



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΔΠΜΣ
ΣΥΣΤΗΜΑΤΑ ΑΥΤΟΜΑΤΙΣΜΟΥ

Πρόβλεψη συμβάντων βλαβών σε εξοπλισμό
παραγωγής με τεχνικές μηχανικής μάθησης

ΕΥΣΤΑΘΙΟΣ ΡΑΠΤΗΣ

Επιβλέπων Καθηγητής
Πανώριος Μπενάρδος

Αθήνα, 2023

Abstract

In the modern era of automation and industrial processes, the need for effective fault detection and diagnosis techniques has become increasingly critical. Timely and accurate identification of faults is essential for ensuring the smooth operation, safety, and reliability of complex systems, such as manufacturing plants, power grids, transportation networks, and more. Traditional methods for fault detection often rely on rule-based systems or statistical approaches, which may struggle to handle the complexity and variability of real-world systems.

This work addresses the topic of detecting plant failure events in advance by utilizing time series data of sensor measurements and control reference signals for multiple control components of a plant, as well as additional measurements from plant zones. The dataset consists of time series data collected over a period of three to four years, with approximately one sample recorded every 15 minutes. Plant failure events are characterized by a start time, end time, and failure code, with only faults of types 1-5 considered relevant, while code 6 represents other faults not of interest. The objective is to develop a predictive model that can accurately forecast future failure events of types 1-5 and predict their timing based on historical data.

The proposed approach aims to leverage machine learning techniques to learn patterns and anomalies in the data, enabling early detection and proactive maintenance strategies to mitigate the risk of costly breakdowns. This thesis discusses the specific challenges associated with this problem, such as data availability, data preprocessing, feature engineering, model building, imbalanced classification, and model interpretability, and proposes potential solutions to address these challenges. The results obtained from the developed predictive model are presented, demonstrating its effectiveness in predicting future failure events and contributing to the improvement of plant reliability and operational efficiency. Namely, favourable values are attained in critical parameters such as the precision of non-failure events and the recall of failure events.

Keywords— Fault Prediction, Machine Learning, Multilabel classification, Imbalanced Learning

Περίληψη

Στη σύγχρονη εποχή του αυτοματισμού και των βιομηχανικών διαδικασιών, η ανάγκη για αποτελεσματικές τεχνικές ανίχνευσης και διάγνωσης σφαλμάτων έχει γίνει όλο και πιο κρίσιμη. Ο έγκαιρος και ακριβής εντοπισμός σφαλμάτων είναι απαραίτητος για τη διασφάλιση της ομαλής λειτουργίας, της ασφάλειας και της αξιοπιστίας πολύπλοκων συστημάτων, όπως εργοστάσια παραγωγής, δίκτυα ηλεκτρικής ενέργειας, δίκτυα μεταφορών και άλλα. Οι παραδοσιακές μέθοδοι πρόβλεψης σφαλμάτων βασίζονται συχνά σε συστήματα που βασίζονται σε κανόνες ή σε στατιστικές προσεγγίσεις, οι οποίες δυσκολεύονται να διαχειριστούν την πολυπλοκότητα και τη μεταβλητότητα των συστημάτων του πραγματικού κόσμου.

Η παρούσα εργασία ασχολείται με το θέμα της πρόβλεψης συμβάντων βλάβης ενός εργοστασίου, χρησιμοποιώντας δεδομένα χρονοσειρών μετρήσεων αισθητήρων και σημάτων αναφοράς ελέγχου για πολλαπλά στοιχεία ελέγχου του εργοστασίου, καθώς και πρόσθετες μετρήσεις από τις ζώνες του εργοστασίου. Το σύνολο δεδομένων αποτελείται από δεδομένα χρονοσειράς που συλλέγονται για μια περίοδο τριών έως τεσσάρων ετών, με περίπου ένα δείγμα να καταγράφεται κάθε 15 λεπτά. Τα συμβάντα βλάβης της εγκατάστασης χαρακτηρίζονται από χρόνο έναρξης, χρόνο λήξης και κωδικό βλάβης, ενώ μόνο οι βλάβες των τύπων 1-5 θεωρούνται σχετικές, καθώς ο κωδικός 6 αντιπροσωπεύει άλλες βλάβες που δεν ενδιαφέρουν. Ο στόχος είναι να αναπτυχθεί ένα μοντέλο που να μπορεί να προβλέπει με ακρίβεια τα μελλοντικά συμβάντα βλάβης των τύπων 1-5 και να προβλέπει το χρόνο τους με βάση τα ιστορικά δεδομένα. Η προτεινόμενη προσέγγιση αποσκοπεί στην αξιοποίηση τεχνικών μηχανικής μάθησης για την εκμάθηση προτύπων και ανωμαλιών στα δεδομένα, επιτρέποντας την έγκαιρη πρόγνωση και την εφαρμογή στρατηγικών προληπτικής συντήρησης για τον μετριασμό του κινδύνου δαπανηρών βλαβών.

Η εργασία εξετάζει τις συγκεκριμένες προκλήσεις που σχετίζονται με το πρόβλημα αυτό, όπως η διαθεσιμότητα δεδομένων, η προετοιμασία των δεδομένων, η μηχανική των ανεξάρτητων μεταβλητών, η ποιότητα των δεδομένων, η μοντελοποίηση του προβλήματος, οι μη σταθμισμένες κλάσεις και η ερμηνευσιμότητα του μοντέλου, και προτείνει πιθανές λύσεις για την αντιμετώπιση αυτών των προκλήσεων. Παρουσιάζονται τα αποτελέσματα που προέκυψαν από το μοντέλο πρόβλεψης που αναπτύχθηκε, καταδεικνύοντας την αποτελεσματικότητά του, αλλά και τις αδυναμίες του στην πρόβλεψη μελλοντικών συμβάντων βλάβης, συμβάλλοντας στη βελτίωση της αξιοπιστίας και της λειτουργικής αποδοτικότητας της του εργοστασίου. Ειδικότερα, ενθαρρυντικές τιμές επετεύχθησαν σε κρίσιμες παραμέτρους όπως το precision των συμβάντων μη-βλάβης καθώς και το recall των συμβάντων βλάβης.

Λέξεις Κλειδιά— Πρόβλεψη Βλαβών, Μηχανική Μάθηση, Ταξινόμηση Πολλαπλών Κλάσεων, Μη Σταθμισμένες Κλάσεις

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Πανώριο Μπενάρδο. Παρά τον μεγάλο φόρτο εργασίας του, ήταν πάντα πρόθυμος να με βοηθήσει σε ότι χρειαζτώ και να με καθοδηγήσει, παρέχοντάς μου τις απαραίτητες επιστημονικές του γνώσεις.

Πάνω από όλα, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στην οικογένειά μου για την ακλόνητη στήριξή τους καθ' όλη τη διάρκεια τόσο του ακαδημαϊκού όσο και προσωπικού μου βίου. Η αγάπη, η ενθάρρυνση και η κατανόησή τους ήταν ανεκτίμητες για μένα. Είμαι ευγνώμον για την υπομονή και την κατανόησή τους κατά τις αμέτρητες ώρες που αφιέρωσα για την έρευνα και τη συγγραφή αυτής της εργασίας. Η σταθερή πίστη τους στις ικανότητές μου ήταν πηγή κίνητρου και έμπνευσης. Θα ήθελα επίσης να τους ευχαριστήσω για τη δημιουργία ενός ευνοϊκού και υποστηρικτικού περιβάλλοντος που μου επέτρεψε να επικεντρωθώ στις ακαδημαϊκές μου προσπάθειες.

Περιεχόμενα

1	Εισαγωγή	15
2	Προετοιμασία δεδομένων	19
2.1	Αρχική επεξεργασία	19
2.2	Καθορισμός μεταβλητής-στόχου	23
3	Διερευνητική Ανάλυση Συνόλου Δεδομένων	26
3.1	Διερεύνηση ανεξάρτητων μεταβλητών	27
3.2	Συσχετίσεις ανεξάρτητων μεταβλητών	41
3.3	Διερεύνηση μεταβλητής-στόχου	44
4	Προεπεξεργασία δεδομένων	47
4.1	Feature engineering	47
4.2	Στάθμιση κλάσεων	49
5	Ανάπτυξη προβλεπτικού μοντέλου	54
5.1	Αλγόριθμοι	54
5.2	Μετρικές αξιολόγησης μοντέλου	59
5.3	Ρύθμιση υπερπαραμέτρων	61
5.4	Αποτελέσματα	63
5.5	Συγκριτική ανάλυση κόστους	67
6	Συμπεράσματα και μελλοντικές προεκτάσεις	69

Λίστα σχημάτων

1.1	Βήματα πρόβλεψης βλαβών με τεχνικές μηχανικής μάθησης	15
2.1	Επισκόπηση λειτουργίας μοντέλου ταξινόμησης	23
3.1	Μακροσκοπική απεικόνιση της μεταβλητής S1	27
3.2	Μικροσκοπική απεικόνιση της μεταβλητής S1	28
3.3	Μακροσκοπική και μικροσκοπική απεικόνιση των μεταβλητών S2,S3,S4	29
3.4	Κατανομή των τιμών της μεταβλητής S1	31
3.5	Boxplot των τιμών της μεταβλητής S1	32
3.6	Κατανομή των τιμών της μεταβλητής S2	33
3.7	Boxplot των τιμών της μεταβλητής S2	34
3.8	Κατανομή των τιμών της μεταβλητής S3	35
3.9	Boxplot των τιμών της μεταβλητής S3	35
3.10	Κατανομή των τιμών της μεταβλητής S4	36
3.11	Boxplot των τιμών της μεταβλητής S4	36
3.12	Κατανομή των μετρήσεων S1,S2,S3,S4	37
3.13	Boxplot των μετρήσεων S1,S2,S3,S4	38
3.14	Boxplot των μετρήσεων S1,S2,S3,S4	38
3.15	Κατανομή και boxplot της μεταβλητής E2	39
3.16	Boxplot των μετρήσεων S1,S2,S3,S4	40
3.17	Χρονοσειρά των R1,R2,R3,R4	40
3.18	Γράφημα συσχετίσεων	42
3.19	Πολλαπλό scatterplot	43
3.20	Κατανομή βλαβών	44
3.21	Ανισορροπία κλάσεων	45
3.22	Χρονική εξέλιξη σφαλμάτων	45
4.1	Απεικόνιση της τεχνικής SMOTE	52
5.1	logistic function	55
5.2	KNN	57
5.3	Πίνακας σύγκρισης προβλέψεων μοντέλου ταξινόμησης	60
5.4	Results, fault type : 1	64
5.5	Results, fault type : 2	64
5.6	Results, fault type : 3	65

5.7	Results, fault type : 4	65
5.8	Results, fault type : 5	66

Λίστα πινάκων

2.1	Αρχική μορφή δεδομένων, αρχείο a	19
2.2	Αρχική μορφή δεδομένων, αρχείο b	20
2.3	Αρχική μορφή δεδομένων, αρχείο c	21
2.4	Επεξεργασμένα δεδομένα	22
3.1	Πίνακας περιγραφικών στατιστικών	27
3.2	Πίνακας συσχετίσεων	42
5.1	Ρύθμιση υπερπαραμέτρων - Βλάβη τύπου 1	62
5.2	Ρύθμιση υπερπαραμέτρων - Βλάβη τύπου 2	62
5.3	Ρύθμιση υπερπαραμέτρων - Βλάβη τύπου 3	62
5.4	Ρύθμιση υπερπαραμέτρων - Βλάβη τύπου 4	62
5.5	Ρύθμιση υπερπαραμέτρων - Βλάβη τύπου 5	63

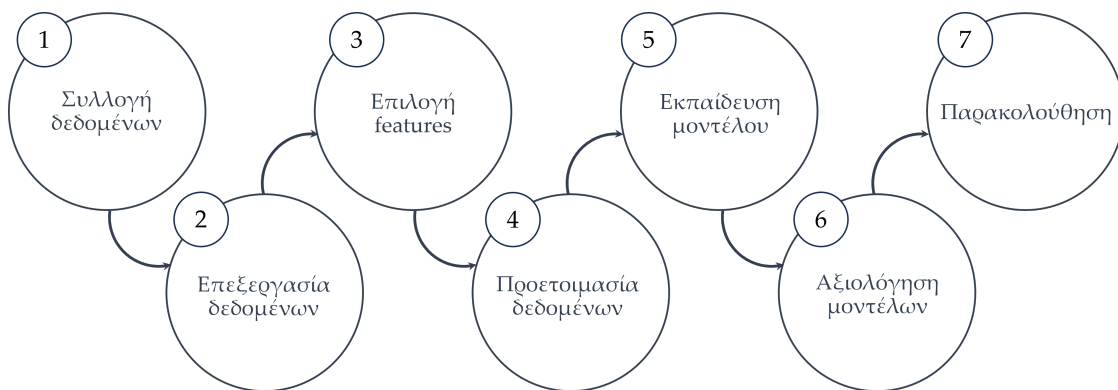
Κεφάλαιο 1

Εισαγωγή

Η ανίχνευση σφαλμάτων με χρήση μηχανικής μάθησης είναι μια προσέγγιση που αξιοποιεί προηγμένους αλγόριθμους για τον αυτόματο εντοπισμό ανωμαλιών ή αποκλίσεων από την κανονική συμπεριφορά του συστήματος. Περιλαμβάνει την εκπαίδευση μοντέλων σε ιστορικά δεδομένα, επιτρέποντάς τους να μάθουν πρότυπα και χαρακτηριστικά που σχετίζονται με την κανονική λειτουργία του συστήματος. Όταν εισάγονται νέα δεδομένα, τα μοντέλα μπορούν να ανιχνεύσουν αποκλίσεις από τα γνωστά μοτίβα και να τις επισημάνουν ως πιθανά σφάλματα ή ανωμαλίες.

Η πρόοδος που έχει επιτευχθεί στον τομέα του βιομηχανικού Internet of Things (IoT) [1] και στην ανάλυση μεγάλου όγκου δεδομένων [2] έχει γεννήσει μία μεγάλη προσδοκία για πιο αποδοτική λειτουργία της βιομηχανίας και επομένως οικονομικά οφέλη μέσα από προβλεπτικά μοντέλα [3,4] σε ένα μεγάλο εύρος εφαρμογών. Τέτοια προβλεπτικά μοντέλα αποτελούν την αιχμή του δόρατος για την νέα βιομηχανική επανάσταση που βιώνουμε αυτές τις δεκαετίες, εξαιτίας της επεκτασιμότητάς, της ευελιξίας και της ραγδαίας εφαρμογής τους.

Η διαδικασία πρόβλεψης σφαλμάτων με χρήση μηχανικής μάθησης περιλαμβάνει συνήθως τα ακόλουθα βήματα:



Σχήμα 1.1: Βήματα πρόβλεψης βλαβών με τεχνικές μηχανικής μάθησης

- Συλλογή δεδομένων: Συλλέγονται σχετικά δεδομένα από το υπό εξέταση σύστημα ή

διαδικασία. Τα δεδομένα αυτά μπορεί να περιλαμβάνουν μετρήσεις αισθητήρων, λειτουργικές παραμέτρους, σήματα ελέγχου και κάθε άλλη πληροφορία που μπορεί να βοηθήσει στον χαρακτηρισμό της συμπεριφοράς του συστήματος.

- Επεξεργασία δεδομένων: Τα δεδομένα που συλλέγονται συνήθως δεν είναι στην βέλτιστη μορφοποίηση για να αναλυθούν και να εφαρμοστούν σε αυτά μέθοδοι μηχανικής μάθησης. Έτσι, πρέπει να λάβει χώρα το λεγόμενο “data cleaning”, η προεπεξεργασία δηλ. που αντιμετωπίζει τυχόν κενά στα δεδομένα, ακραίες τιμές και θόρυβο.
- Εξαγωγή και επιλογή ανεξάρτητων μεταβλητών ή χαρακτηριστικών (features): Από τα δεδομένα εξάγονται χαρακτηριστικά για τη σύλληψη σημαντικών πληροφοριών που μπορούν να διακρίνουν μεταξύ κανονικής και ελαττωματικής συμπεριφοράς. Τα χαρακτηριστικά αυτά μπορεί να είναι για παράδειγμα στατιστικά μέτρα ή άλλες αναπαραστάσεις που αφορούν τα ειδικά χαρακτηριστικά του εκάστοτε προβλήματος. Μπορούν επίσης να χρησιμοποιηθούν τεχνικές επιλογής χαρακτηριστικών για τον εντοπισμό των πιο σχετικών χαρακτηριστικών που συμβάλλουν στην ανίχνευση σφαλμάτων.
- Προετοιμασία δεδομένων: Τα επεξεργασμένα δεδομένα χωρίζονται σε σύνολα δεδομένων εκπαίδευσης (train) και δοκιμής (test). Στα δεδομένα προστίθενται επίσης και πληροφορίες κανονικής ή ελαττωματικής λειτουργίας (label/target). Οι ελαττωματικές περιπτώσεις μπορούν να ληφθούν από ιστορικά αρχεία σφαλμάτων ή να δημιουργηθούν τεχνητά με την εισαγωγή γνωστών σφαλμάτων στο σύστημα. Τα δεδομένα εκπαίδευσης πρέπει να είναι αντιπροσωπευτικά της κανονικής και της ελαττωματικής συμπεριφοράς του συστήματος.
- Εκπαίδευση μοντέλου: Διάφοροι αλγόριθμοι μηχανικής μάθησης μπορούν να εφαρμοστούν για την εκπαίδευση μοντέλων ανίχνευσης σφαλμάτων. Οι δημοφιλείς αλγόριθμοι περιλαμβάνουν τις μηχανές διανυσμάτων υποστήριξης (SVM), τη λογιστική παλινδρόμηση, τα νευρωνικά δίκτυα και άλλους πολλούς αλγόριθμους. Τα μοντέλα εκπαιδεύονται, με στόχο να συλλάβουν τα μοτίβα που σχετίζονται με την κανονική συμπεριφορά του συστήματος και να τα διακρίνουν από τα μη φυσιολογικά μοτίβα.
- Αξιολόγηση μοντέλων: Τα εκπαιδευμένα μοντέλα αξιολογούνται χρησιμοποιώντας το σύνολο δεδομένων δοκιμής, το οποίο περιέχει άγνωστες για το μοντέλο περιπτώσεις. Διάφορες μετρικές αξιολόγησης χρησιμοποιούνται για την αξιολόγηση της απόδοσης των μοντέλων στην ανίχνευση σφαλμάτων.
- Παρακολούθηση: Αφού εκπαιδευτεί και αξιολογηθεί το μοντέλο ανίχνευσης σφαλμάτων, μπορεί να αναπτυχθεί στο επιχειρησιακό περιβάλλον. Τα δεδομένα πραγματικού χρόνου από το σύστημα παρακολουθούνται συνεχώς και όταν εντοπίζονται ανωμαλίες, μπορούν να γίνουν οι κατάλληλες ενέργειες, όπως η δημιουργία συναγερμών, η ενεργοποίηση παρεμβάσεων συντήρησης ή η έναρξη διορθωτικών μέτρων.

Στην παρούσα εργασία, το επίκεντρο είναι η λειτουργία ενός εργοστασίου και η ικανότητα να προβλέπονται γεγονότα βλάβης του εργοστασίου. Στόχος είναι η αξιοποίηση των διαθέσιμων

δεδομένων, συμπεριλαμβανομένων χρονοσειρών μετρήσεων αισθητήρων, σημάτων αναφοράς ελέγχου και πρόσθετων μετρήσεων από τις ζώνες της εγκατάστασης, για την πρόβλεψη μελλοντικών συμβάντων βλάβης των τύπων 1-5 και τον προσδιορισμό του χρόνου εμφάνισής τους. Η πρόγνωση βλαβών επιτρέπει τη βελτιστοποίηση των προγραμμάτων συντήρησης, τη μείωση του χρόνου διακοπής λειτουργίας και τη βελτίωση της απόδοσης της εγκατάστασης.

Το διαθέσιμα δεδομένα αντλήθηκαν από το υλικό του διαγωνισμού PHM 2015 Data Challenge [9] και εμπεριέχουν πληροφορίες σχετικά με τον εξοπλισμό, τις ζώνες και τα συμβάντα βλάβης 33 εργοστασίων. Κάθε εργοστάσιο χαρακτηρίζεται από τον αριθμό των εξαρτημάτων και των ζωνών που διαθέτει, ενώ όλα τα εργοστάσια καταγράφουν βλάβες από το ίδιο σταθερό σύνολο κωδικών βλάβης. Τα σφάλματα τύπου 1-5 είναι αυτά που μας απασχολούν, ενώ ο κωδικός 6 περιλαμβάνει δευτερεύοντα σφάλματα που δεν αποτελούν αντικείμενο του προβλεπτικού μας μοντέλου. Με ρυθμό δειγματοληψίας υψηλής συχνότητας περίπου ένα δείγμα κάθε 15 λεπτά, τα δεδομένα της χρονοσειράς καλύπτουν μια περίοδο περίπου τριών έως τεσσάρων ετών, παρέχοντας ένα σημαντικό ιστορικό πλαίσιο για την ανάλυση.

Για κάθε εξάρτημα διαθέτουμε μια χρονοσειρά μετρήσεων 4 αισθητήρων S_1, S_2, S_3, S_4 και μια χρονοσειρά σημάτων αναφοράς ελέγχου R_1, R_2, R_3, R_4 . Αυτές οι μετρήσεις δυναμικά καταγράφουν διάφορες λειτουργικές παραμέτρους, όπως για παράδειγμα θερμοκρασία, πίεση, ροή, τάση ή ταχύτητα περιστροφής κλπ, οι οποίες ωστόσο δεν αποκαλύφθηκαν από τους ιθύνοντες του διαγωνισμού. Με την ανάλυση αυτών των δεδομένων, καθίσταται δυνατή η παρακολούθηση της συμπεριφοράς των επιμέρους εξαρτημάτων και ο εντοπισμός ανωμαλιών ή αποκλίσεων από τα αναμενόμενα πρότυπα. Οι αλλαγές στις ενδείξεις των αισθητήρων μπορούν να χρησιμεύσουν ως δείκτες πιθανών συμβάντων αστοχίας. Ως εκ τούτου, αυτά τα δεδομένα διαδραματίζουν κρίσιμο ρόλο στην εργασία πρόγνωσης, επιτρέποντας την ανίχνευση συγκεκριμένων τύπων βλάβης.

Εκτός από τα δεδομένα των εξαρτημάτων, διαθέτουμε επίσης δεδομένα που αντιπροσωπεύουν μετρήσεις από σταθερό αριθμό ζωνών του εργοστασίου. Κάθε ζώνη καλύπτει ένα ή περισσότερα εξαρτήματα και συμβάλλει στην κατανόηση της συνολικής απόδοσης του εργοστασίου. Για παράδειγμα, εάν υπάρχουν τρεις ζώνες στο σύνολο δεδομένων, αυτές οι μετρήσεις παρέχουν πληροφορίες για τη συλλογική συμπεριφορά και τις αλληλεπιδράσεις μεταξύ των εξαρτημάτων σε κάθε ζώνη. Η ανάλυση των δεδομένων σε επίπεδο ζώνης βοηθά στην καταγραφή των χωρικών σχέσεων, των εξαρτήσεων και των συσχετίσεων που ενδέχεται να επηρεάζουν τα συμβάντα βλάβης. Με την εξέταση των δεδομένων ζώνης παράλληλα με τα δεδομένα των εξαρτημάτων, καθίσταται δυνατός ο εντοπισμός μοτίβων και ανωμαλιών που δεν είναι άμεσα εμφανείς κατά την εξέταση εξαρτημάτων μεμονωμένα. Τα δεδομένα σε επίπεδο ζώνης ενισχύουν την προγνωστική ικανότητα προσφέροντας μια ευρύτερη προοπτική και διευκολύνοντας τον εντοπισμό βλαβών που μπορεί να περιλαμβάνουν πολλαπλά εξαρτήματα ή να παρουσιάζουν χωρικά χαρακτηριστικά.

Τέλος, τα δεδομένα περιλαμβάνουν συμβάντα βλάβης του εργοστασίου, καθένα από τα οποία χαρακτηρίζεται από το χρόνο έναρξης, το χρόνο λήξης και τον κωδικό βλάβης. Αυτά τα συμβάντα βλάβης αντιπροσωπεύουν ιστορικά περιστατικά βλαβών εντός του εργοστασίου. Οι κωδικοί βλάβης επιτρέπουν την κατηγοριοποίηση και ταξινόμηση των βλαβών για σκοπούς ανάλυσης των διαφορετικών βλαβών, χωρίς να ανανδεικνύουν όμως χρήσιμα διαφοροποιητικά χαρακτηριστικά τους. Με τον ορισμό ενός σταθερού συνόλου κωδικών βλαβών σε όλες τις εγκαταστάσεις, το σύνολο δεδομένων διασφαλίζει τη συνοχή στην παρακολούθηση και την ανάλυση των συμβάντων βλάβης. Η ανάλυση ιστορικών δεδομένων είναι ζωτικής σημασίας για την κατανόηση των παραγόντων που οδηγούν σε βλάβη και την εκπαίδευση προγνωστικών μοντέλων.

Όπως προαναφέρθηκε, δεν είναι διαθέσιμες περαιτέρω χρήσιμες (όπως φυσικοχημικές, μηχανολογικές ή ηλεκτρολογικές) πληροφορίες για την μεγάλη πλειοψηφία των δεδομένων, όπως οι μετρήσεις και οι βλάβες. Έτσι, ένα σημαντικό εργαλείο σε προβλήματα μηχανικής μάθησης, όπως η αξιοποίηση επιπλέον γνώσεων πάνω στο αντικείμενο, καθίσταται απρόσιτο και το πρόβλημα λαμβάνει μια περισσότερο αμιγώς μαθηματική μορφή.

Στο επόμενο κεφάλαιο θα οριστεί το γενικό πλαίσιο του προβλήματος και της φύσης των δεδομένων, η πρωταρχική επεξεργασία αυτών, ώστε να είναι σε μια αξιοποιήσιμη μορφή, αλλά και ο τρόπος καθορισμού της μεταβλητής-στόχου. Στην συνέχεια, θα διεξαχθεί μια διερευνητική στατιστική ανάλυση των δεδομένων, αλλά και μια εξέταση της αλληλεξάρτησης των διαφορετικών μεταβλητών, ώστε να εξαχθούν συμπεράσματα για την αξιοποίηση αυτών. Στο 4ο κεφάλαιο, παρουσιάζεται η τελική ρύθμιση των ανεξάρτητων μεταβλητών του μοντέλου, αλλά και ο τρόπος αντιμετώπισης του προβλήματος των μη σταθμισμένων κλάσεων, ενώ στο τελευταίο κεφάλαιο, οι πιθανοί αλγόριθμοι του μοντέλου, η ρύθμιση των υπερπαραμέτρων τους και η παρουσίαση των τελικών αποτελεσμάτων.

Κεφάλαιο 2

Προετοιμασία δεδομένων

2.1 Αρχική επεξεργασία

Λόγω του μεγάλου όγκου των δεδομένων, η ανάλυση και διερεύνηση των δεδομένων θα γίνει αρχικά για ένα εργοστάσιο και στη συνέχεια τα αποτελέσματα θα επεκταθούν στο σύνολο των δεδομένων. Θα μελετήσουμε ως παράδειγμα το εργοστάσιο 1 το οποίο αποτελείται από 6 εξαρτήματα και 3 ζώνες.

Τα δεδομένα δίνονται σε 3 διαφορετικά αρχεία. Το αρχείο *a* περιέχει πληροφορίες για τα διάφορα εξαρτήματα. Συγκεκριμένα για κάθε χρονική στιγμή μέτρησης έχουμε για τα διάφορα εξαρτήματα τις τιμές που έδωσαν οι αισθητήρες S1,S2,S3,S4 και τα σήματα αναφοράς ελέγχου R1,R2,R3,R4. Τα δεδομένα στην αρχική τους μορφή φαίνονται στον παρακάτω πίνακα:

Component	Time	S1	S2	S3	S4	R1	R2	R3	R4
1	2009-08-18 18:12:00	711	630	69	600	689	20	40	1
5	2009-08-18 18:14:27	725	460	101	705	689	20	40	1
6	2009-08-18 18:14:31	711	505	69	678	689	20	40	1
1	2009-08-18 18:14:43	705	630	69	600	689	20	40	1
2	2009-08-18 18:14:47	734	516	101	671	689	20	40	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
6	2012-11-09 22:29:26	725	595	68	622	700	40	40	1
1	2012-11-09 22:29:36	718	727	5	539	700	40	40	1
2	2012-11-09 22:29:42	732	741	5	529	700	40	40	1
3	2012-11-09 22:29:48	738	504	69	679	700	40	40	1
4	2012-11-09 22:30:00	739	556	0	647	700	40	40	1

Πίνακας 2.1: Αρχική μορφή δεδομένων, αρχείο *a*

Επίσης έχουμε το αρχείο *b* που μας δείχνει για τις διάφορες χρονικές στιγμές μετρήσεων τη σωρευτική ενέργεια που καταναλώνεται και τη στιγμιαία ισχύ στις ζώνες του εργοστασίου που καλύπτουν ένα ή περισσότερα εξαρτήματα. Τα δεδομένα στην αρχική τους μορφή φαίνονται στον παρακάτω πίνακα:

Zone	Time	E1	E2
2	2009-08-18 18:14:56	233.72	31.52
3	2009-08-18 18:14:56	629.44	49.20
1	2009-08-18 18:15:02	863.16	80.72
2	2009-08-18 18:29:41	241.44	31.41
3	2009-08-18 18:29:42	641.31	45.38
⋮	⋮	⋮	⋮
3	2012-11-09 22:14:52	792063.13	43.38
1	2012-11-09 22:15:01	1034577.47	53.73
2	2012-11-09 22:29:54	242517.44	10.08
3	2012-11-09 22:29:54	792073.16	38.48
1	2012-11-09 22:30:03	1034590.59	48.56

Πίνακας 2.2: Αρχική μορφή δεδομένων, αρχείο *b*

Όπως είναι φανερό, το E1 αποτελεί ένα συνεχώς αυξανόμενο μέγεθος που όχι μόνο εμπεριέχει πληροφορία από χρονικές περιόδους εκτός του 3ώρου της συγκεκριμένης δειγματοληψίας, αλλά είναι και συνάρτηση αποκλειστικά του E2 και ως εκ τούτου το αποκλείουμε από τα δεδομένα εισόδου του μοντέλου. Αυτή η περίπτωση αποτελεί χαρακτηριστικό παράδειγμα, όπου οι γνώσεις του πραγματικού προβλήματος βοηθούν στην βέλτιστη ρύθμιση του μαθηματικού μοντέλου.

Τέλος μας δίνεται το αρχείο *c* που μας παρέχει πληροφορίες σχετικά με τις χρονικές στιγμές που παρατηρούνται συμβάντα βλάβης για τα οποία έχουμε το χρόνο έναρξης και λήξης της βλάβης, καθώς και τον τύπο της βλάβης. Τα δεδομένα στην αρχική τους μορφή φαίνονται στον παρακάτω πίνακα:

Όπως φαίνεται από τους παραπάνω πίνακες, οι διάφορες μετρήσεις λαμβάνονται περιοδικά, περίπου κάθε ένα τέταρτο με μια πολύ μικρή απόκλιση μερικών δευτερολέπτων ανά μέτρηση. Επειδή είναι κρίσιμο να συνδυαστούν τα δεδομένα, θεωρούμε ότι οι μετρήσεις γίνονται ακριβώς ανά ένα τέταρτο. Με άλλα λόγια κάνουμε μια διακριτοποίηση του χρόνου.

Start_Time	End_Time	Type
2009-08-31 21:44:47	2009-08-31 22:29:47	6
2009-08-31 13:59:36	2009-08-31 22:44:54	6
2009-09-01 07:44:53	2009-09-01 08:14:55	6
2009-09-01 08:44:55	2009-09-01 09:29:53	6
2009-09-01 09:44:56	2009-09-01 15:30:01	6
⋮	⋮	⋮
2012-11-08 14:14:43	2012-11-08 15:14:55	1
2012-11-08 17:44:44	2012-11-08 17:59:45	2
2012-11-08 19:44:54	2012-11-08 21:00:01	2
2012-11-08 23:59:25	2012-11-09 00:14:47	1
2012-11-09 12:14:23	2012-11-09 13:29:36	1

Πίνακας 2.3: Αρχική μορφή δεδομένων, αρχείο c

Δεδομένου ότι κάθε εργοστάσιο αποτελείται από διαφορετικό αριθμό εξαρτημάτων για καθένα από τα οποία έχουμε 4 χρονοσειρές μετρήσεων αισθητήρων S_1, S_2, S_3, S_4 και 4 χρονοσειρές σημάτων αναφοράς ελέγχου R_1, R_2, R_3, R_4 θα πρέπει με κάποιον τρόπο να παρακάμψουμε το γεγονός αυτό χωρίς να παραλείψουμε κρίσιμες πληροφορίες για τη λειτουργία του εργοστασίου. Το ίδιο ισχύει και με τα δεδομένα για την ενέργεια E_1, E_2 που αφορούν τις διάφορες ζώνες του εργοστασίου, το πλήθος των οποίων διαφέρει από εργοστάσιο σε εργοστάσιο. Έτσι, θα είμαστε σε θέση να αναλύσουμε τα δεδομένα όλων των εργοστασίων με τον ίδιο τρόπο και να αναπτύξουμε μοντέλα που μπορούν να εφαρμοστούν σε διαφορετικά εργοστάσια ασχέτως με το πόσα εξαρτήματα και ζώνες υπάρχουν στο κάθε εργοστάσιο.

Εφόσον ο αριθμός αυτός των μετρητικών διατάξεων σε κάθε εξάρτημα της παραγωγής είναι το μόνο κοινό στοιχείο ανάμεσα στα διαφορετικά εργοστάσια, γίνεται η υπόθεση πως λαμβάνει χώρα η ίδια μέτρηση σε όλα τα εξαρτήματα. Επομένως, ομαδοποιούμε τις μετρήσεις υπολογίζοντας το μέσο $S_1, \dots, S_4, R_1, \dots, R_4$ μεταξύ των εξαρτημάτων και το μέσο E_1, E_2 μεταξύ των ζωνών. Αν ορίσουμε ως S_{ij} τις μετρήσεις των αισθητήρων, R_{ij} τις τιμές αναφοράς και E_{mn} τις ενεργειακές μετρήσεις, με $i \in \{1, 2, 3, 4\}$, $m \in \{1, 2\}$, $j \in \{1, \dots, c\}$ ο αριθμός του εξαρτήματος και $n \in \{1, \dots, z\}$ ο αριθμός της ζώνης του εργοστασίου. Όπως έχει προαναφερθεί, ο αριθμός των εξαρτημάτων c και των ζωνών z , ποικίλουν σε κάθε εργοστάσιο. Έτσι, θα έχουμε τις νέες ανεξάρτητες μεταβλητές που προκύπτουν:

$$S_i = \frac{\sum_{j=1}^c S_{ij}}{c} \quad \forall i \in \{1, 2, 3, 4\}$$

$$R_i = \frac{\sum_{j=1}^c R_{ij}}{c} \quad \forall i \in \{1, 2, 3, 4\}$$

$$E_m = \frac{\sum_{n=1}^z E_{mn}}{z} \quad \forall m \in \{1, 2\}$$

Στη συνέχεια ενώνουμε τα δύο αρχεία που περιέχουν τις μετρήσεις και αποκτούμε σε πρώτη φάση τα δεδομένα που αργότερα θα χρησιμοποιηθούν για την εκπαίδευση μοντέλων μηχανικής μάθησης. Η μορφή των δεδομένων φαίνεται στον παρακάτω πίνακα:

Time	S1	S2	S3	S4	R1	R2	R3	R4	E2
2009-08-18 18:15:00	725.2	554.17	85	647	689	20	40	1	53.8
2009-08-18 18:30:00	719.8	556.33	79.67	645.67	689	20	40	1	51.2
2009-08-18 18:45:00	719.2	542.83	90.33	654.33	689	20	40	1	53.7
2009-08-18 19:00:00	713.7	536.83	90.33	658	689	20	40	1	53.4
2009-08-18 19:15:00	712	528.83	85	662.83	689	20	40	1	53.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2012-11-09 21:30:00	734.7	654.67	41.17	5840	700	40	40	1	34.6
2012-11-09 21:45:00	730.5	606.5	41.17	614.33	700	40	40	1	37.1
2012-11-09 22:00:00	730.7	640.33	30.5	593.33	700	40	40	1	30.4
2012-11-09 22:15:00	733.7	663.67	41.83	578.5	700	40	40	1	35.8
2012-11-09 22:30:00	731.8	640.67	41.17	593	700	40	40	1	32.4

Πίνακας 2.4: Επεξεργασμένα δεδομένα

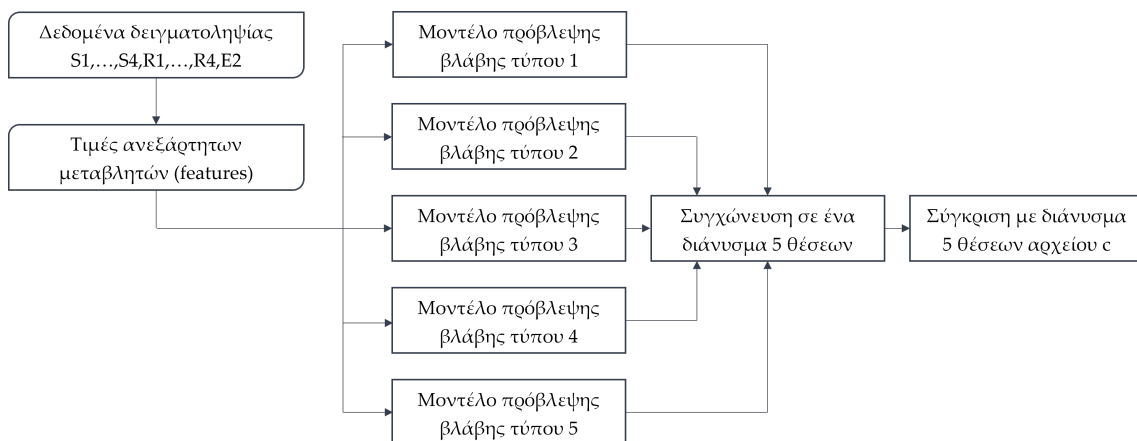
Στη συνέχεια εντοπίζουμε χρονικές στιγμές που δεν έχουμε μετρήσεις (missing values) και γεμίζουμε τις τιμές που λείπουν με τις μετρήσεις της αμέσως προηγούμενης χρονικής στιγμής που έχουμε δεδομένα. Ουσιαστικά, με τον τρόπο αυτό θεωρούμε ότι δεν άλλαξε κάτι τις στιγμές που δεν έχουμε δεδομένα και οι μετρήσεις έμειναν όπως ήταν. Άλλες προσεγγίσεις μπορούν να χρησιμοποιηθούν όπως για παράδειγμα η μέθοδος της γραμμικής παρεμβολής, ωστόσο δεν έχει μεγάλη σημασία καθώς στα δεδομένα μας οι τιμές που λείπανε ήταν 39 μέσα σε εκατοντάδες χιλιάδες δεδομένα.

2.2 Καθορισμός μεταβλητής-στόχου

Στα προβλήματα ταξινόμησης, η μεταβλητή-στόχος παίζει καθοριστικό ρόλο στον καθορισμό του στόχου του μοντέλου. Η μεταβλητή-στόχος σε ένα μοντέλο ταξινόμησης αντιπροσωπεύει το αποτέλεσμα ή την ετικέτα κλάσης που το μοντέλο στοχεύει να προβλέψει. Είναι μια κατηγορική μεταβλητή που αποδίδει κάθε περίπτωση ή παρατήρηση σε μια συγκεκριμένη κλάση ή κατηγορία. Λαμβάνει διακριτές τιμές που αντιστοιχούν σε διαφορετικές κλάσεις ή κατηγορίες. Σε κάθε περίπτωση στο σύνολο δεδομένων αποδίδεται μια ενιαία ετικέτα κλάσης με βάση τα χαρακτηριστικά ή τα γνωρίσματα που σχετίζονται με την εν λόγω περίπτωση.

Για να διευκολυνθεί η εκπαίδευση του μοντέλου, οι κατηγορικές μεταβλητές-στόχοι συχνά κωδικοποιούνται ως αριθμητικές τιμές. Η διαδικασία αυτή ονομάζεται κωδικοποίηση ετικέτας. Ωστόσο, είναι ζωτικής σημασίας να σημειωθεί ότι η αριθμητική κωδικοποίηση δεν υπονοεί κάποια εγγενή διάταξη ή μέγεθος μεταξύ των κλάσεων. Μετά την απαραίτητη επεξεργασία των δεδομένων, προχωράμε στο αρχείο c που περιέχει την πληροφορία για τις χρονικές στιγμές συμβάντων βλάβης και τον τύπο κάθε βλάβης. Συνδυάζοντάς την πληροφορία αυτή με τα μέχρι τώρα δεδομένα των μετρήσεων υπολογίζουμε για κάθε χρονική στιγμή μέτρησης το χρόνο μέχρι την επόμενη βλάβη τύπου i για κάθε $i \in \{1, 2, 3, 4, 5\}$.

Υπάρχουν δύο κύριοι τύποι προβλημάτων ταξινόμησης: η δυαδική ταξινόμηση και η ταξινόμηση πολλαπλών κατηγοριών. Στη δυαδική ταξινόμηση, η μεταβλητή-στόχος έχει δύο πιθανές κλάσεις ή ετικέτες. Παραδείγματα περιλαμβάνουν την ταξινόμηση των μηνυμάτων ηλεκτρονικού ταχυδρομείου ως spam ή ham ή την πρόβλεψη του κατά πόσον μια συναλλαγή με πιστωτική κάρτα είναι δόλια ή μη δόλια. Στην ταξινόμηση πολλαπλών κλάσεων, η μεταβλητή-στόχος έχει περισσότερες από δύο διαφορετικές κλάσεις. Για παράδειγμα, η ταξινόμηση εικόνων ζώων σε κατηγορίες όπως σκύλος, γάτα ή άλογο. Στην περίπτωσή μας, θα αντιμετωπίζαμε ένα πρόβλημα δυαδικής ταξινόμησης εάν προβλέπαμε απλώς ανάμεσα σε βλάβη και μη-βλάβη, ενώ όταν προβλέπουμε και το είδος της βλάβης έχουμε ταξινόμηση πολλαπλών κλάσεων.



Σχήμα 2.1: Επισκόπηση λειτουργίας μοντέλου ταξινόμησης

Ένα στοιχείο που θα παίζει καθοριστικό ρόλο στην ανάλυσή μας, είναι η πληροφορία ότι οι βλάβες διαφορετικού τύπου είναι ανεξάρτητα γεγονότα. Αυτό σημαίνει ότι μέσα σε ένα δεδομένο χρονικό διάστημα (πχ ένα τέταρτο, τρεις ώρες κλπ) μπορούν να συμβούν ένας οι περισσότεροι τύποι βλάβης ακόμα και ταυτόχρονα χωρίς η μία βλάβη να εξαρτάται από την άλλη. Έτσι, μπορούμε να διαχωρίσουμε το αρχικό μας πρόβλημα σε πέντε παράλληλα προβλήματα ως εξής: Κατασκευή μοντέλου πρόβλεψης για το αν θα συμβεί βλάβη τύπου 1 εντός τριώρου, κατασκευή διαφορετικού μοντέλου πρόβλεψης για το αν θα συμβεί βλάβη τύπου 2 εντός τριώρου κοκ. Στο τέλος ένα συνδυαστικό μοντέλο θα δίνει μια τελική πρόβλεψη για το τι τύποι βλάβης θα συμβούν εντός τριώρου από τη στιγμή που λαμβάνουμε τις μετρήσεις.

Με τον παραπάνω τρόπο μετατρέπουμε το αρχικό πρόβλημα πολλαπλών κλάσεων και πολλαπλών ετικετών σε πέντε δυαδικά μοντέλα μειώνοντας έτσι δραματικά την πολυπλοκότητα του προβλήματος. Σε κάθε χρονική στιγμή μέτρησης, αποδίδουμε λοιπόν μια ετικέτα 0 ή 1 για κάθε $i \in \{1,2,3,4,5\}$, όπου 0 σημαίνει ότι δεν θα συμβεί βλάβη τύπου i μέσα στο επόμενο τρίωρο και 1 σημαίνει ότι θα συμβεί βλάβη τύπου i μέσα στο επόμενο τρίωρο.

Επίσης, η κατανομή των κλάσεων στη μεταβλητή-στόχο μπορεί να ποικίλλει. Το πρόβλημα μη σταθμισμένων κλάσεων αναφέρεται σε καταστάσεις όπου υπάρχει μεγάλη ανομοιότητα στην συχνότητα εμφάνισης των διαφορετικών κλάσεων. Ο χειρισμός αυτής της ανομοιότητας αποτελεί σημαντική πρόκληση στα μοντέλα ταξινόμησης, προκειμένου να διασφαλιστεί ότι το μοντέλο δεν μεροληπτεί προς την πλειοψηφούσα κλάση και μπορεί να προβλέψει αποτελεσματικά τη μειοψηφούσα κλάση.

Τέλος, θα πρέπει να τονιστεί ένα εξαιρετικά κρίσιμο στοιχείο για τα όσα θα παρουσιαστούν παρακάτω. Γνωρίζουμε εξ αρχής ότι τα δεδομένα του προβλήματος, δηλαδή οι μετρήσεις και οι βλάβες έχουν μια σχέση εξάρτησης μέσα σε ένα χρονικό πλαίσιο 3 ωρών. Αυτό σημαίνει ότι οι βλάβες δεν εξαρτώνται από τις τιμές των αισθητήρων και των αναφορών ελέγχου που συνέβησαν παραπάνω από 3 ώρες πριν.

Εδώ μπορούμε να χρησιμοποιήσουμε δυο βασικές προσεγγίσεις. Η πρώτη είναι η εξής: Να φτιάξουμε ένα μοντέλο το οποίο δεδομένων των μετρήσεων του τελευταίου τριώρου να προβλέπει αν πρόκειται να συμβεί βλάβη στο αμέσως επόμενο τέταρτο. Αυτή η προσέγγιση εξετάστηκε ενδελεχώς, αλλά απορρίφθηκε διότι απέτυχε να επιφέρει ικανοποιητικά αποτελέσματα καθώς δημιουργεί εξαιρετικά ανισόρροπες κλάσεις που οποιοδήποτε μοντέλο αποτύγχανε να προβλέψει. Επιπλέον η πρόβλεψη σε ορίζοντα 15 λεπτών απαιτεί εξαιρετική ακρίβεια που ο έντονος θόρυβος στις μετρήσεις (όπως θα δείξουμε παρακάτω) καθιστά αδύνατον οποιοδήποτε μοντέλο να επιτύχει.

Έτσι ακολουθήθηκε η δεύτερη προσέγγιση που είναι η εξής:

Να φτιαχτεί ένα μοντέλο το οποίο, δεδομένων των μετρήσεων της εκάστοτε χρονικής στιγ-

μής, θα προβλέπει αν θα συμβεί βλάβη μέσα στο επόμενο τρίωρο. Το πλεονέκτημα αυτής της προσέγγισης είναι ότι αξιοποιεί την πληροφορία σχετικά με την εξάρτηση σφαλμάτων και μετρήσεων εντός τριώρου και ταυτόχρονα καθιστά τις δύο κλάσεις (θα συμβεί ή δεν θα συμβεί βλάβη) περισσότερο ισορροπημένες διότι το μοντέλο θα εκπαιδευτεί με τρόπο ώστε να «βλέπει» την επερχόμενη βλάβη σε κάθε χρονική στιγμή μέτρησης μέσα σε ένα χρονικό ορίζοντα τριών ωρών (12 μετρήσεις) πριν αυτή συμβεί. Το μειονέκτημα αυτής της προσέγγισης είναι ότι έχει λιγότερη χρονική ακρίβεια ως προς το πότε η βλάβη θα συμβεί. Ωστόσο, σε ένα πραγματικό περιβάλλον η γνώση ότι ένα σφάλμα θα συμβεί εντός τριώρου είναι αρκετά ικανοποιητική καθώς υπάρχει ο απαραίτητος χρόνος ώστε οι υπεύθυνοι να προβούν στις απαραίτητες ενέργειες αποτροπής της βλάβης.

Κεφάλαιο 3

Διερευνητική Ανάλυση Συνόλου Δεδομένων

Η επιτυχής εφαρμογή των αλγορίθμων μηχανικής μάθησης εξαρτάται από την ποιότητα και τις γνώσεις που προκύπτουν από την ανάλυση των δεδομένων. Η ανάλυση δεδομένων είναι το θεμελιώδες βήμα που προηγείται της εφαρμογής των αλγορίθμων μηχανικής μάθησης, προσφέροντας πολύτιμες πληροφορίες για τη δομή, τα μοτίβα και τα χαρακτηριστικά του συνόλου δεδομένων.

Η ανάλυση δεδομένων μας επιτρέπει να κατανοήσουμε τη δομή, τις μεταβλητές και τις σχέσεις των δεδομένων. Εξερευνώντας τα δεδομένα μέσω περιγραφικών στατιστικών, οπτικοποιήσεων και άλλων τεχνικών, μπορούμε να εντοπίσουμε τάσεις, κατανομές και μεμονωμένα χαρακτηριστικά των επιμέρους μεταβλητών. Η κατανόηση των δεδομένων μας βοηθά να κατανοήσουμε το πλαίσιο και το νόημα πίσω από τα δεδομένα, διασφαλίζοντας ότι λαμβάνουμε τεκμηριωμένες αποφάσεις καθ' όλη τη διάρκεια της διαδικασίας μηχανικής μάθησης. Επίσης, η ανάλυση δεδομένων βοηθά στην αποκάλυψη πιθανών μεροληπτικών τάσεων στο σύνολο δεδομένων, όπως υποεκπροσωπούμενες κλάσεις και αλληλεξαρτήσεις μεταβλητών. Είναι φανερό λοιπόν ότι αυτή η πρώτη φάση της διερευνητικής ανάλυσης είναι ζωτικής σημασίας ώστε να εφαρμοστεί το σωστό μοντέλο, οι κατάλληλες μέθοδοι και να αποφευχθούν εγκαίρως τυχόν εμπόδια.

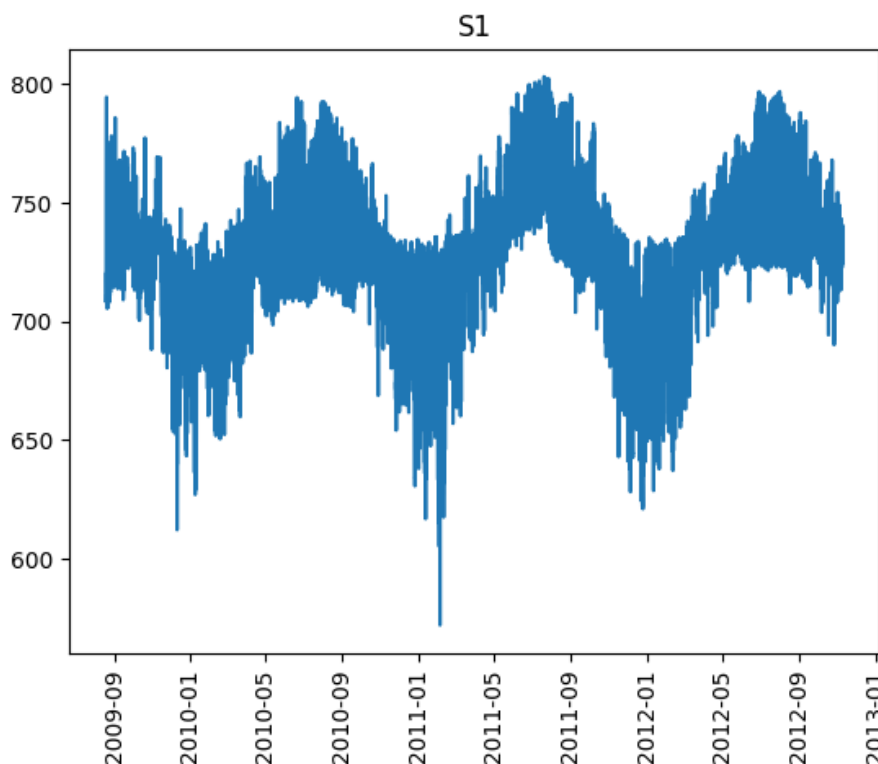
3.1 Διερεύνηση ανεξάρτητων μεταβλητών

Αρχικά, θα παρουσιαστούν σε αυτήν την ενότητα τα βασικά χαρακτηριστικά όλων των μεταβλητών των δεδομένων μας, όπως οι χρονοσειρές (σε μακρο- και μικροσκοπικό επίπεδο) τους, οι κατανομές τους, καθώς και στατιστικά στοιχεία όπως μέσος όρος και τυπική απόκλιση. Στον παρακάτω πίνακα φαίνονται συγκεντρωτικά στατιστικές πληροφορίες για τα δεδομένα μας, ενώ στη συνέχεια θα περάσουμε στην επιμέρους παρουσίαση των μεμονωμένων μεταβλητών.

	S1	S2	S3	S4	R1	R2	R3	R4	E2
mean	727.6	641.7	22.1	592.4	705.9	53.1	63.5	1.4	24.3
std	25.8	56.1	26.7	34.7	13.8	30.8	35.4	0.5	14.4
min	571.8	460.7	0	271	655.7	20	20	1	3.9
25%	715.2	599	1	568.7	700	29	40	1	0.9
50%	728.2	640.8	5	592.8	700	29	40	1	23.6
75%	739.2	679.2	40.3	618.8	709	92.3	99	2	3.9
max	803.2	1234.5	95.7	704.2	790.7	110	139	2.2	70.8

Πίνακας 3.1: Πίνακας περιγραφικών στατιστικών

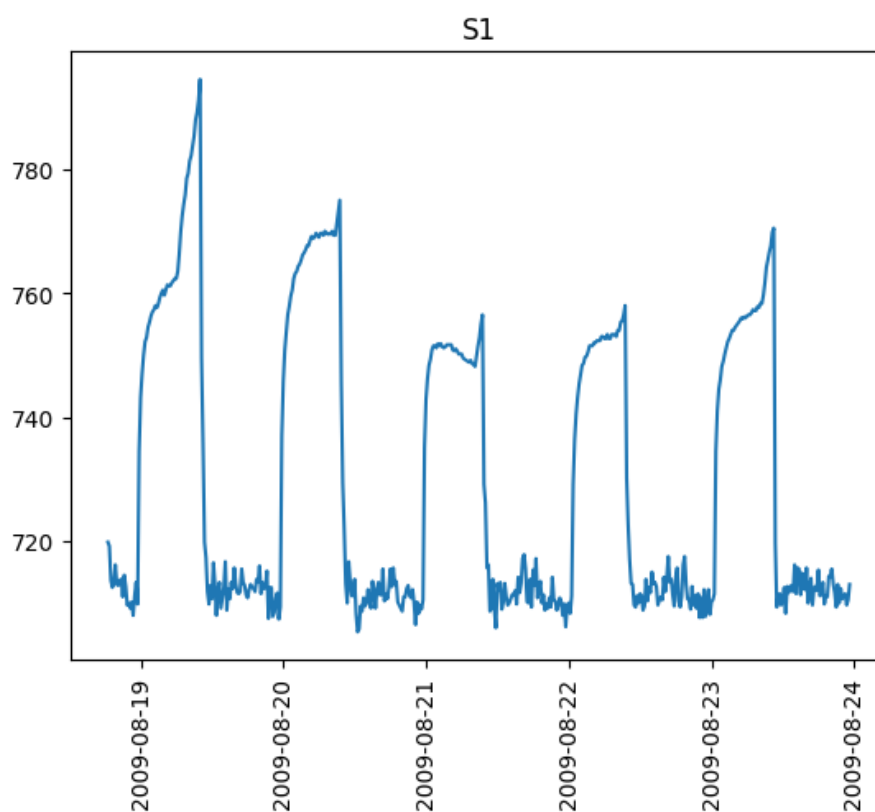
Ξεκινάμε με την πρώτη μεταβλητή, την S1. Το πρώτο γράφημα δείχνει την μακροχρόνια συμπεριφορά του:



Σχήμα 3.1: Μακροσκοπική απεικόνιση της μεταβλητής S1

Βλέπουμε ότι μακροπρόθεσμα το S1 παρουσιάζει μια έντονη εποχικότητα, καθώς στην αρχή του έτους οι τιμές είναι χαμηλές, φτάνοντας στα μέγιστα επίπεδά τους κατά το τρίτο τρίμηνο. Μπορούμε να υποθέσουμε ότι η μέτρηση αυτή είναι κάποιο περιβαλλοντικό στοιχείο όπως θερμοκρασία που μπορεί να παρουσιάζει αυτή την εποχικότητα μέσα στον χρόνο, ή να επηρεάζεται από κάποιο τέτοιο στοιχείο. Επιπροσθέτως, βλέπουμε τα χαρακτηριστικά μιας στάσιμη χρονοσειράς, που κινείται δηλαδή γύρω από ένα σταθερό μέσο, ο οποίος βρίσκεται ανάμεσα στο 700 και το 750, κάτι που επιβεβαιώνεται και από τον μέσο όρο που είναι 727.6, ενώ παρατηρούμε και κάποιες έκτοπες τιμές που θα φανούν και αργότερα στο αντίστοιχο box-plot.

Ας εξετάσουμε όμως το S1 και σε πιο μικροσκοπικό επίπεδο:

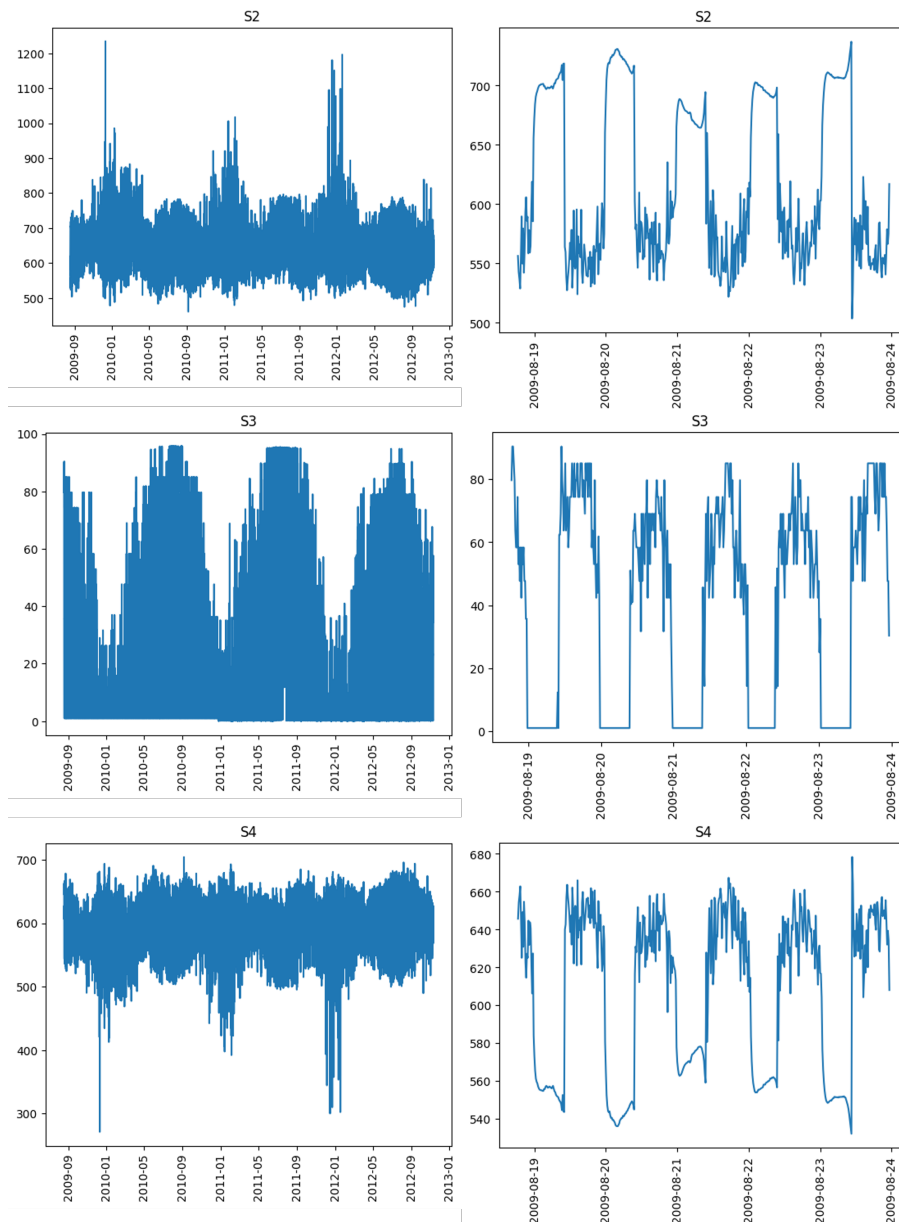


Σχήμα 3.2: Μικροσκοπική απεικόνιση της μεταβλητής S1

Παρατηρώντας το S1 σε ένα διάστημα 5 ημερών βλέπουμε ότι κι εδώ υπάρχει ξεκάθαρη περιόδικότητα από μέρα σε μέρα. Συγκεκριμένα βλέπουμε ότι στις πρώτες ώρες της ημέρας οι τιμές αυξάνονται κατακόρυφα μέχρι περίπου τα μισά της ημέρας. Τότε κάνουν μια ξαφνική κορύφωση και μειώνονται πάλι. Παρόλο που από μέρα σε μέρα το μοτίβο είναι ξεκάθαρο, υπάρχουν μικρές διαφοροποιήσεις. Αυτές οι διαφοροποιήσεις μπορεί να είναι τυχαίες και να οφείλονται σε διάφορους εξωγενείς παράγοντες που επηρεάζουν τη λειτουργία της μονάδας, ωστόσο μπορεί να σηματοδοτούν και κάποια επερχόμενη βλάβη.

Παρόμοια βραχυπρόθεσμη συμπεριφορά παρουσιάζουν και οι υπόλοιπες μετρήσεις S2,S3,S4 από την άποψη ότι παρουσιάζουν ένα ημερήσιο μοτίβο. Η μεταβλητή S2 βλέπουμε ότι παίρνει ομοίως τις μεγαλύτερες τιμές της στο πρώτο μισό της ημέρας, ενώ έχει μεγαλύτερη διακύμανση σε σχέση με την S1 στο υπόλοιπο της ημέρας. Αντιθέτως, εύκολα παρατηρείται ότι οι άλλες δύο μεταβλητές S3 και S4 λαμβάνουν χαμηλές τιμές τις πρώτες ώρες και υψηλότερες στην υπόλοιπη διάρκεια της ημέρας.

Μακροσκοπικά, οι S2 και S4 δεν παρουσιάζουν κάποιο ετήσιο μοτίβο, κάτι που εμφανίζεται αντιθέτως στην χρονοσειρά της S3. Περισσότερα έκτοπα σημεία παρουσιάζονται πάνω από τον μέσο όρο στις μεταβλητές S1 και S4, ενώ στις υπόλοιπες εμφανίζονται περισσότερες έντονες διακυμάνσεις σε τιμές υψηλότερες του μέσου όρου, όπως θα δούμε και στα αντίστοιχα boxplots στη συνέχεια.



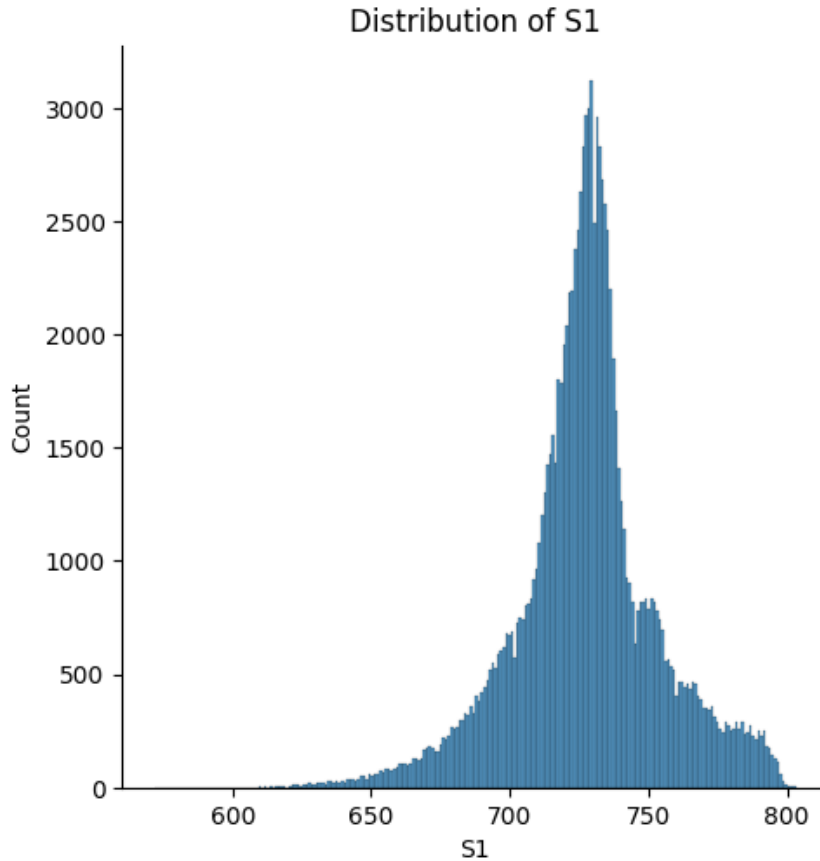
Σχήμα 3.3: Μακροσκοπική και μικροσκοπική απεικόνιση των μεταβλητών S2,S3,S4

Η έντονη περιοδική και εποχιακή συμπεριφορά που παρατηρείται στις μετρήσεις των αισθητήρων από διάφορα εξαρτήματα και ζώνες ενός εργοστασίου μπορεί να αποδοθεί σε διάφορους παράγοντες:

- **Κύκλοι παραγωγής:** Τα εργοστάσια έχουν συχνά καλά καθορισμένους κύκλους παραγωγής και διαδικασίες που παρουσιάζουν περιοδική συμπεριφορά. Οι κύκλοι αυτοί μπορεί να σχετίζονται με βάρδιες, προγράμματα εργασίας, ρουτίνες συντήρησης ή παραγωγή παρτίδων. Οι μετρήσεις των αισθητήρων μπορεί να καταγράφουν διακυμάνσεις στους ρυθμούς παραγωγής, την κατανάλωση ενέργειας ή τη χρήση πόρων, οδηγώντας σε περιοδικά μοτίβα.
- **Λειτουργία εξοπλισμού:** Διαφορετικά εξαρτήματα και ζώνες εντός ενός εργοστασίου μπορεί να αντιστοιχούν σε συγκεκριμένο εξοπλισμό ή μηχανήματα. Η λειτουργία αυτού του εξοπλισμού μπορεί να ακολουθεί τακτικά πρότυπα ή κύκλους, όπως ακολουθίες εκκίνησης και τερματισμού λειτουργίας, προγράμματα συντήρησης ή βάρδιες λειτουργίας. Αυτοί οι λειτουργικοί κύκλοι μπορούν να αντικατοπτρίζονται στις μετρήσεις του αισθητήρα ως περιοδική συμπεριφορά.
- **Περιβαλλοντικές συνθήκες:** Το περιβάλλον του εργοστασίου, συμπεριλαμβανομένων παραγόντων όπως η θερμοκρασία, η υγρασία και ο φωτισμός, μπορεί να παρουσιάζει εποχιακές διακυμάνσεις. Οι εποχιακές αλλαγές, όπως οι καιρικές συνθήκες ή οι ώρες φωτός της ημέρας, μπορούν να επηρεάσουν τη συμπεριφορά ορισμένων διεργασιών ή να επηρεάσουν την αποδοτικότητα του εξοπλισμού. Αυτές οι περιβαλλοντικές μεταβολές μπορεί να συμβάλλουν στα παρατηρούμενα εποχιακά μοτίβα στις μετρήσεις των αισθητήρων.
- **Ζήτηση ενέργειας:** Η χρήση ενέργειας σε ένα εργοστάσιο μπορεί να παρουσιάζει περιοδική και εποχιακή συμπεριφορά λόγω παραγόντων όπως η ζήτηση παραγωγής, οι αλλαγές στα χρονοδιαγράμματα εργασίας ή οι διακυμάνσεις στις περιβαλλοντικές συνθήκες. Ορισμένες διαδικασίες παραγωγής ή εξοπλισμός μπορεί να έχουν διαφορετικές ενεργειακές απαιτήσεις, με αποτέλεσμα περιοδικά ή εποχιακά πρότυπα στα δεδομένα κατανάλωσης ενέργειας.
- **Ρουτίνες συντήρησης και καθαρισμού:** Οι τακτικές δραστηριότητες συντήρησης, τα χρονοδιαγράμματα καθαρισμού ή οι επιθεωρήσεις μπορούν να εισάγουν περιοδικά μοτίβα στις μετρήσεις των αισθητήρων. Οι δραστηριότητες αυτές μπορεί να εκτελούνται σε τακτά χρονικά διαστήματα, οδηγώντας σε αξιοσημείωτες διακυμάνσεις στα δεδομένα.
- **Ανθρώπινη συμπεριφορά και πρότυπα εργασίας:** Οι ανθρώπινοι παράγοντες παίζουν ρόλο στις λειτουργίες του εργοστασίου και τα πρότυπα εργασίας μπορούν να εισάγουν περιοδικότητα στις μετρήσεις του αισθητήρα. Οι αλλαγές βάρδιας, τα διαλείμματα ή τα χρονοδιαγράμματα των εργαζομένων μπορούν να επηρεάσουν τους ρυθμούς παραγωγής, τη χρήση των πόρων ή άλλες πτυχές που καταγράφονται από τους αισθητήρες. Το γεγονός λοιπόν πως για το συγκεκριμένο σύνολο δεδομένων δε δίνονται επαρκείς

πληροφορίες είναι εξαιρετικά σημαντική πρόκληση γιατί δεν μπορούν να αξιολογηθούν οι πιθανές αιτίες με όσα παρατηρούνται στις μετρήσεις.

Προχωράμε τώρα στις κατανομές των μετρήσεων S1, S2, S3, S4.



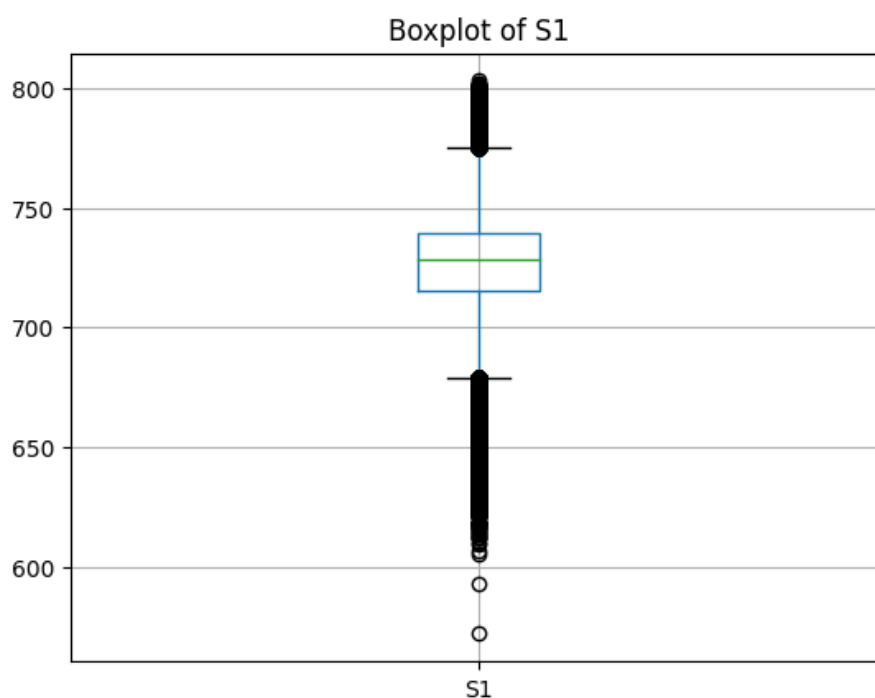
Σχήμα 3.4: Κατανομή των τιμών της μεταβλητής S1

Η κατανομή του S1 είναι αρκετά συμμετρική και ομοιάζει με κανονική κατανομή. Αυτό φαίνεται επίσης από την εγγύτητα του μέσου όρου (727.6) και της διαμέσου (728.2), καθώς και από το ότι το ένα τέταρτο των μετρήσεων ανήκουν στο διάστημα $[MO-12.4, MO]$ και άλλο ένα τέταρτο των μετρήσεων στο διάστημα $[MO, MO+11.6]$. Παρατηρείται λίγο μεγαλύτερη συγκέντρωση των μετρήσεων κοντά στον μέσο όρο σε σχέση με μια κανονική κατανομή, όπου περίπου το 19% των μετρήσεων απέχει έως μισή τυπική απόκλιση από τον μέσο όρο, αλλά και διεύρυνση της καμπύλης προς τις χαμηλότερες τιμές. Στη μηχανική μάθηση, συχνά προτιμάται οι μεταβλητές να ακολουθούν μια κανονική κατανομή για διάφορους λόγους [13] :

- Απλοποιεί τις παραδοχές μοντελοποίησης: Πολλοί αλγόριθμοι μηχανικής μάθησης υποθέτουν ότι τα δεδομένα ακολουθούν κανονική κατανομή. Για παράδειγμα, η γραμμική παλινδρόμηση, η λογιστική παλινδρόμηση και οι αλγόριθμοι Gaussian naïve Bayes υποθέτουν ότι η μεταβλητή-στόχος είναι κανονικά κατανομημένη ή ότι η σχέση μεταξύ των ανεξάρτητων μεταβλητών και της μεταβλητής-στόχου είναι γραμμική όταν υποθέτουν κανονικότητα.

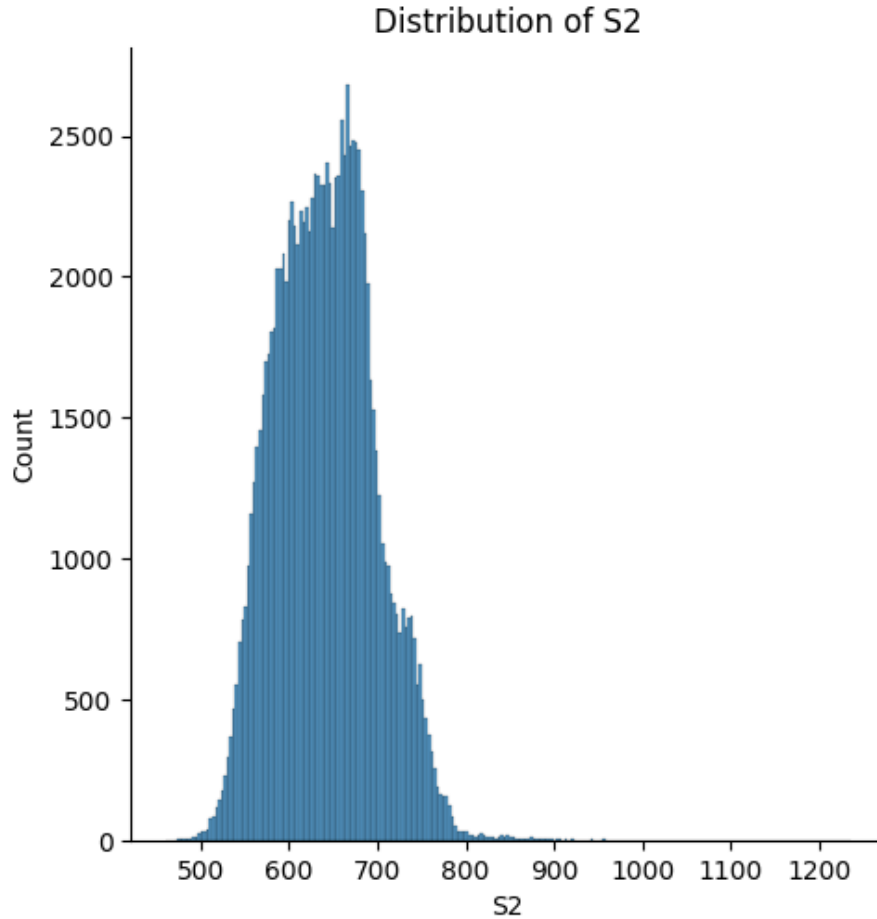
- Στατιστικές μέθοδοι: Η κανονική κατανομή έχει πολλές επιθυμητές ιδιότητες. Επιτρέπει τη χρήση διαφόρων στατιστικών μεθόδων, όπως ο έλεγχος υποθέσεων, τα διαστήματα εμπιστοσύνης και η εκτίμηση των παραμέτρων. Εάν οι ανεξάρτητες μεταβλητές είναι κανονικά κατανομημένες, διευκολύνεται η εφαρμογή αυτών των στατιστικών τεχνικών για την ανάλυση και την ερμηνεία των δεδομένων.
- Απλοποιεί την κλιμάκωση χαρακτηριστικών (feature scaling): Ορισμένοι αλγόριθμοι μηχανικής μάθησης, όπως η γραμμική παλινδρόμηση, είναι ευαίσθητοι στην κλίμακα των ανεξάρτητων μεταβλητών. Όταν οι μεταβλητές κατανομονται κανονικά, αυτό συχνά συνεπάγεται ότι βρίσκονται σε παρόμοια κλίμακα, γεγονός που διευκολύνει την εφαρμογή τεχνικών feature scaling, όπως η τυποποίηση (μέσος όρος 0 και τυπική απόκλιση 1).
- Σταθερότητα της εκτίμησης: Σε ορισμένες περιπτώσεις, η παραδοχή της κανονικότητας των μεταβλητών μπορεί να οδηγήσει σε πιο σταθερή εκτίμηση των παραμέτρων του μοντέλου. Για παράδειγμα, η εκτίμηση μέγιστης πιθανοφάνειας χρησιμοποιείται συνήθως σε πολλά μοντέλα και προϋποθέτει κανονικότητα των δεδομένων. Όταν τα χαρακτηριστικά ακολουθούν κανονική κατανομή, μπορεί να οδηγήσει σε πιο εύρωστες και σταθερές εκτιμήσεις παραμέτρων.

Επιστρέφοντας στην μεταβλητή S1, η συγκέντρωση περισσότερων μετρήσεων κοντά στον μέσο όρο, έχει σαν αποτέλεσμα την εμφάνιση περισσότερων έκτοπων σημείων εκατέρωθεν του μέσου όρου, στο αντίστοιχο boxplot, ενώ οι περισσότερες παρατηρούνται στην δεξιά ουρά όπως προαναφέραμε.



Σχήμα 3.5: Boxplot των τιμών της μεταβλητής S1

Σε ότι αφορά την μεταβλητή $S2$, βλέπουμε άλλη μια αρκετά συμμετρική κατανομή με πολύ μικρή απόκλιση μέσου όρου (641.7) και διαμέσου (640.8).

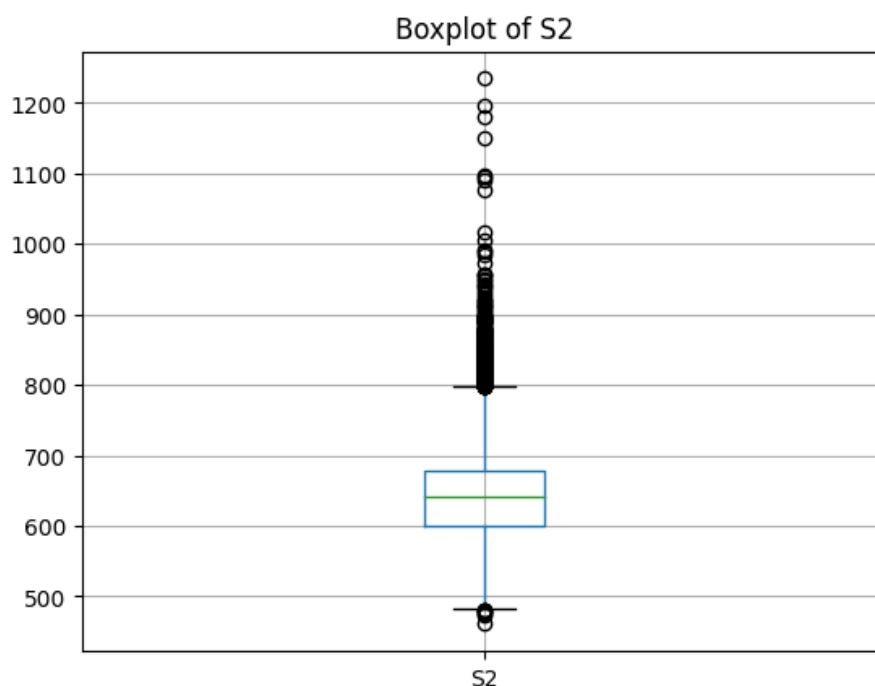


Σχήμα 3.6: Κατανομή των τιμών της μεταβλητής $S2$

Εδώ παρατηρείται, ωστόσο, μια μεγαλύτερη ασυμμετρία (skewness) ως προς τις έκτοπες τιμές, οι οποίες εμφανίζονται κατα κύριο λόγο στην δεξιά ουρά της κατανομής, κάτι που μαρτυρά και η ασυμμετρία της ελάχιστης (460.7) και της μέγιστης τιμής (1234.5). Στο σημείο αυτό θα πρέπει να γίνει λόγος σχετικά με την ύπαρξη ακραίων τιμών σε προβλήματα μηχανικής μάθησης.

Οι ακραίες τιμές στις ανεξάρτητες μεταβλητές μπορούν να έχουν διαφορετικές επιπτώσεις σε ένα πρόβλημα μηχανικής μάθησης και η χρησιμότητά τους εξαρτάται από το συγκεκριμένο πλαίσιο και τη φύση των ακραίων τιμών. Οι ακραίες μερικές φορές υποδηλώνουν σφάλματα στη συλλογή ή τη μέτρηση δεδομένων. Σε τέτοιες περιπτώσεις, οι ακραίες τιμές δεν είναι χρήσιμες και μπορούν να εισάγουν θόρυβο ή μεροληψία στο μοντέλο. Ορισμένοι αλγόριθμοι μηχανικής μάθησης είναι ευαίσθητοι στις ακραίες τιμές και μπορεί να επηρεαστούν αρνητικά από την παρουσία τους. Οι ακραίες τιμές μπορούν να διαστρεβλώσουν τις σχέσεις μεταξύ των

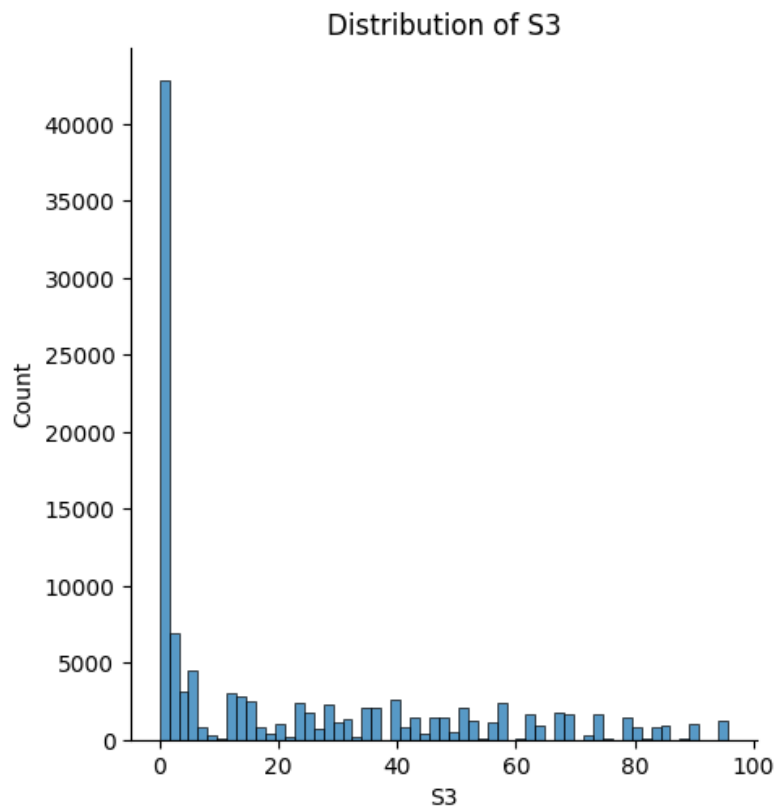
χαρακτηριστικών και της μεταβλητής-στόχου, οδηγώντας σε μη βέλτιστη απόδοση του μοντέλου. Μπορεί επίσης να έχουν σημαντική επίδραση στη διαδικασία μάθησης του μοντέλου. Εάν οι ακραίες τιμές αντιπροσωπεύουν θόρυβο ή άσχετα μοτίβα, μπορεί να οδηγήσουν σε υπερπροσαρμογή (overfitting), όπου το μοντέλο απομνημονεύει τις ακραίες τιμές αντί να μαθαίνει τα υποκείμενα μοτίβα. Αυτό μπορεί να οδηγήσει σε κακή γενίκευση σε άγνωστα δεδομένα.



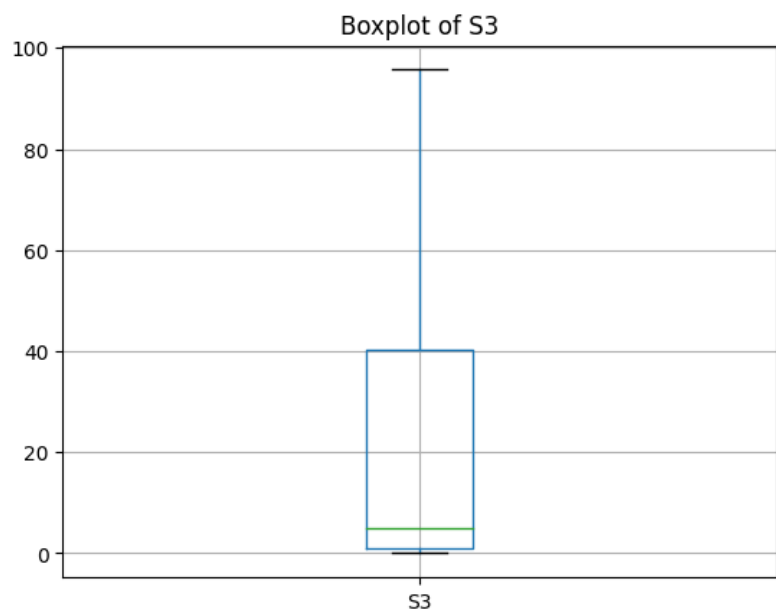
Σχήμα 3.7: Boxplot των τιμών της μεταβλητής S2

Από την άλλη, οι ακραίες τιμές μπορεί να αντιπροσωπεύουν ακραίες ή σπάνιες περιπτώσεις που παρουσιάζουν ιδιαίτερο ενδιαφέρον ή σημασία σε ορισμένες εφαρμογές. Σε αυτές τις περιπτώσεις, οι ακραίες τιμές παρέχουν πολύτιμες πληροφορίες για ανωμαλίες ή ιδιαίτερες συμπεριφορές εντός των δεδομένων. Στην δική μας περίπτωση από τη φύση του προβλήματος που μελετάμε, μπορούμε να καταλάβουμε ότι οι ακραίες τιμές αποτελούν πιθανότατα ανωμαλίες στις χρονοσειρές των μετρήσεων πράγμα που μπορεί να βοηθήσει το μοντέλο να καταλάβει ότι υπάρχει κάποια επερχόμενη βλάβη.

Συνεχίζοντας με την μεταβλητή S3, παρατηρούμε ότι η κατανομή της είναι αρκετά διαφορετική καθώς όπως είδαμε νωρίτερα οι τιμές του συγκεκριμένου αισθητήρα είναι κοντά στο μηδέν για αρκετά μεγάλα χρονικά διαστήματα, περιοδικά, κάθε μέρα. Η ασυμετρία αυτή φανερώνεται επίσης και από την ουσιώδη διαφορά ανάμεσα σε μέσο όρο (22.1) και διάμεσο (1). Λόγω αυτής της ασυμετρίας παρατηρείται και αρκετά μεγάλη τυπική απόκλιση (26.7), η οποία συμβάλει στο να μην φαίνονται outliers στο boxplot της μεταβλητής αυτής.

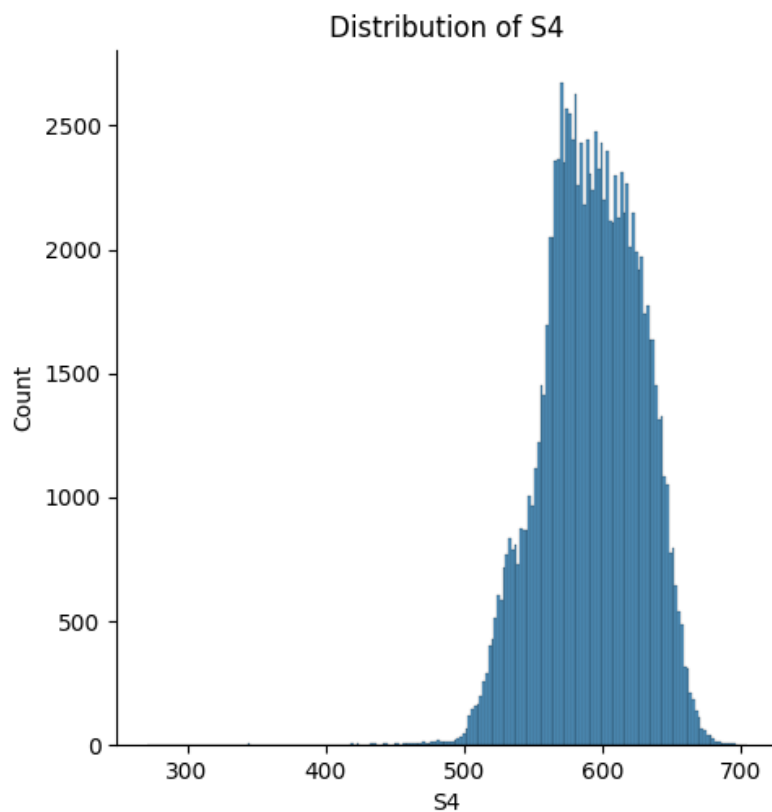


Σχήμα 3.8: Κατανομή των τιμών της μεταβλητής S3

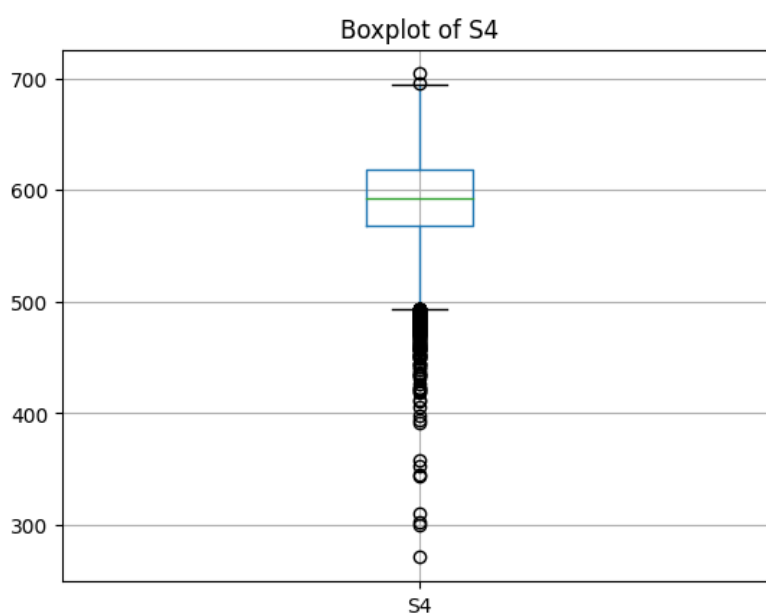


Σχήμα 3.9: Boxplot των τιμών της μεταβλητής S3

Τέλος για τις μετρήσεις του αισθητήρα S4 έχουμε την ακόλουθη κατανομή:



Σχήμα 3.10: Κατανομή των τιμών της μεταβλητής S4

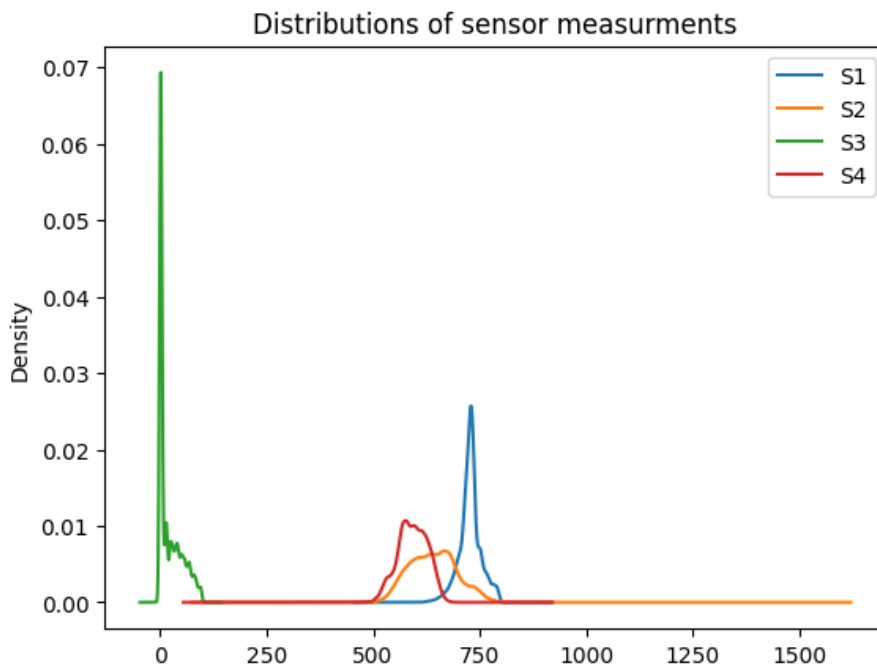


Σχήμα 3.11: Boxplot των τιμών της μεταβλητής S4

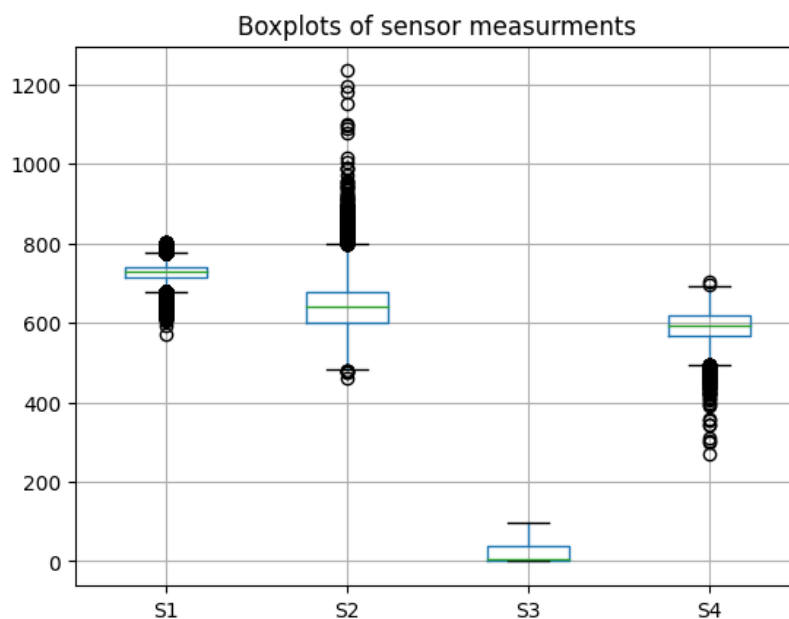
Και εδώ η κατανομή είναι αρκετά συμμετρική με την διάμεσο (592.8) να βρίσκεται πολύ κοντά στον μέσο όρο (592.4), ενώ παρατηρείται μία ασυμμετρία ως προς τις έκτοπες τιμές, ομοίως με την μεταβλητή S2, αλλά στην αριστερή ουρά της κατανομής, όπως φαίνεται και από το αντίστοιχο boxplot.

Ας σημειωθεί σε αυτό το σημείο, ότι παρατηρούνται πολλές ομοιότητες στις χρονοσειρές, τόσο σε μακροσκοπικό όσο και σε μικροσκοπικό επίπεδο, στις κατανομές αλλά και στα boxplot των μεταβλητών S2 και S4. Στις μεν κατανομές παρατηρείται συμμετρία ως προς τον μέσο όρο, ενώ στα υπόλοιπα μια αξονική συμμετρία ως προς τον οριζόντιο άξονα, οι οποίες οφείλονται στην αλληλεξάρτηση των δύο αυτών μεταβλητών που θα διαπιστώσουμε στα επόμενα κεφάλαια.

Συγκεντρωτικά οι κατανομές των μετρήσεων από τους 4 αισθητήρες φαίνονται στα παρακάτω γραφήματα:

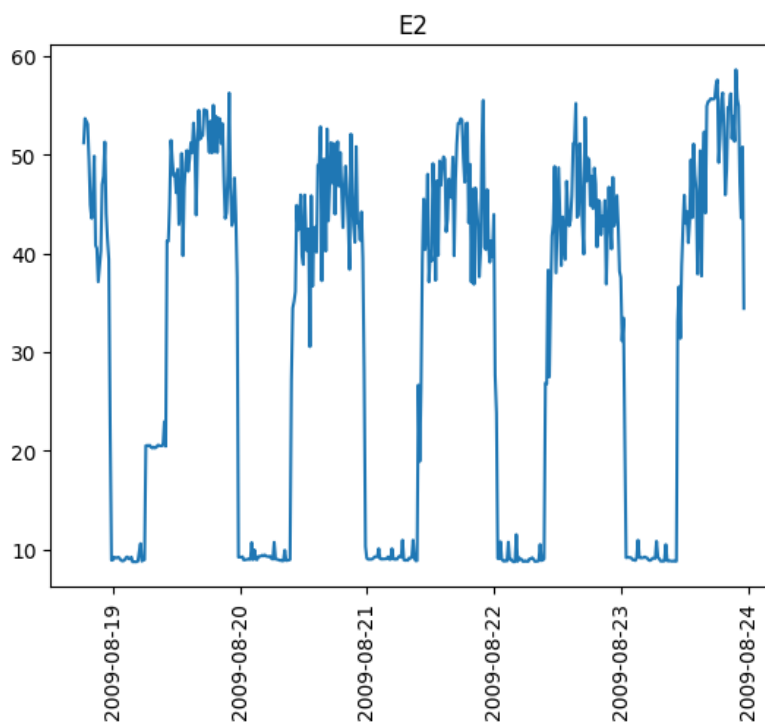


Σχήμα 3.12: Κατανομή των μετρήσεων S1,S2,S3,S4



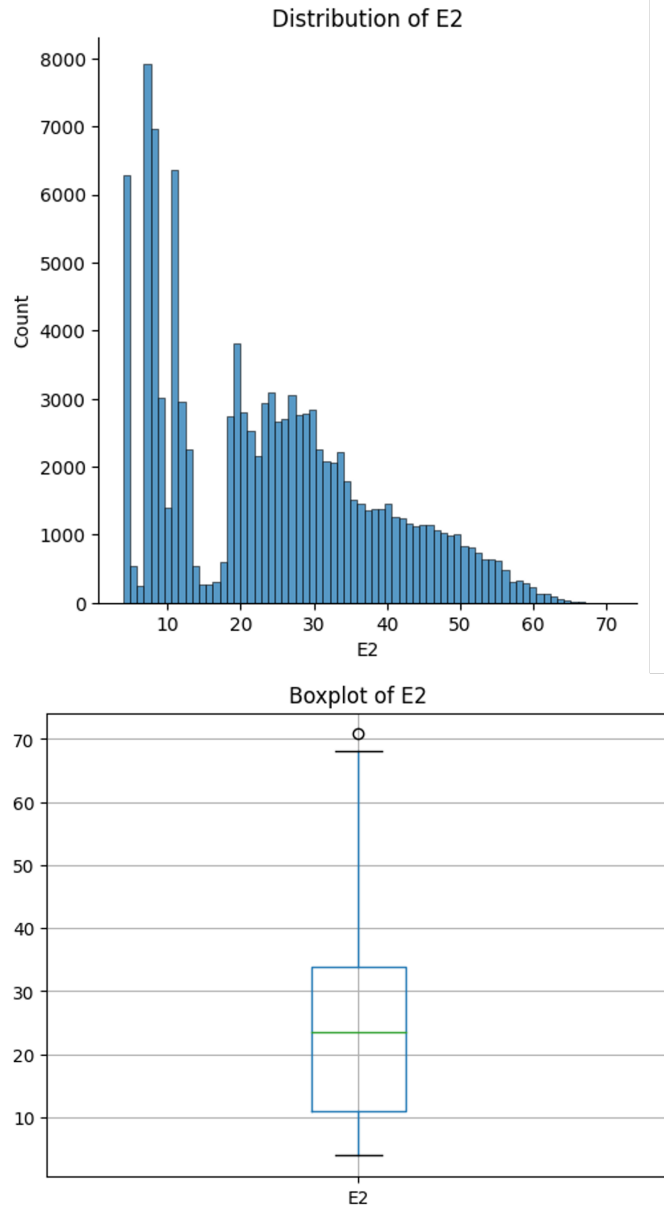
Σχήμα 3.13: Boxplot των μετρήσεων S1,S2,S3,S4

Προχωράμε τώρα στην εξερεύνηση του E2. Σε ημερήσια βάση παρατηρείτε στασιμότητα των μετρήσεων σε χαμηλά επίπεδα τις πρώτες ώρες και μια αύξηση αργότερα μέσα στην ημέρα, ενώ σε ετήσια βάση παρατηρούνται μεγαλύτερες τιμές στο τρίτο τρίμηνο του έτους, ομοίως με τις μεταβλητές S1 και S3.

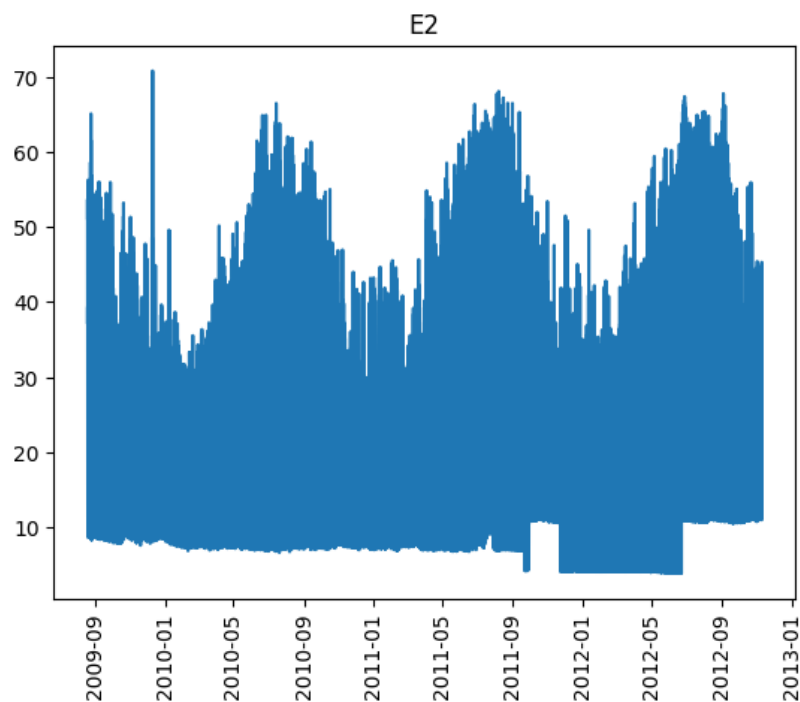


Σχήμα 3.14: Boxplot των μετρήσεων S1,S2,S3,S4

Τόσο μακροπρόθεσμα όσο και βραχυπρόθεσμα και το E2 παρουσιάζει περιοδικότητα και σαφές μοτίβο ενώ η κατανομή της μοιάζει περισσότερο με μίξη κατανομών γεγονός που υποδηλώνει ότι υπάρχει κάποιος παράγοντας που επηρεάζει τις τιμές που παίρνει η συγκεκριμένη μέτρηση. Ένας τέτοιος παράγοντας μπορεί να σχετίζεται με την ύπαρξη ή τον ερχομό κάποιας βλάβης, κάτι που τα μοντέλα μηχανικής μάθησης θα μπορέσουν να συλλάβουν.

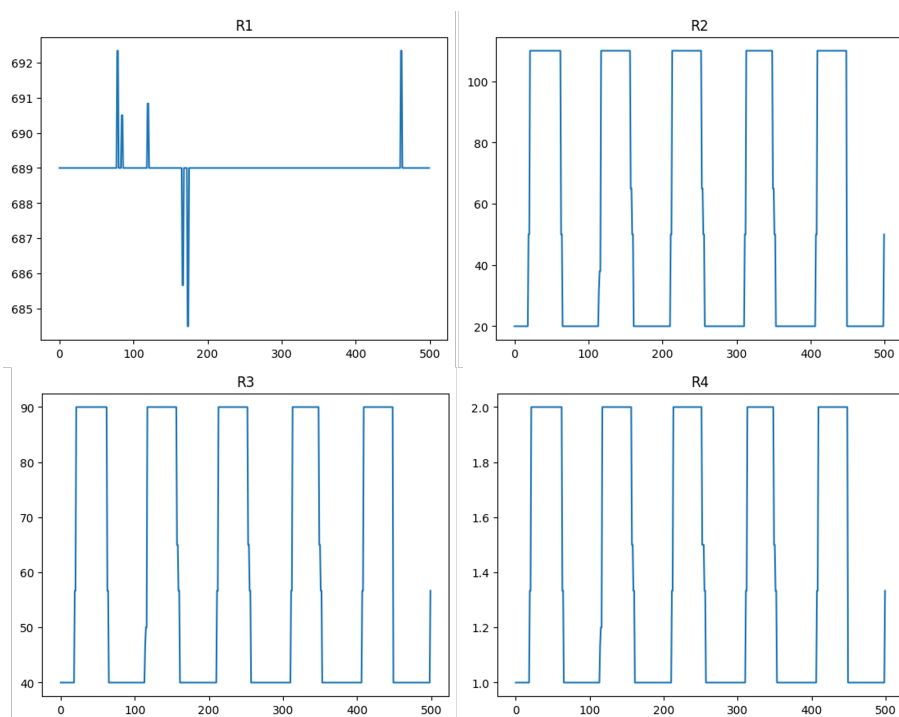


Σχήμα 3.15: Κατανομή και boxplot της μεταβλητής E2



Σχήμα 3.16: Boxplot των μετρήσεων S1,S2,S3,S4

Σημαντικό ρόλο θα παίξουν επίσης και οι αναφορές ελέγχου R1,R2,R3,R4. Ας δούμε τη συμπεριφορά τους μέσα σε ένα διάστημα περίπου 5 ημερών:



Σχήμα 3.17: Χρονοσειρά των R1,R2,R3,R4

Παρατηρούμε ότι τα R2,R3,R4 παρουσιάζουν παρόμοια συμπεριφορά, δηλαδή παίρνουν εναλλάξ τιμές στις οποίες παραμένουν για σταθερά χρονικά διαστήματα, κάτι αναμενόμενο για παραμέτρους λειτουργίας μιας γραμμής παραγωγής. Διαφορετική συμπεριφορά παρουσιάζει το R1 που είναι σταθερό κάνοντας απότομες μεταβολές προς άλλες τιμές οι οποίες διαρκούν για λίγο, κάτι που μπορεί ενδεχομένως να φανερώνει κάποια βλάβη στο μοντέλο μας. Παρόλο που οι τιμές των αναφορών ελέγχου δεν παρουσιάζουν μεγάλη μεταβλητότητα, μπορεί να είναι πολύ χρήσιμες για την προβλεπτική ικανότητα ενός μοντέλου μηχανικής μάθησης και, πράγματι, βελτίωσαν τα αποτελέσματα του μοντέλου όταν συμπεριλήφθηκαν στις εισόδους του.

3.2 Συσχετίσεις ανεξάρτητων μεταβλητών

Εξαιρετικά σημαντική είναι επίσης και η αλληλεξάρτηση των ανεξάρτητων μεταβλητών του μοντέλου (features). Η υψηλή συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών μπορεί να οδηγήσει σε πολυσυγγραμμικότητα, η οποία συμβαίνει όταν δύο ή περισσότερες μεταβλητές συσχετίζονται σε μεγάλο βαθμό μεταξύ τους. Η πολυσυγγραμμικότητα μπορεί να δημιουργήσει δυσκολίες στην ερμηνεία του μοντέλου και μπορεί να επηρεάσει τη σταθερότητα και την αξιοπιστία των συντελεστών του μοντέλου ή των κατατάξεων σπουδαιότητας των μεταβλητών. Μπορεί, επίσης, να δυσχεράνει τον προσδιορισμό των επιμέρους συνεισφορών των συσχετιζόμενων μεταβλητών στη μεταβλητή-στόχο [16].

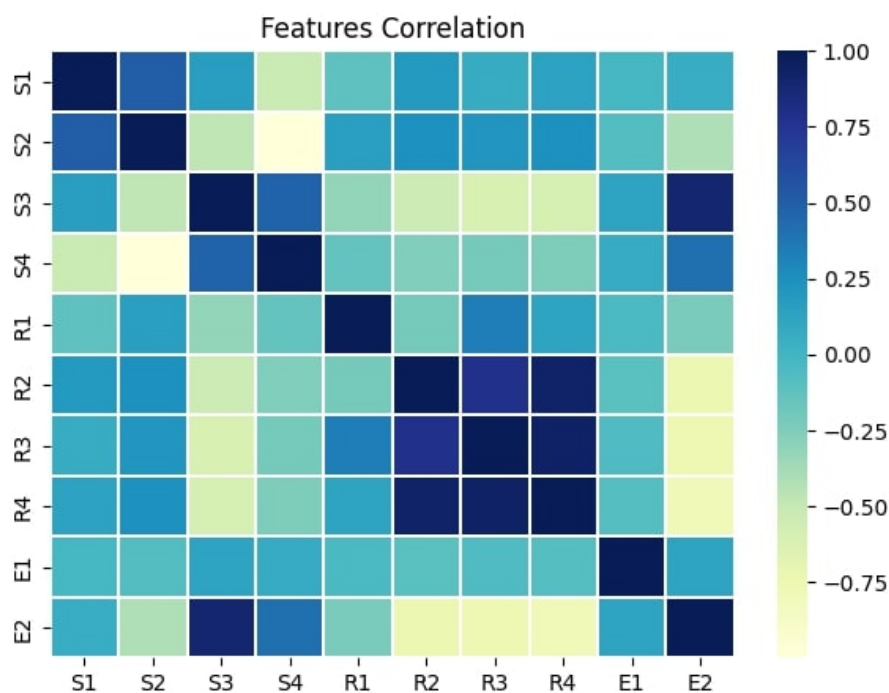
Επίσης οι έντονα συσχετιζόμενες μεταβλητές μπορεί να παρέχουν πλεονάζουσες ή επικαλυπτόμενες πληροφορίες. Η συμπερίληψη τέτοιων μεταβλητών στο μοντέλο δεν προσθέτει απαραίτητα επιπλέον αξία και μπορεί να οδηγήσει σε υπερπροσαρμογή. Τα περιττά χαρακτηριστικά μπορεί να αυξήσουν την πολυπλοκότητα του μοντέλου χωρίς να βελτιώσουν σημαντικά την προβλεπτική απόδοση.

Επιπλέον, οι συσχετιζόμενες μεταβλητές μπορούν να εισάγουν αστάθεια στις προβλέψεις του μοντέλου. Μικρές αλλαγές ή διαταραχές στα δεδομένα μπορούν να οδηγήσουν σε σημαντικές διακυμάνσεις στις προβλέψεις του μοντέλου, καθώς το μοντέλο μπορεί να βασίζεται σε μεγάλο βαθμό στις συσχετισμένες μεταβλητές. Αυτό μπορεί να καταστήσει το μοντέλο ευαίσθητο στο θόρυβο και να μειώσει την ικανότητα γενίκευσής του, το οποίο επηρεάζει την απόδοση του μοντέλου μηχανικής μάθησης.

Οι γραμμικές συσχετίσεις των μεταβλητών (συντελεστής Pearson) στο πρόβλημά μας φαίνονται στον παρακάτω πίνακα με ακρίβεια 2 δεκαδικών ψηφίων, καθώς επίσης και στο σχήμα 3.18 που ακολουθεί:

	S1	S2	S3	S4	R1	R2	R3	R4	E2
S1	1.00	0.50	0.16	-0.52	-0.12	0.19	0.07	0.14	0.06
S2	0.50	1.00	-0.47	-1.00	0.15	0.25	0.21	0.24	-0.41
S3	0.16	-0.47	1.00	0.47	-0.31	-0.54	-0.61	-0.60	0.90
S4	-0.52	-1.00	0.47	1.00	-0.15	-0.25	-0.22	-0.25	0.41
R1	-0.12	0.15	-0.31	-0.15	1.00	-0.22	0.34	0.13	-0.23
R2	0.19	0.25	-0.54	-0.25	-0.22	1.00	0.78	0.92	-0.74
R3	0.07	0.21	-0.61	-0.22	0.34	0.78	1.00	0.94	-0.76
R4	0.14	0.24	-0.60	-0.25	0.13	0.92	0.94	1.00	-0.79
E2	0.06	-0.41	0.90	0.41	-0.23	-0.74	-0.76	-0.79	1.00

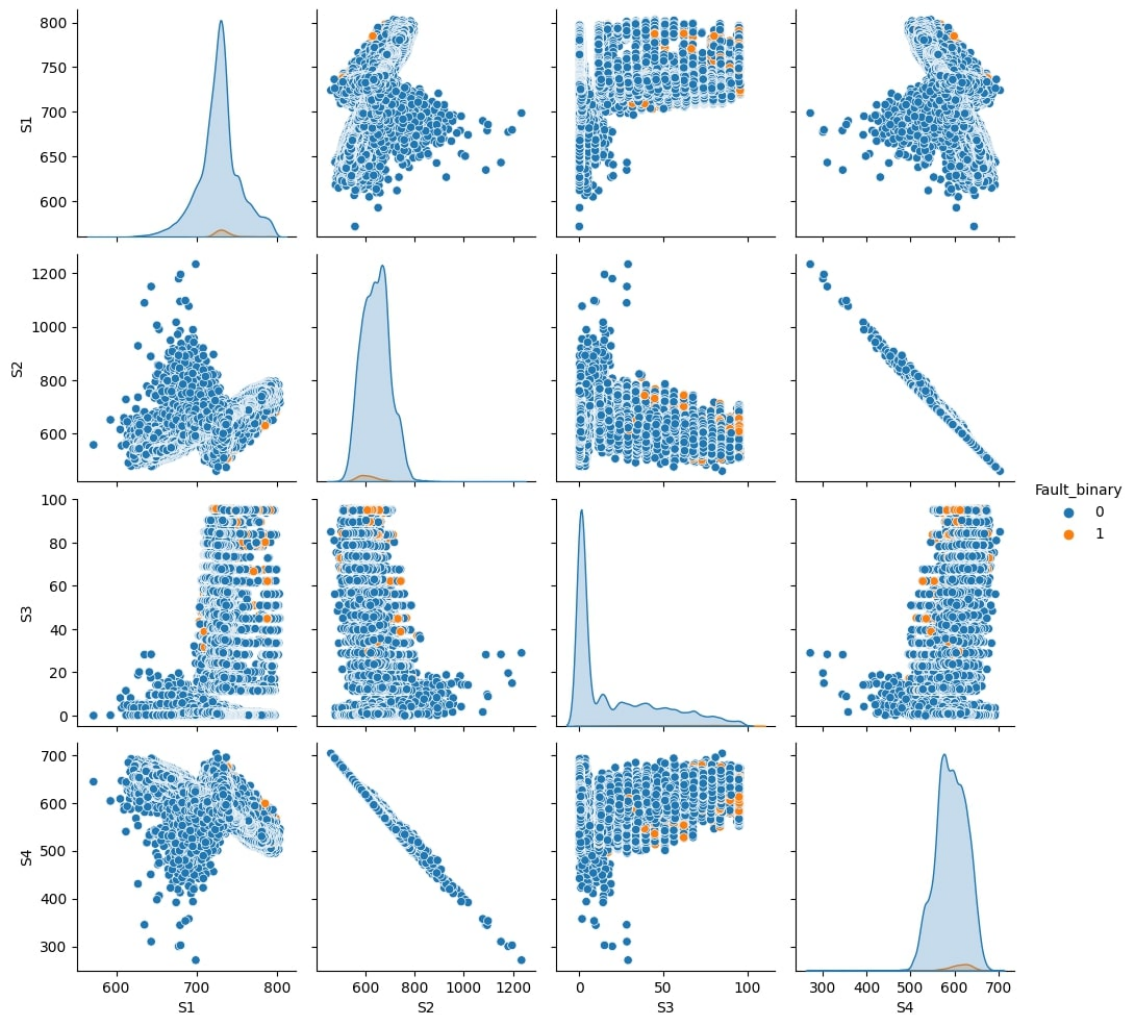
Πίνακας 3.2: Πίνακας συσχετίσεων



Σχήμα 3.18: Γράφημα συσχετίσεων

Βλέπουμε ότι σε γενικές γραμμές δεν υπάρχει μεγάλη συσχέτιση μεταξύ μεταξύ των περισσότερων μετρήσεων. Δηλαδή η συσχέτισή τους κυμαίνεται από -0.5 έως 0.5 . Είναι λογικό να υπάρχει αυτή η μικρή συσχέτιση καθώς οι μετρήσεις αφορούν διαφορετικά πεδία του ίδιου εργοστασίου. Μεγάλη συσχέτιση είτε θετική είτε αρνητική παρατηρείται μεταξύ των S2, S4, των E2, S3 και των R2,R3 με το R4. Αυτό θα παίξει ρόλο παρακάτω όταν, γίνει η επιλογή των ανεξάρτητων μεταβλητών που θα χρησιμοποιηθούν στα μοντέλα μηχανικής μάθησης.

Το παρακάτω γράφημα παρουσιάζει μια διαφορετική οπτικοποίηση των συσχετίσεων μεταξύ των μετρήσεων των τεσσάρων αισθητήρων. Στα κελιά εκτός της διαγωνίου του συγκεκριμένου σχήματος απεικονίζονται τα διαγράμματα διασποράς για όλα τα ζευγάρια των μεταβλητών, δηλαδή σημειώνονται οι τιμές που λαμβάνουν ταυτόχρονα δύο αισθητήρες. Φανερή είναι και εδώ η μεγάλη εξάρτηση μεταξύ των αισθητήρων 2 και 4, αφού για κάθε τιμή στον οριζόντιο άξονα αντιστοιχεί μόνο μία τιμή στον κατακόρυφο.

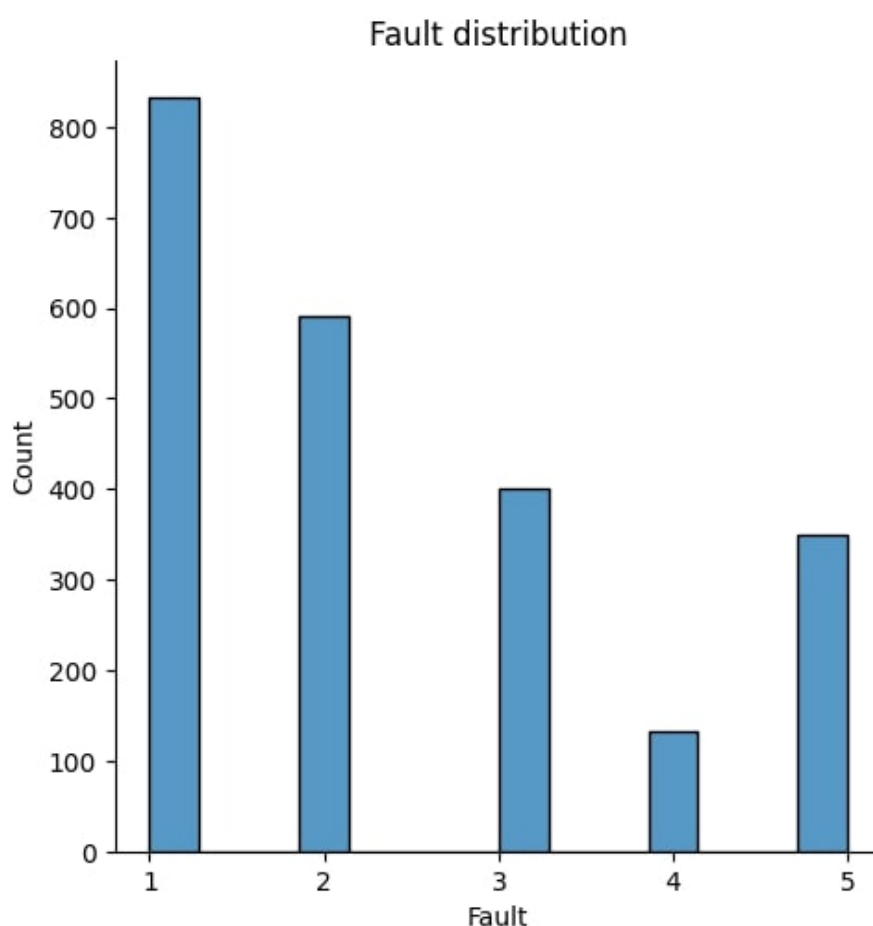


Σχήμα 3.19: Πολλαπλό scatterplot

Πέρα από την έντονη συσχέτιση των S2, S4, αξιοσημείωτο είναι ότι στα παραπάνω γραφήματα φαίνεται να σχηματίζονται clusters, ειδικά σε όσα περιέχουν το S1. Με πορτοκαλί βλέπουμε τα σημεία στα οποία υπάρχει σφάλμα οποιουδήποτε τύπου. Παρατηρούμε λοιπόν ότι οι παρατηρήσεις που αφορούν τα σφάλματα συγκεντρώνονται στις περισσότερες περιπτώσεις σε ένα από τα δύο clusters. Αυτό μπορεί να αποδειχθεί εξαιρετικά χρήσιμο για την προβλεπτική ικανότητα των μοντέλων, διότι σημαίνει ότι οι βλάβες έχουν κάποια σχέση με συγκεκριμένες ομάδες τιμών κάποιων μεταβλητών.

3.3 Διερεύνηση μεταβλητής-στόχου

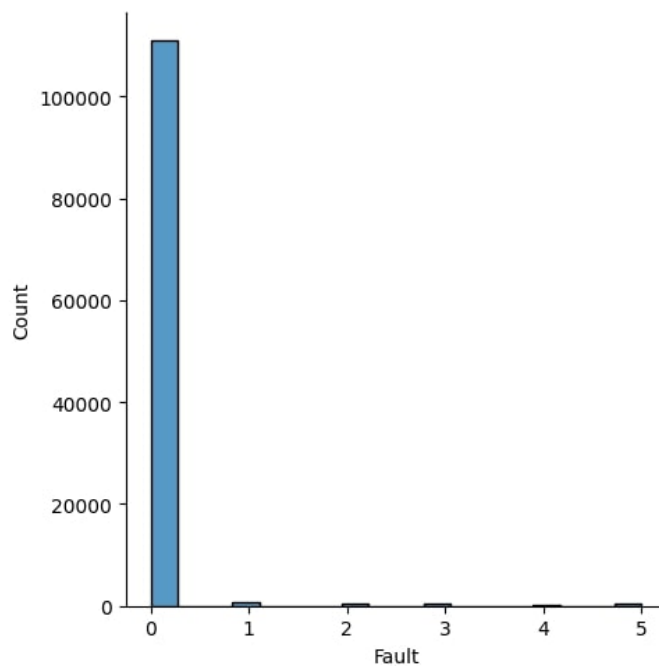
Σε ότι αφορά τις μεταβλητές-στόχους, η κατανομή τους είναι η ακόλουθη:



Σχήμα 3.20: Κατανομή βλαβών

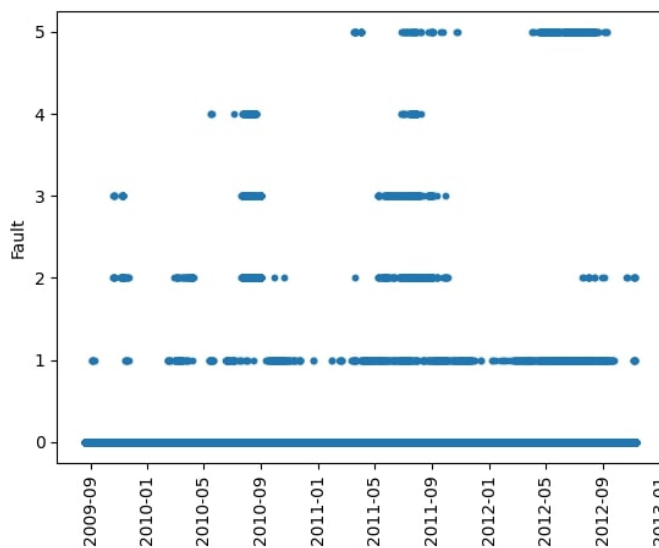
Παρατηρούμε ότι τα περισσότερα σφάλματα είναι τύπου 1, ενώ τα λιγότερα είναι τύπου 4. Το βασικό πρόβλημα που ήδη φαίνεται είναι ότι υπάρχουν μερικές εκατοντάδες περιπτώσεις σφαλμάτων για κάθε τύπο, τη στιγμή που τα δεδομένα μας είναι εκατοντάδες χιλιάδες.

Η ανισοροπία αυτή φαίνεται στο παρακάτω διάγραμμα:



Σχήμα 3.21: Ανισοροπία κλάσεων

Τέλος παρουσιάζουμε ένα ακόμη γράφημα που αφορά το πώς συμβαίνουν σφάλματα κατά τη διάρκεια της χρονικής περιόδου για την οποία έχουμε δεδομένα:



Σχήμα 3.22: Χρονική εξέλιξη σφαλμάτων

Η πρώτη παρατήρηση έχει να κάνει με το γεγονός ότι τα σφάλματα δεν συμβαίνουν ομοιόμορφα κατά τη διάρκεια του χρόνου. Τα σφάλματα τύπου 5 ξεκινούν από το δεύτερο μισό της υπό εξέταση χρονικής περιόδου. Αυτό έχει σημασία για την εκπαίδευση του μοντέλου. Θα πρέπει το οποιοδήποτε μοντέλο να εκπαιδευτεί μέσα σε μια χρονική περίοδο όπου υπάρχουν όλα τα είδη σφαλμάτων. Στην παρούσα εργασία αυτό θα επιτευχθεί καθώς, θα χρησιμοποιήσουμε πολλά διαφορετικά εργοστάσια για την εκπαίδευση του μοντέλου, οπότε θα υπάρχει εκπροσώπηση όλων των σφαλμάτων.

Η δεύτερη παρατήρηση είναι ότι εκτός από τα σφάλματα τύπου 1, τα υπόλοιπα φαίνεται ότι συμβαίνουν «μαζεμένα», δηλαδή μετά από μεγάλες χρονικές περιόδους «ηρεμίας» παρατηρούνται εξάρσεις όπου σφάλματα ίδιου τύπου συμβαίνουν συχνά για πολύ καιρό. Αυτό είναι λογικό διότι μπορεί να παίρνει καιρό μέχρι οι διαχειριστές του εργοστασίου να εξετάσουν, να αντιληφθούν και να επισκευάσουν τις ζημιές που αποτελούν την αιτία των σφαλμάτων. Αυτό είναι άλλη μία ένδειξη του πόσο αναγκαία μπορεί να αποδειχθεί η ανάπτυξη μεθόδων μηχανικής μάθησης για την έγκαιρη πρόγνωση τέτοιων καταστάσεων.

Κεφάλαιο 4

Προεπεξεργασία δεδομένων

4.1 Feature engineering

Όπως είδαμε προηγουμένως, το R4 είναι πολύ έντονα συσχετισμένο τόσο με το R2 όσο και με το R3. Ως «μεγάλη» θεωρούμε τη συσχέτιση άνω του 0.9 κατά απόλυτη τιμή. Επομένως αφαιρούμε το R4 από το σύνολο των μεταβλητών μας καθώς η πληροφορία που περιέχει είναι σε μεγάλο βαθμό κοινή με την πληροφορία που περιέχεται στα R2, R3. Ομοίως, καθώς τα S2,S4 έχουν -0.999 συσχέτιση, δεν χρειαζόμαστε και τα δύο κι έτσι συμπεριλαμβάνουμε μόνο το S2 στο σύνολο των εισόδων του μοντέλου. Επιπλέον, όπως έχουμε δει, ήδη έχει αφαιρεθεί το E1 από το σύνολο των μεταβλητών μας. Έτσι μένουμε με τα S1, S2, S3, R1, R2, R3,E2.

Από αυτές δεν συμπεριλαμβάνονται αυτούσιες οι μεταβλητές αναφοράς R1, R2, R3 επειδή παρουσιάζουν διακριτή συμπεριφορά με απόλυτα περιοδικά μοτίβα, αλλά διατηρούνται σαν πληροφορία μέσω των εξής επιπλέον μεταβλητών: R1·S1, R2·S2, R3·S3.

Η δημιουργία μιας νέας ανεξάρτητης μεταβλητής με τη λήψη του γινομένου δύο ανεξάρτητων μεταβλητών αποτυπώνει τις αλληλεπιδράσεις και τις μη γραμμικές σχέσεις, παρέχοντας πρόσθετες πληροφορίες στο μοντέλο μηχανικής μάθησης. Η προσέγγιση αυτή μπορεί να ενισχύσει την απόδοση του μοντέλου και να επιτρέψει μια πιο ολοκληρωμένη κατανόηση των υποκείμενων σχέσεων στα δεδομένα. Επομένως, αν λάβουμε υπόψιν το γεγονός ότι τα R είναι αναφορές ελέγχου για τα S, η σχέση που τα συνδέει παίζει σημαντικό ρόλο.

Έγινε η επιλογή να χρησιμοποιηθούν 7 ανεξάρτητες μεταβλητές παρόλο που θεωρητικά θα μπορούσαμε να κατασκευάσουμε και να χρησιμοποιήσουμε πολλές περισσότερες διότι είναι σημαντικό να αποφευχθεί η «κατάρρα της διαστατικότητας» (curse of dimensionality):

Η κατάρρα της διαστατικότητας αναφέρεται στα προβλήματα και τους περιορισμούς που προκύπτουν με δεδομένα υψηλής διάστασης. Καθώς αυξάνεται ο αριθμός των ανεξάρτητων μεταβλητών ή αλλιώς της διάστασης του χώρου των features, αναδύονται διάφορα ζητήματα που μπορούν να επηρεάσουν την απόδοση και την αποτελεσματικότητα των αλγορίθμων μηχανικής

μάθησης [21]. Τα προβλήματα αυτά είναι:

- **Αυξημένη αραιότητα δεδομένων:** Σε χώρους υψηλών διαστάσεων, τα διαθέσιμα δεδομένα καθίστανται αραιά. Καθώς αυξάνεται ο αριθμός των διαστάσεων, ο αριθμός των σημείων δεδομένων που απαιτούνται για τη διατήρηση ενός αντιπροσωπευτικού δείγματος αυξάνεται εκθετικά. Αυτό συμβαίνει διότι η κατανομή των σημείων στον χώρο των features, γίνεται αραιότερη και είναι δυσκολότερο για το μοντέλο να αναγνωρίσει στατιστικά στοιχεία, αποστάσεις και πυκνότητες στον πολυδιάστατο χώρο, τα οποία είναι θεμελιώδη για πολλούς αλγόριθμους μηχανικής μάθησης. Αυτή η αραιότητα καθιστά δύσκολη τη γενίκευση από τα διαθέσιμα δεδομένα και μπορεί να οδηγήσει σε υπερπροσαρμογή.
- **Αυξημένη υπολογιστική πολυπλοκότητα:** Με μεγαλύτερο αριθμό διαστάσεων, η υπολογιστική πολυπλοκότητα της επεξεργασίας και ανάλυσης των δεδομένων αυξάνεται δραματικά. Πολλοί αλγόριθμοι παρουσιάζουν εκθετική ή σχεδόν εκθετική αύξηση των υπολογιστικών απαιτήσεων καθώς αυξάνεται η διασταση. Αυτό μπορεί να καταστήσει υπολογιστικά δαπανηρή ή ανέφικτη την εκπαίδευση μοντέλων ή την εκτέλεση ορισμένων λειτουργιών σε δεδομένα υψηλών διαστάσεων.
- **Περιττά και άσχετα χαρακτηριστικά:** Καθώς αυξάνεται η διαστατικότητα, γίνεται πιο πιθανό ορισμένα από τα χαρακτηριστικά να είναι περιττά ή άσχετα με τη διαδικασία μάθησης. Τα περιττά χαρακτηριστικά παρέχουν περιττές πληροφορίες, ενώ τα άσχετα χαρακτηριστικά εισάγουν θόρυβο και μπορούν να παραπλανήσουν τον αλγόριθμο μάθησης. Αυτά τα χαρακτηριστικά μπορούν να εμποδίσουν την απόδοση του μοντέλου και να αυξήσουν τον κίνδυνο υπερπροσαρμογής.
- **Ζητήματα υπερπροσαρμογής και γενίκευσης:** Η κατάρτα της διαστατικότητας επιτείνει τον κίνδυνο υπερπροσαρμογής. Με υψηλότερο αριθμό διαστάσεων, τα μοντέλα μπορεί να προσαρμόζονται πολύ στενά στα δεδομένα εκπαίδευσης, χάνοντας την ικανότητά τους να γενικεύουν σε αθέατα δεδομένα.

Η κλιμάκωση των μεταβλητών (feature scaling) είναι επίσης ένα κρίσιμο βήμα προεπεξεργασίας στη μηχανική μάθηση που περιλαμβάνει το μετασχηματισμό και την κανονικοποίηση των χαρακτηριστικών ενός συνόλου δεδομένων. Στόχος της είναι να διασφαλίσει ότι όλα τα χαρακτηριστικά έχουν συγκρίσιμη κλίμακα και κατανομή, γεγονός που συμβάλλει στη βελτίωση της απόδοσης και της σταθερότητας των μοντέλων μηχανικής μάθησης.

Η κλιμάκωση μεταβλητών είναι απαραίτητη επειδή πολλοί αλγόριθμοι μηχανικής μάθησης είναι ευαίσθητοι στη σχετική κλίμακα των χαρακτηριστικών. Όταν τα χαρακτηριστικά έχουν διαφορετικές κλίμακες, οι αλγόριθμοι που βασίζονται σε συγκρίσεις απόστασης ή μεγέθους μπορεί να μεροληπτούν προς τα χαρακτηριστικά με μεγαλύτερες κλίμακες. Αυτό μπορεί να οδηγήσει

σε ανακριβείς προβλέψεις ή παρερμηνείες της σημασίας των διαφορετικών χαρακτηριστικών στο μοντέλο. Οι πιο γνωστές τεχνικές κλιμάκωσης είναι οι εξής:

- **Τυποποίηση:** Η τυποποίηση μετασχηματίζει τις μεταβλητές ώστε να έχουν μέσο όρο 0 και τυπική απόκλιση 1. Πραγματοποιείται με την αφαίρεση του μέσου όρου κάθε μεταβλητής και τη διαίρεση με την τυπική του απόκλιση:

$$z = \frac{x - \sigma}{\mu}$$

Οι μετασχηματισμένες μεταβλητές έχουν κατανομή Gauss. Η τυποποίηση διατηρεί το σχήμα της κατανομής και είναι χρήσιμη όταν τα δεδομένα δεν έχουν γνωστό ή σταθερό εύρος.

- **Κανονικοποίηση:** Η κανονικοποίηση κλιμακώνει τις μεταβλητές σε ένα συγκεκριμένο εύρος, συνήθως μεταξύ 0 και 1. Επιτυγχάνεται αφαιρώντας την ελάχιστη τιμή κάθε μεταβλητής και διαιρώντας με το εύρος:

$$z = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Η κανονικοποίηση εξασφαλίζει ότι όλες οι μεταβλητές έχουν την ίδια κλίμακα, καθιστώντας τις άμεσα συγκρίσιμες. Αυτή η τεχνική είναι κατάλληλη όταν η κατανομή των δεδομένων δεν είναι απαραίτητα κανονική και όταν υπάρχει ένα γνωστό ή επιθυμητό εύρος για τα χαρακτηριστικά.

Στην δική μας περίπτωση η κατάλληλότερη τεχνική κλιμάκωσης είναι η κανονικοποίηση, καθώς όπως είδαμε νωρίτερα δεν έχουν όλες οι μεταβλητές την κανονική κατανομή. Επιπλέον στα δεδομένα μας υπάρχουν όπως είδαμε κάποια outliers. Με την κανονικοποίηση διατηρείται το σχετικό μέγεθος αυτών των outliers που μπορεί να βοηθήσουν στην προβλεπτική ικανότητα του μοντέλου καθώς εκφράζουν ανωμαλίες στη χρονοσειρά των μετρήσεων.

4.2 Στάθμιση κλάσεων

Το πρόβλημα των μη σταθμισμένων κλάσεων είναι αρκετά σύνηθες στη μηχανική μάθηση. Αναφέρεται σε μια κατάσταση όπου μια κλάση, γνωστή ως πλειοψηφική κλάση, έχει σημαντικά μεγαλύτερο αριθμό περιπτώσεων στο σύνολο των διαθέσιμων δεδομένων σε σύγκριση με την άλλη κλάση ή τις άλλες κλάσεις, γνωστές ως μειοψηφικές κλάσεις.

Συνέπειες

Οι μη σταθμισμένες κλάσεις μπορούν να δημιουργήσουν διάφορα προβλήματα σε εργασίες

μηχανικής μάθησης, όπως η μεροληπτική απόδοση του μοντέλου, οι ανακριβείς προβλέψεις και η δυσκολία εκμάθησης της μειοψηφικής κλάσης. Πιο συγκεκριμένα:

- **Μεροληπτική απόδοση του μοντέλου:** Οι περισσότεροι αλγόριθμοι μηχανικής μάθησης είναι σχεδιασμένοι για τη βελτιστοποίηση της συνολικής ακρίβειας. Με την παρουσία ανισορροπίας κλάσεων, το μοντέλο μπορεί να είναι προκατειλημμένο προς την πλειοψηφούσα κλάση. Ως αποτέλεσμα, μπορεί να επιτύχει υψηλή ακρίβεια προβλέποντας απλώς την πλειοψηφική κλάση για τις περισσότερες περιπτώσεις, ενώ έχει κακή απόδοση για τη μειοψηφική κλάση. Αυτή η μεροληψία μπορεί να οδηγήσει σε παραπλανητικές μετρήσεις απόδοσης και σε ανακριβή κατανόηση της αποτελεσματικότητας του μοντέλου.
- **Κακή ευαισθησία στην τάξη μειονότητας:** Σε μη ισορροπημένα σύνολα δεδομένων, η κλάση μειονότητας παρουσιάζει συχνά μεγαλύτερο ενδιαφέρον, καθώς αντιπροσωπεύει την κλάση-στόχο ή την κλάση ενδιαφέροντος. Ωστόσο, λόγω του περιορισμένου αριθμού περιπτώσεων από την κλάση μειονότητας, το μοντέλο μπορεί να δυσκολευτεί να μάθει τα διακριτικά χαρακτηριστικά της. Κατά συνέπεια, η ευαισθησία ή η ανάκληση του μοντέλου για την κλάση της μειονότητας μπορεί να είναι χαμηλή, οδηγώντας σε αυξημένο αριθμό ψευδώς αρνητικών αποτελεσμάτων.
- **Μειωμένη ικανότητα γενίκευσης:** Η ανισορροπία τάξεων μπορεί να επηρεάσει αρνητικά την ικανότητα γενίκευσης των μοντέλων μηχανικής μάθησης. Δεδομένου ότι η πλειοψηφούσα κλάση κυριαρχεί στα δεδομένα εκπαίδευσης, το μοντέλο μπορεί να γίνει υπερβολικά προκατειλημμένο προς αυτήν, με αποτέλεσμα να μειώνεται η ικανότητα γενίκευσης σε ανθέατες περιπτώσεις ή η καλή απόδοση σε νέα δεδομένα. Το ζήτημα αυτό είναι ιδιαίτερα έντονο όταν η μειοψηφική τάξη παρουσιάζει σημαντικές διαφορές από την πλειοψηφική τάξη, καθιστώντας δύσκολη την εκμάθηση των ιδιαίτερων προτύπων τους από το μοντέλο.
- **Μεροληψία συλλογής δεδομένων:** Η ανισορροπία τάξεων μπορεί να εισάγει μεροληψίες στη συλλογή και την ανάλυση δεδομένων. Για παράδειγμα, εάν ένα ιατρικό διαγνωστικό μοντέλο εκπαιδευτεί σε ανισόρροπα δεδομένα, μπορεί να διαγνώσει δυσανάλογα λανθασμένα την τάξη της μειονότητας. Αυτό μπορεί να έχει σημαντικές συνέπειες στον πραγματικό κόσμο, όπως ο λανθασμένος εντοπισμός σπάνιων ασθενειών ή ανακριβείς προβλέψεις σε κρίσιμες καταστάσεις.
- **Μη κατατοπιστικές μετρικές αξιολόγησης:** Οι παραδοσιακές μετρικές αξιολόγησης, όπως η ακρίβεια, μπορεί να είναι παραπλανητικές παρουσία ανισορροπίας τάξεων. Για παράδειγμα, ένα μοντέλο που προβλέπει πάντα την πλειοψηφική τάξη μπορεί να επιτύχει υψηλή ακρίβεια λόγω της ανισορροπίας της τάξης, αλλά παρέχει μικρή αξία στην καταγραφή της μειονοτικής τάξης. Ως εκ τούτου, εναλλακτικές μετρικές αξιολόγησης, όπως η ακρίβεια, η ανάκληση, το F1-score και η περιοχή κάτω από την καμπύλη ROC (receiver operating characteristic curve) είναι συχνά πιο κατατοπιστικές όταν έχουμε να κάνουμε με ανισόρροπα σύνολα δεδομένων.

Αντιμετώπιση

Ένας κλασικός τρόπος αντιμετώπισης του προβλήματος της ανισορροπίας κλάσεων είναι η αναδειγματοληψία. Αυτή μπορεί να γίνει με δύο τρόπους:

- Υπερδειγματοληψία της μειονοτικής τάξης: Αυτό περιλαμβάνει την αύξηση του αριθμού των περιπτώσεων στην τάξη μειονότητας. Τεχνικές όπως η τυχαία υπερδειγματοληψία, η SMOTE (Synthetic Minority Over-sampling Technique) ή η ADASYN (Adaptive Synthetic Sampling) μπορούν να χρησιμοποιηθούν για τη δημιουργία συνθετικών δειγμάτων ή την αντιγραφή υφιστάμενων δειγμάτων της κλάσης μειονότητας.
- Υποδειγματοληψία της πλειοψηφικής τάξης: Αυτό περιλαμβάνει τη μείωση του αριθμού των περιπτώσεων της πλειοψηφικής κλάσης. Τεχνικές υποδειγματοληψίας όπως η τυχαία υποδειγματοληψία ή τα Cluster Centroids μπορούν να βοηθήσουν στη μείωση του μεγέθους του συνόλου δεδομένων της πλειοψηφικής κλάσης ώστε να ταιριάζει με το μέγεθος της μειονοτικής κλάσης.

Οι τεχνικές επαναδειγματοληψίας δεδομένων αποσκοπούν στη δημιουργία ενός πιο ισορροπημένου συνόλου δεδομένων εκπαίδευσης, το οποίο μπορεί να βοηθήσει τον ταξινομητή να μάθει τα πρότυπα και να κάνει καλύτερες προβλέψεις για την τάξη της μειονότητας. Ωστόσο, είναι σημαντικό να σημειωθεί ότι αυτές οι τεχνικές μπορεί επίσης να εισάγουν κάποιες προκαταλήψεις ή απώλεια πληροφοριών, οπότε απαιτείται προσεκτική εξέταση και πειραματισμός.

Άλλες τεχνικές αντιμετώπισης είναι οι εξής:

- Βαρύτητα κλάσης: Η στάθμιση των κλάσεων είναι μια άλλη προσέγγιση για την αντιμετώπιση της ανισορροπίας των κλάσεων. Αναθέτοντας υψηλότερα βάρη στην κλάση της μειονότητας και χαμηλότερα βάρη στην κλάση της πλειοψηφίας κατά τη διάρκεια της εκπαίδευσης του μοντέλου, ο ταξινομητής εστιάζει περισσότερο στην ορθή πρόβλεψη των περιπτώσεων της κλάσης της μειονότητας. Οι περισσότεροι αλγόριθμοι μηχανικής μάθησης παρέχουν μια επιλογή για τον καθορισμό των βαρών κλάσης και τα βάρη αυτά μπορούν να προσαρμοστούν με βάση την κατανομή της κλάσης.
- Προσαρμογή κατώφλιου (Threshold adjustment): Στη δυαδική ταξινόμηση, το κατώφλι απόφασης καθορίζει το σημείο στο οποίο οι προβλεπόμενες πιθανότητες ταξινομούνται στη μία ή την άλλη κλάση. Από προεπιλογή, το κατώφλι έχει οριστεί στο 0.5, αλλά η προσαρμογή του κατώφλιου μπορεί να βοηθήσει στην επίτευξη καλύτερης ισορροπίας μεταξύ ακρίβειας και ανάκλησης για την κλάση μειονότητας.

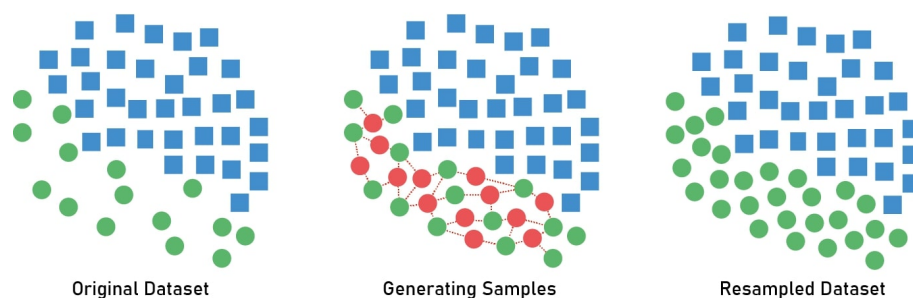
Στην παρούσα εργασία η τεχνική αντιμετώπισης της ανισορροπίας κλάσεων που επιλέχθηκε είναι SMOTE (Synthetic Minority Over-sampling Technique). Η SMOTE είναι μια δημοφιλής τεχνική που χρησιμοποιείται για την αντιμετώπιση της ανισορροπίας των κλάσεων στη μηχανική μάθηση. Είναι ειδικά σχεδιασμένη για την υπερδειγματοληψία της μειονοτικής κλάσης με τη δημιουργία συνθετικών παραδειγμάτων αντί για την απλή αντιγραφή υφιστάμενων περιπτώσεων. Η SMOTE λειτουργεί με τη δημιουργία συνθετικών δειγμάτων κατά μήκος των τμημάτων γραμμής που συνδέουν ζεύγη περιπτώσεων της κλάσης μειονότητας. Για την επίτευξή του ακολουθείται ο παρακάτω αλγόριθμος:

- Βρες τη μειοψηφική κλάση.
- Επίλεξε τυχαία μια περίπτωση από την κλάση μειονότητας ως σημείο εκκίνησης.
- Βρες τους k πλησιέστερους γείτονες της επιλεγμένης περίπτωσης από την μειοψηφική κλάση.
- Επίλεξε τυχαία έναν από τους k πλησιέστερους γείτονές του.
- Δημιούργησε ένα συνθετικό δείγμα με παρεμβολή μεταξύ της επιλεγμένης περίπτωσης και του γείτονά της. Αυτό γίνεται με τον υπολογισμό της διαφοράς μεταξύ των τιμών features του επιλεγμένου παραδείγματος και του γείτονά του, τον πολλαπλασιασμό της με έναν τυχαίο αριθμό μεταξύ 0 και 1. Για παράδειγμα, αν έχουμε δύο περιπτώσεις A και B και θέλουμε να δημιουργήσουμε ένα συνθετικό δείγμα, υπολογίζουμε τη διαφορά μεταξύ των τιμών features των A και B , την πολλαπλασιάζουμε με έναν τυχαίο αριθμό r και προσθέτουμε το αποτέλεσμα στο A . Ο τύπος είναι:

$$synthetic_sample = A + r \cdot (B - A)$$

- Επαναλάβετε τα βήματα 2 έως 5 για έναν επιθυμητό αριθμό επαναλήψεων ή έως ότου επιτευχθεί το επιθυμητό επίπεδο ισορροπίας των κλάσεων.

Synthetic Minority Oversampling Technique



Σχήμα 4.1: Απεικόνιση της τεχνικής SMOTE

Η τεχνική SMOTE παράγει συνθετικά δείγματα που βρίσκονται στο χώρο των features μεταξύ των υφιστάμενων περιπτώσεων της κλάσης μειοψηφίας. Με τη δημιουργία αυτών των συνθετικών παραδειγμάτων, η τεχνική SMOTE αυξάνει αποτελεσματικά τον αριθμό των περιπτώσεων της μειοψηφικής κλάσης, αντιμετωπίζοντας το πρόβλημα της ανισορροπίας των κλάσεων. Αυτό επιτρέπει στον ταξινομητή να μαθαίνει από ένα πιο ισορροπημένο σύνολο δεδομένων και να κάνει καλύτερες προβλέψεις για την κλάση μειοψηφίας που στην περίπτωση μας είναι και το ζητούμενο δηλαδή να εντοπίζονται έγκαιρα οι βλάβες.

Φυσικά η υπερδειγματοληψία αφορά μόνο τα δεδομένα εκπαίδευσης ώστε το μοντέλο να μάθει όσο το δυνατόν καλύτερα και τις δύο κλάσεις. Η αξιολόγηση του μοντέλου θα γίνει σε άγνωστα και ανισόροπα δεδομένα, διότι αυτό είναι το ρεαλιστικό σενάριο.

Κεφάλαιο 5

Ανάπτυξη προβλεπτικού μοντέλου

5.1 Αλγόριθμοι

Αρχικά, θα γίνει μια αναλυτική επισκόπηση όλων των αλγορίθμων που χρησιμοποιήθηκαν στην παρούσα εργασία. Από πιο κλασικές μεθόδους όπως η λογιστική παλινδρόμηση και η KNN, έως πιο εξελιγμένες μεθόδους όπως τα δίκτυα LSTM, θα εξετάσουμε σημαντικές πτυχές των αλγορίθμων αυτών γύρω από την λειτουργία τους, τα δυνατά αλλά και τα αδύναμα σημεία τους.

Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση (logistic regression) είναι μια στατιστική τεχνική που χρησιμοποιείται ευρέως για προβλήματα δυαδικής ταξινόμησης σε διάφορους τομείς, όπως η ιατρική, η οικονομία, το μάρκετινγκ και οι κοινωνικές επιστήμες. Είναι μέλος της οικογένειας των γενικευμένων γραμμικών μοντέλων και παρέχει μια πρακτική και ερμηνεύσιμη προσέγγιση για την πρόβλεψη της πιθανότητας εμφάνισης ενός γεγονότος. Η λογιστική παλινδρόμηση επεκτείνει τις αρχές της γραμμικής παλινδρόμησης στον τομέα της δυαδικής ταξινόμησης εφαρμόζοντας έναν μετασχηματισμό γνωστό ως λογιστική συνάρτηση (που ονομάζεται επίσης σιγμοειδής συνάρτηση). Η λογιστική συνάρτηση αντιστοιχίζει την έξοδο της γραμμικής παλινδρόμησης σε μια τιμή μεταξύ 0 και 1, που αντιπροσωπεύει την πιθανότητα μια παρατήρηση να ανήκει στη θετική κλάση. Η μαθηματική διατύπωση είναι η εξής:

$$p = 1/(1 + e^{-z})$$

όπου p είναι η πιθανότητα, z είναι ο γραμμικός συνδυασμός των προβλεπτικών παραγόντων και των αντίστοιχων συντελεστών τους. Η λογιστική παλινδρόμηση προϋποθέτει αρκετές βασικές παραδοχές για την παροχή αξιόπιστων αποτελεσμάτων:

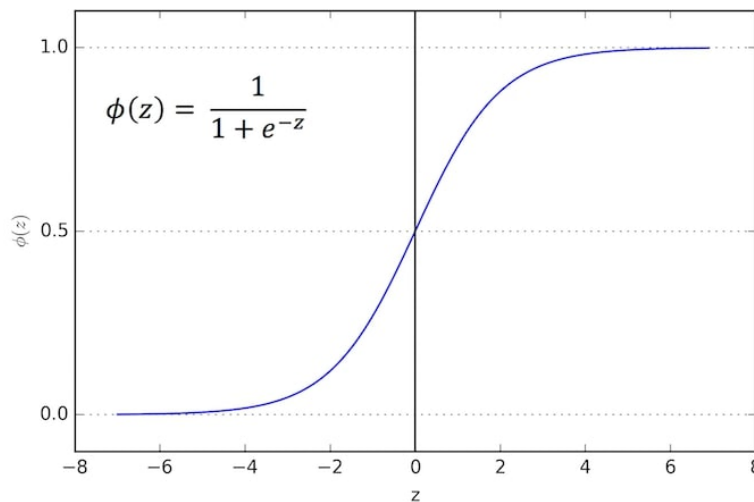
- Γραμμικότητα: Η σχέση μεταξύ των ανεξάρτητων μεταβλητών και των λογαριθμικών αποδόσεων της εξαρτημένης μεταβλητής είναι γραμμική.
- Ανεξαρτησία των σφαλμάτων: Οι παρατηρήσεις είναι ανεξάρτητες μεταξύ τους.

- Απουσία πολυσυγγραμμικότητας: Οι μεταβλητές πρόβλεψης δεν συσχετίζονται σε μεγάλο βαθμό μεταξύ τους.
- Μεγάλο μέγεθος δείγματος: Υπάρχουν επαρκή δεδομένα για την ακριβή εκτίμηση των παραμέτρων του μοντέλου.
- Δεν υπάρχουν ακραίες τιμές: Οι ακραίες τιμές επηρεάζουν τα αποτελέσματα

Το μοντέλο λογιστικής παλινδρόμησης εκπαιδεύεται με μια διαδικασία που ονομάζεται εκτίμηση μέγιστης πιθανοφάνειας, όπου ο στόχος είναι να βρεθούν οι βέλτιστες τιμές για τους συντελεστές που μεγιστοποιούν την πιθανότητα παρατήρησης των δεδομένων. Για την εύρεση αυτών των βέλτιστων τιμών παραμέτρων μπορούν να χρησιμοποιηθούν διάφοροι αλγόριθμοι βελτιστοποίησης, όπως gradient descent ή Newton-Raphson.

Η λογιστική παλινδρόμηση, όπως και κάθε άλλο μοντέλο μηχανικής μάθησης, είναι επιρρεπής σε υπερπροσαρμογή ή υποπροσαρμογή. Η υπερπροσαρμογή συμβαίνει όταν το μοντέλο συλλαμβάνει θόρυβο ή άσχετα μοτίβα από τα δεδομένα εκπαίδευσης, με αποτέλεσμα την κακή γενίκευση σε άγνωστα δεδομένα. Τεχνικές κανονικοποίησης, μπορούν να χρησιμοποιηθούν για τον μετριασμό της υπερπροσαρμογής. Από την άλλη πλευρά, η υποπροσαρμογή συμβαίνει όταν το μοντέλο είναι πολύ απλοϊκό για να συλλάβει τα υποκείμενα πρότυπα. Ένας καλός σχεδιασμός των ανεξαρτητων μεταβλητών και η προσαρμογή της πολυπλοκότητας του μοντέλου μπορούν να βοηθήσουν στην αντιμετώπιση της υποπροσαρμογής.

Η λογιστική παλινδρόμηση μπορεί να χειριστεί τόσο κατηγορικά όσο και συνεχή προβλήματα. Οι κατηγορικές μεταβλητές κωδικοποιούνται συνήθως με τεχνικές όπως το 1-hot, όπου κάθε κατηγορία αντιπροσωπεύεται από μια δυαδική στήλη. Οι συνεχείς προγνωστικοί παράγοντες χρησιμοποιούνται απευθείας στο μοντέλο μετά από κατάλληλη κλιμάκωση.



Σχήμα 5.1: logistic function

Η υπερπαράμετρος που θα ρυθμιστεί είναι η σταθερά C . Η παράμετρος C ελέγχει την ποσότητα της κανονικοποίησης που εφαρμόζεται στο μοντέλο λογιστικής παλινδρόμησης. Μια μικρότερη τιμή του C υποδηλώνει μια ισχυρότερη κανονικοποίηση, που σημαίνει ότι το μοντέλο θα προσπαθήσει να διατηρήσει τους συντελεστές (βάρη) πιο κοντά στο μηδέν. Από την άλλη πλευρά, μια μεγαλύτερη τιμή του C μειώνει την ισχύ της κανονικοποίησης, επιτρέποντας στο μοντέλο να προσαρμόζεται περισσότερο στα δεδομένα εκπαίδευσης. Οι τιμές που θα δοκιμαστούν είναι 0.001, 0.01, 0.1, 1, 10, 100, 1000, καθώς προσπαθούμε να συμπεριλάβουμε διαφορετικές τάξεις μεγέθους.

Naive Bayes

Ο ταξινομητής Naive Bayes είναι ένας δημοφιλής και ισχυρός αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για εργασίες ταξινόμησης. Είναι ιδιαίτερα αποτελεσματικός σε προβλήματα ταξινόμησης κειμένου, όπως το φιλτράρισμα ανεπιθύμητων μηνυμάτων, η ανάλυση συναισθήματος και η κατηγοριοποίηση εγγράφων. Παρά την απλότητά του, ο ταξινομητής Naive Bayes έχει επιδείξει σταθερά ισχυρές επιδόσεις σε διάφορους τομείς.

Ο ταξινομητής Naive Bayes υποθέτει ότι οι ανεξάρτητες μεταβλητές είναι κατά συνήθη ανεξάρτητες μεταξύ τους δεδομένης της κλάσης που ανήκουν. Αυτή η υπόθεση απλοποιεί τον υπολογισμό του όρου πιθανότητας στο θεώρημα του Bayes. Παρόλο που αυτή η υπόθεση σπάνια ισχύει στην πράξη, ο ταξινομητής Naive Bayes εξακολουθεί να αποδίδει αρκετά καλά σε πολλά σενάρια του πραγματικού κόσμου.

Ο ταξινομητής Naive Bayes εκπαιδεύεται με την εκτίμηση των παραμέτρων από το σύνολο δεδομένων εκπαίδευσης. Τα βασικά βήματα είναι τα εξής:

- Υπολογισμός των εκ των προτέρων πιθανοτήτων κάθε κλάσης με βάση την κατανομή των κλάσεων στα δεδομένα εκπαίδευσης.
- Εκτίμηση πιθανοφάνειας κάθε ανεξάρτητης μεταβλητής με δεδομένη κάθε ετικέτα κλάσης.
- Υπολογισμός των εκ των υστέρων πιθανοτήτων για κάθε κλάση χρησιμοποιώντας το θεώρημα του Bayes.
- Ταξινόμηση με βάση την υψηλότερη εκ των υστέρων πιθανότητα ως την προβλεπόμενη κλάση για ένα δεδομένο σημείο δεδομένων.

Σε αυτή την εργασία θα χρησιμοποιήσουμε το Naive Bayes σαν ένα μοντέλο βάσης (baseline model) έτσι ώστε να κρίνουμε αν το μοντέλο μας τα πηγαίνει καλύτερα, θα είναι δηλαδή ένας πήχης για εμάς.

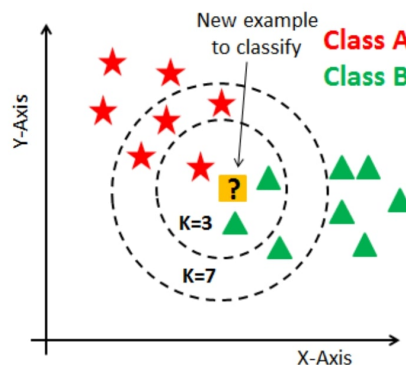
K Nearest Neighbours

Ο ταξινομητής K-Nearest Neighbors (KNN) είναι ένας δημοφιλής και ευέλικτος αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται τόσο για εργασίες ταξινόμησης όσο και για εργασίες παλινδρόμησης. Πρόκειται για μια μη παραμετρική μέθοδο που κάνει προβλέψεις με βάση την ομοιότητα μεταξύ του σημείου δεδομένων εισόδου και των k πλησιέστερων γειτόνων του. Ο KNN είναι γνωστός για την απλότητα και τη διαισθητική του έννοια, γεγονός που τον καθιστά κατάλληλη επιλογή για διάφορες εφαρμογές.

Ο ταξινομητής KNN είναι ένας τύπος αλγορίθμου μάθησης που βασίζεται στη μνήμη. Αντί να μαθαίνει ένα συγκεκριμένο μοντέλο από τα δεδομένα εκπαίδευσης, αποθηκεύει ολόκληρο το σύνολο των δεδομένων εκπαίδευσης στη μνήμη και το χρησιμοποιεί για να κάνει προβλέψεις για νέα, άγνωστα σημεία δεδομένων. Όταν του δίνεται μια νέα είσοδος, ο KNN αναζητά τους k πλησιέστερους γείτονες στο σύνολο δεδομένων εκπαίδευσης με βάση μια επιλεγμένη μετρική απόστασης. Η ετικέτα κλάσης ή η τιμή του νέου σημείου δεδομένων καθορίζεται κατά πλειοψηφία (στην ταξινόμηση) ή με μέσο όρο (στην παλινδρόμηση) μεταξύ των k πλησιέστερων γειτόνων του.

Για τη μέτρηση της ομοιότητας ή της απόστασης μεταξύ σημείων δεδομένων, ο KNN βασίζεται σε μετρικές απόστασης. Οι μετρικές απόστασης που χρησιμοποιούνται συνήθως περιλαμβάνουν την Ευκλείδεια απόσταση, την απόσταση Manhattan, την απόσταση Minkowski και την απόσταση Hamming. Η επιλογή της μετρικής απόστασης εξαρτάται από τη φύση των δεδομένων και του εκάστοτε προβλήματος.

Η τιμή του k , που αντιπροσωπεύει τον αριθμό των γειτόνων που λαμβάνονται υπόψη, επηρεάζει σημαντικά την απόδοση του ταξινομητή KNN. Μια μικρότερη τιμή του k (π.χ. $k = 1$) οδηγεί σε πιο ευέλικτα όρια, αλλά μπορεί να είναι ευαίσθητη στο θόρυβο. Από την άλλη πλευρά, μια μεγαλύτερη τιμή του k (π.χ. $k = 10$) παράγει πιο ομαλά όρια αλλά μπορεί να παραβλέψει τοπικά μοτίβα. Η βέλτιστη τιμή του k καθορίζεται συχνά μέσω πειραματισμού και διασταυρούμενης επικύρωσης.



Σχήμα 5.2: KNN

Long Short-Term Memory (LSTM)

Το νευρωνικό δίκτυο είναι ένα δίκτυο από απλούς υπολογιστικούς κόμβους (νευρώνες, νευρώνια), διασυνδεδεμένους μεταξύ τους. Είναι εμπνευσμένο από το Κεντρικό Νευρικό Σύστημα (ΚΝΣ), το οποίο προσπαθεί να προσομοιώσει.

Οι νευρώνες είναι τα δομικά στοιχεία του δικτύου. Κάθε τέτοιος κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων από διαφορετικές πηγές (είτε από άλλους νευρώνες, είτε από το περιβάλλον), επιτελεί έναν υπολογισμό με βάση αυτές τις εισόδους και παράγει μία έξοδο. Η εν λόγω έξοδος είτε κατευθύνεται στο περιβάλλον, είτε τροφοδοτείται ως είσοδος σε άλλους νευρώνες του δικτύου. Υπάρχουν τρεις τύποι νευρώνων: οι νευρώνες εισόδου, οι νευρώνες εξόδου και οι υπολογιστικοί νευρώνες ή κρυμμένοι νευρώνες. Οι νευρώνες εισόδου δεν επιτελούν κανέναν υπολογισμό, μεσολαβούν απλώς ανάμεσα στις περιβαλλοντικές εισόδους του δικτύου και στους υπολογιστικούς νευρώνες. Οι νευρώνες εξόδου διαχετεύουν στο περιβάλλον τις τελικές αριθμητικές εξόδους του δικτύου. Οι υπολογιστικοί νευρώνες πολλαπλασιάζουν κάθε είσοδό τους με το αντίστοιχο συναπτικό βάρος και υπολογίζουν το ολικό άθροισμα των γινομένων. Το άθροισμα αυτό τροφοδοτείται ως όρισμα στη συνάρτηση ενεργοποίησης, την οποία υλοποιεί εσωτερικά κάθε κόμβος. Η τιμή που λαμβάνει η συνάρτηση για το εν λόγω όρισμα είναι και η έξοδος του νευρώνα για τις τρέχουσες εισόδους και βάρη.

Τα LSTM είναι μορφή τεχνητών ανατροφοδοτούμενων νευρωνικών δικτύων που χρησιμοποιούνται στον τομέα της βαθιάς μάθησης. Σε αντίθεση με τα συνηθέστερα νευρωνικά δίκτυα πρόσθιας τροφοδότησης, τα ανατροφοδοτούμενα νευρωνικά δίκτυα διαθέτουν συνδέσμους ανατροφοδότησης που τους επιτρέπουν να επεξεργάζονται δεδομένα ως σειρές (συνήθως χρονοσειρές όπως δεδομένα ήχου ή βίντεο). Τα LSTM χρησιμοποιούνται σε εφαρμογές όπως την αυτόματη ανάγνωση χειρογράφων, την αναγνώριση ομιλίας και την ανίχνευση ανωμαλιών σε δικτυακές επικοινωνίες.

Η αρχιτεκτονική των LSTM αναπτύχθηκε ως λύση στο πρόβλημα εξαφάνισης κλίσεων που σημειώνεται στα παραδοσιακά ανατροφοδοτούμενα νευρωνικά. Η βασική της μονάδα απαρτίζεται από ένα κελί και πύλες εισόδου, εξόδου και λήθης. Το κελί κράταει πληροφορίες από προηγούμενες θέσεις χρόνου ενώ οι τρεις πύλες ρυθμίζουν την ροή πληροφορίας εντός και εκτός του κελιού. Έτσι, το δίκτυο μπορεί να μοντελοποιήσει τις σχέσεις χρονοσειρών με καθυστερήσεις άγνωστης διάρκειας μεταξύ σημαντικών γεγονότων. Αυτή η σχετικά χαμηλή ευαισθησία στην χρονική απόσταση πληροφοριών αποτελεί και το βασικό πλεονέκτημα των LSTM σε σχέση με τα παραδοσιακά ανατροφοδοτούμενα, τα κρυφά Μαρκοβιανά μοντέλα και άλλες μεθόδους μάθησης σειρών.

Οι παράμετροι που μπορούν να ρυθμιστούν είναι:

- Ρυθμός εκμάθησης: Αυτή η υπερπαράμετρος ελέγχει το μέγεθος του βήματος ή το ρυθμό με τον οποίο το μοντέλο μαθαίνει από τις ενημερώσεις κλίσης κατά τη διάρκεια της εκπαίδευσης. Ένας πολύ υψηλός ρυθμός μάθησης μπορεί να προκαλέσει απόκλιση του μοντέλου, ενώ ένας πολύ χαμηλός ρυθμός μάθησης μπορεί να οδηγήσει σε αργή σύγκλιση. Οι συνήθειες στρατηγικές περιλαμβάνουν την αναζήτηση σε πλέγμα ή τη χρήση προσαρμοστικών μεθόδων ρυθμού μάθησης όπως η Adam ή η RMSprop.
- Αριθμός κρυφών επιπέδων και μονάδων: Η αρχιτεκτονική ενός νευρωνικού δικτύου, συμπεριλαμβανομένου του αριθμού των κρυφών στρωμάτων και του αριθμού των μονάδων σε κάθε στρώμα, είναι μια άλλη σημαντική υπερπαράμετρος. Η προσθήκη περισσότερων στρωμάτων ή μονάδων μπορεί να αυξήσει την ικανότητα του μοντέλου να μαθαίνει σύνθετα μοτίβα, αλλά πάρα πολλά στρώματα ή μονάδες μπορεί να οδηγήσουν σε υπερπροσαρμογή. Συχνά είναι απαραίτητη η εκτέλεση μιας αναζήτησης για την εύρεση του βέλτιστου βάθους και πλάτους του δικτύου.
- Μέγεθος παρτίδας (Batch size): Το μέγεθος παρτίδας καθορίζει τον αριθμό των δειγμάτων που διαδίδονται μέσω του δικτύου πριν από την εκτέλεση μιας ενημέρωσης βάρους. Μεγαλύτερα μεγέθη παρτίδας μπορούν να παρέχουν μια πιο σταθερή εκτίμηση κλίσης, ενώ μικρότερα μεγέθη παρτίδας μπορούν να εισάγουν περισσότερο θόρυβο. Η επιλογή ενός κατάλληλου μεγέθους παρτίδας εξαρτάται από τους περιορισμούς μνήμης, την υπολογιστική απόδοση και την απόδοση γενίκευσης.

5.2 Μετρικές αξιολόγησης μοντέλου

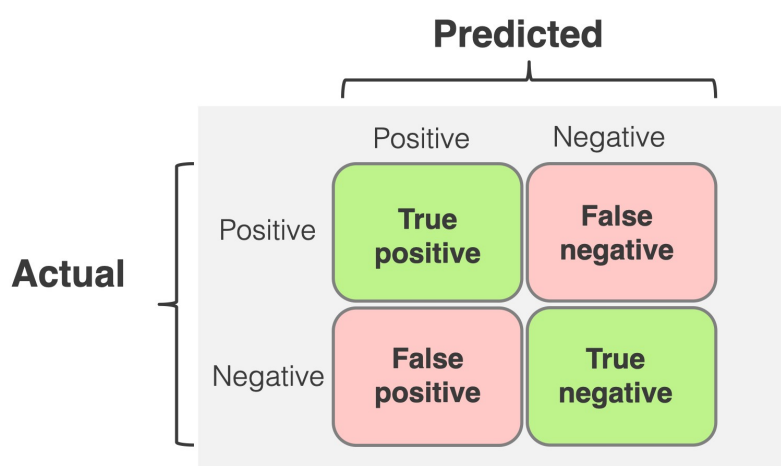
Η αξιολόγηση του μοντέλου είναι ένα κρίσιμο βήμα στη μηχανική μάθηση. Περιλαμβάνει την αξιολόγηση της απόδοσης και της ικανότητας γενίκευσης ενός εκπαιδευμένου μοντέλου σε αθέατα δεδομένα. Η κατάλληλη αξιολόγηση μοντέλου διασφαλίζει ότι το επιλεγμένο μοντέλο είναι αποτελεσματικό και αξιόπιστο για την πραγματοποίηση ακριβών προβλέψεων ή την παροχή πολύτιμων πληροφοριών.

Η αξιολόγηση μοντέλων εξυπηρετεί πολλαπλούς σκοπούς στη μηχανική μάθηση:

- Αξιολόγηση επιδόσεων: Παρέχει πληροφορίες σχετικά με την ακρίβεια και την αξιοπιστία του.
- Επιλογή μοντέλου: Βοηθά στη σύγκριση και την επιλογή του μοντέλου με τις καλύτερες επιδόσεις μεταξύ πολλαπλών υποψηφίων.
- Ρύθμιση υπερπαραμέτρων: Καθοδηγεί την επιλογή των βέλτιστων υπερπαραμέτρων αξιολογώντας την επίδρασή τους στην απόδοση του μοντέλου.

- Αξιολόγηση γενίκευσης: Προσδιορίζει την ικανότητα του μοντέλου να κάνει ακριβείς προβλέψεις σε νέα, αθέατα δεδομένα.

Για προβλήματα ταξινόμησης, χρησιμοποιούνται συνήθως διάφορες μετρικές αξιολόγησης για την αξιολόγηση της απόδοσης του μοντέλου, βασισμένες στους ορισμούς των αληθώς θετικών (true positive)/αρνητικών (true negative) και των ψευδώς θετικών (false positive)/αρνητικών (false negative) παρατηρήσεων που εξηγούνται στο παρακάτω σχήμα.



Σχήμα 5.3: Πίνακας σύγκρισης προβλέψεων μοντέλου ταξινόμησης

Ορισμένες βασικές μετρικές περιλαμβάνουν:

- Accuracy: Το ποσοστό των σωστών προβλέψεων επί του συνολικού αριθμού των προβλέψεων.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

- Precision: Η αναλογία των αληθώς θετικών αποτελεσμάτων (σωστά προβλεπόμενα θετικά αποτελέσματα) επί του συνόλου των προβλεπόμενων θετικών αποτελεσμάτων.

$$Pre = \frac{TP}{TP + FP}$$

- Recall: Το ποσοστό των αληθώς θετικών αποτελεσμάτων επί όλων των πραγματικών θετικών αποτελεσμάτων.

$$Rec = \frac{TP}{TP + FN}$$

- F1 Score: Ο αρμονικός μέσος όρος των precision και recall, παρέχοντας ένα ισορροπημένο μέτρο.

$$F1 = \frac{2 \cdot Pre \cdot Rec}{Pre + Rec}$$

Η κάθε μια από αυτές τις μετρικές έχει την δική της σημασία και ανάλογα με την εφαρμογή μπορεί να θέλουμε να βελτιστοποιήσουμε μια σε βάρος μιας άλλης, όπως θα δούμε και στη συνέχεια. Σε προβλήματα σαν το δικό μας όπου υπάρχει πολύ άνιση κατανομή των κλάσεων προτιμάται σαν πιο αντικειμενικό μέτρο σύγκρισης ενός μοντέλου το F1 Score καθώς συνδυάζει την πληροφορία από τις άλλες μετρικές.

5.3 Ρύθμιση υπερπαραμέτρων

Για την εύρεση των βέλτιστων υπερπαραμέτρων θα ακολουθήσουμε την εξής διαδικασία: Αρχικά, για κάθε έναν από τους αλγορίθμους μηχανικής μάθησης που περιγράψαμε πιο πάνω θα ορίσουμε τις υπερπαραμέτρους που θα ρυθμιστούν. Για κάθε υπερπαραμέτρο θα ορίσουμε ένα σύνολο από πιθανές τιμές. Για κάθε συνδυασμό τιμών των υπερπαραμέτρων θα εκπαιδεύσουμε το μοντέλο στα δεδομένα εκπαίδευσης και θα αξιολογήσουμε την απόδοση του στα δεδομένα επικύρωσης. Όποιος συνδυασμός υπερπαραμέτρων δώσει τα καλύτερα αποτελέσματα θα θεωρηθεί ο βέλτιστος. Μεταξύ των διαφορετικών αλγορίθμων μηχανικής μάθησης όποιος δίνει τα καλύτερα αποτελέσματα θα θεωρηθεί ο βέλτιστος μαζί με τις βέλτιστες υπερπαραμέτρους του. Λόγω της έντονης ανομοιομορφία των κλάσεων η μετρική με την οποία θα γίνει η αξιολόγηση θα είναι το F1 Score.

Για τα δεδομένα εκπαίδευσης θα χρησιμοποιήσουμε τα δεδομένα του εργοστασίου 1, ήτοι 113.221 παραδείγματα, ενώ για δεδομένα επικύρωσης θα χρησιμοποιήσουμε αυτά του εργοστασίου 6 δηλ. 55.990 παραδείγματα. Είναι σημαντικό τα δεδομένα στα οποία αξιολογείται ένα μοντέλο να είναι ανθετά ώστε να αποφευχθούν φαινόμενα υπερπροσαρμογής. Επιλέχθηκε το εργοστάσιο 6, καθώς σε αυτό υπάρχουν βλάβες όλων των ειδών σε αντίθεση με άλλα εργοστάσια που κάποιες από τις βλάβες 1-5 δεν εμφανίζονται.

Υπενθυμίζεται ότι το αρχικό πρόβλημα πολλαπλής ταξινόμησης (πρόβλεψη τύπου βλάβης 1-5) έχει αναχθεί σε πέντε δυαδικά προβλήματα, ένα για κάθε τύπο βλάβης. Επομένως η διαδικασία εύρεσης βέλτιστων υπερπαραμέτρων και βέλτιστου μοντέλου που περιγράψαμε πιο πάνω θα επαναληφθεί για καθένα από τα πέντε δυαδικά προβλήματα. Συγκεκριμένα, η λογιστική παλινδρόμηση πέτυχε τα καλύτερα αποτελέσματα

Τα αποτελέσματα από τη ρύθμιση των υπερπαραμέτρων για κάθε τύπο βλάβης φαίνονται στους παρακάτω πίνακες:

Model	Best Hyperparameters	Score
Naive Bayes	–	0.52
Logistic Regression	$C = 0.01$	0.6
KNN	$n = 7$	0.57
LSTM	Layers=2, Nodes=100, LR=0.001, BS=32	0.58

Πίνακας 5.1: Ρύθμιση υπερπαραμέτρων - Βλάβη τύπου 1

Model	Best Hyperparameters	Score
Naive Bayes	–	0.55
Logistic Regression	$C = 0.01$	0.61
KNN	$n = 11$	0.59
LSTM	Layers=2, Nodes=50, LR=0.001, BS=64	0.57

Πίνακας 5.2: Ρύθμιση υπερπαραμέτρων - Βλάβη τύπου 2

Model	Best Hyperparameters	Score
Naive Bayes	–	0.49
Logistic Regression	$C = 0.01$	0.59
KNN	$n = 3$	0.58
LSTM	Layers=3, Nodes=100, LR=0.0001, BS=64	0.58

Πίνακας 5.3: Ρύθμιση υπερπαραμέτρων - Βλάβη τύπου 3

Model	Best Hyperparameters	Score
Naive Bayes	–	0.5
Logistic Regression	$C = 0.01$	0.55
KNN	$n = 5$	0.54
LSTM	Layers=2, Nodes=200, LR=0.01, BS=32	0.53

Πίνακας 5.4: Ρύθμιση υπερπαραμέτρων - Βλάβη τύπου 4

Model	Best Hyperparameters	Score
Naive Bayes	–	0.54
Logistic Regression	$C = 0.01$	0.63
KNN	$n = 3$	0.57
LSTM	Layers=2, Nodes=500, LR=0.001, BS=64	0.59

Πίνακας 5.5: Ρύθμιση υπερπαραμέτρων - Βλάβη τύπου 5

Από την παραπάνω διαδικασία προέκυψε ότι για όλα τα είδη βλάβης το μοντέλο που πετυχαίνει το καλύτερο F1 Score είναι η λογιστική παλινδρόμηση και μάλιστα σε όλες τις περιπτώσεις με την ίδια υπερπαραμέτρο $C=0.01$.

Βασικό ερώτημα που προκύπτει είναι πώς το LSTM δεν καταφέρνει να τα πάει καλύτερα από τη λογιστική παλινδρόμηση. Μια πιθανή εξήγηση είναι ότι έτσι όπως έχει μοντελοποιηθεί το πρόβλημα, ένα μοντέλο προσπαθεί να προβλέψει αν θα υπάρξει βλάβη μέσα στο επόμενο τρίωρο. Το LSTM, προκειμένου να κάνει μια τέτοια πρόβλεψη, χρησιμοποιεί δεδομένα από το παρελθόν. Έτσι τα δεδομένα που χρησιμοποιεί μπορεί να απέχουν πάνω από τρεις ώρες από τη βλάβη που επιχειρεί να προβλέψει και επομένως να είναι ανεξάρτητα. Το γεγονός αυτό ενδεχομένως δημιουργεί σύγχυση στο μοντέλο και οδηγεί σε λανθασμένες προβλέψεις.

Επιπλέον πρέπει να σημειωθεί ότι το LSTM και γενικά τα νευρωνικά δίκτυα έχουν έναν πολύ μεγάλο αριθμό από υπερπαραμέτρους που μπορούν να ρυθμιστούν. Στη συγκεκριμένη εργασία δοκιμάστηκαν πολύ λίγες διαφορετικές τιμές για τις διάφορες προς ρύθμιση υπερπαραμέτρους, καθώς δεν υπήρχε η υπολογιστική δύναμη για κάτι καλύτερο.

5.4 Αποτελέσματα

Εφόσον έχουμε επιλέξει τα μοντέλα που θα χρησιμοποιηθούν και έχουμε ρυθμίσει τις υπερπαραμέτρους, τους μπορούμε να προχωρήσουμε στο τελικό στάδιο αυτής της εργασίας. Χρησιμοποιούμε τώρα όλα τα διαθέσιμα δεδομένα μας από τα 33 εργοστάσια. Από αυτά τα 26 χρησιμοποιούνται για την εκπαίδευση του τελικού μοντέλου και τα 8 για την τελική αξιολόγηση.

Για τα 26 εργοστάσια εκπαίδευσης ακολουθούμε ακριβώς την ίδια προετοιμασία δεδομένων που δείξαμε στα παραπάνω. Στο τέλος ενώνουμε τα δεδομένα σε ένα ενιαίο σύνολο δεδομένων. Σπάμε αυτό το σύνολο δεδομένων σε 5 σύνολα, ένα για κάθε μεταβλητή-στόχο. Εκπαιδεύουμε ένα μοντέλο λογιστικής παλινδρόμησης για κάθε μεταβλητή στόχο σε ισορροπημένα σύνολα ακολουθώντας τη μέθοδο SMOTE.

Στο τέλος συνδυάζουμε τα 5 μοντέλα σε ένα ενιαίο μοντέλο που δέχεται ως είσοδο τις τιμές των ανεξάρτητων μεταβλητών και επιστρέφει 5 δυαδικές μεταβλητές που εκφράζουν το αν το μοντέλο προβλέπει βλάβη τύπου 1-5 μέσα στο επόμενο τρίωρο.

Εφαρμόζουμε το συνδυαστικό μοντέλο στα δεδομένα αξιολόγησης που αφορούν τα υπόλοιπα 8 εργοστάσια. Σημειώνουμε ότι αυτά τα δεδομένα δεν τα έχει δει ποτέ το μοντέλο και επομένως η επίδοση του μοντέλου είναι ρεαλιστική. Φυσικά οι μεταβλητές-στόχοι σε αυτά τα δεδομένα είναι ανισόρροπες. Ας δούμε τα αποτελέσματα, ξεχωριστά για κάθε μεταβλητή στόχο:

Fault Type: 1				
	precision	recall	f1-score	support
0.0	0.96	0.60	0.73	506100
1.0	0.15	0.73	0.24	48059
accuracy			0.61	554159
macro avg	0.55	0.66	0.49	554159
weighted avg	0.89	0.61	0.69	554159

Σχήμα 5.4: Results, fault type : 1

Βλέπουμε ένα συνολικό accuracy 0.61 καθώς και ότι το μοντέλο προβλέπει σωστά το 60% των περιπτώσεων μη-βλάβης τύπου 1 και το 73% των περιπτώσεων βλάβης τύπου 1. Επίσης, από τις φορές που η πρόβλεψη είναι μη-βλάβη το 96% όντως δεν υπάρχει βλάβη, ενώ από τις φορές που η πρόβλεψη είναι βλάβη μόνο το 15% είναι όντως βλάβη. Προχωράμε στις βλάβες τύπου 2:

Fault Type: 2				
	precision	recall	f1-score	support
0.0	0.97	0.67	0.79	516229
1.0	0.13	0.70	0.23	37930
accuracy			0.67	554159
macro avg	0.55	0.68	0.51	554159
weighted avg	0.91	0.67	0.75	554159

Σχήμα 5.5: Results, fault type : 2

Βλέπουμε ένα συνολικό accuracy 0.67 καθώς και ότι το μοντέλο προβλέπει σωστά το 67% των περιπτώσεων μη-βλάβης τύπου 2 και το 70% των περιπτώσεων βλάβης τύπου 2. Επίσης, από τις φορές που η πρόβλεψη είναι μη-βλάβη το 97% όντως δεν υπάρχει βλάβη, ενώ από τις

φορές που η πρόβλεψη είναι βλάβη μόνο το 13% είναι όντως βλάβη. Προχωράμε στις βλάβες τύπου 3:

```

Fault Type: 3
      precision    recall  f1-score   support

 0.0         1.00      0.72      0.84     545515
 1.0         0.05      0.93      0.09       8644

 accuracy          0.72     554159
 macro avg         0.52      0.83      0.47     554159
 weighted avg      0.98      0.72      0.82     554159
    
```

Σχήμα 5.6: Results, fault type : 3

Βλέπουμε ένα συνολικό accuracy 0.72 καθώς και ότι το μοντέλο προβλέπει σωστά το 72% των περιπτώσεων μη-βλάβης τύπου 3 και το 93% των περιπτώσεων βλάβης τύπου 3. Επίσης, από τις φορές που η πρόβλεψη είναι μη-βλάβη το 100% όντως δεν υπάρχει βλάβη, ενώ από τις φορές που η πρόβλεψη είναι βλάβη μόνο το 5% είναι όντως βλάβη. Προχωράμε στις βλάβες τύπου 4:

```

Fault Type: 4
      precision    recall  f1-score   support

 0.0         0.99      0.86      0.92     532343
 1.0         0.18      0.77      0.30       21816

 accuracy          0.86     554159
 macro avg         0.59      0.82      0.61     554159
 weighted avg      0.96      0.86      0.90     554159
    
```

Σχήμα 5.7: Results, fault type : 4

Βλέπουμε ένα συνολικό accuracy 0.86 καθώς και ότι το μοντέλο προβλέπει σωστά το 86% των περιπτώσεων μη-βλάβης τύπου 4 και το 77% των περιπτώσεων βλάβης τύπου 4. Επίσης, από τις φορές που η πρόβλεψη είναι μη-βλάβη το 99% όντως δεν υπάρχει βλάβη, ενώ από τις φορές που η πρόβλεψη είναι βλάβη μόνο το 18% είναι όντως βλάβη. Προχωράμε στις βλάβες τύπου 5:

Fault Type: 5				
	precision	recall	f1-score	support
0.0	0.99	0.79	0.88	541240
1.0	0.08	0.76	0.14	12919
accuracy			0.79	554159
macro avg	0.54	0.77	0.51	554159
weighted avg	0.97	0.79	0.86	554159

Σχήμα 5.8: Results, fault type : 5

Βλέπουμε ένα συνολικό accuracy 0.79 καθώς και ότι το μοντέλο προβλέπει σωστά το 79% των περιπτώσεων μη-βλάβης τύπου 5 και το 76% των περιπτώσεων βλάβης τύπου 5. Επίσης, από τις φορές που η πρόβλεψη είναι μη-βλάβη το 99% όντως δεν υπάρχει βλάβη, ενώ από τις φορές που η πρόβλεψη είναι βλάβη μόνο το 8% είναι όντως βλάβη.

Σε όλες τις περιπτώσεις παρατηρούμε το ίδιο μοτίβο. Το μοντέλο τα πηγαίνει πάρα πολύ καλά στις περιπτώσεις μη-βλάβης τόσο στο precision όσο και στο recall, ενώ στην στις περιπτώσεις βλάβης έχουμε πολύ καλά αποτελέσματα στο recall αλλά όχι τόσο καλά στο precision.

Αυτό οφείλεται στην ανισορροπία των κλάσεων. Εφόσον δεν είχαμε πολλά παραδείγματα από την κλάση βλάβης κατασκευάσαμε τεχνητά δεδομένα μέσω της μεθόδου SMOTE. Με αυτό τον τρόπο «πιέσαμε» το μοντέλο να μάθει οπωσδήποτε να προβλέπει τις βλάβες με αποτέλεσμα συχνά να προβλέπει βλάβες εκεί που δεν υπάρχουν. Αυτό ωστόσο δεν είναι απαραίτητα κακό και ο λόγος είναι ο εξής: θέλουμε να έχουμε ένα μοντέλο που να έχει δύο βασικά χαρακτηριστικά. Το πρώτο είναι όταν έρχεται βλάβη να την προβλέπει επιτυχώς. Αυτό ισοδυναμεί με το να έχει να έχει υψηλό recall η κλάση 1 για κάθε μεταβλητή-στόχο. Αυτό επιτυγχάνεται καθώς το μέσο recall στην κλάση 1 είναι 77.8%. Το δεύτερο είναι όταν το μοντέλο προβλέπει ότι δεν έρχεται βλάβη να μπορούμε να το εμπιστευτούμε και όντως να μην έρχεται βλάβη. Αυτό ισοδυναμεί με το να έχει το μοντέλο υψηλό precision στην κλάση 0. Και αυτό επιτυγχάνεται καθώς το μέσο precision στην κλάση 0 είναι 98,2%!

Το μοντέλο μας επιτυγχάνει ένα μέσο precision της τάξης του 11.8% στις περιπτώσεις των βλαβών, πράγμα που σημαίνει ότι μόλις περίπου μία στις δέκα προβλέψεις του μοντέλου που αφορούν βλάβη, ακολουθούνται όντως από μία. Φυσικά και η πρόβλεψη βλάβης εκεί που δεν υπάρχει πραγματικά βλάβη δεν είναι κάτι επιθυμητό, όμως στην πράξη, είναι προτιμότερο από το να έχουμε απρόβλεπτες βλάβες και διακοπή της παραγωγής. Στην συνέχεια θα διαπιστώσουμε την αλήθεια αυτού του ισχυρισμού μέσα από μια κοστολογική σύγκριση ανάμεσα στην υιοθέτηση και μη του προβλεπτικού μοντέλου αυτού.

5.5 Συγκριτική ανάλυση κόστους

Ας υποθέσουμε ότι στο συγκεκριμένο σύστημα παραγωγής έχουμε κατα μέσο όρο z βλάβες ανά έτος, έστω επίσης x_f το κόστος της αποκατάστασης μιας βλάβης, ενώ x_m το κόστος της προληπτικής συντήρησης. Τότε, στην περίπτωση υιοθέτησης του προβλεπτικού μοντέλου, θα έχουμε τα αντίστοιχα συνολικά κόστη αποκατάστασης βλαβών που το μοντέλο δεν κατάφερε να προβλέψει c_f και προληπτικής συντήρησης όταν το μοντέλο προβλέπει λανθασμένα βλάβη c_m (για ένα έτος):

$$c_f = z \cdot (100\% - Prec_{nf}) \cdot x_f$$

$$c_m = z \cdot (100\% - Prec_f) \cdot x_m$$

όπου $Prec_{nf}$ το precision για τις περιπτώσεις μη-βλάβης, ενώ $Prec_f$ για τις περιπτώσεις βλάβης. Πιο συγκεκριμένα, $(100\% - Prec_{nf})$ είναι το ποσοστό των περιπτώσεων βλαβών όπου το μοντέλο μας προβλέπει μη-βλάβη, οπότε και επιβαρυνόμαστε με το κόστος της αποκατάστασης της βλάβης, ενώ $(100\% - Prec_f)$ το ποσοστό των περιπτώσεων μη-βλάβης που το μοντέλο μας προβλέπει βλάβη, οπότε και επωμιζόμαστε το κόστος της προληπτικής συντήρησης. Επίσης, θα έχουμε το όφελος της διαφοράς του κόστους όταν το μοντέλο προβλέπει επιτυχώς μια επερχόμενη βλάβη:

$$p = z \cdot Prec_f \cdot (x_f - x_m)$$

Εξισώνοντας τα δύο κόστη με το κέρδος έχουμε:

$$z \cdot (100\% - Prec_{nf}) \cdot x_f + z \cdot (100\% - Prec_f) \cdot x_m = z \cdot Prec_f \cdot (x_f - x_m)$$

όπου ο ετήσιος αριθμός βλαβών απλοποιείται, καθώς είναι κοινός παράγοντας και έχουμε:

$$(100\% - 98.2\%) \cdot x_f + (100\% - 11.8\%) \cdot x_m = 98.2\% \cdot (x_f - x_m) \Rightarrow \frac{x_f}{x_m} = 1.934$$

Όπως καταλαβαίνουμε, αν ο λόγος των δύο κοστών είναι μεγαλύτερος από 1.934, η εφαρμογή του προβλεπτικού μας μοντέλου μας είναι πολύ πιθανόν να αποδώσει οικονομικά οφέλη σε

μακροχρόνιο ορίζοντα. Προφανώς, αυτή πρόκειται για μια απλουστευμένη ανάλυση κόστους, καθώς δεν λάβαμε υπόψιν διαφοροποιήσεις μεταξύ των τύπων των βλαβών, απλά δείχνει το γενικό πλαίσιο στο οποίο θα μπορούσε να κριθεί η εφαρμογή του μοντέλου σε ένα πραγματικό σύστημα παραγωγής. Αυτός ο ενδεικτικός λόγος, πάντως, αποτελεί μια πολύ ενθαρρυντική ένδειξη για το μοντέλο μας, καθώς είναι συχνό στην βιομηχανία το κόστος της αποκατάστασης μια βλάβης να είναι πολλαπλάσιο αυτού της συντήρησης, πόσο δε μάλλον αν λάβει κανείς υπ' όψιν την ζημία από την διακοπή της παραγωγής και όχι μόνο τα άμεσα οικονομικά κόστη.

Κεφάλαιο 6

Συμπεράσματα και μελλοντικές προεκτάσεις

Η μηχανική μάθηση έχει αναδειχθεί σε ένα ισχυρό εργαλείο για την ανίχνευση βλαβών, που μπορεί να φέρει επανάσταση στον τρόπο με τον οποίο οι βιομηχανίες προσεγγίζουν τη συντήρηση και την αξιοπιστία. Μέσω της ικανότητάς τους να μαθαίνουν μοτίβα και ανωμαλίες από τεράστιες ποσότητες δεδομένων, τα μοντέλα μηχανικής μάθησης έχουν αποδειχθεί αποτελεσματικά στην αυτοματοποίηση των διαδικασιών ανίχνευσης βλαβών, στην ενίσχυση της αποδοτικότητας και στη μείωση του χρόνου διακοπής λειτουργίας.

Η εφαρμογή αλγορίθμων μηχανικής μάθησης προσφέρει πολλά βασικά πλεονεκτήματα στην ανίχνευση σφαλμάτων. Πρώτον, τα μοντέλα αυτά μπορούν να χειριστούν μεγάλα και πολύπλοκα σύνολα δεδομένων, εξάγοντας πολύτιμες πληροφορίες και εντοπίζοντας κρυμμένα μοτίβα, άτι δύσκολο μέσω χειροκίνητης ανάλυσης. Επιπλέον, οι αλγόριθμοι μηχανικής μάθησης μπορούν να προσαρμόζονται και να μαθαίνουν από νέα δεδομένα, βελτιώνοντας συνεχώς την απόδοση και την ακρίβειά τους με την πάροδο του χρόνου. Έτσι, ένα μοντέλο σαν αυτό που κατασκευάσαμε μπορεί με τον καιρό να εκπαιδευτεί περαιτέρω με νέα δεδομένα και να προσαρμοστεί με τρόπο ώστε να αυξήσει την απόδοσή του.

Συνοπτικά, στην συγκεκριμένη εργασία στόχος ήταν η δημιουργία ενός προβλεπτικού μοντέλου για πολλαπλούς τύπους βλαβών μιας παραγωγικής μονάδας. Αρχικά, καταστρώθηκε ένα γενικό σχέδιο σύμφωνα με τις πρωτογενείς πληροφορίες του προβλήματος όπως ότι πρόκειται για μια ταξινόμηση πολλαπλών κλάσεων, την ανεξαρτησία των βλαβών κλπ. Στην συνέχεια έλαβε χώρα μια αρχική επεξεργασία των δεδομένων έτσι ώστε να έρθουν σε μια ομογενή και αξιοποιήσιμη μορφή, θεμελιώδεις κομμάτι οποιουδήποτε μοντέλου μηχανικής μάθησης.

Στη συνέχεια, προχωρήσαμε σε μια ανάλυση όλων των δεδομένων, τόσο εισόδου όσο και της μεταβλητής στόχου, ώστε να κατανοήσουμε κύρια χαρακτηριστικά τους όπως κατανομές, έκτοπα σημεία, αλληλεξαρτήσεις κλπ, στοιχεία θεμελιώδη για το επόμενο βήμα, που είναι η επιλογή των απαραίτητων ανεξάρτητων μεταβλητών του μοντέλου μας. Εδώ ήταν εμφανής και

η σχεδόν παντελής έλλειψη γνώσεων για το σύστημα και την φύση των μετρήσεων, πράγμα που συχνά αξιοποιείται σε πραγματικές εφαρμογές και βοηθάει σε μεγάλο βαθμό το μοντέλο στην επίτευξη καλύτερων αποτελεσμάτων. Χαρακτηριστικό παράδειγμα της χρησιμότητας αυτής αποτελεί η γνώση για τις μετρήσεις E1 και E2, η οποία φανέρωσε την αλληλεξαρτησία τους κάτι που δεν θα φαινόταν από στοιχεία όπως οι κατανομές ή η αλληλεξάρτησή τους.

Ένα ουσιώδες εμπόδιο που αντιμετωπίσαμε στην παρούσα εργασία είναι αυτό των μη σταθμισμένων κλάσεων, όπως είναι σύνηθες σε προβλήματα ανίχνευσης βλαβών, όπου οι περιπτώσεις μη-βλάβης πλειοψηφούν κατά πολύ στο σύνολο των δεδομένων μας. Για την αντιμετώπιση αυτού του φαινομένου, επιλέχθηκε η μέθοδος SMOTE, η οποία δημιουργεί περιπτώσεις βλαβών στα δεδομένα μας, λαμβάνοντας πληροφορίες από τις ήδη υπάρχουσες περιπτώσεις, και όχι με υποδειγματοληψία της πλειοψηφούσας κλάσης, καθώς αυτό μπορεί να απέκλειε σημαντική πληροφορία από το μοντέλο μας.

Τέλος, ορίσαμε κάποιους αλγόριθμους με τους οποίους θα προσπαθούσαμε να εκπαιδεύσουμε το μοντέλο μας, από πιο θεμελιώδεις όπως η λογιστική παλινδρόμηση και ο Naive Bayes έως πιο εξελιγμένους όπως ο αλγόριθμος KNN και τα δίκτυα LSTM. Στους αλγόριθμους αυτούς υπάρχουν συγκεκριμένες υπερπαραμέτροι που είναι βασικές για την απόδοση του εκάστοτε αλγόριθμου, για τις οποίες βρέθηκαν οι βέλτιστες τιμές, ενώ στη συνέχεια συγκρίθηκαν οι αποδόσεις των αλγόριθμων αυτών. Η λογιστική παλινδρόμηση έφερε τα καλύτερα αποτελέσματα με μέσο recall στις περιπτώσεις βλάβης ίσο με 77.8% και precision στις περιπτώσεις μη-βλάβης ίσο με 98.2%.

Το αρνητικό σημείο του μοντέλου μας είναι το χαμηλό precision στην κλάση βλάβης, το ποσοστό των φορών που όντως επέρχεται βλάβη όταν το μοντέλο μας το προβλέπει που είναι ίσο με 11.8%. Το να προβλέπει κανείς περισσότερες φορές βλάβες απ' όσες συμβαίνουν, αλλά να προβλέπει επιτυχώς όλες σχεδόν τις βλάβες είναι κάτι γενικά ωφέλιμο, αν θεωρήσει κανείς ότι το κόστος της αποκατάστασης της βλάβης είναι μεγαλύτερο από αυτό της προληπτικής συντήρησης.

Όπως είδαμε και στο τελευταίο μέρος του προηγούμενου κεφαλαίου, θα πρέπει να διεξάγεται μια κοστολογική έρευνα στο εκάστοτε σύστημα παραγωγής, όπου θα λαμβάνονται υπ' όψιν ο εξοπλισμός, τα άμεσα κόστη αποκατάστασης βλάβης και προληπτικής συντήρησης, τα έμμεσα κόστη διακοπής της παραγωγής, ενώ μπορεί να εφαρμόζονται επιπλέον έλεγχοι από τους υπεύθυνους της παραγωγής για το αν θα εμπιστεύονται μια πρόβλεψη βλάβης του μοντέλου.

Συμπερασματικά, η μηχανική μάθηση προσφέρει τεράστιες δυνατότητες στην ανίχνευση βλαβών, επιτρέποντας στις βιομηχανίες να μεταβούν από τις αντιδραστικές στις προληπτικές στρατηγικές συντήρησης. Με περαιτέρω πρόοδο στην ανάπτυξη αλγόριθμων, τη διαθεσιμότητα δεδομένων και την επιστημονική γνώση στον τομέα, τα συστήματα ανίχνευσης βλαβών που βασίζονται στη μηχανική μάθηση θα συνεχίσουν να εξελίσσονται, οδηγώντας σε βελτιωμένη

αξιοπιστία, παραγωγικότητα και εξοικονόμηση κόστους σε ένα ευρύ φάσμα βιομηχανιών. Αξιοποιώντας τη δύναμη της μηχανικής μάθησης, οι βιομηχανίες μπορούν να ξεκλειδώσουν νέες δυνατότητες στη βελτιστοποίηση της συντήρησης και της απόδοσης του εξοπλισμού, ενισχύοντας τελικά τη λειτουργική αριστεία και την ικανοποίηση των πελατών.

Βιβλιογραφία

1. Accenture and GE. (2015). *Industrial Internet Insights Report. Industrial Insights Report*. Available: https://www.accenture.com/us-en/_acnmedia/Accenture/next-gen/reassembling-industry/pdf/Accenture-Industrial-Internet-Changing-Competitive-Landscape-Industries.pdf
2. Lee J. (2015). *Industrial Big Data, Mechanical Industry Press, China*
3. Millman (1995) N. (ed, 2015). "Big data to unlock value from the industrial internet of things.,"
4. Alter A., Banerjee P., Daugherty P.E., and Negm W., (ed, 2014). "Driving Unconventional Growth through the Industrial Internet of Things.,"
5. Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer.
6. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
7. Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
8. Guyon, I., Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157-1182.
9. Japkowicz, N., Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5), 429-449.
10. 2015 PHM Data Challenge. (2015). Available: <https://www.phmsociety.org/events/conference/phm/15/data-challenge>.
11. He, H., Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284.
12. Liu, Y., Shen, D. (2018). Multimodal fusion and classification of Alzheimer's disease based on multi-task linear programming. *Pattern Recognition*, 83, 371-383.
13. Burges J.C.C. (2003). Some Notes on Applied Mathematics for Machine Learning. *Advanced Lectures on Machine Learning*, 21-40
14. Tang, J., Alelyani, S., Liu, H. (2014). Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, 37-64.

15. Huang, G., Ding, X., Zhou, H. (2012). Optimization of SVM hyperparameters using particle swarm optimization for imbalanced datasets. *Neurocomputing*, 96, 105-117.
16. Aggarwal, C. C. (2015). *Outlier Analysis*. Springer.
17. Davis, J., Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233-240). ACM.
18. Adeli E., Zhou Q. Pfefferbaum A. Sullivan E.V. Fei-Fei L. Niebles J.C. Pohl K. M. (2021). Representation Learning with Statistical Independence to Mitigate Bias *IEEE Winter Conf Appl Comput Vis. 2021 Jan.*, 2512–2522.
19. Kotsiantis, S., Pintelas, P. (2004). Machine learning for handling data imbalance: A case study. *European Journal of Operational Research*, 166(2), 491-502.
20. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.
21. Chawla, N. V. (2010). Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook* (pp. 875-886). Springer.
22. Bellman, R. (1966) Dynamic programming. *Science*, 153(3731), 34-37.
23. Veropoulos, K., Campbell, C., Cristianini, N. (1999). Controlling the sensitivity of support vector machines. In *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 55-60).
24. Chen, C., Liaw, A., Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*, 110.
25. Chawla, N. V., Lazarevic, A., Hall, L. O., Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 107-119). Springer.
26. Raskutti, B., Wong, F. (2004). Feature selection with prior feature probabilities. In *Proceedings of the 21st International Conference on Machine Learning* (pp. 81). ACM.
27. Kubat, M., Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the 14th International Conference on Machine Learning* (pp. 179-186).
28. Kubat, M., Holte, R. C., Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30(2-3), 195-215.
29. Chen, C., Li, T. (2006). Adaptive noise reduction for outlier detection. In *Proceedings of the 2006 SIAM International Conference on Data Mining* (pp. 612-616). SIAM.

30. Chawla, N. V., Setiono, R., Hossain, L. (2005). Robust decision trees for data mining from imbalanced data sets. In *Proceedings of the 2nd International Conference on Industrial Applications of Holonic and Multi-Agent Systems* (pp. 22-27). Springer.
31. Kubat, M., Holte, R. C., Matwin, S. (1997). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30(2-3), 195-215.
32. Japkowicz, N., Stephen, S. (2001). Learning from imbalanced data sets: A comparison of various strategies. In *Proceedings of the AAAI Workshop on Learning from Imbalanced Data Sets* (Vol. 68, No. 1).
33. Liu, Y., Wu, J., Zhou, S. (2016). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46(2), 270-283.
34. Chawla, N. V., Kegelmeyer, W. P. (2005). Learning ensembles from bites: A scalable and accurate approach. *Journal of Machine Learning Research*, 6(Aug), 489-511.
35. Fernández, A., García, S. (2010). An insight into imbalanced big data classification. *Knowledge-Based Systems*, 42, 97-108.
36. Li, J., Cheng, K., Wang, S., Morstatter, F. (2018). Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6), 1-45.
37. Chen, S., He, H., Sun, G. (2009). Boosting for imbalanced data learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(4), 964-976.
38. Batista, G. E., Prati, R. C., Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20-29.
39. Liu, Y., Wu, J., Zhou, S. (2009). Exploratory under-sampling for class-imbalance learning. In *Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery* (Vol. 6, pp. 273-277). IEEE.
40. Fernández, A., López, V., Galar, M., del Jesus, M. J., Herrera, F. (2012). Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-Based Systems*, 25, 13-21.
41. Garcia, S., Herrera, F. (2008). An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, 9(Jan), 2677-2694.
42. Garcia, S., Herrera, F. (2010). Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. *Evolutionary Computation*, 19(1), 1-28.
43. Chawla, N. V., Lazarevic, A., Hall, L. O., Bowyer, K. W. (2003). *SMOTEBoost: Improving prediction of the minority class in boosting*. Springer.