ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΗΣ ΣΧΕΔΙΑΣΗΣ ΚΑΙ ΠΑΡΑΓΩΓΗΣ

http://www.eee.uniwa.gr
http://www.idpe.uniwa.gr
Θηβών 250, Αθήνα-Αιγάλεω 12241
Τηλ: +30 210 538-1614
Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών
Τεχνητή Νοημοσύνη και Βαθιά Μάθηση
https://aidl.uniwa.gr/

UNIVERSITY OF WEST ATTICA
FACULTY OF ENGINEERING
DEPARTMENT OF ELECTRICAL & ELECTRONICS ENGINEERING
DEPARTMENT OF INDUSTRIAL DESIGN AND PRODUCTION ENGINEERING
http://www.eee.uniwa.gr
http://www.idpe.uniwa.gr
250, Thivon Str., Athens, GR-12241, Greece
Tel: +30 210 538-1614
Master of Science in
Artificial Intelligence and Deep Learning
https://aidl.uniwa.gr/

Master of Science Thesis

Detection of hydraulic oil leaks using Artificial Intelligence

Student: Kogioumtzidis Georgios
Registration Number: AIDL-0006

MSc Thesis Supervisor

Papageorgas Panagiotis
Professor

ATHENS-EGALEO, February 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΗΣ ΣΧΕΔΙΑΣΗΣ ΚΑΙ ΠΑΡΑΓΩΓΗΣ

http://www.eee.uniwa.gr
http://www.idpe.uniwa.gr
Θηβών 250, Αθήνα-Αιγάλεω 12241
Τηλ: +30 210 538-1614
Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών
Τεχνητή Νοημοσύνη και Βαθιά Μάθηση
https://aidl.uniwa.gr/

UNIVERSITY OF WEST ATTICA
FACULTY OF ENGINEERING
DEPARTMENT OF ELECTRICAL & ELECTRONICS ENGINEERING
DEPARTMENT OF INDUSTRIAL DESIGN AND PRODUCTION ENGINEERING
http://www.eee.uniwa.gr
http://www.idpe.uniwa.gr
250, Thivon Str., Athens, GR-12241, Greece
Tel: +30 210 538-1614
Master of Science in
Artificial Intelligence and Deep Learning
https://aidl.uniwa.gr/

Μεταπτυχιακή Διπλωματική Εργασία

Ανίχνευση διαρροής υδραυλικού ελαίου με χρήση Τεχνητής Νοημοσύνης

Φοιτητής: Κογιουμτζίδης Γεώργιος
ΑΜ: AIDL-0006

Επιβλέπων Καθηγητής

Παπαγέωργας Παναγιώτης
Καθηγητής

ΑΘΗΝΑ-ΑΙΓΑΛΕΩ, ΦΕΒΡΟΥΑΡΙΟΣ 2023

This MSc Thesis has been accepted, evaluated and graded by the following committee:

Supervisor	Member	Member
Panagiotis Papageorgas Digitally signed by Panagiotis Papageorgas Date: 2023.02.27 10:56:34 +02'00'	Dimitrios Pyromalis Digitally signed by Dimitrios Pyromalis Date: 2023.02.28 18:01:16 +02'00'	Michail Papoutsidakis Digitally signed by Michail Papoutsidakis Date: 2023.02.27 22:54:36 +02'00'
Papageorgas Panagiotis	Pyromalis Dimitrios	Papoutsidakis Michail
Professor	Assistant Professor	Associate Professor
Electrical & Electronics Engineering Department	Dept. Industrial Design and Production Engineering	Dept. Industrial Design and Production Engineering
University of West Attica	University of West Attica	University of West Attica

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος Κογιουμτζίδης Γεώργιος του Ανδρέα, με αριθμό μητρώου AIDL-0006, μεταπτυχιακός φοιτητής του ΔΠΜΣ «Τεχνητή Νοημοσύνη και Βαθιά Μάθηση» του Τμήματος Ηλεκτρολόγων και Ηλεκτρονικών Μηχανικών και του Τμήματος Μηχανικών Βιομηχανικής Σχεδίασης και Παραγωγής, της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής,

δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της μεταπτυχιακής διπλωματικής εργασίας και κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολο τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος. Η εργασία δεν έχει κατατεθεί στο πλαίσιο των απαιτήσεων για τη λήψη άλλου τίτλου σπουδών ή επαγγελματικής πιστοποίησης πλην του παρόντος.
Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του διπλώματός μου.
Επιθυμώ την απαγόρευση πρόσβασης στο πλήρες κείμενο της εργασίας μου μέχρι 3/2023 και έπειτα από αίτηση μου στη Βιβλιοθήκη και έγκριση του επιβλέποντος καθηγητή.

Ο Δηλών
Κογιουμτζίδης Γεώργιος

The University of West Attica and Georgios Kogioumtzidis
February, 2023

Declaration of the author of this MSc thesis

I, Georgios, Andreas Kogioumtzidis with the following student registration number: AIDL-0006, postgraduate student of the MSc programme in "Artificial Intelligence and Deep Learning", which is organized by the Department of Electrical and Electronic Engineering and the Department of Industrial Design and Production Engineering of the Faculty of Engineering of the University of West Attica, hereby declare that:

I am the author of this MSc thesis and any help I may have received is clearly mentioned in the thesis. Additionally, all the sources I have used (e.g., to extract data, ideas, words or phrases) are cited with full reference to the corresponding authors, the publishing house or the journal; this also applies to the Internet sources that I have used. I also confirm that I have personally written this thesis and the intellectual property rights belong to myself and to the University of West Attica. This work has not been submitted for any other degree or professional qualification except as specified in it.

Any violations of my academic responsibilities, as stated above, constitutes substantial reason for the cancellation of the conferred MSc degree.

I wish to deny access to the full text of my MSc thesis until 3/2023, following my application to the Library of UNIWA and the approval from my supervisor.

The author
Kogioumtzidis Georgios

Ευχαριστίες

Για την εκπόνηση της παρούσας διπλωματικής εργασίας η οποία εκπονήθηκε στο πλαίσιο της φοίτησης μου στο Πρόγραμμα Μεταπτυχιακών Σπουδών «Τεχνητή Νοημοσύνη και Βαθιά Μάθηση» του Τμήματος Ηλεκτρολόγων και Ηλεκτρονικών Μηχανικών, θα ήθελα ιδιαιτέρως να ευχαριστήσω τους επιβλέποντες καθηγητές Παπαγέωργα Παναγιώτη και Πυρομάλη Δημήτριο για την ευκαιρία που μου έδωσαν να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα, καθώς και για την πολύτιμη βοήθεια και καθοδήγηση που μου προσέφεραν με οποιονδήποτε τρόπο στην συγγραφή και ολοκλήρωση της.

Ιδιαίτερες ευχαριστίες οφείλω στο Καθηγητή του Πανεπιστημίου Πελοποννήσου (Τμήμα Πληροφορικής και Τηλεπικοινωνιών) Γεώργιο – Όθων Γλεντή, ο οποίος συνέβαλε σημαντικά παραχωρώντας το σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα διπλωματική εργασία.

Abstract

The use of infrastructure and the installation of fluid transfer piping combined with poor maintenance and poor operating conditions due to high pressures can result in leaks. These leaks can cause disasters that can be environmental as well as financial. In order to solve such problems in time and to proceed to timely repair or control of the situation it is necessary to use models in hardware and software collaboration. These give real-time data and can predict the occurrence of a leakage as well as identify the possible location. The research presented in this paper proposes an AI model for a real-time monitoring system capable of detecting in time the leakage that will occur in a pressurized fluid pipeline. The results of the model are then validated with values obtained from sensors. The model is based on accelerometers placed inside the network on the outside of the pipelines. The signals obtained from the accelerometers have been analyzed and normalized to enable them to be input into the model to be used. Decision Trees and Random Forrest were chosen as optimal. Experiments were carried out on approximately 28m long pipeline made of carbon steel with an outer diameter of 48.3 mm and a thickness of 1.5 mm. Artificial leaks were used to provide the data and results.

Keywords

Machine learning, leak detection, artificial intelligence, supervised learning algorithms, acoustic sensors.

Περίληψη

Η χρήση υποδομών καθώς και η εγκατάσταση σωληνώσεων μεταφοράς ρευστών σε συνδυασμό με την ανεπαρκή συντήρηση και τις κακές συνθήκες λειτουργίας που οφείλονται σε υψηλές πιέσεις, μπορεί να επιφέρουν διαρροές. Οι διαρροές αυτές μπορούν να προκαλέσουν καταστροφές που μπορεί να είναι περιβαλλοντικές καθώς και οικονομικές. Για να λυθούν γρήγορα τέτοια προβλήματα και να προχωρήσουν σε έγκαιρη επισκευή ή έλεγχο της κατάστασης, είναι απαραίτητη η χρήση μοντέλων σε συνεργασία χρήσης hardware και software. Αυτά δίνουν στοιχεία σε πραγματικό χρόνο και μπορούν να προβλέψουν το συμβάν μιας διαρροής, όπως επίσης και να προσδιορίσουν το πιθανό σημείο. Η έρευνα που παρουσιάζεται σε αυτή την εργασία προτείνει ένα μοντέλο Τεχνητής Νοημοσύνης για ένα σύστημα παρακολούθησης, σε πραγματικό χρόνο, ικανό να εντοπίζει έγκαιρα την διαρροή που θα προκύψει σε έναν αγωγό ρευστών υπό πίεση. Στη συνέχεια γίνεται επιβεβαίωση των αποτελεσμάτων του μοντέλου με τιμές που λαμβάνονται από αισθητήρες. Το μοντέλο βασίζεται σε επιταχυνσιόμετρα τοποθετημένα εντός του δικτύου στην εξωτερική πλευρά των αγωγών. Τα σήματα που έχουν ληφθεί από τα επιταχυνσιόμετρα έχουν αναλυθεί και έχουν υποστεί κανονικοποίηση για να μπορέσουν να εισαχθούν στο μοντέλο που θα χρησιμοποιηθεί. Επιλέχθηκαν τα Decision Trees και τα Random Forrest σαν βέλτιστα. Τα πειράματα πραγματοποιήθηκαν σε αγωγούς μήκους περίπου 28 μέτρων από ανθρακοχάλυβα με εξωτερική διάμετρο 48,3 mm και πάχος 1,5mm. Χρησιμοποιήθηκαν τεχνητές διαρροές για την παροχή των στοιχείων και των αποτελεσμάτων.

Λέξεις – κλειδιά

Μηχανική μάθηση, ανίχνευση διαρροής, τεχνητή νοημοσύνη, αλγόριθμοι επιτηρούμενης μάθησης, ακουστικοί αισθητήρες.

Table of Contents

Λίστα πινάκων

Λίστα σχημάτων

Ακρωνύμια

TN: Τεχνητή Νοημοσύνη

IEEE: The Institute for Electrical and Electronics Engineers

RTTM: Real time transient model

SVM: Support Vector Machines

MSE: Mean Square Error

KNN: K-Nearest Neighbors

RBF: Radial Basis Function

SVC: Support Vector Classification

LBFGS: Limited-memory Broyden–Fletcher–Goldfarb–Shanno

Εισαγωγή

Μια βιομηχανική μονάδα αποβλέπει στη μεγιστοποίηση της παραγωγής της, και ως εκ τούτου στην ελαχιστοποίηση των βλαβών των μηχανών της και κατ' επέκταση του χρόνου που θα βρίσκεται εκτός λειτουργίας. Η έγκαιρη ανίχνευση σε σωληνώσεις μεταφοράς υδραυλικού ελαίου συμβάλει προς το σκοπό αυτό. Αυτό μπορεί να επιτευχθεί με μοντέλα κατηγοριοποίησης Τεχνητής Νοημοσύνης. Παρακάτω παραθέτουμε πιο συγκεκριμένα τους βασικούς πυλώνες της διπλωματικής εργασίας.

Αντικείμενο της Διπλωματικής

Τα τελευταία χρόνια παρατηρείται στις βιομηχανίες, οι οποίες δραστηριοποιούνται στο χώρο της παραγωγής υλικού, να υφίσταται απώλειες στη παραγωγική τους ικανότητα λόγω των δυσλειτουργιών που παρουσιάζουν οι μηχανές. Ο χρόνος αλλά και το κόστος επιδιόρθωσης της βλάβης της μηχανής συντελούν σημαντικά σε αυτό. Στα πλαίσια της παρούσας εργασίας θα εξετάσουμε συγκεκριμένα τις διαρροές υδραυλικού ελαίου εντός των αγωγών, και θα προτείνουμε μια μέθοδο βασιζόμενη σε Τεχνητή Νοημοσύνη για την έγκαιρη ανίχνευση τέτοιων διαρροών.

Δοκιμάζοντας διαφορετικούς αλγόριθμους μηχανικής μάθησης, βρίσκουμε το αποδοτικότερο μοντέλο με την υψηλότερη ακρίβεια ανίχνευσης διαρροής αλλά και πρόβλεψης. Ως μέθοδος ανίχνευσης διαρροής χρησιμοποιείται η ακουστική μέθοδος με χρήση επιταχυνσιόμετρων, τα σήματα των οποίων αναλύονται και εισάγονται ως σύνολο δεδομένων στα μοντέλα μηχανικής μάθησης. Αφού περαιτέρω αναλυθούν και δοκιμαστούν κάποιοι αλγόριθμοι μηχανικής μάθησης θα επιλεχτεί το μοντέλο με το υψηλότερο ποσοστό ακρίβειας πρόβλεψης.

Σκοπός και στόχοι

Με την έγκαιρη ανίχνευση και πρόβλεψη διαρροών στους αγωγούς, στοχεύουμε στο περιορισμό της παρουσίασης βλαβών των μηχανών και κατ' επέκταση στη μείωση του κόστους συντήρησης αλλά και στη μείωση του χρόνου και του κόστους επιδιόρθωσης, ενώ παράλληλα αποβλέπουμε στην αύξηση της παραγωγής. Στη παρούσα μελέτη γίνονται χρήση διαφορετικών αλγορίθμων μηχανικής μάθησης με σκοπό τη κατανόηση και αντίστοιχα τη χρήση τους σε προβλήματα διαρροής ρευστών σε αγωγούς. Καθώς η ΤΝ μπαίνει ολοένα και περισσότερο στη ζωή μας, η χρήση της λαμβάνει μέρος όλο και περισσότερο σε εφαρμογές χρόνο με το χρόνο. Εδώ χρησιμοποιείται η μηχανική μάθηση για να δούμε αν μπορεί αποδοτικά να αναγνωρίσει μια διαρροή έγκαιρα και αποτελεσματικά. Στόχος είναι μεταγενέστερα, προσθέτοντας και άλλες παραμέτρους κατά την εκπαίδευση του μοντέλου να είναι σε θέση να προβλέπει μια διαρροή που μπορεί να συμβεί.

Μεθοδολογία

Θα δοκιμαστούν διάφορα μοντέλα μηχανικής μάθησης της επιτηρούμενης μάθησης με σκοπό να εντοπιστεί το αποδοτικότερο μοντέλο για τον εντοπισμό διαρροής υδραυλικού ελαίου σε έναν αγωγό. Τα δεδομένα (σύνολο δεδομένων) που χρησιμοποιεί αυτή η μεταπτυχιακή εργασία έχουν παρθεί από τον κύριο Γεώργιο- Όθων Γλεντή, Καθηγητή του Τμήματος Πληροφορικής και Τηλεπικοινωνιών του Πανεπιστημίου της Πελοποννήσου[1]. Αυτά θα υποστούν κανονικοποίηση (Normalization) ώστε να γίνει πιο σωστά ο διαχωρισμός των κλάσεων. Επίσης θα χρησιμοποιηθούν τόσο για την εκπαίδευση και αξιολόγηση όσο και για τη δοκιμή του μοντέλου, ώστε να επαληθευτεί η αποτελεσματικότητα του μοντέλου. Η γλώσσα προγραμματισμού που θα χρησιμοποιηθεί είναι η python διότι είναι η πιο οικεία στην χρήση μηχανικής μάθησης λόγω των βιβλιοθηκών που υπάρχουν διαθέσιμες.

Καινοτομίες

Οι εφαρμογές που χρησιμοποιούν τη μηχανική μάθηση σε αγωγούς ρευστών επιτρέπουν στους υπολογιστές να αναλύουν γρήγορα και με ακρίβεια τεράστιες ποσότητες δεδομένων. Αυτό περιλαμβάνει τη δυνατότητα να "διαβάζουν" με ακρίβεια τα σήματα και το θόρυβο στα δεδομένα. Μετά τη συλλογή και ανάλυση αυτών των δεδομένων δημιουργούνται μοντέλα που είναι ικανά να κάνουν προβλέψεις με υψηλή ακρίβεια. Με την εφαρμογή αυτών των μοντέλων θα μπορούσαν να μειωθούν μελλοντικά οι μεγάλες απώλειες που υφίσταται η βιομηχανία όσον αφορά τις διαρροές αγωγών. Σε μια αντίθετη περίπτωση χρήσης μόνο αισθητήρων συνδεδεμένων με συστήματα PLC και SCADA θα περιορίζονταν στη δυνατότητα ανίχνευσης μόνο και όχι στην πρόβλεψη. Με τη χρήση Τεχνητής Νοημοσύνης αυτό είναι εφικτό.

Οργάνωση Κειμένου

Στο πρώτο κεφάλαιο παρουσιάζουμε υπάρχουσες τεχνολογίες για την ανίχνευση διαρροών σε αγωγούς και αναλύονται διάφορα είδη συστημάτων, τα οποία χρησιμοποιούν κλασικές μεθόδους. Πιο συγκεκριμένα θα εξετάσουμε συστήματα ανίχνευσης διαρροής βασισμένα σε ηλεκτρονικούς υπολογιστές, εσωτερικά και εξωτερικά συστήματα ανίχνευσης διαρροής.

Στο δεύτερο κεφάλαιο παρουσιάζουμε κλασικές τεχνολογίες της ΤΝ, στις οποίες βασίζονται οι σύγχρονες μέθοδοι και παρουσιάζουμε μοντέλα για τη ταξινόμηση και παλινδρόμηση.

Στο τρίτο κεφάλαιο προχωράμε στο σχεδιασμό και στην υλοποίηση του μοντέλου αφού πρώτα διεξάγουμε πολυάριθμα πειράματα με μοντέλα ταξινόμησης τα οποία αναφέρθηκαν στο κεφάλαιο 2. Παρουσιάζουμε τα αποτελέσματα ταξινόμησης μαζί με διάφορους δείκτες αποδοτικότητας και ακρίβειας. Από τη διεξαγωγή των πειραμάτων επιλέγουμε εκείνο το μοντέλο το οποίο παρουσιάζει μέγιστη αποδοτικότητα.

Τέλος παρουσιάζουμε τα συμπεράσματα στα οποία καταλήγουμε με τη λήξη της πειραματικής φάσης και αναφέρουμε πιθανές βελτιώσεις που μπορεί να γίνουν σε μελλοντική έρευνα.

1 ΚΕΦΑΛΑΙΟ 1: (Συστήματα Ανίχνευσης διαρροής αγωγών)

Τα τελευταία χρόνια λόγω των αυξανόμενων διαρροών σε αγωγούς, είναι πλέον αναγκαίο να χρησιμοποιούνται συστήματα ανίχνευσης διαρροών. Τα συστήματα ανίχνευσης διαρροών δεν μπορούν να αποτρέψουν τις διαρροές που προκαλούνται, αλλά μπορούν να βοηθήσουν στην ελαχιστοποίηση των συνεπειών μιας διαρροής και το εύρος των καταστροφών της. Για να επιτευχθεί βέβαια η σωστή λειτουργία των συστημάτων ανίχνευσης διαρροών σε αγωγούς είναι απαραίτητο να χρησιμοποιούνται τεχνολογίες ανίχνευσης διαρροών έτσι ώστε να εντοπίζεται έγκαιρα και να προλαμβάνεται μια διαρροή. Σημαντικό ρόλο παίζει εξίσου και η διασφάλιση επαρκών πόρων για τη συνεχή συντήρηση των συστημάτων ανίχνευσης.
Οι τεχνολογίες ανίχνευσης διαρροών χωρίζονται σε δυο κατηγορίες, σε εξωτερικούς και εσωτερικούς μεθόδους.

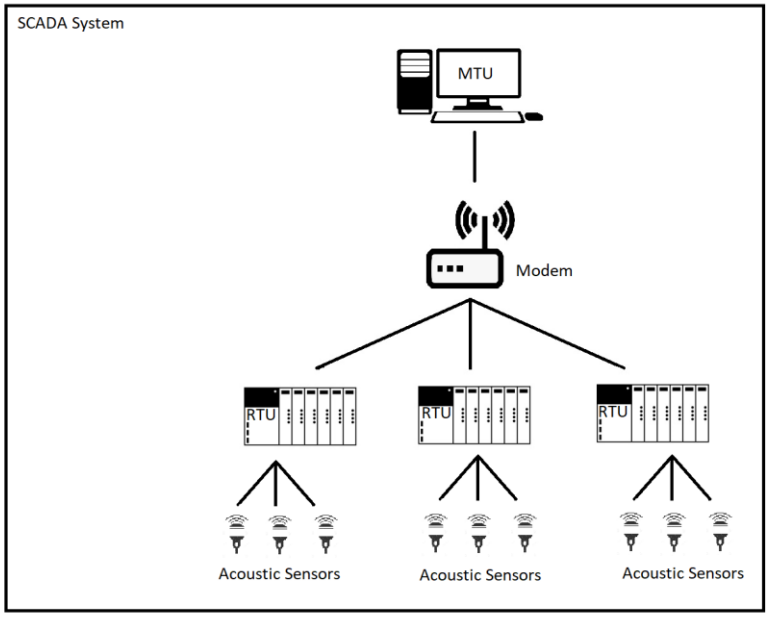
1.1 Συστήματα ανίχνευσης διαρροής βασισμένα σε Η/Υ

1.1.1 Όργανα μέτρησης.

Η ανίχνευση διαρροών είναι μία από τις εργασίες που μπορούν να εκτελέσουν τα συστήματα που βασίζονται σε υπολογιστή. Τα συστήματα αυτά έχουν συνήθως δύο κύρια στοιχεία: τα όργανα και έναν εποπτικό υπολογιστή με σχετικό λογισμικό και πρωτόκολλα επικοινωνίας. Η χρήση υπολογιστικών συστημάτων στην παρακολούθηση αγωγών επιτρέπει την ταχύτερη συλλογή, ανάλυση και δράση του μεγαλύτερου όγκου δεδομένων. Λόγω αυτών των παραγόντων η πλειονότητα των σύγχρονων συστημάτων αγωγών κάνει χρήση κάποιου τύπου παρακολούθησης μέσω υπολογιστή με τη λειτουργία του συστήματος σε εμπορικά διαθέσιμο ή ειδικά κατασκευασμένο λογισμικό. Κατά τη μέτρηση παραμέτρων όπως η πίεση αγωγού, η θερμοκρασία, η ροή και οι ιδιότητες του προϊόντος, τα όργανα περιλαμβάνουν μετρητές ροής, μετατροπείς πίεσης, αισθητήρες και καλώδια τα οποία τοποθετούνται κατά μήκος του αγωγού (είτε εξωτερικά είτε εσωτερικά).Είναι ζωτικής σημασίας η επιλογή της βέλτιστης διάταξης απόδοσης για ένα συγκεκριμένο περιβάλλον λειτουργίας, διότι η ευαισθησία και η ακρίβεια των εγκατεστημένων οργάνων χρησιμεύουν ως οι κύριοι περιοριστικοί παράγοντες για την αποτελεσματικότητα[2].

1.1.2 Σύστημα SCADA

Το SCADA είναι ένα σύστημα επικοινωνιών που βασίζεται σε υπολογιστή. Ονομάζεται σύστημα εποπτικού ελέγχου και συλλογής δεδομένων (**S**upervisory **C**ontrol **A**nd **D**ata **A**cquisition) το οποίο παρακολουθεί, επεξεργάζεται, μεταδίδει και εμφανίζει τα δεδομένα του αγωγού. Είναι δυνατόν να χρησιμοποιηθούν συστήματα SCADA απευθείας για την ανίχνευση διαρροών και για την υποστήριξη ενός συστήματος ανίχνευσης διαρροής. Συνήθως, ένα σύστημα ανίχνευσης διαρροής αγωγού χρησιμοποιεί τις πληροφορίες που παράγονται από ένα σύστημα SCADA για να βοηθήσει στον προσδιορισμό του κατά πόσον είναι πιθανό να συμβεί μια διαρροή. Τα συστήματα SCADA χρησιμοποιούν προγραμματισμένους λογικούς ελεγκτές (PLC), απομακρυσμένες τερματικές μονάδες (RTU) και άλλες ηλεκτρονικές συσκευές μέτρησης. Αυτές είναι τοποθετημένες σε σημεία ειδικά προστατευμένα και είναι υπεύθυνα για τη συλλογή δεδομένων σε πραγματικό χρόνο. Υπάρχουν πολυάριθμοι τρόποι επικοινωνίας με αυτές τις συσκευές, όπως μικροκύματα, κινητή τηλεφωνία, δορυφορικά συστήματα, μισθωμένες γραμμές κλπ. αλλά τα πιο δημοφιλή μέσα επικοινωνίας είναι η επικοινωνία μέσω της κινητής τηλεφωνίας και τα επίγεια και δορυφορικά ραδιοσυστήματα[2].



Σχήμα 1. Δομή συστήματος SCADA

Στη παραπάνω εικόνα απεικονίζεται ένα παράδειγμα δομής ενός συστήματος SCADA. Όπως φαίνεται στη κορυφή, απεικονίζεται η κύρια τερματική μονάδα, η οποία είναι και η μονάδα διεπαφής με τον χρήστη. Η MTU επικοινωνεί με τα RTUs μέσω της κινητής τηλεφωνίας χρησιμοποιώντας το Modem για να επιτευχθεί η επικοινωνία. Τα αντίστοιχα RTUs βρίσκονται σε απομακρυσμένο χώρο και είναι συνδεδεμένα μεταξύ τους όπως και με τα όργανα μέτρησης ή τους αισθητήρες. Η επικοινωνία μεταξύ των RTUs μπορεί να πραγματοποιείται μέσω Ethernet, Profinet ή και Profibus. Αυτό εξαρτάται από τον τύπο των εξαρτημάτων και περιφερειακών που έχουν επιλεχθεί κατά την εγκατάσταση. Στην παραπάνω περίπτωση οι αισθητήρες είναι συνήθως ακουστικοί αισθητήρες ανίχνευσης διαρροής.

1.1.2.1 Κύρια τερματική Μονάδα (Master Terminal Unit)

Σε ένα σύστημα SCADA, μια κύρια τερματική μονάδα (MTU) είναι μια συσκευή που στέλνει εντολές σε απομακρυσμένες τερματικές μονάδες (RTU) ή PLCs που βρίσκονται μακριά από το κέντρο ελέγχου. Η MTU συλλέγει επίσης τα απαραίτητα δεδομένα, τα αποθηκεύει, τα επεξεργάζεται και τα εμφανίζει ως γραφήματα, καμπύλες και πίνακες για να βοηθήσει στις αποφάσεις και στις λήψεις των κατάλληλων ενεργειών[3]. Το πρωτόκολλο επικοινωνίας είναι γνωστό ως η ανταλλαγή μηνυμάτων μεταξύ των συσκευών πεδίου και της MTU. Η MTU ζητά διαδοχικά δεδομένα από κάθε συσκευή και θα ζητήσει αυτόματα πληροφορίες από την πρώτη συσκευή όταν σαρώσει την τελευταία συσκευή, ξεκινώντας έναν ατελείωτο κύκλο ερωτηματολογίου. Η κύρια τερματική μονάδα (MTU) θεωρείται η καρδιά του συστήματος σε ένα σύστημα SCADA.

1.1.2.2 Απομακρυσμένη τερματική Μονάδα (Remote Terminal Unit)

Μια RTU είναι μια συσκευή ελέγχου που συχνά τοποθετείται σε απομακρυσμένη περιοχή και αντιπροσωπεύει ένα βασικό κομμάτι ενός μεγάλου συστήματος ανίχνευσης διαρροής. Η κύρια λειτουργία μιας RTU είναι η παρακολούθηση και η διαχείριση των οργάνων μέτρησης που είναι τοποθετημένοι στο σύστημα, όπως βαλβίδες, ενεργοποιητές, αισθητήρες και άλλα. Παρόλο που η επικοινωνία μεταξύ της MTU και της RTU είναι αμφίδρομη, η κύρια διάκριση είναι ότι η RTU δεν μπορεί να ξεκινήσει την επικοινωνία σε αντίθεση με τα MTU, απλώς συλλέγει και αποθηκεύει δεδομένα από τα σημεία που έχουν τοποθετηθεί τα όργανα. Για παράδειγμα, αναφέρουν όταν αλλάζουν οι συνθήκες στο σημείο των οργάνων ή σε μια χρονοπρογραμματισμένη βάση χωρίς να ερωτώνται. Ο ρυθμός αναζήτησης έχει άμεσο αντίκτυπο στα συστήματα ανίχνευσης διαρροής που εξαρτώνται από το σύστημα SCADA για τη λήψη δεδομένων λειτουργίας. Προγράμματα στο εσωτερικό της MTU ενεργοποιούν τον δίαυλο επικοινωνίας είτε από εντολές του χειριστή, είτε αυτόματα και ξεκινούν την επικοινωνία μεταξύ της MTU και των RTU. Οι εντολές που στέλνουν οι MTUs στις RTUs ξεκινούν αυτόματα. Η RTU αποστέλλει τις ζητούμενες πληροφορίες όταν η κύρια τερματική μονάδα (MTU) τις ζητά. Συνεπώς, η MTU θεωρείται ως master και η RTU ως Slave. Αφού λάβει τα κατάλληλα δεδομένα, η MTU συνδέεται με τις συσκευές διεπαφής χειριστή[3].

1.2 Εσωτερικά συστήματα ανίχνευσης διαρροής

Εσωτερικοί μέθοδοι χαρακτηρίζονται αυτοί, οι οποίοι χρησιμοποιούν εξόδους αισθητήρων για την παρακολούθηση των εσωτερικών παραμέτρων ενός αγωγού όπως η πίεση, η θερμοκρασία, το ιξώδες, η πυκνότητα, ο ρυθμός ροής και η ηχητική ταχύτητα προϊόντος. Παρακάτω θα αναλύσουμε και τις μεθόδους που είναι βασισμένες σε εσωτερικά συστήματα ανίχνευσης διαρροής. Αυτές είναι: η μέθοδος ισοζυγίου όγκου ,η μέθοδος αρνητικής πίεσης, η στατιστική ανάλυση, η μοντελοποίηση μεταβατικής κατάστασης σε πραγματικό χρόνο και η παρακολούθηση ρυθμού μεταβολής πίεσης.

1.2.1 Μέθοδος ισοζυγίου όγκου (Volume Balance Method)

Η ανίχνευση διαρροής μέσω Volume balance χρησιμοποιείται ευρέως λόγω της απλότητας της αλλά και της άμεσης συσχέτισης με τις διαρροές. Οι μετρητές ροής θα πρέπει να είναι τοποθετημένοι σε όλες τις εισόδους και εξόδους ενός αγωγού και τα δεδομένα μέτρησής τους να αποστέλλονται σε τακτά χρονικά διαστήματα δειγματοληψίας σε ένα σύστημα εποπτικού ελέγχου και συλλογής (SCADA)[4]. Ο τρόπος λειτουργίας είναι σχετικά απλός, ένα ρευστό εισέρχεται σε ένα τμήμα αγωγού και παραμένει εκεί μέχρι να εξέλθει από το τμήμα εξόδου του αγωγού. Το ρευστό που εισέρχεται και εξέρχεται από ένα τυπικό κυλινδρικό δίκτυο αγωγών μπορεί να μετρηθεί. Η είσοδος και η έξοδος καταγράφονται στα δύο άκρα του τμήματος του αγωγού. Όταν οι ροές όγκου είναι ισορροπημένες τότε δεν υπάρχει διαρροή, επομένως μια διαφορά μεταξύ των ροών όγκου που βρίσκονται υπό παρατήρηση στα δύο άκρα του αγωγού υποδηλώνει την ύπαρξη διαρροής[5].

Μπορούν να χρησιμοποιηθούν διάφορες τεχνικές ισοζυγίου όγκου:

- Απλή μέθοδος ισοζυγίου όγκου (Simple volume balance), όπου η συνολική ροή εισόδου αφαιρείται από τη συνολική ροή εξόδου χωρίς να λαμβάνεται υπόψη η θερμοκρασία, η πίεση του αγωγού ή τη σύνθεση του προϊόντος.
- Βελτιωμένο ισοζύγιο όγκου (Enhanced Volume Balance), στο οποίο το ισοζύγιο όγκου προσαρμόζεται για τις αλλαγές στο απόθεμα του αγωγού βάση τα χαρακτηριστικά του ρευστού, την πίεση και τη θερμοκρασία κατά μήκος του αγωγού[4]

Το ισοζύγιο όγκου σε σύγκριση με άλλες τεχνικές ανίχνευσης διαρροών είναι ιδιαίτερα αποτελεσματικό στον εντοπισμό μικρών διαρροών και μπορεί να χρησιμοποιηθεί σε αγωγούς όπως αερίου, υγρού και πολλαπλών φάσεων με την προϋπόθεση να υπάρχουν διαθέσιμοι μετρητές ροής. Η μέτρηση της ροής σε κάθε άκρο της γραμμής ή του τμήματος του αγωγού δεν μπορεί να εντοπίσει τη θέση της διαρροής γι' αυτό τον λόγο οι διαρροές εντοπίζονται αργά. Ένας από τους μεγαλύτερους περιορισμούς αυτής της μεθόδου είναι η αβεβαιότητα των οργάνων, όπου η ευαισθησία στις τυχαίες διαταραχές και στη δυναμική των αγωγών μπορεί να οδηγήσει μερικές φορές σε ψευδείς συναγερμούς όταν είναι συνδεδεμένο με ένα σύστημα SCADA[6].

1.2.2 Μέθοδος αρνητικής πίεσης (Negative Pressure)

Σύμφωνα με αυτή τη μέθοδο όταν εμφανίζεται διαρροή σε κάποιο αγωγό, δημιουργείται μια πτώση πίεσης στη θέση διαρροής παράγοντας ένα αρνητικό κύμα πίεσης. Αρχικά λαμβάνεται πρώτα ως κριτήριο αναφοράς η αρχική πίεση του αγωγού από αισθητήρες που είναι τοποθετημένοι στα δύο άκρα του αγωγού. Οι αισθητήρες αυτοί μπορούν να εντοπίσουν τη θέση της διαρροής σύμφωνα με τη μεταβολή του σήματος πίεσης και το διάστημα μεταξύ των κυμάτων και του αρνητικού κύματος πίεσης που προκαλείται από τη διαρροή. Όταν το αρνητικό κύμα πίεσης φτάνει στο τελικό άκρο του αγωγού, θα προκαλέσει την πτώση πίεσης της εισόδου του σταθμού και στη συνέχεια της πίεσης εξόδου του σταθμού[7]. Επειδή η μέθοδος των αρνητικών κυμάτων πίεσης δεν βασίζεται στο υλικό του συστήματος, δεν απαιτεί τη δημιουργία μαθηματικού μοντέλου. Ωστόσο, για να είναι αποτελεσματική αυτή, απαιτείται η διαρροή να είναι γρήγορη και ξαφνική, διότι, ιδίως στην περίπτωση μιας σταδιακής διαρροής, δεν υπάρχει σαφές αρνητικό κύμα πίεσης[6]. Η νέα γενιά τεχνολογίας αρνητικών κυμάτων λαμβάνει

δείγματα δεδομένων πίεσης σε υψηλές συχνότητες και τα στέλνει στον κεντρικό διακομιστή για ανάλυση. Υπάρχουν ολοκληρωμένοι αλγόριθμοι που ανιχνεύουν μικρές διαρροές και ελαχιστοποιούν τους ψευδείς συναγερμούς λόγω λειτουργικών αλλαγών. Η μέθοδος κύματος αρνητικής πίεσης μπορεί να εφαρμοστεί σε έναν αγωγό υγρού. Υλοποιείται κατά βάση σε έναν αυτόνομο υπολογιστή και μπορεί να ανιχνεύσει μικρές έως μεγάλες διαρροές μέσα σε λίγα λεπτά, επίσης μπορεί να παρέχει την ακριβή θέση διαρροών σε ακτίνας 100 μέτρων. Υπολογίζεται ότι οι εκτιμήσεις του μεγέθους της διαρροής δεν θα είναι ακριβείς εάν δεν χρησιμοποιούνται ροόμετρα. Εάν οι πτώσεις πιέσεως από τις λειτουργικές αλλαγές μοιάζουν με τις πτώσης πιέσεως των διαρροών, μπορεί να προκύψουν ψευδείς συναγερμοί. Επίσης εάν μια διαρροή δεν εντοπιστεί τη στιγμή που θα εμφανιστεί, μπορεί να παραμείνει μη ανιχνεύσιμη[4].

1.2.3 Στατιστική ανάλυση (Statistical Analysis)

Η μέθοδος στατιστικής ανάλυσης μπορεί να εφαρμοστεί σε αγωγούς αερίου, υγρών και πολυφασικών αγωγών. Συνήθως εφαρμόζεται σε έναν αυτόνομο υπολογιστή ή ένα σύστημα SCADA. Σε αυτή την μέθοδο για την ανίχνευση διαρροών μπορεί να χρησιμοποιηθούν διαφορετικά επίπεδα στατιστικής ανάλυσης. Πριν από τον προσδιορισμό του κατά πόσον έχει εμφανιστεί μια ανωμαλία, φιλτράρονται τα δεδομένα ροής ή πίεσης στο ένα άκρο με σκοπό τη μείωση των επίπεδων θορύβου. Έτσι θα διευκρινιστεί εάν έχει συμβεί κάποια ανωμαλία και αν υπάρχει κάποια διαρροή. Ο βαθμός της στατιστικής συμμετοχής ποικίλλει σε μεγάλο βαθμό, ανάλογα με τις διάφορες μεθόδους που βασίζονται σε εσωτερικά συστήματα ανίχνευσης διαρροής. Δυο διαδεδομένοι μέθοδοι που βασίζονται στην στατιστική ανάλυση είναι η Ανάλυση Σημείων Πίεσης (Pressure Point Analysis) και το Στατιστικό Ισοζύγιο Όγκου (Statistical Volume Balance).

- Ανάλυση σημείων πίεσης
  Η μέθοδος ανάλυσης σημείου πίεσης είναι μια τεχνική ανίχνευσης διαρροών που βασίζεται στη στατιστική ιδιοτήτων των μετρούμενων πιέσεων σε διάφορα σημεία κατά μήκος του αγωγού. Η διαρροή προσδιορίζεται μέσω της σύγκρισης των μετρούμενων τιμών έναντι της τρέχουσας στατιστικής τάσης των προηγούμενων μετρήσεων[8]. Ένα συμβάν διαρροής υποδεικνύεται εάν η στατιστική πίεση των νέων εισερχόμενων δεδομένων είναι σημαντικά χαμηλότερη από την προηγούμενη τιμή ή χαμηλότερη από ένα προκαθορισμένο όριο. Με βάση το γεγονός ότι μια διαρροή προκαλεί πάντοτε άμεση μείωση της πίεσης στο σημείο διαρροής, η τεχνική αυτή θεωρείται ένας από τους ταχύτερους τρόπους εντοπισμού διαρροών σε αγωγούς. Η τεχνική ανάλυσης σημείων πίεσης χρειάζεται απλώς σήματα πίεσης από ένα ή περισσότερα σημεία ανίχνευσης, επομένως τα πλεονεκτήματα της περιλαμβάνουν χαμηλό κόστος εγκατάστασης και εύκολη συντήρηση. Επιπλέον, έχει την ικανότητα να βρίσκει μικροσκοπικές ρωγμές που δεν είναι ανιχνεύσιμες με άλλες τεχνικές. Ωστόσο, ο εντοπισμός των σημείων διαρροής με αυτή τη στρατηγική αποτελεί πρόκληση, με αποτέλεσμα η εφαρμογή αυτής της μεθόδου να είναι πολύ περιορισμένη.

- Στατιστικό Ισοζύγιο Όγκου
  Το σύστημα στατιστικού ισοζυγίου όγκου λειτουργεί αποτελεσματικά για όλους τους αγωγούς. Χρησιμοποιεί μια βελτιωμένη μέθοδο ισοζύγιο όγκου σε συνδυασμό με τη τεχνική δοκιμής διαδοχικής αναλογίας πιθανοτήτων (Sequential Probability Ratio Test) για να παρέχει αξιόπιστη ανίχνευση διαρροών. Το σύστημα προσδιορίζει εάν το βελτιωμένο ισοζύγιο όγκου έχει αυξηθεί με μια προκαθορισμένη πιθανότητα με τον υπολογισμό του λόγω της πιθανότητας διαρροής προς την πιθανότητα μη διαρροής. Μια διαρροή συνήθως έχει ως αποτέλεσμα πτώση της πίεσης του αγωγού και διαφορά στους ρυθμούς ροής. Τέτοιου είδους μοτίβα αναγνωρίζονται από το σύστημα στατιστικού ισοζυγίου όγκου και ο εντοπισμός των διαρροών βασίζεται σε υπολογισμούς πιθανότητας που εκτελούνται σε τακτά χρονικά διαστήματα δειγματοληψίας. Συνήθως οι αλλαγές κατάστασης λειτουργίας στο δίκτυο δημιουργεί πτώση στην πίεση και στην ροή του αγωγού. Για αυτό το λόγο θα πρέπει να εξισορροπούνται αυτές οι διακυμάνσεις του δικτύου εκτός αν υπάρχει όντως συμβάν διαρροής. Η ικανότητα αυτή να μαθαίνει το σύστημα π.χ. τις αλλαγές κατάστασης λειτουργίας δικτύου, έχει ως αποτέλεσμα το υψηλό επίπεδο αξιοπιστίας και ο συνδυασμός με τη τεχνική δοκιμής διαδοχικής αναλογίας πιθανοτήτων και της αναγνώρισης μοτίβων, να οδηγεί σε χαμηλό αριθμό ψευδών συναγερμών. Μέσα σε λίγα λεπτά, μπορούν να εντοπιστούν μικρές, μεσαίες και μεγάλες διαρροές ανεξάρτητα του τύπου των ρευστών που ρέει στους αγωγούς[9].

1.2.4 Μοντελοποίηση μεταβατικής κατάστασης σε πραγματικό χρόνο (RTTM)

Η μέθοδος RTTM βασίζεται στην απόκλιση των μετρούμενων τιμών και των προβλεπόμενων συνθηκών μοντελοποίησης από το μεταβατικό μοντέλο προσομοίωσης. Μπορεί να εφαρμοστεί σε αγωγούς αερίου, υγρών και πολυφασικών αγωγών, εάν υπάρχουν διαθέσιμες οι μετρήσεις στις εισόδους και εξόδους του αγωγού από αισθητήρες όπως ροής, πίεσης και θερμοκρασίας αφού προϋπόθεση για να χρησιμοποιηθεί αυτή η μέθοδος είναι τα δεδομένα μέτρησης. Το μοντέλο μπορεί να προσαρμοστεί έτσι ώστε να είναι σε θέση να διακρίνει διαρροές, λάθος μετρήσεις που οφείλονται σε λάθη των οργάνων ή και κανονικά μεταβατικά φαινόμενα. Λόγω της έντονης απαίτησης υπολογισμού, υλοποιείται σε έναν αυτόνομο υπολογιστή καθώς απαιτεί πολλές διαδικτυακές μετρήσεις. Οι μετρήσεις αυτές προέρχονται από την εγκατάσταση πολλών οργάνων, όπου ο αριθμός αυτών αντιστοιχεί και στην υψηλότερη ακρίβεια στο μοντέλο. Το μοντέλο αυτό βασίζεται σε όργανα που λειτουργούν σωστά και είναι ρυθμισμένα (βαθμονομημένα) για βέλτιστη απόδοση. Τα λάθος ρυθμισμένα όργανα μπορεί να οδηγήσουν σε ψευδείς συναγερμούς διαρροής. Το σύστημα προσομοιώνει τις συνθήκες του αγωγού χρησιμοποιώντας προηγμένη μηχανική ρευστών, υδραυλική μοντελοποίηση σε υπολογιστή, όπως επίσης κάνει υπολογισμούς που περιλαμβάνουν τη διατήρηση της ορμής και της ενέργειας, καθώς και ένα σύνολο από εξισώσεις για υπολογισμό της ροής[10]. Για την ανίχνευση διαρροής και προσδιορισμό της θέσης της, υπολογίζεται το προφίλ πίεσης-ροής του αγωγού από τις τιμές, οι οποίες μετρώνται, στην είσοδο του αγωγού. Στην συνέχεια υπολογίζεται το προφίλ πίεσης-ροής του αγωγού από τις αντίστοιχες τιμές στην έξοδο του αγωγού. Αυτά τα δυο προφίλ που υπολογίστηκαν αναλύονται και με σύνθετους υπολογισμούς προσδιορίζεται η διαρροή και το ακριβής σημείο της. Τέλος, εφόσον τα χαρακτηριστικά που παρατηρήθηκαν διαφέρουν με την πρόβλεψη του συστήματος RTTM, τότε σηματοδοτείται συναγερμός για διαρροή[2]. Η χρησιμοποίηση του RTTM είναι χρονοβόρα και δαπανηρή, καθώς απαιτεί την εγκατάσταση πολλαπλών οργάνων, συστηματική εκπαίδευση των ελεγκτών

και συχνή συντήρηση του συστήματος, ως εκ τούτου χρησιμοποιείται κυρίως από μεγάλες εταιρείες αγωγών.

1.2.5 Παρακολούθηση ρυθμού μεταβολής πίεσης/ροής (Rate of change monitoring)

Η παρακολούθηση του ρυθμού της πίεσης ή της ροής είναι μια ακόμη μέθοδος ανίχνευσης διαρροής και μπορεί να εφαρμοστεί σε όλους του αγωγούς υγρού, αερίου καθώς και σε οποιοδήποτε αγωγό είναι εφικτή η μέτρηση ροής και πίεσης. Σε αυτή την τεχνολογία οι αλλαγές των τιμών της πίεσης και της ροής στην είσοδο και έξοδο του αγωγού σε ένα καθορισμένο χρονικό διάστημα, συγκρίνονται με τις αντίστοιχες τιμές των ορίων που έχουν προκαθοριστεί από τον χρήστη. Εφόσον αυτά τα όρια ξεπεραστούν δημιουργείται συναγερμός διαρροής.Η εφαρμογή μετατροπέα πίεσης κατά μήκος του αγωγού σηματοδοτεί μια διαρροή παρακολουθώντας τα κύματα σημάτων που δημιουργούνται και μεταδίδονται κατά την πτώση της πίεσης μέσα στον αγωγό [11]. Αυτή η μέθοδος διαρροής μπορεί να υλοποιηθεί στο πλαίσιο ενός συστήματος SCADA.

1.3 Εξωτερικά συστήματα ανίχνευσης διαρροής

Οι εξωτερικοί μέθοδοι είναι οι μέθοδοι που βασίζονται στην ανίχνευση διαρροών από ειδικές συσκευές που είναι εγκατεστημένες στο εξωτερικό τμήμα του αγωγού. Αυτές οι συσκευές παρακολουθούν και σηματοδοτούν διαρροή όταν υπάρχει ανωμαλία στο περιβάλλον του αγωγού. Παραδείγματα αυτών των μεθόδων ανίχνευσης διαρροής είναι: Ακουστική μέθοδος ανίχνευσης διαρροής, ανίχνευση διαρροής μέσω οπτικής ίνας, υπέρυθρη θερμογραφία, μέθοδος ανίχνευσης υγρών, βιολογική μέθοδος και μέθοδος ανίχνευσης ατμών. Οι μέθοδοι αυτοί θα συζητηθούν στις επόμενες ενότητες.

1.3.1 Ακουστική μέθοδος ανίχνευσης διαρροής (Acoustic sensing)

Έχουν αναπτυχθεί τελευταία πολυάριθμες τεχνικές ανίχνευσης και εντοπισμού διαρροών, συμπεριλαμβανομένων εκείνων που βασίζονται στο ισοζύγιο μάζας/όγκου, στα αρνητικά κύματα πίεσης, στα μεταβατικά μοντέλα και στις οπτικές ίνες. Μια νέα μέθοδος εύρεσης διαρροών είναι η τεχνική ακουστικής εκπομπής. Η ιδέα πίσω από την ανίχνευση διαρροών σε αγωγούς με τη χρήση της τεχνολογίας ακουστικών εκπομπών είναι ότι το υγρό που διαφεύγει παράγει ένα ακουστικό σήμα όταν διέρχεται από έναν αγωγό. Οι **αισθητήρες** που είναι τοποθετημένοι στο εξωτερικό του αγωγού παρακολουθούν τις θέσεις και τις εντάσεις του θορύβου στο εσωτερικό του αγωγού. Με τη χρήση αυτών των δεδομένων δημιουργείται ένας βασικός "ακουστικός χάρτης" της γραμμής. Το ακουστικό σήμα που προκύπτει από μια διαρροή αναγνωρίζεται και εξετάζεται από τους επεξεργαστές του συστήματος. Εφόσον υπάρχει αλλαγή στο βασικό ακουστικό προφίλ σηματοδοτεί συναγερμό. Κοντά στο σημείο της διαρροής, το σήμα που λαμβάνεται είναι ισχυρότερο, καθιστώντας δυνατό τον εντοπισμό της θέσης της διαρροής [2]. Επίσης ο ακριβής προσδιορισμός της θέσης προσδιορίζεται υπολογίζοντας τη χρονική υστέρηση(lag) που δημιουργείται μεταξύ των ακουστικών σημάτων που ανιχνεύονται από τους αισθητήρες. Η μέθοδος αυτή είναι κατάλληλη για σωλήνες με χαμηλή ροή και υψηλή εσωτερική πίεση. Στην ανίχνευση θαμμένων αγωγών, η ακουστική τεχνολογία εκπομπής έχει

μεγαλύτερη ευαισθησία από άλλες τεχνολογίες ανίχνευσης και την καθιστά πολύ ακριβή κατά τον εντοπισμό της πηγής διαρροής[12].

Η ακουστική μέθοδος για την ανίχνευση διαρροών μπορεί να χωριστεί σε δύο κατηγορίες: **ενεργητική** και **παθητική**. Ακούγοντας τις ηχώ των ηχητικών παλμών που αντανακλώνται λόγω διαρροής, οι ενεργητικές μέθοδοι μπορούν να ανιχνεύσουν ελαττώματα αγωγών. Από την άλλη, με τη παθητική μέθοδο εντοπίζονται διαρροές ακούγοντας τις μεταβολές στον ήχο που παράγονται από τα κύματα πίεσης στους αγωγούς.

Τα συνήθη όργανα ακουστικής ανίχνευσης διαρροών είναι συσκευές ακουστικής ακρόασης ή ράβδοι (συνήθως τοποθετημένοι στο θαμμένο αγωγό), επιταχυνσιόμετρα, συσχετιστές θορύβου διαρροής και συστήματα υδροφώνων (για τοποθέτηση στο εσωτερικό των αγωγών), όπως για παράδειγμα ακουαφόνια, τα οποία απαιτούν άμεση επαφή με τις βαλβίδες ή τους κρουνούς, ή γεώφωνα, τα οποία εντοπίζουν τις διαρροές που βρίσκονται στη επιφάνεια του εδάφους. Τα όργανα αυτά παρέχουν ικανοποιητικές επιδόσεις μόνο στην περίπτωση μεταλλικών σωλήνων και έχουν περιορισμένες δυνατότητες σε πλαστικούς σωλήνες[13]. Εάν κατά την χρήση αυτής της μεθόδου χρησιμοποιούνται μαλακοί σωλήνες, όπως αυτοί που κατασκευάζονται από πλαστικό, η στρατηγική αυτή δεν λειτουργεί αποτελεσματικά, καθώς είναι πιο ελαστικοί και επιβραδύνουν τα ηχητικά κύματα κατά 300-600 m/sec[14].Σε αυτή την περίπτωση τα επιταχυνσιόμετρα προτιμώνται προκειμένου να ξεπεραστούν οι αδυναμίες της ακουστικής μεθόδου καθώς σαν αισθητήρες είναι πιο ακριβής στον εντοπισμό και την ανίχνευση διαρροών σε πλαστικό αγωγό. Επιπλέον, τα όργανα αυτά μπορεί να χάσουν την ικανότητα τους να εντοπίζουν διαρροές σε αγωγούς που μεταφέρουν φυσικό αέριο καθώς και σε βαθιά θαμμένους υπόγειους σωλήνες[15].

Τα **επιταχυνσιόμετρα** είναι ηλεκτρομηχανικά όργανα που μπορούν να ανιχνεύουν την επιτάχυνση. Ως εκ τούτου, χρησιμοποιούνται για τη μέτρηση αυτών των δυνάμεων. Οι δυνάμεις επιτάχυνσης διακρίνονται σε δύο μορφές, τις στατικές και τις δυναμικές δυνάμεις. Οι στατικές δυνάμεις που σχετίζονται με τις δυνάμεις του πλανήτη γη στα αντικείμενα πάνω στον πλανήτη, ενώ οι δυναμικές οφείλονται σε δυνάμεις λόγω της κίνησης ή της δόνησης[16]. Για τη δημιουργία ενός επιταχυνσιόμετρου πρέπει να προστεθεί το κύκλωμα που απαιτείται για τη μετατροπή σημάτων σε τάση. Πρόσθετες προσεγγίσεις περιλαμβάνουν, μεταξύ άλλων, τη χρήση του φαινομένου της πιεζοαντίστασης, των φυσαλίδων θερμού αέρα καθώς και του φωτός. Η πιο ευρέως προσέγγιση είναι τα επιταχυνσιόμετρα που χρησιμοποιούν το πιεζοηλεκτρικό φαινόμενο. Αυτά περιλαμβάνουν μικροσκοπικές κρυσταλλικές δομές, οι οποίες κατά τη πίεση που υφίστανται από τις επιταχυντικές δυνάμεις, παράγεται τάση. Αυτή η τάση είναι το σήμα που χρειάζεται κάποιος για να αναλύσει και στην συνέχεια να εντοπίσει μία διαρροή που θα συμβεί σε κάποια χρονική στιγμή. Συνήθως κατά την χρήση των επιταχυνσιομέτρων για τον εντοπισμό και την ανίχνευση διαρροών χρησιμοποιείται ανάλυση ετερο-συσχέτισης για την ανάλυση των σημάτων ακουστικής και δόνησης.

1.3.2 Ανίχνευση διαρροής μέσω οπτικής ίνας (Fiber optic sensing)

Στο πλαίσιο αυτής της τεχνικής εγκαθίστανται αισθητήρες οπτικών ινών κατά μήκος της εξωτερικής πλευράς του αγωγού για την ανίχνευση διαρροών υγρών, πολυφασικών ρευστών ή και για διαρροή φυσικού αερίου. Σε αυτά τα συστήματα μεταβάλλεται η μετάδοση του φωτός μέσω του καλωδίου οπτικών ινών όταν μια ουσία εισέρχεται στην επίστρωση του καλωδίου.

Οι επιστρώσεις στο καλώδιο οπτικών ινών αλληλεπιδρούν με τις διαρροές υδρογονανθράκων για να αλλάξουν ένα τμήμα των διαθλαστικών ιδιοτήτων του καλωδίου. Αυτή η αλλαγή του δείκτη διάθλασης εντοπίζεται με τη χρήση παλμικού λέιζερ υψηλής συχνότητας, το οποίο μπορεί επίσης να χρησιμοποιηθεί για την ανίχνευση μικρών διαρροών[10]. Δημοφιλείς τεχνικές ανίχνευσης διαρροών μέσω οπτικής ίνας που χρησιμοποιούνται έως και σήμερα είναι η **κατανεμημένη ακουστική ή κατανεμημένη ανίχνευση κραδασμών**(Distributed Acoustic or Vibration Sensing), η **κατανεμημένη ανίχνευση θερμοκρασίας** (Distributed Temperature Sensing) και η **κατανεμημένη ανίχνευση παραμόρφωσης**. Ανάλογα με τον τύπο του εδάφους, τη μετατόπιση του καλωδίου και τις συνθήκες του εδάφους, η κάθε μέθοδος θα αντιδράσει με διαφορετικό τρόπο σε κάθε συμβάν διαρροής και θα χρειαστεί διαφορετικό χρόνο για να σημάνει συναγερμό στη διαρροή που συνέβη.

Κατανεμημένη ακουστική ή κατανεμημένη ανίχνευση κραδασμών

Η συσκευή εξετάζει την οπισθοσκέδαση Rayleigh που εμφανίζεται φυσικά με την άντληση συνεκτικών ενεργειακών παλμών λέιζερ στις οπτικές ίνες του καλωδίου που είναι εγκατεστημένο παράλληλα με τον αγωγό. Ο προσδιορισμός της θέσης μιας διαρροής υπολογίζεται με τη χρονομέτρηση του χρονικού διαστήματος μεταξύ της εκπομπής του παλμού λέιζερ και της ανίχνευσης της ανάκλασης του. Η ακουστηκότητα αυτή μειώνεται όσο απομακρύνεται κανείς από μια διαρροή[9].

Κατανεμημένη ανίχνευση θερμοκρασίας

Η ιδέα πίσω από αυτή την τεχνολογία είναι ότι όταν ένας αγωγός παρουσιάζει διαρροή ή ένα υγρό υδρογονανθράκων εισχωρεί στο καλώδιο επικάλυψης, η θερμοκρασία του καλωδίου θα αλλάξει. Οι ανωμαλίες κατά μήκος του αγωγού μπορούν να εντοπιστούν με την παρακολούθηση των διακυμάνσεων της θερμοκρασίας στο καλώδιο οπτικών ινών[17]. Τα συστήματα θέρμανσης και οι διαρροές αργού πετρελαίου έχουν συνήθως ως αποτέλεσμα τη τοπική θέρμανση κοντά στον αγωγό καθώς και διαρροές αγωγών φυσικού αερίου προκαλούν τοπική ψύξη στη γη κοντά στο σημείο διαρροής. Τέτοιες διακυμάνσεις θερμότητας μπορούν να εντοπιστούν από το καλώδιο οπτικών ινών. Η παρακολούθηση αγωγών σε μεγάλες αποστάσεις καθίσταται δυνατή χάρη στον νηματοειδή σχεδιασμό των οπτικών ινών και τις ιδιότητες χαμηλών απωλειών. Οι τεχνικές κατανεμημένης ανίχνευσης μπορούν να παρέχουν την συνεχή μέτρηση της θερμοκρασίας της ίνας σε συνάρτηση με την απόσταση[4].

1.3.3 Υπέρυθρη θερμογραφία (Infrared Thermography)

Η υπέρυθρη θερμογραφία είναι μια μέθοδος που βασίζεται στην υπέρυθρη εικόνα και χρησιμοποιεί υπέρυθρες κάμερες που εμφανίζουν υπέρυθρο φως στην περιοχή 900-1400 nm για την ανίχνευση αλλαγών θερμοκρασίας στο περιβάλλον του αγωγού. Τέτοιοι μέθοδοι που βασίζονται στον μηχανισμό υπέρυθρης θερμογραφίας μπορούν να χρησιμοποιηθούν για την ανίχνευση διαρροής σε αγωγούς. Ένα θερμογράφημα είναι η εικόνα που παράγεται από μια υπέρυθρη κάμερα θερμογραφίας. Η μέθοδος αυτή έχει αποκτήσει ευρεία αποδοχή για την παρακολούθηση αγωγών λόγω της ικανότητας της να μετρά τις μεταβολές της θερμοκρασίας σε πραγματικό χρόνο και με ανέπαφο τρόπο[18]. Οι μετρήσεις θερμοκρασίας είναι ένας από τους συνηθέστερους παράγοντες σε αγωγούς αερίου για την ανίχνευση διαρροής, καθώς οι διαρροές αερίου συνήθως προκαλούν μη φυσιολογική κατανομή της θερμοκρασίας. Η

υπέρυθρη θερμογραφία μπορεί να χρησιμοποιηθεί για ποικίλες εφαρμογές ως ανέπαφο και μη επεμβατικό εργαλείο παρακολούθησης της κατάστασης. Οι θερμικές κάμερες είναι αποτελεσματικά εργαλεία για την ανίχνευση αντικειμένων ποικίλων σχημάτων και υλικών. Έχουν τη δυνατότητα παρακολούθησης αντικειμένων από απόσταση και παρέχουν μια οπτική αναπαράσταση που χρησιμοποιεί διαφορετικά χρώματα σε διαφορετικές κατανομές θερμοκρασίας της περιοχής αυτής. Με τη χρήση θερμικής κάμερας μπορεί να αναγνωριστούν ανωμαλίες στο περιβάλλον του αγωγού, καθώς οι περιοχές με υψηλές ή και χαμηλές θερμοκρασίες θα απεικονίζονται με διαφορετικό χρώμα στην θερμική εικόνα. Υπάρχουν δύο τύποι θερμογραφίας: η **ενεργητική θερμογραφία** και η **παθητική θερμογραφία**[19]. Η ενεργητική θερμογραφία δείχνει την περιοχή ενδιαφέροντος με θερμική αντίθεση φόντου, σε αντίθεση με την παθητική θερμογραφία, όπου η περιοχή ενδιαφέροντος επικεντρώνεται στη μεταβολή της θερμοκρασίας και στο φόντο. Η ανέπαφη και μη επεμβατική, σε πραγματικό χρόνο, μέτρηση της θερμοκρασίας προσφέρει μια λύση στον εντοπισμό διαρροής αγωγών αφού σε αντίθετη περίπτωση για την ανίχνευση θερμοκρασίας θα γινόταν χρήση των συμβατικών τεχνικών μέτρησης θερμοκρασίας, όπως των αντιστάσεων θερμοκρασίας και των θερμοστοιχείων (thermocouple). Τα κύρια εξαρτήματα για την δημιουργία του συστήματος είναι μια βάση κάμερας, μια υπέρυθρη κάμερα και μια μονάδα απεικόνισης για την προβολή της υπέρυθρης θερμικής εικόνας. Η χρήση της τεχνικής αυτής επιφέρει ποιοτική και αποδοτική απεικόνιση της διαρροής , χρόνο απόκρισης και εύκολη χρήση. Παρόλα αυτά η εγκατάσταση καμερών μεγάλης ανάλυσης έχει υψηλό κόστος[17].

1.3.4 Μέθοδος ανίχνευσης υγρών (Liquid sensing method)

Η ανίχνευση διαρροών με ανίχνευση υγρών διατίθεται ως ένα πλήρες σύστημα ανίχνευσης και εντοπισμού διαρροών που περιλαμβάνει όλο το απαραίτητο υλικό και λογισμικό. Για κάθε εφαρμογή χρησιμοποιείται ένας συγκεκριμένος τύπος καλωδίου ανάλογα με το υγρό που παρακολουθείται. Διατίθεται λογισμικό διεπαφής ελεγκτή για την παροχή πληροφοριών σε πραγματικό χρόνο όσον αφορά την ανίχνευση διαρροής και την καταγραφή ιστορικών δεδομένων. Οι συγκεκριμένοι τύποι καλωδίων επιλέγονται για κάθε εφαρμογή με βάση το συγκεκριμένο υγρό που παρακολουθείται. Τα καλώδια ανίχνευσης υγρών θάβονται κάτω ή κοντά σε έναν αγωγό και είναι κατασκευασμένα έτσι ώστε να αντανακλούν τις μεταβολές στους μεταδιδόμενους ενεργειακούς παλμούς που προκαλούνται από τις διαφορές αντίστασης που οφείλονται από την επαφή με τα υγρά του αγωγού. Ένας μικροεπεξεργαστής μεταδίδει συνεχώς ασφαλείς ενεργειακούς παλμούς κατά μήκος της σύνδεσης και στη συνέχεια λαμβάνει τους παλμούς που ανακλώνται. Στη μνήμη του μικροεπεξεργαστή διατηρείται ένας χάρτης βασικής ανάκλασης με βάση τη συγκεκριμένη εγκατάσταση καλωδίου. Όταν υπάρχει διαρροή το υγρό εισέρχεται στο καλώδιο και μεταβάλλει τη σύνθετη αντίσταση του, η οποία αλλάζει το μοτίβο ανάκλασης που επιστρέφει στον μικροεπεξεργαστή. Αυτή η αλλαγή στο μοτίβο ανάκλασης σηματοδοτεί διαρροή στο σημείο που υπάρχει μεταβολή της σύνθετης αντίστασης. Στην συνέχεια ο μικροεπεξεργαστής καταγράφει την θέση της διαρροής και δημιουργεί συναγερμό διαρροής. Η ανίχνευση διαρροών με ανίχνευση υγρών προσφέρει υψηλή ακρίβεια στον προσδιορισμό της θέσης της διαρροής, όπως επίσης εύκολη εγκατάσταση, συντήρηση και διαχείριση του λογισμικού. Η επιλογή αυτής της μεθόδου απαιτεί υψηλό κόστος εγκατάστασης και σημαντικές ανάγκες για καλωδίωση ισχύος και σήματος[2].

1.3.5 Βιολογική μέθοδος ανίχνευσης διαρροής. (Biological leak detection method)

Ο όρος "βιολογικοί τρόποι ανίχνευσης διαρροών" αναφέρεται στη παραδοσιακή διαδικασία για την ανίχνευση διαρροής ρευστών σε έναν αγωγό με τη χρήση του ανθρώπινου παράγοντα, εκπαιδευμένων σκύλων, όπως επίσης και τη χρήση τεχνολογίας όπως τα μη επανδρωμένα αεροσκάφη. Βασικός παράγοντας σε αυτή τη μέθοδο είναι η εκπαίδευση του προσωπικού για να είναι σε θέση να αναγνωρίσουν μια διαρροή στον αγωγό. Σε κάθε περίπτωση θα πρέπει να ελέγχεται ο αγωγός κατά μήκος από τους εκπαιδευμένους παρατηρητές και να εντοπίζονται έγκαιρα ανωμαλίες στο περιβάλλον του αγωγού. Τέτοιες ανωμαλίες μπορεί να είναι διαρροές, μη επιθυμητές μυρωδιές, ραγίσματα καθώς και θόρυβοι που μπορεί να σηματοδοτήσουν βλάβη στον αγωγό αν δεν αντιμετωπιστούν έγκαιρα. Όσον αφορά τις μυρωδιές που προέρχονται από κάποιο ράγισμα του αγωγού μπορεί να εντοπιστούν από εκπαιδευμένα σκυλιά καθώς έχουν τη δυνατότητα να αντιλαμβάνονται μέσω της όσφρησης από μεγάλες αποστάσεις μυρωδιές που μπορεί να σημαίνουν κάποια διαρροή. Συνήθως σε τέτοιου είδους ανωμαλίες τα εκπαιδευμένα σκυλιά έχουν μεγαλύτερη αποδοτικότητα από τον ανθρώπινο παράγοντα. Τα τελευταία χρόνια έχει αλλάξει ο τρόπος λειτουργίας ανίχνευσης διαρροών λόγω της γρήγορης ανάπτυξης των τηλεχειριζόμενων οχημάτων (μη επανδρωμένα αεροσκάφη, μη επανδρωμένα υποβρύχια). Έχει αποδειχθεί η ανθεκτικότητα τους κατά την εκτέλεση επιθεώρησης σε συνθήκες οι οποίες είναι επικίνδυνες για τον ανθρώπινο παράγοντα καθώς και για τα εκπαιδευμένα σκυλιά. Μπορούν να βρίσκονται σε λειτουργία επιθεώρησης σε σημεία που είναι επικίνδυνα και δεν είναι προσβάσιμα από τους ανθρώπους καθώς θα έθεταν σε κίνδυνο την υγεία τους. Τα μη επανδρωμένα οχήματα έχουν το πλεονέκτημα ότι διαθέτουν ένα απομακρυσμένο σύστημα λειτουργίας, γεγονός που τα καθιστά ιδανικά για επιθεώρηση σε επικίνδυνο και απομακρυσμένο περιβάλλον. Ένα ακόμα πλεονέκτημα τους είναι ότι έχουν χαμηλό κόστος συντήρησης. Από την άλλη τα μειονεκτήματα των μη επανδρωμένων οχημάτων είναι η περιορισμένη απόδοση σε κακές καιρικές συνθήκες καθώς και το υψηλό κόστος αγοράς. Θα χρειαστεί επιπρόσθετα η εκπαίδευση του χειριστή ώστε να μπορεί να χειρίζεται και να εντοπίζει τις ανωμαλίες που μπορεί να προκύψουν στον αγωγό.

1.3.6 Μέθοδος ανίχνευσης ατμών (Vapor Sensing)

Αν και μπορούν να χρησιμοποιηθούν σε αγωγούς, οι συσκευές ανίχνευσης αερίων υδρογονανθράκων χρησιμοποιούνται συνήθως σε συστήματα δεξαμενών αποθήκευσης. Τα συστήματα παρακολούθησης ατμών είναι μια σχετικά απλή ιδέα για την ανίχνευση διαρροών. Η εγκατάσταση ενός δευτερεύοντος αγωγού πρέπει να γίνει σε όλο το μήκος του αγωγού για να χρησιμοποιηθεί η τεχνική ανίχνευσης διαρροών με ανίχνευσης ατμών. Ο αγωγός μπορεί είτε να περικλείει πλήρως τον αγωγό είτε να είναι ένας διάτρητος σωλήνας μικρής διαμέτρου προσαρτημένος σε αυτόν. Επίσης είναι στεγανός και γεμάτος με αέρα (Μία Ατμόσφαιρα).

Για την ανίχνευση της παρουσίας διαρροής, δείγματα αερίων αέρα αναρροφώνται στο σωλήνα και αξιολογούνται με τη χρήση αισθητήρων ατμών υδρογονανθράκων. Ανεξάρτητα από το αν ο σωλήνας είναι εγκατεστημένος στον αέρα, στο νερό ή στο έδαφος, μετά από ορισμένο χρονικό διάστημα, το εσωτερικό του σωλήνα παράγει μια ακριβή εικόνα των υλικών που το περιβάλλουν. Για την ανάλυση των στοιχείων στον σωλήνα, ο συσσωρευμένος αέρας στο σωλήνα ωθείται με σταθερή ταχύτητα από μια αντλία προς μια μονάδα ανίχνευσης, η οποία καταγράφει την παρατηρούμενη στάθμη σε συνάρτηση με το χρόνο. Αισθητήρες αερίου είναι

τοποθετημένοι στο άκρο του σωλήνα όπου βρίσκεται η μονάδα ανίχνευσης. Αυτοί σηματοδοτούν την ύπαρξη διαρροής για κάθε αύξηση της συγκέντρωσης αερίου. Το μέγεθος της διαρροής μπορεί να προσδιοριστεί συγκρίνοντας το ύψος της κορυφής, το οποίο είναι ανάλογο της συγκέντρωσης της ουσίας (μια μικρή διαρροή παράγει μια μικρή κορυφή και μια μεγάλη διαρροή παράγει μια μεγάλη κορυφή)[11].

Για τον προσδιορισμό της θέσης της διαρροής, μια ακριβής ποσότητα δοκιμαστικού αερίου εγχέεται στην ηλεκτρολυτική κυψέλη της ανιχνεύσιμης γραμμής στο τέλος πριν από κάθε λειτουργία άντλησης. Μαζί με τον αέρα, το αέριο αυτό διανύει όλο το μήκος του σωλήνα αισθητήρα. Το αέριο δοκιμής παράγει ένα δείκτη αρχής ή ένα δείκτη τέλους καθώς κινείται μέσω της συσκευής ανίχνευσης. Υποδεικνύοντας ότι ολόκληρη η ποσότητα αέρα που περιέχεται στο σωλήνα αισθητήρα έχει περάσει από το σταθμό μέτρησης, η άφιξη της λειτουργεί ως δείκτης ελέγχου. Έτσι, ο δείκτης τέλους παρέχει πληροφορίες για ολόκληρο το μήκος του σωλήνα αισθητήρα. Με τη χρήση του λόγου της διάρκειας διαδρομής της κορυφής διαρροής προς εκείνη του δείκτη τέλους υπολογίζεται η ακριβής θέση διαρροής[20]. Λόγω των πρακτικών προκλήσεων της εγκατάστασης ενός συστήματος σε όλο το μήκος ενός αγωγού, οι σωλήνες ανίχνευσης ατμών χρησιμοποιούνται συνήθως μόνο σε μικρές γραμμές. Όπως επίσης με αυτή τη μέθοδος ανίχνευσης διαρροής μπορεί μια διαρροή να ανιχνευτεί σε μεγάλο χρονικό διάστημα, κάνοντας τη μη αποτελεσματική για χρήση σε υποθαλάσσιους αγωγούς.

2 ΚΕΦΑΛΑΙΟ 2: (Τεχνητή Νοημοσύνη - Μηχανική μάθηση)

2.1 Μηχανική Μάθηση

Η μηχανική μάθηση έχει οριστεί το 1950 από το πρωτοπόρο Τεχνητής Νοημοσύνης Arthur Samuel ως "το πεδίο σπουδών που δίνει στους υπολογιστές τη δυνατότητα να μαθαίνουν χωρίς να προγραμματιστούν από τον άνθρωπο"[21].

Μηχανική μάθηση είναι ένα υποκατάστατο/κομμάτι της ΤΝ. Χρησιμοποιεί μεγάλο όγκο δεδομένων ως παραδείγματα για το πως μπορεί να επιλυθεί ένα πρόβλημα και χρησιμοποιώντας τα μοντέλα μηχανικής μάθησης δημιουργεί το επιθυμητό αποτέλεσμα με υψηλή ακρίβεια πρόβλεψης. Σκοπός της είναι η δυνατότητα να μιμηθεί τον ανθρώπινο εγκέφαλο, όπως η ικανότητα της μάθησης και πρόβλεψης και στη συνέχεια βελτίωσης, στοχεύοντας στην υψηλή ακρίβεια πρόβλεψης. Τέτοια δεδομένα μπορεί να είναι χαρακτήρες, αριθμοί, εικόνες, τιμές από αισθητήρες όπως και πιο ευαίσθητα δεδομένα, τραπεζικές συναλλαγές , ηλεκτρονικές αλληλογραφίες, φωτογραφίες ανθρώπων, ιστορικά αγορών και αλλά. Αυτά τα δεδομένα συλλέγονται και χρησιμοποιούνται ως δεδομένα εκπαίδευσης και δεδομένα δοκιμής. Όσο μεγαλύτερος είναι ο όγκος των δεδομένων τόσο πιο υψηλή θα είναι η ακρίβεια στις προβλέψεις. Για να επιτευχθεί αυτό επιλέγεται το κατάλληλο μοντέλο μηχανικής μάθησης και η κατάλληλη αλγοριθμική τεχνική η οποία μπορεί να ρυθμιστεί αλλάζοντας τις παραμέτρους που την απαρτίζουν, έτσι ώστε να αυξηθεί η ακρίβεια της πρόβλεψης του μοντέλου και να είναι τα αποτελέσματα ποιο εύστοχα. Αφού τα δεδομένα περαστούν από το αλγόριθμο μάθησης γίνεται ένας έλεγχος των δεδομένων δοκιμής με αυτά της εκπαίδευσης και υπολογίζεται η ακρίβεια της πρόβλεψης. Το έτοιμο μοντέλο που έχει δημιουργηθεί θα μπορεί να χρησιμοποιηθεί για την ίδια εργασία μελλοντικά για διαφορετικό σετ δεδομένων.

Η λειτουργία της μηχανικής μάθησης μπορεί να είναι παραστατική, όπου το σύστημα χρησιμοποιεί τα δεδομένα για να εξηγήσει τι συνέβη, προβλέψιμη όπου το σύστημα χρησιμοποιεί τα δεδομένα για να προβλέπει ένα ενδεχόμενο, και συντομογραφική όπου το σύστημα χρησιμοποιεί τα δεδομένα για να προτείνει προτάσεις για το τι ενέργεια θα εκτελέσει[22].

Η μηχανική μάθηση χρησιμοποιείται σε πολύπλοκες εργασίες με τον τρόπο που το λύνει και ο άνθρωπος. Τέτοιες εργασίες είναι η αναγνώριση μιας οπτικής σκηνής όπως η αναγνώριση προσώπου, η αναγνώριση χειρόγραφων χαρακτήρων όπως επίσης και πιο πολύπλοκες εργασίες όπου είναι η αναγνώριση και ο αποκλεισμός ανεπιθύμητων αλληλογραφιών, η σύσταση προϊόντων ή υπηρεσιών και η ανίχνευση ενός συμβάντος από την ανάγνωση σημάτων από αισθητήρες κάποιου συστήματος. Στην δικιά μας περίπτωση γίνεται ανάγνωση των σημάτων από ακουστικά αισθητήρια που καταγράφουν αν συνέβη διαρροή σε σύστημα ροής ρευστών, και στην συνέχεια εκτελείται μια προγραμματισμένη ενέργεια.

Υπάρχουν υποκατηγορίες (Μοντέλα μηχανικής μάθησης) που απαρτίζουν την μηχανική μάθηση, οι οποίες χρησιμοποιούν διάφορες αλγοριθμικές τεχνικές. Ανάλογα με την εργασία που θα εκτελεστεί επιλέγεται και το κατάλληλο μοντέλο. Αυτό βασίζεται κυρίως στο πως θα λαμβάνεται η μάθηση ή θα τροφοδοτείται αυτή στο μοντέλο μηχανικής μάθησης. Οι πιο διαδεδομένες υποκατηγορίες είναι η Επιτηρούμενη μάθηση (Supervised Learning), η μη Επιτηρούμενη μάθηση (Unsupervised Learning) και η Ενισχυτική μάθηση (Reinforcement Learning).Κάθε ένα από αυτά τα μοντέλα μηχανικής μάθησης αποτελείται από αλγοριθμικές τεχνικές οι οποίες είναι υπεύθυνες για την εκπαίδευση και πρόβλεψη της επιθυμητής εξόδου.

Τέτοιες τεχνικές επιτηρούμενης μάθησης είναι οι Linear και Polynomial Regression, Decision Trees, Random Forest, KNN, Logistic Regression, SVM και Naïve-Bayes. Γνωστές αλγοριθμικές τεχνικές μη επιτηρούμενης μάθησης είναι ο SVD,PCA και K-means.

2.1.1 Επιτηρούμενη μάθηση (Supervised Learning)

Στην επιτηρούμενη μάθηση (Supervised Learning) το μοντέλο έχει εκπαιδευτεί με προκαθορισμένα παραδείγματα εξόδου. Με απλά λόγια δίνονται τα δεδομένα εκπαίδευσης και τα αντίστοιχα δεδομένα ετικέτες, τα οποία περιγράφουν την επιθυμητή έξοδο για κάθε ένα από τα δεδομένα εκπαίδευσης. Χρησιμοποιώντας την μέθοδο αυτή, το μοντέλο μπορεί να εκπαιδευτεί συγκρίνοντας τις πραγματικές εξόδους με τις αντίστοιχες εξόδους που έχει προβλέψει έτσι ώστε να βελτιωθεί βρίσκοντας τα σφάλματα και πετυχαίνοντας μεγαλύτερη ακρίβεια πρόβλεψης όταν δίνονται νέα δεδομένα. Ως εκ τούτου η επιτηρούμενη μάθηση χρησιμοποιεί μοτίβα για να προβλέψει τις τιμές των ετικετών όπως επίσης και τις ετικέτες που δεν έχουν προκαθοριστεί. Στη πλειοψηφία στις εφαρμογές επιτηρούμενης μάθησης, ο κύριος στόχος είναι να αναπτυχθεί η τέλεια συνάρτηση πρόβλεψης h(x).Στην πράξη αυτό σημαίνει ότι στην εκπαίδευση γίνεται χρήση μαθηματικών αλγορίθμων, οι οποίοι παραμετροποιώντας τους βελτιώνεται η συνάρτηση τους, έτσι ώστε δίνοντας ένα δεδομένο εισόδου x (π.χ τιμές αναλογικών σημάτων από αισθητήρες) να προβλέπεται με ακρίβεια η τιμή h(x) (π.χ μέγεθος διαρροής σε αγωγό ρευστών). Η τιμή x θα μπορούσε να είναι παραπάνω από μια είσοδο π.χ x2 για πίεση υψηλής σε αγωγό ρευστών , η x3 για θερμοκρασία ρευστών. Η σωστή επιλογή των τιμών εισόδου είναι ένα σημαντικό κομμάτι κατά τη σχεδίαση της μηχανικής μάθησης[23].

2.1.2 Μη επιτηρούμενη μάθηση (Unsupervised Learning)

Στην μη επιτηρούμενη μάθηση (Unsupervised Learning) το μοντέλο εκπαιδεύεται χωρίς προκαθορισμένα παραδείγματα εξόδου. Σε αυτή τη περίπτωση δεν υπάρχουν δεδομένα ετικέτες και ο αλγόριθμος εκπαίδευσης ψάχνει να βρει ομοιότητες και διαφορές μεταξύ των δεδομένων εισόδου. Η συνήθης χρήση της μη επιτηρούμενης μάθησης είναι για την ανίχνευση κρυφών μοτίβων σε πολύπλοκα δεδομένα και η εξερευνητική ανάλυση των δεδομένων. Μπορεί η χρήση της μη επιτηρούμενης μάθησης για την ανίχνευση κρυφών μοτίβων εντός των δεδομένων εκπαίδευσης να είναι μονόδρομος αλλά επίσης μπορεί να χρησιμοποιηθεί έτσι ώστε ο αλγόριθμος του μοντέλου αυτόματα να ανιχνεύει και να κατηγοριοποιεί τα δεδομένα εισόδου, ακόμα και αν είναι σε αρχική μορφή(Raw Data)[23]. Γνωστές εφαρμογές της μη επιτηρούμενης μάθησης είναι για τη ανίχνευση απάτης σε πιστωτικές κάρτες. Με αυτές ελέγχει μεγάλο όγκο συναλλαγών και αν εντοπιστεί κάποια ανωμαλία στις συναλλαγές καταχωρείται ως παραβάτης, άλλες γνωστές εφαρμογές της μη επιτηρούμενη μάθησης είναι τα συνιστώμενα συστήματα. Κάνοντας εξόρυξη σε μεγάλο όγκο δεδομένων προτιμήσεων του χρήστη, ανιχνεύουν και είναι σε θέση να προτείνουν τις προτιμήσεις του χρήστη. Στην πτυχιακή εργασία θα μπορούσε να χρησιμοποιηθεί για τον εντοπισμό του μεγέθους της διαρροής και να το κατηγοριοποιήσει σε μικρή διαρροή, μεσαία διαρροή , μεγάλη διαρροή ή και καθόλου διαρροή. Στην δική μας περίπτωση αυτά είναι ήδη κατηγοριοποιημένα για αυτό τον λόγο δεν θα χρησιμοποιηθεί αυτή η μέθοδος.

2.1.3 Ενισχυτική μάθηση (Reinforcement Learning)

Η ενισχυτική μάθηση (Reinforcement Learning) είναι ένα ειδικό μοντέλο της επιτηρούμενης μάθησης της οποίας οι επιθυμητές έξοδοι είναι άγνωστοι. Υπάρχουν 2 παράμετροι που το χαρακτηρίζουν. Ο ένας είναι ο εκπαιδευτής, ο οποίος παρέχει την πληροφορία για το αν είναι σωστά ή λάθος τα αποτελέσματα και ο άλλος ο μαθητής, στόχος του οποίου είναι να φτάσει στην επιθυμητή έξοδο μέσω των ανταμοιβών που λαμβάνει κάθε φορά που φτάνει στο επιθυμητό αποτέλεσμα. Το μοντέλο αυτό προσδιορίζει το πως ένας τεχνητός πράκτορας (πραγματικός ή ρομπότ) μπορεί να μάθει να επιλέγει ενέργειες προκειμένου να φτάσει στην αναμενόμενη ανταμοιβή. Επειδή όμως είναι βασισμένο μόνο στη πληροφορία ως προς το εάν η εκτίμηση της εξόδου είναι κοντά με την πραγματική, η καθορισμένη σωστή απάντηση μπορεί να μην είναι γνωστή στον εκπαιδευτή ή και στον μαθητή. Η ενισχυτική μάθηση είναι μια διαδικασία που ανταμείβει την σωστή έξοδο και τιμωρεί τη λανθασμένη. Αυτό την καθιστά πιο αργή διαδικασία σε σχέση τις άλλες μεθόδους[24]. Γνωστές εφαρμογές της ενισχυτικής μάθησης είναι στα ηλεκτρονικά παιχνίδια, όπου η ανταμοιβή δίνεται όταν κερδίζεις ένα παιχνίδι όπως επίσης και όταν κερδίζεται κάποιος αντίπαλος. Στην πτυχιακή εργασία δεν θα χρησιμοποιηθεί η μέθοδος αυτή διότι δεν εξυπηρετεί την εργασία.

2.1.4 Ταξινόμηση

Στη ΤΝ και στην μηχανική μάθηση, με την ταξινόμηση εννοούμε την δυνατότητα της μηχανής να μπορεί να καταχωρεί το κάθε δεδομένο εισόδου σε μια προκαθορισμένη κλάση. Αυτό συμβαίνει βρίσκοντας μοτίβα στα δεδομένα εκπαίδευσης και καταχωρώντας τα στην αντίστοιχη κλάση ετικέτες που έχει προκαθορίσει ο χρήστης. Η ταξινόμηση μπορεί να είναι δυαδικής (οι κλάσεις είναι 2 (0 και 1)) ή πολλαπλής (όπου οι κλάσεις είναι παραπάνω από 2).Παραδείγματα δυαδικής ταξινόμησης είναι η αναγνώριση μιας φωτογραφίας για το εάν είναι άνθρωπος ή όχι. Στη δικιά μας εργασία με δυαδική ταξινόμηση θα μπορούσε να γίνει ανίχνευση εάν υπάρχει διαρροή η όχι. Παραδείγματα πολλαπλής ταξινόμησης είναι η αναγνώριση προσώπου και καταχώρηση σε πιο άτομο αντιστοιχεί. Στην πτυχιακή εργασία θα γίνει αναγνώριση διαρροής και καταχώρηση στη κλάση ανάλογα το μέγεθος της (μικρή, μεσαία, μεγάλη ή και καθόλου).

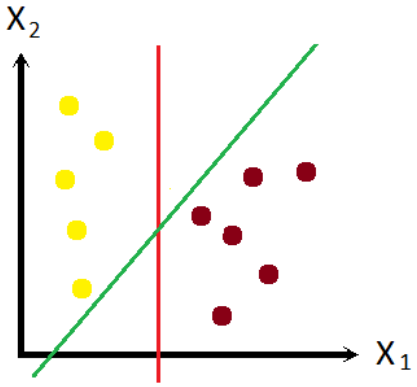
2.2 Support Vector Machine

Οι **μηχανές διανυσμάτων υποστήριξης (SVM)** είναι ένας από του πιο διαδεδομένους αλγόριθμους της επιτηρούμενης μάθησης ο οποίος λύνει προβλήματα ταξινόμησης(Classification) και παλινδρόμησης (Regression). Η συνηθέστερη χρήση του σε προβλήματα μηχανικής μάθησης είναι προβλήματα ταξινόμησης διότι έχει την τάση να μην γίνεται υπέρ-προσαρμογή (overfit) κατά την εκπαίδευση. Ο σκοπός του αλγόριθμου αυτού είναι να δημιουργήσει την καλύτερη γραμμή ή αλλιώς τα όρια απόφασης που μπορεί να διαχωρίσει τον n-διαστάσεων χώρο σε κλάσεις, έτσι ώστε να μπορεί στο μέλλον να διαχωρίσει και να τοποθετεί στην σωστή κλάση τα δεδομένα εισόδου. Αυτή η καλύτερη γραμμή ή αλλιώς όριο απόφασης ονομάζεται υπέρ-επίπεδο (hyperplane). Το υπέρ-επίπεδο τοποθετείται στην μέγιστη απόσταση από τα σημεία δεδομένων. Τα δεδομένα που βρίσκονται στην ελάχιστη απόσταση από το υπέρ-επίπεδο ονομάζονται διανύσματα υποστήριξης. Η απόσταση των διανυσμάτων υποστήριξης από το υπέρ-επίπεδο ονομάζεται περιθώριο(margin). Λόγω της κοντινής τους θέσης στην συγκεκριμένη θέση του υπέρ-επιπέδου, η επιρροή τους είναι μεγαλύτερη από άλλα

σημεία δεδομένων[25]. Όπως παρατηρούμε στο παρακάτω σχήμα τα διανύσματα υποστήριξης από τις δυο κλάσεις που βρίσκονται κοντά στις γραμμές είναι 4,ένα από τη μια κλάση(κίτρινη) και 3 από τη δεύτερη(κόκκινη).

Οι μηχανές διανυσμάτων υποστήριξης χωρίζονται σε 2 κατηγορίες, **στις γραμμικές μηχανές διανυσμάτων υποστήριξης** (Linear SVM) και στις **μη-γραμμικές μηχανές διανυσμάτων υποστήριξης** (Non-linear SVM)[26].

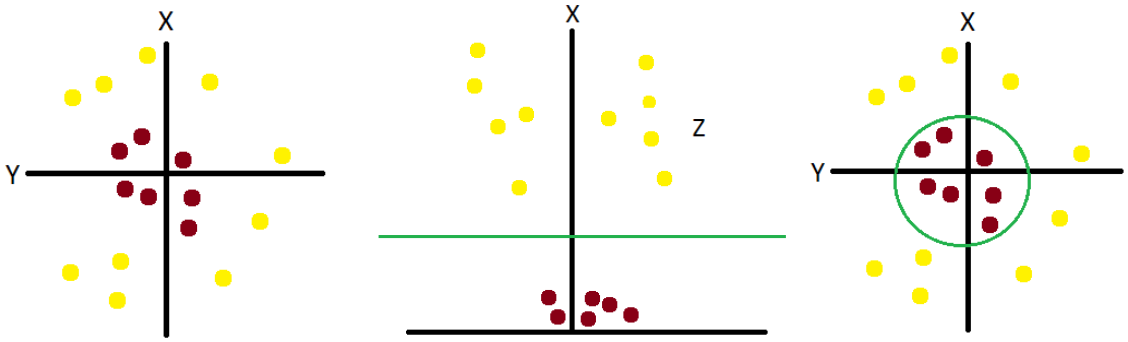
Οι **γραμμικές μηχανές διανυσμάτων υποστήριξης** χρησιμοποιούνται για προβλήματα που ο γραμμικός διαχωρισμός δεδομένων είναι απαραίτητος. Τα δεδομένα εκπαίδευσης χωρίζονται σε 2 κλάσεις μέσω μιας γραμμής διαχωρισμού. Τέτοιο είδος διαχωρισμός δεδομένων χαρακτηρίζεται ως γραμμικός διαχωρισμός δεδομένων.



Σχήμα 2. Απεικόνιση δεδομένων στο δισδιάστατο χώρο και διαχωρισμός των κλάσεων

Στο σχήμα 2 είναι τα δεδομένα τοποθετημένα στο δισδιάστατο χώρο και εκεί φαίνονται οι 2 κλάσεις όπου ανάμεσα σε αυτές βρίσκεται η διαχωριστική γραμμή που τις διαχωρίζει. Όπως βλέπουμε είναι εύκολο να τα διαχωρίσουμε απλά τοποθετώντας μια διαχωριστική γραμμή.

Οι **μη-γραμμικές μηχανές διανυσμάτων υποστήριξης** χρησιμοποιούνται για προβλήματα που ο διαχωρισμός δεδομένων δεν είναι γραμμικός. Που σημαίνει ότι τα δεδομένα εκπαίδευσης δεν μπορούν να διαχωριστούν και να τοποθετηθούν σε κλάσεις με μια διαχωριστική γραμμή. Τέτοιο είδος διαχωρισμού δεδομένων χαρακτηρίζεται ως μη-γραμμικός διαχωρισμός δεδομένων.



Σχήμα 3. Απεικόνιση δεδομένων στο δισδιάστατο και τρισδιάστατο χώρο και τα βήματα για το διαχωρισμό κλάσεων

Στο σχήμα 3 φαίνονται οι 2 κλάσεις όπου παρατηρείται ότι δεν είναι γραμμικά τοποθετημένες. Σε τέτοιες περιπτώσεις πρέπει να προστεθούν παραπάνω χώροι. Στην παραπάνω περίπτωση πηγαίνει στο τρισδιάστατο χώρο. Ως εκ τούτου διαχωρίζονται οι κλάσεις της με μια κυκλική ακτίνα, στην συγκεκριμένη περίπτωση, των μη-γραμμικών δεδομένων.

Στη μηχανική μάθηση οι μηχανισμοί πυρήνων είναι κατηγορία αλγόριθμων για ανάλυση μοτίβων. Ένας από αυτούς είναι οι μηχανές διανυσμάτων υποστήριξης. Αυτό δίνει την δυνατότητα να μετατρέπει τον χώρο των δεδομένων εισόδου σε χώρο υψηλότερης διάστασης.

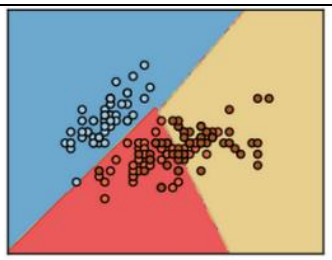
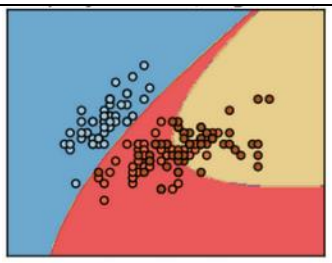
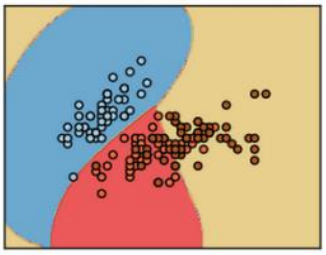
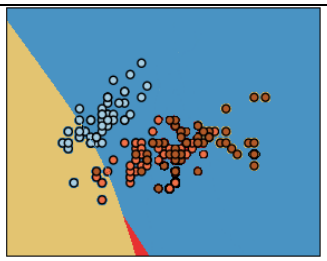
Ο χώρος εισόδου X αποτελείται από τα x και x'[26].

$$\Phi(\chi_\iota)$$

και αντιπροσωπεύει τη συνάρτηση πυρήνα που μετατρέπει τον χώρο εισόδου σε χώρο υψηλότερης διάστασης, έτσι ώστε να μην αντιστοιχίζεται κάθε σημείο δεδομένων[26]. Η συνάρτηση πυρήνα καθορίζεται ως

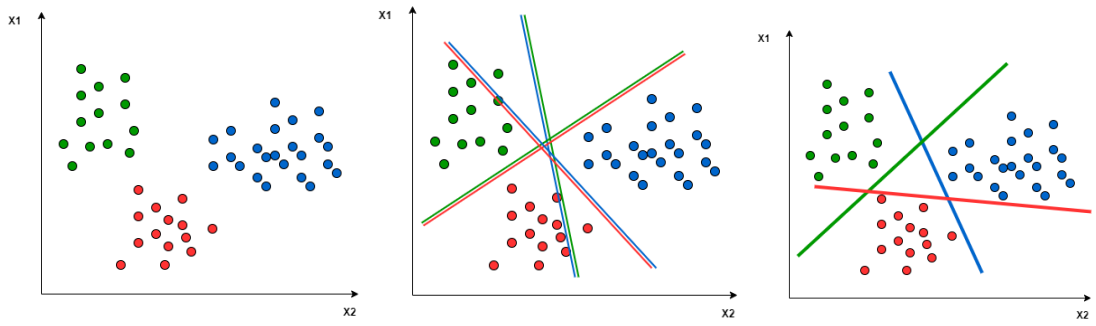
$$k(x, x')$$

Ένα βασικό χαρακτηριστικό των μηχανών διανυσμάτων υποστήριξης είναι η επιλογή του πυρήνα.(Kernel). Ανάλογα τον πυρήνα που θα επιλεχθεί, ο αλγόριθμος κάνει σύνθετους μετασχηματισμούς δεδομένων για να μεγιστοποιήσει τα όρια διαχωρισμού μεταξύ των σημείων των δεδομένων εκπαίδευσης ανάλογα τις κλάσεις ή αλλιώς τις ετικέτες που έχουν οριστεί από το χρήστη. Οι πιο διαδεδομένοι πυρήνες των μηχανών διανυσμάτων υποστήριξης απεικονίζονται στον παρακάτω πίνακα.

Γραμμική συνάρτηση (Linear Function)	$k(x_i, x_j) = x_i * x_j$	
Πολυωνυμική συνάρτηση (Polynomial Function)	$k(x_i, x_j) = (1 + x_i * x_j)^d$	
Λειτουργία ακτινικής βάσης (Radial Basis Function)	$k(x_i, x_j) = \exp\left(-\gamma \left\|\left\|x_i - x_j\right\|\right\|^2\right)$	
Σιγμοειδής συνάρτηση (Sigmoid Function)	$k(x_i, x_j) = tahn(ax^T y + c)$	

Πίνακας 1. Πυρήνες των μηχανών διανυσμάτων υποστήριξης

Οι γραμμικές μηχανές διανυσμάτων υποστήριξης μπορούν, όπως αναφέρθηκε και παραπάνω, να λύσουν προβλήματα ταξινόμησης πολλαπλών κλάσεων. Στην πραγματικότητα ο αλγόριθμος δεν υποστηρίζει ταξινόμηση πολλαπλών κλάσεων. Αυτό που υποστηρίζει είναι η δυαδική ταξινόμηση (2 κλάσεων). Για να επιτευχθεί αυτό θα χρησιμοποιηθεί ο ίδιος κανόνας, το πρόβλημα πολλαπλής ταξινόμησης να διασπαστεί σε πολλαπλές δυαδικές ταξινομήσεις κλάσεων και στην συνέχεια να εκτελεστεί διαδοχικά δυαδική ταξινόμηση σε κάθε ζεύγος κλάσεων. Αυτή η μέθοδος ονομάζεται ένας εναντίων ενός (one vs one).Μια άλλη μέθοδος πολλαπλής ταξινόμησης είναι η μέθοδος ένας εναντίων των υπόλοιπων (one vs rest) όπου εδώ το πρόβλημα πολλαπλής ταξινόμησης διασπάται πάλι σε πολλαπλές δυαδικές ταξινομήσεις κλάσεων, οι οποίες εκτελούνται για κάθε κλάση ξεχωριστά. Αν υποθέσουμε ότι σε ένα σύνολο δεδομένων έχουμε x κλάσεις, τότε στην μέθοδο που ονομάζεται ένας εναντίων ενός θα γίνει χρήση του αλγόριθμου $\frac{x(x-1)}{2}$ φορές. Στην μέθοδο ένας εναντίων των υπόλοιπων αντίστοιχα θα γίνει χρήση του αλγόριθμου x φορές.



Σχήμα 4. Απεικόνιση δυο προσεγγίσεων για ταξινόμηση σε κλάσεις [27]

Βλέποντας ένα παράδειγμα προβλήματος πολλαπλής ταξινόμησης (σχήμα 4 α) έχουμε τρεις κλάσεις πράσινο, κόκκινο και μπλε. Χρησιμοποιώντας την μέθοδο ένας εναντίων ενός (σχήμα 4 β) παρατηρείται ότι το υπέρ-επίπεδο τοποθετείται ανάμεσα σε κάθε 2 κλάσεις αγνοώντας την τρίτη κλάση κάθε φορά αντίστοιχα. Χρησιμοποιώντας την μέθοδο ένας εναντίων των υπολοίπων (σχήμα 4 γ) παρατηρείται ότι το υπέρ-επίπεδο χωρίζει κάθε κλάση μεταξύ των υπολοίπων ταυτόχρονα. Ο διαχωρισμός λαμβάνει υπόψην όλα τα δεδομένα χωρίς να αγνοήσει κάποια κλάση όπως παρατηρήθηκε στην μέθοδο ένας εναντίων ενός. Η χρήση της μεθόδου ένας εναντίων ενός προτιμάται συνήθως σε περίπτωση που το σύνολο των δεδομένων είναι μεγάλο και έχει πολλές κλάσεις. Η μέθοδος ένας εναντίων των υπολοίπων για το λόγο ότι δεν μπορεί να διαχειριστεί μεγάλο σύνολο δεδομένων καθίσταται πιο γρήγορη σε μικρά σύνολα δεδομένων, αφού πραγματοποιείται εκπαίδευση σε μικρότερο αριθμό κλάσεων.

Πίσω από τις γραμμικές μηχανές διανυσμάτων υποστήριξης κρύβεται μια μαθηματική έννοια, αυτή καθιστά εφικτή την υλοποίηση του. Αρχικά οι γραμμικές μηχανές διανυσμάτων υποστήριξης σχεδιάστηκαν για προβλήματα δυαδικής ταξινόμησης. Στόχος του μοντέλου είναι να βρει την συνάρτηση που θα μεγιστοποιεί τα περιθώρια μεταξύ των σημείων των δεδομένων των δύο κλάσεων και του υπέρ-επιπέδου. Αυτή η ιδανική συνάρτηση του υπέρ-επιπέδου μπορεί να βρεθεί με την αρχική μορφή του αλγορίθμου.

$$\min E_0\,(w,\xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi\rho \quad [26]$$

το οποίο εξαρτάται από τον παρακάτω τύπο.

$$yp\left(w^T x_p + \Theta\right) \geq 1 - \xi_p, p = 1, \dots, N,$$

$$\xi p \geq 0, p = 1, \dots, N, \quad [26]$$

όπου $\boldsymbol{\xi} = (\xi_1 \dots, \xi_N)^T$ και $\boldsymbol{\xi_p}$ είναι χαλαρές μεταβλητές (slack variables), **w** είναι τα βάρη και το **θ** η κλίση, τα οποία είναι χρήσιμα για τον υπολογισμό του υπέρ-επιπέδου. Το $y_p \in \{-1,+1\}$ περιγράφει την έξοδο της ταξινόμησης. Το C είναι μία ρυθμιστική παράμετρος η οποία δημιουργεί προβλήματα κατά τον υπολογισμό των περιθωρίων στην αρχική μορφή του αλγόριθμου.

Υπάρχουν βελτιωμένοι αλγόριθμοι γραμμικών μηχανών διανυσμάτων υποστήριξης όπως ο L1 και L2. Ο L1 βελτιώνει και επιλύει τα παραπάνω προβλήματα χωρίς περιορισμούς και

χρησιμοποιείται κατά βάση λόγω της ταχύτητας του στα προβλήματα ταξινόμησης και τη δυνατότητα ταξινόμησης με μικρότερο αριθμό διανυσμάτων ταξινόμησης[26]. Ο L2 είναι πολύ δημοφιλές για προβλήματα ταξινόμησης 2 κλάσεων.

$$\boldsymbol{L1} \quad min_w J(w) = \frac{1}{2} w^T w + C \sum_{j=1}^{l} \max\left(1 - y_j w^T x_j, 0\right)$$

$$\boldsymbol{L2} \quad min_w J(w) = \frac{1}{2} w^T w + C \sum_{j=1}^{l} \left[\max\left(1 - y_j w^T x_j, 0\right)\right]^2$$

2.3 Naïve Bayes

Ο Naïve Bayes είναι ένα μοντέλο μηχανικής μάθησης το οποίο βασίζεται πάνω στη πιθανότητα μέσω των στατιστικών μεθόδων για προβλήματα ταξινόμησης. Ο Naïve Bayes χρησιμοποιείται σε πολύπλοκες εφαρμογές και αυτό λόγω της απλότητας του, που επιτρέπει όλες οι μεταβλητές του να επηρεάσουν την τελική απόφαση λόγω της ανεξαρτησίας του. Αυτή η απλότητα έχει ως αποτέλεσμα μικρότερη χρήση υπολογιστικής ισχύς και υψηλή ακρίβεια πρόβλεψης, κάνοντας την ελκυστική για την χρήση της σε διάφορες εφαρμογές, όπως επίσης και σε μεγάλα σύνολα δεδομένων[28]. Σε σχέση με άλλους αλγόριθμους είναι γρήγορος κατά τη διαδικασία πρόβλεψης και μπορεί να χρησιμοποιηθεί για real-time προβλέψεις. Άλλα οφέλη του αλγόριθμου είναι η εύκολη διαχείριση δεδομένων με θόρυβο, με άσχετα χαρακτηριστικά, με ελλιπή δεδομένα, όπως επίσης και με μικρό αριθμό δεδομένων εκπαίδευσης. Ένα βασικό μειονέκτημα του Naïve Bayes είναι ότι εάν στις μεταβλητές κατηγορίας των δεδομένων δοκιμής δεν εντοπιστεί η μεταβλητή στα δεδομένα εκπαίδευσης τότε το μοντέλο θα το εκλάβει ως πιθανότητα 0 με αποτέλεσμα να μην μπορεί να γίνει πρόβλεψη. Αυτό μπορεί να λυθεί όμως με μια τεχνική που ονομάζεται Laplace estimation

Η μαθηματική έννοια που χαρακτηρίζει τον Naïve Bayes είναι

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \qquad [28]$$

Όπου

X δεδομένα άγνωστης κλάσης X=(x_1 ,x_2,…,x_n), όπου τα x_1 ,x_2,…,x_n αντιπροσωπεύουν τα χαρακτηριστικά των δεδομένων.

c πιθανότητα της κλάσης

$P(c|x)$ μεταγενέστερη πιθανότητα(posterior probability) της προβλεπόμενης κλάσης c

$P(x|c)$ πιθανότητα (likelihood) της προβλεπόμενης κλάσης.

$P(x)$ πιθανότητα x

$P(c)$ προβλεπόμενη προηγούμενη πιθανότητα (prior probability)

Αντικαθιστώντας και κάνοντας χρήση του Chain Rule καταλήγουμε στον παρακάτω τύπο. Για όλες τις καταχωρίσεις στο σύνολο δεδομένων ο παρονομαστής παραμένει σταθερός. Επομένως θα μπορούσε αφαιρεθεί και να εφαρμοστεί η αναλογικότητα όπως παρατηρείται παρακάτω.

$$P(c|X) = \frac{P(x_1|c) * P(x_2|c) * \dots * P(x_n|c) * P(c)}{P(x_1)P(x_2) \dots P(x_n)}$$

$$P(c|X)\alpha\ P(c) \prod_{i=1}^{n} P(x_i|c)$$

Ο Naïve Bayes χρησιμοποιείται επίσης και σε προβλήματα παραπάνω από 2 κλάσεις, επομένως χρησιμοποιώντας τον παρακάτω τύπο μπορεί να βρεθεί η κλάση c με την μεγαλύτερη πιθανότητα βάση των προγνωστικών[29].

$$c = argmax_c P(c) \prod_{i=1}^{n} P(x_i|c)$$

Γνωστή μέθοδος του αλγόριθμου Naïve Bayes είναι η μέθοδος πολυωνυμικό Naïve Bayes (Multinomial),η οποία είναι ένα μοντέλο συμβάντων που συνήθως χρησιμοποιείται σε προβλήματα ταξινόμησης εγγράφων. Τα χαρακτηριστικά των διανυσμάτων αντιπροσωπεύουν τις συχνότητες που έχουν δημιουργήσει συγκεκριμένα συμβάντα εξαιτίας της πολυωνυμικής κατανομής, το πόσο συχνά δηλαδή παρουσιάζεται μια λέξη στο έγγραφο. Άλλη μέθοδος είναι η Bernoulli Naive Bayes, η οποία είναι παρόμοια με την Multinomial με την διαφορά ότι εδώ οι προβλέψεις είναι μεταβλητές Boolean (Σωστό(1)/Λάθος(0)). Δηλαδή αν μια λέξη εμφανίζεται στο κείμενο η όχι. Επίσης γνωστή μέθοδος για προβλέψεις με δεδομένα συνεχή στο χρόνο που σχετίζονται με κάθε κλάση και όχι διακριτές είναι η γκαουσιανή Naïve Bayes(Gaousian) μέθοδος. Οι τιμές αυτές κατανέμονται σύμφωνα με την κανονική ή την γκαουσιανή κατανομή. Κατά την δημιουργία της παρουσιάζεται μια καμπύλη σε σχήμα καμπάνα, η οποία είναι συμμετρική ως προς το μέσο όρο των τιμών των χαρακτηριστικών[30].

Η πιθανότητα των χαρακτηριστικών δίνεται από τον παρακάτω τύπο:

$$P(x_i|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(x_i-\mu_c)^2}{2\sigma_c^2}\right)$$ [29]

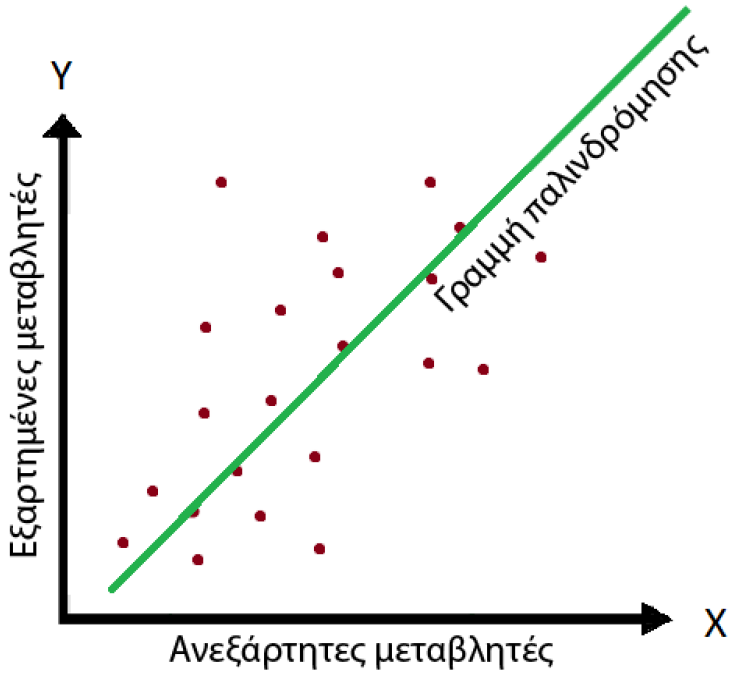
Όπου

μ_c ο μέσος όρος των τιμών στο x που σχετίζεται με την κλάση c

σ_c^2 η διορθωμένη διακύμανση Bessel των τιμών που σχετίζεται με την κλάση c

2.4 Γραμμική παλινδρόμηση (Linear Regression)

Η γραμμική παλινδρόμηση είναι εξίσου ένα μοντέλο μηχανικής μάθησης, το οποίο χρησιμοποιεί τεχνικές και εξετάζει τις σχέσεις μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Υπολογίζοντας τη σχέση μεταξύ των μεταβλητών, το μοντέλο μπορεί να προβλέψει μια εξαρτημένη μεταβλητή βάση ενός η παραπάνω ανεξάρτητων μεταβλητών[31]. Χρησιμοποιώντας την σχέση αυτών των μεταβλητών βρίσκει την ιδανική γραμμή την οποία και χρησιμοποιεί για την πρόβλεψη, όπως φαίνεται στο σχήμα 5. Αυτή η γραμμή ονομάζεται γραμμική παλινδρόμηση και στόχος του αλγόριθμου είναι να βρεθεί η ιδανική γραμμή με το μικρότερο σφάλμα μεταξύ των εξαρτημένων και ανεξάρτητων μεταβλητών .



Σχήμα 5. Αναπαράσταση δεδομένων χρησιμοποιώντας γραμμική παλινδρόμηση

Τα σύνολα δεδομένων που χρησιμοποιούνται συνήθως στην γραμμική παλινδρόμηση αποτελούνται από συνεχή δεδομένα (Continuous data). Στην πτυχιακή εργασία το σύνολο δεδομένων δεν αποτελείται από συνεχή δεδομένα, παρόλα αυτά θα δοκιμαστεί η γραμμική παλινδρόμηση για να αποδειχτεί η χαμηλή ακρίβεια πρόβλεψης λόγω της γραμμικότητας. Γνωστά μοντέλα παλινδρόμησης εκτός της γραμμικής παλινδρόμησης είναι και η Πολυωνυμική Παλινδρόμηση (Polynomial Regression) όπως και η Λογιστική Παλινδρόμηση (Logistic regression). Συνήθης χρήση της πολυωνυμικής παλινδρόμησης είναι η προγνωστική ανάλυση όπως για προβλέψεις τιμών προϊόντων, μισθών εργαζομένων, πωλήσεων και γενικά για προβλέψεις που κάνουν χρήση συνεχών δεδομένων.

Η μαθηματική έννοια που χαρακτηρίζει την γραμμική παλινδρόμηση είναι

$$y = b_0 + b_1 x + e$$

Όπου

y εξαρτημένη μεταβλητή

x ανεξάρτητη μεταβλητή

b_0 είναι το σημείο όπου η ευθεία τέμνει τον άξονα y'y

b_1 συντελεστής γραμμικής παλινδρόμησης.

e σφάλμα

Τα x και y είναι οι τιμές των δεδομένων εκπαίδευσης και τα b_0,b_1είναι τα βάρη της συνάρτησης. Ανάλογα τις τιμές που έχουν, δίνουν και την ανάλογη γραμμή παλινδρόμησης. Στόχος είναι να υπολογιστούν αυτές οι τιμές, έτσι ώστε να ελαχιστοποιηθεί το σφάλμα και να επιλεχθεί η καλύτερη γραμμή. Αυτό επιτυγχάνεται με την συνάρτηση του κόστους. Για να βρεθεί το σφάλμα πρέπει να υπολογιστεί το άθροισμα του τετραγώνου του σφάλματος και βελτιώνοντας τις παραμέτρους κατάλληλα πετυχαίνεται η μείωση του σφάλματος. Η παρακάτω συνάρτηση μας δίνει το κόστος συνάρτησης.

$$MSE = \frac{1}{2}\sum_{i=1}^{m}(y_i - (b_1 x_i + b_0))^2$$

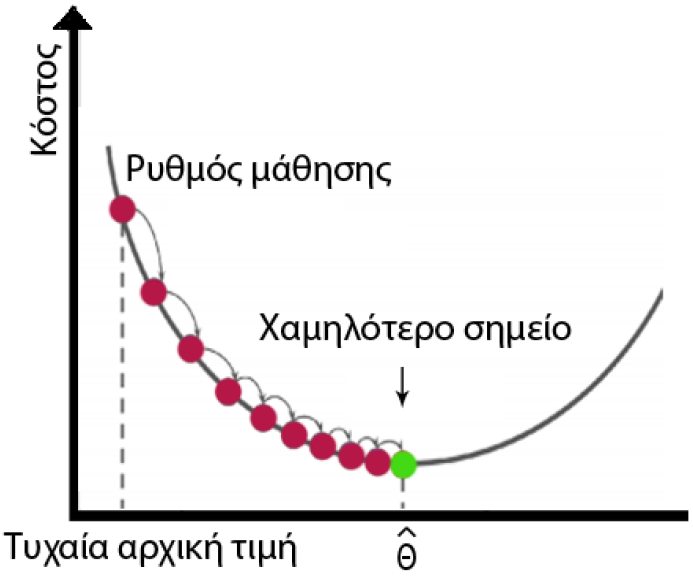
Όπου

MSE συνάρτηση κόστους ή αλλιώς η συνάρτηση σφάλματος

$b_1 x_i + b_0$ η προβλεπόμενη μεταβλητή

y_i η πραγματική μεταβλητή

Αφού έχει υπολογιστεί η συνάρτηση κόστους, πρέπει να ελαχιστοποιηθεί για να επιτευχθεί μεγαλύτερη ακρίβεια πρόβλεψης. Μπορεί να ελαχιστοποιηθεί με την χρήση του Gradient descent υπολογίζοντας την κλίση της συνάρτησης κόστους. Η κύρια χρήση του αλγόριθμου Gradient descent είναι η τυχαία επιλογή τιμών του συντελεστή γραμμής και η συνεχής ενημέρωση αυτών των τιμών με στόχο την ελαχιστοποίηση της συνάρτησης κόστους[32]. Όπως φαίνεται στην σχήμα 6, παρατηρείται ένα υπόδειγμα του Gradient descent. Στόχος είναι να φτάσει στο χαμηλότερο σημείο από αριστερά προς τα δεξιά. Τα μεγάλα βήματα που θα κάνει κάθε φορά στην ενημέρωση ή αλλιώς επανάληψη ονομάζεται ρυθμός μάθησης(learning rate). Η επιλογή του ρυθμού είναι ένα σημαντικό κομμάτι διότι αν επιλεχθεί μεγάλο μπορεί να ξεπεράσει το ελάχιστο σημείο. Το ιδανικό είναι, όπως φαίνεται και στο σχήμα, να επιλεχθεί στην αρχή μεγάλος ρυθμός και στην συνέχεια να μειώνεται ώστε έτσι να είναι εφικτή η επίτευξη στο χαμηλότερο σημείο.



Σχήμα 6. Παράδειγμα της ελαχιστοποίησης του κόστους χρησιμοποιώντας Gradient Decent

Η μαθηματική έννοια του Gradient Decent είναι [33]

Για θ_0

$$\theta_0 = \theta_0 - a\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x_i) - yi)$$

Για θ_1

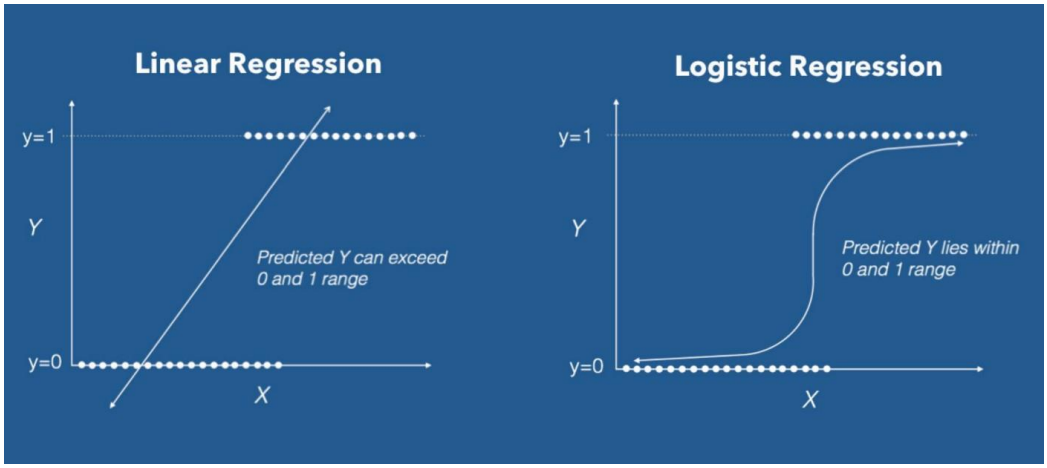
$$\theta_1 = \theta_1 - a\frac{1}{m}\sum_{i=1}^{m}((h_\theta(x_i) - yi)x_i)$$

Όπου

$(h_\theta(x_i) - yi)$	αντιπροσωπεύει την παράγωγο της συνάρτησης κόστους.
α	είναι ο ρυθμός μάθησης.
$\theta_{0,1}$	είναι οι παράμετροι της γραμμής παλινδρόμησης θ_0 η τομή της y και θ_1 είναι η κλήση της γραμμής παλινδρόμησης.

2.5 Λογιστική παλινδρόμηση (Logistic Regression)

Η λογιστική παλινδρόμηση όπως και η γραμμική είναι ένα μοντέλο της μηχανικής μάθησης. Σε αντίθεση με την γραμμική παλινδρόμηση, που χρησιμοποιείται για προβλήματα παλινδρόμησης, η λογιστική χρησιμοποιείται για προβλήματα ταξινόμησης. Ο αλγόριθμος βασίζεται στη λογική της πιθανότητας. Προβλέπει την έξοδο μιας εξαρτημένης κατηγορικής μεταβλητής και την επιστρέφει ως μια τιμή πιθανότητας χρησιμοποιώντας τη σιγμοειδή συνάρτηση(sigmoid function). Δίνει τη πιθανότητα του αποτελέσματος που μπορεί να είναι (ναι, όχι η 0, 1) με τιμές που κυμαίνονται μεταξύ του 0 και του 1. Στη παρακάτω εικόνα απεικονίζεται η διαφορά της γραμμικής παλινδρόμησης, όπου χρησιμοποιεί την γραμμή παλινδρόμησης για πρόβλεψη. Οι εξαρτημένες μεταβλητές μπορούν να πάρουν τιμές που ξεπερνούν τα όρια 0 και 1 που έχουμε στην λογιστική παλινδρόμηση.



Σχήμα 7. Διαφορές Γραμμικής παλινδρόμησης και Λογιστικής παλινδρόμησης [34]

Η μαθηματική έννοια που χαρακτηρίζει την σιγμοειδή συνάρτηση είναι

$$\sigma(y) = \frac{1}{1 + e^{-y}}$$

Όπου

$\sigma(y) = \sigma(b_0 + b_1 x)$ η μεταβλητές εισόδου.

Οι τιμές b_0, b_1 και x έχουν την ίδια σημασία όπως και στην γραμμική παλινδρόμηση και ανεξαρτήτως τις τιμές που έχουν είναι είτε αρνητικές είτε θετικές δίνοντας ως έξοδο τιμές που περιορίζονται μεταξύ του 0 και 1. Οι τιμές που θα έχει η y θα είναι μεταξύ του 0 και του 1.

$$h_\theta(x) = \sigma(y)$$

Όπου

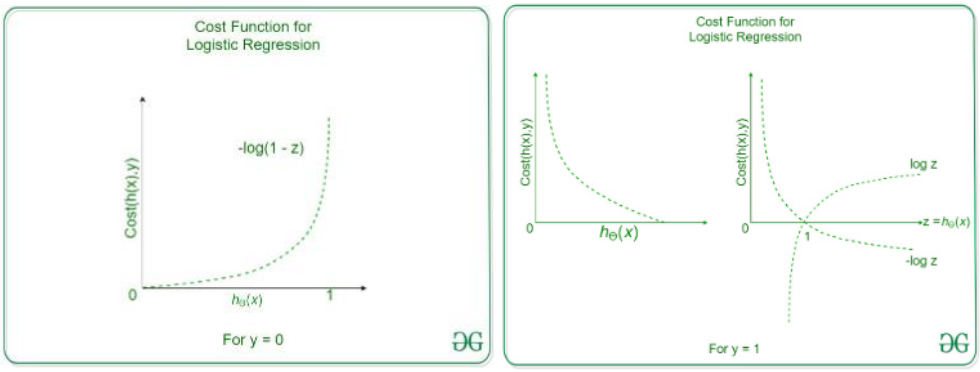
$h_\theta(x)$ η πιθανότητα να συμβεί το γεγονός (Προβλεπτική μεταβλητή),

σ(y) σιγμοειδής συνάρτηση.

Ένα από τα πλεονεκτήματα της λογιστικής παλινδρόμησης είναι η δυνατότητα να διαχειριστεί σύνολα δεδομένων που αποτελούνται από συνεχή ή και διακριτές τιμές και να αντιστοιχήσει τις προβλέψεις με πιθανότητες. Η συνάρτηση κόστους είναι πιο πολύπλοκη από ότι στην γραμμική παλινδρόμηση και περιορίζεται μεταξύ τιμών 0 και 1. Ο διαχωρισμός της κλάσης καθορίζεται από την τιμή κατωφλιού που έχει οριστεί, σε ποια τιμή μεταξύ του 0 και του 1.

Όπως αναφέρθηκε και στην γραμμική παλινδρόμηση η συνάρτηση κόστους ελαχιστοποιεί το σφάλμα έτσι ώστε να υπάρχει υψηλή ακρίβεια με χαμηλό σφάλμα. Η συνάρτηση κόστους δίνεται από το παρακάτω τύπο

$$C(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)), & if\ y = 1 \\ -\log(1 - h_\theta(x)), & if\ y = 0 \end{cases}$$



Σχήμα 8. Απεικόνιση συμπεριφοράς της συνάρτησης κόστους στη λογιστική παλινδρόμηση [35]

Απλοποιώντας τον παραπάνω τύπο η συνάρτηση κόστους καταλήγει

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^i \log\left(h_\theta(x(i))\right) + (1 - y^i)\log\left(1 - h_\theta(x(i))\right)\right]$$

Όπως φαίνεται από τα παραπάνω γραφήματα (σχήμα 8) το κόστος θα είναι ίσο με το 0 αν y= 1 αλλά αν το hθ(x) ίσο με 0 τότε το κόστος απειρίζεται. Η συνάρτηση κόστους έχει μια λογική

και αυτό επειδή το -log(x) αυξάνεται κατά πολύ όταν η τιμή του x πλησιάζει το 0, οπότε το κόστος θα είναι μεγάλο αν το μοντέλο προβλέψει πιθανότητα κοντά στο 0 για μια θετική περίπτωση. Επίσης θα είναι πολύ μεγαλύτερο αν το μοντέλο προβλέψει κοντά στο 1 για μια αρνητική περίπτωση. Άρα το -log(x) θα είναι κοντά στο 0 όταν η τιμή x είναι κοντά στο 1, οπότε το κόστος θα είναι κοντά στο 0 για μια πρόβλεψη κοντά στο 0 σε μια αρνητική περίπτωση ή κοντά στο 1 για μια θετική [36].

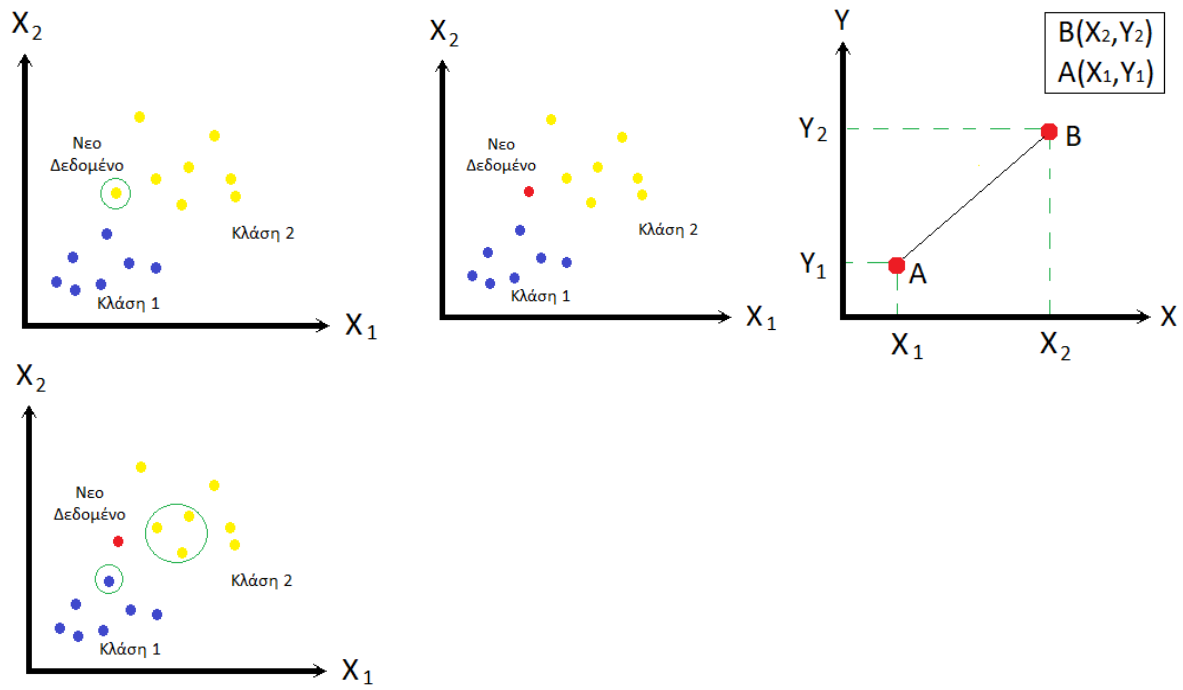
Όπως στην γραμμική παλινδρόμηση έτσι και στην λογιστική πρέπει να ελαχιστοποιηθεί η τιμή του κόστους περαιτέρω. Αυτό μπορεί να πραγματοποιηθεί χρησιμοποιώντας το Gradient Descent όπου θα γίνεται ταυτόχρονη αντικατάσταση όλων τον παραμέτρων διαδοχικά.

$$\theta_j - a\frac{\partial}{\partial\theta_j}J(\theta) = \theta_j - a\sum_{i=1}^{m}\left[h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x_j^{(i)}\Big]$$

2.6 Κ-Κοντινότεροι γείτονες(KNN)

Ο αλγόριθμος Κ-Κοντινότεροι γείτονες είναι ένας από τους απλούστερους αλγορίθμους μηχανικής μάθησης που βασίζεται στην τεχνική επιβλεπόμενης μάθησης. Μπορεί να χρησιμοποιηθεί για τα προβλήματα ταξινόμησης όπως επίσης και για προβλήματα παλινδρόμησης. Ο αλγόριθμος KNN υποθέτει βάση της ομοιότητας δεδομένων που υπάρχουν σε κοντινή απόσταση. Με απλά λόγια, δίνοντας δεδομένα ως είσοδο, ο αλγόριθμος συγκρίνει τις ομοιότητες , την απόσταση μεταξύ των νέων δεδομένων με των ήδη υπάρχοντων δεδομένων και τα ταξινομεί στην κατηγορία που μοιάζει περισσότερο με τις υπάρχουσες κατηγορίες. Κατά την εκπαίδευση των δεδομένων ο αλγόριθμος δεν μαθαίνει από το σύνολο εκπαίδευσης αμέσως, αλλά αποθηκεύει το σύνολο δεδομένων και κατά τη στιγμή της ταξινόμησης εκτελεί μια ενέργεια στο σύνολο δεδομένων. Αυτή η ενέργεια είναι να ταξινομεί τα νέα δεδομένα σε μια κλάση που τα δεδομένα είναι όμοια με αυτά των νέων δεδομένων που δόθηκαν[37].

Θέλοντας να λύσουμε ένα πρόγραμμα ταξινόμησης χρησιμοποιούμε τον αλγόριθμο KNN, οπότε θέτουμε ένα νέο δεδομένο που θέλουμε να προβλεφθεί η κλάση του. Θα πρέπει να προεπιλεχθεί ο αριθμός **K**, ο οποίος καθορίζει πόσοι γείτονες θα χρησιμοποιηθούν κατά την εκτέλεση του αλγορίθμου.(σχήμα 9 βήμα 1) Στη συνέχεια επιλέγεται ο αριθμός **K** γειτόνων και υπολογίζεται η ευκλείδεια απόσταση αυτών των γειτόνων(σχήμα 9 βήμα 2). Αφού υπολογιστεί η ευκλείδεια απόσταση μεταξύ των κοντινών σημείων τότε φανερώνονται οι κοντινοί γείτονες και οι κλάσεις που ανήκουν (σχήμα 9 βήμα 3).Στη Κλάση 2 ανήκουν τρεις γείτονες σε αντίθεση με τη κλάση 1 όπου ανήκει μόνο ένας. Το νέο δεδομένο τοποθετείται στην κλάση με τα περισσότερα κοντινά σημεία. (σχήμα 9 βήμα 4)



Σχήμα 9. Βήματα ταξινόμησης δεδομένων χρησιμοποιώντας τον αλγόριθμο KNN

Υπάρχουν πολλοί μέθοδοι εκτίμησης της απόστασης, και ανάλογα με το θέμα που εξετάζεται, πρέπει να επιλέγεται η κατάλληλη μέθοδος. Ωστόσο, μια ευρέως διαδεδομένη και γνωστή επιλογή είναι η ευθεία απόσταση (γνωστή και ως ευκλείδεια απόσταση).

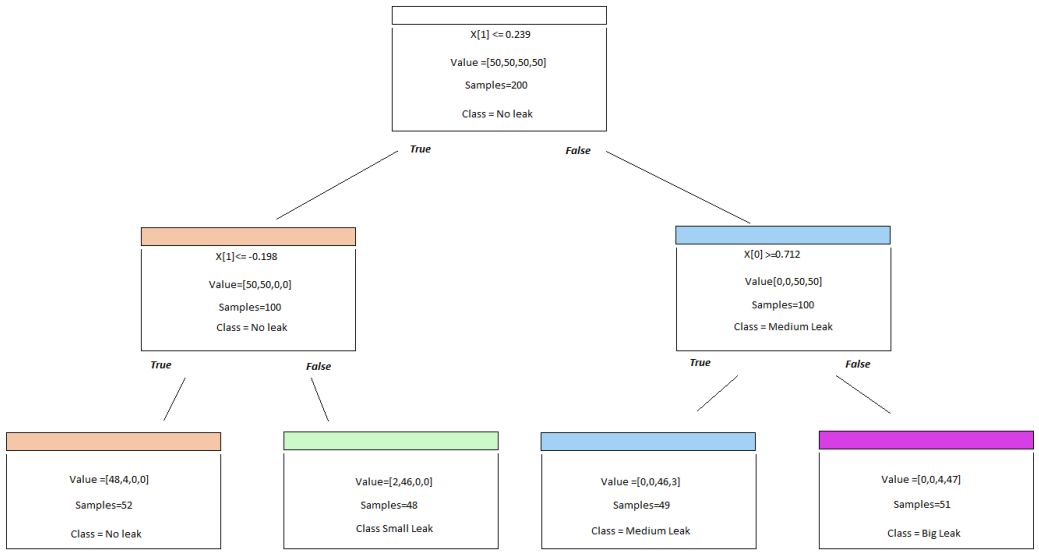
Η μαθηματική έννοια που χαρακτηρίζει την ευκλείδεια συνάρτηση

$$A_1, A_2 = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$ [37]

Παρόλο που ο αλγόριθμος είναι απλός και εύκολος στην υλοποίηση του, καθώς δεν χρειάζεται να γίνουν ρυθμίσεις στις παραμέτρους του αλγόριθμου όταν δημιουργείται το μοντέλο, το κύριο μειονέκτημα του είναι ότι γίνεται σημαντικά αργό καθώς αυξάνεται ο όγκος των δεδομένων του και το καθιστά μη πρακτική επιλογή σε περιβάλλοντα όπου οι προβλέψεις πρέπει να γίνονται γρήγορα[38].

2.7 Δέντρα απόφασης (Decision Trees)

Τα Δέντρα απόφασης είναι ένα κομμάτι της μηχανικής μάθησης και αποτελούν βασικά χαρακτηριστικά των Random Forest. Η χρήση τους δεν περιορίζεται μόνο σε προβλήματα παλινδρόμησης αλλά και σε προβλήματα ταξινόμησης. Θεωρούνται ένας από τους πιο ισχυρούς αλγόριθμους λόγω της δυνατότητας να διαχειριστούν πολύπλοκα σύνολα δεδομένων. Ο τρόπος λειτουργίας των δέντρων απόφασης είναι ίδιος με τον τρόπο σκέψης των ανθρώπων. Για αυτό το λόγο είναι εύκολο να διαχειριστούν τα δεδομένα εισόδου και δημιουργώντας μια λογική αποφάσεων καταλήγουν σε μια πρόβλεψη υψηλής ακρίβειας. Το όνομα τους σχετίζεται με τον τρόπο λειτουργίας τους, όπου ένα δέντρο αποτελείται από κόμβους, κλαδιά και φύλλα. Ο κάθε κόμβος αντιπροσωπεύει ένα χαρακτηριστικό από το σύνολο δεδομένων. Κάθε κλαδί αντιπροσωπεύει μία απόφαση και κάθε φύλλο αντιπροσωπεύει την έξοδο. Αυτή μπορεί να είναι μία τιμή πρόβλεψης για προβλήματα παλινδρόμησης ή κατηγορική τιμή για προβλήματα ταξινόμησης [39]. Για να δημιουργηθούν τα δέντρα απόφασης χρησιμοποιούνται αλγόριθμοι όπως ο CART (Classification and Regression Trees), ο οποίος χρησιμοποιεί τον Gini ή και ο ID3 που χρησιμοποιεί entropy και information gain. Στην περίπτωση χρήσης του αλγόριθμου CART τα κλαδιά που θα δημιουργηθούν από τον κόμβο είναι πάντα δυο σε αντίθεσή με τον αλγόριθμο ID3 όπου μπορούν να είναι και παραπάνω. Ένα παράδειγμα ενός προβλήματος ταξινόμησης για εύρεση διαρροής με χρήση του αλγόριθμου δέντρων απόφασης απεικονίζεται στην παρακάτω εικόνα 3.1.



Σχήμα 10. Αναπαράσταση ενός παραδείγματος ταξινόμησης των Δέντρων αποφάσεων

Το σχήμα 10 απεικονίζει ένα παράδειγμα ταξινόμησης και τον τρόπο απόφασης πρόβλεψης. Για την εύρεση και την ταξινόμηση μεγέθους της διαρροής, η διαδικασία χωρίζεται μέσω των κλαδιών (branches) σε επίπεδα και υπό-επίπεδα (depths) και ξεκινάει από τον αρχικό κόμβο που βρίσκεται στην αρχή του δέντρου. (depth = 0). Ελέγχεται αν το πρώτο χαρακτηριστικό (X[1]<=0,239, τιμές δεύτερης στήλης του συνόλου δεδομένων) είναι αληθές. Εφόσον είναι, προχωράει στο δεύτερο υπό επίπεδο(depth =1) αριστερά και ξανά ελέγχεται το επόμενο χαρακτηριστικό (X[1] <=-0,198).Αφού έχει επιλεχθεί και έχει καταλήξει στο φύλλο (leaf) ελέγχει επίσης αν υπάρχει άλλο υπό επίπεδο (depth = 2), αν όχι τότε ελέγχεται τι κλάση είναι

και στην συνέχεια δίνεται ως απάντηση στην πρόβλεψη του προβλήματος. Η ίδια διαδικασία και αν επιλεγόταν η δεξιά πλευρά. Θα γινόταν έλεγχος για το αν ισχύει το χαρακτηριστικό X[0] (τιμές πρώτης στήλης του συνόλου δεδομένων), και θα συνέχιζε όπως και στην αριστερή διαδρομή που ακολουθήθηκε προηγουμένως. Ένα βασικό στοιχείο είναι τα samples, τα οποία αναπαριστούν τον αριθμό εκπαίδευσης στον κάθε κόμβο. Στο σχήμα 10 παρατηρείται ότι έτρεξε η εκπαίδευση 100 φορές για τιμή X[1] <=0,239 και από αυτές 52 είχαν τιμή X[1] <= -0,198. Επίσης η μεταβλητή Value αντιπροσωπεύει τον αριθμό των στοιχείων σε κάθε κλάση. Όπως παρατηρείται στον κάτω αριστερά κόμβο, όπου ο αλγόριθμος έχει ταξινομήσει από τα 52 τα 48 στοιχεία σωστά στη πρώτη κλάση που δεν υπάρχει διαρροή(no leak) και 46 από τα 48 στοιχεία σωστά για τη δεύτερη κλάση που έχουμε μικρού μεγέθους διαρροή(small leak).

Όπως αναφέρθηκε στην αρχή για να πραγματοποιηθεί και να τρέξει ένα δέντρο απόφασης πρέπει να επιλεχθεί ένας αλγόριθμος όπου θα το κάνει εφικτό. Στο παράδειγμα θα επιλεχθεί ο CART λόγω του ότι θελουμε να λύσουμε ένα πρόβλημα ταξινόμησης. Οπότε σε κάθε κόμβο υπολογίζεται ο αριθμός Gini, ο οποίος προέρχεται από τον υπολογισμό της πρόσμειξης (impurity). Ένας κόμβος είναι καθαρός(pure) όταν τρέχουν όλες οι εκπαιδεύσεις από τον συγκεκριμένο κόμβο και ανήκει σε μια κατηγορία (π.χ no leak). Αυτός ο αριθμός υπολογίζεται χρησιμοποιώντας την μαθηματική έννοια

$$G_i = 1 - \sum_{k=1}^{n} Pi, k^2$$

Όπου $P_{i,k}$ είναι ο λόγος των εμφανίσεων της κλάσης K μεταξύ των εκπαιδεύσεων στο i^{th} κόμβο [36]

Αν θέλαμε να υπολογίσουμε το αριθμό Gini στην περίπτωση No Leak στο παράδειγμα θα είχαμε

$$1 - \left(\frac{48}{52}\right)^2 - \left(\frac{4}{52}\right)^2 - \left(\frac{0}{52}\right)^2 - \left(\frac{0}{52}\right)^2 \approx 0.142.$$

Ο αλγόριθμος Entropy είναι μία ακόμη μέθοδος που είναι απαραίτητη για να τρέξει ένα δέντρο απόφασης. Η εντροπία είναι ένα μέτρο της διαταραχής ή της αβεβαιότητας. Στόχος των μοντέλων μηχανικής μάθησης γενικά είναι η μείωση της αβεβαιότητας[40]. Συνήθως στην μηχανική μάθηση χρησιμοποιείται για να μετρηθεί η πρόσμειξη (impurity) που σημαίνει ότι ένα σύνολο εντροπίας είναι 0 όταν περιέχει παραδείγματα εκπαίδευσης που αφορούν μια κλάση.

Η μαθηματική έννοια που την χαρακτηρίζει είναι

$$H_i = -\sum_{\substack{k=1 \\ P_i, k \neq 0}}^{n} Pi, k^2 \quad [36]$$

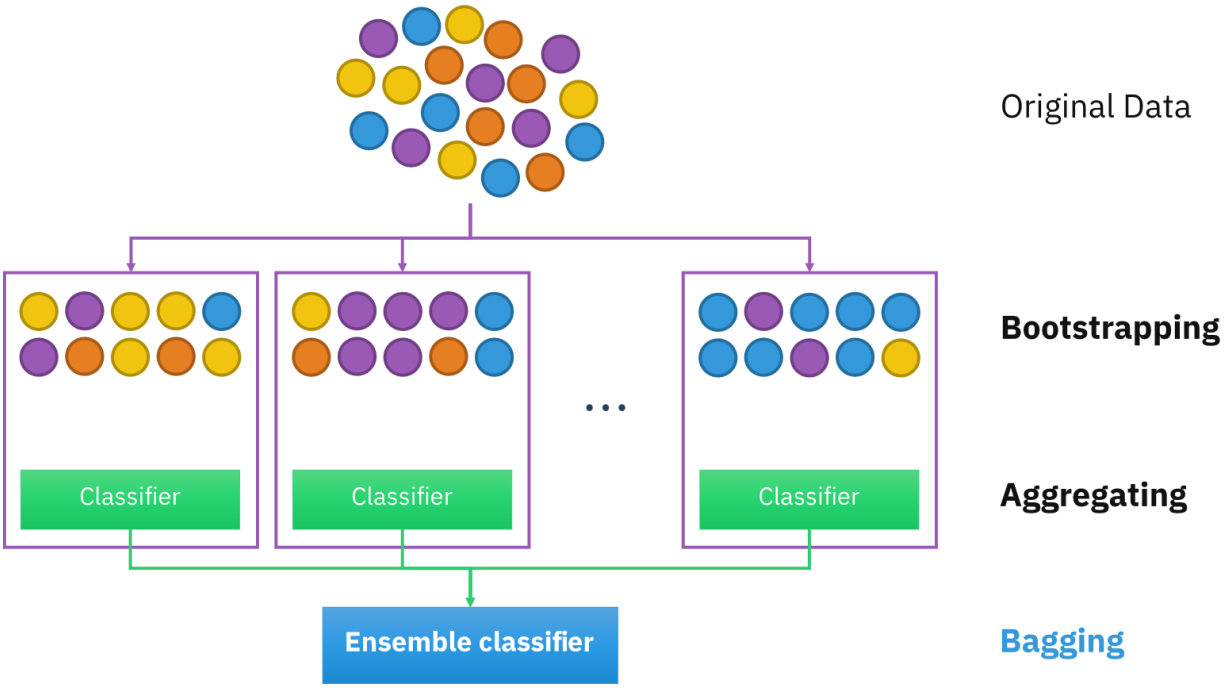
Και αν θέλαμε να υπολογίσουμε το αριθμό Entropy στην περίπτωση No Leak στο παράδειγμα θα είχαμε

$$-\left(\frac{48}{52}\right)\log_2\left(\frac{48}{52}\right) - \left(\frac{4}{52}\right)\log_2\left(\frac{4}{52}\right) \approx 0.391.$$

2.8 Τυχαία δάση(Random Forest)

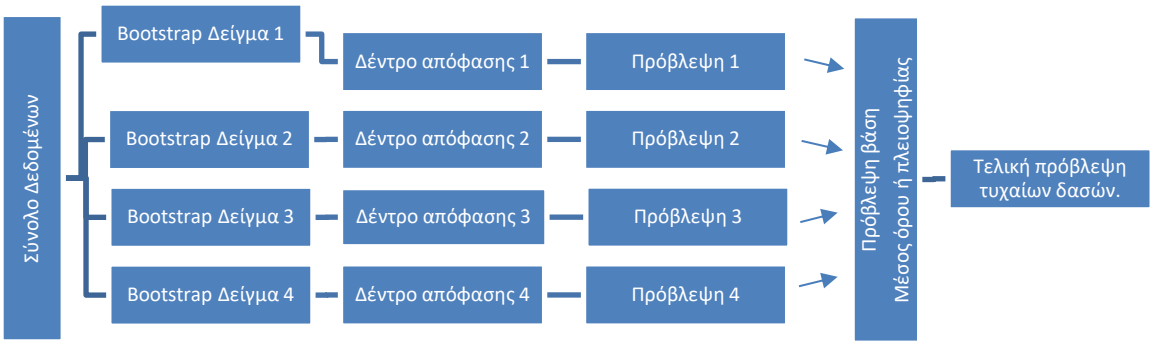
Τα τυχαία δάση είναι μια μέθοδος επιτηρούμενης μάθησης της μηχανικής μάθησης. Χρησιμοποιείται σε προβλήματα ταξινόμησης όπως και σε προβλήματα παλινδρόμησης. Είναι ένα από τους πιο ισχυρούς αλγόριθμους της μηχανικής μάθησης, και αυτό λόγω μιας τεχνικής που ονομάζεται **ensemble learning (Bagging)**. Είναι μία διαδικασία χρήσης πολλαπλών μοντέλων μηχανικής μάθησης με στόχο την υψηλότερη ακρίβεια πρόβλεψης συγκριτικά με ένα μονό μοντέλο μηχανικής μάθησης. Το **Bagging** γνωστό και ως Bootstrap Aggregation είναι η τεχνική που χρησιμοποιούν τα Τυχαία δάση. Επιλέγει ένα τυχαίο δείγμα από το σύνολο δεδομένων. Στην συνέχεια κάθε μοντέλο δημιουργείται από τα δείγματα (Bootstrap Samples) που παρέχονται από τα αρχικά δεδομένα με αντικατάσταση που είναι γνωστή ως row sampling. Αυτή η ενέργεια του row sampling με την αντικατάσταση ονομάζεται **bootstrap**. Το κάθε μοντέλο εκπαιδεύεται παράλληλα και εξάγει αποτελέσματα τα οποία συλλέγονται. Η πλειοψηφία μέσα από τους συνδυασμούς αυτών των αποτελεσμάτων δίνει το τελικό αποτέλεσμα. Η ενέργεια που περιλαμβάνει το συνδυασμό όλων τον αποτελεσμάτων και την δημιουργία της εξόδου βάση της πλειοψηφίας ονομάζεται aggregation[41].

Στη παρακάτω εικόνα παρατηρείται ένα πάράδειγμα του **ensemble learning (Bagging)**



Σχήμα 11. Απεικόνιση της δομής της τεχνικής Bagging[42]

Στην περίπτωση των τυχαίων δασών κατασκευάζονται ταυτόχρονα πολλά δέντρα απόφασης κατά την εκπαίδευση και στη συνέχεια υπολογίζεται ο μέσος όρος ή το αποτέλεσμα βάση της πλειοψηφίας από τη τελική πρόβλεψη των μεμονωμένων δέντρων απόφασης που κατασκευάστηκαν. Αυτή η τεχνική επιτυγχάνει υψηλή ακρίβεια πρόβλεψης αλλά και διορθώνει τα προβλήματα overfitting που υπάρχουν στα δέντρα απόφασης. Αυτό έχει ως αποτέλεσμα υψηλότερη απόδοση σε προβλήματα παλινδρόμησης και ταξινόμησης συγκριτικά με τα δέντρα απόφασης. Επίσης είναι ικανό να διαχειριστεί μεγάλα σύνολα δεδομένων με πολλές διαστάσεις. Στην παρακάτω εικόνα απεικονίζεται η δομή των τυχαίων δασών.



Σχήμα 12. Παράδειγμα της δομής των Τυχαίων δασών

Αναλύοντας το παραπάνω σχήμα παρατηρείται η διαδικασία των τυχαίων δασών από το αρχικό στάδιο όπου είναι το σύνολο δεδομένων μέχρι και το τελικό στάδιο, όπου υπολογίζεται η τελική πρόβλεψη. Αν υποθέσουμε την επίλυση ενός προβλήματος ταξινόμησης. Όπως παρατηρείται στο σχήμα 12 λαμβάνονται τυχαία 4 δείγματα Bootstrap 1-4 από το σύνολο δεδομένων. Το καθένα από αυτά θα εκπαιδευτεί με ένα συγκεκριμένο μοντέλο όπου στην προκειμένη περίπτωση είναι τα δέντρα απόφασης. Το καθένα από αυτά θα εκπαιδευτεί ξεχωριστά με τα δεδομένα που έχει από το δείγμα Bootstrap και θα καταλήξει σε μία πρόβλεψη. Αυτή η πρόβλεψη, για παράδειγμα από το δέντρο απόφασης 1, θα μπορούσε να αφορά σε μια μικρή διαρροή στη περίπτωση μας. Αν έχουμε στο δέντρο απόφασης 2 και 3 επίσης πρόβλεψη για μικρή διαρροή και στο δέντρο απόφασης 4 πρόβλεψη για καθόλου διαρροή, τότε θα ληφθεί σαν τελική πρόβλεψη αυτή που θα συγκεντρώσει το μέγιστο αριθμό ψήφων από τα επιμέρους δέντρα απόφασης. Άρα η τελική απόφαση θα είναι ότι υπάρχει μικρή μεγέθους διαρροή αφού έχουμε 3 ψήφους για μικρή διαρροή και 1 για καθόλου διαρροή.

Τα τυχαία δάση είναι ένας από τους πιο ευρέως διαδεδομένους αλγορίθμους που χρησιμοποιούν **ensemble learning** και είναι τόσο αποτελεσματικός λόγω της χρήσης του Bootstrap. Αυτό μειώνει τη διακύμανση (**Variance)** του τελικού μοντέλου. Χαμηλό Variance σημαίνει και χαμηλή υπερπροσαρμογή (**overfitting)** στη εκπαίδευση των δεδομένων. Overfitting έχουμε όταν το μοντέλο προσπαθεί να εξηγήσει μικρές παραλλαγές στο σύνολο δεδομένων. Κατά την δημιουργία του συνόλου δεδομένων π.χ. στην δειγματοληψία μπορεί να συμπεριληφθούν και κάποια ανεπιθύμητα σήματα, όπως ο θόρυβος ή και ανεπιθύμητες ακραίες τιμές, τις οποίες μπορεί το μοντέλο να της απομνημονεύσει και να προσαρμόζεται πολύ στενά στο σύνολο εκπαίδευσης, με συνέπεια να μη μπορεί να γενικεύσει καλά τα νέα δεδομένα και να αδυνατεί να ταξινομήσει αποτελεσματικά τα δεδομένα. Η υπερπροσαρμογή μπορεί να επηρεάσει το τελικό αποτέλεσμα. Με την επιλογή των τυχαίων δασών με αντικατάσταση (Row Sampling) μειώνεται η επίδραση τέτοιων ανεπιθύμητων σημάτων που θα είχαν επιρροή στο τελικό αποτέλεσμα[43].

3 ΚΕΦΑΛΑΙΟ 3: (Σχεδιασμός)

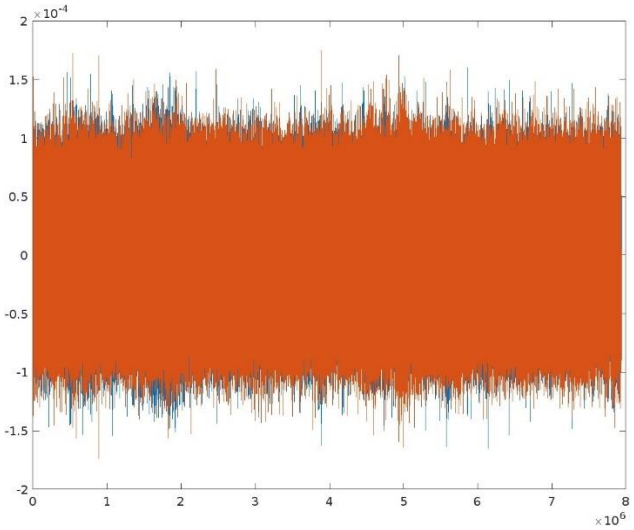
3.1 Ανάλυση σημάτων

Για τη πτυχιακή εργασία έχουν ληφθεί ως σύνολο δεδομένων τα σήματα που καταγράφηκαν από ακουστικούς αισθητήρες. Αυτά τα σήματα περιγράφουν αν συνέβη διαρροή στο σύστημα ροής ρευστών. Η πίεση του συστήματος είναι περίπου 5 bar. Για την καταγραφή έχουν χρησιμοποιηθεί δυο ακουστικοί αισθητήρες(επιταχυνσιόμετρα), οι οποίοι τοποθετήθηκαν ο ένας στα 1.8 μέτρα αριστερά από το σημείο αναφοράς (0) και ο δεύτερος στα 300 εκατοστά δεξιά από το σημείο αναφοράς, όπως φαίνεται και στο σχήμα 13. Όταν πραγματοποιείται μια διαρροή, αυτή βρίσκεται εντός αυτής της απόστασης.

Σχήμα 13. Απεικόνιση της θέσης των αισθητήρων στον αγωγό

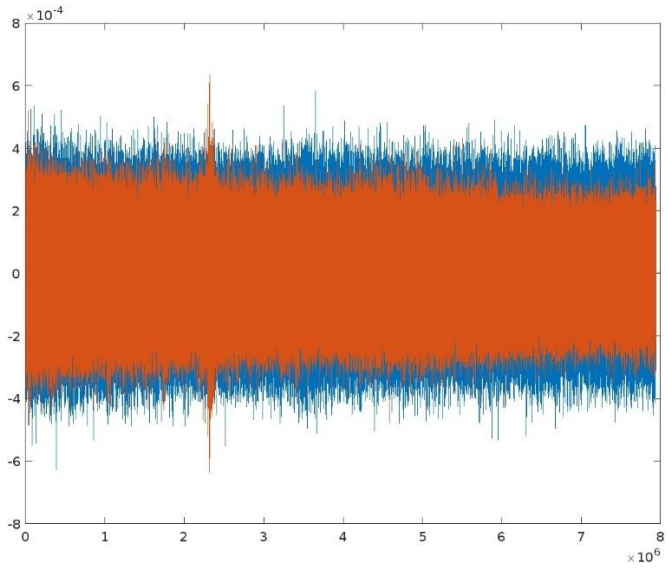
Έχει γίνει δειγματοληψία στα 44100Hz για 180 δευτερόλεπτα κάθε καταγραφή. Οι καταγραφές είναι τέσσερις. Μια χωρίς διαρροή (background noise) και τρεις με διαρροή, χαμηλή(small leak),μεσαία(medium leak) και υψηλή (large leak).Τα σήματα καταγραφής έχουν ληφθεί σε αρχεία MATLAB με επέκταση .mat. Για αυτό και έχει χρησιμοποιηθεί το MATLAB για την εμφάνιση και ανάλυση αυτών των σημάτων.

Όπως παρατηρείτε στο σχήμα 14, απεικονίζονται τα σήματα των δυο αισθητήρων όπου δεν υπάρχει διαρροή. Στο σχήμα η καταγραφή είναι 180 δευτερολέπτων μοιρασμένα περίπου κάθε 22.5 δευτερόλεπτα στον άξονα x. Το εύρος τιμών κυμαίνεται περίπου στο - 0.00014 και στο 0.00014.



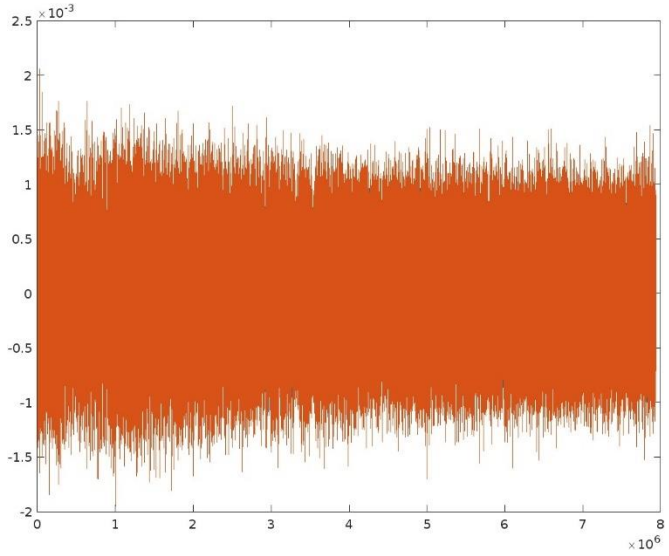
Σχήμα 14. Απεικόνιση σημάτων χωρίς την ύπαρξη διαρροής

Στο σχήμα 15, απεικονίζονται τα σήματα των δυο αισθητήρων όπου υπάρχει μικρή διαρροή. Όπως και στο σήμα χωρίς διαρροής, έτσι και στα σήματα διαρροών, η διάρκεια των απεικονίσεων είναι 180 δευτερόλεπτα. Εδώ το εύρος τιμών κυμαίνεται στο -0,0004 και στο -0,0004.



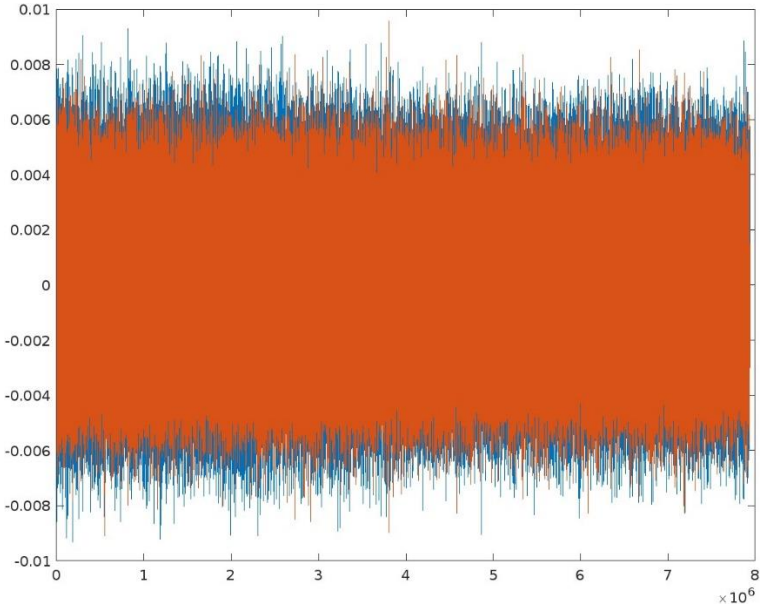
Σχήμα 15. Απεικόνιση σημάτων σε συμβάν μικρού μεγέθους διαρροής

Στο σχήμα 16, απεικονίζονται τα σήματα των δυο αισθητήρων όπου υπάρχει μεσαίου μεγέθους διαρροή. Το εύρος τιμών κυμαίνεται μεταξύ - 0.0012 και 0.0012.



Σχήμα 16. Απεικόνιση σημάτων σε συμβάν μεσαίου μεγέθους διαρροής

Στο γράφημα 17 απεικονίζονται τα σήματα των δυο αισθητήρων όπου υπάρχει μεγάλη διαρροή. Το εύρος τιμών κυμαίνεται μεταξύ– 0.006 και 0.006.



Σχήμα 17. Απεικόνιση σημάτων σε συμβάν μεγάλου μεγέθους διαρροής

Βλέποντας, αντιλαμβάνεται κάποιος εύκολα τις κλάσεις και μπορεί να τις ξεχωρίσει χωρίς να δυσκολευτεί. Αυτό θα πρέπει όμως να μπορεί να πραγματοποιηθεί και από την μηχανή. Η λέξη κλειδί είναι η Μηχανική μάθηση, όπου η μηχανή θα διαβάσει τα σήματα(σύνολο δεδομένων) που έχουν εισαχθεί ως είσοδο και με την βοήθεια ανθρώπου (επιβλεπόμενη μάθηση) θα χωριστούν σε κλάσεις. Έτσι η μηχανή θα μπορέσει να εκπαιδευτεί και να είναι σε θέση να προβλέψει σε ποια από τις παραπάνω κλάσεις αντιστοιχούν τα σήματα. Π.χ. καμία διαρροή κλάση 0 (τιμές από - 0.00014 έως και 0.00014),μικρή διαρροή κλάση 1(τιμές από -0,0004 έως και -0,0004), μεσαία διαρροή κλάση 2 (τιμές από - 0.0012 έως και 0.0012) και μεγάλη διαρροή κλάση 3(– 0.006 έως και 0.006).

3.2 Σύνολο Δεδομένων (dataset)

Για το πρακτικό κομμάτι της πτυχιακής και τη χρήση των μοντέλων χρησιμοποιήθηκε η γλώσσα προγραμματισμού python στη πλατφόρμα Colab της google. Η επιλογή αυτή βασίστηκε στο γεγονός ότι υπάρχει πλειάδα βιβλιοθηκών που έχουν βελτιστοποιηθεί ως προς τη χωρική και χρονική πολυπλοκότητα των αλγορίθμων που εκτελούν. Για τη δημιουργία του dataset χρησιμοποιήθηκε η βιβλιοθήκη Pandas.

Αφού εισαχθούν όλες οι κλάσεις από τον χρήστη με τις παραπάνω τιμές στα επιμέρους σύνολα δεδομένων κάθε διαρροής, συγχωνεύονται όλα και δημιουργείται ένα σύνολο δεδομένων για να μπορέσουν να αναλυθούν και στη συνέχεια να τροποποιηθούν και να χωριστούν σε δεδομένα εκπαίδευσης και δοκιμής. Στο πίνακα 2, φαίνεται ένα παράδειγμα του συνόλου δεδομένων. Η πρώτη στήλη περιέχει τις τιμές του αριστερού αισθητήρα, η δεύτερη στήλη τις τιμές του δεξιού αισθητήρα και η τρίτη στήλη αντιστοιχεί στην διαρροή.

	1	2	3
0	0.000036	-0.000019	0
1	0.000002	-0.000041	0
2	0.000004	-0.000062	0
3	-0.000006	-0.000036	0
4	-0.000007	-0.000030	0
...	...	...	...
31751991	-0.001585	0.000115	3
31751992	-0.000746	0.000035	3
31751993	-0.000002	0.000244	3
31751994	0.000069	0.000832	3
31751995	-0.000790	0.001504	3

Πίνακας 2. Αναπαράσταση του συνόλου δεδομένων στην Python μετά την είσοδο και την αντιστοίχιση τους σε κλάσεις

Η τροποποίηση που θα υποστούν είναι η κανονικοποίηση (Normalization) ώστε να είναι πιο υψηλή η ακρίβεια πρόβλεψης από το σύστημα. Με τη κανονικοποίηση τα δεδομένα τροποποιούνται στις τιμές τους, έτσι ώστε να χρησιμοποιούν μια κοινή γραμμή. Αυτό βοηθάει στον καλύτερο διαχωρισμό των κλάσεων όπως επίσης γίνεται και πιο κατανοητό προς το χρήστη.

Η κανονικοποίηση που χρησιμοποιήθηκε στο σύνολο δεδομένων είναι StandardScaler χρησιμοποιώντας το κώδικα στο παρακάτω σχήμα 18. Σκοπός είναι να μετασχηματιστούν τα δεδομένα έτσι ώστε κατά την κατανομή τους να έχουν μέση τιμή 0 και απόκλιση (standard deviation) 1.

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler().fit(x_train)
x_train = scaler.transform(x_train)
x_test = scaler.transform(x_test)
```

Σχήμα 18. Αναπαράσταση στη Python η κανονικοποίηση StandardScaler

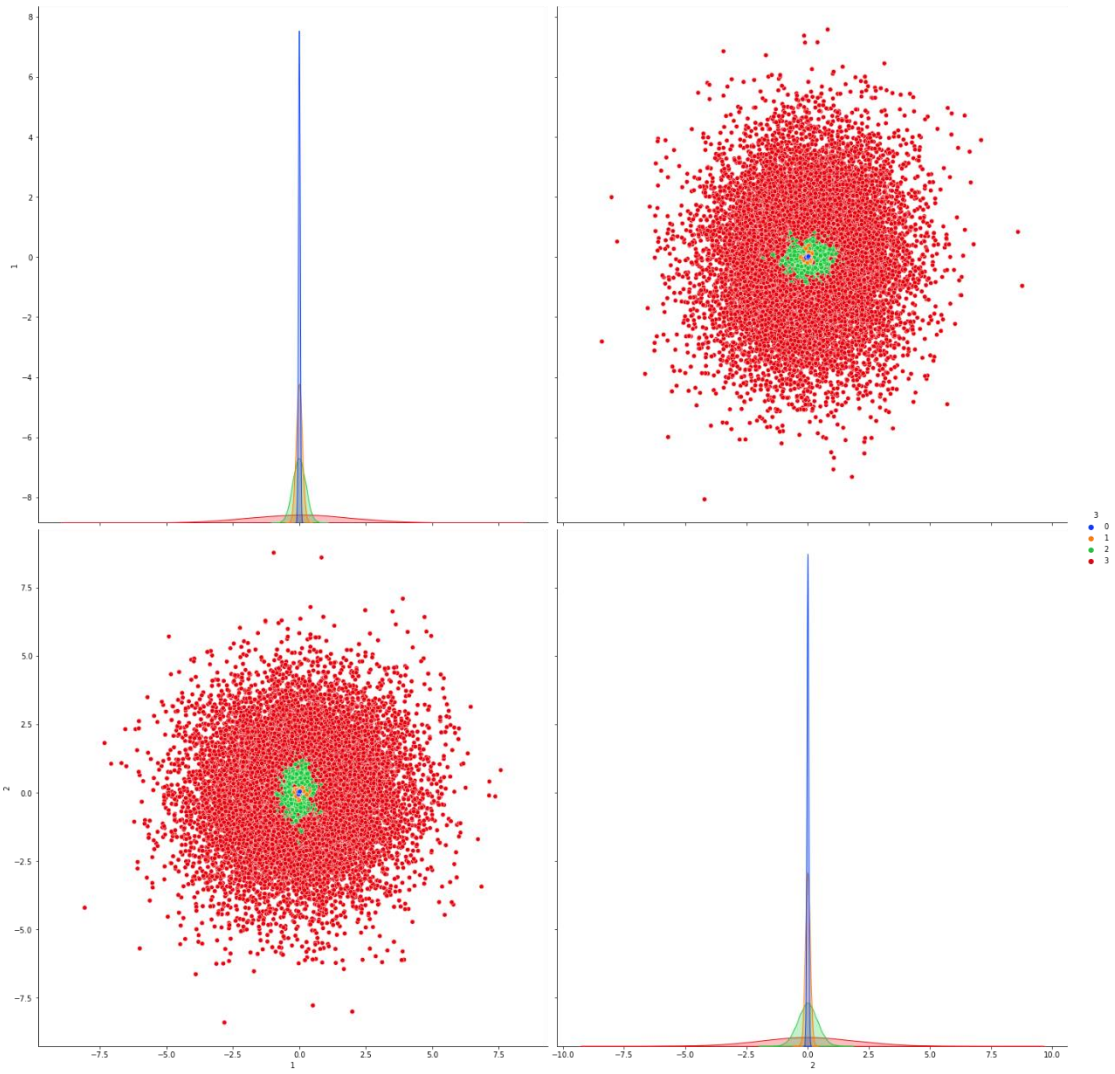
Παρακάτω στο σχήμα 19 φαίνεται ένα κομμάτι δεδόμένων που έχει πραγματοποιηθεί η κανονικοποίηση των δεδομένων. Στα αριστερά είναι τα δεδομένα πριν την ομαλοποίηση και δεξιά τα δεδομένα μετα τη κανονικοποίηση.

1	2
0.000255	-0.000167
0.000292	0.000145
-0.000046	0.000013
-0.000037	0.000010
0.000028	-0.000060
...	...
-0.000303	-0.001499
0.000059	-0.000481
-0.000012	0.000025
0.000028	0.000032
0.000014	-0.001908

1	2
0.268371	-0.191068
0.307707	0.163996
-0.059698	0.013935
-0.050092	0.010272
0.020273	-0.069777
...	...
-0.339859	-1.707076
0.053897	-0.548306
-0.022439	0.027096
0.020662	0.035372
0.005603	-2.172232

Σχήμα 19. Τιμές αισθητήρων πριν και μετά την κανονικοποίηση.

Στη συνέχεια εκτελέστηκε η seaborn(pairplot). Είναι μια βιβλιοθήκη με δυνατότητα επιλογής του τρόπου απεικόνισης των δεδομένων. Γίνεται χρήση αυτής για να επαληθευτεί αν λειτούργησε η κανονικοποίηση στο σύνολο των δεδομένων, όπως επίσης αν απεικονίζονται σωστά και διαχωρίζονται οι κλάσεις. Στο σχήμα 20, απεικονίζεται η εκτέλεση της seaborn. Έχουν εμφανιστεί 4 διαγράμματα, 2 για κάθε αισθητήρα. Στο διάγραμμα διασποράς(scatterplot) και των 2 αισθητήρων παρατηρείται ότι τα δεδομένα που έχουν χρησιμοποιηθεί είναι μη-γραμμικά κατανεμημένα. Οι κλάσεις είναι εύκολα διαχωρίσιμες, όπως φαίνεται, με τη κλάση 0 μπλε χρώμα να βρίσκεται στο κέντρο 0 του άξονα x και y , η κλάση 1 πορτοκαλί χρώμα να βρίσκεται μεταξύ -0.5 και 0.5, η κλάση 2 πράσινη βρίσκεται μεταξύ -0.1 και 0.1 και 3 η κόκκινη κλάση μεταξύ -8 και 8. Στα διαγράμματα πυκνότητας (densityplot) παρατηρούνται οι τιμές όλων των κλάσεων που έχουν σημείο αναφοράς το 0 στον άξονα x. Στη κλάση 3 γίνεται εύκολα αντιληπτό το ανώτατο σημείο (peak) που είναι κοντά στο -9 του άξονα y. Αντίστοιχα και στις άλλες κλάσεις φαίνονται τα ανώτατα σημεία. Κλάση 2 περίπου στο -6, κλάση 1 περίπου στο -4 και κλάση 0 περίπου στο 8.



Σχήμα 20. Αναπαράσταση διαγράμματος διασποράς(scatterplot)

Κατανοώντας το παραπάνω σχήμα 20 γίνεται αντιληπτό ότι αφού τα δεδομένα είναι μη γραμμικά κατανεμημένα, τα μοντέλα μηχανικής μάθησης που στηρίζονται σε μη γραμμικά δεδομένα θα είναι αυτά που θα πετύχουν υψηλότερη ακρίβεια πρόβλεψης. Στη πτυχιακή εργασία θα δοκιμαστούν όλα τα μοντέλα της επιβλεπόμενης μάθησης. Ακόμα θα δοκιμαστούν και μέθοδοι μοντέλων που στηρίζονται σε προβλέψεις δεδομένων γραμμικής κατανομής δεδομένων. Εκεί περιμένουμε να δούμε χαμηλή ακρίβεια πρόβλεψης εξαιτίας της γραμμικότητας των δεδομένων.

3.3 Εκπαίδευση Δεδομένων

Για να εφαρμοστούν τα μοντέλα που παρουσιάστηκαν στην προηγούμενη ενότητα πρέπει να χωριστεί το σύνολο δεδομένων που δημιουργήθηκε σε δεδομένα εκπαίδευσης (train set) και σε δεδομένα δοκιμής(test data). Στην πτυχιακή εργασία επιλέχθηκε τυχαίος διαχωρισμός του συνόλου δεδομένων σε 85% δεδομένων εκπαίδευσης και 15% δεδομένων δοκιμής. Για τη δοκιμή των αλγορίθμων χρησιμοποιήθηκε μικρότερος αριθμός δεδομένων για εκπαίδευση λόγω του ότι 31.751.996 δεδομένα απαιτούν μεγάλη υπολογιστική ισχύ και μεγάλο χρόνο

εκπαίδευσης. Έτσι επιλέχθηκαν 158.760 δεδομένα, τα οποία είναι η επιλογή με βήμα κάθε 200 γραμμές από το αρχικό σύνολο δεδομένων. Αυτό θα μας βοηθήσει να πάρουμε μία εικόνα για το ποιος αλγόριθμος θα παρουσιάσει τη καλύτερη απόδοση και στη συνέχεια να βρεθούν οι υπερ-παράμετροι που θα οδηγήσουν στη βελτίωση της ακρίβειας πρόβλεψης. Για την εύρεση αυτών θα χρησιμοποιηθεί ο GridSearchCV, όπου είναι μια τεχνική για την εύρεση των καταλληλότερων παραμέτρων. Θα δούμε παρακάτω ποιοι παράμετροι επιλέχθηκαν και ποιοι έβγαλαν τη μεγαλύτερη απόδοση μετά την εκτέλεση του GridSearchCV.

3.4 Εύρεση αποδοτικότερου μοντέλου

Για να βρεθεί το πιο αποδοτικό μοντέλο, χρησιμοποιήθηκαν όλα τα μοντέλα που παρουσιάστηκαν στην προηγούμενη ενότητα. Όπως προαναφέρθηκε, χρησιμοποιείται μικρότερος αριθμός δεδομένων για να γίνει πιο γρήγορα η εκπαίδευση και αφού βρεθεί ο πιο αποδοτικός τότε θα αυξηθεί και ο αριθμός των δεδομένων του συνόλου δεδομένων.

3.4.1 Naive Bayes

Πρώτο μοντέλο που επιλέχθηκε είναι ο Naive Bayes. Σε αυτό το μοντέλο υπάρχουν πολλαπλοί μέθοδοι, οι οποίοι δοκιμάζονται όλοι για να συμπεράνουμε ποια μέθοδος είναι η αποδοτικότερη. Την υψηλότερη απόδοση την πετύχαμε με την μέθοδο Gaussian (πίνακας 3), όπου και αυτό είναι λογικό διότι τα δεδομένα μας είναι μη γραμμικά κατανεμημένα.

Μεθόδοι Naive Bayes	Ευστοχία
Gaussian	81%
Categorical	24%
Multinomial	29%
Bernoulli	24%
Complement	14%

Πίνακας 3. Αποτελέσματα διαφορετικών μεθόδων του μοντέλου Naive Bayes

Για την χρήση των μεθόδων Categorical, Multinomial, Bernoulli και Complement ήταν απαραίτητο να γίνει κανονικοποίηση δεδομένων σε MinMax. Αυτό σημαίνει ότι κάθε δεδομένο του συνόλου δεδομένων έχει μεταφραστεί έτσι ώστε να βρίσκεται στο εύρος τιμών μεταξύ μηδέν και ένα. Όπως παρατηρείται, οι μέθοδοι αυτοί πέτυχαν πολύ χαμηλό ποσοστό ακρίβειας πρόβλεψης. Δοκιμάστηκαν επίσης τεχνικές για εύρεση υπερ-παραμέτρων, παρόλα αυτά δεν προέκυψε βελτίωση στην απόδοση του μοντέλου.

Στο παρακάτω πίνακα 4 παρουσιάζεται η αναφορά ταξινόμησης (Classification Report) της μεθόδου Gaussian. Η κλάση 3 (μεγάλη διαρροή) γίνεται από τον αλγόριθμο ευκολότερα διαχωρίσιμη με ποσοστό 94% και η κλάση 1 (μικρή διαρροή) ως λιγότερο διαχωρίσιμη με ποσοστό 70%.

	precision	recall	f1-score	support
0	0.77	0.92	0.84	5786
1	0.70	0.69	0.70	5961
2	0.81	0.73	0.77	6008
3	0.97	0.91	0.94	6059
accuracy			0.81	23814
macro avg	0.82	0.81	0.81	23814
weighted avg	0.82	0.81	0.81	23814

Πίνακας 4. Αναφορά ταξινόμησης της μεθόδου Gaussian

3.4.2 Support Vector Machines

Στην συνέχεια ακολούθησε το μοντέλο Μηχανές Διανυσμάτων Υποστήριξης. Εδώ επιλέχθηκε ο γκαουσιανός πυρήνας (Rbf) μιας και τα δεδομένα μας είναι μη γραμμικά. Πάραυτα δοκιμάστηκαν διάφορες παραλλαγές του SVM classifier όπως ένας εναντίων ενός (one vs one) και ένας εναντίων όλων (one vs rest), ο κοινός SVC καθώς και ο nuSVC. Ο nuSVC χρησιμοποιεί την παράμετρο nu σε αντίθεση με την παράμετρο C που χρησιμοποιείται στον απλό SVC. Η παράμετρος nu είναι ένα ανώτερο όριο για το σφάλμα εκπαίδευσης και οι τιμές που μπορεί να πάρει είναι από 0 έως και 1. Από την άλλη, η παράμετρος C στον απλό SVC είναι η παράμετρος ποινής του όρου σφάλματος. Ελέγχει το συμβιβασμό μεταξύ ομαλού ορίου απόφασης και ορθής ταξινόμησης των σημείων εκπαίδευσης. Τέλος δοκιμάστηκαν, ο αλγόριθμος με παράμετρο Poly όπως επίσης και ο LinearSVC.

	Μέθοδοι SVM	Ευστοχία
one vs rest	**SVC rbf:**	63%
	LinearSVC	33%
one vs one	**SVC rbf**	74%
	nu SVC rbf	80%
	nu SVC Poly	32%

Πίνακας 5. Αποτελέσματα διαφορετικών μεθόδων του μοντέλου SVM

Όπως παρατηρείται στα παραπάνω αποτελέσματα (πίνακας 5), την υψηλότερη ακρίβεια πρόβλεψης πέτυχε ο αλγόριθμος nuSVC rbf με ακρίβεια 80%. Κατά την επιλογή παραμέτρων επιλέχθηκε ο γκαουσιανός πυρήνας και η τιμή 0.41 για την παράμετρο nu. Στη περίπτωση του nu SVC Poly,ο αλγόριθμος είχε χαμηλότερη ακρίβεια πρόβλεψης με ποσοστό 32%. Κατά την χρήση αυτού του μοντέλου απαιτείται πολύ χρόνος κατά την εκπαίδευση σε σχέση με τα άλλα μοντέλα και αυτό δημιουργεί πρόβλημα όταν υπάρχουν μεγάλα σύνολα δεδομένων. Επίσης μεγάλο χρόνο κατά την εκπαίδευση του μοντέλου χρειάστηκε και κατά την χρήση της παραλλαγής του μοντέλου για τη μέθοδο ένας εναντίων όλων (one vs rest). Εδώ είχαμε ακρίβεια πρόβλεψης 63% , χαμηλότερη από ότι είχαμε με την μέθοδο ένας εναντίων ενός(one vs one) όπου η ακρίβεια πρόβλεψης ήταν 74%. Ακολούθησε η LinearSVC με ποσοστό 33% όπου είναι επίσης χαμηλό και λογικό αφού τα δεδομένα μας δεν είναι γραμμικά κατανεμημένα.

3.4.3 Γραμμική παλινδρόμηση

Επόμενη δοκιμή ήταν ο αλγόριθμος γραμμικής παλινδρόμησης. Γνωρίζοντας ότι τα δεδομένα μας δεν είναι γραμμικά κατανεμημένα δεν θα μπορούσαμε να βγάλουμε κάποιο αποτέλεσμα, καθώς η γραμμική παλινδρόμηση χρησιμοποιείται κατά κύριο λόγο για δεδομένα γραμμικά κατανεμημένα.

Παρόλα αυτά δοκιμάστηκε η γραμμική παλινδρόμηση για εγκυκλοπαιδικούς λόγους και είχε ως αποτέλεσμα MSE(συνάρτηση σφάλματος) = 1.24 για τα δεδομένα Test, το οποίο χρησιμοποιείται για τον υπολογισμό της απόδοσης του αλγόριθμου και είναι ο μέσος όρος του τετραγώνου της διαφοράς μεταξύ των παρατηρούμενων και των προβλεπόμενων τιμών μιας μεταβλητής.

Επίσης, εκτυπώνοντας τις προβλέψεις (σχήμα 21) που κατάφερε να υπολογίσει ήταν σε μορφή που δεν μπορούσαν να φανούν χρήσιμες για την ακριβή πρόβλεψη. Οι κλάσεις κατανέμονται διαφορετικά από ότι έχουν καθοριστεί εξαρχής. Δηλαδή αντί για 0, 1, 2 και 3 έχουμε ενδιάμεσους αριθμούς των κλάσεων όπως απεικονίζονται παρακάτω.

```
array([1.50239413, 1.50535733, 1.50243569, ..., 1.49536978, 1.50217169,
1.5009067 ])
```

Σχήμα 21. Αποτελέσματα προβλέψεων μετά την εκτέλεση της γραμμικής παλινδρόμησης

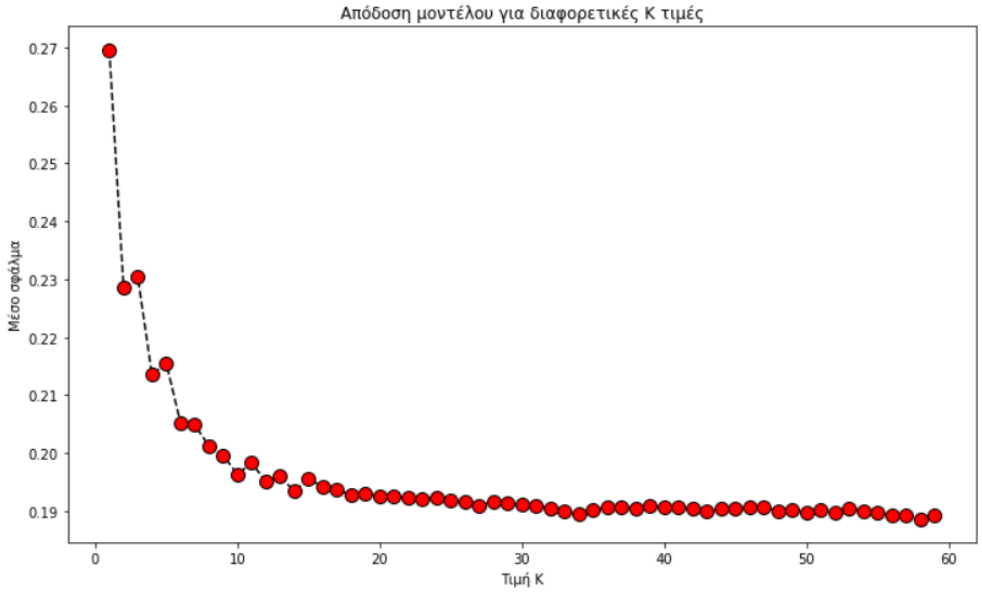
3.4.4 Λογιστική παλινδρόμηση

Μετά ακολούθησε η λογιστική παλινδρόμηση όπου και εδώ τα αποτελέσματα ήταν ανεπαρκή, καθώς η ακρίβεια πρόβλεψης ήταν μόλις 35%. Στη βιβλιοθήκη της sklearn, χρησιμοποιήθηκε η παράμετρος για χρήση της πολυωνυμικής λογιστικής παλινδρόμησης η οποία είναι μια μέθοδος για προβλήματα ταξινόμησης. Μία επίσης χρήσιμη παράμετρος που επιλέχθηκε κατά την εκπαίδευση για την ελαχιστοποίηση της συνάρτησης κόστους ήταν η παράμετρος **solver:** lbfgs **,** η οποία επιλέγεται για προβλήματα ταξινόμησης και είναι γρήγορη κατά την εκπαίδευση των συνόλων δεδομένων. Άλλοι παράμετροι για προβλήματα ταξινόμησης είναι οι newton-cg, sag και saga. Για μικρά σύνολα δεδομένων είναι καλή επιλογή η παράμετρος liblinear, καθώς και σε μεγάλα η sag και saga.

Logistic Regression: 35%

3.4.5 Κ-Κοντινότεροι γείτονες

Επόμενο μοντέλο που χρησιμοποιήθηκε είναι ο αλγόριθμος KNN. Σε αυτό τον αλγόριθμο η πιο χρήσιμη παράμετρος βελτιοποίησης της απόδοσης του μοντέλου είναι η παράμετρος **n_neighbors** η οποία καθορίζει πόσοι γείτονες θα χρησιμοποιηθούν κατά την εκπαίδευση. Εδώ με την χρήση της εντολής **for** για την εύρεση του καλύτερου K τιμής = n_neighbors καταφέραμε να προσδιορίσουμε τη τιμή με τη μεγαλύτερη ακρίβεια πρόβλεψης.



Σχήμα 22. Αναπαράσταση αποτελεσμάτων μέσου σφάλματος για 60 τιμές K

Όπως απεικονίζεται παραπάνω(σχήμα 22), η καλύτερη τιμή K που επιλέχθηκε είναι η K=59 όπου είχαμε Ελάχιστο σφάλμα = 0.189 και η ακρίβεια πρόβλεψης περίπου 81%

KNN: 81%

Επίσης δοκιμάστηκε και η παράμετρος **weights: uniform** και **distance,** αλλά η παράμετρος uniform είχε καλύτερα αποτελέσματα και πετύχαμε ακρίβεια προβλέψεις 81%. Με την παράμετρο uniform όλα τα σημεία σε κάθε γειτονιά έχουν ισοδύναμη κατανομή βάρους. Με

την παράμετρο distance οι κοντινότεροι γείτονες στο query point (σημείο ερωτήματος) θα έχουν μεγαλύτερη επιρροή από τους γείτονες που βρίσκονται αντίστοιχα μακριά από το σημείο.

3.4.6 Δέντρα απόφασης

Στην συνέχεια δοκιμάστηκε ο επόμενος αλγόριθμος, τα Δέντρα απόφασης, τα οποία παρέχουν μια εξαιρετικά αποτελεσματική δομή μέσω της οποίας μπορούμε να θέσουμε τις επιλογές και να διερευνήσουμε τα πιθανά αποτελέσματα μέσω των επιλογών που προεπιλέξαμε. Στην περίπτωση μας δοκιμάστηκαν ως παράμετροι **criterion**: **entropy** και **gini** καθώς και η παράμετρος **max depth.** Η παράμετρος criterion είναι η συνάρτηση για την μέτρηση της ποιότητας μιας διάσπασης. Η παράμετρος max depth ορίζει το μέγιστο βάθος του δέντρου. Το καλύτερο αποτέλεσμα ήταν 80.8% με τις επιλεγμένες παραμέτρους entropy και max depth= 9. Δοκιμάστηκαν επίσης και άλλοι παράμετροι όπως ο gini, καθώς και άλλες τιμές max depth. Παρόλα αυτά τα αποτελέσματα είχαν χαμηλότερη ακρίβεια.

Decision Trees: ~81%

3.4.7 Τυχαία δάση

Τελευταίος αλγόριθμος που δοκιμάστηκε ήταν τα τυχαία δάση. Όπως προαναφέρθηκε και στο προηγούμενο κεφάλαιο τα τυχαία δάση κατασκευάζουν ταυτόχρονα δέντρα απόφασης που τρέχουν παράλληλα αποφεύγοντας την υπερπροσαρμογή (overfitting) δεδομένων σε μεγάλα σύνολα δεδομένων. Η ακρίβεια πρόβλεψης ήταν στα 81% και επιτεύχθηκε με τη ρύθμιση υπερ-παραμέτρων του αλγόριθμου. Σημαντική παράμετρος στα τυχαία δάση είναι η παράμετρος **n_estimators,** η οποία προσδιορίζει πόσα δέντρα θα δημιουργηθούν κατά την εκτέλεση του αλγόριθμου. Άλλοι παράμετροι που επιλέχθηκαν για να φτάσουμε σε αυτό το αποτέλεσμα ήταν η **max_depth** και **criterion,** ίδιοι παράμετροι όπως και στα Decision trees, η **bootstrap** στην οποία προκαθορίζεται αν θα χρησιμοποιηθούν δείγματα bootstrap κατά την δημιουργία των δέντρων ,και η **min_samples_split** με την οποία καθορίζουμε τον ελάχιστο αριθμό δειγμάτων που απαιτείται για την διάσπαση ενός κόμβου.

Οι παράμετροι που επιλέχθηκαν ήταν:

- n_estimators= 250
- max_depth= 10
- criterion= gini
- min_samples_split= 5

```
rfc = RandomForestClassifier(n_estimators=250,min_samples_split=2,max_depth=10)
rfc.fit(x_train, y_train)
y_pred = rfc.predict(x_test)
dt1=metrics.accuracy_score(y_test,y_pred)
dt1
```

Σχήμα 23. Απεικόνιση εκτέλεσης του κώδικα Τυχαίων Δασών στη Python με τις επιλεγμένες παραμέτρους

Random Forest Classifier: 81%

Δοκιμάστηκαν και άλλες τιμές. Παρόλα αυτά, με την χρήση του αλγόριθμου grid καταλήξαμε στις τιμές οι οποίες αποδίδουν μεγαλύτερη ακρίβεια πρόβλεψης.

3.5 Αποτελέσματα

Αφού δοκιμάστηκαν όλοι οι αλγόριθμοι μηχανικής μάθησης που αναφέρθηκαν στις παραπάνω ενότητες, καταλήξαμε σε ικανοποιητικά αποτελέσματα χωρίς την εύρεση των καταλληλότερων υπερ-παραμέτρων μέσω εξειδικευμένων αλγόριθμων, όπως του Grid Search. Αυτό μας έδωσε μια πρώτη εικόνα για το ποσοστό ακρίβειας που μπορεί να πετύχει κάθε αλγόριθμος. Στο παρακάτω πίνακα βλέπουμε τα αποτελέσματα των αλγορίθμων που δοκιμάστηκαν στη διπλωματική εργασία.

Αλγόριθμος	Ευστοχία
NB Gaussian	81%
NB Categorical	24%
NB Multinomial	29%
NB Bernoulli	24%
NB Complement	14%
SVM SVC rbf	63%
SVM LinearSVC	33%
SVM SVC rbf	74%
SVM nu SVC rbf	80%
SVM nu SVC Poly	32%
Logistic Regression	35%
KNN	81%
Decision Trees	>81%
Random Forest Classifier	81%

Πίνακας 6. Αποτελέσματα ακρίβειας πρόβλεψης των αλγορίθμων

Με βάση τα πειράματα που διεξάχθηκαν στις παραπάνω ενότητες, οι αλγόριθμοι που αποδίδουν καλύτερα με τα υψηλότερα αποτελέσματα είναι ο Naive Bayes, με την υπόθεση ότι τα δεδομένα εκπαίδευσης προέρχονται από μια κατανομή Gauss, ο KNN, Decision Trees και Random Forest Classifier καθώς πέτυχαν ποσοστό ακρίβειας ~81%. Έχοντας μια εικόνα ποιοι από τους αλγόριθμους μπορούν να πετύχουν υψηλό ποσοστό ακρίβειας, προχωρήσαμε στο επόμενο βήμα αυτό της βελτιστοποίησης των αλγόριθμων αυτών με τη εύρεση των καταλληλότερων υπερ-παραμέτρων, καθώς θέλουμε να βελτιώσουμε την αποδοτικότητα τους. Για τον αλγόριθμο KNN είχαμε ήδη βρει τη καταλληλότερη και βασική υπερ-παράμετρο n_neighbors, η οποία ήταν για K=59 και πετύχαμε το 81%. Τον αλγόριθμο Naïve Bayes (Gaussian) τον απορρίψαμε διότι ακόμα και με αλλαγές των υπερ-παραμέτρων δεν καταφέραμε να ξεπεράσουμε το ποσοστό ακρίβειας του 81%.

Αφού επιλέχθηκαν οι 2 αλγόριθμοι με τα υψηλότερα ποσοστά ακρίβειας, το επόμενο βήμα είναι να δοκιμαστούν με ξεχωριστό Test Dataset καθώς θέλουμε να το προσομοιώσουμε με κανονικές συνθήκες και να δούμε την συμπεριφορά του σε δεδομένα που δεν έχουν συναντήσει στην εκπαίδευση τους.

Μετά τη προσθήκη του εξωτερικού Test Dataset εκτελέστηκαν οι 2 αλγόριθμοι Random Forest Classifier και Decision Trees

```
rfc = RandomForestClassifier(n_estimators=250,min_samples_split=2,max_depth=10)
rfc.fit(x_train, y_train)
y_pred = rfc.predict(x_test1)
dt1=metrics.accuracy_score(ytest1,y_pred)
dt1
```

Σχήμα 24. Απεικόνιση εκτέλεσης του κώδικα Τυχαίων Δασών στη Python με τις επιλεγμένες παραμέτρους αλλά διαφορετικό Test Dataset

Random Forest Classifier: 81.4%

```
tree1 = DecisionTreeClassifier(criterion= 'entropy',max_depth=9)
tree1.fit(x_train, y_train)
y_pred = tree1.predict(x_test1)
tree=metrics.accuracy_score(ytest1,y_pred)
tree
```

Σχήμα 25. Απεικόνιση εκτέλεσης του κώδικα Δέντρων αποφάσεων στη Python με διαφορετικό Test Dataset

Decision Trees: 81.9%

Παρατηρείται ότι τα ποσοστά ακρίβειας βελτιώθηκαν περίπου 1%, +0.4 για τα Random Forest και +0,9 για τα Decision Trees. Παρόλα αυτά εμείς θέλουμε να δοκιμάσουμε αν μπορούμε να αυξήσουμε το ποσοστό περαιτέρω βρίσκοντας τις καταλληλότερες υπερ-παραμέτρους αυτών των αλγορίθμων. Για αυτό το βήμα χρησιμοποιήθηκε ο Grid Search. Οι καταλληλότερες τιμές υπερ-παραμέτρων που βρέθηκαν μετά από μια διαδικασία ωρών μέσω του GridSearchCV απεικονίζονται στο παρακάτω σχήμα.

```
CV_rfc.best_params_

{'max_depth': 10,
 'max_features': 'auto',
 'min_samples_leaf': 2,
 'n_estimators': 79}
```

Σχήμα 26. Αποτελέσματα των καταλληλότερων τιμών μετά την εκτέλεση του GridSearchCV

Στην συνέχεια οι παράμετροι που βρέθηκαν από το Grid Search επιλέχθηκαν ως παράμετροι στον Random Forest Classifier. Το ίδιο συνέβη και στον αλγόριθμο Decision Trees όπου εκεί είχαμε ως καταλληλότερους παραμέτρους το Criterion=entropy και max_depth = 12.

```
rfc = RandomForestClassifier(n_estimators=79,bootstrap=True,max_depth=10,crite-
rion='gini',min_samples_leaf= 2, min_samples_split=4, max_features= "auto",
random_state = 42)
rfc.fit(x_train, y_train)
y_pred = rfc.predict(x_test1)
rfc=metrics.accuracy_score(ytest1,y_pred)
rfc
```

Σχήμα 27. Απεικόνιση εκτέλεσης του κώδικα Τυχαίων Δασών στη Python με τις καταλληλότερες παραμέτρους

Random Forest Classifier: 82.4%

Τα αποτελέσματα ήταν ικανοποιητικά φτάνοντας το 82.4% για τον αλγόριθμο Random Forest Classifier.

```
tree1 = DecisionTreeClassifier(criterion= 'entropy',max_depth=12)
tree1.fit(x_train, y_train)
y_pred = tree1.predict(x_test1)
tree=metrics.accuracy_score(ytest1,y_pred)
tree
```

Σχήμα 28. Απεικόνιση εκτέλεσης του κώδικα Δέντρων αποφάσεων στη Python με τις καταλληλότερες παραμέτρους

Decision Trees: 83.4%

Στην περίπτωση των Decision Trees είχαμε ικανοποιητικά αποτελέσματα φτάνοντας κοντά στο 83.4%, ξεπερνώντας το ποσοστό του Random Forest Classifier. Κάνοντας μια δοκιμή εξάγαμε το Confusion Matrix (σχήμα 28), με το οποίο μπορούμε να αξιολογήσουμε την ακρίβεια μιας ταξινόμησης.

```
array([[45,  4,  0,  0],
       [ 1, 36, 13,  0],
       [ 0,  6, 43,  1],
       [ 0,  2,  6, 42]])
```

Σχήμα 29. Εκτέλεση Confusion Matrix και προβολή αποτελεσμάτων

Επίσης εκτελέστηκε και μια αναφορά ταξινόμησης (Classification _report) για να παρατηρηθεί η ποιότητα των προβλέψεων από τον αλγόριθμο ταξινόμησης, δηλαδή πόσες προβλέψεις ήταν Αληθείς και πόσες Ψευδείς.

	precision	recall	f1-score	support
0	0.98	0.92	0.95	49
1	0.75	0.72	0.73	50
2	0.69	0.86	0.77	50
3	0.98	0.84	0.90	50
accuracy			0.83	199
macro avg	0.85	0.83	0.84	199
weighted avg	0.85	0.83	0.84	199

Πίνακας 7. Αναφορά ταξινόμησης των Decision Trees

Όπως φαίνεται στο Σχήμα 29 και στο πίνακα 7 αντίστοιχα, οι προβλέψεις που έκανε για τη κλάση 0 (καθόλου διαρροή) ο αλγόριθμος ήταν 45. Τέσσερις προβλέψεις τις ταξινόμησε σε λάθος κλάσεις. Στην αναφορά ταξινόμησης απεικονίζεται το ποσοστό ακρίβειας όπου ήταν στο 95% για την κλάση 0. Για την κλάση 1 (μικρή διαρροή) είχαμε αντίστοιχα 36 σωστές προβλέψεις και 14 λάθος ταξινομημένες με ποσοστό ακρίβειας 73%. Στην κλάση 2(μεσαία διαρροή) είχαμε 43 σωστές προβλέψεις και 7 λάθος με ποσοστό ακρίβειας 77% . Και τέλος στην τελευταία κλάση 3(μεγάλη διαρροή) είχαμε 42 σωστές προβλέψεις και 8 σε λάθος κλάσεις κατανεμημένες και το ποσοστό ακριβείας στα 90%.

4 ΣΥΜΠΕΡΑΣΜΑΤΑ

Τα αποτελέσματα αυτής της μεταπτυχιακής εργασίας ήταν ικανοποιητικά κάνοντας χρήση μηχανικής μάθησης με ποσοστό ακρίβειας να ανέρχεται κοντά στο 84%, πράγμα που υποδεικνύει ότι χρησιμοποιώντας αυτή τη μέθοδο μπορεί να βοηθήσει θετικά σε παρόμοια προβλήματα διαρροής ρευστών. Το μοντέλο στο οποίο καταλήξαμε είναι τα Δέντρα Αποφάσεων με ποσοστό ακρίβειας 83.4% και αποδεκτή ικανότητα γενίκευσης. Παρατηρήσαμε ότι, ένα μικρό ποσοστό δεδομένων εκπαίδευσης παρουσίασε κοινά στατιστικά χαρακτηριστικά, με αποτέλεσμα να υπάρχουν μερικές λάθος κατηγοριοποιήσεις. Δεδομένου αυτής της παρατήρησης, πιστεύουμε ότι ένα πληρέστερο και στατιστικά ”ισχυρότερο” σύνολο δεδομένων θα βοηθούσε να πετύχουμε καλύτερα αποτελέσματα. Στην προεπεξεργασία δεδομένων εκπαίδευσης χρησιμοποιήθηκε η μέθοδος της κανονικοποίησης δεδομένων (StandarScaler). Μία πιθανή μελλοντική επέκταση του μοντέλου θα μπορούσε να είναι η πρόβλεψη διαρροών προτού αυτές συμβούν με τη χρήση μοντέλων βαθιάς μάθησης και των επιμέρους χαρακτηριστικών(features), όπως αισθητήρες πίεσης, ροής και θερμοκρασίας.

Bibliography – References – Online sources

[1] K. Angelopoulos and G. O. Glentis, “Test and measurement assisted leak vibration signal analysis for leakages in metallic pipelines,” in *in 24th Pan-Hellenic Conference on Informatics (PCI 2020)*, Nov. 2020, pp. 214–218. doi: 10.1145/3437120.3437309.

[2] S. of A. Dept. of Environmental Conservation, “Technical Review of Leak Detection Technologies Volume I Crude Oil Transmission Pipelines,” Anchorage, AK, Aug. 1999. Accessed: Jan. 31, 2023. [Online]. Available: http://www.state.ak.us/dec

[3] Electricalvoice, “Master Terminal Units (MTU) in SCADA Systems,” Nov. 07, 2017. https://electricalvoice.com/master-terminal-units-mtu-in-scada-systems/#:~:text=Master%20terminal%20units%20%28MTU%29%20in%20SCADA%20system%20is,human%20interface%20and%20helps%20to%20take%20control%20decisions. (accessed Jan. 31, 2023).

[4] J. Zhang, H. P., and M. Twomey, “An Overview of Pipeline Leak Detection Technologies,” Oct. 2017.

[5] M. A. Adegboye, W. K. Fung, and A. Karnik, “Recent advances in pipeline monitoring and oil leakage detection technologies: Principles and approaches,” *Sensors (Basel)*, vol. 19, no. 11. MDPI AG, Jun. 04, 2019. doi: 10.3390/s19112548.

[6] J. Wan, Y. Yu, Y. Wu, R. Feng, and N. Yu, “Hierarchical leak detection and localization method in natural gas pipeline monitoring sensor networks,” *Sensors (Basel)*, vol. 12, no. 1, pp. 189–214, Jan. 2012, doi: 10.3390/s120100189.

[7] H. Lu, T. Iseley, S. Behbahani, and L. Fu, “Leakage detection techniques for oil and gas pipelines: State-of-the-art,” *Tunnelling and Underground Space Technology*, vol. 98. Elsevier Ltd, Apr. 01, 2020. doi: 10.1016/j.tust.2019.103249.

[8] A. b. M. Akib, N. b. Saad, and v. Asirvadam, “Pressure point analysis for early detection system. In Proceedings of the 2011 IEEE 7th International Colloquium on Signal Processing and its Applications,” 2011, pp. 103–107.

[9] J. Zhang and M. Twomey, “Introduction to Pipeline Leak Detection,” Oct. 2017.

[10] R. Cramer, D. Shaw, R. Tulalian, P. Angelo, and M.Stuijvenberg, “Detecting and Correcting Pipeline Leaks Before They Become a Big Problem,” *Mar Technol Soc J*, vol. 49, Apr. 2014, doi: 10.2118/167874-MS.

[11] G. Geiger, “State-of-the-art in leak detection and localization Leak Detection and Localisation in Pipelines View project,” Dec. 2006. [Online]. Available: https://www.researchgate.net/publication/290631637

[12] B. Qin, Z. Yunping, F. Min, and S. Xiaojian, “Leakage detection technology of oil and gas transmission pipelines and its development trend,” *Petrol. Eng. Construct*, vol. 33, pp. 19–23, 2007.

[13] D. De, J. Mashford, and S. Burn, “Computer Aided Leak Location and Sizing in Pipe Networks,” *Urban Water Security Res. Alliance Tech. Rep.*, 2010.

[14] M. Ferrante and B. Brunone, "Pipe system diagnosis and leak detection by unsteady-state tests. 2. Wavelet analysis," *Adv Water Resour*, vol. 26, no. 1, pp. 107–116, 2003, doi: https://doi.org/10.1016/S0309-1708(02)00102-1.

[15] W. Liang, L. Zhang, Q. Xu, and C. Yan, "Gas pipeline leakage detection based on acoustic technology," *Eng Fail Anal*, vol. 31, pp. 1–7, 2013, doi: https://doi.org/10.1016/j.engfailanal.2012.10.020.

[16] Dimension Engineering, "A beginner's guide to accelerometers," *http://www.dimensionengineering.com/info/accelerometers*, Dec. 07, 2015.

[17] M. M. Adegboye, W. k. Fung, and A. Karnik, *Recent Advances in Pipelines Monitoring and Oil Leakage Detection Technologies: Principles and Approaches*. 2019. doi: 10.20944/preprints201905.0041.v1.

[18] S. Bagavathiappan, B. B. Lahiri, T. Saravanan, J. Philip, and T. Jayakumar, "Infrared thermography for condition monitoring – A review," *Infrared Phys Technol*, vol. 60, pp. 35–55, 2013, doi: https://doi.org/10.1016/j.infrared.2013.03.006.

[19] M. Manekiya and A. Pachiyappan, "Leakage detection and estimation using IR thermography," in *In Proceedings of the 2016 International Conference on Communication and Signal Processing (ICCSP)*, Apr. 2016, pp. 1516–1519. doi: 10.1109/ICCSP.2016.7754411.

[20] Framatome ANP GmbH, "LEOS – Leak Detection and Location System." 1998. Accessed: Feb. 01, 2023. [Online]. Available: https://www.researchgate.net/publication/242384640_Leak_Detection_and_Locating_-_A_Survey

[21] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM J. Res. Dev.*, vol. 44, pp. 206–227, 1967.

[22] S. Brown, "Machine learning, explained," *https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained*, Apr. 21, 2021.

[23] T. Tiwari, T. Tiwari, and S. Tiwari, "How Artificial Intelligence, Machine Learning and Deep Learning are Radically Different?," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 8, no. 2, Feb. 2018, doi: 10.23956/ijarcsse.v8i2.569.

[24] K. Du and M. Swamy, "Reinforcement Learning," 2014, pp. 547–561. doi: 10.1007/978-1-4471-5571-3_18.

[25] M. Hossain, *Support Vector Machine**. 2022.

[26] JavaTpoint, "Support Vector Machine Algorithm." https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm (accessed Feb. 02, 2023).

[27] Baeldung, "Multiclass Classification Using Support Vector Machines," Nov. 11, 2022. https://www.baeldung.com/cs/svm-multiclass-classification (accessed Feb. 02, 2023).

[28] A. Wibawa *et al.*, "Naïve Bayes Classifier for Journal Quartile Classification," *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, vol. 7, no. 2, pp. 91–99, Jun. 2019, doi: 10.3991/ijes.v7i2.10659.

[29] R. Gandhi, "Naive Bayes Classifier," May 05, 2018. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c (accessed Feb. 02, 2023).

[30] N Kumar, "Naive Bayes Classifiers," Jan. 10, 2023. https://www.geeksforgeeks.org/naive-bayes-classifiers/ (accessed Feb. 02, 2023).

[31] K. Kumari and S. Yadav, "Linear regression analysis study," *Journal of the Practice of Cardiovascular Sciences*, vol. 4, p. 33, Jan. 2018, doi: 10.4103/jpcs.jpcs_8_18.

[32] JavaTpoint, "Linear Regression in Machine Learning." https://www.javatpoint.com/linear-regression-in-machine-learning (accessed Feb. 02, 2023).

[33] Humanunsupervised, "Regression. Cost Function. Hypothesis. Gradient", Accessed: Feb. 02, 2023. [Online]. Available: https://www.humanunsupervised.com/post/regression-univariate-cost-function-hypothesis-gradient-descent

[34] Datacamp, "Logistic Regression," Apr. 2018. https://www.datacamp.com/tutorial/logistic-regression-R (accessed Feb. 03, 2023).

[35] M. Gupta, "ML Cost function in Logistic Regression," May 06, 2019. https://www.geeksforgeeks.org/ml-cost-function-in-logistic-regression/#discuss (accessed Feb. 02, 2023).

[36] J. Dawani, *Hands-On Mathematics for Deep Learning: Build a Solid Mathematical Foundation for Training Efficient Deep Neural Networks*. Packt Publishing, 2020. [Online]. Available: https://books.google.de/books?id=Ujl1zQEACAAJ

[37] JavaTpoint, "K-Nearest Neighbor(KNN) Algorithm for Machine Learning." https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning (accessed Feb. 02, 2023).

[38] O. Harrison, "Machine Learning Basics with the K-Nearest Neighbors Algorithm," *Towards Data Science*, Sep. 10, 2018. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 (accessed Jan. 31, 2023).

[39] M. Sanjeevi, "Decision Trees Algorithms," Oct. 2017, Accessed: Feb. 02, 2023. [Online]. Available: https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1

[40] Towardsdatascience, "How Decision Trees Make Decisions," Jan. 11, 2019. https://towardsdatascience.com/entropy-how-decision-trees-make-decisions-2946b9c18c8 (accessed Feb. 02, 2023).

[41] AnalyticsVidhya, “Understanding Random Forest,” Jun. 17, 2021. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/ (accessed Feb. 02, 2023).

[42] Wikipedia, “Bootstrap. Aggregating. Ensemble Bagging”, Accessed: Feb. 02, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Bootstrap_aggregating#/media/File:Ensemble_Bagging.svg

[43] Medium, “Understanding Random Forests,” Mar. 03, 2019. https://medium.com/@harshdeepsingh_35448/understanding-random-forests-aa0ccecdbbbb (accessed Feb. 02, 2023).