ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΟΛΟΓΝΩ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ & ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Σχεδιασμός Πληροφοριακού Συστήματος Εκπαίδευσης Μοντέλων Μηχανικής Μάθησης για Ενεργειακές Εφαρμογές

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Συμεών Χορόζογλου

Επιβλέπων: Χρυσόστομος Δούκας
Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2023

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΟΛΟΓΝΩ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ & ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Σχεδιασμός Πληροφοριακού Συστήματος Εκπαίδευσης Μοντέλων Μηχανικής Μάθησης για Ενεργειακές Εφαρμογές

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Συμεών Χορόζογλου

Επιβλέπων: Χρυσόστομος Δούκας
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 10η Μαρτίου 2023.

.....	.....	.....
Δούκας Χ.	Ψαρράς Ι.	Ασκούνης Δ.
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2023

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΟΛΟΓΝΩ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ & ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

.....

Συμεών Χορόζογλου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Στην παρούσα διπλωματική εργασία στοχεύουμε στην ανάπτυξη ενός πληροφοριακού συστήματος για την εκπαίδευση μοντέλων μηχανικής μάθησης πάνω σε ενεργειακά δεδομένα. Η εξάπλωση της χρήσης των φωτοβολταϊκών για την παραγωγή ενέργειας αλλά και η προσπάθεια για μείωση κατανάλωσης ενέργειας απαιτούν σύγχρονα συστήματα πρόβλεψης και επεξεργασίας των αντίστοιχων δεδομένων. Αυτό ενισχύεται και από την πολύ μεγάλη συγκέντρωση διαθέσιμης πληροφορίας μέσα από τη χρήση αισθητήρων και έξυπνων συσκευών. Στο σύστημα που υλοποιούμε θέλουμε να παρέχουμε τη δυνατότητα πρόβλεψης για την παραγωγή ενέργειας φωτοβολταϊκών, πρόβλεψης εξοικονόμησης ενέργειας σε κτήρια και βιομηχανίες ύστερα από εργασίες ανακαίνισης, καθώς και ταξινόμησης διαφόρων ενεργειακών έργων ανακαίνισης σε αντίστοιχες κλάσεις. Για τον σκοπό αυτό αναπτύχθηκε η κατάλληλη μεθοδολογία που υποδεικνύει τα βήματα τα οποία πρέπει να ακολουθηθούν για τη διαδικασία της πρόβλεψης. Στη διαδικασία αυτή συμπεριλαμβάνονται ο καθορισμός των δεδομένων και η προεπεξεργασία τους, ο σχεδιασμός κατάλληλων μοντέλων νευρωνικών δικτύων και αλγορίθμων μηχανικής μάθησης, η επιλογή των υπερπαραμέτρων των μοντέλων, η εκπαίδευσή τους, η αξιολόγηση των μοντέλων και η οπτική αναπαράσταση των προβλέψεών τους. Ο χρήστης μπορεί να επεμβαίνει σημαντικά σε κάθε βήμα της διαδικασίας, λαμβάνοντας αποφάσεις για τα δεδομένα που θα εξαχθούν, τον τρόπο που θα επεξεργαστούν, τα μοντέλα που θα εκπαιδευτούν καθώς και για τις τιμές που θα λάβουν οι υπερπαραμέτροι των επιλεγμένων μοντέλων. Η μεθοδολογία αυτή ενσωματώνεται στο πληροφοριακό σύστημα που σχεδιάζουμε, προσφέροντας έτσι μια φιλική διεπαφή στο χρήστη για την ευκολότερη μελέτη αυτών των προβλημάτων. Με αυτόν τον τρόπο καθίσταται εύκολη η διαδικασία πρόβλεψης και η αξιολόγηση των μοντέλων.

Λέξεις Κλειδιά: Πληροφοριακό σύστημα, μηχανική μάθηση, ταξινόμηση, παλινδρόμηση, πρόβλεψη, νευρωνικά δίκτυα, επεξεργασία δεδομένων, φωτοβολταϊκά, εξοικονόμηση ενέργειας

Abstract

The target of the current thesis is the development of an information system for training machine learning models using energy data. The need to enhance the energy production from solar panels, along with the efforts made for reducing the superfluous consumption of energy, require innovative support systems that process the corresponding data and make useful predictions. The necessity to implement such information systems is reinforced by the enormous amount of available data, collected from smart meters, sensors, and smart devices. The system we develop integrates various machine learning models, which are designed for predicting energy production of a dataset associated with solar panels, classifying energy projects in appropriate classes, and predicting energy savings of some refurbishments in buildings and industries. Aforementioned problems belong to classification, regression and forecasting categories. Taking this parameter into account we propose the methodology which leads to accurate predictions, and we analyze extensively each step followed. The procedure of making accurate predictions includes definition of the problem, data pre-processing, creation of machine learning and neural networks models, hyperparameter tuning, models' training, models' evaluation, and visualization of predicted values. Users can intervene in every step by extracting the desired features from the data, determining the way data will be pre-processed, selecting the models and tuning their hyperparameters. This methodology is integrated into the designed information system, offering to users a friendly interface to examine energy problems and machine learning models.

Keywords: Information system, machine learning, classification, regression, forecasting, neural networks, data pre-processing, solar panels, energy savings

Ευχαριστίες

Με την ολοκλήρωση αυτής της διπλωματικής εργασίας κλείνει ένας πολύ σημαντικός κύκλος, αυτός των προπτυχιακών μου σπουδών. Σε αυτό το σημείο θα ήθελα να ευχαριστήσω τους ανθρώπους που με βοήθησαν, με καθοδήγησαν και με στήριξαν σε αυτήν τη διαδρομή.

Πρώτα απ' όλα θα ήθελα να ευχαριστήσω τον υπεύθυνο καθηγητή της διπλωματικής μου, κ. Χρυσόστομο Δούκα για την ευκαιρία που μου έδωσε να εκπονήσω τη διπλωματική μου στο Εργαστήριο Συστημάτων Αποφάσεων και Διοίκησης. Είναι μεγάλη μου χαρά που συνεργάστηκα μαζί του στα πλαίσια προπτυχιακών μαθημάτων της σχολής αλλά κυρίως για την εκπόνηση της παρούσας διπλωματικής εργασίας. Ακόμα, θα ήθελα να ευχαριστήσω όλα τα μέλη του εργαστηρίου που με βοήθησαν με την καθοδήγησή τους και τις ιδέες τους. Η ανταπόκρισή τους για την επίλυση των αποριών που προέκυπταν στα πλαίσια της εργασίας ήταν πάντα άμεση και ουσιαστική.

Επίσης θα ήθελα να ευχαριστήσω τους καθηγητές κ. Ιωάννη Ψαρρά και κ. Δημήτριο Ασκούνη για τις γνώσεις που μου προσέφεραν στα προπτυχιακά μαθήματα που παρακολούθησα και για τη συμμετοχή τους στην εξέταση αυτής της διπλωματικής.

Επιπλέον θα ήθελα να ευχαριστήσω τους φίλους και τους ανθρώπους που γνώρισα στη διάρκεια των φοιτητικών μου χρόνων. Οι στιγμές που μοιραστήκαμε θα μου μείνουν για πάντα ως μερικές από τις ομορφότερες αναμνήσεις της φοιτητικής μου ζωής.

Τέλος δεν θα μπορούσα να μην αναφερθώ στα μέλη της οικογένειάς μου. Η υποστήριξή τους σε κάθε μου βήμα αποτελεί σημαντική βοήθεια για την εκπλήρωση των στόχων μου.

Περιεχόμενα

ΕΥΡΕΤΗΡΙΟ ΕΙΚΟΝΩΝ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Εισαγωγή

Τα τελευταία χρόνια γίνονται σημαντικές προσπάθειες ώστε το ποσοστό της ενέργειας που παράγεται από Ανανεώσιμες Πηγές να αυξηθεί με τον Ευρωπαϊκό Οργανισμό Περιβάλλοντος να θέτει ως στόχο μέχρι το 2030 το 32% της ενέργειας που καταναλώνεται να προέρχεται από αυτές, στόχος ο οποίος μπορεί ακόμα και σε σύντομο χρονικό διάστημα να αναθεωρηθεί προς τα πάνω[1]. Παράλληλα, γίνονται μια σειρά από παρεμβάσεις και ανακατασκευές σε κτήρια παλιάς τεχνολογίας ώστε να μειωθεί όσο το δυνατόν περισσότερο η ενέργεια που καταναλώνουν και οι ρύποι που εκπέμπουν. Η ανάγκη για μείωση των ρύπων και η στροφή σε πράσινες μορφές ενέργειας έρχεται σε μία εποχή που ανθίζει η επιστήμη των δεδομένων και η μηχανική μάθηση.

Η προσπάθεια για μείωση των ρύπων συμπίπτει με την εποχή των big data [1]. Η ευρεία χρήση των smart meters σήμερα, η παρακολούθηση της παραγωγής των φωτοβολταϊκών, των αιολικών πάρκων καθώς και των υπόλοιπων ανανεώσιμων πηγών, οδηγεί στη δημιουργία μεγάλου όγκου δεδομένων που έχουν άμεση σχέση με τον ενεργειακό τομέα. Επίσης, η συνεχής καταγραφή καιρικών και γεωγραφικών δεδομένων συνεισφέρει σημαντικά σε τομείς πρόβλεψης παραγωγής από τις ΑΠΕ, καθώς έχουν άμεση συσχέτιση. Δημιουργείται λοιπόν η ανάγκη για σωστή αξιοποίηση και επεξεργασία αυτών των δεδομένων με σκοπό την καλύτερη δυνατή πρόβλεψη παραγωγής ενέργειας από τις ΑΠΕ, την πρόβλεψη εξοικονόμησης ενέργειας σε ανακατασκευασμένα κτήρια, την ταξινόμησή τους σε αντίστοιχες ενεργειακές κλάσεις καθώς και άλλες εφαρμογές στον ενεργειακό χώρο που έχουν ως κοινό παρονομαστή τη μείωση των ρύπων.

Η αξιοποίηση των νέων τεχνολογιών μπορεί να συμβάλλει καθοριστικά σε αυτές τις προσπάθειες με την ανάπτυξη μεθόδων και μοντέλων βασισμένα στη τεχνητή νοημοσύνη με στόχο την ακριβή παρακολούθηση και πρόβλεψη ενεργειακών ζητημάτων. Η μετάβαση και ο εκσυγχρονισμός των ηλεκτρικών δικτύων σε έξυπνα δίκτυα (smart grids) απαιτεί την ανάπτυξη έξυπνων συστημάτων διαχείρισης ενέργειας [2], με την τεχνητή νοημοσύνη να διαδραματίζει ιδιαίτερα σημαντικό ρόλο. Η δημιουργία και ανάπτυξη πληροφοριακών συστημάτων που αξιοποιούν σωστά τα δεδομένα που αναφέραμε και ενσωματώνουν μεθοδολογίες και μοντέλα μηχανικής μάθησης μπορούν να προσφέρουν σημαντικές λύσεις στις ενεργειακές προκλήσεις που αντιμετωπίζουμε σήμερα. Την ανάγκη για ανάπτυξη τέτοιων πληροφοριακών συστημάτων, που κάνουν χρήση αλγορίθμων μηχανικής μάθησης εξετάζουμε σε αυτήν τη διπλωματική εργασία.

[1] Renewable energy targets (europa.eu)

1.2 Στόχος της Διπλωματικής

Στην παρούσα διπλωματική προχωρήσαμε στην μελέτη ορισμένων ενεργειακών προβλημάτων και πώς μπορούμε να τα αντιμετωπίσουμε αξιοποιώντας τη διαθέσιμη πληροφορία. Όπως τονίστηκε παραπάνω, η μετάβαση σε έξυπνα δίκτυα, η περαιτέρω ανάπτυξη των ανανεώσιμων πηγών ενέργειας και η μείωση των ρύπων μπορούν να υποστηριχτούν από αντίστοιχα πληροφοριακά συστήματα. Το περιεχόμενο αυτής της διπλωματικής εργασίας είναι η ανάπτυξη ενός πληροφοριακού συστήματος που θα δέχεται και θα επεξεργάζεται δεδομένα σχετικά με προβλήματα εξοικονόμησης ενέργειας ύστερα από εργασίες ανακαίνισης, καθώς και παραγωγής ενέργειας φωτοβολταϊκών. Τα ακριβή δεδομένα των προβλημάτων αναλύονται στο κεφάλαιο 4. Παράλληλα, αναπτύχθηκαν διάφορα μοντέλα νευρωνικών δικτύων και μηχανικής μάθησης τα οποία θα εκπαιδευτούν και θα δοκιμαστούν στα αντίστοιχα ενεργειακά σύνολα δεδομένων που διαθέτουμε. Τα μοντέλα αυτά ενσωματώθηκαν στο πληροφοριακό σύστημα και διατίθενται στο χρήστη για αξιολόγηση.

Στην εφαρμογή αυτή, ο χρήστης θα μπορεί επίσης να κάνει μία σειρά από επιλογές σχετικά με την προεπεξεργασία των δεδομένων, την επιλογή των μοντέλων και τις υπερπαραμέτρους τους πριν οδηγηθεί στις προβλέψεις που θα πραγματοποιηθούν και στα αποτελέσματα που θα προκύψουν. Συνεπώς θέλουμε να δημιουργήσουμε ένα εργαλείο το οποίο θα παρέχει ένα φιλικό περιβάλλον για τους χρήστες ώστε να υλοποιούν τις επιθυμητές τους προβλέψεις, αλλά ταυτόχρονα θα βοηθά στην κατανόηση και αξιολόγηση δημοφιλών αλγορίθμων μηχανικής μάθησης και αρχιτεκτονικών νευρωνικών δικτύων, καθώς και στην επιρροή που έχουν οι διαφορετικές τιμές για τις διαθέσιμες υπερπαραμέτρους στην επίδοση των εξεταζόμενων μοντέλων.

Το εργαλείο αυτό θα απευθύνεται τόσο σε χρήστες που επιθυμούν να πραγματοποιήσουν γρήγορα και εύκολα διάφορες προβλέψεις χωρίς απαραίτητα να διαθέτουν το αντίστοιχο υπόβαθρο πάνω στους τομείς της τεχνητής νοημοσύνης. Σημειώνεται και στη συνέχεια της εργασίας το γεγονός ότι υπάρχει έλλειψη στον τομέα της πληροφορικής όσον αφορά τη διάθεση προσωπικού με το απαραίτητο τεχνικό υπόβαθρο για την ανάλυση των δεδομένων. Η δημιουργία αυτού του εργαλείου θα συνεισφέρει στην αξιοποίηση πληροφορίας πάνω σε ενεργειακά ζητήματα χωρίς να απαιτούνται σημαντικές γνώσεις στην επιστήμη δεδομένων.

Χρήστες που έχουν μεγαλύτερη εξοικείωση με την επιστήμη των δεδομένων και με μοντέλα μηχανικής μάθησης μπορούν να κάνουν μία πιο εκτεταμένη χρήση του πληροφοριακού συστήματος, καθώς θα μπορούν να κατανοήσουν καλύτερα τη σημασία των υπερπαραμέτρων των μοντέλων που μπορούν να οριστούν και να εξάγουν συμπεράσματα για την απόδοση των αλγορίθμων και των νευρωνικών δικτύων που χρησιμοποιούνται.

1.3 Δομή της Διπλωματικής

Η παρούσα διπλωματική εργασία αποτελείται από έξι κεφάλαια. Παρακάτω παρουσιάζουμε συνοπτικά το περιεχόμενο αυτών των κεφαλαίων:

- Το πρώτο κεφάλαιο είναι το κεφάλαιο εισαγωγής, όπου παρουσιάζουμε μία σύντομη εισαγωγή στα προβλήματα που εξετάζουμε και στους στόχους και τη συμβολή αυτής της εργασίας.

- Στο δεύτερο κεφάλαιο κάνουμε μια πιο εκτενή παρουσίαση των πεδίων που θα μελετήσουμε σε αυτήν την εργασία. Ταυτόχρονα αναφέρουμε ορισμένες από τις προκλήσεις που αντιμετωπίζουμε σήμερα στο χώρο της ενέργειας και των big data.

- Στο τρίτο κεφάλαιο παραθέτουμε το θεωρητικό υπόβαθρο που χρειάζεται στα πλαίσια της εργασίας. Παρουσιάζουμε ορισμένους δημοφιλείς αλγορίθμους μηχανικής μάθησης που χρησιμοποιούμε για προβλήματα ταξινόμησης και παλινδρόμησης και ορισμένες αρχιτεκτονικές νευρωνικών δικτύων. Ακόμη περιγράφουμε τις μετρικές απόδοσης και σφάλματος, βάσει των οποίων αξιολογήθηκαν τα μοντέλα.

- Στο τέταρτο κεφάλαιο αναλύουμε τη μεθοδολογία που αναπτύχθηκε σχετικά με τη διαδικασία πρόβλεψης. Περιγράφουμε τα επιμέρους βήματα που πρέπει να ακολουθηθούν, όπως ο καθορισμός των δεδομένων, η προεπεξεργασία τους, η επιλογή των μοντέλων, ο καθορισμός των υπερπαραμέτρων τους, η αξιολόγησης της επίδοσης των μοντέλων και η οπτικοποίηση των προβλέψεων.

- Στο πέμπτο κεφάλαιο παρουσιάζουμε τα εργαλεία και τις βιβλιοθήκες που χρησιμοποιήσαμε για την ανάπτυξη του πληροφοριακού συστήματος, καθώς και τα διαγράμματα που σχεδιάστηκαν για τον τρόπο χρήσης του και τις λειτουργίες του. Επίσης, παραθέτουμε στιγμιότυπα από τις σελίδες που μπορεί να περιηγηθεί ο χρήστης αλλά και ορισμένα αποτελέσματα που προκύπτουν για κάποια προβλήματα και ορισμένες επιλεγμένες τιμές. Το πληροφοριακό σύστημα ενσωματώνει τη μεθοδολογία που αναπτύχθηκε στο κεφάλαιο τέσσερα.

- Στο έκτο κεφάλαιο παρουσιάζουμε τα συμπεράσματα που προέκυψαν και σημειώνουμε ορισμένες διαπιστώσεις σχετικά με μελλοντικές επεκτάσεις που μπορούν να πραγματοποιηθούν.

ΚΕΦΑΛΑΙΟ 2

ΔΙΑΤΥΠΩΣΗ ΠΡΟΒΛΗΜΑΤΟΣ

2.1 Εισαγωγή

Όπως περιγράψαμε στο εισαγωγικό κεφάλαιο, η μετάβαση στην αύξηση της παραγωγής ενέργειας από ΑΠΕ καθώς και η ανάγκη για εξοικονόμηση ενέργειας αποτελούν πολύ σημαντικές προκλήσεις. Η ανάπτυξη και η ενσωμάτωση νέων τεχνολογιών ακόμα και σε απλές συσκευές δημιουργεί έναν τεράστιο όγκο διαθέσιμων δεδομένων, τα οποία αν επεξεργαστούν και αξιοποιηθούν με τον κατάλληλο τρόπο μπορεί να οδηγήσουν σε πολύ χρήσιμα συμπεράσματα σχετικά με την πρόβλεψη παραγωγής ή ζήτησης ενέργειας, καθώς και σε άλλους σημαντικούς τομείς.

Η επεξεργασία και αξιοποίηση αυτών των δεδομένων με την εφαρμογή αλγορίθμων μηχανικής μάθησης αποτελεί σημαντικό πεδίο έρευνας. Παρ' όλα αυτά είναι αρκετά σύνηθες φαινόμενο να μην υπάρχει η κατάλληλη τεχνογνωσία για την αξιοποίηση αυτών των δεδομένων. Συνεπώς θα ήταν χρήσιμη μία εφαρμογή η οποία μπορεί να οδηγήσει ακόμα και έναν άπειρο χρήστη στην εξαγωγή χρήσιμων συμπερασμάτων σε σημαντικά προβλήματα, όπως η πρόβλεψη παραγωγής ενέργειας φωτοβολταϊκών ή ακόμα και να διευκολύνει έναν πιο έμπειρο χρήστη να συγκρίνει την απόδοση διάφορων αλγορίθμων μηχανικής μάθησης και νευρωνικών δικτύων δοκιμάζοντας διαφορετικές υπερπαραμέτρους στα μοντέλα που θέλει να εξετάσει.

Στο κεφάλαιο αυτό θα παρουσιάσουμε πιο αναλυτικά τον ρόλο των big data και της τεχνητής νοημοσύνης στον χώρο της ενέργειας. Σημειώνουμε πώς η ενσωμάτωση αυτών των πεδίων σε ένα πληροφοριακό σύστημα μπορεί να προσφέρει ουσιαστικές λύσεις και αναφέρουμε τις ομάδες των ανθρώπων (stakeholders) καθώς και τον τρόπο με τον οποίο αυτές μπορούν να επωφεληθούν από την ανάπτυξη ενός τέτοιου συστήματος.

2.2 Big Data και διαχείριση ενέργειας

Η ραγδαία ανάπτυξη στην τεχνολογία αισθητήρων, η ευρεία χρήση smart meters και συστημάτων cloud computing όπως επίσης και η χρήση μίας σειράς έξυπνων συσκευών έχει οδηγήσει στη συσσώρευση πολύ μεγάλου όγκου πληροφορίας πάνω σε ενεργειακά δεδομένα [3]. Ακόμη, η ανάπτυξη των έξυπνων δικτύων (smart grids) δημιουργεί συλλογή δεδομένων μέχρι και από απλές οικιακές συσκευές. Συνεπώς χρειάζεται να δημιουργηθούν αποτελεσματικά συστήματα πρόβλεψης ζήτησης και παραγωγής ενέργειας για την καλύτερη δυνατή εξυπηρέτηση των έξυπνων δικτύων.

Πέρα από τους αισθητήρες, τη χρήση smart meters, οικιακών συσκευών και άλλων τεχνολογικών συστημάτων καταγραφής κατανάλωσης και παραγωγής ενέργειας, τα ενεργειακά big data εμπλουτίζονται

και από πολλά άλλα πεδία, όπως καιρικά ή γεωγραφικά δεδομένα. Αυτά τα δεδομένα παίζουν μεταξύ άλλων πολύ σημαντικό ρόλο στην πρόβλεψη παραγωγής ενέργειας από ανανεώσιμες πηγές, καθώς και στην πρόβλεψη ζήτησης ενέργειας. Είναι σαφές ότι περιοχές με έντονη ηλιοφάνεια ή ισχυρούς ανέμους αποτελούν ευνοϊκά σημεία για εγκατάσταση φωτοβολταϊκών ή αιολικών πάρκων, ενώ περιοχές με ακραίες θερμοκρασίες απαιτούν την αντίστοιχη κατανάλωση ενέργειας για τη ρύθμιση της θερμοκρασίας στα κτήρια. Για παράδειγμα, οι αστικές θερμικές νησίδες (urban heat islands) [4] που αναπτύσσονται στα αστικά κέντρα απαιτούν μεγάλη κατανάλωση ενέργειας για λόγους κλιματισμού. Ιδιαίτερα για την περίπτωση της Αθήνας, το ενεργειακό φορτίο που απαιτείται για λόγους κλιματισμού κτηρίων μπορεί μέχρι και να τριπλασιαστεί λόγω του φαινομένου των αστικών θερμικών νησίδων [5].

Η σωστή εκμετάλλευση των big data μπορεί να βοηθήσει τις βιομηχανίες ενέργειας να διαχειριστούν καλύτερα μία σειρά από προβλήματα, όπως η διαχείριση της παραγωγής από ανανεώσιμες πηγές, η διαχείριση του κόστους λειτουργίας και η αντίστοιχη βελτίωση των υπηρεσιών που παρέχονται. Η αλματώδης αύξηση στη συλλογή δεδομένων δημιουργεί και μία σειρά από προκλήσεις στον κλάδο της πληροφορικής όπως:

- Συλλογή υψηλής ποιότητας δεδομένων: Είναι πολύ σημαντικό να μπορούμε να συγκεντρώσουμε δεδομένα των οποίων οι τιμές που συλλέχθηκαν ανταποκρίνονται στην πραγματικότητα και ακολουθούν συγκεκριμένες προδιαγραφές. Είναι σύνηθες φαινόμενο τα σύνολα δεδομένα που συλλέγονται να περιέχουν πολλές κενές τιμές ή μη ρεαλιστικές τιμές. Για παράδειγμα αναφέρουμε ότι εσωτερικές βλάβες σε όργανα μέτρησης θερμοκρασίας μπορεί να οδηγήσουν σε καταγραφή λανθασμένων θερμοκρασιών.

- Ενσωμάτωση δεδομένων: Με τον όρο ενσωμάτωση δεδομένων (data integration) [6] περιγράφουμε το πρόβλημα της δημιουργίας μίας ενιαίας όψης δεδομένων που προέρχονται από δύο ή περισσότερες πηγές. Η ενσωμάτωση πληροφορίας από διαφορετικές πηγές μπορεί να αποδειχθεί μία αρκετά δύσκολη διαδικασία. Προβλήματα όπως:

  - δεδομένα σε διαφορετική μορφοποίηση
  - δεδομένα σε διαφορετική κλίμακα
  - δεδομένα με κενές και λάθος τιμές
  - διπλότυπα δεδομένα
  - δεδομένα που διαχωρίζονται με διαφορετικά σύμβολα

  δημιουργούν σοβαρά εμπόδια στην ενοποίηση των δεδομένων.

- Θέματα ασφάλειας και ιδιωτικότητας: Η συλλογή, διαχείριση και αποθήκευση δεδομένων κυρίως όσον αφορά τους οικιακούς χρήστες απαιτεί ιδιαίτερη προσοχή καθώς πρέπει να λαμβάνεται υπόψιν η ισχύουσα νομοθεσία κάθε κράτους γύρω από την προστασία των προσωπικών δεδομένων. Στην Ευρώπη ισχύει ο Γενικός Κανονισμός για την Προστασία των Δεδομένων

(GDPR) [7] που πρέπει να λαμβάνεται υπόψιν τόσο για τη συλλογή, όσο και για την επεξεργασία των δεδομένων.

- Εύρεση εξειδικευμένου προσωπικού. Η ραγδαία ανάπτυξη νέων τεχνολογιών σε συνδυασμό με την πολύ μεγάλη ανάγκη για την επεξεργασία και ανάλυση δεδομένων δημιουργεί πολλά κενά για κάλυψη θέσεων εργασίας στον χώρο του data analysis. Αν αναλογιστούμε το γεγονός ότι μελετάμε ενεργειακά δεδομένα, τότε η εύρεση του αντίστοιχου προσωπικού το οποίο θα πρέπει να διαθέτει και γνώσεις πάνω στον ενεργειακό τομέα καθίσταται ακόμα πιο δύσκολη υπόθεση.

2.3 Τεχνητή νοημοσύνη και ενέργεια

Η τεχνητή νοημοσύνη καταλαμβάνει μεγάλο μέρος της σύγχρονης ζωής με το διαδίκτυο των πραγμάτων (Internet of things) να εξαπλώνεται διαρκώς δίνοντας καλύτερη δυνατότητα σύνδεσης στις οικιακές συσκευές με το χρήστη. Το διαδίκτυο των πραγμάτων αναφέρεται σε πράγματα που διαθέτουν την ικανότητα να επικοινωνούν μεταξύ τους καθώς και με άλλες συσκευές [8]. Παραδείγματα της εξάπλωσης του διαδικτύου των πραγμάτων είναι η αυτόματη ρύθμιση της θερμοκρασίας και κατανάλωσης ηλεκτρικής ενέργειας, η δυνατότητα που έχει πλέον ένας χρήστης να ρυθμίζει κάμερες, φώτα ή κλιματιστικά από απόσταση μέσω υπολογιστικών συστημάτων, αλλά και μία σειρά από αισθητήρες που βρίσκονται σε οικιακές συσκευές και παρέχουν χρήσιμα δεδομένα.

Η εξάπλωση όλων αυτών των συσκευών και η συνεχής εξέλιξη του διαδικτύου των πραγμάτων έχει οδηγήσει στην κατασκευή όλο και περισσότερων έξυπνων σπιτιών. Τα έξυπνα σπίτια διαθέτουν ένα σύστημα αυτοματισμού (home automation) που μπορεί να δέχεται φωνητικές εντολές και να ελέγχει οικιακές λειτουργίες, όπως ο φωτισμός και ο κλιματισμός αλλά ακόμα και συστήματα ελέγχου όπως συναγερμοί. Η κατασκευή έξυπνων σπιτιών μπορεί να συμβάλει στην καλύτερη παρακολούθηση και εξοικονόμηση ενεργειακής κατανάλωσης.

Σε αυτό το κεφάλαιο κάναμε επίσης και μία σύντομη αναφορά στα smart grids. Ένα smart grid χρησιμοποιεί smart meters και άλλες συσκευές για να παρακολουθεί τις ενεργειακές ανάγκες των καταναλωτών, την παραγωγή από ανανεώσιμες πηγές ενέργειας, τα έξυπνα συστήματα διανομής ενέργειας και τις περιόδους υψηλής και χαμηλής ενεργειακής κατανάλωσης. Τα έξυπνα ηλεκτρικά δίκτυα έχουν επίσης ως στόχο:

- να μειώσουν τις ενεργειακές απώλειες
- να μειώσουν το ενεργειακό αποτύπωμα της παραγωγής ενέργειας
- να παρέχουν στο χρήστη πληροφορίες και συμβουλές σχετικά με την ενεργειακή του κατανάλωση
- να βελτιώσουν τις παρεχόμενες υπηρεσίες
- να βελτιστοποιήσουν τη διαχείριση της ενέργειας

Είναι λοιπόν σαφές ότι τα συστήματα παρακολούθησης και ελέγχου που αναπτύσσονται στα έξυπνα δίκτυα υποστηρίζονται από την ενσωμάτωση τεχνητής νοημοσύνης. Η συγκέντρωση πληροφορίας σε ένα τέτοιο δίκτυο παίζει πολύ σημαντικό ρόλο στην αποδοτική λειτουργία του. Η πρόβλεψη ζήτησης ενέργειας από τους καταναλωτές αλλά και η πρόβλεψη παραγωγής ενέργειας από ανανεώσιμες πηγές μπορεί να πραγματοποιηθεί αποτελεσματικά με χρήση μοντέλων μηχανικής μάθησης και νευρωνικών δικτύων. Ακόμη, η ανάπτυξη πληροφοριακών συστημάτων που κάνουν χρήση αυτών των τεχνολογιών μπορεί να λειτουργήσει υποστηρικτικά στην ανάγκη για αποτελεσματική διαχείριση πληροφορίας και ενέργειας στα έξυπνα δίκτυα.

2.4 Stakeholders

Η παρακολούθηση ζήτησης και παραγωγής ενέργειας, η ταξινόμηση κτηρίων σε ενεργειακές κλάσεις όπως και άλλα ενεργειακά προβλήματα που περιγράφηκαν παραπάνω έχουν πολλούς ενδιαφερόμενους φορείς. Η ανάπτυξη εργαλείων που ενσωματώνει μοντέλα μηχανικής μάθησης για τα προβλήματα που σημειώθηκαν βοηθά αρκετές εμπλεκόμενες ομάδες να βγάλουν χρήσιμα συμπεράσματα με σκοπό τη λήψη αποφάσεων και την αντιμετώπιση ενεργειακών ζητημάτων. Πιο συγκεκριμένα, στο πληροφοριακό σύστημα που αναπτύξαμε βοηθάμε στην ταξινόμηση έργων σε ενεργειακές κλάσεις, στην πρόβλεψη εξοικονόμησης ενέργειας σε ορισμένους τομείς όπως οι βιομηχανίες και τα κτήρια, που προκύπτει ύστερα από αντίστοιχες παρεμβάσεις, καθώς και στην πρόβλεψη παραγωγής ενέργειας φωτοβολταϊκών. Τα ενδιαφερόμενα μέρη ενός τέτοιου πληροφοριακού συστήματος παρουσιάζονται ακολούθως:

- Στελέχη επιχειρήσεων ενέργειας. Εταιρείες που δραστηριοποιούνται στο χώρο των ανανεώσιμων πηγών ενέργειας ή ακόμα και μικρότερες εταιρείες που ασχολούνται με την ενεργειακή αναβάθμιση κτηρίων θα μπορούν να παρέχουν στα στελέχη τους ένα εργαλείο για γρήγορη ανάλυση των δεδομένων τους.

- Οικονομικά στελέχη επιχειρήσεων. Τα κόστη ανακατασκευών εταιρικών κτηρίων και βιομηχανιών πρέπει να μελετηθούν σε σχέση με τη μελλοντική μείωση των εξόδων που θα προκύψει από μικρότερη κατανάλωση ενέργειας.

- Μελλοντικοί Επενδυτές. Άτομα ή επιχειρήσεις που ενδιαφέρονται να επενδύσουν στο χώρο των φωτοβολταϊκών έχουν τη δυνατότητα να εξετάσουν την πρόβλεψη παραγωγής ενέργειας που μπορούν να επιτύχουν.

- Κρατικοί Οργανισμό και Υπηρεσίες. Οργανισμοί και Υπηρεσίες του κράτους που δραστηριοποιούνται στον τομέα της ενέργειας και του περιβάλλοντος μπορούν να μελετήσουν τη συνεισφορά των φωτοβολταϊκών στην παραγωγή ενέργειας ή να προβλέψουν την εξοικονόμηση

ενέργειας για ανακαίνιση κτηρίων. Πολλά προγράμματα της ευρωπαϊκής ένωσης διαθέτουν πόρους με στόχο την καλύτερη ενεργειακή απόδοση των κτηρίων και την αναβάθμισή της ενεργειακής τους κλάσης.

- Οικιακοί χρήστες. Καταναλωτές που ενδιαφέρονται να προχωρήσουν σε ανακαίνιση των σπιτιών τους με στόχο την εξοικονόμηση ενέργειας αλλά και την αναβάθμιση της ενεργειακής κλάσης των κατοικιών τους.

2.5 Επίλογος

Σε αυτό το κεφάλαιο παρουσιάσαμε μερικές από τις προκλήσεις που αντιμετωπίζουμε στο χώρο των δεδομένων και της ενέργειας. Αναφερθήκαμε στην τεράστια συλλογή δεδομένων που έχουμε στη διάθεσή μας σήμερα καθώς και σε ορισμένες από τις πηγές δεδομένων στο χώρο της ενέργειας, όπως smart meters, οικιακές συσκευές, καιρικά και γεωγραφικά δεδομένα, αλλά και γενικά δεδομένα που συλλέγονται από καταγραφή παραγωγής ενεργειακών πηγών. Παράλληλα, σημειώσαμε κάποια από τα προβλήματα που συναντάμε στη διαχείριση των big data, αναφέροντας βασικά ζητήματα όπως η ενσωμάτωση πληροφορίας, η προστασία των προσωπικών δεδομένων, η διάθεση ποιοτικών δεδομένων και η εύρεση κατάλληλα εξειδικευμένου προσωπικού για την αξιοποίησή τους, ιδιαίτερα στον τομέα της ενέργειας που εξετάζουμε.

Επιπλέον παρουσιάστηκε η χρήση συστημάτων που διαθέτουν τεχνητή νοημοσύνη και είναι ιδιαίτερα διαδεδομένα στον ενεργειακό τομέα. Τονίστηκε ο ρόλος του διαδικτύου των πραγμάτων στη μείωση κατανάλωσης της ενέργειας μέσα από την εξάπλωση των έξυπνων συσκευών. Στη συνέχεια είδαμε την ανάγκη για μετάβαση σε έξυπνα ηλεκτρικά δίκτυα, τα οποία θα συμβάλλουν στη μείωση της κατανάλωσης, και τονίσαμε τη συμβολή που έχει η τεχνητή νοημοσύνη για την ανάπτυξη και βελτίωσή τους.

Τέλος είδαμε τα εμπλεκόμενα μέρη που θα ωφεληθούν από ένα πληροφοριακό σύστημα για την εκπαίδευση μοντέλων μηχανικής μάθησης το οποίο θα παρέχει τη δυνατότητα προβλέψεων παραγωγής ενέργειας από φωτοβολταϊκά, πρόβλεψη εξοικονόμησης ενέργειας ανακαινισμένων κτηρίων αλλά και ταξινόμηση ενεργειακών έργων σε κλάσεις. Οι εμπλεκόμενες ομάδες που αναφέραμε είναι τα στελέχη ενεργειακών εταιρειών, οικονομικά στελέχη, μελλοντικοί επενδυτές, κρατικές υπηρεσίες αλλά και οικιακοί χρήστες.

ΚΕΦΑΛΑΙΟ 3

ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

3.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο είδαμε μεταξύ άλλων το ρόλο των δεδομένων και της τεχνητής νοημοσύνης στην ενέργεια. Σημειώσαμε την ανάγκη για ύπαρξη συστημάτων τα οποία θα ενσωματώνουν μοντέλα μηχανικής μάθησης, τα οποία θα παρέχουν ουσιαστικές λύσεις για μία σειρά από ενεργειακά ζητήματα, όπως μοντέλα ταξινόμησης έργων σε ενεργειακές κλάσεις ή πρόβλεψης ζήτησης και παραγωγής ενέργειας.

Σε αυτό το κεφάλαιο αναλύουμε διάφορους αλγορίθμους μηχανικής μάθησης και αρχιτεκτονικές νευρωνικών δικτύων που χρησιμοποιούνται στο πληροφοριακό σύστημα που σχεδιάσαμε. Παράλληλα θα παρουσιάσουμε και τις μετρικές βάσει των οποίων αξιολογούμε την επίδοση των μοντέλων που σχεδιάσαμε.

Η μηχανική μάθηση αποτελεί πεδίο της τεχνητής νοημοσύνης που ασχολείται με τη μελέτη και την κατασκευή αλγορίθμων, οι οποίοι μπορούν να μαθαίνουν από δεδομένα και να κάνουν προβλέψεις ή να λαμβάνουν αποφάσεις σχετικά με αυτά. Με τον όρο τεχνητή νοημοσύνη αναφερόμαστε σε υπολογιστικά συστήματα τα οποία μιμούνται στοιχεία της ανθρώπινης συμπεριφοράς και υποδηλώνουν ευφυΐα. Τα νευρωνικά δίκτυα αποτελούν πεδίο της μηχανικής μάθησης και είναι δίκτυα υπολογιστικών κόμβων, διασυνδεδεμένων μεταξύ τους. Μερικά από τα πεδία εφαρμογής της τεχνητής νοημοσύνης είναι η επεξεργασία φωνής και γλώσσας, ιατρικές εφαρμογές όπως διάγνωση ασθενειών, οικονομικές εφαρμογές καθώς και πληθώρα άλλων τομέων της σύγχρονης τεχνολογίας.

3.2 Κατηγορίες προβλημάτων μηχανικής μάθησης

Η μηχανική μάθηση αποτελείται από τρεις κύριες κατηγορίες, την επιβλεπόμενη μάθηση, την μη επιβλεπόμενη μάθηση και την ενισχυτική μάθηση. Η επιβλεπόμενη μάθηση αναφέρεται σε προβλήματα όπου τα σύνολα δεδομένων εκπαίδευσης περιέχουν τόσο τα χαρακτηριστικά εισόδου όσο και την επιθυμητή τιμή εξόδου. Ο σκοπός της επιβλεπόμενης μάθησης είναι η παραγωγή μίας συνάρτησης ταξινόμησης ή πρόβλεψης, η οποία θα συσχετίζει τις τιμές χαρακτηριστικών ενός συνόλου δεδομένων σε ανάλογες τιμές ή κλάσεις. Η συνάρτηση αυτή καλείται μοντέλο ταξινόμησης, μοντέλο πρόβλεψης ή απλά ταξινομητής [9]. Οι δύο κατηγορίες επιβλεπόμενης μάθησης είναι η ταξινόμηση και η παλινδρόμηση.

Ταξινόμηση: Στα προβλήματα ταξινόμησης το υπολογιστικό σύστημα καλείται να αποδώσει στο δείγμα που εξετάζει μία διακριτή τιμή. Για παράδειγμα, σε μία δυαδική ταξινόμηση η απόφαση που θα έπρεπε να

λάβει το σύστημα είναι αν ένα δείγμα ανήκει στην κλάση 0 ή 1. Παρ' όλα αυτά, ένα πρόβλημα ταξινόμησης μπορεί να έχει δείγματα που πρέπει να ταξινομηθούν σε παραπάνω από δύο κλάσεις, όπως η τοποθέτηση σε ενεργειακές κλάσεις ανακαινισμένων έργων, που μελετάμε σε αυτήν την εργασία.

Παλινδρόμηση: Σε αυτήν την κατηγορία προβλημάτων οι αλγόριθμοι που αναπτύσσονται έχουν ως στόχο να προβλέψουν συνεχείς τιμές για τα δείγματα που εξετάζονται. Οι τιμές που προβλέπονται πρέπει να είναι όσο το δυνατόν πιο κοντά στις πραγματικές τιμές εξόδου των δειγμάτων. Παραδείγματα προβλημάτων παλινδρόμησης είναι η πρόβλεψη της τιμής μίας μετοχής ή η πρόβλεψη παραγωγής ενέργειας ενός φωτοβολταϊκού.

Η μη επιβλεπόμενη μάθηση μελετά τον τρόπο με τον οποίο υπολογιστικά συστήματα μαθαίνουν να αναπαριστούν μη χαρακτηρισμένα δεδομένα εισόδου σε μία δομή που αντανακλά τη στατιστική απεικόνιση των δεδομένων εισόδου [10]. Η ενισχυτική μάθηση είναι το πρόβλημα όπου ένα σύστημα πρέπει να μάθει να συμπεριφέρεται μέσα από την αλληλεπίδραση με το περιβάλλον του [11]. Στο πληροφοριακό σύστημα που αναπτύξαμε τα προβλήματα που παρουσιάζονται ανήκουν στην κατηγορία της επιβλεπόμενης μάθησης.

3.3 Αλγόριθμοι

Σε αυτήν την ενότητα μελετάμε ορισμένους από τους πιο γνωστούς αλγόριθμους που χρησιμοποιούνται στην επιβλεπόμενη μάθηση. Στα πλαίσια της μηχανικής μάθησης έχει αναπτυχθεί ένα ευρύ φάσμα αλγορίθμων, όμως εδώ θα αναφέρουμε αυτούς που δοκιμάστηκαν για τις ανάγκες της διπλωματικής, όπως τα δέντρα αποφάσεων, τα τυχαία δάση, τη μέθοδο gradient boosting, τους Κ-κοντινότερους γείτονες (Knn), τις μηχανές διανυσματικής στήριξης (SVM) και τη γραμμική παλινδρόμηση. Ακόμη θα μελετήσουμε ορισμένες αρχιτεκτονικές νευρωνικών δικτύων και θα αναφερθούμε στις μετρικές απόδοσης και σφάλματος που χρησιμοποιούμε για την αξιολόγηση των μοντέλων μας.

3.3.1 Δέντρα Αποφάσεων

Τα δέντρα αποφάσεων είναι μία μη παραμετρική μέθοδος επιβλεπόμενης μάθησης που χρησιμοποιείται για προβλήματα ταξινόμησης και παλινδρόμησης. Τα μοντέλα που κατασκευάζονται προβλέπουν την τιμή εξόδου μέσα από απλούς κανόνες οι οποίοι προκύπτουν με βάση τα χαρακτηριστικά των δεδομένων. Στα δέντρα αποφάσεων τα δείγματα χωρίζονται με βάση αυτούς τους κανόνες, μέχρι να καταλήξουν σε κάποιο κόμβο φύλλο του δέντρου, ο οποίος για ένα πρόβλημα ταξινόμησης θα συνεπάγεται και την αντίστοιχη κλάση. Τα δέντρα αποφάσεων έχουν το πλεονέκτημα ότι είναι εύκολα στην κατανόηση και τη δημιουργία και η πολυπλοκότητά τους είναι χαμηλή.

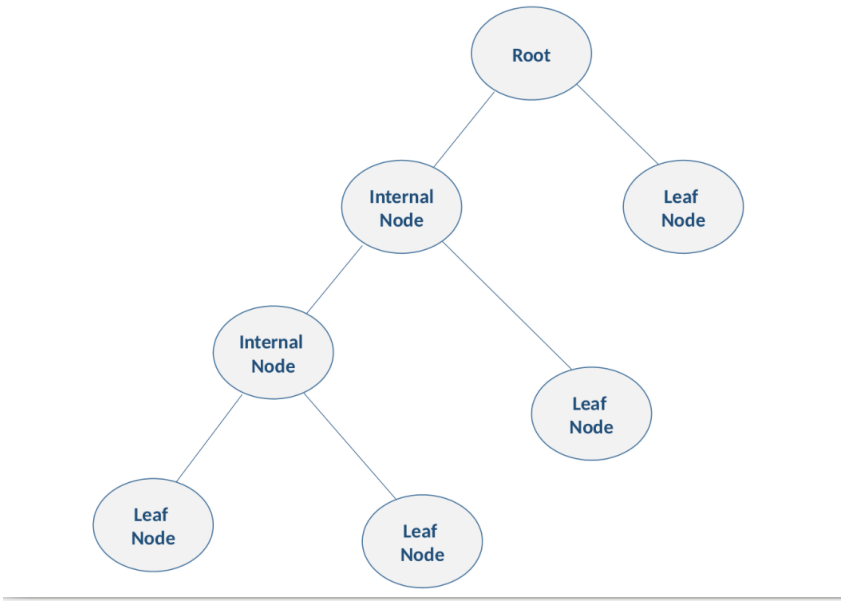

Figure 3.1 Decision Tree

Παρ' όλα αυτά, υπάρχει ο κίνδυνος το μοντέλο που θα δημιουργηθεί να υπερπορσαρμόζεται στα δεδομένα που δέχεται και να παρουσιάζει μεγάλη διακύμανση (overfittng). Η υπερπροσαρμογή είναι ένα βασικό θέμα στην επιβλεπόμενη μάθηση που μας εμποδίζει να γενικεύουμε τα μοντέλα από τα δεδομένα που έχουν εκπαιδευτεί σε δεδομένα που δεν έχουν δει [12]. Παρακάτω παρουσιάζονται οι εικόνες τριών μοντέλων, όπου στην πρώτη περίπτωση το μοντέλο δεν προσαρμόζεται καλά στα δεδομένα εκπαίδευσης και παρουσιάζει μεγάλη προκατάληψη (underfitting), στη δεύτερη περίπτωση το μοντέλο θα λειτουργήσει καλύτερα σε νέα δεδομένα και στην τρίτη παρατηρείται υπερπροσαρμογή στα δεδομένα εκπαίδευσης.

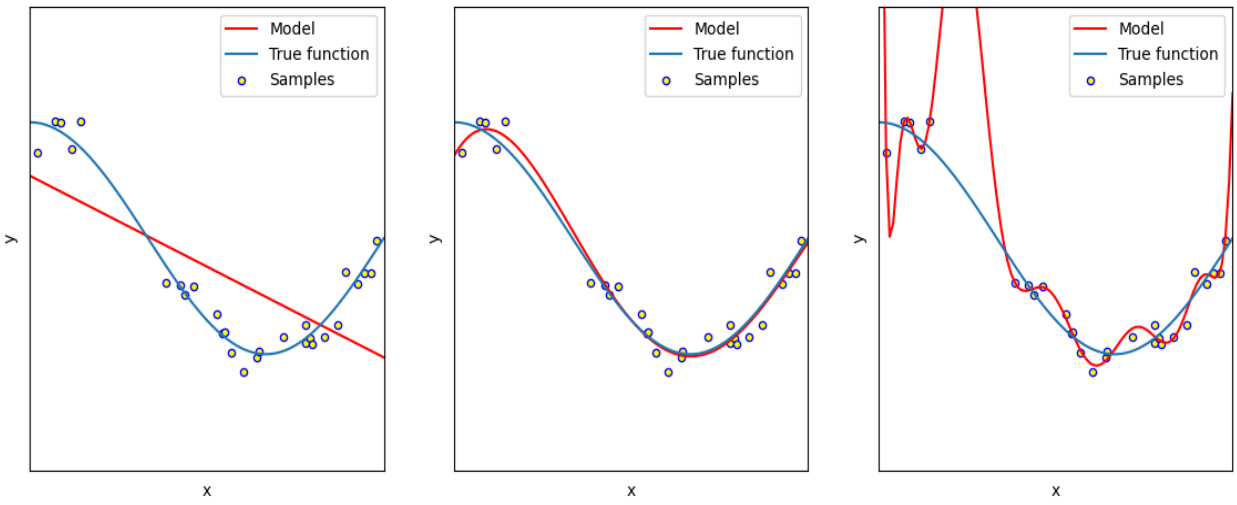

Figure 3.2 Underfitting, well-designed model, overfitting

Αν για το μοντέλο που σχεδιάσουμε επιλέξουμε το βάθος του δέντρου να είναι μεγάλο τότε υπάρχει ο κίνδυνος της υπερπροσαρμογής, ενώ αν επιλέξουμε μικρό βάθος τότε το μοντέλο μπορεί να μην προσαρμόζεται καλά στα δεδομένα εκπαίδευσης. Ο διαχωρισμός των δειγμάτων στους κόμβους των δέντρων περιγράφεται παρακάτω [13]:

Έστω D_m τα δείγματα που βρίσκονται στον κόμβο m με χαρακτηριστικά x_i, i=1,2,..,n και διάνυσμα τιμών εξόδου y. Για κάθε υποψήφιο διαχωρισμό θ = (j, t_m) όπου j κάποιο χαρακτηριστικό και t_m το κατώφλι για τον κόμβο m, διαχωρίζουμε τα δείγματα στο δεξί και στο αριστερό παιδί του κόμβου με βάση τις ακόλουθες σχέσεις:

$$D_l(\theta) = \{(x, y)|x_j \leq t_m\}$$

$$D_r(\theta) = D_m \setminus D_l(\theta)$$

Η αποτελεσματικότητα ενός υποψήφιου διαχωρισμού γίνεται βάση μίας συνάρτησης κόστους. Η εύρεση του βέλτιστου δέντρου απόφασης είναι αρκετά δύσκολη διαδικασία, γι' αυτό και αρκετά μοντέλα δέντρων απόφασης βασίζονται σε ευριστικούς ή σε άπληστους αλγορίθμους για την κατασκευή τους.

3.3.2 Τυχαία Δάση (Random Forest)

Τα τυχαία δάση προσφέρουν σημαντική βελτίωση στις επιδόσεις προβλημάτων επιβλεπόμενης μάθησης και έχουν ως πλεονέκτημα την αποφυγή της υπερεκπαίδευσης. Τα τυχαία δάση είναι ένας συνδυασμός από δέντρα αποφάσεων επιτρέποντας στα μοντέλα που χρησιμοποιούν αυτήν την τεχνική να έχουν καλύτερη γενίκευση. Σε ένα πρόβλημα ταξινόμησης τα δέντρα αυτής της μεθόδου έχουν από μία ψήφο και στο τέλος το δείγμα ταξινομείται στη δημοφιλέστερη κλάση [14]. Η επιλογή περισσότερων δέντρων μπορεί να έχει ως αποτέλεσμα καλύτερες προβλέψεις, όμως με αυτόν τον τρόπο αυξάνεται η πολυπλοκότητα. Είναι σύνηθες να χρησιμοποιείται η τεχνική bagging ή αλλιώς bootstrap aggregation [15] ώστε κάθε δέντρο απόφασης να λαμβάνει ως είσοδο ένα υποσύνολο από δείγματα με τυχαίο τρόπο, ενδεχομένως κάποιο δείγμα να εμφανίζεται και παραπάνω από μία φορά, και όχι το σύνολο των δειγμάτων εκπαίδευσης.

3.3.3 Gradient Boosting

Η τεχνική gradient boosting είναι μία ευρέως χρησιμοποιούμενη μέθοδος που συνδυάζει περισσότερα μοντέλα (ensemble) τόσο για προβλήματα ταξινόμησης όσο και για προβλήματα παλινδρόμησης. Για προβλήματα ταξινόμησης, η μέθοδος αυτή λαμβάνει ένα σύνολο από αδύναμους ταξινομητές, όπως δέντρα αποφάσεων, και επικεντρώνεται στην κατασκευή ενός πιο ισχυρού ταξινομητή.

Σε αντίθεση με τα τυχαία δάση όπου δημιουργούνται παράλληλα τυχαία δέντρα και το αποτέλεσμα προκύπτει από έναν μέσο όρο, η κατασκευή ενός μοντέλου gradient boosting γίνεται προσθέτοντας τους αδύναμους ταξινομητές ακολουθιακά. Σε κάθε στάδιο η προσθήκη ενός ταξινομητή λαμβάνει υπόψιν το σφάλμα που έχει υπολογιστεί μέχρι και το προηγούμενο στάδιο και εκπαιδεύεται με βάση αυτό [16]. Υπάρχουν διάφορες υλοποιήσεις αυτής της μεθόδου, όπως η XGBoost, που όμως η κλιμακωσημότητά τους δεν είναι ικανοποιητική. Η υλοποίηση LightGBM [17] επιτρέπει ταχύτερη εκπαίδευση των δεδομένων και υψηλή αποτελεσματικότητα.

3.3.4 Κ-πλησιέστεροι γείτονες (k-nearest neighbors)

Ο αλγόριθμος k-nearest neighbors είναι και αυτός μία μη παραμετρική μέθοδος επιβλεπόμενης μάθησης τόσο για προβλήματα ταξινόμησης όσο και για προβλήματα παλινδρόμησης. Πρόκειται για έναν από τους πιο απλούς και θεμελιώδεις αλγορίθμους που χρησιμοποιείται ιδιαίτερα σε προβλήματα ταξινόμησης [18]. Οι k κοντινότεροι γείτονες καθορίζονται με μία συνάρτηση απόστασης, με την Ευκλείδια απόσταση να είναι η ευρύτερα χρησιμοποιούμενη. Η Ευκλείδια απόσταση δύο σημείων x και y δίνεται από τον παρακάτω τύπο:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

Η επιλογή του αριθμού k παίζει σημαντικό ρόλο στην κατασκευή του μοντέλου, καθώς ένας μικρός αριθμός k μπορεί να προκαλέσει υπερπροσαρμογή του μοντέλου ενώ ένας μεγάλος αριθμός μπορεί να οδηγήσει στην κατασκευή ενός μοντέλου που δίνει λίγη σημασία στα δεδομένα εκπαίδευσης. Πέρα από την επιλογή του αριθμού των κοντινότερων γειτόνων και της συνάρτησης που αυτοί θα υπολογιστούν, σημαντική παράμετρος είναι αν οι γείτονες έχουν όλοι το ίδιο βάρος στην απόφαση ταξινόμησης ή το βάρος τους κυμαίνεται ανάλογα με την απόστασή τους από το δείγμα.

3.3.5 Μηχανές Διανυσματικής Στήριξης SVM

Μία ακόμα μέθοδος ταξινόμησης και παλινδρόμησης με μεγάλη αποτελεσματικότητα και ευρεία χρήση είναι οι μηχανές διανυσματικής στήριξης [19]. Για την περίπτωση της ταξινόμησης σε δύο διαχωρίσιμες κλάσεις οι μηχανές διανυσματικής στήριξης έχουν ως στόχο την εύρεση ενός υπερεπιπέδου (hyperplane)

$$g(x) = w^T x + w_0 = 0$$

με x_i, i = 1,2,…,N τα διανύσματα χαρακτηριστικών του συνόλου δεδομένων εκπαίδευσης, το οποίο ταξινομεί σωστά τα δείγματα. Παράλληλα αυτό το υπερεπίπεδο πρέπει να αφήνει το μέγιστο δυνατό περιθώριο (margin) και από τις δύο κλάσεις ώστε το μοντέλο να μην δείχνει προτίμηση σε κάποια από τις δύο.

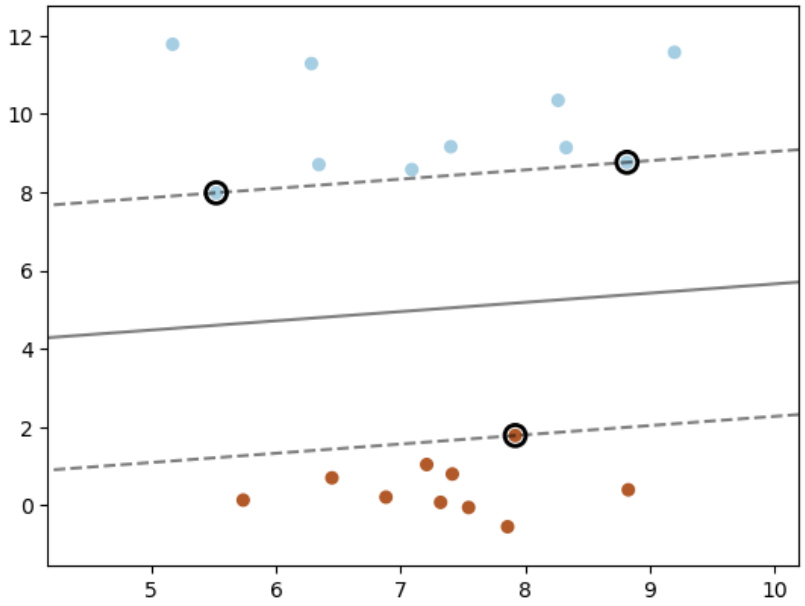

Figure 3.3 Hyperplane

Τα δείγματα ταξινομούνται στην μία κλάση στην περίπτωση που ισχύει η σχέση:

$$g(x) = w^T x + w_0 \geq 1$$

ενώ ταξινομούνται στην δεύτερη κλάση στην περίπτωση που ισχύει:

$$g(x) = w^T x + w_0 \leq -1$$

Τα διανύσματα τα οποία βρίσκονται πάνω σε ένα από τα δύο υπερεπίπεδα που σημειώνονται με διακεκομμένες γραμμές, δηλαδή για τα διανύσματα για τα οποία ισχύει η σχέση:

$$g(x) = w^T x + w_0 = \pm 1$$

ονομάζονται διανύσματα στήριξης [20].

3.3.6 Γραμμική Παλινδρόμηση

Τα μοντέλα γραμμικής παλινδρόμησης χρησιμοποιούνται για να προβλέψουν μία τιμή σε προβλήματα παλινδρόμησης. Η τιμή που θέλουμε να προβλέψουμε ονομάζεται εξαρτημένη μεταβλητή ενώ χαρακτηρίζουμε ανεξάρτητες μεταβλητές, τις μεταβλητές που χρησιμοποιούμε για να γίνει η πρόβλεψη. Για παράδειγμα, αν θέλουμε να προβλέψουμε την τιμή ενός σπιτιού, η τιμή του σπιτιού θα ήταν η εξαρτημένη μεταβλητή και τα τετραγωνικά, το έτος κατασκευής και ο όροφος θα αποτελούσαν τις ανεξάρτητες μεταβλητές. Τα μοντέλα γραμμικής παλινδρόμησης είναι αρκετά ευαίσθητα στις ακραίες τιμές, γι' αυτό και είναι καλό να γίνεται προεπεξεργασία στα δεδομένα εκπαίδευσης που δίνονται σε αυτά τα μοντέλα. Η εξίσωση ενός μοντέλου γραμμικής παλινδρόμησης περιγράφεται από τον ακόλουθο τύπο:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

Όπου y_i είναι η τιμή που προβλέπουμε, x_i, i=1,2,..,N οι ανεξάρτητες μεταβλητές και ε το σφάλμα της εξίσωσης $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$ από την τιμή των δειγμάτων. Παρακάτω φαίνεται ένα παράδειγμα μοντέλου γραμμικής παλινδρόμησης.

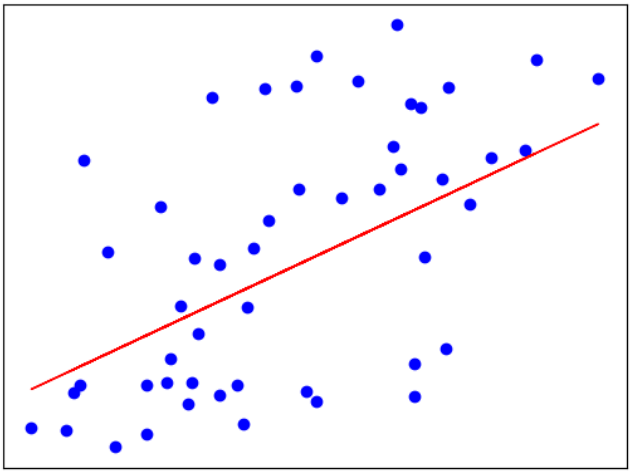

Figure 3.4 Linear Regression Model

3.4 Νευρωνικά Δίκτυα

Τα τεχνητά νευρωνικά δίκτυα ή αλλιώς νευρωνικά δίκτυα είναι αρχικώς εμπνευσμένα από μελέτες που έχουν γίνει σχετικά με τους μηχανισμούς που το βιολογικό νευρικό σύστημα και συγκεκριμένα ο ανθρώπινος εγκέφαλος επεξεργάζονται μία πληροφορία [21]. Ένας νευρώνας αποτελεί μία μονάδα επεξεργασίας πληροφορίας, όπου δέχεται μία είσοδο και παράγει μία έξοδο με βάση μία συνάρτηση ενεργοποίησης. Πιο αναλυτικά, ένα σύνολο συνάψεων με βάρη w_{jk} πολλαπλασιάζεται με κάποια σήματα εισόδου w_j και το άθροισμά τους μαζί με μία πόλωση b_k δίνεται ως είσοδος στον νευρώνα k. Στη συνέχεια ο νευρώνας παράγει μία έξοδο y μέσα από μία συνάρτηση ενεργοποίησης φ. Παρακάτω δίνουμε τις εξισώσεις για την είσοδο και την έξοδο του νευρώνα.

$$u_k = \sum_{j=1}^{n} w_{jk} x_j$$

$$y_k = \varphi(u_k + b_k)$$

Στο σχήμα 3.5 παρουσιάζουμε και μία αναπαράσταση του μοντέλου ενός νευρώνα.

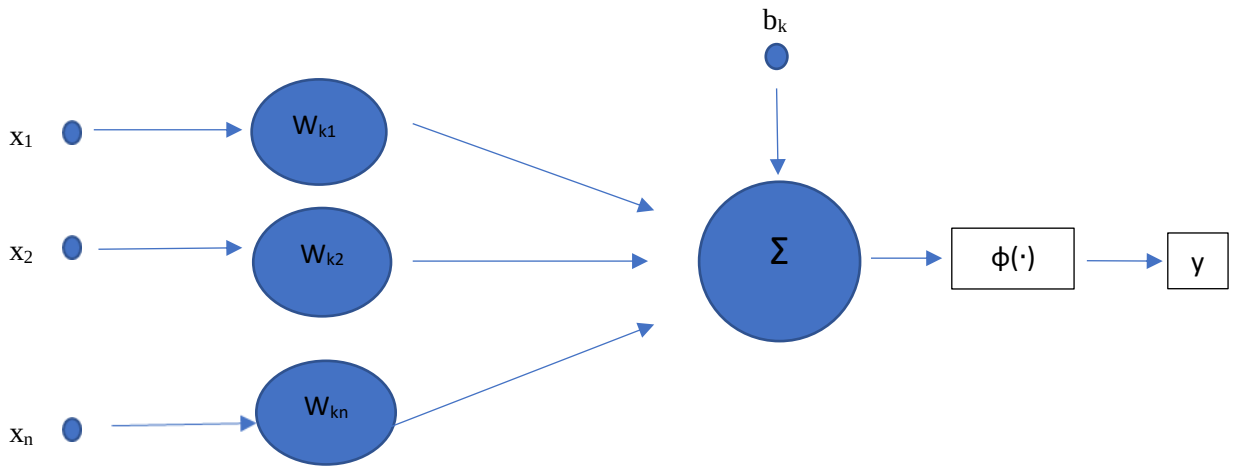

Figure 3.5 Neuron Model

Ένα νευρωνικό δίκτυο μπορεί να αποτελείται από πολλά στρώματα τέτοιων νευρώνων, όπου η έξοδος ενός νευρώνα αποτελεί είσοδο για έναν ή παραπάνω νευρώνες του επόμενου στρώματος. Στην περίπτωση ενός νευρωνικού δικτύου με πολλά τέτοια στρώματα γίνεται λόγος για βαθύ νευρωνικό δίκτυο.

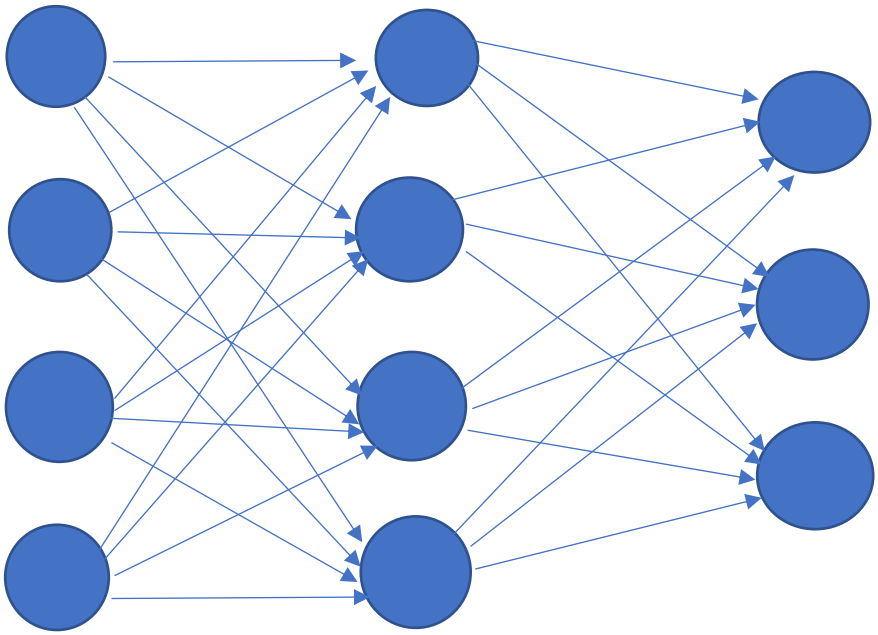

Figure 3.6 Neural Network's Layers

3.4.1 Συναρτήσεις Ενεργοποίησης

Όπως αναφέραμε προηγουμένως η έξοδος ενός νευρώνα προκύπτει ύστερα από την εφαρμογή μίας συνάρτησης ενεργοποίησης. Η συνάρτηση ενεργοποίησης παίζει μεγάλο ρόλο στη διαδικασία εκπαίδευσης, ιδιαίτερα στα βαθιά νευρωνικά δίκτυα [22]. Θα παρουσιάσουμε τρεις από αυτές, οι οποίες χρησιμοποιήθηκαν και στο πληροφοριακό σύστημα που παρουσιάζουμε, τη συνάρτηση ReLU (rectified linear unit), τη συνάρτηση Sigmoid και τη συνάρτηση Softmax. Η συνάρτηση ReLU [23] αποτελεί την πιο δημοφιλή συνάρτηση ενεργοποίησης όσον αφορά στη χρήση της σε βαθιά νευρωνικά δίκτυα. Η συνάρτηση ReLU ορίζεται ως εξής:

$$f(x) = \begin{cases} x & if \quad x > 0 \\ 0 & otherwise \end{cases}$$

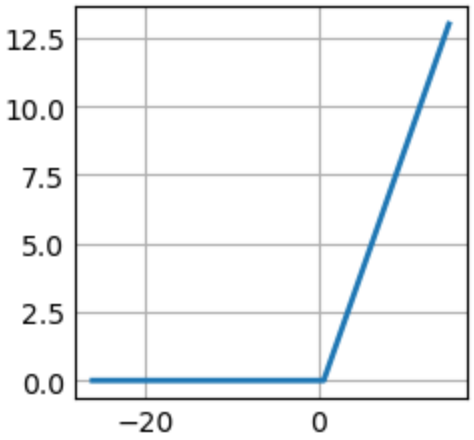

Figure 3.7 ReLU

Η συνάρτηση Sigmoid δίνεται από τον ακόλουθο τύπο:

$$S(x) = \frac{1}{1 + e^{-x}}$$

Την αναπαριστάμε στο ακόλουθο σχήμα:

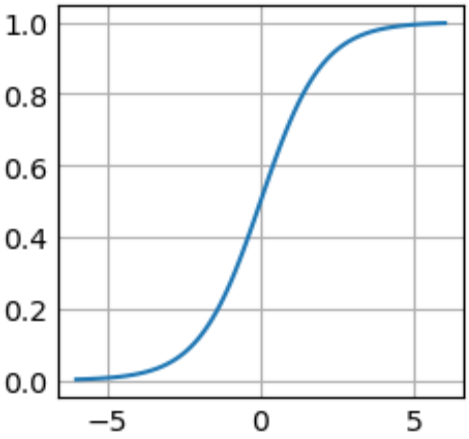

Figure 3.8 Sigmoid

Και τέλος η συνάρτηση Softmax παρουσιάζεται παρακάτω:

$$S(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

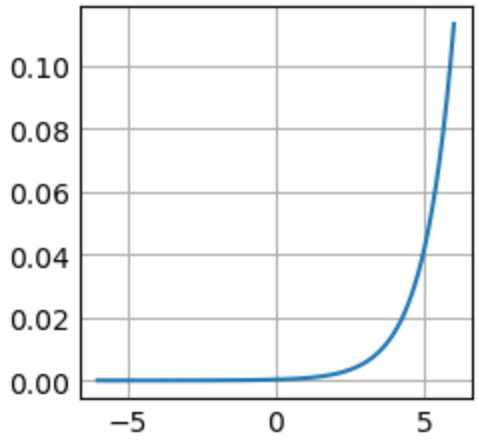

Figure 3.9 Softmax

3.4.2 Αλγόριθμος Backpropagation

Ο αλγόριθμος backpropagation βρίσκεται στον πυρήνα των νευρωνικών δικτύων και αποτελεί μία πολύ σημαντική μέθοδο για τον υπολογισμό των βαρών και των σταθερών που ελαχιστοποιούν το κόστος της συνάρτησης σφάλματος [24]. Κατά τη διαδικασία εκπαίδευσης λαμβάνουμε το σφάλμα που προκύπτει στην έξοδο του δικτύου σε σχέση με την πραγματική έξοδο χρησιμοποιώντας μία μετρική, τις οποίες θα δούμε στην επόμενη ενότητα. Στόχος είναι να ελαχιστοποιήσουμε αυτό το σφάλμα μεταβάλλοντας τα βάρη και τις σταθερές (biases). Για μία συνάρτηση πολλών μεταβλητών η κλίση της μας δίνει την κατεύθυνση που πρέπει να κινηθούμε για να αυξηθεί η τιμή της συνάρτησης. Στην περίπτωση μίας συνάρτησης κόστους πολλών μεταβλητών $C(x_1, x_2,\ldots,x_n)$ ενδιαφερόμαστε για την κατεύθυνση που μειώνεται η τιμή της συνάρτησης, συνεπώς μας ενδιαφέρει η αρνητική κλίση (gradient descent) $-\nabla C(x_1,x_2,\ldots,x_n)$. Σε κάθε βήμα του αλγορίθμου backpropagation υπολογίζουμε την gradient descent με βάση τα δεδομένα εκπαίδευσης και ανανεώνουμε τα βάρη και τις σταθερές με βάση κάποιο βήμα η που ορίζουμε. Η τελική τιμή των βαρών και σταθερών προκύπτει ύστερα από την τελευταία επανάληψη του αλγορίθμου.

3.4.3 Συνελικτικά Νευρωνικά Δίκτυα (CNNs)

Τα συνελικτικά νευρωνικά δίκτυα αποτελούν έναν από τους σημαντικότερους τομείς στη βαθιά μηχανική μάθηση με εντυπωσιακά αποτελέσματα σε πεδία όπως η όραση υπολογιστών και η επεξεργασία φωνής και γλώσσας [25]. Ένα συνελικτικό νευρωνικό δίκτυο έχει ενδιάμεσα στρώματα (hidden layers) τα οποία βασίζονται στη συνέλιξη (convolutional layers). Μία τέτοια αρχιτεκτονική προσφέρει σημαντικά πλεονεκτήματα σε σύγκριση με ένα απλό νευρωνικό δίκτυο όπως η μείωση των υπολογισμών που απαιτούνται. Σε κάθε convolutional layer εφαρμόζεται κάποιο φίλτρο, το οποίο έχει κάποιες διαστάσεις σε μορφή πίνακα. Παραδείγματος χάριν, μπορούμε να θεωρήσουμε έναν πίνακα 3X3 σαν ένα φίλτρο, το οποίο θα πολλαπλασιαστεί με αντίστοιχους πίνακες 3X3 των δεδομένων εισόδων. Αυτός ο πολλαπλασιασμός πινάκων έχει ως στόχο να εξάγει χρήσιμα συμπεράσματα, όπως θα μπορούσε να είναι η ανίχνευση ακμών ή γωνιών σε δεδομένα εικόνας. Είναι σύνηθες σε ένα συνελικτικό νευρωνικό δίκτυα να υπάρχουν ενδιάμεσα στρώματα όπως pooling layers, τα οποία έχουν ως στόχο να μειώσουν τις διαστάσεις εισόδου. Στα μοντέλα που σχεδιάσαμε χρησιμοποιούμε max pooling layers. Ένα max pooling layer 2X2 θα κρατούσε μόνο την μεγαλύτερη τιμή εισόδου στις διαστάσεις 2X2 που θα εφαρμοζόταν. Το average pooling layer θα επέστρεφε το μέσο όρο των τιμών για τις τιμές εισόδου που θα εφαρμοζόταν. Παρακάτω παρουσιάζουμε την αρχιτεκτονική ενός συνελικτικού νευρωνικού δικτύου, όπου υπάρχουν δύο ενδιάμεσα συνελικτικά στρώματα και δύο max pooling layers.

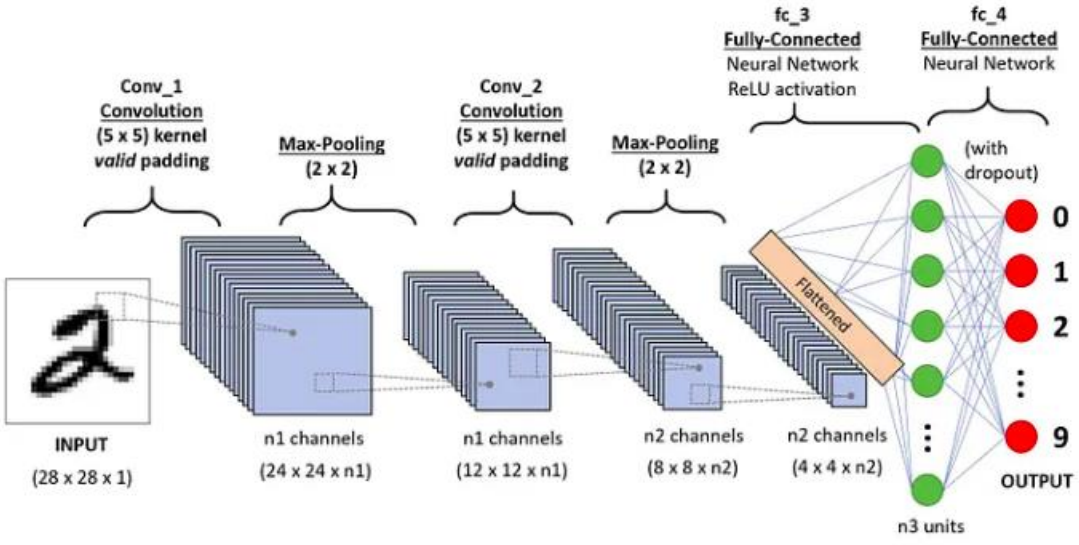

Figure 3.10 A CNN to classify handwritten digits, Source[2]: towardsdatascience

[2] https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

3.4.4 Νευρωνικά Δίκτυα με Ανάδραση (RNNs)

Τα νευρωνικά δίκτυα με ανάδραση (RNNs) έχουν την ιδιότητα, σε αντίθεση με τα νευρωνικά δίκτυα που είδαμε μέχρι εδώ, να διατηρούν μία μορφή μνήμης σχετικά με προγενέστερες πληροφορίες. Αυτή η ιδιότητα είναι πολύ σημαντική σε περιπτώσεις όπου μία πρόβλεψη είναι άμεση συνέπεια κάποιας προηγούμενης κατάστασης. Παραδείγματος χάριν, αν θέλουμε να προβλέψουμε την επόμενη λέξη μίας πρότασης οι αμέσως προηγούμενες λέξεις παίζουν πολύ σημαντικό ρόλο στην προσπάθεια εύρεσης της λέξης. Τα RNNs είναι δίκτυα με βρόχους ανατροφοδότησης όπου επιτρέπουν τη διατήρηση της πληροφορίας.

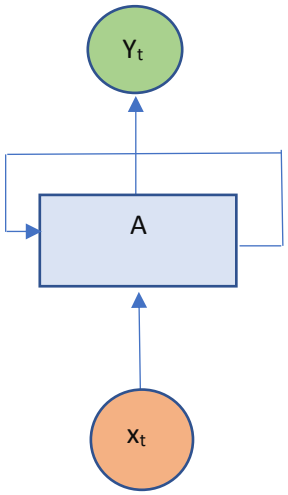

Figure 3.11 Network Loop

Όπως βλέπουμε στο παραπάνω σχήμα μία είσοδος x_t δίνεται σε ένα νευρωνικό δίκτυο A το οποίο παράγει μία Y_t έξοδο. Ύστερα θα ανατροφοδοτηθεί με το επόμενο στιγμιότυπο εισόδου x_{t+1} και την έξοδο Y_t που προέκυψε στο προηγούμενο βήμα. Μία δημοφιλής υποκατηγορία RNNs είναι τα νευρωνικά δίκτυα με ανάδραση διπλής κατεύθυνσης (BRNNs).

Τα νευρωνικά δίκτυα BRNNs μπορούν να εκπαιδευτούν χρησιμοποιώντας ταυτόχρονα πληροφορία τόσο από το παρελθόν όσο και από κάποια μελλοντική κατάσταση [26]. Με αυτόν τον τρόπο τα BRNNs καταφέρνουν να αποκτούν αυξημένη πληροφορία για την εκπαίδευσή τους. Η βασική διαδικασία για τη δημιουργία ενός BRNN είναι ο χωρισμός των νευρώνων σε δύο μέρη, ένα μέρος σε αυτούς που δείχνουν μία προηγούμενη κατάσταση και ένα μέρος σε αυτούς που δείχνουν μία μελλοντική κατάσταση.

3.4.5 Δίκτυα Μακράς Βραχύχρονης Μνήμης (LSTM)

Στην ενότητα αυτή θα παρουσιάσουμε τα νευρωνικά δίκτυα μακράς βραχύχρονης μνήμης (LSTMs). Τα κλασικά νευρωνικά δίκτυα με ανάδραση αποτυγχάνουν να διατηρήσουν πληροφορία για στιγμιότυπα που βρίσκονται πιο μακριά από 5 με 10 βήματα [27]. Τα νευρωνικά δίκτυα LSTMs αποτελούν μία υποκατηγορία RNN δικτύων τα οποία έχουν τη δυνατότητα να μαθαίνουν μακροχρόνιες εξαρτήσεις κατά την εκπαίδευσή τους. Η ροή πληροφορίας στην τρέχουσα κατάσταση (cell state) επηρεάζεται από τρεις πύλες, μία πύλη εισόδου (input gate), μία πύλη εξόδου (output gate) και μία πύλη όπου αποφασίζεται η πληροφορία που θα αποκλειστεί (forget gate). Οι πύλες αυτές βοηθούν στη διατήρηση ή απομάκρυνση πληροφορίας από την τρέχουσα κατάσταση.

3.5 Μετρικές απόδοσης και σφάλματος

Σε αυτήν την ενότητα αναφέρουμε τις μετρικές αξιολόγησης και σφάλματος που χρησιμοποιήσαμε για την αξιολόγηση των μοντέλων μας.

Accuracy Score

Η πρώτη μετρική που χρησιμοποιούμε για το πρόβλημα ταξινόμησης είναι η ακρίβεια (accuracy). Πρόκειται για μία πολύ απλή μετρική που μας δίνει το σύνολο των σωστά ταξινομημένων δειγμάτων προς το συνολικό αριθμό των δειγμάτων.

$$accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

Precision

Η δεύτερη μετρική αξιολόγησης που χρησιμοποιούμε είναι η μετρική precision. Για το παράδειγμα της δυαδικής ταξινόμησης έστω ότι έχουμε δύο κλάσεις Positive και Negative. Η μετρική precision ορίζεται ως το σύνολο των δειγμάτων που ταξινομήθηκαν στην κλάση Positive (True Positive) προς το σύνολο όλων των δειγμάτων που ταξινομήθηκαν στην κλάση Positive (True Positive + False Positive).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Μπορούμε ομοίως να γενικεύσουμε για περισσότερες από δύο κλάσεις. Σε αυτήν την περίπτωση θα επιλέξουμε να πάρουμε τον μέσο όρο για κάθε τιμή Precision που υπολογίστηκε για την αντίστοιχη κλάση.

Recall

Η τρίτη μετρική που χρησιμοποιούμε είναι η μετρική Recall. Σε αυτήν την περίπτωση αξιολογούμε το ποσοστό των δειγμάτων που ταξινομήθηκαν στην κλάση Positive προς το σύνολο των δειγμάτων που ανήκουν στην κλάση Positive,

$$Recall = \frac{True\ Positive}{Total\ Positive}$$

Όμοια και σε αυτήν την περίπτωση μπορούμε να γενικεύσουμε για παραπάνω από δύο κλάσεις και να υπολογίσουμε το μέσο όρο για κάθε κλάση.

F1 score

Η μετρική F1 score ορίζεται ως ο αρμονικός μέσος των μετρικών precision και recall,

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

με x_i να είναι οι τιμές εξόδου των δεδομένων και y_i οι τιμές που προβλέφθηκαν.

Mean Absolute Error (MAE)

Για τα προβλήματα παλινδρόμησης και πρόβλεψης η πρώτη μετρική που χρησιμοποιούμε είναι το μέσο απόλυτο σφάλμα (MAE) [28]. Εκφράζεται ως ο μέσος όρος της διαφοράς των τιμών των δειγμάτων από τις τιμές που προβλέφθηκαν

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Mean Squared Error (MSE)

Η δεύτερη μετρική σφάλματος με βάση την οποία αξιολογούμε τα μοντέλα μας είναι το μέσο τετραγωνικό σφάλμα το οποίο ορίζεται ως εξής:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}$$

RMSE

Η τρίτη μετρική αποτελεί την τετραγωνική ρίζα της μετρικής MSE,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}}$$

όπου y_i είναι οι τιμές που προβλέφθηκαν και x_i οι πραγματικές τιμές.

R2

Η μετρική R2 εκφράζεται από τον ακόλουθο τύπο:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Όπου $SS_{res} = \sum(x_i - y_i)^2$

Και $SS_{tot} = \sum(x_i - \bar{x})^2$

3.6 Επίλογος

Σε αυτό το κεφάλαιο είδαμε το θεωρητικό υπόβαθρο που χρειάζεται για την ανάπτυξη των μοντέλων που χρησιμοποιούνται στο πληροφοριακό σύστημα. Αναφέραμε τις κατηγορίες προβλημάτων μηχανικής μάθησης και επεκταθήκαμε στα προβλήματα επιβλεπόμενης μάθησης που μελετώνται κατά βάση σε αυτήν την εφαρμογή. Στα πλαίσια προβλημάτων ταξινόμησης και παλινδρόμησης έχει αναπτυχθεί πληθώρα αλγορίθμων, όμως η επιλογή των συγκεκριμένων αλγορίθμων που παρουσιάσαμε οφείλεται στο ότι είναι αυτοί που χρησιμοποιήθηκαν στα πλαίσια της διπλωματικής εργασίας.

Ακόμη είδαμε κάποιες αρχιτεκτονικές τεχνητών νευρωνικών δικτύων και στοιχεία της εσωτερικής δομής τους. Επιπλέον, σημειώσαμε ορισμένες συναρτήσεις ενεργοποίησης και παρουσιάσαμε τον αλγόριθμο backpropagation που χρησιμοποιείται για την ελαχιστοποίηση των σφαλμάτων στα νευρωνικά δίκτυα. Τέλος παρουσιάσαμε τις μετρικές accuracy score, precision, recall, f1score, mean absolute error, mean squared error, root mean squared error και R2 με τις οποίες θα αξιολογηθούν τα μοντέλα που δημιουργήσαμε και παρουσιάζουμε στη συνέχεια.

Στο επόμενο κεφάλαιο θα δούμε τα βήματα για τη διαδικασία πρόβλεψης που ακολουθούνται σχετικά με τα τέσσερα προβλήματα που θέλουμε να μελετήσουμε. Η μεθοδολογία που αναπτύσσεται είναι αυτή που θα βασιστούμε για τη δημιουργία του πληροφοριακού συστήματος. Θα παρουσιάσουμε αναλυτικά τις ενέργειες που απαιτούνται στην εκτέλεση κάθε βήματος. Ακόμη θα αναφερθούμε σε ορισμένες από τις υπερπαραμέτρους των μοντέλων που ενσωματώνουμε στο πληροφοριακό σύστημα και στον τρόπο που αυτές επηρεάζουν την επίδοση των μοντέλων που σχεδιάζουμε.

ΚΕΦΑΛΑΙΟ 4

ΜΕΘΟΔΟΛΟΓΙΑ

4.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο είδαμε αρκετούς αλγορίθμους μηχανικής μάθησης και κάποιες δημοφιλείς αρχιτεκτονικές νευρωνικών δικτύων. Παράλληλα εξετάσαμε και τις μετρικές με τις οποίες αξιολογούνται προβλήματα ταξινόμησης, παλινδρόμησης και πρόβλεψης. Σε αυτό το κεφάλαιο θα αναπτύξουμε τη μεθοδολογία βάσει της οποίας αναπτύχθηκε το πληροφοριακό σύστημα και θα παρουσιάσουμε τα βήματα που ακολουθήθηκαν.

Στην ενότητα 4.2 θα καθορίσουμε τα προβλήματα που πρόκειται να εξετάσουμε, σημειώνοντας αν πρόκειται για πρόβλημα ταξινόμησης, παλινδρόμησης ή πρόβλεψης καθώς και τον στόχο που θέλουμε να προβλέψουμε. Για τα προβλήματα αυτά θα δούμε τα δεδομένα που απαιτούνται να συλλεχθούν σε αντίστοιχα αρχεία, βάσει των οποίων θα γίνουν οι προβλέψεις.

Στην ενότητα 4.3 θα παρουσιάσουμε τον τρόπο που τα αρχικά δεδομένα θα μετατραπούν σε χαρακτηριστικά χρήσιμα για τα μοντέλα, αλλά θα αναφέρουμε και τη δημιουργία νέων χαρακτηριστικών μέσα από τα αρχικά χαρακτηριστικά που θα εξαχθούν.

Στην ενότητα 4.4 βλέπουμε τον τρόπο με τον οποίο προεπεξεργάζονται τα δεδομένα. Οι τιμές των χαρακτηριστικών χρειάζεται να κανονικοποιηθούν και να μετασχηματιστούν ώστε να καταστεί δυνατή η αποτελεσματική επεξεργασία τους από τα μοντέλα. Παράλληλα τονίζουμε την ανάγκη για απαλοιφή των ακραίων τιμών και παρουσιάζουμε τη μέθοδο που ακολουθούμε για να αφαιρέσουμε τις ακραίες τιμές από τα δεδομένα εισόδου.

Στην ενότητα 4.5 παρουσιάζουμε τα μοντέλα που δημιουργήθηκαν, λαμβάνοντας υπόψιν την κατηγορία των προβλημάτων που μελετάμε. Για κάθε κατηγορία γίνεται χρήση των αντίστοιχων αλγορίθμων.

Στην ενότητα 4..6 σημειώνουμε τη σημασία των υπερπαραμέτρων για την εκπαίδευση των μοντέλων και βλέπουμε τον τρόπο με τον οποίο θα χωριστούν τα δεδομένα που δίνονται ως είσοδος, σε δεδομένα εκπαίδευσης και δεδομένα δοκιμής.

Στην ενότητα 4.7 παρουσιάζουμε τον τρόπο με τον οποίο αξιολογούνται τα μοντέλα που σχεδιάσαμε. Στο κεφάλαιο 3 μελετήσαμε τις μετρικές απόδοσης και σφάλματος που πρόκειται να χρησιμοποιήσουμε ανάλογα με την κατηγορία που ανήκει κάθε πρόβλημα. Τα βήματα αυτά παρουσιάζονται στην ακόλουθη εικόνα ώστε να γίνει πιο κατανοητή η διαδικασία που ακολουθήθηκε για τα βήματα της πρόβλεψης,

ΒΗΜΑΤΑ ΔΙΑΔΙΚΑΣΙΑΣ ΠΡΟΒΛΕΨΗΣ

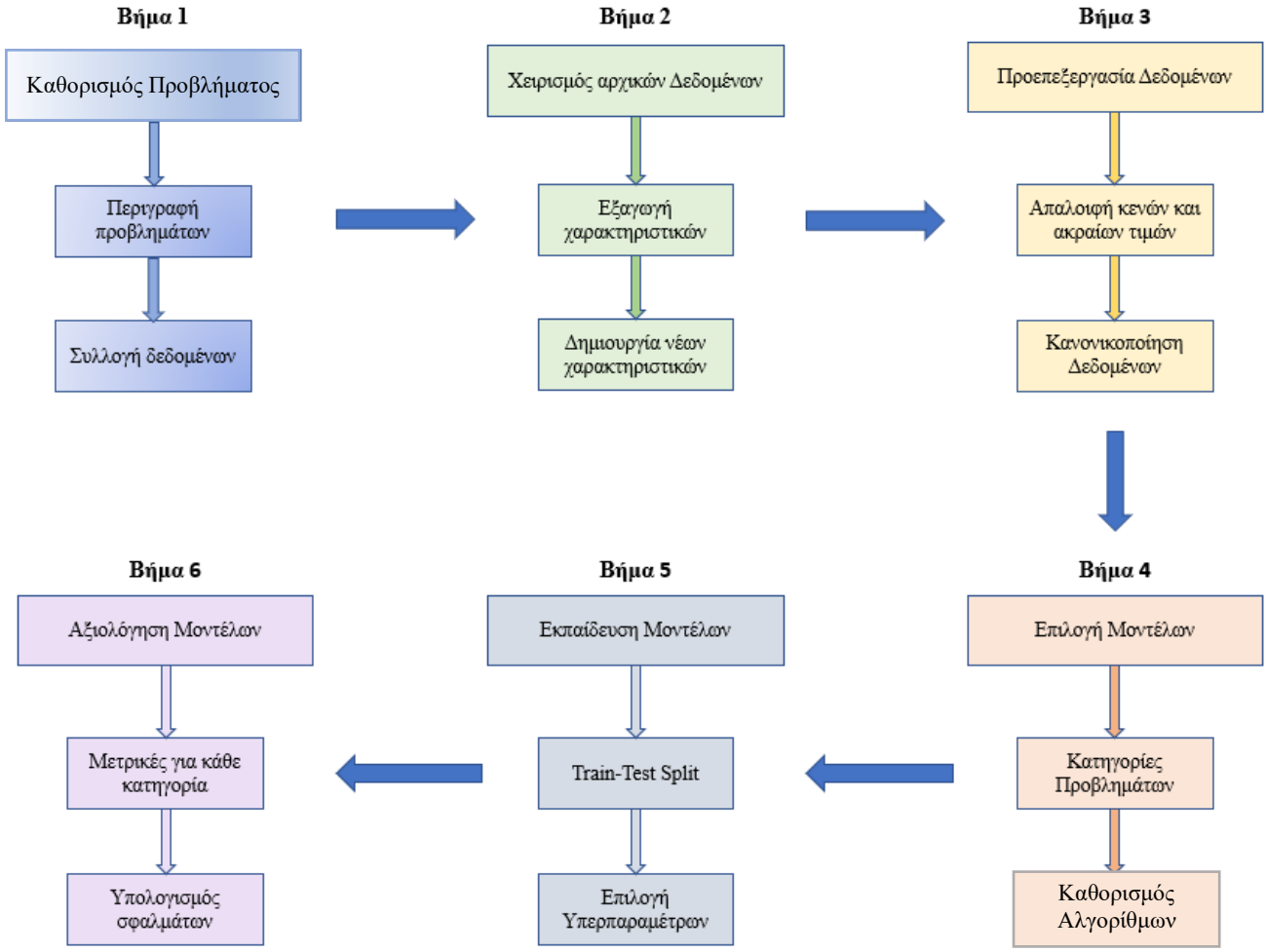

Figure 4.1 Prediction's Steps

Κάθε βήμα που απεικονίζεται στο παραπάνω σχήμα αναλύεται στις επόμενες ενότητες μαζί με τις επιμέρους ενέργειες που απαιτούνται. Με την ολοκλήρωση κάθε βήματος προχωράμε διαδοχικά στο επόμενο μέχρι την ολοκλήρωση της πρόβλεψης.

4.2 Καθορισμός Προβλήματος

Το πρώτο βήμα στη διαδικασία πρόβλεψης είναι αυτό του καθορισμού προβλήματος. Για κάθε πρόβλημα που θέλουμε να μελετήσουμε πρέπει να αποφασίσουμε ποιος είναι ο στόχος της πρόβλεψης και να συλλέξουμε τα ανάλογα δεδομένα. Η κατάταξη του προβλήματος σε κάποια κατηγορία είναι σημαντική διαδικασία για τα επόμενα βήματα του σχεδιασμού των μοντέλων και στις μετρικές αξιολόγησης ή σφάλματος. Τα μοντέλα και οι αλγόριθμοι που θα χρησιμοποιηθούν είναι διαφορετικοί για προβλήματα ταξινόμησης, παλινδρόμησης αλλά και για το σχεδιασμό νευρωνικών δικτύων με στόχο την πρόβλεψη κάποιας τιμής.

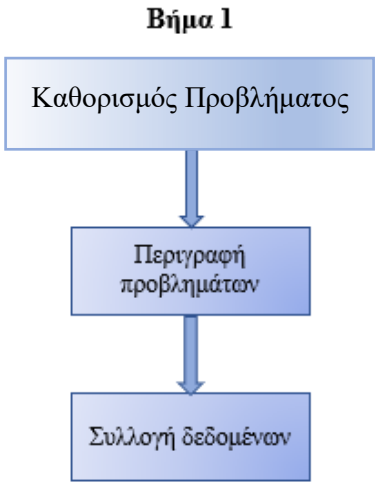

Figure 4.2 Step 1

Στο πληροφοριακό σύστημα ενσωματώσαμε τέσσερα προβλήματα τα οποία παρέχονται στο χρήστη προς μελέτη. Το πρώτο πρόβλημα που εξετάζουμε είναι η κατάταξη ενεργειακών έργων σε τρεις κλάσεις. Πρόκειται προφανώς για ένα πρόβλημα ταξινόμησης, συνεπώς θα χρησιμοποιηθούν και οι αντίστοιχοι αλγόριθμοι που αναφέραμε στο προηγούμενο κεφάλαιο . Στο δεύτερο πρόβλημα θέλουμε να προβλέψουμε την εξοικονόμηση ενέργειας που προκύπτει ύστερα από εργασίες ανακατασκευής σε τομείς όπως η βιομηχανία ή κτήρια. Η τιμή που θα προβλέψουμε είναι μία συνεχής τιμής, άρα πρόκειται για ένα πρόβλημα παλινδρόμησης. Το τρίτο πρόβλημα αποτελεί κι αυτό ένα πρόβλημα παλινδρόμησης, καθώς θέλουμε να προβλέψουμε την παραγωγή ενέργειας φωτοβολταϊκών. Στο τέταρτο πρόβλημα θέλουμε πάλι να προβλέψουμε την παραγωγή ενέργειας φωτοβολταϊκών, αλλά αυτήν τη φορά θέλουμε να χρησιμοποιήσουμε ιστορικά δεδομένα. Συνεπώς στην τελευταία αυτήν περίπτωση χρειάζεται να σχεδιάσουμε και να εκπαιδεύσουμε ορισμένα μοντέλα νευρωνικών δικτύων. Οι μετρικές που θα αξιολογηθούν οι επιδόσεις των μοντέλων θα είναι οι αντίστοιχες με την κατηγορία που μελετάμε. Για να ολοκληρωθεί η διαδικασία του πρώτου βήματος χρειάζεται να συλλέξουμε και τα κατάλληλα δεδομένα

που εξυπηρετούν το στόχο μας. Τα δεδομένα πάνω στα οποία θα εργαστούμε καθορίζουν τους αλγορίθμους που θα επιλέξουμε, καθώς η μορφή των δεδομένων, η ποσότητά τους και άλλοι παράγοντες επηρεάζουν την επίδοση των αλγορίθμων. Η επιλογή των χαρακτηριστικών θα προκύψει από τα αρχικά δεδομένα που θα συλλεχθούν και η προεπεξεργασία των δεδομένων θα πραγματοποιηθεί πάνω στις τιμές των χαρακτηριστικών που θα εξαχθούν.

4.3 Χειρισμός αρχικών Δεδομένων

Σε αυτήν την ενότητα θα παρουσιάσουμε τα αρχικά σύνολα δεδομένων που επεξεργαστήκαμε. Θα δούμε τα αρχικά πεδία που θέλουμε να περιέχουν τα csv αρχεία που θα ανεβάζει ο χρήστης στο πληροφοριακό σύστημα. Από αυτά τα αρχεία θα εξάγουμε τα χαρακτηριστικά που χρησιμοποιούνται για την εκπαίδευση των μοντέλων.

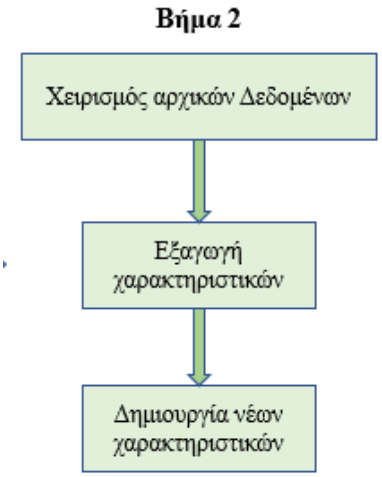

Figure 4.3 Step 2

Επίσης σημειώνουμε ποια πεδία δεν προσφέρουν στην εκπαίδευση των μοντέλων και κατά συνέπεια δεν τα εξάγουμε. Ακόμη αναφέρουμε τα καινούρια χαρακτηριστικά που δημιουργήθηκαν για την εκπαίδευση των μοντέλων του τέταρτου προβλήματος. Τα πεδία που θα εξάγουμε αποτελούν και τα διαθέσιμα χαρακτηριστικά που μπορούν να επιλεχθούν από το χρήστη στο πληροφοριακό σύστημα**.** Το βήμα αυτό είναι πολύ σημαντικό για τη διαδικασία της πρόβλεψης, καθώς αν δεν εξαχθούν σωστά τα χαρακτηριστικά από τα αρχικά δεδομένα τότε να μοντέλα δεν θα εκπαιδευτούν αποτελεσματικά. Στις παρακάτω υποενότητες παρουσιάζουμε τα πεδία που χρειάζεται να περιέχουν τα csv αρχεία για να προχωρήσουμε στην εξαγωγή χαρακτηριστικών.

4.3.1 ML-Based Renovation Classification

Όπως αναφέραμε στην προηγούμενη ενότητα, σε αυτό το πρόβλημα θέλουμε να ταξινομήσουμε ενεργειακά projects σε τρεις κλάσεις. Το αρχικό σύνολο δεδομένων είναι ένα αρχείο csv [29], το οποίο διαθέτει τις ακόλουθες στήλες που παρουσιάζονται στην επόμενη σελίδα:

- Project_ID
- Cost
- Energy_Consumption_Reduction (MWh)
- Building_Year
- Country_City
- Planned_CO2_Reduction
- Energy_Consumption_before (MWh)
- Total_Heating_Area (m^2)
- Floors
- Labels

Η κλάση που ανήκει κάθε έργο βρίσκεται στη στήλη Labels, επομένως αποτελεί την πρόβλεψη στόχο για αυτό το πρόβλημα. Πέρα από τις στήλες Country_City και Project_ID, οι υπόλοιπες χρησιμοποιούνται για να εξάγουμε τα χαρακτηριστικά του προβλήματος.

4.3.2 Estimating Energy Savings of EE refurbishments

Σε αυτό το πρόβλημα προβλέπουμε την εξοικονόμηση ενέργειας σε έργα ανακατασκευής. Το αρχείο csv με το αρχικό σύνολο δεδομένων [30] περιέχει τα ακόλουθα πεδία:

- Alias
- Sector
- InvestmentValue (€)
- NetAnnualSaving (€)
- Country
- Lifetime
- EnergySavings (kWh)
- Category

Η πρόβλεψη γίνεται με τις τιμές της στήλης EnergySavings. Εξάγουμε τα χαρακτηριστικά των υπόλοιπων στηλών για την εκπαίδευση των μοντέλων, πέρα από το πεδίο Alias το οποίο αποτελεί ένα αναγνωριστικό για κάθε ανακαίνιση και δεν συμβάλει στην εκπαίδευση των μοντέλων.

4.3.3 Mid-Term Weather-based PV production forecasting

Για την πρόβλεψη παραγωγής ενέργειας φωτοβολταϊκών διαθέτουμε το παρακάτω σύνολο δεδομένων [31] με τα ακόλουθα αρχικά πεδία:

- datetime
- Humidity
- Temperature
- cloudcover
- windspeed (km/h)
- Solar (W/m^2)
- Diffuse_Solar (W/m^2)
- Production (kWh)
- year
- month
- day
- timestamp

Στην περίπτωση αυτή εξάγουμε τα χαρακτηριστικά από όλα τα πεδία του συνόλου δεδομένων και θέλουμε να προβλέψουμε τις τιμές παραγωγής ενέργειας των φωτοβολταϊκών με βάση τη στήλη Production του συνόλου δεδομένων.

4.3.4 Short-Term Weather-based PV production forecasting

Σε αυτό το πρόβλημα χρησιμοποιούμε το ίδιο σύνολο δεδομένων με το παραπάνω πρόβλημα. Σε αυτήν την περίπτωση όμως θέλουμε να προβλέψουμε την τιμή παραγωγής των φωτοβολταϊκών με βάση ιστορικά δεδομένα [32]. Γι’ αυτό το πρόβλημα δημιουργούμε επιπλέον νέα χαρακτηριστικά από αυτά που εξήχθησαν. Αρχικά, από το χαρακτηριστικό timestamp λαμβάνουμε μόνο την ώρα και στην συνέχεια φτιάχνουμε δύο νέα χαρακτηριστικά, τα sin_hour και cos_hour, που περιέχουν έναν ημιτονικό και έναν συνημιτονικό μετασχηματισμό αντίστοιχα των τιμών αυτού του πεδίου.

4.4 Προεπεξεργασία Δεδομένων

Το επόμενο βήμα στη διαδικασία πρόβλεψης είναι αυτό της προεπεξεργασίας δεδομένων. Αφού επιλέξουμε τα χαρακτηριστικά με τα οποία θα εργαστούμε, χρειάζεται να επεξεργαστούμε τις τιμές τους ώστε να είναι στην κατάλληλη μορφή για να δοθούν ως είσοδοι στα μοντέλα μας. Παρακάτω παρουσιάζουμε τους μετασχηματισμούς που πραγματοποιήσαμε.

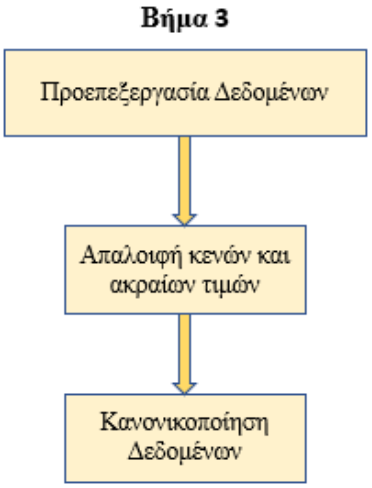

Figure 4.4 Step 3

Απαλοιφή κενών και ακραίων τιμών

Μία σημαντική παράμετρος στην προεπεξεργασία δεδομένων είναι ο χειρισμός των κενών τιμών. Αν και τα δεδομένα που διαθέτουμε δεν είχαν πεδία με κενές τιμές, είναι ένα αρκετά σύνηθες φαινόμενο που αντιμετωπίζουμε στην επιστήμη των δεδομένων. Μία απάντηση σε αυτό το πρόβλημα είναι η διαγραφή των σειρών που περιέχουν κάποια κενή τιμή σε μία στήλη. Στην περίπτωση όμως που υπάρχει κάποια στήλη που περιέχει μεγάλο αριθμό κενών τιμών θα χάναμε μεγάλο ποσοστό πληροφορίας. Σε μία τέτοια περίπτωση, θα μπορούσαμε να αφαιρέσουμε τελείως τη στήλη με τις πολλές κενές τιμές, χάνοντας ένα χαρακτηριστικό εκπαίδευσης διατηρώντας όμως περισσότερα δείγματα. Μία άλλη λύση είναι η απόδοση τιμών στα κενά πεδία, ιδιαίτερα σε περιπτώσεις που το ποσοστό κενών τιμών είναι χαμηλό. Μέθοδοι όπως η απόδοση τιμών με βάση τον υπολογισμό της μέσης τιμής ενδέχεται να οδηγήσουν σε μία λανθασμένη προτίμηση στα δεδομένα, γι' αυτό έχουν αναπτυχθεί μέθοδοι που λαμβάνουν κι άλλες παραμέτρους υπόψιν [33].

Η εξάλειψη των ακραίων τιμών είναι επίσης πολύ σημαντική διαδικασία στην προεπεξεργασία δεδομένων. Για να αποκλείσουμε τις ακραίες τιμές χρησιμοποιήσαμε την μέθοδο IQR (interquartile method). Η μέθοδος IQR περιγράφεται από τα ακόλουθα βήματα:

- Χωρίζουμε το σύνολο δεδομένων σε δύο ταξινομημένα μέρη με βάση την τιμή που θέλουμε να προβλέψουμε
- Για κάθε μέρος βρίσκουμε τα μεσαία δείγματα M1, M2
- Βρίσκουμε τη διαφορά Δ των δύο μεσαίων δειγμάτων
- Θέτουμε κάτω όριο, $lower\ bound = M1 - 1.5\Delta$ και άνω $upper\ bound = M2 + 1.5\Delta$
- Αφαιρούμε όσα δείγματα έχουν τιμές πέρα από το άνω ή κάτω όριο

Κανονικοποίηση Δεδομένων

Το επόμενο βήμα στην προεπεξεργασία δεδομένων είναι να κανονικοποιήσουμε τις τιμές των δεδομένων. Οι δύο μέθοδοι που χρησιμοποιούμε σε αυτήν την εργασία είναι η Min-Max Normalization και η Z-score Normalization. Η μέθοδος Min-Max Normalization μετασχηματίζει τα δεδομένα σε διαφορετική κλίμακα, στη συγκεκριμένη περίπτωση θέσαμε την ελάχιστη τιμή 0 και την μέγιστη 1. Ο τύπος για την αλλαγή τιμών παρουσιάζεται παρακάτω:

$$X_{new}[:,k] = \frac{X[:,k] - \min\left(X[:,k]\right)}{\max(X[:,k]) - \min\left(X[:,k]\right)}$$

Οι καινούριες τιμές για το χαρακτηριστικό k προκύπτουν αν αφαιρέσουμε από κάθε τιμή του χαρακτηριστικού την ελάχιστη τιμή και στη συνέχεια διαιρέσουμε με τη διαφορά μέγιστης και ελάχιστης τιμής. Για τη μέθοδο Z-score Normalization αφαιρούμε τη μέση τιμή από τις τιμές των δειγμάτων και διαιρούμε με την τυπική απόκλιση:

$$x_{new} = \frac{x - \bar{x}}{\sigma}$$

Μετατροπή Κατηγορικών Τιμών

Υπάρχουν πεδία στα δεδομένα μας που αντί για αριθμητικές τιμές περιέχουν χαρακτήρες. Αυτό θα μας δημιουργούσε πρόβλημα, καθώς στα μοντέλα που σχεδιάσαμε οι τιμές των πεδίων πρέπει να περιέχουν αριθμητικές τιμές για να δοθούν ως δεδομένα εισόδου. Για παράδειγμα, τα πεδία Sector, Country και Category περιέχουν χαρακτήρες. Το πεδίο Sector χαρακτηρίζει τον τομέα που γίνεται η ανακαίνιση για το δεύτερο πρόβλημα. Η μετατροπή θα γίνει αντιστοιχώντας μία ακέραια τιμή για κάθε κατηγορία τομέα. Αν για παράδειγμα το πεδίο Sector διέθετε δέκα διαφορετικές κατηγορίες, κάθε κατηγορία θα αντιστοιχούσε σε έναν ακέραιο αριθμό από το μηδέν έως το εννέα.

4.5 Επιλογή Μοντέλων

Το επόμενο βήμα που ακολουθείται στη διαδικασία πρόβλεψης είναι η επιλογή των μοντέλων που θα αναπτυχθούν και θα έχει στη διάθεσή του ο χρήστης του πληροφοριακού συστήματος να επιλέξει. Η ανάπτυξη κάθε μοντέλου σχετίζεται με την κατηγορία που ανήκει το πρόβλημα, όπως είπαμε διαφορετικοί αλγόριθμοι θα χρησιμοποιηθούν για προβλήματα ταξινόμησης, διαφορετικοί για προβλήματα παλινδρόμησης και θα αναπτυχθούν αντίστοιχες αρχιτεκτονικές νευρωνικών δικτύων για πρόβλεψη με βάση ιστορικά δεδομένα.

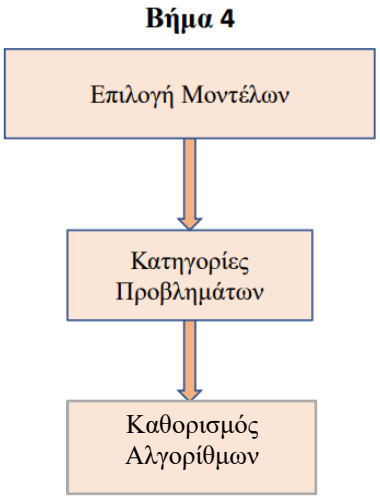

Figure 4.5 Step 4

Για το πρώτο πρόβλημα της ταξινόμησης των έργων σε ενεργειακές κλάσεις αναπτύξαμε τέσσερα μοντέλα που βασίζονται στους αλγορίθμους Random Forest, Decision Tree, KNN και στη μέθοδο Support Vector Machine. Για το πρόβλημα παλινδρόμησης της πρόβλεψης εξοικονόμησης ενέργειας ανακαινισμένων τομέων κατασκευάσαμε τρία μοντέλα που βασίζονται σε δύο μεθόδους Gradient Boosting των βιβλιοθηκών XGBoost και LightGBM, τις οποίες θα δούμε στο Κεφάλαιο 5, και στον αλγόριθμο Random Forest. Για το πρόβλημα της πρόβλεψης παραγωγής ενέργειας φωτοβολταϊκών αναπτύξαμε τέσσερα μοντέλα τα οποία βασίζονται το ένα στον αλγόριθμο Random Forest, δύο από αυτά σε μοντέλα γραμμικής παλινδρόμησης και ένα στη μέθοδο Support Vector Machine. Τέλος για την πρόβλεψη παραγωγής ενέργειας φωτοβολταϊκών βάσει ιστορικών δεδομένων κατασκευάστηκαν τέσσερα νευρωνικά δίκτυα, τα οποία χρησιμοποιούν LSTM, Bidirectional, και Convolutional layers.

4.6 Εκπαίδευση Μοντέλων

Το πέμπτο βήμα στη διαδικασία της πρόβλεψης είναι η εκπαίδευση των μοντέλων. Σε αυτό το βήμα χωρίζουμε τα δείγματα σε αυτά που θα χρησιμοποιήσουν τα μοντέλα για να εκπαιδευτούν και σε αυτά που θα δοκιμαστούν και προχωράμε στην επιλογή των διαθέσιμων υπερπαραμέτρων για κάθε μοντέλο. Παρακάτω παρουσιάζουμε τα δύο επιμέρους βήματα που απαιτούνται σε αυτό το στάδιο της διαδικασίας πρόβλεψης.

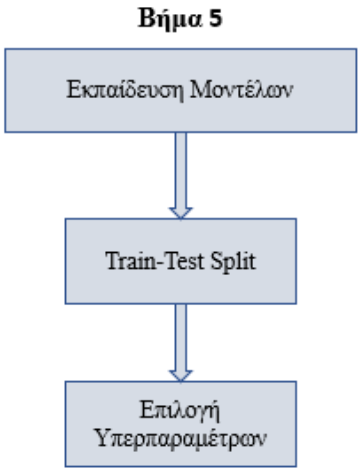

Figure 4.6 Step 5

Train-Test Split

Αφού επιλέξουμε τα μοντέλα που θα χρησιμοποιήσουμε ακολουθεί η εκπαίδευσή τους. Αρχικά θα χωρίσουμε το σύνολο δεδομένων σε δεδομένα εκπαίδευσης και δεδομένα δοκιμής. Μία συνήθης πρακτική είναι το 70%-80% των δεδομένων να χρησιμοποιείται για την εκπαίδευση των μοντέλων και το υπόλοιπο 20%-30% για την αξιολόγησή τους. Φυσικά ο χρήστης έχει τη δυνατότητα να διαμορφώσει τα ποσοστά αυτά όπως επιθυμεί. Ακόμη μπορούμε να αποφασίσουμε αν ο διαχωρισμός των δεδομένων θα γίνει με τυχαίο ή όχι τρόπο. Σε περιπτώσεις όπως η πρόβλεψη παραγωγής ενέργειας φωτοβολταϊκών, η εποχικότητα των δεδομένων παίζει σημαντικό ρόλο, συνεπώς τα δεδομένα δεν θα μπορούσαν να χωριστούν με τυχαίο τρόπο. Στα άλλα δύο προβλήματα χωρίζουμε τυχαία τα δεδομένα βοηθώντας έτσι στην αποφυγή του overfittng.

Επιλογή υπερπαραμέτρων

Η επιλογή των υπερπαραμέτρων καθορίζει σε μεγάλο βαθμό την απόδοση των μοντέλων. Στο πληροφοριακό σύστημα δίνουμε την επιλογή στο χρήστη εφόσον το επιθυμεί να επιλέξει ο ίδιος κάποιες από τις υπερπαραμέτρους εκπαίδευσης. Αυτό δεν είναι βέβαια απαραίτητο καθώς μπορεί να προχωρήσει και στο επόμενο βήμα της αξιολόγησης διατηρώντας τις προκαθορισμένες τιμές που έχουμε θέσει. Για το πρώτο πρόβλημα οι διαθέσιμες υπερπαράμετροι για τον αλγόριθμο Random Forest είναι οι n_estimators, criterion, max_depth και random state. Αυτές προσδιορίζουν αντίστοιχα τον αριθμό των ταξινομητών που θα κατασκευαστούν, το κριτήριο αξιολόγησης του χωρισμού των κόμβων, το βάθος που θα έχουν οι ταξινομητές-δέντρα που θα κατασκευαστούν καθώς και μία τυχαία κατάσταση. Για τη μέθοδο SVM οι διαθέσιμες υπερπαράμετροι είναι η C, όπου προσδιορίζει το επιθυμητό περιθώριο που θα έχουν τα υπερεπίπεδα, η υπερπαράμετρος kernel, όπου δηλώνει τη μέθοδο που θα χρησιμοποιηθεί για τον μετασχηματισμό των δεδομένων εισόδου σε μορφή όπου θα είναι δυνατόν να κατασκευαστούν υπερεπίπεδα για τον διαχωρισμό τους, και η υπερπαράμετρος gamma , όπου υποδεικνύει το εύρος της ακτίνας επιρροής κάθε μεμονωμένου δείγματος. Για τον αλγόριθμο Decision Tree έχουμε επίσης τις υπερπαραμέτρους criterion και max depth όπως επίσης και την υπερπαράμετρο splitter που καθορίζει τη στρατηγική με την οποία θα χωριστούν οι κόμβοι του δέντρου. Για το τέταρτο μοντέλο χρησιμοποιήθηκε ο αλγόριθμος KNN με τις υπερπαραμέτρους που μπορούν να αποδοθούν τιμές να είναι οι neighbors, Weights και Algorithm. Αυτές προσδιορίζουν αντίστοιχα τον αριθμό των K κοντινότερων δειγμάτων, το βάρος που έχει κάθε γείτονας ανάλογα με την απόσταση από το εξεταζόμενο δείγμα και τον αλγόριθμο που υπολογίζεται η απόσταση από τους κοντινότερους γείτονες.

Για το δεύτερο πρόβλημα ο αλγόριθμος Random Forest χρησιμοποιεί τις ίδιες υπερπαραμάτρους που περιγράφηκαν παραπάνω ενώ οι δύο μέθοδοι Gradient Boosting που αναπτύχθηκαν διαθέτουν στο χρήστη την επιλογή του βαθμού εκμάθησης learning rate και το βάθος των δέντρων που θα κατασκευαστούν max depth. Για το τρίτο πρόβλημα οι υπερπαράμετροι για τον αλγόριθμο Random Forest και για τη μέθοδο SVM είναι ίδιες με αυτές που περιγράφηκαν παραπάνω, ενώ για το μοντέλο linear regression παρέχουμε την υπερπαράμετρο n_jobs που καθορίζει τον αριθμό των νημάτων ή διαδικασιών που δημιουργούνται και για το μοντέλο least-angle regression παρέχουμε τη σταθερά alpha και τον μέγιστο αριθμό επαναλήψεων max iter.

Τέλος, για το τέταρτο πρόβλημα και τα νευρωνικά δίκτυα ο χρήστης μπορεί να επιλέξει για κάθε ένα από τα μοντέλα ξεχωριστά το βήμα εκμάθησης learning rate, τη μετρική σφάλματος loss που θέλουμε να ελαχιστοποιήσουμε κατά τη διαδικασία εκπαίδευσης, τη συνάρτηση ενεργοποίησης activation που θα χρησιμοποιηθεί καθώς και τον αριθμό των επαναλήψεων που τα νευρωνικά δίκτυα θα εκπαιδευτούν σε ολόκληρο το σύνολο δεδομένων, epochs.

4.7 Αξιολόγηση Μοντέλων

Το τελευταίο βήμα στη διαδικασία πρόβλεψης είναι η αξιολόγηση των μοντέλων για τις προβλέψεις που πραγματοποιήθηκαν. Γι' αυτόν το σκοπό χρησιμοποιούμε τις κατάλληλες μετρικές απόδοσης και σφάλματος. Το βήμα 6 αποτελεί και το τελευταίο βήμα στην διαδικασία της πρόβλεψης που αναπτύξαμε.

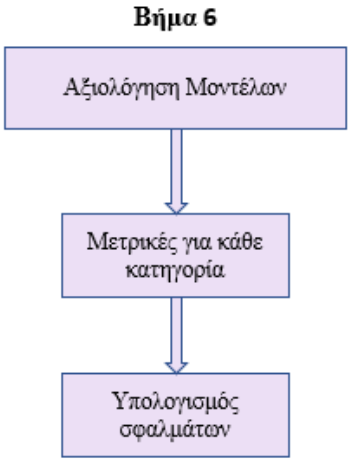

Figure 4.7 Step 6

Για το πρόβλημα ταξινόμησης οι μετρικές που αξιολογούνται τα μοντέλα είναι οι accuracy score, precision, recall και F1score. Για τα υπόλοιπα τρία προβλήματα χρησιμοποιείται η μετρική απόδοσης R2 καθώς και η μετρικές σφάλματος Mean Absolute Error (MAE), Mean Squared Error (MSE) και Root Mean Squared Error (RMSE). Επίσης, πέρα από το πρώτο πρόβλημα, αναπαριστάμε γραφικά τις τιμές που προβλέφθηκαν με τις πραγματικές τιμές εξόδου των δειγμάτων. Όλες οι μετρικές απόδοσης και σφάλματος έχουν παρουσιαστεί στο κεφάλαιο 3.

Το βήμα της αξιολόγησης των μοντέλων παίζει ιδιαίτερα σημαντικό ρόλο στο πληροφοριακό σύστημα. Οι μετρικές απόδοσης και σφάλματος μας δίνουν τη δυνατότητα να δούμε πόσο αποτελεσματικά λειτουργούν τα μοντέλα αλλά και να συγκρίνουμε τις μεταξύ τους επιδόσεις. Αν και το βήμα 6 αποτελεί το τελευταίο βήμα στη διαδικασία πρόβλεψης, ο χρήστης θα μπορούσε να γυρίσει πάλι πίσω στο βήμα 5 να επιλέξει διαφορετικές τιμές για τις υπερπαραμέτρους και να διαπιστώσει πως αυτές επηρεάζουν τα αποτελέσματα των προβλέψεων.

4.8 Επίλογος

Σε αυτό το κεφάλαιο είδαμε αναλυτικά τη μεθοδολογία που αναπτύχθηκε και τα βήματα διαδικασίας πρόβλεψης που ακολουθούνται. Για κάθε βήμα είδαμε τις αντίστοιχες ενέργειες που εκτελούνται, καθώς και ποιες δυνατότητες δίνονται στο χρήστη για την εκπαίδευση των μοντέλων. Η διαδικασία πρόβλεψης ξεκινά με τον καθορισμό των προβλημάτων που μελετάμε και τη συλλογή των δεδομένων που απαιτούνται για το κάθε πρόβλημα. Η διαδικασία συνεχίζεται με τα χαρακτηριστικά που εξάγονται από αυτά τα δεδομένα αλλά και τα νέα χαρακτηριστικά που δημιουργούμε. Έπειτα σημειώνουμε τα επιμέρους βήματα που χρειάζονται για την προεπεξεργασία των δεδομένων, όπως η εξάλειψη των ακραίων τιμών και η κανονικοποίηση των τιμών των χαρακτηριστικών. Στο επόμενο βήμα παρουσιάζομε τα διαθέσιμα μοντέλα για κάθε πρόβλημα και αναφέρουμε τους αλγορίθμους που χρησιμοποιούμε. Στη συνέχεια ακολουθεί ο διαχωρισμός του συνόλου δεδομένων, η επιλογή των υπερπαραμέτρων και η εκπαίδευση των μοντέλων. Τέλος, αναφέρουμε τις μετρικές που χρησιμοποιούνται για κάθε πρόβλημα και μας δείχνουν την επίδοση που είχαν τα μοντέλα. Αυτό αποτελεί και το τελευταίο βήμα στη διαδικασία της πρόβλεψης.

Στο επόμενο κεφάλαιο θα αναφέρουμε αρχικά τα εργαλεία και τις βιβλιοθήκες που χρησιμοποιήθηκαν για την ανάπτυξη του πληροφοριακού συστήματος. Στη συνέχεια θα παραθέσουμε τα UML διαγράμματα που σχεδιάστηκαν για όλες τις ενέργειες που μπορεί να εκτελέσει ο χρήστης. Τέλος θα δούμε στιγμιότυπα από την εφαρμογή για τις σελίδες που δημιουργήθηκαν και θα παρουσιάσουμε ορισμένα αποτελέσματα που προκύπτουν ακολουθώντας τα βήματα της διαδικασίας που περιγράφηκαν σε αυτήν την ενότητα. Θα συγκρίνουμε τις επιδόσεις που έχουν τα μοντέλα για κάποια προβλήματα και θα δούμε τα διαφορετικά αποτελέσματα που προκύπτουν για την επιλογή διαφορετικών τιμών των υπερπαραμέτρων.

ΚΕΦΑΛΑΙΟ 5

ΠΛΗΡΟΦΟΡΙΑΚΟ ΣΥΣΤΗΜΑ

5.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο αναλύσαμε τη μεθοδολογία και τα βήματα για τη διαδικασία πρόβλεψης που ακολουθούμε, καθώς και τις ενέργειες που θέλουμε να πραγματοποιούνται στην εφαρμογή μας, τις επιλογές που παρέχονται στο χρήστη σχετικά με την εκπαίδευση των μοντέλων και την αξιολόγησή τους. Για κάθε βήμα είδαμε τις επιλογές που θέλουμε να έχει στη διάθεσή του ο χρήστης για να προεπεξεργάζεται όπως επιθυμεί εκείνος τα δεδομένα και να διαλέγει τα μοντέλα που θέλει να χρησιμοποιήσει στις προβλέψεις του.

Σε αυτό το κεφάλαιο περιγράφουμε το πληροφοριακό σύστημα που αναπτύχθηκε με βάση την παραπάνω μεθοδολογία. Αρχικά θα παρουσιάσουμε στην ενότητα 5.2 τα εργαλεία και τις βιβλιοθήκες που χρησιμοποιήθηκαν για την ανάπτυξη του συστήματος . Στη συνέχεια θα παραθέσουμε στην ενότητα 5.3 τα διαγράμματα UML που δημιουργήθηκαν για τις ανάγκες σχεδιασμού του πληροφοριακού συστήματος. Στην ενότητα 5.4 θα παρουσιάσουμε τις σελίδες που δημιουργήσαμε ενσωματώνοντας τη μεθοδολογία που αναπτύχθηκε στο προηγούμενο κεφάλαιο. Θα απεικονίσουμε τα στιγμιότυπα για την αρχική σελίδα καθώς και τις σελίδες για το ανέβασμα των αρχείων, την επιλογή των μοντέλων, την προεπεξεργσία των δεδομένων, την επιλογή των υπερπαραμέτρων και την παρουσίαση των αποτελεσμάτων. Ακόμη θα δούμε κάποια στιγμιότυπα που αφορούν στα αποτελέσματα ίδιων μοντέλων για διαφορετικές υπερπαραμέτρους, αλλά και τα διαγράμματα που δημιουργούνται αναφορικά με τα αποτελέσματα των μετρικών αξιολόγησης.

5.2 Εργαλεία και βιβλιοθήκες που χρησιμοποιήθηκαν

Σε αυτήν την ενότητα θα παρουσιάσουμε τα εργαλεία και τις βιβλιοθήκες που χρειάστηκαν για την ανάπτυξη του πληροφοριακού συστήματος. Παρουσιάζουμε τις γλώσσες, τις βιβλιοθήκες αλλά και το web framework που χρησιμοποιήθηκε.

Python

Η γλώσσα προγραμματισμού Python αποτελεί μία από τις πιο δημοφιλείς γλώσσες προγραμματισμού σήμερα. Είναι μία γλώσσα υψηλού επιπέδου, γενικού σκοπού η οποία προσφέρει σημαντικά πλεονεκτήματα στους χρήστες, όπως οι βιβλιοθήκες που διαθέτει και η ευκολία στην κατανόηση και τη χρήση. Χρησιμοποιείται σε πολλές εφαρμογές τεχνητής νοημοσύνης και μηχανικής

μάθησης, καθώς και σε άλλες επιστημονικές εφαρμογές που απαιτούνται πολύπλοκοι μαθηματικοί υπολογισμοί.

Javascript

Η γλώσσα προγραμματισμού Javascript είναι μία γλώσσα προγραμματισμού που συγκαταλέγεται ανάμεσα στις κύριες τεχνολογίες για την ανάπτυξη διαδικτυακών εφαρμογών μαζί με τις HTML και CSS. Οι πιο δημοφιλείς φυλλομετρητές διαθέτουν μία μηχανή Javascript ώστε να εκτελείται ο κώδικας στις συσκευές των χρηστών.

HTML και CSS

Η γλώσσα HTML (HyperText Markup Language) είναι η κύρια γλώσσα σήμανσης για τις ιστοσελίδες. Τα στοιχεία της HTML αποτελούνται από ετικέτες, συνήθως ανά ζεύγη με την πρώτη ετικέτα να είναι η ετικέτα έναρξης και η δεύτερη η ετικέτα λήξης. Ένας φυλλομετρητής (web browser) διαβάζει HTML έγγραφα και τα συνθέτει σε μορφή περιεχομένου ιστοσελίδων. Η γλώσσα CSS (Cascading Style Sheets) χρησιμοποιείται για την εμφάνιση, τη διάταξη και το στυλ μορφοποίησης ενός εγγράφου που έχει γραφτεί σε μία γλώσσα σήμανσης όπως η HTML.

Pandas

Η βιβλιοθήκη Pandas [34] είναι μία βιβλιοθήκη της Python που χρησιμοποιείται για την ανάλυση και διαχείριση δεδομένων. Προσφέρει αρκετές δομές δεδομένων, όπως τη δομή DataFrame, και μεθόδους για το χειρισμό τους. Η ευκολία και η ταχύτητα που παρέχει η βιβλιοθήκη pandas την καθιστά ένα από τα σημαντικότερα εργαλεία στον τομέα της ανάλυσης δεδομένων. Στην παρούσα εργασία χρησιμοποιούμε τη βιβλιοθήκη για να διαβάσουμε τα δεδομένα που χρειαζόμαστε από τα csv αρχεία, αλλά και για διάφορες μετατροπές που πραγματοποιούμε στα δεδομένα μας.

Numpy

Η βιβλιοθήκη Numpy είναι άλλη μία πολύ χρήσιμη βιβλιοθήκη της Python που προσφέρει υψηλή επίδοση στην εκτέλεση αριθμητικών υπολογισμών [35] και χρησιμοποιείται ιδιαίτερα για πράξεις μεταξύ πινάκων. Η δομή Numpy array, όπου πρόκειται για έναν ν-διάστατο πίνακα, προσφέρει σημαντικά πλεονεκτήματα στους χρήστες σε σχέση με τις λίστες της Python, όπως ταχύτητα, ευκολία στη χρήση, συναρτήσεις για τον χειρισμό των Numpy arrays και καταναλώνει επίσης λιγότερη μνήμη συγκριτικά με τη χρήση των λιστών της Python.

Scikit-learn

Η βιβλιοθήκη scikit-learn παρέχει υλοποιήσεις αλγορίθμων μηχανικής μάθησης για τη γλώσσα Python, όπως οι αλγόριθμοι knn, svm, random forest, decision tree, linear regression και άλλοι που χρησιμοποιούμε στο πληροφοριακό μας σύστημα, καθώς και υλοποιήσεις για τις μετρικές που χρησιμοποιούμε και αναφέραμε στα προηγούμενα κεφάλαια. Ακόμη οι συναρτήσεις για την κανονικοποίηση των δεδομένων έχουν ληφθεί από αυτήν τη βιβλιοθήκη.

Tensorflow

Η βιβλιοθήκη Tensorflow [36] χρησιμοποιείται για προβλήματα μηχανικής μάθησης, τεχνητής νοημοσύνης και κυρίως για το σχεδιασμό και εκπαίδευση σε βαθιά νευρωνικά δίκτυα. Αναπτύχθηκε από την ομάδα Google Brains και μπορεί να ενσωματωθεί από πολλές γλώσσες προγραμματισμού όπως Python, C++, Javascript και Java. Στο πληροφοριακό μας σύστημα τα νευρωνικά δίκτυα που σχεδιάσαμε για το τέταρτο πρόβλημα έχουν υλοποιηθεί με αυτήν τη βιβλιοθήκη.

XGBoost και LightGBM

Οι βιβλιοθήκες XGBoost και LightGBM είναι δύο βιβλιοθήκες που προσφέρουν υλοποιήσεις πάνω σε gradient boosting μεθόδους για προβλήματα μηχανικής μάθησης. Η βιβλιοθήκη XGBoost συμβάλει στη λύση πολλών προβλημάτων πάνω στην επιστήμη δεδομένων με υψηλή αποτελεσματικότητα και υποστηρίζει την εκτέλεση του κώδικα σε κατανεμημένα συστήματα, ενώ η βιβλιοθήκη LightGBM προσφέρει σημαντικά πλεονεκτήματα στην ταχύτητα εκπαίδευσης, στην κατανάλωση μνήμης, στον χειρισμό δεδομένων μεγάλης κλίμακας και υποστηρίζει και αυτή την εκτέλεση κώδικα σε κατανεμημένα συστήματα.

Django

Το εργαλείο Django είναι ένα εύχρηστο web framework για την ανάπτυξη ιστοσελίδων χρησιμοποιώντας τη γλώσσα Python. Βασίζεται στην αρχιτεκτονική MVC. Διαθέτει μία σχεσιακή βάση δεδομένων που ονομάζεται Model, ένα σύστημα για την επεξεργασία HTTP αιτημάτων με αντίστοιχες μεθόδους που ονομάζεται View και έναν αποστολέα URL που ονομάζεται Controller. Πολλές γνωστές εφαρμογές έχουν αναπτυχθεί με το Django web Framework καθώς προσφέρει σημαντικά πλεονεκτήματα όπως η απλότητα και η ταχύτητα με την οποία μπορεί να αναπτυχθεί μία εφαρμογή.

5.3 Διαγράμματα UML

Σε αυτήν την ενότητα παρουσιάζουμε μέσα από δύο διαγράμματα UML, ένα use case diagram και ένα sequence diagram, την αλληλεπίδραση του χρήστη με το πληροφοριακό σύστημα και τις ενέργειες που μπορεί να εκτελέσει. Τα διαγράμματα UML μας βοηθούν ιδιαίτερα στον τρόπο με τον οποίο πρόκειται να σχεδιάσουμε ένα σύστημα.

5.3.1 Use case diagram

Το πρώτο διάγραμμα που σχεδιάσαμε για τις ανάγκες της εφαρμογή είναι ένα use case διάγραμμα. Τα διαγράμματα use case χρησιμοποιούνται για τη γραφική αναπαράσταση των πιθανών αλληλεπιδράσεων του χρήστη με το σύστημα. Μας βοηθούν σημαντικά να καταλάβουμε με απλό τρόπο τη λειτουργία του συστήματος που αναπτύσσουμε. Όπως βλέπουμε, οι διαθέσιμες ενέργειες που μπορεί να εκτελέσει ο χρήσης είναι η επιλογή του προβλήματος που θέλει να μελετήσει, το ανέβασμα του αντίστοιχου αρχείου, επιλογές σχετικά με την προεπεξεργασία των δεδομένων, επιλογή των μοντέλων και των υπερπαραμέτρων τους.

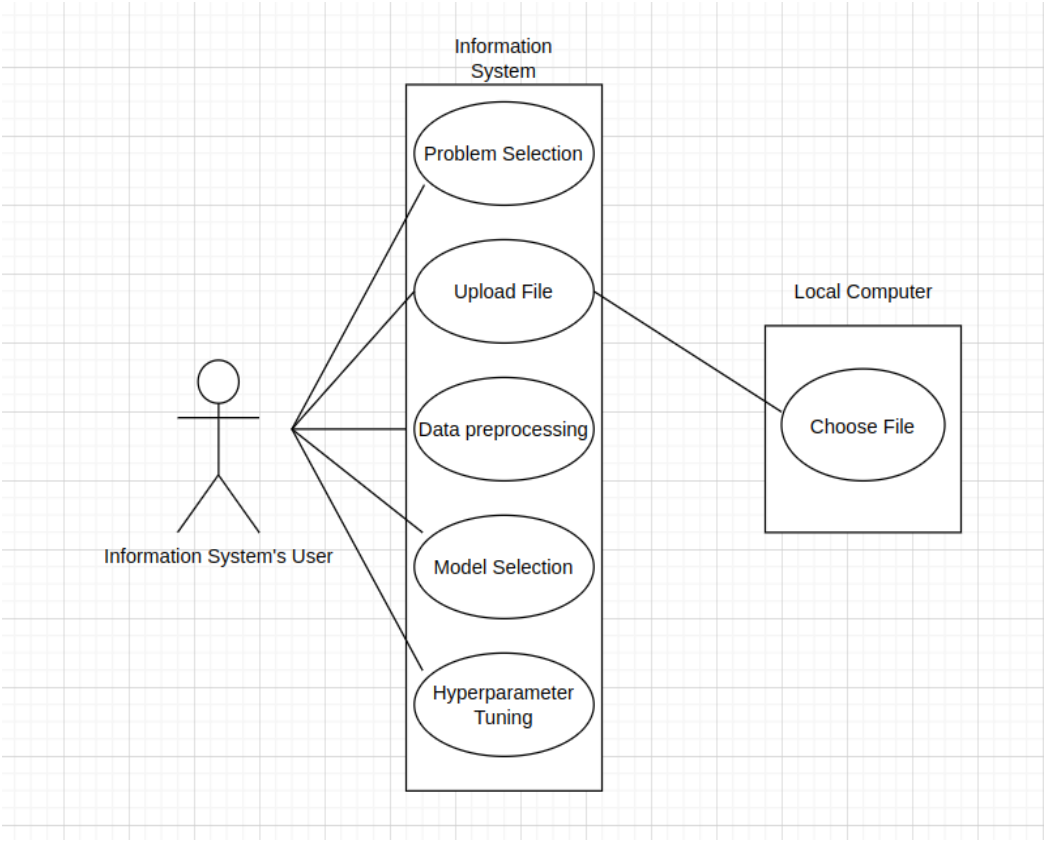

Figure 5.1 Use case

5.3.2 Sequence diagram

Το Sequence διάγραμμα μας δείχνει την ακολουθιακή εκτέλεση των λειτουργιών ανάμεσα στα συνεργαζόμενα πεδία. Όπως βλέπουμε, ο χρήστης αρχικά διαλέγει το πρόβλημα το οποίο θέλει να εξετάσει. Αυτό θα έχει ως αποτέλεσμα να μεταβεί στην αντίστοιχη σελίδα με το πρότυπο csv που πρέπει να δώσει ως είσοδο. Όταν πατήσει το εικονίδιο Choose File του δίνεται η δυνατότητα να ανεβάσει το αρχείο που επιθυμεί από τον τοπικό του υπολογιστή. Αυτό το αρχείο φορτώνεται στο πληροφοριακό σύστημα και εμφανίζεται ένα μήνυμα ότι το αρχείο φορτώθηκε επιτυχώς. Η επόμενη ενέργεια που ακολουθεί είναι η επιλογή για την προεπεξεργασία των δεδομένων και η επιλογή των μοντέλων που θα χρησιμοποιηθούν. Ανάλογα με την επιλογή των μοντέλων εμφανίζονται οι διαθέσιμες υπερπαράμετροι για ορισμό των τιμών τους και με βάση τις τιμές που θα θέσει ο χρήστης θα πραγματοποιηθούν οι ανάλογες προβλέψεις και θα εμφανιστούν τα αποτελέσματα που προκύπτουν.

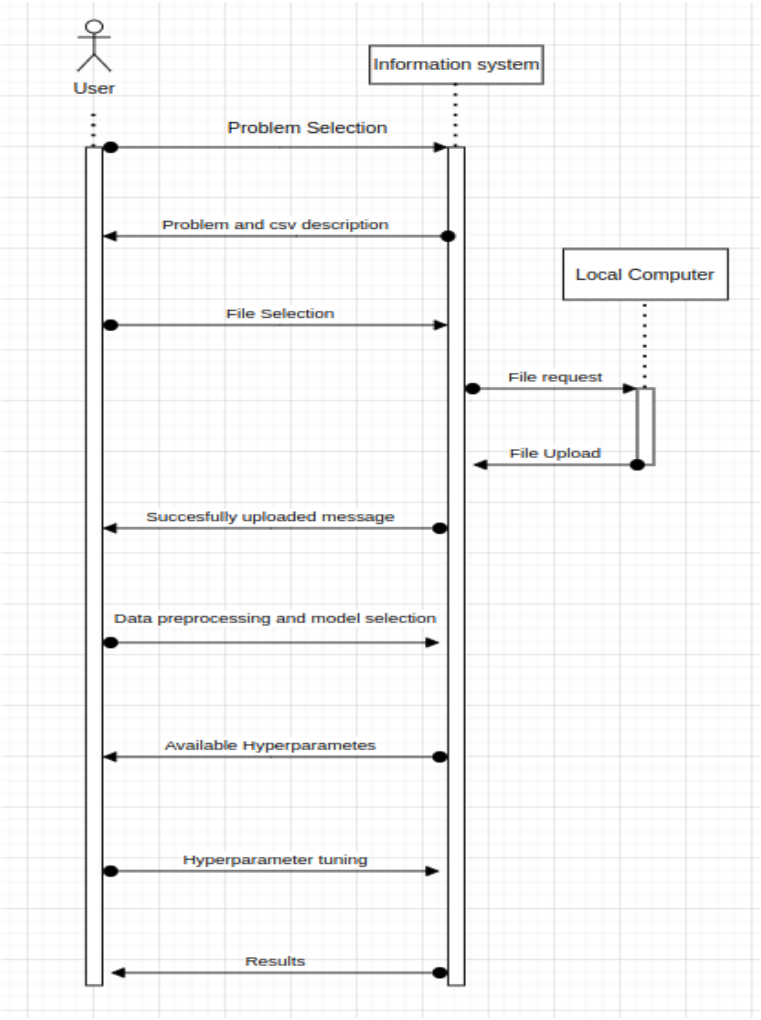

Figure 5.2 Sequence

5.4 Παρουσίαση πληροφοριακού συστήματος

Σε αυτήν την ενότητα παρουσιάζουμε το user interface, τις λειτουργικότητες του πληροφοριακού συστήματος και τα στιγμιότυπα των σελίδων που μπορεί να περιηγηθεί ο χρήστης. Όπως έχουμε αναφέρει, στόχος της εργασίας είναι η δημιουργία ενός περιβάλλοντος όπου ο χρήστης θα μπορεί να επεξεργαστεί και να δοκιμάσει με ευκολία τα τέσσερα προβλήματα που παρουσιάσαμε στα προηγούμενα κεφάλαια. Η αρχική σελίδα εμφανίζει στο χρήστη τα τέσσερα διαθέσιμα προβλήματα μαζί με τον τίτλο του κάθε προβλήματος, αλλά και το εικονίδιο με το μενού πάνω αριστερά στη σελίδα όπως φαίνεται στο ακόλουθο στιγμιότυπο.

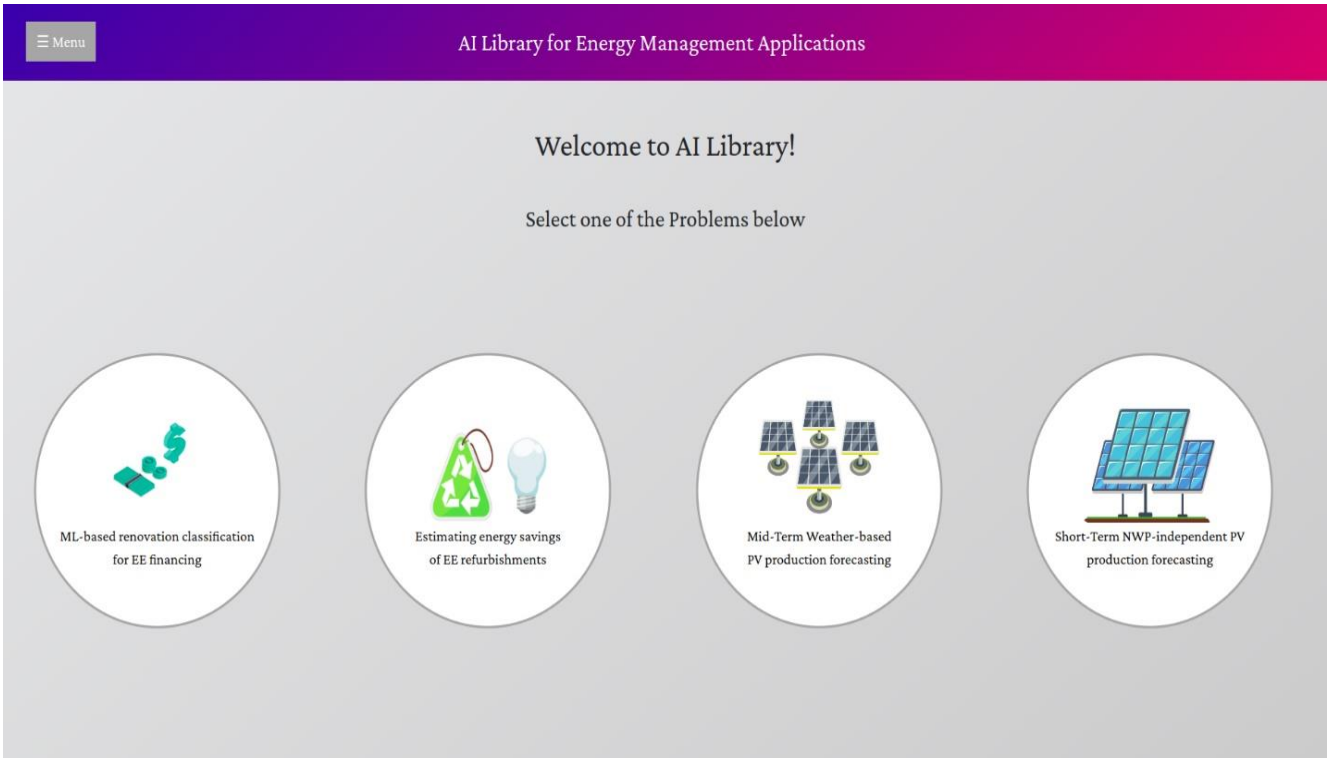

Figure 5.3 Home page

Από το μενού ο χρήστης μπορεί να επιλέξει τη σελίδα About, όπου ακολουθεί μία σύντομη περιγραφή για το κάθε πρόβλημα. Μία πιο εκτενής περιγραφή του συνόλου δεδομένου που απαιτείται ακολουθεί στη σελίδα upload του κάθε προβλήματος. Παρακάτω απεικονίζουμε τα στιγμιότυπα για τη σελίδα About που λήφθηκαν. Η πρώτη εικόνα δείχνει την περιγραφή για το πρόβλημα της ταξινόμησης ενεργειακών έργων σε κλάσεις.

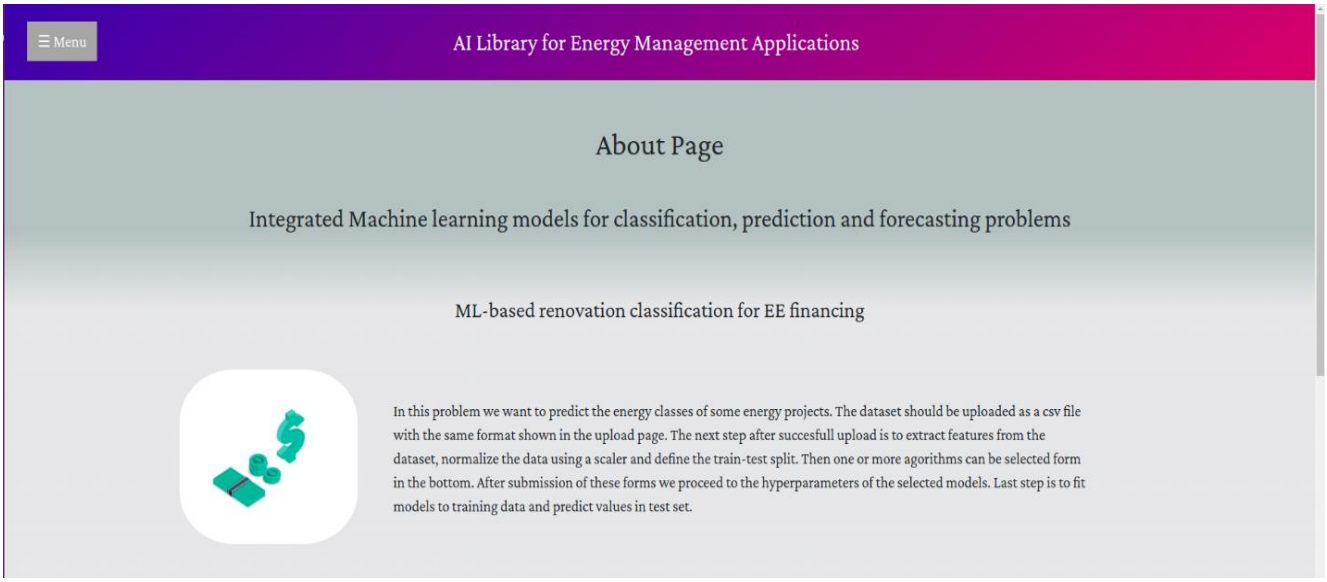

Figure 5.4 About 1

Στα επόμενα δύο στιγμιότυπα φαίνεται η συνέχεια της σελίδας About με τα άλλα τρία προβλήματα που είναι διαθέσιμα στην αρχική σελίδα.

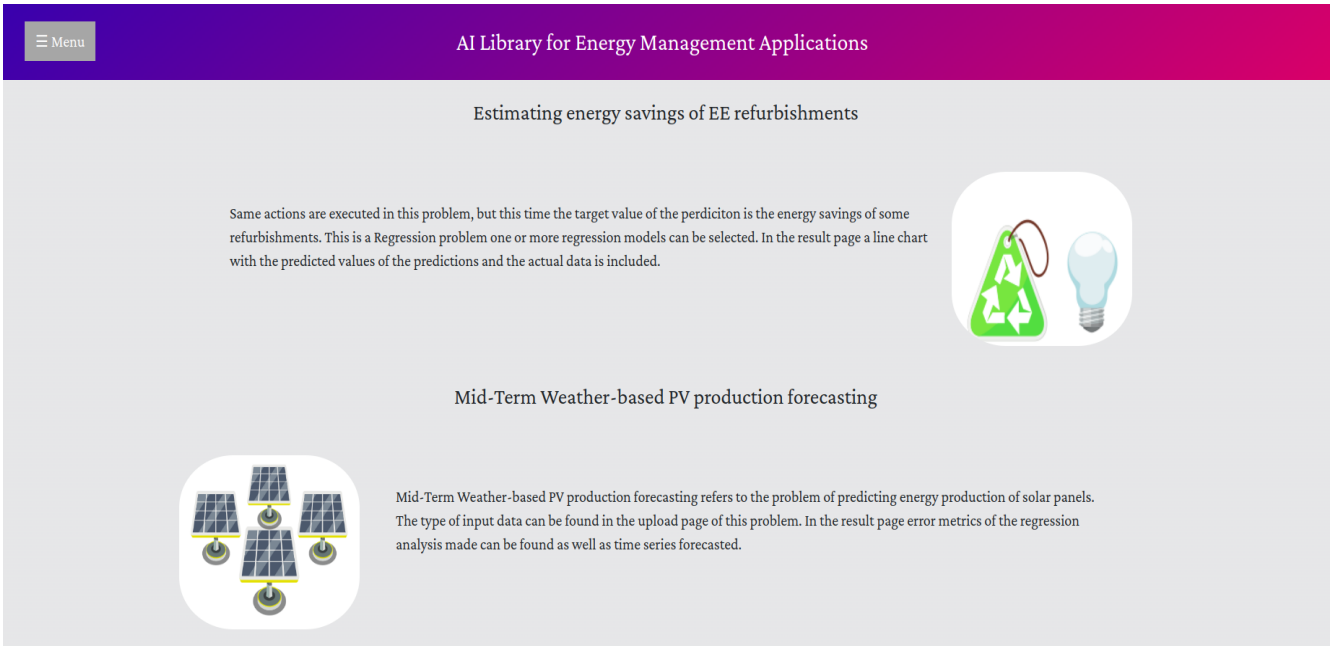

Figure 5.5 About 2

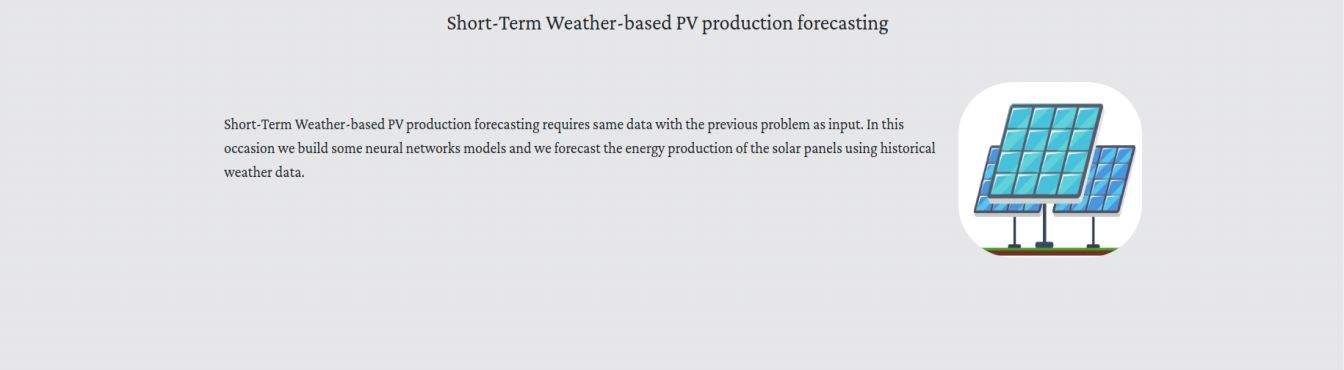

Figure 5.6 About 3

Η επιλογή κάθε προβλήματος οδηγεί στην αντίστοιχη σελίδα όπου περιγράφεται ο στόχος της πρόβλεψης με βάση τα χαρακτηριστικά του συνόλου δεδομένων. Παράλληλα εμφανίζεται ένα πρότυπο csv με τα ονόματα των στηλών και τον τύπο των δεδομένων, το οποίο πρέπει να ακολουθήσει ο χρήστης ώστε να είναι εφικτή η σωστή προεπεξεργασία των δεδομένων. Κάτω από το πρότυπο csv εμφανίζεται ένα κουμπί choose file για την επιλογή του αρχείου που θέλει να ανεβάσει ο χρήστης. Αφού επιλέξει ο χρήστης το αρχείο πατώντας στο κουμπί upload τα δεδομένα περνάνε στο πληροφοριακό σύστημα μέσα από τη μέθοδο POST του αντικειμένου HTTP. Το Django χρησιμοποιεί δύο μεθόδους για τη διεπαφή του χρήστη με το σύστημα μέσα από φόρμες, την GET και την POST. Παρακάτω παρουσιάζουμε την αντίστοιχη σελίδα που περιγράψαμε.

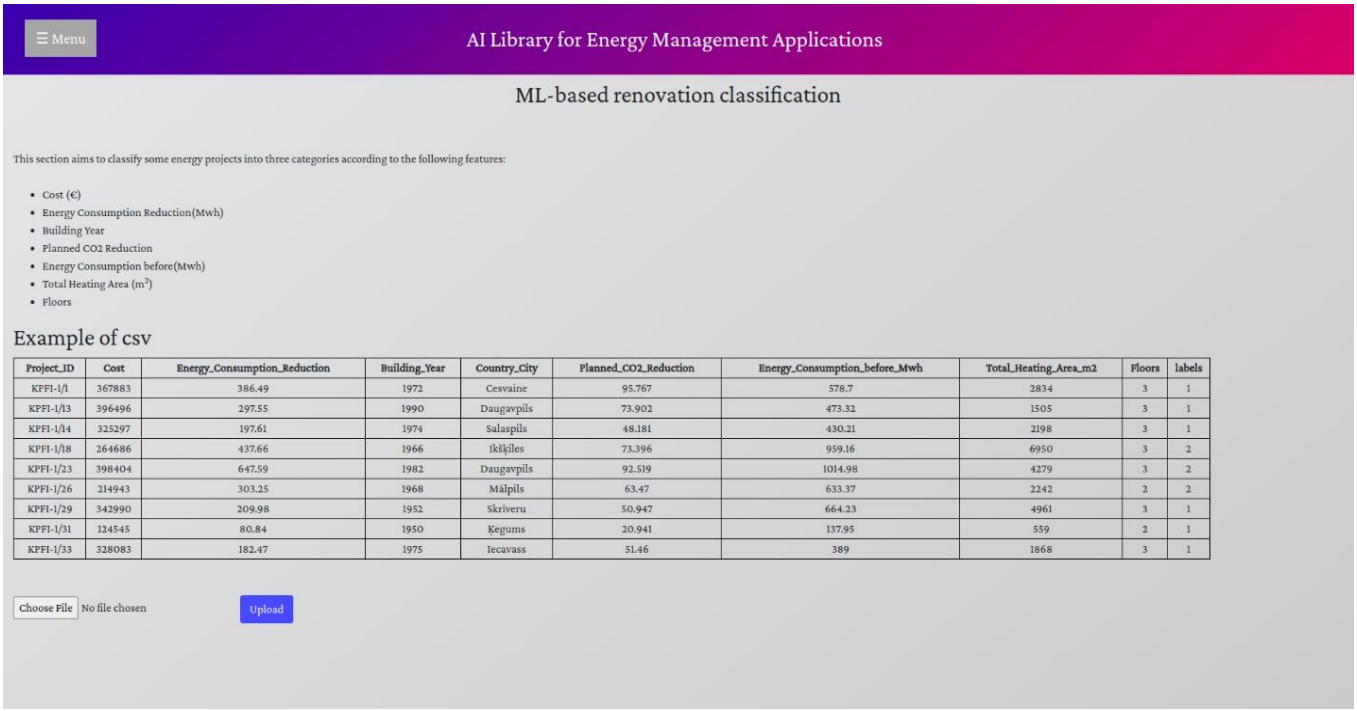

Figure 5.7 Upload Page

Αφού ολοκληρωθεί επιτυχώς το ανέβασμα του αρχείου εμφανίζεται το αντίστοιχο μήνυμα στον χρήστη μαζί με την επιλογή Get Started όπου θα τον οδηγήσει στην επόμενη σελίδα. Η επιλογή αυτή φαίνεται στο παρακάτω στιγμιότυπο.

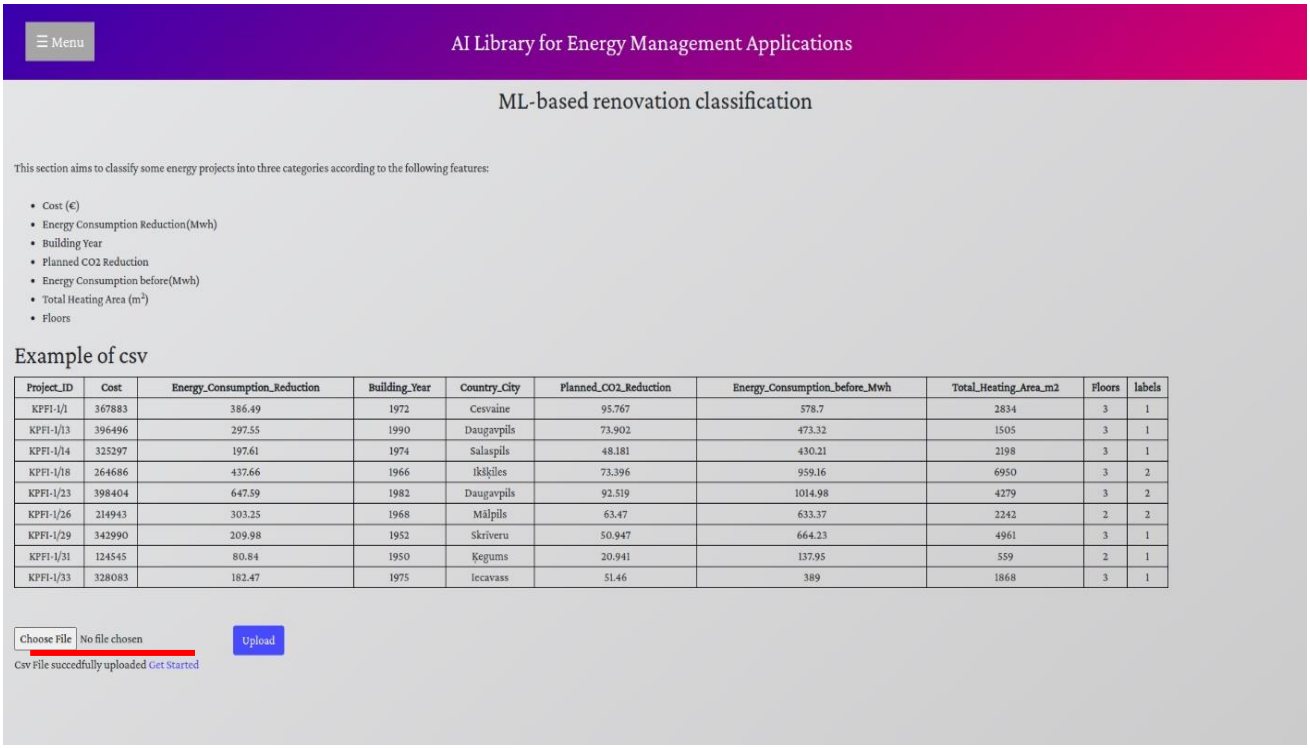

Figure 5.8 Get Started

Το πάτημα της επιλογής Get Started οδηγεί στην κλήση της αντίστοιχης όψης όπου θα επιστρέψει μέσω της μεθόδου render το αντίστοιχο template, στην προκειμένη περίπτωση το template για την προεπεξεργασία των δεδομένων και την επιλογή των αλγορίθμων, μαζί με κάποιο περιεχόμενο. Η επόμενη σελίδα που θα εμφανιστεί στο χρήστη του επιτρέπει να διαλέξει τα χαρακτηριστικά που θα χρησιμοποιηθούν για την εκπαίδευση των μοντέλων, τον τρόπο με τον οποίο θα κανονικοποιηθούν τα δεδομένα, το ποσοστό που θα διαχωριστούν τα δεδομένα σε δεδομένα εκπαίδευσης και στα δείγματα που θα αξιολογήσουμε τα μοντέλα μας καθώς και τους αλγορίθμους που θέλουμε να χρησιμοποιήσουμε. Οι φόρμες που σχεδιάσαμε περιέχουν προεπιλεγμένες τιμές ώστε ο χρήστης να μπορεί να προχωρήσει κατευθείαν στην επόμενη σελίδα. Αρχικά έχουν επιλεχθεί όλα τα διαθέσιμα χαρακτηριστικά και οι αλγόριθμοι, με τον χρήστη να μπορεί να αποεπιλέξει ότι δεν θέλει να χρησιμοποιηθεί.

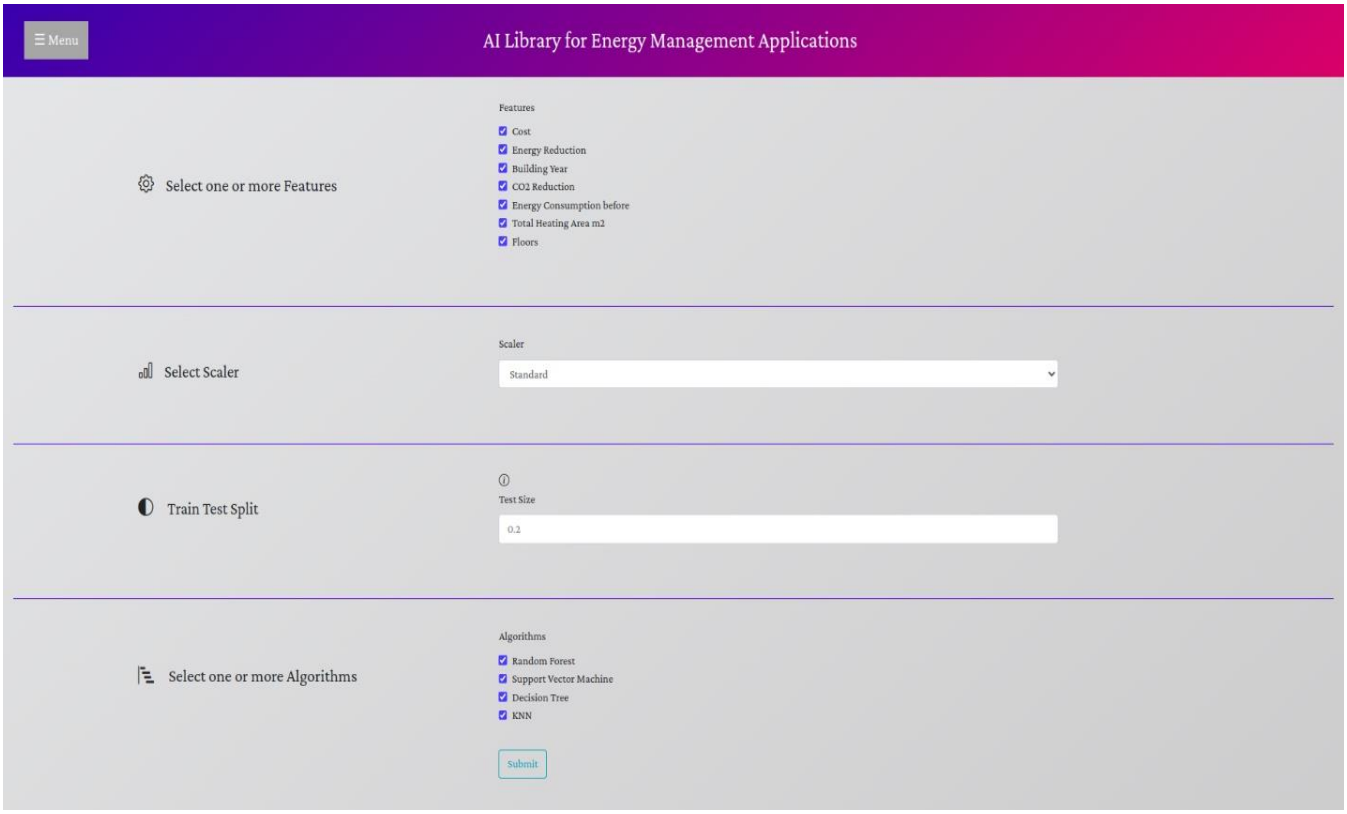

Figure 5.9 Model Selection and Data Preprocessing

Πατώντας το κουμπί Submit οι πληροφορίες που έχουν συμπληρωθεί σε αυτές τις φόρμες μεταβιβάζονται εξυπηετώντας ένα POST request. Οι πληροφορίες αυτές διατηρούνται στο σύστημα ενώ στη συνέχεια εμφανίζεται η επόμενη σελίδα με τις υπερπαραμέτρους για τους αλγορίθμους που επιλέξαμε. Οι διαθέσιμες υπερπαράμετροι εμφανίζονται ως φόρμες με βάση την επιλογή των αλγορίθμων που έγινε στο προηγούμενο βήμα. Οι φόρμες έχουν πάλι ορισμένες αρχικές προεπιλεγμένες τιμές επιτρέποντας παράλληλα στο χρήστη να κάνει τις αλλαγές που επιθυμεί και να δοκιμάσει τις τιμές που θέλει να ορίσει αυτός. Παρακάτω παρουσιάζουμε δύο στιγμιότυπα που αφορούν την ίδια σελίδα με τις υπερπαραμέτρους για το πρόβλημα που έχουμε επιλέξει.

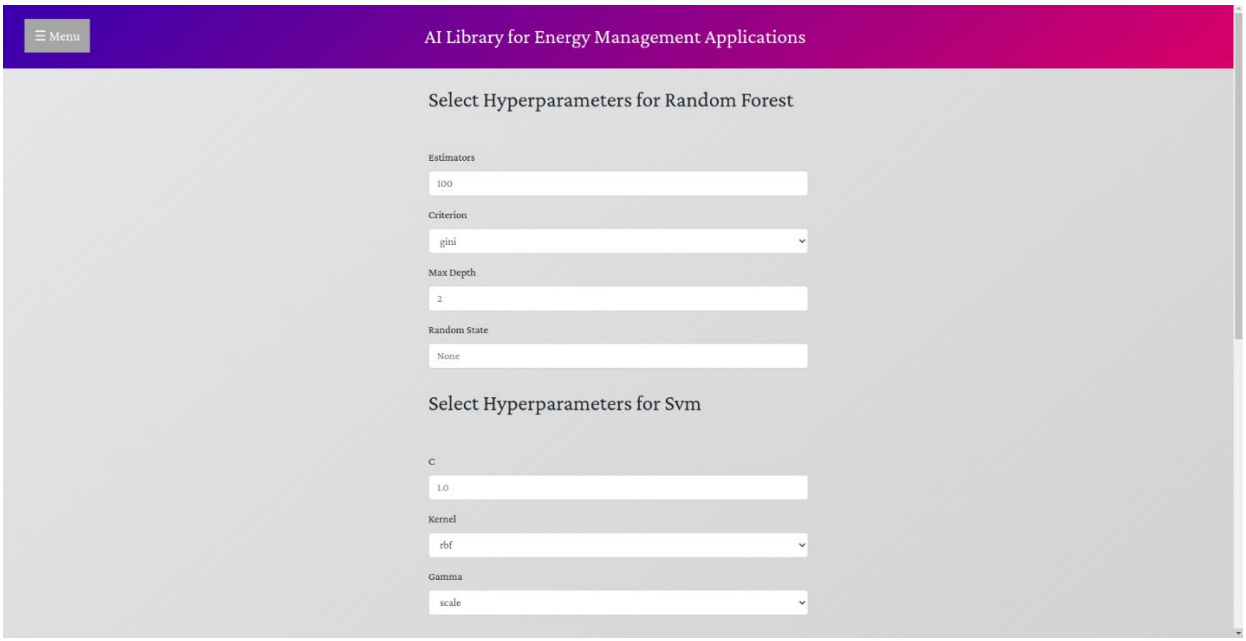

Figure 5.10 Hyperparameter Tuning 1

Στο παραπάνω στιγμιότυπο φαίνονται οι διαθέσιμες υπερπαράμετροι για τους αλγορίθμους Random Forest και για τη μέθοδο Support Vector Machine ενώ στο στιγμιότυπο που έπεται φαίνονται οι υπερπαράμετροι για τους υπόλοιπους αλγορίθμους που επιλέχθηκαν.

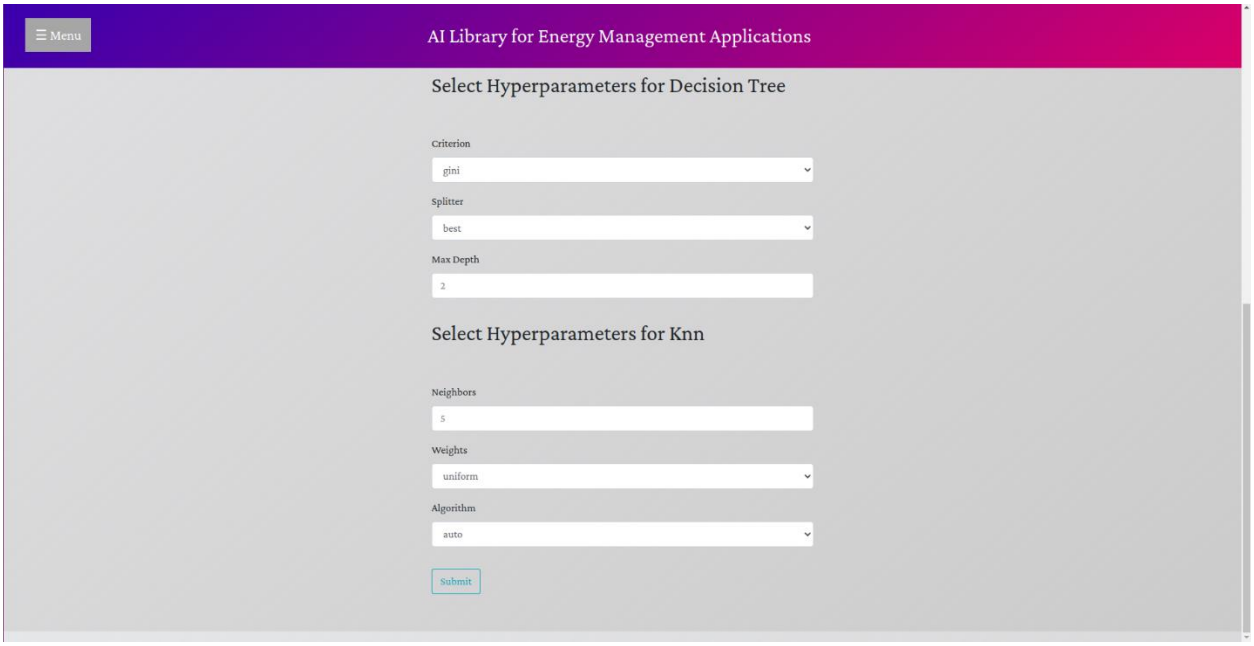

Figure 5.11 Hyperparameter Tuning 2

Με το πάτημα του κουμπιού Submit ξεκινάει η εκπαίδευση των μοντέλων. Για τα προβλήματα που εξετάζουμε έχει δημιουργηθεί η αντίστοιχη συνάρτηση που δέχεται ως ορίσματα όλες τις επιλογές που έχει κάνει ο χρήστης. Συγκεκριμένα δέχεται μία λίστα με τα μοντέλα που θα χρησιμοποιηθούν, το ποσοστό διαχωρισμού των δεδομένων εκπαίδευσης, τη μέθοδο με την οποία θα κανονικοποιηθούν τα δεδομένα καθώς και μία λίστα με τις τιμές των υπερπαραμέτρων των αλγορίθμων. Η συνάρτηση θα επιστρέψει ένα λεξικό με τις μετρικές που υπολογίστηκαν. Η επόμενη σελίδα που θα επιστραφεί είναι η σελίδα των αποτελεσμάτων μαζί με το περιεχόμενο του λεξικού. Τα αποτελέσματα θα οπτικοποιηθούν σε διαγράμματα με βάση τις μετρικές του αντίστοιχου προβλήματος. Παρακάτω ακολουθεί το στιγμιότυπο για τις τέσσερεις μετρικές που υπολογίστηκαν για το πρόβλημα ταξινόμησης, την μετρική accuracy score, την Precision, την Recall και την F1score.

Figure 5.12 Metrics

Η δομή της εφαρμογής και η ροή της πληροφορίας γίνεται με τον ίδιο τρόπο και για τα υπόλοιπα τρία προβλήματα. Η διαφορά βρίσκεται στα διαθέσιμα μοντέλα που μπορεί να επιλέξει ο χρήστης, συνεπώς και στις υπερπαραμέτρους που έχει κάθε μοντέλο. Επίσης, και τα χαρακτηριστικά των δεδομένων καθώς και τα σύνολα δεδομένων είναι διαφορετικά. Επιπλέον, τα άλλα τρία προβλήματα είναι προβλήματα παλινδρόμησης και πρόβλεψης, συνεπώς χρησιμοποιούνται διαφορετικές μετρικές για την αξιολόγησή τους. Μία ακόμη διαφορά είναι στο χρόνο που απαιτείται για την εκπαίδευση των μοντέλων για τα διαφορετικά προβλήματα. Για παράδειγμα, αν επιλέξουμε μεγάλο αριθμό εποχών στο τέταρτο πρόβλημα

τότε θα χρειαστεί να περιμένουμε λίγη ώρα μέχρι να εκπαιδευτούν τα μοντέλα. Στα υπόλοιπα τρία προβλήματα παριστάνουμε διαγραμματικά τις τιμές που προβλέφθηκαν σε σχέση με τις πραγματικές τιμές που υπάρχουν στα σύνολα δεδομένων. Παρακάτω παρουσιάζουμε τις χρονοσειρές που προκύπτουν από την πρόβλεψη των τιμών των επιλεγμένων μοντέλων για το πρόβλημα της μεσοπρόθεσμης παραγωγής ενέργειας των φωτοβολταϊκών. Το διάγραμμα αφορά τα πρώτα 100 δείγματα.

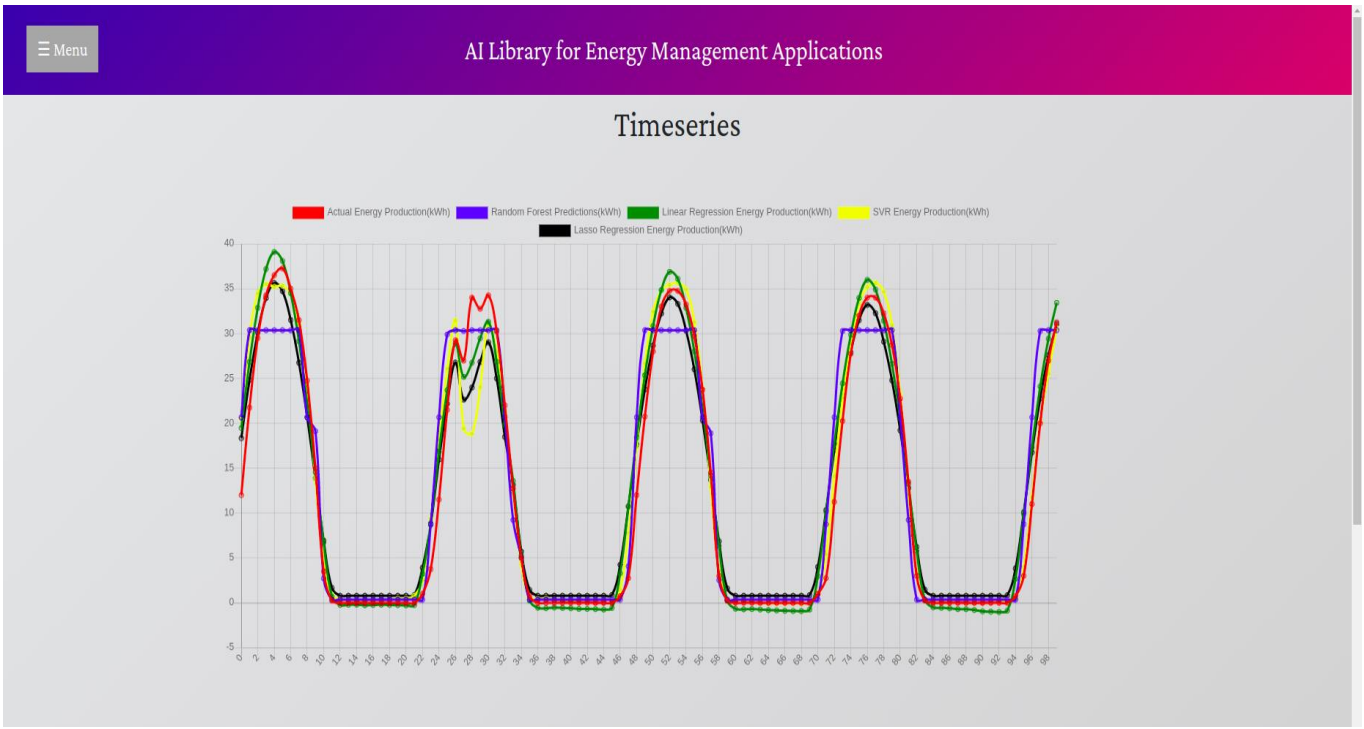

Figure 5.13 Regression Predictions

Το επόμενο στιγμιότυπο παρουσιάζει τις μετρικές που χρησιμοποιήθηκαν για τα υπόλοιπα τρία προβλήματα, την Mean Absolute Error, την Mean Squared Error, την Root Mean Squared Error και την μετρική R2. Όπως μπορούμε να διαπιστώσουμε, η αναπαράσταση των μετρικών μας δίνει τη δυνατότητα να αποκτήσουμε μία ακριβή εικόνα για την απόδοση των μοντέλων στο σύνολο του δείγματος που δοκιμάστηκαν.

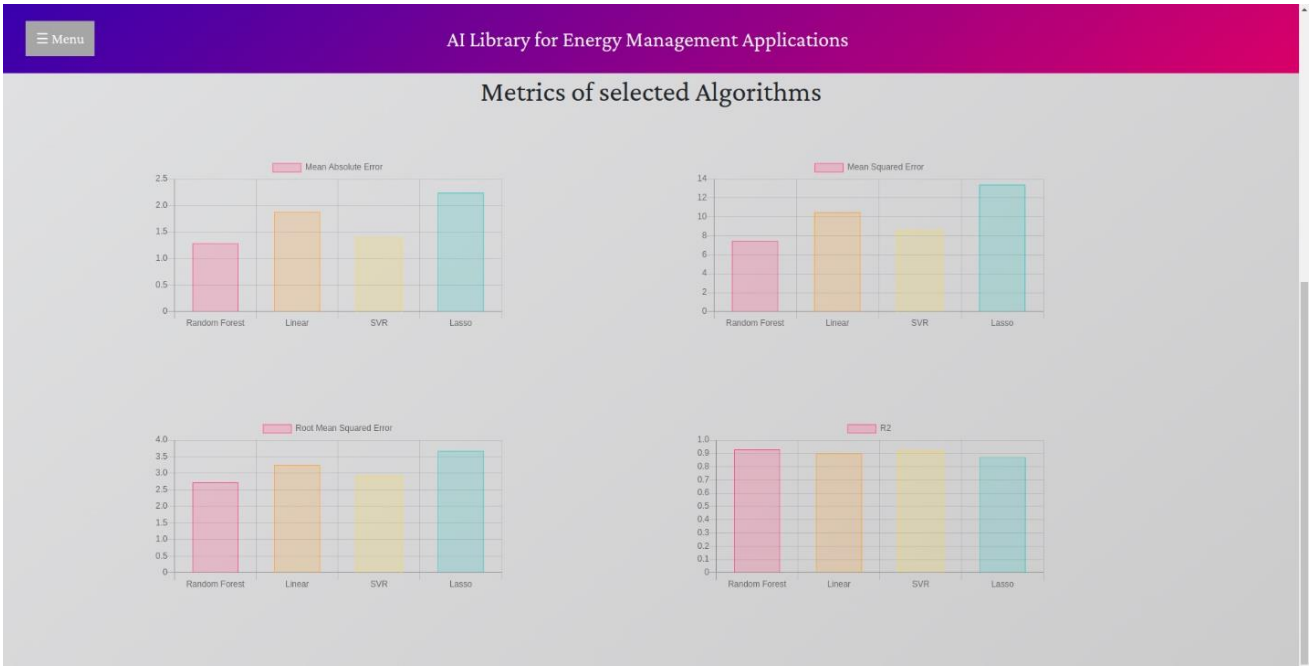

Figure 5.14 Regression metrics

Ο χρήστης έχει τη δυνατότητα κάνοντας hover πάνω στις μπάρες των μετρικών να δει την αριθμητική τιμή που υπολογίστηκε στρογγυλοποιημένη σε δύο δεκαδικά ψηφία. Στην παρακάτω εικόνα βλέπουμε ότι η ακρίβεια που υπολογίστηκε για τη μέθοδο Support Vector Machine είναι 0.92.

Figure 5.15 Accuracy Score value

Στη συνέχεια θα δείξουμε δύο παραδείγματα για την επιρροή που έχουν οι υπερπαράμετροι των μοντέλων στην απόδοσή τους. Παρακάτω παρουσιάζονται δύο εικόνες με προβλέψεις του αλγορίθμου Random Forest για το πρόβλημα της μεσοπρόθεσμης πρόβλεψης παραγωγής ενέργειας φωτοβολταϊκών. Στη φόρμα για την επιλογή των αλγορίθμων έχει επιλεχθεί ο Random Forest ώστε να εμφανίζεται μόνο αυτό το μοντέλο στο διάγραμμα μαζί με τα πραγματικά δεδομένα και να είναι πιο ευδιάκριτα τα αποτελέσματα που προκύπτουν. Στην πρώτη περίπτωση έχουμε αφήσει την προεπιλογή της υπερπαραμέτρου max depth στην τιμή 2 ενώ στη δεύτερη περίπτωση ορίζουμε το μέγιστο βάθος των δέντρων που θα κατασκευαστούν στην τιμή 5.

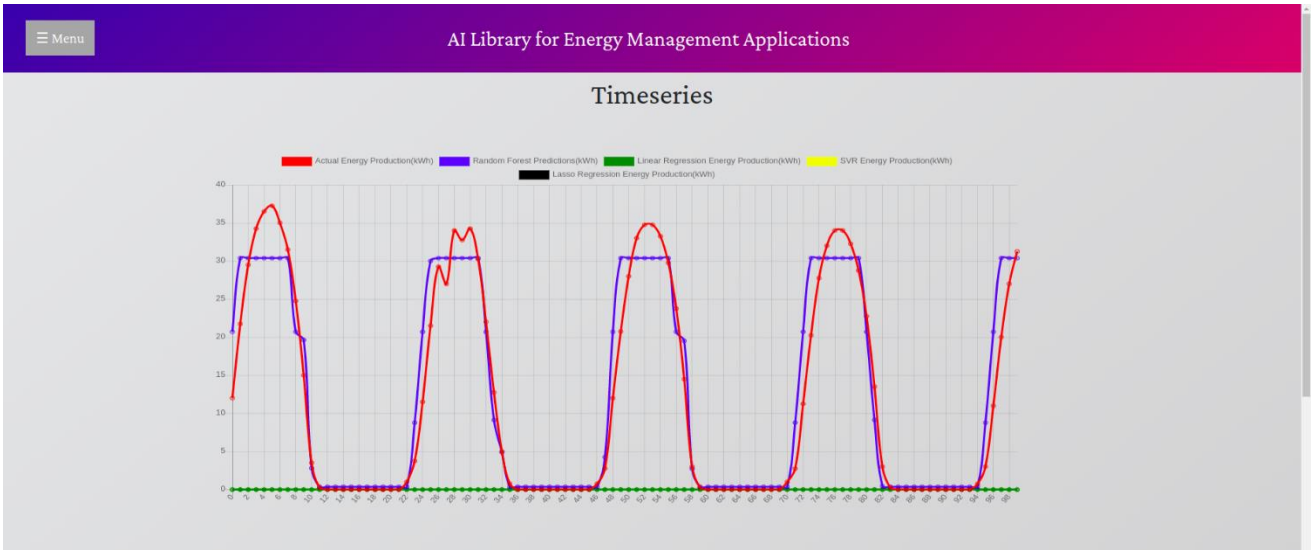

Figure 5.16 max depth=2

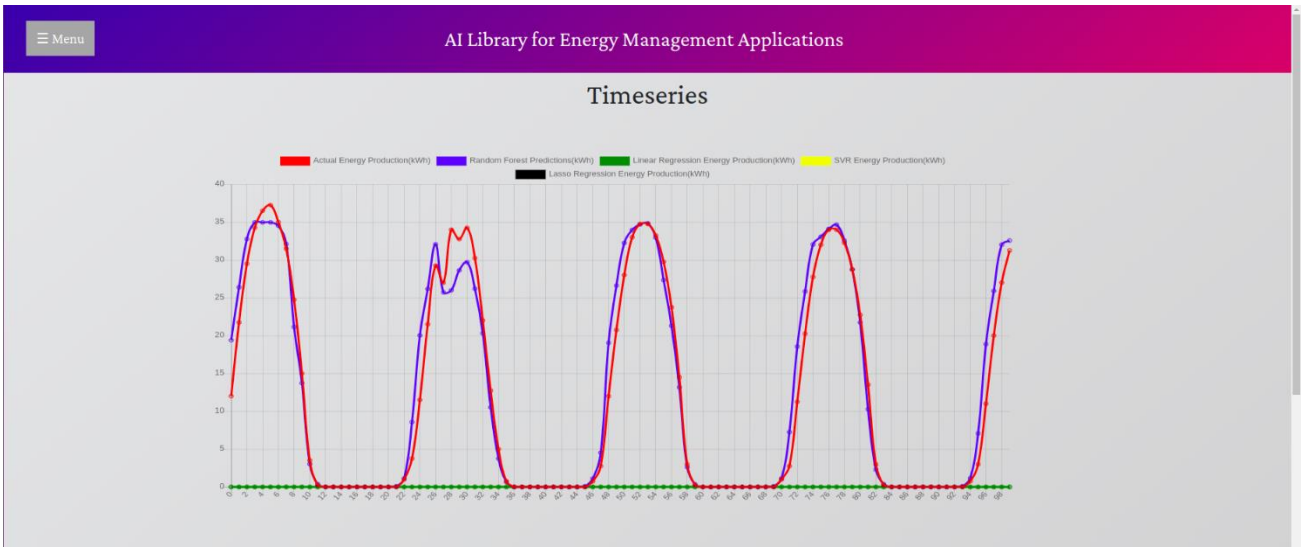

Figure 5.17 max depth=5

Όπως βλέπουμε το μοντέλο προσαρμόζεται καλύτερα με μέγιστο βάθος 5 και προβλέπει τιμές πιο κοντά στα πραγματικά δεδομένα. Αυτό είναι λογικό να συμβαίνει όπως είδαμε και στο κεφάλαιο 3, καθώς η κατασκευή δέντρων με μικρό βάθος δεν προσαρμόζεται καλά στα δεδομένα εκπαίδευσης του προβλήματος. Τέλος παρουσιάζουμε συγκριτικές αποδόσεις για τα νευρωνικά δίκτυα που σχεδιάσαμε στο τέταρτο πρόβλημα. Στην πρώτη εικόνα φαίνονται οι αποδόσεις των μοντέλων για τις προεπιλεγμένες τιμές των υπερπαραμέτρων. Ενδεικτικά αναφέρουμε ότι το βήμα εκμάθησης στο LSTM μοντέλο έχει οριστεί στην τιμή 0.001 και εκπαιδεύεται για 5 εποχές. Αναφέρουμε αυτές τις δύο υπερπαραμέτρους γιατί θα τις μεταβάλλουμε σε επόμενο σενάριο εκπαίδευσης.

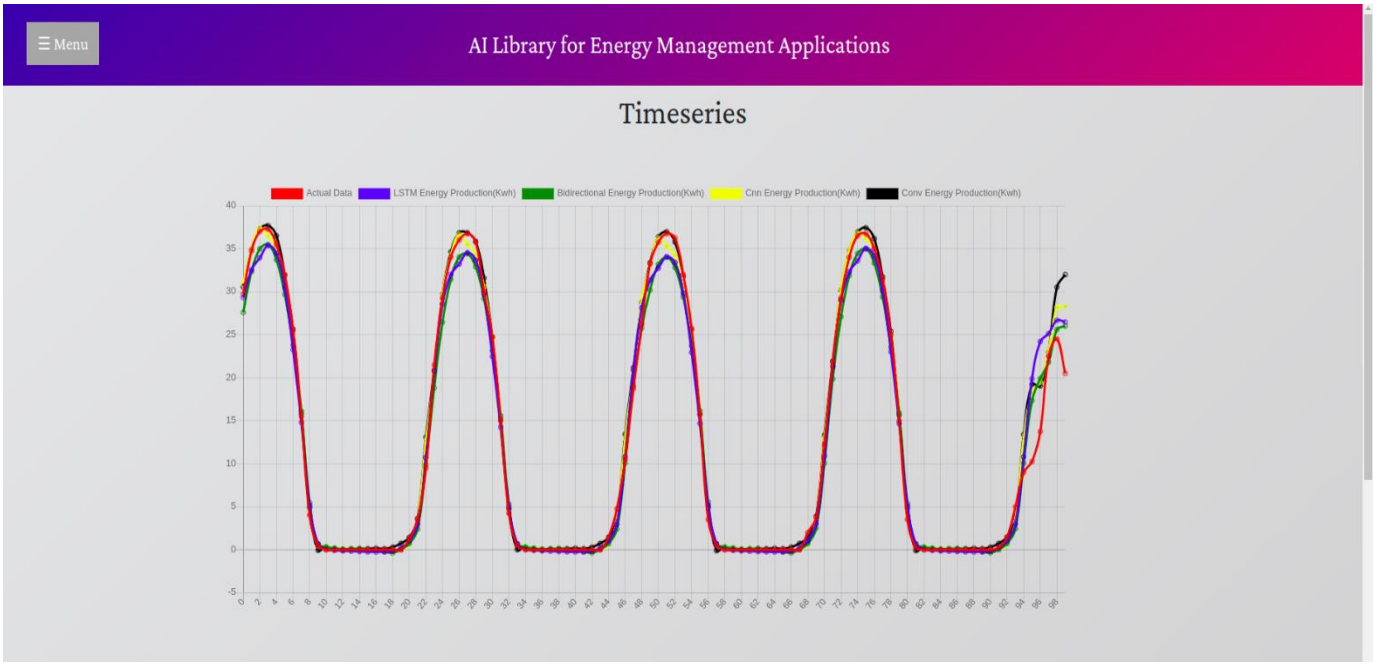

Figure 5.18 default values

Όπως προκύπτει από το στιγμιότυπο οι τιμές προβλέψεων του νευρωνικού δικτύου που χρησιμοποιεί μόνο στρώματα LSTM, το οποίο απεικονίζεται με την μπλε γραμμή, έχουν τη μεγαλύτερη απόκλιση από τις τιμές των πραγματικών δεδομένων, οι οποίες αναπαρίστανται με την κόκκινη γραμμή και παρουσιάζουν τη χειρότερη συμπεριφορά ανάμεσα στα εξεταζόμενα μοντέλα. Ακολούθως παρουσιάζουμε και τις μετρικές των μοντέλων που υπολογίστηκαν για τις συγκεκριμένες υπερπαραμέτρους.

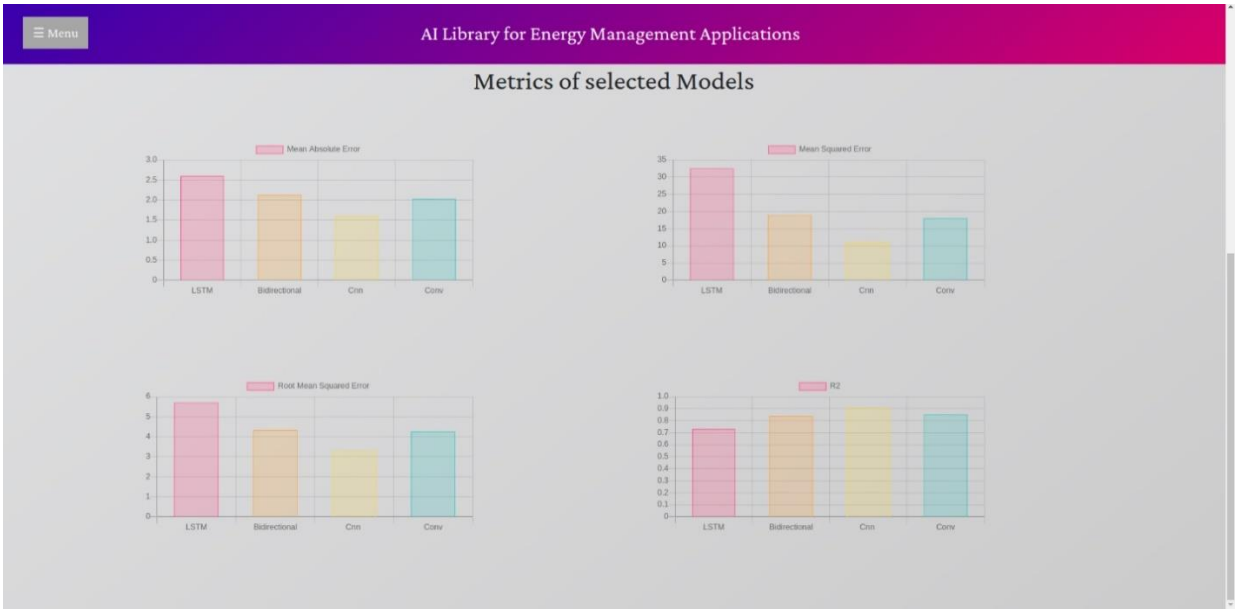

Figure 5.19 metrics for default values

Και η δεύτερη εικόνα επιβεβαιώνει τη χαμηλή επίδοση του μοντέλου LSTM καθώς οι μετρικές σφάλματος παρουσιάζουν υψηλές τιμές. Επανερχόμενοι πάλι στον ορισμό τον υπερπαραμέτρων, θέτουμε αυτή τη φορά το βήμα εκμάθησης για το μοντέλο LSTM στην τιμή 0.005 ενώ το εκπαιδεύουμε για 20 εποχές. Στα υπόλοιπα μοντέλα διατηρούμε τις προκαθορισμένες τιμές στις υπερπαραμέτρους εκπαίδευσης, δηλαδή σε 5 εποχές και βήμα εκμάθησης 0.001 όπως και προηγουμένως. Η εικόνα παρουσιάζει τις τιμές που προβλέφθηκαν

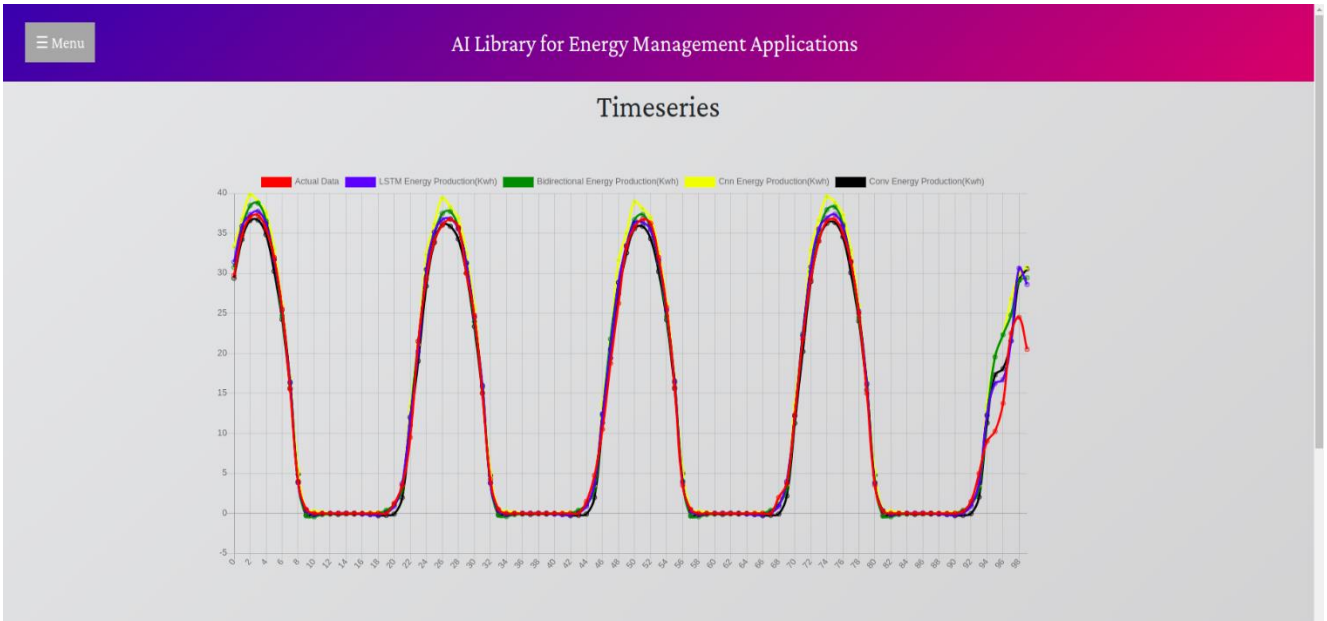

Figure 5.20 Better LSTM

Είναι φανερό ότι η εκπαίδευση του LSTM μοντέλου έχει γίνει με μεγαλύτερη επιτυχία σε αυτήν την περίπτωση και οι προβλέψεις των τιμών είναι πολύ κοντά στα πραγματικά δεδομένα. Παρακάτω παρουσιάζονται και οι μετρικές που υπολογίστηκαν.

Figure 5.21 Improved metrics

Η διαφορά απόδοσης ανάμεσα στα μοντέλα είναι εμφανής. Το μοντέλο LSTM αποκτά καλύτερη δυνατότητα να προβλέπει τιμές για τα δεδομένα που εξετάζουμε τόσο σε σχέση με τα άλλα μοντέλα, όσο και σε σχέση με τον εαυτό του για τις παραμέτρους που εξετάστηκαν στην πρώτη περίπτωση. Οι μετρικές σφάλματος στη δεύτερη περίπτωση είναι αρκετά χαμηλότερες ενώ επιτυγχάνει και την υψηλότερη τιμή στην μετρική R2.

5.5 Επίλογος

Σε αυτό το κεφάλαιο παρουσιάσαμε τα εργαλεία, τις βιβλιοθήκες και τα διαγράμματα που σχεδιάστηκαν για τις ανάγκες της εφαρμογής. Ο κύριος όγκος του κεφαλαίου όμως αφιερώνεται στην παρουσίαση των οθονών της εφαρμογής. Δείξαμε την αρχική σελίδα και τη σελίδα about που περιέχει σύντομες πληροφορίες σχετικά με τα προβλήματα. Επίσης απεικονίστηκε και η σελίδα με τις φόρμες για την επιλογή των χαρακτηριστικών, τις επιλογές σχετικά με την προεπεξεργασία των δεδομένων και την

επιλογή των μοντέλων. Η τελευταία σελίδα πριν την παρουσίαση των αποτελεσμάτων αφορούσε την επιλογή των υπερπαραμέτρων.

Επιπλέον παρουσιάσαμε συνοπτικά ένα σενάριο χρήσης της εφαρμογής που περιλαμβάνει τις βασικές λειτουργικότητες του συστήματος. Φυσικά υπάρχουν πάρα πολλές επιλογές που μπορεί να εξετάσει ο χρήστης με την επιλογή των μοντέλων και των υπερπαραμέτρων τους και να λάβει μια σειρά από διαφορετικά αποτελέσματα. Επικεντρωθήκαμε στην περιήγηση του χρήστη στην εφαρμογή για το πρόβλημα της ταξινόμησης, παραθέτοντας όμως και τα στιγμιότυπα των αποτελεσμάτων για το πρόβλημα της παλινδρόμησης. Η επιλογή κάθε προβλήματος οδηγεί σε διαφορετικά μοντέλα, τα οποία είδαμε και αναλύσαμε στο κεφάλαιο 3, στις αντίστοιχες υπερπαραμέτρους που έχουμε επιλέξει και στην αξιολόγησή τους με τις ανάλογες μετρικές. Επιλέξαμε ορισμένες από τις υπερπαραμέτρους που διαθέτει η βιβλιοθήκη scikit-learn και κάποιες για τα νευρωνικά δίκτυα που κατασκευάσαμε. Η επιλογή έγινε με βάση την επιρροή που έχουν οι υπερπαράμετροι αυτοί στην αποτελεσματικότητα των μοντέλων και παραθέσαμε το παράδειγμα του αλγορίθμου Random Forest, όπου η μεταβολή της τιμής της υπερπαραμέτρου μέγιστο βάθος από 2 σε 5 προκάλεσε σημαντική διαφοροποίηση στις προβλέψεις των τιμών παραγωγής ενέργειας για τα φωτοβολταϊκά. Τέλος εξετάσαμε τη σημασία και το ρόλο των υπερπαραμέτρων στα μοντέλα νευρωνικών δικτύων που σχεδιάσαμε και είδαμε πως μπορεί να μεταβληθεί σε μεγάλο βαθμό η απόδοση των μοντέλων με βάση τις τιμές που ορίζουμε.

ΚΕΦΑΛΑΙΟ 6

ΣΥΜΠΕΡΑΣΜΑΤΑ

6.1 Σύνοψη

Η ανάγκη για αντιμετώπιση των περιβαλλοντικών προκλήσεων που αντιμετωπίζουμε σήμερα απαιτεί μεγαλύτερη εκμετάλλευση των ανανεώσιμων πηγών ενέργειας και μείωση της ενεργειακής κατανάλωσης. Η ανάγκη αυτή έρχεται σε μία εποχή όπου υπάρχει τεράστιος όγκος δεδομένων, ο οποίος εφόσον αξιοποιηθεί αποτελεσματικά μπορεί να προσφέρει σημαντικές λύσεις.

Σε αυτήν τη διπλωματική εργασία προχωρήσαμε στην ανάπτυξη ενός πληροφοριακού συστήματος για τη διαχείριση τεσσάρων προβλημάτων ενέργειας. Παρουσιάσαμε το ρόλο των δεδομένων και της τεχνητής νοημοσύνης στη διαχείριση της ενέργειας και είδαμε αλγορίθμους και μοντέλα νευρωνικών δικτύων τα οποία μπορούν να συνεισφέρουν σημαντικά σε προβλήματα πρόβλεψης και ταξινόμησης. Επίσης, είδαμε τον τρόπο και τη διαδικασία που πρέπει να ακολουθούμε για να επεξεργαζόμαστε τα αρχικά δεδομένα που διαθέτουμε ώστε να μετασχηματίζονται σε δεδομένα κατάλληλα για την εκπαίδευση των μοντέλων. Παράλληλα, σημειώσαμε ορισμένες από τις μετρικές απόδοσης και σφάλματος οι οποίες μας βοηθούν στην αξιολόγηση των μοντέλων μας, ανάλογα με την κατηγορία που ανήκει το πρόβλημα που μελετάμε.

Στη συνέχεια προχωρήσαμε στην ανάπτυξη της μεθοδολογίας και την παρουσίαση των βημάτων για τη διαδικασία της πρόβλεψης. Για κάθε βήμα ξεχωριστά είδαμε τις ενέργειες που πρέπει να εκτελεστούν ώστε να οδηγηθούμε από το αρχικό σύνολο δεδομένων σε αποτελέσματα που προσφέρουν χρήσιμα συμπεράσματα για την απόδοση των μοντέλων. Υπογραμμίσαμε επίσης στο κομμάτι της ανάπτυξης των μοντέλων τη σημασία που έχουν οι υπερπαράμετροί τους για την αποτελεσματική τους επίδοση στις προβλέψεις που πραγματοποιούνται.

Ακόμη, αναφέραμε τα εργαλεία και τις βιβλιοθήκες που χρησιμοποιήθηκαν για την ανάπτυξη του πληροφοριακού συστήματος, τα διαγράμματα UML με βάση τα οποία προχωρήσαμε στον αρχικό σχεδιασμό της εφαρμογής και παρουσιάσαμε τη διεπαφή του χρήστη με το πληροφοριακό σύστημα. Τέλος είδαμε τις διαθέσιμες ενέργειες που μπορεί να εκτελέσει ο χρήστης κατά την περιήγησή του στην εφαρμογή και παραθέσαμε τα αντίστοιχα στιγμιότυπα για κάποια από τα σενάρια λειτουργίας και τις διαθέσιμες επιλογές που μπορεί ο χρήστης να πραγματοποιήσει.

6.2 Μελλοντικές Επεκτάσεις

Η ανάπτυξη της παρούσας διπλωματικής εργασίας οδήγησε σε ορισμένες διαπιστώσεις σχετικά με μελλοντικές επεκτάσεις που θα μπορούσαν να εφαρμοστούν στο πληροφοριακό σύστημα που υλοποιήθηκε. Κατά τη διαδικασία ανάπτυξης προέκυπταν συνεχώς νέοι προβληματισμοί όσον αφορά τις δυνατότητες ενσωμάτωσης περαιτέρω λειτουργιών στο πληροφοριακό σύστημα. Η μεθοδολογία που αναπτύξαμε στο κεφάλαιο 4 θα μπορούσε να εμπλουτιστεί και από άλλα βήματα. Στη συνέχεια παρουσιάζουμε τις κύριες επεκτάσεις που θα συνέβαλαν στην ανάπτυξη του περιεχομένου του πληροφοριακού συστήματος και τις πρόσθετες δυνατότητες και λειτουργίες που αυτές θα παρείχαν στο χρήστη.

Ενσωμάτωση περισσότερων προβλημάτων και μοντέλων

Σε αυτή τη διπλωματική εργασία προχωρήσαμε στη μελέτη τεσσάρων προβλημάτων και αναπτύξαμε τα αντίστοιχα μοντέλα. Παρ' όλα αυτά κι άλλα προβλήματα όπως η βέλτιστη διαχείριση ενέργειας για έξυπνα δίκτυα θα μπορούσαν να προστεθούν στην παρούσα εφαρμογή με το σχεδιασμό και υλοποίηση των αντίστοιχων μεθόδων και μοντέλων. Επίσης, οι αλγόριθμοι που μελετήσαμε είναι ορισμένοι μόνο από όσους έχουν αναπτυχθεί για τις κατηγορίες προβλημάτων που μελετάμε. Υπάρχει η δυνατότητα να αναπτυχθούν επιπλέον μοντέλα δίνοντας κι άλλες επιλογές στον χρήστη. Επιπλέον τα μοντέλα διαθέτουν κι άλλες υπερπαραμέτρους οι οποίες θα μπορούσαν να προστεθούν στο πληροφοριακό σύστημα. Στο πληροφοριακό σύστημα που σχεδιάσαμε επικεντρωθήκαμε στις κύριες υπερπαραμέτρους που προσφέρουν οι βιβλιοθήκες που χρησιμοποιήσαμε για τους αλγορίθμους που έχουμε επιλέξει να ενσωματώσουμε.

Λειτουργίες και διαγράμματα

Όσον αφορά τις λειτουργίες, θα μπορούσαν να προστεθούν ακόμη περισσότερες σε αυτό το πληροφοριακό σύστημα όπως η δυνατότητα να αποθηκεύει ο χρήστης τα μοντέλα που σχεδιάστηκαν και να τα επαναχρησιμοποιεί χωρίς να χρειάζεται να ορίζει ξανά τις ίδιες τιμές για τις υπερπαραμέτρους ή να προχωρά στην απόδοση τιμών κατά την διάρκεια της προεπεξεργασίας δεδομένων. Επιπλέον θα μπορούσε να προστεθεί μία σελίδα σχετικά με το πρόβλημα της ταξινόμησης ενεργειακών έργων όπου ο χρήστης αφού εκπαίδευε τα μοντέλα θα μπορούσε να συμπληρώσει μία φόρμα με τιμές για τα πεδία των διαθέσιμων χαρακτηριστικών ώστε να λάβει μία μεμονωμένη πρόβλεψη για κάποιο έργο που εξετάζει. Το ίδιο θα μπορούσε να συμβεί και για το πρόβλημα της εξοικονόμησης ενέργειας σε έργα που πραγματοποιούνται εργασίες ανακατασκευής, όπου εκεί θα θέλαμε να προβλέψουμε την τιμή της ενέργειας που εξοικονομείται με βάση τη συμπληρωμένη φόρμα των χαρακτηριστικών. Ακόμη, είναι δυνατό να υπάρξει η προσθήκη διαγραμμάτων στην τελευταία σελίδα των αποτελεσμάτων τα οποία θα μπορούσαν να προσφέρουν στο χρήστη μία ακόμα μεγαλύτερη εικόνα όσον αφορά στην αξιολόγηση των προβλέψεων και στη σύγκριση

των μοντέλων. Τέλος, το πληροφοριακό σύστημα που υλοποιήσαμε θα μπορούσε να εξαχθεί ως διαδικτυακή εφαρμογή.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., ... & Widom, J. (2011). Challenges and opportunities with Big Data 2011-1.

[2] Fang, X., Misra, S., Xue, G., & Yang, D. (2011). Smart grid—The new and improved power grid: A survey. IEEE communications surveys & tutorials, 14(4), 944-980.

[3] Zhou, K., Fu, C., & Yang, S. (2016). Big data driven smart energy management: From big data to big insights. Renewable and sustainable energy reviews, 56, 215-225.

[4] Rizwan, A. M., Dennis, L. Y., & Chunho, L. I. U. (2008). A review on the generation, determination and mitigation of Urban Heat Island. Journal of environmental sciences, 20(1), 120-128.

[5] Santamouris, M., Papanikolaou, N., Livada, I., Koronakis, I., Georgakis, C., Argiriou, A., & Assimakopoulos, D. N. (2001). On the impact of urban climate on the energy consumption of buildings. Solar energy, 70(3), 201-216.

[6] Lenzerini, M. (2002, June). Data integration: A theoretical perspective. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 233-246).

[7] Li, H., Yu, L., & He, W. (2019). The impact of GDPR on global technology development. Journal of Global Information Technology Management, 22(1), 1-6.

[8] Atzori, L., Iera, A., & Morabito, G. (2010). The internet of things: A survey. Computer networks, 54(15), 2787-2805.

[9] Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. Machine learning techniques for multimedia: case studies on organization and retrieval, 21-49.

[10] Dayan, P., Sahani, M., & Deback, G. (1999). Unsupervised learning. The MIT encyclopedia of the cognitive sciences, 857-859.

[11] Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. Journal of artificial intelligence research, 4, 237-285.

[12] Ying, X. (2019, February). An overview of overfitting and its solutions. In Journal of physics: Conference series (Vol. 1168, p. 022022). IOP Publishing.

[13] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

[14] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32

[15] Breiman, L. (1996). Bagging predictors. Machine learning, 24, 123-140.

[16] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21.

[17] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

[18] Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883.

[19] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their applications, 13(4), 18-28.

[20] Theodoridis, S., & Koutroumbas, K. (2006). Pattern recognition. Elsevier.

[21] Bishop, C. M. (1994). Neural networks and their applications. Review of scientific instruments, 65(6), 1803-1832.

[22] Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. arXiv preprint arXiv:1710.05941.

[23] Agarap, A. F. (2018). Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375.

[24] Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. Proceedings of the IEEE, 78(10), 1550-1560.

[25] Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems.

[26] Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE transactions on Signal Processing, 45(11), 2673-2681.

[27] Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. Neural computation, 12(10), 2451-2471.

[28] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 30(1), 79-82.

[29] Sarmas, E., Spiliotis, E., Marinakis, V., Koutselis, T., & Doukas, H. (2022). A meta-learning classification model for supporting decisions on energy efficiency investments. Energy and Buildings, 258, 111836.

[30] Sarmas, E., Spiliotis, E., Dimitropoulos, N., Marinakis, V., & Doukas, H. (2023). Estimating the Energy Savings of Energy Efficiency Actions with Ensemble Machine Learning Models. Applied Sciences, 13(4), 2749.

[31] Sarmas, E., Strompolas, S., Marinakis, V., Santori, F., Bucarelli, M. A., & Doukas, H. (2022). An Incremental Learning Framework for Photovoltaic Production and Load Forecasting in Energy Microgrids. Electronics, 11(23), 3962.

[32] Sarmas, E., Dimitropoulos, N., Marinakis, V., Mylona, Z., & Doukas, H. (2022). Transfer learning strategies for solar power forecasting under data scarcity. Scientific Reports, 12(1), 14643.

[33] Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). A gentle introduction to imputation of missing values. Journal of clinical epidemiology, 59(10), 1087-1091

[34] McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. Python for high performance and scientific computing, 14(9), 1-9.

[35] Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. Computing in science & engineering, 13(2), 22-30.

[36] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016, November). Tensorflow: a system for large-scale machine learning. In Osdi (Vol. 16, No. 2016, pp. 265-283).