ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Ανάπτυξη Μοντέλου Πρόβλεψης Παραγωγής Ενέργειας σε ΦΒ Εγκατάσταση Μέσω Ολοκληρωμένης Ανάλυσης Πολλαπλών Ροών Δεδομένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μαρίνα-Νίκη Ι. Τσόπελα

Επιβλέπων: Ιωάννης Ψαρράς

Καθηγητής Ε.Μ.Π

Αθήνα, Ιούλιος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Ανάπτυξη Μοντέλου Πρόβλεψης Παραγωγής Ενέργειας σε ΦΒ Εγκατάσταση Μέσω Ολοκληρωμένης Ανάλυσης Πολλαπλών Ροών Δεδομένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μαρίνα-Νίκη Ι. Τσόπελα

Επιβλέπων: Ιωάννης Ψαρράς

Καθηγητής Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την Ιουλίου 2015

…….	…….	…….
Ιωάννης Ψαρράς	Δημήτριος Ασκούνης	Χρυσόστομος Δούκας
Καθηγητής ΕΜΠ	Αν. Καθηγητής ΕΜΠ	Επ. Καθηγητής ΕΜΠ

Αθήνα, Ιούλιος 2015

.....

Μαρίνα-Νίκη Ι. Τσόπελα

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

ΠΡΟΛΟΓΟΣ

Η εκπόνηση της διπλωματικής εργασίας πραγματοποιήθηκε κατά το ακαδημαϊκό έτος 2014-2015, την περίοδο Μαρτίου2015 – Ιουλίου 2015. Η εργασία σχετίζεται θεματικά με την ερευνητική δραστηριότητα του Εργαστηρίου Συστημάτων Αποφάσεων και Διοίκησης, το οποίο υπάγεται στον Τομέα Ηλεκτρικών Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιο Πολυτεχνείου.

Αρχικά θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Ιωάννη Ψαρρά για την ανάθεση του θέματος καθώς επίσης και τον κ. Χρυσόστομο Δούκα για την καθοδήγηση που μου προσέφερε. Επιπλέον, ιδιαίτερες ευχαριστίες οφείλω στον κ. Γεώργιο Αναστασόπουλο για την άριστη καθοδήγησή του και τις πολύτιμες συμβουλές του. Τις ευχαριστίες μου θα ήθελα να εκφράσω και στην υποψήφια διδάκτορα του ΕΜΠ Φαίδρα Δέδε για την βοήθειά της στην περάτωση της εργασίας.

Αθήνα, Ιούλιος 2015

Μαρίνα-Νίκη Ι. Τσόπελα

ΠΕΡΙΛΗΨΗ

Η απελευθέρωση των αγορών ενέργειας, σε συνδυασμό με την ταχεία είσοδο των ΑΠΕ στα συστήματα ηλεκτρικής ενέργειας, έχουν καταστήσει αναγκαία τη δημιουργία μοντέλων πρόβλεψης, για την παραγωγή της ηλεκτρικής ενέργειας. Η μεταβλητότητα και η περιορισμένη προβλεψιμότητα είναι τα κύρια μειονεκτήματα των ΑΠΕ και συγκεκριμένα ενός ΦΒ συστήματος, τα οποία πρέπει να αντιμετωπίσουν οι διαχειριστές των συστημάτων. Προκειμένου να διεισδύσει μία ΦΒ διάταξη σε ένα σύστημα ηλεκτρικής ενέργειας, είναι απαραίτητο να διακρίνεται από αξιοπιστία και αποτελεσματικότητα. Αυτά τα δύο χαρακτηριστικά, μπορεί να τα εξασφαλίσει ένα μοντέλο πρόβλεψης, καθώς θα προβλέψει αστοχίες και σφάλματα του ΦΒ συστήματος.

Η παραγωγή ηλεκτρικής ενέργειας, μέσω ενός ΦΒ, αποτελεί μη στάσιμη διαδικασία, ενώ η συμπεριφορά του εξαρτάται από ένα πλήθος παραγόντων, με κυριότερες τις καιρικές συνθήκες. Αυτή η εξάρτηση αποτελεί το λόγο, για τον οποίο η πλειονότητα των μοντέλων πρόβλεψης, που χρησιμοποιούνται για ηλιακή παραγωγή, λαμβάνουν δεδομένα εισόδου από μοντέλα προβλέψεων καιρού, τα οποία παρέχουν εκτιμήσεις των μελλοντικών καιρικών συνθηκών. Έχουν αναπτυχθεί διάφορες τεχνικές, για την πρόβλεψη της ηλεκτρικής ενέργειας. Ένα μοντέλο, το οποίο περιγράφει ικανοποιητικά τη σχέση ανάμεσα στην παραγωγή και τις καιρικές συνθήκες, είναι το μοντέλο γραμμικής παλινδρόμησης.

Αντικείμενο της παρούσας διπλωματικής εργασίας, είναι η ανάπτυξη ενός μοντέλου γραμμικής παλινδρόμησης, για την πρόβλεψη της παραγωγής ενός ΦΒ συστήματος. Στόχος μας, είναι η δημιουργία ενός μοντέλου πρόβλεψης, που θα χαρακτηρίζεται από υψηλά επίπεδα ακρίβειας, προκειμένου να αποφευχθούν αστοχίες και αύξηση λειτουργικού κόστους. Για το λόγο αυτό, επιλέχθηκε η υλοποίηση του μοντέλου να γίνει στο λογισμικό περιβάλλον του Rapidminer. Η υλοποίηση γίνεται με χρήση ιστορικών δεδομένων, πολλαπλών πηγών

Αρχικά επιχειρείται η προετοιμασία των δεδομένων, ώστε να μπορεί το μοντέλο να τα διαχειριστεί. Για να επιλεχθεί η σχέση που περιγράφει καλύτερα την παραγωγή ενέγειας, εξετάζεται ένα μεγάλο πλήθος μεγεθών και των δυνατών συνδυασμών τους. Στη συνέχεια, πραγματοποιείται η δημιουργία του μοντέλου, καθώς προκύπτει, πως η παραγωγή εξαρτάται με αρκετά ικανοποιητική ακρίβεια από την ακτινοβολία, μέσω μιας γραμμικής σχέσης. Τέλος, προτείνεται η καταλληλότερη χρήση του μοντέλου, προκειμένου η πρόβλεψη της παραγόμενης ισχύος να βελτιώνεται διαρκώς.

Λέξεις κλειδιά: παραγωγή ενέργειας, φωτοβολταϊκή εγκατάσαση, πολλαπλές ροές δεδομένων, προβλεπτικό μοντέλο παραγωγής ενέργειας, γραμμική παλινδρόμηση, Rapidminer.

ABSTRACT

The deregulation of the energy markets, combined with the rapid entrance of Renewable Energy Sources in electric systems have made compulsory the creation of predictive models for the production of electricity. The variability and limited predictability are the main disadvantages of Renewable Energy, which must be faced by the administrator of the system. In order for a PV system to enter a power system, it is necessary to be characterized by high levels of credibility and effectiveness. These two features can be secured by a predictive model, which will predict failures and errors in the PV system.

Energy production through a PV is a non-stationary process, and the result depends on a number of factors. Usually the production presents a high dependence with the weather. This dependence is the reason why the majority of prediction models for PV energy production, are using forecasting models as data input. Various techniques have been developed in order to predict electric energy production. A model that adequately describes the relationship between production and the weather, is the linear regression model.

The subject of this thesis, is the development of a linear regression model, to predict the production of a PV system. Our goal is to create a model characterized by high levels of accuracy in order to avoid faults and increase operating costs. Therefore, we choose to implement the model through the software environment of Rapidminer. The implementation is achieved with a large amount of past data.

Initially we attempt to prepare the data in order to enable the model and then we create the model, assuming that the production depends on the radiation via a linear equation. Finally, we propose the most appropriate use of the linear model in order to improve the output.

Key words: energy production, photovoltaic system, multiple data streams, predictive model for energy production, Rapidminer, linear regression

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΙΝΑΚΑΣ ΣΧΗΜΑΤΩΝ

ΠΙΝΑΚΑΣ ΠΙΝΑΚΩΝ

ΠΙΝΑΚΑΣ ΕΞΙΣΩΣΕΩΝ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Αντικείμενο της διπλωματικής εργασίας

Η Ενέργεια αποτελεί εδώ και χρόνια ένα πολύ σημαντικό παγκόσμιο ζήτημα. Ο τομέας της ενέργειας είναι αναμφίβολα η βάση κάθε μοντέρνας οικονομίας. Η σπουδαιότητά της είναι τόσο θεωρητική όσο και πρακτική, ειδικά στις μέρες μας, όπου οι κυβερνήσεις αντιμετωπίζουν την πρόκληση για την επίτευξη συγκεκριμένων στόχων, που έχουν σκοπό τον εφοδιασμό βιώσιμων και περιβαλλοντικά ασφαλών μορφών ενέργειας. Οι ενεργειακές τεχνολογίες κεντρίζουν το επιστημονικό, μηχανικό, πολιτικό, κοινωνικό, οικολογικό και οικονομικό ενδιαφέρον σε όλο τον κόσμο.

Έρευνες, σχετικά με τις κλιματολογικές συνθήκες στη γη, έχουν δείξει, ότι η θερμοκρασία της επιφάνειάς της έχει αυξηθεί σημαντικά τους τελευταίους δύο αιώνες. Υπεύθυνες, για αυτό το φαινόμενο, είναι οι αυξανόμενες συγκεντρώσεις των αερίων του θερμοκηπίου στην ατμόσφαιρα. Τέτοια συμπεράσματα, έχουν ευαισθητοποιήσει την κοινή γνώμη και έχουν οδηγήσει τα κράτη σε προώθηση μέτρων για την μείωση της εκπομπής αερίων ρύπων. Καθώς όμως ο άνθρωπος συνεχίζει να επωφελείται της ενέργειας, που προσφέρει η χρήση ορυκτών καυσίμων, το πρόβλημα των επικίνδυνων αερίων εντείνεται. Επιπλέον, το τρέχον ενεργειακό σύστημα, που βασίζεται σε τεράστιο βαθμό στα ορυκτά καύσιμα, δεν είναι πλέον βιώσιμο, καθώς εξαντλούνται τα αποθέματα. Λαμβάνοντας λοιπόν υπόψη τους παραπάνω περιβαλλοντικούς λόγους, καθώς και την αναμενόμενη αύξηση της ζήτησης ηλεκτρικής ενέργειας, γίνεται καθημερινά προσπάθεια στη μείωση της κατανάλωσης ενέργειας, στη δημιουργία αποδοτικότερων ενεργειακών συστημάτων και στην παραγωγή ενέργειας από ανανεώσιμες πηγές ενέργειας.[19]

Οι πιο διαδεδομένες τεχνολογίες ανανεώσιμων πηγών ενέργειας, για την παραγωγή ηλεκτρικής ενέργειας, είναι: τα φωτοβολταϊκά συστήματα, οι ανεμογεννήτριες, οι υδροηλεκτρικοί σταθμοί, η βιομάζα και οι γεωθερμικοί σταθμοί παραγωγής. Πολλές από τις τεχνολογίες αυτές έχουν αναπτυχθεί, σε τέτοιο βαθμό, ώστε να θεωρούνται αποδοτικές και βιώσιμες. Το μόνο μειονέκτημα που παρουσιάζουν αυτές οι μορφές ενέργειας, είναι η μη ελεγχόμενη παραγωγή τους. Αυτό οφείλεται στο γεγονός ότι, η πρωτογενής πηγή ενέργειας (π.χ ήλιος) παρουσιάζει μεταβλητότητα και περιορισμένη προβλεψιμότητα, σε αντίθεση με τους συμβατικούς σταθμούς παραγωγής (π.χ. λιγνιτικούς, φυσικού αερίου, πετρελαϊκούς, κ.α.), η παραγωγή των οποίων είναι σε μεγάλο βαθμό προγραμματιζόμενη και ελεγχόμενη από το Διαχειριστή του Συστήματος. [20]

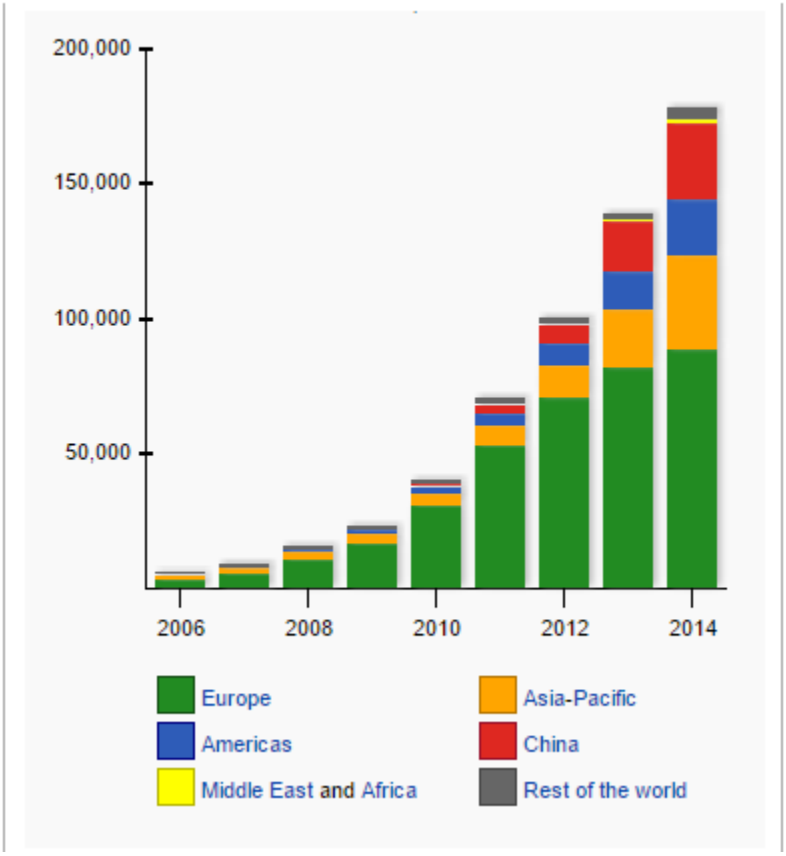
Η παρούσα εργασία αναφέρεται σε μία πρωτογενή πηγή ενέργειας, την ηλιακή, που αξιοποιείται στα φωτοβολταϊκά συστήματα.

Μέσω των φωτοβολταϊκών, η ηλιακή ενέργεια που προέρχεται από τον ήλιο, μετατρέπεται σε ηλεκτρική. Η γη συλλαμβάνει το ένα δισεκατομμυριοστό της εκπεμπόμενης ηλιακής ακτινοβολίας, που όμως αντιστοιχεί σε τεράστια ενεργειακή ποσότητα, αν αναλογιστούμε ότι η ηλιακή ενέργεια που φτάνει στη γη, σε μία εβδομάδα, είναι περίπου ίση με τη

συνολικά αποθηκευμένη ενέργεια όλων των καυσίμων του πλανήτη. Η ηλιακή ενέργεια, στο σύνολό της, είναι πρακτικά ανεξάντλητη, αφού προέρχεται από τον ήλιο και ως εκ τούτου δεν υπάρχουν περιορισμοί χώρου και χρόνου για την εκμετάλλευσή της [20]. Μπορούμε να χωρίσουμε την εκμετάλλευσή της σε τρείς κατηγορίες εφαρμογών: τα παθητικά ηλιακά συστήματα, τα ενεργητικά ηλιακά συστήματα και τα φωτοβολταϊκά συστήματα. Στα πλαίσια της συγκεκριμένης εργασίας, εξετάζουμε ένα φωτοβολταϊκό σύστημα, οπότε οι υπόλοιπες κατηγορίες, δεν θα μας απασχολήσουν.

Αν και όλη η γη δέχεται την ηλιακή ακτινοβολία, η ποσότητα που μπορεί να αξιοποιήσει ένα φωτοβολταϊκο σύστημα, εξαρτάται κυρίως από τη γεωγραφική του θέση, την εποχή και την ακτινοβολία, την υγρασία, τη νέφωση και την θερμοκρασία κάθε χρονική στιγμή.

Στις ημέρες μας, τα φωτοβολταϊκά, είναι μία ταχέως αναπτυσσόμενη αγορά.. Η παγκόσμια ανάπτυξη των φωτοβολταϊκών, εδώ και δύο δεκαετίες, όπως φαίνεται και από το παρακάτω διάγραμμα, ακολουθεί εκθετική τροχιά. Σε αυτό το διάστημα, τα φωτοβολταϊκά έχουν εξελιχθεί, από μία αγορά εφαρμογών μικρής κλίμακας, σε μία από τις κυριότερες πηγές ενέργειας.[21]



Σχήμα 1 Αθροιστική αύξηση φωτοβολταϊκών [21]

Αυτή η, μεγάλης κλίμακας, εξάπλωση της παραγωγής ηλεκτρικής ενέργειας, μέσω της ηλιακής ακτινοβολίας, έχει οδηγήσει σε μία μεγάλη αύξηση της διείσδυσης της ηλιακής ενέργειας σε πολλά δίκτυα ηλεκτρικής ενέργειας. Σε πολλές περιπτώσεις, αυτή η μορφή ηλεκτρικής ενέργειας, μπορεί να καλύψει μεγάλα ποσοστά της συνολικής ζήτησης. Κάτι τέτοιο θα μπορούσε να είναι προβληματικό για τους χειριστές των συστημάτων ηλεκτρικής ενέργειας επειδή, σε αντίθεση με τις συμβατικές πηγές ενέργειας, η ηλιακή είναι συνεχώς μεταβαλλόμενη. Είναι σχεδόν αδύνατο για τον άνθρωπο να ελέγξει εξ ολοκλήρου την παραγωγή ηλεκτρικής ενέργειας από πρωτογενή πηγή ενέργειας. Συνεπώς, η ανάπτυξη και χρήση προηγμένων εργαλείων πρόβλεψης, της παραγόμενης ενέργειας από φωτοβολταϊκά, κρίνεται πλέον απαραίτητη για τη διασφάλιση της ασφαλούς και αποδοτικής λειτουργίας των σύγχρονων συστημάτων ηλεκτρικής ενέργειας. Μπορούμε δηλαδή να χρησιμοποιήσουμε τις προβλέψεις της παραγόμενης ισχύος στον προγραμματισμό των μονάδων παραγωγής και στην ελαχιστοποίηση των λειτουργικών δαπανών. Οι προβλέψεις της ηλιακής ενέργειας θεωρούνται αναγκαίες και μερικές φορές υποχρεωτικές, σε συστήματα ηλεκτρικής ενέργειας, όπου είτε η διείσδυση είναι υψηλή είτε δεν υπάρχει δυνατότητα για επιπλέον εφεδρεία. Η ανάγκη για μια ακριβή πρόβλεψη της ηλιακής ενέργειας αναγνωρίζεται σήμερα από τη βιομηχανία παραγωγής και μεταφοράς ηλεκτρικής ενέργειας, ώστε να υπάρχει δυνατότητα διείσδυσης μεγάλης κλίμακας, χωρίς προβλήματα.

1.2 Η πρόβλεψη της παραγωγής ενέργειας από φωτοβολταϊκά συστήματα

Η πρόβλεψη της ηλιακής παραγωγής, στοχεύει στο να παρέχει στους τελικούς χρήστες τις εκτιμήσεις της πιθανής διαθέσιμης ενέργειας, σε μια δεδομένη στιγμή στο μέλλον. Η μορφή της εξόδου των μοντέλων πρόβλεψης, που θα χρησιμοποιήσουμε, είναι οι προβλέψεις σημείων. Δηλαδή για κάθε χρονικό βήμα πρόβλεψης παρέχεται μία τιμή της παραγόμενης ηλιακής ενέργειας. Επομένως, τα πιθανοτικά μοντέλα πρόβλεψης, που έχουν αναπτυχθεί, παρέχουν όλες τις πιθανές τιμές που μπορεί να λάβει η ηλιακή παραγωγή, μία συγκεκριμένη ώρα.

Διαφορετικές χρήσεις της πρόβλεψης της παραγωγής ενός φωτοβολταϊκού, απαιτούν διαφορετικούς τύπους προβλέψεων. Οι προβλέψεις μπορεί να βρίσκουν εφαρμογή σε ένα μόνο φωτοβολταϊκό σύστημα ή ακόμα και σε ένα μεγάλο αριθμό ΦΒ συστημάτων που εκτείνονται σε μία γεωγραφική περιοχή. Οι προβλέψεις μπορεί να εστιάζουν στην ισχύ της εξόδου του ΦΒ συστήματος ή στον ρυθμό μεταβολής της. Ανάλογα λοιπόν με το τι θέλουμε να προβλέψουμε, αλλάζει και η μέθοδος που χρησιμοποιούμε. Οι μέθοδοι πρόβλεψης εξαρτώνται επίσης από τα εργαλεία και την πληροφορία που έχουμε διαθέσιμη, όπως είναι τα δεδομένα που δεχόμαστε από μετεωρολογικούς σταθμούς και δορυφόρους.[30]

Η ηλιακή ενέργεια αποτελεί μία από τις πιο γρήγορα αναπτυσσόμενες ανανεώσιμες πηγές ενέργειας. Οι δύο κύριες προκλήσεις, για τα υψηλά ποσοστά διείσδυσης των φωτοβολταϊκών συστημάτων, είναι η μεταβλητότητα και η αβεβαιότητα, δηλαδή το γεγονός ότι η έξοδος του φωτοβολταϊκού παρουσιάζει μεταβλητότητα σε όλες τις χρονικές κλίμακες (από δευτερόλεπτα έως και χρόνια) και η ίδια η μεταβλητότητα είναι δύσκολο να προβλεφθεί. Για την αντιστάθμιση αυτής της μεταβλητότητας απαιτούνται μοντέλα πρόβλεψης της παραγόμενης ενέργειας από την ηλιακή ακτινοβολία, για τις επόμενες ώρες, κυρίως για τη διαχείριση αλλά και το εμπόριο της ενέργειας.

1.3 Φάσεις υλοποίησης

Αντικείμενο της συγκεκριμένης εργασίας είναι η μελέτη μίας φωτοβολταϊκής εγκατάστασης, με στόχο την πρόβλεψη της παραγόμενης ηλεκτρικής ενέργειας. Η ανάπτυξη του μοντέλου έγινε μέσω του λογισμικού Rapidminer. Οι λόγοι, για τους οποίους επιλέχθηκε το συγκεκριμένο εργαλείο, αναλύονται στο Κεφάλαιο 2.

Πρώτα έγινε η εγκατάσταση του λογισμικού Rapidminer στον υπολογιστή.

Στη συνέχεια, για την χρησιμοποίηση του RapidMiner Studio, ζητείται η δημιουργία μιας καινούργιας αποθήκης, όπως φαίνεται και στο λογισμικό *repository*. Αρχικά, δημιουργήθηκες μιας *τοπική αποθήκη (local repository)* στον υπολογιστή.

Τα επόμενα βήματα αφορούν τη δημιουργία του μοντέλου πρόβλεψης και θα παρουσιαστούν αναλυτικά στα επόμενα κεφάλαια. Συνοπτικά αναφέρεται, πως ο καθορισμός των διαδικασιών ανάλυσης με το RapidMiner Studio γίνεται με τη βοήθεια και το συνδυασμό τελεστών και ρυθμίζοντας παραμέτρους.

Οι διεργασίες μπορούν να δημιουργηθούν από ένα μεγάλο αριθμό τυχαίων τελεστών και είναι δυνατό να παρουσιαστούν από ένα γράφημα διεργασίας (flow design). Το RapidMiner Studio ελέγχει διαρκώς τη διαδικασία που αναπτύσσεται, έτσι ώστε να μην υπάρχουν συντακτικά λάθη και ταυτόχρονα προτείνει λύσεις. Αυτό είναι εφικτό από το meta-data transformation, το οποίο επιτρέπει την έγκαιρη εύρεση λύσης, σε περιπτώσεις όπου ο συνδυασμός των χειριστών δεν είναι ο κατάλληλος (quick fixes). Επιπλέον, το RapidMiner Studio, προσφέρει τη δυνατότητα καθορισμού κρίσιμων σημείων (breakpoints) και κατά συνέπεια την εικονική επιθεώρηση κάθε ενδιάμεσου σημείου.

Από μόνο του, το RapidMiner Studio, διαθέτει πάνω από 1.500 λειτουργίες, για κάθε επαγγελματική ανάλυση δεδομένων, υποστηρίζει τα 3-D γραφήματα, τους πίνακες διασποράς και επιτρέπει στον χρήστη να προσαρμόζει πλήρως τα δεδομένα και τα γραφήματά του.

1.4 Δομή της εργασίας

Ο σκοπός της παρούσας εργασίας είναι να παρουσιάσει ένα πλαίσιο με οδηγίες, για κάποιον που επιθυμεί να αναπτύξει ένα μοντέλο πρόβλεψης της παραγόμενης ηλεκτρικής ενέργειας από την ηλιακή ακτινοβολία. Συγκεκριμένα, παρουσιάζει την υλοποίηση ενός μοντέλου πρόβλεψης, μέσω του στατιστικού λογισμικού Rapidminer και όλα τα στάδια της δημιουργίας του. Δηλαδή τη συλλογή των παρελθοντικών δεδομένων, τη διαχείριση και τις προσαρμογές των δεδομένων, τις σχέσεις της απόδοσης του στατιστικού μοντέλου με τις μεταβλητές εισόδου που εφαρμόζουμε και τέλος την απόδοση του παραγόμενου μοντέλου. Ακολουθεί μία σύντομη αναφορά στα κεφάλαια της εργασίας.

Στο *Κεφάλαιο 1*, πραγματοποιήθηκε σύντομη αναφορά στα χαρακτηριστικά και τη σπουδαιότητα της ηλιακής ενέργειας, καθώς επίσης και στην αξία δημιουργίας μοντέλων πρόβλεψης. Αναλύονται οι φάσεις υλοποίησης και παρουσιάζεται η δομή της εργασίας.

Στο *Κεφάλαιο 2,* παρουσιάζονται τα πλεονεκτήματα, που προσφέρει η εξόρυξη δεδομένων στα συστήματα ηλεκτρικής ενέργειας, καθώς επίσης και οι μέθοδοι ανάλυσης, που μπορούν να εφαρμοστούν, μέσω της εξόρυξης δεδομένων. Επίσης, γίνεται αναφορά στην ιστορική πορεία του Rapidminer, στους λόγους για την επιλογή του Rapidminer, ως το στατιστικό εργαλείο για τη δημιουργία του μοντέλου πρόβλεψης έναντι άλλων ανταγωνιστικών λογισμικών, αναλύονται τα χαρακτηριστικά του και συγκρίνονται με άλλα δημοφιλή λογισμικά.

Στο *Κεφάλαιο 3*, γίνεται μια παρουσίαση της μεθοδολογίας, προκειμένου να καταλήξουμε στην εμπειρική αλλά και στατιστική ανάλυση των δεδομένων. Επίσης, παρουσιάζονται διαδικασίες προ-επεξεργασίας των αρχικών δεδομένων, έτσι ώστε να έρθουν σε μορφή κατάλληλη για ανάλυση. Αναλύονται και εξετάζονται οι κυριότεροι στατιστικοί δείκτες ανάλυσης της ακρίβειας της πρόβλεψης, ιδιαίτερα σημαντικοί, προκειμένου να μπορεί να διεξαχθεί η αξιολόγηση των μοντέλων πρόβλεψης.

Στο *Κεφάλαιο 4,* παρουσιάζεται η εφαρμογή της μεθοδολογίας για τη διαδικασία πρόβλεψης. Αναλύονται όλα τα στάδια της προετοιμασίας της χρονοσειράς, και αναλύεται η αντιμετώπιση των προβλημάτων που εμφανίστηκαν κατά τη διαδικασία. Παρουσιάζονται όλοι οι τελεστές που χρησιμοποιήθηκαν και αναφέρονται οι ρόλοι και τα ιδιαίτερα χαρακτηριστικά τους.

Στο *Κεφάλαιο 5*, παρουσιάζεται η υλοποίηση του μοντέλου πρόβλεψης. Τα αποτελέσματα αναλύονται και ελέγχονται, ενώ υπολογίζεται η ακρίβεια της πρόβλεψης.

Στο *Κεφάλαιο 6*, το οποίο αποτελεί και τον επίλογο της διπλωματικής εργασίας, αναδεικνύονται τα κυριότερα συμπεράσματα, που προέκυψαν από τα προηγούμενα κεφάλαια της παρούσας μελέτης. Επιπλέον, καταγράφονται οι προοπτικές για περαιτέρω έρευνα και βελτιώσεις, που θα καταστήσουν τη μεθοδολογία πιο αποδοτική.

ΚΕΦΑΛΑΙΟ 2

Επισκόπηση Τεχνολογικών Επιλογών για την Εξόρυξη Δεδομένων

2.1 Εισαγωγή

Υπάρχουν πολλοί ορισμοί για την έννοια της εξόρυξης δεδομένων. Στα πλαίσια αυτής της εργασίας, εξόρυξη δεδομένων θεωρείται η εφαρμογή αλγορίθμων και στατιστικών μεθόδων σε πραγματικά δεδομένα. Υπάρχουν διάφοροι κλάδοι εξόρυξης δεδομένων στην επιστήμη, στη μηχανική, στα οικονομικά και οπουδήποτε αλλού, όπου μία παρόμοια διαδικασία θα μπορούσε να φανεί χρήσιμη.[18]

Η εξόρυξη δεδομένων, αποτελείται από μια ομάδα τεχνικών, που έχουν σκοπό την ανάκτηση πληροφοριών από τα δεδομένα. Για την εκτέλεση διαδικασιών εξόρυξης δεδομένων υπάρχουν πολλές εφαρμογές, κάποιες από τις οποίες είναι: RapidMiner, Weka, R, orange, SAS και SPSS.

Η συγκεκριμένη εργασία, ασχολείται μόνο με το RapidMiner. Στα υπόλοιπα εργαλεία θα γίνει μία απλή αναφορά, με σκοπό την σύγκρισή τους με το Rapidminer.

2.2 Σημασία της Εξόρυξης Δεδομένων

Η εξόρυξη δεδομένων αποτελεί μία βασική διαδικασία εξόρυξης γνώσης, από μεγάλες βάσεις δεδομένων. Αναζητά συσχετίσεις, κανόνες, πρότυπα και πληροφορίες από ένα μεγάλο όγκο δεδομένων. Με την ανάκτηση αυτής της πληροφορίας, ο άνθρωπος, έχει στα χέρια του, όλα τα απαραίτητα εργαλεία προκειμένου να κάνει μία πρόβλεψη για μία μελλοντική κατάσταση. Η σημασία της εξόρυξης δεδομένων μπορεί να προσδιοριστεί από την ευρεία εφαρμογή, που βρίσκουν τα αποτελέσματα της, όπως η λήψη αποφάσεων. Ένας άλλος τομέας, που έχει απασχολήσει ιδιαίτερα τα ενεργειακά συστήματα, είναι η δυνατότητα δημιουργίας ενός προφίλ παραγωγής, ιδιαίτερα για τα τις ανανεώσιμες πηγές ενέργειας, όπου η πρόβλεψη είναι πολύ δύσκολη. Μέσω της εξόρυξης δεδομένων είναι δυνατή η εξαγωγή πληροφοριών, σχετικά με την παραγωγή ενός φωτοβολταϊκού συστήματος, καθώς επίσης και οι συσχετίσεις της παραγωγής με τις κλιματολογικές συνθήκες. Με αυτό τον τρόπο, οδηγούμαστε στην ακριβή πρόβλεψη μελλοντικής παραγωγής. Είναι σημαντικό να αναφερθεί πως, στις μέρες μας, έχουν κάνει την εμφάνισή τους συστήματα, που προσφέρουν ενεργειακές υπηρεσίες, μέσω διαδικτύου και στηρίζονται στην επεξεργασία ενεργειακών δεδομένων πραγματικού χρόνου.

2.3 Μοντέλα και Μέθοδοι Εξόρυξης Δεδομένων

Η εξόρυξη δεδομένων χωρίζεται σε κατηγορίες, ανάλογα με τον τρόπο που αναλύονται τα δεδομένα. Η εργασία επικεντρώνεται στην ανάλυση της πρόβλεψης, η οποία χρησιμοποιεί μεθόδους και τεχνικές προβλέψεων, για να προβλεφθεί ένα αποτέλεσμα. Περιλαμβάνει την ανάλυση μεμονωμένων περιπτώσεων, με σκοπό την πρόβλεψη παρόμοιων χαρακτηριστικών τους, η οποία δεν μπορεί να παρατηρηθεί απευθείας.

Εκτός από την ανάλυση της πρόβλεψης, μία άλλη εναλλακτική, είναι η περιγραφική ανάλυση, η οποία έχει σαν στόχο την ανακάλυψη διατάξεων στα δεδομένα. Η προγνωστική, μαζί με την περιγραφική ανάλυση, ονομάζονται «ανακάλυψη γνώσης από δεδομένα» (knowledge discovery in data, KDD). Η ανακάλυψη διατάξεων στα δεδομένα μπορεί να είναι εξαιρετικά χρήσιμη, αλλά συνήθως είναι δυσκολότερο να ληφθεί άμεσο όφελος από την περιγραφική ανάλυση, σε σχέση με την προγνωστική. Αυτό συμβαίνει διότι, οι προβλέψεις μπορούν να χρησιμοποιηθούν άμεσα για τη λήψη αποφάσεων, οι οποίες θα μεγιστοποιήσουν το όφελος του υπεύθυνου λήψης της απόφασης. Είναι πολύ σημαντικό να διασαφηνιστεί η διαφορά ανάμεσα σε μία πρόβλεψη και μία απόφαση. Η εξόρυξη δεδομένων επιτρέπει την πρόβλεψη, αλλά οι προβλέψεις είναι χρήσιμες, μόνο στην περίπτωση που επιτρέπουν τη λήψη αποφάσεων, οι οποίες θα έχουν καλύτερα αποτελέσματα, θα μεγιστοποιήσουν το κέρδος, όπου μεγιστοποίηση κέρδους γενικά σημαίνει μεγιστοποίηση αποδοτικότητας .

Οι αλγόριθμοι εξόρυξης δεδομένων απαρτίζονται από τα παρακάτω στοιχεία: [26]

1. Μέθοδος αναζήτησης

   Με τη σειρά της, η μέθοδος αναζήτησης χωρίζεται σε δύο κατηγορίες:
   - στην αναζήτηση παραμέτρων, για την εύρεση ελεύθερων παραμέτρων ώστε να βελτιστοποιηθεί το μοντέλο και
   -στην αναζήτηση μοντέλου, για την εύρεση κατάλληλου, για κάθε μία τέτοια δομή, που εφαρμόζει τις κατάλληλες παραμέτρους.

2. Παράσταση μοντέλου

   Για την δημιουργία ενός μοντέλου, μέσω της εξόρυξης δεδομένων, απαιτείται εκτενής αναπαράσταση, σε συνδυασμό με παραδείγματα και επαρκή χρόνο εκπαίδευσης. Αν οι παραπάνω προϋποθέσεις δεν πληρούνται τότε το μοντέλο, που θα παραχθεί, δεν θα είναι ακριβές.

3. Αξιολόγηση μοντέλου

   Η αξιολόγηση του μοντέλου γίνεται βάση ορισμένων κριτηρίων, τα οποία πρέπει να πληρούν τις προδιαγραφές της διαδικασίας KDD. Για παράδειγμα, τα μοντέλα

πρόβλεψης συχνά αξιολογούνται από την εμπειρική ακρίβεια πρόβλεψης σε κάποιο σύνολο ελέγχου. [32]

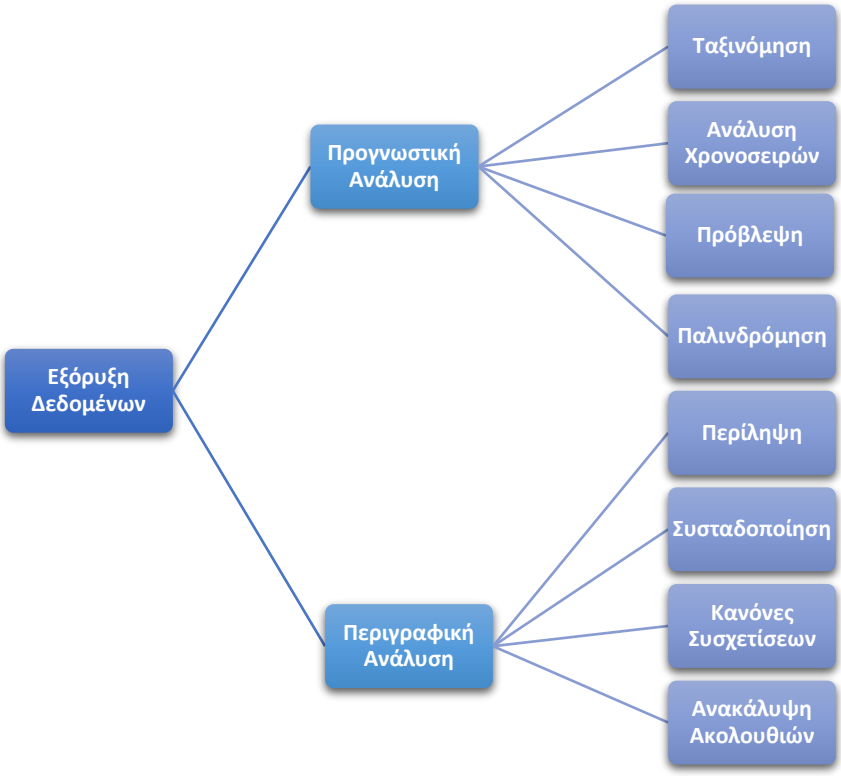
Η εξόρυξη δεδομένων, υποδιαιρείται σε περαιτέρω μεθόδους data mining, ανάλογα με την ανάλυση που εφαρμόζεται.

Η προβλεπτική ανάλυση, απαρτίζεται από τις εξής μεθόδους:

- Ταξινόμηση (Classification)
- Ανάλυση Χρονοειρών (Time Series Analysis)
- Πρόβλεψη (Prediction)
- Παλινδρόμηση (Regression)

Ενώ η περιγραφική ανάλυση, περιέχει τις ακόλουθες:

- Περίληψη (Summarization)
- Συσταδοποίηση (Clustering)
- Κανόνες Συσχετίσεων (Association Rules)
- Ανακάλυψη Ακολουθιών (Sequential Pattern Discovery)



Σχήμα 2 Ταξινόμηση των Τεχνικών Εξόρυξης Δεδομένων

- Ταξινόμηση

Η ταξινόμηση αποτελεί μια μέθοδο εξόρυξης δεδομένων, βασιζόμενη στην εκμάθηση μηχανής (machine learning). Χρησιμοποιείται για να κατατάξει τα δεδομένα σε προκαθορισμένες κλάσεις. Συνήθως, χρησιμοποιεί μαθηματικές μεθόδους, όπως δένδρα αποφάσεων, νευρωνικά δίκτυα, Naïve Bayes μοντέλα και μηχανές διανυσμάτων υποστήριξης (SVM).

- Ανάλυση Χρονοσειρών

Η ανάλυση χρονοσειρών στηρίζεται στην υπόθεση ότι, η τιμή ενός μεγέθους ακολουθεί ένα πρότυπο, που επαναλαμβάνεται στο χρόνο. Η συγκεκριμένη ανάλυση είναι αδύνατη χωρίς την ύπαρξη ιστορικών δεδομένων προηγούμενων και σταθερών χρονικών περιόδων. Τα απαραίτητα βήματα, για τη συγκεκριμένη ανάλυση, είναι η εξέταση της δομής μιας χρονοσειράς, η εύρεση ομοιοτήτων και η χρήση διαγραμμάτων χρονοσειρών, με στόχο την πρόβλεψη μελλοντικών τιμών.

- Πρόβλεψη

Η πρόβλεψη είναι η μέθοδος, η οποία ανακαλύπτει συσχετίσεις μεταξύ ανεξάρτητων μεταβλητών και συσχετίσεις ανάμεσα σε ανεξάρτητες και εξαρτημένες μεταβλητές. Η πρόβλεψη συνδέεται άμεσα με την ταξινόμηση, αφού οποιαδήποτε τεχνική της ταξινόμησης μπορεί να προσαρμοστεί και στην πρόβλεψη. Αυτό συμβαίνει διότι, στην πρόβλεψη, τα ιστορικά δεδομένα, χρησιμοποιούνται ως παραδείγματα εκπαίδευσης και η τιμή της μεταβλητής που προβλέπεται είναι ήδη γνωστή. Στη συνέχεια, με τη χρήση των ιστορικών δεδομένων, δημιουργείται ένα μοντέλο πρόβλεψης, το οποίο φανερώνει την τάση της προβλεπόμενης μεταβλητής. Τέλος, το μοντέλο εφαρμόζεται στα καινούρια δεδομένα και προκύπτει μία πρόβλεψη.

- Παλινδρόμηση

Η παλινδρόμηση, σαν μέθοδος, έχει σκοπό τον προσδιορισμό συσχετίσεων ανάμεσα σε μία εξαρτημένη μεταβλητή και μία ή περισσότερες ανεξάρτητες, δηλαδή την ανάλυση μιας συνάρτησης $y = f(x)$. Κάποιες, αρκετά συνηθισμένες, μέθοδοι παλινδρόμησης είναι η λογιστική παλινδρόμηση, τα δένδρα παλινδρόμησης και τα νευρωνικά δίκτυα. [26]

Στη συγκεκριμένη εργασία, η μέθοδος παλινδρόμησης, που θα χρησιμοποιηθεί, είναι η γραμμική παλινδρόμηση, η οποία θα αναλυθεί στο επόμενο κεφάλαιο.

- Περίληψη

Η περίληψη περιγράφει τα δεδομένα, μέσω κάποιων χαρακτηριστικών και αντιπροσωπευτικών πληροφοριών τους. Με τη βοήθεια κάποιων στατιστικών μεθόδων, όπως είναι το ιστόγραμμα, το διάγραμμα διασποράς, η τυπική απόκλιση και η διακύμανση, γίνεται πιο εύκολη η κατανόηση των γνωρισμάτων των δεδομένων.

- Συσταδοποίηση

Η μέθοδος της συσταδοποίησης είναι η διαδικασία ομαδοποίησης ενός συνόλου στοιχείων, με τέτοιο τρόπο, ώστε τα αντικείμενα που ανήκουν στην ίδια ομάδα (συστάδα) να παρουσιάζουν περισσότερες ομοιότητες μεταξύ τους παρά με αντικείμενα άλλης ομάδας. Οι συστάδες μπορεί να είναι αμοιβαία αποκλειόμενες ή επικαλυπτόμενες. [28]

- Κανόνες Συσχετίσεων

Οι κανόνες συσχετίσεων αποτελούν μία μέθοδο ανακάλυψης συσχετίσεων, μεταξύ μεταβλητών, σε μεγάλες βάσεις δεδομένων. Έχει σκοπό την αναζήτηση ισχυρών προτύπων, μέσα στις βάσεις δεδομένων. Στη συγκεκριμένη μέθοδο στηρίχτηκε και η ανάλυση του καλαθιού αγοράς (market basket analysis), σύμφωνα με την οποία, αναλύοντας τα προϊόντα που περιέχει ένα καλάθι του σουπερμάρκετ, είναι δυνατό να προβλεφθεί κάθε επιπλέον αγορά του καταναλωτή.

- Ανακάλυψη ακολουθιών

Η μέθοδος της ανακάλυψης ακολουθιών, έχει σαν στόχο, τον καθορισμό προτύπων στα δεδομένα, η συσχέτιση των οποίων βασίζεται στο χρόνο. [26] Η ακολουθιακή ανάλυση είναι δυνατή διότι, οι χρονοσειρές και οι ακολουθίες στηρίζονται σε χρονικά συνεχόμενα δεδομένα, τα οποία, με τη σειρά τους, βασίζονται σε εξαρτημένες μεταξύ τους παρατηρήσεις.

2.4 Το εργαλείο Rapidminer

Για την υλοποίηση ενός μοντέλου πρόβλεψης σε ένα σύστημα ηλεκτρικής ενέργειας απαιτείται η αξιοποίηση της διαθέσιμης γνώσης. Στην περίπτωση ενός ΣΗΕ, η διαθέσιμη γνώση είναι τα δεδομένα και συγκεκριμένα τα ιστορικά δεδομένα. Ανάλογα με τα διαθέσιμα μέσα, είναι δυνατό η διαδικασία παραγωγής προβλέψεων να γίνει αυτόματα, μέσα από εξειδικευμένα λογισμικά μέσα. Ένα τέτοιο μέσο είναι και το Rapidminer.

2.4.1 Ιστορική πορεία του Rapidminer

Το RapidMiner αποτελεί ένα ισχυρό οπτικό εργαλείο για την εξόρυξη δεδομένων, την εκμάθηση μηχανής (learning machine) και την πρόβλεψη αναλύσεων.

Ήταν αρχικά γνωστό ως YALE (Yet Another Learning Environment) και αναπτύχθηκε το 2001 από τους Ralf Klinkenberg, Ingo Mierswa, και Simon Fischer στο Artificial Intelligence Unit του Πολυτεχνείου του Dortmund. Στόχος τους ήταν η δημιουργία ενός εργαλείου εξόρυξης δεδομένων, το οποίο θα ήταν πιο ευέλικτο και πολύ πιο ισχυρό από τα εργαλεία που είχαν διατεθεί στην αγορά, μέχρι τότε. Καθώς ξεκίνησε ως λογισμικό

ανοιχτού κώδικα, με ελεύθερη πρόσβαση σε κάθε ενδιαφερόμενο, το YALE προσέγγισε γρήγορα πολλούς χρήστες. Μέχρι το 2006, το ενδιαφέρον για τις συμβουλευτικές υπηρεσίες και την εκπαίδευση που προσέφερε το YALE είχε αυξηθεί κατακόρυφα. Αυτό οδήγησε τους Ralf Klinkenberg και Ingo Mierswa να ιδρύσουν μία δική τους εταιρεία, την Rapid-I, η οποία παρείχε επαγγελματική υποστήριξη, εκπαίδευση, συμβουλευτικές υπηρεσίες, λύσεις και ακόμα περισσότερες υπηρεσίες στους τομείς των προβλεπτικών αναλύσεων και της εξόρυξης δεδομένων και κειμένου, για τους χρήστες του λογισμικού. Προκειμένου να ανταπεξέλθουν καλύτερα στις απαιτήσεις παγκόσμιων εταιρειών, με μεγάλες ποσότητες δεδομένων, αναγκάστηκαν να επαναπροσδιορίσουν από την αρχή το YALE και να το μετονομάσουν σε RapidMiner. Με αυτό το όνομα περιγράφονταν καλύτερα τα ιδιαίτερα χαρακτηριστικά του λογισμικού, δηλαδή η εφαρμογή ανάπτυξης ταχείας εξόρυξης δεδομένων.[4]

2.4.2 Χαρακτηριστικά του Rapidminer

Το RapidMiner, το οποίο αποτελεί το κυρίαρχο προϊόν της Rapid-I, θεωρείται ηγετικό σύστημα ανοικτού κώδικα παγκόσμιας εμβέλειας, για επαγγελματική εξόρυξη δεδομένων. Καλύπτει όλα τα στάδια της διαδικασίας εξόρυξης δεδομένων, από την μεταφορά και μετατροπή δεδομένων, ως την περιγραφική και προβλεπτική μοντελοποίηση, ανάπτυξη αξιολόγηση μοντέλου. Τα βασικά χαρακτηριστικά του είναι η τεράστια ευελιξία και το λειτουργικό εύρος, το οποίο υποστηρίζει όλα τα είδη εξόρυξης δεδομένων, εξόρυξης κειμένου, εξόρυξης ήχου, ανάλυσης χρονοσειρών, καθώς και προβλεπτικών και προβλεπτικών αναλύσεων. Η συγκεκριμένη ευελιξία και η ταχεία ανάπτυξη, οδηγούν σε πολύ γρήγορες υλοποιήσεις έργων. Η υλοποίησή του έγινε σε γλώσσα προγραμματισμού Java, αλλά το περιβάλλον του Rapidminer δεν απαιτεί τη γραφή κώδικα από τον χρήστη.

2.4.3 Σύγκριση

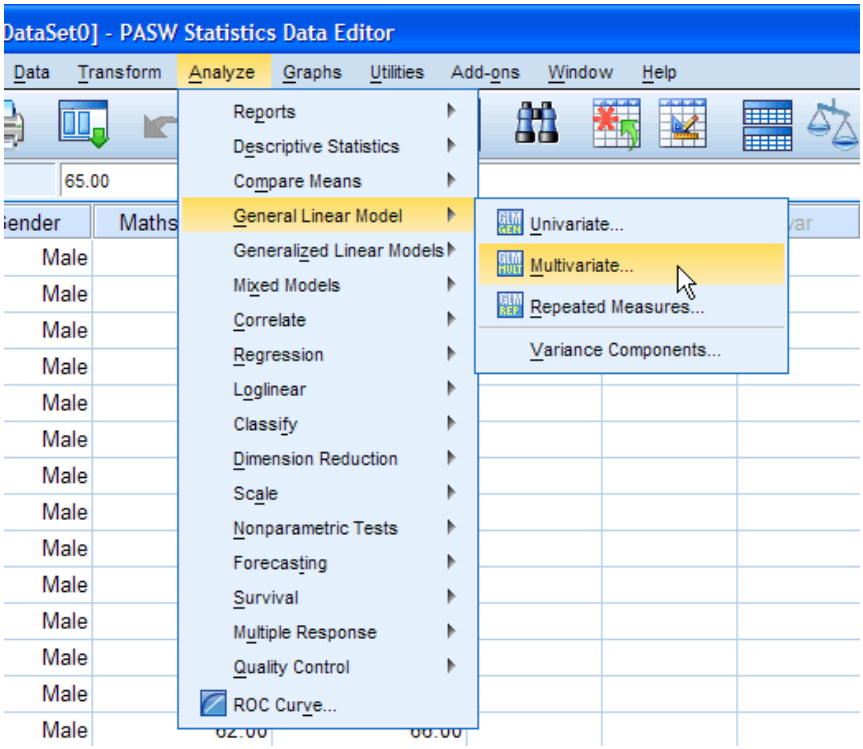
Παρακάτω, αναλύονται μερικές λειτουργίες που διαθέτει το RapidMiner.

Αρχικά, έχει τη δυνατότητα να φορτώσει ολόκληρο το σύνολο των δεδομένων στη μνήμη - εφόσον το μέγεθος της μνήμης το επιτρέπει – και να εκτελέσει χρονικά αποτελεσματική προ-επεξεργασία (preprocessing) και εξόρυξη στην μνήμη, για παράδειγμα: CSVExampleSource, DatabaseExampleSource και CSVReader, DatabaseReader. Εναλλακτικά, μπορεί να διαβάσει τα δεδομένα σε κομμάτια. Δηλαδή, να διαβάσει τη βάση δεδομένων, γραμμή προς γραμμή ή ακόμα και αρχείο προς αρχείο και επομένως να δουλέψει στη βάση δεδομένων ή σε μεγάλη συλλογή εγγράφων και αρχείων. Για παράδειγμα: CachedDataBaseExampleSource, FiIterator.[3]

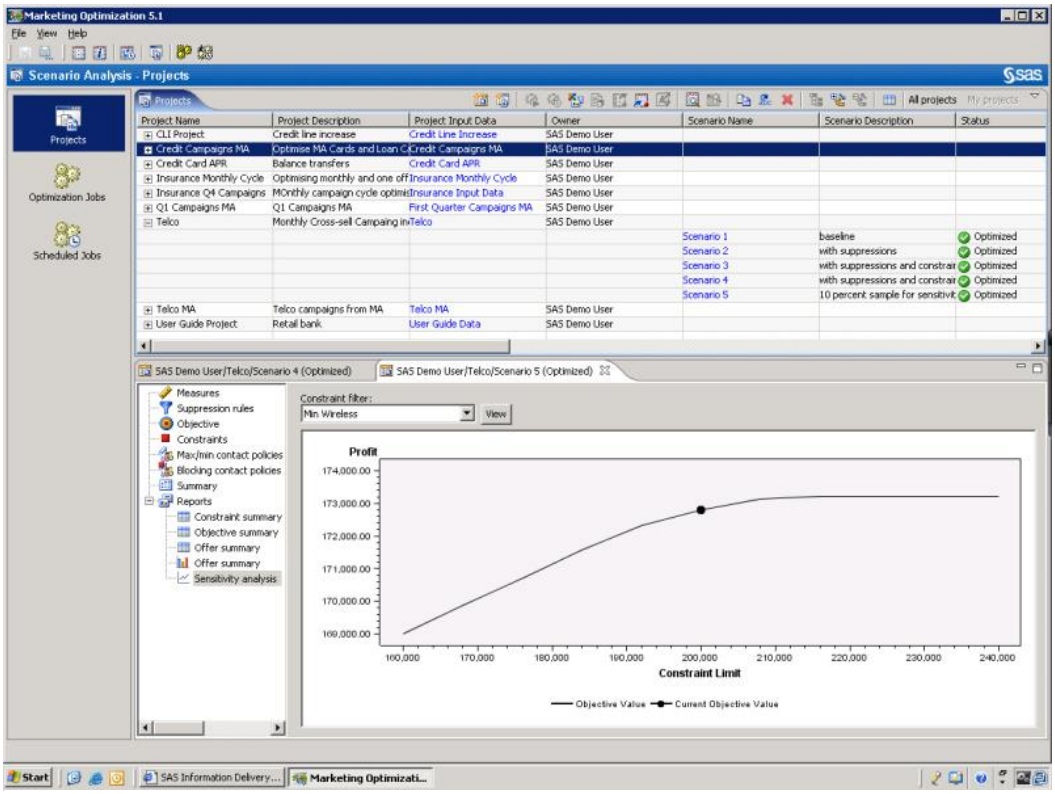
2.4.3.1 *Σύγκριση Rapidminer με εμπορικά προϊόντα*

Εκτός από τα λειτουργικά του χαρακτηριστικά, το RapidMiner, συγκριτικά με τα εμπορικά προϊόντα, όπως είναι το SAS και το SPSS, υπερτερεί καθώς είναι εντελώς δωρεάν και ανοιχτού κώδικα (open source). Πέρα από τις πτυχές των επιχειρήσεων, το RapidMiner έχει και άλλα πλεονεκτήματα, όπως η τεράστια ευελιξία στη διαδικασία σχεδιασμού. Επομένως, δίνει τη δυνατότητα στους χρήστες του, να γνωρίζουν, οποιαδήποτε στιγμή, τις τροποποιήσεις που επιδέχονται τα δεδομένα και να αναζητούν ελεύθερα λύσεις σε δύσκολα προβλήματα που αντιμετωπίζουν, κατά τη διάρκεια εκτέλεσης του προγράμματος. Λύσεις είναι εύκολο να βρεθούν μέσω του ανοιχτού φόρουμ της Rapid-I και μέσω του οδηγού χρήσης του RapidMiner.[2]

Επίσης, καθώς αποτελεί προϊόν πανεπιστημίου και βασίζεται σε ερευνητικό μοντέλο, παρέχει σε καθηγητές, μαθητές και ερευνητές ένα ξεχωριστό προϊόν, προκειμένου να καλυφθούν οι απαιτητικές τους ανάγκες.



Σχήμα 3 Λογισμικό SPSS [13]



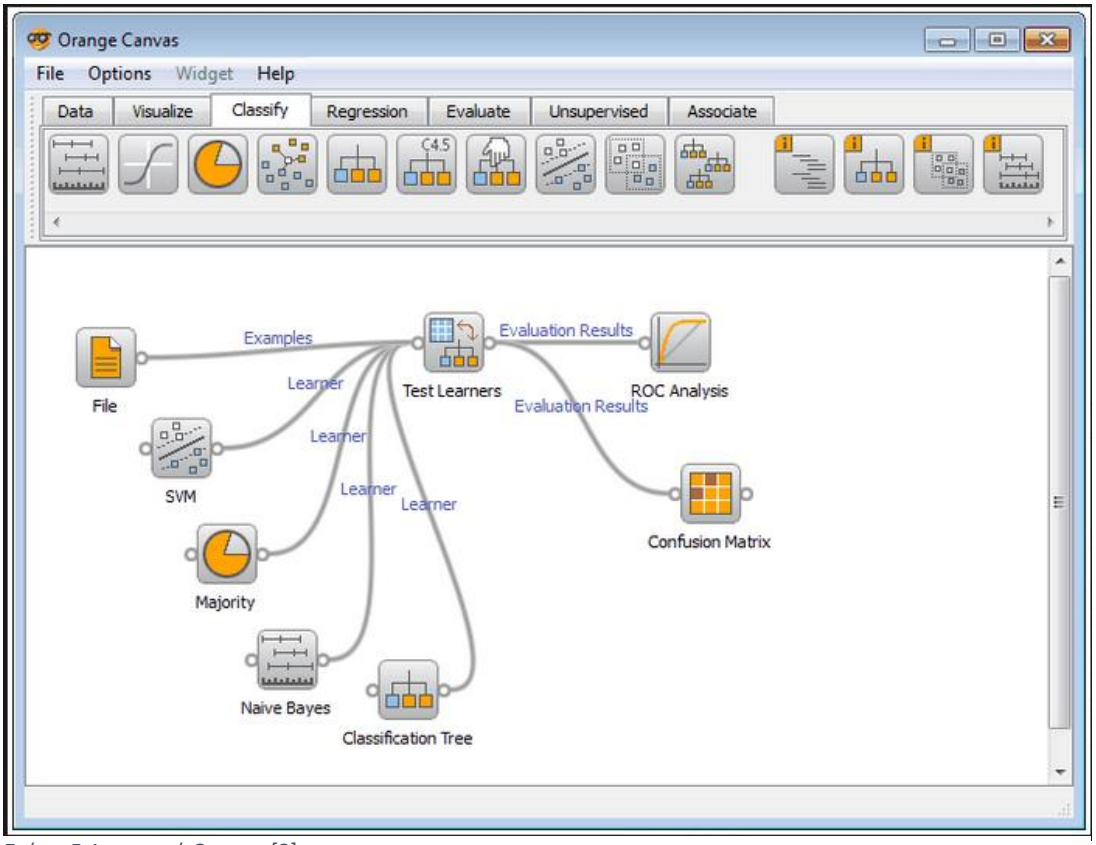
Σχήμα 4 Λογισμικό SAS [15]

2.4.3.2 ***Σύγκριση Rapidminer με δωρεάν προϊόντα***

Περισσότερο ενδιαφέρον όμως παρουσιάζει η σύγκριση του RapidMiner με τα υπόλοιπα δωρεάν εργαλεία που κυκλοφορούν, όπως το orange, το R και το Weka.

Orange:

Το orange αποτελεί λογισμικό ανοιχτού κώδικα για την εξόρυξη, οπτικοποίηση και ανάλυση δεδομένων, για αρχάριους και έμπειρους χρήστες. Περιλαμβάνει μια μεγάλη εργαλειοθήκη, με διαθέσιμο σύνολο στοιχείων για την προ-επεξεργασία δεδομένων, τη δυνατότητα βαθμολόγησης και φιλτραρίσματος, τη μοντελοποίηση, την αξιολόγηση μοντέλων και άλλες τεχνικές εξερεύνησης. Το πρόγραμμα υλοποιείται σε C++ και Python και είναι διατίθεται σε όλες τις δημοφιλείς πλατφόρμες (Linux, Mac, Windows). Επικεντρώνεται σε παραδοσιακούς συμβολικούς αλγορίθμους, αντί για αριθμητικούς. Ιδρύθηκε το 1996, στο Πανεπιστήμιο της Λιουμπλιάνα και στο Ινστιτούτο Jozef Stefan.[7]



Σχήμα 5 Λογισμικό Orange [8]

R:

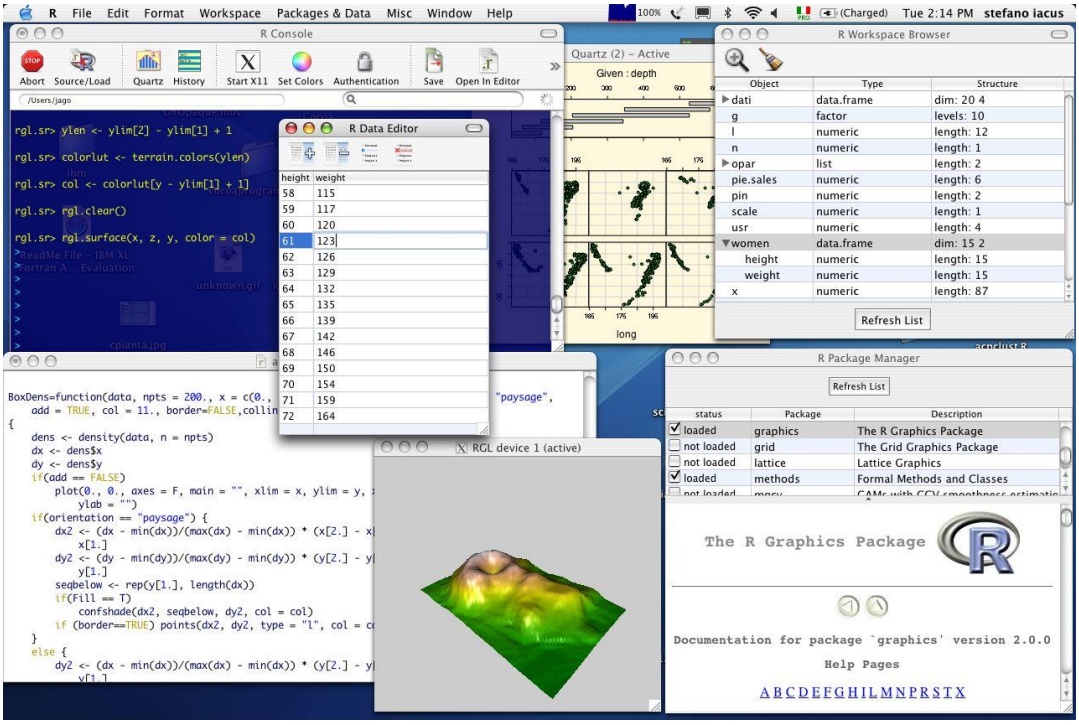
[9],[10]Πρόκειται για λογισμικό ανοιχτού κώδικα, διαθέσιμο για όλες τις πλατφόρμες. Ιδρύθηκε από τους Ross Ithaka και Robert Gentleman, στο πανεπιστήμιο του Άουκλαντ (Auckland), στη Νέα Ζηλανδία.

R είναι μια γλώσσα προγραμματισμού και ένα περιβάλλον λογισμικού, για στατιστικούς υπολογισμούς και γραφικά. Πρόκειται για ένα πλήρως σχεδιασμένο και συνεκτικό σύστημα, μέσα στο οποίο εφαρμόζονται στατιστικές τεχνικές. Η γλώσσα R χρησιμοποιείται ευρέως από στατιστικολόγους για την ανάπτυξη του στατιστικού λογισμικού και την ανάλυση δεδομένων.

Παρέχει ένα ευρύ φάσμα στατιστικών (γραμμικών και μη γραμμικών μοντέλων, κλασικούς στατιστικούς ελέγχους, ανάλυση χρονοσειρών, ταξινόμηση, ομαδοποίηση κ.α) και γραφικών τεχνικών και είναι εξαιρετικά επεκτάσιμη. Η R μπορεί να διαχειριστεί οπτικοποίηση και ανάλυση δεδομένων ως 16 TB. Ένα από τα δυνατά σημεία της R είναι η ευκολία, παραγωγής καλά σχεδιασμένων σεναρίων, συμπεριλαμβανομένων των μαθηματικών συμβόλων και τύπων, όπου χρειάζεται. Για παράδειγμα, έμπειροι χρήστες μπορούν να γράψουν κώδικα σε γλώσσα C, για την άμεση διαχείριση αντικειμένων της R. Οι καθορισμένες επιλογές, στο δευτερεύοντα σχεδιασμό των γραφικών, επιτρέπουν στο χρήστη να συνεχίζει να διατηρεί πλήρη έλεγχο.

Η R είναι μια ολοκληρωμένη συλλογή λογισμικών λειτουργιών, για το χειρισμό των δεδομένων, τον υπολογισμό και τη γραφική τους απεικόνιση. Περιλαμβάνει:

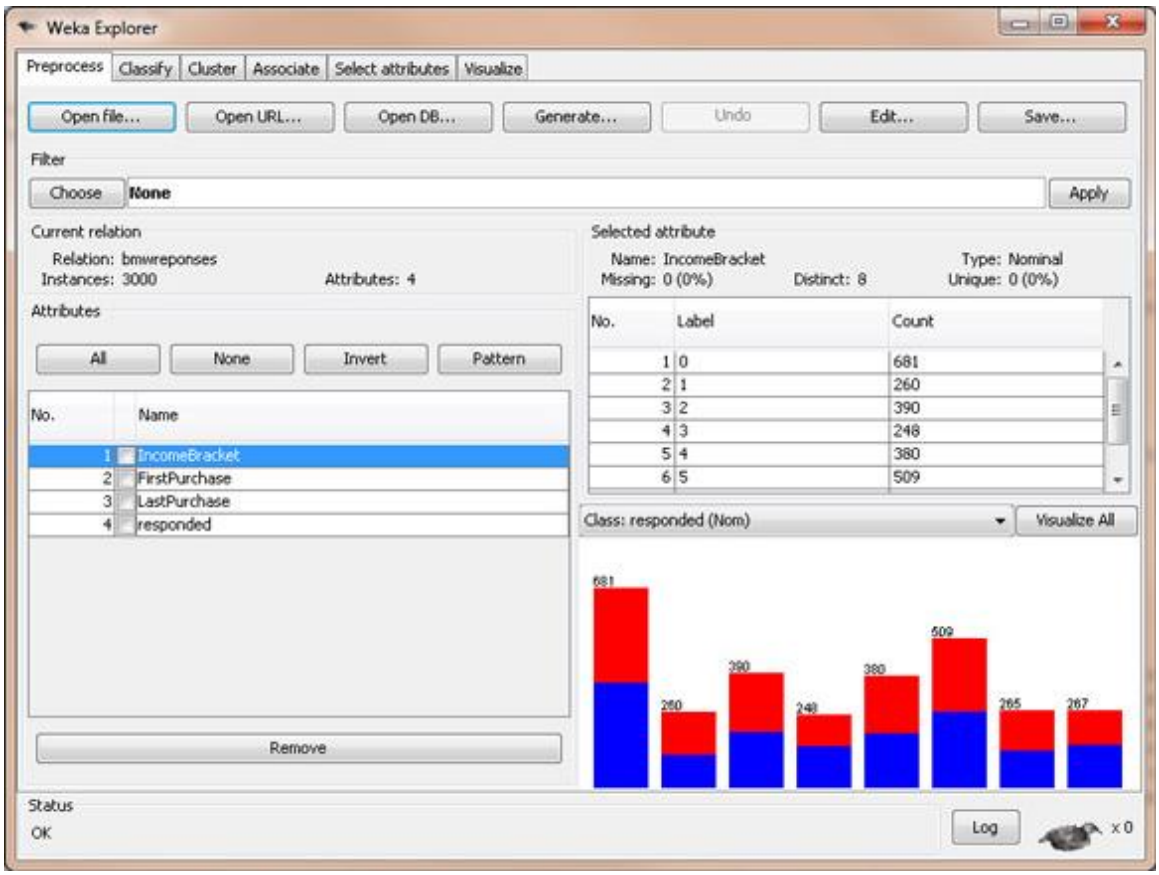
- Έναν αποτελεσματικό χειρισμό των δεδομένων και εγκαταστάσεις αποθήκευσης
- Μια συλλογή τελεστών, για υπολογισμούς σε πίνακες, για την εξόρυξη κειμένου και γραφικών
- Μια μεγάλη, συνεκτική και ολοκληρωμένη συλλογή των ενδιάμεσων εργαλείων, για την ανάλυση δεδομένων
- Γραφικές λειτουργίες, για την ανάλυση δεδομένων και την απεικόνιση, είτε στην οθόνη είτε σε έντυπη μορφή
- Μία καλά ανεπτυγμένη, απλή και αποτελεσματική γλώσσα προγραμματισμού, που περιλαμβάνει βρόχους, συναρτήσεις, καθορισμένους από το χρήστη και λειτουργίες εισόδου και εξόδου δεδομένων.



Σχήμα 6 Λογισμικό R [11]

Weka:

Το Weka, άρχισε να αναπτύσσεται το 1993, στο πανεπιστήμιο Waikato της Νέας Ζηλανδίας. Σήμερα, υποστηρίζει αρκετές τυπικές εργασίες εξόρυξης δεδομένων, πιο συγκεκριμένα, προ-επεξεργασία δεδομένων, ομαδοποίηση, ταξινόμηση, οπτικοποίηση και επιλογή χαρακτηριστικών. Όλες οι τεχνικές στηρίζονται στην υπόθεση ότι, τα δεδομένα είναι διαθέσιμα, ως ένα ενιαίο αρχείο ή μία σχέση, όπου κάθε στοιχείο δεδομένου περιγράφεται από ένα σταθερό αριθμό μεταβλητών. Μερικά χαρακτηριστικά του Weka είναι, η ελεύθερη διαθεσιμότητα, η φορητότητα, καθώς είναι συμβατό σχεδόν με κάθε σύγχρονη πλατφόρμα πληροφορικής και η εύκολη χρήση, χάρη στο γραφικό περιβάλλον και στις κατανοητές τεχνικές προ-επεξεργασίας και μοντελοποίησης δεδομένων που διαθέτει. Οι αλγόριθμοι που χρησιμοποιεί μπορούν, είτε να εφαρμοστούν απευθείας στο σετ δεδομένων είτε να κληθούν από δικό μας κώδικα Java.[5]



Σχήμα 7 Λογισμικό Weka [12]

Λόγω της μεγάλης ομοιότητας του Rapidminer με τα Weka, απαιτείται περαιτέρω ανάλυση ανάμεσα στις διαφορές τους.

Συγκρίνοντας αυτά τα δύο εργαλεία, επισημαίνεται ότι επικρατεί η λανθασμένη άποψη ότι το Rapidminer είναι μια εκδοχή του Weka, καθόσον είναι δυνατή η λειτουργία του

Rapidminer και μετά τη διαγραφή του αρχείου weka.jar. Επιπλέον, κατά την μελέτη των τελεστών του Rapidminer, παρατηρείται ότι, μόνο οι 100 προέρχονται από το Weka, οι υπόλοιποι 400, οι οποίοι δεν είναι διαθέσιμοι στο Weka, διαθέτουν πολλές πηγές δεδομένων, μεθόδους προ-επεξεργασίας, καθώς και τεχνικές επικύρωσης και οπτικοποίησης. Το Rapidminer, πέρα από διαφορετικό περιβάλλον εργασίας για το χρήστη και πληθώρα τελεστών, εμφανίζει και κάποιες επιπλέον διαφορές σε σχέση με το Weka.[6],[2]

- Δύναμη και ευελιξία: αν και το Weka είναι ένα εργαλείο, πολύ εύκολο στη χρήση του, δεν είναι αρκετά ευέλικτο και ισχυρό, ώστε να ανταπεξέλθει στις απαιτήσεις πολύπλοκων σύγχρονων διαδικασιών. Από την άλλη πλευρά, το Rapidminer παρέχει περισσότερα βήματα για την ανάλυση δεδομένων – με τη βοήθεια της πληθώρας τελεστών, όπως αναφέρεται παραπάνω – και περισσότερες δυνατότητες συνδυασμών.

- Ευχρηστία: η ροή δεδομένων στο Rapidminer είναι πάντα ίδια με μια δομή δένδρου με βάση, σε αντίθεση με μία διάταξη γράφημα, όπως είναι το Knowledge Flow του Weka. Επίσης, το Rapidminer μπορεί να εξασφαλίσει αυτόματες επικυρώσεις και αυτόματες διαδικασίες βελτιστοποίησης, για την εξόρυξη δεδομένων μεγάλης κλίμακας, σε αντίθεση με τα διαδραστικά γραφήματα. Επιπλέον, παρέχονται μακριά ονόματα παραμέτρων, τα οποία είναι πιο κατανοητά και βοηθούν στη γραφική διασύνδεση του χρήστη και στον ορισμό των βημάτων και των παραμέτρων. Η διάταξη με δομή δένδρου με βάση, μαζί με μια ισχυρότερη έννοια του συναρμολογούμενου(modular) (ένα ωραίο παράδειγμα είναι ο τελεστής cross validation στο Rapidminer, ο οποίος επιτρέπει την αυθαίρετη επεξεργασία αποτελεσμάτων), επιτρέπει σημεία διακοπής και τον εύκολο ορισμό των επαναχρησιμοποιημένων δομικών στοιχείων.

- Απόδοση: Το Rapidminer μπορεί να διαχειριστεί μεγαλύτερα σύνολα δεδομένων από το Weka. Όμως, η κατανάλωση της μνήμης συνεχίζει να κυμαίνεται σε χαμηλότερα επίπεδα, καθιστώντας την εσωτερική αναπαράσταση δεδομένων πιο αποτελεσματική, σε συνδυασμό με πολυεπίπεδη προβολή δεδομένων.

- Επιδεκτικότητα: η εσωτερική διαχείριση των δεδομένων του Rapidminer, επιτρέπει την εφαρμογή μεγάλης ποσότητας μεθόδων εξόρυξης και εκμάθησης δεδομένων, απευθείας, σε μία εξωτερική βάση δεδομένων. Με αυτό τον τρόπο επιτυγχάνεται η εφαρμογή γραμμικών μεθόδων μάθησης - όπως είναι το Perceptron – απευθείας σε μία βάση δεδομένων, χωρίς να φορτωθούν τα δεδομένα στη μνήμη. Αυτό επιτρέπει εξόρυξη δεδομένων μεγάλης κλίμακας.

- Ενσωμάτωση: η εργαλειοθήκη του Weka μπορεί να ενσωματωθεί σε διαφορετικά προϊόντα λογισμικού, ωστόσο, αν ενωθούν οι διάφορες διαδικασίες εξόρυξης δεδομένων στο ίδιο προϊόν, θα πρέπει να μετατραπούν εκ νέου τα δεδομένα ξανά

και ξανά (με όλα τα μειονεκτήματα που συνεπάγεται αυτό, όπως αναφέρθηκαν παραπάνω). Αντίθετα, στο Rapidminer τα δεδομένα εμφανίζονται σε στρώσεις και έτσι επιτρέπεται η ενσωμάτωση των διαφόρων κλάδων της ανάλυσης σε ένα ενιαίο προϊόν, χωρίς να αντιγράφονται και να μετατρέπονται εκ νέου τα δεδομένα κάθε φορά. Αυτό ενισχύει την ενσωμάτωση, ακόμα και αν ο κώδικας είναι λίγο πιο πολύπλοκος.

- Προ-επεξεργασία: υπάρχουν πάρα πολλές μέθοδοι για την προ-επεξεργασία, την εξόρυξη και τον μετασχηματισμό δεδομένων, διαθέσιμες στο Rapidminer. Πολύ περισσότερες από αυτές που υπάρχουν στο Weka, για την ανάλυση δεδομένων. Όλες οι φάσεις μίας ανάλυσης ενσωματώνονται σε μία διαδικασία/εργαλείο, διευκολύνοντας κατά πολύ την ανάλυση.

Συνολική Σύγκριση:

Πίνακας 1 Πίνακας σύγκρισης εργαλείων εξόρυξης δεδομένων [15]

Διαδικασία	**R-programming**	**Weka**	**Orange**	**Rapidminer**
Διαχωρισμός του συνόλου δεδομένων σε σύνολα εκπαίδευσης και δοκιμής	✓ (Περιορισμένες μέθοδοι διαχωρισμού)	✓ (Περιορισμένες μέθοδοι διαχωρισμού)	✓(Περιορισμένες μέθοδοι διαχωρισμού)	✓ (Περιορισμένες μέθοδοι διαχωρισμού)
Περιγραφή κλιμάκωσης	✓	X (δεν μπορεί να αποθηκεύσει παραμέτρους κλιμάκωσης για να τις εφαρμόσει σε μελλοντικά σύνολα δεδομένων)	X (δεν υπάρχουν μέθοδοι κλιμάκωσης)	✓
Περιγραφή επιλογής	X (δεν υπάρχουν μέθοδοι)	✓ (δεν είναι μέρος του Knowledge Flow)	X (δεν υπάρχουν μέθοδοι)	✓
Βελτιστοποίηση παραμέτρων για την εκμάθηση μηχανής/στατιστικές μεθόδους	X (δεν είναι αυτόματο)	X (δεν είναι αυτόματο)	X (δεν είναι αυτόματο)	✓

Επιβεβαίωση μοντέλου με χρήση διασταυρωμένης επικύρωσης και/ή ανεξάρτητα σύνολα επικύρωσης	✓ (περιορισμένος αριθμός μεθόδων για την μέτρηση σφαλμάτων)	✓ (δεν μπορεί να αποθηκεύσει το μοντέλο γι'αυτό πρέπει να κατασκευάσουμε από την αρχή το μοντέλο για κάθε μελλοντικό σετ δεδομένων)	✓ (δεν μπορεί να αποθηκεύσει το μοντέλο γι'αυτό πρέπει να κατασκευάσουμε από την αρχή το μοντέλο για κάθε μελλοντικό σετ δεδομένων)	✓

Παρατηρείται ότι το Rapidminer ανταπεξέρχεται πολύ καλύτερα σε όλες τις διαδικασίες, από τα υπόλοιπα δωρεάν εργαλεία εξόρυξης δεδομένων.

2.5 Διακρίσεις Rapidminer

[16] Για όλους τους παραπάνω λόγους, από δημοσκόπηση που έγινε το 2014, στο KDnuggets – το οποίο ασχολείται με την κάλυψη ειδήσεων στον τομέα των Αναλύσεων για Επιχειρήσεις, της Εξόρυξης Δεδομένων και της Επιστήμης Δεδομένων, συμπεριλαμβανομένων συνεντεύξεων από πολλούς εμπειρογνώμονες στους συγκεκριμένους τομείς, το Rapidminer ανακηρύχθηκε το καλύτερο εργαλείο εξόρυξης δεδομένων. Η συγκεκριμένη δημοσκόπηση μετρούσε, πόσο ευρέως χρησιμοποιείται ένα εργαλείο εξόρυξης δεδομένων και πόσο έντονα οι προμηθευτές υποστηρίζουν το εργαλείο τους. Το 2014, το 71% των ψηφοφόρων χρησιμοποιούσε εμπορικό λογισμικό, ενώ το 78% χρησιμοποιούσε δωρεάν λογισμικά. Τα 10 εργαλεία που χρησιμοποιούνται περισσότερο από τους ψηφοφόρους, είναι τα εξής:

1. RapidMiner, 44.2%
2. R, 38.5%
3. Excel, 25.8%
4. SQL, 25.3%
5. Python, 19.5%
6. Weka, 17.0%
7. KNIME, 15.0%
8. Hadoop, 12.7%
9. SAS base, 10.9%
10. Microsoft SQL Server, 10.5%

Ο παρακάτω πίνακας, παρουσιάζει τα αποτελέσματα της δημοσκόπησης, σε φθίνουσα σειρά. Το «%alone» είναι το ποσοστό των ψηφοφόρων, οι οποίοι χρησιμοποιούν μόνο το

αντίστοιχο εργαλείο. Για παράδειγμα, το 35,1% των χρηστών του Rapidminer, χρησιμοποιούν αποκλειστικά αυτό το εργαλείο.

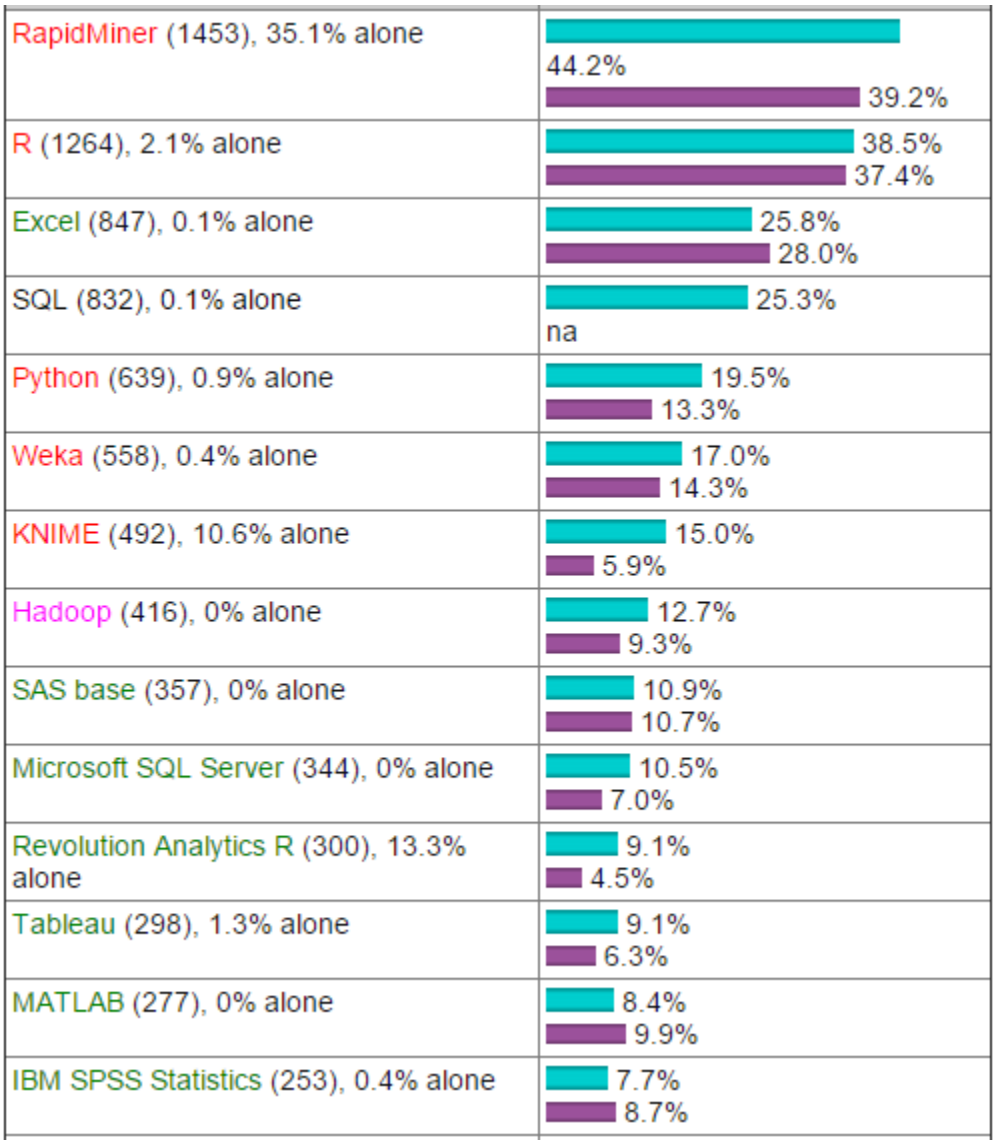
Κόκκινο: Δωρεάν/Ανοιχτού Κώδικα εργαλεία

Πράσινο: Εμπορικά εργαλεία

Ροζ: Hadoop εργαλεία

% χρήστες το 2014

% χρήστες το 2013



SAS Enterprise Miner (235), 1.3% alone	7.2% 5.9%
SAP (including BusinessObjects/Sybase/Hana) (225), 0% alone	6.8% 1.4%
Unix shell/awk/gawk (190), 0% alone	5.8% na
IBM SPSS Modeler (187), 3.2% alone	5.7% 6.1%
Other free analytics/data mining tools (168), 1.8% alone	5.1% 3.4%
Rattle (161), 0% alone	4.9% 4.5%
BayesiaLab (136), 23.5% alone	4.1% 1.0%
Other Hadoop/HDFS-based tools (129), 0% alone	3.9% na
Gnu Octave (128), 0% alone	3.9% 2.9%
JMP (125), 3.2% alone	3.8% 4.1%
KXEN (now part of SAP) (125), 0% alone	3.8% 1.9%
Predixion Software (122), 47.5% alone	3.7% 2.7%
Salford SPM/CART/Random	3.6%
Pig (116), 0% alone	3.5% na
Orange (112), 0% alone	3.4% 3.6%

Σχήμα 8 Αποτελέσματα δημοσκόπησης KDDnuggets 2014 σχετικά με την χρήση κάθε εργαλείου [16]

Όπως φαίνεται και από τα αποτελέσματα, το Rapidminer είχε τους περισσότερους χρήστες και κινητοποιώντας τους καλύτερα από τις υπόλοιπες εταιρείες.

Μία ακόμα διάκριση, που κέρδισε το Rapidminer, είναι η ανάδειξή του από τη Gartner - συμβουλευτική εταιρεία που δραστηριοποιείται και στον τομέα της τεχνολογίας πληροφοριών - το 2015, σε «ηγέτη», μέσω του Μαγικού Τεταρτημορίου.

Το Μαγικό Τεταρτημόριο είναι ένα σύστημα δύο αξόνων, μέσω του οποίου η Gartner αξιολογεί δεκαέξι προμηθευτές προηγμένων πλατφόρμων ανάλυσης. Ο κάθετος άξονας

εκφράζει την ικανότητα του εργαλείου στην εκτέλεση, ενώ ο οριζόντιος, την ολοκλήρωση του οράματος κάθε εταιρείας.[17]

Σχήμα 9 Μαγικό Τεταρτημόριο [17]

Το Rapidminer, ανήκει στους «ηγέτες» και όπως αναφέρεται από την εταιρεία, έφτασε σε αυτή τη θέση, λόγω των παρακάτω χαρακτηριστικών:

- Δυνατά σημεία
  -Η πλατφόρμα του υποστηρίζει ένα εκτενές εύρος και βάθος λειτουργικότητας.
  -Το Rapidminer επιλέγεται συχνότερα από πελάτες, λόγω της ευκολίας στη χρήση που παρουσιάζει, της ευκολίας στην εκτέλεση και τα μηδενικά κόστη άδειας και εγκατάστασης.

-Διαθέτει καινούρια ομάδα ηγεσίας και φιλόδοξους στόχους για διαχείριση μεγάλου όγκου δεδομένων. Από την ταχύτητα ανάπτυξης της εταιρείας, το Rapidminer πλεονεκτεί στην καινοτομία.

- Αδυναμίες
  -Παρότι διαθέτει πολλούς πελάτες, το Rapidminer στερείται μεγάλης προβολής στην αγορά, έξω από την κοινότητα της επιστήμης δεδομένων.
  -Αντιμετωπίζει πολλές δυσκολίες, καθώς η απόδοσή του, για την ικανοποίηση των πελατών και την παράδοση της αξίας των επιχειρήσεων, έχει μειωθεί κατά τους τελευταίους δώδεκα μήνες.
  -Οι πελάτες αντιμετώπισαν κακή εκπαίδευση και έλλειψη τεκμηρίωσης. Τόνισαν επίσης, ως σημεία αδυναμίας, τον τομέα των πωλήσεων και το τμήμα υποστήριξης πελατών.

Όπως φαίνεται από την παραπάνω ανάλυση και σύγκριση, η επιλογή του Rapidminer, για την πραγματοποίηση της συγκεκριμένης εργασίας, ενδείκνυται, καθώς:

- Ο όγκος των δεδομένων, που διατίθεται, είναι αρκετά μεγάλος.
- Το περιβάλλον του εργαλείου είναι πολύ φιλικό προς τον χρήστη, χωρίς την απαίτηση κώδικα, μειώνοντας έτσι τα σφάλματα, μέσω της γραφής, σε κάποια γλώσσα προγραμματισμού.
- Το ίδιο το εργαλείο, προσφέρει πολύ μεγάλο αριθμό τελεστών.
- Η εγκατάστασή του λογισμικού και η λήψη του είναι εντελώς δωρεάν.

ΚΕΦΑΛΑΙΟ 3

Μεθοδολογικό Πλαίσιο Πρόβλεψης Παραγωγής Ενέργειας

3.1 Θεωρητικό Υπόβαθρο

3.1.1 Εισαγωγή

Το αντικείμενο των προβλέψεων, έχει σαν στόχο, την όσο το δυνατόν ακριβέστερη εκτίμηση της ζητούμενης μεταβλητής στο μέλλον. Όπως αναφέρθηκε και στα προηγούμενα κεφάλαια, ο κλάδος των προβλέψεων είναι πολύ σημαντικός για την ισορροπία των συστημάτων ηλεκτρικής ενέργειας, όταν προστίθενται και οι ΑΠΕ. Η παραγωγή των προβλέψεων επιτυγχάνεται με την ανάλυση της διαθέσιμης γνώσης και αφορά μελλοντικά γεγονότα. Για την ανάλυση των ιστορικών δεδομένων, χρησιμοποιήθηκε το εργαλείο Rapidminer, το οποίο κατά κύριο λόγο, αποτελεί εργαλείο εξόρυξης δεδομένων.

3.1.2 Πρόβλεψη παραγωγής ηλεκτρικής ενέργειας

Η πρόβλεψη της παραγωγής ηλεκτρικής ενέργειας αποτελεί σημαντικό ζήτημα του τομέα της επιστήμης των προβλέψεων. Η πρόβλεψη της παραγωγής λειτουργεί ως αρωγός στην ασφάλεια του συστήματος ηλεκτρικής ενέργειας. Στηριζόμενοι σε αυτήν, οι χειριστές, μπορούν να αποφύγουν τυχόν αστοχίες του συστήματος. Για παράδειγμα, σε περίπτωση χαμηλής παραγωγής από τα φωτοβολταϊκά, μπορούν να αναπληρώσουν το έλλειμα από άλλες πηγές. Ή αντίστοιχα, σε περίπτωση αυξημένης παραγωγής, είναι δυνατόν να καλύψουν τη ζήτηση αποκλειστικά από τα φωτοβολταϊκά. Οι ενέργειες αυτές μπορούν να οδηγήσουν στην εξασφάλιση της κάλυψης της ζήτησης και σε οικονομικότερη και ασφαλέστερη παραγωγή ενέργειας. Γενικότερα, ακριβείς προβλέψεις συμβάλλουν άμεσα στην καλύτερη διαχείριση του συστήματος και στην καλύτερη λήψη αποφάσεων.

Κατά τη λήψη αποφάσεων, εμφανίζονται διάφορα προβλήματα, όπως για παράδειγμα ο βέλτιστος και ασφαλής προσδιορισμός δέσμευσης μιας μονάδας και κατανομής ενέργειας. [26]

Η λήψη αποφάσεων, που αφορά τα συστήματα ηλεκτρικής ενέργειας, στοχεύει και στη βραχυπρόθεσμη λειτουργία, αλλά και στη μακροπρόθεσμη, όπως είναι για παράδειγμα η μελλοντική εισαγωγή μιας μονάδας παραγωγής.

Για την ανάπτυξη των κατάλληλων μεθοδολογιών πρόβλεψης, που χρειάζεται ένα ΣΗΕ, απαιτείται μία περαιτέρω ανάλυση.

3.1.3 Κατηγορίες μεθόδων πρόβλεψης

Αρχικά, υπάρχουν τρεις κατηγορίες προβλέψεων: οι ποσοτικές μέθοδοι, οι κριτικές μέθοδοι, και οι τεχνολογικές μέθοδοι.

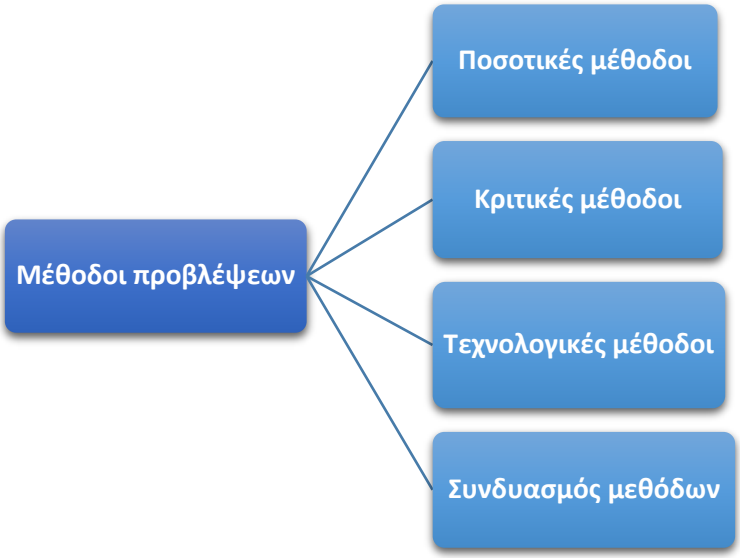
- Ποσοτικές μέθοδοι: Οι στατιστικές προβλέψεις προκύπτουν με την εφαρμογή ποσοτικών μοντέλων, πάνω σε μία σειρά δεδομένων και αναφέρονται στην εφαρμογή στατιστικών μοντέλων χρονοσειρών ή αιτιοκρατικών μοντέλων επί της σειράς αυτής, με σκοπό την αυτοματοποιημένη και συστηματική παραγωγή προβλέψεων [23]. Μειονεκτήματα αυτών των προβλέψεων είναι, η θεώρηση του αμετάβλητου στο χρόνο προτύπου της χρονοσειράς χωρίς να λαμβάνονται υπόψη τυχόν μελλοντικοί παράγοντες και αλλαγές που είναι πιθανό να επηρεάσουν την πρόβλεψη. Ωστόσο, ένα πλεονέκτημά τους, είναι η ευκολία στη χρήση και την παραγωγή αποτελεσμάτων μεγάλης ακρίβειας, μέσα σε σύντομο χρονικό διάστημα.

- Κριτικές μέθοδοι: οι κριτικές μέθοδοι, αντίθετα από τις ποσοτικές, δεν στηρίζονται σε μεγάλο όγκο δεδομένων. Ο κύριος άξονάς τους είναι ο ανθρώπινος παράγοντας, η γνώση, η κρίση και η διαίσθηση ενός ατόμου ή μιας ομάδας ατόμων. Επομένως, η πρόβλεψη μπορεί να προκύψει, είτε από ατομικές μεθόδους είτε από μεθόδους επιτροπής. Αντίθετα από τις ποσοτικές μεθόδους, υπερτερούν καθώς, μπορούν να λάβουν υπόψη τους και να αντιμετωπίσουν ασυνέχειες και ανομοιογένειες στα δεδομένα. Ταυτόχρονα όμως, επειδή στηρίζονται στον ανθρώπινο παράγοντα και στην ανθρώπινη κρίση, χαρακτηρίζονται από προκατάληψη. Για παράδειγμα, αν η τάση του υπεύθυνου είναι απαισιόδοξη, τότε και το μοντέλο θα είναι απαισιόδοξο και αντίθετα.

- Τεχνολογικές μέθοδοι: οι τεχνολογικές μέθοδοι, βρίσκουν εφαρμογή κυρίως σε προβλέψεις με μακροπρόθεσμο ορίζοντα. Χωρίζονται σε δύο υπό-κατηγορίες, τις διερευνητικές μεθόδους και τις κανονιστικές μεθόδους. Οι πρώτες εξετάζουν στο παρελθόν και στο παρόν και έχουν σαν στόχο τη διερεύνηση όλων των πιθανών περιπτώσεων στο μέλλον. Ενώ οι δεύτερες στηρίζονται πρώτα στην εφαρμογή των στόχων και έπειτα εξετάζουν τη δυνατότητα επίτευξης τους, ανάλογα με τους διαθέσιμους πόρους.

- Συνδυασμός μεθόδων: πολύ συχνά η επιλογή μίας μόνο μεθόδου δεν πετυχαίνει την επιθυμητή ακρίβεια. Αυτό συμβαίνει διότι, κάθε μέθοδος επιφέρει και έναν παράγοντα σφάλματος. Προκειμένου να μειωθεί το διάστημα διακύμανσης των

σφαλμάτων, συνδυάζονται οι προβλέψεις, που προκύπτουν από τις παραπάνω μεθόδους.



Σχήμα 10 Κατηγορίες Μεθόδων Πρόβλεψης

3.1.4 Ορίζοντας πρόβλεψης

Μια πρόβλεψη μπορεί να εξυπηρετήσει διάφορους σκοπούς. Για παράδειγμα, μπορεί να είναι απαραίτητη για την παρακολούθηση της καθημερινής λειτουργίας ενός συστήματος ηλεκτρικής ενέργειας ή για μία μελλοντική επένδυση, στην οποία θα προβεί μία επιχείρηση. Επομένως, οι μέθοδοι προβλέψεων χωρίζονται και ανάλογα με τον ορίζοντα πρόβλεψης. Ο ορίζοντας πρόβλεψης παρουσιάζει τις χρονικές περιόδους που πρέπει να προβλεφθούν. Με κριτήριο τον χρονικό ορίζοντα, υπάρχουν τρείς κατηγορίες:

- Βραχυπρόθεσμη πρόβλεψη: η τιμή του ορίζοντα πρόβλεψης είναι σχετικά μικρή. Καθώς προβλέπονται οι επόμενες ώρες έως και μία εβδομάδα. Συμβάλλει στην ενεργειακή διαχείριση ενός συστήματος ηλεκτρικής ενέργειας. Ουσιαστικά, στοχεύει στη ρύθμιση του συστήματος και ελέγχει τον τρόπο με τον οποίο εντάσσονται οι μονάδες παραγωγής. Έτσι αποφεύγονται καταστάσεις υπερφόρτωσης και αστοχίες, και βελτιώνεται η αξιοπιστία του δικτύου.

- Μεσοπρόθεσμη πρόβλεψη: λίγο μεγαλύτερος ορίζοντας πρόβλεψης, από ένα μήνα έως τρία έτη. Συμβάλλει στη συντήρηση των μονάδων παραγωγής και στο χειρισμό των διαθέσιμων συστημάτων ηλεκτρικής ενέργειας. Επιπλέον δίνει προβλέψεις για τη ζήτηση φορτίου, βοηθώντας στη διαπραγμάτευση των συμβάσεων.

- Μακροπρόθεσμη πρόβλεψη: χρονικός ορίζοντας μεγαλύτερος από τρία έτη. Βοηθάει στη λήψη αποφάσεων, σχετικά με την αγορά νέων μονάδων παραγωγής, δηλαδή στο σχεδιασμό επενδύσεων και στη μακροχρόνια ανάπτυξη αλλά και στη συντήρηση των ήδη υπαρχόντων.

Για την παραγωγή ηλεκτρικής ενέργειας από φωτοβολταϊκό, ενδείκνυται μεγάλος ορίζοντας πρόβλεψης (μακροπρόθεσμη πρόβλεψη).

3.1.5 Χρονοσειρές

Χρονοσειρές ή χρονολογικές σειρές είναι ένα σύνολο διαδοχικών παρατηρήσεων, οι οποίες περιγράφουν την πορεία ενός μεγέθους στο χρόνο. Οι παρατηρήσεις λαμβάνονται σε καθορισμένες και ισαπέχουσες χρονικές στιγμές και μπορούν να χρησιμοποιηθούν στην δημιουργία ενός μοντέλου, για την πρόβλεψη των μελλοντικών τιμών της χρονοσειράς.

Όταν οι παρατηρήσεις μεταξύ τους είναι εξαρτημένες και μπορούν να προσδιορίσουν ακριβώς την μελλοντική τιμή, η διαδικασία είναι ντετερμινιστική. Σύμφωνα με αυτό το μοντέλο, υπάρχει ακριβής γνώση των παραγόντων, που επηρεάζουν τη χρονοσειρά.

Αντίθετα, όταν τα μεγέθη επηρεάζονται από τον «τυχαίο παράγοντα», τότε η διαδικασία ονομάζεται στοχαστική. Στοχαστικές είναι όλες οι διαδικασίες στον πραγματικό κόσμο, καθώς τα ιστορικά δεδομένα δεν αρκούν για τον ακριβή προσδιορισμό μιας μελλοντικής τιμής.

3.1.6 Ποιοτικά χαρακτηριστικά χρονοσειρών

[23] «Η συστηματική μελέτη μιας χρονοσειράς ξεκινάει με την επισκόπηση του γραφήματός της, στο πεδίο του χρόνου». Τα βασικά ποιοτικά της χαρακτηριστικά είναι η τάση, η κυκλικότητα, η εποχιακότητα, οι ασυνέχειες και οι μη κανονικές διακυμάνσεις.

1. Τάση
   Η τάση αντιπροσωπεύει την εικόνα της χρονοσειράς και ορίζεται ως η μακροπρόθεσμη μεταβολή των τιμών της. Αν και δεν υπάρχουν αυτόματες τεχνικές, για την ταυτοποίηση της τάσης μιας χρονοσειράς, συνήθως ακολουθείται ανοδική ή πτωτική πορεία σε ευθεία γραμμή ή εκθετική καμπύλη, οπότε είναι εύκολος ο προσδιορισμός της. Η τάση, μπορεί να οριστεί μόνο στην περίπτωση ύπαρξης ικανοποιητικού αριθμού παρατηρήσεων, που επιτρέπει τον ορισμό του κατάλληλου μήκους περιόδου, όπου θα αναζητηθεί η τάση της χρονοσειράς.

2. Κυκλικότητα
   Η κυκλικότητα φανερώνει μια «κυματοειδή» μεταβολή, η οποία οφείλεται σε εξωγενείς παράγοντες. Εμφανίζεται σε μη σταθερές και μεγαλύτερες του έτους περιόδους. Τέλος ο κυκλικός παράγοντας εμφανίζεται στην γραφική παράσταση μιας χρονοσειράς από τις ανόδους και τις καθόδους ανάμεσα στην υψηλότερη και τη χαμηλότερη στάθμη.

3. Εποχιακότητα
   «Η εποχιακότητα ορίζεται σαν μία περιοδική διακύμανση, που έχει σταθερό και μικρότερο του έτους μήκος» [23]. Η συνιστώσα της εποχιακότητας αποτελείται από ενέργειες, που είναι σχετικά σταθερές ως προς το χρόνο, την κατεύθυνση και το μέγεθος. Μπορεί να προσδιοριστεί πολύ εύκολα σε μία χρονοσειρά, ανάλογα με τον τρόπο που επαναλαμβάνονται κάποιες αλλαγές στο χρόνο. Η διαφορά της εποχιακότητας με την κυκλικότητα είναι ότι, η πρώτη έχει σταθερή διάρκεια, σε αντίθεση με τη δεύτερη που μπορεί να αλλάζει από κύκλο σε κύκλο.

4. Ασυνέχειες
   Οι ασυνέχειες αποτελούν εκείνες τις παρατηρήσεις, οι οποίες ξεφεύγουν από την πορεία μιας χρονοσειράς και αποτελούν απότομες αλλαγές, που δεν θα μπορούσαν να προβλεφθούν. Ανάλογα με τη χρονική τους διάρκεια, χωρίζονται σε δύο κατηγορίες, "outliers/special events", οι οποίες δεν επηρεάζουν πολύ τη χρονοσειρά και σε "level-shifts", οι οποίες έχουν μόνιμο χαρακτήρα. Οι πρώτες μπορεί να οφείλονται σε κάποιο ασυνήθιστο και απρόβλεπτο γεγονός και απαιτούν ιδιαίτερη προσοχή. Οι "level-shifts", εμφανίζονται στο μέσο επίπεδο των τιμών της χρονοσειράς και έχουν μεγάλη διάρκεια.

5. Μη κανονικές διακυμάνσεις
   Πρόκειται ουσιαστικά για τις μη κανονικές διακυμάνσεις, που μένουν, όταν απομονωθούν από τη χρονοσειρά τα τρία πρώτα χαρακτηριστικά, που αναφέρθηκαν. Αντιπροσωπεύουν το στοιχείο του σφάλματος μιας τυχαίας μεταβλητής ή κάποιας ασυνέχειας.

3.1.7 Μοντέλα πρόβλεψης

Τα μοντέλα πρόβλεψης αποτελούν τον οδηγό της διαδικασίας παραγωγής προβλέψεων. Όπως αναφέρθηκε και στην παράγραφο 3.2.1, υπάρχουν διάφορες μέθοδοι προβλέψεων (ποσοτικές, κριτικές, τεχνολογικές) και η καθεμία αναλύεται με ένα διαφορετικό μοντέλο. Καθώς το πρόβλημα, αφορά την ανάλυση και ανάκτηση πληροφορίας από ιστορικά δεδομένα, τα οποία τείνουν να διατηρήσουν ένα σταθερό πρότυπο συμπεριφοράς, ενδείκνυνται μόνο οι ποσοτικές μέθοδοι πρόβλεψης. Τα μοντέλα που χαρακτηρίζουν αυτή τη μέθοδο είναι το μοντέλο χρονοσειρών (time series model) και το αιτιοκρατικό μοντέλο (casual relationship/explanatory).

3.1.8 Μοντέλο χρονοσειρών

Το μοντέλο χρονοσειρών (time series model) αποτελεί το πιο διαδεδομένο μοντέλο ποσοτικής ανάλυσης. Βασίζεται στην υπόθεση, ότι η προβλεπόμενη μεταβλητή, , ακολουθεί καθορισμένο και σταθερό πρότυπο, που επαναλαμβάνεται στο χρόνο. Επομένως, η πρόβλεψη επιτυγχάνεται με την αναγνώριση του προτύπου και την προέκτασή του στο χρόνο. Το μοντέλο αναπαρίσταται, ως ένα σύστημα, το οποίο δέχεται σαν είσοδο τα παρελθοντικά δεδομένα X και παράγει ως έξοδο την προβλεπόμενη τιμή Y, για την επόμενη χρονική περίοδο. Απαραίτητη προϋπόθεση του συγκεκριμένου μοντέλου είναι η κατοχή μεγάλου όγκου δεδομένων του υπό πρόβλεψη μεγέθους, προκειμένου να παραχθεί ένα ακριβές μοντέλο. Το μοντέλο των χρονοσειρών είναι κατάλληλο στις περιπτώσεις, όπου η εξέλιξη του εξεταζόμενου μεγέθους εξαρτάται μόνο από τις προηγούμενες τιμές του και όχι από άλλες παραμέτρους. Αυτή η ιδιαιτερότητα όμως αποτελεί και το βασικό μειονέκτημα του μοντέλου.

Ανάλογα με τον συναρτησιακό τύπο, που περιγράφεται το σύστημα, είναι δυνατόν αυτό να διαχωριστεί σε τρεις βασικές μεθόδους: την αποσύνθεση (decomposition), την εξομάλυνση (smoothing) και τις αυτοπαλινδρομικές μεθόδους κινητού μέσου όρου (autoregressive moving average).

3.1.9 Αιτιοκρατικό μοντέλο

Το αιτιοκρατικό μοντέλο (casual relationship model) βασίζεται στην υπόθεση, ότι υπάρχει μία σταθερή σχέση, η οποία συνδέει το προβλεπόμενο μέγεθος (εξαρτημένη μεταβλητή) με ορισμένες παραμέτρους, που το επηρεάζουν (ανεξάρτητες μεταβλητές). Αντίθετα από το μοντέλο χρονοσειρών, στο οποίο, η συνάρτηση που ορίζει το σύστημα ορίζεται από το πρότυπο των ιστορικών δεδομένων, στο αιτιοκρατικό μοντέλο πρώτα προσδιορίζεται η συσχέτιση της εξαρτημένης μεταβλητής με τις ανεξάρτητες και μετά επιχειρείται η πρόβλεψη της εξαρτημένης μεταβλητής, μέσω της συνάρτησης, που συνδέει την έξοδο με την είσοδο του συστήματος.
Μερικά μειονεκτήματα που παρουσιάζει αυτό το μοντέλο είναι η μεγάλη ευαισθησία σε αλλαγές των ανεξάρτητων μεταβλητών, διότι αυτές επηρεάζουν και το εξεταζόμενο μέγεθος, καθώς και ο μεγάλος όγκος δεδομένων που απαιτεί.
Από την άλλη πλευρά, το βασικό πλεονέκτημα του αιτιοκρατικού μοντέλου είναι το γεγονός ότι, μπορεί να προβλεφθεί η έξοδος του συστήματος, δηλαδή η εξεταζόμενη τιμή για διάφορες τιμές των μεταβλητών εισόδου.

3.1.10 Επιλογή Κατάλληλης μεθόδου πρόβλεψης

Στη συγκεκριμένη παράγραφο αναλύονται οι παράγοντες, οι οποίοι επηρεάζουν τη επιλογή της κατάλληλης μεθόδου πρόβλεψης.

- Τα χαρακτηριστικά των δεδομένων

  Οι διαφορετικές μέθοδοι πρόβλεψης εφαρμόζονται αποτελεσματικότερα για διαφορετικά πρότυπα δεδομένων. Για την αναγνώριση ενός προτύπου συμπεριφοράς των εξεταζόμενων δεδομένων, υπάρχουν κάποια βασικά χαρακτηριστικά. Αυτά τα χαρακτηριστικά, όπως αναφέρθηκε και στην παράγραφο 3.3.1, είναι η τάση, η κυκλικότητα, η εποχιακότητα, οι ασυνέχειες και οι μη κανονικές διακυμάνσεις. Σε χρονοσειρές, οι οποίες χαρακτηρίζονται από έντονη τυχαιότητα, προτιμώνται απλές μέθοδοι. Ενώ σε χρονοσειρές, που εμφανίζουν έντονη τάση ή κυκλικότητα, ιδανικότερες είναι οι αυτοπαλινδρομικές μέθοδοι.

- Ορίζοντας πρόβλεψης

  Ο χρονικός ορίζοντας πρόβλεψης μπορεί να είναι είτε βραχυπρόθεσμος, είτε μεσοπρόθεσμος, είτε μακροπρόθεσμος. Σε περίπτωση βραχυπρόθεσμης πρόβλεψης και ύπαρξης δεδομένων με ωριαίο ή ημερήσιο τύπο, τότε προτιμώνται στατιστικές μέθοδοι. Από την άλλη πλευρά, αν πρόκειται για μεσοπρόθεσμη ή μακροπρόθεσμη πρόβλεψη, θα προτιμώνται ποιοτικές μεθόδοι.

- Ευκολία εφαρμογής της μεθόδου

  Είναι γεγονός πως η ευκολία και η απλότητα εφαρμογής μιας μεθόδου παίζουν πολύ σημαντικό ρόλο, καθώς ένα πολύπλοκο μοντέλο απαιτεί χρονοβόρες υπολογιστικές διαδικασίες και είναι πιο δύσκολο στην κατανόησή του από τρίτους.

- Κόστος

  Το κόστος επηρεάζει άμεσα την επιλογή μιας μεθόδου πρόβλεψης και εξαρτάται κυρίως από τον όγκο των δεδομένων και την πολυπλοκότητα του μοντέλου.

- Αξιοπιστία

  Κάθε μοντέλο διαθέτει κάποια όρια αξιοπιστίας. Όσο πιο αυστηρά είναι αυτά τα όρια, τόσο πιο ακριβές είναι το μοντέλο.

Επομένως προτιμάται η εφαρμογή ενός μοντέλου, για την πρόβλεψη της παραγωγής ενός ΦΒ συστήματος, με τη μέθοδο της Πολλαπλής Γραμμικής Παλινδρόμησης.

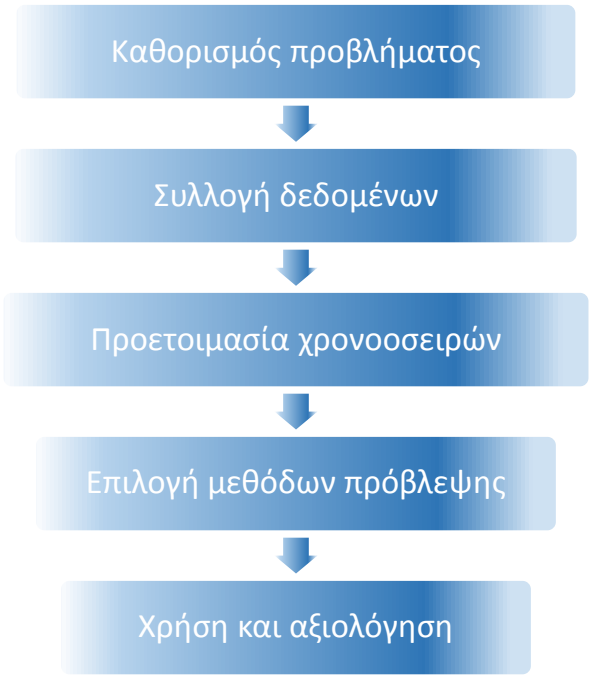
3.1.11 Περιορισμοί προβλεπτικών αναλύσεων

Μετά την επιλογή του εργαλείου, για την εξόρυξη δεδομένων, καθώς επίσης και της προτεινόμενης ανάλυσης, απομένει η περιγραφή της διαδικασίας πρόβλεψης.

Αρχικά απαιτείται η γνώση των ορίων της ανάλυσης πρόβλεψης. Απαραίτητη προϋπόθεση είναι η ύπαρξη ενός συνόλου δεδομένων, επαρκούς μεγέθους και ποιότητας, για εκπαίδευση. Επιπλέον, πρέπει να υπάρχει σαφής ορισμός και ιστορικά παραδείγματα της έννοιας, που θα προβλεφθεί. Για να είναι μία διαδικασία εξόρυξης δεδομένων πετυχημένη, πρέπει επίσης, οι δράσεις που στηρίζονται σε προβλέψεις να είναι ξεκάθαρα καθορισμένες και να έχουν αξιόπιστες συνέπειες κέρδους και μικρές αθέλητες συνέπειες. Τα δεδομένα, τα οποία χρησιμοποιούνται, κατά το στάδιο της εκπαίδευσης, πρέπει να είναι αντιπροσωπευτικά των εξεταζόμενων δεδομένων. Συνήθως τα πρώτα λαμβάνονται από ιστορικά σύνολα δεδομένων, ενώ τα δεύτερα δημιουργούνται στο μέλλον. Αν το φαινόμενο, που θα προβλεφθεί, δεν είναι σταθερό στο χρόνο, τότε οι προβλέψεις είναι πολύ πιθανόν να είναι άχρηστες. Τέλος, για μία πετυχημένη εφαρμογή, πρέπει οι συνέπειες των δράσεων να είναι ουσιαστικά ανεξάρτητες για διαφορετικά παραδείγματα.

Επομένως, για να υλοποιηθεί μια ανάλυση πρόβλεψης, είναι απαραίτητα να ακόλουθα πέντε βήματα: Καθορισμός του προβλήματος, συλλογή των δεδομένων, προετοιμασία χρονοσειρών, επιλογή μεθόδων πρόβλεψης, χρήση και αξιολόγηση των μοντέλων πρόβλεψης.

Σε αυτό το σημείο, ένα διάγραμμα ροής των διαφόρων σταδίων της ανάλυσης, βοηθά στην κατανόηση της δομής του υπόλοιπου κεφαλαίου.



Σχήμα 11 Διάγραμμα ροής της διαδικασίας της πρόβλεψης

3.2 Μεθοδολογία

3.2.1 Καθορισμός του προβλήματος

Το πρώτο στάδιο της διαδικασίας παραγωγής και αξιολόγησης προβλέψεων είναι ο καθορισμός του προβλήματος. Αποτελεί ένα από τα δυσκολότερα στάδια της διαδικασίας, καθώς ο αναλυτής πρέπει να καταλάβει αυτό που χρειάζεται να προβλεφθεί, τους παράγοντες από τους οποίους επηρεάζεται και την μετέπειτα χρήση των προβλέψεων.

Στην παρούσα εργασία, το πρόβλημα που πρέπει αντιμετωπισθεί, είναι η ανάπτυξη μοντέλου πρόβλεψης παραγωγής ενέργειας σε ΦΒ εγκατάσταση. Το μοντέλο πρόβλεψης θα αναπτυχθεί στο περιβάλλον του Rapidminer και η συνολική υλοποίηση στο εργαλείο, θα ακολουθεί πέντε στάδια.

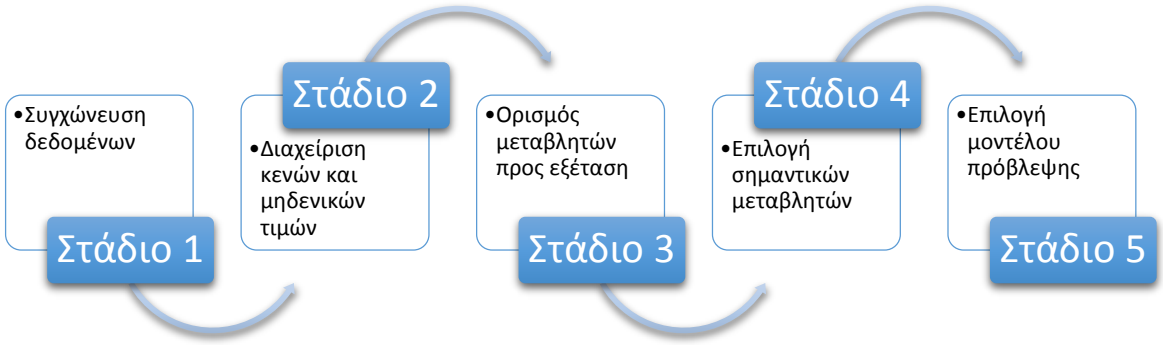


Figure 12 Διάγραμμα ροής σταδίων υλοποίησης

Τα αποτελέσματα μίας τέτοιας ανάλυσης θα μπορούσαν να χρησιμοποιηθούν από ένα δίκτυο ηλεκτρικής ενέργειας, το οποίο διαθέτει μία ΦΒ διάταξη.

Επίσης το επερχόμενο μοντέλο προβλέψεων, θα μπορούσε να χρησιμοποιηθεί αυτούσιο σε μία οποιαδήποτε ΦΒ εγκατάσταση.

3.2.2 Συλλογή δεδομένων

Απαιτείται προσοχή και έμφαση στην ορθή συλλογή και συντήρηση των δεδομένων. Πρέπει να συγκεντρωθούν ιστορικά δεδομένα για την εγκατάσταση που να αναφέρονται σε όσο το δυνατόν μεγαλύτερο χρονικό διάστημα. Στη συγκεκριμένη περίπτωση, τα ιστορικά δεδομένα που ήταν διαθέσιμα, αναφέρονταν σε διάστημα τριών χρόνων.

3.2.3 Προετοιμασία χρονοσειρών

Μέσω της προετοιμασίας της χρονοσειράς επιτυγχάνεται, η προσαρμογή, η εξομάλυνση και γενικότερα η προετοιμασία των δεδομένων, για την εφαρμογή μοντέλων πρόβλεψης. Τα δεδομένα χωρίζονται σε δύο κατηγορίες: στα δεδομένα προς εκπαίδευση και στα δεδομένα προς εξέταση. Τα πρώτα χρησιμοποιούνται στο στάδιο της εκπαίδευσης, για την δημιουργία του μοντέλου, ενώ τα δεύτερα για την εξέταση της ακρίβειας του μοντέλου.

Υπάρχουν δύο τρόποι με τους οποίους μπορεί να αξιολογηθεί ένα μοντέλο:

- In sample
  Σύμφωνα με αυτή τη μέθοδο, η αξιολόγηση του μοντέλου γίνεται στο 100% του συνόλου δεδομένων, δηλαδή και στα δεδομένα προς εκπαίδευση και στα δεδομένα προς εξέταση.

- Out of sample
  Στη συγκεκριμένη μέθοδο, η αξιολόγηση του μοντέλου γίνεται μόνο βάση των δεδομένων προς εξέταση, δηλαδή του 20% του συνολικού όγκου.

Στα πλαίσια της εργασίας, η αξιολόγηση γίνεται με τη μέθοδο "out of sample"

Στη συγκεκριμένη παράγραφο αναλύεται η εποπτευόμενη μάθηση (supervised learning), ο καθαρισμός δεδομένων (data cleaning), η κωδικοποίηση (recoding), τα ελλιπή δεδομένα (missing data) και η έννοια της υπέρ-προσαρμογής (overfitting).

3.2.3.1 *1 Εποπτευόμενη μάθηση (supervised learning)*

Ο στόχος ενός αλγορίθμου, εποπτευόμενης μάθησης, είναι η απόκτηση ενός ταξινομητή, μέσα από τα παραδείγματα εκπαίδευσης. Ο ταξινομητής είναι το απαιτούμενο εργαλείο, για τη δημιουργία προβλέψεων , πάνω στα εξεταζόμενα δεδομένα.

Κάθε παράδειγμα εκπαίδευσης και εξέτασης παρουσιάζεται ως ένα διάνυσμα σειράς σταθερού μήκους *p*. Κάθε στοιχείο του διανύσματος, το οποίο αντιπροσωπεύει ένα παράδειγμα, ονομάζεται χαρακτηριστική αξία. Μπορεί να είναι ένας πραγματικός αριθμός ή μία τιμή οποιουδήποτε άλλου τύπου. Ένα σύνολο, που χρησιμοποιείται για εκπαίδευση, είναι ένα σύνολο διανυσμάτων, με γνωστές τις τιμές των δεδομένων. Υπάρχει διαφορά ανάμεσα σε ένα χαρακτηριστικό, που είναι μία ολόκληρη στήλη και μία χαρακτηριστική τιμή.

Για ένα εξεταζόμενο παράδειγμα, ο άγνωστος είναι η τιμή *y*,. Η έξοδος του ταξινομητή είναι μια υπόθεση για το *y*, μια προβλεπόμενη τιμή. Συνήθως, κάθε ετικέτα της τιμής y είναι πραγματικός αριθμός. Σε αυτή την περίπτωση, η εποπτευόμενη μάθηση ονομάζεται «παλινδρόμηση» και ο ταξινομητής ονομάζεται «μοντέλο παλινδρόμησης». Η λέξη «ταξινομητής» χρησιμοποιείται κυρίως, όταν οι τιμές είναι διακεκριμένες. Στις περισσότερες περιπτώσεις, υπάρχουν μόνο δύο τιμές.

Αν διατίθενται *n* παραδείγματα εκπαίδευσης και κάθε παράδειγμα περιλαμβάνει τιμές για *p* διαφορετικά χαρακτηριστικά, τότε τα δεδομένα εκπαίδευσης είναι ένας πίνακας με *n* γραμμές και *p* στήλες, μαζί με ένα διάνυσμα στήλης τιμών *y*.

3.2.3.2 *Εξακρίβωση και Καθαρισμός Δεδομένων (Data Validation and Cleaning)*

Καθώς ξεκινάει η εξόρυξη δεδομένων, δεν είναι απολύτως σαφή τα δεδομένα. Σημαντική τεκμηρίωση μπορεί να λείπει και τα δεδομένα μπορεί να προέρχονται από διάφορες πηγές, οι οποίες τα παράγουν με διαφορετικό τρόπο. Ως εκ τούτου, το πρώτο βήμα, σε ένα έργο εξόρυξης δεδομένων, είναι ο εντοπισμός των λαθών και η μείωση του πλήθους αυτών. Αυτό το στάδιο, είναι συνήθως το πιο χρονοβόρο και το δυσκολότερο, καθώς απαιτεί εμπειρία, διαίσθηση, κρίση και συναναστροφή με πολλούς άλλους ανθρώπους.

Επικύρωση δεδομένων σημαίνει επιβεβαίωση της αξιοπιστίας αυτών, ενώ καθαρισμός δεδομένων σημαίνει, διόρθωση τυχόν λάθη που υπάρχουν. Πολύ συχνά, δεν είναι δυνατή η αξιολόγηση μιας τιμής ως σωστής ή λανθασμένης και στην περίπτωση που είναι λάθος, είναι απίθανο να βρεθεί η σωστή τιμή. Επιπλέον, υπάρχει η πιθανότητα τα λάθη να είναι τόσα πολλά, που το κόστος διόρθωσής τους να είναι απαγορευτικό. Ωστόσο, υπάρχει η δυνατότητα, ορισμένα δεδομένα να είναι σωστά ή λάθος, επειδή έχουν ή δεν έχουν περάσει μια σειρά από ελέγχους. Όταν τα δεδομένα έχουν μεγάλη πιθανότητα λάθους, η απλούστερη προσέγγιση για τον καθαρισμό είναι απλά η απόρριψή τους, καθώς πιο εξελιγμένες και εξειδικευμένες μέθοδοι καθαρισμού μπορεί να μην είναι επωφελείς. Γενικά, περισσότερη προσοχή πρέπει να δοθεί στην εξακρίβωση των δεδομένων παρά στην επισκευή τους.

Μία ακόμα μέθοδος καθαρισμού δεδομένων είναι η συγχώνευση εγγραφών, που αναφέρονται στην ίδια οντότητα και δεν θα έπρεπε να είναι χωρισμένες. Οι ξεχωριστές εγγραφές συνήθως οφείλονται στις διακυμάνσεις της παρουσίασης, όπως είναι για παράδειγμα οι διαφορετικοί τρόποι, με τους οποίους γράφεται η ίδια διεύθυνση. Ένα πρώτο βήμα προς την επικύρωση δεδομένων είναι η εξέταση αναφορών, οι οποίες περιέχουν τα βασικά στατιστικά στοιχεία (ελάχιστο, μέγιστο, πιο συχνές τιμές κλπ.) για κάθε μεταβλητή. Σε αυτή τη μέθοδο, πρέπει να είναι γνωστός ο λόγος εμφάνισης των συχνότερων τιμών. Μεταβλητές, οι οποίες θα έπρεπε να έχουν μοναδικές τιμές, πρέπει να επιβεβαιώνεται, ότι πράγματι οι τιμές τους είναι μοναδικές.

Επιπλέον, αν ένα σύνολο δεδομένων, περιέχει *m* μεταβλητές, τότε υπάρχουν *m(m-1)/2* πιθανά ζευγάρια μεταβλητών. Επομένως, θα ήταν σχεδόν αδύνατο η διαπίστωση ανωμαλιών στη συσχέτιση ανάμεσα σε όλα τα ζευγάρια. Ωστόσο, μία μέθοδος που θα μπορούσε να εφαρμοστεί θα ήταν η εξέταση της συσχέτισης ανάμεσα σε συγκεκριμένες μεταβλητές, όπου συσχετίσεις δεν θα έπρεπε να υπάρχουν ή θα ήταν μικρές.

Ένα ακόμα σημαντικό βήμα της εξακρίβωσης δεδομένων είναι ο έλεγχος, των κατανομών, που θα έπρεπε να είναι παρόμοιες. Πολύ συχνά σύνολα, τα οποία θα έπρεπε να είναι στατιστικά ίδια, παρουσιάζουν κάποιες διαφορές.

3.2.3.3 *Κωδικοποίηση δεδομένων*

Στα πραγματικά δεδομένα, υπάρχει μεγάλη πολυπλοκότητα και μεταβλητότητα. Κάποιες μεταβλητές έχουν πραγματική τιμή, άλλες αριθμητική, αλλά όχι πραγματική τιμή. Πολλές

μεταβλητές περιλαμβάνουν κατηγορίες, για παράδειγμα για έναν φοιτητή η μεταβλητή «Έτος» μπορεί να έχει τις τιμές, πρωτοετής, δευτεροετής και τελειόφοιτος. Συνήθως τα ονόματα, τα οποία χρησιμοποιούνται για τις διαφορετικές τιμές μιας τέτοιας μεταβλητής, επηρεάζουν τους αλγορίθμους εξόρυξης δεδομένων, αλλά είναι πολύ σημαντικά για την ανθρώπινη κατανόηση. Επίσης, μερικές φορές οι μεταβλητές του συγκεκριμένου τύπου, μπορεί να έχουν ονόματα, τα οποία φαίνονται σαν αριθμητικές τιμές, για παράδειγμα ταχυδρομικοί κώδικες. Σε αυτή την περίπτωση η διαχείρισή τους είναι αρκετά δύσκολη.

Δύσκολη είναι επίσης η διαχείριση μεταβλητών, οι οποίες υφίστανται μόνο όταν συνδυάζονται με άλλες, όπως για παράδειγμα είναι η μεταβλητή «ημέρα» στο συνδυασμό «ημέρα/μήνας/έτος». Επιπλέον, σημαντικές μεταβλητές (δηλαδή μεταβλητές οι οποίες έχουν προγνωστική δύναμη), μπορεί να προκύπτουν έμμεσα από άλλες μεταβλητές. Για παράδειγμα, η ημέρα της εβδομάδας μπορεί να είναι προβλεπτική μεταβλητή, αλλά μόνο οι μεταβλητές ημέρα/μήνας/έτος να δίνονται στα αρχικά δεδομένα, όμως κανένας αλγόριθμος δεν μπορεί να διαχωρίσει την ημέρα της εβδομάδας αυτόματα, ως συνάρτηση της μεταβλητής «ημέρα/μήνας/έτος». Επομένως, μέσω της κατανόησης της διαδικασίας πρέπει να βρεθούν οι μεταβλητές που έχουν προβλεπτική δύναμη, και έπειτα να γραφτεί ένας κώδικα, ο οποίος θα καθιστά αυτές τις μεταβλητές σημαντικές.

Ακόμη και με την πολυπλοκότητα των μεταβλητών, πολλές διαστάσεις συνήθως αγνοούνται, όπως για παράδειγμα, μονάδες όπως τα ευρώ και διαστάσεις όπως τα κιλά.

Κάποιοι αλγόριθμοι εκπαίδευσης μπορούν να διαχειριστούν μόνο μεταβλητές, που σαν τιμές έχουν κατηγορίες.

Κάποιοι άλλοι αλγόριθμοι εκπαίδευσης, μπορούν να διαχειριστούν μόνο μεταβλητές με πραγματικές τιμές. Γι' αυτούς, οι μεταβλητές με κατηγορίες πρέπει να γίνουν αριθμητικές. Οι τιμές μας δυαδικής μεταβλητής μπορούν να κωδικοποιηθούν ως 0 ή 1. Συνήθως ενδείκνυται η κωδικοποίηση του μηδενός ως «όχι» ή «ψέμα» και της μονάδας ως «ναι» ή «αλήθεια». Συνήθως ο καλύτερος τρόπος κωδικοποίησης μεταβλητής, η οποία έχει *k* διαφορετικές κατηγορίες, είναι η χρήση *k* μεταβλητών πραγματικής τιμής. Για την ν-οστή κατηγορηματική τιμή, ορίζεται η ν-οστή από αυτές τις μεταβλητές, ίση με και τις υπόλοιπες *k-1* ίσες με 0.

Οι κατηγορηματικές μεταβλητές, με πάνω από 20 τιμές, είναι δύσκολα διαχειρίσιμες. Συνήθως η ανθρώπινη παρέμβαση είναι αναγκαία για την σωστή κωδικοποίηση. Για παράδειγμα, αν ένα αρχείο περιλαμβάνει αριθμούς τηλεφώνου, θα πρέπει να απομονωθούν τα πρώτα πέντε ψηφία, τα οποία υποδεικνύουν τις περιοχές της Ελλάδος.

Ένας έξυπνος τρόπος κωδικοποίησης παραγόντων πρόβλεψης είναι, η αντικατάσταση κάθε διακριτής τιμής από το μέσο όρο της κατηγορίας, στην οποία αναφέρεται. Για παράδειγμα, αν η μέση τιμή *y* είναι 20 για του άνδρες και 16 για τις γυναίκες, αυτές οι τιμές θα μπορούσαν να αντικαταστήσουν τις τιμές «άνδρας» και «γυναίκα» μιας μεταβλητής φύλου.

Ωστόσο, ο κλασικός τρόπος κωδικοποίησης μιας διακριτής μεταβλητής με *m* τιμές είναι, η εισαγωγή *m-1* δυαδικών μεταβλητών. Με αυτή την τυποποιημένη προσέγγιση, ο αλγόριθμος εκπαίδευσης μπορεί να μάθει έναν συντελεστή για κάθε νέα μεταβλητή, που αντιστοιχεί σε μία βέλτιστη αριθμητική τιμή για την αντίστοιχη διακριτή τιμή.

Τέλος, όταν επαναλαμβάνεται η επεξεργασία και κωδικοποίηση δεδομένων, δεν πρέπει να ληφθούν υπόψη τα δεδομένα προς εξέταση. Στην περίπτωση, που η επαναλαμβανόμενη επεξεργασία είναι με οποιονδήποτε τρόπο βασισμένη στα δεδομένα προς εξέταση, τότε αυτά τα δεδομένα είναι έμμεσα διαθέσιμα από την εκπαίδευση, κάτι το οποίο μπορεί να οδηγήσει σε υπέρ-προσαρμογή(overfitting). Ακόμα πιο σημαντικό είναι να μην χρησιμοποιηθούν οι τιμές των εξεταζόμενων δεδομένων πριν ή κατά τη διάρκεια της εκπαίδευσης. Όταν μια διακριτή τιμή αντικαθίσταται από το μέσο όρο της κατηγορίας, στην οποία αναφέρεται, η μέση τιμή πρέπει να προέρχεται αποκλειστικά και μόνο από τον κλάδο της εκπαίδευσης.

3.2.3.4 ***Ελλιπή δεδομένα***

Ένα σύνηθες πρόβλημα, είναι τα ελλιπή δεδομένα, όταν δηλαδή η τιμή μιας δοθείσας μεταβλητής, προς εξέταση ή προς εκπαίδευση, δεν είχε καταγραφεί και αποθηκευτεί στη βάση δεδομένων. Συνήθως, οι τιμές που λείπουν, υποδεικνύονται με ερωτηματικό «?». Άλλες φορές, υποδεικνύονται με σειρά ψηφίων, όπως το 0 και μοιάζουν με έγκυρες τιμές. Σε αυτές τις περιπτώσεις, δεν πρέπει να μπερδεύονται οι τιμές που λείπουν με έγκυρες, κανονικές τιμές.

Οι κενές τιμές μπορεί να οφείλονται σε αστοχία του πληροφοριακού συστήματος, το οποίο τροφοδοτεί με τα δεδομένα ή ακόμα και σε λάθος ανθρώπινο χειρισμό.[23]

Ορισμένοι αλγόριθμοι εκπαίδευσης μπορούν να αντιμετωπίσουν τα ελλιπή δεδομένα, εσωτερικά. Παρ’ όλα αυτά, η πιο απλή προσέγγιση από την πλευρά μας είναι, η απόρριψη όλων των παραδειγμάτων (σειρών) που περιέχουν κενές τιμές. Αυτή η μέθοδος, διατήρησης μόνο σειρών χωρίς ελλείψεις, ονομάζεται «πλήρης ανάλυση κατά περίπτωση». Μία εξίσου εύκολη, αλλά διαφορετική προσέγγιση είναι, η απόρριψη όλων των μεταβλητών (στηλών), οι οποίες περιέχουν ελλείψεις. Ωστόσο, αυτές οι μέθοδοι δεν ενδείκνυνται στη συγκεκριμένη εργασία, καθώς εξαλείφουν πολλά χρήσιμα δεδομένα εκπαίδευσης.

Row No.	datetime	tmpf	dwpf	relh	drct	sknt	clouds
1	2012-01-01T	?	?	?	?	?	0
2	2012-01-01T	?	?	?	?	?	0
3	2012-01-01T	?	?	?	?	?	0
4	2012-01-01T	5	5	100	290	154.332	0
5	2012-01-01T	6	5	93.300	300	154.332	1
6	2012-01-01T	5	4	93.240	300	154.332	1
7	2012-01-01T	6	5	93.300	0	0	1
8	2012-01-01T	8	7	93.400	0	0	1
9	2012-01-01T	10	8	87.370	0	0	1
10	2012-01-01T	12	8	76.500	0	0	1
11	2012-01-01T	13	8	71.640	?	0.514	1
12	2012-01-01T	14	8	67.120	0	0	1
13	2012-01-01T	14	8	67.120	0	0	0
14	2012-01-01T	15	8	62.920	0	0	0
15	2012-01-01T	15	8	62.920	?	0.514	0
16	2012-01-01T	16	9	63.150	?	0.514	0
17	2012-01-01T	16	8	59.010	200	205.776	0
18	2012-01-01T	16	8	59.010	200	308.664	0
19	2012-01-01T	17	8	55.370	?	102.888	0
20	2012-01-01T	17	7	51.720	190	25.722	0
21	2012-01-01T	17	8	55.370	?	102.888	0

Σχήμα 13 Ελλιπή δεδομένα αναπαρίστανται με «?»

Αν τελικά επιλεγεί η διατήρηση της μεταβλητής, με τις τιμές που λείπουν, τότε ακολουθείται μία από τις παρακάτω διαδικασίες εκτίμησης της ελλείπουσας τιμής, ανάλογα με την περίπτωση:

- Εύρεση κενής τιμής, από άλλες πηγές ή απευθείας ορισμός αυτής. Για να είναι ακριβής αυτή η μέθοδος, πρέπει να υπάρχει απόλυτη εγκυρότητα της δευτερεύουσας πηγής ή να υπάρχει ασφαλής κριτική εκτίμηση για το ύψος στο οποίο κυμάνθηκε.
- Ορίζεται η κενή τιμή ίση με το ημιάθροισμα (μέσος όρος) της προηγούμενης και της επόμενης παρατήρησης. Αυτή η διαδικασία μπορεί να εφαρμοστεί μόνο σε περιπτώσεις, όπου τα δεδομένα δεν χαρακτηρίζονται από εποχιακή συμπεριφορά.
- Αντίθετα, στην περίπτωση που τα δεδομένα παρουσιάζουν σαφή εποχιακή συμπεριφορά, τότε η κενή τιμή ορίζεται, ως ο μέσος όρος των τιμών των αντίστοιχων περιόδων. Για παράδειγμα, αν τα δεδομένα αποτελούνται από την

παραγωγή ενός ΦΒ συστήματος, και παρατηρηθούν κενές τιμές τον Ιούλιο, τότε οι τιμές μπορούν να γεμίσουν από το μέσο όρο των υπόλοιπων Ιουλίων. [23]

Όπως αναφέρεται και σε επόμενα κεφάλαια, στη συγκεκριμένη εργασία, οι τιμές που λείπουν γέμισαν μέσω γραμμικής παρεμβολής. Οι λόγοι για αυτή την επιλογή θα παρουσιαστούν αργότερα.

Το γεγονός ότι μια τιμή λείπε, μπορεί να εξαρτάται από το ίδιο το σύστημα, το οποίο τροφοδοτεί τα δεδομένα ή ακόμα και από την τιμή μιας άλλης μεταβλητής. Για παράδειγμα, αν η ταχύτητα του ανέμου είναι μηδέν, τότε η κατεύθυνση του ανέμου είναι απροσδιόριστη.

3.2.3.5 ***Υπέρ-προσαρμογή***

Σε μία πραγματική εφαρμογή της εποπτευόμενης μάθησης, υπάρχει ένα σύνολο παραδειγμάτων εκπαίδευσης με τιμές, καθώς και ένα σύνολο παραδειγμάτων εξέτασης με άγνωστες τιμές. Σκοπός της εξόρυξης δεδομένων για πρόβλεψη είναι, η πρόβλεψη των εξεταζόμενων δεδομένων.

Ωστόσο, κατά την έρευνα και τους πειραματισμούς, απαιτείται η μέτρηση της απόδοσης κάθε χρησιμοποιούμενου αλγορίθμου. Προκειμένου αυτή να μετρηθεί, χρησιμοποιείται ένα εξεταζόμενο σύνολο, το οποίο διαθέτει γνωστές τιμές. Εκπαιδεύεται ο ταξινομητής, του χρησιμοποιούμενου συνόλου, εφαρμόζεται πάνω στο εξεταζόμενο σύνολο και έπειτα μετριέται η απόδοση, συγκρίνοντας τις προβλεπόμενες τιμές, που προκύπτουν, με τις πραγματικές τιμές (οι οποίες δεν ήταν διαθέσιμες στον αλγόριθμο που εφαρμόστηκε κατά το στάδιο της εκπαίδευσης).

Είναι σχεδόν απαραίτητος ο υπολογισμός της απόδοσης ενός ταξινομητή, πάνω σε ένα ανεξάρτητο εξεταζόμενο σύνολο. Κάθε αλγόριθμος, ο οποίος χρησιμοποιείται στην εκπαίδευση, ψάχνει μοτίβα ανάμεσα στα δεδομένα εκπαίδευσης, δηλαδή συσχετίσεις ανάμεσα στις μεταβλητές. Κάποια από τα μοτίβα, τα οποία ανακαλύπτονται, μπορεί να είναι ψευδή, δηλαδή μπορεί να είναι έγκυρα για τα δεδομένα εκπαίδευσης, εξαιτίας του τυχαίου τρόπου, με τον οποίο αυτά επιλέχτηκαν από το δείγμα, αλλά δεν είναι έγκυρα ή όχι τόσο ισχυρά, σε ολόκληρο το δείγμα. Ένας ταξινομητής, ο οποίος στηρίζεται σε αυτά τα ψευδή μοτίβα, θα έχει υψηλότερη ακρίβεια στα δεδομένα εκπαίδευσης, παρά στο υπόλοιπο δείγμα. Μόνο όταν η ακρίβεια μετριέται σε ένα ανεξάρτητο σύνολο εξέτασης, είναι μία δίκαιη εκτίμηση της ακρίβειας του ταξινομητή σε ολόκληρο το δείγμα. Το φαινόμενο, κατά το οποίο στηριζόμαστε σε μοτίβα, τα οποία είναι ισχυρά μόνο στα δεδομένα εκπαίδευσης, ονομάζεται «υπέρ-προσαρμογή». Ο χρήστης μπορεί να επιλέξει ποιες μεταβλητές θα χρησιμοποιηθούν. Επίσης μπορεί να τρέξει έναν αλγόριθμο πολλές φορές και να μετρήσει την απόδοσή του αλλάζοντας κάποιες ρυθμίσεις του.

3.2.4 Επιλογή μεθόδων πρόβλεψης

Για μακροπρόθεσμες προβλέψεις, όπως είναι και η περίπτωσή που εξετάζει η εργασία, τα ιδανικότερα μοντέλα είναι τα μοντέλα παλινδρόμησης.

3.2.4.1 *Γραμμική παλινδρόμηση*

Το μοντέλο πρόβλεψης, που χρησιμοποιείται, για να προβλέπει κάθε μισή ώρα την τιμή της παραγόμενης ηλεκτρικής ενέργειας από μία φωτοβολταϊκή εγκατάσταση, είναι μοντέλο παλινδρόμησης, και συγκεκριμένα γραμμικής παλινδρόμησης.

Η παλινδρόμηση περιλαμβάνει διαδικασίες και τεχνικές μοντελοποίησης και ανάλυσης διαφόρων μεταβλητών, όπου το ζητούμενο είναι η εύρεση συσχετίσεων, μεταξύ μιας εξαρτημένης και μίας ή και περισσοτέρων ανεξάρτητων μεταβλητών [23]. Δηλαδή, η παλινδρόμηση παρουσιάζει την αλλαγή της τιμής της εξαρτημένης μεταβλητής καθώς μία ανεξάρτητη μεταβλητή μεταβάλλεται και όλες οι υπόλοιπες παραμένουν σταθερές.

Ένας ακόμα λόγος χρήσης της ανάλυσης της παλινδρόμησης, είναι και η εύρεση μιας εκτιμούμενης τιμής της εξαρτημένης μεταβλητής. Καθώς οι ανεξάρτητες μεταβλητές διατηρούνται σταθερές, είναι δυνατός ο υπολογισμός μίας μέσης τιμής της προσδοκώμενης εξαρτημένης μεταβλητής. Η εκτιμούμενη τιμή υπολογίζεται, πολύ εύκολα, μέσω μίας μαθηματικής σχέσης των ανεξάρτητων μεταβλητών, η οποία καλείται εξίσωση παλινδρόμησης.

Αν και η παλινδρόμηση χρησιμοποιείται ευρέως σε θέματα πρόβλεψης, ο κύριος λόγος χρησιμοποίησής της είναι η ανάλυση και κατανόηση των σχέσεων, μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών. Ειδικότερα, με τη βοήθεια συγκεκριμένων δεικτών, που προκύπτουν μέσα από την παλινδρόμηση, διαπιστώνεται ποιες, από τις εξεταζόμενες μεταβλητές, είναι συσχετισμένες επαρκώς με την ανεξάρτητη μεταβλητή, αλλά και τη μορφή αυτής της συσχέτισης. [23]

Υπάρχουν δύο μοντέλα παλινδρόμησης: η απλή γραμμική και η πολλαπλή γραμμική παλινδρόμηση.

Στη μέθοδο της απλής γραμμικής παλινδρόμησης, υποθέτουμε ότι η εξαρτημένη μεταβλητή (μεταβλητή πρόβλεψης) εξαρτάται γραμμικά από μία άλλη ανεξάρτητη μεταβλητή. Θεωρώντας Y την εξαρτημένη μεταβλητή και X την ανεξάρτητη, αυτή η εξάρτηση εκφράζεται από τη συνάρτηση παλινδρόμησης, η οποία έχει τύπο:

$$Y = a \cdot X + b$$

Αυτή είναι η εξίσωση της ευθείας. Επομένως στην απλή παλινδρόμηση, το σχήμα είναι μία ευθεία γραμμή, όπου α είναι το σημείο τομής της ευθείας με τον άξονα των εξαρτημένων μεταβλητών και b είναι η κλίση της ευθείας.

Στις περιπτώσεις που η εξαρτημένη μεταβλητή εξαρτάται από περισσότερες από μία ανεξάρτητες μεταβλητές, το μοντέλο της απλής παλινδρόμησης μπορεί να γενικευθεί, μέσω της τεχνικής της πολλαπλής παλινδρόμησης, ώστε να συμπεριλάβει όλες τις μεταβλητές, που επηρεάζουν την τιμή της μεταβλητής πρόβλεψης. Και πάλι, στην πολλαπλή παλινδρόμηση, θεωρούμε γραμμική σχέση ανάμεσα στην εξαρτημένη μεταβλητή και τις ανεξάρτητες και επιχειρείται η πρόβλεψη της τιμή της, βάσει των τιμών των υπολοίπων. Στη συγκεκριμένη εργασία, η ανεξάρτητη προβλεπόμενη μεταβλητή, είναι η παραγωγή ηλεκτρικής ενέργειας, μέσω του φωτοβολταϊκού, κάθε μισή ώρα και οι ανεξάρτητες μεταβλητές, είναι οι καιρικές συνθήκες.

Η γενική μορφή της πολλαπλής παλινδρόμησης είναι:

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \cdots + b_k X_k + e$$

Εξίσωση 1: Γενική εξίσωση Πολλαπλής παλινδρόμησης

Η μεταβλητή *Y* εκφράζει την εξαρτημένη μεταβλητή, ενώ οι μεταβλητές X_1, X_2, έως X_k εκφράζουν τις ανεξάρτητες μεταβλητές. Οι συντελεστές b_0, b_1, έως b_k είναι σταθερές παράμετροι. Ο σταθερός όρος b_0 είναι η τιμή του *Y*, αν όλες οι μεταβλητές X_i πάρουν την τιμή μηδέν. Ο συντελεστής b_i είναι η ποσότητα, κατά την οποία η εξαρτημένη μεταβλητή *Y* αυξάνεται, εάν αυξηθεί το X_i κατά μία μονάδα και όλες οι υπόλοιπες ανεξάρτητες μεταβλητές παραμείνουν αμετάβλητες. Τέλος το e δηλώνει τον τυχαίο παράγοντα, ο οποίος θεωρείται κανονικά κατανεμημένος γύρω από το μηδέν.

Αν θεωρηθεί ότι το *Y* εξαρτάται από δύο μεταβλητές *X*, τότε η μεταβλητή Y παριστάνεται στο επίπεδο.

Προχωρώντας την εξάρτηση του *Y* σε παραπάνω μεταβλητές, το σχήμα περιπλέκεται, και δημιουργείται ένα υπέρ-επίπεδο που είναι δύσκολο να περιγραφεί. Στην πράξη η πολλαπλή γραμμική παλινδρόμηση παίρνει την εξής μορφή:

$$Y = b_0 + b_1 \cdot X_{1,i} + b_2 \cdot X_{2,i} + \cdots + b_k X_{k,i} + e_i$$

Εξίσωση 2: Πολλαπλή γραμμική παλινδρόμηση

3.2.4.2 ***Υπολογισμός των συντελεστών απλής γραμμικής παλινδρόμησης***

Όπως αναφέρεται και στο επόμενο κεφάλαιο, το μοντέλο, που θα αναπτυχθεί, θα είναι η απλή γραμμική παλινδρόμηση. Προκειμένου να υπολογιστούν οι συντελεστές παλινδρόμησης, εφαρμόζεται η μέθοδος ελαχίστων τετραγώνων. Διατίθενται λοιπόν, n πραγματικές τιμές της εξαρτημένης μεταβλητής, επομένως, και n τιμές, που προκύπτουν από την εξίσωση παλινδρόμησης (y=ax+b) ($\widehat{Y}_i$). Τότε οι συντελεστές a, b προκύπτουν από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των διαφορών των τιμών Y_i, από τις τιμές $\widehat{Y}_i$, όπως φαίνεται και στον παρακάτω τύπο:

$$(a, b) \mid \min[\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2]$$

Εξίσωση 3: Μέθοδος ελαχίστων τετραγώνων

Μέσω της ευθείας παλινδρόμησης, υπολογίζονται οι εκτιμήσεις της εξαρτημένης μεταβλητής, δηλαδή το $\widehat{Y}_i$ και κατόπιν υπολογίζονται και οι αποκλίσεις e_i, ως η διαφορά $(Y_i - \widehat{Y}_i)$. Σύμφωνα με τη μέθοδο ελαχίστων τετραγώνων, ελαχιστοποιείται το άθροισμα των τετραγώνων των σφαλμάτων e_i. Με αυτό τον τόπο υπολογίζονται οι συντελεστές της συνάρτησης παλινδρόμησης Τελικά, οι συντελεστές a, b προκύπτουν ως εξής:

$$b = \frac{\frac{\sum_{i=1}^{n} X_i \cdot Y_i}{n} - \bar{X} \cdot \bar{Y}}{\frac{\sum_{i=1}^{n} X_i^2}{n} - \bar{X}^2}$$

Εξίσωση 4: Εύρεση Συντελεστής b απλής γραμμικής παλινδρόμησης

$$a = \bar{Y} - b \cdot \bar{X}$$

Εξίσωση 5: Εύρεση Συντελεστής a απλής γραμμικής παλινδρόμησης

Όπου $\bar{X}$ και $\bar{Y}$ οι μέσες τιμές των διανυσμάτων X, Y και n ο αριθμός των παρατηρήσεων (δεδομένων). [23]

3.2.5 Χρήση και αξιολόγηση των μοντέλων πρόβλεψης

Από τη στιγμή που έχει δημιουργηθεί το μοντέλο, αυτό χρησιμοποιείται για τον υπολγισμό των ζητούμενων προβλέψεων.

Η παρακολούθηση του προτύπου της χρονοσειράς και των σφαλμάτων της πρόβλεψης, συμβάλλει στην έγκαιρη αντιμετώπιση διορθωτικών αλλαγών και στην εξάλειψη προκατάληψης στις τελικές προβλέψεις.[23]

Ένας τρόπος για να διαπιστωθεί, αν το επιλεγμένο μοντέλο και οι παραχθείσες προβλέψεις είναι ικανοποιητικά, είναι ο χρόνος. Καθώς δηλαδή θα έρχονται οι καινούριες μετρήσεις της μεταβλητής *Y*, αυτές θα συγκρίνονται με τις εκτιμούμενες.

Ένας άλλος τρόπος άμεσης αξιολόγησης του μοντέλου, ο διαχωρισμός των δεδομένων σε δεδομένα προς εκπαίδευση και δεδομένα προς εξέταση. Αυτός ο διαχωρισμός θα πρέπει να γίνει κατά 80%-20%. Με τον τρόπο αυτό, εφαρμόζεται το μοντέλο, στα δεδομένα που προς εξέταση και συγκρίνονται άμεσα οιπροβλεπόμενες τιμές με τις πραγματικές.

Τέλος, η αξιολόγηση και μέτρηση της ακρίβειας του μοντέλου μπορεί να επιτευχθεί και με εξειδικευμένους στατιστικούς δείκτες σφαλμάτων. Από την προηγούμενη παράγραφο ορίστηκε το σφάλμα της πρόβλεψης, να περιγράφεται από τον τύπο

$$e_i = (Y_i - \widehat{Y}_\iota)$$

Εξίσωση 6: Σφάλμα πρόβλεψης

Όπου Y_i η πραγματική τιμή της προς πρόβλεψη παρατήρησης και $\widehat{Y}_\iota$ η τιμή της πρόβλεψης.

Επίσης διατίθενται n παρατηρήσεις (δεδομένα).

Οι στατιστικοί δείκτες σφάλματος, οι οποίοι παρέχουν διάφορες πληροφορίες, σχετικά με το μοντέλο, αναλύονται παρακάτω. [26]

- Μέσο σφάλμα (Mean error)

$$ME = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_\iota)$$

Εξίσωση 7: Μέσο σφάλμα

Όπως φαίνεται και από τον παραπάνω τύπο, το μέσο σφάλμα προκύπτει, ως ο μέσος όρος των σφαλμάτων και εκφράζει ένα μέτρο συστηματικότητας του σφάλματος. Όταν ο δείκτης λαμβάνει τιμές κοντά στο μηδέν, αυτό σημαίνει ότι τα σφάλματα χαρακτηρίζονται από τυχαιότητα. Οι θετικές τιμές, μεταφράζονται ως απαισιοδοξία στις προβλέψεις, ενώ οι αρνητικές ως αισιοδοξία.

- Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_\iota)^2$$

Εξίσωση 8: Μέσο τετραγωνικό σφάλμα

Ο δείκτης αποτελεί μέτρο της ακρίβειας του μοντέλου πρόβλεψης. Λόγω του τετραγώνου, επηρεάζεται έντονα από μεγάλα σφάλματα. Όσο πιο μικρό είναι το μέσο τετραγωνικό σφάλμα, τόσο πιο ακριβές είναι το μοντέλο. Οι μονάδες του είναι οι μονάδες της χρονοσειράς, υψωμένες στο τετράγωνο.

- Ρίζα του Μέσου Τετραγωνικού Σφάλματος (Root Mean Squared Error)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_\iota)^2}$$

Εξίσωση 9: Ρίζα του μέσου τετραγωνικού σφάλματος

Η ρίζα του μέσου τετραγωνικού σφάλματος, είναι απλά η τετραγωνική ρίζα του MSE. Σχηματικά, πρόκειται για την απόσταση ενός σημείου από την ευθεία, που ορίζει η εξίσωση απλής γραμμικής παλινδρόμησης. Το ιδιαίτερο χαρακτηριστικό αυτού του δείκτη, που τον καθιστά εξαιρετικά χρήσιμο, είναι ότι έχει τις ίδιες μονάδες με τη χρονοσειρά.

- Συντελεστής R^2 (Squared Correlation)

Η συσχέτιση των τιμών $\hat{Y}$ (οι οποίες προκύπτουν από την εξίσωση παλινδρόμησης) και των πραγματικών τιμών Y συμβολίζεται με R. Ο συντελεστής αυτός είναι πάντα θετικός, παίρνοντας τιμές στο διάστημα 0 έως 1. Το τετράγωνο του R καλείται squared correlation. Για τον υπολογισμό του R^2, χρησιμοποιείται η εξίσωση:

$$R^2 = \frac{ερμηνευθείσα\ διακύμανση\ των\ τιμών\ Y}{συνολική\ διακύμανση\ των\ τιμών\ Y} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Εξίσωση 10: Squared correlation

Εκφράζει το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής, το οποίο ερμηνεύεται από τη διακύμανση των τιμών της ανεξάρτητης. Υποδεικνύει .την ποιότητα προσαρμογής της γραμμής παλινδρόμησης στα δεδομένα. Δεν έχει μονάδα μέτρησης και το εύρος των τιμών του είναι από μηδέν έως ένα. Όσο πιο κοντά βρίσκεται στη μονάδα, τόσο μεγαλύτερο ποσοστό διακύμανσης της εξαρτημένης μεταβλητής εξηγείται από την παλινδρόμηση, δηλαδή τόσο μεγαλύτερη γραμμική συσχέτιση υπάρχει ανάμεσα στην εξαρτημένη με την ανεξάρτητη μεταβλητή. [36]

Όταν $R^2 = 0$, τότε η ευθεία παλινδρόμησης απεικονίζει επακριβώς τη σχέση μεταξύ της ανεξάρτητης με την εξαρτημένη μεταβλητή. Όταν η ευθεία παλινδρόμησης είναι παράλληλη στον άξονα της ανεξάρτητης μεταβλητής, τότε η κλίση της είναι και ο συντελεστής R^2 είναι επίσης μηδέν. Δηλαδή, δεν υπάρχει καθόλου γραμμική σχέση μεταξύ ανεξάρτητης και εξαρτημένης μεταβλητής. [31]

ΚΕΦΑΛΑΙΟ 4

Προ-επεξεργασία και Ανάλυση Δεδομένων

4.1 Εισαγωγή

Στο συγκεκριμένο κεφάλαιο θα παρουσιαστεί η ανάλυση δεδομένων και η σχεδίαση του μοντέλου πρόβλεψης στο Rapidminer.

Το RapidMiner διαθέτει ένα φιλικό περιβάλλον εργασίας για τον χρήστη, όπου οι αναλύσεις ρυθμίζονται σε μία *όψη διαδικασίας (process view)*. Χρησιμοποιεί μια σπονδυλωτή έννοια, όπου το κάθε βήμα μιας ανάλυσης (για παράδειγμα ένα βήμα προ-επεξεργασίας ή μια διαδικασία μάθησης) απεικονίζεται από έναν *τελεστή (*operator*)* στη διαδικασία ανάλυσης. Αυτοί οι τελεστές έχουν πύλες εισόδου και εξόδου, μέσω των οποίων επικοινωνούν με άλλους τελεστές, προκειμένου να λάβουν δεδομένα εισόδου ή να μεταδώσουν τα διαμορφωμένα δεδομένα ή τα μοντέλα, που έχουν δημιουργήσει, στους τελεστές που ακολουθούν. Αν διάφοροι τελεστές είναι διασυνδεδεμένοι, τότε μιλάμε για *διαδικασία (process).*

Όλοι οι τελεστές (operators) που είναι διαθέσιμοι στο RapidMiner παρουσιάζονται στις παρακάτω ομάδες:

Έλεγχος διαδικασίας (Process Control): Τελεστές όπως βρόχοι και κλάδοι υπό όρους, οι οποίοι μπορούν να ελέγξουν τη ροή της διαδικασίας.

Χρησιμότητα (Utility): Βοηθητικοί τελεστές, οι οποίοι μαζί με τον χειριστή "Subprocess" περιλαμβάνουν και σημαντικούς μακρο-τελεστές.

Πρόσβαση στις αποθήκες (Repository Access): Περιλαμβάνει τελεστές για την ανάγνωση και τη γραφή στις αποθήκες.

Εισαγωγή (Import): περιλαμβάνει ένα μεγάλο αριθμό τελεστών για την ανάγνωση δεδομένων, από εξωτερικές πηγές, όπως αρχεία, βάσεις δεδομένων κ.α.

Εξαγωγή (Export): Περιλαμβάνει ένα μεγάλο αριθμό τελεστών για τη γραφή δεδομένων σε αρχεία, βάσεις δεδομένων και άλλα.

Μετατροπή δεδομένων (Data Transformation): Ίσως είναι η σημαντικότερη ομάδα στην ανάλυση, όσον αφορά το μέγεθος και τη συνάφεια. Όλοι οι τελεστές, για τη μετατροπή δεδομένων και μετά-δεδομένων, βρίσκονται σε αυτή την ομάδα.

Μοντελοποίηση (Modelling): Περιλαμβάνει τις διαδικασίες εξόρυξης δεδομένων, όπως είναι η μέθοδος ταξινόμησης, η μέθοδος παλινδρόμησης, η ομαδοποίηση, οι σταθμίσεις, η συσχέτιση και η ανάλυση ομοιότητας, όπως και τελεστές, για να εφαρμόσει τα μοντέλα, που δημιουργήθηκαν σε καινούρια σύνολα δεδομένων.

Αξιολόγηση (Evaluation): Τελεστές, οι οποίοι μπορούν να υπολογίσουν την ποιότητα ενός μοντέλου.

Η χρήση των τελεστών είναι πολύ εύκολη, καθώς χρειάζεται απλά η επιλογή ενός από το «*Operators View»* και να η τοποθέτησή του στη διαδικασία με "drag and drop". Συνδέονται μεταξύ τους, με τη δημιουργία μιας γραμμή ανάμεσα στις πύλες εισόδου (input) και εξόδου (output) τους.

Ένας τελεστής ορίζεται από διάφορες παραμέτρους:

- Την περιγραφή των αναμενόμενων εισόδων
- Την περιγραφή των παρεχόμενων εξόδων
- Την ενέργεια, που εκτελείται από τον τελεστή στις εισόδους, η οποία τελικά οδηγεί στις εξόδους
- Έναν αριθμό παραμέτρων, οι οποίες ελέγχουν την ενέργεια, που εκτελείται

Οι παράμετροι αυτές, θα παρουσιαστούν για κάθε τελεστή που χρησιμοποιείται ξεχωριστά.

Ένα μεγάλο πλεονέκτημα του RapidMiner είναι, ότι προσφέρει μεγάλη ευελιξία και ελευθερία. Για παράδειγμα μια διαδικασία, η οποία έχει ήδη δημιουργηθεί, μπορεί να χρησιμοποιηθεί ξανά σε ένα παρόμοιο πρόβλημα. Ένα μοντέλο, το οποίο έχει παραχθεί, ήδη μία φορά, μπορεί να εφαρμοστεί ξανά, ώστε τα αποτελέσματα μιας ανάλυσης να αναζητήσουν τη μέθοδο, η οποία προσφέρει μέγιστη επιτυχία.

Τα βήματα, που θα ακολουθούνται για την διαδικασία παραγωγής και αξιολόγησης προβλέψεων, είναι ακριβώς ίδια με αυτά που ορίσθηκαν στο Κεφάλαιο 3.

4.2 Καθορισμός προβλήματος

Το πρόβλημα, είναι η ανάπτυξη μοντέλου πρόβλεψης παραγωγής ενέργειας, σε ΦΒ εγκατάσταση, μέσω ολοκληρωμένης ανάλυσης πολλαπλών ροών δεδομένων. Η μεταβλητή, που χρειάζεται να προβλεφθεί, είναι η παραγωγή του ΦΒ συστήματος.

4.3 Συλλογή δεδομένων

Τα ιστορικά δεδομένα που ήταν διαθέσιμα, αναφέρονταν σε διάστημα τριών χρόνων και περιλάμβαναν τιμές για την παραγωγή του ΦΒ καθώς και τιμές για τις καιρικές συνθήκες, που επηρέασαν τη παραγωγή, σε αυτό το διάστημα.

Το σύνολο των δεδομένων της παραγωγής περιλαμβάνει τιμές για την παραγόμενη ισχύ του ΦΒ και την ακτινοβολία, κάθε μία ώρα, για ένα εικοσιτετράωρο. Από την άλλη πλευρά, τα δεδομένα του καιρού, περιλαμβάνουν τιμές για τη θερμοκρασία, την

κατεύθυνση του αέρα, την υγρασία, τα σύννεφα, την κάλυψη του ουρανού και το σημείο δρόσου, κάθε μισή ώρα, από τις πέντε προ μεσημβρίας έως τις εννέα μετά μεσημβρίας.

4.4 Προετοιμασία χρονοσειρών

Το συγκεκριμένο βήμα, καλύφθηκε εξ' ολοκλήρου στο Rapidminer.

Η διαδικασία υλοποίησής του, αποτελείται από τα στάδια ένα έως τέσσερα όπως φαίνονται στο Σχήμα 12 (παράγραφος 3.2.1).

4.4.1 Συγχώνευση δεδομένων

4.4.1.1 *Ανάγνωση δεδομένων*

Πρώτη ενέργεια που χρειάστηκε, ήταν η ανάγνωση των δεδομένων. Τα δύο σύνολα δεδομένων δόθηκαν από το σύστημα, σε μορφή αρχείων CSV. Τα CSV (Comma-Separated Values) αρχεία περιέχουν δεδομένα (αριθμούς και κείμενο), στη μορφή απλού κειμένου. Κάθε γραμμή του αρχείου αποτελεί μία καταχώρηση δεδομένων. Κάθε καταχώρηση διαθέτει ένα ή και περισσότερα πεδία, χωρισμένα με κόμμα.

Το πρώτο αρχείο «δεδομένα παραγωγής ενέργειας» περιέχει την παραγωγή ηλεκτρικής ενέργειας από το φωτοβολταϊκό και την ακτινοβολία του ήλιου, κάθε μία ώρα, για ένα εικοσιτετράωρο, από 2/1/2012 έως 1/12/2014.

Λόγω του τεράστιου όγκου των δεδομένων, θα παρουσιαστεί μόνο ένα μέρος του αρχείου. Τα δεδομένα του πρώτου πεδίου αποτελούν την ακτινοβολία, του δεύτερου πεδίου την παραγωγή και το τρίτο πεδίο είναι η ημερομηνία με την ώρα. Για παράδειγμα, η πρώτη καταχώρηση, φανερώνει ότι, στις 2 Ιανουαρίου 2012 και ώρα 00:00 είχαμε ακτινοβολία 0,426 και παραγωγή 0. Ενώ, στις 2 Ιανουαρίου 2012 και ώρα 10:00 π.μ. η ακτινοβολία περιγράφεται από τον αριθμό 193,983 και η παραγωγή από τον αριθμό 1211,455. Φανερώνεται λοιπόν μία πολύ μεγάλη διαφορά ανάμεσα σε αυτές τις δύο καταχωρήσεις, παρότι αναφέρονται στην ίδια ημέρα. Αυτό συμβαίνει διότι, η πρώτη καταχώρηση είναι το βράδυ, όπου φυσικά η ηλιακή ακτινοβολία είναι σχεδόν μηδενική, άρα δεν έχουμε και παραγωγή ηλεκτρικής ενέργειας, ενώ η δεύτερη καταχώρηση είναι το πρωί, όπου η ηλιακή ακτινοβολία είναι σε αρκετά υψηλά επίπεδα. Διαφορά επίσης υπάρχει στην παραγωγή από μήνα σε μήνα.

0.426,0,2012-01-02T00:00:00Z,
1.016,0,2012-01-02T01:00:00Z,
0.968,0,2012-01-02T02:00:00Z,
1.096,0,2012-01-02T03:00:00Z,
1.033,0,2012-01-02T04:00:00Z,
0.979,0,2012-01-02T05:00:00Z,
0.995,0,2012-01-02T06:00:00Z,
1.022,0,2012-01-02T07:00:00Z,
1.288,0,2012-01-02T08:00:00Z,
23.335,97.386,2012-01-02T09:00:00Z,
193.983,1211.455,2012-01-02T10:00:00Z,

Σχήμα 14 Δεδομένα παραγωγής ΦΒ συστήματος

Όπως φαίνεται και από τον πίνακα με τα δεδομένα, οι μεταβλητές ακτινοβολία και παραγωγής παίρνουν σαν τιμές πραγματικούς αριθμούς.

Το δεύτερο αρχείο «δεδομένα καιρού» περιέχει τα δεδομένα, τα οποία δίνονταν από το σύστημα, κάθε μισή ώρα, από τις πέντε προ μεσημβρίας, μέχρι τις εννέα μετά μεσημβρίας σχετικά με τις καιρικές συνθήκες, από 1/1/2012 έως 30/12/2014.

datetime	tmpf	dwpf	relh	drct	sknt	clouds
2012-01-01T05:00:00Z						0
2012-01-01T05:30:00Z						0
2012-01-01T06:00:00Z						0
2012-01-01T06:30:00Z	5	5	100	290	154.332	0
2012-01-01T07:00:00Z	6	5	93.3	300	154.332	1
2012-01-01T07:30:00Z	5	4	93.24	300	154.332	1
2012-01-01T08:00:00Z	6	5	93.3	0	0	1
2012-01-01T08:30:00Z	8	7	93.4	0	0	1
2012-01-01T09:00:00Z	10	8	87.37	0	0	1
2012-01-01T09:30:00Z	12	8	76.5	0	0	1

Σχήμα 15 Δεδομένα καιρού

Στην πρώτη στήλη, φαίνεται η ημερομηνία μαζί με την ώρα (datetime), στη δεύτερη στήλη η θερμοκρασία (temperature) "tmpf" σε βαθμούς Κελσίου, στην τρίτη στήλη το σημείο δρόσου πάλι σε βαθμούς Κελσίου (dewpoint) "dwpf", στην τέταρτη στήλη η υγρασία (relative humidity) "relh" σε επί τοις εκατό τιμές, στην πέμπτη στήλη η κατεύθυνση του ανέμου σε μοίρες (direction) "drct", στην έκτη στήλη η κάλυψη του ουρανού (sky coverage) "sknt" και στην τελευταία στήλη τα σύννεφα "clouds".

Όπως φαίνεται, και από τον πίνακα, το σύστημα δίνει για τη θερμοκρασία το σημείο δρόσου, την κατεύθυνση του ανέμου και τα σύννεφα, ακέραιες τιμές.

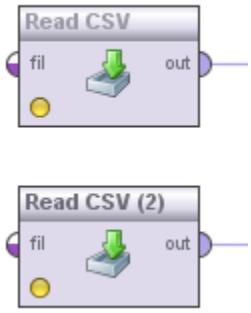
Συγκεκριμένα, για τα σύννεφα, οι τιμές που εμφανίζονται, είναι αριθμοί από το 0 έως το 9. Αυτοί οι αριθμοί ορίζουν, πόσο καλυμμένος είναι ο ουρανός από σύννεφα, σε μία κλίμακα, που ονομάζεται οχτάρια (okta). Αυτή η κλίμακα περιγράφει την ποσότητα της κάλυψης από σύννεφα, σε οποιαδήποτε τοποθεσία. Ξεκινώντας από το μηδέν (απόλυτα καθαρός ουρανός) και καταλήγοντας στο οχτώ (απόλυτα συννεφιασμένος). Επίσης, πολλές φορές χρησιμοποιείται άλλη μία μονάδα, το εννέα που σημαίνει, πως δεν φαίνεται καθόλου ο ουρανός, συνήθως εξαιτίας υψηλών επιπέδων ομίχλης ή πολύ χιονιού. [34]

Αντίστοιχα, για τις μεταβλητές υγρασία και κάλυψη του ουρανού, οι τιμές είναι πραγματικές.

Μετά τη συλλογή των δεδομένων, παρατηρείται ότι αυτά αναφέρονται σε διαφορετικές χρονικές περιόδους, συγκεκριμένα η μεταβλητή «δεδομένα παραγωγής ενέργειας», διαθέτει δεδομένα από 2/1/2012 έως 1/12/2014, ενώ η μεταβλητή «δεδομένα καιρού», από 1/1/2012 έως 30/12/2014. Αυτό συνεπάγεται, ότι το πρώτο υπολείπεται δεδομένων, σχετικά με τις ημερομηνίες. Η δεύτερη διαφορά είναι, ότι η μεταβλητή «δεδομένα παραγωγής ενέργειας», διαθέτει τιμές, κάθε μία ώρα, ενώ η «δεδομένα καιρού», κάθε μισή. Οπότε και πάλι το πρώτο αρχείο υπολείπεται δεδομένων. Παρατηρώντας λίγο πιο προσεκτικά τα δεδομένα, εμφανίζεται άλλη μια διαφορά, η οποία αφορά τις ώρες, κατά τις οποίες λαμβάνουμε τα δεδομένα. η πρώτη μεταβλητή δίνει δεδομένα για ένα εικοσιτετράωρο, ενώ η δεύτερη, από τις πέντε προ μεσημβρίας, έως τις εννέα μετά μεσημβρίας.

Για να αντιμετωπιστούν αυτές οι διαφορές μεταξύ των δύο συνόλων, θα χρησιμοποιηθούν κάποιους τελεστές στη συνέχεια του κεφαλαίου.

Όσον αφορά στην υλοποίηση της ανάγνωσης δεδομένων, μέσω του RapidMiner, χρησιμοποιείται δύο φορές ο ίδιος τελεστής, προκειμένου να διαβαστούν τα δύο αρχεία. Ο τελεστής ονομάζεται *"Read CSV"* και ανήκει στους τελεστές Εισαγωγής (Import) – Δεδομένα (Data). Το αρχείο, που θα διαβάσει κάθε τελεστής ορίζεται, τοποθετώντας το σωστό μονοπάτι, που οδηγεί σε αυτό. Επίσης, προκειμένου να διαβάσει ο τελεστής, σωστά, τα δεδομένα, από το παράθυρο των παραμέτρων, ορίζουμε τον τύπο κάθε μεταβλητής.



Σχήμα 17 Τελεστές "Read CSV"

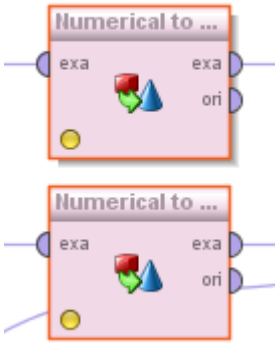
Σχήμα 16 Παράμετροι "Read CSV"

Σχήμα 18 Παράμετροι "Read CSV (2)"

Με αυτό τον τρόπο, στην έξοδο του κάθε τελεστή, εμφανίζονται τα δεδομένα των αρχείων, όπως ακριβώς δόθηκαν από το σύστημα, χωρισμένα σε στήλες, όπου κάθε στήλη είναι και μία μεταβλητή.

4.4.1.2 *Κωδικοποίηση δεδομένων - Μετατροπή τύπου τιμών*

Αυτό το βήμα εκτέλεσης του προγράμματος, αφορά τη μετατροπή των αριθμητικών τιμών σε πραγματικές. Θέλουμε κυρίως, να μετατρέψουμε τις ακέραιες μεταβλητές σε πραγματικές. Αυτό το κάνουμε διότι, κάποιοι τελεστές, που θα χρησιμοποιήσουμε στη συνέχεια, δεν μπορούν να διαχειριστούν αριθμητικές τιμές γενικά, παρά μόνο πραγματικές. Χρησιμοποιείται ο ίδιος τελεστής, δύο φορές (ένας για κάθε αρχείο), ο οποίος ονομάζεται *"Numerical to Real"*. Ανήκει στους τελεστές Μετατροπή δεδομένων (Data Transformation) – Μετατροπή τύπου (Type conversion). Και σε αυτή την περίπτωση από τις παραμέτρους του, γίνεται η επιλογή της μετατροπής όλων των αριθμητικών τιμών – στην περίπτωση των δεδομένων που διαθέτουμε, τις ακέραιες - σε πραγματικές. Επίσης, σε περίπτωση που λείπει μία τιμή, η καινούρια πραγματική τιμή εξακολουθεί να λείπει.



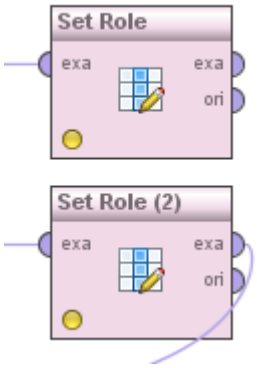
Σχήμα 19 Τελεστές "Numerical to Real"

Ως είσοδο (πύλη "exa") δέχεται ένα σύνολο δεδομένων, στην περίπτωσή η παρούσας ανάλυσης, τι μεταβλητές «δεδομένα παραγωγής ενέργειας», «δεδομένα καιρού» και δίνει σαν έξοδο (πύλη "exa") το αρχείο με τις αριθμητικές μεταβλητές, που έχουν μετατραπεί σε πραγματικές.

4.4.1.3 ***Ορισμός ρόλου***

Στο σημείο αυτό, χρησιμοποιείται ο τελεστής *"Set Role"* από τους Μετατροπή δεδομένων (Data Transformation) – Τροποποίηση ονόματος και ρόλου (Name and role modification), προκειμένου να ορισθεί ο ρόλος μίας μεταβλητής.

Ο ρόλος μίας μεταβλητής αναπαριστά το ρόλο, που λαμβάνει η συγκεκριμένη μεταβλητή στο σύνολο δεδομένων. Υπάρχουν δύο είδη μεταβλητών: αυτές, που απλά περιγράφουν ένα παράδειγμα «απλές» (regular) και αυτές, που ορίζουν κάθε παράδειγμα ξεχωριστά «σημαντικές» (special). Οι μεταβλητές επίσης μπορούν να υιοθετήσουν διάφορους ρόλους. Για παράδειγμα, οι σημαντικές μεταβλητές, μπορούν να είναι "ID", οι οποίες ορίζουν ξεκάθαρα το παράδειγμα που μας ενδιαφέρει. Επιπλέον, μπορούν να είναι "weight" οι οποίες χαρακτηρίζουν το βάρος κάθε παραδείγματος. Αλλάζοντας τον ρόλο μίας μεταβλητής, αλλάζει και ο ρόλος της σε μία διαδικασία. Κάθε μεταβλητή, μπορεί να έχει ακριβώς έναν ρόλο. Κάθε σύνολο δεδομένων, που χρησιμοποιείται σε μία διαδικασία, μπορεί να έχει πολλές σημαντικές μεταβλητές, αλλά ένας «σημαντικός» ρόλος δεν μπορεί να επαναληφθεί.



Σχήμα 21 Τελεστής "Set role"

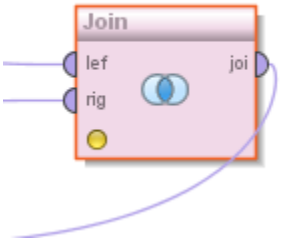
Σχήμα 20 Παράμετρος "Set role"

Όπως φαίνεται και στο Σχήμα 20, από τις παραμέτρους του τελεστή ορίζεται, και για τα δύο αρχεία, η μεταβλητή "*datetime*" ως ID. Με αυτό τον τρόπο, η συγκεκριμένη μεταβλητή μετατρέπεται στο αναγνωριστικό του συνόλου δεδομένων. Αυτή η μετατροπή της μεταβλητής datetime από «απλή» σε ID χρησιμεύει στον τελεστή, που ακολουθεί.

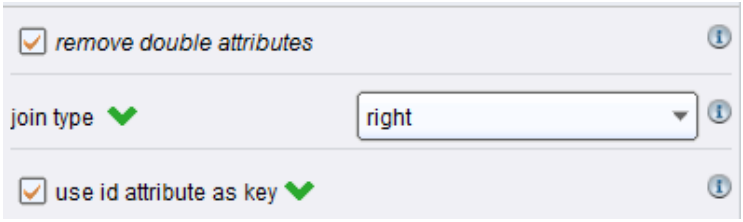
Κάθε τελεστής δέχεται ως είσοδο (πύλη "exa") τα αρχεία, που έχουν διαβαστεί και δίνει σαν έξοδο (πύλη "exa") το σύνολο δεδομένων κάθε αρχείου, με αλλαγμένο το ρόλο της μεταβλητής datetime.

4.4.1.4 *Συγχώνευση δεδομένων*

Μέχρι αυτό το σημείο, έχει πραγματοποιηθεί η κωδικοποίηση των δύο αρχείων ξεχωριστά. Για τη δημιουργία του μοντέλου πρόβλεψης, είναι απαραίτητη η ένωση των δύο αρχείων σε ένα ενιαίο και ο ορισμός των παρατηρήσεων με βάση μία κοινή μεταβλητή. Αυτό επιτυγχάνεται με τον τελεστή *"Join"*, από τους τελεστές Μετατροπή δεδομένων (Data Transformation) – Ορισμός λειτουργιών (Set Operations).



Σχήμα 23 Τελεστής Join



Σχήμα 22 Παράμετροι τελεστή "Join"

Ο συγκεκριμένος τελεστής δέχεται στις δύο πύλες εισόδου τα δύο αρχεία. Στην παρούσα διαδικασία, στην πύλη "lef", δέχεται τη μεταβλητή «δεδομένα παραγωγής ενέργειας», ενώ στην "rig" τη «δεδομένα καιρού».

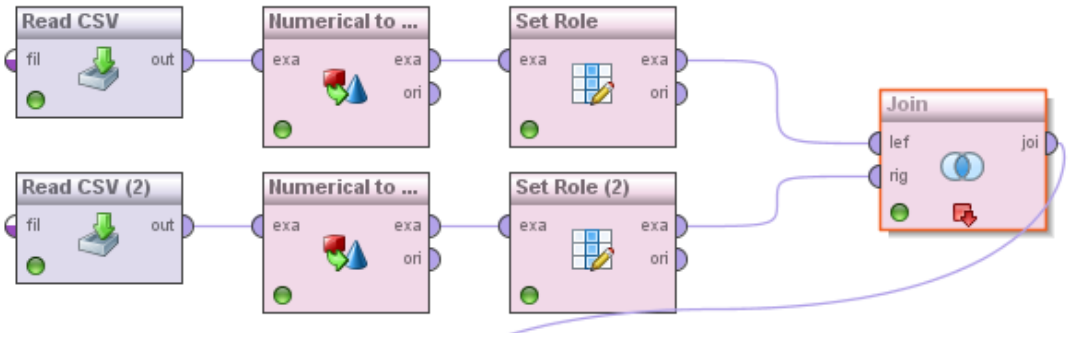
Αυτός ο τελεστής ενώνει δύο σύνολα δεδομένων, χρησιμοποιώντας μεταβλητές-κλειδιά. Ταυτόσημες τιμές, των μεταβλητών-κλειδιά, υποδεικνύουν ταιριαστά σύνολα. Μία μεταβλητή, η οποία έχει ρόλο ID, επιλέγεται αυτόματα ως μεταβλητή-κλειδί. Η μόνη κοινή μεταβλητή, που διαθέτουν τα δύο αρχεία, είναι η ημερομηνία/ώρα (dateime) και γι αυτό το λόγο΄, στον προηγούμενο τελεστή, ορίσθηκε και στα δύο αρχεία, ως ID , ώστε να χρησιμοποιηθεί αυτόματα ως μεταβλητή-κλειδί.

Η δυσκολία είναι, πως από το αρχείο «δεδομένα παραγωγής ενέργειας», λαμβάνονται τιμές, κάθε μία ώρα, ενώ από το «δεδομένα καιρού», κάθε μισή και, το δεύτερο δίνει δεδομένα για περισσότερες ημέρες αλλά λιγότερες ώρες. Για την αντιμετώπιση αυτών των διαφορών, υπάρχει η δυνατότητα διαγραφής των τιμών των μεταβλητών, που εμφανίζονται στις και μισή κάθε ώρας και όσων δεν εμφανίζονται τις κοινές ημέρες και ώρες. Με αυτό τον τρόπο όμως θα χάνονταν πολλά και σημαντικά δεδομένα. Επομένως, επιλέγεται η διατήρηση όλων των δεδομένων και κατά τη συγχώνευση των δύο, να εμφανίζεται ένα σύνολο δεδομένων, το οποίο θα δίνει τιμές, κάθε μισή ώρα από τις πέντε προ μεσημβρίας έως τις εννέα μετά μεσημβρίας για τις ημέρες από 1/1/2012 έως 30/12/2014. Αυτός ο τρόπος συγχώνευσης, επιτυγχάνεται επιλέγοντας από τις παραμέτρους του τελεστή την επιλογή join type – "right". Έτσι διατηρήθηκαν όλα τα δεδομένα (δηλαδή όλες οι τιμές datetime) του δεύτερου αρχείου και η συγχώνευση έγινε με βάση αυτό.

Ένα καινούριο πρόβλημα, που εμφανίζεται από αυτή την επιλογή, είναι ότι στα δεδομένα «δεδομένα παραγωγής ενέργειας» (radiation και power), θα εμφανιστούν κενές τιμές, δηλαδή ερωτηματικά «?», στις σειρές, όπου η τιμή της ώρας στην μεταβλητή datetime είναι και μισή και στις ημέρες που λείπουν.

Αυτό το πρόβλημα θα αντιμετωπιστεί παρακάτω με κατάλληλους τελεστές..

Σε αυτό το σημείο, γραφικά, η διαδικασία, στην *Προοπτική Σχεδιασμού* του Rapidminer, είναι η ακόλουθη (Στάδιο 1):



Σχήμα 24 Διαδικασία μέχρι τελεστή "Join" (Στάδιο 1, συγχώνευση δεδομένων)

Επίσης, το σύνολο δεδομένων, που προκύπτει μετά την συγχώνευση, δηλαδή κατά την έξοδο αυτού του σταδίου, στο RapidMiner, θα έχει την παρακάτω μορφή:

datetime	radiation	power	tmpf	dwpf	relh	drct	sknt	clouds
2012-01-01T19:30:00Z	?	?	?	?	?	?	?	0
2012-01-01T20:00:00Z	?	?	?	?	?	?	?	0
2012-01-01T20:30:00Z	?	?	?	?	?	?	?	0
2012-01-01T21:00:00Z	?	?	?	?	?	?	?	0
2012-01-02T05:00:00Z	0.979	0	?	?	?	?	?	0
2012-01-02T05:30:00Z	?	?	?	?	?	?	?	0
2012-01-02T06:00:00Z	0.995	0	?	?	?	?	?	0
2012-01-02T06:30:00Z	?	?	11	6	71.260	280	411.552	0
2012-01-02T07:00:00Z	1.022	0	11	5	66.480	270	25.722	0
2012-01-02T07:30:00Z	?	?	11	6	71.260	290	462.996	1
2012-01-02T08:00:00Z	1.288	0	12	6	66.700	280	462.996	1
2012-01-02T08:30:00Z	?	?	12	6	66.700	260	360.108	1
2012-01-02T09:00:00Z	23.335	97.386	13	7	66.910	260	360.108	1
2012-01-02T09:30:00Z	?	?	13	7	66.910	270	25.722	1
2012-01-02T10:00:00Z	193.983	1211.455	14	7	62.690	270	360.108	1

Σχήμα 25 Δεδομένα μετά τον τελεστή "Join"

4.4.2 Διαχείριση κενών και μηδενικών τιμών

Σε αυτή παρουσιάζεται το τέταρτο στάδιο της υλοποίησης, σύμφωνα με το Σχήμα 12

4.4.2.1 *Αντικατάσταση ελλιπών τιμών*

Όπως φαίνεται και από το Σχήμα 25, το σύνολο των δεδομένων διαθέτει πολλές ελλιπείς τιμές. Αυτές οι τιμές, είτε δημιουργήθηκαν από την συγχώνευση του προηγούμενου σταδίου, είτε προϋπήρχαν, αφού δόθηκαν έτσι από το σύστημα.

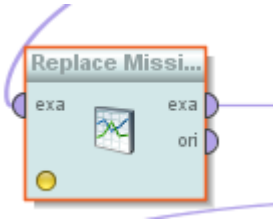
Όπως αναφέρθηκε και στην παράγραφο 3.7.3.4 υπάρχουν διάφοροι τρόποι, για την αντιμετώπιση των κενών τιμών.

- Διαγραφή όλων των παραδειγμάτων (σειρών), όπου λείπουν οι τιμές. Η εφαρμογή αυτής της μεθόδου θα ήταν καταστροφική για την ανάλυσή, καθώς θα διαγράφονταν σχεδόν τα μισά δεδομένα.
- Εκτίμηση ενός ύψους, στο οποίο κυμάνθηκαν οι κενές τιμές. Εξαιτίας όμως του μεγάλου όγκου των δεδομένων, κάτι τέτοιο θα ήταν υπερβολικά χρονοβόρο και μη αποτελεσματικό για την διαδικασία. Αν αναλογιστούμε, πως στο σύνολο διαθέτουμε περίπου 26.000 τιμές (17 ώρες- κάθε μισή ώρα- κάθε ημέρα για 3 χρόνια), για 8 μεταβλητές, και πως τα δεδομένα μας χαρακτηρίζονται από εποχιακή συμπεριφορά (δηλαδή το γέμισμα όλων των τιμών με μία ίδια θα ήταν πολύ

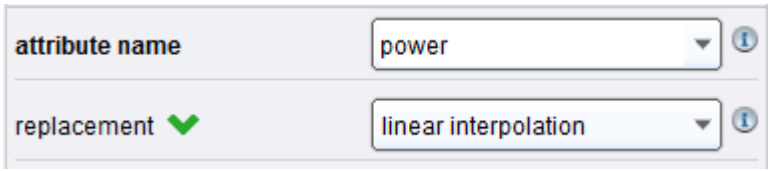
δύσκολο), τότε καταλαβαίνουμε πως το γέμισμα των κενών τιμών από την μεριά μας θα ήταν σχεδόν αδύνατο.

- Αφού η χρονοσειρά χαρακτηρίζεται από εποχιακή συμπεριφορά, υπάρχει η δυνατότητα ορισμού των κενών τιμών, ως ο μέσος όρος των τιμών των αντίστοιχων περιόδων. Κάτι τέτοιο θα ήταν για ακόμα μία φορά σχεδόν αδύνατο, λόγω του μεγάλου όγκου των δεδομένων. Οι ελλιπείς τιμές είναι σίγουρα περισσότερες από 13.000 (26.000/2), διότι 13.000 είναι μόνο οι κενές τιμές των μεταβλητών «παραγωγή» και «ακτινοβολία». Επομένως, η εύρεση του μέσου όρου των τιμών των αντίστοιχων περιόδων, για κάθε κενή τιμή, θα αύξανε την πολυπλοκότητα και αποτελεσματικότητα της διαδικασίας μας. Η υψηλή πολυπλοκότητα είναι ανεπιθύμητη, διότι καθιστά την διαδικασία αργή και μη κατανοητή από τρίτους.

Για την επίλυση του προβλήματος των κενών τιμών και τη συνέχιση της ανάλυσης, χρησιμοποιείται ο τελεστής *"Replace Missing Values"*, ο οποίος βρίσκεται στους τελεστές Μετατροπή Δεδομένων (Data Transformation) – Σειρές (Series). Ο συγκεκριμένος τελεστής γεμίζει τις κενές τιμές στις μεταβλητές του, σύμφωνα με τους εξής τρόπους: με καθορισμένη τιμή, με την προηγούμενη ή την επόμενη τιμή, εφαρμόζοντας γραμμική παρεμβολή στις γειτονικές τιμές.
Οι παραπάνω επιλογές φαίνονται στο παράθυρο των παραμέτρων.



Σχήμα 27 Τελεστής "Replace missing Values"



Σχήμα 26 Παράμετροι τελεστή "Replace missing values"

Δέχεται σαν είσοδο (πύλη "exa") ένα σύνολο δεδομένων και δίνει, ως έξοδο (πύλη "exa"), το ίδιο σύνολο, με γεμάτα τα ελλιπή δεδομένα.

Επιλέγεται το γέμισμα των κενών τιμών, με τη μέθοδο της γραμμικής παρεμβολής. Επιλέχθηκε αυτή η μέθοδος, διότι οι τιμές κάθε μεταβλητής λαμβάνονται ανά μισή ώρα, δηλαδή δύο γειτονικές τιμές, μεταξύ τους, έχουν πολύ μικρή διαφορά, στην ώρα και άρα πολύ μικρή διαφορά και στην τιμή. Αν και δεν είναι απόλυτα ακριβής ο συγκεκριμένος τρόπος, είναι ο πιο ασφαλής συγκριτικά με τους άλλους δύο. Αν για παράδειγμα, επιλέγαμε την αντικατάσταση των κενών τιμών με μία καθορισμένη τιμή, τότε θα εμφανιζόταν ένας πολύ μεγάλος παράγοντας σφάλματος, καθώς θα έπρεπε να γεμίσουν με την ίδια τιμή δεδομένα, τα οποία χαρακτηρίζονται από εποχιακή συμπεριφορά. Δηλαδή, θα έπρεπε να

ορισθεί το ίδιο ύψος παραγωγής και για χειμερινές και για καλοκαιρινές ημέρες, για παράδειγμα. Αντίστοιχα, αν επιλέγαμε να γεμίσουν με την επόμενη ή την προηγούμενη τιμή, θα υπήρχε μεγάλη πιθανότητα αντιστοίχισης της τιμής της ακτινοβολίας, στις εννέα μετά μεσημβρίας, με την τιμή της στις εννέα προ μεσημβρίας.

Μία άλλη παράμετρος, που απαιτεί προσοχή στον συγκεκριμένο τελεστή, είναι η μεταβλητή, της οποίας οι τιμές θα οριστούν. Όπως φαίνεται και στο Σχήμα 26, στο σημείο "attribute name" (όνομα μεταβλητής) έχει τοποθετηθεί η μεταβλητή "power", δηλαδή η παραγωγή του ΦΒ. Με αυτό τον τρόπο, οι κενές τιμές της μεταβλητής "power" θα γεμίσουν μέσω της γραμμικής παρεμβολής.

Το ίδιο ακριβώς εφαρμόστηκε και για τις υπόλοιπες έξι μεταβλητές (ακτινοβολία, κατεύθυνση του ανέμου, υγρασία, σημείο δρόσου, κάλυψη ουρανού και θερμοκρασία), οι οποίες περιέχουν ελλιπή δεδομένα. Επομένως, τοποθετήθηκαν στη σειρά άλλοι έξι ίδιοι τελεστές ("Replace missing value"), με τη μόνη διαφορά, ότι στο παράθυρο των παραμέτρων τους ορίζεται το όνομα της μεταβλητής που εξετάζεται κάθε φορά.

Η μεταβλητή σύννεφα δεν περιέχει κενές τιμές, οπότε δεν χρειάστηκε επιπλέον χρήση του τελεστή.

Μέχρι αυτό το σημείο, έχουν χρησιμοποιηθεί εφτά τελεστές "Replace Missing value" και έχουν γεμίσει οι κενές τιμές με γραμμική παρεμβολή. Προκειμένου όμως να επιτευχθεί γραμμική παρεμβολή, είναι υποχρεωτική η ύπαρξη γειτονικών τιμών.

Όπως έχει αναφερθεί παραπάνω, μετά τον τελεστή "Join" τα δύο αρχεία έγιναν ένα σύμφωνα με το δεύτερο, δηλαδή διατηρήθηκαν όλες οι τιμές της μεταβλητής "datetime" του δεύτερου αρχείου. Επομένως, για την ημέρα 01/01/2012 και όλο τον μήνα Δεκέμβριο του 2014, οι μεταβλητές "power" και "radiation" εξακολουθούν να μην έχουν τιμές, καθώς είναι τα άκρα του συνόλου δεδομένων. Από τη στιγμή λοιπόν, που η μέθοδος της γραμμικής παρεμβολής δεν μπορεί να εφαρμοστεί, τότε επιλέγουμε να μην γεμίσουμε τις τιμές, έτσι ώστε, με έναν επόμενο τελεστή, να διαγραφεί το συγκεκριμένο υποσύνολο των δεδομένων. Η απώλεια του συγκεκριμένου υποσυνόλου δεν επηρεάζει την ανάλυση καθώς αποτελεί ένα πολύ μικρό κομμάτι του μεγάλου όγκου των δεδομένων που διατίθενται.

4.4.2.2 ***Φιλτράρισμα παρατηρήσεων***

Από την παράγραφο 4.4.5, έχουν παραμείνει κάποιες κενές τιμές, στα άκρα του συνόλου δεδομένων, για τις μεταβλητές της παραγωγής (power) και της ακτινοβολίας (radiation). Όπως αναφέρθηκε και στην προηγούμενη παράγραφο, αυτές οι κενές τιμές, επιλέγεται να διαγραφούν τελείως από τα δεδομένα, καθώς είναι πολύ λίγες συγκριτικά με τον συνολικό όγκο. Για την πραγματοποίηση αυτής της ενέργειας, απαιτείται ένας τελεστής, ο οποίος θα φιλτράρει τα δεδομένα και θα διαγράψει αυτά, που δεν χρειάζονται. Ένας τέτοιος τελεστής είναι ο "Filter Examples", ο οποίος βρίσκεται στο σύνολο των τελεστών Μετατροπή Δεδομένων (Data Transformation) – Φιλτράρισμα (Filtering).

Ο συγκεκριμένος τελεστής επιλέγει, ποιες παρατηρήσεις (δηλαδή γραμμές) ενός συνόλου δεδομένων θα διατηρηθούν και ποιες θα απομακρυνθούν. Οι παρατηρήσεις, οι οποίες ικανοποιούν τις δοθείσες συνθήκες, διατηρούνται, ενώ οι υπόλοιπες διαγράφονται.

Στην διαδικασία που υλοποιείται, είναι απαραίτητα η διαγραφή κενών τιμών από τις μεταβλητές "power" και "radiation".

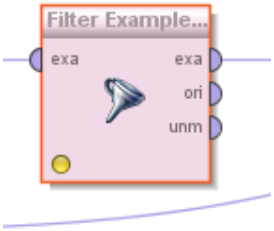
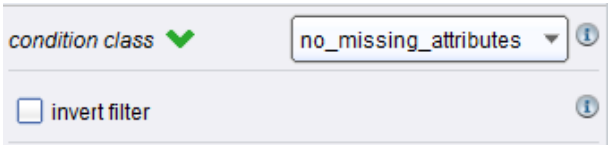


Figure 29 Τελεστής "Filter Examples"



Σχήμα 28 Παράμετροι τελεστή "Filter Examples" για τη διαχείριση κενών τιμών

Στην είσοδό του τελεστή (πύλη "exa"), εισέρχεται το σύνολο δεδομένων, και επιστρέφεται, από την έξοδο (πύλη "exa"), ένα καινούριο σύνολο δεδομένων, το οποίο περιλαμβάνει μόνο τις παρατηρήσεις (γραμμές), που ικανοποιούν τη συνθήκη, που ορίσθηκε στο παράθυρο των παραμέτρων.

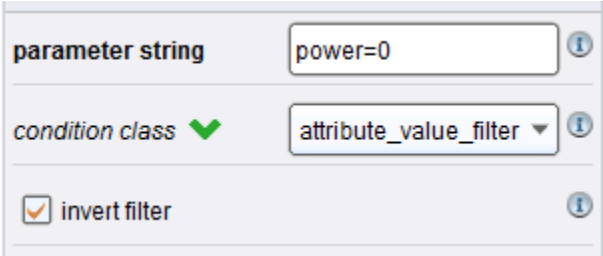
Επιλέγεται η διατήρηση των παρατηρήσεων, που δεν περιέχουν κενές τιμές. (Σχήμα 29)

Τελικά, στην έξοδο του δεύτερου τελεστή "Filter Examples", διατηρείται το σύνολο δεδομένων, με τιμές που ξεκινάνε από τις 02/01/2012 και καταλήγουν την 01/12/2014.

4.4.2.3 *Χειρισμός μηδενικών τιμών*

Από την εμφάνιση των δεδομένων που υπάρχουν μέχρι αυτό το σημείο, παρατηρείται ότι, η παραγωγή μηδενίζεται τις βραδινές ώρες. Αυτές οι μηδενικές τιμές θα επηρεάσουν το μοντέλο και τη διαδικασία εξαγωγής των βαρών. Καθώς αυτές οι μηδενικές τιμές θα επηρεάσουν την αποτελεσματικότητα του μοντέλου, πρέπει να αφαιρεθούν. Με αυτή την ενέργεια όμως θα χαθούν δεδομένα. η συγκεκριμένη διαγραφή δεν επηρεάζει την ανάλυση καθώς και στο μέλλον, η παραγωγή εξακολουθεί να είναι μηδέν το βράδυ, επομένως δεν χρειάζεται πρόβλεψη για αυτές τις ώρες.

Αυτή η διαδικασία θα επιτευχθεί με έναν ακόμα τελεστή "Filter Examples". Η ανάλυσή του είναι ίδια με αυτή που εμφανίζεται στην παράγραφο 4.4.2.2., με διαφορετικές παραμέτρους.



Σχήμα 30 Παράμετροι τελεστή "Filter Examples" (2)

Στην παράμετρο "condition class", ορίζεται η συνθήκη, την οποία πρέπει να ικανοποιούν τα παραδείγματα, για να διατηρηθούν. Επιλέγεται η "attribute_value_filter" (φίλτρο τιμής μεταβλητής). Με αυτή την επιλογή εμφανίζεται και η παράμετρος "parameter string" (συμβολοσειρά παραμέτρου). Σε αυτή την παράμετρο ορίζεται η τελική συνθήκη, η οποία είναι: "power=0". Επομένως, διατηρούνται μόνο τα παραδείγματα (σειρές), τα οποία έχουν παραγωγή ίση με μηδέν. Δηλαδή, επιλέγονται μόνο οι βραδινές ώρες. Εμείς όμως επιθυμούμε να τις διαγράψουμε αυτές τις σειρές. Όπως βλέπουμε και στις παραμέτρους του τελεστή, υπάρχει η επιλογή "invert filter" (αντιστροφή φίλτρου), η οποία αντιστρέφει το φίλτρο. Επομένως, επιλέγοντάς την, διαγράφονται, όσα παραδείγματα ικανοποιούν τη συνθήκη. Με αυτό τον τρόπο μετακινούνται οι βραδινές ώρες από τα δεδομένα.

Τέλος, το σύνολο των τελεστών, που τοποθετήθηκαν σε σειρά, φαίνεται παρακάτω (Στάδιο 2):

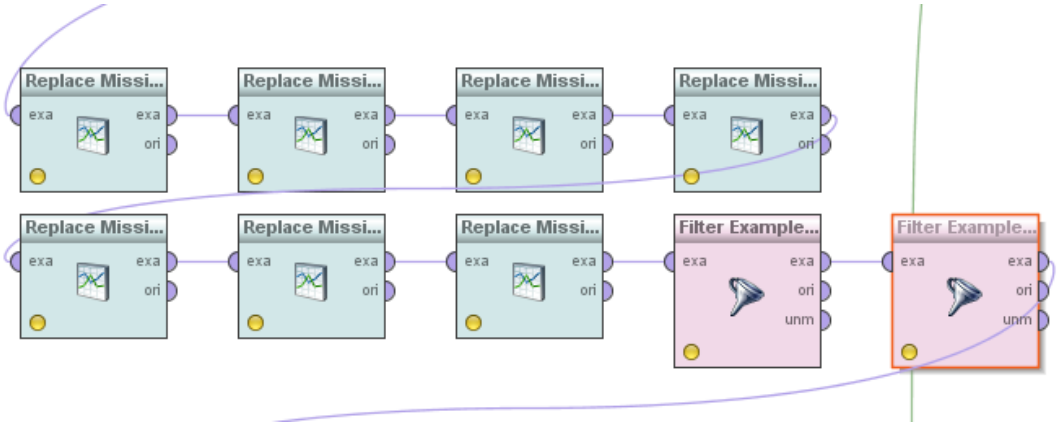


Figure 31 Διαδικασία από τελεστή "Replace Missing Values" έως "Filter Examples" (Στάδιο 2, διαχείριση κενών και μηδενικών τιμών)

Ο πρώτος τελεστής δέχεται στην είσοδο του την έξοδο του τελεστή "Join" και ο τελευταίος συνδέεται με το υπόλοιπο κύκλωμα ανάλυσης.

Ως αποτέλεσμα του δεύτερου σταδίου, εμφανίζεται ένα σύνολο δεδομένων χωρίς την ύπαρξη κενών και μηδενικών τιμών.

Row No.	datetime	radiation	power	tmpf	dwpf	relh	drct	sknt	clouds
1	2012-01-021	12.312	48.693	12	6	66.700	260	360.108	1
2	2012-01-021	23.335	97.386	13	7	66.910	260	360.108	1
3	2012-01-021	108.659	654.420	13	7	66.910	270	25.722	1
4	2012-01-021	193.983	1211.455	14	7	62.690	270	360.108	1
5	2012-01-021	260.093	2207.302	15	7	58.770	290	617.328	1
6	2012-01-021	326.203	3203.148	15	6	54.860	310	565.884	1
7	2012-01-021	450.330	5711.928	15	6	54.860	280	77.166	1
8	2012-01-021	574.456	8220.707	16	7	55.120	270	617.328	1
9	2012-01-021	596.552	8715.976	15	6	54.860	260	823.104	1
10	2012-01-021	618.647	9211.245	16	6	51.450	240	462.996	3

Figure 32 Σύνολο δεδομένων μετά το χειρισμό μηδενικών και κενών τιμών

4.4.3 Ορισμός μεταβλητών προς εξέταση

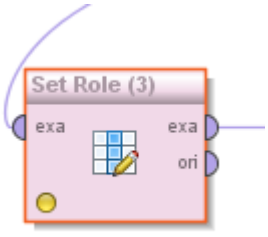
Σε αυτή την παράγραφο αναλύεται το Στάδιο 3 της προετοιμασίας χρονοσειράς (Σχήμα 12).

4.4.3.1 *Ορισμός ρόλου (2)*

Προκειμένου να προχωρήσει η ανάλυσή, χρειάζεται η επαναχρησιμοποίηση του τελεστή *"Set Role"*, για τον ορισμό του ρόλου μίας μεταβλητής.

Η λειτουργία του τελεστή είναι ακριβώς η ίδια, όπως παρουσιάστηκε στην παράγραφο 4.4.1.3, με τη μόνη διαφορά ότι τώρα αλλάζει ο ρόλος της μεταβλητής "power".

Μετατρέπεται η συγκεκριμένη μεταβλητή, από «απλή» (regular) σε «σημαντική» (special) και πιο συγκεκριμένα σε «ετικέτα» (label), όπως φαίνεται στο Σχήμα 33.



Σχήμα 34 Τελεστής "Set Role" (3)

Σχήμα 33 Παράμετροι τελεστή "Set Role" (3)

Όταν μία μεταβλητή οριστεί ως «ετικέτα» (label), τότε λειτουργεί ως μεταβλητή-στόχος για τους τελεστές εκμάθησης (όπως είναι για παράδειγμα ο τελεστής "Decision Tree"). Οι «ετικέτες» ταυτοποιούν το παράδειγμα, με οποιοδήποτε τρόπο και πρέπει να προβλεφθούν για τα καινούρια παραδείγματα, τα οποία δεν έχουν ακόμα χαρακτηριστεί. Δηλαδή, μία μεταβλητή «ετικέτα» (label) αποτελεί τη μεταβλητή, που πρόκειται να προβλεφθεί.

Καθώς αντικείμενο της συγκεκριμένης ανάλυσης είναι η πρόβλεψη της παραγωγής του ΦΒ, ορίζεται η μεταβλητή "power" ως «ετικέτα».

4.4.3.2 *Ορισμός μεταβλητών προς εξέταση*

Επόμενο βήμα της ανάλυσης και προετοιμασίας της χρονοσειράς, είναι ο ορισμός μεταβλητών προς εξέταση, δηλαδή η δημιουργία νέων μεταβλητών. Οι καινούριες μεταβλητές αποτελούν προϊόντα των ήδη υπαρχόντων και η δημιουργία τους εξυπηρετεί στην ανακάλυψη της καλύτερης περιγραφής της παραγωγής και επομένως στη δημιουργία αποδοτικότερου μοντέλου πρόβλεψης.

Για τη δημιουργία καινούριων μεταβλητών, χρησιμοποιείται ο τελεστής "Generate Attributes", από τους τελεστές Μετατροπή Δεδομένων (Data Transformation) – Μείωση και Μετατροπή του Συνόλου των Μεταβλητών (Attribute Set Reduction and Transformation) – Δημιουργία (Generation).

Ο συγκεκριμένος τελεστής δημιουργεί νέες, οριζόμενες από τον χρήστη, μεταβλητές ,χρησιμοποιώντας μαθηματικές εκφράσεις. Τα ονόματα των μεταβλητών του συνόλου δεδομένων, μπορούν να χρησιμοποιηθούν ως μεταβλητές στις μαθηματικές εκφράσεις, που εφαρμόζονται. Κατά την εφαρμογή του τελεστή, αυτές οι εκφράσεις αξιολογούνται.

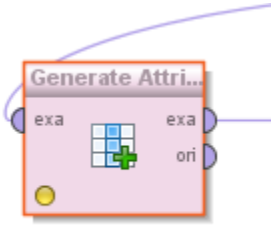
Αποτέλεσμα του τελεστή δεν είναι μόνο η δημιουργία νέων στηλών, για τις καινούριες μεταβλητές, αλλά επίσης και το γέμισμα αυτών των στηλών, με τις κατάλληλες τιμές. Αν μια μεταβλητή, η οποία χρησιμοποιείται σε μία μαθηματική έκφραση, δεν είναι ορισμένη, τότε ολόκληρη η έκφραση γίνεται απροσδιόριστη και στην θέση της αποθηκεύεται ένα ερωτηματικό "?". Είναι σημαντικό να αναφερθούν κάποιοι περιορισμοί που υπάρχουν, σχετικά με τα ονόματα των μεταβλητών που δημιουργούνται, ώστε να δουλέψει σωστά ο τελεστής. Για παράδειγμα, δεν επιτρέπεται τα ονόματα να περιέχουν παρενθέσεις, κενά και άνω παύλες. Επίσης, δεν επιτρέπεται τα ονόματα των καινούριων μεταβλητών να συμπίπτουν με ονόματα τελεστών και σταθερών, όπως είναι το «π» (pi).

Ο τελεστής "Generate Attributes" υποστηρίζει ένα μεγάλο αριθμό συναρτήσεων και έτσι επιτρέπει τη γραφή πλούσιων εκφράσεων. Δηλαδή, υποστηρίζονται όλες οι μαθηματικές (πρόσθεση, αφαίρεση, πολλαπλασιασμός, διαίρεση) και λογικές πράξεις (>, <, <=, >=, ==, !=, 1, &&, ‖). Επιπλέον, υποστηρίζει συναρτήσεις εκθετικές, λογαριθμικές, τριγωνομετρικές, στατιστικές και πολλές ακόμα. Η διαδικασία που παρουσιάζεται σε αυτή την εργασία, περιορίζεται κυρίως σε μαθηματικές συναρτήσεις και συναρτήσεις κειμένου. Από τις συναρτήσεις κειμένου, αυτές που παρουσιάζουν περισσότερο ενδιαφέρον, είναι οι: κόψιμο ("cut") και ανάλυση ("parse").

Η συνάρτηση "cut" ορίζεται ως: "cut(μεταβλητή, πρώτο ψηφίο, μήκος)". Χρησιμοποιείται ουσιαστικά, για να κόψει την υπό-ακολουθία συγκεκριμένου «μήκους» από τη θέση που δείχνει το «πρώτο ψηφίο». Η αρίθμηση ξεκινάει από το 0. Για παράδειγμα cut("text", 1, 2) μας δίνει "ex".

Αντίστοιχα, η συνάρτηση "parse", μετατρέπει τη συγκεκριμένη συμβολοσειρά ονομαστικής (nominal) τιμής σε αριθμό.

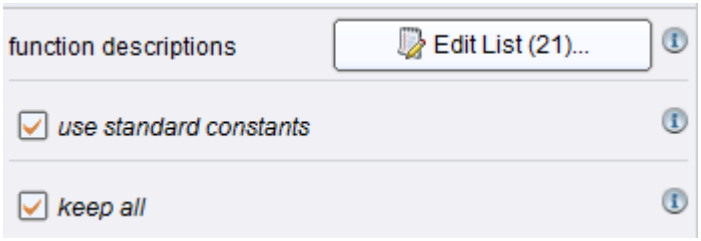
Ο τελεστής περιέχει πολλές ακόμα συναρτήσεις, οι οποίες όμως ξεφεύγουν από τα όρια της εργασίας, γι' αυτό και δεν θα αναφερθούν.



Σχήμα 35 Τελεστής "Generate Attributes"

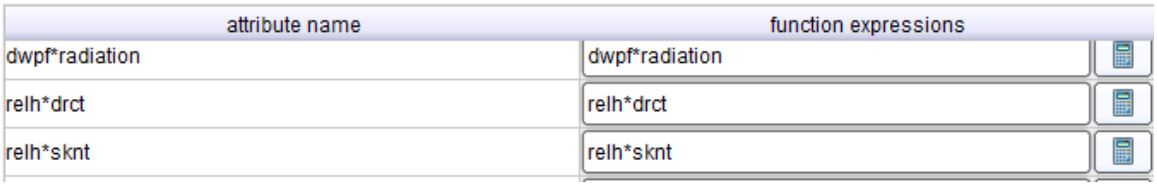
Όπως φαίνεται και από το Σχήμα 35, η είσοδος (πύλη "exa") του τελεστή, είναι το σύνολο δεδομένων και η έξοδος (πύλη "exa") το σύνολο δεδομένων, με τις καινούριες μεταβλητές.

Αρχικά, χρησιμοποιείται ένας τελεστής "Generate Attributes", για την παραγωγή όλων των δυνατών συνδυασμών, μεταξύ των ανεξάρτητων μεταβλητών. Οι ανεξάρτητες μεταβλητές είναι εφτά, επομένως οι δυνατοί συνδυασμοί είναι εικοσιένα, όπως εμφανίζεται και στο Σχήμα 36, στην παράμετρο "function descriptions".



Σχήμα 36 Παράμετροι τελεστή "Generate Attributes" (1)

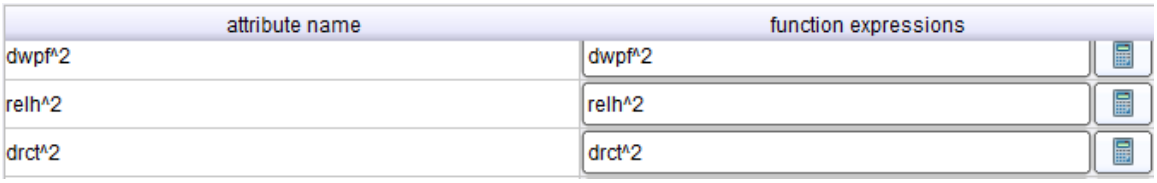
Ένα παράδειγμα αυτών των συνδυασμών φαίνεται και στο Σχήμα 37



Σχήμα 37 Καινούριες μεταβλητές τελεστή "Generate Attributes" (1)

Οι καινούριες μεταβλητές, ονομάζονται (στήλη "attribute name") σύμφωνα με τον τύπο με τον οποίο ορίζονται (στήλη "function expressions"). Για παράδειγμα, στην πρώτη γραμμή έχει ορισθεί ο συνδυασμός της ακτινοβολίας με το σημείο δρόσου. Αυτή η μεταβλητή θα ονομάζεται "dwpf*radiation" – με αυτό το όνομα θα εμφανιστεί και στο σύνολο δεδομένων – και ορίζεται ως ο πολλαπλασιασμός της ακτινοβολίας επί το σημείο δρόσου.

Επόμενη ενέργεια είναι ο ορισμός καινούριων μεταβλητών, ως οι δυνάμεις των ήδη υπαρχόντων. Δηλαδή, η ύψωση όλων των ανεξάρτητων μεταβλητών στο τετράγωνο και στην τρίτη. Αυτές οι καινούριες μεταβλητές είναι δεκατέσσερις, καθώς οι αρχικές μεταβλητές είναι εφτά (7*2=14). Ένα παράδειγμα αυτών των καινούριων μεταβλητών φαίνεται στο Σχήμα 38.



Σχήμα 38 Καινούριες μεταβλητές τελεστή "Generate Attributes" (2)

Όπως και προηγουμένως, το όνομα κάθε μεταβλητής φαίνεται στην πρώτη στήλη και ο τύπος με τον οποίο ορίζεται, στη δεύτερη.

Μέχρι στιγμής έχει ολοκληρωθεί η δημιουργία καινούριων μεταβλητών μέσω του συνδυασμού των ήδη υπαρχόντων. Οι μόνες μεταβλητές, που δεν έχουν χρησιμοποιηθεί, είναι η παραγωγή και η ημερομηνία με την ώρα.

Η παραγωγή είναι η μεταβλητή που θέλουμε να προβλέψουμε, επομένως δεν θα χρησιμοποιηθεί.

Η ημερομηνία με την ώρα (datetime) είναι μη αριθμητική μεταβλητή και περιέχει τέσσερις μεταβλητές (έτος, μήνας, ημέρα, ώρα) σε μία, με τη μορφή που φαίνεται παρακάτω:

datetime
2014-05-08T10:00:00Z
2014-05-08T10:30:00Z
2014-05-08T11:00:00Z
2014-05-08T11:30:00Z

Σχήμα 39 Μεταβλητή "datetime"

Ο διαχωρισμός της μεταβλητής "datetime" σε επιμέρους ανεξάρτητες μεταβλητές, είναι απαραίτητος για τον καλύτερο προσδιορισμό της παραγωγής. Θα χρησιμοποιηθεί δηλαδή τελεστές "Generate Attributes", ακόμα πέντε φορές.

Στον πρώτο τελεστή διαχωρίζεται το έτος, ο μήνας και η ημέρα. Χρησιμοποιείται δηλαδή η συνάρτηση "cut". Όπως φαίνεται και στο Σχήμα 39, το έτος ξεκινάει από το ψηφίο μηδέν και διαθέτει τέσσερα ψηφία. Ο μήνας ξεκινάει από το πέντε και διαθέτει δύο ψηφία και η ημέρα ξεκινάει από το οκτώ και διαθέτει και αυτή δύο ψηφία. Οπότε, οι μεταβλητές θα δημιουργηθούν σύμφωνα με τους παρακάτω τύπους:

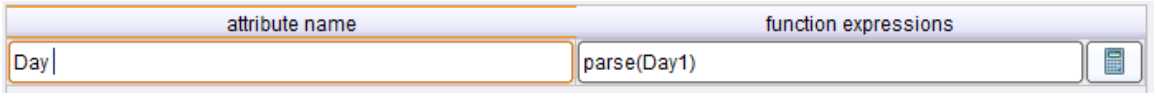
attribute name	function expressions
Year	cut(datetime,0,4)
Month1	cut(datetime,5,2)
Day1	cut(datetime,8,2)

Σχήμα 40 Καινούριες μεταβλητές τελεστή "Generate Attributes" (3) με χρήση της συνάρτησης "cut"

Επειδή, από την έξοδο του τελεστή, ο τύπος των μεταβλητών μήνας και ημέρα δεν είναι αναγνωρίσιμος, θα χρησιμοποιηθούν άλλοι δύο τελεστές "Generate Attribute" και συγκεκριμένα η συνάρτηση "parse", ώστε να μετατραπούν σε κατάλληλη μορφή

attribute name	function expressions
Month	parse(Month1)

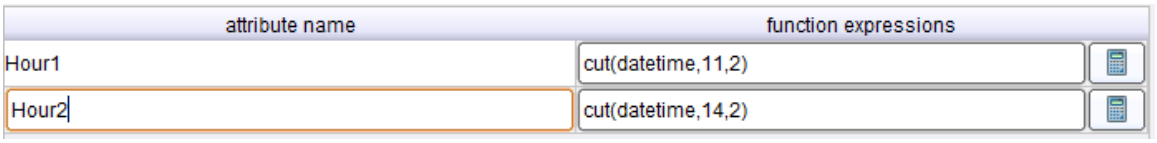
Σχήμα 41 Καινούρια μεταβλητή (month) τελεστή "Generate Attributes" (4) με χρήση της συνάρτησης "parse"



Σχήμα 42 Καινούρια μεταβλητή τελεστή (day) "Generate Attributes" (5) με χρήση της συνάρτησης "parse"

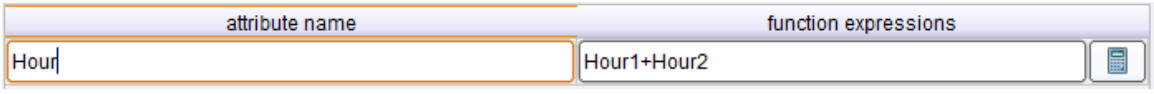
Η τελευταία μεταβλητή, η οποία πρέπει να δημιουργηθεί, είναι η ώρα. Χρησιμοποιούνται δύο τελεστές "Generate Attributes", ο πρώτος θα κόψει τις ώρες και τα λεπτά από τη μεταβλητή "datetime" και ο δεύτερος θα ενώσει αυτές τις δύο μεταβλητές σε μία ενιαία "Hour". Με αυτό τον τρόπο, αφαιρείται η άνω και κάτω τελεία, που υπάρχει στην αρχική μεταβλητή "datetime" (π.χ 2012-01-02T14:30:00Z) και εμποδίζει στη μετέπειτα δημιουργία του μοντέλου.

ξαναγυρνώντας στο Σχήμα 39, παρατηείται πως η ώρα ξεκινάει στο ψηφίο ένδεκα και έχει μήκος δύο ψηφία, ενώ τα λεπτά ξεκινούν στο δεκατέσσερα και περιλαμβάνουν και αυτά δύο ψηφία. Οπότε, ο πρώτος τελεστής, δημιουργεί τις μεταβλητές "Hour1" και "Hour2", μέσω της συνάρτησης "cut".



Σχήμα 43 Καινούριες μεταβλητές (Hour1, Hour2) τελεστή "Generate Attributes" (6) με χρήση της συνάρτησης "cut"

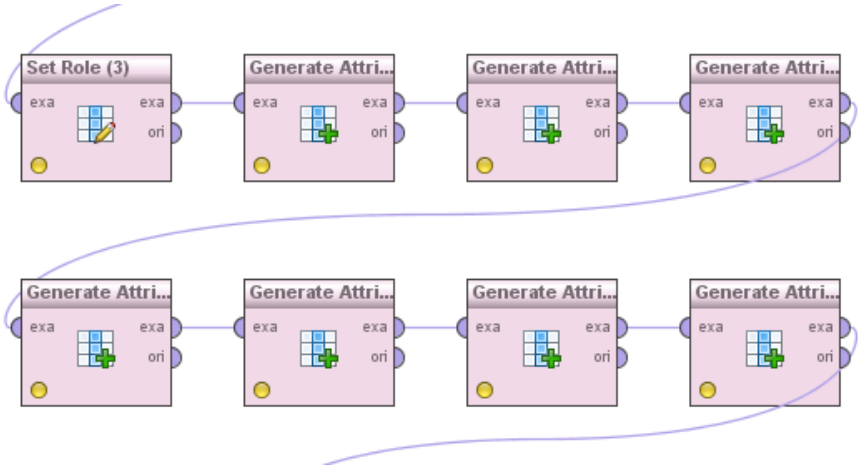
Προκειμένου όμως να προκύψει μία ενιαία μεταβλητή, απαιτείατια ένας ακόμα τελεστής "Generate Attributes", ο οποίοσθα δημιουργήσει τη μεταβλητή "Hour".



Σχήμα 44 Καινούρια μεταβλητή (Hour) τελεστή "Generate Attributes" (7)

Τελικά, το σύνολο των δεδομένων, με την εισαγωγή αυτών των τελεστών "Generate Attributes", από εφτά ανεξάρτητες-απλές μεταβλητές, πλέον έχει πενήντα απλές μεταβλητές.

η υλοποίηση του σταδίου 3, όπως εμφανίζεται στο Rapidminer είναι η ακόλουθη:



Σχήμα 45 Διαδικασία από τελεστή "Set Role" μέχρι "Generate Attributes (7)" (Στάδιο 3, ορισμός μεταβλητών προς εξέταση)

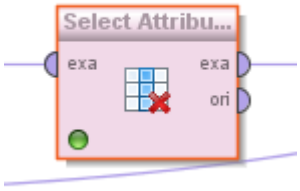
4.4.4 Επιλογή σημαντικών μεταβλητών

Το τέταρτο στάδιο της προετοιμασίας χρονοσειράς (Σχήμα 12), ονομάζεται «Επιλογή σημαντικών μεταβλητών».

4.4.4.1 *Χρήση βοηθητικών μεταβλητών*

Στην παράγραφο 4.4.3.2, εμφανίστηκε η ανάγκη κατασκευής κάποιων βοηθητικών μεταβλητών, οι οποίες εξυπηρέτησαν στην δημιουργία των "Month", "Day", "Hour". Αυτές οι βοηθητικές μεταβλητές είναι οι "Day1", "Month1", "Hour1", "Hour2". Από τη στιγμή που κατασκευάστηκαν οι ανεξάρτητες μεταβλητές που θα χρησιμοποιηθούν στη διαδικασία εξαγωγής βαρών, δεν απαιτείται η διατήρηση των βοηθητικών μεταβλητών, καθώς επηρεάζουν το μοντέλο, ως ανεξάρτητες. Γι αυτό το λόγο διαγράφονται, με τη βοήθεια του τελεστή "Select Attributes", από τους τελεστές Μετατροπή Δεδομένων (Data Transformation) – Μείωση και Μετατροπή του Συνόλου των Μεταβλητών (Attribute Set Reduction and Transformation) – Επιλογή (Selection).

Ο τελεστής "Select Attributes" επιτρέπει τον διαχωρισμό των μεταβλητών σε αυτές που θα διαγραφούν και σε αυτές που θα διατηρηθούν. Διατίθενται διάφορα φίλτρα, τα οποία βοηθούν στην επιλογή. Μόνο οι επιλεγμένες μεταβλητές θα περάσουν στην έξοδο του τελεστή και οι υπόλοιπες θα απομακρυνθούν από τα δεδομένα.

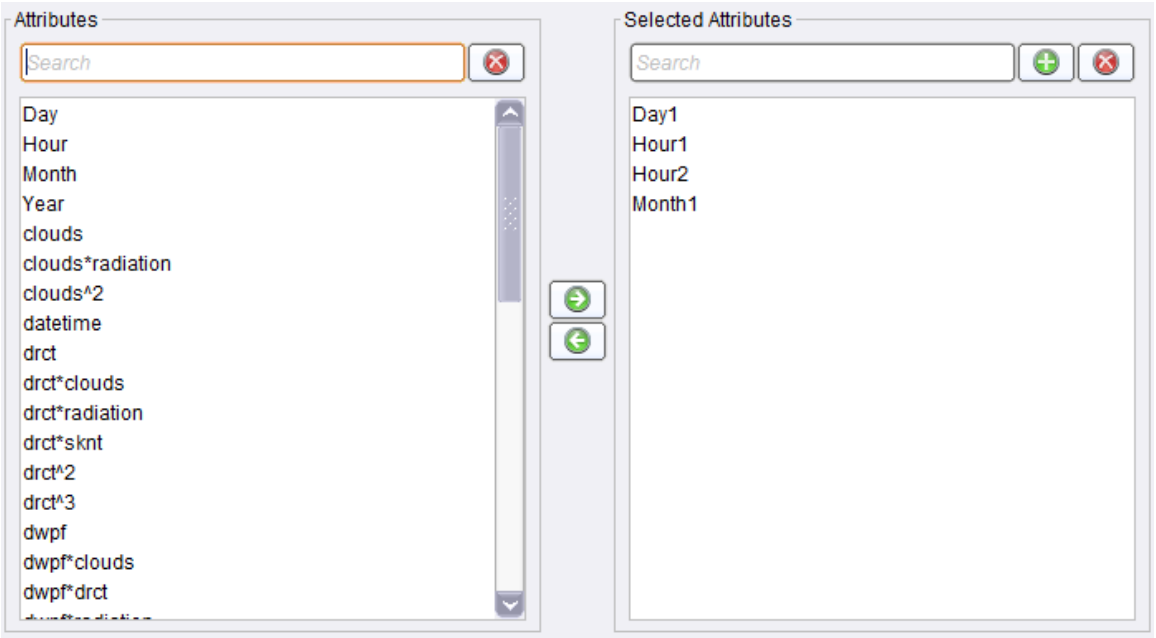


Σχήμα 47 Τελεστής "Select Attributes"

Σχήμα 46 Παράμετροι τελεστή "Select Attributes"

Αναλύοντας τις παραμέτρους του τελεστή, η παράμετρος“attribute filter type” επιτρέπει, τον καθορισμό φίλτρου (μεθόδου), σύμφωνα με το οποίο επιλέγονται οι μεταβλητές. Για την διαδικασία που αναλύεται, επιλέγεται η κατηγορία του υποσυνόλου “subset”, η οποία επιτρέπει την επιλογή πολλαπλών μεταβλητών, από μία λίστα. Όλες οι μεταβλητές του συνόλου δεδομένων εμφανίζονται σε αυτή τη λίστα και είναι πολύ εύκολος ο διαχωρισμός των επιθυμητών από τις ανεπιθύμητες μεταβλητές.

Με αυτή την επιλογή, εμφανίζεται άλλη μία παράμετρος, η οποία φαίνεται και στο Σχήμα 46, με το όνομα “attributes”. Αυτή η παράμετρος ανοίγει ένα παράθυρο με δύο στήλες. Όλες οι μεταβλητές φαίνονται στην αριστερή στήλη και μπορούν να μεταφερθούν στη δεξιά στήλη, η οποία είναι η λίστα με τις επιλεγμένες μεταβλητές, που θα περάσουν στην έξοδο του τελεστή. Όσες μεταβλητές παραμείνουν στην αριστερή στήλη, θα διαγραφούν από τα δεδομένα μας. Αυτή η διαδικασία φαίνεται στο Σχήμα 48.



Σχήμα 48 Επιλογή μεταβλητών του τελεστή "Select Attributes"

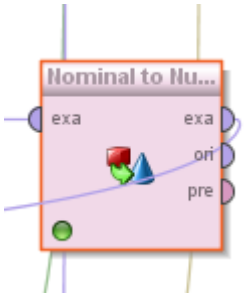
Ολόκληρη η διαδικασία, που αναφέρθηκε παραπάνω, μπορεί να αντιστραφεί με την επιλογή της παραμέτρου "invert filter". Δηλαδή, όσες μεταβλητές περνάνε στη δεξιά στήλη, θα αφαιρεθούν. Επομένως, από τη στιγμή που είναι μαρκαρισμένη αυτή η επιλογή (Σχήμα 46) και οι μεταβλητές "Day1", "Hour1", "Hour2", "Month1" βρίσκονται στη δεξιά στήλη, επιτυγχάνεται η διαγραφή τους από τα δεδομένα.

Τελικά, η έξοδος (πύλη "exa") του τελεστή, θα είναι το σύνολο δεδομένων, χωρίς τις βοηθητικές μεταβλητές.

4.4.4.2 *Αλλαγή τύπου μεταβλητής*

Παρατηρείται από την έξοδο του τελευταίου τελεστή ("Select Attributes"), ότι σχεδόν όλες οι απλές μεταβλητές, από τις οποίες θα προκύψει και η πρόβλεψη της παραγωγής, έχουν πραγματικές τιμές, εκτός από τις μεταβλητές "Year", "Month", "Day", "Hour". Από τη στιγμή που ο τελεστής που θα χρησιμοποιηθεί αργότερα, για τον ορισμό του μοντέλου, δεν μπορεί να διαχειριστεί μη αριθμητικές τιμές, είναι απαραίτητη η αλλαγή του τύπου αυτών των μεταβλητών.

Γι αυτό το λόγο, χρησιμοποιείται ο τελεστής "Nominal to Numerical", από τη λίστα Μετατροπή Δεδομένων (Data Transformation) – Μετατροπή Τύπου (Type Conversion). Αυτός ο τελεστής αλλάζει τον τύπο των μη αριθμητικών μεταβλητών σε αριθμητικές. Επίσης, αντιστοιχίζει όλες τις τιμές αυτών των μεταβλητών σε αριθμητικές τιμές.

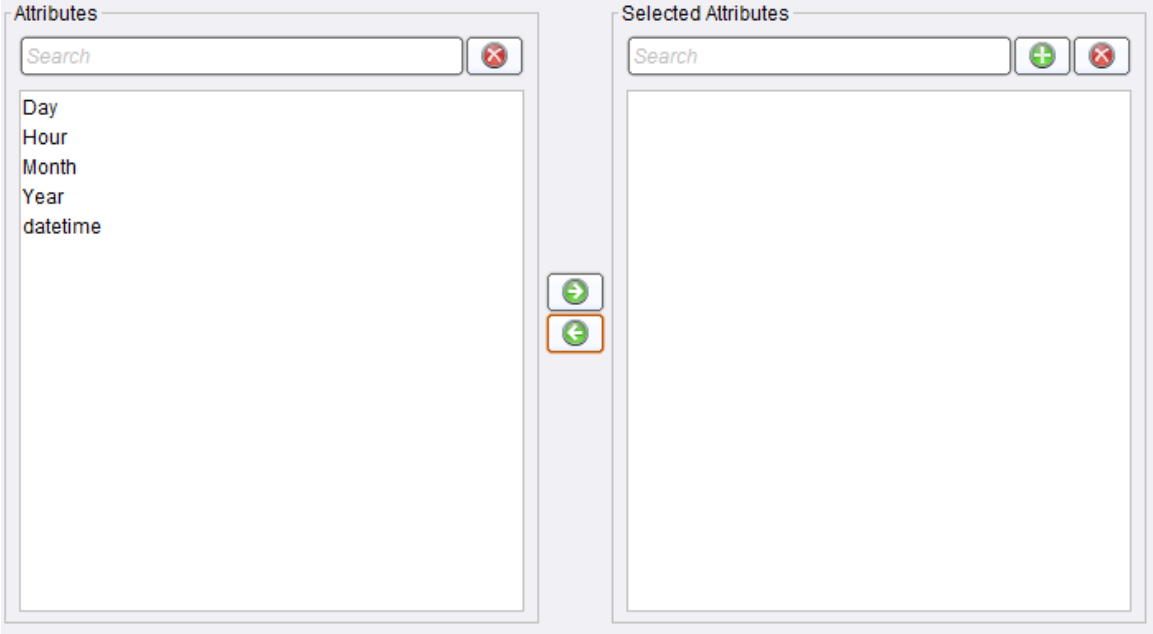


Σχήμα 49 Τελεστής "Nominal to Numerical"

Σχήμα 50 Παράμετροι τελεστή "Nominal to Numerical"

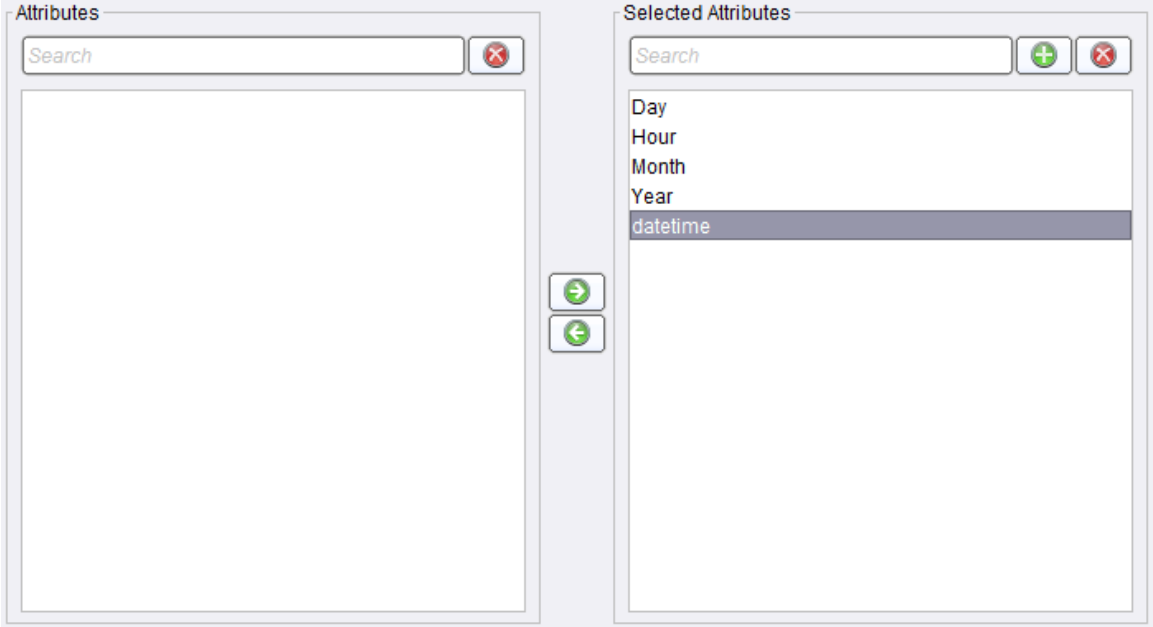
Από τις παραμέτρους του, για τον τύπο φίλτρου (attribute filter type), επιλέγεται υποσύνολο "subset". Όπως και στον προηγούμενο τελεστή, που εξετάστηκε, αυτή η εναλλακτική επιτρέπει την επιλογή πολλών μεταβλητών, από μία λίστα, η οποία περιέχει όλες τις μεταβλητές του συνόλου δεδομένων. Με αυτή την επιλογή εμφανίζεται άλλη μία παράμετρος η οποία ονομάζεται "attributes" και επιτρέπει την επιλογή των μεταβλητών,

που πρέπει να μετατραπούν σε αριθμητικές. Με την παράμετρο αυτή ανοίγει στο παράθυρο εργασίας μία λίστα, η οποία περιέχει όλες τις μη αριθμητικές μεταβλητές, όπως φαίνεται στο Σχήμα 51.



Σχήμα 51 Μη αριθμητικές μεταβλητές τελεστή "Nominal to Numerical"

Για την αλλαγή όλων των επιθυμητών μεταβλητών σε αριθμητικές, είναι αναγκαία η μεταφορά τους στη δεξιά στήλη. Οπότε, θα εμφανιστεί η εικόνα που φαίνεται στο Σχήμα 52.



Σχήμα 52 Επιλεγμένες μη αριθμητικές μεταβλητές τελεστή "Nominal to Numerical"

Η τελευταία παράμετρος, η οποία απαιτεί προσοχή, είναι η "coding type". Αυτή η παράμετρος εκφράζει τον κώδικα, σύμφωνα με τον οποίο μετατρέπεται ο τύπος των μεταβλητών. Για τις ανάγκες της εργασίας, επιλέγεται η εναλλακτική "unique integers" (μοναδικοί ακέραιοι), σύμφωνα με την οποία, οι τιμές των μη αριθμητικών τιμών αντιμετωπίζονται ως ίσες. Έτσι, η μη αριθμητική μεταβλητή μετατρέπεται σε μεταβλητή, πραγματικής τιμής.

Τελικά, στην έξοδο (πύλη "exa") του τελεστή, εμφανίζεται το σύνολο δεδομένων, όπως ήταν στην είσοδό (πύλη "exa"), με τις μεταβλητές "Year", "Month", "Day", "Hour" αριθμητικές.

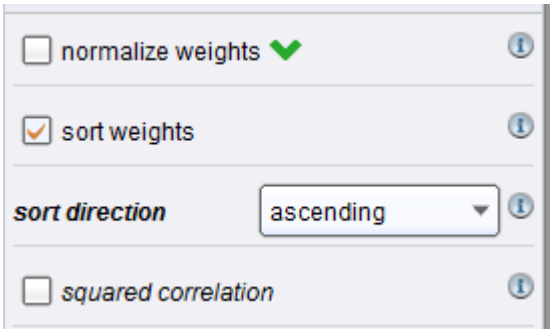
4.4.4.3 *Βάρη μέσω συσχέτισης*

Μέχρι αυτό το σημείο έχει δημιουργηθεί ένα σύνολο δεδομένων, με πενήντα απλές-ανεξάρτητες μεταβλητές. Για την δημιουργία του μοντέλου, είναι πολύ δύσκολο, η διαχείριση τόσο μεγάλου πλήθους μεταβλητών. Γι' αυτό το λόγο, πρέπει να υπολογιστούν τα βάρη των μεταβλητών προκειμένου να αποκλειστούν κάποιες από την πρόβλεψη. Το βάρος κάθε μεταβλητής, εκφράζει το βαθμό συσχέτισης της εξαρτημένης μεταβλητής "power" από κάθε ανεξάρτητη μεταβλητή. Από τους τελεστές Μοντελοποίηση (Modelling) – Βάρη Μεταβλητών (Attribute Weighting), επιλέγεται ο τελεστής "Weight by Correlation".

Ο τελεστής αυτός εκτιμά τη σχέση των μεταβλητών, υπολογίζοντας την τιμή της συσχέτισης, για κάθε μεταβλητή του συνόλου δεδομένων, ως προς την μεταβλητή «ετικέτα» (label), δηλαδή την παραγωγή. Όσο μεγαλύτερο το βάρος, τόσο πιο σημαντική θεωρείται η μεταβλητή.

Η συσχέτιση είναι ένας αριθμός ανάμεσα στο -1 και το +1, που μετράει το βαθμό του συνδέσμου, μεταξύ δύο μεταβλητών (ας τις ονομάσουμε X και Y). Μία θετική τιμή για τη συσχέτιση δείχνει ανάλογη σχέση ανάμεσα στις δύο μεταβλητές. Για παράδειγμα, υψηλές τιμές της μεταβλητής X ,συνήθως, συνδέονται με υψηλές τιμές της μεταβλητής Y και αντίστοιχα, χαμηλές τιμές της X συνδέονται με χαμηλές τιμές της Y. Από την άλλη πλευρά, μία αρνητική συσχέτιση, σημαίνει και αντιστρόφως ανάλογη σχέση. Δηλαδή, υψηλές τιμές για την X συνδέονται με χαμηλές τιμές για την Y και αντίστροφα.

Σχήμα 53 Τελεστής "Weight by Correlation"



Σχήμα 54 Παράμετροι τελεστή "Weight by Correlation"

Η παράμετρος “sort weights” δηλώνει, την επιθυμία ταξινόμησης των βαρών. Στην περίπτωση που αυτή η παράμετρος έχει οριστεί ως αληθής τότε το είδος της ταξινόμησης ορίζεται από την παράμετρο “sort direction”. Επιλέγεται η ταξινόμηση να είναι αύξουσα. Δηλαδή, πρώτη να εμφανίζεται η μεταβλητή, με το μικρότερο βάρος και τελευταία αυτή με το μεγαλύτερο.

Το ιδιαίτερο χαρακτηριστικό του συγκεκριμένου τελεστή είναι οι πύλες εξόδου, που προσφέρει.

Η πύλη “exa” δίνει το σύνολο δεδομένων, όπως ακριβώς ήρθε στην είσοδο του τελεστή.

Η πύλη “wei” δίνει τα βάρη των μεταβλητών, ως προς την μεταβλητή «ετικέτα» (label), δηλαδή την παραγωγή “power”, την οποία θέλουμε να προβλέψουμε. Το αποτέλεσμα αυτής της εξόδου θα είναι τα βάρη, πενήντα μεταβλητών, σε αύξουσα σειρά, όπως φαίνεται στο Σχήμα 55.

attribute	weight
Year	0.003
tmpf*relh	0.006
dwpf	0.014
Month	0.031
Day	0.033
dwpf^3	0.040
relh*drct	0.041
drct^3	0.052
dwpf^2	0.056
drct*clouds	0.068
sknt*clouds	0.070
tmpf*clouds	0.117
drct^2	0.129
dwpf*relh	0.153
tmpf*dwpf	0.157
dwpf*drct	0.170
dwpf*clouds	0.188
clouds	0.239
clouds^2	0.254
drct	0.277
relh*clouds	0.307
relh*sknt	0.320
dwpf*sknt	0.351
drct*sknt	0.355
tmpf^3	0.366
tmpf^2	0.378
clouds*radiation	0.380
tmpf	0.386
sknt^2	0.417
tmpf*drct	0.439
sknt	0.466
Hour	0.467
relh^3	0.518
tmpf*sknt	0.527
relh	0.533
relh^2	0.533
dwpf*radiation	0.647
sknt*radiation	0.719
drct*radiation	0.846
relh*radiation	0.866
radiation^3	0.891
tmpf*radiation	0.896
radiation^2	0.953
radiation	0.994

Σχήμα 55 Βάρη μεταβλητών από τον τελεστή "Weight by Correlation"

Όπως παρατηρείται, η ανεξάρτητη μεταβλητή, με το μεγαλύτερο βάρος, είναι η ακτινοβολία (radiation). Δηλαδή, η παραγωγή εξαρτάται, σε πολύ μεγάλο βαθμό, από την ακτινοβολία.

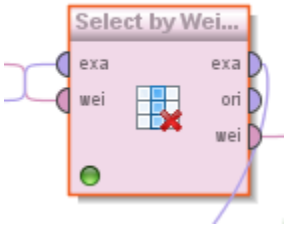
4.4.4.4 *Επιλογή μέσω Βαρών*

Από τη στιγμή που οι μεταβλητές, οι οποίες επηρεάζουν περισσότερο την παραγωγή, είναι γνωστές, πρέπει να διαχωριστούν από το υπόλοιπο σύνολο δεδομένων και να χρησιμοποιηθούν για τη δημιουργία του μοντέλου.

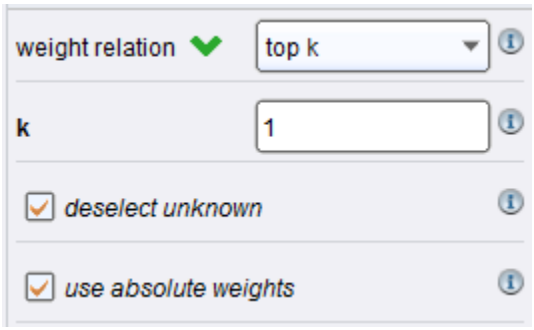
Επιλέγεται ο τελεστής "Select by Weights", από τους Μετατροπή Δεδομένων (Data Transformation) – Μείωση και Μετατροπή του Συνόλου των Μεταβλητών (Attribute Set Reduction and Transformation) – Επιλογή (Selection).

Αυτός ο τελεστής επιλέγει, από ένα σύνολο δεδομένων, μόνο τις μεταβλητές, των οποίων τα βάρη ικανοποιούν τα καθορισμένα κριτήρια, σε σχέση με τα βάρη εισόδου.

Τα βάρη εισόδου εισέρχονται στον τελεστή, μέσω της πύλης εισόδου "wei". Από την πύλη εισόδου "exa" εισέρχεται το σύνολο δεδομένων.



Σχήμα 56 Τελεστής "Select by Weights"



Σχήμα 57 Παράμετροι τελεστή "Select by Weights"

Η παράμετρος "weight relation" ορίζει τη συνθήκη, σύμφωνα με την οποία, θα διατηρηθούν οι μεταβλητές. Για παράδειγμα, η συνθήκη θα μπορούσε να ορίζει, να διατηρηθούν μόνο οι μεταβλητές, των οποίων το βάρος είναι μεγαλύτερο από μία τιμή. Για το μοντέλο που θα δημιουργηθεί, επιλέγεται η συνθήκη "top k" και όπου k θέτουμε τον αριθμό ένα. Αυτό σημαίνει ότι θα διατηρηθεί μόνο η μεταβλητή, με το μεγαλύτερο βάρος.

Η πύλη "wei" παραδίδει τα βάρη των μεταβλητών, όπως ακριβώς εισήχθησαν στην είσοδο, δηλαδή από τον τελεστή "Weight by Correlation" (συνδέεται στην έξοδο της συνολικής διαδικασίας). Ενώ η πύλη "exa" δίνει το σύνολο δεδομένων, διατηρώντας μόνο τις μεταβλητές, οι οποίες ικανοποίησαν τη συνθήκη.

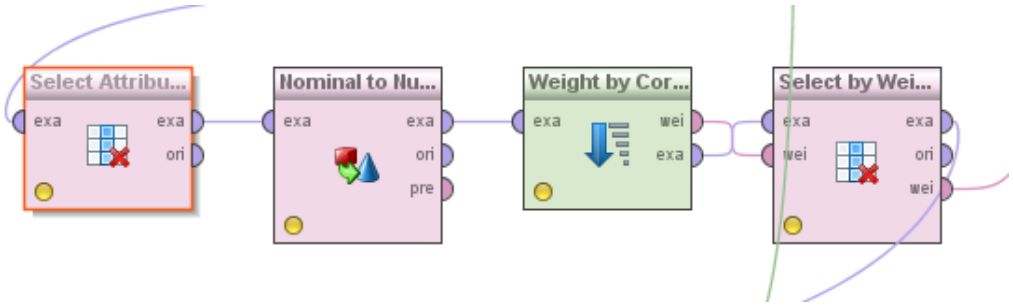
Κοιτώντας την έξοδο, εμφανίζεται ένα σύνολο δεδομένων, το οποίο διαθέτει μόνο μία ανεξάρτητη μεταβλητή, αυτή με το μεγαλύτερο βάρος. Για το σύνολο δεδομένων που διατίθεται, αυτή η μεταβλητή είναι η ακτινοβολία (Σχήμα 58).

Row No.	datetime	power	radiation
1	2012-01-021	48.693	12.312
2	2012-01-021	97.386	23.335
3	2012-01-021	654.420	108.659
4	2012-01-021	1211.455	193.983
5	2012-01-021	2207.302	260.093
6	2012-01-021	3203.148	326.203
7	2012-01-021	5711.928	450.330
8	2012-01-021	8220.707	574.456
9	2012-01-021	8715.976	596.552
10	2012-01-021	9211.245	618.647

Σχήμα 58 Διατήρηση μεταβλητής με το μεγαλύτερο βάρος από τελεστή "Select by Weights"

Παρατηρούμε ότι η παραγωγή ενέργειας από τη ΦΒ εγκατάσταση περιγράφεται πολύ καλά (συσχέτιση R=0.994) από την ηλιακή ακτινοβολία. Επομένως, Επιλέγουμε τη διατήρηση μόνο της συγκεκριμένης μεταβλητής..

Επομένως το τελικό στάδιο της προετοιμασίας της χρονοσειράς (Στάδιο 4), φαίνεται στο Σχήμα 59.



Σχήμα 59 Διαδικασία ανάλυσης από τελεστή "Select Attributes" έως "Select by Weights" (Στάδιο 4, επιλογή σημαντικών μεταβλητών)

Τέλος, η συνολική διαδικασία της προετοιμασίας της χρονοσειράς, περιλαμβάνει τα στάδια 1,2,3 και 4. Η σύνδεσή τους γίνεται από τις πύλες εισόδου και εξόδου του καθενός. Δηλαδή η έξοδος του σταδίου k (όπου k = 1, 2, 3) είναι είσοδος για το στάδιο k+1.

ΚΕΦΑΛΑΙΟ 5

Υλοποίηση Μοντέλου Πρόβλεψης

5.1 Εισαγωγή

Ένα μοντέλο πρόβλεψης αποτελεί τη διαδικασία, που ακολουθείται, προκειμένου να παραχθούν προβλέψεις. Υπάρχει μεγάλη ποικιλία μοντέλων και η επιλογή του κατάλληλου είναι ιδιαίτερα σημαντική.

Όπως αναφέρθηκε και στην παράγραφο 3.2.4.2, το μοντέλο, που θα χρησιμοποιηθεί, είναι το μοντέλο γραμμικής παλινδρόμησης. Η γραμμική παλινδρόμηση κατασκευάζει μία γραμμική σχέση, ανάμεσα στην ανεξάρτητη μεταβλητή και τις εξαρτημένες. Αποτελεί το ιδανικό μοντέλο πρόβλεψης, διότι, καθώς το μοντέλο μας θα δέχεται τα δεδομένα πραγματικού χρόνου, είναι πολύ εύκολο να προβλέψει, μέσω μιας γραμμικής σχέσης, την τιμή της εξαρτημένης μεταβλητής.

Από το προηγούμενο κεφάλαιο επιλέχθηκε η διατήρηση μόνο μίας ανεξάρτητης μεταβλητής, της ακτινοβολίας, καθώς είχε το μεγαλύτερο βάρος, δηλαδή μπορεί να περιγράψει πολύ καλά την παραγωγή. Επομένως, η παραγωγή θα εξαρτάται μόνο από μια μεταβλητή (την ακτινοβολία) και η γραμμική παλινδρόμηση θα είναι απλή με συνάρτηση:

$$y = a \cdot x + b = a \cdot radiation + b$$

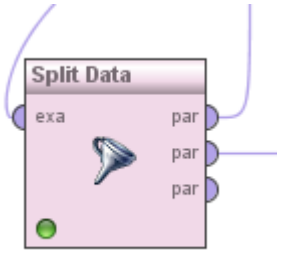
5.2 Επιλογή μοντέλου πρόβλεψης

Αυτή η παράγραφος θα αποτελέσει το 5^{ο} και τελευταίο Στάδιο της υλοποίησης, στο Rapidminer (Σχήμα 12).

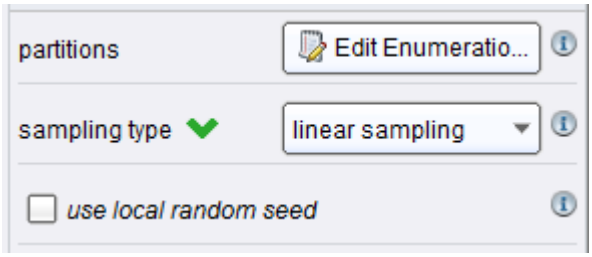
5.2.1 Διαχωρισμός δεδομένων

Το πρώτο βήμα της εφαρμογής του μοντέλου πρόβλεψης είναι ο διαχωρισμός των δεδομένων. Προκειμένου να ελεγχθεί η απόδοση του μοντέλου, χωρίζονται τα δεδομένα σε δεδομένα εκπαίδευσης και δεδομένα προς εξέταση. Τα δεδομένα εκπαίδευσης χρησιμοποιούνται, για την εκπαίδευση και δημιουργία του μοντέλου, ενώ τα δεδομένα προς εξέταση χρησιμοποιούνται για την αξιολόγηση του μοντέλου. Απαιτείται δηλαδή ο διαχωρισμός του συνόλου δεδομένων σε δύο υποσύνολα.

Για το διαχωρισμό των δεδομένων χρησιμοποιείται ο τελεστής “Split Data”, που ανήκει στους Μετατροπή Δεδομένων (Data Transformation) – Φιλτράρισμα (Filtering) – Δειγματοληψία (Sampling).



Σχήμα 60 Τελεστής "Split Data"



Σχήμα 61 Παράμετροι τελεστή "Split Data"

Ο τελεστής "Split Data" παράγει τον επιθυμητό αριθμό υποσυνόλων του συνολικού συνόλου δεδομένων, που δίνουμε σαν είσοδο. Διαμελίζει δηλαδή το σύνολο δεδομένων σε υποσύνολα, ανάλογα με καθορισμένα μεγέθη.

Δέχεται το σύνολο δεδομένων ως είσοδο (πύλη "exa") και δίνει τα υποσύνολα από τις πύλες εξόδου (πύλες "par"). Ανάλογα με το πόσα υποσύνολα διαθέτουμε, τόσες πύλες εξόδου χρησιμοποιούμε. Ο αριθμός των υποσυνόλων και το σχετικό τους μέγεθος καθορίζονται από την παράμετρο "partitions". Το άθροισμα της αναλογίας όλων των υποσυνόλων πρέπει να δίνει τη μονάδα. Η παράμετρος "sampling type" υποδεικνύει τον τρόπο, με τον οποίο τα παραδείγματα (σειρές) κατανέμονται στα τελικά υποσύνολα.

Από το Σχήμα 60 φαίνεται, η επιλογή της γραμμικής μεθόδου (linear sampling) για τον διαχωρισμό των δεδομένων. Αυτή η επιλογή, διαχωρίζει το σύνολο δεδομένων σε υποσύνολα, χωρίς να αλλάζει τη σειρά των παραδειγμάτων. Δηλαδή, δημιουργούνται υποσύνολα, τα οποία περιέχουν διαδοχικά παραδείγματα (σειρές).

Τον αριθμό των υποσυνόλων δεν τον ορίζουμε άμεσα εμείς οι χρήστες, αλλά υπολογίζεται, αυτόματα, από τον αριθμό των αναλογιών, που εισάγουμε στην παράμετρο "partitions". Για το μοντέλο που πρόκειται να δημιουργηθεί, επιλέξαμε να χωρίσουμε τα δεδομένα σε 80% δεδομένα εκπαίδευσης και 20% δεδομένα προς εξέταση, όπως ακριβώς φαίνεται και στο Σχήμα 62.

ratio
0.8
0.2

Σχήμα 62 Αναλογία δεδομένων για τον τελεστή "Split Data"

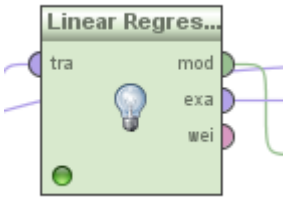
Τελικά, με αυτό τον τρόπο, καταφέραμε να πάρουμε από την πάνω έξοδο το 80% των δεδομένων και να τα χρησιμοποιήσουμε, για να φτιάξουμε και να εκπαιδεύσουμε το μοντέλο και από την κάτω είσοδο το 20%, για να το αξιολογήσουμε.

5.2.2 Εφαρμογή Γραμμικής Παλινδρόμησης

Το μοντέλο, που θα χρησιμοποιηθεί, αποτελεί μοντέλο παλινδρόμησης και πιο συγκεκριμένα μοντέλο γραμμικής παλινδρόμησης.

Η παλινδρόμηση είναι η τεχνική που χρησιμοποιείται για μία αριθμητική πρόβλεψη. Αποτελεί στατιστική μέτρηση, η οποία καθορίζει τη σχέση μεταξύ μιας εξαρτημένης μεταβλητής (όπως στην περίπτωσή μας είναι η παραγωγή), με μία σειρά από ανεξάρτητες. Το μοντέλο της γραμμικής παλινδρόμησης προσπαθεί να μοντελοποιήσει αυτή την εξάρτηση, τοποθετώντας μια γραμμική συνάρτηση στα ιστορικά δεδομένα.

Ο τελεστής, που υποστηρίζει αυτό το μοντέλο και ουσιαστικά υπολογίζει τη συνάρτηση γραμμικής παλινδρόμησης, είναι ο "Linear Regression". Ανήκει στους τελεστές Μοντελοποίηση (Modelling) – Ταξινόμηση και Παλινδρόμηση (Classification and Regression) – Ταίριασμα Συνάρτησης (Function Fitting).



Σχήμα 63 Τελεστής "Linear Regression"

Η είσοδος του τελεστή (πύλη "tra") δέχεται το υποσύνολο, που περιέχει τα δεδομένα προς εκπαίδευση. Δηλαδή, συνδέεται με την πάνω πύλη εξόδου του προηγούμενου τελεστή "Split Data". Επιπλέον, ο τελεστής δεν μπορεί να διαχειριστεί μη αριθμητικές τιμές και αυτός είναι ο λόγος, που στο προηγούμενο κεφάλαιο, δηλαδή στο προηγούμενο βήμα της διαδικασίας της πρόβλεψης, χρησιμοποιήθηκε ένας ξεχωριστός τελεστής, για την μετατροπή όλων των μεταβλητών σε αριθμητικές (παράγραφος 4.4.4.2).

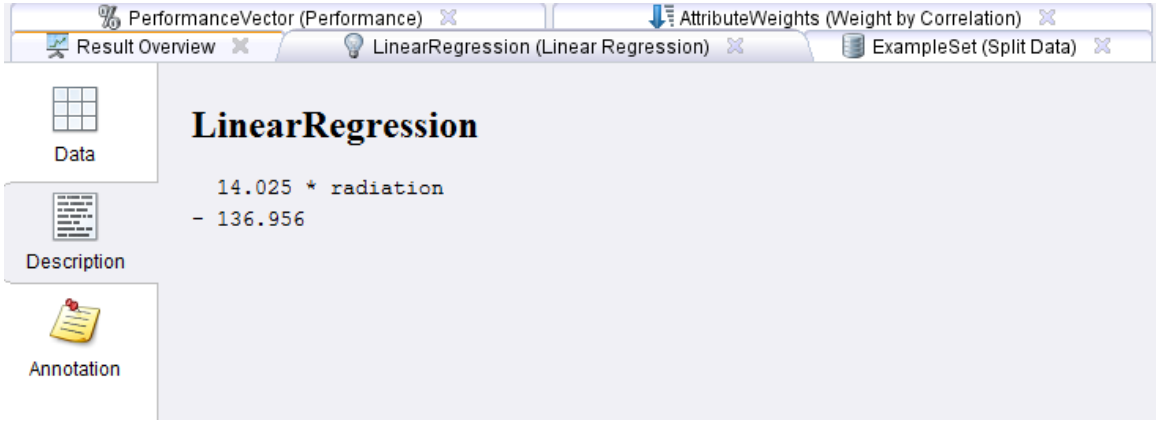
Έξοδο του τελεστή θα δώσουν οι δύο πάνω πύλες εξόδου ("mod", "exa").

Η πρώτη (πύλη "mod") δίνει το μοντέλο γραμμικής παλινδρόμησης, δηλαδή, εμφανίζει τους συντελεστές της συνάρτησης $y = a \cdot x + b$. Το μοντέλο, το οποίο λαμβάνεται, μπορεί να εφαρμοστεί σε άγνωστα δεδομένα, προκειμένου να γίνει η πρόβλεψη.

Από την δεύτερη πύλη (πύλη "exa") εμφανίζεται το υποσύνολο δεδομένων προς εκπαίδευση, όπως ακριβώς εισήχθη στην πύλη εισόδου.

Αν τρέξουμε τη διαδικασία, μέχρι αυτό το σημείο, θα δούμε στα αποτελέσματα της πύλης "mod" τους συντελεστές του μοντέλου (Σχήμα 64) καθώς και κάποιους δείκτες, οι οποίοι χαρακτηρίζουν τις μεταβλητές.

Σε αυτή την παράγραφο θα ασχοληθούμε μόνο με την περιγραφή (Description) του μοντέλου. Τα δεδομένα (Data) που εξάγει, δηλαδή οι δείκτες των μεταβλητών, θα παρουσιαστούν και θα σχολιασθούν αναλυτικά στο κεφάλαιο 6.



Σχήμα 64 Αποτελέσματα τελεστή "Linear Regression"

Η πρώτη τιμή είναι ο συντελεστής της ακτινοβολίας και η δεύτερη είναι η σταθερά. Δηλαδή η συνάρτηση παλινδρόμησης παίρνει τη μορφή:

$$y = 14.025 \cdot x - 136.956 \rightarrow$$
$$power = 14.025 \cdot radiation - 136.956$$

Εξίσωση 11: Εξίσωση μοντέλου απλής γραμμικής παλινδρόμησης

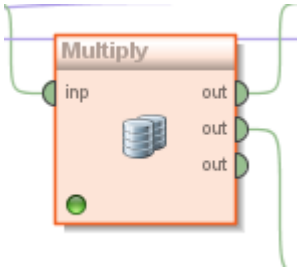
Η παραπάνω συνάρτηση αποτελεί το μοντέλο πρόβλεψης και μέσω αυτής, προβλέπεται κάθε μελλοντική τιμή της παραγωγής του φωτοβολταϊκού, αρκεί να είναι γνωστή κάθε χρονική στιγμή η ακτινοβολία.

Για τον συγκεκριμένο τελεστή δεν αναφερθήκαμε καθόλου στις παραμέτρους του. Από τις παραμέτρους του μπορούμε να ορίσουμε κάποιες συνθήκες/όρια, σύμφωνα με τα οποία, γίνεται η επιλογή των σημαντικών ανεξάρτητων μεταβλητών του μοντέλου. Από τη στιγμή όμως που διαθέτουμε μόνο μία ανεξάρτητη μεταβλητή, την ακτινοβολία, δεν χρειάστηκε να ορίσουμε κάποια συνθήκη.

5.2.3 Χρήση Πολλαπλασιαστή

Σε αυτό το σημείο απαραίτητη είναι η χρήση ενός τελεστή, ο οποίος θα αντιγράφει τα δεδομένα εισόδου, που δέχεται στην είσοδο και θα τα εξάγει, όσες φορές επιθυμεί ο χρήστης.

Κατάλληλος είναι ο τελεστής "Multiply", από τους Έλεγχος Διαδικασίας (Process Control). Ο τελεστής αυτός αντιγράφει, οτιδήποτε εισέρχεται από την πύλη εισόδου του στις πύλες εξόδου του. Καθώς περισσότερες πύλες συνδέονται, περισσότερα αντίγραφα δημιουργούνται. Το ιδιαίτερο χαρακτηριστικό του τελεστή είναι, ότι μία αλλαγή στις μεταβλητές ενός αντίγραφου δεν έχει καμία επίπτωση στα υπόλοιπα, ενώ μία αλλαγή στα δεδομένα ενός αντίγραφου επηρεάζει άμεσα και τα υπόλοιπα. Για παράδειγμα, όταν μία μεταβλητή αλλάζει ή προστίθεται σε ένα αντίγραφο του συνόλου δεδομένων, τότε αυτή η αλλαγή δεν έχει καμία επίδραση στα υπόλοιπα. Αντίθετα, αν μετατραπούν τα δεδομένα σε ένα αντίγραφο, τότε και όλα τα υπόλοιπα αντίγραφα, τα οποία δημιουργούνται από τον τελεστή, επηρεάζονται.



Σχήμα 65 Τελεστής "Multiply"

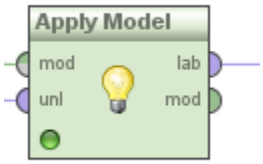
Στην ανάλυσή που εξετάζουμε, ο τελεστής "Multiply" δέχεται το μοντέλο, που δημιουργήθηκε στον τελεστή "Linear Regression" και παράγει δύο πανομοιότυπα μοντέλα στις εξόδους του.

Στόχος είναι η χρησιμοποίηση του ενός αντιγράφου στην εμφάνιση του μοντέλου, στα συνολικά αποτελέσματα και του άλλου στην πρόβλεψη και στην αξιολόγηση. Δηλαδή, το πρώτο αντίγραφο χρησιμοποιείται για την εμφάνιση όλων των χαρακτηριστικών του μοντέλου, στα αποτελέσματα, ενώ το δεύτερο συνδέεται με τον επόμενο τελεστή, που θα αναλύσουμε, προκειμένου να αξιολογηθεί το μοντέλο.

5.2.4 Εφαρμογή Μοντέλου Πρόβλεψης

Από τη στιγμή που έχει δημιουργηθεί μοντέλο μας, πρέπει να εφαρμοστεί στα δεδομένα. Χρησιμοποιείται ο τελεστής "Apply Model", τον οποίο βρίσκουμε στους Μοντελοποίηση (Modeling) – Εφαρμογή Μοντέλου (Model Application).

Ένα μοντέλο, πρώτα εκπαιδεύεται σε ένα σύνολο δεδομένων και οποιαδήποτε πληροφορία, σχετική με αυτά τα δεδομένα, μαθαίνεται από το μοντέλο. Στη συνέχεια, το ίδιο μοντέλο εφαρμόζεται σε ένα δεύτερο σύνολο δεδομένων για πρόβλεψη. Ουσιαστική λειτουργία του τελεστή είναι η εφαρμογή του ήδη γνωστού και εκπαιδευμένου μοντέλου, σε ένα σύνολο δεδομένων. Όλες οι απαραίτητες παράμετροι είναι αποθηκευμένες στο εσωτερικό του τελεστή, επομένως δεν χρειάζεται να εφαρμοστεί καμία ρύθμιση.



Σχήμα 66 Τελεστής "Apply Model"

Ο συγκεκριμένος τελεστής χρησιμοποιείται για την αξιολόγηση του μοντέλου και την πρόβλεψη της παραγωγής. Δέχεται το αντίγραφο του μοντέλου, από την πύλη εισόδου "mod", ενώ από την πύλη "unl", δέχεται το σύνολο δεδομένων, όπως παράχθηκε από τον τελεστή "Split Data" (παράγραφος 5.2.1), δηλαδή δέχεται το 20% του συνολικού όγκου. Είναι σημαντικό τα σύνολα δεδομένων, που εισέρχονται από κάθε πύλη εισόδου, να έχουν τον ίδιο αριθμό και την ίδια σειρά μεταβλητών. Επίσης, είναι απαραίτητο, ο τύπος και ο ρόλος των μεταβλητών να είναι και αυτός ίδιος.

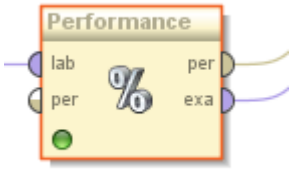
Γενικά, όσον αφορά τις πύλες εξόδου ενός τελεστή "Apply Model", η πύλη "lab" δίνει το ενημερωμένο σύνολο δεδομένων, μετά την εφαρμογή του μοντέλου στα δεδομένα εισόδου. Υπάρχουν κάποιες πληροφορίες, που προστίθενται στο σύνολο δεδομένων της εισόδου, προτού καταλήξει αυτό στην έξοδο. Για παράδειγμα, όταν εφαρμόζεται ένα μοντέλο πρόβλεψης, σε ένα σύνολο δεδομένων, μέσω του τελεστή "Apply Model", μία καινούρια μεταβλητή, με ρόλο πρόβλεψης, προστίθεται στα δεδομένα. Αυτή, αποθηκεύει τις προβλεπόμενες τιμές της μεταβλητής, που έχει το ρόλο «ετικέτα» (label). Από την άλλη πλευρά, η πύλη "mod" εμφανίζει το μοντέλο, όπως ακριβώς εισήχθη στην είσοδο του τελεστή.

Για τον τελεστή της ανάλυσής μας, η πύλη "mod" δεν συνδέεται πουθενά, καθώς δίνει το ίδιο αποτέλεσμα με το πρώτο αντίγραφο του πολλαπλασιαστή (Multiply), ενώ η πύλη

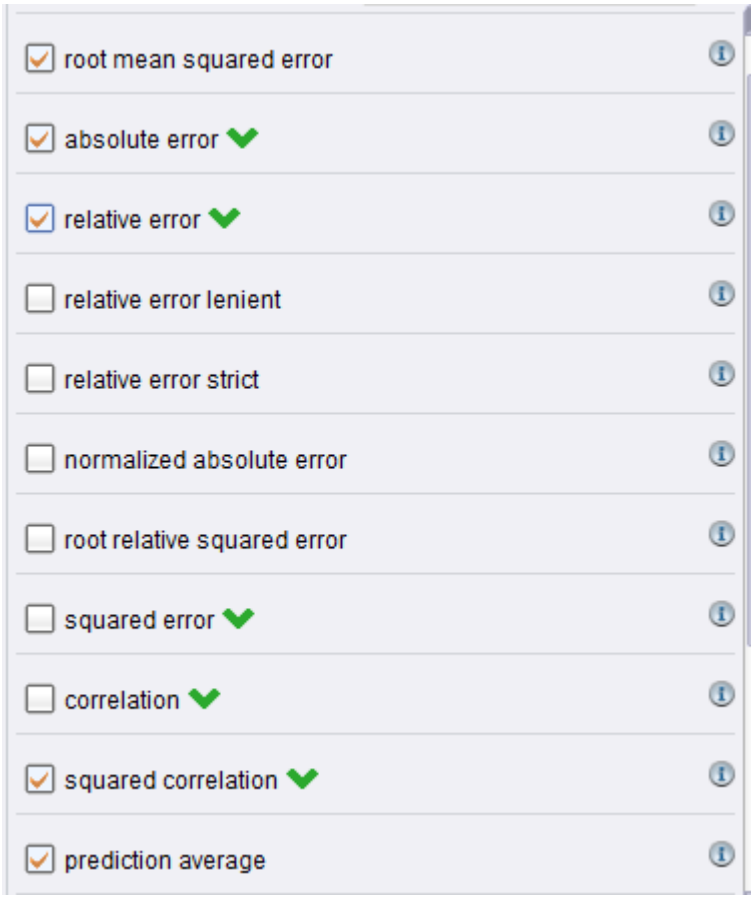
εξόδου "lab" οδηγείται, όπως θα αναλυθεί και στην επόμενη παράγραφο, σε έναν άλλο τελεστή, προκειμένου να αξιολογηθεί το μοντέλο. Η έξοδος που δίνει η "lab", είναι το ενημερωμένο σύνολο δεδομένων. Δηλαδή, είναι το σύνολο δεδομένων, όπως ακριβώς μπήκε στην είσοδο του τελεστή, μαζί με την πρόβλεψη της παραγωγής, ως καινούρια μεταβλητή.

5.2.5 Αξιολόγηση Απόδοσης μοντέλου

Όλη αυτή η ανάλυση και η σχεδίαση στο Rapidminer έχει σκοπό τη δημιουργία ενός αποδοτικού μοντέλου, για την πρόβλεψη της παραγωγής του ΦΒ. Την απόδοση του μοντέλου θα τη δώσει ο τελευταίος τελεστής, ο οποίος ονομάζεται "Performance" και ανήκει στους Αξιολόγηση (Evaluation) – Μέτρηση απόδοσης (Performance Measurement) – Ταξινόμηση και Παλινδρόμηση (Classification and Regression). Χρησιμοποιείται για την αξιολόγηση της απόδοσης παλινδρομικών πακέτων και αποδίδει μία λίστα κριτηρίων απόδοσης.



Σχήμα 67 Τελεστής "Performance"



Σχήμα 68 Παράμετροι τελεστή "Performance"

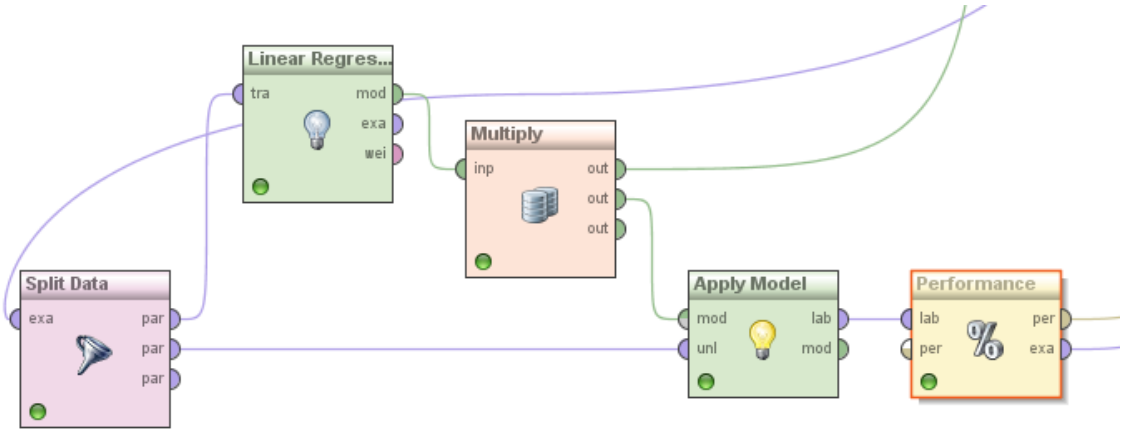
Η αξιολόγηση πραγματοποιείται με τη μέθοδο “out of sample”, σύμφωνα με την οποία η αξιολόγηση γίνεται στο 20% του συνολικό συνόλου δεδομένων.

Ο τελεστής δέχεται σαν είσοδο (πύλη “lab”) το ενημερωμένο σύνολο δεδομένων, δηλαδή αυτό που έχει και την πρόβλεψη. Έξοδο θα πάρουμε και από τις δύο πύλες του. Η πύλη “exa” μας δίνει αμετάβλητο το ενημερωμένο σύνολο δεδομένων και εμείς τη συνδέουμε στη συνολική έξοδο της διαδικασίας. Η πύλη “per”, δίνει ένα διάνυσμα απόδοσης, το οποίο ουσιαστικά είναι μία λίστα με τιμές, για τα κριτήρια απόδοσης. Το διάνυσμα απόδοσης υπολογίζεται από τη μεταβλητή «ετικέτα» και τη μεταβλητή της πρόβλεψης και τα αποτελέσματα,, εξαρτώνται από τις παραμέτρους, που επιλέχθηκαν. Δηλαδή όσες παράμετροι είναι επιλεγμένες (Σχήμα 68), αυτές θα εμφανιστούν και στα αποτελέσματα.

Όπως φαίνεται και από το Σχήμα 68, επιλέχθηκε η εμφάνιση των δεικτών: ρίζα μέσης τετραγωνικής απόκλισης “root mean squared error”, απόλυτο σφάλμα “absolute error”, σχετικό σφάλμα “relative error”, συντελεστής αυτοσυσχέτισης “squared correlation” και μέση πρόβλεψη “prediction average”, ως οι πιο σημαντικοί δείκτες για την αξιολόγηση ενός παλινδρομικού μοντέλου.

Οι παραπάνω δείκτες, καθώς και οι τιμές των αποτελεσμάτων τους, θα αναλυθούν στην επόμενη παράγραφο.

Τελικά το Στάδιο 5 που αποτελεί την επιλογή του μοντέλου πρόβλεψης, υλοποιήθηκε στο Rapidminer με τον παρακάτω τρόπο:



Σχήμα 69 Διαδικασία ανάλυσης από τελεστή "Filter Examples" έως "Performance" (Στάδιο 5, επιλογή μοντέλου πρόβλεψης)

Η συνολική υλοποίηση στο Rapidminer περιλαμβάνει τα στάδια 1 έως 5, συνδεδεμένα μεταξύ τους σε σειρά.

5.3 Παρουσίαση αποτελεσμάτων

Το μοντέλο που αναπτύχθηκε είναι μοντέλο γραμμικής παλινδρόμησης και συγκεκριμένα πρόκειται για απλή γραμμική παλινδρόμηση. Προκειμένου να εφαρμοστεί η μέθοδος απλής γραμμικής παλινδρόμησης είναι απαραίτητα η επιλογή της ανεξάρτητης μεταβλητής, η οποία επηρεάζει περισσότερο την παραγωγή. Όπως φάνηκε και στην παράγραφο 4.4.4.3, το μεγαλύτερο βάρος το είχε η ακτινοβολία (μεταβλητή "radiation"). Δηλαδή η παραγωγή εξαρτάται σε μεγάλο βαθμό από την ακτινοβολία. Η εξάρτηση της παραγωγή από την ακτινοβολία, περιγράφεται μέσω της εξίσωσης γραμμικής παλινδρόμησης, που αποτέλεσε έξοδο του τελεστή "Linear Regression":

$$power = 14.025 \cdot radiation - 136.956$$

Τα υπόλοιπα αποτελέσματα που έδωσε ο συγκεκριμένος τελεστής, φαίνονται στο Σχήμα 70 και αποτελούν τον *πίνακα συντελεστών της παλινδρόμησης*.

Attribute	Coefficient	Std. Error	Std. Coeffici...	Tolerance	t-Stat	p-Value	Code
radiation	14.025	0.011	0.994	1	1227.720	0	****
(Intercept)	-136.956	5.746	?	?	-23.836	0	****

Σχήμα 70 Πίνακας Συντελεστών Παλινδρόμησης

- Αρχικά, από τον δείκτη *coefficient* δίνονται οι συντελεστές του μοντέλου. Ένας θετικός συντελεστής δηλώνει ανάλογη σχέση μεταξύ ανεξάρτητης και εξαρτημένης μεταβλητής. Δηλαδή στην περίπτωσή του μοντέλου που αναπτύχθηκε, όπου ο συντελεστής της ακτινοβολίας (radiation) είναι θετικός, οποιαδήποτε αύξηση στην ακτινοβολία ισοδυναμεί με αύξηση στην παραγωγή (power).

- Επίσης, δίνονται τιμές για τον δείκτη *std error*, ο οποίος υποδεικνύει το "standard error" (σφάλμα) των συντελεστών. Χρησιμοποιείται επίσης για τη δημιουργία διαστημάτων εμπιστοσύνης, όπου διάστημα εμπιστοσύνης είναι ένα διάστημα εκτίμησης μιας παραμέτρου. Για παράδειγμα το διάστημα εμπιστοσύνης της ακτινοβολίας κατασκευάζεται από τον τύπο (14.025 ± k·0.011), όπου k είναι μία σταθερά η οποία εξαρτάται από τον βαθμό εμπιστοσύνης. Δηλαδή θεωρώντας βαθμό εμπιστοσύνης 95% τότε k = 1.96.

- Από τον δείκτη *Std Coefficient*, αντιλαμβανόμαστε πόσο συμβάλλει μία ανεξάρτητη μεταβλητή στην πρόβλεψη. Ο συγκεκριμένος δείκτης παίρνει τιμές από το μηδέν έως το ένα και όσο πιο κοντά στη μονάδα βρίσκεται τόσο σημαντικότερη είναι η ανεξάρτητη μεταβλητή. Από τη στιγμή που η παραγωγή εξαρτάται μόνο από μία μεταβλητή, την ακτινοβολία, τότε είναι λογικό να

εμφανίζεται τόσο υψηλό Std Coefficient. Αυτό σημαίνει ότι η μεταβλητή "radiation" συμβάλλει πολύ σημαντικά στην πρόβλεψη.

- Ο επόμενος δείκτης ονομάζεται Tolerance (ανεκτικότητα) και υποδεικνύει πόσο συσχετίζονται μεταξύ τους οι ανεξάρτητες μεταβλητές και σε ποιο βαθμό. Όσο η τιμή τείνει στο ένα τόσο χαμηλότερη συγγραμικότητα έχουμε. Εφόσον το μοντέλο διαθέτει μόνο μία μεταβλητή είναι λογικό η ανεκτικότητα να είναι ίση με τη μονάδα.

- Ακόμα ένας δείκτης που εμφανίζεται είναι ο στατιστικός δείκτης t ή όπως ονομάζεται στα αποτελέσματα *t-Stat.* Ο στατιστικός δείκτης t για ένα συγκεκριμένο τελεστή , αποτελεί την εκτίμηση σημαντικότητας του τελεστή με την παρουσία όλων των υπόλοιπων μεταβλητών [23]. Στο παράδειγμά μας διαθέτουμε μόνο μία μεταβλητή και γι' αυτό βγαίνει πολύ υψηλός ο δείκτης.

- Ο δείκτης *p-value* αποτελεί το ελάχιστο επίπεδο σημαντικότητας. Δηλαδή δείχνει αν μία ανεξάρτητη μεταβλητή δεν έχει στατιστικώς σημαντική προβλεπτική ικανότητα με την παρουσία άλλων μεταβλητών. Επομένως, μια μη σημαντική αξία *p* μπορεί να χρησιμοποιηθεί στην απομάκρυνση ορισμένων ανεξάρτητων μεταβλητών από την συνάρτηση πολλαπλής γραμμικής παλινδρόμησης. Ορίζουμε ένα κατώτερο επίπεδο σημαντικότητας, σύμφωνα με το οποίο γίνεται η σύγκριση κάθε τιμής *p*. Αν η τιμή *p* μίας μεταβλητής είναι μεγαλύτερη από το κατώτερο επίπεδο που ορίσαμε, τότε αυτή η μεταβλητή δεν είναι τόσο σημαντική για την εξίσωσή. Στην περίπτωσή της μεταβλητής "radiation" η τιμή του p-value, δεν ξεπερνάει αυτό το επίπεδο και επομένως θεωρείται ότι έχει υψηλή προβλεπτική ικανότητα.

- Ο τελευταίος δείκτης ονομάζεται *Code* και με τη βοήθεια του συμβόλου "*" , δείχνει πόσο σημαντική είναι κάθε μεταβλητή για την πρόβλεψη. Εξαρτάται άμεσα από τον δείκτη p-value. Όσα περισσότερα αστεράκια έχει μια μεταβλητή, τόσο σημαντικότερη είναι. Για άλλη μια φορά φαίνεται ότι η ακτινοβολία είναι αρκετά σημαντική.

Με την ανάλυση των προηγούμενων κεφαλαίων, είχαν συνδεθεί τέσσερις εξόδους στη συνολική διαδικασία.

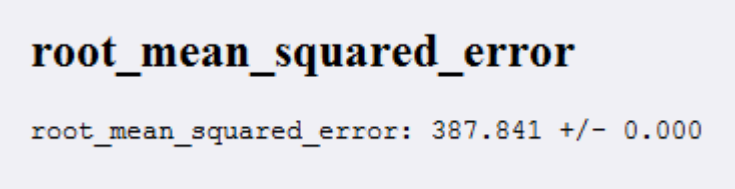
Η πρώτη έξοδος, προερχόταν από τον τελεστή "Select by Weights" και αφορούσε τα βάρη όλων των ανεξάρτητων μεταβλητών (Σχήμα 55)

Η δεύτερη έξοδος, προέρχεται από τον τελεστή "Multiply" και παρουσιάζει το μοντέλο όπως δημιουργήθηκε στον προηγούμενο τελεστή "Linear Regression". Αυτή η έξοδος περιεγράφηκε παραπάνω.

Η Τρίτη και η τέταρτη έξοδος προέρχονται και οι δύο από τις πύλες εξόδου του τελευταίου τελεστή, "Performance".

Από την πύλη "per" παίρνουμε τους δείκτες απόδοσης, τους οποίους επιλέξαμε να εμφανίζονται.

- Root Mean Squared Error (Ρίζα του Μέσου Τετραγωνικού Σφάλματος)



Σχήμα 71 Ρίζα του Μέσου Τετραγωνικού Σφάλματος (Root Mean Squared Error)

RMSE = ± 387.841

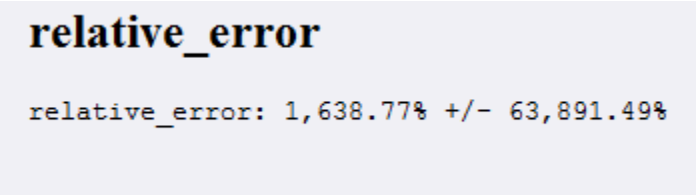
Αυτός ο δείκτης δείχνει το εύρος τιμών στο οποίο θα προβλέψει το μοντέλο και έχει μονάδες μέτρησης τις ίδιες με την παραγωγή. Αν για παράδειγμα η μεταβλητή "power" έχει πραγματική τιμή 5000, το μοντέλο θα προβλέψει 5000 ± 387.841.

- Absolute Error (Απόλυτο Σφάλμα)

Σχήμα 72 Απόλυτο Σφάλμα (Absolute Error)

Αποτελεί τη μέση απόλυτη απόκλιση της πρόβλεψης από την πραγματική τιμή. Οι μονάδες του είναι ίδιες με της παραγωγής.

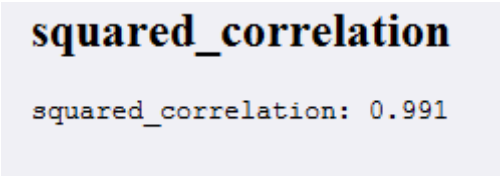
- Relative Error (Σχετικό Σφάλμα)



Σχήμα 73 Σχετικό Σφάλμα (Relative Error)

Πρόκειται για τη μέση απόλυτη απόκλιση της πρόβλεψης από την πραγματική τιμή προς τον αριθμό των παρατηρήσεων.

- Squared Correlation (Συντελεστής R^2)



Σχήμα 74 Συντελεστής R^2 (Squared Correlation)

Υποδεικνύει .την ποιότητα προσαρμογής της γραμμής παλινδρόμησης στα δεδομένα.

Στο συγκεκριμένο μοντέλο, $R^2 = 0.991$, το οποίο είναι πολύ ικανοποιητικό νούμερο. Αυτό σημαίνει ότι το 99.1% της διακύμανσης των δεδομένων ερμηνεύεται από την εξίσωση της παλινδρόμησης που έδωσε το μοντέλο [23].

Η μικρή απόκλιση που εμφανίζεται, μπορεί να οφείλεται σε τυχαίους παράγοντες των οποίων η πρόβλεψη είναι αδύνατη, σε μη γραμμική συσχέτιση μεταξύ των δύο μεταβλητών και στην εξάρτηση της παραγωγής από περισσότερες μεταβλητές.

- Prediction Average (Μέση Πρόβλεψη)

Σχήμα 75 Μέση Πρόβλεψη (Prediction Average)

Δίνει τον μέσο όρο όλων των προβλέψεων. Προστίθενται όλες οι προβλέψεις και διαιρούνται δια το πλήθος τους.

Συγκεντρωτικά, η απόδοση του μοντέλου εμφανίζεται στον παρακάτω πίνακα:

Πίνακας 2 Δείκτες απόδοσης μοντέλου πρόβλεψης

Root Mean Squared Error	± 387.841
Absolute Error	280.084 + / - 268.279
Relative Error	1,638.77% + / - 63,891.49%
Squared Correlation	0.991
Prediction Average	5839.981 + / - 4032.623

Από την πύλη εξόδου "exa", λαμβάνεται το σύνολο δεδομένων μαζί με την μεταβλητή της πρόβλεψης. Δηλαδή προκύπτει:

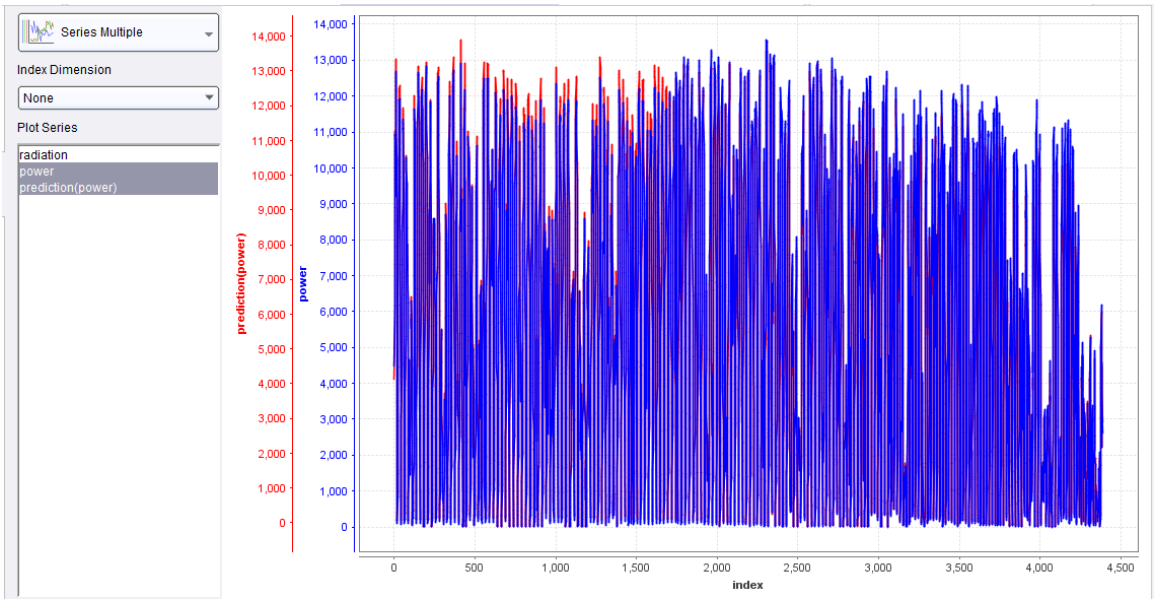
Row No.	datetime	power	prediction(p...	radiation
1	2014-05-08T10:00:00Z	4469.617	4135.811	304.652
2	2014-05-08T10:30:00Z	6140.148	5778.442	421.773
3	2014-05-08T11:00:00Z	7810.678	7421.073	538.894
4	2014-05-08T11:30:00Z	8898.084	8654.550	626.842
5	2014-05-08T12:00:00Z	9985.491	9888.028	714.790
6	2014-05-08T12:30:00Z	10457.208	10573.131	763.638
7	2014-05-08T13:00:00Z	10928.926	11258.235	812.487
8	2014-05-08T13:30:00Z	10071.440	10424.669	753.053

Σχήμα 76 Μεταβλητή της πρόβλεψης στο σύνολο δεδομένων

Η σύγκριση της πραγματικής παραγωγής με την προβλεπόμενη είναι εφικτή και από αυτή τη έξοδο. Αυτός είναι και ο λόγος που χωρίστηκαν τα δεδομένα σε δεδομένα προς εξέταση και δεδομένα προς εκπαίδευση. Φαίνεται οι προβλέψεις να είναι σχετικά κοντά με τις

πραγματικές τιμές. Προσθέτοντας και τον πολύ υψηλό συντελεστή R^2 που προέκυψε, μπορούμε να πούμε πως το μοντέλο μας έχει πολύ καλή απόδοση. Δηλαδή είναι δυνατή η πρόβλεψη της παραγωγής ηλεκτρικής ενέργειας του ΦΒ με μεγάλο ποσοστό επιτυχίας, για τα επόμενα χρόνια.

Ένας ακόμα τρόπος, πιο γραφικός, εμφάνισης της καλής απόδοσης του μοντέλου, είναι να τοποθετώντας την πραγματική παραγωγή και την πρόβλεψη πάνω στους ίδιους άξονες.



Σχήμα 77 Γραφική αναπαράσταση αποτελεσμάτων

Οι μπλε τιμές είναι οι πραγματικές τιμές της παραγωγής του ΦΒ, ενώ οι κόκκινες οι προβλεπόμενες. Παρατηρούνται κάποιες φυσιολογικές μικρές αποκλίσεις ανάμεσα στην προβλεπόμενη και την πραγματική παραγωγή.

Αυτή η υψηλή απόδοση του μοντέλου, οφείλεται κυρίως στο γεγονός ότι είχαμε στη διάθεσή μας μεγάλο όγκο δεδομένων και έκανε δυνατή τη δημιουργία ενός ικανοποιητικού μοντέλου.

ΚΕΦΑΛΑΙΟ 6

Συμπεράσματα και Προοπτικές

6.1 Συμπεράσματα

Η πρόβλεψη της παραγόμενης ισχύος από μία ανανεώσιμη πηγή ενέργειας (όπως είναι ένα φωτοβολταϊκό σύστημα) αποτελεί το ισχυρότερο εργαλείο με το οποίο μπορούν να εφοδιαστούν τα συστήματα ηλεκτρικής ενέργειας. Η ύπαρξη εκτιμήσεων με ακρίβεια είναι απαραίτητη για τη λειτουργία του συστήματος. Στο πλαίσιο αυτό στην παρούσα εργασία παρουσιάζεται και εφαρμόζεται μια μεθοδολογία πρόβλεψης της παραγωγής ηλεκτρικής ενέργειας από ΦΒ εγκατάσταση μέσω ανάλυσης πολλαπλών ροών δεδομένων.

Η εξόρυξη δεδομένων αφορά την εξόρυξη γνώσης από σύνολα δεδομένων σχετικά με χαρακτηριστικά και μοτίβα που μπορεί να εμφανίζουν. Το Rapidminer ως εργαλείο εξόρυξης δεδομένων, διαθέτει πολύ μεγάλο λειτουργικό εύρος και ευελιξία. Δεν απαιτεί τη χρήση κώδικα και περιλαμβάνει πληθώρα τελεστών και έτοιμων συναρτήσεων. Ένα μικρό δείγμα αυτών των τελεστών παρουσιάστηκε και στο πειραματικό κομμάτι της εργασίας. Κάποιες από τις λειτουργίες που υποστηρίζει το Rapaidminer, παρουσιάστηκαν στο Κεφάλαιο 2.

Στη συνέχεια, έγινε μία σύγκριση του Rapidminer με κάποια παρόμοια εργαλεία που κυκλοφορούν και αναφέρονται κάποια δυνατά σημεία του που το έχουν καταστήσει ηγέτη στον τομέα της εξόρυξης δεδομένων και των τεχνικών προβλέψεων. Επιπλέον, γίνεται μια βιβλιογραφική ανασκόπηση των βασικών μεθόδων πρόβλεψης, της έννοιας της χρονοσειράς και των παραγόντων οι οποίοι επηρεάζουν την επιλογή του κατάλληλου μοντέλου πρόβλεψης

Αναπτύσσεται μια μεθοδολογία πρόβλεψης για την παραγωγή ηλεκτρικής ενέργειας από μία ΦΒ εγκατάσταση, που αποτελείται από 2 βασικά στάδια. Το στάδιο της προ-επεξεργασίας δεδομένων και το στάδιο δημιουργίας του τελικού μοντέλου πρόβλεψης. Στο στάδιο της προ-επεξεργασίας, τα διαθέσιμα δεδομένα τα οποία συλλέχθηκαν σε διάστημα που εκτείνεται σε τρία ημερολογιακά έτη, υπόκεινται σε χειρισμό μηδενικών και κενών τιμών. Κατασκευάστηκε πλήρες σύνολο δυνατών συνδυασμών δεδομένων. Δημιουργήθηκαν 50 συνδυασμοί, είτε μέσω του πολλαπλασιασμού όλων των ανεξάρτητων μεταβλητών μεταξύ τους, είτε μέσω της ύψωσης των μεταβλητών σε κάποια δύναμη. Επιπλέον, μέσω της διεξαγωγής βαρών, επιλέχθηκε η μεταβλητή η οποία περιγράφει καλύτερα την παραγωγή.

Παρατηρήθηκε ότι ο χειρισμός μηδενικών και κενών τιμών εξαρτάται από τη φύση του προβλήματος. Στη συγκεκριμένη περίπτωση μηδενική τιμή ισοδυναμεί με βραδινή ώρα και επομένως μηδενική ακτινοβολία και παραγωγή. Στην περίπτωση των κενών τιμών, οι παρατηρήσεις γέμισαν με γραμμική παρεμβολή διότι όσο υπάρχει ηλιοφάνεια υπάρχει και παραγωγή.

Κατά την δημιουργία του τελικού μοντέλου, το αρχικό σύνολο δεδομένων χωρίζεται σε 2 υποσύνολο, εκ των οποίων το πρώτο χρησιμεύει για την εκπαίδευση του μοντέλου ενώ το

δεύτερο για τον έλεγχο της προβλεπτικής του ικανότητας. Το τελικό μοντέλο που προέκυψε είναι το y = 14.025·x -136.956

Στη συνέχεια, το μοντέλο αξιολογήθηκε μέσω κάποιων στατιστικών δεικτών. Τα αποτελέσματα που προέκυψαν κρίνονται ιδιαίτερα ικανοποιητικά με τον συντελεστή R^2 να έχει τιμή 99.1%.

Η πρόβλεψη προκύπτει κοντά στην πραγματική τιμή και αυτό φαίνεται και από τη γραφική αναπαράσταση των αποτελεσμάτων.

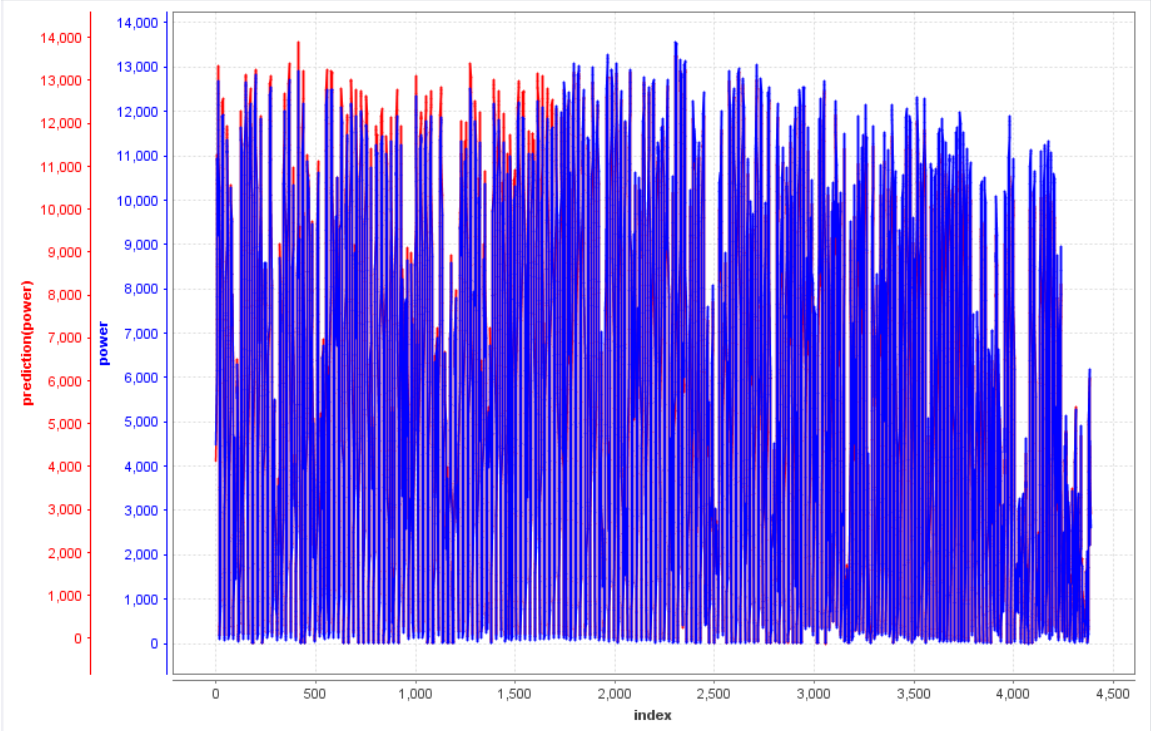


Figure 78 Γραφική αναπαράσταση αποτελεσμάτων μοντέλου πρόβλεψης

6.2 Προοπτικές

Κάποιες βελτιώσεις που θα βοηθούσαν στην ακρίβεια της πρόβλεψης, αφορούν το μοντέλο που χρησιμοποιήσαμε και το σύνολο δεδομένων που διαθέτουμε.

Υπάρχουν δύο είδη γραμμικής παλινδρόμησης: η απλή και η πολλαπλή. Στην παρούσα εργασία, επιλέχθηκε η εφαρμογή της απλής παλινδρόμησης, διατηρώντας έτσι τη μεταβλητή με το μεγαλύτερο βάρος. Μια διαφορετική προσέγγιση θα ήταν η επιλογή της πολλαπλής παλινδρόμησης. Ένας τρόπος για την εφαρμογή ενός μοντέλου πολλαπλής παλινδρόμησης, θα ήταν ο ορισμός ενός κατώτατου ορίου για τα βάρη των μεταβλητών και η διατήρηση μόνο αυτών που βρίσκονται πάνω από το όριο. Με αυτό τον τρόπο, η παραγωγή μας εξαρτιόταν και από άλλες ανεξάρτητες μεταβλητές και ίσως βελτιωνόταν η απόδοση του μοντέλου.

Μία άλλη αλλαγή που θα μπορούσε να εφαρμοστεί, διατηρώντας πάντα την ιδέα της πολλαπλής γραμμικής παλινδρόμησης, αφορά τον τελεστή "Weight by Correlation" (παράγραφος 4.4.4.3). Ο συγκεκριμένος τελεστής υπολογίζει το βάρος κάθε μεταβλητής, θεωρώντας πως οι υπόλοιπες δεν υπάρχουν. Δηλαδή θεωρεί πως δεν υπάρχει καμία συσχέτιση μεταξύ των μεταβλητών. Αντί όμως να υπολογίζεται το βάρος κάθε ανεξάρτητης μεταβλητής με έναν ξεχωριστό τελεστή, μία άλλη εναλλακτική θα ήταν η εισαγωγή όλων των ανεξάρτητων μεταβλητών στο μοντέλο της γραμμικής παλινδρόμησης και με βάση το δείκτη p-value να υπολογίζει μόνο του το μοντέλο ποιες μεταβλητές είναι σημαντικές για την πρόβλεψή. Σε αυτή την περίπτωση, θα έπρεπε να οριστεί ένα όριο στον δείκτη p-value και όσες μεταβλητές το ξεπερνούσαν θα διαγράφονταν από την εξίσωση παλινδρόμησης και συνεπώς από το μοντέλο.

Πέρα από τις καιρικές συνθήκες, υπάρχουν και άλλοι παράγοντες οι οποίοι επηρεάζουν την απόδοση ενός ΦΒ. Τέτοιοι παράγοντες, είναι για παράδειγμα η παλαιότητα του ΦΒ και η ανακλαστικότητα του πλαισίου. Επομένως μία βελτίωση του μοντέλου πρόβλεψης θα περιλάμβανε επιπλέον δεδομένα, δηλαδή επιπλέον ανεξάρτητες μεταβλητές οι οποίες θα καθόριζαν την παραγωγή.

Η εφαρμογή που αναπτύξαμε θα ήταν ιδιαίτερα χρήσιμη στην περίπτωση των real-time data. Δηλαδή θα μπορούσε το μοντέλο που δημιουργήσαμε να δέχεται δεδομένα πραγματικού χρόνου και να αυτό-διορθώνεται. Για παράδειγμα, εάν το μοντέλο, με την πάροδο του χρόνου εμφάνιζε όλο και μεγαλύτερες αποκλίσεις, με αποτέλεσμα να πέφτει και η απόδοσή του, θα μπορούσε να ξαναγυρνάει στο εσωτερικό του τελεστή "Linear Regression" και να επαναπροσδιορίζει τους συντελεστές της παλινδρομικής εξίσωσης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Mesut Ozkan (2011), "The comparison of data mining tools", Department of Computer Engineering İstanbul Kültür University

[2] rapidminer website, "Why did you choose rapidminer", τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε:
(https://rapidminer.com/)

[3] rapidminer website, "Difference between weka and Rapidminer", τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε:
(https://rapidminer.com/)

[4] Data Mining, Analytics, Big Data and Data Science (kdnuggets) website, "Interview with Rapidminer's Ingo Mierswa, Ralf Klinkenberg", part1, Feb 2010, τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε:
(http://www.kdnuggets.com/2010/02/f-interview-rapid-i-founders.html)

[5] Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, Sally Jo Cunningham, "Weka: Practical Machine Learning Tools and Techniques with Java Implementations", Department of Computer Science, University of Waikato, New Zealand

[6] Data Mining, Analytics, Big Data and Data Science (kdnuggets) website,"Rapidminer is not a version of Weka", Ingo Mierswa, 05 Dec 2007, τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε:
(http://www.kdnuggets.com/news/2007/n24/5i.html)

[7] Orange website, τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε:
(http://orange.biolab.si/)

[8] Orange website, "Orange screenshots" τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε:
(http://orange.biolab.si/screenshots/)

[9] John Fox and Robert Andersen, "Using the R statistical computing environment to teach social statistics courses", Department of Sociology, McMaster University, τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε:
(http://www.unt.edu/rss/Teaching-with-R.pdf)

[10] R-project website, "What is R?", τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε:
(http://www.r-project.org/about.html)

[11] R-project, "R screnshots", τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε:

(https://www.r-project.org/screenshots/screenshots.html)

[12] International Business Machines (IBM) Developer Works website, "Data mining with WEKA, part 2: Classification and clustering", Michael Abernethy , τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε: (http://www.ibm.com/developerworks/library/os-weka2/)

[13] laerd statistics website, "One way MANOVA in SPSS", τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε: (https://statistics.laerd.com/spss-tutorials/one-way-manova-using-spss-statistics.php)

[14] Statistical Analysis System (SAS) The Power to Know, "SAS Marketing Optimization: Plan, prioritize and optimize communications to maximize profits", τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε: (http://www.sas.com/offices/NA/canada/en/solutions/crm/mktopt/)

[15] slideshare website, "Data Mining tools (R, Weka, Rapidminer, Orange)", Mayur Surani, 26 Feb 2015, τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε: (http://www.slideshare.net/MayurSurani/data-mining-tools-45159317)

[16] Data Mining, Analytics, Big Data and Data Science (kdnuggets) website,"KDnuggets 15th Annual Analytics, Data Mining, Data Science Software Poll: Rapidminer Continues to Lead", Gregory Piatesky, 07 Jun 2014, τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε: (http://www.kdnuggets.com/2014/06/kdnuggets-annual-software-poll-rapidminer-continues-lead.html)

[17] Gartner website, "Magic Quadrant for Advanced Analytics Platforms", τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε: (http://www.gartner.com/technology/reprints.do?id=12AHPOU0&ct=150225&st=sb)

[18] Charles Elkan, "Predictive analytics and data mining" May 28, 2013

[19] Γ.Σιδεράτος, (2010), «Ανάπτυξη Μοντέλων Πρόβλεψης Παραγωγής Αιολικής Ισχύος Με Χρήση Νευρωνικών Δικτύων και Τεχνικών Ασαφούς Λογικής», Διδακτορική Εργασία, Εθνικό Μετσόβιο Πολυτεχνείο Σχόλη Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Τομέας Ηλεκτρικής Ισχύος, Αθήνα.

[20] Χ.Καραμέρος, (2013), «Πιθανοτική Πρόβλεψη Ηλιακής Παραγωγής με Χρήση Artmap» Διπλωματική Εργασία, Εθνικό Μετσόβιο Πολυτεχνείο Σχόλη Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Τομέας Ηλεκτρικής Ισχύος, Αθήνα.

[21] European Photovoltaic Industry Association (EPIA) website, "Global Market Outlook for Photovoltaics 2014-2018", τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε:

(http://www.epia.org/fileadmin/user_upload/Publications/EPIA_Global_Market_Outlook_for_Photovoltaics_2014-2018_-_Medium_Res.pdf)

[22] Frauhofer Institute for Solar Energy Systems (ISE) website, "Photovoltaics report", Freiburg, 24 October 2014, τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε: (http://www.ise.fraunhofer.de/de/downloads/pdf-files/aktuelles/photovoltaics-report-in-englischer-sprache.pdf)

[23] Φ.Πετρόπουλος & Β.Ασημακόπουλος,(2013), «Επιχειρησιακές Προβλέψεις», εκδόσεις Συμμετρία, Αθήνα.

[24] Gerard E. Dallal, "How to read the Output from Multiple Linear Regression Analyses", τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε: (http://www.jerrydallal.com/lhsp/regout.htm)

[25] Rapidminer website, "Rapidminer Products,Studio", τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε: (https://rapidminer.com/products/studio/)

[26] Π.Λαδάς, (2014), «Βραχυπρόθεσμη Πρόβλεψη Ενεργειακής Ζήτησης, Προσεγγίσεις Βασισμένες στη Μηχανική Μάθηση», Διπλωματική Εργασία,Εθνικό Μετσόβιο Πολυτεχνείο Σχόλη Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Τομέας Ηλεκτρονικών Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων, Αθήνα.

[27] zentut website,"Data Mining Techniques", τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε: (http://www.zentut.com/data-mining/data-mining-techniques/)

[28] Kenneth D. Bailey (1994), "Typologies and Taxonomies, an Introduction to Classification Techniques", University of California Los Angeles.

[29] Christopher J Matheus, Philip K. Chan, Gregory Pialetsky-Shapiro, "Systems for Knowledge Discovery in Databases", GTE laboratories Incorporated, Waltham.

[30] "Photovoltaic and Solar Forecasting State of the Art Report IEA PVPS Task 14, Subtask 3.1, Oct 2013" Authors Sophie Pelland, Jan Remund, Jan Kleissl, Takashi Oozeki, Karel De Brabandere.

[31] «Συσχέτιση και Γραμμική Παλινδρόμηση», Αλεξάνδρειο Τεχνολογικό Εκπαιδευτικό Ίδρυμα Θεσσαλονίκης Τμήμα Πληροφορικής Εργαστήριο Θεωρία Πιθανοτήτων και Στατιστική, τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε: (http://aetos.it.teithe.gr/~vkostogl/files/Statistiki/ARXEIA%20THEORIAS/ERG-STAT_Simeioseis%20Palindromisi.pdf)

[32] Usama M. Fayyad, Gregory Piattsky-Shapiro, Padhraic Smyth, Ramasamy Uthurusamy, "Advances in Knowledge Discovery and Data Mining", American Association for Artificial Intelligence, Jan 1996.

[33] Butler analytics website, "Rapidminer review" (2015), τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε: (http://butleranalytics.com/rapidminer-review/)

[34] British Broadcasting Corporation (bbc) , "Geography, Synoptic charts and weather", τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε: (http://www.bbc.co.uk/bitesize/standard/geography/weather_climate/synoptic_charts/revision/2/)

[35] Data Mining, Analytics, Big Data and Data Science (kdnuggets) website, "Free Data mining Software: Rapidminer 4.0 (formerly YALE)", Ingo Mierswa, τελευταία πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε: (http://www.kdnuggets.com/news/2007/n15/8i.html)

[36] «Ποσοτικές Μέθοδοι στη Διοίκηση Επιχειρήσεων» website, «Δειγματοληπτική Συσχέτιση» πρόσβαση 27 Ιουλίου 2015, διαθέσιμο σε: (http://androulakis.bma.upatras.gr)