάτων

Κατάλογος Πινάκων

Συντομογραφίες & Ακρωνύμια

AQI	Air Quality Index
ARIMA	Autoregressive Integrated Moving Average
CNN	Convolution Neural Network
FFT	Fast Fourier Transform
LSTM	Long Short Term Memory
MLP	Multi Layer Perceptron
MSE	Mean Squared Error
RF	Random Forest
RMSE	Root Mean Square Error
SVM	Support Vector Machine

Εισαγωγή

Με την τεχνολογική και οικονομική ανάπτυξη των αστικών περιοχών, ανακύπτουν σοβαρά περιβαλλοντικά προβλήματα, όπως η ατμοσφαιρική ρύπανση. Πολλοί ερευνητές έχουν επισιτίσει την προσοχή και έχουν επικεντρωθεί σε αυτά τα προβλήματα λόγω των επιπτώσεων στην ανθρώπινη υγεία. Οι πληροφορίες πρόβλεψης της ποιότητας του αέρα είναι ένας από τους καλύτερους τρόπους μέσω των οποίων οι άνθρωποι μπορούν να ενημερωθούν ώστε να είναι πιο προσεκτικοί για τα σοβαρά ζητήματα υγείας και να προστατεύσουν την ανθρώπινη υγεία που προκαλείται από την ατμοσφαιρική ρύπανση. Σε πολλές αστικές περιοχές η ατμοσφαιρική ρύπανση αποτελεί μείζον περιβαλλοντικό ζήτημα. Σχεδόν καμία αστική περιοχή δεν ακολουθεί πλήρως τις κατευθυντήριες γραμμές για την ποιότητα του αέρα που έχει θέσει ο Παγκόσμιος Οργανισμός Υγείας ΠΟΥ(Limb[1]; WHO[2]). Προκειμένου να αμβλυνθούν οι επιπτώσεις των αυξημένων ρύπων, είναι ανάγκη να διαδοθεί η ευαισθητοποίηση των πολιτών ώστε να περιορίζουν τις υπαίθριες δραστηριότητές τους σε περίπτωση κακών ατμοσφαιρικών συνθηκών (Salnikov et al.[3]). Παράλληλα, είναι επίσης ζωτικής σημασίας η ανάπτυξη στατιστικών μοντέλων που μπορούν να εκτιμούν και να προβλέπουν αποτελεσματικά τη συγκέντρωση των ρύπων. Η μοντελοποίηση της ατμοσφαιρικής ρύπανσης ασχολείται με τις συγκεντρώσεις των ρύπων, τα χαρακτηριστικά τους και τη σύνδεσή τους με τις περιφερειακές μετεωρολογικές συνθήκες για περαιτέρω ερευνητικές εργασίες και επιστημονικές εφαρμογές (Daly[4]et al).. Με τη μοντελοποίηση της ατμοσφαιρικής ρύπανσης μπορεί κανείς να εκτιμήσει το επίπεδο της ατμοσφαιρικής ρύπανσης και να αξιολογήσει τις επιπτώσεις της στο περιβάλλον και την ανθρώπινη υγεία (Brunekreef et al.[5]). Επιπλέον, λαμβάνοντας υπόψη τη σχέση των πηγών εκπομπών με τους ατμοσφαιρικούς ρύπους καθώς και με τις περιφερειακές και μετεωρολογικές παραμέτρους, ο ρόλος των μοντέλων αυτών είναι απαραίτητος (Lelieveld et al[6]). Εκτός από τον προσδιορισμό των πραγματικών πηγών εκπομπών, οι μελλοντικές λύσεις μετριασμού είναι η άλλη σημαντική συμβολή της μοντελοποίησης της ατμοσφαιρικής ρύπανσης. Οι τεχνικές μοντελοποίησης της ατμοσφαιρικής ρύπανσης χωρίζονται κυρίως σε τρεις τύπους: 1) ατμοσφαιρική χημεία, 2) διασπορά 3) μηχανική μάθηση. Για να αντιμετωπιστούν οι περιορισμοί των παραδοσιακών μοντέλων, οι προσεγγίσεις μηχανικής μάθησης που βασίζονται σε στατιστικούς αλγορίθμους φαίνονται πολλά υποσχόμενες. Αντί να λαμβάνουν υπόψη φυσικές και χημικές διεργασίες, τα στατιστικά μοντέλα βασίζονται αυστηρά σε ιστορικά δεδομένα για να κάνουν προβλέψεις

ατμοσφαιρικής ρύπανσης. Η παλινδρόμηση, οι χρονοσειρές και ο αυτοπαλινδρομικός ολοκληρωμένος κινητός μέσος όρος (ARIMA) είναι οι πιο συνηθισμένες στατιστικές προσεγγίσεις που εφαρμόζονται στον τομέα της επιστήμης και της μηχανικής του περιβάλλοντος. Παρόλο που τα μοντέλα που βασίζονται στην παλινδρόμηση μπορούν να παρέχουν καλά αποτελέσματα, εντούτοις, η μη γραμμική συμπεριφορά των ατμοσφαιρικών ρύπων και άλλα επιδραστικά περιφερειακά χαρακτηριστικά οδηγούν σε ένα πολύ σύνθετο σύστημα δημιουργίας ατμοσφαιρικών ρύπων (Brunelli *et al.*[7]). Για το λόγο αυτό, προηγμένες προσεγγίσεις μηχανικής μάθησης(Νευρωνικά δίκτυα – βαθιά μάθηση, μηχανές διανυσμάτων υποστήριξης, κ.α.) είναι γνωστοί λόγω της ικανότητάς τους να ξεπερνούν αποτελεσματικά το ζήτημα καταγραφής της τάσης μη γραμμικότητας στη μοντελοποίηση της ατμοσφαιρικής ρύπανσης. Έχει αποδειχθεί από μελέτες στη μοντολοποίηση της ποιότητας του αέρα τα Νευρωνικά δίκτυα προτιμώνται έναντι των κλασσικών στατιστικών μεθόδων, για την ικανότητα τους να αποδίδουν καλύτερες επιδόσεις και να χειρίζονται τη μη γραμμικότητα και την πολυπλοκότητα των αρχείων καταγραφής των ρύπων. Ωστόσο η εφαρμογή τους δείχνει ότι τα μοντέλα πρόβλεψης που βασίζονται σε νευρωνικά δίκτυα παρουσιάζουν διάφορα προβλήματα όπως τοπικά ελάχιστα, η κατάλληλη αρχιτεκτονική δικτύου, υπέρ-προσαρμογή και η μικρή γενίκευση. Σε γενικές γραμμές τα μοντέλα στατιστικής – μηχανικής μάθησης παρουσιάζουν καλύτερη προγνωστική απόδοση σε σχέση με τα παραδοσιακά μοντέλα ανάπτυξης. Η ακριβής πρόβλεψη δεδομένων χρονοσειρών για την ποιότητα του αέρα είναι ένας ερευνητικός τομέας που έχουν καταβληθεί πολλές προσπάθειες από τους ερευνητές για τη δημιουργία μοντέλων ικανών να προσαρμόσουν τις υποκείμενες χρονοσειρές. Συχνά, η πρόβλεψη της ποιότητας του αέρα περιλαμβάνει ένα θορυβώδες και περιορισμένο όγκο ιστορικών δεδομένων. Επιπλέον, η πρόβλεψη μιας μεμονωμένης εξαρτάται συνήθως από πολλά γεγονότα που εξαρτώνται το ένα από το άλλο. Τα μοντέλα αναγκάζονται τότε να περιλαμβάνουν ειδικά προσαρμοσμένες τεχνικές για να αντιμετωπίσουν εσφαλμένα ή ελλιπή δεδομένα Αυτά τα πολύπλοκα προβλήματα καθιστούν δύσκολη τη γενίκευση της λύσης ώστε να μπορεί να μεταφερθεί και σε άλλες τοποθεσίες. Εκτός αυτού, ο αέρας αλλάζει γρήγορα σε μικρά χρονικά διαστήματα, με τα ωριαία δεδομένα να είναι πιο αβέβαια σε σύγκριση με τις μηνιαίες και ετήσιες τάσεις και την εποχικότητα. Η έλλειψη και η κακή ποιότητα δεδομένων, η χαμηλή χωρική ανάλυση των σημείων δεδομένων και το κόστος των αισθητήρων υψηλής ποιότητας προστίθενται στον κατάλογο των προβλημάτων. Στόχος της εργασίας μας είναι να

δημιουργήσουμε διαφορετικά μοντέλα πρόβλεψης για δύο διαφορετικούς ρύπους: PM2.5, PM 10 και τον δείκτη ποιότητας αέρα.

1.1 Αντικείμενο εργασίας

Η ατμοσφαιρική ρύπανση είναι ένα μείζον ζήτημα για τις σύγχρονες πόλεις που επηρεάζονται κυρίως από οχήματα μεταφοράς και βιομηχανικές εγκαταστάσεις και συγκαταλέγεται στους σημαντικότερους παγκόσμιους παράγοντες κινδύνου για πρόωρη θνησιμότητα. Με βάση στατιστικά στοιχεία του Παγκόσμιου Οργανισμού Υγείας, το 90% των ανθρώπων εκτίθεται στον αέρα που περιέχει υψηλά επίπεδα ρύπων, περιστασιακά πολύ πέρα από τα αντίστοιχα όρια, με αποτέλεσμα περισσότερους από επτά εκατομμύρια θανάτους ετησίως από καρκίνο του πνεύμονα, εγκεφαλικό επεισόδιο, καρδιαγγειακές και άλλες ασθένειες λόγω της ατμοσφαιρικής ρύπανσης του περιβάλλοντος (εξωτερικού χώρου) και των νοικοκυριών. Σε απάντηση αυτής της απειλής, οι πόλεις χρησιμοποιούν ολοένα και περισσότερο τεχνολογίες του Διαδικτύου των Πραγμάτων για την υλοποίηση λύσεων παρακολούθησης της ατμοσφαιρικής ρύπανσης μέσω της τακτικής συλλογής δεδομένων περιβάλλοντος και του συμπερασμού της κατάστασης της ποιότητας του αέρα. Οι λύσεις αυτές, ενώ υστερούν σε ακρίβεια σε σύγκριση με τις συμβατικές συσκευές παρακολούθησης του αέρα, ενισχύουν σε μεγάλο βαθμό την περιοχή κάλυψης διευκολύνοντας την απόκτηση μεγάλων ποσοτήτων δεδομένων και επιτρέποντας μία χωρικά λεπτομερή και αυξημένης ευαισθησίας ανίχνευση μεταβολών στην ατμοσφαιρική ρύπανση. Ως αποτέλεσμα, πολλές έξυπνες υπηρεσίες πόλης μπορούν να πραγματοποιηθούν είτε αντιδραστικά είτε προληπτικά.

1.2 Δομή εργασίας

Κεφάλαιο 2: Το κεφάλαιο 2 παρουσιάζει το θεωρητικό υπόβαθρο μαζί με τις μεθόδους και τεχνικές για την κατανόηση του πεδίου του θέματος.

Κεφάλαιο 3: Γίνεται βιβλιογραφική επισκόπηση των ερευνητικών εργασιών με μεθόδους μηχανικής μάθησης για την ατμοσφαιρική ρύπανση και την πρόβλεψη της ποιότητας του αέρα.

Κεφάλαιο 4: Στο κεφάλαιο 4 παρουσιάζεται τη μεθοδολογία υλοποίησης των μοντέλων μηχανικής μάθησης.

Κεφάλαιο 5: Στο κεφάλαιο αυτό περιγράφονται τα πειράματα και η εφαρμογή των μοντέλων μηχανικής μάθησης, επίσης καταγράφονται και αναλύονται τα αποτελέσματα των πειραμάτων.

Θεωρητικό Υπόβαθρο

Το κεφάλαιο αυτό αναφέρεται στο θεωρητικό υπόβαθρο για να γίνουν κατανοητοί οι όροι-έννοιες και τεχνικές που χρησιμοποιούνται στην εργασία. Η ενότητα 2.1 αναφέρατε στην παρακολούθησης της ποιότητας του αέρα. Η ενότητα 2.2 περιγράφει τον ορισμό του δείκτη ποιότητας του αέρα και πως υπολογίζεται. Η ενότητα 2.3 περιγράφει τους αισθητήρες IoT που χρησιμοποιούνται. Στην ενότητα 2.4 περιγράφονται τα βασικά στοιχεία μιας χρόνο-σειράς. Στην ενότητα 2.5 παρουσιάζονται τρις στατιστικές προσεγγίσεις για την πρόβλεψη χρόνο-σειρών. Τέλος η ενότητα 2.6 παρουσιάζει εισαγωγικά τις τεχνικές μηχανικής μάθησης που εφαρμόζονται για την πρόβλεψη, και στις υπό-ενότητες που ακολουθούν εμβαθύνει σε κλασσικές μεθόδους μηχανικής μάθησης, καθώς και βαθιάς μάθησης με νευρωνικά δίκτυα.

2.1 Παρακολούθηση της ποιότητας του αέρα

Η ατμοσφαιρική ρύπανση (ΑΡ) είναι ένα από τα σοβαρότερα και σημαντικότερα περιβαλλοντικά προβλήματα για τις σύγχρονες πόλεις που επηρεάζονται κυρίως από οχήματα μεταφοράς και βιομηχανικές εγκαταστάσεις και συγκαταλέγεται στους σημαντικότερους παγκόσμιους παράγοντες κινδύνου για πρόωρη θνησιμότητα. Τα αιωρούμενα σωματίδια (ΑΣ), το διοξείδιο του αζώτου (NO2) και το όζον (O3) προκαλούν τις σημαντικότερες βλάβες στην ανθρώπινη υγεία. Οι υψηλές συγκεντρώσεις ατμοσφαιρικής ρύπανσης έχουν επιπτώσεις, ιδίως στους κατοίκους των αστικών περιοχών. Η ατμοσφαιρική ρύπανση συνεπάγεται επίσης σημαντικές οικονομικές επιπτώσεις, καθώς μειώνει το προσδόκιμο ζωής, αυξάνει τις ιατρικές δαπάνες και μειώνει την παραγωγικότητα στο σύνολο της οικονομίας μέσω της απώλειας εργάσιμων ημερών λόγω προβλημάτων υγείας.

Με βάση στατιστικά στοιχεία του Παγκόσμιου Οργανισμού Υγείας, το 90% των ανθρώπων εκτίθεται στον αέρα που περιέχει υψηλά επίπεδα ρύπων, περιστασιακά πολύ πέρα από τα αντίστοιχα όρια, με αποτέλεσμα περισσότερους από επτά εκατομμύρια θανάτους ετησίως από καρκίνο του πνεύμονα, εγκεφαλικό επεισόδιο, καρδιαγγειακές και άλλες ασθένειες λόγω της ατμοσφαιρικής ρύπανσης του περιβάλλοντος (εξωτερικού χώρου) και των νοικοκυριών.

Ως αποτέλεσμα, οι άνθρωποι αναζητούν καλύτερους τρόπους για να παρακολουθούν την ποιότητα του αέρα στο άμεσο περιβάλλον τους, ώστε να λαμβάνουν τα κατάλληλα μέτρα,

όπως το να φορούν μάσκες ή να μένουν στο σπίτι. Ενώ υπάρχουν πολλές εφαρμογές για smart-phone που αναφέρουν δημόσια διαθέσιμα δεδομένα για την ποιότητα του αέρα σε επίπεδο πόλης ή περιοχής, δεν μπορούν να αναφέρουν την πραγματική ποιότητα του αέρα που αναπνέουν οι άνθρωποι, η οποία είναι πολύ πιο σημαντική. Αυτό είναι ιδιαίτερα σημαντικό δεδομένου ότι περνάμε τον περισσότερο χρόνο μας σε κλειστούς χώρους όπως τα σπίτια και τα γραφεία, όπου η ποιότητα του αέρα μπορεί να αποκλίνει σημαντικά από την εξωτερική.

– *Χρήση σταθμών παρακολούθησης*

Λαμβάνοντας υπόψη τη σημασία της ποιότητας του αέρα για τις ανθρώπινες ζωές, ο Παγκόσμιος Οργανισμός Υγείας (ΠΟΥ) έχει αναπτύξει κατευθυντήριες γραμμές για τη μείωση των επιπτώσεων της ατμοσφαιρικής ρύπανσης στη δημόσια υγεία, θέτοντας τα όρια των συγκεντρώσεων διαφόρων ατμοσφαιρικών ρύπων, μερικοί από τους οποίους είναι το όζον σε επίπεδο εδάφους (O3), το διοξείδιο του αζώτου (NO2) και το διοξείδιο του θείου (SO2) . Παραδοσιακά, οι συγκεντρώσεις των ατμοσφαιρικών ρύπων μετρούνται με τη χρήση σταθμών παρακολούθησης της ποιότητας του αέρα (AQM), οι οποίοι είναι ιδιαίτερα αξιόπιστοι, ακριβείς, και μπορούν να μετρήσουν ένα ευρύ φάσμα ρύπων με τη χρήση τυποποιημένων αναλυτών. Ωστόσο, οι σταθμοί αυτοί έχουν τρία βασικά μειονεκτήματα: 1) τη σημαντική υποδομή που απαιτείται για την εγκατάστασή τους λόγω του ογκώδους μεγέθους τους, 2) τις περίπλοκες λειτουργικές απαιτήσεις, π.χ. πρόσβαση σε ηλεκτρικό δίκτυο, θέρμανση/ψύξη και ασφαλή καταφύγια, και 3) το απαγορευτικό κόστος απόκτησης, εγκατάστασης και τακτικής συντήρησης και βαθμονόμησης. Αυτά τα μειονεκτήματα μειώνουν τον αριθμό των εγκαταστάσεων και οδηγούν σε αραιά κατανεμημένα δίκτυα AQM με δεδομένα ατμοσφαιρικής ρύπανσης περιορισμένης χωρικής ανάλυσης .

– *Δίκτυο βασισμένο σε έξυπνους αισθητήρες(**IoT**).*

Σε απάντηση αυτής της κατάτασης που έχει δημιουργηθεί οι αστικές περιοχές χρησιμοποιουν όλο και περισσότερο τεχνολογίες του διαδικτύου των πραγμάτων.

Οι συσκευές Internet-of-Things (IoT) γίνονται όλο και πιο διαδεδομένες και ορισμένες από αυτές, όπως οι αισθητήρες, παράγουν συνεχή δεδομένα χρονοσειράς, δηλαδή δεδομένα ροής. Αυτά τα δεδομένα ροής IoT αποτελούν μία από τις πηγές μεγάλων δεδομένων και απαιτούν προσεκτική εξέταση για αποτελεσματική επεξεργασία και ανάλυση δεδομένων.

Τα πρόσφατα μοντέλα συστημάτων προτείνονται συνεχώς με βάση αισθητήρες που βασίζονται στο IoT. Η μέτρηση του AQ από τους αισθητήρες με τη χρήση των δεδομένων που συλλέγονται μπορεί να διαδραματίσει ενεργητικό ρόλο στη διαχείριση των πόλεων. Σε πολλές πόλεις η λήψη αποφάσεων μπορεί να γίνει ακόμη πιο γρήγορα και απλά από ποτέ, με τη βοήθεια αισθητήρων που συλλέγουν δεδομένα.

2.2 Δείκτης ποιότητας αέρα

Ο δείκτης ατμοσφαιρικής ρύπανσης αναγνωρίζεται ως δείκτης ποιότητας αέρα(AQI – Air Quality Index) στις περισσότερες χώρες. Οι ειδικοί ορίζουν τον AQI ως το σημείο αναφοράς για την κατάσταση της ποιότητας του αέρα σε μια συγκεκριμένη περιοχή.

Η ποιότητα του αέρα (AQI) είναι ένας δείκτης που δημιουργήθηκε για να αναφέρει την ποιότητα του αέρα, μετρώντας πόσο καθαρός ή ανθυγιεινός είναι ο αέρας και ποιες συναφείς επιπτώσεις στην υγεία μπορεί να αποτελούν ανησυχία, ιδίως για τις ομάδες κινδύνου. Επικεντρώνεται στις επιπτώσεις στην υγεία που μπορούν να εμφανιστούν μέσα σε λίγες ώρες ή ημέρες μετά την έκθεση σε μολυσμένο αέρα.

Οι παράμετροι που πρέπει να υπολογιστούν για την πρόβλεψη της ποιότητας του αέρα είναι οι εξής: Όζον, Μονοξείδιο του άνθρακα, Διοξείδιο του θείου, διοξείδιο του αζώτου, αιωρούμενα σωματίδια PM 10 με διάμετρο μικρότερη από 10μm , αιωρούμενα σωματίδια PM 2.5 με διάμετρο μικρότερη από 2,5 μm. Χρησιμοποιεί μια κλίμακα που κυμαίνεται από 0 έως 100, αποτελούμενες από 6 κατηγορίες που περιγράφουν τα επίπεδα ατμοσφαιρικής ρύπανσης().Υπάρχουν έξι κατηγορίες AQI, δηλαδή: Πράσινο – Καλή(0-50), Κίτρινο – Μέτρια(51-100), Πορτοκαλί – Ανθυγιεινή για ευαίσθητες ομάδες(101-150), Κόκκινο – Ανθυγιεινή(151 – 200), Μωβ – Πολύ Ανθυγιεινή(201- 300) και Καφέ – Επικίνδυνη(301 – 500) όπως φαίνεται στην παρακάτω εικόνα. Οι πληροφορίες πρόβλεψης της ποιότητας του αέρα είναι ένας από τους καλύτερους τρόπους μέσω των οποίων οι άνθρωποι μπορούν να ενημερωθούν ώστε να είναι πιο προσεκτικοί για τα σοβαρά ζητήματα υγείας και να προστατευτούν από την ατμοσφαιρική ρύπανση.

Ο υπολογισμός του Δείκτη Ποιότητας Αέρα Ο δείκτης AQI (I), ο οποίος είναι γραμμική συνάρτηση της συγκέντρωσης των ρύπων, υπολογίζεται από την εξίσωση 1, όπου επιτυγχάνεται η μετατροπή των συγκεντρώσεων των ρύπων σε αριθμούς.

$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}} (C_{high} - C_{low}) + I_{low}$$

(1)

"I" είναι ο δείκτης ποιότητας αέρα AQI,

"C" είναι η μετρούμενη συγκέντρωση ρύπων,

"Clow" το μικρότερο όριο συγκέντρωσης στην κατηγορία στην οποία αντιστοιχεί ο ρύπος C,

"Chigh" το μεγαλύτερο όριο της συγκέντρωσης, στην κατηγορία στην οποία αντιστοιχεί ο ρύπος C,

"Ilow" η μικρότερη τιμή του δείκτη ποιότητας του αέρα που ανήκει στο μικρότερο όριο του ρύπου "Clow"

και "Ihigh" Η μεγαλύτερη τιμή του δείκτη ποιότητας αέρα που αντιστοιχεί στο μεγαλύτερο όριο του ρύπου "Chigh".

AQI Breakpoint Table

Category	AQI	$PM_{2.5}$ ($\mu g/m^3$) 24hr avg	PM_{10} ($\mu g/m^3$) 24hr avg
Good	0--50	0--12.0	0--54
Moderate	51--100	12.1--35.4	55--154
Unhealthy for Sensitive Groups	101--150	35.5--55.4	155--254
Unhealthy	151--200	55.5--150.4	255--354
Very Unhealthy	201--300	150.5--250.4	355--424
Hazardous	301--500	250.5--500.4	425--604

Εικόνα 2.1: Κατηγορίες του δείκτη ποιότητας του αέρα

Παράμετροι Ατμοσφαιρικών Ρύπων	
Όζον(O3)	Σχηματίζεται από χημικές αντιδράσεις μεταξύ αερίων και οξειδίων, παρουσία ηλιακού φωτός με το όζον.
Μονοξείδιο του άνθρακα(CO)	Οχήματα και οποιαδήποτε μορφής καύσης
Διοξείδιο του θείου(SO2)	Βιομηχανία, οχήματα(χρήση καυσίμου με μεγάλη περιεκτικότητα θείου)
Διοξείδιο του Αζώτου(NO2)	Εργοστάσια και μεταφορικά μέσα
PM 2.5	Σωματιδιακή αιωρούμενη ύλη μικρότερη από 2,5 μικρόμετρα
PM 10	Σωματιδιακή αιωρούμενη ύλη με διάμετρο μικρότερη από 10 μικρόμετρα

Πίνακας 2.1: Βασικοί ατμοσφαιρική ρύποι. Οι ατμοσφαιρικοί ρύποι που εξετάζουμε στην εργασία είναι τα αιωρούμενα σωματίδια PM 2.5 και PM 10.

2.3 IoT αισθητήρες

Οι αισθητήρες είναι το τελευταίο επίπεδο σε όλα τα συστήματα IOT, έρχονται σε άμεση επαφή με το περιβάλλον και επηρεάζονται από αυτό. Συλλέγουν πληροφορίες από το περιβάλλον και τις μεταδίδουν στο επόμενο επίπεδο του συστήματος. Όντας η πηγή πληροφοριών, αποτελούν το πιο κρίσιμο τμήμα ενός συστήματος παρακολούθησης. Επί του παρόντος, υπάρχει μεγάλος αριθμός αισθητήρων για την παρακολούθηση της ποιότητας του αέρα. Οι αισθητήρες αυτοί ποικίλλουν ανάλογα με τον μορφή αέριο/στερεό) του ρύπου που στοχεύουν, την τεχνολογία που χρησιμοποιείται για τη μέτρησή του και τον ίδιο τον ρύπο.

Αισθητήρες Gaz

Υπάχει ένα μεγάλο ποσοστό των ρύπων που είναι επιβλαβείς για την ανθρώπινη υγεία, και είναι παρόντες στο περιβάλλον σε αέρια μορφή. Για την παρακολούθηση αυτών των ουσιών, ένα ευρύ φάσμα φάσμα αισθητήρων χαμηλού κόστους είναι διαθέσιμο στην αγορά. Αυτοί οι αισθητήρες χρησιμοποιούν διαφορετικές τεχνολογίες για τη συλλογή των συγκεντρώσεων των επικίνδυνων σωματιδίων στον αέρα. Μεταξύ των πιο ευρέως χρησιμοποιούμενων τεχνολογίες στα συστήματα παρακολούθησης της ποιότητας του αέρα είναι οι στερεάς κατάστασης και οι ηλεκτροχημικοί αισθητήρες.

1) Αισθητήρες αερίων στερεάς κατάστασης: Ένας αισθητήρας στερεάς κατάστασης χρησιμοποιεί μεταλλικά επιφάνειες οξειδίων για την αφαίρεση της συγκέντρωσης ενός συγκεκριμένου αερίου του περιβάλλοντος. Τα αέρια του περιβάλλοντος διαχωρίζονται σε

ιόντα φορτιστή ή συμπλέγματα που κάνουν τα ηλεκτρόνια να συσσωρεύονται στην επιφάνεια των οξειδίων μετάλλων όταν εκτίθενται σε αυτά. Αυτή η συσσώρευση δημιουργεί μια μετρήσιμη μεταβολή της αγωγιμότητας που χρησιμοποιείται για τον υπολογισμό της συγκέντρωσης των αερίων. Προστίθεται θερμαντικό στοιχείο για την αύξηση την αντίδραση και την κανονική θερμοκρασία [23]. Τα πλεονεκτήματα αυτών των αισθητήρων είναι το μικρό τους μέγεθος, η υψηλή ευαισθησία και το χαμηλό κόστος. Ωστόσο, πάσχουν από περιορισμένη ακρίβεια μέτρησης, μεγάλο χρόνο προθέρμανσης και προβλήματα μακροχρόνιας σταθερότητας [23].

2) Ηλεκτροχημικοί αισθητήρες αερίων: Βασίζονται στον αντίδραση οξείδωσης-αναγωγής που λαμβάνει χώρα μεταξύ του αισθητήρα και των μορίων του αερίου του περιβάλλοντος. Η αντίδραση αυτή παράγει ένα ηλεκτρικό σήμα (ρεύμα) ανάλογο της συγκέντρωσης του του αερίου του περιβάλλοντος [2].

Αισθητήρες σωματιδίων

Λαμβάνοντας υπόψη τις απειλές για την υγεία που προκαλούνται από τον πολλαπλασιασμό των μικρό-σωματιδίων στον αέρα , είναι ζωτικής σημασίας η ακριβείς συλλογή και τακτικές μετρήσεις της συγκέντρωσής τους. Για το σκοπό αυτό, υπάρχουν διάφορες τεχνολογίες αισθητήρων χαμηλού κόστους.

Το που χρησιμοποιούνται ευρύτερα στον τομέα της IOT είναι οι εξής:

1) Διασπορά φωτός: Σε αυτόν τον τύπο ανιχνευτή, ένα λέιζερ υψηλής ενέργειας ως πηγή φωτός σκεδάζεται από το σωματίδιο που διέρχεται μέσω ενός θαλάμου ανίχνευσης. Στη συνέχεια, το μέγεθος και ο αριθμός των σωματιδίων από την ανάλυση της έντασης του σκεδαζόμενου φωτός [24].

2) Φωτοσκίαση (νεφελόμετρο): Ένα νεφελόμετρο είναι ένα όργανο που χρησιμοποιεί μια λυχνία LED εγγύς υπέρυθρης ακτινοβολίας και ένα πυρίτιο ανιχνευτή για τη μέτρηση του μεγέθους και της συγκέντρωσης μάζας των PM σε στον ατμοσφαιρικό αέρα αναλύοντας τις εντάσεις του σκεδαζόμενου φωτός και του σχήματος του μοτίβου σκέδασης [24]. Σε αυτά τα συστήματα, οι αισθητήρες εγκαθίστανται σε μια ηλεκτρονική πλακέτα που είναι επιφορτισμένη με την παροχή ενέργειας και τη συλλογή πληροφοριών. Για το σκοπό αυτό, η επιλογή της κατάλληλης πλακέτας είναι ένα κρίσιμο σημείο στο σύστημα παρακολούθησης της ποιότητας του αέρα.

2.4 Χρόνο-σειρές

Ο πρωταρχικός στόχος της ανάλυσης χρονοσειρών είναι η ανάπτυξη μαθηματικών μοντέλων που παρέχουν αληθοφανείς περιγραφές για δειγματοληπτικά δεδομένα, όπως αυτά που συναντώνται σε σύνολα δεδομένων. Προκειμένου να δοθεί ένα στατιστικό πλαίσιο για την περιγραφή του χαρακτήρα των δεδομένων που φαινομενικά κυμαίνονται με τυχαίο τρόπο με την πάροδο του χρόνου, υποθέτουμε ότι μια χρόνο-σειρά μπορεί να οριστεί ως μια συλλογή τυχαίων μεταβλητών που διατάσσονται σύμφωνα με τη σειρά με την οποία λαμβάνονται στο χρόνο. Για παράδειγμα, μπορούμε να θεωρήσουμε μια χρόνο-σειρά ως μια ακολουθία τυχαίων μεταβλητών, x1, x2, x3,... , όπου η τυχαία μεταβλητή x1 δηλώνει την τιμή που λαμβάνει η σειρά κατά την πρώτη χρονική στιγμή, η μεταβλητή x2 δηλώνει την τιμή για τη δεύτερη χρονική περίοδο, x3 δηλώνει την τιμή για την τρίτη χρονική περίοδο, κ.ο.κ. Γενικά, μια συλλογή τυχαίων μεταβλητών, {xt}, με δείκτη t αναφέρεται ως στοχαστική διαδικασία. Το t είναι συνήθως διακριτό και κυμαίνεται μεταξύ των ακεραίων t = 0, ±1, ±2, ..., ή κάποιο υποσύνολο των ακεραίων. Οι παρατηρούμενες τιμές μιας στοχαστικής διαδικασίας αναφέρονται ως υλοποίηση της στοχαστικής διαδικασίας δηλαδή της χρόνο-σειράς. Μια χρόνο-σειρά υπολογίζεται ότι επηρεάζεται από τέσσερις κύριες συνιστώσες: Τάση, Εποχιακή, Κυκλική και Ακανόνιστη. Η τάση είναι μια γενική τάση μεταβολών μιας χρόνο-σειράς, με μια μακροπρόθεσμη κίνηση αύξησης, μείωσης ή στασιμότητας. Οι εποχικές μεταβολές συνδέονται με αλλαγές κατά τη διάρκεια των εποχών του έτους, όπου το κλίμα και ο καιρός αποτελούν σημαντικούς παράγοντες. Το κυκλικό μέρος περιγράφει τις διαφορές ως μεσοπρόθεσμες μεταβολές σε μια χρόνο-σειρά, που προκαλούνται από συνθήκες με κυκλικό χαρακτήρα. Η τελευταία συνιστώσα, η

ακανόνιστη, είναι τυχαίες μεταβολές ή ο λεγόμενος θόρυβος, οι οποίες δεν είναι τυπικές και δεν είναι σε θέση να περιγραφούν από τα προηγούμενα μέρη.

2.5 Στατιστικές μέθοδοι

Οι στατιστικές προσεγγίσεις χρησιμοποιούνται παραδοσιακά για να προβλέψεις. Μία από τις πιο ευρέως αναγνωρισμένες διαδικασίες για την πρόβλεψη χρονολογικών σειρών είναι η αυτοπαλινδρομική ολοκληρωμένη Κινητός Μέσος Όρος (ARIMA) [6], η οποία έχει χρησιμοποιηθεί σε πολλές μελέτες. Αυτή η τεχνική έχει αποδειχθεί ότι είναι πολύ αποτελεσματική όταν τα δεδομένα είναι στάσιμα και δεν παρουσιάζουν ανοδικές ή

καθοδικές τάσεις. Ωστόσο, η αποτελεσματικότητα της ARIMA παραμένει σχετική μόνο για γραμμικά συστήματα, καθιστώντας ακατάλληλη σε πολλές περιπτώσεις.

2.5.1 ARIMA

Το μοντέλο ARIMA είναι μια δημοφιλής στατιστική μέθοδος για την πρόβλεψη χρόνο-σειρών), χρησιμοποιεί ιστορικές πληροφορίες για να κάνει προβλέψεις και ως εκ τούτου αποτλεί μια δημοφιλή και ευέλικτη κατηγορία πρόβλεψης ενός μοντέλου. Δεδομένου ότι αυτός ο τύπος μοντέλου χρησιμοποιείται ως βάση για πιο σύνθετα μοντέλα, ονομάζεται βασική τεχνική πρόβλεψης. Είναι επίσης γνωστή ως προσέγγιση Box-Jenkins. Το ίδιο το ακρωνύμιο περιγράφει τα χαρακτηριστικά του μοντέλου. Οι συνιστώσες είναι η αυτό-παλινδρομική **(**AR**),** η ολοκληρωμένη συνιστώσα και ο κινητός μέσος όρος (MA). Η αυτό-παλινδρόμηση αντιπροσωπεύει τις επιδράσεις των προηγούμενων παρατηρήσεων. Η ολοκληρωμένη συνιστώσα αντιπροσωπεύει τις τάσεις και τις εποχιακές τάσεις. Ο κινητός μέσος όρος αντιπροσωπεύει τις επιδράσεις των προηγούμενων τυχαίων σφαλμάτων. Η μαθηματική αναπαράσταση του μοντέλου ARIMA δίνεται στην εξίσωση (1) [31].

$$y_t' = c + \emptyset_1 y_{t-1}' + \ldots + \emptyset_1 y_{t-p}' + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \theta_1 \varepsilon_t \qquad (1)$$

όπου *y;* είναι η σειρά διαφορών. Οι προγνωστικοί παράγοντες στη δεξιά πλευρά αποτελούνται από τις καθυστερημένες τιμές του y και τα καθυστερημένα σφάλματα. Το μοντέλο έχει τρεις παραμέτρους (p, *d, q),* όπου,

p = τάξη του αυτό-παλίνδρομου μέρους,

d = βαθμός πρώτης διαφοράς,

q = τάξη του κινητού μέσου όρου.

Το μοντέλο ARIMA χρησιμοποιεί διαφοροποιημένα δεδομένα για να τα καταστήσει στάσιμα, πράγμα που σημαίνει ότι υπάρχει συνέπεια των δεδομένων με την πάροδο του χρόνου. Το μοντέλο μπορεί να δημιουργηθεί χρησιμοποιώντας εποχικές και μη εποχικές διαμορφώσεις. Ένα εποχικό μοντέλο πρέπει να λαμβάνει υπόψη των αριθμό των γεγονότων σε κάθε εποχή, εκτός από τους αυτό-παλίνδρομους διαφορικούς μέσους όρους της κάθε εποχής.

2.5.2 Γρήγορος μετασχηματισμός Fourier

Ο FFT είναι μια παραλλαγή του Διακριτού Μετασχηματισμού Fourier (DFT) με μόνο διαφορά ότι ο FFT είναι υπολογιστικά ταχύτερος (Press et al., 2002).

Υπολογιστικά, ο DFT είναι της τάξης του O(N2), ενώ ο FFT είναι της τάξης του της τάξης του O(N.log2N). Εάν ðx0; x1; x2.:; xN1Þ υποδηλώνουν μια χρονοσειρά, η DFT Hn δίνεται από τη σχέση

$$H_n = \sum_{k=0}^{N-1} x_k e^{2\pi i k n/N} \qquad (1)$$

η οποία μπορεί να αντιστραφεί με αντίστροφο μετασχηματισμό Fourier ως εξής:

$$x_k = (1/N) \sum_{n=0}^{N-1} H_n e^{-2\pi i k n/N} \qquad (2)$$

Η εξίσωση (1) είναι περιοδική στο n με περίοδο N. Συνεπώς, η συχνότητα κυμαίνεται από 1/2 έως 1/2 στο διακριτό διάστημα n/N. Τώρα η εκτιμήσεις του περιοδογράμματος του φάσματος ισχύος σε διαφορετικά συχνότητες δίνεται ως εξής[25]

$$\begin{aligned} P(0) &= P(f_0) = \tfrac{1}{N^2}|H_0|^2 \\ P(f_k) &= \tfrac{1}{N^2}\left[|H_k|^2 + |H_{N-k}|^2\right] \end{aligned} \qquad (3)$$

Όπου $f_k(=\frac{k}{N})$ ορίζεται μόνο για τη μηδενική και τη θετική συχνότητα (επίσης, ο μετασχηματισμός Fourier (1) είναι συμμετρικός, δηλ. οι Hk και H-k έχουν την ίδια τιμή). Στην παρούσα μελέτη, έχουμε απεικονίσει την ισχύ σε συνάρτηση με την περίοδο.

Η επιθεώρηση του φάσματος ισχύος μας βοηθά να εντοπίσουμε και να επιλέξουμε τις συχνότητες (περιόδους) που υποτίθεται ότι είναι κυρίαρχες στο χρονοσειρά. Αφού πάρουμε μόνο τον επιλεγμένο αριθμό των κυρίαρχων συχνότητες για μια συγκεκριμένη χρονοσειρά, κατασκευάσαμε το μετασχηματισμό Fourier και πήραμε το αντίστροφό του. Στην παρούσα μελέτη, η χρονική σειρά που προέκυψε μετά τον αντίστροφο μετασχηματισμό Fourier υπό επιλεγμένες συχνότητες έχει ονομαστεί ως συνιστώσα FFT της χρονοσειράς.

2.5.3 Μέθοδος Theta

Η μέθοδος Theta(Θήτα), βασίζεται στην έννοια της τροποποίησης της τοπικής καμπυλότητας μιας χρονοσειράς μέσω ενός συντελεστή Theta, ο οποίος εφαρμόζεται απευθείας στις δεύτερες διαφορές των δεδομένων. Οι σειρές που προκύπτουν διατηρούν τη μέση τιμή και την κλίση των αρχικών δεδομένων αλλά όχι την καμπυλότητά τους. Αυτές οι νέες χρονοσειρές ονομάζονται γραμμές Theta. Το κύριο χαρακτηριστικό τους είναι η βελτίωση της προσέγγισης της μακροπρόθεσμης συμπεριφοράς των δεδομένων ή η αύξηση των βραχυπρόθεσμων χαρακτηριστικών, ανάλογα με την τιμή του συντελεστή Theta. Η προτεινόμενη μέθοδος αποσυνθέτει την αρχική χρονοσειρά σε δύο ή περισσότερες διαφορετικές γραμμές Theta. Αυτές προεκτείνονται χωριστά και οι επακόλουθες προβλέψεις συνδυάζονται.

Το μοντέλο βασίζεται στην έννοια της τροποποίησης των τοπικών καμπυλοτήτων της χρονοσειράς. Αυτή η μεταβολή προκύπτει από έναν συντελεστή, που ονομάζεται συντελεστής Θήτα ο οποίος εφαρμόζεται απευθείας στις δεύτερες διαφορές της χρονοσειράς:

$$X''_{\text{new}}(\theta) = \theta \cdot X''_{\text{data}}, \quad \text{where } X''_{\text{data}} = X_t - 2X_{t-1} + X_{t-2} \text{ at time } t.$$

Εάν οι τοπικές καμπυλότητες μειωθούν σταδιακά, τότε η χρονοσειρά αποκλιμακώνεται. Όσο μικρότερη είναι η τιμή του συντελεστή Theta, τόσο μεγαλύτερος είναι ο βαθμός αποκλιμάκωσης. Στην ακραία περίπτωση όπου Θ=0 η χρονοσειρά μετατρέπεται σε γραμμική παλινδρόμηση. Η προοδευτική μείωση των διακυμάνσεων μειώνει τις απόλυτες διαφορές μεταξύ των διαδοχικών όρων της παράγωγης σειράς και σχετίζεται, από ποιοτική άποψη, με την εμφάνιση μακροπρόθεσμων τάσεων στα δεδομένα.

αν η τοπική καμπυλότητα αυξηθεί (Θ>1), τότε η χρονοσειρά επεκτείνεται. Όσο μεγαλύτερος είναι ο βαθμός επέκτασης, τόσο μεγαλύτερη είναι η μεγέθυνση της βραχυπρόθεσμης συμπεριφοράς. Ακολουθώντας αυτή τη διαδικασία, δημιουργείται ένα σύνολο νέων χρονοσειρών, οι λεγόμενες γραμμές Theta. Η τοποθέτηση αυτών των γραμμών σε σχέση με τα αρχικά δεδομένα μπορεί να γίνει με πολλούς διαφορετικούς τρόπους. Εάν η προσαρμογή είναι μια διαδικασία εκτίμησης, τότε ο μέσος όρος και η κλίση των γραμμών –Theta παραμένουν οι ίδιες σε σχέση με εκείνες των αρχικών δεδομένων. Η γενική

διατύπωση της μεθόδου είναι η εξής: Η αρχική χρονοσειρά αποσυντίθεται σε δύο ακόμη γραμμές Theta. Κάθε μία από τις Theta-γραμμές προεκτείνεται χωριστά και οι προβλέψεις απλώς συνδυάζονται. Οποιαδήποτε μέθοδος πρόβλεψης μπορεί να χρησιμοποιηθεί για την προβολή μιας γραμμής Theta. Για κάθε ορίζοντα πρόβλεψης μπορεί να χρησιμοποιηθεί διαφορετικός συνδυασμός γραμμών Theta. Αυτό αποδεικνύεται εξετάζοντας μια από τις απλούστερες περιπτώσεις στην οποία η αρχική χρονοσειρά αναλύεται σε δύο γραμμές Theta, όπως φαίνεται παρακάτω:

Θ = 0 και Θ =2
Data = ½(L(Θ=0) + L(Θ=2))

Όπου L(Θ=0) αντιπροσωπεύει την παράμετρο Θ ίση με μηδέν. Η πρώτη γραμμή theta(θ=0) είναι η γραμμική παλινδρόμηση των δεδομένων και η δεύτερη σειρά έχει διαφορά ακριβώς διπλάσια από την αρχική χρόνο-σειρά. Αυτή είναι μια περίπτωση όπου συν θέτονται δύο, ακραίες και συμμετρικές ως προς το 1, γραμμές theta. Η πρώτη συνιστώσα L(Θ=0) περιγράφει τη χρόνο-σειρά μέσω γραμμικής παλινδρόμησης, η δεύτερη συνιστώσα έχει διπλασιάσει τις τοπικές καμπυλότητες μεγαλώνοντας τη βραχυπρόθεσμη συμπεριφορά. Η πρώτη γραμμή theta εξάγεται με τον συνήθη τρόπο μια γραμμική τάση. Η δεύτερη εξάγεται μέσω απλής εκθετικής εξομάλυνσης.Ένας απλός συνδυασμός αυτών των δύο προβλέψεων παράγει την τελική πρόβλεψη του μοντέλου Θήτα για τη συγκεκριμένη χρόνο-σειρά.

Τα βήματα που ακολουθούν περιγράφουν την παραγωγή προβλέψεων χρονοσειρών:

Βήμα 1. (Αποεποχικοποίηση) Οι χρονοσειρές αποεποχικοποιήθηκαν μέσω της κλασικής μεθόδου πολλαπλασιαστικής αποσύνθεσης.

Βήμα 2. (Αποσύνθεση) Η χρόνο-σειρά χωρίστηκε σε δύο γραμμές Theta, η μια γραμμή δείχνει τη γραμμική παλινδρόμηση(theta = 0) και η δεύτερη γραμμή την theta(θ=2).

Βήμα 3. (Πρόβλεψη) Η ευθεία γραμμή(γραμμική παλινδρόμηση) επεκτείνεται με τον γνωστό τρόπο. Η δεύτερη γραμμή επεκτείνεται μέσω της μεθόδου απλής εκθετικής εξομάλυνσης.

Βήμα 4. (Συνδυασμός) Οι προβλέψεις που προέκυψαν από την παρέκταση των δύο γραμμών συνδυάστηκαν με ίσα βάρη.

Βήμα 5. (Εποχικοποίηση) Οι προβλέψεις επαναπροσδιορίστηκαν.

2.6 Τεχνικές μηχανικής μάθησης

Πολλές πρόσφατες μελέτες βασίζονται στους αλγόριθμους μηχανικής μάθησης, οι οποίοι έχουν το πλεονέκτημα ότι δεν απαιτούν ρητό μαθηματικό μοντέλο. Επιπλέον, μόλις εκπαιδευτούν, οι προβλέψεις είναι σχεδόν στιγμιαίες, σε αντίθεση με τους στατιστικές προσεγγίσεις. Τα νευρωνικά δίκτυα(βαθιά μάθηση) είναι μία κορυφαία διαδικασία σε αυτή την κατηγορία. Βασίζεται σε ένα επαναλαμβανόμενη αρχιτεκτονική που μπορεί να επεξεργαστεί χρονοσειρές και είναι ικανή να μαθαίνει και να θυμάται(π.χ. LSTM) δεδομένα εισόδου για μεγάλο χρονικό διάστημα. Αυτό οδηγεί στην ευρεία χρήση πρόβλεψης χρονοσειρών.

2.6.1 Μηχανική μάθηση

Η μηχανική μάθηση (ML) είναι η επιστημονική μελέτη των αλγορίθμων και των στατιστικών μοντέλων που χρησιμοποιούν τα συστήματα υπολογιστών για να εκτελέσουν μια συγκεκριμένη εργασία χωρίς να έχουν προγραμματιστεί ρητά. Το κύριο πλεονέκτημα της χρήσης της μηχανικής μάθησης είναι ότι, μόλις ένας αλγόριθμος μάθει τι πρέπει να κάνει με τα δεδομένα, μπορεί να λειτουργεί αυτόματα.

Οι αλγόριθμοι μηχανικής μάθηση; χωρίζονται σε τρία είδη τεχνικών ανάλογα με τα προβλήματα που αντιμετωπίζουν:

- Επιβλεπόμενη Μάθηση(supervised learning)
- Μη Επιβλεπόμενη Μάθηση(unsupervised learning)
- Ενισχυτική Μάθηση(reinforcement learning)

Στην επιβλεπόμενη μάθηση, ο αλγόριθμος αυτός αποτελείται από μια μεταβλητή-στόχο/αποτέλεσμα (ή εξαρτημένη μεταβλητή) που πρέπει να προβλεφθεί από ένα δεδομένο σύνολο προβλεπτικών παραγόντων (ανεξάρτητες μεταβλητές). Χρησιμοποιώντας αυτό το σύνολο μεταβλητών, δημιουργούμε μια συνάρτηση που αντιστοιχίζει τις εισόδους στις επιθυμητές εξόδους. Η διαδικασία εκπαίδευσης συνεχίζεται έως ότου το μοντέλο επιτύχει ένα επιθυμητό επίπεδο ακρίβειας στα δεδομένα εκπαίδευσης. Υπάρχουν δύο κύριοι τύποι προβλημάτων επιβλεπόμενης μάθησης: είναι η ταξινόμηση που περιλαμβάνει την πρόβλεψη μιας κλάσης με ετικέτα, και η παλινδρόμηση που περιλαμβάνει την πρόβλεψη αριθμητική τιμή. Παραδείγματα επιβλεπόμενης μάθησης: Δένδρο απόφασης, τυχαίο δάσος, μηχανές διανυσμάτων υποστήριξης.

Η μάθηση χωρίς επίβλεψη, γνωστή και ως μη επιβλεπόμενη μηχανική μάθηση, χρησιμοποιεί αλγορίθμους μηχανικής μάθησης για την ανάλυση και ομαδοποίηση μη χαρακτηρισμένων συνόλων δεδομένων. Αυτοί οι αλγόριθμοι ανακαλύπτουν κρυμμένα μοτίβα ή ομαδοποιήσεις δεδομένων χωρίς την ανάγκη ανθρώπινης παρέμβασης. Οι αλγόριθμοι μη επιβλεπόμενης μάθησης περιλαμβάνουν προβλήματα συσχέτισης, τα οποία προσπαθούν να περιγράψουν τμήματα των δεδομένων, και προβλήματα ομαδοποίησης, τα οποία επιδιώκουν να εντοπίσουν ομαδοποιήσεις.

Η ενισχυτική μάθηση (RL) είναι ένας τύπος τεχνικής μηχανικής μάθησης που επιτρέπει σε έναν πράκτορα να μαθαίνει σε ένα διαδραστικό περιβάλλον με δοκιμή και σφάλμα χρησιμοποιώντας ανατροφοδότηση από τις δικές του ενέργειες και εμπειρίες.

Τα μοντέλα μηχανικής μάθησης που μελετώνται και εφαρμόζονται στην εργασία ανήκουν στην επιβλεπόμενη μάθηση.

2.6.2 Παλινδρόμηση τυχαίο δάσος(RFR)

Τα τυχαία δάση είναι ένας τύπος αλγορίθμου μηχανικής μάθησης που χρησιμοποιείται για εργασίες ταξινόμησης και παλινδρόμησης. Ένα μοντέλο ταξινόμησης λαμβάνει δεδομένα εισόδου και τα κατατάσσει σε μία από διάφορες κατηγορίες. Για παράδειγμα, δεδομένου ενός συνόλου εικόνων που αποτελείται από εικόνες σκύλων και γατών, ένας ταξινομητής θα μπορούσε να χρησιμοποιηθεί για να προβλέψει αν κάθε εικόνα είναι σκύλος ή γάτα. Με λίγα λόγια, ένας αλγόριθμος τυχαίου δάσους λειτουργεί με τη δημιουργία πολλαπλών δέντρων απόφασης, καθένα από τα οποία βασίζεται σε ένα τυχαίο υποσύνολο των δεδομένων. Τα δέντρα αποφάσεων είναι ένας τύπος αλγορίθμου που κάνει προβλέψεις εξετάζοντας τις εισόδους δεδομένων και καθορίζοντας σε ποια κατηγορία ανήκουν. Τα τυχαία δάση προχωρούν ένα βήμα παραπέρα δημιουργώντας πολλαπλά δέντρα αποφάσεων και στη συνέχεια υπολογίζοντας τον μέσο όρο των αποτελεσμάτων τους. Αυτό συμβάλλει στη μείωση της πιθανότητας υπερπροσαρμογής, δηλαδή όταν ο αλγόριθμος λειτουργεί καλά μόνο στα δεδομένα εκπαίδευσης και όχι στα νέα δεδομένα. Τα τυχαία δάση είναι ένα ισχυρό εργαλείο για τη μηχανική μάθηση και μπορούν να χρησιμοποιηθούν για ποικίλες εργασίες, όπως η αναγνώριση προσώπου, η ανίχνευση απάτης, η πρόβλεψη της συμπεριφοράς των καταναλωτών και οι προβλέψεις για το χρηματιστήριο.

Ο αλγόριθμος Random Forest(Τυχαίο δάσος-RF) ενσωματώνει δέντρα ταξινόμησης και παλινδρόμησης. Κάθε δέντρο δημιουργείται με τη χρήση τυχαίων διανυσμάτων. Για το

μοντέλο ταξινομητή με βάση το τυχαίο δάσος, οι βασικές παράμετροι είναι ο αριθμός των δέντρων απόφασης, καθώς και ο αριθμός των χαρακτηριστικών στο τυχαίο υποσύνολο σε κάθε κόμβο του δέντρου. Κατά την διάρκεια της εκπαίδευσης του μοντέλου, καθορίζεται πρώτα ο αριθμός των δέντρων, ένα μεγάλος αριθμός είναι καλή επιλογή, αλλά απαιτεί περισσότερο χρόνο για τον υπολογισμό. Ένας μικρότερος αριθμός χαρακτηριστικών οδηγεί σε μεγαλύτερη μείωση της διακύμανσης, αλλά σε μεγάλη αύξηση της μεροληψίας. Ο αριθμός των χαρακτηριστικών μπορεί να οριστεί από τη παρακάτω σχέση:

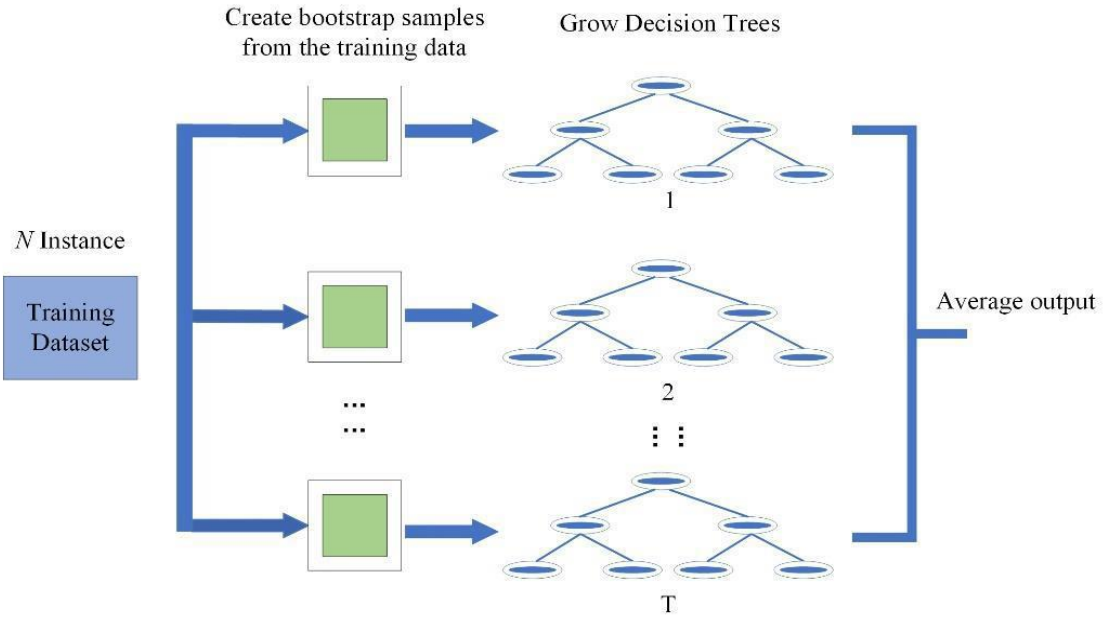
$N_F = \sqrt{M}$, όπου M είναι ο συνολικός αριθμός των χαρακτηριστικών.

Το μοντέλο παλινδρόμησης τυχαίου δάσους παρουσιάζεται στο σχήμα 2.

Υποθέτοντας ότι το μοντέλο περιλαμβάνει T δέντρα παλινδρόμησης για την πρόβλεψη, η τελική έξοδος παλινδρόμησης του μοντέλου δίνεται από την παρακάτω εξίσωση:

$$H_{(\chi)} = \frac{1}{T}\sum_{i=1}^{T} h_i(x)$$

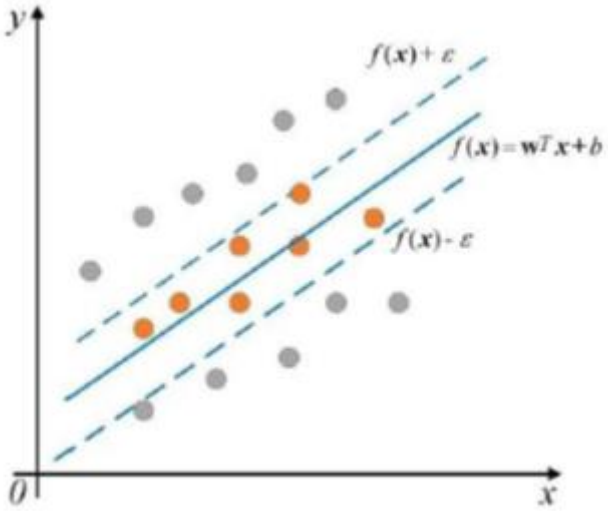
όπου T είναι ο αριθμός των δέντρων παλινδρόμησης, $h_i(x)$ είναι η έξοδος του i-οστού δέντρου παλινδρόμησης ((h_i) στο δείγμα x. Από τα πρ4απάνω προκύπτει, πρόβλεψη του Τυχαίου Δάσους είναι ο μέσος όρος των προβλεπόμενων τιμών όλων των δέντρων.



Σχήμα 1: Διάγραμμα παλινδρόμησης τυχαίου δάσους

2.6.3 Μηχανές Διανυσμάτων Παλινδρόμησης (SVR)

Οι μηχανές διανυσμάτων υποστήριξης (SVM) εισήχθησαν στο [8], για προβλήματα ταξινόμησης. Ο στόχος είναι να αναζητηθεί το βέλτιστο διαχωριστικό υπερεπίπεδο μεταξύ των κλάσεων. Τα σημεία που βρίσκονται στα όρια των κλάσεων ονομάζονται διανύσματα υποστήριξης, και ο ενδιάμεσος χώρος ονομάζεται υπερεπίπεδο- όταν ένας γραμμικός διαχωριστής δεν είναι σε θέση να ανακαλύψει λύση, τα σημεία δεδομένων προβάλλονται σε έναν χώρο υψηλότερων διαστάσεων, όπου τα προηγούμενα μη γραμμικά διαχωρίσιμα σημεία γίνονται γραμμικά διαχωρίσιμα, χρησιμοποιώντας συναρτήσεις πυρήνα. Η όλη εργασία μπορεί να διατυπωθεί ως τετραγωνικό πρόβλημα βελτιστοποίησης που μπορεί να επιλυθεί με ακριβείς τεχνικές. Στη συνέχεια αναπτύχθηκε μια εναλλακτική συνάρτηση απωλειών, η οποία επέτρεψε επίσης την εφαρμογή του SVM σε προβλήματα παλινδρόμησης. Η παλινδρόμηση διανυσμάτων υποστήριξης (SVR) έχει εφαρμοστεί στον τομέα της πρόβλεψης χρονοσειρών, με πολύ καλά αποτελέσματα. Οι μηχανές διανυσμάτων υποστήριξης (SVM) είναι μοντέλα μάθησης με επίβλεψη, που χρησιμοποιούνται σε προβλήματα ταξινόμησης και παλινδρόμησης. Στον τύπο παλινδρόμησης SVR, το σύνολο των δεδομένων εκπαίδευσης περιλαμβάνει μεταβλητές πρόβλεψης και παρατηρούμενες τιμές απόκρισης. Ο στόχος είναι να βρεθεί μια συνάρτηση f(x) που αποκλίνει από το yn (ετικέτες δείγματος) κατά μια τιμή no μεγαλύτερη από " (μεροληψία) για κάθε σημείο εκπαίδευσης x - δηλαδή να παραμείνει όσο το δυνατόν πιο επίπεδη. Επομένως, η SVR είναι επίσης γνωστή ως παλινδρόμηση σωλήνων. Το σχηματικό της διάγραμμα παρουσιάζεται στο Σχήμα 1.



Σχήμα2: Διάγραμμα παλινδρόμησης διανυσμάτων υποστήριξης

Η εξίσωση που περιγράφη την γραμμική παλινδρόμηση για το SVM είναι η εξής:

$$f(x) = \sum_{n=1}^{N} (a_n - a_n^*)(x_n^T x) + b. \qquad (1)$$

Όπου x είναι το διάνυσμα χαρακτηριστικών εισόδου, b είναι η παράμετρος απόστασης, an και a n είναι οι εισαγόμενες πολλαπλασιαστές Lagrange. Ωστόσο, ορισμένα προβλήματα παλινδρόμησης δεν μπορούν να περιγραφούν επαρκώς με τη χρήση ενός γραμμικό μοντέλο. Σε αυτή την περίπτωση, μπορούμε να λάβουμε ένα μη γραμμικό μοντέλο SVR αντικαθιστώντας το γινόμενο τελείας x με μια μη γραμμική συνάρτηση πυρήνα K(x1, x2) = <φ(x1), φ(x2)>, όπου φ(x) είναι ένας μετασχηματισμός που απεικονίζει το x σε ένα χώρο υψηλής διάστασης. Επομένως, η τελική λύση για τη μη γραμμική SVR μπορεί να προκύψει ως εξής:

$$f(x) = \sum_{n=1}^{N} (a_n - a_n^*)K(x, x_n) + b. \qquad (2)$$

2.6.4 Νευρωνικά δίκτυα

Τα νευρωνικά δίκτυα αποτελούνται από ένα νευρωνικό δίκτυο, είναι ένας παράλληλος επεξεργαστής με κατανεμημένη αρχιτεκτονική που αποτελείται από απλές μονάδες επεξεργασίας και είναι ικανός να αποθηκεύει γνώση και να την καθιστά διαθέσιμη για χρήση. Αυτό ταιριάζει με τον ανθρώπινο εγκέφαλο σε δύο σημεία: α) το δίκτυο παίρνει γνώση από το περιβάλλον, β) μέσω της μάθησης και η δύναμη των συνδέσεων μεταξύ των νευρώνων, που είναι το συναπτικό βάρος, χρησιμοποιείται για τη διατήρηση της αποκτηθείσας γνώσης. Το εύρος τιμών του πλάτους εξόδου του νευρώνα γράφεται ως το κλειστό διάστημα [0,1] ή [-1,1] Το μοντέλο του νευρώνα στο σχήμα περιλαμβάνει επίσης μια εξωτερικά εφαρμοσμένη μεροληψία που συμβολίζεται ως . Ως αποτέλεσμα της απόκλισης, η συνάρτηση ενεργοποίησης αυξάνεται ή μειώνεται ανάλογα με το αν είναι θετική ή αρνητική. Περιγραφή του νευρώνα k που αντιπροσωπεύεται στο προηγούμενο σχήμα με τις ακόλουθες εξισώσεις:

$$u_k = \sum_{j=1}^{m} w_{kj} x_j$$

και

$$Y_\kappa = \frac{\varphi}{(u_k + b_k)}$$

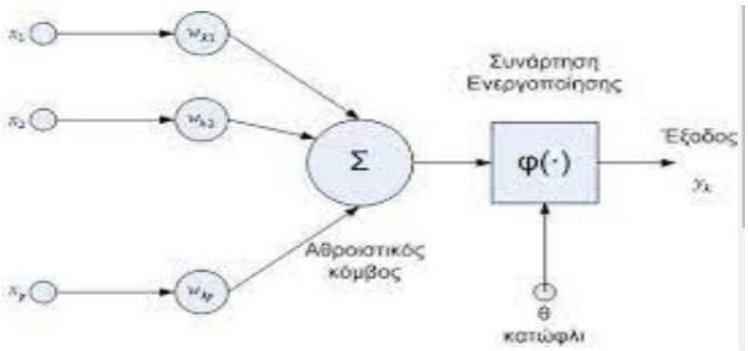
Όπου $x_1, x_2, \dots \dots \dots, x_k$ είναι τα σήματα εισόδου $w_{k1}, w_{k2}, \dots \dots \dots, w_{km}$ είναι αντίστοιχα τα συναπτικά βάρη του νευρώνα k, u_k είναι η έξοδος του γραμμικού συνδυαστεί που οφείλεται

στα σήματα εισόδου, b_k είναι η πόλωση $\varphi(\cdot)$ είναι η συνάρτηση ενεργοποίησης, και y_k είναι το σήμα εξόδου του νευρώνα.

Η σιγμοειδής συνάρτηση, της οποίας το γράφημα μοιάζει με <<S>> είναι η πιο κοινή μορφή συνάρτησης ενεργοποίησης που χρησιμοποιείται στα νευρωνικά δίκτυα. Ορίζεται ως μια αύξουσα συνάρτηση που παρουσιάζει μια ισορροπία μεταξύ γραμμικής και μη γραμμικής συμπεριφοράς. Ένα παράδειγμα σιγμοειδούς συνάρτησης είναι η λογιστική συνάρτηση, η οποία ορίζεται ως εξής: 1

$$\varphi(u) = \frac{1}{1 + \exp(-au)}$$

όπου a είναι η παράμετρος κλίσης της σιγμοειδούς συνάρτησης. Ενώ η συνάρτηση κατωφλίου έχει τιμή μηδέν ή ένα, η σιγμοειδής συνάρτηση μπορεί να πάρει τιμές από ένα συνεχές πεδίο τιμών από μηδέν έως ένα (1). Η σιγμοειδής συνάρτηση είναι διαφορίσιμη, ενώ η συνάρτηση κατωφλίου δεν είναι.



Σχήμα 3: Μη γραμμικό μοντέλο τεχνητού νευρώνα

2.6.5 Πολυστρωματικά Perceptrons(MLP)

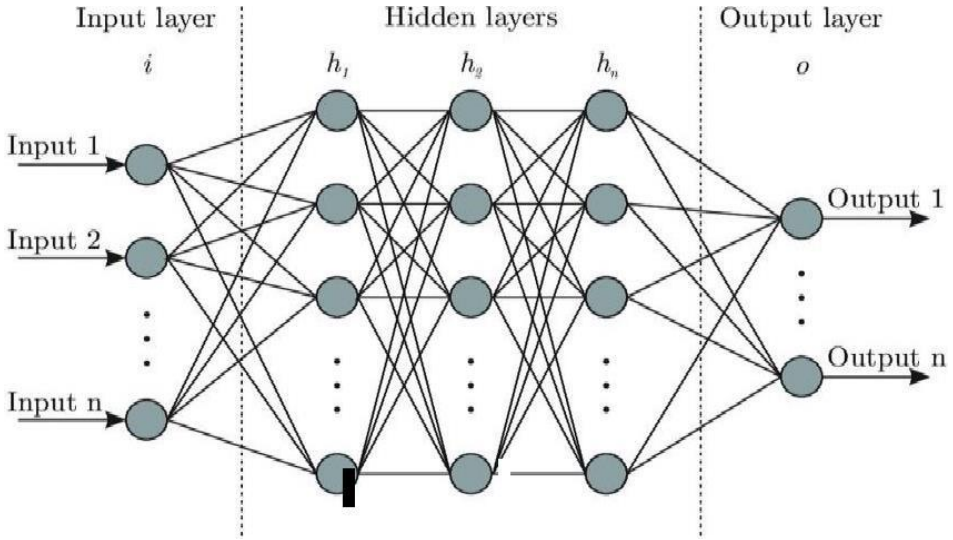
Γενικά, τα νευρωνικά δίκτυα όπως τα Multilayer Perceptrons ή MLPs παρέχουν δυνατότητες που προσφέρονται από λίγους αλγορίθμους, όπως ανθεκτικότητα στο θόρυβο. Τα νευρωνικά δίκτυα είναι ανθεκτικά στο θόρυβο στα δεδομένα εισόδου και στη συνάρτηση απεικόνισης και μπορούν να υποστηρίξουν τη μάθηση και την πρόβλεψη ακόμη και με την παρουσία ελλιπών τιμών. Τα νευρωνικά δίκτυα δεν κάνουν ισχυρές υποθέσεις σχετικά με τη συνάρτηση χαρτογράφησης και μαθαίνουν εύκολα γραμμικές και μη γραμμικές σχέσεις. Μπορεί να καθοριστεί ένας αυθαίρετος αριθμός χαρακτηριστικών εισόδου, παρέχοντας άμεση υποστήριξη για πολυμεταβλητές προβλέψεις. Επίσης μπορεί να

καθοριστεί ένας αυθαίρετος αριθμός τιμών εξόδου, παρέχοντας άμεση υποστήριξη για προβλέψεις πολλαπλών βημάτων, ακόμη και για προβλέψεις πολλαπλών μεταβλητών. Για αυτές τις δυνατότητες και μόνο, τα νευρωνικά δίκτυα τροφοδότησης μπορεί να είναι χρήσιμα για την πρόβλεψη χρονοσειρών. Το πολυστρωματικό perceptron (MLP) είναι ένα εξελιγμένο τεχνητό νευρωνικό δίκτυο. Αποτελείται από πολλαπλά perceptrons όπως στο σχήμα 4. Αποτελούνται από ένα στρώμα εισόδου που λαμβάνει το σήμα, ένα στρώμα εξόδου που λαμβάνει μια απόφαση ή πρόβλεψη σχετικά με την είσοδο και έναν αυθαίρετο αριθμό κρυφών στρωμάτων που χρησιμεύουν ως η πραγματική υπολογιστική μηχανή του MLP[9]. Τα MLP με ένα μόνο κρυφό στρώμα μπορούν να προσεγγίσουν οποιαδήποτε συνεχή συνάρτηση.

Το perceptron είναι ένας γραμμικός ταξινομητής, δηλαδή ένας αλγόριθμος που ταξινομεί την είσοδο χρησιμοποιώντας μια ευθεία γραμμή για να χωρίσει δύο κατηγορίες[9]. Ένα διάνυσμα χαρακτηριστικών x συνήθως πολλαπλασιάζεται με βάρη w και προστίθεται σε ένα bias b: y = w * x + b.

$$y = {}^{\varphi(\Sigma^{n}}{}_{i=1}\, wi_{\,xi} + b) \qquad (1)$$

όπου w είναι το διάνυσμα των βαρών, x είναι το διάνυσμα των εισόδων, b είναι η προκατάληψη και phi είναι η μη γραμμική συνάρτηση ενεργοποίησης.



Σχήμα 4: αρχιτεκτονική πολυστρωματικού νευρωνικού δικτύου perceptron.

2.6.6 Δίκτυα Μακράς Βραχείας Μνήμης(LSTM)

Το LSTM είναι ένας τύπος επαναλαμβανόμενου νευρωνικού δικτύου, σχεδιασμένος για να αντιμετωπίζει τη μακροπρόθεσμη μάθηση εξάρτησης. Μια μονάδα LSTM αποτελείται από ένα κύτταρο, μια πύλη εισόδου και μια πύλη εξόδου και μια πύλη λήθης. Το πλεονέκτημα αυτής της αρχιτεκτονικής είναι η δυνατότητα να αφήνει το δίκτυο να μαθαίνει πότε να εφαρμόζει ένα ευρύτερο πλαίσιο και στη συνέχεια να καθορίζει πότε να βασίζεται στη μακροπρόθεσμη ή βραχυπρόθεσμη μνήμη, καθιστώντας δυνατή την εκμάθηση πιο σύνθετων λειτουργιών. Τα δίκτυα μακράς βραχίας μνήμης ή LSTM προσθέτουν τον ρητό χειρισμό της σειράς μεταξύ των παρατηρήσεων κατά τη μάθηση μιας συνάρτησης απεικόνισης από τις εισόδους στις εξόδους, που δεν προσφέρεται από τα MLP ή τα CNN. Είναι ένας τύπος νευρωνικού δικτύου που προσθέτει εγγενή υποστήριξη για δεδομένα εισόδου που αποτελούνται από ακολουθίες παρατηρήσεων. Τα επαναλαμβανόμενα νευρωνικά δίκτυα προσθέτουν άμεσα υποστήριξη για δεδομένα ακολουθιών εισόδου. Αυτή η ικανότητα των LSTM έχει χρησιμοποιηθεί με μεγάλη επιτυχία σε πολύπλοκα προβλήματα επεξεργασίας φυσικής γλώσσας, όπως η νευρωνική μηχανική μετάφραση, όπου το μοντέλο πρέπει να μάθει τις πολύπλοκες αλληλεπιδράσεις μεταξύ των λέξεων τόσο εντός μιας δεδομένης γλώσσας όσο και μεταξύ των γλωσσών κατά τη μετάφραση από τη μία γλώσσα στην άλλη. Το πιο σχετικό πλαίσιο των παρατηρήσεων εισόδου για την αναμενόμενη έξοδο μαθαίνεται και μπορεί να αλλάζει δυναμικά. Το μοντέλο μαθαίνει τόσο μια αντιστοίχιση από τις εισόδους στις εξόδους όσο μαθαίνει ποιο πλαίσιο από την ακολουθία εισόδου είναι χρήσιμο για την αντιστοίχιση και μπορεί να αλλάξει δυναμικά αυτό το πλαίσιο ανάλογα με τις ανάγκες.

2.6.7 Συνελικτικά Νευρωνικά Δίκτυα(CNN)

Πρόκειται για μια υποκατηγορία των τεχνικών νευρωνικών δικτύων που έχει αποδειχθεί πολύ αποτελεσματικό και έχει μεγάλη επιτυχία στην αναγνώριση προσώπων, αντικειμένων και οδικών πινακίδων, ενώ ταυτόχρονα χρησιμοποιείται σε ρομπότ και αυτόνομα οχήματα. Αυτά τα δίκτυα αποτελούνται από νευρώνες με τα ίδια βάρη και συναπτικές εισόδους με τα απλά νευρωνικά δίκτυα. Η διαφορά των συνελικτικών δικτύων σε σύγκριση με τα απλά δίκτυα οφείλεται στη δομή τους και στον τρόπο με τον οποίο προωθούν πληροφορίες. Η ικανότητα των CNNs να μαθαίνουν και να εξάγουν αυτόματα χαρακτηριστικά από ακατέργαστα δεδομένα εισόδου μπορεί να εφαρμοστεί σε προβλήματα πρόβλεψης χρονοσειρών. Μια ακολουθία παρατηρήσεων μπορεί να αντιμετωπιστεί σαν μια

μονοδιάστατη εικόνα την οποία ένα μοντέλο CNN μπορεί να διαβάσει και να απομονώσει στα πιο σημαντικά στοιχεία. Επίσης, γίνεται εκμάθηση χαρακτηριστικών, αυτόματος εντοπισμός, εξαγωγή των σημαντικότερων χαρακτηριστικών από τα ακατέργαστα δεδομένα εισόδου που αφορούν άμεσα το πρόβλημα πρόβλεψης που μοντελοποιείται. Τα CNN αποκτούν τα πλεονεκτήματα των Multilayer Perceptrons για την πρόβλεψη χρονοσειρών, δηλαδή την υποστήριξη πολυμεταβλητών εισόδων, πολυμεταβλητών εξόδων και την εκμάθηση αυθαίρετων αλλά πολύπλοκων λειτουργικών σχέσεων, αλλά δεν απαιτούν από το μοντέλο να μαθαίνει απευθείας από τις παρατηρήσεις υστέρησης. Αντ' αυτού, το μοντέλο μπορεί να μάθει μια αναπαράσταση από μια μεγάλη ακολουθία εισόδου που είναι πιο σχετική για το πρόβλημα πρόβλεψης.

Βιβλιογραφική επισκόπηση

Οι παραδοσιακές τεχνικές που βασίζονται στις πιθανότητες και τη στατιστική είναι πολύ περίπλοκες και λιγότερο αποτελεσματικές. Τα μοντέλα πρόβλεψης ποιότητας αέρα(AQI) που βασίζονται στην Μηχανική Μάθηση (ML) έχουν αποδειχθεί πιο αξιόπιστα και συνεπή. Οι προηγμένες τεχνολογίες και οι αισθητήρες έκαναν τη συλλογή δεδομένων πιο εύκολη και ακριβή. Οι ακριβείς και αξιόπιστες προβλέψεις μέσω τέτοιων τεράστιων περιβαλλοντικών δεδομένων απαιτούν αυστηρή ανάλυση την οποία μόνο οι αλγόριθμοι μηχανικής μάθησης (ML) μπορούν να αντιμετωπίσουν αποτελεσματικά. Ειδικότερα, οι αλγόριθμοι ML έχουν χρησιμοποιηθεί ευρέως για την πρόβλεψη της ποιότητας του αέρα. Λόγω των υψηλών μη γραμμικών διεργασιών που αφορούν τις συγκεντρώσεις των ρύπων και της μερικώς γνωστής δυναμικής τους, είναι πολύ δύσκολο να παραχθεί ένα μοντέλο ικανό να προβλέψει τέτοιου είδους γεγονότα [26]. Τα μοντέλα μηχανικής μάθησης(ML) είναι ένα παράδειγμα μη παραμετρικών και μη γραμμικών μοντέλων που αξιοποιούν μόνο τις ιστορικές πληροφορίες για να μάθουν την κρυφή σχέση μεταξύ των δεδομένων [27]. Σε γενικές γραμμές, οι προσεγγίσεις ML, όπως τα τεχνητά νευρωνικά δίκτυα (ANN), Βαθιά Μάθηση, και οι μηχανές διανυσμάτων υποστήριξης (SVM), έχει αποδειχθεί ότι υπερτερούν των παραδοσιακών τεχνικών όταν προβλέπουν χρόνο-σειρές (TS) με υψηλό επίπεδο μη γραμμικότητας.

3.1 Σχετικές εργασίες

D´ıaz-Robles et al. [28] πραγματοποίησαν μια εμπειρική μελέτη με την εφαρμογή ενός υβριδικού μοντέλου που χρησιμοποιεί ANNs και ARIMA για την πρόβλεψη της ποιότητας του αέρα στη Χιλή, και συγκεκριμένα των μετρήσεων P10. Τα μοντέλα συνδυάστηκαν για να συλλάβουν τα διαφορετικά πρότυπα μέσα στα δεδομένα: το μοντέλο ARIMA για να συλλάβει τη γραμμικότητα του συνόλου δεδομένων και τα ANNs για να συλλάβουν τη μη γραμμικότητα από τα κατάλοιπα του μοντέλου ARIMA. Οι συγγραφείς κατέληξαν στο συμπέρασμα ότι το μοντέλο που προκύπτει έχει υψηλή ικανότητα γενίκευσης και υπερτερεί τόσο του ARIMA όσο και των ANNs που χρησιμοποιούνται μεμονωμένα.

Οι S. S. Ganesh et al. [1] παρουσίασαν μοντέλα πολλαπλής παλινδρόμησης για την πρόβλεψη του AQI [4]. Εφάρμοσαν μοντέλα παλινδρόμησης όπως το SVR και το μοντέλο πολλαπλής γραμμικής παλινδρόμησης για την πρόβλεψη του δείκτη ποιότητας του

αέρα(AQI). Μεταξύ αυτών, το SVR παρουσίασε υψηλό επίπεδο απόδοσης. Για την αξιολόγηση της απόδοσης των μοντέλων παλινδρόμησης εξέτασαν κριτήρια αξιολόγησης όπως MAE, Mean absolute percentange error (MAPE), Correlation coefficient (R), RMSE και Index of agreement (IA).

Οι Huixiang Liu et al. [2] παρουσίασαν μια ερευνητική εργασία στην οποία χρησιμοποιήθηκαν μοντέλα παλινδρόμησης για την πρόβλεψη της ποιότητας του αέρα [11]. Εφάρμοσαν μοντέλα μηχανικής μάθησης όπως το SVR και το Random forest regression (RFR) για την πρόβλεψη του AQI. Μεταξύ των δύο μοντέλων, το μοντέλο RFR είχε καλύτερη απόδοση από το μοντέλο SVR, επειδή η χρονική πολυπλοκότητα του SVR αυξήθηκε κυβικά με την αύξηση του αριθμού των δειγμάτων. Χρησιμοποίησαν μετρήσεις επιδόσεων όπως το RMSE, το συντελεστή.

Οι Mauro Castelli et al. [3] παρουσίασαν μια ερευνητική εργασία στην οποία χρησιμοποίησαν το SVR για την πρόβλεψη της ποιότητας του αέρα στην Καλιφόρνια [1]. Η μελέτη αφορά τη χρήση του SVR για τον ακριβή υπολογισμό του AQI και των συγκεντρώσεων των ρύπων. Η μελέτη παρήγαγε ένα εξαιρετικά κατάλληλο μοντέλο για την ωριαία πρόβλεψη του AQI, ακριβείς συγκεντρώσεις ρύπων όπως o3, co2 και so2, καθώς και την ωριαία ατμοσφαιρική ρύπανση στην περιοχή. Χρησιμοποίησαν το MAE, το κανονικοποιημένο μέσο τετραγωνικό σφάλμα (NMSE) και το RMSE ως μετρικές αξιολόγησης της απόδοσης. Οι Bing-chun liu et al. (2017) παρουσίασαν μια ερευνητική εργασία στην οποία χρησιμοποίησαν πολυδιάστατο συνεργατικό μοντέλο SVR [10]. Σε αυτή την εργασία τα δεδομένα που χρησιμοποιήθηκαν ήταν συνεργατικά δεδομένα ποιότητας αέρα πολλαπλών πόλεων. Καθώς τα δεδομένα που χρησιμοποιήθηκαν ήταν από περισσότερες από μία πόλεις, η εκπαίδευση είναι πολύπλοκη που οδηγεί σε αύξηση του χρόνου εκπαίδευσης. Οι τιμές RMSE για τα σύνολα δεδομένων εκπαίδευσης και δοκιμής ήταν μικρότερες από 12 και κατέληξαν στο συμπέρασμα ότι το SVR είναι ισχυρό και αποτελεί ένα εφαρμοσμένο μοντέλο για την πρόβλεψη του AQI. Χρησιμοποίησαν τις μετρικές επιδόσεων όπως το RMSE και το Mean ab- solute percentange error (MAPE). Η εργασία των Khaled et al.[5] επικεντρώνεται στο σύστημα παρακολούθησης και στην ενότητα πρόβλεψης. Διερευνώνται τρεις αλγόριθμοι μηχανικής μάθησης (ML) για τη δημιουργία ακριβών μοντέλων πρόβλεψης για ένα βήμα και πολλά βήματα μπροστά από τις συγκεντρώσεις του όζοντος σε επίπεδο εδάφους (O3), του διοξειδίου του αζώτου (NO2) και του διοξειδίου του θείου (SO2). Αυτοί οι αλγόριθμοι ML είναι οι μηχανές διανυσμάτων

υποστήριξης, τα δέντρα μοντέλων M5P και τα τεχνητά νευρωνικά δίκτυα (ANN). Επιδιώκονται δύο τύποι μοντελοποίησης: μονό-μεταβλητές και πολύ-μεταβλητές. Με βάση εκτεταμένα πειράματα, ο M5P υπερτερεί έναντι άλλων αλγορίθμων για όλα τα αέρια σε όλους τους ορίζοντες όσον αφορά το NRMSE και το PTA λόγω της αποτελεσματικότητας της δενδρικής δομής και της ισχυρής ικανότητας γενίκευσης. Από την άλλη πλευρά, ο ANN πέτυχε τα χειρότερα αποτελέσματα λόγω της κακής ικανότητας γενίκευσης όταν εκπαιδεύεται σε μικρό σύνολο δεδομένων με πολλά χαρακτηριστικά που οδηγεί σε ένα πολύπλοκο δίκτυο που υπέρ-προσαρμόζει τα δεδομένα, ενώ είχε τον SVM καλύτερο από τον ANN στην περίπτωσή αυτή λόγω της προσαρμοστικότητάς του με δεδομένα υψηλών διαστάσεων. Τα μέτρα που χρησιμοποιούνται για την αξιολόγηση είναι η ακρίβεια της τάσης πρόβλεψης και το μέσο τετραγωνικό σφάλμα (RMSE). Τα αποτελέσματα δείχνουν ότι η χρήση διαφορετικών χαρακτηριστικών στην πολύ-μεταβλητή μοντελοποίηση με τον αλγόριθμο M5P αποδίδει τις καλύτερες επιδόσεις πρόβλεψης. Οι Freeman et al. [6] χρησιμοποιούν *επαναλαμβανόμενα νευρωνικά δίκτυα* (RNN) με LSTM για να προβλέψουν τις τοπικές 8ωρες μέσες συγκεντρώσεις όζοντος με βάση τις ωριαίες μετρήσεις των σταθμών παρακολούθησης του αέρα. Πραγματοποιήθηκε μια φάση προ-επεξεργασίας για τον υπολογισμό των ελλιπών δεδομένων, τον εντοπισμό ακραίων τιμών και την επιλογή χαρακτηριστικών με δέντρα απόφασης. Τα ωριαία ποιοτικά και μετεωρολογικά δεδομένα του αέρα συλλέχθηκαν με τη χρήση αναλυτών οπτικής φασματοσκοπίας απορρόφησης OPSIS differ- ential που τοποθετήθηκαν κοντά σε ένα τοπικό πανεπιστήμιο στο Κουβέιτ για την εκπαίδευση και την πρόβλεψη τιμών έως και τρεις ημέρες. Τα νευρωνικά δίκτυα feedforward και ARIMA συγκρίθηκαν με την προτεινόμενη μέθοδο χρησιμοποιώντας τα μέτρα απόδοσης RMSE και MAE. Στην εργασία των Kok κ.ά. [7] προτείνουν ένα μοντέλο βαθιάς μάθησης που βασίζεται σε δίκτυα *μακράς βραχίας μνήμης* (LSTM) προκειμένου να κάνουν προ-διαγραφές για την ατμοσφαιρική ρύπανση με τα δεδομένα από την ανάλυση έξυπνων πόλεων IoT. Η δομή του δικτύου αποτελείται από ένα στρώμα εισόδου, ένα κρυφό στρώμα με 24 μονάδες LSTM και ένα στρώμα εξόδου, με μέγεθος δέσμης 50 και 100 εποχών. Τα πειράματα χρησιμοποίησαν μια βάση δεδομένων με 17568 περιπτώσεις σε διαστήματα πέντε λεπτών και τα χαρακτηριστικά όζον, PM, μονοξείδιο του άνθρακα, διοξείδιο του θείου, διοξείδιο του αζώτου, γεωγραφικό μήκος, γεωγραφικό πλάτος και χρονοσφραγίδα. Το όζον και το διοξείδιο του αζώτου προβλέπονται και τα μοντέλα αξιολογούνται με τη χρήση hold-out στο 70% της εκπαίδευσης και στο 30% της δοκιμής και τα αποτελέσματα συγκρίνονται με το SVM χρησιμοποιώντας RMSE και μετρικές *μέσου*

απόλυτου σφάλματος (MAE). Στην εργασία Rishanti et al.[8], δοκιμάστηκαν μοντέλα μηχανικής μάθησης για την πρόβλεψη της ατμοσφαιρικής ρύπανσης στις έξυπνες πόλεις της Μαλαισίας λόγω της σοβαρότητας της ατμοσφαιρικής ρύπανσης στις αστικές περιοχές. Έτσι, το Random Forest έδωσε την υψηλότερη ακρίβεια στην πρόβλεψη των PM 2.5στις έξυπνες πόλεις της Μαλαισίας σε σχέση με το νευρωνικό δίκτυο πολλαπλών επιπέδων. Το Random Forest έχει ακρίβεια 97%, ενώ το νευρωνικό δίκτυο πολλαπλών επιπέδων έχει ακρίβεια 92%. Παρόλο που το Random Forest έδωσε καλύτερη ακρίβεια από το MLP, καθώς αυξάνεται ο αριθμός των νευρώνων στα κρυφά στρώματα, το MLP αυξάνει την ακρίβειά του στην πρόβλεψη των τιμών. Από την άλλη πλευρά, καθώς αυξάνεται ο αριθμός των δέντρων απόφασης στο Random Forest, η ακρίβεια μειώνεται. Στην εργασία, των Nath et al. [9] συγκρίνουν στατιστικές (auto-regressive, Holt- Winters, εποχιακή ARIMA και Prophet) και βαθιές μεθόδους μάθησης (LSTM, LSTM auto-encoder, Bi-LSTM, convolution LSTM) για την προ- dict PM2.5 και τις 10 συγκεντρώσεις PM τους επόμενους μήνες. Τα δεδομένα ελήφθησαν μεταξύ 2016 και 2020 από έναν σταθμό στο Victoria Memorial Hall στην Καλκούτα της Ινδίας. Το στατιστικό μοντέλο Holt-Winters είχε καλύτερες επιδόσεις για τα RMSE και MAE από τα μοντέλα βαθιάς μάθησης. Bihter Das et al.[10]: Στην παρούσα μελέτη, συγκρίθηκαν τα μοντέλα MLP, RNN και LSTM για την πρόβλεψη ατμοσφαιρικών ρύπων όπως τα PM10 και SO2 , προκειμένου να αναπτυχθεί μια ακριβής, αποτελεσματική και εύκολη τεχνική. Τα πειραματικά αποτελέσματα έδειξαν ότι το LSTM υπερείχε καλύτερα σε σύγκριση με τα μοντέλα MLP και RNN. Το LSTM εκτιμά τους ατμοσφαιρικούς ρύπους PM10 και SO2 πιο κοντά στις πραγματικές τιμές. Η μελέτη αυτή επιβεβαιώνει ότι τα μοντέλα βαθιάς μάθησης μπορούν να χρησιμοποιηθούν πλήρως για την πρόβλεψη των πιο σημαντικών ατμοσφαιρικών ρύπων που επηρεάζουν περισσότερο την ανθρώπινη υγεία και το περιβάλλον. Ο Patra [11] χρησιμοποιεί μοντέλα multi-layer perceptron (MLP), support vector regression (SVR) και autoregressive integrated moving average (ARIMA) για πρόβλεψη ενός μήνα για το CO και το NO 2με τη δημόσια βάση δεδομένων AirQuality που λαμβάνεται από το UCI Machine Learning Repository [15] με 390 περιπτώσεις ημερήσιων μέσων αποκρίσεων από μια συλλογή 5 χημικών αισθητήρων οξειδίων μετάλλων που είναι εγκατεστημένοι σε μια συσκευή πολλαπλών χημικών αισθητήρων για την ποιότητα του αέρα. Ο συγγραφέας καταλήγει στο συμπέρασμα ότι τα καλύτερα αποτελέσματα, που παρουσιάζονται σε όρους μέσου τετραγωνικού σφάλματος (RMSE), επιτυγχάνονται με MLP με αρχιτεκτονική (4 8 1) για το CO και με αρχιτεκτονική (10 2 1) για το NO 2. Οι Kaya et al.[12] εκτιμούν το PM10 με deep flexible sequential

(DFS), ένα υβριδικό βαθιάς μάθησης μοντέλο που περιλαμβάνει LSTM και CNN, με τις μετρικές MAE και RMSE για 4, 12 και 24 μεγέθη παραθύρων σε τέσσερις ξεχωριστούς σταθμούς μέτρησης της Κωνσταντινούπολης, Τουρκία. Οι Cai et al. [29] συνέκριναν τα αποτελέσματα που προέκυψαν με τη χρήση ενός πολυγραμμικού μοντέλου παλινδρόμησης με αυτά που επιτεύχθηκαν από ένα ANN κατά την πρόβλεψη της ωριαίας συγκέντρωσης ατμοσφαιρικών ρύπων, καταλήγοντας στο συμπέρασμα ότι τα ANN παράγουν πιο αξιόπιστα αποτελέσματα. Οι Madhuri et al. [30] ανέφεραν ότι η ταχύτητα του ανέμου, η κατεύθυνση του ανέμου, η υγρασία και η θερμοκρασία έπαιζαν σημαντικό ρόλο στη συγκέντρωση των ατμοσφαιρικών ρύπων. Οι συγγραφείς χρησιμοποίησαν τεχνικές ML με επίβλεψη για την πρόβλεψη του AQI και διαπίστωσαν ότι ο αλγόριθμος RF παρουσίασε τα λιγότερα σφάλματα ταξινόμησης.

Μεθοδολογία

Στην παρούσα ενότητα περιγράφεται η προτεινόμενη μεθοδολογία για την κατασκευή μοντέλων πρόβλεψης χρονοσειρών για τις συγκεντρώσεις *PM 2.5*. PM 10 και AQI, η σύγκριση των μοντέλων και η επιλογή του καλύτερου μοντέλου λαμβάνοντας υπόψη τις προβλέψεις επόμενης ώρας, και 24h. Ο κύριος στόχος της προτεινόμενης μεθοδολογίας είναι η σύγκριση διαφορετικών αρχιτεκτονικών βαθιάς μάθησης μεταξύ τους και με πιο συμβατικές- στατιστικές μεθόδους μηχανικής μάθησης.

4.1 Περιγραγή δεδομένων

Το σύνολο των δεδομένων προέρχονται από την ανοικτή βάση δεδομένων https://openaq.org/#/locations?page=1&countries=GR..Αποτελείται από ωριαίες μετρήσεις ρύπανσης για περίοδο από τον Απρίλιο του 2020 και ολόκληρο το 2021 με κέντρο λήψης την περιοχή GR0048A σε αστικό περιβάλλον στην Ελλάδα. Το σύνολο δεδομένων αποτελείται από τις μετρήσεις αιωρούμενων σωματιδίων.

Τα δεδομένα που συλλέχθηκαν αφορούσαν επίπεδα ρύπων και περιλαμβάνουν την ημερομηνία-ώρα, τη συγκέντρωση αιωρούμενων σωματιδίων(PM 2.5 και PM 10), πληροφορίες για το γεωγραφικό σημείο με συντεταγμένες, καθώς και των κωδικό της περιοχής. Τα δεδομένα που συλλέχθηκαν αποτελούνται από τα ακόλουθα χαρακτηριστικά: locationId(Κωδικός τοποθεσίας), location(τοποθεσία) DateTime(, ημερομηνία- ώρα λήψης δεδομένων), PM 2.5(συγκέντρωση PM 2.5), PM 10(συγκέντρωση PM 10), latitude(γεωγραφικό πλάτος), longitude(γεωγραφικό μήκος). Εφαρμόζουμε τη μέθοδο αντικατάστασης μέσων όρων η οποία αντικαθιστά όλες τις ελλιπείς τιμές για αριθμητικά χαρακτηριστικά σε ένα σύνολο δεδομένων με τους μέσους όρους από τα δεδομένα εκπαίδευσης, και τις τιμές NAN με μηδέν.

4.2 Προ-επεξεργασία δεδομέων

Στη διαδικασία αυτή, τα δεδομένα καθαρίζονται, όπως με την αφαίρεση ακραίων τιμών και ανωμαλιών. *Τα* δεδομένα προετοιμάζονται επίσης ώστε να είναι σε κατάλληλη μορφή για τη χρήση από τους αλγορίθμους μηχανικής μάθησης. Η ποιότητα των δεδομένων και η αντιπροσωπευτικότητά τους είναι τα πρώτα και κυριότερα σημεία που εγγυώνται την επιτυχή δημιουργία μοντέλων πρόβλεψης. Το στάδιο της προ-επεξεργασίας των δεδομένων συχνά επηρεάζει την ικανότητα γενίκευσης ενός αλγορίθμου μηχανικής μάθησης. Η προ-επεξεργασία

δεδομένων περιλαμβάνει συνήθως τα ακόλουθα βήματα: υπολογισμό ελλιπών δεδομένων, αφαίρεση η τροποποίηση ακραίων τιμών, και μετασχηματισμό δεδομένων(κανονικοποίηση). Ενώ τα δύο πρώτα είναι χρήσιμα για να έχουμε πιο ακριβή και πλήρη σύνολα δεδομένων, το τρίτο χρησιμοποιείται για να έχουμε ομοιόμορφα κατανεμημένα δεδομένα και να ελαχιστοποιήσουμε την μεταβλητότητα.

Το αρχικό σύνολο δεδομένων δεν περιέχει μεταβλητές για τον υπολογισμό του AQI, οπότε υπολογίστηκε με χρήση τύπου. Για τον υπολογισμό τις τιμής του AQI, παίρνουμε τη μέση τιμή για τις συγκεντρώσεις των σωματιδίων PM 2.5 , PM 10 για περίοδο 24h(οι καταχωρήσεις είναι ανά ώρα στο σύνολο του dataset) σύμφωνα με τη σχέση 1 Κεφ.2. Η γενικευμένη μέθοδος αποτελείται τουλάχιστον από τρεις παράγοντες που συμβάλουν στον AQI, εκ των οποίων ο ένας πρέπει να είναι ένα επίπεδο PM(PM 2.5, PM10). Λόγω των περιορισμένων χαρακτηριστικών στην τοποθεσία δημιουργήσαμε, ο τύπος που δημιουργήσαμε για να υπολογίσουμε το επίπεδο AQI σε κάθε γραμμή του καινούργιου συνόλου δεδομένων. Άλλα χαρακτηριστικά όπως PM2.5_SubIndex, PM10_SubIndex ,Checks, AQI_calculated, AQI_bucket_calculated ήταν απαραίτητα για τον υπολογισμό Δείκτη Ποιότητας Αέρα και την προσθήκη τους σε νέο αρχείο AQI.CSV. το νέο πεδίο AQI_bucket_calculated αντιστοιχίζεται σε μια προκαθορισμένη κατάσταση από 1- 6, ανάλογα με τις τιμές του AQI. Το παρακάτω σχήμα δείχνει τις προκαθορισμένες καταστάσεις του AQI.

301 – 500	Hazardous
201 – 300	Very Unhealthy
151 – 200	Unhealthy
101 – 150	Unhealthy for Sensitive Groups
51 – 100	Moderate
0 – 50	Good

Εικόνα 4.1: Προκαθορισμένα επίπεδα AQI.

4.3 Επιλογή χαρακτηριστικών

Το βήμα αυτό αφορά την επιλογή των χαρακτηριστικών που θα συμπεριληφθούν στη διαδικασία πρόβλεψης μαζί με κάθε αέριο-στόχο, όπως η θερμοκρασία, η υγρασία και η ημέρα της εβδομάδας.

4.4 Διαχωρισμός συνόλου δεδομένων

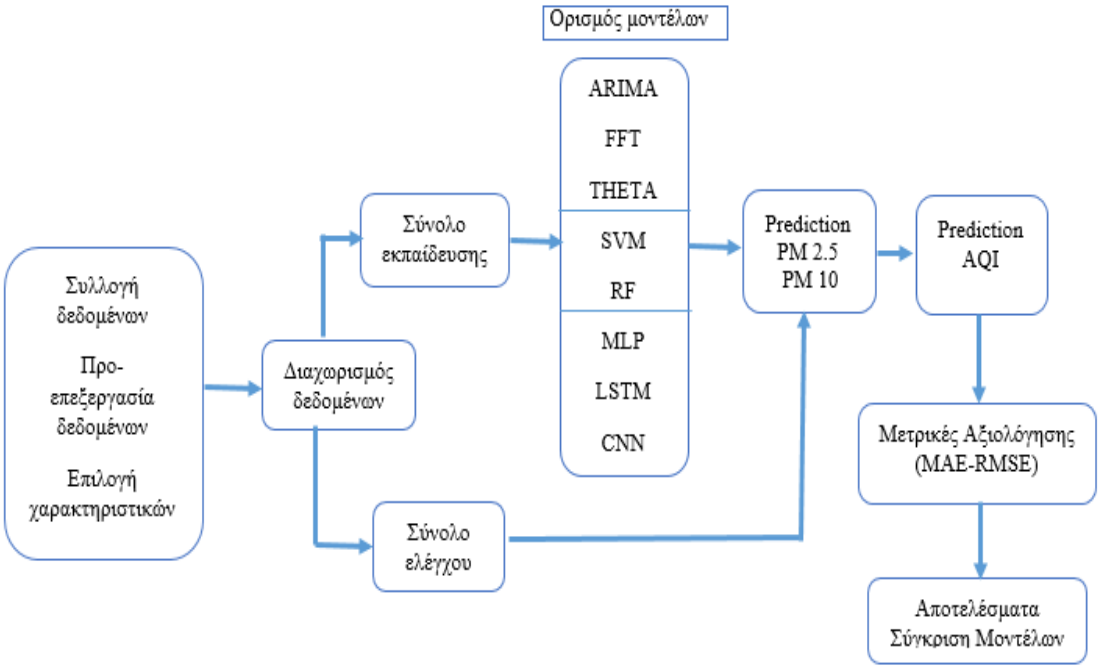
Για την αξιολόγηση της απόδοσης του μοντέλου, το σύνολο δεδομένων χωρίζεται σε δύο σύνολα δεδομένων εκπαίδευσης και δοκιμής. Επιλέγεται ένα τμήμα για εκπαίδευση που αντιστοιχεί στο 80% του αρχικού συνόλου δεδομένων, και το υπόλοιπο 20% για έλεγχο. Το σύνολο δεδομένων εκπαίδευσης προορίζεται για την εκπαίδευση του μοντέλου και αποτελείται από το πιο σημαντικό μέρος του συνόλου δεδομένων, χρησιμοποιείται για τον υπολογισμό των προβλεπόμενων τιμών. το σύνολο δεδομένων δοκιμής περιέχει ακατέργαστα δεδομένα για το εκπαιδευμένο μοντέλο, χρησιμοποιείται για την αξιολόγησης της ακρίβειας. Αυτός ο πρώτος διαχωρισμός γίνεται διατηρώντας τη χρόνο-σειρά συνεχή, οπότε το σύνολο ελέγχου είναι οι τελευταίες μετρήσεις. Στόχος είναι η γενίκευση του μέτρου απόδοσης από την εκμάθηση στο σύνολο ελέγχου σε προβλέψεις από ακατέργαστα δεδομένα. Κατά συνέπεια, τα δεδομένα μπορούν να χρησιμοποιηθούν με ασφάλεια για πρόβλεψη με ορίζοντα μιας και είκοσι τεσσάρων ωρών, και του δείκτη ποιότητας αέρα εφόσον η κατανομή παραμένει σταθερή(ιστορικά δεδομένα).

4.5 Μοντέλα πρόβλεψης

σε αυτό το βήμα, τα μοντέλα μαθαίνουν για να χρησιμοποιηθούν για να αποδώσουν μελλοντικές τιμές σε ένα χαρακτηριστικό-στόχο απαρατήρητων περιπτώσεων δεδομένων με βάση τις ιστορικές τιμές των διακριτικών χαρακτηριστικών.την παρούς αεργασία χρησιμοποιήσαμε τα εξής μοντέλα μηχανικής μάθησης με επίβλεψη: Random Forest, SVM, ARIMA,FFT(Fourier), Theta, MLP, LSTM.,CNN. Μετά την επιλογή του κατάλληλου μοντέλου και των παραμέτρων αυτού, ακολουθεί η αξιολόγηση της ακρίβειας της πρόβλεψης που προκύπτει με κριτήρια αξιολόγησης, όπως το Μέσο Τετραγωνικό Σφάλμα, και η Ρίζα Μέσου Τετραγωνικού Σφάλματος.

Η διαδικασία που περιλαμβάνει τα παραπάνω βήματα για την κατασκευή και εφαρμογή μοντέλων βασισμένων στην ML για την πρόβλεψη τιμών απαρατήρητων δεδομένων-στόχων απεικονίζεται στο Σχήμα 1. Κατά την εκπαίδευση, συλλέγονται δεδομένα με γνωστές τιμές-στόχους, επιλέγεται ένα υποσύνολο χαρακτηριστικών και στη συνέχεια χρησιμοποιείται για την

κατασκευή ενός μοντέλου πρόβλεψης. Υπάρχουν πολλά υποσύνολα χαρακτηριστικών που επιλέγονται και διάφοροι αλγόριθμοι ML που χρησιμοποιούνται- ως εκ τούτου, υπάρχουν διάφοροι προγνωστικοί παράγοντες που μπορούν να εκπαιδευτούν. Κατά τη δοκιμή, επικυρώνονται και αξιολογούνται τα παραγόμενα μοντέλα από τη φάση της εκπαίδευσης. Στόχος της εργασίας μας είναι να δημιουργήσουμε διαφορετικά μοντέλα πρόβλεψης για δύο διαφορετικούς ρύπους: PM2.5, PM 10 και για τον δείκτη ποιότητας αέρα.



Εικόνα 4.2: Διάγραμμα ροής της διαδικασίας πρόβλεψης

4.6 Ορίζοντες πρόβλεψης

Ο ορίζοντας πρόβλεψης είναι ένας αριθμός μελλοντικών γιθα τις οποίες γίνεται η πρόβλεψη. Συχνά καθορίζεται με τη φύση της μεταβλητής που αναλύεται. Σημειακή εκτίμηση ή σημειακή πρόβλεψη είναι ένας αριθμός(τιμή) που αντιπροσωπεύει την καλύτερη εκτίμηση μιας μελλοντικής τιμής της μεταβλητής που αναλύεται. Διάστημα πρόβλεψης είναι το διάστημα στο οποίο αναμένεται να βρίσκονται οι μελλοντικές παρατηρήσεις με ορισμένο επίπεδο πιθανότητας.

(1) Προκειμένου να εφαρμοστούν μοντέλα μηχανικής μάθησης σε προβλήματα πρόβλεψης, η χρόνο-σειρά πρέπει να μετασχηματιστεί σε ένα πίνακα στο οποίο η τιμή σχετίζεται με το χρονικό παράθυρο(υστερήσεις) που προηγείται. (2) Στο πλαίσιο μιας χρόνο-σειράς μια υστέρηση σε σχέση με ένα χρονική βήμα t, ορίζεται ως οι τιμές τις σειράς σε προηγούμενα χρονικά βήματα. Για παράδειγμα, υστέρηση 1 είναι η τιμή στο χρονικό βήμα t-1 και η

υστέρηση(m) είναι η τιμή στο επόμενο χρονικό βήμα t-m. Αυτός ο τύπος μετασχηματισμού επιτρέπει επίσης την προσθήκη πρόσθετων μεταβλητών. Μόλις τα δεδομένα μετασχηματιστούν στο νέο σχήμα μπορεί να εκπαιδευτεί οποιοδήποτε μοντέλο παλινδρόμησης για να προβλέψει την επόμενη τιμή(βήμα) της χρόνο-σειράς. Κατά την εκπαίδευση του μοντέλου, κάθε χρόνο-σειρά θεωρείται ξεχωριστεί περίπτωση δεδομένων, όπου οι τιμές στις υστερήσεις (3)1,2, ……1 θεωρούνται προγνωστικοί δείκτες για την ποσότητα- στόχο της χρόνο-σειράς στο χρονικό βήμα p+1.

Επιλέγουμε αξιόπιστα μοντέλα για την πρόβλεψη της επόμενης ώρας με ορίζοντα την επόμενη ώρα(ωριαία πρόβλεψη), η οποία είναι σύμφωνη με άλλες εργασίες που έχουν αναπτύξει παρόμοια συστήματα[]. Αλλά και η πρόβλεψη των συγκεντρώσεων των ρύπων με ορίζοντα 24 ωρών είναι χρήσιμη προκειμένου να προειδοποιούνται οι κάτοικοι τις αστικής περιοχής για πιθανούς κινδύνους λόγω των υψηλών επιπέδων, γεγονός που πρέπει να επηρεάσει τη συμπεριφορά τους προκειμένου να μειώσουν την έκθεση τους αποφεύγοντας σχεδόν κάθε δραστηριότητα σε μια συγκεκριμένη περιοχή για κάποια ώρα. Σύμφωνα με τα παραπάνω, ορίζοντας πρόβλεψης 24 ωρών είναι πιθανώς η καλύτερη επιλογή, η οποία εξισορροπεί την ακρίβεια πρόβλεψης, με την χρησιμότητα μιας έγκυρης προειδοποίησης σε ένα αστικό περιβάλλον. Ο ορίζοντας πρόβλεψης ένα δείχνει ότι πρέπει να προβλεφθεί ένα βήμα στο μέλλον δηλαδή να οριστεί η αμέσως επόμενη τιμή για τα αιωρούμενα σωματίδια PM 2.5 & PM 10 ως ετικέτα για την αντίστοιχη περίπτωση . Το μέγεθος του βήματος ορίζεται σε ένα προκειμένου να μετακινηθεί ένα βήμα για το επόμενο παράθυρο. Στα αρχικά δεδομένα κάθε σημείο δεδομένων dti αντιστοιχίζεται σε dtjk όπου i είναι ο δείκτης γραμμής και οι τιμές ορίζονται από μηδέν έως το μέγεθος των αρχικών δεδομένων είναι n, j είναι ο δείκτης γραμμής των δεδομένων, και οι τιμές κυμαίνονται από μηδέ μέχρι το μέγεθος του παραθύρου, και k είναι ο δείκτης που περιλαμβάνει τιμές από μηδέν έως το μέγεθος του παραθύρου -1(Μ.Π. -1). Ορίζοντας πρόβλεψης 24 ωρών μπορούμε να χρησιμοποιήσουμε τρεις ημέρες 72 ώρες για να προβλέψουμε τις επόμενες 24 ώρες.

4.7 Μετρικές Αξιολόγησης

Για την αξιολόγηση των μοντέλων εφαρμόζεται ένα σύνολο πολλαπλών μετρικών επιδόσεων. Αυτά τα μέτρα είναι στατιστικά κριτήρια που μπορούν να χρησιμοποιηθούν για τη μέτρηση και την παρακολούθηση της απόδοσης ενός μοντέλου. Τα πιο δημοφιλή κριτήρια αξιολόγησης είναι

το μέσο απόλυτο σφάλμα (MAE), το μέσο τετραγωνικό σφάλμα (RMSE). Το MAE είναι ο αριθμητικός μέσος όρος της διαφοράς μεταξύ των πραγματικών και των προβλεπόμενων τιμών, ενώ το RMSE είναι η τετραγωνική ρίζα του μέσου όρου της τετραγωνικής διαφοράς μεταξύ της τιμής στόχου και της τιμής που προβλέπει το μοντέλο. Είναι η τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος. Εδώ, τα σφάλματα είναι οι διαφορές μεταξύ των προβλεπόμενων τιμών (τιμές που προβλέπονται από το μοντέλο παλινδρόμησής) και των πραγματικών τιμών μιας μεταβλητής.

Η μαθηματική αναπαράσταση των μέτρων αξιολόγησης MAE, RMSE δίνεται από την παρακάτω σχέσεις:

MAE

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \quad (1)$$

RMSE

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

Τα μοντέλα μηχανικής μάθησης επικυρώνονται συγκρίνοντας τις μετρήσεις αξιολόγησης. Όσο χαμηλότερα είναι τα MAE, RMSE το μοντέλο μηχανικής μάθησης αποδίδει καλύτερα.

Μοντελοποίηση & Πειράματα

Εδώ, παρουσιάζουμε μια διάφορα μοντέλα μηχανικής μάθησης για να συγκρίνουμε τις επιδόσεις τους στις προβλέψεις της ποιότητας του αέρα. Παρουσιάζεται πρώτα την εφαρμογή της ικανότητας του μοντέλου να προβλέπει κανονικά πρότυπα ατμοσφαιρικών ρύπων και, δεύτερον, τον δείκτη ποιότητας του αέρας. Αποφασίστηκε να εφαρμοστούν οκτώ τεχνικές πρόβλεψης, η καθεμία με το δικό της μοναδικό χαρακτηριστικό, οι οποίες αναγνωρίστηκαν ως δυνητικά πλεονεκτικές προσεγγίσεις για την πρόβλεψη της ποιότητας του αέρα. Οι στατιστικές μέθοδοι ARIMA FFT(Fast Fourier Transformation) και Theta έχουν εφαρμοστεί σε προβλήματα χρονοσειρών με αξιόπιστα αποτελέσματα στο παρελθόν, κυρίως το Arima. Τα βαθιά νευρωνικά δίκτυα το τυχαίο δάσος και η μηχανές διανυσμάτων υποστήριξης είχαν χρησιμοποιηθεί στην πρόσφατη βιβλιογραφία με ισχυρά αποτελέσματα σε προβλήματα πρόβλεψης της ποιότητας του αέρα. Μια έκδοση του CCN συμπεριλήφθηκε λόγω των προβλεπτικών ικανοτήτων του σε προβλήματα χρονοσειρών. Τα νευρωνικά δίκτυα, MLP, CNN και LSTM, χρειάστηκαν το μεγαλύτερο χρονικό διάστημα για να βρουν τις βέλτιστες παραμέτρους, ενώ οι στατιστικές τεχνικές περιλάμβαναν λιγότερες υπέρ-παραμέτρους και δεν ήταν τόσο ευαίσθητες όσον αφορά τη ρύθμιση στα νευρωνικά δίκτυα.

5.1 Random Forest Regression

Ο Random Forest χρησιμοποιείται για την πρόβλεψη χρόνο-σειρών, αν και απαιτεί πρώτα να μετατραπεί το σύνολο δεδομένων χρόνο-σειρών σε πρόβλημα μάθησης με επίβλεψη. Αυτό απαιτεί επίσης τη χρήση μιας εξειδικευμένης τεχνικής για την αξιολόγηση του μοντέλου που ονομάζεται επικύρωση walk-forward, καθώς η αξιολόγηση του μοντέλου με τη χρήση k-πτυχών διασταυρούμενης επικύρωσης θα οδηγούσε σε λανθασμένα αποτελέσματα. Στην επικύρωση walk-forward, το σύνολο δεδομένων χωρίζεται πρώτα σε σύνολα εκπαίδευσης και δοκιμής επιλέγοντας ένα σημείο αποκοπής.

Η παρακάτω συνάρτηση εκτελεί την επικύρωση walk-forward.

```
# walk-forward validation for univariate data
def walk_forward_validation_MSE(data, n_test, error_def):
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
predictions = list()
train, test = train_test_split(data, n_test)
history = [x for x in train]
for i in range(len(test)):
testX, testy = test[i, :-1], test[i, -1]
yhat = random_forest_forecast(history, testX)
predictions.append(yhat)
history.append(test[i])
print('>expected=%.1f, predicted=%.1f' % (testy, yhat))
if error_def == "mse":
error = mean_squared_error(test[:, -1], predictions)
if error_def =="rmse":
error = sqrt(mean_squared_error(test[:, -1], predictions))
if error_def == "mae":
error = mean_absolute_error(test[:, -1], predictions)
return error, test[:, -1], predictions
data = series_to_supervised(values, n_in=6)
rmse, y, yhat = walk_forward_validation_MSE(data, 1, "rmse")
print('RMSE: %.3f' % rmse)
mae, y, yhat = walk_forward_validation_MSE(data, 12, "mae")
print('MAE: %.3f' % mae)
pyplot.plot(y, label='Actual')
#pyplt.plot(yhat, color="red")
pyplot.plot(yhat, label='Predicted')
pyplot.legend()
pyplot.show()
```

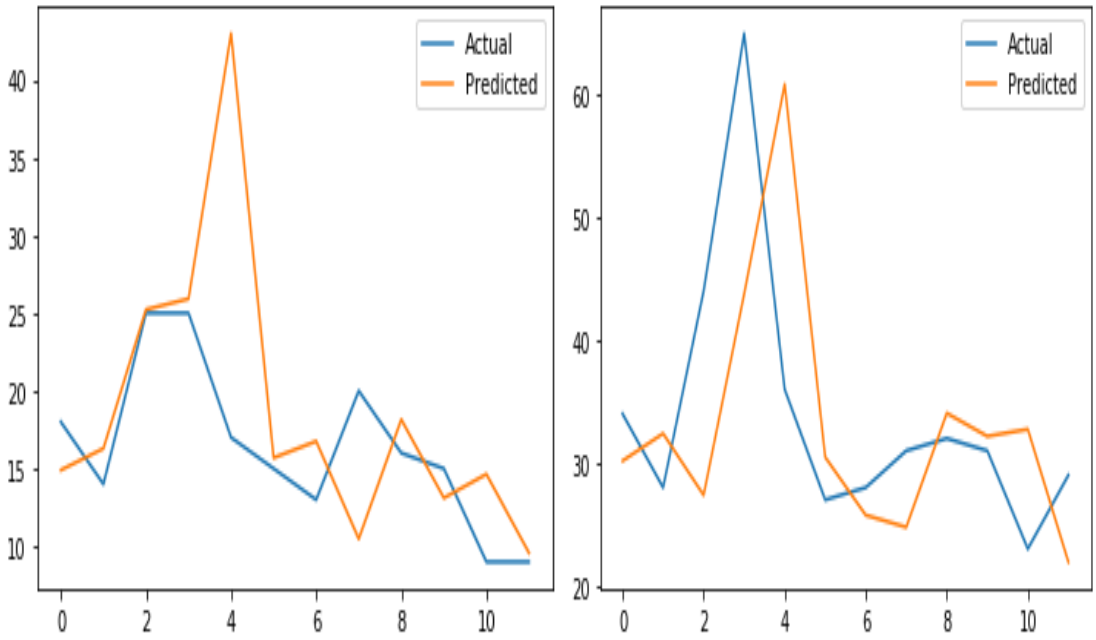
Η συνάρτηση train_test_split() καλείται για να χωρίσει το σύνολο δεδομένων σε σύνολα train και test. Μπορούμε να χρησιμοποιήσουμε την κλάση RandomForestRegressor για να κάνουμε μια πρόβλεψη ενός βήματος. Η παρακάτω συνάρτηση random_forest_forecast() το υλοποιεί αυτό, λαμβάνοντας ως είσοδο το σύνολο δεδομένων εκπαίδευσης και τη σειρά εισόδου δοκιμής, προσαρμόζοντας ένα μοντέλο και κάνοντας μια πρόβλεψη ενός βήματος.

```
# fit an random forest model and make a one step prediction
  def random_forest_forecast(train, testX):
  train = np.asarray(train)
  trainX, trainy = train[:, :-1], train[:, -1]
  model = RandomForestRegressor(n_estimators=1000)
  model.fit(trainX, trainy)
  yhat = model.predict([testX])
  return yhat[0]
```

Τα δεδομένα υποβάλονται σε προ-επεξεργασία και εκαίδευονται με τον αλγόριθμο Random Forest Reggresion για την πρόβλεψη δύο διαφορετικών ρύπων(αιωρούμενα σωματίδια) PM 2.5, PM 10, καθώς και για την πρόβλεψη του δείκτη ποιότητας του αέρα. Για τα πειράματα υποθέσαμε τα χαρακτηριτικά PM 2.5 και PM 10 ως είσοδο για να πάρουμε ως τελική έξοδο το χαρακτηριστικό AQI του δείκτη ποιότητας αέρα. Η εκτέλεση του παραπάνω κώδικα αναφέρει τις αναμενόμενες και τις προβλεπόμενες τιμές για κάθε βήμα στο σύνολο δοκιμών και, στη συνέχεια, το RMSE, MAE για όλες τις προβλεπόμενες τιμές.

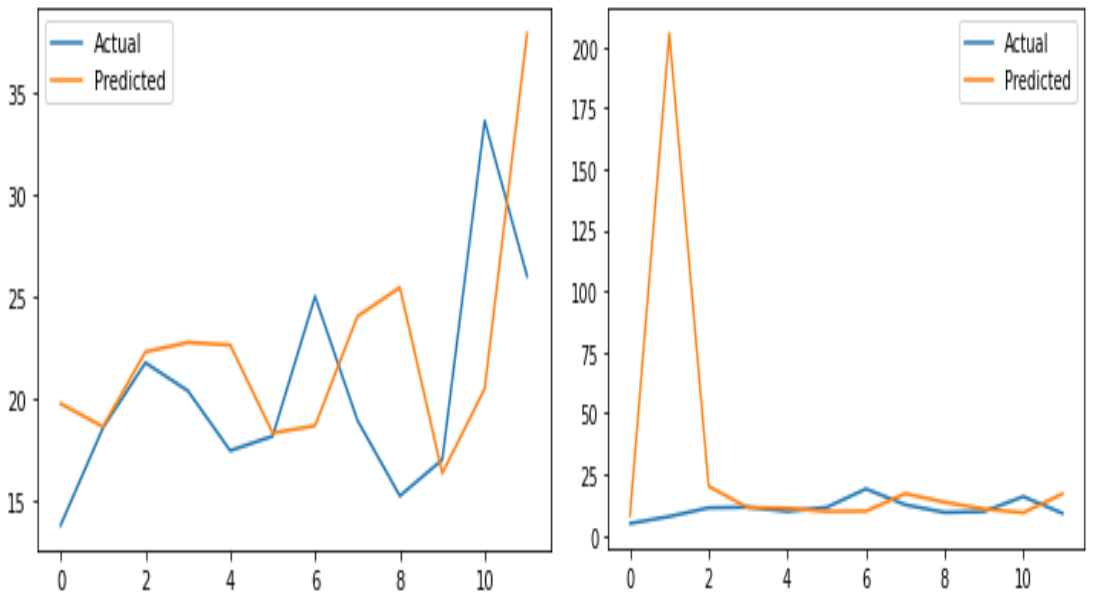
Το μοντέλο RFR εφαρμόστηκε χρησιμοποιώντας το ίδιο σύνολο δεδομένων για τις παραμέτρους PM 2.5, PM10, και διαφορετικό για το Δείκτη Ποιότητας Αέρα(AQI). Τα αποτελέσματα που προέκυψαν για τα αιωρούμενα σωματίδια PM 2.5, PM10 με εφαρμογή του μοντέλου RFR παρουσιάζονται στην εικόνα 3.1, ενώ τα αποτελέσματα για το Δείκτη ποιότητας του αέρα παρουσιάζονται στην εικόνα 3.2. Αυτό το σχήμα αποκαλύπτει ότι το μοντέλο RFR μπορεί να προβλέψει τις παρατηρούμενες τιμές PM 2.5, PM10, AQI με καλή ακρίβεια. Λαμβάνοντας υπόψη τις ακραίες τιμές, οι προβλεπόμενες τιμές έχουν μεγαλύτερη συχνότητα ακραίων τιμών σε σύγκριση με τις παρατηρούμενες, γεγονός που υποδηλώνει ότι το μοντέλο RFR τείνει να υποεκτιμά κάπως τις προβλεπόμενες τιμές. Τα στατιστικά στοιχεία σφάλματος πρόβλεψης για τα αιωρούμενα σωματίδια PM 2.5, PM 10, και του δείκτη ποιότητας του αέρα παρουσιάζονται στο πίνακα 5,1.

Στην εικόνα 5.1 απεικονίζεται το διάγραμμα διασποράς ωριαίας πρόβλεψης των αιωρούμενων σωματιδίων PM 2.5 & PM10 μεταξύ πραγματικών και προβλεπόμενων τιμών του RFR μοντέλου, για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



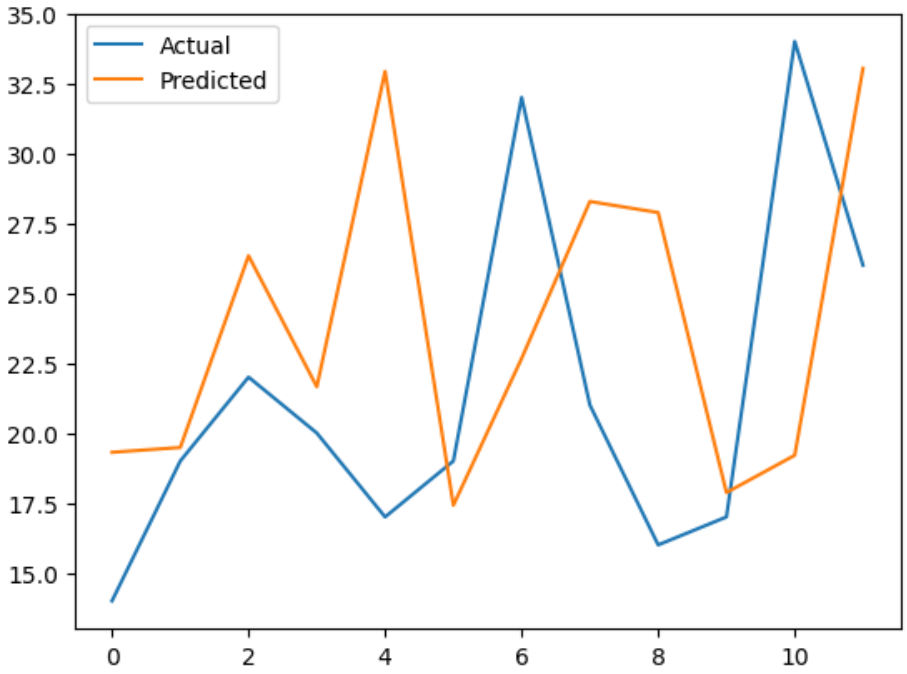
Εικόνα 5.1: Ωριαία Πρόβλεψη αιωρούμενων σωματιδίων PM 2.5, PM 10 με χρήση του μοντέλου πρόβλεψης RFR.

Στην εικόνα 5.2 απεικονίζεται το διάγραμμα διασποράς ημερήσιας πρόβλεψης των αιωρούμενων σωματιδίων PM 2.5 & PM10 μεταξύ πραγματικών και προβλεπόμενων τιμών του RFR μοντέλου, για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



Εικόνα 5.2: Ημερήσια Πρόβλεψη αιωρούμενων σωματιδίων PM 2.5, PM 10 με χρήση του μοντέλου πρόβλεψης RFR.

Στην εικόνα 5.3 απεικονίζεται το διάγραμμα διασποράς του δείκτη ποιότητας αέρα μεταξύ πραγματικών και προβλεπόμενων τιμών του RFR μοντέλου για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



Εικόνα 5.3: Πρόβλεψη του Δίκτη Ποιότητας του Αέρα(AQI), με χρήση του μοντέλου πρόβλεψης RFR.

Μετρικές	PM 2.5		PM 10		AQI
MAE	4.687	20.559	4.711	5.132	6.708
RMSE	1.437	7.731	0.728	11.956	7.314

Πίνακας 5.1: σύγκριση μετρικών αξιολόγησης παραμέτρων PM 2.5, PM10(πρόβλεψη επόμενης ώρας και 24 ωρών), AQI για το μοντέλο RFR, για ολόκληρο το σύνολο δεδομένων δοκιμής.

5.2 Support Vector Regression

Το SVM έχει τρεις υπέρ-παραμέτρους που πρέπει να καθοριστούν από τον χρήστη: τη συνάρτηση τύπου πυρήνα, τη σταθερά κανονικοποίησης *C* και τη μέγιστη επιτρεπόμενη απόκλιση *ε*. Ο διαχωρισμός χρόνο-σειρών σε συνδυασμό με την αναζήτηση τυχαίου πλέγματος χρησιμοποιήθηκε για να προκύψουν οι βέλτιστοι αριθμοί τόσο για το *C* όσο και για το *ε*, 2παρόμοια με ό,τι έγινε στο [22]. Σύμφωνα με το [21], το κατάλληλο εύρος για την παράμετρο *C* θα πρέπει να κυμαίνεται μεταξύ 10 και 100. Για το διαχωρισμό των

δεδομένων σε σύνολα δεδομένων εκπαίδευσης και δοκιμής, χρησιμοποιήθηκε μια μέθοδος από τη βιβλιοθήκη Scikit-learn, η οποία επιτρέπει την επιλογή του μεγέθους της δοκιμής και του μεγέθους της εκπαίδευσης ως ποσοστό. Τα δεδομένα που προέκυψαν αποθηκεύονται στις μεταβλητές test και train. Το 80% του συνόλου δεδομένων χρησιμοποιήθηκε για τους σκοπούς της εκπαίδευσης της μεθόδου μηχανικής μάθησης. Το υπόλοιπο 20% των δεδομένων χρησιμοποιήθηκε για τη δοκιμή του αλγορίθμου.

Παράδειγμα του κώδικα όπου τα δεδομένα χωρίζονται σε σύνολο δεδομένων εκπαίδευσης και δοκιμής:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.200)
```

Για να ληφθούν αυτές οι τιμές, ήταν απαραίτητο να δοκιμαστούν διάφορες τιμές για κάθε παράμετρο με τη μέθοδο .fit(). Η μέθοδος αυτή μπορεί να λάβει δύο ορίσματα: train_x, που αντιπροσωπεύει τους δείκτες για το σύνολο δεδομένων εκπαίδευσης, και train_y, που αντιπροσωπεύει τις τιμές-στόχους για το σύνολο δεδομένων εκπαίδευσης. Πρόκειται για πραγματικούς αριθμούς στην παλινδρόμηση. Στην παρούσα μέθοδο εξετάζεται ο rbf πυρήνας (PM 2.5: SVR(C=5) – PM 10: SVR(C=50). Στη συνέχεια, αφού εκπαιδευτεί το μοντέλο, μπορεί να χρησιμοποιηθεί για την πραγματοποίηση προβλέψεων. Προκειμένου να προβλεφθούν τα δεδομένα, χρησιμοποιείται η εφαρμογή της μεθόδου .predict() που οδηγεί στη μεταβλητή y_pred με βάση το σύνολο δεδομένων δοκιμής. Όταν ολοκληρωθεί η πρόβλεψη, το σύνολο δεδομένων δοκιμής συγκρίνεται με το προβλεπόμενο σύνολο δεδομένων με τον υπολογισμό του μέσου τετραγωνικού σφάλματος και του μέσου απόλυτου σφάλματος. Τα στατιστικά στοιχεία σφάλματος πρόβλεψης για τις συγκεντρώσεις των σωματιδίων, και του δείκτη ποιότητας του αέρα, που προέκυψαν από την εφαρμογή του μοντέλου SVR παρουσιάζονται στον πίνακα 5.2.

```
test_y = test_y.reshape((len(test_y), 1))
inv_y = np.concatenate((test_y, test_x), axis=1)
inv_y = scaler.inverse_transform(inv_y)
inv_y = inv_y[:, 0]
yhat_svm = model_svm.predict(test_x)
yhat_svm = yhat_svm.reshape((len(yhat_svm), 1))
# invert scaling for forecast
inv_yhat_svm = np.concatenate((yhat_svm, test_x), axis=1)
inv_yhat_svm = scaler.inverse_transform(inv_yhat_svm)
inv_yhat_svm = inv_yhat_svm[:, 0]
rmse = np.sqrt(mean_squared_error(inv_y, inv_yhat_svm))
mae = mean_absolute_error (inv_y, inv_yhat_svm))
print('Test RMSE: %.3f' % rmse)
print('Test MAE: %.3f' % mae)
```

Το μοντέλο μηχανικής μάθησης SVM εφαρμόστηκε χρησιμοποιώντας το ίδιο σύνολο δεδομένων για τις παραμέτρους PM 2.5, PM10, και διαφορετικό για το Δείκτη Ποιότητας Αέρα(AQI). Τα αποτελέσματα των προβλέψεων που προέκυψαν με την χρήση του μοντέλου SVM παρουσιάζονται στο πίνακα 5.2.

Μετρικές	**PM 2.5**	**PM 10**	**AQI**
MAE	0.678	0.287	8.165
RMSE	1.814	4.314	8.732

Πίνακας 5.2: σύγκριση μετρικών αξιολόγησης παραμέτρων PM 2.5, PM10(πρόβλεψη επόμενης ώρας και 24 ωρών), AQI για το μοντέλο SVR, για ολόκληρο το σύνολο δεδομένων δοκιμής.

5.3 ARIMA

Ο αλγόριθμος ARIMA χρησιμοποιήθηκε στη διαδικασία εκτίμησης των παραμέτρων PM10, PM2.5, και AQI. Η βιβλιοθήκη statsmodels παρέχει τη δυνατότητα προσαρμογής ενός μοντέλου ARIMA. Ένα μοντέλο ARIMA μπορεί να δημιουργηθεί χρησιμοποιώντας τη βιβλιοθήκη statsmodels ως εξής: Ορίζουμε το μοντέλο καλώντας την ARIMA() και περνώντας τις παραμέτρους p, d και q. Το μοντέλο προετοιμάζεται στα δεδομένα εκπαίδευσης καλώντας τη συνάρτηση fit(). Οι προβλέψεις μπορούν να γίνουν καλώντας τη συνάρτηση predict() και καθορίζοντας το δείκτη του χρόνου ή των χρόνων που πρέπει να προβλεφθούν. Το 80% του συνόλου δεδομένων χρησιμοποιήθηκε για τους σκοπούς της εκπαίδευσης της μεθόδου μηχανικής μάθησης. Το υπόλοιπο 20% των δεδομένων χρησιμοποιήθηκε για τη δοκιμή του αλγορίθμου προκειμένου να εκτιμηθεί η επόμενη τιμή, σύμφωνα με τον αλγόριθμο που παρουσιάζεται παρακάτω:

```
# split into train and test sets
  X = df.values
  size = int(len(X) * 0.8)
  train, test = X[0:size], X[size:len(X)]
  history = [x for x in train]
  predictions = list()
  # walk-forward validation
  for t in range(len(test)):
  model = ARIMA(history, order=(5,1,0))
  model_fit = model.fit()
  output = model_fit.forecast()
  yhat = output[0]
  predictions.append(yhat)
  obs = test[t]
  history.append(obs)
  print('predicted=%f, expected=%f' % (yhat, obs))
  # evaluate forecasts
  from math import sqrt
  from sklearn.metrics import
  mean_absolute_error,mean_squared_error
  rmse = sqrt(mean_squared_error(test, predictions))
  mae = mean_absolute_error(test, predictions)
  print('Test RMSE: %.3f' % rmse)
  print("Test MAE : %.3f "% mae)
```

Το μοντέλο μηχανικής μάθησης ARIMA εφαρμόστηκε χρησιμοποιώντας το ίδιο σύνολο δεδομένων για τις παραμέτρους PM 2.5, PM10, και διαφορετικό για το Δείκτη Ποιότητας Αέρα(AQI). Τα αποτελέσματα των προβλέψεων που προέκυψαν με την χρήση του μοντέλου ARIMA παρουσιάζονται στο πίνακα 5.3.

Μετρικές	**PM 2.5**	**PM 10**	**AQI**
MAE	0.596	0.198	7.732
RMSE	1.875	1.256	7.425

Πίνακας 5.3: σύγκριση μετρικών αξιολόγησης παραμέτρων PM 2.5, PM10(πρόβλεψη επόμενης ώρας και 24 ωρών), AQI για το μοντέλο ARIMA, για ολόκληρο το σύνολο δεδομένων δοκιμής.

5.4 Fast Fourier Transformation

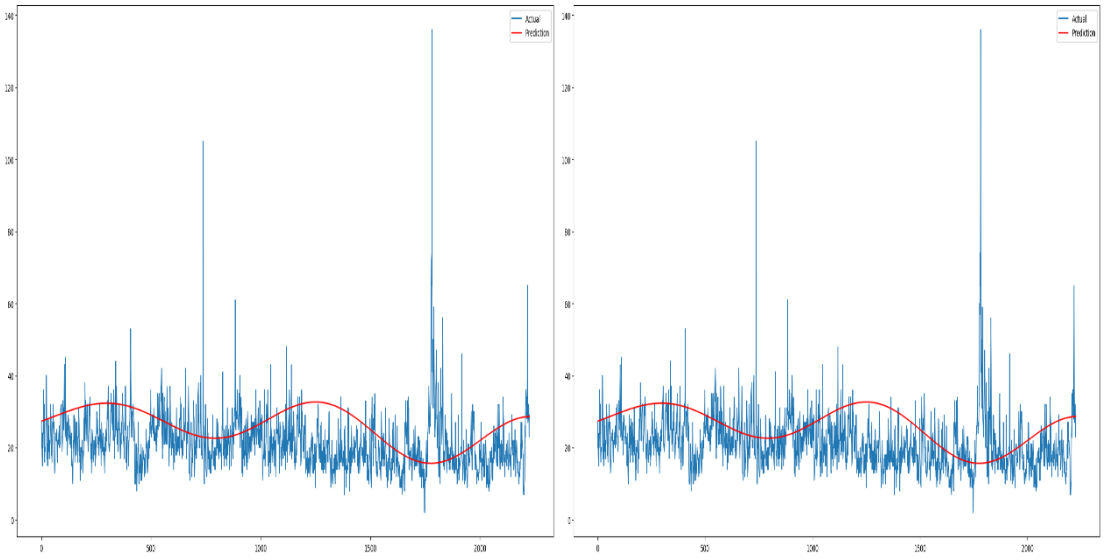
Ο αλγόριθμος FFT χρησιμοποιήθηκε στη διαδικασία εκτίμησης των παραμέτρων PM10, PM2.5, και AQI. Η βιβλιοθήκη FFT παρέχει τη δυνατότητα προσαρμογής ενός μοντέλου FFT το μοντέλο προετοιμάζεται στα δεδομένα εκπαίδευσης. Οι προβλέψεις μπορούν να

γίνουν καλώντας τη συνάρτηση fft(x, n_predict) και καθορίζοντας το δείκτη του χρόνου ή των χρόνων που πρέπει να προβλεφθούν. Το 80% του συνόλου δεδομένων χρησιμοποιήθηκε για τους σκοπούς της εκπαίδευσης της μεθόδου μηχανικής μάθησης. Το υπόλοιπο 20% των δεδομένων χρησιμοποιήθηκε για τη δοκιμή, σύμφωνα με τον αλγόριθμο που παρουσιάζεται παρακάτω:

```
def fft(x, n_predict):
n = x.size
n_harm = 10
t = np.arange(0, n)
p = np.polyfit(t, x, 1)
x_notrend = x - p[0] * t
x_freqdom = fft.fft(x_notrend)
f = fft.fftfreq(n)
indexes = list(range(0,n))
# sort indexes by frequency, lower -> higher
indexes.sort(key = lambda i: np.absolute(f[i]))
t = np.arange(0, n + n_predict)
restored_sig = np.zeros(t.size)
for i in indexes[:1 + n_harm * 2]:
ampli = np.absolute(x_freqdom[i]) / n # amplitude
phase = np.angle(x_freqdom[i]) # phase
restored_sig += ampli * np.cos(2 * np.pi * f[i] * t + phase) return restored_sig + p[0] * t
import numpy as np
from sklearn.metrics import mean_absolute_error,mean_squared_error
rmse = np.sqrt(mean_squared_error(test, pred[t:]))
mae = mean_absolute_error(test, pred[t:])
print('Test RMSE: %.3f' % rmse)
print("Test MAE : %.3f "% mae)
```

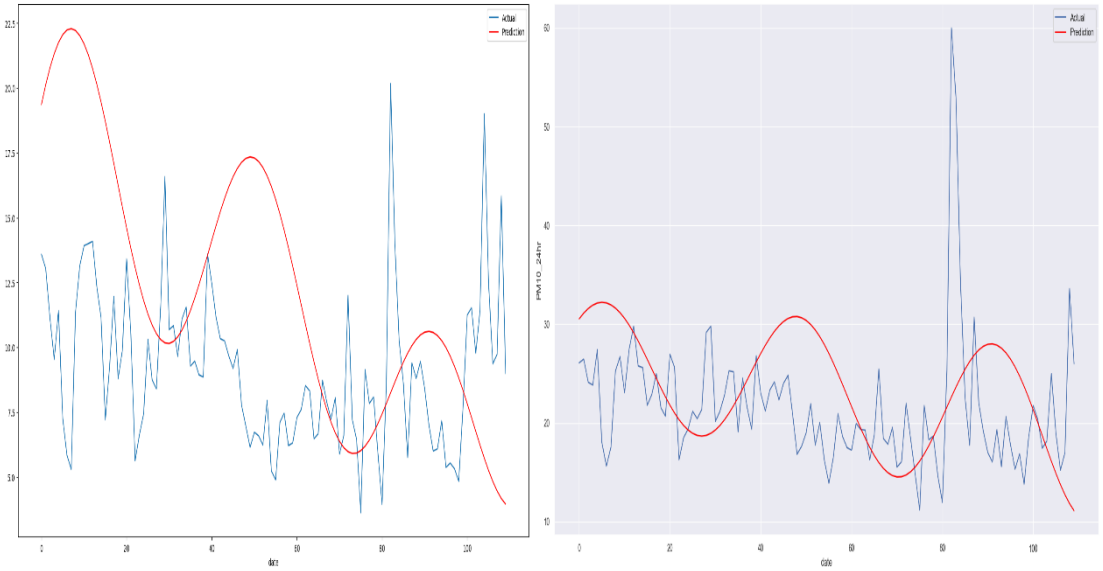
Το μοντέλο μηχανικής μάθησης FFT εφαρμόστηκε χρησιμοποιώντας το ίδιο σύνολο δεδομένων για τις παραμέτρους PM 2.5, PM10, και διαφορετικό για το Δείκτη Ποιότητας Αέρα(AQI). Τα αποτελέσματα που προέκυψαν για τα αιωρούμενα σωματίδια PM 2.5, PM10 με εφαρμογή του μοντέλου FFT ωριαίες και ημερήσιες προβλέψεις παρουσιάζονται στις εικόνες 5.4 κι 5.5, ενώ τα αποτελέσματα για το Δείκτη ποιότητας του αέρα παρουσιάζονται στην εικόνα 5.6. Το μοντέλο επιτυγχάνει μέτρια έως κακή πρόβλεψη. Οι λαμβανόμενες προβλέψεις έχουν απόσταση από τις παρατηρούμενες. Ωστόσο οι παρατηρούμενες τιμές παρουσιάζουν μεγαλύτερη ποσότητα ακραίων τιμών που τα μοντέλα πρόβλεψης τείνουν να υπό-εκτιμούν. Τα στατιστικά στοιχεία σφάλματος πρόβλεψης για τα αιωρούμενα σωματίδια PM 2.5, PM 10, και του δείκτη ποιότητας του αέρα παρουσιάζονται στο πίνακα 5.4.

Στην εικόνα 5.4 απεικονίζεται το διάγραμμα διασποράς ωριαίας πρόβλεψης των αιωρούμενων σωματιδίων PM 2.5 & PM10 μεταξύ πραγματικών και προβλεπόμενων τιμών του Fourier μοντέλου, για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



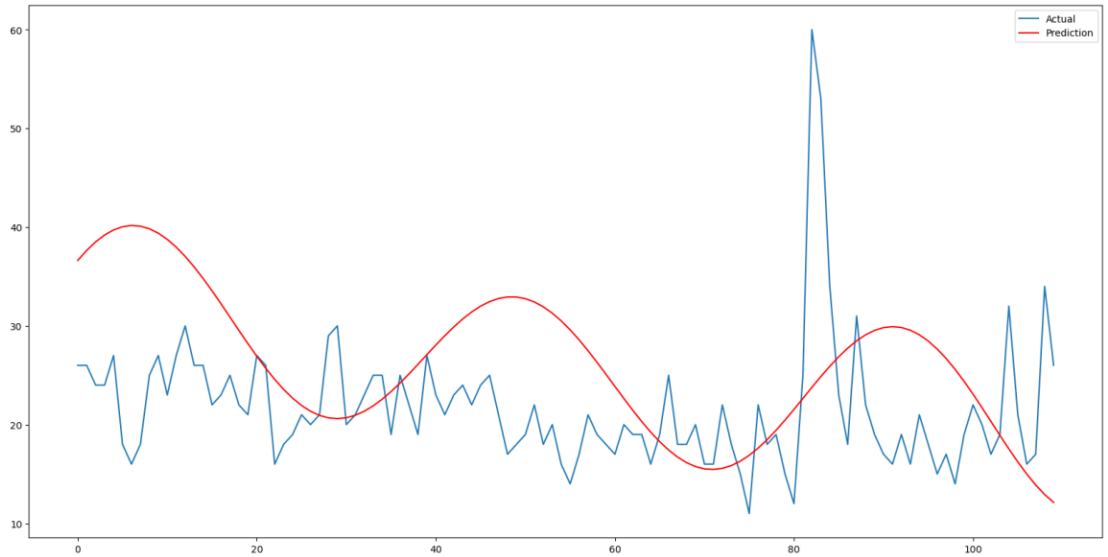
Εικόνα 5.4: Ωριαία Πρόβλεψη αιωρούμενων σωματιδίων PM 2.5, PM 10 με χρήση του μοντέλου πρόβλεψης Fourier.

Στην εικόνα 5.5 απεικονίζεται το διάγραμμα διασποράς ημερήσιας πρόβλεψης των αιωρούμενων σωματιδίων PM 2.5 & PM10 μεταξύ πραγματικών και προβλεπόμενων τιμών του Fourier μοντέλου, για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



Εικόνα 5.5: Ημερήσια Πρόβλεψη αιωρούμενων σωματιδίων PM 2.5, PM 10 με χρήση του μοντέλου πρόβλεψης Fourier.

Στην εικόνα 5.6 απεικονίζεται το διάγραμμα διασποράς του δείκτη ποιότητας αέρα μεταξύ πραγματικών και προβλεπόμενων τιμών του Fourier μοντέλου για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



Εικόνα 5.6: Πρόβλεψη του Δίκτη Ποιότητας του Αέρα(AQI), με τη χρήση του μοντέλου πρόβλεψης FFT.

Μετρικές	**PM 2.5**		**PM 10**		**AQI**
MAE	5.956	5.143	7.866	6.116	7.939
RMSE	7.363	6.464	10.180	8.331	10.203

Πίνακας 5.4: σύγκριση μετρικών αξιολόγησης παραμέτρων PM 2.5, PM10(πρόβλεψη επόμενης ώρας και 24 ωρών), AQI για το μοντέλο Fourier, για ολόκληρο το σύνολο δεδομένων δοκιμής.

5.5 Theta

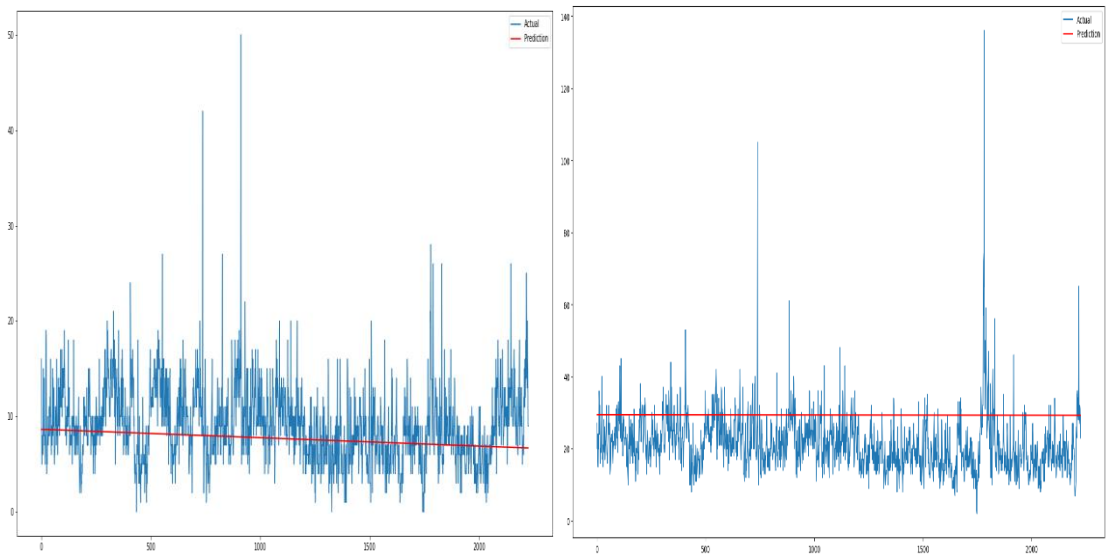
Για την διαδικασία εκτίμησης των παραμέτρων PM10, PM2.5, και AQI, εφαρμόστηκε το μοντέλο theta. Ένα μοντέλο Theta μπορεί να δημιουργηθεί χρησιμοποιώντας τη βιβλιοθήκη ThetaModel. Το μοντέλο προετοιμάζεται στα δεδομένα εκπαίδευσης καλώντας τη συνάρτηση fit(). Οι προβλέψεις μπορούν να γίνουν καλώντας τη συνάρτηση forecast. Το 80% του συνόλου δεδομένων χρησιμοποιήθηκε για τους σκοπούς της εκπαίδευσης της μεθόδου μηχανικής μάθησης. Το υπόλοιπο 20% των δεδομένων χρησιμοποιήθηκε για τη δοκιμή του αλγορίθμου προκειμένου να εκτιμηθεί η επόμενη τιμή, σύμφωνα με τον αλγόριθμο που παρουσιάζεται παρακάτω:

```
history = list(train)
predictions=[]
from tqdm import tqdm
for x in tqdm(range(len(test))) :
tm = ThetaModel(np.array(history),period =len(np.array(history)))
res = tm.fit(use_mle=True)
history.append(test[x])
predictions.append(res.forecast(1).values[0])
rmse = np.sqrt(mean_squared_error(test, res.forecast(11127-t).values))
mae = mean_absolute_error(test, res.forecast(11127-t).values)
print('Test RMSE: %.3f' % rmse)
print("Test MAE : %.3f "% mae)
```

Το μοντέλο μηχανικής μάθησης Theta εφαρμόστηκε χρησιμοποιώντας το ίδιο σύνολο δεδομένων για τις παραμέτρους PM 2.5, PM10, και διαφορετικό για το Δείκτη Ποιότητας Αέρα(AQI). Τα αποτελέσματα που προέκυψαν για τα αιωρούμενα σωματίδια PM 2.5, PM10(πρόβλεψη ωριαία, ημερήσια) με εφαρμογή του μοντέλου Theta παρουσιάζονται στις εικόνες 5.7-5.8, ενώ τα αποτελέσματα για το Δείκτη ποιότητας του αέρα παρουσιάζονται στην εικόνα 5.9. Το μοντέλο επιτυγχάνει μέτρια έως κακή πρόβλεψη. Οι λαμβανόμενες προβλέψεις έχουν απόσταση(είναι ευθεία γραμμή) από τις παρατηρούμενες. Ωστόσο οι

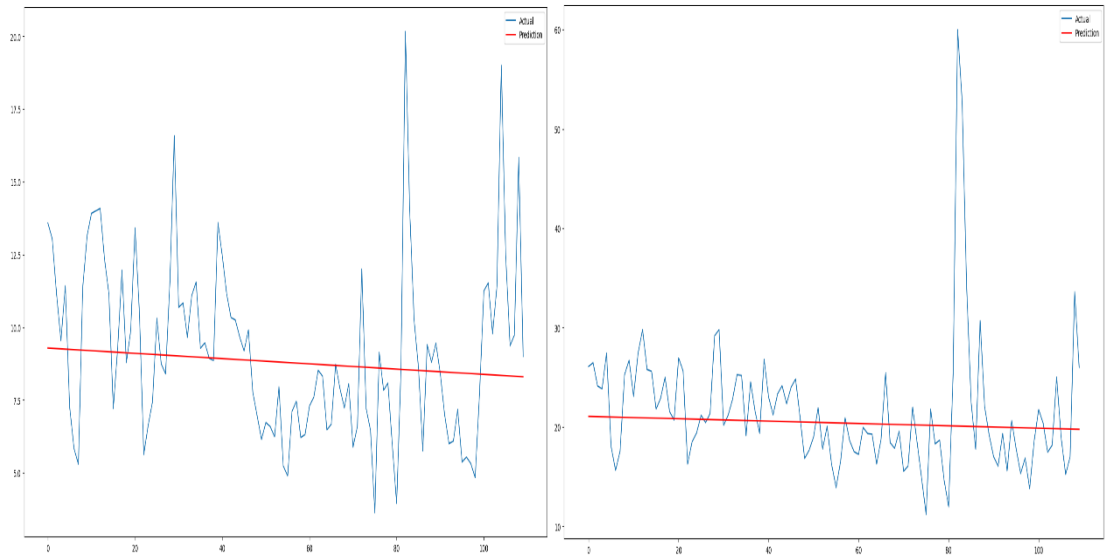
παρατηρούμενες τιμές παρουσιάζουν μεγαλύτερη ποσότητα ακραίων τιμών που τα μοντέλα πρόβλεψης τείνουν να υπό-εκτιμούν. Τα στατιστικά στοιχεία σφάλματος πρόβλεψης για τα αιωρούμενα σωματίδια PM 2.5, PM 10, και του δείκτη ποιότητας του αέρα παρουσιάζονται στο πίνακα 5.5.

Στην εικόνα 5.7 απεικονίζεται το διάγραμμα διασποράς ωριαίας πρόβλεψης των αιωρούμενων σωματιδίων PM 2.5 & PM10 μεταξύ πραγματικών και προβλεπόμενων τιμών του Theta μοντέλου, για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



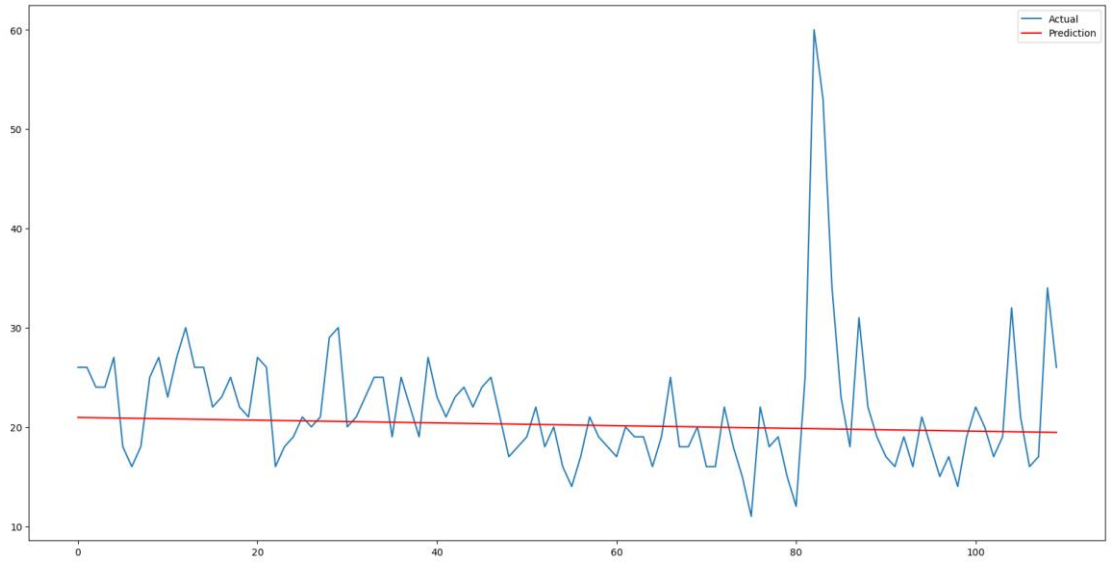
Εικόνα 5.7: Ωριαία Πρόβλεψη αιωρούμενων σωματιδίων PM 2.5, PM 10 με χρήση του μοντέλου πρόβλεψης Theta.

Στην εικόνα 5.8 απεικονίζεται το διάγραμμα διασποράς ημερήσιας πρόβλεψης των αιωρούμενων σωματιδίων PM 2.5 & PM10 μεταξύ πραγματικών και προβλεπόμενων τιμών του Theta μοντέλου, για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



Εικόνα 5.8: Ημερήσια Πρόβλεψη αιωρούμενων σωματιδίων PM 2.5, PM 10 με χρήση του μοντέλου πρόβλεψης Theta.

Στην εικόνα 5.9 απεικονίζεται το διάγραμμα διασποράς του δείκτη ποιότητας αέρα μεταξύ πραγματικών και προβλεπόμενων τιμών του Theta μοντέλου για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



Εικόνα 5.9: Πρόβλεψη του Δίκτη Ποιότητας του Αέρα(AQI), με τη χρήση του μοντέλου πρόβλεψης Theta.

Μετρικές	**PM 2.5**		**PM 10**		**AQI**
MAE	3.037	2.253	9.394	4.017	2.271
RMSE	4.159	2.836	11.102	6.536	2.993

Πίνακας 5.5: σύγκριση μετρικών αξιολόγησης παραμέτρων PM 2.5, PM10(πρόβλεψη επόμενης ώρας και 24 ωρών), AQI για το μοντέλο Theta, για ολόκληρο το σύνολο δεδομένων δοκιμής.

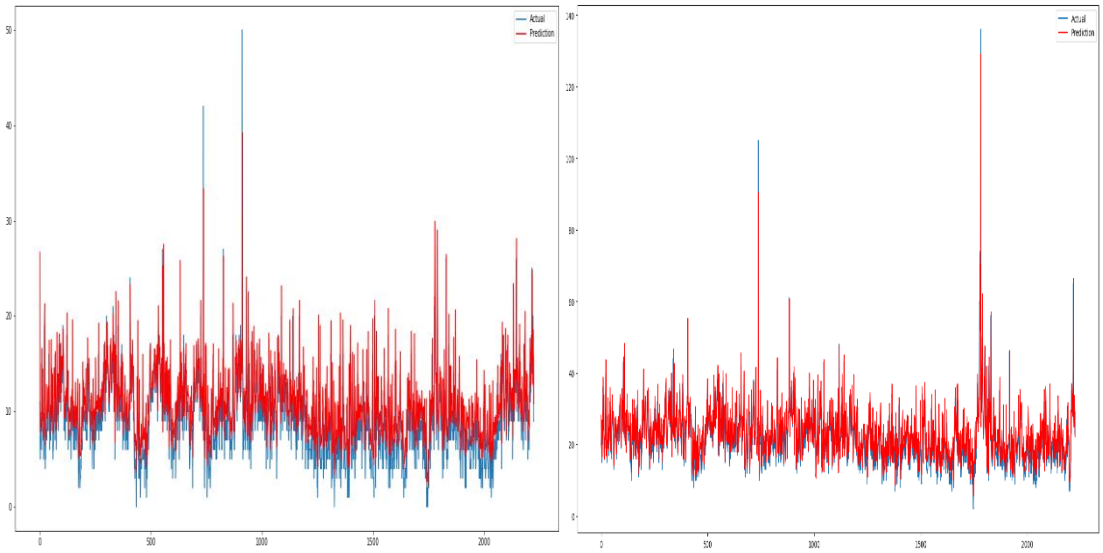
5.6 Multi Layer Perceptron

Το μοντέλο προετοιμάζεται στα δεδομένα εκπαίδευσης καλώντας τη συνάρτηση fit(). Οι προβλέψεις μπορούν να γίνουν καλώντας τη συνάρτηση predict() και καθορίζοντας το δείκτη του χρόνου ή των χρόνων που πρέπει να προβλεφθούν. Το 80% του συνόλου δεδομένων χρησιμοποιήθηκε για τους σκοπούς της εκπαίδευσης της μεθόδου μηχανικής μάθησης. Το υπόλοιπο 20% των δεδομένων χρησιμοποιήθηκε για τη δοκιμή του αλγορίθμου προκειμένου να εκτιμηθεί η επόμενη τιμή, σύμφωνα με τον αλγόριθμο που παρουσιάζεται παρακάτω:

```
initializer = tf.keras.initializers.HeNormal()
model = Sequential()
# model.add(Flatten())
model.add(Dense(100,activation='relu',input_dim =5))
model.add(Dense(50,activation='relu'))
model.add(Dense(1))
model.fit(X_train , y_train, epochs=100, batch_size=32,validation_data =
(X_test,y_test))
pred = model.predict(X_test)
rmse = np.sqrt(mean_squared_error(y_test,pred.reshape(len(pred))))
mae = mean_absolute_error(y_test, pred.reshape(len(pred)))
print('Test RMSE: %.3f' % rmse)
print("Test MAE : %.3f "% mae)
```

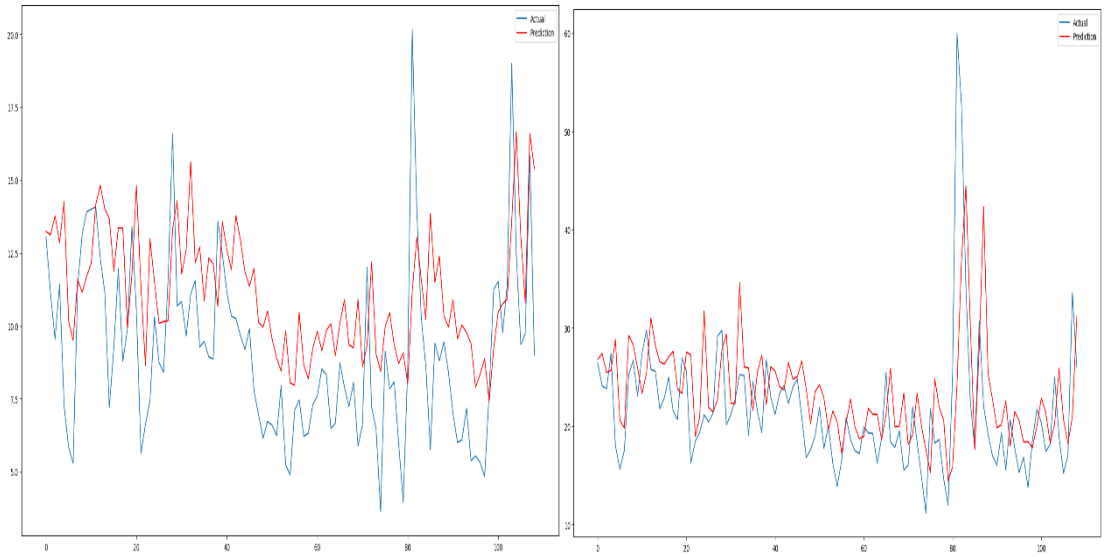
Το σύνολο δεδομένων που εφαρμόστηκε είναι το ίδιο όπως και για τα προηγούμενα μοντέλα. Τα αποτελέσματα που προέκυψαν με τη χρήση του μοντέλου MLP παρουσιάζονται στις εικόνες 5.10 – 5.11. Από τα σχήματα είναι αρκετά σαφές το MLP προβλέπει με μεγάλη ακρίβεια της συγκεντρώσεις PM 2.5, PM 10, AQI, οι προβλεπόμενες τιμές είναι παρόμοιες με τις πραγματικές, και παράγει σταθερή απόδοση. Τα στατιστικά στοιχεία σφάλματος πρόβλεψης για τις συγκεντρώσεις των σωματιδίων, και του δείκτη ποιότητας του αέρα, παρουσιάζονται στον πίνακα 5.6.

Στην εικόνα 5.10 απεικονίζεται το διάγραμμα διασποράς ωριαίας πρόβλεψης των αιωρούμενων σωματιδίων PM 2.5 & PM10 μεταξύ πραγματικών και προβλεπόμενων τιμών του MLP μοντέλου, για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



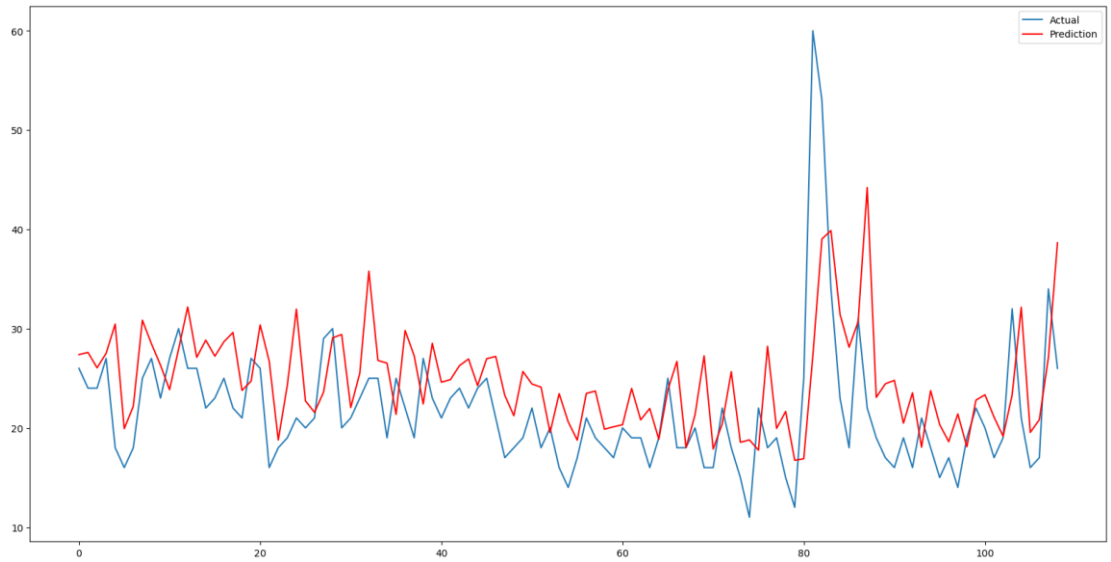
Εικόνα 5.10: Ωριαία Πρόβλεψη αιωρούμενων σωματιδίων PM 2.5, PM 10 με χρήση του μοντέλου πρόβλεψης MLP.

Στην εικόνα 5.11 απεικονίζεται το διάγραμμα διασποράς ημερήσιας πρόβλεψης των αιωρούμενων σωματιδίων PM 2.5 & PM10 μεταξύ πραγματικών και προβλεπόμενων τιμών του MLP μοντέλου, για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



Εικόνα 5.11: Ημερήσια Πρόβλεψη αιωρούμενων σωματιδίων PM 2.5, PM 10 με χρήση του μοντέλου πρόβλεψης MLP.

Στην εικόνα 5.12 απεικονίζεται το διάγραμμα διασποράς του δείκτη ποιότητας αέρα μεταξύ πραγματικών και προβλεπόμενων τιμών του MLP μοντέλου για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



Εικόνα 5.12: Πρόβλεψη του Δίκτη Ποιότητας του Αέρα(AQI), με τη χρήση του μοντέλου πρόβλεψης MLP.

Μετρικές	**PM 2.5**		**PM 10**		**AQI**
MAE	5.781	2.711	4.756	4.464	4.835
RMSE	3.877	3.211	2.891	6.365	6.112

Πίνακας 5.6: σύγκριση μετρικών αξιολόγησης παραμέτρων PM 2.5, PM10(πρόβλεψη επόμενης ώρας και 24 ωρών), AQI για το μοντέλο MLP, για ολόκληρο το σύνολο δεδομένων δοκιμής.

5.7 Long Short-Term Memory

Τα LSTM μπορούν να χρησιμοποιηθούν για τη μοντελοποίηση προβλημάτων πρόβλεψης μονό-μεταβλητών χρόνο-σειρών. Πρόκειται για προβλήματα που αποτελούνται από μία μόνο σειρά παρατηρήσεων και απαιτείται ένα μοντέλο που μαθαίνει από τη σειρά παρελθοντικών παρατηρήσεων για να προβλέψει την επόμενη τιμή της ακολουθίας.

Ένα Vanilla LSTM είναι ένα μοντέλο LSTM που έχει ένα μόνο κρυφό στρώμα μονάδων LSTM και ένα στρώμα εξόδου που χρησιμοποιείται για να κάνει μια πρόβλεψη. Στην εφαρμογή του μοντέλου LSTM χρησιμοποιήθηκε η βιβλιοθήκη Keras από το πλαίσιο TensorFlow. Ορίσουμε ένα μοντέλο Vanilla LSTM για τη μόνο-μεταβλητή πρόβλεψη χρόνο-σειρών ως εξής.

```
#define model
model = Sequential()
model.add(LSTM(100, input_shape=(train_X.shape[1], train_X.shape[2])))
model.add(Dropout(0.02))
# model.add(Dropout(0.3))
model.add(Dense(1))
model.compile(loss='mean_squared_error', optimizer='adam')
```

Το κλειδί στον ορισμό είναι η μορφή της εισόδου, δηλαδή τι αναμένει το μοντέλο ως είσοδο για κάθε δείγμα όσον αφορά τον αριθμό των χρονικών βημάτων και τον αριθμό των χαρακτηριστικών. Εργαζόμαστε με μια μονό-μεταβλητή σειρά, οπότε ο αριθμός των χαρακτηριστικών είναι ένας, για μια μεταβλητή. Το σχήμα της εισόδου για κάθε δείγμα καθορίζεται στο όρισμα input_shape στον ορισμό του πρώτου κρυμμένου στρώματος. Σχεδόν πάντα έχουμε πολλαπλά δείγματα, επομένως, το μοντέλο θα περιμένει το στοιχείο

εισόδου των δεδομένων εκπαίδευσης να έχει τις διαστάσεις ή το σχήμα: [samples, timesteps, features].

Σε αυτή την περίπτωση, ορίζουμε ένα μοντέλο με 100 μονάδες LSTM στο κρυφό επίπεδο και ένα επίπεδο εξόδου που προβλέπει μια απλή αριθμητική τιμή. Μετά την προετοιμασία και επεξεργασία των δεδομένων εφαρμόστηκε η μέθοδος fit(). Το μοντέλο προσαρμόζεται χρησιμοποιώντας την αποδοτική έκδοση Adam της στοχαστικής καθόδου κλίσης και βελτιστοποιείται χρησιμοποιώντας τη συνάρτηση απώλειας μέσου τετραγωνικού σφάλματος ή "mse".

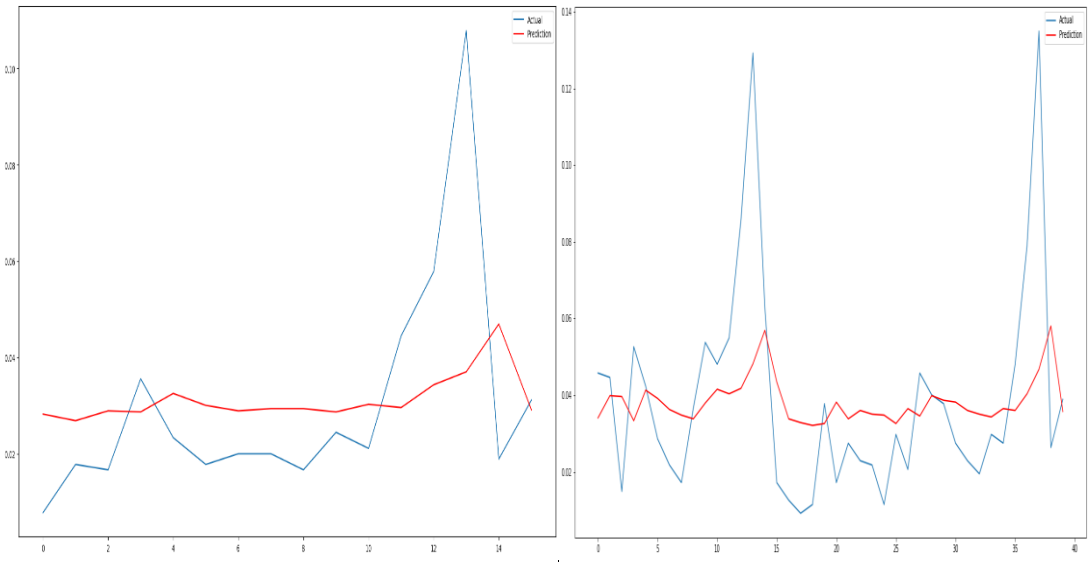
```
model.fit(train_X, train_y, epochs=5, batch_size=16,
validation_data=(test_X,
  test_y), verbose=2, shuffle=False)
```

Αφού προσαρμοστεί το μοντέλο, μπορούμε να το χρησιμοποιήσουμε για να κάνουμε πρόβλεψη, όπως φαίνεται στον παρακάτω κώδικα:

```
# make a prediction
  yhat = model.predict(test_X)
  test_X = test_X.reshape((test_X.shape[0], 1))
  # invert scaling for forecast
  inv_yhat = np.concatenate((yhat, test_X), axis=1)
  inv_yhat = scaler.inverse_transform(inv_yhat)
  inv_yhat = inv_yhat[:,0]
  # invert scaling for actual
  test_y = test_y.reshape((len(test_y), 1))
  inv_y = np.concatenate((test_y, test_X), axis=1)
  inv_y = scaler.inverse_transform(inv_y)
  inv_y = inv_y[:,0]
  rmse = np.mean_squared_error(np.array([inv_y[-1]]), np.array([inv_yhat[-1]]
 print('Test RMSE: %.3f' % rmse)
 print("Test MAE : %.3f "% mae)
```

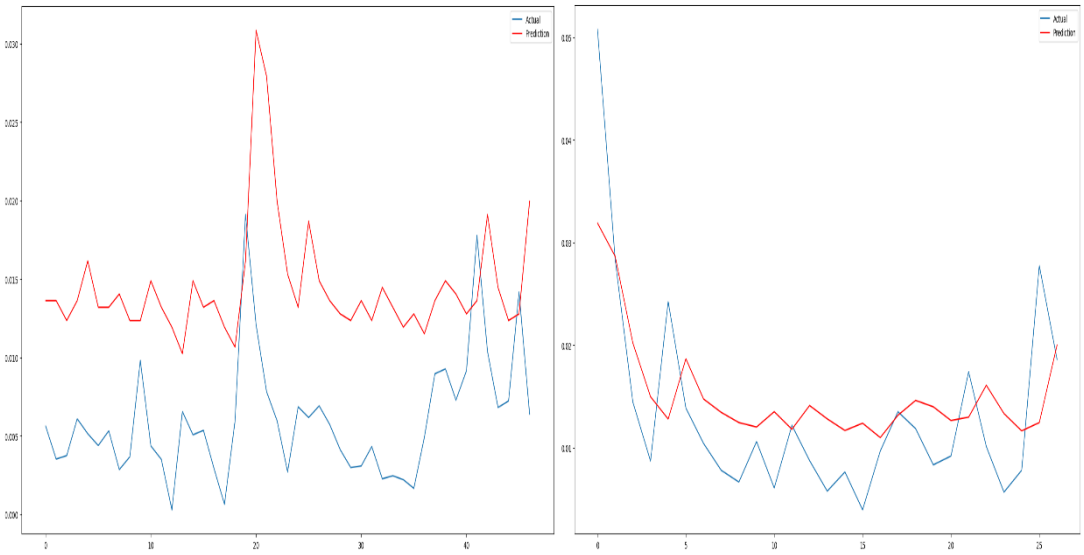
Το σύνολο δεδομένων που εφαρμόστηκε είναι το ίδιο όπως και για τα προηγούμενα μοντέλα. Τα αποτελέσματα που προέκυψαν με τη χρήση του μοντέλου LSTM παρουσιάζονται στις εικόνες 5.13 & 5.14. Από τα σχήματα είναι αρκετά σαφές το μοντέλο LSTM σε γενικές γραμμές προβλέπει με κάποια ακρίβεια της συγκεντρώσεις PM 2.5, PM 10, AQI, καθώς υπάρχουν διαστήματα με αποκλήσει των προβλεπόμενων τιμών από τις πραγματικές. Τα στατιστικά στοιχεία σφάλματος πρόβλεψης για τις συγκεντρώσεις των σωματιδίων, και του δείκτη ποιότητας του αέρα, παρουσιάζονται στον πίνακα 5.6.

Στην εικόνα 5.13 απεικονίζεται το διάγραμμα διασποράς ωριαίας πρόβλεψης των αιωρούμενων σωματιδίων PM 2.5 & PM10 μεταξύ πραγματικών και προβλεπόμενων τιμών του LSMT μοντέλου, για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



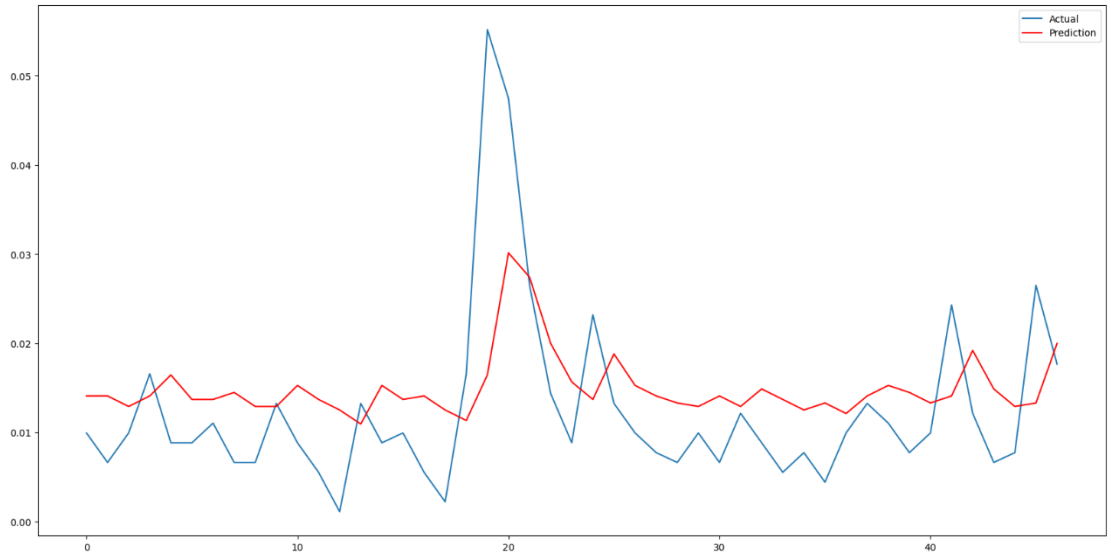
Εικόνα 5.13: Ωριαία Πρόβλεψη αιωρούμενων σωματιδίων PM 2.5, PM 10 με χρήση του μοντέλου πρόβλεψης LSTM.

Στην εικόνα 5.13 απεικονίζεται το διάγραμμα διασποράς ημερήσιας πρόβλεψης των αιωρούμενων σωματιδίων PM 2.5 & PM10 μεταξύ πραγματικών και προβλεπόμενων τιμών του LSTM μοντέλου, για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



Εικόνα 5.14: Ημερήσια Πρόβλεψη αιωρούμενων σωματιδίων PM 2.5, PM 10 με χρήση του μοντέλου πρόβλεψης LSTM.

Στην εικόνα 5.15 απεικονίζεται το διάγραμμα διασποράς του δείκτη ποιότητας αέρα μεταξύ πραγματικών και προβλεπόμενων τιμών του LSTM μοντέλου για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



Εικόνα 5.15: Πρόβλεψη του Δίκτη Ποιότητας του Αέρα(AQI), με τη χρήση του μοντέλου πρόβλεψης LSTM.

Μετρικές	**PM 2.5**		**PM 10**		**AQI**
MAE	2.791	4.838	2.726	4.611	2.354
RMSE	3.877	5.870	2.678	5.798	2.678

Πίνακας 5.7: σύγκριση μετρικών αξιολόγησης παραμέτρων PM 2.5, PM10(πρόβλεψη επόμενης ώρας και 24 ωρών), AQI για το μοντέλο LSTM, για ολόκληρο το σύνολο δεδομένων δοκιμής.

5.8 Convolutional Neural Network

Εφαρμόζουμε μονό-μεταβλητή χρόνο-σειρά οπότε ο αριθμός των χαρακτηριστικών είναι ένα, για μια μεταβλητή. Τα CNN δεν θεωρεί ότι τα δεδομένα έχουν χρονικά βήματα, αλλά τα αντιμετωπίζει ως μια ακολουθία πάνω στην οποία μπορούν να εκτελεστούν πράξεις συνελικτικής ανάγνωσης, όπως μια μονοδιάστατη εικόνα. Ορίζουμε ένα στρώμα συνελικτικής ανάλυσης με 64 φίλτρα και μέγεθος πυρήνα 3. Ακολουθεί ένα στρώμα συγκέντρωσης μεγίστων και ένα πυκνό στρώμα για την ερμηνεία του χαρακτηριστικού εισόδου. Ορίζεται ένα στρώμα εξόδου που προβλέπει μια απλή αριθμητική τιμή. . Μετά την προετοιμασία και επεξεργασία των δεδομένων εφαρμόστηκε η μέθοδος fit(). Το μοντέλο προσαρμόζεται χρησιμοποιώντας την αποδοτική έκδοση Adam της στοχαστικής καθόδου κλίσης και βελτιστοποιείται χρησιμοποιώντας τη συνάρτηση απώλειας μέσου τετραγωνικού σφάλματος ή "mse". Αφού οριστεί το μοντέλο, μπορούμε να το προσαρμόσουμε στο σύνολο δεδομένων εκπαίδευσης.

```
fit(X_train , y_train, epochs=100, batch_size=32,validation_data =
```

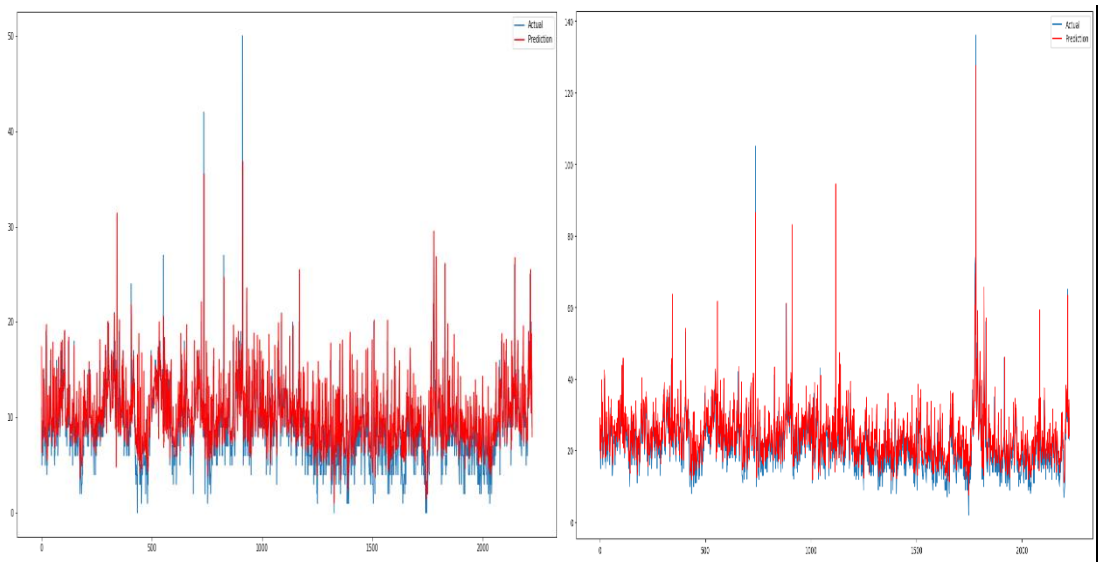
Εόσον προσαρμοστεί το μοντέλο, μπορούμε να το χρησιμοποιήσουμε για να κάνουμε πρόβλεψη:

```
pred = model.predict(X_test)
 rmse = np.sqrt(mean_squared_error(y_test,pred.reshape(len(pred))))
 mae = mean_absolute_error(y_test, pred.reshape(len(pred)))
 print('Test RMSE: %.3f' % rmse)
 print("Test MAE : %.3f "% mae)
```

Το σύνολο δεδομένων που εφαρμόστηκε είναι το ίδιο όπως και για τα προηγούμενα μοντέλα. Τα αποτελέσματα που προέκυψαν με τη χρήση του μοντέλου CNN παρουσιάζονται στις εικόνες 5.16 – 5.17 – 5.18 . Από τα σχήματα είναι αρκετά σαφές το CNN προβλέπει με μεγάλη ακρίβεια της συγκεντρώσεις PM 2.5, PM 10, AQI, οι προβλεπόμενες τιμές είναι παρόμοιες με τις πραγματικές, και παράγει σταθερή απόδοση.

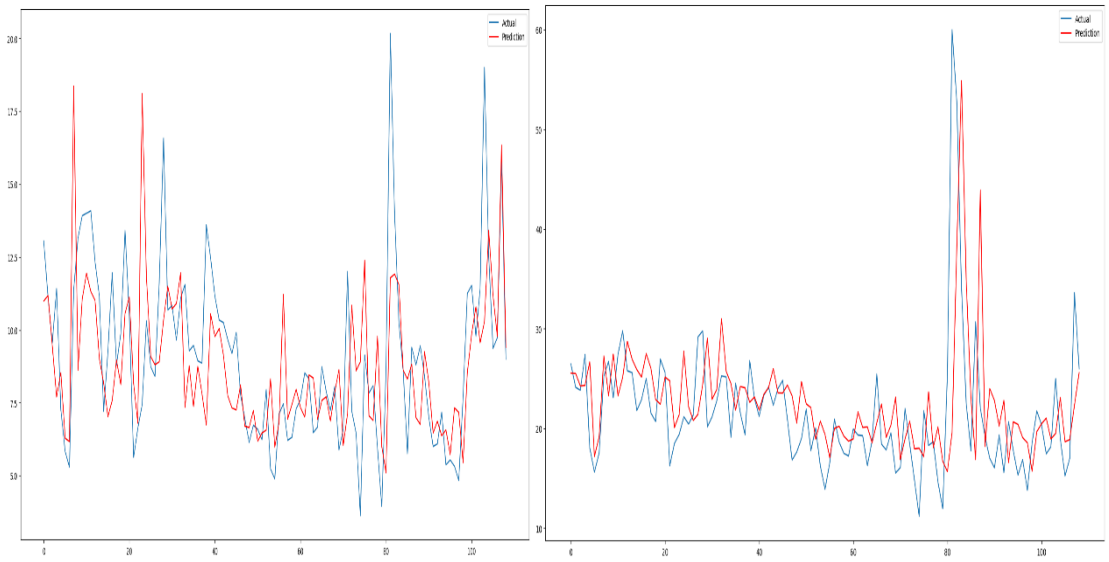
Τα στατιστικά στοιχεία σφάλματος πρόβλεψης για τις συγκεντρώσεις των σωματιδίων, και του δείκτη ποιότητας του αέρα, παρουσιάζονται στον πίνακα 5.8..

Στην εικόνα 5.16 απεικονίζεται το διάγραμμα διασποράς ωριαίας πρόβλεψης των αιωρούμενων σωματιδίων PM 2.5 & PM10 μεταξύ πραγματικών και προβλεπόμενων τιμών του CNN μοντέλου, για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



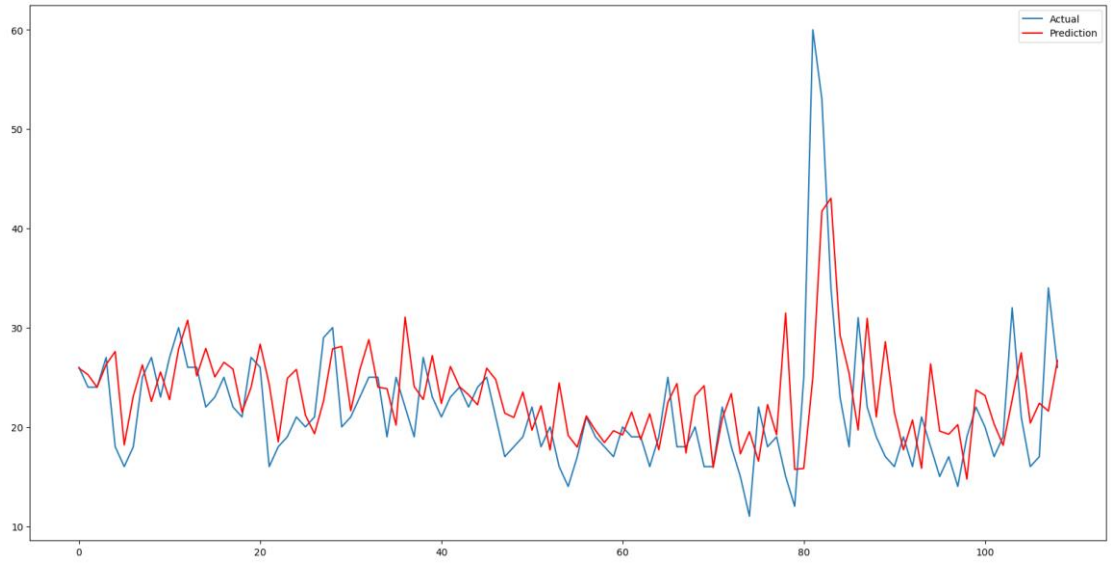
Εικόνα 5.16: Ωριαία Πρόβλεψη αιωρούμενων σωματιδίων PM 2.5, PM 10 με χρήση του μοντέλου πρόβλεψης CNN.

Στην εικόνα 5.17 απεικονίζεται το διάγραμμα διασποράς ημερήσιας πρόβλεψης των αιωρούμενων σωματιδίων PM 2.5 & PM10 μεταξύ πραγματικών και προβλεπόμενων τιμών του CNN μοντέλου, για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



Εικόνα 5.17: Ημερήσια Πρόβλεψη αιωρούμενων σωματιδίων PM 2.5, PM 10 με χρήση του μοντέλου πρόβλεψης CNN.

Στην εικόνα 5.18 απεικονίζεται το διάγραμμα διασποράς του δείκτη ποιότητας αέρα μεταξύ πραγματικών και προβλεπόμενων τιμών του CNN μοντέλου για ολόκληρο το σύνολο δεδομένων δοκιμής του δείκτη ποιότητας αέρα. Η Μπλε γραμμή δείχνει τις παρατηρούμενες τιμές και η πορτοκαλί γραμμή τις παραγόμενες προβλέψεις.



Εικόνα 5.18: Πρόβλεψη του Δίκτη Ποιότητας του Αέρα(AQI), με τη χρήση του μοντέλου πρόβλεψης CNN.

Μετρικές	**PM 2.5**		**PM 10**		**AQI**
MAE	2.704	1.864	4.756	4.226	4.305
RMSE	3.719	2.682	6.999	6.649	6.122

Πίνακας 5.8: σύγκριση μετρικών αξιολόγησης παραμέτρων PM 2.5, PM10(πρόβλεψη επόμενης ώρας και 24 ωρών), AQI για το μοντέλο CNN, για ολόκληρο το σύνολο δεδομένων δοκιμής.

5.9 Συγκριτικά αποτελέσματα

Στην παρουσα ενότητα παρουσιάζουμε μια διάφορα είδη μοντέλων μηχανικής μάθησης για να συγκρίνουμε τις επιδόσεις τους στις προβλέψεις δύο ατμοσφαιρικών ρύπων και της ποιότητας του αέρα. Παρουσιάζεται πρώτα μια σύγκριση της ικανότητας του μοντέλου να προβλέπει κανονικά πρότυπα ατμοσφαιρικών ρύπων και, δεύτερον, την ακρίβεια του μοντέλου για το δείκτη ποιότητας του αέρα. Εφαρμόζουμε οκτώ τεχνικές πρόβλεψης, η καθεμία με το δικό της μοναδικό χαρακτηριστικό, οι οποίες αναγνωρίστηκαν ως δυνητικά πλεονεκτικές προσεγγίσεις για την πρόβλεψη της ποιότητας του αέρα.

Συγκρίνονται τα αποτελέσματα των προβλέψεων όλων των μοντέλων μηχανικής μάθησης. Τα μοντέλα εκπαιδεύονται στα ίδια δεδομένα εκπαίδευσης και δοκιμής για τις συγκεντρώσεις των ρύπων PM 2.5, PM 10 και διαφορετικό για AQI, πίνακας 6.1. Η αξιολόγηση των αποτελεσμάτων παρουσιάζεται σε τρία μέρη: Το πρώτο περιλαμβάνει το σφάλμα παλινδρόμησης των προβλέψεων μια ώρα μπροστά, για τα σωματίδια PM 2.5, PM 10 πίνακα 5.9. Το σφάλμα περιλαμβάνει MAE, RMSE για κάθε ρύπο Η δεύτερη αξιολόγηση παρουσιάζει τα αποτελέσματα των ημερίσιων προβλέψεων για σωματίδια PM 2.5 - PM 10 που βρίσκονται στον Πίνακα 5.10. Τέλος, η πρόβλεψη του δείκτη ποιότητας του αέρα παρουσιάζονται στον πίνακα 5.11.

Μοντέλα	**PM 2.5**		**PM 10**	
	RMSE	**MAE**	**RMSE**	**MAE**
ARIMA	1.875	0.596	1.256	0.198
Theta	4.159	3.037	11.102	9.394
FFT	7.363	5.956	10.180	6.116
RFR	1.437	4.687	0.728	4.711
SVR	0.839	0.678	4.314	0.287
MLP	3.877	5.781	2.891	4.756
LSTM	3.877	2.791	2.678	2.726
CNN	3.719	2.704	6.999	4.756

Πίνακας 5.9 : Αποτελέσματα Ωριαίας Πρόβλεψης PM 2.5, PM10.

Στον πίνακα 5.9 παρουσιάζονται τα αποτελέσματα του πειράματος. Διαπιστώσαμε ότι το μοντέλο που κατασκευάστηκε με τον αλγόριθμο SVM έχει τα μικρότερα MAE και RMSE. Το FFT και τα υψηλότερα μέτρα απόδοσης MAE και RMSE.

Μοντέλα	**PM 2.5**		**PM 10**	
	RMSE	**MAE**	**RMSE**	**MAE**
ARIMA	5.547	4.382	6.785	5.682
Theta	2.836	2.253	6.536	4.017
FFT	6.464	5.143	8.331	6.116
RFR	7.731	20.559	11.956	5.132
SVR	4.345	3.965	5.346	4.272
MLP	3.211	2.711	6.365	4.464
LSTM	5.870	4.838	5.798	4.611
CNN	2.682	1.864	6.649	4.226

Πίνακας 5.10 : Αποτελέσματα ημερήσιας Πρόβλεψης PM2.5, PM 10.

Στον πίνακα 5.10 παρουσιάζονται τα αποτελέσματα του πειράματος. Διαπιστώσαμε ότι το μοντέλο που κατασκευάστηκε με τον αλγόριθμο CNN έχει τα μικρότερα μέτρα απόδοσης MAE και RMSE, ενώ το μοντέλο RFR και τα υψηλότερα μέτρα απόδοσης MAE και RMSE.

Μοντέλα	**AQI**	
	RMSE	**MAE**
ARIMA	7.425	7.732
Theta	2.993	2.271
FFT	10.203	7.939
RFR	7.314	4.687
SVR	8.732	8.165
MLP	6.112	4.835
LSTM	2.678	2.354
CNN	6.122	4.305

Πίνακας 5.11 : Αποτελέσματα Πρόβλεψης AQI

Στον πίνακα 5.11 παρουσιάζονται τα αποτελέσματα του πειράματος. Διαπιστώσαμε ότι το μοντέλο που κατασκευάστηκε με τον αλγόριθμο LSMT έχει τα μικρότερα μέτρα απόδοσης MAE και RMSE, ενώ το μοντέλο FFT και τα υψηλότερα μέτρα απόδοσης MAE και RMSE.

Επίλογος

6.1 Συμπεράσματα

Στην παρούσα εργασία, παρουσιάζονται τα προτεινόμενα μοντέλα μηχανικής μάθησης για την ανάλυση της ατμοσφαιρικής ρύπανσης σε αστική περιοχή της Ελλάδας. χρησιμοποιήσει την ανοικτή βάση δεδομένων openaq.org. Η πρόβλεψη των αιωρούμενων σωματιδίων PM2.5 και PM 10, και του δείκτη ποιότητας του αέρα γίνεται με τη χρήση μοντέλων μηχανικής μάθησης με βάση τους στατιστικούς υπολογισμούς των μετρικών όπως MAE, RMSE.

Μελετήθηκαν μια σειρά από βασικά στοιχεία που σχετίζονται με την πρόβλεψη της ποιότητας του αέρα, συμπεριλαμβανομένης της ανάλυσης των χωροχρονικών προτύπων, της διερεύνησης των βασικών μεταβλητών, των μεθόδων μηχανικής μάθησης για την πρόβλεψη χρονοσειρών και παρόμοιων προκλήσεων στον τομέα της ποιότητας του αέρα. Εφαρμόστηκαν πολυάριθμες τεχνικές, που κυμαίνονται από στατιστικές προσεγγίσεις έως πιο πρόσφατες εξελίξεις στη μηχανική μάθηση. Τα βαθιά νευρωνικά δίκτυα Πολυεπίπεδα Perceptron, τα Δίκτυα Μακράς Βραχίας Μνήμης(LSTM) και τα Συνελικτικά Νευρωνικά Δίκτυα είναι οι αρχιτεκτονικές που παρατηρούνται περισσότερο στη βιβλιογραφία. Η βαθιά μάθηση έχει αποδείξει την υψηλή απόδοση της εκμάθησης κρυφών σχέσεων πολύπλοκων προβλημάτων, ενώ η πιο εξειδικευμένη αρχιτεκτονική συνελλικτικών δικτύων, LSTM και MLP, έχει αποδειχθεί πολύτιμο εργαλείο για την πρόβλεψη χρονοσειρών.

Στόχος της παρούσας εργασίας είναι η αξιολόγηση της απόδοσης των μεθόδων μηχανικής μάθησης για την πρόβλεψη της ποιότητας σε ελληνικό αστικό περιβάλλον. Ξεκινήσαμε με την ανάλυση συνόλων δεδομένων της συγκεκριμένης αστικής περιοχής, συμπεριλαμβανομένων των ατμοσφαιρικών ρύπων. Περαιτέρω, δημιουργήσαμε περισσότερα χαρακτηριστικά με στατιστική επιλογή χαρακτηριστικών και δοκιμάσαμε πολλαπλές σύγχρονες τεχνικές μηχανικής μάθησης όπως βαθιά μάθηση. Υλοποιήθηκαν, βελτιστοποιήθηκαν, εκπαιδεύτηκαν και δοκιμάστηκαν διάφορα μοντέλα μηχανικής και στατιστικής μάθησης για να προσδιοριστούν τα δυνατά και αδύνατα σημεία της πρόβλεψης της ποιότητας του αέρα. Η μεθοδολογία και οι λεπτομέρειες υλοποίησης των συγκεκριμένων μοντέλων παρουσιάζονται στο κεφάλαιο 4.

Στην εφαρμογή των μοντέλων μηχανικής μάθησης δείξαμε ότι το Arima με την SVM έχουν την καλύτερη απόδοση στην πρόβλεψη της συνολικής ποιότητας του αέρα για όλους τους

ρύπους που μελετήθηκαν (PM2.5, PM10). Επιπλέον, διαπιστώσαμε ότι η τα μοντέλα βαθιάς μάθησης παρουσιάζουν μεθόδους που δείχνουν ότι οι πραγματικές τιμές και οι προβλεπόμενες τιμές είναι πολύ κοντά η μία στην άλληκαι έχουν καλυτερη απόδοση στην ημερήσια πρόβλεψη(CNN) και στο δείκτη ποιότητας αέρα(LSTM) . Όλα τα μοντέλα μηχανικής μάθησης εξαρτώνται σε μεγάλο βαθμό από τις υπερπαραμέτρους τους. Μια μικρή αλλαγή μπορεί να επηρεάσει τα αποτελέσματα προς οποιαδήποτε κατεύθυνση της απόδοσης. Στην παρούσα εργασία, καταβάλλεται προσπάθεια για τη ρύθμιση των παραμέτρων κάθε συγκεκριμένου μοντέλου. Επιπλέον, η αναζήτηση υπερπαραμέτρων που χρησιμοποιείται είναι μια απλή προσέγγιση και θα μπορούσε να χρησιμοποιηθεί με πιο κατάλληλη μέθοδος. Τα μοντέλα νευρωνικών δικτύων βαθιάς μάθησης είναι πιο ευάλωτα σε αλλαγές παραμέτρων από άλλα και είναι επιθυμητό να εισαχθούν πιο εύρωστες αρχιτεκτονικές για την αποφυγή αυτού του προβλήματος. Η ασταθής εκπαίδευση ωστόσο, αποτελεί πρόβλημα για τις περισσότερες τεχνικές βαθιάς μάθησης και είναι ένα συνεχές πεδίο έρευνας. Συμπερασματικά, η παρούσα εργασία διαπίστωσε ότι υπάρχουν πολλαπλοί αξιόλογοι αλγόριθμοι μηχανικής μάθησης για την πρόβλεψη της ποιότητας του αέρα. Μέσω του συνδυασμού των δεδομένων των ρύπων, με μια στατιστική τεχνική σχεδιασμού χρονικών-χωρικών χαρακτηριστικών, παρέχουμε ένα υψηλότερο επίπεδο πληροφοριών των δεδομένων που χρησιμοποιούνται για τα μοντέλα μηχανικής μάθησης. Η προσέγγιση πρόβλεψης μονό-μεταβλητής λειτουργεί καλά σε συνδυασμό με το εκτεταμένο σύνολο χαρακτηριστικών. Οι επιλεγμένες μέθοδοι παρουσιάζουν υψηλές επιδόσεις για την πρόβλεψη ξεχωριστών ρύπων με διάφορους ορίζοντες πρόβλεψης. Παρουσιάζουμε αποτελέσματα που δείχνουν ότι η μέθοδος ξεπέρασε κάθε μοντέλο στις περισσότερες περιπτώσεις.

6.2 Μελλοντικές επεκτάσεις

Ως μελλοντική εργασία, είναι η επιλογή συνόλου δεδομένων και μεταβλητών - λαμβάνοντας υπόψη ένα μεγάλο σύνολο δεδομένων με περισσότερες παραμέτρους και μετρήσεις, το οποίο μπορεί να υποστηρίξει ακριβέστερα μοντέλα πρόβλεψης για ατμοσφαιρικούς ρύπους O3, SO2 και NO2. Μπορούμε περαιτέρω να συνθέσουμε δύο ή περισσότερους αλγορίθμους μηχανικής μάθησης και να επεξεργαστούμε μεγάλα δεδομένα για να λάβουμε πιο ακριβή αποτελέσματα.

Βιβλιογραφία

[1] Limb, M. (2016). Half of wealthy and 98% of poorer cities breachair quality guidelines. BMJ: Br. Med. J., 353 (15 pages).

[2] WHO, (2016) Air Pollution Levels Rising in Many of the World's Poorest Cities.

[3] Salnikov, V.G.; Karatayev, M.A., (2011). Impact of air pollution on human health: Focusing on Rudnyi Altay industrial area. Am. J. Environ. Sci., 7(3): 286-294 (9 pages).

[4] Daly, A.; Zannetti, P. (2007). Air pollution modeling--An overview. Ambient air pollut., 15-28 (14 pages).

[5] Brunekreef, B.; Holgate, S. T. (2002). Air pollution and health. The lancet., 360(9341): 1233-1242 (10 pages).

[6] Lelieveld, J.; Evans, J.S.; Fnais, M.; Giannadaki, D.; Pozzer, A., (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature, 525,367 (14 pages).

[7] Brunelli, U.; Piazza, V.; Pignato, L.; Sorbello, F.; Vitabile, S., (2007). Two-days ahead prediction of daily maximum concentrations of SO2, O3, PM10, NO2, CO in the urban area of Palermo, Italy. Atmos. Environ., 41, 2967-2995 (29 pages).

[8] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no.3, pp. 273–297, 1995.

[9] Kitchenham, B. Charters, S. in: E.T. Report (Ed.) (2007),., Guidelines for Performing Systematic Literature *Reviews in Software Engineering, Engineering 2 EBSE* 2007-001 1051.

[10] S. S. Ganesh, S. H. Modali, S. R. Palreddy, and P. Arulmozhivarman, "Fore- casting air quality index using regression models: στο *2017 International Conference on Trends in Electronics and Infor- matics (ICEI)*, 2017, pag. 248-254.

[11] Huixiang Liu 1 , Qing Li 1 , Dongbing Yu 2 and Yu Gu. Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms. *Applied*

Sciences 2019.

[12] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, "A machine learning approach to predict air quality in california," *Complexity*, vol. 2020, 2020.

[13] Bing-Chun Liu, Arihant Binaykia , Pei-Chann Chang, Manoj Kumar Tiwari , ChengChin Tsao.Urban air quality forecasting based on multidimensional collaborative Support Vector Regression (SVR): A case study of BeijingTianjin-Shijiazhuang. *Urban air quality forecasting* PLOS ONE | https://doi.org/10.1371/journal.pone.0179763.

[14] Khaled Bashir Shaban, Abdullah Kadri, and Eman Rezk. Urban Air Pollution Monitoring System With Forecasting Models. *IEEE SENSORS JOURNAL, VOL. 16,* NO. 8, APRIL 15, 2016.

[15] B.S. Freeman, G. Taylor, B. Gharabaghi, J. Thé, Forecasting air quality time series using deep learning, J. Air Waste Manage. Assoc.68 (8) (2018) 866-886.

[16] I. Kok, M. Simsek, S. Ozdemir, A deep learning model for air quality prediction in smart cities, in: IEEE International Conference on Big Data, 2017, p. 1983-1990.

[17] Rishanti Murugan, Naveen Palanichamy. Smart City Air Quality Prediction using Machine Learning. *Proceedings of the Fifth International Conference on Intelligent* Computing and Control Systems (ICICCS 2021).

[18] P. Nath, P. Saha, A. Middya, S. Roy, Long-term time-series pollution forecast usingstatisticalanddeeplearningmethods,NeuralComput.Appl. (2021).

[17] Bihter Das, Ömer Osman Dursun, Suat Toraman. Prediction of air pollutants for air quality using deep learning methods in a metropolitan city. *Urban Climate, ISSN: 2212-*0955, Vol: 46, Page: 101291.

[19] S.Patra,Timeseriesforecastingofairpollutantconcentrationlevelsusing machinelearning,Adv.Comput.Sci.Inf.Technol.4(5)(2017)280-284.

[20] K. Kaya, S. Gunduz Oguducu, Deep Flexible Sequential (DFS) model for air pollution forecasting, Sci.Rep.10(2020)1-12.

[21] L. Cao and F. Tay, "Financial forecasting using support vector machines," Neural Computing & Applications, vol. 10, no. 2, 2001a.

[22] P. Hajek and V. Olej, "Predicting common air quality index - the case of Czech microregions," Aerosol and Air Quality Research, vol. 15, no. 2, pp. 544-555, 2015.

[23] S. Capone et al., “Solid state gas sensors: state of the art and future activities,” Journal of Optoelectronics and Advanced Materials, vol. 5, no. 5, 2003.

[24] W. Yi et al., “A survey of wireless sensor network based air pollution monitoring systems,” Sensors, vol. 15, no. 12, 2015.

[25] Press, W.H., Teukolsky, S.A., Vellerling, W.T., Flannery, B.P., 2002. Numerical Recipes in Cþþ: the Art of Scientific Computing. Cambridge University Press, Cambridge.

[26] U. Brunelli, V. Piazza, L. Pignato, F. Sorbello, and S. Vitabile, “&ree hours ahead prevision of SO2 pollutant concentration using an Elman neural based forecaster,” Building and Environment, vol. 43, no. 3, pp. 304–314, 2008.

[27] G. Bontempi, S. Taieb, Y. Le Borgne, and D. Loshin, “Machine learning strategies for time series forecasting,” in Business Intelligence, pp. 59–73, Springer, Berlin, Germany, 2013.

[28] L. A. D´ıaz-Robles, J. C. Ortega, J. S. Fu et al., “A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: the case of Temuco, Chile,”Atmospheric Environment, vol. 42, no. 35, pp. 8331–8340, 2008.

[29] M. Cai, Y. Yin, and M. Xie, “Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach,” Transportation Research Part D: Transport and Environment, vol. 14, no. 1, pp. 32–41, 2009.

[30] Kamalapurkar S (2020) Air pollution forecasting using supervised machine learning. Int J Sci Technol Res 9(4):118-123.

[31**]**. Zhu, D., Cai, C. , Yang, T., Zhou, X.: A machin e lea rn in g approach for air quality predictio n: model regularization and optimization. Big Data Cognit. Comput. 2(1) (2018).

Υπεύθυνη Δήλωση Συγγραφέα:

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν.1599/1986, η παρούσα εργασία αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε

που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης.