ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Π.Μ.Σ.: “Ψηφιακές Επικοινωνίες & Δίκτυα”

“Το RapidMiner ως Εργαλείο εφαρμογών Big Data Analytics”

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
ΦΟΙΤΗΤΗΣ: ΤΖΕΚΑΣ ΑΛΕΞΑΝΔΡΟΣ
Α.Μ.: ΜΨΕ 1615
ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΑΠΟΣΤΟΛΟΣ ΜΗΛΙΩΝΗΣ

ΑΘΗΝΑ 2018

UNIVERSITY OF PIRAEUS

DEPARTMENT OF DIGITAL SYSTEMS

Postgraduate Programme: “Digital Communications & Networks”

"RapidMiner as a Big Data Analytics Application Tool"

MASTER’S THESIS
STUDENT: TZEKAS ALEXANDROS
A.M.: ΜΨΕ 1615
SUPERVISOR: APOSTOLOS MILIONIS

ATHENS 2018

“Το RapidMiner ως Εργαλείο εφαρμογών Big Data Analytics”

Τζέκας Αλέξανδρος
Πτυχίο Φυσικής, Ε.Κ.Π.Α. 2016

Διπλωματική Εργασία
υποβαλλόμενη για τη μερική εκπλήρωση των απαιτήσεων του
ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΤΙΤΛΟΥ ΣΠΟΥΔΩΝ ΣΤΙΣ ΨΗΦΙΑΚΕΣ ΕΠΙΚΟΙΝΩΝΙΕΣ & ΔΙΚΤΥΑ

"RapidMiner as a Big Data Analytics Application Tool"

Tzekas Alexandros
Degree in Physics, National & Kapodistrian University of Athens

Master's thesis
submitted for the partial fulfillment of the requirements of the
POSTGRADUATE STUDY OF DIGITAL COMMUNICATIONS & NETWORKS

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον κ. Απόστολο Μηλιώνη, επίκουρο Καθηγητή του τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς, για την καθοδήγηση και υποστήριξη κατά την εκπόνηση της παρούσας διπλωματικής εργασίας, αλλά και για την πολύ καλή συνεργασία που είχαμε όλο αυτό το χρονικό διάστημα.

ΤΖΕΚΑΣ ΑΛΕΞΑΝΔΡΟΣ
Αθήνα 2018

Special Thanks

I would like to express my sincere thanks to Mr. Apostolos Milionis, assistant professor of the Digital Systems Department of Piraeus University, for guidance and support in the preparation of this diploma thesis, but also for the very good cooperation we have had throughout this period.

TZEKAS ALEXANDROS
Athens 2018

ΠΕΡΙΛΗΨΗ

Η ραγδαία αύξηση της τεχνολογίας και οι ρηξικέλευθες προτάσεις της Πληροφορικής Επιστήμης έχουν επιφέρει ριζικές αλλαγές στην καθημερινότητα και κυρίως στους διάφορους τομείς της επαγγελματικής μας δραστηριότητας. Η Αναλυτική Μεγάλων Δεδομένων (Big Data Analytics) καθώς και η Εξόρυξη Δεδομένων (Data Mining) προμηνύουν μία νέα Εποχή για την Ανθρωπότητα όπου η στατιστική επεξεργασία πληροφοριών με σκοπό την εξαγωγή ποιοτικών συμπερασμάτων αλλά και προβλέψεων θα αποτελεί στοιχείο απόλυτα συνυφασμένο με την επιχειρηματική δραστηριότητα, τη βιομηχανία, τη ναυτιλία, την ιατρική και γενικότερα με οιαδήποτε έκφανση του πολιτισμού μας.

Αντικείμενο της παρούσας Διπλωματικής Εργασίας είναι η χρήση ενός open source Framework για την επεξεργασία Big Data. Στόχος μας, η δημιουργία ενός Μοντέλου Πρόβλεψης που θα χαρακτηρίζεται από μεγάλη ακρίβεια καθώς και η τοποθέτηση Συναγερμών ούτως ώστε να ειδοποιείται ο ενδιαφερόμενος σε περίπτωση που κάποια ή κάποιες απ' τις μεταβλητές του εξεταζόμενου Συστήματος λάβουν τιμές εκτός των επιτρεπτών ορίων. Προς την επίτευξη των παραπάνω, επιλέχθηκε το λογισμικό περιβάλλον του RapidMiner, στο οποίο εισήχθη ένα Data Set υπό μορφή Excel από τις μετρήσεις που συλλέχθηκαν σε χρονικό διάστημα ενός χρόνου και οι οποίες αφορούν στη λειτουργία ενός Φωτοβολταϊκού Πάρκου.

Η Διπλωματική Εργασία χωρίζεται ως εξής: Στο πρώτο Κεφάλαιο επιχειρείται μία θεωρητική εισαγωγή στη Σύγχρονη Πληροφορική (Data Analytics, Big Data, Data Mining, IoT, Cloud), στο δεύτερο παρατίθενται κάποιες προαπαιτούμενες γνώσεις που κρίνονται απαραίτητες για την εκτέλεση της Πειραματικής Διαδικασίας η οποία πραγματοποιείται στο τρίτο και τέταρτο Κεφάλαιο (όπου καλούμαστε να προβλέψουμε την Παραγόμενη Ισχύ βάσει της προσπίπτουσας Ακτινοβολίας, να θέσουμε κάποιους Συναγερμούς ως προς τη Θερμοκρασία, την Υγρασία και τον Άνεμο και να εκτιμήσουμε την κριτική ταινιών από ενδεχόμενους Users καθώς και τους πιο πιθανούς συνδυασμούς αγοράς προϊόντων σε supermarket), και τέλος στο πέμπτο Κεφάλαιο έχουν καταχωρηθεί κάποια γενικά Συμπεράσματα που προέκυψαν μέσω της Πειραματικής Διαδικασίας.

Λέξεις Κλειδιά

Big Data Analytics, Data Analytics, Data Mining, KDD, Text Mining, Web Mining, Image Mining, Picture Mining, Video Mining, Music Mining, Time Series Data Mining, Spatial Data Mining, Data Integration, Data Selection, Data Cleansing & Normalization, Data Transformation, Pattern Evaluation, Knowledge Presentation, Predictive Data Mining Tasks, Descriptive Data Mining Tasks, Data Analytics Models, Descriptive Analytics, Diagnostic Analytics, Predictive Analytics, Prescriptive Analytics, Classification, Clustering, Association, Regression, Forecasting, Time Series Analysis, Sequential Pattern Discovery, Summarization, Support, Confidence, Neural Networks, Feedforward Neural Networks, SVM, Naive Bayes, k-NN, Decision Trees, Γενετικός Αλγόριθμος, Internet of Things, Cloud, Apache Mahout, Apache Hadoop, Apache Spark, Talent, Orange, Rattle, Java, R, Scala, Python, Φωτοβολταϊκό Πάρκο, Rapid Miner, Rapid Miner Studio, Rapid Miner Server, Repository, Operators.

ABSTRACT

The rapid growth of technology and the groundbreaking proposals of Informatics have brought about radical changes in everyday life and especially in the various areas of our professional activity. Big Data Analytics, as well as Data Mining, promulgate a new era for humanity where the statistical processing of information to produce qualitative conclusions and predictions will be an element inherent in business activity, industry, shipping, medicine and more generally in any aspect of our culture.

The subject of this Diploma Thesis is to use an open source Framework for editing Big Data. Our goal is to create a highly accurate forecasting model as well as an installation of Alarms to alert the interested party if some of the variables of the System under consideration are outside the permissible limits. To achieve the above, the RapidMiner software environment was selected, in which a Data Set in Excel format was introduced from the measurements collected over a period of one year, which concern the operation of a Photovoltaic Park.

The Diploma Thesis is divided as follows: The first chapter attempts a theoretical introduction to Modern Informatics (Data Analytics, Big Data, Data Mining, IoT, Cloud), the second provides some prerequisite knowledge that is necessary for the implementation of the Experimental Process which is done in the third and fourth Chapter (where we are supposed to predict Output Power on the basis of incident Radiation, to set some Temperatures, Humidity and Wind Alarms, and to predict the film rating of potential users and the most likely combinations of supermarket products) and, finally, in the fifth Chapter some General Conclusions that have emerged through the Experimental Procedure have been registered.

Keywords

Big Data Analytics, Data Analytics, Data Mining, KDD, Text Mining, Web Mining, Image Mining, Picture Mining, Video Mining, Music Mining, Time Series Data Mining, Spatial Data Mining, Data Integration, Data Selection, Data Cleansing & Normalization, Data Transformation, Pattern Evaluation, Knowledge Presentation, Predictive Data Mining Tasks, Descriptive Data Mining Tasks, Data Analytics Models, Descriptive Analytics, Diagnostic Analytics, Predictive Analytics, Prescriptive Analytics, Classification, Clustering, Association, Regression, Forecasting, Time Series Analysis, Sequential Pattern Discovery, Summarization, Support, Confidence, Neural Networks, Feedforward Neural Networks, SVM, Naive Bayes, k-NN, Decision Trees, Γενετικός Αλγόριθμος, Internet of Things, Cloud, Apache Mahout, Apache Hadoop, Apache Spark, Talent, Orange, Rattle, Java, R, Scala, Python, Photovoltaic Park, Rapid Miner, Rapid Miner Studio, Rapid Miner Server, Repository, Operators.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

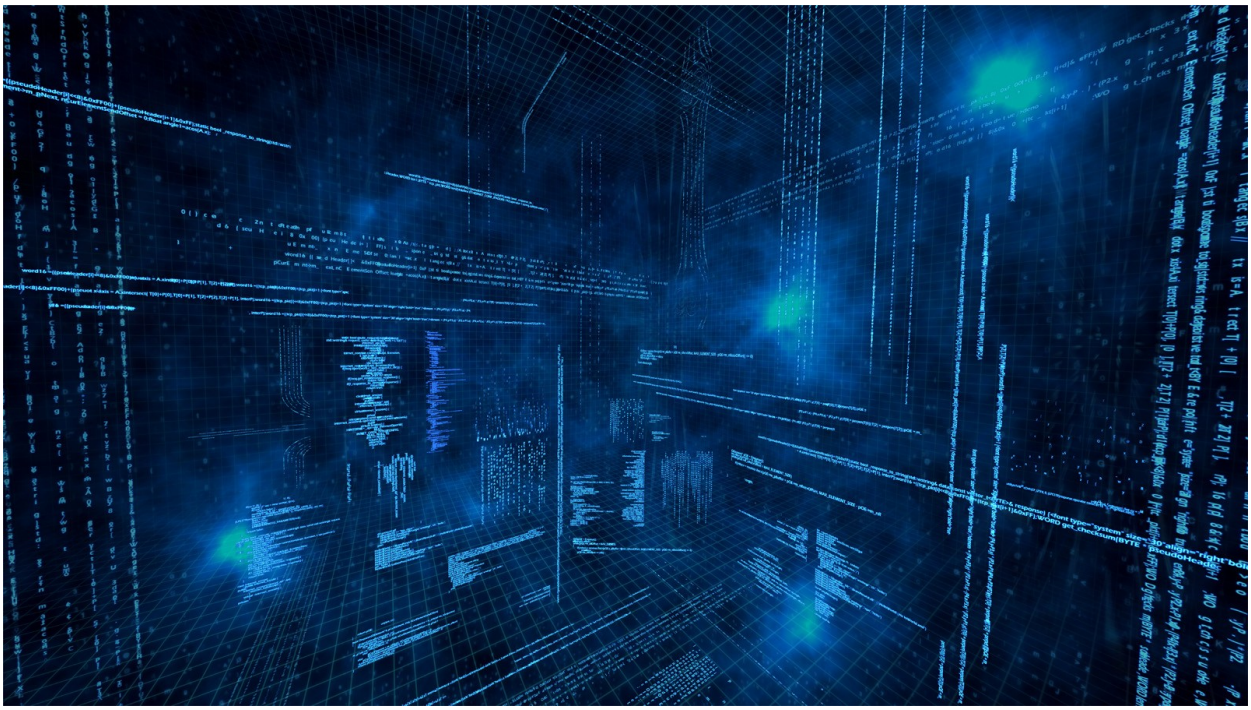
ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΜΑΘΗΜΑΤΙΚΩΝ ΣΧΕΣΕΩΝ

#	ΕΞΙΣΩΣΗ
1	$\text{Support} = \sigma(X \bigcap Y) / N$
2	$\text{Confidence} = \sigma(X \bigcap Y) / \sigma(X)$
3	$\text{Lift}(X,Y) = P(X \bigcap Y) / [P(X) \cdot P(Y)]$
4	$P(t)=\{x_1^2 ,....,x_n^t\}$
5	$4(1H) \mapsto 4He + \text{energy} + 2\ \text{neutrinos}$
6	$E = Mc^2$
7	$P = \varepsilon \cdot \sigma \cdot T \cdot (4 \cdot \pi \cdot R_H^2)$
8	$G_{sc} = P / (4 \cdot \pi \cdot R_o^2)$
9	$G_{on} = Gsc \cdot [1 + 0{,}033 \cdot \cos((360 \cdot n) / 365)]$
10	$I = I_b + I_d$
11	$I_T = I_{b,T} + I_{d,T} + I_{refl,T}$
12	$\theta_z = 90° - \alpha_s$
13	$m = \cos\theta_z^{-1}$
14	$AM_\alpha = AM \cdot (P / P_o)$
15	$\delta = 23{,}45 \cdot \sin[360 \cdot (284 + n) / 365]$
16	$t_s = t_c + \theta / 15 - T_c + 3{,}82 \cdot \{0{,}00075 + 0{,}001868 \cdot \cos[360 \cdot (n-1) / 365] - 0{,}032077 \cdot \sin[360 \cdot (n-1) / 365] - 0{,}014615 \cdot \cos[2 \cdot 360 \cdot (n-1) / 365] - 0{,}004089 \cdot \sin[2 \cdot 360 \cdot (n-1) / 365]\}$
17	$\omega = (t_s - 12) \cdot 15$
18	$K_T = I / I_{oh}$
19	$I_b = I - I_d$
20	$I_T = I_{b,T} + I_{d,T} + I_{refl,T}$
21	$E = h \cdot f = (h \cdot c) / \lambda$
22	$\lambda_g = 1{,}238 / E_g$
23	$RE = \sum_{i=1}^{n} \lvert p_i - d_i \rvert / (\sum_{i=1}^{n} d_i / n)$
24	$AE = \sum_{i=1}^{n} \lvert y_i - p_i \rvert / n$
25	$R^2 = [\sum_{i=1}^{n} (y_i - \langle d \rangle)^2] / [\sum_{i=1}^{n} (d_i - \langle d \rangle)^2]$
26	$_n$

	$RMSE = \{[\sum_{i=1}^{n}(y_i - p_i)^2]/ n\}^{1/2}$
27	$MSE = [\sum_{i=1}^{n}(y_i - p_i)^2]/ n$
28	$ME = [\sum_{i=1}^{n}(y_i - p_i)] / n$

ΚΕΦΑΛΑΙΟ 1: ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ



1.1 Εισαγωγή

Σ’ αυτό το κεφάλαιο θα αναφερθούμε σε κάποιες απ’ τις πλέον καινοτόμες και σύγχρονες τεχνολογικές τάσεις της πληροφορικής, δίνοντας μεγαλύτερη έμφαση στα Big Data Analytics και στο Data Mining. Ακολουθούν πέντε υποκεφάλαια, καθένα απ’ τα οποία καλύπτει ένα διαφορετικό πεδίο της θεωρητικής γνώσης που είναι απαραίτητη για την εκπόνηση της συγκεκριμένης διπλωματικής εργασίας, ξεκινώντας απ’ τα Data Analytics, τα Big Data Analytics, το Data Mining, το IoT και καταλήγοντας σε μία μικρής έκτασης αναφορά στο Cloud. Για την πληρέστερη κατανόηση των παραπάνω, παρατίθενται παραδείγματα εφαρμογής αυτών στην καθημερινότητα καθώς και μαθηματικά μοντέλα και εικονικές αναπαραστάσεις όπου αυτό κρίνεται προσοδοφόρο.

1.2 Data Analytics

Τα Data Analytics είναι η διεργασία με την οποία εξετάζουμε διάφορα σετ δεδομένων με τη χρήση εξειδικευμένων συστημάτων και λογισμικού έχοντας ως σκοπό την εξαγωγή ποσοτικών και ποιοτικών συμπερασμάτων αλλά και χρήσιμων προβλέψεων. Η τεχνολογία των Data Analytics εφαρμόζεται σήμερα σε πάρα πολλές εμπορικές βιομηχανίες με σκοπό την καλύτερη λήψη αποφάσεων που βασίζεται στην καλύτερη προγενέστερη πληροφόρηση. Επίσης χρησιμοποιείται από ερευνητές για την επιβεβαίωση ή την απόρριψη επιστημονικών μοντέλων, θεωριών και υποθέσεων.

Αποτελεί αδιαμφισβήτητο γεγονός πως οι εταιρείες οι οποίες ασχολούνται ενεργά με τον τομέα των Data Analytics, αυξάνουν τα κέρδη τους, βελτιώνουν την επιχειρησιακή τους λειτουργία, οργανώνουν αποδοτικότερες και πιο στοχευμένες ως προς το αγοραστικό τους κοινό διαφημιστικές καμπάνιες, βελτιώνουν σε μεγάλο βαθμό την εξυπηρέτηση των πελατών τους, αντιδρούν ταχύτερα σε αλλαγές της αγοράς και τέλος, ως συνέπεια αυτών, καταφέρνουν να έχουν ανταγωνιστικό πλεονέκτημα. Συναρτήσει με την φύση της εφαρμογής, τα δεδομένα τα οποία αναλύονται μπορούν να είναι είτε “ιστορικά” (δηλαδή πληροφορίες που βρίσκονται αποθηκευμένες επί παραδείγματι σε κάποια Βάση), είτε real-time δεδομένα τα οποία υπόκεινται σε real-time επεξεργασία με τους κατάλληλους μηχανισμούς. Επιπροσθέτως, μπορούν να γίνουν και μελέτες από μία μίξη εσωτερικών συστημάτων και εξωτερικών πηγών δεδομένων.

Οι μεθοδολογίες των Data Analytics σχετίζονται με την ανεύρεση ακολουθιών συμπεριφοράς και σχέσεων ανάμεσα στα εξεταζόμενα δεδομένα στα οποία μπορεί να εφαρμοστεί στατιστική ανάλυση και να αποδειχθεί αν μία υπόθεση εργασίας για ένα σετ δεδομένων είναι αληθής ή ψευδής. Τα Data Analytics συνήθως χωρίζονται σε δύο βασικές κατηγορίες:

- την ποσοτική ανάλυση των δεδομένων η οποία περιλαμβάνει συνήθως αριθμητικά δεδομένα με μετρήσιμες μεταβλητές που δύνανται να μετρηθούν ή/και να συγκριθούν μέσω στατιστικής,
- και την ποιοτική ανάλυση αυτών η οποία είναι περισσότερο ερμηνευτικού χαρακτήρα και επικεντρώνεται στην κατανόηση του περιεχομένου των δεδομένων

(συνήθως μη αριθμητικών) όπως είναι για παράδειγμα κείμενα, εικόνες, αρχεία ήχου και βίντεο.

Από την άλλη το κομμάτι που έχει να κάνει με το business intelligence, παρέχει στους διευθυντές και σε άλλα μέλη της εταιρείας, την δυνατότητα να έχουνε πρακτική πληροφορία σε σχέση με κρίσιμους στατιστικούς δείκτες απόδοσης όπως είναι οι λειτουργίες της εταιρείας, το πελατολόγιο κτλ. Οι σύγχρονοι οργανισμοί χρησιμοποιούν όλο και περισσότερο ατομικής χρήσης εργαλεία, τα οποία επιτρέπουν στους διευθυντές, στους ειδικούς αναλυτές και στο λοιπό προσωπικό να διεξάγουν τα δικά τους queries και να φτιάχνουν από μόνοι τους τις δικές τους αναφορές.

Πιο προηγμένες τεχνικές Data Analytics περιλαμβάνουν το Data Mining, το οποίο δίνει τη δυνατότητα αναγνώρισης καινούργιων τάσεων, ακολουθιών συμπεριφοράς και σχεσιακών μοντέλων. Τα predictive analytics αναφέρονται στους προβλεπτικούς αλγόριθμους και συνήθως προσπαθούν να προβλέψουν την πιθανή συμπεριφορά ενός πελάτη, αστοχίες στους εξοπλισμούς και πιθανά μελλοντικά γεγονότα. Έτσι ακριβώς λειτουργεί και το machine learning, ένας κλάδος που εμπίπτει στο πεδίο της τεχνητής νοημοσύνης και ο οποίος χρησιμοποιεί αυτοματοποιημένους αλγορίθμους για να διαβάσει τα σετ δεδομένων στον ελάχιστο δυνατό χρόνο. Τα Big Data analytics έχουν άμεση σχέση με το data mining, που αναφέρθηκε νωρίτερα και συνήθως εμπίπτουν σε σετ δεδομένων, τα οποία έχουν μη δομημένα ή ημι-δομημένα δεδομένα σαν εγγραφές. Αντίστοιχα το text mining είναι ένα μέσο για ανάλυση αναγνώσιμων αρχείων, μηνυμάτων ηλεκτρονικού ταχυδρομείου, μηνυμάτων κινητής τηλεφωνίας κτλ.

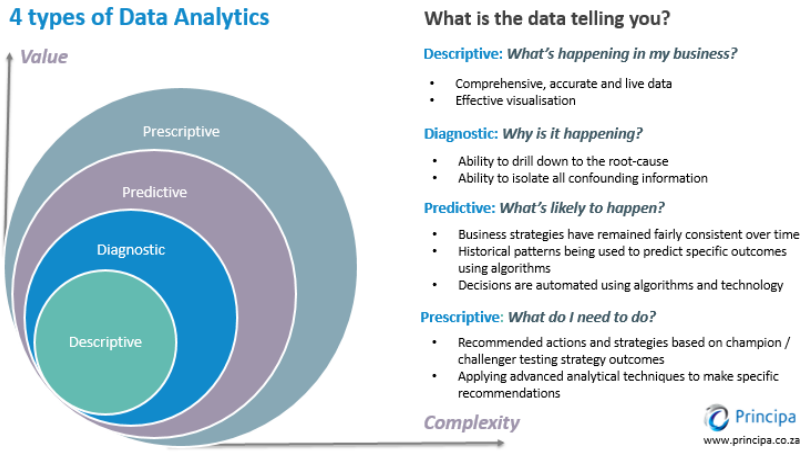
Σήμερα ένα μεγάλο μέρος των σύγχρονων επιχειρήσεων αναζητά υποστήριξη στην υπηρεσία των Data Analytics (εικόνα 1).
Απλά παραδείγματα αποτελούν οι τραπεζικοί οργανισμοί οι οποίοι εφαρμόζουν παρακολουθήσεις ακολουθίας συμπεριφοράς στα συστήματα αναλήψεων των λογαριασμών και των χρεώσεων των πιστωτικών καρτών, για να αποτρέψουν πιθανές απάτες ή και κλοπή προσωπικών δεδομένων. Ακόμα ένα παράδειγμα αποτελούν οι εταιρείες ηλεκτρονικού εμπορίου, αλλά και οι πάροχοι υπηρεσιών μάρκετινγκ, οι οποίοι αναλύουν τα click-stream για να συγκεντρώσουν πολύτιμες πληροφορίες σχετικά με την ταυτότητα των πελατών τους και τις αναζητήσεις τους ως προς τα προϊόντα της εταιρείας. Αντιστοίχως οι πάροχοι κινητής τηλεφωνίας εξετάζουν την κίνηση αλλά και τον όγκο των δεδομένων των πελατών τους σε σχέση με την πρόγνωση του καιρού ή με τις μετακινήσεις του πληθυσμού, για να προβλέψουν τυχόν αστοχίες στην επικοινωνία λόγω συγκέντρωσης μεγάλου αριθμό χρηστών κάτω από την ίδια κυψέλη (bootleneck).

Οι διαδικασίες που χρησιμοποιούνται σε εφαρμογές Data Analytics περιλαμβάνουν πολύ περισσότερες διεργασίες από την απλή ανάλυση των δεδομένων όπως για παράδειγμα την συλλογή, την ενσωμάτωση και τη δοκιμή των αναλυτικών μοντέλων με σκοπό την αξιολόγηση αυτών. Συνήθως, η διαδικασία ξεκινά με τη συλλογή των δεδομένων, τα οποία οι ερευνητές θέτουν υπό ανάλυση ώστε να αναγνωριστεί η πληροφορία που αποτελεί το αντικείμενο της έρευνας. Στην περίπτωση που η συλλογή δεδομένων από διαφορετικά συστήματα κρίνεται απαραίτητη, χρησιμοποιούνται ρουτίνες ενσωμάτωσης˙ τα δεδομένα μετατρέπονται σε μία κοινή μορφή και εν συνεχεία φορτώνονται σε ένα σύστημα ανάλυσης όπως είναι για παράδειγμα ο Hadoop cluster, η NoSQL κ.α..

Ακολουθεί η επίλυση προβλημάτων που σχετίζονται με την ποιότητα και την ακρίβεια της ανάλυσης των εφαρμογών. Αυτή η διαδικασία περιλαμβάνει την διεξαγωγή του data

profiling και του data cleansing, και στοχεύει στην εξάλειψη της πιθανότητας τα δεδομένα να είναι ασυνεχή και στην διαγραφή πιθανών λαθών που προέρχονται από διπλοεγγραφές. Αντίστοιχη εργασία διεξάγεται και στο data preparation με σκοπό τη διαχείριση και την οργάνωση των δεδομένων σύμφωνα με τις data governance πολιτικές που έχει προεπιλεγεί να εφαρμοστούν. Από αυτό το σημείο και μετά ξεκινάει η πραγματική διαδικασία, κατά την οποία καθορίζεται το αναλυτικό μοντέλο που θα χρησιμοποιηθεί.

Σε κάποιες περιπτώσεις οι εφαρμογές analytics μπορούν να ενεργοποιήσουν αυτόματα δράσεις, όπως για παράδειγμα κατά την αγορά ή πώληση μετοχών σε ένα χρηματιστήριο εφόσον αυτό κρίνεται απαραίτητο. Σε οποιαδήποτε άλλη περίπτωση το τελευταίο βήμα της διαδικασίας είναι η κοινοποίηση των εξαγόμενων αποτελεσμάτων στους προϊσταμένους και στους διευθυντές της εταιρείας ώστε να καταστούν χρήσιμα ως προς τη χάραξη στρατηγικής. Αυτό συνήθως γίνεται με την βοήθεια του data visualization, το οποίο είναι μία διαδικασία που παράγει και δημιουργεί διαγράμματα και στατιστικές αναλύσεις, ώστε να είναι κατανοητό το τελικό αποτέλεσμα. Τα data visualization είναι συνήθως διαθέσιμα στο business intelligence dashboard της εφαρμογής και μπορούν να απεικονίσουν τα δεδομένα αλλά και τις αλλαγές αυτών σε πραγματικό χρόνο, για όσο χρονικό διάστημα παρατηρείται καινούργιες πληροφορίες να ρέουν προς το σύστημα.



Εικόνα 1: Οι τέσσερις τύποι των Data Analytics

1.3 Big Data Analytics

1.3.1 Εισαγωγή

Τα τελευταία χρόνια έχει παρατηρηθεί ένα αυξανόμενο ενδιαφέρον για την Διαχείριση Μεγάλων Δεδομένων. Η πρώτη εμφάνιση του όρου Big Data έγινε το 1997 από τους επιστήμονες της NASA. Ανέφεραν ότι αδυνατούσαν να αναπαραστήσουν γραφικά (visualization) τα σύνολα δεδομένων που κατείχαν (data sets), καθώς ήταν τόσο μεγάλα που ήταν ακατόρθωτο να τα αποθηκεύσουν στη κύρια μνήμη, στον τοπικό δίσκο και σε εξωτερικό σκληρό δίσκο. Έτσι δήλωσαν ότι αντιμετωπίζουν πρόβλημα Μεγάλων Δεδομένων. Οι τελευταίες τεχνολογικές εξελίξεις κυρίως στον τομέα των επικοινωνιών και των ολοκληρωμένων κυκλωμάτων έχουν δώσει την δυνατότητα να δημιουργηθούν μηχανισμοί παρακολούθησης των λειτουργιών ενός οργανισμού σε πολύ λεπτομερές επίπεδο. Η λεπτομερής αυτή ψηφιοποίηση των διαδικασιών παραγωγής έχουν καταστήσει μεγάλους οργανισμούς αλλά και εταιρείες μικρού μεγέθους ικανούς να παράγουν τεράστιους όγκους δεδομένων με πολύ ταχείς ρυθμούς. Τα δεδομένα αυτά κρύβουν πολύτιμη γνώση καθώς η ανάλυση τους μπορεί να οδηγήσει σε σημαντικές βελτιστοποιήσεις της παραγωγής.

Ψηφιακά δεδομένα συναντώνται πλέον παντού: σε κάθε τομέα, σε κάθε οικονομία, σε κάθε οργανισμό και χρήστη της ψηφιακής τεχνολογίας. Τα μεγάλα δεδομένα έλκουν όλο και περισσότερο το ενδιαφέρον των ηγετών από όλους τους τομείς, ενώ οι καταναλωτές προϊόντων και υπηρεσιών αναμένεται να ωφεληθούν από την αξιοποίησή τους. Η ικανότητα αποθήκευσης, συγκέντρωσης, συνδυασμού δεδομένων και η χρήση των αποτελεσμάτων για την εκπόνηση λεπτομερών αναλύσεων έχει γίνει πολύ πιο προσιτή και εφικτή. Με λιγότερο από 600 $, κάποιος μπορεί να αγοράσει μια μονάδα δίσκου με ικανότητα να αποθηκεύσει όλη τη μουσική του κόσμου. Επίσης, τα μέσα εξόρυξης γνώσης από τα δεδομένα σημειώνουν σημαντική βελτίωση, καθώς τα διαθέσιμα λογισμικά για την εφαρμογή τεχνικών αυξανόμενης πολυπλοκότητας συνδυάζονται με την αυξανόμενη υπολογιστική ισχύ. Επιπλέον, η δυνατότητα παραγωγής, επικοινωνίας, μερισμού και πρόσβασης δεδομένων έχει εκτοξευθεί από την αύξηση του αριθμού των ατόμων, των συσκευών και των αισθητήρων, που συνδέονται σήμερα σε ψηφιακά δίκτυα. Το 2010, περισσότερα από 4 δισεκατομμύρια άνθρωποι, ή το 60 % του παγκόσμιου πληθυσμού, χρησιμοποιούσαν κινητά τηλέφωνα, και περίπου το 12 % από αυτούς ήταν κάτοχοι smartphone (η διείσδυση των οποίων αυξάνεται κατά περισσότερο από 20 τοις εκατό το χρόνο). Περισσότεροι από 30 εκατομμύρια δικτυωμένοι κόμβοι αισθητήρων βρίσκονται πλέον στους κλάδους μεταφορών, αυτοκινητοβιομηχανίας, επιχειρήσεων κοινής ωφέλειας, καθώς και σε τομείς του λιανικού εμπορίου. Ο αριθμός αυτών των αισθητήρων αυξάνεται σε ποσοστό άνω του 30 %.

Πολλές τεχνολογικές καινοτομίες έχουν οδηγήσει σε δραματική αύξηση των δεδομένων και στη συλλογή αυτών. Αυτός είναι ο λόγος που τα μεγάλης κλίμακας δεδομένα αποτελούν πλέον περιοχή των στρατηγικών επενδύσεων για τους IT οργανισμούς.

Δεν είναι όμως μόνο οι οργανισμοί που παράγουν τεράστιους όγκους δεδομένων. Ακόμη και σε μικρότερη κλίμακα οργάνωσης, στο επίπεδο του ατόμου, η παραγωγή δεδομένων είναι πρωτόγνωρη. Οι περισσότεροι άνθρωποι διαθέτουν έναν ψηφιακό εαυτό, ως προβολή των δραστηριοτήτων τους στα κοινωνικά δίκτυα. Η Google εκτιμά ότι κάθε δύο

μέρες το ψηφιακό υλικό που δημιουργείται από τους χρήστες είναι ισομεγέθες με το έντυπο υλικό που παρήγαγε η ανθρωπότητα από την αρχή της γραφής μέχρι το 2003. Έκρηξη στον όγκο των παραγόμενων δεδομένων παρατηρείται ακόμη στην επιστημονική έρευνα. Τομείς, όπως η ιατρική, η αστρονομία, η μετεωρολογία αλλά και η βιολογία χάρη στις νέες τεχνολογίες, τα νέα τηλεσκόπια, τους νέους και φτηνούς αισθητήρες και τα νέα μηχανήματα για την αποκωδικοποίηση DNA μπορούν και παράγουν όγκους δεδομένων που δεν είναι δυνατόν να αντιμετωπιστούν με τις υπάρχουσες υποδομές. Μάλιστα έχει παρατηρηθεί πως οι ρυθμοί αύξησης είναι εκθετικής κατανομής. Έτσι προβλέπεται για τα επόμενα χρόνια μια ακόμη μεγαλύτερη “έκρηξη πληροφορίας”. Αν και δεν υπάρχει ένα όριο μεγέθους δεδομένων πάνω από το οποίο να δύναται να θεωρηθούν Big Data, συνήθως με το συγκεκριμένο όρο αναφερόμαστε σε όγκους δεδομένων που κυμαίνονται από μερικά terabytes έως δεκάδες ή και εκατοντάδες zetabytes (1.073.741.824 terabytes)

Παρακάτω παραθέτουμε κάποιες πληροφορίες σχετικά με την εντυπωσιακή αύξηση των δεδομένων τα τελευταία χρόνια:

- Το 2011, η ανθρωπότητα δημιούργησε πάνω από 1,2 τρισεκατομμύρια GigaBytes δεδομένων.
- Ο όγκος των δεδομένων αναμένεται να αυξηθεί 50 φορές μέχρι το 2020.
- Η Google λαμβάνει πάνω από 2.000.000 ερωτήματα αναζήτησης κάθε λεπτό.
- 72 ώρες βίντεο προστίθενται στο YouTube κάθε λεπτό.
- Υπάρχουν 217 νέοι χρήστες του Ίντερνετ κάθε λεπτό.
- Οι χρήστες του Twitter στέλνουν πάνω από 100.000 tweets κάθε λεπτό (που είναι πάνω από 140 εκατομμύρια ανά ημέρα).
- Εταιρείες, και οργανισμοί λαμβάνουν 34.000 “likes” σε κοινωνικά δίκτυα κάθε λεπτό.
- Η International Data Corporation (IDC) προβλέπει ότι η αγορά για την τεχνολογία των μεγάλης κλίμακας δεδομένων και υπηρεσιών θα φτάσει τα 16,9 εκατομμύρια δολάρια.

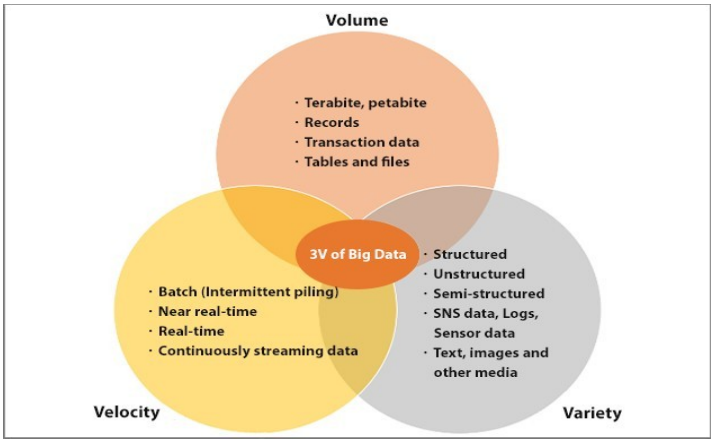
1.3.2 Ορισμοί των Big Data

Η Gartner, η μεγαλύτερη επιχείρηση στον κόσμο που ασχολείται με την τεχνολογική έρευνα και τη συμβουλευτική, το 2012 έδωσε τον εξής ορισμό: «Τα Big Data είναι υψηλού όγκου, υψηλής ταχύτητας ή υψηλής ποικιλίας στοιχεία που απαιτούν αποδοτικές και καινοτόμες μορφές επεξεργασίας πληροφοριών». Στα Big Data συγκαταλέγονται όλες οι πληροφορίες των social media που είναι προσβάσιμες από το ευρύ κοινό και βρίσκονται στο Διαδίκτυο, δηλαδή φωτογραφίες, video και κείμενα, καθώς και όλα τα «απόρρητα δεδομένα» των διαφόρων εταιριών αλλά και των κυβερνήσεων. Έτσι, η Gartner πρότεινε έναν ορισμό που περιλαμβάνει τα τρία Vs: Volume, Velocity, Variety ή αλλιώς τον Όγκο, την Ταχύτητα και την Ποικιλία. Πρόκειται για έναν ορισμό που εστιάζει στο μέγεθος. Η έκθεση επισημαίνει το αυξανόμενο μέγεθος των δεδομένων, το αυξανόμενο ποσοστό παραγωγής τους και το αυξανόμενο εύρος των μορφών που εφαρμόζονται.
Παρακάτω αναφερόμαστε περιληπτικά σε καθέναν απ’ τους τρεις παραπάνω όρους:

- Volume: Ο όγκος των δεδομένων που καλούμαστε να διαχειριστούμε. Με την πάροδο των ετών η αναλογική πληροφορία (έντυπης μορφής πληροφορία)

άρχισε να αντικαθίσταται όλο και περισσότερο με ψηφιακή. Πλέον,έχουμε τεράστιες ποσότητες δεδομένων σε μορφή video, ήχου και εικόνων τόσο σε επιστημονικές εφαρμογές όσο και στα ευρέως πλέον διαδεδομένα κοινωνικά δίκτυα. Είναι πολύ συνηθισμένο μια επιχείρηση να διαθέτει Terabytes ή ακόμα και Petabytes αποθηκευτικού χώρου. Χαρακτηριστικά, η αποθηκευμένη πληροφορία παγκοσμίως αυξάνεται με ρυθμούς τετραπλάσιους από ότι η παγκόσμια οικονομία, ενώ η επεξεργαστική δύναμη των υπολογιστών εννέα φορές γρηγορότερα. Ξεπερνώντας σιγά σιγά τα προβλήματα εύρεσης επαρκούς χώρου αποθήκευσης, νέα ζητήματα αναδύονται όπως η ανάγκη συσχέτισης των Big Data και η δυνατότητα αλίευσης πολύτιμης αξίας.

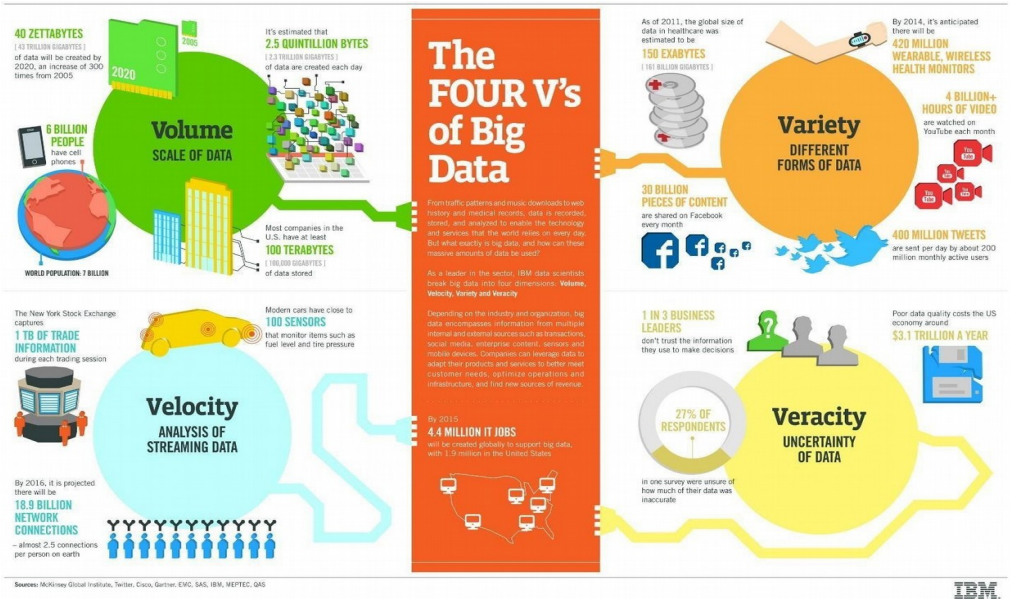
- Velocity: Ο όρος Velocity αναφέρεται στον ταχύτατο ρυθμό με τον οποίο εισέρχονται νέα δεδομένα αλλά και ανανεώνονται τα ήδη υπάρχοντα. Επιπλέον, σχετίζεται με τον χρόνο που απαιτείται για την επεξεργασία και την ανάλυση τους κατά την είσοδό τους στο σύστημα. Πρέπει να σημειωθεί σ' αυτό το σημείο πως δεν αρκεί να μπορούμε να αναλύουμε τα δεδομένα και να εξάγουμε πληροφορία σε πραγματικό χρόνο, αλλά επιπλέον είναι αναγκαίο να εκτελούμε και όλες τις λειτουργίες που ενεργοποιούνται από αυτά (σε πραγματικό χρόνο) ώστε να μην καθυστερεί η όλη διαδικασία. Για παράδειγμα μια εφαρμογή που παρακολουθεί τις τιμές των εταιρειών στο χρηματιστήριο είναι πιθανόν να χρειαστεί να προβεί σε αγοραπωλησίες μετοχών αν επί παραδείγματι αυτές υπερβούν κάποια προκαθορισμένη τιμή.
- Variety: Με τον όρο Variety αναφερόμαστε στο μεγάλο εύρος διαφορετικών τύπων δεδομένων που καλούμαστε να επεξεργαστούμε. Ο όρος ανταποκρίνεται στην ανάγκη να καταχωρούμε, να επεξεργαζόμαστε και να συνδυάζουμε δεδομένα διαφορετικών πηγών. Αυτό έχει ως αποτέλεσμα να ερχόμαστε αντιμέτωποι όχι μόνο με διαφορετικούς τύπους δεδομένων, αλλά και με διαφορετική δομή μεταξύ των ιδίων τύπων. Σε πρώτη φάση, δημιουργείται έτσι η ανάγκη να ενσωματωθούν δεδομένα αυστηρώς δομημένα (structured), ημιδομημένα (semi-structured) και αδόμητα (unstructured). Σε δεύτερη φάση, ακόμα και αν οι πηγές μας χρησιμοποιούν αυστηρή δόμηση των δεδομένων, πιθανόν να χρησιμοποιούν διαφορετική σημασιολογία, ή να υπάρχει ασυμβατότητα μεταξύ τους.



Εικόνα 2: Τα τρία Vs

Ο ορισμός αυτός έχει επαναληφθεί από τη NIST (Nist Big Dataprogram, 2013) και διευρυνθεί από την IBM (IBM, 2013) για να συμπεριλάβει και ένα τέταρτο V˙ τη Veracity

(πιστότητα). Η Veracity αναφέρεται στις προκαταλήψεις, το θόρυβο και τη messiness ή την αξιοπιστία των δεδομένων. Λόγω των πολλών και διαφορετικών μορφών των Μεγάλων Δεδομένων, η ποιότητα και η ακρίβεια είναι παράμετροι που δύσκολα ελέγχονται και αξιολογούνται. Καθώς, όμως, η αξία των Big Data έγκειται κυρίως στο γεγονός ότι μπορούν να οδηγήσουν σε καλύτερες αποφάσεις, είναι απαραίτητο να διαθέσουμε ένα μεγάλο χρονικό διάστημα στη διαλογή και τον καθαρισμό τους. Επί παραδείγματι, δεν μπορούμε να θεωρήσουμε ένα tweet ως αξιόπιστο δεδομένο προς ανάλυση, καθώς ενδέχεται να είναι ψευδές, να υπάρχουν πολλοί λογαριασμοί από τον ίδιο χρήστη κλπ. Εν τούτοις, η ανάλυση των Big Data και η σύγχρονη τεχνολογία, μας επιτρέπουν πλέον να συνεργαστούμε με αυτούς τους τύπους των δεδομένων.



Εικόνα 3: Τα τέσσερα Vs

Η Oracle αποφεύγει την χρήση των Vs για να καταλήξει σε έναν ορισμό. Υποστηρίζει ότι τα μεγάλα στοιχεία είναι η δημιουργία αξίας από παραδοσιακές σχεσιακές βάσεις δεδομένων με στόχο τη λήψη επιχειρηματικών αποφάσεων, η οποία είναι εμπλουτισμένη με νέες πηγές μη- δομημένων δεδομένων. Οι νέες αυτές πηγές περιλαμβάνουν blogs, social media, δίκτυα αισθητήρων, δεδομένα εικόνας και άλλες μορφές δεδομένων, τα οποία ποικίλλουν σε μέγεθος, δομή, μορφή και άλλους παράγοντες. Σύμφωνα με την Oracle τα Big Data είναι το αποτέλεσμα που προκύπτει από την ένταξη πρόσθετων πηγών δεδομένων ώστε να αυξηθούν οι ήδη υπάρχουσες λειτουργίες. Αξίζει να σημειωθεί ότι ο ορισμός της Oracle εστιάζει στην υποδομή ενώ δίδεται ιδιαίτερη βαρύτητα σε μια σειρά από τεχνολογίες όπως: NoSQL, Hadoop, HDFS, R και σχεσιακές βάσεις δεδομένων. Το μειονέκτημα του εν λόγω ορισμού είναι ότι δεν διαθέτει σαφή ποσοτικά κριτήρια.

Η Intel είναι μία από τις λίγες επιχειρήσεις που παρέχουν ποσοτικά στοιχεία στη βιβλιογραφία τους. Αντί να δώσει έναν ορισμό όπως έκαναν οι προαναφερθέντες οργανισμοί, περιγράφει τα Big Data ποσοτικοποιώντας τις εμπειρίες των επιχειρηματικών εταίρων της. Επισημαίνει ότι οι οργανισμοί οι οποίοι μελετήθηκαν ασχολούνται εκτενώς με μη- δομημένα δεδομένα και δίνουν έμφαση στη διεξαγωγή αναλύσεων των δεδομένων

τους τα οποία παράγονται με ρυθμό 500 terabytes ανά εβδομάδα . Τέλος, ισχυρίζεται ότι ο πιο σύνηθες τύπος δεδομένων που συναντάται είναι οι επιχειρηματικές συναλλαγές που είναι αποθηκευμένες σε σχεσιακές βάσεις δεδομένων (σύμφωνα με τον ορισμό της Oracle), και ακολουθούν τα έγγραφα, τα e- mail, τα blogs και τα social media.

Η Microsoft παρέχει ένα ιδιαίτερα περιεκτικό ορισμό: “Big Data είναι ο όρος που χρησιμοποιείται όλο και περισσότερο για να περιγράψει τη διαδικασία εφαρμογής σημαντικής υπολογιστικής ισχύος - την τελευταία λέξη της μηχανικής μάθησης και της τεχνητής νοημοσύνης - σε μαζικά και εξαιρετικά πολύπλοκα σύνολα πληροφοριών”. Ο ορισμός αυτός καθιστά σαφές ότι τα Big Data απαιτούν σημαντική υπολογιστική ισχύ. Η σημασία της υπολογιστικής ισχύος αναφέρθηκε και σε προηγούμενους ορισμούς, αλλά δεν ορίστηκε με ακρίβεια. Επιπλέον, ο ορισμός αυτός εισάγει δύο τεχνολογίες: την μηχανική μάθηση (machine learning) και την τεχνητή νοημοσύνη που είχαν αγνοηθεί από προηγούμενους ορισμούς. Αυτό, ως εκ τούτου, εισάγει την ιδέα ότι υπάρχουν μια σειρά από σχετιζόμενες τεχνολογίες που είναι ζωτικής σημασίας συστατικά του τελικού ορισμού.

Η Google Trends αναφέρει τους ακόλουθους όρους σε σχέση με τα Big Data: Data Analytics, Hadoop, NoSQL, Google, IBM, και Oracle. Από αυτούς τους όρους μια σειρά από τάσεις είναι εμφανείς.

- Πρώτον, ότι τα Big Data είναι άρρηκτα συνδεδεμένα με την ανάλυση δεδομένων και την εξαγωγή γνώσης απ’ αυτά.
- Δεύτερον, είναι σαφές ότι υπάρχουν μια σειρά από σχετιζόμενες τεχνολογίες όπως φαίνεται και από τον ορισμό της Microsoft, δηλαδή τις NoSQL και Apache Hadoop.
- Τέλος, είναι προφανές ότι υπάρχει ένας αριθμός οργανισμών, κυρίως βιομηχανικών οργανισμών που σχετίζονται με Big Data.

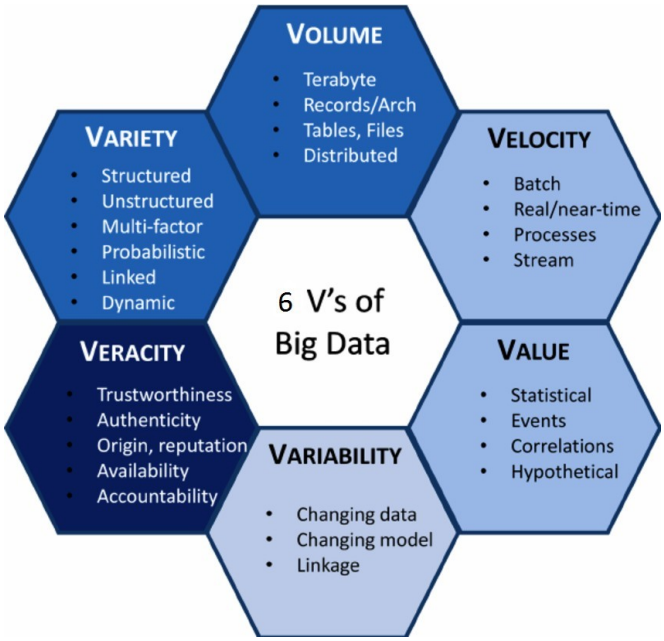
Όπως επισημαίνεται από την Google Trends, υπάρχει μια σειρά από τεχνολογίες που συχνά αναφέρεται ότι εμπλέκονται με τα Big Data. Αποθήκες δεδομένων NoSQL όπως Amazon, Dynamo, Cassandra, CouchDB, Mongo DB κ.ά. παίζουν κρίσιμο ρόλο στην αποθήκευση μεγάλου όγκου μη δομημένων και ιδιαίτερα μεταβαλλόμενων δεδομένων (για τη χρήση των χώρων αποθήκευσης δεδομένων NoSQL υπάρχει μια σειρά εργαλείων ανάλυσης και μεθόδων, συμπεριλαμβανομένων των Map Reduce, NLP, στατιστικού προγραμματισμού, machine learning και visualization).

Τέλος, το 2014 η Wikipedia περιέγραφε τα Big Data ως έναν ευρύτερο όρο για οιοδήποτε συλλογή συνόλων δεδομένων τόσο μεγάλων και σύνθετων που είναι δύσκολο να επεξεργαστούν χρησιμοποιώντας χειροκίνητα εργαλεία ή παραδοσιακές εφαρμογές επεξεργασίας δεδομένων.

1.3.3 Το Μοντέλο των 6 Vs

Αν στους όρους Volume, Velocity, Variety και Veracity που αναφέρουμε προηγουμένως, προσθέσουμε τη Value (Αξία) και τη Variability (Μεταβλητότητα) θα πάρουμε το Μοντέλο των 6 Vs. Καθώς η πρόσβαση σε μεγάλα δεδομένα είναι άχρηστη αν δε μπορούμε να τη μετατρέψουμε σε αξία, πολλοί υποστηρίζουν ότι η Value είναι το πιο σημαντικό V των Big Data. Οι οργανισμοί καλούνται να επιλέξουν την πιο αποτελεσματική από πλευράς κόστους λύση με στόχο την αξιοποίηση της πληροφορίας που θα οδηγεί στην έγκαιρη και

όσο το δυνατό πιο σωστή λήψη αποφάσεων, δίνοντας τη μεγαλύτερη δυνατή αξία στην επιχείρηση.



Εικόνα 4: Τα έξι Vs

1.3.4 Το Μοντέλο των 8 Vs

Εικόνα 5: Τα οχτώ Vs

Το πλέον πολύπλοκο μοντέλο είναι αυτό των 8 Vs που περιέχει τους εξής όρους:

- Volume
- Velocity
- Variety
- Veracity
- Value
- Variability
- Viscosity
- Virality

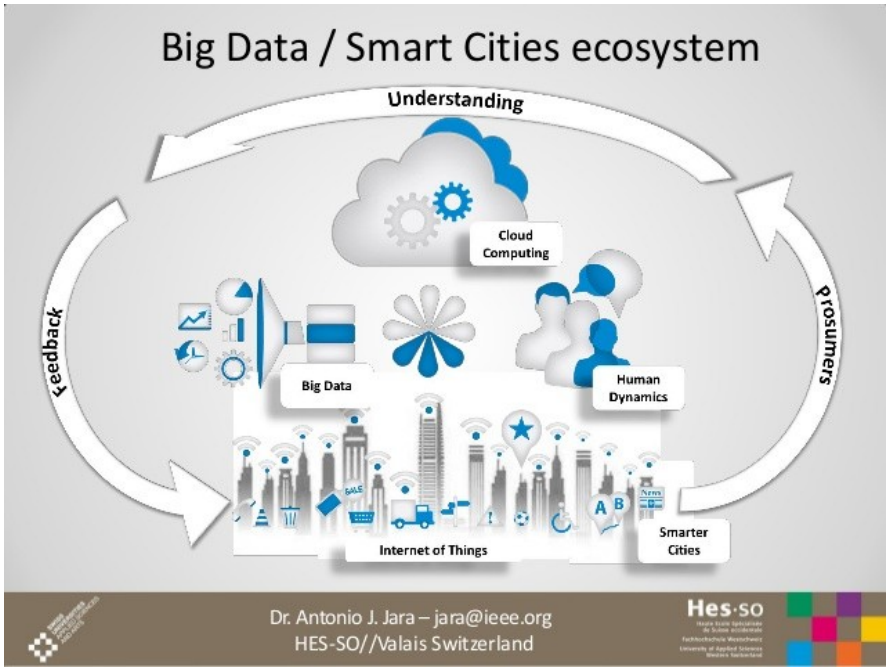
1.3.5 Εφαρμογές

Εικόνα 6: Εφαρμογές των Big Data

Σ΄ αυτό το εδάφιο παραθέτουμε κάποιες απ' τις εφαρμογές των Big Data Analytics στους διάφορους τομείς της ανθρώπινης δραστηριότητας. Φυσικά, τα παραδείγματα είναι πάρα πολλά και προφανώς δεν περιορίζονται σ' αυτά που θα αναφέρουμε στα πλαίσια τούτης της διπλωματικής.

Α) Εφαρμογές των Big Data στην Έξυπνη Πόλη

Η εφαρμογή τεχνολογιών μεγάλων δεδομένων για την έξυπνη πόλη επιτρέπει την αποτελεσματική αποθήκευση και επεξεργασία δεδομένων για την παραγωγή πληροφοριών που μπορούν να βελτιώσουν τις υπηρεσίες έξυπνων πόλεων. Επιπλέον, τα μεγάλα δεδομένα βοηθούν τους υπεύθυνους λήψης αποφάσεων να σχεδιάσουν οποιαδήποτε επέκταση σε υπηρεσίες και πόρους έξυπνων πόλεων. Ως προς την επίτευξη των στόχων και την προώθηση υπηρεσιών σε έξυπνες πόλεις, χρειάζονται τα σωστά εργαλεία και μέθοδοι για αποτελεσματική και προσοδοφόρα ανάλυση δεδομένων. Αυτά τα εργαλεία και οι μέθοδοι μπορούν να ενθαρρύνουν τη συνεργασία και την επικοινωνία μεταξύ φορέων, να παρέχουν υπηρεσίες σε πολλούς τομείς στην έξυπνη πόλη, καθώς και να βελτιώσουν τις εμπειρίες των πελατών και τις επιχειρηματικές ευκαιρίες.



Εικόνα 7: Το Οικοσύστημα “Big Data/Έξυπνη Πόλη”

Έξυπνο Ηλεκτρικό Δίκτυο

Η ταχεία κατανομή των έξυπνων ηλεκτρικών δικτύων επέτρεψε στους ερευνητές να ενσωματώσουν, να αναλύσουν και να χρησιμοποιήσουν δεδομένα παραγωγής και κατανάλωσης ενέργειας σε πραγματικό χρόνο, καθώς και άλλα είδη περιβαλλοντικών δεδομένων. Η βελτίωση της ενεργειακής απόδοσης και των ευφυών υπηρεσιών αναμένεται να οδηγήσει σε υψηλή επενδυτική αποδοτικότητα της υπάρχουσας υποδομής έξυπνου δικτύου. Σε ένα περιβάλλον έξυπνου δικτύου, δημιουργούνται μεγάλα ποσά δεδομένων από διαφορετικές πηγές, όπως οι συνήθειες χρήσης των ηλεκτρικών ρευμάτων των χρηστών, τα δεδομένα μετρήσεων φάσης και τα δεδομένα κατανάλωσης ενέργειας μεταξύ άλλων. Η αποτελεσματική χρήση των μεγάλων δεδομένων που συλλέγονται από το περιβάλλον έξυπνου δικτύου δύναται να βοηθήσει τους υπεύθυνους για τη λήψη αποφάσεων να καταλήξουν σε ορθές αποφάσεις όσον αφορά το επίπεδο παροχής ηλεκτρικής ενέργειας έτσι ώστε να ανταποκρίνεται ταυτόχρονα και στις απαιτήσεις του χρήστη. Τα αναλυτικά στοιχεία των έξυπνων δικτύων μπορούν επίσης να βοηθήσουν να προβλεφθεί η ανάγκη παροχής ηλεκτρικού ρεύματος στο μέλλον. Επιπλέον, η ανάλυση των δεδομένων έξυπνων δικτύων μπορεί να συμβάλει στην επίτευξη των στρατηγικών στόχων μέσω ειδικών σχεδίων τιμολόγησης που συμφωνούν με τα μοντέλα εφοδιασμού, ζήτησης και παραγωγής.

Υγεία

Στις μέρες μας, τα δεδομένα που αφορούν ασθενείς παράγονται με εκθετικό ρυθμό. Οι πληροφορίες αυτές ωστόσο έχουν διαφορετικές μορφές και πρότυπα. Τα Big Data Analytics μας επιτρέπουν να αποκωδικοποιήσουμε ολόκληρα τμήματα DNA μέσα σε λίγα λεπτά γεγονός που διευκολύνει στην εύρεση νέων θεραπειών και στη καλύτερη κατανόηση και πρόβλεψη διαφόρων ασθενειών. Ο ταχύς ρυθμός αύξησης του παγκόσμιου πληθυσμού διευκόλυνε τις γρήγορες αλλαγές στα μοντέλα θεραπείας και πολλές αποφάσεις πίσω από αυτές τις αλλαγές οφείλονται σε δεδομένα. Τα κατάλληλα εργαλεία

ανάλυσης μπορούν να επιτρέψουν στους ειδικούς της ιατρικής περίθαλψης να συλλέγουν και να αναλύουν τα δεδομένα των ασθενών, τα οποία μπορούν επίσης να χρησιμοποιηθούν από ασφαλιστικούς οργανισμούς και οργανισμούς διαχείρισης. Επιπλέον, η σωστή ανάλυση των μεγάλων δεδομένων της ιατρικής περίθαλψης μπορεί να βοηθήσει στην πρόβλεψη των επιδημιών, των θεραπειών και των ασθενειών, καθώς και στη βελτίωση της ποιότητας ζωής και την αποφυγή του αποτρέψιμου θανάτου. Το άθροισμα και η σταθερή φύση των πληροφοριών που συγκεντρώνονται για τα θέματα υγείας συγκεκριμένων ασθενών μπορούν να αυξηθούν μέσω έξυπνων συσκευών, οι οποίες συνδέονται με το σπίτι ή τις κλινικές για να παρακολουθούν συμπεριφορές ώστε να βοηθήσουν στην κατανόηση των αρχείων των ασθενών. Επιπλέον, οι αναλύσεις μεγάλων ποσοτήτων ιατρικών δεδομένων μπορούν να επιτρέψουν στους ιατρούς να ανιχνεύσουν τα προειδοποιητικά σημάδια μιας σοβαρής ασθένειας κατά τη διάρκεια του πρώιμου σταδίου της θεραπείας, γεγονός που μπορεί να σώσει εκατοντάδες ζωές.
Να σημειωθεί ότι οι τεχνικές Μεγάλων Δεδομένων χρησιμοποιούνται ήδη για την παρακολούθηση νεογνών που είτε γεννήθηκαν πρόωρα είτε είναι άρρωστα. Ειδικότερα, με την καταγραφή και ανάλυση κάθε καρδιακού παλμού και της αναπνοής κάθε μωρού, η μονάδα είναι σε θέση να αναπτύσσει αλγορίθμους που μπορούν να προβλέψουν λοιμώξεις 24 ώρες πριν εμφανιστούν οποιαδήποτε σωματικά συμπτώματα. Με αυτόν τον τρόπο, η ομάδα μπορεί να παρέμβει έγκαιρα και να σώσει τις ζωές των βρεφών.

Μεταφορές

Τα μοτίβα που λαμβάνονται από τα μεγάλα ποσά δεδομένων κίνησης μπορούν να συμβάλουν στη βελτίωση των συστημάτων μεταφοράς από την άποψη της ελαχιστοποίησης της κυκλοφοριακής συμφόρησης παρέχοντας εναλλακτικές διαδρομές και περιορίζοντας τον αριθμό των ατυχημάτων. Τα δεδομένα που παράγονται από τα συστήματα μεταφορών μπορούν επίσης να συμβάλουν στη βελτιστοποίηση των μεταφορών εμπορευμάτων.
Η εταιρεία διεθνών ταχυμεταφορών UPS, άρχισε να καταγράφει και να μελετά τις κινήσεις πακέτων και συναλλαγών από τις αρχές της δεκαετίας του 1980. Σήμερα συγκεντρώνει δεδομένα για 16,3 εκατομμύρια πακέτα ημερησίως, για 8,8 εκατομμύρια πελάτες, με μέσω όρο 39,6 εκατομμύρια αιτήματα παρακολούθησης πακέτου καθημερινά. Το μεγαλύτερο τμήμα των μεγάλων δεδομένων που διαθέτει προέρχεται από τηλεματικούς αισθητήρες τοποθετημένους στα οχήματά της. Τα δεδομένα αξιοποιούνται τόσο για την καθημερινή εποπτεία και μέτρηση της αποδοτικότητας αλλά και για τη διαμόρφωση της βέλτιστης δομής των δρομολογίων. Συνέπεια, της εν λόγω πρακτικής ήταν η εξοικονόμηση πάνω από 8.4 εκατομμύρια γαλονιών καυσίμων το 2011, και η μείωση κατά 85 εκατομμύρια μίλια των καθημερινών δρομολογίων. Η UPS εκτιμά ότι η μείωση ενός μιλίου ανά οδηγό την ημέρα, εξοικονομεί κόστος 30.000.000 δολαρίων το χρόνο.

Έξυπνη Διακυβέρνηση

Τα Big Data Analytics μπορούν να διαδραματίσουν σημαντικό ρόλο στη διευκόλυνση της έξυπνης διακυβέρνησης. Οι οργανισμοί με κοινά ενδιαφέροντα μπορούν εύκολα να εντοπιστούν μέσω ανάλυσης δεδομένων γεγονός που ενδέχεται να οδηγήσει στη μεταξύ τους συνεργασία. Η συνεργασία αυτή δύναται με τη σειρά της να συμβάλει στην ανάπτυξη των χωρών που σχετίζονται με τους εν λόγω οργανισμούς. Επιπλέον, τα Big data Analytics μπορούν να βοηθήσουν τις κυβερνήσεις να καθιερώσουν και να εφαρμόσουν ικανοποιητικές πολιτικές καθώς γνωρίζουν ήδη τις ανάγκες των ανθρώπων όσον αφορά την υγεία, την κοινωνική μέριμνα, την εκπαίδευση και ούτω καθεξής. Επιπλέον, ο λόγος της ανεργίας μπορεί να μειωθεί με την ανάλυση των μεγάλων δεδομένων των διαφόρων εκπαιδευτικών ιδρυμάτων.

Παραδείγματα Έξυπνων Πόλεων που εφαρμόζουν Big Data Analytics

- Η Στοκχόλμη εφήρμοσε πρόσφατα έξυπνη διαχείριση και έξυπνες εφαρμογές για την αντιμετώπιση των προβλημάτων της κυκλοφορίας και του περιβάλλοντος. Πραγματοποιήθηκαν μισό εκατομμύριο καταχωρήσεις κλάδων αποβλήτων, βαρών και τοποθεσιών.
- Το έργο Infoshare της Περιφέρειας του Ελσίνκι είναι μία από τις πρωτοποριακές ανοιχτές πλατφόρμες αστικών δεδομένων. Το 2013, διατέθηκαν περισσότερες από 1030 βάσεις δεδομένων που καλύπτουν ένα ευρύ φάσμα αστικών φαινομένων, όπως οι μεταφορές, η οικονομία, οι συνθήκες διαβίωσης, η απασχόληση και η ευημερία. Η πλατφόρμα κέρδισε πρόσφατα το Ευρωπαϊκό Βραβείο Καινοτομίας στη Δημόσια Διοίκηση στην κατηγορία της ενδυνάμωσης των πολιτών. Η έκθεση της κριτικής επιτροπής υποδεικνύει ότι το άνοιγμα των πληροφοριών λήψης αποφάσεων μέσω ηλεκτρονικού συστήματος διαχείρισης περιπτώσεων παρέχει στους πολίτες μια μεγάλη ευκαιρία να συμμετάσχουν σημαντικά στη λήψη αποφάσεων.
- Η Κοπεγχάγη κατατάσσεται στην όγδοη θέση από τον Boyd Cohen σε έναν κατάλογο έξυπνων ευρωπαϊκών πόλεων. Η Κοπεγχάγη έχει ως στόχο να καταστεί η πρώτη ουδέτερη ως προς τον άνθρακα πρωτεύουσα μέχρι το 2025. Ως εκ τούτου, η πόλη εφαρμόζει επί του παρόντος μια σειρά από νέες και καινοτόμες λύσεις στους τομείς των μεταφορών, των αποβλήτων, του νερού, της θέρμανσης και των εναλλακτικών πηγών ενέργειας με σκοπό να εξυπηρετήσει αυτό το στόχο και να βελτιώσει τη βιωσιμότητά του μέσω ενός μεγάλου αριθμού πρωτοβουλιών. Η Κοπεγχάγη διαθέτει ένα εκτεταμένο δίκτυο ποδηλατοδρόμων, το οποίο ακόμη διευρύνεται. Η λύση για το ποδήλατο είναι ενσωματωμένη σε μια ευρεία έννοια βελτίωσης της κυκλοφορίας στην πόλη, όπως για παράδειγμα, εύκολη μετάβαση από ποδήλατα σε δημόσιες συγκοινωνίες και παροχή επαρκών χώρων στάθμευσης γι αυτά.

Β) Άλλες εφαρμογές

Τηλεπικοινωνίες
Στα τηλεφωνικά κέντρα συλλέγονται πολύ μεγάλες ποσότητες αδόμητων και δομημένων δεδομένων. Η χαρτογράφηση και ταξινόμηση των κλήσεων παρέχει τη δυνατότητα εντοπισμού σφαλμάτων και αδυναμιών στις σχετικές υποδομές.

Εμπόριο
Η eBay, μία από τις μεγαλύτερες ηλεκτρονικές πλατφόρμες δημοπρασιών στον κόσμο, καταγράφει συναλλαγές με περισσότερους από 108 εκατομμύρια πελάτες ετησίως ενώ εισπράττει πάνω από 250 εκατομμύρια αιτήματα στον ιστόχωρό της ημερησίως. Επίσης, στο εμπόριο ηλεκτρονικών συσκευών, εκτιμάται ότι πωλείται ένα κινητό τηλέφωνο κάθε 5 δευτερόλεπτα. Είναι εμφανής, λοιπόν, ο όγκος πληροφορίας που αποθηκεύεται τόσο για τους πελάτες όσο και για τα προϊόντα καθώς και η δυνατότητα εξόρυξης γνώσης αναφορικά με τις συνήθειες και τις τάσεις, που μπορεί να συνδράμει στη διαμόρφωση σχετικών πολιτικών. Ακόμη και στο λιανεμπόριο γίνεται αξιοποίηση μεγάλων δεδομένων μελετώντας πληροφορίες από ΜΜΕ για τον εντοπισμό του βέλτιστου σημείου εγκατάστασης ενός καταστήματος.

Αντιμετώπιση καταστροφών
Μέσα από τη συγκέντρωση πληροφορίας που μπορεί να προέρχεται είτε από ΜΜΕ είτε από απλούς πολίτες, δίδεται η δυνατότητα χαρτογράφησης του τόπου όπου λαμβάνει χώρα μια καταστροφή, αξιολόγησης της σοβαρότητάς της και χάραξη της άριστης διαδρομής για την τάχιστη άφιξη των σχετικών υπηρεσιών. Η γρήγορη πρόγνωση του τυφώνα Irene στην Φλόριντα το 2011, η οποία βασίστηκε στην ανάλυση μεγάλων γεωχωρικών δεδομένων, ελαχιστοποίησε τις συνέπειές του αφού έδωσε το χρόνο για τη λήψη όλων των αναγκαίων μέτρων.

Φυσικοί πόροι
Ο κλάδος του πετρελαίου θεωρείται από τους πρώτους που άρχισαν να ασχολούνται με τα μεγάλα δεδομένα. Πετρελαϊκές εταιρείες και κυβερνήσεις κάνουν χρήση και ανάλυση τεράστιων ποσοτήτων δεδομένων σχετικά με τη σεισμική δραστηριότητα σε όλη την υφήλιο με σκοπό την εξερεύνηση και εξόρυξη πετρελαίου.

Διαδίκτυο
Ιστότοποι όπως το facebook και το twitter συγκεντρώνουν πάνω από 25 και 12 terabytes δεδομένων αντίστοιχα. Η Google μέσω των διάφορων εφαρμογών της (mail, google drive, google earth κ.λ.π.) συγκεντρώνει δεδομένα όγκου άνω των 80 terabytes ημερησίως. Η ανάλυση των δεδομένων των χρηστών τους είναι ο οδηγός της διαμόρφωσης της στρατηγικής τους στόχευσης.

Κατανόηση και στόχευση πελατών
Αποτελεί μια από τις μεγαλύτερες και δημοφιλέστερες περιοχές χρήσης των Μεγάλων Δεδομένων σήμερα. Οι εταιρείες είναι πρόθυμες να εμπλουτίσουν τα παραδοσιακά σύνολα δεδομένων τους, με τα δεδομένα των κοινωνικών μέσων και περιήγησης των μηχανών αναζήτησης, ώστε να αποκτήσουν μια πιο ολοκληρωμένη εικόνα των πελατών τους. Απώτερος και βασικός στόχος τους, είναι να δημιουργήσουν μοντέλα πρόβλεψης. Μπορούμε να θυμηθούμε το παράδειγμα του λιανοπωλητή Target στις ΗΠΑ. Τα Target είναι τώρα σε θέση να προβλέψουν με μεγάλη ακρίβεια, πότε κάποιος από τους πελάτες τους περιμένει μωρό και να του παρουσιάσουν αντίστοιχες προσφορές.
Χρησιμοποιώντας Μεγάλα Δεδομένα, εταιρίες τηλεπικοινωνιών μπορούν πλέον να προβλέψουν καλύτερα τυχόν απώλειες πελατών και να προβούν σε πρόταση δελεαστικών πακέτων πριν οι πελάτες εκδηλώσουν δυσαρέσκεια και επιθυμία αποχώρησης από τον πάροχό τους. Με τον ίδιο τρόπο, οι ασφαλιστικές εταιρείες αυτοκινήτων είναι σε θέση να κατανοήσουν το πόσο καλά οδηγούν οι πελάτες τους, και να προτείνουν αντίστοιχα συμβόλαια. Ακόμη και σε προεκλογικές εκστρατείες η εκάστοτε παράταξη μπορεί να επωφεληθεί από την ανάλυση συνόλων Big Data και να εντατικοποιήσει τις εκστρατείες της σε περιοχές που τα Μεγάλα Δεδομένα “αποκαλύπτουν” χαμηλά ποσοστά.

Έξυπνη Επιχείρηση
Τα Μεγάλα Δεδομένα χρησιμοποιούνται επίσης για τη βελτιστοποίηση των επιχειρηματικών διαδικασιών. Οι έμποροι λιανικής πώλησης είναι σε θέση να βελτιστοποιούν τα αποθέματά τους με βάση τις προβλέψεις που δημιουργούνται από τα δεδομένα των μέσων κοινωνικής δικτύωσης, τις τάσεις αναζήτησης Ιστού και τις προβλέψεις καιρού. Μια συγκεκριμένη επιχειρηματική διεργασία που χρησιμοποιεί αρκετά τα Μεγάλα Δεδομένα και την ανάλυση αυτών, είναι η αλυσίδα εφοδιασμού ή η βελτιστοποίηση της διαδρομής παράδοσης. Εδώ, οι γεωγραφικοί αισθητήρες αναγνώρισης και εντοπισμού των ραδιοσυχνοτήτων χρησιμοποιούνται για την

παρακολούθηση των εμπορευμάτων ή των οχημάτων διανομής ώστε να βελτιστοποιήσουν τα δρομολόγια τους με την ενσωμάτωση ζωντανών δεδομένων οδικής κυκλοφορίας.

Βελτίωση αθλητικής επίδοσης

Τα πιο δημοφιλή αθλήματα σήμερα, στηρίζονται στα Μεγάλα Δεδομένα. Χαρακτηριστικό παράδειγμα εφαρμογής είναι το εργαλείο IBM Slam Tracker για τουρνουά τένις. Αυτό που κάνει ουσιαστικά το IBM Slam Tracker είναι ότι χρησιμοποιεί αναλύσεις βίντεο για να παρακολουθείται η απόδοση του κάθε παίκτη σε ποδοσφαιρικό αγώνα. Διαθέτει επίσης και τεχνολογία αισθητήρων σε αθλητικό εξοπλισμό όπως μπάλες μπάσκετ. Επιπλέον, πολλές μεγάλες αθλητικές ομάδες παρακολουθούν τους αθλητές έξω από το αθλητικό περιβάλλον, χρησιμοποιώντας έξυπνη τεχνολογία για να παρατηρούν τη διατροφή και τον ύπνο, καθώς και τη συναισθηματική ευεξία τους με βάση στοιχεία από διάφορες συνομιλίες στα κοινωνικά μέσα.

Επιστήμες και Έρευνα

Τα πεδία της Επιστήμης και της Έρευνας μετασχηματίζονται από τις νέες δυνατότητες που παρέχουν τα Μεγάλα Δεδομένα. Ας πάρουμε ως παράδειγμα το CERN, το ευρωπαϊκό κέντρο ερευνών πυρηνικής φυσικής με το μεγάλο Επιταχυντή Αδρονίων (ο μεγαλύτερος και ισχυρότερος επιταχυντής σωματιδίων στον κόσμο). Κατά τα πειράματα που εκτελεί παράγει τεράστιες ποσότητες δεδομένων. Το κέντρο δεδομένων του CERN έχει 65.000 επεξεργαστές για να αναλύσει 30 petabytes δεδομένων. Ωστόσο, χρησιμοποιεί την υπολογιστική δύναμη χιλιάδων υπολογιστών, οι οποίοι διανέμονται σε 150 κέντρα δεδομένων σε όλο τον κόσμο, για να αναλύσει τα δεδομένα του.

Βελτιστοποίηση Μηχανών

Τα Μεγάλα Δεδομένα βοηθούν τις μηχανές και συσκευές να γίνονται πιο έξυπνες και αυτόνομες. Για παράδειγμα, το Toyota Prius είναι εξοπλισμένο με κάμερες, GPS καθώς και ισχυρούς υπολογιστές και αισθητήρες για την ασφαλή οδήγηση στο δρόμο χωρίς την παρέμβαση του ανθρώπου. Μπορούμε να χρησιμοποιήσουμε ακόμα και εργαλεία Μεγάλων Δεδομένων για τη βελτιστοποίηση της απόδοσης των υπολογιστών και των αποθηκών δεδομένων.

Προσωπική Χρήση

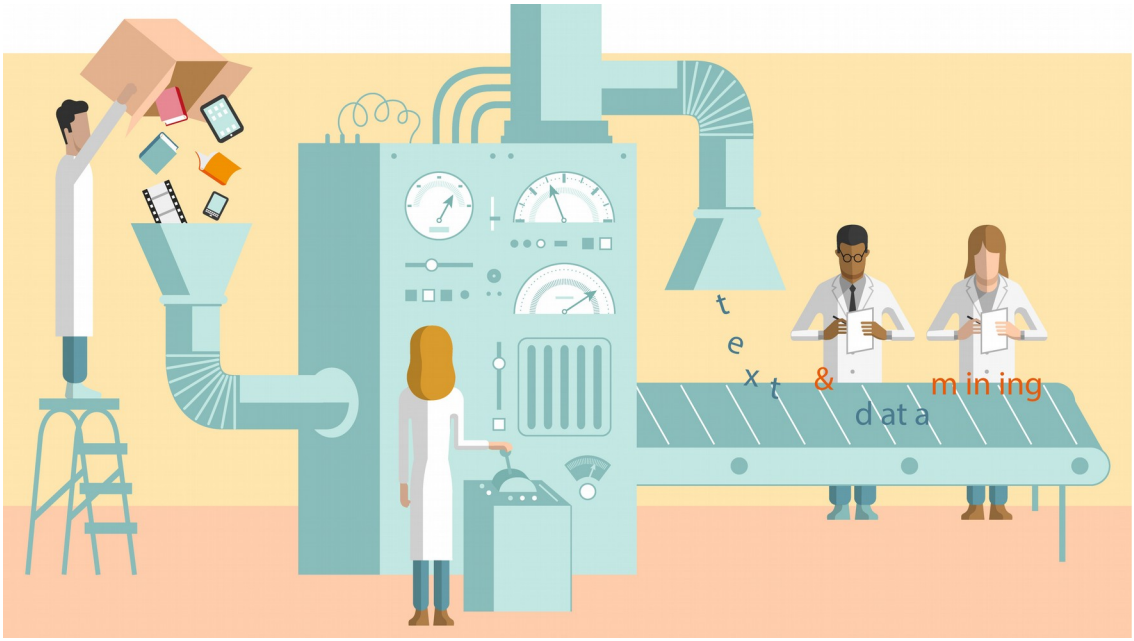
Τα Μεγάλα Δεδομένα δεν αφορούν μόνο τις επιχειρήσεις και τις κυβερνήσεις, αλλά και όλους εμάς ξεχωριστά. Μπορούμε πλέον να επωφεληθούμε από τα δεδομένα που δημιουργούνται από φορητές συσκευές όπως τα smart watches. Ας πάρουμε ως παράδειγμα το Up band from Jawbone, όπου το περιβραχιόνιο συλλέγει δεδομένα σχετικά με την κατανάλωση θερμίδων, τα επίπεδα δραστηριότητας και τα μοτίβο του ύπνου μας. Ενώ δίνει πλούσιες πληροφορίες σχετικά με τα άτομα, η πραγματική του αξία έγκειται στην ανάλυση των δεδομένων. Στη περίπτωση του Jawbone, η εταιρεία συγκεντρώνει εδώ και 60 χρόνια δεδομένα ύπνου κάθε βράδυ.

Οικονομικές Συναλλαγές

Η τελευταία κατηγορία εφαρμογής Μεγάλων Δεδομένων είναι αυτή των χρηματοοικονομικών συναλλαγών. Οι συναλλαγές υψηλής συχνότητας (High Frequency Trading - HFT) είναι μια περιοχή όπου τα Μεγάλα Δεδομένα βρίσκουν μεγάλη χρήση σήμερα. Αλγόριθμοι Μεγάλων Δεδομένων χρησιμοποιήθηκαν για τη λήψη επενδυτικών αποφάσεων. Σήμερα, η πλειοψηφία των μετοχών που ανταλλάσσονται πραγματοποιείται μέσω αλγορίθμων Μεγάλων Δεδομένων. Αυτοί λαμβάνουν υπόψη τους τα σήματα από τα

δίκτυα των social media και τις ειδησεογραφικές ιστοσελίδες προκειμένου να αγοράζουν και να πωλούν τις μετοχές σε κλάσματα του δευτερολέπτου.

1.4 Data Mining



Εικόνα 8: Data Mining

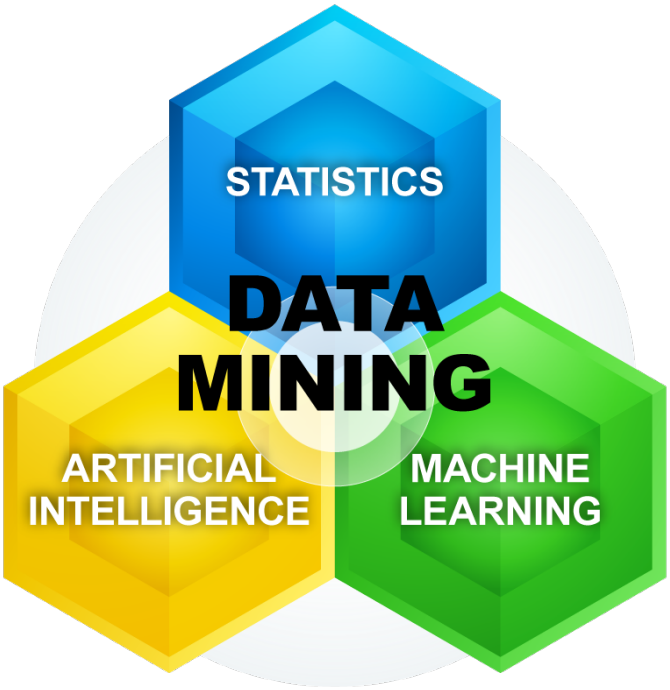
1.4.1 Εισαγωγή

Ζούμε σε μια εποχή όπου καθημερινά συλλέγονται τεράστιες ποσότητες δεδομένων που οργανώνονται σε υπολογιστικά συστήματα και αποθηκεύονται σε βάσεις δεδομένων. Η ανάλυση αυτών των δεδομένων αποτελεί μια ιδιαίτερα σημαντική ανάγκη, καθώς οδηγεί στην εξαγωγή χρήσιμης πληροφορίας. Οι τεχνικές εξόρυξης δεδομένων (data mining) ικανοποιούν την ανάγκη αυτή.

1.4.2 Ορισμός

Εξόρυξη δεδομένων ονομάζεται η διαδικασία εύρεσης προτύπων από ένα μεγάλο όγκο δεδομένων με σκοπό να εξαχθεί ενδιαφέρουσα, μη προφανής και χρήσιμη πληροφορία. Βασικός στόχος της διαδικασίας αυτής είναι να αποδοθεί κατανοητή πληροφορία που θα βοηθήσει στην εξαγωγή σημαντικών συμπερασμάτων και στην λήψη αποφάσεων. Σε

γενικές γραμμές, η εξόρυξη δεδομένων μετατρέπει τη συλλογή των δεδομένων σε γνώση. Εφαρμόζεται σε ένα μεγάλο εύρος επιχειρήσεων και οργανισμών που καλύπτουν τομείς όπως η ιατρική, η οικονομία, η πολιτική, το marketing κ.ά.. Είναι ευρέως γνωστή μέθοδος για την έρευνα συμπεριφοράς του καταναλωτικού κοινού και για την ανακάλυψη πεποιθήσεων από μελέτη δεδομένων ερωτηματολογίων.



Εικόνα 9: Data Mining και άλλοι Επιστημονικοί Τομείς

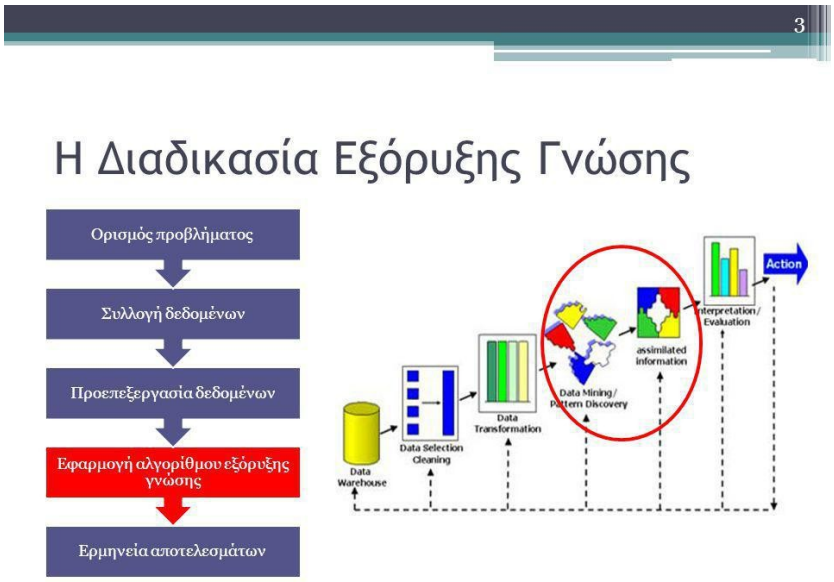
1.4.3 Data Mining & Knowledge Discovery in Data Bases

Στη διεθνή βιβλιογραφία υπάρχει μια γενικότερη σύγχυση ανάμεσα στους όρους «Εξόρυξη Γνώσης» (Data mining) και «Ανεύρεση γνώσης στις βάσεις δεδομένων» (Knowledge Discovery in Data Bases, KDD). Σε πολλές περιπτώσεις αξίζει να σημειωθεί ότι οι δύο αυτοί όροι ταυτίζονται, ενώ στην πραγματικότητα η εξόρυξη δεδομένων αποτελεί τμήμα της ανεύρεσης γνώσης, συγκροτώντας το πυρήνα αυτής.

Η εξόρυξη γνώσης διαθέτει ένα ευρύ πεδίο υπολογιστικών μεθόδων που μεταξύ άλλων περιλαμβάνει τη Στατιστική Ανάλυση (Statistical Analysis), τα Δένδρα Αποφάσεων (Decision Trees), τα Νευρωνικά Δίκτυα (Neural Networks), την Εξαγωγή Κανόνων (Rule Induction) και τη Γραφική Οπτικοποίηση (Graphic Visualization). Τέτοιες μέθοδοι χρησιμοποιούνται για την εύρεση συσχετίσεων, προτύπων και δομών σε μεγάλες και διαρκώς αυξανόμενες βάσεις δεδομένων. Ειδικά η εύρεση εργαλείων είναι ένα ιδιαίτερα σημαντικό εξαγόμενο της εξόρυξης δεδομένων μέσω σχέσεων μεταξύ των χαρακτηριστικών των βάσεων δεδομένων.

Η εξόρυξη γνώσης βοηθά τις σύγχρονες εταιρείες να εστιάζουν στα πιο σημαντικά στοιχεία από τις αποθήκες δεδομένων τους και να προβλέπουν μελλοντικές τάσεις και συνήθειες ώστε να μπορούν να παίρνουν σωστές αποφάσεις.

1.4.4 Διαδικασία Εξόρυξης Γνώσης



Εικόνα 10: Διαδικασία Εξόρυξης Δεδομένων

Η εξόρυξη δεδομένων περιλαμβάνει ολόκληρη τη διαδικασία από την συλλογή των δεδομένων μέχρι την προβολή και την εφαρμογή πρότυπων ευρημάτων σε νέες, άγνωστες δομές δεδομένων. Η διαδικασία περιλαμβάνει:

- τεχνικές για την προεπεξεργασία δεδομένων
- το πραγματικό σύστημα εξόρυξης δεδομένων (σύστημα DM)
- την οπτικοποίηση και επικύρωση των δεδομένων και
- την ερμηνεία και την αξιολόγηση των δεδομένων που οδηγούν στη γνώση.

Τα βήματα του data mining είναι συνοπτικά:

- Η Ενσωμάτωση των Δεδομένων (Data Integration): αρχικά συλλέγονται τα δεδομένα από διάφορες πηγές (Excel, MS Access, Oracle, SQL Server, csv, rdf stores, XML κ.λ.π.) σε μια μοναδική πηγή δεδομένων που ονομάζεται Target Data/ Database, με τη χρήση κάποιας τεχνολογίας όπως SPARQL, SQL, PYTHON κλπ.
- Η Επιλογή των Δεδομένων (Data Selection): σε αυτό το βήμα επικεντρωνόμαστε μόνο σε εκείνα τα δεδομένα που μας ενδιαφέρουν.
- Ο Καθαρισμός και η Κανονικοποίηση των Δεδομένων (Data Cleansing and Normalization): τα δεδομένα που εισάγονται από τις διάφορες πηγές μπορεί να έχουν διαφορετική μορφή από την Target Database. Για τον λόγο αυτό πρέπει να καθαρίσουμε τα δεδομένα χρησιμοποιώντας τον κατάλληλο αλγόριθμο.

- Ο Μετασχηματισμός των Δεδομένων (Data Transformation): εν συνεχεία, τα δεδομένα προετοιμάζονται και μετατρέπονται σε τυπική μορφή.
- Η Εξόρυξη των Δεδομένων (Data Mining): αποτελεί το σημαντικότερο βήμα της διαδικασίας. Σ' αυτό το στάδιο αναλύουμε και προσδιορίζουμε τον τύπο του αλγόριθμου εξόρυξης δεδομένων που είναι κατάλληλος για τα δεδομένα που συλλέχθηκαν και στη συνέχεια εφαρμόζουμε αλγόριθμους για το προσδιορισμό των κρυφών προτύπων. Παραδείγματος χάριν με την χρήση του k-means Clustering Algorithm κατηγοριοποιούνται τα δεδομένα σε ομάδες ανάλογα με τη συσχέτιση τους.
- Η Αξιολόγηση Μοτίβου (Pattern Evaluation): το πρότυπο που εντοπίσαμε από τα δεδομένα ερμηνεύεται και αξιολογείται στη συνέχεια για να αποκτηθούν γνώσεις από αυτό.
- Η Παρουσίαση Γνώσης (Knowledge Presentation): αποτελεί το στόχο της τεχνικής εξόρυξης δεδομένων, όπου η γνώση που συλλέγεται από την παραπάνω διαδικασία οπτικοποιείται, αξιολογείται και λαμβάνεται υπόψη για την λήψη κρίσιμων αποφάσεων.

Πολλές φορές κάποια από τα παραπάνω βήματα μπορούν να συνδυαστούν μεταξύ τους για το καλύτερο δυνατό αποτέλεσμα.

Από τα παραπάνω, λοιπόν, συμπεραίνουμε ότι η εξόρυξη δεδομένων αποτελεί το κλειδί για την ανεύρεση γνώσης. Παρόλα αυτά, δεν καταλαμβάνει παρά μόνο ένα μικρό μέρος της όλης προσπάθειας, δεδομένου της πολυπλοκότητας της. Σε αυτό το σημείο αξίζει να σημειωθεί ότι ο χρήστης, εκμεταλλευόμενος την επαναληπτική μορφή της διαδικασίας ανεύρεσης γνώσης, έχει την δυνατότητα να τροποποιήσει τα μέτρα αξιολόγησης, να τελειοποιήσει την διαδικασία της εξόρυξης, να επιλέξει νέα δεδομένα, να τροποποιήσει περαιτέρω τα ήδη υπάρχοντα ή να ενσωματώσει στη βάση νέα από καινούργιες πηγές, με τελικό στόχο την εξαγωγή διαφορετικών και ακόμη πιο κατάλληλων αποτελεσμάτων.

1.4.5 Τύποι Εξόρυξης Δεδομένων

Οι Τύποι της εξόρυξης δεδομένων κατηγοριοποιούνται ως εξής:

- Data Mining: περιλαμβάνει την ανάλυση αριθμητικών και απόλυτων δεδομένων που είναι αποθηκευμένα σε μεγάλα και πολύπλοκα σύνολα δεδομένων. Συχνά ο όρος αυτός χρησιμοποιείται για να περιγράψει πιο εξειδικευμένες τεχνικές, όπως η εξόρυξη κειμένου, ιστού ή χώρου.

- Text Mining: ο συγκεκριμένος τύπος εξόρυξης περιλαμβάνει αλγορίθμους για την ανάλυση λεξικών και γραμματικών πτυχών των κειμένων. Το πραγματικό κείμενο αναλύεται σε συγκεκριμένες δομές όπου τα πρότυπα και οι βασικές πληροφορίες του κειμένου καταγράφονται, ομαδοποιούνται και ταξινομούνται χρησιμοποιώντας τις μεθόδους εξόρυξης δεδομένων. Τα πρώτα εργαλεία εξόρυξης κειμένου μπορούσαν να καταγράφουν τα περιεχόμενα και τις δομές απλών εγγράφων κειμένου, όπως τα έγγραφα του Microsoft Word και του Acrobat PDF, ενώ πλέον

έχουν τη δυνατότητα να σαρώνουν και να αναλύουν το αδόμητο κείμενο σε s, memos, έρευνες, συνομιλίες, σημειώσεις, φόρουμ και παρουσιάσεις.

- Web Mining: συνιστά την εφαρμογή μεθόδων εξόρυξης δεδομένων σε πληροφορίες που συλλέγονται στο Διαδίκτυο. Στην εξόρυξη ιστού, γίνεται διάκριση μεταξύ:

    - εξόρυξης περιεχομένου ιστού, η οποία είναι η ανάλυση του περιεχομένου του ιστοτόπου,
    - εξόρυξη δομής διαδικτύου ή σχέσεων, δηλαδή ανάλυση εισερχόμενων και εξερχόμενων υπερσυνδέσμων ιστοτόπων ,
    - εξόρυξη χρήσης ιστού, η οποία καταγράφει και αναλύει την αλληλεπίδραση των χρηστών με τους ιστοτόπους μέσω της σάρωσης αρχείων καταγραφής.

- Image Mining: ο στόχος των τεχνικών εξόρυξης εικόνας είναι η ανάλυση και εξαγωγή χωρικών μοτίβων σε δεδομένα εικόνας τα οποία δεν αποθηκεύονται ρητά στις εικόνες. Η εξαγωγή των μοτίβων γίνεται με διάφορους τρόπους όπως π.χ με την αναγνώριση της ύπαρξης και κατανομής των χρωμάτων, της υφής, του σχήματος, των αποστάσεων και των εντάσεων στα δεδομένα εικόνας.

- Picture, Video Data & Music Mining: οι εν λόγω τεχνικές εξόρυξης χρησιμοποιούνται όλο και περισσότερο για να αναγνωρίζουν χαρακτηριστικά σε εικόνες, βίντεο και μουσικά δεδομένα. Αυτός ο τύπος εξόρυξης είναι ο μόνος που μπορεί να αντιμετωπίσει τις τεράστιες ποσότητες περίπλοκων δεδομένων που δημιουργούνται, για παράδειγμα, στο Google ή στο YouTube. Ιδιαίτερη σημασία έχει η γρήγορη ενεργοποίηση των περιεχομένων ανάκτησης εικόνων/ βίντεο, η ευρετηρίαση, ταξινόμηση και παρακολούθηση.

- Time Series Data Mining: σε αυτό το σύστημα εξόρυξης δεδομένων, οι χρονικές σχέσεις διερευνώνται μέσω μιας ειδικής λειτουργίας απόστασης όπως η δυναμική χρονική στρέβλωση. Ο στόχος είναι να αναγνωριστούν ομοιότητες κατά τη διάρκεια της χρονοσειράς, ακόμα και όταν τα παρόμοια χαρακτηριστικά μετατοπίζονται κατά την πορεία της διαδικασίας.

- Spatial Data Mining: επιδιώκει την ανακάλυψη μοτίβων σε μεγάλα, πολυδιάστατα σύνολα χωρικών δεδομένων, τα οποία δημιουργούνται με τεχνικές τηλεπισκόπησης κατά την παρατήρηση της Γης. Εφόσον στην ανάλυση μοτίβων ενσωματωθούν, πέραν των χωρικών δεδομένων, και πρόσθετες χρονολογικές σειρές, τότε χρησιμοποιείται ο όρος εξόρυξη χωρο-χρονικών δεδομένων.

1.4.6 Στόχοι Data Mining

Οι μέθοδοι εξόρυξης στοχεύουν στην ανακάλυψη στοιχειών που θα είναι χρήσιμα για τους οργανισμούς και τις επιχειρήσεις. Πληροφορίες για τυποποιημένες μορφές όπως για παράδειγμα, ότι υπάρχουν πελάτες που θα ψωνίσουν περισσότερο από δύο φορές σε περίοδο εκπτώσεων ή προσφορών, ή ότι είναι πιθανό να αγοράσουν τουλάχιστον μια φορά κατά την διάρκεια των εορταστικών ημερών, του Πάσχα και των Χριστουγέννων, είτε

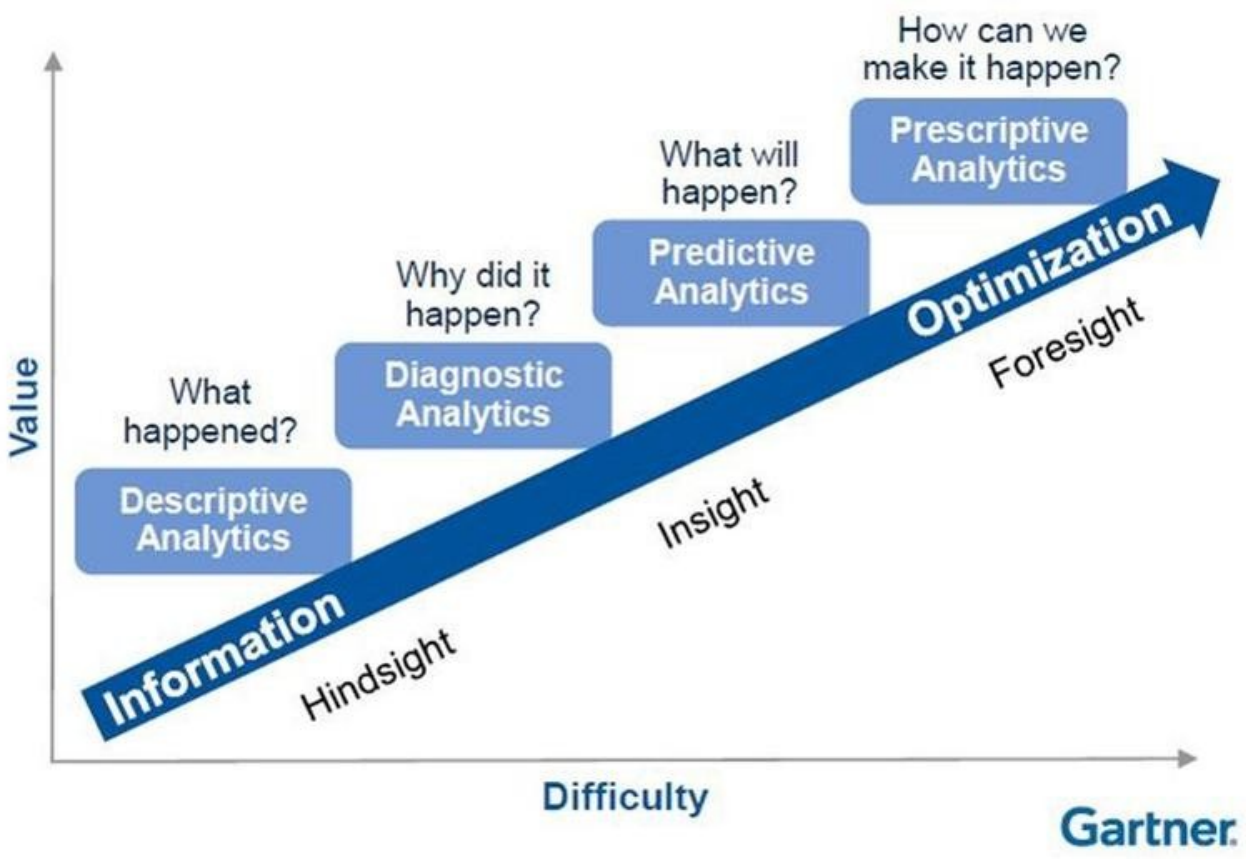
για συσχετίσεις όπως ότι εάν ένας πελάτης αγοράσει dvd player τότε πιθανότατα να αγοράσει και κάποια άλλη ηλεκτρονική συσκευή, μπορεί να αποτελέσουν καθοριστικούς παράγοντες για τη λήψη αποφάσεων όσον αφορά τη λειτουργία μιας εμπορικής επιχείρησης. Επί παραδείγματι δύναται να ληφθούν αποφάσεις σχετικά με το ωράριο, το ύψος και τη διάρκεια των εκπτώσεων, ακόμη και για την τοποθέτηση των προϊόντων μέσα στα καταστήματα. Παράλληλα, τέτοιου είδους πληροφορίες χρησιμοποιούνται για τον προγραμματισμό χρήσης πρόσθετων αποθηκευτικών χώρων ή και για τον σχεδιασμό διαφορετικών στρατηγικών μάρκετινγκ. Τα στελέχη της επιχείρησης, που είναι υπεύθυνα για την λήψη των αποφάσεων εκμεταλλεύονται τις δυνατότητες του Data Mining και μετατρέπουν τις γνώσεις σε επιτυχή αποτελέσματα. Παρακάτω περιγράφονται και αναλύονται οι στόχοι της εξόρυξης δεδομένων.

- Πρόβλεψη: Περιλαμβάνει την χρήση μερικών μεταβλητών ή χαρακτηριστικών μιας βάσης δεδομένων για την πρόβλεψη άγνωστων ή μελλοντικών τιμών χρήσιμων μεταβλητών. Με άλλα λόγια, οι διαδικασίες πρόβλεψης της εξόρυξης δεδομένων (Predictive Data Mining Tasks), προσπαθούν να κάνουν εκτιμήσεις βγάζοντας συμπεράσματα από τα διαθέσιμα δεδομένα. Η προσπάθεια πρόβλεψης μελλοντικών συμπεριφορών έχει ως στόχο να ληφθούν αποφάσεις που να μεγιστοποιούν το κέρδος και να προλαμβάνουν δυσάρεστες καταστάσεις. Τα αποτελέσματα της εξόρυξης μπορεί να είναι πληροφορίες σχετικές με το ύψος των πωλήσεων ενός καταστήματος για μια συγκεκριμένη χρονική περίοδο, αλλά και αν το κλείσιμο μιας γραμμής παραγωγής θα είχε θετική επίδραση στις πωλήσεις. Συγχρόνως σε επιστημονικό επίπεδο, η μελέτη παλαιότερων σεισμικών φαινόμενων ίσως να οδηγούσε στην πρόβλεψη σεισμικής δραστηριότητας.
- Αναγνώριση: Σε αυτή τη φάση οι τυποποιημένες μορφές των δεδομένων χρησιμοποιούνται για να δείξουν την ύπαρξη μιας δραστηριότητας ή ενός γεγονότος.
- Περιγραφή: Είναι η διαδικασία η οποία επικεντρώνεται στην ανακάλυψη προτύπων και αναπαριστά τα δεδομένα μιας πολύπλοκης βάσης δεδομένων με όσο το δυνατό πιο κατανοητό και αξιοποιήσιμο τρόπο. Με άλλα λόγια, οι περιγραφικές διαδικασίες της εξόρυξης δεδομένων (Descriptive Data Mining Tasks) περιγράφουν τις γενικές ιδιότητες των υπαρχόντων διαθέσιμων δεδομένων.
- Ταξινόμηση: Σε αυτό το στάδιο έχουμε διαχωρισμό των στοιχείων, με αποτέλεσμα να προκύπτουν διαφορετικές κατηγορίες ή κλάσεις. Για παράδειγμα, οι πελάτες ενός σούπερ μάρκετ είναι δυνατόν να χωριστούν σε παρορμητικούς, πιστούς ή αλλιώς όπως θα λέγαμε κανονικούς, σπάνιους και σε φίλους των εκπτώσεων και προσφορών. Κατά την ανάλυση των πωλήσεων αυτή η κατηγοριοποίηση χρησιμοποιείται για να ληφθούν αποφάσεις, ώστε να προσελκυστούν περισσότεροι πελάτες ανεξαρτήτως κατηγορίας.
- Βελτιστοποίηση: Μεταξύ των άλλων, σκοπός της εξόρυξης γνώσης είναι η βέλτιστη χρήση κάποιων πόρων κάτω από περιορισμούς. Τέτοιοι πόροι μπορεί να είναι ο χρόνος, ο χώρος, το χρήμα και η μεγιστοποίηση κάποιων μεγεθών, όπως είναι τα κέρδη και οι πωλήσεις. Σε αυτή την περίπτωση η εξόρυξη γνώσης έχει κοινά σημεία με την επιχειρησιακή έρευνα.

1.4.7 Data Mining Models

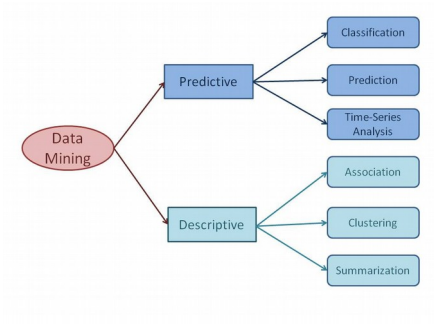
Τα τέσσερα βασικά μοντέλα ανάλυσης των δεδομένων που σχετίζονται με τη διαδικασία του data mining είναι τα εξής:

- Μοντέλο Περιγραφικής Ανάλυσης (Descriptive Analytics): η περιγραφική ανάλυση ζευγαρώνει ακατέργαστα δεδομένα από πολλαπλές πηγές, προκειμένου να δώσει πολύτιμες πληροφορίες σχετικά με το παρελθόν. Η συντριπτική πλειοψηφία των στατιστικών που χρησιμοποιούνται εμπίπτουν σε αυτήν την κατηγορία. Με την χρήση αυτού του μοντέλου δίνεται απάντηση σχετικά με το τι έχει συμβεί σε οποιαδήποτε χρονική στιγμή και έτσι οι αναλυτές κατανοούν πώς οι συμπεριφορές του παρελθόντος μπορούν να επηρεάσουν τα μελλοντικά αποτελέσματα. Ωστόσο, τα ευρήματα που εξάγονται με την συγκεκριμένη μέθοδο δείχνουν εάν κάτι είναι λάθος ή σωστό, χωρίς να εξηγούν το γιατί. Για το λόγο αυτό, το Descriptive Model συνδυάζεται συνήθως και με άλλους τύπους εξόρυξης και ανάλυσης δεδομένων.
- Μοντέλο Διαγνωστικής Ανάλυσης (Diagnostic Analytics): με τη χρήση αυτής της μεθόδου ανάλυσης, τα ιστορικά δεδομένα μπορούν να μετρηθούν με άλλα δεδομένα για να απαντηθεί το ερώτημα γιατί συνέβη ένα γεγονός. Η διαγνωστική ανάλυση δίνει τη δυνατότητα ανίχνευσης εξαρτήσεων και ταυτοποίησης προτύπων, παρέχοντας τελικώς μια βαθιά γνώση ενός συγκεκριμένου προβλήματος.
- Μοντέλο Προγνωστικής Ανάλυσης (Predictive Analytics): οι προγνωστικές αναλύσεις χρησιμοποιούν τα ευρήματα των παραπάνω περιγραφικών και διαγνωστικών αναλύσεων για την ανίχνευση τάσεων, συστάδων και εξαιρέσεων και για την κατανόηση και πρόβλεψη του μέλλοντος. Με το μοντέλο αυτό απαντάται το ερώτημα τι είναι πιθανό να συμβεί. Παρά τα πολυάριθμα πλεονεκτήματα που προσφέρει η προγνωστική ανάλυση, πρέπει να γίνεται κατανοητό ότι η πρόβλεψη είναι απλώς μια εκτίμηση του μέλλοντος με βάσει τις πιθανότητες και όχι βεβαιότητα. Η ακρίβεια της πρόβλεψης εξαρτάται σε μεγάλο βαθμό από την ποιότητα των δεδομένων και τη σταθερότητα της κατάστασης, συνεπώς απαιτεί προσεκτική επεξεργασία και συνεχή βελτιστοποίηση.
- Μοντέλο Κανονιστικής Ανάλυσης (Prescriptive Analytics): στο εν λόγω μοντέλο ανάλυσης χρησιμοποιούνται εξελιγμένα εργαλεία και τεχνολογίες για να απαντηθεί το ερώτημα πώς μπορούμε να κάνουμε κάτι να συμβεί. Η κανονιστική ανάλυση επωφελείται από τα αποτελέσματα των περιγραφικών και προγνωστικών αναλύσεων για να υποδείξει τις διαθέσιμες επιλογές σχετικά με τον τρόπο αξιοποίησης μιας μελλοντικής ευκαιρίας ή την αποτροπή ή ελαχιστοποίηση ενός μελλοντικού κινδύνου και δείχνει την επίδραση που έχει η κάθε απόφαση. Με αυτή τη μέθοδο λαμβάνονται συνεχώς νέα δεδομένα ώστε να επαναπροσδιορίζεται και να βελτιώνεται η ακρίβεια της πρόβλεψης ενώ η χρήση της οδηγεί στην επιλογή των βέλτιστων στρατηγικών αποφάσεων.



Εικόνα 11: Data Mining Models

1.4.8 Μέθοδοι Data Mining



Εικόνα 12: Data Mining Methods

Στο data mining υπάρχουν τρεις κύριες συνιστώσες: αναζήτηση μοντέλου, παράσταση μοντέλου και αξιολόγηση μοντέλου. Οι βασικές μέθοδοι αναζήτησης μοντέλου ψάχνουν και για παραμέτρους και για μοντέλα.

Στην αναζήτηση παραμέτρων ο αλγόριθμος ψάχνει τις ελεύθερες παραμέτρους που βελτιστοποιούν την απόδοση του τελικού μοντέλου. Για απλά προβλήματα η αναζήτηση είναι εύκολη, αλλά για γενικά μοντέλα μία κλειστή λύση δεν είναι εφικτή, και χρησιμοποιούνται μέθοδοι όπως η συζυγής κατάβαση δυναμικού στον αλγόριθμο Back-Propagation για τα Νευρωνικά Δίκτυα.

Η αναζήτηση μοντέλου από την άλλη ψάχνει για το κατάλληλο μοντέλο ή την οικογένεια μοντέλων και για κάθε μία τέτοια δομή που βρίσκει εφαρμόζει έπειτα την αναζήτηση για τις κατάλληλες παραμέτρους του.

Αυτές οι δύο αναζητήσεις είναι χρονοβόρες όταν το μέγεθος του χώρου αναζήτησης είναι μεγάλο και οι υλοποιήσεις τους επωφελούνται ιδιαίτερα από τις τεχνικές παραλληλισμού.

Ένα προβλεπτικό μοντέλο (Predictive Model) κάνει μια πρόβλεψη για τις τιμές των δεδομένων, χρησιμοποιώντας γνωστά αποτελέσματα που έχει βρει από άλλα δεδομένα. Η μοντελοποίηση πρόβλεψης μπορεί να γίνει με βάση τη χρήση ιστορικών δεδομένων. Οι εργασίες εξόρυξης γνώσης από δεδομένα για το χτίσιμο ενός προβλεπτικού μοντέλου περιλαμβάνονται στα παρακάτω υποκεφάλαια.

1.4.8.1 Classification

Αυτή η τεχνική εξόρυξης δεδομένων, επιδιώκει, με την εφαρμογή αλγορίθμων ταξινόμησης σε ένα σύνολο δεδομένων, να εξάγει κανόνες συναρτήσει των οποίων θα γίνει η αντιστοίχιση ενός αντικειμένου με βάση τα χαρακτηριστικά του, σε ένα προκαθορισμένο σύνολο κλάσεων. Το σύνολο των κανόνων που εξάγονται ονομάζεται ταξινομητής (classifier). Η ταξινόμηση προκύπτει από τη διαδικασία εξόρυξης κανόνων ταξινόμησης (Mining Classification Rules).

Κατά τη διαδικασία της κατηγοριοποίησης γίνεται χρήση ενός μέρους των δεδομένων, που ονομάζονται Training Data. Στη συνέχεια, γίνεται χρήση Training Samples για να επιβεβαιωθεί η ακρίβεια του μοντέλου κατηγοριοποίησης που εξήχθη.

Τα Δέντρα Απόφασης (Decision Trees) και τα Νευρωνικά Δίκτυα (Neural Networks) αποτελούν δύο από τους βασικότερους Αλγορίθμους κατηγοριοποίησης. Με τη χρήση Νευρωνικών Δικτύων ως ταξινομητές η κατηγοριοποίηση ανάγεται σε ένα πρόβλημα Density Estimation ή Discrimination ή και Regression. Άλλοι Αλγόριθμοι που χρησιμοποιούνται συχνά είναι οι k-NN και οι Αλγόριθμοι Bayes.

1.4.8.2 Clustering

Η συσταδοποίηση χρησιμοποιείται για να προσδιορίσει τους φυσικούς σχηματισμούς ομάδων από τα δεδομένα βάσει ενός συνόλου κοινών ιδιοτήτων, με τη χρήση διαφόρων αλγορίθμων συσταδοποίησης. Πιο συγκεκριμένα, γίνεται καταμερισμός ενός ετερογενούς πληθυσμού, σε περισσότερες ετερογενείς συστάδες. Με άλλα λόγια, συσταδοποίηση είναι

η εύρεση ομάδων με τέτοιο τρόπο έτσι ώστε τα αντικείμενα σε κάθε ομάδα να είναι όμοια μεταξύ τους, αλλά ταυτόχρονα να διαφέρουν από αντικείμενα που βρίσκονται σε άλλες ομάδες.

Ο k-means αλγόριθμος αποτελεί το βασικότερο αλγόριθμο συσταδοποίησης. Αυτός ο αλγόριθμος έχει ως βασικό στόχο τη βελτιστοποίηση μιας συνάρτησης κόστους. Αρχικά, υπάρχουν k ομάδες, όπου κάθε ομάδα αντιπροσωπεύεται από ένα μέσο διάνυσμα. Μέσα από διαδοχικές επαναλήψεις τα δεδομένα κατατάσσονται σε μία ομάδα σύμφωνα με την ομοιότητα που εμφανίζουν με το μέσο της ομάδας αυτής.

Για να γίνει πιο κατανοητή η συγκεκριμένη μέθοδος, αναφέρουμε το παρακάτω παράδειγμα:

Έστω ένα σύνολο δεδομένων πελατών που περιέχει δύο ιδιότητες: ηλικία και εισόδημα. Ένας αλγόριθμος ομαδοποίησης συγκεντρώνει το σύνολο δεδομένων με βάση αυτές τις ιδιότητες. Η ομάδα 1 περιέχει το νεώτερο πληθυσμό με χαμηλό εισόδημα. Η ομάδα 2 περιέχει τους μέσης ηλικίας πελάτες με υψηλό εισόδημα. Η ομάδα 3 τους μεγαλύτερης ηλικίας με χαμηλό εισόδημα και ούτω καθεξής.

1.4.8.3 Association

Η Συσχέτιση είναι μια από τις βασικότερες μεθόδους εξόρυξης δεδομένων από μεγάλες βάσεις δεδομένων. Με τον όρο αυτό χαρακτηρίζονται οι κανόνες με βάση τους οποίους εκφράζονται οι συσχετίσεις μεταξύ αντικειμένων σε ένα σύνολο δεδομένων και οι οποίοι έχουν τη μορφή $X \rightarrow Y$, δηλαδή κατά πόσο η εμφάνιση ενός συνόλου δεδομένων X (Left Hand Side) έχει ως αποτέλεσμα την εμφάνιση του συνόλου Y (Right Hand Side). Ισχύουν τα εξής:

- $X \subseteq I$
- $Y \subseteq I$
- $X \cap Y = \varnothing$

Τα δύο σύνολα δεδομένων συσχετίζονται μεταξύ τους καθώς η ύπαρξη του ενός, οδηγεί στην ύπαρξη του άλλου. Οι κανόνες προκύπτουν με τη διαδικασία Εξόρυξης Κανόνων Συσχέτισης (Association Rule Mining). Στην καθημερινότητα, οι κανόνες συσχέτισης εφαρμόζονται ευρέως, κυρίως για την έρευνα συμπεριφοράς των καταναλωτών. Μεγάλη έμφαση έχει δοθεί στην εφαρμογή των κανόνων συσχέτισης για την κατανόηση και την ανάλυση του καλαθιού αγοράς (Market Basket Analysis).

Με την εφαρμογή μεθόδων εξόρυξης δεδομένων υπάρχει δυνατότητα να προκύψει ένας τεράστιος όγκος από κανόνες συσχέτισης. Ένας κανόνας συσχέτισης θεωρείται ικανοποιητικός όταν προσφέρει γνώση στον ερευνητή. Για να καθοριστεί ποιο κανόνες είναι σημαντικοί χρησιμοποιούνται τα μέτρα ενδιαφέροντος (interesting measures).

Τα χαρακτηριστικά γνωρίσματα που καθιστούν έναν κανόνα σημαντικό είναι:

- Ο κανόνας που προκύπτει να είναι εύκολα κατανοητός από τον άνθρωπο.
- Ο κανόνας που προκύπτει να έχει ικανοποιητικό βαθμό βεβαιότητας.
- Ο κανόνας που προκύπτει να είναι χρήσιμος και να προσδίδει επιπλέον πληροφορία και γνώση.

Με τα μέτρα ενδιαφέροντος (interesting measures) καθορίζεται το πόσο σημαντικός και ενδιαφέρον είναι ένας κανόνας που προκύπτει από την εξόρυξη δεδομένων. Η σημαντικότητα ενός κανόνα καθορίζεται μέσω δύο μετρικών στατιστικών μεταβλητών˙ των Support και Confidence:

- Support: αντιπροσωπεύει το ποσοστό των συναλλαγών που περιέχουν και το X και το Y στοιχειοσύνολο.

  $$\text{Support} = \sigma(X \bigcap Y) / N \qquad (1)$$

  όπου N είναι το σύνολο των δοσοληψιών.

- Confidence: αποτελεί μια πιθανότητα υπό συνθήκη, P(Y/X), δηλαδή την πιθανότητα μια συναλλαγή που περιέχει το X να περιέχει επίσης και το Y.

  $$\text{Confidence} = \sigma(X \bigcap Y) / \sigma(X) \qquad (2)$$

Με τον όρο Συχνό Στοιχειοσύνολο (Frequent Itemset) χαρακτηρίζονται σύνολα από Στοιχεία (Items) που εμφανίζονται συχνά μαζί, σε ένα σύνολο συναλλαγών. Σε ένα Frequent Itemset το Support είναι μεγαλύτερο από ένα ελάχιστο κατώφλι (min Support Threshold) το οποίο μεταβάλλεται ανάλογα με τη φύση του προβλήματος.

Ένα άλλο μέτρο είναι το lift και εκφράζει το λόγο του παρατηρούμενου Support προς το αναμενόμενο που θα είχαμε αν τα στοιχειοσύνολα X και Y ήταν ανεξάρτητα. Αν το μέτρο lift ενός κανόνα συσχέτισης ισούται με ένα συνεπάγεται ότι τα στοιχειοσύνολα X και Y είναι ασυσχέτιστα. Όταν δύο στοιχειοσύνολα είναι ασυσχέτιστα, δεν μπορεί να εξαχθεί κάποιος κανόνας ο οποίος να συμπεριλαμβάνει τα δύο αυτά γεγονότα. Αν η τιμή του μέτρου lift είναι μεγαλύτερου του ενός τότε μπορούμε να γνωρίζουμε το βαθμό στον οποίο συσχετίζονται μεταξύ τους τα δύο γεγονότα. Όσο μεγαλύτερη είναι η τιμή του lift, τόσο πιο πιθανό είναι η ύπαρξη του X και του Y ταυτόχρονα σε μία συναλλαγή να μη συμβαίνει τυχαία.

$$\text{Lift}(X,Y) = P(X \bigcap Y) / [P(X) \cdot P(Y)] \qquad (3)$$

Απ’ τα μέτρα ενδιαφέροντος μπορούν να εξαχθούν τα εξής συμπεράσματα:

- Μικρή τιμή του μέτρου Support σημαίνει ότι ο κανόνας έχει μικρό ενδιαφέρον, καθώς αφορά σε ένα μικρό αριθμό συναλλαγών και επομένως μπορεί να εξαιρεθεί.
- Ένας κανόνας με μικρό Support υπάρχει πιθανότητα να εμφανίζεται τυχαία.
- Το μέτρο Confidence μετρά την αξιοπιστία. Όσο μεγαλύτερο είναι το μέτρο του Confidence, τόσο μεγαλύτερη είναι η πιθανότητα εμφάνισης του στοιχειοσυνόλου Y σε κανόνα που περιέχει το στοιχειοσύνολο X.
- Κανόνες που προέρχονται από το ίδιο στοιχειοσύνολο, έχουν το ίδιο Support.

Οι κανόνες συσχέτισης μπορούν να κατηγοριοποιηθούν ως εξής:

- Boolean κανόνες συσχέτισης: Αναφέρονται στην ύπαρξη ή μη ενός αντικειμένου σε έναν κανόνα συσχέτισης. Για παράδειγμα αν ένας πελάτης Χ αγοράσει έναν υπολογιστή, θα αγοράσει και λογισμικό Antivirus.
- Ποσοτικοί κανόνες συσχέτισης: Περιγράφουν συσχετίσεις μεταξύ ποσοτικών αντικειμένων.
- Κανόνες συσχέτισης μίας ή πολλών διαστάσεων: Διαχωρίζονται με βάση τον αριθμό των ιδιοτήτων που περιλαμβάνουν.
- Κανόνες συσχέτισης επιπέδου: Προκύπτουν από την ύπαρξη ιεραρχικών επιπέδων ενός Item. Για παράδειγμα για την ιδιότητα “ηλικία” μπορεί να προκύψουν διαφορετικοί κανόνες συσχέτισης για κάθε ηλικιακό διαστήματα τιμών.

Το πρόβλημα της εύρεσης κανόνων συσχέτισης εστιάζεται στην εύρεση όλων των κανόνων που έχουν μία καθορισμένη από το χρήστη ελάχιστη τιμή Support και Confidence. Χρησιμοποιώντας για είσοδο ένα σύνολο από Τ συναλλαγές, λαμβάνονται σαν έξοδος όλοι οι κανόνες που έχουν Support μεγαλύτερο από ένα κατώφλι min_Support και Confidence μεγαλύτερο από ένα κατώφλι min_Confidence. Οι τιμές των κατωφλιών έχουν οριστεί εκ των προτέρων.
Για την εύρεση κανόνων συσχέτισης ακολουθούνται τα εξής βήματα:

- Παράγονται όλοι οι πιθανοί κανόνες συσχέτισης,
- Υπολογίζεται το Support και το Confidence για κάθε έναν κανόνα που έχει παραχθεί,
- Εξαιρούνται οι κανόνες με μικρότερο Support και Confidence από τα κατώφλια min_Support και min_Confidence που έχουν οριστεί.

Αν υπάρχουν “n” διαφορετικά στοιχεία (Items) τότε ισχύει ότι:

- Ο συνολικός αριθμός στοιχειοσυνόλων θα είναι 2^n
- Ο συνολικός αριθμός των πιθανών κανόνων συσχέτισης είναι: $3^n - 2^{n+1} + 1$

1.4.8.4 Regression

Η Παλινδρόμηση (Regression) είναι μια ευρέως χρησιμοποιημένη στατιστική τεχνική μοντελοποίησης για την έρευνα της συσχέτισης μεταξύ μίας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Έτσι η παλινδρόμηση μπορεί να απεικονίζει ένα στοιχειώδες δεδομένο x σε μια πραγματική μεταβλητή πρόβλεψης y.

Η παλινδρόμηση περιλαμβάνει την εκμάθηση μιας συνάρτησης y=f(x) που εκτελεί αυτήν την απεικόνιση καθορίζοντας τη βέλτιστη συνάρτηση (γραμμική, μη-γραμμική, πολυωνυμική κλπ.) για τη μοντελοποίηση των δεδομένων. Όταν χρησιμοποιείται για παρεμβολή (interpolation) σημείων σε ενδιάμεσα τμήματα μπορεί να κάνει και κατηγοριοποίηση.

Η συνάρτηση παλινδρόμησης προβλέπει την συνάρτηση συμμετοχής του ανύσματος x στην κλάση με τιμή y. Η γραμμική παλινδρόμηση $y = \alpha_1 \cdot \chi_1 + \alpha_2 \cdot \chi_2 + ... + \alpha_\nu \cdot \chi_\nu$

υποθέτει γραμμικές συσχετίσεις και μπορεί να βρει έτσι μία διαχωριστική συνάρτηση που τέμνει έναν υπόχωρο σε δύο περιοχές κλάσεων.

Τα Τεχνητά νευρωνικά δίκτυα χρησιμοποιούνται ευρύτατα για εκτίμηση σημείων, ή και για εκτίμηση συνάρτησης, παλινδρόμηση, πρόβλεψη και κατηγοριοποίηση. Στην αξιολόγηση μοντέλων υπάρχει το στάνταρ Mean Squared Error και η Cross Entropy Loss Function για την παλινδρόμηση και κατηγοριοποίηση αντίστοιχα. Δένδρα Παλινδρόμησης, Κανόνες και Regression Splines χρησιμοποιούνται επίσης στην Προβλεπτική Μοντελοποίηση (Predictive Modeling) αν και μπορούν επίσης να εφαρμοστούν και στην Περιγραφική Μοντελοποίηση.

Η παλινδρόμηση μπορεί να εφαρμοστεί σε διάφορους τομείς, όπως στη μετεωρολογία για να προβλεφθούν οι ταχύτητες ανέμου με βάση τη θερμοκρασία, την πίεση αέρα, και την υγρασία.

1.4.8.5 Time Series Analysis

Στην ανάλυση χρονολογικών σειρών ή χρονοσειρών, μελετάται η τιμή ενός γνωρίσματος καθώς μεταβάλλεται στο χρόνο. Οι τιμές λαμβάνονται σε ίσα χρονικά διαστήματα (ημερήσια, εβδομαδιαία, ωριαία, κ.λ.π.). Για να παρασταθούν οπτικά οι χρονοσειρές χρησιμοποιείται ένα διάγραμμα χρονοσειρών.

Υπάρχουν τρεις βασικές λειτουργίες που πραγματοποιούνται στην ανάλυση χρονοσειρών.

- χρησιμοποιούνται μονάδες μέτρησης απόστασης για να καθοριστεί η ομοιότητα ανάμεσα σε διαφορετικές χρονοσειρές,
- εξετάζεται η δομή της χρονοσειράς για να καθοριστεί και να κατηγοριοποιηθεί η συμπεριφορά της,
- γίνεται χρήση διαγραμμάτων χρονοσειρών για την πρόβλεψη μελλοντικών τιμών.

Τέλος μία πρόσφατη λειτουργία είναι η εύρεση των ίδιων των κατηγοριών των χρονοσειρών.

1.4.8.6 Sequential Pattern Discovery

Η ανακάλυψη προτύπων ακολουθιών χρησιμοποιείται για να καθοριστούν σειριακά πρότυπα στα δεδομένα.

Μια ακολουθία αποτελείται από μια σειρά διακριτών τιμών (ή καταστάσεων). Μια ακολουθία DNA είναι μια μακρά ακολουθία από τέσσερα διαφορετικά μέρη: Αδενίνη (Adenine), Θυμίνη (Thymine), Κυτοσίνη (Cytosine) και Γουανίνη (Guanine). Μια επιλογή στον Παγκόσμιο Ιστό είναι μια ακολουθία από ιστοσελίδες. Οι αγορές πελατών μπορούν επίσης να διαμορφωθούν ως στοιχεία ακολουθίας. Χαρακτηριστικό παράδειγμα αποτελεί

ένας πελάτης που αγοράζει αρχικά έναν υπολογιστή, έπειτα ένα μικρόφωνο και τελικά μια Web κάμερα.

Το κοινό των μεθόδων ανάλυσης ακολουθίας και συσχετισμού (Association) είναι ότι κάθε μεμονωμένη περίπτωση περιέχει ένα σύνολο στοιχείων ή καταστάσεων. Η διαφορά τους είναι ότι οι μέθοδοι ανάλυσης ακολουθίας αναλύουν τις μεταβάσεις καταστάσεων ενώ η μέθοδος συσχετισμού θεωρεί κάθε στοιχείο ίσο και ανεξάρτητο. Σύμφωνα με τη μέθοδο ανάλυσης ακολουθίας, το να αγοράσει κάποιος έναν υπολογιστή προτού αγοράσει μικρόφωνο είναι μια διαφορετική ακολουθία από το να αγοράσει μικρόφωνο πριν από έναν υπολογιστή. Για έναν αλγόριθμο συσχετισμού, αυτά θεωρούνται όμοια.

Η ανάλυση ακολουθίας είναι ένας σχετικά νέος τρόπος εξόρυξης δεδομένων. Είναι αρκετά σημαντική σε δύο κυρίως τύπους εφαρμογών: Ανάλυση Παγκόσμιου Ιστού και Ανάλυση DNA. Υπάρχουν διάφορες τεχνικές ανάλυσης ακολουθίας διαθέσιμες όπως οι αλυσίδες Markov κ.α..

1.4.8.7 Forecasting

Η πρόβλεψη είναι μια ακόμα σημαντική μέθοδος εξόρυξης δεδομένων. Μπορεί να βοηθήσει στην απάντηση ερωτημάτων όπως:

- Ποια θα είναι η αξία ενός αποθεματικού αύριο;
- Ποιο θα είναι το ποσοστό πωλήσεων αναψυκτικών για τον επόμενο μήνα κ.λ.π.;

Τα δεδομένα εισόδου είναι τύπου χρονικής σειράς. Οι μέθοδοι πρόβλεψης εξετάζουν γενικές τάσεις και περιοδικότητα. Ως δημοφιλέστερη μέθοδος πρόβλεψης θεωρείται η ARIMA, η οποία υλοποιεί τη μεθοδολογία Auto Regressive Integrated Moving Average Model.

1.4.8.8 Summarization

Η παρουσίαση συνόψεων ή συνοπτικών μοντέλων αφορά μεθόδους που βρίσκουν και απεικονίζουν τα δεδομένα σε υποσύνολα τους. Η σύνοψη χαρακτηρίζει τα δεδομένα και παράγει αντιπροσωπευτικές πληροφορίες σχετικά με τις βάσεις δεδομένων. Αυτό είναι χρήσιμο στην κατανόηση της μεγαλύτερης αξίας μερικών γνωρισμάτων έναντι άλλων.

Βασικές έννοιες της στατιστικής όπως ο μέσος, η διακύμανση, η τυπική απόκλιση αποτελούν απλά μοντέλα ενός πληθυσμού. Το “ταίριασμα” ενός πληθυσμού σε μία κατανομή παρέχει ένα καλύτερο μοντέλο δεδομένων.

1.4.9 Αλγόριθμοι Data Mining

1.4.9.1 Επιλογή Αλγόριθμου

Η επιλογή του καλύτερου αλγόριθμου που θα χρησιμοποιηθεί για ένα συγκεκριμένο αναλυτικό έργο μπορεί να είναι μια πρόκληση για έναν αναλυτή διότι ενώ υπάρχει η δυνατότητα χρήσης διαφορετικών αλγορίθμων για την ίδια εργασία, κάθε αλγόριθμος παράγει ένα διαφορετικό αποτέλεσμα. Επιπλέον, κάποιοι αλγόριθμοι είναι σε θέση να φέρουν εις πέρας διαφορετικά project. Για παράδειγμα τα Decision Trees μπορούν να αξιοποιηθούν όχι μόνο για την πρόβλεψη, αλλά και για τη μείωση του αριθμού των στηλών σε ένα σύνολο δεδομένων καθώς δύναται να εντοπίσουν στήλες που δεν επηρεάζουν το τελικό μοντέλο εξόρυξης.

1.4.9.2 Neural Networks

Εικόνα 13: Neural Networks

Η μελέτη των τεχνητών νευρωνικών δικτύων ξεκίνησε από τις προσπάθειες για προσομοίωση βιολογικών νευρωνικών συστημάτων καθώς οι πρώτες αρχές και λειτουργίες τους βασίζονται και εμπνέονται από το νευρικό σύστημα ζώντων οργανισμών.

Η λειτουργία των νευρωνικών δικτύων προσπαθεί να συνδυάσει τον τρόπο σκέψης του ανθρώπινου εγκεφάλου με τον αφηρημένο τρόπο μαθηματικής σκέψης. Ο ανθρώπινος εγκέφαλος αποτελείται κυρίως από νευρικά κύτταρα τα οποία ονομάζονται νευρώνες που διασυνδέονται με άλλους νευρώνες μέσω νηματοειδών ινών, τους νευρίτες. Νευρωνικό σύστημα καλείται ένα κύκλωμα διασυνδεδεμένων νευρώνων. Στην περίπτωση τεχνητών νευρώνων πρόκειται για ένα αφηρημένο αλγοριθμικό κατασκεύασμα το οποίο εμπίπτει στον τομέα υπολογιστικής νοημοσύνης και στόχος του είναι η επίλυση κάποιου υπολογιστικού προβλήματος.

Τα τεχνητά νευρωνικά έχουν αναπτυχθεί τις τελευταίες δεκαετίες. Ο κλάδος των νευρωνικών με τη σημερινή του μορφή έχει διαχωριστεί εντελώς από τον τομέα της βιολογίας και πλέον χρησιμοποιείται για να λύσει άλλου είδους προβλήματα. Μέχρι τώρα

ερευνητές έχουν χρησιμοποιήσει επιτυχημένα τα τεχνητά νευρωνικά δίκτυα για να μοντελοποιήσουν και να προβλέψουν την σχέση μεταξύ της αντίστασης του εδάφους και του μήκους των τοποθετημένων ηλεκτροδίων στο έδαφος, για να συσχετίσουν την ειδική αντίσταση του εδάφους με την συχνότητα διάδοσης του ρεύματος και την αντίσταση του εδάφους, ενώ έχουν αναπτυχθεί μοντέλα νευρωνικών δικτύων και για το σχεδιασμό συστημάτων γείωσης που αποτελούνται από κατακόρυφες ράβδους.

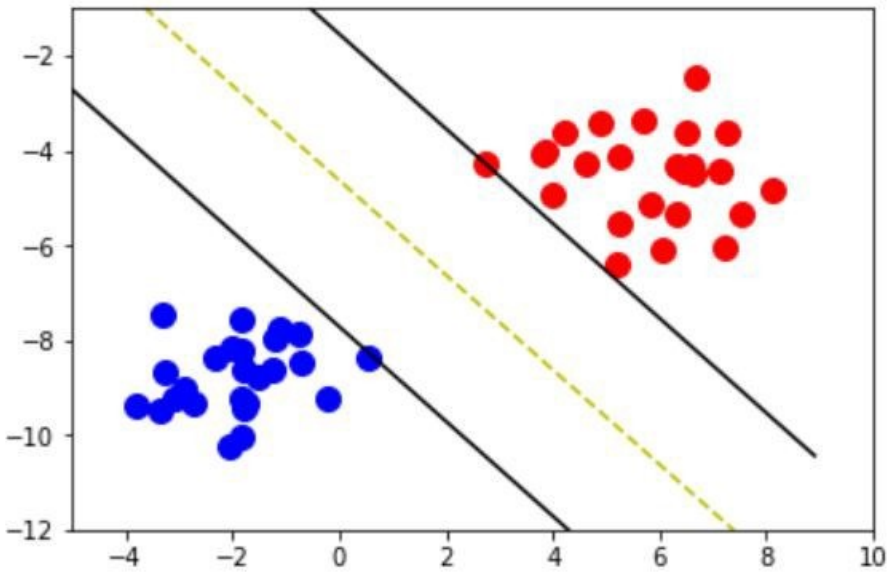
Ένα νευρωνικό δίκτυο περιλαμβάνει μια σειρά από εξαρτημένους επεξεργαστές ή κόμβους. Οι κόμβοι αυτοί συνδέονται με άλλους κόμβους οι οποίοι είναι οργανωμένοι σε στρώματα (layers). Οι προβλέψεις γίνονται συναρτήσει των μεταβλητών εισόδου που εισρέουν στο δίκτυο και των βαρών που σχετίζονται με τις συνδέσεις του δικτύου.

1.4.9.3 Feedforward Neural Networks

Στις περισσότερες περιπτώσεις, ένα τεχνητό νευρωνικό δίκτυο είναι ένα προσαρμοζόμενο σύστημα που αλλάζει τη δομή του βασιζόμενο σε εξωτερικές ή εσωτερικές πληροφορίες που εισρέουν μέσα από το δίκτυο κατά τη διάρκεια της εκμάθησης. Τα σύγχρονα νευρωνικά δίκτυα χρησιμοποιούνται συνήθως για τη μοντελοποίηση σύνθετων σχέσεων μεταξύ εισόδου και εξόδου.

Ένα Feedforward Neural Network είναι ένα τεχνητό νευρωνικό δίκτυο στο οποίο όλες οι συνδέσεις μεταξύ των μονάδων δεν αποτελούν ένα κατευθυνόμενο κύκλο. Σε αυτά τα δίκτυα οι πληροφορίες κινούνται μόνο από τους κόμβους εισόδου προς τους κόμβους εξόδου μέσω των κρυφών κόμβων (hidden nodes) όταν αυτοί είναι διαθέσιμοι. Επίσης δεν υπάρχουν κύκλοι και βρόγχοι στο δίκτυο.

1.4.9.4 SVM



Εικόνα 14: Support Vector Machines

Οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines) είναι μοντέλα επιβλεπομένης μάθησης με σχετικούς αλγορίθμους ανάλυσης δεδομένων και αναγνώρισης μοτίβων, που χρησιμοποιούνται για την ταξινόμηση και την ανάλυση παλινδρόμησης. Ο βασικός SVM παίρνει ένα σύνολο από δεδομένα εισόδου και προβλέπει, για καθένα απ' αυτά, μία από τις δύο δυνατές κλάσης κατηγοριοποίησης. Λαμβάνοντας υπόψη ένα σύνολο παραδειγμάτων εκπαίδευσης, ένας αλγόριθμος εκπαίδευσης SVM χτίζει ένα μοντέλο που εκχωρεί νέα παραδείγματα σε μία από τις δύο κλάσεις. Ένα μοντέλο SVM είναι μια αναπαράσταση των παραδειγμάτων ως σημεία στο χώρο, χαρτογραφώντας με τέτοιο τρόπο τα παραδείγματα των ξεχωριστών κλάσεων ώστε να διαχωρίζονται από ένα σαφές κενό που είναι όσο το δυνατόν πιο ευρύ. Τα νέα παραδείγματα στη συνέχεια χαρτογραφούνται και αυτά στον ίδιο χώρο και ταξινομούνται σε κλάση ανάλογα με τη θέση τους.

Εκτός από την εκτέλεση γραμμικής ταξινόμησης, οι αλγόριθμοι SVM μπορούν να εκτελέσουν αποτελεσματικά μη-γραμμική ταξινόμηση αλλάζοντας τον τύπο του πυρήνα.

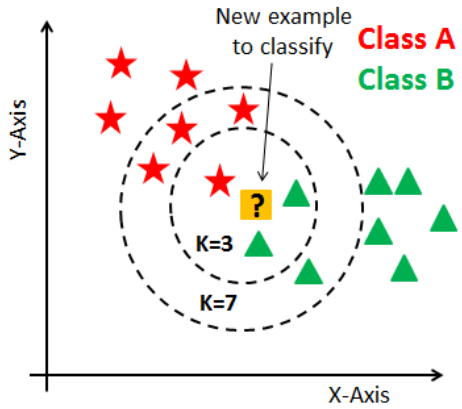
Αναλυτικότερα, ένας αλγόριθμος SVM κατασκευάζει ένα υπέρ-επίπεδο ή ένα σύνολο υπέρ-επίπεδων, που μπορούν να χρησιμοποιηθούν για την ταξινόμηση, την παλινδρόμηση καθώς και για άλλες εργασίες. Ένας καλός διαχωρισμός επιτυγχάνεται από το υπέρ-επίπεδο που έχει τη μεγαλύτερη απόσταση από το πλησιέστερο σημείο δεδομένων εκπαίδευσης οποιασδήποτε κλάσης (ονομάζεται και λειτουργικό περιθώριο ή functional margin), δεδομένου ότι σε γενικές γραμμές όσο μεγαλύτερο το περιθώριο τόσο χαμηλότερο το σφάλμα γενίκευσης του ταξινομητή.

1.4.9.5 Naïve Bayes

Ο ταξινομητής Naïve Bayes είναι ένας πιθανολογικός ταξινομητής βασιζόμενος στην εφαρμογή του θεωρήματος Bayes με ισχυρές υποθέσεις ανεξαρτησίας. Με απλούς όρους, ο ταξινομητής Naïve Bayes υποθέτει ότι η παρουσία (ή απουσία) ενός συγκεκριμένου χαρακτηριστικού της μιας κλάσης δεν έχει σχέση με την παρουσία (ή απουσία) οποιοδήποτε άλλου χαρακτηριστικού, δεδομένης της μεταβλητής της κλάσης.

Ο συγκεκριμένος ταξινομητής μπορεί να εκπαιδευτεί πολύ αποτελεσματικά σε επιτηρούμενο περιβάλλον μάθησης.

1.4.9.6 k-NN



Εικόνα 15: k-NN

Ο αλγόριθμος k-πλησιέστερου γείτονα (k-NN) είναι ένας απ' τους πιο απλούς αλγόριθμους της μηχανικής μάθησης. Με τη χρήση του k-NN κάθε αντικείμενο κατηγοριοποιείται ανάλογα με την κλάση στην οποία ανήκει η πλειοψηφία των πλησιέστερων γειτόνων του. Οι πλησιέστεροι γείτονες καθορίζονται απ' την τιμή του k, που είναι πάντοτε θετική και συνήθως μικρή. Αν k = 1, τότε το αντικείμενο απλώς καταχωρείται στην κλάση του πρώτου γείτονα.

Οι γείτονες λαμβάνονται από ένα σύνολο αντικειμένων για τα οποία η σωστή κατάταξη είναι γνωστή. Αυτό μπορεί να θεωρηθεί ως το σύνολο εκπαίδευσης για τον αλγόριθμο, αν και κανένα ρητό στάδιο εκπαίδευσης δεν απαιτείται.

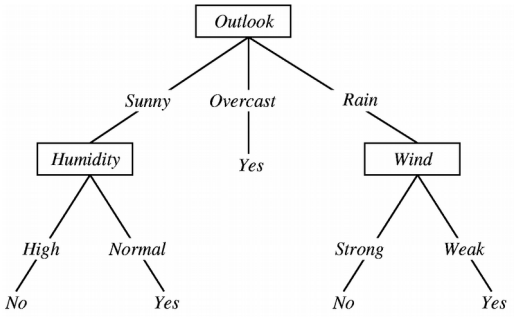
Τα παραδείγματα εκπαίδευσης είναι διανύσματα σε ένα πολυδιάστατο χώρο χαρακτηριστικών, το καθένα με μια ετικέτα κλάσης. Η φάση της εκπαίδευσης του αλγορίθμου, αποτελείται μόνο από την αποθήκευση των χαρακτηριστικών διανυσμάτων και των ετικετών κλάσης των δειγμάτων εκπαίδευσης.

Στη φάση της κατηγοριοποίησης, το k αποτελεί μια καθορισμένη σταθερά από το χρήστη, και ένα μη κατηγοριοποιημένο διάνυσμα (unlabeled vector) ταξινομείται στην κλάση στην οποία ανήκουν οι περισσότεροι k γείτονες του.

Συνήθως για την εύρεση των k κοντινότερων γειτόνων χρησιμοποιείται η Ευκλείδεια απόσταση. Ωστόσο αυτό ισχύει μόνο για συνεχείς μεταβλητές. Σε περιπτώσεις όπως η ταξινόμηση κειμένου, χρησιμοποιούνται άλλα είδη μέτρησης όπως η απόσταση Hamming. Συχνά, η ακρίβεια ταξινόμησης του k-NN μπορεί να βελτιωθεί σημαντικά αν η μέτρηση απόστασης γίνει μέσω εξειδικευμένων αλγορίθμων, όπως ο αλγόριθμος πλησιέστερου γείτονα μεγάλου περιθωρίου (Large Margin Nearest Neighbor) ή ο αλγόριθμος ανάλυσης γειτονικών στοιχείων (Neighbourhood Components Analysis).

Η καλύτερη επιλογή του k εξαρτάται κυρίως από τα δεδομένα. Γενικά, μεγαλύτερες τιμές του k μειώνουν την επίδραση του θορύβου κατά την ταξινόμηση, αλλά κάνουν τα όρια μεταξύ των τάξεων λιγότερο διακριτά. Ένα καλό k μπορεί να επιλεγεί από διάφορες έξυπνες ευρετικές τεχνικές, όπως για παράδειγμα το Cross-Validation.

1.4.9.7 Decision Trees



Εικόνα 16: Decision Trees

Τα δέντρα απόφασης παράγονται από αλγορίθμους που προσδιορίζουν διάφορους τρόπους διάσπασης ενός συνόλου δεδομένων σε μικρότερα τμήματα. Αυτά τα τμήματα σχηματίζουν ένα ανεστραμμένο δέντρο απόφασης που έχει έναν αρχικό κόμβο (ρίζα) στην κορυφή του. Το αντικείμενο της ανάλυσης εκφράζεται σε αυτόν τον κόμβο- ρίζα ως μια απλή, μονοδιάστατη απεικόνιση του περιβάλλοντος του δέντρου απόφασης. Το όνομα του πεδίου των δεδομένων που είναι το αντικείμενο της ανάλυσης εμφανίζεται συνήθως, μαζί με την κατανομή των τιμών που περιέχονται σε αυτόν τον τομέα.

Μόλις εξαχθεί η σχέση μεταξύ των δεδομένων, τότε ένας ή περισσότεροι κανόνες απόφασης που περιγράφουν τη σχέση των δεδομένων μπορούν να συνταχθούν. Κάθε κανόνας ορίζει μια εγγραφή από το σύνολο δεδομένων σε έναν κόμβο ή σε ένα κλαδί με βάση την τιμή του σε ένα από τα πεδία ή τις στήλες του συνόλου δεδομένων. Τα πεδία ή οι στήλες που χρησιμοποιούνται για τη δημιουργία του κανόνα ονομάζονται είσοδοι. Οι κανόνες διάσπασης εφαρμόζονται ο ένας μετά τον άλλο, με την δημιουργία κλαδιών μέσα στα υπάρχοντα κλαδιά με αποτέλεσμα την σύσταση του χαρακτηριστικού ανεστραμμένου δέντρου απόφασης.

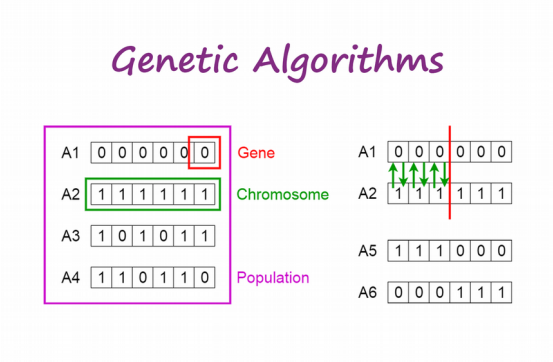
Ένας κόμβος που περιέχει και άλλους κόμβους πιο κάτω από αυτόν ονομάζεται πρόγονος. Ομοίως απόγονοι είναι οι κόμβοι που βρίσκονται ιεραρχικά κάτω από τον

πρόγονο. Οι τελευταίοι κόμβοι κάτω από το δέντρο απόφασης ονομάζονται φύλλα (ή τερματικοί κόμβοι) και κάθε φύλλο αποτελεί μια κλάση. Για κάθε φύλλο, ο κανόνας απόφασης παρέχει μία μοναδική διαδρομή για να εισέλθουν τα δεδομένα στην κλάση που ορίζει. Όλοι οι κόμβοι, συμπεριλαμβανομένων των φύλλων, έχουν αμοιβαίους αποκλειστικούς κανόνες εκχώρησης. Ως αποτέλεσμα, εγγραφές από το σύνολο δεδομένων μπορούν να βρεθούν σε έναν μόνο κόμβο τη φορά. Μόλις οι κανόνες απόφασης καθορισθούν, είναι πλέον δυνατόν να εφαρμοσθούν για την πρόβλεψη και την ταξινόμηση σε νέα δεδομένα που δεν έχουν χρησιμοποιηθεί στην εκπαίδευση.

Μία νέα μορφή δέντρων απόφασης είναι τα Random Forests. Τα τυχαία δάση είναι πολύ-δέντρα που χρησιμοποιούν τυχαία δείγματα από το σύνολο δεδομένων και τεχνικές επαναστάθμισης για την ανάπτυξη πολλαπλών δέντρων που, όταν συνδυαστούν, οδηγούν σε πιο αξιόπιστες προβλέψεις.

Για να δημιουργήσουμε ένα δέντρο απόφασης από ένα συγκεκριμένο σύνολο δεδομένων χρησιμοποιείται ένας αλγόριθμος που ονομάζεται επαγωγέας (inducer). Υπάρχουν αρκετοί επαγωγείς όπως ο ID3, ο C4.5 και ο αλγόριθμος CART.

1.4.9.8 Γενετικός Αλγόριθμος



Εικόνα 17: Γενετικοί Αλγόριθμοι

Ο όρος γενετικός αλγόριθμος χρησιμοποιήθηκε για πρώτη φορά από τον John Holland, του οποίου το βιβλίο Adaptation in Natural and Aritificial Systems (1975) συνέβαλε στη δημιουργία και την σημερινή ανάπτυξη που έχει ο τομέας της έρευνας και της εφαρμογής των γενετικών αλγορίθμων.

Ο Holland φαντάστηκε ότι κάποιες ιδέες και λειτουργίες που εφαρμόζει η φύση στα συστήματά της θα μπορούσαν να έχουν αποτελέσματα, αν ενσωματώνονταν σε αλγόριθμους για υπολογιστές, ώστε να προκύψουν αποδοτικές τεχνικές επίλυσης δύσκολων προβλημάτων. Αποτέλεσμα της εργασίας του Holland ήταν οι Γενετικοί Αλγόριθμοι, μια πολλά υποσχόμενη τεχνική αναζήτησης και βελτιστοποίησης.

Οι γενετικοί αλγόριθμοι χρησιμοποιούν ορολογία δανεισμένη από το χώρο της Γενετικής. Κατ' αναλογία με τα έμβια όντα, αναφέρονται σε άτομα ή γονότυπα μέσα σε έναν πληθυσμό. Πολύ συχνά αυτά τα άτομα καλούνται επίσης και χρωμοσώματα. Τα

χρωμοσώματα αποτελούνται από διάφορα στοιχεία που ονομάζονται γονίδια και είναι διατεταγμένα σε γραμμική ακολουθία. Κάθε γονίδιο επηρεάζει την κληρονομικότητα ενός ή περισσότερων χαρακτηριστικών. Τα γονίδια που επηρεάζουν συγκεκριμένα χαρακτηριστικά γνωρίσματα του ατόμου βρίσκονται σε συγκεκριμένες θέσεις του χρωματοσώματος που καλούνται τόποι (loci). Κάθε χαρακτηριστικό γνώρισμα του ατόμου (όπως για παράδειγμα το χρώμα μαλλιών) έχει την δυνατότητα να εμφανιστεί με διάφορες μορφές, ανάλογα με την κατάσταση στην οποία βρίσκεται το αντίστοιχο γονίδιο που το επηρεάζει. Οι διαφορετικές αυτές καταστάσεις, που μπορεί να πάρει το γονίδιο, καλούνται αλληλόμορφα.

Κάθε γονότυπος, που στις περισσότερες περιπτώσεις είναι ένα μόνο χρωμόσωμα, αναπαριστά μια πιθανή λύση σε ένα πρόβλημα. Το μεταφρασμένο περιεχόμενο του συγκεκριμένου χρωμοσώματος καλείται φαινότυπος και καθορίζεται από τον χρήστη, ανάλογα με τις ανάγκες και τις απαιτήσεις του. Μια διαδικασία εξέλιξης που εφαρμόζεται πάνω σε έναν πληθυσμό χρωμοσωμάτων αντιστοιχεί σε μία εκτενή αναζήτηση μέσα σε ένα χώρο από πιθανές λύσεις. Απαραίτητη προϋπόθεση για την επιτυχημένη έκβαση μιας τέτοιας αναζήτησης αποτελεί η εξισορρόπηση δύο διαδικασιών που είναι προφανώς αντικρουόμενες˙ της εκμετάλλευσης και διατήρησης των καλύτερων λύσεων και της όσο το δυνατόν καλύτερης εξερεύνησης όλου του διαστήματος.

Κατά τη διάρκεια της τελευταίας δεκαετίας, το ενδιαφέρον για τις διαδικασίες βελτιστοποίησης έχει αυξηθεί σε τόσο μεγάλο βαθμό, ώστε υπάρχουν πολύπλοκα και με πολύ αυστηρούς περιορισμούς προβλήματα, που επιδέχονται μόνο προσεγγιστικές λύσεις από τους σημερινούς υπολογιστές. Οι γενετικοί αλγόριθμοι αποσκοπούν στην εξυπηρέτηση τέτοιου είδους προβλημάτων. Εάν και ανήκουν στην κατηγορία των στοχαστικών αλγορίθμων, διαφέρουν σε πολύ μεγάλο βαθμό από τους αλγόριθμους που εφαρμόζουν τυχαίες μεθόδους αναζήτησης και βελτιστοποίησης, αφού είναι σε θέση να συνδυάζουν στοιχεία από άμεσες αλλά και από στοχαστικές τεχνικές. Αυτός είναι και ο κύριος λόγος για τον οποίον θεωρούνται οι πιο εύρωστοι από τις υπάρχουσες μεθόδους άμεσης αναζήτησης. Ένα άλλο εξίσου σημαντικό χαρακτηριστικό τους είναι ότι διατηρούν έναν πληθυσμό πιθανών λύσεων πάνω στον οποίο πειραματίζονται, σε αντίθεση με άλλες μεθόδους που επεξεργάζονται ένα μόνο σημείο του διαστήματος αναζήτησης.

Ο γενετικός αλγόριθμος πραγματοποιεί αναζήτηση σε διάφορες κατευθύνσεις με το να διατηρεί έναν πληθυσμό από πιθανές λύσεις και να υποστηρίζει καταγραφή και ανταλλαγή πληροφοριών μεταξύ αυτών των κατευθύνσεων. Ο πληθυσμός υφίσταται μια προσομοιωμένη γενετική εξέλιξη. Σε κάθε γενιά, οι σχετικά καλές λύσεις αναπαράγονται, ενώ οι υπόλοιπες αφαιρούνται. Ο διαχωρισμός και η αξιολόγηση των διαφόρων λύσεων γίνεται με τη βοήθεια μιας αντικειμενικής συνάρτησης ή συνάρτησης ικανότητας (objective ή fitness function), η οποία παίζει το ρόλο του περιβάλλοντος μέσα στο οποίο εξελίσσεται ο πληθυσμός.

Κατά την διάρκεια της επαναληπτικής εκτέλεσης t, ο γενετικός αλγόριθμος διατηρεί έναν πληθυσμό από πιθανές λύσεις:

$$P(t)=\{x_1^2 ,....,x_n^t \} \qquad (4)$$

Κάθε λύση x_i^t αξιολογείται και δίνει ένα μέτρο της καταλληλότητας και ορθότητάς της. Αφού ολοκληρωθεί η αξιολόγηση όλων των στοιχείων του πληθυσμού, δημιουργείται ένας

πληθυσμός που προκύπτει από την επιλογή των πιο κατάλληλων στοιχείων της προηγούμενης γενιάς. Μερικά μέλη από τον καινούριο πληθυσμό υφίστανται μετατροπές με τη βοήθεια των διαδικασιών της μετάλλαξης και της διασταύρωσης σχηματίζοντας νέες πιθανές λύσεις. Η διασταύρωση συνδυάζει τα στοιχεία δύο χρωμοσωμάτων των γονέων για να δημιουργήσει δύο νέους απογόνους.

Για παράδειγμα, έστω ότι οι γονείς αναπαριστώνται με διανύσματα πέντε διαστάσεων (a1, b1, c1, d1, e1) και (a2, b2, c2, d2, e2) , τότε οι απόγονοι (με σημείο διασταύρωσης —crossover point = 2) είναι οι (a1,b1, c2 ,d2 ,e2) και (a2, b2, c1, d1, e1). Αν οι μεταβλητές στα παραπάνω διανύσματα είναι δυαδικές, τότε κάθε διάνυσμα αναπαριστά την τιμή μιας μεταβλητής, δηλαδή ένα χρωμόσωμα. Στην περίπτωση που είναι πραγματικές, τότε καθεμία είναι ένα χρωμόσωμα, δηλαδή κάθε διάνυσμα αναπαριστά τις τιμές πολλών μεταβλητών και συνεπώς αποτελεί ένα γονότυπο.

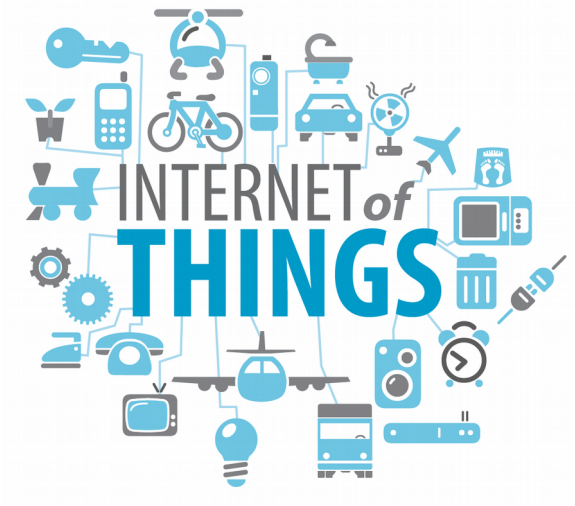
Η διαδικασία της μετάλλαξης αλλάζει αυθαίρετα ένα ή περισσότερα γονίδια του συγκεκριμένου χρωμοσώματος εξυπηρετώντας την εισαγωγή νέων πιθανών λύσεων, διαφορετικών από τις υπάρχουσες, στον ήδη υπάρχοντα πληθυσμό.

Ο γενετικός αλγόριθμος για ένα συγκεκριμένο πρόβλημα πρέπει να αποτελείται από τα παρακάτω πέντε τμήματα:

- Μια γενετική αναπαράσταση των πιθανών λύσεων του προβλήματος.
- Έναν τρόπο δημιουργίας του αρχικού πληθυσμού των πιθανών λύσεων.
- Μια αντικειμενική συνάρτηση αξιολόγησης που παίζει το ρόλο του περιβάλλοντος, κατατάσσοντας τις λύσεις με βάση την καταλληλότητά τους.
- Γενετικούς τελεστές που μετατρέπουν τη σύνθεση των απογόνων.
- Τιμές για διάφορες παραμέτρους όπως μέγεθος πληθυσμού, πιθανότητες εφαρμογής των γενετικών τελεστών κ.λ.π..

1.5 IoT

1.5.1 Εισαγωγή



Εικόνα 18: IoT

Το Internet of Things αναδεικνύεται ως μία από τις σημαντικότερες τάσεις που διαμορφώνουν την ανάπτυξη τεχνολογιών στον τομέα ICT. Η μετάβαση από ένα διαδίκτυο που χρησιμοποιείται για τη διασύνδεση συσκευών τελικού χρήστη σε ένα διαδίκτυο για τη διασύνδεση φυσικών αντικειμένων που επικοινωνούν μεταξύ τους ή / και με ανθρώπους για να προσφέρουν μια συγκεκριμένη υπηρεσία, περιλαμβάνει την ανάγκη να επανεξεταστούν εκ νέου μερικές από τις συμβατικές μεθόδους που χρησιμοποιούνται συνήθως στη δικτύωση, την πληροφορική και την παροχή / διαχείριση υπηρεσιών (networking, computing and service provisioning / management).

Από ένα εννοιολογικό πλαίσιο, το IoT βασίζεται σε τρεις πυλώνες, που σχετίζονται με την ικανότητα των έξυπνων αντικειμένων: (i) να αναγνωρίζονται (anything identifies itself), (ii) να επικοινωνούν (anything communicates) και (iii) να αλληλεπιδρούν (anything interacts) είτε μεταξύ τους, δημιουργώντας δίκτυα διασυνδεδεμένων αντικειμένων είτε με end-users είτε ακόμη και με άλλες οντότητες στο δίκτυο. Η ανάπτυξη τεχνολογιών και λύσεων για την πραγματοποίηση ενός τέτοιου οράματος αποτελεί την κύρια πρόκληση που αντιμετωπίζουμε.

Το IoT είναι άμεσα συνυφασμένο με την έννοια των “έξυπνων αντικειμένων”. Ως έξυπνα αντικείμενα (ή πράγματα) ορίζουμε τις οντότητες που πληρούν τις εξής προϋποθέσεις:

- Έχουν μια φυσική ενσάρκωση και ένα σύνολο συναφών φυσικών χαρακτηριστικών (π.χ. μέγεθος, σχήμα κ.λ.π.).
- Έχουν ένα ελάχιστο σύνολο λειτουργιών επικοινωνίας, όπως τη δυνατότητα να ανακαλύπτουν και να δέχονται εισερχόμενα μηνύματα και να απαντούν σε αυτά.
- Διαθέτουν ένα μοναδικό αναγνωριστικό.

- Συσχετίζονται με τουλάχιστον ένα όνομα και μία διεύθυνση. Το όνομα είναι μια ανθρώπινη αναγνώσιμη περιγραφή του αντικειμένου. Η διεύθυνση είναι μια ακολουθία που μπορεί να διαβαστεί από το μηχάνημα και μπορεί να χρησιμοποιηθεί για την επικοινωνία με το αντικείμενο.
- Διαθέτουν κάποιες βασικές δυνατότητες που κυμαίνονται από την ικανότητα αντιστοίχισης ενός εισερχόμενου μηνύματος μ' ένα δεδομένο ίχνος (όπως σε παθητικά RFID) έως την ικανότητα εκτέλεσης πολύπλοκων υπολογισμών, συμπεριλαμβανομένων εργασιών εντοπισμού και διαχείρισης δικτύου.
- Έχουν στη διάθεσή τους μέσα για τη μέτρηση φυσικών μεγεθών (π.χ. θερμοκρασία, φως, επίπεδο ηλεκτρομαγνητικής ακτινοβολίας κ.λ.π.) ή/ και για την ενεργοποίηση δράσεων που επηρεάζουν τη φυσική πραγματικότητα (ενεργοποιητές).

Το Internet of Things μπορεί να θεωρηθεί ως ένα εξαιρετικά δυναμικό και ριζικά κατανεμημένο δικτυωμένο σύστημα, το οποίο αποτελείται από ένα πολύ μεγάλο αριθμό έξυπνων αντικειμένων που παράγουν και καταναλώνουν πληροφορίες. Η ικανότητα διασύνδεσης με τη φυσική σφαίρα επιτυγχάνεται μέσω της παρουσίας συσκευών ικανών να ανιχνεύουν φυσικά φαινόμενα και να τα μεταφράζουν σε μια ροή δεδομένων (παρέχοντας έτσι πληροφορίες για το τρέχον περιβάλλον), καθώς και μέσω της παρουσίας συσκευών ικανών να ενεργοποιούν δράσεις που έχουν αντίκτυπο σ' αυτήν (μέσω κατάλληλων ενεργοποιητών). Δεδομένου ότι η επεκτασιμότητα αναμένεται να αποτελέσει μείζον ζήτημα λόγω της εξαιρετικά μεγάλης κλίμακας του συστήματος που προκύπτει και λαμβάνοντας επίσης υπόψη την υψηλή δυναμική του όλου εγχειρήματος (καθώς τα έξυπνα αντικείμενα μπορούν να κινηθούν και να δημιουργήσουν ad hoc σύνδεση, ακολουθώντας απρόβλεπτα πρότυπα), η ενσωμάτωση της αυτοδιαχείρισης αναμένεται να αποτελέσει μείζονα κινητήρια δύναμη στην ανάπτυξη ενός συνόλου πιθανών λύσεων.

Από την άποψη της εξυπηρέτησης, το κύριο ζήτημα σχετίζεται με τον τρόπο ενσωμάτωσης (ή σύνθεσης) των λειτουργιών ή / και των πόρων που παρέχονται από έξυπνα αντικείμενα (σε πολλές περιπτώσεις υπό τη μορφή ροής δεδομένων) στις υπηρεσίες. Το γεγονός αυτό απαιτεί τον ορισμό: (i) αρχιτεκτονικών και μεθόδων για την εικονικοποίηση (virtualization) αντικειμένων με τη δημιουργία μιας τυποποιημένης αναπαράστασης έξυπνων αντικειμένων στον ψηφιακό τομέα, ικανών να εμποδίσουν την ετερογένεια των συσκευών / πόρων και (ii) μεθόδων για την ομαλή ενσωμάτωση και σύνθεση των πόρων/ υπηρεσιών των αντικειμένων σε υπηρεσίες προστιθέμενης αξίας (value-added services) για τους τελικούς χρήστες.

Το όραμα του Internet of Things παρέχει μια μεγάλη σειρά ευκαιριών στους χρήστες, τους κατασκευαστές και τις εταιρείες. Οι τεχνολογίες IoT θα έχουν ευρεία εφαρμογή σε πολλούς παραγωγικούς τομείς, όπως π.χ. στην παρακολούθηση του περιβάλλοντος, στην ιατρική περίθαλψη, στη διαχείριση αποθεμάτων και προϊόντων, στη στήριξη στο χώρο εργασίας και στο σπίτι, στην ασφάλεια και στην επιτήρηση. Από την οπτική γωνία του χρήστη, το IoT θα επιτρέψει τη δημιουργία ενός μεγάλου αριθμού νέων υπηρεσιών που θ' ανταποκρίνονται πάντα στις ανάγκες των χρηστών και οι οποίες θα είναι προσβάσιμες και αξιοποιήσιμες απ' το ευρύ κοινό σε καθημερινή βάση. Η εμφάνιση του Internet of Things θα προκαλέσει μια μετατόπιση στην παροχή υπηρεσιών από το σημερινό όραμα των always- on services χαρακτηριστικών της εποχής του Παγκοσμίου Ιστού, στις always-responsive situated services, σχεδιασμένες και διευθετημένες σε πραγματικό χρόνο, για να ανταποκριθούν σε μια συγκεκριμένη ανάγκη κατανοώντας το πλαίσιο στο οποίο ανήκει ο χρήστης.

Ενώ το όραμα του IoT θα απαιτήσει σημαντική πρόοδο σε ορισμένους τομείς των ICT, η υλοποίησή του πιθανότατα θα ακολουθήσει μια διαδοχική διαδικασία, ξεκινώντας από τις υπάρχουσες τεχνολογίες και εφαρμογές. Ειδικότερα, το IoT πιθανότατα θα επεκταθεί ξεκινώντας από τεχνολογίες αναγνώρισης όπως η RFID (Radio Frequency Identification) οι οποίες χρησιμοποιούνται ήδη ευρέως σε πολλές εφαρμογές. Ταυτόχρονα, στην πορεία ανάπτυξής του, το IoT θα βασιστεί πιθανώς σε προσεγγίσεις, όπως ασύρματα δίκτυα αισθητήρων (ως μέσο συλλογής δεδομένων) και αρχιτεκτονικές προσανατολισμένες στις υπηρεσίες (Service-oriented Architectures (SoA)) όπως η αρχιτεκτονική λογισμικού για την επέκταση των Web-based υπηρεσιών μέσω των δυνατοτήτων του IoT.

Συνοψίζοντας, μπορούμε να προσδιορίσουμε τα βασικά χαρακτηριστικά σε επίπεδο συστήματος, τα οποία πρέπει να υποστηρίξει το Internet of Things:

- Ετερογένεια συσκευών: Το IoT θα χαρακτηρίζεται από μεγάλη ετερογένεια όσον αφορά τις συσκευές που συμμετέχουν στο σύστημα, οι οποίες αναμένεται να παρουσιάζουν πολύ διαφορετικές δυνατότητες τόσο από υπολογιστική όσο και από επικοινωνιακή άποψη. Η διαχείριση τέτοιου υψηλού επιπέδου ετερογένειας θα πρέπει να υποστηρίζεται τόσο αρχιτεκτονικά όσο και σε πρωτόκολλα.

- Ευελιξία: Καθώς τα καθημερινά αντικείμενα συνδέονται με μια παγκόσμια υποδομή πληροφοριών, τα ζητήματα κλιμάκωσης προκύπτουν σε διάφορα επίπεδα, μεταξύ των οποίων: (i) ονοματοδοσία και διευθυνσιοδότηση (εξαιτίας του μεγάλου μεγέθους του προκύπτοντος συστήματος), (ii) επικοινωνία δεδομένων και δικτύωση/ διασύνδεση μεταξύ μεγάλου αριθμού φορέων, iii) διαχείριση πληροφορίας και γνώσης (λόγω της δυνατότητας οικοδόμησης ενός digital αντιστοίχου σε οποιαδήποτε οντότητα ή / και φαινόμενο στη φυσική σφαίρα) και iv) παροχή και διαχείριση υπηρεσιών (λόγω του μαζικού αριθμού των επιλογών εκτέλεσης υπηρεσιών (service execution options) που θα μπορούσαν να είναι διαθέσιμες και της ανάγκης να διαχειριστούν ετερογενείς πόρους).

- Πανταχού παρούσα ανταλλαγή δεδομένων μέσω ασύρματων τεχνολογιών εγγύτητας (proximity wireless technologies): Στο IoT, θα διαδραματίσουν εξέχοντα ρόλο οι τεχνολογίες ασύρματων επικοινωνιών, οι οποίες θα επιτρέψουν τη δικτύωση των έξυπνων αντικειμένων. Η πανταχού παρούσα υιοθέτηση του ασύρματου μέσου ανταλλαγής δεδομένων μπορεί να δημιουργήσει προβλήματα όσον αφορά τη διαθεσιμότητα του ραδιοφάσματος, προωθώντας την υιοθέτηση δυναμικών ραδιοσυστημάτων.

- Energy-optimized solutions: Για μια ποικιλία φορέων διαδικτύου, η ελαχιστοποίηση της ενέργειας που θα δαπανηθεί για σκοπούς επικοινωνίας / computing θα αποτελέσει πρωταρχικό στόχο. Ενώ οι τεχνικές που σχετίζονται με τη συγκομιδή ενέργειας (μέσω π.χ. πιεζοηλεκτρικών υλικών ή μικρο- ηλιακών συλλεκτών) θα απαλλάξουν τις συσκευές από τους περιορισμούς που επιβάλλονται από τις λειτουργίες των μπαταριών, η ενέργεια θα είναι πάντοτε ένας σπάνιος πόρος. Με τον τρόπο αυτό θα καταστεί όλο και πιο ελκυστική η ανάγκη να ευρεθούν λύσεις που τείνουν να βελτιστοποιούν τη χρήση ενέργειας (ακόμη και με επιδόσεις).

- Localization and tracking capabilities: Καθώς οι οντότητες στο διαδίκτυο μπορούν να αναγνωριστούν και διαθέτουν δυνατότητες ασύρματης επικοινωνίας μικρής εμβέλειας, γίνεται δυνατή η παρακολούθηση της θέσης (και της κίνησης) έξυπνων αντικειμένων στη φυσική σφαίρα. Αυτό είναι ιδιαίτερα σημαντικό για εφαρμογές στην εφοδιαστική και στη διαχείριση του κύκλου ζωής των προϊόντων όπου έχουν ήδη υιοθετηθεί εκτεταμένες τεχνολογίες RFID.

- Δυνατότητες αυτο-οργάνωσης: Η πολυπλοκότητα και η δυναμική που πολλά σενάρια IoT πιθανότατα θα παρουσιάσουν ως προς τη διανομή πληροφοριών στο σύστημα, θα καταστήσουν τα έξυπνα αντικείμενα (ή ένα υποσύνολο αυτών) ικανά να αντιδράσουν αυτόνομα σε ένα ευρύ φάσμα διαφορετικών καταστάσεων, προκειμένου να ελαχιστοποιηθεί η ανθρώπινη παρέμβαση. Κατόπιν αιτημάτων των χρηστών, οι κόμβοι του Διαδικτύου θα οργανώνονται αυτόνομα σε μεταβατικά ad hoc δίκτυα, παρέχοντας τα βασικά μέσα για την ανταλλαγή δεδομένων και την εκτέλεση συντονισμένων εργασιών.

- Σημασιολογική διαλειτουργικότητα και διαχείριση δεδομένων: Το IoT θα ασχολείται με την ανταλλαγή και την ανάλυση τεράστιων ποσοτήτων δεδομένων. Προκειμένου να μετατραπούν σε χρήσιμες πληροφορίες και να εξασφαλιστεί η διαλειτουργικότητα μεταξύ διαφορετικών εφαρμογών, είναι απαραίτητο να παρέχονται επαρκείς και τυποποιημένες μορφές, μοντέλα και σημασιολογική περιγραφή του περιεχομένου τους (μεταδεδομένα), χρησιμοποιώντας σαφώς καθορισμένες γλώσσες και μορφές. Αυτό θα επιτρέψει στις εφαρμογές IoT να υποστηρίξουν την αυτοματοποιημένη συλλογιστική, ένα βασικό χαρακτηριστικό που θα επιτρέψει την επιτυχή υιοθέτηση μιας τέτοιας τεχνολογίας σε ευρεία κλίμακα.

- Ενσωματωμένοι μηχανισμοί ασφάλειας και προστασίας της ιδιωτικής ζωής: Λόγω της στενής αλληλεπίδρασης με τη φυσική σφαίρα, η τεχνολογία του IoT πρέπει να είναι ασφαλής και να προστατεύει την ιδιωτικότητα εκ σχεδιασμού. Αυτό σημαίνει ότι η ασφάλεια θα πρέπει να θεωρείται βασική ιδιότητα σε επίπεδο συστήματος και να λαμβάνεται υπόψη στο σχεδιασμό αρχιτεκτονικών και μεθόδων για λύσεις IoT. Αυτό αναμένεται να αποτελέσει βασική προϋπόθεση για την αποδοχή του IoT από τους χρήστες και την ευρεία υιοθέτηση της τεχνολογίας.

1.5.2 Εφαρμογές

Οι εφαρμογές του IoT βρίσκονται ακόμη σε πρώιμο στάδιο. Ωστόσο, η χρήση του IoT εξελίσσεται και αυξάνεται ταχέως. Αρκετές εφαρμογές IoT αναπτύσσονται ή εφαρμόζονται σε διάφορες βιομηχανίες, συμπεριλαμβανομένης της παρακολούθησης του περιβάλλοντος, της ιατρικής περίθαλψης, της διαχείρισης αποθεμάτων και παραγωγής, της αλυσίδας εφοδιασμού τροφίμων (food supply chain (FSC)), της μεταφοράς, της υποστήριξης στο χώρο εργασίας και της κατοικίας, της ασφάλειας και της επιτήρησης. Παρακάτω παραθέτουμε κάποια παραδείγματα:

- Smart Homes/Smart Buildings: Η οργάνωση κτιρίων με προηγμένες τεχνολογίες IoT μπορεί να συμβάλει τόσο στη μείωση της κατανάλωσης πόρων που σχετίζονται με τα κτίρια (ηλεκτρική ενέργεια, νερό), όσο και στη βελτίωση του επιπέδου διαβίωσης των ανθρώπων που κατοικούν σ' αυτά, είτε πρόκειται για εργαζόμενους σε κτίρια γραφείων ή για ενοικιαστές ιδιωτικών κατοικιών. Ο αντίκτυπος είναι δυνατός τόσο από οικονομικής απόψεως (μειωμένες επιχειρησιακές δαπάνες) όσο και από κοινωνικής (μείωση του αποτυπώματος άνθρακα που συνδέεται με τα κτίρια το οποίο αποτελεί βασικό παράγοντα για τις παγκόσμιες εκπομπές αερίων που σχετίζονται με το φαινόμενο του θερμοκηπίου). Στην εφαρμογή αυτή, βασικό ρόλο διαδραματίζουν οι αισθητήρες, οι οποίοι χρησιμοποιούνται τόσο για την παρακολούθηση της κατανάλωσης πόρων, όσο και για την ανίχνευση των αναγκών των σημερινών χρηστών. Ένα τέτοιο σενάριο ενσωματώνει διάφορα υποσυστήματα και συνεπώς απαιτεί υψηλό επίπεδο τυποποίησης για τη διασφάλιση της διαλειτουργικότητας. Είναι επίσης αναγκαία η ικανότητα λογικής με κατανεμημένο και συνεργατικό τρόπο και η ενεργοποίησή της, προκειμένου να διασφαλιστεί ότι οι αποφάσεις που λαμβάνονται σχετικά με τους υπό έλεγχο πόρους (π.χ. ενεργοποίηση / απενεργοποίηση φωτισμού, θέρμανσης, ψύξης κλπ.) είναι σύμφωνες με τις ανάγκες και προσδοκίες των χρηστών, οι οποίες με τη σειρά τους συνδέονται αυστηρά με τις δραστηριότητες που αναλαμβάνουν ή σκοπεύουν να αναλάβουν.

- Smart Cities: Στις "Έξυπνες Πόλεις" (Smart Cities) αναφερθήκαμε στο υποκεφάλαιο των Big Data Analytics και ως εκ τούτου θα περιοριστούμε στο να προσθέσουμε μερικές επιπλέον υλοποιήσεις που άπτονται του ενδιαφέροντός μας. Μία απ' αυτές τις εφαρμογές είναι το σύστημα "έξυπνων συσκευών σταθμεύσεως", (βασισμένο σε τεχνολογίες RFID και αισθητήρων) το οποίο μπορεί να επιτρέψει την παρακολούθηση των διαθέσιμων χώρων στάθμευσης και την παροχή στους οδηγούς αυτοματοποιημένων συμβουλών, βελτιώνοντας έτσι την κινητικότητα στην αστική περιοχή. Ακόμη, οι αισθητήρες μπορούν να παρακολουθούν τη ροή της κυκλοφορίας οχημάτων σε αυτοκινητόδρομους και να ανακτούν συγκεντρωτικές πληροφορίες όπως η μέση ταχύτητα και ο αριθμός των αυτοκινήτων. Οι αισθητήρες θα μπορούσαν να ανιχνεύουν το επίπεδο ρύπανσης του αέρα, να ανακτούν πληροφορίες σχετικά με την αιθαλομίχλη, όπως το επίπεδο διοξειδίου του άνθρακα, του PM10 κ.λ.π., και να παρέχουν αυτές τις πληροφορίες σε φορείς υγείας. Τέλος, οι αισθητήρες θα μπορούσαν να χρησιμοποιηθούν σε ιατροδικαστικές διαδικασίες, ανιχνεύοντας παραβιάσεις και διαβιβάζοντας τα σχετικά στοιχεία στις υπηρεσίες επιβολής του νόμου προκειμένου να εντοπίσουν τον παραβάτη ή να αποθηκεύσουν τις πληροφορίες που θα παρασχεθούν σε περίπτωση ατυχήματος για την επακόλουθη ανάλυση.

- Περιβαλλοντική παρακολούθηση: Η τεχνολογία IoT μπορεί να εφαρμοστεί κατάλληλα σε εφαρμογές περιβαλλοντικής παρακολούθησης. Στην περίπτωση αυτή, ο ρόλος- κλειδί αφορά στην ικανότητα να ανιχνεύονται, με τρόπο κατανεμημένο και αυτοδιαχειριζόμενο, φυσικά φαινόμενα και διαδικασίες (π.χ. θερμοκρασία, αιολική ενέργεια, βροχόπτωση, ύψος ποταμού), καθώς και να ενσωματώνονται αποτελεσματικά τέτοια ετερογενή δεδομένα σε παγκόσμιες εφαρμογές. Η επεξεργασία πληροφοριών σε πραγματικό χρόνο, σε συνδυασμό με την ικανότητα ενός μεγάλου αριθμού συσκευών να επικοινωνούν μεταξύ τους, παρέχει μια σταθερή πλατφόρμα για την ανίχνευση και την παρακολούθηση των

“ανωμαλιών” που μπορούν να οδηγήσουν σε κίνδυνο για τη ζωή των ανθρώπων και των ζώων. Η εκτεταμένη ανάπτυξη των μικροσκοπικών συσκευών μπορεί να επιτρέψει την πρόσβαση σε κρίσιμες ζώνες, όπου η παρουσία ανθρώπινων φορέων δεν αποτελεί βιώσιμη επιλογή (π.χ. ηφαιστειακές περιοχές, ωκεάνιες αβύσσους, απομακρυσμένες περιοχές), από όπου οι πληροφορίες μπορούν να κοινοποιηθούν σε ένα σημείο λήψης αποφάσεων προκειμένου να ανιχνευθούν “ανώμαλες” συνθήκες. Από αυτή την άποψη, οι τεχνολογίες IoT μπορούν να επιτρέψουν την ανάπτυξη μιας νέας γενιάς συστημάτων παρακολούθησης και υποστήριξης αποφάσεων, παρέχοντας βελτιωμένη χωρητικότητα και την ικανότητα επεξεργασίας και εξαγωγής συμπερασμάτων σε πραγματικό χρόνο σε σχέση με τις τρέχουσες λύσεις.
Μια άλλη περίπτωση περιβαλλοντικής παρακολούθησης αποτελεί και η ανίχνευση πυρκαγιάς. Όταν μια ακολουθία αισθητήρων ανιχνεύει την πιθανή παρουσία πυρκαγιάς (μέσω π.χ. αισθητήρων θερμοκρασίας), ένας συναγερμός αποστέλλεται απευθείας στην πυροσβεστική υπηρεσία σε σύντομο χρονικό διάστημα (αξιοποιώντας τα προηγμένα χαρακτηριστικά επικοινωνίας της πλατφόρμας IoT), μαζί με άλλες παραμέτρους που είναι χρήσιμες στη λήψη αποφάσεων, όπως η περιγραφή της περιοχής που υπόκειται στη φωτιά, η πιθανή παρουσία ανθρώπων, εύφλεκτων υλικών κλπ. Είναι σαφές ότι η ταχεία αντίδραση έχει ως συνέπεια τη διάσωση ανθρώπινων ζωών, την άμβλυνση των ζημιών στην ιδιοκτησία ή τη βλάστηση και γενικά τη μείωση του βαθμού καταστροφής.
Στην Κίνα, οι ετικέτες RFID και / ή οι bar codes συνδέονται με τα προϊόντα πυρόσβεσης για την ανάπτυξη βάσεων δεδομένων σχετικά με τα προϊόντα πυροπροστασίας και τα συστήματα διαχείρισης. Οι ερευνητές στην Κίνα χρησιμοποιούν επίσης τεχνολογίες IoT για την κατασκευή αυτόματων συστημάτων συναγερμού πυρκαγιάς, προκειμένου να αυξήσουν την αποτελεσματικότητα διαχείρισης της πυρκαγιάς και γενικότερα των καταστάσεων έκτακτης ανάγκης.
Πολλά άλλα σενάρια που σχετίζονται με την προστασία των πολιτών μπορούν να επωφεληθούν από τις τεχνολογίες IoT (περιοχή σήραγγας, σεισμός, τσουνάμι κ.λ.π.), όπου η δυνατότητα πρόσβασης σε περιβαλλοντικά δεδομένα σε πραγματικό χρόνο επιτρέπει την υιοθέτηση αποδοτικών στρατηγικών συντονισμού μεταξύ των ομάδων διάσωσης.

- Διαχείριση αποθεμάτων και προϊόντων: Οι τεχνολογίες βιοαισθητήρων σε συνδυασμό με την τεχνολογία RFID μπορούν να επιτρέψουν τον έλεγχο της διεξαγωγής των διαδικασιών παραγωγής, της ποιότητας του τελικού προϊόντος και της πιθανής υποβάθμισης του προϊόντος στο ράφι, π.χ. στη βιομηχανία τροφίμων. Για παράδειγμα, οι συσκευές RFID μπορούν να χρησιμοποιηθούν για τον εντοπισμό και την παρακολούθηση του προϊόντος, ενώ οι βιοαισθητήρες μπορούν να παρακολουθούν παραμέτρους όπως η θερμοκρασία και η βακτηριακή σύνθεση, προκειμένου να διασφαλιστεί η απαιτούμενη ποιότητα.

- Ασφάλεια και επιτήρηση: Η επιτήρηση ασφάλειας έχει γίνει μια αναγκαιότητα για κτίρια επιχειρήσεων, εμπορικά κέντρα, χώρους στάθμευσης αυτοκινήτων και πολλούς άλλους δημόσιους χώρους. Τα σενάρια εσωτερικής ασφάλειας αντιμετωπίζουν παρόμοιες απειλές, αν και σε διαφορετική κλίμακα. Οι τεχνολογίες IoT μπορούν να χρησιμοποιηθούν για να βελτιώσουν σημαντικά την απόδοση των σημερινών λύσεων, παρέχοντας φθηνότερες και λιγότερο επεμβατικές εναλλακτικές λύσεις σε σχέση με την ευρεία χρήση καμερών, διατηρώντας

ταυτόχρονα την ιδιωτικότητα των χρηστών. Οι αισθητήρες περιβάλλοντος μπορούν να χρησιμοποιηθούν για την παρακολούθηση της παρουσίας επικίνδυνων χημικών ουσιών. Οι αισθητήρες που παρακολουθούν τη συμπεριφορά των ανθρώπων μπορούν να χρησιμοποιηθούν για να εκτιμηθεί η παρουσία ανθρώπων που δρουν με ύποπτο τρόπο. Συνεπώς, μπορούν να κατασκευαστούν αποδοτικά συστήματα έγκαιρης προειδοποίησης. Η προσωπική ταυτοποίηση μέσω RFID ή παρόμοιων τεχνολογιών είναι επίσης μια επιλογή. Ωστόσο, σε πολλές χώρες οι ενώσεις χρηστών διαμαρτύρονται έντονα για την παραβίαση της ιδιωτικής ζωής που θα μπορούσε να προκύψει από την ευρεία υιοθέτηση μιας τέτοιας τεχνολογίας.

- Χρήση του IoT σε FSC: Η σημερινή FSC είναι εξαιρετικά κατανεμημένη και πολύπλοκη. Έχει μεγάλη γεωγραφική και χρονική κλίμακα, πολύπλοκες διαδικασίες λειτουργίας και μεγάλο αριθμό ενδιαφερομένων. Η πολυπλοκότητα έχει προκαλέσει πολλά ζητήματα στη διαχείριση της ποιότητας, τη λειτουργική αποτελεσματικότητα και τη δημόσια ασφάλεια των τροφίμων. Οι τεχνολογίες IoT προσφέρουν ελπιδοφόρες δυνατότητες για την αντιμετώπιση των προκλήσεων ιχνηλασιμότητας, προβολής και ελέγχου. Μια τυπική λύση IoT για το FSC (το λεγόμενο Food-IoT) περιλαμβάνει τρία μέρη: α) τις συσκευές πεδίου όπως οι κόμβοι WSN, οι ετικέτες RFID, οι τερματικοί σταθμοί διεπαφής χρήστη κ.λ.π. β) το σύστημα κορμού όπως βάσεις δεδομένων, διακομιστές και πολλά είδη τερματικών που συνδέονται με κατανεμημένα δίκτυα υπολογιστών κ.λ.π. και γ) τις υποδομές επικοινωνίας όπως WLAN, κυψελοειδές, δορυφορικό, ηλεκτρικό δίκτυο, Ethernet κ.λ.π. Καθώς το σύστημα IoT προσφέρει πανταχού παρούσα ικανότητα δικτύωσης, όλα αυτά τα στοιχεία μπορούν να διανεμηθούν σε ολόκληρο το FSC. Επιπλέον, το IoT προσφέρει αποτελεσματικές λειτουργίες εντοπισμού και παρακολούθησης της διαδικασίας παραγωγής τροφίμων. Το τεράστιο ποσό των ακατέργαστων δεδομένων μπορεί να εξορύσσεται και να αναλύεται περαιτέρω για να βελτιωθεί η επιχειρηματική διαδικασία και να στηριχθεί η λήψη αποφάσεων. Τα Big Data Analytics μπορεί να χρησιμοποιηθούν για την ανάλυση του τεράστιου όγκου δεδομένων που συλλέγονται από την FSC.

- Χρήση του IoT για ασφαλέστερη εξόρυξη: Η ασφάλεια της εξόρυξης αποτελεί μεγάλη ανησυχία για πολλές χώρες λόγω του βαθμού επικινδυνότητας που χαρακτηρίζει την εργασία στα υπόγεια ορυχεία. Προκειμένου να αποφευχθούν και να μειωθούν τα ατυχήματα στον τομέα της εξόρυξης, είναι απαραίτητο να χρησιμοποιηθούν τεχνολογίες του IoT για την ανίχνευση σημάτων καταστροφών, ούτως ώστε να καταστεί δυνατή η έγκαιρη προειδοποίηση, η πρόγνωση καταστροφών και η βελτίωση της ασφάλειας της υπόγειας παραγωγής. Με τη χρήση τεχνολογίας RFID και άλλων ασύρματων τεχνολογιών και συσκευών ασύρματης επικοινωνίας για την αποτελεσματική επικοινωνία μεταξύ επιφάνειας και υπόγειου εδάφους, οι επιχειρήσεις εξόρυξης μπορούν να εντοπίσουν την τοποθεσία των υπόγειων ανθρακωρύχων και να αναλύσουν κρίσιμα δεδομένα που συλλέγονται από τους αισθητήρες. Μια άλλη χρήσιμη εφαρμογή είναι η χρήση χημικών και βιολογικών αισθητήρων για την έγκαιρη ανίχνευση και διάγνωση ασθενειών. Αυτοί οι χημικοί και βιολογικοί αισθητήρες μπορούν να χρησιμοποιηθούν για την απόκτηση βιολογικών πληροφοριών από το ανθρώπινο σώμα και τα όργανα του και για την ανίχνευση επιβλαβούς σκόνης, επιβλαβών αερίων και άλλων περιβαλλοντικών κινδύνων. Μια πρόκληση είναι ότι οι

ασύρματες συσκευές χρειάζονται ενέργεια και θα μπορούσαν ενδεχομένως να εξαπολύσουν αέρια στο ορυχείο. Απαιτούνται περισσότερες έρευνες σχετικά με τα χαρακτηριστικά ασφαλείας των συσκευών IoT που χρησιμοποιούνται στην εξορυκτική παραγωγή.

1.6 Cloud

1.6.1 Ορισμός

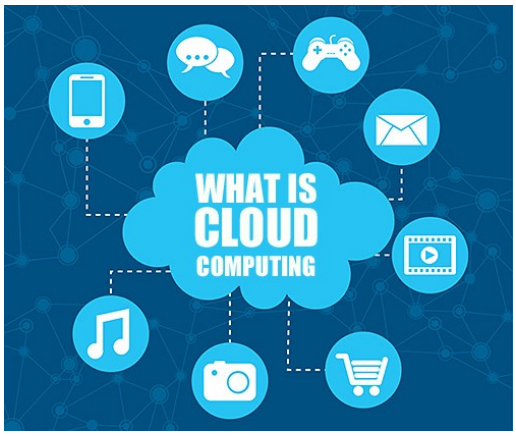
Εικόνα 19: Cloud Technology

Cloud Computing ονομάζεται η κατ' αίτηση διαδικτυακή κεντρική διάθεση υπολογιστικών πόρων (όπως δίκτυο, εξυπηρετητές, εφαρμογές και υπηρεσίες) με υψηλή ευελιξία, ελάχιστη προσπάθεια από τον χρήστη και υψηλή αυτοματοποίηση.

Στο Υπολογιστικό Νέφος η αποθήκευση, η επεξεργασία και η χρήση δεδομένων, λογισμικού και υπηρεσιών γίνεται διαδικτυακά, μέσω απομακρυσμένων υπολογιστών σε κεντρικά Datacenter. Έτσι οι χρήστες εξοικονομούν πόρους από την αγορά και συντήρηση λογισμικού και τη συντήρηση ακριβών εξυπηρετητών και εγκαταστάσεων αποθήκευσης δεδομένων.

Υπηρεσίες όπως η κατ' αίτηση παροχή εικονικών μηχανών, το διαδικτυακό ηλεκτρονικό ταχυδρομείο ή τα κοινωνικά δίκτυα συχνά βασίζονται στην τεχνολογία του Υπολογιστικού Νέφους.

1.6.2 Χαρακτηριστικά του Cloud

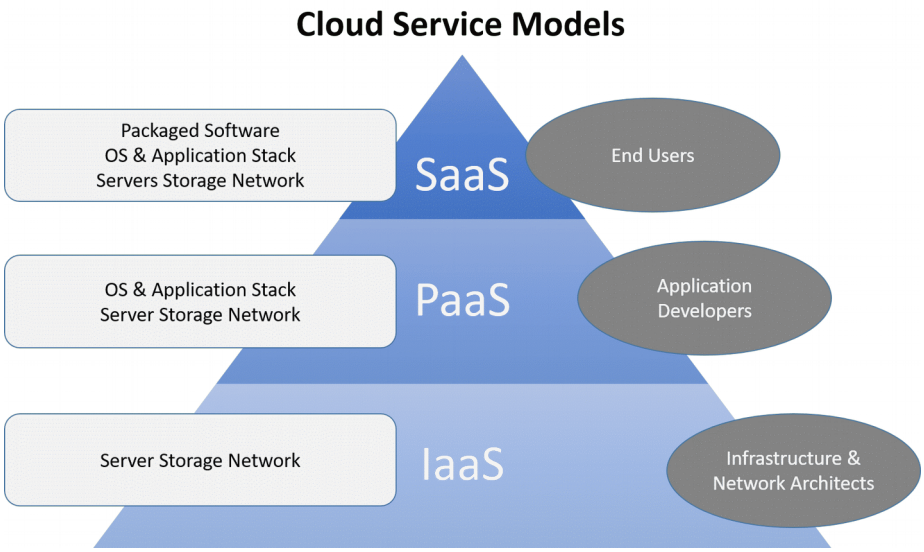


Εικόνα 20: Cloud Computing

- Χαμηλό κόστος:
  Το Cloud computing εξαλείφει το κόστος κεφαλαίου για την αγορά υλικού και λογισμικού και τη δημιουργία και λειτουργία δεδομένων και κέντρων δεδομένων.
- Ταχύτητα:
  Οι περισσότερες υπηρεσίες cloud computing παρέχονται με αυτοεξυπηρέτηση και κατ' απαίτηση, οπότε τα χρηματικά ποσά αυτών των υπολογιστικών πόρων, μπορούν να διατεθούν σε λίγα λεπτά, τυπικά με μερικά μόνο κλικ του ποντικιού, παρέχοντας στις επιχειρήσεις μεγάλη ευελιξία και δίνοντας τους τη δυνατότητα να ασκήσουν το σχεδιασμό χωρητικότητας όπως επιθυμούν την κάθε φορά.
- Παγκόσμια κλίμακα:
  Τα πλεονεκτήματα των υπηρεσιών Cloud computing περιλαμβάνουν τη δυνατότητα μιας ελαστικής κλίμακας. Αυτό σημαίνει ότι παρέχεται το σωστό ποσό πόρων πληροφορικής για την εκάστοτε ανάγκη (για παράδειγμα, περισσότερη ή λιγότερη υπολογιστική ισχύς, χώρος αποθήκευσης, εύρος ζώνης) μόνο όταν χρειάζεται και στην επιθυμητή επιλεγμένη γεωγραφική θέση.
- Παραγωγικότητα:
  Τα τοπικά datacenters απαιτούν συνήθως πολλές ρυθμίσεις "στοίβαξης", εγκατάστασης λογισμικού, επιδιόρθωσης λογισμικού και άλλες χρονοβόρες εργασίες διαχείρισης της πληροφορικής. Το Cloud computing εξαλείφει την ανάγκη για πολλά από αυτά τα καθήκοντα, έτσι ώστε οι ομάδες IT να μπορούν να μειώσουν τον χρόνο που χρειάζεται για την επίτευξη των πιο σημαντικών επιχειρηματικών στόχων τους.
- Απόδοση:
  Οι μεγαλύτερες υπηρεσίες cloud computing λειτουργούν σε ένα παγκόσμιο δίκτυο ασφαλών κέντρων δεδομένων, τα οποία αναβαθμίζονται τακτικά. Αυτό προσφέρει πολλά πλεονεκτήματα, συμπεριλαμβανομένης της μειωμένης καθυστέρησης του δικτύου ως προς τις εφαρμογές τους.
- Αξιοπιστία:
  Το cloud computing καθιστά ευκολότερη και λιγότερο δαπανηρή τη δημιουργία

αντιγράφων ασφαλείας των δεδομένων, την αποκατάσταση καταστροφών και την άμεση αποκατάσταση της λειτουργικότητας της επιχείρησης, επειδή τα δεδομένα μπορούν να αντικατοπτρίζονται σε πολλαπλές τοποθεσίες στο δίκτυο του παροχέα υπηρεσιών cloud.

1.6.3 Υπηρεσιακά Μοντέλα



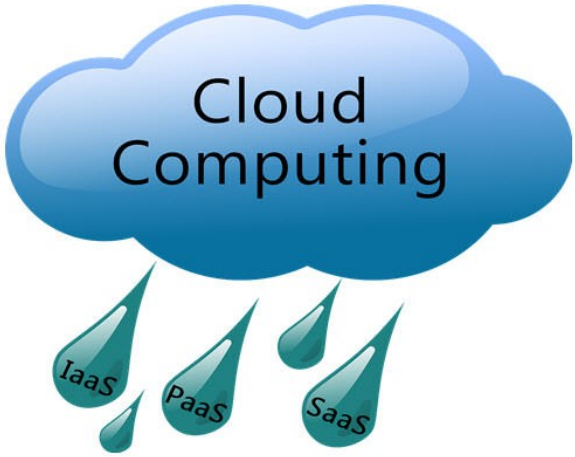
Εικόνα 21: Υπηρεσιακά Μοντέλα Cloud

α) SaaS: Software as Service
Το λογισμικό ως υπηρεσία (SaaS), είναι μια μέθοδος για την παράδοση εφαρμογών λογισμικού μέσω του διαδικτύου, κατ' απαίτηση και συνήθως με συνδρομή. Με το SaaS, οι πάροχοι Cloud φιλοξενούν και διαχειρίζονται την εφαρμογή λογισμικού και την υποκείμενη υποδομή και χειρίζονται οποιαδήποτε συντήρηση, όπως αναβαθμίσεις λογισμικού και ενημερωμένες εκδόσεις ασφαλείας. Οι χρήστες συνδέονται με την εφαρμογή μέσω του διαδικτύου, συνήθως με ένα πρόγραμμα περιήγησης στο τηλέφωνό τους, στο tablet τους ή στον υπολογιστή τους.

β) PaaS: Platform as Service
Η πλατφόρμα ως υπηρεσία (PaaS) αναφέρεται σε υπηρεσίες cloud computing που παρέχουν ένα περιβάλλον κατά παραγγελία για την ανάπτυξη, τον έλεγχο, την παράδοση και τη διαχείριση εφαρμογών λογισμικού. Το PaaS έχει σχεδιαστεί για να διευκολύνει τους προγραμματιστές να δημιουργούν γρήγορα εφαρμογές ιστού ή κινητής τηλεφωνίας χωρίς να ανησυχούν για τη δημιουργία ή τη διαχείριση της υποκείμενης υποδομής διακομιστών, χώρου αποθήκευσης, δικτύου και βάσεων δεδομένων που απαιτούνται για την ανάπτυξη.

γ) IaaS: Infrastructure as Service
Η πιο βασική κατηγορία υπηρεσιών cloud computing. Με το IaaS, μπορούμε να νοικιάσουμε υποδομές πληροφορικής (διακομιστές και εικονικές μηχανές (VM),

αποθηκευτικούς χώρους, δίκτυα, λειτουργικά συστήματα) από έναν πάροχο Cloud με βάση την υπηρεσία πληρωμών.



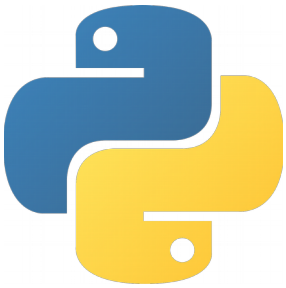
Εικόνα 22: Cloud Computing 2

ΚΕΦΑΛΑΙΟ 2: ΠΡΟΕΤΟΙΜΑΣΙΑ ΠΕΙΡΑΜΑΤΙΚΗΣ ΔΙΑΔΙΚΑΣΙΑΣ

2.1 Γλώσσες Προγραμματισμού

Για την επεξεργασία και ανάλυση των Big Data, υπάρχει ανάγκη επιλογής των κατάλληλων γλωσσών προγραμματισμού. Οι κυριότερες γλώσσες που χρησιμοποιούνται στα Big Data Analytics είναι οι R, Python και Scala, ενώ σε κάποιες εφαρμογές χρησιμοποιούνται οι Java, RegEx και Xpath.

2.1.1 Python



Εικόνα 23: Python Logo

Η Python είναι μια υψηλού επιπέδου γλώσσα προγραμματισμού η οποία δημιουργήθηκε το 1990. Ο κύριος στόχος της είναι η αναγνωσιμότητα του κώδικά της και η ευκολία χρήσης της. Το συντακτικό της επιτρέπει στους προγραμματιστές να εκφράσουν έννοιες σε λίγες γραμμές κώδικα. Διακρίνεται μεταξύ των άλλων γλωσσών λόγω του ότι διαθέτει πολλές βιβλιοθήκες. Ένα ακόμη αξιοσημείωτο χαρακτηριστικό της αποτελεί η ταχύτητα εκμάθησης της.

2.1.2 R

Εικόνα 24: R Logo

Η R είναι γλώσσα προγραμματισμού και περιβάλλον που παρέχει στον χρήστη τη δυνατότητα να κάνει υπολογιστική στατιστική και γραφήματα. Παρέχει ακόμη, τα απαραίτητα εργαλεία προκειμένου να υλοποιηθεί μια στατιστική ανάλυση. Μερικά απ' αυτά τα εργαλεία είναι τα εξής:

- δημιουργία τυχαίων δειγμάτων
- διακριτές και συνεχείς μεταβλητές (Poisson, Gamma, Exponential)
- έλεγχοι υποθέσεων
- στατιστικά τεστ (Kolmogorov-Smirnoff)
- δημιουργία γραφημάτων (ιστόγραμμα, qq plot, pie chart, bar chart)

2.1.3 Scala

Εικόνα 25: Scala Logo

Η Scala είναι μια γλώσσα προγραμματισμού πολλαπλών παραδειγμάτων που σχεδιάστηκε για να ενσωματώσει χαρακτηριστικά του αντικειμενοστραφούς και του συναρτησιακού προγραμματισμού. Το όνομα Scala προέρχεται από την αγγλική φράση "scalable language", που δηλώνει ότι έχει σχεδιαστεί για να μπορεί να αναπτύσσεται παράλληλα με τις ανάγκες των χρηστών της.

2.1.4 Java

Εικόνα 26: Java Logo

Η Java είναι μια αντικειμενοστραφής γλώσσα προγραμματισμού που σχεδιάστηκε από την εταιρεία πληροφορικής Sun Microsystems. Ένα από τα βασικά πλεονεκτήματά της έναντι των περισσότερων άλλων γλωσσών είναι η ανεξαρτησία του λειτουργικού συστήματος και της πλατφόρμας. Παρακάτω παρατίθενται συνοπτικά κάποια απ’ τα σημαντικότερα χαρακτηριστικά της Java:

- Απλή: Στόχος της ομάδας της Sun που ανέπτυξε την Java, ήταν μια γλώσσα εύκολη στη χρήση, που δεν απαιτεί πολλή εξάσκηση και εκπαίδευση. Οι περισσότεροι προγραμματιστές στις μέρες μας δουλεύουν είτε με τη C είτε με τη C++. Έτσι, μολονότι η C++ δεν ήταν η κατάλληλη για το αρχικό σχέδιο, η Java σχεδιάστηκε βάσει της C++, με σκοπό να γίνει όσο το δυνατόν περισσότερο κατανοητή. Η Java παραλείπει πολλά από τα σπανίως χρησιμοποιούμενα και δυσκολονόητα χαρακτηριστικά της C++, που δεν ωφελούν και πολύ την ευελιξία της γλώσσας. Προστέθηκαν δε διεργασίες, όπως η αυτόματη συλλογή των "σκουπιδιών" (automatic garbage collection), διευκολύνοντας τον προγραμματισμό.
  Μια κοινή πηγή πολυπλοκότητας της C++ και της C είναι η διαχείριση της μνήμης. Με την καινούργια διεργασία της αυτόματης συλλογής "σκουπιδιών", που συνιστάται από την περιοδική αποδέσμευση της μνήμης που δεν χρησιμοποιείται, μεγάλο μέρος από την δουλεία των προγραμματιστών αυτοματοποιείται και μειώνονται τα bugs. Ένα πλεονέκτημα της Java που οφείλεται στην απλότητα της είναι το μέγεθος των απαραίτητων εργαλείων. Ο Java interpreter και οι βασικές βιβλιοθήκες είναι μικρές και ο κώδικας της Java είναι τόσο περιορισμένος σε μέγεθος που μπορεί άνετα να τρέξει σε οιαδήποτε μικρή μηχανή και να κατέβει από το δίκτυο.
- Αντικειμενοστραφής: Λέγοντας ότι μία γλώσσα προγραμματισμού είναι αντικειμενοστραφής, εννοούμε ότι η τεχνική σχεδιασμού ενός προγράμματος συγκεντρώνεται σε αντικείμενα. Ένα αντικείμενο είναι ο συνδυασμός δεδομένων, διαδικασιών και λειτουργιών με βασική ιδιότητα την απόκρυψη του συνδυασμού αυτού. Το κάθε αντικείμενο, δηλαδή, αντιμετωπίζεται σαν ένα "μαύρο κουτί". Τα αντικείμενα δεν είναι ανεξάρτητα μεταξύ τους, αλλά βρίσκονται σε σχέση αλληλεξάρτησης με τα υπόλοιπα. Υπάρχει η έννοια της κληρονομικότητας μεταξύ των αντικειμένων, δηλαδή ένα αντικείμενο μπορεί να κληρονομήσει δεδομένα από άλλα αντικείμενα. Οι γλώσσες αντικειμενοστραφή προγραμματισμού είναι γλώσσες υψηλού επιπέδου, αφαιρετικές, αποτελεσματικές, γρήγορες και χρησιμοποιούνται για την δημιουργία μεγάλων και σημαντικών εφαρμογών.
- Συμβατή με Δίκτυα: Η Java έχει μια μεγάλη βιβλιοθήκη από ρουτίνες για την επιτυχημένη συνεργασία με τα πρωτόκολλα HTTP και FTP. Κατ' αυτόν τον τρόπο, οι δικτυακές συνδέσεις δημιουργούνται ευκολότερα από ότι με τη C ή τη C++. Τα προγράμματα σε Java μπορούν να έχουν πρόσβαση μέσω δικτύου σε αντικείμενα, με την ίδια άνεση που ένας χρήστης προσπελάζει ένα τοπικό σύστημα αρχείων.
- Σταθερή: Η Java προορίζεται για την σύνταξη προγραμμάτων που θα είναι αξιόπιστα από όλες τις πλευρές. Δίνεται έμφαση στον από νωρίς έλεγχο για πιθανά προβλήματα και στον έλεγχο σε πραγματικό χρόνο που αποσκοπεί στην εξάλειψη καταστάσεων που προκαλούν λάθη. Η μεγαλύτερη διαφορά μεταξύ Java και C/C++ είναι το γεγονός ότι η Java έχει ένα μοντέλο δεικτών που εξαφανίζει την πιθανότητα της επαναχρησιμοποίησης της μνήμης και την καταστροφή των δεδομένων. Αντί για αριθμητικούς δείκτες (pointer arithmetic), η Java έχει πραγματικούς πίνακες (true arrays). Οι προγραμματιστές της Java δεν έχουν να φοβηθούν την ακούσια (ή μη) τροποποίηση της μνήμης, γιατί δεν υπάρχουν δείκτες (pointers). Εξάλλου, τα προγράμματα σε Java δεν μπορούν να αποκτήσουν μη εγκεκριμένη πρόσβαση στην μνήμη.
- Ασφαλής: Η Java προορίζεται για χρήση σε ανοικτά, δικτυωμένα περιβάλλοντα. Γι' αυτό το λόγο, ιδιαίτερη προσοχή έχει δοθεί στην ασφάλεια που παρέχει η γλώσσα. Η Java επιτρέπει την κατασκευή προγραμμάτων ελεύθερων από ιούς των οποίων η

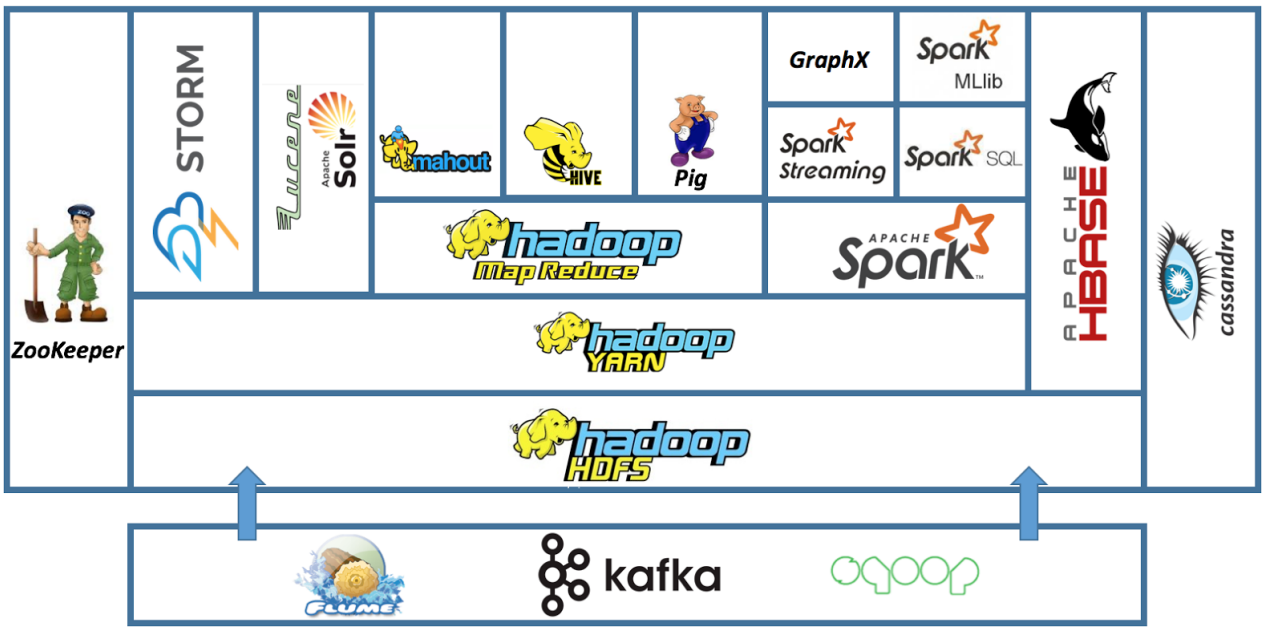
τροποποίηση είναι αδύνατη. Οι τεχνικές πιστοποίησης ταυτότητας βασίζονται στην ασύμμετρη κρυπτογραφία. Υπάρχει μεγάλη σχέση μεταξύ του τρόπου διαχείρισης της μνήμης και της παρεχόμενης ασφάλειας. Αλλαγές στην σημασιολογία των δεικτών της μνήμης κάνουν αδύνατη τη μη έγκυρη πρόσβαση στα δεδομένα της μνήμης. Με αυτόν τον τρόπο καταπολεμούνται οι περισσότεροι ιοί.

- Ουδέτερη της Υποκείμενης Αρχιτεκτονικής: Η Java έχει σχεδιαστεί για να υποστηρίζει δικτυακές εφαρμογές. Ένα δίκτυο, όμως, αποτελείται από ποικιλία διαφορετικών συστημάτων, με διαφορετικές CPU και λειτουργικά συστήματα. Για να μπορούν οι Java εφαρμογές να εκτελούνται παντού στο δίκτυο, το πρόγραμμα Java πρέπει να περάσει από δύο διαδικασίες ώστε να καταλήξει σε εκτελέσιμη μορφή. Πρώτα ο μεταγλωττιστής μετατρέπει τον πηγαίο κώδικα του προγράμματος σε μία ενδιάμεση γλώσσα που καλείται Java bytecodes. Τα Java bytecodes είναι ανεξάρτητα της πλατφόρμας και με χρήση του ερμηνευτή (interpreter) κάθε bytecode εντολή μετατρέπεται σε κατάλληλη δυαδική μορφή για να τρέξει στον εκάστοτε υπολογιστή. Η μεταγλώττιση (compilation) συμβαίνει μόνο μια φορά για κάθε Java πρόγραμμα ενώ η ερμηνεία (interpretation) γίνεται κάθε φορά που το πρόγραμμα εκτελείται. Τα Java bytecodes μπορούμε να τα φανταστούμε σαν τη γλώσσα μηχανής για την Java Virtual Machine (JVM). Κάθε Java ερμηνευτής (π.χ. ένας Web browser που μπορεί να τρέχει applets) είναι μια λογισμική εφαρμογή της Java Virtual Machine. Η JVM αναλαμβάνει να μετατρέψει τα bytecodes σε κατάλληλη εκτελέσιμη μορφή, ανάλογα με το υποκείμενο software και hardware. Η τεχνική που περιγράφηκε παραπάνω καλείται "write once, run anywhere". Το Java πρόγραμμα μεταγλωττίζεται μία φορά σε Java bytecodes με το μεταγλωττιστή της Java. Έπειτα, τα bytecodes μπορούν να τρέξουν σε οιαδήποτε μηχανή που έχει μία εφηρμοσμένη JVM.
- Φορητή: Σε αντίθεση με την C/C++ δεν υπάρχουν καθόλου χαρακτηριστικά που εξαρτώνται από την CPU του υπολογιστή. Έτσι, τα μεγέθη των πρωταρχικών τύπων δεδομένων είναι καθορισμένα και η συμπεριφορά τους είναι παντού η ίδια. Για παράδειγμα, "int" σημαίνει πάντα έναν 32 bit ακέραιο και "float" πάντα αντιπροσωπεύει έναν 32 bit floating αριθμό.
- Interpreted: Τα Java bytecodes μεταφράζονται σε πραγματικό χρόνο σε εντολές μηχανής που εξαρτώνται από την εκάστοτε πλατφόρμα, και δεν αποθηκεύονται πουθενά. Η διαδικασία είναι γρήγορη και αποτελεσματική. Μαζί με τα bytecodes μεταφέρονται πληροφορίες που μπορούν να χρησιμοποιηθούν κατά την εκτέλεση και παρέχουν τη βάση για τους ελέγχους που πραγματοποιεί ο linker. Επίσης τα προγράμματα γίνονται πιο επιδεκτικά σε debugging διαδικασίες.
- Υψηλής Απόδοσης: Η διαδικασία παραγωγής των εντολών μηχανής είναι απλή και γρήγορη. Ο κώδικας που προκύπτει είναι αποτελεσματικός. Ο μεταγλωττιστής από την μεριά του εφαρμόζει αυτόματη κατανομή των καταχωρητών (automatic register allocation) όταν παράγει τα bytecodes. Η τελική μορφή του κώδικα (εκτελέσιμη δυαδική μορφή) είναι μικρή σε μέγεθος και ταχύτατη στην εκτέλεση.
- Multithreaded: Τα προγράμματα σε Java έχουν την δυνατότητα να αντιμετωπίζουν πολλές καταστάσεις – διαδικασίες ταυτόχρονα. Αντίθετα, η C και C++ είναι single-threaded γλώσσες. Τα πλεονεκτήματα του multithreading είναι η καλύτερη αλληλεπιδραστική ανταπόκριση και η άμεση αντίδραση σε πραγματικό χρόνο.
- Δυναμική: Η Java θεωρείται δυναμικότερη γλώσσα από την C ή C++ καθώς έχει αναπτυχθεί για να προσαρμοστεί σε ένα εξελισσόμενο περιβάλλον. Οι βιβλιοθήκες εργαλείων αναπτύσσονται ελεύθερα με την πρόσθεση νέων μεθόδων και μεταβλητών, χωρίς να επηρεάζονται οι ήδη υπάρχουσες εφαρμογές.

2.2 Εργαλεία για Big Data Analytics & Data Mining

Όπως συμβαίνει σχεδόν πάντοτε σε τομείς χρήσης λογισμικού, υπάρχει και στη χρήση των Big Data η δυνατότητα επιλογής μεταξύ χρήσης ελεύθερου λογισμικού, αλλά και εμπορικών λύσεων, οι οποίες απαιτούν τη χρήση οικονομικών πόρων. Η πλατφόρμα που θα επιλεχθεί, σε κάθε περίπτωση, θα πρέπει να χειρίζεται την εισαγωγή, την επεξεργασία, την αποθήκευση και την αναζήτηση των δεδομένων, καθώς επίσης να παρέχει δυνατότητες ανάλυσής τους. Στο σημείο αυτό θα πραγματοποιηθεί μία παρουσίαση των βασικών επιλογών που παρέχονται, πλην της πλατφόρμας του Rapid Miner που θα αναλυθεί στο επόμενο εδάφιο.

2.2.1 Apache Hadoop



Εικόνα 27: Hadoop

Πρόκειται για ένα open-source software το οποίο χρησιμοποιείται για κατανεμημένη αποθήκευση και επεξεργασία Big Data χρησιμοποιώντας το μοντέλο Map-Reduce.

Το Hadoop αποτελείται από τα εξής δομικά στοιχεία:

• Το Hadoop Common Utilities που περιέχει βασικές βιβλιοθήκες και λειτουργίες που απαιτούνται από τα υπόλοιπα στοιχεία.

• Το Hadoop Distributed File System (HDFS) που διαχειρίζεται την αποθήκευση κατανεμημένων δεδομένων.

• Το Hadoop YARN Framework, το οποίο αποτελεί μία πλατφόρμα διαχείρισης πόρων και ουσιαστικά, είναι υπεύθυνο για τη διαχείριση των υπολογιστικών πόρων σε συστάδες και για τον προγραμματισμό των εφαρμογών των χρηστών.

• Το Hadoop Map-Reduce που αποτελεί υλοποίηση του μοντέλου Map-Reduce για κατανεμημένη επεξεργασία μεγάλης κλίμακας δεδομένων.

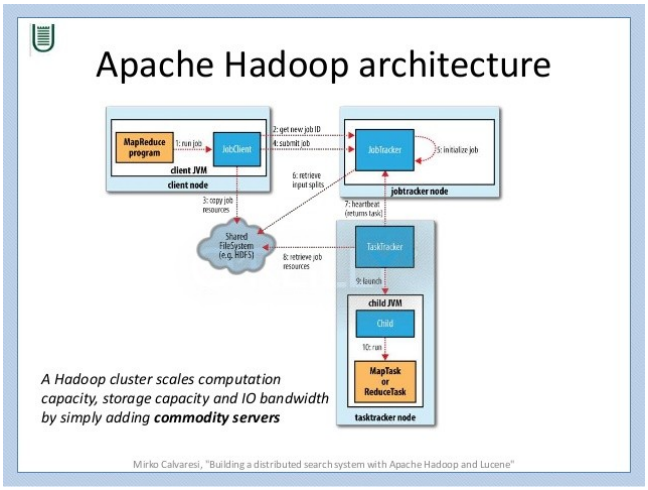
Ένα υπολογιστικό σύστημα που εκτελεί την εφαρμογή Hadoop αποτελείται από υπολογιστικές συστάδες (clusters) οι οποίες απαρτίζονται από εμπορικό υλικό (commodity hardware). Η δομή του Hadoop βασίζεται στην υπόθεση ότι οι αστοχίες υλικού (hardware failures), δηλαδή οι δυσλειτουργίες στα ηλεκτρονικά στοιχεία των υπολογιστικών συστημάτων - είναι συχνές κατά τη διαχείριση μεγάλου όγκου δεδομένων και οφείλει η ίδια η εφαρμογή να τις διαχειρίζεται αποδοτικά.

Ο πυρήνας του Hadoop αποτελείται από ένα τμήμα αποθήκευσης, γνωστό ως Hadoop Distributed File System (HDFS) και ένα τμήμα επεξεργασίας, που βασίζεται στο μοντέλο Map-Reduce που αναφέρθηκε προηγουμένως. Το Hadoop χωρίζει τα δεδομένα σε μεγάλα blocks και τα κατανέμει μεταξύ διαφόρων υπολογιστικών κόμβων που συνιστούν το υπολογιστικό σύστημα. Στη συνέχεια, μεταφέρει τον κώδικα που πρόκειται να εκτελεστεί στους κόμβους ώστε να πραγματοποιηθεί παράλληλη, δηλαδή ταυτόχρονη επεξεργασία των δεδομένων στους κόμβους αυτούς. Ουσιαστικά, διενεργείται αξιοποίηση της ιδιότητας data locality (τοπικότητα) και οι κόμβοι διαχειρίζονται τα επιμέρους δεδομένα στα οποία έχουν πρόσβαση.

Η χρήση της παραλληλίας αποτελεί κλασσική προσέγγιση βελτίωσης της αποδοτικότητας εφαρμογών λογισμικού. Μάλιστα έχει αποδειχθεί πως η χρήση της τοπικότητας των δεδομένων κατ’ αυτόν τον τρόπο από εμπορικά συστήματα, παρέχει καλύτερα αποτελέσματα από αυτά που προσφέρουν εξελιγμένοι υπερυπολογιστές (supercomputers), οι οποίοι βασίζονται σε παράλληλα συστήματα αρχείων όπου ο υπολογισμός και τα δεδομένα διαμοιράζονται μέσω υψηλής ταχύτητας δικτύου. Το μοντέλο Map-Reduce είναι ένα μοντέλο προγραμματισμού αποτελούμενο από μια συνάρτηση απεικόνισης (Map) η οποία φιλτράρει και διατάσσει τα δεδομένα στους κόμβους και μία συνάρτηση μείωσης (Reduce), η οποία πραγματοποιεί υπολογισμούς καταμέτρησης στους υπολογιστικούς κόμβους.

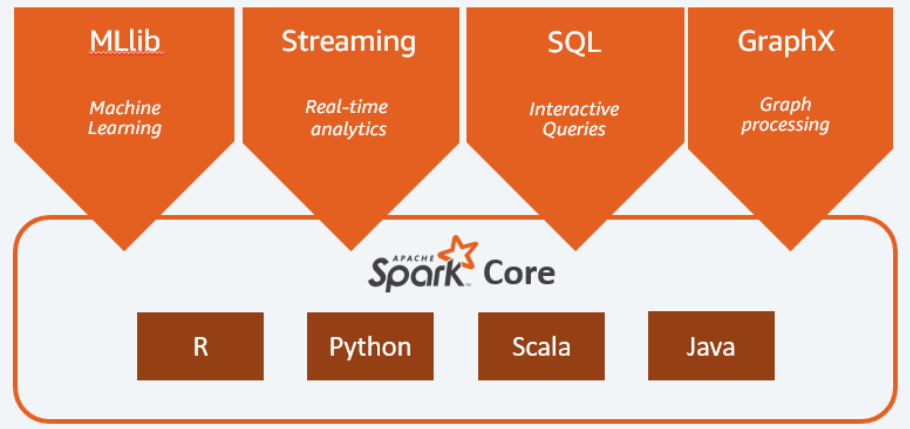
Κλασσικό παράδειγμα χρήσης του μοντέλου Map-Reduce αποτελεί ένα πρόγραμμα μέτρησης λέξεων σε ένα σύνολο αρχείων. Στην περίπτωση αυτή, η συνάρτηση Map απεικονίζει το σύνολο των αρχείων τμηματικά σε κάθε υπολογιστικό κόμβο και η συνάρτηση Reduce μετράει το πλήθος των λέξεων σε κάθε υπολογιστικό κόμβο, λαμβάνοντάς το ζητούμενο αθροιστικά.

Τέλος, αξίζει να σημειωθεί πως αν και η βασική δομή του Hadoop συνίσταται από τα στοιχεία που ήδη αναφέρθηκαν, συχνά χρησιμοποιούνται επεκτάσεις από την Apache που εμπλουτίζουν τις δυνατότητες του Hadoop, αναλόγως την περίσταση, οι σημαντικότερες από τις οποίες είναι: Apache HBase, Apache Pig, Apache Hive, Apache Phoenix, Apache Spark, Apache ZooKeeper, Apache Flume, Apache Sqoop, Apache Storm.



Εικόνα 28: Hadoop Architecture

2.2.2 Apache Spark



Εικόνα 29: Spark

Αποτελεί επίσης λογισμικό ελεύθερου κώδικα για επεξεργασία Big Data. Δημιουργήθηκε αρχικά στο Πανεπιστήμιο Berkeley, της California και στη συνέχεια παραχωρήθηκε αφιλοκερδώς στην Apache Software Foundation. Είναι μεταγενέστερο του Hadoop και ουσιαστικά προσφέρει κάποια πλεονεκτήματα, τα οποία θα αναλυθούν εν συνέχεια. Το Spark προσφέρει στον προγραμματιστή μία διεπαφή (Interface) επικεντρωμένη σε μία δομή δεδομένων, γνωστή ως Resilient Distributed Dataset ή RDD. Πρόκειται για μια συλλογή κατανεμημένων αντικειμένων σε ένα σύνολο υπολογιστικών κόμβων η οποία διασφαλίζει αποτελεσματική διαχείριση αστοχιών υλικού, όπως ακριβώς το Hadoop.

Ο λόγος δημιουργίας του Spark είναι ορισμένοι δομικοί περιορισμοί που επιβάλλονται από το μοντέλο Map-Reduce του Hadoop. Συγκεκριμένα, απαιτείται γραμμική ροή δεδομένων ως είσοδος από το δίσκο σε κατανεμημένα συστήματα, κατάλληλη επεξεργασία σύμφωνα

με τις συναρτήσεις Map και Reduce και τέλος, γραμμικού χρόνου αποθήκευση των δεδομένων στο δίσκο. Αντίθετα, το Spark παρέχει τη δυνατότητα πραγματοποίησης των υπολογισμών σε διαμοιραζόμενη μνήμη, όπου η ταχύτητα είναι σημαντικά μεγαλύτερη από την αντίστοιχη στο δίσκο. Με τον τρόπο αυτό, καθίσταται δυνατή η εφαρμογή επαναληπτικών αλγορίθμων που πραγματοποιούν πολλαπλές φορές πρόσβαση στα δεδομένα σε κάθε επανάληψη, χωρίς αυτό να συμβαίνει εις βάρος του χρόνου υπολογισμού, αφού ο χρόνος πρόσβασης σε δεδομένα μνήμης είναι ταχύτερος. Σύμφωνα μάλιστα με πληροφορίες που βρίσκονται στην επίσημη ιστοσελίδα του Spark, είναι δυνατόν να εκτελεστούν εφαρμογές έως και 100 φορές ταχύτερα στη μνήμη και έως 10 φορές ταχύτερα στο δίσκο, εν συγκρίσει με το Hadoop.

Το Apache Spark πέραν του τρόπου διαχείρισης των Big Data, προσφέρει τις εξής βασικές επεκτάσεις:

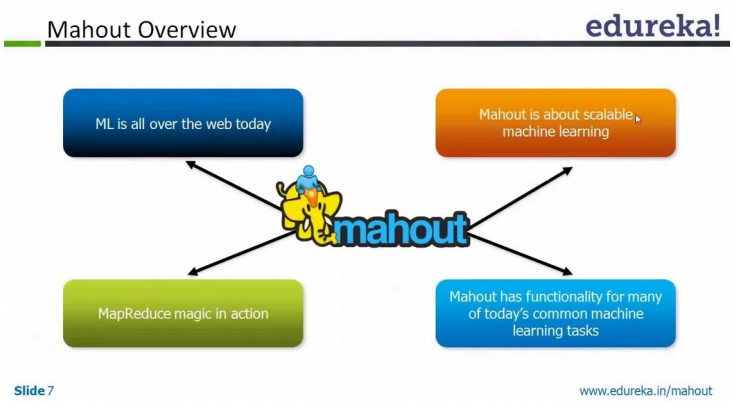
• Spark SQL: Επιτρέπει ερωτήματα (queries) σε δεδομένα με χρήση SQL, σε συνδυασμό με τις γλώσσες προγραμματισμού Java, Scala, Python και R.

• Park Streaming: Καθιστά εφικτή την επεξεργασία δεδομένων που εισέρχονται στο σύστημα ενώ βρίσκονται ήδη σε εξέλιξη υπολογισμοί στα προηγούμενα δεδομένα. Αυτό το χαρακτηριστικό είναι πολύ σημαντικό, καθώς στο Hadoop δεν μπορούν να προστίθενται νέα δεδομένα κατά τη διάρκεια της επεξεργασίας. Υποστηρίζονται οι γλώσσες προγραμματισμού Java, Scala και Python.

• MLlib: Πρόκειται για μία βιβλιοθήκη μηχανικής μάθησης (Machine Learning Library) η οποία δίνει τη δυνατότητα εκτέλεσης αλγορίθμων αυτού του είδους έως και 100 φορές ταχύτερα από το Hadoop.

• GraphX: Παρέχει ένα API (Application Programming Interface) για τα δεδομένα σε μορφή γραφημάτων, επιτρέποντας μάλιστα υπολογισμούς με χρήση επαναληπτικών αλγορίθμων με αποδοτικό τρόπο.

2.2.3 Apache Mahout



Εικόνα 30: Mahout

Το Apache Mahout είναι ένα έργο του Apache Software Foundation για την παραγωγή δωρεάν εφαρμογών κατανεμημένων ή άλλως κλιμακούμενων αλγορίθμων μηχανικής

μάθησης που εστιάζονται κυρίως στους τομείς φιλτραρίσματος, ομαδοποίησης και ταξινόμησης.

Πολλές από τις εφαρμογές του Apache Mahout χρησιμοποιούν την πλατφόρμα Apache Hadoop. Το Mahout παρέχει επίσης βιβλιοθήκες Java για κοινές λειτουργίες μαθηματικών (που εστιάζονται στη γραμμική άλγεβρα και στα στατιστικά στοιχεία) και τις πρωτότυπες συλλογές Java.

Πρέπει να σημειωθεί ότι τω εν προκειμένω έργο βρίσκεται σε εξέλιξη˙ ο αριθμός των εφαρμοσμένων αλγορίθμων έχει αυξηθεί ταχύτατα, ωστόσο εξακολουθούν να υπάρχουν σοβαρές ελλείψεις.

2.2.4 Talend

Εικόνα 31: Talend

Το Talend είναι μια πλατφόρμα ανοικτού κώδικα, βασισμένη στο μοντέλο Hadoop που προσφέρει μια σειρά προϊόντων διαχείρισης Big Data. Το βασικότερο στοιχείο είναι το Master Data Management (MDM), το οποίο έχει τη δυνατότητα να επεξεργάζεται δεδομένα σε πραγματικό χρόνο, να αξιοποιεί άλλες εφαρμογές, να ενσωματώνει τα δεδομένα τους και να εκτελεί διάφορες διαδικασίες, όπως εκτιμήσεις της ποιότητας των Big Data. Παρέχεται δωρεάν και προσφέρει αρκετές δυνατότητες πράγμα το οποίο το καθιστά καλή επιλογή για την κάλυψη πολλών περιπτώσεων.

2.2.5 Cassandra

Εικόνα 32: Cassandra Logo

Το Apache Cassandra είναι ένα εξαιρετικά ισχυρό σύστημα κατανεμημένων βάσεων δεδομένων ανοιχτού κώδικα που λειτουργεί εξαιρετικά καλά για να χειρίζεται τεράστιους όγκους αρχείων που διανέμονται σε πολλούς διακομιστές βασικών προϊόντων. Είναι μια από τις πιο αποτελεσματικές βάσεις δεδομένων NoSQL που είναι διαθέσιμες σήμερα στην αγορά.

Χαρακτηριστικά:

- Είναι μια column- oriented βάση δεδομένων.
- Είναι μία εξαιρετικά συνεπής πλατφόρμα, ανεκτική σε σφάλματα και κλιμακωτή.
- Δημιουργήθηκε στο Facebook και αργότερα έγινε open source.
- Το μοντέλο δεδομένων βασίζεται στο Google Bigtable.
- Ο κατανεμημένος σχεδιασμός βασίζεται στο Amazon Dynamo.

2.2.6 Orange

Εικόνα 33: Orange Logo

Η Orange είναι μία βιβλιοθήκη αντικειμένων πυρήνα και ρουτινών της C++ και περιλαμβάνει μία μεγάλη ποικιλία αλγορίθμων Machine Learning & Data Mining ενώ παράλληλα περιέχει ρουτίνες για εισαγωγή και χειρισμό δεδομένων. Επίσης, με τη βοήθεια μιας συλλογής από modules της Python το περιβάλλον της επιτρέπει τη δημιουργία κώδικα για γρήγορη προτυποποίηση νέων αλγορίθμων και έλεγχο συστημάτων.

Επιπροσθέτως, η Orange συμπεριλαμβάνει ένα μεγάλο αριθμό widget (συστατικά) γραφικών που χρησιμοποιούν μεθόδους της βιβλιοθήκης πυρήνα και των υπό-προγραμμάτων (modules) της Orange. Με τη χρήση του οπτικού προγραμματισμού, τα widget μπορούν να συγκεντρωθούν σε μία εφαρμογή με ένα εργαλείο οπτικού προγραμματισμού που ονομάζεται Orange Canvas.

Όλα αυτά μαζί καθιστούν την Orange, ένα περιεκτικό, βασισμένο σε συστατικά πλαίσιο για μηχανική εκμάθηση και εξόρυξη δεδομένων, προορισμένο για έμπειρους χρήστες και ερευνητές στη μηχανική εκμάθηση που θέλουν να αναπτύξουν τους δικούς τους αλγόριθμους χρησιμοποιώντας όσο το δυνατόν περισσότερο κώδικα, αλλά και για αρχάριους, που έχουν τη δυνατότητα να χρησιμοποιήσουν ένα παντοδύναμο, αλλά ταυτόχρονα εύκολο στη χρήση περιβάλλον οπτικού προγραμματισμού.

Η Orange παρέχει ένα ποικιλόμορφο περιβάλλον για τους υπεύθυνους ανάπτυξης, τους ερευνητές αλλά και για όσους επιθυμούν να εντρυφήσουν στην εξόρυξη δεδομένων. Χάρη στην Python, μία γλώσσα συγγραφής σεναρίων (scripting language) νέας γενιάς και περιβάλλοντος προγραμματισμού, τα σενάρια (scripts) για την εξόρυξη δεδομένων είναι μεν απλά, αλλά ισχυρά.

Για ακόμη πιο γρήγορη προτυποποίηση, η Orange υιοθετεί μία προσέγγιση βασισμένη σε συστατικά: η μέθοδος ανάλυσης μπορεί να υλοποιηθεί με τη χρήση ενός υπάρχοντος αλγορίθμου και αντικατάσταση κάποιων βασικών συστατικών του με άλλα, καινούργια.

Ότι είναι τα συστατικά της Orange για τη συγγραφή σεναρίων, είναι και τα widgets της Orange για τον οπτικό προγραμματισμό. Τα widgets χρησιμοποιούν ένα ειδικά σχεδιασμένο μηχανισμό επικοινωνίας για μεταβαλλόμενα αντικείμενα όπως σύνολα δεδομένων, λίστες χαρακτηριστικών, τεχνικές εκμάθησης, ταξινομητές, και άλλα, επιτρέποντας την εύκολη κατασκευή αρκετά πολύπλοκων σχημάτων εξόρυξης δεδομένων, χρησιμοποιώντας προσεγγίσεις και τεχνικές τελευταίας τεχνολογίας. Η βασική αρχή της Orange δεν είναι να καλύψει κάθε μέθοδο και άποψη στη μηχανική εκμάθηση και εξόρυξη δεδομένων, αλλά να καλύψει σε βάθος και σχολαστικά αυτές που υλοποιούνται, δημιουργώντας τες από επαναχρησιμοποιήσιμα συστατικά τα οποία έμπειροι χρήστες μπορούν να αλλάξουν ή και να αντικαταστήσουν με καινούργια.

2.2.7 Rattle

Το Rattle είναι ένα πακέτο λογισμικού ανοιχτού κώδικα που παρέχει μία γραφική διεπιφάνεια χρήστη για εξόρυξη δεδομένων με χρήση της προγραμματιστικής στατιστικής γλώσσας R. Σχεδιάστηκε ειδικά για να διευκολύνει τη μετάβαση από την απλή και βασική εξόρυξη δεδομένων, που υπάρχει απαραίτητα στις διεπιφάνειες χρήστη, στην εξελιγμένη ανάλυση δεδομένων, χρησιμοποιώντας μία ισχυρή στατιστική γλώσσα.

Το Rattle ενώνει μία πλειάδα πακέτων R τα οποία είναι απαραίτητα για κάποιον που ασχολείται με τον τομέα της εξόρυξης δεδομένων, αλλά συχνά δύσκολα προς χρήση για έναν αρχάριο. Δεν είναι απαραίτητη η κατανόηση της R για να ξεκινήσει κανείς να χρησιμοποιεί το Rattle˙ αυτό θα γίνει σιγά-σιγά, με την ολοένα και αυξανόμενη επιτήδευση στην εκτέλεση έργων εξόρυξης δεδομένων.

Η διεπιφάνεια χρήστη του Rattle παρέχει μία πρώτη εικόνα στη δύναμη της R ως εργαλείο για το Data Mining. Το Rattle χρησιμοποιείται ως λογισμικό εκμάθησης της γλώσσας R. Υπάρχει μία καρτέλα καταγραφής κώδικα (Log Code tab), η οποία αναπαράγει τον κώδικα R που χρησιμοποιήθηκε για οποιαδήποτε δραστηριότητα στη διεπιφάνεια χρήστη και η οποία μπορεί να αντιγραφεί και να επικολληθεί. Επίσης, το Rattle δύναται να χρησιμοποιηθεί για στατιστική ανάλυση ή παραγωγή μοντέλων. Επιτρέπει στο σύνολο δεδομένων να χωριστεί σε δεδομένα εκπαίδευσης, διασταύρωσης και ελέγχου. Το σύνολο δεδομένων εν συνεχεία μπορεί να προβληθεί και να τεθεί υπό επεξεργασία.

Το Rattle βασίζεται σε μία εκτεταμένη συλλογή πακέτων ανάπτυξης R, γεγονός το οποίο αποτελεί και μία ακόμη απόδειξη της δύναμης της R. Μερικά από αυτά είναι τα εξής: ada, arules, doBy, ellipse, fBasics, fpc, gplots, Hmisc, kernlab, mice, network, party, playwith, pmml, random Forest, reshape, rggobi, RGtk2, ROCR, RODBC και rpart. Τα παραπάνω είναι διαθέσιμα από το CRAN (Comprehensive R Archive Network).

Η Rattle χρησιμοποιεί τη γραφική διεπιφάνεια χρήστη Gnome, όπως αυτή παρέχεται από το πακέτο RGtk2. Τρέχει δε σε όλα τα συστήματα, συμπεριλαμβανομένων των GNU/Linux, Macintosh OS/X και MS/Windows.

2.3 Rapid Miner

2.3.1 Εισαγωγή

Το RapidMiner είναι μια open source πλατφόρμα λογισμικού για Data Analytics που υλοποιήθηκε το 2001 από τους Ralf Klinkenberg, Ingo Mierswa και Simon Fischer στο Τμήμα Τεχνητής Νοημοσύνης του πανεπιστημίου του Dortmund και παρέχει ένα ολοκληρωμένο περιβάλλον για την προετοιμασία δεδομένων, τη μηχανική μάθηση, την εξόρυξη δεδομένων και την ανάλυση προγνωστικών. Χρησιμοποιείται για επιχειρηματικές και εμπορικές εφαρμογές καθώς και για έρευνα, εκπαίδευση, κατάρτιση, ταχεία ανάπτυξη πρωτοτύπων και ανάπτυξη εφαρμογών και υποστηρίζει όλα τα στάδια της διαδικασίας μηχανικής μάθησης, συμπεριλαμβανομένης της προετοιμασίας δεδομένων, της

απεικόνισης των αποτελεσμάτων, της επικύρωσης του μοντέλου και της βελτιστοποίησης. Η γλώσσα προγραμματισμού βάσει της οποίας λειτουργεί το συγκεκριμένο λογισμικό είναι η Java.

Εκτός από CSV το Rapid Miner δέχεται ένα σύνολο από τύπους αρχείων όπως PDF, XML, JSON, XLS. Υπάρχουν οδηγοί (Wizards) που διευκολύνουν την εισαγωγή δεδομένων από γνωστούς τύπους αρχείων. Μια δεύτερη εναλλακτική πηγή δεδομένων μπορεί να είναι μια βάση δεδομένων και οι πίνακες της. Είναι εφικτό να συνδέσουμε το Rapid Miner με όλες τις γνωστές σχεσιακές βάσεις δεδομένων Mysql, Microsoft Access, IBM DB2, Postgres, Oracle και αρκετές άλλες. Υπάρχουν οι κατάλληλοι οδηγοί ώστε να δημιουργήσουμε μια σύνδεση με την εκάστοτε βάση δεδομένων είτε πρόκειται για τοπική εγκατάσταση είτε για πρόσβαση σε κάποια απομακρυσμένη βάση. Τέλος το Rapid Miner μας δίνει την δυνατότητα να πάρουμε δεδομένα και από μη σχεσιακές βάσεις δεδομένων (NoSQL). Τα πιο δημοφιλή παραδείγματα NoSQL συστημάτων που μπορούμε να συνδέσουμε είναι τα MongoDB και Cassandra.

Είναι σημαντικό να αναφέρουμε ότι με την προσθήκη ενός Extension δίδεται η δυνατότητα επεξεργασίας και τελικά δημιουργίας δεδομένων από HTML που μπορούμε να πάρουμε ως απάντηση σε ένα HTTP ερώτημα. Το Extension Web Mining είναι αυτό που επεκτείνει τη λειτουργικότητα του Rapid Miner με σκοπό να μπορούμε να έχουμε πρόσβαση μεταξύ άλλων σε RSS Feed, HTML, API και Web Services.

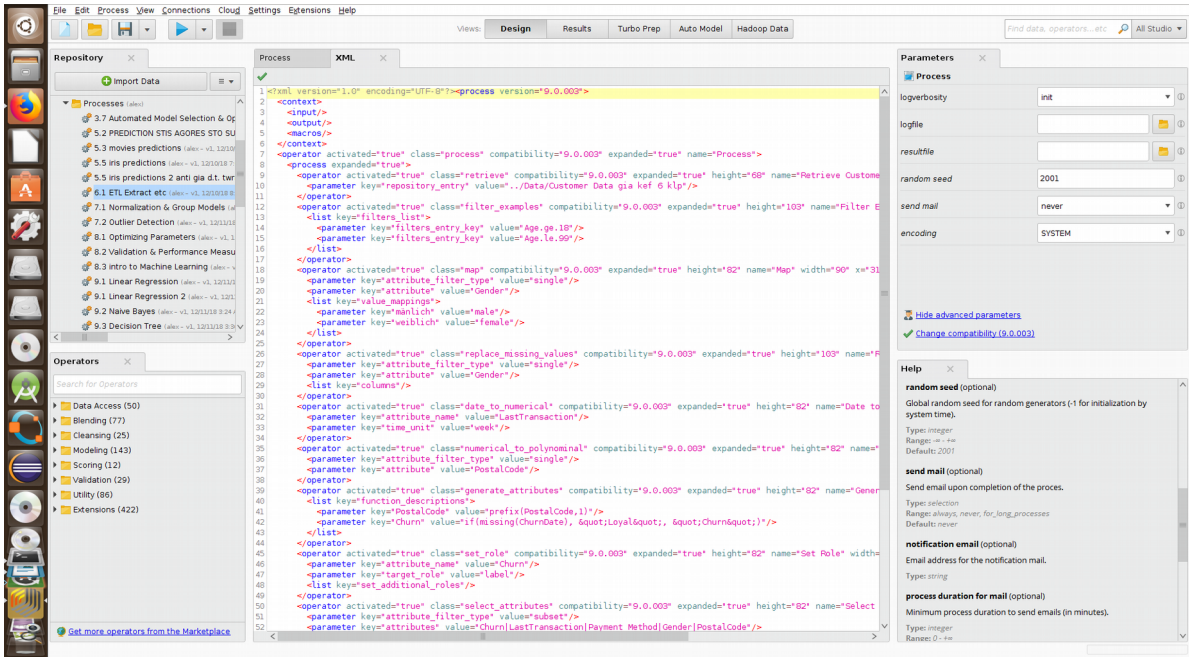
Τα δεδομένα παραδειγμάτων ή στιγμιότυπων στο RapidMiner περιγράφονται με τη χρήση ενός XML εγγράφου. Αυτό το αρχείο περιγραφής των χαρακτηριστικών περιέχει πληροφορίες για τον τύπο των δεδομένων και την πηγή τους. Τα σύνολα δεδομένων μπορούν να διανεμηθούν σε αρκετά αρχεία. Αυτό είναι ιδιαίτερα χρήσιμο αν η ετικέτα είναι αποθηκευμένη σε δικό της αρχείο. Στην περίπτωση που κάποιος δε θέλει να χρησιμοποιήσει τη βασική μορφή δεδομένων (XML) του RapidMiner, δίνεται η δυνατότητα χρήσης κάποιου τελεστή ειδικής μορφής, που μπορεί να διαβάσει αρχεία Arff, csv (comma separated value), bibtex, dBase, C4.5 και πολλά άλλα.

Το RapidMiner χρησιμοποιεί ένα μοντέλο πελάτη / εξυπηρετητή με τον Server να προσφέρεται είτε ως δημόσια είτε ως ιδιωτική υποδομή cloud.

Διαθέτει τρεις εκδόσεις (ανάλογα με το μέγεθος των δεδομένων που θέλουμε να αναλύσουμε και των αριθμό των χρηστών που συμμετέχουν στο project):

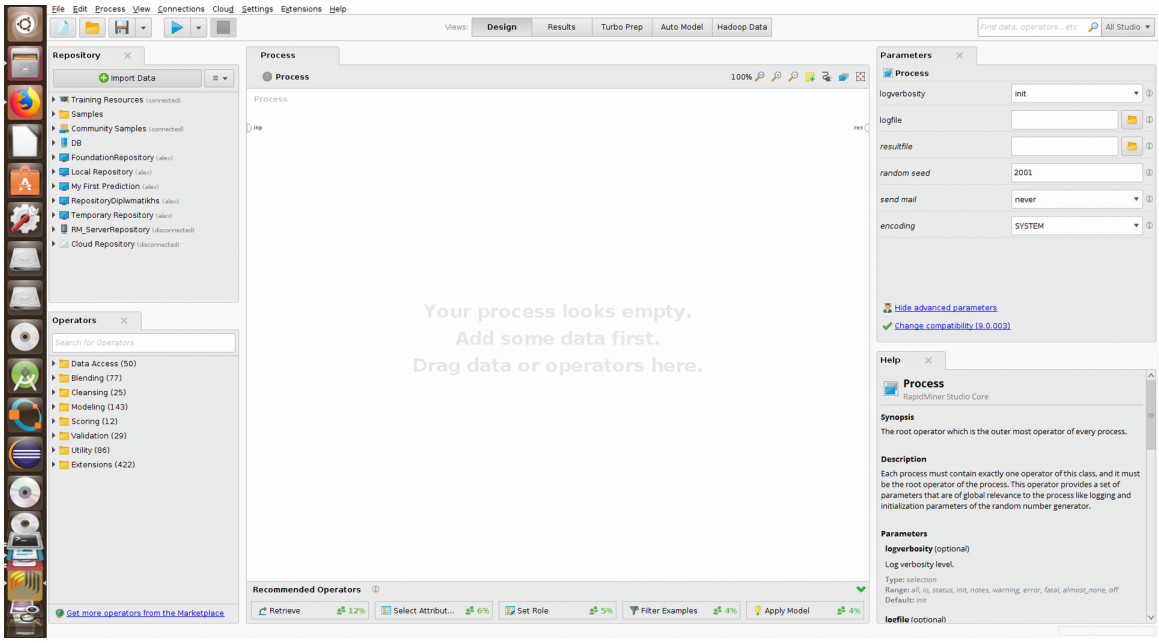
- **Rapid Miner Studio**:

  α) Μπορεί να εγκατασταθεί σε προσωπικό υπολογιστή.
  β) Παρέχει ένα GUI για σχεδίαση και εκτέλεση αναλυτικών ροών εργασίας οι οποίες καλούνται Processes και αποτελούνται από πολλαπλούς Operators. Η χρήση των Operators για την επεξεργασία των δεδομένων καθιστά τη γραφή κώδικα μη απαραίτητη.
- **Rapid Miner Server**: εγκαθίσταται και λειτουργεί σε Server και είναι ικανή να δεχθεί παράλληλη επεξεργασία δεδομένων από πολλούς χρήστες. Τα δεδομένα δύνανται να έχουν μεγαλύτερο μέγεθος απ' ότι στην πρώτη έκδοση.
- **Rapid Miner Radoop**: επεξεργάζεται δεδομένα ενός Cluster υπολογιστών. Σε αυτήν την περίπτωση το Rapid Miner εγκαθίσταται σε ένα σύστημα Hadoop και μπορεί να εκτελέσει ανάλυση πολύ μεγάλων δεδομένων.



Εικόνα 34: XML

2.3.2 Rapid Miner Studio



Εικόνα 35: Rapid Miner Studio

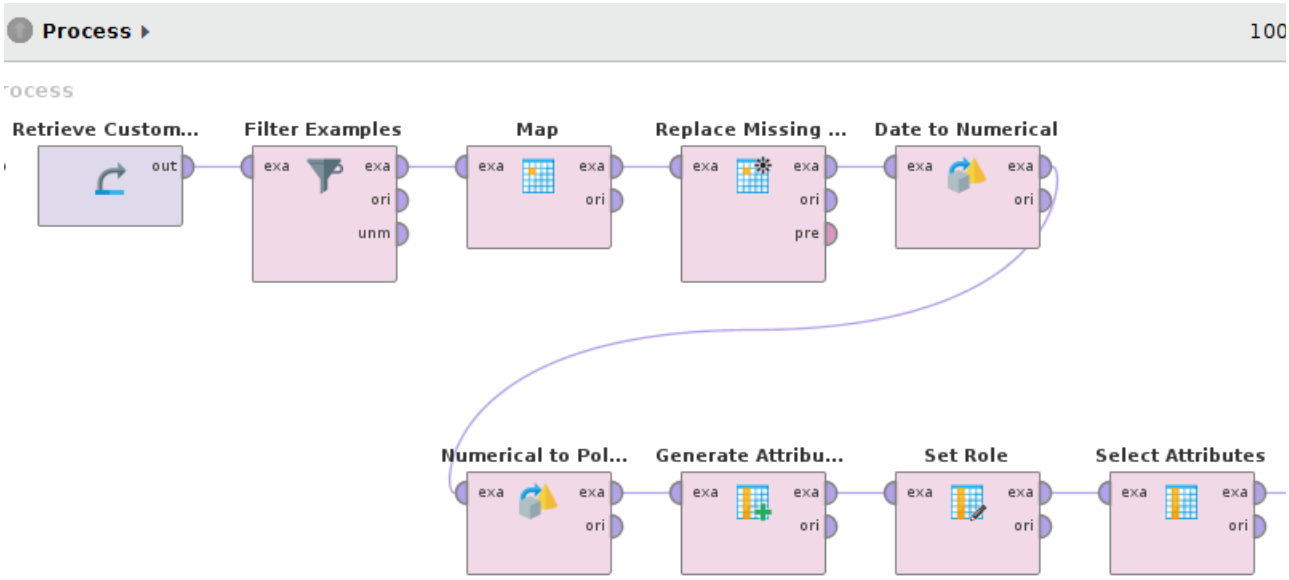
Στην παραπάνω εικόνα παρουσιάζεται το περιβάλλον του RapidMiner. Παρατηρούμε τα εξής στοιχεία:

- Στο πάνω μέρος υπάρχει μία μπάρα όπου μπορούμε να επιλέξουμε ανάμεσα στο Design (Σχεδίαση), τα Results (Αποτελέσματα), το Turbo Prep, το Auto Model και το Hadoop Data.
- Πάνω, αριστερά βρίσκεται το Repository όπου έχουμε τη δυνατότητα αποθήκευσης δεδομένων και διεργασιών.
- Ακριβώς από κάτω συναντάμε τους Operators οι οποίοι όπως βλέπουμε είναι ταξινομημένοι σε 7 κατηγορίες˙ Data Access (Πρόσβαση σε Δεδομένα), Blending (Μετασχηματισμός Δεδομένων), Cleansing (Καθαρισμός Δεδομένων), Modeling (Μοντελοποίηση), Scoring (Αξιολόγηση), Validation (Επικύρωση), Utility (Χρησιμότητα) συν την κατηγορία των Extensions (τα κατεβάζουμε απ' το Rapid Miner Marketplace). Μέσα στους ομώνυμους φακέλους μπορούμε να εντοπίσουμε και να επιλέξουμε κάθε φορά τον κατάλληλο και να τον σύρουμε κεντρικά στην επιφάνεια σχεδιασμού με Drag and Drop. Κάθε Operator εκτελεί μία μόνο εργασία μέσα στο Process και η έξοδος του αποτελεί την είσοδο για τον επόμενο.
- Στην περιοχή στο κέντρο γίνεται η σχεδίαση του Process ενώ στο κάτω μέρος η πλατφόρμα μας προτείνει πιθανούς Operators.
- Πάνω, δεξιά γίνεται η παραμετροποίηση για κάθε επιλεγμένο Operator.
- Τέλος, για να εκτελέσουμε το Process πατάμε το play. Αν επιθυμούμε η εκτέλεση να γίνει στο Server, πηγαίνουμε στο βελάκι και επιλέγουμε "Run Process on Server". Όταν η διαδικασία φτάσει στο τέλος, τα αποτελέσματα εμφανίζονται αυτόματα. Αυτό μπορεί να γίνει με στατιστική απόδοση, με δένδρο απόφασης και με πολλούς άλλους τρόπους.

Το Rapid Miner επιλέγει αυτόματα τη λειτουργία εμφάνισης αποτελεσμάτων (Results Mode). Αν θέλουμε να εξετάσουμε μια διαδικασία πιο προσεκτικά, μπορούμε να τοποθετήσουμε σημεία παύσης πριν και μετά από κάθε Operator. Σε αυτή τη περίπτωση, κάθε φορά που ο έλεγχος φτάνει σε σημείο παύσης, παρουσιάζονται ενδιάμεσα αποτελέσματα, κατά όμοιο τρόπο με το pop-up μήνυμα στο τέλος της διαδικασίας. Μπορεί κανείς επίσης να δει και να μελετήσει το διάγραμμα χρήσης της μνήμης και τη μπάρα προόδου.

Εικόνα 36: Rapid Miner Market Place

2.3.3 Rapid Miner Operators

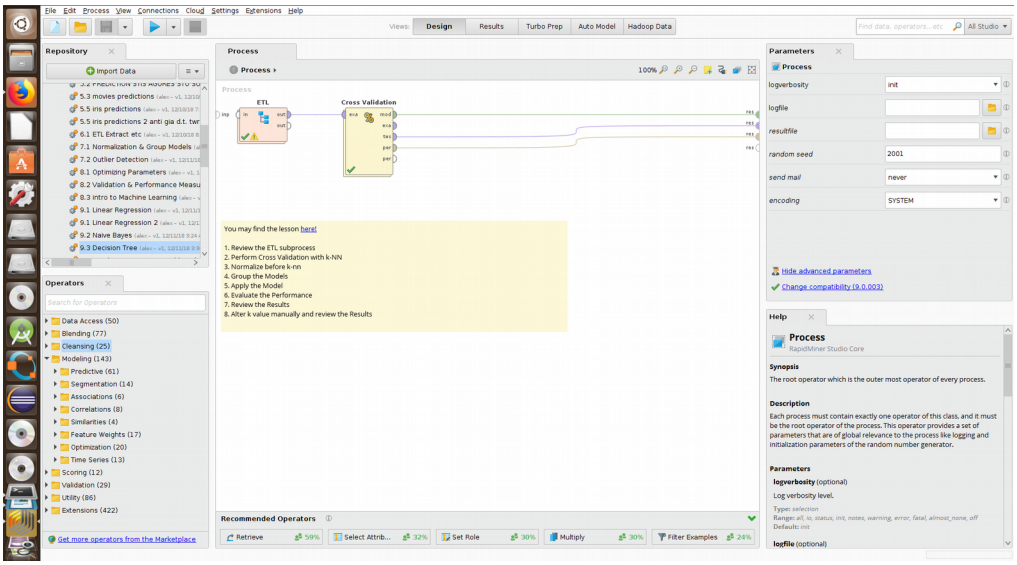


Εικόνα 37: Rapid Miner Process & Operators

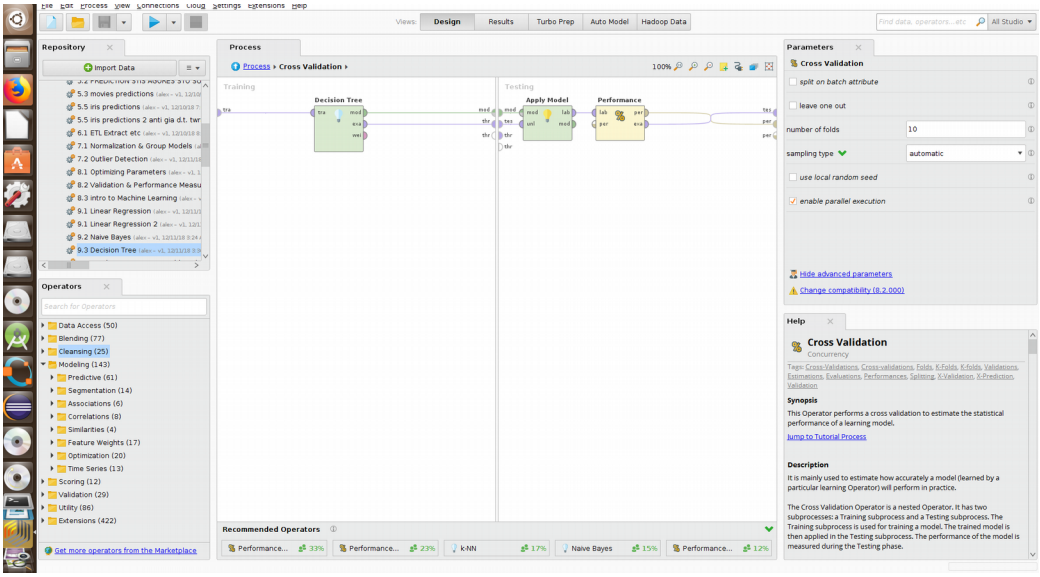
Το RapidMiner παρέχει πάνω από 400 Operators. Παρακάτω παραθέτουμε μερικούς εξ αυτών:

- **Αλγόριθμοι μηχανικής εκμάθησης:** ένας τεράστιος αριθμός σχημάτων εκμάθησης για έργα παλινδρόμησης και κατηγοριοποίησης, συμπεριλαμβανομένων των SVM, των Decision Trees και αλγόριθμων εκμάθησης κανόνων, των αλγορίθμων οκνηρής εκμάθησης, Bayesian εκμάθησης και λογιστικής εκμάθησης. Αρκετοί αλγόριθμοι για εξόρυξη κανόνων συσχέτισης και συσταδοποίησης περιλαμβάνονται στο RapidMiner. Επιπλέον, έχουν προστεθεί αρκετά σχήματα μετά-εκμάθησης, συμπεριλαμβανομένου και του Bayesian Boosting.
- **Τελεστές "προετοιμασίας" δεδομένων:** διακριτοποίηση, φιλτράρισμα παραδειγμάτων και ιδιοτήτων, αναπλήρωση χαμένων και άπειρων τιμών, κανονικοποίηση, αφαίρεση άχρηστων ιδιοτήτων, δειγματοληψία, μείωση πολυδιαστατικότητας και πολλά άλλα.
- **Τελεστές χαρακτηριστικών:** αλγόριθμοι επιλογής, όπως επιλογή προς τα εμπρός, απαλοιφή προς τα πίσω, και πολλοί γενετικοί αλγόριθμοι, τελεστές για εξαγωγή χαρακτηριστικών από χρονικές σειρές, στάθμιση και συνάφεια χαρακτηριστικών και παραγωγή νέων.
- **Μετά-τελεστές:** τελεστές βελτιστοποίησης για σχεδιασμό διαδικασιών, π.χ. επαναλήψεις δεδομένων ή τεχνικές βελτιστοποίησης παραμέτρων.

- **Τελεστές αξιολόγησης απόδοσης:** διασταύρωση και άλλες τεχνικές αξιολόγησης, πολλά κριτήρια απόδοσης κατηγοριοποίησης και παλινδρόμησης, τελεστές για βελτιστοποίηση παραμέτρων σε εσώκλειστους τελεστές ή αλυσίδες τελεστών κ.α.
- **Τελεστές οπτικοποίησης:** τελεστές καταγραφής και παρουσίασης αποτελεσμάτων. Online δημιουργία 2D και 3D γραφημάτων των δεδομένων που σχετίζονται με μοντέλα εκμάθησης και άλλα διαδικαστικά αποτελέσματα.
- **Τελεστές εισόδου/ εξόδου:** ευέλικτοι τελεστές για είσοδο και έξοδο δεδομένων, τεχνική υποστήριξη αρκετών τύπων αρχείων, συμπεριλαμβανομένων των Arff, C4.5, csv, bibtex, dBase, και απευθείας ανάγνωση δεδομένων από βάσεις δεδομένων.

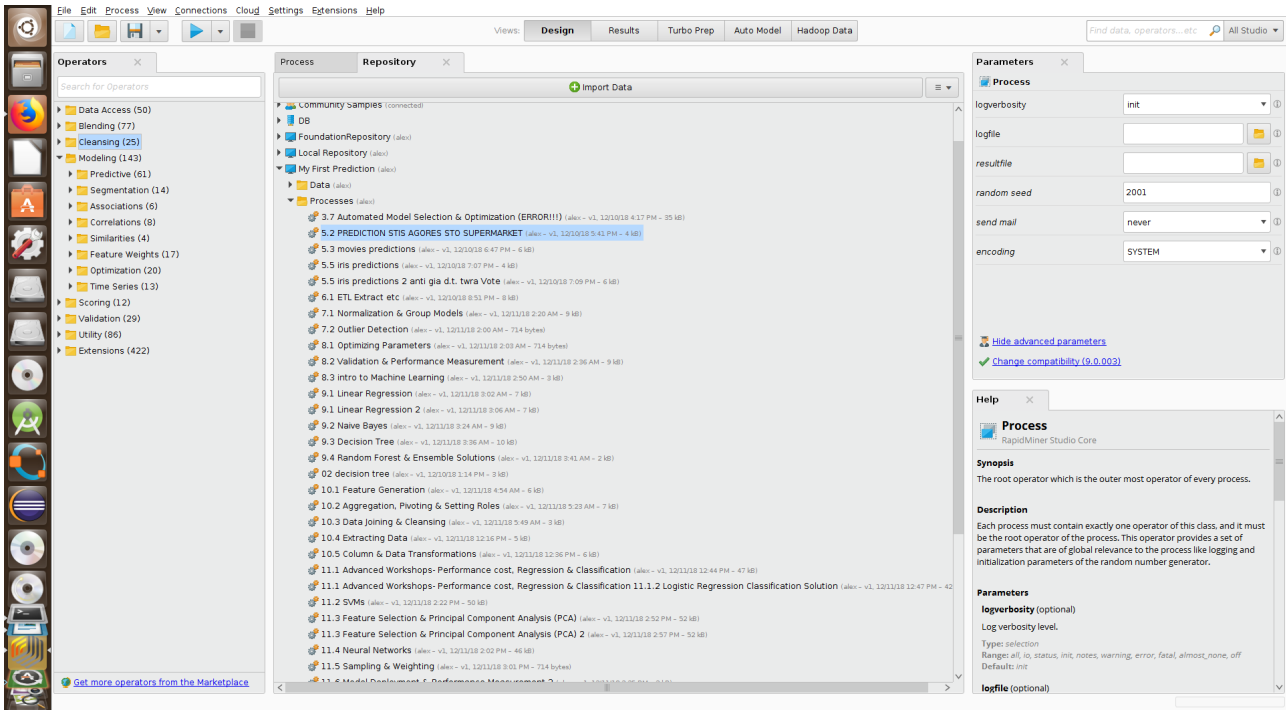


Εικόνα 38: Decision Tree Process



Εικόνα 39: Decision Tree Subprocess

2.3.4 Rapid Miner Repository



Εικόνα 40: Rapid Miner Repository

Η αποθήκευση των δεδομένων και των διαδικασιών στο Repository προσφέρει πολλά πλεονεκτήματα:

- Δεδομένα, διαδικασίες, αποτελέσματα και αναφορές αποθηκεύονται σε τοποθεσίες που υποδεικνύονται ως σχετικές μεταξύ τους, διευκολύνοντας έτσι το χρήστη να εντοπίζει αυτό που αναζητά.
- Το άνοιγμα και φόρτωμα των αρχείων δεν απαιτεί περαιτέρω ρυθμίσεις. Τα δεδομένα μπορούν να ανοιχτούν, διαβαστούν ή ενσωματωθούν στη διαδικασία με ένα απλό πάτημα του ποντικιού. Μπορεί έτσι ο χρήστης να έχει μία επισκόπηση των αποθηκευμένων δεδομένων, των χαρακτηριστικών και υποσημειώσεων τους οποιαδήποτε στιγμή, χωρίς να πρέπει να ανοίξει ξεχωριστά το αρχείο.
- Όλα τα δεδομένα εισόδου/ εξόδου αλλά και τα ενδιάμεσα αποτελέσματα υπομνηματίζονται με μετά-πληροφορίες. Αυτό εξασφαλίζει τη συνέπεια και ακεραιότητα των δεδομένων και κάνει εφικτές τις διαδικασίες εγκυροποίησης κατά τη διάρκεια της ανάπτυξης, καθώς και την παροχή συστατικών βοήθειας με ευαισθησία περιεχομένου. Το Repository μπορεί να βρίσκεται είτε σε τοπικό είτε σε κοινό σύστημα αρχείων ή ακόμα και να είναι διαθέσιμο από τον εξωτερικό διακομιστή ανάλυσης του RapidMiner, τον Rapid Analytics.

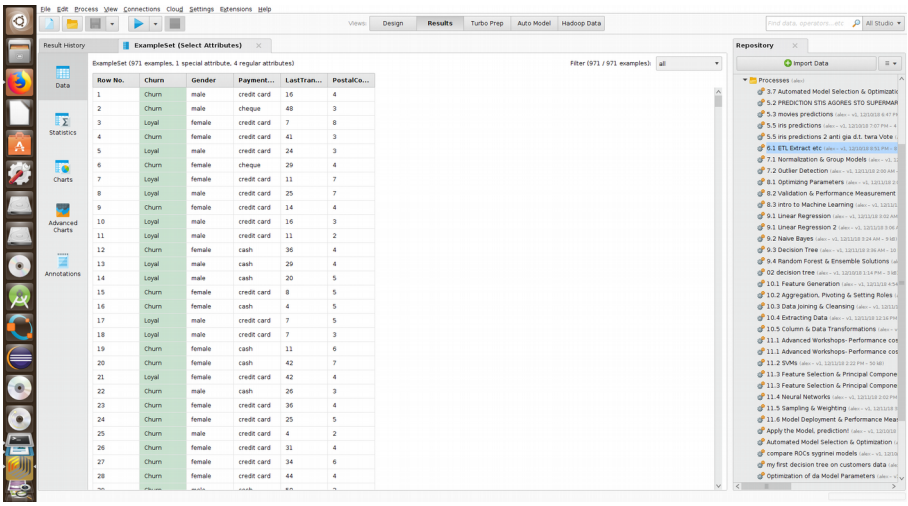
2.3.5 Μορφή Εμφάνισης των Δεδομένων

Στο Rapid Miner κάθε αποτέλεσμα προβάλεται μέσα από τη δικιά του κάρτα αρχείου. Υπάρχουν πολλοί διαφορετικοί τρόποι για την παρουσίαση των αποτελεσμάτων. Για παράδειγμα, για τα σύνολα δεδομένων υπάρχουν τρεις επιλογές εμφάνισης˙ η εμφάνιση των μετά-δεδομένων και στατιστικών, η εμφάνιση των ίδιων των δεδομένων και η εμφάνιση διαφορετικών οπτικοποιήσεων των δεδομένων.

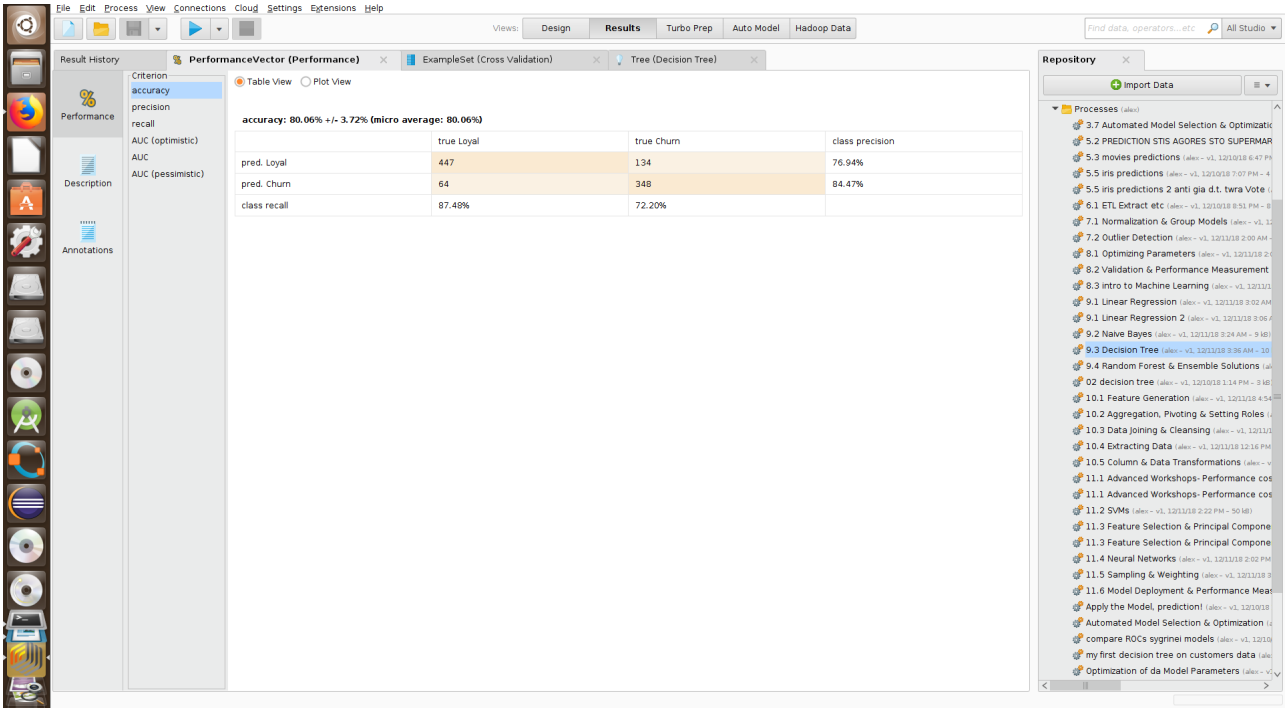
- **Πίνακες (table):** Μια από τις πιο συνηθισμένες μορφές εμφάνισης πληροφορίας στο RapidMiner είναι αυτή του πίνακα. Πρέπει να σημειωθεί πως οι πίνακες δε χρησιμοποιούνται μόνο για την εμφάνιση συνόλων δεδομένων αλλά και για την εμφάνιση μετά-δεδομένων, παραγόντων που επηρεάζουν τη στάθμιση, μητρών όπως οι συσχετίσεις μεταξύ γνωρισμάτων και πολλών άλλων. Αυτές οι μορφές εμφανίσεων έχουν συνήθως τον όρο Table στο όνομα τους, ειδικά αν υπάρχει φόβος σύγχυσης.

- **Διαγράμματα (Plots):** Ένα από τα δυνατότερα χαρακτηριστικά του RapidMiner είναι οι πολυάριθμες μέθοδοι οπτικοποίησης δεδομένων, πινάκων, μοντέλων και άλλων αποτελεσμάτων που βρίσκονται στη καρτέλα Plot View. Υπάρχουν δυο λειτουργίες 3D διαγραμμάτων ενσωματωμένες στο RapidMiner˙ η πρώτη παράγει 3D έγχρωμα διαγράμματα, με δυνατότητα περιστροφής με χρήση του ποντικιού ενώ δεύτερη είναι η λειτουργία 2D έγχρωμου διαγράμματος. Υπάρχουν επίσης και συστατικά για σχεδιασμό διαγραμμάτων διασποράς και ιστογραμμάτων.

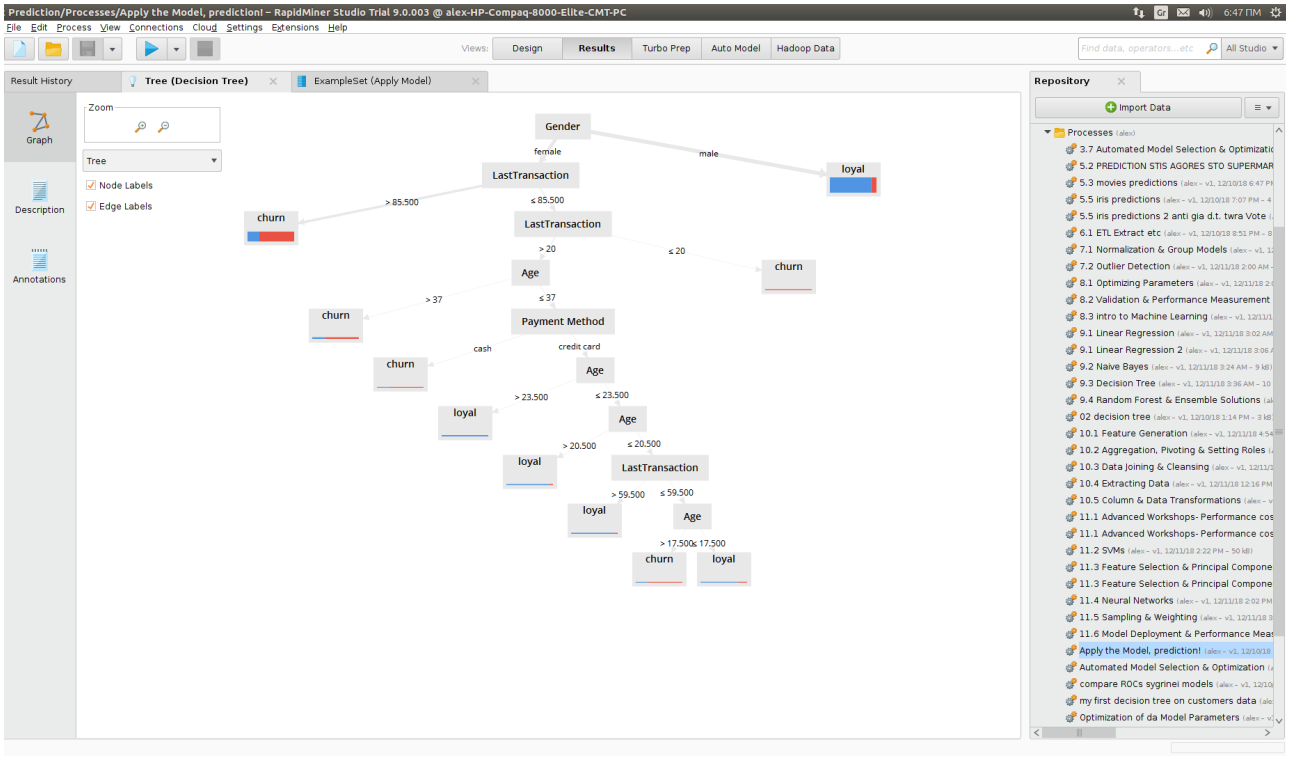
- **Γραφήματα (Graphs):** Τα γραφήματα είναι μια ακόμη μορφή εμφάνισης που συναντάται αρκετά συχνά στο RapidMiner. Με τον όρο γραφήματα εννοούμε κυρίως όλες τις οπτικοποιήσεις που απεικονίζουν κόμβους και τις μεταξύ τους σχέσεις (μπορεί να είναι κόμβοι εντός μιας ιεραρχικής συσταδοποίησης ή κόμβοι ενός δέντρου απόφασης).



Εικόνα 41: Decision Tree Example Set

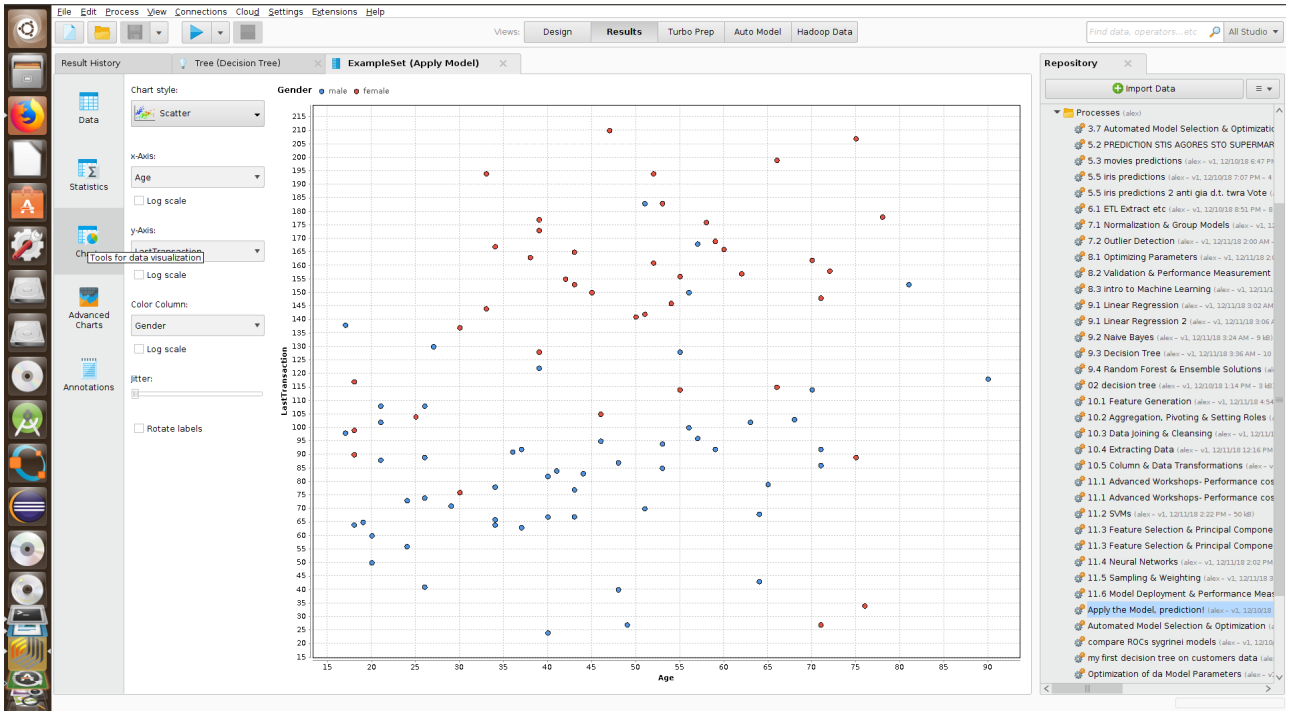


Εικόνα 42: Decision Tree Performance



Εικόνα 43: Decision Tree

Εικόνα 44: Ραβδόγραμμα



Εικόνα 45: Charts

2.3.6 Πλεονεκτήματα απ’ τη Χρήση του Rapid Miner

Το RapidMiner προσφέρει μία απλή διεπιφάνεια χρήστη, όπου όλοι οι τελεστές εμφανίζονται σε κατηγορίες στο παράθυρο προβολής τελεστών. Η εισαγωγή δεδομένων γίνεται απλά και γρήγορα, είτε με την εισαγωγή τελεστή ανάγνωσης αρχείου, είτε με την εισαγωγή τελεστή παραγωγής νέων δεδομένων. Δίνεται η δυνατότητα επιλογής ανάμεσα σε πολλούς τελεστές κατηγοριοποίησης των δεδομένων και οπτικοποίησης των αποτελεσμάτων. Στο παράθυρο της κυρίως διαδικασίας, μπορούμε να εισάγουμε τους τελεστές που θέλουμε και να δημιουργήσουμε το επιθυμητό μοντέλο. Η παραμετροποίηση των αλγορίθμων γίνεται επίσης πολύ εύκολα, αφού υπάρχει ενσωματωμένο παράθυρο στη διεπιφάνεια χρήστη όπου εμφανίζονται οι παράμετροι του κάθε τελεστή, έτσι ώστε να μπορεί ο χρήστης να προβεί στη ρύθμιση τους ανά πάσα στιγμή. Τα αποτελέσματα αλλά και τα ίδια τα δεδομένα εμφανίζονται στη προβολή αποτελεσμάτων του RapidMiner, όπου δίνονται πολλές επιλογές σχετικά με τον τρόπο εμφάνισης και προβολής τους. Τέλος, παρέχεται online τεκμηρίωση και παραδείγματα για την καλύτερη κατανόηση του RapidMiner και των λειτουργιών του.

2.4 Ηλιακή Ενέργεια & Φωτοβολταϊκό Φαινόμενο

2.4.1 Ηλιακή Ενέργεια

Ο ήλιος είναι η βασική πηγή ζωής στον πλανήτη μας. Σχεδόν όλες οι μορφές παράγωγης ενέργειας είναι συσχετισμένες έμμεσα ή άμεσα με την ηλιακή. Στον παρακάτω πίνακα, όπου παρουσιάζεται η Μέση Ένταση στην επιφάνεια της Γης συναρτήσει των κύριων Πηγών Ενέργειας του Κλιματικού Συστήματος, φαίνεται ξεκάθαρα η υπεροχή της ηλιακής έναντι των υπολοίπων.

Πηγές Ενέργειας	Μέση Ένταση (W / m^2)
Η/Μ Ακτινοβολία	240 ⇒ 99,97 %
Ενεργειακά Σωματίδια	0,001
Γεωθερμία	0,06
Ανθρωπογενείς Πηγές	0,02

Πίνακας 1: Πηγές Ενέργειας και Μέση Ένταση

Η ηλιακή ενέργεια προέρχεται από τον πυρήνα του Ήλιου, όπου γίνεται η θερμοπυρηνική σύντηξη ατόμων υδρογόνου και ατόμων ηλίου. Κάθε δευτερόλεπτο αυτής της διεργασίας, 630 εκατομμύρια τόνοι υδρογόνου μετατρέπονται σε 635 εκατομμύρια τόνους ηλίου.

$$4(1H) \mapsto 4He + energy + 2\ neutrinos \qquad (5)$$

Οι υπόλοιποι 5 εκατομμύρια τόνοι (έλλειμμα μάζας) μετατρέπονται σε ηλεκτρομαγνητική ενέργεια, η οποία εκλύεται από την επιφάνεια του Ήλιου στο διάστημα. Ο υπολογισμός γίνεται μέσω της γνωστής εξίσωσης:

$$E = Mc^2 \qquad (6)$$

Ο Ήλιος έχει ανομοιογενή κατανομή θερμοκρασίας και πολύ διαφορετικές θερμοκρασίες στα διάφορα στρώματά του, από $16 \cdot 10^6$ Kelvin βαθμούς στον πυρήνα του, έως μερικές χιλιάδες ή εκατομμύρια βαθμούς στο εξωτερικό του (φωτόσφαιρα και στέμμα, αντίστοιχα). Η θερμοκρασία αυξάνεται από τη φωτόσφαιρα προς τα έξω (χρωμόσφαιρα και στέμμα).

Ως ακτινοβολούσα επιφάνεια του Ήλιου θεωρείται η φωτόσφαιρα διότι το στέμμα έχει πολύ μικρή πυκνότητα (αραιό). Η μέση θερμοκρασία της φωτόσφαιρας είναι περίπου 6000 Kelvin και η παραγόμενη Ισχύς του Ήλιου στη φωτόσφαιρα υπολογίζεται απ' τον παρακάτω τύπο:

$$P = \varepsilon \cdot \sigma \cdot T \cdot (4 \cdot \pi \cdot R_H^2) \qquad (7)$$

όπου:

- ε: η Ικανότητα Εκπομπής του Ήλιου που ισούται με 1
- σ: η Σταθερά Boltzmann που ισούται με 5,67 10-8 $W \cdot m^{-2} \cdot grad^{-4}$
- R_h: η Μέση Ακτίνα του Ήλιου = $6,96 \cdot 10^8$ m
- $P = 3,91 \cdot 10^{23}$ kW

Το ενδιαφέρον εστιάζεται πρωτίστως στην ακτινοβολία που περιλαμβάνει μήκη κύματος από 0.25 έως 3.0 μm, το τμήμα εκείνο της ηλεκτρομαγνητικής ακτινοβολίας που περιέχει την περισσότερη ενέργεια που εκλύεται από τον ήλιο.

2.4.2 Χαρακτηριστικά Ηλιακής Ακτινοβολίας

Α) Ένταση ηλιακής ακτινοβολίας:

Είναι η ποσότητα που ισούται με την ηλιακή ενέργεια που προσπίπτει σε μία επιφάνεια ανά μονάδα χρόνου προς το εμβαδόν της επιφάνειας αυτής. Η μονάδα μέτρησής της είναι *Joule / (seconds * (meters ^ 2))* και επειδή *Joule / second = Watt* (W= Μονάδα Μέτρησης της Ισχύος) καταλήγουμε στην απλούστερη έκφραση W/m^2.

Β) Ηλιακή σταθερά G_{sc}:

Η ένταση της ηλιακής ακτινοβολίας που εκπέμπει ο ήλιος πάνω σε μια επιφάνεια κάθετη στη διεύθυνση διάδοσής της στα όρια της ατμόσφαιρα της γης (και η οποία βρίσκεται στη μέση απόσταση ήλιου-γης), είναι γνωστή ως ηλιακή σταθερά G_{sc} και ισούται με 1366,25 $\pm$ 0,71 W/m^2. Υπολογίζεται δε απ' την εξίσωση:

$$G_{sc} = P / (4 \cdot \pi \cdot R_o^2) \qquad (8)$$

όπου R_o (Μέση Απόσταση Γης- Ήλιου) = $1,5 \cdot 10^8$. Να σημειωθεί ότι η Γη έχει πολύ μικρή ενεργό επιφάνεια απορρόφησης της ηλιακής ακτινοβολίας που προσεγγίζεται περισσότερο απ' αυτήν ενός δισδιάστατου δίσκου και όχι μιας τρισδιάστατης σφαίρας, γεγονός το οποίο συνεπάγεται ένα πολύ μικρό μέρος της συνολικής εκπεμπόμενης Έντασης να φτάνει στη Γη.

Γ) Ένταση Ηλιακής Ακτινοβολίας εκτός Ατμόσφαιρας σε Επίπεδο Κάθετο ως προς τις Ακτίνες:

Η Ένταση της ηλιακής ακτινοβολίας σε μια περιοχή μεταβάλλεται

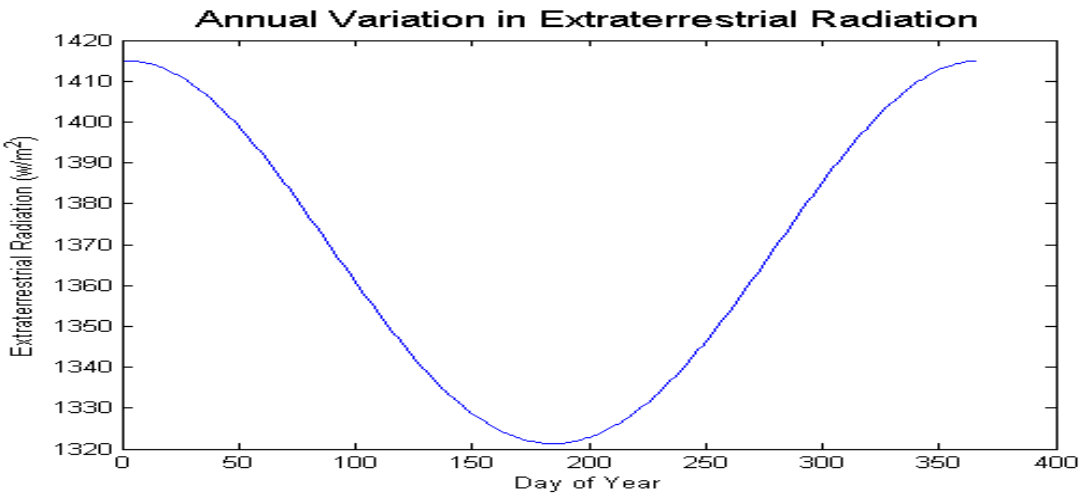
- κατά τη διάρκεια του εικοσιτετραώρου, λόγω περιστροφής της γης περί τον άξονά της,
- και κατά την διάρκεια του έτους, λόγω περιστροφής της γης γύρω από τον ήλιο σε ελλειπτική τροχιά.

Η μεταβολή της απόστασης της γης από τον ήλιο έχει ως αποτέλεσμα τη μεταβολή της ακτινοβολίας εκτός της ατμόσφαιρας.
Αν G_{on} είναι η ένταση της ακτινοβολίας εκτός της ατμόσφαιρας σε επίπεδο κάθετο προς τις ακτίνες του ήλιου, τη n-ιοστή μέρα του έτους, θα είναι:

$$G_{on} = Gsc \cdot [1 + 0,033 \cdot \cos((360 \cdot n) / 365)] \qquad (9)$$

Όπου:

- G_{sc}: ηλιακή σταθερά
- n: η n-ιοστή μέρα του έτους, με n=1 την πρώτη Ιανουαρίου και n=365 την τριακοστή πρώτη Δεκεμβρίου



Εικόνα 46: Ετήσιες μεταβολές στην ακτινοβολία στα ατμοσφαιρικά όρια

Δ) Συνιστώσες ηλιακής ακτινοβολίας:

Όταν η ηλιακή ακτινοβολία εισέλθει στην ατμόσφαιρα, ένα μέρος της προσπίπτει σε μόρια αέρα, νερού και σκόνης με αποτέλεσμα να διαχέεται και ένα άλλο απορροφάται από O_3,

H_2O και CO_2. Κατά τις μεσημεριανές ώρες μιας ανέφελης ημέρας, περίπου το 25% της ηλιακής ακτινοβολίας σκεδάζεται ή απορροφάται καθώς διασχίζει την ατμόσφαιρα και ως συνέπεια αυτού, μόνο το 75 % της ηλιακής ακτινοβολίας προσπίπτει στο έδαφος χωρίς να έχει υποστεί κάποια επίδραση. Αυτή η συνιστώσα της ακτινοβολίας ονομάζεται ακτινική ή άμεση ακτινοβολία.

Ως προς το ποσοστό της ακτινοβολίας που σκεδάζεται, ένα μέρος της χάνεται στο διάστημα ενώ το υπόλοιπο φτάνει στο έδαφος με διαφορετική πλέον κατεύθυνση. Αυτή η ακτινοβολία ονομάζεται διάχυτος ακτινοβολία.

Η ολική ακτινοβολία που προσπίπτει σε ένα οριζόντιο επίπεδο είναι το άθροισμα της ακτινικής και της διάχυτης ακτινοβολίας.

$$I = I_b + I_d \qquad (10)$$

Όπου:

- I: η ολική ηλιακή ακτινοβολία σε οριζόντιο επίπεδο στην επιφάνεια της γης
- I_b: η ακτινική συνιστώσα της ηλιακής ακτινοβολίας σε οριζόντιο επίπεδο
- I_d: η διάχυτη συνιστώσα της ηλιακής ακτινοβολίας σε οριζόντιο επίπεδο

Εάν η υπό εξέταση επιφάνεια είναι κεκλιμένη, τότε στον προηγούμενο ορισμό πρέπει να προσθέσουμε την ηλιακή ακτινοβολία που ανακλάται από το έδαφος.

$$I_T = I_{b,T} + I_{d,T} + I_{refl,T} \qquad (11)$$

Όπου:
I_T: η ολική ηλιακή ακτινοβολία σε κεκλιμένο επίπεδο στην επιφάνεια της γης
$I_{b,T}$: η ακτινική συνιστώσα της ηλιακής ακτινοβολίας σε κεκλιμένο επίπεδο
$I_{d,T}$: η διάχυτη συνιστώσα της ηλιακής ακτινοβολίας σε κεκλιμένο επίπεδο
$I_{refl,T}$: η ανακλώμενη συνιστώσα της ηλιακής ακτινοβολίας σε κεκλιμένο επίπεδο

Ε) Παράμετροι εξαρτώμενοι απ' τη σχετική θέση Γης- Ήλιου:

- Γεωγραφικό πλάτος φ: Το γεωγραφικό πλάτος του τόπου εγκατάστασης εκφράζει τη γωνία που σχηματίζει ο τόπος με τον ισημερινό και είναι $\varphi \in [-90^\circ, 90^\circ]$. Οι θετικές τιμές της γωνίας φ αντιστοιχούν στο βόρειο ημισφαίριο, ενώ οι αρνητικές τιμές στο νότιο ημισφαίριο.
- Γεωγραφικό μήκος του τόπου θ: Το γεωγραφικό μήκος του τόπου εγκατάστασης εκφράζει τη γωνία που σχηματίζει ο τόπος με τον πρώτο μεσημβρινό (αστεροσκοπείο του Greenwich). Ισχύει $\theta \in [-180^\circ, 180^\circ]$. Οι θετικές τιμές της γωνίας θ αντιστοιχούν στο δυτικό ημισφαίριο, ενώ οι αρνητικές τιμές στο ανατολικό ημισφαίριο.
- Γωνία ανύψωσης ηλίου (α_s) και αζιμούθιος γωνία ηλίου (γ_s):
    - Η γωνία ανύψωσης ήλιου είναι η γωνία μεταξύ του ορίζοντα και της ευθείας του σημείου του παρατηρητή με τον ήλιο.
    - Η αζιμούθιος γωνία ήλιου είναι η γωνιακή απόκλιση της προβολής της ακτινικής συνιστώσας στο οριζόντιο επίπεδο από το νότο.

Κατά την διάρκεια της ημέρας, η α_s και η γ_s μεταβάλλονται συνεχώς καθώς ο ήλιος διατρέχει τον ουρανό. Για την αζιμούθιο ισχύει:

- Για νότιο προσανατολισμό $\gamma_s=0$
- Για γωνίες δυτικά από το νότο παίρνει θετικές τιμές.
- Για γωνίες ανατολικά από το νότο παίρνει αρνητικές τιμές.

- Ζενιθιακή γωνία του ήλιου θ_z:

$$\theta_z = 90° - \alpha_s \qquad (12)$$

- Αέριος Μάζα AM: Αέριος μάζα ονομάζεται ο λόγος της ατμοσφαιρικής μάζας που διαπερνά η ακτινική συνιστώσα της ακτινοβολίας προς τη μάζα που θα διαπερνούσε αν ο ήλιος βρισκόταν στο ζενίθ, δηλαδή στην κατακόρυφο. Αυτό συνεπάγεται ότι το AM=1 στο επίπεδο της θάλασσας όταν ο ήλιος είναι στο ζενίθ. Για $\theta_z \in (0°, 70°)$ στο επίπεδο της θάλασσας μια καλή προσέγγιση (m) της αερίου μάζας (AM) δίνεται από τον τύπο:

$$m = \cos\theta_z^{-1} \qquad (13)$$

Για μεγαλύτερες γωνίες πρέπει να ληφθεί υπόψη και η καμπυλότητα της γης.

- Απόλυτη Αέριος Μάζα AM_α: υπολογίζεται πολλαπλασιάζοντας την τιμή της AM με το πηλίκο της ατμοσφαιρικής πίεσης στο σημείο εξέτασης (P) προς την ατμοσφαιρική πίεση στο επίπεδο της θάλασσας (P_o):

$$AM_\alpha = AM \cdot (P / P_o) \qquad (14)$$

- Απόκλιση ηλίου δ: Ο σημαντικότερος παράγοντας που διαμορφώνει την Ένταση της ηλιακής ακτινοβολίας στα όρια της ατμόσφαιρας είναι η σχετική θέση του ήλιου και της γης. Η θέση του ήλιου παίρνει πολύ διαφορετικές τιμές, ως αποτέλεσμα της μεταβολής της δ, δηλαδή της γωνίας που σχηματίζεται ανάμεσα στην ευθεία που ενώνει το κέντρο της γης με το κέντρο του ήλιου, και στο επίπεδο του ισημερινού. Οι τιμές της απόκλισης του ήλιου είναι θετικές για το βόρειο ημισφαίριο και αρνητικές για το νότιο.

  Άμεση συνέπεια των διαφορετικών τιμών της δ κατά τη διάρκεια του έτους είναι το γεγονός ότι οι κυκλικές τροχιές διαγράφονται βορειότερα στον ουρανό το καλοκαίρι, δηλαδή ο ήλιος ανατέλλει νωρίτερα και δύει αργότερα σε σχέση με τις άλλες εποχές του χρόνου όσον αφορά το βόρειο ημισφαίριο. Παράλληλα διαμορφώνονται οι αντίστοιχες μετεωρολογικές και κλιματολογικές συνθήκες που επικρατούν στις διάφορες εποχές. Χρήσιμα μεγέθη για τη γενική εκτίμηση της καθημερινής και της εποχιακής διακύμανσης της ακτινοβολίας σε ένα τόπο, είναι η θεωρητική ηλιοφάνεια, δηλαδή το χρονικό διάστημα από την ανατολή μέχρι τη δύση του ήλιου, καθώς και η μέση πραγματική ηλιοφάνεια που δείχνει το μέσο όρο των ωρών που ο ήλιος δεν καλύπτεται από σύννεφα. Επίσης, ο αριθμός των ημερών με ηλιοφάνεια, στη διάρκεια των οποίων ο ήλιος δεν καλύπτεται από σύννεφα, καθώς και των

ανήλιων ημερών, που ο ήλιος καλύπτεται από σύννεφα σε ολόκληρο το διάστημα της ημέρας.

Η ηλιακή απόκλιση δίνεται από την παρακάτω σχέση:

$$\delta = 23{,}45 \cdot \sin[360 \cdot (284 + n) / 365] \qquad (15)$$

Όπου:

- n: η n-ιοστή μέρα του έτους, με n=1 την πρώτη Ιανουαρίου και n=365 την τριακοστή πρώτη Δεκεμβρίου

Είναι δ=0˚ κατά την εαρινή ισημερία (21 Μαρτίου) και την φθινοπωρινή ισημερία (21 Σεπτεμβρίου), οπότε ο ήλιος ανατέλλει ακριβώς στην ανατολή και δύει στη δύση. Επίσης είναι δ=23,45˚ την εικοστή πρώτη Ιουνίου (θερινό ηλιοστάσιο) και δ=-23,45˚ την 21 Δεκεμβρίου (χειμερινό ηλιοστάσιο).

- Ηλιακός χρόνος t_s και ωριαία γωνία ηλίου ω: Ο ηλιακός χρόνος είναι ο χρόνος που μετράται με βάση τη φαινόμενη κίνηση του ήλιου στον ουρανό και δε συμπίπτει με τον τοπικό χρόνο του ωρολογίου. Οι δύο χρόνοι συνδέονται με τη σχέση:

$$t_s = t_c + \theta / 15 - T_c + 3{,}82 \cdot \{0{,}00075 + 0{,}001868 \cdot \cos[360 \cdot (n - 1) / 365] - 0{,}032077 \cdot \sin[360 \cdot (n - 1) / 365] - 0{,}014615 \cdot \cos[2 \cdot 360 \cdot (n - 1) / 365] - 0{,}004089 \cdot \sin[2 \cdot 360 \cdot (n - 1) / 365]\} \qquad (16)$$

Όπου :

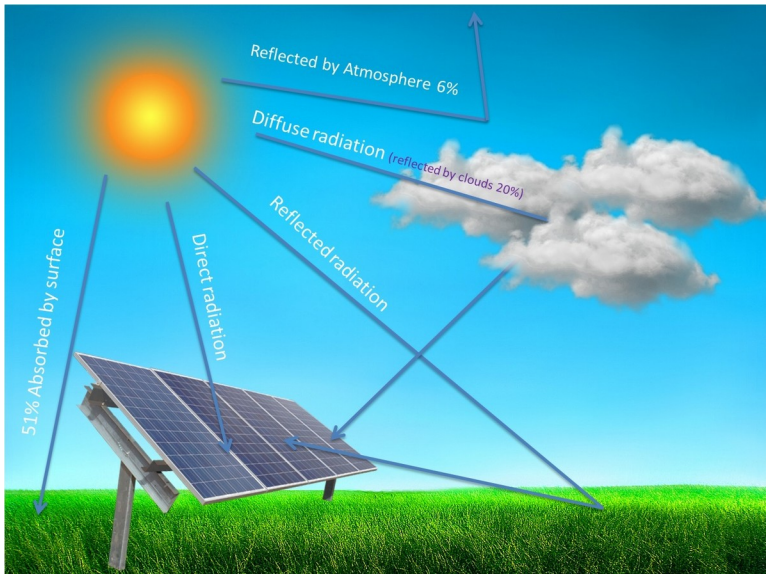
- t_s: ο ηλιακός χρόνος σε ώρες
- t_c: ο τοπικός χρόνος σε ώρες
- θ: το γεωγραφικό μήκος της θέσης του παρατηρητή σε μοίρες
- T_C: η τοπική ωρολογιακή ζώνη σε σχέση με GMT σε ώρες (για την Ελλάδα +2 ώρες)
- n: η n-ιοστή μέρα του έτους, με n=1 την πρώτη Ιανουαρίου και n=365 την τριακοστή πρώτη Δεκεμβρίου

Η ημερήσια κίνηση της γης απεικονίζεται με την ημερήσια περιστροφή της ουράνιας σφαίρας περί τον πολικό άξονα και η στιγμιαία θέση του ήλιου περιγράφεται από την ωριαία γωνία ω, η οποία ορίζεται ως η γωνία μεταξύ του μεσημβρινού που περνάει από τον ήλιο και του μεσημβρινού που περνάει από τη θέση της εγκατάστασης. Είναι ω=0˚ κατά το ηλιακό μεσημέρι και αυξάνει με ρυθμό 15˚/h. Είναι ω<0˚ για π.μ. και ω>0˚ για μ.μ. Η ωριαία γωνία του ήλιου ω υπολογίζεται από τη σχέση:

$$\omega = (t_s - 12) \cdot 15 \qquad (17)$$

Όπου ω, η ωριαία γωνία του ήλιου σε ώρες.

Στ) Συλλέκτης Ηλιακής Ενέργειας:



Εικόνα 47: Συνιστώσες Ηλιακής Ακτινοβολίας

Ο Συλλέκτης Ηλιακής Ενέργειας εξαρτάται απ' τα εξής γεωμετρικά στοιχεία:

- Γωνία πρόσπτωσης της ηλιακής ακτινοβολίας ψ: Η γωνία ψ ορίζεται ως η γωνία μεταξύ της ακτινικής συνιστώσας της ακτινοβολίας που προσπίπτει σε μια επιφάνεια και της καθέτου στην επιφάνεια αυτή.

- Προσανατολισμός του συλλέκτη (β), αζιμούθια γωνία επιφάνειας (γ): Ένα από τα σημαντικότερα στοιχεία κάθε συστήματος που εκμεταλλεύεται την ηλιακή ενέργεια είναι ο προσανατολισμός του ηλιακού συλλέκτη σε σχέση με την κατεύθυνση της ηλιακής ακτινοβολίας. Όπως η θέση του ήλιου στον ουρανό, έτσι και ο προσανατολισμός ενός επίπεδου στην επιφάνεια της γης περιγράφεται από δύο γωνίες, την κλίση (β) και την αζιμούθια γωνία επιφάνειας (γ).

    - Η κλίση του συλλέκτη (β) είναι η δίεδρη γωνία που σχηματίζεται ανάμεσα στο επίπεδο του συλλέκτη και στον ορίζοντα και μπορεί να πάρει τιμές από 0˚ μέχρι 180˚. Για γωνίες β>90˚ το επίπεδο του συλλέκτη είναι στραμμένο προς τα κάτω.

    - Η αζιμούθια γωνία επιφάνειας του συλλέκτη (γ) είναι η γωνία που σχηματίζεται πάνω στο οριζόντιο επίπεδο ανάμεσα στην προβολή της κατακόρυφου του συλλέκτη και στον τοπικό μεσημβρινό βορρά-νότου. Παίρνει τιμές από -180˚ έως 180˚. Η γωνία -180˚ (που συμπίπτει με την +180˚) αντιστοιχεί σε τοποθέτηση του συλλέκτη προς το βορρά, η γωνία -90˚ προς την ανατολή, η γωνία 0˚ προς το νότο και η γωνία 90˚ προς τη δύση.

Προφανώς, η πυκνότερη ισχύς μιας δέσμης ηλιακής ακτινοβολίας, πάνω σε ένα επίπεδο συλλέκτη θα υφίσταται όταν η επιφάνειά του είναι κάθετη προς τη κατεύθυνση της ακτινοβολίας, δηλαδή όταν η γωνία πρόσπτωσης ψ είναι 0˚. Η συνθήκη όμως αυτή δεν είναι εύκολο να εξασφαλιστεί καθώς ο ήλιος συνεχώς μετακινείται κατά τη διάρκεια της ημέρας. Προς εξασφάλιση τούτης της συνθήκης έχουν κατασκευαστεί μηχανικές διατάξεις που επαναπροσανατολίζουν συνεχώς τον συλλέκτη (π.χ. με τη βοήθεια ηλεκτρονικού υπολογιστή ή φωτοκύτταρων), ώστε η επιφάνεια του να αντικρίζει πάντα κάθετα τον ήλιο. Οι διατάξεις όμως αυτές είναι πολύπλοκες και δαπανηρές. Έτσι, η χρήση τους δικαιολογείται μόνον σε περιπτώσεις ειδικών εφαρμογών.

Στις συνηθισμένες περιπτώσεις οι συλλέκτες τοποθετούνται σε σταθερή κλίση και αζιμούθια γωνία, που επιλέγονται έτσι ώστε η γωνία πρόσπτωσης της ηλιακής ακτινοβολίας να είναι όσο το δυνατό μικρότερη, κατά τη διάρκεια του έτους.

Οι επίπεδοι συλλέκτες χρησιμοποιούν την άμεση, τη διάχυτη και την ανακλώμενη ακτινοβολία. Η επιλογή του ευνοϊκού προσανατολισμού και της κλίσης του συλλέκτη είναι το σημαντικότερο μέτρο για τη βελτίωση του ηλιακού κέρδους, αφού επηρεάζει σημαντικά το ποσό της ηλιακής ενέργειας που προσπίπτει στην επιφάνεια του συλλέκτη. Στο βόρειο ημισφαίριο, η βέλτιστη κλίση του συλλέκτη, για τη διάρκεια του έτους, είναι ίση με τον γεωγραφικό παράλληλο του τόπου, και η αζιμούθια γωνία είναι 0˚ (προς το νότο). Λόγω όμως της μεταβολής της απόκλισης του ήλιου (δ) στη διάρκεια του έτους, η βέλτιστη κλίση του συλλέκτη είναι διαφορετική για κάθε εποχή. Έτσι, αν επιδιώκεται να παράγει το σύστημα όσο το δυνατόν περισσότερη ενέργεια στη διάρκεια του καλοκαιριού, η κλίση του συλλέκτη επιλέγεται περίπου 10˚ ως 15˚ μικρότερη από την παράλληλο του τόπου, ενώ για τον χειμώνα η κλίση επιλέγεται περίπου 10˚ ως 15˚ μεγαλύτερη από την παράλληλο του τόπου. Αν στο έδαφος υπάρχει επιφάνεια με μεγάλο συντελεστή ανάκλασης (π.χ. χιόνι) απαιτείται μεγαλύτερη κλίση. Ο βέλτιστος προσανατολισμός για την αζιμούθια γωνία επιφάνειας είναι νότιος (γ=0˚), ενώ απόκλιση κατά 20°-30° από το νότο έχει μικρή επίδραση.

Τα παραπάνω ισχύουν για τη συλλογή της άμεσης ηλιακής ακτινοβολίας που έρχεται σαν δέσμη από τον ήλιο. Για τις άλλες, από ενεργειακή άποψη λιγότερο σημαντικές, μορφές της ηλιακής ακτινοβολίας, ο κυριότερος παράγοντας είναι η απόλυτη τιμή της κλίσης του συλλέκτη, ανεξάρτητα από τη θέση του ήλιου. Έτσι, όσο η κλίση απέχει περισσότερο από το οριζόντιο, τόσο μεγαλύτερο ποσό ανακλώμενης ακτινοβολίας από το έδαφος δέχεται ο συλλέκτης, αλλά και τόσο μικρότερο ποσό διάχυτης ακτινοβολίας από τον ουρανό.

Για παράδειγμα, σε περιοχές με υγρό κλίμα, όπου λόγω των σταγονιδίων του νερού στην ατμόσφαιρα, ένα μεγάλο μέρος της ηλιακής ακτινοβολίας διαχέεται στον ουρανό, η βέλτιστη κλίση του ηλιακού συλλέκτη για τη διάρκεια ολόκληρου του έτους είναι περίπου 10 - 15% μικρότερη από τη γωνία του τοπικού γεωγραφικού πλάτους. Έτσι, ο συλλέκτης αντικρίζει περισσότερο τον ουρανό και δέχεται αφθονότερα τη διάχυτη ακτινοβολία.

Ζ) Η Ένταση της Ηλιακής Ακτινοβολίας στο Έδαφος και επίδραση της Ατμόσφαιρας σ’ αυτήν:

Η ένταση της ηλιακής ακτινοβολίας εκτός από τις μεταβολές που υφίσταται λόγω της κίνησης της γης περί τον ήλιο, και οι οποίες μπορούν να υπολογιστούν με βάση γεωμετρικές σχέσεις, υφίσταται έντονες διακυμάνσεις λόγω της παρεμβολής της

ατμόσφαιρας. Οι διακυμάνσεις αυτές διακρίνονται ανάλογα με το αν πρόκειται για αίθρια ατμόσφαιρα, δηλαδή δεν υπάρχουν νέφη, ή μη αίθρια (νεφελώδη) ατμόσφαιρα, χωρίς όμως να μπορεί να γίνει σαφής διάκριση. Ορίζεται ο συντελεστής καθαρότητας της ατμόσφαιρας ως ο λόγος της μετρούμενης ολικής έντασης της ακτινοβολίας σε οριζόντιο επίπεδο (I) και της ακτινοβολίας εκτός της ατμόσφαιρας στην ίδια θέση (I_{oh}).

$$K_T = I / I_{oh} \qquad (18)$$

Ο συντελεστής K_T αποτελεί χαρακτηριστικό του τόπου εγκατάστασης και συχνά δίδεται σε πίνακες ως μέση μηνιαία τιμή. Έχει αποδειχθεί στατιστικά ότι ανεξάρτητα από το γεωγραφικό πλάτος, το ποσοστό του χρόνου κατά τον οποίο η συνολική ημερήσια ακτινοβολία είναι μικρότερη ή ίση από ορισμένη τιμή εξαρτάται άμεσα από την τιμή του K_T. Είναι επίσης διαπιστωμένο, ότι όσο περισσότερο αίθρια είναι η ατμόσφαιρα, τόσο μεγαλύτερη είναι η ένταση της ακτινοβολίας και μικρότερο το ποσοστό της διάχυτης επί της ολικής. Από σχέσεις συσχέτισης διάχυτης και ολικής ακτινοβολίας, υπολογίζεται ο λόγος της διάχυτης ακτινοβολίας I_d προς την ολική I, ως συνάρτηση του συντελεστή καθαρότητας K_T.

Υπολογίζοντας τη διάχυτη ακτινοβολία, εύκολα υπολογίζουμε και την ακτινική (I_b). Σε οριζόντιο επίπεδο η ηλιακή ακτινοβολία έχει δύο συνιστώσες, τη διάχυτη και την ακτινική οπότε:

$$I_b = I - I_d \qquad (19)$$

Στην περίπτωση κεκλιμένου επιπέδου, η ηλιακή ακτινοβολία στην επιφάνεια της γης αποτελείται από τρεις συνιστώσες, την ακτινική που προέρχεται από τον ηλιακό δίσκο, τη διάχυτη που προέρχεται από τον ουράνιο θόλο και την ανακλώμενη που προέρχεται από το έδαφος της γύρω περιοχής:

$$I_T = I_{b,T} + I_{d,T} + I_{refl,T} \qquad (20)$$

Η ακτινική ακτινοβολία που δέχεται το κεκλιμένο επίπεδο εξαρτάται από την γωνία πρόσπτωσης ψ. Η διάχυτη και ανακλώμενη ακτινοβολία από την άλλη, δεν εξαρτώνται από τον προσανατολισμό του επιπέδου και ούτε προέρχονται από όλο τον ουράνιο θόλο ή το έδαφος της γύρω περιοχής. Η διάχυτη ακτινοβολία που δέχεται το κεκλιμένο επίπεδο προέρχεται μόνο από το τμήμα του ουρανού που βρίσκεται σε οπτική επαφή με το επίπεδο.

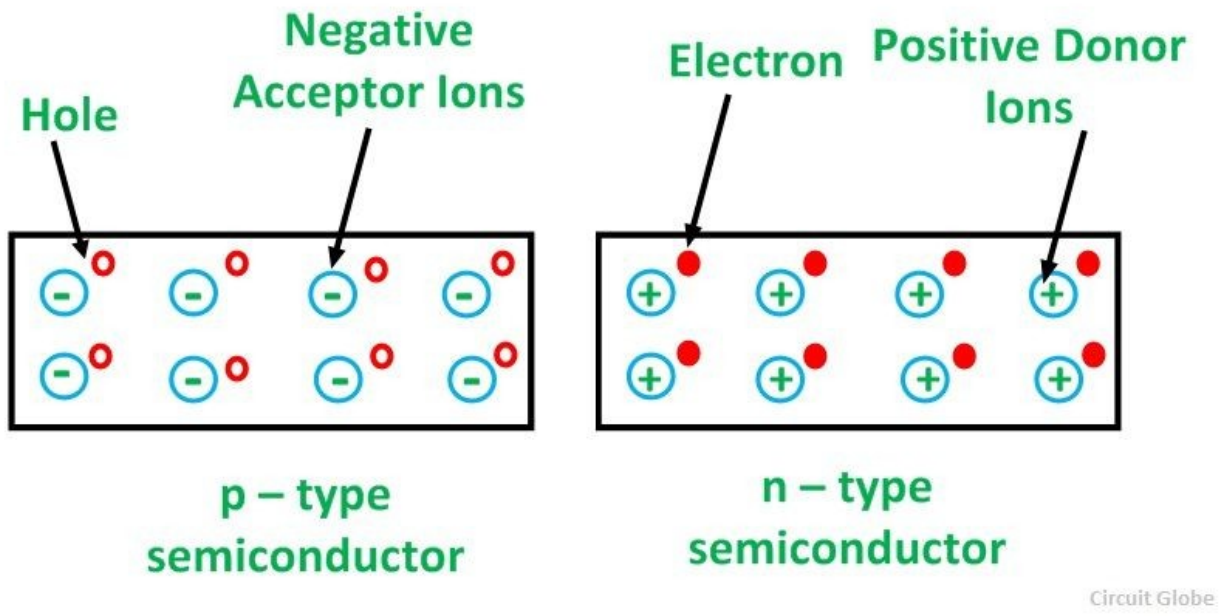
2.4.3 Φωτοβολταϊκό Φαινόμενο

Όταν η ηλιακή ακτινοβολία προσπίπτει σε μια επιφάνεια είτε ανακλάται, είτε την διαπερνά, είτε απορροφάται από το υλικό της επιφάνειας. Στην τελευταία περίπτωση το συνηθέστερο αποτέλεσμα είναι η μετατροπή της απορροφούμενης Φωτεινής Ενέργειας σε Θερμική που συνεπάγεται την αύξηση της Μέσης Κινητικής Ενέργειας, λόγω χαοτικής κίνησης, των δομικών στοιχείων του υλικού. Έτσι, όταν πλησιάζουμε το χέρι μας κοντά σε μία φωτεινή

πηγή παρατηρούμε ότι ζεσταίνεται. Επίσης, τις μέρες με μεγαλύτερη ηλιοφάνεια έχουμε υψηλότερες θερμοκρασίες.

Παρόλα αυτά υπάρχουν υλικά τα οποία έχουν την ιδιότητα να μετατρέπουν την ενέργεια των προσπιπτόντων φωτονίων σε Ηλεκτρική. Αυτά τα υλικά είναι οι ημιαγωγοί και σε αυτά οφείλεται η τεράστια τεχνολογική πρόοδος που έχει συντελεστεί τα τελευταία χρόνια στον τομέα της ηλεκτρονικής και συνεπακόλουθα στον ευρύτερο χώρο της πληροφορικής και των τηλεπικοινωνιών.

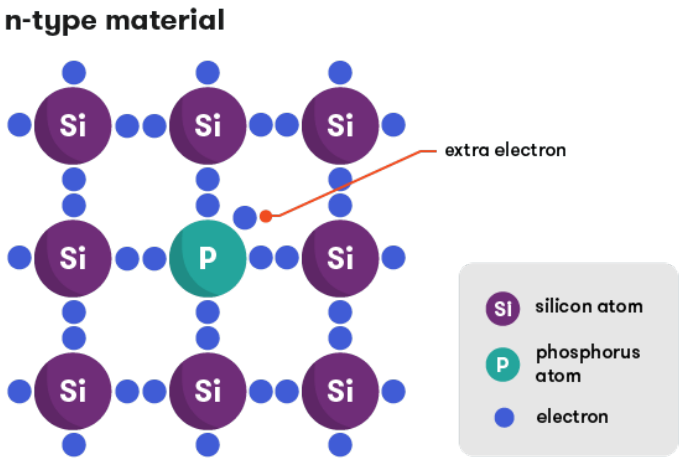
Α) Ημιαγωγοί Τύπου Ν και Τύπου Ρ:



Εικόνα 48: Ημιαγωγοί Τύπου Ρ και Τύπου Ν

- Οι ημιαγωγοί τύπου Ν δημιουργούνται όταν σε έναν ημιαγωγό όπως το πυρίτιο ή το γερμάνιο προστεθεί πολύ μικρή ποσότητα ενός στοιχείου της πέμπτης ομάδας του περιοδικού πίνακα. Τα στοιχεία που συνήθως χρησιμοποιούνται ως προσμίξεις είναι το αρσενικό, ο φωσφόρος και το αντιμόνιο ενώ η ποσότητα που απαιτείται είναι της τάξης των μερικών μερών στο εκατομμύριο, δηλαδή σε κάθε ένα εκατομμύριο άτομα πυριτίου ή γερμανίου να αντιστοιχούν μερικά άτομα αρσενικού ή φωσφόρου. Τα άτομα της πρόσμιξης ενσωματώνονται στην κρυσταλλική δομή του ημιαγωγού, και καταλαμβάνουν θέσεις των ατόμων του σχηματίζοντας ομοιοπολικούς δεσμούς με τα γειτονικά άτομα. Επειδή τα άτομα της πέμπτης ομάδας του περιοδικού πίνακα έχουν πέντε ηλεκτρόνια στη στοιβάδα σθένους, όταν καταλάβουν μία θέση σε ένα άτομο του ημιαγωγού θα χρησιμοποιήσουν τα τέσσερα για το σχηματισμό ομοιοπολικών δεσμών και θα παραμείνει αδιάθετο ένα ηλεκτρόνιο, το οποίο θα περιφέρεται γύρω από τον πυρήνα της πρόσμιξης. Το ηλεκτρόνιο αυτό μπορεί, σε θερμοκρασία δωματίου, να απομακρυνθεί πολύ πιο εύκολα από ότι ένα ηλεκτρόνιο στον ενδογενή ημιαγωγό. Επειδή το πεντασθενές στοιχείο πρόσμιξης "δίνει" στον ημιαγωγό ηλεκτρόνια, ονομάζεται δότης (donor). Στη συνέχεια το άτομο της πρόσμιξης ιονίζεται και αποκτά θετικό φορτίο. Επειδή η απομάκρυνση του

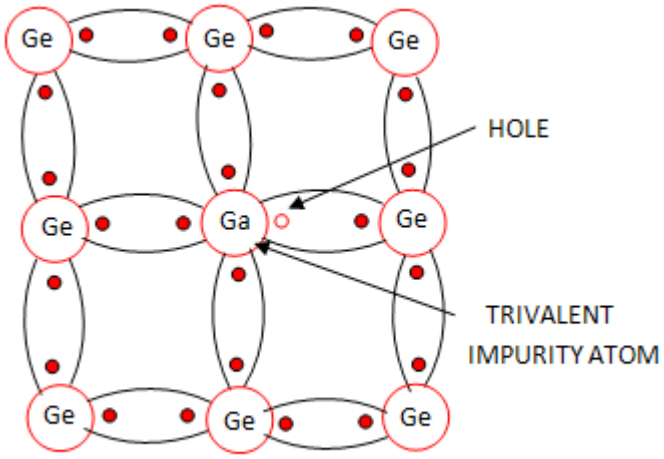
ηλεκτρονίου από το δότη είναι πολύ πιο εύκολη από ότι από ένα άτομο του ημιαγωγού, ο "δανεισμός" ενός ηλεκτρονίου από κάποιο γειτονικό άτομο θα είναι δύσκολος. Επιπλέον η πιθανότητα να βρίσκεται κοντά ένας άλλος δότης που θα μπορούσε εύκολα να "δανείσει" ένα ηλεκτρόνιο είναι αμελητέα. Αυτό έχει ως αποτέλεσμα το θετικό φορτίο να παραμένει ακίνητο στο δότη και στον ημιαγωγό να κινούνται τα ελεύθερα ηλεκτρόνια. Βέβαια, η απομάκρυνση ηλεκτρονίων από τα άτομα του ημιαγωγού δεν πρέπει να αποκλεισθεί αλλά πρέπει να σημειωθεί ότι ο αριθμός τους είναι πολύ μικρός σε σχέση με τον αριθμό των ηλεκτρονίων που προέρχονται από τους δότες. Έτσι η προσθήκη δοτών έχει ως αποτέλεσμα να υπάρχουν πολλά ελεύθερα ηλεκτρόνια και πολύ λίγες οπές στον ημιαγωγό. Συνεπώς, σε ένα ημιαγωγό τύπου N το ηλεκτρικό ρεύμα μεταφέρεται κυρίως από ένα είδος φορτίων, τα ηλεκτρόνια, τα οποία ονομάζονται και φορείς πλειονότητας η πλειοψηφίας (majority carriers). Αντίθετα οι οπές στους ημιαγωγούς τύπου N ονομάζονται φορείς μειονότητας η μειοψηφίας (minority carriers). Τέλος, η αύξηση της συγκέντρωσης των δοτών σε ένα ημιαγωγό έχει ως αποτέλεσμα την αύξηση της συγκέντρωσης των ηλεκτρονίων και συνεπώς της αγωγιμότητας του.



Εικόνα 49: Ημιαγωγός Τύπου N

- Οι ημιαγωγοί τύπου P δημιουργούνται όταν σε έναν ημιαγωγό όπως, το πυρίτιο η το γερμάνιο, προστεθεί πολύ μικρή ποσότητα ενός στοιχείου της τρίτης ομάδας του περιοδικού πίνακα. Τα στοιχεία που χρησιμοποιούνται συνήθως ως προσμίξεις είναι το βάριο, το γάλλιο και το ίνδιο ενώ η ποσότητα που απαιτείται είναι, όπως και στους ημιαγωγούς τύπου N, της τάξης των μερικών μερών στο εκατομμύριο. Τα άτομα της πρόσμιξης καταλαμβάνουν θέσεις των ατόμων του ημιαγωγού. Επειδή τα άτομα της τρίτης ομάδας του περιοδικού πίνακα έχουν τρία ηλεκτρόνια στη στοιβάδα σθένους, όταν καταλάβουν μία θέση ενός ατόμου του ημιαγωγού, θα χρησιμοποιήσουν όλα τα ηλεκτρόνια σθένους για το σχηματισμό ομοιοπολικών δεσμών. Έτσι θα παραμείνει ένα γειτονικό άτομο του ημιαγωγού, το οποίο θα απαιτεί ένα ηλεκτρόνιο για να σχηματίσει την πλήρη δομή των οκτώ ηλεκτρονίων στην εξωτερική στιβάδα του. Το απαιτούμενο αυτό ηλεκτρόνιο θα το "δανειστεί από κάποιο γειτονικό άτομο του ημιαγωγού. Το ηλεκτρόνιο που θα καταλάβει, με αυτό τον τρόπο, την κενή θέση θα ιονίσει με αρνητικό φορτίο το άτομο της τρίτης ομάδας

του περιοδικού πίνακα. Η διαδικασία αυτή αντιστοιχεί με την "απελευθέρωση" μιας οπής και επειδή τα άτομα αυτά αποδέχονται ένα ηλεκτρόνιο ονομάζονται αποδέκτες (acceptors). Η δομή των δεσμών ενός αποδέκτη κάνει πολύ πιο εύκολη την απελευθέρωση μιας οπής από ότι μπορεί να συμβεί σε ένα άτομο του ημιαγωγού σε θερμοκρασία δωματίου. Το ηλεκτρόνιο που έχει καλύψει το έλλειμμα του δεσμού παραμένει ακίνητο στον αποδέκτη. Όπως συμβαίνει στους ημιαγωγούς τύπου Ν έτσι και στους ημιαγωγούς τύπου Ρ υπάρχουν και ελεύθερα ηλεκτρόνια, των οποίων όμως η συγκέντρωση είναι πολύ μικρότερη από αυτήν των οπών. Έτσι η προσθήκη αποδεκτών έχει ως αποτέλεσμα να υπάρχουν πολλές οπές και πολύ λίγα ελεύθερα ηλεκτρόνια στον ημιαγωγό. Σε έναν ημιαγωγό τύπου Ρ οι φορείς πλειονότητας είναι οι οπές, ενώ οι φορείς μειονότητας είναι τα ηλεκτρόνια. Τέλος, η αύξηση της συγκέντρωσης των αποδεκτών σε ένα ημιαγωγό έχει ως αποτέλεσμα την αύξηση της συγκέντρωσης των οπών και συνεπώς της αγωγιμότητας του.



Εικόνα 50: Ημιαγωγός Τύπου Ρ

Β) Διαδικασία Μετατροπής Ηλιακής Ενέργειας σε Ηλεκτρική:

Τα ηλιακά στοιχεία, που είναι δίοδοι ημιαγωγού με τη μορφή δίσκου, δέχονται ηλιακή ακτινοβολία. Κάθε φωτόνιο της ακτινοβολίας φέρει ενέργεια που δίδεται απ' τη σχέση:

$$E = h \cdot f = (h \cdot c) / \lambda \qquad (21)$$

όπου:

- h: η σταθερά του Plank που ισούται με $6{,}3 \cdot 10^{-34}$ J
- f: η συχνότητα εκπομπής
- c: η ταχύτητα διάδοσης
- λ: το μήκος κύματος

Αν η ενέργεια είναι ίση ή μεγαλύτερη από το Ενεργειακό Διάκενο του ημιαγωγού (E_g), τότε υπάρχει η δυνατότητα να απορροφηθεί από ένα χημικό δεσμό και να ελευθερώσει ένα ηλεκτρόνιο. Απ' την προηγούμενη εξίσωση συμπεραίνουμε ότι για ένα συγκεκριμένο υλικό και επομένως για ένα συγκεκριμένο E_g, υπάρχει μία μέγιστη τιμή λ_g πάνω απ' την οποία το φωτοηλεκτρικό φαινόμενο δεν παρατηρείται. Αν το Ενεργειακό διάκενο είναι σε μονάδες eV και το λ μετριέται σε μm, τότε θα ισχύει:

$$\lambda_g = 1{,}238 / E_g \qquad (22)$$

Για ακτινοβολία που περιέχει φωτόνια με μήκη κύματος μικρότερα απ' την παραπάνω τιμή, δημιουργείται μια περίσσεια από ζεύγη φορέων πέρα από τις συγκεντρώσεις που αντιστοιχούν στις συνθήκες ισορροπίας. Οι φορείς αυτοί, καθώς κυκλοφορούν στο στερεό, ενδέχεται να βρεθούν στην περιοχή της επαφής p-n (που δημιουργείται στην επιφάνεια επαφής ενός ημιαγωγού τύπου p και ενός ημιαγωγού τύπου n), οπότε θα δεχθούν την επίδραση του ηλεκτροστατικού της πεδίου. Με τον τρόπο αυτό, τα ελεύθερα ηλεκτρόνια εκτρέπονται προς το τμήμα τύπου n και οι οπές εκτρέπονται προς το τμήμα τύπου p, με αποτέλεσμα να δημιουργηθεί μια διαφορά δυναμικού ανάμεσα στους ακροδέκτες των δύο τμημάτων της διόδου.

Η εκδήλωση της τάσης αυτής ανάμεσα στις δύο όψεις του φωτιζόμενου δίσκου, η οποία αντιστοιχεί σε ορθή πόλωση της διόδου, ονομάζεται φωτοβολταϊκό φαινόμενο. Η διάταξη αποτελεί μια πηγή ρεύματος που διατηρείται όσο διαρκεί η πρόσπτωση του ηλιακού φωτός πάνω στην επιφάνεια του στοιχείου. Όταν ένα φωτοβολταϊκό στοιχείο δέχεται κατάλληλη ακτινοβολία, διεγείρεται παράγοντας ηλεκτρικό ρεύμα, που η τιμή του θα είναι ανάλογη προς τα φωτόνια που απορροφά.

2.5 Φωτοβολταϊκό Πάρκο

Εικόνα 51: Φωτοβολταϊκό Πάρκο

Σ' αυτό το υποκεφάλαιο θα παραθέσουμε τις Μονάδες Εγκατάστασης Ελέγχου και Παρακολούθησης ενός Φωτοβολταϊκού Πάρκου και θα αναλύσουμε καθεμία απ' αυτές συνοπτικά.

Α) Πάνελ:

- Ένα ηλιακό πάνελ (solar panel, photovoltaic module) είναι ένα σύνολο από ηλιακές κυψέλες και πλήθος άλλων προστατευτικών και λειτουργικών επιστρώσεων τα οποία είναι εγκατεστημένα επάνω σε πλαίσιο αλουμινίου (μέσο στήριξης). Το φύλλο αυτό αποτελεί την πλάτη του πάνελ. Η εμπρός όψη των στοιχείων καλύπτεται από προστατευτικό φύλλο γυαλιού ή διαφανούς πλαστικού. Τα δύο φύλλα εμπρός και πίσω συγκρατιούνται μεταξύ τους με τη βοήθεια ταινίας από συνθετικό ελαστικό και συσφίγγονται με περιμετρικό μεταλλικό περίβλημα, κατασκευή που εξασφαλίζει την απαραίτητη μηχανική αντοχή, τις υποδοχές στήριξης και την αυξημένη στεγανότητα για προστασία από την υγρασία.
- Τα ηλιακά πάνελ μπορούν να χρησιμοποιηθούν ως μονάδες που τελικά θα συγκροτήσουν ένα μεγαλύτερο Φωτοβολταϊκό σύστημα ώστε να παράγουν και να προμηθεύουν ηλεκτρική ενέργεια για οικιακές ή εμπορικές εφαρμογές.
- Συνδέονται σε σειρά μέχρι να αποκτήσουμε το επιθυμητό αποτέλεσμα σε έξοδο ρεύματος και παράλληλα μέχρι να αποκτήσουμε την επιθυμητή ένταση.
- Όταν ένα Φωτοβολταϊκό στοιχείο δεν δέχεται την προσπίπτουσα στο Φωτοβολταϊκό πάνελ ηλιακή ακτινοβολία (μερική σκίαση) σταματάει να παράγει ηλεκτρικό ρεύμα. Επειδή όμως τα Φωτοβολταϊκά στοιχεία συνδέονται σε σειρά και μοντελοποιούνται ως δίοδοι, το στοιχείο με το μικρότερο ρεύμα βραχυκύκλωσης επιβάλλει στο πάνελ το δικό του ρεύμα και συνεπώς το παραγόμενο ηλεκτρικό ρεύμα τείνει στο μηδέν. Ταυτόχρονα, το σκιασμένο Φωτοβολταϊκό στοιχείο δέχεται το άθροισμα των τάσεων όλων των άλλων στοιχείων ανάστροφα και αρχίζει να υπερθερμαίνεται. Το φαινόμενο αυτό προκαλεί τελικά την καταστροφή του στοιχείου το οποίο αποκτά μία καφετί απόχρωση.
- Η προστασία της εγκατάστασης εξασφαλίζεται με την παράλληλη σύνδεση μιας διόδου παράκαμψης (by-pass diode) σε κάθε ομάδα Φωτοβολταϊκών στοιχείων του πλαισίου. Η δίοδος διατηρεί την ανάστροφη τάση σε μια σειρά Φωτοβολταϊκών στοιχείων σε συγκεκριμένη τιμή έτσι ώστε ακόμα και αν κάποιο στοιχείο σκιαστεί ή ακόμα και καταστραφεί να μπορεί να λειτουργεί κανονικά το πάνελ.
- Η θερμοκρασία είναι μια σημαντική παράμετρος λειτουργίας ενός φωτοβολταϊκού συστήματος. Ο συντελεστής θερμοκρασίας για την τάση ανοικτού κυκλώματος είναι κατά προσέγγιση ίσος με -2,3 mV/°C για κάθε ηλιακό στοιχείο. Ο συντελεστής τάσης μιας βασικής μονάδας είναι επομένως αρνητικός και πολύ μεγάλος από τη στιγμή που συνδέονται σε σειρά 33 έως 36 ηλιακά στοιχεία. Ο συντελεστής ρεύματος, από την άλλη πλευρά, είναι θετικός και μικρός +6 μA/°C περίπου ανά τετραγωνικό εκατοστό της βασικής μονάδας. Είναι σημαντικό να σημειώσουμε ότι η τάση καθορίζεται από τη θερμοκρασία λειτουργίας των ηλιακών στοιχείων, η οποία διαφέρει από τη θερμοκρασία περιβάλλοντος.
  Η θερμοκρασιακή κατάσταση μιας ηλιακής πυριτικής κυψέλης επηρεάζει ζωτικά την ισχύ που αποδίδει σε μια συγκεκριμένη ηλιακή ακτινοβολία. Με την ανύψωση της θερμοκρασίας της ηλιακής κυψέλης, η τάση εξόδου μειώνεται με ρυθμό περίπου 2 mV/°C. Η μείωση αυτή αντισταθμίζεται μερικώς από την αύξηση της έντασης του ρεύματος με ρυθμό περίπου 0,5 mA/°C. Τελικά η ανύψωση της θερμοκρασίας της ηλιακής κυψέλης μειώνει την αποδιδόμενη από αυτή ισχύ κατά 0,3%/°C ($P = V \cdot I$). Πρέπει, λοιπόν, να προνοήσουμε ώστε το σύστημα να υφίσταται ένα σωστό αερισμό.
- Τύποι Πάνελ:

    - Μονοκρυσταλλικό πυρίτιο: Υψηλό κόστος κατασκευής, βαθμός απόδοσης της κυψέλης 16- 18%, βαθμός απόδοσης του πλαισίου 12-14%.
    - Πολυκρυσταλλικό πυρίτιο: Χαμηλότερο κόστος κατασκευής, αλλά και μικρότερος βαθμός απόδοσης σε σχέση με το μονοκρυσταλλικό.
    - Άμορφο πυρίτιο: Πολύ χαμηλό κόστος κατασκευής αλλά ο βαθμός απόδοσης είναι περίπου ο μισός (8-9% ανα κυψέλη) του κρυσταλλικού.
    - Λεπτά υμένια: Νέος τύπος από υλικά όπως Ινδιο, Κάδμιο, Τελλούριο, Αρσενικούχο Γάλλιο.
- Η ισχύς των Φωτοβολταϊκών Πάνελ, δίνεται σαν μέγιστη ισχύς από τους κατασκευαστές σε κάποιες συνθήκες οι οποίες ονομάζονται «τυποποιημένες συνθήκες ελέγχου» (standard testing conditions, STC). Αυτές είναι:
    - Θερμοκρασία κυψέλης: 25°C
    - Ηλιακή ακτινοβολία στο επίπεδο του πάνελ: 1000 W/m^2
    - AM 1.5: αντιστοιχεί σε φάσμα ηλιακής ακτινοβολίας όταν ο ήλιος είναι 45° επάνω από τον ορίζοντα. Όταν ο ήλιος είναι στο μέγιστο σημείο του έχουμε AM1.
- Στο βόρειο ημισφαίριο τα πάνελ τοποθετούνται προς το νότο ενώ στο νότιο προς το βορρά.
- Η ένταση της προσπίπτουσας ηλιακής ακτινοβολίας σε μία επιφάνεια είναι μέγιστη όταν η επιφάνεια είναι κάθετη προς την κατεύθυνση της ακτινοβολίας δηλαδή όταν η γωνία πρόσπτωσης είναι 0˚. Όμως, δεδομένου ότι ο ήλιος συνεχώς μετακινείται κατά τη διάρκεια της ημέρας η συγκεκριμένη συνθήκη εξασφαλίζεται μόνο με κινητές βάσεις στήριξης.
- Απαιτείται αξιολόγηση της περιοχής και προσεκτική επιλογή του χώρου εγκατάστασης προκειμένου να εξασφαλιστεί η βέλτιστη απόδοση του συστήματος καθώς ακόμα και η παραμικρή σκίαση επηρεάζει σε μεγάλο ποσοστό την ενεργειακή απόδοση του πάνελ. Ανεπιθύμητη σκίαση μπορεί να προκύψει από δέντρα, θάμνους, οικίσκους, κολώνες κλπ. Μεγαλύτερο πρόβλημα σκίασης για το βόρειο ημισφαίριο παρουσιάζεται κατά τους χειμερινούς μήνες όπου το ύψος του ήλιου είναι χαμηλότερα και οι σκιές γίνονται μακρύτερες.
  Ένας πρακτικός κανόνας είναι να εξασφαλίσουμε ότι η ελάχιστη απόσταση μεταξύ των πάνελ και του εμποδίου να είναι ίση με το διπλάσιο του ύψους του εμποδίου.

Β) Φωτοβολταϊκή γεννήτρια:

Το βασικό συστατικό κάθε φωτοβολταϊκής εγκατάστασης είναι η Φ/Β γεννήτρια, που αποτελείται από τους ηλιακούς συλλέκτες με τα Φ/Β ηλιακά στοιχεία.

Γ) Μπαταρίες:

Συσσωρευτές (μπαταρίες): Ένα αξιόπιστο σύστημα πρέπει να παρέχει επαρκή ηλεκτρική ενέργεια για την ικανοποίηση της ζήτησης και στα χρονικά διαστήματα που δεν υπάρχει αντίστοιχη ηλιακή ακτινοβολία (νυχτερινές ώρες, συννεφιασμένες ημέρες και χρονικές αιχμές της κατανάλωσης). Τα Φωτοβολταϊκά συστήματα που είναι συνδεδεμένα με κεντρικά ηλεκτρικά δίκτυα διανομής, αντλούν από αυτά την απαιτούμενη συμπληρωματική ηλεκτρική ενέργεια. Επίσης διοχετεύουν προς τα δίκτυα την ενδεχόμενη περίσσεια της παραγόμενης Φωτοβολταϊκής ηλεκτρικής ενέργειας, όταν υπερβαίνει την κατανάλωση του συστήματος. Όμως τα απομονωμένα αυτόνομα Φωτοβολταϊκά συστήματα δεν έχουν αυτή τη δυνατότητα ενεργειακής ανταλλαγής. Επομένως, χρειάζεται να αποθηκεύουν μια

ποσότητα από την περίσσεια της ηλεκτρικής τους παραγωγής, ώστε να χρησιμοποιηθεί όταν η ζήτηση είναι μεγαλύτερη από την παραγωγή της γεννήτριας. Αυτόν, ακριβώς, το ρόλο διαδραματίζει η μπαταρία.

Ως προς την απαλλαγή του συστήματος από την περίσσεια της παραγόμενης ηλεκτρικής ενέργειας, πέρα από τη ζήτηση της κατανάλωσης και τη δυνατότητα της αποθήκευσης, αυτή αντιμετωπίζεται με τη διοχέτευσή της στη γη ή σε ηλεκτρικές αντιστάσεις.

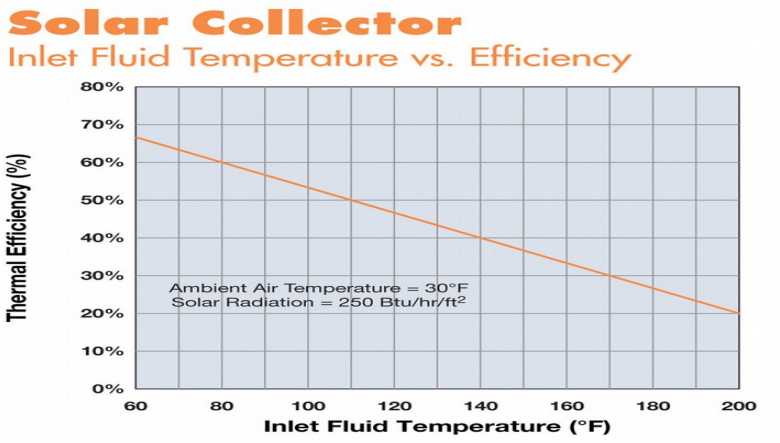
Δ) Ελεγκτής φόρτισης (Charge Controller):

Ρόλος του είναι να παρέχει στις μπαταρίες ηλεκτρική ενέργεια από το ηλιακό πλαίσιο, με έναν τρόπο που αποτρέπει το ηλιακό πάνελ από την υπερφόρτωση των μπαταριών.

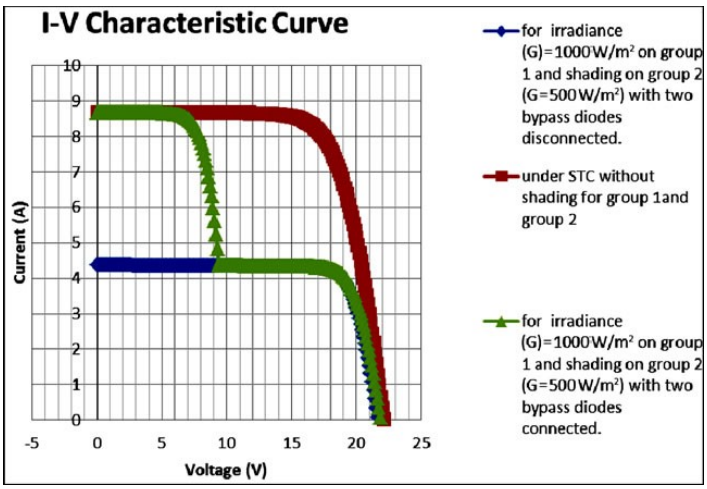
Ε) Μετατροπείς (Inverters):

Μία Φωτοβολταϊκή συστοιχία, ανεξάρτητα από το μέγεθος και το είδος της κατασκευής της, μπορεί να παράγει ηλεκτρική ενέργεια μόνο συνεχούς ρεύματος. Υπάρχουν εφαρμογές για τις οποίες το συνεχές ρεύμα είναι κατάλληλο όπως για παράδειγμα η φόρτιση μπαταριών η οποία μπορεί να πραγματοποιηθεί συνδέοντας την μπαταρία απευθείας σε Φωτοβολταϊκό πλαίσιο χωρίς να μεσολαβεί κάποιο στάδιο μετατροπής της ενέργειας. Αντίθετα, σε περιπτώσεις που το σύστημα τροφοδοτεί φορτία εναλλασσομένου ρεύματος ή παρέχει ενέργεια στο ηλεκτρικό δίκτυο (230 Vrms/ 50Hz) απαιτείται η χρήση inverter. Οι Inverters μετατρέπουν τη συνεχή τάση (DC) που παράγεται από τη γεννήτρια σε εναλλασσόμενη (AC) τάση. Διαιρούνται σε δύο κατηγορίες:

- Inverters αυτόνομων Φωτοβολταϊκών συστημάτων
- Inverters συνδεδεμένοι στο δίκτυο



Εικόνα 52: Inlet Fluid Temperature vs Efficiency



Εικόνα 53: by-pass diode

Στ) Σήματα:

Σε κάθε εφαρμογή ελέγχου, ελέγχεται ένας αριθμός σημάτων με τη χρήση κατάλληλων αισθητήριων στοιχείων, ανάλογα με τα παρατηρήσιμα φυσικά μεγέθη. Το αισθητήριο δίνει κάποιου τύπου ψηφιακό (δυαδικό ή διακριτό) ή αναλογικό (συνεχές) σήμα, ή πιο πολύπλοκα σήματα με τη μορφή κάποιου πρωτοκόλλου, πχ σειριακό. Ο τύπος του σήματος εξαρτάται από το παρατηρούμενο φυσικό μέγεθος και το αισθητήριο που χρησιμοποιείται [CITATION Bro96 \l 1033]. Οι ελεγκτές ενός συστήματος ελέγχου και αυτοματισμού μπορούν να ελέγχουν και να παρακολουθούν ταυτόχρονα ψηφιακά, αναλογικά και σειριακά δεδομένα από αισθητήρες, ενεργοποιητές και συσκευές. Στην εφαρμογή ελέγχου και διαχείρισης ενέργειας που υλοποιήθηκε χρησιμοποιήθηκαν τα εξής σήματα:

α) Σήμα ενέργειας: Το σήμα της ενέργειας είναι αναλογικό και παράγεται από το γινόμενο του ηλεκτρικού δυναμικού με την ένταση κάθε δεδομένη χρονική στιγμή. Ένα κελί μπορεί να παράγει περίπου 0,5V τάση ανεξαρτήτως μεγέθους. Τα περισσότερα solar panels αποτελούνται από 24 cells που παράγουν 12V τάση. Το ρεύμα που παράγει ένα τέτοιο πάνελ είναι συνήθως 4 Amperes. Τα πάνελ όταν συνδέονται μεταξύ τους σειριακά, αυξάνεται η τελική παραγόμενη τάση και ισοδυναμεί με το άθροισμά τους ενώ όταν συνδέονται παράλληλα, η τάση μένει η ίδια και αυξάνεται το ρεύμα το οποίο προκύπτει από το άθροισμά τους.

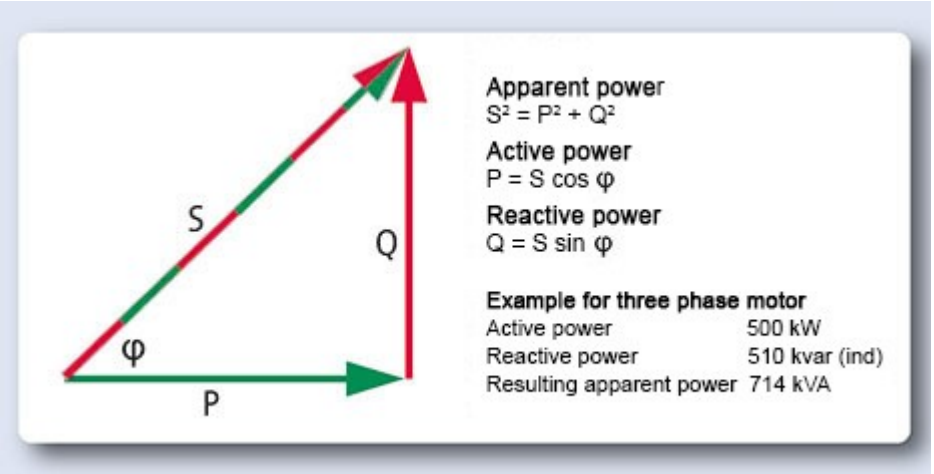
β) Σήμα θερμοκρασίας: Ένας αισθητήρας μετρά την θερμοκρασία της εγκατάστασης. Το σήμα είναι αναλογικό και χρησιμοποιείται για τον έλεγχο της ομαλής λειτουργίας του συστήματος σε ακραίες συνθήκες περιβάλλοντος. Η ιδανική θερμοκρασία για τα φωτοβολταϊκά πάρκα είναι οι 25°C (77°F). Όταν η θερμοκρασία αυξάνεται το ρεύμα (current) αυξάνεται ελάχιστα, ενώ η τάση (voltage) μειώνεται πολύ γρήγορα. Αυτό έχει ως αποτέλεσμα μια γενικά χαμηλότερη απόδοση ενέργειας. Ένας γενικός κανόνας είναι ότι η απόδοση ενός κελιού πέφτει κατά 0,3% για κάθε 1°C πάνω από τους 25°C. Επιπλέον, ένας αισθητήρας μετρά τη θερμοκρασία περιβάλλοντος.

γ) Σήμα υγρασίας: Ένας αισθητήρας μετρά την υγρασία του περιβάλλοντος. Πρόκειται για αναλογικό σήμα και χρησιμοποιείται για τον έλεγχο της ομαλής λειτουργίας του συστήματος σε ακραίες συνθήκες περιβάλλοντος.

δ) Σήμα ανέμου: Ένας αισθητήρας μετρά τον άνεμο του περιβάλλοντος. Πρόκειται για αναλογικό σήμα και χρησιμοποιείται για τον έλεγχο της ομαλής λειτουργίας του συστήματος σε ακραίες συνθήκες περιβάλλοντος.

ε) Σήματα εναλλασσόμενου ρεύματος: Το εναλλασσόμενο ρεύμα αποτελείται από τρία συστατικά:

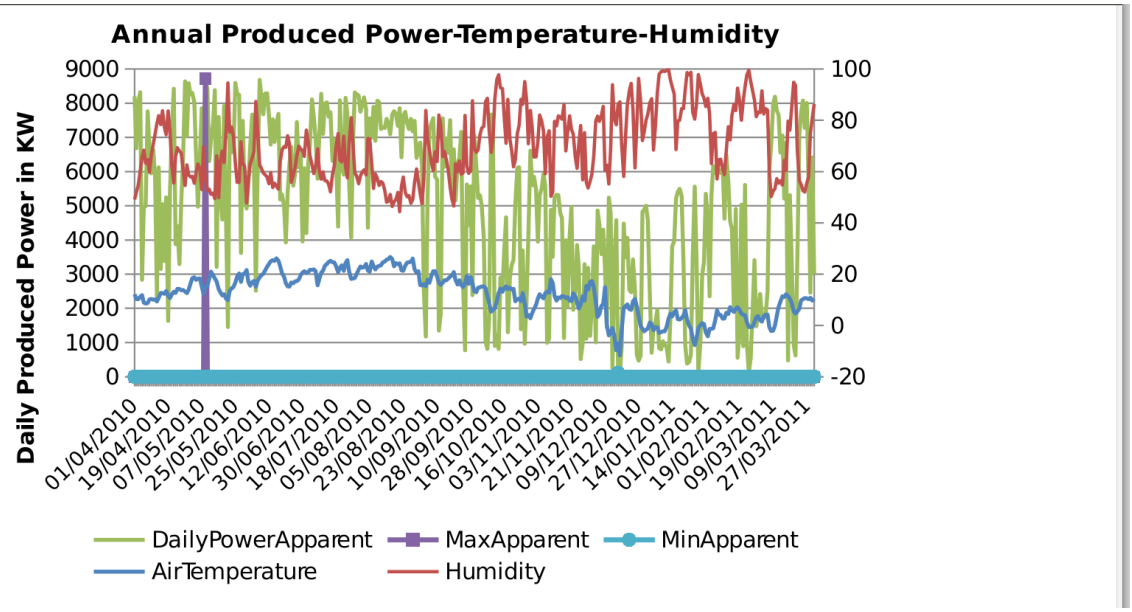
- Real/Active Power: Η ποσότητα που καταναλώνεται. Μετριέται σε Watts.
- Apparent Power: Η ποσότητα ενέργειας που παραδίδεται από μία πηγή σε ένα φορτιστή. Πρέπει να είναι πάντα μεγαλύτερη από αυτή που χρειάζεται μία συσκευή για να δουλέψει. Μετριέται σε Volt-Amperes (va).
- Reactive Power: Η ποσότητα ενέργειας που επιστρέφει στη πηγή σε κάθε κύκλο, λόγω της αποθηκευμένης ενέργειας. Μετριέται σε reactive Volt-Amperes (var).



Εικόνα 54: Εναλλασσόμενο Ρεύμα

Η αναλογία της ενέργειας που καταναλώνεται προς την ενέργεια που παράγεται από μία πηγή ονομάζεται συντελεστής δύναμης (power factor - PF). Ο συντελεστής μπορεί να έχει και αρνητική τιμή στη περίπτωση που παράγεται ενέργεια η οποία τροφοδοτεί το δίκτυο. Το σύστημα που έχει υλοποιηθεί έχει δοκιμαστεί με πραγματικά δεδομένα. Αυτά έχουν συλλεχθεί κατά τη διάρκεια περιόδου ενός έτους από μια πραγματική εγκατάσταση φωτοβολταϊκών που συνέλλεγε και παρακολουθούσε τα παραπάνω σήματα. Στην εικόνα που παρατίθεται παρακάτω απεικονίζεται η ετήσια ενέργεια που παράγεται στο πάρκο σε σχέση με την εξωτερική θερμοκρασία και την υγρασία στο χώρο του πάρκου. Η μέγιστη απόδοση παρουσιάζεται το Μάιο ενώ η ελάχιστη στα μέσα του Δεκέμβρη. Η απόδοση του πάρκου αυξάνεται κατά τη διάρκεια των καλοκαιρινών μηνών λόγω του ευνοϊκού κλίματος. Προφανώς η μεταβολή της απόδοσης είναι ανάλογη με τη μεταβολή της θερμοκρασίας. Η

ελάχιστη απόδοση έχει καταγραφεί κατά τη διάρκεια της ημέρας του χρόνου με την ελάχιστη θερμοκρασία. Βέβαια, δε συμβαίνει το ίδιο με την μέγιστη απόδοση. Αυτό σημαίνει ότι η θερμοκρασία δεν είναι ο μοναδικός παράγοντας που επηρεάζει την απόδοση ενός ηλιακού πάρκου. Τέλος, παρατηρείται ότι η συνολική ετήσια ενέργεια που φωτοβολταϊκό πάρκο έφθασε τα 1,82 MW.

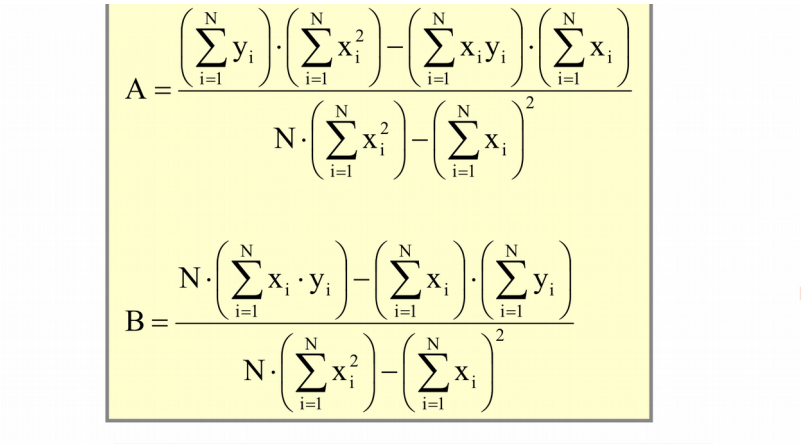


Εικόνα 55: Ετήσια Παραγόμενη Ενέργεια - Θερμοκρασία – Υγρασία

2.6 Μέθοδος Ελαχίστων Τετραγώνων

Έστω ότι έχουμε κάποιες μετρήσεις και κάθε μέτρηση αντιστοιχεί σ' ένα σημείο $\Sigma_i(x_i, y_i)$ σε ένα δισδιάστατο, ορθοκανονικό σύστημα αξόνων (X, Y). Για να χαράξουμε την ευθεία που ικανοποιεί αυτά τα σημεία με τον καλύτερο δυνατό τρόπο, χρησιμοποιούμε τη Μέθοδο Ελαχίστων Τετραγώνων.

Οι παράμετροι A, B της ευθείας $y = A + (B \cdot x)$, υπολογίζονται απ΄ τις εξισώσεις που παρουσιάζονται στην παρακάτω εικόνα:



$$A = \frac{\left(\sum_{i=1}^{N} y_i\right) \cdot \left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i y_i\right) \cdot \left(\sum_{i=1}^{N} x_i\right)}{N \cdot \left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)^2}$$

$$B = \frac{N \cdot \left(\sum_{i=1}^{N} x_i \cdot y_i\right) - \left(\sum_{i=1}^{N} x_i\right) \cdot \left(\sum_{i=1}^{N} y_i\right)}{N \cdot \left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)^2}$$

Εικόνα 56: Υπολογισμός Παραμέτρων Ευθείας Γραμμικής Παλινδρόμησης

2.7 Σφάλματα

Όταν αναφερόμαστε σε πειραματικές διαδικασίες και μετρήσεις που είναι αποτέλεσμα αυτών, πάντοτε πρέπει να λαμβάνουμε υπόψη μας τα σφάλματα που τις συνοδεύουν και βάσει των οποίων αξιολογούνται. Έτσι, αν σ' ένα πείραμα ελεύθερης πτώσης σώματος Α από ύψος h = 5 m απ' το επίπεδο της θάλασσας, προκύψει g = 9.81 m/s^2 και το αντίστοιχο σφάλμα είναι δg = 4.40 m/s^2, συνεπάγεται ότι η g παίρνει τιμές στο διάστημα Δ, όπου Δ = [5.41, 14.21] και το σχετικό σφάλμα ισούται με n= δg/ g $\simeq$ 45 %. Αυτό σημαίνει πως η μέτρησή μας είναι ανακριβής, παρόλο που η πειραματική τιμή της σταθεράς g ταυτίζεται με τη θεωρητική.

Γενικά, ποσοστά σχετικού σφάλματος μικρότερα από 10 % καθιστούν τα αποτελέσματά μας έγκυρα και αξιοποιήσιμα. Ωστόσο, η κάθε περίπτωση είναι διαφορετική και εν τέλει η ίδια η φύση του προβλήματος είναι αυτή που ορίζει ποιες τιμές σχετικού σφάλματος θα γίνουν αποδεκτές.

Παρακάτω, παραθέτουμε συνοπτικά μερικά απ' τα σημαντικότερα είδη σφαλμάτων (p_i = οι τιμές απ' τις προβλέψεις μας ή απ' τις πειραματικές μας μετρήσεις):

- Mean Error:

$$ME = [\sum_{i=1}^{n}(y_i - p_i)] / n \qquad (23)$$

- Mean Squared Error:

$$MSE = [\sum_{i=1}^{n}(y_i - p_i)^2] / n \qquad (24)$$

  - Επηρεάζεται πολύ απ' τα μεγάλα σφάλματα λόγω του τετραγώνου.

- Root Mean Squared Error:

$$RMSE = \{[\sum_{i=1}^{n}(y_i - p_i)^2]/ n\}^{1/2} \qquad (25)$$

  - Έχει ίδιες μονάδες με το μέγεθος που εξετάζουμε.
  - Αν π.χ. RMSE = 3 και η πραγματική τιμή y = 8, τότε το μοντέλο μας θα προβλέψει p = 8 ± 3.

- Squared Correlation (R^2):

$$R^2 = [\sum_{i=1}^{n}(y_i - \langle d \rangle)^2] / [\sum_{i=1}^{n}(d_i - \langle d \rangle)^2] \qquad (26)$$

  όπου d_i = τα δεδομένα μας (ή οι μετρήσεις μας) και $(y_i - \langle d \rangle)$ είναι η απόσταση του σημείου της ευθείας παλινδρόμησης y_i μείον τη μέση τιμή των δεδομένων <d>.

  - $R \in [0,1]$
  - Δείχνει την ποιότητα προσαρμογής της παλινδρόμησης στα δεδομένα.
  - Αν $R^2 = 1 \Rightarrow$ τέλεια εφαρμογή.

- Absolute Error:

$$AE = \sum_{i=1}^{n}|y_i - p_i| / n \qquad (27)$$

- Relative Error:

$$RE = \sum_{i=1}^{n}|p_i - d_i| / (\sum_{i=1}^{n}d_i / n) \qquad (28)$$

ΚΕΦΑΛΑΙΟ 3: ΠΕΙΡΑΜΑΤΙΚΗ ΔΙΑΔΙΚΑΣΙΑ

3.1 Εγκατάσταση του Rapid Miner Studio

Για την εγκατάσταση του Rapid Miner Studio προαπαιτείται η εγκατάσταση της Java 8. Στη διεύθυνση:
https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html μπορούμε να επιλέξουμε την έκδοση της Java που είναι κατάλληλη για το λειτουργικό μας σύστημα. Σε περίπτωση που χρησιμοποιούμε Linux, η εγκατάσταση μπορεί να πραγματοποιηθεί μέσω Τερματικού εκτελώντας την εντολή **su**, πληκτρολογώντας το Root Password και τέλος με τη χρήση της εντολής **apt-get install oracle-java8-installer**.

Εφόσον η εγκατάσταση ήταν επιτυχής, πηγαίνουμε στην ιστοσελίδα: https://rapidminer.com/get-started/ και αφού συμπληρώσουμε τα στοιχεία μας και επιλέξουμε το λειτουργικό σύστημα που μας ενδιαφέρει, κατεβάζουμε το αρχείο: rapidminer_studio_9.1.0.zip.

Στο φάκελο rapidminer-studio ανοίγουμε το τερματικό και εκτελούμε την εντολή: **./RapidMiner-Studio.sh**. Αφού συμπληρώσουμε ότι μας ζητάει, μπορούμε πλέον να χρησιμοποιήσουμε το Rapid Miner Studio για την διεξαγωγή της Πειραματικής Διαδικασίας.

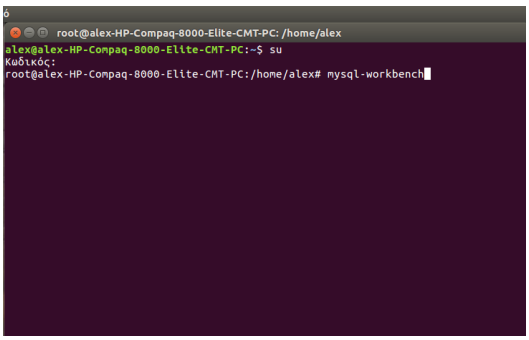
3.2 Εγκατάσταση του Rapid Miner Server

Προαπαιτούμενα:

- Δημιουργία Schema και Database στο MySQL Workbench
- Java 8

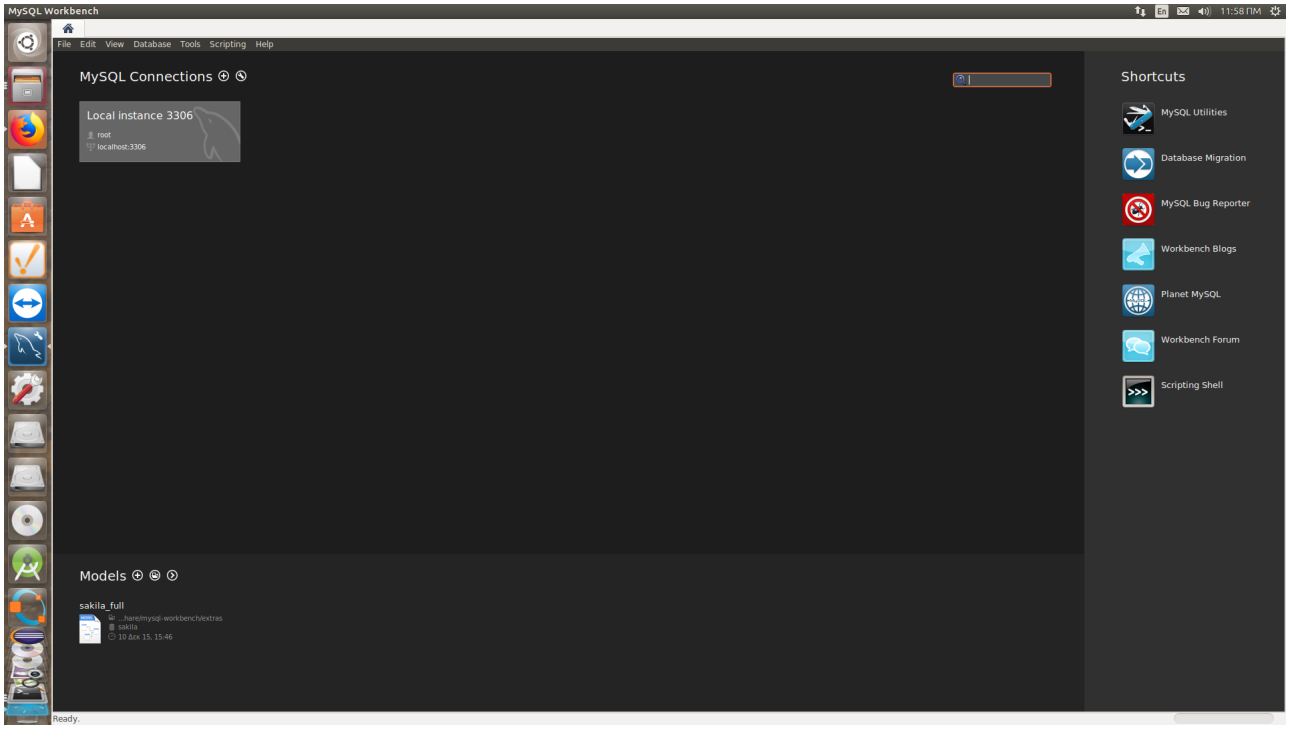
3.2.1 MySQL Workbench

Πηγαίνουμε στην ιστοσελίδα: https://dev.mysql.com/downloads/mysql/ και κατεβάζουμε τον MySQL Community Server 8.0.13. Ανοίγουμε το Τερματικό και εκτελούμε **su** → **MySQL workbench**.



Εικόνα 57: Εκτέλεση MySQL Workbench μέσω Τερματικού

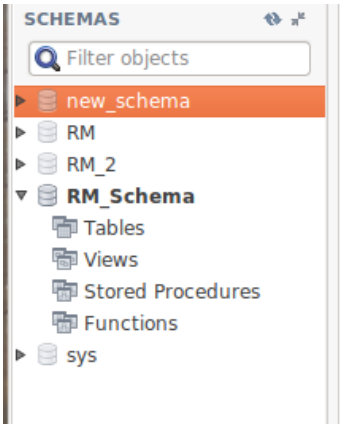
Κατά αυτόν τον τρόπο ανοίγουμε την πλατφόρμα του MySQL Workbench.



Εικόνα 58: MySQL Workbench

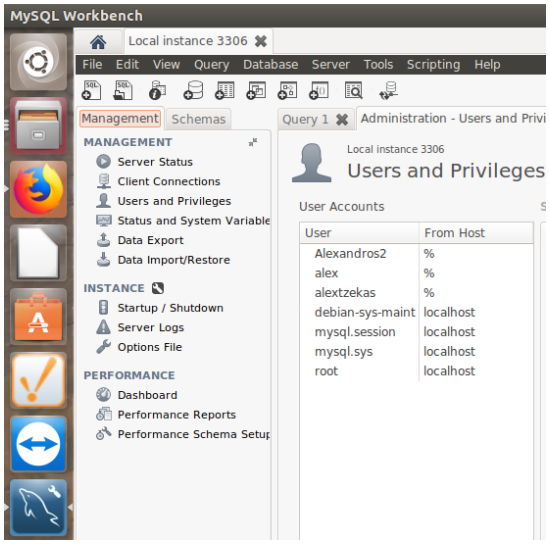
Ακολουθούμε τα παρακάτω Βήματα:

- Πηγαίνουμε στον κατάλογο με τα Schemas και φτιάχνουμε ένα νέο Shema.



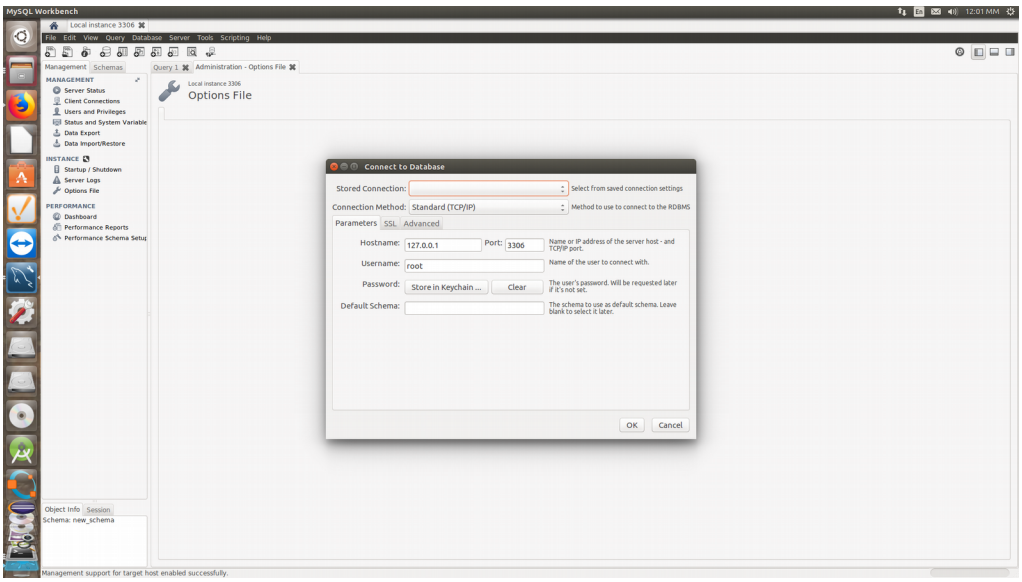
Εικόνα 59: MySQL Workbench Schemas

- Πηγαίνουμε στον κατάλογο Management/ Users and Priviledges και δημιουργούμε ένα νέο user τον οποίον και συσχετίζουμε με το νέο Shema.



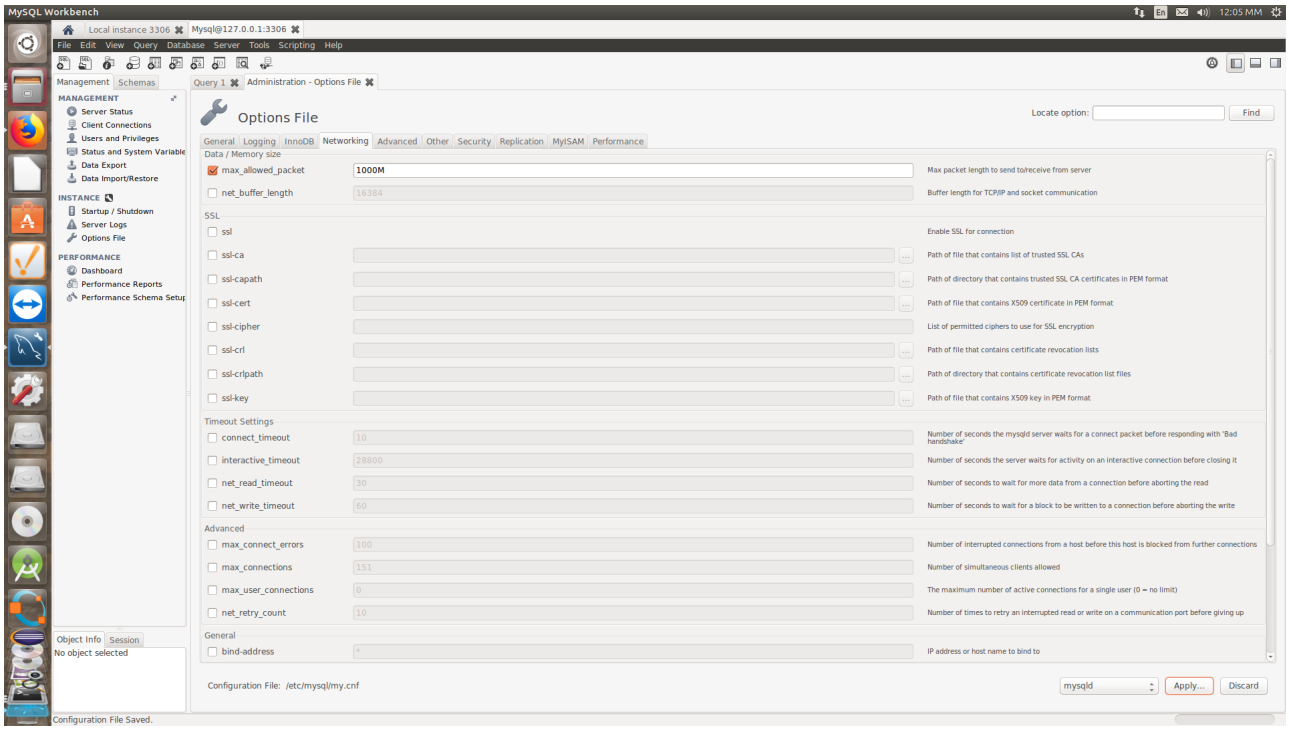
Εικόνα 60: MySQL Workbench Schemas Users & Priviledges

- Συνδεόμαστε με Database.



Εικόνα 61: MySQL Workbench Schemas Connect to Database

- Πηγαίνουμε Options File/ Networking και επιλέγουμε το max_allowed_packet > = 256M.



Εικόνα 62: MySQL Workbench Schemas Users & Priviledges

3.2.2 Java 8

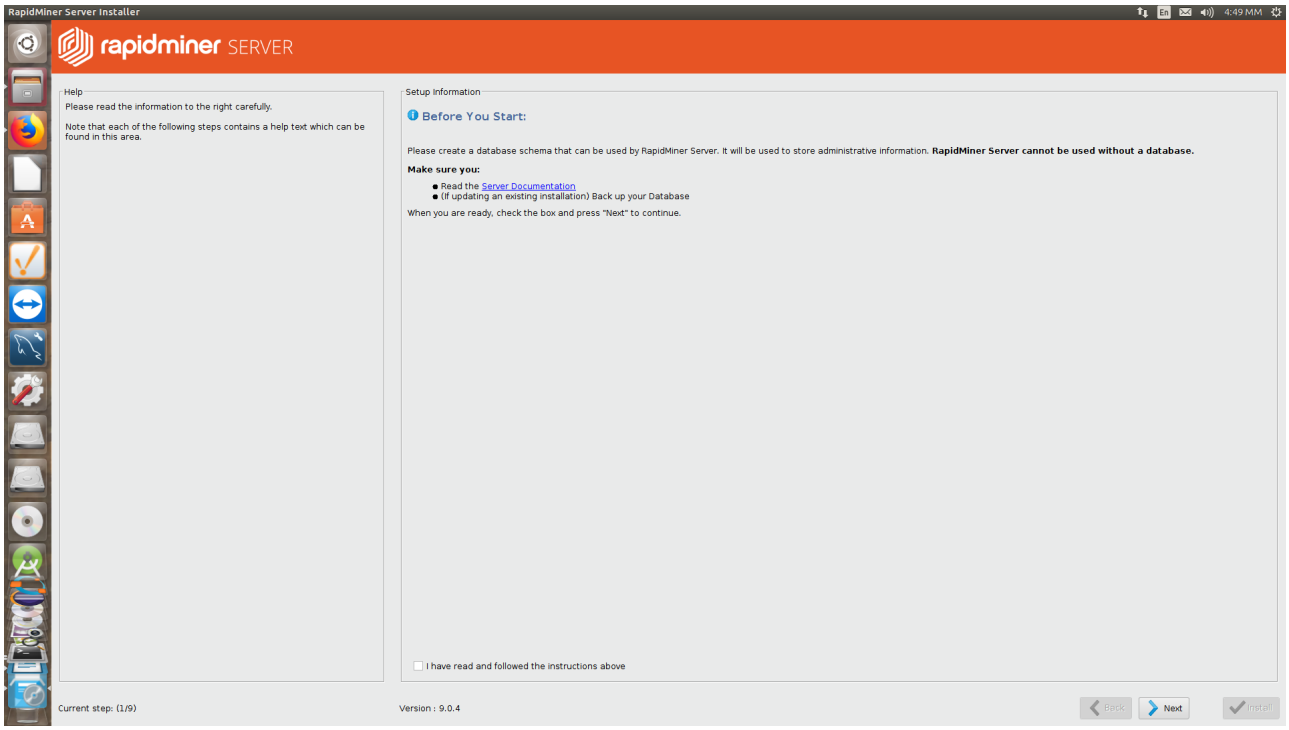
Έχουμε ήδη εγκαταστήσει τη Java 8, ωστόσο για να εντοπίσει την τοποθεσία της ο Rapid Miner Installer, τον οποίον τρέχουμε ως Root, θα πρέπει να αντιγράψουμε τον *mysql-connector-java-8.0.11.jar* στον Root. Αυτό επιτυγχάνεται στο Τερματικό με την εκτέλεση των παρακάτω εντολών:

- **su**
- **cp mysql-connector-java-8.0.11.jar /root/**

3.2.3 Rapid Miner Installer

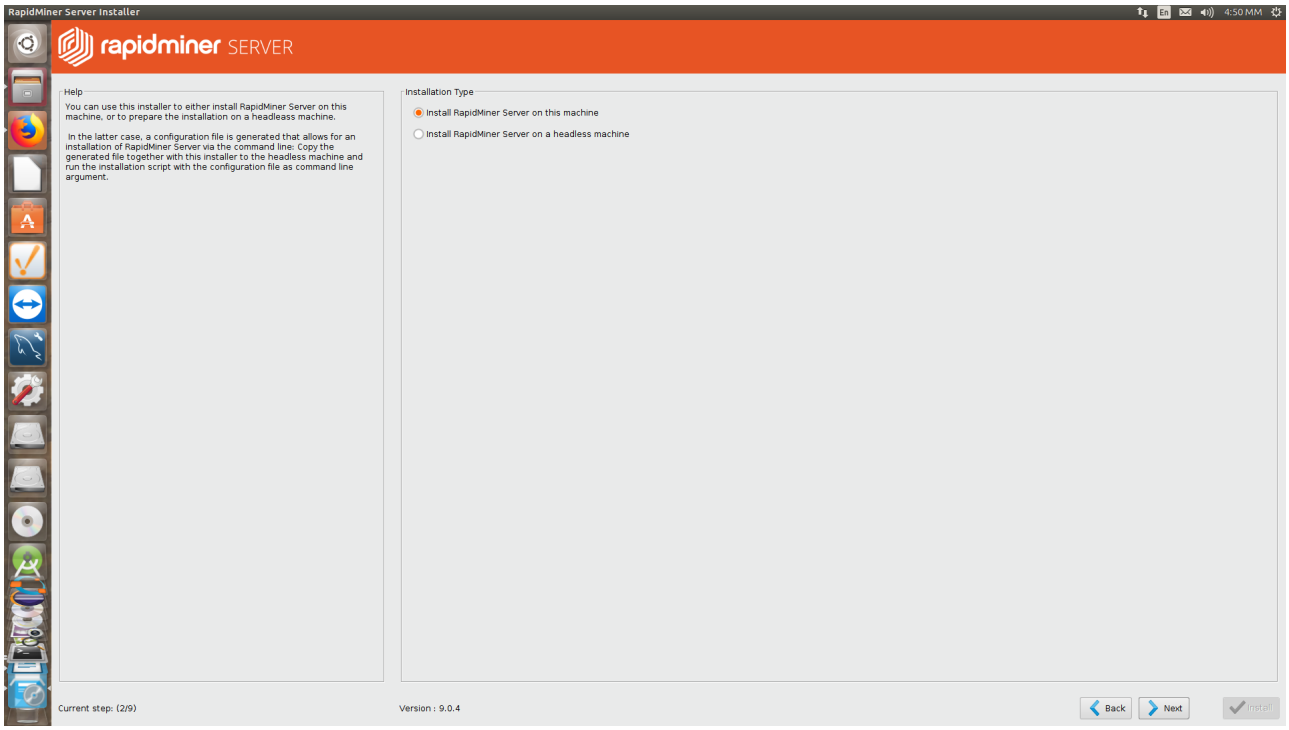
Παρακάτω παραθέτουμε σε Εικόνες τη διαδικασία εγκατάστασης του Rapid Miner Server.

Βήμα 1:



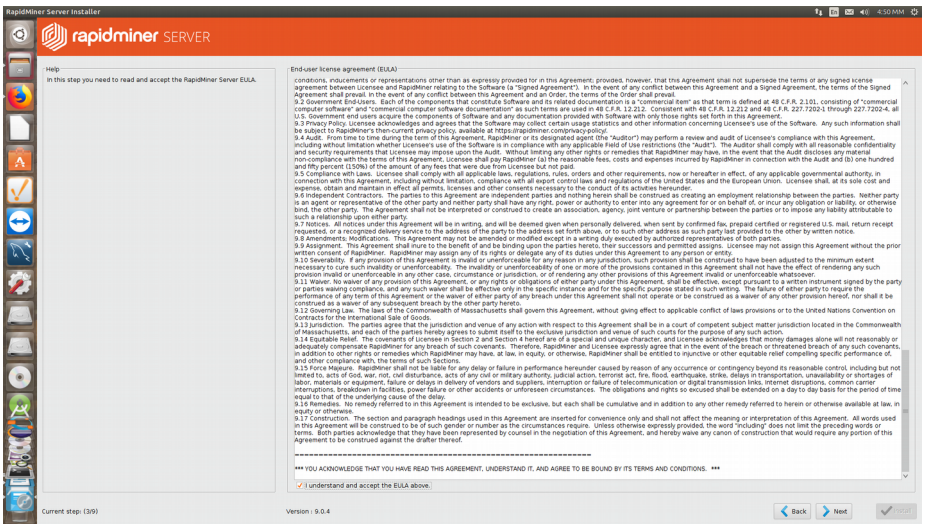
Εικόνα 63: Rapid Miner Server Installer 1

Βήμα 2:



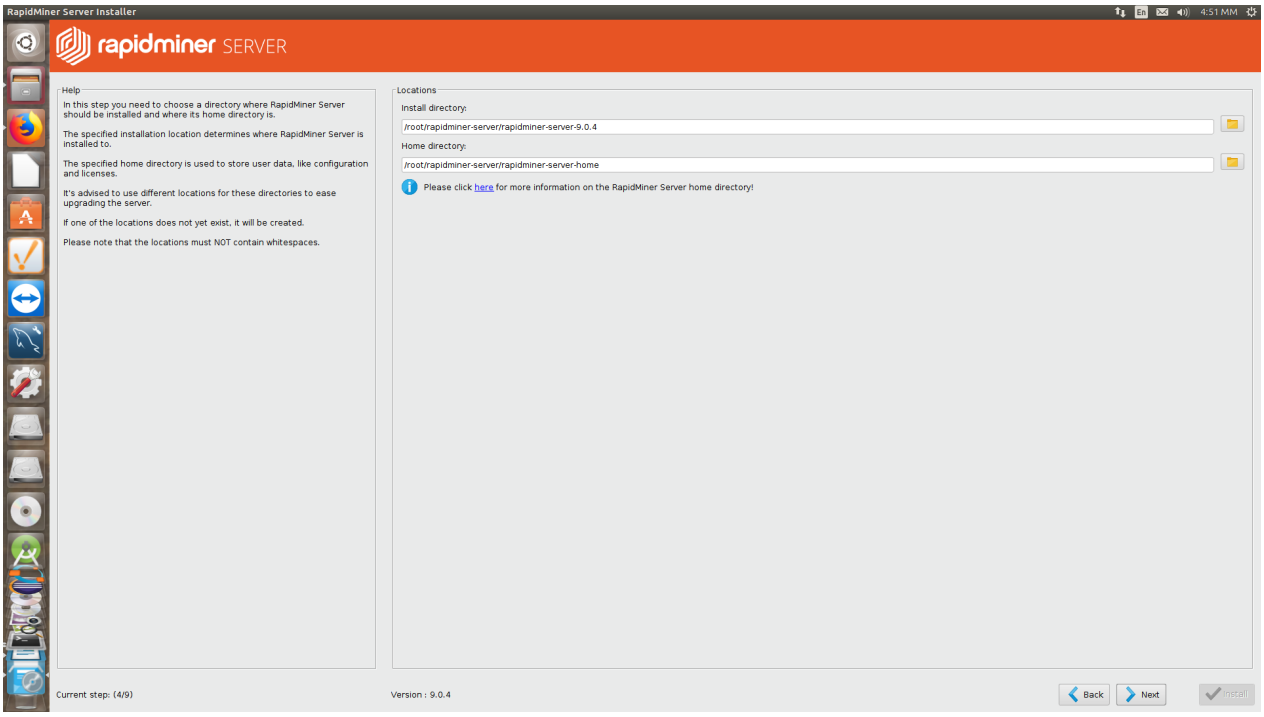
Εικόνα 64: Rapid Miner Server Installer 2

Βήμα 3:



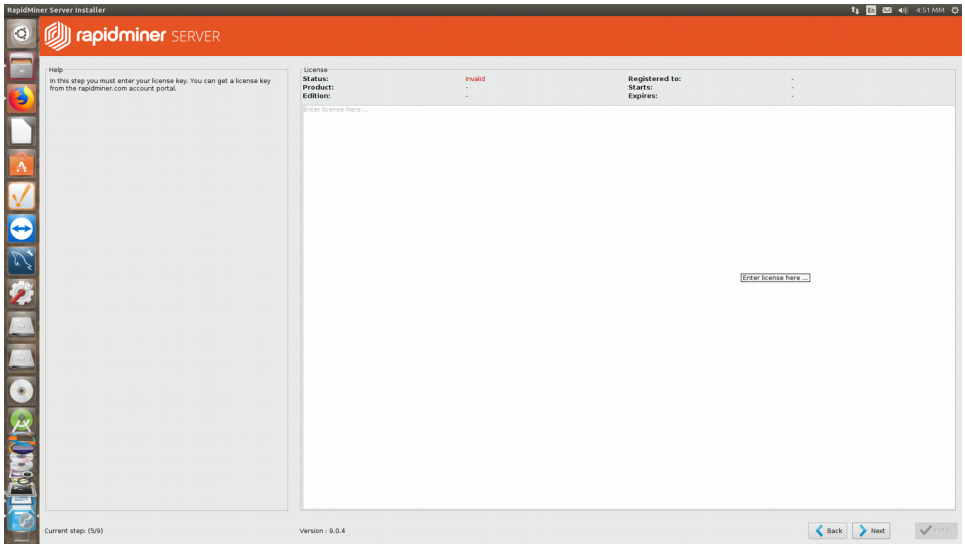
Εικόνα 65: Rapid Miner Server Installer 3

Βήμα 4:



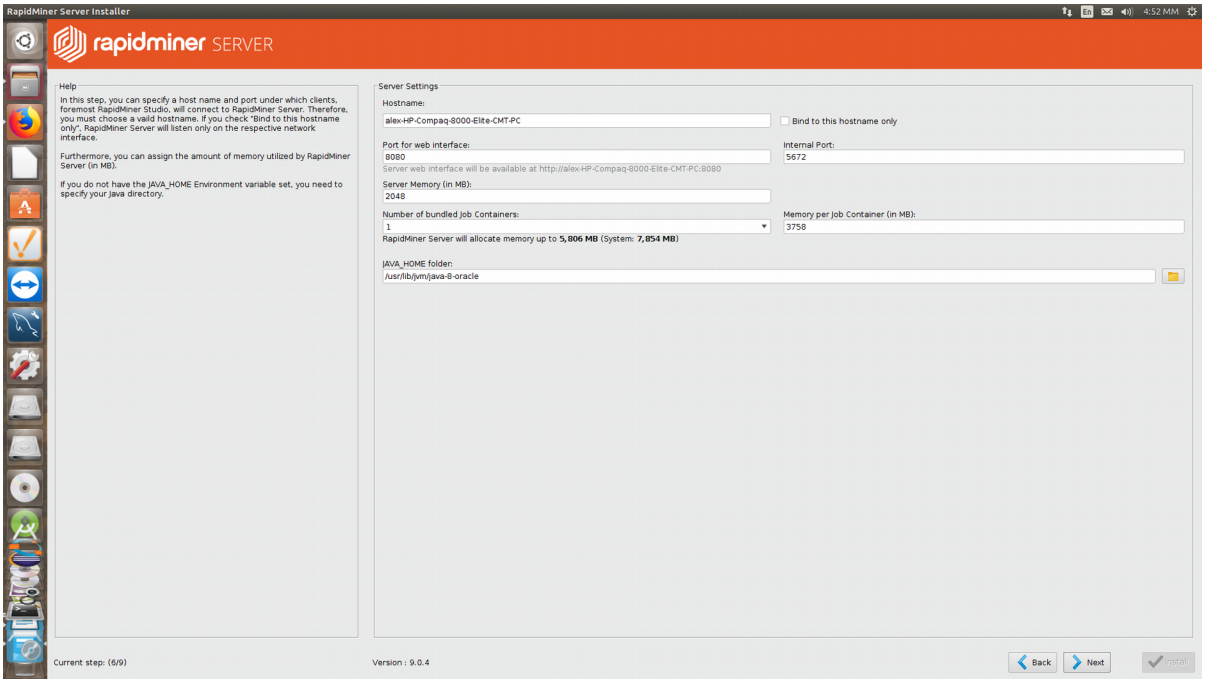
Εικόνα 66: Rapid Miner Server Installer 4

Βήμα 5:



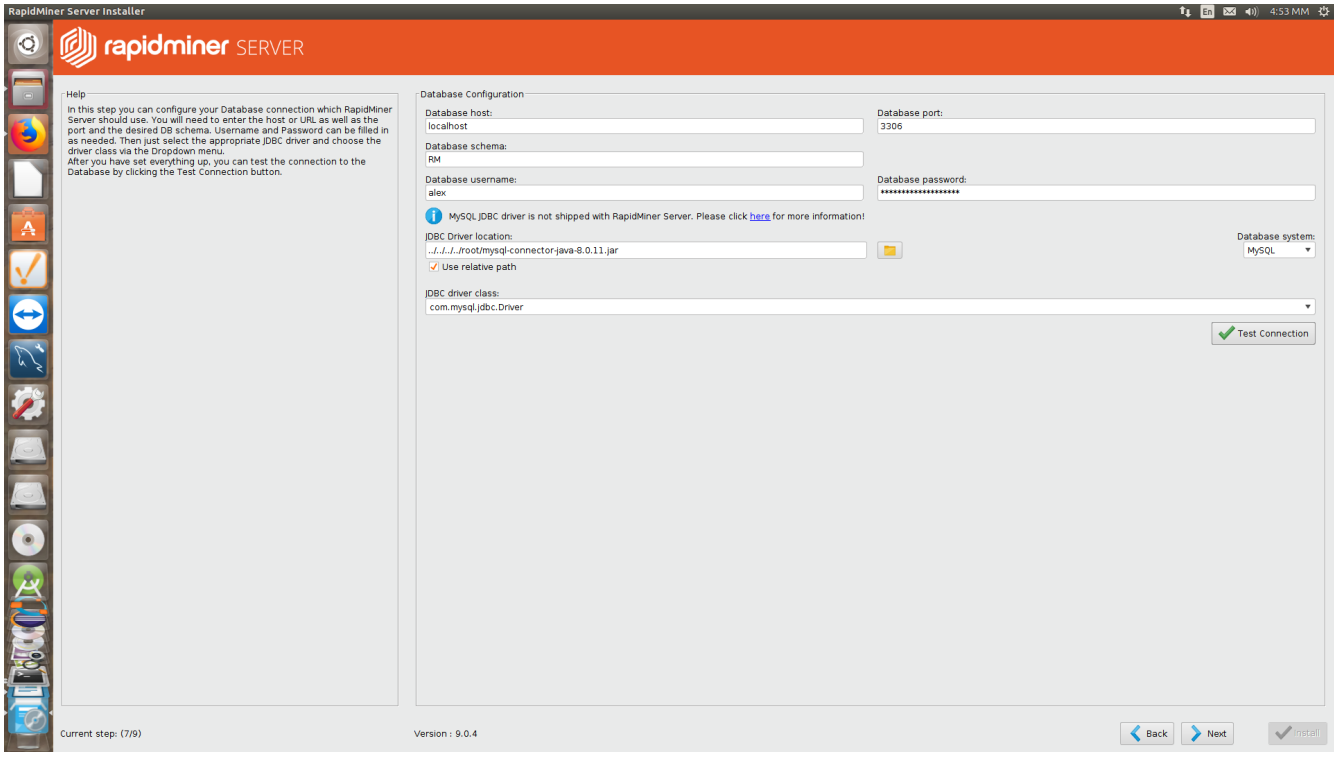
Εικόνα 67: Rapid Miner Server Installer 5

Βήμα 6:



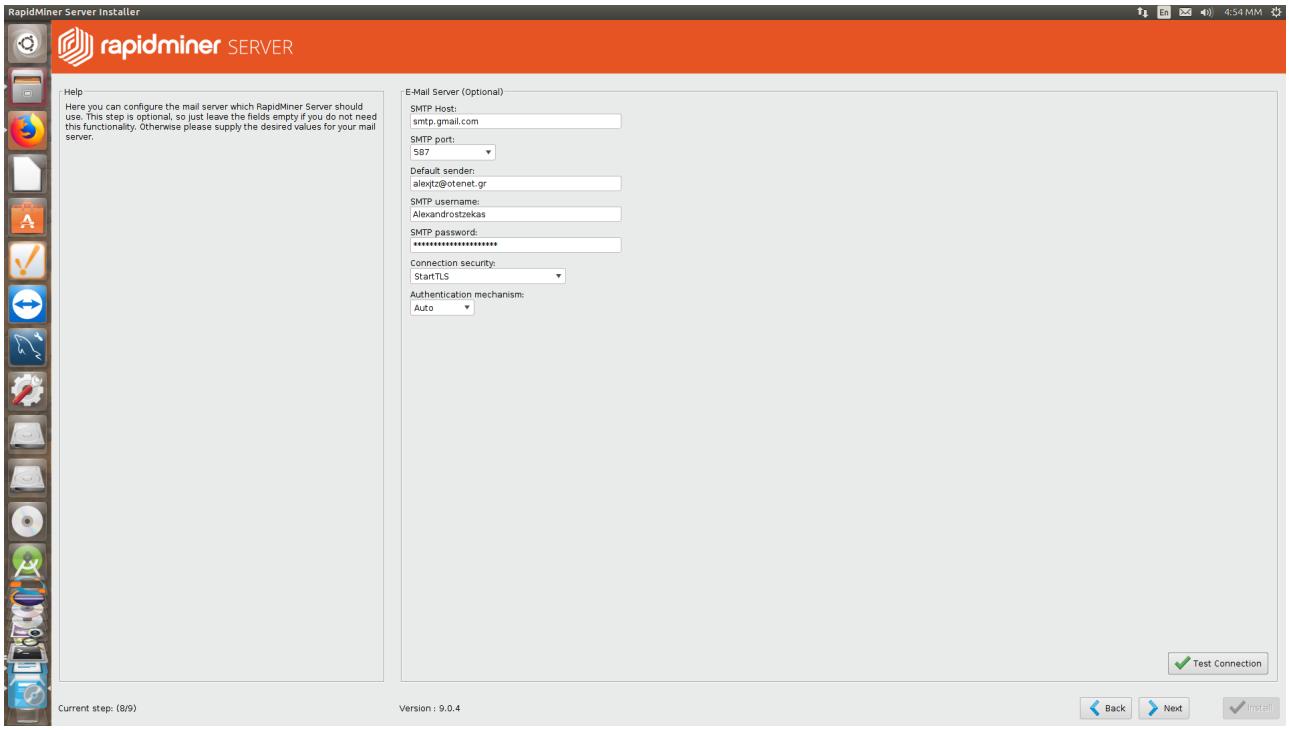
Εικόνα 68: Rapid Miner Server Installer 6

Βήμα 7:



Εικόνα 69: Rapid Miner Server Installer 7

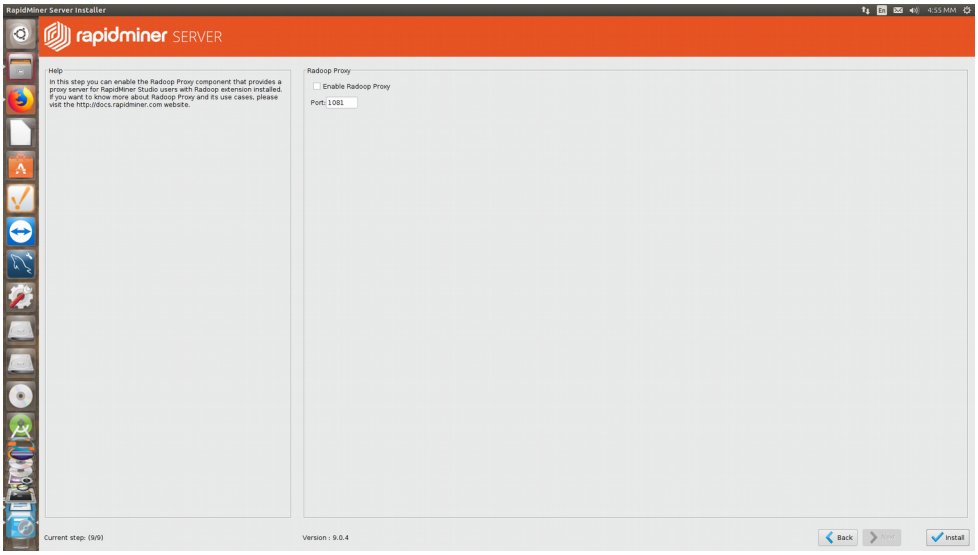
Βήμα 8:



Εικόνα 70: Rapid Miner Server Installer 8

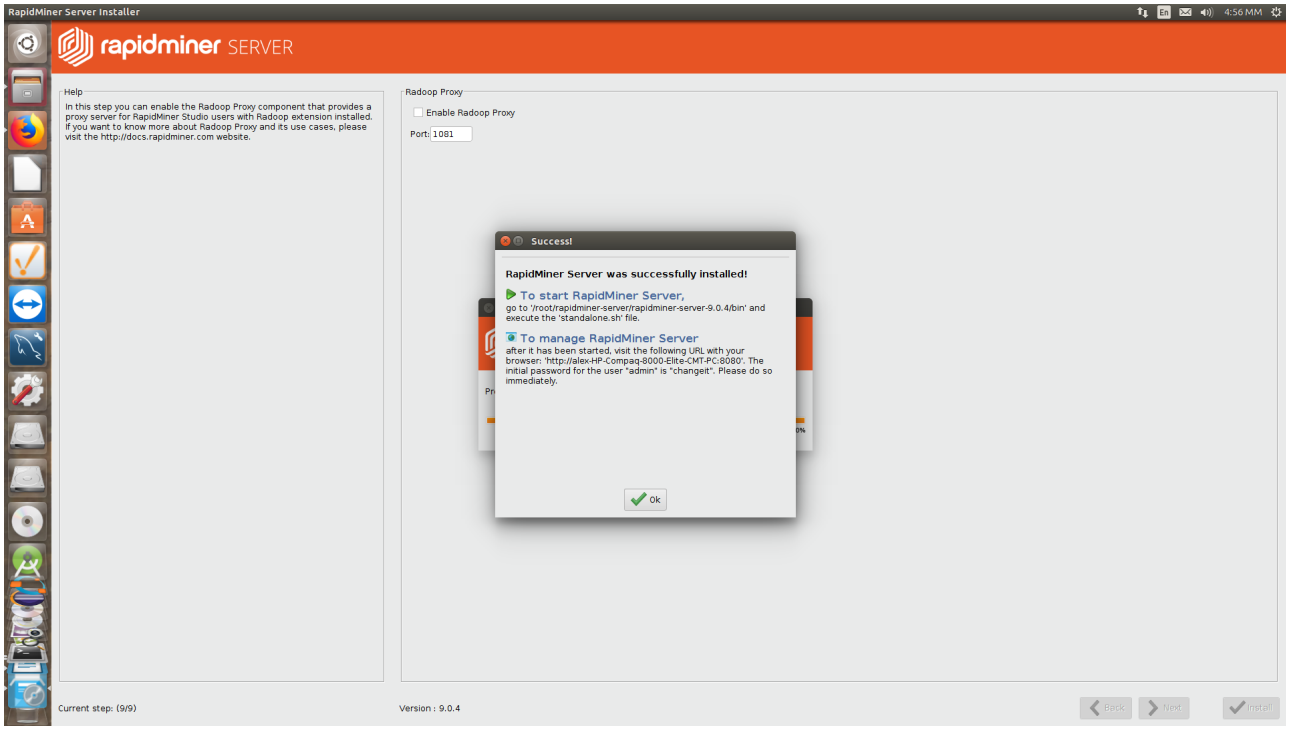
- SMTP Host: Εξαρτάται απ' τον πάροχο των email που χρησιμοποιούμε. Για παράδειγμα, αν έχουμε gmail θα είναι **smtp.gmail.com**.
- SMTP port: Το βρίσκουμε απ' τις ρυθμίσεις λογαριασμού email.
- Default sender: Πληκτρολογούμε ένα έγκυρο email της αρεσκείας μας.
- SMTP username: Χρησιμοποιούμε το username του email μας πλην της καταλήξεως **gmail.com** (για gmail).
- SMTP password: Χρησιμοποιούμε το password του email μας.
- Connection security: Στις ρυθμίσεις λογαριασμού email, το ορίζουμε ως **StartTLS**.

Βήμα 9:



Εικόνα 71: Rapid Miner Server Installer 9

Βήμα 10:



Εικόνα 72: Rapid Miner Server Installer 10

Βήμα 11:

Απ' το Σύστημα μας δίδεται ως username η λέξη **admin** και ως password η λέξη **changeit**, υποδεικνύοντας μας μ' αυτόν τον τρόπο ότι, για την ασφάλεια μας, πρέπει να το αλλάξουμε.

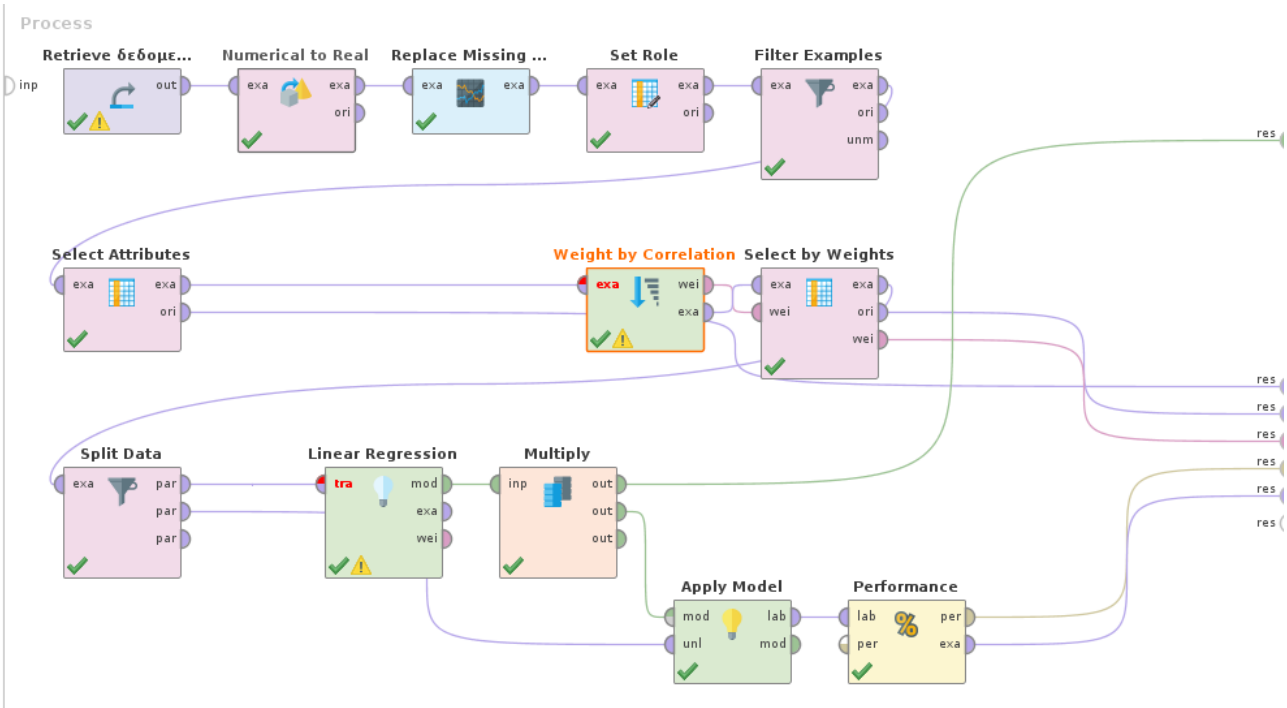
Για να ολοκληρώσουμε την εγκατάσταση του Server, πηγαίνουμε στην τοποθεσία: home/alex/rapidminer-server/rapidminerserver-9.0.4/bin και ανοίγουμε το τερματικό. Εν συνεχεία εκτελούμε τις παρακάτω εντολές:

- **su**
- **./standalone.sh**

Συνδεόμαστε στο web interface (διεπαφή ιστού) και πατάμε **Start installation now**. Όταν ολοκληρωθεί η εγκατάσταση, εμφανίζεται η αρχική σελίδα του διακομιστή RapidMiner.

3.3 Πείραμα Πρόβλεψης

Σε τούτο το εδάφιο θα επιχειρήσουμε να προβλέψουμε την παραγόμενη ισχύ βάσει της έντασης της προσπίπτουσας ηλιακής ακτινοβολίας. Η παρακάτω εικόνα εμπεριέχει όλα τα βήματα της διαδικασίας που ακολουθήθηκε, τα οποία θα περιγραφούν αναλυτικά εν συνεχεία.



Εικόνα 73: Πείραμα Συσχέτισης Παραγόμενης Ισχύος με Ένταση Ακτινοβολίαςhttps://www.lifo.gr/now/culture/227921/hokusai-kai-alloi-megalofyeis-iapones-kallitexnes-se-mia-monadiki-ekthesi-stin-athina?fbclid=IwAR2ZxgNQpG5kGydsV96BSuqu_h9oAK2jopv5oj6BYKj-kFELYr0B3M2bdPY

Βήματα

- Αρχικά μεταφέρουμε στην Οθόνη του Process, τα δεδομένα που μας ενδιαφέρουν με χρήση Drag & Drop απ' τον φάκελο Data όπου τα έχουμε καταχωρήσει.

- Χρησιμοποιούμε τον operator **Numerical to Real** διότι κάποιοι τελεστές δεν μπορούν να διαχειριστούν παρά μόνο πραγματικές τιμές. Στις Παραμέτρους, επιλέγουμε attribute filter type = all, ούτως ώστε να συμπεριληφθούν όλες οι μεταβλητές.

- Προσθέτουμε τον **Replace Missing Values (Series)** με σκοπό την επίλυση του προβλήματος των κενών τιμών. Επιλέγεται το γέμισμα των κενών τιμών, με τη μέθοδο της γραμμικής παρεμβολής με τη λογική ότι οι διαφορές μεταξύ δύο γειτονικών τιμών είναι αρκετά μικρές. Αν και η συγκεκριμένη μέθοδος δεν είναι απόλυτα ακριβής, εν τούτοις είναι η πλέον ασφαλής (για παράδειγμα, αν επιλέγαμε την αντικατάσταση των κενών τιμών με μία καθορισμένη τιμή, τότε θα εμφανιζόταν ένας πολύ μεγάλος παράγοντας σφάλματος, καθώς θα αποκτούσαν την ίδια τιμή δεδομένα, τα οποία χαρακτηρίζονται από εποχιακή συμπεριφορά). Τέλος, στο attribute filter type βάζουμε ένα subset μεταβλητών στο οποίο περιέχονται όλες πλην των MINUTES.

- Επόμενος τελεστής είναι ο **Set Role**, όπου θέτουμε attribute name = Total apparent power, target role = label (υποδηλώνοντας ότι η εν λόγω μεταβλητή έχει ιδιαίτερη σημασία στο πείραμά μας) και set additional roles → edit list:

attribute name	target role
MINUTES	id

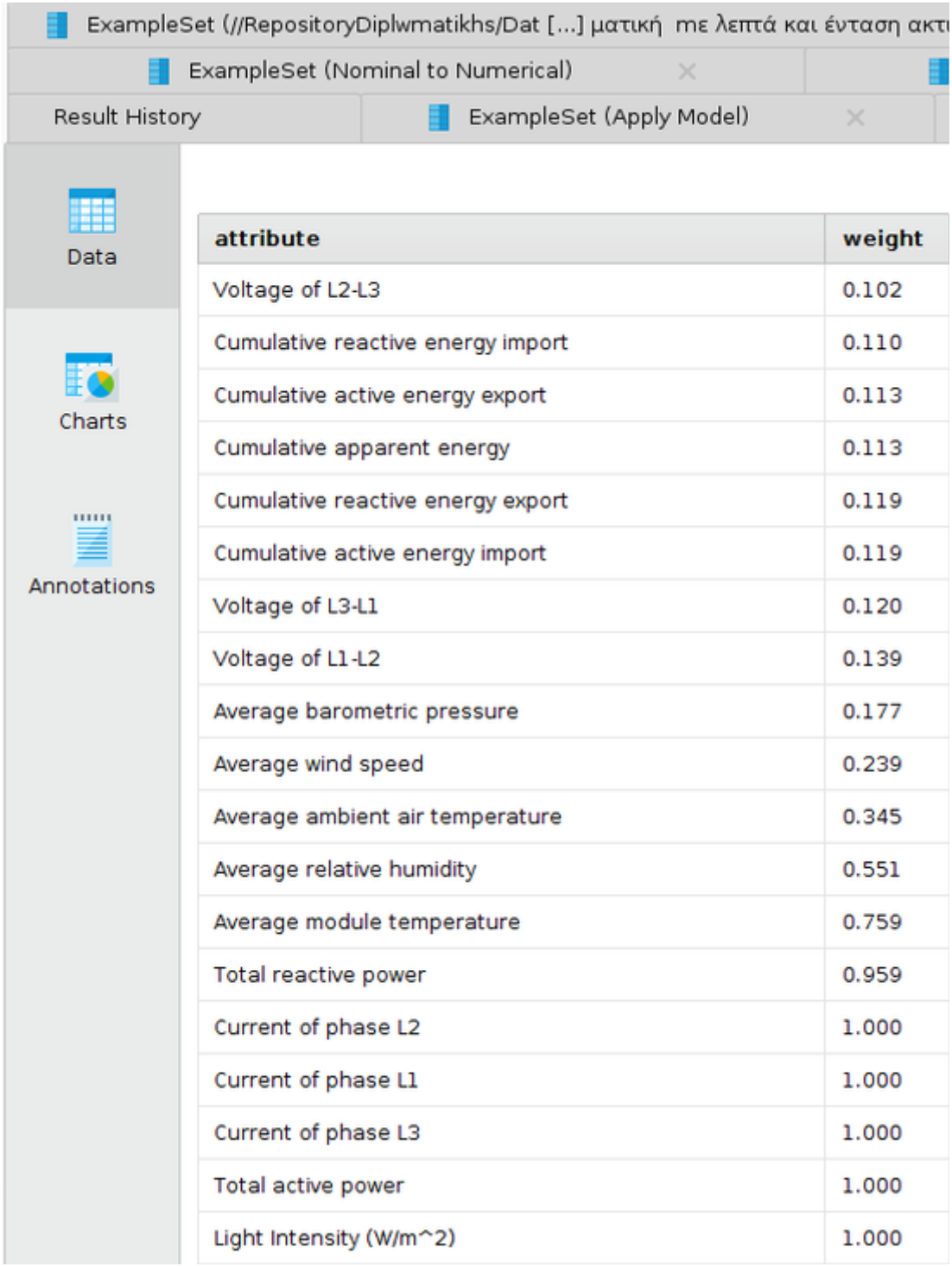
Πίνακας 2: edit list (Set Role)

- Με τον **Filter Examples** απορρίπτουμε τις σειρές για τις οποίες ισχύει Total apparent power < 10 καθώς αυτές αντιστοιχούν σε νυχτερινές ώρες κατά τις οποίες η παραγωγή ισχύος είναι αμελητέα. Αυτό το επιτυγχάνουμε ως εξής:

  - parameter string: Total apparent power >= 10
  - condition class: attribute_value_filter

- Χρησιμοποιούμε τον **Select Attributes** με invert selection (αντιστροφή) και επιλέγουμε attribute filter type = single και attribute = Date/Time έτσι ώστε να μην συμπεριληφθεί η συγκεκριμένη μεταβλητή στη συνέχεια της διαδικασίας.

- Με τον τελεστή **Weight by Correlation** υπολογίζουμε την τιμή της συσχέτισης, για κάθε μεταβλητή του συνόλου δεδομένων, ως προς την μεταβλητή που έχουμε ορίσει ως label (όσο μεγαλύτερο το βάρος, τόσο πιο σημαντική θεωρείται η μεταβλητή). Η συσχέτιση είναι ένας αριθμός ανάμεσα στο -1 και το +1. Μία θετική τιμή για τη συσχέτιση δείχνει ανάλογη σχέση ανάμεσα στις δύο μεταβλητές (υψηλές τιμές της μεταβλητής X συνδέονται με υψηλές τιμές της μεταβλητής Y και αντίστροφα). Αντιθέτως, μία αρνητική συσχέτιση, συνεπάγεται αντιστρόφως ανάλογη σχέση (δηλαδή, υψηλές τιμές για την X συνδέονται με χαμηλές τιμές για την Y και αντίστροφα).
  Στις Παραμέτρους επιλέγουμε sort weights και βάζουμε sort direction = ascending. Έτσι, ο πίνακας που θα προκύψει θα έχει αύξουσα ταξινόμηση.



attribute	weight
Voltage of L2-L3	0.102
Cumulative reactive energy import	0.110
Cumulative active energy export	0.113
Cumulative apparent energy	0.113
Cumulative reactive energy export	0.119
Cumulative active energy import	0.119
Voltage of L3-L1	0.120
Voltage of L1-L2	0.139
Average barometric pressure	0.177
Average wind speed	0.239
Average ambient air temperature	0.345
Average relative humidity	0.551
Average module temperature	0.759
Total reactive power	0.959
Current of phase L2	1.000
Current of phase L1	1.000
Current of phase L3	1.000
Total active power	1.000
Light Intensity (W/m^2)	1.000

Εικόνα 74: Weight by Correlation

Απ' τον παραπάνω πίνακα παρατηρούμε ότι τα μεγαλύτερα βάρη αντιστοιχούν στην Ένταση της Ακτινοβολίας, στα Ρεύματα και στην Ολική Ενεργό Ισχύ. Προφανώς, ενδιαφερόμαστε μόνο για την Ένταση της Ακτινοβολίας καθώς τα υπόλοιπα μεγέθη αφενός ανήκουν στο ίδιο το σύστημα και είναι εξ ορισμού αλληλοεξαρτώμενα και αφετέρου δεν είναι παρά αποτελέσματα αυτής (όπως και η Παραγόμενη Ισχύς την οποία κι εξετάζουμε).

Αξίζει να σημειωθεί πως το ιδιαίτερο χαρακτηριστικό του συγκεκριμένου τελεστή είναι οι πύλες εξόδου, που προσφέρει. Έτσι, η πύλη “**exa**” δίνει το σύνολο δεδομένων, όπως ακριβώς ήρθε στην είσοδο του τελεστή ενώ η πύλη “**wei**” δίνει τα βάρη των μεταβλητών, ως προς την μεταβλητή label.

- Ακριβώς δίπλα απ’ τον τελεστή **Weight by Correlation** θα προσθέσουμε τον **Select by Weights**. 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.
  Επιλέγουμε:

  - weight relation = top k
  - **k = 1** (ώστε να επιλεχθεί μόνο η Ένταση Ακτινοβολίας (Light Intensity) όπως αναφέραμε και προηγουμένως)
  - deselect unknown
  - use absolute weights

  Η πύλη “**wei**” παραδίδει τα βάρη των μεταβλητών, όπως ακριβώς εισήχθησαν στην είσοδο, δηλαδή από τον τελεστή “Weight by Correlation” ενώ η πύλη “**exa**” δίνει το σύνολο δεδομένων, διατηρώντας μόνο τις μεταβλητές, οι οποίες ικανοποίησαν τη συνθήκη.

- Σ’ αυτό το στάδιο θα πρέπει να χωρίσουμε τα δεδομένα μας σε δύο υποσύνολα:
  (α) στα Training Data (για τη δημιουργία και την εκπαίδευση του μοντέλου) και (β) στα Testing Data (για την αξιολόγηση του μοντέλου). Προς τούτο θα χρησιμοποιήσουμε τον **Split Data** operator. Η παράμετρος sampling type του εν λόγω τελεστή, υποδεικνύει τον τρόπο, με τον οποίο τα παραδείγματα κατανέμονται στα τελικά υποσύνολα. Επιλέγουμε linear sampling που διαχωρίζει το σύνολο δεδομένων σε υποσύνολα, χωρίς να αλλάζει τη σειρά των παραδειγμάτων και στο partitions → Edit Enumeration:

ratio
0.8 (για εκπαίδευση)
0.2 (για αξιολόγηση)

Πίνακας 3: ratio (Split Data)

- Ως μοντέλο επιλέγουμε αυτό της Γραμμικής Παλινδρόμησης (**Linear Regression**) καθώς επιλέχθηκε μόνο μία παράμετρος ως βάση για την πρόβλεψη. Ισχύουν τα παρακάτω:

  - Η είσοδος “**tra**” συνδέεται με την πρώτη έξοδο του **Split Data** operator και δέχεται τα Training Data.

- Η έξοδος “**mod**” εμφανίζει τους συντελεστές του μοντέλου γραμμικής παλινδρόμησης. Συνδέεται δε με την είσοδο “**inp**” του επόμενου τελεστή που είναι ο **Multiply**.
- Η έξοδος “**exa**” εμφανίζει το υποσύνολο Training Data, όπως ακριβώς εισήχθη στην πύλη εισόδου και συνδέεται με την είσοδο “**unl**” του **Apply Model** operator.
- Τις Παραμέτρους τις αφήνουμε ως έχουν.

- Προσθέτουμε έναν **Multiply** operator. Κάνουμε τις εξής συνδέσεις:

  - Είσοδος “**inp**” με την έξοδο “**mod**” του **Linear Regression**.
  - Πρώτη έξοδος με “**res**”.
  - Δεύτερη έξοδος με την είσοδο “**mod**” του **Apply Model**.

- Για την Αξιολόγηση του μοντέλου αλλά και για την εξαγωγή Προβλέψεων, εφαρμόζουμε τον τελεστή **Apply Model**. Ο συγκεκριμένος τελεστής, δέχεται το μοντέλο απ’ την είσοδο “**mod**” και το υποσύνολο Testing Data για την αξιολόγησή του απ’ την είσοδο “**unl**”.

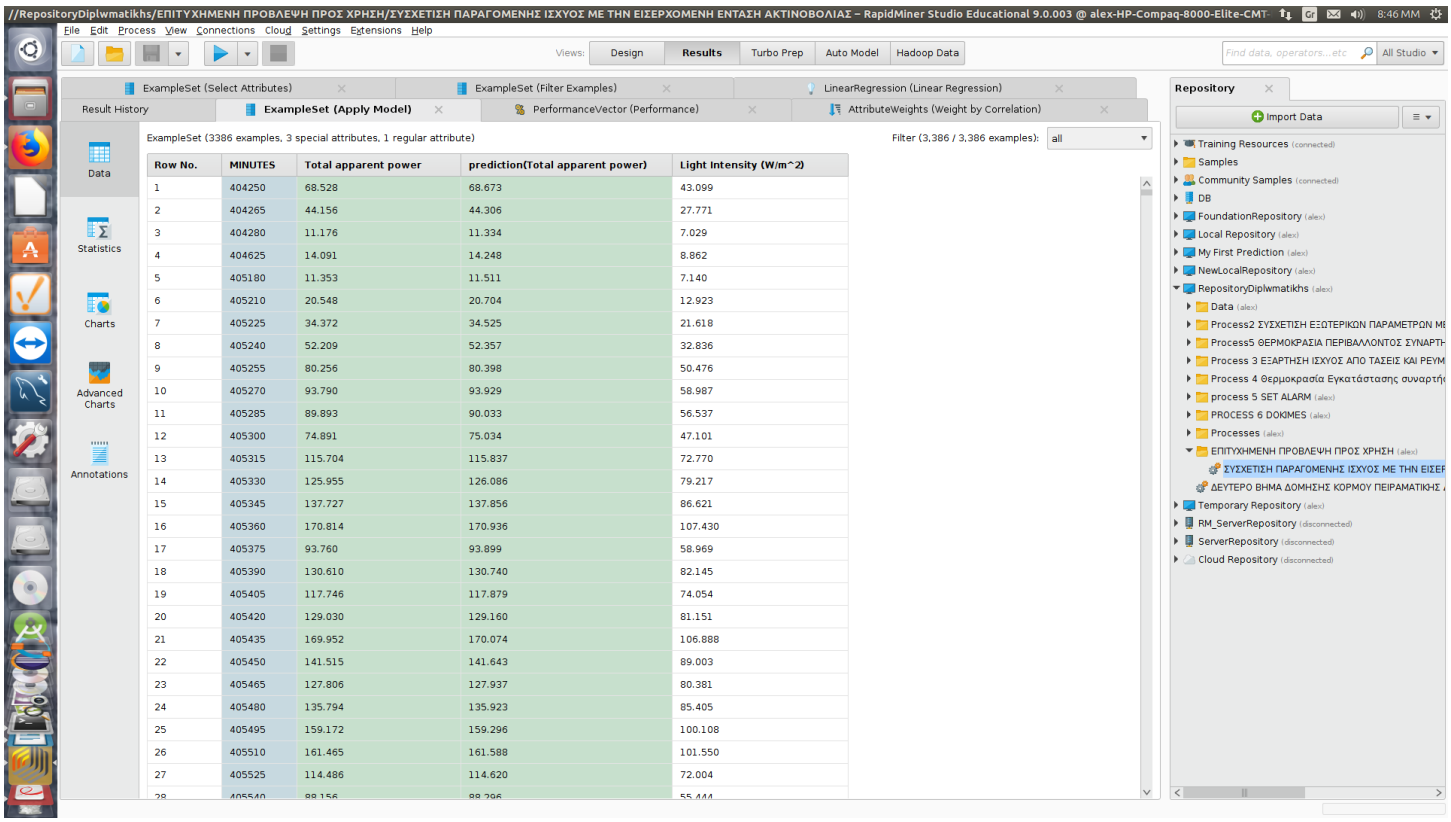
- Έπειτα, μέσω της εξόδου “**lab**”, ο **Apply Model** συνδέεται με τον **Performance** operator, ο οποίος δέχεται τα δεδομένα απ’ την πρώτη είσοδό του που είναι η “**lab**”. Με τη χρήση του συγκεκριμένου τελεστή γίνεται η ποσοτικοποίηση της αξιολόγησης του μοντέλου μας βάσει των σφαλμάτων που θα επιλέξουμε στις παραμέτρους. Τα σφάλματα που επιλέξαμε καθώς και οι τιμές αυτών παρατίθενται στον επόμενο πίνακα:

ΤΙΤΛΟΣ ΣΦΑΛΜΑΤΟΣ	**ΤΙΜΗ**	**ΣΧΟΛΙΑ**
root_mean_squared_error	0.109	Πολύ μικρό.
absolute_error	0.096 ± 0.050	Πολύ μικρό.
relative_error	0.21% ± 0.37 %	Πολύ μικρό.
Squared_Correlation	1	Τέλεια εφαρμογή του μοντέλου πρόβλεψης στα πειραματικά δεδομένα.

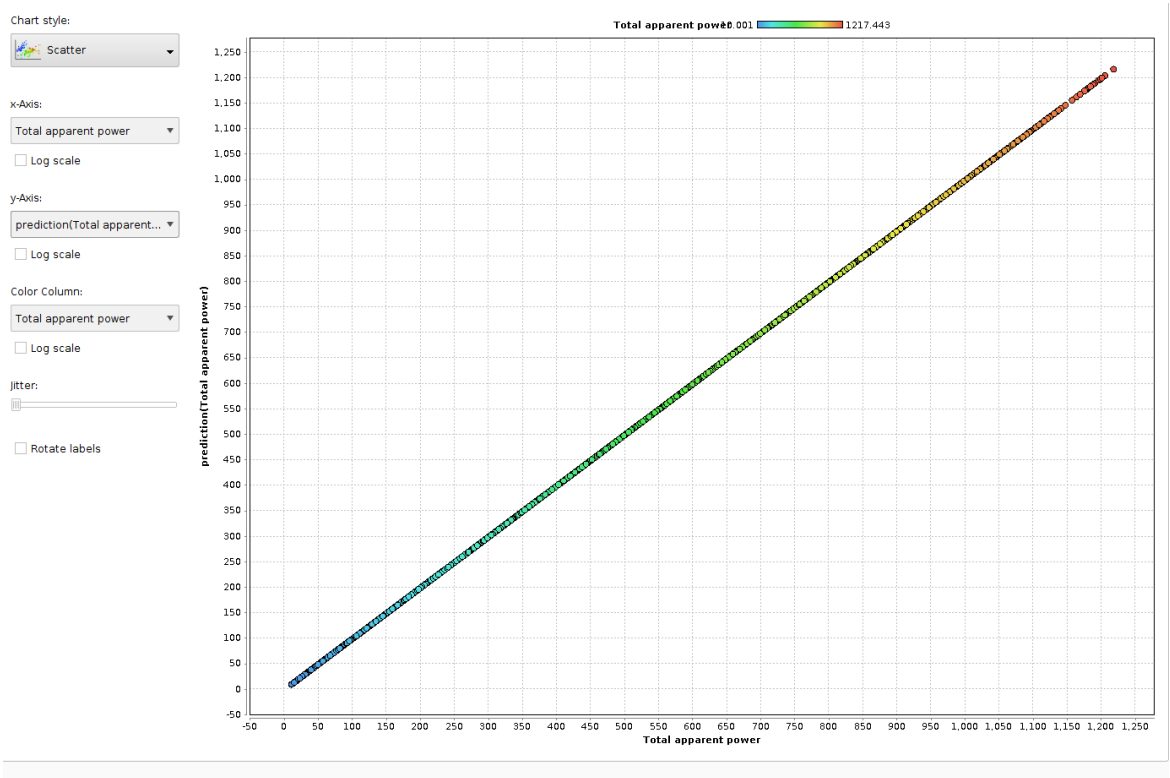
Πίνακας 4: Τιμές Σφαλμάτων Πειραματικής Διαδικασίας

Απ’ τα παραπάνω έπεται ότι η επιλογή του μοντέλου ήταν επιτυχημένη.

- Τέλος, πατάμε το “run” για να λάβουμε τα αποτελέσματά μας:



Εικόνα 75: ExampleSet (Apply Model)



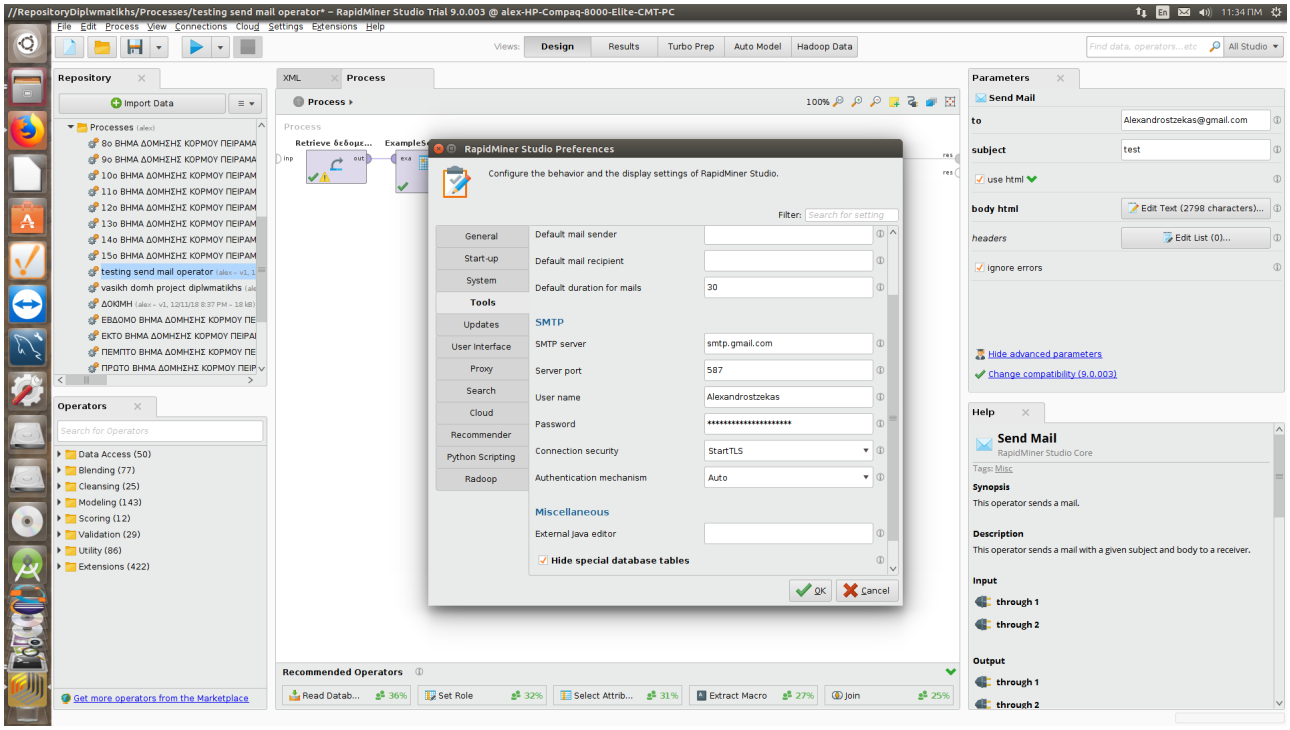
Εικόνα 76: ExampleSet (Apply Model) Διάγραμμα

Απ' το παραπάνω διάγραμμα παρατηρούμε ότι οι προβλέψεις μας είναι επιτυχείς καθώς η ευθεία που σχηματίζεται απ' τα σημεία **Σ(Total apparent power, prediction(Total**

apparent power)) έχει κλίση 45 μοιρών. Αυτό σημαίνει ότι γνωρίζοντας την προβλεπόμενη τιμή για κάποια μελλοντική μέτρηση ουσιαστικά ξέρουμε και την πραγματική τιμή που θα προκύψει κατά αυτήν τη μέτρηση. Φυσικά, στην προκειμένη περίπτωση, η διεξαγωγή των πειραματικών μετρήσεων έγινε υπό “φυσιολογικές” καιρικές συνθήκες. Εάν είχαν παρουσιαστεί ακραία καιρικά φαινόμενα, όπως παραδείγματος χάριν πολύ ισχυροί άνεμοι, ενδέχεται να παρουσιάζονταν σφάλματα που θα επηρέαζαν σημαντικά τις παρατηρηθείσες τιμές με αποτέλεσμα το μοντέλο μας να χρειαζόταν περισσότερες παραμέτρους για να προβεί σε ακριβείς προβλέψεις. Ωστόσο, υπό τις συγκεκριμένες συνθήκες, η επιφόρτιση του μοντέλου με επιπλέον στοιχεία θα είχα ως αποτέλεσμα την, άνευ λόγου, επιβάρυνσή του και την ανάγκη χρήσης μεγαλύτερης υπολογιστικής ισχύος (σε άλλα πειράματα που δεν παρουσιάζονται στην Διπλωματική παρουσιάστηκαν προβλήματα στο σύστημα λόγω πολλών παραμέτρων).

3.4 Alarms

Για τη διεξαγωγή της συγκεκριμένης εφαρμογής, πρέπει πρώτα να καθορίσουμε το SMTP εντός του Rapid Miner Studio. Προς τούτο, πηγαίνουμε στα Settings → Preferences → Tools → SMTP και ακολουθούμε την ίδια διαδικασία που είχαμε ακολουθήσει και στο υποκεφάλαιο 3.2 ως προς τα email.



Εικόνα 77: Configuring email on Rapid Miner Studio

Πηγαίνουμε στο φάκελο Data στο Repository της Διπλωματικής, διαλέγουμε τα δεδομένα του πειράματος και τα μεταφέρουμε στο process με drag & drop. Εν συνεχεία τα συνδέουμε με τον **Replace Missing Values (Series)** operator ώστε να απαλείψουμε τυχόν απροσδιόριστες τιμές (προαιρετικά). Ακολουθεί ο **Branch** operator που διαθέτει δύο εισόδους˙ την cod (condition) και την inp. Επιλέγουμε την πρώτη θύρα ώστε να ενεργοποιηθεί η Δράση του Τελεστή: if_then_else. Στις παραμέτρους condition type επιλέγουμε atribute_value_filter και από κάτω, στο condition value, γράφουμε Average

module temperature <= 60. Ο **Branch** θα εξετάσει αν η μέγιστη τιμή της συγκεκριμένης μεταβλητής πληρεί την παραπάνω συνθήκη και ανάλογα με το αποτέλεσμα θα μας παραπέμψει σε μία απ' τις δύο περιοχές που υπάρχουν στο εσωτερικό του˙ την περιοχή Then και την περιοχή Else.

- Περιοχή Then: Την αφήνουμε κενή.

- Περιοχή Else: Τοποθετούμε τον **Send Mail** operator και στις παραμέτρους του βάζουμε:
  α) to: "το email μας"
  β) subject: temperature ALARM
  γ) body plain → Edit Text:

"ΠΡΟΣΟΧΗ!!! Το Μέγιστο της Θερμοκρασίας της Εγκατάστασης ξεπέρασε το επιτρεπτό όριο των 60 Βαθμών Κελσίου."

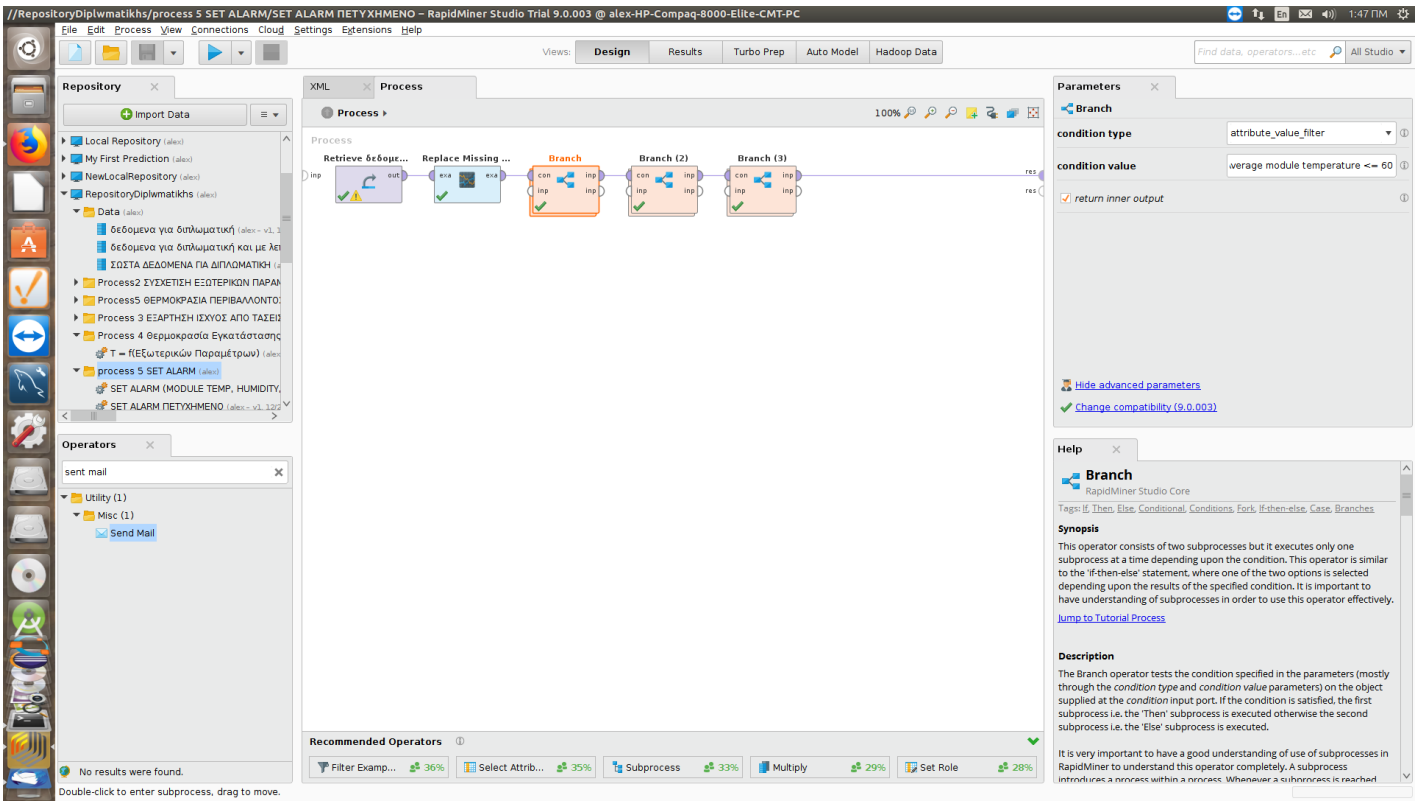
Επαναλαμβάνουμε την ίδια διαδικασία για την Υγρασία θέτοντας ως συνθήκη:

Average relative humidity <= 98

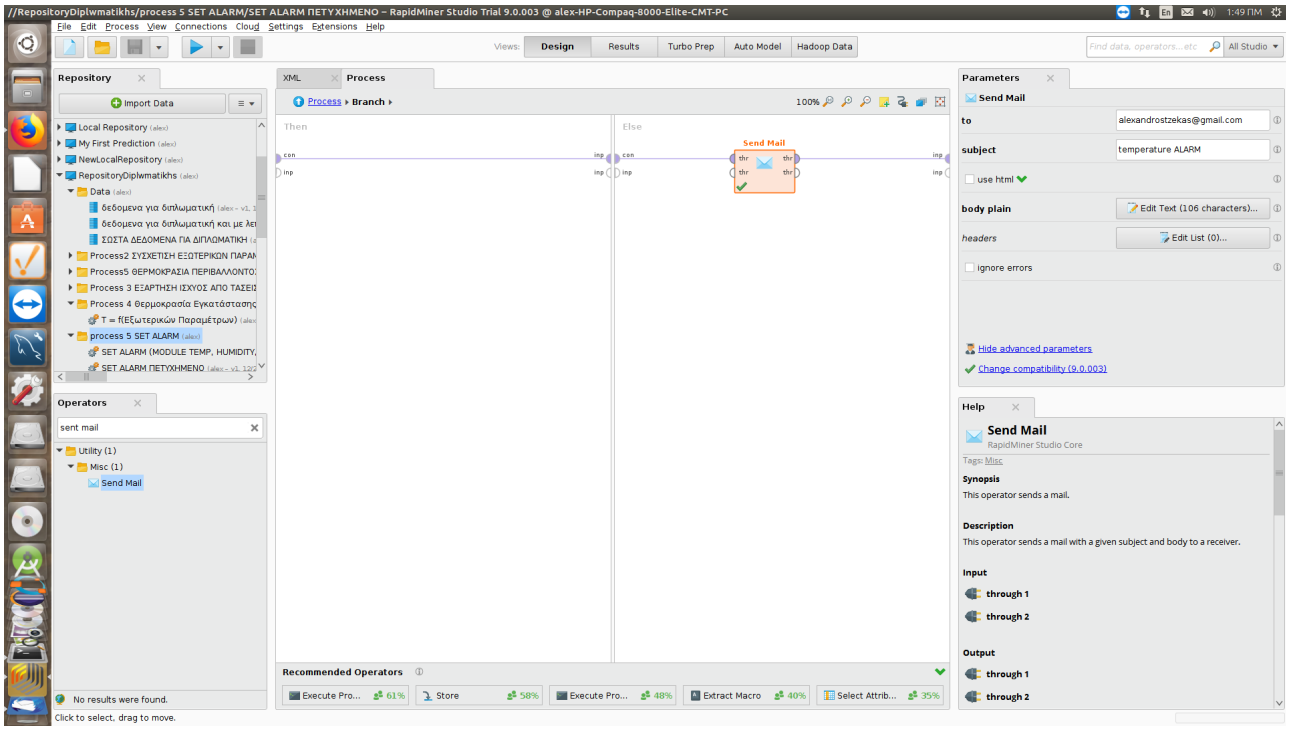
και για την Ταχύτητα του Ανέμου με συνθήκη:

Average wind speed <= 11

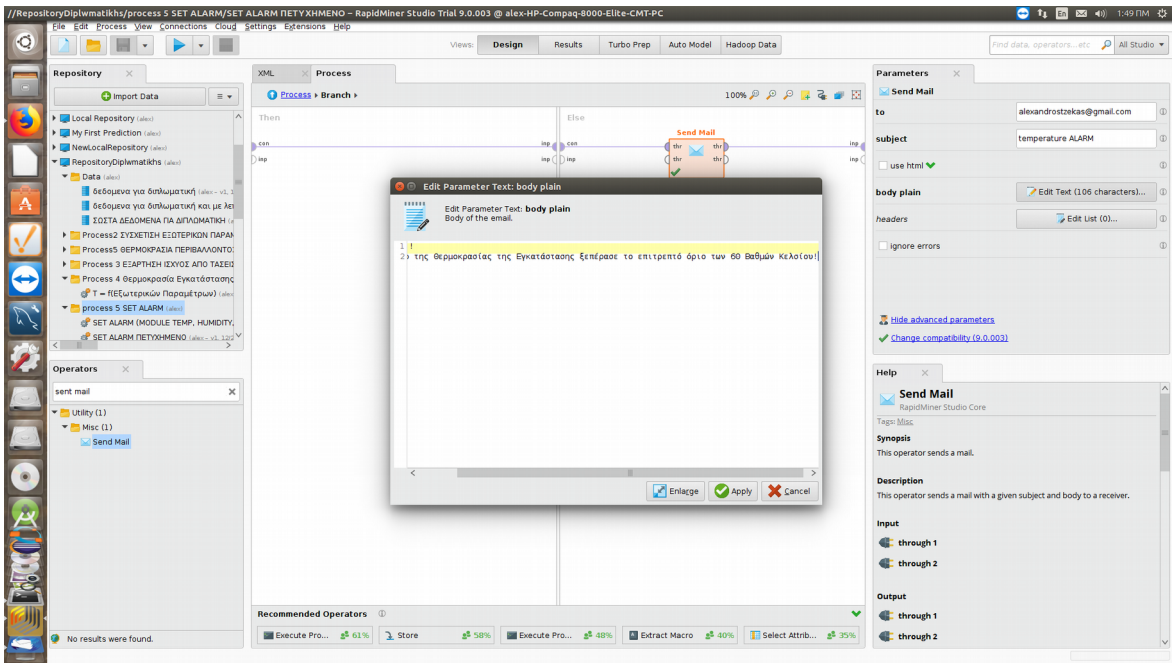
(Οι οριακές τιμές σχετίζονται με τις παρατηρηθείσες μέγιστες τιμές που λάβαμε απ' τα statistics)



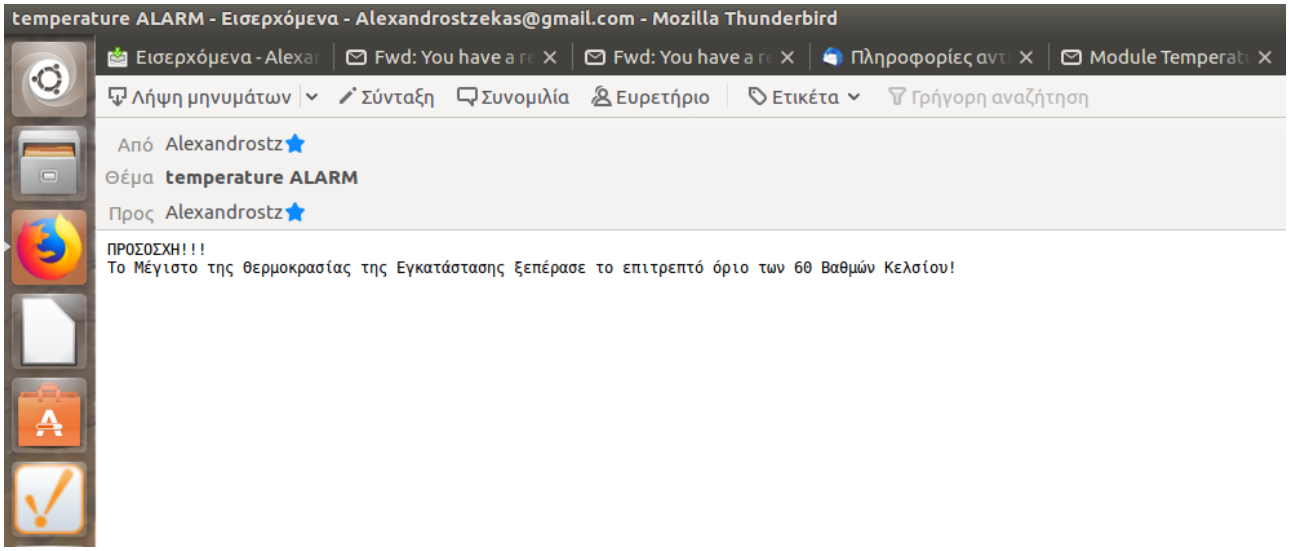
Εικόνα 78: Set Alarm Process/ Temperature Alarm



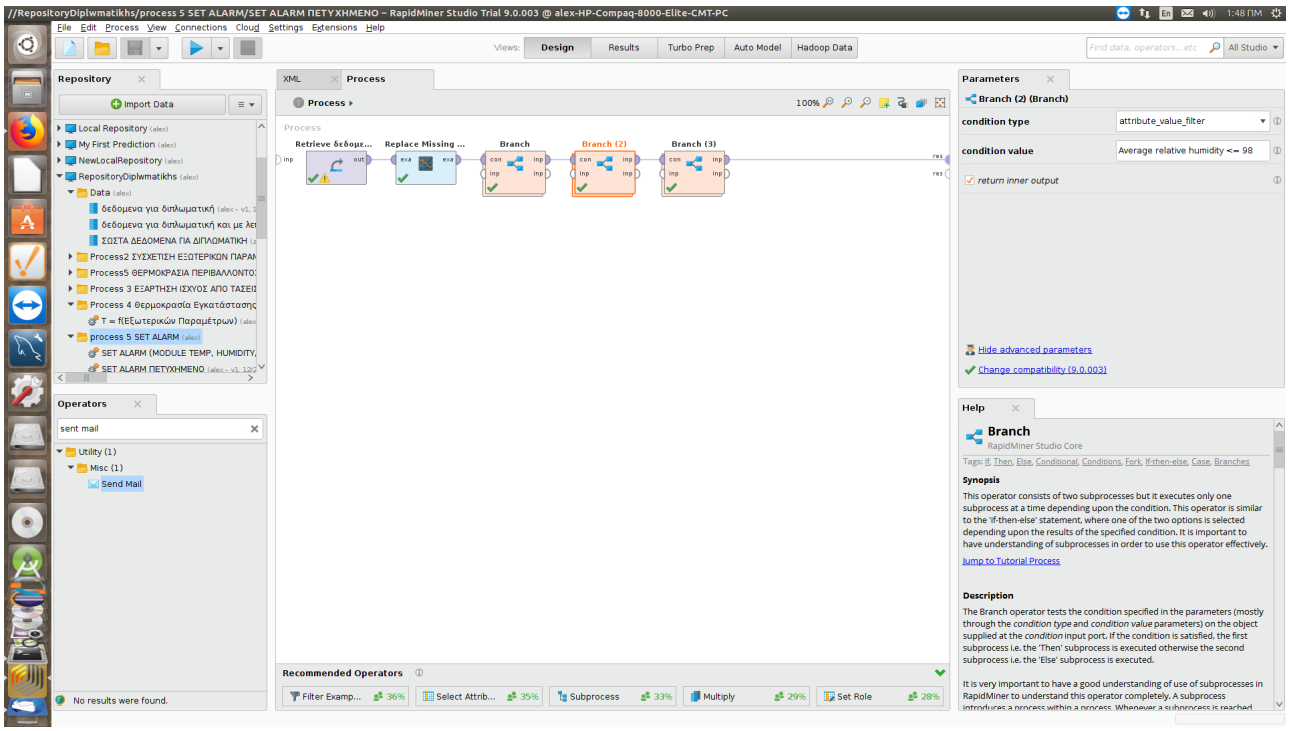
Εικόνα 79: Set Alarm Subprocess1/ Temperature Alarm



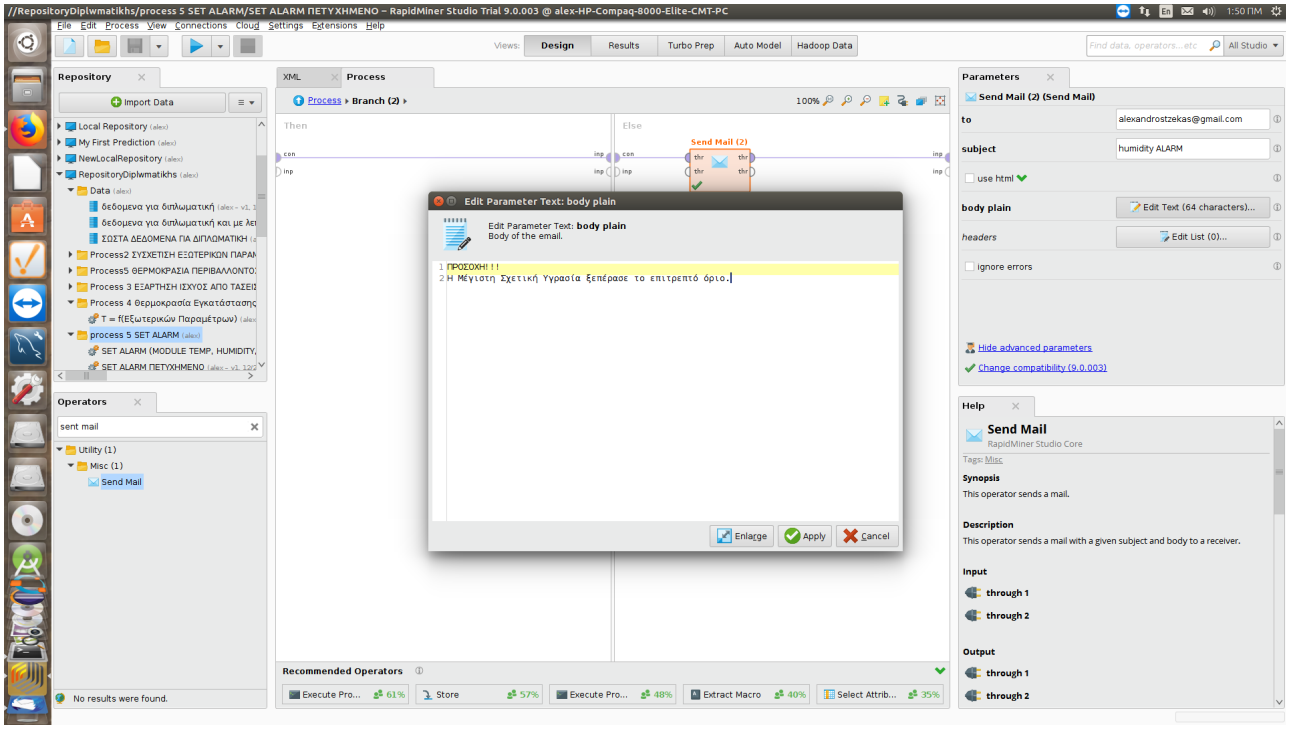
Εικόνα 80: Set Alarm Subprocess1/ Edit Parameter Text (Temperature)



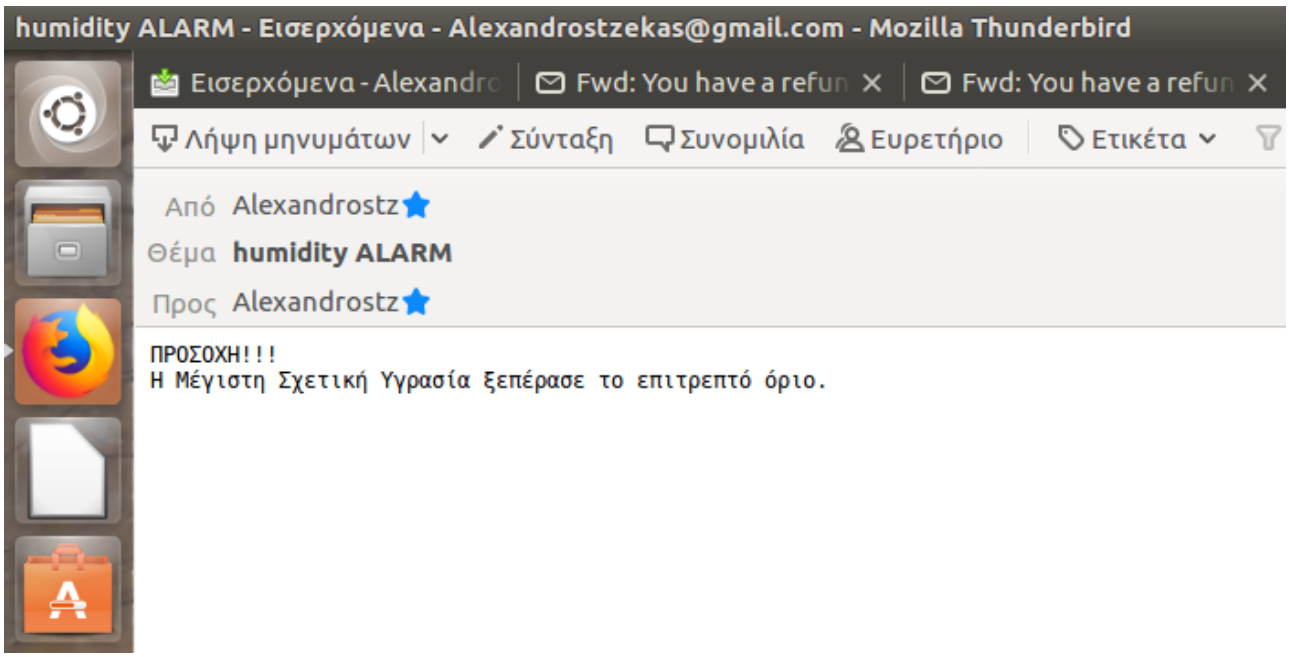
Εικόνα 81: Alarm Email (Temperature)



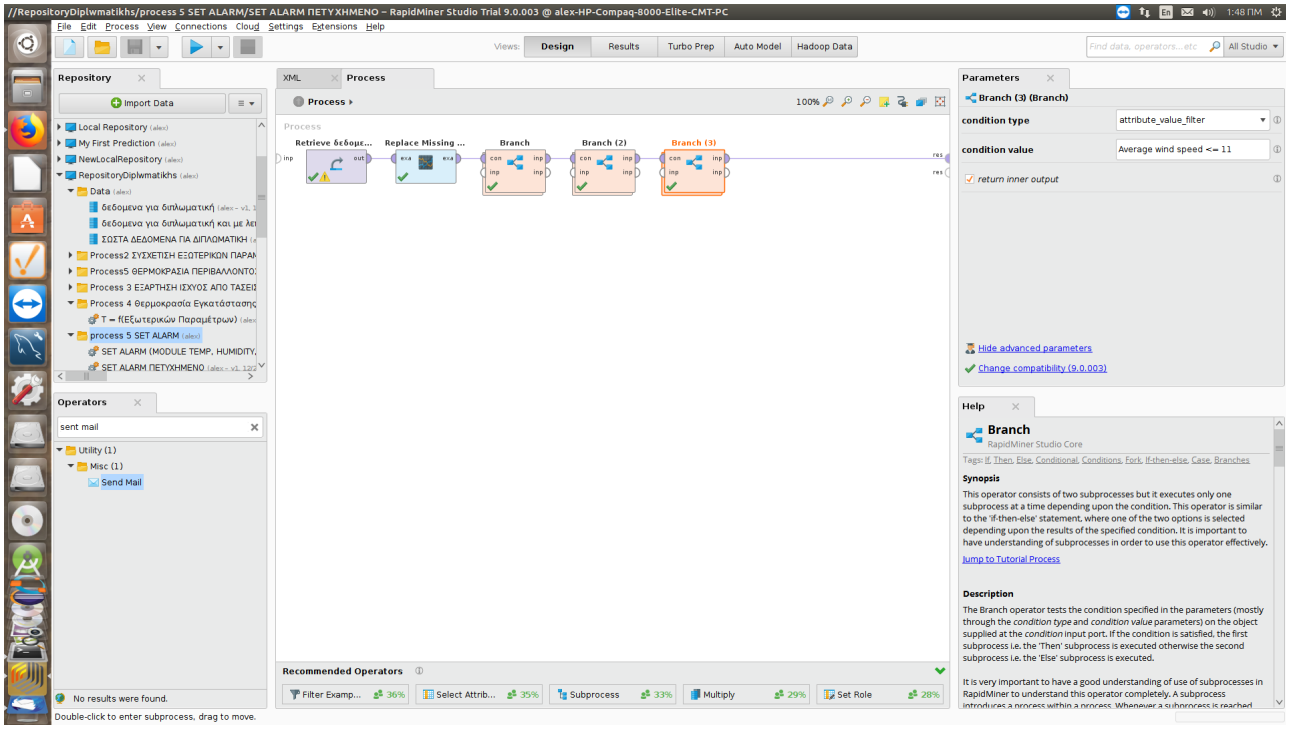
Εικόνα 82: Set Alarm Process/ Humidity Alarm



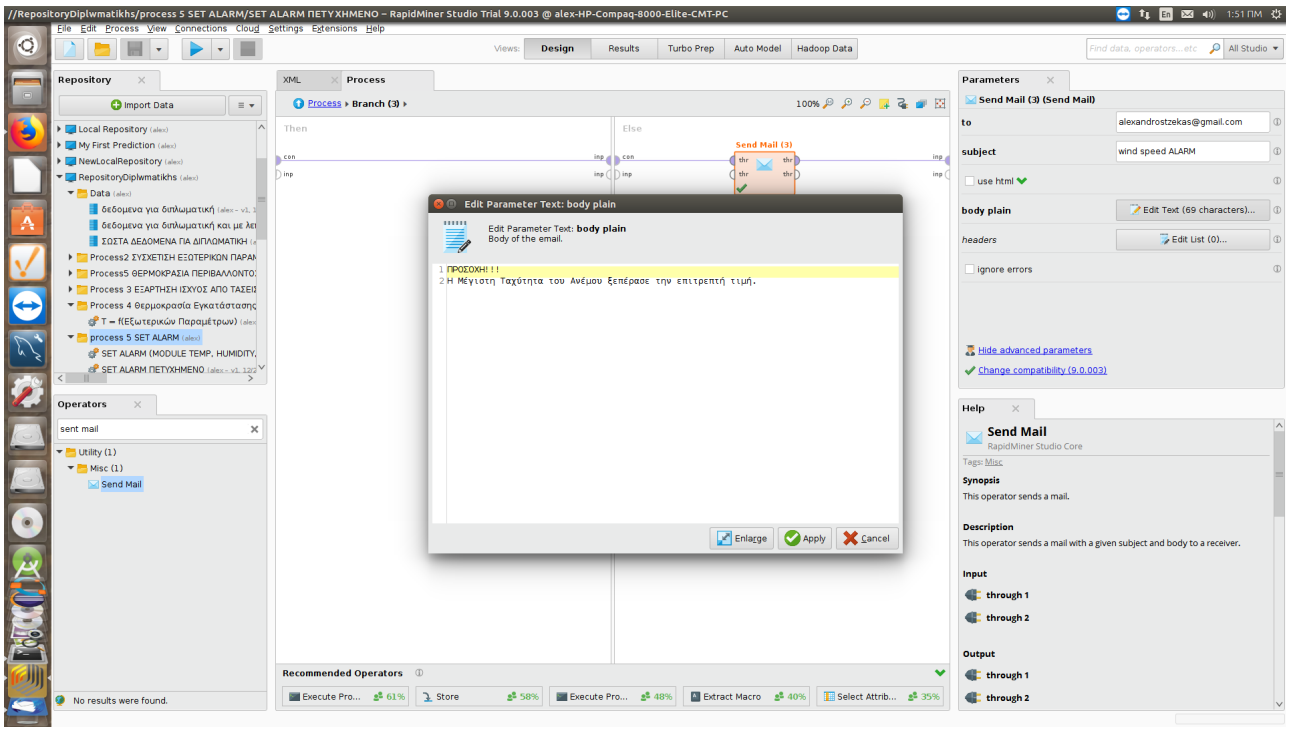
Εικόνα 83: Set Alarm Subprocess2/ Edit Parameter Text (Humidity)



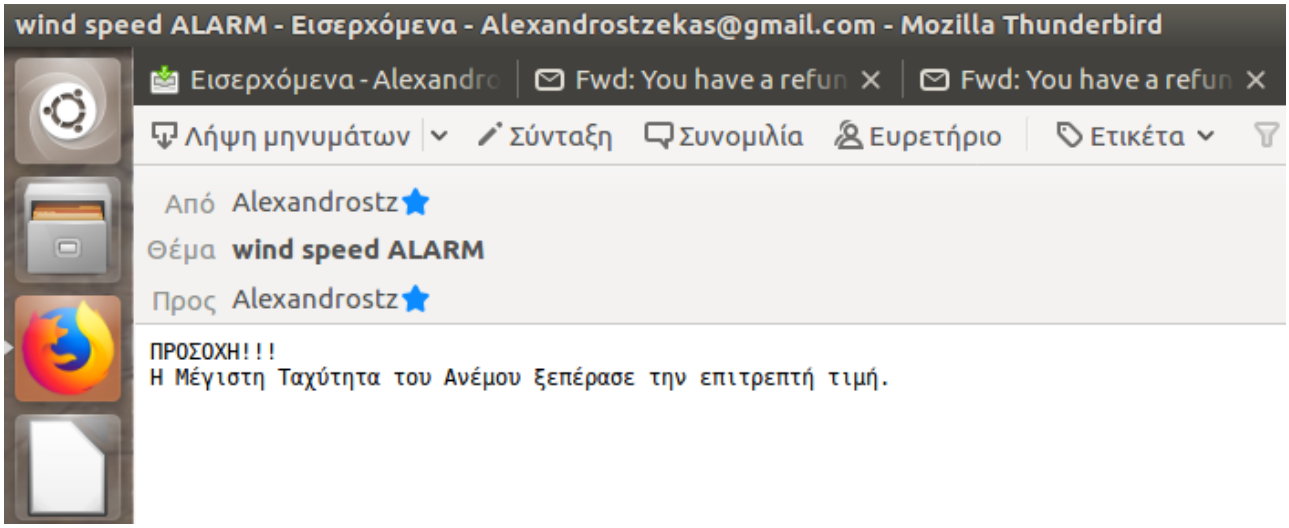
Εικόνα 84: Alarm Email (Humidity)



Εικόνα 85: Set Alarm Process/ Wind Speed Alarm



Εικόνα 86: Set Alarm Subprocess3/ Edit Parameter Text (Wind Speed)



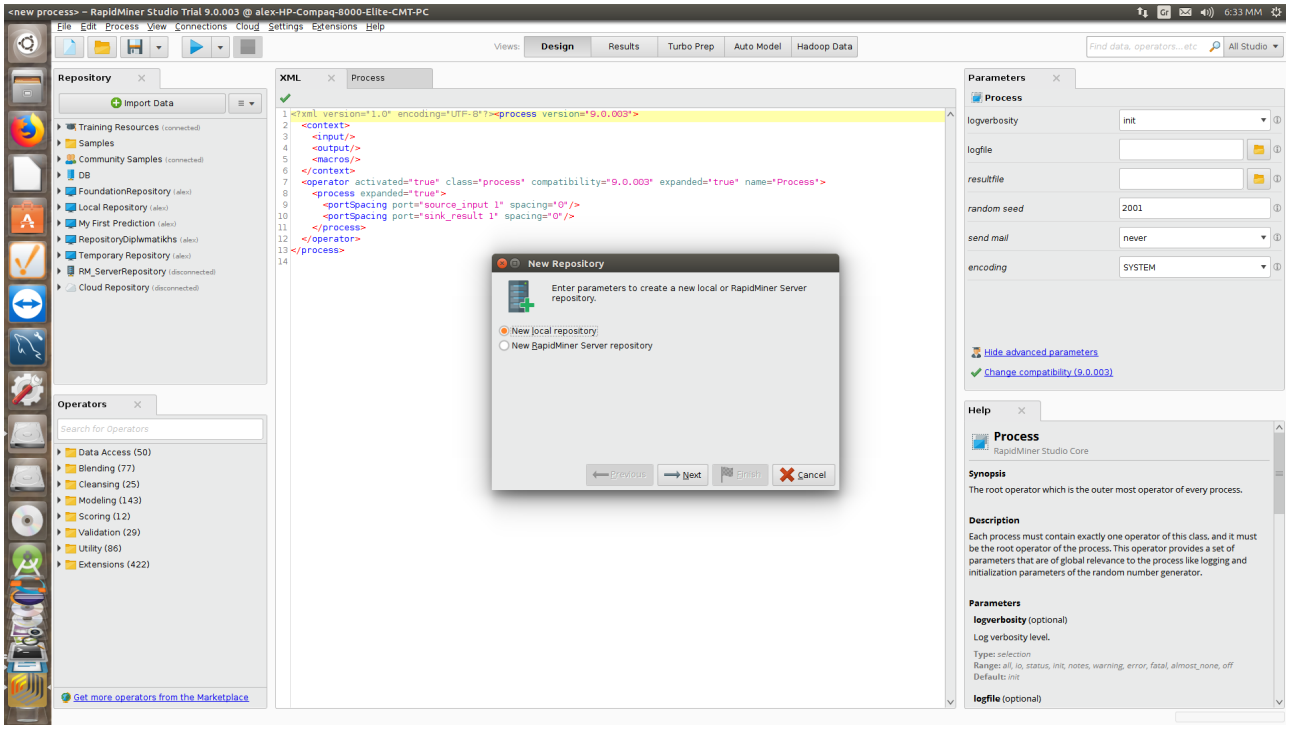
Εικόνα 87: Alarm Email (Wind Speed)

3.5 Εκτέλεση της Διαδικασίας στο Rapid Miner Server

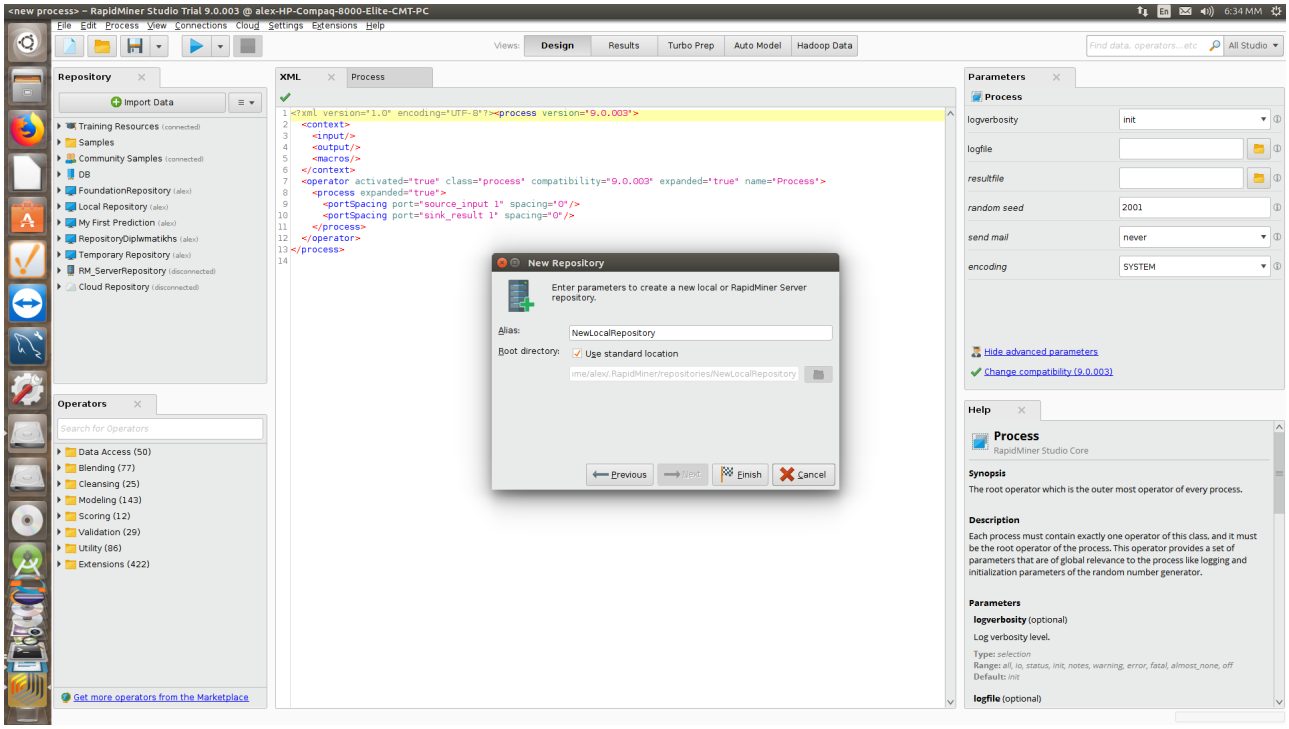
Τα process που περιγράψαμε στα δύο προηγούμενα υποκεφάλαια μπορούμε να τα εκτελέσουμε και στο Rapid Miner Server.

Πρώτα θα συνδέσουμε το Rapid Miner Studio με τον Rapid Miner Server:

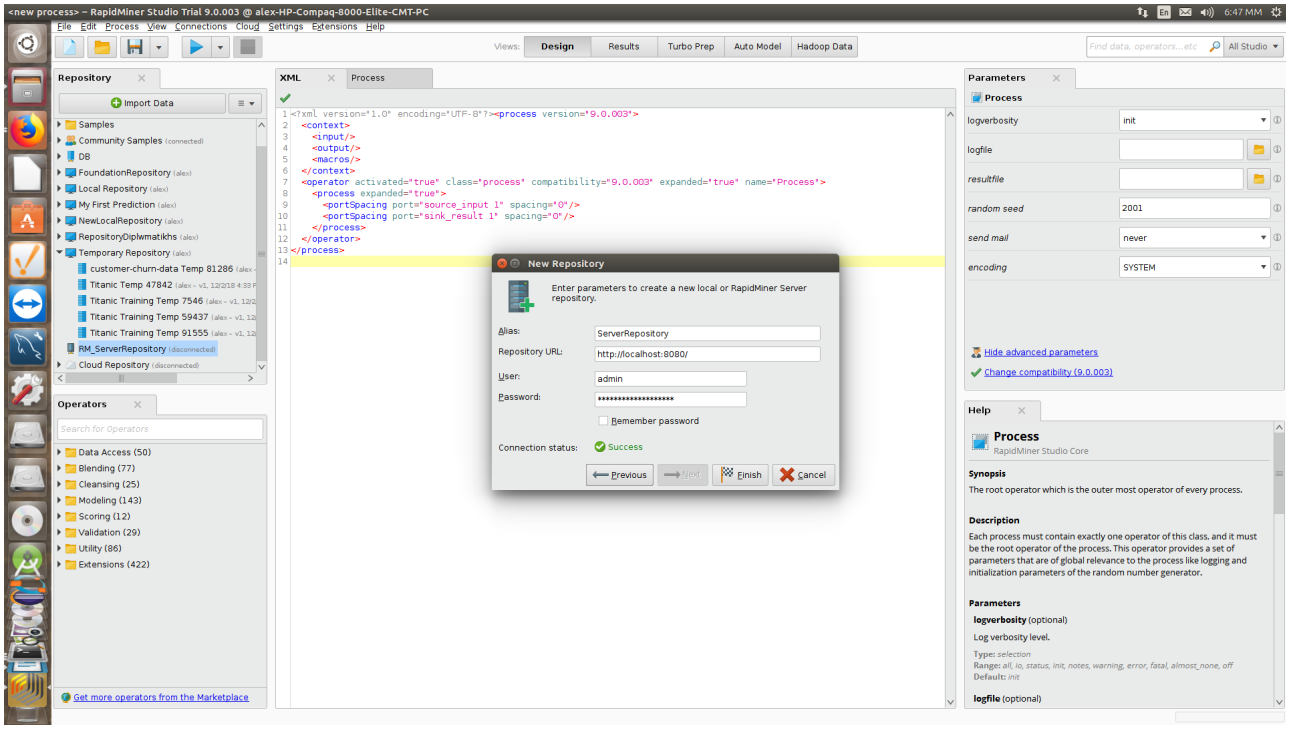
- Πηγαίνουμε στην επιλογή Connections → Connection to Database και συνδεόμαστε με τη Βάση Δεδομένων που χρησιμοποιήσαμε για τον Server.
- Πηγαίνουμε στην επιλογή Create repository → New RapidMiner Server repository και συμπληρώνουμε τα στοιχεία που μας ζητάει.



Εικόνα 88: Rapid Miner Server Repository 1

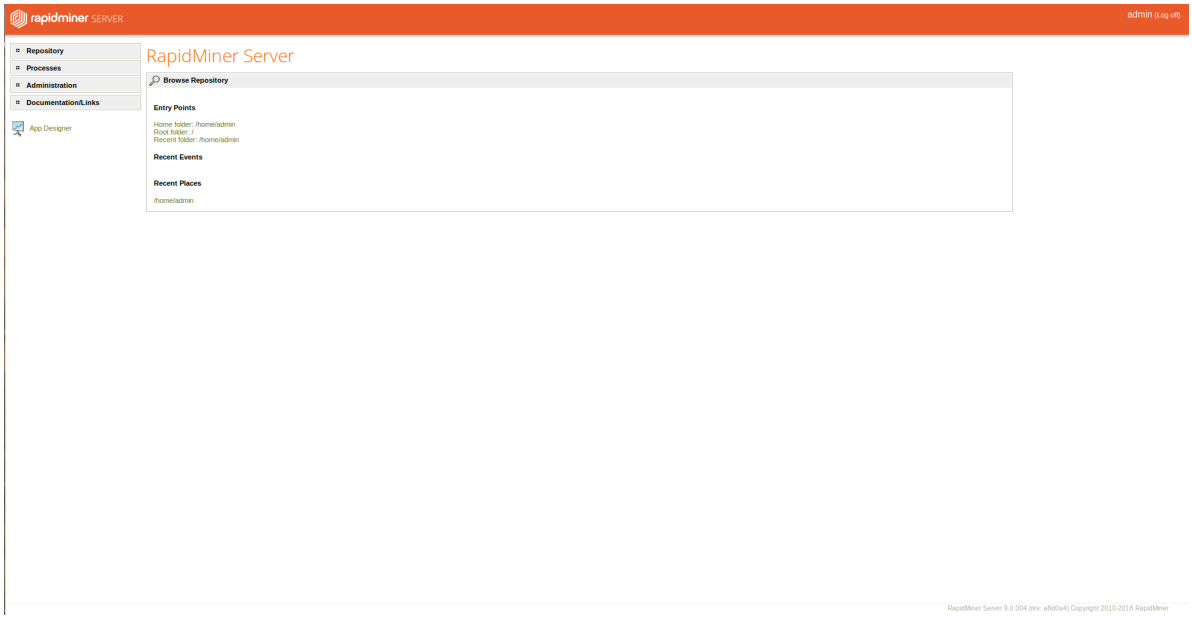


Εικόνα 89: Rapid Miner Server Repository 2



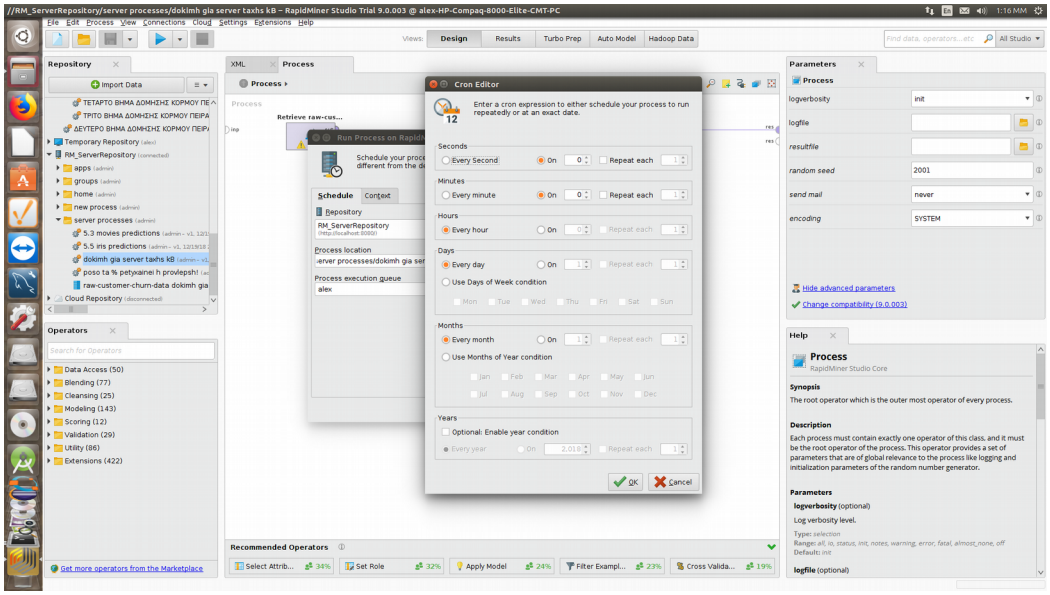
Εικόνα 90: Rapid Miner Server Repository 3

Αποθηκεύουμε στο Repository που κατασκευάσαμε, το process που θέλουμε να τρέξουμε στον Server και επιλέγουμε “**Run Process on Server**”.
Ανοίγουμε την αρχική σελίδα του διακομιστή RapidMiner, ανοίγουμε τον κατάλογο Processes και ελέγχουμε αν η μεταφορά του process ήταν επιτυχής.



Εικόνα 91: Rapid Miner Server

Τέλος, μπορούμε να προγραμματίσουμε κάθε πότε θα εκτελείται το process στον Server. Προς τούτο, επιλέγουμε “**Schedule Process on Server**”. Μπορούμε επί παραδείγματι να καθορίσουμε ότι το process θα εκτελείται κάθε 15 λεπτά, στις 05/01/2019. Φυσικά, στην περίπτωσή μας αυτό δεν έχει κάποιο ιδιαίτερο νόημα καθώς τα δεδομένα μας είναι αμετάβλητα.



Εικόνα 92: Scheduling Process on Rapid Miner Server

ΚΕΦΑΛΑΙΟ 4: ΕΠΙΠΡΟΣΘΕΤΕΣ ΠΡΟΒΛΕΨΕΙΣ

4.1 Πρόβλεψη Κριτικής Ταινιών

Έστω x = ένας τυχαίος user και y = μία τυχαία ταινία. Θέλουμε μία συνάρτηση f(x, y) που να σχετίζεται με τις παραπάνω μεταβλητές και εκχωρώντας τιμές σε αυτές να μπορεί να προβλέψει τη βαθμολογία που θα απέδιδε ένας συγκεκριμένος χρήστης σε μία συγκεκριμένη ταινία. Ισχύει δε f(x, y) ∈ [1,5]. Προς τούτο θα χρησιμοποιήσουμε δύο σύνολα δεδομένων:

1) Ένα σύνολο με τα ονόματα των ταινιών, καθένα απ' τα οποία αντιστοιχεί σε μία συγκεκριμένη movieid,

Row No.	MovieId	Tittle	date
1	1	Toy Story (1995)	01-Jan-1995
2	2	GoldenEye (1995)	01-Jan-1995
3	3	Four Rooms (1995)	01-Jan-1995
4	4	Get Shorty (1995)	01-Jan-1995
5	5	Copycat (1995)	01-Jan-1995
6	6	Shanghai Triad (Yao a yao yao dao waipo qiao) (1995)	01-Jan-1995
7	7	Twelve Monkeys (1995)	01-Jan-1995
8	8	Babe (1995)	01-Jan-1995
9	9	Dead Man Walking (1995)	01-Jan-1995
10	10	Richard III (1995)	22-Jan-1996
11	11	Seven (Se7en) (1995)	01-Jan-1995
12	12	Usual Suspects, The (1995)	14-Aug-1995
13	13	Mighty Aphrodite (1995)	30-Oct-1995
14	14	Postino, Il (1994)	01-Jan-1994
15	15	Mr. Holland's Opus (1995)	29-Jan-1996
16	16	French Twist (Gazon maudit) (1995)	01-Jan-1995
17	17	From Dusk Till Dawn (1996)	05-Feb-1996
18	18	White Balloon, The (1995)	01-Jan-1995
19	19	Antonia's Line (1995)	01-Jan-1995
20	20	Angels and Insects (1995)	01-Jan-1995
21	21	Muppet Treasure Island (1996)	16-Feb-1996
22	22	Braveheart (1995)	16-Feb-1996
23	23	Taxi Driver (1976)	16-Feb-1996
24	24	Rumble in the Bronx (1995)	23-Feb-1996
25	25	Birdcage, The (1996)	08-Mar-1996
26	26	Brothers McMullen, The (1995)	01-Jan-1995
27	27	Bad Boys (1995)	01-Jan-1995
28	28	Apollo 13 (1995)	01-Jan-1995

Εικόνα 93: MovieId & Titles Data

2) ένα σύνολο με τις κριτικές των χρηστών ως προς τις παραπάνω ταινίες.

ExampleSet (100000 examples, 0 special attributes, 4 regular attributes)

Row No.	UserId	MovieId	Rating	TimeStamp
1	196	242	3	881250949
2	186	302	3	891717742
3	22	377	1	878887116
4	244	51	2	880606923
5	166	346	1	886397596
6	298	474	4	884182806
7	115	265	2	881171488
8	253	465	5	891628467
9	305	451	3	886324817
10	6	86	3	883603013
11	62	257	2	879372434
12	286	1014	5	879781125
13	200	222	5	876042340
14	210	40	3	891035994
15	224	29	3	888104457
16	303	785	3	879485318
17	122	387	5	879270459
18	194	274	2	879539794
19	291	1042	4	874834944
20	234	1184	2	892079237
21	119	392	4	886176814
22	167	486	4	892738452
23	299	144	4	877881320
24	291	118	2	874833878
25	308	1	4	887736532
26	95	546	2	879196566
27	38	95	5	892430094

Εικόνα 94: Ratings Data

Ακολουθούμε τα παρακάτω βήματα:

- Εισάγουμε τα Data Sets που αναφέραμε παραπάνω τα οποία τα κατεβάζουμε απ' το διαδίκτυο και συγκεκριμένα απ' το link MovieLens/ GroupLens → older datasets → MovieLens 100k Dataset → Index of unzipped files → **u.data** & **u.item**.

- Χρησιμοποιούμε έναν τελεστή **Join** για να ενώσουμε τα δύο σύνολα σ' ένα. Στην είσοδο "**lef**" (left) συνδέουμε την έξοδο των **Ratings Data** ενώ στην "**rig**" (right) αυτήν των **Movies Data**. Ως join type επιλέγουμε inner και στα key attributes (επιλέγουμε την attribute που θα χρησιμοποιηθεί για τη συγχώνευση):

left key attributes	right key attributes
MovieId	MovieId

Πίνακας 5: key attributes, Join operator

- Προσθέτουμε έναν **Select Attributes** operator. Βάζουμε attribute filter type = single και attribute = TimeStamp. Επιλέγουμε invert selection ούτως ώστε να μην συμπεριληφθεί η παραπάνω παράμετρος στη συνέχεια της διαδικασίας.

- Εν συνεχεία, θα χρησιμοποιήσουμε τον τελεστή **Set Role** και θα θέσουμε ως label την attribute Rating. Στο set additional roles → edit list:

attribute name	target role
UserId	user identification
MovieId	movie identification

Πίνακας 6: set additional roles

- Με τον operator **Split Data** θα χωρίσουμε τα δεδομένα μας σε δύο υποσύνολα: το πρώτο (Training Data), το οποίο αποτελεί το 95 % του αρχικού Data Set, θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου πρόβλεψης που θα εφαρμόσουμε στο επόμενο βήμα ενώ το δεύτερο (Testing Data) για την αξιολόγηση αυτού απ' τον τελεστή **Apply Model**.

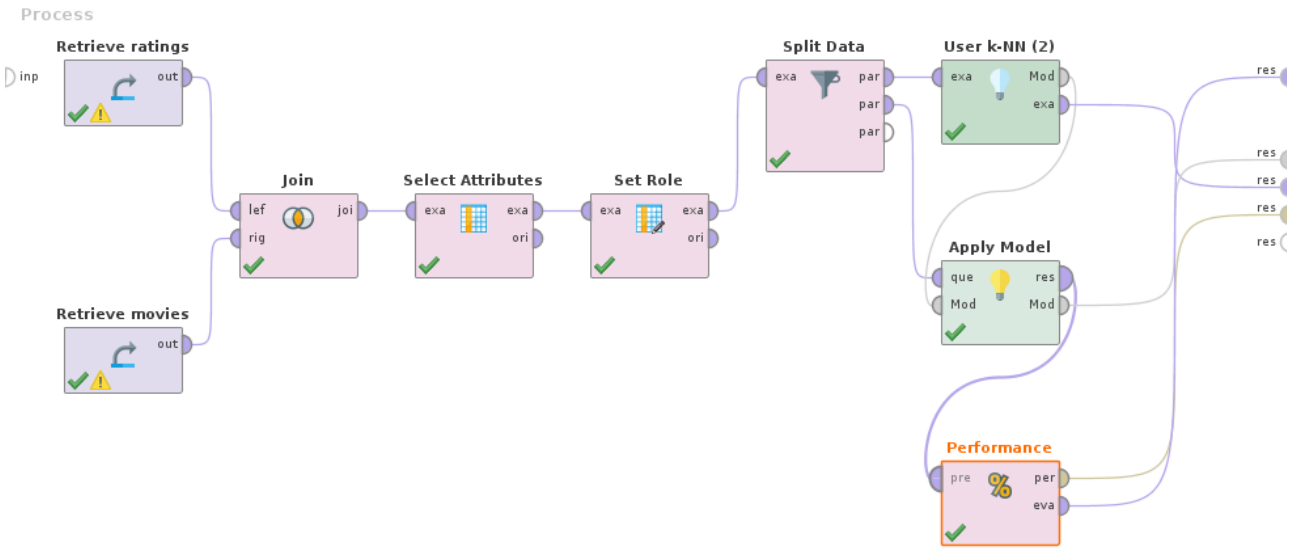
- Την πρώτη έξοδο "**par**" (partition) του **Split Data** τη συνδέουμε με την είσοδο "**exa**" (example set) του **User k-NN** (Αλγόριθμος k- Πλησιέστερων Γειτόνων). Αφήνουμε τις Παραμέτρους ως έχουν, δηλαδή **k = 80** (ούτως ώστε ο αριθμός των γειτόνων που θα λάβουμε υπόψη μας ως προς την πρόβλεψη να μην είναι ούτε πολύ μεγάλος αλλά ούτε και πολύ μικρός), **Min Rating = 1** (εφόσον οι βαθμολογίες ξεκινούν απ' το 1) και **Range = 4** (καθώς ο μέγιστος βαθμός είναι το 5). Τέλος, ενώνουμε την έξοδο "**Mod**" (Model) με την είσοδο "**Mod**" του **Apply Model** και την έξοδο "**exa**" με την τελική έξοδο της διαδικασίας, "**res**", ούτως ώστε να εμφανιστεί στα results ο πίνακας με τις βαθμολογίες που προέβλεψε το μοντέλο μας.

- Για την Αξιολόγηση του μοντέλου αλλά και για την εξαγωγή Προβλέψεων, εφαρμόζουμε τον τελεστή **Apply Model**. Ο συγκεκριμένος τελεστής, δέχεται το μοντέλο απ' την είσοδο "**Mod**" και το υποσύνολο Testing Data για την αξιολόγησή του απ' την είσοδο "**que**" (query set).

- Έπειτα, μέσω της εξόδου “**res**”, ο **Apply Model** συνδέεται με τον **Performance (Performance (Rating Prediction))** operator, ο οποίος δέχεται τα δεδομένα απ’ την είσοδό του που είναι η “**pre**” (predictions). Στις Παραμέτρους, βάζουμε **Min Rating = 1** και **Range = 4** (για τους λόγους που προαναφέραμε). Τέλος χρησιμοποιούμε τις δύο εξόδους: “**per**” (performance) και “**eva**” (evaluation measures) (ώστε να λάβουμε στα αποτελέσματα έναν πίνακα σφαλμάτων) και κατά αυτόν τον τρόπο ολοκληρώνουμε το process.

- Το **RMSE (Root Mean Square Error) = 0,937**.

- Παραθέτουμε το Process και τα σχετικά Results:



Εικόνα 95: Διαδικασία Πρόβλεψης Βαθμολογίας Ταινιών

ExampleSet (95000 examples, 3 special attributes, 2 regular attributes) Filter (95,000 / 95,000 example

Row No.	Rating	UserId	MovieId	Tittle	date
1	3	196	242	Kolya (1996)	24-Jan-1997
2	3	186	302	L.A. Confidential (1997)	01-Jan-1997
3	1	22	377	Heavyweights (1994)	01-Jan-1994
4	2	244	51	Legends of the Fall (1994)	01-Jan-1994
5	4	298	474	Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1963)	01-Jan-1963
6	2	115	265	Hunt for Red October, The (1990)	01-Jan-1990
7	5	253	465	Jungle Book, The (1994)	01-Jan-1994
8	3	305	451	Grease (1978)	01-Jan-1978
9	3	6	86	Remains of the Day, The (1993)	01-Jan-1993
10	2	62	257	Men in Black (1997)	04-Jul-1997
11	5	286	1014	Romy and Michele's High School Reunion (1997)	25-Apr-1997
12	5	200	222	Star Trek: First Contact (1996)	22-Nov-1996
13	3	210	40	To Wong Foo, Thanks for Everything! Julie Newmar (1995)	01-Jan-1995
14	3	224	29	Batman Forever (1995)	01-Jan-1995
15	3	303	785	Only You (1994)	01-Jan-1994
16	5	122	387	Age of Innocence, The (1993)	01-Jan-1993
17	2	194	274	Sabrina (1995)	01-Jan-1995
18	4	291	1042	Just Cause (1995)	01-Jan-1995
19	2	234	1184	Endless Summer 2, The (1994)	01-Jan-1994
20	4	119	392	Man Without a Face, The (1993)	01-Jan-1993
21	4	299	144	Die Hard (1988)	01-Jan-1988
22	2	291	118	Twister (1996)	10-May-1996
23	2	95	546	Broken Arrow (1996)	09-Feb-1996
24	5	38	95	Aladdin (1992)	01-Jan-1992
25	2	102	768	Casper (1995)	01-Jan-1995
26	4	63	277	Restoration (1995)	01-Jan-1995
27	5	160	234	Jaws (1975)	01-Jan-1975
28	3	50	246	Chasing Amy (1997)	01-Jan-1997

Εικόνα 96: Αποτελέσματα Πρόβλεψης Βαθμολογίας Ταινιών

Δεδομένου του RMSE συμπεραίνουμε ότι η πρόβλεψη μας θα έχει ένα σφάλμα πρακτικά ίσο με $\pm$ 1. Αυτό συνεπάγεται πως εάν για παράδειγμα έχουμε προβλέψει ότι ο User 196 θα βαθμολογήσει τη Movie 242 με 3, στην πραγματικότητα η βαθμολογία που θα δώσει θα κυμαίνεται από 2 έως 4 με πιο πιθανή την τιμή 3. Δεδομένου του γεγονότος ότι η βαθμολογία κυμαίνεται από 1 έως 5, το σχετικό σφάλμα θα είναι 20 %, ένα σφάλμα που δε δύναται να θεωρηθεί αμελητέο. Προφανώς, αυτό οφείλεται στην ιδιαίτερη φύση του συγκεκριμένου προβλήματος καθώς εξαρτάται απ' τις υποκειμενικές και πολλές φορές ευμετάβλητες προτιμήσεις του εκάστοτε user. Ένα μοντέλο που θα είχε ως είσοδο περισσότερα δεδομένα για την ιδιοσυγκρασία του κάθε χρήστη και θα προέβαινε στην αναλυτική επεξεργασία αυτών, θα έδινε μεν καλύτερα αποτελέσματα, θα απαιτούσε δε πολύ μεγαλύτερη υπολογιστική ισχύ και τα κατάλληλα μέσα για την ταυτόχρονη παρακολούθηση της δραστηριότητας των χρηστών που υπάγονται στο σύστημά μας (αυτό θα μπορούσε να γίνει με τη βοήθεια των σύγχρονων μέσων κοινωνικής δικτύωσης όπως το fb ή και με τη χρήση των δεδομένων που συλλέγονται από μηχανές αναζήτησης όπως η google).

4.2 Το Καλάθι της Νοικοκυράς (Act of Cross Selling)

Σε αυτό το πείραμα σκοπός μας είναι να συσχετίσουμε την αγορά ενός προϊόντος Α με αυτήν ενός προϊόντος Β. Για να γίνει πλήρως κατανοητή η διαδικασία που θα ακολουθηθεί θα αναφέρουμε ένα απλουστευμένο παράδειγμα το οποίο χαρακτηρίζεται από έναν πολύ μικρό αριθμό αριθμό πελατών και προϊόντων (πέντε και στις δύο περιπτώσεις). Παραθέτουμε τον παρακάτω πίνακα:

Transaction	Pasta	Cheese	Pasta Sauce	Wine	Milk
Customer 1	1	1	1	0	0
Customer 2	1	0	1	0	0
Customer 3	1	1	0	1	0
Customer 4	0	0	0	0	1
Customer 5	1	1	1	1	0

Πίνακας 7: Act of Cross Selling Example

Α) Σχετική Συχνότητα (Πιθανότητα):

Είδος Τροφής	Σχετική Συχνότητα
Pasta	80 %
Cheese	60 %
Sauce	60 %
Wine	40 %
Milk	20 %

Πίνακας 8: Σχετική Συχνότητα

Β) Support:

Έστω ότι θέλουμε να υπολογίσουμε την πιθανότητα ένας πελάτης **x** ν’ αγοράσει Pasta συν ένα απ’ τα υπόλοιπα τρόφιμα του παραπάνω πίνακα (Support). Υπάρχουν οι εξής συνδυασμοί:

Συνδυασμοί	Support
{Pasta, Cheese}	3/5 = 60 %
{Pasta, Sauce}	3/5 = 60 %
{Pasta, Wine}	2/5 = 40 %
{Pasta, Milk}	0 %

Πίνακας 9: Support

Γ) Confidence:

Έστω ότι θέλουμε να υπολογίσουμε την πιθανότητα κάποιος που αγόρασε Pasta (δεδομένο) να αγόρασε επίσης και Cheese. Είναι:

{Pasta} → {Cheese} = # {Pasta & Cheese} / (# Pasta) = 75 %

Με την ίδια λογική μπορούν να υπολογιστούν και τα υπόλοιπα Confidence.

Τώρα μπορούμε να προχωρήσουμε στην εκτέλεση του πειράματος:

- Κατεβάζουμε απ' το διαδίκτυο το Dataset Extended Bakery (5000-out2.csv). Εν συνεχεία το κάνουμε import στο RapidMiner και το αποθηκεύουμε στον φάκελο με τα Δεδομένα μας. Απ' τον συγκεκριμένο φάκελο έχουμε τη δυνατότητα να το εισάγουμε, με τη χρήση Drag & Drop, στο process για να το επεξεργαστούμε, όποτε επιθυμούμε.

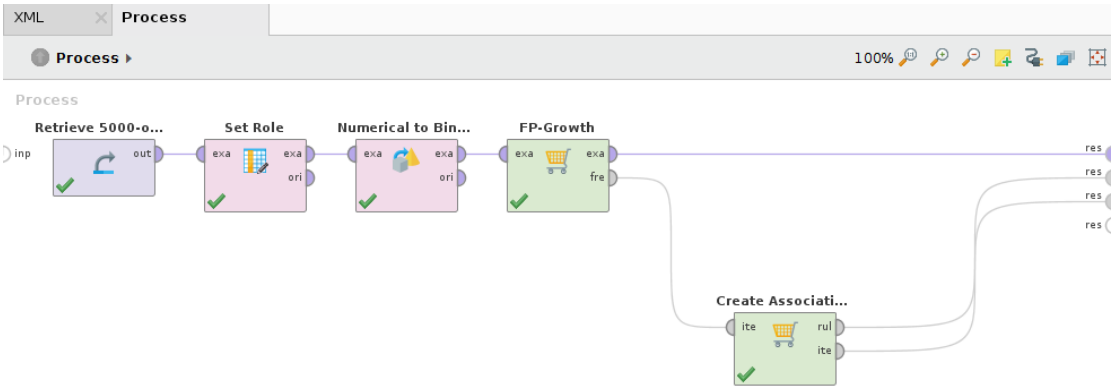
- Χρησιμοποιούμε τον **Set Role** operator και θέτουμε attribute name = customers, target role = id.

- Προσθέτουμε τον τελεστή **Numerical to Binominal** (μετατρέπει τα 0 → false (δεν αγόρασε το προϊόν) και τα 1 → true (αγόρασε το προϊόν)) και επιλέγουμε attribute filter type = all.

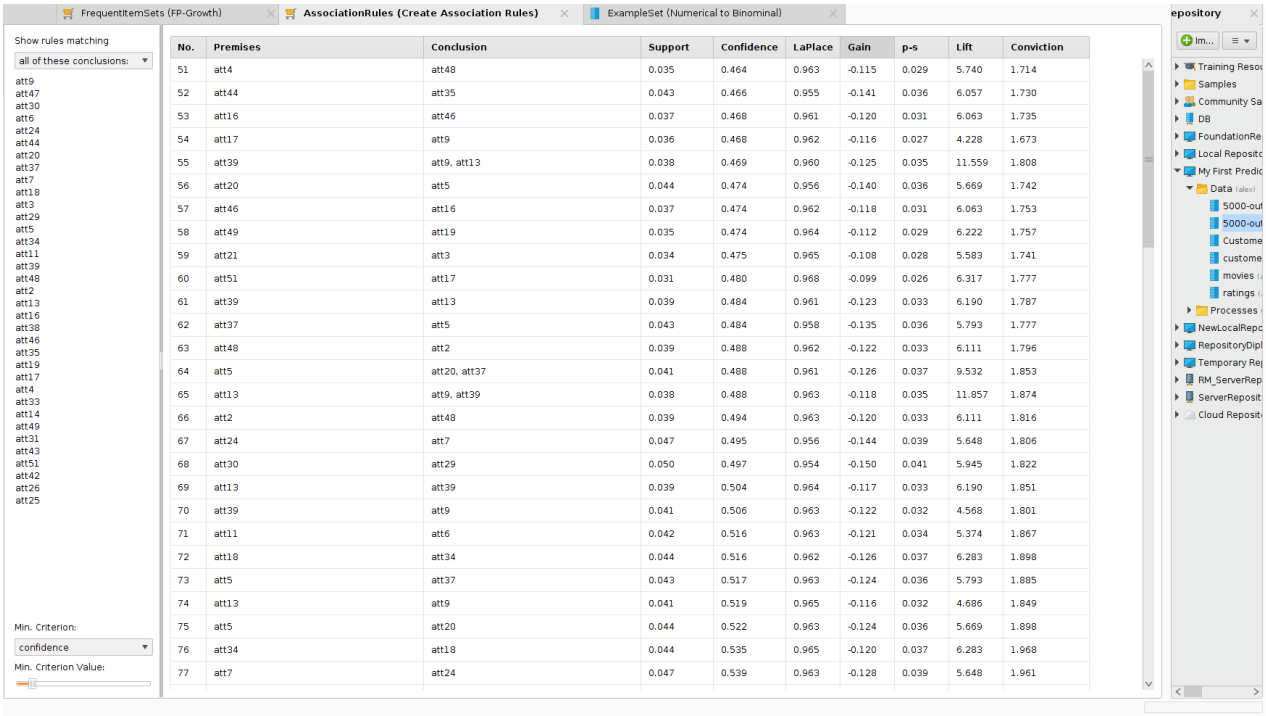
- Χρησιμοποιούμε τον **FP-Growth** operator ο οποίος είναι ένας αποτελεσματικός αλγόριθμος για τον υπολογισμό συχνά συνυπαρχόντων στοιχείων σε μια βάση δεδομένων συναλλαγών. Στις Παραμέτρους βάζουμε minSupport = 0,04. Με αυτόν τον τρόπο μπορούμε να συμπεριλάβουμε ένα αρκετά μεγάλο πλήθος συνδυασμών.

- Τέλος, με τον τελεστή **Create Association Rules**, δημιουργούμε κανόνες συσχετισμού ανάμεσα στα διάφορα στοιχεία του Dataset. Θέλουμε Confidence = 0,4, δηλαδή από 40 % και πάνω ούτως ώστε να μη λάβουμε υπόψη μας συσχετίσεις που δεν έχουν αξιοσημείωτη ισχύ.

- Παρακάτω παραθέτουμε το Process και τα Results:

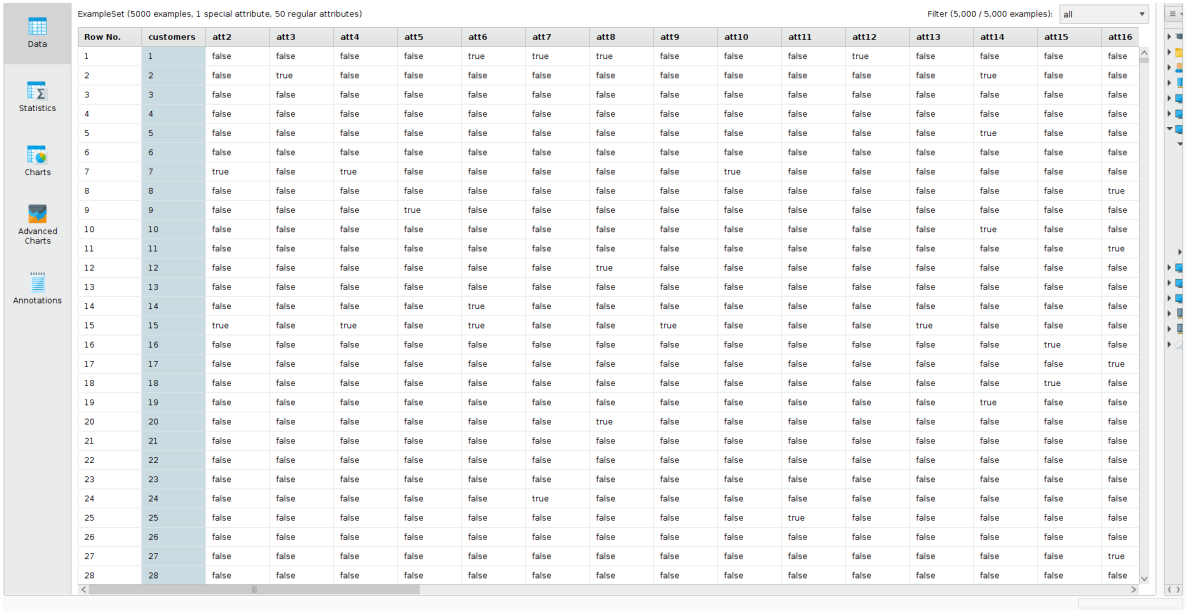


Εικόνα 97: Act of Cross Selling Process



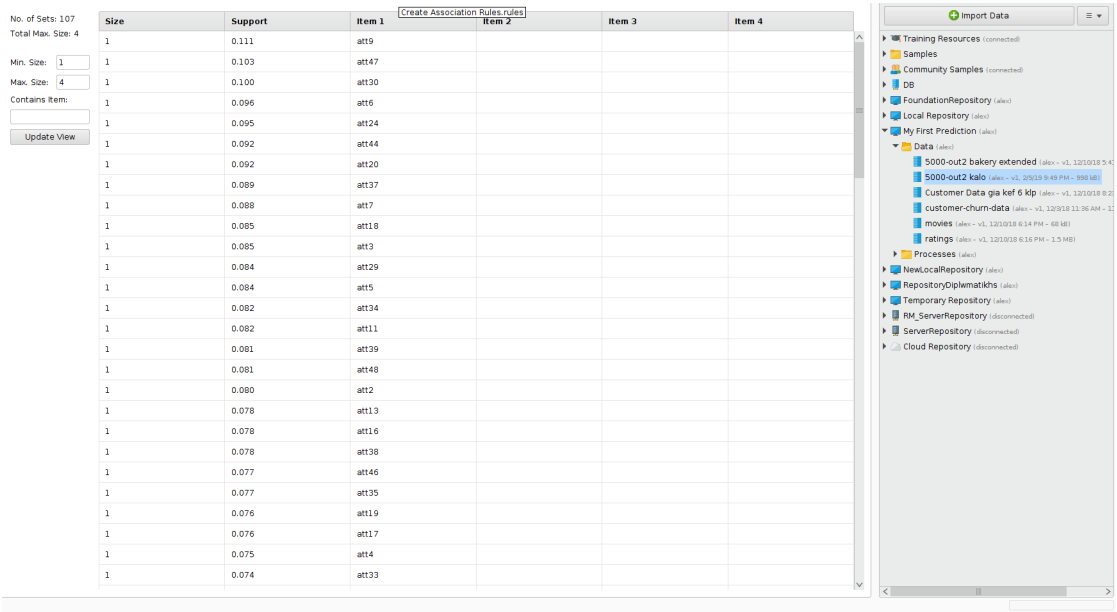
Εικόνα 98: Association Rules

Στην εικόνα 98 είναι καταγεγραμμένοι όλοι οι συσχετισμοί προϊόντων που έχουν Confidence >= 40 % (το κατώφλι που δηλώσαμε στον τελεστή **Create Association Rules**). Επί παραδείγματι αν ένας πελάτης αγοράσει το att4 συνεπάγεται πως υπάρχει μία πιθανότητα 46,4 % να αγοράσει και το att48. Κατά αυτόν τον τρόπο προκύπτουν κανόνες με τους οποίους μπορούμε να κατανοήσουμε τις τάσεις των καταναλωτών και να υλοποιήσουμε στρατηγικές για τη βέλτιστη προώθηση των προϊόντων μας.

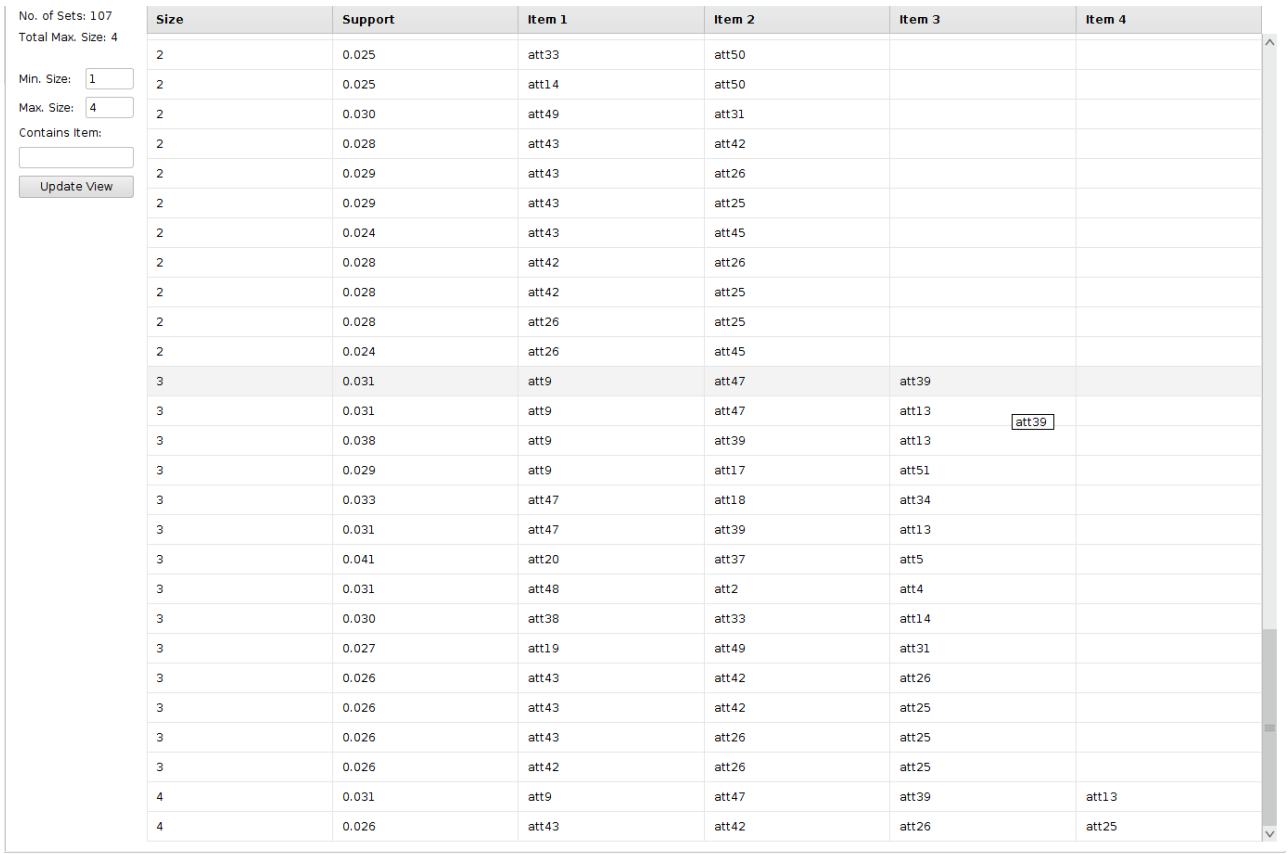


Εικόνα 99: ExampleSet (Numerical to Binominal)

Στην εικόνα 99 βλέπουμε για κάθε έναν απ' τους customers, ποια προϊόντα έχει επιλέξει (δηλώνονται στον πίνακα ως true).



Εικόνα 100: FP-Growth



Size	Support	Item 1	Item 2	Item 3	Item 4
2	0.025	att33	att50		
2	0.025	att14	att50		
2	0.030	att49	att31		
2	0.028	att43	att42		
2	0.029	att43	att26		
2	0.029	att43	att25		
2	0.024	att43	att45		
2	0.028	att42	att26		
2	0.028	att42	att25		
2	0.028	att26	att25		
2	0.024	att26	att45		
3	0.031	att9	att47	att39	
3	0.031	att9	att47	att13	
3	0.038	att9	att39	att13	
3	0.029	att9	att17	att51	
3	0.033	att47	att18	att34	
3	0.031	att47	att39	att13	
3	0.041	att20	att37	att5	
3	0.031	att48	att2	att4	
3	0.030	att38	att33	att14	
3	0.027	att19	att49	att31	
3	0.026	att43	att42	att26	
3	0.026	att43	att42	att25	
3	0.026	att43	att26	att25	
3	0.026	att42	att26	att25	
4	0.031	att9	att47	att39	att13
4	0.026	att43	att42	att26	att25

Εικόνα 101: FP-Growth 2

Με min Support = 0,04, λαμβάνουμε τον παραπάνω πίνακα (εικόνες 101 και 102) που διαθέτει σύνολα από ένα έως και τέσσερα αντικείμενα. Αν είχαμε ορίσει min Support = 0,01 θα προέκυπτε ένας πίνακας με σύνολα από ένα έως και πέντε αντικείμενα ενώ αν ήταν min Support = 0,15 από ένα έως και τρία.

Η παραπάνω διαδικασία θα μπορούσε να ακολουθηθεί για οιοδήποτε σύστημα που σχετίζεται με την προώθηση προϊόντων και το marketing. Ακόμη, μεγάλες επιχειρήσεις που θέλουν να επεκταθούν και σε άλλους τομείς, μπορούν να συμβουλευτούν τέτοιου είδους αλγορίθμους για την ορθότερη λήψη αποφάσεων.

ΚΕΦΑΛΑΙΟ 5: ΣΥΜΠΕΡΑΣΜΑΤΑ

Απ’ την πειραματική διαδικασία (Κεφάλαια 3 και 4) συμπεραίνουμε πως η χρήση λογισμικών Αναλυτικής και Εξόρυξης Δεδομένων όπως το RapidMiner, μπορεί να αποδειχθεί εξαιρετικά προσοδοφόρα σε μεγάλο εύρος δραστηριοτήτων υψίστης κοινωνικής, ανθρωπιστικής και οικονομικής σημασίας. Με ανάλογες εφαρμογές είμαστε σε θέση να εξάγουμε ποιοτικά συμπεράσματα για φαινόμενα τα οποία μας είναι άγνωστα και τα οποία δεν δύναται να εξεταστούν με τις κλασικές πειραματικές μεθόδους (π.χ. χαοτικά φαινόμενα όπου δεν μπορούν να προσδιοριστούν οι βασικές παράμετροι ή για τα οποία δεν είναι δυνατόν να διεξαχθούν πειράματα υπό συνθήκες πλήρως ελεγχόμενες). Ακόμη, η συνεισφορά των Μοντέλων Πρόβλεψης στην Ιατρική όσον αφορά σοβαρές ασθένειες που χρίζουν ταχύτατης αντιμετώπισης, στο Marketing για τη σωστή προώθηση προϊόντων αλλά και για την αποτελεσματικότερη στόχευση ως προς τις ομάδες πελατών, στο Εμπόριο για την κατανόηση της αγοράς και των μεταβολών των αναγκών του αγοραστικού κοινού καθώς και στη Ναυτιλία για την εξασφάλιση ασφαλέστερης μεταφοράς επιβατών και προϊόντων, είναι προφανής. Τέλος, η τοποθέτηση συναγερμών (Set Alarms) μπορεί να χρησιμοποιηθεί σε τομείς όπως η Ναυτιλία ούτως ώστε να προληφθούν και να αποφευχθούν καταστροφές που ενδέχεται να έχουν τόσο οικονομικό κόστος όσο και απώλειες ανθρώπινων ζωών.

Βιβλιογραφία

[1] J. Belissent, “Getting Clever About Smart Cities: New Opportunities Require New Business Models,” Forrester Research Inc., New York, NY, USA, 2010.
[2] L. Atzori, A. Iera, and G. Morabito, “The Internet of Things: A survey,” Comput. Netw., vol. 54, no. 15, pp. 2787–2805, 2010.
[3] Smart Networked Objects and Internet of Things, white paper, Association Instituts Carnot, Greece, 2011.
[4] T. S. Lopez, D. C. Ranasinghe, M. Harrison, and D. McFarlane, “Adding sense to the Internet of Things an architecture framework for smart object systems,” Pers. Ubiquitous Comput., vol. 16, no. 3, pp. 291–308, 2012.
[5] J. Yick, B. Mukherjee, and D. Ghosal, “Wireless sensor network survey,” Comput. Netw., vol. 52, pp. 2292–2330, 2008.
[6] J.-P. Vasseur and A. Dunkels, Interconnecting Smart Objects with IP: The Next Internet. New York: Elsevier, 2010.
[7] J. Heidemann, D. Estrin, R. Govindan, and S. Kumar, “Next century challenges: Scalable coordination in sensor networks,” in Proc. 5th Annu. ACM/IEEE Int. Conf. Mobile Comput. Netw., Seattle, WA, USA, 1999, pp. 263–270.
[8] A. Dunkels, F. Osterlind, and Z. He, “An adaptive communication architecture for wireless sensor networks,” in Proc. 5th ACM Conf. Networked Embedded Sensor Syst. (SenSys), Sydney, Australia, Nov. 2007.
[9] J. Jin, Y. W. Law, W. H. Wang, and M. Palaniswami, “A hierarchical transport architecture for wireless sensor networks,” in Proc. 4th Int. Conf. Intell. Sensors Sensor Netw. Inf. Proc. (ISSNIP), Sydney, Australia, Dec. 2008.
[10] J. Hui and D. Culler, “IP is dead, long live IP for wireless sensor networks,” in Proc. 6th ACM Conf. Networked Embedded Sensor Syst. (SenSys), Raleigh, NC, USA, Nov. 2008.
[11] J. Jin, J. Gubbi, T. Luo, and M. Palaniswami, “Network architecture and QoS issues in the Internet of Things for a smart city,” in Proc. 12th Int. Symp. Commun. Inf. Technol. (ISCIT), Gold Coast, Australia, Oct. 2012.
[12] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, “Internet of Things (IoT): A vision, architectural elements, and future directions,” Future Gener. Comput. Syst., vol. 29, pp. 1645–1660, 2013.
[13] M. Srivastava, T. Abdelzaher, and B. Szymanski, “Human-centric sensing,” Philos. Trans. Roy. Soc. A, vol. 370, no. 1958, pp. 176–197, 2012.
[14] R. V. Kulkarni, A. Forster, and G. K. Venayagamoorthy, “Computational intelligence in wireless sensor networks: A survey,” IEEE Commun. Surveys Tutorials, vol. 13, no. 1, pp. 68–96, Feb. 2011.
[15] M. Conti, S. Chong, S. Fdida, W. Jia, H. Karl, Y.-D. Lin et al., “Research challenges towards the Future Internet,” Comput. Commun., vol. 34, no. 18, pp. 2115–2134, Dec. 2011.
[16] J. Rexford and C. Dovrolis, “Future Internet architecture: Clean-slate versus evolutionary research,” Commun. ACM, vol. 53, no. 9, pp. 36–40, Sep. 2010.
[17] A. Puech, A. Venet, and C. Gepner. (2012, May). Smart Grains Parking Solutions [Online]. Available: http://www.smartgrains.com.
[18] S. Lee, D. Yoon, and A. Ghosh, “Intelligent parking lot application using wireless sensor networks,” in Proc. Int. Symp. Collaborative Technol. Syst. (CTS), 2008, pp. 48–57.

[19] N. M. M. K. Chowdhury and R. Boutaba, “Network virtualization: State of the art and research challenges,” IEEE Commun. Mag., vol. 47, no. 7, pp. 20–26, Jul. 2009.
[20] E. D. Poorter, I. Moerman, and P. Demeester, “Enabling direct connectivity between heterogeneous objects in the Internet of Things through a network-service-oriented architecture,” EURASIP J. Wireless Commun. Netw., vol. 2011, p. 61, 2011.
[21] H. Miedema, “Annoyance caused by environmental noise: Elements for evidence-based noise policies,” J. Social Issues, vol. 63, no. 1, pp. 41–57, 2007.
[22] C. Pham and P. Cousin, “Streaming the sound of smart cities: Experimentations on the SmartSantander test-bed,” in Proc. 2013 IEEE Int. Conf. GreenCom-iThings-CPSCom, Beijing, China, Aug. 2013.
[23] J. Gubbi, S. Marusic, Y. W. Law, A. S. Rao, and M. Palaniswami, “A pilot study of urban noise monitoring architecture using wireless sensor networks,” in Proc. Int. Conf. Advances Comput. Commun. Informat. (ICACCI), Bangkok, Thailand, Sep. 2013.
[24] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, “Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility,” Future Gener. Comput. Syst., vol. 25, pp. 599–616, 2009.
[25] M. Palaniswami, S. Marusic, J. Gubbi, and Y. W. Law, “Noise mapping: Designing an urban information architecture to record and map noise pollution,” Univ. Melbourne, Melbourne, Australia, Tech. Rep., May 2011.
[26] A. Gluhak, S. Krco, M. Nati, D. Pfisterer, N. Mitton, and T. Razafindralambo, “A survey on facilities for experimental Internet of Things research,” IEEE Commun. Mag., vol. 49, no. 11, pp. 58–67, Nov. 2011.
[27] Networking and Information Technology Research and Development (NITRD) Program. (2015). Smart and Connected Communities Framework. [Online]. Available: https://www.nitrd.gov/sccc/materials/scccframework.pdf.
[28] Y. Sun, Y. Xia, H. Song, and R. Bie, ‘‘Internet of things services for small towns,’’ in Proc. Int. Conf. Identificat., Inf. Knowl. Internet Things (IIKI), Oct. 2014, pp. 92–95.
[29] A. J. Jara, Y. Sun, H. Song, R. Bie, D. Genooud, and Y. Bocchi, ‘‘Internet of Things for cultural heritage of smart cities and smart regions,’’ in Proc. IEEE 29th Int. Conf. Adv. Inf. Netw. Appl. Workshops (WAINA), Mar. 2015, pp. 668–675.
[30] H. Song, R. Srinivasan, T. Sookoor, and S. Jeschke, Smart Cities: Foundations and Principles. Hoboken, NJ, USA: Wiley, 2016.
[31] U.S. Department of Transportation. (2013). Livability 101. [Online]. Available: http://www.dot.gov/livability/101.
[32] UNESCO. (2014). Protecting Our Heritage and Fostering Creativity.
[Online]. Available:
http://en.unesco.org/themes/protecting-our-heritage-and-fostering creativity.
[33] S. Simon. (2015). In the company of scholars: ‘Are We Losing Our Past or Our Future? Sustainable Preservation of Cultural Heritage’. [Online]. Available: http://ipch.yale.edu/news/company-scholars-are-we-losing-our-past-or-our-future
sustainable-preservation-cultural.
[34] N. D. Milder and A. Dane, ‘‘Revitalizing small towns: Resolving downtown challenges,’’ Econ. Develop. J., 2013. [Online]. Available:
http://www.rural-design.org/resource/revitalizing-small-towns-resolving-downtown
challenges.
[35] Annex, ‘‘World commission on environment and development,’’ Oxford Univ. Press, New York, NY, USA, 1987. Tech. Rep. A/42/427, 1987.
[36] Sustainable Communities. (2016). About Sustainable Communities. [Online]. Available: http://www.sustainable.org/about.

[37] A. T. Campbell et al., ‘‘The rise of people-centric sensing,’’ IEEE Internet Comput., vol. 12, no. 4, pp. 12–21, Jul./Aug. 2008.
[38] A. T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, and R. A. Peterson, ‘‘People-centric urban sensing,’’ in Proc. 2nd Annu. Int. Workshop Wireless Internet (WICON), New York, NY, USA, 2006, Art. no. 18. [Online]. Available: http://doi.acm.org/10.1145/ 1234161.1234179.
[39] R. K. Ganti, F. Ye, and H. Lei, ‘‘Mobile crowdsensing: Current state and future challenges,’’ IEEE Commun. Mag., vol. 49, no. 11, pp. 32–39, Nov. 2011.
[40] M. Pouryazdan, B. Kantarci, T. Soyata, and H. Song, ‘‘Anchor-assisted and vote-based trustworthiness assurance in smart city crowdsensing,’’ IEEE Access, vol. PP, no. 99, pp. 1–1, doi: 10.1109/ACCESS.2016.2519820.
[41] Y. Sun et al., ‘‘Organizing and querying the big sensing data with event-linked network in the Internet of Things,’’ Int. J. Distrib. Sensor Netw., vol. 2014, 2014, Art. no. 218521.
[42] Y. Sun and A. J. Jara, ‘‘An extensible and active semantic model of information organizing for the Internet of Things,’’ Pers. Ubiquitous Comput., vol. 18, no. 8, pp. 1821 1833, 2014.
[43] Y. Sun, H. Yan, C. Lu, R. Bie, and Z. Zhou, ‘‘Constructing the Web of events from raw data in the Web of things,’’ Mobile Inf. Syst., vol. 10, no. 1, pp. 105–125, 2014.
[44] Y. Sun, C. Lu, R. Bie, and J. Zhang, ‘‘Semantic relation computing theory and its application,’’ J. Netw. Comput. Appl., vol. 59, pp. 219–229, Jan. 2016.
[45] H. Zhuge and Y. Sun, ‘‘The schema theory for semantic link network,’’ Future Generat. Comput. Syst., vol. 26, no. 3, pp. 408–420, 2010.
[46] Cyber Physical Systems Vision Statement, Netw., Inf. Technol. Res., Develop. Program., Washington, DC, USA, Jun. 2015.
[47] E. D. Simmon et al., ‘‘A vision of cyber-physical cloud computing for smart networked systems,’’ Nat. Inst. Standards Technol. (NIST), Tech. Rep. 7951, 2013.
[48] I. Butun, M. Erol-Kantarci, B. Kantarci, and H. Song, ‘‘Cloud-centric multi-level authentication as a service for secure public safety device networks,’’ IEEE Commun. Mag., vol. 54, no. 4, Apr. 2016.
[49] H. Song, G. Fink, S. Jeschke, and G. Rosner, Security and Privacy in Cyber-Physical Systems: Foundations and Applications. Chichester, U.K.: Wiley, 2016.
[50] K. Ashton. (2009, Jun.). Internet of things. RFID J. [Online]. Available: http://www.rfidjournal.com/articles/view?4986
[51] R. van Kranenburg, The Internet of Things: A Critique of Ambient Technology and the All-Seeing Network of RFID. Amsterdam, The Netherlands: Institute of Network Cultures, 2007.
[52] R. van Kranenburg, E. Anzelmo, A. Bassi, D. Caprio, S. Dodson, and M. Ratto, “The internet of things,” in Proc. 1st Berlin Symp. Internet Soc., Berlin, Germany, 2011, pp. 25–27.
[53] Y. Li, M. Hou, H. Liu, and Y. Liu, “Towards a theoretical framework of strategic decision, supporting capability and information sharing under the context of Internet of Things,” Inf. Technol. Manage., vol. 13, no. 4, pp. 205–216, 2012.
[54] L. Tan and N. Wang, “Future internet: The internet of things,” in Proc. 3rd Int. Conf. Adv. Comput. Theory Eng. (ICACTE), Chengdu, China, Aug. 20–22, 2010, pp. V5-376–V5-380.
[55] X. Jia, O. Feng, T. Fan, and Q. Lei, “RFID technology and its applications in internet of things (IoT),” in Proc. 2nd IEEE Int. Conf. Consum. Electron., Commun. Netw. (CECNet), Yichang, China, Apr. 21–23, 2012, pp. 1282–1285.
[56] C. Sun, “Application of RFID technology for logistics on internet of things,” AASRI Procedia, vol. 1, pp. 106–111, 2012.

[57] E. W. T. Ngai, K. K. Moon, F. J. Riggins, and C. Y. Yi, “RFID research: An academic literature review (1995–2005) and future research directions,” Int. J. Prod. Econ., vol. 112, no. 2, pp. 510–520, 2008.
[58] S. Li, L. Xu, and X. Wang, “Compressed sensing signal and data acquisition in wireless sensor networks and internet of things,” IEEE Trans. Ind. Informat., vol. 9, no. 4, pp. 2177–2186, Nov. 2013.
[59] W. He and L. Xu, “Integration of distributed enterprise applications: A survey,” IEEE Trans. Ind. Informat., vol. 10, no. 1, pp. 35–42, Feb. 2014.
[60] D. Uckelmann, M. Harrison, and F. Michahelles, “An architectural approach towards the future internet of things,” in Architecting the Internet of Things. D. Uckelmann, M. Harrison, and F. Michahelles, Eds., New York, NY, USA: Springer, 2011, pp 1–24.
[61] S. Li, L. Xu, X. Wang, and J. Wang, “Integration of hybrid wireless networks in cloud services oriented enterprise information systems,” Enterp. Inf. Syst., vol. 6, no. 2, pp. 165–187, 2012.
[62] L. Wang, L. Xu, Z. Bi, and Y. Xu, “Data filtering for RFID and WSN integration,” IEEE Trans. Ind. Informat., vol. 10, no. 1, pp. 408–418, Feb. 2014.
[63] L. Ren, L. Zhang, F. Tao, X. Zhang, Y. Luo, and Y. Zhang, “A methodology towards virtualization-based high performance simulation platform supporting multidisciplinary design of complex products,” Enterp. Inf. Syst., vol. 6, no. 3, pp. 267–290, 2012.
[64] F. Tao, Y. Laili, L. Xu, and L. Zhang, “FC-PACO-RM: A parallel method for service composition optimal-selection in cloud manufacturing system,” IEEE Trans. Ind. Informat., vol. 9, no. 4, pp. 2023–2033, Nov. 2013.
[65] Q. Li, Z. Wang, W. Li, J. Li, C. Wang, and R. Du, “Applications integration in a hybrid cloud computing environment: Modelling and platform,” Enterp. Inf. Syst., vol. 7, no. 3, pp. 237–271, 2013.
[66] D. Bandyopadhyay and J. Sen, “Internet of things: Applications and challenges in technology and standardization,” Wireless Pers. Commun., vol. 58, no. 1, pp. 49–69, 2011.
[67] ITU NGN-GSI Rapporteur Group, Requirements for Support of USN Applications and Services in NGN Environment, Geneva, Switzerland: International Telecommunication Union (ITU), 2010.
[68] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, “Internet of things: Vision, applications and research challenges,” Ad Hoc Netw., vol. 10, no. 7, pp. 1497–1516, 2012.
[69] O. Vermesan, P. Friess, and P. Guillemin. (2009). Internet of things strategic research roadmap. The Cluster of European Research Projects [Online]. Available: http://www.internet-of-things-research.eu/pdf/ IoT_Cluster_Strategic_Research_Agenda_2009.pdf, accessed on Oct. 1, 2013.
[70] H. Sundmaeker, P. Guillemin, and P. Friess, Vision and Challenges for Realizing the Internet of Things. Brussels, Belgium: European Commission, 2010.
[71] K. Voigt. (2012). China Looks to Lead the Internet of Things [Online]. Available: http://www.cnn.com/2012/11/28/business/china-internet-of-things/, accessed on Oct. 1, 2013.
[72] H. Zhang and L. Zhu, “Internet of things: Key technology, architecture and challenging problems,” in Proc. 2011 IEEE Int. Conf. Comput. Sci. Autom. Eng. (CSAE), Shanghai, China, Jun. 10–12, pp. 507–512.
[73] S. Wang, L. Li, K. Wang, and J. Jones, “E-business system integration: A systems perspective,” Inf. Technol. Manage., vol. 13, no. 4, pp. 233–249, 2012.
[74] F. Tao, H. Guo, L. Zhang, and Y. Cheng, “Modelling of combinable relationship-based composition service network and the theoretical proof of its scale-free characteristics,” Enterp. Inf. Syst., vol. 6, no. 4, pp. 373–404, 2012.

[75] L. Xu, W. Viriyasitavat, P. Ruchikachorn, and A. Martin, “Using propositional logic for requirements verification of service workflow,” IEEE Trans. Ind. Informat., vol. 8, no. 3, pp. 639–646, Aug. 2012.
[76] D. Paulraj, S. Swamynathan, and M. Madhaiyan, “Process model-based atomic service discovery and composition of composite semantic web services using web ontology language for services,” Enterp. Inf. Syst., vol. 6, no. 4, pp. 445–471, 2012.
[77] H. Panetto and J. Cecil, “Information systems for enterprise integration, interoperability and networking: Theory and applications,” Enterp. Inf. Syst., vol. 7, no. 1, pp. 1–6, 2013.
[78] W. Viriyasitavat, L. Xu, and A. Martin, “SWSpec, service workflow requirements specification language: The formal requirements specification in service workflow environments,” IEEE Trans. Ind. Informat., vol. 8, no. 3, pp. 631–638, Aug. 2012.
[79] S. Hachani, L. Gzara, and H. Verjus, “A service-oriented approach for flexible process support within enterprises: An application on PLM systems,” Enterp. Inf. Syst., vol. 7, no. 1, pp. 79–99, 2013.
[80] L. Xu, “Enterprise Systems: State-of-the-art and future trends,” IEEE Trans. Ind. Informat., vol. 7, no. 4, pp. 630–640, Nov. 2011.
[81] M. C. Domingo, “An overview of the internet of things for people with disabilities,” J. Netw. Comput. Appl., vol. 35, no. 2, pp. 584–596, 2012.
[82] C. H. Liu, B. Yang, and T. Liu, “Efficient naming, addressing and profile services in Internet-of-Things sensory environments,” Ad Hoc Netw., to be published.
[83] Y. Wu, Q. Z. Sheng, and S. Zeadally, “RFID: Opportunities and challenges,” in Next-Generation Wireless Technologies, N. Chilamkurti, Ed. New York, NY, USA: Springer, 2013, ch. 7, pp. 105–129.
[84] E. Ilie-Zudor, Z. Kemeny, F. van Blommestein, L. Monostori, and A. Van der Meulen, “A survey of applications and requirements of unique identification systems and RFID techniques,” Comput. Ind., vol. 62, no. 3, pp. 227–252, 2011.
[85] C. Han, J. M. Jornet, E. Fadel, and I. F. Akyildiz, “A cross-layer communication module for the internet of things,” Comput. Netw., vol. 57, no. 3, pp. 622–633, 2013.
[86] D. Guinard, V. Trifa, S. Karnouskos, P. Spiess, and D. Savio, “Interacting with the soa-based internet of things: Discovery, query, selection, and on-demand provisioning of web services,” IEEE Trans. Serv. Comput., vol. 3, no. 3, pp. 223–235, Jul./Sep. 2010.
[87] K. Gama, L. Touseau, and D. Donsez, “Combining heterogeneous service technologies for building an internet of things middleware,” Comput. Commun., vol. 35, no. 4, pp. 405–417, 2012.
[88] D. Romero, G. Hermosillo, A. Taherkordi, R. Nzekwa, R. Rouvoy, and F. Eliassen, “RESTful integration of heterogeneous devices in pervasive environments,” in Distributed Applications and Interoperable Systems. Berlin, Germany: Springer-Verlag, 2010, ch. 01, pp. 1–4.
[89] H. Zhou, The Internet of Things in the Cloud: A Middleware Perspective. Boca Raton, FL, USA: CRC Press, 2012.
[90] L. Atzori, A. Iera, G. Morabito, and M. Nitti, “The social internet of things (SIoT)-when social networks meet the internet of things: Concept, architecture and network characterization,” Comput. Netw., vol. 56, no. 16, pp. 3594–3608, 2012.
[91] M. K. Lim, W. Bahr, and S. Leung, “RFID in the warehouse: A literature analysis (1995–2010) of its applications, benefits, challenges and future trends,” Int. J. Prod. Econ., vol. 145, no. 1, pp. 409–430, 2013.
[92] Q. Zhu, R. Wang, Q. Chen, Y. Liu, and W. Qin, “IoT gateway: Bridging wireless sensor networks into internet of things,” in Proc. IEEE/IFIP 8th Int. Conf. Embedded Ubiquitous Comput. (EUC), Hong Kong, China, Dec. 11–13, 2010, pp. 347–352.

[93] Y. Liu and G. Zhou, “Key technologies and applications of internet of things,” in Proc. 2012, 5th Int. Conf. Intell. Comput. Technol. Autom. (ICICTA), Zhangjiajie, China, pp. 187 200.
[94] H. Cervantes and R. S. Hall, “Automating service dependency management in a service-oriented component model,” in Proc. 6th Workshop Compon.- Based Softw. Eng., Portland, Oregon, USA, May 2003, pp. 1–5.
[95] J. I. Vazquez, A. Almeida, I. Doamo, X. Laiseca, and P. Orduña, “Flexeo: An architecture for integrating wireless sensor networks into the internet of things,” in Proc. 2008, 3rd Symp. Ubiquitous Comput. Ambient Intell., Salamanca, Spain, 2009, pp. 219–228.
[96] C. Flügel and V. Gehrmann, “Scientific workshop 4: Intelligent objects for the internet of things: Internet of things-application of sensor networks in logistics,” Commun. Comput. Inf. Sci., vol. 32, pp. 16–26, 2009.
[97] Z. Pang, Q. Chen, J. Tian, L. Zheng, and E. Dubrova, “Ecosystem analysis in the design of open platform-based in-home healthcare terminals towards the internet-of-things,” in Proc. 2013, 15th Int. Conf. Adv. Commun. Technol. (ICACT), Pyeongchang, Korea, pp. 529–534.
[98] H. Alemdar and C. Ersoy, “Wireless sensor networks for healthcare: A survey,” Comput. Netw., vol. 54, no. 15, pp. 2688–2710, 2010.
[99] I. Plaza, L. Martín, S. Martin, and C. Medrano, “Mobile applications in an aging society: Status and trends,” J. Syst. Softw., vol. 84, no. 11, pp. 1977–1988, 2011.
[100] Z. Pang, Q. Chen, W. Han, and L. Zheng, “Value-centric design of the internet-of-things solution for food supply chain: Value creation, sensor portfolio and information fusion,” Inf. Syst. Front., to be published.
[101] Q. Wei, S. Zhu, and C. Du, “Study on key technologies of internet of things perceiving mine,” Procedia Eng., vol. 26, pp. 2326–2333, 2011.
[102] B. Karakostas, “A DNS architecture for the internet of things: A case study in transport logistics,” Procedia Comput. Sci., vol. 19, pp. 594–601, 2013.
[103] H. Zhou, B. Liu, and D. Wang, “Design and research of urban intelligent transportation system based on the internet of things,” Commun. Comput. Inf. Sci., vol. 312, pp. 572–580, 2012.
[104] E. Qin, Y. Long, C. Zhang, and L. Huang, “Cloud computing and the internet of things: Technology innovation in automobile service,” LNCS 8017, New York, NY, USA, 2013, pp. 173–180.
[105] Y. Zhang, B. Chen, and X. Lu, “Intelligent monitoring system on refrigera- tor trucks based on the internet of things,” Wireless Commun. Appl., vol. 72, pp. 201–206, 2012.
[106] C. G. Keller, T. Dang, H. Fritz, A. Joos, C. Rabe, and D. M. Gavrila, “Active pedestrian safety by automatic braking and evasive steering,” IEEE Trans. Intell. Transp. Syst., vol. 12, no. 4, pp. 1292–1304, Dec. 2011.
[107] Y. C. Zhang and J. Yu, “A study on the fire IOT development strategy,” Procedia Eng., vol. 52, pp. 314–319, 2013.
[108] Z. Ji and A. Qi, “The application of internet of things (IOT) in emergency management system in China,” in Proc. 2010 IEEE Int. Conf. Technol. Homeland Security (HST), pp. 139–142.
[109] S. Wang, Z. Zhang, Z. Ye, X. Wang, X. Lin, and S. Chen, “Application of environmental internet of things on water quality management of urban scenic river,” Int. J. Sustain. Develop. World Ecol., vol. 20, no. 3, pp. 216–222, 2013.
[110] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, “Context aware computing for the internet of things: A survey,” IEEE Commun. Surveys Tuts., to be published.

[111] F. Wang, B. Ge, L. Zhang, Y. Chen, Y. Xin, and X. Li, “A system framework of security management in enterprise systems,” Syst. Res. Behav. Sci., vol. 30, no. 3, pp. 287–299, 2013.
[112] J. Li, J. Yang, Y. Zhao, and B. Liu, “A top-down approach for approxi- mate data anonymisation,” Enterp. Inf. Syst., vol. 7, no. 3, pp. 272–302, 2013.
[113] Y. Xing, L. Li, Z. Bi, M. Wilamowska-Korsak, and L. Zhang, “Operations research (OR) in service industries: A comprehensive review,” Syst. Res. Behav. Sci., vol. 30, no. 3, pp. 300–353, 2013.
[114] J. Wan and J. Jones, “Managing IT service management implementation complexity from the perspective of the Warfield version of systems science,” Enterp. Inf. Syst., vol. 7, no. 4, pp. 490–522, 2013.
[115] R. Roman, P. Najera, and J. Lopez, “Securing the internet of things,” Computer, vol. 44, no. 9, pp. 51–58, 2011.
[116] L. Li, “Technology designed to combat fakes in the global supply chain,” Bus. Horizons, vol. 56, no. 2, pp. 167–177, 2013.
[117] S. L. Ting and W. H. Ip, “Combating the counterfeits with web portal technology,” Enterp. Inf. Syst., to be published.
[118] J. Clarke, R. Castro, A. Sharma, J. Lopez, and N. Suri, “Trust & security RTD in the internet of things: Opportunities for international cooperation,” in Proc. 1st Int. Conf. Security of Internet of Things, Kollam, India, 2012, pp. 172–178.
[119] L. Xu, “Introduction: Systems science in industrial sectors,” Syst. Res. Behav. Sci., vol. 30, no. 3, pp. 211–213, 2013.
[120] F. Li, C. Jin, Y. Jing, M. Wilamowska-Korsak, and Z. Bi, “A rough programming model based on the greatest compatible classes and synthesis effect,” Syst. Res. Behav. Sci., vol. 30, no. 3, pp. 229–243, 2013.
[121] Y. Lin, X. Duan, C. Zhao, and L. Xu, Systems Science Methodological Approaches. Boca Raton, FL, USA: CRC Press, 2013
[122] L. Atzori, D. Carboni, and A. Iera, “Smart things in the social loop: Paradigms, technologies, and potentials,” Ad Hoc Netw., to be published.
[123] L. Xu, “Information architecture for supply chain quality management,” in Int. J. Prod. Res., vol. 49, no. 1, pp. 183–198, 2011.
[124] J. Z. Sun, “Towards the web of things: Open research issues and the BAS-AMI use case,” Lect. Notes Electr. Eng., vol. 144, pp. 1–8, 2012.
[125] D. Guinard, V. Trifa, F. Mattern, and E. Wilde, “From the internet of things to the web of things: Resource-oriented architecture and best practices,” in Architecting the Internet of Things. New York, NY, USA: Springer, 2011, pp 97–129.
[126] F. Xia, “Wireless sensor technologies and applications,” Sensors, vol. 9, no. 11, pp. 8824–8830, 2009.
[127] E. Yaacoub, A. Kadri, and A. Abu-Dayya, “Cooperative wireless sensor networks for green internet of things,” in Proc. 8th ACM Symp. QoS Security Wireless Mobile Netw., Paphos, Cyprus, 2012, pp. 79–80.
[128] A. Arsénio, H. Serra, R. Francisco, F. Nabais, J. Andrade, and E. Serrano, “Internet of Intelligent Things: Bringing artificial intelligence into things and communication networks,” Stud. Comput. Intell., vol. 495, pp. 1–37, 2014.
[129] J. O. Kephart and D. M. Chess, “The vision of autonomic computing,” IEEE Computer, vol. 36, no. 1, pp. 41–50, Jan. 2003.
[130] G. Kortuem, F. Kawsar, D. Fitton, and V. Sundramoorthy, “Smart objects as building blocks for the internet of things,” IEEE Internet Comput., vol. 14, no. 1, pp. 44–51, Jan./Feb. 2010.

[131] Y. Ding, Y. Jin, L. Ren, and K. Hao, “An intelligent self-organization scheme for the internet of things,” IEEE Comput. Intell. Mag., vol. 8, no. 3, pp. 41–53, Aug. 2013.
[132] B. P. Rao, P. Saluia, N. Sharma, A. Mittal, and S. V. Sharma, “Cloud computing for internet of things & sensing based applications,” in Proc. 2012 6th Int. Conf. Sens. Technol. (ICST), Kolkata, West Bangal, India, pp. 374–380.
[133] S. Fang, L. Xu, H. Pei, and Y. Liu, “An integrated approach to snowmelt flood forecasting in water resource management,” IEEE Trans. Ind. Informat., vol. 10, no. 1, pp. 548–558, Feb. 2014.
[134] F.J. Batlles, M.A. Rubio, J. Tovar, F.J. Olmo, L. Alados-Arboledas: Empirical modeling of hourly direct irradiance by means of hourly global irradiance, 2000 Elsevier Science Ltd
[135] J.W.H. Betcke, E.A. Alsema: Evaluation of the performance of PV system designs in The Netherlands in 1997-2000, University of Utrecht
[136] Christopher P. Cameron, William E. Boyson, Daniel M. Riley, Sandia National Laboratories: Comparison of PV system performance-model predictions with measured PV system performance, 33rd IEEE PVSC, San Diego, CA 2008
[137] Thomas Carlsson: Experimental Setup for Full Scale Field Tests of CdTe and CIS Thin−film PV Modules, M.S. Thesis, Department of Engineering Physics and Mathematics, Helsinki University of Technology
[138] Anna J. Carr: A Detailed Performance Comparison of PV Modules of Different Technologies and The Implications for PV System Design Methods, July 2005
[139] Widalys De Soto: Improvement and Validation of a Model for Photovoltaic Array Performance, M.S. Thesis, Mechanical Engineering, University of Wisconsin-Madison, 2004
[140] W. De Soto, S.A. Klein and W.A. Beckman : Improvement and Validation of a Model for Photovoltaic Array Performance, 2006
[141] John A. Duffie and William A. Beckman: Solar Engineering of Thermal Processes, 2 nd edition, Wiley, New York, NY, 1991
[142] K. Emery, J. Burdick, Y. Caiyem, D. Dunlavy, H. Field, B. Kroposki, T. Moriarty, L. Ottoson, S. Rummel, T. Strand, and M.W. Wanlass National Renewable Energy Laboratory: Temperature dependence of photovoltaic cells, modules and systems
[143] DG Erbs, SA Klein and JA. Duffie: Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation. Solar Energy 1982;28:293–302
[144] A. Hunter Fanney, Brian P. Dougherty and Mark W. Davis: Evaluating Building Integrated Photovoltaic Performance Models, 29 th IEEE Photovoltaic Specialists Conference (PVSC) May 20-24 th , 2002
[145] K.H. Hussein, I. Muta, T. Hoshino and M. Osakada: Maximum photovoltaic power tracking: an algorithm for rapidly changing atmospheric conditions, IEEE 1995
[146] Bryan Fry: Simulation of Grid-Tied Building Integrated Photovoltaic Systems, M.S. Thesis, Mechanical Engineering, University of Wisconsin-Madison, 1998
[147] A. Goetzberger and V.U. Hoffmann: Photovoltaic Solar Energy Generation, Springer
[148] B.T. Griffith and P.G. Ellis: Photovoltaic and Solar Thermal Modeling with the EnergyPlus Calculation Engine, July 2004 NREL
[149] David L. King, Jay A. Kratochvil, and William E. Boyson, Sandia National Laboratories: Stabilization and Performance characteristics of commercial amorphous-silicon PV modules