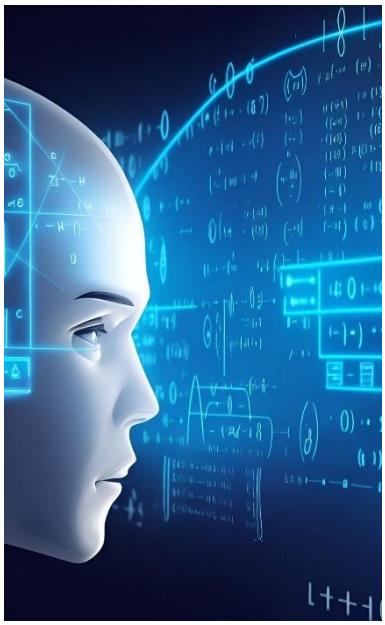
ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΦΥΣΙΚΗΣ

Διπλωματική Εργασία

Αριθμητικά μοντέλα, στατιστική ανάλυση και τεχνικές συμβολικής παλινδρόμησης στη μελέτη και την εξαγωγή ιδιοτήτων καινοτόμων υλικών

ΔΗΜΗΤΡΙΟΣ ΑΓΓΕΛΗΣ



Υπεβλήθη για την εκπλήρωση μέρους των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Φυσικής

ΛΑΜΙΑ 2023

Εγκρίθηκε από τα Μέλη της Τριμελούς Εξεταστικής Επιτροπής:

Πρώτος Εξεταστής (Επιβλέπων) — Φίλιππος Σοφός
Επίκουρος Καθηγητής, Τμήμα Φυσικής, Πανεπιστήμιο Θεσσαλίας

Δεύτερος Εξεταστής — Θεόδωρος Καρακασίδης
Καθηγητής, Τμήμα Φυσικής, Πανεπιστήμιο Θεσσαλίας

Τρίτος Εξεταστής — Άγγελος Αβραμόπουλος
Επίκουρος Καθηγητής, Τμήμα Φυσικής, Πανεπιστήμιο Θεσσαλίας

Ευχαριστίες

Πρώτα απ’ όλα, θέλω να ευχαριστήσω τον επιβλέποντα της διπλωματικής εργασίας μου, Επίκουρο Καθηγητή κ. Σοφό Φίλιππο, για τη στήριξη και τη συμπαράστασή του κατά τη διάρκεια της δουλειάς μου, όσο και για την εμπιστοσύνη που μου έδειξε θέτοντας από την αρχή υψηλούς στόχους και εξασφαλίζοντας την επίτευξή τους, μέσω της βοήθειας και καθοδήγησής του. Επίσης, ευχαριστώ και τα υπόλοιπα μέλη της εξεταστικής επιτροπής της διπλωματικής εργασίας μου, Καθηγητή κ. Καρακασίδη Θεόδωρο και Επίκουρο Καθηγητή κ. Αβραμόπουλο Άγγελο για την προσεκτική ανάγνωση της εργασίας μου, για τις πολύτιμες υποδείξεις τους καθώς και για την προθυμία να επιλύσουν οποιοδήποτε πρόβλημα πρόκυπτε, παρέχοντας πάντα σωστή καθοδήγηση. Οφείλω ευχαριστίες στον Επίκουρο Καθηγητή κ. Πετρόπουλο Νικόλαο, για τις πολύτιμες συμβουλές που μου έδωσε σε όλη τη διάρκεια της φοίτησής μου, καθώς και σε όλους του καθηγητές του τμήματος του μεταπτυχιακού «Εφαρμοσμένη Φυσική». Ευχαριστώ τις φίλες μου Τριανταφύλλου Χαρούλα και Φαροπούλου Ειρήνη για την ηθική υποστήριξή τους. Επίσης, ευχαριστώ την αδερφή μου Μαρία-Αθανασία για την κατανόησή της, ιδιαίτερα κατά τη διάρκεια των τελευταίων μηνών της προσπάθειάς μου. Πάνω απ’ όλα, είμαι ευγνώμων στους γονείς μου, Ιωάννη και Ελένη για την ολόψυχη αγάπη και υποστήριξή τους όλα αυτά τα χρόνια. Αφιερώνω αυτή την εργασία στην μητέρα μου, στον πατέρα μου και στην αδερφή μου.

Δημήτριος Αγγελής

Αριθμητικά μοντέλα, στατιστική ανάλυση και τεχνικές συμβολικής παλινδρόμησης στη μελέτη και την εξαγωγή ιδιοτήτων καινοτόμων υλικών

Δημήτριος Αγγελής

Πανεπιστήμιο Θεσσαλίας, Τμήμα Φυσικής, 2023

Επιβλέπων Καθηγητής: Σοφός Φίλιππος, Επίκουρος Καθηγητής

Περίληψη

Η υιοθέτηση της μηχανικής μάθησης στο πεδίο της επιστήμης των υλικών τα τελευταία χρόνια, έχει αποτελέσει ένα σημαντικό υπολογιστικό εργαλείο, καθώς έχει τη δυνατότητα να εξάγει ακριβή αποτελέσματα με ελάχιστο υπολογιστικό κόστος, έναντι των υπολογιστικών προσομοιώσεων αλλά και των χρονοβόρων και κοστοβόρων πειραμάτων. Παρά την αδιαμφισβήτητη κυριαρχία των τεχνικών μηχανικής μάθησης, αδύνατο σημείο αποτελεί το γεγονός ότι τα περισσότερα μοντέλα αδυνατούν να αποτυπώσουν τη φυσική ερμηνεία της συμπεριφοράς του συστήματος στις προβλέψεις τους. Η συμβολική παλινδρόμηση, από την άλλη, μια μέθοδος η οποία βασίζεται στις αρχές του γενετικού προγραμματισμού, είναι σε θέση να εξάγει αναλυτικές εξισώσεις οι οποίες αντλούν πληροφορίες από τα δεδομένα και είναι σε θέση να αποτυπώσει τη φυσική διεργασία του συστήματος, χωρίς να έχει γνώση για το υπό μελέτη σύστημα. Στόχος της παρούσας εργασίας, είναι η εξαγωγή εξισώσεων που περιγράφουν ιδιότητες μεταφοράς ρευστών όπως ο συντελεστής διάχυσης, το ιξώδες διάτμησης και η θερμική αγωγιμότητα και να συγκριθούν με τις ήδη υπάρχουσες εξισώσεις στη βιβλιογραφία που χρησιμοποιούν αρκετοί ερευνητές. Τα αποτελέσματα έδειξαν ότι υπάρχει η δυνατότητα να ληφθούν πιο απλές σε έκφραση εξισώσεις από αυτές που χρησιμοποιούνται, με συγκρίσιμη ακρίβεια, μόνο αξιοποιώντας δεδομένα, ενώ ταυτόχρονα αποτυπώνεται η φυσική συμπεριφορά των συντελεστών.

Λέξεις Κλειδιά: *Μηχανική Μάθηση, Συμβολική Παλινδρόμηση, Πρόβλεψη ιδιοτήτων ρευστών, συντελεστής διάχυσης, ιξώδες διάτμησης, θερμική αγωγιμότητα*

Numerical models, statistical analysis and symbolic regression techniques in the investigation and extraction of novel material properties

Dimitrios Angelis

University of Thessaly, Department of Physics, 2023

Supervisor: Sofos Filippos, Assistant Professor

Abstract

The embracing of machine learning by the field of materials science over the last few years, has proven to be a valuable computational tool, as it is equipped with the capacity to provide accurate results, at only a fraction of the computational cost that computer simulations, as well as costly and time-consuming experiments require. Notwithstanding the dominance of machine learning techniques, capturing the physical interpretation of the system's behavior constitutes a weak point of most models during their predictions. Symbolic regression on the other hand, a method based on the principles of genetic programming, does not consider *a priori* knowledge of the investigated system, acquire information by handling data and produces analytical expressions that provide a physics-inspired overlook. The aim of this study is equation extraction towards fluid properties prediction, such as diffusion coefficient, shear viscosity and thermal conductivity, and compare them with well-established equations found in the literature that researchers are most familiar with. Results show that it is possible to provide simpler forms than the currently used equations, with comparable accuracy, solely by utilizing data, while at the same time those equations are bound to the physical behavior of the system.

Keywords: *Machine Learning, Symbolic Regression, Fluids properties prediction, Diffusion coefficient, Shear viscosity, Thermal Conductivity*

Πίνακας Περιεχομένων

Κεφάλαιο 1 Εισαγωγή 1

Κεφάλαιο 2 Μηχανική Μάθηση 5

2.1 Κατηγορίες και χαρακτηριστικά 5

2.2 Προεπεξεργασία δεδομένων 8

2.3 Παλινδρόμηση 13

2.3.1 Γραμμική παλινδρόμηση 14

2.3.2 Λογιστική παλινδρόμηση 25

2.4 Νευρωνικά δίκτυα 30

2.5 Δέντρα απόφασης 35

2.6 Βαθιά Μάθηση 41

Κεφάλαιο 3 Συμβολική Παλινδρόμηση 43

3.1 Διατύπωση προβλήματος 44

3.2 Εξελικτική Υπολογιστική 46

3.3 Εξελικτικοί Αλγόριθμοι 47

3.4 Γενετικός Προγραμματισμός 49

3.5 Διαδικασία μέσω Γενετικού Προγραμματισμού 50

3.6 Χαρακτηριστικά 56

Κεφάλαιο 4 Μηχανική Μάθηση στην Επιστήμη των Υλικών 60

4.1 Μοντελοποίηση σε πολλαπλή κλίμακα 61

4.2 Πρόβλεψη ιδιοτήτων και σχεδιασμός υλικών 63

Κεφάλαιο 5 Εφαρμογή 65

5.1 Θεωρητικές σχέσεις ιδιοτήτων μεταφοράς 66

5.2 Υλοποίηση Συμβολικής Παλινδρόμησης **68**

5.2.1 Συντελεστής Διάχυσης 70

5.2.2 Ιξώδες διάτμησης 75

5.2.3 Θερμική αγωγιμότητα 81

Κεφάλαιο 6 Συμπεράσματα και προτάσεις για περαιτέρω έρευνα 86

Βιβλιογραφία 88

Παράρτημα Α 101

104

Παράρτημα Β 105

Κατάλογος Πινάκων

Πίνακας 1: Σετ δεδομένων για παράδειγμα βασικών στατιστικών μέτρων. 9
Πίνακας 2: Σετ δεδομένων για παράδειγμα γραμμικής παλινδρόμησης. 16
Πίνακας 3: Σετ δεδομένων για παράδειγμα γραμμικής παλινδρόμησης. 18
Πίνακας 4: Σετ δεδομένων για παράδειγμα γραμμικής παλινδρόμησης. 18
Πίνακας 5: Σετ δεδομένων για παράδειγμα λογιστικής παλινδρόμησης. 26
Πίνακας 6: Σετ δεδομένων για παράδειγμα δέντρου απόφασης 1. 37
Πίνακας 7: Σετ δεδομένων για παράδειγμα δέντρου απόφασης 2. 39
Πίνακας 8: Εκφράσεις συντελεστή διάχυσης $\boldsymbol{D}$ για όλες τις καταστάσεις του ρευστού. 73
Πίνακας 9: Αποτελέσματα στατιστικών μέτρων της κάθε εξίσωσης του Πίνακα 8 στα δεδομένα. . 73
Πίνακας 10: Εκφράσεις ιξώδους διάτμησης $\boldsymbol{ns}$ για όλες τις καταστάσεις του ρευστού. 79
Πίνακας 11: Αποτελέσματα στατιστικών μέτρων της κάθε εξίσωσης του Πίνακα 10 στα δεδομένα. 80
Πίνακας 12: Εκφράσεις θερμικής αγωγιμότητας $\boldsymbol{\lambda}$. 84
Πίνακας 13: Αποτελέσματα στατιστικών μέτρων της κάθε εξίσωσης του Πίνακα 12 στα δεδομένα. 85

Κατάλογος Σχημάτων

Σχήμα 1: Κατηγορίες Τεχνητής Νοημοσύνης (Artificial Intelligence). 3

Σχήμα 2: Κατηγορίες Μηχανικής Μάθησης: Επιβλεπόμενη Μάθηση (Supervised Learning), Μη επιβλεπόμενη Μάθηση (Unsupervised Learning), Ενισχυτική Μάθηση (Reinforcement Learning). 6

Σχήμα 3: Διάφορες δυνατότητες της κάθε κατηγορίας Μηχανικής Μάθησης. 7

Σχήμα 4: Παράδειγμα ταιριάσματος μοντέλων στα δεδομένα. a) Κανονική προσαρμογή του μοντέλου σε ένα τυχαίο σετ δεδομένων. b) Υπερπροσαρμογή του μοντέλου στα δεδομένα. c) Υποπροσαρμογή του μοντέλου στα δεδομένα. 9

Σχήμα 5: Μορφή Θηκογράμματος (box-plot). 12

Σχήμα 6: Παράδειγμα γραμμικής παλινδρόμησης. Ο άξονας των Χ αποτυπώνει την κάθε ώρα, όπου 0 είναι η αρχική μέτρηση στις 12:00 και 8 είναι η τελική μέτρηση στις 20:00. Ο άξονας των Υ αποτυπώνει τον αριθμό των χρημάτων που υπάρχουν στο ταμείο. 17

Σχήμα 7: Επίλυση απλής γραμμικής παλινδρόμησης μέσω Python. Η τιμή του Intercept αντιστοιχεί στο $\boldsymbol{a}$ και η τιμή του Hours στο $\boldsymbol{b}$. 22

Σχήμα 8: Παράδειγμα ταιριάσματος (fitting) της εξίσωσης απλής γραμμικής παλινδρόμησης $\boldsymbol{Y = 100 + 50 * X}$ στο σετ δεδομένων του Πίνακα 2. 23

Σχήμα 9: Επίλυση απλής γραμμικής παλινδρόμησης μέσω Python. Οι τιμές του Intercept και DifferenceInRating αντιστοιχούν στις τιμές $\boldsymbol{a}$ και $\boldsymbol{b}$. 27

Σχήμα 10: Σιγμοειδής καμπύλη. Το αποτέλεσμα της για όλα τα $\boldsymbol{x}$ είναι στο διάστημα [0,1]. Για μία τιμή $\boldsymbol{x > 0}$, η καμπύλη οριζοντοποιείται και δίνει πάντα 1, ενώ αντίστοιχα για μία τιμή $\boldsymbol{x < 0}$ η καμπύλη οριζοντοποιείται και δίνει πάντα 0. 29

Σχήμα 11: Γενική μορφή ενός νευρωνικού δικτύου (artificial neural network). 32

Σχήμα 12: Συνδυασμός της έως τώρα πληροφορίας σε έναν κόμβο για την τροποποίησή της σε άλλη. 33

Σχήμα 13: Παράδειγμα δέντρου απόφασης 1. 37

Σχήμα 14: Παράδειγμα δέντρου απόφασης 2. 38

Σχήμα 15: Η θέση της Συμβολικής Παλινδρόμησης στο υπερσύνολο της Εξελικτικής Υπολογιστικής. 44

Σχήμα 16: Γνωστά και άγνωστα χαρακτηριστικά προβλημάτων. 45

Σχήμα 17: Παραδείγματα συμβολικών εκφράσεων $\boldsymbol{S1}$ και $\boldsymbol{S2}$. 50

Σχήμα 18: Παραδείγματα συμβολικών εκφράσεων $\boldsymbol{S3}$ και $\boldsymbol{S4}$. 52

Σχήμα 19: Διαδικασία συμβολικής παλινδρόμησης μέσω γενετικού προγραμματισμού. 54
Σχήμα 20: Παράδειγμα Pareto front. 56
Σχήμα 21: Πολλαπλή κλίμακα της Επιστήμης των Υλικών και μέθοδοι που χρησιμοποιούνται. 62
Σχήμα 22: Παραδείγματα γραφημάτων για τις σχέσεις ανάμεσα στη μεταβλητή $\boldsymbol{F}$ και την $\boldsymbol{F}*$. (a) Γράφημα διασποράς (scatter plot). (b) Γραμμή ταυτότητας (identity line). (c) Γράφημα ταυτότητας (identity plot).....70
Σχήμα 23: Συμπεριφορά των δεδομένων συντελεστή διάχυσης $\boldsymbol{D}$. (a) Διάγραμμα φάσης $\boldsymbol{\rho} - \boldsymbol{T}$. Η κόκκινη περιοχή απεικονίζει την αέρια κατάσταση (gas), η πράσινη περιοχή την υγρή (liquid) ενώ όλη η άλλη περιοχή είναι η υπερκρίσιμη (super critical). (b) Διάγραμμα $\boldsymbol{\rho} - \boldsymbol{D}$. (c) Διάγραμμα $\boldsymbol{T} - \boldsymbol{D}$. (d) Τιμές συντελεστή συσχέτισης για τις μεταβλητές $\boldsymbol{\rho}, \boldsymbol{T}, \boldsymbol{D}$.....72
Σχήμα 24: Γράφημα ταυτότητας (identity plot) της Εξ. $\boldsymbol{F8}$ του Πίνακα 8.....75
Σχήμα 25: Συμπεριφορά των δεδομένων. (a) Διάγραμμα φάσης $\boldsymbol{\rho} - \boldsymbol{T}$. (b) Διάγραμμα $\boldsymbol{\rho} - \boldsymbol{ns}$. (c) Διάγραμμα $\boldsymbol{T} - \boldsymbol{ns}$. (d) Τιμές συντελεστή συσχέτισης για τις μεταβλητές $\boldsymbol{\rho}, \boldsymbol{T}, \boldsymbol{ns}$.....77
Σχήμα 26: Γράφημα ταυτότητας (identity plot) για την Εξ. $\boldsymbol{F1}$ (a) και $\boldsymbol{F2}$ (b) του Πίνακα 10.....78
Σχήμα 27: Γραφήματα ταυτότητας (identity plots) για τις εξισώσεις $\boldsymbol{F7}$ (Σχ. a) και $\boldsymbol{F8}$ (Σχ. b) του Πίνακα 10.....80
Σχήμα 28: Συμπεριφορά των δεδομένων. (a) Διάγραμμα φάσης $\boldsymbol{\rho} - \boldsymbol{T}$. (b) Διάγραμμα $\boldsymbol{\rho} - \boldsymbol{\lambda}$. (c) Διάγραμμα $\boldsymbol{T} - \boldsymbol{\lambda}$. (d) Τιμές συντελεστή συσχέτισης για τις μεταβλητές $\boldsymbol{\rho}, \boldsymbol{T}, \boldsymbol{\lambda}$. 82
Σχήμα 29: Γραφήματα ταυτότητας (identity plots) για τις εξισώσεις. $\boldsymbol{F4}$ και $\boldsymbol{F6}$ του Πίνακα 12....85
Σχήμα 30: : Γραφήματα διασποράς για τις εξισώσεις $\boldsymbol{Fi}\ \boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{9}$ του Πίνακα 8 (συντελεστής διάχυσης $\boldsymbol{D}$). Στις εξισώσεις όπου ο άξονας δεν ξεκινάει από το σημείο (0,0) αλλά από το (-10,-10) είναι ένδειξη ότι κάποια τιμή της πρόβλεψης έχει λάβει αρνητική τιμή.....102
Σχήμα 31: Γραφήματα διασποράς για τις εξισώσεις $\boldsymbol{Fi}\ \boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{9}$ του Πίνακα 10 (ιξώδες διάτμησης $\boldsymbol{ns}$).....103
Σχήμα 32: Γραφήματα διασποράς για τις εξισώσεις $\boldsymbol{Fi}\ \boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{9}$ του Πίνακα 12 (θερμική αγωγιμότητα $\boldsymbol{\lambda}$).....104

Κεφάλαιο 1 Εισαγωγή

Η τεχνολογική εξέλιξη έχει επιφέρει τεράστια αλλαγή στον τρόπο διεξαγωγής των επιστημονικών υπολογισμών και από μία καθαρά εμπειρική ή/και θεωρητική προσέγγιση, οι επιστήμονες κατάφεραν να προσεγγίζουν θέματα του ενδιαφέροντος τους με τη χρήση των υπολογιστών. Η συνεχής συλλογή δεδομένων από συσκευές, πειραματικές διαδικασίες, προσομοιώσεις, εξισώσεις κ.α., έφερε στο φως την ανάγκη για εφεύρεση μιας νέας τεχνικής για την αξιοποίηση αυτών των μεγάλων δεδομένων (big data). Η επιστήμη των δεδομένων (data science) εστιάζει στη μελέτη της συμπεριφοράς των δεδομένων με τη χρήση αριθμητικών μεθόδων, στατιστικών μεθόδων αλλά και την εύρεση μαθηματικών σχέσεων, οι οποίες είναι ικανές να περιγράψουν αυτή τη συμπεριφορά και αρκετές φορές να χρησιμοποιηθούν για την επαλήθευση των φυσικών νόμων, μεταξύ άλλων.

Ο όγκος αυτών των δεδομένων είναι τόσο μεγάλος, ώστε να είναι πρακτικά αδύνατο για έναν άνθρωπο να τα διαχειριστεί και να τα κατανοήσει με μη αυτόματο τρόπο. Αυτός είναι ένας σημαντικός λόγος για τον οποίο οι επιστήμονες και οι μηχανικοί βρήκαν καταφύγιο στην Τεχνητή Νοημοσύνη (Artificial Intelligence). Μάλιστα, η επίδραση της τεχνητής νοημοσύνης και των κλάδων της στη διαχείριση των δεδομένων και εξαγωγή πληροφοριών έχει αποδειχθεί τόσο ισχυρή, ώστε πλέον να αποτελεί αναπόσπαστο κομμάτι στην υπολογιστική διαδικασία στους περισσότερους κλάδους της μηχανικής και της υπολογιστικής επιστήμης (Sofos, Stavrogiannis, et al., 2022). Η τεχνητή νοημοσύνη είναι ένα υποσύνολο της πληροφορικής, στο οποίο οι τεχνικές που χρησιμοποιούνται επιτρέπουν στους ανθρώπους και τις μηχανές να βγάλουν εις πέρας εργασίες όπου μόνο οι οντότητες με νοημοσύνη μπορούν. Τέτοιες ενέργειες περιλαμβάνουν την αντίληψη (perceive), την κατανόηση (reason) και τη δράση (act) (Patrick, 1992). Επίσης, περιλαμβάνει διάφορους κλάδους όπου ο κάθε κλάδος ειδικεύεται σε ένα κομμάτι και προσεγγίζει τα διάφορα θέματα με βάση τους δικούς του αλγορίθμους (βλ. Σχ. 1). Ο κλάδος που έχει ξεχωρίσει και εφαρμόζεται συχνά για ανάλυση

δεδομένων, είναι η Μηχανική Μάθηση (Machine Learning) λόγω της ικανότητας που έχει να λαμβάνει υπόψη παλαιότερες προσεγγίσεις και μέσω αυτών να ενισχύει τις επόμενες προβλέψεις (Sofos, Stavrogiannis, et al., 2022).

Αρκετές φορές, η μηχανική μάθηση έχει παρερμηνευθεί ως κοινό πεδίο με τη στατιστική ανάλυση. Η διαφορά τους είναι πως στην πρώτη κατασκευάζονται αλγόριθμοι για να ταξινομήσουν ή να κάνουν πρόβλεψη για τα δεδομένα, ενώ στη δεύτερη πραγματοποιείται μόνο περιγραφή των δεδομένων. Οι αλγόριθμοι αυτοί, δέχονται ως είσοδο δεδομένα και προσπαθούν να βρουν μοτίβα ή σχέσεις μέσα σε αυτά προκειμένου να κάνουν κάποια πρόβλεψη. Πλέον, ολοένα και περισσότεροι ερευνητές επιλέγουν μία μέθοδο μηχανικής μάθησης για μελέτη σε διάφορους τομείς, όπως για παράδειγμα στην αυτόνομη οδήγηση (Shalev-Shwartz et al., 2016), στις ηλιακές κυψέλες (Zhong et al., 2022), χημεία (Janet et al., 2019) ή στη μελέτη της ηλιακής ακτινοβολίας (Quej et al., 2017), μεταξύ άλλων.

Παρά το γεγονός ότι η άνοδος των μεθόδων που βασίζονται σε δεδομένα έχει βοηθήσει έως ένα βαθμό τη μελέτη πολύπλοκων συστημάτων, όταν γίνονται προσπάθειες αποκρυπτογράφησης την φυσικών νόμων που διέπουν ένα σύστημα, υπάρχουν ορισμένα εμπόδια (Angelis et al., n.d.). Από τη μία μεριά, έχουν γίνει επιτυχώς ενέργειες ταύτισης των λύσεων μερικών διαφορικών εξισώσεων με φυσικούς νόμους, μέσω των Φυσικο-Ενημερωμένων Νευρωνικών δικτύων (Physics-Informed Neural Networks (PINNs)) (Cai et al., 2022) ή και των PI-GP (Raissi et al., 2017) , B-PINNs (L. Yang et al., 2021). Από την άλλη μεριά, το αποτέλεσμα της εισόδου σε ένα νευρωνικό δίκτυο (βλ. Κεφ. 2.4) είναι προκαθορισμένο, γεγονός που σημαίνει ότι η χαρτογράφηση της συμπεριφοράς των δεδομένων μπορεί να είναι ιδιαίτερα περίπλοκη (Tohme et al., 2022). Ενώ για τους υπολογιστές μία τέτοια διαδικασία φαίνεται απλή, για τους ανθρώπους ίσως είναι εξαιρετικά απαιτητικό να βγάλουν νόημα (Papastamatiou et al., 2022). Τέτοιου είδους μοντέλα χαρακτηρίζονται ως μοντέλα μαύρου-κουτιού (black-box models) (Asadzadeh et al., 2021) και η απουσία της φυσικής σημασίας σε αυτά τα μοντέλα, αποτέλεσε τον κινητήριο μοχλό για την απομάκρυνση και τη στροφή σε άλλα (Angelis et al., n.d.).

Στον αντίποδα, η Συμβολική Παλινδρόμηση (Symbolic Regression) είναι μία μέθοδος μηχανικής μάθησης η οποία έχει κατασκευαστεί με βάση την Εξελικτική Υπολογιστική (Evolutionary Computing). Διαφέρει από τα μοντέλα μαύρου κουτιού καθώς έχει τη

Σχήμα 1: Κατηγορίες Τεχνητής Νοημοσύνης (Artificial Intelligence).

δυνατότητα να παρουσιάζει τη φυσική σημασία της συμπεριφοράς των δεδομένων, καθώς δημιουργεί μία εξίσωση χωρίς να δέχεται κάποιου είδους περιορισμούς. Η ικανότητα αυτή είναι ιδιαίτερα χρήσιμη στις φυσικές επιστήμες, όπου η αποδοχή της συμβολικής παλινδρόμησης φαίνεται σε εφαρμογές της που έχουν σημειωθεί στη χημική μηχανική (Farizhandi et al., 2020; Neumann et al., 2020), στην εύρεση ιδιοτήτων υλικών (Bomarito et al., 2022; Karakasidis et al., 2022; M. Wang et al., 2022), στα υλικά κατασκευών (Chaabene & Nehdi, 2021; Ren et al., 2022; Rezaei et al., 2022) και στην αστρονομία ή αστροφυσική (Delgado et al., 2022; Manzi & Vasile, 2020; Matchev et al., 2022; Shao et al., 2022; Wadekar et al., 2022), μεταξύ άλλων.

Σκοπός της εργασίας είναι η εφαρμογή της συμβολικής παλινδρόμησης για την πρόβλεψη μακροσκοπικών ιδιοτήτων ενός ρευστού, μέσω σχέσεων από τη νανοκλίμακα. Πιο συγκεκριμένα, γίνεται εστίαση σε ιδιότητες μεταφοράς όπως ο συντελεστής διάχυσης (diffusion coefficient), το ιξώδες διάτμησης (shear viscosity) και η θερμική αγωγιμότητα (thermal conductivity) για το λόγο ότι ρυθμίζουν τους ρυθμούς μεταφοράς μάζας, ορμής και ενέργειας, και έχουν διαχρονικά προσελκύσει το ενδιαφέρον της ερευνητικής κοινότητας (Sofos et al., 2009). Οι προαναφερθείς ιδιότητες είναι δύσκολο να βρεθούν και προσεγγίζονται μόνο μέσω εμπειρικών σχέσεων ή πολύπλοκων αριθμητικών μεθόδων στην ατομική κλίμακα. Έχουν σημειωθεί αρκετές ενέργειες αποκρυπτογράφησης των παραπάνω ιδιοτήτων μέσω μεθόδων μηχανικής μάθησης, αλλά χωρίς μία φυσική ερμηνεία των αποτελεσμάτων, η διεργασία θεωρείται ατελής (Papastamatiou et al., 2022). Η εφαρμογή της συμβολικής

παλινδρόμησης έχει ως σκοπό να ξεδιαλύνει τις σχέσεις των παραπάνω ιδιοτήτων και να τις παρουσιάσει ως συνάρτηση μακροσκοπικών ιδιοτήτων όπως η πυκνότητα και η θερμοκρασία.

Στα κεφάλαια που ακολουθούν, θα αναλυθεί η μηχανική μάθηση και θα παρουσιαστεί ο τρόπος λειτουργίας των αλγόριθμων που χρησιμοποιούνται περισσότερο (Κεφ. 2), θα γίνει πλήρης διερεύνηση της συμβολικής παλινδρόμησης (Κεφ. 3), θα παρουσιαστεί η ένταξη των μεθόδων μηχανικής μάθησης, η συνεισφορά της συμβολικής παλινδρόμησης αλλά και ο τρόπος που πραγματοποιούνται οι μελέτες στην επιστήμη των υλικών (Κεφ. 4), προτού διεξαχθεί η εφαρμογή της συμβολικής παλινδρόμησης για την εύρεση ιδιοτήτων ενός υλικού, όπου προτείνονται πιο απλές εξισώσεις από αυτές που χρησιμοποιούνται έως τώρα (Κεφ. 5). Τέλος, η εργασία ολοκληρώνεται με συμπεράσματα αλλά και προτεινόμενες κατευθύνσεις για περαιτέρω έρευνα (Κεφ. 6).

Κεφάλαιο 2 Μηχανική Μάθηση

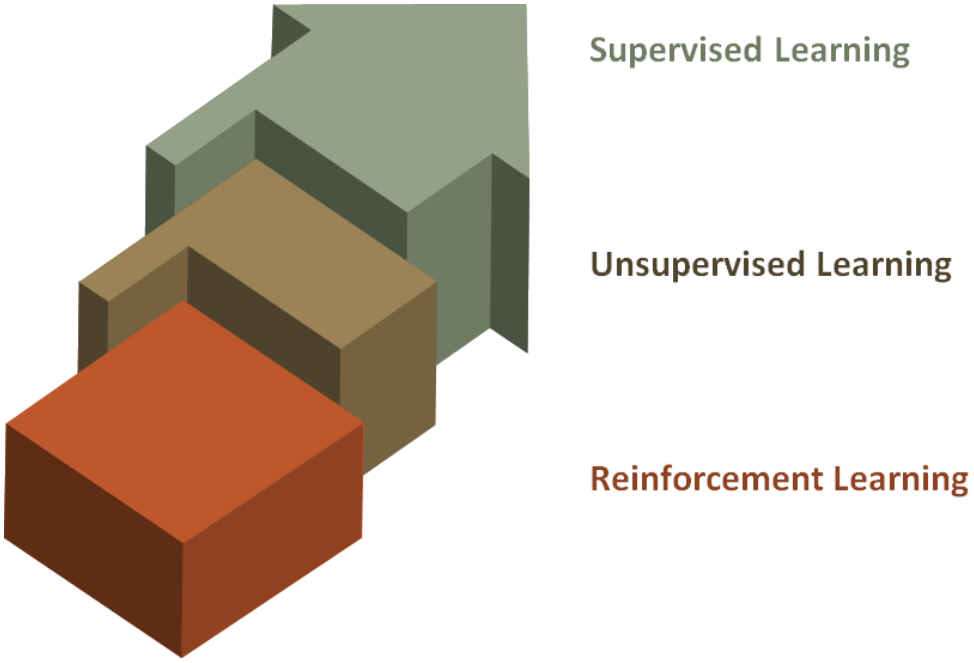
Για την εφαρμογή μιας μεθόδου μηχανικής μάθησης, είναι απαραίτητο να υπάρχουν δεδομένα. Τα δεδομένα εισάγονται σε έναν αλγόριθμο και, ανάλογα με το είδος του αλγόριθμου, εξάγονται διαφορετικά αποτελέσματα. Το πρώτο βήμα για την υλοποίηση μίας τέτοιας διαδικασίας ξεκινά πριν την είσοδο των δεδομένων στον αλγόριθμο, στο στάδιο της προεπεξεργασίας. Ανάλογα με το ζητούμενο, υπάρχουν κατηγορίες στη μηχανική μάθηση όπου εκτελούνται διαφορετικές λειτουργίες ή ειδικεύονται σε διαφορετικό σκοπό. Το κοινό χαρακτηριστικό των αλγορίθμων που χρησιμοποιούνται, είναι ότι ο εκάστοτε αλγόριθμος εκπαιδεύεται πάνω στα δεδομένα και προσπαθεί να μάθει τη συμπεριφορά τους, ώστε να επιχειρήσει περαιτέρω προβλέψεις ή να προχωρήσει σε ταξινόμησή τους.

Στο κεφάλαιο αυτό, θα παρουσιαστούν αναλυτικά οι δημοφιλείς κατηγορίες της μηχανικής μάθησης, ο τρόπος και ο λόγος που τα δεδομένα χρειάζονται προεπεξεργασία και τέλος θα παρουσιαστεί ο τρόπος λειτουργίας των αλγορίθμων που χρησιμοποιούνται περισσότερο. Υπάρχουν απλές μέθοδοι, όπως η γραμμική παλινδρόμηση, πιο σύνθετες οι οποίες βασίζονται σε Νευρωνικά Δίκτυα (Neural Networks) και τα Δέντρα Αποφάσεων (Decision Trees) και, άλλες, ακόμα πιο απαιτητικές υπολογιστικά και περίπλοκες στην υλοποίηση, όπως η Βαθιά Μάθηση (Deep Learning)..

2.1 Κατηγορίες και χαρακτηριστικά

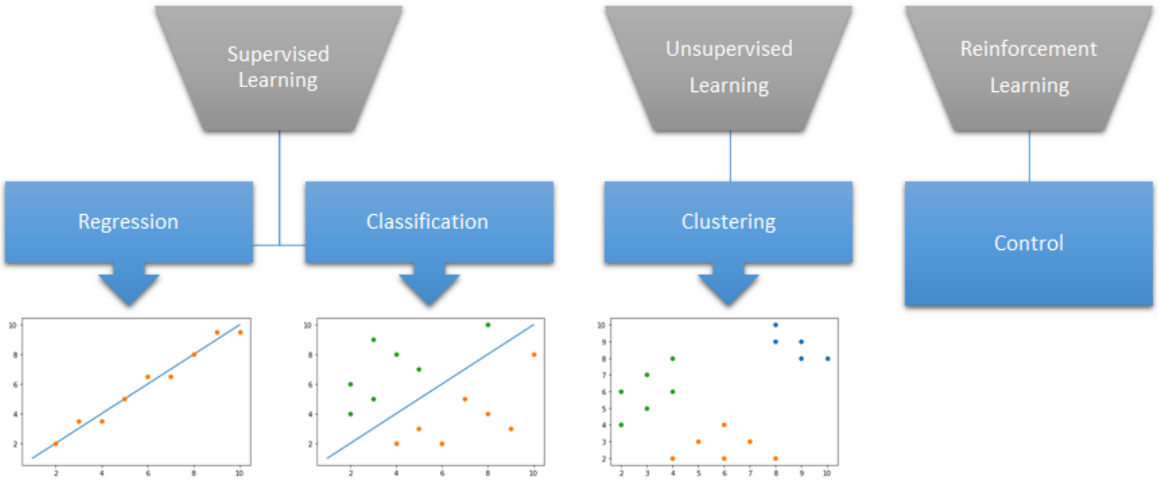
Στη μηχανική μάθηση αναγνωρίζονται τρείς κατηγορίες (Raschka, 2015) (βλ. Σχ. 2):

1. Επιβλεπόμενη Μάθηση (Supervised Learning).
2. Μη επιβλεπόμενη Μάθηση (Unsupervised Learning).
3. Ενισχυτική Μάθηση (Reinforcement Learning).



Σχήμα 2: Κατηγορίες Μηχανικής Μάθησης: Επιβλεπόμενη Μάθηση (Supervised Learning), Μη επιβλεπόμενη Μάθηση (Unsupervised Learning), Ενισχυτική Μάθηση (Reinforcement Learning).

Στην επιβλεπόμενη μάθηση είναι γνωστά τα δεδομένα που εισάγονται και εξάγονται από τον αλγόριθμο. Δηλαδή, ο αλγόριθμος γνωρίζει το αποτέλεσμα των δεδομένων εισόδου και κατά τη διαδικασία εκπαίδευσης προσπαθεί να βρει μία σχέση η οποία να συνδέει την τιμή στην είσοδο με την τιμή στην έξοδο. Αντίθετα, στη μη επιβλεπόμενη μάθηση δεν υπάρχουν "ταμπέλες" σε δεδομένα και οι αλγόριθμοι προσπαθούν να βρουν κρυμμένες διασυνδέσεις ανάμεσα σε αυτά (Chatzilygeroudis et al., 2021). Στην ενισχυτική μάθηση, η ανάγκη ύπαρξης ενός ικανού σετ δεδομένων ελαχιστοποιείται, καθώς το μοντέλο δημιουργεί από μόνο του δεδομένα και, επιπλέον, προσπαθεί από μόνο του να βελτιστοποιείται, πετυχαίνοντας πιο εύστοχες προβλέψεις (Alexiadis, 2019). Κάποιες από τις δυνατότητες της κάθε κατηγορίες φαίνονται στο Σχ. 3. Τέλος, ειδική αναφορά θα πρέπει να γίνει και σε μία υβριδική κατηγορία όπου παρατηρείται ένας συνδυασμός της επιβλεπόμενης μάθησης και της μη επιβλεπόμενης μάθησης, την ημι-επιβλεπόμενη μάθηση (Semi-Supervised Learning). Σε αυτή την κατηγορία, υπάρχουν δεδομένα με "ταμπέλες" εισόδου/εξόδου αλλά και χωρίς, ενώ σκοπός της διαδικασίας είναι η αναγνώριση του τρόπου που μπορεί να επηρεαστεί η διαδικασία εκπαίδευσης από αυτή τη μίξη δεδομένων και η κατασκευή κατάλληλων αλγόριθμων για εκμετάλλευση αυτού του συστήματος (F. Hu & Hao, 2012).



Σχήμα 3: Διάφορες δυνατότητες της κάθε κατηγορίας Μηχανικής Μάθησης.

Ο χαρακτηρισμός ενός μοντέλου μηχανικής μάθησης ως επιτυχημένου, συνδέεται άμεσα με την ποιότητα και την ποσότητα των διαθέσιμων δεδομένων (Angelis et al., n.d.). Η συγκέντρωση δεδομένων συνήθως γίνεται από βάσεις δεδομένων στο διαδίκτυο, ανατρέχοντας στη βιβλιογραφία, είτε δημιουργώντας ένα σετ δεδομένων από το μηδέν μέσω μετρήσεων σε πειράματα ή προσομοιώσεων (π.χ., Μοριακής Δυναμικής) (Guo et al., 2021). Επίσης, αναφορά πρέπει να γίνει και στο είδος των δεδομένων καθώς έχουν παρατηρηθεί δεδομένα σε διακριτή μορφή (κείμενα), συνεχή (τανυστές) ή και γραφήματα με βάρη (Wei et al., 2019). Αρκετές φορές, όμως, η πρόσβαση σε δεδομένα δεν εγγυάται την άμεση χρήση τους, καθώς μπορεί να βρίσκονται σε διαφορετική μορφή και να χρειαστεί η τροποποίησή τους σε μία κοινή μορφή (Sofos, Stavrogiannis, et al., 2022) μέσω προεπεξεργασίας.

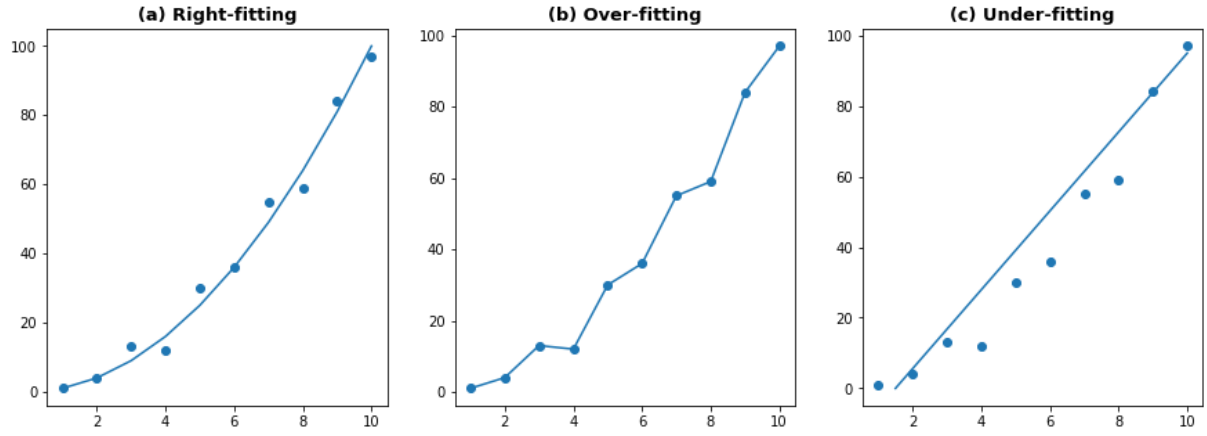
Μία αδυναμία των μοντέλων μηχανικής μάθησης, είναι η ευαισθησία στην υπερπροσαρμογή ή υπερμοντελοποίηση (overfitting). Η υπερπροσαρμογή δημιουργείται όταν το εκάστοτε μοντέλο έχει εκπαιδευτεί σε ένα σετ δεδομένων, τόσο ώστε να αδυνατεί να λειτουργήσει σε προβλέψεις εκτός αυτού του σετ. Πιο συγκεκριμένα, η προσαρμογή στα δεδομένα έχει πραγματοποιηθεί σε μεγαλύτερο βαθμό από αυτόν που θα έπρεπε δημιουργώντας έτσι περίπλοκες λύσεις, που παρουσιάζουν, όμως, καλό ταίριασμα στα δεδομένα. Από τη μία μεριά το σφάλμα που παρατηρείται είναι ελάχιστο, από την άλλη όμως το αποτέλεσμα του αλγόριθμου δεν μπορεί να γενικευθεί, γεγονός που καθιστά αδύνατη την εφαρμογή του σε διαφορετικό σετ δεδομένων καθώς το σφάλμα που θα παρατηρηθεί θα είναι μεγάλο. Σε τέτοιες

περιπτώσεις, ο αλγόριθμος κρίνεται αναξιόπιστος και δεν υπολογίζεται για περαιτέρω χρήση. Επίσης, παγίδες κατά τη χρήση ενός μοντέλου μηχανικής μάθησης κρύβονται στην αποφυγή της υπερπροσαρμογής. Για να αποφύγουν οι ερευνητές το προαναφερθέν φαινόμενο επιλέγουν μία μικρή περιοχή των δεδομένων ως δεδομένα εκπαίδευσης του μοντέλου, αλλά προβλήματα προκύπτουν από το γεγονός ότι το μοντέλο δεν μπορεί να επιτύχει ικανοποιητική ακρίβεια, με αποτέλεσμα να υποπίπτει σε υποπροσαρμογή (underfitting). Έτσι, ιδιαίτερη προσοχή χρειάζεται στο χτίσιμο ή την επιλογή του μοντέλου, προκειμένου να επιτύχει κανείς ένα κατάλληλο ταίριασμα στα δεδομένα (right fitting). Η αποφυγή της υποπροσαρμογής βρίσκεται στην αποφυγή της επιλογής ενός μοντέλου που δεν παρουσιάζει καλό ταίριασμα με τα δεδομένα. Ενδεικτικές αλλά αποτελεσματικές μέθοδοι αποφυγής υπερπροσαρμογής αποτελούν οι hold-out, k-fold cross-validation και regularization (Meng et al., 2020). Μία απεικόνιση των φαινομένων υπερπροσαρμογής, υποπροσαρμογής και σωστού ταιριάσματος του μοντέλου στα δεδομένα φαίνεται στο Σχ. 4. Πιο συγκεκριμένα, η προσέγγιση του μοντέλου του Σχ. 4a φαίνεται πιο πιθανή να περιγράφει ένα σετ, έναντι της απλής προσέγγισης που παρατηρείται στο Σχ. 4c και της υπερπροσέγγισης του Σχ. 4b.

Για να γίνει μία πρώτη εκτίμηση της αποδοτικότητας του μοντέλου, χωρίζεται το σετ δεδομένων σε ένα ποσοστό που καλείται σετ εκπαίδευσης και σε ένα άλλο ποσοστό που καλείται σετ ελέγχου. Η ιδέα πίσω από αυτό το διαχωρισμό είναι να εκπαιδευτεί ο αλγόριθμος σε ένα μέρος των δεδομένων (συνήθως ένα μεγάλο ποσοστό) και να γίνει ένας έλεγχος σε ένα άλλο σετ δεδομένων (το μικρότερο ποσοστό) το οποίο δεν έχει επεξεργαστεί ο αλγόριθμος. Έτσι, θα δοθεί μια εικόνα για το σφάλμα που θα παρατηρηθεί σε δεδομένα τα οποία δε γνώριζε ο αλγόριθμος κατά την εκπαίδευσή του.

2.2 Προεπεξεργασία δεδομένων

Η ελεύθερη πρόσβαση σε δεδομένα δεν σημαίνει απαραίτητα ότι τα δεδομένα αυτά έχουν και την κατάλληλη μορφή για να χρησιμοποιηθούν απευθείας σε ένα μοντέλο μηχανικής μάθησης. Αρκετές φορές υπάρχουν τιμές που δεν έχουν συμπληρωθεί, εμφανίζονται ακραίες τιμές, δεδομένα με διαφορετική μορφή και άλλα παρόμοια φαινόμενα. Αυτά τα μη καθαρά



Σχήμα 4: Παράδειγμα ταιριάσματος μοντέλων στα δεδομένα. a) Κανονική προσαρμογή του μοντέλου σε ένα τυχαίο σετ δεδομένων. b) Υπερπροσαρμογή του μοντέλου στα δεδομένα. c) Υποπροσαρμογή του μοντέλου στα δεδομένα.

δεδομένα επηρεάζουν αρνητικά τις τεχνικές μηχανικής μάθησης για την εξαγωγή προβλέψεων και τις κάνουν να καταλήγουν σε προβλέψεις που απέχουν πολύ από την πραγματικότητα. Πριν παρουσιαστούν τα βήματα αντιμετώπισης αυτών των περιπτώσεων, θα γίνει εισαγωγή βασικών στατιστικών μέτρων για την κατανόηση των προβλημάτων που δημιουργούνται (π.χ., της ακραίας τιμής).

Η συλλογή των δεδομένων γίνεται σε πίνακες, όπου η στήλη του κάθε πίνακα αντιπροσωπεύει το είδος της μεταβλητής και οι γραμμές τις μετρήσεις (τις καταστάσεις). Για παράδειγμα, ένα σετ δεδομένων φαίνεται στον Πίνακα 1, όπου έχει καταγραφεί η θερμοκρασία σε μία πόλη σε διάστημα μιας εβδομάδας. Για λόγους απλότητας, υποθέτουμε ότι η θερμοκρασία κατά τη διάρκεια της ημέρας παρέμενε σταθερή. Κάθε στήλη αντιπροσωπεύει το είδος της μεταβλητής που μελετάται. Η πρώτη στήλη αναφέρεται στον αριθμό των μετρήσεων, η δεύτερη στην ημέρα που έγινε η μέτρηση, ενώ η τρίτη στη θερμοκρασία.

Πίνακας 1: Σετ δεδομένων για παράδειγμα βασικών στατιστικών μέτρων.

	Ημέρα	**Θερμοκρασία (σε κελσίου)**
1	Δευτέρα	25
2	Τρίτη	24
3	Τετάρτη	27
4	Πέμπτη	25

	Ημέρα	Θερμοκρασία (σε κελσίου)
5	Παρασκευή	28
6	Σάββατο	25
7	Κυριακή	22

Μια πρώτη παρατήρηση σχετικά με τις μεταβλητές του προβλήματος, οι οποίες είναι:

1. **Ποιοτικές**: οι τιμές δεν είναι αριθμοί.
2. **Ποσοτικές**: οι τιμές είναι αριθμοί. Επιπλέον, οι ποσοτικές μεταβλητές χωρίζονται σε:
    a. **Διακριτές**: παίρνουν μεμονωμένες τιμές (π.χ., αριθμός ημερών βροχής σε ένα μήνα).
    b. **Συνεχής**: παίρνουν οποιαδήποτε τιμή μέσα σε ένα διάστημα πραγματικών αριθμών (π.χ., ατμοσφαιρική πίεση).

Τα στατιστικά μέτρα που θα μας απασχολήσουν, είναι:

- **Μέση τιμή** (Average value, $\bar{x}$), είναι η τιμή στην οποία βρίσκονται γύρω της τα δεδομένα. Προσοχή χρειάζεται όταν υπάρχουν μεγάλες τιμές στο δείγμα διότι θα επηρεαστεί. Η σχέση που περιγράφει τη μέση τιμή είναι:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{2.1}$$

  όπου i είναι η κάθε κατάσταση και n ο αριθμός των καταστάσεων ή αλλιώς των παρατηρήσεων.
- **Διάμεσος** (Median, M_d), ορίζεται ως η κεντρική τιμή όλων των δεδομένων. Αυτό σημαίνει πως αφήνει αριστερά και δεξιά της το 50% των μετρήσεων. Η μαθηματική σχέση που την περιγράφει, είναι:

$$M_d = X_{\frac{N+1}{2}} \; N = Περιττός \quad ή \quad M_d = \frac{X_{\frac{N}{2}} + X_{\frac{N+1}{2}}}{4} \; N = Άρτιος, \tag{2.2}$$

όπου N είναι ο συνολικός αριθμός των παρατηρήσεων.

- **Εύρος** ($Range$), ορίζεται ως η απόσταση της μεγαλύτερης τιμής των παρατηρήσεων ενός δείγματος, από τη μικρότερη:

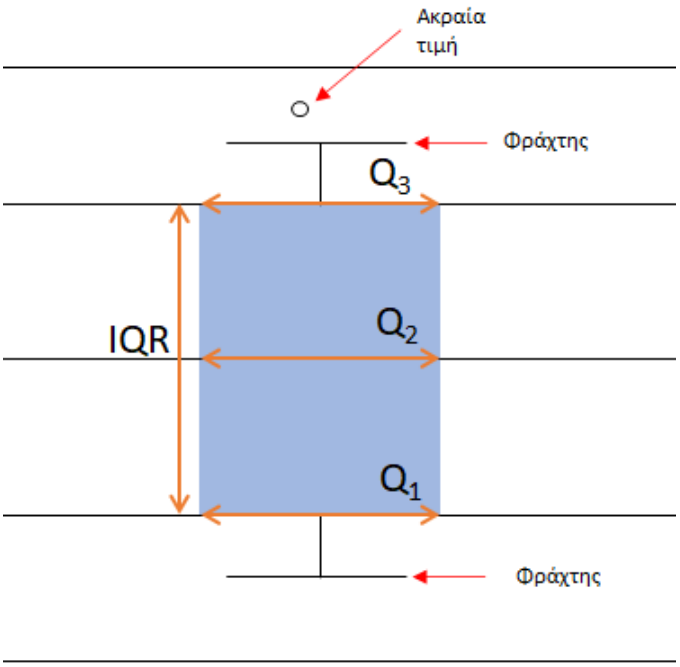
$$Range = X_{max} - X_{min} \tag{2.3}$$

- **Εκατοστημόρια** (ή εκατοστιαία σημεία), είναι μία γενίκευση της έννοιας της διαμέσου. Το p-οστό εκατοστημόριο από ένα σύνολο, ορίζεται η τιμή για την οποία εάν διατάξουμε τις τιμές σε αύξουσα σειρά, p% των παρατηρήσεων βρίσκονται αριστερά της και (100-p)% από δεξιά της. Τα εκατοστημόρια που χρησιμοποιούμε πιο συχνά είναι τα:
  - **Πρώτο τεταρτημόριο (Q_1) ⇨ 25%**
  - **Δεύτερο τεταρτημόριο (Q_2) ⇨ 50%** (=διάμεσος)
  - **Τρίτο τεταρτημόριο (Q_3) ⇨ 75%**

- **Ενδοτεταρτημοριακό εύρος** (IQR), ορίζεται ως η διαφορά του τρίτου τεταρτημόριου από το πρώτο και δίνεται από τη μαθηματική σχέση:

$$IQR = Q_3 - Q_1 \tag{2.4}$$

- **Θηκόγραμμα** (box-plot). Με βάση τα τεταρτημόρια κατασκευάζονται τα θηκογράμματα, τα οποία είναι γραφήματα όπου συνοψίζονται περιγραφικά μέτρα όπως η διάμεσος, τα τεταρτημόρια, το ενδοτεταρτημοριακό εύρος και το πιο σημαντικό στην περίπτωσή μας, οι ακραίες τιμές. Ένα παράδειγμα θηκογράμματος, φαίνεται στο Σχ. 5, όπου απεικονίζονται τα τεταρτημόρια μαζί με το ενδοτεταρτημοριακό εύρος. Η τιμή που αντιστοιχεί στο φράχτη είναι 1.5 φορές το ενδοτεταρτημοριακό εύρος, ενώ οποιαδήποτε τιμή εμφανιστεί πέρα από το φράχτη, θεωρείται ύποπτη ή αλλιώς παράτυπη ή αλλιώς ακραία τιμή. Τα θηκογράμματα χρησιμοποιούνται για τον έλεγχο της



Σχήμα 5: Μορφή Θηκογράμματος (box-plot).

ορθότητας των δεδομένων, καθώς η εμφάνιση μίας παράτυπης τιμής προϊδεάζει για εσφαλμένη καταχώρηση στα δεδομένα. Ιδιαίτερη προσοχή θέλει το γεγονός ότι οι παράτυπες τιμές δεν είναι απαραίτητα λάθος, ο όρος παράτυπη υποδηλώνει ότι δεν είναι συνηθισμένη. Έτσι, καθώς παρατηρηθεί μία παράτυπη τιμή, πρώτο βήμα είναι ο έλεγχος αυτής της τιμής, διότι μπορεί να έχει καταχωρηθεί λάθος η τιμή της, το εργαλείο που πραγματοποίησε τη μέτρηση να έχει υποστεί κάποια ζημιά, αλλά αρκετές φορές οι παράτυπες τιμές είναι τιμές που αντιστοιχούν σε πραγματικές παρατηρήσεις, όπου τότε επιλέγεται από τον εκάστοτε ερευνητή εάν επιθυμεί να τις λάβει υπόψη ή να τις διαχειριστεί διαφορετικά.

Μία άλλη τεχνική για την εύρεση μίας ακραίας τιμής, είναι περισσότερο εμπειρική. Τα δεδομένα ταξινομούνται σε αύξουσα ή φθίνουσα σειρά και ελέγχεται η μέγιστη και ελάχιστη τιμή του δείγματος για τυχόν εσφαλμένη καταχώρηση. Αξίζει να σημειωθεί πως αν και αυτή η τεχνική είναι γρήγορη, δεν επιλέγεται συχνά.

Τα βήματα που χρειάζονται για να καθαρίσουν τα δεδομένα και έτσι να οδηγήσουν σε πιο ασφαλείς προβλέψεις, περιλαμβάνουν τα στάδια (Swamynathan, 2019):

- Συλλογή δεδομένων.

- Αφαίρεση θορύβου και κατάλληλη διαχείριση των ακραίων τιμών.
- Αφαίρεση διπλότυπων τιμών στα δεδομένα και ασυνεπειών, μέσω αποδοχής πρότερης γνώσης της συμπεριφοράς του συστήματος.
- Τρόπο που θα αντιμετωπιστούν οι κενές τιμές στα δεδομένα.

Ενώ οι πιο σύνηθες τρόποι για την αντιμετώπιση των κενών δεδομένων είναι (Swamynathan, 2019):

- Διαγραφή των καταστάσεων όπου οι τιμές δεν είναι πλήρως συμπληρωμένες. Η περίπτωση αυτή συνήθως υιοθετείται όταν το πλήθος των καταστάσεων είναι μικρό συγκριτικά με το σετ δεδομένων.
- Συμπλήρωση των κενών τιμών με τη μέση τιμή/μέσο όρο των υπόλοιπων τιμών.
- Τυχαία συμπλήρωση της κενής τιμής με μία άλλη ίδιας φύσης. Προτιμάται η τεχνική αυτή όταν οι τιμές που λείπουν είναι λίγες και έτσι δε θα είναι στατιστικά σημαντική η μελέτη της συγκεκριμένης κατάστασης.
- Πρόβλεψη της τιμής από ένα άλλο μοντέλο.

2.3 Παλινδρόμηση

Η επιστήμη των δεδομένων, έχει ως στόχο την απόκτηση γνώσης του υπό μελέτη συστήματος μελετώντας τη συμπεριφορά των δεδομένων. Η παλινδρόμηση, είναι μία διαδικασία κατά την οποία εκτιμώνται οι σχέσεις που υπάρχουν στα δεδομένα. Χωρίζεται σε υποκατηγορίες όπου η κάθε κατηγορία ειδικεύεται σε διαφορετικό είδος μεταβλητών ή χρησιμοποιείται για διαφορετικό σκοπό. Συνήθως, η γραμμική παλινδρόμηση (linear regression) χρησιμοποιείται για την πρόβλεψη συνεχών μεταβλητών, ενώ η λογιστική παλινδρόμηση (logistic regression) για διακριτές (Ayyadevara, 2018). Αξίζει να σημειωθεί πως υπάρχουν τεχνικές όπως αυτές των νευρωνικών δικτύων, δέντρων απόφασης και άλλων, όπου είναι ικανές να προβλέψουν συνεχείς και διακριτές μεταβλητές, αλλά η αναφορά και ανάλυση τους γίνεται παρακάτω. Στο σημείο αυτό, θα αναλυθεί η γραμμική και η λογιστική παλινδρόμηση.

2.3.1 Γραμμική παλινδρόμηση

Γραμμική παλινδρόμηση είναι η τεχνική αποτύπωσης της στατιστικής σχέσης μεταξύ δύο ή περισσοτέρων μεταβλητών σε μία ευθεία γραμμή (ή περίπου ευθεία), όπου η αλλαγή της τιμής μίας μεταβλητής θα προκαλέσει ανάλογη αλλαγή σε άλλη μεταβλητή. Η χρήση μίας τέτοιας τεχνικής, βοηθά στην εκτίμηση μίας άγνωστης μεταβλητής από μία γνωστή. Αρκετές φορές, οι μεταβλητές αυτές αναφέρονται ως εξαρτώμενες και ανεξάρτητες, όπου με τον όρο εξαρτώμενη μεταβλητή γίνεται αναφορά στη μεταβλητή που γίνεται η πρόβλεψη, ενώ ανεξάρτητη είναι η μεταβλητή που χρησιμοποιήθηκε για την πρόβλεψη. Παραδείγματος χάριν, καθώς η θερμοκρασία που υπάρχει στο περιβάλλον μειώνεται, ο αριθμός των ατόμων που χρησιμοποιούν μπουφάν αυξάνεται. Στην προκειμένη περίπτωση, η θερμοκρασία είναι η ανεξάρτητη μεταβλητή που χρησιμοποιήθηκε για να γίνει η πρόβλεψη του αριθμού των μπουφάν που χρησιμοποιούνται, δηλαδή της εξαρτημένης μεταβλητής. Ιδιαίτερη προσοχή όμως, απαιτείται στον κατάλληλο ορισμό της αιτίας που το προκαλεί αυτό, διότι ενώ φαίνεται λογικό ότι φορούν όλο και περισσότεροι άνθρωποι μπουφάν καθώς η θερμοκρασία μειώνεται, το αντίθετο δεν αξιολογείται το ίδιο, διότι η θερμοκρασία δεν επηρεάζεται από τα ρούχα που φοράει κάποιος.

Αυτή η μεταβολή της μίας μεταβλητής λόγω μεταβολής της άλλης, υποδηλώνει ότι αυτές οι δύο μεταβλητές είναι κατά μία έννοια συσχετισμένες. Στη στατιστική χρησιμοποιούνται οι συντελεστές συσχέτισης για να προσδιορίσουν ακριβώς την εξάρτηση μίας μεταβλητής από την άλλη. Ένας τέτοιος συντελεστής συσχέτισης είναι ο Pearson r_{xy}, όπου υπολογίζει τη συσχέτιση της μεταβλητής x έναντι της y, με μαθηματική εξίσωση:

$$r_{xy} = \frac{A}{\sqrt{B}} \tag{2.5}$$

όπου,

$$A = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \quad (2.6)$$

$$B = \sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad (2.7)$$

Οι τιμές του συντελεστή συσχέτισης κυμαίνονται στο διάστημα [-1,1]. Μία τιμή κοντά στο -1 υποδηλώνει ότι υπάρχει αρνητική συσχέτιση στις μεταβλητές. Δηλαδή, όταν αυξάνει για παράδειγμα η μεταβλητή x, τότε η μεταβλητή y ελαττώνεται. Αντίθετα, μία τιμή κοντά στο 1 σημαίνει θετική εξάρτηση στις δύο μεταβλητές και κατά την αύξηση της μίας μεταβλητής παρατηρείται αύξηση και στην άλλη. Τέλος, τιμές κοντά στο 0 φανερώνουν μη συσχέτιση μεταξύ των δύο μεταβλητών καθώς η μεταβολή στη μία μεταβλητή δεν επηρεάζει τη δεύτερη.

Όταν η αλλαγή σε μία μεταβλητή επηρεάζεται από παραπάνω από ένα παράγοντες, η παλινδρόμηση ονομάζεται πολυμεταβλητή γραμμική παλινδρόμηση (multivariate linear regression), έναντι της απλής γραμμικής παλινδρόμησης (simple linear regression) που υπάρχει επηρεασμός της μεταβλητής από μόνο έναν παράγοντα. Ο φορμαλισμός μίας απλής γραμμικής παλινδρόμησης φαίνεται από την εξίσωση:

$$Y = a + b * X \quad (2.8)$$

όπου Yείναι η μεταβλητή που επιχειρείται η πρόβλεψη (εξαρτώμενη μεταβλητή), X είναι η μεταβλητή που χρησιμοποιείται για να γίνει η πρόβλεψη (ανεξάρτητη μεταβλητή), a είναι ο σταθερός συντελεστής (bias) και b είναι η κλίση της ευθείας (slope).

Για την κατανόηση των όρων a και b, ακολουθεί ένα παράδειγμα. Έστω ότι γίνεται πρόβλεψη των χρημάτων που έχει ένα εστιατόριο στο ταμείο του από τις 12:00 και μετά, ενώ την ίδια στιγμή υποθέτουμε ότι η επιχείρηση δεν έχει έξοδα ή φόρους, έτσι ό,τι μπαίνει στο ταμείο της μένει. Στην προκειμένη περίπτωση, με βάση την Εξ. (2.8), Y είναι τιμή της εξαρτημένης μεταβλητής, δηλαδή της πρόβλεψης που επιχειρείται, X είναι η ανεξάρτητη τιμή, δηλαδή τα χρήματα στο ταμείο της επιχείρησης, ενώ a είναι τα χρήματα που είχε στο ταμείο το εστιατόριο στις 12:00 (π.χ., 100 ευρώ). Υποθέτουμε ότι τα λεφτά του εστιατορίου, αυξάνονται

με σταθερό ρυθμό 50 ευρώ κάθε ώρα που περνάει, έως τις 20:00 που αποφασίζει να κλείσει. Ένα σχεδιάγραμμα που να αποτυπώνει αυτή τη σχέση φαίνεται στο Σχ. 6.

Η κλίση της ευθείας υπολογίζεται από τη μαθηματική σχέση:

$$b = \frac{Y_{max} - Y_{min}}{X_{max} - X_{min}} \tag{2.9}$$

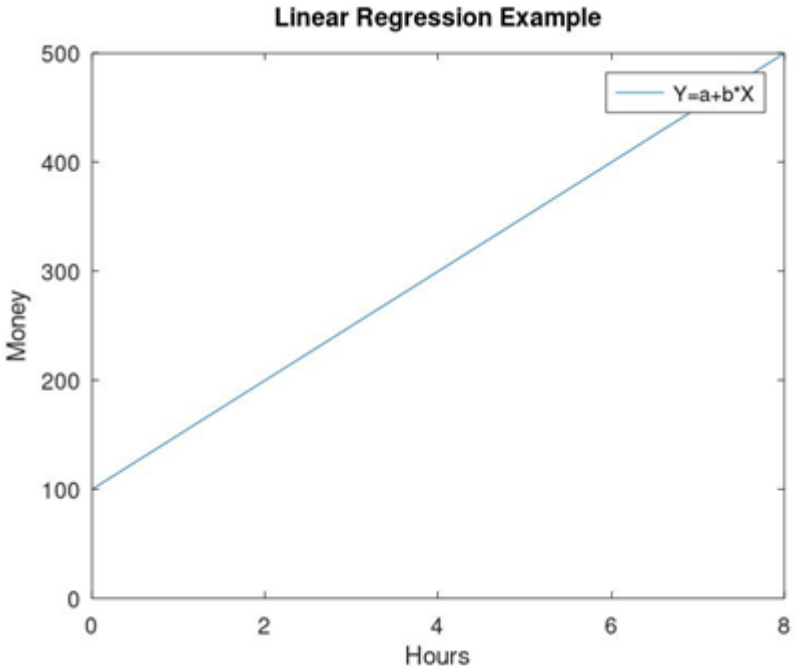
όπου Y_{max}, Y_{min}, X_{max} , X_{min} είναι οι μέγιστες και ελάχιστες τιμές των δεδομένων Y και X αντίστοιχα. Το συνολικό σετ δεδομένων του παραπάνω προβλήματος, αναπαρίσταται στον Πίνακα 2. Η μηδενική ώρα είναι η ώρα που ξεκινάει η πρόβλεψη (12:00), κάθε ώρα που περνάει αναγράφεται με τον αριθμό της, έως την όγδοη ώρα που σταματά η πρόβλεψη (20:00). Για να λύσει κανείς το πρόβλημα της απλής γραμμικής παλινδρόμησης του παραπάνω παραδείγματος και να βρει τους συντελεστές χρειάζεται τουλάχιστον δύο τιμές για τα X και Y. Άρα, με βάση τον Πίνακα 2, παίρνουμε τις τιμές:

$$Y_{max} = 500, Y_{min} = 100, X_{max} = 8, X_{min} = 0$$

$$Y_{max} = 500, Y_{min} = 100, X_{max} = 8, X_{min} = 0$$

Πίνακας 2: Σετ δεδομένων για παράδειγμα γραμμικής παλινδρόμησης.

Ώρα	Χρήματα
0	100
1	150
2	200
3	250
4	300
5	350
6	400
7	450
8	500



Σχήμα 6: Παράδειγμα γραμμικής παλινδρόμησης. Ο άξονας των X αποτυπώνει την κάθε ώρα, όπου 0 είναι η αρχική μέτρηση στις 12:00 και 8 είναι η τελική μέτρηση στις 20:00. Ο άξονας των Y αποτυπώνει τον αριθμό των χρημάτων που υπάρχουν στο ταμείο.

Όπου $X_{min} = 0$ εννοούμε την ώρα 12:00 και $X_{max} = 8$ εννοούμε την ώρα 20:00, ενώ $Y_{max} = 500$, Y_{min} =100 είναι τα μέγιστα και τα ελάχιστα χρήματα στο ταμείο του εστιατορίου. Αντικαθιστώντας τα δεδομένα στην Εξ. (2.9) προκύπτει:

$$b = \frac{500 - 100}{8 - 0} = \frac{400}{8} \Rightarrow b = 50.$$

Ένας πιο γενικός τρόπος επίλυσης του παραπάνω προβλήματος είναι ο εξής. Έστω ότι έχουμε δεδομένα μόνο για τις ώρες 0 και 1 (δηλαδή τις ώρες 12:00 και 13:00). Τότε το συνολικό σετ δεδομένων θα ήταν όπως του Πίνακα 3. Έχοντας ως βασική υπόθεση ότι μετράμε τα χρήματα του εστιατορίου με βάση το πόσα χρήματα υπάρχουν στο ταμείο, αντικαθιστούμε τα δεδομένα του Πίνακα 3 στην Εξ. 2.8 και προκύπτει:

$$100 = a + b * 0$$
$$150 = a + b * 1$$

Λύνοντας το παραπάνω σύστημα εξισώσεων, προκύπτει:

$$a = 100, \quad b = 50\,.$$

Πίνακας 3: Σετ δεδομένων για παράδειγμα γραμμικής παλινδρόμησης**.**

Ώρα	**Χρήματα**
0	100
1	150

Άρα κάνοντας την υπόθεση ότι στο ταμείο του εστιατορίου στις 12:00 (ώρα 0) υπήρχαν 100 ευρώ και στις 13:00 (ώρα 1) 150 ευρώ, παρατηρούμε πως κατά τη λύση ο σταθερός συντελεστής a ισούται με 100 και η κλίση της ευθείας b ισούται με 50. Εάν συνεχίσουμε για τις υπόλοιπες ώρες όπως στο προηγούμενο παράδειγμα, ενώ σαν γνώση της συμπεριφοράς των δεδομένων έχουμε μόνο τον Πίνακα 3 και κατ' επέκταση τις τιμές των a και b όπως υπολογίστηκαν παραπάνω, τότε η προσέγγιση των χρημάτων που θα έχει το εστιατόριο στο ταμείο του, φαίνεται στον Πίνακα 4.

Πίνακας 4: Σετ δεδομένων για παράδειγμα γραμμικής παλινδρόμησης.

Ώρα	**Χρήματα (προσέγγιση)**	**Χρήματα (πραγματική τιμή)**	**Απόκλιση από πραγματική τιμή**
0	100	100	0
1	150	150	0
2	200	200	0
3	250	250	0
4	300	300	0
5	350	350	0
6	400	400	0
7	450	450	0
8	500	500	0

Το πρόβλημα λύνεται χρησιμοποιώντας μόνο δύο καταστάσεις (ώρες 0,1) και μπορεί εύκολα να συνεχιστεί για όσες ώρες χρειαστεί με τη μικρότερη δυνατή απόκλιση από την πραγματική τιμή. Το θέμα που δημιουργείται με την εφαρμογή της παραπάνω τεχνικής στην καθημερινότητα, είναι ότι τα δεδομένα που χρησιμοποιούνται δεν είναι τόσο ξεκάθαρα. Επίσης, για να χρησιμοποιηθεί όλο το σετ δεδομένων και να δοθεί μία πλήρης εικόνα της συμπεριφοράς των δεδομένων και όχι μόνο δύο τιμών, όπως έγινε προηγουμένως, χρησιμοποιούνται δείκτες ακρίβειας που υπολογίζουν το σφάλμα της εκτίμησης στα δεδομένα, όπως το συνολικό τετραγωνικό σφάλμα (sum of squared error) το οποίο θέλουμε να ελαχιστοποιήσουμε (ordinary least squares method).

Στο συνολικό τετραγωνικό σφάλμα, υπολογίζεται το άθροισμα των διαφορών του τετραγώνου της πραγματικής τιμής και της τιμής που προβλέφθηκε σε όλες τις παρατηρήσεις, με μαθηματική σχέση:

$$SSE = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 \qquad (2.10)$$

όπου $\widehat{Y}$ είναι η τιμή που προβλέπεται, Y είναι η επιθυμητή (πραγματική), ο αριθμός n αντιστοιχεί στον αριθμό των καταστάσεων που μελετήθηκαν και i είναι η ατομική κατάσταση.

Η διαφορά της προβλεπόμενης τιμής από την πραγματική παρουσιάζει την απόσταση της πρόβλεψης από την πραγματική τιμή. Ο λόγος για τον οποίο επιλέγεται το τετράγωνο της διαφοράς και όχι απλά η διαφορά, είναι για την αποφυγή ενός ακούσιου λάθους. Για παράδειγμα, εάν έχουμε Y σωστή τιμή σε μία κατάσταση, τότε η προβλεπόμενη τιμή $\widehat{Y}$, μπορεί να είναι είτε μικρότερη, είτε μεγαλύτερη, είτε ίση με τη σωστή απάντηση. Εάν η προβλεπόμενη τιμή είναι μικρότερη της σωστής απάντησης, τότε η διαφορά μεταξύ των δύο θα είναι θετική. Αντίθετα, αν η προβλεπόμενη τιμή είναι μεγαλύτερη από τη σωστή τιμή, τότε η διαφορά θα είναι αρνητική, ενώ αν είναι ίσες η διαφορά θα είναι μηδέν. Εάν προσθέσουμε όμως όλα τα σφάλματα, υπάρχει η περίπτωση το συνολικό σφάλμα να βγάλει ένδειξη μηδέν, καθώς αλληλοεξουδετερώνονται τα θετικά σφάλματα με τα αρνητικά. Έτσι, για να αποφύγουμε αυτού του είδους την περίπτωση, επιλέγουμε να τετραγωνίσουμε το σφάλμα. Η τακτική που ακολουθείται με τη χρήση ενός τέτοιου μέτρου ακρίβειας είναι η ελαχιστοποίησή του, μέσω

εύρεσης των βέλτιστων τιμών των παραμέτρων a και b, έτσι ώστε η προβλεπόμενη τιμή να είναι όσο το δυνατόν πιο κοντά στην αληθινή.

Η τεχνική για να ληφθούν οι βέλτιστες τιμές (gradient descent) συνοψίζεται σε 6 βήματα (Ayyadevara, 2018):

1. Θέτουμε τυχαίες τιμές στις σταθερές a και b.
2. Υπολογίζουμε το συνολικό τετραγωνικό σφάλμα μέσω της Εξ. 2.10 για όλες τις καταστάσεις στα δεδομένα (αυτή είναι γνωστή και ως συνάρτηση κόστους).
3. Αλλάζουμε ελάχιστα τις τιμές στις σταθερές.
4. Ελέγχουμε αν με την αλλαγή στις τιμές των σταθερών επηρεάστηκε το συνολικό τετραγωνικό σφάλμα.
5. Εάν το συνολικό τετραγωνικό σφάλμα μειώθηκε, τότε επαναλαμβάνουμε τη διαδικασία του βήματος 3. Δηλαδή, αν είχαμε επιλέξει να αυξήσουμε ελάχιστα τις μεταβλητές, το επαναλαμβάνουμε. Στην περίπτωση που το συνολικό τετραγωνικό σφάλμα αυξηθεί, επιλέγουμε την αντίθετη διαδικασία από αυτή που ακολουθήθηκε στο βήμα 3. Δηλαδή, αν είχαμε αυξήσει τις τιμές των μεταβλητών, τώρα τις μειώνουμε.
6. Επαναλαμβάνουμε τα βήμα 2 έως 4, μέχρι το συνολικό τετραγωνικό σφάλμα να είναι το μικρότερο δυνατό.

Για την απεικόνιση μίας γραμμικής παλινδρόμησης σε ένα σετ δεδομένων, θα χρειαστεί ο ορισμός κάποιων στατιστικών μέτρων, όπως:

- **Υπολείμματα** ($Residuals$). Ορίζεται ως η διαφορά της πραγματικής τιμής από την προβλεπόμενη, ενώ το άθροισμά τους δίνει την κατανομή των σφαλμάτων.

$$Residuals = Y - \hat{Y} \tag{2.11}$$

Ενώ η απόκλιση των υπολειμμάτων (Residual Deviance), φαίνεται στην Εξ. (2.10).

- **Μηδενική Απόκλιση** $(Null\ Deviance)$. Ορίζεται ως η απόκλιση, κατά τη μη χρήση ανεξάρτητων μεταβλητών στο μοντέλο.

$$Null\ Deviance = \sum_{i=1}^{n} (Y - \overline{Y})^2 \tag{2.12}$$

όπου, $\overline{Y}$ είναι ο μέσος όρος της εξαρτώμενης μεταβλητής.

- **R τετράγωνο** (R Squared - R^2). Αποτυπώνει τη συσχέτιση μεταξύ των προβλεπόμενων μεταβλητών και των πραγματικών. Μας δείχνει δηλαδή πόσο καλά ταιριάζει το μοντέλο μας με τη συμπεριφορά των δεδομένων. Παίρνει τιμές από το 0 έως το 1, ενώ μία τιμή κοντά στο 1 υποδηλώνει καλύτερο ταίριασμα στα δεδομένα. Συνήθως, υπολογίζεται με το να τετραγωνιστεί η συσχέτιση μεταξύ των πραγματικών τιμών των εξαρτημένων μεταβλητών και των τιμών των προβλέψεων των εξαρτημένων μεταβλητών. Άλλη σχέση που το υπολογίζει είναι:

$$R^2 = 1 - (ResidualDeviance - NullDeviance) \tag{2.13}$$

Όταν η τιμή της απόκλισης των υπολειμμάτων είναι αρκετά μικρότερη σε σχέση με την τιμή της μηδενικής απόκλισης, έχει γίνει καλή προσέγγιση στις προβλεπόμενες τιμές και το R^2 είναι μεγάλο.

Έως τώρα, είδαμε έναν τρόπο επίλυσης του παραδείγματος απλής γραμμικής παλινδρόμησης μέσω υπολογισμού της κλίσης της ευθείας b από την Εξ. (2.9) και υπολογισμού του σταθερού συντελεστή a από τα χρήματα που υπήρχαν στο ταμείο τη μηδενική ώρα. Μία άλλη λύση, δίνεται από μέσω της επίλυσης ενός συστήματος εξισώσεων μέσω δεδομένων του Πίνακα 3. Τώρα, θα επιλύσουμε το πρόβλημα απλής γραμμικής παλινδρόμησης (εύρεση των a και b) μέσω της γλώσσας προγραμματισμού Python και θα απεικονίσουμε τα αποτελέσματα. Στο σετ δεδομένων του Πίνακα 2 γίνεται προσέγγιση της λύσης μέσω ελαχιστοποίησης του σφάλματος (τεχνική ordinary least squares) (βλ. Σχ. 7) και υπολογίζεται η τιμή του σταθερού συντελεστή $a = 100$ και της κλίσης της ευθείας $b = 50$. Επίσης, υπολογίστηκε το

```
==========================================
Dep. Variable:                   Money
Model:                             OLS
Method:                  Least Squares
Date:                 Thu, 19 Jan 2023
Time:                         14:31:36
No. Observations:                    9
Df Residuals:                        7
Df Model:                            1
Covariance Type:            nonrobust
==========================================
                 coef    std err
------------------------------------------
Intercept    100.0000   1.02e-13
Hours         50.0000   2.15e-14
==========================================
```

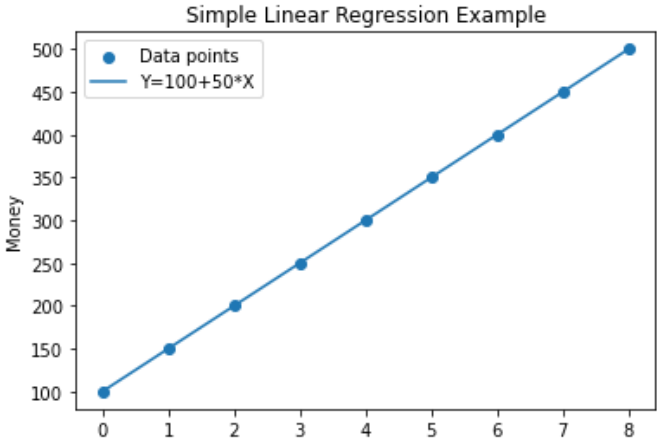
Σχήμα 7: Επίλυση απλής γραμμικής παλινδρόμησης μέσω Python. Η τιμή του Intercept αντιστοιχεί στο $\boldsymbol{a}$ και η τιμή του Hours στο $\boldsymbol{b}$.

τυπικό σφάλμα της κάθε προσέγγισης το οποίο είναι ένα μέτρο αποτύπωσης της μεταβολής των τιμών a και b. Για παράδειγμα, για τον σταθερό συντελεστή a η πραγματική του τιμή μπορεί να κυμαίνεται $1.02e-13$ μακριά από την τιμή 100, ενώ για το λόγο ότι αυτή η μεταβολή είναι αρκετά μικρή, σπάνια λαμβάνεται υπόψη. Όμοια, για την τιμή της κλίσης b.

Επίσης, η απεικόνιση των δεδομένων αλλά και η ευθεία απλής γραμμικής παλινδρόμησης για το παραπάνω παράδειγμα:

$$Y = 100 + 50 * X \tag{2.14}$$

αποτυπώνεται στο Σχ. 8. Το πρώτο που παρατηρούμε είναι ότι η ευθεία γραμμή ταυτίζεται με τα δεδομένα. Η ευθεία γραμμή είναι η ευθεία της Εξ. (2.14) και παρατηρούμε ότι η περιγραφή των δεδομένων γίνεται έμπιστα, καθώς περνάει από το κάθε δεδομένο. Πράγματι, για το λόγο ότι δεν υπάρχει απόκλιση της προβλεπόμενης τιμής από την πραγματική, περιμέναμε το R^2 να είναι το μέγιστο (=1) το οποίο υπολογίστηκε 0.9999999999999996. Όσο μεγαλύτερο και πιο κοντά στο 1 είναι η τιμή του R^2, τόσο καλύτερο ταίριασμα θα παρατηρείται στα δεδομένα. Στην προκειμένη περίπτωση, τα δεδομένα ήταν σχεδιασμένα έτσι ώστε να είναι γραμμικά συσχετισμένα. Στην πράξη όμως, τα μοντέλα που χρησιμοποιούνται για την πρόβλεψη των τιμών δεν έχουν τόσο καλή προσέγγιση και η τιμή του R^2 είναι πιο χαμηλή.



Σχήμα 8: Παράδειγμα ταιριάσματος (fitting) της εξίσωσης απλής γραμμικής παλινδρόμησης $\boldsymbol{Y = 100 + 50 * X}$ στο σετ δεδομένων του Πίνακα 2.

Διάφορα προβλήματα κατά τη χρήση της απλής γραμμικής παλινδρόμησης, προκύπτουν:

- Όταν η σχέση ανάμεσα στα δεδομένα δεν είναι πάντα γραμμική. Στο προηγούμενο παράδειγμα, είχαμε αναφερθεί στα χρήματα που υπάρχουν στα ταμεία του εστιατορίου για τις ώρες 12:00 έως 20:00. Εάν με το μοντέλο της απλής γραμμικής παλινδρόμησης επιχειρούσε κανείς να προβλέψει τα έσοδα στις 21:00, θα έβρισκε μία τιμή μακριά από την πραγματική, για το λόγο ότι το εστιατόριο έκλεισε στις 20:00 και τα χρήματα παρέμειναν ίδια έως τις 21:00. Όμοια, η τιμή που θα προβλεπόταν για τις 22:00 θα ήταν ακόμη πιο μακριά από την πραγματική διότι και στις 22:00 ήταν κλειστό το εστιατόριο. Άρα, στην προκειμένη περίπτωση παρατηρείται μία γραμμική αύξηση έως μία συγκεκριμένη τιμή όπου πέρα από αυτή η τιμή είναι σταθερή.
- Όταν υπάρχει μία ακραία τιμή στις ανεξάρτητες μεταβλητές επηρεάζει αρκετά τις τιμές των παραμέτρων κατά τη διαδικασία ελαχιστοποίησης του μέσου σφάλματος.

Τακτικές για την αποφυγή τέτοιου είδους περιπτώσεων, περιλαμβάνουν:

- Η ανάλυση της συμπεριφοράς των μεταβλητών και η επιλογή τμημάτων από το σύνολο όπου μπορεί να εφαρμοστεί απλή γραμμική παλινδρόμηση. Όπως για παράδειγμα, υπάρχουν κάποιες ώρες μέσα στην ημέρα όπου το εστιατορίου του προηγούμενου παραδείγματος παραμένει κλειστό. Έτσι, εάν πετάξουμε εκτός αυτές τις ώρες μπορεί να εφαρμοστεί απλή γραμμική παλινδρόμηση στις επόμενες. Επίσης, ορισμένα διαστήματα ημερών μέσα στο χρόνο ενδέχεται να έχουν περισσότερη ή και

λιγότερη πελατεία. Έτσι, μπορεί να προσεγγιστεί κάθε διάστημα ξεχωριστά για το λόγο ότι θα παρουσιάζει διαφορετική κλίση (αλλού θα είναι πιο απότομη η αύξηση ενώ αλλού λιγότερο απότομη).

- Κανονικοποίηση ακραίων τιμών έως το ενενηκοστό ένατο ποσοστιαίο σημείο καθώς τότε οι παρατηρήσεις με αρκετά παράξενη συμπεριφορά δε θα επηρεάζουν το αποτέλεσμα κατά πολύ.

Σε αντίθεση με την απλή γραμμική παλινδρόμηση, μία πιο περίπλοκη σχέση σε δεδομένα είναι πιο πιθανή να υπάρχει και αντί να παρατηρείται συσχέτιση μόνο σε δύο, να διακρίνεται για περισσότερες μεταβλητές. Τότε όπως προαναφέρθηκε, η μέθοδος παλινδρόμησης ονομάζεται πολυμεταβλητή γραμμική παλινδρόμηση, ενώ ο φορμαλισμός που την αποτυπώνει είναι:

$$Y = a + w_1 \cdot X_1 + w_2 \cdot X_2 + .. \qquad (2.15)$$

όπου w_i $(i = 1,2, \dots)$ είναι τα βάρη της κάθε ανεξάρτητης μεταβλητής αντίστοιχα και a ο σταθερός συντελεστής. Για παράδειγμα, εάν τα χρήματα που υπήρχαν στο ταμείο της επιχείρησης ήταν περισσότερα εάν η ημέρα ήταν Σάββατο ή Κυριακή, ενώ λιγότερα τις καθημερινές, τότε η Εξ. (2.15) θα πάρει τη μορφή της εξίσωσης:

$$Y = a + w_1 * X_1 + w_2 * X_2$$

όπου a είναι ο σταθερός συντελεστής, X_1 είναι η μεταβλητή που αντιστοιχεί στην κάθε ώρα, ενώ X_2 είναι η μεταβλητή που αντιστοιχεί στην ημέρα που βρισκόμαστε. Τέλος, w_1 και w_2 είναι τα βάρη των μεταβλητών X_1 και X_2 αντίστοιχα, όπου η τιμή τους θα αντικατοπτρίζει τη σημαντικότητα της μεταβλητής στο αποτέλεσμα. Κατά την επίλυση του προβλήματος πολυμεταβλητής γραμμικής παλινδρόμησης μέσω ελαχιστοποίησης του συνολικού τετραγωνικού σφάλματος, όμοια με προηγουμένως ψάχνουμε τις κατάλληλες τιμές των a, w_1, w_2 έτσι ώστε το συνολικό τετραγωνικό σφάλμα να είναι το μικρότερο δυνατό.

Διάφορα προβλήματα κατά τη χρήση της πολυμεταβλητής γραμμικής παλινδρόμησης, προκύπτουν:

- Από την πολυσυγγραμικότητα (multicollinearity), στην οποία υπάρχει η πιθανότητα οι ανεξάρτητες μεταβλητές να είναι συσχετισμένες μεταξύ τους.
- Οι τιμές των συντελεστών (a, b ή $a, w_1, w\_\{2\}, \dots$) είναι αρκετά μεγάλες. Σε αυτή την περίπτωση, ακόμη και μία ελάχιστη μεταβολή της ανεξάρτητης μεταβλητής θα προκαλέσει μεγάλη αλλαγή στην προβλεπόμενη τιμή. Έτσι, προτιμάται μία μικρή τιμή στους συντελεστές, χωρίς να απορριφθεί το ενδεχόμενο σε ορισμένες περιπτώσεις να χρειαζόμαστε μεγαλύτερες τιμές στους συντελεστές.
- Τα δεδομένα είναι ελάχιστα. Όπως στο προηγούμενο παράδειγμα χρησιμοποιήσαμε δύο μόνο καταστάσεις στα δεδομένα και βρήκαμε μία εξίσωση να τα περιγράφει, έτσι βρέθηκε και μία εξίσωση χρησιμοποιώντας εννέα καταστάσεις στα δεδομένα. Το ερώτημα που προκύπτει εδώ, είναι εάν αισθάνεται κανείς πιο σίγουρος έχοντας μία εξίσωση που προέκυψε μελετώντας εννέα καταστάσεις στα δεδομένα ή μία εξίσωση που προέκυψε μελετώντας μόλις δύο καταστάσεις. Προφανώς θα επιλέξουμε τη λύση που προέκυψε από τα περισσότερα δεδομένα διότι θα υπάρχει πιο σφαιρική εικόνα στη συμπεριφορά των δεδομένων. Έτσι, εάν αντίστοιχα είχαμε μία εξίσωση που προέκυψε από εκατό καταστάσεις στα δεδομένα, θα την προτιμούσαμε από εκείνη με τις εννέα. Ο λόγος είναι ότι οι λύσεις που προκύπτουν μελετώντας πολλές καταστάσεις, είναι πιο εύκολο να γενικευτούν διότι μία μελέτη σε λίγα δεδομένα μπορεί να έχει υποπέσει στην "εξαίρεση του κανόνα" και έτσι να μην εφαρμόζεται πάντα.

2.3.2 Λογιστική παλινδρόμηση

Η Λογιστική παλινδρόμηση σε αντίθεση με τη γραμμική παλινδρόμηση, χρησιμοποιείται για την πρόβλεψη διακριτών τιμών (π.χ., εάν έγινε σεισμός ή όχι). Σε τέτοιου είδους προβλήματα το αποτέλεσμα της πρόβλεψης θα είναι είτε 1 (εάν το συμβάν έγινε) είτε 0 (εάν το συμβάν δεν έγινε). Ενώ η γραμμική παλινδρόμηση είναι ικανή να εφαρμοστεί με επιτυχία σε αρκετά

προβλήματα, οι αδυναμίες της φαίνονται όταν η προβλεπόμενη τιμή θέλουμε να είναι 0 ή 1. Για παράδειγμα, στον Πίνακα 5 υπάρχουν παρατηρήσεις που αναφέρονται στη διαφορά δυναμικής ομάδων πετοσφαίρισης και στο αν η γηπεδούχος ομάδα σημείωσε νίκη κατά την αναμέτρησή τους. Σημειώνουμε ότι νικητής σε έναν αγώνα πετοσφαίρισης είναι όποιος κερδίσει πρώτος τρία σετ. Έτσι, χωρίς να υπάρχει περίπτωση ισοπαλίας, τα μοναδικά αποδεκτά αποτελέσματα είναι είτε 1 (νίκη γηπεδούχου) είτε 0 (ήττα γηπεδούχου).

Πίνακας 5: Σετ δεδομένων για παράδειγμα λογιστικής παλινδρόμησης.

Δυναμική Ομάδων	**Νίκη Γηπεδούχου**
300	1
500	1
-200	0

Κατά την επίλυση του παραπάνω προβλήματος ως πρόβλημα γραμμικής παλινδρόμησης με τη μέθοδο ordinary least squares (βλ. Σχ. 9) προκύπτει η Εξ. (2.16), όπου με τη μεταβλητή Y αναφερόμαστε στην περίπτωση νίκης της γηπεδούχου ομάδας και με X στη διαφορά δυναμικής των ομάδων.

$$Y = 0.359 + 0.0015 * X \tag{2.16}$$

Αν στην Εξ. (2.16) αντικαταστήσουμε τις δυναμικές των ομάδων του Πίνακα 5, θα πάρουμε:

$$Y = 0.359 + 0.0015 * 300 \Rightarrow Y = 0.809$$
$$Y = 0.359 + 0.0015 * 500 \Rightarrow Y = 1.109$$
$$Y = 0.359 + 0.0015 * (-200) \Rightarrow Y = 0.059$$

Ένα πρόβλημα παρατηρείται στην δεύτερη παρατήρηση, όπου η διαφορά στη δυναμική όταν ισούται με 500, έδωσε τιμή μεγαλύτερη του 1. Επειδή όμως αυτό δεν έχει λογική, διότι η γηπεδούχος ομάδα είτε κέρδισε ($Y = 1$) είτε έχασε ($Y = 0$), θα ήταν ιδιαίτερα βολικό όλες οι

```
==========================================
Dep. Variable:               HomeWins
Model:                            OLS
Method:                 Least Squares
Date:                Thu, 19 Jan 2023
Time:                        18:09:58
No. Observations:                   3
Df Residuals:                       1
Df Model:                           1
Covariance Type:            nonrobust
==========================================
                          coef    std err
------------------------------------------
Intercept              0.3590      0.158
DifferenceInRating     0.0015      0.000
==========================================
```

Σχήμα 9: Επίλυση απλής γραμμικής παλινδρόμησης μέσω Python. Οι τιμές του Intercept και DifferenceInRating αντιστοιχούν στις τιμές $\boldsymbol{a}$ και $\boldsymbol{b}$.

τιμές πάνω από το 1 να στρογγυλοποιούνται στο 1 και όλες οι τιμές κάτω από το 0 να στρογγυλοποιούνται στο 0. Άρα τα ερωτήματα σε τέτοιου είδους περιπτώσεις συνοψίζονται σε:

- Μήπως η γραμμική παλινδρόμηση δεν είναι ικανή να περιγράψει ένα τέτοιο φαινόμενο;
- Μήπως κάποια άλλη μορφή παλινδρόμησης μπορεί να αξιοποιήσει καλύτερα τα δεδομένα μέσω μίας διαφορετικής συνάρτησης;

Τα μειονεκτήματα της γραμμικής παλινδρόμησης στην πρόβλεψη διακριτών τιμών, μεταξύ άλλων είναι (Ayyadevara, 2018):

- Η υπόθεση ότι οι μεταβλητές σχετίζονται γραμμικά. Ενώ, καθώς παρατηρείται μία μεγαλύτερη διαφορά στη δυναμική των ομάδων λόγω διάφορων παραγόντων (όπως ποιότητα παικτών, εμπειρία, κλπ.), οι πιθανότητες νίκης αυξάνουν εκθετικά.
- Δεν υπολογίζεται η πιθανότητα λάθους. Στα αθλήματα συχνά αναφέρεται ο όρος "outsider". Ο όρος αυτός αναφέρεται στην ομάδα με τη μικρότερη δυναμική ενώ αν καταφέρει να κερδίσει το "outsider" (έστω και με μικρότερες πιθανότητες), η γραμμική παλινδρόμηση αδυνατεί να προβλέψει το ενδεχόμενο.
- Υποθέτει ότι η πιθανότητα να συμβεί ένα γεγονός αυξάνει γραμμικά, καθώς αυξάνεται η τιμή της ανεξάρτητης μεταβλητής. Από τη στιγμή που η διαφορά της δυναμικής των ομάδων είναι μεγάλη (π.χ., >600), δεν έχει και τόσο μεγάλη σημασία για την

πιθανότητα νίκης αν η διαφορά είναι 700 ή 800, καθώς η διαφορά της μίας ομάδας από την άλλη είναι εξαιρετικά μεγάλη (το ίδιο ισχύει και για -700, -800 για την περίπτωση ήττας).

Το μεγαλύτερο μειονέκτημα από όλα τα παραπάνω είναι η υπόθεση ότι όλες οι μεταβλητές σχετίζονται γραμμικά, ενώ στην πραγματικότητα λίγες από αυτές το κάνουν. Η λογιστική παλινδρόμηση δεν υιοθετεί τον περιορισμό της γραμμικής συσχέτισης των παρατηρήσεων και μέσω της σιγμοειδής καμπύλης, προβάλλει τα αποτελέσματά της στο εύρος [0,1]. Η σιγμοειδής καμπύλη, δίνεται από τη σχέση:

$$S(t) = \frac{1}{1 + e^{-t}} \tag{2.17}$$

όπου για μεγάλες τιμές του t, το εκθετικό είναι κοντά στο 0, άρα το $S(t)$ είναι κοντά στο 1, ενώ για μικρές τιμές του t, το εκθετικό είναι αρκετά μεγάλο, άρα ο παρονομαστής είναι αρκετά μεγάλος και έτσι το $S(t)$ είναι κοντά στο 0. Η αναπαράσταση της σιγμοειδής καμπύλης φαίνεται στο Σχ. 10. Αυτό που παρατηρούμε είναι ότι όλα τα αποτελέσματα της καμπύλης της Εξ. (2.17) είναι στο διάστημα [0,1], ενώ από ένα t και μετά οριζοντοποιείται. Στα θετικά t οριζοντοποιείται στην τιμή 1, ενώ στα αρνητικά t οριζοντοποιείται στο 0.

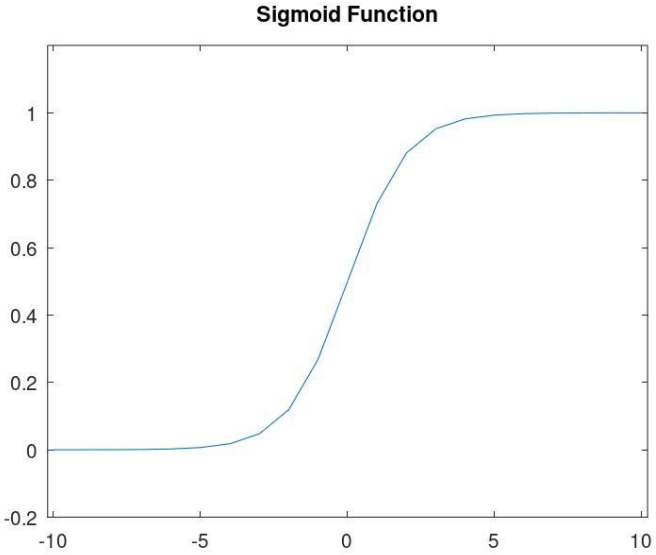
Όμοια, η λογιστική παλινδρόμηση, δίνεται από τη σχέση:

$$Y = \frac{1}{1 + e^{-(a+b*X)}} \tag{2.18}$$

ενώ σε περιπτώσεις πολλών μεταβλητών, το εκθετικό της Εξ. (2.18) παίρνει τη μορφή της Εξ. (2.15).

Κατά τη γραμμική παλινδρόμηση, καθώς κάναμε μετάβαση από το $X = 0$ στο $X = 1$, είδαμε ότι η αύξηση που γινόταν στο Y, ήταν όσο και η τιμή της κλίσης της καμπύλης b. Για να καταλάβει κανείς την ραγδαία μεταβολή όμως στη λογιστική παλινδρόμηση, ας πάρουμε την ευθεία:

$$Y = 1 + 2 * X \tag{2.19}$$



Σχήμα 10: Σιγμοειδής καμπύλη. Το αποτέλεσμα της για όλα τα $\boldsymbol{x}$ είναι στο διάστημα [0,1]. Για μία τιμή $\boldsymbol{x} > \mathbf{0}$, η καμπύλη οριζοντοποιείται και δίνει πάντα 1, ενώ αντίστοιχα για μία τιμή $\boldsymbol{x} < \mathbf{0}$ η καμπύλη οριζοντοποιείται και δίνει πάντα 0.

Ας χρησιμοποιήσουμε αυτή την ευθεία σαν γραμμική παλινδρόμηση και σαν λογιστική παλινδρόμηση. Άρα, για να τη χρησιμοποιήσουμε σαν λογιστική παλινδρόμηση, ας αντικαταστήσουμε την ευθεία της Εξ. (2.19) στην Εξ. (2.18):

$$Y = \frac{1}{(1 + e^{-(1+2\cdot X)})} \tag{2.20}$$

Ας αντικαταστήσουμε τώρα τις τιμές $X = 0,1,2$ στις εξισώσεις (2.19) και (2.20).

$$X = 0 \quad Y = 1 + 2 * 0 \Rightarrow Y = 1$$
$$X = 1 \quad Y = 1 + 2 * 1 \Rightarrow Y = 3$$
$$X = 2 \quad Y = 1 + 2 * 2 \Rightarrow Y = 5$$
$$X = 0 \quad Y = \frac{1}{(1 + e^{-(1+2\cdot 0)})} \Rightarrow Y = 0.7311$$
$$X = 1 \quad Y = \frac{1}{(1 + e^{-(1+2*1)})} \Rightarrow Y = 0.9526$$
$$X = 2 \quad Y = \frac{1}{(1 + e^{-(1+2*2)})} \Rightarrow Y = 0.9933$$

Αυτό που παρατηρούμε είναι πως στην περίπτωση της γραμμικής παλινδρόμησης γίνεται μία συνεχή αύξηση όσο και η τιμή της κλίσης της ευθείας (στην προκειμένη περίπτωση $b = 2$), ενώ στην περίπτωση της λογιστικής παλινδρόμησης, μετά από μόλις 2 τιμές, η τιμή του Y έχει φτάσει να οριζοντοποιηθεί σε μία τιμή πολύ κοντά στο 1. Αντίστοιχα, για τις τιμές $X = 0, -1, -2$ θα βρίσκαμε:

$$X = 0 \quad Y = \frac{1}{(1 + e^{-(1+2*0)})} \Rightarrow Y = 0.7311$$

$$X = -1 \quad Y = \frac{1}{(1 + e^{-(1+2*-1)})} \Rightarrow Y = 0.2689$$

$$X = -2 \quad Y = \frac{1}{(1 + e^{-(1+2*-2)})} \Rightarrow Y = 0.047426$$

μία τιμή πολύ κοντά στο 0 δηλαδή. Η μεταβολή στην τιμή του Y που έγινε από το 0 στο 1 ($0.7311 \rightarrow 0.9526$), δεν ήταν ίδια με τη μεταβολή που έγινε από το 0 στο -1 ($0.7311 \rightarrow 0.2689$). Αυτό συμβαίνει διότι η μεταβολή του Y κατά την αλλαγή της τιμής του X επηρεάζεται από την εξίσωση, ενώ την ίδια στιγμή η τιμή του 0.7311 όταν $X = 0$ μπορεί να ερμηνευθεί ως η πιθανότητα το Y να είναι 1, όταν το $X = 0$ (Ayyadevara, 2018).

2.4 Νευρωνικά δίκτυα

Ένα νευρωνικό δίκτυο είναι ένα μοντέλο που κατατάσσεται στην επιβλεπόμενη μάθηση και είναι σχεδιασμένο έτσι ώστε να αντιδρά όπως ο εγκέφαλος ενός ανθρώπου, ενώ οι εφαρμογές του επεκτείνονται τόσο σε προβλήματα παλινδρόμησης όσο και ταξινόμησης. Η ανακάλυψη μίας τέτοιας αρχιτεκτονικής προήλθε από το γεγονός ότι η γραμμική και η λογιστική παλινδρόμηση δεν μπορούν να περιγράψουν με επιτυχία όλων των ειδών τα προβλήματα, διότι σε περιπτώσεις όπου η συμπεριφορά των δεδομένων είναι μη γραμμική, χρειάζονται πιο περίπλοκες προσεγγίσεις για την κάθε πρόβλεψη (Ayyadevara, 2018). Η γενική μορφή

ενός νευρωνικού δικτύου παρουσιάζεται στο Σχ. 11. Κατά τη σχεδίαση ενός νευρωνικού δικτύου, τέσσερα στοιχεία είναι εκείνα που χρειάζονται διευκρίνηση (Benardos & Vosniakos, 2007):

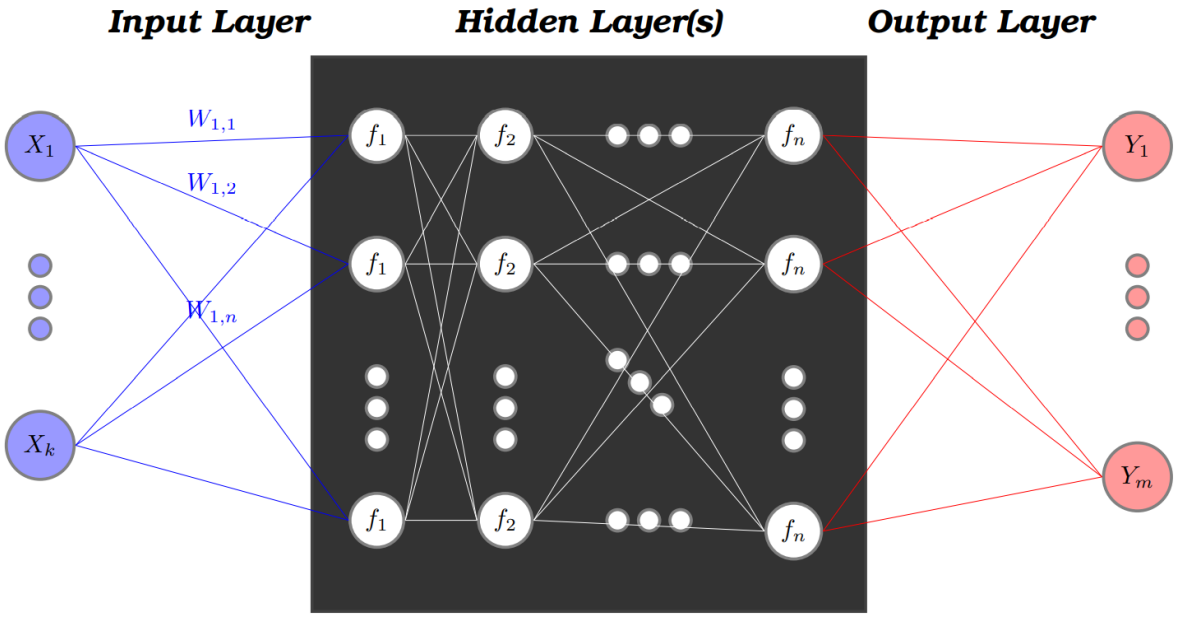
- Ο αριθμός των επιπέδων.
- Ο αριθμός των νευρώνων σε κάθε επίπεδο.
- Η συνάρτηση ενεργοποίησης του κάθε επιπέδου.
- Ο τρόπος εκπαίδευσης του δικτύου.

Στη γενική μορφή ενός νευρωνικού δικτύου, παρατηρούνται τριών ειδών επίπεδα:

1. Επίπεδο εισόδου (Input Layer).
2. Κρυφά επίπεδα (Hidden Layers).
3. Επίπεδο εξόδου (Output Layer).

Στο επίπεδο εισόδου εισάγονται οι ανεξάρτητες μεταβλητές και είναι υπεύθυνο να μεταφέρει τις παραμέτρους εισόδου για να πραγματοποιηθεί η εκπαίδευση του μοντέλου. Το επίπεδο εξόδου είναι εκείνο το οποίο εμφανίζει τα αποτελέσματα ή αλλιώς τις εξαρτημένες μεταβλητές, ενώ τα κρυφά επίπεδα είναι εκείνα τα οποία συνδέουν την είσοδο με την έξοδο, ενώ ταυτόχρονα αξίζει να σημειωθεί ότι δεν είναι ορατά εκτός του δικτύου (Wei et al., 2019). Η αξία των κρυφών επιπέδων φαίνεται κατά τη διαχείριση μιας πολύ περίπλοκης συμπεριφοράς δεδομένων ή μιας συμπεριφοράς που δεν είναι πολύ προφανής, καθώς ο αριθμός τους βοηθά στην ενσωμάτωση μη γραμμικότητας, αποφεύγοντας έτσι αποτελέσματα που βασίζονται σε μία πολύ μεγάλη γραμμική εξίσωση (Ayyadevara, 2018). Το πλήθος των κρυφών επιπέδων σε αρκετά απλές περιπτώσεις, μπορεί να είναι ίσο με ένα, ενώ συνήθως επιλέγεται μέσω της μεθόδου trial and error (Sofos, Stavrogiannis, et al., 2022).

Σε κάθε επίπεδο υπάρχουν κόμβοι (νευρώνες) οι οποίοι είναι συνδεδεμένοι μέσω ακμών για να διαρρέει η πληροφορία, απαρτίζοντας έτσι ένα δίκτυο (Chaabene et al., 2020). Κάθε πληροφορία μέσα στο δίκτυο (όμοια με τη γραμμική/λογιστική παλινδρόμηση), επηρεάζεται από ένα βάρος (weight) το οποίο δηλώνει τη δύναμη ή αλλιώς τη σημαντικότητα που

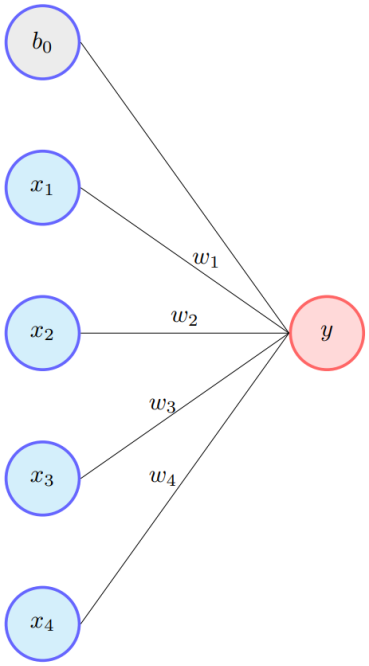


Σχήμα 11: Γενική μορφή ενός νευρωνικού δικτύου (artificial neural network).

έχει μία παράμετρος εισόδου στο αποτέλεσμα και από ένα σταθερό συντελεστή (bias). Το δίκτυο στοχεύει να βελτιστοποιήσει τις παραμέτρους του (βάρη και σταθερούς συντελεστές) για να πετύχει καλύτερες προβλέψεις. Καθώς η πληροφορία μεταφέρεται από το ένα επίπεδο στο άλλο, συναντά τους νευρώνες, όπου ο κάθε νευρώνας συλλέγει όλη την πληροφορία η οποία υπήρχε στο προηγούμενο επίπεδο (βλ. Σχ. 12) και τη συνδυάζει όπως φαίνεται στην Εξ. (2.23). Όταν οι συνδέσεις μεταξύ των νευρώνων δε σχηματίζουν κύκλο, το δίκτυο μεταφέρει την πληροφορία μόνο προς τα εμπρός και ονομάζεται δίκτυο τροφοδοσίας προς τα εμπρός (feed-forward).

Εφόσον συλλεχθεί και συνδυαστεί η πληροφορία, το αποτέλεσμά της χρησιμοποιείται ως δεδομένο σε μία συνάρτηση ενεργοποίησης, της οποίας ο ρόλος είναι να επιλέξει εάν πρέπει ο συγκεκριμένος νευρώνας να ενεργοποιηθεί ή όχι. Η συνάρτηση ενεργοποίησης τροποποιεί την έως τώρα πληροφορία και ανάλογα με το αποτέλεσμά της θα γίνει ενεργοποίηση ή όχι του νευρώνα για να τη μεταφέρει τελικά στο επόμενο επίπεδο (Sharma et al., 2017). Για αυτό το λόγο, ανάλογα το ζητούμενο χρησιμοποιείται και διαφορετικού είδους συνάρτηση ενεργοποίησης, όπως για παράδειγμα η Sigmoid όπως φαίνεται στην Εξ. (2.17) ή η $Tanh$ με μαθηματική σχέση:

$$Tanh = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.21}$$



Σχήμα 12: Συνδυασμός της έως τώρα πληροφορίας σε έναν κόμβο για την τροποποίησή της σε άλλη.

Κάθε νευρώνας μπορεί να περιγραφεί από τους παρακάτω τύπους (Sofos, Stavrogiannis, et al., 2022):

$$y_k = f(u_k + b_k) \tag{2.22}$$

$$u_k = \sum_{i=1}^{N} w_{k_i} x_i \tag{2.23}$$

Οι μεταβλητές $x_1, x_2, \ldots$ είναι οι παράμετροι εισόδου, με $w_1, w_2, \ldots$ αναγράφονται τα βάρη κάθε κόμβου, u_k είναι το γραμμικό αποτέλεσμα του συνδυασμού από τα βάρη, b_k είναι ο σταθερός συντελεστής, f είναι η συνάρτηση ενεργοποίησης και y_k το αποτέλεσμα του νευρώνα. Στο τελευταίο επίπεδο, μέσω μίας μετασχηματικής συνάρτησης (transformation function), εκτιμάται το αποτέλεσμα (Wei et al., 2019).

Για να βγει το αποτέλεσμα από τη συνάρτηση ενεργοποίησης, θα χρειαστεί επίσης και η τιμή του σταθερού συντελεστή (bias). Ο σταθερός συντελεστής μπορεί να θεωρηθεί ως ένα μέσο για να μετρήσει την ευκολία του νευρώνα να ενεργοποιηθεί (Nielsen, 2015). Ο σταθερός συντελεστής, όπως και τα βάρη μαθαίνονται από τον αλγόριθμο κατά τη διαδικασία εκπαίδευσης ενώ αξίζει να κρατήσει κάποιος πως αν ένας νευρώνας έχει ένα πολύ μεγάλο

σταθερό συντελεστή, τότε θα ήταν πολύ πιο εύκολο η συνάρτηση ενεργοποίησης να βγάλει μία τιμή η οποία θα ενεργοποιήσει τελικά τον νευρώνα. Η συνάρτηση ενεργοποίησης βοηθά να εφαρμόσουμε μη-γραμμικότητα και έτσι να είναι σε θέση το μοντέλο μας να εκπαιδευτεί και να κάνει προβλέψεις σε πιο απαιτητικές καταστάσεις (Ayyadevara, 2018). Επίσης, ένα μεγάλο πλήθος κρυφών επιπέδων ενισχύει την εκπαίδευση σε πολύπλοκες συμπεριφορές καθώς επιτυγχάνεται μία υψηλή μη γραμμικότητα.

Ο τρόπος εκπαίδευσης του δικτύου είναι η διαδικασία κατά την οποία ρυθμίζονται τα βάρη και οι σταθεροί συντελεστές, με τις τεχνικές της εμπρόσθιας διάδοσης (forward propagation) και της οπίσθιας διάδοσης (back propagation) να είναι αυτές που χρησιμοποιούνται περισσότερο. Κατά την εμπρόσθια διάδοση, δίνονται τιμές στα βάρη και υπολογίζεται το αποτέλεσμα στο τελικό επίπεδο, ενώ για την πρώτη διάδοση όπου τα βάρη δεν έχουν κάποια τιμή, τους δίνεται μία τυχαία. Συνήθως, η επιλογή που εφαρμόζεται περισσότερο, είναι να τους δοθεί μία τυχαία τιμή μεταξύ του 0 και του 1, χωρίς αυτό να σημαίνει πως κατά το πέρας της εκπαίδευσης η τιμή του βάρους υποχρεωτικά θα είναι σε αυτό το εύρος. Κατά την οπίσθια διάδοση, υπολογίζεται η απόσταση της πρόβλεψης από την πραγματική τιμή (το σφάλμα της δηλαδή) και επαναπροσδιορίζονται τα βάρη ξεκινώντας αυτή τη φορά από το τελευταίο επίπεδο και δουλεύοντας προς τα πίσω, με σκοπό να ελαχιστοποιηθεί το σφάλμα. Η εμπρόσθια διάδοση είναι εκείνη η οποία θα μας βοηθήσει να εκτιμήσουμε το σφάλμα, ενώ η οπίσθια διάδοση είναι εκείνη που θα μας βοηθήσει να ελαχιστοποιήσουμε το σφάλμα (Ayyadevara, 2018).

Από τη στιγμή που θα μεταβληθεί κάποιο βάρος ελάχιστα (είτε προς τα πάνω είτε προς τα κάτω), το μέσο σφάλμα θα αλλάξει και εκείνο (είτε θα αυξηθεί είτε θα μειωθεί). Εάν σε ένα βάρος αυξηθεί η τιμή του και παρατηρηθεί μείωση στο μέσο σφάλμα, τότε επιλέγεται να ακολουθηθεί μία πορεία αύξησης του βάρους ενώ εάν κατά την αύξηση της τιμής του βάρους παρατηρηθεί μία αύξηση στο μέσο σφάλμα, τότε θα προτιμηθεί η αντίθετη πορεία. Ακόμη και αν παρατηρηθούν μικρές αλλαγές στο μέσο σφάλμα, μπορεί να εξακριβωθεί η κατεύθυνση όπου θα κινηθεί η εκπαίδευση (μείωση ή αύξηση της τιμής του βάρους), αλλά και κατά το πόσο (το μέγεθος της τιμής) θα αλλάξει. Το μέγεθος της αλλαγής της τιμής του βάρους θα εκτιμηθεί από την αλλαγή που θα παρατηρηθεί στο μέσο σφάλμα. Εάν κατά την ελάχιστη αλλαγή στο βάρος παρατηρηθεί μεγάλη αλλαγή στο μέσο σφάλμα, τότε μπορεί να

επιλεγεί μία μεγαλύτερη τροποποίηση του βάρους, ενώ αν παρατηρηθεί μικρή αλλαγή, τότε θα προτιμηθεί μία μικρή τροποποίηση της τιμής.

Κατά την ανανέωση των τιμών στα βάρη, πρέπει να αποφασίσουμε το μέγεθος στη μεταβολή των τιμών. Εάν αποφασίσουμε να μην αλλάξουμε όλα τα βάρη με τη μία αλλά να τα ανανεώσουμε με πιο αργό ρυθμό, καταφέρνουμε να έχει περισσότερη σταθερότητα το μοντέλο. Ο ρυθμός με τον οποίο πραγματοποιείται η ανανέωση στα βάρη είναι ο ρυθμός μάθησης.

2.5 Δέντρα απόφασης

Τα δέντρα απόφασης (decision trees) χρησιμοποιούνται για να προβλέψουν ή να ταξινομήσουν ένα γεγονός. Περιέχουν γραφήματα δεντρικής μορφής (με κόμβους και ακμές) όπου κάθε κόμβος αντικατοπτρίζει ένα τεστ στα δεδομένα, ενώ στα κλαδιά παρουσιάζεται το αποτέλεσμα του τεστ (Rahman et al., 2021). Τα χαρακτηριστικά ενός δέντρου απόφασης, περιλαμβάνουν:

- *Αρχικός κόμβος* (root node). Βρίσκεται στη ρίζα του δέντρου και περιλαμβάνει όλο το υπό μελέτη δείγμα.
- *Κόμβος απόφασης* (decision node). Είναι ένας υπο-κόμβος ο οποίος χωρίζεται σε περαιτέρω κόμβους.
- *Τερματικό Κόμβος* (leaf node). Ξεκινώντας από τη ρίζα του δέντρου, απαντώνται οι διάφορες συνθήκες προκειμένου να φτάσει κανείς στον τελικό κόμβο (στα φύλλα του δέντρου) όπου υπάρχει το τελικό αποτέλεσμα και η πρόβλεψη του μοντέλου.
- *Διαχωρισμός* (Splitting). Η διαδικασία κατά την οποία ένας κόμβος χωρίζεται σε υποκόμβους με βάση μία συνθήκη.
- *Κλαδί* (Branch/Sub-tree). Ένα κλάσμα του δέντρου (μία περιοχή του).
- *Γονιός & παιδί κόμβος* (Parent & child node). Όταν ένας κόμβος χωρίζεται παραδείγματος χάριν σε δύο υπο-κόμβους, τότε οι δύο υπο-κόμβοι είναι τα παιδία του γονιού (του αρχικού κόμβου που χωρίστηκε).

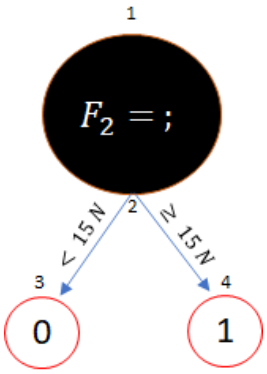
Στο Σχ. 13 φαίνεται η μορφή ενός δέντρου απόφασης. Οι αριθμοί που αναγράφονται πάνω και κάτω από κάθε κόμβο, χρησιμοποιούνται μόνο για αναφορά. Ο κόμβος με τον αριθμό 1, είναι ο αρχικός κόμβος. Ο αριθμός 2 αναφέρεται στον διαχωρισμό που υπέστη ο κόμβος με βάση μία συνθήκη, ενώ οι κόμβοι 3 και 4 είναι οι τερματικοί κόμβοι, που παρουσιάζουν το αποτέλεσμα. Για παράδειγμα, έστω ότι έχουμε ένα σώμα μάζας m, στο οποίο ασκείτε μία σταθερή δύναμη προς τα δεξιά, με μέτρο $F_1 = 15N$. Σε αυτό το σώμα, ασκείται επίσης μία μεταβλητή, δεύτερη δύναμη F_2 με φορά αντίθετη της F_1. Στα δεδομένα του Πίνακα 6, υπάρχουν διάφορες τιμές που αντιστοιχούν στο μέτρο της δύναμης F_2. Στόχος είναι, εφαρμόζοντας το δέντρο απόφασης του Σχ. 13, να ταξινομηθούν οι τιμές της δύναμης F_2 που δίνουν συνισταμένη δύναμη μικρότερη ή ίση του 0.

Ένα σώμα το οποίο βρίσκεται σε ισορροπία, έχει συνισταμένη δυνάμεων ίσο με το μηδέν (πρώτος νόμος του Νεύτωνα) (Freedman et al., 1996):

$$\Sigma F = 0$$

Για λόγους απλότητας, υποθέτουμε ότι δεν ασκείται κατακόρυφη δύναμη στο σώμα. Έτσι, η συνισταμένη δύναμη του Y άξονα είναι ίση με 0, ενώ η συνισταμένη δύναμη του X άξονα εξαρτάται από τις δυνάμεις F_1 και F_2. Για να είναι το σώμα σε ισορροπία (είτε ακίνητο, είτε κινούμενο σε ευθεία έχοντας σταθερή ταχύτητα), εφόσον ασκείται μία σταθερή δύναμη $F_1 = 15N$ στο σώμα μάζας m, πρέπει το μέτρο της μεταβλητής δύναμης F_2 να είναι ίσο με το μέτρο της δύναμης F_1. Στην περίπτωση όπου το μέτρο της δύναμης F_2 δεν είναι ίσο με το μέτρο της δύναμης F_1, τότε το σώμα θα τεθεί σε κίνηση.

Ας πάρουμε θετική κατεύθυνση, τη φορά της δύναμης F_1. Για να είναι η συνθήκη «θετική συνισταμένη» αληθής, πρέπει $F_2 < F_1$. Σε αυτές τις περιπτώσεις, η τιμή που θα εξάγει το μοντέλο του δέντρου απόφασης θα είναι 0. Στην αντίθετη περίπτωση όπου η συνισταμένη των δυνάμεων είναι αρνητική ($F_2 > F_1$) ή σε ισορροπία ($F_2 = F_1$), το μοντέλο θα εξάγει την τιμή 1.



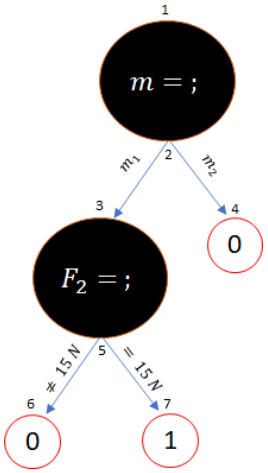
Σχήμα 13: Παράδειγμα δέντρου απόφασης 1.

Η διαδικασία λήψης των αποτελεσμάτων, ξεκινά από τον αρχικό κόμβο και ανάλογα με τις συνθήκες διαχωρισμού, ακολουθείται μία πορεία έως έναν τερματικό κόμβο, όπου εκεί εκτιμάται το αποτέλεσμα. Στο συγκεκριμένο παράδειγμα, η πρώτη και μοναδική συνθήκη που συναντάται, βρίσκεται στον κόμβο με αριθμό 1, όπου για κάθε ένα από τα δεδομένα, γίνεται η ερώτηση για την τιμή του μέτρου της F_2. Έτσι, ανάλογα με την απάντηση θα επιλεχθεί διαφορετική διαδρομή που θα δώσει και διαφορετική απάντηση. Στην τρίτη στήλη του Πίνακα 6, αναγράφεται η πρόβλεψη του μοντέλου για τις τιμές των δεδομένων.

Πίνακας 6: Σετ δεδομένων για παράδειγμα δέντρου απόφασης 1.

	F_2	**Πρόβλεψη δέντρου απόφασης 1**
1	7	0
2	15	1
3	17	1
4	2	0
5	21	1
6	19	1
7	5	0

Ένα άλλο παράδειγμα δέντρου απόφασης, φαίνεται στο Σχ. 14. Οι αριθμοί που εμφανίζονται πάνω και κάτω από κάθε κόμβο χρησιμοποιούνται μόνο για αναφορά. Ο κόμβος με



Σχήμα 14: Παράδειγμα δέντρου απόφασης 2.

αριθμό 1 είναι ο αρχικός κόμβος, ο κόμβος με αριθμό 3 είναι ένας κόμβος απόφασης, οι κόμβοι με αριθμούς 4,6,7 είναι τερματικοί, ενώ με τους αριθμούς 2 και 5 παρατηρείται η διαδικασία διαχωρισμού.

Σε αυτό το παράδειγμα, έστω ότι υπάρχουν δύο σώματα μάζας m_1 και m_2, σε απόσταση τέτοια ώστε το ένα σώμα να μην επηρεάζεται από το άλλο. Επίσης, στα δύο σώματα ασκείται μία σταθερή δύναμη ίση με $F_1 = 15N$ προς τα δεξιά και δύο μεταβλητές δυνάμεις F_2 και F_3 αντίστοιχα προς τα αριστερά, με την ιδιαιτερότητα ότι τα μέτρα τους είναι ίσα (δηλαδή $F_2 = F_3$). Χρησιμοποιώντας ένα δέντρο απόφασης όπως του Σχ. 14, προσπαθούμε να αξιολογήσουμε τις περιπτώσεις όπου το σώμα μάζας m_1 θα παρουσιάσει συνισταμένη δύναμη ίση με το 0. Έτσι, όλες οι περιπτώσεις που οδηγούν σε συνισταμένη δύναμη διάφορη του μηδενός για το σώμα μάζας m_1 θα πάρουν την τιμή 0, ενώ εκείνες που δίνουν αποτέλεσμα ισορροπίας του σώματος μάζας m_1 ($F_2 = F_1$) θα πάρουν την τιμή 1.

Στα δεδομένα του Πίνακα 7, αναγράφονται οι τιμές δυνάμεων F_2 και F_3, το σώμα μάζας στο οποίο έχει καταγραφεί η μέτρηση, καθώς και η πρόβλεψη του δέντρου απόφασης. Το παρόν πρόβλημα θα μπορούσε να χωριστεί σε δύο σετ δεδομένων όπου το κάθε ένα θα αφορούσε το δικό του σώμα και η διαδικασία επίλυσης να ήταν παρόμοια με του πρώτου δέντρου απόφασης. Ο λόγος που επιλέχθηκε αυτός ο τρόπος επίλυσης, είναι για την περιγραφή των εννοιών που έχουν να κάνουν με το δέντρο απόφασης.

Η διαδικασία ξεκινά από τον αρχικό κόμβο, όπου απαντάται η πρώτη απόφαση. Στη συγκεκριμένη περίπτωση η πρώτη απόφαση ταξινομεί τις περιπτώσεις σε εκείνες που

αφορούν το σώμα μάζας m_1 και σε εκείνες που αφορούν το σώμα μάζας m_2. Όπως αναφέραμε προηγουμένως, στόχος είναι να βρεθούν οι περιπτώσεις όπου το σώμα μάζας m_1 θα βρίσκεται σε ισορροπία. Συνεπώς, όλες οι καταστάσεις που αφορούν το σώμα μάζας m_2 και όλες οι καταστάσεις όπου θα επιφέρουν θετική συνισταμένη δυνάμεων για το σώμα μάζας m_1 θα πάρουν τιμή 0 . Αντίθετα, οι τιμές που θα σταθεροποιήσουν το σώμα μάζας m_1 θα πάρουν την τιμή 1. Η διαφορά από το προηγούμενο παράδειγμα δέντρου απόφασης, είναι πως ως απάντηση σε μία συνθήκη μπορεί να δοθεί μία άλλη συνθήκη. Η συνθήκη του κόμβου με αριθμό 3, χωρίζει τα δεδομένα σε εκείνα όπου η τιμή είναι διάφορη του 15 ή ίση με 15. Άρα, η μοναδική περίπτωση που λαμβάνει τιμή 1, είναι η περίπτωση με αριθμό 2.

Πίνακας 7: Σετ δεδομένων για παράδειγμα δέντρου απόφασης 2.

	Σώμα μάζας	$F_2 = F_3$	**Πρόβλεψη δέντρου απόφασης 2**
1	m_1	7	0
2	m_1	15	1
3	m_2	17	0
4	m_1	2	0
5	m_2	21	0
6	m_1	19	0
7	m_2	15	0

Κατά το σχεδιασμό ενός δέντρου απόφασης, συχνά εμφανίζονται θέματα υπερπροσαρμογής (Ayyadevara, 2018), καθώς αν για παράδειγμα δημιουργήσουμε πάρα πολλά επίπεδα (στο δεύτερο παράδειγμα του Σχ. 14 έχουμε 2 επίπεδα), το δέντρο θα έχει τόσα πολλά κλαδιά που θα υπερπροσαρμοστεί στο σετ δεδομένων και δε θα είναι καθόλου εύκολο να γενικευτεί η λύση. Επίσης, εάν κατασκευαστεί ο αλγόριθμος με τρόπο ώστε να καταλήγουν ελάχιστες παρατηρήσεις σε ένα τερματικό κόμβο, το δέντρο μας είναι αρκετά πιθανό να είναι μεγάλο με αρκετά κλαδιά, το οποίο με τη σειρά θα του είναι αρκετά πιθανό να οδηγήσει σε υπερπροσαρμογή. Τέλος, εάν σε έναν κόμβο απόφασης παρατηρηθεί μικρός αριθμός παρατηρήσεων που θα εξεταστούν, όμοια με προηγουμένως θα καταλήξουμε να έχουμε ελάχιστες παρατηρήσεις σε τερματικό κόμβο. Στην τελευταία περίπτωση, το πρόβλημα δημιουργείται

έμμεσα και όχι άμεσα. Από την άλλη μεριά, τακτικές αποφυγής υπερπροσαρμογής στα δέντρα αποφάσεων, μεταξύ άλλων περιλαμβάνουν (Ayyadevara, 2018):

- Να τεθεί περιορισμός ως προς τον ελάχιστο αριθμό παρατηρήσεων που θα καταλήξουν σε έναν τερματικό κόμβο (π.χ., έναν μεγάλο αριθμό).
- Να οριστεί ένα μέγιστο μήκος στα επίπεδα που θα δημιουργηθούν στο δέντρο.
- Όμοια με το πρώτο, να τεθεί περιορισμός ως προς τον ελάχιστο αριθμό παρατηρήσεων που θα καταλήξουν σε έναν κόμβο απόφασης.

Για το λόγο ότι τα δέντρα απόφασης συχνά υπερπροσαρμόζονται, επιλέγονται κάποιες παραλλαγές τους, όπως ο αλγόριθμος Random Forest. Ο αλγόριθμος random forest αξιοποιεί πολλά δέντρα αποφάσεων, όπου το κάθε δέντρο απόφασης εστιάζει σε ένα δείγμα των δεδομένων (Ayyadevara, 2018). Στον αλγόριθμο random forest, η επιλογή του δείγματος των δεδομένων γίνεται στην τύχη (random) ενώ η ανάλυσή του γίνεται μέσω πολλαπλών δέντρων απόφασης (δημιουργία δάσους – forest). Ο τρόπος λειτουργίας του αλγόριθμου random forest, είναι να ερευνά τυχαία υποσύνολα του σετ δεδομένων και ως απάντηση να δίνει το μέσο όρο των προβλέψεων. Η μέθοδος αυτή ονομάζεται Bagging ή αλλιώς bootstrap aggregating.

Η εξαγωγή των χαρακτηριστικών (feature extraction), είναι ένας τρόπος μετατροπής δεδομένων σε αριθμητικά χαρακτηριστικά (διατηρώντας την πληροφορία του αρχικού σετ δεδομένων) έτσι ώστε να επεξεργαστούν από ένα μοντέλο μηχανικής μάθησης, γεγονός που έχει αποδειχθεί πιο αποτελεσματικό έναντι μίας απλής εφαρμογής του μοντέλου μηχανικής μάθησης σε ένα σετ δεδομένων (MathWorks, n.d.). Αλγόριθμοι όπως τα νευρωνικά δίκτυα ή τα δέντρα απόφασης, για το λόγο ότι επεξεργάζονται την πληροφορία σε ένα μικρό πλήθος κόμβων, χαρακτηρίζονται ως ρηχή μάθηση. Όπως αναφέρουν οι Wei et al. (Wei et al., 2019), ενώ η εφαρμογή τους είναι επιτυχής σε διάφορους τομείς της επιστήμης των υλικών, αρκετές φορές οι προσεγγίσεις Συναρτησιακής Θεωρίας Πυκνότητας (Density Functional Theory) προσδίδουν μεγαλύτερη ακρίβεια. Επιπρόσθετα αναφέρουν ότι τέτοιου είδους αλγόριθμοι βασίζονται σε χειροκίνητη εξαγωγή χαρακτηριστικών, απαιτούν ερευνητές με γνώση του συστήματος προκειμένου να κατασκευάσουν κατάλληλες αναπαραστάσεις για τα δεδομένα

εισόδου και αυτό οδηγεί στη μείωση της ακρίβειας του μοντέλου. Ως συνέπεια, συζητήσεις ήρθαν στο φως για τη δημιουργία πιο ισχυρών μεθόδων, όπως η βαθιά μάθηση.

2.6 Βαθιά Μάθηση

Σε αντίθεση με τους αλγόριθμους ρηχής μάθησης, η βαθιά μάθηση προωθεί την πληροφορία σε πολλαπλά επίπεδα για την απόκτηση καλύτερης γνώσης του συστήματος και την εξαγωγή προβλέψεων (Guo et al., 2021). Τα μοντέλα βαθιάς μάθησης αξιοποιούν έννοιες της γραμμικής άλγεβρας, της θεωρίας πιθανοτήτων αλλά και προγραμματιστικών τεχνικών σε έναν αριθμό κρυφών επιπέδων ο οποίος μπορεί να ξεπεράσει τους χίλιους (Dimiduk et al., 2018). Σύμφωνα με τους Wei et al., η εφαρμογή μοντέλων ρηχής μάθησης σε καταστάσεις μη γραμμικής ταξινόμησης δεν είναι ικανοποιητική. Αντίθετα, τα μοντέλα βαθιάς μάθησης χρησιμοποιούν μηχανισμούς αυτόματης εξαγωγής χαρακτηριστικών και κατασκευάζουν χαρακτηριστικά χαμηλού επιπέδου, προκειμένου να δημιουργηθούν πιο περίπλοκες αναπαραστάσεις υψηλού επιπέδου από τις κατηγορίες χαρακτηριστικών και τα ξεπερνούν σε επίδοση.

Ένας τύπος δικτύου βαθιάς μάθησης είναι το βαθή συνελικτικό νευρωνικό δίκτυο (deep convolutional neural network (Krizhevsky et al., 2017)). Το συνελικτικό νευρωνικό δίκτυο είναι ένα νευρωνικό δίκτυο εμπρόσθιας τροφοδοσίας, το οποίο συνδυάζει το δίκτυο με διακριτή συνέλιξη για επεξεργασία εικόνων και δίνει τη δυνατότητα τοποθέτησης μίας εικόνας ως είσοδο στο δίκτυο, ελαχιστοποιώντας έτσι την ανάγκη για συνεχείς μετατροπές στο είδος των δεδομένων, όπως γίνεται συνήθως στους αλγόριθμους αναγνώρισης εικόνας (Wei et al., 2019).

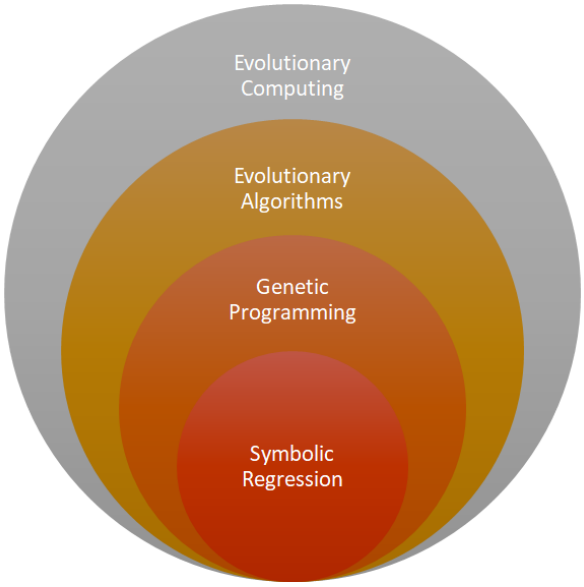
Από τη μία μεριά, τα δίκτυα βαθιάς μάθησης (deep neural networks) προσδίδουν δυνατότητες οι οποίες δεν εμφανίζονται εξ ολοκλήρου στα απλά νευρωνικά δίκτυα. Από την άλλη μεριά, ο μεγάλος χρόνος εκπαίδευσης των δικτύων βαθιάς μάθησης αλλά και το γεγονός ότι τα αποτελέσματά τους δε συμβαδίζουν με την αποτύπωση φυσικών εννοιών είχε ως

αποτέλεσμα να ξεπεραστεί η απόδοση τους σε ορισμένες περιπτώσεις από αλγόριθμους Ρηχής Γνώσης (Wei et al., 2019).

Κεφάλαιο 3 Συμβολική Παλινδρόμηση

Η συμβολική παλινδρόμηση είναι η εξαγωγή μίας μαθηματικής εξίσωσης η οποία είναι ικανή να περιγράψει ένα σετ δεδομένων. Σε αντίθεση με τις κλασικές μορφές παλινδρόμησης (π.χ., γραμμική), στη συμβολική παλινδρόμηση δεν είναι γνωστή ούτε η μορφή, ούτε οι παράγοντες της εξίσωσης. Έτσι, αντί να μεταβάλλονται οι σταθερές για να επιτευχθεί μια καλή συμφωνία της εξίσωσης με τα δεδομένα, γίνεται μία προσπάθεια να βρεθεί και η μορφή της εξίσωσης και οι παράμετροί της ταυτόχρονα (Tohme et al., 2022). Το γεγονός ότι η συμβολική παλινδρόμηση ενεργεί με ελάχιστους (έως καθόλου) περιορισμούς, θέτει βάσεις για μία αντικατάσταση των μοντέλων μαύρου κουτιού με άλλες που ενεργούν αποκλειστικά με βάση τα δεδομένα και προσφέρουν αποτελέσματα που μπορούν να αποκαλύψουν τη φυσική σημασία του μηχανισμού τους (Papastamatiou et al., 2022).

Η προσέγγιση της συμβολικής παλινδρόμησης μπορεί να επιτευχθεί από διάφορους δρόμους (Angelis et al., n.d.), αλλά συνήθως υλοποιείται από τον γενετικό προγραμματισμό. Ο Γενετικός Προγραμματισμός (Genetic Programming) βρίσκεται σε έναν τομέα των Εξελικτικών Αλγόριθμων (Evolutionary Algorithms), οι οποίοι με τη σειρά τους κατατάσσονται στην Εξελικτική Υπολογιστική (Evolutionary Computing), όπου αποτελεί ένα υποσύνολο της τεχνητής νοημοσύνης (βλ. Σχ. 15). Στο Κεφάλαιο αυτό, θα γίνει εισαγωγή στο είδος του προβλήματος προς επίλυση, θα παρουσιαστούν τα σύνολα της εξελικτικής υπολογιστικής, εξελικτικών αλγόριθμων, γενετικού προγραμματισμού, θα αναλυθεί η διαδικασία υλοποίησης της συμβολικής παλινδρόμησης μέσω γενετικού προγραμματισμού αλλά και μέσω ενός μη εξελικτικού αλγόριθμου. Τέλος, θα παρουσιαστούν διάφορα χαρακτηριστικά που την έχουν κάνει να ξεχωρίσει στις φυσικές επιστήμες, καθώς και τα υπέρ/κατά της χρήσης της συγκεκριμένης τεχνικής.

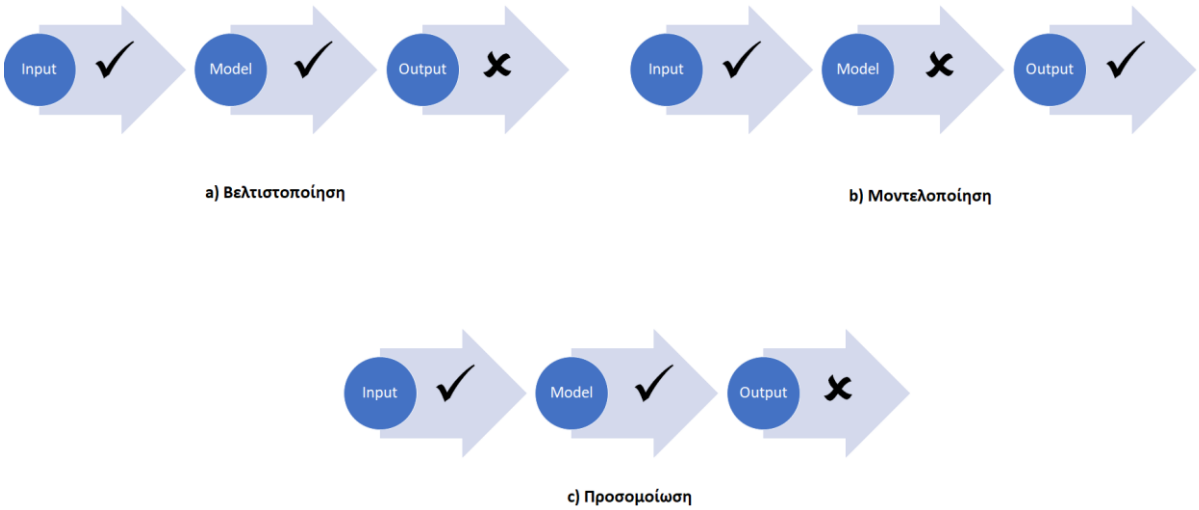


Σχήμα 15: Η θέση της Συμβολικής Παλινδρόμησης στο υπερσύνολο της Εξελικτικής Υπολογιστικής.

3.1 Διατύπωση προβλήματος

Κατά την επίλυση διάφορων προβλημάτων, χρειάζεται η εισαγωγή χαρακτήρων, ένα μοντέλο ικανό να τα περιγράψει και το αποτέλεσμα της επεξεργασίας (η πρόβλεψη δηλαδή). Υπάρχουν διάφορα είδη προβλημάτων τα οποία επιλύονται με διαφορετικούς τρόπους, για παράδειγμα:

- **Βελτιστοποίησης** (Optimization). Σε προβλήματα τέτοιου είδους, είναι γνωστό το μοντέλο που περιγράφει ένα φαινόμενο και το αποτέλεσμα, ενώ γίνονται ενέργειες να βρεθεί η είσοδος στο μοντέλο (βλ. Σχ. 16a).
- **Μοντελοποίησης** (Modelling). Σε αυτά τα προβλήματα είναι γνωστή η είσοδος και η έξοδος του μοντέλου, ενώ ψάχνεται το μοντέλο το οποίο θα δώσει το σωστό αποτέλεσμα για την κάθε είσοδο (βλ. Σχ. 16b).
- **Προσομοίωσης** (Simulation). Στα προβλήματα προσομοιώσεων γνωρίζονται οι είσοδοι και το μοντέλο, ενώ απλά με την εφαρμογή των εισόδων στο μοντέλο προσεγγίζεται το αποτέλεσμα (βλ. Σχ. 16c).



Σχήμα 16: Γνωστά και άγνωστα χαρακτηριστικά προβλημάτων.

Η συμβολική παλινδρόμηση όπως προαναφέρθηκε, βρίσκει μία έκφραση που περιγράφει τα δεδομένα. Άρα, εύκολα παρατηρεί κανείς ότι είναι πρόβλημα μοντελοποίησης. Τα προβλήματα μοντελοποίησης από την άλλη, μπορούν να μετατραπούν σε προβλήματα βελτιστοποίησης εάν οριστεί το σφάλμα που προκύπτει από το μοντέλο, ως μία ποσότητα η οποία πρέπει να ελαχιστοποιηθεί (Eiben et al., 2003). Το πρώτο βήμα για να βρεθεί το μοντέλο, είναι ο ορισμός τη μορφής του μοντέλου, όπως για παράδειγμα ένας κώδικας Java, ένα δέντρο απόφασης ή ένα νευρωνικό δίκτυο μεταξύ άλλων (Eiben et al., 2003).

Εφόσον, επιλεγεί το εργαλείο αποτύπωσης της λύσης, δημιουργείται ένα σύνολο από όλες τις πιθανές λύσεις του προβλήματος και αναζητούνται μέσα σε αυτό στοιχεία τα οποία χαρτογραφούν επιτυχώς τα δεδομένα. Στο πρόβλημα βελτιστοποίησης, αναζητείται το στοιχείο (μέσα σε αυτό το σύνολο) το οποίο έχει περιγράψει τις περισσότερες καταστάσεις σωστά. Έτσι, η λύση του προβλήματος βελτιστοποίησης μπορεί να χρησιμοποιηθεί ως λύση του προβλήματος μοντελοποίησης. Άρα, για να επιλυθεί ένα πρόβλημα μοντελοποίησης, πρέπει να βρεθεί η λύση μέσα από το σύνολο των λύσεων. Αυτό το σύνολο των λύσεων, αναφέρεται ως χώρος αναζήτησης (search space), ο οποίος είναι τεράστιος. Ανάλογα με τη φύση του προβλήματος, ο χώρος αναζήτησης περιέχει όλες τις πιθανές λύσεις του υπό μελέτη προβλήματος, όπου σε περιπτώσεις βελτιστοποίησης θα είναι η είσοδος στο μοντέλο, ενώ σε περιπτώσεις μοντελοποίησης θα είναι διάφορα μοντέλα.

Σε δεύτερο επίπεδο, χρειάζεται ο ορισμός της λύσης του προβλήματος. Στα προβλήματα μοντελοποίησης, ως λύση ορίζεται η ποσότητα η οποία παράγει τα σωστά αποτελέσματα, για κάθε είσοδο που της έχει δοθεί. Κατά την εφαρμογή μίας τέτοιας διαδικασίας όμως, αναζητείται το μοντέλο το οποίο δίνει τα περισσότερες σωστές απαντήσεις (αν όχι όλες). Άρα, από πρόβλημα μοντελοποίησης, μέσω αυτής της τεχνικής μπορεί να μετατραπεί σε πρόβλημα βελτιστοποίησης.

3.2 Εξελικτική Υπολογιστική

Η εξελικτική υπολογιστική (Eiben et al., 2003) είναι ένας κλάδος της πληροφορικής οποίος βασίζεται σε στοιχεία από τη θεωρία εξέλιξης του Δαρβίνου. Πιο συγκεκριμένα, έστω ότι έχουμε ένα πληθυσμό σε ένα περιβάλλον που αποτελείται από πολλά και διάφορα άτομα, τα οποία με τη σειρά τους πρέπει να επιβιώσουν και να αναπαραχθούν. Το πόσο καλά ταιριάζουν τα άτομα αξιολογείται από το πόσο καλά έχουν επιτύχει τους σκοπούς τους. Αυτό με τη σειρά του -δηλαδή το πόσο άξια είναι τα άτομα- χρησιμοποιείται ως συνθήκη που θα τους επιτρέψει να επιβιώσουν και να αναπαραχθούν σε ένα νέο πληθυσμό. Όσο πιο άξιο είναι ένα άτομο τόσο πιο πολλές πιθανότητες έχει να συνεχιστεί το "DNA" του.

Η γενική εικόνα που επικρατεί στην εξελικτική υπολογιστική είναι ότι υπάρχει ένας πληθυσμός και μέσα σε αυτόν τον πληθυσμό υπάρχει ένας αριθμός ατόμων. Κάθε ένα από τα άτομα αξιολογείται από το πόσο καλά μπορεί να προσαρμοστεί στο εκάστοτε περιβάλλον και εάν αυτό το άτομο αξιολογηθεί σχετικά καλά έναντι του υπόλοιπου πληθυσμού, επιλέγεται να αναπαραχθεί. Καθώς επιλέγονται και αναπαράγονται τα άξια άτομα, θα δημιουργηθούν μεταλλάξεις τους, οι οποίες θα δημιουργήσουν ένα νέο πληθυσμό που θα αποτελείται από νέα άτομα, τα οποία νέα άτομα με τη σειρά τους πρέπει να αξιολογηθούν.

3.3 Εξελικτικοί Αλγόριθμοι

Οι εξελικτικοί αλγόριθμοι (Back, 1996) προσπαθούν να υλοποιήσουν την επιλογή και την αναπαραγωγή άξιων ατόμων σε έναν πληθυσμό. Αυτό με τη σειρά του θα φέρει αύξηση της ποιότητας που παρατηρείται στα άτομα. Άρα, αυτό που πραγματοποιείται είναι η δημιουργία ενός τυχαίου πληθυσμού ο οποίος θα εμπεριέχει τις πιθανές λύσεις του προβλήματος και στον οποίο θα εφαρμοστεί ένα μέτρο σύγκρισης για να βρεθούν εκείνες που παρουσιάζουν καλύτερο ταίριασμα στα δεδομένα. Αφού εντοπιστούν οι πιο υποσχόμενες λύσεις, χρησιμοποιούνται για τη δημιουργία ενός νέου πληθυσμού μέσω των διαδικασιών του ανασυνδυασμού (recombination - crossover) και της μετάλλαξης (mutation). Κατά τη διαδικασία του ανασυνδυασμού, επιλέγονται δύο ή περισσότερα άτομα τα οποία χαρακτηρίζονται ως γονείς (parents) και παράγουν ένα ή περισσότερα άτομα τα οποία χαρακτηρίζονται ως παιδιά (children). Η διαδικασία της μετάλλαξης είναι σχεδόν ίδια με αυτή του ανασυνδυασμού, με τη διαφορά ότι πραγματοποιείται σε ένα άτομο και συμβάλει στη δημιουργία ενός άλλου ατόμου.

Εφόσον πραγματοποιηθούν οι διαδικασίες του ανασυνδυασμού και της μετάλλαξης, έχει δημιουργηθεί ένας νέος πληθυσμός από υποσχόμενες λύσεις, στον οποίο πρέπει να αξιολογηθούν τα άτομά του. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να βρεθεί ένα άτομο, του οποίου το ταίριασμα στα δεδομένα είναι ικανοποιητικό (αν όχι τέλειο) ή μέχρι να ορίσει ο κάθε ερευνητής ένα όριο στο αριθμό των πληθυσμών που θα δημιουργηθούν. Με το να επαναλαμβάνονται οι διαδικασίες της επιλογής, ανασυνδυασμού και μετάλλαξης, προσεγγίζεται η βέλτιστη λύση καθώς το ταίριασμα των λύσεων θα είναι όλο και μεγαλύτερο στον κάθε πληθυσμό που θα δημιουργείται.

Αξίζει να σημειωθεί πως η όλη διαδικασία βασίζεται στην τυχαιότητα (στοχαστική). Για παράδειγμα, μπορεί να γίνεται η αξιολόγηση των ατόμων με βάση ένα χαρακτηριστικό (το ταίριασμα στο περιβάλλον), αλλά η επιλογή των υποσχόμενων ατόμων γίνεται στην τύχη. Αυτό με τη σειρά του, προσδίδει μία πιθανότητα ακόμη και σε άτομα με μικρότερη αξία να επιλεγούν για να αναπαραχθούν. Επίσης, κατά τις διαδικασίες του ανασυνδυασμού και της μετάλλαξης, τα μέρη τα οποία συνδυάζονται από τους γονείς για να δημιουργηθούν τα παιδιά επιλέγονται στην τύχη.

Υπάρχουν διάφορα είδη εξελικτικών αλγόριθμων που υλοποιούν την παραπάνω διαδικασία, με αυτά του γενετικού προγραμματισμού, γενετικών αλγορίθμων (Genetic Algorithms) και εξελικτικού προγραμματισμού (Evolutionary Programming) να είναι αυτά που χρησιμοποιούνται περισσότερο. Ανάλογα το ζητούμενο, χρησιμοποιείται διαφορετικού είδους αλγόριθμος, ενώ οι διαφορές στο κάθε είδος παρατηρούνται στον τρόπο που κωδικοποιούν τη διαδικασία (τι μορφή θα της δώσουν) και σε διαφορετικές επιλογές του τρόπου που θα πραγματοποιηθεί ο ανασυνδυασμός, η μετάλλαξη, η συνάρτηση που θα αξιολογήσει τα δεδομένα και η επιλογή των ατόμων μεταξύ άλλων.

Συνοψίζοντας, τα πιο βασικά χαρακτηριστικά των εξελικτικών αλγόριθμων, είναι (Eiben et al., 2003):

- *Αναπαράσταση* (Representation). Ο τρόπος που θα κωδικοποιήσουν τη διαδικασία (π.χ., δεντρική μορφή (tree structure), δυαδική (binary), κ.α.), όπου η κάθε μορφή χρησιμοποιείται για διαφορετικού είδους σκοπό.
- *Συνάρτηση αξιολόγησης* (Evaluation function-Fitness function). Ο τρόπος που θα εξεταστεί το ταίριασμα των ατόμων στο περιβάλλον.
- *Πληθυσμός* (Population). Το σύνολο των ατόμων.
- *Τρόπος επιλογής γονέα* (Parent selection).
- *Παραλλαγή των ατόμων* (Variation operators). Μηχανισμοί ανασυνδυασμού και μετάλλαξης.
- *Επιλογή επιζώντων* (Survivor selection).

Σε κάθε είδος εξελικτικού αλγόριθμου χρησιμοποιούνται διαφορετικοί μηχανισμοί αυτών των χαρακτηριστικών, ενώ ιδιαίτερη προσοχή χρειάζεται σε δύο σημεία (Eiben et al., 2003):

1. Οι διαδικασίες του ανασυνδυασμού και της μετάλλαξης να έχουν την ίδια αναπαράσταση με αυτή του προβλήματος για να είναι εφικτές.
2. Η διαδικασία της επιλογής δεν απαιτεί τον παραπάνω περιορισμό καθώς δρα αποκλειστικά στο μέτρο που θα χρησιμοποιηθεί για το ταίριασμα.

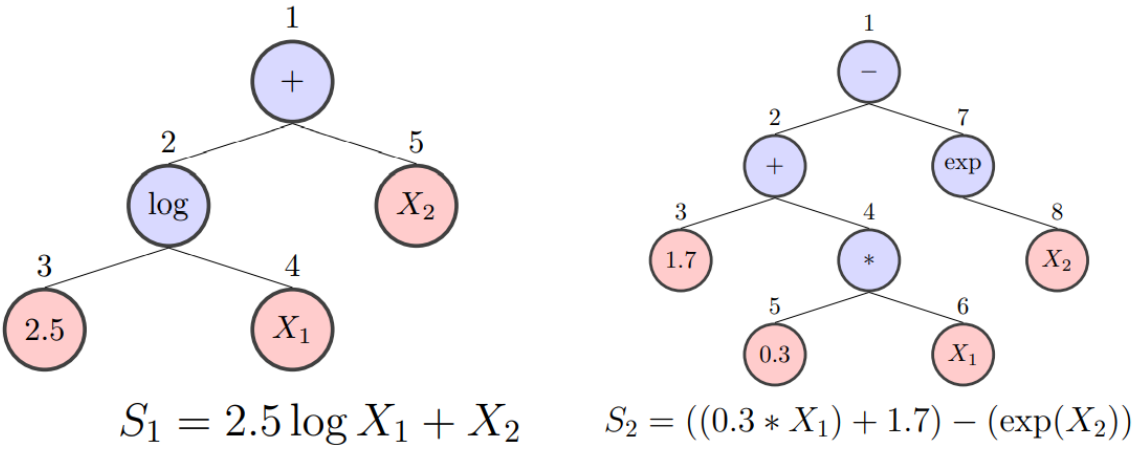
3.4 Γενετικός Προγραμματισμός

Ο γενετικός προγραμματισμός (Koza, 1994) μεταφράζει τα άτομα που υπάρχουν σε έναν πληθυσμό μέσω μίας δεντρικής μορφής. Κάποια εργαλεία που μπορούν να το υλοποιήσουν προγραμματιστικά είναι το πακέτο Glyph στην Python (Quade et al., 2018) ή το GPTIPS toolbox στη MatLab (Searson et al., 2010). Στη δεντρική μορφή αυτή, υπάρχουν ιεραρχικές συνθέσεις μεταξύ αρχικών συναρτήσεων (primitive functions) και τερματικών (terminals), τα οποία ανάλογα με τη φύση του προβλήματος που μελετάται, παίρνουν διαφορετική μορφή. Οι αρχικές συναρτήσεις, συνήθως είναι μαθηματικές συναρτήσεις ή λογικές συναρτήσεις μεταξύ άλλων, ενώ στα τερματικά χρησιμοποιούνται οι μεταβλητές εισόδου του εκάστοτε προβλήματος ή κάποια σταθερά τιμή.

Προκειμένου να γίνει εξαγωγή μίας εξίσωσης ικανή να περιγράψει τη συμπεριφορά των δεδομένων, ο γενετικός προγραμματισμός ψάχνει έναν πολύ μεγάλο χώρο από μαθηματικές εξισώσεις σε μία εξελικτική διαδικασία. Στην εξελικτική αυτή διαδικασία, δημιουργούνται διάφορες συμβολικές εκφράσεις (symbolic expressions) και συγκρίνονται. Οι εκφράσεις με χαμηλή ακρίβεια ή μεγάλη πολυπλοκότητα εγκαταλείπονται, ενώ οι υποσχόμενες εκφράσεις συνδυάζονται έτσι ώστε κάνοντας ένα βήμα τη φορά να χτίζεται μία πιο ακριβής τελική έκφραση.

Στα δένδρα απόφασης, υπήρχαν διάφοροι κόμβοι και ακμές όπου σε κάθε κόμβο, πλην του τερματικού, υπήρχε μία συνθήκη και ανάλογα με την απάντηση, ακολουθείται διαφορετική διαδρομή. Η δεντρική μορφή του γενετικού προγραμματισμού από την άλλη, αποτελείται από αρχικές συναρτήσεις και τερματικούς κόμβους. Αντί για συνθήκη που πρέπει να απαντηθεί στους κόμβους του δέντρου (πλην των τερματικών), υπάρχουν μαθηματικά σύμβολα ή συναρτήσεις (π.χ. , $+$, $-$,$\div$, exp, κ.α.), ενώ στους τερματικούς κόμβους υπάρχουν είτε σταθερές (π.χ., 0.1) είτε κάποια από τις μεταβλητές εισόδου (π.χ., X_1). Ο συνδυασμός των αρχικών συναρτήσεων και τερματικών κόμβων, δίνει την τελική συμβολική εξίσωση σε δεντρική μορφή όπως φαίνεται για παράδειγμα στο Σχ. 17 για τη συμβολική έκφραση S_1.

Οι αριθμοί που φαίνονται πάνω από κάθε κόμβο χρησιμοποιούνται μόνο για αναφορά και δεν έχουν καμία σημασία για την υλοποίηση της διαδικασίας. Οι κόμβοι με



Σχήμα 17: Παραδείγματα συμβολικών εκφράσεων S_1 και S_2.

αριθμούς 3,4 και 5 είναι οι τερματικοί κόμβοι όπου περιέχουν μία σταθερά (2.5) και δύο παραμέτρους εισόδου (X_1, X_2), ενώ οι κόμβοι με αριθμό 1 και 2 είναι οι αρχικές συναρτήσεις οι οποίες περιέχουν βασικά μαθηματικά σύμβολα και συναρτήσεις ($+$, log). Οι διάφοροι συνδυασμοί των τερματικών κόμβων και των αρχικών συναρτήσεων (συμβολική έκφραση) είναι τα άτομα, από έναν πληθυσμό ο οποίος έχει πολλές συμβολικές εκφράσεις. Στόχος του γενετικού προγραμματισμού είναι να συγκρίνει όλες τις εκφράσεις και να εντοπίσει εκείνες που παρουσιάζουν καλύτερο ταίριασμα στα δεδομένα, έτσι ώστε να τις συνδυάσει, μέχρι να βρει μία πολύ καλή έκφραση.

3.5 Διαδικασία μέσω Γενετικού Προγραμματισμού

Ο γενετικός προγραμματισμός ξεκινάει δημιουργώντας τυχαία έναν πληθυσμό από συμβολικές εκφράσεις και τις συγκρίνει, για να συνδυάσει τα υποσχόμενα μέρη και να δημιουργήσει μία τελική εξίσωση η οποία θα έχει ένα καλό ταίριασμα στα δεδομένα. Ο αριθμός των συμβολικών εκφράσεων που θα δημιουργηθούν τυχαία, είναι προκαθορισμένος από τον εκάστοτε χρήστη. Η εύρεση των κατάλληλων μερών είναι αρκετά περίπλοκη και χρονοβόρα διαδικασία, καθώς η σύγκριση γίνεται σε ένα φαινομενικά άπειρο χώρο που περιέχει όλα τα μαθηματικά σύμβολα και σταθερές. Για αυτό το λόγο, ανάλογα τη φύση του προβλήματος που μελετάται, χρειάζεται και το κατάλληλο μέτρο σύγκρισης για την εύρεση του βέλτιστου

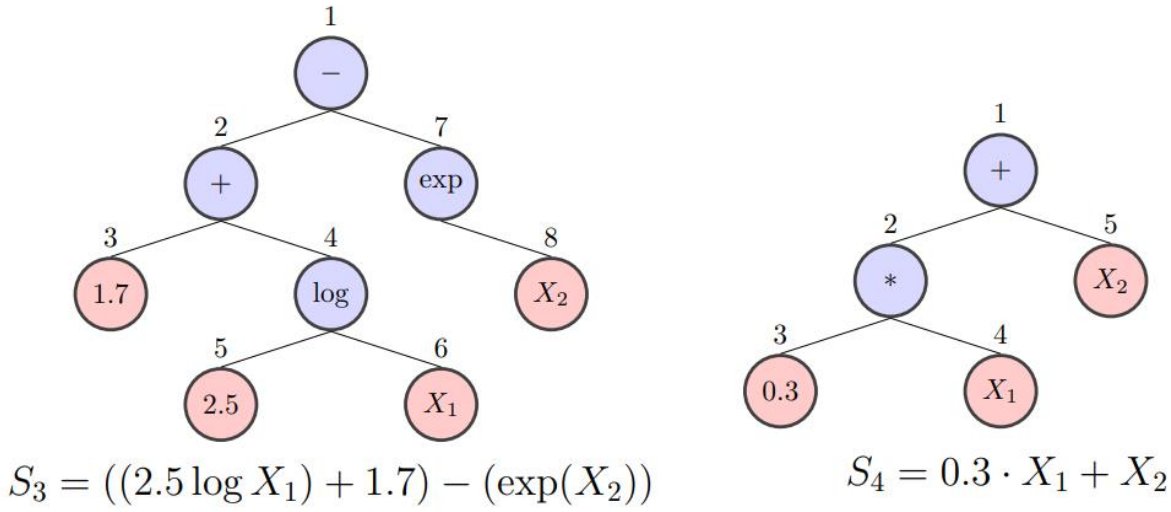
ταιριάσματος. Κατά τη διαδικασία εύρεσης εξίσωσης, η αποτελεσματικότητα εκτιμάται από τον αριθμό των δεδομένων που χειρίζονται σωστά και τα μέρη των εκφράσεων βαθμολογούνται ανάλογα με το σφάλμα που θα εμφανίσουν. Τα μέτρα σφάλματος που χρησιμοποιούνται συνήθως είναι το άθροισμα τετραγώνων των σφαλμάτων (Sum of the Squares Error, Εξ. (2.10)) ή το μέσο τετραγωνικό σφάλμα (Mean Squared Error) με μαθηματική εξίσωση:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{3.1}$$

όπου εδώ, όπως και στο SSE της Εξ. (2.10), n αντιστοιχεί στον αριθμό των καταστάσεων που έχουν ελεγχθεί, i είναι η ατομική κατάσταση, $\hat{Y}$ είναι η τιμή που προβλέφθηκε από το μέρος της έκφρασης και Y είναι η σωστή απάντηση.

Ο έλεγχος και η σύγκριση των εκφράσεων όπως αναφέρθηκε προηγουμένως γίνεται σε ένα πολύ μεγάλο μαθηματικό χώρο, ενώ το γεγονός ότι οι εκφράσεις δημιουργούνται τυχαία, ενισχύει τις εικασίες ότι οι αρχικές εκφράσεις θα συνεισφέρουν έως ένα μικρό ποσοστό στην τελική εξίσωση καθώς το σφάλμα που θα παρουσιάζουν είναι αρκετά μεγάλο. Όμως, κάποιες εκφράσεις θα εμφανίσουν καλύτερο (ίσως ελάχιστα καλύτερο) ταίριασμα στα δεδομένα. Αυτή τη διαφορά είναι που εκμεταλλεύεται και αξιοποιεί συνεχώς ο γενετικός προγραμματισμός, με το να συνδυάζει τις "καλές" εκφράσεις και να διαγράφει τις "κακές". Από τη στιγμή που μία συμβολική έκφραση αξιολογηθεί κατάλληλα ώστε να "επιζήσει", ο συνδυασμός της με μία άλλη γίνεται τυχαία. Για το λόγο ότι γίνεται τυχαία, υπάρχει περίπτωση η νέα εξίσωση που θα δημιουργηθεί να διαφέρει ως προς το σχήμα της δεντρικής μορφής αλλά και ως προς το βάθος του δέντρου.

Για παράδειγμα, έστω ότι οι συμβολικές εκφράσεις S_1 και S_2 (βλ. Σχ. 17) έχουν κριθεί κατάλληλες να "επιζήσουν" από το τεστ που έγινε στα δεδομένα για το ταίριασμα και πρέπει να συνδυαστούν. Τα μέρη τα οποία θα συνδυαστούν επιλέγονται τυχαία, έστω ότι από την έκφραση S_1 επιλέχθηκε ο κόμβος με αριθμό 2 (συνδυασμός 2-3-4) και από την έκφραση S_2 επιλέχθηκε ο κόμβος 4 (συνδυασμός 4-5-6). Τότε, οι δύο αυτές εκφράσεις ανταλλάσσουν τα μέρη τους και δημιουργούν έτσι νέες εκφράσεις S_3 και S_4 (βλ. Σχ. 18). Μόλις ολοκληρωθεί ο συνδυασμός των εξισώσεων, οι πρόγονοι των συμβολικών εκφράσεων S_3 και S_4 αφαιρούνται



Σχήμα 18: Παραδείγματα συμβολικών εκφράσεων S_3 και S_4.

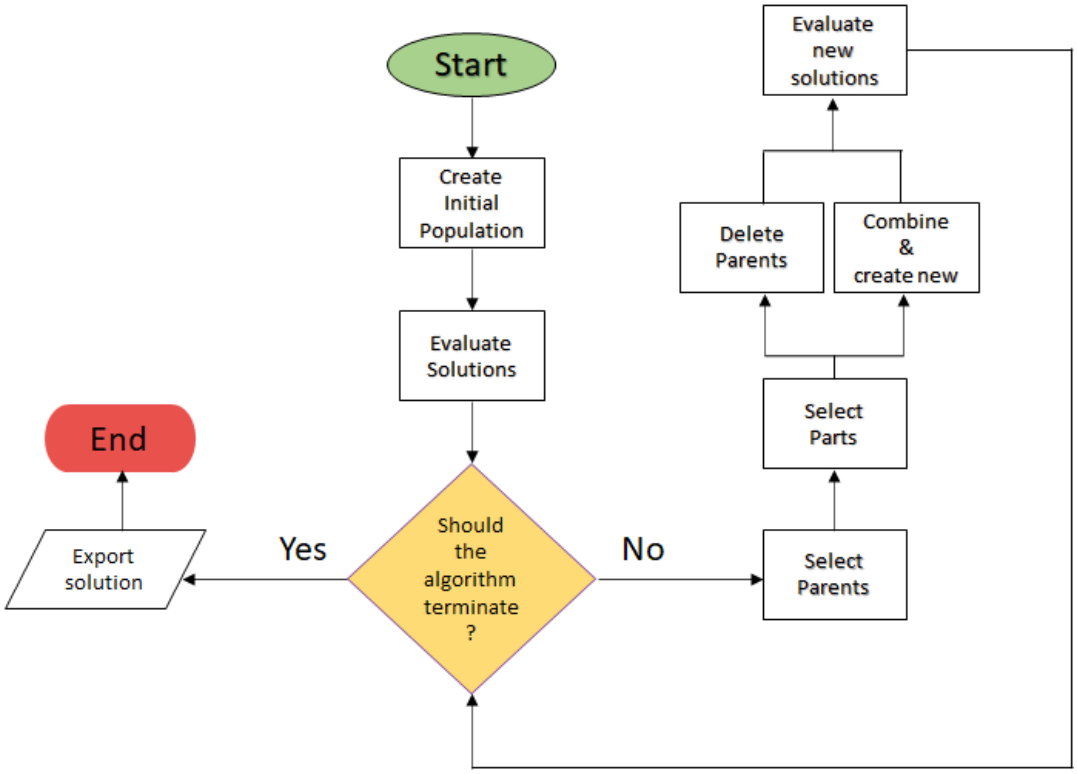
ταυτόχρονα από τον πληθυσμό. Αυτή η ανταλλαγή των υπομερών των εξισώσεων είχε ως αποτέλεσμα τη δημιουργία νέων εξισώσεων, ενώ αξίζει να σημειωθεί πως αν και το βάθος των δύο δέντρων παρέμεινε το ίδιο είναι καθαρά σύμπτωση, καθώς ο έλεγχος και ο συνδυασμός των εξισώσεων πραγματοποιείται σε χιλιάδες εξισώσεις.

Συνδυάζοντας παράλληλα τα μέρη των υποσχόμενων εκφράσεων, ο αλγόριθμος πλησιάζει τη βέλτιστη εξίσωση κάνοντας ένα βήμα τη φορά, καθώς μειώνεται το μέσο σφάλμα με την αφαίρεση των μερών που παρουσιάζουν κακή απόδοση, σε μία διαδικασία που θεωρείται εξελικτική, μέχρι την απόκτηση μίας τελικής "δυνατής" εξίσωσης. Ο αλγόριθμος θα τερματίσει σε ένα σημείο το οποίο έχει προκαθορίσει ο χρήστης. Το σημείο αυτό μπορεί να είναι ο ορισμός ενός μέγιστου αριθμού πληθυσμών που θα δημιουργηθούν, όταν το σφάλμα μηδενιστεί (τέλειο ταίριασμα στα δεδομένα) ή το σημείο όπου μία εξίσωση θα παρουσιάσει στατιστικό σφάλμα μικρότερο από μία σταθερά. Στην πράξη όμως επιλέγεται ένας συνδυασμών των παραπάνω για να καλυφθεί κάθε ενδεχόμενο. Επίσης, αρκετοί επιλέγουν να εξάγει ο αλγόριθμος παραπάνω από μία εξίσωση (συνήθως με διαφορετικές πολυπλοκότητες) και να επιλέξουν την εξίσωση που επιθυμούν. Σε αυτή την περίπτωση, χρειάζεται προσοχή στην επιλογή της τελικής εξίσωσης, καθώς εάν επιλεγεί μία πάρα πολύ απλή εξίσωση, είναι αρκετά πιθανό να έχει υποπέσει σε υποπροσαρμογή και να παρουσιάζει ένα όχι και τόσο καλό ταίριασμα στα δεδομένα. Από την άλλη, μία εξίσωση με υψηλή πολυπλοκότητα, εκτός από το ότι θα είναι δύσκολο να έχει φυσική σημασία, είναι πιθανό να έχει υποπέσει σε

υπερπροσαρμογή και να εμποδίζεται έτσι η εφαρμογή της σε ένα διαφορετικό σετ δεδομένων ή και η γενίκευσή της (Papastamatiou et al., 2022). Ο όρος πολυπλοκότητα συνδέεται άμεσα με τον αριθμό των κόμβων που υπάρχουν στη δεντρική μορφή του γενετικού προγραμματισμού. Το μήκος και το βάθος του δέντρου της συμβολικής εξίσωσης, μπορεί να ρυθμιστεί ανάλογα με τις ανάγκες του κάθε χρήστη, έτσι ώστε να εξάγει εξισώσεις ρυθμισμένης πολυπλοκότητας.

Συνοψίζοντας, ο γενετικός προγραμματισμός ξεκινάει με το να δημιουργήσει έναν πληθυσμό από τυχαίες συμβολικές εκφράσεις των οποίων το πλήθος και το μέγεθος καθορίζεται από τον χρήστη. Ένας τρόπος απεικόνισης μίας συμβολικής έκφρασης είναι μέσω ενός διαγράμματος δένδρου το οποίο περιέχει αρχικές συναρτήσεις και τερματικούς κόμβους. Επιπλέον, ο γενετικός προγραμματισμός κάνει μία σύγκριση στα μέρη των εκφράσεων σχετικά με το πόσο ταιριάζουν στο σετ δεδομένων, επιλέγει τις υποσχόμενες εκφράσεις, συνδυάζει τυχαία υπομέρη τους και ταυτόχρονα διαγράφει τις εξισώσεις με χαμηλή αποτελεσματικότητα. Ανάλογα τη φύση του προβλήματος που μελετάται υπό τον γενετικό προγραμματισμό, θα διαφέρει και το μέτρο σύγκρισης που θα χρησιμοποιηθεί. Στην προκειμένη περίπτωση που ήταν η εξαγωγή εξίσωσης για την περιγραφή δεδομένων, μπορούσε να χρησιμοποιηθεί το άθροισμα τετραγώνων των σφαλμάτων (Εξ. (2.10)) ή το μέσο τετραγωνικό σφάλμα (Εξ. (3.1)). Τέλος, φτάνοντας στο σημείο να τερματίσει ο αλγόριθμος, εξάγει τον αριθμό εξισώσεων που του έχει ζητηθεί. Μία σχηματική αναπαράσταση της παραπάνω διαδικασίας απεικονίζεται στο Σχ. 19.

Εύκολα αντιλαμβάνεται κανείς πως κατά τη διαδικασία της συμβολικής παλινδρόμησης, πρέπει να βελτιστοποιηθούν οι παράμετροι που αντιστοιχούν σε παραπάνω από μία μεταβλητές. Αυτού του είδους οι ενέργειες που βελτιστοποιούν αρκετά αντικείμενα, αναφέρονται ως multi-objective optimization. Η επιλογή των εξαγόμενων εξισώσεων γίνεται μέσω του pareto frontier (pareto front). Πιο συγκεκριμένα, έστω ότι προσπαθούμε να ελαχιστοποιήσουμε δύο εξισώσεις F_1 και F_2. Καθώς ελαχιστοποιείται η εξίσωση F_1, υπάρχει η πιθανότητα η F_2 να μην είναι η ελάχιστη δυνατή. Ενώ αντίστροφα, όταν η εξίσωση F_2 είναι η ελάχιστη δυνατή, να μην ισχύει το ίδιο για την F_1. Εάν στοχεύσουμε στη λύση της μίας εξίσωσης, η λύση είναι μοναδική, ενώ εάν προσπαθήσουμε να ελαχιστοποιήσουμε και τις δύο



Σχήμα 19: Διαδικασία συμβολικής παλινδρόμησης μέσω γενετικού προγραμματισμού.

εξισώσεις ταυτόχρονα, η λύση που προκύπτει δεν είναι μοναδική καθώς υπάρχουν διάφοροι συμβιβασμοί (trade-offs) που πρέπει να λάβουμε υπόψη (Van Veldhuizen et al., 1998). Όλες εκείνες οι λύσεις που δεν μπορούμε να προσδιορίσουμε εάν η μία είναι καλύτερη από την άλλη, ονομάζονται pareto optimal.

Προτού παρουσιαστεί ο ρόλος των pareto optimal λύσεων στο pareto front, ας ξεκινήσουμε με τον ορισμό των μεγεθών που θα χρειαστούν, όπως (Hua et al., 2021):

- *Objective Optimization Problems*: Ένα πρόβλημα βελτιστοποίησης πολλαπλών αντικειμένων, μπορεί να περιγραφεί από τις μαθηματικές σχέσεις:

$$min\ F(x) = \left(f_1(x), f_2(x), \dots, f_m(x)\right)^T,$$

τέτοιο ώστε

$$x = (x_1, x_2, \dots, x_n)^T \in D,$$

όπου x είναι ένα διάνυσμα με n μεταβλητές απόφασης και $D \subset R^2$ είναι ο χώρος αποφάσεων. Με m γίνεται αναφορά στον αριθμό των αντικειμένων και $F(x) \in \boldsymbol{\Lambda} \subset R^2$ είναι το διάνυσμα αντικειμένου με m αντικείμενα και $\boldsymbol{\Lambda}$ είναι ο χώρος αντικειμένων.

- *Pareto Dominance*: Εάν $x_A, x_B \in D$ είναι δύο διανύσματα αποφάσεων, τότε το x_A κυριαρχεί (dominates) του $x_B (x_A < x_B)$, ανν:

$$f_i(x_A) \leq f_i(x_B), \qquad \forall i = 1,2,\dots,m$$
$$f_i(x_A) < f_i(x_B), \qquad \exists j = 1,2,\dots,m$$

  Το διάνυσμα x_A ονομάζεται μη-κυριαρχούμενο (non-dominated) και το διάνυσμα x_B ονομάζεται κυριαρχημένο (dominated). Ως Pareto optimal solution, ορίζουμε τη λύση για την οποία δεν υπάρχει κάποια εφικτή λύση που να την κυριαρχεί.

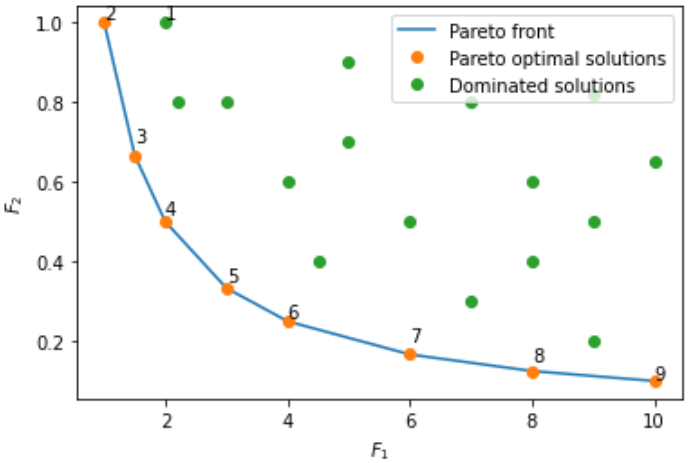
- *Pareto Set & Pareto Front*: Η ένωση όλων των λύσεων Pareto optimal ονομάζεται Pareto set:

$$PS = \{x_{PS} \in D \mid \nexists x \in D, x < x_{ps}\},$$

  ενώ η χαρτογράφηση του Pareto set στο χώρο τον αντικειμένων, απαρτίζει το Pareto front:

$$PF = \{F(x) \in \boldsymbol{\Lambda} \mid x \in PS\}$$

Συνοψίζοντας, διάφορες λύσεις παρουσιάζονται σε έναν χώρο (feasible solutions). Ένα τέτοιο παράδειγμα φαίνεται στο Σχ. 20. Μέσω των εξελικτικών αλγόριθμων βελτιώνονται οι λύσεις που βελτιστοποιούν τις παραμέτρους (Pareto improvement). Για παράδειγμα, η λύση με αριθμό 2 έναντι της 1 στο Σχ. 20, μπορεί να κράτησε σταθερή την τιμή για το ένα αντικείμενο (F_2), αλλά κατάφερε να ελαχιστοποιήσει το άλλο (F_1). Συνεπώς, η λύση 2 μπορεί να θεωρηθεί ότι κυριαρχεί έναντι της 1, ή αλλιώς πιο γενικά, η λύση 1 είναι κυριαρχημένη. Pareto optimal λύσεις (ή Pareto efficient), ονομάζονται εκείνες όπου δεν επιδέχονται καμία βελτιστοποίηση



Σχήμα 20: Παράδειγμα Pareto front.

ή αλλιώς δεν επιδέχονται καμία κυριάρχηση. Τέτοιου είδους λύσεις είναι εκείνες με αριθμούς $(2, \dots, 9)$. Όλες οι λύσεις που βρίσκονται πάνω στην καμπύλη Pareto (Pareto front) δεν μπορούν να αξιολογηθούν από το ποια είναι καλύτερη από την άλλη. Αυτό που προσφέρουν, είναι συμβιβασμοί ανάμεσα στις επιμέρους μεταβλητές ή στα επιμέρους αντικείμενα που μελετώνται.

Στη συμβολική παλινδρόμησης, στους άξονες X, Y του Σχ. 20, επιλέγεται η πολυπλοκότητα και η ακρίβεια. Έτσι, ανάλογα με το βαθμό πολυπλοκότητας της εξίσωσης που θα επιλεχθεί, στον άλλον άξονα θα εμφανίζεται η ακρίβεια της εξίσωσης. Η καμπύλη Pareto που θα σχηματιστεί, θα δίνει διάφορες λύσεις για διάφορες πολυπλοκότητες εξισώσεων, που παρουσιάζουν τη μεγαλύτερη ακρίβεια στα δεδομένα. Από αυτές τις λύσεις που δεν υπάρχει καλύτερη προσέγγιση, επιλέγεται η εξαγόμενη εξίσωση.

3.6 Χαρακτηριστικά

Το μεγάλο πλεονέκτημα της συμβολικής παλινδρόμησης είναι το γεγονός ότι δε χρειάζεται να έχει φυσική γνώση του υπό μελέτη συστήματος, επειδή ενεργεί αποκλειστικά σε δεδομένα (Kabliman et al., 2019). Η δυνατότητα αυτή είναι ιδιαίτερα σημαντική στις φυσικές

επιστήμες και στην επιστήμη των υλικών διότι η αποκρυπτογράφηση αρκετών φυσικών νόμων μέσω εξισώσεων δημιουργεί μία δομή με την οποία οι περισσότεροι επιστήμονες και μηχανικοί είναι οικείοι. Ταυτόχρονα, η τεχνική της συμβολικής παλινδρόμησης ξεχωρίζει από τα υπόλοιπα είδη παλινδρομήσεων καθώς δεν εγκλωβίζεται μόνο στη βελτιστοποίηση των σταθερών των παραμέτρων αλλά δημιουργεί την εξίσωση από το μηδέν, χωρίς να έχει την παραμικρή προκατάληψη για το σύστημα που μελετάται (Sofos, Charakopoulos, et al., 2022).

Κάποιοι, κατά την υλοποίηση της συμβολικής παλινδρόμησης επιλέγουν να οριοθετήσουν το χώρο αναζήτηση σε ένα μέρος των συμβόλων καθώς θέτουν διάφορους περιορισμούς στον αλγόριθμο (συνήθως με το να αποδεχθούν κάποια γνώση για το υπό μελέτη σύστημα) (Brence et al., 2021; França, 2018; Kronberger, de Franca, et al., 2022). Έτσι, μπορεί να επιβληθεί μία συγκεκριμένη μορφή στις εξισώσεις (π.χ., μονοτονία), να ελαττωθεί το υπολογιστικό κόστος και να αυξηθεί η ακρίβεια (Angelis et al., n.d.).

Αξιοσημείωτη ακόμα, είναι η συνεισφορά της συμβολικής παλινδρόμησης στη βαθμονόμηση παραμέτρων (calibration of parameters). Η βαθμονόμηση αρκετών παραμέτρων ενός μοντέλου που βασίζονται βασίζονται στις συνθήκες επεξεργασίας, αποτελεί ένα σύνηθες πρόβλημα (Kronberger, Kabliman, et al., 2022), καθώς για να έρθουν σε συμφωνία με τις μετρήσεις, ταιριάζονται σε αυτές. Για αυτό το λόγο, δεν είναι απόλυτα γνωστός το ποσοστό εξάρτησης, ή αλλιώς ο βαθμός εξάρτησης της βαθμονόμησης των παραμέτρων (Kabliman et al., 2021). Μέσω συμβολικής παλινδρόμησης, έχουν προταθεί δύο τεχνικές (explicit & implicit) για να αντιμετωπιστούν τέτοιου είδους θέματα (Kronberger, Kabliman, et al., 2022).

Επίσης, οι δυνατότητες της συμβολικής παλινδρόμησης δεν περιορίζονται μόνο σε μία απλή εξαγωγή εξισώσεων. Εξίσου σημαντική είναι ικανότητα να ξεδιαλύνει κάποιες σχέσεις που υπάρχουν σε ένα σετ δεδομένων, στις οποίες υπήρχε μερική γνώση των φυσικών νόμων οι οποίοι είναι υπεύθυνοι για τη συμπεριφορά του συστήματος (Vaddireddy et al., 2020) ή ήταν εν μέρη ασαφής (Gilpin, 2021). Άλλοι, χρησιμοποιούν τη συμβολική παλινδρόμηση για να επιβεβαιώσουν ήδη υπάρχοντες φυσικούς νόμους (Schmidt & Lipson, 2009) (π.χ., νόμος βαρύτητας (Lemos et al., 2022)). Σε αυτές τις περιπτώσεις, δεν είναι απλή η εξαγωγή και επαλήθευση φυσικών νόμων διότι υπεραπλουστευμένα σετ δεδομένων ή η έλλειψη κατάλληλων μέτρων σύγκρισης μπορεί να οδηγήσει σε λανθασμένα αποτελέσματα (Matsubara et al., 2022).

Για να εμπιστευτεί κανείς τις εξισώσεις που δημιουργούνται, χρειάζεται κατανόηση της φυσικής συμπεριφοράς του συστήματος, ειδικά σε περιπτώσεις περίπλοκων και μη γραμμικών δυναμικών συστημάτων (Haider et al., 2021). Από τη μία μεριά, αρκετές εξισώσεις ή φυσικοί νόμοι περίπλοκων συστημάτων είναι δύσκολο να βρεθούν ή είναι ελάχιστες, και μέσω μοντέλων που δρουν μέσω δεδομένων εξάγονται διαφορικές εξισώσεις για την αναγνώριση αυτών των νόμων (Cao & Zhang, 2022). Από την άλλη μεριά, ο θόρυβος που υπάρχει στα δεδομένα ή η έλλειψη δεδομένων αποτελεί ένα κύριο πρόβλημα κατά την υλοποίηση ενός μοντέλου που ενεργεί βάσει δεδομένων (Gilpin, 2021). Ο θόρυβος μπορεί εν μέρει να αντιμετωπιστεί βάσει του κατάλληλου μοντέλου (Angelis et al., n.d.), ενώ το γεγονός ότι η συμβολική παλινδρόμηση έχει αποδειχθεί πιο αποτελεσματική έναντι άλλων μοντέλων όταν δεδομένα είναι σε έλλειψη (Wilstrup & Kasak, 2021), επιλέγεται με μεγαλύτερη ευκολία για τη μελέτη φυσικών συστημάτων. Θα πρέπει επίσης να αναφερθεί, ότι η εξαγωγή εξισώσεων διευκολύνει την ενσωμάτωση τους σε άλλες τεχνικές (π.χ., finite element solver) (Kabliman et al., 2019) και ενισχύεται έτσι η αποδοχή της συμβολικής παλινδρόμησης από την επιστημονική κοινότητα, διότι όχι μόνο μπορεί να βρει, να διαλευκάνει και να επαληθεύσει φυσικούς νόμους, αλλά μπορεί να χρησιμοποιηθεί και σε συνδυασμό με άλλες τεχνικές.

Αντίθετα, το βασικό μειονέκτημα της συμβολικής παλινδρόμησης έναντι άλλων μοντέλων, είναι το υπολογιστικό κόστος για την εξαγωγή των εξισώσεων (Danai & Cava, 2021). Ο αριθμός των εξισώσεων και των μερών τους που πρέπει να αξιολογηθούν είναι αρκετά μεγάλος (Danai & Cava, 2021). Για αυτό το λόγο, προτιμάται η εφαρμογή της σε προβλήματα όπου ο αριθμός των παραμέτρων εισόδου είναι σχετικά μικρός (Wadekar et al., 2020). Παρ' όλα αυτά, μια τεχνική που αξιοποιείται τακτικά είναι η αναγνώριση των κύριων παραμέτρων εισόδου σε ένα σύστημα (συνήθως από ένα άλλο μοντέλο, όπως π.χ., random forest (Karakasidis et al., 2022)) και έπειτα η εφαρμογή της συμβολικής παλινδρόμησης.

Τέλος, προσοχή θα πρέπει να δοθεί στο μέγεθος των εξαγόμενων εκφράσεων. Η συμβολική παλινδρόμηση χαρακτηρίζεται ως μία ερμηνεύσιμη μέθοδος, σε αντίθεση με άλλες μεθόδους μαύρου κουτιού. Το χαρακτηριστικό της ερμηνευσιμότητας συνήθως ταυτίζεται με το μέγεθος των εξαγόμενων εκφράσεων, διότι εάν εξάγει μία πάρα πολύ μεγάλη εξίσωση με πολλές παραμέτρους, τότε θα είναι εξαιρετικά δύσκολο να κατανοηθεί ο μηχανισμός του συστήματος. Μία παρενέργεια της συμβολικής παλινδρόμησης είναι το "φούσκωμα" των

εξισώσεων (bloat) (Bomarito et al., 2022). Σε αυτό το "φούσκωμα" παρατηρείται μία ραγδαία αύξηση στο μέγεθος και την πολυπλοκότητα των εξισώσεων ενώ το ταίριασμα στα δεδομένα αυξάνει ελάχιστα (Bomarito et al., 2022).

Κεφάλαιο 4 Μηχανική Μάθηση στην Επιστήμη των Υλικών

Η ισορροπία που έπρεπε να τηρηθεί στις υπολογιστικές προσομοιώσεις μεταξύ υπολογιστικού κόστους και ακρίβειας έχει υποσκελιστεί από την άνοδο της υπολογιστικής ισχύς σε συνδυασμό με τις μεθόδους που βασίζονται σε δεδομένα και την ενίσχυσή τους μέσω της τεχνητής νοημοσύνης. Η σύνδεση της επιστήμης δεδομένων και της μηχανικής μάθησης στην επιστήμη των υλικών (μεταξύ άλλων κλάδων) είναι φανερή από τον αυξανόμενο αριθμό εργασιών που σχετίζονται με το θέμα (Zhou et al., 2019). Η μηχανική μάθηση έχει τη δυνατότητα να κάνει σχετικά γρήγορες προβλέψεις χωρίς να στερείται ακρίβεια (Schmid et al., 2021) και να έχει καλύτερη υπολογιστική απόδοση από τις αριθμητικές προσομοιώσεις (Goh et al., 2021).

Η χρήση των μεθόδων μηχανικής μάθησης έγκειται στην εύρεση και την πρόβλεψη ιδιοτήτων των ήδη υπαρχόντων υλικών, γεγονός το οποίο με τη σειρά του θα συμβάλει στη δημιουργία νέων, καινοτόμων υλικών με εξειδικευμένες ιδιότητες (Angelis et al., n.d.). Πλέον, γραφικά δεδομένα, δεδομένα που βρίσκονται στη βιβλιογραφία ή ακόμη και δεδομένα από πειράματα και προσομοιώσεις, αποθηκεύονται σε ηλεκτρονικές βάσεις δεδομένων (π.χ., Inorganic Crystal Structure Database (Belsky et al., 2002), Open Quantum Materials Database (Kirklin et al., 2015)) σε διαφορετικές μορφές. Κινήματα όπως παραδείγματος χάριν η Πληροφορική Υλικών (Materials Informatics) (Lopez-Bezanilla & Littlewood, 2020) τα οποία αξιοποιούν χαρακτηριστικά του σύγχρονου περιβάλλοντος, όπως είναι η διαχείριση βάσεων δεδομένων, υπολογισμοί πολλαπλής κλίμακας και πειραμάτων μεταξύ άλλων, δημιουργήθηκαν για την ανακάλυψη καινοτόμων υλικών σε ένα λογικό υπολογιστικό κόστος (Angelis et al., n.d.).

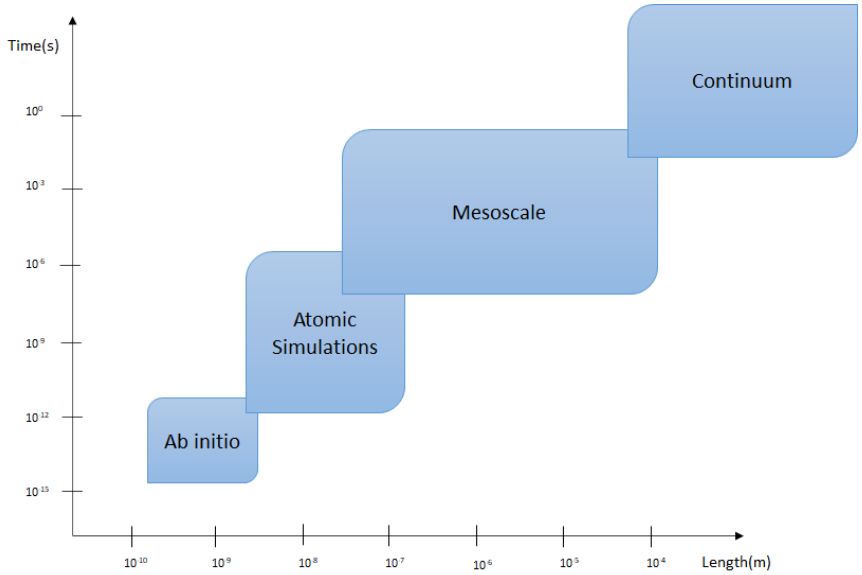
Στο Κεφάλαιο αυτό αναλύεται, στη συνέχεια, ο τρόπος που πραγματοποιούνται οι υπολογισμοί σε πολλαπλή κλίμακα και η εφαρμογή των μεθόδων μηχανικής μάθησης για την πρόβλεψη ιδιοτήτων και το σχεδιασμό υλικών.

4.1 Μοντελοποίηση σε πολλαπλή κλίμακα

Η μελέτη και η έρευνα των υλικών, είτε αυτά είναι σε στερεά είτε σε υγρή μορφή, πραγματοποιείται σε πολλαπλές κλίμακες χρόνου και μεγέθους όπου η κάθε κλίμακα ενσωματώνει ευρήματα της προηγούμενης (Angelis et al., n.d.). Στο Σχ. 21, απεικονίζεται η προσέγγιση της μελέτης των υλικών καθώς και μέθοδοι που χρησιμοποιούνται.

Η μικρότερη κλίμακα όπου πραγματοποιούνται μελέτες είναι η ατομική κλίμακα (atomic scale). Σε αυτή την κλίμακα, μέθοδοι πρώτης αρχής (ab-initio, first principles) πραγματοποιούνται απευθείας βάσει κβαντομηχανικών υπολογισμών και χωρίς να χρειάζονται υποθέσεις ούτε ενίσχυση από πειραματικά δεδομένα, με σκοπό να βρουν μία σχέση ικανή να περιγράψει την ενέργεια του υπό μελέτη συστήματος, την Επιφάνεια Δυναμικής Ενέργειας (Potential Energy Surface). Το πρόβλημα που δημιουργείται είναι πως οι υπολογισμοί αυτοί, βασίζονται σε λύσεις της εξίσωσης Schrödinger και αυτή αποτελεί μία όχι και τόσο πρακτική διαδικασία όταν εφαρμόζεται σε αρκετά συστήματα που μελετώνται καθημερινά (Angelis et al., n.d.). Έως ένα βαθμό όμως, μπορούν να προσεγγιστούν μέσω της ενσωμάτωσης της Συναρτησιακής Θεωρίας Πυκνότητας (Behler & Parrinello, 2007). Από τη μία μεριά, οι μέθοδοι ατομικής κλίμακας έχουν παρουσιάσει ικανοποιητική ακρίβεια στις κβαντικές προσεγγίσεις, αλλά από την άλλη, είναι περιορισμένες ως προς τον υπολογιστικό χρόνο που απαιτούν ή το μέγεθος των προσομοιώσεων (Angelis et al., n.d.). Λόγω των περιορισμών των μεθόδων ατομικής κλίμακας, δημιουργήθηκαν αρκετές μέθοδοι σωματιδίων για να προσπελάσουν τα εμπόδια.

Οι σωματιδιακές μέθοδοι, όπως για παράδειγμα Μοριακής Δυναμικής (Molecular Dynamics), Υδροδυναμικής Λείων Σωματιδίων (Smoothed Particle Hydrodynamics) και Δυναμικής Διασκορπιστικών Σωματιδίων (Dissipative Particle Dynamics), μπορούν να εφαρμοστούν σε διάφορες κλίμακες χρόνου και μεγέθους. Τα χαρακτηριστικά τέτοιων ειδών προσεγγίσεων, εμπεριέχουν την θεώρηση ενός σωματιδίου ως υλικό σημείο αλλά και ως σημείο προσέγγισης (Angelis et al., n.d.). Πιο συγκεκριμένα, στην ατομική κλίμακα μέσω των προσεγγίσεων της Μοριακής Δυναμικής, ένα σωματίδιο θεωρείται ως ένα άτομο ή μόριο, στη μέσο-κλίμακα μέσω της Δυναμικής Διασκορπιστικών Σωματιδίων ως ένα σμήνος ατόμων ή μορίων, ενώ στη μακρο-κλίμακα μέσω της Υδροδυναμικής Λείων Σωματιδίων ως μία μικρή περιοχή (Ye et al., 2019).



Σχήμα 21: Πολλαπλή κλίμακα της Επιστήμης των Υλικών και μέθοδοι που χρησιμοποιούνται.

Η ένταξη των υπολογιστικών προσομοιώσεων στις μελέτες των υλικών, κάλυψε διάφορα κενά που υπήρχαν στις πειραματικές προσεγγίσεις, όπως για παράδειγμα κάποιες περιοχές στην ατομική κλίμακα που δεν μπορούσαν να εκτιμηθούν (Sofos et al., 2009). Πλέον, η εισχώρηση της μηχανικής μάθησης στην επιστήμη των υλικών αποτελεί μία κοινή προσέγγιση, καθώς επιφέρει πιο ακριβή αποτελέσματα σε ελάχιστο χρόνο συγκριτικά με τις προσομοιώσεις ή τις πειραματικές διαδικασίες. Επίσης, η μηχανική μάθηση έχει τη δυνατότητα να ενσωματώσει τις έως τώρα τεχνικές στις προσεγγίσεις της και να ενισχύσει περισσότερο τις προβλέψεις της.

Για παράδειγμα, η συμβολή της μηχανικής μάθησης στις προσομοιώσεις πολλαπλής κλίμακας, είναι καθοριστική πλέον στη δημιουργία των ενδοατομικών δυναμικών μηχανικής μάθησης (ML interatomic potentials)(Mishin, 2021; Y. Yang et al., 2021). Μάλιστα, αναφορές φέρουν τα ενδοατομικά δυναμικά μηχανικής μάθησης που έχουν εκπαιδευτεί από προσομοιώσεις Μοριακής Δυναμικής πρώτων αρχών (ab-initio Molecular Dynamics), να είναι ικανά να γεφυρώσουν τις προσομοιώσεις Μοριακής Δυναμικής, Συναρτησιακής Θεωρίας Πυκνότητας (DFT) και Πεπερασμένων Στοιχείων (Finite Element) και έτσι να επεκτείνουν τις δυνατότητες της έρευνας στα υλικά (Stephenson et al., 2018). Επιπρόσθετα, η εφαρμογή των τεχνικών συμβολικής παλινδρόμησης στα ενδοατομικά δυναμικά, επέφερε την εξαγωγή απλών εξισώσεων μειώνοντας έτσι το ρίσκο υπερπροσαρμογής, όπως αναφέρεται για παράδειγμα στο μοντέλο (Hernandez et al., 2019) η δημιουργία ενδοατομικών δυναμικών πολλαπλών

σωμάτων γρήγορα και με ακρίβεια, ενώ την ίδια στιγμή βασίζονται σε θεμελιώδεις φυσικές αρχές και είναι ικανά να χρησιμοποιηθούν σε πολλαπλή κλίμακα.

Η σύνδεση που επιτυγχάνεται ανάμεσα στις κλίμακες μέσω του συνδυασμού των κλασσικών μεθόδων και αυτών της μηχανικής μάθησης, επιφέρει πιο ακριβής προβλέψεις (Angelis et al., n.d.). Επίσης, πρέπει να σημειωθεί η αναγνώριση των τεχνικών συμβολικής παλινδρόμησης ως ένα εξαιρετικό εργαλείο για τη μελέτη σε πολλαπλές κλίμακες (Angelis et al., n.d.), καθώς η μορφή των αναλυτικών εξισώσεων έχει χαρακτηριστεί αρκετά ελκυστική, τόσο για τη μελέτη ιδιοτήτων στη μικρο-κλίμακα, όσο και στη μακρο-κλίμακα (Reimann et al., 2019).

4.2 Πρόβλεψη ιδιοτήτων και σχεδιασμός υλικών

Ο τρόπος με τον οποίο γίνονται οι μελέτες στα υλικά έχει αλλάξει, με την επιλογή μοντέλων μηχανικής μάθησης για την εξαγωγή ιδιοτήτων υλικών, καθώς η ανάπτυξη των βάσεων δεδομένων για ιδιότητες υλικών έχει αναπτυχθεί ιδιαίτερα (Hart et al., 2021). Πιο συγκεκριμένα, πραγματοποιούνται προσεγγίσεις όσον αφορά τις ιδιότητες των υλικών και της συσχέτισής τους με τις συνθήκες επεξεργασίας, έχοντας ως σκοπό τη δημιουργία νέων υπολογιστικών μοντέλων (Kronberger, Kabliman, et al., 2022). Μία περίπτωση τέτοιων μοντέλων απαρτίζουν τα συστατικά (constitutive) μοντέλα όπου περιγράφουν σε διαφορετικές συνθήκες τη συμπεριφορά που παρατηρείται στο υλικό, το οποίο συμβάλει στη δημιουργία μίας σχέσης τάσης-παραμόρφωσης (stress-strain) για την εξαγωγή των νόμων που διέπουν το σύστημα (Zhang et al., 2017) (π.χ., μελέτη πλαστικότητας (Bomarito et al., 2021)). Άλλες τεχνικές εστιάζουν στη δημιουργία εξισώσεων από το μηδέν για τη πρόβλεψη ιδιοτήτων υλικών (π.χ., διάχυση Lennard-Jones υγρού (Alam et al., 2022)).

Μέσω διάφορων συνδυασμών στοιχείων του περιοδικού πίνακα, μπορεί ο οποιοσδήποτε να δημιουργήσει ένα υλικό, θεωρητικά τουλάχιστον (Angelis et al., n.d.). Η έννοια της σταθερότητας (stability) συναντάται σε μορφή κριτηρίου για να εκτιμηθεί η κατασκευή του θεωρητικού αυτού υλικού. Οι μέθοδοι συμβολικής παλινδρόμησης έχουν αποδειχθεί ως ένα

αξιόπιστο μέσο στην ανακάλυψη και τη σχεδίαση υλικών, καθώς μπορούν να αναγνωρίσουν αναπαραστάσεις χημικών ενώσεων (descriptors (Seko et al., 2018)) και να δημιουργήσουν νέες που να σχετίζονται με τη σταθερότητα (He & Zhang, 2021). Κατά τη δημιουργία και το σχεδιασμό νέων υλικών οι ερευνητές προσπαθούν να δημιουργήσουν υλικά με στοχευμένες ιδιότητες, αλλά αυτό αρκετές φορές είναι εξαιρετικά απαιτητική διαδικασία καθώς υπάρχει η πιθανότητα η ιδιότητα-στόχος, να παρατηρείται σε συγκεκριμένη, μοναδική δομή, ή άλλες ιδιότητες του υλικού να χρειάζεται να έρθουν σε συμφωνία για να επιτευχθεί καλύτερη απόδοση (Hautier, 2019).

Η αναγνώριση των παραμέτρων που είναι υπεύθυνες για τη λειτουργικότητα αλλά και των εξαρτήσεών τους, κρίνεται αναγκαία προκειμένου να βελτιστοποιηθούν (Y. Wang et al., 2019). Ένα τέτοιο παράδειγμα αναφέρεται σε μία εφαρμογή της συμβολικής παλινδρόμησης (He & Zhang, 2021), όπου κατασκευάστηκαν ακριβείς αναπαραστάσεις χημικών ενώσεων χωρίς να υπάρχει γνώση του υπό μελέτη συστήματος πριν την εφαρμογή της, οι οποίες καθορίζουν τη σταθερότητα των MXene. Άλλοι, δημιούργησαν αναπαραστάσεις χημικών ενώσεων όπου συνέβαλαν στην ανακάλυψη νέων καταλυτών (Weng et al., 2020). Μέσω αυτών των παραδειγμάτων ενισχύονται οι θεωρίες περί επιτάχυνσης της ανακάλυψης νέων υλικών χωρίς να υπάρχει πρότερη γνώση για το σύστημα, μέσω της δημιουργίας αναπαραστάσεων χημικών ενώσεων από τη συμβολική παλινδρόμηση (Angelis et al., n.d.)

Κεφάλαιο 5 Εφαρμογή

Στο παρόν κεφάλαιο, επιχειρείται μέσω της συμβολικής παλινδρόμησης ο υπολογισμός ιδιοτήτων μεταφοράς ρευστών από δεδομένα που προέρχονται από προσομοιώσεις στη νανοκλίμακα. Οι ιδιότητες μεταφοράς καθορίζουν τους μηχανισμούς μεταφοράς της μάζας, της ορμής και της ενέργειας στα υλικά. Από τη διεθνή βιβλιογραφία (Bugel & Galliero, 2008; Meier et al., 2004) συλλέχθηκαν τρία σετ δεδομένων. Οι ανεξάρτητες μεταβλητές είναι τα μακροσκοπικά μεγέθη της πυκνότητας (ρ) και της θερμοκρασίας (T), ενώ εξαρτημένες μεταβλητές είναι οι τρεις ιδιότητες μεταφοράς, δηλαδή ο συντελεστής διάχυσης (D), το ιξώδες διάτμησης (n_s) και η θερμικής αγωγιμότητας (λ). Η διαδικασία για την εύρεση των παραπάνω ιδιοτήτων μεταφοράς είναι εξαιρετικά απαιτητική πειραματικά, ενώ οι προσομοιώσεις που εκτελούνται συνήθως με Μοριακή Δυναμική, είναι ιδιαίτερα χρονοβόρες και απαιτούν λεπτομερή στατιστική επεξεργασία. Στη βιβλιογραφία συναντώνται αρκετές εμπειρικές εξισώσεις, οι οποίες μπορούν να δώσουν τη γενική εικόνα των ιδιοτήτων αυτών.

Όπως γίνεται αντιληπτό, το πεδίο έρευνας είναι κατάλληλο για τη διερεύνηση της εφαρμογής της μεθόδου της συμβολικής παλινδρόμησης, με στόχο την εξαγωγή απλών εκφράσεων για τις τρεις ιδιότητες μεταφοράς, οι οποίες θα εμπεριέχουν επαρκή φυσική ερμηνεία. Στη συνέχεια, παρουσιάζονται οι εξισώσεις που πηγάζουν από τη στατιστική μηχανική και συνήθως χρησιμοποιούνται για τον υπολογισμό των ιδιοτήτων μεταφοράς, ενώ αμέσως μετά επιχειρείται η εφαρμογή της συμβολικής παλινδρόμησης μέσω του λογισμικού HeuristicLab (Wagner et al., 2014) για την εύρεση εξισώσεων που εκφράζουν τις ιδιότητες μεταφοράς σε σχέση με την πυκνότητα και τη θερμοκρασία.

Η επιλογή των εξισώσεων συμβολικής παλινδρόμησης, έγινε με βάση τη φυσική ερμηνεία που μπορούν να προσφέρουν για να περιγράψουν τη συμπεριφορά του ρευστού, το ελάχιστο σφάλμα και τη χαμηλή πολυπλοκότητα, έτσι ώστε να αποφευχθούν φαινόμενα υπερπροσαρμογής. Ένα επιπλέον μέτρο σύγκρισης που βοήθησε στην επιλογή των εξισώσεων,

βασίζεται στα γραφήματα ταυτότητας (identity plots). Τέλος, πραγματοποιήθηκε μία προσπάθεια εύρεση της εξίσωσης που προσδίδει μεγαλύτερη ακρίβεια.

5.1 Θεωρητικές σχέσεις ιδιοτήτων μεταφοράς

Οι έρευνες των ιδιοτήτων μεταφοράς, όπως του συντελεστή διάχυσης, του ιξώδους μεταφοράς και της θερμικής αγωγιμότητας μέσω υπολογιστικών προσομοιώσεων, έχουν απασχολήσει διαχρονικά αρκετούς ερευνητές (Y. Hu et al., 1996; Liu et al., 2006; Murad et al., 1993). Οι σχέσεις που περιγράφουν τις παραπάνω ιδιότητες, βρίσκονται στην εργασία (Sofos et al., 2009).

Ο συντελεστής διάχυσης μπορεί να περιγραφεί από τη σχέση Einstein (Εξ. (5.1)) ή τη σχέση Green-Kubo (Εξ. (5.2)).

$$D = \lim_{t\to\infty} \frac{1}{2dNt} < \sum_{j=1}^{N} \left[r_j(t) - r_j(0)\right]^2 > \tag{5.1}$$

$$D = \frac{1}{3N} \int_0^\infty < \sum_{j=1}^{N} vj(0) * v_j(t) > \tag{5.2}$$

όπου r_j είναι το διάνυσμα θέσης του j^{th} ατόμου, d είναι η διάσταση του προβλήματος και v_j είναι το διάνυσμα ταχύτητας του j^{th} ατόμου.

Το ιξώδες διάτμησης από την άλλη, μπορεί να περιγραφεί από την Εξ. (5.3), όπου J_p^{xy} είναι το στοιχείο εκτός διαγωνίου του μικροσκοπικού τανυστή τάσης (βλ. Εξ. (5.4)), όπου $u(r_{ij})$ είναι το δυναμικό Lennard-Jones του i ατόμου, r_{ij} είναι το διάνυσμα θέσης μεταξύ των ατόμων i και j και v_i^j είναι το j-στοιχείο της ταχύτητας του ατόμου i.

$$n_s = \frac{1}{Vk_BT} \int_0^\infty dt < J_p^{xy}(t) \ * \ J_p^{xy}(0) > \tag{5.3}$$

$$J_p^{xy} = \sum_{i=1}^{N} m_i \, v_i^x v_i^y - \sum_{i=1}^{N} \sum_{j>1}^{N} r_{ij}^x \frac{\partial u(r_{ij})}{\partial r_{ij}^y} \tag{5.4}$$

Τέλος, η θερμική αγωγιμότητα δίνεται από την Εξ. (5.5), όπου είναι το ολοκλήρωμα της συνάρτησης χρονικής αυτοσυσχέτισης της μικροσκοπικής ροής θερμότητας J_q^x (βλ. Εξ. (5.6)), όπου v_i είναι το μέγεθος της ταχύτητας του i ατόμου και $\boldsymbol{I}$ ο μοναδιαίος πίνακας.

$$\lambda = \frac{1}{V k_B T_2} \int_0^{\infty} dt < J_q^x(t) \ * \ J_q^x(0) > \tag{5.5}$$

$$J_q^x = \frac{1}{2} \sum_{i=1}^{N} m_i (v_i)^2 \, v_i^x - \sum_{i=1}^{N} \sum_{j>1}^{N} \left[r_{ij}^x : \frac{\partial u(r_{ij})}{\partial r_{ij}^x} - \boldsymbol{I} * u(r_{ij}) \right] * \ u_i^x \tag{5.6}$$

Οι σχέσεις αυτές προέρχονται από τη στατιστική μηχανική και μπορεί να θεωρηθούν ιδιαίτερα περίπλοκες, καθώς περιλαμβάνουν μικροσκοπικές ιδιότητες του συστήματος σε ατομικό επίπεδο, όπως θέσεις και ταχύτητες σωματιδίων αλλά και τα δυναμικά αλληλεπίδρασης μεταξύ τους. Ως συνέπεια, γίνεται αντιληπτό ότι η υλοποίηση υπολογιστικών προσομοιώσεων απαιτεί άριστη γνώση της φυσικής συμπεριφοράς του συστήματος προκειμένου να αποφευχθούν σφάλματα.

Με την εισαγωγή της μηχανικής μάθησης στην πρόβλεψη ιδιοτήτων, από την άλλη, είναι δυνατή η μερική ή ολική αντικατάσταση πολύπλοκων και χρονοβόρων διαδικασιών, με ικανοποιητική ακρίβεια. Οι τεχνικές μηχανικής μάθησης όμως, αν και εφαρμόζονται από μία ευρεία γκάμα ερευνητών στην επιστήμη των υλικών, χωρίς την αποτύπωση της φυσικής διεργασίας του συστήματος, θεωρούνται ελλιπείς (Papastamatiou et al., 2022). Στην κατεύθυνση αυτή, στόχος της εργασίας είναι να αναζητηθούν απλές σχέσεις για τις ιδιότητες μεταφοράς των ρευστών, οι οποίες θα κουβαλούν τη φυσική ερμηνεία του συστήματος, βασιζόμενες αποκλειστικά στα διαθέσιμα δεδομένα.

5.2 Υλοποίηση Συμβολικής Παλινδρόμησης

Η χρήση της συμβολικής παλινδρόμησης πραγματοποιείται για την εξαγωγή σχέσεων που περιγράφουν το συντελεστή διάχυσης D, το ιξώδες διάτμησης n_s και τη θερμική αγωγιμότητα λ, ως συναρτήσεις της θερμοκρασίας T και της πυκνότητας ρ. Η υλοποίησή της, έγινε μέσω του λογισμικού HeuristicLab. Πιο συγκεκριμένα, 80% των δεδομένων και στις τρεις περιπτώσεις χρησιμοποιήθηκε για εκπαίδευση του αλγορίθμου, ενώ το υπόλοιπο 20% χρησιμοποιήθηκε για έλεγχο στα δεδομένα, με τυχαία επιλογή από το σετ δεδομένων. Το λογισμικό HeuristicLab λειτουργεί με βάση τη διαδικασία του γενετικού προγραμματισμού που περιεγράφηκε στο Κεφ. 3.5. Τεχνικές λεπτομέρειες του προγράμματος παρατίθενται στο Παράρτημα.

Τα μαθηματικά σύμβολα και οι συναρτήσεις που χρησιμοποιούνται ως δομικά στοιχεία των εξισώσεων είναι:

- *Πρόσθεση* $(+)$.
- *Αφαίρεση* $(-)$.
- *Πολλαπλασιασμό* $(*)$.
- *Διαίρεση* $(\div)$.
- *Εκθετικό* (exp).
- *Λογαριθμικό* (log).
- *Δύναμη* (^).

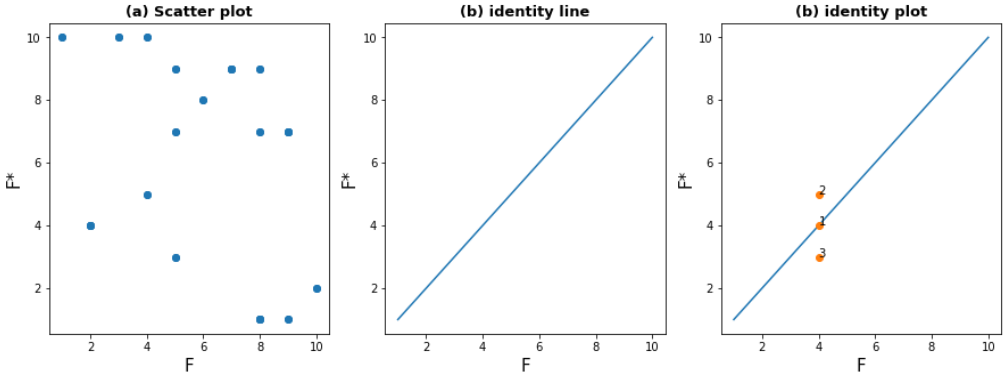
Η προσέγγιση της συμβολικής παλινδρόμησης μέσω του γενετικού προγραμματισμού βασίζεται στην τυχαιότητα. Έτσι, κάθε ολοκλήρωση της διαδικασίας λόγο μη ντετερμινισμού, θα επιφέρει διαφορετική εξίσωση που περιγράφει τα δεδομένα. Οι εξισώσεις που προκύπτουν όμως, δεν αποκαλύπτουν όλες τη φυσική σημασία της διεργασίας, ενώ κάποιες από αυτές είναι προϊόν υπερπροσαρμογής στα δεδομένα. Κάποιες εξισώσεις όμως, εμφανίζονται σχεδόν σε κάθε επανάληψη της διαδικασίας παρά την τυχαιότητα που επικρατεί στον γενετικό προγραμματισμό και τις μεταλλάξεις των εξισώσεων. Αυτό το γεγονός με τη σειρά του προϊδεάζει ότι οι εξισώσεις αυτές, είναι αρκετά πιθανό να συμβάλουν έως ένα βαθμό στην

πραγματική σχέση που αποτυπώνει τη συμπεριφορά του συστήματος. Για την αποφυγή της τυχαιότητας και της υπερπροσαρμογής, η τεχνική της συμβολικής παλινδρόμησης εφαρμόστηκε για κάθε σετ δεδομένων 9 φορές ενώ η επιλογή των εξισώσεων έγινε με βάση:

- τη φυσική ερμηνεία του συστήματος,
- την πολυπλοκότητά τους,
- το ποσοστό σφάλματος που παρουσιάζουν (MSE, Εξ. (3.1)),
- τα γραφήματα ταυτότητας.

Το γράφημα ταυτότητας είναι η ένωση ενός γραφήματος διασποράς και της γραμμής ταυτότητας (βλ. Σχ. 22). Ένα γράφημα διασποράς (scatter plot) είναι ένας τρόπος απεικόνισης της σχέσης μεταξύ μεταβλητών σε ένα καρτεσιανό επίπεδο. Ένα παράδειγμα γραφήματος διασποράς φαίνεται στο Σχ. 22a. Πιο συγκεκριμένα, αναπαρίσταται η σχέση ανάμεσα στην πραγματική τιμή μίας μεταβλητής F και της πρόβλεψης που έγινε από ένα μοντέλο F^*. Στον άξονα των X (οριζόντιος άξονας) βρίσκονται οι τιμές που έχουν σημειωθεί για τη μεταβλητή F και στον άξονα των Y (κατακόρυφος άξονας) βρίσκονται οι προβλεπόμενες τιμές μέσω ενός μοντέλου (π.χ., μία εξίσωση). Η ευθεία στη διαγώνιο (identity line, $y = x$, βλ. Σχ. 22b), βρίσκεται για να προσδιορίσει την επιτυχία του μοντέλου στην πρόβλεψη. Ο τρόπος που λειτουργεί το γράφημα αυτό, είναι με το να αναπαριστά ένα σημείο με συντεταγμένες (X, Y) όπου η τιμή του X είναι η πραγματική τιμή μίας μεταβλητής και η τιμή του Y είναι η πρόβλεψη που πραγματοποιήθηκε από το εκάστοτε μοντέλο. Εάν αυτό το σημείο παρατηρηθεί επάνω στην ευθεία της διαγώνιου τότε σημαίνει ότι $X = Y$, άρα το μοντέλο έχει κάνει αποτελεσματική πρόβλεψη.

Για παράδειγμα, στο Σχ. 22c έχουν σημειωθεί 3 σημεία και η γραμμή ταυτότητας. Εάν για μία μεταβλητή F η πραγματική τιμή της αντιστοιχεί στην τιμή 4 και μέσω ενός μοντέλου εκτιμηθεί ότι η πραγματική τιμή είναι 4 ($F^* = 4$), τότε θα πρέπει το σημείο αυτό (4,4) να βρίσκεται πάνω στην ευθεία $y = x$ όπως φαίνεται για το σημείο με αριθμό 1. Αντίθετα, εάν η πραγματική τιμή για τη μεταβλητή $F = 4$ και μέσω ενός μοντέλου, η τιμή εκτιμηθεί $F^* = 5$, τότε το μοντέλο έχει υπερεκτιμήσει την τιμή της μεταβλητής F και δεν βρίσκεται πάνω στη διαγώνιο. Πιο συγκεκριμένα, θα βρίσκεται στο σημείο (4,5), πάνω από τη διαγώνιο δηλαδή, όπως και συμβαίνει για το σημείο με αριθμό 2. Τέλος, εάν μέσω ενός μοντέλου, η τιμή



Σχήμα 22: Παραδείγματα γραφημάτων για τις σχέσεις ανάμεσα στη μεταβλητή $\boldsymbol{F}$ και την $\boldsymbol{F}^*$. (a) Γράφημα διασποράς (scatter plot). (b) Γραμμή ταυτότητας (identity line). (c) Γράφημα ταυτότητας (identity plot).

εκτιμηθεί $F^* = 3$, ενώ η πραγματική τιμή είναι $F = 4$, τότε το μοντέλο έχει υποεκτιμήσει την τιμή της μεταβλητής F, δεν βρίσκεται πάνω στην ευθεία $y = x$, θα βρίσκεται στο σημείο (4,3) και θα φαίνεται όπως το σημείο με αριθμό 3. Στην πράξη, δε μας ενδιαφέρει εάν το μοντέλο έχει υποεκτιμήσει ή υπερεκτιμήσει τη συγκεκριμένη μεταβλητή, μας ενδιαφέρει εάν την έχει προβλέψει επιτυχώς ($y = x$) ή ανεπιτυχώς ($y \neq x$), καθώς και σε περίπτωση αποτυχίας το μέγεθος της απόκλισης.

Για κάθε μία από τις ιδιότητες μεταφοράς, τα γραφήματα ταυτότητας της κάθε εξίσωσης βρίσκονται στο παράρτημα, ενώ παρακάτω χρησιμοποιήθηκαν κάποια από αυτά. Αξίζει να σημειωθεί πως σε κάθε γράφημα ταυτότητας, τα δεδομένα που χρησιμοποιήθηκαν για εκπαίδευση του μοντέλου φαίνονται με πορτοκαλί χρώμα, ενώ εκείνα που χρησιμοποιήθηκαν για έλεγχο με μπλε.

5.2.1 Συντελεστής Διάχυσης

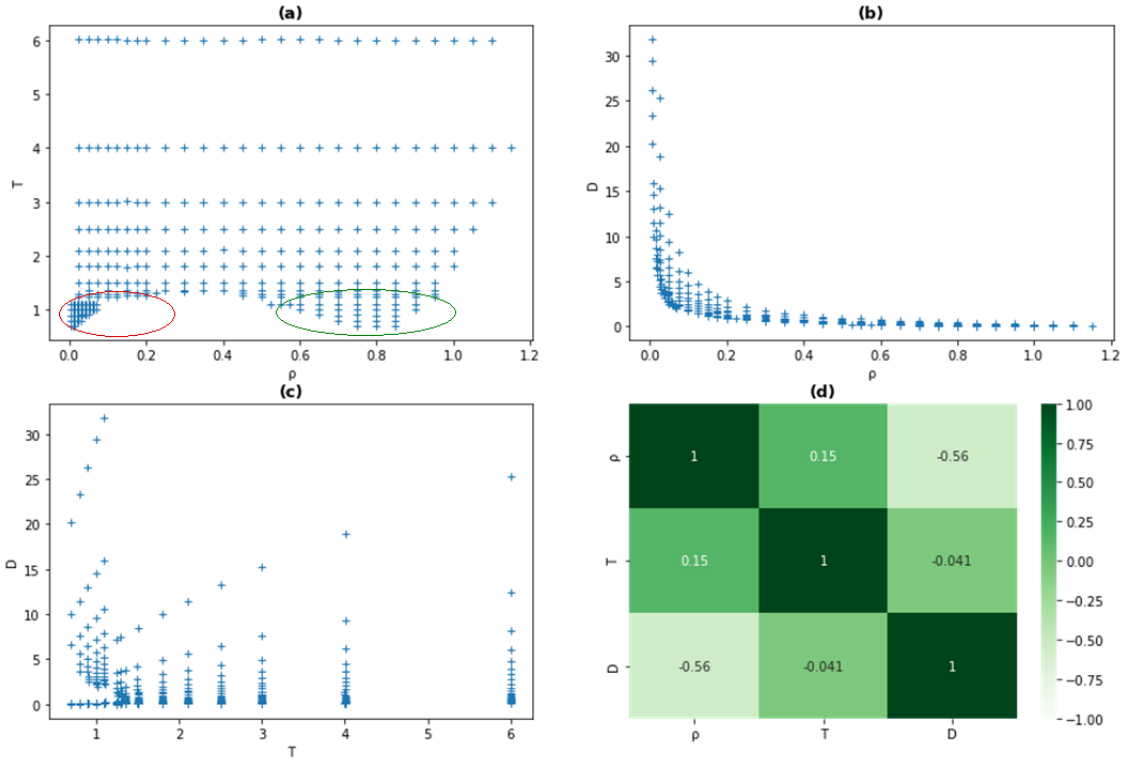
Το σετ δεδομένων για τον συντελεστή διάχυσης D αποτελείται από 315 παρατηρήσεις, οι οποίες βρίσκονται σε όλες τις καταστάσεις του ρευστού (αέρια (gas), υγρή (liquid) και υπερκρίσιμη (supercritical) μορφή). Προτού γίνει η εξαγωγή εξισώσεων, επιδιώκεται μία διερεύνηση στο σετ δεδομένων για να εμφανίσει τυχόν σχέσεις μεταξύ των δεδομένων.

Από το Σχ. 23, αυτά που παρατηρούμε, είναι:

- **Σχήμα a**: Τα δεδομένα βρίσκονται σε όλες τις καταστάσεις του ρευστού. Η κόκκινη περιοχή αντιστοιχεί στην αέρια μορφή (ατμός), η πράσινη περιοχή στην υγρή μορφή, ενώ όλα τα υπόλοιπα στην υπερκρίσιμη περιοχή (super critical) (Papastamatiou et al., 2022).
- **Σχήμα b**: Για σταθερή θερμοκρασία T, παρατηρείται ότι ο συντελεστής διάχυσης D είναι αντιστρόφως ανάλογος της πυκνότητας ρ.
- **Σχήμα c**: Για σταθερή πυκνότητα ρ, παρατηρείται ότι ο συντελεστής διάχυσης D είναι γραμμικά εξαρτημένος της θερμοκρασίας T.
- **Σχήμα d**: Ενώ οι μεταβλητές T και ρ δεν είναι συσχετισμένες σε ένα μεγάλο βαθμό, παρατηρείται μία αρνητική συσχέτιση μεταξύ του συντελεστή διάχυσης D και της πυκνότητας ρ, δηλαδή μικρή πυκνότητα (πιο κοντά στην αέρια κατάσταση) σημαίνει μεγάλη τιμή για τον συντελεστή διάχυσης, ενώ μεγάλη πυκνότητα (πιο κοντά στην στερεά κατάσταση) σημαίνει μικρή τιμή για τον συντελεστή διάχυσης. Η αρνητική συσχέτιση ανάμεσα στις μεταβλητές $D - \rho$, ήταν αναμενόμενη λόγω της φυσικής τάσης που υπάρχει στο σύστημα να ελαττώνεται η κινητικότητα των ατόμων καθώς αυξάνεται η πυκνότητα (Papastamatiou et al., 2022).

Οι εξαγόμενες εκφράσεις του Πίνακα 8, έχουν ως στόχο τη γενίκευση της συμπεριφοράς του συντελεστή διάχυσης έναντι όλων των καταστάσεων του ρευστού. Οι εξισώσεις αυτές, είναι εκείνες όπου από μία συνεχή βελτίωση, βρέθηκαν στην καμπύλη Pareto. Οι σχέσεις που παρατηρήθηκαν ανάμεσα στα δεδομένα, βοηθούν έως ένα βαθμό την ερμηνεία των εξισώσεων, έτσι ώστε να γίνει η επιλογή σε εκείνες που εμφανίζουν για παράδειγμα αρνητική συσχέτιση του συντελεστή διάχυσης με την πυκνότητα και γραμμική εξάρτηση του συντελεστή διάχυσης με τη θερμοκρασία.

Από τον Πίνακα 8, βλέπουμε ότι κυριαρχούν δύο μορφές εξισώσεων, με χαρακτηριστικές περιπτώσεις τις εξισώσεις F_8 και F_1. Παρά την τυχαιότητα που επικρατεί στη διαδικασία, μία θετική συσχέτιση ανάμεσα στις μεταβλητές $D - T$ και μία αρνητική συσχέτιση ανάμεσα στις μεταβλητές $D - \rho$, παρατηρήθηκε σε όλες τις εξισώσεις, ενώ οι διαφορές που



Σχήμα 23: Συμπεριφορά των δεδομένων συντελεστή διάχυσης $\boldsymbol{D}$. (a) Διάγραμμα φάσης $\boldsymbol{\rho - T}$. Η κόκκινη περιοχή απεικονίζει την αέρια κατάσταση (gas), η πράσινη περιοχή την υγρή (liquid) ενώ όλη η άλλη περιοχή είναι η υπερκρίσιμη (super critical). (b) Διάγραμμα $\boldsymbol{\rho - D}$. (c) Διάγραμμα $\boldsymbol{T - D}$. (d) Τιμές συντελεστή συσχέτισης για τις μεταβλητές $\boldsymbol{\rho, T, D}$.

παρατηρούνται έχουν να κάνουν με τις μεταβολές στις τιμές των σταθερών. Ο μικρός συντελεστής της θερμοκρασίας, υποδηλώνει τη μικρή εξάρτηση του συντελεστή διάχυσης από αυτή. Πράγμα το οποίο περιμέναμε, λόγω του χαμηλού συντελεστή συσχέτισης που παρατηρήθηκε στο Σχ. 23d.

Σε όλες τις εξισώσεις του Πίνακα 8, παρατηρείται αύξηση στην τιμή του συντελεστή διάχυσης καθώς μειώνεται η τιμή της πυκνότητας. Αυτό με τη σειρά του εξυπηρετεί τη φυσική συμπεριφορά του συστήματος, καθώς με τη μείωση της πυκνότητας και προσέγγιση της αέριας κατάστασης, είναι αναμενόμενο η τιμή του συντελεστή διάχυσης να αυξάνεται (Papastamatiou et al., 2022). Επιπρόσθετα, για το λόγο ότι η θερμοκρασία σε όλες τις εξισώσεις φαίνεται να είναι ανάλογη του συντελεστή διάχυσης, εύκολα βλέπουμε ότι η μείωση στη θερμοκρασία θα επιφέρει μείωση στο συντελεστή διάχυσης. Το φαινόμενο αυτό είναι απόλυτα λογικό, διότι καθώς η θερμοκρασία πλησιάζει το μηδέν, η κίνηση των σωματιδίων ελαχιστοποιείται έως τη θερμοκρασία $T = 0$ όπου δεν υπάρχει κίνηση των σωματιδίων. Έτσι λοιπόν, βλέπουμε ότι όλες οι εξισώσεις αποτυπώνουν τη φυσική συμπεριφορά του

συστήματος. Βέβαια, το ερώτημα που προκύπτει είναι ως προς το ποια εξίσωση θα ήταν πιο αποτελεσματική να επιλέξουμε.

Από τα στοιχεία του Πίνακα 9, όλες οι τιμές που αντιστοιχούν στο R^2 είναι ικανοποιητικές. Στα μέσα σφάλματα που παρατηρούμε από την άλλη υπάρχουν αποκλίσεις μεταξύ των εξισώσεων. Για παράδειγμα, αν και στις εξισώσεις F_1 και F_8 το R^2 που παρατηρείται είναι υψηλό, το μέσο σφάλμα που παρατηρείται στην F_1 είναι σχεδόν 7 φορές το μέσο σφάλμα της F_8. Αυτό έχει ως αποτέλεσμα τη χαμηλότερη απόδοση της Εξ. F_1 έναντι της F_8 σε θέματα όπου η ακρίβεια των υπολογισμών είναι το κύριο μέλημα. Από την άλλη μεριά, η Εξ F_8 παρουσιάζει μικρό μέσο σφάλμα σε σύγκριση με τις υπόλοιπες εξισώσεις του Πίνακα 9 και το R^2 είναι αρκετά ικανοποιητικό.

Πίνακας 8: Εκφράσεις συντελεστή διάχυσης $\boldsymbol{D}$ για όλες τις καταστάσεις του ρευστού.

Αριθμός	**Μαθηματική εξίσωση**
F_1	$0.131965165 * \frac{T}{\rho} - 0.12563$
F_2	$0.131965257 * \frac{T}{\rho} - 0.12563$
F_3	$0.131961368 * \frac{T}{\rho} - 0.12563$
F_4	$\frac{0.0363038+0.106631*T}{\rho} - 0.0982058$
F_5	$0.0131963691 * \frac{T}{\rho} - 0.12563$
F_6	$\frac{(0.0363062 + 0.106632\ T)}{\rho} - 0.098209$
F_7	$0.13196199 * \frac{T}{\rho} - 0.12563$
F_8	$\frac{0.0672683+0.233244*T}{\rho} + 0.0716851$
F_9	$0.131964486 * \frac{T}{\rho} - 0.12563$

Πίνακας 9: Αποτελέσματα στατιστικών μέτρων της κάθε εξίσωσης του Πίνακα 8 στα δεδομένα.

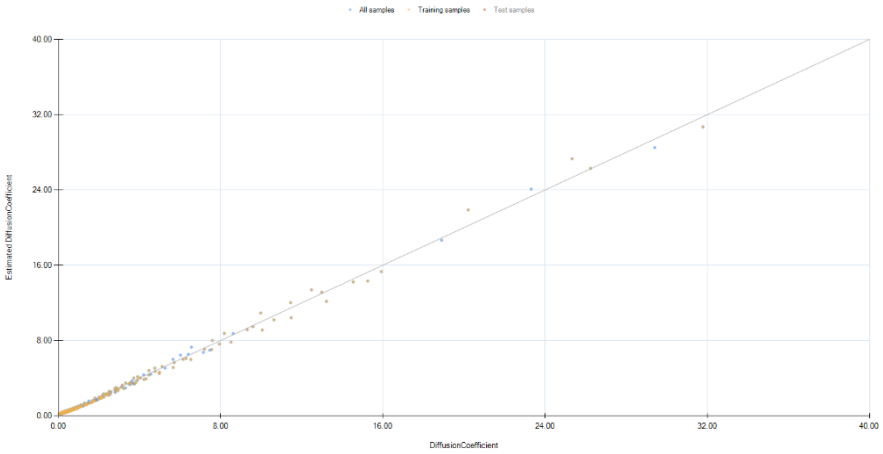
Αριθμός	$\boldsymbol{R^2}$ **(train)**	$\boldsymbol{R^2}$ **(test)**	**MSE (train)**	**MSE (test)**
F_1	0.97719807774222467	0.98641383042100028	0.41326224202057127	0.4514671130096824
F_2	0.97719807774222467	0.98641383042100028	0.4132622420205711	0.45146711300968312
F_3	0.97719807774222467	0.98641383042100006	0.41326224202057138	0.45146711300967773

Αριθμός	R^2 (train)	R^2 (test)	MSE (train)	MSE (test)
F_4	0.99476553143503033	0.99716450557825886	0.094869554877451245	0.078112859818194672
F_5	0.97719807774222467	0.98641383042100006	0.41326224202057155	0.45146711300968162
F_6	0.99476553143503033	0.99716450557825931	0.09486955487745112	0.078112859818195435
F_7	0.97719807774222422	0.98641383042099984	0.41326224202057155	0.4514671130096804
F_8	0.99547148675373331	0.99776090021096731	0.082074814395671075	0.062025551574790352
F_9	0.97719807774222467	0.98641383042100028	0.41326224202057149	0.45146711300968412

Σε όλες τις εξαγόμενες σχέσεις του Πίνακα 8, παρατηρείται η πρόσθεση ενός αρνητικού συντελεστή. Αυτό έχεις ως αποτέλεσμα, σε κάποιους συνδυασμούς τιμών $\rho - T$, η τιμή του D να είναι αρνητική ενώ στο σετ δεδομένων υπάρχουν μόνο θετικές τιμές. Απορρίπτοντας τις εξισώσεις $F_i, i = (1,2,3,5,7,9)$, απαλείφονται οι περιπτώσεις εμφάνισης αρνητικού συντελεστή διάχυσης (στο παράρτημα παρατίθεται σχετικό σχήμα όπου παρουσιάζονται οι εξισώσεις που πρόβλεψαν για κάποιες τιμές αρνητικό συντελεστή διάχυσης). Άρα, από τις εναπομείναντες εξισώσεις $F_j, j = (4,6,8)$ προσπαθούμε να βρούμε την καλύτερη δυνατή. Παρατηρώντας τις τιμές των εξισώσεων για το R^2 βλέπουμε ότι είναι ικανοποιητικό για όλες και επιπλέον οι αποκλίσεις στο μέσο σφάλμα είναι μικρές σε σχέση με τις υπόλοιπες. Η Εξ. F_8 όμως είναι εκείνη που παρουσιάζει το μικρότερο δυνατό μέσο σφάλμα, συνοδευόμενη από το μεγαλύτερο δυνατό R^2 και αποτελεί την εξίσωση που θα εστιάσουμε.

Στο Σχ. 24 φαίνεται το γράφημα ταυτότητας για την Εξ. F_8. Αυτό που παρατηρούμε, είναι πως το μοντέλο έδειξε ευελιξία προσεγγίζοντας ικανοποιητικά τόσο τα δεδομένα εκπαίδευσης όσο και τα «άγνωστα» δεδομένα ελέγχου. Επιπλέον, στις χαμηλές τιμές του συντελεστή διάχυσης η προβλέψεις του μοντέλου είναι ακριβής, ενώ από ένα σημείο και μετά εμφανίζονται αποκλίσεις και στα δύο είδη δεδομένων. Οι αποκλίσεις στα δεδομένα σημαίνουν ότι το χρησιμοποιημένο μοντέλο της Εξ. F_8 δεν είναι ικανό να περιγράψει εξ ολοκλήρου τη συμπεριφορά του συντελεστή διάχυσης στο σετ δεδομένων που χρησιμοποιήθηκε. Μία πιο περίπλοκη εξίσωση, θα είναι σε θέση να ταιριάξει καλύτερα στα δεδομένα και να παρουσιάσει πιο ακριβή περιγραφή της συμπεριφοράς του συντελεστή διάχυσης.

Βέβαια, οι απλουστευμένες μορφές των εξισώσεων του Πίνακα 8 μπορούν εύκολα να γενικευτούν καθώς μία απλή εξίσωση είναι λιγότερο πιθανό να είναι προϊόν υπερπροσαρμογής. Συγκρίνοντας τις εξισώσεις του Πίνακα 8 που εκφράζουν το συντελεστή διάχυσης ως



Σχήμα 24: Γράφημα ταυτότητας (identity plot) της Εξ. $\boldsymbol{F_8}$ του Πίνακα 8.

συνάρτηση των μακροσκοπικών μεγεθών της θερμοκρασίας και της πυκνότητας, έναντι των εξισώσεων στατιστικής μηχανικής που χρησιμοποιούνται ευρέως (εξισώσεις 5.1-5.2), συμπεραίνουμε ότι η ικανοποιητική ακρίβεια των εξισώσεων συμβολικής παλινδρόμησης καθώς και η υπεραπλουστευμένη τους μορφή, τις καθιστούν ελκυστικές για την αξιοποίησή τους σε περισσότερα θέματα. Επίσης, η αποτύπωση της φυσικής συμπεριφοράς του συστήματος σε όλες τις καταστάσεις του ρευστού μέσω μακροσκοπικών μεγεθών καθιστά πιο εύκολη την κατανόηση του όλου συστήματος από τον εκάστοτε ερευνητή, από ότι μία εξίσωση με μικροσκοπικά μεγέθη που έχουν να κάνουν με ταχύτητες, θέσεις και δυναμικά σωματιδίων.

5.2.2 Ιξώδες διάτμησης

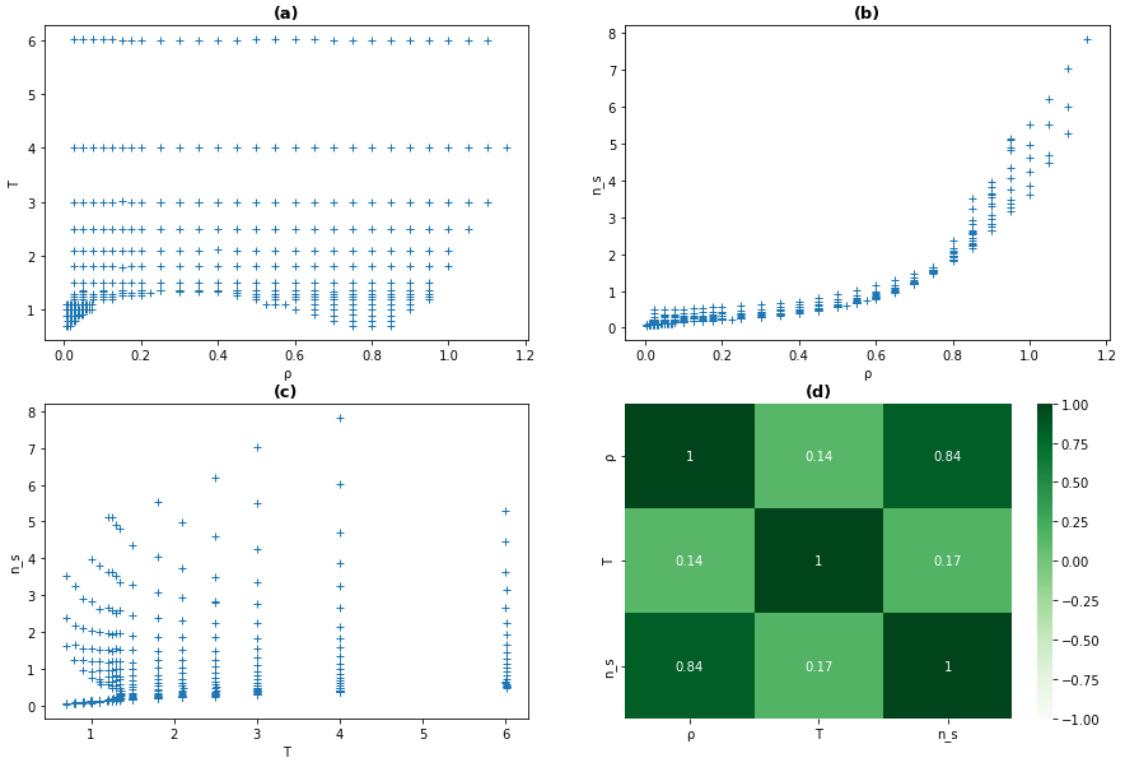
Το σετ δεδομένων για το ιξώδες διάτμησης n_s, αποτελείται από 339 παρατηρήσεις. Όμοια με προηγουμένως, επιχειρείται μία πρώτη προσέγγιση αποκρυπτογράφησης της συμπεριφοράς των δεδομένων.

Πιο συγκεκριμένα, βάσει του Σχ. 25:

- **Σχήμα a**: Τα δεδομένα καλύπτουν όλες τις μορφές του ρευστού (αέρια, υγρή και υπερκρίσιμη).
- **Σχήμα b**: Για σταθερή θερμοκρασία T, η σχέση ανάμεσα στο ιξώδες διάτμησης n_s και στην πυκνότητα ρ παρουσιάζει μικρή μεταβολή στις μικρές και μεσαίες τιμές της πυκνότητας ($\rho < 0.5$) ενώ για μεγαλύτερες τιμές στην πυκνότητα ($\rho > 0.5$) μία ραγδαία αύξηση στην τιμή του ιξώδους διάτμησης.
- **Σχήμα c**: Για σταθερή πυκνότητα ρ, παρατηρείται μία σχεδόν γραμμική σχέση ανάμεσα στα n_s και T.
- **Σχήμα d**: Ενώ οι μεταβλητές ρ και T φαίνεται να μην είναι σημαντικά συσχετισμένες, το ίδιο δεν ισχύει για τη σχέση ανάμεσα στο n_s και την ρ. Η μεγάλη τιμή του συντελεστή συσχέτισης ήταν αναμενόμενη αν συμπεριληφθεί η εικόνα της εξάρτησης που παρατηρήθηκε στο Σχ. 25b. Αυτό δηλώνει ότι ενώ για μικρές τιμές της πυκνότητας (κοντά στην αέρια κατάσταση) το ιξώδες διάτμησης παρουσιάζει μία χαμηλή τιμή, σε μεγάλες τιμές της πυκνότητας (κοντά στη στερεά κατάσταση) η τιμή από το ιξώδες διάτμησης θα είναι υψηλή.

Οι εξισώσεις που λήφθηκαν για την περιγραφή του ιξώδους διάτμησης n_s έναντι των μακροσκοπικών μεγεθών της πυκνότητας ρ και της θερμοκρασίας T, παρουσιάζονται στον Πίνακα 10. Επίσης, τα στατιστικά μέτρα της κάθε εξίσωσης στο σετ δεδομένων, φαίνονται στον Πίνακα 11. Κάποιες εξισώσεις φαίνονται πάρα πολύ απλές και πιθανώς να έχουν υποπέσει σε υποπροσαρμογή, άλλες αποτυπώνουν τις παραπάνω παραδοχές ενώ άλλες είναι λίγο πιο περίπλοκες σε μορφή.

Για παράδειγμα, η Εξ. F_2 είναι πάρα πολύ απλή, ενώ παράλληλα η μεταβλητή της θερμοκρασίας δεν εμφανίζεται καθόλου. Αυτή η εξίσωση παρουσιάζει το μικρότερο R^2 ενώ το μέσο σφάλμα της είναι το μεγαλύτερο από όλες, άρα απορρίπτεται. Οι εξισώσεις που παρουσιάζουν βελτιωμένα στατιστικά μέτρα έναντι των υπολοίπων, είναι εκείνες όπου παρατηρείται μία εκθετική εξάρτηση του ιξώδους διάτμησης από την πυκνότητα, συνοδευόμενο από

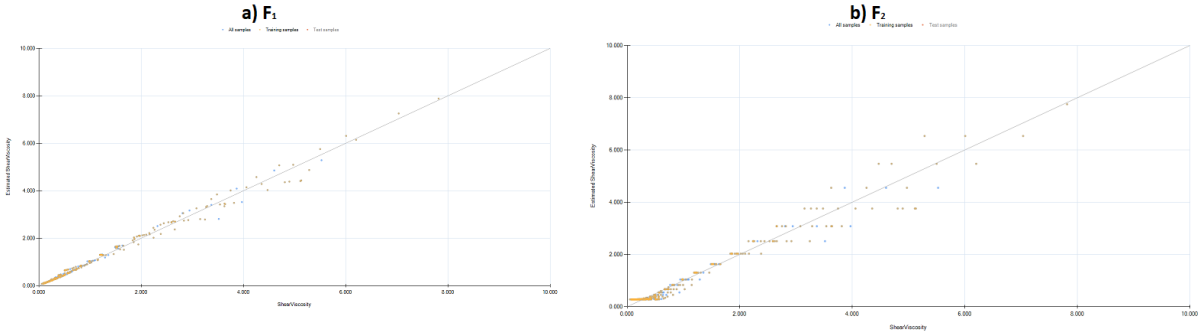


Σχήμα 25: Συμπεριφορά των δεδομένων. (a) Διάγραμμα φάσης $\boldsymbol{\rho - T}$. (b) Διάγραμμα $\boldsymbol{\rho - n_s}$. (c) Διάγραμμα $\boldsymbol{T - n_s}$. (d) Τιμές συντελεστή συσχέτισης για τις μεταβλητές $\boldsymbol{\rho, T, n_s}$.

μία μικρή εξάρτηση της θερμοκρασίας. Ένα χαρακτηριστικό παράδειγμα είναι η Εξ. F_1 του Πίνακα 10 όπου παρουσιάζει αρκετά καλές τιμές τόσο για το R^2 όσο και για το MSE. Επίσης, εξισώσεις όπου η πυκνότητα ρ είναι υψωμένη σε δύναμη, όπως για παράδειγμα οι εξισώσεις F_5 και F_6 παρουσιάζουν ικανοποιητικά αποτελέσματα. Αναμενόμενο γεγονός, λόγω της υψηλής συσχέτισης που παρατηρείται ανάμεσα στην πυκνότητα και το ιξώδες διάτμησης.

Τέλος, οι εξισώσεις F_7 και F_8 παρουσιάζουν ελάχιστα αυξημένη πολυπλοκότητα έναντι των υπολοίπων (παραμένουν όμως πιο ερμηνεύσιμες από τις εξισώσεις (5.3) και (5.4)) και ως συνέπεια τα στατιστικά μέτρα τους είναι καλύτερα από των υπολοίπων. Μένει να ελεγχθούν και τα γραφήματα ταυτότητας.

Στο Σχ. 26b, φαίνεται το γράφημα ταυτότητας για την Εξ. F_2. Παρατηρούμε ότι το μοντέλο δεν περιγράφει με ακρίβεια τα δεδομένα καθώς φαίνονται αρκετές αποκλίσεις. Γεγονός το οποίο περιμέναμε λόγω χαμηλής πολυπλοκότητας της εξίσωσης (απουσίαζε μέχρι και η μεταβλητή της θερμοκρασίας) και καταλήγουμε στην απόρριψή της. Αντίθετα, η ικανότητα μίας εξίσωση με μεγαλύτερη πολυπλοκότητα όπως η Εξ. F_1, δεν αποτυπώνεται μόνο στις καλύτερες τιμές των μέτρων ακριβείας αλλά και μέσω του Σχ. 26a καθώς φαίνεται μία καλύτερη



Σχήμα 26: Γράφημα ταυτότητας (identity plot) για την Εξ. $\boldsymbol{F_1}$ (a) και $\boldsymbol{F_2}$ (b) του Πίνακα 10.

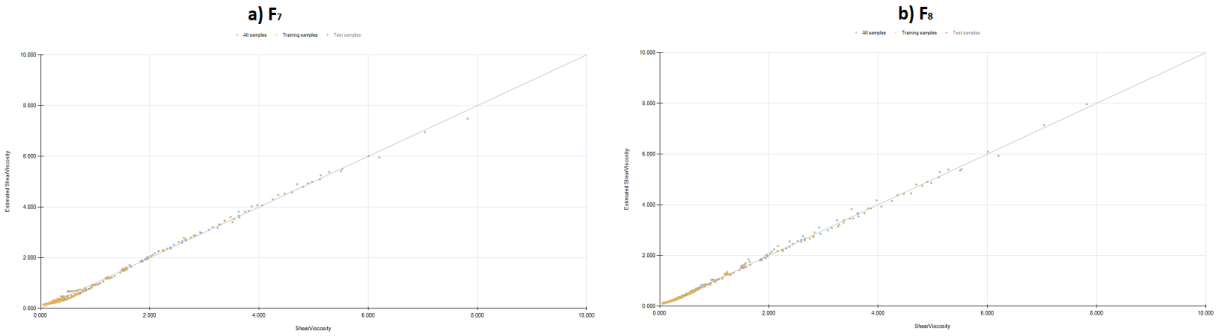
συμφωνία με τα δεδομένα. Για τιμές $n_s < 1$ παρατηρούμε ότι η ταύτιση είναι ικανοποιητική ενώ για $n_s > 1$ εμφανίζονται πιο τακτικές αποκλίσεις. Άρα αυτό που παρατηρούμε συγκρίνοντας τις δύο εξισώσεις, είναι ότι με την άνοδο της πολυπλοκότητας, αυξήθηκε και η ακρίβεια πρόβλεψης του μοντέλου. Τέλος, οι εξισώσεις F_7 και F_8 παρουσιάζουν υψηλές τιμές στο R^2 και χαμηλές τιμές στο MSE.

Συγκρίνοντας τα γραφήματα ταυτότητας για τις εξισώσεις F_7 και F_8 στο Σχ. 27, παρατηρούμε ότι για τιμές $n_s < 1$ και τα δύο μοντέλα έρχονται σε καλή συμφωνία με τα δεδομένα, με την Εξ. F_8 να παρουσιάζει ελάχιστα καλύτερη. Για τιμές $n_s > 1$ παρατηρείται μεγαλύτερο πλήθος αποκλίσεων, το οποίο σημαίνει ότι οι εξισώσεις αυτές δεν αποκαλύπτουν εντελώς τη συμπεριφορά από το ιξώδες διάτμησης και η πραγματική του συμπεριφορά θα περιγράφεται από μία πιο περίπλοκη εξίσωση. Ωστόσο, το γεγονός ότι παρουσιάζουν ένα πολύ υψηλό R^2 το οποίο συνοδεύεται από ένα χαμηλό MSE και έχουν μικρότερη πολυπλοκότητα από τις εξισώσεις (5.3) και (5.4), τις καθιστά ικανοποιητικές.

Επιπλέον, αξίζει να σημειωθεί το γεγονός ότι οι προσεγγίσεις πραγματοποιούνται σε ένα μικρό σετ δεδομένων (ένα κανονικό σετ ξεπερνά τις 1000-1500 τιμές στα δεδομένα) και χωρίς γνώση του υπό μελέτη συστήματος, αποτυπώνονται οι διασυνδέσεις ανάμεσα στις μεταβλητές με τη μορφή εξισώσεων, οι οποίες με τη σειρά τους είναι πιο εύκολα ερμηνεύσιμες από εκείνες που χρησιμοποιούσαν οι ερευνητές έως τώρα.

Πίνακας 10: Εκφράσεις ιξώδους διάτμησης n_s για όλες τις καταστάσεις του ρευστού.

Αριθμός	Μαθηματική εξίσωση
F_1	$-0.032115 + 0.0334553e^{(0.360302\rho(14.708-0.41177T))} + 0.105055T$
F_2	$0.27698 + 4.27481\rho^4$
F_3	$\frac{(-3.4218*\rho - 0.61477*T + T*\rho*0.72759 + 0.55182)}{e^{(-1.4863*\rho)}} * \frac{(-2.5727*\rho + e^{1.5779*\rho} + 0.0000001346)*(-0.17988)}{e^{-1.4863*\rho}} + 0.05454$
F_4	$0.18485 + 0.41432\rho + 2.78634\rho^4 + \frac{(2.96141\rho^7)}{(-0.00121249 + T)}$
F_5	$-\frac{(-0.391831 + \rho^5(-34.7762 + 0.147653T) - 1.0183T)}{(5.16084 + T)}$
F_6	$-0.16877 + 0.0041106(-6.0252 - 6.8735T - 190.983\rho^5)\log 0.000567964T^2$
F_7	$0.011353\,(\,33.366 - 0.9798\,T)\,(\,0.290025\,e^{\frac{7.66111\rho^4+0.185023}{T+0.872118}} + 9.47143\,\rho^4 - 9.2732) + 3.4697$
F_8	$-0.049116 + e^{(2.8158\rho)}\left(0.0190059 - 0.21252\rho + \frac{(0.246438\rho^6)}{T} + 0.0155619T\right)$
F_9	$\frac{\left(4.20135(-0.00306634 + 0.0592015T + \rho^4T)\right)}{(-0.0987941 + T)}$



Σχήμα 27: Γραφήματα ταυτότητας (identity plots) για τις εξισώσεις $\boldsymbol{F_7}$ (Σχ. a) και $\boldsymbol{F_8}$ (Σχ. b) του Πίνακα 10.

Πίνακας 11: Αποτελέσματα στατιστικών μέτρων της κάθε εξίσωσης του Πίνακα 10 στα δεδομένα.

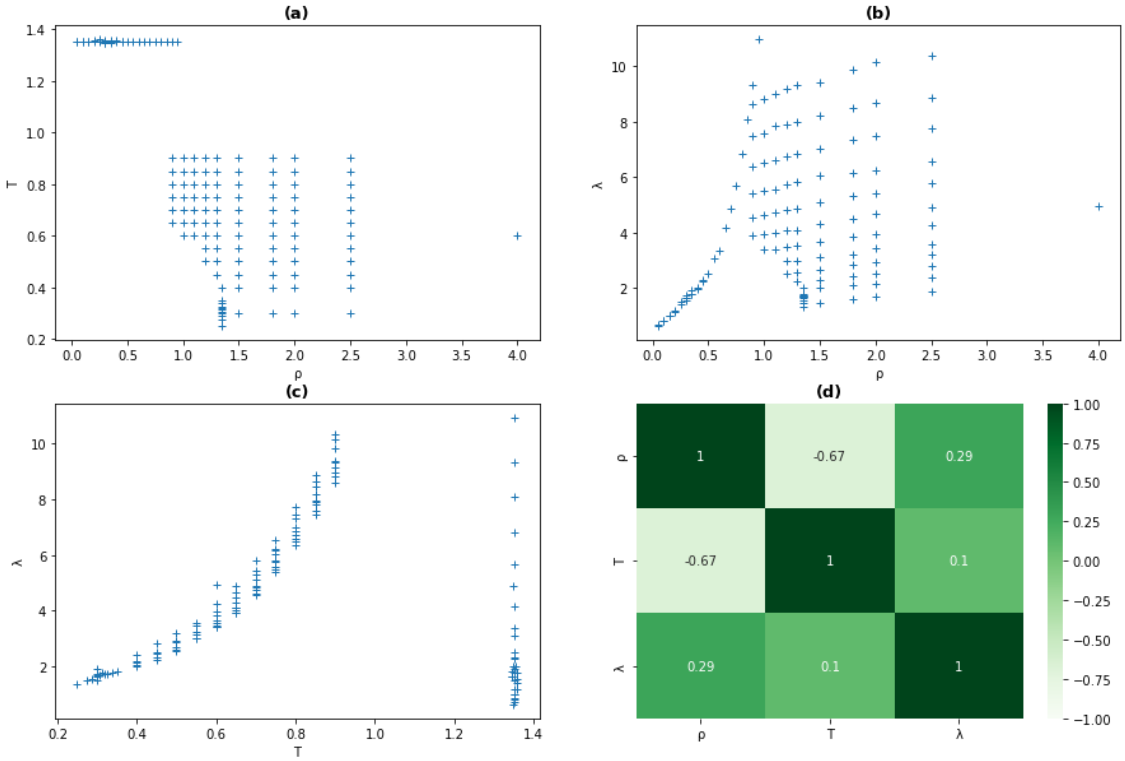
Αριθμός	R^2 (train)	R^2 (test)	MSE (train)	MSE (test)
F_1	0.99164294631206529	0.987807320994339	0.016269123055019864	0.016730456634206595
F_2	0.96251309628358939	0.95511854237652039	0.072977758943258031	0.065076787875294057
F_3	0.98122591742071208	0.97125829684669118	0.036548509933411182	0.040238025611963435
F_4	0.98849085605403275	0.984701888429639	0.022405465623036476	0.021046547031746882
F_5	0.98809972324309048	0.979874759558822	0.023166904770092159	0.027944475308802647
F_6	0.98999397628269659	0.98456512381675865	0.019479261139997234	0.021401075881335625
F_7	0.99710611047713693	0.99535954003726379	0.005744179190216761	0.0061332836602836895
F_8	0.99826499888254028	0.99755831697325936	0.0034438623987476328	0.0033422557523709196
F_9	0.96955501624236751	0.9711672359429715	0.060431277961905255	0.0549520121162562

5.2.3 Θερμική αγωγιμότητα

Για τον υπολογισμό της θερμικής αγωγιμότητας, το σετ δεδομένων που χρησιμοποιήθηκε αποτελούταν από 130 μετρήσεις. Ακολουθώντας την ίδια μέθοδο με τις προηγούμενες δύο ιδιότητες μεταφοράς, επιχειρείται η αναγνώριση τυχόν σχέσεων στα δεδομένα. Βάσει του Σχ. 28:

- **Σχήμα a**: Λόγω του μικρού σετ δεδομένων, δεν παρατηρούνται μετρήσεις σε όλες τις μορφές του ρευστού. Ένα ποσοστό παρατηρήσεων φαίνεται να βρίσκεται στην υγρή περιοχή και ένα άλλο στην υπερκρίσιμη.
- **Σχήμα b**: Για σταθερό T και μικρές τιμές της πυκνότητας ($\rho < 1$), δηλαδή πιο κοντά στην αέρια κατάσταση, παρατηρείται μία δυνατή εξάρτηση ανάμεσα στην πυκνότητα και τη θερμική αγωγιμότητα καθώς για μικρές αλλαγές της τιμής της πυκνότητας, η τιμή της θερμικής αγωγιμότητας εκτοξεύεται, ενώ για μεγαλύτερες τιμές ($\rho > 1$), δηλαδή πιο κοντά στη στερεά κατάσταση, μία γραμμική εξάρτηση.
- **Σχήμα c**: Για σταθερό ρ και μικρές τιμές της θερμοκρασίας ($T < 1$) παρατηρείται μία εκθετική αύξηση στη θερμική αγωγιμότητα.
- **Σχήμα d**: Η θερμοκρασία T και η πυκνότητα ρ φαίνεται να μην είναι σημαντικά συσχετισμένες με τη θερμική αγωγιμότητα. Αντίθετα, παρατηρείται έως ένα βαθμό μία αρνητική συσχέτιση ανάμεσα στη θερμοκρασία και την πυκνότητα.

Το γεγονός ότι η θερμική αγωγιμότητα αποτελείται από ένα πολύ μικρό σετ δεδομένων, καθιστά πιο δύσκολη τη προσέγγιση ενός μοντέλου, ικανό να την περιγράψει. Οι εξισώσεις που προέκυψαν για τη θερμική αγωγιμότητα φαίνονται στον Πίνακα 12. Βάσει του Πίνακα 13, προσπαθούμε να προσανατολιστούμε στις εκφράσεις που εμφανίζουν μικρότερο σφάλμα. Η Εξ. F_7 παρουσιάζει απαγορευτικά μέτρα ακρίβειας τόσο για το MSE όσο και για το R^2 και απορρίπτεται ως επιλογή. Επίσης η Εξ. F_8 παρουσιάζει ικανοποιητικά MSE και R^2



Σχήμα 28: Συμπεριφορά των δεδομένων. (a) Διάγραμμα φάσης $\boldsymbol{\rho - T}$. (b) Διάγραμμα $\boldsymbol{\rho - \lambda}$. (c) Διάγραμμα $\boldsymbol{T - \lambda}$. (d) Τιμές συντελεστή συσχέτισης για τις μεταβλητές $\boldsymbol{\rho, T, \lambda}$.

όσον αφορά τα δεδομένα εκπαίδευσης, αλλά το ίδιο δεν ισχύει για τα δεδομένα ελέγχου, καθώς τα μέτρα ακρίβειας μας οδηγούν στην απόρριψή της. Οι εξισώσεις F_1 και F_2 εμφανίζουν μία υψηλή εξάρτηση της θερμικής αγωγιμότητας από τη θερμοκρασία, καθώς εμφανίζονται οι όροι T^{15} και T^{26} στις αντίστοιχες εξισώσεις. Από τη μία μεριά η τιμή του συντελεστή της θερμοκρασίας είναι μικρή, πράγμα που αποτυπώνει τη μικρή συσχέτιση μεταξύ των μεταβλητών λ και T όπως φαίνεται στο Σχ. 28. Από την άλλη μεριά, για το λόγο ότι η θερμοκρασία είναι υψωμένη σε μία μεγάλη δύναμη (όπως είδαμε 15 και 26 αντίστοιχα), έχει ως αποτέλεσμα η παραμικρή μεταβολή στη θερμοκρασία να επηρεάσει ανάλογα ή αντιστρόφως ανάλογα τη θερμική αγωγιμότητα, παρά το μικρή τιμή του συντελεστή της θερμοκρασίας. Έτσι, θα περιμέναμε μία συσχέτιση κοντά στην τιμή 1 ή -1 στο Σχ. 28d. Πράγμα το οποίο δεν ισχύει και οι εξισώσεις απορρίπτονται.

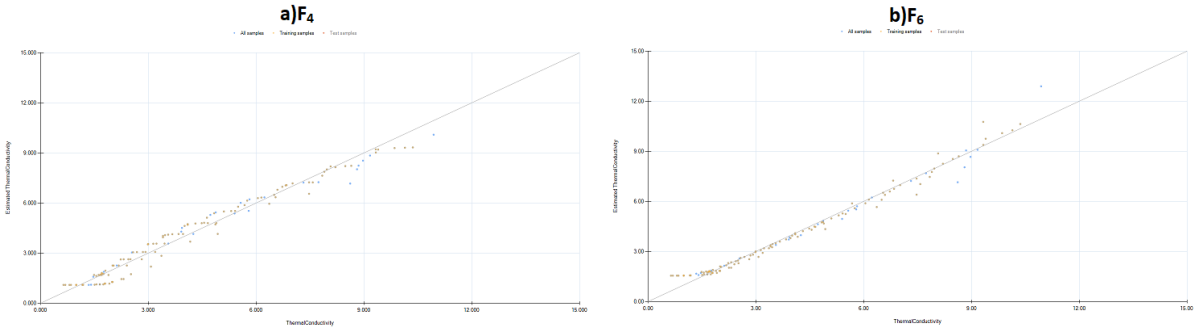
Το παράδοξο όμως, είναι ότι οι εξισώσεις F_1 και F_2 εμφανίζουν μικρότερα στατιστικά μέτρα από κάποιες άλλες. Εάν η επιλογή της τελικής εξίσωσης γινόταν μόνο με βάση το σφάλμα που παρουσιάζουν, θα έπρεπε να της δεχθούμε ως πιθανή επιλογή. Αντίθετα, εφαρμόζοντας τέσσερα κριτήρια, οδηγηθήκαμε στην απόρριψή τους.

Εάν εστιάσουμε στα μέτρα ακριβείας τόσο για τα δεδομένα εκπαίδευσης όσο και για τα δεδομένα ελέγχου, οι εξισώσεις οι οποίες φαίνεται να είναι πιο κοντά στην εκτίμηση της θερμικής αγωγιμότητας είναι οι F_4 και F_6. Τα MSE που παρουσιάζουν και οι δύο είναι τα μικρότερα σε σύγκριση με τις υπόλοιπες εξισώσεις, ενώ ταυτόχρονα το R^2 έχει μία ικανοποιητική τιμή. Βέβαια, η Εξ. F_4 φαίνεται υπεραπλουστευμένη ενώ η μεταβλητή της θερμοκρασίας δεν εμφανίζεται καθόλου. Για το λόγο ότι το σετ δεδομένων είναι μικρό και δεν έχουμε αρκετή γνώση για το σύστημα, μπορούμε να δεχθούμε την Εξ. F_4 επειδή εμφανίζει καλύτερα αποτελέσματα από άλλες εξισώσεις. Συγκρίνοντας τα γραφήματα ταυτότητας των δύο εξισώσεων (βλ. Σχ. 29), παρατηρούμε ότι αποκλίσεις υπάρχουν και στα δύο μοντέλα. Στις τιμές 1.5 – 7.5 της θερμικής αγωγιμότητας λ, παρατηρούμε ότι η Εξ. F_6 έχει κάνει καλύτερη προσέγγιση διότι εμφανίζει λιγότερες αποκλίσεις έναντι της Εξ. F_4. Βέβαια για τιμές $\lambda > 7.5$, αποκλίσεις είναι φανερές και στα δύο μοντέλα. Συνεπώς, ερευνώντας συλλογικά τις εξισώσεις του Πίνακα 12, πιο αποτελεσματική φαίνεται να είναι η Εξ. F_6 αν και πρέπει να σημειωθεί πως οι προσεγγίσεις της συμβολικής παλινδρόμησης έγιναν με ελάχιστη γνώση, χρησιμοποιώντας μόλις 130 μετρήσεις.

Η ικανότητα του μοντέλου πρόβλεψης για τη θερμική αγωγιμότητα λ, δεν είναι της ίδιας δυναμικής με των δύο προηγούμενων ιδιοτήτων μεταφοράς D και n_s. Ωστόσο, χρησιμοποιώντας ελάχιστη πληροφορία η προσέγγιση δεν αξιολογείται κακή. Αξίζει να σημειωθεί, πως και στις τρεις περιπτώσεις (D, n_s, λ) δημιουργήθηκαν μοντέλα χωρίς να αποδεχθούμε την οποιαδήποτε γνώση της συμπεριφοράς του συστήματος, δρώντας μόνο σε λιγοστές μετρήσεις. Τα μοντέλα που κατασκευάστηκαν δεν αποτελούν καθαυτού απόδειξη αποτύπωσης της συμπεριφοράς της εκάστοτε ιδιότητας μεταφοράς, αλλά λόγω αποτύπωσης της φυσικής συμπεριφοράς του συστήματος και των τιμών των στατιστικών μέτρων ακριβείας, οι εκτιμήσεις τους είναι ικανοποιητικές.

Πίνακας 12: Εκφράσεις θερμικής αγωγιμότητας $\boldsymbol{\lambda}$.

Αριθμός	Μαθηματική εξίσωση
F_1	$1.479 + \frac{2.6011 * T}{-1.7905 + 0.181153 * T + \frac{0.00145662 * T^{15}}{\rho^5}}$
F_2	$1.18938 + \frac{1}{\left(0.020492 + 0.00132161 log^2\left(e^{(8.71226\rho)} + 0.191047 T^{26}\right)\right)}$
F_3	$\frac{1}{\left(2.5427 * T - 1.6501 * \rho + \frac{1.9545 * \rho}{(0.9398 * T)^2} + (0.81679 * T)^6 * 0.50201 - 1.5442\right) * \left(\frac{1}{-0.000019696 * \rho} + 6.3463\right) * (-0.000013332)}$
F_4	$1.0906 + 0.000393888\rho^6$
F_5	$\left(\frac{714.93 * \rho - 46.254}{3.6207 * T - e^{(1.3027 * \rho)} + e^{(1.2419 * T)} + \frac{(1.9644 * \rho - 0.22024) * (0.94414 * \rho - 1.4896 * T)}{-0.1253 * T} + 11.078} + 1.8015\right) * 0.068486 * \rho + 0.5442$
F_6	$\frac{(1.77818 * 10^{-6} + 8.80732 * 10^{-6} * \rho) * \rho^2 * T^3}{\rho^2 * 0.90782 * \rho * T^2 + 0.979298 T^3} + 1.5466$
F_7	$4.4647 * \rho * T + 0.29764$
F_8	$\frac{0.72436 * T(\rho(0.244056 - 0.956311 * T) + \rho^2(7.33751 * T - 1.87257) - 0.646252)}{(\rho - 0.732042 * T + 0.0300181) log(\rho) - 0.115508 * \rho + 0.395653 * T + 0.212666}$
F_9	$\frac{0.93594 + (1915500 T^3 \rho^2)}{\left(8023(-3.55859 T^3 + 16.837 T^5 + (16.7705 + T^2)\rho^2)\right)}$



Σχήμα 29: Γραφήματα ταυτότητας (identity plots) για τις εξισώσεις. F_4 και F_6 του Πίνακα 12.

Πίνακας 13: Αποτελέσματα στατιστικών μέτρων της κάθε εξίσωσης του Πίνακα 12 στα δεδομένα.

Αριθμός	R^2 (train)	R^2 (test)	MSE (train)	MSE (test)
F_1	0.97051899958003429	0.97646572639987006	0.18937557910809061	0.18789571509589437
F_2	0.95614424665058206	0.95161719767803776	0.28171393675442835	0.4818998750205743
F_3	0.95647922552471565	0.91876942977752318	0.27956215026908809	0.70201681686550144
F_4	0.97279738125489257	0.97340898192518721	0.17474005646778135	0.255040906163639
F_5	0.94800358919414	0.95715271057666551	0.33400665742785496	0.36730754572310542
F_6	0.98359592247051153	0.96552015381072531	0.10537402522395795	0.28447486847750003
F_7	0.58749551354860308	0.38936873480112094	2.6497837554221637	5.404802029192795
F_8	0.96047480409478758	0.89783293703384	0.25389595866092296	0.90185645312072127
F_9	0.96742763860765768	0.91362016289139647	0.20923339485506032	0.73545780148965945

Κεφάλαιο 6 Συμπεράσματα και προτάσεις για περαιτέρω έρευνα

Η ένταξη των μεθόδων μηχανικής μάθησης στην επιστήμη των υλικών, ενίσχυσε τις προβλέψεις καθώς αυξήθηκε η ακρίβεια των προσεγγίσεων και μειώθηκε το υπολογιστικό κόστος. Βέβαια, σε αρκετά μοντέλα μηχανικής μάθησης (μοντέλα μαύρου κουτιού) το αποτέλεσμα δεν συνάγει με τη φανέρωση της φυσικής διεργασίας του συστήματος και ως συνέπεια παρουσιάζουν ελλιπείς απαντήσεις. Η συμβολική παλινδρόμηση από την άλλη, δεν υιοθετεί τους περιορισμούς που θέτουν τα παραπάνω μοντέλα και δημιουργεί μία εξίσωση η οποία είναι ικανή να φανερώσει τη φυσική σημασία του συστήματος. Επιπλέον, η ικανότητα να εξάγει εκφράσεις μόνο ελέγχοντας τα δεδομένα, μειώνει την ανάγκη αποδοχής κάποιου είδους γνώσης για το υπό μελέτη σύστημα. Τα χαρακτηριστικά αυτά, είναι ιδιαίτερα χρήσιμα σε περιπτώσεις όπου είναι δύσκολο να βρεθούν οι εξισώσεις που περιγράφουν ένα σύστημα ή υπάρχει μερική γνώση της συμπεριφοράς του ή υπάρχουν μόνο εμπειρικές σχέσεις.

Ο πιο σύνηθες τρόπος υλοποίησης της συμβολικής παλινδρόμησης, είναι μέσω του γενετικού προγραμματισμού. Εκεί, δημιουργείται ένα σύνολο εκφράσεων που παρουσιάζονται σε δεντρική μορφή, πραγματοποιείται ένας έλεγχος των εξισώσεων μέσω στατιστικών μέτρων και συνδυάζονται οι πιο ελπιδοφόρες, με σκοπό τη δημιουργία μίας «δυνατής» εξίσωσης, σε μία εξελικτική διαδικασία. Επίσης, δίνεται η δυνατότητα επιλογής της πολυπλοκότητας της εξίσωσης. Έτσι, ανάλογα με την επιθυμητή πολυπλοκότητα, θα επιλεχθεί διαφορετική εξίσωση από την καμπύλη Pareto.

Στην εφαρμογή του Κεφ. 5, προσεγγίστηκαν ιδιότητες μεταφοράς όπως ο συντελεστής διάχυσης, το ιξώδες διάτμησης και η θερμική αγωγιμότητα. Αξιοποιώντας μόνο παρατηρήσεις για τις παραπάνω ιδιότητες, κατασκευάστηκαν εξισώσεις πιο απλές από αυτές που χρησιμοποιούνται ευρέως, σε χαμηλό υπολογιστικό κόστος και συγκρίσιμη ακρίβεια. Επιπρόσθετα, οι εξισώσεις αυτές ήταν ικανές να περιγράψουν τη φυσική συμπεριφορά των ιδιοτήτων. Για το συντελεστή διάχυσης και το ιξώδες διάτμησης οι εξισώσεις που βρέθηκαν ήταν

ικανοποιητικές, ενώ για τη θερμική αγωγιμότητα λόγω του μικρότερου αριθμού παρατηρήσεων, η ακρίβεια δεν ήταν στα ίδια επίπεδα με των άλλων δύο. Οι δυνατότητες αυτές, ανοίγουν το δρόμο για την εφαρμογή της μεθόδου σε υπολογισμό περισσοτέρων ιδιοτήτων στα ρευστά.

Εξίσου σημαντικό, είναι το γεγονός ότι η μορφή των αποτελεσμάτων της συμβολικής παλινδρόμησης (συμβολική έκφραση), βοηθά την εισαγωγή τους σε άλλες μεθόδους όπως για παράδειγμα οι υπολογιστικές προσομοιώσεις. Στην εφαρμογή του Κεφ. 5, δημιουργήθηκε ένα μακροσκοπικό μοντέλο το οποίο είχε εκπαιδευτεί από δεδομένα προσομοιώσεων στη νανοκλίμακα. Έτσι, γίνεται πιο εύκολη η μετάβαση από τη μία κλίμακα στην άλλη, ενώ θα ήταν ιδιαίτερα ενδιαφέρον η περίπτωση σύνδεσης των υπολογιστικών προσομοιώσεων με των τεχνικών συμβολικής παλινδρόμησης. Σε τέτοια περίπτωση, τα θετικά στοιχεία των δύο τεχνικών θα μπορούσαν να συνδυαστούν σε μία κοινή, όπου οι υπολογιστικές προσομοιώσεις θα προμηθεύουν με υψηλής ποιότητας δεδομένα το μοντέλο μηχανικής μάθησης, για να παρουσιαστούν προβλέψεις υψηλότερης ακρίβειας που να αποτυπώνουν βέβαια τη φυσική σημασία του συστήματος.

Έως τώρα, όλες οι τεχνικές συμβολικής παλινδρόμησης που έχουν παρουσιαστεί είναι είτε ευριστικές (heuristic), είτε παρουσιάζονται αποτελεσματικές κάτω από συγκεκριμένες προϋποθέσεις. Ως συνέπεια, προσπάθειες επιχειρούνται για τη βελτιστοποίηση της μεθόδου. Ενδιαφέρον, θα είχε η προσπάθεια εξέλιξης της διαδικασίας της συμβολικής παλινδρόμησης, προκειμένου να αποκτήσει την ικανότητα ο υπολογιστής να εξάγει την συνολικά «καλύτερη» εξίσωση από άποψη ισορροπίας πολυπλοκότητας-ακρίβειας, πιθανώς με την εισαγωγή περισσότερων μέτρων ακρίβειας και κατά μία έννοια να αυτοματοποιηθεί η διαδικασία, προκειμένου να ελαχιστοποιηθεί η συμβολή του ανθρώπινου παράγοντα.

Βιβλιογραφία

Alam, T. M., Allers, J. P., Leverant, C. J., & Harvey, J. A. (2022). Symbolic regression development of empirical equations for diffusion in Lennard-Jones fluids. *The Journal of Chemical Physics*, *157*(1), 014503. https://doi.org/10.1063/5.0093658

Alexiadis, A. (2019). Deep multiphysics: Coupling discrete multiphysics with machine learning to attain self-learning in-silico models replicating human physiology. *Artificial Intelligence in Medicine*, *98*, 27–34. https://doi.org/10.1016/j.artmed.2019.06.005

Angelis, D., Sofos, F., & Karakasidis, T. (n.d.). Artificial Intelligence in physical sciences: Symbolic Regression trends and perspectives. *Submitted*.

Asadzadeh, M. Z., Gänser, H.-P., & Mücke, M. (2021). Symbolic regression based hybrid semiparametric modelling of processes: An example case of a bending process. *Applications in Engineering Science*, *6*, 100049. https://doi.org/10.1016/j.apples.2021.100049

Ayyadevara, V. K. (2018). Pro machine learning algorithms. *Apress: Berkeley, CA, USA*.

Back, T. (1996). *Evolutionary algorithms in theory and practice: Evolution strategies, evolutionary programming, genetic algorithms*. Oxford university press.

Behler, J., & Parrinello, M. (2007). Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.*, *98*(14), 146401. https://doi.org/10.1103/PhysRevLett.98.146401

Belsky, A., Hellenbrandt, M., Karen, V. L., & Luksch, P. (2002). New developments in the Inorganic Crystal Structure Database (ICSD): Accessibility in support of materials research

and design. *Acta Crystallographica Section B*, *58*(3 Part 1), 364–369. https://doi.org/10.1107/S0108768102006948

Benardos, P. G., & Vosniakos, G.-C. (2007). Optimizing feedforward artificial neural network architecture. *Engineering Applications of Artificial Intelligence*, *20*(3), 365–382. https://doi.org/10.1016/j.engappai.2006.06.005

Bomarito, G. F., Leser, P. E., Strauss, N. C. M., Garbrecht, K. M., & Hochhalter, J. D. (2022). Bayesian Model Selection for Reducing Bloat and Overfitting in Genetic Programming for Symbolic Regression. *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 526–529. https://doi.org/10.1145/3520304.3528899

Bomarito, G. F., Townsend, T. S., Stewart, K. M., Esham, K. V., Emery, J. M., & Hochhalter, J. D. (2021). Development of interpretable, data-driven plasticity models with symbolic regression. *Computers & Structures*, *252*, 106557. https://doi.org/10.1016/j.compstruc.2021.106557

Brence, J., Todorovski, L., & Džeroski, S. (2021). Probabilistic grammars for equation discovery. *Knowledge-Based Systems*, *224*, 107077. https://doi.org/10.1016/j.knosys.2021.107077

Bugel, M., & Galliero, G. (2008). Thermal conductivity of the Lennard-Jones fluid: An empirical correlation. *Chemical Physics*, *352*(1), 249–257. https://doi.org/10.1016/j.chemphys.2008.06.013

Cai, S., Mao, Z., Wang, Z., Yin, M., & Karniadakis, G. E. (2022). Physics-informed neural networks (PINNs) for fluid mechanics: A review. *Acta Mechanica Sinica*, 1–12. https://doi.org/10.1007/s10409-021-01148-1

Cao, W., & Zhang, W. (2022). Data-driven and physical-based identification of partial differential equations for multivariable system. *Theoretical and Applied Mechanics Letters*, *12*(2), 100334. https://doi.org/10.1016/j.taml.2022.100334

Chaabene, W. B., Flah, M., & Nehdi, M. L. (2020). Machine learning prediction of mechanical properties of concrete: Critical review. *Construction and Building Materials*, *260*, 119889. https://doi.org/10.1016/j.conbuildmat.2020.119889

Chaabene, W. B., & Nehdi, M. L. (2021). Genetic programming based symbolic regression for shear capacity prediction of SFRC beams. *Construction and Building Materials*, *280*, 122523. https://doi.org/10.1016/j.conbuildmat.2021.122523

Chatzilygeroudis, K., Hatzilygeroudis, I., & Perikos, I. (2021). Machine Learning Basics. In *Intelligent Computing for Interactive System Design: Statistics, Digital Signal Processing, and Machine Learning in Practice* (1st ed., pp. 143–193). Association for Computing Machinery. https://doi.org/10.1145/3447404.3447414

Danai, K., & Cava, W. G. L. (2021). Controller design by symbolic regression. *Mechanical Systems and Signal Processing*, *151*, 107348. https://doi.org/10.1016/j.ymssp.2020.107348

Delgado, A. M., Wadekar, D., Hadzhiyska, B., Bose, S., Hernquist, L., & Ho, S. (2022). Modelling the galaxy–halo connection with machine learning. *Monthly Notices of the Royal Astronomical Society*, *515*(2), 2733–2746. https://doi.org/10.1093/mnras/stac1951

Dimiduk, D. M., Holm, E. A., & Niezgoda, S. R. (2018). Perspectives on the impact of machine learning, deep learning, and artificial intelligence on materials, processes, and structures engineering. *Integrating Materials and Manufacturing Innovation*, *7*(3), 157–172. https://doi.org/10.1007/s40192-018-0117-8

Eiben, A. E., Smith, J. E., & others. (2003). *Introduction to evolutionary computing* (Vol. 53). Springer.

Farizhandi, A. A. K., Zhao, H., Chen, T., & Lau, R. (2020). Evaluation of material properties using planetary ball milling for modeling the change of particle size distribution in a gas-solid fluidized bed using a hybrid artificial neural network-genetic algorithm approach. *Chemical Engineering Science*, *215*, 115469. https://doi.org/10.1016/j.ces.2020.115469

França, F. O. de. (2018). A greedy search tree heuristic for symbolic regression. *Information Sciences*, *442–443*, 18–32. https://doi.org/10.1016/j.ins.2018.02.040

Freedman, R. A., Sandin, T., & Ford, A. L. (1996). *University physics*. Addison-Wesley.

Gilpin, W. (2021). *Chaos as an interpretable benchmark for forecasting and data-driven modelling*. https://doi.org/10.48550/ARXIV.2110.05266

Goh, G. D., Sing, S. L., & Yeong, W. Y. (2021). A review on machine learning in 3D printing: Applications, potential, and challenges. *Artificial Intelligence Review*, *54*(1), 63–94. https://doi.org/10.1007/s10462-020-09876-9

Guo, K., Yang, Z., Yu, C.-H., & Buehler, M. J. (2021). Artificial intelligence and machine learning in design of mechanical materials. *Mater. Horiz.*, *8*(4), 1153–1172. https://doi.org/10.1039/D0MH01451F

Haider, C., de França, F. O., Burlacu, B., & Kronberger, G. (2021). *Using Shape Constraints for Improving Symbolic Regression Models*. arXiv. https://doi.org/10.48550/ARXIV.2107.09458

Hart, G. L., Mueller, T., Toher, C., & Curtarolo, S. (2021). Machine learning for alloys. *Nature Reviews Materials*, *6*(8), 730–755. https://doi.org/10.1038/s41578-021-00340-w

Hautier, G. (2019). Finding the needle in the haystack: Materials discovery and design through computational ab initio high-throughput screening. *Computational Materials Science*, *163*, 108–116. https://doi.org/10.1016/j.commatsci.2019.02.040

He, M., & Zhang, L. (2021). Machine learning and symbolic regression investigation on stability of MXene materials. *Computational Materials Science*, *196*, 110578. https://doi.org/10.1016/j.commatsci.2021.110578

Hernandez, A., Balasubramanian, A., Yuan, F., Mason, S. A., & Mueller, T. (2019). Fast, accurate, and transferable many-body interatomic potentials by symbolic regression. *Npj Computational Materials*, *5*(1), 1–11. https://doi.org/10.1038/s41524-019-0249-1

Hu, F., & Hao, Q. (2012). *Intelligent sensor networks: The integration of sensor networks, signal processing and machine learning*. Taylor & Francis. https://doi.org/10.1201/b14300

Hu, Y., Wang, H., Guo, Y., & Zheng, L. (1996). Simulation of lubricant rheology in thin film lubrication Part I: simulation of Poiseuille flow. *Wear*, *196*(1), 243–248. https://doi.org/10.1016/0043-1648(95)06914-3

Hua, Y., Liu, Q., Hao, K., & Jin, Y. (2021). A Survey of Evolutionary Algorithms for Multi-Objective Optimization Problems With Irregular Pareto Fronts. *IEEE/CAA Journal of Automatica Sinica*, *8*(2), 303–318. https://doi.org/10.1109/JAS.2021.1003817

Janet, J. P., Liu, F., Nandy, A., Duan, C., Yang, T., Lin, S., & Kulik, H. J. (2019). Designing in the face of uncertainty: Exploiting electronic structure and machine learning models for discovery in inorganic chemistry. *Inorganic Chemistry*, *58*(16), 10592–10606. https://doi.org/10.1021/acs.inorgchem.9b00109

Kabliman, E., Kolody, A. H., Kommenda, M., & Kronberger, G. (2019). Prediction of stress-strain curves for aluminium alloys using symbolic regression. *AIP Conference Proceedings*, *2113*(1), 180009. https://doi.org/10.1063/1.5112747

Kabliman, E., Kolody, A. H., Kronsteiner, J., Kommenda, M., & Kronberger, G. (2021). Application of symbolic regression for constitutive modeling of plastic deformation. *Applications in Engineering Science*, *6*, 100052. https://doi.org/10.1016/j.apples.2021.100052

Karakasidis, T. E., Sofos, F., & Tsonos, C. (2022). The Electrical Conductivity of Ionic Liquids: Numerical and Analytical Machine Learning Approaches. *Fluids*, *7*(10). https://doi.org/10.3390/fluids7100321

Kirklin, S., Saal, J. E., Meredig, B., Thompson, A., Doak, J. W., Aykol, M., Rühl, S., & Wolverton, C. (2015). The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *Npj Computational Materials*, *1*(1), 1–15. https://doi.org/10.1038/npjcompumats.2015.10

Koza, J. R. (1994). Genetic programming as a means for programming computers by natural selection. *Statistics and Computing*, *4*(2), 87–112. https://doi.org/10.1007/BF00175355

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kronberger, G., de Franca, F. O., Burlacu, B., Haider, C., & Kommenda, M. (2022). Shape-Constrained Symbolic Regression—Improving Extrapolation with Prior Knowledge. *Evolutionary Computation*, *30*(1), 75–98. https://doi.org/10.1162/evco_a_00294

Kronberger, G., Kabliman, E., Kronsteiner, J., & Kommenda, M. (2022). Extending a physics-based constitutive model using genetic programming. *Applications in Engineering Science*, *9*, 100080. https://doi.org/10.1016/j.apples.2021.100080

Lemos, P., Jeffrey, N., Cranmer, M., Ho, S., & Battaglia, P. (2022). *Rediscovering orbital mechanics with machine learning*. arXiv. https://doi.org/10.48550/ARXIV.2202.02306

Liu, Q.-X., Jiang, P.-X., & Xiang, H. (2006). Molecular dynamics study of the thermal conductivity of nanoscale argon films. *Molecular Simulation*, *32*(8), 645–649. https://doi.org/10.1080/08927020600880794

Lopez-Bezanilla, A., & Littlewood, P. B. (2020). Growing field of materials informatics: Databases and artificial intelligence. *MRS Communications*, *10*(1), 1–10. https://doi.org/10.1557/mrc.2020.2

Manzi, M., & Vasile, M. (2020). Orbital anomaly reconstruction using deep symbolic regression. *71st International Astronautical Congress*.

Matchev, K. T., Matcheva, K., & Roman, A. (2022). Analytical Modeling of Exoplanet Transit Spectroscopy with Dimensional Analysis and Symbolic Regression. *The Astrophysical Journal*, *930*(1), 33. https://doi.org/10.3847/1538-4357/ac610c

MathWorks. (n.d.). https://www.mathworks.com/discovery/feature-extraction.html

Matsubara, Y., Chiba, N., Igarashi, R., Taniai, T., & Ushiku, Y. (2022). *Rethinking Symbolic Regression Datasets and Benchmarks for Scientific Discovery*. arXiv. https://doi.org/10.48550/ARXIV.2206.10540

Meier, K., Laesecke, A., & Kabelac, S. (2004). Transport coefficients of the Lennard-Jones model fluid. II Self-diffusion. *The Journal of Chemical Physics*, *121*(19), 9526–9535. https://doi.org/10.1063/1.1786579

Meng, L., McWilliams, B., Jarosinski, W., Park, H.-Y., Jung, Y.-G., Lee, J., & Zhang, J. (2020). Machine learning in additive manufacturing: A review. *Jom*, *72*(6), 2363–2377. https://doi.org/10.1007/s11837-020-04155-y

Mishin, Y. (2021). Machine-learning interatomic potentials for materials science. *Acta Materialia*, *214*, 116980. https://doi.org/10.1016/j.actamat.2021.116980

Murad, S., Ravi, P., & Powles, J. G. (1993). A computer simulation study of fluids in model slit, tubular, and cubic micropores. *The Journal of Chemical Physics*, *98*(12), 9771–9781. https://doi.org/10.1063/1.464356

Neumann, P., Cao, L., Russo, D., Vassiliadis, V. S., & Lapkin, A. A. (2020). A new formulation for symbolic regression to identify physico-chemical laws from experimental data. *Chemical Engineering Journal*, *387*, 123412. https://doi.org/10.1016/j.cej.2019.123412

Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 25). Determination press San Francisco, CA, USA.

Papastamatiou, K., Sofos, F., & Karakasidis, T. E. (2022). Machine learning symbolic equations for diffusion with physics-based descriptions. *AIP Advances*, *12*(2), 025004. https://doi.org/10.1063/5.0082147

Patrick, H. (1992). *Winston. Artificial Intelligence*. Addison-Wesley (now 3rd Edition).

Quade, M., Gout, J., & Abel, M. (2018). *Glyph: Symbolic Regression Tools*. arXiv. https://doi.org/10.48550/ARXIV.1803.06226

Quej, V. H., Almorox, J., Arnaldo, J. A., & Saito, L. (2017). ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment. *Journal of Atmospheric and Solar-Terrestrial Physics*, *155*, 62–70. https://doi.org/10.1016/j.jastp.2017.02.002

Rahman, J., Ahmed, K. S., Khan, N. I., Islam, K., & Mangalathu, S. (2021). Data-driven shear strength prediction of steel fiber reinforced concrete beams using machine learning approach. *Engineering Structures*, *233*, 111743. https://doi.org/10.1016/j.engstruct.2020.111743

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2017). Machine learning of linear differential equations using Gaussian processes. *Journal of Computational Physics*, *348*, 683–693. https://doi.org/10.1016/j.jcp.2017.07.050

Raschka, S. (2015). *Python machine learning*. Packt publishing ltd.

Reimann, D., Nidadavolu, K., ul Hassan, H., Vajragupta, N., Glasmachers, T., Junker, P., & Hartmaier, A. (2019). Modeling Macroscopic Material Behavior With Machine Learning Algorithms Trained by Micromechanical Simulations. *Frontiers in Materials*, *6*. https://doi.org/10.3389/fmats.2019.00181

Ren, J., Zhang, L., Zhao, H., Zhao, Z., & Wang, S. (2022). Determination of the fatigue equation for the cement-stabilized cold recycled mixtures with road construction waste materials based on data-driven. *International Journal of Fatigue*, *158*, 106765. https://doi.org/10.1016/j.ijfatigue.2022.106765

Rezaei, H., Zarfam, P., Golafshani, E. M., & Amiri, G. G. (2022). Seismic fragility analysis of RC box-girder bridges based on symbolic regression method. *Structures*, *38*, 306–322. https://doi.org/10.1016/j.istruc.2021.12.058

Schmid, M., Altmann, D., & Steinbichler, G. (2021). A Simulation-Data-Based Machine Learning Model for Predicting Basic Parameter Settings of the Plasticizing Process in Injection Molding. *Polymers*, *13*(16). https://doi.org/10.3390/polym13162652

Schmidt, M., & Lipson, H. (2009). Distilling Free-Form Natural Laws from Experimental Data. *Science*, *324*(5923), 81–85. https://doi.org/10.1126/science.1165893

Searson, D. P., Leahy, D. E., & Willis, M. J. (2010). GPTIPS: an open source genetic programming toolbox for multigene symbolic regression. *Proceedings of the International Multiconference of Engineers and Computer Scientists*, *1*, 77–80.

Seko, A., Togo, A., & Tanaka, I. (2018). Descriptors for Machine Learning of Materials Data. In I. Tanaka (Ed.), *Nanoinformatics* (pp. 3–23). Springer Singapore. https://doi.org/10.1007/978-981-10-7617-6_1

Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2016). *Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving*. arXiv. https://doi.org/10.48550/ARXIV.1610.03295

Shao, H., Villaescusa-Navarro, F., Genel, S., Spergel, D. N., Anglés-Alcázar, D., Hernquist, L., Davé, R., Narayanan, D., Contardo, G., & Vogelsberger, M. (2022). Finding Universal Relations in Subhalo Properties with Artificial Intelligence. *The Astrophysical Journal*, *927*(1), 85. https://doi.org/10.3847/1538-4357/ac4d30

Sharma, S., Sharma, S., & Athaiya, A. (2017). Activation functions in neural networks. *Towards Data Science*, *6*(12), 310–316.

Sofos, F., Charakopoulos, A., Papastamatiou, K., & Karakasidis, T. E. (2022). A combined clustering/symbolic regression framework for fluid property prediction. *Physics of Fluids*, *34*(6), 062004. https://doi.org/10.1063/5.0096669

Sofos, F., Karakasidis, T., & Liakopoulos, A. (2009). Transport properties of liquid argon in krypton nanochannels: Anisotropy and non-homogeneity introduced by the solid walls. *International Journal of Heat and Mass Transfer*, *52*(3), 735–743. https://doi.org/10.1016/j.ijheatmasstransfer.2008.07.022

Sofos, F., Stavrogiannis, C., Exarchou-Kouveli, K. K., Akabua, D., Charilas, G., & Karakasidis, T. E. (2022). Current Trends in Fluid Research in the Era of Artificial Intelligence: A Review. *Fluids*, *7*(3). https://doi.org/10.3390/fluids7030116

Stephenson, D., Kermode, J. R., & Lockerby, D. A. (2018). Accelerating multiscale modelling of fluids with on-the-fly Gaussian process regression. *Microfluidics and Nanofluidics*, *22*(12), 1–12. https://doi.org/10.1007/s10404-018-2164-z

Swamynathan, M. (2019). *Mastering machine learning with python in six steps: A practical implementation guide to predictive data analytics using python*. Apress.

Tohme, T., Liu, D., & Youcef-Toumi, K. (2022). *GSR: A Generalized Symbolic Regression Approach*. arXiv. https://doi.org/10.48550/ARXIV.2205.15569

Udrescu, S.-M., & Tegmark, M. (2019). *AI Feynman: A Physics-Inspired Method for Symbolic Regression*. https://doi.org/10.48550/ARXIV.1905.11481

Vaddireddy, H., Rasheed, A., Staples, A. E., & San, O. (2020). Feature engineering and symbolic regression methods for detecting hidden physics from sparse sensor observation data. *Physics of Fluids*, *32*(1), 015113. https://doi.org/10.1063/1.5136351

Van Veldhuizen, D. A., Lamont, G. B., & others. (1998). Evolutionary computation and convergence to a pareto front. *Late Breaking Papers at the Genetic Programming 1998 Conference*, 221–228.

Wadekar, D., Thiele, L., Villaescusa-Navarro, F., Hill, J. C., Cranmer, M., Spergel, D. N., Battaglia, N., Anglés-Alcázar, D., Hernquist, L., & Ho, S. (2022). *Augmenting astrophysical scaling relations with machine learning: Application to reducing the SZ flux-mass scatter*. arXiv. https://doi.org/10.48550/ARXIV.2201.01305

Wadekar, D., Villaescusa-Navarro, F., Ho, S., & Perreault-Levasseur, L. (2020). *Modeling assembly bias with machine learning and symbolic regression*. arXiv. https://doi.org/10.48550/ARXIV.2012.00111

Wagner, S., Kronberger, G., Beham, A., Kommenda, M., Scheibenpflug, A., Pitzer, E., Vonolfen, S., Kofler, M., Winkler, S., Dorfer, V., & Affenzeller, M. ". (2014). Architecture and Design of the HeuristicLab Optimization Environment. In R. Klempous, J. Nikodem, W. Jacak, & Z. Chaczko (Eds.), *Advanced Methods and Applications in Computational Intelligence* (pp. 197–261). Springer International Publishing. https://doi.org/10.1007/978-3-319-01436-4_10

Wang, M., Chen, C., & Liu, W. (2022). Establish algebraic data-driven constitutive models for elastic solids with a tensorial sparse symbolic regression method and a hybrid feature selection technique. *Journal of the Mechanics and Physics of Solids*, *159*, 104742. https://doi.org/10.1016/j.jmps.2021.104742

Wang, Y., Wagner, N., & Rondinelli, J. M. (2019). Symbolic regression in materials science. *MRS Communications*, *9*(3), 793–805. https://doi.org/10.1557/mrc.2019.85

Wei, J., Chu, X., Sun, X.-Y., Xu, K., Deng, H.-X., Chen, J., Wei, Z., & Lei, M. (2019). Machine learning in materials science. *InfoMat*, *1*(3), 338–358. https://doi.org/10.1002/inf2.12028

Weng, B., Song, Z., Zhu, R., Yan, Q., Sun, Q., Grice, C. G., Yan, Y., & Yin, W.-J. (2020). Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts. *Nature Communications*, *11*(1), 1–8. https://doi.org/10.1038/s41467-020-17263-9

Wilstrup, C., & Kasak, J. (2021). *Symbolic regression outperforms other models for small data sets*. arXiv. https://doi.org/10.48550/ARXIV.2103.15147

Yang, L., Meng, X., & Karniadakis, G. E. (2021). B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. *Journal of Computational Physics*, *425*, 109913. https://doi.org/10.1016/j.jcp.2020.109913

Yang, Y., Zhao, L., Han, C.-X., Ding, X.-D., Lookman, T., Sun, J., & Zong, H.-X. (2021). Taking materials dynamics to new extremes using machine learning interatomic potentials. *Journal of Materials Informatics*, *1*(2), 10. https://doi.org/10.20517/jmi.2021.001

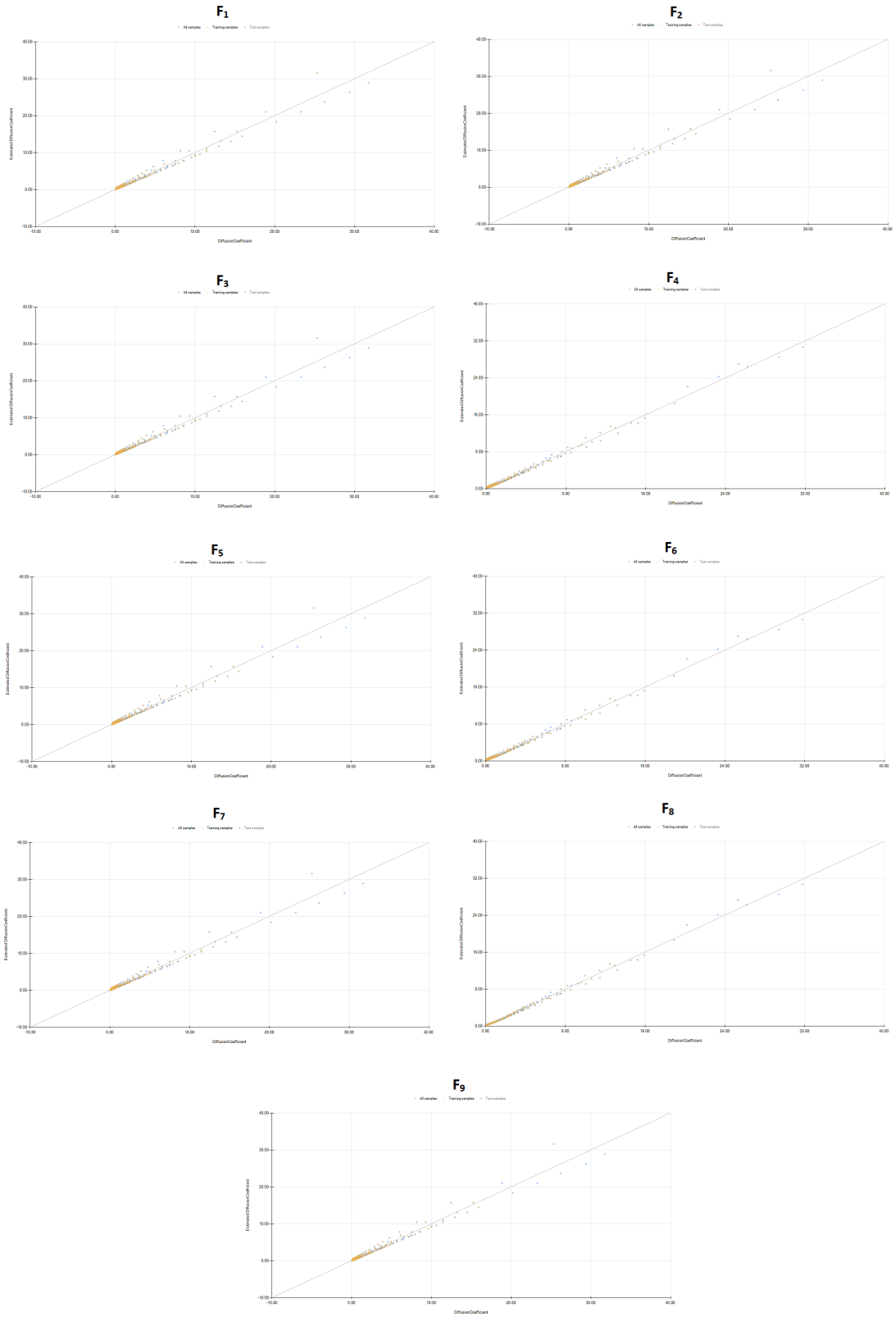
Ye, T., Pan, D., Huang, C., & Liu, M. (2019). Smoothed particle hydrodynamics (SPH) for complex fluid flows: Recent developments in methodology and applications. *Physics of Fluids*, *31*(1), 011301. https://doi.org/10.1063/1.5068697

Zhang, X., Chen, Z., & Liu, Y. (2017). Chapter 6—Constitutive Models. In X. Zhang, Z. Chen, & Y. Liu (Eds.), *The Material Point Method* (pp. 175–219). Academic Press. https://doi.org/10.1016/B978-0-12-407716-4.00006-5

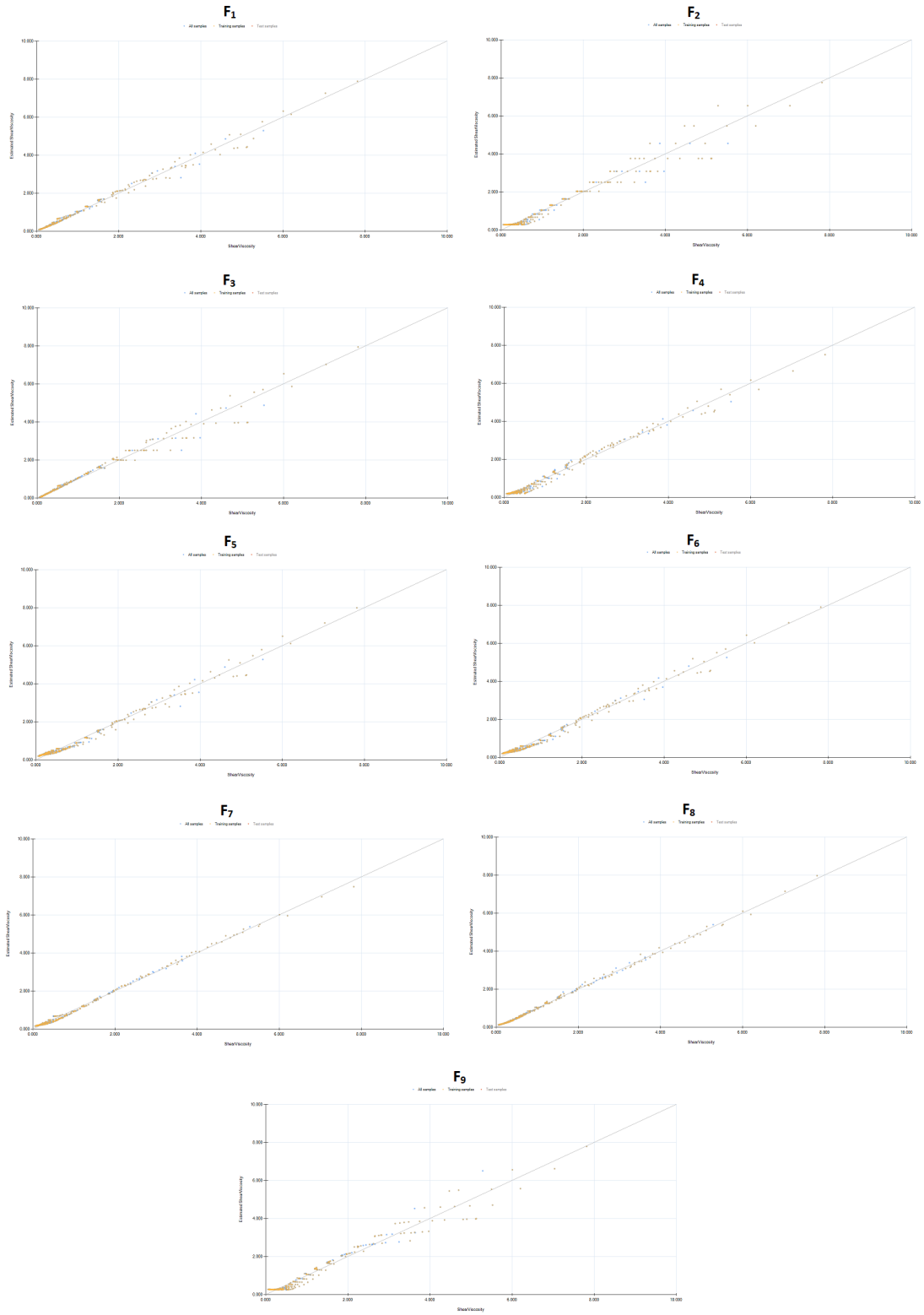
Zhong, S., Yap, B. K., Zhong, Z., & Ying, L. (2022). Review on Y6-Based Semiconductor Materials and Their Future Development via Machine Learning. *Crystals*, *12*(2). https://doi.org/10.3390/cryst12020168

Zhou, T., Song, Z., & Sundmacher, K. (2019). Big Data Creates New Opportunities for Materials Research: A Review on Methods and Applications of Machine Learning for Materials Design. *Engineering*, *5*(6), 1017–1026. https://doi.org/10.1016/j.eng.2019.02.011

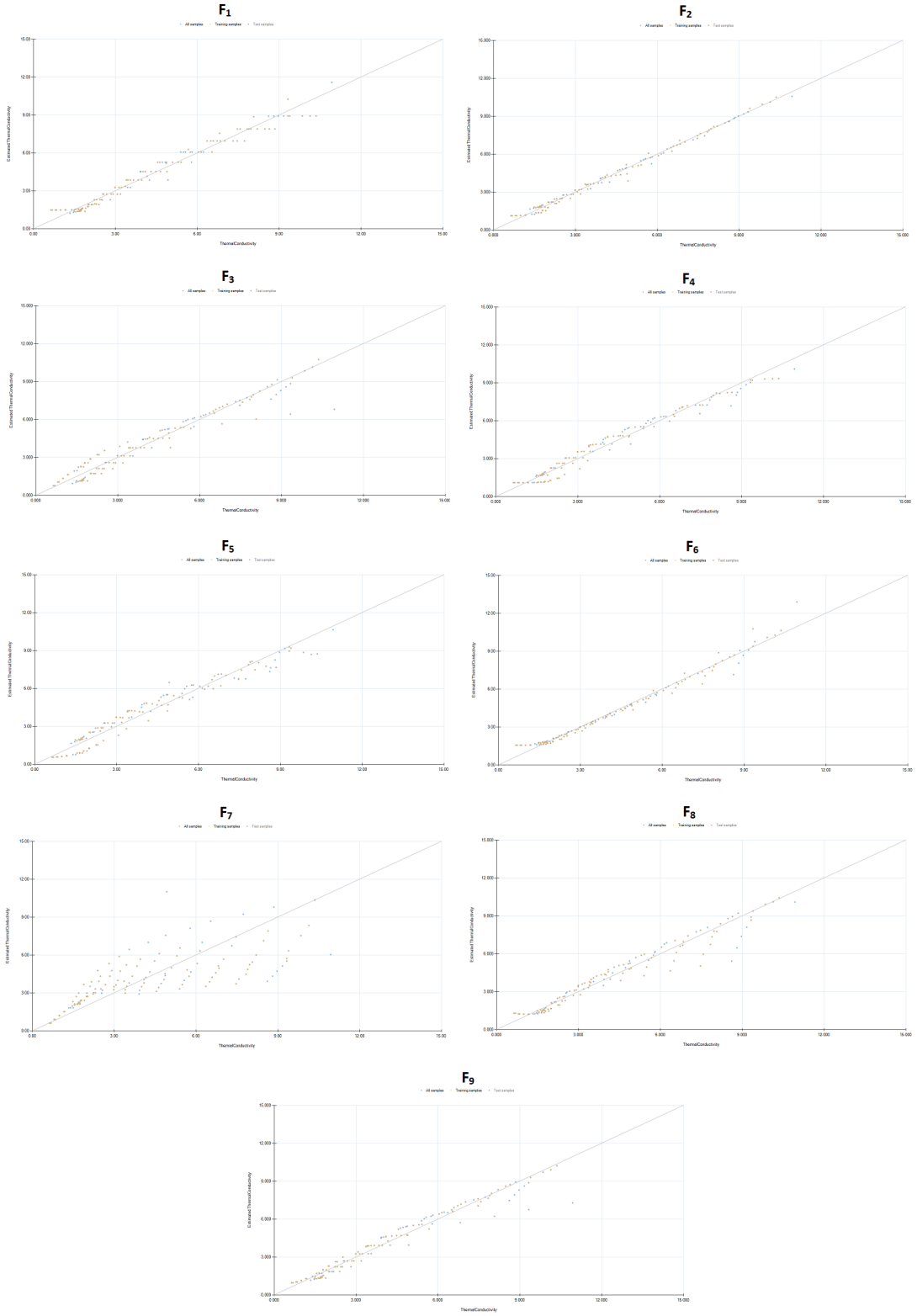
Παράρτημα Α



Σχήμα 30: : Γραφήματα διασποράς για τις εξισώσεις $\boldsymbol{F_i}$ $(\boldsymbol{i = 1, \dots, 9})$ του Πίνακα 8 (συντελεστής διάχυσης $\boldsymbol{D}$). Στις εξισώσεις όπου ο άξονας δεν ξεκινάει από το σημείο (0,0) αλλά από το (-10,-10) είναι ένδειξη ότι κάποια τιμή της πρόβλεψης έχει λάβει αρνητική τιμή.



Σχήμα 31: Γραφήματα διασποράς για τις εξισώσεις $\boldsymbol{F_i}$ $\boldsymbol{(i = 1, \dots, 9)}$ του Πίνακα 10 (ιξώδες διάτμησης $\boldsymbol{n_s}$).



Σχήμα 32: Γραφήματα διασποράς για τις εξισώσεις $\boldsymbol{F_i}$ $(\boldsymbol{i = 1, \dots, 9})$ του Πίνακα 12 (θερμική αγωγιμότητα $\boldsymbol{\lambda}$).

Παράρτημα Β

Το ελεύθερο λογισμικό HeuristicLab υλοποιεί τη διαδικασία της συμβολικής παλινδρόμησης βάσει του γενετικού προγραμματισμού. Τα σύμβολα και οι εξισώσεις που χρησιμοποιήθηκαν για τις αρχικές συναρτήσεις και των τριών περιπτώσεων (D, N_s, λ) περιλαμβάνουν ($+, -, \div, *$, exp, log, ^) ενώ στους τερματικούς κόμβους επιλέχθηκαν σταθερές και οι μεταβλητές ρ και T. Κατά την είσοδο των δεδομένων, επιλέχθηκε να ανακατευτεί το σετ δεδομένων (επιλογή shuffle dataset) ενώ το 80% των δεδομένων χρησιμοποιήθηκε για εκπαίδευση και το υπόλοιπο 20% για έλεγχο. Το μέγιστο βάθος του δέντρου για τη συμβολική έκφραση ορίστηκε στα 12 επίπεδα ενώ το μέγιστο μήκος σε 150. Ο αριθμός των ατόμων στον πληθυσμό επιλέχθηκε να είναι ίσος με 5000 ενώ ο αριθμός των πληθυσμών που θα δημιουργηθούν ίσος με 300. Τέλος, η πιθανότητα μετάλλαξης επιλέχθηκε να είναι 15%, ενώ κατά την εξαγωγή της εξίσωσης, χρησιμοποιήθηκε η επιλογή simplify solution για να απλουστεύσει την παραγόμενη εξίσωση.