ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ & ΔΙΟΙΚΗΣΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ ΜΕ ΘΕΜΑ:

«Ανάπτυξη και αξιολόγηση μοντέλων πρόβλεψης της κατανάλωσης φυσικού αερίου για οικιακούς και μικρούς βιομηχανικούς καταναλωτές»

ΝΙΚΟΛΑΟΣ ΔΟΥΛΦΗΣ
Α.Μ. 2001010052

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΜΑΤΣΑΤΣΙΝΗΣ ΝΙΚΟΛΑΟΣ

ΧΑΝΙΑ
ΣΕΠΤΕΜΒΡΙΟΣ 2008

Η παρούσα διπλωματική αφιερώνεται στην οικογένειά μου και κυρίως στους γονείς μου, Κώστα και Σίσσυ.

ΕΥΧΑΡΙΣΤΙΕΣ

Στο σημείο αυτό θα ήθελα να εκφράσω τις ευχαριστίες μου σε όσους συνέβαλαν στην προσπάθειά μου με οποιονδήποτε τρόπο.

Αρχικά, ευχαριστώ θερμά το Διευθυντή Τιμολόγησης και Καταμέτρησης, Φώτη Τσιτσιρίγκο και τον Κώστα Μενάγια, της Εταιρείας Παροχής Αερίου Αττικής Α.Ε., για τις πολύτιμες συμβουλές, το ενδιαφέρον τους και την καθοδήγησή που μου προσέφεραν για την επιτυχή ολοκλήρωση της διπλωματικής μου εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή Νικόλαο Ματσατσίνη για την καθοδήγηση και τις εποικοδομητικές παρατηρήσεις του.

Τέλος, ευχαριστώ τους φίλους μου για τις υπέροχες στιγμές που μου προσέφεραν όλα αυτά τα χρόνια και για τις αναμνήσεις που θα πάρω μαζί μου φεύγοντας από τα Χανιά.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ……...4
ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ……..5
1.1 ΑΝΤΙΚΕΙΜΕΝΟ ΜΕΛΕΤΗΣ……5
1.2 Η ΑΝΑΓΚΑΙΟΤΗΤΑ ΤΗΣ ΠΡΟΒΛΕΨΗΣ ΓΕΝΙΚΩΣ……6
1.3 ΓΕΝΙΚΑ ΓΙΑ ΤΟ ΦΥΣΙΚΟ ΑΕΡΙΟ……..9
1.3.1 ΓΕΝΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ……..…….....9
1.3.2 ΙΣΤΟΡΙΚΑΣΤΟΙΧΕΙΑ……...……9
1.3.3 ΧΡΗΣΕΙΣ - ΕΦΑΡΜΟΓΕΣ ΦΥΣΙΚΟΥ ΑΕΡΙΟΥ……10
1.3.4 ΟΦΕΛΗ ΑΠΟ ΤΗ ΧΡΗΣΗ ΦΥΣΙΚΟΥ ΑΕΡΙΟΥ……11
1.3.5 ΓΕΝΙΚΑ ΣΤΟΙΧΕΙΑ – Η ΑΓΟΡΑ ΦΥΣΙΚΟΥ ΑΕΡΙΟΥ ΣΤΗΝ ΕΛΛΑΔΑ……13
1.3.6 ΓΕΝΙΚΑ ΓΙΑ ΤΙΣ ΕΤΑΙΡΕΙΕΣ ΕΚΜΕΤΑΛΛΕΥΣΗΣ ΦΥΣΙΚΟΥ ΑΕΡΙΟΥ ΣΤΗΝ ΕΛΛΑΔΑ…….15
1.3.7 ΤΟ ΣΥΣΤΗΜΑ ΜΕΤΑΦΟΡΑΣ ΦΥΣΙΚΟΥ ΑΕΡΙΟΥ ΣΤΗΝ ΕΛΛΑΔΑ……17
1.3.8 ΤΟ ΣΥΣΤΗΜΑ ΔΙΑΝΟΜΗΣ ΦΥΣΙΚΟΥ ΑΕΡΙΟΥ ΣΤΗΝ ΕΛΛΑΔΑ……...18

ΚΕΦΑΛΑΙΟ 2: ΥΦΙΣΤΑΜΕΝΗ ΚΑΤΑΣΤΑΣΗ ΚΑΙ ΠΡΟΒΛΗΜΑΤΑ……...19
2.1 ΓΕΝΙΚΑ……...19
2.2 ΥΦΙΣΤΑΜΕΝΗ ΚΑΤΑΣΤΑΣΗ……..19
2.3 Η ΑΝΑΓΚΑΙΟΤΗΤΑ ΤΗΣ ΠΡΟΒΛΕΨΗΣ ΕΙΔΙΚΑ ΓΙΑ ΜΙΑ ΕΤΑΙΡΕΙΑ ΔΙΑΝΟΜΗΣ……....20
2.4 Ο ΥΦΙΣΤΑΜΕΝΟΣ ΤΡΟΠΟΣ ΠΡΟΒΛΕΨΗΣ ΤΗΣ ΚΑΤΑΝΑΛΩΣΗΣ……...22

ΚΕΦΑΛΑΙΟ 3: ΚΑΘΟΡΙΣΜΟΣ ΚΑΙ ΕΠΙΛΟΓΗ ΚΡΙΤΗΡΙΩΝ……23
3.1 ΑΝΑΛΥΣΗ ΠΑΡΑΓΟΝΤΩΝ – ΚΡΙΤΗΡΙΩΝ……...23
3.2 ΠΑΡΑΓΟΝΤΕΣ ΠΟΥ ΔΙΑΜΟΡΦΩΝΟΥΝ ΤΗ ΘΕΡΜΟΚΡΑΣΙΑ ΣΤΟ ΕΣΩΤΕΡΙΚΟ ΤΗΣ ΟΙΚΙΑΣ…….24
3.2.1 ΕΞΑΡΤΗΣΗ ΑΠΟ ΤΙΣ ΚΛΙΜΑΤΟΛΟΓΙΚΕΣ ΣΥΝΘΗΚΕΣ……..24
3.2.2 ΕΞΑΡΤΗΣΗ ΑΠΟ ΤΙΣ ΕΝΕΡΓΕΙΑΚΕΣ ΑΠΩΛΕΙΕΣ ΤΟΥ ΚΤΙΡΙΟΥ……...25
3.3 ΚΡΙΤΗΡΙΑ ΠΟΥ ΑΦΟΡΟΥΝ ΤΗ ΘΕΡΜΟΚΡΑΣΙΑ ΣΤΟ ΕΣΩΤΕΡΙΚΟ ΤΗΣ ΟΙΚΙΑΣ……..26
3.4 ΠΑΡΑΓΟΝΤΕΣ ΚΑΙ ΚΡΙΤΗΡΙΑ ΠΟΥ ΑΦΟΡΟΥΝ ΤΗΝ ΤΑΣΗ ΚΑΤΑΝΑΛΩΣΗΣ……..28
3.5 ΜΗ ΕΠΙΛΕΧΘΕΝΤΑ ΚΡΙΤΗΡΙΑ……..29

ΚΕΦΑΛΑΙΟ 4: ΕΠΙΛΟΓΗ ΚΑΙ ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ……31
4.1 ΕΙΣΑΓΩΓΗ……..31
4.2 ΔΕΔΟΜΕΝΑ ΠΟΥ Η ΣΥΓΚΕΝΤΡΩΣΗ ΤΟΥΣ ΕΙΝΑΙ ΑΝΕΦΙΚΤΗ……...31
4.3 ΔΕΔΟΜΕΝΑ ΠΟΥ ΑΠΟΚΤΗΘΗΚΑΝ ΑΠΟ ΜΕΤΕΩΡΟΛΟΓΙΚΟΥΣ ΣΤΑΘΜΟΥΣ……32
4.4 ΔΕΔΟΜΕΝΑ ΠΟΥ ΑΠΟΚΤΗΘΗΚΑΝ ΑΠΟ ΤΗΝ ΕΤΑΙΡΕΙΑ ΔΙΑΝΟΜΗΣ ΦΥΣΙΚΟΥ ΑΕΡΙΟΥ……..41

ΚΕΦΑΛΑΙΟ 5: ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ……...48
5.1 ΕΙΣΑΓΩΓΗ……48
5.2 Η ΔΙΑΔΙΚΑΣΙΑ ΕΠΙΛΥΣΗΣ ΕΝΟΣ ΠΡΟΒΛΗΜΑΤΟΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ……...52
5.3 Η ΠΡΟΕΤΟΙΜΑΣΙΑ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΣΤΗΝ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ……..52
5.4 ΕΠΙΛΟΓΗ ΤΟΥ ΚΑΤΑΛΛΗΛΟΥ ΑΛΓΟΡΙΘΜΟΥ ΜΑΘΗΣΗΣ……... 53
5.5 ΕΚΠΑΙΔΕΥΣΗ ΜΕΣΩ ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ……..….55
5.6 ΤΟ ΛΟΓΙΣΜΙΚΟ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΚΑΙ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ……...56
5.7 ΜΕΤΡΑ ΑΞΙΟΛΟΓΗΣΗΣ ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ ΤΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ……57
5.7.1. ΣΦΑΛΜΑ ΠΡΟΒΛΕΨΗΣ……...57
5.7.2 ΜΕΣΗ ΑΠΟΛΥΤΗ ΑΠΟΚΛΙΣΗ (MEAN ABSOLUTE DEVIATION –MAD)……57
5.7.3 ΜΕΣΟ ΑΠΟΛΥΤΟ ΕΚΑΤΟΣΤΙΑΙΟ ΣΦΑΛΜΑ (MEAN ABSOLUTE PERCENTAGE ERROR)..57
5.7.4 ΜΕΣΟ ΤΕΤΡΑΓΩΝΙΣΜΕΝΟ ΣΦΑΛΜΑ (MEAN SQUARED ERROR –MSE)……58
5.7.5 ΜΕΣΟ ΤΕΤΡΑΓΩΝΙΚΟ ΣΦΑΛΜΑ (ROOT MEAN SQUARED ERROR –RMSE)……58
5.7.6 ΤΟ ΤΥΠΙΚΟ ΣΦΑΛΜΑ ΕΝΟΣ ΔΙΑΣΤΗΜΑΤΟΣ ΕΜΠΙΣΤΟΣΥΝΗΣ (STANDARD ERROR)…..58
5.7.7 ΠΙΝΑΚΑΣ ΣΥΓΧΥΣΗΣ ΚΑΙ ΠΑΡΑΓΩΓΑ ΜΕΤΡΑ ΕΚΤΙΜΗΣΗΣ ΣΦΑΛΜΑΤΟΣ……59
5.7.7.1 ΤΟ ΜΕΤΡΟ ΤΩΝ ΠΡΑΓΜΑΤΙΚΑ ΘΕΤΙΚΩΝ ΠΑΡΑΔΕΙΓΜΑΤΩΝ……...59
5.7.7.2 ΤΟ ΜΕΤΡΟ ΤΩΝ ΕΣΦΑΛΜΕΝΑ ΘΕΤΙΚΩΝ ΠΑΡΑΔΕΙΓΜΑΤΩΝ……59
5.7.7.3 ΤΟ ΜΕΤΡΟ ΤΗΣ ΕΠΑΝΑΚΛΗΣΗΣ……..60
5.7.7.4 ΤΟ ΜΕΤΡΟ F……..60

ΚΕΦΑΛΑΙΟ 6: ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΟ ΛΟΓΙΣΜΙΚΟ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ WEKA……...61
6.1 ΕΙΣΑΓΩΓΗ ΣΤΟ ΛΟΓΙΣΜΙΚΟ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ WEKA……..61
6.2 ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ……63
6.3 ΔΗΜΙΟΥΡΓΙΑ ΤΟΥ ΣΩΜΑΤΟΣ ΕΚΠΑΙΔΕΥΣΗΣ ΚΑΙ ΑΞΙΟΛΟΓΗΣΗΣ……64
6.4 ΑΞΙΟΛΟΓΗΣΗ ΚΑΙ ΕΠΙΛΟΓΗ ΤΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ……65
6.5 ΕΠΙΛΟΓΗ ΑΛΓΟΡΙΘΜΩΝ……...67
6.6 ΠΑΡΟΥΣΙΑΣΗ ΕΠΙΛΕΧΘΕΝΤΩΝ ΑΛΓΟΡΙΘΜΩΝ WEKA……..68
6.6.1 ΑΛΓΟΡΙΘΜΟΙ ΤΩΝ Κ ΠΙΟ ΚΟΝΤΙΝΩΝ ΓΕΙΤΟΝΩΝ (ΑΛΓΟΡΙΘΜΟΙ IBK, KSTAR)……68
6.6.2 ΠΑΛΙΝΔΡΟΜΗΣΗ (ΑΛΓΟΡΙΘΜΟΙ LINEAR REGRESSION, SMOREG, M5RULES)……70
6.6.3 ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ (ΑΛΓΟΡΙΘΜΟΙ REPTREE, M5P, DECISION TABLE)……71
6.6.4 ΤΕΧΝΗΤΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ (ΑΛΓΟΡΙΘΜΟΙ MULTILAYER PERCEPTOR)……...74
6.6.4.1 ΤΟΠΟΛΟΓΙΑ PERCEPTRON……75
6.6.4.2 ΤΟΠΟΛΟΓΙΑ MULTI-LAYER PERCEPTRON……...76
6.6.5 ΑΛΓΟΡΙΘΜΟΙ ΜΕΤΑ (ΑΛΓΟΡΙΘΜΟΙ BAGGING, REGRESSION BY DISCRETIZATION)…..77
6.6.5.1 Ο ¨ΜΕΤΑ¨ ΑΛΓΟΡΙΘΜΟΣ ΕΜΦΩΛΙΑΣΗ (BAGGING)……77
6.7 ΑΞΙΟΛΟΓΗΣΗ ΚΑΙ ΕΠΙΛΟΓΗ ΚΑΤΑΛΛΗΛΟΤΕΡΩΝ ΑΛΓΟΡΙΘΜΩΝ……78

ΚΕΦΑΛΑΙΟ 7: ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΛΓΟΡΙΘΜΩΝ ΓΙΑ ΔΙΑΦΟΡΕΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ…….87
7.1 ΕΙΣΑΓΩΓΗ……..87
7.2 ΕΦΑΡΜΟΓΗ ΑΛΓΟΡΙΘΜΩΝ ΣΤΟ ΣΥΝΟΛΟ ΤΩΝ ΔΕΔΟΜΕΝΩΝ……..87
7.3 ΕΦΑΡΜΟΓΗ ΑΛΓΟΡΙΘΜΩΝ ΑΝΑ ΚΑΤΗΓΟΡΙΑ ΧΡΗΣΗΣ……..94
7.3.1 ΟΙΚΙΑΚΟ - ΑΤΟΜΙΚΗ ΘΕΡΜΑΝΣΗ (IN1)……...95
7.3.2 ΟΙΚΙΑΚΟ - ΜΑΓΕΙΡΕΜΑ, ΖΕΣΤΟ ΝΕΡΟ, ΘΕΡΜΑΝΣΗ (IN2)……...95
7.3.3 ΟΙΚΙΑΚΟ - ΚΕΝΤΡΙΚΗ ΘΕΡΜΑΝΣΗ (IN3)……..96
7.3.4 ΕΜΠΟΡΙΚΟ - ΘΕΡΜΑΝΣΗ (<5500NM3/Y) (IN4)……96
7.3.5 ΕΜΠΟΡΙΚΟ - ΘΕΡΜΑΝΣΗ (>5500NM3/Y) (IN5)……....97
7.3.6 ΕΜΠΟΡΙΚΟ - ΛΟΙΠΕΣ ΧΡΗΣΕΙΣ & ΘΕΡΜΑΝΣΗ (IN6)……97
7.3.7 ΑΝΑΛΥΣΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ…….....98
7.4 ΕΦΑΡΜΟΓΗ ΑΛΓΟΡΙΘΜΩΝ ΑΝΑ ΚΑΤΗΓΟΡΙΑ ΚΑΤΑΝΑΛΩΤΩΝ……100
7.4.1 ΑΘΛΗΤΙΚΕΣ ΕΓΚΑΤΑΣΤΑΣΕΙΣ (AD01)……101
7.4.2 ΆΛΛΕΣ ΥΠΗΡΕΣΙΕΣ (AD02)……101
7.4.3 ΑΝΑΨΥΚΤΗΡΙΟ (AD03)……...……102
7.4.4 ΑΡΤΟΠΟΙΕΙΟ - ΖΑΧΑΡΟΠΛΑΣΤΕΙΟ (AD04)…….....102
7.4.5 ΑΝΑΨΥΚΤΗΡΙΟ (AD05)…….....103
7.4.6 ΒΙΟΜΗΧΑΝΙΑ (AD06)……..103
7.4.7 ΒΙΟΤΕΧΝΙΑ (AD07)……..104
7.4.8 ΓΡΑΦΕΙΟ – ΕΤΑΙΡΕΙΑ (AD08)…….....104
7.4.9 ΔΗΜΟΣΙΟ ΚΤΗΡΙΟ (AD09)…….....105
7.4.10 ΕΜΠΟΡΙΚΟ ΚΑΤΑΣΤΗΜΑ (AD10) ...……..105
7.4.11 ΕΡΓΑΣΤΗΡΙΟ (AD11)……106
7.4.12 ΕΣΤΙΑΤΟΡΙΟ - ΜΕΖΕΔΟΠΩΛΕΙΟ (AD12)……...106
7.4.13 ΘΕΑΤΡΟ - ΣΙΝΕΜΑ (AD13)……...107
7.4.14 ΙΑΤΡΕΙΟ (AD14)……..107
7.4.15 ΙΔΙΩΤΗΣ (AD15)…….....108
7.4.16 ΙΕΡΟΣ ΝΑΟΣ (AD16)……...108
7.4.17 ΚΑΘΑΡΙΣΤΗΡΙΑ (AD17)…….....109
7.4.18 ΚΑΦΕ - ΜΠΑΡ (AD18)……109
7.4.19 ΚΕΝΤΡΙΚΗ ΘΕΡΜΑΝΣΗ (AD19)……...110
7.4.20 ΝΟΣΟΚΟΜΕΙΟ - ΚΛΙΝΙΚΗ (AD20)……..110
7.4.21 ΞΕΝΟΔΟΧΕΙΟ (AD21)……111
7.4.22 ΟΙΚΟΣ ΕΥΓΗΡΙΑΣ - ΟΡΦΑΝΟΤΡΟΦΕΙΟ (AD22)……111
7.4.23 ΠΑΝΕΠΙΣΤΗΜΙΟ (AD23)……...112
7.4.24 ΠΟΛΥΚΑΤΟΙΚΙΑ (AD24)……...112
7.4.25 ΠΡΕΣΒΕΙΑ (AD25)……...113
7.4.26 ΣΧΟΛΕΙΟ (AD26)…….....113
7.4.27 ΣΧΟΛΗ - ΕΚΠΑΙΔΕΥΤΗΡΙΟ (AD27)……....114
7.4.28 ΤΡΑΠΕΖΑ (AD28)……...114
7.4.29 ΑΝΑΛΥΣΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ……115

ΣΥΜΠΕΡΑΣΜΑΤΑ…….....118

ΒΙΒΛΙΟΓΡΑΦΙΑ……121

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία ασχολείται με την ανάπτυξη και αξιολόγηση μοντέλων πρόβλεψης της αναμενόμενης κατανάλωσης φυσικού αερίου για οικιακούς και μικρούς επαγγελματικούς καταναλωτές για την περιοχή της Αττικής.

Στο πρώτο κεφάλαιο παρουσιάζονται οι στόχοι της μελέτης, η αναγκαιότητα ύπαρξης μιας τέτοιας έρευνας. Επίσης, γίνεται μια παρουσίαση των χαρακτηριστικών του φυσικού αερίου και της αγοράς του στην Ελλάδα. Στο δεύτερο κεφάλαιο παρουσιάζονται οι λόγοι που καθιστούν αναγκαία, την ύπαρξη εργαλείων πρόβλεψης της κατανάλωσης και αναλύονται οι υφιστάμενες μέθοδοι και τα προβλήματα που προκύπτουν. Στο τρίτο κεφάλαιο επιχειρείται μια ανάλυση των παραγόντων που επηρεάζουν την κατανάλωση και παρουσιάζονται όλα τα επιλεγμένα κριτήρια.

Το τέταρτο κεφάλαιο περιλαμβάνει τη διαδικασία συλλογής των απαιτούμενων δεδομένων και παρουσιάζεται η ανάλυσή τους. Στο πέμπτο κεφάλαιο παρουσιάζεται η τεχνική της μηχανικής μάθησης και γίνεται ανάλυση της μεθοδολογίας αυτής. Στο έκτο κεφάλαιο παρουσιάζεται το λογισμικό μηχανικής μάθησης που χρησιμοποιήθηκε, περιγράφεται η διαδικασία προετοιμασίας των δεδομένων και επιλέγονται οι καταλληλότεροι αλγόριθμοι. Ακόμα, παρουσιάζεται, βήμα βήμα, όλη η διαδικασία που ακολουθήθηκε κατά την επεξεργασία των δεδομένων.

Στο έβδομο κεφάλαιο παρουσιάζονται τα αποτελέσματα της επεξεργασίας των δεδομένων που προέκυψαν και γίνεται ανάλυση των αποτελεσμάτων. Τέλος, παρουσιάζονται όλα τα συμπεράσματα που προέκυψαν από την αξιολόγηση των αλγορίθμων και την ανάλυση των αποτελεσμάτων της πρόβλεψης της κατανάλωσης φυσικού αερίου.

ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ

1.1 ΑΝΤΙΚΕΙΜΕΝΟ ΜΕΛΕΤΗΣ

Σκοπός της παρούσης διπλωματικής εργασίας είναι η ανάπτυξη και αξιολόγηση μοντέλων πρόβλεψης της αναμενόμενης κατανάλωσης φυσικού αερίου για την περιοχή της Αττικής. Συγκεκριμένα μελετάται το πρόβλημα της πρόβλεψης της κατανάλωσης φυσικού αερίου για οικιακούς και μικρούς επαγγελματικούς καταναλωτές μιας εταιρείας διανομής. Η ανάπτυξη των μοντέλων αιτιολογείται από την αναγκαιότητα ύπαρξης εργαλείων, τα οποία θα είναι σε θέση να υποστηρίζουν τις αποφάσεις της επιχείρησης, με επιστημονικό τρόπο.

Η διαδικασία σχεδιασμού των προτεινόμενων μοντέλων προϋποθέτει την ανάπτυξη και μελέτη μιας μεθοδολογίας πρόβλεψης, γεγονός που είναι ιδιαίτερα σημαντικό, δεδομένης της πληθώρας διαφορετικής φιλοσοφίας μαθηματικών μοντέλων και εργαλείων πρόβλεψης. Η χρήση μοντέλων πρόβλεψης πολλαπλών μεταβλητών κρίνεται απαραίτητη λόγω της επιρροής της κατανάλωσης φυσικού αερίου από ένα σύνολο ανεξάρτητων παραγόντων. Το θέμα είναι ιδανικό για την εφαρμογή και αξιοποίηση διάφορων μεθοδολογιών data mining όπως για παράδειγμα τα νευρωνικά δίκτυα, τα συστήματα ασαφούς λογικής αλλά και άλλες μεθοδολογίες, για τον υπολογισμό της αναμενόμενης κατανάλωσης φυσικού αερίου.

Βασική προϋπόθεση της ανάπτυξης και αξιολόγησης μοντέλων πρόβλεψης είναι η ανάλυση των δεδομένων και των παραγόντων που επηρεάζουν την κατανάλωση. Τα δεδομένα που χρησιμοποιούνται στην εργασία αφορούν την εταιρεία διανομής φυσικού αερίου Αττικής. Λόγω του θέματος της εργασίας προκύπτουν κάποιες ιδιαίτερες προκλήσεις για το σχεδιασμό του μοντέλου όπως η αστάθεια / μεταβλητότητα της κατανάλωσης, η ισχυρή εξάρτηση από τις καιρικές συνθήκες, η έλλειψη αξιόπιστων ιστορικών στοιχείων.

Στόχος είναι να αναπτυχθούν και να αξιολογηθούν μοντέλα πρόβλεψης που μέσω της διεξαγωγής χρήσιμων συμπερασμάτων, να επιτυγχάνεται μακροπρόθεσμα η καλύτερη εξυπηρέτηση των πελατών και η βελτιστοποίηση του συστήματος διανομής σε συνδυασμό με την ελαχιστοποίηση των δεσμευμένων κεφαλαίων και του κόστους. Είναι επομένως απαραίτητη η ανάπτυξη και αξιολόγηση αξιόπιστων μοντέλων για την πρόβλεψη της κατανάλωσης.

1.2 Η ΑΝΑΓΚΑΙΟΤΗΤΑ ΤΗΣ ΠΡΟΒΛΕΨΗΣ ΓΕΝΙΚΩΣ

Οι εταιρείες που σχετίζονται με το φυσικό αέριο χρησιμοποιούν μοντέλα πρόβλεψης της κατανάλωσης για διάφορους λόγους. Γενικά, μια πρόβλεψη της κατανάλωσης του φυσικού αερίου είναι σημαντική και εξυπηρετεί σε διάφορους τομείς τις εταιρείες διανομής και μεταφοράς του αλλά και τις αρμόδιες κρατικές υπηρεσίες.

Για το κράτος η πρόβλεψη είναι γενικά χρήσιμη :

α) Για τον προγραμματισμό των ενεργειακών επενδύσεων. Η πρόβλεψη κατανάλωσης είναι απαραίτητη για το κράτος καθώς αποτελεί προϋπόθεση για τον καλύτερο προγραμματισμό των ενεργειακών επενδύσεων που αφορούν το φυσικό αέριο. Εάν παραδείγματος χάρη, προβλέπεται αύξηση της κατανάλωσης για τα επόμενα χρόνια, τότε το κράτος θα πρέπει να εξετάσει το ενδεχόμενο αύξησης των επενδύσεων για το φυσικό αέριο. Οι επενδύσεις αυτές γίνονται με σκοπό την καλύτερη αξιοποίηση του φυσικού αερίου και αφορούν, για παράδειγμα, την κατασκευή εγκαταστάσεων αποθήκευσης φυσικού αερίου σε διάφορες απομακρυσμένες περιοχές της χώρας, την αγορά και εγκατάσταση εξοπλισμού και μέσων μεταφοράς σε νησιά αλλά και για άλλους λόγους.

β) Για την αποτίμηση της πορείας των δεσμεύσεων. Ως μέλος της Ευρωπαϊκής Ένωσης αλλά και πολλών άλλων διεθνών οργανισμών, η χώρα μας είναι δεσμευμένη να τηρεί συμφωνίες που αφορούν τη διάδοση της χρήσης των ανανεώσιμων πηγών ενέργειας. Η πρόβλεψη της κατανάλωσης θα βοηθήσει στην παρακολούθηση της πορείας της διάδοσης του φυσικού αερίου και των δεσμεύσεων στο πλαίσιο των διεθνών ενεργειακών και περιβαλλοντικών συμφωνιών. Μέσω της πρόβλεψης της κατανάλωσης μπορεί να γίνει μια εκτίμηση της πορείας και της επάρκειας των ενεργειών που υιοθετούνται για την τήρηση των δεσμεύσεων.

γ) Για τη μελέτη πιθανών αλλαγών στο θεσμικό πλαίσιο. Είναι πιθανό να απαιτούνται τυχόν τροποποιήσεις στο θεσμικό πλαίσιο που αφορά το ενεργειακό σύστημα της χώρας ανάλογα με την πορεία της κατανάλωσης του φυσικού αερίου και της διείσδυσής του στην εγχώρια αγορά. Τροποποιήσεις που ενδεχομένως να αφορούν τους νόμους και τις διατάξεις σχετικά με το φυσικό αέριο.

Για μια εταιρεία μεταφοράς η πρόβλεψη είναι γενικά χρήσιμη :

α) Για τη διαμόρφωση ή επέκταση του δικτύου μεταφοράς. Είναι σημαντικό να υπάρχει μια εκτίμηση της κατανάλωσης για να είναι δυνατόν να καθοριστεί μελλοντικά η τοπολογία και τα χαρακτηριστικά του δικτύου μεταφοράς εκ μέρους της εταιρείας μεταφοράς φυσικού αερίου. Με τη χρήση των αποτελεσμάτων της πρόβλεψης της κατανάλωσης φυσικού αερίου, η εταιρεία είναι σε θέση να ελέγχει την ικανότητα του δικτύου μεταφοράς για την κάλυψη της μελλοντικής ζήτησης και να προβλέψει τις πιθανές επεκτάσεις ή ενισχύσεις που χρειάζονται να πραγματοποιηθούν.

β) Για τη διασφάλιση της ενεργειακής τροφοδοσίας. Η εταιρεία μεταφοράς οφείλει να εγγυάται την αδιάλειπτη τροφοδοσία σε φυσικό αέριο στους πελάτες της. Η χώρα μας βρίσκεται πολύ μακριά από τα κέντρα παραγωγής φυσικού αερίου και προμηθεύεται φυσικό αέριο από ξένες χώρες όπως η Ρωσία και η Αλγερία. Η πρόβλεψη της κατανάλωσης χρησιμεύει για να διασφαλιστεί η απρόσκοπτη μελλοντική τροφοδοσία φυσικού αερίου καθώς η εταιρεία μεταφοράς φυσικού αερίου θα είναι σε θέση να προμηθευτεί και να διατηρεί επαρκή αποθέματα φυσικού αερίου ώστε να μπορεί να καλύψει κάθε μελλοντική ζήτηση.

γ) Για τη δέσμευση των απαραίτητων κεφαλαίων. Η πρόβλεψη της κατανάλωσης θα βοηθήσει την εταιρεία μεταφοράς να εκτιμήσει και να δεσμεύσει ή συγκεντρώσει τα απαραίτητα κεφαλαία που χρειάζονται για την προμήθεια των απαραίτητων ποσοτήτων φυσικού αερίου.

Για μια εταιρεία διανομής η πρόβλεψη είναι γενικά χρήσιμη :

α) Για την τιμολόγηση των πελατών. Η εταιρεία μεταφοράς χρειάζεται να προσδιορίσει τις τιμές που θα χρεώνει για την παροχή φυσικού αερίου ώστε να καλύπτεται το κόστος και να μένει κέρδος. Η πρόβλεψη της κατανάλωσης βοηθά στο να εκτιμηθούν οι προβλεπόμενες πωλήσεις ώστε σε συσχέτιση με το κόστος να καθοριστούν οι τιμές που χρεώνει η εταιρεία με σκοπό τη μεγιστοποίηση του μέγιστου κέρδους. Η πρόβλεψη της μηνιαίας κατανάλωσης είναι πολύ χρήσιμη για την τιμολόγηση των πελατών ιδιαίτερα για την εκτίμηση της κατανάλωσης όταν δεν είναι δυνατή η καταγραφή των ενδείξεων των μετρητών καθώς προκύπτουν προβλήματα και καθυστερήσεις στην καταγραφή των ενδείξεων. Επίσης η πρόβλεψη της μηνιαίας κατανάλωσης είναι πολύ χρήσιμη για τη σωστή μηνιαία κατανομή της κατανάλωσης εντός του διμηνιαίου λογαριασμού του πελάτη. Συγκεκριμένα, επειδή η τιμή του αερίου αλλάζει ανά μήνα και η καταγραφή των ενδείξεων των μετρητών και η τιμολόγηση γίνεται ανά δίμηνο, χρειάζεται να γνωρίζουμε ποια ποσότητα αντιστοιχεί σε κάθε μήνα, ώστε να γίνεται σωστά η τιμολόγηση.

β) Για το στρατηγικό και επιχειρησιακό σχεδιασμό. Ανάλογα με τις προβλεπόμενες πωλήσεις, κάθε εταιρεία αναπτύσσεται στην αγορά αναλόγως, καθορίζει τις μελλοντικές στρατηγικές που πρόκειται να υιοθετήσει και οργανώνει και εκσυγχρονίζει τις διαδικασίες που ακολουθεί. Με την πρόβλεψη της κατανάλωσης, η εταιρεία μπορεί να προβεί σε ενέργειες με σκοπό τη βελτίωση της λειτουργίας της μέσω της αξιοποίησης νέων τεχνολογιών, της καλύτερης οργάνωσης, της αξιοποίησης του ανθρώπινου δυναμικού ώστε να καλύπτονται οι ανάγκες και να βελτιωθεί η αποτελεσματικότητά της.

γ) Για την κατασκευή ή επέκταση των δικτύων διανομής. Είναι σημαντικό να υπάρχει μια εκτίμηση της κατανάλωσης για να είναι δυνατόν να καθοριστεί μελλοντικά η τοπολογία και τα χαρακτηριστικά του δικτύου διανομής εκ μέρους της εταιρείας διανομής φυσικού αερίου. Με τη χρήση των αποτελεσμάτων της πρόβλεψης της κατανάλωσης φυσικού αερίου, η εταιρεία είναι σε θέση να ελέγχει την ικανότητα του δικτύου διανομής για την κάλυψη της μελλοντικής ζήτησης και να προβλέψει τις πιθανές επεκτάσεις ή ενισχύσεις που χρειάζονται να πραγματοποιηθούν.

δ) Για τη αξιολόγηση των παραγόντων που καθορίζουν την κατανάλωση. Δίνεται η ευκαιρία να αναγνωριστούν οι παράγοντες που επηρεάζουν πιο πολύ την κατανάλωση ώστε να μπορεί η εταιρεία να προβεί σε διορθωτικές ενέργειες. Επίσης, μειώνεται η αβεβαιότητα που προκαλεί η μεταβολή των εξωτερικών συνθηκών όπως π.χ. οι καιρικές συνθήκες, διάφοροι αστάθμητοι παράγοντες.

ε) Για τη διαχείριση κίνδυνων. Κάθε εταιρεία έχει την ανάγκη για κάλυψη από τους κινδύνους αγοράς μέσω της διαδικασίας του προσδιορισμού, της ανάλυσης και της ανάπτυξης στρατηγικών για αντιμετώπιση των κινδύνων. Η πρόβλεψη της κατανάλωσης βοηθά στον προσδιορισμό και την εκτίμηση πιθανών κινδύνων για την εταιρεία.

ζ) Για τη διεξαγωγή ενός επιτυχημένου μάρκετινγκ. Το μάρκετινγκ κατέχει υψηλή θέση στις προτεραιότητες κάθε εταιρείας. Ακολουθώντας μια σωστή στρατηγική μάρκετινγκ, η εταιρεία μπορεί να αυξήσει τα κέρδη της και να θωρακιστεί ενάντια στις πτωτικές τάσεις και τις μεταβολές της αγοράς. Με ένα αποτελεσματικό μάρκετινγκ το κοινό ενημερώνεται καλύτερα για τα προϊόντα, προσελκύονται με αυτό το τρόπο νέοι πελάτες και αυξάνεται η εμπιστοσύνη των ήδη υπαρχόντων πελατών. Η πρόβλεψη της κατανάλωσης είναι κλειδί για τη διεξαγωγή ενός επιτυχημένου μάρκετινγκ και βοηθά στον καλύτερο προσδιορισμό της αγοράς - στόχου και του προφίλ των πελατών.

η) Για την παρακολούθηση των συμβάσεων με τους προμηθευτές. Η πρόβλεψη είναι ένα εργαλείο απαραίτητο για τον προγραμματισμό της κατανάλωσης και την παρακολούθηση της σύμβασης με την εταιρεία μεταφοράς. Καθώς η εταιρεία διανομής αγοράζει αέριο από μια εταιρεία μεταφοράς, πρέπει να παρέχει στην εταιρεία μεταφοράς μια πρόβλεψη των μελλοντικών αναγκών της έτσι ώστε η εταιρεία μεταφοράς να δύναται να προγραμματίσει τη διάθεση του απαραίτητου αερίου. Εάν το λάθος της πρόβλεψης είναι μεγάλο, η εταιρεία μεταφοράς ενδέχεται να επιβάλλει οικονομικές ποινές εις βάρος της εταιρείας διανομής λόγω αθέτησης των ρητρών που αφορούν κυρίως την ετήσια κατανάλωση, αλλά και τις μέγιστες ωριαίες ποσότητες.

1.3 ΓΕΝΙΚΑ ΓΙΑ ΤΟ ΦΥΣΙΚΟ ΑΕΡΙΟ

1.3.1 Γενικά χαρακτηριστικά

Το Φυσικό Αέριο είναι ένα μίγμα αερίων υδρογονανθράκων, κυρίως μεθάνιο, δηλαδή ο ελαφρύτερος υδρογονάνθρακας, είναι πολύ καθαρό, χωρίς προσμίξεις και θειούχα συστατικά, και κάνει τέλεια καύση στις κατάλληλες συσκευές. Συναντάται στη φύση σε κοιτάσματα μόνο του ή συνυπάρχει με κοιτάσματα πετρελαίου και εξάγεται από τις υπόγειες κοιλότητες υπό υψηλή πίεση και μεταφέρεται προς τους τόπους όπου πρόκειται να χρησιμοποιηθεί όπως είναι, χωρίς την ανάγκη περαιτέρω επεξεργασίας. Τα μεγαλύτερα γνωστά αποθέματα φυσικού αερίου βρίσκονται στις χώρες της πρώην Σοβιετικής Ένωσης (Ρωσία, χώρες της Κασπίας), στη Μέση Ανατολή, στις ΗΠΑ, στη Βενεζουέλα, στην Αλγερία και στη Νιγηρία.

Το φυσικό αέριο είναι άχρωμο και άοσμο. Η χαρακτηριστική του οσμή δίνεται τεχνητά ώστε να γίνεται αντιληπτό σε τυχόν διαρροές. Είναι η καθαρότερη πηγή πρωτογενούς ενέργειας, μετά τις ανανεώσιμες μορφές. Τα μεγέθη των εκπεμπόμενων ρύπων είναι σαφώς μικρότερα σε σχέση με τα συμβατικά καύσιμα, ενώ η βελτίωση του βαθμού απόδοσης μειώνει τη συνολική κατανάλωση καυσίμου και συνεπώς περιορίζει την ατμοσφαιρική ρύπανση.

1.3.2 Ιστορικά στοιχεία

Οι πρώτες προσπάθειες εκμετάλλευσης του Φυσικού Αερίου ως ενεργειακής πηγής χρονολογούνται από τις αρχές του 20ου αιώνα, όταν στις ΗΠΑ έγιναν οι πρώτες γεωτρήσεις και κατασκευάστηκαν οι πρώτοι υποτυπώδεις αγωγοί μεταφοράς. Όμως το παγκόσμιο ενεργειακό τοπίο μετεβλήθη άρδην λόγω της ενεργειακής κρίσης κατά τη διάρκεια της δεκαετίας του 70'. Η προσπάθεια για μείωση της εξάρτησης της οικονομίας από το πετρέλαιο είχε σαν αποτέλεσμα την ισχυροποίηση της θέσης του φυσικού αερίου. Έτσι, κατά την περίοδο 1970- 1996, η κατανάλωση φυσικού αερίου τριπλασιάστηκε και έκτοτε συνεχίζει να αυξάνεται με ταχύ ρυθμό.

Σημαντικό ρόλο για την αποδοχή του φυσικού αερίου διαδραμάτισε η εκτεταμένη επιβάρυνση του περιβάλλοντος από την καύση των συμβατικών καυσίμων. Σε μια εποχή όπου τα σοβαρά προβλήματα ατμοσφαιρικής ρύπανσης απαιτούν ενεργειακές επιλογές που συνδυάζουν την οικονομική ανάπτυξη με την περιβαλλοντική προστασία, είναι λογικό η ζήτηση του φυσικού αερίου να αυξάνεται ταχύτατα.

Στην Ελλάδα πρόδρομος του φυσικού αερίου ήταν το φωταέριο. Το διέθεσε στην αγορά, για πρώτη φορά το 1857, η Γαλλική Εταιρεία Φωταερίου. Η Δημοτική Επιχείρηση Φωταερίου (ΔΕΦΑ) συνέχισε να προμηθεύει τους καταναλωτές της με φωταέριο μέχρι το 1984. Τη χρονιά αυτή έγινε η σύνδεση με τα Ελληνικά Διυλιστήρια Ασπροπύργου (ΕΛ.Δ.Α.) και άρχισε η τροφοδότηση του δικτύου της ΔΕΦΑ με ναφθαέριο το οποίο χρησιμοποιήθηκε μέχρι το 1997. Το 1983 καταρτίζεται η πρώτη προμελέτη για το φυσικό αέριο στην Ελλάδα. Η μελέτη γίνεται για λογαριασμό της τότε Δημόσιας Επιχείρησης Πετρελαίου (ΔΕΠ) και το 1987 υπογράφεται η πρώτη διακρατική συμφωνία μεταξύ Ελλάδας και Ρωσίας για την προμήθεια φυσικού αερίου.

Ακολουθούν συμφωνίες της ΔΕΠ με την ρωσική Sojuzgazexport, σήμερα Gazexport, και με την Sonatrach της Αλγερίας. Το Σεπτέμβριο του 1988 ιδρύεται η Δημόσια Επιχείρηση Αερίου (ΔΕΠΑ) ως θυγατρική εταιρία της Δημόσιας Επιχείρησης Πετρελαίου. Η ΔΕΠΑ επιφορτίζεται με την ευθύνη μιας μεγάλης ενεργειακής επένδυσης και αναλαμβάνει την εισαγωγή, την μεταφορά και τη διανομή Φυσικού Αερίου σε όλη την Ελλάδα. Το Μάιο του 1997 ολοκληρώνεται το σύστημα μεταφοράς και τίθεται σε πλήρη λειτουργία και αρχίζει η συνεχής τροφοδότηση της Ελλάδας με ρωσικό αέριο.

1.3.3 Χρήσεις - Εφαρμογές φυσικού αερίου

Το φυσικό αέριο μπορεί να χρησιμοποιηθεί:

- **Για οικιακή χρήση.**

  Για κεντρική θέρμανση της πολυκατοικίας, ατομική θέρμανση του διαμερίσματος ή της μονοκατοικίας, μαγείρεμα, θέρμανση νερού, κλιματισμό.

- **Για επαγγελματική χρήση.**

  Για θέρμανση, μαγείρεμα, παραγωγή ζεστού νερού, παραγωγή ατμού, κλιματισμό, συμπαραγωγή ηλεκτρικής και θερμικής ενέργειας.

- **Για βιομηχανική χρήση.**

  Για κάλυψη θερμικών αναγκών για όλες τις παραγωγικές διαδικασίες, συμπαραγωγή ηλεκτρικής και θερμικής ενέργειας, κλιματισμό.

1.3.4 Οφέλη από τη χρήση φυσικού αερίου

Οικονομικά και ενεργειακά οφέλη

Η χρήση του φυσικού αερίου έχει πολλά οφέλη για τους καταναλωτές αλλά και για το περιβάλλον. Καταρχάς, είναι μια πολύ οικονομική μορφή ενέργειας καθώς είναι πιο οικονομικό από τις ανταγωνιστικές μορφές ενέργειας όπως το πετρέλαιο και το ηλεκτρικό ρεύμα. Για παράδειγμα στην χώρα μας, το φυσικό αέριο είναι 20% πιο οικονομικό σε σχέση με το πετρέλαιο και 60% σε σχέση με το ηλεκτρικό ρεύμα σε μηνιαία βάση. Επίσης, το φυσικό αέριο παρουσιάζει ένα ακόμη πλεονέκτημα μια και η υψηλότερη απόδοση της καύσης του έχει σαν αποτέλεσμα την εξοικονόμηση ενέργειας μέχρι και 30%.

Πολύ γρήγορα ο καταναλωτής ή επιχείρηση αποσβένει την επένδυση που έκανε και που αφορά την αγορά του απαραίτητου εξοπλισμού και το αρχικό κόστος της εγκατάστασης. Η μέτρηση της κατανάλωσης γίνεται με ακρίβεια βάση των ενδείξεων των εγκατεστημένων μετρητών. Επίσης, η πληρωμή των λογαριασμών του φυσικού αερίου πραγματοποιείται μετά την κατανάλωσή του σε αντίθεση με το πετρέλαιο που εξοφλείται κατά την παραλαβή. Για την χρήση του δεν απαιτείται αποθηκευτικός χώρος αφού λόγω των δικτύων διανομής δεν υπάρχουν προβλήματα παραγγελίας, μεταφοράς, παραλαβής και αποθήκευσης.

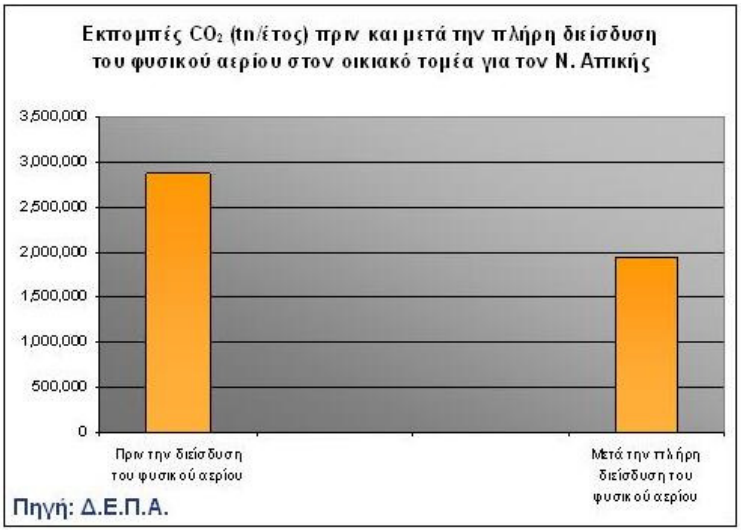
Γενικά, το φυσικό αέριο χρησιμοποιείται για την παραγωγή ηλεκτρικής ενέργειας. Με τη χρήση του φυσικού αερίου, όμως, απευθείας σε οικιακές και βιομηχανικές χρήσεις αποφεύγονται οι απώλειες μετατροπής και μεταφοράς του φυσικού αερίου. Η χρησιμοποίηση φυσικού αερίου σε μονάδες παραγωγής ηλεκτρικής ενέργειας συνδυασμένου κύκλου θα έχει ως αποτέλεσμα τη σημαντική αύξηση του βαθμού απόδοσης παραγωγής ηλεκτρισμού σε 52-55% έναντι 35-40% των συμβατικών ηλεκτροπαραγωγικών σταθμών. Λόγω της "καθαρότητας" των προϊόντων καύσης του φυσικού αερίου, μπορεί να χρησιμοποιηθεί απευθείας σε βιομηχανικές εφαρμογές χωρίς την παρεμβολή εναλλακτών που έχουν ως συνέπεια ενεργειακές απώλειες.

Συγκριτικά με το πετρέλαιο, υπάρχει πολύ μεγαλύτερη επάρκεια αποθεμάτων φυσικού αερίου (περί τα 350 τρισ. κυβικά μέτρα) και η τροφοδοσία της παγκόσμιας ενεργειακής αγοράς είναι κατά πολύ ασφαλέστερη. Η χρήση του φυσικού αερίου έχει σημαντικές θετικές επιδράσεις στη δομή του ενεργειακού ισοζυγίου μιας χώρας, μια και θα μειωθεί η εξάρτησή από το πετρέλαιο. Ως εκ τούτου, η επιλογή του φυσικού αερίου ως κύριας ενεργειακής πηγή αποτελεί παγκόσμια στρατηγική επιλογή.

Οφέλη για το περιβάλλον

Το σημαντικότερο όφελος όμως είναι ότι το φυσικό αέριο, λόγω της μορφής και της σύστασής του, είναι ένα κατεξοχήν οικολογικό καύσιμο. Οι ρύποι που εκλύονται στο περιβάλλον κατά την καύση του είναι σημαντικά μειωμένοι σε σχέση με αυτούς που παράγονται από τα άλλα καύσιμα. Αξίζει να σημειωθεί ότι το φυσικό αέριο, ενώ παγκοσμίως συμμετέχει στην κατανάλωση καυσίμων κατά 22.5%, ευθύνεται μόνο για το 16% των εκπομπών διοξειδίου του άνθρακα. Για την παραγωγή ίσου ποσού ενέργειας, το φυσικό αέριο εκπέμπει 30% λιγότερο διοξείδιο του άνθρακα σε σχέση με το πετρέλαιο και 50% λιγότερο διοξείδιο του άνθρακα σε σχέση με τον άνθρακα. Το στοιχείο αυτό είναι ιδιαίτερα σημαντικό δεδομένου ότι το διοξείδιο του άνθρακα είναι ένας από τους βασικούς παράγοντες δημιουργίας του Φαινομένου του Θερμοκηπίου.

Όσον αφορά στους άλλους παραγόμενους ρύπους, το φυσικό αέριο εκπέμπει 2 φορές λιγότερο μονοξείδιο του άνθρακα και 4.700 φορές λιγότερο διοξείδιο του θείου από το μαζούτ. Μετά την πλήρη διείσδυση του φυσικού αερίου στην αγορά της χώρας μας παρατηρήθηκε μια εμφανής πτώση των εκπομπών διοξειδίου του άνθρακα στην ατμόσφαιρα όπως φαίνεται και από το παρακάτω διάγραμμα.

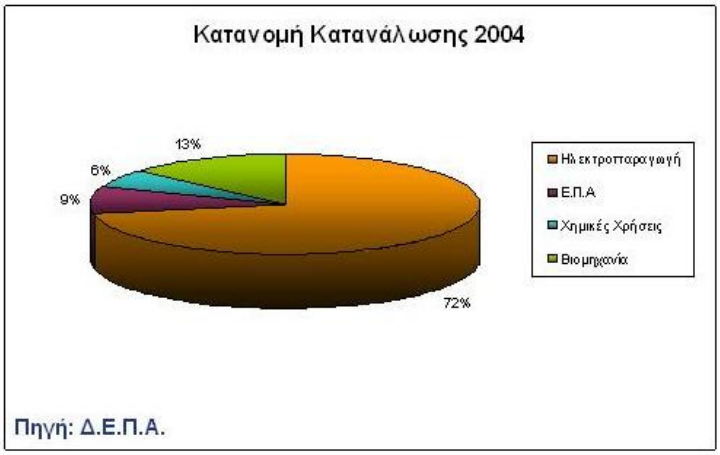


Σε σύγκριση με το ντίζελ η εκπομπή μονοξειδίου του άνθρακα και διοξειδίου του θείου είναι μειωμένη κατά 2,3 και 733 φορές αντίστοιχα. Επιπλέον, η καύση του φυσικού αερίου είναι καθαρή, με αποτέλεσμα τη μειωμένη εκπομπή αιθάλης και αιωρούμενων σωματιδίων. Η μειωμένη εκπομπή οξειδίων του θείου περιορίζει τη δημιουργία του φαινομένου της όξινης βροχής, προστατεύοντας με αυτό τον τρόπο τα δάση και τους βιότοπους. Ταυτόχρονα, λόγω της χαμηλής εκπομπής διοξειδίου του θείου, επιβραδύνεται η φθορά των αρχαίων μνημείων από τη γυψοποίηση του μαρμάρου.

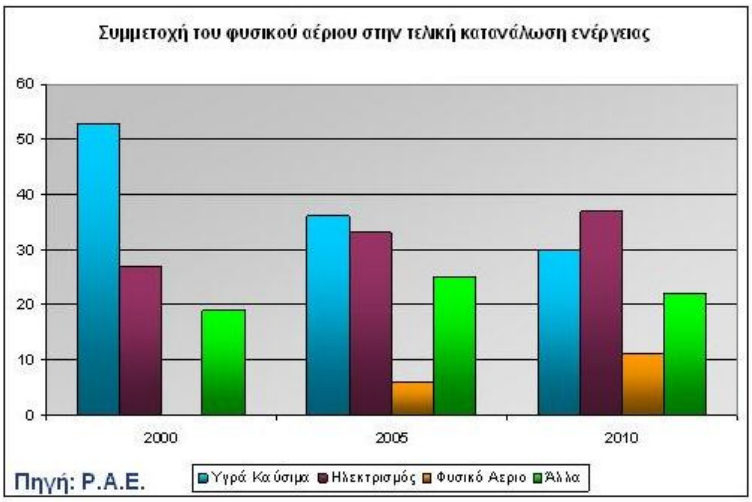
Το φυσικό αέριο, κατά συνέπεια, έχει αναδειχθεί στο κατεξοχήν οικολογικό καύσιμο του 21ου αιώνα και η ισχυροποίηση της θέσης του στη ενεργειακή αγορά της Ελλάδας προωθεί ουσιαστικά εκείνο το είδος ανάπτυξης που συμβαδίζει με την προστασία της φύσης και του περιβάλλοντος.

1.3.5 Γενικά στοιχεία – Η αγορά φυσικού αερίου στην Ελλάδα

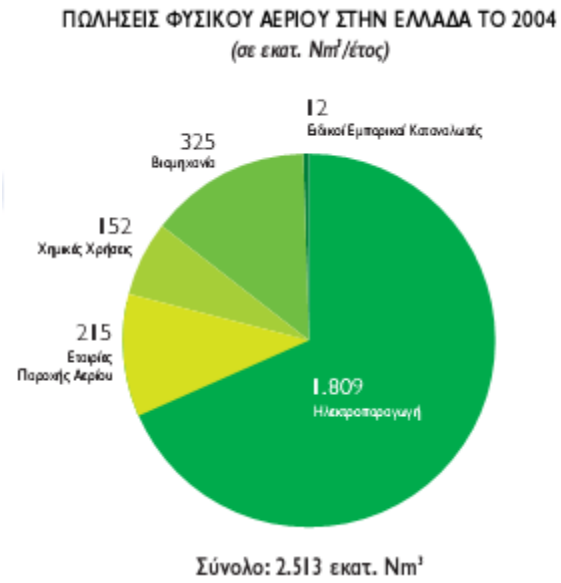
Ο συνδυασμός των πλεονεκτημάτων που χαρακτηρίζουν το φυσικό αέριο το κατατάσσουν ως το ταχύτερα αναπτυσσόμενο σε όλο τον κόσμο. Εξαιτίας των χαρακτηριστικών και των πλεονεκτημάτων του φυσικού αερίου αποφασίστηκε η εισαγωγή του στην Ελλάδα στα πλαίσια της προσπάθειας εκσυγχρονισμού και βελτίωσης του ενεργειακού ισοζυγίου, αλλά και για τη διαφοροποίηση των ενεργειακών πηγών της χώρας μας. Οι τομείς στους οποίους χρησιμοποιείται πρωτίστως το φυσικό αέριο είναι η ηλεκτροπαραγωγή και δευτερευόντως η βιομηχανία, ο οικιακός και τριτογενής τομέας όπως φαίνεται και από την κατανομή της κατανάλωσης για το 2004.



Με τη δημιουργία της υποδομής για τη μεταφορά και τη διανομή του φυσικού αερίου, άνοιξε ο δρόμος για τη συμμετοχή του στο ενεργειακό ισοζύγιο της χώρας. Ιδιαίτερα μετά την απελευθέρωση της αγοράς ηλεκτρικής ενέργειας, οι ανάγκες της χώρας μας για φυσικό αέριο αυξήθηκαν ραγδαία. Το φυσικό αέριο υπολογίζεται να καλύπτει πάνω από το 15% των ενεργειακών αναγκών μέχρι το 2010. Στο παρακάτω διάγραμμα φαίνεται το ποσοστό συμμετοχής του φυσικού αερίου για την παραγωγή ενέργειας στη χώρας μας καθώς και μια εκτίμηση για το 2010.



Στην ελληνική αγορά η διανομή του φυσικού αερίου στον οικιακό τομέα ξεκίνησε τα τελευταία έξι χρόνια. Το 2005 παρατηρήθηκε αύξηση της κατανάλωσης φυσικού αερίου σε σχέση με τα επίπεδα του 2003 κατά 94,1% και κατά 39,3% σε σχέση με την κατανάλωση του 2004. Στο παρακάτω διάγραμμα παρουσιάζεται η κατανομή των πωλήσεων φυσικού αερίου στην Ελλάδα για το 2004.



1.3.6 Γενικά για τις εταιρείες εκμετάλλευσης φυσικού αερίου στην Ελλάδα

Εταιρεία μεταφοράς

Δ.Ε.Π.Α. Α.Ε.

Ιδρύθηκε το 1988, με σκοπό την ανάπτυξη της υποδομής για την εισαγωγή, τη μεταφορά και τη διάθεση του φυσικού αερίου στην Ελλάδα. Σκοπός της εταιρείας είναι η ανάπτυξη του συστήματος φυσικού αερίου με τρόπο που να διασφαλίζει την αδιάλειπτη και ασφαλή τροφοδοσία των καταναλωτών και η επιτάχυνση της διείσδυσης του φυσικού αερίου σε όλους τους ενεργειακούς τομείς. Η Δ.Ε.Π.Α. εισάγει φυσικό αέριο από τη ρωσική εταιρεία Gasexport και την αλγερινή Sonatrach και ασχολείται με την πώληση φυσικού αερίου στις εταιρείες παροχής αερίου (Ε.Π.Α.) και σε καταναλωτές με κατανάλωση μεγαλύτερη από 10 εκατομμύρια Nm3 ετησίως.

Γενικά, η Δ.Ε.Π.Α. είναι υπεύθυνη για την :

- Εισαγωγή, μεταφορά, αποθήκευση φυσικού αερίου.
- Κατασκευή και εκμετάλλευση του Εθνικού Συστήματος Μεταφοράς φυσικού αερίου.
- Πώληση φυσικού αερίου σε μεγάλους καταναλωτές, με ετήσια κατανάλωση άνω των 10 εκατομ. κυβικών μέτρων (κ.μ.).
- Πώληση φυσικού αερίου σε Εταιρίες Παροχής Αερίου.
- Διανομή φυσικού αερίου σε περιοχές όπου δεν έχουν συσταθεί Εταιρίες Παροχής Αερίου.

Η Δ.Ε.Π.Α. κατασκεύασε το Σύστημα Μεταφοράς Υψηλής Πίεσης, ένα σύστημα αγωγών φυσικού αερίου που διατρέχουν τη χώρα, καθώς και έναν υπερσύγχρονο σταθμό υποδοχής υγροποιημένου φυσικού αερίου στη νήσο Ρεβυθούσα, κοντά στην Αθήνα. Μετά τη σύνδεση του ελληνικού Συστήματος με το αντίστοιχο ρωσικό, η Δ.Ε.Π.Α. ανέλαβε τη λειτουργία, τη συντήρηση και την επέκτασή του και σε άλλες περιοχές της χώρας.

Η Δ.Ε.Π.Α. ίδρυσε τις Ε.Π.Α. (Εταιρία Διανομής Αερίου) που είναι υπεύθυνες για τη διανομή φυσικού αερίου στις διάφορες περιοχές της χώρας. Μέχρι σήμερα έχουν ιδρυθεί η Ε.Π.Α. Αττικής, η Ε.Π.Α. Θεσσαλονίκης και η Ε.Π.Α. Θεσσαλίας.

Εταιρείες διανομής

Ε.Π.Α. **Αττικής**

Στην Αττική ο αποκλειστικός διανομέας φυσικού αερίου είναι η Ε.Π.Α. Αττικής (Εταιρεία Παροχής Αερίου Αττικής) ή αλλιώς Φυσικό Αέριο Αττικής. Η εταιρεία ιδρύθηκε το 2001, με σκοπό να καθιερώσει την πιο αποδοτική, οικονομική και οικολογική μορφή ενέργειας, το φυσικό αέριο.

Η Ε.Π.Α. Αττικής είναι υπεύθυνη

- Για την ανάπτυξη, λειτουργία, συντήρηση και εμπορική εκμετάλλευση του δικτύου διανομής στον νομό Αττικής.
- Για την πώληση φυσικού αερίου στον οικιακό, τριτογενή και βιομηχανικό τομέα.
- Για τις διαδικασίες που σχετίζονται με το φυσικό αέριο (εγκατάσταση, τιμολόγηση, εξυπηρέτηση, ενημέρωση).

Στον οικιακό τομέα η Ε.Π.Α. Αττικής προμηθεύει περίπου 110.000 νοικοκυριά, στο βιομηχανικό τομέα προμηθεύει βιομηχανίες με κατανάλωση μικρότερη των 10GWh και στον τριτογενή τομέα προμηθεύει επιχειρήσεις του ξενοδοχειακού κλάδου, νοσοκομεία, κολυμβητήρια, πανεπιστήμια, εστίες και άλλες μικρότερες επιχειρήσεις.

Ε.Π.Α. Θεσσαλονίκης

Στη Θεσσαλονίκη ο αποκλειστικός διανομέας φυσικού αερίου είναι η Ε.Π.Α. Θεσσαλονίκης ή αλλιώς Αέριο Θεσσαλονίκης. Η εταιρεία ιδρύθηκε το 2000 και έχει τις ίδιες ευθύνες όπως η Ε.Π.Α. Αττικής.

Ε.Π.Α. Θεσσαλίας

Στη Θεσσαλία ο αποκλειστικός διανομέας φυσικού αερίου είναι η Ε.Π.Α. Θεσσαλίας ή αλλιώς Αέριο Θεσσαλίας. Η εταιρεία ιδρύθηκε το 2000 και έχει τις ίδιες ευθύνες όπως η Ε.Π.Α. Αττικής.

1.3.7 Το σύστημα μεταφοράς φυσικού αερίου στην Ελλάδα

Το φυσικό αέριο εισάγεται στη χώρα μας μέσω αγωγών υψηλής πίεσης (70bar). Το σύστημα μεταφοράς φυσικού αερίου αποτελείται από τα εξής βασικά τμήματα :

- Κεντρικός αγωγός μεταφοράς αερίου υψηλής πίεσης (70 bar),από τα ελληνοβουλγαρικά σύνορα μέχρι την Αττική, συνολικού μήκους 512 χλμ.

- Κλάδοι μεταφοράς υψηλής πίεσης προς την ανατολική Μακεδονία και Θράκη, τη Θεσσαλονίκη, το Βόλο και την Αττική, συνολικού μήκους 440 χλμ.

- Μετρητικοί και ρυθμιστικοί σταθμοί για τη μέτρηση της παροχής αερίου και τη ρύθμιση της πίεσης,.

- Σύστημα τηλεχειρισμού, ελέγχου λειτουργίας και τηλεπικοινωνιών.

- Κέντρα λειτουργίας και συντήρησης, στην Αττική, τη Θεσσαλονίκη, τη Θεσσαλία και στην Ξάνθη.

- Εγκαταστάσεις Υγροποιημένου Φυσικού Αερίου (Υ.Φ.Α.) στη νήσο Ρεβυθούσα.

Τερματικός σταθμός Ρεβυθούσας

Οι εγκαταστάσεις υγροποιημένου φυσικού αερίου (Υ.Φ.Α.) στη νήσο Ρεβυθούσα, στον κόλπο των Μεγάρων Αττικής, αποτελούν τμήμα της βασικής υποδομής του δικτύου μεταφοράς. Πρόκειται για μία σύγχρονη μονάδα που αποσκοπεί στην αποθήκευση υγροποιημένου αερίου για την κάλυψη των αιχμών ζήτησης φυσικού αερίου και στην αύξηση της αξιοπιστίας του συστήματος. Περιλαμβάνει δύο δεξαμενές υγροποιημένου φυσικού αερίου, συνολικής χωρητικότητας 130.000 εκ. κ.μ., εγκαταστάσεις ελλιμενισμού δεξαμενόπλοιων, κρυογονικές εγκαταστάσεις και αεροποιητές για την επαναεριοποίηση του υγροποιημένου αερίου.

Για την τροφοδοσία του Συστήματος Μεταφοράς, έχει κατασκευαστεί δίδυμος αγωγός που συνδέει τη Ρεβυθούσα με την ακτή της Αγ. Τριάδας. Το Δεκέμβριο του 1999 ολοκληρώθηκε η κατασκευή του Τερματικού Σταθμού, ενώ από το Φεβρουάριο του 2000 ο Σταθμός βρίσκεται σε πλήρη λειτουργία. Για τη μεταφορά του Υγροποιημένου Φυσικού Αερίου, έχει ναυλωθεί δεξαμενόπλοιο χωρητικότητας 29.500 κ.μ.

Σε συνέχεια της απόφασης της Δ.Ε.Π.Α. για αύξηση της παροχετευτικής δυναμικότητας του Τερματικού Σταθμού από 270 m^3/h Υ.Φ.Α. σε 1000 m^3/h Υ.Φ.Α., η Δ.Ε.Π.Α. ολοκλήρωσε τόσο τον Βασικό Σχεδιασμό του Έργου, όσο και τις απαραίτητες μελέτες Ασφάλειας και Περιβαλλοντικών επιπτώσεων.

1.3.8 Το σύστημα διανομής φυσικού αερίου στην Ελλάδα

Το σύστημα διανομής αποτελείται από τα εξής βασικά τμήματα :

- Δίκτυα μέσης πίεσης (19bar) σε Αττική, Θεσσαλονίκη, Λάρισα, Βόλο, Οινόφυτα, Πλατύ Ημαθίας, Σέρρες, Ξάνθη, Καβάλα, Κομοτηνή

- Δίκτυα χαμηλής πίεσης (4bar) σε Αττική, Θεσσαλονίκη, Λάρισα, Βόλο, Κομοτηνή

Η Δ.Ε.Π.Α. μέσω των Εταιριών Διανομής Αερίου έχει παραχωρήσει στις θυγατρικές της Εταιρίες Παροχής Αερίου Αττικής, Θεσσαλονίκης, Θεσσαλίας τη χρήση των δικτύων μέσης και χαμηλής πίεσης που αφορούν τις αντίστοιχες περιοχές.

Επίσης η Δ.Ε.Π.Α., θα επεκτείνει τα δίκτυά της στις περιοχές της Ανατολικής Στερεάς Ελλάδας και Ευβοίας, Ανατολικής Μακεδονίας και Θράκης καθώς και στην περιοχή της Κεντρικής Μακεδονίας, και προτίθεται να παραχωρήσει τη χρήση των νέων δικτύων σε νέες ΕΠΑ που θα ιδρυθούν σε αυτές τις περιοχές.

ΚΕΦΑΛΑΙΟ 2: ΥΦΙΣΤΑΜΕΝΗ ΚΑΤΑΣΤΑΣΗ ΚΑΙ ΠΡΟΒΛΗΜΑΤΑ

2.1 ΓΕΝΙΚΑ

Στο κεφάλαιο αυτό γίνεται μια ανάλυση της υφιστάμενης κατάστασης της εταιρείας διανομής φυσικού αερίου Αττικής και παρουσιάζονται οι λόγοι που καθιστούν αναγκαία την ανάπτυξη ενός μοντέλου πρόβλεψης της κατανάλωσης. Επίσης, προσδιορίζονται όλοι οι παράγοντες που σχετίζονται και επηρεάζουν την κατανάλωση φυσικού αερίου.

2.2 ΥΦΙΣΤΑΜΕΝΗ ΚΑΤΑΣΤΑΣΗ

Η εταιρεία διανομής φυσικού αερίου κατηγοριοποιεί τους πελάτες της σε δύο μεγάλες κατηγορίες. Η πρώτη κατηγορία περιλαμβάνει τους μεγάλους βιομηχανικούς πελάτες (B2B) με κατανάλωση μεγαλύτερη από 1100MWh ετησίως ενώ η δεύτερη κατηγορία (B2C) περιλαμβάνει τους πελάτες με κατανάλωση μικρότερη από 1100MWh ετησίως. Η καταγραφή των ενδείξεων των μετρητών πραγματοποιείται μηνιαίως για τους μεγάλους βιομηχανικούς πελάτες και η εταιρεία εκδίδει μηνιαία τιμολόγια. Για τους οικιακούς και μικρούς βιομηχανικούς καταναλωτές, η καταγραφή των ενδείξεων των μετρητών πραγματοποιείται διμηνιαίως και η εταιρεία εκδίδει τιμολόγια ανά δίμηνο.

Η χρέωση των πελατών προκύπτει από το άθροισμα της αξίας χρέωσης ενέργειας και της αξίας χρέωσης ισχύος. Η αξία χρέωσης ενέργειας είναι η αξία της ποσότητας του φυσικού αερίου που καταναλώθηκε σε συγκεκριμένο χρονικό διάστημα και η αξία χρέωσης ισχύος είναι η τιμή της σταθερής χρέωσης που έχει προϋπολογιστεί από την εταιρεία διανομής βάσει της εγκατεστημένης ισχύος. Η εταιρεία είναι σε θέση να γνωρίζει μηνιαία, τη κατανάλωση όλων των μεγάλων βιομηχανικών πελατών και μέρους των οικιακών και μικρών βιομηχανικών πελατών (όσων πελατών η καταγραφή των ενδείξεων του μετρητή γίνεται τον παρόντα μήνα).

Για τους πελάτες που η κατανάλωσή τους δεν είναι δυνατόν να προκύψει από την καταγραφή των ενδείξεων των μετρητών αφού αυτή γίνεται ανά δίμηνο, πραγματοποιείται εκτίμηση της κατανάλωσης από την εταιρεία διανομής. Στην ουσία δεν είναι απαραίτητη για την εταιρεία μια μελλοντική πρόβλεψη αν και αυτό είναι δυνατόν εξυπηρετείται από το μοντέλο αλλά ένας όσο το δυνατόν πιο ακριβής υπολογισμός της μηνιαίας κατανάλωσης των οικιακών και μικρών βιομηχανικών πελατών.

2.3 Η ΑΝΑΓΚΑΙΟΤΗΤΑ ΤΗΣ ΠΡΟΒΛΕΨΗΣ ΕΙΔΙΚΑ ΓΙΑ ΜΙΑ ΕΤΑΙΡΕΙΑ ΔΙΑΝΟΜΗΣ

Οι εταιρείες διανομής φυσικού αερίου χρησιμοποιούν διάφορες μεθόδους πρόβλεψης της κατανάλωσης και για διάφορους σκοπούς. Ειδικότερα, για την εταιρεία διανομής η πρόβλεψη της κατανάλωσης είναι απαραίτητη :

α) Για την τιμολόγηση των πελατών.

Η πρόβλεψη της μηνιαίας κατανάλωσης είναι πολύ χρήσιμη για την τιμολόγηση των πελατών για τους παρακάτω λόγους.

- Για την εκτίμηση της κατανάλωσης όταν δεν είναι δυνατή η καταγραφή των ενδείξεων των μετρητών. Συγκεκριμένα, η καταγραφή των ενδείξεων των μετρητών πραγματοποιείται από καταμετρητές ανά δίμηνο. Προκύπτουν όμως προβλήματα και καθυστερήσεις στην καταγραφή των ενδείξεων που συνήθως προέρχονται είτε από το φόρτο εργασίας, είτε από τεχνικές δυσκολίες, είτε από τον μη επαρκή αριθμό διαθέσιμων καταμετρητών, είτε από τις καιρικές συνθήκες, είτε από άλλους. Επίσης, μειώνεται η αβεβαιότητα που προκαλεί η μεταβολή των εξωτερικών συνθηκών όπως π.χ. οι καιρικές συνθήκες, διάφοροι αστάθμητοι παράγοντες

- Για την επιβεβαίωση των αποτελεσμάτων των τιμολογήσεων. Συγκεκριμένα έχει διαπιστωθεί ότι η λήψη των ενδείξεων για διάφορους λόγους (άγνοια ή αδιαφορία ή οκνηρία των καταμετρητών) σε ορισμένες περιπτώσεις δεν είναι αξιόπιστη και για αυτό το λόγο είναι αναγκαία η ανάπτυξη του παρόντος μοντέλου ώστε με αξιόπιστο και γρήγορο τρόπο να διεξάγεται έλεγχος.

- Για τη σωστή μηνιαία κατανομή της κατανάλωσης εντός του διμηνιαίου λογαριασμού του πελάτη. Συγκεκριμένα, επειδή η τιμή του αερίου αλλάζει ανά μήνα και η καταγραφή των ενδείξεων των μετρητών και η τιμολόγηση γίνεται ανά δίμηνο, χρειάζεται να γνωρίζουμε ποια ποσότητα αντιστοιχεί σε κάθε μήνα, ώστε να γίνεται σωστά η τιμολόγηση.

β) Για τα οικονομικά της εταιρείας διανομής.

Η πρόβλεψη της μηνιαίας κατανάλωσης είναι πολύ χρήσιμη για τα οικονομικά της επιχείρησης. Κατά κύριο λόγο, η πρόβλεψη βοηθά στην εκτίμηση των μηνιαίων εσόδων της εταιρείας διανομής ειδικά όταν δεν έχουν ολοκληρωθεί οι τιμολογήσεις. Η εταιρεία διανομής υποχρεούται να γνωρίζει την ποσότητα και την αξία της ατιμολόγητης κατανάλωσης. Όπως αναφέρθηκε παραπάνω, για τους οικιακούς και μικρούς επαγγελματικούς καταναλωτές, η καταγραφή των ενδείξεων των μετρητών πραγματοποιείται διμηνιαίως και επομένως είναι δύσκολο για την εταιρεία διανομής να υπολογίζει την μηνιαία ποσότητα φυσικού αερίου που πουλά στους καταναλωτές της.

Επίσης, με τη μηνιαία πρόβλεψη της κατανάλωσης η εταιρεία διανομής είναι σε θέση να υπολογίσει με περισσότερη ακρίβεια τα έσοδα, τα αποτελέσματα και τον προϋπολογισμό ανά μήνα και να γνωρίζει την ποσότητα του φυσικού αερίου που πωλήθηκε στους καταναλωτές, παράγοντας με αυτόν τον τρόπο ακριβή εμπορικά – οικονομικά αποτελέσματα.

γ) Για τη διεξαγωγή ενός επιτυχημένου μάρκετινγκ.

Το μάρκετινγκ κατέχει υψηλή θέση στις προτεραιότητες κάθε εταιρείας. Ακολουθώντας μια σωστή στρατηγική μάρκετινγκ, η εταιρεία διανομής φυσικού αερίου μπορεί να αυξήσει τα κέρδη της και να θωρακιστεί ενάντια στις πτωτικές τάσεις και τις μεταβολές της αγοράς. Η πρόβλεψη της κατανάλωσης είναι κλειδί για τη διεξαγωγή ενός επιτυχημένου μάρκετινγκ και βοηθά στον καλύτερο προσδιορισμό της αγοράς - στόχου και της συμπεριφοράς των καταναλωτών.

δ) Για το στρατηγικό και επιχειρησιακό σχεδιασμό.

Ανάλογα με τις προβλεπόμενες πωλήσεις, κάθε εταιρεία αναπτύσσεται στην αγορά αναλόγως, καθορίζει τις μελλοντικές στρατηγικές και ενέργειες που πρόκειται να εφαρμόσει και οργανώνει και εκσυγχρονίζει τις διαδικασίες που ακολουθεί. Με την πρόβλεψη της κατανάλωσης, η εταιρεία μπορεί να προβεί σε ενέργειες με σκοπό τη βελτίωση της λειτουργίας της μέσω της αξιοποίησης νέων τεχνολογιών, της καλύτερης οργάνωσης, της αξιοποίησης του ανθρώπινου δυναμικού ώστε να καλύπτονται οι ανάγκες και να βελτιωθεί η αποτελεσματικότητά της.

Επίσης, μέσω της πρόβλεψης της κατανάλωσης, επιτυγχάνεται η παραγωγή αξιόπιστων μελλοντικών δεδομένων για την επεξεργασία σε όλα τα επίπεδα προγραμματισμού. Δίνεται η ευκαιρία να αναγνωριστούν και να αξιολογηθούν οι παράγοντες που επηρεάζουν πιο πολύ την κατανάλωση ώστε να μπορεί η εταιρεία να προβεί στις απαραίτητες ενέργειες και να επιστήσει την προσοχή της σε συγκεκριμένες ομάδες-στόχους ή παράγοντες. Επίσης, ένα μοντέλο πρόβλεψης είναι πολύ χρήσιμο για την εκτίμηση πιθανών σεναρίων για διάφορα δεδομένα και για την υποστήριξη των αποφάσεων της επιχείρησης.

ε) Για τη διαχείριση κίνδυνων.

Κάθε εταιρεία έχει την ανάγκη για κάλυψη από τους κινδύνους αγοράς μέσω της διαδικασίας του προσδιορισμού, της ανάλυσης και της ανάπτυξης στρατηγικών για αντιμετώπιση των κινδύνων. Η πρόβλεψη της κατανάλωσης βοηθά στον προσδιορισμό και την εκτίμηση πιθανών κινδύνων για την εταιρεία.

η) Για την παρακολούθηση των συμβάσεων με τους προμηθευτές.

Η πρόβλεψη είναι ένα εργαλείο απαραίτητο για τον προγραμματισμό και την παρακολούθηση της σύμβασης με την εταιρεία μεταφοράς. Καθώς η εταιρεία διανομής αγοράζει αέριο από μια εταιρεία μεταφοράς, πρέπει να παρέχει στην εταιρεία μεταφοράς μια πρόβλεψη των μελλοντικών αναγκών της έτσι ώστε η εταιρεία μεταφοράς να δύναται να προγραμματίσει τη διάθεση του απαραίτητου αερίου. Εάν το λάθος της πρόβλεψης είναι μεγάλο, η εταιρεία μεταφοράς ενδέχεται να επιβάλλει οικονομικές ποινές εις βάρος της εταιρείας διανομής λόγω αθέτησης των ρητρών που αφορούν κυρίως την ετήσια ή ημερήσια κατανάλωση, αλλά και τις μέγιστες ωριαίες ποσότητες. Με αυτό τον τρόπο, η εταιρεία έχει την δυνατότητα να ελαχιστοποιεί την επιβολή ρητρών.

θ) Για την ενημέρωση των ενδιαφερομένων.

Το μοντέλο που αναπτύσσεται σε αυτή την εργασία έχει σαν αποτέλεσμα την πρόβλεψη της μηνιαίας κατανάλωσης των οικιακών πελατών. Με μια κατάλληλη προσαρμογή, το μοντέλο είναι δυνατόν να υπολογίζει τη μηνιαία κατανάλωση ενός υποψήφιου καταναλωτή αφού εισαχθούν σε αυτό τα απαραίτητα δεδομένα που τον αφορούν. Η εταιρεία διανομής μπορεί να αξιοποιήσει και με αυτό τον τρόπο το μοντέλο, παρέχοντας στον υποψήφιο καταναλωτή τον υπολογισμό της προβλεπόμενης μηνιαίας κατανάλωσής του. Με αυτό το τρόπο ο υποψήφιος καταναλωτής φυσικού αερίου θα έχει τη δυνατότητα να ενημερωθεί για την προβλεπόμενη μηνιαία κατανάλωση του αλλά και να συγκρίνει αν τον συμφέρει η χρήση του φυσικού αερίου σε σχέση με τα ανταγωνιστικά προϊόντα. Ο υπολογισμός, λοιπόν, της κατανάλωσης φυσικού αερίου για τον κάθε πελάτη θα μπορούσε να προσφερθεί στους υποψήφιους καταναλωτές μέσω του διαδικτύου σαν μια εύκολη, γρήγορη και πρωτοποριακή υπηρεσία της εταιρείας φυσικού αερίου για την ενημέρωση των ενδιαφερομένων και για πρόσκτηση νέων πελατών.

2.4 Ο ΥΦΙΣΤΑΜΕΝΟΣ ΤΡΟΠΟΣ ΠΡΟΒΛΕΨΗΣ ΤΗΣ ΚΑΤΑΝΑΛΩΣΗΣ

Η εταιρεία διανομής χρησιμοποιεί την παρούσα στιγμή δύο τρόπους για να πραγματοποιήσει την μηνιαία πρόβλεψη όπου ο ένας βασίζεται στον άλλο. Και οι δύο τρόποι δεν αποτελούν μοντέλα πρόβλεψης αλλά κάνουν επεξεργασία στοιχείων.

Ο πρώτος τρόπος παίρνει σαν δεδομένη την κατανάλωση των μεγάλων βιομηχανικών πελατών και μέρους των οικιακών και μικρών βιομηχανικών (αφού αυτές οι κατηγορίες τιμολογούνται ανά μήνα) και αφαιρεί αυτή την ποσότητα φυσικού αερίου από την μηνιαία ποσότητα που προμηθεύτηκε από την εταιρεία μεταφοράς. Το αποτέλεσμα που προκύπτει δίνει την μηνιαία ποσότητα φυσικού αερίου που ενώ η εταιρεία πούλησε στους οικιακούς και μικρούς βιομηχανικούς πελάτες της μέχρι στιγμής δεν έχει τιμολογήσει.

Ο δεύτερος τρόπος κάνει χρήση των προηγούμενων καταναλώσεων φυσικού αερίου για κάθε κωδικό καταναλωτή. Ύστερα, γίνεται πρόβλεψη της κατανάλωσης ανάλογα με τις υπολογισμένες βαθμοημέρες θέρμανσης για την εκτιμώμενη περίοδο, συνδυάζοντας τα στοιχεία των προηγούμενων καταναλώσεων και αντίστοιχων βαθμοημερών.

ΚΕΦΑΛΑΙΟ 3: ΚΑΘΟΡΙΣΜΟΣ ΚΑΙ ΕΠΙΛΟΓΗ ΚΡΙΤΗΡΙΩΝ

3.1 ΑΝΑΛΥΣΗ ΠΑΡΑΓΟΝΤΩΝ – ΚΡΙΤΗΡΙΩΝ

Αρχικά, η μελέτη ξεκίνησε από τον καθορισμό των παραγόντων εκείνων που επηρεάζουν την καθημερινή κατανάλωση φυσικού αερίου. Οι παράγοντες αυτοί είναι εκείνοι που είναι υπεύθυνοι για το πότε, γιατί και πόσο φυσικό αέριο καταναλώνεται σε κάθε κτίριο.

Στην Ελλάδα το φυσικό αέριο χρησιμοποιείται κυρίως για θέρμανση, είτε αναφερόμαστε σε οικιακούς, είτε σε μικρούς επαγγελματικούς καταναλωτές. Οι καταναλωτές επιδιώκουν να εξασφαλίσουν ένα άνετο εσωτερικό κλίμα πλήρως προσαρμοσμένο στις ανάγκες τους. Η θερμική άνεση των χρηστών ενός κτιρίου αποτελεί μια υποκειμενική αίσθηση που βασίζεται σ ένα σύνολο παραγόντων μεταξύ των οποίων είναι η θερμοκρασία, τα ρεύματα αέρα, η υγρασία του αέρα αλλά και ο ρουχισμός των ατόμων. Η θερμική άνεση των ατόμων στο χώρο είναι ουσιαστικά και αυτή που έχει τη μεγαλύτερη σημασία αφού σχετίζεται, απόλυτα, με την κατανάλωση ενέργειας. Ο κάθε ένοικος καθορίζει την θερμική άνεση που αισθάνεται σε ένα χώρο με τροποποίηση των συνθηκών λειτουργίας της εγκατάστασης θέρμανσης. Επομένως, η κατανάλωση φυσικού αερίου για την παραγωγή θερμότητας, συνδέεται ευθέως με την θερμική άνεση που επιδιώκουν να απολαύσουν, οι χρήστες.

Η καθημερινή κατανάλωση φυσικού αερίου επηρεάζεται από πληθώρα παραγόντων και θα πρέπει να επιλεχθούν αυτοί που επηρεάζουν ποιο πολύ την κατανάλωση. Εν συνεχεία, θα πρέπει να επιλεχθούν τα κριτήρια εκείνα που πρέπει να ληφθούν υπόψη και να συγκεντρωθούν ιστορικά δεδομένα ώστε να αρχίσει η επεξεργασία αυτών. Τα κριτήρια θα πρέπει να επιλεχθούν με γνώμονα τη διαθεσιμότητά και τη δυνατότητα συγκέντρωσής δεδομένων για ένα επαρκή χρονικό διάστημα. Η σημαντικότερη πηγή δεδομένων είναι η αίτηση παροχής φυσικού αερίου που συμπληρώνει ο καταναλωτής. Οπότε τα κριτήρια θα πρέπει να επιλεχθούν με βάση τη δυνατότητα απόκτησης δεδομένων από την αίτηση.

Βέβαια ελλοχεύουν και αρκετοί κίνδυνοι κατά την διάρκεια επιλογής κριτηρίων. Αν επιλεχθεί πολύ μικρό πλήθος κριτηρίων υπάρχει ο κίνδυνος να μην είναι ακριβές το εξαγόμενο αποτέλεσμα. Στην αντίθετη περίπτωση, δηλαδή να επιλεχθεί μεγάλο πλήθος κριτηρίων, η βελτίωση στην ακρίβεια του αποτελέσματος μπορεί να είναι αμελητέα. Οπότε είναι σημαντικό να επιλεχθούν εκείνα τα κριτήρια που πραγματικά επηρεάζουν την κατανάλωση στο μεγαλύτερο βαθμό, ώστε το επιθυμητό αποτέλεσμα να προκύψει με το λιγότερο δυνατό κόπο και τη μεγαλύτερη δυνατή ακρίβεια.

Για την καλύτερη κατανόηση και διαχείριση των παραγόντων που επηρεάζουν την κατανάλωση φυσικού αερίου, θεωρήθηκε σωστό να χωριστούν σε δύο μεγάλες κατηγορίες.

- Η **πρώτη κατηγορία** περιλαμβάνει τους παράγοντες που επηρεάζουν τη θερμοκρασία στο εσωτερικό της οικίας.

- Η **δεύτερη κατηγορία** ανήκουν οι παράγοντες που επηρεάζουν την τάση για κατανάλωση φυσικού αερίου και προσδιορίζουν τη ζήτησή του.

3.2 ΠΑΡΑΓΟΝΤΕΣ ΠΟΥ ΔΙΑΜΟΡΦΩΝΟΥΝ ΤΗ ΘΕΡΜΟΚΡΑΣΙΑ ΣΤΟ ΕΣΩΤΕΡΙΚΟ ΤΗΣ ΟΙΚΙΑΣ

Όπως αναφέρθηκε, η θερμοκρασία στο εσωτερικό του κτιρίου καθορίζεται από τους χρήστες ώστε να επιτυγχάνεται η θερμική τους άνεση. Αναλόγως της εσωτερικής θερμοκρασίας της οικίας, ο καταναλωτής κρίνει εάν θα ενεργοποιήσει ή όχι, το σύστημα θέρμανσης της οικίας του.

Η θερμοκρασία στο εσωτερικό ενός κτιρίου εξαρτάται από διάφορες παραμέτρους. Κυρίως, εξαρτάται από τις εξωτερικές κλιματολογικές συνθήκες που διαμορφώνονται αναλόγως της εποχής, της κλιματολογικής ζώνης-περιοχής και των ιδιαιτεροτήτων της. Επίσης, διαμορφώνεται από τις ενεργειακές απώλειες ή κέρδη του κτιρίου ανάλογα με το σχεδιασμό και τα υλικά της κατασκευής του.

3.2.1 Εξάρτηση από τις κλιματολογικές συνθήκες

Η εσωτερική θερμοκρασία του κτιρίου εξαρτάται από τις εξωτερικές κλιματολογικές συνθήκες που διαμορφώνονται αναλόγως της εποχής, της κλιματολογικής ζώνης-περιοχής και των ιδιαιτεροτήτων της. Τα βασικά χαρακτηριστικά που ορίζουν τις εξωτερικές κλιματολογικές συνθήκες είναι τα εξής:

- Θερμοκρασία
- Σχετική υγρασία
- Ταχύτητα του ανέμου
- Κατεύθυνση του ανέμου
- Βροχόπτωση
- Ηλιακή ακτινοβολία

3.2.2 Εξάρτηση από τις ενεργειακές απώλειες του κτιρίου

Η εξάρτηση από τις κλιματολογικές συνθήκες ενός κτιρίου διαμορφώνεται, επίσης, από τις ενεργειακές απώλειες ή κέρδη του κτιρίου ανάλογα με το σχεδιασμό και τα υλικά της κατασκευής του. Οι ενεργειακές απώλειες και κέρδη ενός κτιρίου προκύπτουν από τους παρακάτω λόγους:

- **Απώλειες λόγω μετάδοσης θερμότητας και εξαερισμού στο εξωτερικό περιβάλλον**. Οι απώλειες αυτές οφείλονται στο γεγονός ότι διαφεύγει θερμότητα από το κτίριο προς το εξωτερικό περιβάλλον παρόλο τις μονώσεις του κτιρίου. Οι απώλειες αυτές εξαρτώνται από την ποιότητα της θερμομόνωσης του κελύφους, τον αριθμό των ανοιγμάτων (θύρες, παράθυρα), τη θέση του διαμερίσματος (ισόγειο, ρετιρέ), το σχεδιασμό και το εμβαδόν του κτιρίου, τη διαρρύθμιση των εσωτερικών χώρων αλλά και από ένα σύνολο άλλων παραμέτρων.

- **Απώλειες ή κέρδη από γειτονικές ζώνες λόγω μετάδοσης της θερμότητας και εξαερισμού**. Οι απώλειες αυτές οφείλονται στο γεγονός ότι διαφεύγει θερμότητα από το κτίριο προς το έδαφος ή γειτονικά κτίρια. Οι απώλειες αυτές εξαρτώνται από τη γειτνίαση με κτίρια, τη θέση του διαμερίσματος (ισόγειο, ρετιρέ), τον τύπο της οικίας (πολυκατοικία, μονοκατοικία) και το εμβαδόν του χώρου.

- **Εσωτερικά θερμικά κέρδη**. Σημαντική πηγή θερμότητας στα κτίρια είναι οι άνθρωποι και οι δραστηριότητές τους. Εν γένει τα εσωτερικά θερμικά φορτία προκαλούνται από τους ανθρώπους (θερμική ενέργεια που εκλύεται λόγω του μεταβολισμού), τα φωτιστικά σώματα και τις ηλεκτρικές συσκευές που εκλύουν θερμότητα κατά τη λειτουργία τους. Τα θερμικά φορτία ή κέρδη που προκύπτουν σε ένα χώρο εξαρτώνται από τον αριθμό των ατόμων μέσα σε αυτό, το ωράριο παραμονής τους στο χώρο και το είδος της δραστηριότητάς τους. Αντίστοιχα, εξαρτώνται από τη χρήση του κτιρίου και το είδος, αριθμό και χρόνο λειτουργίας όλων των ηλεκτρικών συσκευών.

- **Ηλιακά κέρδη**. Η θερμότητα που συλλέγεται από τον ήλιο κατά τη διάρκεια της ημέρας διαμορφώνει κατά πολύ την εσωτερική θερμοκρασία του κτιρίου. Μπορεί να είναι σε τέτοια ποσότητα, που να προκαλέσει υπερθέρμανση ακόμα και το χειμώνα. Παράλληλα, κατά τη διάρκεια της νύχτας δεν υπάρχει η αντίστοιχη πηγή θερμότητας (ήλιος). Η θερμική μάζα του κτιρίου απορροφά την περίσσεια ενέργεια κατά τη διάρκεια της ημέρας και να την αποδίδει σταδιακά στο χώρο αργότερα, έτσι ώστε να μην υπάρχουν σημαντικές θερμοκρασιακές διακυμάνσεις κατά τη διάρκεια του εικοσιτετραώρου. Η ποσότητα της θερμικής μάζας εξαρτάται από το βάρος και το είδος των δομικών υλικών που βρίσκονται εσωτερικά του κτιρίου, την ποιότητα των μονώσεων αλλά και από τον προσανατολισμό του κτιρίου.

3.3 ΚΡΙΤΗΡΙΑ ΠΟΥ ΑΦΟΡΟΥΝ ΤΗ ΘΕΡΜΟΚΡΑΣΙΑ ΣΤΟ ΕΣΩΤΕΡΙΚΟ ΤΗΣ ΟΙΚΙΑΣ

Με βάση τους παραπάνω παράγοντες που διαμορφώνουν τη θερμοκρασία εντός του κτιρίου, επιλέχθηκαν τα κριτήρια εκείνα που πρέπει να ληφθούν υπόψη και να συγκεντρωθούν ιστορικά δεδομένα ώστε να αρχίσει η επεξεργασία αυτών. Τα κριτήρια θα πρέπει να επιλεχθούν με γνώμονα τη διαθεσιμότητά και τη δυνατότητα συγκέντρωσής δεδομένων για ένα επαρκή χρονικό διάστημα. Τα επιλεχθέντα κριτήρια είναι τα παρακάτω:

I. **Θερμοκρασία περιβάλλοντος** (C^o). Η εξωτερική θερμοκρασία είναι το βασικότερο κριτήριο αφού επηρεάζει κατά το μεγαλύτερο ποσοστό την εσωτερική θερμοκρασία του κτιρίου. Η θερμοκρασία εκφράζεται σε βαθμούς Κελσίου (C^o).

II. **Σχετική υγρασία** (%). Η υγρασία είναι υδρατμοί που περιέχονται στον αέρα. Ανάλογα με το ποσοστό υδρατμών που περιέχεται στον αέρα, η υγρασία επηρεάζει το κέλυφος του κτιρίου. Η υγρασία μπορεί να εισχωρήσει μέσα στα δομικά υλικά του κελύφους του κτιρίου ανάλογα με τη σύσταση και το βαθμό προστασίας τους. Με αυτό τον τρόπο το κέλυφος του κτιρίου διατηρείται υγρό και μειώνεται η θερμική ενέργεια του κτιρίου. Η σχετική υγρασία εκφράζεται επί τοις %.

III. **Ταχύτητα και διεύθυνση του ανέμου**. Κάθε κτίριο επηρεάζεται από τον άνεμο που φυσά γύρω ή πάνω σε αυτό. Από τον άνεμο ασκούνται πιέσεις που εξαρτώνται από το σχήμα του κτιρίου, τον περιβάλλοντα χώρο και τα παρακείμενα κτίρια. Η ταχύτητα και η διεύθυνση του ανέμου σε συνδυασμό με τη θερμοκρασία και την υγρασία του, επηρεάζουν το κέλυφος του κτιρίου και ενδέχεται να μειώσουν τη θερμική ενέργεια του κτιρίου.. Η ταχύτητα του ανέμου εκφράζεται σε μέτρα ανά δευτερόλεπτο (m/s), ενώ η διεύθυνση του ανέμου χαρακτηρίζεται από το σημείο του ορίζοντα από όπου πνέει ο άνεμος και εκφράζεται σε μοίρες που εύκολα μεταφράζονται σε σύμβολα ανεμολογίου.

IV. **Ηλιακή ακτινοβολία** (W/m^2). Η ηλιακή ακτινοβολία εισέρχεται μέσα από τα ανοίγματα του κτιρίου και συλλέγεται από τη μάζα του κτιρίου και αποτελεί μια από της μεγαλύτερες πηγές θερμότητας. Όσο αυξάνεται η ηλιακή ακτινοβολία, τόσο αυξάνονται τα θερμικά φορτία που δέχεται το κτίριο, ανάλογα βέβαια με το σχεδιασμό και τα υλικά κατασκευής και θερμομόνωσης του. Η ηλιακή ακτινοβολία εξαρτάται, επίσης από τον προσανατολισμό του κτιρίου. Για παράδειγμα, τα νότια ανοίγματα και οι νότιες προσόψεις κτιρίων, δέχονται μεγάλα ποσά ηλιακής ακτινοβολίας. Η ηλιακή ακτινοβολία εκφράζεται σε W/m^2.

V. **Εμβαδόν κτιρίου ή διαμερίσματος** (m^2). Το εμβαδόν είναι ένα σημαντικό κριτήριο αφού μας δίνει στοιχεία για το μέγεθος του κτιρίου ή διαμερίσματος. Η ηλιακή ακτινοβολία που προσπέφτει στην οροφή του κτιρίου εξαρτάται από το εμβαδόν του κτιρίου. Επίσης, οι θερμικές απώλειες προς το έδαφος και τον αέρα, εξαρτώνται από το εμβαδόν. Το εμβαδόν του κτιρίου ή διαμερίσματος εκφράζεται σε m^2.

VI. **Τύπος κτιρίου και θέση διαμερίσματος ή άλλου χώρου στο κτίριο**. Οι θερμικές απώλειες εξαρτώνται από τον τύπο του κτιρίου π.χ. μονώροφο ή πολυώροφο κτίριο. Στην περίπτωση του μονώροφου κτιρίου, οι θερμικές απώλειες είναι αυξημένες καθώς το δάπεδο και η οροφή είναι σε άμεση επαφή με το περιβάλλον. Στην περίπτωση του πολυώροφου κτιρίου, οι θερμικές απώλειες είναι μειωμένες για τους ενδιάμεσους ορόφους. Αντίθετα, στους ακραίους ορόφους (υπόγεια, ισόγεια, ρετιρέ), οι θερμικές απώλειες είναι αυξημένες.

VII. **Βαθμοημέρες θέρμανσης**. Η απλούστερη μέθοδος για την εκτίμηση της ενεργειακής κατανάλωσης ενός κτιρίου είναι η μέθοδος των βαθμοημερών (Degree Days), που αφορά τις περιόδους θέρμανσης και ψύξης ενός κτιρίου. Η μέθοδος αυτή χρησιμοποιείται ευρέως για τον υπολογισμό και τη σύγκριση της κατανάλωσης θερμότητας. Όλες οι αντίστοιχες έρευνες, χρησιμοποιούν σχεδόν αποκλειστικά αυτή τη μέθοδο για τον υπολογισμό της απαιτούμενης θερμότητας και ως συνέπεια, για τον υπολογισμό της κατανάλωσης φυσικού αερίου. Πεποίθηση του μελετητή, όμως, είναι ότι η μέθοδος αυτή είναι πολύ γενική και δεν οδηγεί σε ιδιαίτερα ακριβή αποτελέσματα αν χρησιμοποιηθεί μόνη της. Για τη βελτίωση των αποτελεσμάτων, χρησιμοποιούνται τα κριτήρια που επηρεάζουν κατά το μέγιστο την κατανάλωση και η μέθοδος αυτή χρησιμοποιείται ως ένα επιπλέον ισόβαθμο κριτήριο. Η μέθοδος αυτή έχει ένα μεγάλο πλεονέκτημα. Συνδυάζει τη θερμοκρασία που επικρατεί στο εσωτερικό του κτιρίου με τη θερμοκρασία που επιθυμεί ο χρήστης για τη θερμική του άνεση. Η μέθοδος είναι απλή. Το σύστημα θέρμανσης ενεργοποιείται όταν η εσωτερική θερμοκρασία της οικίας είναι χαμηλότερη της επιθυμητής και δεν ενεργοποιείται ή απενεργοποιείται στην αντίθετη περίπτωση. Οι βαθμοημέρες θέρμανσης που μας ενδιαφέρουν, αφορούν τις περιόδους θέρμανσης ενός κτιρίου και υπολογίζονται με απλό τρόπο.

3.4 ΠΑΡΑΓΟΝΤΕΣ ΚΑΙ ΚΡΙΤΗΡΙΑ ΠΟΥ ΑΦΟΡΟΥΝ ΤΗΝ ΤΑΣΗ ΚΑΤΑΝΑΛΩΣΗΣ

Οι χρήστες ενός κτιρίου καθορίζουν τη θερμοκρασία στο εσωτερικό του κτιρίου ώστε να επιτύχουν τη θερμική τους άνεση. Η επιλογή του χρήστη, όμως, για το αν θα ενεργοποιήσει το σύστημα θέρμανσης, δεν βασίζεται μόνο στη θερμοκρασία που επικρατεί στο χώρο αλλά συνδέεται με πολλούς άλλους παράγοντες που έχουν σχέση με τις καταναλωτικές συνήθειες του χρήστη, την επιθυμητή εσωτερική θερμοκρασία και άλλα.

Η τάση κατανάλωσης φυσικού αερίου επηρεάζεται, εκτός από το προφίλ του καταναλωτή, από το μέγεθος και τη χρήση του κτιρίου ή χώρου, όπως επίσης και από τη χρήση που προορίζεται να έχει το φυσικό αέριο.

- **Εφαρμογές φυσικού αερίου**. Η κατανάλωση εξαρτάται σε πολύ μεγάλο βαθμό από τις εφαρμογές που έχει (κεντρική ή ατομική θέρμανση, θέρμανση νερού, μαγείρεμα). Η κατανάλωση βέβαια, όσον αφορά τους οικιακούς καταναλωτές, δεν διαφοροποιείται ιδιαίτερα από με την πρόσθεση υπηρεσιών όπως η θέρμανση νερού και το μαγείρεμα.

- **Χρήση φυσικού αερίου** (οικιακή, επαγγελματική). Για τον προσδιορισμό της ζήτησης του φυσικού αερίου πολύ σημαντικός παράγοντας είναι το είδος της χρήσης του φυσικού αερίου. Όπως είναι λογικό η κατανάλωση από τους επαγγελματικούς καταναλωτές είναι σημαντικά μεγαλύτερη των οικιακών και πολύ πιο σταθερή χρονικά.

- **Τύπος καταναλωτή**. Ο τύπος καταναλωτή δηλαδή αν είναι ιδιώτης, επαγγελματίας, επιχείρηση έχει πολύ μεγάλη σημασία αφού ανάλογα με τον τύπο του μπορεί να γίνει πιο ακριβής η πρόβλεψη.

- **Ιστορικά στοιχεία κατανάλωσης**. Χρήσιμα στοιχεία για τον υπολογισμό της κατανάλωσης αποτελούν τα ιστορικά στοιχεία κατανάλωσης αφού μπορούμε να έχουμε μια πρώτης τάξης ιδέα του καταναλωτικού προφίλ του κάθε καταναλωτή.

- **Αιτούμενη εγκατεστημένη ισχύς.** Η εγκατεστημένη ισχύς φυσικού αερίου προσδιορίζει το μέγιστο ωριαίο φορτίο κατανάλωσης και από αυτό το στοιχείο μπορούμε να έχουμε μια ιδέα των ενεργειακών απαιτήσεων του κτηρίου. Η εγκατεστημένη ισχύς υπολογίζεται από το μηχανικό που σχεδιάζει το κτήριο οπότε πρόκειται για μια αρκετά αξιόπιστη μελέτη. Ο παράγοντας αυτός σε συνδυασμό με άλλα στοιχεία μπορεί να οδηγήσει στη διεξαγωγή αποτελεσμάτων με ξεχωριστό ενδιαφέρον.

3.5 ΜΗ ΕΠΙΛΕΧΘΕΝΤΑ ΚΡΙΤΗΡΙΑ.

Παρόλο που υπάρχουν και άλλοι παράγοντες που διαμορφώνουν τη θερμοκρασία εντός του κτιρίου και θα μπορούσαν να χρησιμοποιηθούν σαν κριτήρια, τελικά δεν επιλέχθηκαν κυρίως λόγω της δυσκολίας ανεύρεσης δεδομένων αλλά και επειδή η ακρίβεια που θα προσέφεραν αυτά τα κριτήρια στο επιδιωκόμενο αποτέλεσμα θα ήταν ανεπαίσθητη.

- **Προσανατολισμός κτιρίου ή διαμερίσματος**. Ανάλογα με τον κύριο προσανατολισμό του κτιρίου, κυμαίνεται και η προσπίπτουσα ηλιακή ακτινοβολία που όσο αυξάνεται, τόσο αυξάνονται τα θερμικά φορτία που δέχεται το κτίριο. Για παράδειγμα, ο βόρειος προσανατολισμός δέχεται ελάχιστη ηλιακή πρόσπτωση το πρωί και το βράδυ. Αντίθετα, τα ανατολικά και δυτικά ανοίγματα δέχονται μεγάλα ποσά ηλιακής ακτινοβολίας το καλοκαίρι (ενώ το χειμώνα πολύ μικρά). Παρόλο που ο παράγοντας αυτός έχει μεγάλη σημασία και θα μπορούσε να χρησιμοποιηθεί σαν κριτήριο, τελικά δεν επιλέχθηκε γιατί είναι εξαιρετικά δύσκολο να συγκεντρωθούν δεδομένα. Η μόνη πηγή αυτών των δεδομένων, θα ήταν η αίτηση σύνδεσης του καταναλωτή αλλά ένα τέτοιο στοιχείο δεν θα μπορούσε να συμπεριληφθεί στην αίτηση. Ο καταναλωτής πολλές φορές δεν γνωρίζει τον προσανατολισμό του κτιρίου και δεν μπορεί αυτό να απαιτηθεί. Είναι μια πολύ ειδική πληροφορία, άγνωστη από πολλούς καταναλωτές. Επίσης, η ακρίβεια που θα προσέφερε αυτό το κριτήριο στο αποτέλεσμα θα ήταν ανεπαίσθητη. Οπότε το κριτήριο δεν επιλέχθηκε, λόγω δυσκολίας συλλογής δεδομένων και χαμηλής επίδρασης στο αποτέλεσμα.

- **Αριθμός ανοιγμάτων οικίας.** Όσο ποιο πολλά ανοίγματα έχει ένα κτίριο (δηλαδή θύρες, παράθυρα) και σε συνδυασμό με το υλικό της κατασκευής τους, τόσο ποιο πολύ αυξάνονται οι απώλειες θερμότητας λόγω της μετάδοσης της προς το εξωτερικό περιβάλλον. Παρόλο που ο παράγοντας αυτός έχει κάποια σημασία και θα μπορούσε να χρησιμοποιηθεί σαν κριτήριο, τελικά δεν επιλέχθηκε γιατί είναι εξαιρετικά δύσκολο να συγκεντρωθούν δεδομένα. Η μόνη πηγή αυτών των δεδομένων, θα ήταν η αίτηση σύνδεσης του καταναλωτή αλλά ένα τέτοιο στοιχείο δεν θα μπορούσε να συμπεριληφθεί στην αίτηση. Ο καταναλωτής πολλές φορές δεν γνωρίζει τον ακριβή αριθμό ανοιγμάτων του κτιρίου ούτε τις διαστάσεις τους ή τα υλικά κατασκευής τους και δεν μπορεί αυτό να απαιτηθεί. Είναι μια πολύ ειδική πληροφορία, άγνωστη από πολλούς καταναλωτές. Επίσης, η ακρίβεια που θα προσέφερε αυτό το κριτήριο στο αποτέλεσμα θα ήταν ανεπαίσθητη. Οπότε το κριτήριο δεν επιλέχθηκε, λόγω δυσκολίας συλλογής δεδομένων και χαμηλής επίδρασης στο αποτέλεσμα.

- **Ποιότητα μονώσεων**. Οι απώλειες λόγω μετάδοσης θερμότητας και εξαερισμού στο εξωτερικό περιβάλλον είναι σημαντικές για τη θέρμανση ενός κτιρίου. Οι απώλειες αυτές εξαρτώνται από την ποιότητα της θερμομόνωσης του κελύφους του κτιρίου. Οι πληροφορίες, που θα μπορούσαν να εξαχθούν από τη συλλογή και επεξεργασία δεδομένων για αυτό το κριτήριο, θα ήταν πολύ σημαντικές. Θα φανερωνόταν πόσο επηρεάζεται η κατανάλωση φυσικού αερίου για θέρμανση, από την ποιότητα της θερμομόνωσης του κτιρίου. Θα έπρεπε να γίνει συλλογή δεδομένων για όλους τους υπάρχοντες χρήστες φυσικού αερίου και να βρεθεί τρόπος ανάκτησης για τους νέους χρήστες. Επειδή το κριτήριο είναι ποιοτικό, είναι πολύ δύσκολο να συγκεντρωθούν αυτά τα δεδομένα. Ο μόνος τρόπος θα ήταν να συμπεριληφθεί στην αίτηση, ερώτηση που θα ζητούσε από τον καταναλωτή να κατατάξει το κτίριο σε κάποιες κατηγορίες με κριτήριο την παλαιότητά του (για παράδειγμα χρονολογία κατασκευής). Από τη χρονολογία κατασκευής του κτιρίου σε συνδυασμό με τους τότε ισχύοντες κανονισμούς κατασκευής, θα μπορούσε να προκύψει μια ικανοποιητική περιγραφή των μονώσεων του. Η μεθοδολογία αυτή είναι βέβαια ανέφικτη καθώς δεν εξάγει ακριβή στοιχεία, πολύ καταναλωτές δεν γνωρίζουν την χρονολογία κατασκευής του κτιρίου που κατοικούν είτε είναι ιδιοκτήτες, είτε όχι, αλλά το κυριότερο είναι ότι τα κτίρια πολλές φορές δεν κατασκευάζονταν με βάση τις προδιαγραφές που οφείλουν να τηρήσουν για ποικίλους λόγους. Οπότε το κριτήριο δεν επιλέχθηκε, κυρίως λόγω της δυσκολίας συλλογής δεδομένων.

- **Άλλοι παράγοντες.** Είναι φυσικό πως υπάρχουν πολλοί παράγοντες που παρουσιάζουν ξεχωριστό ενδιαφέρον και σχετίζονται με την κατανάλωση φυσικού αερίου. Για παράδειγμα, πολλοί βιομηχανικοί και εμπορικοί πελάτες διακόπτουν τη λειτουργία των επιχειρήσεων τους κατά τη διάρκεια των Σαββατοκύριακων ή του καλοκαιριού όπως επίσης παρατηρούνται αυξομειώσεις της κατανάλωσης λόγω γιορτών και αργιών και κατά συνέπεια η ημέρα της εβδομάδας θεωρείται παράγοντας επίσης. Στην περίπτωση των κατοικιών μια παρόμοια παράμετρος είναι το αν η κατοικία χρησιμοποιείται ως μόνιμη ή ως μη μόνιμη. Μια άλλη σχετική παράμετρος είναι η πρόσφατη τάση κατανάλωσης, η οποία αντιπροσωπεύει την γενική απαίτηση για φυσικό αέριο. Η οικονομική ικανότητα των καταναλωτών επηρεάζει σημαντικά την κατανάλωση δεδομένου ότι οι καταναλωτές επιζητούν μια οικονομική μορφή ενέργειας ώστε το κόστος των ενεργειακών τους αναγκών να παραμένει όσο το δυνατόν χαμηλότερο. Επομένως οι τιμές του φυσικού αερίου σε σχέση με τις ανταγωνιστικές τιμές καυσίμων (πετρέλαιο θέρμανσης, ηλεκτρικό ρεύμα) και ο δείκτης οικονομικής ανάπτυξης αποτελούν παράγοντες που επηρεάζουν την ζήτηση φυσικού αερίου.

ΚΕΦΑΛΑΙΟ 4: ΕΠΙΛΟΓΗ ΚΑΙ ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ

4.1 ΕΙΣΑΓΩΓΗ

Αφού αναλύθηκαν οι παράγοντες που επηρεάζουν την κατανάλωση φυσικού αερίου και καθορίστηκαν τα κριτήρια, επόμενο βήμα είναι η συγκέντρωση των απαιτούμενων δεδομένων. Δεδομένα τύπου: χρήσεις και εφαρμογές φυσικού αερίου, ιστορικά στοιχεία κατανάλωσης, τύπος καταναλωτή και αιτούμενη εγκατεστημένη ισχύς υπήρχαν σε βάση δεδομένων της εταιρείας διανομής Φυσικό Αέριο Αττικής. Τα υπόλοιπα δεδομένα όμως έπρεπε να συγκεντρωθούν από άλλες πηγές.

Θερμοκρασία περιβάλλοντος	Δεδομένα που συγκεντρώθηκαν, ύστερα από επεξεργασία, από μετεωρολογικούς σταθμούς
Σχετική υγρασία	
Ταχύτητα του ανέμου	
Διεύθυνση του ανέμου	
Ηλιακή ακτινοβολία	
Βαθμοημέρες θέρμανσης	
Εμβαδόν κτιρίου ή διαμερίσματος	Δεδομένα που η συγκέντρωση τους δεν ήταν εφικτή.
Τύπος κτιρίου και θέση διαμερίσματος ή άλλου χώρου στο κτίριο	
Κωδικός καταναλωτή	Δεδομένα που αποκτήθηκαν από βάση δεδομένων της εταιρείας διανομής.
Ταχυδρομικός κώδικας καταναλωτή	
Κατηγορία καταναλωτή	
Εφαρμογές & Χρήση φυσικού αερίου	
Ιστορικά στοιχεία κατανάλωσης	
Αιτούμενη εγκατεστημένη ισχύς	

4.2 ΔΕΔΟΜΕΝΑ ΠΟΥ Η ΣΥΓΚΕΝΤΡΩΣΗ ΤΟΥΣ ΕΙΝΑΙ ΑΝΕΦΙΚΤΗ.

Δεν κατέστει δυνατόν να συγκεντρωθούν δεδομένα για κάποια κριτήρια. Τα δεδομένα που δεν συγκεντρώθηκαν αφορούν τα κριτήρια: 'Εμβαδόν κτιρίου ή διαμερίσματος' και 'Τύπος κτιρίου και θέση διαμερίσματος ή άλλου χώρου στο κτίριο'. Τα δεδομένα αυτά δεν υπήρχαν μέσα στη βάση δεδομένων της εταιρείας διανομής και δεν συλλέγονταν γιατί είναι σχετικά δύσκολη η συγκέντρωσή τους. Απαιτεί να αναδιαμορφωθεί η αίτηση που κάνει ο καταναλωτής στην εταιρεία διανομής ώστε να επιτρέπεται η συλλογή αυτού του τύπου των δεδομένων. Ακόμα και αν το πράξει μελλοντικά η εταιρεία, δεν θα υπάρχουν αυτά τα στοιχεία για όλους τους παλιούς καταναλωτές. Επομένως, παρόλο που ενδεχομένως αυτά τα δύο κριτήρια θα αύξαναν την ακρίβεια της πρόβλεψης, η συγκέντρωσή τους και η αξιοποίησή τους ήταν αδύνατη.

4.3 ΔΕΔΟΜΕΝΑ ΠΟΥ ΑΠΟΚΤΗΘΗΚΑΝ ΑΠΟ ΜΕΤΕΩΡΟΛΟΓΙΚΟΥΣ ΣΤΑΘΜΟΥΣ

Τα μετεωρολογικά κριτήρια προκύπτουν από τους παράγοντες που διαμορφώνουν τη θερμοκρασία στο εσωτερικό ενός κτηρίου. Παρόλο που ο πιο σημαντικός παράγοντας είναι, όπως είναι φυσικό, η θερμοκρασία, σε μορφή βαθμοημερών θέρμανσης, επιλέγουμε να συγκεντρωθούν δεδομένα για κάθε μετεωρολογικό κριτήριο.

Τα κριτήρια για τα οποία πρέπει να συλλεχθούν δεδομένα είναι τα εξής:

1. Θερμοκρασία περιβάλλοντος (C^o)
2. Σχετική υγρασία (%)
3. Ταχύτητα ανέμου
4. Διεύθυνση ανέμου
5. Ηλιακή ακτινοβολία (W/m^2)
6. Βαθμοημέρες θέρμανσης (προκύπτουν από τη Θερμοκρασία Περιβάλλοντος)

Καθώς ο χώρος μελέτης μας είναι το σύνολο της Αττικής, κρίθηκε σκόπιμη η συλλογή δεδομένων από διαφορετικά σημεία της, έτσι ώστε τα αποτελέσματα της πρόβλεψης μας να είναι πιο ακριβή. Σημαντικός παράγοντας που διαφοροποιεί τις μετεωρολογικές συνθήκες από περιοχή σε περιοχή, είναι η γεωγραφική θέση της Αττικής. Η Αττική περικλείεται τόσο από βουνά (Πεντέλη, Υμηττός, Πάρνηθα) όσο και από θάλασσα με αποτέλεσμα να παρατηρούνται μεγάλες μετεωρολογικές αποκλείσεις. Για παράδειγμα μια περιοχή κοντά σε θάλασσα (π.χ. Πειραιάς) έχει αυξημένη υγρασία και παρουσιάζει διαφορετική κλιματολογική συμπεριφορά σε σχέση με μια πιο ορεινή περιοχή (π.χ. Πεντέλη) δηλαδή περιοχές που βρίσκονται σε μεγάλο υψόμετρο ή περιοχές που βρίσκονται μακριά από θάλασσα. Επίσης, ένας ακόμα παράγοντας είναι η πυκνότητα της συγκέντρωσης των κτηρίων σε μια περιοχή.

Σε πυκνοκατοικημένες περιοχές είναι λογικό να παρατηρείται αύξηση της θερμοκρασίας λόγω της γειτνίασης των κτηρίων και της μείωσης της διείσδυσης του ανέμου με αποτέλεσμα να μειώνονται οι θερμικές απώλειες των κτηρίων. Επομένως, οι μετεωρολογικές συνθήκες διαφέρουν πολύ από περιοχή σε περιοχή μέσα στην περιοχή της Αττικής.

Στον παρακάτω πίνακα φαίνεται η μέγιστη θερμοκρασιακή απόκλιση που παρατηρείται ανάμεσα σε δύο περιοχές της Αττικής για τις 03 Μαΐου 2005.

	Θερμοκρασία	Υγρασία	Ταχύτητα Ανέμου	Διεύθυνση Ανέμου	Ηλιακή Ακτινοβολία	Βαθμοημέρες Θέρμανσης
Ελάχιστη τιμή	16,36	320,10	1,96	227,97	50,65	1,96
Μέγιστη τιμή	20,04	324,75	2,05	218,07	50,93	0

Όπως παρατηρούμε υπάρχει, για τη συγκεκριμένη ημέρα, απόκλιση 3,68 βαθμών κελσίου. Διαφορά αρκετή ώστε σε μια περιοχή της Αττικής να ενεργοποιήσει ένας καταναλωτής το σύστημα θέρμανσης ενώ σε άλλη όχι. Αυτό φαίνεται επίσης και από τις βαθμοημέρες θέρμανσης. Στην περιοχή της Αττικής που έχει 20 βαθμούς κελσίου δεν χρειάζεται να σπαταληθεί ενέργεια για θέρμανση ενώ στην άλλη περιοχή απαιτείται ενέργεια τόση ώστε να καλύψει τη θερμοκρασιακή διαφορά των 2 βαθμών κελσίου που απαιτείται για την θερμική άνεση του καταναλωτή. Άρα δεν μπορούμε να χρησιμοποιήσουμε τα ίδια κλιματολογικά δεδομένα για όλες τις περιοχές της Αττικής.

Είναι σκόπιμη επομένως, η συλλογή δεδομένων για διαφορετικές περιοχές της Αττικής. Η αρχική προσέγγιση ήταν η αναζήτηση ισοκλιματικών ζωνών της Αττικής δηλαδή ζωνών που περιλαμβάνουν περιοχές που παρουσιάζουν ίδιο κλιματολογικό προφίλ. Μετά από εκτενή έρευνα στο Πανεπιστήμιο Αθηνών και το Εθνικό Μετσόβειο Πολυτεχνείο δεν κατέστει δυνατόν να βρεθούν έρευνες με αυτό το αντικείμενο καθώς όπως διαπιστώσαμε και πληροφορηθήκαμε, δεν έχει εκπονηθεί καμία αντίστοιχη μελέτη για την Αττική.

Ο μόνος τρόπος για την εύρεση έγκυρων κλιματολογικών δεδομένων είναι η συγκέντρωση δεδομένων από μετεωρολογικούς σταθμούς της Αττικής. Τα δεδομένα πρέπει να προέρχονται από μετεωρολογικούς σταθμούς δημόσιων υπηρεσιών ώστε να είναι επίσημα, συνεχόμενα (χωρίς κενά ή κενές εγγραφές), ακριβή και να διατίθενται για την χρονική περίοδο που μας ενδιαφέρει. Έπειτα από αναζήτηση υπηρεσιών και μετεωρολογικών σταθμών του ενδιαφέροντός μας, βρέθηκαν οι ακόλουθοι.

	Περιοχή		**Φορέας**
1	Αττική	Ελληνικό	Ε.Μ.Υ.
2	Αττική	Ν. Φιλαδέλφεια	Ε.Μ.Υ.
3	Αττική	Άνω Λιόσια	Ε.Μ.Υ
4	Αττική	Τατόι	Ε.Μ.Υ.
5	Αττική	Πειραιάς	Ε.Μ.Υ.
6	Αττική	Πεντέλη	Ε.Α.Α.
7	Αττική	Θησείο	Ε.Α.Α.
8	Αττική	Ζωγράφου	ΜΕΤΕΟΝΕΤ
9	Αττική	Μενίδι	ΜΕΤΕΟΝΕΤ
10	Αττική	Ψυτάλλεια	ΜΕΤΕΟΝΕΤ
11	Αττική	Ηλιούπολη	ΜΕΤΕΟΝΕΤ
12	Αττική	Γαλάτσι	ΜΕΤΕΟΝΕΤ

Ε.Μ.Υ. : Εθνική Μετεωρολογική Υπηρεσία
Ε.Α.Α. : Εθνικό Αστεροσκοπείο Αθηνών
ΜΕΤΕΟΝΕΤ : Μετεωρολογική Υπηρεσία τμήματος Πολιτικών Μηχανικών Ε.Μ.Π.

Έπειτα από επικοινωνία με τις προαναφερθείσες υπηρεσίες και αίτηση παροχής δεδομένων με τη βοήθεια του επιβλέποντος καθηγητή, αποκτήθηκαν τα μετεωρολογικά δεδομένα ανά ημέρα, για τους παράγοντες που μας αφορούν. Τα δεδομένα επιλέχθηκε να συγκεντρωθούν ανά ημέρα καθώς με αυτό τον τρόπο θα ήταν εφικτός ο υπολογισμός των βαθμοημερών θέρμανσης. Η καταγεγραμμένες καταναλώσεις φυσικού αερίου από την εταιρεία διανομής αφορούν διαφορετικές χρονικές περιόδους με διαφορετική ημερομηνία έναρξης της καταμέτρησης. Οπότε είναι επιβεβλημένη η συγκέντρωση των δεδομένων ανά ημέρα ώστε να είναι εύκολοι οι υπολογισμοί. Επομένως, αν γνωρίζουμε τις ημερομηνίες που αναφέρεται η κατανάλωση, μπορούμε να αθροίσουμε τις βαθμοημέρες για κάθε μέρα τις χρονικής περιόδου και να βρούμε τις συνολικές βαθμοημέρες θέρμανσης.

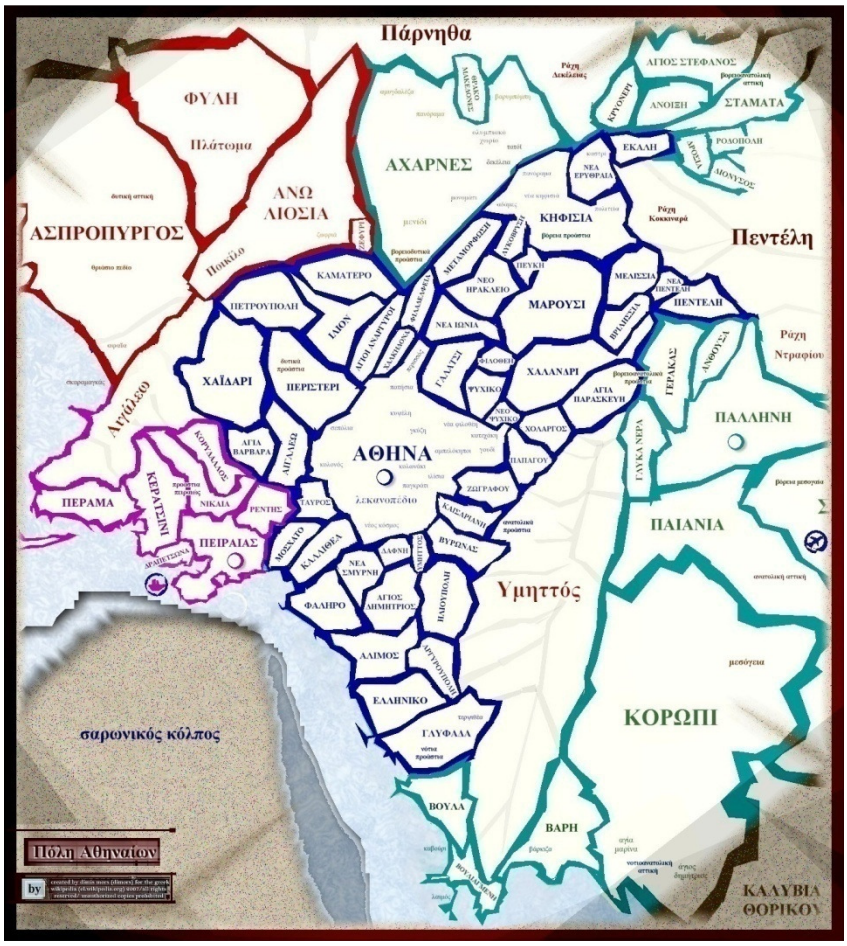
Η διάθεση των δεδομένων ήταν δωρεάν και αφορούσε την χρονική περίοδο από 01 Ιανουαρίου 2005 έως 30 Μαρτίου 2008. Επιλέχθηκε η συγκεκριμένη περίοδος καθώς για αυτήν υπάρχουν όλα τα απαιτούμενα δεδομένα κατανάλωσης φυσικού αερίου από την εταιρεία διανομής, είναι αρκετά εκτενής (3,2 έτη), τα κλιματολογικά δεδομένα είναι συνεχόμενα (χωρίς κενά ή κενές εγγραφές) και ακριβή. Μετά τη συλλογή των κλιματολογικών δεδομένων για αυτούς τους μετεωρολογικούς σταθμούς, παρατηρήθηκε ότι η κάλυψη τους δεν εξασφαλίζει μεγάλη ακρίβεια. Οι σταθμοί απέχουν πολύ μεταξύ τους και δεν καλύπτουν όλες τις περιοχές της Αττικής με αποτέλεσμα και πάλι να παρατηρούνται μεγάλες θερμοκρασιακές αποκλείσεις ανάμεσα στα όρια κάλυψης των μετεωρολογικών σταθμών.

Στην παραπάνω φωτογραφία παρουσιάζεται η γεωγραφική θέση κάθε μετεωρολογικού σταθμού στην περιοχή της Αττικής. Παρατηρούμε ότι η κεντρική νοτιοδυτική περιοχή του λεκανοπεδίου καλύπτεται μόνο από έναν σταθμό, αυτόν του Θησείου. Η θέση των μετεωρολογικών σταθμών δεν είναι η κατάλληλη για την ακριβή συλλογή μετεωρολογικών δεδομένων.

Ένα άλλο πρόβλημα είναι ότι ελλείψει ορισμού ισοκλιματικών ζωνών, δεν είναι δυνατόν να οριστούν ποιες περιοχές καλύπτει κάθε μετεωρολογικός σταθμός, δηλαδή σε ποιες περιοχές ισχύουν οι ίδιες κλιματολογικές συνθήκες όπως αυτές που ισχύουν για τον σταθμό καταγραφής μετεωρολογικών δεδομένων.

Καθότι δεν υπήρχε κάποιος τρόπος εύκολης συλλογής κλιματολογικών δεδομένων για ισοκλιματικές περιοχές της Αττικής, κρίθηκε απαραίτητη η αναζήτηση μιας μεθόδου που θα χρησιμοποιούσε δεδομένα από διαφορετικούς μετεωρολογικούς σταθμούς με εκροή μια συνισταμένη κλιματολογικών συνθηκών που θα όριζαν κάθε ορισμένη κλιματολογική περιοχή.

Αρχικά, ερευνήθηκε το κατά πόσο θα μπορούσαν να χρησιμοποιηθούν οι δήμοι και τα όριά τους, για την εφαρμογή της μεθόδου. Έπειτα από έρευνα παρατηρήθηκε ότι ακόμα και μέσα στα όρια των δήμων της Αττικής υπάρχουν αρκετά μεγάλες κλιματολογικές αποκλίσεις. Αυτό συμβαίνει γιατί κάποιοι δήμοι καλύπτουν μεγάλη έκταση με αποτέλεσμα να υπάρχουν διακυμάνσεις στο υψόμετρο ή στην πυκνότητα της συγκέντρωσης των κτηρίων.



Στην απεικόνιση της προηγούμενης σελίδας παρουσιάζονται οι δήμοι της Αττικής. Είναι προφανές ότι δεν μπορούν να χρησιμοποιηθούν τα όρια των δήμων καθώς, όπως φαίνεται παραπάνω, υπάρχουν δήμοι (π.χ. Δήμος Αθηναίων, Αχαρνών κ.α.) που είναι μεγάλοι σε έκταση και που στα όριά τους παρουσιάζονται αρκετά μεγάλες κλιματολογικές αποκλίσεις.

Στον παρακάτω πίνακα παρουσιάζεται η μέγιστη θερμοκρασιακή απόκλιση που παρατηρείται ανάμεσα στον ίδιο δήμο της Αττικής (Δήμος Αθηναίων) για τις 03 Μαΐου 2005.

	Θερμοκρασία	Υγρασία	Ταχύτητα Ανέμου	Διεύθυνση Ανέμου	Ηλιακή Ακτινοβολία	Βαθμοημέρες Θέρμανσης
Ελάχιστη τιμή	17,14	320,67	1,96	219,97	50,89	0,86
Μέγιστη τιμή	18,56	324,15	2,05	218,07	51,01	0

Όπως παρατηρούμε από τον πίνακα, υπάρχει, για τη συγκεκριμένη ημέρα, απόκλιση 1,42 βαθμών κελσίου. Διαφορά αρκετή ώστε μέσα στον ίδιο δήμο να ενεργοποιήσει ένας καταναλωτής το σύστημα θέρμανσης ενώ σε άλλη όχι. Αυτό φαίνεται επίσης και από τις βαθμοημέρες θέρμανσης. Στην περιοχή του Δήμου Αθηναίων που έχει 18,56 βαθμούς κελσίου ίσως να μην χρειαστεί να σπαταληθεί ενέργεια για θέρμανση ενώ στην άλλη περιοχή απαιτείται ενέργεια τόση ώστε να καλυφθεί η θερμοκρασιακή διαφορά των 0,56 βαθμών κελσίου που απαιτείται για την θερμική άνεση του καταναλωτή. Άρα δεν μπορούμε να χρησιμοποιήσουμε τα ίδια κλιματολογικά δεδομένα για το σύνολο του κάθε δήμου.

Μετά από αναζήτηση άλλης ομαδοποίησης περιοχών που παρουσιάζουν κοινή κλιματολογική συμπεριφορά, καταλήξαμε στους ταχυδρομικούς κώδικες. Τα Ελληνικά Ταχυδρομεία (ΕΛ.ΤΑ.) έχουν ορίσει διαφορετικούς ταχυδρομικούς κωδικούς ακόμα και ανάμεσα στα όρια των εκάστοτε δήμων. Δηλαδή υπάρχουν δήμοι που είναι χωρισμένοι ακόμα και σε 8 ταχυδρομικούς κώδικες για την καλύτερη οργάνωση και εξυπηρέτηση των καταναλωτών και τη διευκόλυνση των ταχυδρομικών υπηρεσιών. Για παράδειγμα ο Δήμος Χαλανδρίου είναι χωρισμένος σε 4 διαφορετικούς ταχυδρομικούς κώδικες. Η ομαδοποίηση αυτή κρίθηκε η καταλληλότερη για την εφαρμογή της επιλεγμένης μεθόδου για την εύρεση της συνισταμένης κλιματολογικής συμπεριφοράς.

Οι ταχυδρομικοί κώδικες που μας ενδιαφέρουν είναι 242 και έχουν επιλεχθεί αυτοί για τους οποίους υπάρχει δίκτυο και καταναλωτές φυσικού αερίου. Για τους ταχυδρομικούς κώδικες αυτούς βρέθηκαν επίσης οι συντεταγμένες τους από τα Ελληνικά Ταχυδρομεία. Οι συντεταγμένες απευθύνονται στο κεντρικότερο σημείο της περιοχής που καλύπτει ο κάθε ταχυδρομικός κώδικας.

Παρακάτω παρουσιάζονται όλοι οι ταχυδρομικοί κώδικες που επιλέχθηκαν για την περιοχή της Αττικής.

10431, 10432, 10433, 10434, 10435, 10436, 10437, 10438, 10439, 10440, 10441, 10442, 10443, 10444, 10445, 10446, 10447, 10551, 10552, 10553, 10554, 10555, 10556, 10557, 10558, 10559, 10560, 10561, 10562, 10563, 10564, 10671, 10672, 10673, 10674, 10675, 10676, 10677, 10678, 10679, 10680, 10681, 10682, 10683, 11141, 11142, 11143, 11144, 11145, 11146, 11147, 11251, 11252, 11253, 11254, 11255, 11256, 11257, 11361, 11362, 11363, 11364, 11471, 11472, 11473, 11474, 11475, 11476, 11521, 11522, 11523, 11524, 11525, 11526, 11527, 11528, 11631, 11632, 11633, 11634, 11635, 11636, 11676, 11741, 11742, 11743, 11744, 11745, 11851, 11852, 11853, 11854, 11855, 12131, 12132, 12133, 12134, 12135, 12136, 12137, 12241, 12242, 12243, 12244, 12351, 12461, 12462, 13121, 13122, 13123, 13231, 13451, 13561, 13562, 13671, 13672, 13675, 14121, 14122, 14123, 14162, 14231, 14232, 14233, 14234, 14235, 14341, 14342, 14343, 14451, 14452, 14561, 14562, 14563, 14564, 14671, 15121, 15122, 15123, 15124, 15125, 15126, 15127, 15231, 15232, 15233, 15234, 15235, 15237, 15341, 15342, 15343, 15451, 15452, 15561, 15562, 15662, 15669, 15771, 15772, 15773, 16121, 16122, 16231, 16232, 16233, 16341, 16342, 16343, 16344, 16345, 16346, 16451, 16452, 16561, 16562, 16673, 16674, 16675, 16777, 17121, 17122, 17123, 17124, 17234, 17235, 17236, 17237, 17341, 17342, 17343, 17455, 17456, 17561, 17562, 17563, 17564, 17671, 17672, 17673, 17674, 17675, 17676, 17778, 18120, 18121, 18122, 18153, 18233, 18344, 18345, 18346, 18450, 18451, 18452, 18453, 18454, 18531, 18532, 18533, 18534, 18535, 18536, 18537, 18538, 18539, 18540, 18541, 18542, 18543, 18544, 18545, 18546, 18547, 18548, 18648, 18755, 18756, 18757, 18758, 18863, 99999.

Οπότε αντί των 78 δήμων της Αττικής, επιλέγουμε να χρησιμοποιήσουμε 242 ταχυδρομικούς κώδικες. Με αυτή την ομαδοποίηση η ακρίβεια που θα επιτευχθεί θα είναι πολύ μεγαλύτερη και πιο αντιπροσωπευτική καθώς η έκταση των περιοχών θα είναι πιο μικρή με αποτέλεσμα να παρουσιάζεται κοινή κλιματολογική συμπεριφορά εντός της ίδιας περιοχής.

Έπειτα από συνεργασία και προτάσεις της υπηρεσίας ΜΕΤΕΟΝΕΤ του τμήματος Πολιτικών Μηχανικών του Εθνικού Μετσόβιου Πολυτεχνείου, επιλέχθηκε η καταλληλότερη μέθοδος για την εύρεση των συνισταμένων κλιματολογικών δεδομένων έπειτα από συγκερασμό των δεδομένων των 12 επιλεγμένων μετεωρολογικών σταθμών. Η μέθοδος που επιλέχθηκε, υπολογίζει από τα 3 κοντινότερα σημεία για τα οποία διατίθενται μετεωρολογικά στοιχεία, το μετεωρολογικά δεδομένα για το σημείο που ψάχνουμε. Αυτό γίνεται υπολογίζοντας τις αποστάσεις των 3 σημείων και ύστερα, την απόσταση εκφρασμένη αντιστρόφως ανάλογα, επί τις εκατό. Έπειτα, υπολογίζονται τα μετεωρολογικά δεδομένα πολλαπλασιάζοντας κάθε μετεωρολογικό στοιχείο του συγκεκριμένου σημείου με το ποσοστό που αντιστοιχεί στο ανάλογο μετεωρολογικό σημείο.

Για την εφαρμογή της μεθόδου βρέθηκαν οι συντεταγμένες των επιλεχθέντων μετεωρολογικών σταθμών και των ταχυδρομικών κωδικών της Αττικής.

	Περιοχή		Φορέας	Υψόμετρο	Συντεταγμένες
1	Αττική	Ελληνικό	Ε.Μ.Υ.	15μ	37°53'59", 23°43'58"
2	Αττική	Ν. Φιλαδέλφεια	Ε.Μ.Υ.	138μ	38°02'59", 23°40'01"
3	Αττική	Άνω Λιόσια	Ε.Μ.Υ	258μ	38°07'22", 23°33'49"
4	Αττική	Τατόι	Ε.Μ.Υ.	235μ	38°06'00", 23°46'58"
5	Αττική	Πειραιάς	Ε.Μ.Υ.	2,1μ	37°56'00", 23°38'00"
6	Αττική	Πεντέλη	Ε.Α.Α.	509μ	38°02'55", 23°51'53"
7	Αττική	Θησείο	Ε.Α.Α.	107μ	37°58'23", 23°43'06"
8	Αττική	Ζωγράφου	ΜΕΤΕΟΝΕΤ	181μ	37°58'37", 23°47'12"
9	Αττική	Μενίδι	ΜΕΤΕΟΝΕΤ	210μ	38°06'23", 23°44'02"
10	Αττική	Ψυτάλλεια	ΜΕΤΕΟΝΕΤ	20μ	37°56'31", 23°35'13"
11	Αττική	Ηλιούπολη	ΜΕΤΕΟΝΕΤ	206μ	37°55'05", 23°45'39"
12	Αττική	Γαλάτσι	ΜΕΤΕΟΝΕΤ	176μ	38°01'45", 23°45'26"

Οι συντεταγμένες των μετεωρολογικών σταθμών και των ταχυδρομικών κωδικών εκφράστηκαν σε ακτίνια ανά γεωγραφικό πλάτος και μήκος. Ο τρόπος υπολογισμού των ακτινίων είναι ο εξής. Χρησιμοποιούμε την συνάρτηση του Microsoft Excel, Radians και μέσα στη συνάρτηση αθροίζουμε το γινόμενο των μοιρών με 3600, των λεπτών με 60 και τα δευτερόλεπτα για τα γεωγραφικά πλάτη και μήκη ξεχωριστά των συντεταγμένων των σημείων που βρίσκονται οι 12 μετεωρολογικοί σταθμοί και οι 242 ταχυδρομικοί κώδικες.

C1	D1	E1	F1	G1	H1
Lat			Lon		
37	53	59	23	43	58

=RADIANS((C1*3600+D1*60+E1)/3600)
=RADIANS((F1*3600+G1*60+H1)/3600)

Μετά υπολογίζουμε την απόσταση (σε χιλιόμετρα) του κάθε ταχυδρομικού κώδικα από κάθε μετεωρολογικό σταθμό. Αυτό γίνεται με την παρακάτω συνάρτηση.

=(ACOS(SIN(Y2)*SIN(I2)+COS(Y2)*COS(I2)*COS(J2-Z2))*180/PI()*60)*1,854

I2	Γεωγραφικό πλάτος μετεωρολογικού σταθμού (σε ακτίνια)
J2	Γεωγραφικό μήκος μετεωρολογικού σταθμού (σε ακτίνια)
Y2	Γεωγραφικό πλάτος ταχυδρομικού κώδικα (σε ακτίνια)
Z2	Γεωγραφικό μήκος ταχυδρομικού κώδικα (σε ακτίνια)

Έπειτα επιλέγουμε τις 3 κοντινότερες αποστάσεις δηλαδή τους 3 κοντινότερους μετεωρολογικούς σταθμούς σε κάθε ταχυδρομικό κωδικό. Έπειτα υπολογίζουμε το ποσοστό χρησιμοποίησης των κλιματολογικών δεδομένων του εκάστοτε μετεωρολογικού σταθμού. Το ποσοστό αυτό θα είναι αντιστρόφως ανάλογο της απόστασης. Με τον παρακάτω τύπο υπολογίζουμε το κάθε ποσοστό.

=((AR2+AS2+AT2)/AR2)/(((AR2+AS2+AT2)/AR2)+((AR2+AS2+AT2)/AS2)+((AR2+AS2+AT2)/AT2))

AR2	Απόσταση 1ου κοντινότερου Μετεωρ. Σταθμού από Ταχ. Κώδικα
AS2	Απόσταση 2ου κοντινότερου Μετεωρ. Σταθμού από Ταχ. Κώδικα
AT2	Απόσταση 3ου κοντινότερου Μετεωρ. Σταθμού από Ταχ. Κώδικα

Αποτέλεσμα 1ο	Ποσοστό ανάλογα του οποίου θα αντλήσουμε μετεωρολογικά δεδομένα από τον 1ο κοντινότερο Μετεωρ. Σταθμό.
Αποτέλεσμα 2ο	Ποσοστό ανάλογα του οποίου θα αντλήσουμε μετεωρολογικά δεδομένα από τον 2ο κοντινότερο Μετεωρ. Σταθμό.
Αποτέλεσμα 3ο	Ποσοστό ανάλογα του οποίου θα αντλήσουμε μετεωρολογικά δεδομένα από τον 3ο κοντινότερο Μετεωρ. Σταθμό.

Το αποτέλεσμα θα είναι το ποσοστό ανάλογα του οποίου θα αντλήσουμε μετεωρολογικά δεδομένα για κάθε ταχυδρομικό κώδικα. Για παράδειγμα, για τον ταχυδρομικό κώδικα 10431, θα αντλήσουμε δεδομένα κατά 64,71% από τον μετεωρολογικό σταθμό του Θησείου, κατά 18,25% από το σταθμό του Ζωγράφου και κατά 17,04% από το σταθμό του Γαλατσίου.

Στη συνέχεια πολλαπλασιάζουμε κάθε μετεωρολογικό στοιχείο του συγκεκριμένου σταθμού με το ποσοστό που αντιστοιχεί στον ανάλογο μετεωρολογικό σταθμό και κατασκευάζουμε τη βάση δεδομένων για κάθε ταχυδρομικό κωδικό. Για παράδειγμα, για τον ταχυδρομικό κώδικα 10431, θα έχουμε:

Μετεωρολογικός σταθμός	Θησείο	Ζωγράφου	Γαλάτσι
Ποσοστό άντλησης δεδομένων από τον συγκεκριμένο μετεωρολογικό σταθμό.	64,7%	18,3%	17,0%
Θερμοκρασία	Θερμ. σταθμού * 64,7%	Θερμ. σταθμού * 18,3%	Θερμ. σταθμού * 17,0%
Ηλιοφάνεια	Ηλιοφ. σταθμού * 64,7%	Ηλιοφ. σταθμού * 18,3%	Ηλιοφ. σταθμού * 17,0%
Ταχύτητα Ανέμου	Ταχ. Ανέμου σταθμού * 64,7%	Ταχ. Ανέμου σταθμού * 18,3%	Ταχ. Ανέμου σταθμού * 17,0%
Διεύθυνση Ανέμου	Διεύθ. Ανέμου σταθμού * 64,7%	Διεύθ. Ανέμου σταθμού * 18,3%	Διεύθ. Ανέμου σταθμού * 17,0%

Για την κατασκευή της βάσης δεδομένων χρησιμοποιήθηκαν κατά κόρων οι εντολές του Microsoft Excel, vlookup και match, για την αντιστοίχηση των ποσοστών με τον κάθε ταχυδρομικό κωδικό. Επειδή οι βάσεις δεδομένων ήταν τεράστιες, η διαδικασία αυτή ήταν χρονοβόρα και περίπλοκη.

Έπειτα υπολογίζουμε τις βαθμοημέρες θέρμανσης για κάθε ημέρα και για κάθε ταχυδρομικό κώδικα βάση της θερμοκρασίας. Η κλασσική μέθοδος των βαθμοημερών για την εκτίμηση των ενεργειακών απαιτήσεων για θέρμανση βασίζεται στην υπόθεση ότι, σε μακρές χρονικές περιόδους, (π.χ. μήνας, έτος) τα ηλιακά θερμικά κέρδη, καθώς και τα εσωτερικά θερμικά κέρδη ενός κτιρίου, αντισταθμίζουν τις θερμικές του απώλειες, όταν η εξωτερική θερμοκρασία είναι 18.3ºC ή υψηλότερη. Επίσης, ότι η κατανάλωση ενέργειας είναι ανάλογη της διαφοράς ανάμεσα στη μέση εξωτερική θερμοκρασία της ημέρας και αυτή των 18.3ºC. Αυτή η βασική υπόθεση εκφράζεται μαθηματικά σε μία εξίσωση, στην οποία η κατανάλωση ενέργειας μίας χρονικής περιόδου είναι ανάλογη του αριθμού των βαθμοημερών της περιόδου, οι οποίες υπολογίζονται με βάση τη θερμοκρασία των 18.3ºC. Σύμφωνα με την μέθοδο τον υπολογισμό των ημερησίων βαθμοημερών θέρμανσης, χρησιμοποιείτε η παρακάτω εξίσωση:

$$HDD = Tb - Tm$$

Όπου T_b είναι η θερμοκρασία βάσης και T_m είναι η μέση ημερήσια θερμοκρασία περιβάλλοντος. Όπως είναι λογικό μόνο τα θετικά αποτελέσματα έχουν υπόσταση. Η θερμοκρασία βάσης (18.3^o) προσδιορίζετε ως η τιμή της εξωτερικής θερμοκρασίας η οποία για συγκεκριμένες τιμές της εσωτερικής θερμοκρασίας του κτιρίου η ολικές θερμοκρασιακές απώλειες είναι ίσες με το θερμικά κέρδη (από τον ήλιο, τους ενοίκους, τα φώτα κτλ.).

Όπως γίνεται αντιληπτό η μέθοδος αυτή απαιτεί την γνώση των μέσων ημερήσιων τιμών της θερμοκρασίας. Για αυτό το λόγο, όπως προαναφέρθηκε, επιλέχθηκε να συλλεχθούν μετεωρολογικά δεδομένα ανά ημέρα. Οπότε με τον υπολογισμό των βαθμοημερών θέρμανσης για κάθε μέρα, μπορούμε να αθροίσουμε τις ημερήσιες βαθμοημέρες και να βρούμε τις συνολικές βαθμοημέρες θέρμανσης για την περίοδο που μας ενδιαφέρει.

Επομένως δημιουργήθηκαν ακριβείς βάσεις δεδομένων για κάθε ταχυδρομικό κωδικό που περιλαμβάνουν τη θερμοκρασία, ηλιοφάνεια, ταχύτητα ανέμου, διεύθυνση ανέμου, ηλιακή ακτινοβολία και βαθμοημέρες θέρμανσης.

4.4 ΔΕΔΟΜΕΝΑ ΠΟΥ ΑΠΟΚΤΗΘΗΚΑΝ ΑΠΟ ΤΗΝ ΕΤΑΙΡΕΙΑ ΔΙΑΝΟΜΗΣ ΦΥΣΙΚΟΥ ΑΕΡΙΟΥ

Από την εταιρεία διανομής φυσικού αερίου συλλέχθηκαν όλα τα δεδομένα που αφορούν τους καταναλωτές για τα παρακάτω κριτήρια.

- **Εφαρμογές φυσικού αερίου** (κεντρική ή ατομική θέρμανση, θέρμανση νερού, μαγείρεμα).
- **Χρήση φυσικού αερίου** (οικιακή, επαγγελματική).
- **Τύπος καταναλωτή**. (ιδιώτης, επαγγελματίας, είδος επιχείρησης)
- **Ιστορικά στοιχεία κατανάλωσης**.
- **Αιτούμενη εγκατεστημένη ισχύς.**

Το δείγμα αποτελείται από **39.668** καταναλωτές για τους οποίους έχουν πραγματοποιηθεί **492.259** μετρήσεις κατανάλωσης. Για κάθε καταναλωτή έχουν συγκεντρωθεί οι καταναλώσεις φυσικού αερίου και έχουν υπολογιστεί οι βαθμοημέρες θέρμανσης συναρτήσει του πλήθους των ημερών της κάθε χρονικής περιόδου.

Στοιχεία δεδομένων δείγματος
Κωδικός καταναλωτή
Ταχυδρομικός κώδικας
Κατηγορία χρήσης φυσικού αερίου
Κατηγορία καταναλωτή
Κατανάλωση Φυσικού Αερίου
Βαθμοημέρες θέρμανσης
Πλήθος ημερών

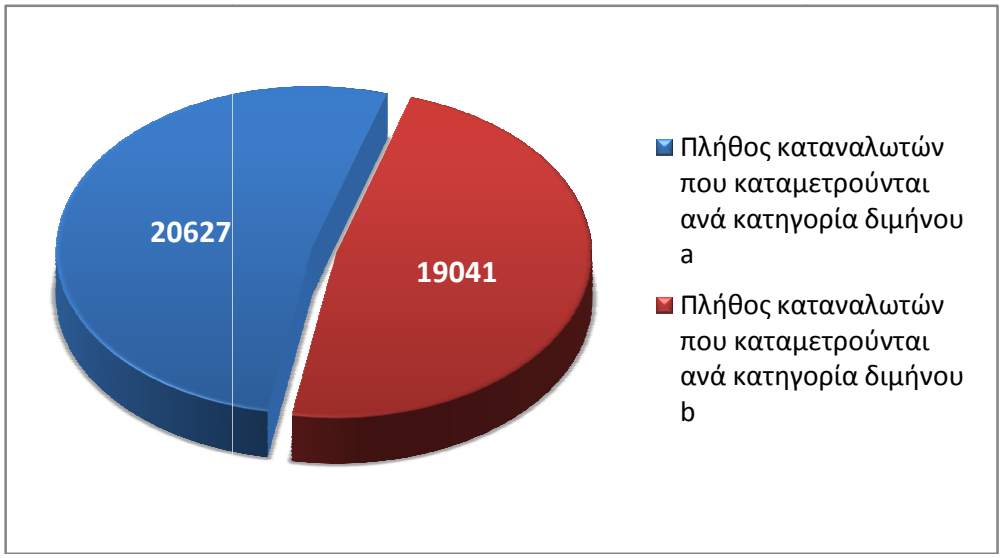
Το δείγμα περιλαμβάνει 39.668 κωδικούς καταναλωτών και 237 ταχυδρομικούς κώδικες της ευρύτερης περιοχής της Αττικής. Η βάση δεδομένων περιέχει 492.259 μετρήσεις κατανάλωσης διάφορων χρονικών περιόδων.

Παρακάτω ακολουθεί μια ανάλυση των δεδομένων που αποκτήθηκαν. Για κάθε καταμέτρηση κατανάλωσης φυσικού αερίου διατίθενται τα ακόλουθα δεδομένα:

1. **Κωδικός πελάτη (Costumer)**. Παίρνει τιμές από 4000000081 ως 4000050463. Υπάρχουν στοιχεία για 39668 πελάτες.

2. **Ταχυδρομικός κώδικας καταναλωτή (Zip code):** Βάση του ταχ. κώδικα έχουν υπολογισθεί οι βαθμοημέρες θέρμανσης και γνωρίζουμε την περιοχή που μένει. Η βάση δεδομένων περιέχει 237 ταχ. κώδικες από 10431 ως 19600.

3. **Ημερομηνίες κατανάλωσης (From Date - To Date):** Οι ημερομηνίες στις οποίες αναφέρεται η μετρηθείσα κατανάλωση. Από αυτές τις ημερομηνίες μπορούμε να καθορίσουμε σε ποιο δίμηνο του έτους αντιστοιχεί η κατανάλωση. Από τους 39668 καταναλωτές του δείγματος, οι 20627 καταναλωτές έχουν καταμετρηθεί ανά δίμηνο, αρχής γενομένης του Νοεμβρίου 2004 (που ονομάζουμε δίμηνο a). Οι υπόλοιποι 19041 καταναλωτές έχουν καταμετρηθεί ανά δίμηνο, αρχής γενομένης του Δεκεμβρίου 2004 (που ονομάζουμε δίμηνο b).



Παρακάτω παρουσιάζονται η κωδική ονομασία που δόθηκε για κάθε δίμηνο.

Δίμηνα Κατηγορίας a			Δίμηνα Κατηγορίας b		
Μήνες	Έτος	Κωδικός	Μήνες	Έτος	Κωδικός
11 - 12	2004	**04f_a**	12 - 01	2004	**04f_b**
01 - 02	2005	**05a_a**	02 - 03	2005	**05a_b**
03 - 04	2005	**05b_a**	04 - 05	2005	**05b_b**
05 - 06	2005	**05c_a**	06 - 07	2005	**05c_b**
07 - 08	2005	**05d_a**	08 - 09	2005	**05d_b**
09 - 10	2005	**05e_a**	10 - 11	2005	**05e_b**
11 - 12	2005	**05f_a**	12 - 01	2005	**05f_b**
01 - 02	2006	**06a_a**	02 - 03	2006	**06a_b**
03 - 04	2006	**06b_a**	04 - 05	2006	**06b_b**
05 - 06	2006	**06c_a**	06 - 07	2006	**06c_b**
07 - 08	2006	**06d_a**	08 - 09	2006	**06d_b**
09 - 10	2006	**06e_a**	10 - 11	2006	**06e_b**
11 - 12	2006	**06f_a**	12 - 01	2006	**06f_b**
01 - 02	2007	**07a_a**	02 - 03	2007	**07a_b**
03 - 04	2007	**07b_a**	04 - 05	2007	**07b_b**
05 - 06	2007	**07c_a**	06 - 07	2007	**07c_b**
07 - 08	2007	**07d_a**	08 - 09	2007	**07d_b**
09 - 10	2007	**07e_a**	10 - 11	2007	**07e_b**
11 - 12	2007	**07f_a**	12 - 01	2007	**07f_b**
01 - 02	2008	**08a_a**	02 - 03	2008	**08a_b**

4. **Διάστημα κατανάλωσης (Days):** Το διάστημα, σε ημέρες, της μετρούμενης κατανάλωσης.

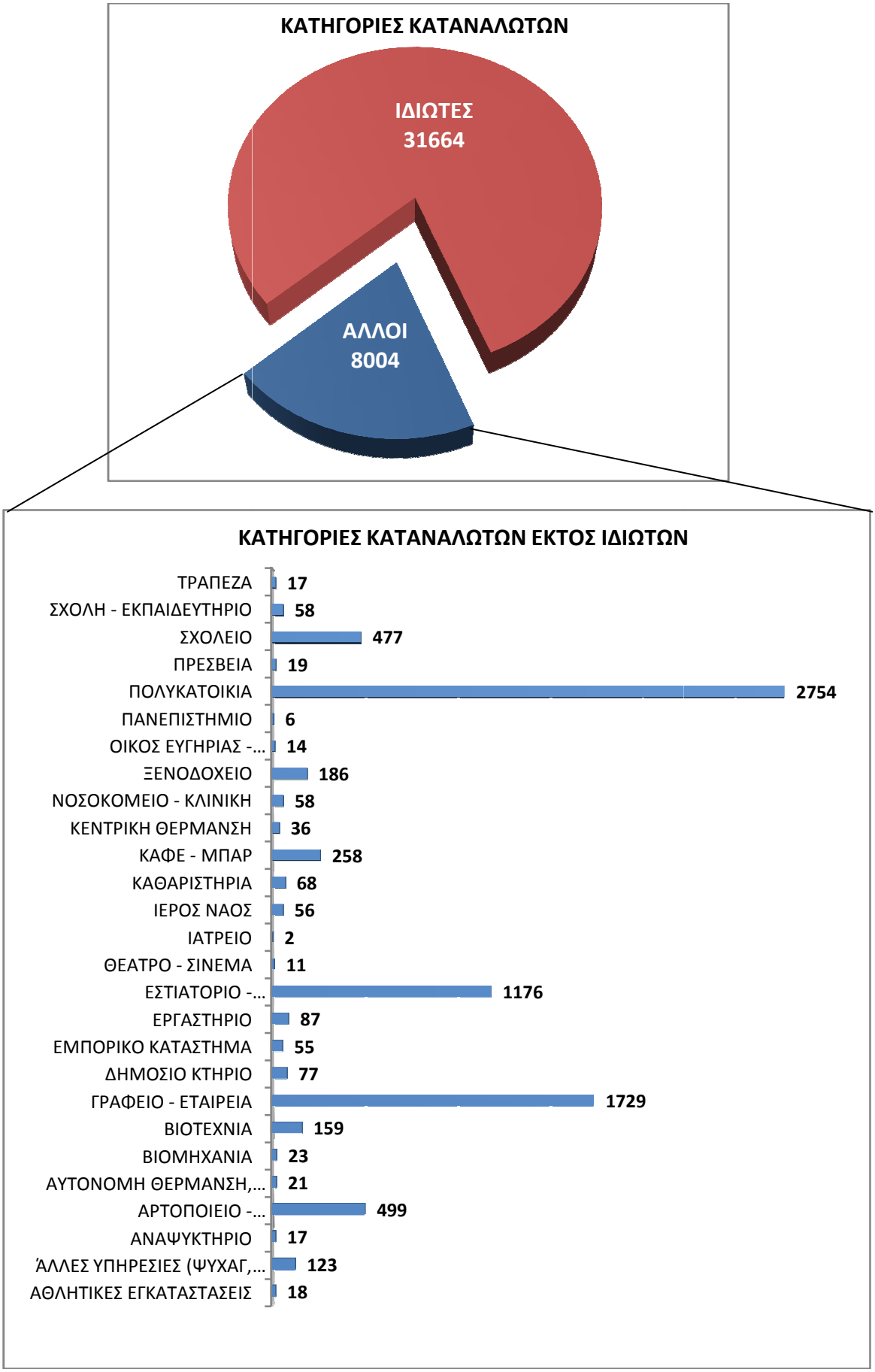
5. **Βαθμοημέρες Θέρμανσης (Degree Days):** Οι βαθμοημέρες θέρμανσης υπολογίστηκαν βάση του κάθε ταχυδρομικό κωδικού και ανάλογα το χρονικό διάστημα της μετρηθείσας κατανάλωσης. Για παράδειγμα, αν η περίοδος που αναφέρεται η κατανάλωση είναι το δίμηνο αρχές Ιανουαρίου - τέλος Φεβρουαρίου, τότε αθροίζουμε τις ημερήσιες βαθμοημέρες για το συγκεκριμένο ταχ. κωδικό για να βρούμε τις συνολικές βαθμοημέρες θέρμανσης για την περίοδο που μας ενδιαφέρει.

6. **Κατανάλωση (Consumption):** Η κατανάλωση φυσικού αερίου για ένα συγκεκριμένο διάστημα.

7. **Εγκατεστημένη Ισχύς (Capacity charge):** Μέγιστη δυνατότητα παροχής φυσικού αερίου. Το στοιχείο αυτό είναι το μόνο ενδεικτικό ποσοτικό στοιχείο που φανερώνει την ποσότητα φυσικού αερίου που υπολογίσθηκε ότι θα καταναλώνει το κτήριο. Είναι ενδεικτικό του μεγέθους, της κατασκευής του κτηρίου, της θερμομόνωσης του άλλα και άλλων στοιχείων που δεν μπορούν να βρεθούν για αυτά δεδομένα.

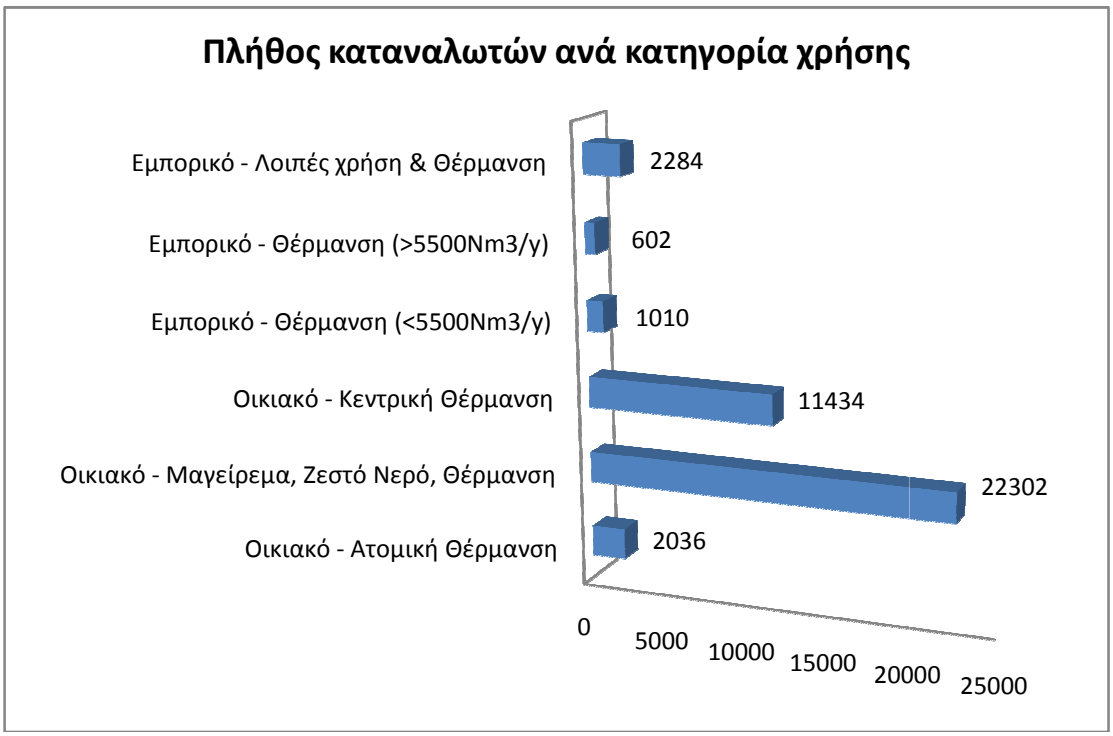
8. **Κατηγορίες Καταναλωτή (Account_Desc):** Η κατηγορία που εντάσσεται ο πελάτης. Αρχικά υπήρχαν πάνω από 50 κατηγορίες πελατών. Μετά από ομαδοποίηση κατηγοριών που παρουσιάζουν ίδιο καταναλωτικό προφίλ καταλήξαμε σε 28 κατηγορίες πελατών οι οποίες είναι:

Κατηγορίες Καταναλωτών	Κωδικός Κατηγορίας	Πλήθος Καταναλωτών Δείγματος	Πλήθος Μετρήσεων Κατανάλωσης
ΑΘΛΗΤΙΚΕΣ ΕΓΚΑΤΑΣΤΑΣΕΙΣ	**ad01**	18	255
ΑΛΛΕΣ ΥΠΗΡΕΣΙΕΣ (ΨΥΧΑΓ, ΚΛΠ)	**ad02**	123	2067
ΑΝΑΨΥΚΤΗΡΙΟ	**ad03**	17	208
ΑΡΤΟΠΟΙΕΙΟ - ΖΑΧΑΡΟΠΛΑΣΤΕΙΟ	**ad04**	499	7756
ΑΥΤΟΝΟΜΗ ΘΕΡΜΑΝΣΗ, ΜΑΓ., ΖΝΧ	**ad05**	21	388
ΒΙΟΜΗΧΑΝΙΑ	**ad06**	23	311
ΒΙΟΤΕΧΝΙΑ	**ad07**	159	2623
ΓΡΑΦΕΙΟ - ΕΤΑΙΡΕΙΑ	**ad08**	1729	16254
ΔΗΜΟΣΙΟ ΚΤΗΡΙΟ	**ad09**	77	903
ΕΜΠΟΡΙΚΟ ΚΑΤΑΣΤΗΜΑ	**ad10**	55	579
ΕΡΓΑΣΤΗΡΙΟ	**ad11**	87	1407
ΕΣΤΙΑΤΟΡΙΟ - ΜΕΖΕΔΟΠΩΛΕΙΟ	**ad12**	1176	17586
ΘΕΑΤΡΟ - ΣΙΝΕΜΑ	**ad13**	11	189
ΙΑΤΡΕΙΟ	**ad14**	2	16
ΙΔΙΩΤΗΣ	**ad15**	31664	401123
ΙΕΡΟΣ ΝΑΟΣ	**ad16**	56	802
ΚΑΘΑΡΙΣΤΗΡΙΑ	**ad17**	68	1188
ΚΑΦΕ - ΜΠΑΡ	**ad18**	258	4131
ΚΕΝΤΡΙΚΗ ΘΕΡΜΑΝΣΗ	**ad19**	36	653
ΝΟΣΟΚΟΜΕΙΟ - ΚΛΙΝΙΚΗ	**ad20**	58	997
ΞΕΝΟΔΟΧΕΙΟ	**ad21**	186	3368
ΟΙΚΟΣ ΕΥΓΗΡΙΑΣ - ΟΡΦΑΝΟΤΡΟΦΕΙΟ	**ad22**	14	184
ΠΑΝΕΠΙΣΤΗΜΙΟ	**ad23**	6	117
ΠΟΛΥΚΑΤΟΙΚΙΑ	**ad24**	2754	24510
ΠΡΕΣΒΕΙΑ	**ad25**	19	288
ΣΧΟΛΕΙΟ	**ad26**	477	3232
ΣΧΟΛΗ - ΕΚΠΑΙΔΕΥΤΗΡΙΟ	**ad27**	58	873
ΤΡΑΠΕΖΑ	**ad28**	17	256

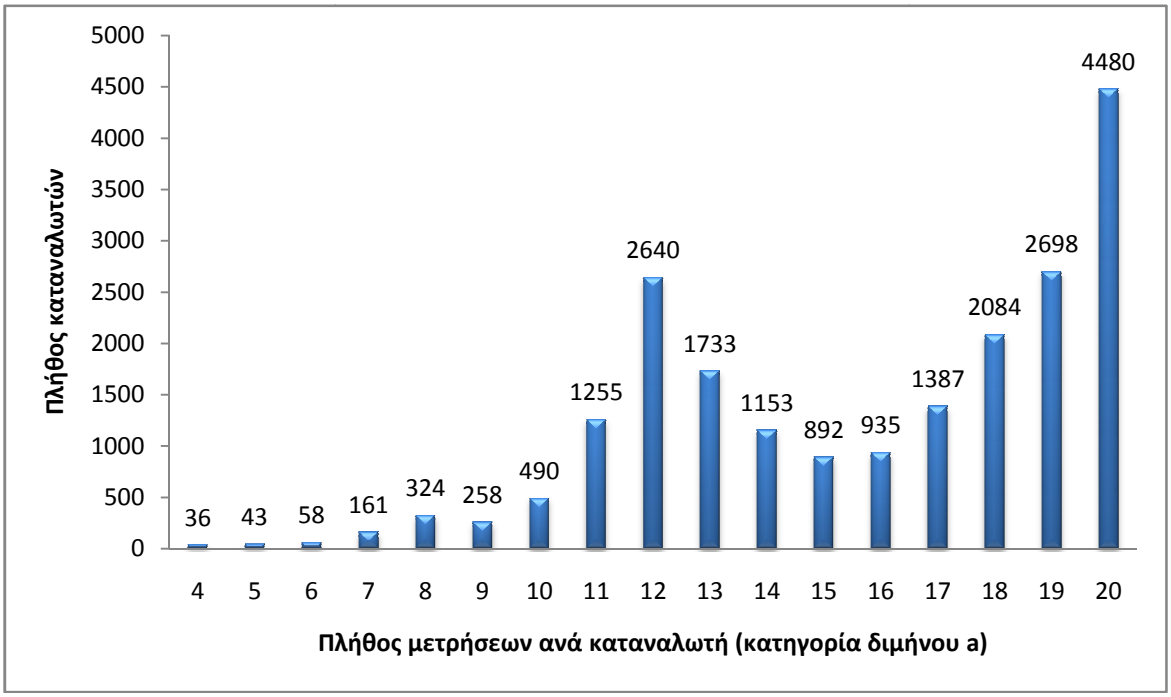

ΚΑΤΗΓΟΡΙΕΣ ΚΑΤΑΝΑΛΩΤΩΝ
ΙΔΙΩΤΕΣ
31664
ΑΛΛΟΙ
8004
ΚΑΤΗΓΟΡΙΕΣ ΚΑΤΑΝΑΛΩΤΩΝ ΕΚΤΟΣ ΙΔΙΩΤΩΝ
ΤΡΑΠΕΖΑ 17
ΣΧΟΛΗ - ΕΚΠΑΙΔΕΥΤΗΡΙΟ 58
ΣΧΟΛΕΙΟ 477
ΠΡΕΣΒΕΙΑ 19
ΠΟΛΥΚΑΤΟΙΚΙΑ 2754
ΠΑΝΕΠΙΣΤΗΜΙΟ 6
ΟΙΚΟΣ ΕΥΓΗΡΙΑΣ -... 14
ΞΕΝΟΔΟΧΕΙΟ 186
ΝΟΣΟΚΟΜΕΙΟ - ΚΛΙΝΙΚΗ 58
ΚΕΝΤΡΙΚΗ ΘΕΡΜΑΝΣΗ 36
ΚΑΦΕ - ΜΠΑΡ 258
ΚΑΘΑΡΙΣΤΗΡΙΑ 68
ΙΕΡΟΣ ΝΑΟΣ 56
ΙΑΤΡΕΙΟ 2
ΘΕΑΤΡΟ - ΣΙΝΕΜΑ 11
ΕΣΤΙΑΤΟΡΙΟ -... 1176
ΕΡΓΑΣΤΗΡΙΟ 87
ΕΜΠΟΡΙΚΟ ΚΑΤΑΣΤΗΜΑ 55
ΔΗΜΟΣΙΟ ΚΤΗΡΙΟ 77
ΓΡΑΦΕΙΟ - ΕΤΑΙΡΕΙΑ 1729
ΒΙΟΤΕΧΝΙΑ 159
ΒΙΟΜΗΧΑΝΙΑ 23
ΑΥΤΟΝΟΜΗ ΘΕΡΜΑΝΣΗ,... 21
ΑΡΤΟΠΟΙΕΙΟ -... 499
ΑΝΑΨΥΚΤΗΡΙΟ 17
ΆΛΛΕΣ ΥΠΗΡΕΣΙΕΣ (ΨΥΧΑΓ,... 123
ΑΘΛΗΤΙΚΕΣ ΕΓΚΑΤΑΣΤΑΣΕΙΣ 18

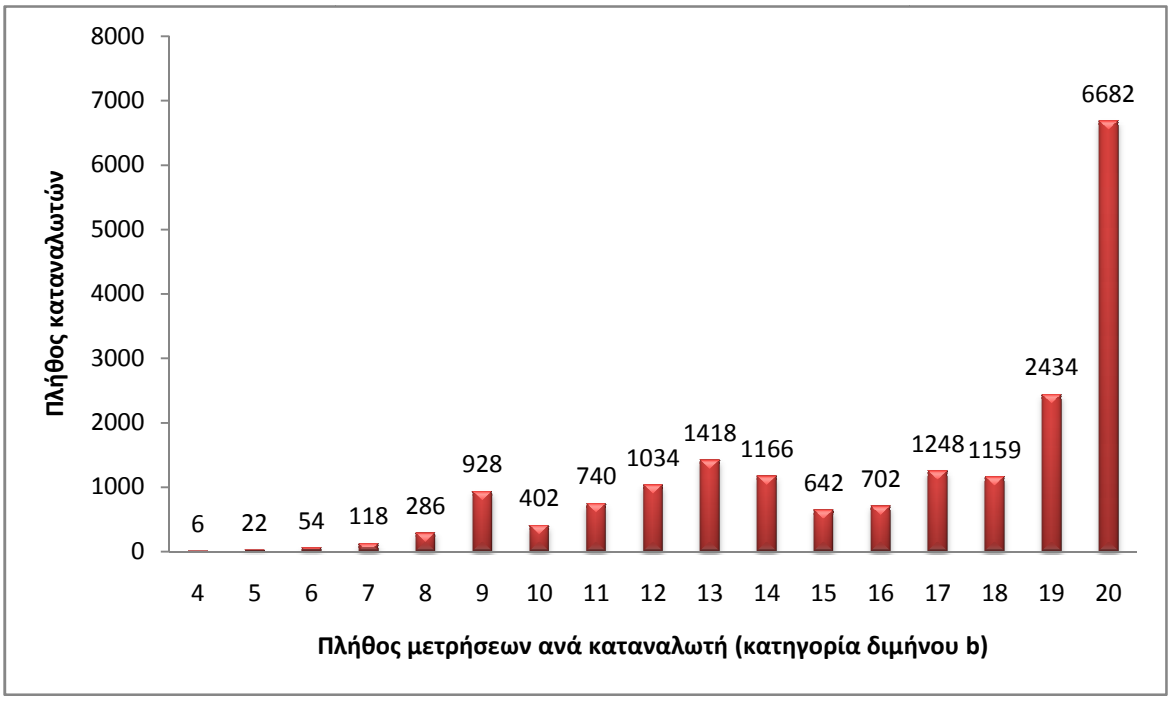
9. **Κατηγορίες χρήσης φυσικού αερίου. (Installation Type Desc):** Χρήση φυσικού αερίου. Περιλαμβάνει 6 κατηγορίες: 1. Οικιακό - Μαγείρεμα, Ζεστό Νερό, Θέρμανση, 2. Οικιακό - Ατομική Θέρμανση, 3. Κεντρική Θέρμανση, 4. Εμπορικό - θέρμανση (<5500Nm3/y), 5. Εμπορικό - θέρμανση (>5500Nm3/y), 6. Εμπορικό - Λοιπές χρήσεις & θέρμανση.

Κατηγορίες Χρήσης	Κωδικός Κατηγορίας	Πλήθος Καταναλωτών Δείγματος	Πλήθος Μετρήσεων Κατανάλωσης
Οικιακό - Ατομική Θέρμανση	in1	2036	25286
Οικιακό - Μαγείρεμα, Ζεστό Νερό, Θέρμανση	in2	22302	280220
Οικιακό - Κεντρική Θέρμανση	in3	11434	130736
Εμπορικό - Θέρμανση (<5500Nm3/y)	in4	1010	15654
Εμπορικό - Θέρμανση (>5500Nm3/y)	in5	602	5483
Εμπορικό - Λοιπές χρήσεις & Θέρμανση	in6	2284	34880



Από τα διαθέσιμα στοιχεία για κάθε καταναλωτή απουσιάζουν μετρήσεις κατανάλωσης για διάφορους λόγους. Αυτό σημαίνει ότι υπάρχει διαφορετικό πλήθος διμηνιαίων μετρήσεων για κάθε καταναλωτή. Στους παρακάτω πίνακες εμφανίζονται το πλήθος των καταναλωτών συναρτήσει του πλήθους των μετρήσεων ανά καταναλωτή. Για παράδειγμα, από το σύνολο των 20627 καταναλωτών της κατηγορίας διμήνου a, οι 4480 έχουν καταμετρηθεί 20 φορές.





ΚΕΦΑΛΑΙΟ 5: ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

5.1 ΕΙΣΑΓΩΓΗ

Στα πλαίσια της επιστημονικής περιοχής Εξόρυξης Δεδομένων (Data Mining) ή αλλιώς «Ανακάλυψης Γνώσης σε Βάσεις Δεδομένων» (Knowledge Discovery in Databases) η μηχανική μάθηση αποτέλεσε κινητήριο μοχλό για την ανάπτυξη της ως άνω περιοχής και αναπόσπαστο τμήμα της. Αποτελεί έναν κλάδο του γενικότερου επιστημονικού χώρου της τεχνητής νοημοσύνης που προσφέρει εργαλεία που αυτόματα επεξεργάζονται και αναλύουν μεγάλες ποσότητες δεδομένων, παράγοντας χρήσιμη γνώση.

Δύο είναι οι μεγάλοι στόχοι των προβλημάτων με τα οποία ασχολείται η μηχανική μάθηση και ως προέκτασή της η Εξόρυξη Δεδομένων:

1. Η πρόβλεψη
2. Η ανακάλυψη γνώσης

Η ανακάλυψη γνώσης από δεδομένα (Knowledge Discovery in Data –KDD) είναι η μη τετριμμένη διαδικασία για την εξαγωγή έγκυρων, πρωτότυπων, πιθανώς χρήσιμων και οπωσδήποτε κατανοητών προτύπων (patterns) μέσα στα δεδομένα. Η Εξόρυξη Γνώσης (Data Mining) είναι η χρήση αλγορίθμων και τεχνικών για την εξαγωγή προτύπων κατά τη διάρκεια της διαδικασίας KDD.

Κύριες εργασίες της διαδικασίας εξόρυξης γνώσης είναι η εύρεση συσχετίσεων μεταξύ των δεδομένων (κανόνες συσχέτισης), η κατηγοριοποίηση σε προκαθορισμένες κλάσεις (δέντρα απόφασης, νευρωνικά δίκτυα, κατηγοριοποίηση Bayesian) και η συσταδοποίηση / ομαδοποίηση (ιεραρχικοί, διαμεριστικοί, με βάση την πυκνότητα).

- **Κανόνες Συσχέτισης (Association Rules)**

Έστω I ένα σύνολο από αντικείμενα (items). Ένας κανόνας συσχέτισης έχει τη μορφή $X \Rightarrow Y$ όπου $X,Y \subseteq I$ και $X \cap Y = \emptyset$. Το X ονομάζεται head (κεφαλή) ή LHS (left-hand side) ή antecedent (προηγούμενο) του κανόνα. Το Y ονομάζεται body (σώμα) ή RHS (right-hand side) ή consequent (επακόλουθο) του κανόνα. Ένας κανόνας σχετίζεται με διάφορα μέτρα ποιότητας όπως:

α) Η υποστήριξη (support) του κανόνα ορίζεται ως το ποσοστό των δοσοληψιών που περιέχουν τα αντικείμενα ($X \cup Y$) ή αλλιώς η πιθανότητα $P(X \cup Y)$.

β) Η εμπιστοσύνη (confidence) του κανόνα ορίζεται ως το ποσοστό των δοσοληψιών του X που περιέχουν και το Y ή αλλιώς, η εξαρτημένη πιθανότητα $P(X \cup Y|X)= P(X \cup Y)/P(X)$.

- **Κατηγοριοποίηση (Classification)**

  Ανάπτυξη ενός μοντέλου πρόβλεψης της κλάσης των στιγμιότυπων ενός προβλήματος. Το μοντέλο χτίζεται με βάση ένα σύνολο δεδομένων εκπαίδευσης (training set). Η απόδοση του μοντέλου αξιολογείται με βάση ένα σύνολο δεδομένων ελέγχου (test set). Εποπτευόμενη μάθηση: χρειάζεται να γνωρίζουμε την κλάση των στιγμιότυπων εκπαίδευσης.

- **Συσταδοποίηση (Clustering)**

  Διαχωρισμός των δεδομένων σε ομάδες / συστάδες έτσι ώστε για κάθε εγγραφή που περιλαμβάνει μία συστάδα, η ομοιότητα της με οποιασδήποτε εγγραφή από την ίδια συστάδα να είναι μεγαλύτερη από την ομοιότητα της με οποιασδήποτε εγγραφή από άλλες συστάδες. Στην μη εποπτευόμενη μάθηση δεν γνωρίζουμε την κλάση στην οποία ανήκουν τα δεδομένα εκπαίδευσης. Μας δίνεται ένα σύνολο μετρήσεων, παρατηρήσεων κλπ. με στόχο να ανακαλύψουμε κλάσεις ή ομάδες μέσα στα δεδομένα.

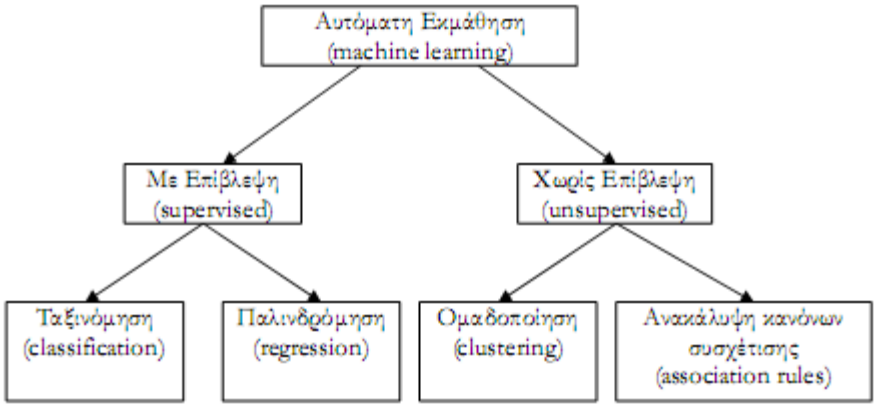
Συνήθως η ανακάλυψη γνώσης αποτελεί ένα πρότερο στάδιο της πρόβλεψης, στις περιπτώσεις όπου τα δεδομένα είναι ακατάλληλα για πρόβλεψη. Επίσης δρα συμπληρωματικά στην πρόβλεψη, όντας εγγύτερα στην υποστήριξη αποφάσεων παρά στη λήψη αποφάσεων. Από την άλλη πλευρά η πρόβλεψη είναι απαραίτητη σε πολλές εφαρμογές (π.χ. για τη μετεωρολογία, τη σεισμολογία, τις οικονομικές επιστήμες) ακόμη και όταν λειτουργεί ως μαύρο κουτί, όταν δηλαδή δεν παρέχει πληροφορίες που να την υποστηρίζουν και να την δικαιολογούν.

Τα τελευταία χρόνια πληθαίνουν οι εφαρμογές της μηχανικής μάθησης. Τεχνικές όπως οι γενετικοί αλγόριθμοι, τα νευρωνικά δίκτυα, η ασαφής λογική, τα έμπειρα συστήματα, τα δέντρα απόφασης και άλλες μέθοδοι από τους τομείς της επεξεργασίας σημάτων και από την αναγνώριση προτύπων και την στατιστική, αντιμετωπίζουν με επιτυχία τη δύσκολη φύση πάμπολλων προβλημάτων.

Καθώς, η μηχανική μάθηση (machine learning) είναι ένα υποπεδίο της τεχνητής νοημοσύνης, ασχολείται με τη σχεδίαση αυτόματων διαδικασιών ικανών να μαθαίνουν από παραδείγματα. Η εκμάθηση από παραδείγματα αφορά στην παραγωγή μιας λογικής περιγραφής των αναγκαίων και ικανών συνθηκών που αντιστοιχούν σε μια κλάση αντικειμένων, π.χ. σε έναν κανόνα δοθείσης μιας γλώσσας αναπαράστασης.

Μία σημαντική φροντίδα είναι η εύρεση των αναγκαίων συμβιβασμών ανάμεσα στην πολυπλοκότητα των κανόνων και στην προσαρμογή των δεδομένων, έτσι ώστε να αποφευχθεί η υπερβολική προσαρμογή (overfitting) και να καταστεί δυνατή η ερμηνεία των αποτελεσμάτων.

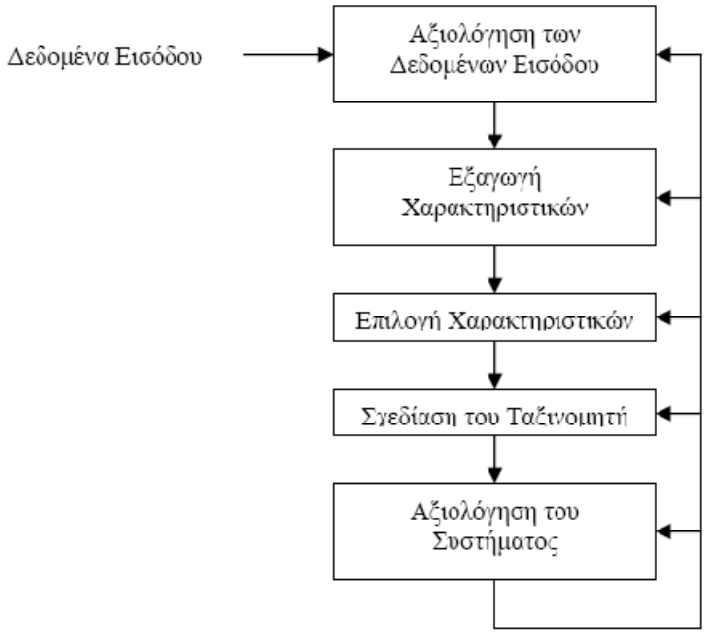
Η μηχανική μάθηση περιλαμβάνει μία ιεραρχία μεθοδολογιών, όπως απεικονίζονται στο παρακάτω σχήμα.



Στο πρώτο επίπεδο χωρίζεται σε δύο κλάδους: την εκμάθηση με επίβλεψη και την εκμάθηση χωρίς επίβλεψη. Στην πρώτη περίπτωση σε κάθε παράδειγμα είναι γνωστή η τιμή ή η κατηγορία στην οποία ανήκει η παρατήρηση y, ενώ στη δεύτερη είναι άγνωστη και η μηχανική μάθηση επιχειρεί να την προσεγγίσει ομαδοποιώντας τα παραδείγματα. Η πρόβλεψη της αυριανής μέσης θερμοκρασίας με δεδομένες μετρήσεις της μέσης ημερήσιας θερμοκρασίας είναι για παράδειγμα ένα πρόβλημα με επίβλεψη, γιατί στα παραδείγματα του παρελθόντος υπάρχει η ακριβής τιμή της μετρούμενης θερμοκρασίας. Αντίθετα το πρόβλημα της εύρεσης της κατηγορίας ενός νέου προϊόντος σε σχέση με τα υπόλοιπα προϊόντα σε ένα πολυκατάστημα με βάση τα χαρακτηριστικά του και τις πωλήσεις του είναι ένα συνηθισμένο παράδειγμα προβλήματος χωρίς επίβλεψη.

Το ζητούμενο στην περίπτωση της μάθησης υπό επίβλεψη είναι η κατασκευή ενός μοντέλου (ή αλλιώς μιας υπόθεσης) που αναπαριστά τη γνώση που παρέχεται μέσω της εμπειρίας E και το οποίο στη συνέχεια χρησιμοποιείται για την αξιολόγηση νέων στιγμιοτύπων. Κατά κανόνα, οι προβλέψεις του προκύπτοντος μοντέλου (οι τιμές της συνάρτησης που προσεγγίζει τη συνάρτηση-στόχο) θα επαληθεύονται (θα ισούνται με την τιμή της συνάρτησης-στόχου) για την πλειοψηφία από τα στοιχεία που περιλαμβάνονται στην E, τα οποία λέγονται και στιγμιότυπα εκπαίδευσης. Μία θεμελιώδης υπόθεση στην οποία στηρίζονται οι περισσότεροι αλγόριθμοι και η θεωρία μηχανικής μάθησης είναι πως η κατανομή των στιγμιοτύπων εκπαίδευσης είναι αντιπροσωπευτική της γενικής κατανομής των στιγμιοτύπων στον υπό μοντελοποίηση χώρο. Όπως είναι φυσικό, οι προβλέψεις ενός μοντέλου (ταξινομητή) για μελλοντικά (άγνωστα) στιγμιότυπα είναι περισσότερο αξιόπιστες αν τα στιγμιότυπα εκπαίδευσης ακολουθούν παρόμοια κατανομή με αυτή των μελλοντικών. Η διεργασία της ανακάλυψης γνώσης και της μηχανικής μάθησης είναι μία διαδικασία πολλαπλών σταδίων, η οποία περιλαμβάνει την προετοιμασία των δεδομένων, την αναζήτηση προτύπων και την αξιολόγηση της γνώσης που ανακτάται από τα δεδομένα.

Τα βασικά στάδια ενός λογισμικού μηχανικής μάθησης απεικονίζονται στο παρακάτω σχήμα.



Με μαθηματικούς όρους, έστω ένα δείγμα δεδομένων εκπαίδευσης L. Στα προβλήματα με επίβλεψη αυτό αποτελείται από n ζευγάρια παρατηρήσεων (y1, x1), …, (yn,xn), όπου xi ∈RP και yi ∈R (i = 1, …, n). Στα προβλήματα χωρίς επίβλεψη δεν υπάρχει η μεταβλητή y και το σύνολο L αποτελείται από n παρατηρήσεις x1, …, xn, όπου xi ∈RP. Η μεταβλητή Y ονομάζεται εξαρτημένη μεταβλητή και οι μεταβλητές Xi ονομάζονται ανεξάρτητες, ή αλλιώς ιδιότητες του προβλήματος.

Η εκμάθηση με επίβλεψη αφορά σε δύο περαιτέρω κατηγορίες προβλημάτων, την παλινδρόμηση και την ταξινόμηση. Στην πρώτη η μεταβλητή y είναι μία συνεχής μεταβλητή ενώ στη δεύτερη μία διακριτή κατηγορική μεταβλητή. Το παράδειγμα της θερμοκρασίας είναι πρόβλημα παλινδρόμησης όταν η θερμοκρασία μετράται σαν πραγματικός συνεχής αριθμός, ενώ είναι πρόβλημα κατάταξης αν μετράται σαν θερμοκρασιακή ζώνη (π.χ. κρύο/ ζέστη).

Τέλος, τα προβλήματα χωρίς επίβλεψη χωρίζονται σε ομαδοποίησης και ανακάλυψης κανόνων. Η μεθοδολογία της ομαδοποίησης επιχειρεί να ξεχωρίσει τα δεδομένα σε άτυπες ομάδες, με βάση τις αποστάσεις ανάμεσα στις ιδιότητες των παραδειγμάτων. Το παράδειγμα της κατάταξης ενός νέου προϊόντος σε ομάδα που αναφέρθηκε προηγουμένως ανήκει στην πρώτη κατηγορία.

Η ανακάλυψη κανόνων από την άλλη διερευνά τα δεδομένα για να ανακαλύψει κανόνες που διέπουν τις ιδιότητες των παραδειγμάτων. Αν για παράδειγμα είχαμε στα χέρια μας ένα σύνολο δεδομένων που περιελάμβανε τις αγορές προϊόντων σε ένα κατάστημα από κάθε πελάτη, τότε ένας αλγόριθμος ανακάλυψης κανόνων θα συμπέρανε κανόνες του τύπου ‘26% των πελατών που αγόρασαν ψυγεία, αγόρασαν και ηλεκτρική σκούπα’, ή ‘68% των πελατών που αγόρασαν σοκολατάκια και αγόρασαν επίσης σαμπάνια, πλήρωσαν με πιστωτική κάρτα’. Η πλειοψηφία των προβλημάτων ανήκει στην κατηγορία των προβλημάτων εκμάθησης με επίβλεψη.

5.2 Η ΔΙΑΔΙΚΑΣΙΑ ΕΠΙΛΥΣΗΣ ΕΝΟΣ ΠΡΟΒΛΗΜΑΤΟΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Η διαδικασία λύσεως ενός προβλήματος Μηχανικής Μάθησης είναι μία αλληλεπιδραστική και επαναληπτική διαδικασία, η οποία περιλαμβάνει πλήθος βημάτων στα οποία θα πρέπει να ληφθούν αποφάσεις από τον χρήστη.

Αρχικά, κατανοούμε όλες τις πλευρές του συγκεκριμένου προβλήματος συμπεριλαμβανομένης οποιασδήποτε σχετικής προηγούμενης γνώσης για το πρόβλημα και τους στόχους των τελικών χρηστών για την αρχική επιλογή των δεδομένων εκπαίδευσης. Η κατανόηση όλων των πτυχών του προβλήματος είναι σημαντικό βήμα για τη περαιτέρω συνέχιση της όλης διαδικασίας.

Εν συνεχεία, σημαντικός είναι ο ρόλος της προετοιμασίας των δεδομένων που αφορά την εξάλειψη των απλών σφαλμάτων και την ορθή αποκατάσταση των εγγραφών (στιγμιότυπων).

Η διαδικασία αυτή είναι πολύ σημαντική διότι ασυνεπή και λανθασμένα δεδομένα οδηγούν σε ανακριβή αποτελέσματα. Μια περίπτωση που εμφανίζεται συχνά η ανάγκη διόρθωσης των δεδομένων είναι τα πεδία που περιέχουν κείμενο. Ένα όνομα πόλης, για παράδειγμα, μπορεί να είναι γραμμένο με ποικίλους τρόπους (π.χ. διάφορες συντομεύσεις).

5.3 Η ΠΡΟΕΤΟΙΜΑΣΙΑ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΣΤΗΝ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Μία σημαντική υπόθεση για την επιτυχία μίας εφαρμογής μηχανικής μάθησης είναι η σωστή εκμετάλλευση των παρεχόμενων δεδομένων. Όλοι οι αλγόριθμοι της μηχανικής μάθησης απαιτούν από αυτά να είναι ενοποιημένα σε ένα αρχείο εισόδου ακολουθώντας συγκεκριμένη τυποποίηση. Το ζητούμενο είναι τα παραδείγματα να αντιπροσωπεύονται από τις γραμμές του αρχείου και οι ιδιότητες από τις στήλες. Ο καθορισμός και η επιλογή του συνόλου των μεταβλητών που περιγράφουν τα δεδομένα αποτελεί κρίσιμο βήμα στη διαδικασία μάθησης αφού σε μεγάλο βαθμό επηρεάζει την αναπαράσταση των δεδομένων και τα παραγόμενα αποτελέσματα.

Επιλέγουμε το σύνολο των δεδομένων ή επικεντρώνουμε το ενδιαφέρον μας στις μεταβλητές ή τα δεδομένα στα οποία πρόκειται να γίνει η όλη επεξεργασία και να εξαχθεί η γνώση και τα τελικά συμπεράσματα. Όπως είπαμε κάθε στοιχείο αποτελεί ένα πολυδιάστατο διάνυσμα, όπου κάθε διάσταση εκφράζει και μια μεταβλητή. Ο πίνακας παρουσιάζει τις κατηγορίες των μεταβλητών.

<u>Ποσοτικές μεταβλητές</u>
i. Συνεχείς (π.χ.: το βάρος ενός ανθρώπου)
ii. Διακριτές (π.χ.: το πλήθος των μαθητών)
<u>Ποιοτικές μεταβλητές</u>
i. Κατηγορικές (π.χ.: η οικογενειακή εξέταση)
ii. Διάταξης (π.χ.: η επίδοση ενός μαθητή: Καλώς/Λίαν καλώς/ Άριστα)

Πολλές δυσκολίες προκύπτουν προς την κατεύθυνση της ενοποίησης των δεδομένων. Αυτά μπορούν να υπάρχουν σε διάφορες μορφές και σε διαφορετικά αποθηκευτικά μέσα. Μπορεί να είναι κατανεμημένες ή συγκεντρωμένες, σε αρχεία κειμένου, σε βάσεις δεδομένων, σε λογιστικά φύλλα, ταξινομημένες ή αταξινόμητες, χρονικά διατεταγμένες ή χωρικά διεσπαρμένες.

5.4 ΕΠΙΛΟΓΗ ΤΟΥ ΚΑΤΑΛΛΗΛΟΥ ΑΛΓΟΡΙΘΜΟΥ ΜΑΘΗΣΗΣ

Η επιλογή του καταλληλότερου αλγορίθμου μάθησης για το δοθέν πρόβλημα περιλαμβάνει αποφάσεις για το ποιο μοντέλο και ποιες παράμετροι είναι οι πλέον κατάλληλες να χρησιμοποιηθούν. Είναι γνωστό ότι ένα πρόβλημα μπορεί να λυθεί συγκριτικά καλά χρησιμοποιώντας διάφορες μεθόδους, π.χ. διαφορετικούς τύπους τεχνητών νευρωνικών δικτύων, δέντρων απόφασης, άλλων ταξινομητών, κ.λπ. Επιπλέον, μπορεί να συμβεί μια μέθοδος να προβλέπει καλύτερα ορισμένα μέρη του διαστήματος των περιπτώσεων από τις άλλες. Κατά συνέπεια, η επιλογή της καταλληλότερης μεθόδου για την τελική λύση είναι ένα περίπλοκο πρόβλημα.

Ακολούθως περιγράφονται οι πιο συνήθεις μέθοδοι εκτίμησης ακρίβειας για την επιλογή του βέλτιστου μοντέλου. Η ακρίβεια (accuracy) του μοντέλου προκύπτει από τη διαίρεση του αθροίσματος των στιγμιότυπων που σωστά ανατέθηκαν στην συγκεκριμένη κατηγορία και των στιγμιοτύπων που σωστά απορρίφθηκαν από την συγκεκριμένη κατηγορία, με το άθροισμα των στιγμιοτύπων που σωστά ανατέθηκαν στην συγκεκριμένη κατηγορία, των στιγμιοτύπων που λανθασμένα ανατέθηκαν στην συγκεκριμένη κατηγορία, των στιγμιοτύπων που λανθασμένα απορρίφθηκαν από την συγκεκριμένη κατηγορία και των στιγμιοτύπων που σωστά απορρίφθηκαν από την συγκεκριμένη κατηγορία.

Η πιο απλή μέθοδος αξιολόγησης της ακρίβειας είναι η εκτίμηση μέσω δείγματος ελέγχου. Σε αυτήν, η αρχική συλλογή των δεδομένων χωρίζεται σε δύο ξένα υποσύνολα, ένα σύνολο εκπαίδευσης (training set) και ένα σύνολο ελέγχου (test set). Ο αλγόριθμος μάθησης δέχεται το πρώτο για την εκπαίδευσή του και "εξετάζεται" στο δεύτερο, χωρίς να του αποκαλύπτονται φυσικά οι γνωστές (και θεωρούμενες σωστές) απαντήσεις. Στην περίπτωση που υπάρχουν παράμετροι του αλγορίθμου που πρέπει να συντονισθούν, έτσι ώστε να επιλεχθούν αυτές που δίνουν την υψηλότερη αποτελεσματικότητα, το σύνολο εκπαίδευσης πρέπει να χωριστεί επιπλέον σε ένα "πραγματικό" σύνολο εκπαίδευσης και ένα σύνολο επικύρωσης (validation set), το οποίο χρησιμοποιείται "εσωτερικά" από τον αλγόριθμο για την εκτίμηση των βέλτιστων παραμέτρων. Μόνο όταν αυτές σταθεροποιηθούν, εξετάζεται ο αλγόριθμος πάνω στο σύνολο ελέγχου.

Σε αυτήν την περίπτωση η στατιστική μπορεί να δώσει κάποια διαστήματα μέσα στα οποία ανήκει η ακρίβεια με πιθανότητα (1 - α) 100%, τα οποία καλούνται διαστήματα εμπιστοσύνης. Με εφαρμογή της διαδικασίας που δίνει τα διαστήματα εμπιστοσύνης μπορεί να δοθεί ένα διάστημα, όπου με πιθανότητα 95% η τιμή πραγματικού σφάλματος ανήκει σ' αυτό. Το διάστημα εμπιστοσύνης λοιπόν του σφάλματος είναι:

$$\left[error_S(h) - 1.96\sqrt{\frac{error_S(h)(1-error_S(h))}{n}}, error_S(h) + 1.96\sqrt{\frac{error_S(h)(1-error_S(h))}{n}} \right]$$

Άρα η έκφραση που πρέπει να εκτιμηθεί για την σύγκριση αλγορίθμων δίνεται από την έκφραση:

$$error_{T_0}(L_A(S_0)) - error_{T_0}(L_B(S_0))$$

όπου η έκφραση L(S) συμβολίζει τον ταξινομητή που εξάγεται από έναν αλγόριθμο L που εκπαιδεύτηκε πάνω σε ένα σύνολο εκπαίδευσης S και ο συμβολισμός S ⊂ D σημαίνει ότι η επιλογή του S γίνεται σύμφωνα με την κατανομή D. Για να εκτιμηθεί η διαφορά του σφάλματος υπολογίζεται λοιπόν η μέση τιμή

$$\bar{\delta} \equiv \underset{S \subset D_0}{E}\left[error_D(L_A(S)) - error_D(L_B(S)) \right]$$

δίνοντας ένα διάστημα εμπιστοσύνης για τη διαφορά του σφάλματος $\bar{\delta} \pm t_{N,k-1} s_{\bar{\delta}}$

όπου $t_{N,k-1}$ είναι μια σταθερά της στατιστικής και $s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)}\sum_{i=1}^{k}(\delta_i - \bar{\delta})^2}$

Αφού οι υποθέσεις προέρχονται από εκπαίδευση πάνω στα ίδια σύνολα, οι μόνες διαφορές που παρατηρούνται στις υποθέσεις οφείλονται μόνο στους αλγορίθμους μάθησης.

Το πρόβλημα με τη προηγούμενη μέθοδο είναι πως εξαρτάται έντονα από τον τυχαίο διαχωρισμό σε δύο σύνολα. Αν τύχει το δείγμα που θα επιλεγεί για σύνολο εκπαίδευσης να είναι αρκετά αντιπροσωπευτικό του συνόλου ελέγχου, τα αποτελέσματα θα είναι υπερεκτιμημένα, ενώ το αντίθετο θα συμβεί αν δεν είναι αντιπροσωπευτικό. Μια μέθοδος που μειώνει το πρόβλημα είναι η διασταυρωμένη επικύρωση (cross-validation – CV). Σε αυτήν, η συλλογή χωρίζεται τυχαία σε k ξένα μεταξύ τους τμήματα ίσου (περίπου) μεγέθους (folds). Στην συνέχεια γίνονται k επαναλήψεις. Σε κάθε επανάληψη, ένα από τα τμήματα είναι το σύνολο ελέγχου και τα υπόλοιπα k-1 αποτελούν το σύνολο εκπαίδευσης. Αφού κληθεί ο αλγόριθμος μάθησης για κάθε τμήμα, η εκτιμώμενη τιμή του μέτρου είναι ο μέσος όρος του για όλα τα τμήματα (βλέπε Εικόνα 5). Με τον τρόπο αυτό, ακόμα κι αν σε κάποια τμήματα γίνεται υπερεκτίμηση, για μεγάλο k αναμένεται ισόποση σχεδόν υποεκτίμηση, με αποτέλεσμα ο μέσος όρος να είναι πιο κοντά στην πραγματική τιμή. Ειδική περίπτωση της διασταυρωμένης επικύρωσης είναι η εξαντλητική μέθοδος (leaveone-out), για την οποία τα σύνολα ελέγχου αποτελούνται από ένα μόνο στιγμιότυπο.

Μια άλλη παραλλαγή είναι η στρωματοποιημένη διασταυρωμένη επικύρωση (stratified CV), στην οποία τα τμήματα δε δημιουργούνται τελείως τυχαία, αλλά διατηρούν την ίδια σχεδόν κατανομή στιγμιοτύπων για κάθε κλάση με αυτή της αρχικής συλλογής.

5.5 ΕΚΠΑΙΔΕΥΣΗ ΜΕΣΩ ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Η εκπαίδευση των δικτύων επιτυγχάνεται με την χρήση κάποιων αλγορίθμων που καλούνται αλγόριθμοι εκπαίδευσης ή μάθησης (training or learning algorithms). Γενικά, υπάρχουν πολλοί διαφορετικοί αλγόριθμοι για την εκπαίδευση δικτύων και αρκετές παραλλαγές τους, οι οποίοι όμως μπορούν να ταξινομηθούν σε τρεις γενικές κατηγορίες: Επιβλεπόμενη (Supervised), Ενισχυτική (Reinforcement) και Μη-επιβλεπόμενη (Unsupervised).

- Στην επιβλεπόμενη μάθηση υποθέτουμε την παρουσία ενός δασκάλου κατά τη διάρκεια της διαδικασίας της εκπαίδευσης. Έτσι κάθε δείγμα που χρησιμοποιείται για την εκπαίδευση του δικτύου (δείγμα εκπαίδευσης - training pattern) αποτελείται από ένα δείγμα εισόδου (input pattern) και ένα δείγμα επιθυμητής εξόδου (desired or target output pattern). Κατά τη διάρκεια της εκπαίδευσης γίνεται σύγκριση μεταξύ της εξόδου που υπολογίζει το δίκτυο δίνοντας του σαν είσοδο το δείγμα εισόδου και της επιθυμητής εξόδου όπως αυτή καθορίζεται από το δείγμα επιθυμητής εξόδου, προκειμένου να καθοριστεί το λάθος. Το λάθος στη συνέχεια χρησιμοποιείται για να μεταβληθούν οι ελεύθερες παράμετροι του δικτύου (δηλαδή τα βάρη και τα κατώφλια) έτσι ώστε να βελτιωθεί η απόδοσή του. Η διαδικασία της διόρθωσης της τιμής των ελευθέρων παραμέτρων επαναλαμβάνεται έως ότου οι τιμές τους συγκλίνουν σε ένα σύνολο τιμών τέτοιο ώστε το δίκτυο να είναι ικανό να επιτελέσει το έργο για το οποίο εκπαιδευόταν.

- Στην ενισχυτική μάθηση υποθέτουμε και πάλι την ύπαρξη δασκάλου, αλλά σε αντίθεση με την επιβλεπόμενη μάθηση, δεν παρέχουμε στο δίκτυο τη σωστή απάντηση (δηλαδή το επιθυμητό διάνυσμα εξόδου). Αντί αυτού, δίνουμε μια ένδειξη για το αν η έξοδος που υπολογίζεται είναι σωστή ή λάθος. Το δίκτυο θα πρέπει, στη συνέχεια, να χρησιμοποιήσει αυτήν την πληροφορία, που του παρέχει ο δάσκαλος, προκειμένου να βελτιώσει την απόδοσή του. Τυπικά, δίνεται μια ανταμοιβή, ενισχύοντας τα βάρη των κόμβων που δίνουν σωστή απάντηση και μια τιμωρία ελαττώνοντας τις τιμές των βαρών σε αυτούς τους κόμβους που δίνουν λάθος απάντηση.

- Στην μη-επιβλεπόμενη μάθηση δεν υπάρχει κάποιος δάσκαλος για να εφοδιάσει το δίκτυο με την σωστή απάντηση, δηλαδή τα δείγματα εκπαίδευσης αποτελούνται μόνο από τα δείγματα εισόδου και δεν περιέχουν δείγματα επιθυμητής εξόδου. Στην περίπτωση αυτή, λοιπόν, το σύστημα πρέπει να μάθει ανακαλύπτοντας και προσαρμόζοντας τον εαυτό του σε κάποια δομικά χαρακτηριστικά των διανυσμάτων εισόδου. Αυτό γίνεται ανακαλύπτοντας κάποιες στατιστικές κανονικότητες και ομαδοποιήσεις των δειγμάτων εισόδου. Ένα τέτοιο είδος μάθησης επιτυγχάνεται με την ενίσχυση επιλεγμένων βαρών προκειμένου το διάνυσμα εξόδου να ταιριάζει σε κεντρικά πρωτότυπα δείγματα εκπαίδευσης που είναι αντιπροσωπευτικά ενός συνόλου από παρόμοια δείγματα.

5.6 ΤΟ ΛΟΓΙΣΜΙΚΟ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΚΑΙ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ

Στόχος είναι η γρήγορη ανάπτυξη ενός πρωτοτύπου συστήματος που θα δώσει άμεσα στοιχεία γνώσης ώστε να μπορέσουμε στη συνέχεια να προχωρήσουμε στη συγχώνευσή τους. Επομένως χρειαζόμαστε ένα λογισμικό που θα συγκεντρώνει μια πληθώρα αλγορίθμων εξόρυξης γνώσης.

Το πρόγραμμα που θα χρησιμοποιήσουμε για την επίλυση του προβλήματος είναι το WEKA (Waikato Environment for Knowledge Analysis). Το WEKA είναι ένα περιβάλλον ανάπτυξης εφαρμογών μηχανικής μάθησης και εξόρυξης γνώσης, το οποίο αναπτύχθηκε στο πανεπιστήμιο του Waikato στη Νέα Ζηλανδία. Είναι γραμμένο σε Java, ώστε να μπορεί να χρησιμοποιηθεί με όσο το δυνατόν περισσότερα λειτουργικά συστήματα, και διατίθεται ελεύθερα (συμπεριλαμβανομένου του πηγαίου κώδικα). Παρέχει ένα ευρύ σύνολο από υλοποιήσεις αλγορίθμων μηχανικής μάθησης (τόσο για κατηγοριοποίηση όσο και για συσταδοποίηση, clustering) καθώς και μηχανισμούς για προ-επεξεργασία δεδομένων και μετα-επεξεργασία αποτελεσμάτων.

Το αποτέλεσμα των αλγορίθμων εξόρυξης γνώσης περιλαμβάνει κανόνες συσχέτισης, συστάδες, σχήματα ταξινόμησης κλπ. και εκφράζεται με τη μορφή μιας δήλωσης γνώσης. Ο χρήστης έχει τη δυνατότητα να χρησιμοποιήσει τις υλοποιήσεις των αλγορίθμων είτε από τη γραμμή εντολών είτε από το γραφικό περιβάλλον το οποίο προσφέρει το WEKA, ενώ ο προγραμματιστής μπορεί να καλέσει τις υλοποιήσεις των αλγορίθμων από τα δικά του προγράμματα. Έτσι το WEKA μπορεί να λειτουργήσει σαν μια βιβλιοθήκη υλοποιήσεων αλγορίθμων μηχανικής μάθησης, που μπορεί να χρησιμοποιηθεί για την δημιουργία νέων προγραμμάτων. Επίσης, καθώς παρέχει μια πλήρη βιβλιοθήκη με κώδικα για αξιολόγηση αποτελεσμάτων (π.χ. για διασταυρωμένη επικύρωση), μπορούν πολύ εύκολα να συγκριθούν νέες μέθοδοι με ήδη υπάρχουσες.

5.7 ΜΕΤΡΑ ΑΞΙΟΛΟΓΗΣΗΣ ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ ΤΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Ο έλεγχος της ποιότητας της προσαρμογής των δεδομένων στους παραγόμενους κανόνες γίνεται με τον υπολογισμό και την εκτίμηση του σφάλματος. Υπάρχουν πολλά μέτρα εκτίμησης του σφάλματος στη βιβλιογραφία, εδώ όμως θα αναφερθούν τα βασικά από αυτά που έχουν και την πιο ευρεία βάση χρήσης. Τα πρώτα μέτρα αναφέρονται σε προβλήματα τύπου παλινδρόμησης και ταξινόμησης, ενώ τα τελευταία μόνο σε προβλήματα τύπου ταξινόμησης.

5.7.1. Σφάλμα πρόβλεψης

Το σφάλμα πρόβλεψης e_k για το παράδειγμα k είναι η απλή διαφορά ανάμεσα στην πραγματική μετρήσιμη τιμή της μεταβλητής y και στην προσαρμοσμένη από αλγόριθμο τιμή προσέγγισης w:

$$e_k = y_k - w_k$$

Το αντίστοιχο εκατοστιαίο σφάλμα πρόβλεψης είναι ο λόγος του σφάλματος προς την πραγματική τιμή:

$$pe_k = \frac{e_k}{y_k}$$

5.7.2 Μέση απόλυτη απόκλιση (Mean Absolute Deviation –MAD)

Η μέση απόλυτη απόκλιση για ένα σύνολο δεδομένων L είναι η μέση τιμή για την απόλυτη τιμή του σφάλματος πρόβλεψης σε όλα τα παραδείγματα του συνόλου L:

$$MAD = \frac{\sum_{k=1}^{n} |e_k|}{n}$$

όπου n το πλήθος των παραδειγμάτων στο σύνολο L.

5.7.3 Μέσο απόλυτο εκατοστιαίο σφάλμα (Mean Absolute Percentage Error)

Το μέσο απόλυτο εκατοστιαίο σφάλμα για ένα σύνολο δεδομένων L είναι η μέση τιμή του απόλυτου εκατοστιαίου σφάλματος πρόβλεψης σε όλα τα παραδείγματα του συνόλου L:

$$MAPE = \frac{\sum_{k=1}^{n} |pe_k|}{n}$$

5.7.4 Μέσο τετραγωνισμένο σφάλμα (Mean Squared Error –MSE)

Το μέσο τετραγωνισμένο σφάλμα είναι η μέση τιμή του τετραγώνου του σφάλματος πρόβλεψης σε όλα τα παραδείγματα του συνόλου L:

$$MSE = \frac{\sum_{k=1}^{n} \left| e_k^2 \right|}{n}$$

5.7.5 Μέσο τετραγωνικό σφάλμα (Root Mean Squared Error –RMSE)

Το μέσο τετραγωνικό σφάλμα είναι η ρίζα του αντίστοιχου μέσου τετραγωνισμένου σφάλματος MSE:

$$RMSE = \sqrt{\frac{\sum_{k=1}^{n} \left| e_k^2 \right|}{n}}$$

5.7.6 Το τυπικό σφάλμα ενός διαστήματος εμπιστοσύνης (Standard Error – SE)

Στα προβλήματα ταξινόμησης όπου η εξαρτημένη μεταβλητή Y είναι κατηγορική, το σφάλμα πρόβλεψης είναι πάντα ένας ακέραιος αριθμός. Σε όλες τις περιπτώσεις τέτοιων προβλημάτων υπολογίζεται το ποσοστό σφάλματος ê ως ο λόγος του αριθμού των παραδειγμάτων που ταξινομήθηκαν εσφαλμένα από τον αλγόριθμο προς το πλήθος αυτών:

$$\hat{e} = \frac{1_{\{y_k \neq w_k\}}}{n}$$

Αν e είναι το πραγματικό ποσοστό σφάλματος του αλγόριθμου, τότε με διάστημα εμπιστοσύνης 1-α ισχύει για το e:

$$\hat{e} - z_{\alpha/2} SE \leq e \leq \hat{e} + z_{\alpha/2} SE$$

Για 95% διάστημα εμπιστοσύνης, z0.025 = 1.96 και έτσι:

$$\hat{e} - 1.96 SE \leq e \leq \hat{e} + 1.96 SE$$

5.7.7 Πίνακας σύγχυσης και παράγωγα μέτρα εκτίμησης σφάλματος

Ειδικά στα προβλήματα τύπου ταξινόμησης όπου η μεταβλητή εξόδου μπορεί να είναι τύπου ποιότητας, να δέχεται δηλαδή διακριτές τιμές, κατασκευάζεται ο πίνακας σύγχυσης (confusion table) για να υποδείξει τα σφάλματα ταξινόμησης του εκάστοτε αλγόριθμου ανά τιμή της εξαρτημένης μεταβλητής.

Αν δηλαδή Y είναι η πραγματική μεταβλητή εξόδου με δυνατές τιμές στο σύνολο {y1, .., yk} και Y' η εκτίμηση αυτής που κατασκευάζει το μοντέλο του εκάστοτε αλγόριθμου με δυνατές τιμές στο σύνολο {y'1, .., y'k}, k ο αριθμός των διακριτών τιμών της Y και της Y', τότε ο πίνακας σύγχισης είναι η μήτρα:

$$CT = \begin{vmatrix} sum(y'_1 | Y = y_1) & \dots & sum(y'_k | Y = y_1) \\ \dots & \dots & \dots \\ sum(y'_1 | Y = y_k) & \dots & sum(y'_k | Y = y'_k) \end{vmatrix}$$

5.7.7.1 Το μέτρο των πραγματικά θετικών παραδειγμάτων

Τα πραγματικά θετικά (true positive) παραδείγματα για την κλάση i∈[1,..,k] ορίζονται ως ο αριθμός των παραδειγμάτων που ταξινομήθηκε σωστά από τον αλγόριθμο, δηλαδή:

$$tp_i = sum(y'_i | Y = y_i)$$

5.7.7.2 Το μέτρο των εσφαλμένα θετικών παραδειγμάτων

Τα εσφαλμένα θετικά παραδείγματα (false positive) για την κλάση i∈[1,..,k] είναι ο αριθμός των παραδειγμάτων που εκτιμήθηκαν ότι ανήκουν στην κλάση i ενώ δεν ανήκουν σε αυτήν:

$$fp_i = sum(y'_i | Y \neq y_i)$$

Η ακρίβεια (precision) για την κλάση i∈[1,..,k] ορίζεται ως ο λόγος των πραγματικά θετικών ως προς το άθροισμα πραγματικά θετικών και εσφαλμένα θετικών παραδειγμάτων:

$$prc_i = \frac{sum(y'_i | Y = y_i)}{sum(y'_i | Y = y_i) + sum(y'_i | Y \neq y_i)}$$

5.7.7.3 Το μέτρο της επανάκλησης

Η επανάκληση (recall) για την κλάση i∈[1,..,k] ορίζεται ως ο λόγος των πραγματικά θετικών ως προς τον αριθμό των συνολικά θετικών παραδειγμάτων:

$$rc_i = \frac{sum(y'_i | Y = y_i)}{sum(y'_i)}$$

Γενικά υπάρχει ένας συμβιβασμός ανάμεσα στις υψηλές τιμές του μέτρου ακρίβειας και στις επίσης υψηλές τιμές του μέτρου επανάκλησης.

5.7.7.4 Το μέτρο f

Το μέτρο f είναι ένας συνδυασμός της ακρίβειας και της επανάκλησης. Ορίζεται ως ο αρμονικός μέσος των δύο, και θεωρείται ένα καλό συγκεντρωτικό μέτρο για την ικανότητα ταξινόμησης της κλάσης i:

$$f_i = \frac{2 \cdot rc \cdot prc}{rc + prc}$$

ΚΕΦΑΛΑΙΟ 6: ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΟ ΛΟΓΙΣΜΙΚΟ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ WEKA

6.1 ΕΙΣΑΓΩΓΗ ΣΤΟ ΛΟΓΙΣΜΙΚΟ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ WEKA

Στόχος είναι η γρήγορη ανάπτυξη ενός πρωτοτύπου συστήματος που θα δώσει άμεσα στοιχεία γνώσης ώστε να μπορέσουμε στη συνέχεια να προχωρήσουμε στη συγχώνευσή τους. Επομένως χρειαζόμαστε ένα λογισμικό που θα συγκεντρώνει μια πληθώρα αλγορίθμων εξόρυξης γνώσης.

Το πρόγραμμα που θα χρησιμοποιήσουμε για την επίλυση του προβλήματος είναι το WEKA (Waikato Environment for Knowledge Analysis). Το WEKA είναι ένα περιβάλλον ανάπτυξης εφαρμογών μηχανικής μάθησης και εξόρυξης γνώσης, το οποίο αναπτύχθηκε στο πανεπιστήμιο του Waikato στη Νέα Ζηλανδία. Είναι γραμμένο σε Java, ώστε να μπορεί να χρησιμοποιηθεί με όσο το δυνατόν περισσότερα λειτουργικά συστήματα, και διατίθεται ελεύθερα (συμπεριλαμβανομένου του πηγαίου κώδικα).

Το πρόγραμμα Weka περιλαμβάνει 4 βασικές λειτουργίες:

- Simple CLI: Εκτέλεση εντολών από command window
- Explorer: Το περιβάλλον με τις βασικές λειτουργίες επεξεργασίας δεδομένων.
- Experimenter: Το περιβάλλον με δυνατότητες σύνταξης μεγαλύτερων πειραμάτων που περιλαμβάνουν τη συγκριτική μελέτη διαφορετικών μοντέλων.
- Knowledge Flow: Προσφέρει τη δυνατότητα ανάπτυξης σύνθετων μοντέλων με μορφή μπλοκ διαγράμματος.

Το WEKA είναι ένα λογισμικό που επιτρέπει τη λύση προβλημάτων μηχανικής μάθησης με τη χρήση ποικιλίας αλγορίθμων, συμπεριλαμβανομένων των νευρωνικών δικτύων. Επιλέχθηκε η χρήση του γιατί παρέχει ένα ευρύ σύνολο από υλοποιήσεις αλγορίθμων μηχανικής μάθησης (τόσο για κατηγοριοποίηση όσο και για συσταδοποίηση, clustering) καθώς και μηχανισμούς για προ-επεξεργασία δεδομένων και μετα-επεξεργασία αποτελεσμάτων. Περιέχει, λοιπόν, υλοποιημένες μεθόδους για:

- Προεπεξεργασία Δεδομένων
- Ταξινόμηση
- Συσταδοποίηση
- Εύρεση Κανόνων Συσχέτισης

Το WEKA είναι λογισμικό που επιτρέπει την επίδειξη της λύσης προβλημάτων μηχανικής μάθησης με τη χρήση ποικιλίας αλγορίθμων, συμπεριλαμβανομένων των νευρωνικών δικτύων.

Επιτρέπει την είσοδο δεδομένων με μία απλή μορφοποίηση, στην οποία πρώτα παρατίθενται οι τύποι δεδομένων (στήλες) ενός προβλήματος μάθησης και μετά παρατίθενται τα στιγμιότυπα δεδομένων (γραμμές). Η τελευταία στήλη χρησιμοποιείται για την κλάση (τάξη) ενός στιγμιοτύπου. Σε διάφορα σύνολα δεδομένων δεν είναι ασυνήθιστο να υπάρχουν και γραμμές σχολίων μέσα στην περιγραφή των δεδομένων.

Το αποτέλεσμα των αλγορίθμων εξόρυξης γνώσης περιλαμβάνει κανόνες συσχέτισης, συστάδες, σχήματα ταξινόμησης κλπ. και εκφράζεται με τη μορφή μιας δήλωσης γνώσης. Ο χρήστης έχει τη δυνατότητα να χρησιμοποιήσει τις υλοποιήσεις των αλγορίθμων είτε από τη γραμμή εντολών είτε από το γραφικό περιβάλλον το οποίο προσφέρει το WEKA, ενώ ο προγραμματιστής μπορεί να καλέσει τις υλοποιήσεις των αλγορίθμων από τα δικά του προγράμματα. Έτσι το WEKA μπορεί να λειτουργήσει σαν μια βιβλιοθήκη υλοποιήσεων αλγορίθμων μηχανικής μάθησης, που μπορεί να χρησιμοποιηθεί για την δημιουργία νέων προγραμμάτων. Επίσης, καθώς παρέχει μια πλήρη βιβλιοθήκη με κώδικα για αξιολόγηση αποτελεσμάτων (π.χ. για διασταυρωμένη επικύρωση), μπορούν πολύ εύκολα να συγκριθούν νέες μέθοδοι με ήδη υπάρχουσες.

Το WEKA έχει πλήρη on-line τεκμηρίωση, τόσο όσον αφορά στη χρήση του όσο και όσον αφορά στις δυνατές αλλαγές και στην επέκτασή του. Η εγκατάσταση του WEKA γίνεται με την απλή εκτέλεση ενός συμπιεσμένου αρχείου (μεγέθους περίπου 8 MB) που περιέχει το σύστημα, σύμφωνα και με τις αναλυτικές οδηγίες εγκατάστασης που δίνονται στην ιστοσελίδα του κατασκευαστή του. Πριν την εγκατάσταση του WEKA είναι σκόπιμο να έχει ολοκληρωθεί η εγκατάσταση της Java.

6.2 ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Η βασική μορφοποίηση των δεδομένων για επεξεργασία μέσω του προγράμματος Weka χαρακτηρίζεται ως τύπου ARFF. Τα αρχεία αυτά περιέχουν το σύνολο των δεδομένων και πρέπει να έχουν συγκεκριμένο format και να αποθηκεύονται με την επέκταση .arff. Το περιεχόμενο ενός τέτοιου αρχείου μπορεί να διαβαστεί χρησιμοποιώντας το notepad και φαίνεται ακολούθως.

Εισαγωγή ονόματος αρχείου δεδομένων

- @RELATION Training Set 1

Όνομα του προβλήματός μας

Εισαγωγή ονομάτων χαρακτηριστικών και ο τύπος τους (numeric χαρακτηριστικά = numeric, nominal χαρακτηριστικά = { })

- @ATTRIBUTE Instalat { }
- @ATTRIBUTE Zip code { }
- @ATTRIBUTE Capacity charge numeric
- @ATTRIBUTE Ac.Desc. { }
- @ATTRIBUTE In.Type { }
- @ATTRIBUTE Days { }
- @ATTRIBUTE C.D. { }
- @ATTRIBUTE Temp. numeric
- @ATTRIBUTE Hum. numeric
- @ATTRIBUTE W.Vel. numeric
- @ATTRIBUTE W.dir. numeric
- @ATTRIBUTE Sol numeric
- @ATTRIBUTE HDD numeric
- @ATTRIBUTE class Consumption numeric

Ονόματα χαρακτηριστικών και τύπος τους

Δήλωση κλάσεων προς αναγνώριση

Εισαγωγή αρχείου δεδομένων

Δήλωση δεδομένων και κλάσης που αντιστοιχούν

- @DATA
- 4000000081, 15234, 2.5, Ad15, In08, 65, A4f_a, 12, 320, 2, 227, 50, 238, 25

Εκτός από τη μορφή arff, το πρόγραμμα αναγνωρίζει και τη μορφή αρχείου “comma separated values” (csv).

6.3 ΔΗΜΙΟΥΡΓΙΑ ΤΟΥ ΣΩΜΑΤΟΣ ΕΚΠΑΙΔΕΥΣΗΣ ΚΑΙ ΑΞΙΟΛΟΓΗΣΗΣ

Για τη δημιουργία των δεδομένων εκπαίδευσης και αξιολόγησης επιλέξαμε 6000 καταμετρήσεις κατανάλωσης, δημιουργώντας 3 ενδεικτικά σετ δεδομένων εκπαίδευσης των 2000 καταμετρήσεων. Αυτό έγινε για να ανακαλύψουμε ποια χαρακτηριστικά επηρεάζουν κατά το μέγιστο την κατανάλωση φυσικού αερίου και για να καθορίσουμε ποιοι αλγόριθμοι είναι οι καταλληλότεροι για επεξεργασία των δεδομένων μας.

Τα επιλεγμένα dataset προς ανάλυση ονομάζονται "training set 1", "training set 2", "training set 3". Τα 2000 υποδείγματα (instances) που αποτελούν το κάθε αρχείο έχουν προκύψει από τη συγκέντρωση μετεωρολογικών και καταναλωτικών δεδομένων που περιγράφτηκαν στο προηγούμενο κεφάλαιο.

Τα χαρακτηριστικά (attributes) είναι είτε αριθμητικά (numeric) είτε ονομαστικά (nominal). Αριθμητικά ονομάζονται τα δεδομένα που είναι ποσοτικά δηλαδή λαμβάνουν αριθμητικές συνεχείς τιμές, ενώ τα ονομαστικά είναι ποιοτικά δεδομένα (π.χ. ονόματα, καθορισμένες ονομασίες). Παρακάτω παραθέτονται τα χαρακτηριστικά αυτά τα οποία έχουν αναλυθεί στο προηγούμενο κεφάλαιο.

Χαρακτηριστικά (Attributes)	Τύπος δεδομένων
1. Κωδικός καταναλωτή (Instalat.)	ονομαστικό (nominal)
2. Ταχυδρομικός κώδικας καταναλωτή (Zip code)	ονομαστικό (nominal)
3. Εγκατεστημένη Ισχύς (Capacity charge)	ονομαστικό (nominal)
4. Κατηγορία καταναλωτή (Ac.Desc.)	ονομαστικό (nominal)
5. Κατηγορία χρήσης φυσικού αερίου (In.Type)	ονομαστικό (nominal)
6. Διάστημα ημερών κατανάλωσης (Days)	αριθμητικά (numeric)
7. Κατηγορία Διμήνου (C.D.)	ονομαστικό (nominal)
8. Θερμοκρασία περιβάλλοντος (Temp.)	ονομαστικό (nominal)
9. Σχετική υγρασία (Hum.)	αριθμητικά (numeric)
10. Ταχύτητα ανέμου (W.Vel.)	αριθμητικά (numeric)
11. Διεύθυνση ανέμου (W.dir.)	αριθμητικά (numeric)
12. Ηλιακή ακτινοβολία (Sol.)	αριθμητικά (numeric)
13. Βαθμοημέρες θέρμανσης (HDD)	αριθμητικά (numeric)
14. Κατανάλωση Φυσικού Αερίου (Consumption)	αριθμητικά (numeric)

Το δέκατο τέταρτο χαρακτηριστικό, attribute (class) είναι το στοιχείο βάση του οποίου ταξινομούνται τα υποδείγματα και είναι αριθμητικό (numeric). Θα πρέπει επίσης να αναφερθεί πως δεν έχουμε καθόλου άγνωστες τιμές (missing values).

6.4 ΑΞΙΟΛΟΓΗΣΗ ΚΑΙ ΕΠΙΛΟΓΗ ΤΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Αρχικά θα εξετάζουμε 19 χαρακτηριστικά (attributes) για να παρατηρήσουμε ποια από αυτά είναι τα πιο σημαντικά και επηρεάζουν κατά το μέγιστο βαθμό την κατανάλωση φυσικού αερίου.

Θα πάμε στο Weka Explorer και στην καρτέλα με την ονομασία Select Attributes (επιλογή χαρακτηριστικών). Θα τρέξουμε κάποιους αλγόριθμους ώστε να ανακαλύψουμε ποια χαρακτηριστικά (attributes) είναι τα πιο σημαντικά και επηρεάζουν κατά το μέγιστο βαθμό την κατανάλωση φυσικού αερίου.

Από τους δώδεκα Αξιολογητές που διαθέτει το Weka στις βιβλιοθήκες του, μόνο οι τρεις μπορούν να επεξεργαστούν τα δεδομένα μας καθώς το στοιχείο που θέλουμε να προβλέψουμε, δηλαδή η κατανάλωση παίρνει αριθμητικές τιμές (αριθμητικό, numeric). Επομένως, θα τρέξουμε τους εξής αλγόριθμους επιλογής των σημαντικότερων χαρακτηριστικών.

1. Αξιολογητής Χαρακτηριστικού: SfsSubsetEval σε συνδυασμό με μέθοδο GreedyStepwise. Προσδιορίζει την αξία ενός μέρους των χαρακτηριστικών, λαμβάνοντας υπόψη την ατομική ικανότητα πρόβλεψης κάθε χαρακτηριστικού σε σχέση με τη διαφορά ανάμεσά τους.

2. Αξιολογητής Χαρακτηριστικού: ChiSquaredAttributeEval σε συνδυασμό με μέθοδο Ranker. Προσδιορίζει την αξία του χαρακτηριστικού υπολογίζοντας την αξία της chi-squared στατιστικής αναλόγως της κλάσης.

3. Αξιολογητής Χαρακτηριστικού: ReliefFAttributeEval σε συνδυασμό με μέθοδο Ranker. Προσδιορίζει την αξία του χαρακτηριστικού υπολογίζοντας επανειλημμένως δείγματα υποδειγμάτων (instances), λαμβάνοντας υπόψη το κάθε χαρακτηριστικό της σχετικότερου υποδείγματος της ίδιας ή διαφορετική κλάσεις.

Attribute Evaluator	Method
SfsSubsetEval	GreedyStepwise
ChiSquaredAttributeEval	Ranker
ReliefFAttributeEval	Ranker

Τρέξαμε τα δεδομένα και των τριών σετ δεδομένων εκπαίδευσης με κάθε μέθοδο και τα αποτελέσματα παρατίθενται στον παρακάτω πίνακα.

	SfsSubsetEval	ChiSquaredAttributeEval	ReliefFAttributeEval
Training set 1	Επιλεγμένα χαρακτηριστικά: 1,2,4,5,13,14	Επιλεγμένα χαρακτηριστικά: 1,2,3,4,5,13,14	Επιλεγμένα χαρακτηριστικά: 1,2,3,4,5,13,14
Training set 2	Επιλεγμένα χαρακτηριστικά: 1,2,4,5,13,14	Επιλεγμένα χαρακτηριστικά: 1,2,4,5,13,14	Επιλεγμένα χαρακτηριστικά: 1,2,4,5,13,14
Training set 3	Επιλεγμένα χαρακτηριστικά: 1,2,4,5,13,14	Επιλεγμένα χαρακτηριστικά: 1,2,4,5,6,13,14	Επιλεγμένα χαρακτηριστικά: 1,2,4,5,6,13,14

Παρατηρούμε ότι τα κυριότερα χαρακτηριστικά, όπως έδειξαν οι αλγόριθμοι αξιολόγησης είναι τα εξής:

Αριθμός Χαρακτηριστικού	Χαρακτηριστικό (Attribute)
1	Κωδικός καταναλωτή (Instalat.)
2	Ταχυδρομικός κώδικας καταναλωτή (Zip code)
4	Κατηγορία καταναλωτή (Ac.Desc.)
5	Κατηγορία χρήσης φυσικού αερίου (In.Type)
13	Βαθμοημέρες θέρμανσης (HDD)
14	Κατανάλωση Φυσικού Αερίου (Consumption)

Επομένως μπορούμε να αφαιρέσουμε τα υπόλοιπα οκτώ χαρακτηριστικά από το δείγμα καθώς δεν επηρεάζουν σε μεγάλο βαθμό την κατανάλωση. Δηλαδή τα υπόλοιπα attributes δεν μας δίνουν κάποια αξιοποιήσιμη πληροφορία και επομένως μπορούν να παραλειφθούν. Αυτό θα γίνει χρησιμοποιώντας την εντολή Remove στην καρτέλα με την ονομασία Preprocess, επιλέγοντας και διαγράφοντας τα χαρακτηριστικά που πρέπει να παραληφθούν.

Εφόσον το επιθυμούμε, για να βελτιώσουμε το αποτέλεσμα της πρόβλεψης (σε μικρό βαθμό όμως) μπορούμε να χρησιμοποιήσουμε τα υπόλοιπα δεδομένα.

6.5 ΕΠΙΛΟΓΗ ΑΛΓΟΡΙΘΜΩΝ

Αρχικά θα εξετάζουμε όλους τους αλγόριθμους που υπάρχουν στις βιβλιοθήκες του WEKA για να ανακαλύψουμε ποιοι από αυτούς είναι σε θέση να δώσουν ικανοποιητικά αποτελέσματα πρόβλεψης της κατανάλωσης φυσικού αερίου.

Διάφορες προσεγγίσεις με σκοπό την αυτόματη εξαγωγή κανόνων από σώμα δεδομένων έχουν αναπτυχθεί με διαφορετική συμπεριφορά και απόδοση η καθεμία σχετικά με το είδος των στοιχείων προς εκπαίδευση, ή την συνάφεια των χαρακτηριστικών με την προς μοντελοποίηση κλάση, ή τέλος με το πρόβλημα των ελλιπών δεδομένων. Για το λόγω αυτό επιλέξαμε να εξετάσουμε τα δεδομένα μας με αλγόριθμους διαφορετικών οικογενειών όσων αφορά το τρόπο χειρισμού των δεδομένων από τα οποία θα προκύψουν οι κανόνες των μοντέλων για την εκτίμηση των προσωδιακών παύσεων.

Θα πάμε στο Weka Explorer και στην καρτέλα με την ονομασία Classify (κατηγοριοποίηση). Θα τρέξουμε όλους αλγόριθμους και για τα τρία ενδεικτικά σετ δεδομένων εκπαίδευσης. Τα επιλεγμένα dataset προς ανάλυση ονομάζονται “training set 1”, “training set 2”, “training set 3”.

Από τους εβδομήντα Ταξινομητές που διαθέτει το Weka στις βιβλιοθήκες του, μόνο οι έντεκα μπορούν να επεξεργαστούν τα δεδομένα μας καθώς το στοιχείο που θέλουμε να προβλέψουμε, δηλαδή η κατανάλωση, παίρνει αριθμητικές τιμές (αριθμητικό, numeric). Αρκετοί αλγόριθμοι δεν τρέχουν διότι το class attribute είναι numeric και όχι nominal. Επομένως, θα τρέξουμε τους εξής αλγόριθμους επιλογής των σημαντικότερων χαρακτηριστικών και για τα τρία σετ δεδομένων εκπαίδευσης.

Επιλεχθέντες Αλγόριθμοι Weka
M5P
REPTree
Linear Regression
Multilayer Perceptor
SMOreg
IBk
KStar
Bagging
Regression By Discretization
Decision Table
M5Rules

Ακολουθεί μια παρουσίαση των επιλεχθέντων αλγορίθμων και της κατηγορίας (οικογένειας) που ανήκουν.

6.6 ΠΑΡΟΥΣΙΑΣΗ ΕΠΙΛΕΧΘΕΝΤΩΝ ΑΛΓΟΡΙΘΜΩΝ WEKA

Παρακάτω παρουσιάζονται οι αλγόριθμοι του WEKA που θα χρησιμοποιήσουμε.

6.6.1 Αλγόριθμοι των k πιο κοντινών γειτόνων (Αλγόριθμοι IBk, KStar)

Η μέθοδος των κοντινότερων γειτόνων βρίσκεται ανάμεσα στις πιο δημοφιλείς μεθόδους ταξινόμησης. Αντιπροσωπεύει την παλαιότερη (μη-παραμετρική) γενιά μεθόδων που προτάθηκαν για αυτό το πρόβλημα και έχει ενδελεχώς εφαρμοστεί στο πεδίο της στατιστικής και ειδικά σε αυτό της αναγνώρισης προτύπων. Τελευταία έχει ανανεωθεί το ενδιαφέρον για αυτή στην κοινότητα της μηχανικής εκμάθησης.

Παρόλη τη βασική της απλότητα και το γεγονός ότι πολλές εναλλακτικές προηγμένες τεχνικές έχουν αναπτυχθεί μετά την πρώτη της εμφάνιση, η μέθοδος των κοντινότερων γειτόνων παραμένει ακόμα η πιο επιτυχημένη για πολλά προβλήματα ταξινόμησης.

Αυτός ο αλγόριθμος ανήκει σε μια κατηγορία αλγόριθμων που βασίζονται στα στιγμιότυπα (instances) του ιστορικού συνόλου δεδομένων για να προσεγγίσουν τη λύση σε ένα νέο στιγμιότυπο. Ο αλγόριθμος αυτός, αντίθετα με άλλους, δεν κατασκευάζει ένα μοντέλο από τα δεδομένα εκπαίδευσης, για να το εφαρμόσει στα δεδομένα εξέτασης για την ταξινόμηση ονοματικών (nominal) ή την πρόβλεψη συνεχών (continiues) κατηγοριών.

Ο αλγόριθμος αυτός αναθέτει σε ένα αντικείμενο άγνωστης κλάσης την πλειοψηφική κλάση που προκύπτει από k αντικείμενα εγνωσμένης κλάσης κοντύτερα σε αυτό. Επομένως δεν κατασκευάζει αρχικά κάποιο μοντέλο. Αυτού του είδους η μάθηση καλείται "μάθηση βασισμένη-σε-περίπτωση" (Instance Based learning, IB), ή "οκνηρή εκμάθηση" (Lazy Learning).

Η έννοια της εγγύτητας συνήθως εκφράζεται από ένα κριτήριο απόστασης στον ευκλείδειο χώρο με άξονες τις μεταβλητές εισόδου των αντικειμένων (τις ιδιότητες). Στην πλειονότητα των περιπτώσεων εφαρμογής του αλγόριθμου σε σχετικά περιορισμένο όγκο δεδομένων η απλή ευκλείδεια απόσταση προτιμάται λόγω απλότητας υλοποίησης.

Αν: x είναι το νέο στιγμιότυπο του οποίου η κλάση ζητείται να προσδιοριστεί,
x_v είναι το ν-ιοστό «κοντινό» στο x στιγμιότυπο από το σύνολο στιγμιότυπων γνωστής κλάσης,
μ είναι ο αριθμός ιδιοτήτων εισόδου των στιγμιότυπων x και x_v,

τότε η ευκλείδεια απόσταση $\Delta(x,x_v)$ μεταξύ τους ορίζεται ως:

$$\Delta(x, x_v) = \left[\sum (x_i - x_{vi})^2 \right]^{1/2}$$

Οι βασισμένοι-σε-περίπτωση μέθοδοι εκμάθησης αποθηκεύουν απλά τα παραδείγματα κατάρτισης και αναβάλλουν τη γενίκευση (που χτίζει ένα μοντέλο) έως ότου πρέπει να ταξινομηθεί μια νέα περίπτωση ή να γίνει μια πρόβλεψη. (Αυτό εξηγεί ένα άλλο όνομα για τις μεθόδους IB - οκνηρή εκμάθηση - δεδομένου ότι αυτές οι μέθοδοι καθυστερούν την επεξεργασία έως ότου πρέπει να ταξινομηθεί μια νέα περίπτωση). Το πρότυπο που χτίζεται με την IB διαδικασία δεν είναι ένα γενικό μοντέλο που χρησιμοποιεί όλα τα στοιχεία κατάρτισης, αλλά μάλλον ένα τοπικό πρότυπο που περιλαμβάνει μόνο μερικές από τις περιπτώσεις. Οι μέθοδοι IB χρησιμοποιούνται τόσο για ταξινόμηση όσο και για παλινδρόμηση (regression) σε συνεχή δεδομένα.

Οι πιο σημαντικές μέθοδοι αυτής της κατηγορίας είναι η μέθοδος κοντινών γειτόνων (nearest neighbors), η τοπικά σταθμισμένη παλινδρόμηση (locally weighted regression), και ο συλλογισμός κατά-περίπτωση (case-base reasoning). Όσον αφορά την παρούσα διατριβή από την ομάδα αυτών των αλγορίθμων χρησιμοποιήσαμε την K-περιπτώσεις μάθηση (Ibk), η οποία αναπαριστά την αποκτηθείσα γνώση ως συλλογή περιπτώσεων εκπαίδευσης ή αλλιώς στιγμιοτύπων (instances). Είναι μια μορφή μάθησης με επόπτευση (supervised learning) από τα στιγμιότυπα εκπαίδευσης. Κρατά μια πλήρη μνήμη των στιγμιοτύπων κατάρτισης και ταξινομεί τις νέες περιπτώσεις χρησιμοποιώντας τις πιο παρόμοιες περιπτώσεις κατάρτισης. Μια νέα περίπτωση ταξινομείται έπειτα από την εύρεση της περίπτωσης με την υψηλότερη ομοιότητα και την χρησιμοποίηση της κατηγορίας της ως πρόβλεψη.

Για αυτό το λόγο, ο IBk χαρακτηρίζεται από μια πολύ χαμηλή προσπάθεια εκπαίδευσης από τα δεδομένα. Αυτό οδηγεί από την μια μεριά σε ανάγκη για μεγάλο χώρο αποθήκευση λόγο του ότι διατηρεί όλα τα στιγμιότυπα κατάρτισης στην μνήμη. Επιπλέον, λαμβάνει χώρα η σύγκριση μιας νέας υπόθεσης με όλες τις ήδη υπάρχουσες, το οποίο οδηγεί σε ένα υψηλό κόστος υπολογισμού για την ταξινόμηση. Ο αλγόριθμος αυτός χρησιμοποιεί ένα μέτρο απόστασης για να προβλέψει, σαν κατηγορία του στιγμιότυπου εξέτασης, την κατηγορία του κοντινότερου στιγμιότυπου εκπαίδευσης που θα βρει. Η συνάρτηση ομοιότητας (similarity function) που χρησιμοποιεί ο IBk για k στιγμιότυπα είναι,

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

Όσον αφορά την περίπτωση συνεχών χαρακτηριστικών η συνάρτηση f παίρνει την μορφή,

$$f(x_i, y_i) = (x_i - y_i)^2$$

όπου κάθε στιγμιότυπο περιγράφεται από n χαρακτηριστικά.

Η IBk μέθοδος ταξινόμησης είναι όμοια με την κοντινότερου-γείτονα εκτός από το γεγονός ότι κανονικοποιεί τα χαρακτηριστικά των στιγμιοτύπων, επεξεργάζεται τις περιπτώσεις επαυξητικά, και έχει μια απλή πολιτική για την ανοχή των ελλειπουσών τιμών (missing values). Επιπλέον, ο IBk σώζει μόνο τις λανθασμένα-ταξινομημένες περιπτώσεις και υιοθετεί μια πολιτική "αναμονή και βλέπει" για την συλλογή των στοιχείων σχετικά με τον καθορισμό του ποια από τα χαρακτηριστικά αναμένεται να αποδώσουν καλύτερα κατά τη διάρκεια της ταξινόμησης. Το μόνο όριο στην πολυπλοκότητα αυτής της μεθόδου αυτόματης εκμάθησης από δεδομένα είναι το όριο στη δυνατότητα να αποθηκευτούν οι περιπτώσεις. Το WEKA διαθέτει στη βιβλιοθήκη αλγορίθμων τους αλγορίθμους IBk και Kstar.

6.6.2 Παλινδρόμηση (Αλγόριθμοι Linear Regression, SMOreg, M5Rules)

Αν Y είναι η εξαρτημένη μεταβλητή εξόδου και υπάρχουν k ανεξάρτητες μεταβλητές εισόδου, το μοντέλο που αποδίδει τη σχέση τους έχει την εξής μορφή :

$$Y_i = g(x_{i1}, x_{i2}, ..., x_{ik}) + \varepsilon_i$$

όπου $(x_{i1}, x_{i2}, ..., x_{ik})$ ένα σύνολο προκαθορισμένων τιμών των ανεξάρτητων μεταβλητών, Y_i η αντίστοιχη τυχαία τιμή της εξαρτημένης μεταβλητής που έχει μέση τιμή $E(Y) = g(x_{i1}, x_{i2}, ..., x_ik)$, και ε_i είναι μία τυχαία μεταβλητή με μέση τιμή μηδέν, που δίνει το τυχαίο λάθος όταν γίνεται εκτίμηση ή πρόβλεψη της Y_i από τη μέση τιμή της, δηλαδή την τυχαία μεταβλητότητα της Y_i γύρω από τη μέση τιμή της λόγω της ύπαρξης μη ελεγχόμενων "τυχαίων" παραγόντων.

Αυτό που επιθυμούμε είναι το τυχαίο σφάλμα της πρόβλεψης:

$$D = Y - g(x_{i1}, x_{i2}, ..., x_{ik})$$

να είναι με κάποια στατιστική έννοια ελάχιστο. Συνήθως σαν τέτοιο χρησιμοποιείται η μέση τιμή του τετραγώνου του, E(D2), το καλούμενο μέσο τετραγωνισμένο λάθος (MSE). Αυτό αποδεικνύεται ότι ελαχιστοποιείται αν διαλέξουμε σαν $g(x_{i1}, x_{i2}, ..., x_{ik})$ τη μέση υπό συνθήκες τιμή του Y, $E(Y \mid X_1, X_2, ..., X_k)$. Έτσι με δεδομένο σετ παρατηρήσεων X1 = xi1, Xk = xik, η βέλτιστη πρόβλεψη για την τ.μ. Y είναι

$$g(x_{i1}, x_{i2}, ..., x_{ik}) = E(Y \mid X_1 = x_{i1}, X_2 = x_{i2}, ..., X_k = x_{ik}).$$

Βέβαια αυτή η πρόβλεψη θα διαφέρει από την πραγματική τιμή της μέτρησης Y = yi κατά yi - g(xi1, xi2, ..., xik). Επειδή όμως στην πράξη η εύρεση κατανομών υπό συνθήκες και των μέσων τιμών τους είναι δύσκολη, χρησιμοποιούμε συγκεκριμένες μορφές για τη συνάρτηση g, όπως

$$g(x_1, x_2, ..., x_k) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

και προσδιορίζουμε τους άγνωστους παράγοντες β0, β1, ..., βk έτσι ώστε να ελαχιστοποιούν το μέσο τετραγωνικό σφάλμα της πρόβλεψης E(D2). Η συνάρτηση g(x1, x2, ..., xk) λέγεται συνάρτηση παλινδρόμησης (regression function). Οι άγνωστοι συντελεστές β0, β1,…, βk λέγονται παράμετροι της συνάρτησης παλινδρόμησης. Άλλες συνήθεις μορφές της συνάρτησης παλινδρόμησης είναι:

$$g(x1, x2, ..., xk) = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 ... + \beta_k x_k^k$$

$$g(x1, x2, ..., xk) = \beta_0 + \beta_1 e^x + \beta_2 e^{2x} .. + \beta_k e^{kx}$$

Η επιλογή της κατάλληλης συνάρτησης προκύπτει παρατηρώντας τη γραφική απεικόνιση των ζευγών (x1, y1), (x2, y2), ..., (xn, yn) των μετρήσεων στο δισδιάστατο επίπεδο. Τα χρήσιμα αυτά διαγράμματα λέγονται διαγράμματα διασκόρπισης (scatter diagrams).

Η πιο απλή περίπτωση παλινδρόμησης είναι η γραμμική παλινδρόμηση. Η υπόθεση πίσω από αυτό τον τύπο παλινδρόμησης είναι ότι η εξαρτημένη μεταβλητή Y είναι γραμμικός συνδυασμός των εξαρτημένων μεταβλητών X_1, X_2, …, X_k, δηλαδή

$$Y=\alpha+\beta(X)+\varepsilon$$

Το σφάλμα της παλινδρόμησης ε είναι μία τυχαία μεταβλητή με την εξής ιδιότητα:

$$\mu_\varepsilon = 0$$

6.6.3 Δέντρα Απόφασης (Αλγόριθμοι REPTree, M5P, Decision Table)

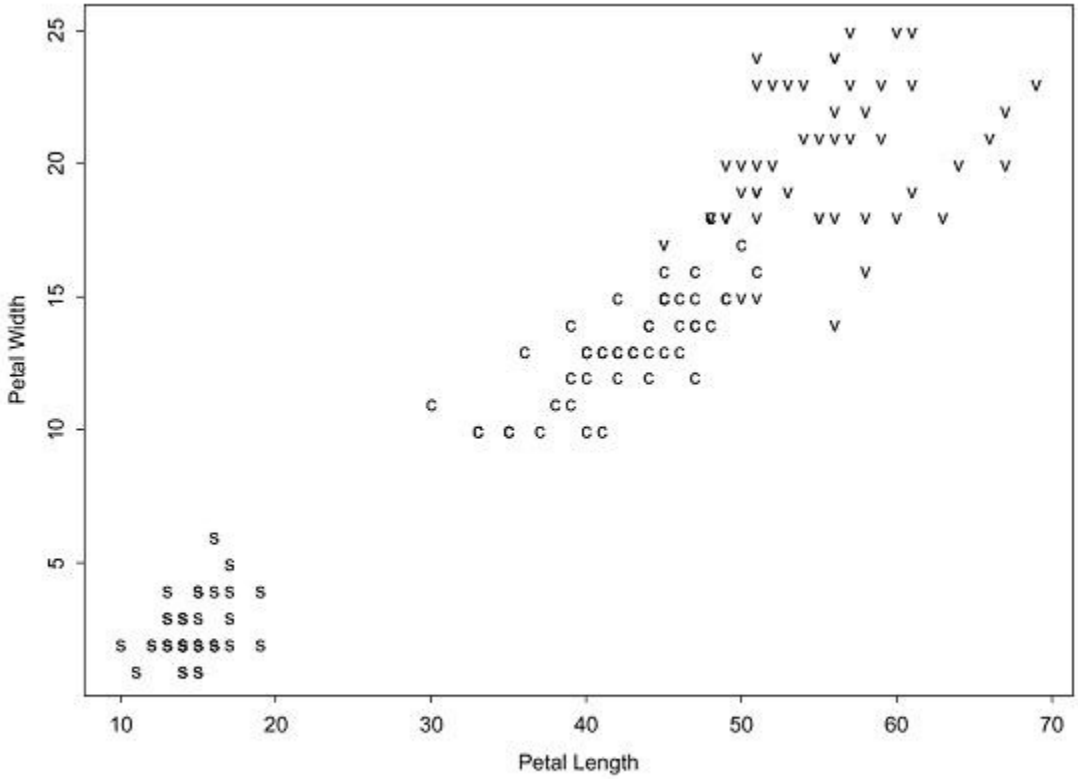
Τα δέντρα απόφασης είναι προϊόντα της επαγωγικής διαδικασίας εκμάθησης. Ένα δέντρο απόφασης επάγεται από ένα σύνολο εκμάθησης, που αποτελείται από αντικείμενα. Κάθε αντικείμενο περιγράφεται πλήρως από ένα σύνολο ιδιοτήτων και από μια ετικέτα κλάσης (class label). Οι ιδιότητες μπορούν να είναι διάταξης ή κατηγορίας. Για παράδειγμα, οι ακέραιες τιμές είναι διάταξης, ενώ οι δυαδικές είναι κατηγορίας. Ο στόχος ενός αλγόριθμου μηχανικής μάθησης αναφορικά με την κατασκευή ενός μοντέλου από ένα σύνολο δεδομένων, είναι η εύρεση ή η μεγαλύτερη δυνατή προσέγγιση της πραγματικής αντιστοίχισης ανάμεσα στο σύνολο των ιδιοτήτων και στην ετικέτα κλάσης. Ένα σύνολο εκπαίδευσης ελεύθερο θορύβου είναι αυτό στο οποίο όλα τα αντικείμενα προκύπτουν με τη χρήση της αντιστοίχισης αυτής.

Ένα δέντρο απόφασης περιέχει έναν ή περισσότερους εσωτερικούς κόμβους και έναν ή περισσότερους κόμβους-φύλλα. Όλοι οι εσωτερικοί κόμβοι έχουν δύο ή περισσότερους κόμβους-παιδιά. Όλοι οι μη τερματικοί κόμβοι περιέχουν τεμαχισμούς στους οποίους ελέγχεται η τιμή μιας μαθηματικής ή μιας λογικής έκφρασης των ιδιοτήτων. Οι κορυφές από έναν εσωτερικό κόμβο t στους κόμβους-παιδιά του είναι σημασμένες με τα ξεχωριστά αποτελέσματα της δοκιμασίας στον κόμβο t. Κάθε κόμβος-φύλλο έχει μια ετικέτα κλάσης προσαρτημένη. Ο αριθμός των κλάσεων είναι πεπερασμένος. Ένας κόμβος-φύλλο M ονομάζεται καθαρός όταν όλα τα δείγματα του συνόλου εκμάθησης κάτω από τον κόμβο αυτό ανήκουν στην ίδια κλάση.

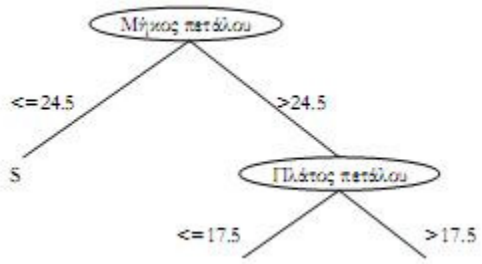
Η κατασκευή ενός δέντρου από το σύνολο εκμάθησης είναι μια επαγωγική διαδικασία. Η περισσότερο χρησιμοποιούμενη κατηγορία επαγωγικής εκμάθησης για την κατασκευή δέντρων απόφασης είναι η επαγωγική κατασκευή δέντρων απόφασης από την κορυφή προς τα κάτω (Top Down Induction of Decision Trees –TDIDT). Αρχίζοντας με ένα κενό δέντρο και με το σύνολο εκπαίδευσης, ο ακόλουθος αλγόριθμος εφαρμόζεται μέχρι κανένας τεμαχισμός να μην είναι εφικτός:. Αν όλα τα παραδείγματα εκπαίδευσης στον τρέχοντα κόμβο t ανήκουν στην κατηγορία (τάξη) c, δημιούργησε ένα κόμβο-παιδί του t και σταμάτησε. Αλλιώς, βαθμολόγησε κάθε ένα από τα σύνολα των πιθανών τεμαχισμών S, χρησιμοποιώντας ένα κριτήριο καταλληλότητας. Επέλεξε τον βέλτιστο τεμαχισμό s* σαν τον έλεγχο στον τρέχοντα κόμβο, και δημιούργησε τόσους κόμβους-παιδιά όσα ξεχωριστά αποτελέσματα παράγονται από τον τεμαχισμό s*. Προσάρτησε ετικέτες στις κορυφές ανάμεσα στον κόμβο-γονέα και στους κόμβους-παιδιά με τα αποτελέσματα του s*, και κατένειμε τα δεδομένα εκπαίδευσης χρησιμοποιώντας το s* στους κόμβους-παιδιά.

Ένα απλό παράδειγμα εφαρμογής του αλγόριθμου είναι το πρόβλημα της ταξινόμησης του άνθους του λουλουδιού της ίριδας σε ένα από τα τρία γνωστά είδη: iris setosa, iris versicolor, iris virginica. Ο πρώτος που έθεσε το πρόβλημα ήταν ο στατιστικός R.A. Fisher, σε μία μελέτη του γύρω από τη χρήση πολλαπλών ανεξάρτητων μεταβλητών στα προβλήματα ταξινόμηση. Το πρόβλημα περιέχει 150 παραδείγματα ανθέων του λουλουδιού, τέσσερις ανεξάρτητες μεταβλητές για το μήκος και πλάτος των πετάλων και των σεπάλων του άνθους, και μία εξαρτημένη που αντιστοιχεί στο σωστό είδος στο οποίο ανήκει κάθε άνθος από τα παραδείγματα.

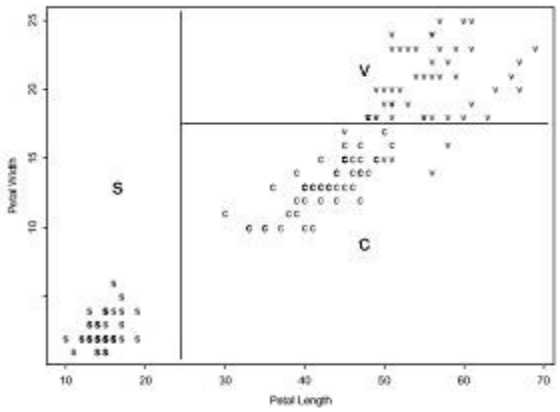
Η παρακάτω εικόνα παρουσιάζει την κατανομή των τάξεων κατά μήκος των αξόνων των δύο πρώτων ανεξάρτητων μεταβλητών, που όπως θα δειχθεί από το δέντρο που κατασκεύασε ο αλγόριθμος είναι αυτές που επηρεάζουν περισσότερο το πρόβλημα. Οι τάξεις συμβολίζονται με ένα μικρό γράμμα του αγγλικού αλφάβητου, c για το είδος versicolor, v για το είδος virginica, s για το είδος setosa. Το είδος setosa είναι γραμμικά διαχωρίσιμο, τα άλλα δύο όμως όχι.



Κατανομή των κλάσεων στο παράδειγμα ταξινόμησης του άνθους της ίριδας



Το δέντρο απόφασης του προβλήματος ταξινόμησης της ίριδας



Ο τεμαχισμός του προβλήματος της ίριδας από το δέντρο απόφασης

Το δέντρο που κατασκεύασε ο αλγόριθμος παρουσιάζεται στην πρώτη εικόνα. Μόνο δύο από τις τέσσερις ιδιότητες του συνόλου δεδομένων ήταν αρκετές για να ταξινομήσουν με ακρίβεια τα παραδείγματα σε μία από τις τρεις διαθέσιμες κλάσεις της εξαρτημένης μεταβλητής. Στη δεύτερη εικόνα φαίνεται ο δισδιάστατος χώρος που ορίζουν οι δύο αυτές ιδιότητες (μήκος και πλάτος πετάλου) και πως κάθε κόμβος του δέντρου τεμαχίζει αυτό τον χώρο στα δύο, ώστε να γίνεται ο καλύτερος δυνατός διαχωρισμός κλάσεων ανάμεσα στα φύλλα του δέντρου.

Αρκετοί δημοφιλείς αλγόριθμοι είναι βασισμένοι στην TDIDT, όπως ο ID3 και ο C4.5.

Ο M5 αλγόριθμος είναι ένα υβρίδιο διασταύρωσης δέντρου απόφασης και γραμμικής παλινδρόμησης. Δουλεύει αρχικά σαν δέντρο απόφασης τμηματοποιώντας το διανυσματικό χώρο που ορίζουν οι ιδιότητες σε φύλλα, σε κάθε ένα από τα οποία εκπαιδεύεται ένα ξεχωριστό μοντέλο γραμμικής παλινδρόμησης, με το υποσύνολο των παραδειγμάτων του συνόλου εκπαίδευσης που αντιστοιχούν σε εκείνο το φύλλο.

6.6.4 Τεχνητά Νευρωνικά Δίκτυα (Αλγόριθμοι Multilayer Perceptor)

Τα τεχνητά νευρωνικά δίκτυα είναι ένας κλάδος της τεχνητής νοημοσύνης και προήλθαν από την επιστημονική προσπάθεια να μοντελοποιηθεί η λειτουργία των νευρώνων του εγκεφάλου. Αποτελούνται από ένα δίκτυο από κόμβους που λέγονται νευρώνες και διασυνδέσεις μεταξύ νευρώνων που ονομάζονται συνάψεις. Ένα βιολογικό νευρωνικό δίκτυο είναι μία μη γραμμική υψηλής παραλληλοποίησης μηχανή που χαρακτηρίζεται από αξιοπιστία και ανοχή στα λάθη.

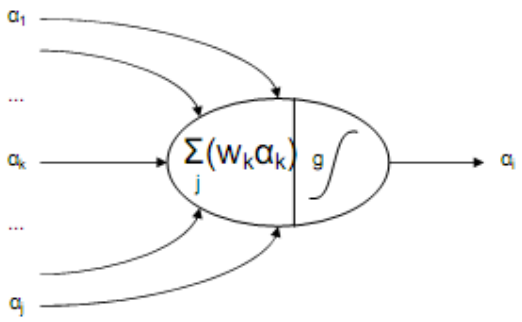
Ένας βιολογικός νευρώνας μπορεί να έχει μέχρι και 10000 διαφορετικές εισόδους, και μπορεί να στέλνει την έξοδό του (την παρουσία ή απουσία μιας βραχυπρόθεσμης διέγερσης) σε πολλούς άλλους νευρώνες. Οι νευρώνες είναι διασυνδεδεμένοι σε τρισδιάστατα πρότυπα. Ο βιολογικός νευρώνας είναι κατά πολύ πιο αργός σε σχέση με μία λογική πύλη σε σιλικόνη. Αλλά η ταχύτητα της επαγωγικής διαδικασίας παραγωγής μιας απόφασης είναι πολύ ταχύτερη σε βιολογικά δίκτυα από ότι στους ταχύτερους ηλεκτρονικούς υπολογιστές.

Ένας τεχνητός νευρώνας αi όπως παρουσιάζεται στην εικόνα 3-7, επιτελεί τα αμέσως ακόλουθα:

- Δέχεται ένα αριθμό από j εισόδους α_k (είτε από πραγματικά δεδομένα εισόδου, είτε από εξόδους από άλλους νευρώνες του νευρωνικού δικτύου). Κάθε είσοδος έρχεται διαμέσου μιας σύναψης και έχει ένα μέτρο ισχύος (ένα βάρος).

- Αθροίζει τα γινόμενα βάρους και τιμής εισόδου για κάθε είσοδο

- Παράγει μία έξοδο α_i, περνώντας το άθροισμα διαμέσου μιας συνάρτησης ενεργοποίησης g (βηματική συνάρτηση με μία τιμή κατωφλίου, στην πιο απλή εκδοχή της).



Ο τεχνητός νευρώνας

Τα ΤΝΔ είναι ένα είδος πολύ-επεξεργαστικού υπολογιστικού συστήματος, με τα εξής χαρακτηριστικά:

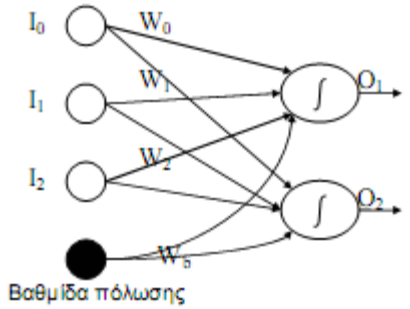
- Απλά υπολογιστικά στοιχεία
- Μεγάλο βαθμό διασύνδεσης μεταξύ τους
- Απλά βαθμωτά μηνύματα
- Προσαρμοστική διασύνδεση μεταξύ των στοιχείων

Διακρίνονται αρκετές τοπολογίες νευρωνικών δικτύων, κάθε μία από τις οποίες έχει τα δικά της χαρακτηριστικά.

6.6.4.1 Τοπολογία perceptron

Η περίπτωση του perceptron υπήρξε μία από τις πιο ενθουσιώδεις εξελίξεις κατά τις πρώιμες ημέρες της αναγνώρισης προτύπων. Αποτελείται από ένα δίκτυο στοιχειωδών επεξεργαστών που μπορούν να μάθουν να αναγνωρίζουν και να ταξινομούν πρότυπα αυτόνομα. Οι επεξεργαστές αυτοί είναι απλά στοιχεία στοιχισμένα σε ένα επίπεδο. Το κλασσικό perceptron ενός επιπέδου, όταν του δοθούν δύο κλάσεις προτύπων, προσπαθεί να βρει ένα γραμμικό όριο διαχωρισμού ανάμεσα στις δύο κλάσεις. Αν τα δύο σύνολα προτύπων είναι όντως γραμμικά διαχωρίσιμα, τότε ο αλγόριθμος του perceptron εγγυάται να βρει το ν-διάστατο όριο διαχωρισμού ανάμεσα στις δύο κλάσεις σε ένα πεπερασμένο αριθμό βημάτων. Εντούτοις αν ο χώρος των προτύπων δεν είναι γραμμικά διαχωρίσιμος, τότε το perceptron θα αποτύχει.

Ένας νευρώνας αρκεί για το βασικό κορμό ενός απλού perceptron. Επίσης μία από τις εισόδους πρέπει να είναι πάντα ενεργοποιημένη (βαθμίδα πόλωσης – bias unit), ώστε να υπάρχει μία ελάχιστη διέγερση για την ενεργοποίηση της εξόδου. Ένα παράδειγμα perceptron με 3 εισόδους και δύο νευρώνες παρουσιάζεται στην παρακάτω εικόνα:



Παράδειγμα τοπολογίας perceptron

6.6.4.2 Τοπολογία Multi-layer perceptron

Ένα perceptron ενός επιπέδου είναι προφανώς ακατάλληλο για καταστάσεις με πολλές κλάσεις και μη γραμμικά όρια διαχωρισμού, αυτό υπήρξε και το κίνητρο για την ανακάλυψη των δικτύων πολλών επιπέδων από perceptrons. Τα επίπεδα νευρώνων σε αυτά τα δίκτυα είναι πλήρως συνδεδεμένα, αλλά απουσιάζουν οι συνδέσεις μεταξύ των νευρώνων του ίδιου επιπέδου. Τα δίκτυα MLP χρησιμοποιούν επιθεωρημένη εκπαίδευση, υλοποιημένη σε δύο φάσεις. Στην προωθούμενη φάση, υπολογίζεται η έξοδος της ζώνης εξόδου του δικτύου. Στην backward φάση τα βάρη ρυθμίζονται ώστε να μειώσουν το λάθος ανάμεσα στις παρατηρούμενες και στις επιθυμητές εξόδους.

Ένα MLP θεωρείται ότι διαθέτει ικανότητα γενίκευσης, δηλαδή μπορεί να παράξει σωστή (ή σχεδόν σωστή) έξοδο δεχόμενο είσοδο που δεν χρησιμοποιήθηκε στη φάση της εκπαίδευσης. Το δίκτυο μπορεί να ιδωθεί σαν μια μη γραμμική αντιστοιχία εισόδων-εξόδων, και η εκπαιδευτική διαδικασία σαν μια συνάρτηση προσαρμογής σε ένα ορισμένο σύνολο δεδομένων. Ένα MLP παρουσιάζεται αποδοτικό στην παρεμβολή, καθώς οι συνεχείς συναρτήσεις ενεργοποίησης παράγουν συνεχείς συναρτήσεις εξόδου. Επίσης επειδή η διαδικασία εκμάθησης έχει να κάνει με ψάξιμο σε ένα σύνθετο χώρο, είναι συχνά χρονοβόρα.

Τα MLP έχουν εφαρμοστεί σε πολλές περιπτώσεις, από απλή ταξινόμηση, προσέγγιση συναρτήσεων, μέχρι αναγνώριση φωνής και αναγνώριση στόχου (σε στρατιωτικές εφαρμογές). Λόγω της διάδοσής τους στην επιστημονική κοινότητα, πολλές είναι και οι εφαρμογές σε προβλήματα της περιβαλλοντικής πληροφορικής.

6.6.5 Αλγόριθμοι Meta (Αλγόριθμοι Bagging, Regression By Discretization)

Τα "Μετα" μαθησιακά σχήματα, είναι μια σχετικά νέα μέθοδος για το συνδυασμό πολλαπλών ταξινομητών . Η θεμελιώδης ιδέα είναι η κατασκευή πολλών "εμπειρογνωμόνων" ("experts") για την ανάδειξη της πλειοψηφούσας γνώμης. Εμπειρογνώμονες είναι τα μοντέλα που δημιουργήθηκαν με τεχνικές μηχανική μάθησης. Η διαφορά ανάμεσα στους ¨Μετα¨ αλγόριθμους και τα δέντρα απόφασης είναι ότι οι ¨Μετα¨ αλγόριθμοι προσδιορίζουν ποιος ταξινομητής θα χρησιμοποιηθεί, αντί να προβλέπουν την αξία της τάξης απευθείας (όπως τα δέντρα απόφασης).

Πλεονέκτημα των ¨Μετα¨ αλγορίθμων είναι ότι συχνά βελτιώνουν σημαντικά την προβλεπτική ικανότητα ενώ μειονέκτημα αποτελεί το γεγονός ότι τα εξαγόμενα είναι πολύ δύσκολο να αναλυθούν. Οι ¨Μετα¨ αλγόριθμοι εφαρμόζονται τόσο για ταξινόμηση, όσο και για αριθμητική πρόβλεψη

6.6.5.1 Ο ¨Μετα¨ αλγόριθμος Εμφωλίαση (bagging)

Συνδυάζει προβλέψεις μέσω καταμέτρησης ψήφων / εύρεσης μέσου όρου και αποτελεί μια απλή μέθοδο όπου κάθε μοντέλο λαμβάνει ισοδύναμη βαρύτητα.

Αρχικά γίνεται δειγματοληψία αρκετών συνόλων εκπαίδευσης μεγέθους n (αντί του ενός συνόλου μεγέθους n), ακολουθεί η κατασκευή ενός ταξινομητή για κάθε σύνολο εκπαίδευσης και τέλος συνδυάζονται οι προβλέψεις των ταξινομητών.

Όταν το σχήμα εκμάθησης είναι ασταθές (unstable), η εμφωλίαση βελτιώνει σχεδόν πάντα την αποδοτικότητα καθώς διαφορετικά, μικρές αλλαγές στα δεδομένα εκπαίδευσης μπορούν να επιφέρουν σημαντικές μεταβολές στο μοντέλο όπως για παράδειγμα στα δένδρα απόφασης.

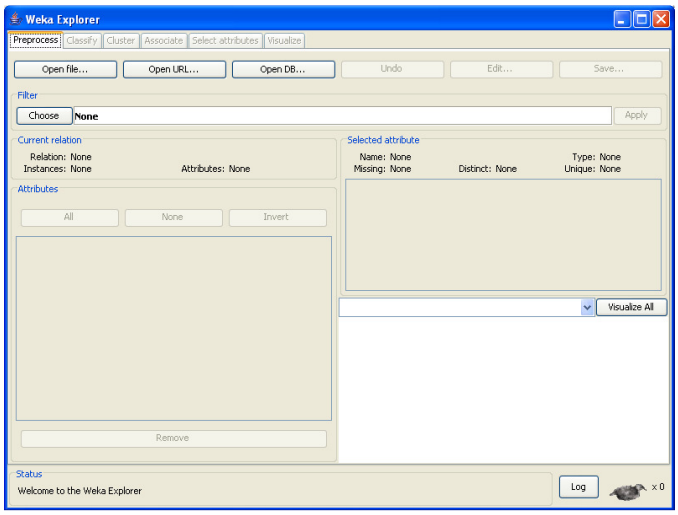
Η εμφωλίαση λειτουργεί καθώς μειώνει τη διακύμανση μέσω της καταμέτρησης ψήφων / εύρεσης μέσου όρου. Σε κάποιες παθολογικές καταστάσεις το συνολικό σφάλμα μπορεί να αυξηθεί. Συνήθως, η βελτίωση είναι ανάλογη του αριθμού των ταξινομητών. Για αριθμητικά χαρακτηριστικά: εύρεση μέσου όρου και όχι σύστασης πλειοψηφίας.

Για την ανάλυση της επιρροής επί της απόδοσης καθενός συνόλου εκπαίδευσης γίνεται η υπόθεση άπειρων ταξινομητών, κατασκευασμένων με βάση διαφορετικά σύνολα εκπαίδευσης μεγέθους n. Για κάθε μαθησιακό σχήμα, υπολογίζονται:

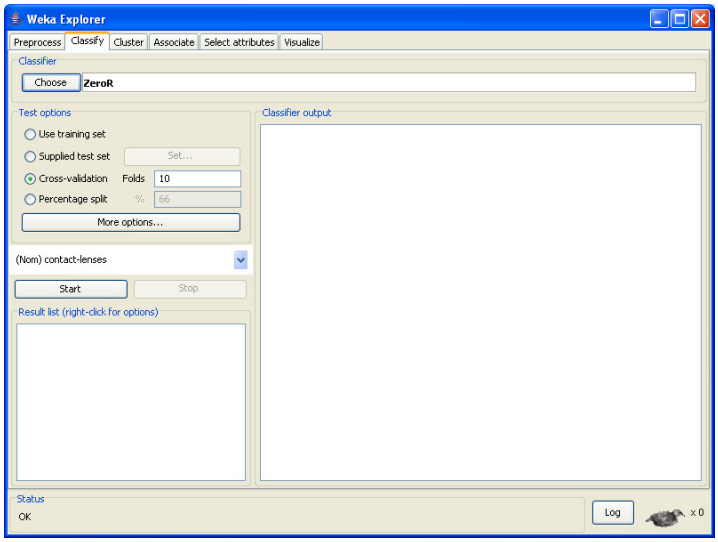
- Bias (προκατάληψη) = αναμενόμενο σφάλμα του μετα-ταξινομητή σε νέα δεδομένα.
- Variance (διακύμανση) = αναμενόμενο σφάλμα λόγω του συγκεκριμένου συνόλου εκπαίδευσης που χρησιμοποιήθηκε.
- Συνολικό αναμενόμενο σφάλμα: bias + variance.

6.7 ΑΞΙΟΛΟΓΗΣΗ ΚΑΙ ΕΠΙΛΟΓΗ ΚΑΤΑΛΛΗΛΟΤΕΡΩΝ ΑΛΓΟΡΙΘΜΩΝ

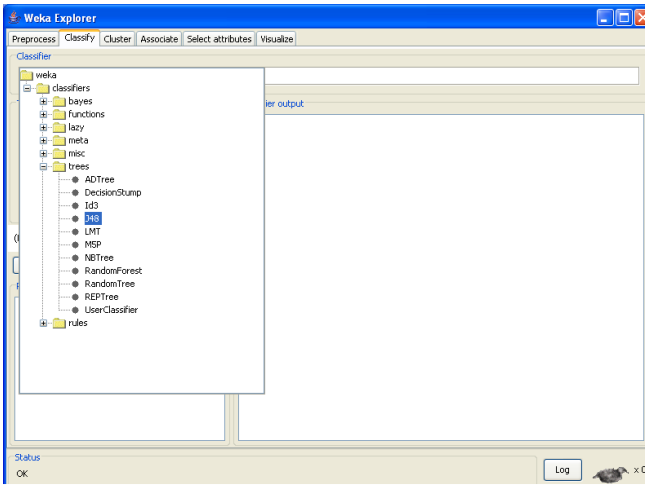
Θα εξετάζουμε τους έντεκα αλγόριθμους που επιλέξαμε, για να ανακαλύψουμε ποιοι από αυτούς είναι οι ικανότεροι να προβλέψουν την κατανάλωση φυσικού αερίου. Θα προσπαθήσουμε να καταλήξουμε στους τρεις καλύτερους αλγόριθμους ώστε να είναι εύκολη η επεξεργασία των δεδομένων κατά την επεξεργασία μεγάλων βάσεων δεδομένων. Θα εφαρμόσουμε τους έντεκα αλγόριθμους και θα επιλέξουμε εκείνους με το μικρότερο σφάλμα και υψηλότερο συντελεστή συσχέτισης.



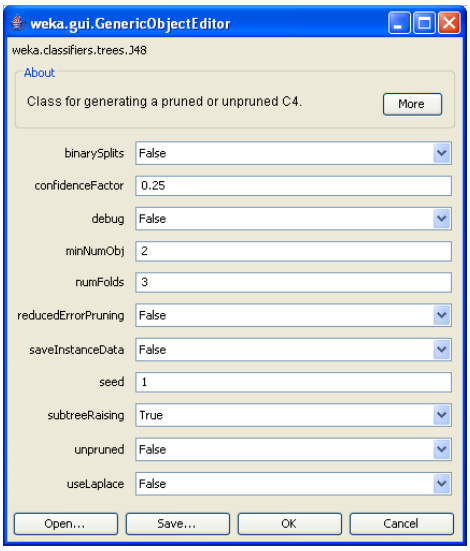
Για να τρέξουμε τους αλγόριθμους θα πάμε στο Weka Explorer και στην καρτέλα με την ονομασία Classify (κατηγοριοποίηση). Θα τρέξουμε τους έντεκα επιλεγμένους αλγόριθμους για τα επιλεγμένα χαρακτηριστικά και για τα τρία ενδεικτικά σετ δεδομένων εκπαίδευσης. Τα επιλεγμένα dataset προς ανάλυση ονομάζονται “training set 1”, “training set 2”, “training set 3”.



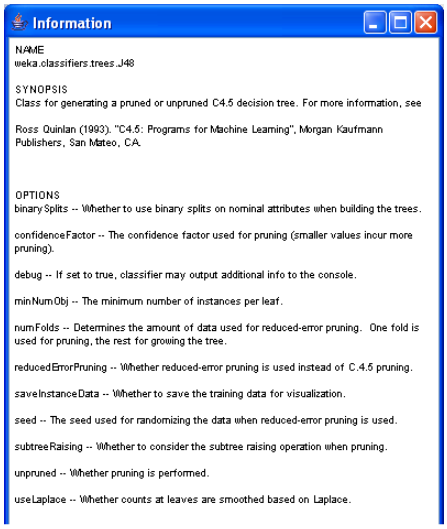
Για κάθε αρχείο δεδομένων εκπαίδευσης, θα πηγαίνουμε στην επιλογή classifier > choose, επιλέγοντάς κάθε φορά τον αλγόριθμο που θέλουμε να τρέξουμε.



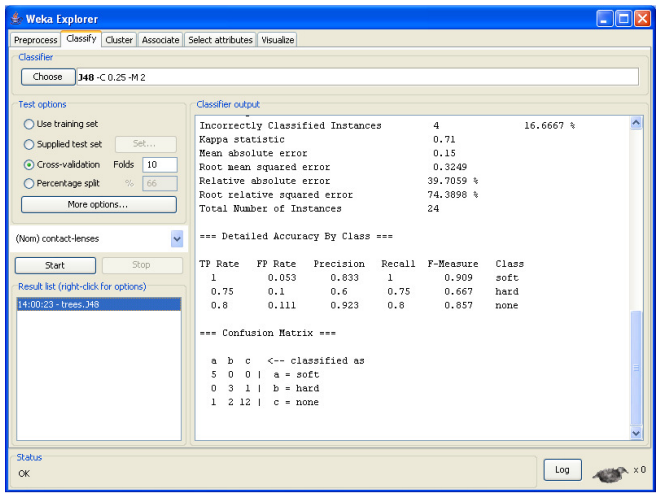
Κάνοντας διπλό κλικ πάνω στην επιλεγμένη μέθοδο εμφανίζεται το παράθυρο παραμέτρων της μεθόδου.



Επιλέγοντας more εμφανίζεται μία σύντομη οθόνη με περιγραφές των παραμέτρων της εικονιζόμενης μεθόδου.



Επιλέγοντας start ξεκινά η διαδικασία ανάπτυξης του μοντέλου και στην δεξιά πλευρά της οθόνης “classifier output” εμφανίζονται τα αποτελέσματα.



Εάν επιλεγεί μία άλλη μέθοδος, και εκτελεστεί η ανάλυση, τότε αυτή προστίθεται στο παράθυρο result list, διατηρώντας και τα αποτελέσματα της προηγούμενης ανάλυσης.

Τώρα θα τρέξουμε τους διάφορους αλγόριθμους μηχανικής μάθησης που χρησιμοποιούνται για ταξινόμηση (classification). Επιλέγουμε τον αλγόριθμο weka.classifiers (κουμπί "Choose" πάνω αριστερά). Θα εξετάσουμε τα σετ εκπαίδευσης δυο τρόπους.

Η πρώτη μέθοδος είναι η επιλογή του πειράματος (test options) το πεδίο percentage split 66%. Όταν ξεκινήσει το πείραμα (κουμπί start), θα χρησιμοποιηθεί ο αλγόριθμος στο 66% του σετ δεδομένων που θα αξιολογηθεί ως προς την ορθότητα των αποφάσεων του στο υπόλοιπο 34% του σετ δεδομένων.

Η δεύτερη μέθοδος είναι το Cross-validation, η τεχνική της διασταυρωμένης επικύρωσης όπου ο ταξινομητής αποτιμάτε, χρησιμοποιώντας τον αριθμό των folds που εισάγονται στο ανάλογο πεδίο. Η τεχνική αυτή χωρίζει το σύνολο των δεδομένων σε 10 ίσα μέρη (10-fold cross validation) και επαναλαμβάνει τα πειράματα 10 φορές. Σε κάθε επανάληψη τα διανύσματα που παριστάνουν τα παράθυρα των 9 μερών αποτελούν τα δεδομένα εκπαίδευσης και τα διανύσματα που παριστάνουν τα παράθυρα του 10ου μέρους αποτελούν τα δεδομένα αξιολόγησης. (Σε κάθε επανάληψη χρησιμοποιείται διαφορετικό μέρος για τα δεδομένα αξιολόγησης.) Έτσι όλα τα μέρη χρησιμοποιούνται τελικά τόσο για εκπαίδευση όσο και για αξιολόγηση, χωρίς όμως ποτέ το σύστημα να αξιολογείται σε δεδομένα που έχουν χρησιμοποιηθεί για την εκπαίδευσή του. Τα τελικά αποτελέσματα είναι ο μέσος όρος των αποτελεσμάτων των επαναλήψεων.

Τα αποτελέσματα περιλαμβάνουν τις εξής πληροφορίες:

1. Run information: Πληροφορίες σχετικά με τις επιλογές του learning scheme, relation name, instances, attributes και το test mode που σχετίζονται με τη διαδικασία.

2. Classifier model (full training set): Μια textual αναπαράσταση του classification μοντέλου που δημιουργήθηκε σε όλα τα training data.

3. Τα αποτελέσματα

4. Summary: Λίστα στατιστικών για το πώς έγινε η πρόβλεψη την true class των instances κάτω από το επιλεγμένο test mode.

5. Detailed Accuracy By Class: Μια πιο λεπτομερής αναφορά ανά class για την ακρίβεια πρόβλεψης του classifier.

6. Confusion Matrix: Δείχνει πόσα instances αντιστοιχίζονται σε κάθε class. Τα elements δείχνουν τον αριθμό των test examples των οποίων η ακριβής class είναι η γραμμή και των οποίων η προβλεπόμενη class είναι η στήλη.

Επομένως θα τρέξουμε τα επιλεγμένα dataset (“training set 1”, “training set 2”, “training set 3”) των 2000 υποδειγμάτων (instances) έκαστος. Τα αποτελέσματα παρουσιάζονται στους παρακάτω πίνακες.

Training Set 1, μέθοδος Cross Validation (10 folds)

Αλγόριθμοι / Δείκτες	M5P	REPTree	Linear Regression	Multilayer Perceptor	SMOreg	IBk4	KStar	Bagging	Regression By Discret.	Decision Table	M5Rules
Correlation coefficient	0,9301	0,8425	0,8446	0,9056	0,8061	0,9309	0,9314	0,8997	0,9008	0,8385	0,9286
Mean absolute error	466,5	644,6	912,9	670,4	940,3	303,97	297,5	528,3	620,1	628,7	446,1
Root mean squared error	936,1	1389,0	1361,3	1131,7	1508,8	624,41	621,5	1109,6	1106,5	1409,0	943,4
Relative absolute error	24,96%	34,48%	48,84%	35,86%	50,30%	25,95	25,40%	28,26%	33,17%	33,63%	23,86%
Root relative squared error	36,69%	54,43%	53,35%	44,35%	59,13%	36,56	36,39%	43,49%	43,37%	55,22%	36,97%
Total Number of Instances	2038	2038	2038	2038	2038	2038	2038	2038	2038	2038	2038

Training Set 1, μέθοδος Percentage Split

Αλγόριθμοι / Δείκτες	M5P	REPTree	Linear Regression	Multilayer Perceptor	SMOreg	IBk4	KStar	Bagging	Regression By Discret.	Decision Table	M5Rules
Correlation coefficient	0,9470	0,8562	0,8531	0,8862	0,8425	0,9503	0,942	0,9142	0,9083	0,8017	0,9529
Mean absolute error	636,2	678,6	963,9	797,3	954,8	296,3	286,6	745,9	691,2	806,3	565,3
Root mean squared error	1070,9	1540,9	1504,9	1331,6	1573,3	534,1	574,5	1443,0	1337,2	1816,5	1047,0
Relative absolute error	32,70%	34,88%	49,55%	40,98%	49,08%	25,99%	25,13%	38,30%	35,52%	41,45%	29,06%
Root relative squared error	38,50%	55,40%	54,11%	47,88%	56,57%	31,43%	33,82%	51,88%	48,80%	65,30%	37,65%
Total Number of Instances	2038	2038	2038	2038	2038	2038	2038	2038	2038	2038	2038

Training Set 2, μέθοδος Cross Validation (10 folds)

Αλγόριθμοι / Δείκτες	M5P	REPTree	Linear Regression	Multilayer Perceptor	SMOreg	IBk3	KStar	Bagging	Regression By Discret.	Decision Table	M5Rules
Correlation coefficient	0,904	0,818	0,821	0,880	0,783	0,924	0,9302	0,874	0,875	0,815	0,902
Mean absolute error	513,2	709,1	1004,2	737,4	1034,4	280,8	280,5	581,1	682,1	691,5	490,8
Root mean squared error	1029,7	1527,9	1497,4	1244,9	1659,6	633,6	608,6	1220,6	1217,2	1549,9	1037,8
Relative absolute error	27,45%	37,93%	53,72%	39,45%	55,33%	26,14%	26,11%	31,09%	36,49%	36,99%	26,25%
Root relative squared error	40,36%	59,87%	58,69%	48,79%	65,04%	38,35%	36,84%	47,84%	47,71%	60,75%	40,67%
Total Number of Instances	2017	2017	2017	2017	2017	2017	2017	2017	2017	2017	2017

Training Set 2, μέθοδος Percentage Split

Αλγόριθμοι / Δείκτες	M5P	REPTree	Linear Regression	Multilayer Perceptor	SMOreg	IBk3	KStar	Bagging	Regression By Discret.	Decision Table	M5Rules
Correlation coefficient	0,920	0,832	0,829	0,861	0,818	0,9069	0,9244	0,888	0,882	0,779	0,926
Mean absolute error	699,8	746,5	1060,3	877,0	1050,3	282,1	276,1	820,5	760,3	886,9	621,8
Root mean squared error	1178,0	1695,0	1655,4	1464,8	1730,6	706,8	641,6	1587,3	1470,9	1998,2	1151,7
Relative absolute error	35,97%	38,37%	54,51%	45,08%	53,99%	27,03%	26,45%	42,13%	39,07%	45,60%	31,96%
Root relative squared error	42,35%	60,94%	59,52%	52,67%	62,23%	42,42%	38,50%	57,07%	53,68%	71,83%	41,41%
Total Number of Instances	2017	2017	2017	2017	2017	2017	2017	2017	2017	2017	2017

Training Set 3, μέθοδος Cross Validation (10 folds)

Αλγόριθμοι / Δείκτες	M5P	REPTree	Linear Regression	Multilayer Perceptor	SMOreg	IBk4	KStar	Bagging	Regression By Discret.	Decision Table	M5Rules
Correlation coefficient	0,9441	0,8551	0,8573	0,9192	0,8182	0,9455	0,9455	0,9132	0,9143	0,8511	0,9425
Mean absolute error	453,2	626,2	886,9	651,3	913,5	129,1	128,98	513,2	602,4	610,8	433,4
Root mean squared error	909,4	1349,4	1322,5	1099,4	1465,8	381,9	381,94	1078,0	1075,0	1368,9	916,6
Relative absolute error	24,25%	33,50%	47,45%	34,84%	48,87%	16,68	16,72	27,45%	32,22%	32,67%	23,18%
Root relative squared error	35,64%	52,88%	51,83%	43,09%	57,45%	32,55	32,55	42,25%	42,13%	53,65%	35,92%
Total Number of Instances	2014	2014	2014	2014	2014	2014	2014	2014	2014	2014	2014

Training Set 3, μέθοδος Percentage Split

Αλγόριθμοι / Δείκτες	M5P	REPTree	Linear Regression	Multilayer Perceptor	SMOreg	IBk4	KStar	Bagging	Regression By Discret.	Decision Table	M5Rules
Correlation coefficient	0,9612	0,8690	0,8659	0,8995	0,8551	0,9537	0,9293	0,9279	0,9219	0,8137	0,9672
Mean absolute error	618,1	659,3	936,4	774,6	927,6	123,5	139,1	724,6	671,5	783,3	549,2
Root mean squared error	1040,4	1497,0	1462,0	1293,6	1528,5	345,9	425,4	1401,9	1299,1	1764,7	1017,2
Relative absolute error	31,77%	33,89%	48,14%	39,81%	47,68%	16,31%	18,36%	37,21%	34,51%	40,27%	28,23%
Root relative squared error	37,40%	53,82%	52,57%	46,52%	54,96%	30,10%	37,02%	50,40%	47,41%	63,44%	36,57%
Total Number of Instances	2014	2014	2014	2014	2014	2014	2014	2014	2014	2014	2014

Μετά από την εξέταση και των έντεκα αλγορίθμων που είχαν αρχικά επιλεχτεί, ανακαλύψαμε ποιοι από αυτούς είναι οι ικανότεροι να προβλέψουν την κατανάλωση φυσικού αερίου. Καταλήξαμε στους τρεις καλύτερους αλγόριθμους, μετά από επιλογή εκείνων με το μικρότερο σφάλμα και τον υψηλότερο συντελεστή συσχέτισης. Καλύτερα αποτελέσματα δίνουν οι αλγόριθμοι που υλοποιήθηκαν με τη μέθοδο είναι Cross-validation (επιλογή: folds 10) όπου ο ταξινομητής αποτιμάτε, χρησιμοποιώντας τον αριθμό των folds που εισάγονται στο ανάλογο πεδίο.

Παρακάτω παρουσιάζονται τα αποτελέσματα για τους τρεις καλύτερους αλγορίθμους με τη μέθοδο Cross Validation.

Training Set 1

	IBk4	***KStar***	***Bagging***
Correlation coefficient	0,9309	0,9314	0,8997
Mean absolute error	303,97	297,5	528,3
Root mean squared error	624,41	621,5	1109,6
Relative absolute error	25,95	25,40%	28,26%
Root relative squared error	36,56	36,39%	43,49%
Total Number of Instances	2038	2038	2038
Duration (Seconds)	3	19	32

Training Set 2

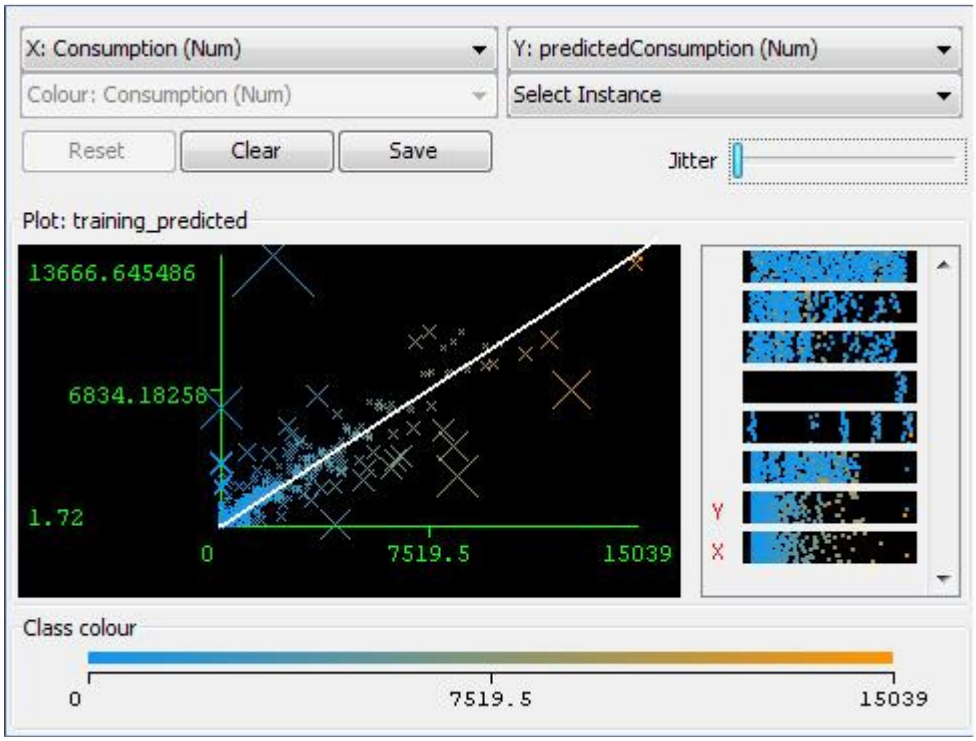
	IBk3	***KStar***	***Bagging***
Correlation coefficient	0,924	0,9302	0,874
Mean absolute error	280,8	280,5	581,1
Root mean squared error	633,6	608,6	1220,6
Relative absolute error	26,14%	26,11%	31,09%
Root relative squared error	38,35%	36,84%	47,84%
Total Number of Instances	2017	2017	2017
Duration (Seconds)	3	20	35

Training Set 3

	IBk4	***KStar***	***Bagging***
Correlation coefficient	0,9455	0,9455	0,9132
Mean absolute error	129,1	128,98	513,2
Root mean squared error	381,9	381,94	1078,0
Relative absolute error	16,68	16,72	27,45%
Root relative squared error	32,55	32,55	42,25%
Total Number of Instances	2014	2014	2014
Duration (Seconds)	3	19	34

Παρατηρούμε ότι τρεις αλγόριθμοι δίνουν συνεχώς καλύτερο αποτέλεσμα πρόβλεψης, μετά την εφαρμογή και των έντεκα αλγορίθμων. Καλύτερος αλγόριθμος φαίνεται να είναι ο IBk των κ-κοντινότερων γειτόνων. Ακολουθούν οι αλγόριθμοι Kstar και Bagging.

Το αποτέλεσμα της πρόβλεψης είναι ικανοποιητικότατο. Για τον αλγόριθμος IBk ο συντελεστής αυτοσυσχέτισης (Correlation coefficient) είναι άνω του 0,9. Αυτό σημαίνει ότι η κατανάλωση με την αναμενόμενη κατανάλωση που προέκυψε από την πρόβλεψη στο αντίστοιχο διάγραμμα, τείνει να είναι μια διαγώνιος ευθεία που όσο πλησιάζει στη μονάδα (1) τόσο καλύτερη είναι η πρόβλεψη. Επομένως όσο τείνει στη μονάδα ο συντελεστής αυτοσυσχέτισης τόσο καλύτερο είναι το αποτέλεσμα που δίνει ο αλγόριθμος. Στην παρακάτω εικόνα παρουσιάζεται το διάγραμμα κατανάλωσης – προβλεφθείσας κατανάλωσης για τον αλγόριθμο IBk για το σετ δεδομένων Training Set 1 με συντελεστή αυτοσυσχέτισης 0,9309.



Οι παραπάνω αλγόριθμοι είναι οι καταλληλότεροι για έναν ακόμη λόγο. Συγκριτικά με τους υπόλοιπους, χρειάζονται το λιγότερο χρόνο για την επεξεργασία των δεδομένων. Αυτό είναι πολύ σημαντικό ειδικά για την επεξεργασία μεγάλων βάσεων δεδομένων.

ΚΕΦΑΛΑΙΟ 7: ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΛΓΟΡΙΘΜΩΝ ΤΟΥ WEKA ΓΙΑ ΔΙΑΦΟΡΕΤΙΚΕΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

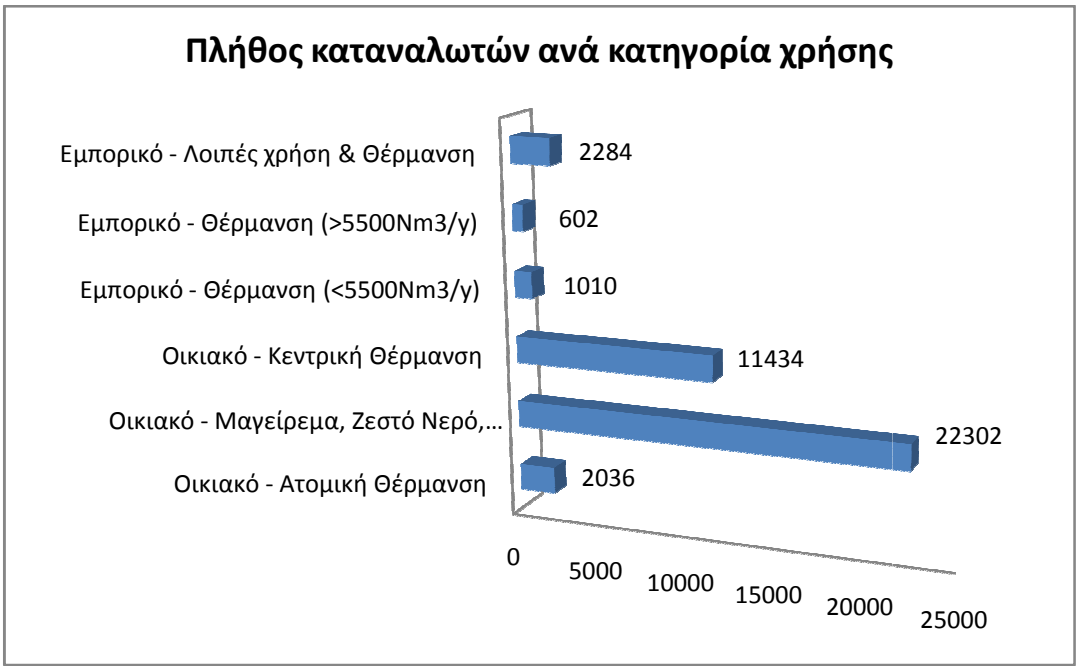
7.1 ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο αυτό θα περιγράψουμε τα πειράματα που πραγματοποιήθηκαν και τα αποτελέσματά τους. Αρχικά, θα εφαρμοσθούν οι τρεις επιλεχθέντες βέλτιστοι αλγόριθμοι για το σύνολο των δεδομένων και μετά ανά κατηγορία καταναλωτών και ανά κατηγορία χρήσης φυσικού αερίου. Αυτό θα γίνει για την εξαγωγή συμπερασμάτων για κάθε κατηγορία αλλά και για την αξιολόγηση της ικανότητας των αλγορίθμων να προβλέπουν την κατανάλωση φυσικού αερίου.

7.2 ΕΦΑΡΜΟΓΗ ΑΛΓΟΡΙΘΜΩΝ ΣΤΟ ΣΥΝΟΛΟ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Αρχικά, δημιουργήθηκε η βάση δεδομένων όλων των καταναλωτών με έξι χαρακτηριστικά (attributes). Τα χαρακτηριστικά αυτά επιλέχθηκαν βάση των αποτελεσμάτων αξιολόγησης χαρακτηριστικών που παρουσιάστηκε σε προηγούμενο κεφάλαιο. Το επιλεγμένα dataset προς ανάλυση ονομάστηκε “consumption ad_all”. Η βάση δεδομένων αποτελείται από 39.668 καταναλωτές για τους οποίους έχουν πραγματοποιηθεί 492.259 μετρήσεις κατανάλωσης διάφορων χρονικών περιόδων και 237 ταχυδρομικούς κώδικες της ευρύτερης περιοχής της Αττικής. Για κάθε καταναλωτή έχουν συγκεντρωθεί οι καταναλώσεις φυσικού αερίου και έχουν υπολογιστεί οι βαθμοημέρες θέρμανσης συναρτήσει του πλήθους των ημερών της κάθε χρονικής περιόδου.

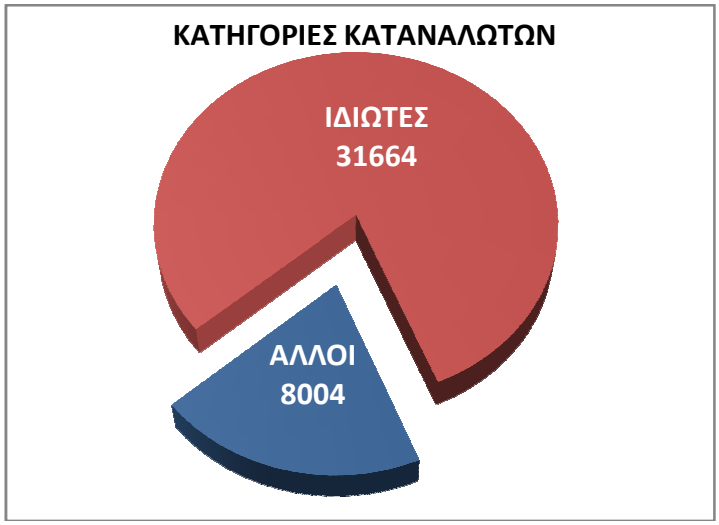
Παρακάτω παρουσιάζεται το πλήθος των καταναλωτών που ανήκει σε κάθε κατηγορία χρήσης φυσικού αερίου.



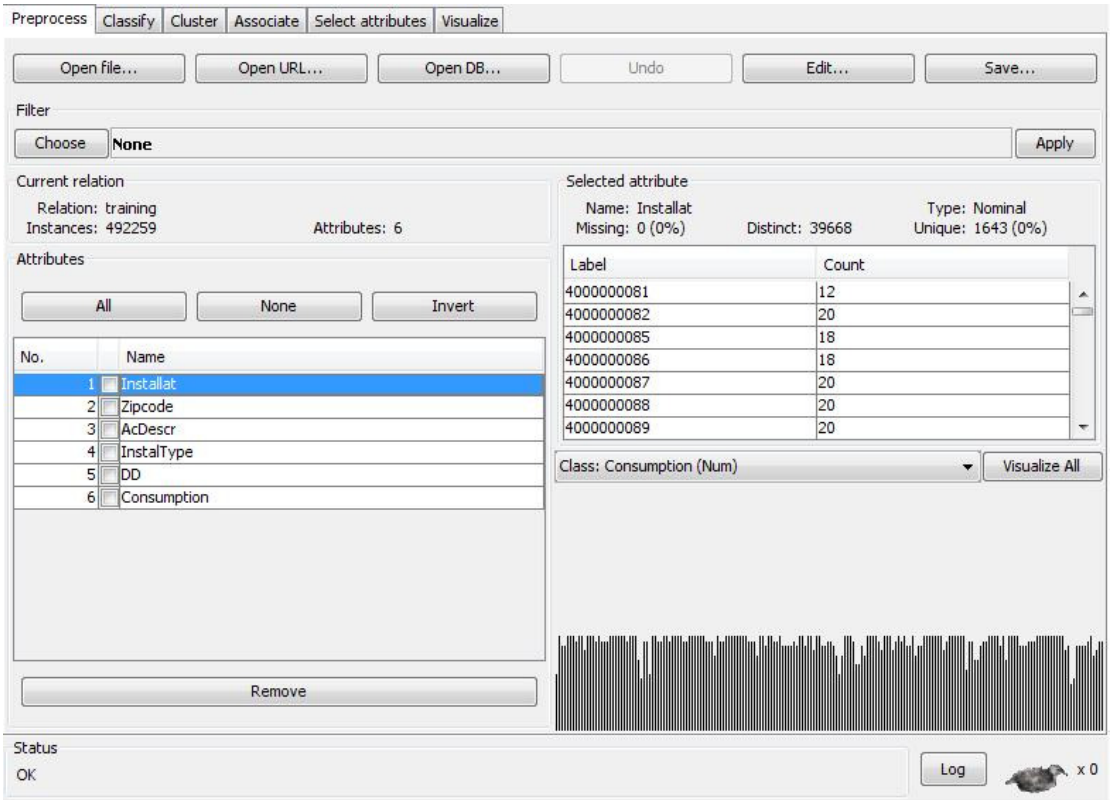
Παρακάτω παρουσιάζεται το πλήθος των καταναλωτών που ανήκει σε κάθε κατηγορία καταναλωτή.

Κατηγορίες Καταναλωτών	Κωδικός Κατηγορίας	Πλήθος Καταναλωτών
ΑΘΛΗΤΙΚΕΣ ΕΓΚΑΤΑΣΤΑΣΕΙΣ	**ad01**	**18**
ΑΛΛΕΣ ΥΠΗΡΕΣΙΕΣ (ΨΥΧΑΓ, ΚΛΠ)	**ad02**	**123**
ΑΝΑΨΥΚΤΗΡΙΟ	**ad03**	**17**
ΑΡΤΟΠΟΙΕΙΟ - ΖΑΧΑΡΟΠΛΑΣΤΕΙΟ	**ad04**	**499**
ΑΥΤΟΝΟΜΗ ΘΕΡΜΑΝΣΗ, ΜΑΓ., ΖΝΧ	**ad05**	**21**
ΒΙΟΜΗΧΑΝΙΑ	**ad06**	**23**
ΒΙΟΤΕΧΝΙΑ	**ad07**	**159**
ΓΡΑΦΕΙΟ - ΕΤΑΙΡΕΙΑ	**ad08**	**1729**
ΔΗΜΟΣΙΟ ΚΤΗΡΙΟ	**ad09**	**77**
ΕΜΠΟΡΙΚΟ ΚΑΤΑΣΤΗΜΑ	**ad10**	**55**
ΕΡΓΑΣΤΗΡΙΟ	**ad11**	**87**
ΕΣΤΙΑΤΟΡΙΟ - ΜΕΖΕΔΟΠΩΛΕΙΟ	**ad12**	**1176**
ΘΕΑΤΡΟ - ΣΙΝΕΜΑ	**ad13**	**11**
ΙΑΤΡΕΙΟ	**ad14**	**2**
ΙΔΙΩΤΗΣ	**ad15**	**31664**
ΙΕΡΟΣ ΝΑΟΣ	**ad16**	**56**
ΚΑΘΑΡΙΣΤΗΡΙΑ	**ad17**	**68**
ΚΑΦΕ - ΜΠΑΡ	**ad18**	**258**
ΚΕΝΤΡΙΚΗ ΘΕΡΜΑΝΣΗ	**ad19**	**36**
ΝΟΣΟΚΟΜΕΙΟ - ΚΛΙΝΙΚΗ	**ad20**	**58**
ΞΕΝΟΔΟΧΕΙΟ	**ad21**	**186**
ΟΙΚΟΣ ΕΥΓΗΡΙΑΣ - ΟΡΦΑΝΟΤΡΟΦΕΙΟ	**ad22**	**14**
ΠΑΝΕΠΙΣΤΗΜΙΟ	**ad23**	**6**
ΠΟΛΥΚΑΤΟΙΚΙΑ	**ad24**	**2754**
ΠΡΕΣΒΕΙΑ	**ad25**	**19**
ΣΧΟΛΕΙΟ	**ad26**	**477**
ΣΧΟΛΗ - ΕΚΠΑΙΔΕΥΤΗΡΙΟ	**ad27**	**58**
ΤΡΑΠΕΖΑ	**ad28**	**17**

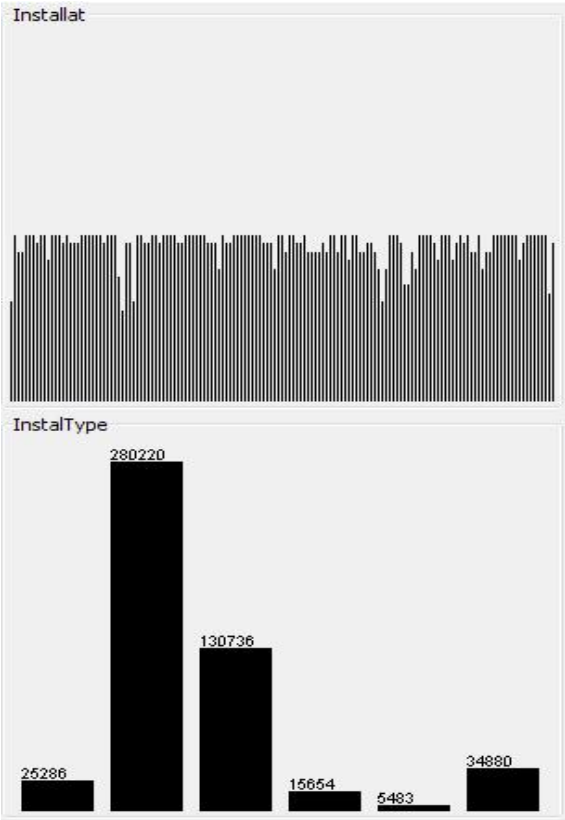
Παρατηρούμε ότι η πλειοψηφία των καταναλωτών ανήκει στην κατηγορία ¨ΙΔΙΩΤΕΣ¨.

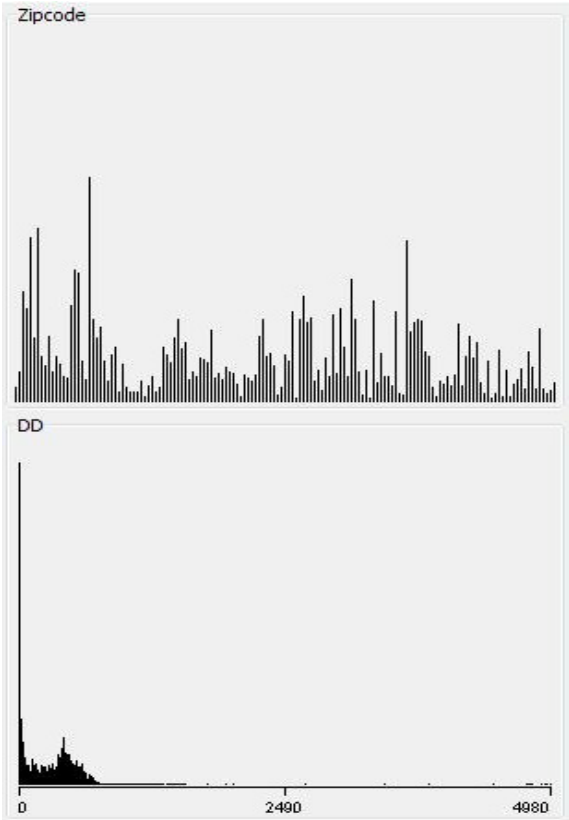


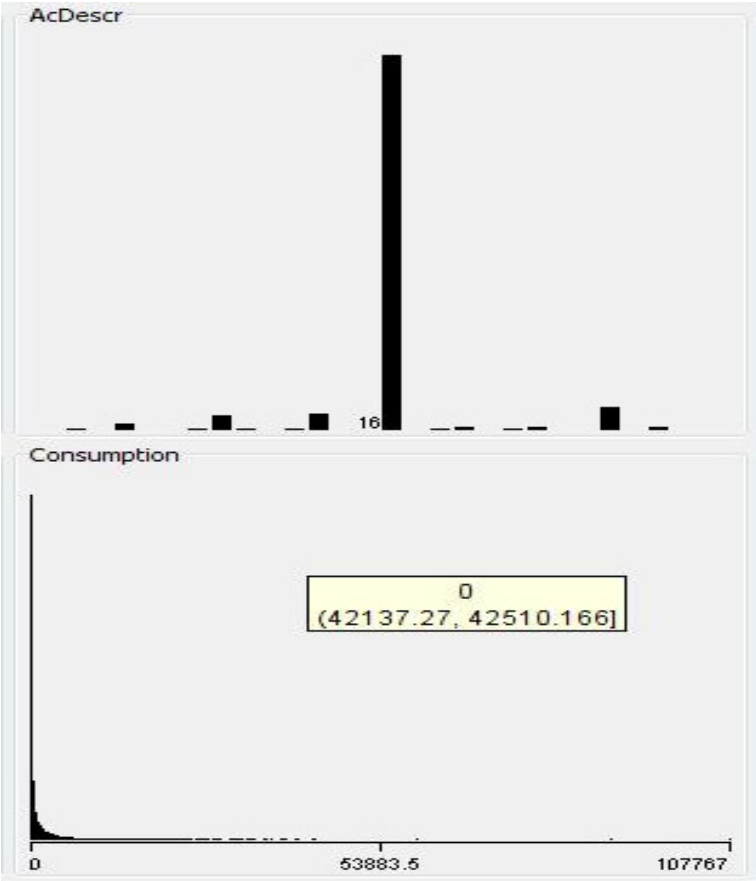
Θα πάμε στο Weka Explorer και στην καρτέλα με την ονομασία Preprocess, θα ανοίγουμε το αρχείο των δεδομένων προς ανάλυση δηλαδή το αρχείο “consumption_ad_all”. Στην καρτέλα preprocess επιλέγουμε open file > data > consumption_ad_all.arff.



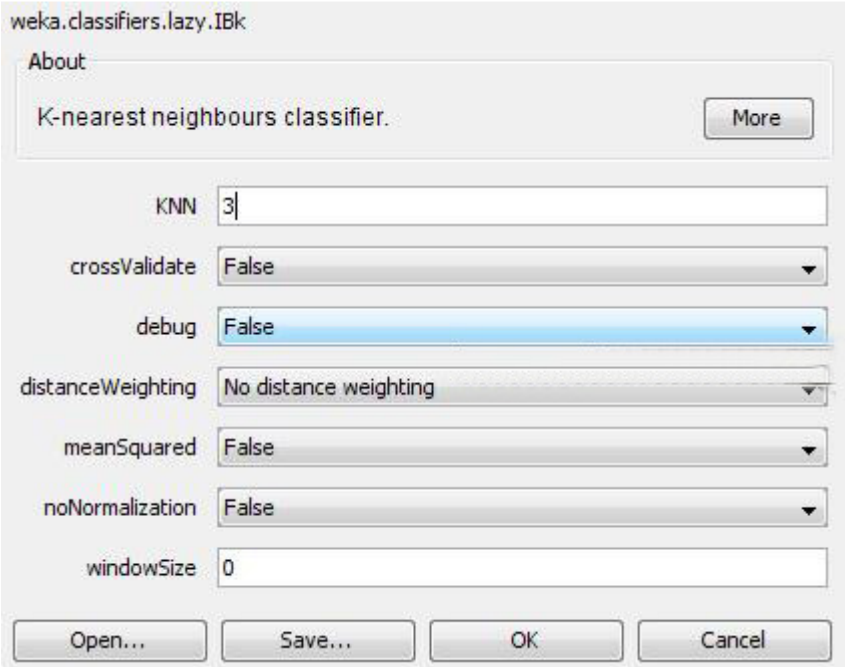
Στην επιλογή ‘Visualize All’ εμφανίζονται τα στατιστικά της βάσης δεδομένων.



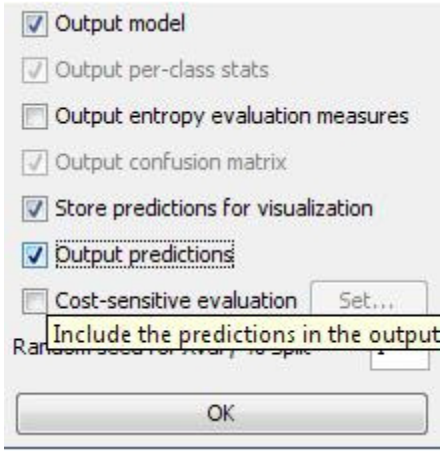




Θα εφαρμόσουμε του επιλεχθέντες αλγόριθμους IBk, Kstar και Bagging με την τεχνική Cross Validation. Ο αλγόριθμος IBk, θα παραμετροποιηθεί ώστε να βγάζει καλύτερα αποτελέσματα. Στην επιλογή KNN θα εισάγουμε τον αριθμό 3 αντί για 1. Με την επιλογή αυτή ο αλγόριθμος IBk, χρησιμοποιεί τους τρεις κοντινότερους γείτονες αντί για έναν.



Επίσης στην καρτέλα 'More Options' επιλέγουμε 'Output Predictions'. Με την επιλογή αυτή, μετά από το πέρας της επεξεργασίας του αλγορίθμου, θα εμφανιστούν οι καταναλώσεις, η προβλεφθείσα κατανάλωση και η απόκλισή τους.



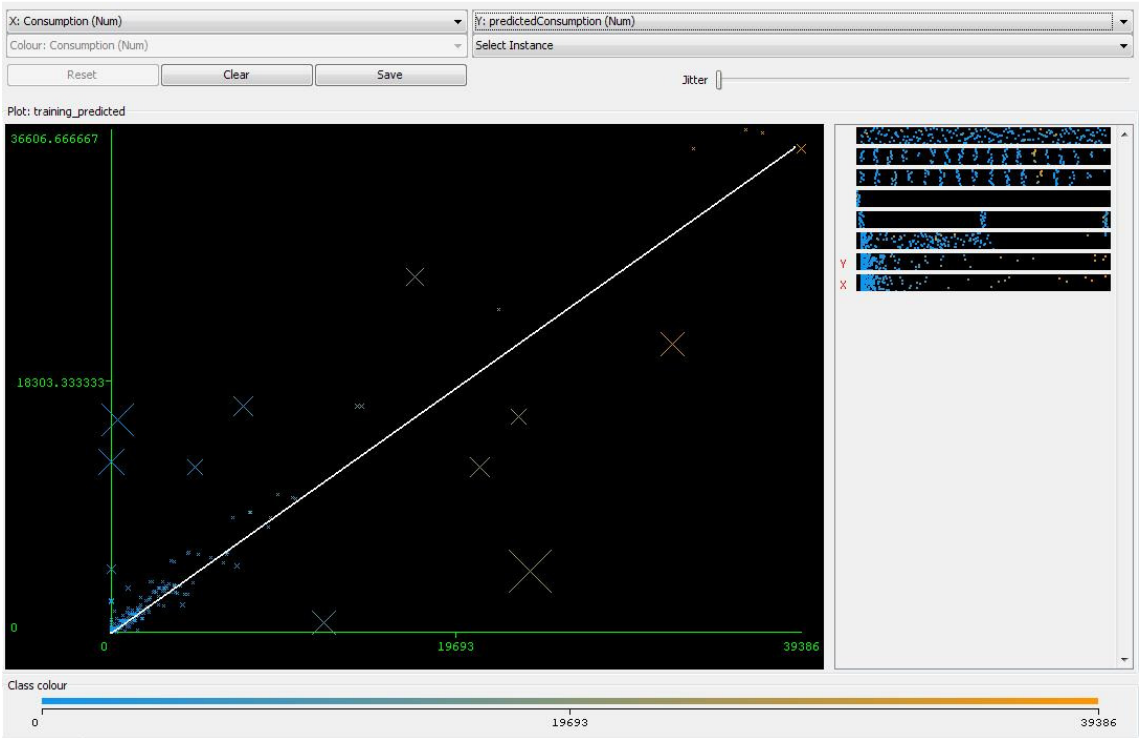
Η βάση δεδομένων είναι πολύ μεγάλη και η εφαρμογή των αλγορίθμων αποδείχτηκε χρονοβόρα. Παρακάτω παρουσιάζονται ενδεικτικά τα αποτελέσματα της εφαρμογής του αλγορίθμου IBk με KNN 3.

```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 3 -W 0
Relation:     training
Instances:    492259
Attributes:   6
              Installat
              Zipcode
              AcDescr
              InstalType
              DD
              Consumption
Test mode:    10-fold cross-validation


=== Classifier model (full training set) ===

IB1 instance-based classifier
using 3 nearest neighbour(s) for classification

Time taken to build model: 44028 seconds



=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9082
Mean absolute error                      62.551
Root mean squared error                  142.81
Relative absolute error                  31.98 %
Root relative squared error              47.65 %
Total Number of Instances                492259
```

Παρακάτω παρουσιάζονται τα αποτελέσματα της εφαρμογής των τριών αλγορίθμων IBk, Kstar και Bagging με την τεχνική Cross Validation.

	IBk3	KStar	Bagging
Correlation coefficient	**0,9082**	**0,8771**	**0,8345**
Mean absolute error	**62,551**	**66,355**	**72,234**
Root mean squared error	**142,81**	**155,79**	**180,90**
Relative absolute error	**31,98%**	**33,46%**	**39,56%**
Root relative squared error	**47,65%**	**48,12%**	**50,52%**
Total Number of Instances	**492259**	**492259**	**492259**
Time (Hours)	**12:23:12**	**15:58:21**	**26:12:45**

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 3. Ο συντελεστής αυτοσυσχέτισης είναι 0,9082 με μέσο σφάλμα 32,98%. Επειδή στο δείγμα περιλαμβάνονται όλες οι κατηγορίες χρήσεως και καταναλωτών είναι λογικό η πρόβλεψη της κατανάλωσης να μην είναι τόσο καλή όσο με τα δεδομένα εκπαίδευσης των 2000 καταμετρήσεων. Η βάση δεδομένων είναι τεράστια (492.259 καταμετρήσεις) και περιλαμβάνει και καταναλωτές που η κατανάλωσή τους δεν είναι ανάλογη της θερμοκρασίας, όπως επιχειρήσεις και βιομηχανίες. Στην παρακάτω εικόνα παρουσιάζεται το διάγραμμα κατανάλωσης – προβλεφθείσας κατανάλωσης για τον αλγόριθμο IBk με συντελεστή αυτοσυσχέτισης 0,9082.



7.3 ΕΦΑΡΜΟΓΗ ΑΛΓΟΡΙΘΜΩΝ ΑΝΑ ΚΑΤΗΓΟΡΙΑ ΧΡΗΣΗΣ

Όπως παρατηρήσαμε από την εφαρμογή των αλγορίθμων σε όλη τη βάση δεδομένων, τα αποτελέσματα ήταν καλά, με συντελεστή συσχέτισης άνω του 0,9. Παρόλα αυτά είναι πολύ ενδιαφέρον να εφαρμοσθούν οι αλγόριθμοι και ανά κατηγορία χρήσης καθώς οι καταναλωτές παρουσιάζουν ομοειδή καταναλωτική συμπεριφορά. Οι κατηγορίες χρήσεις περιλαμβάνουν ομαδοποιημένους τους καταναλωτές σε έξι βασικές κατηγορίες.

Κατηγορίες Χρήσης	Κωδικός Κατηγορίας	Πλήθος Καταναλωτών Δείγματος	Πλήθος Μετρήσεων Κατανάλωσης
Οικιακό - Ατομική Θέρμανση	in1	2036	25286
Οικιακό - Μαγείρεμα, Ζεστό Νερό, Θέρμανση	in2	22302	280220
Οικιακό - Κεντρική Θέρμανση	in3	11434	130736
Εμπορικό - Θέρμανση (<5500Nm3/y)	in4	1010	15654
Εμπορικό - Θέρμανση (>5500Nm3/y)	in5	602	5483
Εμπορικό - Λοιπές χρήσεις & Θέρμανση	in6	2284	34880

Αρχικά, δημιουργήθηκε η βάση δεδομένων για κάθε κατηγορία χρήσης φυσικού αερίου. Τα επιλεγμένα dataset προς ανάλυση ονομάστηκαν αντίστοιχα “in1”, “in2”, “in3”, “in4”, “in5”, “in6”. Η βάση δεδομένων “in1” αποτελείται από 2.036 καταναλωτές για τους οποίους έχουν πραγματοποιηθεί 25.286 μετρήσεις κατανάλωσης. Η βάση δεδομένων “in2” αποτελείται από 22.302 καταναλωτές για τους οποίους έχουν πραγματοποιηθεί 280.220 μετρήσεις κατανάλωσης. Η βάση δεδομένων “in3” αποτελείται από 11.434 καταναλωτές για τους οποίους έχουν πραγματοποιηθεί 130.736 μετρήσεις κατανάλωσης. Η βάση δεδομένων “in4” αποτελείται από 1.010 καταναλωτές για τους οποίους έχουν πραγματοποιηθεί 15.654 μετρήσεις κατανάλωσης. Η βάση δεδομένων “in5” αποτελείται από 602 καταναλωτές για τους οποίους έχουν πραγματοποιηθεί 5.483 μετρήσεις κατανάλωσης. Η βάση δεδομένων “in6” αποτελείται από 2.284 καταναλωτές για τους οποίους έχουν πραγματοποιηθεί 34.880 μετρήσεις κατανάλωσης.

Παρακάτω παρουσιάζονται τα αποτελέσματα των αλγορίθμων για κάθε κατηγορία.

7.3.1 Οικιακό - Ατομική Θέρμανση (in1). Σε αυτή τη κατηγορία ανήκουν οι οικιακοί καταναλωτές που έχουν ατομική θέρμανση.

	IBk3	KStar	Bagging
Correlation coefficient	0,9003	0,8886	0,8770
Mean absolute error	111,2	127,9	147,1
Root mean squared error	268,5	308,8	355,1
Relative absolute error	30,30%	34,84%	40,07%
Root relative squared error	43,65%	50,20%	57,73%
Total Number of Instances	25286	25286	25286
Time (Seconds)	330	2640	11352

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 3. Ο συντελεστής αυτοσυσχέτισης είναι 0,9003 με μέσο σφάλμα 30,30%. Υπάρχει άμεση εξάρτηση αυτής της κατηγορίας χρήσης με το χαρακτηριστικό Βαθμοημέρες Θέρμανσης. Επομένως, ο αλγόριθμος συσχετίζει εύκολα τα δεδομένα και είναι ικανός να δώσει μια αρκετά ακριβή πρόβλεψη.

7.3.2 Οικιακό - Μαγείρεμα, Ζεστό Νερό, Θέρμανση (in2). Σε αυτή τη κατηγορία ανήκουν οι οικιακοί καταναλωτές που έχουν ατομική θέρμανση, Θέρμανση και Ζέσταμα νερού.

	IBk3	KStar	Bagging
Correlation coefficient	0,8771	0,8657	0,8544
Mean absolute error	66,355	76,3	87,8
Root mean squared error	155,79	179,2	206,0
Relative absolute error	33,46%	38,48%	44,25%
Root relative squared error	48,12%	55,34%	63,64%
Total Number of Instances	280220	280220	280220
Time (Hours)	12:23:12	14:12:59	21:55:42

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 3. Ο συντελεστής αυτοσυσχέτισης είναι 0,8771 με μέσο σφάλμα 33,46%. Υπάρχει εξάρτηση αυτής της κατηγορίας χρήσης με το χαρακτηριστικό Βαθμοημέρες Θέρμανσης. Παρόλα αυτά, επειδή το φυσικό αέριο χρησιμοποιείται και για άλλες λειτουργίες (π.χ. μαγείρεμα), η εξάρτηση από την εξωτερική θερμοκρασία είναι χαμηλότερη με αποτέλεσμα να μειώνεται η ικανότητα πρόβλεψης του αλγορίθμου.

7.3.3 Οικιακό - Κεντρική Θέρμανση (in3). Σε αυτή τη κατηγορία ανήκουν οι οικιακοί καταναλωτές που έχουν κεντρική θέρμανση.

	IBk3	KStar	Bagging
Correlation coefficient	0,9108	0,8990	0,8873
Mean absolute error	361,7	416,0	478,3
Root mean squared error	9503,4	10928,9	12568,2
Relative absolute error	28,70%	33,01%	37,96%
Root relative squared error	39,07%	44,93%	51,67%
Total Number of Instances	130736	130736	130736
Time (Seconds)	1231	9848	42346

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 3. Ο συντελεστής αυτοσυσχέτισης είναι 0,9108 με μέσο σφάλμα 28,70%. Υπάρχει άμεση εξάρτηση αυτής της κατηγορίας χρήσης με το χαρακτηριστικό Βαθμοημέρες Θέρμανσης. Η εξάρτηση από την εξωτερική θερμοκρασία είναι υψηλή με αποτέλεσμα να αυξάνεται η ικανότητα πρόβλεψης του αλγορίθμου λόγω της άμεσης συσχέτισης βαθμοημέρες θέρμανσης - κατανάλωσης. Επομένως, ο αλγόριθμος συσχετίζει εύκολα τα δεδομένα και είναι ικανός να δώσει μια πολύ ακριβής πρόβλεψη.

7.3.4 Εμπορικό - Θέρμανση (<5500Nm3/y) (in4). Σε αυτή τη κατηγορία ανήκουν οι εμπορικοί καταναλωτές που έχουν κατανάλωση μικρότερη των 5500Nm^3/y.

	IBk3	KStar	Bagging
Correlation coefficient	0,8602	0,8490	0,8380
Mean absolute error	101,0	116,2	133,6
Root mean squared error	324,0	372,6	428,5
Relative absolute error	29,35%	33,75%	38,82%
Root relative squared error	51,15%	58,82%	67,65%
Total Number of Instances	15654	15654	15654
Time (Seconds)	131	1048	4506

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 3. Ο συντελεστής αυτοσυσχέτισης είναι 0,8602 με μέσο σφάλμα 29,35%. Δεν υπάρχει τόσο ισχυρή εξάρτηση αυτής της κατηγορίας χρήσης με το χαρακτηριστικό Βαθμοημέρες Θέρμανσης. Σε αυτή την κατηγορία ανήκουν πολλές επιχειρήσεις, γραφεία, μικρές βιοτεχνίες με αποτέλεσμα να χρησιμοποιούν το φυσικό αέριο και για άλλους λόγους εκτός της θέρμανσης. Παρόλα αυτά το αποτέλεσμα είναι σχετικά ικανοποιητικό.

7.3.5 Εμπορικό - Θέρμανση (>5500Nm3/y) (in5). Σε αυτή τη κατηγορία ανήκουν οι εμπορικοί καταναλωτές που έχουν κατανάλωση μεγαλύτερη των 5500Nm^3/y.

	IBk3	KStar	Bagging
Correlation coefficient	0,7452	0,7086	0,6994
Mean absolute error	816,6	669,0	769,4
Root mean squared error	2518,5	2696,0	3100,4
Relative absolute error	40,12%	32,88%	37,81%
Root relative squared error	67,11%	71,86%	82,64%
Total Number of Instances	5483	5483	5483
Time (Seconds)	17	167	1336

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 3. Ο συντελεστής αυτοσυσχέτισης είναι 0,7452 με μέσο σφάλμα 40,12%. Δεν υπάρχει τόσο ισχυρή εξάρτηση αυτής της κατηγορίας χρήσης με το χαρακτηριστικό Βαθμοημέρες Θέρμανσης. Σε αυτή την κατηγορία ανήκουν πολλές μεγάλες επιχειρήσεις, σχετικά μεγάλες βιοτεχνίες και βιομηχανίες με αποτέλεσμα να χρησιμοποιούν το φυσικό αέριο και για άλλους λόγους εκτός της θέρμανσης. Το αποτέλεσμα δεν είναι ικανοποιητικό όσο στις άλλες κατηγορίες για ευνόητους λόγους.

7.3.6 Εμπορικό - Λοιπές χρήσεις & Θέρμανση (in6). Σε αυτή τη κατηγορία ανήκουν όλοι οι εμπορικοί καταναλωτές που δεν ανήκουν στις υπόλοιπες κατηγορίες εμπορικών καταναλωτών.

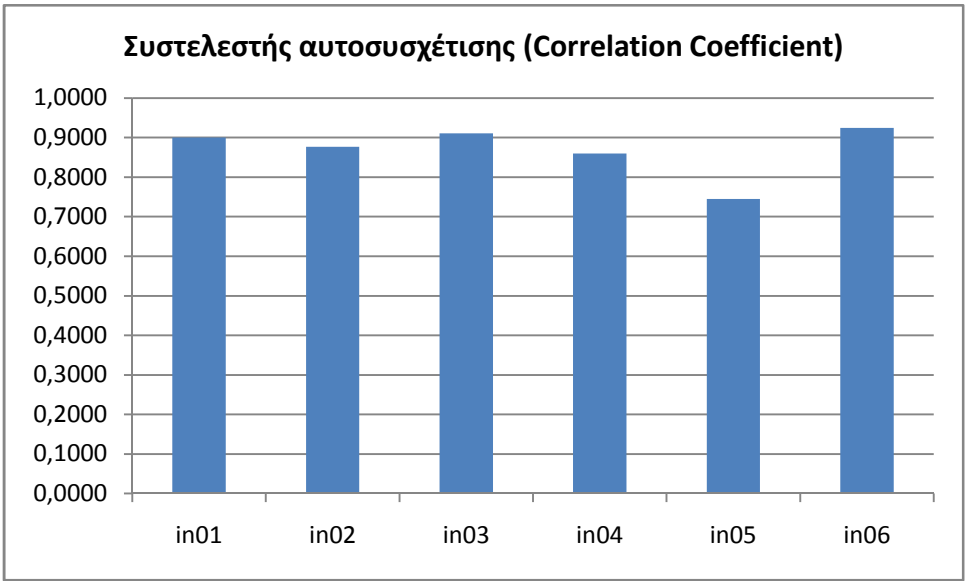
	IBk3	KStar	Bagging
Correlation coefficient	0,9250	0,9130	0,9011
Mean absolute error	361,7	416,0	478,3
Root mean squared error	853,4	981,4	1128,6
Relative absolute error	27,70%	31,86%	36,63%
Root relative squared error	38,01%	43,71%	50,27%
Total Number of Instances	34880	34880	34880
Time (Seconds)	681	5448	23426

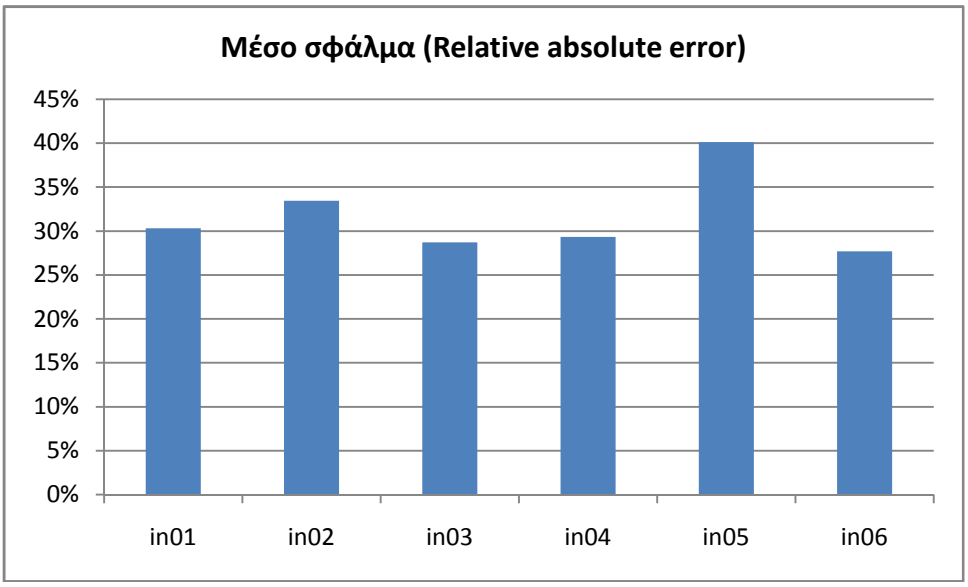
Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 3. Μάλιστα είναι το καλύτερο αποτέλεσμα όλων των κατηγοριών χρήσης. Ο συντελεστής αυτοσυσχέτισης είναι 0,9250 με μέσο σφάλμα 27,70%. Υπάρχει άμεση εξάρτηση αυτής της κατηγορίας χρήσης με το χαρακτηριστικό Βαθμοημέρες Θέρμανσης.

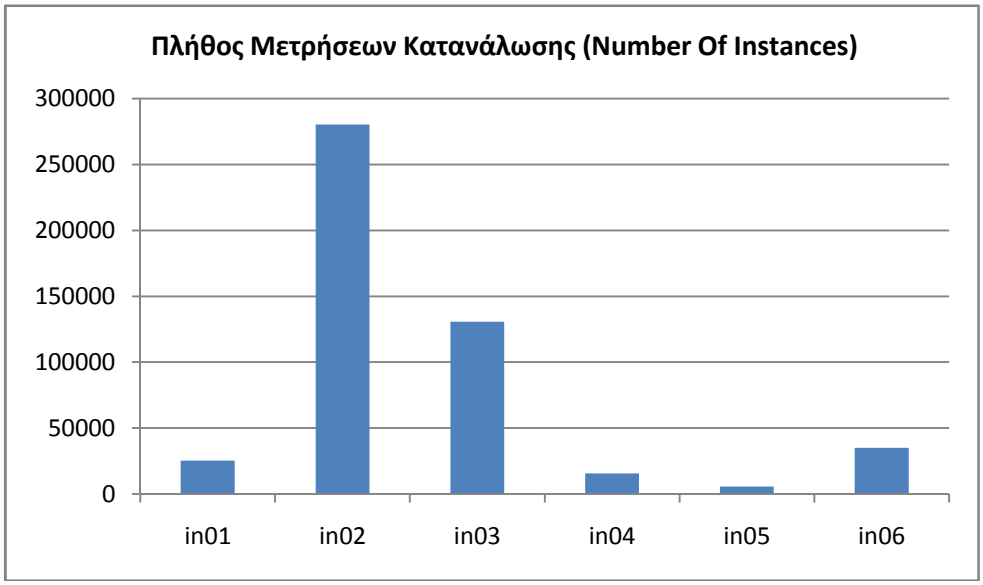
7.3.7 Ανάλυση των αποτελεσμάτων

Η εξάρτηση από την εξωτερική θερμοκρασία είναι υψηλή με αποτέλεσμα να αυξάνεται η ικανότητα πρόβλεψης του αλγορίθμου λόγω της άμεσης συσχέτισης βαθμοημέρες θέρμανσης - κατανάλωσης. Αυτό συμβαίνει γιατί στην κατηγορία αυτή ανήκουν εμπορικοί καταναλωτές που χρησιμοποιούν κυρίως το φυσικό αέριο για θέρμανση. Για παράδειγμα νοσοκομεία, γραφεία κ.α. Επομένως, ο αλγόριθμος συσχετίζει εύκολα τα δεδομένα και είναι ικανός να δώσει μια πολύ ακριβής πρόβλεψη.

Από την εφαρμογή των αλγορίθμων παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει **συνεχώς** ο αλγόριθμος **IBk** με KNN 3. Παρακάτω παρουσιάζονται τα αποτελέσματα και των έξι κατηγοριών χρήσης φυσικού αερίου.







Παρατηρούμε ότι τα αποτελέσματα της πρόβλεψης είναι πολύ καλά στις κατηγορίες που περιλαμβάνουν καταναλωτές που χρησιμοποιούν το φυσικό αέριο κυρίως για θέρμανση. Αντίθετα, στις κατηγορίες που περιλαμβάνουν καταναλωτές που χρησιμοποιούν το φυσικό αέριο για επαγγελματική χρήση (βιοτεχνίες, εργαστήρια), μειώνεται η ικανότητα των αλγορίθμων να προβλέπουν την κατανάλωση.

Επίσης, όσο μειώνεται το πλήθος των μετρήσεων της κατανάλωσης, τόσο μειώνεται η ακρίβεια της πρόβλεψης. Είναι λογικό να αυξάνεται το ποσοστό ακρίβειας της πρόβλεψης, όσο πιο πολλές μετρήσεις υπάρχουν στη βάση δεδομένων ούτος ώστε να υπάρχουν πιο πολλά δεδομένα για την εξαγωγή κανόνων και συσχετίσεων.

7.4 ΕΦΑΡΜΟΓΗ ΑΛΓΟΡΙΘΜΩΝ ΑΝΑ ΚΑΤΗΓΟΡΙΑ ΚΑΤΑΝΑΛΩΤΩΝ

Είναι πολύ ενδιαφέρον να εφαρμοσθούν οι αλγόριθμοι και ανά κατηγορία καταναλωτών καθώς οι καταναλωτές ανά κατηγορίες παρουσιάζουν ομοειδή καταναλωτική συμπεριφορά. Οι κατηγορίες καταναλωτών περιλαμβάνουν ομαδοποιημένους τους καταναλωτές σε είκοσι οκτώ βασικές κατηγορίες.

Κατηγορίες Χρήσης	Κωδικός Κατηγορίας	Πλήθος Καταναλωτών Δείγματος	Πλήθος Μετρήσεων Κατανάλωσης
ΑΘΛΗΤΙΚΕΣ ΕΓΚΑΤΑΣΤΑΣΕΙΣ	**ad01**	18	255
ΑΛΛΕΣ ΥΠΗΡΕΣΙΕΣ	**ad02**	123	2067
ΑΝΑΨΥΚΤΗΡΙΟ	**ad03**	17	208
ΑΡΤΟΠΟΙΕΙΟ – ΖΑΧΑΡΟΠΛ.	**ad04**	499	7756
ΑΥΤΟΝ. ΘΕΡΜΑΝΣΗ, ΜΑΓ., ΖΝΧ	**ad05**	21	388
ΒΙΟΜΗΧΑΝΙΑ	**ad06**	23	311
ΒΙΟΤΕΧΝΙΑ	**ad07**	159	2623
ΓΡΑΦΕΙΟ - ΕΤΑΙΡΕΙΑ	**ad08**	1729	16254
ΔΗΜΟΣΙΟ ΚΤΗΡΙΟ	**ad09**	77	903
ΕΜΠΟΡΙΚΟ ΚΑΤΑΣΤΗΜΑ	**ad10**	55	579
ΕΡΓΑΣΤΗΡΙΟ	**ad11**	87	1407
ΕΣΤΙΑΤΟΡΙΟ - ΜΕΖΕΔΟΠΩΛΕΙΟ	**ad12**	1176	17586
ΘΕΑΤΡΟ - ΣΙΝΕΜΑ	**ad13**	11	189
ΙΑΤΡΕΙΟ	**ad14**	2	16
ΙΔΙΩΤΗΣ	**ad15**	31664	401123
ΙΕΡΟΣ ΝΑΟΣ	**ad16**	56	802
ΚΑΘΑΡΙΣΤΗΡΙΑ	**ad17**	68	1188
ΚΑΦΕ - ΜΠΑΡ	**ad18**	258	4131
ΚΕΝΤΡΙΚΗ ΘΕΡΜΑΝΣΗ	**ad19**	36	653
ΝΟΣΟΚΟΜΕΙΟ - ΚΛΙΝΙΚΗ	**ad20**	58	997
ΞΕΝΟΔΟΧΕΙΟ	**ad21**	186	3368
ΟΙΚΟΣ ΕΥΓΗΡΙΑΣ – ΟΡΦΑΝΟΤΡ.	**ad22**	14	184
ΠΑΝΕΠΙΣΤΗΜΙΟ	**ad23**	6	117
ΠΟΛΥΚΑΤΟΙΚΙΑ	**ad24**	2754	24510
ΠΡΕΣΒΕΙΑ	**ad25**	19	288
ΣΧΟΛΕΙΟ	**ad26**	477	3232
ΣΧΟΛΗ - ΕΚΠΑΙΔΕΥΤΗΡΙΟ	**ad27**	58	873
ΤΡΑΠΕΖΑ	**ad28**	17	256

Αρχικά, δημιουργήθηκαν οι βάσεις δεδομένων για κάθε μια κατηγορία ξεχωριστά. Τα επιλεγμένα dataset προς ανάλυση ονομάστηκαν αντίστοιχα με τις κατηγορίες. Κάθε dataset περιλαμβάνει τόσες καταχωρήσεις κατανάλωσης όσο και το πλήθος μετρήσεων κατανάλωσης που φαίνεται στον παραπάνω πίνακα.

Παρακάτω παρουσιάζονται τα αποτελέσματα των αλγορίθμων για κάθε κατηγορία.

7.4.1 Αθλητικές εγκαταστάσεις (ad01). Σε αυτή τη κατηγορία ανήκουν αθλητικοί χώροι, γήπεδα, στάδια, γυμναστήρια.

	IBk2	KStar	Bagging
Correlation coefficient	0,9203	0,8888	0,8968
Mean absolute error	775,4	921,3	949,6
Root mean squared error	2333,7	2729,0	2634,5
Relative absolute error	24,70%	29,34%	30,24%
Root relative squared error	39,00%	45,61%	44,03%
Total Number of Instances	255	255	255
Time (Seconds)	1	1	2

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 2.

7.4.2 Άλλες υπηρεσίες (ad02). Σε αυτή τη κατηγορία ανήκουν όλοι οι καταναλωτές που δεν ανήκουν στις υπόλοιπες κατηγορίες.

	IBk2	KStar	Bagging
Correlation coefficient	0,9684	0,9638	0,9441
Mean absolute error	166,0	192,2	207,6
Root mean squared error	555,0	624,2	756,8
Relative absolute error	23,93%	27,71%	29,93%
Root relative squared error	25,00%	28,12%	34,10%
Total Number of Instances	2067	2067	2067
Time (Seconds)	2	19	6

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 2.

7.4.3 Αναψυκτήριο (ad03). Σε αυτή τη κατηγορία ανήκουν αναψυκτήρια, περίπτερα, υπαίθριες καφετέριες.

	IBk2	KStar	Bagging
Correlation coefficient	0,7648	0,7875	0,7833
Mean absolute error	137,8	129,9	127,7
Root mean squared error	239,2	213,6	215,6
Relative absolute error	48,95%	46,15%	45,37%
Root relative squared error	68,74%	61,40%	62,00%
Total Number of Instances	208	208	208
Time (Seconds)	1	1	2

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος KStar.

7.4.4 Αρτοποιείο - Ζαχαροπλαστείο (ad04). Σε αυτή τη κατηγορία ανήκουν όλα τα αρτοποιεία, φούρνοι, ζαχαροπλαστεία, εργαστήρια αρτοποιίας & ζαχαροπλαστικής.

	IBk2	KStar	Bagging
Correlation coefficient	0,8891	0,8772	0,9041
Mean absolute error	240,0	383,9	224,4
Root mean squared error	561,0	630,8	520,6
Relative absolute error	27,27%	43,48%	25,39%
Root relative squared error	46,00%	51,82%	42,76%
Total Number of Instances	7756	7756	7756
Time (Seconds)	32	227	1238

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος Bagging.

7.4.5 Αναψυκτήριο (ad05). Σε αυτή τη κατηγορία ανήκουν οι καταναλωτές που χρησιμοποιούν το φυσικό αέριο για αυτόνομη θέρμανση, μαγείρεμα και θέρμανση νερού.

	IBk2	KStar	Bagging
Correlation coefficient	0,8865	0,8666	0,861
Mean absolute error	82,9	78,6	85,1
Root mean squared error	129,5	140,1	142,5
Relative absolute error	48,16%	45,70%	49,47%
Root relative squared error	46,18%	49,93%	50,82%
Total Number of Instances	388	388	388
Time (Seconds)	1	1	1

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 2.

7.4.6 Βιομηχανία (ad06). Σε αυτή τη κατηγορία ανήκουν μικρές βιομηχανίες και μικροί σταθμοί παραγωγής.

	IBk2	KStar	Bagging
Correlation coefficient	0,88	0,8806	0,8937
Mean absolute error	1201,0	1187,1	1203,9
Root mean squared error	2343,0	2324,3	2201,1
Relative absolute error	32,50%	32,10%	32,56%
Root relative squared error	47,87%	47,46%	44,95%
Total Number of Instances	311	311	311
Time (Seconds)	1	1	53

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος Bagging.

7.4.7 Βιοτεχνία (ad07). Σε αυτή τη κατηγορία ανήκουν βιοτεχνίες και επιχειρήσεις παραγωγής προϊόντων σε μικρή κλίμακα.

	IBk2	KStar	Bagging
Correlation coefficient	0,9071	0,9035	0,9261
Mean absolute error	185,0	207,0	173,3
Root mean squared error	793,0	813,0	705,0
Relative absolute error	22,75%	25,42%	21,28%
Root relative squared error	42,43%	43,50%	37,74%
Total Number of Instances	2623	2623	2623
Time (Seconds)	4	30	160

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος Bagging.

7.4.8 Γραφείο – Εταιρεία (ad08). Σε αυτή τη κατηγορία ανήκουν γραφεία εταιρειών, επαγγελματιών, μικρών επιχειρήσεων, δικηγορικών γραφείων και όλα τα συναφή.

	IBk2	KStar	Bagging
Correlation coefficient	0,6772	0,4739	0,4677
Mean absolute error	485,0	1087,0	1165,0
Root mean squared error	1938,0	2694,0	2616,0
Relative absolute error	32,83%	73,00%	78,90%
Root relative squared error	64,45%	89,00%	87,00%
Total Number of Instances	16254	16254	16254
Time (Seconds)	560	1459	11604

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 2.

7.4.9 Δημόσιο κτήριο (ad09). Σε αυτή τη κατηγορία ανήκουν μεγάλα κτήρια που ανήκουν στο ελληνικό δημόσιο και χρησιμοποιούν φυσικό αέριο για θέρμανση αλλά και συμπαραγωγή ηλεκτρικής ενέργειας.

	IBk2	KStar	Bagging
Correlation coefficient	0,7677	0,4939	0,4977
Mean absolute error	485,0	1087,0	1165,0
Root mean squared error	1938,0	2694,0	2616,0
Relative absolute error	32,83%	73,00%	78,90%
Root relative squared error	64,45%	89,00%	87,00%
Total Number of Instances	903	903	903
Time (Seconds)	1	1	104

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 2.

7.4.10 Εμπορικό κατάστημα (ad10). Σε αυτή τη κατηγορία ανήκουν εμπορικά καταστήματα, μαγαζιά, καταστήματα χονδρικής και λιανικής πώλησης.

	IBk2	KStar	Bagging
Correlation coefficient	0,8461	0,574	0,5719
Mean absolute error	232,0	392,0	405,0
Root mean squared error	643,0	959,0	953,0
Relative absolute error	37,00%	62,40%	64,44%
Root relative squared error	55,31%	82,44%	81,91%
Total Number of Instances	579	579	579
Time (Seconds)	1	1	57

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 2.

7.4.11 Εργαστήριο (ad11). Σε αυτή τη κατηγορία ανήκουν πάσης φύσεως εργαστήρια.

	IBk2	KStar	Bagging
Correlation coefficient	0,9407	0,9428	0,9518
Mean absolute error	124,4	130	114,5
Root mean squared error	451,8	440	405,0
Relative absolute error	20,74%	21,7	19,00%
Root relative squared error	34,19%	33,3	30,67%
Total Number of Instances	1407	1407	1407
Time (Seconds)	2	13	167

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος Bagging.

7.4.12 Εστιατόριο - Μεζεδοπωλείο (ad12). Σε αυτή τη κατηγορία ανήκουν εστιατόρια, μεζεδοπωλεία, ταβέρνες, χώροι δεξιώσεων και συνεστιάσεων.

	IBk2	KStar	Bagging
Correlation coefficient	0,9043	0,8783	0,8321
Mean absolute error	360,3	593	713
Root mean squared error	692,4	915	1215
Relative absolute error	32,70%	53,85	61,35
Root relative squared error	43,06%	56,9	71,1
Total Number of Instances	17586	17586	17586
Time (Seconds)	116	1058	5058

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 2.

7.4.13 Θέατρο - Σινεμά (ad13). Σε αυτή τη κατηγορία ανήκουν θέατρα, κινηματογράφοι, θεατρικοί και συναυλιακοί χώροι.

	IBk2	KStar	Bagging
Correlation coefficient	0,8893	0,8988	0,8723
Mean absolute error	189,3	177,0	237,0
Root mean squared error	429,0	410,0	458,0
Relative absolute error	29,00%	27,20%	36,43%
Root relative squared error	45,00%	43,50%	48,53%
Total Number of Instances	189	189	189
Time (Seconds)	1	3	33

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος KStar.

7.4.14 Ιατρείο (ad14). Σε αυτή τη κατηγορία ανήκουν ιδιωτικά ιατρεία.

	IBk2	KStar	Bagging
Correlation coefficient	0,0466	-0,0568	-0,0369
Mean absolute error	6,0	6,6	6,3
Root mean squared error	9,0	10,0	8,3
Relative absolute error	89,00%	99,00%	94,00%
Root relative squared error	104,00%	115,00%	96,00%
Total Number of Instances	16	16	16
Time (Seconds)	1	1	1

Παρατηρούμε ότι τα αποτελέσματα των αλγορίθμων είναι πάρα πολύ κακά και αποτυγχάνουν τελείως να αναπτύξουν κανόνες για αυτή τη κατηγορία. Αυτό συμβαίνει γιατί σε αυτή τη κατηγορία καταναλωτών υπάρχουν μόνο 16 μετρήσεις με αποτέλεσμα την αδυναμία των αλγορίθμων να συσχετίσουν επαρκώς τα δεδομένα. Είναι σκόπιμο να απαλειφθεί η κατηγορία αυτή και να μεταφερθούν οι καταχωρήσεις σε άλλη κατηγορία όπως για παράδειγμα στην κατηγορία «Γραφείο - Εταιρεία».

7.4.15 Ιδιώτης (ad15). Σε αυτή τη κατηγορία ανήκουν όλοι οι οικιακοί καταναλωτές που χρησιμοποιούν το φυσικό αέριο για οικιακή χρήση.

	IBk3	KStar	Bagging
Correlation coefficient	0,9065		
Mean absolute error	652,0		
Root mean squared error	1109,0	Αδύνατη η	επεξεργασία
Relative absolute error	24,02%	των	δεδομένων
Root relative squared error	37,11%		
Total Number of Instances	401123		
Time (Seconds)	23:12:45		

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 3. Μόνο ο αλγόριθμος IBk μπόρεσε να επεξεργαστεί τα δεδομένα. Οι άλλοι δύο αλγόριθμοι δεν ήταν ικανοί να διαχειριστούν τόσο μεγάλη βάση δεδομένων, καθώς οι καταχωρήσεις της κατανάλωσης είναι 401.123.

7.4.16 Ιερός ναός (ad16). Σε αυτή τη κατηγορία ανήκουν ιεροί ναοί, εκκλησίες, μοναστήρια.

	IBk2	KStar	Bagging
Correlation coefficient	0,9234	0,9195	0,7866
Mean absolute error	152,0	185,0	337,0
Root mean squared error	399,0	482,0	764,0
Relative absolute error	23,80%	27,36%	49,90%
Root relative squared error	38,45%	40,84%	64,70%
Total Number of Instances	802	273	273
Time (Seconds)	1	1	22

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 2.

7.4.17 Καθαριστήρια (ad17). Σε αυτή τη κατηγορία ανήκουν τα καθαριστήρια ειδών ρουχισμού και κλινοσκεπασμάτων.

	IBk2	KStar	Bagging
Correlation coefficient	0,8228	0,8536	0,8708
Mean absolute error	354,0	335,0	307,0
Root mean squared error	1013,0	888,0	831,0
Relative absolute error	39,00%	36,00%	33,89%
Root relative squared error	59,00%	52,00%	49,00%
Total Number of Instances	1188	1188	1188
Time (Seconds)	1	1	150

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος Bagging.

7.4.18 Καφέ - Μπαρ (ad18). Σε αυτή τη κατηγορία ανήκουν καφετέριες, μπαρ, κλαμπ, χώροι διασκέδασης, νυχτερινά κέντρα.

	IBk2	KStar	Bagging
Correlation coefficient	0,876	0,8653	0,8528
Mean absolute error	158,0	180,0	170,0
Root mean squared error	426,0	443,0	460,0
Relative absolute error	29,88%	34,15%	32,30%
Root relative squared error	48,40%	50,36%	52,20%
Total Number of Instances	4131	4131	4131
Time (Seconds)	14	112	480

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 2.

7.4.19 Κεντρική θέρμανση (ad19). Σε αυτή τη κατηγορία ανήκουν όλοι οι οικιακοί καταναλωτές που χρησιμοποιούν το φυσικό αέριο για το σύνολο του κτηρίου όπως για παράδειγμα οι πολυκατοικίες.

	IBk2	KStar	Bagging
Correlation coefficient	0,9281	0,9433	0,9202
Mean absolute error	267,8	250,0	342,0
Root mean squared error	537,0	477,0	557,0
Relative absolute error	23,26%	21,74%	29,75%
Root relative squared error	37,88%	33,65%	39,00%
Total Number of Instances	653	653	653
Time (Seconds)	2	3	120

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος KStar.

7.4.20 Νοσοκομείο - Κλινική (ad20). Σε αυτή τη κατηγορία ανήκουν νοσοκομεία, κλινικές, κέντρα θεραπείας, νοσηλευτικοί χώροι.

	IBk2	KStar	Bagging
Correlation coefficient	0,9763	0,9712	0,9661
Mean absolute error	222,0	312,0	285,0
Root mean squared error	618,0	720,0	741,0
Relative absolute error	12,99%	18,28%	16,71%
Root relative squared error	21,66%	25,20%	26,00%
Total Number of Instances	997	997	997
Time (Seconds)	2	9	141

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 2.

7.4.21 Ξενοδοχείο (ad21). Σε αυτή τη κατηγορία ανήκουν όλοι οι ξενοδοχειακές μονάδες.

	IBk2	KStar	Bagging
Correlation coefficient	0,9145	0,9124	0,9207
Mean absolute error	772,0	1117,0	837,0
Root mean squared error	1809,0	2000,0	1726,0
Relative absolute error	25,89%	37,42%	28,00%
Root relative squared error	40,89%	45,22%	39,00%
Total Number of Instances	3368	3368	3368
Time (Seconds)	11	78	434

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος Bagging.

7.4.22 Οίκος ευγηρίας - Ορφανοτροφείο (ad22). Σε αυτή τη κατηγορία ανήκουν οι οίκοι ευγηρίας, τα ορφανοτροφεία και άλλα.

	IBk2	KStar	Bagging
Correlation coefficient	0,9741	0,9817	0,93
Mean absolute error	399,7	348,3	638,0
Root mean squared error	714,0	639,0	1177,0
Relative absolute error	18,18%	15,85%	29,00%
Root relative squared error	22,51%	20,14%	37,00%
Total Number of Instances	184	184	184
Time (Seconds)	1	1	10

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος KStar.

7.4.23 Πανεπιστήμιο (ad23). Σε αυτή τη κατηγορία ανήκουν όλα τα πανεπιστημιακά ιδρύματα και χώροι.

	IBk2	KStar	Bagging
Correlation coefficient	0,9609	0,9582	0,9349
Mean absolute error	101,0	100,0	112,4
Root mean squared error	303,0	342,0	412,0
Relative absolute error	17,18%	17,05%	19,00%
Root relative squared error	27,39%	30,96%	37,25%
Total Number of Instances	117	117	117
Time (Seconds)	1	1	6

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 2.

7.4.24 Πολυκατοικία (ad24). Σε αυτή τη κατηγορία ανήκουν πολυκατοικίες.

	IBk2	KStar	Bagging
Correlation coefficient	0,8673	0,8631	0,8345
Mean absolute error	519,0	561,0	629,0
Root mean squared error	1090,0	1117,8	1490,0
Relative absolute error	31,96%	34,58%	39,96%
Root relative squared error	50,78%	52,00%	56,78%
Total Number of Instances	24510	24510	24510
Time (Seconds)	470	1800	5470

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 2.

7.4.25 Πρεσβεία (ad25). Σε αυτή τη κατηγορία ανήκουν όλες οι πρεσβείες και τα προξενεία.

	IBk2	KStar	Bagging
Correlation coefficient	0,8749	0,8763	0,8494
Mean absolute error	568,0	532,0	668,0
Root mean squared error	1730,0	1674,0	1831,0
Relative absolute error	24,80%	23,26%	29,16%
Root relative squared error	49,00%	48,00%	52,57%
Total Number of Instances	288	288	288
Time (Seconds)	1	2	49

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος KStar.

7.4.26 Σχολείο (ad26). Σε αυτή τη κατηγορία ανήκουν όλα τα ιδιωτικά και δημόσια σχολεία.

	IBk2	KStar	Bagging
Correlation coefficient	0,7332	0,8292	0,5236
Mean absolute error	461,0	342,0	723,0
Root mean squared error	983,0	803,0	1238,8
Relative absolute error	47,39%	35,12%	74,24%
Root relative squared error	68,82%	56,24%	86,69%
Total Number of Instances	3232	3232	3232
Time (Seconds)	5	49	273

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος KStar.

7.4.27 Σχολή - Εκπαιδευτήριο (ad27). Σε αυτή τη κατηγορία ανήκουν όλες οι σχολές, τα φροντιστήρια και τα εκπαιδευτήρια πάσης φύσεως.

	IBk2	KStar	Bagging
Correlation coefficient	0,9016	0,9135	0,8793
Mean absolute error	374,0	331,0	447,0
Root mean squared error	879,9	816,0	960,0
Relative absolute error	27,59%	24,40%	32,93%
Root relative squared error	43,79%	40,60%	47,82%
Total Number of Instances	873	873	873
Time (Seconds)	1	2	119

Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος KStar.

7.4.28 Τράπεζα (ad28). Σε αυτή τη κατηγορία ανήκουν τράπεζες.

	IBk3	KStar	Bagging
Correlation coefficient	0,9167	0,9145	0,9075
Mean absolute error	802,0	765,0	954,0
Root mean squared error	1640,0	1657,0	17171,0
Relative absolute error	26,28%	25,05%	31,27%
Root relative squared error	39,99%	40,40%	41,87%
Total Number of Instances	256	256	256
Time (Seconds)	1	1	10

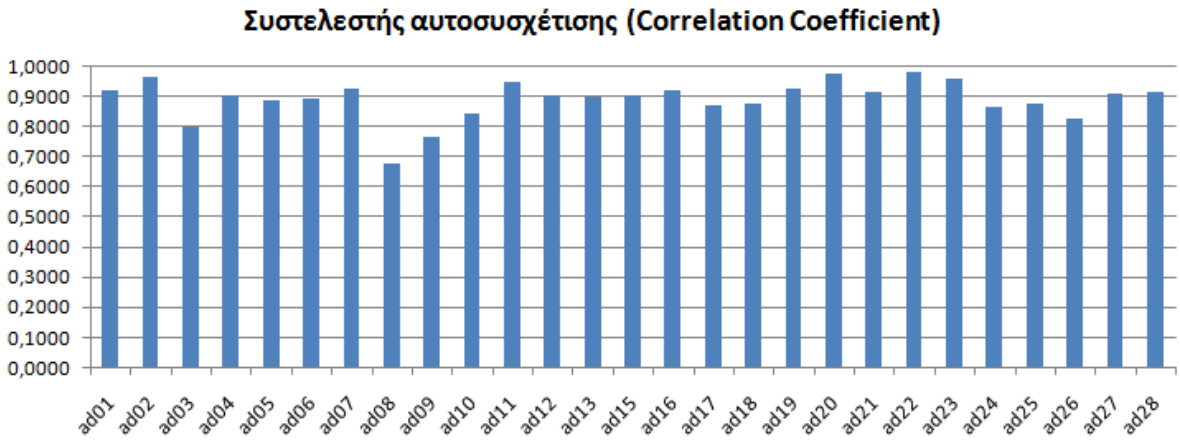
Παρατηρούμε ότι το καλύτερο αποτέλεσμα δίνει ο αλγόριθμος IBk με KNN 3.

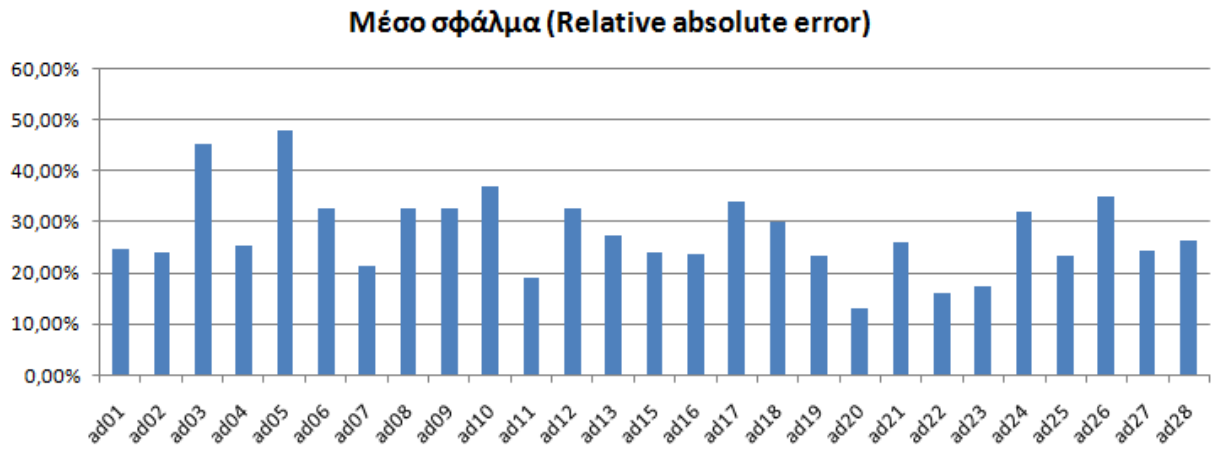
7.4.29 Ανάλυση των αποτελεσμάτων

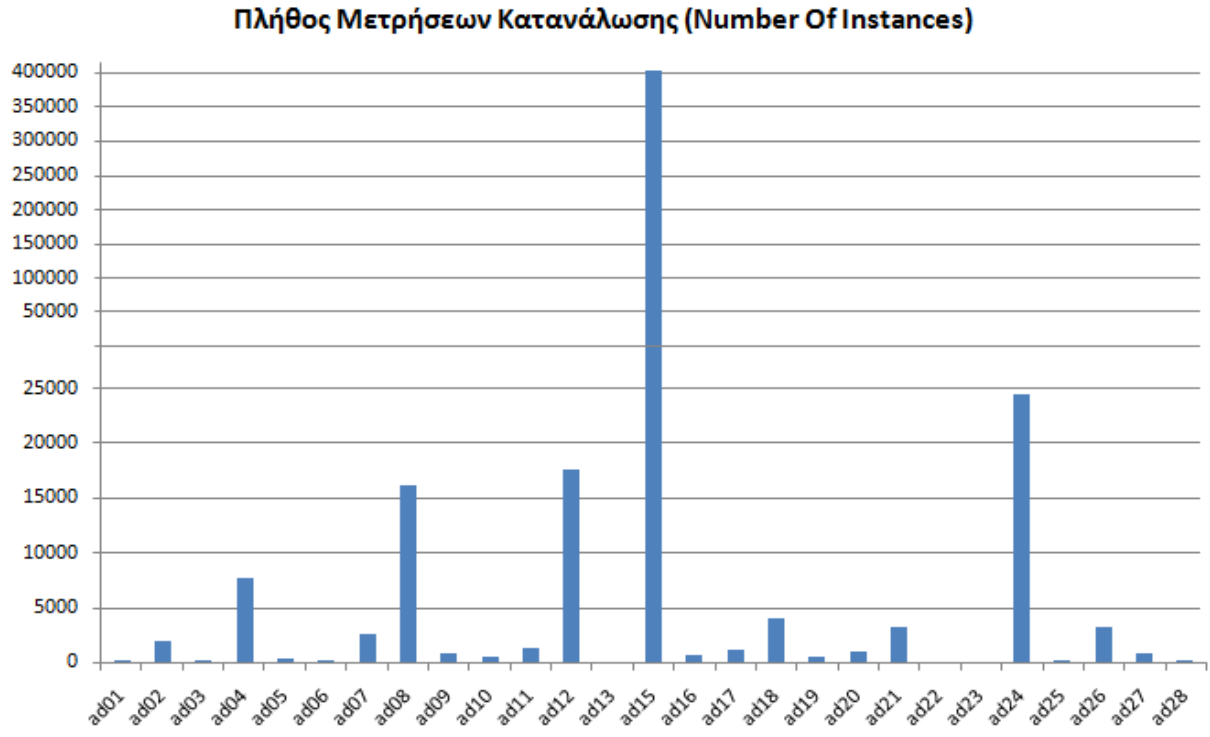
Παρατηρούμε ότι σε κάποιες κατηγορίες έχουμε εξαιρετικά αποτελέσματα, ακόμα και με συντελεστή αυτοσυσχέτισης άνω του 0,95. Τα αποτελέσματα της πρόβλεψης είναι πολύ καλά στις κατηγορίες που περιλαμβάνουν καταναλωτές που χρησιμοποιούν το φυσικό αέριο κυρίως για θέρμανση.

Από την εφαρμογή των αλγορίθμων παρατηρούμε ότι κάθε φορά παρουσιάζεται διαφορετικός αλγόριθμος ικανότερος ανάλογα με την κατηγορία. Παρακάτω παρουσιάζονται τα αποτελέσματα των είκοσι οκτώ κατηγοριών καταναλωτών και των καταλληλότερων αλγόριθμων.

Κατηγορία	*Συντελεστής αυτοσυσχέτισης*	*Μέσο σφάλμα*	*Ικανότερος αλγόριθμος*	*Πλήθος Μετρήσεων Κατανάλωσης*
ad01	0,9203	24,70%	IBk2	255
ad02	0,9684	23,93%	IBk2	2067
ad03	0,7875	46,15%	Kstar	208
ad04	0,9041	25,39%	Bagging	7756
ad05	0,8865	48,16%	IBk2	388
ad06	0,8937	32,56%	Bagging	311
ad07	0,9261	21,28%	Bagging	2623
ad08	0,6772	32,83%	IBk2	16254
ad09	0,7677	32,83%	IBk2	903
ad10	0,8461	37,00%	IBk2	579
ad11	0,9518	19,00%	Bagging	1407
ad12	0,9043	32,70%	IBk2	17586
ad13	0,8988	27,20%	Kstar	189
ad15	0,9065	24,02%	IBk3	401123
ad16	0,9234	23,80%	IBk2	802
ad17	0,8708	33,89%	Bagging	1188
ad18	0,8760	29,88%	IBk2	4131
ad19	0,9281	23,26%	Kstar	653
ad20	0,9763	12,99%	IBk2	997
ad21	0,9145	25,89%	IBk2	3368
ad22	0,9817	15,85%	Kstar	184
ad23	0,9609	17,18%	IBk1	117
ad24	0,8673	31,96%	IBk1	24510
ad25	0,8763	23,26%	Kstar	288
ad26	0,8292	35,12%	Kstar	3232
ad27	0,9135	24,40%	Kstar	873
ad28	0,9167	26,28%	IBk3	256


Συστελεστής αυτοσυσχέτισης (Correlation Coefficient)
1,0000
0,9000
0,8000
0,7000
0,6000
0,5000
0,4000
0,3000
0,2000
0,1000
0,0000
ad01 ad02 ad03 ad04 ad05 ad06 ad07 ad08 ad09 ad10 ad11 ad12 ad13 ad15 ad16 ad17 ad18 ad19 ad20 ad21 ad22 ad23 ad24 ad25 ad26 ad27 ad28


Μέσο σφάλμα (Relative absolute error)
60,00%
50,00%
40,00%
30,00%
20,00%
10,00%
0,00%
ad01 ad02 ad03 ad04 ad05 ad06 ad07 ad08 ad09 ad10 ad11 ad12 ad13 ad15 ad16 ad17 ad18 ad19 ad20 ad21 ad22 ad23 ad24 ad25 ad26 ad27 ad28


Πλήθος Μετρήσεων Κατανάλωσης (Number Of Instances)
400000
350000
300000
250000
200000
150000
100000
50000
25000
20000
15000
10000
5000
0
ad01 ad02 ad03 ad04 ad05 ad06 ad07 ad08 ad09 ad10 ad11 ad12 ad13 ad15 ad16 ad17 ad18 ad19 ad20 ad21 ad22 ad23 ad24 ad25 ad26 ad27 ad28

Παρατηρούμε ότι σε 15 κατηγορίες έχουμε εξαιρετικά αποτελέσματα με συντελεστή αυτοσυσχέτισης άνω του 0,9. Ακόμα, σε 5 κατηγορίες ο συντελεστής ανεβαίνει και πάνω από το 0,95. Αυτό σημαίνει ότι ο αλγόριθμος είναι σε θέση να εκπαιδευτεί πολύ καλά, να συσχετίσει τα δεδομένα και να εξάγει κανόνες που οδηγούν σε μια ικανοποιητικότατη πρόβλεψη.

Κατηγορία	*Συντελεστής αυτοσυσχέτισης*	*Μέσο σφάλμα*	*Ικανότερος αλγόριθμος*	*Πλήθος Μετρήσεων Κατανάλωσης*
ΑΘΛΗΤΙΚΕΣ ΕΓΚΑΤΑΣΤΑΣΕΙΣ	0,9203	24,70%	IBk2	255
ΑΛΛΕΣ ΥΠΗΡΕΣΙΕΣ	**0,9684**	23,93%	IBk2	2067
ΑΡΤΟΠΟΙΕΙΟ – ΖΑΧΑΡΟΠΛ.	0,9041	25,39%	Bagging	7756
ΒΙΟΤΕΧΝΙΑ	0,9261	21,28%	Bagging	2623
ΕΡΓΑΣΤΗΡΙΟ	**0,9518**	19,00%	Bagging	1407
ΕΣΤΙΑΤΟΡΙΟ - ΜΕΖΕΔΟΠΩΛΕΙΟ	0,9043	32,70%	IBk2	17586
ΙΔΙΩΤΗΣ	0,9065	24,02%	IBk3	401123
ΙΕΡΟΣ ΝΑΟΣ	0,9234	23,80%	IBk2	802
ΚΕΝΤΡΙΚΗ ΘΕΡΜΑΝΣΗ	0,9281	23,26%	Kstar	653
ΝΟΣΟΚΟΜΕΙΟ - ΚΛΙΝΙΚΗ	**0,9763**	12,99%	IBk2	997
ΞΕΝΟΔΟΧΕΙΟ	0,9145	25,89%	IBk2	3368
ΟΙΚΟΣ ΕΥΓΗΡΙΑΣ – ΟΡΦΑΝΟΤΡ.	**0,9817**	15,85%	Kstar	184
ΠΑΝΕΠΙΣΤΗΜΙΟ	**0,9609**	17,18%	IBk1	117
ΣΧΟΛΗ - ΕΚΠΑΙΔΕΥΤΗΡΙΟ	0,9135	24,40%	Kstar	873
ΤΡΑΠΕΖΑ	0,9167	26,28%	IBk3	256

Ακόμα, σε 5 κατηγορίες ο συντελεστής ανεβαίνει και πάνω από το 0,95, όπως φαίνεται στον παραπάνω πίνακα. Σημαντικός παράγοντας είναι ότι τα εργαστήρια, τα νοσοκομεία, οι οίκοι ευγηρίας και τα πανεπιστήμια έχουν ανοιχτό το σύστημα θέρμανσης πολλές ώρες οπότε η κατανομή της κατανάλωσης είναι ομοιόμορφη.

Είναι, επίσης εμφανές από τα διαγράμματα, ότι όσο αυξάνεται το μέγεθος του δείγματος, τόσο αυξάνεται η ακρίβεια της πρόβλεψης. Όσο μειώνεται το πλήθος των μετρήσεων της κατανάλωσης, τόσο μειώνεται η ακρίβεια της πρόβλεψης. Είναι λογικό να αυξάνεται το ποσοστό ακρίβειας της πρόβλεψης, όσο πιο πολλές μετρήσεις υπάρχουν στη βάση δεδομένων ούτος ώστε να υπάρχουν πιο πολλά δεδομένα για την εξαγωγή κανόνων και συσχετίσεων. Επίσης, όσο πιο πολλές καταμετρήσεις κατανάλωσης είναι διαθέσιμες για τον ίδιο καταναλωτή, τόσο καλύτερη είναι η πρόβλεψη. Αυτό είναι λογικό, καθώς ο αλγόριθμος έχει περισσότερες πληροφορίες για τον καταναλωτή και μπορεί να συσχετίσει τα δεδομένα για πιο πολλές περιπτώσεις.

Παρατηρούμε επίσης ότι ο συντελεστής συσχέτισης είναι πολύ χαμηλός στην κατηγορία «Γραφείο - Εταιρεία» (ad08). Η κατηγορία αυτή περιλαμβάνει πολλούς καταναλωτές που παρουσιάζουν διαφορετική καταναλωτική συμπεριφορά. Επίσης, τα μεγέθη της κατανάλωσης αλλάζουν. Στην κατηγορία αυτή είναι δυνατόν να ανήκουν μικρά γραφεία αλλά και ολόκληροι όροφοι γραφείων ή εταιρείες. Είναι καλύτερο να χωριστεί η κατηγορία αυτή στα δύο ώστε να ομαδοποιηθούν καλύτερα οι καταναλωτές και να βελτιωθεί η πρόβλεψη.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Σκοπός αυτής της μελέτης ήταν η ανάπτυξη και αξιολόγηση μοντέλων πρόβλεψης της αναμενόμενης κατανάλωσης φυσικού αερίου οικιακών και μικρών βιομηχανικών πελατών για την περιοχή της Αττικής.

Όπως φαίνεται από τα αποτελέσματα των αλγορίθμων έγινε μια πολύ καλή ανάλυση των δεδομένων και των παραγόντων που επηρεάζουν την κατανάλωση. Προκλήσεις και προβλήματα όπως για παράδειγμα η έλλειψη μετεωρολογικών δεδομένων, ξεπεράστηκαν επιτυχώς,.

Επιλέχθηκαν τα πλέον κατάλληλα και σημαντικά κριτήρια (χαρακτηριστικά) που επηρεάζουν στο μέγιστο την κατανάλωση, καθώς σε περίπτωση που επιλεγόταν μικρό πλήθος κριτηρίων θα υπήρχε ο κίνδυνος να μην είναι ακριβές το εξαγόμενο αποτέλεσμα. Στην αντίθετη περίπτωση, δηλαδή να επιλεγόταν μεγάλο πλήθος κριτηρίων, η βελτίωση στην ακρίβεια του αποτελέσματος μπορεί να είναι αμελητέα. Οπότε είναι σημαντικό που επιλέχθηκαν εκείνα τα κριτήρια που πραγματικά επηρεάζουν την κατανάλωση στο μεγαλύτερο βαθμό, ώστε το επιθυμητό αποτέλεσμα να προκύψει με το λιγότερο δυνατό κόπο και τη μεγαλύτερη δυνατή ακρίβεια.

Καθορίστηκαν επιτυχώς ποιοι αλγόριθμοι είναι οι καταλληλότεροι για την επεξεργασία των δεδομένων μας. Εξετάστηκαν πάνω από έντεκα αλγόριθμοι, ύστερα από αρχική επιλογή πενήντα αλγορίθμων, για να ανακαλύψουμε ποιοι από αυτούς είναι οι ικανότεροι να προβλέψουν την κατανάλωση φυσικού αερίου. Τελικά, καταλήξαμε στους τρεις καλύτερους και ικανότερους αλγόριθμους, μετά από επιλογή εκείνων με το μικρότερο σφάλμα και τον υψηλότερο συντελεστή συσχέτισης, αναλογιζόμενοι και το χρόνο επεξεργασίας των δεδομένων. Συγκριτικά με τους υπόλοιπους αλγορίθμους που εξετάστηκαν, οι επιλεχθέντες αλγόριθμοι είναι αυτοί που χρειάζονται το λιγότερο χρόνο για την επεξεργασία μεγάλων βάσεων δεδομένων.

Από τα αποτελέσματα φάνηκε η ισχυρή εξάρτηση της κατανάλωσης από την εξωτερική θερμοκρασία. Αυτό προκύπτει από το γεγονός ότι στην Ελλάδα το φυσικό αέριο χρησιμοποιείται κυρίως για θέρμανση, είτε αναφερόμαστε σε οικιακούς, είτε σε μικρούς επαγγελματικούς καταναλωτές. Οι καταναλωτές επιδιώκουν να εξασφαλίσουν ένα άνετο εσωτερικό κλίμα πλήρως προσαρμοσμένο στις ανάγκες τους.

Κατά την εφαρμογή των τριών επιλεχθέντων αλγορίθμων στο σύνολο των δεδομένων, παρατηρήσαμε ότι ακόμα και σε μια τεράστια βάση δεδομένων (492.259 καταμετρήσεις) τα αποτελέσματα των αλγορίθμων είναι πολύ ακριβή. Συμμεριζόμενοι το γεγονός ότι στη βάση δεδομένων περιλαμβάνονται όλες οι κατηγορίες καταναλωτών και χρήσεως φυσικού αερίου αλλά και οι καταναλωτές που η κατανάλωσή τους δεν είναι ανάλογη της θερμοκρασίας, καταλήγουμε στο ότι οι αλγόριθμοι είναι πολύ ικανοί και επιτυχείς στις προβλέψεις τους.

Κατά την εφαρμογή των τριών επιλεχθέντων αλγορίθμων ανά κατηγορία χρήσης φυσικού αερίου, παρατηρήσαμε ότι οι αλγόριθμοι συσχετίζουν εύκολα τα δεδομένα και είναι ικανοί να δώσουν μια πολύ ακριβής πρόβλεψη. Στις κατηγορίες χρήσης όπου η εξάρτηση από την εξωτερική θερμοκρασία ήταν υψηλή, αυξανόταν η ικανότητα πρόβλεψης του αλγορίθμου λόγω της άμεσης συσχέτισης βαθμοημέρες θέρμανσης - κατανάλωσης.

Όσον αφορά την εφαρμογή των τριών επιλεχθέντων αλγορίθμων ανά κατηγορία καταναλωτών φυσικού αερίου, παρατηρήσαμε ότι σε ορισμένες κατηγορίες οι αλγόριθμοι συσχετίζουν εύκολα τα δεδομένα και είναι ικανοί να δώσουν εξαιρετικά ακριβής προβλέψεις. Στις κατηγορίες καταναλωτών όπου η εξάρτηση από την εξωτερική θερμοκρασία ήταν υψηλή, αυξανόταν η ικανότητα πρόβλεψης του αλγορίθμου λόγω της άμεσης συσχέτισης βαθμοημέρες θέρμανσης - κατανάλωσης.

Παρατηρήσαμε δηλαδή ότι όταν εφαρμόζονται οι αλγόριθμοι ανά κατηγορίες, προκύπτουν καλύτερα αποτελέσματα καθώς οι καταναλωτές παρουσιάζουν ομοειδή καταναλωτική συμπεριφορά. Είναι σκόπιμη, επομένως, η σωστή επιλογή των κατηγοριών και της ομαδοποίησης των καταναλωτών.

Έγινε αντιληπτό ότι όσο πιο μικρή είναι η βάση δεδομένων, τόσο πιο πενιχρή θα είναι η ακρίβεια της πρόβλεψης καθώς οι αλγόριθμοι θα αποτυγχάνουν να αναπτύξουν κανόνες για αυτή τη κατηγορία και να συσχετίσουν επαρκώς τα δεδομένα. Αντίστοιχα όσο αυξάνεται το μέγεθος του δείγματος, εν προκειμένω το πλήθος των μετρήσεων της κατανάλωσης, τόσο αυξάνεται η ακρίβεια της πρόβλεψης καθώς υπάρχουν πιο πολλά δεδομένα για την εξαγωγή κανόνων και συσχετίσεων.

Επίσης, όσο πιο πολλές καταμετρήσεις κατανάλωσης είναι διαθέσιμες για τον ίδιο καταναλωτή, τόσο καλύτερη είναι η πρόβλεψη. Αυτό είναι λογικό, καθώς ο αλγόριθμος έχει περισσότερες πληροφορίες για τον καταναλωτή και μπορεί να συσχετίσει τα δεδομένα για πιο πολλές περιπτώσεις.

Τέλος, αποδείχτηκε ότι ο καταλληλότερος αλγόριθμος είναι ο IBk των k – κοντινότερων γειτόνων, καθώς στην πλειονότητα των περιπτώσεων αποδείχτηκε ακριβέστερος στις προβλέψεις. Παρόλα αυτά, και οι άλλοι δύο αλγόριθμοι φάνηκαν ικανοί για την εξαγωγή ακριβών προβλέψεων αλλά μόνο όταν εφαρμόζονται σε ορισμένες κατηγορίες καταναλωτών.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική βιβλιογραφία

1. Γαλάνης Νικηφόρος, «Μελέτες θέρμανσης - Από τη θεωρία στην επίλυση με Η/Υ», ΣΕΛΚΑ-4Μ, Τεκδοτική.

2. Κ. Παπακώστας, Γ. Τσιλιγκιρίδης, Ν, Κυριάκης, «Βαθμοημέρες Θέρμανσης 50 Ελληνικών Πόλεων», Τεχν. Χρον. Επιστ. Έκδ. ΤΕΕ, IV, τεύχ. 1-2 2005, Tech. Chron. Sci. J. TCG, IV, No 1-2.

3. Α. Ματζαράκης, Χ. Μπαλαφούτης, «Γεωγραφική κατανομή βαθμοημερών θέρμανσης στον Ελληνικό χώρο για ενεργειακή χρήση», 6th Hellenic Conference in Meteorology, Climatology and Atmospheric Physics.

4. Αχ. Κωστούλας, Ν. Θεοδωρόπουλος, Γ. Βόκας, «Υπολογισμός βαθμοημερών θέρμανσης και ψύξεις για Ελληνικές πόλεις με τη χρήση διάφορων μεθοδολογιών».

5. Παπακώστας Κ.Τ, Εκτίμηση της κατανάλωσης ενέργειας για θέρμανση σε κτίρια κατοικιών στην Αθήνα και τη Θεσσαλονίκη, Πρακτικά 7ου Εθνικού Συνεδρίου Ι.Η.Τ για τις Α.Π.Ε, Πάτρα 2002.

6. Παπακώστας Κ.Τ., Εκτίμηση της κατανάλωσης ενέργειας για θέρμανση κατοικιών, με τη μέθοδο βαθμοημερών μεταβλητής βάσης, πρακ. 6ου Εθνικού Συνεδρίου ΙΗΤ για τις Ήπιες Μορφές Ενέργειας, τόμος Α, Αθήνα, 1999.

7. Παπακώστας Κ.Τ., Συμβολή στην εκτίμηση της κατανάλωσης ενέργειας σε συστήματα θέρμανσης και κλιματισμού στην Ελλάδα, με τη χρήση μμεθόδων απλής και πολλαπλής μμέτρησης, διδακτορική διατριβή, Τμήμα Μηχανολόγων Μηχανικών Α.Π.Θ.,Θεσσαλονίκη, 2001.

8. Παπακώστας Κ.Τ., Μοντέλο υπολογισμού βαθμοημερών θέρμανσης και ψύξης με τυχαία βάση. Βαθμοημέρες Αθηνών - Θεσσαλονίκης, πρακ. 6ου Εθνικού Συνεδρίου ΙΗΤ για τις Ήπιες Μορφές Ενέργειας, τόμος Α, Αθήνα, 1999.

9. ASHRAE, Handbook of Fundamentals, American Society of Heating, Refrigerating and Air Conditioning Engineers, New York: 1993.

10. Η. Βαρλάμης, Β. Βασσάλος, «Αρχιτεκτονική σχεδίαση συστήματος ολοκλήρωσης ετερογενούς σημασιολογικής πληροφορίας και γνώσης», Τμ. Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθήνας, Αθήνα.

Διεθνής βιβλιογραφία

1. F. B. Gorucu, P. U. Geri¸ F. Gumrah, "Artificial Neural Network Modeling for Forecasting Gas Consumption", Energy Sources, 26:299–307, 2004.

2. Alireza Khotanzad, Hassan Elragal, Tsun-Liang Lu, "Combination of Artificial Neural-Network Forecasters for Prediction of Natural Gas Consumption", IEEE transactions on neural networks, Vol. 11, no. 2, March 2000.

3. Jakub Siemek, Stanislaw Nagy, Stanislaw Rychlicki, "Estimation of natural-gas consumption in Poland based on the logistic-curve interpretation", Applied Energy 75 (2003) 1–7.

4. F. B. Gorucu, F. Gumrah, "Evaluation and Forecasting of Gas Consumption by Statistical Analysis", Energy Sources, 26:267–276, 2004.

5. Mahmoud A. Kaboudan, Qingfeng Wilson Liu, "Forecasting quarterly U.S. demand for natural gas", ITEM, Volume 2, No 1, 2004, Paper 4.

6. R. Gutierrez, A. Nafidi, R. Gutierrez Sanchez, "Forecasting total natural-gas consumption in Spain by using the stochastic Gompertz innovation diffusion model", Applied Energy 80 (2005) 115–124.

7. S. Gil, J. Deferrari, "Generalized Model of Prediction of Natural Gas Consumption", Journal of Energy Resources Technology, Vol. 126, June 2004.

8. Ronald H. Brown, Brian M. Marx, George F. Corliss, "Mathematical Models for Gas Forecasting", Department of Electrical and Computer Engineering, Marquette University, Milwaukee.

9. S.Gil, J. Deferrari, L. Duperron, "Modelo de predicción de consumo de gas natural y evaluación de consumos máximos en Argentina", Contribución al II Workshop of Dynamics of Social and Economic Systems – Instituto de Integración Latinoamericana – Facultad de Ciencias Sociales de la Universidad Nacional de La Plata – 7 al 10 de Agosto de 2001.

10. James C. Howland III, Mark S. Voss, "Natural Gas Prediction Using The Group Method of Data Handling", Montana State University, Havre, Montana, USA.

11. Nguyen Hoang Viet, Jacek Mandziuk, "Neural and fuzzy neural networks for natural gas consumption prediction", 2003 IEEE XIII Workshop on Neural Networks for Signal Processing.

12. Petr Musilek, Emil Pelikan, Tomas Brabec, Milan Simunek, “Recurrent Neural Network Based Gating for Natural Gas Load Prediction System”, 2006 International Joint Conference on Neural Networks, Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada, July 16-21, 2006.

13. M. Bodn, “A guide to recurrent neural networks and backpropagation”, Report from the NUTEK-supported project AIS-8: Application of Data Analysis with Learning Systems, 1999-2001. Holst, A. (ed.), SICS Technical Report T2002:03, SICS, Kista, Sweden. 2002.

14. L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, “Classification and Regression Trees”, Wadsworth International Group, Belmont, 1984.

15. C. Gilde, “Time Series Analysis And Prediction Using Recurrent Gated Experts”, MSc Thesis, University of Skovde, Sweden, 1996

16. J. R. Quinlan, “Induction of decision trees”, Machine Learning, 1(1), pp. 81–106

17. Han Liu, Ding Liu, Yan-Ming Liang, Gang Zheng, “Research on natural gas load forecasting based on least squares support vector machine”, School of Automation and Information Engineering, Xian University of Technology, China.

18. A. Tzachanis, V. Kampanas, “An experimental Investigation into the Sufficiency of estimating Energy Consumption in Buildings using Constant and Variable Base Degree-Days Concepts”, Department of Mechanical Engineering, Technological Educational Institute of Larissa.

19. Ian H. Witten, Ian H. Witten, «Data Mining, Practical Machine Learning Tools and Techniques», 2005 by Elsevier Inc.

20. David Scuse, Peter Reutemann, «WEKA Experimenter Tutorial for Version 3-4», 2002-2007, David Scuse and University of Waikato.

21. Richard Kirkby, Eibe Frank, «WEKA Explorer User Guide for Version 3-4», 2002-2007 University of Waikato.