ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
& ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Ανάπτυξη Αλγορίθμων αξιολόγησης ανοικτών δεδομένων για συστήματα IoT βασισμένων σε τεχνικές μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΚΩΝΣΤΑΝΤΙΝΟΣ ΔΕΛΗΟΓΛΑΝΗΣ

Επιβλέπων: Λούτα Μαλαματή

Αναπληρώτρια Καθηγήτρια

ΚΟΖΑΝΗ/ΣΕΠΤΕΜΒΡΙΟΣ/2023

HELLENIC DEMOCRACY
UNIVERSITY OF WESTERN MACEDONIA
SCHOOL OF ENGINEERING
DEPARTMENT OF ELECTRICAL
& COMPUTER ENGINEERING

ece.uowm.gr

Development of Algorithms for evaluating open data for IoT system with Machine Learning technics

THESIS

KONSTANTINOS DELIOGLANIS

SUPERVISOR: Louta Malamati

Assistant Professor

KOZANI/SEPTEMBER/2023

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
& ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν. 1599/1986 και τα άρθρα 2,4,6 παρ. 3 του Ν. 1256/1982, η παρούσα Διπλωματική Εργασία με τίτλο "Ανάπτυξη Αλγορίθμων αξιολόγησης ανοικτών δεδομένων για συστήματα IoT βασισμένων σε τεχνικές μηχανικής μάθησης" καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας και αναφέρονται ρητώς μέσα στο κείμενο που συνοδεύουν, και η οποία έχει εκπονηθεί στο Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Δυτικής Μακεδονίας, υπό την επίβλεψη του μέλους του Τμήματος κ. Μαλαματής Λούτα αποτελεί αποκλειστικά προϊόν προσωπικής εργασίας και δεν προσβάλλει κάθε μορφής πνευματικά δικαιώματα τρίτων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή / και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και μόνο.

Υπογραφή Φοιτητή: _____

Περίληψη

Η ανίχνευση ανωμαλιών παίζει κρίσιμο ρόλο στον εντοπισμό μη φυσιολογικής συμπεριφοράς σε δεδομένα αισθητήρων, επιτρέποντας έγκαιρη παρέμβαση και διατήρηση της ακεραιότητας του συστήματος. Συγκεκριμένα, η διαδικασία επιτρέπει την αναγνώριση διαφορετικής συμπεριφοράς από την προβλεπόμενη. Αυτό είναι ζωτικής σημασίας για πολλές εφαρμογές και συστήματα, καθώς η έγκαιρη ανίχνευση τέτοιων ανωμαλιών επιτρέπει την άμεση παρέμβαση και τη διόρθωση τους, εξασφαλίζοντας τη σταθερότητα και την αξιοπιστία του συστήματος.

Σε αυτήν τη μελέτη, εξετάζουμε την επίδοση και την αποτελεσματικότητα διαφόρων τεχνικών ανίχνευσης ανωμαλιών (ταξινομητών) σε δεδομένα αισθητήρων, που βασίζονται στη μηχανική μάθηση όπως οι Μηχανές Διανυσμάτων Υποστήριξης, το Isolation Forest και οι Κ Πλησιέστεροι Γείτονες. Τα δεδομένα τα οποία μελετήθηκαν προήλθαν από δεδομένα αισθητήρων εδάφους, ποιότητας ανακτημένου νερού αλλά και αγρο-μετεωρολογικών σταθμών που είναι εγκατεστημένοι σε πειραματικό αγροτεμάχιο και σε έξοδο βιολογικού καθαρισμού. Η παρούσα μελέτη παρουσιάζει μια σφαιρική ανάλυση που περιλαμβάνει προεπεξεργασία των δεδομένων και αφαίρεση ακραίων τιμών, εκπαίδευση μοντέλων, και αξιολόγηση προβλέψεων. Αρχικά, τα δεδομένα προ επεξεργάζονται και αφαιρούνται οι ακραίες τιμές. Στη συνέχεια, εκπαιδεύονται οι ταξινομητές στα προ-επεξεργασμένα δεδομένα. Τέλος, μέσω των πειραμάτων, οι διάφοροι ταξινομητές συγκρίνονται για τους διάφορους αισθητήρες επιτρέποντας την κατανόηση των προτύπων απόδοσης και των τάσεων μεταξύ των ταξινομητών και των αισθητήρων, τον εντοπισμό συσχετίσεων, παραλλαγών και δυνητικά ακραίων σημείων στις μετρήσεις απόδοσης χρησιμοποιώντας το scatter plot.

Λέξεις Κλειδιά

Ανίχνευση ανωμαλιών, Δεδομένα αισθητήρων, Απόδοση ταξινομητή, Προ-επεξεργασία δεδομένων, Εκπαίδευση Μοντέλων, Αξιολόγηση Προβλέψεων

Abstract

The detection of anomalies plays a critical role in identifying abnormal behavior in sensor data, allowing timely intervention and maintaining the integrity of the system. Specifically, the process allows for recognizing behavior that differs from what is expected. This is vital for many applications and systems, as early detection of such anomalies enables immediate intervention and their correction, ensuring the stability and reliability of the system.

A comprehensive analysis is presented, encompassing data preprocessing, outlier removal, model training, and prediction evaluation. The data under study originated from soil sensors, recovered water quality, and agrometeorological stations installed in an experimental agricultural plot and at the outlet of biological treatment. Through experiments, different classifiers were compared for various sensors, enabling an understanding of performance patterns and trends among classifiers and sensors, identifying correlations, variations, and potentially extreme points in performance measurements using the scatter plot.

Keywords

Anomaly detection, Sensor data, Classifier performance, Scatter plot visualization, Outlier scores, Data preprocessing, Model training

Ευχαριστίες

Θα ήθελα να εκφράσω τις ευχαριστίες μου προς την επιβλέπουσα καθηγήτρια της διπλωματικής μου εργασίας, την αναπληρώτρια καθηγήτρια του Τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Δυτικής Μακεδονίας, κ. Λούτα Μαλαματή για την πολύτιμη καθοδήγηση της και την εξαιρετική συνεργασία.

Επιπλέον, θα ήθελα να ευχαριστήσω θερμά την μηχανικό και υποψήφια Διδάκτορα του Πανεπιστημίου Δυτικής Μακεδονίας, Κωνσταντίνα Μπαντή, για την υπομονή της, την εμπιστοσύνη που μου έδειξε και υποστήριξη που μου παρείχε καθ' όλη τη διάρκεια αυτής της προσπάθειας μέχρι την ολοκλήρωση της διπλωματικής μου εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω όλους τους καθηγητές και τις καθηγήτριες του Τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών για τις πολύτιμες γνώσεις και την υποστήριξη που μου παρείχαν καθ' όλη τη διάρκεια των σπουδών μου.

Τέλος, θέλω από τα βάθη της καρδιάς μου να ευχαριστήσω την οικογένειά μου, για την ανεκτίμητη στήριξή τους, τόσο στην παρούσα εργασία, όσο και στην ολοκλήρωση των σπουδών μου.

Κωνσταντίνος Δεληογλάνης

ΚΟΖΑΝΗ/ΑΥΓΟΥΣΤΟΣ/2023

Περιεχόμενα

Πίνακας Εικόνων

Κεφάλαιο 1: Εισαγωγή

Στον γρήγορα εξελισσόμενο τομέα της τεχνολογίας, η ανάπτυξη και εφαρμογή συστημάτων βασισμένων σε αισθητήρες έχει αποδειχθεί καθοριστική για μια πληθώρα εφαρμογών, από την αυτοματοποίηση του σπιτιού και την παρακολούθηση του περιβάλλοντος μέχρι την υγεία και τον βιομηχανικό έλεγχο. Με την έξαρση των δεδομένων αισθητήρων, η ικανότητα να επεξεργαζόμαστε και να ερμηνεύουμε αποτελεσματικά και αποδοτικά αυτές τις πληροφορίες είναι ζωτικής σημασίας. Έτσι, η ανίχνευση ανωμαλιών αποκτά ουσιαστική σημασία καθώς επιτρέπει την εξαγωγή σημαντικών πληροφοριών και τη διασφάλιση της ακρίβειας και της αξιοπιστίας του συστήματος.

Η ανίχνευση ανωμαλιών αποτελεί έναν κρίσιμο τομέα της επεξεργασίας δεδομένων και της μηχανικής μάθησης, και είναι ένα πρόβλημα που έχει ερευνηθεί εντός διαφόρων τομέων έρευνας και εφαρμογών [1]. Στόχος της είναι η εντοπισμός ανωμαλιών ή αποκλίσεων σε ένα σύνολο δεδομένων, τα οποία συχνά δείχνουν μη φυσιολογική ή αναμενόμενη συμπεριφορά.

Οι τεχνικές ανίχνευσης ανωμαλιών είναι ιδιαίτερα χρήσιμες σε διάφορες εφαρμογές, όπως η ανίχνευση απάτης, η διαχείριση δικτύων, η βιοπληροφορική, και η παρακολούθηση συστημάτων υγείας, μεταξύ άλλων. Οι τεχνικές αυτές κατατάσσονται σε διάφορες κατηγορίες ανάλογα με τον τρόπο λειτουργίας τους, τις προσεγγίσεις που υιοθετούν και τα δεδομένα που αναλύουν. Αυτές μπορούν να είναι εποπτικές, ημι-εποπτικές ή μη εποπτικές, και μπορεί να βασίζονται σε στατιστικές μεθόδους, μηχανική μάθηση ή ακόμα και συνδυασμό των δύο.

Στο πλαίσιο αυτό, θα εξετάσουμε τις βασικές τεχνικές και προσεγγίσεις στην ανίχνευση ανωμαλιών, τα πλεονεκτήματα και τις προκλήσεις τους, καθώς και τους τομείς στους οποίους εφαρμόζονται ευρέως.

Αναγνωρίζοντας τον ζωτικό ρόλο της αποτελεσματικής οπτικοποίησης των αποτελεσμάτων στην ανάλυση δεδομένων, η έρευνα δίνει ιδιαίτερη προσοχή στην παρουσίαση των αποτελεσμάτων. Άλλωστε, ο στόχος δεν ήταν απλώς να παράγει τα αποτελέσματα, αλλά και να διευκολύνει την ερμηνεία και την κατανόησή τους, ένα βήμα που θα μπορούσε να οδηγήσει σε ενημερωμένη λήψη αποφάσεων.

Η διατριβή περιλαμβάνει όλες αυτές τις διαδικασίες, οι οποίες έχουν υλοποιηθεί χρησιμοποιώντας τη γλώσσα προγραμματισμού Python, συμπληρωμένη από μια ποικιλία πακέτων που έχουν επιλεγεί και χρησιμοποιηθεί με προσοχή, το καθένα συμβάλλοντας στον συνολικό στόχο της αποτελεσματικής ανίχνευσης ανωμαλιών.

Ο στόχος αυτής της έρευνας δεν είναι απλά ένας λειτουργικός αλγόριθμος, αλλά μια συστηματική, λεπτομερής έρευνα για την ανάπτυξη και βελτιστοποίηση μοντέλων μηχανικής μάθησης για την ανίχνευση ανωμαλιών. Η παρούσα έρευνα υπογραμμίζει την αξία της εμπεριστατωμένης ανάλυσης και της προσεκτικής σχεδίασης στην επίτευξη αυτού του στόχου, προσφέροντας έναν πολύτιμο οδηγό για ερευνητές, μηχανικούς και

άλλους επαγγελματίες που ενδιαφέρονται για την ανίχνευση ανωμαλιών και την ανάπτυξη σχετικών λύσεων.

1.1 Αντικείμενο της διπλωματικής

Το αντικείμενο αυτής της διπλωματικής εργασίας είναι μια κριτική εξερεύνηση του πώς οι τεχνικές μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για την ανάπτυξη εξειδικευμένων αλγορίθμων για την αξιολόγηση ανοικτών δεδομένων, επικεντρώνοντας ειδικά στα συστήματα IoT (Internet of things - Διαδίκτυο των πραγμάτων).

Τα συστήματα IoT παράγουν μια τεράστια ποσότητα δεδομένων, τα οποία συχνά μπορούν να είναι ανοιχτά και ελεύθερα προσβάσιμα. Ωστόσο, τα αρχικά δεδομένα που συλλέγονται από αυτά τα συστήματα μπορούν να είναι θορυβώδη, αδόμητα και υψηλής διάστασης, καθιστώντας την άμεση εξαγωγή σημαντικών πληροφοριών δύσκολη. Εδώ λοιπόν είναι όπου οι τεχνικές μηχανικής μάθησης μπορούν να παίξουν έναν μετασχηματικό ρόλο. Χρησιμοποιώντας αλγόριθμους που μπορούν να μαθαίνουν και να παίρνουν αποφάσεις από τα δεδομένα, μπορούμε να αναλύουμε, να ερμηνεύουμε και να χρησιμοποιούμε αυτές τις πληροφορίες πιο αποτελεσματικά.

Στο πλαίσιο αυτής της διατριβής, οι αλγόριθμοι που θα αναπτυχθούν θα εξυπηρετούν τον διπλό σκοπό της ανίχνευσης ανωμαλιών και της αξιολόγησης της απόδοσης. Οι ανωμαλίες στα συστήματα IoT μπορούν να οδηγήσουν σε σημαντικές αποτυχίες ή ανεπάρκειες του συστήματος, και η έγκαιρη ανίχνευση τέτοιων ανωμαλιών είναι ζωτική για τη διατήρηση της απόδοσης και της ακεραιότητας του συστήματος.

Μια σύντομη επισκόπηση των βημάτων που εμπλέκονται, βασισμένη στον κώδικα Python και την δομή της πτυχιακής, περιλαμβάνει:

1. **Συλλογή και Προεπεξεργασία Δεδομένων**: Τα δεδομένα ανακτώνται από ένα API χρησιμοποιώντας τη βιβλιοθήκη 'requests' και τα αποτελέσματα μετατρέπονται σε DataFrame της Pandas. Στη συνέχεια, τα δεδομένα καθαρίζονται με την απόρριψη περιττών στηλών και γραμμών με κενές καταχωρήσεις. Το σύνολο δεδομένων στη συνέχεια διαχωρίζεται σε σύνολα εκπαίδευσης και δοκιμής χρησιμοποιώντας τη συνάρτηση 'train_test_split' από το Scikit-learn.
2. **Εντοπισμός Ανωμαλιών**: Χρησιμοποιούνται διάφορα μοντέλα εντοπισμού ανωμαλιών, συμπεριλαμβανομένων του Isolation Forest, K Nearest Neighbors (k-NN), OneClassSVM και ενός Autoencoder. Για κάθε ένα από αυτά τα μοντέλα, το σύνολο εκπαίδευσης χρησιμοποιείται για την προσαρμογή του μοντέλου, και στη συνέχεια το μοντέλο χρησιμοποιείται για την πρόβλεψη ανωμαλιών στο σύνολο δοκιμής. Οι αρχικές βαθμολογίες ανωμαλίας για τα σύνολα εκπαίδευσης και δοκιμής αποθηκεύονται, μαζί με τις προβλεπόμενες ετικέτες και τις βαθμολογίες εμπιστοσύνης για το σύνολο δοκιμής.
3. **Μείωση Διαστατικότητας**: Για την οπτικοποίηση των δεδομένων και των αποτελεσμάτων εντοπισμού, πραγματοποιείται μείωση της διαστατικότητας χρησιμοποιώντας την Ανάλυση Κύριων Συνιστωσών-Principal Component Analysis

(PCA) από το Scikit-learn. Αυτό μειώνει τα δεδομένα αισθητήρων υψηλής διαστατικότητας σε μόλις δύο διαστάσεις.

4. **Οπτικοποίηση**: Τα μειωμένα δεδομένα των αισθητήρων στη συνέχεια απεικονίζονται σε ένα διάγραμμα διασποράς, με διάφορα χρώματα που δείχνουν τα κανονικά και τα ανωμαλικά σημεία δεδομένων. Αυτή η οπτική απεικόνιση βοηθά στην κατανόηση της αποτελεσματικότητας κάθε μοντέλου στην εντοπισμό ανωμαλιών στα δεδομένα.

Κεφάλαιο 2: Θεωρητικό υπόβαθρο

Σε αυτό το κεφάλαιο την εργασίας θα αναφερθούν τα κύρια θεωρητικά στοιχεία της μελέτης και θα αναλυθούν βασικές έννοιες αναγκαίες για την κατανόηση της έρευνας.

2.1 Τεχνητή Νοημοσύνη

Η Τεχνητή Νοημοσύνη αποτελεί έναν από τους πιο δυναμικούς και προκλητικούς τομείς της σύγχρονης επιστήμης. Η προσπάθεια αναπαραγωγής της ανθρώπινης σκέψης μέσω μηχανών δεν είναι εύκολο έργο, αλλά έχει δώσει ήδη ενδιαφέρουσες εφαρμογές.

Όπως τονίζει ο Russell και ο Norvig στο έργο τους "Artificial Intelligence: A Modern Approach" [12], η τεχνητή νοημοσύνη στρέφεται προς την αναπαράσταση των πληροφοριών, τη λήψη αποφάσεων με βάση αυτές και την εκμάθηση από νέες εμπειρίες.

Η ανθρώπινη ικανότητα για ερμηνεία του κόσμου έχει παρασταθεί σαν μοναδική, ωστόσο, με την εξέλιξη της τεχνητής νοημοσύνης, οι μηχανές έχουν αρχίσει να εκτελούν ερμηνευτικές διεργασίες παρόμοιες με τις ανθρώπινες. Το "Deep Learning" [13], όπως περιγράφεται από τον Goodfellow et al., είναι ένα παράδειγμα αυτής της προσέγγισης.

Η αναφορά στον ανθρώπινο εγκέφαλο ως πρωτότυπο για την ανάπτυξη της τεχνητής νοημοσύνης είναι ιδιαίτερα σημαντική. Όπως περιγράφει ο Searle στο δοκίμιο του "Minds, Brains, and Programs" [14], η αναπαράσταση της σκέψης απαιτεί μια βαθύτερη κατανόηση της λειτουργίας του εγκεφάλου.

Η προσομοίωση της ανθρώπινης σκέψης μέσω της τεχνητής νοημοσύνης αποτελεί μια συνεχιζόμενη προσπάθεια. Είναι βέβαιο πως οι επιστήμονες και οι μηχανικοί θα συνεχίσουν να αναζητούν νέους τρόπους για να προσεγγίσουν αυτό το εγχείρημα.

Η έρευνα και η ανάπτυξη στον τομέα της τεχνητής νοημοσύνης χωρίζονται σε δύο κύριους κλάδους:

• Εφαρμοσμένη τεχνητή νοημοσύνη: Αυτός ο κλάδος χρησιμοποιεί τις αρχές της προσομοίωσης της ανθρώπινης σκέψης για την εκτέλεση μιας συγκεκριμένης εργασίας. Τα συστήματα που αναπτύσσονται με βάση την εφαρμοσμένη τεχνητή νοημοσύνη συνήθως έχουν περιορισμένες δυνατότητες και εξειδικευμένους σκοπούς.

• Γενικευμένη τεχνητή νοημοσύνη: Αυτός ο κλάδος αποσκοπεί στην ανάπτυξη μηχανών που μπορούν να αντιμετωπίσουν οποιαδήποτε εργασία, όπως ένας άνθρωπος. Η γενικευμένη τεχνητή νοημοσύνη είναι ένας στόχος που εξακολουθεί να είναι στο προσκήνιο της ερευνητικής προσπάθειας στον τομέα του AI (Τεχνητή Νοημοσύνη-Artificial Intelligence). Απαιτεί πιο πλήρη κατανόηση του ανθρώπινου εγκεφάλου απ' ό,τι έχουμε σήμερα, καθώς και περισσότερη υπολογιστική ισχύ απ' ό,τι είναι συνήθως διαθέσιμο στους ερευνητές. Μια νέα γενιά τεχνολογίας chip υπολογιστών, οι

νευρομορφικοί επεξεργαστές, σχεδιάζεται για την πιο αποτελεσματική εκτέλεση κώδικα προσομοιωτή εγκεφάλου. Συστήματα όπως η γνωσιακή υπολογιστική πλατφόρμα Watson της IBM (International Business Machines Corporation) χρησιμοποιούν προσομοιώσεις υψηλού επιπέδου ανθρώπινων νευρολογικών διεργασιών για να εκτελέσουν ένα ολοένα αυξανόμενο φάσμα εργασιών.

2.2 Εφαρμογή της Τεχνητής Νοημοσύνης

Η τεχνητή νοημοσύνη εμφανίζει επίσης αναπτυσσόμενη παρουσία σε ποικίλους βιομηχανικούς τομείς με βάση την φιλοσοφία της ανθρώπινης σκέψης όπως παρουσιάστηκε παραπάνω, αποτελώντας ένα απαραίτητο εργαλείο για την αποτελεσματικότητα και την καινοτομία.

Στον χρηματοοικονομικό τομέα, η τεχνητή νοημοσύνη συμβάλλει στον εντοπισμό και την πρόληψη απάτης, ενώ ταυτόχρονα ενισχύει την ποιότητα της εξυπηρέτησης πελατών, προβλέποντας τις μελλοντικές τους ανάγκες και προσαρμόζοντας τις υπηρεσίες ανάλογα. Στον κατασκευαστικό τομέα, υποστηρίζει τη διαχείριση του εργατικού δυναμικού και των διαδικασιών παραγωγής, και προβλέπει σφάλματα πριν εμφανιστούν, επιτρέποντας την προγνωστική συντήρηση.
Στην καθημερινή μας ζωή, η επίδραση της τεχνητής νοημοσύνης γίνεται ολοένα και πιο αισθητή. Οι προσωπικοί βοηθοί στα smartphones μας, όπως η Siri και το Google Assistant, χρησιμοποιούν τεχνολογίες τεχνητής νοημοσύνης για να διευκολύνουν την καθημερινότητά μας, ενώ τα αυτόνομα και αυτοοδηγούμενα αυτοκίνητα προετοιμάζουν το έδαφος για ένα μέλλον όπου η τεχνητή νοημοσύνη θα κυριαρχεί στην κινητικότητα και τις μεταφορές.

Η τεχνητή νοημοσύνη έχει εφαρμογές σε πολλούς άλλους τομείς, συμπεριλαμβανομένων της ρομποτικής, των οικονομικών, της υγείας, της ενέργειας, της επιστήμης, της εκπαίδευσης και πολλών άλλων. Καθώς η τεχνολογία προχωρά, η σημασία και οι εφαρμογές της τεχνητής νοημοσύνης αναμένεται να αυξηθούν σημαντικά.

Γενικά, τα συστήματα τεχνητής νοημοσύνης λειτουργούν απορροφώντας μεγάλες ποσότητες δεδομένων εκπαίδευσης με ετικέτα, αναλύοντας τα δεδομένα για συσχετίσεις και μοτίβα και χρησιμοποιώντας αυτά τα μοτίβα για να κάνουν προβλέψεις για μελλοντικές καταστάσεις. Με αυτόν τον τρόπο, ένα chatbot που τροφοδοτείται με παραδείγματα κειμένου μπορεί να μάθει να δημιουργεί ρεαλιστικές ανταλλαγές με ανθρώπους ή ένα εργαλείο αναγνώρισης εικόνων μπορεί να μάθει να αναγνωρίζει και να περιγράφει αντικείμενα σε εικόνες εξετάζοντας εκατομμύρια παραδείγματα. Οι νέες, ταχέως βελτιωμένες τεχνικές τεχνητής νοημοσύνης μπορούν να δημιουργήσουν ρεαλιστικό κείμενο, εικόνες, μουσική και άλλα μέσα.

Ο προγραμματισμός AI εστιάζει σε γνωστικές δεξιότητες που περιλαμβάνουν τα ακόλουθα:

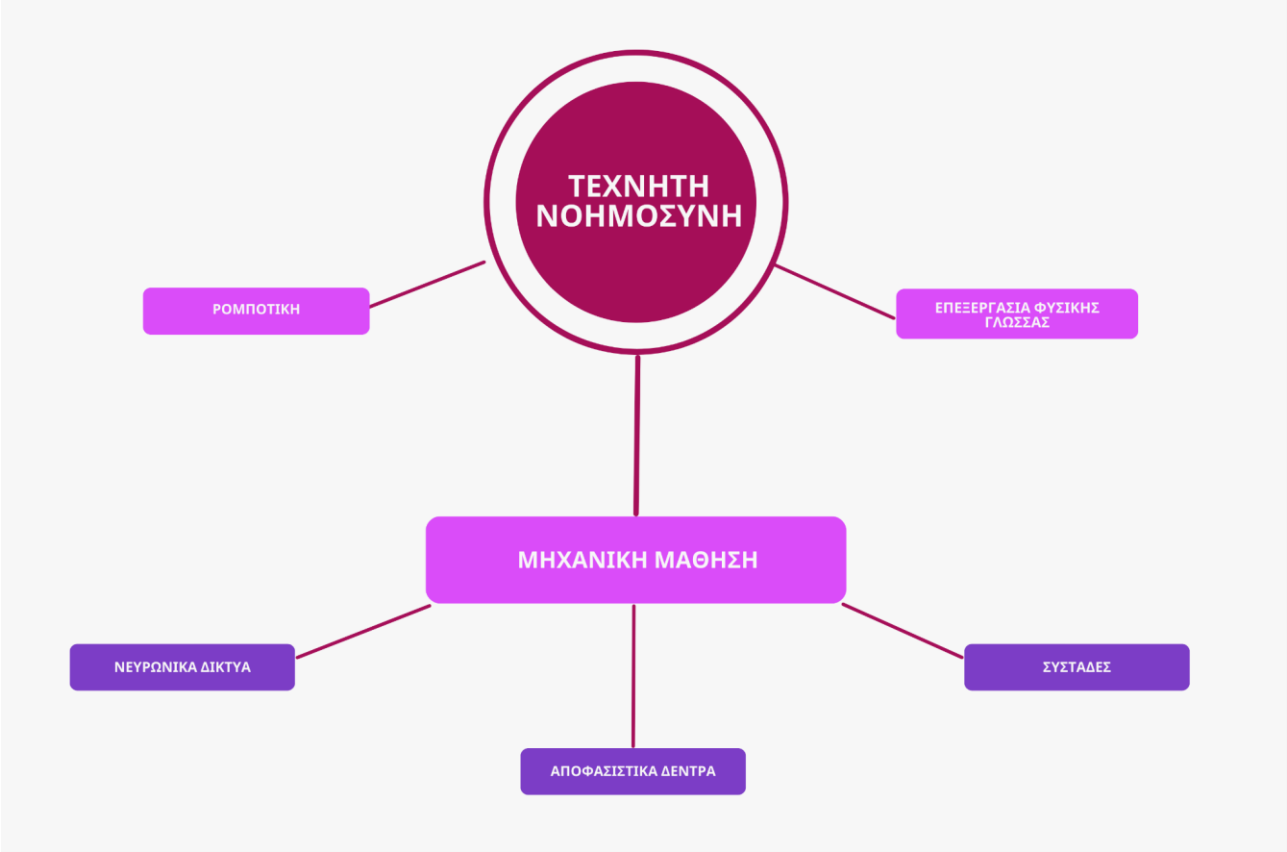
- Μάθηση. Αυτή η πτυχή του προγραμματισμού τεχνητής νοημοσύνης εστιάζει στην απόκτηση δεδομένων και στη δημιουργία κανόνων για το πώς να τα μετατρέψετε σε πληροφορίες που μπορούν να χρησιμοποιηθούν. Οι κανόνες, που ονομάζονται αλγόριθμοι, παρέχουν στις υπολογιστικές συσκευές οδηγίες βήμα προς βήμα για το πώς να ολοκληρωθεί μια συγκεκριμένη εργασία.

- Αιτιολογία. Αυτή η πτυχή του προγραμματισμού AI εστιάζει στην επιλογή του σωστού αλγόριθμου για την επίλυση ενός συγκεκριμένου προβλήματος.

- Αυτο-βελτίωση. Μέσω του προγραμματισμού AI, οι υπολογιστικές συσκευές μπορούν να αναπτύξουν τις δικές τους οδηγίες για να εκπαιδεύσουν τον εαυτό τους για να εκτελέσουν καλύτερα μια εργασία.

- Προβλεπτική ικανότητα. Η προγραμματιστική γλώσσα AI μπορεί να χρησιμοποιηθεί για να αναλύσει τα δεδομένα που συλλέγει και να κάνει προβλέψεις για μελλοντικά γεγονότα.

2.3 Μηχανική Μάθηση



Εικόνα 1. Συσχέτιση Τεχνητής Νοημοσύνης με Μηχανική Μάθηση

Η μηχανική μάθηση είναι ένας κλάδος της τεχνητής νοημοσύνης που επιτρέπει στις μηχανές να μαθαίνουν χωρίς να προγραμματίζονται για αυτόν τον συγκεκριμένο σκοπό. Μια βασική ικανότητα για τη δημιουργία συστημάτων που δεν είναι μόνο έξυπνα, αλλά αυτόνομα και ικανά να αναγνωρίζουν μοτίβα στα δεδομένα για να τα μετατρέψουν σε προβλέψεις. Αυτή η τεχνολογία είναι επί του παρόντος παρούσα σε έναν ατελείωτο αριθμό εφαρμογών, όπως οι προτάσεις Netflix και Spotify, οι έξυπνες απαντήσεις του Gmail ή η φυσική ομιλία Alexa και Siri.

Στο πλαίσιο της εκδήλωσης Open Summit, εκπρόσωποι από τη Google, τη Telefónica και τη BBVA ανέπτυξαν τη συζήτηση σχετικά με τις δυνατότητες της τεχνητής νοημοσύνης στον επιχειρηματικό κόσμο. Συζητήθηκε η ικανότητα της τεχνητής νοημοσύνης να εξάγει ουσιαστική αξία από δεδομένα, αποφέροντας οφέλη τόσο στους ανθρώπους όσο και στην κοινωνία γενικότερα. Η τεχνητή νοημοσύνη προσφέρει δομημένες λύσεις, επιτρέποντας την επεξεργασία δεδομένων σε μεγάλη κλίμακα.[15]

Ο José Luis Espinoza, επιστήμονας δεδομένων στο Μεξικό, τόνισε την ικανότητα της μηχανικής μάθησης να αναγνωρίζει πρότυπα και να μετασχηματίζει δεδομένα σε προγράμματα υπολογιστών, που μπορούν να εκτελούν παρεμβολές σε νέα σύνολα δεδομένων[15]. Αυτή η δυνατότητα βρίσκει εφαρμογές σε διάφορους τομείς, όπως οι

μηχανές αναζήτησης, η ρομποτική, η ιατρική διάγνωση, και ακόμα και στον εντοπισμό απάτης σε πιστωτικές κάρτες.

Παρότι η τεχνητή νοημοσύνη είναι τώρα στο προσκήνιο, λύνοντας προβλήματα στο Go και τους κύβους Rubik, οι ρίζες της βρίσκονται στον παρελθόν. Οι στατιστικές είναι η βάση της αυτοματοποιημένης μάθησης, η οποία χρησιμοποιεί αλγόριθμους για την ανάλυση μεγάλων όγκων δεδομένων, προσπαθώντας να βρει την καλύτερη λύση για κάθε δεδομένο πρόβλημα.

2.4 Είδη Μηχανικής Μάθησης

Εικόνα 2. Είδη Μηχανικής Μάθησης

2.4.1 Εποπτευόμενη/Επιβλεπόμενη Μηχανική Μάθηση

Η επιβλεπόμενη μηχανική μάθηση είναι ένας τύπος τεχνικής μηχανικής μάθησης όπου το μοντέλο μαθαίνει από καλά ετικετοποιημένα εκπαιδευτικά δεδομένα. Αναλυτικά:

Πως λειτουργεί:

- Το μοντέλο εκπαιδεύεται με ένα σετ εκπαίδευσης που περιέχει τα εισερχόμενα δεδομένα (ανεξάρτητες μεταβλητές) και τα αντίστοιχα αποτελέσματα (εξαρτημένες μεταβλητές).
- Στόχος είναι το μοντέλο να αποκτήσει την ικανότητα να προβλέπει την εξαρτημένη μεταβλητή όταν δίνεται μια νέα ανεξάρτητη μεταβλητή.
- Μετά την εκπαίδευση, το μοντέλο δοκιμάζεται με ένα σετ δοκιμών που δεν έχει χρησιμοποιηθεί κατά την εκπαίδευση.

Τύποι επιβλεπόμενης μηχανικής μάθησης:

- **Παλινδρόμηση:** Ο στόχος είναι να προβλέψει έναν συνεχή αριθμό (π.χ., την τιμή ενός σπιτιού με βάση χαρακτηριστικά όπως η τοποθεσία, το μέγεθος, η ηλικία κτλ.).
- **Ταξινόμηση:** Ο στόχος είναι να προβλέψει την κατηγορία ενός σημείου δεδομένων (π.χ., ένα email είναι spam ή όχι).

Πού χρησιμοποιείται:

- **Ανάλυση αισθημάτων:** Η επιβλεπόμενη μάθηση μπορεί να χρησιμοποιηθεί για να κατανοήσει αν το συναίσθημα που εκφράζεται σε ένα κείμενο είναι θετικό, αρνητικό ή ουδέτερο.
- **Αναγνώριση χειρόγραφου:** Η επιβλεπόμενη μάθηση μπορεί να εφαρμοστεί για να αναγνωρίσει χειρόγραφα κείμενα ή αριθμούς.
- **Πρόβλεψη ασθενειών:** Η επιβλεπόμενη μάθηση χρησιμοποιείται επίσης στην ιατρική για την πρόβλεψη της εμφάνισης ασθενειών με βάση διάφορους παράγοντες ρίσκου.
- **Πρόβλεψη τιμών ακινήτων:** Με την επιβλεπόμενη μάθηση, ένα μοντέλο μπορεί να εκπαιδευτεί να προβλέπει την τιμή ενός ακινήτου βάσει χαρακτηριστικών όπως τοποθεσία, μέγεθος, ηλικία κλπ.

2.4.2 Μη-Εποπτευόμενη/Μη-Επιβλεπόμενη Μηχανική Μάθηση

Η μη-επιβλεπόμενη μηχανική μάθηση είναι ένας τύπος μηχανικής μάθησης όπου το μοντέλο προσπαθεί να εξάγει συμπεράσματα από δεδομένα που δεν έχουν προηγουμένως ετικετοποιηθεί ή κατηγοριοποιηθεί. Αναλυτικά:

Πως λειτουργεί:

- Οι αλγόριθμοι μη-επιβλεπόμενης μηχανικής μάθησης προσπαθούν να εντοπίσουν ομοιότητες, πρότυπα ή κλίμακες στα δεδομένα.

- Αντί να προβλέπουν μια εξαρτημένη μεταβλητή, όπως στην επιβλεπόμενη μάθηση, ο στόχος της μη-επιβλεπόμενης μάθησης είναι να αναλύσει τη δομή και τις σχέσεις εντός των δεδομένων.

Τύποι μη-επιβλεπόμενης μηχανικής μάθησης:

- **Ομαδοποίηση:** Ο στόχος είναι να ομαδοποιήσει τα δεδομένα σε διάφορες ομάδες ή clusters με βάση την ομοιότητα (π.χ., k-means, hierarchical clustering).
- **Μείωση διάστασης:** Ο στόχος είναι να μειώσει τον αριθμό των χαρακτηριστικών στα δεδομένα ενώ παράλληλα διατηρείται όσο το δυνατόν περισσότερη πληροφορία (π.χ., Principal Component Analysis (PCA), t-SNE).

Πού χρησιμοποιείται:

- **Συστάσεις:** Η μη-επιβλεπόμενη μάθηση χρησιμοποιείται για τη δημιουργία συστημάτων συστάσεων που προτείνουν προϊόντα ή υπηρεσίες με βάση τα προηγούμενα πρότυπα αγορών ή προτιμήσεων των χρηστών.
- **Ανάλυση κοινωνικών δικτύων:** Η μη-επιβλεπόμενη μάθηση μπορεί να χρησιμοποιηθεί για την κατανόηση των συνδέσεων και της δομής των κοινωνικών δικτύων.
- **Ανακάλυψη γνώσης σε βάσεις δεδομένων (KDD):** Η μη-επιβλεπόμενη μάθηση μπορεί να χρησιμοποιηθεί για την ανακάλυψη νέων γνώσεων σε μεγάλες βάσεις δεδομένων, εντοπίζοντας κρυμμένα πρότυπα ή συσχετίσεις.
- **Αναγνώριση προτύπων:** Η μη-επιβλεπόμενη μάθηση επιτρέπει την αναγνώριση προτύπων σε δεδομένα, όπως η ανίχνευση απάτης ή η αναγνώριση ομάδων σε γενετικά δεδομένα.

2.4.3 Ημι-Εποπτευόμενη/Ημι-Επιβλεπόμενη Μηχανική Μάθηση

Η ημι-επιβλεπόμενη μηχανική μάθηση είναι ένας τύπος μηχανικής μάθησης που επικεντρώνεται στην εκμάθηση από ένα μεικτό σύνολο δεδομένων που περιλαμβάνει δεδομένα με ετικέτα και δεδομένα χωρίς ετικέτα. Αναλυτικά:

Πως λειτουργεί:

- Οι αλγόριθμοι ημι-επιβλεπόμενης μηχανικής μάθησης χρησιμοποιούν ένα μικρό υποσύνολο ετικετοποιημένων δεδομένων μαζί με ένα μεγάλο υποσύνολο μη ετικετοποιημένων δεδομένων για την εκπαίδευση του μοντέλου.

- Συνδυάζει τα πλεονεκτήματα και των δύο ειδών εκπαίδευσης (επιβλεπόμενης και μη-επιβλεπόμενης) για την επίτευξη καλύτερων αποτελεσμάτων.

Τύποι ημι-επιβλεπόμενης μηχανικής μάθησης:

- **Self-training:** Το μοντέλο εκπαιδεύεται αρχικά με τα ετικετοποιημένα δεδομένα και στη συνέχεια χρησιμοποιείται για την ετικέτα σε μη ετικετοποιημένα δεδομένα. Τα νέα ετικετοποιημένα δεδομένα προστίθενται στο σύνολο εκπαίδευσης.

- **Multi-view training:** Εάν τα δεδομένα μπορούν να προβληθούν από διάφορες απόψεις ή για διάφορες "απόψεις" των δεδομένων, το μοντέλο μπορεί να εκπαιδευτεί για κάθε προβολή και έπειτα να συνδυαστεί.

Πού χρησιμοποιείται:

- **Βιολογική επιστήμη και γενετική**: Η ημι-επιβλεπόμενη μάθηση βρίσκει εφαρμογές σε βιοπληροφορική, για παράδειγμα, στην πρόβλεψη της δομής των πρωτεϊνών.

- **Αναγνώριση ανθρώπινης δραστηριότητας**: Χρησιμοποιείται για την αναγνώριση δραστηριότητας σε βίντεο ή σειρές δεδομένων αισθητήρων, όπως τα δεδομένα από τα επιταχυνσιόμετρα στα smartphones.

- **Φωτογραφική ανάλυση**: Χρησιμοποιείται για την ανάλυση και την εξαγωγή πληροφοριών από εικόνες ή βίντεο, όπως η αναγνώριση προσώπων ή ο συνδυασμός εικόνων από διάφορες πηγές.

- **Διάκριση περιεχομένου διαδικτύου**: Χρησιμοποιείται για την αναγνώριση και την κατηγοριοποίηση του περιεχομένου του διαδικτύου, για παράδειγμα, για τη διάκριση των ιστοσελίδων ή των σχολίων ανάμεσα σε χρήσιμα και ενοχλητικά.

Τα πλεονεκτήματα και τα μειονεκτήματα της ημι-επιβλεπόμενης μηχανικής μάθησης είναι παρόμοια με αυτά της επιβλεπόμενης και μη επιβλεπόμενης, με εξαίρεση το γεγονός ότι είναι σε θέση να εκμεταλλευτεί μεγάλο αριθμό μη ετικετοποιημένων δεδομένων, κάτι που μπορεί να είναι ιδιαίτερα χρήσιμο σε περιπτώσεις όπου η ετικέτα των δεδομένων είναι δύσκολη ή δαπανηρή.

2.4.4 Ενισχυτική Μηχανική Μάθηση

Η ενισχυτική μηχανική μάθηση είναι μια τεχνική μηχανικής μάθησης όπου ένας πράκτορας (αλγόριθμος ή πρόγραμμα που λαμβάνει αποφάσεις ή ενέργειες μέσα σε ένα περιβάλλον με σκοπό τη μακροπρόθεσμη μεγιστοποίηση κάποιας μορφής ανταμοιβής) αναλαμβάνει ενέργειες σε ένα περιβάλλον έτσι ώστε να μεγιστοποιήσει κάποια ανταμοιβή. Αναλυτικά:

Πως λειτουργεί:

- Ένας πράκτορας παίρνει αποφάσεις βασισμένες στην κατάσταση του περιβάλλοντος. Η απόφαση ή η ενέργεια επηρεάζει την κατάσταση του περιβάλλοντος και ανάλογα με την αλλαγή, ο πράκτορας λαμβάνει μια ανταμοιβή ή μια ποινή.

- Ο στόχος του πράκτορα είναι να μάθει μια στρατηγική, γνωστή ως πολιτική, η οποία θα του επιτρέπει να παίρνει αποφάσεις που θα μεγιστοποιούν την συνολική ανταμοιβή.

Τύποι ενισχυτικής μηχανικής μάθησης:

- **Model-Free Reinforcement Learning:** Ο πράκτορας μαθαίνει να λαμβάνει αποφάσεις βασισμένες αποκλειστικά στην εμπειρία του, χωρίς να προσπαθεί να κατανοήσει το περιβάλλον του. Παραδείγματα περιλαμβάνουν τις τεχνικές Q-learning και SARSA.
- **Model-Based Reinforcement Learning:** Ο πράκτορας προσπαθεί να κατανοήσει και να μοντελοποιήσει το περιβάλλον του πριν πάρει αποφάσεις.

Πού χρησιμοποιείται:

- **Παιχνίδια:** Η ενισχυτική μάθηση έχει χρησιμοποιηθεί για την ανάπτυξη προγραμμάτων υπολογιστών που μπορούν να ανταγωνιστούν ανθρώπους σε παιχνίδια όπως το Go και το Chess.
- **Αυτόνομα οχήματα:** Χρησιμοποιείται για την εκπαίδευση των συστημάτων πλοήγησης των αυτόνομων οχημάτων.
- **Ρομποτική:** Χρησιμοποιείται για την εκπαίδευση των ρομπότ σε διάφορες εργασίες, συμπεριλαμβανομένης της χειραφέτησης, της εξερεύνησης και της αλληλεπίδρασης με το περιβάλλον.
- **Συστήματα συστάσεων:** Χρησιμοποιείται για την εκπαίδευση των συστημάτων που συστήνουν προϊόντα ή υπηρεσίες στους χρήστες, με βάση την προηγούμενη συμπεριφορά τους.

Το κύριο πλεονέκτημα της ενισχυτικής μάθησης είναι η ικανότητά της να λαμβάνει αποφάσεις με βάση τον προαναφερθέντα σκοπό (maximizing reward), κάτι που την καθιστά ιδιαίτερα επωφελή για προβλήματα όπως η βελτιστοποίηση της πλοήγησης ή η ανάπτυξη προηγμένων συστημάτων για παιχνίδια. Εντούτοις, το μεγαλύτερο μειονέκτημα της ενισχυτικής μάθησης είναι η δυσκολία στην εκπαίδευση και στη στοχοθέτηση, καθώς οι πράκτορες μπορεί να είναι σε θέση να "εξερευνούν" τρόπους για να μεγιστοποιήσουν την ανταμοιβή τους που δεν αντικατοπτρίζουν την επιθυμητή συμπεριφορά.

2.4.5 Εξελικτική Μηχανική Μάθηση

Η εξελικτική μηχανική μάθηση είναι μια κατηγορία αλγορίθμων μηχανικής μάθησης που εμπνέονται από τις θεωρίες της βιολογικής εξέλιξης, όπως η φυσική επιλογή και η γενετική παραλλαγή. Αναλυτικά:

Πως λειτουργεί:

- Οι εξελικτικοί αλγόριθμοι ξεκινούν με ένα αρχικό πληθυσμό λύσεων (ονομάζονται επίσης χρωμοσώματα ή άτομα) για ένα πρόβλημα.
- Κάθε λύση αξιολογείται με βάση μια συνάρτηση καταλληλότητας, η οποία μετράει πόσο καλά λύνει το πρόβλημα.

- Οι καλύτερες λύσεις "αναπαράγονται" μέσω ενός διαδικασίας που απομιμείται την γενετική διασταύρωση και μετάλλαξη, δημιουργώντας μια νέα γενιά λύσεων.
- Αυτή η διαδικασία επαναλαμβάνεται για πολλές γενεές, με την ελπίδα ότι ο πληθυσμός θα "εξελιχθεί" προς καλύτερες λύσεις στο πρόβλημα.

Τύποι εξελικτικής μηχανικής μάθησης:

- **Γενετικοί Αλγόριθμοι:** Αυτοί οι αλγόριθμοι χρησιμοποιούν τεχνικές αναπαραγωγής όπως η διασταύρωση (παίρνει μέρη δύο χρωμοσωμάτων και τα συνδυάζει) και η μετάλλαξη (τυχαία αλλάζει μέρη ενός χρωμοσώματος) για να δημιουργήσει νέα χρωμοσώματα.
- **Στρατηγικές Εξέλιξης:** Χρησιμοποιούν μηχανισμούς προσαρμογής για να διαμορφώνουν τη στρατηγική αναπαραγωγής και μετάλλαξης κατά τη διάρκεια της εξέλιξης.

Πού χρησιμοποιείται:

- **Βελτιστοποίηση:** Εξελικτικοί αλγόριθμοι χρησιμοποιούνται για να βρουν την καλύτερη λύση σε προβλήματα βελτιστοποίησης, όπως το πρόβλημα του πωλητή (Travelling Salesman Problem).
- **Δημιουργία Νευρωνικών Δικτύων:** Μπορούν να χρησιμοποιηθούν για να δημιουργήσουν και να βελτιώσουν τα νευρωνικά δίκτυα, διαδικασία που είναι γνωστή ως νευρωνική εξέλιξη (neuroevolution).
- **Σχεδιασμός Αλγορίθμων:** Χρησιμοποιούνται για να "εξελίξουν" αλγόριθμους που λύνουν συγκεκριμένα προβλήματα, δημιουργώντας και συνδυάζοντας διάφορες δομές και τεχνικές.

Προκλήσεις:

- Χρειάζονται πολλοί υπολογισμοί και επομένως μπορεί να είναι αργοί σε πολύπλοκα προβλήματα.
- Μπορεί να χρειάζονται πολλές γενεές για να βρουν μια ικανοποιητική λύση.
- Μπορεί να καταλήξουν σε τοπικά μέγιστα αντί για το παγκόσμιο μέγιστο.

2.5 Τρόπος Λειτουργίας Τεχνικών Μηχανικής Μάθησης

Ο πιο συνηθισμένος τρόπος λειτουργίας των τεχνικών μηχανικής μάθησης συνήθως περιλαμβάνει τα ακόλουθα βήματα:

1. Συλλογή Δεδομένων:

- Αρχικά, συλλέγεται ένα σύνολο δεδομένων που θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου. Τα δεδομένα αυτά μπορούν να προέρχονται από διάφορες πηγές, όπως βάσεις δεδομένων, ιστοσελίδες, αισθητήρες ή ακόμα και χειρωνακτικές καταγραφές.

2. Επεξεργασία Δεδομένων:

- Τα ακατέργαστα δεδομένα συχνά χρειάζονται επεξεργασία για να γίνουν χρήσιμα. Αυτό μπορεί να περιλαμβάνει τον καθαρισμό από σφάλματα, τη μετατροπή σε κατάλληλες μορφές, την κανονικοποίηση ή την επιλογή σημαντικών χαρακτηριστικών.

3. Επιλογή Μοντέλου:

- Βάσει του προβλήματος που πρέπει να λυθεί, επιλέγεται ένα μοντέλο μηχανικής μάθησης. Υπάρχουν πολλά διαφορετικά μοντέλα, όπως τα δέντρα αποφάσεων, τα νευρωνικά δίκτυα ή οι γραμμικοί παλινδρομητές, ανάλογα με την φύση του προβλήματος.

4. Εκπαίδευση:

- Με τη χρήση των δεδομένων που έχουν συλλεχθεί και επεξεργαστεί, εκπαιδεύεται το μοντέλο ώστε να μάθει τις καλύτερες δυνατές συσχετίσεις μεταξύ εισόδου και εξόδου.

5. Αξιολόγηση:

-Αφού έχει εκπαιδευτεί το μοντέλο, αξιολογείται χρησιμοποιώντας δεδομένα που δεν έχει δει προηγουμένως, για να γίνει κατανοητό πόσο καλά λειτουργεί στην πράξη.

2.6 Σύγκριση μεθόδων

Με τις αρχές της εποπτευόμενης και χωρίς επίβλεψη μάθησης κατανοητές, οι διαφορές μεταξύ τους γίνονται πιο σαφείς.

Η βασική διαφορά ανάμεσα στις δύο προσεγγίσεις είναι ότι η εποπτευόμενη μάθηση χρησιμοποιεί δεδομένα με ετικέτες για την εκπαίδευση αλγορίθμων ταξινόμησης ή πρόβλεψης. Το μοντέλο, δέχεται δεδομένα, προσαρμόζεται και αλλάζει τον τρόπο με τον οποίο αξιολογεί τα διάφορα χαρακτηριστικά των δεδομένων, μέχρι να φτάσει στο επιθυμητό αποτέλεσμα. Η ακρίβεια των μοντέλων αυτών είναι υψηλή, αλλά απαιτείται ανθρώπινη παρέμβαση για την επεξεργασία των δεδομένων. Για παράδειγμα, ένα μοντέλο αυτού του είδους μπορεί να προβλέψει τους χρόνους πτήσης βάσει διαφόρων παραγόντων, αλλά αναγκαστικά πρέπει να επεμβαίνουν άνθρωποι για να ταξινομήσουν τα δεδομένα κατάλληλα.

Από την άλλη πλευρά, τα μοντέλα μάθησης χωρίς επίβλεψη λειτουργούν αυτόνομα, αναζητώντας και εντοπίζοντας δομές μέσα σε δεδομένα χωρίς ετικέτες. Η ανθρώπινη βοήθεια είναι απαραίτητη μόνο για την επικύρωση των εξόδων του μοντέλου. Για

παράδειγμα, όταν κάποιος αγοράζει έναν νέο φορητό υπολογιστή στο διαδίκτυο, ένα μοντέλο αυτού του είδους θα κατανοήσει ότι αυτός ο χρήστης ανήκει σε μια ομάδα που αγοράζει σχετικά προϊόντα. Η επικύρωση αυτών των συμπερασμάτων, ωστόσο, απαιτεί την παρέμβαση ενός αναλυτή δεδομένων.

Ας δούμε πιο αναλυτικά τις διαφορές μεταξύ της εποπτευόμενης και της μη εποπτευόμενης μάθησης σε μερικές βασικές κατηγορίες:

Εποπτευόμενη μάθηση:

1. **Δεδομένα εκπαίδευσης:** Χρειάζεται δεδομένα εκπαίδευσης με ετικέτες, τα οποία αντιστοιχούν σε συγκεκριμένους στόχους ή ετικέτες εξόδου.
2. **Στόχος:** Αναζητά να μάθει μια λειτουργία που θα συνδέει τα χαρακτηριστικά εισόδου με τις ετικέτες εξόδου, βασισμένο στα επισημασμένα δεδομένα εκπαίδευσης.
3. **Εφαρμογές:** Εκτελεί εργασίες ταξινόμησης ή παλινδρόμησης.
4. **Παραδείγματα:** Διανυσματικές μηχανές (SVM), δέντρα αποφάσεων, τυχαία δάση, νευρωνικά δίκτυα.
5. **Εκτίμηση:** Η αξιολόγηση γίνεται συγκρίνοντας τις προβλεπόμενες ετικέτες με αυτές της βασικής αλήθειας, χρησιμοποιώντας μετρήσεις όπως η ακρίβεια, ανάκληση ή το μέσο τετράγωνο σφάλμα.

Μη εποπτευόμενη μάθηση:

1. **Δεδομένα εκπαίδευσης:** Αξιοποιεί δεδομένα εκπαίδευσης χωρίς ετικέτες, όπου οι ετικέτες εξόδου δεν είναι διαθέσιμες.
2. **Στόχος:** Στοχεύει στην ανακάλυψη μοτίβων, δομών ή σχέσεων μέσα στα δεδομένα, χωρίς την ανάγκη για ετικέτες.
3. **Εφαρμογές:** Εκτελεί εργασίες όπως η ομαδοποίηση, ανίχνευση ανωμαλιών, μείωση διαστάσεων και οπτικοποίηση δεδομένων.
4. **Παραδείγματα:** Ομαδοποίηση K-Means, Ιεραρχική ομαδοποίηση, Ανάλυση Κυρίων Συνιστωσών (PCA), Autoencoders.
5. **Εκτίμηση:** Η αξιολόγηση είναι υποκειμενική και βασίζεται στην χρησιμότητα και ερμηνευσιμότητα των εξαγόμενων μοτίβων ή δομών. Μπορεί να χρησιμοποιηθούν μετρήσεις όπως το σκορ σιλουέτας για την ομαδοποίηση.

Η επιλογή ανάμεσα στην εποπτευόμενη και μη εποπτευόμενη μάθηση εξαρτάται από πολλούς παράγοντες, όπως η διαθεσιμότητα επισημασμένων δεδομένων, ο τύπος της εργασίας που πρέπει να εκτελεστεί και τα αναμενόμενα αποτελέσματα.

Παρόλο που η εποπτευόμενη και η μη-εποπτευόμενη μηχανική μάθηση είναι οι δύο βασικές κατηγορίες της μηχανικής μάθησης αξίζει να αναφερθούν και κάποιες περισσότερες πληροφορίες και για τις υπόλοιπες κατηγορίες για την σύγκριση τους.

Ημι-Εποπτευόμενη Μάθηση:

1.Δεδομένα εκπαίδευσης: Χρησιμοποιείται μια συνδυαστική προσέγγιση δεδομένων: μεγάλος όγκος δεδομένων χωρίς ετικέτες σε συνδυασμό με ένα μικρότερο όγκο δεδομένων με ετικέτες.

2.Στόχος: Στοχεύει στο να εκμεταλλευτεί τις πληροφορίες από τα επισημασμένα δεδομένα, ενώ ταυτόχρονα ανακαλύπτει δομές και σχέσεις από τα μη επισημασμένα δεδομένα.

3.Εφαρμογές: Μπορεί να βοηθήσει στη βελτίωση της απόδοσης των μοντέλων ταξινόμησης ή παλινδρόμησης, ειδικά όταν τα διαθέσιμα επισημασμένα δεδομένα είναι περιορισμένα.

4.Παραδείγματα: Label Spreading, Label Propagation, Self-training, Multi-view learning.

5.Εκτίμηση: Καθώς χρησιμοποιείται ένα μίγμα επισημασμένων και μη-επισημασμένων δεδομένων, η αξιολόγηση γίνεται κυρίως βασιζόμενη στα επισημασμένα δεδομένα. Εκτιμώνται οι προβλεπόμενες ετικέτες σε σύγκριση με τις πραγματικές ετικέτες του επισημασμένου υποσυνόλου, χρησιμοποιώντας μετρήσεις όπως η ακρίβεια, ανάκληση ή το μέσο τετράγωνο σφάλμα.

Εξελικτική Μηχανική Μάθηση:

1. Δεδομένα εκπαίδευσης: Ενώ τα δεδομένα εκπαίδευσης μπορούν να είναι επισημασμένα ή όχι, η εξελικτική μηχανική μάθηση δίνει έμφαση στην εύρεση βέλτιστων λύσεων μέσω διαδικασιών που μιμούνται την φυσική εξέλιξη.

2.Στόχος: Στοχεύει στην ανακάλυψη των καλύτερων μοντέλων ή παραμέτρων μέσω διαδικασιών επιλογής, διασταύρωσης, μετάλλαξης και επιβίωσης, βασισμένων σε μια αξιολογική συνάρτηση.

3.Εφαρμογές: Χρησιμοποιείται για τη βελτιστοποίηση παραμέτρων, την επιλογή χαρακτηριστικών, τη δημιουργία νέων μοντέλων και την εξερεύνηση του χώρου λύσεων.

4.Παραδείγματα: Γενετικοί αλγόριθμοι, γενετικός προγραμματισμός, στρατηγικές εξέλιξης, διαφορική εξέλιξη.

5.Εκτίμηση: Η αξιολόγηση βασίζεται στην αξιολογητική συνάρτηση, η οποία ορίζει το πόσο "καλή" είναι μια λύση ή ένα μοντέλο. Αυτή η συνάρτηση μπορεί να είναι οποιοδήποτε κριτήριο απόδοσης, όπως το σφάλμα πρόβλεψης, η ακρίβεια ταξινόμησης ή κάποιος άλλος δείκτης που είναι σχετικός με το ειδικό πρόβλημα που αντιμετωπίζουμε.

Ενισχυτική Μηχανική Μάθηση:

1.Δεδομένα Εκπαίδευσης: Στην ενισχυτική μάθηση, τα "δεδομένα" συχνά παράγονται δυναμικά μέσω της αλληλεπίδρασης του πράκτορα με ένα περιβάλλον. Ο πράκτορας λαμβάνει ανατροφοδότηση μέσω ενισχύσεων, αντί για ετικέτες εξόδου.

2.Στόχος:Στοχεύει στη μακροχρόνια μεγιστοποίηση του αναμενόμενου αθροίσματος ενισχύσεων, με τον πράκτορα να λαμβάνει αποφάσεις που βελτιώνουν τη συνολική του απόδοση στο περιβάλλον.

3.Εφαρμογές: Χρησιμοποιείται σε προβλήματα όπου η απόφαση πρέπει να ληφθεί με βάση την αλληλεπίδραση με ένα περιβάλλον, όπως στα παιχνίδια, την ρομποτική, την οπτικοποίηση και τις συστάσεις.

4.Παραδείγματα: Q-learning, Deep Q Networks (DQN), Proximal Policy Optimization (PPO), Actor-Critic μέθοδοι.

5.Εκτίμηση: Η αξιολόγηση βασίζεται συχνά στην απόδοση του πράκτορα σε εναλλακτικές στρατηγικές ή την ικανότητα του πράκτορα να μεγιστοποιεί την ενίσχυση στο δεδομένο περιβάλλον. Μετρήσεις όπως η μέση ενίσχυση ανά επεισόδιο ή ο χρόνος μάθησης μπορούν να χρησιμοποιηθούν για την αξιολόγηση.

2.7 Προβλήματα

Η μηχανική μάθηση αποτελεί έναν κλάδο της τεχνητής νοημοσύνης που ασχολείται με την κατασκευή αλγορίθμων που μπορούν να "μαθαίνουν" και να βελτιώνουν την απόδοσή τους με την εμπειρία. Οι αλγόριθμοι αυτοί εκπαιδεύονται με τη χρήση δεδομένων, για να δημιουργήσουν ένα μοντέλο που μπορεί να κάνει προβλέψεις ή να πάρει αποφάσεις χωρίς να είναι ρητά προγραμματισμένο για να εκτελεί την εν λόγω εργασία.

Η εφαρμογή της μηχανικής μάθησης έχει πληθώρα εφαρμογών και μπορεί να διευκολύνει σημαντικά την ανάλυση και την επεξεργασία μεγάλων συνόλων δεδομένων. Παρά τα πολλά πλεονεκτήματα που προσφέρει, η χρήση της μηχανικής μάθησης συνοδεύεται από πολλές προκλήσεις και προβλήματα. Ας εξετάσουμε κάποια από τα πιο συνηθισμένα προβλήματα που αντιμετωπίζουμε στη μηχανική μάθηση: ταξινόμηση, πρόβλεψη χρονοσειρών και ανίχνευση ανωμαλιών.

1. **Ταξινόμηση**: Τα προβλήματα ταξινόμησης είναι μια κατηγορία εποπτευόμενης μάθησης, που στόχο έχει να προβλέψουμε την κατηγορία ή την τάξη ενός δείγματος με βάση τα χαρακτηριστικά του. Για παράδειγμα, ένα σύστημα ταξινόμησης μπορεί να προβλέπει εάν ένα email είναι spam ή όχι, ή αν μια εικόνα περιέχει ένα

συγκεκριμένο αντικείμενο. Τα προβλήματα αυτά μπορεί να είναι δυαδικά (δύο κλάσεις) ή πολυταξικά (τρεις ή περισσότερες κλάσεις).

2. **Πρόβλεψη Χρονοσειρών**: Η πρόβλεψη χρονοσειρών αναφέρεται στην ανάλυση μοντέλων δεδομένων ή δομών σε χρονοσειρές για την πρόβλεψη μελλοντικών τιμών. Για παράδειγμα, μπορεί να χρησιμοποιηθεί για την πρόβλεψη των τιμών των μετοχών, της ζήτησης προϊόντων, ή των μετρήσεων του καιρού. Αυτή η προσέγγιση χρησιμοποιεί μεθόδους όπως η αυτοπαλίνδρομη ολισθητική μέση τιμή (ARIMA), τα νευρωνικά δίκτυα και τα μοντέλα μηχανικής μάθησης για να αναγνωρίσουν μοτίβα και τάσεις στα δεδομένα.
3. **Ανίχνευση Ανωμαλιών**: Η ανίχνευση ανωμαλιών αναφέρεται στην εύρεση παρατηρήσεων που δεν ταιριάζουν στα συνήθη μοτίβα ή συμπεριφορά σε ένα σύνολο δεδομένων. Αυτές οι ανωμαλίες μπορεί να οφείλονται σε διάφορους λόγους, όπως σφάλματα μέτρησης, εξαίρεση στην πολιτική, ή απάτη. Για παράδειγμα, η ανίχνευση ανωμαλιών μπορεί να χρησιμοποιηθεί για την εντοπισμός δραστηριοτήτων απάτης στις συναλλαγές με πιστωτικές κάρτες. Οι αλγόριθμοι ανίχνευσης ανωμαλιών μπορεί να είναι εποπτευόμενοι (εάν έχουμε ετικετοποιημένα δεδομένα ανωμαλίας) ή μη εποπτευόμενοι (εάν δεν έχουμε ετικετοποιημένα δεδομένα).

Κεφάλαιο 3: Ανίχνευση Ανωμαλιών

3.1 Τι είναι ανίχνευση ανωμαλιών

Η ανίχνευση ανωμαλιών στη μηχανική μάθηση αναφέρεται στη διαδικασία εντοπισμού σπάνιων ή ασυνήθιστων περιπτώσεων δεδομένων, μοτίβων ή συμπεριφορών που αποκλίνουν σημαντικά από τον κανόνα ή την αναμενόμενη συμπεριφορά σε ένα δεδομένο σύνολο δεδομένων [16]. Οι ανωμαλίες, γνωστές και ως ακραίες τιμές, μπορεί να είναι ενδεικτικές ανώμαλων ή ύποπτων συμβάντων, σφαλμάτων ή ανωμαλιών στα δεδομένα [17].

Ο στόχος της ανίχνευσης ανωμαλιών είναι η αυτόματη διάκριση μεταξύ κανονικών σημείων δεδομένων και ανωμαλιών, χωρίς να απαιτούνται σαφείς ετικέτες ή προηγούμενη γνώση των ανωμαλιών. Είναι συνήθως μια εργασία μάθησης χωρίς επίβλεψη, καθώς η πλειοψηφία των δεδομένων θεωρείται φυσιολογική και οι ανωμαλίες είναι σπάνιες και συχνά άγνωστες[18].

Οι αλγόριθμοι ανίχνευσης ανωμαλιών χρησιμοποιούν διάφορες τεχνικές και μοντέλα για τον εντοπισμό ανωμαλιών με βάση τα εγγενή χαρακτηριστικά των δεδομένων. Μερικές κοινές προσεγγίσεις περιλαμβάνουν:

- Στατιστικές μέθοδοι: Στατιστικές τεχνικές όπως οι βαθμολογίες z [19], η μοντελοποίηση κατανομής Gauss[20] ή ο έλεγχος υποθέσεων [21] μπορούν να χρησιμοποιηθούν για τον εντοπισμό σημείων δεδομένων που αποκλίνουν σημαντικά από τις στατιστικές ιδιότητες των υπολοίπων δεδομένων.
- Μέθοδοι με βάση την απόσταση: Αυτές οι μέθοδοι μετρούν την απόσταση ή την ανομοιότητα μεταξύ των σημείων δεδομένων και χρησιμοποιούν ένα όριο για τον προσδιορισμό των ανωμαλιών. Παραδείγματα περιλαμβάνουν k-πλησιέστερους γείτονες [22], απόσταση Mahalanobis[23] ή αλγορίθμους ομαδοποίησης με βάση την πυκνότητα όπως το DBSCAN [24].
- Μέθοδοι που βασίζονται στη μηχανική μάθηση: Οι εποπτευόμενοι και μη εποπτευόμενοι αλγόριθμοι μηχανικής μάθησης μπορούν να εφαρμοστούν για τον εντοπισμό ανωμαλιών. Οι μη εποπτευόμενες τεχνικές, όπως τα δάση απομόνωσης, ο τοπικός παράγοντας ακραίας θέσης LOF (Local Outlier Factor) ή οι μηχανές διανυσμάτων υποστήριξης μιας κατηγορίας (SVM), μαθαίνουν μοτίβα κανονικών δεδομένων και εντοπίζουν περιπτώσεις που δεν εμπίπτουν σε αυτά τα μοτίβα ως ανωμαλίες. Οι εποπτευόμενες τεχνικές απαιτούν δεδομένα ανωμαλιών με ετικέτα για εκπαίδευση και μπορούν να περιλαμβάνουν αλγόριθμους όπως δέντρα αποφάσεων, τυχαία δάση ή μοντέλα βαθιάς μάθησης, όπως αυτοκωδικοποιητές.
- Ανίχνευση ανωμαλιών χρονοσειρών: Τα δεδομένα χρονοσειρών συχνά εμφανίζουν χρονικά μοτίβα και η ανίχνευση ανωμαλιών σε αυτό το πλαίσιο περιλαμβάνει τον εντοπισμό αποκλίσεων από τα αναμενόμενα μοτίβα με την πάροδο του χρόνου. Μπορούν να χρησιμοποιηθούν τεχνικές όπως κινούμενοι μέσοι όροι.
- Πολυεπίπεδη ανίχνευση ανωμαλιών: Σε μερικές περιπτώσεις, τα δεδομένα που πρέπει να εξεταστούν είναι σύνθετα και πολυδιάστατα, πράγμα που απαιτεί την

εφαρμογή πολλών τεχνικών ανίχνευσης ανωμαλιών σε διάφορα επίπεδα ή διαστάσεις των δεδομένων.

Η επιλογή της μεθόδου ανίχνευσης ανωμαλιών εξαρτάται από τα χαρακτηριστικά των δεδομένων, τη φύση των στοχευμένων ανωμαλιών, τους διαθέσιμους υπολογιστικούς πόρους και τις ειδικές απαιτήσεις της εφαρμογής.

Η ανίχνευση ανωμαλιών έχει διάφορες εφαρμογές σε τομείς, όπως ανίχνευση απάτης, ανίχνευση εισβολής στο δίκτυο, παρακολούθηση συστήματος, ποιοτικός έλεγχος, ανάλυση δεδομένων αισθητήρων και ασφάλεια στον κυβερνοχώρο, μεταξύ άλλων. Διαδραματίζει κρίσιμο ρόλο στον εντοπισμό ασυνήθιστων γεγονότων ή προτύπων που μπορεί να απαιτούν περαιτέρω διερεύνηση ή δράση.

3.2 Αλγόριθμοι Ανίχνευσης Ανωμαλιών

Κάθε αλγόριθμος έχει τα δικά του δυνατά και αδύνατα σημεία και η επιλογή του αλγορίθμου εξαρτάται από τα συγκεκριμένα χαρακτηριστικά των δεδομένων και τη φύση των ανωμαλιών που στοχεύονται. Συχνά συνιστάται η δοκιμή πολλών αλγορίθμων και η σύγκριση της απόδοσής τους σε ένα δεδομένο σύνολο δεδομένων για να προσδιοριστεί η καταλληλότερη προσέγγιση για μια συγκεκριμένη εργασία ανίχνευσης ανωμαλιών.

3.2.1 SVM One-Class (Support Vector Machines)

Οι Μηχανές Διανυσμάτων Υποστήριξης (SVMs) είναι ένας δημοφιλής και ισχυρός τύπος αλγορίθμου μηχανικής μάθησης, ο οποίος γενικά χρησιμοποιείται για ταξινόμηση και παλινδρόμηση. Ωστόσο, μια ειδική παραλλαγή των SVM, γνωστή ως One-Class SVM, χρησιμοποιείται για την ανίχνευση ανωμαλιών.

Ο αλγόριθμος One-Class SVM έχει ως στόχο να μοντελοποιήσει τα "κανονικά" δεδομένα με τον τρόπο που εγκλωβίζει τα περισσότερα σημεία δεδομένων σε μια υπερ-σφαίρα σε υψηλοδιάστατο χώρο. Εδώ, η "κανονικότητα" εκφράζεται ως η εγγύτητα σε αυτήν την υπερ-σφαίρα.

Όταν έρχεται ένα νέο σημείο δεδομένων, το μοντέλο ελέγχει αν βρίσκεται μέσα ή έξω από την υπερ-σφαίρα. Αν βρίσκεται εκτός της υπερ-σφαίρας, τότε το σημείο δεδομένων καταγράφεται ως ανωμαλία. Αυτό επειδή το σημείο δεδομένων δεν ταιριάζει με το "κανονικό" μοτίβο που έχει μάθει το μοντέλο.

Αξίζει να σημειωθεί ότι οι One-Class SVMs είναι ειδικά χρήσιμοι σε περιπτώσεις όπου έχουμε μόνο κανονικά δεδομένα για εκπαίδευση (δηλαδή, μη εποπτευόμενη μάθηση) και δεν γνωρίζουμε πώς μπορεί να φαίνεται μια ανωμαλία. Ο αλγόριθμος εκπαιδεύεται να αναγνωρίζει μόνο την "κανονικότητα" και όλα όσα αποκλίνουν από αυτήν καταγράφονται ως ανωμαλίες.

Ακολουθούν τα βασικά βήματα της διαδικασίας που ακολουθεί ο One-class SVM για τον εντοπισμό ανωμαλιών:

1.Εκπαίδευση μόνο με "Φυσιολογικά" Δεδομένα: Συνήθως, ο One-class SVM εκπαιδεύεται μόνο με δεδομένα που θεωρούνται "φυσιολογικά". Αυτό σημαίνει ότι δεν υπάρχει ανάγκη για δεδομένα ανωμαλιών κατά την φάση εκπαίδευσης.

2. Κατασκευή της Υπερ-σφαίρας: Χρησιμοποιείται ένα υπερεπίπεδο (υποπεριοχή μίας διάστασης λιγότερης από τον περιβάλλοντα χώρο) για να περιβάλλει τα φυσιολογικά δεδομένα. Η ιδέα είναι να βρεθεί μια υπερ-σφαίρα που διαχωρίζει τα δεδομένα από την προέλευση (ή από το πιο κοντινό outlier).

3.Διαφορά Πυρήνα: Ο One-class SVM, όπως και το τυπικό SVM, μπορεί να χρησιμοποιήσει διαφορετικούς πυρήνες (kernels) για τη μετατροπή των δεδομένων σε ένα χώρο υψηλότερων διαστάσεων, όπου τα δεδομένα μπορούν να γίνουν διαχωρίσιμα.

4.Υπολογισμός Τιμής Απόφασης: Όταν ένα νέο σημείο δεδομένων εξετάζεται, υπολογίζεται η τιμή απόφασης. Εάν η τιμή απόφασης είναι κάτω από ένα καθορισμένο κατώφλι, το σημείο θεωρείται ως ανωμαλία.

5. Ρύθμιση Παραμέτρων: Οι παράμετροι, όπως το ν και ο πυρήνας (kernel), πρέπει να ρυθμιστούν με βάση τα δεδομένα και τον επιθυμητό συμβιβασμό μεταξύ ανίχνευσης ανωμαλιών και λανθασμένης ανίχνευσης.

6. Αξιολόγηση: Επειδή ο One-class SVM είναι μη επιβλεπόμενος, η αξιολόγησή του μπορεί να είναι πιο προκλητική. Παρ' όλα αυτά, στην πράξη, μπορούν να χρησιμοποιηθούν δεδομένα που περιέχουν γνωστές ανωμαλίες για να αξιολογηθεί η απόδοση του μοντέλου.

Σε γενικές γραμμές, ο One-class SVM επιτυγχάνει τον εντοπισμό ανωμαλιών προσπαθώντας να "περιβάλλει" τα φυσιολογικά δεδομένα με μια υπερ-σφαίρα, αφήνοντας τις ανωμαλίες εκτός της υπερ-σφαίρας.

3.2.2 Isolation Forest

Ο αλγόριθμος Isolation Forest είναι ένας από τους πιο δημοφιλείς αλγορίθμους ανίχνευσης ανωμαλιών που χρησιμοποιούνται στην εποχή των μεγάλων δεδομένων (big data). Εφαρμόζει μια εντελώς νέα τεχνική ανίχνευσης ανωμαλιών που χρησιμοποιεί τη μηχανική μάθηση χωρίς επίβλεψη και διαφέρει από τις περισσότερες υπάρχουσες τεχνικές.

Πιο συγκεκριμένα, ο αλγόριθμος Isolation Forest δημιουργεί τυχαία δέντρα απομόνωσης (Isolation Trees) μέσα σε ένα δάσος (Forest). Κάθε δέντρο απομονώνει τα δεδομένα, επιλέγοντας τυχαία ένα χαρακτηριστικό και στη συνέχεια επιλέγοντας τυχαία μια τιμή διαχωρισμού μεταξύ των ελάχιστων και μέγιστων τιμών αυτού του χαρακτηριστικού. Η διαδικασία επαναλαμβάνεται μέχρι να απομονωθούν όλα τα δείγματα.

Το κλειδί της τεχνικής είναι ότι τα ανώμαλα σημεία δεδομένων είναι λιγότερο συχνά από τα κανονικά και συνήθως διαφέρουν περισσότερο σε τιμή. Συνεπώς, τα ανώμαλα σημεία

δεδομένων θα απομονώνονται πιο γρήγορα, το οποίο οδηγεί σε δέντρα μικρότερου βάθους. Επομένως, τα δέντρα που απομονώνονται γρήγορα πιθανότατα είναι πιο ανώμαλα.

Για να προβλέψει αν ένα νέο σημείο δεδομένων είναι ανώμαλο, το σημείο τροφοδοτείται σε κάθε δέντρο και εξάγεται η μέση διαδρομή (η μέση διάρκεια του μονοπατιού από τη ρίζα μέχρι το φύλλο). Οι μικρές μέσες διαδρομές υποδηλώνουν ανωμαλία.

Το Isolation Forest εντάσσεται στην κατηγορία των αλγορίθμων ανίχνευσης ανωμαλιών που βασίζονται στην απόσταση[10]. Στην περίπτωση του Isolation Forest, η ανίχνευση ανωμαλιών πραγματοποιείται μέσω της τεχνικής της απομόνωσης, η οποία εκτιμά το πόσο διαφορετικό είναι ένα σημείο δεδομένων σε σχέση με τα υπόλοιπα.

Επιπλέον, το Isolation Forest παρουσιάζει γραμμική χρονική πολυπλοκότητα, κάνοντάς το ιδιαίτερα κατάλληλο για μεγάλα σύνολα δεδομένων. Η συνολική προσέγγιση και η αποτελεσματικότητα του αλγορίθμου τον καθιστούν έναν από τους πιο δημοφιλείς στον τομέα της ανίχνευσης ανωμαλιών.[11]

Ακολουθούν τα βασικά βήματα της διαδικασίας που ακολουθεί το `Isolation Forest` για τον εντοπισμό ανωμαλιών:

1.Τυχαίο Δείγμα: Επιλέγεται ένα τυχαίο υποσύνολο των δεδομένων. Αυτό το βήμα γίνεται για να βελτιώσει την ταχύτητα του αλγορίθμου.

2.Δημιουργία Δένδρου: Στη συνέχεια, δημιουργείται ένα δένδρο απόφασης για το τυχαίο υποσύνολο. Σε κάθε κόμβο του δένδρου, επιλέγεται τυχαία μια διάσταση και ένα τυχαίο όριο, το οποίο χρησιμοποιείται για να διαχωρίσει τα δεδομένα.

3.Επανάληψη: Αυτή η διαδικασία επαναλαμβάνεται πολλές φορές, δημιουργώντας πολλά διαφορετικά δένδρα αποφάσεων, τα οποία μαζί συνθέτουν το "δάσος".

4.Απόσταση απομόνωσης: Για κάθε δείγμα στα δεδομένα, υπολογίζεται η μέση απόσταση που χρειάζεται για να απομονώσει το δείγμα. Αν ένα δείγμα είναι ανωμαλία, τότε συνήθως χρειάζεται λιγότερα βήματα για να απομονωθεί, σε σχέση με ένα "φυσιολογικό" δείγμα.

5. Υπολογισμός Σκορ: Με βάση την προηγούμενη απόσταση απομόνωσης, υπολογίζεται ένα σκορ για κάθε δείγμα. Τα σκορ που είναι κοντά στο 1 υποδεικνύουν ανωμαλίες, ενώ τα σκορ που είναι κοντά στο 0 υποδεικνύουν φυσιολογικά δείγματα.

Η βασική ιδέα του `Isolation Forest` είναι ότι οι ανωμαλίες μπορούν να απομονωθούν πιο εύκολα σε σχέση με τα φυσιολογικά δεδομένα. Κατά συνέπεια, τα δείγματα που απομονώνονται με λιγότερα βήματα στα δένδρα αποφάσεων θεωρούνται ως ανωμαλίες.

3.2.3 Local Outlier Factor (LOF)

Ο αλγόριθμος LOF είναι μια δημοφιλής τεχνική για την ανίχνευση ανωμαλιών που χρησιμοποιεί την τοπική πυκνότητα των δεδομένων. Αντί να υπολογίζει μια παγκόσμια πυκνότητα, όπως πολλοί άλλοι αλγόριθμοι ανίχνευσης ανωμαλιών, ο LOF υπολογίζει την τοπική πυκνότητα, γεγονός που τον καθιστά πιο ευέλικτο στην αντιμετώπιση δεδομένων με διάφορες πυκνότητες.

Ο αλγόριθμος LOF λειτουργεί υπολογίζοντας την "τοπική πυκνότητα" ενός σημείου δεδομένων σε σύγκριση με την πυκνότητα των γειτονικών σημείων. Η "τοπική πυκνότητα" ενός σημείου ορίζεται ως αντίστροφο του μέσου αριθμού των γειτόνων που απαιτούνται για να φτάσουν στο σημείο.

Ο LOF ενός σημείου δεδομένων στη συνέχεια υπολογίζεται ως ο λόγος της τοπικής πυκνότητας του σημείου σε σχέση με την τοπική πυκνότητα των γειτόνων του. Ένα LOF μεγαλύτερο από 1 υποδηλώνει ότι το σημείο έχει σημαντικά χαμηλότερη πυκνότητα από τους γείτονές του και, επομένως, θεωρείται ανώμαλο.

Ο αλγόριθμος LOF έχει το πλεονέκτημα ότι μπορεί να αναγνωρίσει τοπικές ανωμαλίες, δηλαδή ανωμαλίες που είναι ανώμαλες μόνο σε σύγκριση με την τοπική γειτονιά τους, ενώ μπορεί να μην θεωρούνται ανώμαλες αν εξεταστούν σε παγκόσμιο επίπεδο.

3.2.4 Αυτοκωδικοποιητές (Auto encoder)

Οι autoencoders αποτελούν τις θεμελιώδεις αρχιτεκτονικές μη-επιβλεπόμενης μάθησης που χρησιμοποιούνται στην ανίχνευση ανωμαλιών[8]. Είναι σχεδιασμένοι για να ανακατασκευάζουν την είσοδό τους μετά από μια διαδικασία συμπίεσης και αποσυμπίεσης των δεδομένων, και επομένως να μάθουν αποτελεσματικές αναπαραστάσεις των δεδομένων.

Τα βαθιά μη-επιβλεπόμενα μοντέλα που προτείνονται σε αυτή τη περίπτωση για την ανίχνευση ανωμαλιών βασίζονται σε μία από τις ακόλουθες υποθέσεις για τον εντοπισμό των ακραίων τιμών[Chalapathy,[7]:

• Οι "κανονικές" περιοχές στον αρχικό ή κρυφό χώρο χαρακτηριστικών μπορούν να διακριθούν από τις "ανωμαλικές" περιοχές στον αρχικό ή κρυφό χώρο χαρακτηριστικών.

• Η πλειοψηφία των περιπτώσεων δεδομένων είναι κανονική σε σύγκριση με το υπόλοιπο σύνολο δεδομένων.

• Ο μη-επιβλεπόμενος αλγόριθμος ανίχνευσης ανωμαλιών παράγει ένα σκορ ακραίων τιμών των περιπτώσεων δεδομένων βάσει των εγγενών ιδιοτήτων του συνόλου δεδομένων, όπως αποστάσεις ή πυκνότητες. Τα κρυφά επίπεδα του βαθέος νευρωνικού

δικτύου στοχεύουν στη σύλληψη αυτών των εγγενών ιδιοτήτων εντός του συνόλου δεδομένων [9].

Ένας αυτοκωδικοποιητής λοιπόν, αποτελείται από δύο βασικά μέρη: τον κωδικοποιητή και τον αποκωδικοποιητή.

1. Ο κωδικοποιητής μετατρέπει την είσοδο σε έναν κρυμμένο (ή κωδικοποιημένο) αναπαραστατικό χώρο. Η αρχιτεκτονική του κωδικοποιητή επιβάλλει την εύρεση ενός συμπυκνωμένου και ενημερωμένου αναπαραστατικού των δεδομένων εισόδου.
2. Ο αποκωδικοποιητής στη συνέχεια προσπαθεί να ανακατασκευάσει τα αρχικά δεδομένα από αυτήν την κωδικοποιημένη αναπαράσταση.

Οι αυτοκωδικοποιητές μπορούν να χρησιμοποιηθούν για την ανίχνευση ανωμαλιών διότι αντιμετωπίζουν δυσκολία στην ανακατασκευή των ανωμαλιών σε σύγκριση με τα "κανονικά" δεδομένα. Σε ένα κανονικής λειτουργίας αυτοκωδικοποιητή, τα "κανονικά" δεδομένα θα αναδιαμορφωθούν με μικρό σφάλμα, ενώ τα "ανώμαλα" δεδομένα θα έχουν μεγαλύτερα σφάλματα ανακατασκευής, καθώς το μοντέλο δεν έχει μάθει αποτελεσματικά πώς να τα αναπαράγει. Επομένως, οι περιπτώσεις με υψηλά σφάλματα ανακατασκευής μπορούν να επισημανθούν ως ανωμαλίες.

Σε αυτό το σημείο αξίζει να αναφέρουμε περιληπτικά την διαδικασία που ακολουθεί ο Autoencoder για την ανίχνευση ανωμαλιών:

1. Εκπαίδευση του Autoencoder: Ο Autoencoder εκπαιδεύεται χρησιμοποιώντας τα κανονικά δεδομένα ως είσοδο. Ο στόχος είναι να ανακαλυφθεί μια εσωτερική αναπαράσταση (latent representation) των δεδομένων που να περιέχει την βασική δομή και χαρακτηριστικά των κανονικών δειγμάτων.

2. Υπολογισμός των σφαλμάτων ανακατασκευής: Αφού εκπαιδευτεί ο Autoencoder, τα δεδομένα εισόδου περνούν ξανά από το μοντέλο και υπολογίζεται το σφάλμα ανακατασκευής μεταξύ των αρχικών δεδομένων και των ανακατασκευασμένων δεδομένων. Το σφάλμα ανακατασκευής αντιπροσωπεύει το μέτρο της απόκλισης των δεδομένων από το μοντέλο.

3. Καθορισμός ορίου ανωμαλιών: Με βάση τα σφάλματα ανακατασκευής των κανονικών δεδομένων, υπολογίζεται ένα κατώφλι (threshold) για τον καθορισμό των ανωμαλιών. Συνήθως, το κατώφλι ορίζεται ως ένα ποσοστό των υψηλότερων σφαλμάτων ανακατασκευής.

4. Ανίχνευση ανωμαλιών: Τα νέα δεδομένα περνούν από το εκπαιδευμένο Autoencoder και υπολογίζεται το σφάλμα ανακατασκευής. Εάν το σφάλμα υπερβαίνει το καθορισμένο κατώφλι, τότε το δείγμα θεωρείται ανώμαλο.

3.2.5 k-Κοντινότεροι γείτονες (k-NN)

Ο αλγόριθμος των k-Κοντινότερων Γειτόνων (k-NN) είναι ένας απλός αλλά ισχυρός αλγόριθμος που χρησιμοποιείται για τόσο εποπτευόμενη όσο και μη εποπτευόμενη

μάθηση. Στο πλαίσιο της ανίχνευσης ανωμαλιών, η χρήση του k-NN στηρίζεται στην αρχή ότι οι ανωμαλίες είναι μακριά από τα κανονικά δεδομένα.

Ένα βασικό τεχνικό k-NN για την ανίχνευση ανωμαλιών βασίζεται στον παρακάτω ορισμό: Ο βαθμός ανωμαλίας ενός στιγμιότυπου δεδομένων καθορίζεται από την απόστασή του στον k-οστού πλησιέστερο γείτονά του σε ένα δεδομένο σύνολο δεδομένων. Αυτή η βασική τεχνική έχει εφαρμοστεί για την ανίχνευση ναρκών από δορυφορικές εικόνες του εδάφους [2] και για την ανίχνευση σύντομων στροφών (ανωμαλιών) στις περιστροφές DC των μεγάλων συγχρονισμένων γεννητριών ανέμου [3], χρησιμοποιώντας k = 1. Συνήθως εφαρμόζεται ένα όριο στον βαθμό ανωμαλίας για να καθοριστεί αν ένα τεστ είναι ανώμαλο ή όχι. Οι Ramaswamy κλπ. [2000][4], από την άλλη πλευρά, επιλέγουν n στιγμιότυπα με τους μεγαλύτερους βαθμούς ανωμαλίας ως τις ανωμαλίες [1].

Στη συνέχεια, ο αλγόριθμος υπολογίζει την απόσταση ενός σημείου δεδομένων από τους k πλησιέστερους γείτονές του. Αν αυτή η απόσταση είναι σημαντικά μεγαλύτερη σε σύγκριση με τις αποστάσεις των υπόλοιπων σημείων από τους γείτονές τους, το σημείο αυτό μπορεί να θεωρηθεί ως ανωμαλία.

Είναι σημαντικό να σημειωθεί ότι ο αλγόριθμος k-NN εξαρτάται σημαντικά από την επιλογή της τιμής του k. Αν το k είναι πολύ μικρό, ο αλγόριθμος μπορεί να είναι υπερβολικά ευαίσθητος στις ανωμαλίες και μπορεί να παράγει πολλά ψευδώς θετικά αποτελέσματα. Αντίθετα, αν το k είναι πολύ μεγάλο, ο αλγόριθμος μπορεί να παράγει πολλά ψευδώς αρνητικά αποτελέσματα, δηλαδή να μην εντοπίζει πραγματικές ανωμαλίες.

Οι επιδόσεις του k-NN επίσης μπορούν να επηρεαστούν από την κλίμακα και τη φύση των χαρακτηριστικών των δεδομένων, εφόσον η μέτρηση της απόστασης επηρεάζεται από αυτά. Γι' αυτό είναι συχνά απαραίτητη η κανονικοποίηση των χαρακτηριστικών πριν από την εφαρμογή του k-NN.

Η διαδικασία που ακολουθεί ο classifier KNN (Κοντινότερων Γειτόνων) για την ανίχνευση ανωμαλιών περιλαμβάνει τα εξής βήματα:

1. Κατασκευή του μοντέλου: Ο classifier KNN κατασκευάζει ένα μοντέλο χρησιμοποιώντας το σύνολο δεδομένων εκπαίδευσης. Αυτό περιλαμβάνει την ανάθεση μιας τιμής για τον αριθμό των γειτόνων K και την απόσταση που θα χρησιμοποιηθεί για τον υπολογισμό της ομοιότητας μεταξύ των δεδομένων.

2. Υπολογισμός των αποστάσεων: Ο classifier KNN υπολογίζει τις αποστάσεις μεταξύ του κάθε δείγματος ελέγχου και των κοντινότερων γειτόνων του στο σύνολο δεδομένων εκπαίδευσης.

3. Καθορισμός του ορίου ανωμαλιών: Με βάση τις αποστάσεις που υπολογίστηκαν, ο classifier θεωρεί ως ανώμαλα τα δείγματα ελέγχου που έχουν απόσταση μεγαλύτερη από ένα προκαθορισμένο όριο.

4. Κατηγοριοποίηση των δειγμάτων ελέγχου: Τα δείγματα ελέγχου κατηγοριοποιούνται σε δύο κατηγορίες - κανονικά και ανώμαλα - ανάλογα με την απόστασή τους από τους γείτονές τους.

3.2.6 Gaussian Mixture Models (GMM)

Τα Gaussian Mixture Models (GMM) είναι μια πιθανολογική μέθοδος για την αναπαράσταση της κατανομής δεδομένων, που χρησιμοποιείται συχνά για την ανίχνευση ανωμαλιών. Τα GMM αποτελούνται από πολλαπλές κανονικές κατανομές (Gaussians) και μπορούν να κατανεμηθούν σε δεδομένα με πολλαπλές κορυφές.

Κάθε Gaussian σε ένα GMM αντιπροσωπεύει ένα συγκεκριμένο cluster των δεδομένων. Μετά την εκμάθηση του μοντέλου, κάθε Gaussian θα έχει μια κεντρική τιμή (μέσος όρος) και μια διασπορά, που αντιπροσωπεύει την εσωτερική διακύμανση των δεδομένων στο συγκεκριμένο cluster. Το "βάρος" κάθε Gaussian αντιπροσωπεύει το μέρος των συνολικών δεδομένων που προέρχονται από αυτό το cluster.

Η ανίχνευση ανωμαλιών με τη χρήση GMM βασίζεται στην εύρεση των δεδομένων που έχουν σχετικά χαμηλές πιθανότητες υπό το μοντέλο. Αυτό σημαίνει ότι οι ανωμαλίες είναι δεδομένα που δεν είναι καλά ερμηνευμένα από καμία από τις Gaussians στο μοντέλο. Σε πρακτικές εφαρμογές, οι ανωμαλίες συνήθως προσδιορίζονται ως τα δεδομένα που βρίσκονται πέρα από ένα συγκεκριμένο όριο πιθανότητας.

Το GMM είναι ένας ισχυρός αλγόριθμος, αλλά απαιτεί μια καλή εκτίμηση του αριθμού των Gaussians (ή clusters) στα δεδομένα. Αυτό μπορεί να είναι πρόκληση, ειδικά σε μεγάλα ή πολύπλοκα σύνολα δεδομένων.

3.2.7 Χωρική ομαδοποίηση εφαρμογών με θόρυβο βάσει πυκνότητας (DBSCAN)

Ο αλγόριθμος DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [6] είναι ένας μη-παραμετρικός αλγόριθμος ομαδοποίησης βασισμένος στην πυκνότητα. Αυτό σημαίνει ότι ο αλγόριθμος δεν κάνει εκ των προτέρων υποθέσεις σχετικά με τον αριθμό των συστάδων στα δεδομένα, αλλά ανακαλύπτει τις συστάδες βάσει της πυκνότητας των δεδομένων.

Ο DBSCAN λειτουργεί με τον εντοπισμό των περιοχών του χώρου δεδομένων όπου η πυκνότητα των σημείων δεδομένων είναι πολύ ψηλότερη από το περιβάλλον, δηλαδή, τα σημεία δεδομένων είναι πολύ κοντά ο ένας στον άλλον. Αυτές οι περιοχές υψηλής πυκνότητας θεωρούνται συστάδες. Τα σημεία δεδομένων που βρίσκονται σε περιοχές χαμηλής πυκνότητας θεωρούνται θόρυβος ή ανωμαλίες.

Ο DBSCAN έχει δύο βασικές παραμέτρους: την επιθυμητή ελάχιστη πυκνότητα σημείων σε μια συστάδα (MinPts) και την επιθυμητή ακτίνα εντός της οποίας πρέπει να υπάρχει

αυτή η πυκνότητα (ε). Αυτές οι παράμετροι χρησιμοποιούνται για τον προσδιορισμό των συστάδων και των θορύβων.

Αν και ο DBSCAN είναι ιδιαίτερα χρήσιμος για πολλά προβλήματα, η απόδοση του μπορεί να επηρεαστεί από την επιλογή των παραμέτρων ε και MinPts. Μια ακατάλληλη επιλογή των παραμέτρων αυτών μπορεί να οδηγήσει σε υπερβολική ή υποβολική ομαδοποίηση. Επιπλέον, ο DBSCAN μπορεί να αδυνατεί να αναγνωρίσει συστάδες διαφορετικών πυκνοτήτων.

Ενώ λειτουργεί αποτελεσματικά με την ανακάλυψη συστάδων βάσει πυκνότητας, η επιλογή των παραμέτρων γειτονιάς και αριθμού γειτόνων (συνήθως ορίζεται στο 4) είναι κρίσιμη για την αποτελεσματικότητα της ομαδοποίησης[5]. Τα σημεία που βρίσκονται σε περιοχές χαμηλής πυκνότητας καταγράφονται ως θόρυβος ή ανωμαλίες, ενώ οι περιοχές υψηλής πυκνότητας αναγνωρίζονται ως συστάδες.

3.2.8 Robust Random Cut Forest

Το Robust Random Cut Forest (RRCF) είναι ένας αλγόριθμος που ανιχνεύει ανωμαλίες σε ροές δεδομένων εκμεταλλευόμενος την ιδέα των τυχαίων δέντρων αποκοπής. Πρόκειται για έναν μη-παραμετρικό αλγόριθμο, που δηλαδή δεν κάνει υποθέσεις για την υποκείμενη κατανομή των δεδομένων. Αυτό τον καθιστά ιδιαίτερα χρήσιμο σε σενάρια όπου η κατανομή των δεδομένων είναι δύσκολο να καθοριστεί ή αλλάζει με τον χρόνο.

Το RRCF δημιουργεί ένα δάσος από τυχαία δέντρα αποκοπής, κάθε ένα από τα οποία εκπροσωπεί έναν διάνυσμα δεδομένων ως ένα διάνυσμα τυχαίων περικοπών. Στη συνέχεια, μετρά τον αριθμό των περικοπών που απαιτούνται για να απομονωθεί κάθε σημείο. Τα σημεία που απαιτούν λιγότερες περικοπές για να απομονωθούν θεωρούνται πιθανές ανωμαλίες, καθώς αυτά είναι διαφορετικά από την "κανονική" συμπεριφορά που αντιπροσωπεύεται από τη μεγάλη πλειοψηφία των σημείων.

Ένα από τα πλεονεκτήματα του RRCF είναι ότι μπορεί να χειριστεί μεγάλο αριθμό δεδομένων και διαστάσεων με αποδοτικό τρόπο. Αυτό το καθιστά ιδανικό για εφαρμογές πραγματικού χρόνου, όπως η ανίχνευση ανωμαλιών σε ροές δεδομένων. Επιπλέον, ο RRCF είναι ανθεκτικός σε θόρυβο και αποτελεσματικός στην αναγνώριση ανωμαλιών που εμφανίζονται σπάνια ή είναι νέες (δηλαδή δεν έχουν εμφανιστεί προηγουμένως στα δεδομένα).

3.2.9 Απόσταση Mahalanobis

Η απόσταση Mahalanobis είναι μια στατιστική μέτρηση που χρησιμοποιείται για να ποσοτικοποιήσει την απόσταση μεταξύ ενός σημείου και μιας διανυσματικής ομάδας δεδομένων. Είναι ένας πολυδιάστατος δείκτης που λαμβάνει υπόψη τον τρόπο με τον οποίο τα δεδομένα είναι διασπαρμένα ή συσχετισμένα σε πολλές διαστάσεις. Το

σημαντικό σε αυτή τη μέτρηση είναι ότι λαμβάνει υπόψη τις συσχετίσεις των χαρακτηριστικών, σε αντίθεση με την Ευκλείδεια απόσταση.

Η απόσταση Mahalanobis μετριέται από το μέσο όρο των δεδομένων και μπορεί να ερμηνευθεί ως ένα πολλαπλάσιο της τυπικής απόκλισης. Στην ανίχνευση ανωμαλιών, τα σημεία που έχουν μεγάλη απόσταση Mahalanobis από το μέσο όρο των δεδομένων θεωρούνται ανωμαλίες. Αυτό οφείλεται στο γεγονός ότι αυτά τα σημεία είναι σπάνια σύμφωνα με τη στατιστική κατανομή των δεδομένων.

Επίσης, η απόσταση Mahalanobis έχει την ιδιότητα να είναι αναλλοίωτη σε περιστροφές του συντεταγμένου συστήματος και κλίμακες των αξόνων, έτσι είναι ιδιαίτερα χρήσιμη όταν η συσχέτιση μεταξύ των χαρακτηριστικών είναι σημαντική.

3.2.10 Hidden Markov Models (HMM)

Τα Hidden Markov Models (HMM) είναι στατιστικά μοντέλα που χρησιμοποιούνται κυρίως στην επεξεργασία φωνής και την αναγνώριση χειρόγραφων. Χρησιμοποιούνται επίσης σε εφαρμογές όπως η βιοπληροφορική, για την ανάλυση σειρών αμινοξέων, ή στην υπολογιστική λογοτεχνία για την ανάλυση κειμένων.

Ένα HMM περιλαμβάνει μια σειρά "κρυφών" καταστάσεων, κάθε μια από τις οποίες παράγει επιθετικά μια παρατήρηση. Οι μεταβάσεις ανάμεσα σε αυτές τις καταστάσεις διέπονται από μια πιθανότητα, δίνοντας στο μοντέλο τη δυνατότητα να αποτυπώνει χρονικές εξαρτήσεις μεταξύ παρατηρήσεων.

Στο πλαίσιο της ανίχνευσης ανωμαλιών, ένα HMM μπορεί να εκπαιδευτεί σε "κανονικά" δεδομένα. Στη συνέχεια, μια νέα ακολουθία δεδομένων (π.χ., μια σειρά μετρήσεων ή γεγονότων) μπορεί να αξιολογηθεί ως προς την πιθανότητα της να είναι παραγόμενη από το μοντέλο. Αν αυτή η πιθανότητα είναι εξαιρετικά χαμηλή, η ακολουθία μπορεί να θεωρηθεί ως ανωμαλία.

Τα HMMs μπορούν να είναι ιδιαίτερα χρήσιμα για την ανίχνευση ανωμαλιών σε χρονοσειρές δεδομένων, όπου οι χρονικές εξαρτήσεις μεταξύ των σημείων είναι σημαντικές.

3.2.11 Replicator Neural Networks (RNN)

Τα Replicator Neural Networks (RNN) είναι μορφή νευρωνικών δικτύων βαθιάς μάθησης που χρησιμοποιούνται για την ανίχνευση ανωμαλιών. Το όνομα τους αποτυπώνει την ικανότητά τους να "αντιγράφουν" ή να "αναπαράγουν" τα δεδομένα εισόδου, εκπαιδευόμενα να αναδιαμορφώσουν την είσοδο όσο το δυνατόν πιο πιστά.

Ένα RNN εκπαιδεύεται παρατηρώντας ένα μεγάλο σύνολο κανονικών δεδομένων και μαθαίνοντας να αναπαράγει αυτά τα δεδομένα. Η ιδέα είναι ότι το νευρωνικό δίκτυο θα

μάθει τη "δομή" των κανονικών δεδομένων, και έτσι θα μπορεί να ανακατασκευάσει με ακρίβεια τα κανονικά δεδομένα, αλλά θα έχει προβλήματα να ανακατασκευάσει τα ανώμαλα δεδομένα.

Η ανίχνευση ανωμαλιών στη συνέχεια πραγματοποιείται με την εφαρμογή του RNN σε νέα δεδομένα. Εάν το RNN δυσκολεύεται να αναπαράγει ένα νέο σημείο δεδομένων (δηλαδή, εάν το σφάλμα ανακατασκευής είναι μεγάλο), τότε το σημείο δεδομένων είναι πιθανόν να είναι ανωμαλία.

Οι αυτοκωδικοποιητές είναι ένα παράδειγμα RNN που χρησιμοποιείται συχνά για την ανίχνευση ανωμαλιών.

3.2.12 Ανάλυση Κύριων Συνιστωσών (PCA)

Η Ανάλυση Κύριων Συνιστωσών (PCA) είναι μια τεχνική μείωσης διαστάσεων που εφαρμόζεται συχνά σε προβλήματα επεξεργασίας δεδομένων και μηχανικής μάθησης. Το βασικό ιδεολογικό πλαίσιο της PCA είναι να μετατρέψει ένα σύνολο δεδομένων με πολλές πιθανές μεταβλητές σε ένα μικρότερο σύνολο μεταβλητών, κρατώντας την περισσότερη δυνατή πληροφορία. Οι "κύριες συνιστώσες" που παράγει η PCA είναι νέες μεταβλητές που είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών.

Στην ανίχνευση ανωμαλιών, η PCA χρησιμοποιείται για να μειώσει την διάσταση των δεδομένων και να προσπαθήσει να ανακατασκευάσει τα αρχικά δεδομένα από τις μειωμένες διαστάσεις. Αν η ανακατασκευή είναι ακριβής, το δεδομένο θεωρείται "κανονικό". Αν όμως η ανακατασκευή έχει μεγάλο σφάλμα (δηλαδή, απαιτείται μεγάλη "προσπάθεια" για να επαναφέρουμε τα αρχικά δεδομένα από τη μειωμένη διάσταση), τότε το δεδομένο θεωρείται "ανώμαλο".

Αυτό συμβαίνει επειδή οι κύριες συνιστώσες της PCA σχεδιάζονται για να καταγράφουν τις "κανονικές" διακυμάνσεις στα δεδομένα. Επομένως, οι ανωμαλίες, που αποτελούν ασυνήθιστες παραλλαγές, δεν εντάσσονται καλά στο χαμηλό-διάστατο χώρο που κατασκευάζει η PCA.

3.2.13 Ανίχνευση ανωμαλιών με βάση την ομαδοποίηση

Η ανίχνευση ανωμαλιών με βάση την ομαδοποίηση είναι μια προσέγγιση που χρησιμοποιεί αλγορίθμους ομαδοποίησης για να εντοπίσει ανωμαλίες στα δεδομένα. Οι αλγόριθμοι ομαδοποίησης χωρίζουν τα δεδομένα σε ξεχωριστές ομάδες ή "clusters" με βάση την ομοιότητα των δεδομένων. Αυτό μπορεί να βασίζεται σε διάφορα χαρακτηριστικά, όπως η ευκλείδεια απόσταση μεταξύ των σημείων δεδομένων, ή μπορεί να βασίζεται σε πιο πολύπλοκα μοντέλα.

Οι ανωμαλίες μπορούν να εντοπιστούν με διάφορους τρόπους μετά την ομαδοποίηση. Ένας κοινός τρόπος είναι να θεωρήσουμε τα σημεία δεδομένων που δεν ανήκουν σε

κανένα cluster ως ανωμαλίες. Αυτό μπορεί να σημαίνει ότι αυτά τα σημεία είναι πολύ διαφορετικά από όλα τα άλλα σημεία δεδομένων, και επομένως μπορεί να θεωρηθούν ως ανώμαλα.

Μια άλλη προσέγγιση είναι να θεωρήσουμε τα μικρά ή αραιά clusters ως ανωμαλίες. Αυτά τα clusters μπορεί να περιέχουν σημεία δεδομένων που είναι σχετικά διαφορετικά από τα υπόλοιπα, αλλά όχι τόσο διαφορετικά ώστε να μην ανήκουν σε κανένα cluster. Αντίθετα, μπορεί να υπάρχει αρκετή ομοιότητα μεταξύ αυτών των σημείων δεδομένων ώστε να δημιουργηθεί ένα μικρό cluster.

Ωστόσο, η ομαδοποίηση βασίζεται στην υπόθεση ότι οι ανωμαλίες είναι σπάνιες και διαφορετικές από τα κανονικά δεδομένα, πράγμα που ενδέχεται να μην είναι πάντα αληθές. Επιπλέον, η απόδοση της ομαδοποίησης μπορεί να επηρεαστεί από την επιλογή της μετρικής ομοιότητας και του αριθμού των clusters.

3.2.14 Δέντρα αποφάσεων

Τα δέντρα αποφάσεων είναι μοντέλα πρόβλεψης που χρησιμοποιούνται σε στατιστική, τη μηχανική μάθηση και την ανάλυση δεδομένων. Αυτά τα μοντέλα αποτελούνται από σημεία αποφάσεων και φύλλα που αντιπροσωπεύουν την απάντηση ή το αποτέλεσμα. Τα σημεία αποφάσεων αποτελούνται από ερωτήσεις ή τεστ που προκύπτουν από τα δεδομένα και τα φύλλα αποτελούν τις προβλέψεις που προκύπτουν από αυτές τις ερωτήσεις.

Όταν χρησιμοποιούνται για ανίχνευση ανωμαλιών, τα δέντρα αποφάσεων χρησιμοποιούνται για την κατανόηση της "κανονικής" συμπεριφοράς των δεδομένων. Τα δεδομένα που διαφέρουν σημαντικά από αυτή την κανονική συμπεριφορά θεωρούνται ως ανωμαλίες.

Επίσης, τα δέντρα αποφάσεων μπορούν να χρησιμοποιηθούν για την κατανόηση των ασυνήθιστων μονοπατιών ή των χαμηλών συχνοτήτων. Αν ένα συγκεκριμένο μονοπάτι προκύψει πολύ σπάνια, τα δεδομένα που ακολουθούν αυτό το μονοπάτι μπορεί να θεωρηθούν ως ανώμαλα.

Τα τυχαία δάση, από την άλλη πλευρά, είναι μια μέθοδος συνόλου που αποτελείται από πολλά δέντρα αποφάσεων. Τα τυχαία δάση μπορούν να εντοπίσουν ανωμαλίες με παρόμοιο τρόπο, χρησιμοποιώντας την διασπορά των δεδομένων, την απόκλιση από την κανονικότητα ή την ανωμαλία σε μεγάλο αριθμό των δέντρων του δάσους.

Είναι σημαντικό να σημειωθεί ότι η επιλογή του αλγορίθμου εξαρτάται από τα συγκεκριμένα χαρακτηριστικά των δεδομένων, τον τύπο των ανωμαλιών που στοχεύουν και τους διαθέσιμους πόρους και την τεχνογνωσία για υλοποίηση. Ο πειραματισμός και η αξιολόγηση σε συγκεκριμένα σύνολα δεδομένων είναι ζωτικής σημασίας για την επιλογή του καταλληλότερου αλγορίθμου.

Κεφάλαιο 4: Υλοποίηση

Στο παρόν κεφάλαιο περιγράφεται η υλοποίηση της μεθοδολογίας ανίχνευσης ανωμαλιών σε δεδομένα αισθητήρων μέσω αλγορίθμων μη επιβλεπόμενης μάθησης.

Όπως αναφέρεται και σε προηγούμενα κεφάλαια, η μη επιβλεπόμενη μηχανική μάθηση είναι μια τεχνική μηχανικής μάθησης που αναλύει τα εισερχόμενα δεδομένα χωρίς να υπάρχουν προηγούμενες ετικέτες ή ταξινόμηση. Αυτή η τεχνική χρησιμοποιείται ευρέως σε προβλήματα όπως η ομαδοποίηση, η ανίχνευση ανωμαλιών, η εξερεύνηση δεδομένων και η μείωση της διαστατικότητας.

Για το σκοπό αυτής της εργασίας χρησιμοποιείται η μη επιβλεπόμενη μηχανική μάθηση για την ανίχνευση ανωμαλιών. Ο λόγος για τον οποίο χρησιμοποιείται η μη επιβλεπόμενη μάθηση σε αυτήν την περίπτωση είναι διότι, σε πολλές περιπτώσεις, δεν γνωρίζουμε προκαταβολικά ποια δεδομένα είναι ανωμαλίες ή πώς μια ανωμαλία θα μοιάζει. Παράλληλα, συχνά τα δεδομένα που χρησιμοποιούμε για ανίχνευση ανωμαλιών δεν είναι επισημασμένα, πράγμα που καθιστά δύσκολη τη χρήση επιβλεπόμενων τεχνικών μάθησης.

Οι αλγόριθμοι που χρησιμοποιούνται σε αυτόν τον κώδικα, όπως ο Isolation Forest, ο K Nearest Neighbors (k-NN), το OneClassSVM και το Autoencoder, είναι όλοι αλγόριθμοι μη επιβλεπόμενης μηχανικής μάθησης που μπορούν να εντοπίσουν ανωμαλίες χωρίς να χρειάζεται να παρέχουμε ετικέτες.

Παρακάτω θα αναλυθούν και θα περιγράφουν τα εργαλεία που σε τεχνικό επίπεδο χρησιμοποιήθηκαν για την ανάπτυξη του αλγορίθμου και όλων των τεχνικών διαδικασιών για αυτή τη πτυχιακή.

4.1 Εργαλεία

4.1.1 Γλώσσα Προγραμματισμού Python

Η δημοτικότητα της Python στην κοινότητα μηχανικής μάθησης πηγάζει από διάφορους παράγοντες:

- Αναγνωσιμότητα και ευκολία χρήσης: Η καθαρή σύνταξη και η εκφραστική φύση της Python καθιστούν εύκολη την ανάγνωση, τη γραφή και την κατανόηση. Αυτό ενισχύει τη συνεργασία και επιταχύνει την ανάπτυξη στη ροή εργασιών μηχανικής εκμάθησης.
- Πλούσιο Οικοσύστημα: Η Python διαθέτει μια τεράστια συλλογή βιβλιοθηκών και πλαισίων αφιερωμένων στη μηχανική μάθηση, προσφέροντας προ-εφαρμοσμένους αλγόριθμους, εργαλεία χειρισμού δεδομένων και δυνατότητες οπτικοποίησης. Αυτό επιτρέπει στους προγραμματιστές να αξιοποιήσουν τους υπάρχοντες πόρους και να επιταχύνουν τα έργα μηχανικής εκμάθησης.
- Ισχυρή υποστήριξη κοινότητας: Η Python έχει μια μεγάλη και ενεργή κοινότητα προγραμματιστών και ερευνητών στον τομέα της μηχανικής εκμάθησης. Αυτή η

κοινότητα συμβάλλει στην ανάπτυξη βιβλιοθηκών, μοιράζεται γνώσεις μέσω σεμιναρίων και φόρουμ και συνεργάζεται σε έργα ανοιχτού κώδικα.
- Δυνατότητες ενσωμάτωσης: Η Python ενσωματώνεται απρόσκοπτα με άλλες γλώσσες και εργαλεία, καθιστώντας εύκολο τον συνδυασμό μηχανικής εκμάθησης με ανάπτυξη ιστού, διαχείριση βάσεων δεδομένων ή οπτικοποίηση δεδομένων.

Ο συνδυασμός της απλότητας, των ισχυρών βιβλιοθηκών και της υποστηρικτικής κοινότητας της Python την έχει καταστήσει κυρίαρχη γλώσσα για τη μηχανική μάθηση. Έχει γίνει η de facto επιλογή για ερευνητές, επιστήμονες δεδομένων και προγραμματιστές που επιδιώκουν να εξερευνήσουν, να εφαρμόσουν και να αναπτύξουν αποτελεσματικά μοντέλα και εφαρμογές μηχανικής μάθησης.

4.1.2 PyOD

Το PyOD (Python Outlier Detection) είναι μια ολοκληρωμένη βιβλιοθήκη Python ειδικά σχεδιασμένη για ανίχνευση ακραίων στοιχείων σε εργασίες ανάλυσης δεδομένων και μηχανικής μάθησης. Παρέχει ένα ευρύ φάσμα αλγορίθμων και τεχνικών για τον εντοπισμό ανωμαλιών ή ακραίων τιμών σε σύνολα δεδομένων.

Η βιβλιοθήκη PyOD προσφέρει ένα ενοποιημένο και φιλικό προς το χρήστη API, επιτρέποντας στους χρήστες να εφαρμόζουν εύκολα διάφορες μεθόδους ανίχνευσης ακραίων στοιχείων χωρίς να χρειάζεται να εμβαθύνουν στις λεπτομέρειες υλοποίησης. Ενσωματώνεται καλά με άλλες δημοφιλείς βιβλιοθήκες Python όπως οι NumPy, Pandas και scikit-learn, διευκολύνοντας τον απρόσκοπτο χειρισμό δεδομένων και την αξιολόγηση μοντέλων.

Τα βασικά χαρακτηριστικά και οι λειτουργίες του PyOD περιλαμβάνουν:

- **Ευρύ φάσμα αλγορίθμων**: Το PyOD εφαρμόζει μια ποικίλη συλλογή αλγορίθμων ανίχνευσης ακραίων τιμών, συμπεριλαμβανομένων τόσο των παραδοσιακών στατιστικών τεχνικών όσο και των σύγχρονων προσεγγίσεων που βασίζονται στη μηχανική μάθηση. Περιλαμβάνει δημοφιλείς μεθόδους όπως k-NN, Isolation Forest, LOF, PCA και πολλές άλλες.
- **Μέθοδοι συνόλου**: Το PyOD προσφέρει μεθόδους συνόλου που συνδυάζουν πολλαπλούς αλγόριθμους ανίχνευσης ακραίων τιμών για τη βελτίωση της συνολικής απόδοσης και ευρωστίας. Αυτές οι μέθοδοι συνόλου αξιοποιούν τη συλλογική δύναμη λήψης αποφάσεων πολλών μοντέλων για να παρέχουν πιο ακριβή αποτελέσματα ανίχνευσης ανωμαλιών.
- **Outlier Visualization**: Η βιβλιοθήκη περιλαμβάνει εργαλεία οπτικοποίησης που βοηθούν τους χρήστες να κατανοήσουν και να ερμηνεύσουν τις ανωμαλίες που έχουν εντοπιστεί. Παρέχει λειτουργίες σχεδίασης για την οπτικοποίηση της κατανομής δεδομένων, των ορίων απόφασης και των βαθμολογιών ανωμαλιών, επιτρέποντας στους χρήστες να αποκτήσουν γνώσεις για τα υποκείμενα μοτίβα και τα ακραία σημεία στα δεδομένα.
- **Αξιολόγηση μοντέλου**: Το PyOD παρέχει μετρήσεις αξιολόγησης για την αξιολόγηση της απόδοσης των μοντέλων ανίχνευσης ακραίων τιμών. Οι χρήστες

μπορούν να υπολογίσουν διάφορες μετρήσεις, όπως ακρίβεια, ανάκληση, βαθμολογία F1 και εμβαδόν κάτω από την καμπύλη ROC (AUC-ROC) για να μετρήσουν την ποιότητα των ανωμαλιών που ανιχνεύονται.
- **Υποστήριξη για διαφορετικούς τύπους δεδομένων**: Το PyOD έχει σχεδιαστεί για να χειρίζεται διάφορους τύπους δεδομένων, συμπεριλαμβανομένων αριθμητικών, κατηγορικών και μικτών δεδομένων. Προσφέρει λειτουργίες προεπεξεργασίας για το χειρισμό διαφορετικών τύπων δεδομένων, καθιστώντας το ευέλικτο και προσαρμόσιμο σε ένα ευρύ φάσμα πραγματικών σεναρίων.
- **Ενσωμάτωση με αγωγούς μηχανικής εκμάθησης**: Το PyOD ενσωματώνεται απρόσκοπτα με το πλαίσιο διοχέτευσης του scikit-learn, επιτρέποντας στους χρήστες να ενσωματώνουν τον εντοπισμό ακραίων στοιχείων ως βήμα προεπεξεργασίας στις ροές εργασιών μηχανικής εκμάθησης. Αυτό επιτρέπει τον συνδυασμό ανίχνευσης ακραίων τιμών με άλλους αλγόριθμους και τεχνικές μηχανικής μάθησης.

Η βιβλιοθήκη PyOD διατηρείται ενεργά και υποστηρίζεται από μια ζωντανή κοινότητα προγραμματιστών. Χρησιμοποιείται ευρέως σε διάφορους τομείς, όπως ανίχνευση απάτης, ανίχνευση εισβολών, ασφάλεια δικτύου, παρακολούθηση ανωμαλιών και αξιολόγηση ποιότητας δεδομένων. Με το ολοκληρωμένο σύνολο αλγορίθμων και τη φιλική διεπαφή χρήστη, το PyOD απλοποιεί τη διαδικασία ανίχνευσης ακραίων τιμών και δίνει τη δυνατότητα στους χρήστες να αναλύουν αποτελεσματικά και να εντοπίζουν ανωμαλίες στα δεδομένα τους.

4.1.3 Matplotlib

Το Matplotlib είναι μια ευρέως χρησιμοποιούμενη βιβλιοθήκη Python για τη δημιουργία οπτικοποιήσεων και πλοκών υψηλής ποιότητας. Παρέχει ένα ευέλικτο και περιεκτικό σύνολο εργαλείων για τη δημιουργία διαφόρων τύπων γραφημάτων, διαγραμμάτων και γραφημάτων, καθιστώντας το ένα πολύτιμο εργαλείο για την εξερεύνηση, την ανάλυση και την παρουσίαση δεδομένων.

Τα βασικά χαρακτηριστικά και λειτουργίες του Matplotlib περιλαμβάνουν:

- Δυνατότητες σχεδίασης: Το Matplotlib προσφέρει ένα ευρύ φάσμα συναρτήσεων γραφικής παράστασης για τη δημιουργία γραμμικών γραφημάτων, διαγραμμάτων διασποράς, ραβδώσεων, ιστογραμμάτων, διαγραμμάτων πίτας, χάρτες θερμότητας, τρισδιάστατων γραφημάτων και πολλά άλλα. Παρέχει εκτεταμένες επιλογές προσαρμογής, επιτρέποντας στους χρήστες να ελέγχουν πτυχές όπως χρώματα, στυλ γραμμών, δείκτες, ετικέτες αξόνων, τίτλους, θρύλους και σχολιασμούς.
- Αντικειμενοστραφές API: Το Matplotlib παρέχει ένα αντικειμενοστραφή API που δίνει στους χρήστες λεπτομερή έλεγχο των γραφικών τους. Μπορούν να δημιουργήσουν αντικείμενα Figure και Axes, να χειριστούν τις ιδιότητές τους και να προσθέσουν δευτερεύουσες γραφικές παραστάσεις για να δημιουργήσουν σύνθετες διατάξεις. Αυτό το API παρέχει ευελιξία στην προσαρμογή γραφικών παραστάσεων και επιτρέπει πιο προηγμένες τεχνικές σχεδίασης.

- Συμβατότητα και ενοποίηση: Το Matplotlib είναι συμβατό με διάφορα λειτουργικά συστήματα και λειτουργεί άψογα με άλλες βιβλιοθήκες Python όπως οι NumPy και Pandas. Ενσωματώνεται καλά με τα σημειωματάρια Jupyter και μπορεί να χρησιμοποιηθεί σε συνδυασμό με επιστημονικές βιβλιοθήκες υπολογιστών όπως το SciPy και το scikit-learn.
- Έξοδος Ποιότητας Δημοσίευσης: Το Matplotlib επιτρέπει στους χρήστες να δημιουργούν οπτικοποιήσεις έτοιμες για δημοσίευση με έξοδο υψηλής ανάλυσης σε διάφορες μορφές αρχείων, όπως PNG, PDF, SVG και EPS. Η βιβλιοθήκη παρέχει επιλογές για τον έλεγχο του μεγέθους της εικόνας, του DPI (κουκίδες ανά ίντσα) και της αναλογίας διαστάσεων, διασφαλίζοντας ότι τα σχέδια φαίνονται επαγγελματικά σε έντυπα ή ψηφιακά μέσα.
- Διαδραστική σχεδίαση: Το Matplotlib υποστηρίζει διαδραστικές λειτουργίες, επιτρέποντας στους χρήστες να κάνουν ζουμ, μετατόπιση και πλοήγηση σε γραφικά για να εξερευνήσουν τα δεδομένα πιο αποτελεσματικά. Παρέχει εργαλεία για την προσθήκη διαδραστικότητας, όπως συμβουλές εργαλείων, εφέ αιώρησης του ποντικιού και στοιχεία με δυνατότητα κλικ.
- Matplotlib.pyplot: Η ενότητα pyplot του Matplotlib παρέχει μια διεπαφή τύπου MATLAB για γρήγορη και εύκολη δημιουργία γραφικών. Απλοποιεί τη δημιουργία κοινών γραφημάτων με αυτόματη διαχείριση των αντικειμένων Figure και Axes στο παρασκήνιο. Αυτή η ενότητα είναι ιδιαίτερα χρήσιμη για απλές εργασίες σχεδίασης και διαδραστική εξερεύνηση δεδομένων.
- Πλούσιες επιλογές οπτικοποίησης: Το Matplotlib προσφέρει ένα ευρύ φάσμα επιλογών προσαρμογής, συμπεριλαμβανομένου του ελέγχου των χρωμάτων, των στυλ γραμμών, των δεικτών, των σχολιασμών, των πλεγμάτων και των μύθων. Υποστηρίζει διαφορετικούς χρωματικούς χάρτες, γραμμές χρωμάτων και χάρτες χρωμάτων για την αποτελεσματική οπτικοποίηση των αριθμητικών δεδομένων. Παρέχει επίσης υποστήριξη για LaTeX

4.1.4 Numpy

Η NumPy (Αριθμητική Python) είναι μια θεμελιώδης βιβλιοθήκη Python για αριθμητικούς υπολογισμούς. Παρέχει αποτελεσματικές δομές δεδομένων και λειτουργίες για εργασία με μεγάλους, πολυδιάστατους πίνακες και πίνακες. Το NumPy χρησιμεύει ως το θεμέλιο για πολλές άλλες επιστημονικές και σχετικές με δεδομένα βιβλιοθήκες στο οικοσύστημα Python.

Τα βασικά χαρακτηριστικά και λειτουργίες του NumPy περιλαμβάνουν:

- Πολυδιάστατοι πίνακες: Το NumPy εισάγει το αντικείμενο ndarray (N-dimensional array), το οποίο επιτρέπει την αποτελεσματική αποθήκευση και χειρισμό ομοιογενών δεδομένων. Αυτοί οι πίνακες μπορούν να έχουν οποιονδήποτε αριθμό διαστάσεων και να υποστηρίζουν διάφορους τύπους δεδομένων, όπως ακέραιους, κινητήρες και μιγαδικούς αριθμούς.
- Μαθηματικές πράξεις: Το NumPy παρέχει ένα ευρύ φάσμα μαθηματικών συναρτήσεων και πράξεων για χειρισμό πίνακα. Περιλαμβάνει βασικές μαθηματικές πράξεις όπως πρόσθεση, αφαίρεση, πολλαπλασιασμό και διαίρεση,

καθώς και πιο προηγμένες πράξεις όπως τριγωνομετρικές συναρτήσεις, εκθετικές συναρτήσεις, γραμμική άλγεβρα, μετασχηματισμούς Fourier και στατιστικές πράξεις.

- Μετάδοση: Η δυνατότητα μετάδοσης του NumPy επιτρέπει αποτελεσματικές και σιωπηρές λειτουργίες ως προς τα στοιχεία σε συστοιχίες με διαφορετικά σχήματα και μεγέθη. Αυτή η δυνατότητα απλοποιεί την υλοποίηση μαθηματικών υπολογισμών και βελτιώνει την αναγνωσιμότητα κώδικα.
- Καθολικές συναρτήσεις (ufuncs): Τα ufuncs του NumPy είναι συναρτήσεις που λειτουργούν ως προς τα στοιχεία σε πίνακες, κάνοντας τους υπολογισμούς ταχύτερους και πιο συνοπτικούς. Αυτές οι συναρτήσεις είναι βελτιστοποιημένες για αποτελεσματικότητα και είναι διανυσματικές, που σημαίνει ότι μπορούν να επεξεργάζονται πίνακες παράλληλα, με αποτέλεσμα την ταχύτερη εκτέλεση σε σύγκριση με τη χρήση βρόχων.
- Χειρισμός πίνακα: Το NumPy προσφέρει ένα ευρύ φάσμα λειτουργιών για την αναμόρφωση, τον τεμαχισμό, τη συνένωση, το διαχωρισμό και τη στοίβαξη πινάκων. Αυτές οι λειτουργίες παρέχουν ευελιξία στο χειρισμό των διαστάσεων του πίνακα και στην εξαγωγή συγκεκριμένων στοιχείων ή υποσυνόλων δεδομένων.
- Δημιουργία τυχαίων αριθμών: Το NumPy περιλαμβάνει μια ισχυρή μονάδα τυχαίων αριθμών που επιτρέπει τη δημιουργία τυχαίων αριθμών και πινάκων. Παρέχει διάφορες κατανομές πιθανοτήτων, συναρτήσεις τυχαίας δειγματοληψίας και εργαλεία για τη δημιουργία τυχαίων μεταθέσεων και ανακάτεμα δεδομένων.
- Ενσωμάτωση με γλώσσες χαμηλού επιπέδου: Το NumPy έχει σχεδιαστεί για να ενσωματώνεται απρόσκοπτα με άλλες γλώσσες προγραμματισμού, ιδιαίτερα με γλώσσες χαμηλού επιπέδου όπως η C και η Fortran. Αυτή η ενοποίηση επιτρέπει την αποτελεσματική εκτέλεση αριθμητικών υπολογισμών και διευκολύνει τη χρήση υπαρχουσών βιβλιοθηκών και κώδικα γραμμένου σε αυτές τις γλώσσες.
- Βελτιστοποίηση απόδοσης: Η υλοποίηση του NumPy είναι εξαιρετικά βελτιστοποιημένη, επιτρέποντας την αποτελεσματική εκτέλεση αριθμητικών πράξεων. Αξιοποιεί τον υποκείμενο κώδικα C και Fortran για απόδοση, καθιστώντας τον σημαντικά ταχύτερο από την εκτέλεση παρόμοιων υπολογισμών χρησιμοποιώντας δομές και βρόχους δεδομένων της Python.

Το NumPy χρησιμοποιείται ευρέως σε διάφορους τομείς, όπως η επιστημονική έρευνα, η ανάλυση δεδομένων, η μηχανική μάθηση, η επεξεργασία εικόνας και οι προσομοιώσεις. Αποτελεί τη ραχοκοκαλιά πολλών δημοφιλών βιβλιοθηκών στο επιστημονικό οικοσύστημα της Python, όπως το SciPy, τα pandas, το scikit-learn και το TensorFlow. Ο αποτελεσματικός χειρισμός του πίνακα και οι μαθηματικές πράξεις του το καθιστούν θεμελιώδες εργαλείο για την εργασία με αριθμητικά δεδομένα στην Python.

4.2 Απόκτηση Δεδομένων

Στην παρούσα ενότητα, γίνεται αναλυτική αναφορά στην απόκτηση δεδομένων από τους αισθητήρες μέσω της διεπαφής API. Η βιβλιοθήκη `requests` χρησιμοποιείται για την αποστολή αιτημάτων προς το API και τη λήψη των δεδομένων από αυτό. Τα δεδομένα που λαμβάνονται είναι σε μορφή JSON και προσαρμόζονται σε ένα DataFrame για περαιτέρω επεξεργασία. Στη συνέχεια, γίνεται απόρριψη των μη επιθυμητών στηλών, όπως οι στήλες 'timestamp' και 'application_group'. Επίσης, διαγράφονται οι εγγραφές χωρίς τιμές και τα δεδομένα ταξινομούνται βάσει ευρετηρίου. Παρακάτω παρουσιάζεται ένα παράδειγμα του κώδικα που εκτελεί τη διαδικασία αυτή:

```
def get_data(sensor):
    # Λήψη δεδομένων από το API
    sensor_data = requests.get('https://test.uowm.gr/v1/' + sensor + '/all')
    # Μετατροπή σε μορφή JSON
    json_data = json.loads(sensor_data.content)
    # Δημιουργία DataFrame από τα δεδομένα
    df = pandas.DataFrame.from_records(json_data)
    # Απόρριψη μη επιθυμητών στηλών
    df.drop('timestamp', inplace=True, axis=1)
    df.drop('application_group', inplace=True, axis=1)
    # Διαγραφή εγγραφών χωρίς τιμές
    df.dropna(inplace=True)
    # Ταξινόμηση βάσει ευρετηρίου
    df.sort_index(inplace=True)
    # Διαχωρισμός συνόλων εκπαίδευσης και ελέγχου
    train, test = train_test_split(df)
```

4.2.1 Εκπαίδευση Δεδομένων

Σε αυτήν την υποενότητα, παρουσιάζεται η διαδικασία εκπαίδευσης των δεδομένων χρησιμοποιώντας τους αλγορίθμους ανίχνευσης ανωμαλιών. Εισάγονται οι κλάσεις των αλγορίθμων από τη βιβλιοθήκη `pyod`, όπως ο `AutoEncoder`, ο `IForest`, ο `k-NN` και ο `OCSVM` (One-Class Support Vector Machines). Επίσης, χρησιμοποιείται η συνάρτηση `train_test_split` από τη βιβλιοθήκη `sklearn` για τον διαχωρισμό των δεδομένων σε σύνολα εκπαίδευσης και ελέγχου. Τέλος, παρουσιάζεται ένα παράδειγμα του κώδικα που εκτελεί την εκπαίδευση των δεδομένων για κάθε αλγόριθμο:

```
def train_data(train_data, test_data, clf):

    # Προσαρμογή των δεδομένων στο μοντέλο

    clf.fit(train_data)

    # Απόκτηση σκορ ανωμαλιών για το σύνολο εκπαίδευσης

    train_scores = clf.decision_scores_

    # Απόκτηση σκορ ανωμαλιών για το σύνολο ελέγχου

    test_scores = clf.decision_function(test_data)

    # Πρόβλεψη και εμπιστοσύνη πρόβλεψης για το σύνολο ελέγχο.

    test_pred, test_pred_confidence = clf.predict(test_data, return_confidence=True)

    test_results = {

        'train_scores': train_scores,

        'test_scores': test_scores,

        'test_pred': test_pred,

    'test_pred_confidence': test_pred_confidence

    }

    return test_results
```

Αυτός ο κώδικας περιγράφει μια διαδικασία εκπαίδευσης ενός μοντέλου για την ανίχνευση ανωμαλιών και στη συνέχεια αξιολόγησης του μοντέλου σε ένα σύνολο δεδομένων ελέγχου.

4.2.2 Μείωση Διαστάσεων Δεδομένων

Σε αυτήν την υποενότητα, παρουσιάζεται η διαδικασία μείωσης των διαστάσεων των δεδομένων χρησιμοποιώντας τη μέθοδο του PCA (Principal Component Analysis). Εισάγεται η κλάση `PCA` από τη βιβλιοθήκη `sklearn.decomposition`. Στη συνέχεια, παρουσιάζεται ένα παράδειγμα του κώδικα που εκτελεί τη διαδικασία αυτή:

```
def apply_dimensionality_reduction(data):

    # Εκτέλεση PCA για μείωση των διαστάσεων

    pca = PCA(n_components=2)  # Μείωση σε 2 διαστάσεις για scatter plot

    reduced_data = pca.fit_transform(data)

    return reduced_data
```

Στην παραπάνω συνάρτηση `apply_dimensionality_reduction(data)`, χρησιμοποιείται η τεχνική της Ανάλυσης Κυρίων Συνιστωσών (PCA - Principal Component Analysis) για τη μείωση των διαστάσεων των δεδομένων.

Ειδικότερα, αυτό που επιτυγχάνεται είναι τα εξής:

1. Μείωση Διαστάσεων: Οι αρχικές διαστάσεις των δεδομένων μειώνονται σε δύο, έτσι ώστε να μπορούν να απεικονιστούν σε ένα δισδιάστατο χώρο.

2. Διατήρηση της Μεγαλύτερης Δυνατής Πληροφορίας: Η PCA προσπαθεί να διατηρήσει τη μεγαλύτερη δυνατή ποσότητα πληροφορίας από τα αρχικά δεδομένα, επιλέγοντας τις δύο κυριότερες συνιστώσες που κατέχουν τη μεγαλύτερη δυνατή διακύμανση.

3. Οπτικοποίηση: Η μείωση σε δύο διαστάσεις καθιστά εφικτή την οπτικοποίηση των δεδομένων σε ένα scatter plot, κάτι που είναι ιδιαίτερα χρήσιμο για την κατανόηση των δεδομένων, την ανίχνευση ομαδοποιήσεων ή τον εντοπισμό ανωμαλιών.

4.2.3 Απεικόνιση Αποτελεσμάτων

Σε αυτήν την υποενότητα, παρουσιάζεται η διαδικασία απεικόνισης των αποτελεσμάτων χρησιμοποιώντας γραφική απεικόνιση. Χρησιμοποιείται η βιβλιοθήκη `matplotlib.pyplot` για τη δημιουργία των γραφημάτων scatter plot. Επίσης, παρέχεται ένα παράδειγμα του κώδικα που εκτελεί τη διαδικασία αυτή:

```
def plot_results(test_data, y_test_pred, description):

    # Απόκτηση ονομάτων στηλών από το dataset για τη δημιουργία του γραφήματος

    label_names = {

      'feat1': 'Feature 1',

      'feat2': 'Feature 2' }

    # Απεικόνιση των δεδομένων ελέγχου

      plt.figure(figsize=(10, 6))

      plt.scatter(test_data['feat1'], test_data['feat2'], c=y_test_pred, cmap='viridis')

      plt.xlabel(label_names['feat1'])

      plt.ylabel(label_names['feat2'])

      plt.title(description)

      plt.colorbar()

      plt.show()
```

Η παραπάνω συνάρτηση `plot_results` επιτυγχάνει την οπτικοποίηση των δεδομένων `test_data` σε ένα δισδιάστατο χώρο χρησιμοποιώντας ένα scatter plot, με βάση δύο χαρακτηριστικά που αναφέρονται ως 'feat1' και 'feat2'.

Ειδικότερα, αυτό που επιτυγχάνεται με τη συνάρτηση είναι:

1. Οπτικοποίηση Δεδομένων: Παρουσιάζονται τα δεδομένα `test_data` σε ένα scatter plot χρησιμοποιώντας τις στήλες 'feat1' και 'feat2' ως τους δύο άξονες του γραφήματος.

2.Χρωματική Διάκριση: Τα σημεία στο scatter plot χρωματίζονται με βάση τις προβλέψεις `y_test_pred` χρησιμοποιώντας το colormap 'viridis'. Αυτό μπορεί να βοηθήσει στην οπτική διάκριση των διάφορων κατηγοριών ή ομάδων που προκύπτουν από τις προβλέψεις.

3. Ετικέτες και Τίτλος: Η συνάρτηση προσθέτει ετικέτες στους άξονες (με βάση το λεξικό `label_names`) και έναν τίτλο `description` για να καταστήσει το γράφημα πιο ενημερωτικό.

4. Χρωματική Κλίμακα: Προστίθεται μια χρωματική κλίμακα (`colorbar`) για να δείχνει τη σχέση μεταξύ των χρωμάτων των σημείων και των τιμών των προβλέψεων.

Αυτό είναι ένα παράδειγμα υλοποίησης των διαφόρων επιμέρους βημάτων της διαδικασίας ανίχνευσης ανωμαλιών χρησιμοποιώντας τη γλώσσα προγραμματισμού Python. Ο κώδικας παρέχει ένα πλαίσιο για την υλοποίηση της ανίχνευσης ανωμαλιών με αλγόριθμους όπως ο `AutoEncoder`, ο `IForest`, ο `k-NN` και ο `OCSVM.

Κεφάλαιο 5: Αποτελέσματα

5.1 Γενικες πληροφοριες για τα διαγράμματα

Το scatter plot που παράγεται από τη συνάρτηση `plot_results` χρησιμοποιείται για την οπτικοποίηση της κατανομής των σημείων δεδομένων σε έναν δισδιάστατο χώρο. Παρέχει μια γραφική αναπαράσταση της σχέσης μεταξύ δύο μεταβλητών, συνήθως προκύπτοντας από τεχνικές μείωσης της διαστατικότητας όπως η Ανάλυση των Κύριων Συνιστωσών (PCA).

Στο πλαίσιο της ανίχνευσης ανωμαλιών, το scatter plot βοηθά στην οπτικοποίηση του τρόπου με τον οποίο ο αλγόριθμος ταξινομεί τα σημεία δεδομένων ως κανονικά ή ανώμαλα. Κάθε σημείο δεδομένων αναπαρίσταται ως ένα σημείο στο διάγραμμα, με τη θέση του να καθορίζεται από τις τιμές των δύο επιλεγμένων διαστάσεων (π.χ. PC1 και PC2). Το χρώμα του σημείου υποδηλώνει την προβλεπόμενη ετικέτα που έχει ανατεθεί από τον αλγόριθμο, όπου διάφορα χρώματα αντιστοιχούν σε διάφορες κλάσεις ή κατηγορίες.

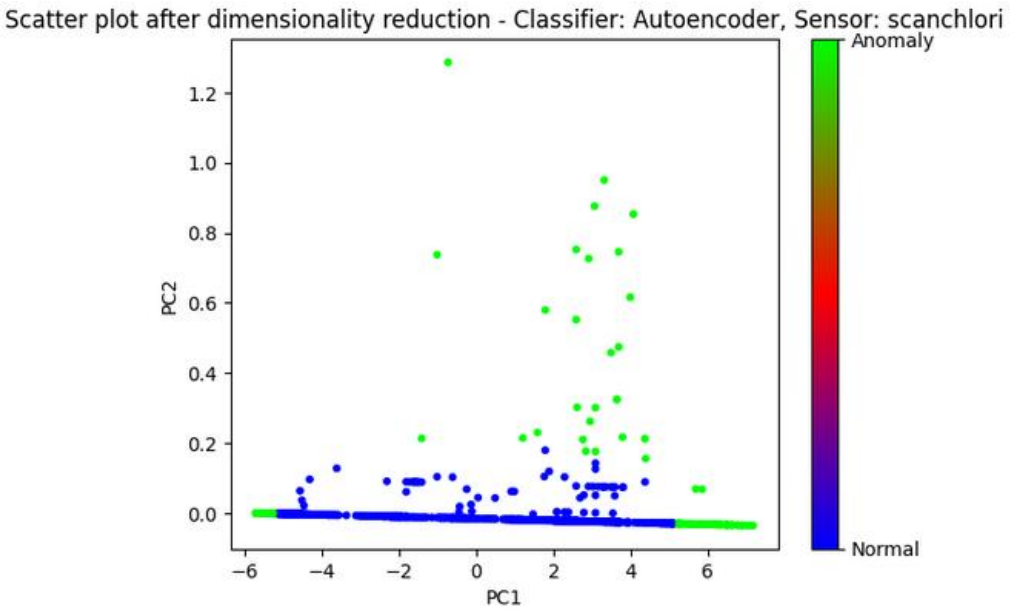
Διερευνώντας τα scatter plots, μπορουμε να παρατηρήσουμε πρότυπα, ομάδες (clusters) ή απομονωμένες περιπτώσεις μεταξύ των κανονικών και ανώμαλων σημείων δεδομένων. Ένας καλός αλγόριθμος ανίχνευσης ανωμαλιών θα έπρεπε ιδανικά να εμφανίζει μια καθαρή διάκριση μεταξύ των δύο κλάσεων, με τα κανονικά σημεία να σχηματίζουν ξεχωριστά σύνολα και τα ανώμαλα σημεία να εμφανίζονται ως εκτός των ομάδων ή απομονωμένα περιστατικά.

Το scatter plot μπορεί επίσης να αποκαλύψει την αποτελεσματικότητα της μείωσης της διαστατικότητας στην καταγραφή της ενδοτερικής δομής των δεδομένων. Επιτρέπει να οπτικοποιήσουμε τα δεδομένα σε έναν χαμηλότερης διάστασης χώρο, διατηρώντας παράλληλα σημαντικές πληροφορίες σχετικά με τον διαχωρισμό μεταξύ κανονικών και ανώμαλων περιπτώσεων.

5.2 Αποτελέσματα

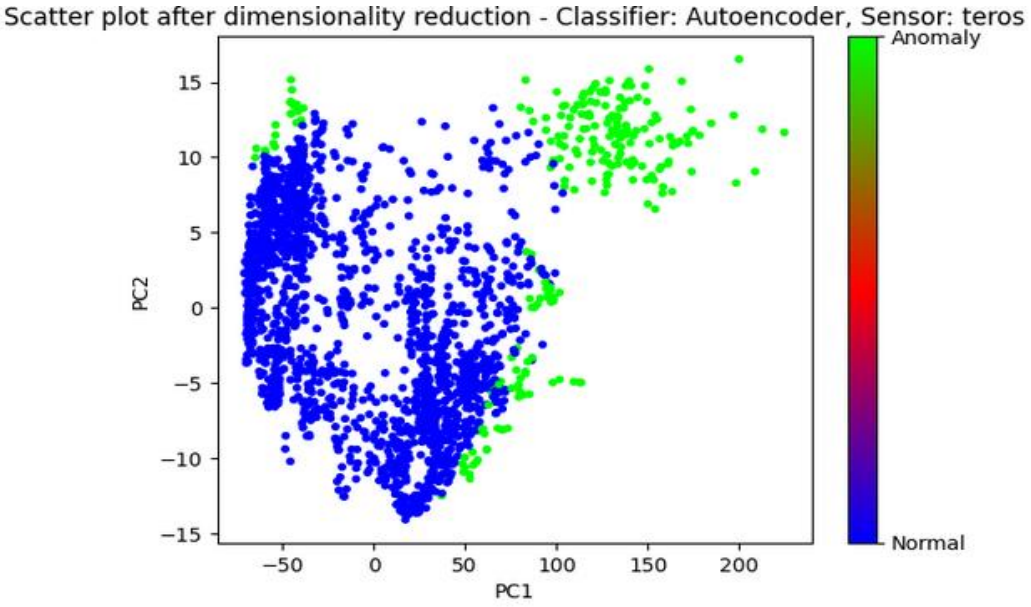
1. **Παράδειγμα 1ο: Ίδιος classifier (Autoencoder) με 6 διαφορετικούς sensors, για σύγκριση μεταξύ των sensors στον ίδιο classifier**

Φαίνεται από την Εικόνα 3, ότι τα περισσότερα δεδομένα (μπλε σημεία) συγκεντρώνονται κοντά στο κέντρο του γραφήματος. Τα πράσινα σημεία, τα οποία ο αλγόριθμος έχει επισημάνει ως ανωμαλίες, βρίσκονται πιο μακριά από το κέντρο της συσσώρευσης των δεδομένων.



Εικόνα 3. Απόδοση του Classifier Autoencoder με Αισθητήρα Scanchlori

Αυτό μπορεί να σημαίνει ότι τα πράσινα σημεία είναι αποκλίσεις από το συνηθισμένο μοτίβο των δεδομένων. Σε ένα πραγματικό σενάριο, αυτό θα μπορούσε να υποδεικνύει πιθανές ανωμαλίες ή προβλήματα που απαιτούν περαιτέρω εξέταση.

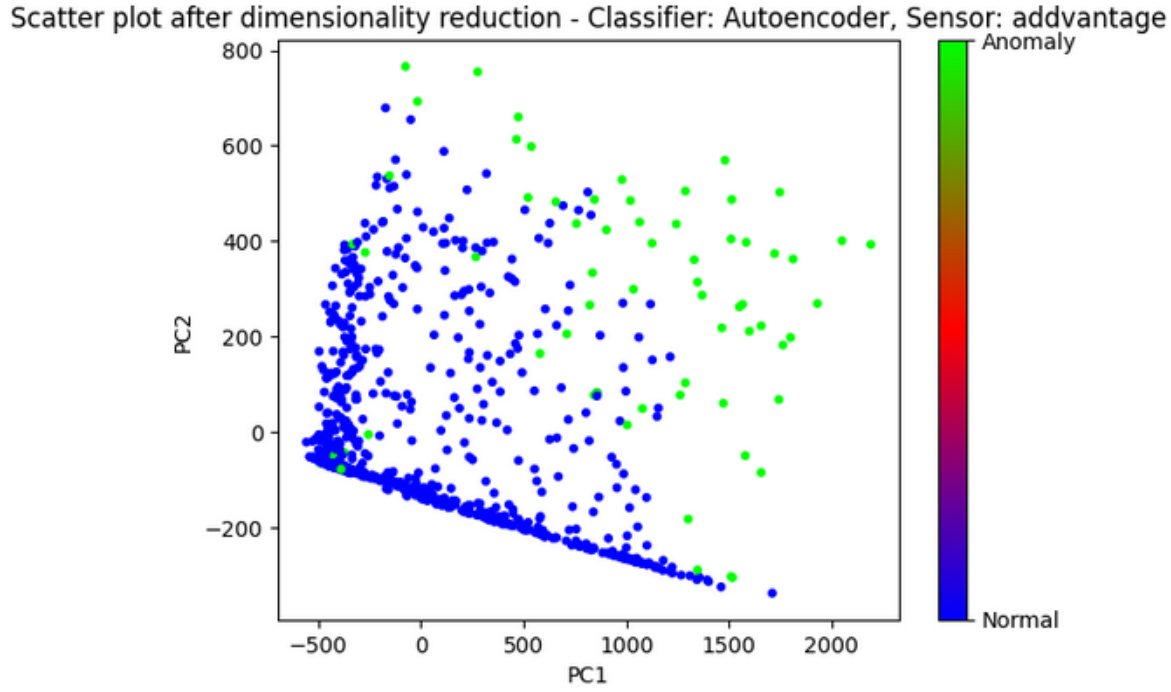


Εικόνα 4. Απόδοση του Classifier Autoencoder με Αισθητήρα Teros

Φαίνεται από την Εικόνα 4, ότι η πλειονότητα των δεδομένων (μπλε σημεία) είναι πιο ομογενή σε σχέση με την Εικόνα 3. Συγκεντρώνονται σε μια σφαιρική περιοχή στο κέντρο του γραφήματος.

Τα πράσινα σημεία, τα οποία ο αλγόριθμος έχει επισημάνει ως ανωμαλίες, φαίνεται να είναι πιο απομονωμένα και σχετικά μακριά από τη συσσώρευση των μπλε σημείων.

Αυτό μπορεί να υποδεικνύει ότι αυτές οι παρατηρήσεις είναι αποκλίσεις από το τυπικό μοτίβο των δεδομένων, που μπορεί να σημαίνει ότι είναι ανωμαλίες ή δεδομένα που παρουσιάζουν μη συνήθη συμπεριφορά. Σε ένα πραγματικό περιβάλλον, αυτό θα μπορούσε να σημαίνει ότι υπάρχουν πιθανά προβλήματα ή συνθήκες που απαιτούν περαιτέρω διερεύνηση ή εξέταση.

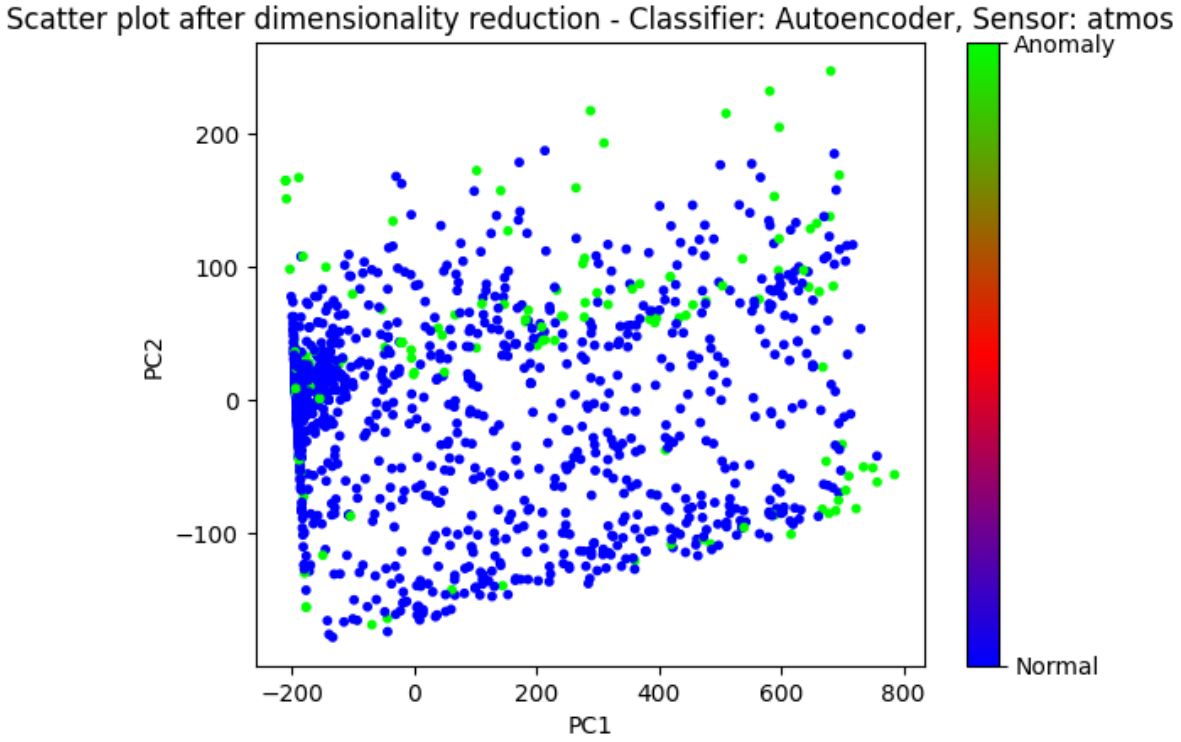


Εικόνα 5. Απόδοση του Classifier Autoencoder με Αισθητήρα Advantage

Η Εικόνα 5, παρουσιάζει κάποιες ιδιαίτερες διαφορές σε σύγκριση με τα προηγούμενα.

Παρατηρούμε ότι η πλειονότητα των δεδομένων (μπλε σημεία) κατανέμεται σε μια σχετικά επίπεδη, λεπτή περιοχή που εκτείνεται από το αριστερό προς το δεξιό μέρος του γραφήματος.
Τα πράσινα σημεία, τα οποία ο αλγόριθμος έχει επισημάνει ως ανωμαλίες, διασκορπίζονται κυρίως γύρω από την κεντρική περιοχή των μπλε σημείων. Αυτό σημαίνει ότι τα πράσινα σημεία είναι απομακρυσμένα από την πυκνότερη περιοχή των μπλε σημείων, καθιστώντας τα πιθανά ανώμαλα σύμφωνα με τον αλγόριθμο.

Μια ερμηνεία σε πραγματικό περιβάλλον θα μπορούσε να είναι ότι τα πράσινα σημεία αντιπροσωπεύουν απρόσμενες καταστάσεις ή συνθήκες που αποκλίνουν σημαντικά από το μέσο όρο των μπλε σημείων. Ενδεχομένως να είναι σημεία που απαιτούν περαιτέρω παρακολούθηση ή εξέταση.

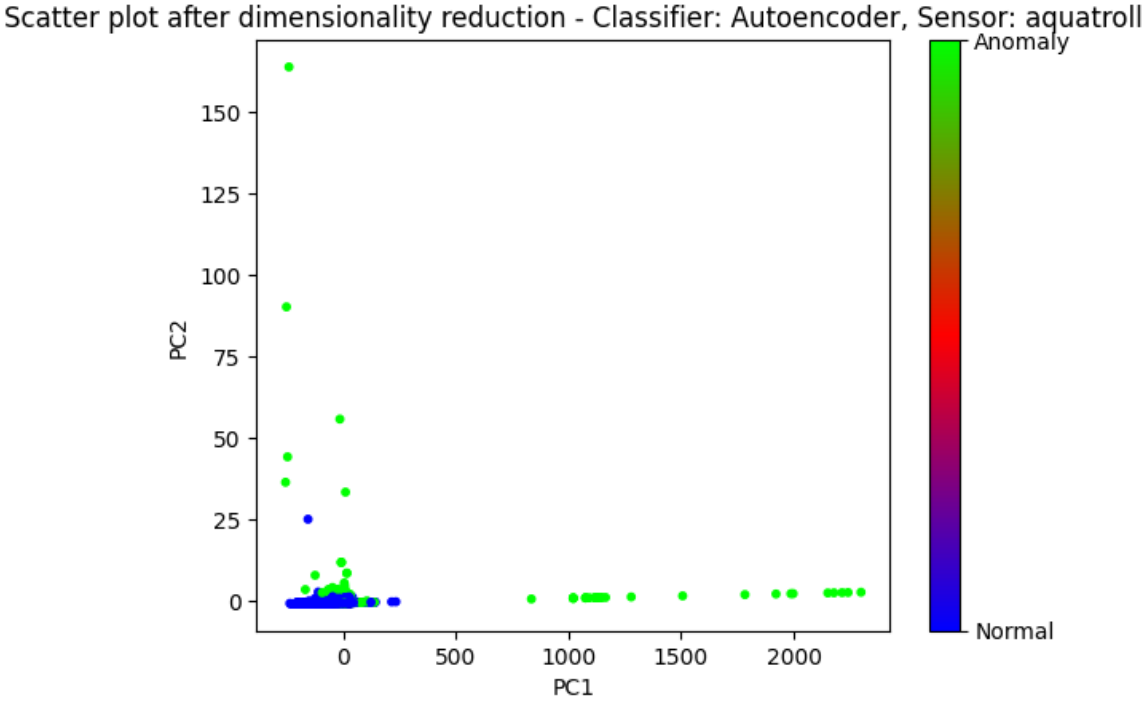


Εικόνα 6. Απόδοση του Classifier Autoencoder με Αισθητήρα Atmos

Αυτή η Εικόνα 6, δείχνει μια σημαντικά διαφορετική κατανομή σε σχέση με τα προηγούμενα.

Η πλειονότητα των δεδομένων (μπλε σημεία) παρουσιάζει μια σχεδόν κυκλική κατανομή στο κέντρο του διαγράμματος. Αυτό μπορεί να υποδηλώνει ότι οι παρατηρήσεις συγκεντρώνονται γύρω από ένα κεντρικό σημείο με ομοιόμορφη κατανομή σε όλες τις κατευθύνσεις.

Τα πράσινα σημεία, τα οποία ο αλγόριθμος έχει επισημάνει ως ανωμαλίες, βρίσκονται στην περίμετρο της κυκλικής περιοχής των μπλε σημείων. Αυτό μπορεί να σημαίνει ότι τα πράσινα σημεία είναι απομακρυσμένα από το κέντρο της πυκνότερης περιοχής των μπλε σημείων, πιθανώς δημιουργώντας μια απόκλιση από την κανονική κατανομή, γεγονός που τα καθιστά πιθανές ανωμαλίες.

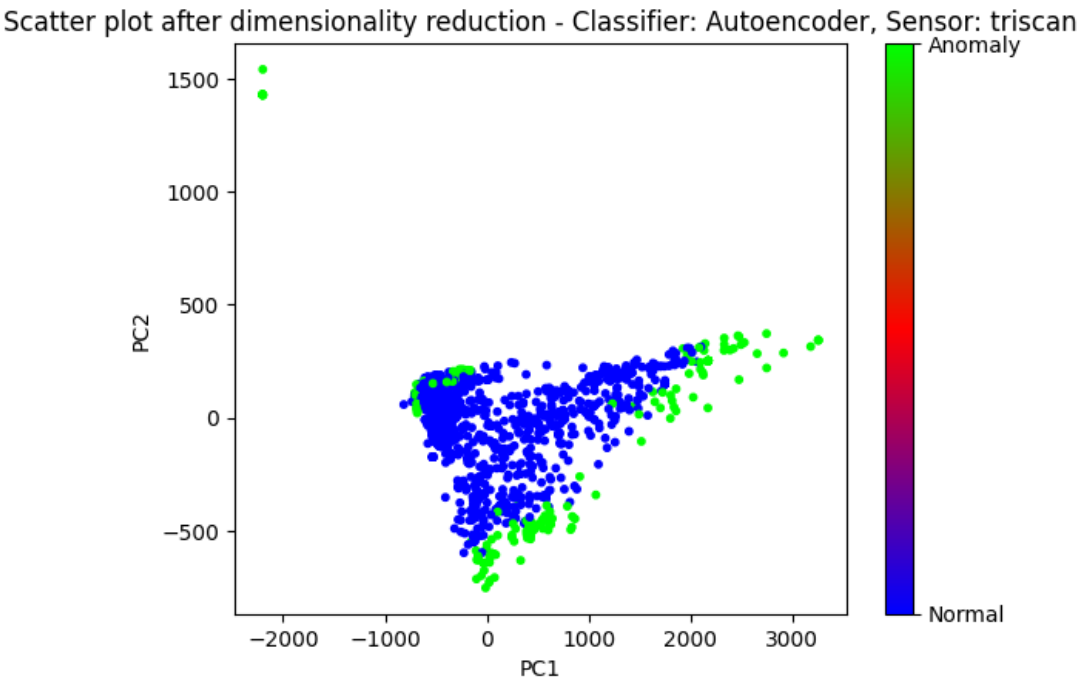


Εικόνα 7. Απόδοση του Classifier Autoencoder με Αισθητήρα Aquatroll

Η Εικόνα 7, δείχνει μια διαφορετική κατανομή σε σχέση με τα προηγούμενα.

Η πλειονότητα των δεδομένων (μπλε σημεία) φαίνεται να έχουν μια διασπορά κυρίως κατά μήκος του άξονα x (PC1), με λιγότερη διασπορά κατά μήκος του άξονα y (PC2). Αυτό μπορεί να υποδηλώνει ότι η διακύμανση των δεδομένων είναι μεγαλύτερη σε μια κατεύθυνση σε σχέση με την άλλη.

Τα πράσινα σημεία, που ο αλγόριθμος έχει επισημάνει ως ανωμαλίες, βρίσκονται εκτός της κύριας ομάδας των μπλε σημείων. Αυτά τα σημεία φαίνεται να αποκλίνουν σημαντικά από την κανονική κατανομή των μπλε σημείων, τόσο στον άξονα x (PC1) όσο και στον άξονα y (PC2).



Εικόνα 8. Απόδοση του Classifier Autoencoder με Αισθητήρα Triscan

Η Εικόνα 8, δείχνει έναν ξεκάθαρο διαχωρισμό μεταξύ των πιθανών ανωμαλιών (πράσινα σημεία) και των κανονικών δεδομένων (μπλε σημεία).

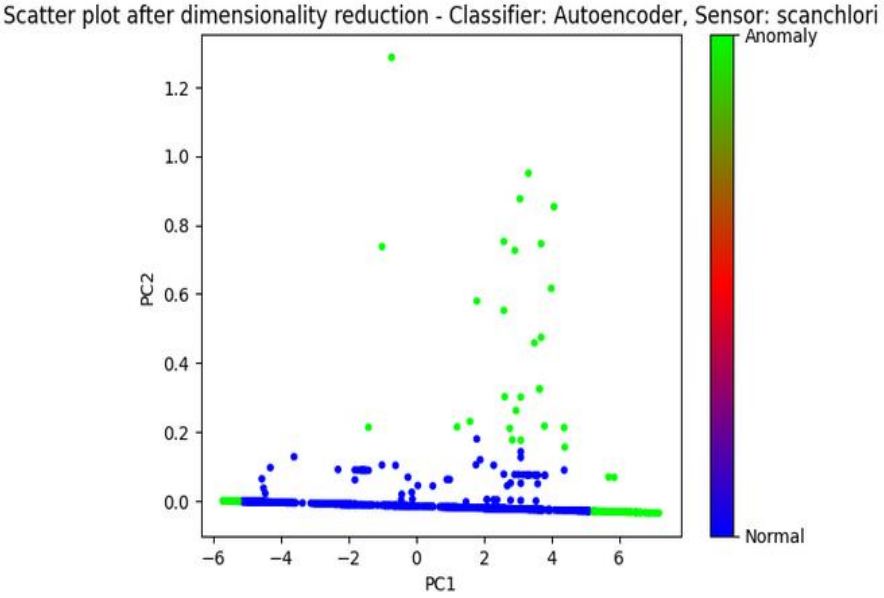
Οι ανωμαλίες είναι συγκεντρωμένες κυρίως σε δύο ξεχωριστές περιοχές: μία μεγαλύτερη συμπυκνωμένη ομάδα στην αριστερή πλευρά του διαγράμματος και μια μικρότερη στη δεξιά πλευρά. Το γεγονός ότι αυτές οι περιοχές είναι σαφώς διαχωρισμένες από το κυρίως σύννεφο των μπλε σημείων υποδηλώνει ότι αυτά τα σημεία πιθανότατα αντιπροσωπεύουν σημαντικές αποκλίσεις από την κανονικότητα.

Αντίθετα, τα κανονικά δεδομένα (μπλε σημεία) εμφανίζονται συγκεντρωμένα σε μια συμπυκνωμένη ομάδα που επεκτείνεται κυρίως κατά μήκος του άξονα x (PC1).

2. Παράδειγμα 2ο: Ίδιος classifier (k-NN) με 6 διαφορετικούς sensors, για σύγκριση μεταξύ διαφορετικών classifier για όλους τους sensors

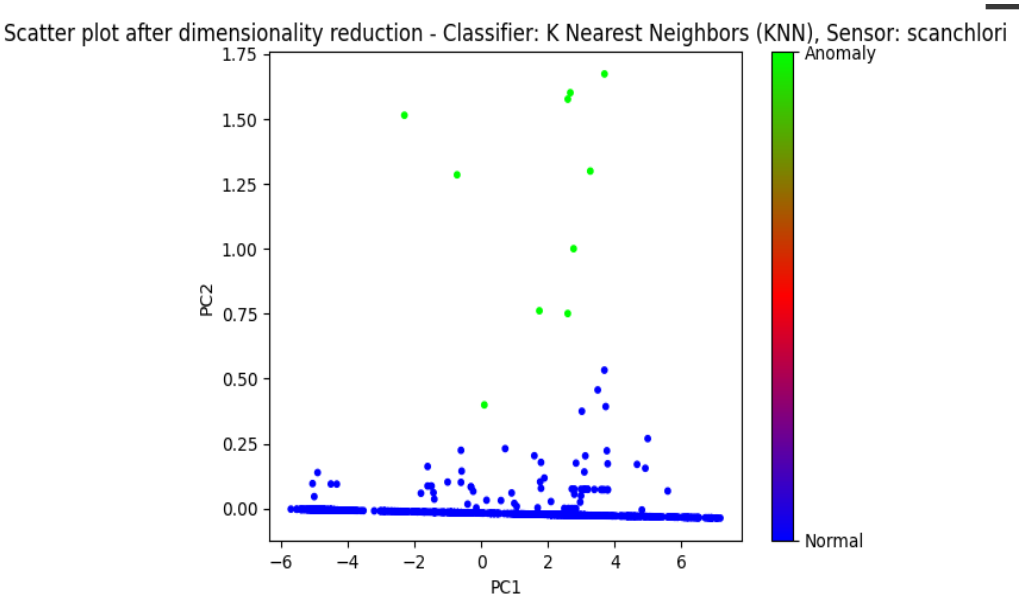
Αρχικά, λαμβάνουμε τα δεδομένα από τον αισθητήρα για τον οποίο θέλουμε να κάνουμε τη σύγκριση. Είναι σημαντικό να διασφαλίσουμε ότι έχουμε τα αντίστοιχα κανονικά και ανώμαλα δείγματα για τους δύο ταξινομητές που θέλουμε να συγκρίνουμε. Έπειτα εκπαιδεύουμε και εφαρμόζουμε τους δύο ταξινομητές στα αντίστοιχα δεδομένα του αισθητήρα, συνεπώς έτσι μπορούμε και ανακτούμε τις προβλέψεις για την κατηγοριοποίηση των δειγμάτων ως κανονικά ή ανώμαλα.

Στη συνέχεια, δημιουργούμε τα scatterplots για τους δύο classifiers. Κάθε δείγμα αναπαρίσταται ως ένα σημείο στο διάγραμμα, με τον άξονα X να αντιστοιχεί σε ένα χαρακτηριστικό και τον άξονα Y να αντιστοιχεί σε ένα διαφορετικό χαρακτηριστικό ή στο ίδιο χαρακτηριστικό που έχει μετατραπεί μέσω μειωμένης διάστασης. Οι κανονικές και ανώμαλες κατηγορίες δειγμάτων εμφανίζονται με διαφορετικά χρώματα και σε αυτή την περίπτωση
Εξετάζουμε τα δύο scatterplots για τον ίδιο sensor και αναλύουμε τις διαφορές στην κατανομή των κανονικών και ανώμαλων δειγμάτων όπως και τον τρόπο με τον οποίο τα δείγματα κατανέμονται στον χώρο των μειωμένων διαστάσεων και παρατηρούμε τη διαχωριστική ικανότητα των δύο classifiers. Εάν υπάρχουν σημαντικές διαφορές στην κατανομή ή στην ορατή απόσταση μεταξύ των κανονικών και ανώμαλων δειγμάτων, αυτό υποδηλώνει τη διαφορετική απόδοση των δύο classifiers στην ανίχνευση ανωμαλιών.



Εικόνα 9. Απόδοση του Classifier Autoencoder με Αισθητήρα Scanchlori

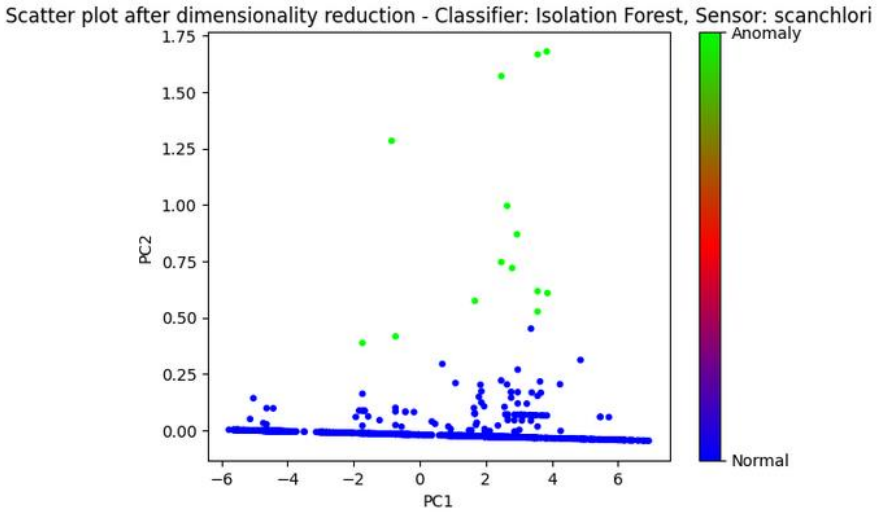
Autoencoder: Από την Εικόνα 9 βλέπουμε ότι ο Autoencoder κατάφερε να διαχωρίσει αρκετά καλά τις ανωμαλίες (πράσινα σημεία) από τα κανονικά δεδομένα (μπλε σημεία). Υπάρχει κάποια επικάλυψη, αλλά γενικά, οι ανωμαλίες είναι καλά οριοθετημένες σε μία ξεχωριστή περιοχή του χάρτη.



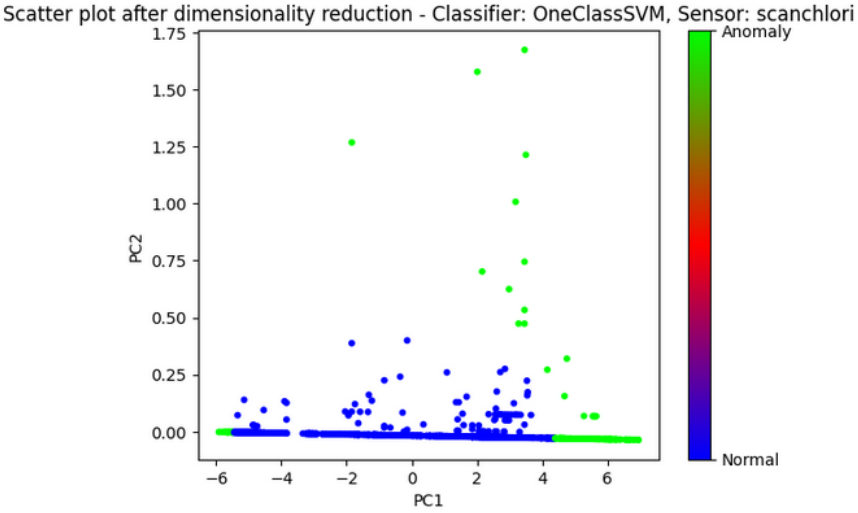
Εικόνα 10. Απόδοση του Classifier kNN με Αισθητήρα Scanchlori

K-Nearest Neighbors (k-NN): Από την Εικόνα 10, φαίνεται ότι ο **k-NN** έχει μεγαλύτερη δυσκολία στο να διαχωρίσει τις ανωμαλίες από τα κανονικά δεδομένα. Οι πράσινες κουκίδες (ανωμαλίες) είναι διάσπαρτες σε ολόκληρο το διάγραμμα, και πολλές φαίνεται να επικαλύπτονται με τις μπλε κουκίδες (κανονικά δεδομένα).

Συνολικά, με βάση αυτά τα δύο διαγράμματα, φαίνεται ότι ο Autoencoder είναι πιο αποτελεσματικός στην ανίχνευση ανωμαλιών σε σύγκριση με τον k-NN για αυτό το συγκεκριμένο σετ δεδομένων.



Εικόνα 11. Απόδοση του Classifier Isolation Forest με Αισθητήρα Scanchlori

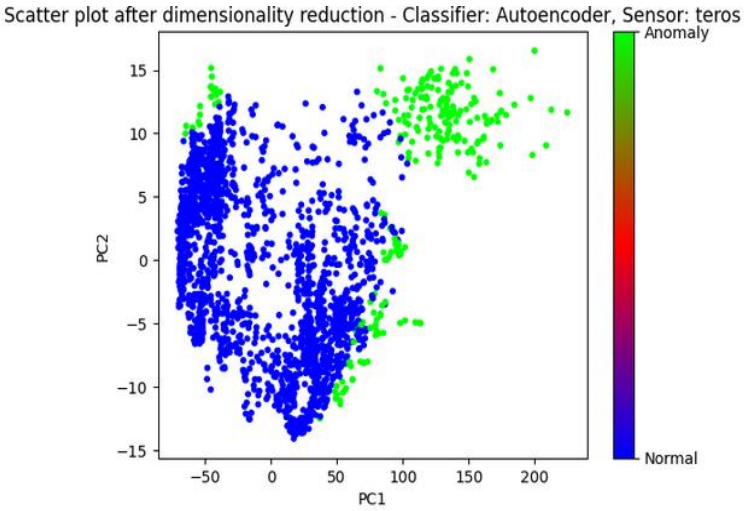


Εικόνα 12. Απόδοση του Classifier OneClassSVM με Αισθητήρα Scanchlori

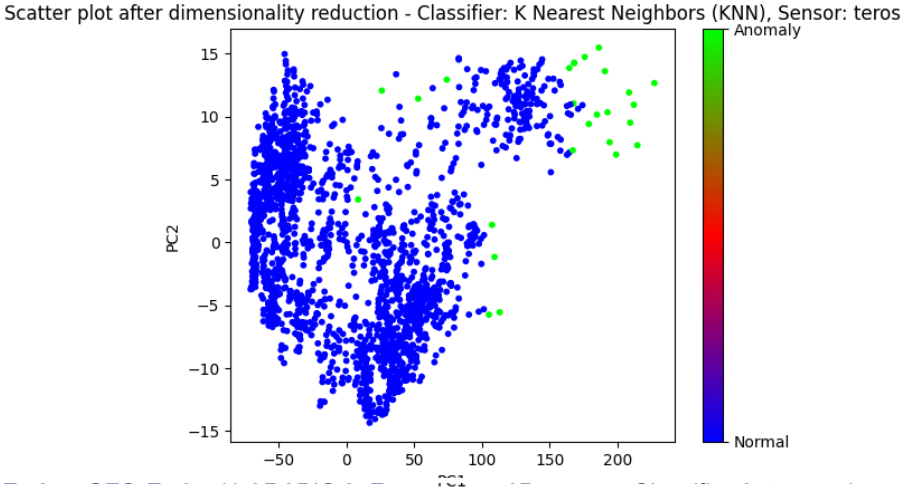
Isolation Forest Τα μπλε σημεία- κανονικά δεδομένα, είναι συγκεντρωμένα σε μια κεντρική σφαιρική περιοχή στον χώρο. Αυτή η πυκνή συγκέντρωση σημείων δείχνει ότι τα κανονικά δεδομένα έχουν παρόμοια χαρακτηριστικά σε αυτές τις δύο κύριες συνιστώσες (PC1 και PC2).Τα πράσινα σημεία-ανωμαλίες, βρίσκονται κυρίως στο περιθώριο της κεντρικής σφαιρικής περιοχής και σε κάποιες περιοχές εκτείνονται πέρα από αυτή. Ο διαχωρισμός τους από την πυκνότερη ομάδα των μπλε σημείων δείχνει ότι αυτά τα σημεία πιθανότατα αντιπροσωπεύουν σημαντικές αποκλίσεις από την κανονικότητα. Στην Εικόνα 11 παρουσιάζεται μια σαφή διαφοροποίηση μεταξύ των πιθανών ανωμαλιών και των κανονικών δεδομένων

One Class SVM. Τα μπλε σημεία- κανονικά δεδομένα, διασπείρονται ευρέως στον χώρο όπως φαίνεται στην Εικόνα 12. Δεν υπάρχει μια σαφής, πυκνή συγκέντρωση τους, αλλά φαίνονται να καλύπτουν σχεδόν ολόκληρο τον χώρο. Τα πράσινα σημεία-ανωμαλιες,

βρίσκονται διάσπαρτα μέσα στην περιοχή των μπλε σημείων. Δεν είναι συγκεντρωμένα σε μία συγκεκριμένη περιοχή, αλλά φαίνονται να είναι ενσωματωμένα με τα μπλε σημεία Στην Εικόνα 12 φαίνεται ότι η διακρίνεια μεταξύ των πιθανών ανωμαλιών και των κανονικών δεδομένων δεν είναι τόσο σαφής. Αυτό θα μπορούσε να καταστήσει δύσκολη την ανίχνευση ανωμαλιών σε αυτό το σύνολο δεδομένων με βάση μόνο αυτές τις δύο κύριες συνιστώσες.



Εικόνα 13. Απόδοση του Classifier Autoencoder με Αισθητήρα Teros

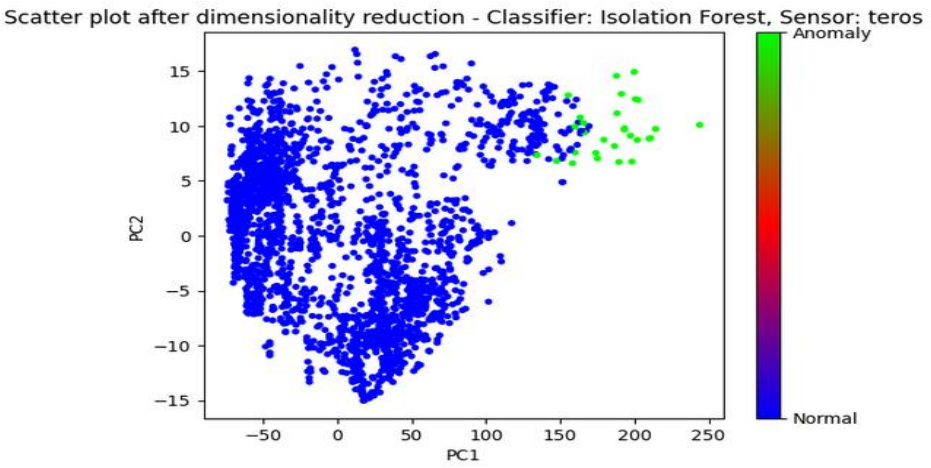


Εικόνα 14. Απόδοση του Classifier kNN με Αισθητήρα Teros

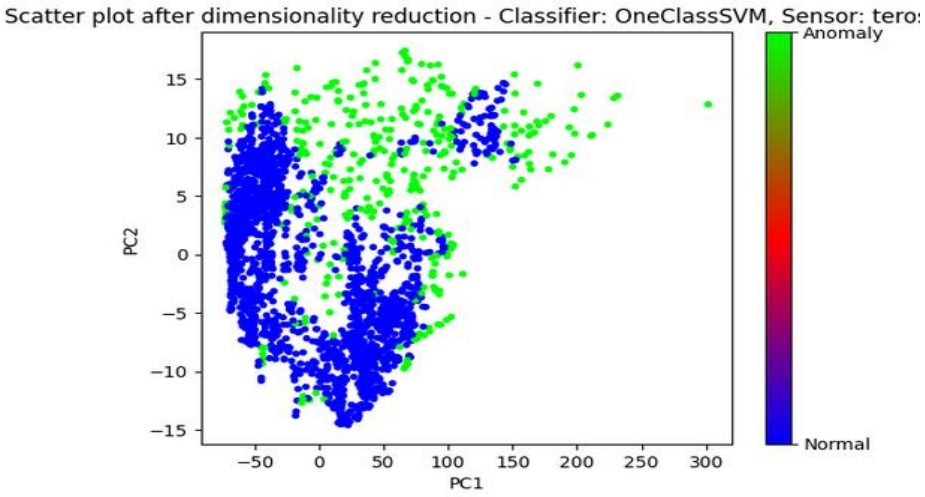
Autoencoder: Στην Εικόνα 13, ο Autoencoder φαίνεται να έχει εξαιρετική απόδοση στην ανίχνευση ανωμαλιών. Οι περισσότερες ανωμαλίες (πράσινα σημεία) βρίσκονται σε ευδιάκριτες, ξεκάθαρα οριοθετημένες περιοχές που δεν έχουν σχεδόν καθόλου επικάλυψη με τα κανονικά δεδομένα (μπλε σημεία). Αυτό σημαίνει ότι ο Autoencoder κατάφερε να αποκαλύψει μια δομή δεδομένων που διαχωρίζει αποτελεσματικά τις ανωμαλίες.

K-Nearest Neighbors (k-NN): Από την Εικόνα 14 , μπορούμε να δούμε ότι ο αλγόριθμος Isolation Forest έχει κάποια δυσκολία στον διαχωρισμό των ανωμαλιών από τα κανονικά δεδομένα. Οι πράσινες κουκίδες (ανωμαλίες) είναι αρκετά διάσπαρτες και έχουν σημαντική επικάλυψη με τις μπλε κουκίδες (κανονικά δεδομένα).

Συνολικά, βάσει αυτών των διαγραμμάτων, φαίνεται ότι ο Autoencoder είναι πιο αποτελεσματικός στην ανίχνευση ανωμαλιών συγκριτικά με τον Isolation Forest για αυτό το συγκεκριμένο σετ δεδομένων.



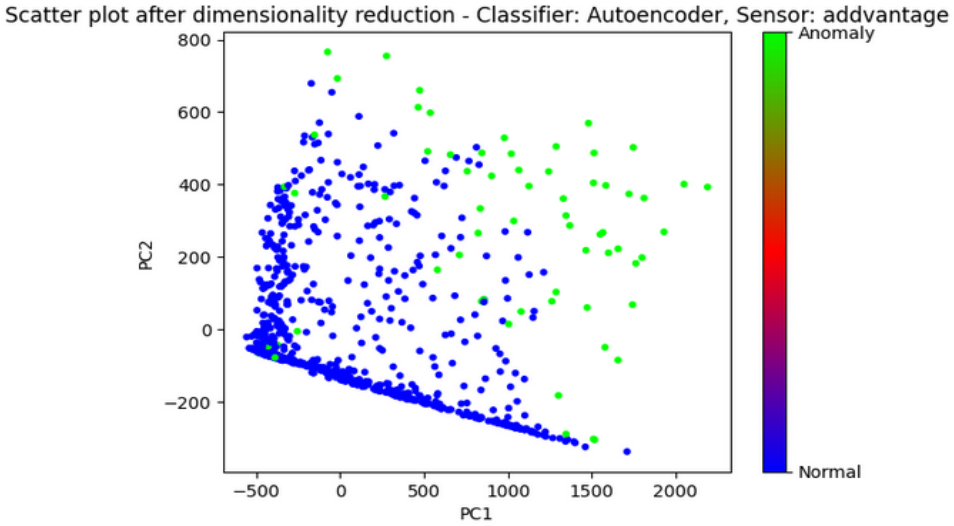
Εικόνα 16. Απόδοση του Classifier Isolation Isolation Forest με Αισθητήρα Teros



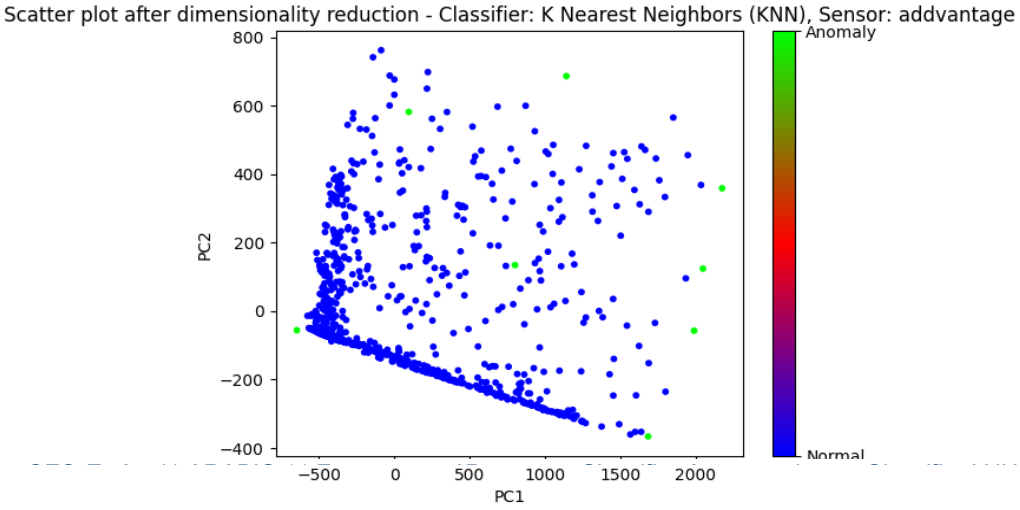
Εικόνα 15. Απόδοση του Classifier OneClassSVM με Αισθητήρα Teros

Isolation Forest. Τα μπλε σημεία-κανονικά δεδομένα, διασπείρονται κυρίως στην περιοχή που βρίσκεται στη μέση του διαγράμματος, δημιουργώντας μια σαφή και συμπαγή συγκέντρωση.Τα πράσινα σημεία-ανωμαλίες, βρίσκονται στις περιφέρειες της κυρίαρχης ομάδας των μπλε σημείων. Είναι εμφανές ότι τα πράσινα σημεία είναι λιγότερο συχνά και είναι σποραδικά διασπαρμένα. Η Εικόνα 15 δείχνει μια σαφή διαχωριστική γραμμή μεταξύ των πιθανών ανωμαλιών και των κανονικών δεδομένων. Οι ανωμαλίες είναι κατανεμημένες σε απομονωμένες περιοχές μακριά από την κυρίαρχη ομάδα των κανονικών δεδομένων. Αυτό σημαίνει ότι η ανίχνευση ανωμαλιών σε αυτό το σύνολο δεδομένων μπορεί να είναι πιο απλή, καθώς υπάρχει μια σαφής διαχωριστική γραμμή μεταξύ των δύο κλάσεων.

OneClassSVM. Τα μπλε σημεία διασπαρμένα. Δεν υπάρχει μια ενιαία περιοχή συγκέντρωσης τους, αλλά φαίνεται να υπάρχει μια κάποια πυκνότητα στο κέντρο του διαγράμματος. Τα πράσινα σημεία φαίνονται να είναι διασκορπισμένα στην περιοχή κοντά στα άκρα του διαγράμματος. Περισσότερο συγκεκριμένα, τα πράσινα σημεία είναι λιγότερο συχνά και φαίνονται να βρίσκονται σε απομονωμένες περιοχές μακριά από την κυρίαρχη πυκνότητα των μπλε σημείων. Στην Εικόνα 16 φαίνεται μια σαφή διαχωριστική γραμμή μεταξύ των πιθανών ανωμαλιών και των κανονικών δεδομένων. Ενώ τα κανονικά δεδομένα είναι διασπαρμένα, οι ανωμαλίες είναι κατανεμημένες σε απομονωμένες περιοχές, κυρίως στα άκρα του διαγράμματος.



Εικόνα 17. Απόδοση του Classifier Autoencoder με Αισθητήρα Addvantage

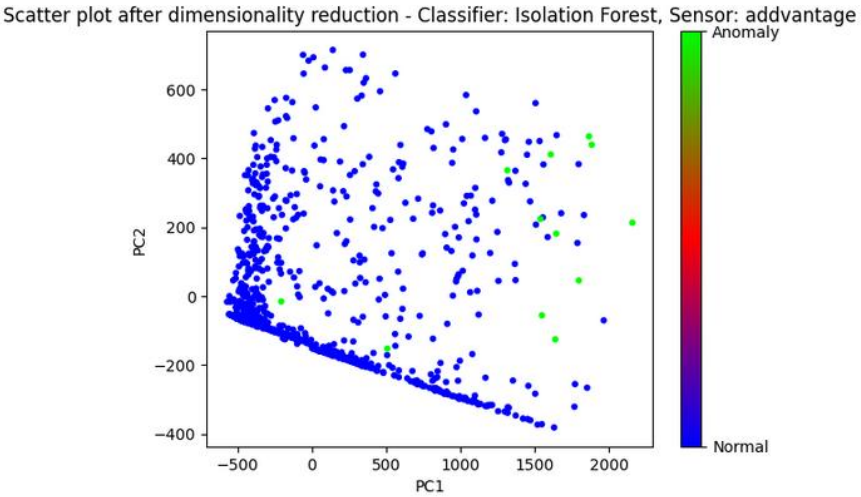


Εικόνα 18. Απόδοση του Classifier kNN με Αισθητήρα Addvantage

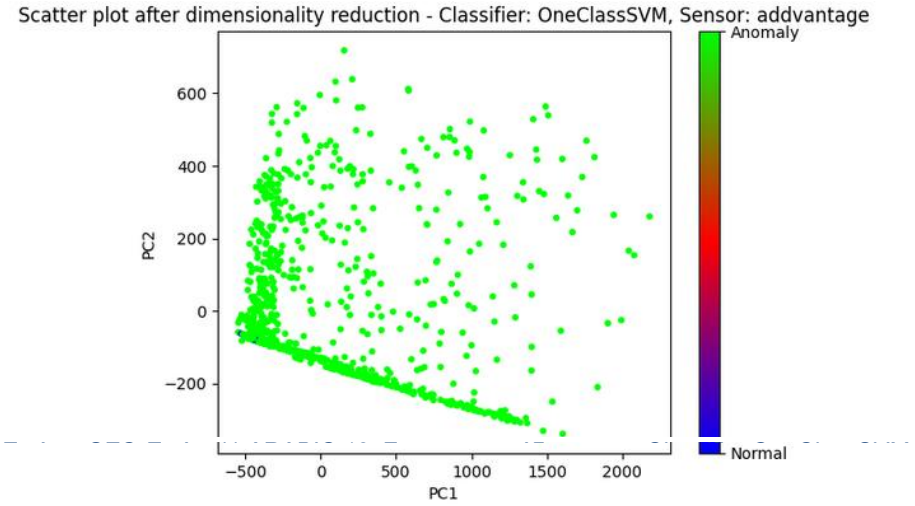
Autoencoder: Όπως και στις προηγούμενες Εικόνες, ο Autoencoder φαίνεται να αποδίδει καλά. Οι ανωμαλίες (πράσινα σημεία) εμφανίζονται ξεκάθαρα οριοθετημένες σε ειδικές περιοχές του χάρτη χαρακτηριστικών και έχουν περιορισμένη επικάλυψη με τα κανονικά δεδομένα (μπλε σημεία).

K Nearest Neighbors (k-NN): Στην Εικόνα 18, το k-NN φαίνεται να έχει δυσκολία στον διαχωρισμό των ανωμαλιών από τα κανονικά δεδομένα. Οι πράσινες κουκίδες (ανωμαλίες) είναι διάσπαρτες σε όλο τον χάρτη χαρακτηριστικών και εμφανίζουν σημαντική επικάλυψη με τα κανονικά δεδομένα (μπλε σημεία).

Βάσει αυτών των διαγραμμάτων, φαίνεται ότι ο Autoencoder είναι πιο αποτελεσματικός στην ανίχνευση ανωμαλιών σε σύγκριση με το k-NN για αυτό το συγκεκριμένο σετ δεδομένων.



Εικόνα 19. Απόδοση του Classifier Isolation Forest με Αισθητήρα Addvantage

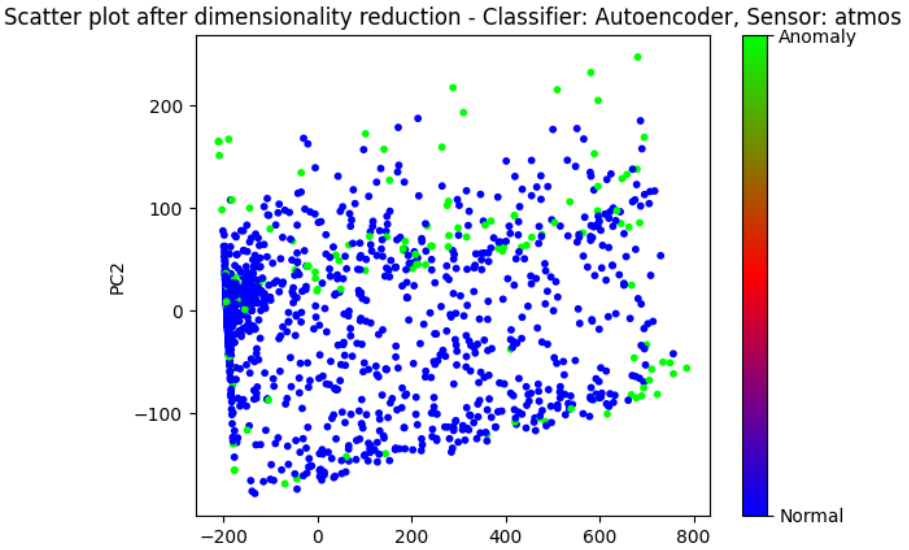


Εικόνα 20. Απόδοση του Classifier OneClassSVM με Αισθητήρα Addvantage

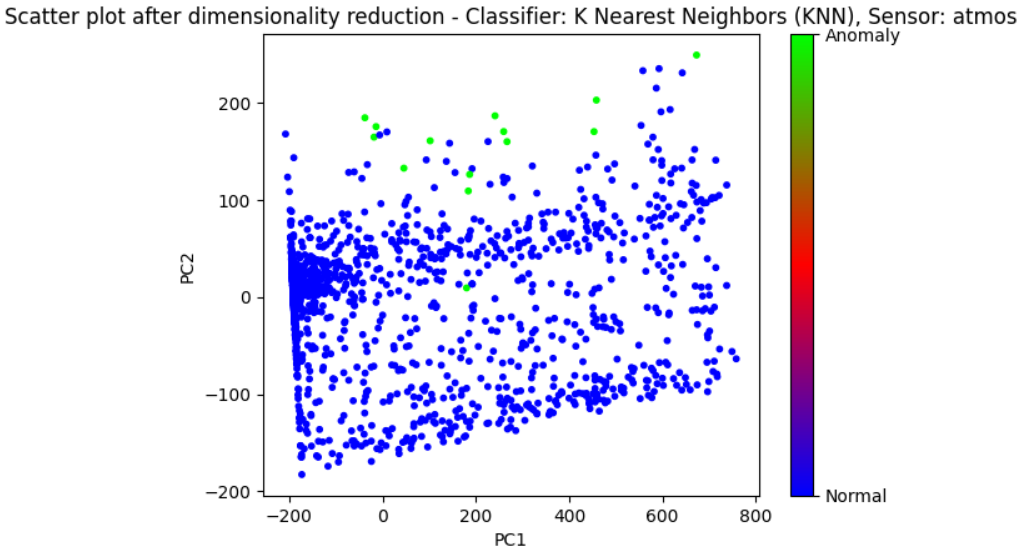
Isolation Forest.Τα μπλε σημεία φαίνονται να καταλαμβάνουν το μεγαλύτερο μέρος του διαγράμματος. Υπάρχει μια ενιαία περιοχή πυκνότητας στο κέντρο του διαγράμματος με κάποια μπλε σημεία που διασπείρονται στα άκρα. Τα πράσινα σημεία είναι διασκορπισμένα στο διάγραμμα, αλλά υπάρχει μια σημαντική συγκέντρωση τους στο κάτω μέρος του διαγράμματος, κάτι που δείχνει μια ξεκάθαρη περιοχή που περιέχει πιθανές ανωμαλίες. Το διάγραμμα δείχνει μια κεντρική συγκέντρωση των κανονικών δεδομένων, με τις ανωμαλίες να είναι πιο ξεκάθαρες και να συγκεντρώνονται κυρίως στο κάτω μέρος του διαγράμματος.

OneClassSVM.Τα μπλε σημεία είναι συγκεντρωμένα σε μια σχετικά στενή περιοχή στο άκρο του διαγράμματος, δημιουργώντας μια έντονη περιοχή πυκνότητας. Τα πράσινα σημεία είναι διασκορπισμένα σε όλο το διάγραμμα, αλλά με σημαντικές συγκεντρώσεις στα ανώτερα και κατώτερα άκρα του διαγράμματος, όπως και στις πλευρές. Οι περιοχές αυτές με τα πράσινα σημεία είναι αρκετά ξεκάθαρες και διακριτές από την κεντρική

περιοχή με τα μπλε σημεία. Το διάγραμμα παρουσιάζει μια ακριανή περιοχή όπου τα κανονικά δεδομένα είναι πυκνά και συγκεντρωμένα. Οι ανωμαλίες είναι κατανεμημένες σε όλο το διάγραμμα, αλλά υπάρχουν περιοχές με σημαντική συγκέντρωση ανωμαλιών, κυρίως στα άκρα και στις πλευρές του διαγράμματος.



Εικόνα 21. Απόδοση του Classifier Autoencoder με Αισθητήρα Atmos



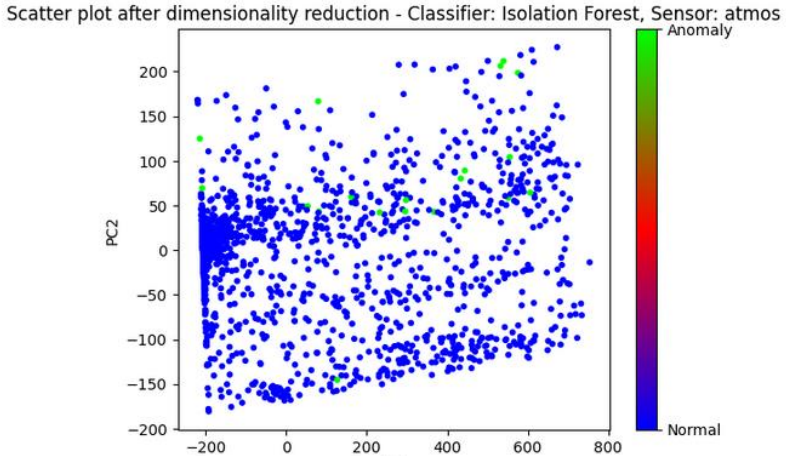
Εικόνα 22. Απόδοση του Classifier kNN με Αισθητήρα Atmos

Autoencoder: Η Εικόνα 21 αναπαριστά τα αποτελέσματα του Autoencoder. Ο Autoencoder προσπαθεί να μοντελοποιήσει τα κανονικά δεδομένα και εντοπίζει τις ανωμαλίες ως τα σημεία που διαφέρουν περισσότερο από το μοντέλο. Τα πράσινα σημεία αναπαριστούν τις ανωμαλίες και φαίνεται να υπάρχει κάποια δυσκολία στο διαχωρισμό τους από τα κανονικά δεδομένα (μπλε σημεία).

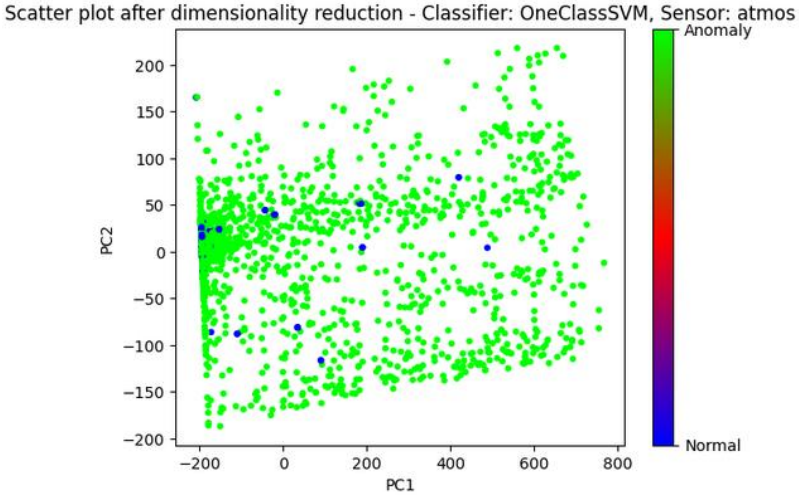
K Nearest Neighbors (k-NN): Στην Εικόνα 22, το k-NN φαίνεται να έχει κάποια επιτυχία στον διαχωρισμό των ανωμαλιών από τα κανονικά δεδομένα, αλλά είναι εμφανές ότι υπάρχει ακόμη σημαντική επικάλυψη ανάμεσα στα πράσινα και τα μπλε σημεία.

Και οι δύο αλγόριθμοι έχουν κάποια επιτυχία στο διαχωρισμό ανωμαλιών από κανονικά δεδομένα, αλλά κανένας δεν έχει αποδώσει τέλεια. Ο Autoencoder φαίνεται να έχει

περισσότερη δυσκολία στο διαχωρισμό των δεδομένων σε σχέση με το k-NN. Η επιλογή του καλύτερου αλγορίθμου εξαρτάται από τις απαιτήσεις του προβλήματος και τις ιδιότητες των δεδομένων, και μπορεί να χρειαστεί περαιτέρω πειραματισμό και ρύθμιση των παραμέτρων.



Εικόνα 23. Απόδοση του Classifier Isolation Forest με Αισθητήρα Atmos

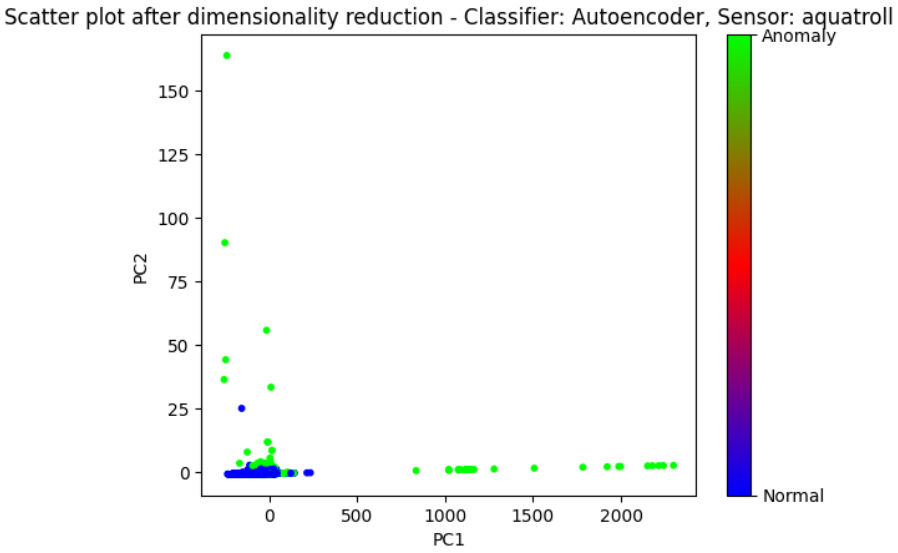


Εικόνα 24. Απόδοση του Classifier OneClassSVM με Αισθητήρα Atmos

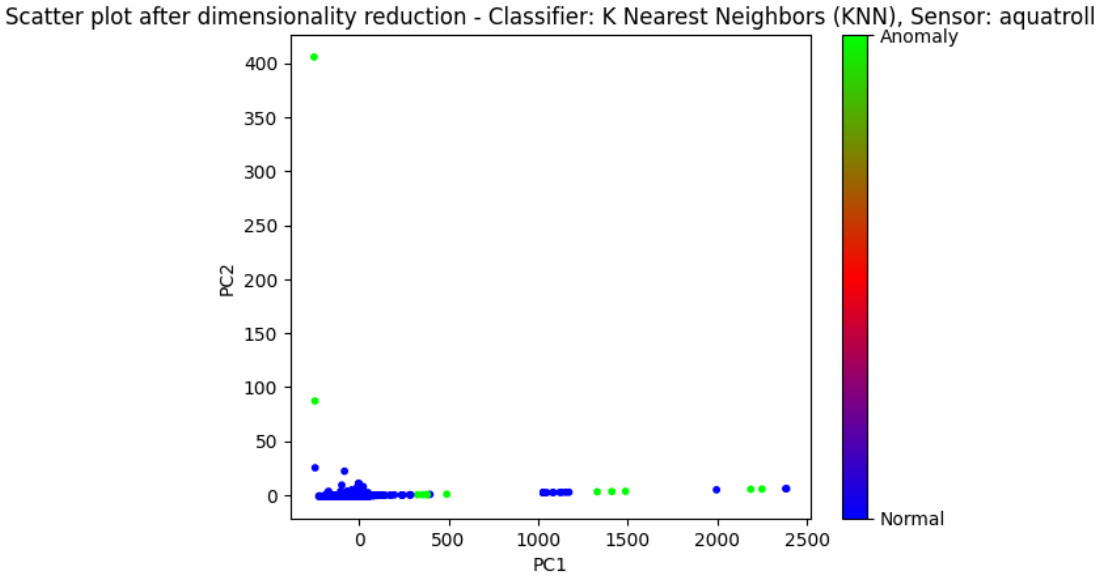
Isolation Forest.Τα μπλε σημεία είναι συγκεντρωμένα στη μέση του διαγράμματος, δημιουργώντας μια σχετικά στενή ομάδα ή "κύριο σύννεφο".Τα πράσινα σημεία βρίσκονται στα περιθώρια του διαγράμματος, μακριά από την κεντρική ομάδα των μπλε σημείων. Κάποια από αυτά τα σημεία είναι πολύ απομακρυσμένα από το κύριο σύννεφο των μπλε σημείων, ενώ άλλα είναι πιο κοντά αλλά ακόμη εκτός του. Σε αυτό το διάγραμμα, τα κανονικά δεδομένα (μπλε σημεία) είναι συγκεντρωμένα στην κεντρική περιοχή, ενώ οι ανωμαλίες (πράσινα σημεία) είναι διασκορπισμένες στα περιθώρια του διαγράμματος.

OneClassSVM. Τα μπλε σημεία είναι συγκεντρωμένα και δημιουργούν τρεις καθαρά οριοθετημένες ομάδες ή συστάδες. Κάθε ομάδα αποτελείται από σημεία που είναι πολύ κοντά μεταξύ τους. Τα πράσινα σημεία βρίσκονται εκτός αυτών των τριών συστάδων. Ορισμένα από αυτά τα σημεία είναι πολύ απομακρυσμένα από τις συστάδες, ενώ άλλα βρίσκονται μεταξύ των συστάδων. Στην Εικόνα 24, τα κανονικά δεδομένα (μπλε σημεία)

σχηματίζουν τρεις χαρακτηριστικές ομάδες. Οι ανωμαλίες (πράσινα σημεία) είναι σημεία που δεν ανήκουν σε καμία από αυτές τις τρεις κεντρικές συστάδες και μπορεί να είναι είτε αρκετά απομακρυσμένα από αυτές είτε μεταξύ τους.



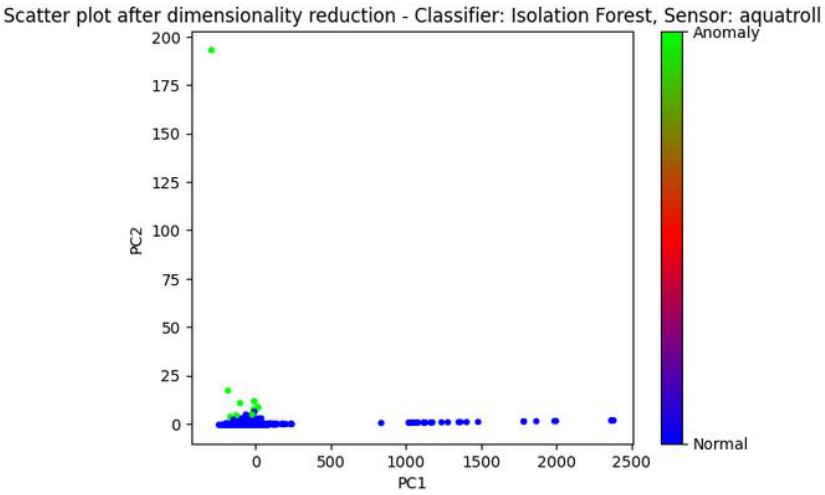
Εικόνα 25. Απόδοση του Classifier Autoencoder με Αισθητήρα Aquatroll



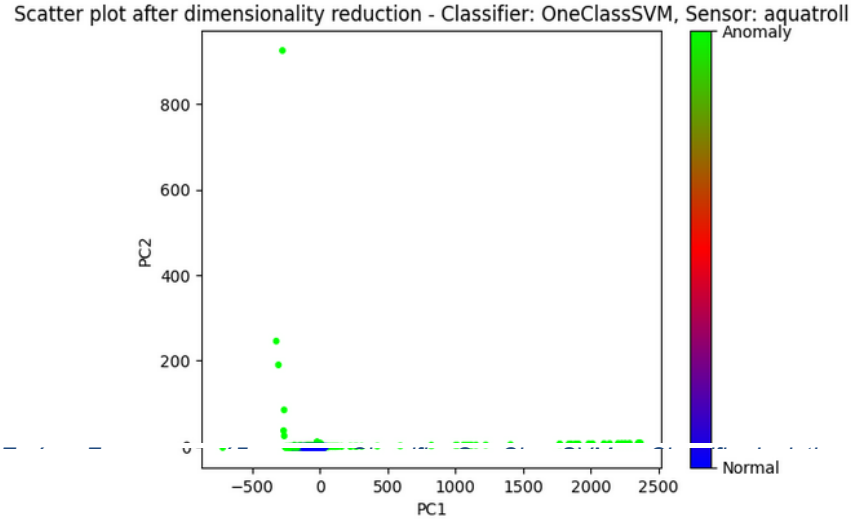
Εικόνα 26. Απόδοση του Classifier kNN με Αισθητήρα Aquatroll

Autoencoder: Το πρώτο διάγραμμα αναπαριστά τα αποτελέσματα του Autoencoder. Παρατηρούμε ότι είναι σε θέση να διαχωρίσει με κάποια επιτυχία τις ανωμαλίες (πράσινα σημεία) από τα κανονικά δεδομένα (μπλε σημεία). Ωστόσο, υπάρχει ακόμη σημαντική επικάλυψη μεταξύ των δύο κατηγοριών.
K Nearest Neighbors (k-NN): Στο δεύτερο διάγραμμα, ο k-NN φαίνεται να έχει κάποια δυσκολία στο διαχωρισμό των ανωμαλιών από τα κανονικά δεδομένα. Αυτό είναι εμφανές από τη σημαντική επικάλυψη μεταξύ των πράσινων και των μπλε σημείων.

Ο Autoencoder φαίνεται να επιδεικνύει καλύτερη απόδοση στο διαχωρισμό των δεδομένων σε σχέση με τον k-NN. Ο k-NN, παρόλο που έχει δυσκολία στο διαχωρισμό των δεδομένων, μπορεί να βελτιωθεί με τη ρύθμιση των παραμέτρων του ή με την εφαρμογή τεχνικών προεπεξεργασίας των δεδομένων.Η επιλογή του καλύτερου αλγορίθμου εξαρτάται από τις απαιτήσεις του προβλήματος και τις ιδιότητες των δεδομένων, και μπορεί να χρειαστεί περαιτέρω πειραματισμός και ρύθμιση των παραμέτρων.



Εικόνα 28. Απόδοση του Classifier Isolation Forest με Αισθητήρα Aquatroll

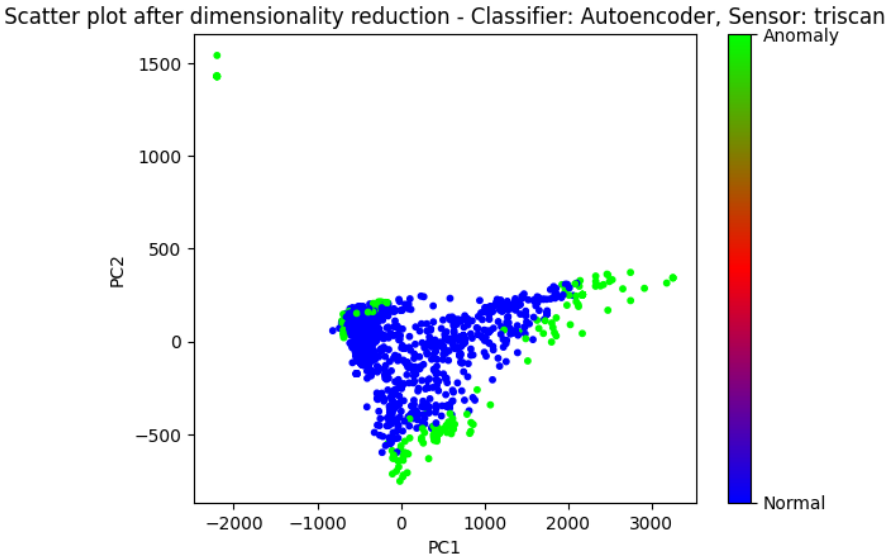


Εικόνα 27. Απόδοση του Classifier OneClassSVM με Αισθητήρα Aquatroll

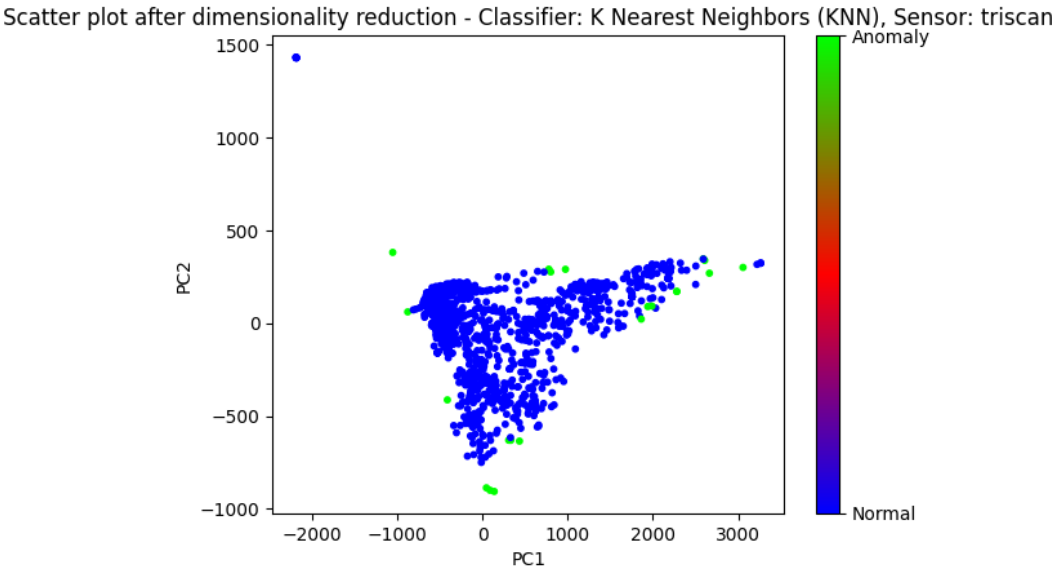
Isolation Forest. Τα μπλε σημεία είναι συγκεντρωμένα και δημιουργούν μια ευδιάκριτη συστάδα στο κέντρο του διαγράμματος. Τα σημεία αυτά είναι συγκεντρωμένα και σχετικά κοντά μεταξύ τους. Τα πράσινα σημεία βρίσκονται εκτός της κεντρικής συστάδας. Τα περισσότερα από αυτά τα σημεία είναι σχετικά απομακρυσμένα από την κεντρική συστάδα, ωστόσο υπάρχουν κάποια που βρίσκονται πολύ κοντά στη συστάδα αλλά χωρίς να ανήκουν εντός αυτής. Σε αυτό το διάγραμμα, τα κανονικά δεδομένα (μπλε σημεία) σχηματίζουν μία σαφή κεντρική συστάδα. Οι ανωμαλίες (πράσινα σημεία) είναι σημεία που δεν ανήκουν εντός της κεντρικής συστάδας και είναι κατανεμημένα σε διάφορες περιοχές γύρω από αυτήν.

OneClassSVM.Τα μπλε σημεία σχηματίζουν δύο κατακόρυφες συστάδες, μία δεξιά και μία αριστερά στο διάγραμμα. Η συστάδα από τη δεξιά πλευρά είναι πιο συμπαγής και συγκεντρωτική, ενώ η συστάδα από την αριστερή πλευρά φαίνεται να είναι λίγο πιο

διασκορπισμένη. Τα πράσινα σημεία είναι διασκορπισμένα στο κέντρο του διαγράμματος μεταξύ των δύο κατακόρυφων συστάδων των μπλε σημείων. Δεν σχηματίζουν κάποια συγκεκριμένη συστάδα και φαίνεται να είναι αρκετά απομακρυσμένα το ένα από το άλλο. Σε αυτό το διάγραμμα, τα κανονικά δεδομένα (μπλε σημεία) σχηματίζουν δύο κατακόρυφες συστάδες, με τη συστάδα στη δεξιά πλευρά να φαίνεται πιο συμπαγής σε σχέση με την αριστερή. Οι ανωμαλίες (πράσινα σημεία) βρίσκονται κυρίως στον κεντρικό χώρο ανάμεσα στις δύο συστάδες των μπλε σημείων και δεν φαίνεται να ακολουθούν κάποιο συγκεκριμένο πρότυπο συσταδοποίησης.



Εικόνα 29. Απόδοση του Classifier Autoencoder με Αισθητήρα Triscan

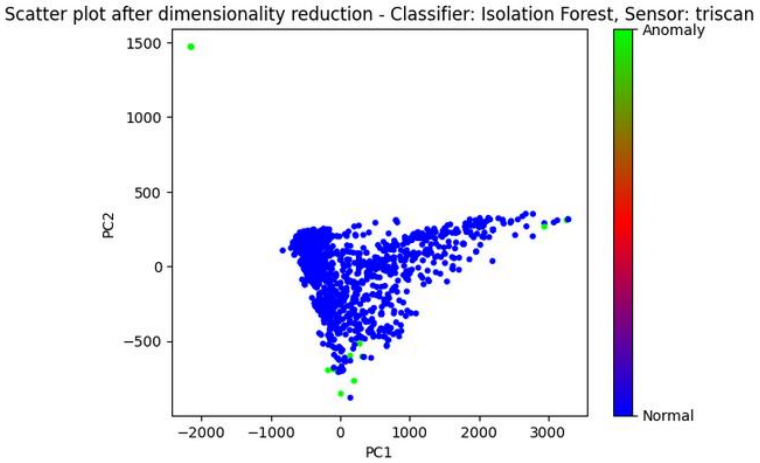


Εικόνα 30. Απόδοση του Classifier kNN με Αισθητήρα Triscan

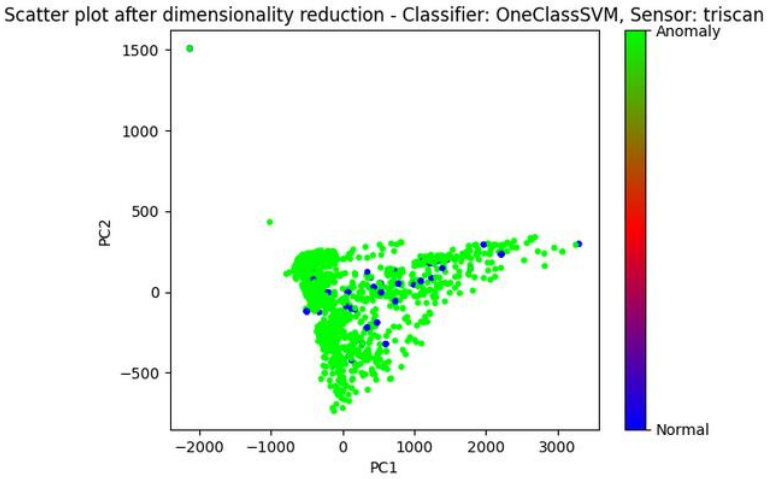
Autoencoder: Το πρώτο διάγραμμα αναπαριστά τα αποτελέσματα του Autoencoder. Παρατηρούμε ότι ο Autoencoder επιτυγχάνει μια καλή διάκριση μεταξύ των ανωμαλιών (πράσινα σημεία) και των κανονικών δεδομένων (μπλε σημεία). Υπάρχει λίγη επικάλυψη, αλλά γενικά, η πλειοψηφία των ανωμαλιών έχει διαχωριστεί επιτυχώς.

K Nearest Neighbors (k-NN): Στο δεύτερο διάγραμμα, ο k-NN δυσκολεύεται στην ορθή διάκριση των ανωμαλιών από τα κανονικά δεδομένα. Τα πράσινα και τα μπλε σημεία είναι σημαντικά επικαλυμμένα.

Ο Autoencoder φαίνεται να επιδεικνύει πολύ καλύτερη απόδοση στο διαχωρισμό των δεδομένων σε σχέση με τον k-NN. Ο k-NN έχει σημαντική επικάλυψη και θα χρειαζόταν περαιτέρω ρύθμιση των παραμέτρων ή προεπεξεργασία των δεδομένων για να βελτιώσει την απόδοσή του. Εδώ, ο Autoencoder φαίνεται να είναι πιο αποτελεσματικός στο να αναγνωρίζει και να διαχωρίζει τις ανωμαλίες από τα κανονικά δεδομένα.



Εικόνα 31. Απόδοση του Classifier Isolation Forest με Αισθητήρα Aquatroll



Εικόνα 32. Απόδοση του Classifier OneClassSVM με Αισθητήρα Aquatroll

Isolation Forest. Τα μπλε σημεία σχηματίζουν μία σχετικά συμπαγή και κυκλική συστάδα στο κέντρο του διαγράμματος. Η συγκέντρωσή τους δηλώνει ότι τα περισσότερα δεδομένα βρίσκονται σε αυτή την περιοχή. Τα πράσινα σημεία είναι διασκορπισμένα στο χώρο γύρω από την κεντρική συστάδα των μπλε σημείων. Αυτά τα σημεία βρίσκονται εκτός της κεντρικής πυκνής συστάδας και συνεπώς αντιπροσωπεύουν τις ανωμαλίες στο διάγραμμα. Σε αυτό το διάγραμμα, τα κανονικά δεδομένα (μπλε σημεία) σχηματίζουν μία κεντρική, συμπαγή και κυκλική συστάδα. Οι ανωμαλίες (πράσινα σημεία) είναι διασκορπισμένες γύρω από αυτή τη συστάδα, δείχνοντας ότι αποκλίνουν από την "τυπική" συμπεριφορά των δεδομένων.

OneClassSVM.Τα μπλε σημεία σχηματίζουν μια σχετικά εκτεταμένη και συνεχή κατανομή που καταλαμβάνει το μεγαλύτερο μέρος του διαγράμματος. Η διασπορά τους φαίνεται να είναι ομοιόμορφη και καλύπτει μεγάλο μέρος του χώρου.Τα πράσινα σημεία είναι λιγότερα σε αριθμό και βρίσκονται στις άκρες του διαγράμματος, περιμετρικά σε σχέση με την κύρια συστάδα των μπλε σημείων. Συγκεκριμένα, βρίσκονται προς τις γωνίες του

διαγράμματος. Σε αυτό το διάγραμμα, τα κανονικά δεδομένα (μπλε σημεία) σχηματίζουν μία εκτεταμένη και ομοιόμορφη κατανομή που καταλαμβάνει το μεγαλύτερο μέρος του διαγράμματος. Οι ανωμαλίες (πράσινα σημεία) είναι πιο διακριτές και βρίσκονται στις άκρες του χώρου, ειδικά κοντά στις γωνίες, δείχνοντας μια απόκλιση από την τυπική συμπεριφορά των δεδομένων.

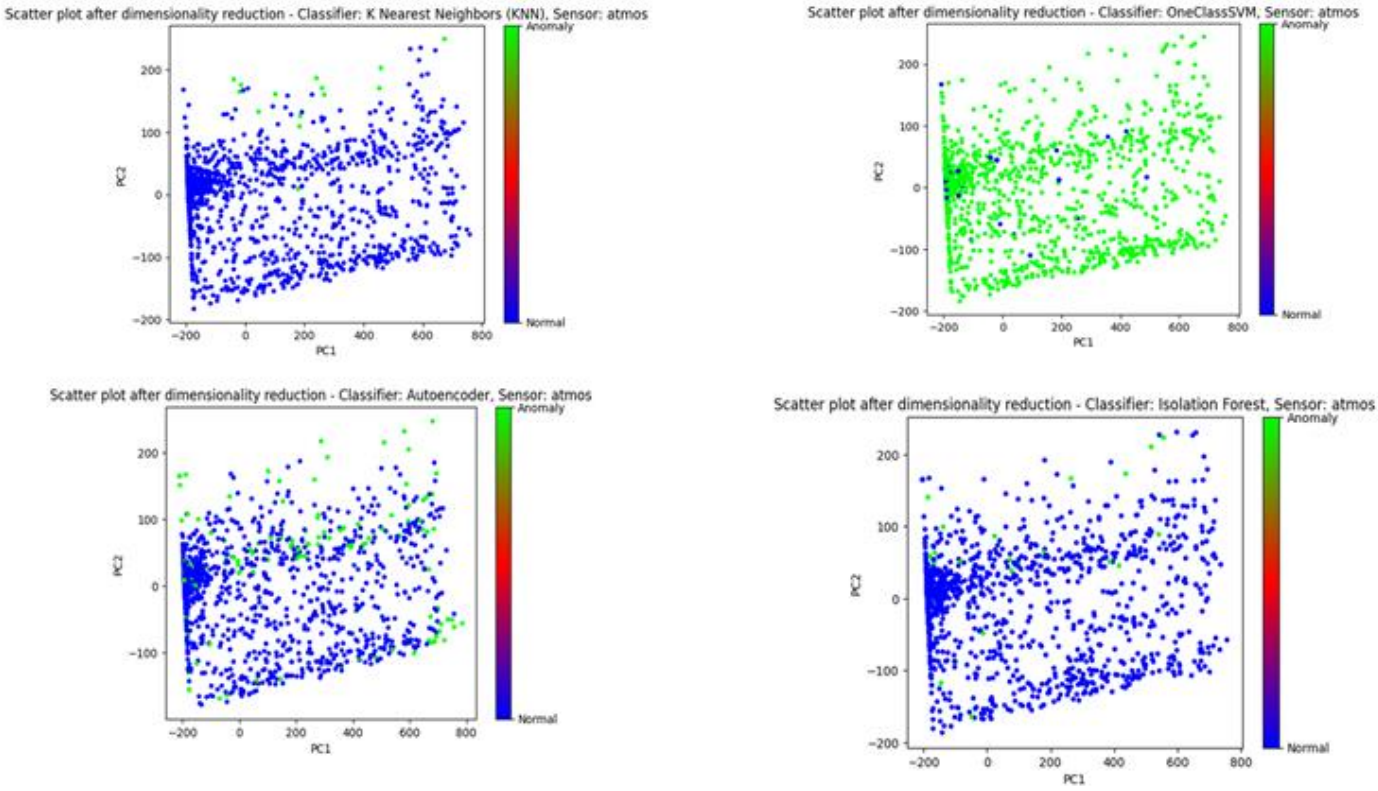
3. Παράδειγμα 3ο:Ίδιος sensor (atmos) για όλους τους classifiers.

Σε αυτή την περίπτωση κάθε scatterplot αναπαριστά την κατανομή των δεδομένων στον χώρο των δύο επιλεγμένων χαρακτηριστικών. Στην συνέχεια, επισημαίνουμε τα ανώμαλα σημεία που έχουν εντοπιστεί από τον classifier τα οποία εμφανίζονται με διαφορετικό χρώμα, ενώ τα κανονικά σημεία εμφανίζονται με άλλο χρώμα και σε αυτή την περίπτωση.

Το επόμενο βήμα είναι να συγκρίνουμε τα scatterplots μεταξύ των αισθητήρων και να παρατηρήσουμε την κατανομή των ανωμαλιών σε κάθε αισθητήρα αξιολογώντας τις διαφορές. Είναι σημαντικό να επισημάνουμε την ποσότητα, την κατανομή και την εξάπλωση των ανωμαλιών σε κάθε scatterplot.

Αναζητούμε τυχόν παρατηρήσεις ή πρότυπα που εμφανίζονται μόνο σε συγκεκριμένους αισθητήρες το οποίο μπορεί να υποδεικνύει ιδιαιτερότητες ή εξαιρέσεις στα δεδομένα που παράγονται από συγκεκριμένους αισθητήρες.

Μπορούμε να προβούμε στην ανίχνευση ανωμαλιών και τυχόν διαφορές μεταξύ των αισθητήρων. Η σύγκριση των scatterplots μεταξύ όλων των αισθητήρων για τον ίδιο classifier μπορεί να μας παρέχει ενδιαφέρουσες πληροφορίες σχετικά με την απόδοση του classifier σε διαφορετικά περιβάλλοντα ή σε διαφορετικές συνθήκες. Μπορούμε να αναδείξουμε δυνατά και αδύναμα σημεία του ταξινομητή, καθώς και να εξάγουμε γενικά συμπεράσματα σχετικά με την αξιοπιστία και την απόδοσή του σε διαφορετικούς αισθητήρες.



Εικόνα 33. Απόδοση όλων των Classiffiers με Αισθητήρα Atmos

1. **Autoencoder**: Το Autoencoder έχει καλή απόδοση στην ανίχνευση των ανωμαλιών, αλλά είναι αδύνατο στο να καταλάβει ορισμένες από τις επικαλυπτόμενες περιοχές μεταξύ των κανονικών και των ανωμαλιών (μπλε και πράσινα σημεία αντίστοιχα).
2. **Isolation Forest**: Το διάγραμμα του Isolation Forest δείχνει μια εξίσου καλή διάκριση, με μερικά πράσινα σημεία (ανωμαλίες) να εμφανίζονται εκτός των περιοχών των μπλε σημείων (κανονικά δεδομένα). Ωστόσο, υπάρχουν και μερικές ανωμαλίες που δεν ανιχνεύονται.
3. **K Nearest Neighbors (k-NN)**: Στο διάγραμμα για τον k-NN, υπάρχει σημαντική επικάλυψη μεταξύ των κανονικών δεδομένων και των ανωμαλιών. Αυτό δείχνει ότι ο k-NN έχει χαμηλή απόδοση στην ανίχνευση ανωμαλιών σε αυτό το συγκεκριμένο σύνολο δεδομένων.
4. **One-Class SVM**: Το διάγραμμα για τον One-Class SVM δείχνει επίσης έντονη επικάλυψη μεταξύ των ανωμαλιών και των κανονικών δεδομένων, παρόμοια με αυτή που βλέπουμε με τον k-NN.

Ο Autoencoder και το Isolation Forest φαίνεται να παρέχουν τα καλύτερα αποτελέσματα σε αυτό το συγκεκριμένο σύνολο δεδομένων, διαχωρίζοντας επιτυχώς την πλειοψηφία των ανωμαλιών από τα κανονικά δεδομένα. Το k-NN και το One-Class SVM, από την άλλη, φαίνεται να έχουν χαμηλή απόδοση, καθώς τα ανώμαλα και τα κανονικά δεδομένα επικαλύπτονται σημαντικά στα αντίστοιχα διαγράμματα.

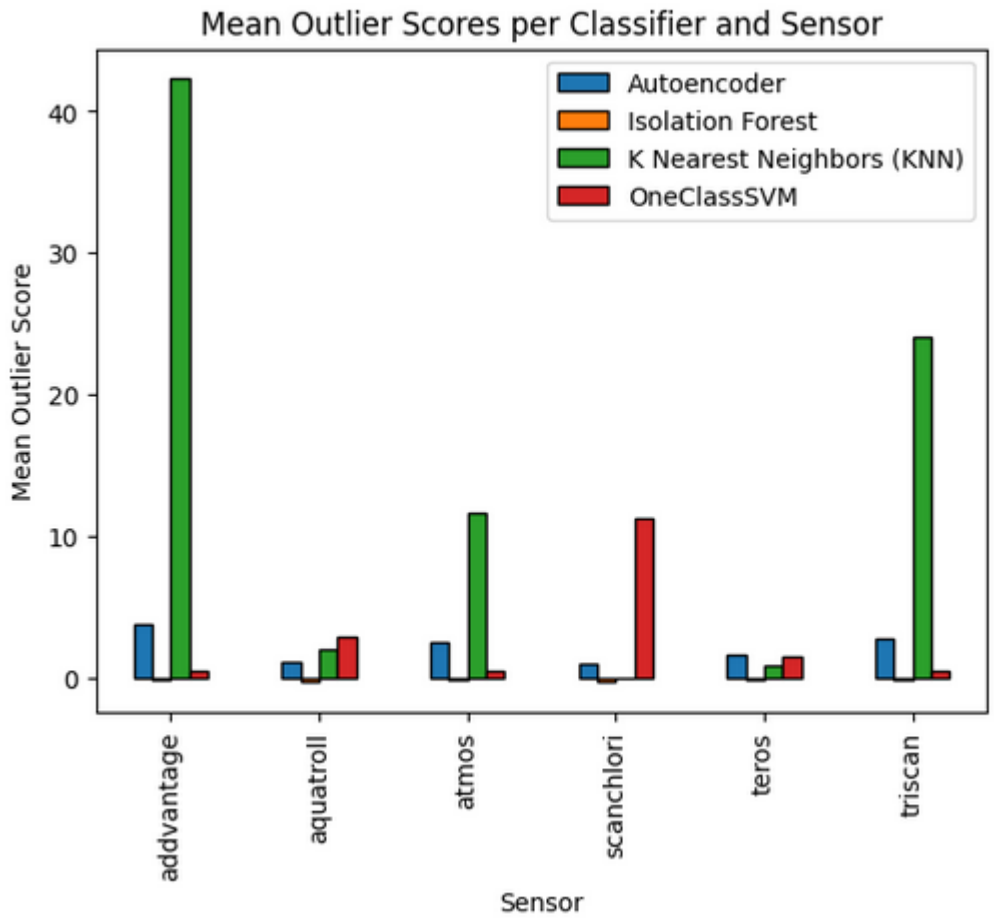
Υπολογισμός outlier scores για κάθε classifier:

Ο λόγος που υπολογίσαμε τα σκορ ανωμαλίας (outlier scores) για κάθε classifier είναι για να μας παρέχουν μια μετρική για το πόσο πιθανό είναι ένα δεδομένο παρατηρούμενο σημείο να είναι ανώμαλο. Τα σκορ ανωμαλίας υπολογίζονται από τους classifiers ανωμαλίας με βάση τα χαρακτηριστικά των παρατηρούμενων σημείων και το μοντέλο που έχουν εκπαιδεύσει. Ο υπολογισμός των σκορ ανωμαλίας μας βοηθά να κατανοήσουμε ποια σημεία θεωρούνται πιθανότερα ως ανώμαλα από τον κάθε ταξινομητή.

Αυτό μας δίνει τη δυνατότητα να συγκρίνουμε την απόδοση των διαφορετικών classifier στην ανίχνευση ανωμαλιών.

Τα σκορ ανωμαλίας μας βοηθούν να κατανοήσουμε ποια σημεία θεωρούνται πιθανότερα ως ανώμαλα από τον κάθε ταξινομητή, δίνοντάς μας τη δυνατότητα να συγκρίνουμε την απόδοση των διαφορετικών ταξινομητών στην ανίχνευση ανωμαλιών. Η ανίχνευση ανωμαλιών παίζει επίσης κρίσιμο ρόλο σε εργασίες όπως η παρακολούθηση της χρήσης πιστωτικών καρτών ή κινητών τηλεφώνων για τον εντοπισμό ξαφνικής αλλαγής στο πρότυπο χρήσης που μπορεί να υποδηλώνει απάτη, όπως κλεμμένη κάρτα ή κλεμμένος χρόνος ομιλίας.[5]

Συνολικά, ο υπολογισμός των σκορ ανωμαλίας μας βοηθά να αξιολογήσουμε την απόδοση των classifiers και να κατανοήσουμε την ικανότητά τους να ανιχνεύουν ανωμαλίες στα δεδομένα μας.



Εικόνα 34. Συγκριση βαθμολογίας απόδοσης Classifiers ανά Αισθητήρα

Στην Εικόνα 34, κάθε μπάρα αντιπροσωπεύει έναν διαφορετικό αισθητήρα, και τα διάφορα χρώματα σε κάθε μπάρα αντιπροσωπεύουν διάφορους ταξινομητές. Αυτό διευκολύνει την σύγκριση του πώς διάφοροι αισθητήρες αποδίδουν κάτω από τον ίδιο ταξινομητή.

Σε αυτή την απεικόνιση, οι βαθμολογίες από διάφορους ταξινομητές δεν είναι απαραίτητα στην ίδια κλίμακα ή συγκρίσιμες λόγω διαφορών στις μεθόδους που χρησιμοποιούνται για τον υπολογισμό των βαθμολογιών. Ως εκ τούτου, υψηλότερες βαθμολογίες δεν υποδηλώνουν απαραίτητα περισσότερες ανωμαλίες, και διάφοροι ταξινομητές μπορεί να κατατάξουν τις ανωμαλίες διαφορετικά.

Το διάγραμμα θα είναι πιο χρήσιμο για τη σύγκριση του πώς διάφοροι αισθητήρες αποδίδουν υπό τον ίδιο ταξινομητή, παρά για τη σύγκριση των ταξινομητών μεταξύ τους.

Όπως αναφέραμε, το συγκεκριμένο διάγραμμα μπάρας παρουσιάζει την μέση βαθμολογία ανωμαλιών για κάθε αισθητήρα, για διάφορους ταξινομητές. Κάθε αισθητήρας εκπροσωπείται από έναν συνδυασμό μπαρών, με κάθε μπάρα να αντιπροσωπεύει τη μέση βαθμολογία ανωμαλίας που δημιουργήθηκε από έναν συγκεκριμένο ταξινομητή.

Από το διάγραμμα, μπορούμε να δούμε ότι για κάθε αισθητήρα (atmos, aquatroll, triscan, scanchlori, teros, addvantage), ο αλγόριθμος K Nearest Neighbors (k-NN) παράγει την υψηλότερη μέση βαθμολογία ανωμαλίας. Αυτό δηλώνει ότι, στο σύνολο των δεδομένων των αισθητήρων, ο k-NN εντοπίζει τη μεγαλύτερη αριθμητική ποσότητα σημείων ως "ανώμαλα", σε σύγκριση με τους άλλους αλγορίθμους.

Επιπλέον, η διαφορά μεταξύ των μέσων βαθμολογιών ανωμαλίας που παράγει ο k-NN σε σχέση με τους υπόλοιπους αλγορίθμους φαίνεται να είναι σημαντική, υποδεικνύοντας ένα σαφές πλεονέκτημα για τον k-NN, τουλάχιστον όσον αφορά την απόδοση στα συγκεκριμένα σύνολα δεδομένων.

Ωστόσο, είναι σημαντικό να θυμόμαστε ότι αυτή η ανάλυση βασίζεται μόνο σε μέσες τιμές και δεν λαμβάνει υπόψη τη διακύμανση των βαθμολογιών ή την ποιότητα των προβλέψεων (δηλαδή, εάν οι ανωμαλίες που εντοπίστηκαν είναι πραγματικά ανωμαλίες). Επίσης, αξίζει να σημειώσουμε ότι διαφορετικοί αλγόριθμοι μπορεί να είναι πιο ή λιγότερο αποτελεσματικοί ανάλογα με το είδος των ανωμαλιών που αναζητούμε, οπότε αυτά τα αποτελέσματα μπορεί να μην είναι γενικά εφαρμόσιμα σε άλλες περιπτώσεις.

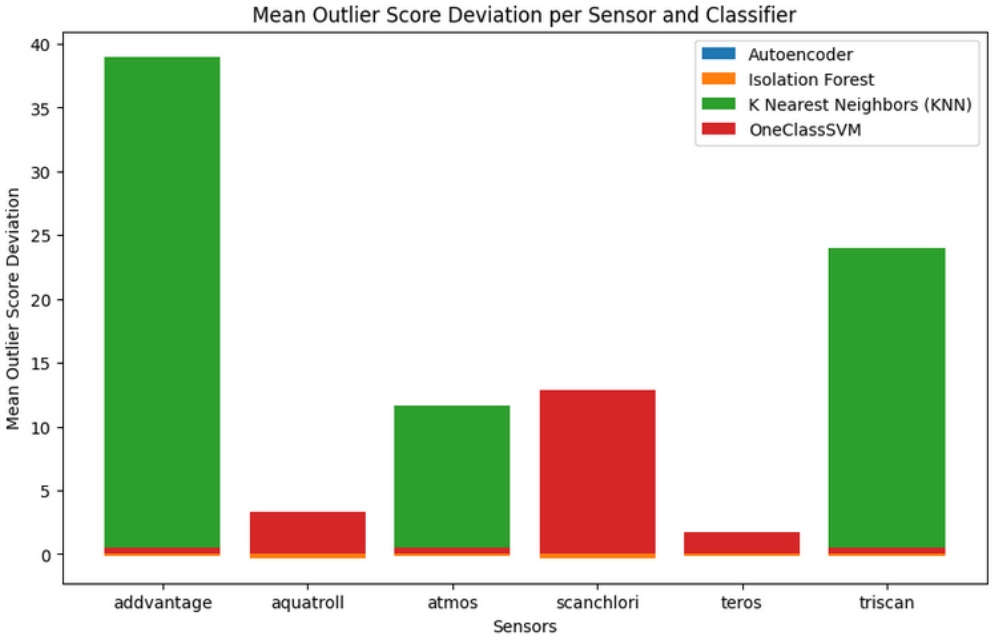
Υπολογισμός Distance-Based Metrics

Μετρικές όπως η μέση απόκλιση ή η τυπική απόκλιση μπορούν να χρησιμοποιηθούν για να αξιολογήσουν πόσο μακριά από τη "νορμαλιότητα" βρίσκονται τα σημεία που χαρακτηρίζονται ως ανώμαλα από τον αλγόριθμο.

Για να υπολογίσουμε μια distance-based μετρική, μπορούμε να υποθέσουμε ότι τα σκορ των ανωμαλιών που παράγονται από τους αλγορίθμους μας αντιπροσωπεύουν "αποστάσεις" από τη νορμαλιότητα. Σε αυτή την περίπτωση, μπορούμε να υπολογίσουμε τη μέση απόκλιση των σκορ των ανωμαλιών για κάθε αισθητήρα.

Σκοπός της Distance-Based Μετρικής: Στη μη επιβλεπόμενη μάθηση, κυρίως σε προβλήματα ανίχνευσης ανωμαλιών, μία κοινή προσέγγιση είναι να υπολογίζουμε την "απόσταση" της κάθε παρατήρησης από το "κέντρο" ή από τις άλλες παρατηρήσεις. Παρατηρήσεις που έχουν μεγάλη απόσταση συχνά θεωρούνται ως ανωμαλίες.

1. **Βασισμένο στον Κώδικα**:
    - Ο **k-NN (K-Nearest Neighbors)** είναι ένα παράδειγμα distance-based αλγορίθμου. Υπολογίζει την απόσταση της κάθε παρατήρησης από τους K πλησιέστερους γείτονές της στον χώρο των χαρακτηριστικών. Αν η μέση απόσταση από αυτούς τους γείτονες είναι πάνω από ένα κατώφλι, η παρατήρηση θεωρείται ανώμαλη.
    - Οι αποστάσεις αυτές, οι οποίες υπολογίζονται από τον k-NN, επηρεάζουν τα outlier scores. Όταν οι αποστάσεις είναι μεγάλες, τα scores είναι υψηλά.
2. **Χρησιμότητα**: Η distance-based μετρική είναι χρήσιμη γιατί προσφέρει έναν άμεσο τρόπο να αξιολογήσουμε το πόσο "κοντά" ή "μακριά" είναι μια παρατήρηση σε σχέση με άλλες παρατηρήσεις.



Εικόνα 35. Μέση απόκλιση βαθμολογίας ανά Αισθητήρα και Classifier

Παρατηρώντας την Εικόνα 35:

1. **Αναπαράσταση**:
   - Τα διαφορετικά χρώματα στην μπάρα δείχνουν το μέσο σκορ ανωμαλίας ανά αισθητήρα για κάθε classifier (Isolation Forest, k-NN, OneClassSVM, Autoencoder).
   - Η ανώτερη μπάρα για κάθε αισθητήρα δείχνει ποιος αλγόριθμος βρήκε την μεγαλύτερη μέση απόσταση (ή το μεγαλύτερο outlier score) για τα δεδομένα αυτού του αισθητήρα.
2. **Συμπεράσματα**:
   - Σε όλους τους αισθητήρες, το k-NN δείχνει την υψηλότερη μέση τιμή ανωμαλίας.
   - Αυτό μπορεί να σημαίνει ότι το k-NN εντοπίζει μεγαλύτερες αποστάσεις μεταξύ των παρατηρήσεων σε σύγκριση με τους άλλους αλγορίθμους, ειδικά σε αυτά τα συγκεκριμένα δεδομένα.
   - Τα υπόλοιπα αλγόριθμοι έχουν μικρότερες τιμές ανωμαλίας σε σύγκριση με το k-NN, με το Autoencoder συχνά να βρίσκεται στην τελευταία θέση, δείχνοντας τη χαμηλότερη μέση τιμή ανωμαλίας.
3. **Σημειώσεις**:
   - Ένα υψηλότερο score δεν σημαίνει απαραίτητα ότι ένας αλγόριθμος είναι "καλύτερος". Μπορεί απλώς να ανιχνεύει ανωμαλίες με διαφορετικό τρόπο ή να είναι πιο ευαίσθητος σε συγκεκριμένους τύπους ανωμαλιών. Η απόδοση πρέπει πάντα να επαληθεύεται με πραγματικά δεδομένα και άλλες μετρικές.

Κεφάλαιο 6: Συμπεράσματα

Η παρούσα διπλωματική εργασία εστίασε στη σημασία της ανίχνευσης ανωμαλιών σε δεδομένα αισθητήρων, καθώς αυτή η διαδικασία διαδραματίζει έναν αποφασιστικό ρόλο στην προστασία των συστημάτων από μη φυσιολογικές συμπεριφορές.

Κατά τη διάρκεια της μελέτης, εξετάστηκαν διάφορες τεχνικές ανίχνευσης ανωμαλιών, με ιδιαίτερη έμφαση στις τεχνικές μηχανικής μάθησης όπως οι Μηχανές Διανυσμάτων Υποστήριξης (SVM), το Isolation Forest και οι Κ Πλησιέστεροι Γείτονες.

Τα αποτελέσματα των πειραμάτων κατέδειξαν ότι οι διάφοροι ταξινομητές παρουσιάζουν διαφορές στην απόδοσή τους, ανάλογα με τους αισθητήρες που χρησιμοποιούνταν. Αυτό επιτρέπει την κατανόηση των προτύπων απόδοσης, των τάσεων και των συσχετίσεων που διαμορφώνονται μεταξύ των ταξινομητών και των αισθητήρων.

Στην συγκεκριμένη περίπτωση δεν μπορεί να κριθεί ακριβώς η απόδοση γιατί ο κάθε ταξινομητής ανιχνευει διαφορετικές ανωμαλίες, και επομένως, οι μετρήσεις απόδοσης μπορεί να είναι παραπλανητικές αν συγκρίνονται απευθείας μεταξύ διαφορετικών μοντέλων. Οι ανωμαλίες που εντοπίζονται από τον ένα ταξινομητή μπορεί να είναι οι πιο προφανείς ή οι πιο κρίσιμες για μια συγκεκριμένη εφαρμογή, ενώ ένας άλλος ταξινομητής μπορεί να εντοπίζει λιγότερο προφανείς αλλά εξίσου σημαντικές ανωμαλίες.

Κάθε ταξινομητής είναι βελτιστοποιημένος για διαφορετικές καταστάσεις και τύπους δεδομένων. Επομένως, η επιλογή του κατάλληλου ταξινομητή απαιτεί μια προσεκτική ανάλυση του προβλήματος και των δεδομένων που έχουμε στα χέρια μας.

Σε γενικές γραμμές, η εργασία αυτή προσφέρει διευκρινίσεις για την απόδοση των τεχνικών ανίχνευσης ανωμαλιών και προτείνει πιθανές κατευθύνσεις για μελλοντική έρευνα. Είναι εμφανές πως η ανίχνευση ανωμαλιών παίζει κεντρικό ρόλο στη διαχείριση και επεξεργασία των δεδομένων αισθητήρων. Οι σύγχρονες τεχνολογίες παρέχουν τη δυνατότητα για την ανίχνευση και την επεξεργασία μεγάλων όγκων δεδομένων, καθιστώντας τον εντοπισμό των ανωμαλιών πιο αποτελεσματικό και ακριβή. Η επιλογή της κατάλληλης τεχνικής για την ανίχνευση ανωμαλιών εξαρτάται σημαντικά από τη φύση των δεδομένων και του εκάστοτε προβλήματος. Είναι σημαντικό να ληφθεί υπόψη η αποτελεσματικότητα των διαφόρων μεθόδων σε σχέση με τον τύπο και τον όγκο των δεδομένων, καθώς και τις ειδικές απαιτήσεις της εφαρμογής.

Η υλοποίηση με την χρήση της γλώσσας προγραμματισμού Python, παρέχει την ευελιξία και τη δυνατότητα επέκτασης, ενισχύοντας τις δυνατότητες ανάλυσης και ανίχνευσης ανωμαλιών. Ο ρόλος της οπτικοποίησης των αποτελεσμάτων είναι εξίσου ζωτικής σημασίας, καθώς βοηθά στην καλύτερη και πιο ολοκληρωμένη ερμηνεία των δεδομένων, ενισχύοντας τη λήψη ενημερωμένων αποφάσεων. Η παρουσίαση των αποτελεσμάτων σε ευανάγνωστη και κατανοητή μορφή είναι απαραίτητη για την επιτυχή εφαρμογή των τεχνικών ανίχνευσης ανωμαλιών στην πράξη.

Η εκπόνηση αυτής της διπλωματικής εργασίας, ως εκ τούτου, αντιμετωπίζει τις προκλήσεις της αξιολόγησης και της επεξεργασίας των τεράστιων όγκων δεδομένων που παράγονται από τα συστήματα IoT, όπως και την ανίχνευση πιθανών ανωμαλιών που μπορούν να επηρεάσουν τη λειτουργικότητα των συστημάτων.

Αξιολόγηση Απόδοσης και Ανίχνευση Ανωμαλιών:

Η παρούσα εργασία επικεντρώνεται επίσης στην ανάπτυξη μεθοδολογιών για την αξιολόγηση της απόδοσης των συστημάτων IoT και την έγκαιρη ανίχνευση ανωμαλιών. Αυτό θα συμβάλει σημαντικά στη διατήρηση της αποδοτικότητας, της ακεραιότητας και της ασφάλειας των συστημάτων.

Γενικά, η παρούσα διπλωματική εργασία επιδιώκει τη δημιουργία μιας ρεαλιστικής και αποτελεσματικής προσέγγισης για την αντιμετώπιση των προκλήσεων της ανίχνευσης ανωμαλιών, εξασφαλίζοντας ταυτόχρονα τη βελτίωση της απόδοσης.

Στο πλαίσιο της παρούσας εργασίας, πραγματοποιήθηκε μια λεπτομερής ανάλυση της διαδικασίας αντιμετώπισης των προκλήσεων αυτών με βάση τον προγραμματισμό Python και τα δεδομένα που προκύπτουν από αυτόν.

Η αποτελεσματική συλλογή και προεπεξεργασία δεδομένων αποδείχθηκε ζωτικής σημασίας για την επίτευξη αξιόπιστων αποτελεσμάτων. Τα εργαλεία 'requests' και Pandas αποδείχθηκαν χρήσιμα για την ανάκτηση και την ταξινόμηση των δεδομένων αντίστοιχα. Ο εντοπισμός ανωμαλιών αποτελεί κεντρικό στοιχείο της εργασίας, όπου τα διάφορα μοντέλα που χρησιμοποιήθηκαν παρείχαν διάφορες προοπτικές όσον αφορά την απόδοση και την ακρίβεια. Η χρήση τεχνικών όπως ο Isolation Forest, k-NN, OneClassSVM και Autoencoder, συνέβαλε στην αποτελεσματικότητα της διαδικασίας.

Η χρήση της PCA για τη μείωση της διαστατικότητας αποδείχθηκε απαραίτητη για την οπτικοποίηση των δεδομένων, επιτρέποντας μια πιο σαφή κατανόηση των δυναμικών των ανωμαλιών στο σύνολο δεδομένων.

Η τελική οπτικοποίηση των δεδομένων προσφέρει μια σαφή εικόνα της αποτελεσματικότητας των εφαρμοζόμενων μεθόδων, καθιστώντας εφικτή την ανάλυση και την επιλογή των καλύτερων τεχνικών για την εντοπισμό ανωμαλιών.

Εν κατακλείδι, η εφαρμογή των παραπάνω μεθόδων στην πράξη αποτελεί ένα σημαντικό βήμα προόδου στην αντιμετώπιση των προκλήσεων, προσφέροντας μια ισχυρή βάση για την περαιτέρω έρευνα και ανάπτυξη στον τομέα αυτό.

Κεφάλαιο 7: Μελλοντικές Επεκτάσεις

Στο πλαίσιο αυτής της ενότητας, εξετάζουμε πιθανές μελλοντικές επεκτάσεις για την εργασία αυτή, που μπορούν να βελτιώσουν την απόδοση του συστήματος ανίχνευσης ανωμαλιών και να επεκτείνουν την εφαρμογή του σε διάφορους τομείς. Ορισμένες από τις μελλοντικές επεκτάσεις που μπορούν να γίνουν εξετάζονται παρακάτω:

1. **Ενσωμάτωση νέων μεθόδων ανίχνευσης ανωμαλιών**: Μπορεί να γίνει έρευνα και εφαρμογή νέων μεθόδων ανίχνευσης ανωμαλιών, πέραν των μεθόδων που χρησιμοποιήθηκαν στην παρούσα εργασία. Παραδείγματα περιλαμβάνουν τη χρήση νευρωνικών δικτύων, μεθόδων βαθιάς μάθησης ή μη επιβλεπόμενων αλγορίθμων μάθησης. Αντί για τους αλγορίθμους που χρησιμοποιήθηκαν στην εργασία, όπως ο IForest και ο k-NN, μπορεί να εξεταστεί η χρήση πιο προηγμένων αλγορίθμων ανίχνευσης ανωμαλιών. Για παράδειγμα, ο αλγόριθμος LOF ή ο αλγόριθμος One-Class Support Vector Machines (OCSVM) μπορούν να εξεταστούν για τη βελτίωση της ακρίβειας και της απόδοσης του συστήματος ανίχνευσης ανωμαλιών.

2. **Βελτιστοποίηση υπαρχουσών μεθόδων**: Μπορούν να γίνουν βελτιώσεις και προσαρμογές στις υπάρχουσες μεθόδους ανίχνευσης ανωμαλιών. Αυτό μπορεί να περιλαμβάνει την επιλογή βέλτιστων παραμέτρων, την αξιολόγηση εναλλακτικών αλγορίθμων ή την εφαρμογή τεχνικών βελτιστοποίησης όπως η διαστασιομείωση ή ο αυτόματος εντοπισμός ακραίων τιμών. Μπορεί να εξεταστεί η εφαρμογή τεχνικών βελτιστοποίησης για τη βελτίωση της επίδοσης του αλγορίθμου. Για παράδειγμα, η βελτιστοποίηση των παραμέτρων του αλγορίθμου, όπως ο αριθμός των γειτόνων στον αλγόριθμο k-NN ή ο αριθμός των δέντρων στον αλγόριθμο IForest, μπορεί να βελτιώσει την απόδοση του συστήματος.

3. Επέκταση σε άλλους τομείς και εφαρμογές: Η εργασία μπορεί να επεκταθεί για να εφαρμοστεί σε διάφορους τομείς και εφαρμογές. Παραδείγματα περιλαμβάνουν την ανίχνευση ανωμαλιών στην ιατρική διάγνωση, την ανίχνευση απάτης σε χρηματοοικονομικές συναλλαγές ή την ανίχνευση κακόβουλου λογισμικού σε συστήματα ασφαλείας.

4. Ενσωμάτωση περισσότερων χαρακτηριστικών: Μπορεί να γίνει η προσθήκη περισσότερων χαρακτηριστικών ή εκτίμησης παραμέτρων για να βελτιωθεί η ακρίβεια της ανίχνευσης ανωμαλιών. Αυτό μπορεί να περιλαμβάνει τη συλλογή πρόσθετων δεδομένων ή την εφαρμογή προηγμένων τεχνικών εξαγωγής χαρακτηριστικών.

5. **Εφαρμογή αυτόματης επεξεργασίας δεδομένων**: Μπορεί να γίνει η ανάπτυξη αυτόματων μεθόδων για την επεξεργασία και την προετοιμασία των δεδομένων πριν από την εκπαίδευση του αλγορίθμου. Αυτό μπορεί να περιλαμβάνει την αυτόματη επιλογή και καθαρισμό των χαρακτηριστικών, την αυτόματη ανίχνευση και αντιμετώπιση των απουσιάζουσων τιμών ή την αυτόματη κλιμάκωση των δεδομένων.

Αυτές οι πιθανές μελλοντικές επεκτάσεις μπορούν να βελτιώσουν την απόδοση και την εφαρμοσιμότητα του συστήματος ανίχνευσης ανωμαλιών και να ανοίξουν νέους

ορίζοντες για την εφαρμογή της ανίχνευσης ανωμαλιών σε ποικίλους τομείς και προβλήματα.

Βιβλιογραφία

[1] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3), 15.

[2] Byers, S., & Raftery, A. E. (1998). Nearest neighbor anomaly detection. Journal of the American Statistical Association 9

[3] Guttormsson, S., Lougheed, J., Wasynczuk, O., & Kawady, T. (1999). K-nearest neighbor algorithm for voltage security assessment. IEEE Transactions on Power Systems, 14(1), 102–107. https://doi.org/10.1109/59.744492

[4] Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (Vol. 29, No. 2, pp. 427-438).

[5] Hodge, V. J., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. Artificial Intelligence Review, 22(2), 85–126.

[6] Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).

[7] Chalapathy, R., & Chawla, S. (2019). Deep Learning for Anomaly Detection: A Survey. arXiv preprint arXiv:1901.03407.

[8] Baldi, P. (2012). Autoencoders, Unsupervised Learning, and Deep Architectures. Proceedings of ICML Workshop on Unsupervised and Transfer Learning, 37–49.

[9] Goldstein, M., & Uchida, S. (2016). A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. PLoS ONE, 11(4), e0152173.

[10]Liu, F.T.; Kai, M.T.; Zhou, Z.H. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 413–422.

[11] Saket Sathe, Arti Ramesh, Improved Anomaly Detection using Isolation Forest with an Application to Water Distribution Systems

[12] Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall.

[13] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

[14] Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences, 3*(3), 417-424.

[15]https://www.bbva.com/en/innovation/machine-learning-what-is-it-and-how-does-it-work/

[16]Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3), 1-58.

[17] Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2), 85-126.

[18] Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13(7), 1443-1471.

[19] Iglewicz, B., & Hoaglin, D. C. (1993). How to detect and handle outliers. *ASQC Quality Press*.

[20] Silverman, B. W. (1986). Density estimation for statistics and data analysis. *Chapman and Hall*.

[21] Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3/4), 591-611.

[22] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21-27.

[23] De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems, 50*(1), 1-18.

[24] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 226-231.

Συντομογραφίες - Αρκτικόλεξα – Ακρωνύμια

βλπ	βλέπε
κ.λπ.	και λοιπά
κ.ο.κ	και ούτω καθεξής
IoT	Internet of things
PCA	Principal Component Analysis
AI	Artificial Intelligence
IBM	International Business Machine
LOF	Local Outlier Factor
SVM	Support Vector Machine
k-NN	κ-Πλησιέστεροι γείτονες

Απόδοση ξενόγλωσσων όρων

Απόδοση	Ξενόγλωσσος όρος
αδερφός	sibling
απορρόφηση	absorption
βάση δεδομένων	database