
Μέθοδοι επίγνωσης κατάστασης ακραίων καιρικών φαινομένων αξιοποιώντας
IoT δίκτυα αισθητήρων και τεχνικές πληθοπορισμού

Πτυχιακή Εργασία



Τμήμα Πληροφορικής
Ιόνιο Πανεπιστήμιο

Στοιχεία Φοιτητή:
Θεοχάρης Παναγιώτης Χαραλαμπίδης Π2018127

Κέρκυρα, 2023

Επιβλέπουσα καθηγήτρια:

Ελένη Χριστοπούλου

Επιτροπή:

Μάρκος Αυλωνίτης

Μιχαήλ Στεφανιδάκης

Περίληψη

Οι εξελίξεις στον τομέα της πληροφορικής καθώς και στον κόσμο του Διαδικτύου των Πραγμάτων μας δίνουν πλέον πρόσβαση σε πληθώρα δεδομένων και μεθόδων ανάλυσης αυτών, ανοίγοντας νέους ορίζοντες στο πώς μπορούμε να βελτιώσουμε την ποιότητα ζωής μας. Σε συνδυασμό με την πρόοδο στον τομέα της τεχνητής νοημοσύνης μπορούμε να αναπτύξουμε μοντέλα μηχανικής μάθησης τα οποία είναι ικανά να προβλέψουν τις εξελίξεις στην καιρική κατάσταση μιας πόλης.

Η παρούσα πτυχιακή εργασία ασχολείται με την μελέτη μεθόδων πρόβλεψης καιρικών φαινομένων στο νησί της Κέρκυρας με τη χρήση μοντέλων μηχανικής μάθησης, πληθοπορισμού και τοπικού IoT δικτύου αισθητήρων καιρικών φαινομένων, καθώς επίσης και την εφαρμογή ορισμένων μεθόδων στην συλλογή και επεξεργασία δεδομένων και την ανάπτυξη μοντέλων μηχανικής μάθησης τόσο για την κατηγοριοποίηση της καιρικής κατάστασης με τα δεδομένα που παρέχονται από το δίκτυο αισθητήρων όσο και για την συμπλήρωση των ιστορικών δεδομένων που χρησιμοποιήθηκαν κατά την εκπαίδευση του μοντέλου κατηγοριοποίησης. Μέσω των μοντέλων αυτών, θα μπορούμε να έχουμε λύσεις έγκαιρης πρόβλεψης και ενημέρωσης για τομείς που χρειάζονται έγκαιρη ενημέρωση σε ακριβή σημεία που δεν καλύπτονται λεπτομερώς από δελτία καιρού, όπως για παράδειγμα στη γεωργία ή τον τουρισμό.

Σε συνδυασμό με το χαμηλό κόστος τέτοιων δικτύων αισθητήρων, η δημιουργία ενός μεγάλου δικτύου σε μια περιοχή ενδιαφέροντος είναι εφικτή και δύναται να παρέχει πολύ χρήσιμες πληροφορίες σε χρήστες που μπορούν να επωφεληθούν σημαντικά από κάτι τέτοιο.

Πίνακας Περιεχομένων

1	Εισαγωγή.....	7
1.1	IoT και Τεχνητή Νοημοσύνη.....	7
1.2	Σκοπός της πτυχιακής εργασίας.....	7
1.3	Στόχος της πτυχιακής εργασίας.....	8
1.4	Δομή της εργασίας.....	8
2	Τεχνητή Νοημοσύνη.....	10
2.1	Μηχανική μάθηση.....	11
2.2	Βαθιά Μάθηση.....	12
3	Διαδίκτυο των Πραγμάτων.....	14
4	Σχετική Βιβλιογραφία.....	28
4.1	Βασική ορολογία μετεωρολογίας.....	28
4.1.1	Θερμοκρασία.....	28
4.1.2	Ατμοσφαιρική Πίεση.....	28
4.1.3	Υγρασία.....	28
4.1.4	Δείκτης δυσφορίας.....	29
4.1.5	Πρόγνωση καιρού.....	30
4.2	Το σύστημα Smart Citizen.....	31
4.3	Πληθοπορισμός.....	35
4.4	Αντίστοιχες εφαρμογές.....	41
4.4.1	Ανίχνευση καιρικών φαινομένων μέσω φωτογραφιών με τη χρήση νευρωνικών δικτύων.....	41
4.4.2	Πρόγνωση καιρού πολλαπλών κατηγοριών από το Twitter χρησιμοποιώντας προσεγγίσεις μηχανικής μάθησης και πληθοπορισμού.....	43
5	Εργαλεία που χρησιμοποιήθηκαν.....	45
6	Μεθοδολογία.....	55
6.1	Δεδομένα.....	55
6.1.1	Κιτ αισθητήρων Smart Citizen.....	55
6.1.2	Ιστορικό σύνολο δεδομένων μέσω Openweather.....	58
6.1.3	Προετοιμασία συνόλου δεδομένων.....	59
6.2	Μοντέλα.....	66
6.2.1	Μοντέλο MICE.....	66
6.3	Decision Tree Classifier.....	70
7	Συμπεράσματα και Μελλοντικές επεκτάσεις.....	77
	Βιβλιογραφία.....	79

Πίνακας εικόνων

Εικόνα 1: Η τεχνητή νοημοσύνη και τα πεδία που εμπίπτουν σε αυτή	10
Εικόνα 2: Ενδεικτικά περιβάλλοντα χρήσης IoT	17
Εικόνα 3: Τιμές Θερμοκρασίας / Υγρασίας για τον υπολογισμό του δείκτη δυσφορίας..	30
Εικόνα 4: Επίπεδα δυσφορίας.....	30
Εικόνα 5: Ένα Smart Citizen Kit με την πλακέτα Urban στο περίβλημά του. Πηγή: https://wikifactory.com/@elgranpri/smart-citizen-kit	32
Εικόνα 6: Η μπροστά και πίσω όψη του Data Board	33
Εικόνα 7: Αισθητήρας Gas Pro Board.....	33
Εικόνα 8: Πίνακας επέκτασης PM Board	34
Εικόνα 9: Ενδεικτικό παράδειγμα χρήσης του PM Board σε συνδυασμό με αισθητήρες αιωρούμενων σωματιδίων και του Gas Pro Board	34
Εικόνα 10: Παραδείγματα CAPTCHAs.....	40
Εικόνα 11: Visual Studio Code.....	46
Εικόνα 12 Google Colab.....	47
Εικόνα 13: Αρχική οθόνη Weka.....	48
Εικόνα 14: Weka Workbench.....	49
Εικόνα 15: Εξίσωση κανονικοποίησης.....	50
Εικόνα 16: Η καρτέλα Classify του Weka	51
Εικόνα 17: Παράδειγμα Classifier Output μέσω του Weka.....	52
Εικόνα 18: Ενδεικτικό δείγμα δεδομένων ενός Smart Citizen Kit σε μορφή CSV.....	56
Εικόνα 19: Η μπροστινή και πίσω όψη του Urban Sensing Board.....	57
Εικόνα 20: Ενδεικτικό δείγμα δεδομένων ενός Smart Citizen Kit σε μορφή JSON.....	57
Εικόνα 21: Εισαγωγή δεδομένων και πρώτα βήματα προετοιμασίας.....	59
Εικόνα 22: Δημιουργία στηλών Unix epoch	60
Εικόνα 23: Συνένωση συνόλων δεδομένων	60
Εικόνα 24: Συμπλήρωση κενών στηλών weather_description.....	60
Εικόνα 25: Παράδειγμα μεθόδων bfill και ffill.....	61
Εικόνα 26: Γράφημα δεδομένων θερμοκρασίας για το διάστημα Αύγουστος-Σεπτέμβριος 2021	61
Εικόνα 27: Μέση ωριαία θερμοκρασία στο διάστημα 13-08-2021 με 1-9-2021	62
Εικόνα 28: Μέση ωριαία υγρασία στο διάστημα 13-8-2021 με 1-9-2021.....	62
Εικόνα 29: Ομαδοποίηση καιρικών φαινομένων.....	63
Εικόνα 30: Κλάσεις συνόλου δεδομένων πριν και μετά την ομαδοποίηση.....	63
Εικόνα 31: Κατανομή παρατηρήσεων πριν την ομαδοποίηση	64
Εικόνα 32: Κατανομή παρατηρήσεων μετά την ομαδοποίηση	64
Εικόνα 33: Πίνακας σύγχυσης του αρχικού συνόλου δεδομένων.....	65
Εικόνα 34: Πίνακας σύγχυσης του ομαδοποιημένου συνόλου δεδομένων.....	65
Εικόνα 35: Ανακατανομή στηλών και αποθήκευση δεδομένων	66
Εικόνα 36: Εισαγωγή συνόλου δεδομένων και φιλτράρισμα των επιθυμητών ημερομηνιών	67
Εικόνα 37: Κλήση των αντικειμένων της MICE.....	67
Εικόνα 38: Διαδικασία παρεμβολής των κενών τιμών.....	68
Εικόνα 39: Δημιουργία γραφήματος με βάση τα καινούρια δεδομένα.....	69
Εικόνα 40: Γράφημα θερμοκρασίας-ημερομηνίας με δεδομένα που προέκυψαν από την τεχνική MICE.....	70

Εικόνα 41: Εισαγωγή δεδομένων και διαγραφή της στήλης timestampDT	72
Εικόνα 42: Διαχωρισμός test-train.....	72
Εικόνα 43: Εκκίνηση εκπαίδευσης του μοντέλου	73
Εικόνα 44: Καμπύλη μάθησης για μοντέλο βάθους 25 επιπέδων	73
Εικόνα 45: Πίνακας σύγκρισης μοντέλου βάθους 25 επιπέδων.....	74
Εικόνα 46: Καμπύλη μάθησης για μοντέλο βάθους 12 επιπέδων.....	74
Εικόνα 47: Πίνακας σύγκρισης για μοντέλο βάθους 12 επιπέδων.....	75
Εικόνα 48: Παραγωγή καμπύλης μάθησης μοντέλου.....	75
Εικόνα 49: Κώδικας παραγωγής του πίνακα σύγκρισης	76

1 Εισαγωγή

1.1 IoT και Τεχνητή Νοημοσύνη

Σήμερα όλο και περισσότερες συσκευές συνδέονται στο διαδίκτυο, δημιουργώντας το λεγόμενο διαδίκτυο των πραγμάτων. Συνοπτικά ο όρος αυτός αναφέρεται στην επέκταση των δυνατοτήτων επεξεργασίας και επικοινωνίας δικτύου σε συσκευές που τυπικά δεν θεωρούνται υπολογιστές (όπως αισθητήρες, οικιακές συσκευές, αυτοκίνητα και άλλα). Χάρη στις εξελίξεις στην υπολογιστική ισχύ, την σμίκρυνση ηλεκτρονικών συσκευών, την ανάπτυξη νέων μεθόδων διασύνδεσης με το διαδίκτυο καθώς και τις εξελίξεις στον τομέα της ανάλυσης δεδομένων, το διαδίκτυο των πραγμάτων ενέχει πολύ σημαντική θέση στην οικονομία και την κοινωνία.[1] Ένα παράδειγμα στον τομέα της υγείας που επικεντρώνεται στη χρήση IoT μετρητών αρτηριακής πίεσης σε ηλικιωμένους και στη συνέπεια τους ως προς την λήψη της κατάλληλης αγωγής τους δείχνει έστω σε πρώιμα στάδια θετική επιρροή τέτοιων τεχνολογιών στην υγεία των ατόμων[2]. Συνδυάζοντας το διαδίκτυο των πραγμάτων με την τεχνητή νοημοσύνη ανοίγονται μπροστά μας απεριόριστες δυνατότητες. Ο όρος Τεχνητή Νοημοσύνη έχει πλέον τεράστια διείσδυση τόσο στην καθημερινότητά μας όσο και σε τομείς με τους οποίους αλληλοεπιδρούμε, όπως σε επιχειρήσεις, υγεία και βιομηχανίες. Η χρήση της τεχνητής νοημοσύνης, για παράδειγμα, στον τομέα της υγείας συναντάται στην ανάλυση ακτινογραφιών θώρακος με σκοπό την ανάλυση πιθανότητας να έχει μολυνθεί ο ασθενής με κορονοϊό. Μια ακόμα εφαρμογή βρίσκεται στον τομέα της γνωστικής ψυχολογίας, όπου με την τεχνητή νοημοσύνη γίνεται δυνατή η προσομοίωση του ανθρώπινου εγκεφάλου ώστε να δοκιμαστούν θεωρίες γύρω από το πώς λειτουργεί σε σχέση με διάφορες βασικές διεργασίες όπως η αναγνώριση προσώπων. Ωστόσο ο όρος είναι γενικότερος και επικαλύπτει τις διακριτές επιστήμες που αποτελούν την τεχνητή νοημοσύνη. Η τεχνητή νοημοσύνη στοχεύει στο να δώσει στις μηχανές και τα προγράμματα τη δυνατότητα να κατανοούν, μαθαίνουν, κρίνουν και να ανταποκρίνονται κατάλληλα σε συνθήκες που δεν έχουν προγραμματιστεί, ή προβλεφθεί από τον προγραμματιστή. [3]

1.2 Σκοπός της πτυχιακής εργασίας

Ο τομέας που θα μας απασχολήσει στην παρούσα εργασία αφορά την χρήση αισθητήρων IoT, μοντέλων μηχανικής μάθησης και τεχνικών πληθοπορισμού μπορούμε να κάνουμε προβλέψεις για την εξέλιξη των καιρικών συνθηκών. Για να γίνει αυτό, θα πρέπει αρχικά να γίνει μελέτη των τεχνικών και τεχνολογιών που χρησιμοποιήθηκαν ή θέλουμε να χρησιμοποιηθούν σε μελλοντικές επεκτάσεις. Οι τεχνολογίες αυτές είναι η μηχανική μάθηση, το crowdsensing (πληθοπορισμός) και το internet of things. Η μελέτη αυτή έδωσε

μια βαθύτερη κατανόηση στο πώς μπορεί να αναπτυχθεί μια τέτοια πλατφόρμα, πώς μπορεί να εκμεταλλευτεί την ανάπτυξη του IoT, και το πώς οι χρήστες μπορούν να συμμετάσχουν στη συλλογή δεδομένων. Στη συνέχεια θα αναλυθεί η πορεία ανάπτυξης των μεθόδων που θα χρησιμοποιήθηκαν για να πραγματοποιηθεί ο σκοπός της πτυχιακής, δηλαδή η συλλογή και επεξεργασία δεδομένων από IoT αισθητήρες, καθώς και η ανάπτυξη και εκπαίδευση των απαραίτητων για τον σκοπό της εργασίας μοντέλων.

1.3 Στόχος της πτυχιακής εργασίας

Σε αυτή την ενότητα θα δούμε τους στόχους της εργασίας και τι θέλουμε να έχουμε πετύχει με την περάτωση αυτής. Το πρόβλημα που μας απασχολεί χωρίζεται σε δύο βασικά τμήματα, τη συλλογή δεδομένων από τους αισθητήρες, και την ανάπτυξη και εκπαίδευση των μοντέλων.

Ο πρώτος στόχος είναι η συλλογή των δεδομένων και ο εμπλουτισμός αυτών μέσω άλλων πηγών ώστε το τελικό σύνολο δεδομένων να παρέχει τις απαραίτητες πληροφορίες για την ακριβή κατηγοριοποίηση των καιρικών φαινομένων, καθώς οι αισθητήρες από τους οποίους συλλέχθηκαν τα δεδομένα δεν παρέχουν πληροφορίες για τα τρέχοντα καιρικά φαινόμενα, παρά μόνο για βασικές καιρικές τιμές (θερμοκρασία, υγρασία, ατμοσφαιρική πίεση και τα λοιπά). Θα πρέπει επίσης να γίνει η απαραίτητη επεξεργασία των δεδομένων ώστε στη φάση εκπαίδευσης του μοντέλου να βεβαιωθούμε πως παρέχεται όσο το δυνατόν καλύτερη πληροφορία.

Ο δεύτερος στόχος αφορά την εύρεση του ιδανικότερου αλγορίθμου μάθησης, την ανάπτυξη ενός μοντέλου κατηγοριοποίησης καιρικών φαινομένων το οποίο θα βασίζεται στον αλγόριθμο που επιλέχθηκε και την απαραίτητη παραμετροποίηση του αλγορίθμου ώστε να επιτευχθεί το καλύτερο δυνατό αποτέλεσμα. Παράλληλα με το πρώτο μοντέλο, αναπτύσσεται και ένα δεύτερο το οποίο αποσκοπεί στην συμπλήρωση των κενών στα δεδομένα, όπου εμφανίζονται, με τη μέγιστη δυνατή πιστότητα με τα πραγματικά δεδομένα.

Τελευταίος στόχος είναι η συμπληρωματική βιβλιογραφική έρευνα πάνω σε μεθόδους πληθοπορισμού και πώς αυτές μπορούν να παρέχουν δεδομένα για τη βελτίωση των μοντέλων μηχανικής μάθησης, συλλέγοντας πληροφορίες από τους ενδιαφερόμενους χρήστες.

1.4 Δομή της εργασίας

Τα επόμενα κεφάλαια διαμορφώνονται ως εξής: Στο δεύτερο κεφάλαιο παρουσιάζεται η τεχνητή νοημοσύνη και τα πεδία που ανήκουν σε αυτή, δηλαδή η μηχανική μάθηση και

η βαθιά μάθηση. Παρουσιάζονται επίσης τα διάφορα είδη αλγορίθμων και μέθοδοι εκπαίδευσης μοντέλων.

Στο τρίτο κεφάλαιο παρουσιάζονται οι βασικοί μετεωρολογικοί όροι που θα μας απασχολήσουν σε επόμενα κεφάλαια.

Στο τέταρτο κεφάλαιο παρουσιάζεται η σχετική βιβλιογραφία καθώς και παρόμοιες εφαρμογές της τεχνητής νοημοσύνης και του διαδικτύου των πραγμάτων στον τομέα της πρόβλεψης καιρικών φαινομένων.

Στο πέμπτο κεφάλαιο γίνεται η ανάλυση των εργαλείων που χρησιμοποιήθηκαν στην παρούσα εργασία.

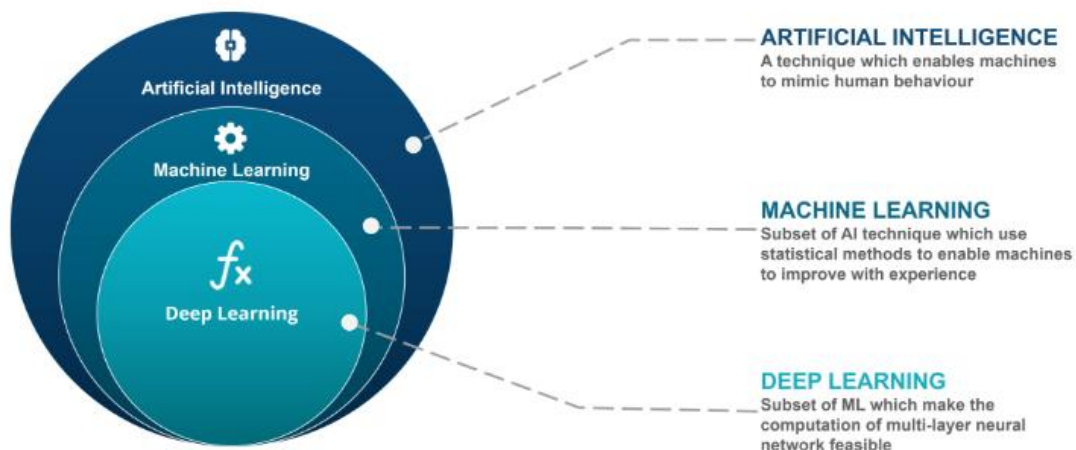
Το έκτο κεφάλαιο αφορά τη διαδικασία συλλογής, εμπλουτισμού και επεξεργασίας των δεδομένων.

Στο έβδομο κεφάλαιο αναλύεται η επιλογή αλγορίθμων μάθησης καθώς και η ανάπτυξη και εκπαίδευση των μοντέλων μάθησης.

Τέλος, στο όγδοο κεφάλαιο έχουμε τα συμπεράσματα και τις πιθανές μελλοντικές επεκτάσεις.

2 Τεχνητή Νοημοσύνη

Η τεχνητή νοημοσύνη ορίζεται ως το ακαδημαϊκό πεδίο το οποίο αποσκοπεί στην ανάπτυξη εφαρμογών που θα επιτρέπουν σε υπολογιστές να αναγνωρίζουν μοτίβα και να συμπεριφέρονται όπως ο ανθρώπινος εγκέφαλος. Τα τελευταία χρόνια η τεχνητή νοημοσύνη έχει εξελιχθεί σημαντικά, οδηγώντας στη δημιουργία δύο νέων πεδίων – υποσύνολα της τεχνητής νοημοσύνης: την μηχανική μάθηση, η οποία θεωρείται υποσύνολο της τεχνητής νοημοσύνης και χρησιμοποιεί μεθόδους στατιστικής ώστε να επιτρέπει στους υπολογιστές να βελτιώνονται συνεχώς με βάση την εμπειρία τους, και την βαθιά μάθηση, υποσύνολο της μηχανικής μάθησης, η οποία στοχεύει στην ανάπτυξη μοντέλων που αναπαριστούν τις ικανότητες του ανθρώπινου εγκεφάλου στο να λύνει προβλήματα, χάρη σε τεχνητά νευρωνικά δίκτυα καθώς και σε τεχνικές προσομοίωσης μάθησης.



Εικόνα 1: Η τεχνητή νοημοσύνη και τα πεδία που εμπίπτουν σε αυτή

Η βαθιά μάθηση μέχρι τώρα έχει επιδείξει την ικανότητά της στο να αναγνωρίζει εικόνες και να επικοινωνεί με φυσικότητα, ενώ προσπάθειες για την ανάπτυξη περισσότερων ικανοτήτων συνεχίζονται, ιδιαίτερα χάρη σε εταιρίες-κολοσσούς που επενδύουν σημαντικά κεφάλαια για το σκοπό αυτό.[3]

Υπάρχουν δύο γενικότερα είδη της τεχνητής νοημοσύνης, τα οποία προκύπτουν από εύρος γνώσεων του συστήματος: Η εφαρμοσμένη τεχνητή νοημοσύνη και η γενικευμένη. Η εφαρμοσμένη τεχνητή νοημοσύνη είναι η πιο συνηθισμένη από τις δύο, και αφορά ευφυή συστήματα ικανά για ένα συγκεκριμένο έργο, όπως για παράδειγμα τα συστήματα οδήγησης οχημάτων χωρίς παρεμβολή οδηγού. Παράδειγμα συστήματος εφαρμοσμένης τεχνητής νοημοσύνης αποτελεί το αυτόνομο όχημα Waymo, το οποίο έχει εκπαιδευτεί

τόσο σε πραγματικές οδηγικές συνθήκες όσο και σε αναπαραστάσεις πραγματικών ατυχημάτων, κάνοντας χρήση τεχνικών προσομοίωσης μάθησης [4]. Στον αντίποδα, η γενικευμένη νοημοσύνη συναντάται πιο σπάνια, και όπως περιγράφει το όνομά της είναι σε θέση να αντιμετωπίζει προβλήματα ή απρόβλεπτες καταστάσεις πρωτόγνωρες για αυτό. Ωστόσο στην παρούσα φάση δεν υπάρχουν συστήματα γενικευμένης νοημοσύνης διαθέσιμα, λόγω της πρωιμότητας του συγκεκριμένου πεδίου.

2.1 Μηχανική μάθηση

Η μηχανική μάθηση, σε συνδυασμό με τα big data ή μέγα-δεδομένα αποτελεί μια ανερχόμενη επιστήμη που χρησιμοποιείται σε πληθώρα εφαρμογών και τομέων. Βασίζεται στην ύπαρξη μεγάλου συνόλου δεδομένων για μια εφαρμογή ώστε να αντληθούν μοτίβα και συσχετίσεις μεταξύ των δεδομένων, ώστε ένα σύστημα να είναι σε θέση να παράγει γνώση και να μπορεί να ανταποκρίνεται κατάλληλα σε νέα άγνωστα δεδομένα. Οι αλγόριθμοι μηχανικής μάθησης χωρίζονται σε τρεις κύριες κατηγορίες ανάλογα με τον ρόλο τους σε ένα σύστημα τεχνητής νοημοσύνης:

- Σε αλγορίθμους λήψης αποφάσεων (decision process). Τέτοιοι αλγόριθμοι χρησιμοποιούν τεχνικές μηχανικής μάθησης για να κατηγοριοποιήσουν επισημασμένα ή μη δεδομένα με απώτερο σκοπό τον εντοπισμό μοτίβων στα δεδομένα αυτά ώστε να βοηθήσουν στη λήψη αποφάσεων.
- Σε αλγορίθμους συνάρτησης σφάλματος (error function), οι οποίοι είναι υπεύθυνοι για την αξιολόγηση της ακρίβειας των προβλέψεων των μοντέλων και
- Σε αλγορίθμους βελτιστοποίησης μοντέλων (model optimization) οι οποίοι εφόσον εντοπιστεί το κατάλληλο για τα δεδομένα μοντέλο γίνεται αναπροσαρμογή στα βάρη, δηλαδή στις ξεχωριστές παραμέτρους που συντελούν στην δημιουργία των προβλέψεων ώστε να επιτευχθεί η μέγιστη δυνατή ακρίβεια.

Στη μηχανική μάθηση υπάρχουν τρεις μέθοδοι εκπαίδευσης των μοντέλων:

- Εποπτευόμενη (supervised) μάθηση, στην οποία η εκπαίδευση γίνεται με τη μελέτη δειγμάτων που παρέχονται σε σύνολο δεδομένων. Σκοπός εδώ είναι να εντοπιστεί η συσχέτιση ανάμεσα στα δείγματα και το γνωστό αποτέλεσμα, ώστε τελικά να εφαρμοστούν οι παρατηρήσεις αυτές σε άγνωστα δεδομένα και να μπορεί ο αλγόριθμος να προβλέπει το αποτέλεσμα που προκύπτει. Παραδείγματα αποτελούν οι αλγόριθμοι regression (παλινδρόμησης) ή κατηγοριοποίησης (classification). Σε αλγορίθμους εποπτευόμενης μάθησης παρέχονται τα δεδομένα εκπαίδευσης καθώς και μια σειρά από επιθυμητά αποτελέσματα, και στόχος του

αλγορίθμου γίνεται να μάθει πως να αναπαράξει τα αποτελέσματα αυτά με βάση τα δεδομένα που έχει.[3] [5]

- Μη-εποπτευόμενη (unsupervised) μάθηση, κατά την οποία ο στόχος είναι απλώς η εξαγωγή μοτίβων από τα σύνολα δεδομένων. Αυτή η μέθοδος είναι πιο κοντά στο πως λειτουργεί ο ανθρώπινος εγκέφαλος. Με άλλα λόγια, παρέχονται απλώς μοτίβα x_i τα οποία θεωρούμε ως δείγματα από μια άγνωστη κατανομή πιθανοτήτων $p_{i[x]}$ και ένα πολύ μικρό σύνολο πληροφοριών ως προς τα δείγματα στα οποία πρέπει να δοθεί περισσότερο βάρος. Παραδείγματα αλγορίθμων που εμπίπτουν στην μη-εποπτευόμενη εκπαίδευση είναι το clustering. [5]
- Ενισχυτική μάθηση (reinforcement learning). Στην ενισχυτική μάθηση εμπίπτει και η θεωρία παιγνίων. Η μέθοδος είναι παρόμοια με την εποπτευόμενη ως προς το γεγονός ότι παρέχεται ένα περιβάλλον εντός του οποίου περιέχονται επιβραβεύσεις ή τιμωρίες για τον αλγόριθμο, και στόχος του είναι ανάλογα με το πρόβλημα να μεγιστοποιήσει τις ανταμοιβές του ή να μειώσει τις ποινές του κατά την διάρκεια ζωής του.[5]

Φυσικά είναι εύλογο να αναρωτηθεί κανείς αν πράγματι η μηχανική μάθηση διαφέρει αρκετά από την τεχνητή νοημοσύνη ώστε να θεωρηθεί ξεχωριστό από εκείνη πεδίο. Αφενός υπάρχει διαφορά στον σκοπό που υπηρετεί το κάθε πεδίο. Ο ρόλος της τεχνητής νοημοσύνης είναι να βελτιώνει το ποσοστό επιτυχίας για ένα έργο που του ανατίθεται. Αφορά έναν υπολογιστή ικανό να προσομοιώνει την ανθρώπινη ευφυΐα για να επιλύσει περίπλοκα προβλήματα, αναζητώντας την ιδανικότερη λύση μέσα από μια σειρά αποφάσεων που καλείται να λάβει. Στον αντίποδα, στόχος της μηχανικής μάθησης είναι η υψηλή ακρίβεια στις προβλέψεις, η μάθηση μέσα από δεδομένα που έχουν συλλεχθεί, και η σύνθεση ή αναπαράσταση γνώσης.

2.2 Βαθιά Μάθηση

Η βαθιά μάθηση αποτελείται από ένα μεγάλο σύνολο μεθόδων που περιλαμβάνουν νευρωνικά δίκτυα, ιεραρχικά πιθανοκρατικά μοντέλα και πληθώρα εποπτευόμενων και μη μεθόδων εκπαίδευσης. Η ανάγκη για τη δημιουργία συστημάτων που μπορούν να λειτουργούν όπως ο ανθρώπινος εγκέφαλος ξεκίνησαν την ανάπτυξη των νευρωνικών δικτύων. Μια σειρά συνεισφορών όπως η ανάπτυξη νευρωνικού δικτύου με μακροχρόνια βραχυπρόθεσμη μνήμη (long short-term memory) σε συνδυασμό με την εμφάνιση πληθώρας επισημασμένων συνόλων δεδομένων υψηλής ποιότητας καθώς και την ικανότητα παράλληλου υπολογισμού με τη χρήση μονάδων υπολογισμού γραφικών οδήγησαν στην βελτίωση ταχύτητας και ακρίβειας στην εκπαίδευση τέτοιων μοντέλων

και άρα στην προώθηση τεχνικών βαθιάς μάθησης. Άλλοι παράγοντες που έπαιξαν ρόλο ήταν τεχνικές επαύξησης δεδομένων. Σε περιπτώσεις όπου το σύνολο δεδομένων είναι εικόνες, τεχνικές επαύξησης δεδομένων αποτελούν μέθοδοι επεξεργασίας εικόνας, όπως περιστροφή, αναστροφή, περικοπή, εισαγωγή θορύβου και άλλες αντίστοιχες μέθοδοι που αποσκοπούν στο να εμπλουτίσουν τα ερεθίσματα στα οποία βασίζονται τα μοντέλα για την δημιουργία ασφαλών προβλέψεων υπό διάφορες συνθήκες που ίσως να μην μπορούσαν να καλυφθούν από το μοντέλο χωρίς την επεξεργασία του συνόλου δεδομένων.[6] Φυσικά υπάρχουν και άλλες τεχνικές επαύξησης δεδομένων που δε βασίζονται μόνο στην επεξεργασία του συνόλου δεδομένων. Η μεταβιβαστική μάθηση (transfer learning) βασίζεται σε μοντέλα που έχουν εκπαιδευτεί σε ένα σύνολο δεδομένων x για μια εργασία, και τα βάρη που έχουν παραχθεί από αυτό το μοντέλο μεταφέρονται σε καινούριο μοντέλο ώστε να εκπαιδευτούν πάνω σε μια παρόμοια εργασία με τη χρήση ενός υποσυνόλου του x , κάτι που βελτιώνει την ταχύτητα και ακρίβεια μάθησης στο καινούριο μοντέλο.

3 Διαδίκτυο των Πραγμάτων

Το Διαδίκτυο των Πραγμάτων (IoT) είναι ένα σημαντικό θέμα στη βιομηχανία της τεχνολογίας, την πολιτική και τον τομέα της μηχανικής, καθώς έχει τη δυνατότητα να μεταμορφώσει διάφορες πτυχές της ζωής μας. Οι συσκευές IoT, όπως οι συσκευές με δυνατότητα σύνδεσης στο Διαδίκτυο, τα εξαρτήματα οικιακού αυτοματισμού και οι συσκευές διαχείρισης ενέργειας, μας οδηγούν προς ένα "έξυπνο σπίτι" με αυξημένη ασφάλεια και ενεργειακή απόδοση. Οι προσωπικές συσκευές IoT, όπως οι φορητές συσκευές παρακολούθησης της φυσικής κατάστασης και της υγείας και οι ιατρικές συσκευές με δυνατότητα δικτύωσης, μεταμορφώνουν τις υπηρεσίες υγειονομικής περίθαλψης. Τα συστήματα IoT, όπως τα δικτυωμένα οχήματα, τα έξυπνα συστήματα κυκλοφορίας και οι αισθητήρες που είναι ενσωματωμένοι σε δρόμους και γέφυρες, μας φέρνουν πιο κοντά στις "έξυπνες πόλεις" με μειωμένη κυκλοφοριακή συμφόρηση και κατανάλωση ενέργειας. Η τεχνολογία IoT προσφέρει επίσης τη δυνατότητα να μετασχηματίζει τη γεωργία, τη βιομηχανία και την παραγωγή και διανομή ενέργειας, αυξάνοντας τη διαθεσιμότητα πληροφοριών κατά μήκος της αλυσίδας αξίας με τη χρήση δικτυωμένων αισθητήρων. Ωστόσο, το IoT εγείρει πολυάριθμα ζητήματα και προκλήσεις που πρέπει να εξεταστούν και να αντιμετωπιστούν για να αξιοποιηθούν τα δυνητικά οφέλη.

Το Διαδίκτυο των Πραγμάτων (IoT) είναι μια ταχέως αναπτυσσόμενη τεχνολογία που έχει τη δυνατότητα να επηρεάσει σημαντικά το Διαδίκτυο και την οικονομία τα επόμενα πέντε έως δέκα χρόνια. Εταιρείες και ερευνητικοί οργανισμοί έχουν προβλέψει διάφορους αριθμούς για τον αντίκτυπο του IoT στο Διαδίκτυο και την οικονομία. Η Huawei προβλέπει 100 δισεκατομμύρια συνδέσεις IoT μέχρι το 2025, ενώ το McKinsey Global Institute προτείνει ότι ο οικονομικός αντίκτυπος του IoT στην παγκόσμια οικονομία μπορεί να είναι μεταξύ 3,9 και 11,1 τρισεκατομμυρίων δολαρίων μέχρι το 2025. Ενώ ορισμένοι θεωρούν το IoT ως έναν επαναστατικό, πλήρως διασυνδεδεμένο κόσμο προόδου, αποτελεσματικότητας και ευκαιριών, άλλοι προειδοποιούν για την επιτήρηση, την προστασία της ιδιωτικής ζωής, τις παραβιάσεις της ασφάλειας και τον εγκλωβισμό των καταναλωτών.

Η συζήτηση "υπόσχεση εναντίον κινδύνου" και η εισροή πληροφοριών μέσω των δημοφιλών μέσων ενημέρωσης και του μάρκετινγκ καθιστούν το IoT ένα πολύπλοκο θέμα προς κατανόηση. Η Εταιρεία Διαδικτύου ανησυχεί για το IoT, καθώς αντιπροσωπεύει μια αυξανόμενη πτυχή του τρόπου με τον οποίο οι άνθρωποι και οι θεσμοί αλληλεπιδρούν με το Διαδίκτυο στην προσωπική, κοινωνική και οικονομική ζωή τους.

Εάν οι μετριοπαθείς προβλέψεις είναι σωστές, η έκρηξη των εφαρμογών του IoT θα μπορούσε να παρουσιάσει μια θεμελιώδη αλλαγή στην εμπλοκή των χρηστών με το Διαδίκτυο και στον αντίκτυπό του, εγείροντας νέα ζητήματα και διαστάσεις των υφιστάμενων προκλήσεων σε θέματα που αφορούν τους χρήστες/καταναλωτές, την τεχνολογία, την πολιτική και το δίκαιο. Το IoT θα έχει πιθανότατα διαφορετικές συνέπειες σε διαφορετικές οικονομίες και περιοχές, δημιουργώντας ένα ευρύ φάσμα ευκαιριών και προκλήσεων σε όλο τον κόσμο.

Ο όρος "Διαδίκτυο των πραγμάτων" (Internet of Things, IoT) εισήχθη για πρώτη φορά το 1999 από τον Βρετανό πρωτοπόρο της τεχνολογίας Kevin Ashton για να περιγράψει ένα σύστημα όπου τα αντικείμενα του φυσικού κόσμου θα μπορούσαν να συνδεθούν στο Διαδίκτυο μέσω αισθητήρων. Ο όρος χρησιμοποιήθηκε αρχικά για να απεικονίσει τη δύναμη της σύνδεσης ετικετών ραδιοσυχνοτικής αναγνώρισης (RFID) σε εταιρικές αλυσίδες εφοδιασμού με το Διαδίκτυο για την καταμέτρηση και τον εντοπισμό αγαθών χωρίς ανθρώπινη παρέμβαση. Σήμερα, το IoT περιλαμβάνει σενάρια όπου η συνδεσιμότητα στο Διαδίκτυο και η υπολογιστική ικανότητα επεκτείνονται σε διάφορα αντικείμενα, συσκευές, αισθητήρες και αντικείμενα καθημερινής χρήσης. Αν και ο όρος "IoT" είναι σχετικά νέος, η ιδέα του συνδυασμού υπολογιστών και δικτύων για την παρακολούθηση και τον έλεγχο συσκευών υπάρχει εδώ και δεκαετίες. Στα τέλη της δεκαετίας του 1970, συστήματα για την εξ αποστάσεως παρακολούθηση μετρητών στο ηλεκτρικό δίκτυο μέσω τηλεφωνικών γραμμών βρισκόνταν ήδη σε εμπορική χρήση.

Οι εξελίξεις στην ασύρματη τεχνολογία επέτρεψαν την ευρεία διάδοση των επιχειρηματικών και βιομηχανικών λύσεων "μηχανή προς μηχανή" (M2M) για την παρακολούθηση και τη λειτουργία του εξοπλισμού. Ωστόσο, πολλές από τις πρώτες λύσεις M2M βασίζονταν σε κλειστά δίκτυα ειδικού σκοπού και σε ιδιόκτητα πρότυπα και όχι σε δίκτυα βασισμένα στο πρωτόκολλο Διαδικτύου (IP) και σε πρότυπα Διαδικτύου. Η πρώτη "συσκευή" Διαδικτύου ήταν μια τοστιέρα με δυνατότητα IP, η οποία μπορούσε να ενεργοποιηθεί και να απενεργοποιηθεί μέσω του Διαδικτύου. Η ανάπτυξη της "δικτύωσης έξυπνων αντικειμένων" συνέβαλε στη δημιουργία των θεμελίων για το σημερινό IoT.

Το Διαδίκτυο των Πραγμάτων (IoT) είναι μια σύγκλιση διαφόρων τεχνολογικών τάσεων και τάσεων της αγοράς που επιτρέπουν τη διασύνδεση περισσότερων και μικρότερων συσκευών με φθινό και εύκολο τρόπο. Η πανταχού παρούσα συνδεσιμότητα, η ευρεία υιοθέτηση της δικτύωσης βάσει IP, η οικονομία της πληροφορικής που καθοδηγείται από τις επενδύσεις της βιομηχανίας, η μικρογραφία, οι εξελίξεις στην ανάλυση δεδομένων και

η υπολογιστική νέφους, έχουν καταστήσει δυνατή τη σύνδεση περισσότερων και μικρότερων συσκευών φθηνά και εύκολα. Αυτές οι τάσεις έχουν οδηγήσει στη συγκέντρωση, συσχέτιση και ανάλυση τεράστιων ποσοτήτων δεδομένων, παρέχοντας νέες ευκαιρίες για την εξαγωγή πληροφοριών και γνώσεων. Η άνοδος του υπολογιστικού νέφους, το οποίο αξιοποιεί απομακρυσμένους, δικτυωμένους υπολογιστικούς πόρους, επιτρέπει σε μικρές και κατανεμημένες συσκευές να αλληλεπιδρούν με ισχυρές δυνατότητες ανάλυσης και ελέγχου στο back-end. Το IoT αντιπροσωπεύει τη σύγκλιση διαφόρων τάσεων υπολογισμού και συνδεσιμότητας και διάφοροι βιομηχανικοί τομείς εξετάζουν τώρα τις δυνατότητες ενσωμάτωσης της τεχνολογίας IoT στα προϊόντα, τις υπηρεσίες και τις λειτουργίες τους.[1] Το IoT μπορεί να υλοποιηθεί μόνο με την ανάπτυξη πολλαπλών τεχνολογιών που καλύπτουν τον τομέα του υλικού, του λογισμικού και των εξαιρετικά ισχυρών εφαρμογών γύρω από κάθε τομέα της βιομηχανίας και του τομέα λειτουργίας. Σε αυτό το πλαίσιο, η παρούσα ενότητα θα παρουσιάσει τους τεχνολογικούς τομείς που επιτρέπουν το IoT και θα προσδιορίσει τις προκλήσεις έρευνας και ανάπτυξης και θα περιγράψει ένα σχέδιο για μελλοντικές ερευνητικές δραστηριότητες για την παροχή πρακτικών και αξιόπιστων λύσεων. Ορισμένοι από τους βασικούς τεχνολογικούς τομείς που θα επιτρέψουν το IoT είναι οι εξής: (i) τεχνολογία αναγνώρισης, (ii) τεχνολογία αρχιτεκτονικής IoT, (iii) τεχνολογία επικοινωνίας, (iv) τεχνολογία δικτύου, (v) τεχνολογία ανακάλυψης δικτύου, (vi) λογισμικά και αλγόριθμοι, (vii) τεχνολογία υλικού, (viii) τεχνολογία επεξεργασίας δεδομένων και σήματος, (ix) τεχνολογία ανακάλυψης και μηχανής αναζήτησης, (x) τεχνολογία διαχείρισης δικτύου σχέσεων, (xi) τεχνολογία αποθήκευσης ισχύος και ενέργειας, (xii) τεχνολογίες ασφάλειας και προστασίας της ιδιωτικής ζωής και (xiii) τυποποίηση. Αυτοί οι βασικοί τεχνολογικοί παράγοντες εξετάζονται συνοπτικά στις ακόλουθες υποενότητες:

- Τεχνολογία Αναγνώρισης:

Η τεχνολογία ταυτοποίησης αντιστοιχίζει ένα μοναδικό αναγνωριστικό (UID) σε μια οντότητα, καθιστώντας την ανακτήσιμη και αναγνωρίσιμη χωρίς ασάφειες. Τα UID μπορούν να δημιουργηθούν ως ενιαία ποσότητα ή ως συλλογή χαρακτηριστικών. Στο όραμα του IoT, τα πράγματα έχουν ψηφιακή ταυτότητα, με ψηφιακά ονόματα και σχέσεις που καθορίζονται στον ψηφιακό τομέα. Η ανάπτυξη του IoT απαιτεί νέες τεχνολογίες για την αντιμετώπιση παγκόσμιων συστημάτων ταυτότητας, τη διαχείριση ταυτότητας, την κωδικοποίηση/κρυπτογράφηση ταυτότητας, τον έλεγχο ταυτότητας και τη

Ενδεικτικά περιβάλλοντα χρήσης IoT		
Περιβάλλον	Περιγραφή	Παραδείγματα
Άνθρωπος	Συσκευές προσαρτημένες ή μέσα στο ανθρώπινο σώμα	Συσκευές (φορητές ή καταπιώμενες συσκευές) παρακολούθησης και διατήρησης ανθρώπινης υγείας και ευεξίας, διαχείριση ασθενειών, βελτίωση φυσικής κατάστασης και παραγωγικότητας)
Σπίτι	Κτήρια όπου ζούν άνθρωποι	Οικιακά συστήματα ασφαλείας
Καταστήματα Λιανικής	Χώροι ενασχόλησης καταναλωτών με το εμπόριο	εστιατόρια, γενικώς χώροι που επισκέπτονται καταναλωτές - self-checkout, βελτιστοποίηση αποθέματος
Γραφεία	Χώροι ενασχόλησης εργαζομένων στη γνώση	Διαχείριση ενέργειας και ασφάλειας γραφείων, βελτιωμένη παραγωγικότητα ακόμα και για εν κινήσει υπαλλήλους
Εργοστάσια	Τυποποιημένα περιβάλλοντα παραγωγής	Χώροι με επαναλαμβανόμενες ρουτίνες εργασίας, συμπεριλαμβανομένων νοσοκομείων και αγροκτημάτων. λειτουργική αποδοτικότητα, βελτιστοποίηση της χρήσης εξοπλισμού και του αποθέματος
Εργοτάξια	Εξατομικευμένα περιβάλλοντα παραγωγής	Εξόρυξη, πετρέλαιο και φυσικό αέριο, κατασκευές. λειτουργική αποδοτικότητα, προληπτική συντήρηση, υγεία και ασφάλεια
Οχήματα	Συστήματα μέσα σε κινούμενα οχήματα	Οχήματα, συμπεριλαμβανομένων αυτοκινήτων, φορτηγών, πλοίων, αεροσκαφών και τρένων. Συντήρηση βάσει κατάστασης, σχεδιασμός βάσει χρήσης, ανάλυση πριν από την πώληση
Πόλεις	Αστικά περιβάλλοντα	Δημόσιοι χώροι και υποδομές σε αστικά περιβάλλοντα. προσαρμοστικός έλεγχος κυκλοφορίας, έξυπνοι μετρητές, περιβαλλοντική παρακολούθηση, διαχείριση πόρων

Εικόνα 2: Ενδεικτικά περιβάλλοντα χρήσης IoT

διαχείριση αποθετηρίων. Επιπλέον, δημιουργούνται παγκόσμιες υπηρεσίες αναζήτησης καταλόγου και ανακάλυψης για εφαρμογές IoT με διάφορα σχήματα μοναδικών αναγνωριστικών.

- Τεχνολογία αρχιτεκτονικής IoT:

Η προσέγγιση της αρχιτεκτονικής προσανατολισμένης στις υπηρεσίες (SOA) χρησιμοποιείται συχνά σε αρχιτεκτονικές ενδιάμεσου λογισμικού για το IoT, επιτρέποντας τη διάσπαση πολύπλοκων συστημάτων σε εφαρμογές με απλούστερα, σαφώς καθορισμένα στοιχεία. Η προσέγγιση αυτή επιτρέπει την

επαναχρησιμοποίηση λογισμικού και υλικού χωρίς την επιβολή συγκεκριμένης τεχνολογίας για την υλοποίηση υπηρεσιών. Η σημασιολογική διαλειτουργικότητα είναι ζωτικής σημασίας για την ουσιαστική επικοινωνία μεταξύ παρόχων και αιτούντων, παρά την ετερογενή φύση των δομών πληροφοριών και των επιχειρηματικών αντικειμένων. Η σημασιολογική διαλειτουργικότητα μπορεί να επιτευχθεί μεταξύ ετερογενών πληροφοριακών συστημάτων μέσω ολοκληρωμένων κοινών μοντέλων πληροφοριών, τα οποία μπορούν να διευκολύνουν τη σημασιολογική διαλειτουργικότητα μεταξύ των συμμετεχουσών εφαρμογών και επιχειρήσεων. Ωστόσο, η προσέγγιση αυτή είναι άκαμπτη και ανελαστική, με αποτέλεσμα να μην είναι ευέλικτη. Εναλλακτικά, η σημασιολογική διαλειτουργικότητα μπορεί να επιτευχθεί με την παροχή κατάλληλων σημασιολογικών διαμεσολαβητών στο τέλος κάθε συμμετέχοντα για τη διευκόλυνση της μετατροπής σε μορφή πληροφορίας που κατανοεί ο συμμετέχων. Τα περισσότερα συστήματα χρησιμοποιούν έναν συνδυασμό από κοινά μοντέλα πληροφοριών που δεν εξαρτώνται από το πλαίσιο και προσεγγίσεις εξειδίκευσης πληροφοριών που αφορούν το πλαίσιο για την επίτευξη σημασιολογικής διαλειτουργικότητας. Οι βασικές απαιτήσεις σχεδιασμού για το IoT περιλαμβάνουν την επεκτασιμότητα, την αρθρωτότητα, την επεκτασιμότητα και τη διαλειτουργικότητα μεταξύ ετερογενών πραγμάτων και των περιβαλλόντων τους. Οι επαγγελματίες της βιομηχανίας έχουν προτείνει την αξιοποίηση των εργασιών στον σημασιολογικό ιστό για την ανάπτυξη ολοκληρωμένων και ανοικτών οντολογιών για την αντιμετώπιση της σημασιολογικής διαλειτουργικότητας στη δυναμική SOA με βάση τη δέσμευση για το σχεδιασμό και την ανάπτυξη εφαρμογών IoT.

- Τεχνολογίες Επικοινωνίας:

Οι εφαρμογές του IoT αποτελούν έναν εκτεταμένο χώρο σχεδιασμού με πολλές διαστάσεις που περιλαμβάνουν διάφορα ζητήματα και παραμέτρους, μερικές από τις οποίες αναφέρονται παρακάτω.

- Εγκατάσταση

- Κινητικότητα - περιστασιακή ή συνεχής που εκτελείται είτε από επιλεγμένα είτε από όλα τα πράγματα στο επιλεγμένο περιβάλλον.

- Κόστος, μέγεθος, πόροι και ενέργεια - πολύ περιορισμένοι πόροι ή απεριόριστοι πόροι.

- Ετερογένεια - ένας μόνο τύπος πράγματος ή ποικίλα σύνολα διαφορετικών ιδιοτήτων και ιεραρχίες.
- Τρόπος επικοινωνίας - επικοινωνία ενός ή πολλών βημάτων (hops).
- Υποδομή - διαφορετικές εφαρμογές αποκλείουν, επιτρέπουν ή απαιτούν τη χρήση σταθερών υποδομών.
- Τοπολογία δικτύου - μονόδρομος, αστέρας, πλέγμα ή πολλαπλών επιπέδων.
- Κάλυψη - αραιή, πυκνή ή πλεονάζουσα.
- Συνδεσιμότητα - συνεχής, περιστασιακή ή σποραδική.
- Μέγεθος δικτύου - από δεκάδες κόμβους έως χιλιάδες.
- Διάρκεια ζωής - λίγες ώρες, αρκετοί μήνες έως πολλά χρόνια.
- Απαιτήσεις QoS - περιορισμοί πραγματικού χρόνου, ανθεκτικότητα στην παραβίαση, διακριτικότητα κ.λπ.

Ένας τόσο εκτεταμένος χώρος σχεδιασμού καθιστά προφανώς την ανάπτυξη εφαρμογών IoT μια περίπλοκη διαδικασία. Μια προσέγγιση μπορεί να είναι να γίνει ο σχεδιασμός για το πιο περιοριστικό σημείο του χώρου σχεδιασμού, π.χ. ελάχιστες δυνατότητες πραγμάτων, υψηλή κινητικότητα κ.λπ. Ωστόσο, συχνά δεν υπάρχει τέτοιο παγκόσμιο ελάχιστο και μπορεί να είναι επιθυμητό να αξιοποιηθούν τα χαρακτηριστικά των διαφόρων σημείων του χώρου σχεδιασμού. Αυτό συνεπάγεται ότι καμία ενιαία πλατφόρμα υλικού και λογισμικού δεν θα να είναι επαρκής για την υποστήριξη ολόκληρου του σχεδιαστικού χώρου. Τα πολύπλοκα και ετερογενή συστήματα θα είναι φυσική απαίτηση.

- Τεχνολογίες Δικτύωσης:

Η ανάπτυξη του IoT απαιτεί την ανάπτυξη κατάλληλης τεχνολογίας δικτύου για τη σύνδεση αντικειμένων στον φυσικό κόσμο και τη σύνδεσή τους στο Διαδίκτυο. Τεχνολογίες όπως η RFID, η ασύρματη επικοινωνία μικρής εμβέλειας και τα δίκτυα αισθητήρων χρησιμοποιούνται για τη συνδεσιμότητα δικτύου, ενώ η έκδοση 6 του πρωτοκόλλου Διαδικτύου (IPv6) επιτρέπει τη διευθυνσιοδότηση, τη σύνδεση και την παρακολούθηση. Η ασφάλεια, η επεκτασιμότητα και η συμβατότητα πολλαπλών πλατφορμών αποτελούν βασικές απαιτήσεις στο παράδειγμα του IoT. Οι τεχνολογίες δικτύων πρέπει να προσφέρουν λύσεις που μπορούν να συνδέσουν

σχεδόν οτιδήποτε στο δίκτυο με μειωμένο κόστος. Η πανταχού παρούσα πρόσβαση στο δίκτυο θα αλλάξει τον τρόπο επεξεργασίας των πληροφοριών.

Το IP παρέχει επικοινωνία από άκρο σε άκρο μεταξύ των συσκευών χωρίς την ανάγκη για μια ενδιάμεση πύλη μετάφρασης πρωτοκόλλου, καθιστώντας πολύπλοκη τη σχεδίαση, τη διαχείριση και την εγκατάσταση. Νέες κλιμακούμενες αρχιτεκτονικές σχεδιασμένες ειδικά για την επικοινωνία πανταχού παρόντων δικτύων αισθητήρων θα επιτρέψουν δίκτυα δισεκατομμυρίων συσκευών. Οι βελτιώσεις στα ασφαλή και αξιόπιστα πρωτόκολλα ασύρματης επικοινωνίας θα επιτρέψουν εφαρμογές κρίσιμης σημασίας για πανταχού παρόντα δίκτυα αισθητήρων που βασίζονται σε ασύρματες αναγνωρίσιμες συσκευές.

Οι μηχανισμοί ανακάλυψης δικτύων είναι απαραίτητοι για την αποτελεσματική διαχείριση δικτύων και επικοινωνιών στο παράδειγμα του IoT. Χωρίς έναν αυτοματοποιημένο μηχανισμό ανακάλυψης, είναι αδύνατο να επιτευχθεί μια κλιμακούμενη και ακριβής ικανότητα διαχείρισης δικτύου. Ένας αυτοματοποιημένος μηχανισμός ανακάλυψης δικτύου μπορεί να εκχωρεί δυναμικά ρόλους σε συσκευές με βάση την έξυπνη αντιστοίχιση με προκαθορισμένα πρότυπα και χαρακτηριστικά, να αναπτύσσει και να εκκινεί αυτόματα ενεργά, παθητικά ή παρακολουθήματα απόδοσης με βάση τους εκχωρημένους ρόλους και χαρακτηριστικά, να ξεκινά, να σταματά, να διαχειρίζεται και να προγραμματίζει τη διαδικασία ανακάλυψης και να πραγματοποιεί αλλαγές σε οποιοδήποτε ρόλο ή προφίλ παρακολούθησης ανά πάσα στιγμή ή να δημιουργεί νέα προφίλ ανάλογα με τις ανάγκες. Οι μηχανισμοί δυναμικής ανακάλυψης δικτύου επιτρέπουν την αλληλεπίδραση μεταξύ συσκευών που δεν είναι προ-ρυθμισμένες και σκληρά κωδικοποιημένες όσον αφορά διευθύνσεις ή τελικά σημεία υπηρεσιών. Παραδείγματα πρωτοκόλλων για την ανακάλυψη σε επίπεδο LAN περιλαμβάνουν τα WS-Discovery, Bonjour και SSDP. Σήμερα υπάρχουν τόσο παθητικοί όσο και δυναμικοί μηχανισμοί ανακάλυψης, ενώ αναπτύσσονται τεχνολογίες για την υλοποίηση δυναμικής ανακάλυψης δεδομένων δικτύου σε πραγματικό χρόνο. Όλες οι υπηρεσίες ανακάλυψης πρέπει να βασίζονται σε μηχανισμούς πιστοποίησης ταυτότητας για την αντιμετώπιση ζητημάτων ιδιωτικότητας ή ασφάλειας.

- **Λογισμικό και Αλγόριθμοι:**

Η δημιουργία εφαρμογών IoT απαιτεί το σχεδιασμό ενός κοινού υποκείμενου λογισμικού για διάφορα περιβάλλοντα και τη δημιουργία συνεκτικών εφαρμογών από διαφορετικές μονάδες. Οι προσπάθειες έρευνας και ανάπτυξης

επικεντρώνονται στην υπηρεσιοστρεφή πληροφορική, η οποία υποστηρίζει διαλειτουργική αλληλεπίδραση μεταξύ μηχανών και αντικειμένων μέσω δικτύων. Η προσέγγιση αυτή βασίζεται στα πρωτόκολλα του Διαδικτύου και ορίζει νέα πρωτόκολλα για την περιγραφή και τη διεύθυνση των περιπτώσεων υπηρεσιών, οργανώνοντας χαλαρά τις υπηρεσίες ιστού σε εικονικά δίκτυα.

- Υλικό:

Η έρευνα της νανοηλεκτρονικής επικεντρώνεται στη σμίκρυνση, το χαμηλό κόστος και την αυξημένη λειτουργικότητα κατά το σχεδιασμό ασύρματων αναγνωρίσιμων συστημάτων. Η τεχνολογία ολοκληρωμένων κυκλωμάτων πυριτίου χρησιμοποιείται για ενισχυμένη μη πτητική μνήμη για την ανίχνευση και την παρακολούθηση παραμέτρων περιβάλλοντος. Απαιτείται περαιτέρω έρευνα σε σχέδια εξαιρετικά χαμηλής ισχύος, χαμηλής τάσης και χαμηλής διαρροής σε τεχνολογίες CMOS υπομικρών RF, λύσεις διαχείρισης ισχύος DC-DC υψηλής απόδοσης, ελεγχόμενη μη πτητική μνήμη εξαιρετικά χαμηλής ισχύος, χαμηλής τάσης και ενσωμάτωση συσκευών RF MEMS και MEMS. Η έμφαση θα δοθεί σε ιδιαίτερα μικροσκοπικά ολοκληρωμένα κυκλώματα, συμπεριλαμβανομένων των πολλαπλών RF, των προσαρμοστικών και αναδιαμορφώσιμων front ends, της μνήμης, του Διαδικτύου των Πραγμάτων (IoT), των πολλαπλών πρωτοκόλλων επικοινωνίας, της ψηφιακής επεξεργασίας και της ασφάλειας. Το IoT θα δημιουργήσει νέες υπηρεσίες και επιχειρηματικές ευκαιρίες για τους παρόχους συστημάτων για την εξυπηρέτηση των απαιτήσεων επικοινωνίας δεκάδων δισεκατομμυρίων συσκευών στο μέλλον.

Οι ετικέτες RFID χρησιμοποιούνται με διάφορους τρόπους, με ετικέτες εξαιρετικά χαμηλού κόστους που έχουν περιορισμένα χαρακτηριστικά και κεντρικές λειτουργίες διαχείρισης δεδομένων. Χρησιμοποιούνται επίσης έξυπνες σταθερές ή κινητές ετικέτες και ενσωματωμένα συστήματα, που προσφέρουν περισσότερες λειτουργίες και τοπικές υπηρεσίες. Οι έξυπνες συσκευές με βελτιωμένη επικοινωνία μεταξύ των συσκευών θα οδηγήσουν σε έξυπνα συστήματα με υψηλό βαθμό ευφυΐας και αυτονομίας, επιτρέποντας την ταχεία ανάπτυξη εφαρμογών IoT και τη δημιουργία νέων υπηρεσιών.

- Δεδομένα και Τεχνολογίες Επεξεργασίας Σήματος:

Βιομηχανικοί φορείς σε διάφορους τομείς έχουν υιοθετήσει την XML ως τη βασική γλώσσα για την τυποποίηση επιχειρηματικών αντικειμένων. Κάθε βιομηχανική

κατηγορία έχει αναπτύξει πρότυπα XML για την έκφραση συμβάσεων, εμπιστοσύνης, διαδικασιών, ροών εργασίας, μηνυμάτων και άλλων σημασιολογικών στοιχείων δεδομένων σε όρους κόμβων και χαρακτηριστικών XML. Αυτά τα λεξιλόγια XML δημοσιεύονται ως γενικευμένοι ορισμοί τύπων εγγράφων (DTD) ή σχήματα XML, εξασφαλίζοντας τη σημασιολογική διαλειτουργικότητα. Πρωτοβουλίες όπως το ISO/IEC 11179 και το Universal Data Element Framework (UDEF) υποστηρίζουν τη σημασιολογική διαλειτουργικότητα μεταξύ δομημένων δεδομένων και λεξικών δεδομένων. Τα πρότυπα που βασίζονται στον σημασιολογικό ιστό από το W3C, όπως τα DAML, RDF και OWL, παρέχουν σημασιολογικά θεμέλια για δυναμικές καταστάσεις που αφορούν την ανακάλυψη επιχειρήσεων και υπηρεσιών. Οι ευφυείς αλγόριθμοι λήψης αποφάσεων πρέπει να ενεργοποιούν δραστηριότητες με βάση τη συσχέτιση μεταξύ γεγονότων, η οποία μπορεί να απαιτεί μετασχηματισμό των ακατέργαστων δεδομένων αισθητήρων. Οι κατάλληλες εργαλειοθήκες και τα πλαίσια, όπως το ESPER και το DROOLS, μπορούν να βοηθήσουν στη διαμόρφωση κανόνων αναγνώσιμων από μηχανήματα για τον καθορισμό ακολουθιών πυροδότησης γεγονότων για συγκεκριμένες δραστηριότητες ή διαδικασίες.

- Τεχνολογίες ανακάλυψης και μηχανών αναζήτησης:

Στο παράδειγμα του IoT, οι πληροφορίες και οι υπηρεσίες σχετικά με τα πράγματα θα είναι κατανεμημένες σε πολλές οντότητες και μπορεί να παρέχονται σε επίπεδο κλάσης (δηλαδή, κοινές πληροφορίες και υπηρεσίες για όλες τις περιπτώσεις πραγμάτων που έχουν την ίδια κλάση) ή σε σειριακό επίπεδο, καθώς και να παρέχονται έγκυρα από τον δημιουργό του πράγματος ή να συνεισφέρονται από άλλες οντότητες, όπως εκείνες που έχουν αλληλεπιδράσει με ένα μεμονωμένο πράγμα στο παρελθόν. Το IoT θα απαιτήσει επίσης την ανάπτυξη υπηρεσιών αναζήτησης ή παραπομπής για τη σύνδεση των πραγμάτων με πληροφορίες και υπηρεσίες και για την υποστήριξη της ασφαλούς πρόσβασης σε πληροφορίες και υπηρεσίες με τρόπο που ικανοποιεί τόσο την ιδιωτικότητα των ατόμων όσο και την εμπιστευτικότητα των επιχειρηματικών πληροφοριών. Τέτοιες αντιστοιχίσεις μεταξύ αιτούντων και παρόχων υπηρεσιών πληροφοριών μπορούν να βασίζονται σε σχέσεις εμπιστοσύνης. Καθώς ένα έξυπνο πράγμα κινείται στον πραγματικό κόσμο, θα συναντήσει νέα περιβάλλοντα και τόσο το έξυπνο πράγμα όσο και άλλα συστήματα που το παρακολουθούν θα χρειάζονται μηχανισμούς αναζήτησης προκειμένου να ανακαλύψουν ποιες δυνατότητες είναι διαθέσιμες στο τοπικό

περιβάλλον του πράγματος. Τέτοιες δυνατότητες μπορεί να είναι η διαθεσιμότητα αισθητήρων, διεπαφές επικοινωνίας δικτύου, εγκαταστάσεις υπολογισμού και επεξεργασίας δεδομένων σε πληροφορίες, καθώς και εγκαταστάσεις για περαιτέρω μεταφορά, χειρισμό, φυσική επεξεργασία ή ειδοποίηση ανθρώπινου χειριστή για προβλήματα.

- **Τεχνολογία Διαχείρισης Δικτύου Σχέσεων.**

Το Internet of Things θα απαιτήσει τη διαχείριση δικτύων που περιέχουν δισεκατομμύρια ετερογενή πράγματα και όπου υπάρχει μεγάλη ποικιλία συσκευών λογισμικού, ενδιάμεσου λογισμικού και υλικού. Οι τεχνολογίες διαχείρισης δικτύων θα πρέπει να αντιμετωπίζουν διάφορα σημαντικά ζητήματα, όπως ασφάλεια, απόδοση και αξιοπιστία. Η διαχείριση δικτύου περιλαμβάνει τη διαχείριση κατανεμημένων βάσεων δεδομένων, αποθετηρίων, αυτόματης ψηφοφορίας των συσκευών δικτύου και γραφικές προβολές σε πραγματικό χρόνο των αλλαγών της τοπολογίας του δικτύου και της κίνησης. Η υπηρεσία διαχείρισης δικτύου χρησιμοποιεί μια ποικιλία εργαλείων, εφαρμογών και συσκευών για να βοηθούν στην παρακολούθηση και τη συντήρηση των δικτύων που εμπλέκονται σε εφαρμογές IoT. Παρόμοια με την υπηρεσίες κοινωνικών δικτύων που ανθούν σήμερα στο διαδίκτυο, θα υπήρχε ανάγκη για πράγματα στο δίκτυο να δημιουργούν σχέσεις μεταξύ τους.

- **Τεχνολογίες ισχύος και αποθήκευσης ενέργειας:**

Τα αυτόνομα πράγματα που λειτουργούν στις εφαρμογές του IoT και εκτελούν παρακολούθηση των γεγονότων χρειάζονται ισχύ και ενέργεια για να εκτελέσουν την απαιτούμενη εργασία. Δεδομένου ότι τα περιβάλλοντα έχουν μεγάλες διακυμάνσεις ανάλογα με το πού και πώς χρησιμοποιείται το πράγμα, η ισχύς μέθοδοι συλλογής ενέργειας μπορεί να ποικίλλουν, π.χ. RF, ηλιακή ενέργεια, ήχος, δόνηση, θερμότητα κ.λπ. Σε καταστάσεις και τοποθεσίες όπου είναι λογικό να υπάρχει μεγάλος αριθμός πραγμάτων με δυνατότητες ανίχνευσης, η χρήση των δικτύων πλέγματος είναι μια καλή πρόταση για την αύξηση της αποδοτικότητας της επικοινωνίας και της ισχύος με τη δυνατότητα προώθησης των μεταδόσεων από το πλησιέστερο πράγμα. Ο αναγνώστης χρειάζεται τότε να βρίσκεται μόνο στην εμβέλεια της άκρης του δικτύου.

Οι τεχνολογίες αποθήκευσης ισχύος και ενέργειας αποτελούν καταλύτες για την ανάπτυξη εφαρμογών IoT. Οι τεχνολογίες αυτές πρέπει να παρέχουν λύσεις

παραγωγής και συγκομιδής ενέργειας υψηλής πυκνότητας ισχύος, οι οποίες, όταν χρησιμοποιούνται με τα σημερινά νανοηλεκτρονικά χαμηλής ισχύος, θα μας επιτρέψουν να σχεδιάσουμε αυτοτροφοδοτούμενες έξυπνες ασύρματες αναγνωρίσιμες συσκευές που βασίζονται σε αισθητήρες.

- Τεχνολογίες ασφάλειας και προστασίας της ιδιωτικής ζωής:

Δύο σημαντικά ζητήματα στο IoT είναι η ιδιωτικότητα των ανθρώπων και η εμπιστευτικότητα των επιχειρηματικών διαδικασιών. Λόγω της κλίμακας ανάπτυξης, της κινητικότητάς τους και συχνά της σχετικά χαμηλής πολυπλοκότητάς τους, το νέφος των πραγμάτων είναι δύσκολο να ελεγχθεί. Για τη διασφάλιση της εμπιστευτικότητας, υπάρχει μεγάλος αριθμός τυποποιημένων τεχνολογιών κρυπτογράφησης προς χρήση. Ωστόσο, η κύρια πρόκληση είναι να γίνουν οι αλγόριθμοι κρυπτογράφησης ταχύτεροι και λιγότερο ενεργοβόροι. Επιπλέον, για τη χρήση ενός συστήματος κρυπτογράφησης θα πρέπει να υπάρχει ένα αποτελεσματικό σύστημα διανομής κλειδιών. Για συστήματα μικρής κλίμακας, η διανομή κλειδιών μπορεί να γίνει στο εργοστάσιο ή κατά τη στιγμή της εγκατάστασης, αλλά για τα ad-hoc δίκτυα, νέα σχήματα διανομής κλειδιών έχουν προταθεί μόνο τα τελευταία χρόνια. Για την προστασία της ιδιωτικής ζωής, η κατάσταση είναι πιο σοβαρή- ένας από τους λόγους είναι η άγνοια (όσον αφορά την προστασία της ιδιωτικής ζωής) του κοινού. Επιπλέον, η τεχνολογία διατήρησης της ιδιωτικής ζωής βρίσκεται ακόμη στα σπάργανα: τα συστήματα που λειτουργούν δεν είναι σχεδιασμένα για συσκευές με περιορισμένους πόρους και δεν έχει ακόμη αναπτυχθεί μια ολιστική άποψη για την ιδιωτική ζωή (π.χ. η άποψη για την ιδιωτική ζωή καθ' όλη τη διάρκεια της ζωής του ατόμου). Η ετερογένεια και η κινητικότητα των "πραγμάτων" στο IoT θα επιτείνει την πολυπλοκότητα της κατάστασης. Επίσης, από νομικής άποψης, ορισμένα ζητήματα δεν είναι ακόμη ξεκάθαρα και χρειάζονται νομική ερμηνεία- παραδείγματα αποτελούν ο αντίκτυπος της τοποθεσίας στη ρύθμιση της ιδιωτικής ζωής και το ζήτημα της ιδιοκτησίας των δεδομένων σε συνεργατικά νέφη "πραγμάτων". Το δίκτυο και η ανωνυμία των δεδομένων μπορούν να αποτελέσουν βάση για την προστασία της ιδιωτικής ζωής, αλλά προς το παρόν, οι τεχνολογίες αυτές υποστηρίζονται κυρίως από αρκετά ισχυρό εξοπλισμό, όσον αφορά την υπολογιστική ισχύ και το εύρος ζώνης. Ένα παρόμοιο επιχείρημα μπορεί να

χρησιμοποιηθεί για την αυθεντικοποίηση των συσκευών και την εδραίωση της εμπιστοσύνης.

- **Τυποποίηση**

Τα πρότυπα θα πρέπει να σχεδιάζονται για να υποστηρίζουν ένα ευρύ φάσμα εφαρμογών και να καλύπτουν κοινές απαιτήσεις από ένα ευρύ φάσμα βιομηχανικών τομέων, καθώς και τις ανάγκες του περιβάλλοντος, της κοινωνίας και των μεμονωμένων πολιτών. Μέσω διαδικασιών συναίνεσης με τη συμμετοχή πολλαπλών ενδιαφερομένων, θα είναι δυνατή η ανάπτυξη τυποποιημένων σημασιολογικών μοντέλων δεδομένων και οντολογιών, κοινών διεπαφών και πρωτοκόλλων, που θα ορίζονται αρχικά σε αφηρημένο επίπεδο και στη συνέχεια με παραδείγματα σύνδεσης με συγκεκριμένες διαπλατφορμικές και διαγλωσσικές τεχνολογίες, όπως XML, ASN.1, υπηρεσίες ιστού κ.λπ. Η χρήση σημασιολογικών οντολογιών και η μηχανικά αναγνώσιμη κωδικοποίηση θα πρέπει να συμβάλει στην αντιμετώπιση των ασάφειας που προκύπτουν από ανθρώπινα λάθη ή διαφορές και παρερμηνείες λόγω διαφορετικών ανθρώπινων γλωσσών σε διαφορετικές περιοχές του κόσμου, καθώς και στη διασταύρωση με πρόσθετες πληροφορίες που είναι διαθέσιμες μέσω άλλων συστημάτων. Απαιτούνται πρότυπα για την αμφίδρομη επικοινωνία και ανταλλαγή πληροφοριών μεταξύ των πραγμάτων, του περιβάλλοντός τους, των ψηφιακών ομολόγων τους στο εικονικό νέφος και των οντοτήτων που ενδιαφέρονται να παρακολουθούν, να ελέγχουν ή να βοηθούν τα πράγματα. Επιπλέον, ο σχεδιασμός των προτύπων για το IoT πρέπει να λαμβάνει υπόψη την αποδοτική και δικαστική χρήση της ενέργειας και της χωρητικότητας του δικτύου, καθώς και τον σεβασμό άλλων περιορισμών, όπως οι υφιστάμενοι κανονισμοί που περιορίζουν τις επιτρεπόμενες ζώνες συχνοτήτων και τα επίπεδα ισχύος για τις επικοινωνίες ραδιοσυχνοτήτων. Καθώς το IoT εξελίσσεται, μπορεί να είναι απαραίτητο να επανεξεταστούν οι εν λόγω κανονιστικοί περιορισμοί και να διερευνηθούν τρόποι για την εξασφάλιση επαρκούς χωρητικότητας για επέκταση, όπως η αναζήτηση πρόσθετης κατανομής ραδιοφάσματος όταν αυτό καθίσταται διαθέσιμο. Μια ιδιαίτερη πρόκληση από την άποψη αυτή είναι η διασφάλιση της παγκόσμιας διαλειτουργικότητας, ιδίως για τα πράγματα και τις συσκευές που κάνουν χρήση του ραδιοφάσματος. Ιστορικά, διάφορες ζώνες του ραδιοφάσματος έχουν διατεθεί για διάφορους σκοπούς, όπως οι ραδιοτηλεοπτικές επικοινωνίες (AM, FM, ψηφιακές ραδιοτηλεοπτικές εκπομπές ήχου, αναλογική επίγεια τηλεόραση, επίγεια ψηφιακή

τηλεόραση), η κινητή τηλεφωνία, η ραδιοφωνία πολιτών, οι επικοινωνίες υπηρεσιών έκτακτης ανάγκης, το ασύρματο διαδίκτυο, το ραδιόφωνο μικρής εμβέλειας. Δυστυχώς, οι κατανομές ζωνών συχνότητων δεν είναι ακριβώς εναρμονισμένες σε όλες τις περιοχές του κόσμου και ορισμένες ζώνες που είναι διαθέσιμες για ένα συγκεκριμένο σκοπό σε μιας περιοχής δεν είναι διαθέσιμα για τον ίδιο σκοπό σε άλλη περιοχή, συχνά επειδή χρησιμοποιούνται για διαφορετικό σκοπό. Η ανακατανομή του ραδιοφάσματος είναι μια αργή διαδικασία, στην οποία συμμετέχουν κυβερνητικές υπηρεσίες, ρυθμιστές και διεθνείς φορείς όπως η Διεθνής Ένωση Τηλεπικοινωνιών (ITU), καθώς και περιφερειακοί φορείς όπως το Ευρωπαϊκό Ινστιτούτο Τηλεπικοινωνιακών Προτύπων (ETSI) ή η Ομοσπονδιακή Επιτροπή Επικοινωνιών (FCC). Χρειάζονται προσεκτικές συζητήσεις για την ελαχιστοποίηση των διαταραχών στους υφιστάμενους χρήστες του ραδιοφάσματος και για τον προγραμματισμό των μελλοντικών αναγκών. Εν τω μεταξύ, πολλές συσκευές IoT που χρησιμοποιούν ραδιοφάσμα θα πρέπει να είναι ικανές να χρησιμοποιούν πολλαπλά πρωτόκολλα και πολλαπλές συχνότητες. Όσον αφορά το παράδειγμα του IoT γενικότερα, μια πολύ ενδιαφέρουσα προσπάθεια τυποποίησης έχει ξεκινήσει επί του παρόντος στο πλαίσιο του ETSI, το οποίο παράγει παγκοσμίως εφαρμόσιμα πρότυπα που σχετίζονται με τις ΤΠΕ. Στο πλαίσιο του ETSI, δημιουργήθηκε η Τεχνική Επιτροπή Machine-to-Machine (M2M) για τη διεξαγωγή δραστηριοτήτων τυποποίησης που σχετίζονται με συστήματα M2M και δίκτυα αισθητήρων. Το M2M είναι ένα πραγματικό παράδειγμα που οδηγεί προς το IoT, αλλά μέχρι στιγμής έχουν υπάρξει πολύ λιγότερες δραστηριότητες τυποποίησης σε αυτόν τον τομέα. Οι στόχοι της επιτροπής ETSI M2M περιλαμβάνουν: ανάπτυξη και συντήρηση μιας αρχιτεκτονικής από άκρο σε άκρο για M2M, ενίσχυση των προσπαθειών τυποποίησης για M2M, συμπεριλαμβανομένης της ενσωμάτωσης δικτύων αισθητήρων, της ονομασίας, της διευθυνσιοδότησης, της θέσης, του QoS, της ασφάλειας, της χρέωσης, της διαχείρισης, της εφαρμογής και των διεπαφών υλικού. Όσον αφορά τις δραστηριότητες της Ομάδας Εργασίας Τεχνολογίας Διαδικτύου (IETF) που σχετίζονται με το IoT, πρόσφατα δημιουργήθηκε η ομάδα Ipv6 over Low Power Wireless Personal Area Networks (6LoWPAN). Η 6LoWPAN έχει ορίσει ένα σύνολο πρωτοκόλλων που μπορούν να χρησιμοποιηθούν για την ενσωμάτωση κόμβων αισθητήρων σε δίκτυα Ipv6. Τα βασικά πρωτόκολλα για την αρχιτεκτονική 6LoWPAN έχουν ήδη καθοριστεί και έχουν κυκλοφορήσει ορισμένα εμπορικά προγράμματα που υλοποιούν αυτή τη σουίτα πρωτοκόλλων. Μια άλλη ομάδα

εργασίας στην IETF με την ονομασία routing over low power and lossy (ROLL) networks, δημιούργησε πρόσφατα το προσχέδιο του πρωτοκόλλου δρομολόγησης RPL. Αυτό θα αποτελέσει τη βάση για τη δρομολόγηση σε δίκτυα χαμηλής ισχύος και απωλειών, συμπεριλαμβανομένου του 6LoWPAN. Είναι σαφές ότι μια αναδυόμενη ιδέα είναι να θεωρηθεί η τυποποίηση του IoT ως αναπόσπαστο μέρος της μελλοντικής διαδικασίας ορισμού και τυποποίησης του Διαδικτύου. Ο ισχυρισμός αυτός διατυπώθηκε επίσης από το σύμπλεγμα ευρωπαϊκών έργων έρευνας και ανάπτυξης για το IoT (CERP-IoT). Όπως παρουσιάζεται στην έκθεσή του, η ενσωμάτωση διαφόρων πραγμάτων σε ευρύτερα δίκτυα, είτε κινητά είτε σταθερά, θα επιτρέψει τη διασύνδεσή τους με το μελλοντικό Διαδίκτυο[7].

4 Σχετική Βιβλιογραφία

4.1 Βασική ορολογία μετεωρολογίας

Η επιστήμη της μετεωρολογίας επικεντρώνεται στην μελέτη και ανάλυση των φαινομένων αυτών προκύπτουν από συνάρτηση η οποία αναλύει τα μετεωρολογικά δεδομένα, τον τόπο, τον χρόνο καθώς και άλλες πληροφορίες. Η ακρίβεια των προγνώσεων αυτών βασίζονται στα μοντέλα που θα χρησιμοποιηθούν. Υπάρχουν διάφορα μοντέλα που χρησιμοποιούνται από τις μετεωρολογικές υπηρεσίες σε Ευρώπη και Αμερική. Τέτοια μοντέλα είναι για παράδειγμα το ICON-EU, το οποίο είναι μια τοπική εκδοχή του παγκόσμιου μοντέλου NWP, το οποίο αναπτύχθηκε και χρησιμοποιείται από την Γερμανική Μετεωρολογική Υπηρεσία (DWD). Ξεκίνησε τη λειτουργία του από τις 15 Ιουνίου 2015 και καλύπτει σχεδόν όλη την περιοχή της Ευρώπης.[8]

4.1.1 Θερμοκρασία

Ως θερμοκρασία, ή θερμοκρασία ατμόσφαιρας ορίζουμε την θερμοκρασία που έχει ο ατμοσφαιρικός αέρας πάνω από μια περιοχή, και η πρόγνωση της βασίζεται στη γνώση της ατμοσφαιρικής πίεσης και της θερμοκρασίας στην περιοχή που μας ενδιαφέρει. Στην παρούσα εργασία όλες οι τιμές που περιγράφουν θερμοκρασία εκφράζονται σε κλίμακα Κελσίου (συμβολίζεται ως C°).

4.1.2 Ατμοσφαιρική Πίεση

Η ατμοσφαιρική πίεση είναι από τα σημαντικότερα μετεωρολογικά στοιχεία καθώς η μεταβολή της επηρεάζει άμεσα τις καιρικές καταστάσεις, καθώς και επηρεάζεται από το υψόμετρο. Έτσι, είναι αναγκαίο να γίνεται αναγωγή στην επιφάνεια της θάλασσας όταν λαμβάνεται υπόψη σε περιοχές υψηλότερες από αυτό της θάλασσας. Η σχέση μέσω της οποίας γίνεται η αναγωγή στην επιφάνεια της θάλασσας έχει ως εξής:

$$P = P_0 \left(1 - \frac{0,0065h}{^{\circ}\text{C} + 0,0065h + 213,75} \right)^{-5,257}$$

Η μονάδα μέτρησης της ατμοσφαιρικής πίεσης έχει οριστεί από τον Διεθνή Μετεωρολογικό Οργανισμό ως το Εκτοπασκάλ (1 hPa).

4.1.3 Υγρασία

Η υγρασία περιγράφει την περιεκτικότητα υδρατμών στον ατμοσφαιρικό αέρα. Εκφράζεται ως ποσοστό επί τοις εκατό (% 100) όπου 100% σημαίνει ότι ο ατμοσφαιρικός

αέρας έχει κορεστεί, δηλαδή περιέχει τη μέγιστη δυνατή ποσότητα υδρατμών και εκφράζεται ως ποσοστό επειδή η περιεκτικότητα υδρατμών έχει άμεση σχέση με τη θερμοκρασία του αέρα. Όσο χαμηλότερη είναι η θερμοκρασία του αέρα τόσο μικρότερη θα είναι η ποσότητα υδρατμών. Όταν μια μάζα υγρού θερμού αέρα ψυχθεί, θα μειωθεί αναγκαστικά η περιεκτικότητα σε υδρατμούς, και έτσι θα δημιουργηθεί είτε νέφος ή δρόσος, δηλαδή η συμπύκνωση κορεσμένου ατμοσφαιρικού αέρα πάνω σε ψυχρά αντικείμενα. Το σημείο δρόσου είναι το σημείο θερμοκρασίας στο οποίο παρατηρείται αυτό το φαινόμενο, και μπορεί να είναι οποιαδήποτε θερμοκρασία πάνω από τους 0 βαθμούς Κελσίου.[9]

4.1.4 Δείκτης δυσφορίας

Ο δείκτης δυσφορίας παρουσιάζει το πώς αντιλαμβανόμαστε τη θερμοκρασία ανά πάσα στιγμή, και βασίζεται σε διάφορους δείκτες για να υπολογιστεί. Μετρείται σε βαθμούς Κελσίου και προκύπτει από την παρακάτω συνάρτηση:

$$HI = -42.379 + 2.04901523T + 10.14333127R - 0.22475541TR - 6.83783 \times 10^{-3}T^2 - 5.481717 \times 10^{-2}R^2 + 1.22874 \times 10^{-3}T^2R + 8.5282 \times 10^{-4}TR^2 - 1.99 \times 10^{-6}T^2R^2$$

Όπου:

- HI ο δείκτης δυσφορίας
- T η θερμοκρασία σε βαθμούς Κελσίου και
- R η σχετική υγρασία ως ποσοστό.[10]

Ο παρακάτω πίνακας δίνει τη δυνατότητα υπολογισμού του δείκτη δυσφορίας εφόσον η θερμοκρασία και η σχετική υγρασία είναι γνωστές τιμές: [10]

ΔΕΙΚΤΗΣ ΘΕΡΜΟΤΗΤΑΣ - HEAT INDEX (C)																
	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%
42°	48	50	52	55	57	59	62	64	66	68	71	73	75	77	80	82
41°	46	48	51	53	55	57	59	61	64	66	68	70	72	74	76	79
40°	45	47	49	51	53	55	57	59	61	63	65	67	69	71	73	75
39°	43	45	47	49	51	53	55	57	59	61	63	65	66	68	70	72
38°	42	44	45	47	49	51	53	55	56	58	60	62	64	66	67	69
37°	40	42	44	45	47	49	51	52	54	56	58	59	61	63	65	66
36°	39	40	42	44	45	47	49	50	52	54	55	57	59	60	62	63
35°	37	39	40	42	44	45	47	48	50	51	53	54	56	58	59	61
34°	36	37	39	40	42	43	45	46	48	49	51	52	54	55	57	58
33°	34	36	37	39	40	41	43	44	46	47	48	50	51	53	54	55
32°	33	34	36	37	38	40	41	42	44	45	46	48	49	50	52	53
31°	32	33	34	35	37	38	39	40	42	43	44	45	47	48	49	50
30°	30	32	33	34	35	36	37	39	40	41	42	43	45	46	47	48
29°	29	30	31	32	33	35	36	37	38	39	40	41	42	43	45	46
28°	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43
27°	27	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
26°	26	26	27	28	29	30	31	32	33	34	35	36	37	38	39	39
25°	25	25	26	27	27	28	29	30	31	32	33	34	34	35	36	37
24°	24	24	24	25	26	27	28	28	29	30	31	32	33	33	34	35
23°	23	23	23	24	25	25	26	27	28	28	29	30	31	32	32	33
22°	22	22	22	22	23	24	25	25	26	27	27	28	29	30	30	31

Εικόνα 3: Τιμές Θερμοκρασίας / Υγρασίας για τον υπολογισμό του δείκτη δυσφορίας

Στον παρακάτω πίνακα φαίνονται 5 διαφορετικές ζώνες δυσφορίας:

< 29 C°	Καθόλου δυσφορία
30 - 34 C°	Ελαφρά αίσθηση δυσφορίας
35 - 39 C°	Έντονη αίσθηση δυσφορίας : να αποφεύγεται η έντονη κόπωση
40 - 45 C°	Υψηλή αίσθηση δυσφορίας : να αποφεύγεται η κόπωση
46 - 53 C°	Σημαντικός κίνδυνος : να αποφεύγεται κάθε δραστηριότητα
≥ 54 C°	Μεγάλος κίνδυνος για την υγεία: πιθανή καρδιακή προσβολή

Εικόνα 4: Επίπεδα δυσφορίας

4.1.5 Πρόγνωση καιρού

Η πρόγνωση του καιρού είναι μια επίμονη διαδικασία, η οποία πραγματοποιείται μέσα από ορισμένες μεθόδους τις οποίες θα αναλύσουμε επιγραμματικά εδώ:

Ο κανόνας της Εμμονής: “Οι τιμές των κινήσεων και των μεταβολών των συστημάτων που επικρατούσαν στο πρόσφατο παρελθόν, τείνουν να διατηρηθούν και στο μέλλον.” Με τον υπολογισμό της ταχύτητας κίνησης των βαρομετρικών συστημάτων και μετώπων μπορούμε να εκτιμήσουμε τον καιρό στην περιοχή που μας ενδιαφέρει καθώς θα επηρεάζεται από τα βαρομετρικά συστήματα που πλησιάζουν.

Ο κανόνας του Γεωστροφικού Ανέμου. “Οι αέριες μάζες μετακινούνται με την ταχύτητα περίπου του Γεωστροφικού Ανέμου”. Τα θερμά μέτωπα κινούνται με ταχύτητες 60%-80%

του Γεωστροφικού Ανέμου, ενώ τα ψυχρά μέτωπα κινούνται με ταχύτητες 80%-90% του ανέμου.

Ο κανόνας της βαρομετρικής τάσης: “Υπολογίζουμε τις τιμές της βαρομετρικής πίεσης στο μέλλον ως αποτέλεσμα του γινομένου της βαρομετρικής τάσης επί τον χρόνο”. Με τον υπολογισμό της μεταβολής της βαρομετρικής πίεσης σε χρονικό πλαίσιο τριών ωρών μπορούμε να προβλέψουμε το αν θα επικρατήσει καλοκαιρία ή κακοκαιρία.

Γίνεται σαφές λοιπόν ότι λόγω δεδομένων ο μόνος κανόνας που θα μας απασχολήσει πρακτικά είναι αυτός της βαρομετρικής τάσης.

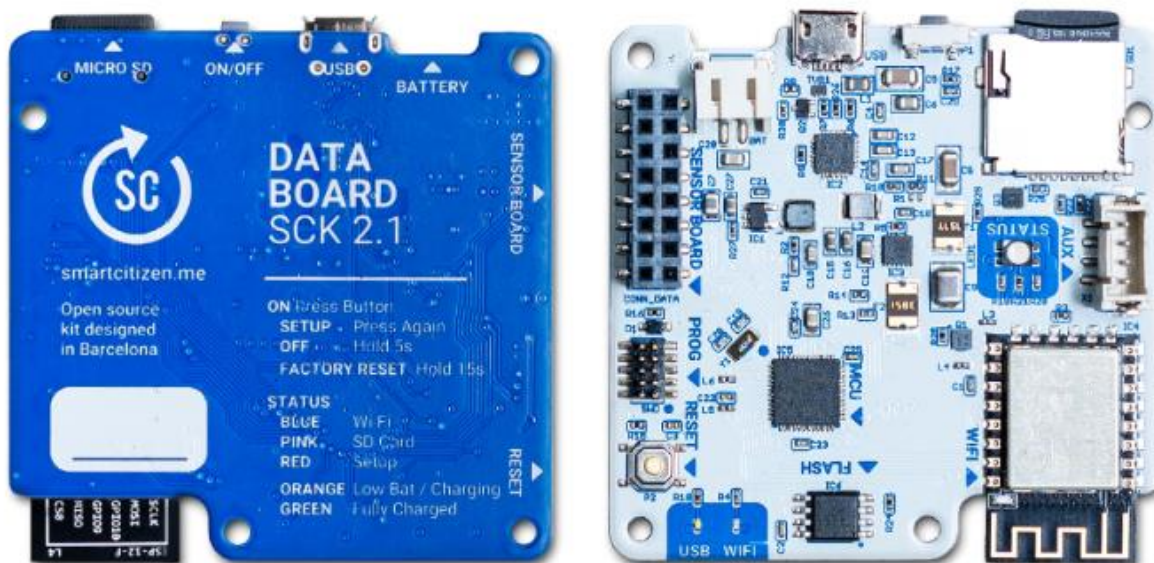
4.2 Το σύστημα Smart Citizen

Τα τελευταία χρόνια υπάρχει ερευνητικό ενδιαφέρον γύρω από τις δυνατότητες φθηνών περιβαλλοντικών αισθητήρων για την παρακολούθηση των επιπέδων μόλυνσης σε πόλεις, με διάφορες μεθόδους κλειστού κώδικα. Το σύστημα Smart Citizen παρέχει μια λύση ανοιχτού κώδικα για το ερευνητικό αυτό πεδίο προσφέροντας μια ευέλικτη και εύκολη στη χρήση της λύση παρακολούθησης διάφορων δεικτών όπως μονοξείδιο του άνθρακα ή επίπεδα θορύβου. Παρέχονται δύο διαφορετικά συστήματα, το Smart Citizen Kit και το Smart Citizen Station. Το πρώτο είναι αυτό που χρησιμοποιήθηκε για τη συλλογή δεδομένων της παρούσας εργασίας και χρησιμοποιείται για την παρακολούθηση ενός περιβάλλοντος για ερευνητικούς σκοπούς, ενώ το Smart Citizen Station είναι ένα πιο εξελιγμένο σύστημα που δύναται να συλλέγει περισσότερα δεδομένα μόλυνσης της ατμόσφαιρας με μεγαλύτερη ακρίβεια.



Εικόνα 5: Ένα Smart Citizen Kit με την πλακέτα Urban στο περιβλήμα του. Πηγή: <https://wikifactory.com/@elgranpri/smart-citizen-kit>

Το σύστημα συμπεριλαμβάνει και λογισμικό για την κεντρική αποθήκευση δεδομένων από πολλαπλά κιτ και δίνεται η δυνατότητα περαιτέρω επέκτασης του συστήματος με την προσθήκη επιπλέον αισθητήρων για συλλογή διάφορων πληροφοριών, όπως η υγρασία χώματος. Τόσο το Kit όσο και το Station βασίζονται για τη λειτουργία τους στο data board, το οποίο χρησιμοποιεί τον επεξεργαστή ARM M0 και είναι επεξεργαστής χαμηλής ενεργειακής κατανάλωσης που συνδυάζει 32 KB μνήμης RAM και 256 KB μνήμης flash, που βοηθούν στη διαχείριση μνήμης και υποστήριξης πολλαπλών αισθητήρων. Το data board συμπεριλαμβάνει επιπλέον μονάδα Wi-Fi για την αποστολή μετρήσεων στο κεντρικό αποθετήριο του Smart Citizen, θύρα κάρτας SD για τοπική αποθήκευση δεδομένων καθώς και 4 MB μνήμης flash για προσωρινή αποθήκευση δεδομένων σε περίπτωση ασταθούς σύνδεσης στο διαδίκτυο. Διατίθεται επίσης μπαταρία τόσο ώστε να συνεχίζεται η συλλογή δεδομένων ακόμα και κατά τη διάρκεια διακοπής ηλεκτροδότησης όσο και για τη λειτουργία των αισθητήρων χωρίς να απαιτείται παροχή ρεύματος, με τη διάρκεια ζωής τους να κυμαίνεται από 24 ώρες έως μια εβδομάδα, αναλόγως τη συχνότητα συλλογής και μέθοδο αποθήκευσης δεδομένων (καταγραφή τοπικά σε SD κάρτα ή αποστολή καταγραφών στο κεντρικό αποθετήριο του Smart Citizen). Στο data board συναντώνται θύρες επέκτασης μέσω των οποίων συνδέονται οι διάφοροι αισθητήρες.

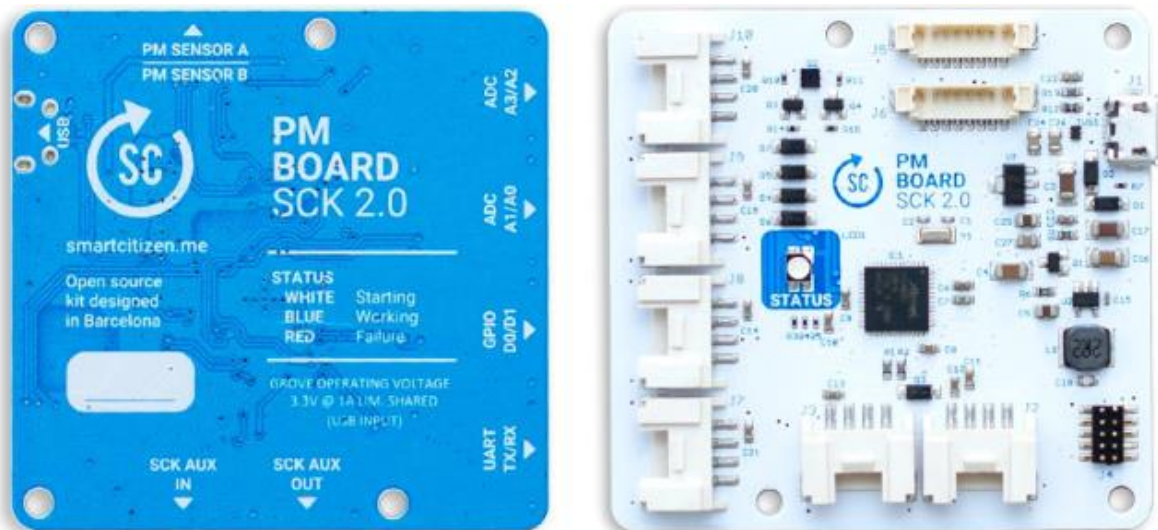


Εικόνα 6: Η μπροστά και πίσω όψη του Data Board

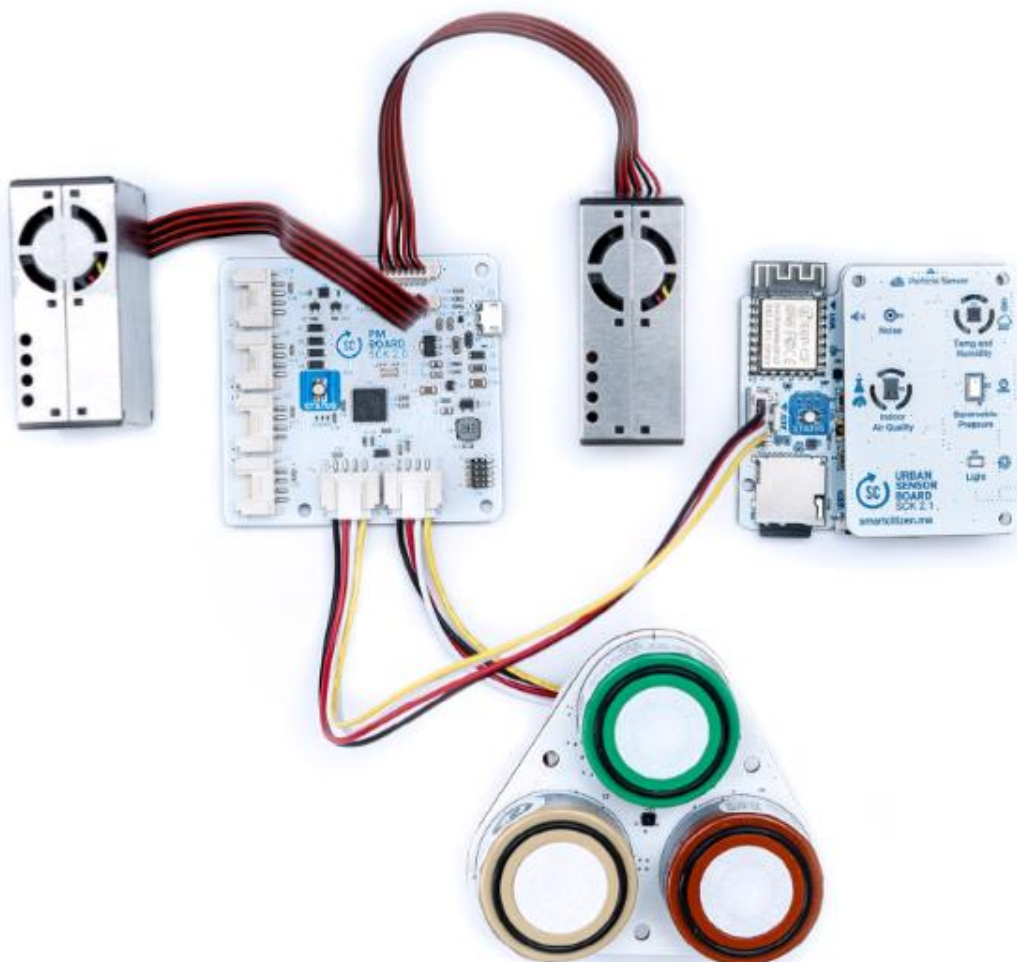
Η κυριότερη επέκταση του Smart Citizen Kit είναι μέσω του Urban Sensor Board (εικόνα 7), το οποίο συλλέγει δεδομένα θερμοκρασίας και υγρασίας, ποιότητας αέρα, βαρομετρικής πίεσης και φωτός. Το Urban Sensor Board φέρει και θύρα επέκτασης μέσω της οποίας μπορεί να συνδεθεί είτε ένας επιπλέον αισθητήρας καταμέτρησης σωματιδιακής ύλης είτε με το PM Sensor Board (εικόνα 7) πάνω στο οποίο μπορεί να συνδεθεί ο αισθητήρας καταμέτρησης σωματιδιακής ύλης και επιπλέον αισθητήρες, όπως ο Gas Pro Board (εικόνα 6) ο οποίος συλλέγει δεδομένα μονοξειδίου του άνθρακα, αζώτου και όζοντος. [11]



Εικόνα 7: Αισθητήρας Gas Pro Board



Εικόνα 8: Πίνακας επέκτασης PM Board



Εικόνα 9: Ενδεικτικό παράδειγμα χρήσης του PM Board σε συνδυασμό με αισθητήρες αωρούμενων σωματιδίων και του Gas Pro Board

4.3 Πληθοπορισμός

Ο πληθοπορισμός είναι ένα αυξανόμενο φαινόμενο στην ακαδημαϊκή έρευνα και πρακτική, όπου μια εταιρεία ή ένα ίδρυμα αναθέτει μια λειτουργία σε ένα απροσδιόριστο δίκτυο ανθρώπων με τη μορφή ανοικτής πρόσκλησης. Η έννοια έχει πολλούς ιδιοσυγκρασιακούς ορισμούς και έχει υποστεί πολλές παρεξηγήσεις. Παρά την αυξανόμενη προσοχή της, κορυφαία περιοδικά έχουν μόλις πρόσφατα αρχίσει να δημοσιεύουν για τον πληθοπορισμό. Οι εφαρμογές του πληθοπορισμού για διάφορες δραστηριότητες διευρύνονται και η σχετική ορολογία επίσης εξελίσσεται. Οι Doan κ.ά. (2011) συγκέντρωσαν έναν κατάλογο επικαλυπτόμενων όρων, όπως η ομότιμη παραγωγή, τα συστήματα που τροφοδοτούνται από τους χρήστες, το περιεχόμενο που παράγεται από τους χρήστες, τα συνεργατικά συστήματα, τα συστήματα κοινοτήτων, τα κοινωνικά συστήματα, η κοινωνική αναζήτηση, τα κοινωνικά μέσα, η συλλογική νοημοσύνη, τα wikinomics, η σοφία του πλήθους, οι έξυπνοι όχλοι, η μαζική συνεργασία και ο ανθρώπινος υπολογισμός.

Ο πληθοπορισμός έχει μακρά ιστορία, με παραδείγματα όπως το "Βραβείο γεωγραφικού μήκους" της βρετανικής κυβέρνησης το 1714 και το Oxford English Dictionary που χρησιμοποίησε 800 αναγνώστες για να καταγράψει λέξεις το 1884. Ο διαγωνισμός λογότυπου της Toyota το 1936 ήταν ένα πρώιμο παράδειγμα χρήσης του πληθοπορισμού στον εταιρικό κόσμο. Ο πληθοπορισμός είναι πιο οργανωμένο μέσω του Διαδικτύου, παρέχοντας εύκολη πρόσβαση σε άτομα από οπουδήποτε στον κόσμο. Ευρείες συστοιχίες μεσαζόντων που έχουν ως στόχο να βοηθήσουν άλλους οργανισμούς να επωφεληθούν από τον πληθοπορισμό λειτουργούν ενεργά τις πλατφόρμες τους που βασίζονται στο Διαδίκτυο, ισχυριζόμενοι ότι επιλύουν επιστημονικά και επιχειρηματικά προβλήματα που δεν μπορούν να λύσουν οι ειδικοί στα τμήματα έρευνας και ανάπτυξης των μεγάλων εταιρειών.

Η έννοια της ανοικτής καινοτομίας εισήχθη από τον Chesbrough (2003) και πολλοί μελετητές πιστεύουν ότι ο πληθοπορισμός αποτελεί σημαντικό τύπο ανοικτής καινοτομίας. Κατά συνέπεια, ο πληθοπορισμός είναι μία από τις πιο συχνά χρησιμοποιούμενες λέξεις-κλειδιά στη βιβλιογραφία της ανοικτής καινοτομίας. Παρόλο που ο αριθμός των ακαδημαϊκών άρθρων για τον πληθοπορισμό έχει αυξηθεί, ο αριθμός των άρθρων ανασκόπησης που συζητούν μελέτες πληθοπορισμού παραμένει πολύ μικρός. Απαιτείται περαιτέρω έρευνα για τη διερεύνηση της υπάρχουσας βιβλιογραφίας σχετικά με τον πληθοπορισμό, συμπεριλαμβανομένων των εφαρμογών, των οφελών και

των προκλήσεων. Στόχος της παρούσας μελέτης είναι να διερευνήσει την ανάπτυξη της βιβλιογραφίας για τον πληθοπορισμός και τις εφαρμογές του σε διάφορους τομείς.

Ο πληθοπορισμός είναι ένα νέο εργαλείο. Αντικαθιστά παλιούς τρόπους λειτουργίας αλλά και τους τροποποιεί. Κατά την εκτενή βιβλιογραφική ανασκόπηση, οι συγγραφείς της παρούσας μελέτης δεν βρήκαν καμία ολοκληρωμένη παρουσίαση των εφαρμογών πληθοπορισμού. Με βάση όλες τις διαθέσιμες πληροφορίες, παρουσιάζεται ένας κατάλογος των υφιστάμενων εφαρμογών crowdsourcing.

- Παραγωγή ιδεών

Οι Schweitzer κ.ά. υποστηρίζουν ότι είναι σημαντικό για τις επιχειρήσεις να κατανοήσουν πότε και πώς να χρησιμοποιούν τον διαγωνισμό ιδεών αντί των συμβατικών τεχνικών έρευνας. Η παραγωγή ιδεών μέσω πληθοπορισμού πραγματοποιείται κυρίως με δύο μορφές: διαγωνισμοί ιδεών και ιδεοληψία με συλλογική νοημοσύνη. Οι διαγωνισμοί ιδεών κάνουν θεμελιώδη χρήση του πληθοπορισμού. Σε έναν διαγωνισμό ιδεών, οργανισμοί ή άτομα καλούν τα πλήθη να υποβάλουν νέες ιδέες και επιλέγονται οι καλύτερες ιδέες. Οι Poetz και Schreier απέδειξαν ότι τα πλήθη μπορούν να ξεπεράσουν τους επαγγελματίες σε πολλά επίπεδα ιδεοληψίας νέων προϊόντων. Μαζί με τις επιχειρήσεις, έχουν εμφανιστεί πολλοί ενδιάμεσοι φορείς για να ζητούν ιδέες από πλήθη για άλλους οργανισμούς και άτομα. Σε διαγωνισμούς ιδεών μέσω crowdsourcing, μια ιδέα ανταμείβεται αν ικανοποιεί τον αναζητητή ιδεών. Μεγάλες εταιρείες όπως η Dell, η IBM, η Procter & Gamble, η Starbucks και η Unilever έχουν στραφεί στον πληθοπορισμό, για παράδειγμα, για να βρουν καλύτερα σχέδια προϊόντων. Έχουν διοργανώσει τόσο διαγωνισμούς ιδεών όσο και εμπλεκόμενους όχλους για τη συλλογική παραγωγή ιδεών. Η ιδεοληψία από εξωτερικές πηγές έχει αναδειχθεί ως πολύτιμη εναλλακτική λύση για την εσωτερική ιδεοληψία. Όταν δημιουργούνται καινοτομίες, αναμένονται προβλέψεις ή λαμβάνονται αποφάσεις, η συλλογική προσπάθεια μεγάλου αριθμού ατόμων είναι συχνά καλύτερη από την εξάρτηση από λίγους ειδικούς. Η ποικιλομορφία του πλήθους ενισχύει σημαντικά τις πιθανότητες παραγωγής χρήσιμων ιδεών. Οι αρχιτεκτονικές συμμετοχής του πλήθους χρειάζονται αναπροσανατολισμό. Η έμφαση πρέπει να μετατοπιστεί από την παραγωγή ιδεών στην εξέλιξη των ιδεών. Τα χρηματικά βραβεία θα πρέπει να είναι πιο ανταγωνιστικά. Τα κίνητρα θα πρέπει να είναι πιο εξατομικευμένα. Το πλήθος θα πρέπει να έχει την ευκαιρία να εκφράζει τις απόψεις του σε όλα τα στάδια της διαδικασίας καινοτομίας. Μια μελέτη του Bayus (2013) διαπίστωσε ότι οι κατά συρροή συνεισφέροντες είναι πιο πιθανό από τους συνεισφέροντες με μία μόνο ιδέα να δημιουργήσουν ιδέες, οι οποίες είναι πολύτιμες για την επιχείρηση που αναζητά ιδέες.

Ωστόσο, η ίδια μελέτη διαπίστωσε ότι οι συνεισφέροντες με επιτυχία στο παρελθόν είναι επιρρεπείς στο να προτείνουν ιδέες παρόμοιες με τις ήδη υλοποιημένες ιδέες τους.

- Microtasking

Το Microtasking είναι ένα θέμα με έντονο ενδιαφέρον στον τομέα του crowdsourcing. Το microtasking ορίζεται ως ένα σύστημα στο οποίο οι χρήστες μπορούν να επιλέγουν και να ολοκληρώνουν μικρές εργασίες για χρηματικές ή μη χρηματικές ανταμοιβές:

(1) Φάση αξιολόγησης: Ποιες διαδικασίες ή έργα μπορούν να αποτελέσουν αντικείμενο crowdsourcing.

(2) Σχεδιασμός εργασιών: Πώς οι εργασίες μπορούν να αποσυντεθούν και να συγκεντρωθούν σε μικροεργασίες.

(3) Ενσωμάτωση: Πώς οι εργασίες μπορούν να ενσωματωθούν στις υπάρχουσες διαδικασίες.

Οι ενδιαμέσες πλατφόρμες μικροεργασιών περιλαμβάνουν τρεις τύπους φορέων: τον εργοδότη, τον φορέα εκμετάλλευσης της πλατφόρμας και τους μικροεργαζόμενους. Ο εργοδότης φροντίζει για τη δημιουργία εργασιών και την εξεύρεση πόρων. Ο φορέας εκμετάλλευσης της πλατφόρμας παρέχει την απαραίτητη υποδομή και λειτουργεί ως διαμεσολαβητής μεταξύ του εργοδότη και των μικροεργαζομένων. Οι μικροεργάτες ολοκληρώνουν τις μικροεργασίες και λαμβάνουν ως αντάλλαγμα χρηματική αποζημίωση. Οι πλατφόρμες μικροεργασιών, για παράδειγμα, η Mechanical Turk της Amazon, έχουν λάβει σημαντική προσοχή από μελετητές, μικροεργαζόμενους και διευθυντές διαφόρων κλάδων. Μεγάλος αριθμός άρθρων έχει δημοσιευτεί με βάση δεδομένα από το Mechanical Turk της Amazon, το οποίο ξεκίνησε το 2005 ως υπηρεσία για την εκτέλεση εργασιών έντασης εργασίας με τη βοήθεια του crowdsourcing. Το Amazon's Mechanical Turk συντονίζει την προσφορά και τη ζήτηση εργασιών που απαιτούν ανθρώπινη νοημοσύνη για να ολοκληρωθούν. Το Microtasking δημιουργεί έναν τρόπο για τη συμμετοχή διαφορετικών πλήθους, με χαμηλό χρηματικό κόστος και συχνά γρήγορο χρόνο απόκρισης. Συνήθως, η ανταμοιβή για την ολοκλήρωση μικροεργασιών κυμαίνεται από 0,01 έως 1,0 δολάρια ανά εργασία. Ένα χαρακτηριστικό παράδειγμα του microtasking συναντάται στην ανάπτυξη αυτόνομων οχημάτων. Τα τελευταία χρόνια, αρκετές μεγάλες και καλά χρηματοδοτημένες εταιρείες έχουν ξεκινήσει την προσπάθεια να φέρουν πλήρως αυτόνομα οχήματα στους δρόμους. Ωστόσο, είναι αβέβαιο αν ο στόχος αυτός μπορεί να επιτευχθεί στο εγγύς μέλλον. Παρά την εντυπωσιακή πρόοδο, έχουν προκύψει προβλήματα που είναι δύσκολο να επιλυθούν, γεγονός που μετριάξει την αισιοδοξία των

μηχανικών που εργάζονται σε αυτή την πρόκληση. Παρά τα δισεκατομμύρια δολάρια που έχουν επενδυθεί στην ανάπτυξη της τεχνολογίας, οι ανταγωνιστές στον τομέα αυτό περιλαμβάνουν κατασκευαστές αυτοκινήτων, εταιρείες τεχνολογίας και νέες εταιρείες από τη Silicon Valley και την Κίνα. Μέχρι το 2018, 55 εταιρείες είχαν εξασφαλίσει άδειες δοκιμής αυτόνομων οχημάτων στην Καλιφόρνια. Οι εταιρείες αυτές χρειάζονται εκατομμύρια επισημασμένες εικόνες ως δεδομένα βασικής αλήθειας για να διδάξουν στους αλγορίθμους τους πώς να βλέπουν. Αυτές οι εικόνες είναι συνήθως φωτογραφίες που λαμβάνονται από βίντεο κυκλοφορίας και σχολιάζονται χειροκίνητα για να γίνουν αναγνώσιμες από μηχανήματα. Η αξιοπιστία των προγνωστικών μοντέλων μηχανικής μάθησης εξαρτάται από την ακρίβεια των δεδομένων εκπαίδευσης, η οποία μπορεί να αυτοματοποιηθεί μόνο εν μέρει. Οι αυτοκινητοβιομηχανίες αναθέτουν την εργασία αυτή σε εξειδικευμένες εταιρείες, όπως πλατφόρμες crowdsourcing ή εταιρείες εξωτερικής ανάθεσης επιχειρηματικών διαδικασιών (BPO), οι οποίες παράγουν εξαιρετικά πλεονάζοντα δεδομένα βασικής αλήθειας. Ενώ οι εταιρείες BPO έχουν μεγαλύτερο έλεγχο των εργαζομένων τους και μπορούν να λειτουργήσουν ως ενδιάμεσος σε ένα εργατικό δυναμικό που μιλάει μόνο μια τοπική γλώσσα, είναι λιγότερο ευέλικτες στην αντιμετώπιση ακραίων διακυμάνσεων της ζήτησης που είναι χαρακτηριστικές της μικροεργασίας.[12]

- Λογισμικό ανοικτού κώδικα

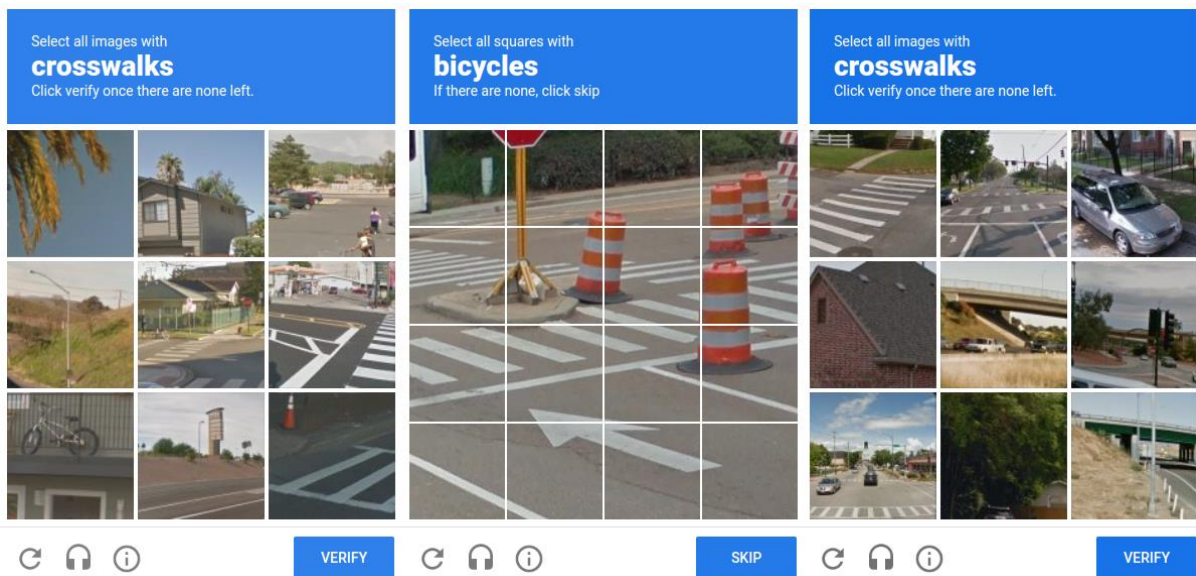
Το Crowdsourcing έχει δώσει μια νέα διάσταση στην ανάπτυξη λογισμικού. Το κίνημα OSS έδειξε ότι ένα δίκτυο παθιασμένων, σπασίκλων εθελοντών μπορεί να γράφει κώδικα, ο οποίος είναι εξίσου χρήσιμος με τον κώδικα που γράφεται από επαγγελματίες σε μια εταιρεία λογισμικού όπως η Microsoft. Στο OSS, τα πλήθη αναπτύσσουν συλλογικά λογισμικό για το γενικότερο καλό των ανθρώπων. Το OSS είναι ένα λογισμικό υπολογιστή, ο πηγαίος κώδικας του οποίου είναι δημόσιος και μπορεί να χρησιμοποιηθεί δωρεάν. Ο Mozilla Firefox, ο Thunderbird, το OpenOffice και το Ubuntu είναι παραδείγματα του μεγάλου αριθμού πλατφορμών OSS που είναι διαθέσιμες σήμερα. Τα έργα OSS αναπτύσσονται με μικρές συνεισφορές από κάθε μέλος μιας μεγάλης κοινότητας συνεισφερόντων. Συνήθως, υπάρχει μια μικρή ομάδα πολύ ενεργών συνεισφερόντων μαζί με μια πολύ μεγάλη ομάδα λανθανόντων συνεισφερόντων που παρακολουθούν την ανάπτυξη και συνεισφέρουν περιστασιακά. Στην ανάπτυξη OSS, η συνεχής ενεργός αξιολόγηση των εισροών από το πλήθος αυξάνει την ποιότητα του λογισμικού. Τα αποτελέσματα της ανάπτυξης OSS, σε πολλές περιπτώσεις, έχουν καλύτερες επιδόσεις από συγκρίσιμα εμπορικά προϊόντα. Πολλές μεγάλες επιχειρήσεις υποστηρίζουν

πλατφόρμες crowdsourcing που επικεντρώνονται στην ανάπτυξη OSS για να έχουν πρόσβαση στο νέο λογισμικό σε αντάλλαγμα. Οι τεχνικές Crowdsourcing μαζί με το ανοικτό, δημόσια προσβάσιμο λογισμικό και τα δεδομένα ήταν χρήσιμες για την ανάπτυξη ισχυρών εργαλείων ενεργειακής ανάλυσης που παρείχαν κρίσιμες, σημαντικές για την πολιτική γνώσεις, ιδίως σε χώρες με χαμηλό εισόδημα. Υποστηρίζεται ότι το OSS και το crowdsourcing είναι συγκεχυμένοι όροι, αλλά υποστηρίζεται ότι το OSS αφορά μια κοινότητα που μοιράζεται κώδικα υπολογιστή για το κοινό καλό, ενώ το crowdsourcing περιλαμβάνει πολλούς συνεισφέροντες αλλά λίγους δικαιούχους.

- Συμμετοχή του κοινού

Η ενίσχυση της συμμετοχής του κοινού αποτελεί βασική προτεραιότητα στο δημόσιο σχεδιασμό και η μαζική ανάθεση είναι ένα εφαρμόσιμο μοντέλο για τη διασφάλιση της συμμετοχής του κοινού σε έργα δημόσιου σχεδιασμού. Το Crowdsourcing είναι χρήσιμο για τη συμμετοχή των πολιτών σε έργα δημόσιου σχεδιασμού, ώστε να αξιοποιηθεί η άγνωστη μακρινή ιδιοφυΐα τους. Η συμμετοχή του κοινού μέσω crowdsourcing μπορεί να εμπλέξει ένα ευρύ φάσμα ανθρώπων και μπορεί να διευκολύνει έναν ανοικτό διάλογο μεταξύ πολιτών και υπευθύνων λήψης αποφάσεων. Πολλοί κυβερνητικοί οργανισμοί, για παράδειγμα η κυβέρνηση Ομπάμα της Αμερικής, έχουν κάνει εκτεταμένη χρήση των νέων τεχνολογιών crowdsourcing για να ενισχύσουν τη συμμετοχή του κοινού σε διάφορες δραστηριότητες, όπως η κοινοτική πρόνοια. Ένα άλλο παράδειγμα είναι η Νέα Ζηλανδία, όπου το κοινοβούλιο ψήφισε το "Wiki Policing Act 2008", το οποίο δίνει στους ψηφοφόρους τη δυνατότητα να συμμετέχουν σε διάλογο με την αστυνομία. Ο αστυνομικός νόμος της Νέας Ζηλανδίας του 1958 παρουσιάστηκε στο κοινό σε μορφή wiki για να τον τροποποιήσει και να τον ξαναγράψει. Αυτή η wiki-έκδοση του νέου αστυνομικού νόμου εγκρίθηκε επίσημα από το κοινοβούλιο της Νέας Ζηλανδίας το 2008. Η συμμετοχή του κοινού μέσω crowdsourcing μπορεί να βοηθήσει στην ανάπτυξη έργων μεγάλης κλίμακας με υψηλή ακρίβεια, όπως το Google Earth. Καθώς το crowdsourcing αντικαθιστά υπηρεσίες που παραδοσιακά κατασκευάζονταν από υψηλά αμειβόμενους επαγγελματίες, μπορούν να δημιουργηθούν ανοικτές βάσεις δεδομένων με βάση το crowdsourcing με πολύ χαμηλό κόστος προς όφελος της ανθρωπότητας. Το χαμηλό κόστος είναι εφικτό επειδή, στο crowdsourcing, το μεγαλύτερο μέρος της εργασίας γίνεται από μεγάλο αριθμό ατόμων που συνεισφέρουν κυρίως χωρίς χρηματική αποζημίωση. Επισημαίνεται ότι το crowdsourcing είναι ένα αποτελεσματικό συμπληρωματικό μέσο δημόσιας συμμετοχής, αλλά δεν μπορεί να αντικαταστήσει τα παραδοσιακά μέσα δημόσιας συμμετοχής. Υποστηρίζεται ότι στον δημόσιο σχεδιασμό το crowdsourcing είναι πιο αποτελεσματικό

από τη συμβατική συμμετοχή των πολιτών. Μορφές συμμετοχής του κοινού στον πληθοπορισμό, εκούσιες και μη, αποτελούν η Wikipedia ή η συμπλήρωση ερωτήσεων CAPTCHA, συνήθως με τη μορφή επιλογής εικόνων που αντιστοιχούν σε μια ιδέα ή μια συγκεκριμένη λέξη, ή παλαιότερες εκδοχές των CAPTCHA όπου ο χρήστης έβλεπε δύο λέξεις παραμορφωμένες και καλούταν να εντοπίσει τις λέξεις που αναγράφονται ως απάντηση. Αυτή η μορφή CAPTCHA χρησιμοποιήθηκε για την ανάπτυξη αλγορίθμων αναγνώρισης χαρακτήρων (OCR), δίνοντας μια γνωστή για το σύστημα λέξη για την επιβεβαίωση πως ο χρήστης είναι πραγματικός και μια δεύτερη, άγνωστη για το σύστημα λέξη η οποία γινόταν γνωστή μετά από πολλές ομόφωνες απαντήσεις χρηστών.[13]



Εικόνα 10: Παραδείγματα CAPTCHAs

- Επιστήμη των πολιτών

Η επιστήμη των πολιτών είναι μια μορφή συνεργατικής έρευνας στην οποία αξιοποιείται η συμμετοχή πλήθους στην επίλυση προβλημάτων του πραγματικού κόσμου. Ένας επιστήμονας-πολίτης συλλέγει και επεξεργάζεται εθελοντικά, για παράδειγμα, δεδομένα για επιστημονική έρευνα. Συνήθως, οι εθελοντές συμμετέχουν σε επιστημονικά έργα λόγω του ενδιαφέροντος και της περιέργειάς τους και επειδή αισθάνονται ότι η βοήθεια στην έρευνα είναι δική τους ευθύνη. Έχει διαπιστωθεί ότι τα κίνητρα των επιστημόνων πολιτών είναι διαχρονικά και δυναμικά. Τα έργα που καθοδηγούνται από επιστήμονες πολίτες αναπτύσσονται σε διάφορους κλάδους, όπως η οικολογία, η γεωλογία, η ιατρική και το περιβάλλον. Η επιστήμη των πολιτών συμβάλλει στη διεξαγωγή επιστημονικών ερευνών όχι μόνο στην καθημερινή επιστήμη αλλά και σε εξαιρετικά περίπλοκες επιστημονικές εργασίες, όπως η γονιδιωματική αλληλουχία. Τα ερευνητικά έργα που χρησιμοποιούν την επιστήμη των πολιτών οργανώνονται συνήθως σε εικονικές

πλατφόρμες. Υπάρχουν πολλές ενδείξεις ότι η επιστήμη των πολιτών με προσεκτικό σχεδιασμό της μελέτης και βέλτιστες συνθήκες μπορεί να παράγει υψηλής ποιότητας και αξιόπιστα αποτελέσματα. Συχνά, έχουν επίσης επιτευχθεί απρόβλεπτες λύσεις. Οι πλατφόρμες των μέσων κοινωνικής δικτύωσης, όπως το Facebook, το Twitter και το YouTube, έχουν αυξηθεί κατακόρυφα ως ένας τρόπος διάδοσης πληροφοριών από μια ποικιλία επίσημων και μη επίσημων πηγών.

- Δημοσιογραφία των πολιτών

Η δημοσιογραφία των πολιτών είναι ένα εναλλακτικό μέσο ενημέρωσης, το οποίο στρέφεται στα πλήθη για τη δημοσιογραφία του Διαδικτύου. Οι πολίτες δημοσιογράφοι βρίσκονται σε έντονο ανταγωνισμό μεταξύ τους, με τους επαγγελματίες δημοσιογράφους και με τους ακαδημαϊκούς. Οι πλατφόρμες Crowdsourcing της δημοσιογραφίας των πολιτών, όπως το Digg.com και το Newsvine, προσφέρουν στους χρήστες τη δυνατότητα να υποβάλλουν, να βαθμολογούν, να προτείνουν και να σχολιάζουν ειδήσεις. Στην αναδυόμενη δημοσιογραφία που χρηματοδοτείται από το πλήθος, οι δωρεές από τους αναγνώστες δημιουργούν μια ισχυρή σχέση μεταξύ δημοσιογράφων και δωρητών, η οποία δημιουργεί μια νέα αίσθηση ευθύνης για τους δημοσιογράφους. Οι γεωγραφικές πληροφορίες που συλλέγονται από εθελοντές έχουν αναδειχθεί σε πολύτιμη εναλλακτική λύση για τα παραδοσιακά γραφεία χαρτογράφησης. Η ποιότητα των δεδομένων αποτελεί μεγάλο μέλημα στη δημοσιογραφία των πολιτών. Η δημοσιογραφία των πολιτών βασίζεται στην ποικιλομορφία των συνεισφερόντων, η οποία μπορεί επίσης να αυξήσει την αξιοπιστία των ειδήσεων.[14]

4.4 Αντίστοιχες εφαρμογές

4.4.1 Ανίχνευση καιρικών φαινομένων μέσω φωτογραφιών με τη χρήση νευρωνικών δικτύων

Οι μετεωρολογικές πληροφορίες επηρεάζουν την καθημερινή ζωή και τις γεωργικές δραστηριότητες, καθιστώντας ζωτικής σημασίας την πρόβλεψη των καιρικών συνθηκών από τα συστήματα υπολογιστών. Η παρούσα μελέτη χρησιμοποιεί τέσσερις κατηγορίες: συννεφιά, βροχή, ηλιοφάνεια και ανατολή. Ο συνδυασμός των μοντέλων GoogLeNet και VGG-16 με νευρωνικά δίκτυα αιχμής (SNN) οδηγεί σε υψηλά ποσοστά ακρίβειας ταξινόμησης 98,48%, 97,58%, 97% και 98,48% για τις κλάσεις συννεφιά, βροχή, ηλιοφάνεια και ανατολή. Η χρήση SNN σε συνδυασμό με μοντέλα βαθιάς μάθησης αποδεικνύεται επίσης επιτυχής. Το κόστος των αισθητήρων με κάμερες έχει αρνητικό αντίκτυπο στην οικονομία της περιοχής. Η τεχνολογία τεχνητής νοημοσύνης αναμένεται να επικρατήσει

περισσότερο στα ενσωματωμένα συστήματα, επιτρέποντας ακριβέστερη ανάλυση και μειωμένο κόστος υλικού. Οι μεγάλες παγκόσμιες εταιρείες χρησιμοποιούν εφαρμογές τεχνητής νοημοσύνης στις τεχνολογίες τους, με τις αρχιτεκτονικές βαθιάς μάθησης να αποτελούν υποκλάδο της τεχνητής νοημοσύνης. Έχουν διεξαχθεί αρκετές μελέτες για την ταξινόμηση εικόνων καιρού με τη χρήση αρχιτεκτονικών βαθιάς μάθησης. Ορισμένες μελέτες ταξινόμησαν εικόνες καιρού πέντε κατηγοριών χρησιμοποιώντας μοντέλα CNN, ενώ άλλες χρησιμοποίησαν μοντέλα επαναλαμβανόμενων νευρωνικών δικτύων (RNN) και μεθόδους μακράς βραχυπρόθεσμης μνήμης (LSTM). Το πιο επιτυχημένο ποσοστό ταξινόμησης μεταξύ διαφορετικών συνόλων δεδομένων ήταν 92,63%. Μια άλλη μελέτη ταξινόμησε εναέριες εικόνες σε δύο τύπους (συννεφιασμένες και ηλιόλουστες) χρησιμοποιώντας την τεχνική του κατευθυντικού ιστογράμματος χρώματος. Η επιτυχία ταξινόμησης που επιτεύχθηκε με μηχανές διανυσμάτων υποστήριξης (SVM) με την τεχνική αύξησης δεδομένων ήταν 98,6%. Μια άλλη μελέτη χώρισε τις εικόνες καιρού σε δύο κατηγορίες χρησιμοποιώντας τα χαρακτηριστικά που προέκυψαν από τα πλήρως συνδεδεμένα στρώματα της αρχιτεκτονικής AlexNet, επιτυγχάνοντας ακρίβεια 91,1%. Μια άλλη μελέτη χρησιμοποίησε δύο διαφορετικά σύνολα δεδομένων, δύο κλάσεων και τεσσάρων κλάσεων, χρησιμοποιώντας τα μοντέλα ResNet και AlexNet. Η μέθοδος SVM πολλαπλών κλάσεων προτιμήθηκε, με την καλύτερη επιτυχία ακρίβειας στο σύνολο δεδομένων δύο κλάσεων να είναι 90% με το μοντέλο ResNet και ένα μέσο ποσοστό ακρίβειας 97% στο σύνολο δεδομένων τεσσάρων κλάσεων. Μια τελική μελέτη χρησιμοποίησε τρεις κλάσεις δεδομένων καιρού και εφάρμοσε τη μέθοδο superpixel και την τεχνική επαύξησης για την ομοιογενή αύξηση των σπόρων εικονοστοιχείων σε κάθε εικόνα. Η επιτυχία της μεθόδου superpixel με 100 σπόρους αυξήθηκε περισσότερο. Η μέθοδος SVM χρησιμοποιήθηκε για την ταξινόμηση εικόνων καιρού, με το μοντέλο ResNet-50 να επιτυγχάνει το καλύτερο συνολικό ποσοστό ακρίβειας 80,7%. Συμπερασματικά, η χρήση μοντέλων νευρωνικών δικτύων συνελκτικού τύπου (CNN) και νευρωνικών δικτύων αιχμής (SNN) στην ταξινόμηση εικόνων καιρού είναι μια νέα προσέγγιση που έχει τη δυνατότητα να βελτιώσει σημαντικά την ακρίβεια και την αποτελεσματικότητα των εικόνων καιρού. Στόχος της μελέτης ήταν η ταξινόμηση εικόνων καιρού με τη χρήση δικτύων CNN και δικτύων τρίτης γενιάς (SNN) για την αυτόματη εξαγωγή χαρακτηριστικών. Ο συνδυασμός CNNs και SNNs βελτίωσε την απόδοση γενίκευσης. Η προτεινόμενη προσέγγιση συνδύασε μοντέλα βαθιάς μάθησης και SNNs για την ταξινόμηση συνόλων δεδομένων με διαφορετικές αναλύσεις. Ο συνδυασμός των χαρακτηριστικών CNN συνέβαλε στην απόδοση ταξινόμησης των SNNs. Το πείραμα περιελάμβανε συνδυασμό μοντέλων CNN και SNN και τα συνδυασμένα χαρακτηριστικά

ταξινομήθηκαν και εκπαιδεύτηκαν με το μοντέλο SNN. Η προτεινόμενη προσέγγιση βελτιώσε την ικανότητα γενίκευσης του μοντέλου. Οι μελλοντικές μελέτες θα χρησιμοποιήσουν διαφορετικά σύνολα δεδομένων και μεθόδους μηχανικής μάθησης σε συνδυασμό με διαφορετικά μοντέλα βαθιάς μάθησης.[15]

4.4.2 Πρόγνωση καιρού πολλαπλών κατηγοριών από το Twitter χρησιμοποιώντας προσεγγίσεις μηχανικής μάθησης και πληθοπορισμού

Η πρόγνωση του καιρού είναι μια κρίσιμη πτυχή της καθημερινής ζωής, με τους χρήστες να τη βιώνουν τέσσερις έως πέντε φορές την ημέρα μέσω διαφόρων πηγών, όπως τα κινητά τηλέφωνα, η τηλεόραση, οι εφημερίδες και τα tweets. Ο προσδιορισμός του κλίματος είναι ένα δημοφιλές θέμα που έχει επηρεάσει σημαντικά τη ζωή και τις δραστηριότητες των ανθρώπων. Ωστόσο, η αξιολόγηση των κλιματικών δεδομένων μπορεί να είναι ένα πρόβλημα διακριτής-συνεχούς επιλογής, καθώς η βροχόπτωση είναι δύσκολο να προβλεφθεί με ακρίβεια λόγω της ενεργητικής βροχόπτωσης και των ποικίλων φυσικών μορφών που εμπλέκονται. Οι βροχοπτώσεις αποτελούν κρίσιμη συνιστώσα της κλιματικής αλλαγής, καθώς είναι ζωτικής σημασίας πηγή νερού για τη ζωή. Η χρήση της γνώσης του κλίματος έχει επίσης σημαντική επίδραση στην κερδοφορία των καθημερινών ανθρώπινων δραστηριοτήτων, επηρεάζοντας τις οικονομικές και κοινωνικές επιπτώσεις. Στην εποχή των μεγάλων δεδομένων, πολλοί εφαρμόζουν μεθόδους πληροφόρησης με ακρίβεια και σε πραγματικό χρόνο, μεταξύ άλλων και για πληροφορίες σχετικά με τον καιρό. Ένα από τα δεδομένα που χρησιμοποιούνται είναι τα δεδομένα από τα μέσα κοινωνικής δικτύωσης, όπως το Twitter. Η παρούσα εργασία αποσκοπεί στην επεξεργασία δεδομένων του Twitter για την παροχή των πιο πρόσφατων πληροφοριών για τον καιρό στο κοινό, εξάγοντας ενδείξεις και θέματα για τον καιρό στην Ινδονησία. Πραγματοποιείται η ταξινόμηση κειμένου με χρήση δεδομένων καιρού, εξάγοντας την ένδειξη και το θέμα του καιρού από ανθρώπους σε όλη την Ινδονησία. Αρκετά μοντέλα μηχανικής μάθησης έχουν χρησιμοποιηθεί με επιτυχία για την ταξινόμηση κειμένων, συμπεριλαμβανομένων των Support Vector Machine (SVM), Multinomial Naive Bayes (MNB) και Logistic Regression (LR). Στην παρούσα μελέτη χρησιμοποιούνται τεχνικές επεξεργασίας φυσικής γλώσσας (NLP), με το Natural Language Toolkit (NLTK) να διαδραματίζει κρίσιμο ρόλο στην εκτέλεση βασικών ασκήσεων tokenization. Για την τελευταία πρόγνωση καιρού στην Ινδονησία χρησιμοποιήθηκαν δεδομένα χρονοσειρών από προηγούμενα έτη. Τα συστήματα πρόβλεψης δεξιοτήτων, όπως το μοντέλο Weather Research and Forecasting (WRF), είναι ικανά να παράγουν προβλέψεις που προσεγγίζουν τα αποτελέσματα των επιχειρησιακών προβλέψεων και να προβλέπουν ακραίες και πολύ

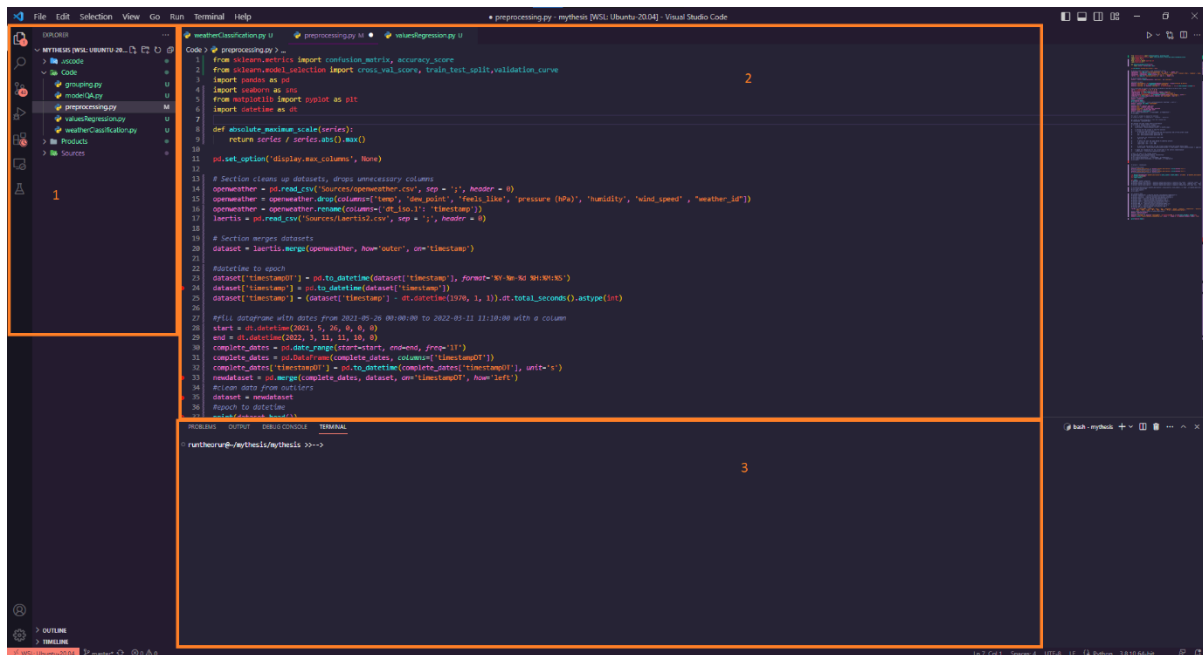
ακραίες εντάσεις βροχόπτωσης. Ωστόσο, η πρόβλεψη της ημερήσιας βροχόπτωσης, ιδίως για τις κατηγορίες βαριάς και πολύ βαριάς βροχόπτωσης, χρειάζεται ακόμη περισσότερη έρευνα και μελέτη για την παραγωγή καλής ποιότητας πρόβλεψης. Το σύνολο δεδομένων του Twitter έχει χρησιμοποιηθεί σε μερικά πειράματα για την πρόβλεψη του χρόνου και του τόπου στον οποίο είναι πιθανό να συμβεί μια συγκεκριμένη μορφή εγκλήματος. Μπορεί να διακρίνει τη σχέση μεταξύ ορισμένων σημείων με την ακρότητα της άποψης και της πραγματικής εγκληματικής δραστηριότητας. Τα πειράματα οδήγησαν στην ανακάλυψη ότι η βροχή, η ελαφριά βροχή και η δυνατή βροχή έχουν κάποιες κοινές λέξεις-κλειδιά, όπως η "βροχή", η οποία επηρεάζει τη διαδικασία υπολογισμού της τιμής TF-IDF. Το μοντέλο SVM αντιμετωπίζει μερικές φορές ασάφεια στα αποτελέσματα της ανάλυσης, ενώ οι μέθοδοι MNB και LR ενδέχεται να παράγουν διαφορετικά αποτελέσματα. [16]

5 Εργαλεία που χρησιμοποιήθηκαν

Για την παρούσα εργασία χρησιμοποιήθηκε μια μεγάλη γκάμα εργαλείων προκειμένου τα στάδια έρευνας και ανάπτυξης των κατάλληλων μεθόδων επεξεργασίας και ανάλυσης δεδομένων καθώς και η εκπαίδευση των μοντέλων να κυλούν όσο πιο ομαλά και γρήγορα γίνεται. Για τον σκοπό αυτό χρησιμοποιήθηκαν διάφορα εργαλεία για τη μελέτη, ανάλυση και προετοιμασία των δεδομένων καθώς και για την επιλογή, ανάπτυξη και εκπαίδευση των μοντέλων μηχανικής μάθησης.

Η συγγραφή του κώδικα που χρησιμοποιήθηκε έγινε μέσω δύο εργαλείων: Του Visual Studio Code και του Google Colab. Η έρευνα για την επιλογή του ιδανικότερου μοντέλου μάθησης έγινε με τη χρήση του Weka.

Το Visual Studio Code είναι editor κώδικα το οποίο αναπτύχθηκε από τη Microsoft και διατίθεται για πλατφόρμες Windows, Linux και macOS. Παρέχει τη δυνατότητα συγγραφής, εκτέλεσης, αποσφαλμάτωσης και αυτόματης συμπλήρωσης κώδικα για πολλές γλώσσες όπως οι C, C++, C#, Python, Go, Java, JavaScript και άλλες.[17] Επιλέχθηκε για διάφορους λόγους για την υλοποίηση της εργασίας, κυρίως λόγω της εξοικείωσης με τον συγκεκριμένο, καθώς και λόγω ταχύτητας, εύκολης αποσφαλμάτωσης μέσω του εργαλείου Debugger αλλά και λόγω της εύκολης διαχείρισης πακέτων και βιβλιοθηκών Python χάρη στην υποστήριξη του Windows Subsystem for Linux (WSL) από τον VS Code. Η χρήση του υποσυστήματος αυτού, που λειτουργεί ως εικονική μηχανή, επιτρέπει στην προσομοίωση ενός Linux συστήματος καταναλώνοντας λιγότερους πόρους σε σχέση με μια παραδοσιακή εικονική μηχανή, ενώ απλουστεύει τη διαχείριση των πακέτων και εργαλείων απαραίτητα για την εκτέλεση κώδικα σε σύγκριση με λύσεις που παρέχονται εγγενώς από το λειτουργικό Windows. Διατίθεται επίσης υποστήριξη Git, επιτρέποντας την χρήση εργαλείων source control (για παράδειγμα GitHub ή GitLab). Η παρακάτω εικόνα δείχνει την αρχική σελίδα του Visual Studio Code μόλις ανοίξουμε ένα project.



Εικόνα 11: Visual Studio Code

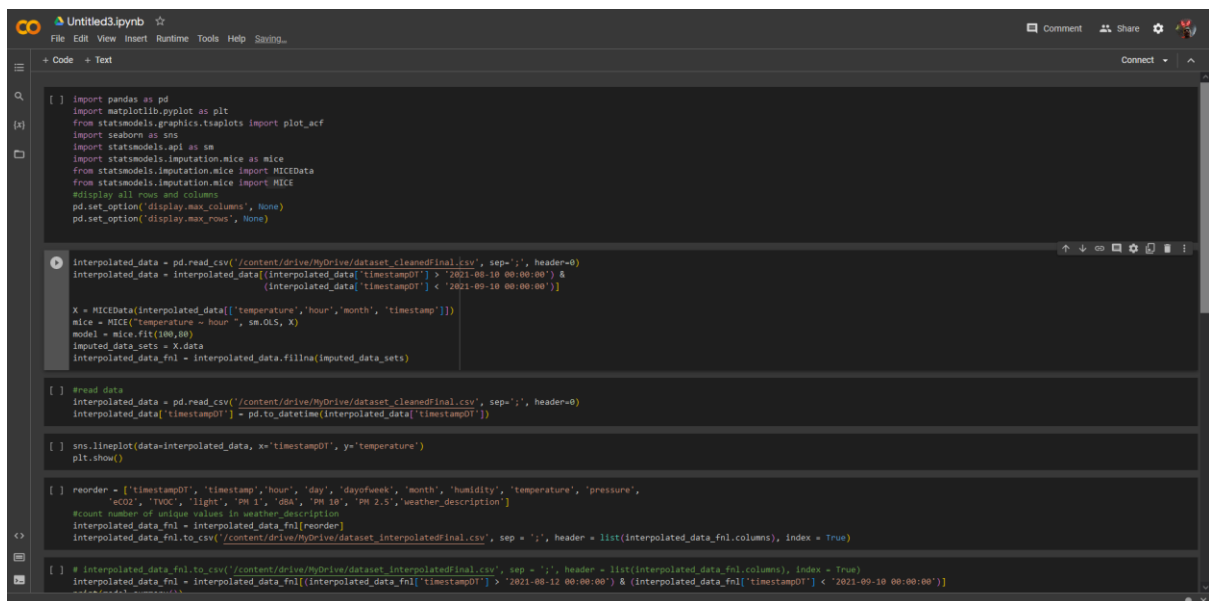
Στην περιοχή (1) βρίσκεται η πλαϊνή μπάρα, στην οποία βρίσκονται οι εξής συντομεύσεις:

- Explorer, μέσω του οποίου βλέπουμε τα ενεργά project και όλους τους φακέλους, υποφακέλους και αρχεία.
- Search, μέσω του οποίου μπορούμε να κάνουμε αναζήτηση σε όλα τα ενεργά project για συγκεκριμένες λέξεις-κλειδιά
- Source Control, στο οποίο βρίσκουμε τις αλλαγές που πραγματοποιήθηκαν και εκκρεμεί ο συγχρονισμός τους με το αποθετήριο τους στην υπηρεσία source control
- Run and Debug, μέσω του οποίου καλούμε τον debugger της αρεσκείας μας και βλέπουμε την εκτέλεση κώδικα καθώς και τις μεταβολές στις μεταβλητές, πίνακες ή και αρχεία που χρησιμοποιούνται στον κώδικα
- Extensions, όπου μπορούμε να εγκαταστήσουμε προεκτάσεις για τον Visual Studio Code, όπως πακέτα αυτόματης συμπλήρωσης κώδικα
- Remote Explorer, χρήσιμο όταν γίνεται χρήση WSL ή άλλων απομακρυσμένων συστημάτων μέσω SSH, και τέλος
- Testing, όπου μπορούμε να δημιουργήσουμε και να εκτελέσουμε μικρά κομμάτια κώδικα με πλασματικά δεδομένα, ως κομμάτι της test-driven ανάπτυξης κώδικα

Η περιοχή (2) είναι ο editor, στον οποίο ανοίγουν τα αρχεία που επιλέγουμε. Εδώ πραγματοποιείται η συγγραφή, επεξεργασία και εκτέλεση κώδικα. Τέλος στο κουτάκι 3 βρίσκουμε το παράθυρο εξόδου στο οποίο δίνονται 4 επιλογές:

- Problems: Αφορά προβλήματα σχετικά με το project
- Output: Αφορά θέματα με το version control, εφόσον γίνεται χρήση του στο project.
- Debug Console: Εδώ εμφανίζονται τα αρχεία καταγραφής του Debugger.
- Terminal: τερματικό, όπου στην περίπτωση μας είναι το τερματικό του WSL και επιτρέπει τη διαχείριση του συστήματος χωρίς να βγούμε από τον editor.

Παράλληλα με το Visual Studio Code, έγινε και συγγραφή κώδικα μέσω του Google Colab. Το Google Colab είναι η διαδικτυακή υπηρεσία παροχής υπολογιστικών πόρων μέσω του Jupyter Notebook, σε συνδυασμό με πρόσβαση σε κάρτες γραφικών που είναι ιδανικές για εφαρμογές τεχνητής νοημοσύνης χάρη στην ικανότητα τους να πραγματοποιούν μεγάλο όγκο υπολογισμών ταυτόχρονα λόγω του μεγάλου αριθμού επεξεργαστών που έχουν. Επιλέχθηκε λόγω της παροχής υπολογιστικών πόρων χωρίς χρέωση, κάτι που ωφέλησε την υλοποίηση της εργασίας λόγω του μειωμένου χρόνου εκπαίδευσης μοντέλων μηχανικής μάθησης.



```

[ ] import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.graphics.tsaplots import plot_acf
import seaborn as sns
import statsmodels.api as sm
from statsmodels.imputation.mice import mice
from statsmodels.imputation.mice import MICEData
from statsmodels.imputation.mice import MICE
#display all rows and columns
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

[ ] Interpolated_data = pd.read_csv('/content/drive/MyDrive/dataset_cleanedfinal.csv', sep=';', header=0)
Interpolated_data[Interpolated_data['timestamp0'] > '2021-08-10 00:00:00'] &
(Interpolated_data['timestamp0'] < '2021-09-10 00:00:00')

X = MICEData(Interpolated_data[['temperature', 'hour', 'month', 'timestamp']])
mice = MICE('temperature ~ hour', sm.OLS, X)
model = mice.fit(100, 10)
imputed_data_sets = X.data
Interpolated_data_fnl = Interpolated_data.fillna(imputed_data_sets)

[ ] #read data
Interpolated_data = pd.read_csv('/content/drive/MyDrive/dataset_cleanedfinal.csv', sep=';', header=0)
Interpolated_data['timestamp0'] = pd.to_datetime(Interpolated_data['timestamp0'])

[ ] sns.lineplot(data=Interpolated_data, x='timestamp0', y='temperature')
plt.show()

[ ] reorder = ['timestamp0', 'timestamp', 'hour', 'day', 'dayofweek', 'month', 'humidity', 'temperature', 'pressure',
'ec02', 'TVOC', 'light', 'PM 1', 'dBa', 'PM 10', 'PM 2.5', 'weather_description']
#count unique values in weather_description
Interpolated_data_fnl = Interpolated_data_fnl[reorder]
Interpolated_data_fnl.to_csv('/content/drive/MyDrive/dataset_interpolatedfinal.csv', sep=';', header = list(Interpolated_data_fnl.columns), index = True)

[ ] # Interpolated_data_fnl.to_csv('/content/drive/MyDrive/dataset_interpolatedfinal.csv', sep=';', header = list(Interpolated_data_fnl.columns), index = True)
Interpolated_data_fnl = Interpolated_data_fnl[(Interpolated_data_fnl['timestamp0'] > '2021-08-12 00:00:00') & (Interpolated_data_fnl['timestamp0'] < '2021-09-10 00:00:00')]

```

Εικόνα 12 Google Colab

Το Google Colab, όντας βασισμένο στο Jupyter Notebook έχει τη μορφή «κελιών» κώδικα, τα οποία μπορούν να εκτελεστούν ανεξάρτητα από τη σειρά στην οποία βρίσκονται, και τα αποτελέσματα της εκτέλεσης κάθε κελιού παραμένουν διαθέσιμα ακόμα και αν αποτύχει η εκτέλεση κάποιου άλλου, κάτι που στο Visual Studio δεν είναι εφικτό, καθιστώντας την αποτυχία εκτέλεσης χρονοβόρων διαδικασιών ακόμα πιο χρονοβόρες, και δίνεται έτσι η δυνατότητα γρήγορων αλλαγών σε παραμέτρους εκπαίδευσης του μοντέλου. Είναι επίσης εφικτή η σύνταξη κειμένου με τη χρήση της γλώσσας Markdown.

Το Google Colab δεν παρέχει ξεχωριστό τερματικό, ωστόσο είναι δυνατή η εκτέλεση εντολών τερματικού (όπως και η προσωρινή εγκατάσταση πακέτων για όσο είναι ενεργό το runtime που χρησιμοποιεί ο χρήστης) με το πρόθεμα «!», επομένως είναι εύκολη η διαχείριση των βιβλιοθηκών που απαιτούνται για την εκτέλεση κώδικα. Η δομή του είναι παρόμοια με αυτή του Visual Studio Code, καθώς στην πλαϊνή μπάρα στα αριστερά βρίσκουμε τις επιλογές για το άνοιγμα του Explorer, την αναζήτηση λέξεων-κλειδιών στον κώδικα, την προβολή των μεταβλητών μετά από κάθε εκτέλεση κελιού και τον πίνακα περιεχομένων. [18]

Για την επιλογή του μοντέλου κατηγοριοποίησης που αναπτύχθηκε, επιλέχθηκε το Weka για τα πρώτα στάδια έρευνας και πειραματισμού για το κατάλληλο μοντέλο. Το Weka, ή Waikato Environment for Knowledge Analysis είναι σουίτα λογισμικού μηχανικής μάθησης, οπτικοποίησης και προεξεπεργασίας δεδομένων γραμμένο σε Java. Αναπτύχθηκε από το πανεπιστήμιο του Waikato στη Νέα Ζηλανδία, παρέχεται δωρεάν, μπορεί να χρησιμοποιηθεί σε κάθε σύγχρονη πλατφόρμα υπολογιστών και διαθέτει ευρεία γκάμα δυνατοτήτων γύρω από την εξόρυξη δεδομένων, όπως προεπεξεργασία, ομαδοποίηση, κατηγοριοποίηση, οπτικοποίηση και παρεμβολή (regression). Στα πλαίσια της παρούσας εργασίας χρησιμοποιήθηκε μόνο για σκοπούς κατηγοριοποίησης, καθώς η προεπεξεργασία είχε υλοποιηθεί μέσω Python. Η αρχική οθόνη του Weka παρέχει 5 επιλογές:

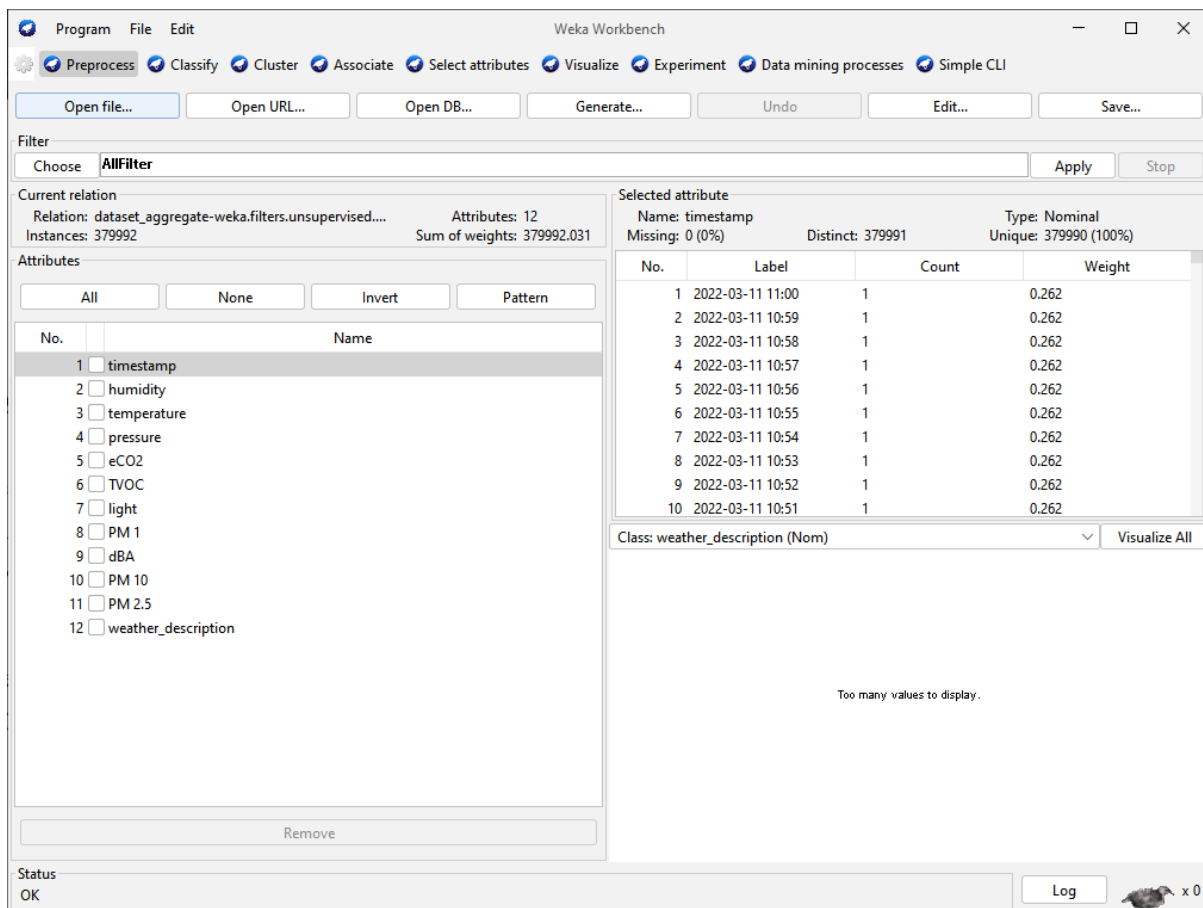


Εικόνα 13: Αρχική οθόνη Weka

- Explorer, το περιβάλλον εξερεύνησης δεδομένων του Weka

- Experimenter, το περιβάλλον πειραματισμού και υλοποίησης στατιστικών δοκιμών ανάμεσα σε διάφορα μοντέλα και μεθόδους μάθησης
- KnowledgeFlow, ένα περιβάλλον παρόμοιο με το Explorer που ωστόσο βρίσκεται σε ένα περιβάλλον γραφικών και drag-and-drop υλοποίησης
- Workbench, το εργαλείο που περιλαμβάνει όλα τα παραπάνω σε καρτέλες μέσω των οποίων ο χρήστης μπορεί να επιλέξει το εργαλείο που χρειάζεται και τέλος
- Simple CLI, το τερματικό του Weka που επιτρέπει την εκτέλεση εντολών ακόμα και σε συστήματα που δεν παρέχουν το δικό τους τερματικό

Στα πλαίσια έρευνας για την παρούσα εργασία χρησιμοποιήθηκε μόνο το workbench, καθώς επιτρέπει την γρήγορη εναλλαγή ανάμεσα στα περιβάλλοντα προετοιμασίας δεδομένων και εκπαίδευσης μοντέλων. Ακολουθεί ενδεικτική απεικόνιση του Weka Workbench, με σύνολο δεδομένων ήδη φορτωμένο.



Εικόνα 14: Weka Workbench

Είναι δυνατή η εισαγωγή δεδομένων μέσω διάφορων αρχείων (όπως csv ή ακόμα και αρχείο τύπου arff που παράγονται από το Weka), ή μέσω βάσεων δεδομένων ή μέσω ενός URL. Δίνεται επίσης η επιλογή δημιουργίας τυχαίων πλασματικών δεδομένων.

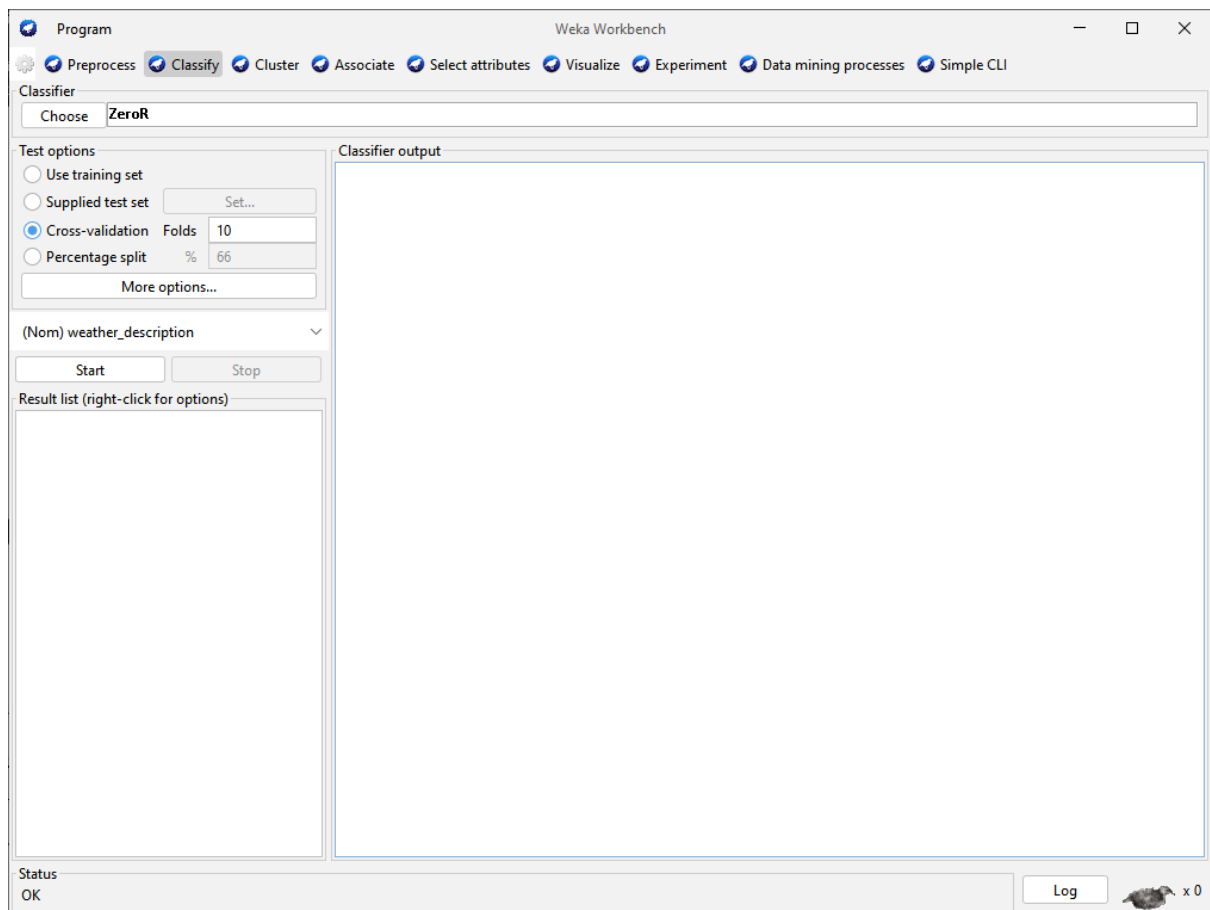
Στη μπάρα «filter» μπορούν να επιλεγθούν διάφορες διαδικασίες προεπεξεργασίας και να εφαρμοστούν στα δεδομένα, όπως για παράδειγμα η κανονικοποίηση των δεδομένων με τη μέθοδο min-max feature scaling, όπου όλα τα αριθμητικά δεδομένα υπολογίζονται με τον τύπο στην εικόνα ώστε να βρεθούν εντός του πεδίου ορισμού [0,1].[19]

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Εικόνα 15: Εξίσωση κανονικοποίησης

Γίνεται επίσης η επιλεκτική εφαρμογή φίλτρων ή/και διαδικασιών σε στήλες των δεδομένων, καθώς επίσης και η διαγραφή στηλών. Στα δεξιά της αρχικής οθόνης υπάρχει η δυνατότητα προβολής των δεδομένων της στήλης που έχει επιλεγθεί, και ανάλογα με τον τύπο δεδομένων θα φαίνονται είτε ως γράφημα μπάρας είτε ως απλό κείμενο (όπως εμφανίζεται και η στήλη timestamp δηλαδή).

Μόλις γίνουν τα απαραίτητα βήματα επεξεργασίας δεδομένων, μπορούμε να προχωρήσουμε στην καρτέλα Classify.



Εικόνα 16: Η καρτέλα Classify του Weka

Στην καρτέλα classify βρίσκουμε αρχικά το κουμπί επιλογής classifier. Το Weka παρέχει μια πλούσια γκάμα από classifiers όπως Bayes, Naive Bayes, Decision Tree Classifier, Random Tree Classifier, kNN (k πλησιέστεροι γείτονες). Αφού επιλέξουμε τον classifier μπορούμε να ορίσουμε το πως θα γίνει η φάση testing του ταξινομητή (classifier):

- «Use training set» : Ο ταξινομητής αξιολογείται με βάση το πόσο καλός είναι στην πρόβλεψη της κλάσης του παραδείγματος πάνω στο οποίο έχει εκπαιδευτεί.
- «Supplied test set» : Ο ταξινομητής αξιολογείται με βάση το πόσο καλός είναι στην πρόβλεψη της κλάσης ενός παραδείγματος που παρέχεται από ξεχωριστό αρχείο που φορτώνεται από τον χρήστη πατώντας την επιλογή «Set...»
- «Cross-validation» : Ο ταξινομητής αξιολογείται με διασταυρούμενη επικύρωση, χρησιμοποιώντας τον αριθμό των αναδιπλώσεων που εισάγονται στο πεδίο κειμένου Folds.
- «Percentage split» : Ο ταξινομητής αξιολογείται με βάση το πόσο καλά προβλέπει ένα ορισμένο ποσοστό των δεδομένων που προορίζονται για δοκιμή. Η ποσότητα

των δεδομένων που προορίζονται για τον σκοπό αυτό εξαρτάται από την τιμή που εισάγεται στο πεδίο %.

Έχοντας ορίσει όλες τις παραμέτρους που επιθυμούμε, πατάμε το Start για να ξεκινήσει η εκπαίδευση του μοντέλου. Με την έναρξη της εκπαίδευσης, η καρτέλα «Result list» εμφανίζει ως νέα καταχώρηση τον ταξινομητή που εκπαιδεύεται, και στη δεξιά καρτέλα, στην «Classifier Output» μπορούμε να παρακολουθήσουμε την εξέλιξη της εκπαίδευσης και μόλις αυτή ολοκληρωθεί να αναλύσουμε τις επιδόσεις του.

```
Time taken to build model: 23.86 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      262225.6172           69.0082 %
Incorrectly Classified Instances    117766.4143           30.9918 %
Kappa statistic                    0.6384
Mean absolute error                 0.1472
Root mean squared error             0.2515
Relative absolute error              60.1121 %
Root relative squared error         71.8747 %
Total Number of Instances          379992.0315

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,214	0,043	0,455	0,214	0,291	0,240	0,808	0,374	clouds
	0,675	0,072	0,611	0,675	0,641	0,579	0,910	0,622	sky is clear
	0,290	0,069	0,412	0,290	0,340	0,257	0,843	0,420	rain
	0,652	0,097	0,529	0,652	0,584	0,510	0,914	0,646	thunderstorm
	1,000	0,041	0,803	1,000	0,891	0,877	0,990	0,871	mist
	1,000	0,001	0,992	1,000	0,996	0,996	1,000	0,989	fog
	1,000	0,039	0,809	1,000	0,894	0,881	0,987	0,829	light intensity drizzle
Weighted Avg.	0,690	0,052	0,659	0,690	0,663	0,620	0,922	0,678	

```

=== Confusion Matrix ===

  a      b      c      d      e      f      g  <-- classified as
11635   18983.46  7412.73  8741.01  2046.34  148.23  5317.85 | a = clouds
5532.31  36618.65  1196.92  2325.72  1238.11  267.74  7105.1 | b = sky is clear
7070.71  2770.39  15743.7  20402.37  7875.29    0  422.11 | c = rain
1307.75  1582.37  13849.04  35374.55  2170.86    0    0 | d = thunderstorm
    0         0         0         0  54284.57    0    0 | e = mist
    0         0         0         0    0  54284.57    0 | f = fog
    0         0         0         0    0    0  54284.57 | g = light intensity drizzle

```

Εικόνα 17: Παράδειγμα Classifier Output μέσω του Weka

Συνοπτικά, τα κυριότερα αποτελέσματα είναι δύο: Η λεπτομερής ανάλυση ακρίβειας ανά κλάση, και ο πίνακας σύγχυσης ανά κλάση. Στη λεπτομερή ανάλυση ακρίβειας ανά κλάση βλέπουμε τις παρακάτω πληροφορίες:

- TP rate (True Positive Rate/ Ποσοστό αληθούς θετικού αποτελέσματος): Ο λόγος των σωστά ταξινομημένων θετικών περιπτώσεων προς το συνολικό αριθμό θετικών περιπτώσεων στο σύνολο δεδομένων.

- FP rate (False Positive Rate/ Ποσοστό ψευδώς θετικών αποτελεσμάτων): Ο λόγος των λανθασμένα ταξινομημένων αρνητικών περιπτώσεων προς τον συνολικό αριθμό των αρνητικών περιπτώσεων στο σύνολο δεδομένων.
- Precision (Ακρίβεια): Ο λόγος των σωστά ταξινομημένων θετικών περιπτώσεων προς τον συνολικό αριθμό των περιπτώσεων που ταξινομούνται ως θετικές. Με άλλα λόγια, μετρά το ποσοστό των θετικών προβλέψεων που είναι σωστές.
- Recall (Ανάκληση): Ο λόγος των σωστά ταξινομημένων θετικών περιπτώσεων προς το συνολικό αριθμό θετικών περιπτώσεων στο σύνολο δεδομένων. Μετρά την ικανότητα του μοντέλου να αναγνωρίζει όλες τις θετικές περιπτώσεις.
- F-measure: Ένας σταθμισμένος αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης. Παρέχει ισορροπία μεταξύ ακρίβειας και ανάκλησης.
- MCC (Matthews Correlation Coefficient / Συντελεστής συσχέτισης Matthews): Ένα μέτρο της ποιότητας της δυαδικής ταξινόμησης που λαμβάνει υπόψη τόσο τα αληθή όσο και τα ψευδή θετικά και αρνητικά αποτελέσματα.
- ROC area (Receiver Operating Characteristic): Η περιοχή κάτω από την καμπύλη ενός διαγράμματος που δείχνει το ποσοστό TP (άξονας y) έναντι του ποσοστού FP (άξονας x) καθώς μεταβάλλεται το κατώφλι απόφασης για την ταξινόμηση. Μετρά τη συνολική απόδοση του μοντέλου.
- PRC area (Precision-Recall Curve): Το εμβαδόν κάτω από την καμπύλη ενός διαγράμματος που δείχνει την ακρίβεια (άξονας y) έναντι της ανάκλησης (άξονας x) καθώς μεταβάλλεται το κατώφλι απόφασης για την ταξινόμηση. Μετρά το συμβιβασμό μεταξύ ακρίβειας και ανάκλησης.

Ακολουθεί ο πίνακας σύγχυσης, ο οποίος έχει και στους δύο του άξονες τις κλάσεις του συνόλου δεδομένων, ωστόσο ο άξονας x (οριζόντια) δείχνει τις πραγματικές κλάσεις κάθε παραδείγματος στο σύνολο δεδομένων, και ο άξονας y (κάθετα) δείχνει την κατηγοριοποίηση του κάθε παραδείγματος. Η κύρια διαγώνιος δείχνει τις αληθείς και σωστές προβλέψεις (δηλαδή το μοντέλο προέβλεψε καιρική συνθήκη «Sky is Clear» και η πραγματική κλάση ήταν ίδια με την πρόβλεψη).

Σε αυτό το σημείο έχει ουσιαστικά ολοκληρωθεί η εκπαίδευση του μοντέλου. Τα επόμενα βήματα μπορούν να είναι είτε η εκπαίδευση νέου ταξινομητή, είτε η μετάβαση στην καρτέλα Preprocess για εκ νέου επεξεργασία του συνόλου δεδομένων, εφόσον κρίνεται απαραίτητο. Εφόσον εντοπιστεί το μοντέλο και ο ταξινομητής της αρεσκείας μας, είναι

δυνατή η αποθήκευση του μοντέλου για μεταγενέστερη χρήση κάνοντας δεξί κλικ στο μοντέλο που θέλουμε να αποθηκεύσουμε και επιλέγοντας «Save Model». Η ανάκληση του μοντέλου γίνεται στην ίδια καρτέλα, κάνοντας δεξί κλικ στον χώρο «Result list» και επιλέγοντας «Load model».

Τέλος, η γλώσσα προγραμματισμού που επιλέχθηκε για την υλοποίηση του απαραίτητου κώδικα της παρούσας εργασίας είναι η Python. Η Python είναι διερμηνευόμενη γλώσσα υψηλού επιπέδου, κυριότερος στόχος της οποίας όταν αναπτύχθηκε ήταν η αναγνωσιμότητα κώδικα και η ευκολία στην εκμάθηση και χρήση. Το συντακτικό της επιτρέπει στην έκφραση εννοιών με λιγότερες γραμμές κώδικα σε σχέση με γλώσσες όπως η Java, και διαθέτει πολλές βιβλιοθήκες για τομείς όπως η ανάλυση δεδομένων, βάσεις δεδομένων, γραφικά περιβάλλοντα χρήστη, επεξεργασία εικόνας και κειμένου, μηχανική μάθηση και ακόμα περισσότερες που απλοποιούν πολλές τυπικές εργασίες. Ωστόσο, ως διερμηνευόμενη γλώσσα είναι σημαντικά πιο αργή σε σχέση με άλλες, μεταγλωττιζόμενες γλώσσες όπως η C. Με άλλα λόγια, η εκτέλεση της Python γίνεται μετατρέποντας μια γραμμή κώδικα σε κώδικα μηχανής, εκτελώντας αυτή τη γραμμή και επαναλαμβάνοντας τη διαδικασία μέχρι να εκτελεστεί ολόκληρο το αρχείο κώδικα. Στον αντίποδα μια μεταγλωττισμένη γλώσσα εκτελείται αφού έχει προηγηθεί η μετατροπή ολόκληρου του κώδικα σε γλώσσα μηχανής. Διερμηνευτές της Python διατίθενται για πολλά λειτουργικά συστήματα.[20] Στην περίπτωση της παρούσας εργασίας, επιλέχθηκε η εγκατάσταση και διαχείριση της μέσω του Windows Subsystem for Linux λόγω της απλής διαδικασίας εγκατάστασης της γλώσσας και συνοδευτικών πακέτων που παρέχουν τα Linux. Χρησιμοποιήθηκαν διάφορες βιβλιοθήκες που βοήθησαν στην αποθήκευση και επεξεργασία δεδομένων σε μορφή πινάκων (γνωστά ως dataframes στη βιβλιοθήκη pandas), την οπτικοποίηση δεδομένων (χάρη στις matplotlib και seaborn).

6 Μεθοδολογία

Στο κεφάλαιο αυτό θα παρουσιαστεί η διαδικασία επιλογής, συλλογής και επεξεργασίας των δεδομένων που είναι απαραίτητα για την ανάπτυξη των μοντέλων μηχανικής μάθησης που θα χρησιμοποιηθούν παρακάτω, καθώς και οι πηγές των δεδομένων. Θα παρουσιαστεί επίσης η διαδικασία επιλογής, ανάπτυξης και εκπαίδευσης των μοντέλων που δημιουργήθηκαν στα πλαίσια της παρούσας εργασίας. Ο στόχος της εργασίας αυτής είναι να αποδειχτεί ότι ένα σύστημα που κάνει χρήση φθηνών IoT αισθητήρων και μοντέλων τεχνητής νοημοσύνης είναι σε θέση να παρέχει καιρικές προβλέψεις.

6.1 Δεδομένα

6.1.1 Κιτ αισθητήρων Smart Citizen

Τα τελευταία χρόνια υπάρχει ερευνητικό ενδιαφέρον γύρω από τις δυνατότητες φθηνών περιβαλλοντικών αισθητήρων για την παρακολούθηση των επιπέδων μόλυνσης σε πόλεις, με διάφορες μεθόδους κλειστού κώδικα. Το σύστημα Smart Citizen παρέχει μια λύση ανοιχτού κώδικα για το ερευνητικό αυτό πεδίο προσφέροντας μια ευέλικτη και εύκολη στη χρήση της λύση παρακολούθησης διάφορων δεικτών όπως μονοξείδιο του άνθρακα ή επίπεδα θορύβου.

Η βασική πηγή δεδομένων για αυτή την εργασία παρέχεται από ένα Smart Citizen Kit σε συνδυασμό με το Urban Board και αισθητήρα αιωρούμενων σωματιδίων του Εργαστηρίου Υπολογιστικής Μοντελοποίησης του Ιονίου πανεπιστημίου, και συλλέγει πληροφορίες για:

- Θερμοκρασία
- Υγρασία
- Ατμοσφαιρική πίεση
- Επίπεδα διαφόρων σωματιδίων στην ατμόσφαιρα (PM10, PM2.5, PM1, CO2)
- Επίπεδα θορύβου (σε dB)
- Επίπεδα φωτός (σε lux)

Τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα εργασία αφορούν το διάστημα 2021-06-09 έως 2022-03-11, ωστόσο με κάποια κενά ενδιάμεσα.

Το kit συλλέγει δεδομένα κάθε λεπτό και η αποθήκευσή τους γίνεται μέσω της αποστολής των δεδομένων στο κεντρικό αποθετήριο του Smart Citizen, προσβάσιμο από τον διαχειριστή του δικτύου στο οποίο ανήκει το kit, από το οποίο τα δεδομένα μπορούν να αποθηκευτούν σε μορφή CSV. Παρακάτω βλέπουμε ένα ενδεικτικό δείγμα των δεδομένων. Σημειώνεται πως τα ονόματα των στηλών στα δεδομένα είναι εξαιρετικά

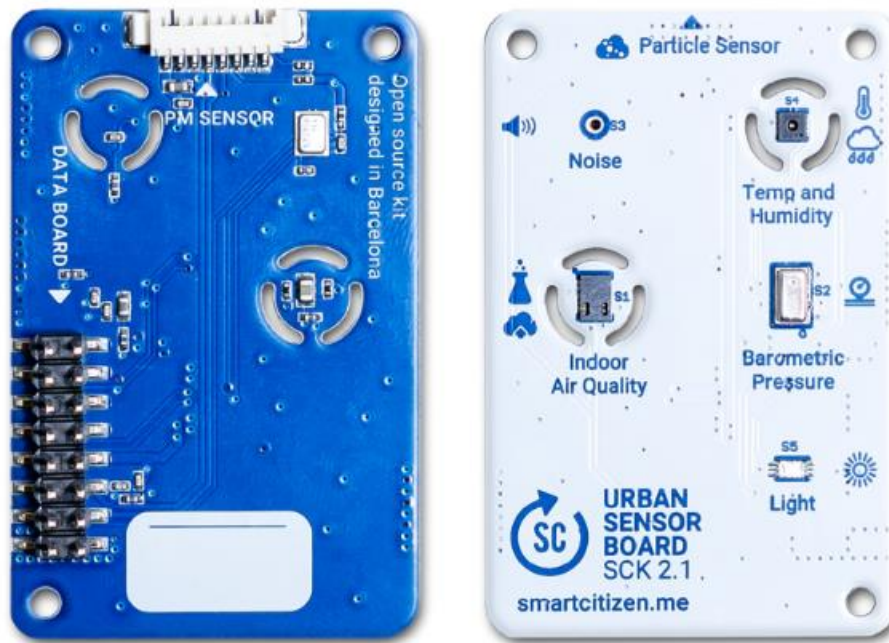
λεπτομερή, καθώς δίνουν ανά είδος πληροφορίας τη μονάδα μέτρησης καθώς και τον κωδικό του αισθητήρα που συλλέγει την πληροφορία αυτή.

```
timestamp.1;humidity in % (SHT31 - Humidity);air temperature in °C (SHT31 - Temperature);barometric pressure in kPa (MPL3115A2 - Barometric Pressure)
2022-03-11 11:10;47.71;9.21;102.89;2160;1045;951;9;84.08;9;9
2022-03-11 11:09;48.58;9.21;102.88;2164;1058;917;20;84.24;40;36
2022-03-11 11:08;47.87;9.13;102.89;2178;1088;821;7;83.53;14;13
2022-03-11 11:07;47.16;9.06;102.89;2183;1107;714;5;84.75;7;7
2022-03-11 11:06;46.61;9.13;102.89;2183;1107;750;3;84.6;9;9
2022-03-11 11:05;46.69;9.13;102.9;2178;1088;738;5;84.08;7;7
2022-03-11 11:04;47.16;9.21;102.9;2172;1075;752;9;84.81;13;13
2022-03-11 11:03;47.37;9.21;102.9;2172;1075;758;6;84.06;11;11
2022-03-11 11:02;47.32;9.13;102.9;2183;1107;787;5;85.1;10;10
2022-03-11 11:01;46.44;9.08;102.9;2199;1138;778;3;84.02;6;6
2022-03-11 11:00;45.61;9.18;102.9;2199;1138;831;9;84.43;14;13
2022-03-11 10:59;45.65;9.21;102.9;2199;1138;814;6;84.44;11;10
2022-03-11 10:58;45.44;9.28;102.91;2183;1107;826;7;84.37;13;13
2022-03-11 10:57;46.56;9.29;102.91;2183;1107;855;5;84.05;11;10
2022-03-11 10:56;46.47;9.29;102.9;2183;1107;885;5;84.09;11;9
```

Εικόνα 18: Ενδεικτικό δείγμα δεδομένων ενός Smart Citizen Kit σε μορφή CSV

Υπάρχει, άτυπα ωστόσο, και η δυνατότητα συλλογής των δεδομένων μέσω ενός API. Με ένα request URL της μορφής «<https://api.smartcitizen.me/v0/devices/xxx>» και το μοναδικό αναγνωριστικό του αισθητήρα που μας ενδιαφέρει μπορούμε να αποκτήσουμε πρόσβαση σε δεδομένα μορφής JSON, τα οποία ωστόσο δε διαθέτουν καμία ιστορικότητα. Είναι δυνατόν μέσω προγραμματιστικών λύσεων ωστόσο η αυτοματοποιημένη συλλογή των δεδομένων, κάτι που θα μπορούσε να επιτρέψει τη συλλογή δεδομένων από πολλούς σταθμούς χωρίς να είναι απαραίτητη η μεσολάβηση του διαχειριστή του εκάστοτε δικτύου αισθητήρων.

Το kit που χρησιμοποιήθηκε για τη συλλογή δεδομένων περιλαμβάνει και το Urban Sensor Board σε συνδυασμό με αισθητήρα αιωρούμενων σωματιδίων, το οποίο είναι υπεύθυνο για τη συλλογή των δεδομένων που χρησιμοποιήθηκαν για την εργασία αυτή.



Εικόνα 19: Η μπροστινή και πίσω όψη του Urban Sensing Board

```

{
  id: 14123,
  uuid: "82b99100-5677-42a3-92bc-f5865fc4c7a",
  name: "LAERTIS_2",
  description: "Smart Citizen Kit 2.1 with Urban Sensor Board",
  state: "has_published",
  postprocessing: null,
  hardware_info: {},
  system_tags: [],
  user_tags: [],
  is_private: false,
  notify_low_battery: false,
  notify_stopped_publishing: false,
  last_reading_at: "2023-02-23T18:43:02Z",
  added_at: "2021-05-26T08:01:12Z",
  updated_at: "2023-02-23T03:00:20Z",
  mac_address: "[FILTERED]",
  owner: {},
  data: {
    recorded_at: "2023-02-23T18:43:02Z",
    added_at: "2023-02-23T18:43:02Z",
    location: {
      ip: null,
      exposure: "outdoor",
      elevation: null,
      latitude: 39.6253346,
      longitude: 19.9192807,
      geohash: "srn1mm8f2c",
      city: "Corfu",
      country_code: "GR",
      country: "Greece"
    },
    sensors: {
      0: {
        id: 113,
        ancestry: "111",
        name: "AMS CCS811 - TVOC",
        description: "Total Volatile Organic Compounds Digital Indoor Sensor",
        unit: "ppb",
        created_at: "2019-03-21T16:43:37Z",
        updated_at: "2019-03-21T16:43:37Z",
        measurement_id: 47,
        uuid: "0c2a1afc-dc08-4066-aacb-0bde6a3ae6f5",
        value: 171
      }
    }
  }
}

```

Εικόνα 20: Ενδεικτικό δείγμα δεδομένων ενός Smart Citizen Kit σε μορφή JSON

6.1.2 Ιστορικό σύνολο δεδομένων μέσω Openweather

Εφόσον ο στόχος ενός εκ των δύο μοντέλων της παρούσας εργασίας είναι η κατηγοριοποίηση καιρικών φαινομένων, είναι απαραίτητο να υπάρχει η πληροφορία καιρικών φαινομένων στα δεδομένα εκπαίδευσης. Χρειάζεται λοιπόν μια πηγή που μας παρέχει πληροφορία αντίστοιχη σε ημέρες με αυτή που συλλέχθηκε από τους αισθητήρες Smart Citizen. Κατά τη διάρκεια της έρευνας για τα δεδομένα αυτά, η υπηρεσία παροχής ιστορικών δεδομένων της Εθνικής Μετεωρολογικής Υπηρεσίας δεν ήταν διαθέσιμη. Μετά από μια σύντομη έρευνα, εντοπίστηκε το Openweather, το οποίο παρέχει δεδομένα με την ίδια μορφή που παρέχονται από το Smart Citizen (και σε μορφή JSON μέσω του API της ιστοσελίδας αλλά και μέσω CSV). Σε αυτό το αρχείο βρίσκεται ωριαία πληροφορία σχετικά με :

- Ημερομηνία και ώρα
- Θερμοκρασία
- Σημείο δρόσου
- Αίσθηση θερμοκρασίας
- Ατμοσφαιρική πίεση
- Υγρασία
- Ταχύτητα ανέμου
- Κωδικός καιρικού φαινομένου
- Περιγραφή καιρικού φαινομένου

Σημειώνεται πως τα ιστορικά δεδομένα παρέχονται ανά ώρα και όχι ανά λεπτό, επομένως θα χρειαστεί να συμπληρωθούν προγραμματιστικά τα κενά αυτά μόλις ολοκληρωθεί η δημιουργία του ενιαίου συνόλου δεδομένων.

Ωστόσο, τα πεδία που είναι απαραίτητα για την εκπαίδευση του μοντέλου είναι μόνο η ημερομηνία και ώρα καθώς και ο κωδικός ή/και η περιγραφή καιρικών φαινομένων, τα οποία ήταν και τα μόνα που παρέμειναν στο σύνολο δεδομένων. Καθώς ο σκοπός της πτυχιακής εργασίας είναι η ανάπτυξη μοντέλου πρόγνωσης καιρικών φαινομένων με βάση τα δεδομένα που συλλέχθηκαν από το Smart Citizen Kit, οι στήλες ταχύτητας ανέμου και σημείου δρόσου απορρίφθηκαν από το τελικό σύνολο δεδομένων. Η πληροφορία που μας ενδιαφέρει περισσότερο, η περιγραφή καιρικού φαινομένου, διατίθεται τόσο σε μορφή κειμένου όσο και σε αριθμητική μορφή. Χάριν ευκολότερης ανάγνωσης των αποτελεσμάτων επιλέχθηκε η διατήρηση της στήλης Weather Description και η διαγραφή της Weather ID.

Συγκεκριμένα, οι κατηγορίες στην περιγραφή καιρικού φαινομένου είναι οι εξής:

- scattered clouds
- broken clouds
- few clouds
- sky is clear
- moderate rain
- light rain
- overcast clouds
- thunderstorm with light rain
- thunderstorm with rain
- thunderstorm
- mist
- fog
- light intensity drizzle
- heavy intensity rain
- thunderstorm with heavy rain
- light intensity shower rain

6.1.3 Προετοιμασία συνόλου δεδομένων

Η προετοιμασία του συνόλου δεδομένων πραγματοποιήθηκε μέσω της γλώσσας προγραμματισμού Python και διαφόρων βοηθητικών πακέτων της.

Το πρώτο στάδιο είναι η αρχική εκκαθάριση των αρχείων και η προετοιμασία για τη συνένωση.

```
openweather = pd.read_csv('Sources/openweather.csv', sep = ';', header = 0)
openweather = openweather.drop(columns=['temp', 'dew_point', 'feels_like', 'pressure (hPa)',
'humidity', 'wind_speed', 'weather_id'])
openweather = openweather.rename(columns={'dt_iso.1': 'timestamp'})
laertis = pd.read_csv('Sources/Laertis2.csv', sep = ';', header = 0)
laertis.columns =
["timestamp", "humidity", "temperature", "pressure", "eCO2", "TVOC", "light", "PM1", "dBA", "PM10",
"PM2.5"]
```

Εικόνα 21: Εισαγωγή δεδομένων και πρώτα βήματα προετοιμασίας

Στο παραπάνω κομμάτι κώδικα γίνονται τρεις διαδικασίες: εισαγωγή δεδομένων σε dataframe, διαγραφή περιττών/διπλότυπων στηλών (κυρίως στο σύνολο δεδομένων του OpenWeather) και μετονομασία στηλών (τόσο μέσω της συνάρτησης rename όσο και μέσω της δημιουργίας του πίνακα dict που περιέχει τις απλουστευμένες ονομασίες στηλών του συνόλου δεδομένων του Smart Citizen).

Στη συνέχεια, γίνεται η δημιουργία μιας επιπλέον στήλης μέσω της μετατροπής των στηλών timestamp σε μορφή unix epoch (η οποία μορφή μετράει πόσα δευτερόλεπτα πέρασαν από την 1^η Ιανουαρίου 1970).

```
openweather['epoch'] = pd.to_datetime(openweather['timestamp']).astype(int) / 10**9  
laertis['epoch'] = pd.to_datetime(laertis['timestamp']).astype(int) / 10**9
```

Εικόνα 22: Δημιουργία στηλών Unix epoch

Η κίνηση αυτή επιλέχθηκε με το σκεπτικό της γενικότερης απλούστευσης των επόμενων βημάτων: της συνένωσης, επεξεργασίας και ευκολότερης εκπαίδευσης των μοντέλων, καθώς κατά την εκτέλεση του κώδικα για την εκπαίδευση του μοντέλου αντιμετωπίστηκε από την Python πρόβλημα στη μετατροπή της στήλης timestamp (που ήταν της μορφής EEEE-MM-HHTΩΩ-ΛΛ-ΔΔ) από string σε float, επομένως κρίθηκε ορθή η μετατροπή της ημερομηνίας σε μορφή που μπορεί να χειριστεί το μοντέλο.

Στη συνέχεια έγινε η συνένωση των δύο συνόλων με την παρακάτω γραμμή κώδικα:

```
dataset = laertis.merge(openweather, how='outer', on='timestamp')
```

Εικόνα 23: Συνένωση συνόλων δεδομένων

Η συνένωση γίνεται χωρίς την αναζήτηση κοινών καταχωρήσεων (λόγω της μεθόδου outer), παρόλα αυτά σε περίπτωση διπλοεγγραφών στη στήλη timestamp (τη μόνη κοινή ανάμεσα στα δύο σύνολα δεδομένων) θα πραγματοποιηθεί συγχώνευση των καταχωρήσεων.

Με τη συνένωση των δύο συνόλων, γίνεται απαραίτητη η διαδικασία συμπλήρωσης των κενών στη στήλη weather_description καθώς η πληροφορία αυτή παρέχεται ανά λεπτό αντί για ωριαία.

```
dataset['weather_description'] = dataset['weather_description'].fillna(method='ffill')  
dataset['weather_description'] = dataset['weather_description'].fillna(method='bfill')
```

Εικόνα 24: Συμπλήρωση κενών στη στήλη weather_description

Με την χρήση της εντολής fillna και τη μέθοδο fill, η οποία γεμίζει κάθε κενό με την τελευταία γνωστή τιμή που εντοπίζει ανά κενό και ανά στήλη. Ένα ελάττωμα της ffill είναι ότι βασίζεται στην ύπαρξη επόμενης τιμής-για να συμπληρώσει το κενό. Αντίστοιχα, η bfill βασίζεται στην ύπαρξη προηγούμενης τιμής για να γίνει η συμπλήρωση, επομένως όποια και από τις δύο να επιλεγεί για τις ανάγκες του συγκεκριμένου συνόλου δεδομένων θα οδηγήσει σε σχεδόν ίδια αποτελέσματα. Στον πίνακα 3-1 γίνεται επίδειξη

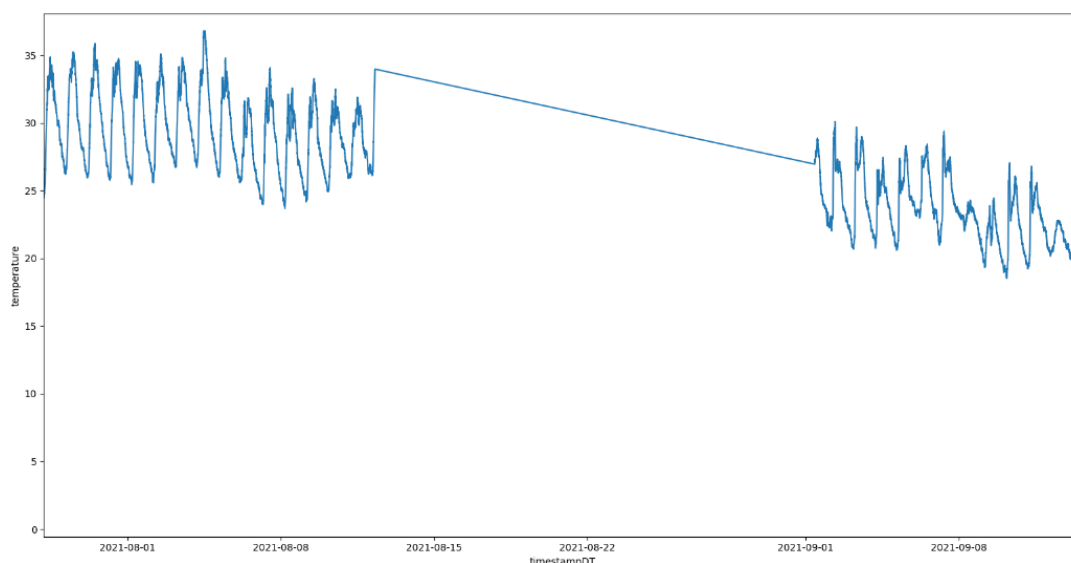
του πως λειτουργούν οι δύο μέθοδοι. Στη στήλη A1 χρησιμοποιήθηκε η μέθοδος fill ενώ στην B1 η μέθοδος bfill. Η κίνηση αυτή γίνεται με βάση την παραδοχή πως τα καιρικά φαινόμενα δεν θα είναι απόλυτα σωστά σε επίπεδο λεπτού της ώρας, ωστόσο δεν υπάρχει κάποιος τρόπος να γίνει μαζική συλλογή των δεδομένων αυτών, ούτε και παρέχονται πιο λεπτομερή δεδομένα από το OpenWeather.

A	B	A1	B1
1	N/A	1	2
N/A	2	1	2
N/A	N/A	1	4
4	4	4	4
N/A	5	4	5
6	N/A	6	N/A

Εικόνα 25: Παράδειγμα μεθόδων bfill και fill

Μαζί με τη διαχείριση των κενών που προκύπτουν από τη διαφορά στην συχνότητα των δεδομένων του Smart Citizen και του Openweather, άλλο ένα σημαντικό κομμάτι στην προετοιμασία του μοντέλου είναι η διαχείριση των μεγάλων κενών που προκύπτουν από την έλλειψη δεδομένων του Smart Citizen.

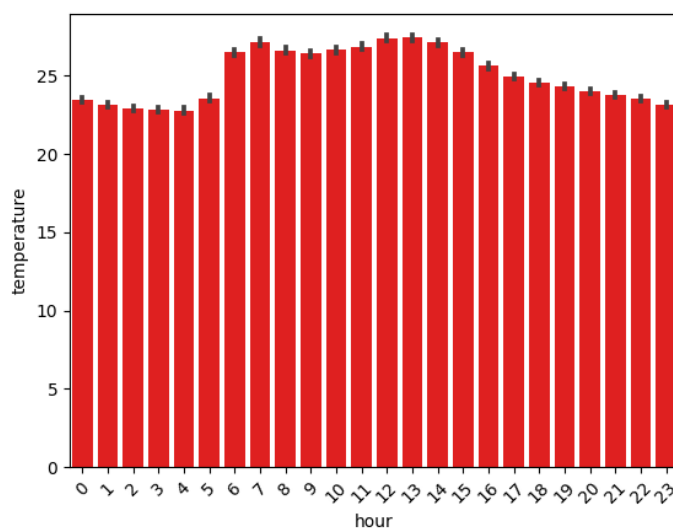
Στο τελικό σύνολο δεδομένων παρουσιάζεται ένα κενό από τις 13 Αυγούστου 2021 έως και την 1η Σεπτεμβρίου 2021, όπως φαίνεται και στο διάγραμμα θερμοκρασίας για εκείνη την περίοδο. Η διαχείριση του κενού αυτού μπορεί να γίνει είτε με συμπλήρωση των δεδομένων που λείπουν είτε με την αφαίρεση των διαστημάτων όπου παρατηρείται έλλειψη δεδομένων.



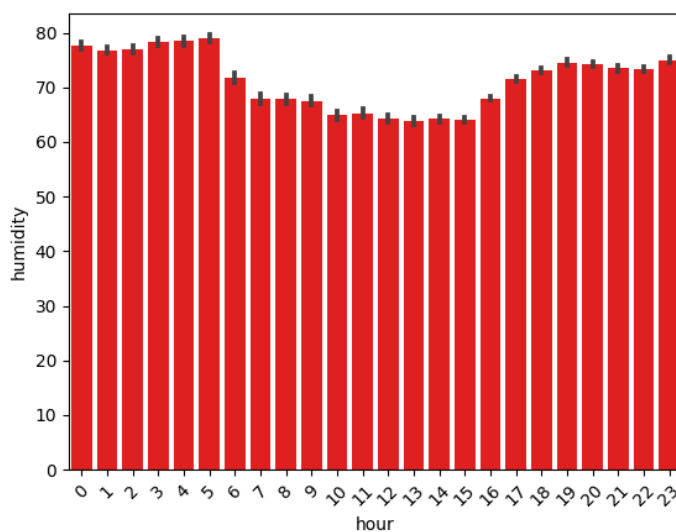
Εικόνα 26: Γράφημα δεδομένων θερμοκρασίας για το διάστημα Αύγουστος-Σεπτέμβριος 2021

Για την συμπλήρωση αυτών των κενών μελετήθηκαν διάφορες μέθοδοι, όπως γραμμική (linear) και λογιστική (logistic) παρεμβολή. Μελετήθηκε επίσης και η τεχνική πολλαπλών υπολογισμών με αλυσιδωτές εξισώσεις (Multiple Imputation with Chained equations, ή MICE)[21].

Οι τεχνικές γραμμικής και λογιστικής παρεμβολής δεν κρίθηκαν κατάλληλες καθώς δεν θα ήταν σε θέση να μνηθούν τα μοτίβα που ακολουθούν τα καιρικά φαινόμενα σε επίπεδο 24ώρου, όπως μπορούμε να δούμε και στα παρακάτω γραφήματα.



Εικόνα 27: Μέση ωριαία θερμοκρασία στο διάστημα 13-08-2021 με 1-9-2021



Εικόνα 28: Μέση ωριαία υγρασία στο διάστημα 13-8-2021 με 1-9-2021

Έτσι αποφασίστηκε η υλοποίηση του μοντέλου MICE ως μια προσπάθεια τόσο συμπλήρωσης των κενών στα δεδομένα, η οποία και θα αναλυθεί παρακάτω. Ωστόσο, λόγω ανακρίβειας στα αποτελέσματα που προέκυπταν, αποφασίστηκε η διαγραφή των κενών από πληροφορία ημερομηνιών στο τελικό σύνολο δεδομένων, ώστε να επιταχυνθεί η διαδικασία ανάπτυξης του μοντέλου.

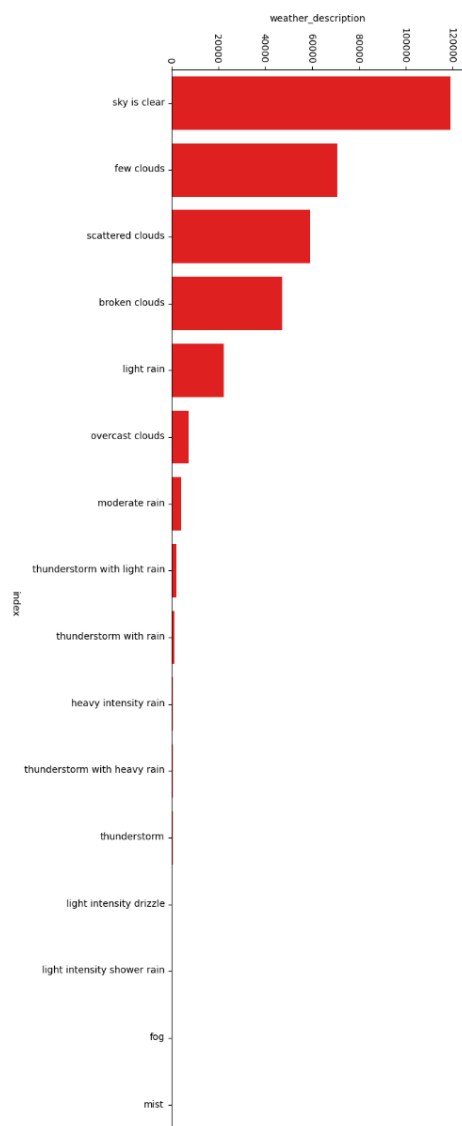
Στο στάδιο πειραματισμού στο Weka έγινε αντιληπτό το γεγονός ότι το μεγάλο σύνολο κλάσεων (δηλαδή το weather description) οδηγούσε σε τεχνητά μειωμένη ακρίβεια, λόγω των πολλαπλών κλάσεων στο σύνολο δεδομένων και των συχνοτήτων εμφάνισης τους όπως φαίνεται στους παρακάτω πίνακες. Ενώ δηλαδή επίσημα η ακρίβεια του συγκεκριμένου συνδυασμού μοντέλου-συνόλου δεδομένων δεν ξεπερνά το 50%, πρακτικά οι κλάσεις που αντιστοιχούν σε νεφελώδη καιρό αν ομαδοποιηθούν θα ωθήσουν το μοντέλο σε ορθότερες κατηγοριοποιήσεις. Έτσι δημιουργήθηκαν δύο εκδόσεις του ίδιου συνόλου δεδομένων με τη μόνη διαφορά ανάμεσα τους να είναι το σύνολο των κλάσεων, από 16 αρχικά σε 6 στο νέο σύνολο δεδομένων σύμφωνα και με τον πίνακα που ακολουθεί και χρησιμοποιήθηκαν για την εκπαίδευση του ίδιου μοντέλου υπό τις ίδιες συνθήκες (Multiclass Classifier με 10 αναδιπλώσεις).

```
dataset['weather_description'] = dataset['weather_description'].replace(['few clouds', 'scattered clouds', 'broken clouds', 'overcast clouds'], 'clouds')
dataset['weather_description'] = dataset['weather_description'].replace(['light rain', 'moderate rain', 'heavy intensity rain', 'light intensity shower rain', 'light intensity drizzle'], 'rain')
dataset['weather_description'] = dataset['weather_description'].replace(['thunderstorm with light rain', 'thunderstorm with rain', 'thunderstorm with heavy rain', 'thunderstorm'], 'thunderstorm')
```

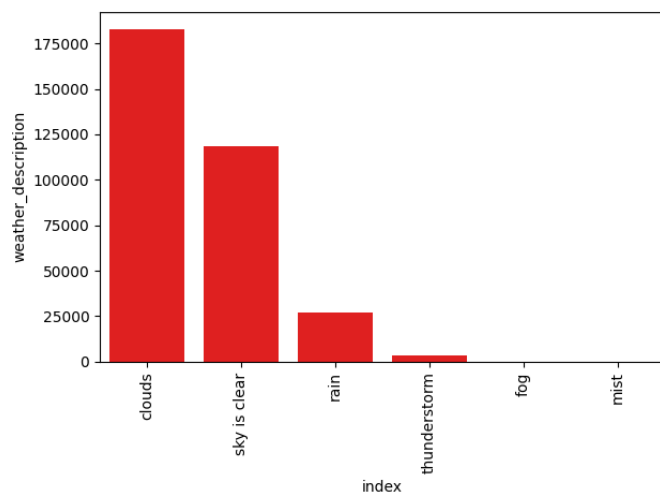
Εικόνα 29: Ομαδοποίηση καιρικών φαινομένων

Αρχικό Weather Description	Ομαδοποιημένο Weather Description
moderate rain	Rain
light rain	
light intensity drizzle	
heavy intensity rain	
thunderstorm with heavy rain	
light intensity shower rain	
sky is clear	Sky is clear
overcast clouds	Clouds
scattered clouds	
broken clouds	
few clouds	
thunderstorm with light rain	Thunderstorm
thunderstorm with rain	
thunderstorm	
mist	Mist
fog	Fog

Εικόνα 30: Κλάσεις συνόλου δεδομένων πριν και μετά την ομαδοποίηση



Εικόνα 31: Κατανομή παρατηρήσεων πριν την ομαδοποίηση



Εικόνα 32: Κατανομή παρατηρήσεων μετά την ομαδοποίηση

Ο αρχικός πίνακας σύγχυσης αποδεικνύει πως το μοντέλο πράγματι έχει δυσκολία στην εξειδικευμένη κατηγοριοποίηση των φαινομένων, και έτσι προκύπτει η ανάγκη πειραματισμού με το σύνολο δεδομένων που περιέχει ομαδοποιημένες κλάσεις.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	<-- classified as	
16370	12356	5563	27102	1	1472	0	0	0	0	0	4	0	0	0	0	0	a=scattered clouds
10834	18919	3013	12622	0	4312	1	0	0	0	0	0	2	0	0	0	0	b=broken clouds
10936	5692	6729	49763	0	612	0	0	0	0	0	8	1	0	0	0	0	c=few clouds
5722	3151	5459	109693	0	27	0	0	0	0	0	36	0	0	0	0	1	d=sky is clear
376	1844	48	204	0	2316	0	0	0	0	0	0	0	0	0	0	0	e=moderate rain
4411	10220	925	2141	0	7607	0	0	0	0	0	0	0	0	0	0	0	f=light rain
1792	2054	1166	2680	0	546	0	0	0	0	0	0	0	0	0	0	0	g=overcast clouds
347	930	13	73	0	847	0	0	0	0	0	0	0	0	0	6	0	h=thunderstorm with light rain
316	315	2	6	1	505	0	0	0	0	0	10	0	0	0	1	0	i=thunderstorm with rain
44	244	0	62	0	0	1	0	0	0	0	0	0	0	0	0	0	j=thunderstorm
0	0	0	0	0	22	0	0	0	0	0	7	0	0	0	0	0	k=mist
7	0	7	19	0	0	0	0	0	0	0	84	0	0	0	0	0	l=fog
0	0	0	177	0	0	0	0	0	0	0	0	0	0	0	0	0	m=light intensity drizzle
118	371	1	2	0	116	0	0	0	0	0	0	0	0	0	1	0	n=heavy intensity rain
161	92	0	3	0	174	0	0	0	0	0	0	0	0	0	5	0	o=thunderstorm with heavy rain
93	10	40	35	0	0	0	0	0	0	0	0	0	0	0	0	0	p=light intensity shower rain

Εικόνα 33: Πίνακας σύγχυσης του αρχικού συνόλου δεδομένων

Πράγματι, με το ομαδοποιημένο σύνολο δεδομένων παρουσιάζεται ο παρακάτω πίνακας σύγχυσης, ο οποίος παρότι δεν παρουσιάζει την εικόνα του ιδανικού μοντέλου, σίγουρα παρουσιάζει μεγαλύτερη ακρίβεια στις προβλέψεις του.

a	b	c	d	e	f	g	<-- classified as
11630	18977	7412	8751	2048	149	5318	a = clouds
5526	36615	1197	2329	1242	269	7108	b = sky is clear
7069	2769	15778	20366	7880	0	422	c = rain
1334	1582	13744	35427	2197	0	0	d = thunderstorm
0	0	0	0	54285	0	0	e = mist
0	0	0	0	0	54285	0	f = fog
0	0	0	0	0	0	54285	g = light intensity drizzle

Εικόνα 34: Πίνακας σύγχυσης του ομαδοποιημένου συνόλου δεδομένων

Πλέον η ακρίβεια στις προβλέψεις του μοντέλου πλησιάζουν το 70%, χωρίς καμία επιπλέον αλλαγή στο σύνολο δεδομένων ή το μοντέλο, θυσιάζοντας όμως τις εξειδικευμένες κατηγορίες καιρού που παρέχει το OpenWeather.

Ωστόσο, στο τελικό μοντέλο που δημιουργήθηκε εξ ολοκλήρου μέσω Python το μοντέλο είχε ικανοποιητικές προβλέψεις ακόμα και χωρίς την ομαδοποίηση, επομένως χάριν πιο λεπτομερών προβλέψεων η ομαδοποίηση δε χρησιμοποιήθηκε στο τελικό σύνολο δεδομένων. Σε αυτό το σημείο έχει ολοκληρωθεί η προετοιμασία του συνόλου δεδομένων. Πραγματοποιούμε μια αποθήκευση του με την παρακάτω εντολή, και μπορούμε να προχωρήσουμε στην ανάπτυξη των μοντέλων, μετά από μία ανακατανομή των στηλών στο dataframe.

```
reorder = ['timestampDT', 'timestamp', 'hour', 'day', 'dayofweek', 'month', 'humidity', 'temperature', 'pressure',
           'eCO2', 'TVOC', 'light', 'PM 1', 'dBa', 'PM 10', 'PM 2.5', 'weather_description']
dataset = dataset[reorder]
dataset.to_csv('Products/dataset_cleanedFinal.csv', sep = ';', header = list(dataset.columns), index = True)
```

Εικόνα 35: Ανακατανομή στήλων και αποθήκευση δεδομένων

6.2 Μοντέλα

Σε αυτό το κεφάλαιο της εργασίας, θα ξεκινήσουμε μια ολοκληρωμένη ανάλυση του συνεχώς εξελισσόμενου τοπίου των μοντέλων τεχνητής νοημοσύνης και μηχανικής μάθησης, εστιάζοντας σε δύο διαφορετικά αλλά σημαίνοντα μοντέλα που αναπτύχθηκαν για συγκεκριμένες εφαρμογές. Το πρώτο υπό εξέταση μοντέλο είναι ένας ταξινομητής δέντρων απόφασης που εκπαιδεύεται για την πρόβλεψη των καιρικών συνθηκών, χρησιμοποιώντας δεδομένα που συλλέγονται μέσω του Smart Citizen Kit. Αυτό το μοντέλο, αξιοποιώντας τη δύναμη των αλγορίθμων μηχανικής μάθησης, στοχεύει στην παροχή ακριβών και έγκαιρων προβλέψεων, επιτρέποντας τη λήψη τεκμηριωμένων αποφάσεων με βάση τα καιρικά πρότυπα. Το δεύτερο μοντέλο, γνωστό ως Multiple Imputation by Chained Equations (MICE), αναπτύχθηκε ως πιλοτικό έργο για την αντιμετώπιση κενών σε σύνολα δεδομένων. Χρησιμοποιώντας μια εξελιγμένη τεχνική καταλογισμού, το μοντέλο MICE συμπληρώνει τα σημεία δεδομένων που λείπουν, βελτιώνοντας έτσι την ποιότητα και την πληρότητα των συνόλων δεδομένων που χρησιμοποιούνται για την ανάλυση. Μέσω μιας σχολαστικής εξέτασης αυτών των δύο μοντέλων, των διαδικασιών ανάπτυξής τους και των πρακτικών επιπτώσεών τους, στοχεύουμε να αποκτήσουμε βαθύτερη κατανόηση των δυνατοτήτων, των περιορισμών και των πιθανών επιπτώσεών τους σε διάφορους τομείς. Επιπλέον, αυτή η διερεύνηση θα συμβάλει στην ευρύτερη κατανόηση των εξελίξεων της τεχνητής νοημοσύνης/μηχανικής μάθησης και θα ανοίξει το δρόμο για μελλοντικές ερευνητικές προσπάθειες για την ανάπτυξη πιο ισχυρών και αποδοτικών μοντέλων.

6.2.1 Μοντέλο MICE

Η τεχνική πολλαπλών υπολογισμών με αλυσιδωτές εξισώσεις βασίζεται στην υπόθεση πως τα δεδομένα που λείπουν είναι τυχαία (ή missing at random, MAR) ωστόσο μπορούν να ανακατασκευαστούν από άλλες γνωστές πληροφορίες (όπου στην περίπτωση μας οι γνωστές πληροφορίες αυτές θα μπορούσαν να είναι η ώρα της ημέρας καθώς και ο μήνας.) Στη διαδικασία MICE, εκτελείται μια σειρά μοντέλων παλινδρόμησης, σύμφωνα με την οποία κάθε μεταβλητή με ελλείποντα δεδομένα μοντελοποιείται υπό την προϋπόθεση ότι εξαρτάται από τις άλλες μεταβλητές των δεδομένων. Η MICE είναι μια εναλλακτική, ευέλικτη προσέγγιση στις διαδικασίες πολλαπλού υπολογισμού, που χρησιμοποιείται σε σύνολα δεδομένων με χιλιάδες παρατηρήσεις και εκατοντάδες

μεταβλητές. Περιλαμβάνει μια σειρά μοντέλων παλινδρόμησης, όπου κάθε μεταβλητή με ελλείποντα δεδομένα μοντελοποιείται υπό την επίδραση των άλλων μεταβλητών στα δεδομένα. Τα πακέτα λογισμικού ποικίλλουν ως προς την εφαρμογή του MICE, με ορισμένα πακέτα να χρησιμοποιούν επίσης ένα πολωνυμικό μοντέλο logit για κατηγορικές μεταβλητές και ένα μοντέλο Poisson για μεταβλητές καταμέτρησης. Η διαδικασία της αλυσιδωτής εξίσωσης μπορεί να αναλυθεί σε τέσσερα γενικά βήματα: Το πρώτο βήμα είναι ένας αρχικός υπολογισμός των τιμών που λείπουν με κάποιες προσωρινές τιμές (που προκύπτουν συνήθως από τον μέσο όρο των δεδομένων που έχουμε διαθέσιμα). Στο δεύτερο βήμα, οι προσωρινές τιμές μιας μεταβλητής κρύβονται (άρα φαίνονται πάλι ως MAR), και στο τρίτο βήμα υπολογίζονται εκ νέου όμως οι υπολογισμοί βασίζονται πλέον στις υπόλοιπες μεταβλητές. Επομένως, στο τέταρτο βήμα οι άγνωστες τιμές (οι οποίες ωστόσο είχαν υπολογιστεί νωρίτερα με τη μέση τιμή της μεταβλητής που εξετάζουμε) αντικαθίστανται από τους υπολογισμούς που προέκυψαν από τις υπόλοιπες μεταβλητές. Στο πέμπτο βήμα, τα βήματα 2 έως 4 επαναλαμβάνονται για κάθε μεταβλητή μέχρι να συμπληρωθούν όλα τα κενά τους. Σε αυτό το σημείο έχει ολοκληρωθεί ένας κύκλος της MICE, άρα στο έκτο βήμα επαναλαμβάνονται τα βήματα 2 έως 5 ώστε να διασφαλιστεί η ακρίβεια στον υπολογισμό των ελλειπουσών τιμών.

Η ανάπτυξη του μοντέλου πραγματοποιήθηκε μέσω του Google Colab, λόγω των εξειδικευμένων πόρων που διαθέτει για την εκπαίδευση τέτοιων μοντέλων. Προκειμένου να επιταχυνθούν οι διαδικασίες εκπαίδευσης οι αρχικές δοκιμές πραγματοποιήθηκαν μόνο για την περίοδο 10 Αυγούστου – 10 Σεπτεμβρίου 2021, η περίοδος δηλαδή στην οποία παρατηρήθηκαν κενά στα δεδομένα.

```
interpolated_data = pd.read_csv('/content/drive/MyDrive/dataset_cleanedFinal.csv', sep=';', header=0)
interpolated_data = interpolated_data[(interpolated_data['timestampDT'] > '2021-08-10 00:00:00') &
                                      (interpolated_data['timestampDT'] < '2021-09-10 00:00:00')]
```

Εικόνα 36: Εισαγωγή συνόλου δεδομένων και φιλτράρισμα των επιθυμητών ημερομηνιών

Εφόσον επιλεγθεί το επιθυμητό υποσύνολο δεδομένων, προχωράμε στην δημιουργία του μοντέλου.

```
X = MICEData(interpolated_data[['temperature', 'hour', 'month', 'timestamp', 'humidity']])
mice = MICE("temperature ~ hour + humidity", sm.OLS, X)
model = mice.fit(100,80)
```

Εικόνα 37: Κλήση των αντικειμένων της MICE

Στην πρώτη γραμμή, δημιουργείται ένα αντικείμενο MICEData. Η κλάση MICEData είναι μια βοηθητική κλάση της βιβλιοθήκης statsmodels.imputation, η οποία προετοιμάζει τα

δεδομένα για τον υπολογισμό. Το `interpolated_data[['temperature','hour','month','timestamp', 'humidity']]` επιλέγει ένα υποσύνολο στηλών από το σύνολο δεδομένων που εισάγαμε. Αυτές οι στήλες περιλαμβάνουν τις στήλες 'temperature', 'hour', 'month', 'timestamp' και 'humidity'. Το αντικείμενο `MICEData` αρχικοποιείται με αυτό το υποσύνολο δεδομένων. Στη δεύτερη γραμμή δημιουργείται ένα αντικείμενο `MICE` με τον τύπο "temperature ~ hour + moisture". Αυτός ο τύπος καθορίζει το μοντέλο υπολογισμού για τη μεταβλητή-στόχο "θερμοκρασία" χρησιμοποιώντας τις μεταβλητές πρόβλεψης "ώρα" και "υγρασία". Εδώ πρακτικά μπορούν να πραγματοποιηθούν διάφοροι συνδυασμοί όμως με βάση τις εικόνες 26 και 27 πως οι μεταβλητές που έχουν την πιο άμεση συσχέτιση με τη μεταβολή της θερμοκρασίας είναι η ώρα της ημέρας (υψηλότερη κατά τις μεσημβρινές ώρες) και η υγρασία (αντιστρόφως ανάλογη με τη θερμοκρασία, όσο υψηλότερη η θερμοκρασία τόσο χαμηλότερη η υγρασία και αντίστροφα). Το όρισμα `sm.OLS` προσδιορίζει την κλάση μοντέλου που θα χρησιμοποιηθεί για τον καταλογισμό, στην προκειμένη περίπτωση, την παλινδρόμηση συνήθων ελαχίστων τετραγώνων (Ordinary Least Squares, OLS) από την βιβλιοθήκη `statsmodels.api`. Το αντικείμενο `MICE` αρχικοποιείται με αυτόν τον τύπο και την κλάση μοντέλου, μαζί με το αντικείμενο `MICEData X` που δημιουργήθηκε προηγουμένως. Τέλος, σε αυτή τη γραμμή, καλείται η μέθοδος `fit` του αντικειμένου `MICE` για την εκτέλεση του υπολογισμού. Τα ορίσματα 100 και 80 καθορίζουν τον αριθμό των επαναλήψεων που πρέπει να εκτελεστούν και τον αριθμό των επαναλήψεων `burn-in`, δηλαδή πόσες επαναλήψεις θα πραγματοποιηθούν πριν ξεκινήσουν οι προβλέψεις του μοντέλου να καταγράφονται. Αυτό γίνεται ώστε να μην επηρεαστούν οι προβλέψεις από τιμές που μπορεί να προκύπτουν τυχαία. Έχουμε ορίσει να γίνουν 100 επαναλήψεις, όμως μόνο οι τελευταίες 20 να είναι αυτές που θα καθορίσουν το μοντέλο. Ο αλγόριθμος `MICE` καταλογίζει επαναληπτικά τις ελλείπουσες τιμές προσαρμόζοντας το καθορισμένο μοντέλο παλινδρόμησης στα παρατηρούμενα δεδομένα και στη συνέχεια καταλογίζοντας τις ελλείπουσες τιμές με βάση τις προβλεπόμενες τιμές από το μοντέλο. Η διαδικασία αυτή επαναλαμβάνεται για τον καθορισμένο αριθμό επαναλήψεων. Το σύνολο δεδομένων που προκύπτει από τον υπολογισμό αποθηκεύεται στη μεταβλητή του μοντέλου.

```
imputed_data_sets = X.data
fml_dataset = interpolated_data.merge(imputed_data_sets, on=['timestamp'], how='outer')
fml_dataset['temperature'] = fml_dataset['temperature_x'].combine_first(fml_dataset['temperature_y'])
fml_dataset.drop(['temperature_x', 'temperature_y'], axis=1, inplace=True)
```

Εικόνα 38: Διαδικασία παρεμβολής των κενών τιμών

Στη συνέχεια γίνεται η συμπλήρωση των κενών τιμών στο σύνολο δεδομένων. Η γραμμή `imputed_data_sets = X.data` περνάει τα δεδομένα που παράχθηκαν από τη `MICE` σε

ένα dataframe για ευκολότερη διαχείριση. Εφόσον τα δεδομένα πλέον είναι σε διαχειρίσιμη μορφή, προχωράμε στη συνένωση αυτού του dataframe με το dataframe που περιέχει το σύνολο δεδομένων που παράχθηκε νωρίτερα, στη στήλη timestamp (ως η μόνη κοινή ανάμεσα στα δύο) και με τη μέθοδο outer. Με αυτή την τεχνική δημιουργούμε τις δύο νέες στήλες temperature_x και temperature_y, όπου η κάθε μια περιέχει αντίστοιχα τα πραγματικά δεδομένα (όσα υπήρχαν) και τις προβλέψεις του μοντέλου MICE. Στην τρίτη γραμμή της εικόνας 35 γίνεται η συγχώνευση των δύο στήλων temperature ώστε οι NaN τιμές να συμπληρωθούν με τις προβλέψεις που προέκυψαν από το MICE.

```
sns.lineplot(data=fnl_dataset, x='timestampDT', y='temperature')  
plt.show()
```

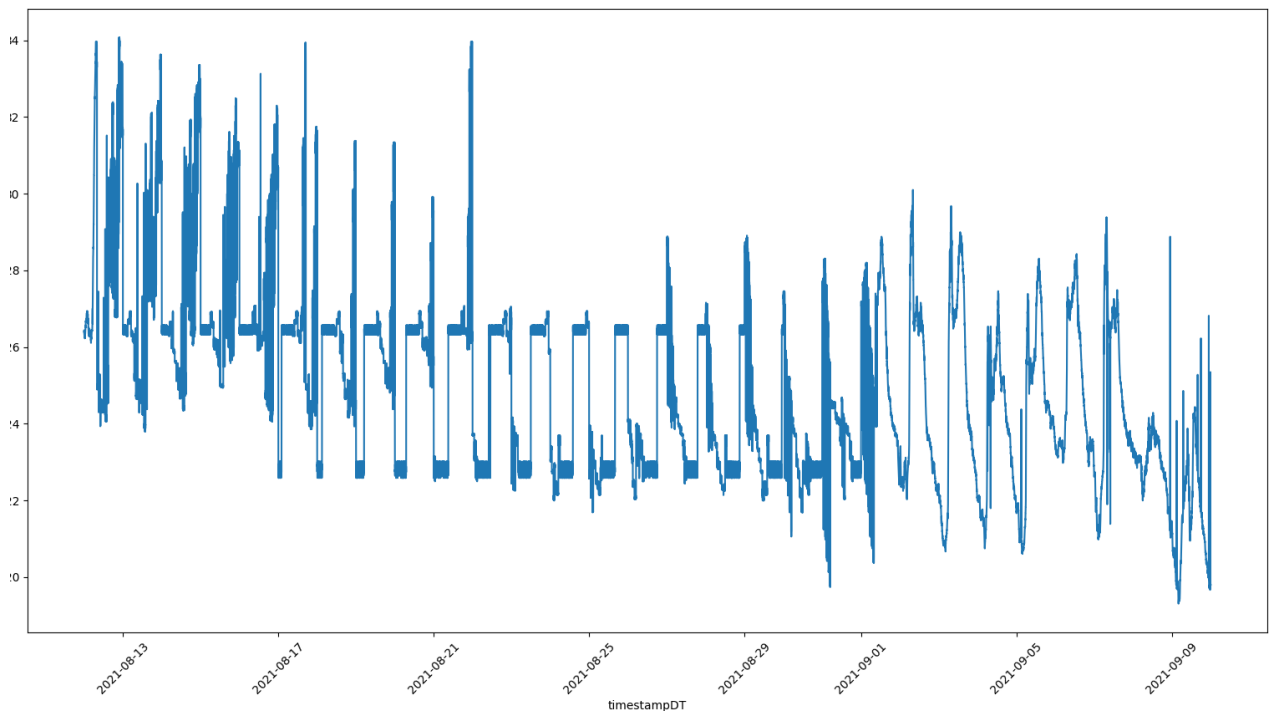
Εικόνα 39: Δημιουργία γραφήματος με βάση τα καινούρια δεδομένα

Τέλος, παράγουμε το νέο γράφημα που απεικονίζει τη θερμοκρασία στο διάστημα 10-08 με 10-09, στο οποίο παρατηρούμε πως το αποτέλεσμα είναι αρκετά θορυβώδες και δεν φαίνεται να ακολουθεί κάποια ιδιαίτερη λογική. Αυτό μπορεί να οφείλεται στον συνδυασμό μεταβλητών που δόθηκε στον τύπο, ή στην ιστορικότητα των δεδομένων. Για αυτό πραγματοποιήθηκε εκ νέου το πείραμα με διαφορετικές παραμέτρους, τις οποίες θα δούμε παρακάτω.

```
X = MICEData(interpolated_data[['temperature', 'hour', 'month', 'timestamp', 'humidity', 'light']])  
mice = MICE("temperature ~ hour + month + humidity + light", sm.OLS, X)  
model = mice.fit(200, 190)
```

Σε αυτό το παράδειγμα, έχουμε προσθέσει επιπλέον την πληροφορία σχετικά με τα επίπεδα φωτός, με τη λογική πως όσο πιο έντονα τα επίπεδα φωτός τόσο μεγαλύτερη θα πρέπει να είναι η θερμοκρασία (πληροφορία που αντικειμενικά ήδη έχουμε με τη στήλη hour, ωστόσο θεωρητικά θα ενισχυθούν οι συσχετίσεις ώρας-θερμοκρασίας με αυτόν τον τρόπο). Σε γενικές γραμμές τα δεδομένα που παράγει η MICE είναι ιδιαίτερα θορυβώδη ακόμα και μετά από διάφορες τροποποιήσεις στο σύνολο επαναλήψεων και burn-in επαναλήψεων. Παρότι το σύνολο δεδομένων που προκύπτει ακολουθεί ένα μοτίβο, είναι αρκετά τυχαίο ακόμα και μετά από διάφορες τροποποιήσεις στο σύνολο επαναλήψεων και burn-in επαναλήψεων. Επομένως, λόγω των έντονων μεταβολών από λεπτό σε λεπτό, η MICE δε θα χρησιμοποιηθεί στον τελικό σύνολο δεδομένων, και τα κενά στα δεδομένα απλώς θα διαγραφούν καθώς ό,τι κέρδος έχουμε από τη συμπλήρωση των κενών αυτών εξαυλώνεται όταν έχουμε δεδομένα που θα μπορούσαν να είναι και ψευδοτυχαία. Θα είχε ενδιαφέρον η προσπάθεια εφαρμογής άλλων

τεχνικών παρεμβολής, ωστόσο γίνεται σαφές πως καταβάλλεται μια τεράστια προσπάθεια πάνω σε ένα θέμα που αποσκοπεί στην παροχή δεδομένων για δύο εβδομάδες παραπάνω για ένα πολύμηνο σύνολο δεδομένων.



Εικόνα 40: Γράφημα θερμοκρασίας-ημερομηνίας με δεδομένα που προέκυψαν από την τεχνική MICE

6.3 Decision Tree Classifier

Ο αλγόριθμος δέντρου αποφάσεων είναι μια ευρέως χρησιμοποιούμενη και διαισθητική τεχνική μηχανικής μάθησης που δημιουργεί ένα δενδροειδές μοντέλο για να κάνει προβλέψεις ή αποφάσεις με βάση τα χαρακτηριστικά εισόδου. Εφαρμόζεται τόσο σε εργασίες ταξινόμησης όσο και σε εργασίες παλινδρόμησης και προσφέρει αρκετά πλεονεκτήματα, συμπεριλαμβανομένης της ερμηνευσιμότητας, της ικανότητας χειρισμού διαφόρων τύπων χαρακτηριστικών και της ανθεκτικότητας σε ελλείποντα δεδομένα και ακραίες τιμές. Σε αυτό το κεφάλαιο, θα εξερευνήσουμε τη γενική λειτουργία του αλγορίθμου δέντρου απόφασης, τη διαδικασία κατασκευής του, τις προαιρετικές τεχνικές κλαδέματος και τον μηχανισμό πρόβλεψης. Επιπλέον, θα συζητήσουμε τα πλεονεκτήματα και τις αδυναμίες των δέντρων απόφασης και θα επισημάνουμε πώς οι μέθοδοι συνόλου μπορούν να βελτιώσουν την απόδοσή τους. Η θεμελιώδης ιδέα πίσω από τον αλγόριθμο δέντρων απόφασης είναι η κατάτμηση των δεδομένων σε μικρότερα υποσύνολα με βάση τις τιμές των χαρακτηριστικών εισόδου. Με την επαναλαμβανόμενη διαίρεση του συνόλου δεδομένων, λαμβάνονται αποφάσεις σε κάθε κόμβο του δέντρου

με βάση τις προκύπτουσες καταταμίσεις. Ο αλγόριθμος ξεκινά με την επιλογή του πιο κατατοπιστικού χαρακτηριστικού από το διαθέσιμο σύνολο χαρακτηριστικών. Τα κριτήρια επιλογής μπορούν να ποικίλλουν, με τα συνήθως χρησιμοποιούμενα μέτρα να περιλαμβάνουν το κέρδος πληροφορίας ή την εντροπία. Μόλις επιλεγεί το χαρακτηριστικό, το σύνολο δεδομένων διαιρείται σε υποσύνολα με βάση τις τιμές του χαρακτηριστικού, δημιουργώντας κόμβους-παιδιά ή κλάδους στο δέντρο. Αυτή η διαδικασία διαχωρισμού συνεχίζεται αναδρομικά για κάθε κόμβο-παιδί μέχρι να ικανοποιηθεί ένα κριτήριο διακοπής. Τα συνήθη κριτήρια διακοπής περιλαμβάνουν την επίτευξη ενός μέγιστου βάθους δέντρου, την ύπαρξη ενός ελάχιστου αριθμού δειγμάτων σε έναν κόμβο ή την επίτευξη καθαρών κλάσεων σε έναν κόμβο-φύλλο (για εργασίες ταξινόμησης). Μετά την κατασκευή του δέντρου απόφασης, μπορεί να εφαρμοστεί ένα προαιρετικό βήμα που ονομάζεται κλάδεμα (pruning) για την καταπολέμηση της υπερπροσαρμογής (overfitting). Η υπερπροσαρμογή συμβαίνει όταν το δέντρο γίνεται πολύ πολύπλοκο και αποδίδει καλά στα δεδομένα εκπαίδευσης αλλά ελάχιστα στα αόρατα δεδομένα. Οι τεχνικές κλαδέματος αποσκοπούν στην απλοποίηση του δέντρου με την αφαίρεση ή τη συγχώνευση κόμβων που συμβάλλουν λιγότερο στη συνολική απόδοση του δέντρου. Αυτό συμβάλλει στη βελτίωση της ικανότητας γενίκευσης του δέντρου, επιτρέποντάς του να κάνει ακριβείς προβλέψεις σε νέα, αόρατα δεδομένα. Για να γίνουν προβλέψεις με τη χρήση του μοντέλου δέντρου αποφάσεων, νέες περιπτώσεις δεδομένων περνούν μέσα από το δέντρο. Ξεκινώντας από τον κόμβο της ρίζας, τα χαρακτηριστικά της περίπτωσης συγκρίνονται με τα κριτήρια διαχωρισμού σε κάθε κόμβο. Με βάση τις τιμές των χαρακτηριστικών, η περίπτωση διασχίζεται στο δέντρο κατά μήκος του κατάλληλου κλάδου μέχρι να φτάσει σε έναν κόμβο φύλλου. Στη συνέχεια, η πρόβλεψη ή η απόφαση που σχετίζεται με τον κόμβο φύλλου αποδίδεται στην περίπτωση. Αυτή η διαδικασία διάσχισης αποτυπώνει τα μοτίβα και τις σχέσεις στα δεδομένα, επιτρέποντας στο δέντρο αποφάσεων να κάνει τεκμηριωμένες προβλέψεις. Τα δέντρα αποφάσεων προσφέρουν πολλά πλεονεκτήματα που συμβάλλουν στη δημοτικότητά τους. Ένα από τα κύρια πλεονεκτήματα είναι η ερμηνευσιμότητά τους. Η προκύπτουσα δενδρική δομή μπορεί εύκολα να οπτικοποιηθεί και να γίνει κατανοητή, επιτρέποντας στους αναλυτές και τους ενδιαφερόμενους να αποκτήσουν εικόνα της διαδικασίας λήψης αποφάσεων. Επιπλέον, τα δέντρα αποφάσεων μπορούν να χειριστούν τόσο κατηγορικά όσο και αριθμητικά χαρακτηριστικά, καθιστώντας τα ευέλικτα για ένα ευρύ φάσμα συνόλων δεδομένων. Μπορούν να αποτυπώσουν πολύπλοκες αλληλεπιδράσεις μεταξύ μεταβλητών, γεγονός που είναι ιδιαίτερα χρήσιμο όταν πρόκειται για μη γραμμικές σχέσεις. Επιπλέον, τα

δέντρα αποφάσεων είναι ανθεκτικά σε ελλείποντα δεδομένα και ακραίες τιμές, καθώς δεν βασίζονται σε τεχνικές κανονικοποίησης δεδομένων ή υπολογισμού. Παρά τα πλεονεκτήματά τους, τα δέντρα αποφάσεων έχουν ορισμένους περιορισμούς. Είναι επιρρεπή σε υπερπροσαρμογή, ιδίως όταν τα δέντρα αναπτύσσονται πολύ βαθιά ή όταν το σύνολο δεδομένων περιέχει θόρυβο ή άσχετα χαρακτηριστικά. Η υπερπροσαρμογή μπορεί να οδηγήσει σε κακή γενίκευση και μη βέλτιστη απόδοση σε αόρατα δεδομένα. Για τον μετριασμό αυτού του ζητήματος, χρησιμοποιούνται συχνά μέθοδοι συνόλου, όπως τα τυχαία δάση ή η ενίσχυση κλίσης. Αυτές οι μέθοδοι συνδυάζουν πολλαπλά δέντρα απόφασης για να βελτιώσουν την ακρίβεια πρόβλεψης και να μειώσουν την υπερπροσαρμογή.

Για την ανάπτυξη του μοντέλου αυτού, στραφήκαμε στο Visual Studio Code. Ξεκινήσαμε με την εισαγωγή του συνόλου δεδομένων που δημιουργήσαμε στην προηγούμενη ενότητα, και τη διαγραφή της στήλης timestampDT, η οποία ήταν στην κλασσική μορφή ημερομηνίας YYYY-MM-DDTHH:MM:SS. Έτσι διατηρήθηκε μόνο η στήλη timestamp που είχε τις ημερομηνίες σε μορφή Unix epoch.

```
source = pd.read_csv('Products/dataset_cleanedFinal.csv', sep=';', header=0)
source = source.drop(columns=['timestampDT'])
```

Εικόνα 41: Εισαγωγή δεδομένων και διαγραφή της στήλης timestampDT

Στη συνέχεια γίνεται ο διαχωρισμός του συνόλου σε δύο νέα dataframes, το X που περιέχει όλο το σύνολο δεδομένων πλην της τελευταίας στήλης που περιέχει τον classifier (δηλαδή τη στήλη weather_description, αυτή που μας ενδιαφέρει) και το y που περιέχει μόνο την στήλη του classifier. Ετοιμάζεται επίσης η μεταβλητή labels με τις μοναδικές κατηγορίες καιρού για χρήση στον πίνακα σύγχυσης παρακάτω. Η τελευταία γραμμή χρησιμοποιεί τη συνάρτηση train_test_split από τη βιβλιοθήκη scikit-learn για να χωρίσει τα δεδομένα σε σύνολα εκπαίδευσης και δοκιμής. Οι μεταβλητές X και y χωρίζονται σε X_train (χαρακτηριστικά εκπαίδευσης), X_test (χαρακτηριστικά δοκιμής), y_train (ετικέτες εκπαίδευσης) και y_test (ετικέτες δοκιμής). Το όρισμα test_size=0,2 καθορίζει ότι το 20% των δεδομένων θα πρέπει να διατεθεί για δοκιμές, ενώ το υπόλοιπο 80% θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου.

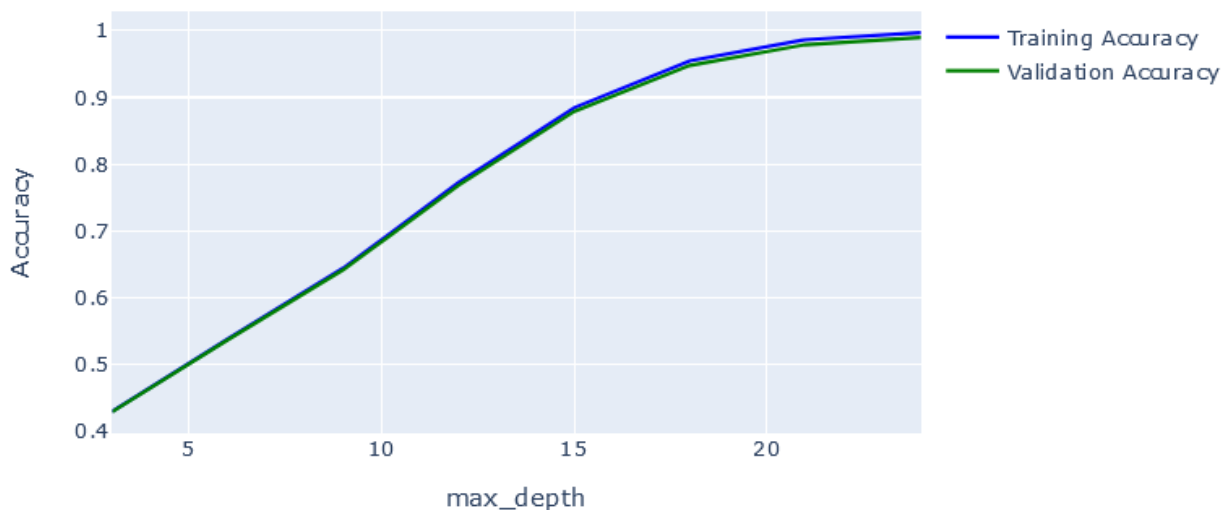
```
X = source.iloc[:, :-1]
y = source.iloc[:, -1]
labels = np.unique(y)
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2)
```

Εικόνα 42: Διαχωρισμός test-train

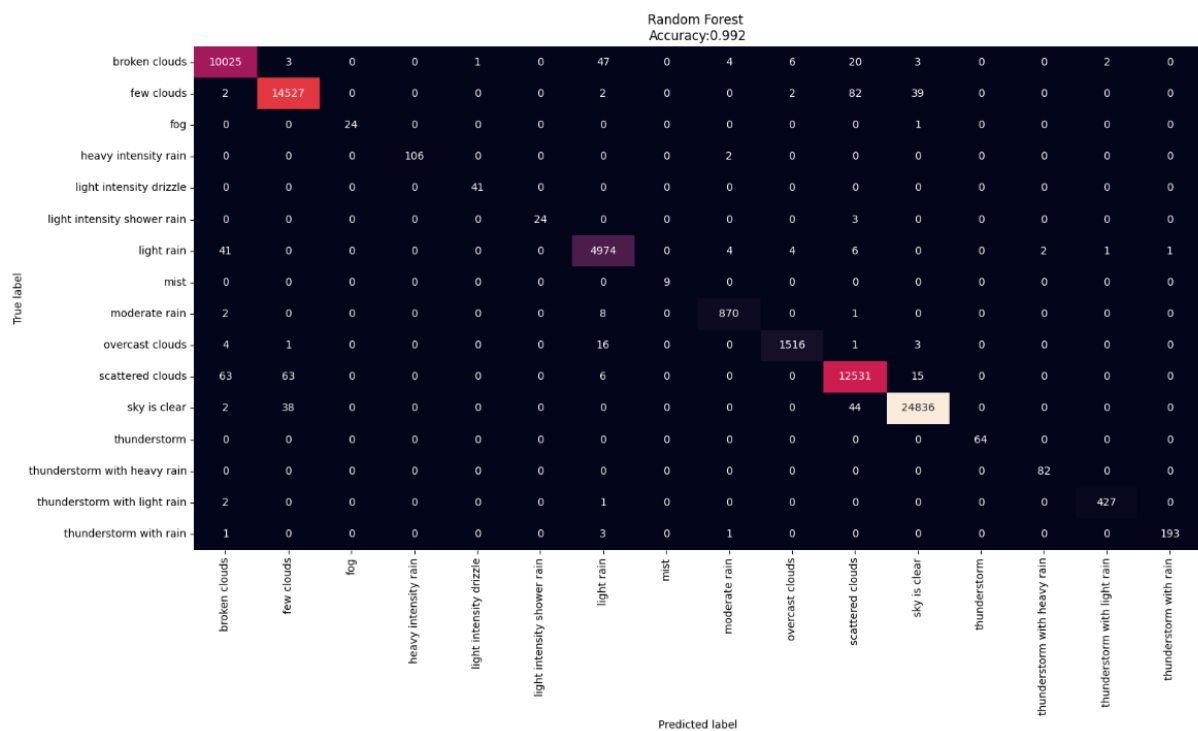
```
dtree_model = DecisionTreeClassifier(max_depth = 25).fit(X_train, y_train)
predictions = dtree_model.predict(X_test)
```

Εικόνα 43: Εκκίνηση εκπαίδευσης του μοντέλου

Η πρώτη γραμμή καλεί την κλάση `DecisionTreeClassifier` από την βιβλιοθήκη `tree` της `scikit-learn`. Η παράμετρος `max_depth` ορίζεται σε 25, η οποία καθορίζει το μέγιστο βάθος ή επίπεδα του δέντρου απόφασης. Στη συνέχεια καλείται η συνάρτηση `fit()` με τα δεδομένα εκπαίδευσης (`X_train` και `y_train`) για την εκπαίδευση του μοντέλου δέντρου απόφασης. Η επιλογή των 25 επιπέδων ορίστηκε με επαναλαμβανόμενες δοκιμές για να βρεθεί το βέλτιστο βάθος για την ιδανικότερη καμπύλη μάθησης. Στην περίπτωση του βάθους 25, η καμπύλη προς το τέλος της εκπαίδευσης χάνει την ανοδική της τάση, κάτι που δείχνει πως πλησιάζει το σημείο στο οποίο δεν υπάρχει κάτι άλλο προς μάθηση για το μοντέλο. Παράλληλα, ο πίνακας σύγκρισης δείχνει μια ακρίβεια στις προβλέψεις της τάξης του 99%.

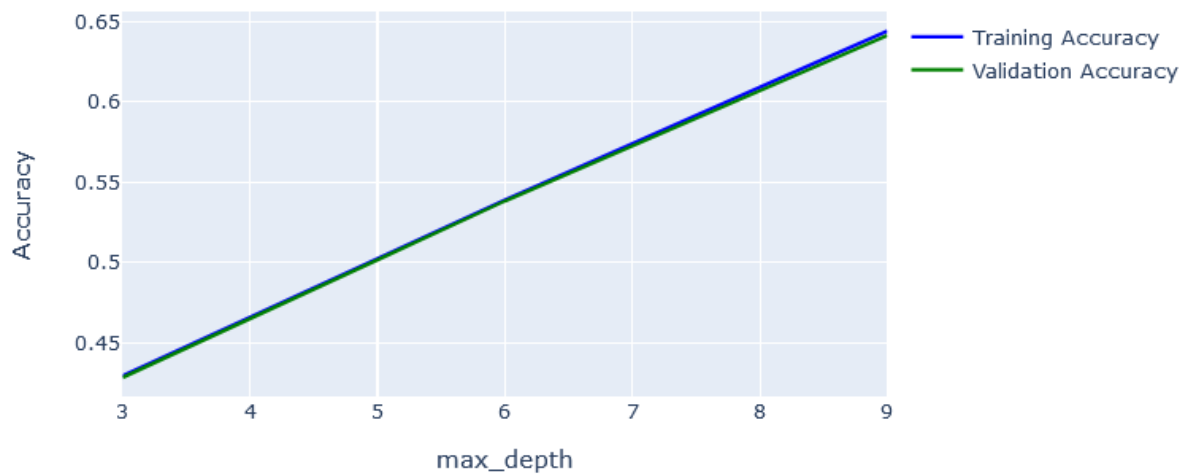


Εικόνα 44: Καμπύλη μάθησης για μοντέλο βάθους 25 επιπέδων

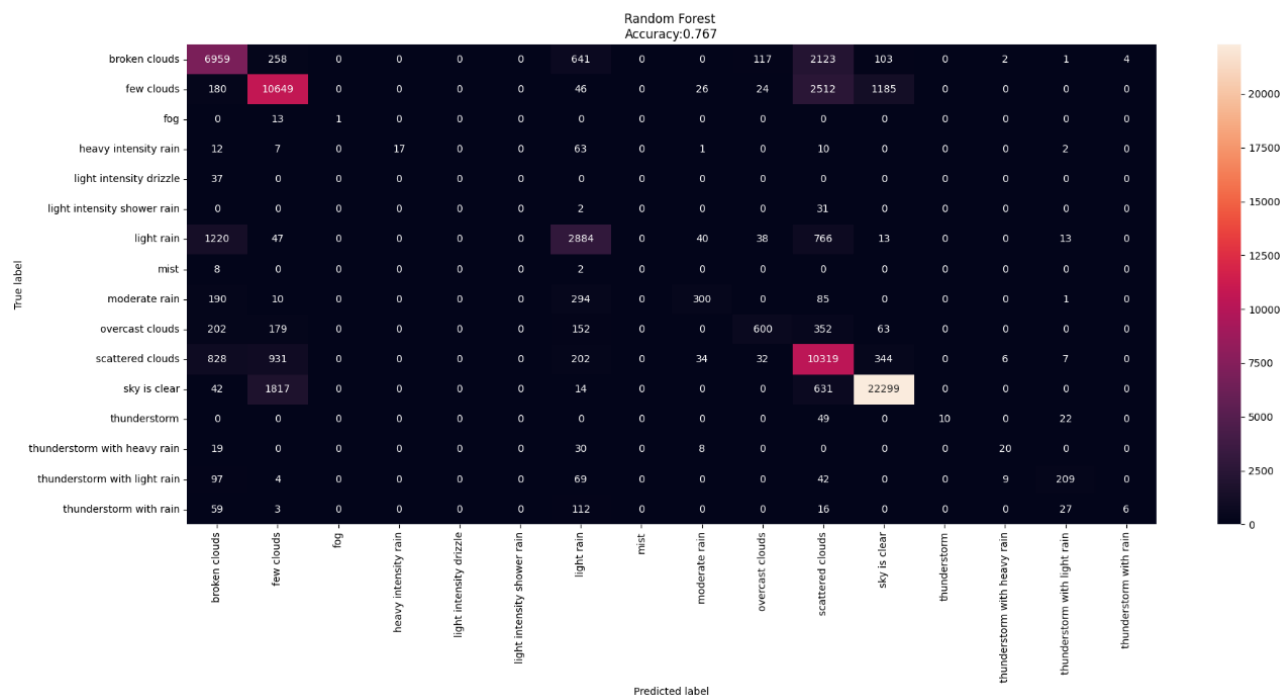


Εικόνα 45: Πίνακας σύγχυσης μοντέλου βάθους 25 επιπέδων

Αν μειωθεί το βάθος του μοντέλου στα μισά, δηλαδή στα 12 επίπεδα, η καμπύλη μάθησης παρουσιάζει σταθερή ανοδική πορεία, υποδεικνύοντας πως τα 12 επίπεδα δεν επαρκούν για να έχουμε ένα επαρκώς εκπαιδευμένο μοντέλο και ο πίνακας σύγχυσης αντικατοπτρίζει την εικόνα αυτή με μια μειωμένη ακρίβεια στα 76%.



Εικόνα 46: Καμπύλη μάθησης για μοντέλο βάθους 12 επιπέδων



Εικόνα 47: Πίνακας σύγχυσης για μοντέλο βάρους 12 επιπέδων

Η παραγωγή της καμπύλης μάθησης γίνεται μέσω της συνάρτησης `plot_validation_curve` της βιβλιοθήκης `model_selection` του πακέτου `sklearn` με τις παρακάτω εντολές.

```
param_range = np.arange(3, 25, 3)
plot_validation_curves(model, X_train, y_train, "max_depth", param_range, 5)
```

Εικόνα 48: Παραγωγή καμπύλης μάθησης μοντέλου

Η συνάρτηση επιστρέφει δύο πίνακες: τον `train_scores` που περιέχουν τα αποτελέσματα εκπαίδευσης για κάθε τιμή στο σύνολο δεδομένων και τον `valid_scores` που περιέχουν τα αποτελέσματα επικύρωσης για κάθε τιμή στο σύνολο δεδομένων. Ο επόμενος κώδικας υπολογίζει τη μέση τιμή και την τυπική απόκλιση των βαθμολογιών `train/validation` στις αναδιπλώσεις και στη συνέχεια σχεδιάζει την καμπύλη επικύρωσης χρησιμοποιώντας τη `matplotlib`. Στη συνέχεια δημιουργούμε τον πίνακα σύγχυσης για το μοντέλο.

```
cm = confusion_matrix(y_test, predictions)
cm_df = pd.DataFrame(cm, index=labels, columns=labels)
plt.figure(figsize=(5.5, 4))
sns.heatmap(cm_df, annot=True, fmt='g')
plt.title('Random Forest \nAccuracy:{0:.3f}'.format(
    accuracy_score(y_test, predictions)))
plt.ylabel('True label')
plt.xlabel('Predicted label')
plt.show(block=True)
```

Εικόνα 49: Κώδικας παραγωγής του πίνακα σύγχυσης

Αξιοσημείωτη είναι η χρήση της συνάρτησης heatmap που κωδικοποιεί χρωματικά τον πίνακα και τον κάνει πολύ εύκολο στην κατανόηση του παρά το μεγάλο του μέγεθος. Η κύρια διαγώνιος χαρακτηρίζεται από πιο έντονα χρώματα σε σχέση με άλλα πεδία, τονίζοντας τις ορθά θετικές κατηγοριοποιήσεις του μοντέλου.

7 Συμπεράσματα και Μελλοντικές επεκτάσεις

Η παρούσα εργασία επικεντρώθηκε στην ανάλυση και μελέτη μεθόδων ανάπτυξης μοντέλων μηχανικής μάθησης για την πρόγνωση και κατηγοριοποίηση των καιρικών φαινομένων για μια πολύ συγκεκριμένη περιοχή ή/και χρήση (όπως στην γεωργία χάρη στους κατάλληλους αισθητήρες του Smart Citizen). Τα μειωμένα κόστη και η ευρεία πρόσβαση στα εργαλεία που χρησιμοποιούνται στα πλαίσια της τεχνητής νοημοσύνης και διαχείρισης δεδομένων IoT συσκευών σημαίνουν πως τέτοια μοντέλα μπορούν να αναπτυχθούν και να χρησιμοποιηθούν παράλληλα με συνήθεις μεθόδους πρόγνωσης καιρικών φαινομένων ώστε να δίνεται η δυνατότητα λήψης αποφάσεων σε τομείς που άλλοτε δεν ήταν εύκολο. Αποδεικνύεται πως ακόμα και με πρόσβαση σε βασικά μετεωρολογικά δεδομένα υπάρχει η δυνατότητα ανάλυσης καιρικών φαινομένων σε βαθμό ικανοποιητικό για χρήση από πολίτες. Μελετήθηκαν παρόμοιες έρευνες και εργασίες που προσεγγίζουν το θέμα με διαφορετικό τρόπο (όπως με τη συλλογή δεδομένων πληθοπορισμού από μέσα κοινωνικής δικτύωσης που είδαμε νωρίτερα), και ο συνδυασμός αυτών των μεθόδων μπορεί να εφαρμοστεί ώστε να επεκταθούν οι δυνατότητες των μοντέλων και διαδικασιών που μελετήθηκαν ώστε να παρέχεται μια ολοκληρωμένη πλατφόρμα για τη συλλογή δεδομένων, την μετάδοση προβλέψεων σε χρήστες και την αξιολόγηση των προβλέψεων από τους χρήστες ή και από ιστορικά δεδομένα.

Η παρούσα εργασία μπορεί να βελτιωθεί μελλοντικά δημιουργώντας μια ενιαία πλατφόρμα διαχείρισης των δεδομένων και βελτίωση των προβλέψεων. Η χρήση για παράδειγμα του Apache Kafka για τη διαχείριση μεγάλου στόλου αισθητήρων Smart Citizen, χάρη στο API του Smart Citizen που είδαμε στην αντίστοιχη ενότητα. Σαφώς είναι απαραίτητη και η ανάπτυξη εφαρμογής μέσω της οποίας θα είναι προσβάσιμη όλη η πληροφορία που συλλέγεται, παράγεται και απαιτείται για το σύστημα.

Συνοψίζοντας:

- Ανάπτυξη βάσης δεδομένων Apache Kafka για τη συλλογή και διατήρηση των δεδομένων από τους αισθητήρες
- Αυτοματοποίηση ροών συλλογής δεδομένων μέσω αυτόματων διεργασιών (cron jobs)
- Ανάπτυξη εφαρμογής για πρόσβαση στις προβλέψεις του συστήματος και συλλογή δεδομένων πληθοπορισμού από χρήστες. Η εφαρμογή θα πρέπει να δίνει τη δυνατότητα επιλογής τοποθεσίας που ενδιαφέρει τον χρήστη και να δίνει την τωρινή καιρική κατάσταση, καθώς και ορισμένες μελλοντικές

προβλέψεις, καθώς και μια απλή διεπαφή για παροχή ανατροφοδότησης από τους χρήστες (είτε σε δυαδική μορφή είτε σε αναλυτική)

- Περαιτέρω μελέτη αλγορίθμων κατάλληλων για χρονοσειρές όπως η MICE για την παροχή μελλοντικών καιρικών τιμών (θερμοκρασία, υγρασία και λοιπά)

Βιβλιογραφία

- [1] K. Rose, S. Eldridge, and L. Chapin, “The Internet of Things: An Overview”.
- [2] Y. Zhang *et al.*, “Adherence with blood pressure monitoring wearable device among the elderly with hypertension: The case of rural China,” *Brain Behav*, vol. 10, no. 6, Jun. 2020, doi: 10.1002/brb3.1599.
- [3] “Has the Future Started? The Current Growth of Artificial Intelligence, Machine Learning, and Deep Learning,” *ijcsm*, pp. 115–123, Jan. 2022, doi: 10.52866/ijcsm.2022.01.01.013.
- [4] J. M. Scanlon, K. D. Kusano, T. Daniel, C. Alderson, A. Ogle, and T. Victor, “Waymo simulated driving behavior in reconstructed fatal crashes within an autonomous vehicle operating domain,” *Accident Analysis & Prevention*, vol. 163, p. 106454, Dec. 2021, doi: 10.1016/j.aap.2021.106454.
- [5] O. Bousquet, U. von Luxburg, and G. Rätsch, Eds., *Advanced lectures on machine learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003 [and] Tübingen, Germany, August 4-16, 2003: revised lectures*. in Lecture notes in computer science, Lecture notes in artificial intelligence, no. 3176. Berlin ; New York: Springer, 2004.
- [6] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *J Big Data*, vol. 6, no. 1, p. 60, Dec. 2019, doi: 10.1186/s40537-019-0197-0.
- [7] D. Bandyopadhyay and J. Sen, “Internet of Things: Applications and Challenges in Technology and Standardization,” *Wireless Pers Commun*, vol. 58, no. 1, pp. 49–69, May 2011, doi: 10.1007/s11277-011-0288-5.
- [8] P. A. Belyakova, A. N. Shikhov, S. I. Perminov, and V. M. Moreydo, “Accuracy assessment of heavy rainfall forecasting in the Western Caucasus with ICON-Eu regional atmospheric model for short-term flood forecasting,” *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 611, no. 1, p. 012049, Dec. 2020, doi: 10.1088/1755-1315/611/1/012049.
- [9] Γ. Τριανταφύλλου, “ΑΝΑΠΤΥΞΗ ΑΥΤΟΝΟΜΟΥ ΣΤΑΘΜΟΥ ΚΑΤΑΓΡΑΦΗΣ & ΠΡΟΒΛΕΨΗΣ ΜΕΤΕΩΡΟΛΟΓΙΚΩΝ ΦΑΙΝΟΜΕΝΩΝ”, [Online]. Available: <https://apothesis.eap.gr/archive/download/ecb8642c-3cbf-4919-b6ad-92d45d794deb.pdf>
- [10] L. P. Rothfusz, “The Heat Index ‘Equation’ (or, More Than You Ever Wanted to Know About Heat Index)”.
- [11] G. Camprodon *et al.*, “Smart Citizen Kit and Station: An open environmental monitoring system for citizen participation and scientific experimentation,” *HardwareX*, vol. 6, p. e00070, Oct. 2019, doi: 10.1016/j.ohx.2019.e00070.
- [12] “Crowdsourced production of AI Training Data: How human workers teach self-driving cars how to see”.
- [13] M. Moradi and M. R. Keyvanpour, “CAPTCHA for crowdsourced image annotation: directions and efficiency analysis,” *AJIM*, vol. 74, no. 3, pp. 522–548, May 2022, doi: 10.1108/AJIM-08-2021-0215.
- [14] M. Hossain and I. Kauranen, “Crowdsourcing: a comprehensive literature review,” *Strategic Outsourcing: An International Journal*, vol. 8, no. 1, pp. 2–22, Feb. 2015, doi: 10.1108/SO-12-2014-0029.
- [15] M. Toğaçar, B. Ergen, and Z. Cömert, “Detection of weather images by using spiking neural networks of deep learning models,” *Neural Comput & Applic*, vol. 33, no. 11, pp. 6147–6159, Jun. 2021, doi: 10.1007/s00521-020-05388-3.
- [16] K. Purwandari, J. W. C. Sigalingging, T. W. Cenggoro, and B. Pardamean, “Multi-class Weather Forecasting from Twitter Using Machine Learning Approaches,” *Procedia Computer Science*, vol. 179, pp. 47–54, 2021, doi: 10.1016/j.procs.2020.12.006.
- [17] “Visual Studio Code,” *Wikipedia*. Jun. 18, 2023. Accessed: Jun. 30, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Visual_Studio_Code&oldid=1160719180
- [18] “colab.google.” <https://colab.google/> (accessed Jun. 30, 2023).

- [19] “Normalization (statistics),” *Wikipedia*. Sep. 11, 2022. Accessed: Jun. 30, 2023. [Online]. Available:
[https://en.wikipedia.org/w/index.php?title=Normalization_\(statistics\)&oldid=1109670121](https://en.wikipedia.org/w/index.php?title=Normalization_(statistics)&oldid=1109670121)
- [20] “Python,” *Βικιπαίδεια*. Mar. 10, 2023. Accessed: Jul. 01, 2023. [Online]. Available:
<https://el.wikipedia.org/w/index.php?title=Python&oldid=9946167>
- [21] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, “Multiple imputation by chained equations: what is it and how does it work?: Multiple imputation by chained equations,” *Int. J. Methods Psychiatr. Res.*, vol. 20, no. 1, pp. 40–49, Mar. 2011, doi: 10.1002/mpr.329.
- [22] “deiktis_thermotitas.pdf.” Accessed: Jun. 13, 2023. [Online]. Available:
https://www.meteo.gr/pdf/deiktis_thermotitas.pdf
- [23] “WekaManual-3-8-3.pdf.” Accessed: Jun. 30, 2023. [Online]. Available: <https://deac-ams.dl.sourceforge.net/project/weka/documentation/3.8.x/WekaManual-3-8-3.pdf>