ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ – WEB INTELLIGENCE

Εφαρμογή μοντέλων μηχανικής μάθησης για πρόβλεψη βροχόπτωσης σε πραγματικό χρόνο χρησιμοποιώντας αριθμητικά μοντέλα και δεδομένα επίγειων παρατηρήσεων

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

των

Κύρου Γεώργιου & Μανώλα Ιωάννη

Επιβλέπων : Κωνσταντίνος Ι. Διαμαντάρας
Καθηγητής, Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων, Διεθνές Πανεπιστήμιο Ελλάδος

Θεσσαλονίκη, Ιούνιος 2023

ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ – WEB INTELLIGENCE

Εφαρμογή μοντέλων μηχανικής μάθησης για πρόβλεψη βροχόπτωσης σε πραγματικό χρόνο χρησιμοποιώντας αριθμητικά μοντέλα και δεδομένα επίγειων παρατηρήσεων

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

των

Κύρου Γεώργιου & Μανώλα Ιωάννη

Επιβλέπων : Κωνσταντίνος Ι. Διαμαντάρας
Καθηγητής, Διεθνές Πανεπιστήμιο Ελλάδος

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 30 Ιουνίου 2023.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Όνομα Επώνυμο	Όνομα Επώνυμο	Όνομα Επώνυμο
Choose an item. ΔΙ.ΠΑ.Ε.	Choose an item. ΔΙ.ΠΑ.Ε.	Choose an item. ΔΙ.ΠΑ.Ε.

Θεσσαλονίκη, Ιούνιος 2023

(Υπογραφή)

.....

Κύρος Γεώργιος

Πληροφορική Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης

(Υπογραφή)

.....

Μανώλας Ιωάννης

Εφαρμοσμένη Πληροφορική Πανεπιστημίου Μακεδονίας

Περίληψη

Η εφαρμογή αλγορίθμων μηχανικής μάθησης σε μεγάλα σύνολα δεδομένων στον τομέα της Μετεωρολογίας βρίσκεται στο απόγειο της τα τελευταία χρόνια. Σήμερα, λόγω της κλιματικής αλλαγής αλλά και των ακραίων καιρικών συνθηκών, κρίνεται αναγκαία η πρόβλεψη με ακρίβεια της ποσότητας της βροχόπτωσης (ή η ένταση της βροχής) για τη σωστή λήψη μέτρων προστασίας της ζωής και των περιουσιών. Η παρούσα εργασία σκοπεύει να εμβαθύνει στην ανάλυση των μετεωρολογικών δεδομένων με την χρήση τεχνικών μηχανικής μάθησης για τη βελτίωση της ικανότητας στην πρόβλεψη βροχοπτώσεων σε πραγματικό χρόνο. Για το σκοπό αυτό, αναλύονται οι σχέσεις μεταξύ θερμοδυναμικών παραμέτρων που προέρχονται από δορυφορικές μετρήσεις και καταγεγραμμένων βροχοπτώσεων από επίγειους μετεωρολογικούς σταθμούς, πάντα σε συνδυασμό με τα αποτελέσματα του αριθμητικού ατμοσφαιρικού μοντέλου. Ο κύριος σκοπός της εργασίας είναι να βρεθούν οι καταλληλότερες σχέσεις μεταξύ των ατμοσφαιρικών συνθηκών και του σχηματισμού νεφώσεων που οδηγούν στην παραγωγή βροχοπτώσεων και να δημιουργηθεί ένα μοντέλο μηχανικής μάθησης για την πρόβλεψη της βροχόπτωσης σε πραγματικό χρόνο. Σε αυτή την εργασία χρησιμοποιήθηκαν διάφορες μέθοδοι και τεχνικές μηχανικής μάθησης, όπως Παλινδρόμηση (Regression), Μηχανική μάθηση σε σύνολα (Ensemble Machine Learning) και Βαθιά μάθηση (Deep Learning). Τέλος, τα αποτελέσματά τους συγκρίνονται προκειμένου να καταλήξουμε στα μοντέλα που ταιριάζουν καλύτερα στο συγκεκριμένο πρόβλημα.

Λέξεις Κλειδιά: « Μηχανική μάθηση, μετεωρολογικά δεδομένα, πρόβλεψη βροχόπτωσης, τυχαίο δάσος, βαθιά μάθηση, νευρωνικά δίκτυα »

Abstract

The application of machine learning (ML) algorithms in large datasets in the field of Meteorology is at the forefront of research. In this context, the use of satellite data to estimate the amount of rainfall is an important field of research, with operational applications. It is important to accurately predict the amount of rainfall (or rain rate) in a particular area for the proper taking of life and property protection measures. The present work intends to deepen the analysis of meteorological data with ML techniques to improve our capacity in short-range forecasting of rainfall. To this end, relationships between thermodynamic parameters derived by satellite measurements and recorded rainfall by in-situ gauges, along with outputs from a numerical atmospheric model are analyzed. The main purpose of the work is to find the best relationships between the atmospheric conditions and the formation of clouds that lead to production of rainfall and build a ML model for nowcasting of rainfall. Several ML methods are used i.e., Multiple Linear Regression, Ensemble Machine Learning, and Deep Learning, and their results are compared in order to find the best fit model.

Keywords: « Machine learning, meteorological data, rainfall prediction, random forest, deep learning, neural networks »

Ευχαριστίες

Η ολοκλήρωση αυτής της εργασίας θα ήταν αδύνατη δίχως την πολύτιμη βοήθεια κάποιων ανθρώπων στους οποίους θα θέλαμε να εκφράσουμε τις θερμότερες ευχαριστίες μας.

Πρώτον από όλους, νιώθουμε την ανάγκη να ευχαριστήσουμε τον επιβλέπων καθηγητή μας, κ. Διαμαντάρα Κωνσταντίνο, ο οποίος μας εμπιστεύτηκε, και μας βοήθησε με αμείωτο ενδιαφέρον καθ' όλη τη διάρκεια υλοποίησης της παρούσας μεταπτυχιακής εργασίας.

Θερμές ευχαριστίες θα θέλαμε να αποδώσουμε επίσης στους γονείς μας, και τους ανθρώπους μας, οι οποίοι μας στήριξαν συνεχώς κατά την διάρκεια αυτής μας της προσπάθειας.

Τέλος, ένα μεγάλο ευχαριστώ και την ευγνωμοσύνη μας στους ερευνητές του Εθνικού Αστεροσκοπείου Αθηνών της μονάδας ΜΕΤΕΟ, και ιδιαίτερα στο Διευθυντή ερευνών κ. Λαγουβάρδο Κωνσταντίνο, τη Διευθύντρια ερευνών κα. Κοτρώνη Βασιλική και Δρ. Ντάφη Σταύρο, για τις κατευθύνσεις τους και για την παροχή δεδομένων.

Πρόλογος

Η εργασία αυτή αποτελεί το τελικό στάδιο ενός μεγάλου και τέλειου κύκλου για το μεταπτυχιακό μας στο Διεθνές Πανεπιστήμιο της Ελλάδος με τίτλο Ευφυείς Τεχνολογίες Διαδικτύου (Web Intelligence). Το θέμα της παρούσας εργασίας επιλέχθηκε σε συμφωνία με τον καθηγητή μας και επιβλέπων αυτής της εργασίας κ. Διαμαντάρα Κωνσταντίνο, μετά από πρότασή μας η οποία συζητήθηκε και αναπτύχθηκε με τον Διευθυντή ερευνών στο Εθνικό Αστεροσκοπείο Αθηνών/ΜΕΤΕΟ, κ. Λαγουβάρδο Κωνσταντίνο.

Τίτλος της εργασίας είναι «Εφαρμογή μοντέλων μηχανικής μάθησης για πρόβλεψη βροχόπτωσης σε πραγματικό χρόνο χρησιμοποιώντας αριθμητικά μοντέλα και δεδομένα επίγειων παρατηρήσεων». Κύριος σκοπός της εργασίας ήταν η καλύτερη και ακριβέστερη πρόβλεψη της βροχής για τις αμέσως επόμενες 3 ώρες με την χρήση τεχνικών και μοντέλων μηχανικής μάθησης.

Για την ολοκλήρωση της εργασίας δρομολογήθηκαν και εφαρμόστηκαν πολλά βήματα και ενέργειες εργασίας. Όλες οι ενέργειες και εργασίες που διενεργήθηκαν σε αυτή την εργασία αναγράφονται σε τίτλους στα περιεχόμενα καθώς και αναλυτικά παρακάτω.

Πίνακας περιεχομένων

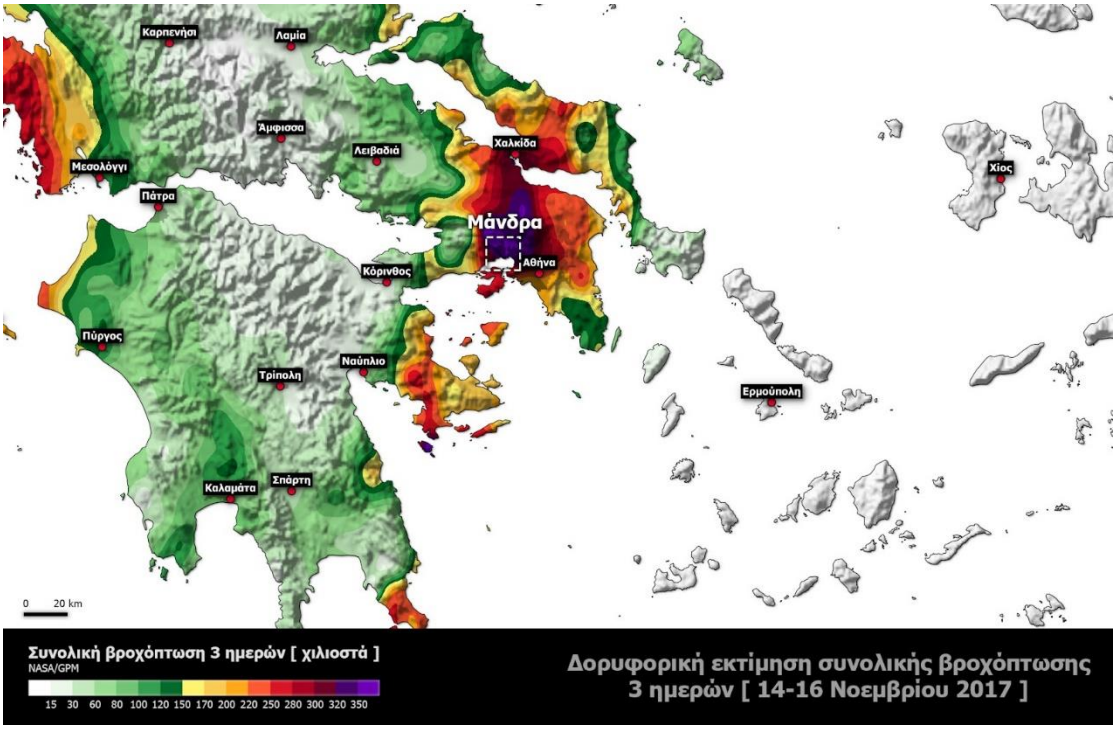
1

Εισαγωγή

Στις μέρες μας, η ανάπτυξη της τεχνολογίας και η αύξηση του όγκου των πληροφοριών καθιστούν την ανάλυση δεδομένων μια πολύπλοκη διαδικασία. Οι εξελισσόμενες προκλήσεις του χειρισμού του όγκου και της πολυπλοκότητας των μεγάλων δεδομένων έχουν οδηγήσει σε αρκετές μελέτες που επικεντρώνονται στην εφαρμογή μοντέλων μηχανικής μάθησης [1]. Η πρόγνωση καιρού βασίζεται σε μεγάλα σύνολα δεδομένων που προέρχονται από ραντάρ, δορυφόρους και επίγειες μετρήσεις, απαιτώντας έτσι αποτελεσματικές τεχνολογίες αποθήκευσης, επεξεργασίας και εξόρυξης δεδομένων [2].

Η βροχόπτωση αποτελεί ένα πολύ βασικό και σημαντικό καιρικό φαινόμενο με μεγάλες επιπτώσεις στο περιβάλλον και στην κοινωνία. Οι βροχοπτώσεις επηρεάζουν πολλές πτυχές των ανθρώπινων δραστηριοτήτων, όπως η γεωργική παραγωγή, οι κατασκευές, ηλεκτροπαραγωγή, δασοκομία και τουρισμός, μεταξύ άλλων [3]. Μεταξύ 1970 και 2019, το 44% όλων των καταστροφών και το 31% όλων των οικονομικών απωλειών παγκοσμίως σχετίζονταν με ακραίες βροχοπτώσεις και πλημμύρες [4]. Οι πλημμύρες και παρόμοια καιρικά φαινόμενα με μεγάλες επιπτώσεις αναμένεται να συμβαίνουν συχνότερα σε όλο τον κόσμο τα επόμενα χρόνια. Σύμφωνα με την τελευταία έκθεση της Διακυβερνητικής Επιτροπής για την Κλιματική Αλλαγή (IPCC), η ένταση των ακραίων βροχοπτώσεων έχει αυξηθεί τα τελευταία χρόνια σε πολλές περιοχές, συμπεριλαμβανομένων των περισσότερων περιοχών της Ευρώπης [5]. Ιδιαίτερα στην Ελλάδα, οι βροχοπτώσεις είναι ένα πολύ σημαντικό φαινόμενο, καθώς η έντασή τους στο παρελθόν έχει προκαλέσει πρωτοφανείς καταστροφές και απώλειες ανθρώπινων ζωών (Εικόνα 1) κατά τη διάρκεια καταστροφικών πλημμυρών [6,7,8].

Κατά συνέπεια, η ακριβής πρόβλεψη της βροχόπτωσης παραμένει μια κρίσιμη πρόκληση λόγω των συνεπειών της. Αρκετοί ατμοσφαιρικοί παράγοντες επηρεάζουν την εμφάνιση και την ένταση των βροχοπτώσεων. Η θερμοκρασία, η υγρασία, η ηλιακή ακτινοβολία, η ατμοσφαιρική πίεση, η μικροφυσική των νεφών είναι μερικοί από τους παράγοντες που επηρεάζουν την εμφάνιση βροχοπτώσεων και την έντασή της [9].



Εικόνα 1. Δορυφορική εκτίμηση της συνολικής βροχόπτωσης που έπεσε στην περιοχή της Μάνδρας στην Αττική το διάστημα 14 με 16 Νοεμβρίου 2017, όπου λόγω των ξαφνικών πλημμυρών χάθηκαν 24 ανθρώπινες ζωές. Πηγή δεδομένων: NASA/GPM [10].

Σε αυτήν την εργασία, εφαρμόσαμε μοντέλα μηχανικής μάθησης, χρησιμοποιώντας ιστορικά δεδομένα μοντέλων καιρού, δορυφορικά δεδομένα και παρατηρήσεις εδάφους από μετεωρολογικούς σταθμούς, με σκοπό την πρόβλεψη της στιγμιαίας βροχόπτωσης εντός των επόμενων 3 ωρών.

1.1 Αντικείμενο διπλωματικής

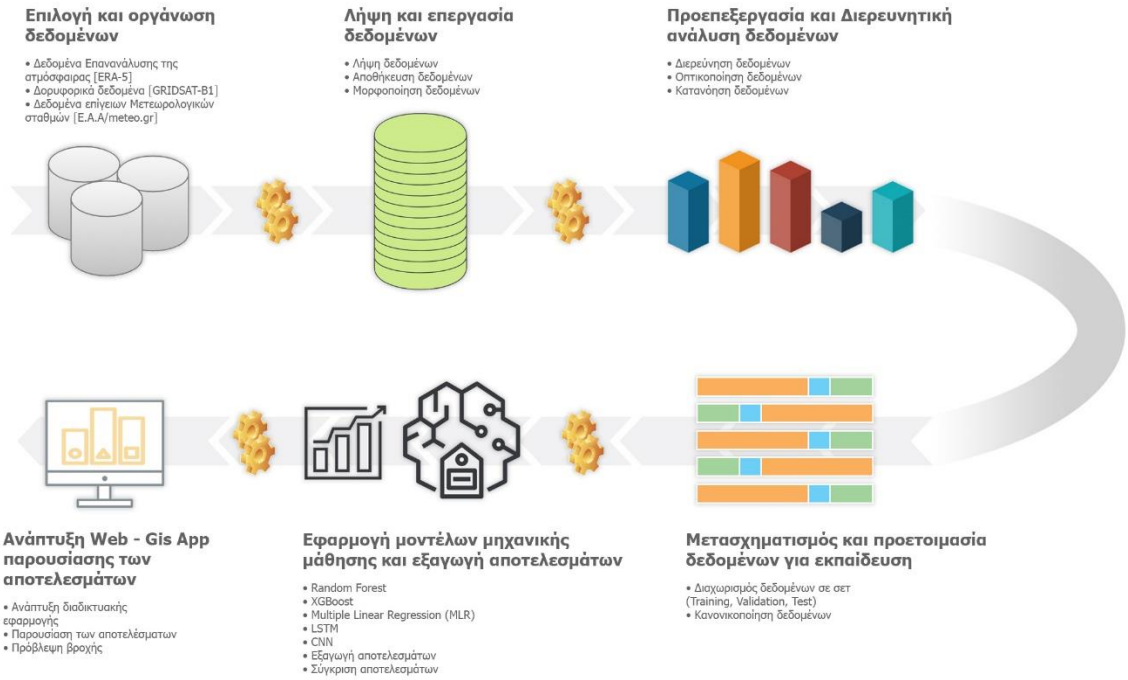
Αντικείμενο αυτής της πτυχιακής εργασίας αποτελεί η πρόβλεψη της στιγμιαίας βροχής το αμέσως επόμενο 3ωρο, με την χρήση μοντέλων και τεχνικών μηχανικής μάθησης από δεδομένα επανανάλυσης της ατμόσφαιρας, δορυφορικά δεδομένα καθώς και δεδομένα από επίγειους μετεωρολογικούς σταθμούς. Η χρήση και η εφαρμογή των μοντέλων μηχανικής μάθησης έχει αποδειχθεί πως βελτιώνει, διορθώνει και επιλύει πολλά προβλήματα της καθημερινότητας μας μέσω των προηγμένων τεχνικών και τεχνολογιών που χρησιμοποιούνται.

Με βάση την βιβλιογραφία και τις σχετικές έρευνες, σε αυτή την εργασία εφαρμόσαμε μοντέλα και τεχνικές μάθησης που αποδεδειγμένα έχουν αποδώσει και έχουν επιφέρει σημαντικές επιτυχίες και αποτελέσματα σε θέματα πρόβλεψης καιρού και παραμέτρων αυτού. Έτσι, επιλέξαμε 5 διαφορετικά μοντέλα μηχανικής μάθησης που θα εφαρμόσουμε και θα συγκρίνουμε την απόδοσή τους και τα αποτελέσματα τους πάνω στο πρόβλημα της πρόβλεψης της βροχής. Αυτά τα μοντέλα είναι:

1. Τυχαίο Δάσος (Random Forest) - Κατηγορία Δέντρα αποφάσεων (Decision Trees)
2. Πολλαπλή γραμμική παλινδρόμηση (MLR) - Κατηγορία Γραμμικής Παλινδρόμησης (Regression)
3. XGBoost - Κατηγορία Ενισχυμένα δέντρα αποφάσεων (Gradient-boosted Decision Trees)
4. Νευρωνικά Δίκτυα Μακράς Βραχύχρονης Μνήμης (LSTM) - Κατηγορία Αναδρομικά νευρωνικά δίκτυα (Recurrent Neural Networks, RNN)
5. Συνελικτικά Νευρωνικά Δίκτυα (CNN) - Κατηγορία Νευρωνικά Δίκτυα Συνέλιξης (Convolutional Neural Networks, CNN)

Για την οργάνωση και επιλογή των δεδομένων συμβουλευτήκαμε τους ειδικούς επιστήμονες του Εθνικού Αστεροσκοπείου Αθηνών/ΜΕΤΕΟ σε θέματα μετεωρολογίας Δρ. Λαγουβάρδο Κωνσταντίνο, Δρ. Ντάφη Σταύρο και την Δρ. Κοτρώνη Βασιλική. Στόχος μας ήταν η επιλογή και η οργάνωση των δεδομένων με τον καλύτερο δυνατό τρόπο. Έτσι, κρίθηκε απαραίτητο να επιλέξουμε 3 πηγές δεδομένων που αυτά ήταν τα δεδομένα επανανάλυσης της ατμόσφαιρας ERA-5, τα δορυφορικά δεδομένα GRIDSAT-B1, καθώς και δεδομένα από τους επίγειους μετεωρολογικούς σταθμούς του Δικτύου Αυτόματων Μετεωρολογικών Σταθμών του Εθνικού Αστεροσκοπείου Αθηνών/ΜΕΤΕΟ. Στόχος ήταν ο καλύτερος δυνατός συνδυασμός των μετεωρολογικών παραμέτρων και συνθηκών της ατμόσφαιρας, βρίσκοντας τις καλύτερες συσχετίσεις μεταξύ αυτών των δεδομένων και της παραμέτρου στόχου πρόβλεψης της βροχής, μέσα από την χρήση και εφαρμογή των προηγμένων τεχνικών μηχανικής μάθησης.

Επιπλέον, πριν φτάσουμε στην ανάπτυξη και εφαρμογή των μοντέλων μηχανικής μάθησης πραγματοποιήθηκε μια ενδελεχής έρευνα και επεξεργασία των δεδομένων με σκοπό να αναλυθούν και να κατανοηθούν οι σχέσεις μεταξύ των δεδομένων, και να επιτευχθεί η σωστή τροποποίηση και μορφοποίηση των δεδομένων για την εκπαίδευση των μοντέλων. Στην Εικόνα 2 παρουσιάζεται το διάγραμμα ροής εργασιών και ενεργειών που έγιναν σε αυτή την εργασία από την αρχή μέχρι και το τέλος αυτής.



Εικόνα 2. Διάγραμμα ροής εργασιών και ενεργειών που συντέλεσαν αυτή την εργασία.

2

Σχετικές εργασίες

Δεδομένου ότι η βελτιστοποίηση της πρόβλεψης των καιρικών φαινομένων και της βροχόπτωσης έχει απασχολήσει πολλές ερευνητικές ομάδες, σε αυτήν την ενότητα θα εξετάσουμε μερικές εργασίες που σχετίζονται άμεσα ή έμμεσα με τη μεθοδολογία και τον τρόπο που έχει επιλεγεί για αυτήν την συγκεκριμένη διπλωματική εργασία.

2.1 Πρόβλεψη βροχοπτώσεων με χρήση ενισχυμένου αναδρομικού νευρωνικού δικτύου (Intensified LSTM)

Σε αυτή την ερευνητική εργασία οι S. Poornima και M. Pushpalatha είχαν ως στόχο την σωστή πρόβλεψη της βροχόπτωσης με την χρήση του αλγοριθμικού μοντέλου μηχανικής μάθησης LSTM. Στόχος τους ήταν να εκπαιδεύσουν το νευρωνικό δίκτυο με ένα σύνολο δεδομένων βροχόπτωσης και αυτό με την σειρά του να καταφέρνει να προβλέπει την βροχόπτωση με όσο το δυνατόν μεγαλύτερη ακρίβεια. Σε συγκρίσεις των αποτελεσμάτων με άλλους αλγορίθμους που χρησιμοποιούνται για τον ίδιο σκοπό όπως τα Arima, RNN αλλά και το απλό LSTM, διαπίστωσαν ότι το ενισχυμένο LSTM επιφέρει τα καλυτέρα αποτελέσματα. Αυτό συμβαίνει κυρίως διότι ο συγκεκριμένος αλγόριθμος έχει την δυνατότητα να διατηρεί περισσότερα δεδομένα στην μνήμη του. Έτσι, κατόρθωσε να έχει συγκεντρωτικά μικρότερες σφάλματα και μεγαλύτερη ακρίβεια για όλες τις εποχές [5].

2.2 Συνελικτικό Νευρωνικό Δίκτυο ConvLSTM: Μια προσέγγιση Μηχανικής Μάθησης για βροχόπτωση σε πραγματικό χρόνο

Στην εργασία τους οι Xingjian Shi, Zhourong Chen, Hao Wang και Dit-Yan Yeung έχουν ως στόχο την πρόβλεψη της βροχόπτωσης στο άμεσο κοντινό χρονικό διάστημα (Nowcasting) με την χρήση μοντέλων μηχανικής μάθησης για μια συγκεκριμένη χρονική

περίοδο. Η συγκεκριμένη ομάδα ανακάλυψε ότι ελάχιστοι έχουν ασχοληθεί με την άμεση πρόβλεψη σε κοντινό χρονικό διάστημα με την χρήση τεχνικών μηχανικής μάθησης και αποφάσισε να ερευνήσει περαιτέρω τον συγκεκριμένο τομέα. Έτσι, η συγκεκριμένη ερευνητική ομάδα εφάρμοσε μια επέκταση του αλγοριθμικού μοντέλου LSTM για την λύση αυτού του προβλήματος, το ConvLSTM. Όπως αποδείχθηκε το ConvLSTM όχι μόνο διατηρεί τα πλεονεκτήματα τον κλασικού LSTM αλλά είναι καταλληλότερο χωροχρονικά λόγω της εγγενούς συνελικτικής δομής του [3].

2.3 Τεχνικές Μηχανικής Μάθησης για την πρόβλεψη της ημερήσιας βροχόπτωσης

Στόχος της εργασίας των Chalachew Muluken Liyew και Haileyesus Amsaya Melese ήταν η πρόβλεψη της ημερήσιας ποσότητας βροχόπτωσης για την βελτίωση της γεωργικής παραγωγής και μείωση των φαινομένων όπως οι πλημύρες. Με αυτό τον τρόπο θα μπορούσε να γίνει ευκολότερη η πρόσβαση των ανθρώπων σε καθαρό πόσιμο νερό και τροφή και επιπλέον θα μειωθεί η θνησιμότητα από τα ακραία καιρικά φαινόμενα. Για να το επιτύχουν αυτό, οι ερευνητές εφάρμοσαν τρία μοντέλα μηχανικής μάθησης, την Πολλαπλή Γραμική Παλινδρόμηση (MLR), το Τυχαίο Δάσος (Random Forest) και το XGBoost. Έπειτα από μια πληθώρα δοκιμών και πειραμάτων με τα ατμοσφαιρικά δεδομένα που διέθεταν, κατέληξαν στο συμπέρασμα ότι ο καταλληλότερος αλγόριθμος μηχανικής μάθησης για την ακριβέστερη πρόβλεψη της ημερήσιας ποσότητας βροχής ήταν ο XGBoost [11].

2.4 Μοντέλο βελτίωσης για προβλέψεις βροχοπτώσεων με χρήση του LSTM με πολλαπλούς μετεωρολογικούς παράγοντες

Η ερευνητική ομάδα των Chang-Jiang Zhang, Jing Zeng, Hui-Yuan Wang, Lei-Ming Ma, Hai Chu έθεσε ως σκοπό της εργασίας τους την βελτίωση της ακρίβειας στην πρόβλεψης του καιρού με την χρήση του μοντέλου μηχανικής μάθησης LSTM. Ως είσοδος για το μοντέλο χρησιμοποιήθηκε μια πληθώρα δεδομένων που θεωρήσαν ότι επηρεάζουν άμεσα τον καιρό σε δωδεκάωρη βάση. Με την διεξαγωγή διάφορων σεναρίων, η συγκεκριμένη ερευνητική ομάδα κατάφερε να κατανοήσει καλυτέρα την πρόγνωση του καιρού και να εντοπίσει σχέσεις και μοτίβα τα οποία επηρεάζουν την έκβαση των καιρικών φαινομένων σε μεγαλύτερό βαθμό. Χάρη σε αυτές τις παρατηρήσεις πραγματοποιήθηκαν διορθώσεις στο μοντέλο LSTM οι οποίες βοήθησαν στην αύξηση της αποτελεσματικότητας του μοντέλου πρόγνωσης [2].

3

Θεωρητικό υπόβαθρο

Σε αυτή την ενότητα αναφέρονται βασικές πληροφορίες και στοιχεία σχετικά με το θεωρητικό υπόβαθρο αυτής της εργασίας. Αρχικά αναλύεται ο όρος της Μηχανικής Μάθησης και στη συνέχεια οι κλάδοι αυτής καθώς και τα μοντέλα και οι τεχνικές που εφαρμόστηκαν σε αυτή την εργασία. Τέλος, παρουσιάζονται και λίγες πληροφορίες σε ότι έχει να κάνει με την Μετεωρολογία και την πρόβλεψη καιρού.

3.1 Μηχανική Μάθηση

Η μηχανική μάθηση είναι ένας κλάδος της επιστήμης των υπολογιστών, συγκεκριμένα του τομέα της τεχνητής νοημοσύνης, στόχος του οποίου είναι ο σχεδιασμός και η ανάπτυξη αλγορίθμων και μοντέλων που επιτρέπουν στους υπολογιστές να αφομοιώσουν και να βελτιώνουν την απόδοσή τους αυτόνομα μέσω της εμπειρίας που αποκτούν κατά την εκπαίδευση τους από τα δεδομένα.

Ως εμπειρία ορίζεται η προσαρμογή στα δεδομένα, δηλαδή στα χαρακτηριστικά και τα μοτίβα που απαιτούνται για την εκτέλεση συγκεκριμένων εργασιών [12]. Η εκπαίδευση δεν περιορίζεται στα αρχικά πεπερασμένα δεδομένα, καθώς οι αλγόριθμοι εξελίσσονται μέσα από τα ίδια τους τα λάθη και γίνονται διαρκώς καλύτεροι και πιο ακριβείς. Παρατηρούμε ότι ο ουσιαστικός στόχος των αλγορίθμων μηχανικής μάθησης είναι να μιμούνται την ανθρώπινη νοημοσύνη και τον τρόπο που αυτή εξελίσσεται διαρκώς μέσα από τα λάθη που βιώνει. Ένα από τα βασικά εργαλεία της μηχανικής μάθησης είναι οι αλγόριθμοι εκμάθησης από τους οποίους αποτελείται. Συγκεκριμένα, οι αλγόριθμοι μάθησης με επίβλεψη (Supervised learning), οι αλγόριθμοι μάθησης χωρίς επίβλεψη (Unsupervised learning) οι συνδυαστικοί αλγόριθμοι ημι-επιβλεπόμενης μάθησης (Semi-Supervised learning) και οι αλγόριθμοι ενισχυτικής μάθησης (Reinforcement Learning) [13].

1. Αλγόριθμοι μάθησης με επίβλεψη (Supervised learning): χρησιμοποιούνται για την εκτίμηση ή ταξινόμηση νέων δεδομένων, βάση του συνόλου ετικετοποιημένων (εξαρτημένων), παραδειγμάτων εκπαίδευσης. Οι αλγόριθμοι αυτοί ακολουθούν συγκεκριμένα μοντέλα μάθησης έτσι ώστε να εξάγουν κανόνες και μοτίβα από τα δεδομένα εκπαίδευσης και να τα εφαρμόσουν σε νέα, μη ετικετοποιημένα δεδομένα. Οι αλγόριθμοι αυτής της κατηγορίας χρησιμοποιούν διάφορες τεχνικές με σκοπό την χρήσης της διαδικασίας της επιβλεπόμενης μάθησης, μερικές είναι η λογιστική παλινδρόμηση (Logistic Regression), δέντρα αποφάσεων (Decision Trees), μηχανές διανυσμάτων υποστήριξης (SVM) και πολλές άλλες.
2. Αλγόριθμοι μάθησης χωρίς επίβλεψη (Unsupervised learning): χρησιμοποιούνται για την ανάλυση και την εξαγωγή μοτίβων από μη ετικετοποιημένα δεδομένα. Συγκεκριμένα, ο αλγόριθμος προσπαθεί να εντοπίσει σχέσεις και μοτίβα ανάμεσα στα δεδομένα. Παραδείγματα τέτοιων αλγορίθμων είναι η ομαδοποίηση k-means, και η ομαδοποίηση Προσδοκιών-Μεγιστοποίησης (Expectation–Maximization).
3. Αλγόριθμοι ημι-επιβλεπόμενης μάθησης (Semi-Supervised learning): είναι μια μέθοδος που συνδυάζει τόσο τα στοιχεία της επιβλεπόμενης αλλά και μη επιβλεπόμενης μάθησης. Στην ημι-επιβλεπόμενη μάθηση, έχουμε ένα σύνολο δεδομένων εκπαίδευσης όπου μόνο ένα μέρος των δεδομένων χρησιμοποιείται σαν ετικέτα. Η λειτουργία του συγκεκριμένου αλγορίθμου βασίζεται στην ιδέα της αξιοποίησης των πληροφοριών από τα δεδομένα που έχουν ετικέτα και των πληροφοριών από τα δεδομένα που δεν έχουν ετικέτα για την βελτίωση της απόδοσης του.
4. Αλγόριθμοι ενισχυτικής μάθησης (Reinforcement learning): ασχολούνται με την μάθηση μέσω του πειραματισμού και της αλληλεπίδρασης με το περιβάλλον που δραστηριοποιούνται. Ο συγκεκριμένος αλγόριθμος μαθαίνει συνεχώς από τις δοκιμές και τα λάθη του, καταφέρνοντας με αυτό τον τρόπο να προσαρμόζεται καταλλήλως στις ανάγκες του προβλήματος και να βελτιώνει συνεχώς την απόδοση του. Αναλόγως τον τρόπο εφαρμογής του, ο αλγόριθμος ενισχυτικής μάθησης λειτουργεί με τρόπο παρόμοιο της εποπτευόμενης μάθησης ή της ημι-εποπτευόμενης προσέγγισης.

3.1.1 Μηχανική Μάθηση & Μετεωρολογία

Σήμερα όσο ποτέ η επίδραση της κλιματικής αλλαγής επιφέρει σημαντικά προβλήματα και ζημιές τόσο στις περιουσίες που χάνονται καθημερινά, όσο και στις ανθρώπινες ζωές που κινδυνεύουν από τα ακραία καιρικά φαινόμενα τον τελευταίων ετών. Έτσι, κρίνεται απαραίτητη η σωστότερη και ακριβέστερη πρόβλεψη των ακραίων καιρικών φαινομένων, αλλά και του καιρού γενικότερα με σκοπό την προστασία των περιουσιών, του περιβάλλοντος και των ανθρώπων.

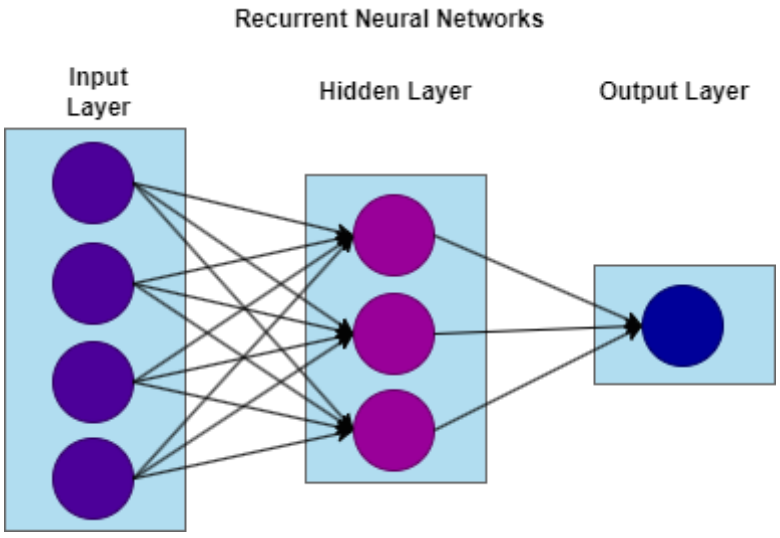
Καθώς οι προγνώσεις του καιρού είναι από την φύση τους πολύπλοκες και επηρεάζονται από πολλούς παράγοντες όπως η θερμοκρασία, η ατμοσφαιρική πίεση, η υγρασία, και οι άνεμοι, οι επιστήμονες του κλάδου της μετεωρολογίας για να επιτύχουν καλύτερα αποτελέσματα στις προβλέψεις τους έχουν ξεκινήσει να κάνουν χρήση της μηχανικής μάθησης. Η χρήση της μηχανικής μάθησης βοηθά στη πρόβλεψη του καιρού, καθιστώντας λιγότερο απαιτητική τη δημιουργία πολύπλοκων πλαισίων προγραμματισμού με λιγότερη προσπάθεια από τους μετεωρολόγους [14]. Πολλά επιχειρησιακά κέντρα πρόγνωσης καιρού εφαρμόζουν τέτοιες τεχνικές για να βελτιώσουν τις προβλέψεις τους όπως η Εθνική Διοίκηση Ωκεανών και Ατμόσφαιρας (NOAA), Το Ευρωπαϊκό Κέντρο Προγνώσεων Καιρού Μεσαίου Εύρους (ECMWF), και η Γαλλική Μετεωρολογική Υπηρεσία (Météo-France) [8].

Οι κύριες μέθοδοι πρόβλεψης της βροχόπτωσης μπορούν να χωριστούν σε δύο κατηγορίες, τους στατιστικούς αλγόριθμους και τους αλγόριθμους μηχανικής μάθησης. Ο πλέον συμβατικός τρόπος πρόβλεψης της βροχόπτωσης είναι με την χρήση των στατιστικών αλγορίθμων. Οι γραμμικές εφαρμογές είναι αυτές που κάνουν χρήση των στατιστικών μεθόδων, ενώ από την άλλη μεριά η μη γραμμικές μέθοδοι χρήζουν εφαρμογής μοντέλων μηχανικής μάθησης για την επίλυση των προβλημάτων τους. Ένα τέτοιου είδους πρόβλημα είναι η πρόβλεψη βροχοπτώσεων. Τέλος, εκτός των κλασσικών γραμμικών και μη γραμμικών μεθόδων για την πρόβλεψη της βροχής, τα νευρωνικά δίκτυα τείνουν να εφαρμόζονται κατά κόρον για τέτοιες εφαρμογές. Τα νευρωνικά δίκτυα που χρησιμοποιούνται για την υλοποίηση ενός αλγορίθμου μηχανικής μάθησης χωρίζονται σε κατηγορίες. Τα αναδρομικά νευρωνικά δίκτυα (RNN) και Συνελικτικά νευρωνικά δίκτυα (CNN) είναι δύο από τις κατηγορίες που χρησιμοποιήσαμε σε αυτή την εργασία διότι έπειτα από ερευνά και πειραματισμούς καταλήξαμε ότι στην περίπτωση του προβλήματος της πρόγνωσης της βροχόπτωσης αποτελούν τις καταλληλότερες κατηγορίες νευρωνικών δικτύων που εφαρμόζουν στο πρόβλημα μας.

Αναλυτικά, στη συγκεκριμένη έρευνα χρησιμοποιήθηκαν τόσο μοντέλα στατιστικών αλγορίθμων όσο και μοντέλα μηχανικής μάθησης. Έτσι, καταφέραμε να συγκρίνουμε και να αναλύσουμε τα αποτελέσματα όλων ώστε να φτάσουμε στο καλύτερο δυνατό αποτέλεσμα. Χρησιμοποιήθηκε το η κλασσική μέθοδος γραμμικής παλινδρόμησης με πολλές μεταβλητές (Multiple Linear Regression) όπως επίσης τα μοντέλα μηχανικής μάθησης βασισμένα στα Δέντρα Αποφάσεων (Random Forest), XGBoost καθώς επίσης και Νευρωνικά Δίκτυα (LSTM και CNN). Παρακάτω αναφέρονται κάποιες θεωρητικές πληροφορίες για τις παραπάνω τεχνικές και μοντέλα μηχανικής μάθησης.

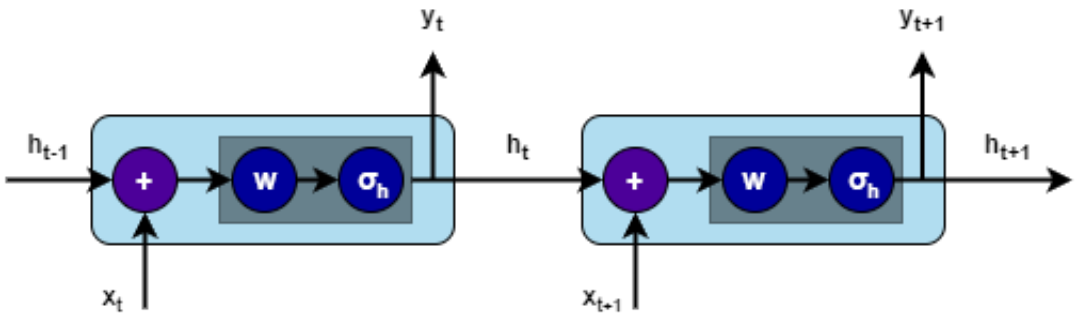
3.1.2 Αναδρομικά Νευρωνικά δίκτυα (RNN)

Τα αναδρομικά νευρωνικά δίκτυα (RNNs) είναι μια κατηγορία νευρωνικών δικτύων που σχεδιάστηκε για την επεξεργασία ακολουθιακών δεδομένων, όπως χρονοσειρές ή κείμενο [5]. Το RNN έχει τη δυνατότητα της εκμάθησης εξαρτήσεων προηγούμενου χρόνου, δηλαδή διατηρεί στην μνήμη του προηγούμενα αποτελέσματα και καταστάσεις ως συμπληρωματικές πληροφορίες κατά την εισαγωγή των νέων δεδομένων, η συγκεκριμένη λειτουργία είναι μείζονος σημασίας για την ορθή πρόβλεψη καιρικών προτύπων. Στην Εικόνα 3 παρατηρούμε την γενική μορφή της αρχιτεκτονικής ενός RNN.



Εικόνα 3. Γενική δομή ενός RNN.

Σε ένα RNN, η πληροφορία κυλά χρονικά από το ένα βήμα στο επόμενο. Κάθε βήμα λαμβάνει ως είσοδο τόσο το τρέχοντα χρονικό βήμα όσο και την κρυφή κατάσταση από το προηγούμενο χρονικό βήμα. Αυτή η εσωτερική κατάσταση ανανεώνεται συνεχώς καθώς το δίκτυο προχωρά, επιτρέποντας την επεξεργασία των προηγούμενων καταστάσεων για την παραγωγή πιο πλούσιων και συνδετικών δεδομένων κατά την είσοδο. Το ποσοστό επιτυχίας της πρόβλεψης ενός RNN εξαρτάται κατά κύριο λόγο από τη συνάρτηση ενεργοποίησης, η οποία παίρνει την τρέχουσα είσοδο και την προηγούμενη κατάσταση ως διάνυσμα εισόδου και προβλέπει την έξοδο με βάση το αποτέλεσμα της κρυφής κατάστασης, όπως μπορούμε να παρατηρήσουμε και στην Εικόνα 4 μια πιο λεπτομερής έκδοση του RNN [15].



Εικόνα 4. Ένα απλό αναδρομικό νευρωνικό δίκτυό (RNN) με δύο εισόδους (t, t + 1).

Η Εξίσωση που χρησιμοποιείται στο παραπάνω σχήμα είναι

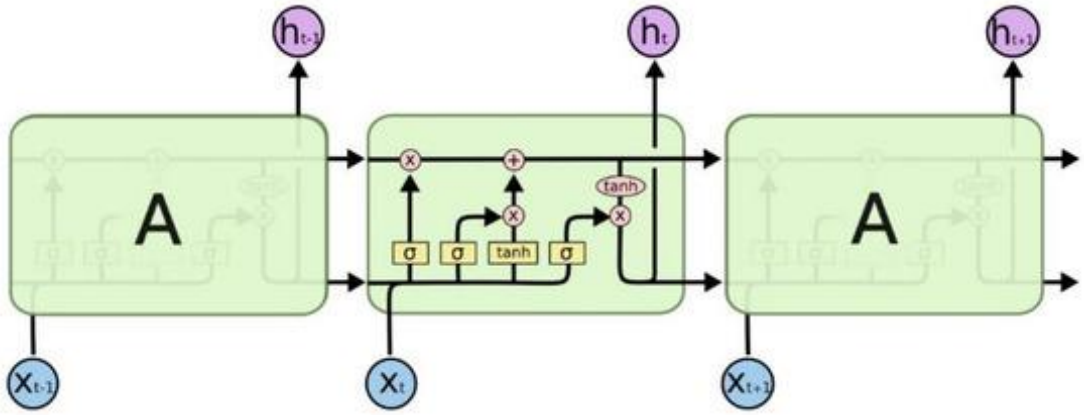
$$y_t = h_t = \sigma_h(W * [h_{t-1}, x_t] + b) \quad (1)$$

Όπου + είναι η σημειακή συνένωση, σ_h είναι οι λειτουργίες ενεργοποίησης τυπικού tanh και W είναι το νευρωνικό με βάρη.

3.1.3 Νευρωνικά Δίκτυα Μακράς Βραχύχρονης Μνήμης (LSTM)

Τα δίκτυα μακράς βραχύχρονης μνήμης ή LSTM (Long Short-Term Memory) είναι ένα μοντέλο αλγορίθμου αναδρομικού νευρωνικού δικτύου (Recurrent neural network - RNN) το οποίο έχει αποδειχθεί ιδιαίτερα ισχυρό στην χρήση δεδομένων με χρονική εξάρτηση μεγάλης εμβέλειας. Το LSTM χρησιμοποιείται κυρίως για την ανάλυση, πρόβλεψη και παραγωγή ακολουθιών δεδομένων. Μερικές από τις πιο γνωστές χρήσεις του LSTM είναι στην αναγνώριση ομιλίας και την μετάφραση, στην δημιουργία κειμένου ακόμη και στην δημιουργία περιγραφής εικόνων ή υπότιτλων σε βίντεο. Το μοντέλο LSTM αναφέρθηκε για πρώτη φορά από τους Schmidhuber και Hochreiter για την επίλυση της χρονικής καθυστέρησης η οποία υπήρχε στο RNN, γνωστό και ως το πρόβλημα της εξαφάνισης ή της εκκρεμότητας των μακροπρόθεσμων εξαρτήσεων [3][4].

Η δομή του LSTM αποτελείται από ένα κελί (cell) και μερικές πύλες (gates). Το κελί είναι μια ειδική μονάδα μνήμης η οποίο ουσιαστικά λειτουργεί ως μέσω αποθήκευσης και συγκράτησης των πληροφοριών της εκάστοτε κατάστασης, οι πύλες του μοντέλου από την άλλη ελέγχουν τη ροή των δεδομένων μέσα και έξω από το κελί. Οι πύλες είναι αυτές που αποφασίζουν για το ποια πληροφορία θα εισαχθεί στη μνήμη (Input Gate) ποια πληροφορία θα απορριφθεί (Forget Gate) καθώς και ποια πληροφορία θα εξαχθεί από τη μονάδα (Output Gate) όπως παρατηρούμε στο Εικόνα 5.



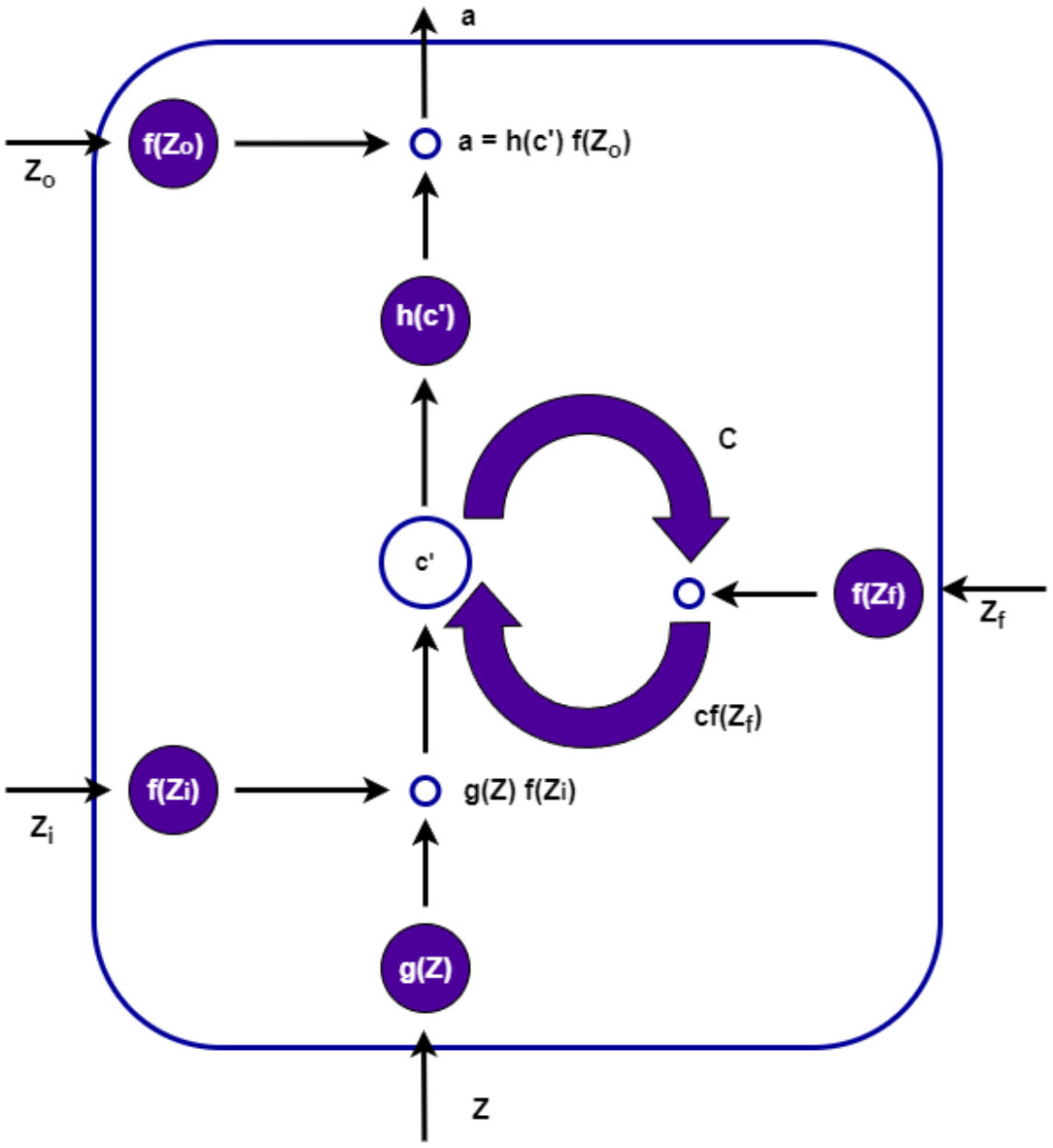
Εικόνα 5. Διάγραμμα δομής δικτύου LSTM [4].

Τα βασικά βήματα του αλγοριθμικού μοντέλου LSTM είναι τα εξής:

1. Υπολογισμός της τρέχουσας εισόδου αλλά και της προηγούμενης κατάστασης.
2. Υπολογισμός της πύλης (Forget Gate), η οποίο είναι η πύλη αυτή που επιλέγει ποια πληροφορία από την που προϋπάρχει στην μνήμη από την προηγούμενη κατάσταση θα ξεχαστεί.
3. Υπολογισμός της πύλης εισόδου (Input Gate), η οποία επιλέγει ποια θα είναι η νέα πληροφορία που θα εισέλθει στην μνήμη.
4. Υπολογισμός της νέας εσωτερικής κατάστασης.
    a. Η προηγούμενη κατάσταση ενημερώνεται με βάση την πύλη (Forget Gate).
    b. Η νέα πληροφορία προστίθεται σύμφωνα με την πύλη εισόδου στην μνήμη.
5. Υπολογισμός της πύλης εξόδου (Output Gate), η οποία καθορίζει ποια πληροφορία θα διεξαχθεί από την μνήμη.
6. Υπολογισμός της τρέχουσας εξόδου, η οποία εξαρτάται από την μνήμη η οποία έχει ενημερωθεί με σύμφωνα με την πύλη εξόδου.

Στην συνέχεια για κάθε στοιχείο της ακολουθίας επαναλαμβάνονται τα βήματα 2 έως 6. Όλα τα στοιχεία επηρεάζουν τόσο την εσωτερική κατάσταση και την μνήμη όσο και την έξοδο του LSTM. Επίσης, η πληροφορία από τα προηγούμενα στοιχεία διατηρείται συνεχώς στην μνήμη.

Στην Εικόνα 6 μπορεί να παρατηρηθεί η ακριβής λειτουργία ενός αλγοριθμικού μοντέλου LSTM όπως την περιεγράφηκε παραπάνω.

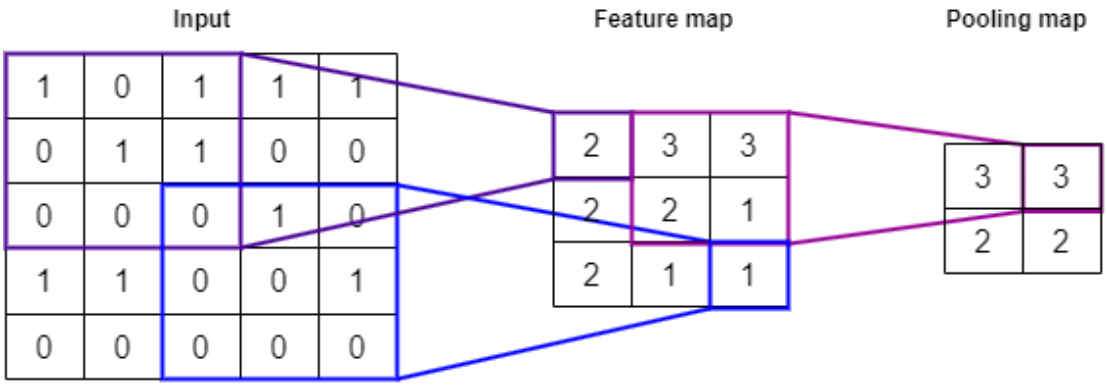


Εικόνα 6. Διάγραμμα δομής μονάδας LSTM [4].

Όπου Το Z είναι η είσοδος και Z_i το σήμα ελέγχου της πύλης εισόδου. Το Z_f είναι το σήμα ελέγχου της πύλης (Forget Gate), το Z_o είναι το σήμα ελέγχου της πύλης εξόδου. $f(x)$ είναι συνάρτηση χρησιμοποιείται για να δώσει το βαθμό ανοίγματος της πύλη, συνήθως χρησιμοποιείται η σιγμοειδής συνάρτηση (Sigmod): $f(x) = \frac{1}{1+e^{-x}}$ με το εύρος τιμών να βρίσκεται εντός [0, 1]. Οι $h(x)$ και $g(x)$ είναι συναρτήσεις ενεργοποίησης.

3.1.4 Συνελικτικά Νευρωνικά Δίκτυα (CNN)

Τα συνελικτικά νευρωνικά δίκτυα (CNNs) είναι μια κατηγορία αλγορίθμων μηχανικής μάθησης που χρησιμοποιούνται κυρίως για την ανάλυση εικόνων και την επεξεργασία σήματος. Το αρχικό επίπεδο ενός τέτοιου αλγορίθμου είναι το συνελικτικό επίπεδο το οποίο αποτελείται από νευρώνες με συγκεκριμένες διαστάσεις που συνθέτουν τον πυρήνα (Kernel) του νευρωνικού δικτύου [7]. Μια από τις κυριότερες λειτουργίες του CNN είναι η συγκέντρωση (Pooling) που βοηθά το μοντέλο μειώνοντας των αριθμό των παραμέτρων. Ο πιο συχνά χρησιμοποιούμενος τρόπος συγκέντρωσης είναι η μέγιστη συγκέντρωση (Max pooling). Σε αυτή τη μέθοδο επιλέγεται και διατηρείται η μέγιστη τιμή σε κάθε περιοχή των χαρακτηριστικών, ενώ οι άλλες τιμές απορρίπτονται. Στην Εικόνα 7 μπορούμε να παρατηρήσουμε την λειτουργία συνέλιξης και της συγκέντρωσης όπως τα περιγράψαμε παραπάνω [15].



Εικόνα 7. Παραδείγματα λειτουργίας συνέλιξης και συγκέντρωσης.

Η εκπαίδευση των συνελικτικών νευρωνικών δικτύων (CNNs) συχνά γίνεται με τη χρήση αλγορίθμων οπισθοδρομικής διάδοσης σφάλματος (Back propagation), οι οποίοι αναπροσαρμόζουν τις παραμέτρους τους για να βελτιώσουν το ενδεχόμενο σφάλμα ανάμεσα στην πραγματική και την εκτιμώμενη τιμή. Τα συνελικτικά νευρωνικά δίκτυα ανά τα χρόνια έχουν πετύχει εξαιρετικά αποτελέσματα σε πολλές εφαρμογές όπως η αναγνώριση αντικειμένων, η αναγνώριση προσώπων και η αναγνώριση προτύπων.

3.1.5 Πολλαπλή Γραμμική Παλινδρόμηση (MLR)

Η πολλαπλή γραμμική παλινδρόμηση (Multiple Linear Regression) είναι μια στατιστική μέθοδος που χρησιμοποιείται για την πρόβλεψη ενός αποτελέσματος βάση άλλων δεδομένων και μεταβλητών, παραδείγματος χάρη για την πρόβλεψη της ημερήσιας ποσότητας βροχόπτωσης, χρησιμοποιώντας τις περιβαλλοντικές μεταβλητές [11]. Στην πολλαπλή γραμμική παλινδρόμηση, χρησιμοποιούμε μια εξαρτημένη μεταβλητή η οποία εξαρτάται

γραμμικά από άλλες ανεξάρτητες ή μη μεταβλητές, δηλαδή υπάρχει μια γραμμική σχέση ανάμεσα σε όλες τις μεταβλητές [16]. Οι ανεξάρτητες μεταβλητές μπορεί να είναι είτε αριθμητικές είτε κατηγορικές. Αυτό που επιχειρεί η πολλαπλή γραμμική παλινδρόμηση είναι να προσαρμόσει μια ευθεία στα δεδομένα, ώστε με αυτό τον τρόπο η απόσταση μεταξύ των πραγματικών τιμών και των προβλεπόμενων να είναι η μικρότερο δυνατή. Μια πολλαπλή γραμμική παλινδρόμηση είναι μια επέκταση της απλής γραμμικής παλινδρόμησης για πολλαπλά δεδομένα και επεξηγηματικές η ανεξάρτητες μεταβλητές X και μια μεταβλητή εξόδου Y [17].

$$x_{i1}, x_{i2}, \dots, x_{ip},\ y_i \qquad i = 1,2,3 \dots n$$

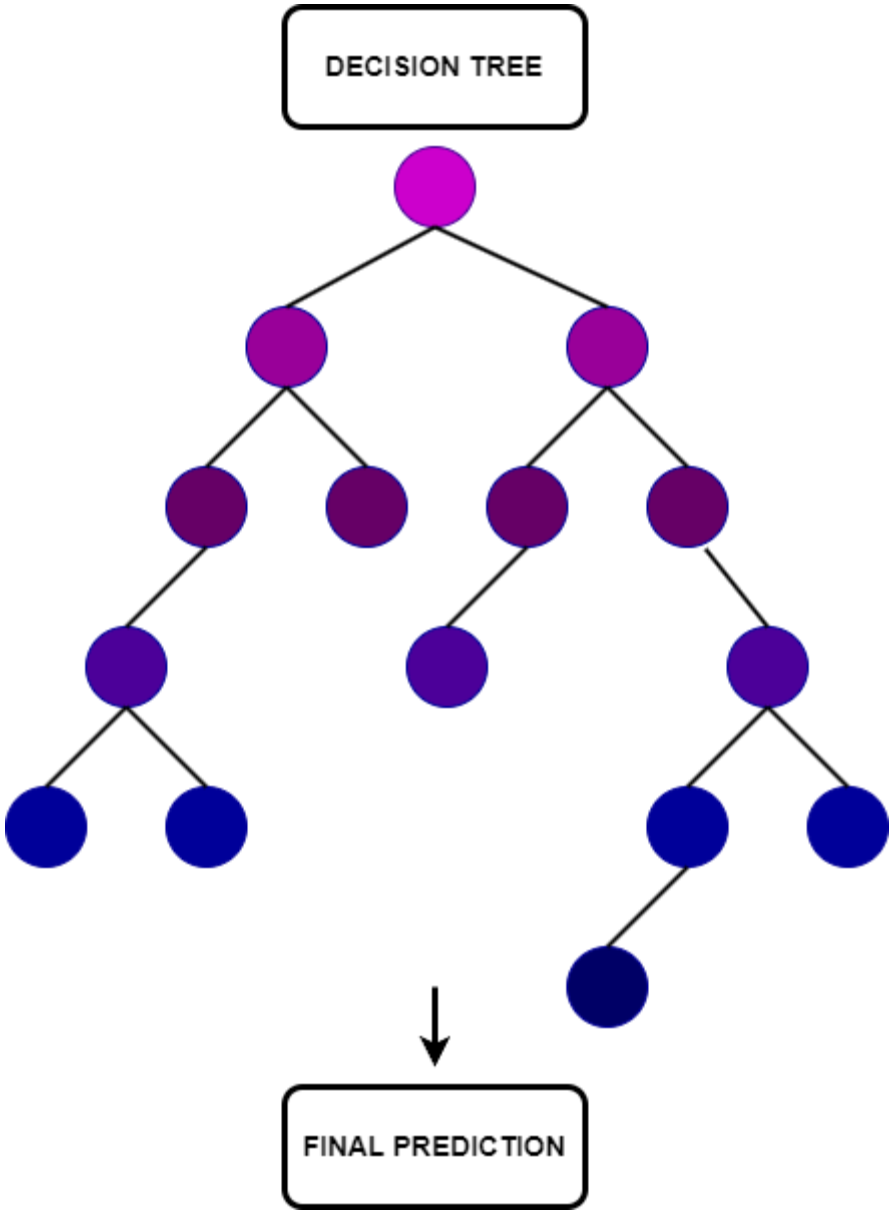
Η γενική εξίσωση πολλαπλής γραμμικής παλινδρόμησης είναι η εξής:

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\ + \varepsilon_i = x_i^T \beta + \varepsilon_i \qquad (2)$$

όπου x_i^T είναι η μετατόπιση του x_i(ανεξάρτητης μεταβλητής), ε_i είναι ο θόρυβος ή το τυχαίο λάθος, β είναι ο συντελεστής παλινδρόμησης και Y_i είναι η εξαρτημένη μεταβλητή.

3.1.6 Δέντρα Αποφάσεων (Decision Trees)

Το Δέντρα αποφάσεων (Decision trees) είναι ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για ταξινόμηση και παλινδρόμηση, ο συγκεκριμένος αλγόριθμος αναπαριστά μια δομή δέντρου για να μοντελοποιήσει τις σχέσεις μεταξύ των μεταβλητών [18]. Το δέντρο ξεκινά από τον πρώτο κόμβο όπου αντιπροσωπεύει το σύνολο των δεδομένων εισόδου, το οποίο ονομάζεται ριζά, στην συνέχεια ξεκινάει να δημιουργεί συνεχώς νέες διαδρομές (διακλαδώσεις) και κόμβους για κάθε πιθανή τιμή και κάθε δυνατή απόφαση, η διαδικασία αυτή επαναλαμβάνεται μέχρι να καταλήξει σε κάποια λύση, μειώνοντας ωστόσο το σύνολο το δεδομένων σε κάθε επανάληψη [19]. Το τελικό αποτέλεσμα θυμίζει ένα δέντρο, όπου έχει πάρει και την ονομασία του ο αλγόριθμος, κάθε φύλλο (κόμβος) αντιπροσωπεύει μια πρόβλεψη και κάθε κλαδί την διαδρομή που έχει ακολουθήσει για να φτάσει σε αυτό το αποτέλεσμα.



Εικόνα 8. Αποτύπωση ενός Δέντρου απόφασης.

Αν και το δέντρο αποφάσεων είναι αρκετά δημοφιλές σαν αλγόριθμος λόγο της απλότητας αλλά και της ευκολίας στην κατανόηση του, αντιμετωπίζει συχνά προβλήματα στην επιλογή του καταλληλότερου κόμβου για διακλάδωση. Για αυτόν ακριβώς τον λόγο χρησιμοποιείται η εντροπία και ο δείκτης Gini [20].

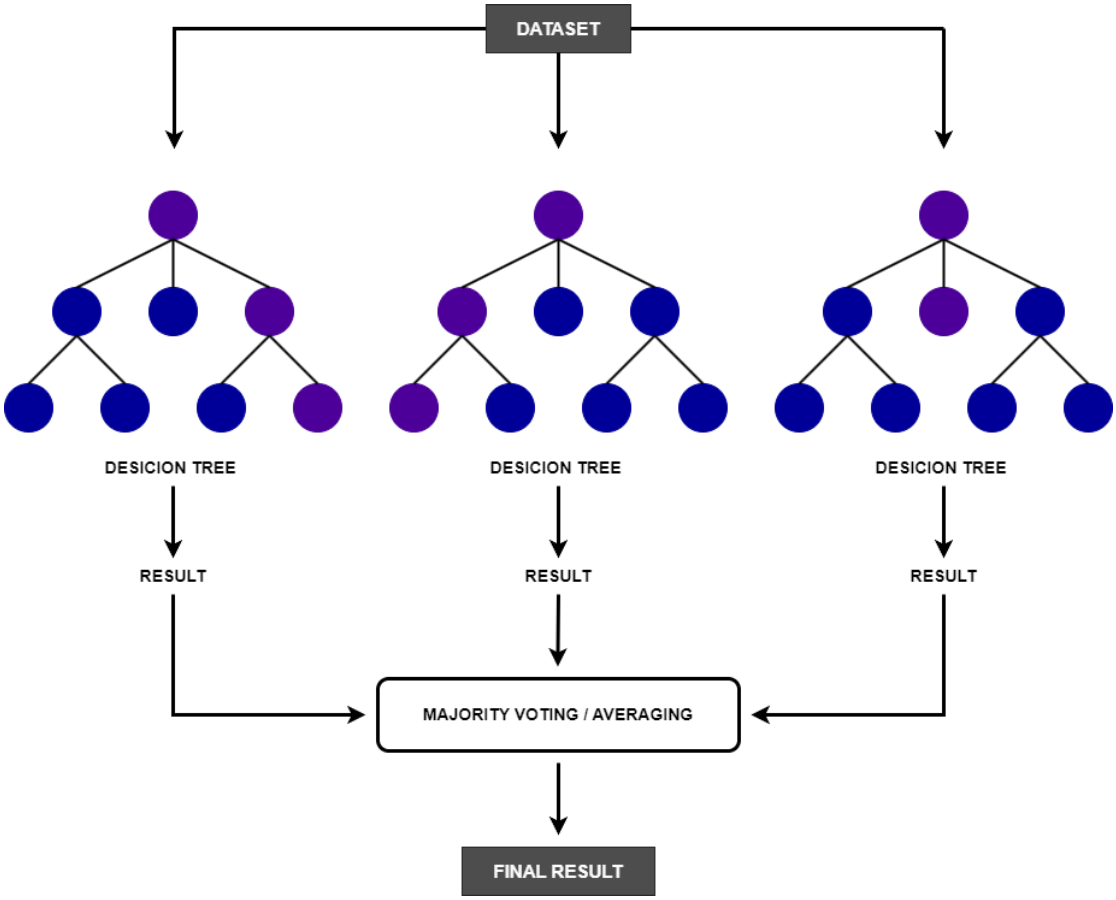
1. Η εντροπία ποσοτικοποιεί την τυχαιότητα σε κάθε ένα σύνολο τιμών μιας κλάσης, με αυτό τον τρόπο τα σύνολα με μεγάλη εντροπία καταλήγουν πολύ διαφορετικά και προσφέρουν λίγες πληροφορίες σε σχέση με τα υπόλοιπα που ανήκουν στο σύνολο. Έτσι, ο αλγόριθμος εντοπίζει διακλαδώσεις που ελαχιστοποιούν την εντροπία, αυξάνοντας έτσι την ομοιογένεια εντός της εκάστοτε κλάσεις.

2. Επιπλέον, ο δείκτης Gini ελέγχει την πιθανότητα της μεταβλητής να ταξινομηθεί εσφαλμένα. Η τιμή του δείκτη βρίσκεται στο διάστημα [0,1] όπου:
   - 0 σημαίνει ότι τα δεδομένα ανήκουν σε μια κλάση.
   - 1 σημαίνει ότι τα δεδομένα κατανέμονται τυχαία μεταξύ των κλάσεων.

   Συνεπώς, κατά την δημιουργία ενός δέντρου αποφάσεων επιλέγεται μεταβλητή με όσο το δυνατόν μικρότερη τιμή στον δείκτη Gini.

3.1.7 Τυχαίο Δάσος (Random Forest)

Το Τυχαίο δάσος (Random Forest) είναι ένας ισχυρός και ακριβής αλγόριθμος μηχανικής μάθησης. Έχει υψηλή απόδοση σε κάθε είδους πρόβλημα ακόμα και σε αυτά με μη γραμμικές σχέσεις. Κυρίως χρησιμοποιείται για προβλήματα ταξινόμησης και παλινδρόμησης. Το τυχαίο δάσος ανήκει στους αλγορίθμους επιβλεπόμενης μάθησης (Supervised learning) και επιτελεί χρήση των μεθόδων συνόλου (Ensemble), που συνδυάζουν τα αποτελέσματα πολλών απλών μοντέλων μηχανικής μάθησης για να βελτιώσουν την ακρίβεια των προβλέψεων [18]. Ουσιαστικά ένα τυχαίο δάσος (Random Forest) αποτελείται από ένα σύνολο από απλά δέντρα αποφάσεων (Decision trees), ο αλγόριθμος δέχεται τα αποτελέσματα όλων των δέντρων και ως τελικό αποτέλεσμα κρατάει το μέσο ορό από τα αποτελέσματα που έχουν δημιουργήσει τα δέντρα απόφασης ή το αποτέλεσμα που εμφανίστηκε τις περισσότερες φορές ανάμεσα στα δέντρα απόφασης ανάλογα με το τι λύση χρειάζεται το κάθε πρόβλημα [19]. Στην Εικόνα 9 μπορούμε να παρατηρήσουμε την δομή ενός αλγορίθμου τυχαίου δάσους και το πως ακριβώς αυτός λειτουργεί.



Εικόνα 9. Αρχιτεκτονική ενός αλγορίθμου Τυχαίου δάσους [21].

Τα βήματα του αλγορίθμου τυχαίου δάσους είναι τα εξής:

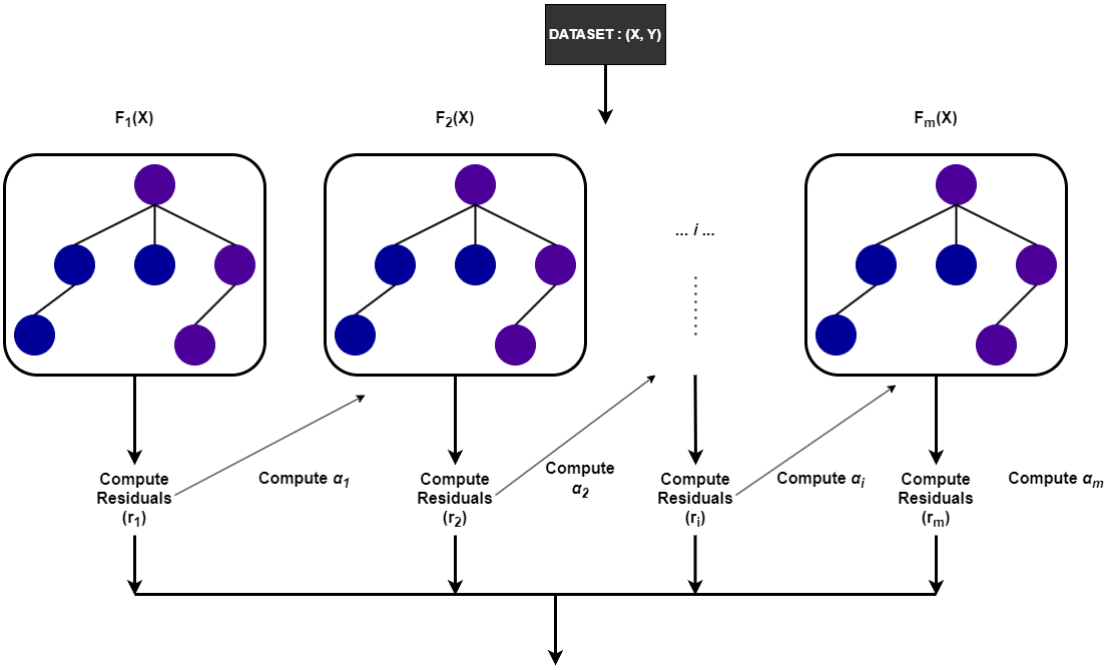
1. Επιλέγεται ένα τυχαίο σημείο από τα δεδομένα εκπαίδευσης.
2. Δημιουργείται ένα δέντρο αποφάσεων συσχετίζοντας τα δεδομένα εκπαίδευσης.
3. Επιλέγεται ένας αριθμός N δέντρων και δημιουργούνται επαναλαμβάνοντας τα βήματα 1 και 2.
4. Δίνεται η εντολή στα δέντρα αποφάσεις να προβλέψουν την μεταβλητή y για το νέο σημείο και το τυχαίο δάσος αντιστοιχεί των μέσο ορό όλων τον προβλέψεων ή την πρόβλεψη που εμφανίστηκε τις περισσότερες φορές στην μεταβλητή y.

Ο αλγόριθμος τυχαίου δάσους είναι από τους κυριότερους αλγόριθμους μηχανικής μάθησης που επιλέγονται για την πρόβλεψη βροχόπτωσης χρησιμοποιώντας ως δεδομένα εισόδου περιβαλλοντικές μεταβλητές ή χαρακτηριστικά. Ένας λόγος για την συχνή χρήση του αλγορίθμου τυχαίου δάσους στην πρόβλεψη της βροχόπτωσης είναι ότι ο συγκεκριμένος αλγόριθμος έχει ως κύριο πλεονέκτημα του την μείωση της πιθανότητας υπερ-προσαρμογής στα δεδομένα (Overfitting) [20]. Έτσι, επιτρέπει την εξάλειψη μεταβλητών που δεν είναι σημαντικές, μπορεί να λειτουργήσει με θόρυβο ή ελλιπή δεδομένα, με συνεχείς κατηγορικές ή αριθμητικές μεταβλητές και με μεγάλο αριθμό μεταβλητών ή υψηλά δείγματα.

3.1.8 XGBoost

Το eXtreme Gradient Boosting η εν συντομία XGBoost είναι ένα μοντέλο επιβλεπόμενης μάθησης οπού βασίζεται στον αλγόριθμο Gradient Boosting και είναι ιδιαίτερα γνωστός για την απόδοσή του σε προβλήματα πρόβλεψης και ταξινόμησης [11]. Συγκεκριμένα ο XGBoost συνδυάζει πολλά μοντέλα εκπαίδευσης, όπως τα δέντρα απόφασης (Random Forest) και γραμμικούς μετασχηματισμούς, για να δημιουργήσει ένα ισχυρότερο και πιο ακριβές μοντέλο ειδικότερα όταν αφορά μικρότερα δείγματα δεδομένων [22].

Το συγκεκριμένο μοντέλο είναι από τα νεότερα μοντέλα μηχανικής μάθησης το οποίο δημιουργήθηκε από τους Carlos Guestrin και Tianqi Chen μόλις το 2011, ενώ λαμβάνει συνεχώς βελτιστοποιήσεις ακόμη και σήμερα. Οι περισσότεροι αλγόριθμοι δέντρου έχουν δύσκολη εφαρμογή καθώς η κατανεμημένη εκπαίδευση είναι κάτι που δεν επιτυγχάνεται εύκολα. Το XGBoost καταφέρνει να λύσει αυτού του είδους τα προβλήματα καθώς παρέχει μια πληθώρα από προηγμένες δυνατότητες, όπως η ρύθμιση της πολυπλοκότητας του μοντέλου με τη χρήση παραμέτρων, η υποστήριξη για ενσωμάτωση και εξαγωγή χαρακτηριστικών και η εκτέλεση παράλληλων υπολογισμών με την χρήση πολυνημάτωσης (Multithreading) του επεξεργαστή, κάτι που τον καθιστά άμεσο και εύστοχο [23].



Εικόνα 10. Αναπαράσταση λειτουργίας μοντέλου XGBoost [24].

$$F_m(X) = F_{m-1}(X) + a_m h_m(X, r_{m-1}) \quad (3)$$

όπου ai και ri είναι παράμετροι τακτοποίησης (regularizes) και τα υπολείμματα (residuals) που υπολογίζονται με το i^{th} δέντρο αντίστοιχα. Το hi είναι μια συνάρτηση που έχει εκπαιδευτεί να προβλέπει τα υπολείμματα (residuals) r_i χρησιμοποιώντας το X με το i^{th} δέντρο.

Για να υπολογίσουμε το ai χρησιμοποιούμε το r_i με την παρακάτω εξίσωση:

$$\arg\frac{min}{a} = \sum_{i=1}^{m} L\big(Y_i, F_{i-1}(X_i) + ah_i(X_i, r_{i-1})\big) \quad (4)$$

όπου $L\big(Y, F(X)\big)$ είναι μια διαφοροποιήσιμη συνάρτηση απώλειας.

3.2 Μετεωρολογία και πρόβλεψη καιρού

Σε αυτή την ενότητα αναφέρονται πολύ λίγες και βασικές πληροφορίες σε ότι έχει να κάνει με την επιστήμη της Μετεωρολογίας και το πεδίο της πρόβλεψης του καιρού, με σκοπό την εξοικείωση του αναγνώστη της παρούσας εργασίας με τα δεδομένα και τις παραμέτρους που αναφέρονται παρακάτω.

3.2.1 Επιστήμη της Μετεωρολογίας

Η Μετεωρολογία αποτελεί ένα πεδίο των θετικών επιστημών και αποτελεί υποσύνολο των ατμοσφαιρικών επιστημών, μαζί με τη Φυσική της Ατμόσφαιρας, την Κλιματολογία, την Ατμοσφαιρική Χημεία. Η Μετεωρολογία επικεντρώνεται στη μελέτη της ατμόσφαιρας, των ατμοσφαιρικών φαινομένων και των ατμοσφαιρικών επιδράσεων στον καιρό μας. Ο καιρός εμφανίζεται σε διαφορετικές κλίμακες του χώρου και του χρόνου. Οι τέσσερις μετεωρολογικές κλίμακες είναι: Μικροκλίμακα (Microscale), Μεσοκλίμακα (Mesoscale), Συνοπτική κλίμακα (Synoptic scale) και Παγκόσμια κλίμακα (Global scale) [25].

- Μικροκλίμακα (Microscale meteorology): Η Μετεωρολογία μικροκλίμακας εστιάζει σε φαινόμενα που κυμαίνονται σε μέγεθος ακτίνας από μερικά εκατοστά έως λίγα χιλιόμετρα και που έχουν μικρή διάρκεια ζωής (λιγότερο από μία ημέρα). Αυτά τα φαινόμενα επηρεάζουν πολύ μικρές γεωγραφικές περιοχές.
- Μεσοκλίμακα (Mesoscale meteorology): Τα φαινόμενα μεσοκλίμακας κυμαίνονται σε ακτίνα μεγέθους από λίγα χιλιόμετρα έως περίπου 1.000 χιλιόμετρα (620 μίλια).
- Συνοπτική κλίμακα (Synoptic scale meteorology): Τα φαινόμενα συνοπτικής κλίμακας καλύπτουν μια έκταση αρκετών εκατοντάδων ή και χιλιάδων χιλιομέτρων. Τα συστήματα υψηλής και χαμηλής πίεσης που εμφανίζονται στις τοπικές μετεωρολογικές προβλέψεις είναι συνοπτικής κλίμακας.
- Παγκόσμια κλίμακα (Global scale meteorology): Τα φαινόμενα παγκόσμιας κλίμακας είναι καιρικά μοτίβα που σχετίζονται με τη μεταφορά θερμότητας, ανέμου και υγρασίας από τις τροπικές περιοχές στους πόλους. Ένα σημαντικό μοτίβο είναι η

παγκόσμια ατμοσφαιρική κυκλοφορία, η μεγάλης κλίμακας κίνηση του αέρα που βοηθά στη διανομή της θερμικής ενέργειας (θερμότητας) σε όλη την επιφάνεια της Γης.

Τα μετεωρολογικά φαινόμενα είναι καιρικά φαινόμενα που μελετούνται και εξηγούνται από την επιστήμη της Μετεωρολογίας. Τα μετεωρολογικά φαινόμενα περιγράφονται και ποσοτικοποιούνται από τις βασικές μεταβλητές της ατμόσφαιρας της Γης:

- Θερμοκρασία
- Ατμοσφαιρική πίεση αέρα
- Υδρατμοί
- Άνεμος

Καθώς και οι διακυμάνσεις και οι αλληλεπιδράσεις αυτών των μεταβλητών και πώς αλλάζουν με την πάροδο του χρόνου [26].

Ο κλάδος της Μετεωρολογίας αποτελεί μια πολύ σημαντική επιστήμη που καθορίζει και συντονίζει πολλές φορές την καθημερινότητα του ανθρώπου και του απλού πολίτη, καθώς καλείται να διαχειριστεί την εξάρτηση του ανθρώπου με τα καιρικά φαινόμενα επί καθημερινής βάσεως. Έτσι, η Μετεωρολογία βρίσκει εφαρμογή σε πάρα πολλούς κλάδους όπως: η πρόγνωση καιρού (Weather forecasting), η Αεροπορική μετεωρολογία (Aviation meteorology), Αγροτική μετεωρολογία (Agricultural meteorology), Υδρομετεωρολογία (Hydrometeorology), Θαλάσσια μετεωρολογία (Maritime meteorology), Στρατιωτική μετεωρολογία (Military meteorology), Περιβαλλοντική μετεωρολογία (Environmental meteorology) καθώς πλέον και στις Ανανεώσιμες πηγές ενέργειας (Renewable energy).

3.2.1.1 Το καιρικό φαινόμενο της βροχής

Σε αυτή την υποενότητα παρουσιάζονται λίγες πληροφορίες για την παράμετρο στόχο πρόβλεψης σε αυτή την εργασία, την βροχή.

Σύμφωνα με την ηλεκτρονική μετεωρολογική εγκυκλοπαίδεια του Εθνικού Αστεροσκοπείου Αθηνών, την οποία επιμελείται η επιστημονική ομάδα του meteo.gr, η βροχή ή βροχόπτωση αποτελεί το φαινόμενο της πτώσης κατακρημνισμάτων από νέφη σε υγρή μορφή και τα οποία φτάνουν στο έδαφος. Πρόκειται για σταγόνες νερού που δημιουργούνται είτε λόγω κορεσμού των νεφών σε υδρατμούς, σε θερμοκρασία πάνω από 0°C, είτε λόγω πτώσης παγοκρυστάλλων που λιώνουν πριν φτάσουν στο έδαφος. Οι σταγόνες της βροχής έχουν διάφορα μεγέθη που κυμαίνονται από 0.5 έως 6 χιλιοστά και η ταχύτητα πτώσης τους εξαρτάται από το βάρος και το μέγεθός τους, και κυμαίνεται από 10 km/h (ασθενής βροχή) έως 40 km/h (πολύ ισχυρή βροχή).

Η μέτρηση του ύψους βροχής γίνεται με ειδικά όργανα που ονομάζονται βροχόμετρα και η μονάδα μέτρησης είναι τα χιλιοστά (mm). Η σωστή μέτρηση του ύψους της βροχής γίνεται σε οριζόντιο επίπεδο την ώρα που εξελίσσεται το φαινόμενο. Ένα χιλιοστό βροχής αντιστοιχεί σε 1 λίτρο νερού σε 1 τετραγωνικό μέτρο ή σε 1000 λίτρα νερού σε 1 στρέμμα. Επίσης, με τα βροχόμετρα μπορεί να υπολογιστεί η έντασης της βροχής με μονάδα μέτρησης τα χιλιοστά ανά ώρα [26].

Αρκετοί ατμοσφαιρικοί παράγοντες επηρεάζουν την εμφάνιση και την ένταση των βροχοπτώσεων. Η θερμοκρασία, η υγρασία, η ηλιακή ακτινοβολία, η ατμοσφαιρική πίεση, η μικροφυσική των νεφών είναι μερικοί από τους παράγοντες που επηρεάζουν την εμφάνιση βροχοπτώσεων και την έντασή της [9].

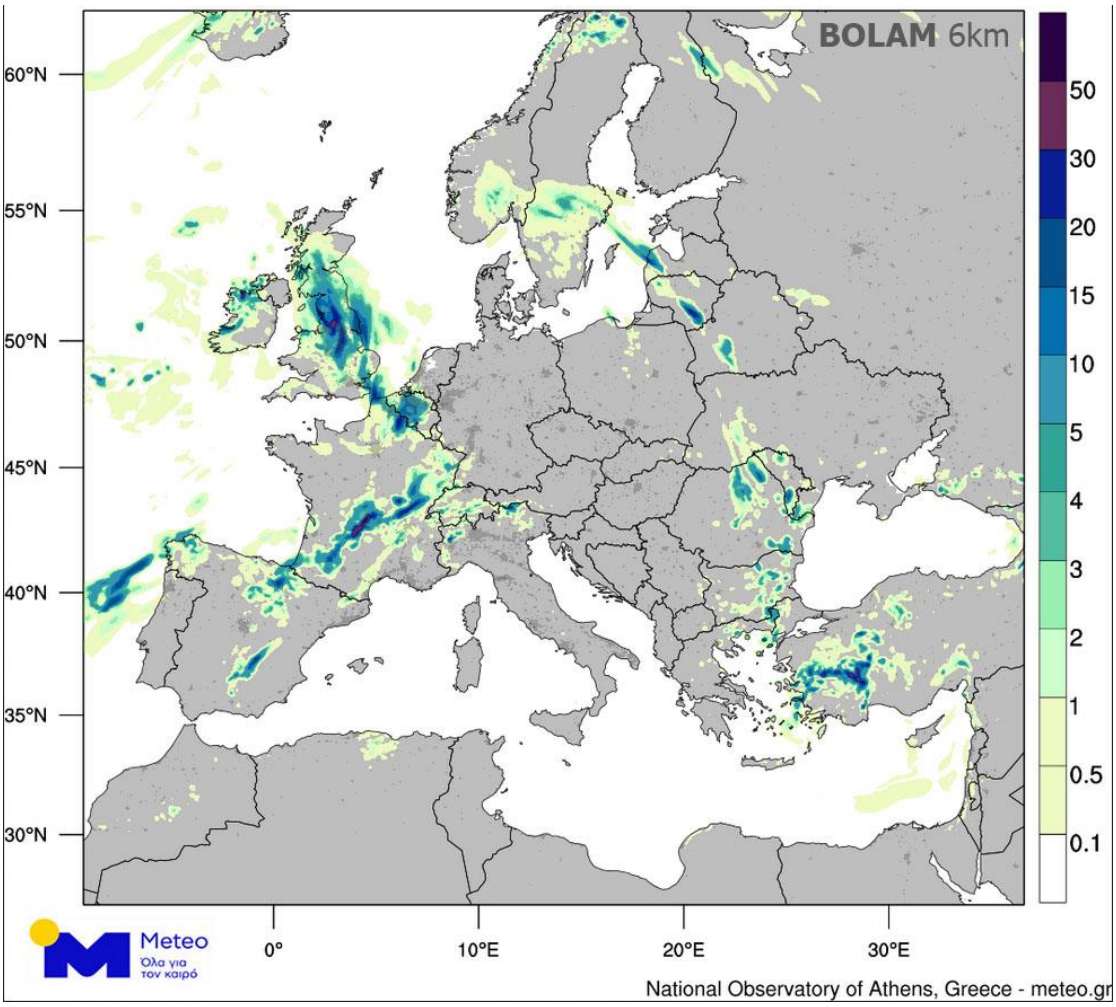
3.2.2 Πρόβλεψη καιρού

Η πρόβλεψη καιρού είναι η πρόβλεψη των ατμοσφαιρικών συνθηκών όπως η θερμοκρασία, η υγρασία, το σημείο δρόσου, η βροχόπτωση και η ταχύτητα του ανέμου σε μια καθορισμένη τοποθεσία. Όργανα όπως βαρόμετρα, θερμόμετρο και ραντάρ χρησιμοποιούνται για τη συλλογή δεδομένων για πρόβλεψη όπως οι τρέχουσες καιρικές συνθήκες, τα καιρικά μοτίβα, η παρακολούθηση της κίνησης του αέρα και των νεφών [27]. Η πρόβλεψη του καιρού αποτελεί μια σημαντική διαδικασία καθώς λαμβάνεται υπόψη για την προστασία της ζωής και της περιουσίας των πολιτών. Επίσης, οι προβλέψεις που βασίζονται στη θερμοκρασία και τις βροχοπτώσεις είναι σημαντικές για τη γεωργία και επομένως για τους εμπόρους στις αγορές εμπορευμάτων. Σε καθημερινή βάση, πολλοί άνθρωποι χρησιμοποιούν μετεωρολογικές προβλέψεις για να καθορίσουν την καθημερινότητα τους. Δεδομένου ότι οι υπαίθριες δραστηριότητες εξαρτώνται σε μεγάλο βαθμό από τα καιρικά φαινόμενα όπως η έντονη βροχόπτωση, το χιόνι και τον άνεμο, οι προβλέψεις μπορούν να χρησιμοποιηθούν για τον προγραμματισμό δραστηριοτήτων γύρω από αυτά τα γεγονότα.

Οι άνθρωποι προσπάθησαν να προβλέψουν τον καιρό ανεπίσημα για χιλιετίες και επίσημα από τον 19ο αιώνα. Οι καιρικές προβλέψεις γινόταν παλαιότερα συλλέγοντας ποσοτικά δεδομένα σχετικά με την τρέχουσα κατάσταση της ατμόσφαιρας, της γης και των ωκεανών και χρησιμοποιώντας τη μετεωρολογία για να προβλέψουν μελλοντικά πώς θα αλλάξει η ατμόσφαιρα και ο καιρός σε ένα δεδομένο μέρος. Με την πάροδο των ετών και από το 1922 ο L.F. Richardson έφερε την επανάσταση στον τομέα της πρόβλεψης του καιρού φέρνοντας στο προσκήνιο την αριθμητική πρόβλεψη καιρού. Η αριθμητική πρόβλεψη καιρού (NWP) ορίζεται ως η παραγωγή μιας πρόβλεψης μέσω ενός ολοκληρωμένου συνόλου μαθηματικών εξισώσεων που περιγράφουν ουσιαστικά όλες τις δυναμικές και φυσικές διεργασίες στην ατμόσφαιρα χρησιμοποιώντας αριθμητικές διαδικασίες [28].

Τα τελευταία χρόνια έχουν αναπτυχθεί πάρα πολλά μοντέλα αριθμητικής πρόβλεψης του καιρού σε παγκόσμια αλλά και πιο τοπική κλίμακα. Μερικά από τα πιο γνωστά μοντέλα αριθμητικής πρόβλεψης του καιρού στα οποία βασίζονται στο μεγαλύτερο ποσοστό οι προγνώσεις καιρού παγκοσμίως είναι το Αμερικανικό αριθμητικό μοντέλο πρόγνωσης καιρού (Global Forecasting System ή GFS) καθώς και το Ευρωπαϊκό αριθμητικό μοντέλο πρόγνωσης καιρού (European Centre for Medium-Range Weather Forecasts ή ECMWF).

Εκτός, των αριθμητικών μοντέλων πρόγνωσης καιρού παγκόσμιας κλίμακας, υπάρχουν και τα αριθμητικά μοντέλα που βασίζονται και εξειδικεύονται σε πιο μικρές περιοχές, επιπέδου χώρας. Στην Ελλάδα, το Εθνικό Αστεροσκοπείο Αθηνών και πιο συγκεκριμένα η μονάδα METEO, έχει αναπτύξει και λειτουργεί από το 1999 το αριθμητικό μοντέλο πρόγνωσης καιρού BOLAM. Η τρέχουσα εγκατάσταση του BOLAM στο Ε.Α.Α/METEO περιλαμβάνει ένα πλέγμα που καλύπτει την Ευρώπη, τη λεκάνη της Μεσογείου και τη Βόρεια Αφρική. Το πλέγμα αποτελείται από 770 x 702 σημεία με οριζόντια ανάλυση πλέγματος 0,06° (~6km), καθώς και 40 κάθετα επίπεδα της ατμόσφαιρας [29].

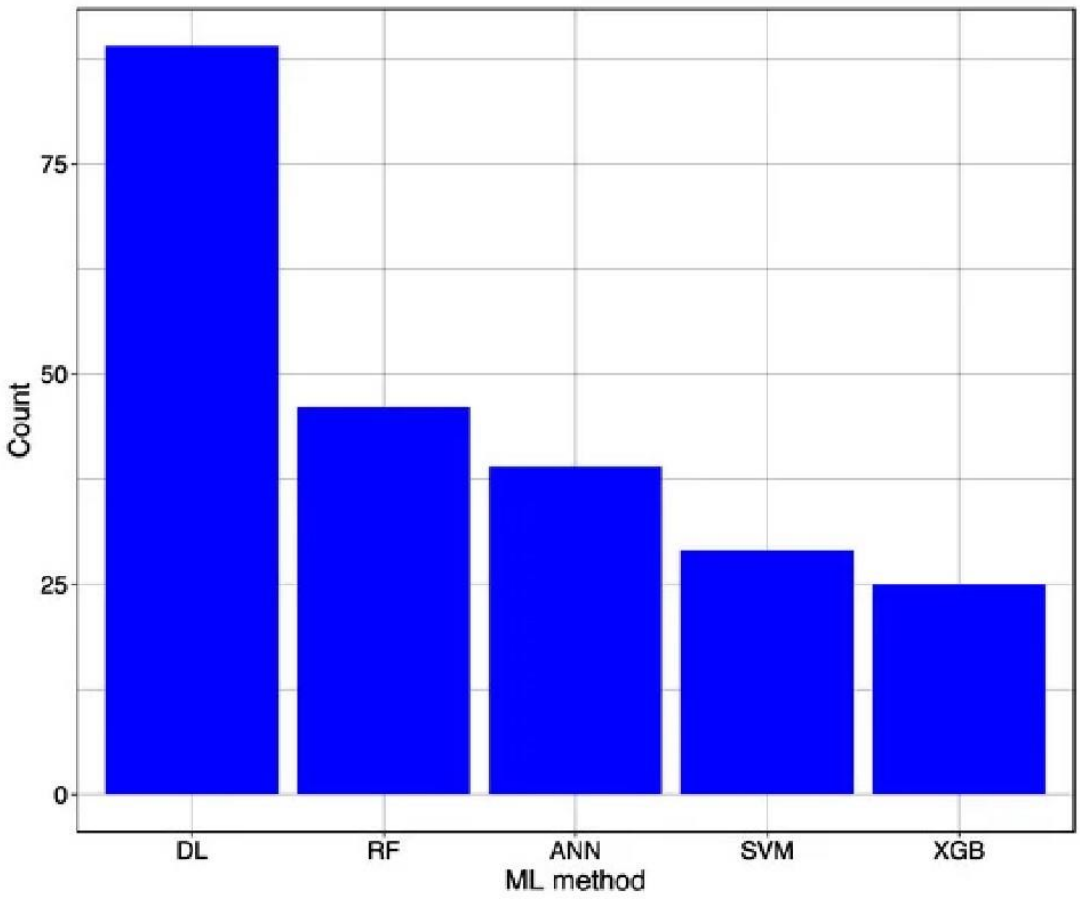


Εικόνα 11. Χάρτης πρόβλεψης της βροχής σε πλέγμα για την Ευρώπη από το αριθμητικό μοντέλο πρόγνωσης καιρού BOLAM του Εθνικού Αστεροσκοπείου Αθηνών/METEO.

3.2.3 Πρόβλεψη καιρού με την χρήση Μηχανικής Μάθησης

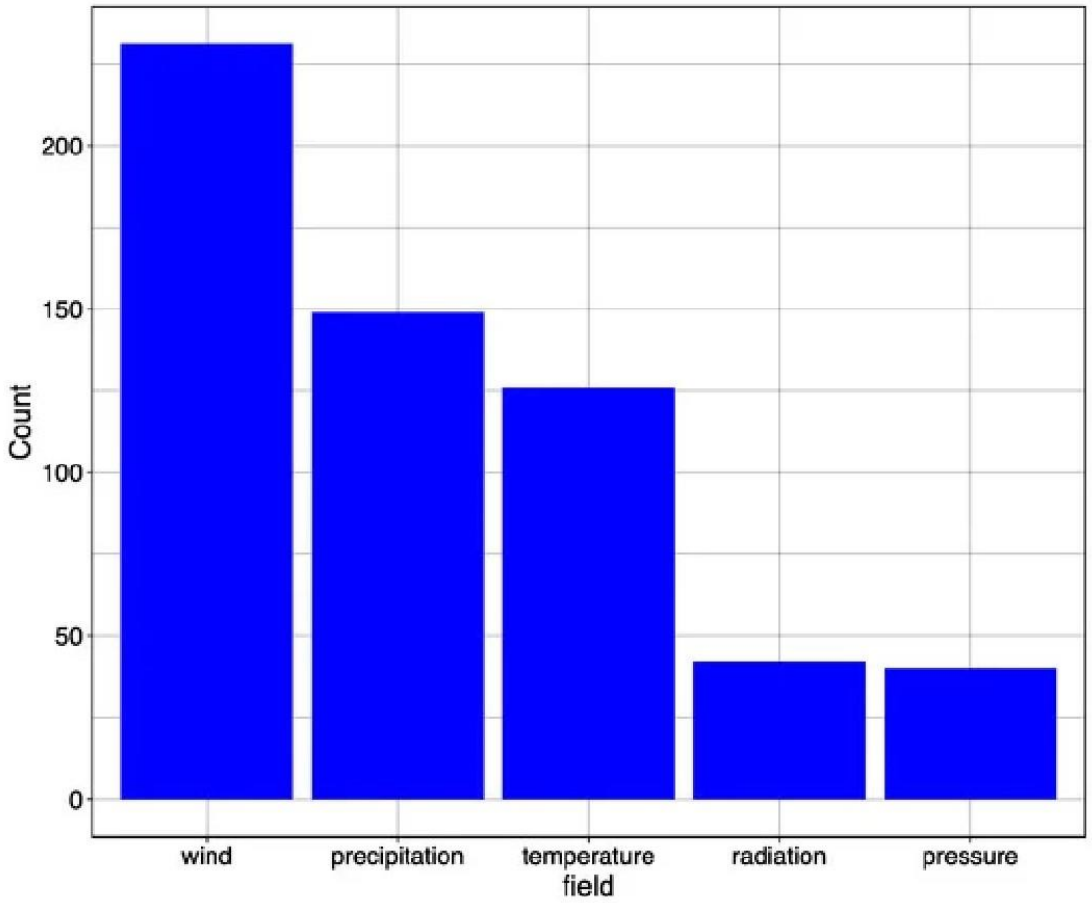
Από τις αρχές του 21ου αιώνα, με την έλευση των μεγάλων δεδομένων, των αποτελεσματικών υπερ-υπολογιστών με μεγάλη υπολογιστική ισχύ και του επιστημονικού ενδιαφέροντος για τις αναδυόμενες νέες μεθόδους, άρχισαν να αυξάνονται οι εφαρμογές της μηχανικής μάθησης σε πολλά προβλήματα και τομείς της επιστήμης. Έτσι, και στον τομέα της Μετεωρολογίας τα τελευταία χρόνια έχει ανθίσει η χρήση της μηχανικής μάθησης και των μοντέλων της για την πρόβλεψη καιρού αλλά και παραμέτρων της όπως η θερμοκρασία, ο άνεμος και η βροχή.

Πολλές τεχνικές και μοντέλα της μηχανικής μάθησης έχουν εφαρμοστεί σε διάφορες χώρες του κόσμου με σκοπό την βελτίωση της πρόβλεψης του καιρού ή κάποιας παραμέτρου. Σύμφωνα με μια πρόσφατη έρευνα από τους Bogdan Bochenek και Zbigniew Ustrnul που έγινε πάνω σε 500 εργασίες σχετικά με μεθόδους μηχανικής μάθησης που εφαρμόστηκαν στον τομέα του κλίματος και της αριθμητικής πρόβλεψης καιρού, διαπιστώθηκε πως για τους επιστήμονες της ατμόσφαιρας, η πιο συχνή και ενδιαφέρουσα ομάδα τεχνικών μηχανικής μάθησης είναι η εποπτευόμενη μάθηση (Supervised learning). Ενώ, οι πιο συχνές μέθοδοι και μοντέλα μηχανικής μάθησης που χρησιμοποιούνται κατά κόρον είναι κατά σειρά οι μέθοδοι Βαθιάς Μάθησης (Deep Learning), τα Τυχαία Δέντρα Αποφάσεων (Random Forest), τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks), οι Μηχανές Διανυσματικής Υποστήριξης (Support Vector Machines) και η τεχνική XGBoost (Εικόνα 12).



Εικόνα 12. Οι πιο συνηθισμένες μέθοδοι μηχανικής μάθησης σε άρθρα που σχετίζονται με την πρόγνωση καιρού [30].

Ενώ, τέλος οι πιο συχνές παράμετροι για τις οποίες οι περισσότεροι ερευνητές εφάρμοσαν μεθόδους μηχανικής μάθησης για την πρόβλεψή τους ήταν είναι ο άνεμος και η βροχόπτωση (Εικόνα 13).



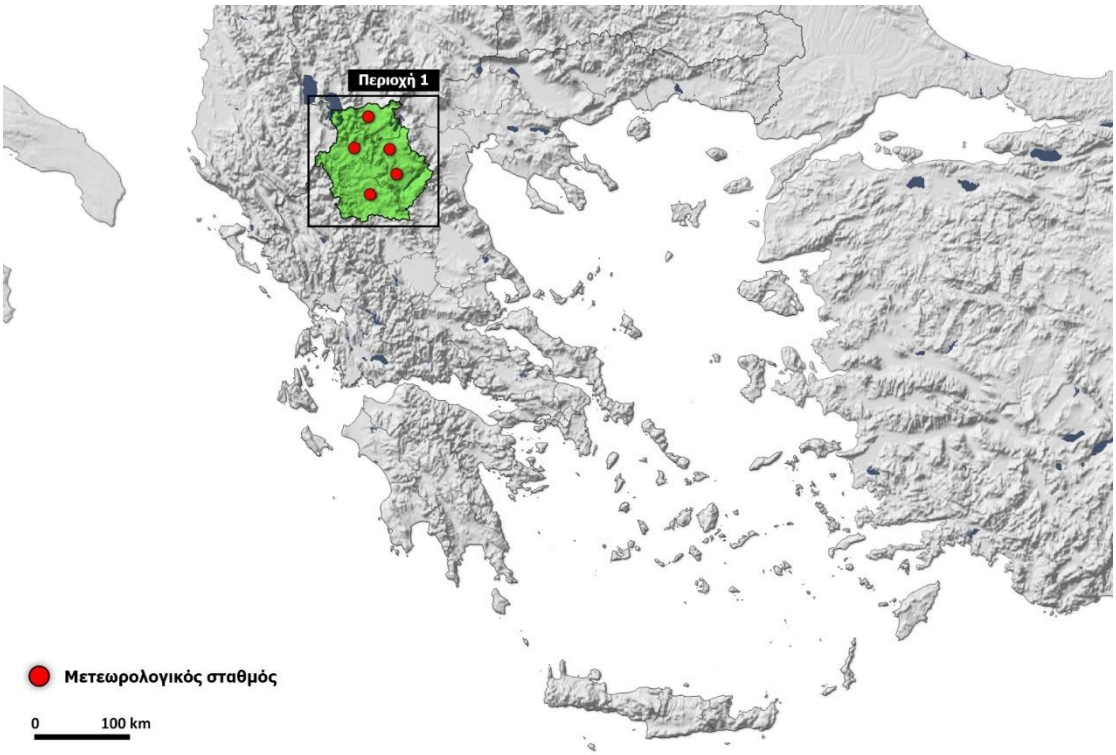
Εικόνα 13. Οι πιο συνηθισμένοι παράμετροι/μετεωρολογικά πεδία σε άρθρα που σχετίζονται με την εφαρμογή μεθόδων μηχανικής μάθησης για την πρόβλεψη καιρού [30].

4

Δεδομένα και τεχνικές

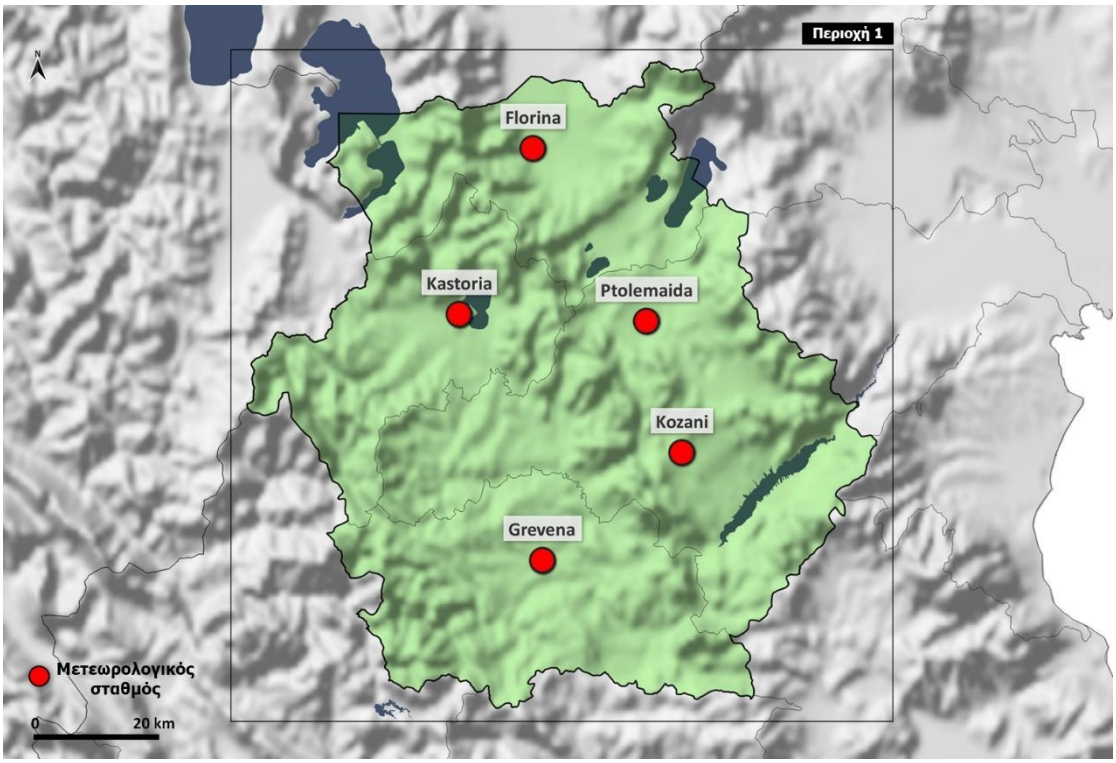
4.1 Περιοχή μελέτης και δεδομένα

Η περιοχή μελέτης είναι η Δυτική Μακεδονία, που βρίσκεται στη Βόρεια Ελλάδα (Εικόνα 14). Είναι μια ορεινή περιοχή έκτασης 9.451 km^2, και αποτελείται κυρίως από οροπέδια με μέσο υψόμετρο πάνω από τα 400-500μ. Η Δυτική Μακεδονία αποτελείται από 4 περιφερειακές ενότητες, αυτές είναι των Γρεβενών, της Καστοριάς, της Κοζάνης και της Φλώρινας.



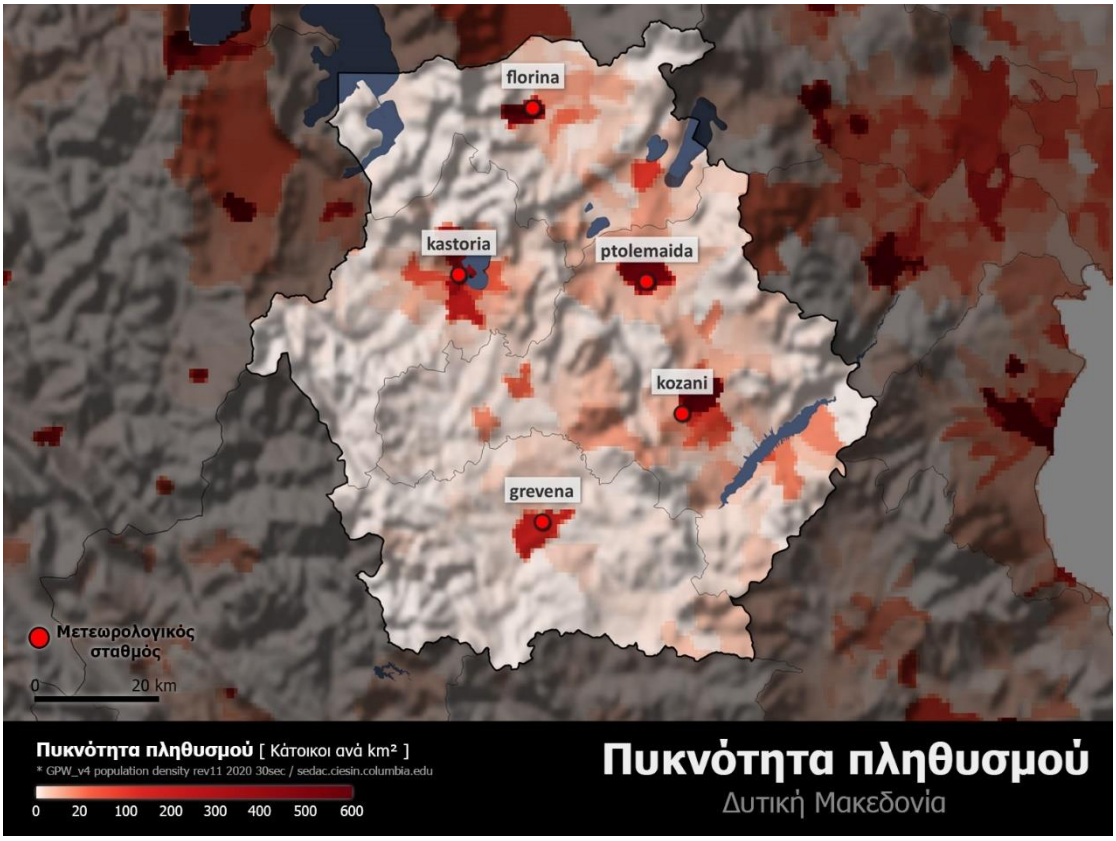
Εικόνα 14. Η περιοχή μελέτης στην Ελλάδα.

Χρησιμοποιήθηκαν δεδομένα από 3 πηγές. Τα αριθμητικά δεδομένα επανανάλυσης μοντέλου ανακτήθηκαν από τη Ευρωπαϊκή Βάση Δεδομένων για το κλίμα Copernicus Climate Database (ERA-5) [11] και τα δορυφορικά δεδομένα από το GRIDSAT-B1 της Εθνικής Διοίκησης για την Ατμόσφαιρα και τους Ωκεανούς (NOAA) [19]. Τέλος, χρησιμοποιήθηκαν τα δεδομένα 5 μετεωρολογικών σταθμών από το Δίκτυο Αυτόματων Μετεωρολογικών Σταθμών του Εθνικού Αστεροσκοπείου Αθηνών/ΜΕΤΕΟ [31] (Εικόνα 15). Περισσότερες πληροφορίες για τις 3 πηγές δεδομένων παρατίθενται παρακάτω στην περιγραφή των δεδομένων.



Εικόνα 15. Η περιοχή μελέτης στην Δυτική Μακεδονία με τους 5 μετεωρολογικούς σταθμούς που χρησιμοποιήθηκαν στην ανάλυση.

Η επιλογή των 5 από τους 38 σταθμούς που είναι εγκατεστημένοι στη Δυτική Μακεδονία [31] βασίστηκε στη διαθεσιμότητα δεδομένων τα έτη 2017 έως το 2020, στη γεωγραφική κατανομή των σταθμών στις περιφερειακές ενότητες και στη γειτνίαση με την πυκνότητα πληθυσμού (Εικόνα 16). Αυτοί οι πέντε μετεωρολογικοί σταθμοί είναι εγκατεστημένοι κοντά στις πέντε μεγαλύτερες πόλεις της Δυτικής Μακεδονίας. Πιο συγκεκριμένα, οι τέσσερις σταθμοί (Γρεβενά, Καστοριά, Κοζάνη, Φλώρινα) είναι οι πρωτεύουσες κάθε περιφερειακής ενότητας αντίστοιχα.



Εικόνα 16. Η περιοχή μελέτης στην Δυτική Μακεδονία με τους 5 μετεωρολογικούς σταθμούς καθώς και την πυκνότητα του πληθυσμού σε αυτή. Πηγή δεδομένων: sedac.ciesin.columbia.edu.

4.1.1 Περιγραφή των δεδομένων

Από το σύνολο δεδομένων μοντέλου επανανάλυσης της ατμόσφαιρας ERA-5, ανακτήθηκαν δεδομένα από παραμέτρους της ατμόσφαιρας σε 20 επίπεδα (ύψη) σε αυτή καθώς και παράμετροι σε μεμονωμένα επίπεδα, όπως η θερμοκρασία, η σχετική υγρασία ο άνεμος κ.α. Από το σύνολο δεδομένων GRIDSAT-B1, χρησιμοποιήθηκε το κανάλι 10.8μm (θερμοκρασία νεφών) και το κανάλι 6.2μm (υδρατμοί). Τα δεδομένα από τους μετεωρολογικούς σταθμούς του Δικτύου Αυτόματων Μετεωρολογικών Σταθμών του Εθνικού Αστεροσκοπείου Αθηνών/ΜΕΤΕΟ, περιλάμβαναν επίγειες μετρήσεις όπως η θερμοκρασία, σχετική υγρασία, ατμοσφαιρική πίεση, ταχύτητα και κατεύθυνση ανέμου και ρυθμό βροχόπτωσης. Στον Πίνακα 1 αναφέρονται οι πηγές δεδομένων και κάποιες πληροφορίες για αυτά.

Πίνακας 1. Χαρακτηριστικά των δεδομένων.

Πηγή δεδομένων	**Τύπος**	**Γεωγραφική περιοχή**	**Χρονικό εύρος**	**Ρυθμός δειγματοληψίας**

ERA5 - Δεδομένα μοντέλου	Δεδομένα σε πλέγμα	19°E, 23°E, 23°N, 23°N	2017-2020	3 ώρες
GRIDSAT-B1 Δορυφορικά δεδομένα	Δεδομένα σε πλέγμα	19°E, 23°E, 23°N, 23°N	2017-2020	3 ώρες
Δεδομένα Μετ. Σταθμών (E.A.A/meteo.gr)	Δεδομένα σε μορφή .txt	Γρεβενά, Καστοριά, Κοζάνη, Πτολεμαΐδα, Φλώρινα	2017-2020	10 λεπτά με μετατροπή σε 3 ώρες με τεχνικές επαναδειγματοληψίας

Εικόνα 17. Οι 3 πηγές δεδομένων της εργασίας μας.

4.1.1.1 Δεδομένα μοντέλου επανανάλυσης της ατμόσφαιρας - ERA5

Στο πλαίσιο της Υπηρεσίας Κλιματικής Αλλαγής Copernicus (C3S), το Ευρωπαϊκό Κέντρο Μεσοπρόθεσμων Μετεωρολογικών Προγνώσεων (ECMWF) παράγει την έκδοση ενός μοντέλου επανανάλυσης δεδομένων της ατμόσφαιρας ERA-5 στο οποίο ενσωματώσει μια λεπτομερή καταγραφή της κατάστασης της ατμόσφαιρας, της επιφάνειας της γης και των ωκεανών από το 1950 και μετά παγκοσμίως. Πρόκειται δηλαδή για ιστορικά δεδομένα επανανάλυσης της κατάστασης στην ατμόσφαιρα αλλά και στο έδαφος. Τα δεδομένα παρέχονται σε οριζόντια ανάλυση 25 χιλιομέτρων, και παρέχονται σε ρυθμό μιας ώρας για τη διάρκεια όλων των ετών. Μέσα σε αυτά τα δεδομένα μπορεί κάποιος να εντοπίσει παραμέτρους της ατμόσφαιρας όπως η θερμοκρασία, η υγρασία, ο άνεμος καθώς και βροχόπτωση μεταξύ άλλων σε παρελθοντικό χρόνο για οποιαδήποτε περιοχή του πλανήτη.

Ο ρόλος των δεδομένων μοντέλου επανανάλυσης στις εφαρμογές παρακολούθησης του κλίματος είναι πλέον ευρέως αναγνωρισμένος. Τα δεδομένα ERA-5 του ECMWF (Dee et al., 2011) χρησιμοποιούνται τακτικά, μαζί με άλλα σύνολα δεδομένων, ως βασικά δεδομένα για την ετήσια αξιολόγηση της κατάστασης του κλίματος του Παγκόσμιου Οργανισμού Μετεωρολογίας (WMO) και στις αξιολογήσεις που πραγματοποιούνται από την Διακυβερνητική Επιτροπής για την Κλιματική Αλλαγή (IPCC). Συνδυάζοντας βέλτιστα επίγειες και δορυφορικές παρατηρήσεις καθώς και δεδομένων μοντέλων, τα δεδομένα

επανανάλυσης ERA-5 παρέχουν μια τεράστια γκάμα πληροφοριών και παραμέτρων για την ατμόσφαιρα και το κλίμα χωρίς απώλειες δεδομένων. Τα δεδομένα ERA-5 έχουν βρει ευρεία εφαρμογή στις ατμοσφαιρικές επιστήμες, ιδίως σε επιχειρησιακά μετεωρολογικά κέντρα όπου, για παράδειγμα, οι επαναναλύσεις χρησιμοποιούνται για την αξιολόγηση του αντίκτυπου της παρατήρησης των αλλαγών του κλίματος, για τη μέτρηση της προόδου στη μοντελοποίηση και τις δυνατότητες αφομοίωσης καθώς επίσης και για εκπαίδευση μοντέλων μηχανικής μάθησης όπου είναι απαραίτητα τα ιστορικά δεδομένα για μια περιοχή. Τα δεδομένα παρέχονται μέσω της σελίδας https://cds.climate.copernicus.eu και είναι προσβάσιμα και από το απλό κοινό [11].

Στην εργασία μας χρησιμοποιήσαμε δεδομένα επανανάλυσης της ατμόσφαιρας ERA-5 με σκοπό την εκπαίδευση μοντέλων μηχανικής μάθησης από ιστορικά δεδομένα της ατμόσφαιρας για την κάθε περιοχή. Σε αυτή την εργασία χρησιμοποιήθηκαν παράμετροι της ατμόσφαιρας που σχετίζονται με την διαδικασία παραγωγής της βροχής. Στον Πίνακα 2 παρουσιάζονται αναλυτικά οι παράμετροι καθώς και τα επίπεδα (ύψη) της ατμόσφαιρας.

Πίνακας 2. Χαρακτηριστικά των δεδομένων επανανάλυσης της ατμόσφαιρας ERA-5.

Παράμετρος	Επίπεδα [hPa]	Μονάδα μέτρησης	Χρονικό βήμα
Pressure Levels			
Temperature	[250 .. 1000] ανά 50 hPa	°C	3 ώρες
Potential Vorticity	[250 .. 1000] ανά 50 hPa	$K\ m^2\ kg^{-1}\ s^{-1}$	3 ώρες
Specific Humidity	[250 .. 1000] ανά 50 hPa	$kg\ kg^{-1}$	3 ώρες
Geopotential Height	[250 .. 1000] ανά 50 hPa	$m^2\ s^{-2}$	3 ώρες
(U,V) Wind Components	[250 .. 1000] ανά 50 hPa	$m\ s^{-1}$	3 ώρες
Single Levels			
Convective Available Potential Energy	-	$J\ kg^{-1}$	3 ώρες
Boundary layer dissipation	-	$J\ m^{-2}$	3 ώρες
Mean sea level pressure	-	Pa	3 ώρες
Total column cloud ice water	-	$kg\ m^{-2}$	3 ώρες

Total column cloud liquid water	-	kg m^{-2}	3 ώρες
Total column snow water	-	kg m^{-2}	3 ώρες
Total column supercooled liquid water	-	kg m^{-2}	3 ώρες
Total column water vapour	-	kg m^{-2}	3 ώρες
Volumetric soil water layer 1	0 - 7cm	m^3 m^{-3}	3 ώρες
Zero degree level	-	m	3 ώρες

Συνολικά από τα δεδομένα ERA-5 χρησιμοποιήσαμε 135 παραμέτρους.

4.1.1.2 Δορυφορικά δεδομένα - GRIDSAT-B1

Οι γεωστατικοί δορυφόροι παρέχουν συνεχείς και υψηλής χρονικής ανάλυσης παρατηρήσεις της Γης από τη δεκαετία του 1970. Το σύνολο δεδομένων Gridded Satellite (GridSat) περιλαμβάνει παρατηρήσεις από το ορατό, το υπέρυθρο κανάλι καθώς και τα υπέρυθρα κανάλια υδρατμών από πολλούς δορυφόρους πολλών χωρών σε όλο τον κόσμο. Το αποτέλεσμα είναι ένα αρχείο κλιματικών δεδομένων που χρησιμοποιείται ήδη από τη μετεωρολογική κοινότητα. Παραδείγματα περιλαμβάνουν την εκ νέου ανάλυση των τροπικών κυκλώνων, τις μελέτες για τις βροχοπτώσεις παγκοσμίως καθώς και για την εκπαίδευση μοντέλων μηχανικής μάθησης. Τα δεδομένα παρέχονται σε οριζόντια ανάλυση 7 χιλιομέτρων και με χρονικό βήμα 3 ωρών.

Τα δεδομένα παρέχονται στον παρακάτω σύνδεσμο: https://www.ncei.noaa.gov/products/gridded-geostationary-brightness-temperature και είναι προσβάσιμα μόνο με την εγγραφή ως μέλος της κοινότητας [19]. Στον Πίνακα 3 παρουσιάζονται αναλυτικά οι παράμετροι από τα δορυφορικά δεδομένα που χρησιμοποιήθηκαν για την εργασία αυτή.

Πίνακας 3. Χαρακτηριστικά των δορυφορικών δεδομένων GRIDSAT-B1.

Παράμετρος	**Μονάδα μέτρησης**	**Χρονικό βήμα**
Brightness Temperature [BT] channel 10.8μm	°C	3 ώρες
Water Vapor [WV] channel 6.2μm	°C	3 ώρες
Deep convection [BT - WV]	°C	3 ώρες

Συνολικά από τα δορυφορικά δεδομένα GRIDSAT-B1 χρησιμοποιήσαμε 3 παραμέτρους.

4.1.1.3 Δεδομένα Μετεωρολογικών Σταθμών - E.A.A/meteo.gr

Τα τελευταία 16 χρόνια, το Ινστιτούτο Περιβαλλοντικής Έρευνας και Βιώσιμης Ανάπτυξης (ΙΕΠΒΑ) του Εθνικού Αστεροσκοπείου Αθηνών (Ε.Α.Α) έχει αναπτύξει και λειτουργεί ένα δίκτυο αυτόματων μετεωρολογικών σταθμών σε όλη την Ελλάδα. Το κίνητρο πίσω από την ανάπτυξη του δικτύου είναι η παρακολούθηση των καιρικών συνθηκών στην Ελλάδα με στόχο να υποστηρίξει όχι μόνο τις ανάγκες έρευνας (παρακολούθηση και ανάλυση καιρού, αξιολόγηση μετεωρολογικών προβλέψεων) αλλά και τις ανάγκες διαφόρων κοινοτήτων του παραγωγικού τομέα (γεωργία, κατασκευές, αναψυχής και τουρισμού, κ.λπ.). Μέχρι και τις αρχές του 2023, λειτουργούν 500 μετεωρολογικοί σταθμοί, οι οποίοι παρέχουν δεδομένα σε πραγματικό χρόνο σε διαστήματα 10 λεπτών [31]. Τα δεδομένα παρέχονται μόνο σε συνεννόηση και επικοινωνία με το Ινστιτούτο και την κεντρική σελίδα του https://meteo.gr.

Σε αυτή την εργασία τα δεδομένα των μετεωρολογικών σταθμών χρησιμοποιήθηκαν ως δεδομένα εκπαίδευσης αλλά και ως δεδομένα επαλήθευσης. Η παράμετρος της βροχής για την οποία είναι ο στόχος της πρόβλεψης μας χρησιμοποιήθηκε ως επαλήθευση. Στον Πίνακα 4 παρουσιάζονται αναλυτικά οι παράμετροι και τα χαρακτηριστικά των δεδομένων των επίγειων μετρήσεων από τους μετεωρολογικούς σταθμούς του Δικτύου Αυτόματων Μετεωρολογικών Σταθμών του Εθνικού Αστεροσκοπείου Αθηνών/METEO.

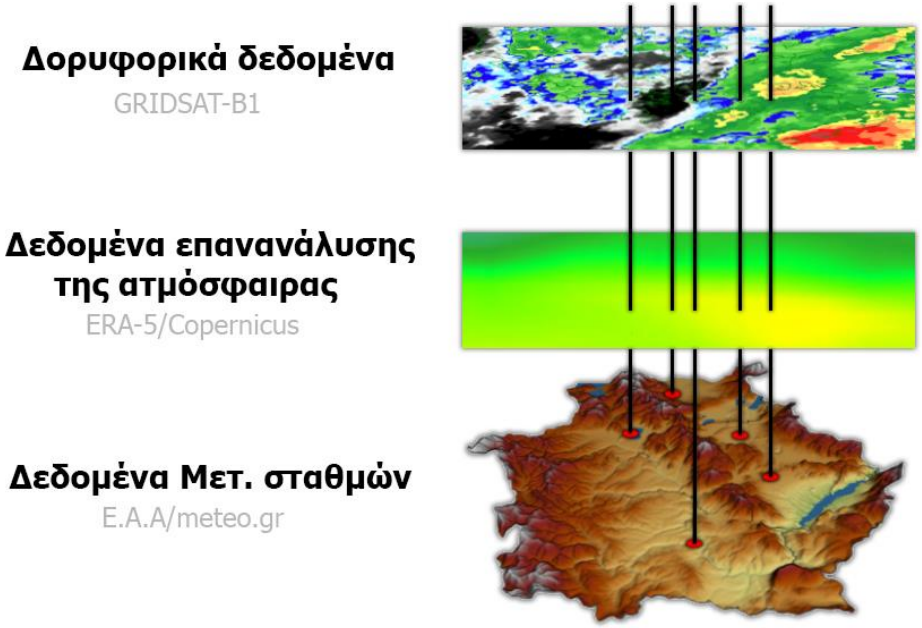
Πίνακας 4. Χαρακτηριστικά των επίγειων μετρήσεων από τους Μετεωρολογικούς Σταθμούς του Δικτύου Αυτόματων Μετεωρολογικών Σταθμών του Εθνικού Αστεροσκοπείου Αθηνών/METEO.

Παράμετρος	**Μονάδα μέτρησης**	**Χρονικό βήμα**
Temperature	°C	10 λεπτά
Humidity	%	10 λεπτά
Dew Point	°C	10 λεπτά
Barometer	hPa	10 λεπτά
Wind Speed	km/h	10 λεπτά
Wind Direction	degrees	10 λεπτά
Rain Rate	mm/h	10 λεπτά
Rain	mm	10 λεπτά

Συνολικά από τα δεδομένα μετεωρολογικών σταθμών του E.A.A/meteo.gr χρησιμοποιήσαμε 8 παραμέτρους. Η παράμετρος της βροχής (rain) χρησιμοποιήθηκε ως επαλήθευση στην εκπαίδευση των μοντέλων μας και πρόκειται για την παράμετρο στόχο.

4.1.2 Ετοιμασία - Προεπεξεργασία δεδομένων

Προκειμένου να πραγματοποιηθεί η μελέτη, πραγματοποιήθηκε ένα σύνολο λειτουργιών και διαδικασιών για την προετοιμασία των δεδομένων [18]. Αρχικά, ήταν βασικό να γίνει δειγματοληψία των σημειακών δεδομένων στο πλησιέστερο σημείο από τα δεδομένα σε πλέγμα (ERA-5 και GRIDSAT-B1) στη θέση των μετεωρολογικών σταθμών (Εικόνα 18).



Εικόνα 18. Η διαδικασία εξαγωγής των δεδομένων από τα δεδομένα σε πλέγμα στα σημεία των μετεωρολογικών σταθμών.

Και οι τρεις πηγές δεδομένων διέθεταν διαφορετικούς ρυθμούς δειγματοληψίας (χρονικό βήμα), επομένως το επόμενο βήμα ήταν να πραγματοποιηθεί μορφοποίηση των δεδομένων σε ένα κοινό χρονικό βήμα δηλαδή σε ρυθμό 3 ωρών. Τα δεδομένα από τους μετεωρολογικούς σταθμούς μετατράπηκαν από διαστήματα 10 λεπτών σε διαστήματα 3 ωρών με μεθόδους συνάθροισης (άθροισμα, μέσος όρος, μέσος), οι οποίες εφαρμόστηκαν 10 λεπτά πριν και μετά το χρονικό βήμα των 3 ωρών. Στην Εικόνα 19 το διάγραμμα ροής με τις εργασίες που εκτελέστηκαν μέχρι την εκπαίδευση των μοντέλων και την παραγωγή των αποτελεσμάτων.

Εικόνα 19. Η ροή των εργασιών/ενεργειών που αναμένονται να εφαρμοστούν στα δεδομένα μέχρι την εκπαίδευση των μοντέλων και την εξαγωγή των αποτελεσμάτων.

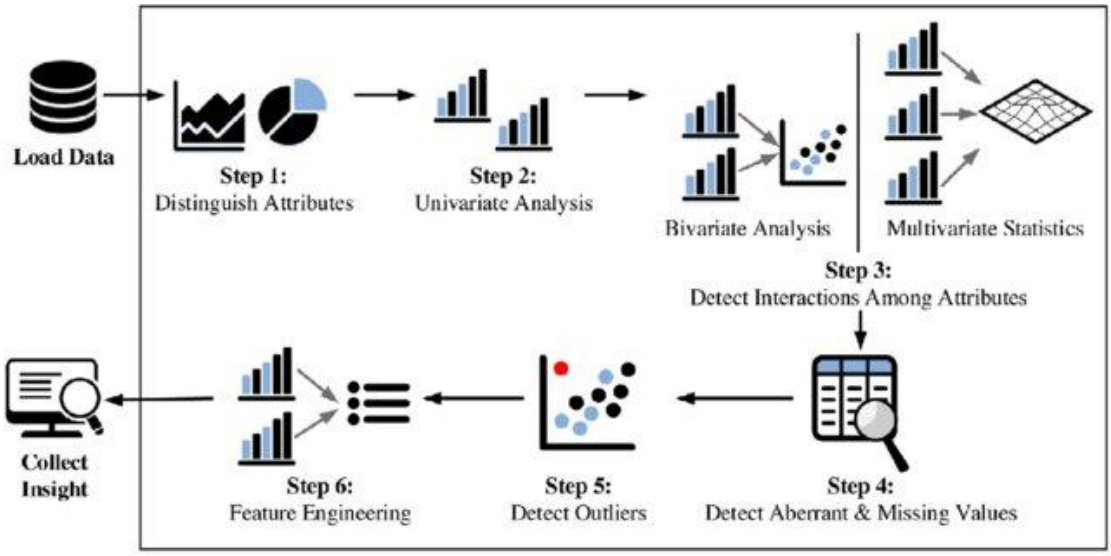
4.1.2.1 Διερευνητική Ανάλυση των Δεδομένων (EDA)

Η διερευνητική ανάλυση δεδομένων (EDA) είναι ένα κρίσιμο βήμα σε οποιαδήποτε έρευνα ανάλυσης δεδομένων. Ο πρωταρχικός στόχος της διερευνητικής ανάλυσης είναι η ανάλυση των δεδομένων για να εξετάσουμε την κατανομή, ακραίες τιμές και ανωμαλίες στο σύνολο δεδομένων με σκοπό την επιλογή της σωστής προσέγγισης στο πρόβλημά μας [32].

Υπάρχουν τέσσερις κύριοι τύποι EDA:

1. Univariate non-graphical: Αυτή είναι η απλούστερη μορφή ανάλυσης δεδομένων, όπου τα δεδομένα που αναλύονται αποτελούνται από μία μόνο μεταβλητή. Ο κύριος σκοπός της μονομεταβλητής ανάλυσης είναι να περιγράψει τα δεδομένα και να βρει μοτίβα που υπάρχουν μέσα σε αυτά.
2. Univariate graphical: Οι μη γραφικές μέθοδοι δεν παρέχουν πλήρη εικόνα των δεδομένων. Απαιτούνται λοιπόν γραφικές παραστάσεις και οπτικοποιήσεις των δεδομένων. Οι συνήθεις τύποι μονομεταβλητών γραφικών περιλαμβάνουν:
    a. Διαγράμματα τα οποία δείχνουν όλες τις τιμές δεδομένων και την κατανομή.
    b. Ιστογράμματα, μια γραφική παράσταση ράβδων στην οποία κάθε ράβδος αντιπροσωπεύει τη συχνότητα (πλήθος) ή την αναλογία περιπτώσεων για ένα εύρος τιμών.
    c. Διαγράμματα με μπάρες, τα οποία απεικονίζουν γραφικά την τα τεταρτημόρια των δεδομένων.
3. Multivariate non-graphical: Τα δεδομένα πολλαπλών μεταβλητών προκύπτουν από περισσότερες από μία μεταβλητές. Οι πολυμεταβλητές μη γραφικές τεχνικές δείχνουν γενικά τη σχέση μεταξύ δύο ή περισσότερων μεταβλητών των δεδομένων μέσω διασταύρωσης πινάκων ή στατιστικών.

4. Multivariate graphical: Τα δεδομένα πολλαπλών μεταβλητών χρησιμοποιούν γραφικά για να εμφανίζουν σχέσεις μεταξύ δύο ή περισσότερων συνόλων δεδομένων.



Εικόνα 20. Τα θεμελιώδη βήματα της διαδικασίας διερευνητικής ανάλυσης δεδομένων (EDA) [33].

Πραγματοποιήσαμε διερευνητική ανάλυση δεδομένων (EDA) χρησιμοποιώντας συνδυασμό των παραπάνω τεχνικών με σκοπό την παροχή συνοπτικών στατιστικών στοιχείων για την παράμετρο που έχουμε ως στόχο να προβλέψουμε, καθώς και για το σύνολο των δεδομένων. Αρχικά, θέλαμε να δούμε και να εξετάσουμε τα ποιοτικά και ποσοτικά χαρακτηριστικά του συνόλου των δεδομένων μας.

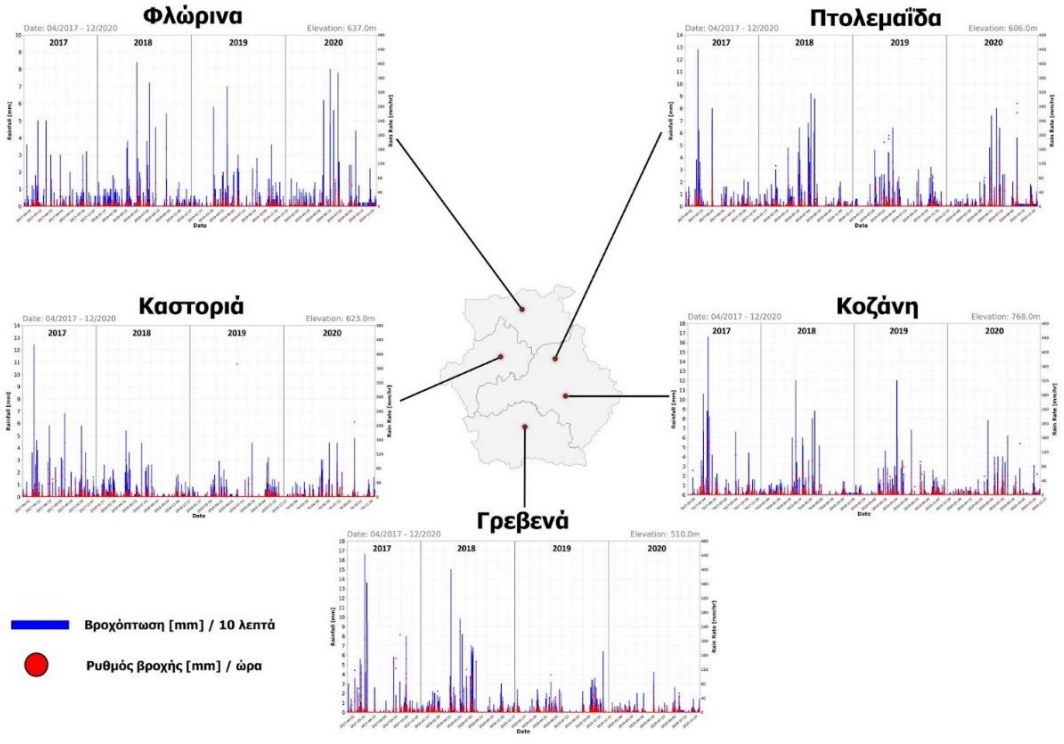
Οπότε, όπως βλέπουμε και στον Πίνακα 5 ο κύβος δεδομένων για κάθε ένα από τα 5 σημεία αποτελείται από 10960 εγγραφές με χρονικό εύρος από 2017-04-02 00:00 μέχρι και τις 2020-12-31 21:00:00. Οι παράμετροι/χαρακτηριστικά είναι 147 και όλες είναι τύπου float64. Όλα τα χαρακτηριστικά είναι συνεχείς μεταβλητές και δεν υπάρχει καμία κατηγορική μεταβλητή στο σύνολο των δεδομένων.

Πίνακας 5. Περιγραφή ενός κύβου δεδομένων από το σύνολο των δεδομένων.

Γραμμές δεδομένων	10960
Στήλες δεδομένων/Χαρακτηριστικά	147
Τύπος δεδομένων	float64
Μέγεθος δεδομένων	~12 mb

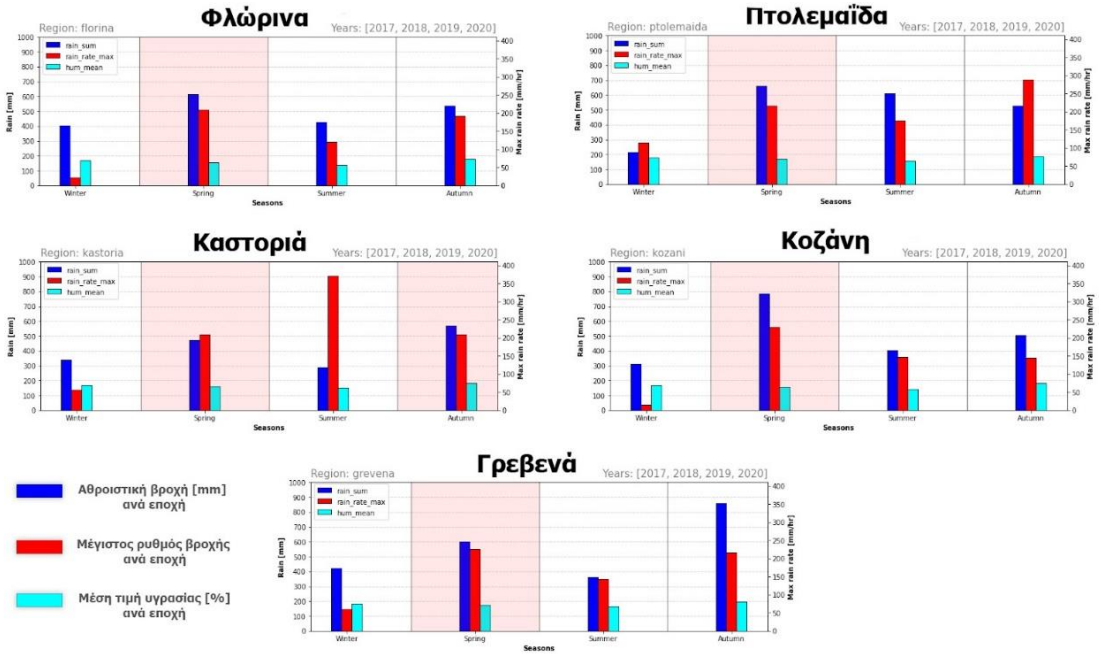
Στη συνέχεια χρειάστηκε να εξετάσουμε την μεταβλητή στόχο που είναι η βροχή, ώστε να εντοπίσουμε κάποιες ακραίες τιμές, καθώς επίσης να εξάγουμε κάποια βασικά συμπεράσματα και στατιστικά στοιχεία για την διακύμανση της. Όπως, παρατηρούμε στην Εικόνα 21 η βροχή αποτελεί μια συνεχή παράμετρο η οποία δέχεται τιμές μεγαλύτερες ή ίσες

από το μηδέν. Παρατηρούμε επίσης πως δεν παρουσιάζεται κάποια τάση στις χρονοσειρές και πως διανέμονται τα ποσά βροχής μέσα στο έτος.



Εικόνα 21. Η κατανομή της βροχής καθώς και του ρυθμού βροχόπτωσης στους 5 μετεωρολογικούς σταθμούς κατά τα έτη 2017-2020.

Επιπλέον, παρατηρώντας και τα στατιστικά στοιχεία ανά εποχή για την παράμετρο της βροχής βλέπουμε πως τα μεγαλύτερα ποσά βροχής καθώς και οι πιο ραγδαίες βροχοπτώσεις σημειώνονται συνήθως την άνοιξη και το φθινόπωρο. Σε όλες τις πόλεις η συνολική βροχή ανά εποχή είναι περισσότερη την άνοιξη, εκτός των Γρεβενών και της Καστοριάς όπου η περισσότερη βροχή μέσα στο έτος εντοπίζεται το φθινόπωρο (Εικόνα 22).



Εικόνα 22. Εποχιακά στατιστικά για την συνολική βροχόπτωση, την μέση υγρασία καθώς και τον μέγιστο ρυθμό βροχόπτωσης στα 5 σημεία των μετεωρολογικών σταθμών κατά τα έτη 2017-2020.

4.1.2.1.1 Έλεγχος στασιμότητας (Stationarity test)

Στη συνέχεια ελέγξαμε την στασιμότητα των χρονοσειρών μας (Stationarity test). Ο έλεγχος στασιμότητας των δεδομένων είναι πολύ σημαντική στην έρευνα όπου οι υποκείμενες μεταβλητές βασίζονται στο χρόνο. Μια στάσιμη χρονοσειρά σημαίνει ότι οι διακυμάνσεις των τιμών της δε διαφοροποιούνται με το χρόνο. Μια μη-στάσιμη χρονοσειρά μπορεί να έχει τάσεις (trends), δηλαδή (αργές) αλλαγές στη μέση τιμή της με το χρόνο. Μια μη-στάσιμη χρονοσειρά μπορεί επίσης να παρουσιάζει περιοδικότητα (periodicity) όταν αναφέρεται σε συγκεκριμένες περιόδους που σχετίζονται με φυσικές εποχές του έτους.

Για το δικό μας σύνολο δεδομένων και την μεταβλητή στόχο που είναι η βροχή εφαρμόσαμε την μέθοδο Augmented Dickey Fuller Test (ADF) (Εξίσωση 5) για να ελέγξουμε την στασιμότητα της χρονοσειράς [16].

$$\Delta Y_t = a + \beta_t + \gamma Y_t + \sum_{j=1}^{p} \left(\delta_j \Delta Y_{t-j}\right) + e_t \qquad (5)$$

Όπου:

- t είναι το χρονικό βήμα (step).
- a είναι μια σταθερά που ονομάζεται σταθερά ολίσθησης (drift).
- β είναι ο συντελεστής της χρονικής τάσης.
- γ είναι ο συντελεστής που παρουσιάζει τη ρίζα της διαδικασίας.

Βασική υπόθεση για να ελέγξουμε την στασιμότητα της χρονοσειράς μας ήταν εάν η τιμή της στατιστικής δοκιμής (Test statistic) είναι μικρότερη από αυτές των 3 επιπέδων των κρίσιμων τιμών (Critical Values), τότε μπορούμε να πούμε ότι η χρονοσειρά είναι στάσιμη (Stationary). Εξετάζοντας τον Πίνακα 6, βλέπουμε πως η τιμή της στατιστικής δοκιμής είναι πολύ μικρότερη από αυτή των 3 επιπέδων των κρίσιμων τιμών (1%, 5%, 10%), οπότε συμπεραίνουμε πως η χρονοσειρά μας για την μεταβλητή της βροχής είναι στάσιμη (Stationary). Η παρούσα υπόθεση θα μπορούσε να αποφευχθεί κοιτώντας μόνο την τιμή της p-value = 0.

Επομένως, καταλήγουμε στο συμπέρασμα πως η μεταβλητή στόχος της βροχής δεν εξαρτάται από τον χρόνο. Στην συνέχεια θα ελέγξουμε αν οι προηγούμενες χρονικές στιγμές επηρεάζουν τις επόμενες με τις τεχνικές της αυτοσυσχέτισης.

Πίνακας 6. Αποτελέσματα από την εφαρμογή της μεθόδου Augmented Dickey-Fuller Test στα δεδομένα του σταθμού των Γρεβενών για τον έλεγχο στασιμότητας της μεταβλητής της βροχής (rain).

Μεταβλητή	**Test Statistic**	**Critical Values**	
Βροχή (Rain)	-31.5	1%	-3.4
		5%	-2.8
		10%	-2.5

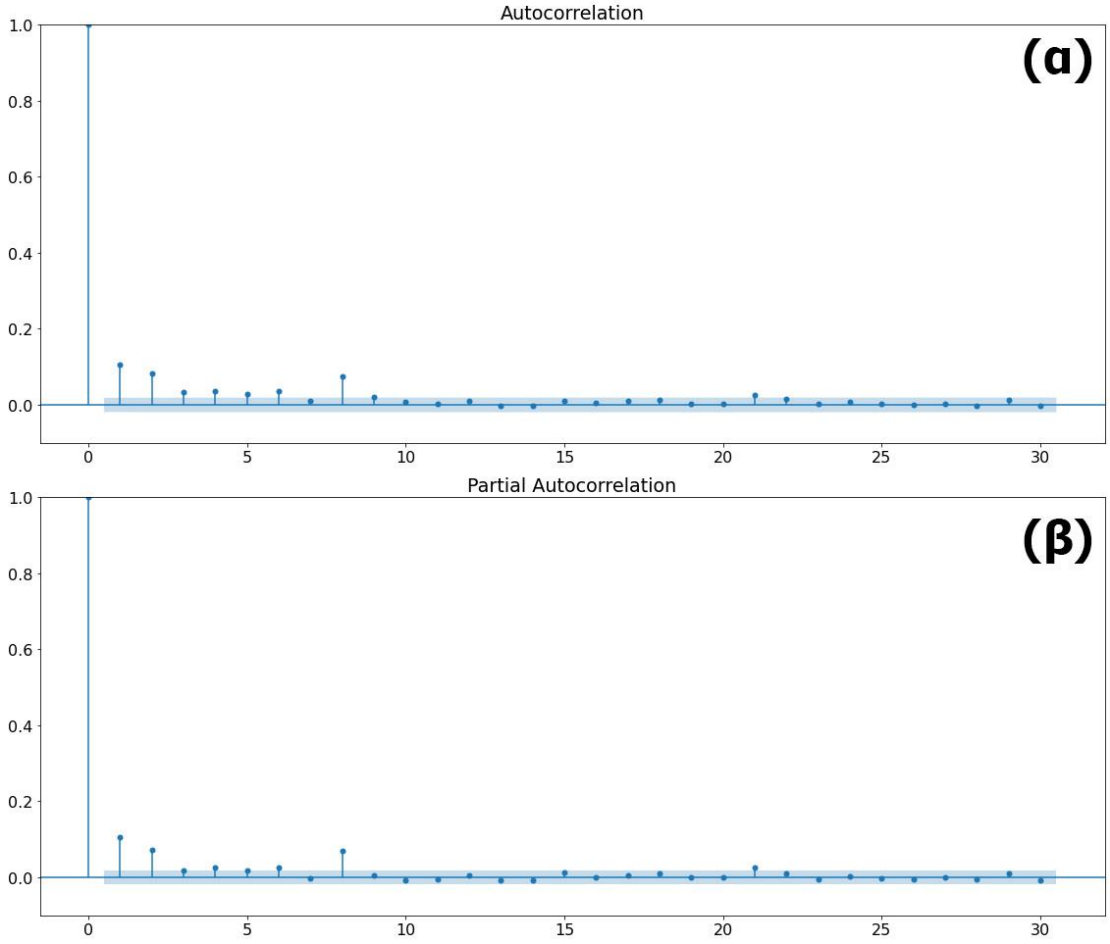
4.1.2.1.2 Συναρτήσεις Συσχέτισης (ACF) και Αυτοσυσχέτισης (PACF)

Ένας ακόμη σημαντικός έλεγχος στην μεταβλητή στόχο ήταν να εξετάσουμε την αυτοσυσχέτιση της σε σχέση με προγενέστερες χρονικές στιγμές. Στην ανάλυση χρονοσειρών ο έλεγχος αυτός γίνεται με δύο γραφικές παραστάσεις που διαμορφώνουν κλασικά στατιστικά μοντέλα: τη συνάρτηση αυτοσυσχέτισης (ACF) και τη συνάρτηση μερικής αυτοσυσχέτισης (PACF).

Αυτές οι δύο συναρτήσεις δείχνουν την ένταση της χρονικής αυτοσυσχέτισης, η καθεμία με τον δικό της τρόπο. Ένα κλασσικό μοντέλο θεωρείται προσαρμοσμένο όταν και οι δύο συναρτήσεις, δεν εμφανίζουν σημαντική αυτοσυσχέτιση. Η συνάρτηση αυτοσυσχέτισης (ACF) λαμβάνεται από τη γραμμική συσχέτιση κάθε x_t της σειράς με τις άλλες προγενέστερες χρονικές στιγμές, όπως x_{t-1}, x_{t-2} και ούτω καθεξής. Η συνάρτηση μερικής αυτοσυσχέτισης (PACF), ωστόσο, λαμβάνει σχεδόν το ίδιο αποτέλεσμα, αφαιρώντας όμως την παρεμβολή των

άλλων τιμών. Στην ACF, για παράδειγμα, η συσχέτιση μεταξύ x_t και x_{t-2} υφίσταται παρεμβολή x_{t-1}. Η PACF αφαιρεί αυτή την παρεμβολή [14].

Στην εργασία μας εφαρμόσαμε τις συναρτήσεις ACF και PACF και στους 5 κύβους δεδομένων για κάθε σταθμό αντίστοιχα. Όπως παρατηρούμε στην Εικόνα 23 παρουσιάζονται τα δύο γραφήματα για την αυτοσυσχέτιση (ACF) στην Εικόνα 23α και για την μερική αυτοσυσχέτιση (PACF) στην Εικόνα 23β. Η μπλε περιοχή απεικονίζει το διάστημα εμπιστοσύνης 95% και είναι ένας δείκτης του ορίου σημαντικότητας. Παρατηρούμε αρχικά πως η τιμή στη χρονική στιγμή t_0 αυτοσυσχετίζεται πλήρως με τον εαυτό της, το οποίο ήταν αναμενόμενο. Επιπλέον, παρατηρούμε πως υπάρχουν αρκετές αυτοσυσχετίσεις που είναι σημαντικά μη μηδενικές. Επομένως, η χρονοσειρά είναι μη τυχαία (non-random). Τέλος, παρατηρούμε πως οι συσχετίσεις μετά την τιμή στην χρονική στιγμή t_1 είναι σημαντικά μικρές και δεν μπορούν να ληφθούν υπόψιν. Οπότε, καταλήγουμε στο συμπέρασμα πως δεν υπάρχει μεγάλη αυτοσυσχέτιση της μεταβλητής στόχου της βροχής με προηγούμενες χρονικές στιγμές, στοιχείο που θα επαληθεύσουμε και στα αποτελέσματα των μοντέλων.



Εικόνα 23. Γραφήματα αυτοσυσχέτισης (ACF) (α) και μερικής αυτοσυσχέτισης (PACF) (β) για τη μεταβλητή της βροχής στο σταθμό των Γρεβενών.

4.1.2.1.3 Συσχετίσεις μεταξύ της μεταβλητής στόχου και των υπόλοιπων χαρακτηριστικών (Correlations)

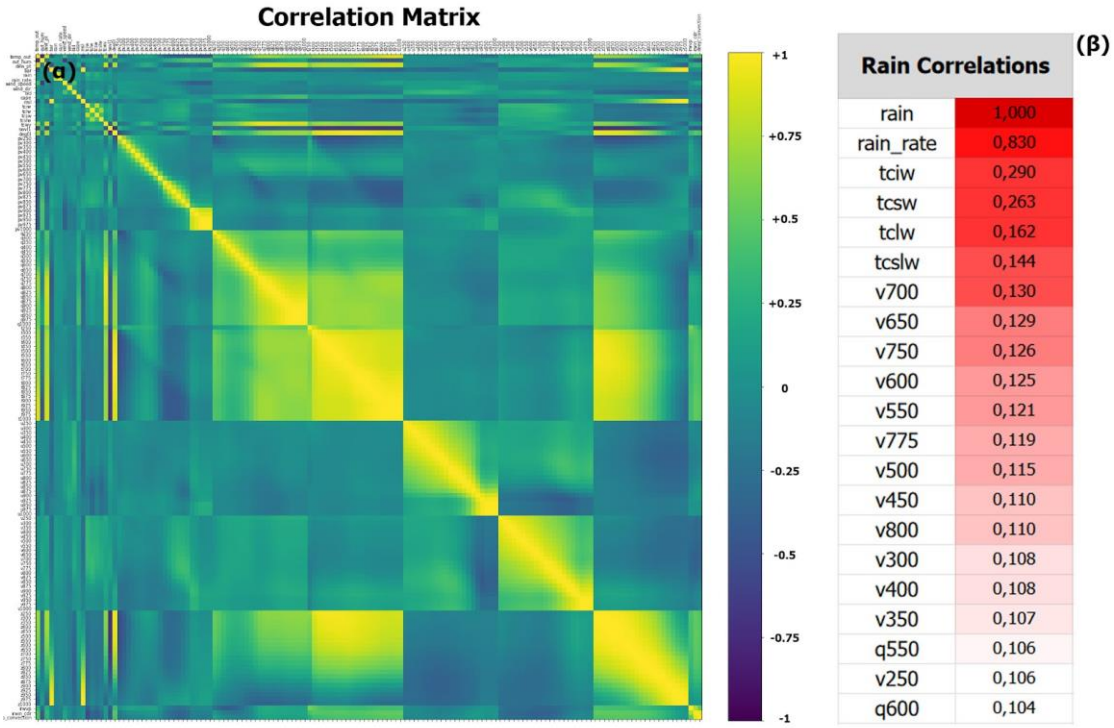
Το τελευταίο βήμα της διερευνητικής ανάλυσης των δεδομένων (EDA) ήταν να εξετάσουμε τις συσχετίσεις μεταξύ του συνόλου των δεδομένων καθώς και της μεταβλητής στόχου της βροχής. Για τον υπολογισμό των συσχετίσεων χρησιμοποιήθηκε η μέθοδος Pearson. Η συσχέτιση Pearson (Εξίσωση 6) μετρά την ισχύ της γραμμικής σχέσης μεταξύ δύο μεταβλητών. Δέχεται τιμές μεταξύ -1 και 1, με τιμή -1 που σημαίνει έντονη αρνητική γραμμική συσχέτιση, 0 δεν υπάρχει συσχέτιση και +1 σημαίνει έντονη θετική συσχέτιση [34].

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (6)$$

Όπου:

- n είναι το μέγεθος του δείγματος.
- x_i, y_i είναι τα 2 χαρακτηριστικά που εξετάζονται αντίστοιχα.
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, η μέση τιμή του x.
- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, η μέση τιμή του y.

Με βάση τα παραπάνω εφαρμόσαμε την μέθοδο Pearson στα σύνολα δεδομένων μας για να εντοπίσουμε και να εξετάσουμε τις συσχετίσεις μεταξύ των μεταβλητών. Όπως βλέπουμε στην Εικόνα 24α υπάρχουν αρκετά υψηλές συσχετίσεις σε μεταβλητές που βρίσκονται στα ίδια ύψη της ατμόσφαιρας καθώς επίσης και σε εντελώς διαφορετικά επίπεδα. Στην Εικόνα 24β παρουσιάζονται τα 20 χαρακτηριστικά με την υψηλότερη θετική συσχέτιση με την μεταβλητή στόχο της βροχής. Και οι δυο πίνακες θα είναι χρήσιμοι κατά την εκπαίδευση των μοντέλων και για τα 5 σημεία, επιτρέποντας μας να επιλέξουμε κάθε φορά διαφορετικές ομάδες χαρακτηριστικών σημειώνοντας κάθε φορά την επιρροή στο αποτέλεσμα.



Εικόνα 24. Πίνακας συσχέτισης των 147 χαρακτηριστικών μεταξύ τους (α) και των 20 χαρακτηριστικών με την μεγαλύτερη συσχέτιση με τη μεταβλητή της βροχής (β).

4.1.2.2 Απώλειες δεδομένων και ακραίες τιμές

Είχαμε έναν μικρό αριθμό απώλειας τιμών, περίπου 10-15 χρονικά βήματα, σχεδόν το 0,001% όλων των δεδομένων. Έτσι, οι τιμές που έλειπαν τις αποκαταστήσαμε με τη μέθοδο της παρεμβολής (interpolation) [32].

Για την ανίχνευση των ακραίων τιμών (outliers), χρησιμοποιήθηκε η μέθοδος «Z-score» [33] (Εξίσωση 7) με όλα εκείνα τα δείγματα που έχουν Z > 3 να απορρίπτονται από το τελικό κύβο δεδομένων.

$$Z = \frac{x_i - \mu}{\sigma} \tag{7}$$

Όπου:

- x_ι η τιμή του συνόλου την χρονική στιγμή i.
- μ ο μέσος όρος του συνόλου.
- σ η τυπική απόκλιση του συνόλου.

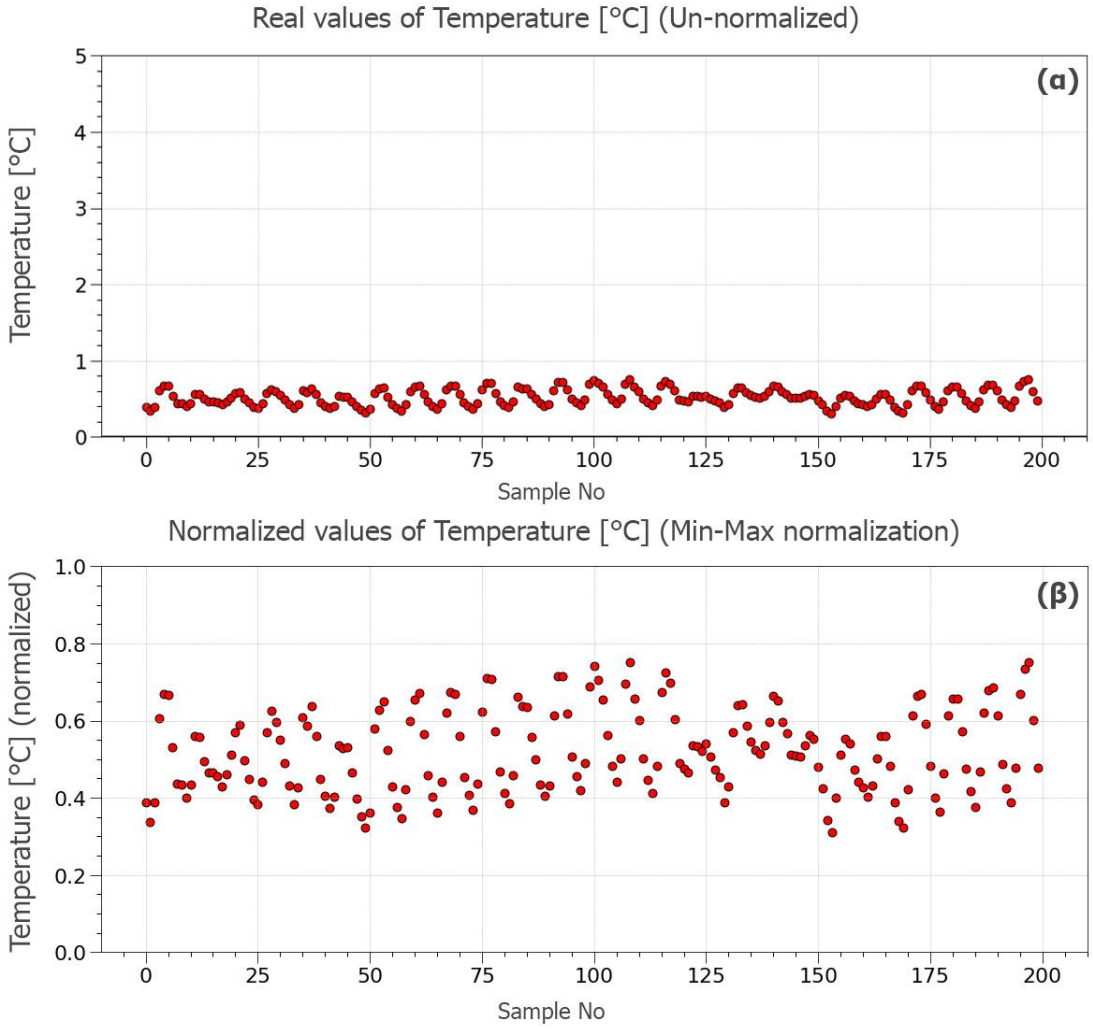
Ως αποτέλεσμα, αφαιρέθηκε μόνο το 0,002% των δεδομένων, οπότε έχουμε έναν τελικό αριθμό 10.940 δειγμάτων για κάθε ένα σημείο.

4.1.2.3 Κανονικοποίηση δεδομένων

Η κανονικοποίηση δεδομένων εφαρμόστηκε μόνο στα δεδομένα για τα Νευρωνικά Δίκτυα του LSTM και CNN. Χρησιμοποιήθηκε η τεχνική min–max (Εξίσωση 8) [16]. Έτσι, όλες οι τιμές των μεταβλητών κανονικοποιήθηκαν μεταξύ 0 και 1. Με αυτόν τον τρόπο, αποφεύχθηκαν χαρακτηριστικά που έπαιρναν πολύ μεγάλες τιμές σε σχέση με τα υπόλοιπα δεδομένα και έτσι είχαν μεγαλύτερη επίδραση στην εφαρμογή των αλγορίθμων μηχανικής μάθησης.

$$z = \frac{x - min(x)}{max(x) - min(x)} \tag{8}$$

Όπου $min(x)$ και $max(x)$ είναι η ελάχιστη και η μέγιστη τιμή, αντίστοιχα. x είναι η τιμή που αναμένεται να κανονικοποιηθεί και z είναι η τελική κανονικοποιημένη τιμή. Μετά την προεπεξεργασία, εφαρμόσαμε αντίστροφη μετατροπή των κανονικοποιημένων τιμών πίσω στις πραγματικές τιμές για την πρόβλεψη και τον υπολογισμό των μετρικών.



Εικόνα 25. Παράδειγμα εφαρμογής της κανονικοποίησης (Normalization) στην μεταβλητή της θερμοκρασίας. Τα μη κανονικοποιημένα δεδομένα στο γράφημα (α) και τα τα κανονικοποιημένα στο γράφημα (β).

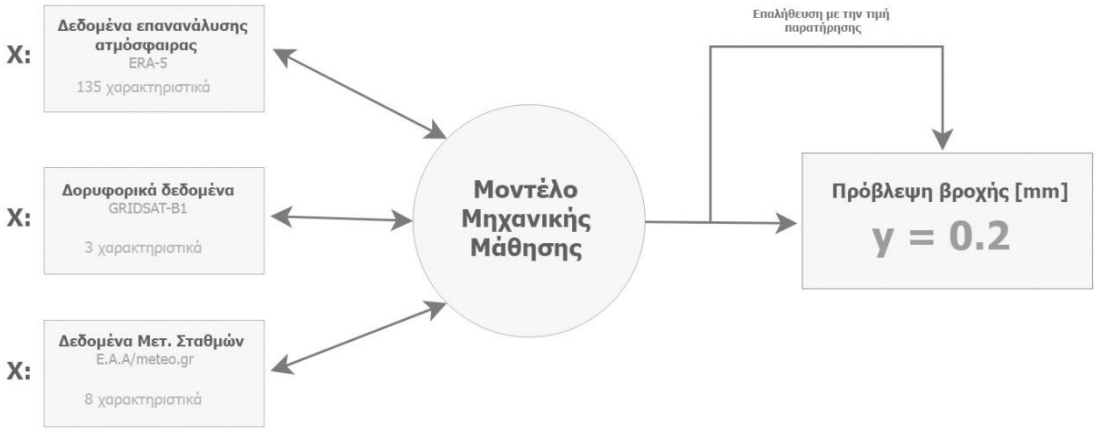
4.2 Μοντέλα Μηχανικής Μάθησης και Μέθοδοι

Τα μοντέλα μηχανικής μάθησης που εφαρμόστηκαν σε αυτήν τη μελέτη είναι το Τυχαίο Δάσος (Random Forest), το XGBoost, η Πολλαπλή Γραμμική Παλινδρόμηση (MLR), τα Νευρωνικά Δίκτυα Μακράς Βραχύχρονης Μνήμης (LSTM) καθώς και τα Συνελικτικά Νευρωνικά Δίκτυα (CNN). Το Random forest είναι ένας εποπτικο αλγόριθμος συνόλου μηχανικής μάθησης που βασίζεται σε προγνωστικά δέντρων αποφάσεων [14]. Εφαρμόζεται τόσο για εφαρμογές ταξινόμησης όσο και για εφαρμογές παλινδρόμησης. Το XGBoost, είναι ένα επεκτάσιμο σύστημα μηχανικής εκμάθησης για ενίσχυση δέντρων αποφάσεων. Είναι μια συγκεκριμένη εφαρμογή της μεθόδου Gradient Boosting που χρησιμοποιεί πιο ακριβείς προσεγγίσεις για να βρει το καλύτερο μοντέλο δέντρου [34,35]. Τα Νευρωνικά Δίκτυα Μακράς Βραχύχρονης Μνήμης (LSTM) είναι ένα τεχνητό νευρωνικό δίκτυο που χρησιμοποιείται στους τομείς της τεχνητής νοημοσύνης και της βαθιάς μάθησης. Ένα LSTM είναι μια παραλλαγή της προσέγγισης μοντελοποίησης που βασίζεται σε νευρωνικά δίκτυα, μια αναβαθμισμένη έκδοση RNN ικανή να μάθει την μακροπρόθεσμη εξάρτηση που υπάρχει σε διάφορα στάδια στα δεδομένα διαδοχικών χρονοσειρών [36].

Κάθε μοντέλο χρησιμοποιεί όλα τα δεδομένα από τις τρείς πηγές (δεδομένα επανανάλυσης ERA-5, δορυφορικά δεδομένα και δεδομένα επίγειων μετεωρολογικών σταθμών) χωρισμένα σε ένα σύνολο εκπαίδευσης (Training set), ένα σύνολο επικύρωσης (Validation set) και ένα σύνολο δοκιμών (Test set). Ο στόχος των μοντέλων εκπαίδευσης ήταν να προβλέψουν τη στιγμιαία βροχόπτωση για τις επόμενες 3 ώρες σε κάθε σταθμό.

4.2.1 Τεχνικές και Μέθοδοι

Σε αυτή την ενότητα θα παρουσιαστούν πληροφορίες και ρυθμίσεις των υπερ-παραμέτρων καθώς και γενικότερα η αρχιτεκτονική για τα μοντέλα μηχανικής μάθησης που εφαρμόστηκαν. Στην Εικόνα 26 παρουσιάζεται μια γενική αρχιτεκτονική και δομή η οποία εφαρμόστηκε σε όλα τα μοντέλα μηχανικής μάθησης.



Εικόνα 26. Γενική αρχιτεκτονική για την εκπαίδευση των μοντέλων.

Για το μοντέλο του του Τυχαίου Δάσους (Random Forest) χρησιμοποιήθηκε η βιβλιοθήκη sklearn στην python. Χρησιμοποιήσαμε την βιβλιοθήκη παλινδρόμησης (Regressor) του Random Forest, και για την εύρεση των καλύτερων δυνατών παραμέτρων του, χρησιμοποιήθηκε η μέθοδος GridSearchCV, η οποία λειτουργεί εποπτικά ψάχνοντας κάθε φορά τον καλύτερο δυνατό συνδυασμό των παραμέτρων. Για την εύρεση των καλύτερων δυνατών παραμέτρων για κάθε σημείο, το GridSearchCV ρυθμίστηκε έτσι ώστε να εφαρμοστεί 30 φορές επαναληπτικά ψάχνοντας τους καλύτερους δυνατούς συνδυασμούς για τις παρακάτω παραμέτρους: max_depth, max_features, min_samples_leaf, min_samples_split, n_estimators με τιμή cross validation 5. Οι τιμές που πρόσφεραν τα καλύτερα δυνατά αποτελέσματα ήταν οι παρακάτω:

- max_depth:10
- max_features:150
- min_samples_leaf:2
- min_samples_split:5
- n_estimators:80

Για το μοντέλο της Πολλαπλής Γραμμικής Παλινδρόμησης (MLR) ελάχιστες ρυθμίσεις χρειάστηκαν. Για την εύρεση την καλύτερης ευθείας χρησιμοποιήσαμε την βιβλιοθήκη και της sklearn στην python.

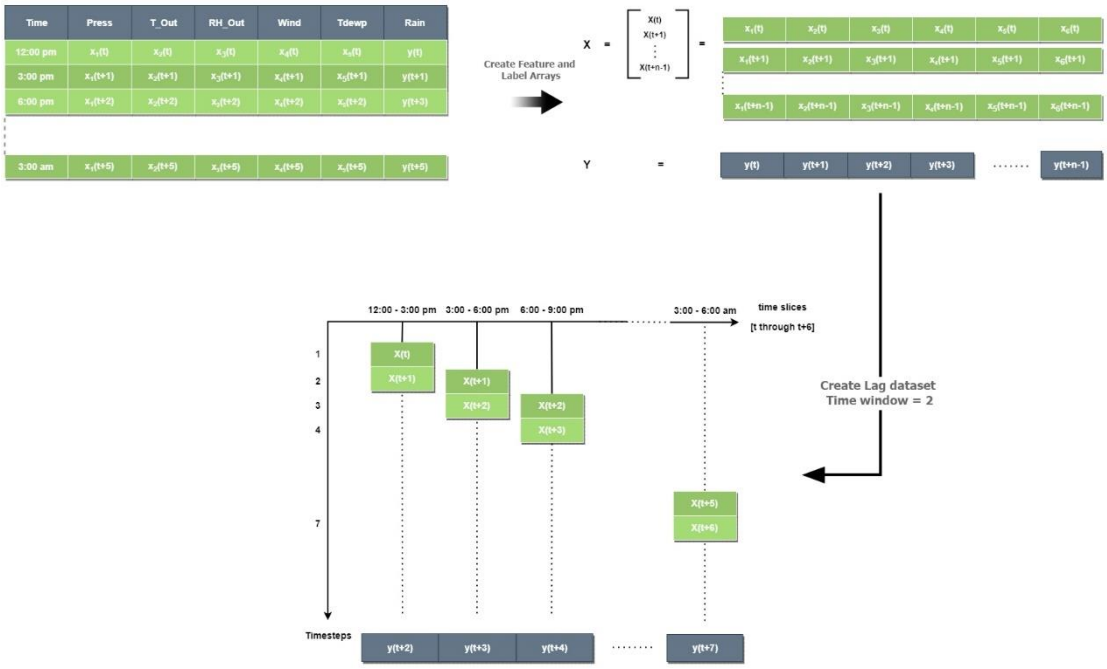
Για το μοντέλο του XGBoost χρησιμοποιήθηκε η ίδια τεχνική με αυτή του Random Forest έχοντας όμως διαφορετικές παραμέτρους για να βρεθεί ο καλύτερος δυνατός συνδυασμός. Επομένως, και μετά από 10 επαναλήψεις του GridSearchCV οι τιμές που πρόσφεραν τα καλύτερα δυνατά αποτελέσματα στο μοντέλο του XGBoost ήταν οι παρακάτω:

- nthread: 4
- max_depth: 2
- colsample_bytree: 0.8

- min_child_weight: 5
- learning_rate: 0.06
- gamma: 0.5
- subsample: 1
- n_estimators: 60

Για το μοντέλο των Νευρωνικών Δικτύων Μακράς Βραχύχρονης Μνήμης (LSTM) χρησιμοποιήθηκε το πακέτο keras της βιβλιοθήκης tensorflow στην python. Το πιο σημαντικό κομμάτι για την είσοδο δεδομένων στο μοντέλο LSTM ήταν η μετατροπή των δεδομένων σε χρονοσειρές με χρονικό παράθυρο. Χρειάστηκε να μοντελοποιήσουμε τα δεδομένα με τρόπο που το δίκτυο μπορεί να μάθει από μια ακολουθία προηγούμενων τιμών.

Πιο συγκεκριμένα, το LSTM αναμένει τα δεδομένα εισόδου σε μια συγκεκριμένη μορφή τρισδιάστατου tensor μεγέθους δείγματος δοκιμής (Test set) ανά χρονικά βήματα με βάση τον αριθμό των χαρακτηριστικών εισόδου. Ως μια εποπτευόμενη προσέγγιση μάθησης, το LSTM απαιτεί χαρακτηριστικά και ετικέτες για να μπορεί να εκπαιδευτεί. Στο πλαίσιο της πρόβλεψης χρονοσειρών, είναι σημαντικό να παρέχονται οι προηγούμενες τιμές ως χαρακτηριστικά και οι μελλοντικές τιμές ως ετικέτες (στόχοι), έτσι ώστε το LSTM να μπορεί να εκπαιδευτεί να προβλέπει το μέλλον. Έτσι, μεταφέρουμε τα δεδομένα χρονοσειρών σε μια δισδιάστατη διάταξη χαρακτηριστικών *X*, όπου τα δεδομένα εισόδου αποτελούνται από επικαλυπτόμενες τιμές καθυστέρησης στον επιθυμητό αριθμό χρονικών βημάτων σε παρτίδες (time window). Δημιουργούμε έναν πίνακα μιας διάστασης *y* που αποτελείται μόνο από τις ετικέτες (μεταβλητής στόχος) ή τις μελλοντικές τιμές που προσπαθούμε να προβλέψουμε για κάθε παρτίδα χαρακτηριστικών εισόδου. Τα δεδομένα εισόδου θα πρέπει επίσης να περιλαμβάνουν τιμές με χρονική καθυστέρηση του *y*, ώστε το δίκτυο να μπορεί επίσης να μάθει από προηγούμενες τιμές της μεταβλητής στόχου. Η διαδικασία περιγράφεται πολύ αναλυτικά στην Εικόνα 27.



Εικόνα 27. Διαδικασία μοντελοποίησης των δεδομένων με χρονική καθυστέρηση (lag dataset) για είσοδο στο νευρωνικό δίκτυο του LSTM.

Η αρχιτεκτονική του LSTM ήταν μια πάρα πολύ απλή αρχιτεκτονική, αφού μετά από πολλές δοκιμές με διαφορετικές τιμές στις υπερ-παραμέτρους καταλήξαμε στις παρακάτω (Πίνακας 7). Χρησιμοποιήθηκε ένα ακολουθιακό δίκτυο (Sequential) αποτελούμενο από ένα απλό επίπεδο LSTM(32), ακολουθούμενο από ένα πυκνό επίπεδο Dense(32) με συνάρτηση ενεργοποίησης (Activation) την ReLu, και ένα ακόμη πυκνό επίπεδο Dense(1) όπου καταλήγει στην προβλεπόμενη τιμή. Για τον ρυθμό εκμάθησης (Learning rate) χρησιμοποιήθηκε η συνάρτηση ReduceLROnPlateau. Ως βελτιστοποιητής του ρυθμού εκμάθησης μετά από δοκιμές ανάμεσα σε RMSprop και Adam καταλήξαμε ως καταλληλότερο και αποδοτικότερο στον δεύτερο. Τέλος, η εκπαίδευση έγινε σε 40 εποχές (epochs) με μέγεθος παρτίδας (batch size) ίσο με 32.

Πίνακας 7. Αρχιτεκτονική και ρυθμίσεις των υπερ-παραμέτρων του μοντέλου LSTM.

LSTM	
LSTM	32
Dense	32
Dense	1
Learning rate	1e-4
Optimizer	Adam

Epochs	40
Batch size	32

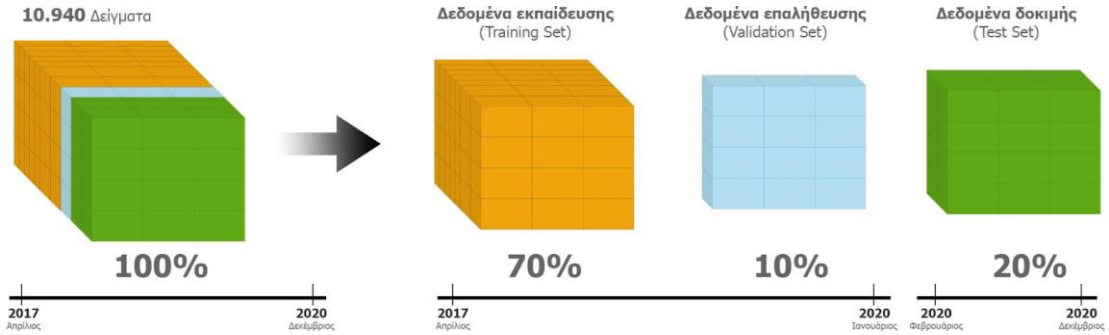
Τέλος, για το μοντέλο των Συνελικτικών Νευρωνικών Δικτύων (CNN) χρησιμοποιήθηκε το πακέτο keras της βιβλιοθήκης tensorflow στην python. Η αρχιτεκτονική του CNN ήταν μια πάρα πολύ απλή αρχιτεκτονική όπως και στο LSTM, αφού μετά από πολλές δοκιμές με διαφορετικές τιμές στις υπερ-παραμέτρους καταλήξαμε στις παρακάτω (Πίνακας 8). Χρησιμοποιήθηκε ένα ακολουθιακό δίκτυο (Sequential) αποτελούμενο από ένα συνελικτικό επίπεδο μιας διάστασης Conv1d(64), ακολουθούμενο από ακόμη 3 συνελικτικά επίπεδα μιας διάστασης και ένα απλό πυκνό επίπεδο Dense(8) με συνάρτηση ενεργοποίησης την ReLu, κλείνοντας με ένα πυκνό επίπεδο Dense(1) όπου καταλήγει στην προβλεπόμενη τιμή μετά από εφαρμογή της ισοπέδωσης στον πίνακα τιμών (Flatten). Για τον ρυθμό εκμάθησης (Learning rate) χρησιμοποιήθηκε ο βελτιστοποιητής Adam με τιμή 1*e*-4. Τέλος, η εκπαίδευση έγινε σε 30 εποχές (epochs) με μέγεθος παρτίδας (batch size) ίσο με 32.

Πίνακας 8. Αρχιτεκτονική και ρυθμίσεις των υπερ-παραμέτρων του μοντέλου CNN.

CNN	
Conv1d	units=64, strides=1, kernel_size=3, padding="causal", activation="relu"
Conv1d	units=32, strides=1, kernel_size=3, padding="causal", activation="relu"
Conv1d	units=16, strides=1, kernel_size=3, padding="causal", activation="relu"
Conv1d	units=8, strides=1, kernel_size=3, padding="causal", activation="relu"
Dense	8
Flatten	-
Learning rate	1e-4
Optimizer	Adam
Epochs	40
Batch size	32

4.2.2 Δεδομένα εκπαίδευσης και δεδομένα δοκιμής (Training and Test sets)

Για την ανάπτυξη και την αξιολόγηση των μοντέλων, τα δεδομένα χωρίστηκαν σε τρία μέρη: 70% για το σύνολο εκπαίδευσης (Training set), 10% για το σύνολο επικύρωσης (Validation set) και 20% για το σύνολο δοκιμής (Test set). Επιλέξαμε να χρησιμοποιήσουμε τη μέθοδο διαχωρισμού σε αναλογία 70-10-20, καθώς έχει αποδειχθεί πως η κακή ισορροπία στην αναλογία στο σετ εκπαίδευσης έχει συνήθως αρνητική επίδραση στην εκτιμώμενη απόδοση του μοντέλου, υποδηλώνοντας ότι είναι καλύτερο να υπάρχει μια καλή ισορροπία μεταξύ των μεγεθών του σετ εκπαίδευσης και σετ επικύρωσης για να υπάρχει αξιόπιστη εκτίμηση της απόδοσης του μοντέλου [37]. Έτσι, το σύνολο εκπαίδευσης (Training set) είναι το τμήμα των δεδομένων που χρησιμοποιείται για την εκπαίδευση του μοντέλου. Το μοντέλο θα πρέπει να παρατηρεί και να μαθαίνει από το σετ εκπαίδευσης, βελτιστοποιώντας οποιαδήποτε από τις παραμέτρους του. Το σετ επικύρωσης (Validation set) χρησιμοποιείται για την αποφυγή υπερβολικής προσαρμογής των μοντέλων (Overfitting). Στη συνέχεια, αφού καθοριστούν οι παράμετροι των μοντέλων, το σύνολο δοκιμών (Test set) χρησιμοποιήθηκε για την τελική αξιολόγηση του μοντέλου [38]. Τα δεδομένα ημερολογιακά χωρίστηκαν από τον Απρίλιο του 2017 έως τον Ιανουάριο του 2020 για τα σύνολα δεδομένων εκπαίδευσης και επικύρωσης (Training & Validation set) και τα υπόλοιπα δείγματα (Φεβρουάριος 2020 έως Δεκέμβριος 2020) χρησιμοποιήθηκαν για το σετ δοκιμών (Test set).



Εικόνα 28. Ο διαχωρισμός των δεδομένων σε αναλογία 70-10-20, των σετ δεδομένων εκπαίδευσης, επαλήθευσης και δοκιμής αντίστοιχα.

4.2.3 Μετρικές αξιολόγησης

Αυτή η ενότητα ορίζει τις μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση των αποτελεσμάτων των αλγορίθμων και των μοντέλων που εφαρμόστηκαν. Η απόδοση κάθε μοντέλου αξιολογήθηκε και συγκρίθηκε χρησιμοποιώντας τρεις διαφορετικές μετρικές: Μέσο Απόλυτο Σφάλμα (MAE) (Εξίσωση 9), Μέσο Τετραγωνικό Σφάλμα (MSE) (Εξίσωση 10), Μέσο Τετραγωνικό Σφάλμα ρίζας (RMSE) (Εξίσωση 11).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (9)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (10)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (11)$$

Όπου *i* είναι ο αύξων αριθμός του δείγματος, y_i η παρατηρούμενη τιμή, y_i η προβλεπόμενη τιμή και *n* είναι ο συνολικός αριθμός των δειγμάτων [39].

5

Αποτελέσματα

Σε αυτό το κεφάλαιο παρουσιάζουμε τα αποτελέσματα των μοντέλων μηχανικής μάθησης που εφαρμόσαμε και συγκρίνουμε την απόδοσή τους με βάση τις 3 μετρικές που αναφέραμε παραπάνω. Στον Πίνακα 9 παρουσιάζονται οι μετρικές και από τα 5 μοντέλα μηχανικής μάθησης για 5 σημεία που εφαρμόσαμε. Ο υπολογισμός των μετρικών στα αποτελέσματα αναφέρονται σε γεγονότα όπου η βροχή ήταν μεγαλύτερη από 0.2 χιλιοστά απορρίπτοντας έτσι στους υπολογισμούς τις περιπτώσεις όπου δεν υπήρξε καταγραφή βροχής.

Πίνακας 9. Μετρικές για την πρόβλεψη της στιγμιαίας βροχόπτωσης (> 0.2 mm) για τις επόμενες 3 ώρες από τα 5 εκπαιδευμένα μοντέλα, που αξιολογήθηκαν στο σύνολο δεδομένων δοκιμής (test set) για τα 5 σημεία.

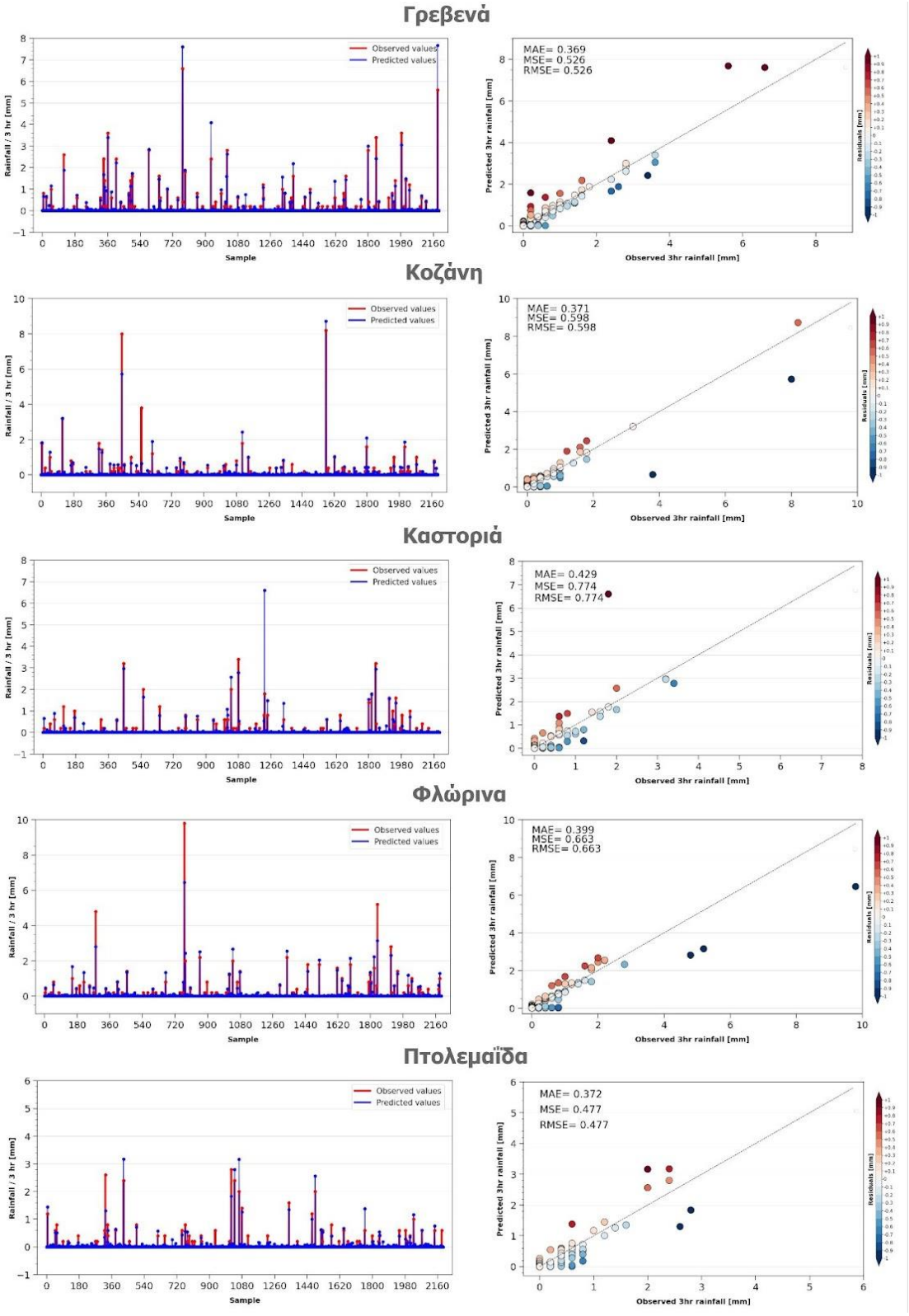
Μοντέλο	**MAE [mm]**	**MSE [mm]**	**RMSE [mm]**
Random Forest (RF)	0.3	0.4	0.6
XGBoost	0.3	0.5	0.7
Multiple Linear Regression (MLR)	0.4	0.6	0.7
CNN	0.6	1.0	1.1
LSTM	0.7	1.2	1.3

Τα αποτελέσματα των μοντέλων Μηχανικής Μάθησης επικυρώθηκαν πραγματοποιώντας την τεχνική επικύρωσης Cross Validation και ανακατασκευάζοντας τα μοντέλα (Fine tuning) μέχρι να καταλήξουμε πάντα στους καλύτερους δυνατούς συνδυασμούς. Προχωρώντας στα αποτελέσματα όπως παρουσιάζονται στον Πίνακα 9, μπορούμε να παρατηρήσουμε εύκολα πως το μοντέλο μηχανικής μάθησης βασιζόμενο στα δέντρα

αποφάσεων Random forest, απέδωσε καλύτερα από υπόλοιπα μοντέλα. Δεύτερο στην κατάταξη ήταν το XGBoost Regressor και τελευταία τα νευρωνικά δίκτυα CNN και LSTM, ενώ στην μέση της κατάταξη είναι η κλασσική μέθοδος της γραμμικής παλινδρόμησης με πολλές μεταβλητές (MLR). Αυτή η κατάταξη μπορεί να εξηγηθεί από την διερευνητική ανάλυση των δεδομένων (EDA) όπως αναλύθηκε παραπάνω. Το πρόβλημά μας φαίνεται να εφαρμόζει καλύτερα σε αλγορίθμους παλινδρόμησης (Regression) από ό,τι στα κλασσικά νευρωνικά δίκτυα, έτσι οι αλγόριθμοι που βασίζονται σε μεθόδους παλινδρόμησης καθώς και σε δέντρα αποφάσεων (MLR, Random forest και XGBoost) προσαρμόστηκαν καλύτερα στο πρόβλημα, δίνοντας έτσι τα καλύτερα αποτελέσματα [19,40].

Επιπλέον, το Random forest απέδωσε καλύτερα από όλα τα υπόλοιπα μοντέλα και σε υψηλές τιμές βροχόπτωσης, φαίνεται δηλαδή να είναι το μοντέλο που είχε συνολικά την μεγαλύτερη σταθερότητα και απόδοση σε όλη την κατανομή των δεδομένων. Τα Νευρωνικά δίκτυα CNN και LSTM παρά τις πολλές παραλλαγές και δοκιμές που δέχτηκαν στις υπερ-παραμέτρους τους, παρουσίασαν κακή προγνωστική ακρίβεια συνολικά στα δεδομένα, έχοντας μόνο μια μικρή εξαίρεση στην κατά τα αλλά αρκετά καλή και σταθερή προγνωστική δυνατότητα στις χαμηλές τιμές βροχής. Τέλος, το XGBoost είχε παρόμοια απόδοση με το μοντέλο Random forest μιας και πρόκειται για παρόμοιας αρχιτεκτονικής μοντέλο, αλλά ήταν χειρότερο σε υψηλές και ακραίες τιμές βροχοπτώσεων.

Στην Εικόνα 29 παρουσιάζονται τα γραφήματα σύγκρισης πραγματικών και προβλεπόμενων τιμών της στιγμιαίας βροχόπτωσης 3 ωρών και στα 5 σημεία από τα αποτελέσματα του μοντέλου Random forest.



Εικόνα 29. Γραφήματα σύγκρισης πραγματικών και προβλεπόμενων τιμών στιγμιαίας βροχόπτωσης 3 ωρών και στα 5 σημεία από τα αποτελέσματα του μοντέλου Random forest.

6

Τεχνικές λεπτομέρειες

Στο κεφάλαιο αυτό αναφέρονται και αναλύονται περισσότερες πληροφορίες και λεπτομέρειες σε ότι έχει να κάνει με θέματα της διπλωματικής που έχουν τεχνικό ενδιαφέρον. Επομένως, στις παρακάτω ενότητες θα αναφέρουμε σε προγραμματιστικά εργαλεία που χρησιμοποιήθηκαν από την αρχή της εργασίας μέχρι και την εκτέλεση της και την εξαγωγή των τελικών αποτελεσμάτων.

6.1 Οργάνωση και επιλογή των δεδομένων

Για την οργάνωση και επιλογή των δεδομένων συμβουλευτήκαμε τους ειδικούς επιστήμονες του Εθνικού Αστεροσκοπείου Αθηνών/ΜΕΤΕΟ σε θέματα μετεωρολογίας Δρ. Λαγουβάρδο Κωνσταντίνο, Δρ. Ντάφη Σταύρο και την Δρ. Κοτρώνη Βασιλική. Στόχος μας ήταν η επιλογή και η οργάνωση των δεδομένων με τον καλύτερο δυνατό τρόπο. Έτσι, κρίθηκε απαραίτητο να επιλέξουμε 3 πηγές δεδομένων που αυτά ήταν τα δεδομένα επανανάλυσης της ατμόσφαιρας ERA5, τα δορυφορικά δεδομένα GRIDSAT-B1, καθώς και δεδομένα από τους επίγειους μετεωρολογικούς σταθμούς του Δικτύου Αυτόματων Μετεωρολογικών Σταθμών του Εθνικού Αστεροσκοπείου Αθηνών/ΜΕΤΕΟ.

6.1.1 Λήψη και διαχείριση των δεδομένων

Στην συνέχεια της διαδικασίας της οργάνωσης και επιλογής των πηγών των δεδομένων προχωρήσαμε στον σχεδιασμό και στην ανάπτυξη ρουτινών για την λήψη και διαχείριση των δεδομένων από τις 3 πηγές που αναφέρθηκαν παραπάνω. Για την λήψη και την διαχείριση των δεδομένων χρησιμοποιήθηκαν οι γλώσσες προγραμματισμού python καθώς και scripts σε περιβάλλον linux.

Πίνακας 10. Πηγές δεδομένων και πληροφορίες για την λήψη των δεδομένων.

Πηγή δεδομένων	**Ιστότοπος**	**Μέγεθος δεδομένων**
Δεδομένα επανανάλυσης της ατμόσφαιρας - ERA5	https://cds.climate.copernicus.eu/cdsapp	~4.5 gb
Δορυφορικά δεδομένα - GRIDSAT-B1	https://www.ncei.noaa.gov/data/geostationary-ir-channel-brightness-temperature-gridsat-b1/access/	~1.2 gb
Δεδομένα Μετ. Σταθμών - E.A.A/METEO	https://meteosearch.meteo.gr/	~420 mb

Ο συνολικός όγκος των δεδομένων ανήλθε στα ~6 gb συνολικά και από τις 3 πηγές, και αποθηκεύτηκαν σε εξωτερικό δίσκο τύπου SSD για την ταχύτερη διαχείριση και επεξεργασία αυτών. Για την λήψη των δεδομένων χρησιμοποιήθηκαν 2 σταθεροί υπολογιστές με λειτουργικά συστήματα Windows και Linux αντίστοιχα.

Πίνακας 11. Δείγμα κώδικα σε python για την λήψη των δορυφορικών δεδομένων.

Λήψη των δορυφορικών δεδομένων

```
# Import libraries
import numpy as np
import pandas as pd
import wget
import urllib.request
from threading import Thread
import os

# Years
years=[2017,2018,2019,2020]

# Read filenames urls from .txt file
df=pd.read_csv(str(year)+".txt",delimiter=',',index_col= False)

# Build url
url="https://www.ncei.noaa.gov/data/geostationary-ir-channel-
brightness-temperature-gridsat-b1/access/"+str(year)+"/"

# Download file and save on folder with name as year
for d in df["file"]:
    os.system('wget -O'+(save_dir+d)+" "+(url+d))
    print("File "+d+" has Downloaded!\n")
```

Πίνακας 12. Δείγμα κώδικα σε python για την λήψη των δεδομένων ERA-5.

Λήψη των δεδομένων ERA-5

```
# Import libraries
import cdsapi

# Calling of cdsapi
c = cdsapi.Client()

# Init settings to retrieve parameters and data
c.retrieve(
    'reanalysis-era5-single-levels',
    {
        'product_type': 'reanalysis',
        'format': 'grib',
        'variable': [
            '10m_u_component_of_wind',
'10m_v_component_of_wind', '2m_dewpoint_temperature',
            '2m_temperature',       'boundary_layer_dissipation',
'convective_available_potential_energy',
            'convective_inhibition', 'mean_sea_level_pressure',
            'snow_albedo',        'total_column_cloud_ice_water',
'total_column_cloud_liquid_water',
            'total_column_snow_water',
'total_column_supercooled_liquid_water',
'total_column_water_vapour',
            'volumetric_soil_water_layer_1',
'zero_degree_level',
        ],
    'year': [
            '2016'
        ],
        'month': ['01', '02', '03',
            '04', '05', '06',
            '07', '08', '09',
            '10', '11', '12'],
        'day': [
            '01', '02', '03',
            '04', '05', '06',
            '07', '08', '09',
            '10', '11', '12',
            '13', '14', '15',
            '16', '17', '18',
            '19', '20', '21',
            '22', '23', '24',
            '25', '26', '27',
            '28', '29', '30',
            '31',
        ],
        'time': [
            '00:00', '03:00', '06:00',
            '09:00', '12:00', '15:00',
            '18:00', '21:00',
        ],
        'area': [
```

```
            42, 19, 38,
            24,
        ],
    },
    'era5_data.grib')
```

Για την λήψη των δεδομένων από τους επίγειους μετεωρολογικούς σταθμούς του Δικτύου Αυτόματων Μετεωρολογικών Σταθμών του Εθνικού Αστεροσκοπείου Αθηνών/ΜΕΤΕΟ χρησιμοποιήθηκε διαδικασία στο server του Εθνικού Αστεροσκοπείου Αθηνών/ΜΕΤΕΟ.

Για την λήψη των δεδομένων ERA-5 χρησιμοποιήθηκε το api που παρέχει το Ευρωπαϊκό Πρόγραμμα Copernicus για την κατάσταση του κλίματος cdsapi [41]. Για την χρήση του χρειάζεται κάποιος να κάνει εγγραφή ως χρήστης στην βάση δεδομένων. Για την λήψη των δορυφορικών δεδομένων δεν απαιτήθηκε εγγραφή παρά μόνο αναφορά στα δεδομένα. Τέλος, για την λήψη και την χρήση των δεδομένων από τους επίγειους μετεωρολογικούς σταθμούς του Δικτύου Αυτόματων Μετεωρολογικών Σταθμών του Εθνικού Αστεροσκοπείου Αθηνών/ΜΕΤΕΟ, δρομολογήσαμε επικοινωνία και λάβαμε ειδική άδεια από τους υπεύθυνους Δρ Λαγουβάρδο Κωνσταντίνο & Δρ. Κοτρώνη Βασιλική.

Η διαδικασία τη λήψης των δεδομένων διήρκησε περίπου 2 εβδομάδες λόγω του μεγάλου όγκου των δεδομένων καθώς και λόγω των επιπλοκών που εμφανίστηκαν στο server των δορυφορικών δεδομένων GRIDSAT-B1.

Πίνακας 13. Μορφή των δεδομένων καθώς και ο τύπος των αρχείων.

Πηγή δεδομένων	**Μορφή δεδομένων**	**Τύπος αρχείου**
Δεδομένα επανανάλυσης της ατμόσφαιρας - ERA5	Δεδομένα σε πλέγμα (Gridded)	.grib
Δορυφορικά δεδομένα - GRIDSAT-B1	Δεδομένα σε πλέγμα (Gridded)	.netcdf
Δεδομένα Μετ. Σταθμών - Ε.Α.Α/ΜΕΤΕΟ	Δεδομένα σε μορφή κειμένου	.txt

Στη συνέχεια της λήψης των αρχείων των δεδομένων προχωρήσαμε σε ενέργειες διαχείρισης και επεξεργασίας των δεδομένων με σκοπό την μετατροπή τους σε μια κοινά αποδεκτή μορφή για την ανάγνωσή τους από την γλώσσα προγραμματισμού της python. Για τις ανάγκες αυτών των ενεργειών χρησιμοποιήθηκαν οι γλώσσες προγραμματισμού της python καθώς και βιβλιοθήκες σε Linux, ιδανικές για επεξεργασία γεωχωρικών δεδομένων. Μερικές

από αυτές ήταν οι: GDAL, cdo και ncl. Κύριο μέλημα ήταν να ενωθούν τα διαφορετικά αρχεία των δεδομένων ERA-5 ανά έτος σε ένα μοναδικό καθώς και των δορυφορικών δεδομένων.

Εικόνα 30. Βιβλιοθήκες που χρησιμοποιήθηκαν για την διαχείριση και την επεξεργασία των γεωχωρικών δεδομένων.

6.2 Ανάγνωση και επεξεργασία των δεδομένων

Για την ανάγνωση και επεξεργασία των δεδομένων χρησιμοποιήθηκε εξ' ολοκλήρου η γλώσσα προγραμματισμού python σε περιβάλλον του Visual Studio Code. Επιπλέον, χρησιμοποιήθηκαν σε μεγάλο βαθμό τεχνολογίες νέφους (Cloud) και πιο συγκεκριμένα της Google και την πλατφόρμα του Google colab, για τον διαμοιρασμό και την κοινοποίηση των εργασιών μεταξύ των δύο συγγραφέων.

Για την επεξεργασία των δεδομένων κρίθηκε αναγκαίο η εγκατάσταση της πλατφόρμας Anaconda. Ευρέως γνωστή πλατφόρμα που χρησιμοποιείται για προβλήματα ανάλυσης δεδομένων. Για την εγκατάσταση της πλατφόρμας σε Windows και Linux ακολουθήθηκαν πιστά οι οδηγίες που παρέχονται στον επίσημο ιστότοπο της πλατφόρμας: https://docs.anaconda.com/free/anaconda/install/windows/.

Πίνακας 14. Προγραμματιστικά εργαλεία και βιβλιοθήκες που χρησιμοποιήθηκαν κατά την εκτέλεση της εργασίας.

Ενέργεια	**Προγραμματιστικά εργαλεία**	**Βιβλιοθήκες**
Ανάγνωση δεδομένων	python, gdal, cdo	os, glob, json, pandas, numpy, xarray
Δειγματοληψία σημειακών δεδομένων	python	pandas, numpy, xarray
Επεξεργασία δεδομένων	python, gdal, cdo	pandas, numpy, xarray, scipy.stats, sklearn
Οπτικοποίηση δεδομένων	python, Qgis	matplotlib, cartopy

Στους παρακάτω πίνακες παρουσιάζονται δείγματα κώδικα από τις ρουτίνες που χρησιμοποιήθηκαν στην γλώσσα προγραμματισμού python για την επεξεργασία των δεδομένων μέχρι την τελική τους μορφή για είσοδο στα μοντέλα μηχανικής μάθησης.

Πίνακας 15. Δείγμα κώδικα σε python για την ανάγνωση και δειγματοληψία των δεδομένων στα 5 σημεία.

Ανάγνωση και δειγματοληψία των δεδομένων στα 5 σημεία

```
# Import libraries
import warnings
import metpy
import numpy as np
import xarray as xr
import pandas as pd
import datetime
import netCDF4 as nc
import glob

# Settings
mode_stations=True
mode_sats=True
mode_era5=True
vars_pl=["z","pv","q","t","u","v"]

# Paths
path_sat = "GRIDSAT-B1/data_m/data_yearly/"
stations=pd.read_csv("paper/stations/stations_info.txt")

# City names
cities=["grevena","kozani","kastoria","florina","ptolemaida"]

# Data columns to get
columns_to_get=["date","temp_out","out_hum","dew_pt","bar","rain
","rain_rate","wind_speed","wind_dir"]
single_l_columns=['time',  'bld',  'cape',  'cin',  'msl',  'asn',
'tciw', 'tclw', 'tcsw', 'tcslw', 'tcwv', 'swvl1', 'deg0l']

# Dictionary settings for aggregation methods
d                                                                =
{"date":"median","temp_out":"median","out_hum":"median","dew_pt"
:"median","bar":"median"\
     ,"rain":"sum","rain_rate":"median","wind_speed":"median","w
ind_dir":"median"}

# Timestamp data
timestamp_all=[
 datetime.time(23, 50),
 datetime.time(0, 0),
 datetime.time(0, 10),
 datetime.time(2, 50),
 datetime.time(3, 0),
 datetime.time(3, 10),
```

```
 datetime.time(5, 50),
 datetime.time(6, 0),
 datetime.time(6, 10),
 datetime.time(8, 50),
 datetime.time(9, 0),
 datetime.time(9, 10),
 datetime.time(11, 50),
 datetime.time(12, 0),
 datetime.time(12, 10),
 datetime.time(14, 50),
 datetime.time(15, 0),
 datetime.time(15, 10),
 datetime.time(17, 50),
 datetime.time(18, 0),
 datetime.time(18, 10),
 datetime.time(20, 50),
 datetime.time(21, 0),
 datetime.time(21, 10)]

timestamp_utc=[
 datetime.time(0, 0),
 datetime.time(3, 0),
 datetime.time(6, 0),
 datetime.time(9, 0),
 datetime.time(12, 0),
 datetime.time(15, 0),
 datetime.time(18, 0),
 datetime.time(21, 0)]

# Routine for readind and parsing data for 5 points (cities)
cnt=0
for city in cities:

    # Read station data and resample from 10min to 3hours.
    data_station   =   pd.read_csv("paper/stations/"+city+".txt",
delimiter="\t")
    data_station["date"]         =         data_station["date"]+"
"+data_station["time"]
    if city=="ptolemaida":
        data_station["date"]                                    =
pd.to_datetime(data_station["date"], format='%d-%m-%Y %H:%M')
    data_station["date"]  =  pd.to_datetime(data_station["date"],
format='%Y-%m-%d %H:%M')
    data_station = data_station[columns_to_get]
    data_station = data_station[data_station["date"]>="2017-04-01
23:50"]
    data_station = data_station.reset_index(drop=True)
    data_station                                                =
data_station[data_station["date"].dt.time.isin(timestamp_all)].r
eset_index(drop=True)
    data_station   =   data_station.groupby(data_station.index   //
3).agg(d)
    data_station["wind_dir"]=data_station["wind_dir"].fillna(0)
    data_station = data_station[:-1]
    #data.to_csv("paper/stations/data_3hr/"+city+".txt",
index=False)
```

```
    # Read satellite data
    if mode_sats:
        ds_gridsat=xr.open_mfdataset(path_sat+'*.nc')
        gridsat_data=ds_gridsat.sel(lon=stations["lon"][cnt],
                                                    lat=stations[
"lat"][cnt],
method='nearest',drop=True).to_dataframe().reset_index()

        gridsat_data  =  gridsat_data[gridsat_data["time"]>="2017-
04-01 23:50"].reset_index(drop=True)

        gridsat_data["irwvp"]=gridsat_data["irwvp"]-273.15
        gridsat_data["irwin_cdr"]=gridsat_data["irwin_cdr"]-
273.15
        gridsat_data=gridsat_data.rename(columns={"time":"date"}
)

    # Read ERA-5 data
    if mode_era5:
        ds_era5_pl=xr.open_dataset("GRIDSAT-
B1/era5_p/all/all_p.nc")
        ds_era5_sl=xr.open_dataset("GRIDSAT-
B1/era5_p/all/sl.nc")
        ds_era5_sla=ds_era5_sl.drop_dims("depth")
        era_5_pressure_pl_data=ds_era5_pl.sel(lon=stations["lon"
][cnt],
                                                    lat=stati
ons["lat"][cnt],
method='nearest',drop=True).to_dataframe().reset_index().pivot_t
able(vars_pl, "time",'plev')
        era_5_pressure_sl_data=ds_era5_sla.sel(lon=stations["lon
"][cnt],
                                                    lat=stati
ons["lat"][cnt],
method='nearest',drop=True).to_dataframe().reset_index()
        era_5_swvl1=ds_era5_sl["swvl1"].sel(lon=stations["lon"][
cnt],
                                                    lat=stati
ons["lat"][cnt],
method='nearest',drop=True).to_dataframe().reset_index()

        era_5_pressure_sl_data=era_5_pressure_sl_data.merge(era_
5_swvl1,on='time')
        era_5_pressure_sl_data=era_5_pressure_sl_data[single_l_c
olumns]


        level_one                                               =
era_5_pressure_pl_data.columns.get_level_values(0).astype(str)
        level_two=(era_5_pressure_pl_data.columns.get_level_valu
es(1)/100).astype(int).astype(str)
        era_5_pressure_pl_data.columns= level_one + level_two

        era5_data=era_5_pressure_sl_data.merge(era_5_pressure_pl
_data, on='time')
        era5_data    =    era5_data[era5_data["time"]>="2017-04-01
23:50"].reset_index(drop=True)
```

```
        era5_data=era5_data.rename(columns={"time":"date"})
        all_data=data_station.merge(era5_data, on='date')
        all_data=all_data.merge(gridsat_data, on='date')
    cnt=+1

    # Save final cube data from 3 data sources for every point
(city)
    all_data.to_csv("paper/final_data/"+city+".txt",
index=False)
```

Πίνακας 16. Δείγμα κώδικα σε python για την επεξεργασία των δεδομένων.

Προεπεξεργασία των δεδομένων

```
# Import libraries
import warnings
import metpy
import numpy as np
import xarray as xr
import pandas as pd
import datetime
import netCDF4 as nc
import glob
from sklearn import preprocessing
from sklearn.preprocessing import MinMaxScaler, StandardScaler,
RobustScaler
from sklearn.model_selection import train_test_split
from scipy.stats import zscore

# Cities
cities=["grevena","kozani","kastoria","florina","ptolemaida"]

t_h=['t250','t300','t350','t400','t450','t500','t550','t600','t6
50','t700','t750','t775','t800','t825','t850','t875',
 't900','t925','t950','t975','t1000']

# Sample Routine for data preprocessing
for city in cities:
    data=pd.read_csv("paper/final_data/"+cities[4]+".txt")

    data.pop("cin")
    data.pop("asn")
    data["msl"]=data["msl"]/100
    data[t_h]=data[t_h]-273.15

    nan_data=data.isna().sum()

    final_data        =        data.interpolate(method='linear',
axis=0).ffill().bfill()
    final_data["deep_convection"]=final_data["irwin_cdr"]-
final_data["irwvp"]
.
.
.
```

```
.
.
```

6.3 *Ανάπτυξη και εφαρμογή των μοντέλων μηχανικής μάθησης*

Για την ανάπτυξη και την εφαρμογή των αλγορίθμων και των μοντέλων μηχανικής μάθησης χρησιμοποιήθηκε η γλώσσα προγραμματισμού python καθώς και οι βιβλιοθήκες της sklearn και tensorflow. Στους παρακάτω πίνακες παρουσιάζεται δείγμα κώδικα για τα 3 βασικά μοντέλα της εργασίας (Random Forest, XGBoost και LSTM).

Πίνακας 17. Δείγμα κώδικα σε python για την εκτέλεση των μοντέλων.

Εφαρμογή και εκτέλεση του μοντέλου Random Forest

```
# Import libraries
import warnings
import metpy
import numpy as np
import xarray as xr
import pandas as pd
import datetime
import netCDF4 as nc
import glob
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import GridSearchCV
from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error,
mean_absolute_error
from sklearn.pipeline import make_pipeline
from scipy.stats import zscore
import matplotlib.pyplot as plt
from matplotlib.lines import Line2D
import matplotlib as mpl

# Cities
cities=["grevena","kozani","kastoria","florina","ptolemaida"]

# Temperature levels
t_h=['t250','t300','t350','t400','t450','t500','t550','t600','t6
50','t700','t750','t775','t800','t825','t850','t875',
 't900','t925','t950','t975','t1000']

# Number of runs
run_N=30
```

```
runs=np.arange(0,run_N,1)
# Run procedure of grid-search for hyperparameters optimization
of Random Forest

for i in range(0,5):
    # Read data
    data=pd.read_csv("paper/final_data/"+cities[i]+".txt")

    data.pop("cin")
    data.pop("asn")
    data["msl"]=data["msl"]/100
    data[t_h]=data[t_h]-273.15

    nan_data=data.isna().sum()

    final_data        =        data.interpolate(method='linear',
axis=0).ffill().bfill()
    final_data["deep_convection"]=final_data["irwin_cdr"]-
final_data["irwvp"]

    final_data.pop("date")
    y=final_data["rain"].values
    final_data.pop("rain")
    x=final_data.values

    # Split data
    x_train, x_test, y_train, y_test = train_test_split(x, y,
test_size = 0.2, random_state=42)

    runs_n=[]
    bootstraps=[]
    max_depths=[]
    max_featuress=[]
    min_samples_leafs=[]
    min_samples_splits=[]
    n_estimatorss=[]

    MAES=[]
    MSES=[]
    RMSES=[]
    R2S=[]
    for run in runs:
        rf = RandomForestRegressor()

        param_grid = {
            'bootstrap': [True],
            'max_depth': [10],
            'max_features': [120,130, 140],
            'min_samples_leaf': [2],
            'min_samples_split': [5],
            'n_estimators': [50,80]
        }# Create a based model
        rf = RandomForestRegressor()# Instantiate the grid search
model
        grid_search = GridSearchCV(estimator = rf, param_grid =
param_grid,
```

```
                                    cv = 4, n_jobs = -1, verbose =
2)

        grid_search.fit(x_train, y_train)

        print ('Best Parameters: ', grid_search.best_params_, '
\n')
        best_grid = grid_search.best_estimator_

        # Run predictions and save results for each station
        y_prediction =  best_grid.predict(x_test)

        predictions_rf=pd.DataFrame(y_prediction)
        predictions_rf[predictions_rf<0]=0
        predictions_rf = predictions_rf.round(2)

        th=0.2
        score=r2_score(y_test[y_test>=th],y_prediction[y_test>=t
h])
        score_r2="R2 score is: "+str(score)
        MAE="MAE                    score                    is:
"+str(mean_absolute_error(y_test[y_test>=th],y_prediction[y_test
>=th]))
        MSE="MSE                    score                    is:
"+str(mean_squared_error(y_test[y_test>=th],y_prediction[y_test>
=th]))
        RMSE="RMSE                   score                    is:
"+str(np.sqrt(mean_squared_error(y_test[y_test>=th],y_prediction
[y_test>=th])))



        runs_n.append(run)
        R2S.append(score)
        MAES.append(mean_absolute_error(y_test[y_test>=th],y_pre
diction[y_test>=th]))
        RMSES.append(np.sqrt(mean_squared_error(y_test[y_test>=t
h],y_prediction[y_test>=th])))
        MSES.append(mean_squared_error(y_test[y_test>=th],y_pred
iction[y_test>=th]))

        bootstraps.append(grid_search.best_params_["bootstrap"])
        max_depths.append(grid_search.best_params_["max_depth"])
        max_featuress.append(grid_search.best_params_["max_featu
res"])
        min_samples_leafs.append(grid_search.best_params_["min_s
amples_leaf"])
        min_samples_splits.append(grid_search.best_params_["min_
samples_split"])
        n_estimatorss.append(grid_search.best_params_["n_estimat
ors"])

        print(score_r2)
        print(MAE)
        print(MSE)
        print(RMSE)
```

```
        data_s=pd.DataFrame({'Observed':                    y_test,
'Predicted':y_prediction}, columns=['Observed', 'Predicted'])
        data_s.to_csv("paper/results/RF/stats/data/"+cities[i]+"
/"+"run_"+str(run)+".txt")

    stats = pd.DataFrame({'run_number': runs_n, 'R2': R2S, 'MAE':
MAES, 'MSE': MSES, 'RMSE': RMSES,
                          'Bootstrap': bootstraps, 'max_depth':
max_depths,
                          'max_features':          max_featuress,
'min_samples_leaf': min_samples_leafs,
                          'min_samples_split':
min_samples_splits, 'n_estimators': n_estimatorss},
                           columns=['run_number',  'R2',  'MAE',
'MSE', 'RMSE', 'Bootstrap', 'max_depth', 'max_features',
                          'min_samples_leaf',
'min_samples_split', 'n_estimators'])

    stats.to_csv('paper/results/RF/stats/'+cities[i]+'.txt',
sep="\t")
```

Πίνακας 18. Δείγμα κώδικα σε python για την εκτέλεση των μοντέλου XGBoost.

Εφαρμογή και εκτέλεση του μοντέλου XGBoost

```
# Import Libraries
import warnings
import metpy
import numpy as np
import xarray as xr
import pandas as pd
import datetime
import netCDF4 as nc
import glob
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.metrics import r2_score
from       sklearn.metrics       import       mean_squared_error,
mean_absolute_error
from sklearn.pipeline import make_pipeline
from xgboost import XGBRegressor
from scipy.stats import zscore
import matplotlib.pyplot as plt
from matplotlib.lines import Line2D
import matplotlib as mpl

# Cities
cities=["grevena","kozani","kastoria","florina","ptolemaida"]

# Temperature levels
t_h=['t250','t300','t350','t400','t450','t500','t550','t600','t6
50','t700','t750','t775','t800','t825','t850','t875',
 't900','t925','t950','t975','t1000']
```

```
run_N=10
runs=np.arange(0,run_N,1)
plot_proc=False

for i in range(0,5):

    # Read data
    data=pd.read_csv("paper/final_data/"+cities[i]+".txt")

    data.pop("cin")
    data.pop("asn")
    data["msl"]=data["msl"]/100
    data[t_h]=data[t_h]-273.15

    nan_data=data.isna().sum()

    final_data        =        data.interpolate(method='linear',
axis=0).ffill().bfill()
    final_data["deep_convection"]=final_data["irwin_cdr"]-
final_data["irwvp"]

    final_data.pop("date")
    y=final_data["rain"].values
    final_data.pop("rain")
    x=final_data.values

    # Split dataset
    x_train, x_test, y_train, y_test = train_test_split(x, y,
test_size = 0.2, random_state=42)
    runs_n=[]
    gammas=[]
    max_depths=[]
    subsamples=[]
    min_child_weights=[]
    colsample_bytrees=[]
    n_estimatorss=[]

    MAES=[]
    MSES=[]
    RMSES=[]
    R2S=[]

    # XGBRegressor fit
    for run in runs:
        xg = XGBRegressor()

        param_grid = {
            'nthread':[4],
            'max_depth': [2],
            'colsample_bytree': [0.6, 0.8],
            'min_child_weight': [2,5],
            'learning_rate': [0.06],
            'gamma': [0.5],
            'subsample': [0.8,1],
            'n_estimators': [50]
        }# Create a based model
```

```
        grid_search = GridSearchCV(estimator = xg, param_grid =
param_grid,
                                   cv = 2, n_jobs = -1, verbose =
True)

        grid_search.fit(x_train, y_train)

        # print the best parameters
        print ('Best Parameters: ', grid_search.best_params_, '
\n')
        best_grid = grid_search.best_estimator_

        # Run predictions and save results for each station
        y_prediction =  best_grid.predict(x_test)

        predictions_rf=pd.DataFrame(y_prediction)
        predictions_rf[predictions_rf<0]=0
        predictions_rf = predictions_rf.round(2)

        th=0.2
        score=r2_score(y_test[y_test>=th],y_prediction[y_test>=t
h])
        score_r2="R2 score is: "+str(score)
        MAE="MAE                        score                    is:
"+str(mean_absolute_error(y_test[y_test>=th],y_prediction[y_test
>=th]))
        MSE="MSE                        score                    is:
"+str(mean_squared_error(y_test[y_test>=th],y_prediction[y_test>
=th]))
        RMSE="RMSE                       score                   is:
"+str(np.sqrt(mean_squared_error(y_test[y_test>=th],y_prediction
[y_test>=th])))

        runs_n.append(run)
        R2S.append(score)
        MAES.append(mean_absolute_error(y_test[y_test>=th],y_pre
diction[y_test>=th]))
        RMSES.append(np.sqrt(mean_squared_error(y_test[y_test>=t
h],y_prediction[y_test>=th])))
        MSES.append(mean_squared_error(y_test[y_test>=th],y_pred
iction[y_test>=th]))

        max_depths.append(grid_search.best_params_["max_depth"])
        colsample_bytrees.append(grid_search.best_params_["colsa
mple_bytree"])
        gammas.append(grid_search.best_params_["gamma"])
        subsamples.append(grid_search.best_params_["subsample"])
        min_child_weights.append(grid_search.best_params_["min_c
hild_weight"])
        n_estimatorss.append(grid_search.best_params_["n_estimat
ors"])

        print(score_r2)
        print(MAE)
        print(MSE)
        print(RMSE)
```

```
        data_s=pd.DataFrame({'Observed':                    y_test,
'Predicted':y_prediction}, columns=['Observed', 'Predicted'])
        data_s.to_csv("paper/results/XGBoost/stats/data/"+cities
[i]+"/"+"run_"+str(run)+".txt")

    stats = pd.DataFrame({'run_number': runs_n, 'R2': R2S, 'MAE':
MAES, 'MSE': MSES, 'RMSE': RMSES,
                          'max_depth':                max_depths,
'colsample_bytree': colsample_bytrees,
                          'min_child_weight':  min_child_weights,
'gamma': gammas,
                          'subsample':                subsamples,
'n_estimators': n_estimatorss},
                           columns=['run_number',           'R2',
'MAE','MSE',     'RMSE',     'max_depth',     'colsample_bytree',
'min_child_weight',
                          'gamma', 'subsample', 'n_estimators'])
    stats.to_csv('paper/results/XGBoost/stats/'+cities[i]+'.txt'
, sep="\t")
```

Πίνακας 19. Δείγμα κώδικα σε python για την εκτέλεση των μοντέλου LSTM.

Εφαρμογή και εκτέλεση του μοντέλου LSTM

```
# Import Libraries
import pandas as pd
import numpy as np
import os
import glob
from tensorflow import keras
import matplotlib.pyplot as plt
import math
import statistics
from statistics import mean
import datetime as dt
from functools import reduce
from keras.callbacks import EarlyStopping
from sklearn.impute import KNNImputer
from math import sqrt
from sklearn.preprocessing import MinMaxScaler, StandardScaler,
RobustScaler
from       sklearn.metrics       import       mean_squared_error,
mean_absolute_error, r2_score
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense, Conv2D, Conv1D,
Bidirectional,      RepeatVector,      AvgPool2D,      MaxPool2D
,GlobalMaxPooling1D,  Flatten  ,  Dropout,  BatchNormalization,
Input,   ZeroPadding2D,    Activation,   ReLU,   MaxPooling1D,
TimeDistributed
from tensorflow.keras.optimizers import Adam, SGD, RMSprop
from tensorflow.keras.regularizers import l2
from  tensorflow.keras.callbacks  import  LearningRateScheduler,
ReduceLROnPlateau

# Split data
```

```
train_size = int(len(x) * 0.8)
test_size = len(x) - train_size
train, test = x.iloc[0:train_size], x.iloc[train_size:len(x)]
print(len(train), len(test))

# Normalize data based on MinMax algorithm
feature_columns = x.columns[0:len(x.columns)-1]

f_transformer = MinMaxScaler(feature_range=(0,1))
rain_transformer = MinMaxScaler(feature_range=(0,1))

f_transformer                                                    =
f_transformer.fit(train[feature_columns].to_numpy())
rain_transformer = rain_transformer.fit(train[['rain']])

train.loc[:,                    feature_columns]                 =
f_transformer.transform(train[feature_columns].to_numpy())
train['rain'] = rain_transformer.transform(train[['rain']])

test.loc[:,                    feature_columns]                  =
f_transformer.transform(test[feature_columns].to_numpy())
test['rain'] = rain_transformer.transform(test[['rain']])

# Def to create lag dataset from previous timesteps for RNN-LSTM
def create_lags_dataset(X, y, time_steps=1):
    Xs, ys = [], []
    for i in range(len(X) - time_steps):
        v = X.iloc[i:(i + time_steps)].values
        Xs.append(v)
        ys.append(y.iloc[i + time_steps])
    return np.array(Xs), np.array(ys)

# Make data cube for input
time_steps = 2

# reshape to [samples, time_steps, n_features]

X_train,   y_train   =   create_lags_dataset(train,   train.rain,
time_steps)
X_test, y_test = create_lags_dataset(test, test.rain, time_steps)

print("size - "+"timesteps_window_size -"+" nof_features")
print(X_train.shape)

from tensorflow.python.framework.func_graph import flatten
try:
    del model
except:
    pass

# Settings of model
model = Sequential()
units=32
```

```
model.add(Bidirectional(LSTM(units,
input_shape=[X_train.shape[1],                X_train.shape[2]],
return_sequences=False)))
model.add((Dense(units, activation="relu")))
model.add(Dense(1))

metrics = ['mae', 'mape']
lr = 1e-4
opt = Adam(learning_rate=lr)
reduce_lr = ReduceLROnPlateau(monitor='val_loss', factor=0.2,
                              patience=5)

# Compile model
model.compile(optimizer= opt, loss="mse")

# Fit network
history = model.fit(X_train, y_train,
                    epochs=40,
                    batch_size=32,
                    validation_split=0.1,
                    verbose=2,
                    shuffle=False,callbacks=[reduce_lr])
print(history)

# Plot history
plt.plot(history.history['loss'], label='train')
plt.plot(history.history['val_loss'], label='test')
plt.legend()
plt.show()

# Make predictions
y_pred = model.predict(X_test)
y_train_inv                                                    =
rain_transformer.inverse_transform(y_train.reshape(-1, 1))
y_test_inv = rain_transformer.inverse_transform(y_test.reshape(-
1, 1))
y_pred_inv = rain_transformer.inverse_transform(y_pred.reshape(-
1, 1))
y_preds_d=y_pred_inv.flatten()[y_test_inv.flatten()>=0.2]
y_tests_d=y_test_inv.flatten()[y_test_inv.flatten()>=0.2]

rmse = sqrt(mean_squared_error(y_tests_d, y_preds_d))
mae = mean_absolute_error(y_tests_d, y_preds_d)
mse = mean_squared_error(y_tests_d, y_preds_d)
score_r2 =r2_score(y_tests_d, y_preds_d)

print('Test R2:', score_r2)
print('Test MAE: %.8f' % mae)
print('Test MSE: %.8f' % mse)
print('Test RMSE: %.8f' % rmse)

predictions_rf=pd.DataFrame(y_pred_inv)
predictions_rf[predictions_rf<0]=0
predictions_rf = predictions_rf.round(2)

# Plot results
```

```
plt.plot(y_test_inv.flatten(), marker='.', label="true")
plt.plot(y_pred_inv.flatten(),      color='red',      marker='.',
label="prediction")
plt.ylabel('Rain amount')
plt.xlabel('Time Step')
plt.legend()
plt.show()
```

Συνοπτικά στην Εικόνα 31 παρουσιάζονται οι τεχνολογίες που χρησιμοποιήθηκαν συνολικά σε αυτή την εργασία από την λήψη, διαχείριση και επεξεργασία των δεδομένων μέχρι και την οργάνωση και εφαρμογή των τεχνικών και των μοντέλων μηχανικής μάθησης.

Εικόνα 31. Λογισμικό, τεχνολογίες και βιβλιοθήκες που χρησιμοποιήθηκαν για την εκτέλεση της εργασίας.

6.4 Ανάπτυξη Web - GIS εφαρμογής για την παρουσίαση των αποτελεσμάτων

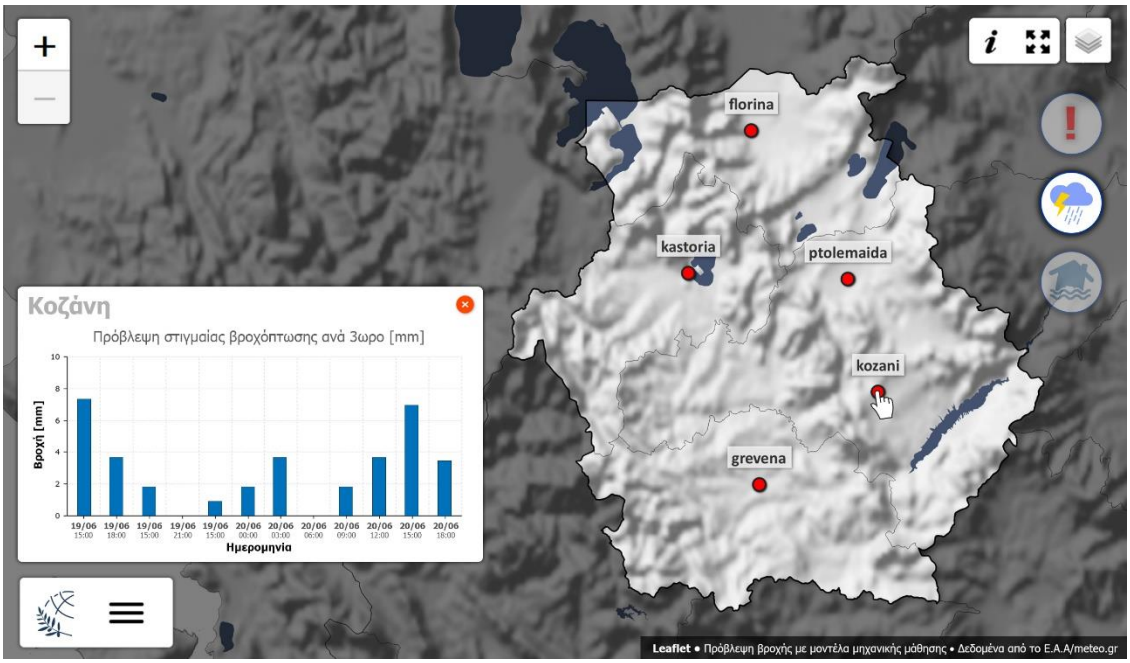
Το τελικό στάδιο αυτής της εργασίας ήταν η ανάπτυξη μιας διαδικτυακής εφαρμογής GIS (Γεωχωρικών δεδομένων) με σκοπό την παρουσίαση των αποτελεσμάτων με βάση το καλύτερο μοντέλο που αναπτύχθηκε. Οι τεχνολογίες και τα προγραμματιστικά εργαλεία που χρησιμοποιήθηκαν για την ανάπτυξη της εφαρμογής αναφέρονται στον Πίνακα 20.

Πίνακας 20. Τεχνολογίες και προγραμματιστικά εργαλεία για την ανάπτυξη της Web-Gis εφαρμογής.

Ενέργειες	**Προγραμματιστικά εργαλεία**
Ανάπτυξη σελίδας	Javascript, HTML, CSS, PHP, python
Οπτικοποίηση των Γεωχωρικών δεδομένων	Javascript, Leaflet
Δημιουργία γραφημάτων – Επιχειρησιακές λειτουργίες	Javascript, AJAX, jQuery
Διοχέτευση δεδομένων	Python, Javascript, PHP

Η εφαρμογή έχει ως σκοπό την παρουσίαση των προβλέψεων βροχής για το επόμενο 3ωρο για 2 ημέρες μπροστά. Τα νέα δεδομένα θα ανασύρονται από τη επιχειρησιακή μονάδα ΜΕΤΕΟ του Εθνικού Αστεροσκοπείου Αθηνών. Επίσης, έχουν τεθεί κάποια όρια με βάση την βιβλιογραφία για τις ακραίες βροχοπτώσεις, έτσι ώστε να αναφέρονται κάποιου είδους ειδοποιήσεις (εικονίδια δεξιά στην Εικόνα 32) στην οθόνη της εφαρμογής. Κάνοντας κλικ σε αυτές ο χρήστης θα μπορεί να εντοπίσει τα σημεία που αναμένεται έντονη βροχόπτωση και την ακριβής χρονική στιγμή.

Τελικός σκοπός της εφαρμογής είναι η άμεση ενημέρωση και έγκαιρη προειδοποίηση των χρηστών/πολιτών για τα ακραία καιρικά φαινόμενα στην περίπτωσή μας η έντονη βροχή.



Εικόνα 32. Στιγμιότυπο της διαδικτυακής εφαρμογής GIS παρουσίασης των αποτελεσμάτων.

Ο σύνδεσμος που οδηγεί στην εφαρμογή είναι ο παρακάτω και θα είναι άμεσα διαθέσιμος

- https://climatebook.gr/feedline/ml_rain

7

Επίλογος

Σε αυτή την ενότητα κλείνουμε την παρούσα εργασία κάνοντας μια μικρή σύνοψη των ενεργειών και των εφαρμογών αυτής της εργασίας καθώς επίσης και την διατύπωση κάποιων συμπερασμάτων με βάση τα αποτελέσματα. Τέλος, κλείνουμε με τις μελλοντικές σκέψεις και επεκτάσεις σε ότι έχει να κάνει με την εξέλιξη της παρούσας εργασίας.

7.1 Σύνοψη και συμπεράσματα

Σε αυτή την εργασία αναπτύχθηκαν και εφαρμόστηκαν μοντέλα μηχανικής μάθησης με σκοπό την πρόβλεψη της στιγμιαίας βροχόπτωσης που αναμένεται για τις αμέσως επόμενες 3 ώρες. Η ανάλυση βασίστηκε σε δεδομένα 4 ετών από μετεωρολογικούς σταθμούς στην περιοχή μελέτης, που βρίσκονται στη Δυτική Μακεδονία στην Ελλάδα. Γενικά, όλα τα μοντέλα κατάφεραν να αποδώσουν αρκετά καλά παρέχοντας συνεπείς προβλέψεις και φαίνεται πως αφομοίωσαν τη σύνθετη σχέση μεταξύ των ατμοσφαιρικών συνθηκών και της παραγωγής βροχοπτώσεων. Τα καλύτερα αποτελέσματα εξήχθησαν από τα μοντέλα σχετικά με τα δέντρα αποφάσεων (Random Forest και XGBoost), καθώς με την κλασσική στατιστική τεχνική της γραμμικής παλινδρόμησης. Λιγότερο καλά απέδωσαν τα τεχνητά νευρωνικά δίκτυα (ANNs) όπως το LSTM και το CNN.

7.2 Μελλοντικές επεκτάσεις

Τα αποτελέσματα αυτής της εργασίας μας δίνουν κίνητρο για περαιτέρω έρευνα για την εφαρμογή μοντέλων μηχανικής μάθησης για την πρόβλεψη βροχοπτώσεων με σκοπό την βελτίωση των υπαρχουσών μεθόδων αριθμητικής πρόβλεψης καιρού. Μια πρωταρχική σκέψη για μελλοντική έρευνα είναι να χρησιμοποιήσουμε βασικά μοντέλα (αριθμητικά μοντέλα πρόγνωσης καιρού) για σύγκριση των προβλέψεων τους με τις προβλέψεις των προτεινόμενων μοντέλων αυτής της εργασίας. Τέλος, ως μελλοντική επέκταση θα μπορούσε να ήταν η

εκπαίδευση των μοντέλων σε όλα τα πλεγματικά σημεία της Ελλάδος με σκοπό την παραγωγή προβλέψεων της βροχής σε επίπεδο πλέγματος κατά μήκος και πλάτος όλης της χώρας.

8

Βιβλιογραφία

[1] Kim, H.-U & Bae, T.-S. (2017). Preliminary study of deep learning-based precipitation prediction. Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography. 35. 423-429. 10.7848/ksgpc.2017.35.5.423.

[2] Zhang, C., Zeng, J., Wang, H., Ma, L., & Chu, H. (2019). Correction model for rainfall forecasts using the LSTM with multiple meteorological factors. In Meteorological Applications (Vol. 27, Issue 1). Wiley. https://doi.org/10.1002/met.1852.

[3] Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems, 28.

[4] Ju, J., & Liu, F.-A. (2021). Multivariate Time Series Data Prediction Based on ATT-LSTM Network. In Applied Sciences (Vol. 11, Issue 20, p. 9373). MDPI AG. https://doi.org/10.3390/app11209373.

[5] Poornima, S., & Pushpalatha, M. (2019). Prediction of Rainfall Using Intensified LSTM Based Recurrent Neural Network with Weighted Linear Units. In Atmosphere (Vol. 10, Issue 11, p. 668). MDPI AG. https://doi.org/10.3390/atmos10110668.

[6] Q. Zhao, Y. Liu, W. Yao and Y. Yao, "Hourly Rainfall Forecast Model Using Supervised Learning Algorithm," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-9, 2022, Art no. 4100509, doi: 10.1109/TGRS.2021.3054582.

[7] Y. R. Sari, E. C. Djamal and F. Nugraha, "Daily Rainfall Prediction Using One Dimensional Convolutional Neural Networks," 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE), Yogyakarta, Indonesia, 2020, pp. 90-95, doi: 10.1109/IC2IE50715.2020.9274572.

[8] Anochi, J. A., de Almeida, V. A., & de Campos Velho, H. F. (2021). Machine Learning for Climate Precipitation Prediction Modeling over South America. In Remote Sensing (Vol. 13, Issue 13, p. 2468). MDPI AG. https://doi.org/10.3390/rs13132468.

[9] F. Manokij, K. Sarinnapakorn and P. Vateekul, "Forecasting Thailand's Precipitation with Cascading Model of CNN and GRU," 2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE), Pattaya, Thailand, 2019, pp. 1-6, doi: 10.1109/ICITEED.2019.8929975.

[10] Skofronick-Jackson, G., Kulie, M., Milani, L., Munchak, S. J., Wood, N. B., & Levizzani, V. (2019). Satellite Estimation of Falling Snow: A Global Precipitation Measurement (GPM) Core Observatory Perspective. In Journal of Applied Meteorology and Climatology (Vol. 58, Issue 7, pp. 1429–1448). American Meteorological Society. https://doi.org/10.1175/jamc-d-18-0124.1.

[11] Liyew, C. M., & Melese, H. A. (2021). Machine learning techniques to predict daily rainfall amount. In Journal of Big Data (Vol. 8, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s40537-021-00545-4.

[12] Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. American journal of epidemiology, 188(12), 2222-2239.

[13] El Naqa, I., & Murphy, M. J. (2015). What is machine learning? (pp. 3-11). Springer International Publishing.

[14] Macabiog, R. E. N., & Dela Cruz, J. C. (2019). Rainfall Predictive Approach for La Trinidad, Benguet using Machine Learning Classification. In 2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM). 2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and

Control, Environment, and Management (HNICEM). IEEE. https://doi.org/10.1109/hnicem48295.2019.9072761.

[15] E. Dehaerne, B. Dey, S. Halder, S. De Gendt and W. Meert, "Code Generation Using Machine Learning: A Systematic Review," in IEEE Access, vol. 10, pp. 82434-82455, 2022, doi: 10.1109/ACCESS.2022.3196347.

[16] Grace, R. K., & Suganya, B. (2020). Machine Learning based Rainfall Prediction. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE. https://doi.org/10.1109/icaccs48705.2020.9074233.

[17] Eberly, L. E. (2007). Multiple Linear Regression. In Topics in Biostatistics (pp. 165–187). Humana Press. https://doi.org/10.1007/978-1-59745-530-5_9.

[18] Naik, A. R., Deorankar, A. V., & Ambhore, P. B. (2020). Rainfall Prediction based on Deep Neural Network: A Review. In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE. https://doi.org/10.1109/icimia48430.2020.9074892.

[19] Oswal, N. (2021). Predicting Rainfall using Machine Learning Techniques. Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.36227/techrxiv.14398304.v1.

[20] Sarasa-Cabezuelo, A. (2022). Prediction of Rainfall in Australia Using Machine Learning. In Information (Vol. 13, Issue 4, p. 163). MDPI AG. https://doi.org/10.3390/info13040163.

[21] What is a Random Forest? Available online: https://www.tibco.com/reference-center/what-is-a-random-forest (accessed on 06 June 2023).

[22] Li, W., Yin, Y., Quan, X., & Zhang, H. (2019). Gene Expression Value Prediction Based on XGBoost Algorithm. In Frontiers in Genetics (Vol. 10). Frontiers Media SA. https://doi.org/10.3389/fgene.2019.01077.

[23] Dong, W., Huang, Y., Lehane, B., & Ma, G. (2020). XGBoost algorithm-based prediction of concrete electrical resistivity for structural health

monitoring. In Automation in Construction (Vol. 114, p. 103155). Elsevier BV. https://doi.org/10.1016/j.autcon.2020.103155.

[24] How XGBoost Works. Available online: https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html (accessed on 10 June 2023).

[25] The Science and Art of Meteorology, National Geographic. Available online: https://education.nationalgeographic.org/resource/science-art-meteorology/ (accessed on 12 June 2023).

[26] Digital meteorological encyclopedia of the National Observatory of Athens, edited by the scientific team of meteo.gr. Available online: https://wiki.meteo.gr/ (accessed on 12 June 2023).

[27] https://www.irjmets.com/uploadedfiles/paper//issue_2_february_2022/18740/final/fin_irjmets1643811133.pdf?fbclid=IwAR3zYtGL16REcguZp1sLu9X_Fy-7k59rzZpdSstsLPiv8G-jmgf-6jFLRN4.

[28] Kimura, R. (2002). Numerical weather prediction. In Journal of Wind Engineering and Industrial Aerodynamics (Vol. 90, Issues 12–15, pp. 1403–1414). Elsevier BV. https://doi.org/10.1016/s0167-6105(02)00261-1.

[29] Lagouvardos, K., Kotroni, V., Koussis, A., Feidas, H., Buzzi, A., & Malguzzi, P. (2003). The Meteorological Model BOLAM at the National Observatory of Athens: Assessment of Two-Year Operational Use. In Journal of Applied Meteorology (Vol. 42, Issue 11, pp. 1667–1678). American Meteorological Society. https://doi.org/10.1175/1520-0450(2003)042<1667:tmmbat>2.0.co;2.

[30] Bochenek, B., & Ustrnul, Z. (2022). Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives. In Atmosphere (Vol. 13, Issue 2, p. 180). MDPI AG. https://doi.org/10.3390/atmos13020180.

[31] Shah, U., Garg, S., Sisodiya, N., Dube, N., & Sharma, S. (2018). Rainfall Prediction: Accuracy Enhancement Using Machine Learning and Forecasting Techniques. In 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC). 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE. https://doi.org/10.1109/pdgc.2018.8745763.

[32] Kala, A., & Vaidyanathan, S. G. (2018). Prediction of Rainfall Using Artificial Neural Network. In 2018 International Conference on Inventive Research in Computing Applications (ICIRCA). 2018 International

Conference on Inventive Research in Computing Applications (ICIRCA). IEEE. https://doi.org/10.1109/icirca.2018.8597421.

[33] Basha, C. Z., Bhavana, N., Bhavya, P., & V, S. (2020). Rainfall Prediction using Machine Learning & Deep Learning Techniques. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE. https://doi.org/10.1109/icesc48915.2020.9155896.

[34] Murat, M., Malinowska, I., Gos, M., & Krzyszczak, J. (2018). Forecasting daily meteorological time series using ARIMA and regression models. In International Agrophysics (Vol. 32, Issue 2, pp. 253–264). Walter de Gruyter GmbH. https://doi.org/10.1515/intag-2017-0007.

[35] Cramer, S., Kampouridis, M., Freitas, A. A., & Alexandridis, A. K. (2017). An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. In Expert Systems with Applications (Vol. 85, pp. 169–181). Elsevier BV. https://doi.org/10.1016/j.eswa.2017.05.029.

[36] Hernández, E., Sanchez-Anguix, V., Julian, V., Palanca, J., & Duque, N. (2016). Rainfall Prediction: A Deep Learning Approach. In Lecture Notes in Computer Science (pp. 151–162). Springer International Publishing. https://doi.org/10.1007/978-3-319-32034-2_13.

[37] Graham, Anosh & Pathak, Ekta & Correspondence, Anosh & Graham. (2017). Time series analysis model to forecast rainfall for Allahabad region. journal of pharmacognosy and phytochemistry. 6. 1418-1421.

[38] Mishra, N., Soni, H. K., Sharma, S., & Upadhyay, A. K. (2018). Development and Analysis of Artificial Neural Network Models for Rainfall Prediction by Using Time-Series Data. In International Journal of Intelligent Systems and Applications (Vol. 10, Issue 1, pp. 16–23). MECS Publisher. https://doi.org/10.5815/ijisa.2018.01.03.

[39] Esteves, J. T., de Souza Rolim, G., & Ferraudo, A. S. (2018). Rainfall prediction methodology with binary multilayer perceptron neural networks. In Climate Dynamics (Vol. 52, Issues 3–4, pp. 2319–2331). Springer Science and Business Media LLC. https://doi.org/10.1007/s00382-018-4252-x.

[40] Khan, M. I., & Maity, R. (2020). Hybrid Deep Learning Approach for Multi-Step-Ahead Daily Rainfall Prediction Using GCM Simulations. In IEEE

Access (Vol. 8, pp. 52774–52784). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/access.2020.2980977.

[41] How to use the CDS API. Available online: https://cds.climate.copernicus.eu/api-how-to (accessed on 15 June 2023).