ΤΕΙ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

MSc «Μηχανικων Η/Υ και Συστηματων – Ευφυη Συστηματα και IoT»

Αποδοτική Ανάλυση Ενεργειακών Δεδομένων IOT

Efficient Energy Data Analysis

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΝΙΚΟΛΑΟΣ Γ. ΤΣΙΚΡΙΚΑΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΠΑΠΑΪΩΑΝΝΟΥ ΑΘΑΝΑΣΙΟΣ

ΛΑΡΙΣΑ

ΣΕΠΤΕΜΒΡΙΟΣ 2020

ΠΕΡΙΛΗΨΗ

Το ευφυές ηλεκτρικό πλέγμα στοχεύει στην αρτιότερη διαχείριση του δικτύου ηλεκτρικής ενέργειας, ενόψει των ανανεώσιμων πηγών ενέργειας που καθιστούν ασταθή την παραγωγή, βάσει της ανάλυσης πληθώρας δεδομένων που προέρχονται από αισθητήρες κατά μήκος του δικτύου και από έξυπνους μετρητές στους τελικούς καταναλωτές. Στην εργασία αυτή μελετώνται αποδοτικές μέθοδοι και υλοποιούνται εργαλεία ανάλυσης των δεδομένων που αφορούν την κατανάλωση ηλεκτρικής ενέργειας, έτσι ώστε οι πάροχοι ενέργειας και οι διαχειριστές του δικτύου να μπορούν να ταιριάζουν αποδοτικότερα την παραγωγή με την κατανάλωση. Αρχικά θα κάνουμε μια σύντομη αναφορά στα μεγάλα δεδομένα αλλά και το ρόλο τους, ενώ στην συνέχεια θα αναφερθούμε στο πρόγραμμα Apache Spark ως προς τον τρόπο λειτουργίας του καθώς και το πώς χρησιμοποιείται. Στην συνέχεια θα αναφερθούμε στους αλγορίθμους στους οποίους θα βασιστούμε για να γράψουμε τα προγράμματα μας σε γλώσσα Python.

Υλοποιούμε με κώδικα Python σε Apache Spark τρεις μεθόδους ανάλυσης δεδομένων που απαντούν σε δημοφιλή ερωτήματα διαχειριστών δικτύου και παροχών ηλεκτρικής ενέργειας. Η πρώτη μέθοδος είναι η μέθοδος της γραμμικής παλινδρόμησης, μια μέθοδο στατιστικής ανάλυσης προβλέποντας την σχέση των μη εξαρτώμενων μεταβλητών με τις εξαρτώμενες μεταβλητές. Η δεύτερη μέθοδος αφορά στην εκτέλεση ενοποιημένων ερωτημάτων τα οποία μπορούν να μας δώσουν σημαντικά συμπεράσματα τόσο στατιστικά αλλά και ερευνητικά. Τέλος, ως τρίτη μέθοδο, θα εφαρμόσουμε την ανίχνευση ανωμαλιών με την μέθοδο μιας κλάσης όχι γραμμική με απεικόνιση σχεδιαγράμματος.

Λέξεις Κλειδιά : Μεγάλα Δεδομένα, Apache Spark, Αλγόριθμοι, Γραμμική Παλινδρόμηση, Ερωτήματα, Ανίχνευση Ανωμαλιών.

ABSTRACT

The intelligent electrical grid aims to better manage the electricity grid, in view of the renewable energy sources that make production unstable, based on the analysis about a lot of data coming from sensors across the grid and from smart meters to final consumers. In this work, efficient methods are studied and tools for data analysis related to electricity consumption are implemented, so that energy providers and network operators can more efficiently match production with consumption. We will first make a brief reference to the big data and their role, while then we will refer to the Apache Spark program in terms of how it works and how it is used. Next we will talk about the algorithms we will rely on to write our programs in Python.

We implement Python code in Apache Spark three data analysis methods and we give answers to popular network administrators about power supply queries. The first method is the linear regression method, a method of statistical analysis predicting the relationship of non-dependent variables with dependent variables. The second method concerns the execution of unified queries which can give us important conclusions both statistically and research. Finally, as a third method, we will apply the detection of anomalies with the method of a non-linear class with a diagrammatic representation.

Keywords: Big Data, Apache Spark, Algorithms, Linear Regression, Queries, Outlier Detection.

ΕΥΧΑΡΙΣΤΙΕΣ

Ως την ελάχιστη δυνατή μνεία με την παρούσα παράγραφο θα ήθελα να ευχαριστήσω όλους όσους συνέβαλαν για την εκπόνηση της εργασίας και ιδιαίτερα :

Τον επιβλέποντα καθηγητή μου, κο Θανάση Παπαϊωάννου για την πολύτιμη υποστήριξή του, και για τις παραγωγικές υποδείξεις του καθώς και το πολύ ωραίο κλίμα συνεργασίας το οποίο διαμόρφωσε.

Τον καθηγητή - συντονιστή κο Βέντζα Δημήτριο κατά την διάρκεια όλης της χρονικής περιόδου των μεταπτυχιακών σπουδών.

Επίσης την συνάδελφο και συμφοιτήτρια κα Γεωργία Σπηλιοπούλου για την βοήθεια που μου προσέφερε.

Τέλος θα ήθελα να ευχαριστήσω όλους τους μεταπτυχιακούς φοιτητές για τις ανταλλαγές απόψεων, και για την σημαντική βοήθεια τους.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Ο ΝΕΟΣ ΡΟΛΟΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Διανύουμε μια νέα εποχή της πληροφόρησης με τις τηλεπικοινωνίες και το διαδίκτυο να έχουν εισέλθει για τα καλά μέσα στην ζωή μας παίζοντας μεγάλο ρόλο στην κοινωνία μας είτε μέσω της κοινωνικής δικτύωσης, η των επιχειρήσεων η της οικονομίας. Άρα είδη έχουμε αρχίσει να ζούμε σε μια εποχή όπου τα δεδομένα παίζουν μεγάλο ρόλο στην ζωή μας. Τομείς, όπως η ιατρική που απαιτεί την ανάλυση του DNA απαιτεί παραγωγή μεγάλου όγκου δεδομένων, τα διαστημικά ταξίδια που πραγματοποιούνται και οι νέες έρευνες που γίνονται μέσω αυτών, νέα εξελιγμένα μηχανήματα που χρησιμοποιούν αισθητήρες είτε στην παραγωγή υλικών – προϊόντων σε μεγάλες βιομηχανίες αλλά και σε συσκευές οποίες έχουν μπει στην καθημερινή μας ζωή.

1.2 ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΤΩΝ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ

Τα μεγάλα δεδομένα για την ενέργεια παρέχουν έναν νέο τρόπο ανάλυσης και κατανόησης της συμπεριφοράς κατανάλωσης ενέργειας των ατόμων, έτσι ώστε να βελτιωθεί η ενεργειακή απόδοση και να προωθηθεί η εξοικονόμηση ενέργειας, και η παραγωγή τους απαιτεί την ανάλυση δεδομένων μέσω διαφόρων τεχνικών αφού, αυτά μπορεί να είναι δομημένα, ημι - δομημένα, η και τελείως αδόμητα από διάφορες πηγές φτάνοντας σε μεγέθη σε terabytes και μερικές φορές έως petabytes.

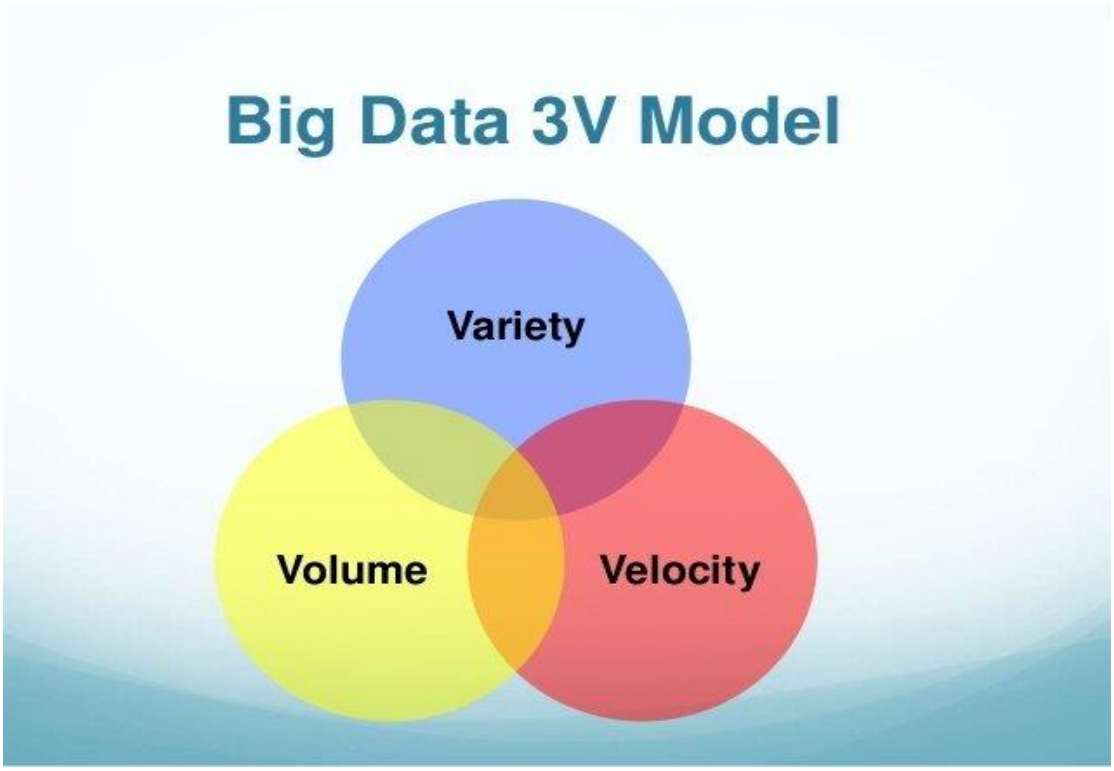
Επίσης γίνεται λόγος για μεγάλο όγκο δεδομένων και διαφορετικό τύπο, τα οποία είναι αδύνατον να επεξεργαστούν με τις παραδοσιακές βάσεις δεδομένων. Η τεχνητή νοημοσύνη, το δίκτυο των πραγμάτων οι τηλεπικοινωνίες είναι μερικές από τις οποίες παράγουν μεγάλα δεδομένα. Τέλος ακόμα ένα χαρακτηριστικό είναι ότι παράγονται σε πραγματικό χρόνο και σε μεγάλη κλίμακα.

Η ανάλυση των μεγάλων δεδομένων επιτρέπει στους αναλυτές, τους ερευνητές και τους επιχειρηματικούς χρήστες να λαμβάνουν καλύτερες και ταχύτερες αποφάσεις χρησιμοποιώντας δεδομένα που ήταν προηγουμένως απρόσιτα ή άχρηστα [2]. Οι επιχειρήσεις μπορούν να χρησιμοποιήσουν

προηγμένες τεχνικές ανάλυσης όπως αναλυτικά κείμενα, μηχανική μάθηση, προγνωστική ανάλυση, εξόρυξη δεδομένων, στατιστικές και επεξεργασία φυσικής γλώσσας για να αποκτήσουν νέες πληροφορίες από ανεξάρτητες πηγές δεδομένων ανεξάρτητα ή μαζί με τα υπάρχοντα δεδομένα επιχείρησης.

1.3 ΜΟΝΤΕΛΟ 3V

Πρόκειται για το μοντέλο το οποίο απεικονίζει τα χαρακτηριστικά των μεγάλων δεδομένων.



Εικόνα 1.1 : Απεικόνιση του μοντέλου 3V.

1.3.1 VOLUME

Εντός του χώρου των κοινωνικών μέσων, για παράδειγμα, ο τόμος αναφέρεται στην ποσότητα των δεδομένων που παράγονται μέσω ισοτόπων, πύλες και ηλεκτρονικές εφαρμογές. Ειδικά για τις εταιρείες B2C, ο Όγκος περιλαμβάνει τα διαθέσιμα δεδομένα που βρίσκονται εκεί έξω και πρέπει να αξιολογηθεί ως προς τη συνάφεια. Εξετάστε τα εξής το Facebook έχει 2 δισεκατομμύρια χρήστες, το YouTube 1 δισεκατομμύριο χρήστες, το Twitter 350 εκατομμύρια χρήστες και το Instagram 700 εκατομμύρια χρήστες. Κάθε

μέρα, αυτοί οι χρήστες συμβάλλουν σε δισεκατομμύρια εικόνες, αναρτήσεις, βίντεο, tweets κλπ. Τώρα μπορείτε να φανταστείτε την ασυνήθιστα μεγάλη ποσότητα - όγκος των δεδομένων που παράγονται κάθε λεπτό και κάθε ώρα.

1.3.2 VELOCITY

Με τη Velocity αναφερόμαστε στην ταχύτητα με την οποία παράγονται τα δεδομένα. Παραμένοντας με το παράδειγμα των κοινωνικών μέσων ενημέρωσης, 900 εκατομμύρια φωτογραφίες φορτώνονται καθημερινά στο Facebook, 500 εκατομμύρια tweets δημοσιεύονται στο Twitter, 0,4 εκατομμύρια ώρες βίντεο μεταφορτώνονται στο Youtube και 3,5 δισεκατομμύρια αναζητήσεις πραγματοποιούνται στο Google. Αυτό είναι σαν μια έκρηξη πυρηνικών δεδομένων. Τα BigData βοηθούν την εταιρεία να κρατήσει αυτή την έκρηξη, να δεχτεί την εισερχόμενη ροή δεδομένων και ταυτόχρονα να την επεξεργαστεί γρήγορα, ώστε να μην δημιουργεί σημεία συμφόρησης.

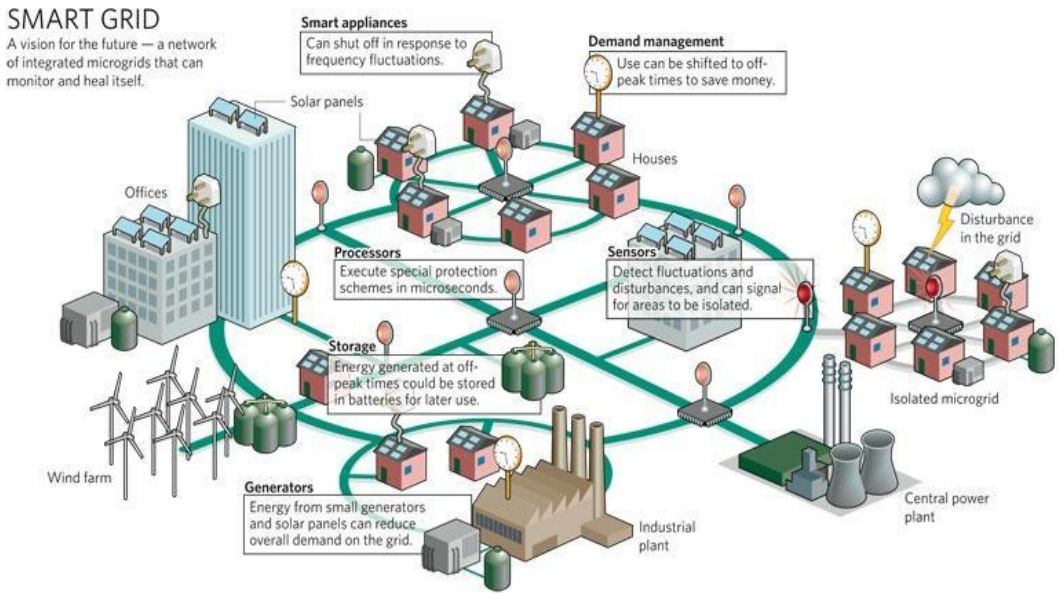
1.3.3 VARIETY

Η ποικιλία στα μεγάλα δεδομένα αναφέρεται σε όλα τα δομημένα και αδόμητα δεδομένα που έχουν τη δυνατότητα να παραχθούν είτε από ανθρώπους είτε από μηχανές. Τα πιο συχνά προστιθέμενα δεδομένα είναι δομημένα text, tweets, εικόνες & βίντεο. Εντούτοις, τα στοιχεία που δεν έχουν δομηθεί όπως τα μηνύματα ηλεκτρονικού ταχυδρομείου, τα φωνητικά μηνύματα, το χειρόγραφο κείμενο, η ανάγνωση ΗΚΓ, η ηχογράφηση κ.λπ., είναι επίσης σημαντικά στοιχεία στο Variety. Η ποικιλία αφορά τη δυνατότητα ταξινόμησης των εισερχόμενων δεδομένων σε διάφορες κατηγορίες.

1.4 ΤΟ ΕΥΦΥΕΣ ΗΛΕΚΤΡΙΚΟ ΠΛΕΓΜΑ

Όταν αναφερόμαστε στο έξυπνο δίκτυο στην ουσία επρόκειτο για ένα ηλεκτρικό δίκτυο το οποίο περιλαμβάνει και εκτελεί διάφορες λειτουργίες επί πρόσθετα με ηλεκτρικούς μετρητές οι οποίοι εμπεριέχουν, έξυπνους μετρητές, έξυπνες συσκευές, ανανεώσιμες πηγές ενέργειας και ενεργειακά αποδοτικούς πόρους [3]. Η ηλεκτρονική ρύθμιση της ισχύος και ο έλεγχος της παραγωγής και της διανομής ηλεκτρικής ενέργειας αποτελούν σημαντικές πτυχές του έξυπνου δικτύου [4]. Το έξυπνο δίκτυο χρησιμοποιεί τεχνολογία

υπολογιστών για τη βελτίωση της επικοινωνίας, της αυτοματοποίησης και της συνδεσιμότητας των διαφόρων στοιχείων του δικτύου ενέργειας, και επίσης βελτιώνει την ηλεκτρική παραγωγή ενέργειας καθώς και της ισχύος. Ένα από τα πλεονεκτήματα που χαρακτηρίζει το έξυπνο δίκτυο είναι ότι έχουν αντικατασταθεί οι παραδοσιακοί αναλογικοί μετρητές με ψηφιακές συσκευές αμφίδρομης επικοινωνίας ώστε να μεταδίδουν αλλά και να δέχονται πίσω πληροφορίες. Τα δεδομένα που συλλέγονται μέσω έξυπνων μετρητών είναι επίσης απαραίτητα για τη λειτουργία του έξυπνου δικτύου. Με την ανάλυση αυτών των δεδομένων, οι μονάδες ηλεκτροπαραγωγής είναι σε θέση να προβλέψουν καλύτερα και να ανταποκριθούν σε περιόδους μέγιστης ζήτησης. Αυτό τους επιτρέπει να μειώσουν την παραγωγή όταν απαιτείται λιγότερη ενέργεια και γρήγορα να αυξάνεται η παραγωγή όταν πλησιάζουν οι περίοδοι αιχμής. Έτσι με τη χρήση των νέων τεχνολογιών υπολογιστών και τεχνολογιών ανάλυσης δεδομένων το έξυπνο δίκτυο βελτιώνει την ευελιξία και την αποδοτικότητα του παραδοσιακού δικτύου, ώστε να αναπτύσσονται νέες τάσεις προς την παραγωγή ενέργειας όπως η αιολική και ηλιακή ενέργεια και κατασκευή νέων τεχνολογιών όπως τα ηλεκτρικά αυτοκίνητα.



Εικόνα 2: Έξυπνο δίκτυο Στο Κοντινό Μέλλον

1.5 ΈΞΥΠΝΟΙ ΜΕΤΡΗΤΕΣ

Οι έξυπνοι μετρητές είναι ηλεκτρονικές συσκευές οι οποίες καταγράφουν πληροφορίες σχετικά με την κατανάλωση της ηλεκτρικής ενέργειας, τα επίπεδα τάσης και διάφορες άλλες μετρήσεις. Οι έξυπνοι μετρητές επικοινωνούν με τον προμηθευτή ανταλλάσσοντας διάφορες πληροφορίες σχετικά με την συμπεριφορά της κατανάλωσης ενέργειας αλλά και το κόστος του πελάτη. Τυπικά θα μπορούσαμε να πούμε ότι οι έξυπνοι μετρητές καταγράφουνε ενέργεια κοντά στον πραγματικό χρόνο σε σύντομα διαστήματα μέσα στην ημέρα. Η επικοινωνία γίνεται μεταξύ του μετρητή και το κεντρικό σύστημα, ενώ η επικοινωνία με το διαδίκτυο μπορεί να γίνεται μέσω ασύρματης επικοινωνίας η μέσω καλωδίωσης με την χρήση προγραμματιζόμενων λογικών ελεγκτών [34].

1.6 ΈΞΥΠΝΟ ΣΠΙΤΙ

Αποκαλείται αλλιώς και οικιακός αυτοματισμός, οποίος χρησιμοποιεί συσκευές οι ο οποίες είναι εγκατεστημένες σε μια οικία και επικοινωνούν με το διαδίκτυο. Σε μια τέτοια εγκατάσταση χρησιμοποιούνται συνήθως συσκευές με αισθητήρες και ενεργοποιητές οι οποίες συνδέονται με το διαδίκτυο των πραγμάτων και μπορεί να γίνει απεικόνιση της λειτουργίας αυτών σε μια οθόνη έτσι ώστε να έχουμε των έλεγχο και πρόσβαση σε αυτές ανάλογα με την χρήση και της ανάγκες που μας εξυπηρετούν. Τέλος η τεχνολογία σχετικά με το έξυπνο σπίτι μας δίνει την δυνατότητα να ελέγχουμε τις συσκευές μας από απομακρυσμένο σημείο με διάφορες συσκευές η με την χρήση εφαρμογών από τα κινητά μας [35].

1.7 ΈΞΥΠΝΟ ΣΠΙΤΙ ΚΑΙ ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ

Η χρήση συσκευών αισθητήρων σε οικιακούς αυτοματισμούς και η συνεχής ροή δεδομένων εξ αυτών οδήγησαν στην δημιουργία των μεγάλων δεδομένων. Πια η χρησιμότητα όμως για αυτόν των όγκο δεδομένων ;

Μέσω του μεγάλου όγκου δεδομένων και μέσω της ανάλυσης αυτών μπορούμε να οδηγηθούμε σε σαφή συμπεράσματα σχετικά με την συμπεριφορά μιας εγκατάστασης έξυπνων συσκευών σε ένα έξυπνο σπίτι. Όπως την μέση κατανάλωση ενέργειας του σπιτιού, τις συσκευές με την μεγαλύτερη κατανάλωση και «εργατοώρες», αλλά και πως μπορούμε να παρέμβουμε στην καλύτερη δυνατή λειτουργία του έξυπνου σπιτιού από τα αποτελέσματα που λαμβάνουμε από τον μεγάλο όγκο δεδομένων ως προς την κατανάλωση ενέργειας και λειτουργικότητας. Συμπερασματικά θα μπορούσαμε να πούμε πως αν δεν είχαμε τόσο τεράστιο όγκο δεδομένων δεν θα μπορούσαμε να έχουμε μια σαφή εικόνα και ποια τα οφέλη σε σχέση με τις

νέες τεχνολογίες όπως για παράδειγμα σε αυτή την περίπτωση ο οικιακός αυτοματισμός.

1.8 ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΣΤΟ ΕΞΥΠΝΟ ΔΙΚΤΥΟ

Η υπολογιστική ανάλυση δεδομένων καθώς και η εξόρυξη τους περιλαμβάνει διάφορες τεχνικές και σχέσεις μεταξύ μεταβλητών βάσης δεδομένων, μηχανικής μάθησης, στατιστικών στοιχείων και όλα αυτά γίνονται με συλλογή δεδομένων διαφορετικής μορφής δομής και ποιότητας. Για αυτό ο όρος αποδοτικότητα δεδομένων που χρησιμοποιείται στο έξυπνο δίκτυο είναι σημαντικός.

1.9 ΑΠΟΔΟΤΙΚΟΤΗΤΑ ΔΕΔΟΜΕΝΩΝ

Η έννοια αποδοτικότητα δεδομένων αναφέρεται στην διαδικασία και στον τρόπο εύκολης χρήσης και πρόσβασης αυτών [4]. Η τοποθεσία στην οποία αποθηκεύονται τα δεδομένα έχει να κάνει με την συνολική αποτελεσματικότητα αυτών, και η αποτελεσματικότητα των δεδομένων στοχεύει στην ευκολότερη πρόσβαση στα πιο συχνά χρησιμοποιούμενα δεδομένα στο δίκτυο, τοποθετώντας τα έτσι σε συσκευές αποθήκευσης υψηλού κόστους και υψηλής ισχύος, μεταφέροντας παλαιότερα δεδομένα αρχειοθέτησης σε βραδύτερες και λιγότερο δαπανηρές εναλλακτικές λύσεις. Με αυτόν τον τρόπο, τα άτομα που εργάζονται στο δίκτυο έχουν ταχύτερη πρόσβαση σε δεδομένα μεγαλύτερης χρησιμότητας χωρίς να προκαλούν ζημία και κόστος στον προϋπολογισμό ενός οργανισμού. Άλλες τεχνικές που βελτιώνουν την αποδοτικότερη αποθήκευση δεδομένων είναι η συμπίεση δεδομένων, η οποία είναι η διαδικασία της συρρίκνωσης των αρχείων στο ελάχιστο δυνατό μέγεθος [5].

Ένα εύρος τεχνικών για αποδοτική αποθήκευση δεδομένων είναι:

- Αποσύνδεση: Η αυτόματη αφαίρεση πολλαπλών αντιγράφων του ίδιου αρχείου.
- Συμπίεση δεδομένων: Η αφαίρεση περιττών πληροφοριών μέσα σε ένα αρχείο, προκειμένου να μειωθεί το μέγεθός του.

- Λεπτή παροχή πόρων αποθήκευσης: Κατανομή μόνο εκείνων των πόρων που χρησιμοποιούνται στην πραγματικότητα, αντί να κρατάτε περισσότερο χώρο από ό, τι χρειάζεστε εκείνη την εποχή.

1.10 ΣΧΕΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΟ ΕΞΥΠΝΟ ΔΙΚΤΥΟ

Η εξάρτηση του έξυπνου δικτύου από τα δεδομένα είναι κάτι παραπάνω απλά από ένα σημείο αναφοράς, η μη σωστή ανάλυση δεδομένων καθιστά ένα έξυπνο δίκτυο ανεπαρκή. Για αυτό με τη σωστή χρήση διαφόρων εργαλείων ανάλυσης για έξυπνα δίκτυα μπορούμε να επιτυγχάνουμε την αξιοπιστία την αποδοτικότητα και σωστή ανάγνωση αλλά και την σωστή χρήση των συμπερασμάτων που παίρνουμε κάθε φορά από τα δεδομένα.

1.11 ΣΤΟΧΟΣ ΤΗΣ ΠΑΡΟΥΣΑΣ ΕΡΓΑΣΙΑΣ

Θέλουμε να δημιουργήσουμε - βελτιώσουμε ένα «εργαλείο» (tool), το οποίο στην συγκεκριμένη περίπτωσης είναι το Apache Spark, με σκοπό να διευκολύνουμε τους χρήστες ως προς την ανάλυση δεδομένων να είναι πιο αποδοτική και πιο εύχρηστη εφαρμόζοντας κάποια βασικά ερωτήματα καθώς και κάποιους μαθηματικούς αλγόριθμους προβλέποντας πιθανά σφάλματα, ενώ όλα αυτά θα έχουν την δυνατότητα να μπορούν να απεικονίζονται σε μια οθόνη (monitor) έτσι ώστε να βοηθά τον χρήστη ως προς την ανάγνωση αποτελεσμάτων αλλά και να μπορεί να οδηγηθεί πιο γρήγορα σε συμπεράσματα.

1.12 ΕΡΓΑΛΕΙΑ ΑΝΑΛΥΣΗΣ ΓΙΑ ΕΞΥΠΝΟ ΔΙΚΤΥΟ

Εκτός από το εργαλείο του Apache Spark που θα χρησιμοποιήσουμε για να κάνουμε ανάλυση σε δεδομένα, υπάρχουν επίσης και άλλες επιλογές χρήσης εργαλείων ανάλυσης δεδομένων.

Μερικά από αυτά είναι:

- Lab View: Ένα ενδιαφέρον εργαλείο (tool) για μηχανικούς αφού έχει μια πληθώρα συνδυαστικής σύνδεσης με όργανα μέτρησης αλλά και την δυνατότητα παραλληλισμού ανάκτησης δεδομένων με αυτά. Επίσης είναι αρκετά εύχρηστο για όσους δεν έχουν αρκετές γνώσεις προγραμματισμού, αλλά δίνει και την δυνατότητα να «συνεργάζεται» άρτια με γλώσσες προγραμματισμού όπως είναι η Python. Είναι ο συνδυασμός απόκτησης δεδομένων, ανάλυσης δεδομένων και παρουσίασης των αποτελεσμάτων, που μεγιστοποιεί πραγματικά τη δύναμη της Εικονικής Οργάνωσης. Η φιλοσοφία του βασίζεται πάνω στο εικονικό όργανο και στην χρήση του ηλεκτρονικού υπολογιστή. Ένα εικονικό όργανο αποτελείται από έναν βιομηχανικό τυποποιημένο υπολογιστή ή σταθμό εργασίας εξοπλισμένο με ισχυρό λογισμικό εφαρμογών, οικονομικά αποδοτικό υλικό όπως πλακέτες (plug - in) και λογισμικό προγράμματος οδήγησης, τα οποία από κοινού εκτελούν τις λειτουργίες των παραδοσιακών οργάνων. Αυτός είναι ο λόγος για τον οποίο οι εφαρμογές και τα προγράμματα που έχουν δημιουργηθεί με το Lab View αναφέρονται ως εικονικά όργανα (VIs) [36].
- Mat Lab: Ένα από τα πιο γνωστά εργαλεία ανάλυσης δεδομένων με πολλά πλεονεκτήματα για τους χρήστες, αφού προσφέρει: χιλιάδες προκαθορισμένες λειτουργίες για στατιστική ανάλυση μηχανική εκμάθηση και επεξεργασία σήματος, ταχεία απόδοση με απλές αλλαγές κώδικα και πρόσθετο υλικό, διαδραστικές και εξαιρετικά προσαρμόσιμες οπτικοποιήσεις δεδομένων και διευρυμένη ανάλυση σε μεγάλα δεδομένα χωρίς μεγάλες αλλαγές κώδικα. Επρόκειτο για μια υψηλού επιπέδου γλώσσα και διαδραστική πλατφόρμα, έχει την δυνατότητα να συμπληρώνει τις λειτουργίες του excel πάνω σε πολύπλοκα θέματα ανάλυσης δεδομένων, ενώ μέσω του εργαλείου του Simulink υπάρχει η δυνατότητα ενός περιβάλλοντος γραφικού προγραμματισμού
- Excel: Η χρήση του Excel είναι ευρέως διαδεδομένη στον κλάδο. Είναι ένα πολύ ισχυρό εργαλείο ανάλυσης δεδομένων και σχεδόν όλες οι μεγάλες και μικρές επιχειρήσεις χρησιμοποιούν το Excel στη λειτουργία

τους καθημερινά [38]. Μέσω του εργαλείου (analysis tool pack) και αφού γίνει η εγκατάσταση του μπορούμε εύκολα να κάνουμε στατιστικές αναλύσεις δεδομένων, γραφική απεικόνιση αυτών αλλά και να κάνουμε πιο εύκολο χειρισμό στα δεδομένα (manipulate data) [38].

ΚΕΦΑΛΑΙΟ 2

ΕΡΓΑΛΕΙΑ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ ΜΕΓΑΛΟΥ ΟΓΚΟΥ

2.1 ΓΕΝΙΚΑ

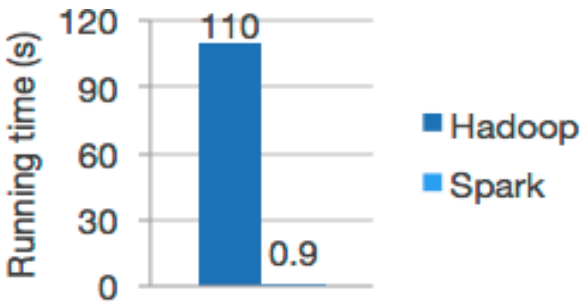
Η ανάλυση(analytics) είναι η ανακάλυψη, η ερμηνεία και η επικοινωνία σημαντικών προτύπων στα δεδομένα . Περιλαμβάνει επίσης την εφαρμογή προτύπων δεδομένων για την αποτελεσματική λήψη αποφάσεων. Με άλλα λόγια, τα αναλυτικά στοιχεία μπορούν να θεωρηθούν ως ο συνδετικός ιστός μεταξύ των δεδομένων και της αποτελεσματικής λήψης αποφάσεων μέσα σε έναν οργανισμό. Ιδιαίτερα πολύτιμη σε περιοχές πλούσιες με καταγεγραμμένες πληροφορίες, τα αναλυτικά στοιχεία βασίζονται στην ταυτόχρονη εφαρμογή στατιστικών στοιχείων, στον προγραμματισμό υπολογιστών και στην έρευνα λειτουργιών για τον ποσοτικό προσδιορισμό της απόδοσης.

Έτσι λοιπόν χρησιμοποιούμε διάφορα εργαλεία για την ανάλυση των δεδομένων όπου χρησιμοποιούνται ευρέως για την παροχή ουσιαστικής ανάλυσης ενός μεγάλου συνόλου δεδομένων. Αυτό το λογισμικό επιλέγεται σε σχέση με την εύρεση των τρεχουσών τάσεων της αγοράς, των προτιμήσεων των πελατών και άλλων πληροφοριών. Μερικά από τα πιο δημοφιλή είναι το Apache Spark το οποίο και θα χρησιμοποιήσουμε στην συγκεκριμένη εργασία, το Windows Azure, το Splice Machine, το Sky tree κλπ.

2.2 APACHE SPARK

Το Apache Spark είναι ένα λογισμικό τύπου (framework) για υπολογισμούς σε clusters υπολογιστικών συστημάτων. Η ανάπτυξη του και η έρευνα του ξεκίνησε το 2009 στο πανεπιστήμιο του UC Berkley. Οι ερευνητές του είχαν εργαστεί παλαιοτέρα για το MapReduce και παρατήρησαν ότι είναι αναποτελεσματικό για διεργασίες που απαιτούν πολλές επαναλήψεις (iterations). Έτσι η φιλοσοφία σχεδίασης του βασίζεται στα γρήγορα ερωτήματα σε επαναληπτικούς αλγορίθμους με το να αποθηκεύονται τα

δεδομένα στην μνήμη RAM και όχι στο σκληρό δίσκο αποθήκευσης έτσι ώστε να επιτυγχάνεται μεγαλύτερη ταχύτητα.



Εικόνα 3: Διαφορά στον χρόνο εκτέλεσης.

Το Apache Spark είναι γραμμένο και βασισμένο πάνω στην γλώσσα προγραμματισμού SCALA παρόλα αυτά το Spark έχει σχεδιαστεί να δίνει την δυνατότητα μέσω API όπως και στο Hadoop την δυνατότητα να εργαστούμε και με άλλες γλώσσες προγραμματισμού PYTHON, JAVA , SCALA και SQL. Τέλος προσφέρει πλήθος ενσωματωμένων βιβλιοθηκών.

```
df = spark.read.json("logs.json") df.where("age> 21")

.select("name.first").show()
```

Εικόνα 2.2 : Παράδειγμα ανάγνωσης κάποιου αρχείου(file).

Το Spark μπορεί να τρέξει σε clusters του Hadoop, σαν αυτόνομο σύστημα, πάνω στο Messos, ή το Kurbenetes. Ακόμα να έχει πρόσβαση στο HDFS, Alluxio, Apache Cassandra, Hbase , hive και σε αρκετές ακόμα βάσεις δεδομένων.

2.2.1 ΕΥΚΑΜΠΤΟ ΚΑΤΑΝΕΜΗΜΕΝΟ ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ

Ο τρόπος λειτουργίας του Spark θα μπορούσαμε να πούμε βασίζεται πάνω στο κατανεμημένο σύνολο δεδομένων (resilient distributed dataset) , δηλαδή ένα εύκαμπτο κατανεμημένο σύνολο δεδομένων το οποίο είναι μια συλλογή στοιχείων μεταξύ όλων των κόμβων ενός συμπλέγματος (cluster) έτσι ώστε να λειτουργούν παράλληλα. Η δημιουργία του (RDD) ξεκινάει από ένα αρχείο το οποίο μπορεί να βρίσκεται, να είναι αποθηκευμένο σε ένα σύστημα αρχείων

του Hadoop η απλά δίνοντας εμείς κατά την δημιουργία ενός προγράμματος. Ακόμα οι χρήστες έχουν την δυνατότητα να διατηρήσουν το (RDD) στην μνήμη σε περίπτωση που θέλουν να το ξανά χρησιμοποιήσουν σε μία παράλληλη λειτουργία. Τέλος υπάρχει η δυνατότητα ανάκτησης του (RDD) σε περίπτωση αποτυχίας των κόμβων.

Μία ακόμη λειτουργία που χαρακτηρίζει το Spark είναι η δυνατότητα να μοιράζει τις μεταβλητές σε περίπτωση παράλληλης λειτουργίας. Από προεπιλογή όταν το Spark τρέχει μία λειτουργία σε παραλληλισμό σε ένα σύνολο εργασιών σε διαφορετικούς κόμβους στέλνει ένα αντίγραφο της μεταβλητής σε κάθε έργο (task) που χρησιμοποιείται από κάποια συνάρτηση. Μερικές φορές οι μεταβλητές πρέπει να μοιράζονται μεταξύ των εργασιών (tasks) η μεταξύ των εργασιών και του προγράμματος οδήγησης. Το Spark υποστηρίζει δύο τύπους διαμερισμού μεταβλητών , στην πρώτη περίπτωση στις μεταβλητές εκπομπής (broadcast variables) οι οποίες μπορούν να χρησιμοποιηθούν για προσωρινή αποθήκευση μιας τιμής στην μνήμη σε όλους του κόμβους. Ο δεύτερος τύπος είναι οι μεταβλητές συσσωρευτών που χρησιμοποιούνται σε περιπτώσεις πρόσθεσης και μέτρησης.

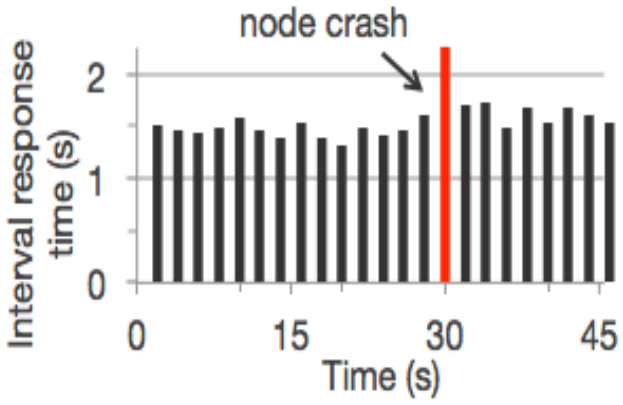
2.2.2 ΛΕΙΤΟΥΡΓΙΕΣ (RDD)

Το (RDD) υποστηρίζει δύο τύπους λειτουργιών τους λεγόμενους μετασχηματισμούς θα μπορούσαμε να πούμε, κατά την πρώτη δημιουργεί ένα νέο σετ δεδομένων από ένα είδη υπάρχων και επιστρέφει μια τιμή στο πρόγραμμα οδήγησης μετά από τους απαραίτητους υπολογισμούς, ενώ κατά την δεύτερη συγκεντρώνει όλα τα στοιχεία του (RDD) και επιστρέφει το τελικό αποτέλεσμα στο πρόγραμμα.

Οι μετασχηματισμοί στο Spark δεν υπολογίζουν αμέσως τα αποτελέσματα τους, αλλά εφαρμόζονται σε μερικά σετ δεδομένων (πχ αρχείο). Οι μετασχηματισμοί κάνουν υπολογισμούς μόνο για απαραίτητες ενέργειες που απαιτούνται και επιστρέφει τα κατάλληλα δεδομένα στο πρόγραμμα οδήγησης, έτσι χρησιμοποιείται ένα σύνολο δεδομένων και επιστρέφει το αποτέλεσμα που επιθυμείται.

2.2.3 STREAMING DATA SPARK

Αναφέρεται στην επεξεργασία ροής που είναι μια λειτουργία που ενσωματωμένη στο Apache Spark και επιτρέπει να γράφουμε εργασίες ροής σε JAVA, SCALA και PYTHON. Το Spark Streaming καθιστά εύκολη την κατασκευή επεκτάσιμων εφαρμογών συνεχούς ροής. Ενώ δίνει την δυνατότητα επαναχρησιμοποίησης του κώδικα η να εκτελεστούν ad-hoc ερωτήσεις σε κατάσταση ροής. Τέλος δίνει την δυνατότητα ανάκτησης μιας χαμένης εργασίας .



Εικόνα 4: Απεικόνιση διαγράμματος σε σχέση με αποτυχία ενός κόμβου και τον χρόνο.

2.2.4 ΒΙΒΛΙΟΘΗΚΗ SPARK MLIB

Το MLIB είναι μία κλιμακωτή βιβλιοθήκη εκμάθησης μηχανών του ApacheSpark παρέχει την δυνατότητα βιβλιοθηκών classification, regression, clustering, και collaborative filtering, για μηχανική εκμάθηση και ανάλυση δεδομένων.

2.2.5 MODULE SPARK SQL

Δίνει την δυνατότητα ανάγνωσης αρχείων τύπου μορφής JSON, η μέσω εφαρμογής Hive, καθώς και ανάκτηση δεδομένων από πηγές SQL.

2.2.6 ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΠΟΥ ΜΑΣ ΠΡΟΣΦΕΡΕΙ ΤΟ SPARK

- Streaming Data: Δυνατότητα επεξεργασίας δεδομένων και ανάλυση αυτό σε πραγματικό χρόνο, επιτρέποντας να γράφουμε εργασίες πάνω σε JAVA, SCALA και PYTHON.
- Machine Learning: Παρέχει την δυνατότητα αλγορίθμων εκμάθησης (MLIB)
- Interactive Analysis: Παρέχει την δυνατότητα της διαδραστικής ανάλυσης έτσι ώστε να είναι αρκετά γρήγορο σε σχέση με το Hadoop. Επίσης συνδέεται με πολλές γλώσσες ανάπτυξης, όπως την PYTHON, SQL και R.
- FogComputing: Ενώ το Spark είναι ίσως πιο γνωστό για τις αναλύσεις δεδομένων, ο επόμενος μεγάλος στόχος της κοινότητας είναι το Ιντερνέτ των πραγμάτων (IoT). Το IoT ενσωματώνει αντικείμενα και συσκευές με μικροσκοπικούς αισθητήρες που επικοινωνούν μεταξύ τους και με τον χρήστη, δημιουργώντας ένα πλήρως διασυνδεδεμένο κόσμο. Αυτός ο κόσμος συγκεντρώνει τεράστια ποσά δεδομένων, τα επεξεργάζεται και προσφέρει επαναστατικά νέα χαρακτηριστικά και εφαρμογές που μπορούν να χρησιμοποιήσουν οι άνθρωποι στην καθημερινότητά τους [8].

2.3 PYSPARK

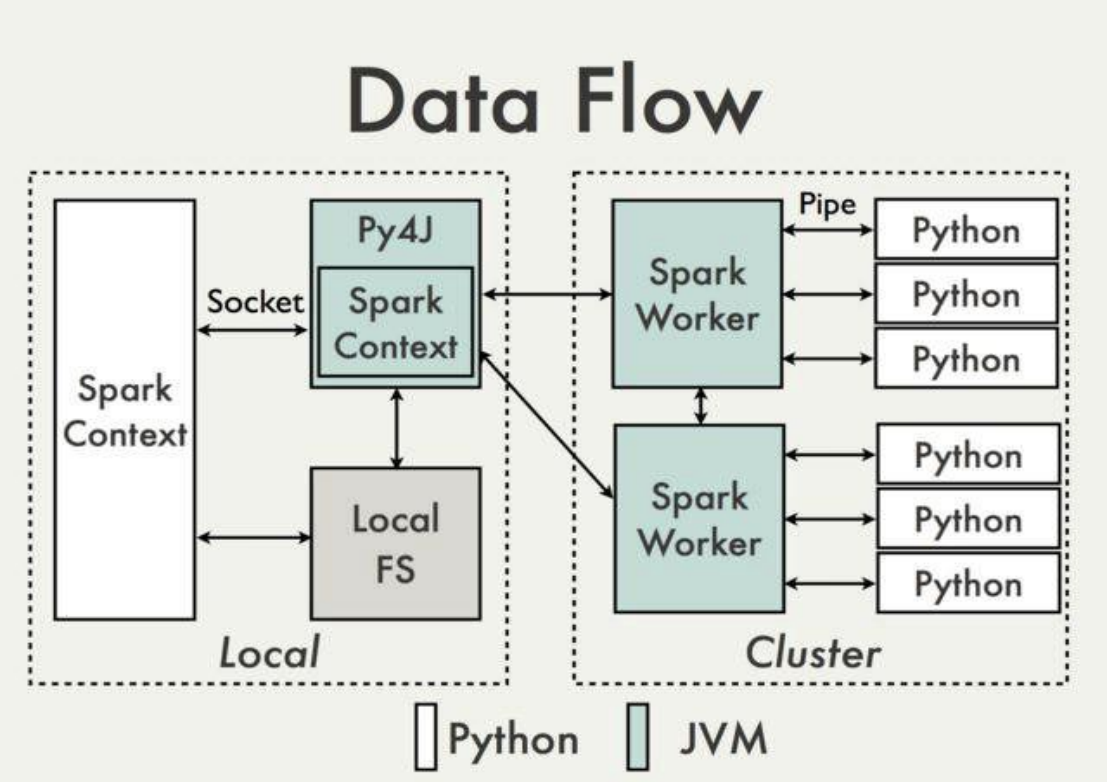
Η κοινότητα ανοιχτού κώδικα έχει αναπτύξει ένα καταπληκτικό εργαλείο για την υποστήριξη της Python για το Apache Spark [9]. Το Pyspark βοηθά τους χρήστες να αλληλεπιδρούν με RDDστο Apache Sparkκαι το Python μέσω της βιβλιοθήκης του Py4j. Υπάρχουν πολλά πλεονεκτήματα που καθιστούν το Pyspark της Apache Sparkκαλύτερο, μερικά από αυτά είναι:

- Ταχύτητα: Είναι 100 φορές πιο γρήγορα από τα παραδοσιακά μεγάλης κλίμακας πλαίσια (framework) επεξεργασίας δεδομένων.
- Powerful Caching: Το απλό επίπεδο προγραμματισμού που προσφέρεται δίνει την δυνατότητα ισχυρής προσωρινής αποθήκευσης και διατήρηση ισχύς των δίσκων.
- Deployment: Μπορεί να αναπτυχθεί και μέσω του Mesos η του Hadoop.

- Real Time: Υπολογισμός σε πραγματικό χρόνο και χαμηλή καθυστέρηση λόγω του υπολογισμού της μνήμης.

2.4 ΑΝΑΦΟΡΑ ΣΤΟ SPARKCONTEXT

Το Spark Context είναι το σημείο εισόδου σε οποιαδήποτε λειτουργία τρέχουμε ένα Apache Spark πρόγραμμα. Όταν εκτελούμε οποιαδήποτε εφαρμογή Spark, ξεκινάει ένα πρόγραμμα οδήγησης, το οποίο έχει την κύρια λειτουργία το Spark Context . Στη συνέχεια, το πρόγραμμα οδήγησης εκτελεί τις λειτουργίες εντός των κόμβων. Το Spark Context χρησιμοποιεί το Py4J για την εκκίνηση ενός JVM και δημιουργεί ένα Java SparkContext. Από προεπιλογή, το PySpark έχει το SparkContext διαθέσιμο ως 'sc', οπότε η δημιουργία ενός νέου Spark Context δεν θα λειτουργήσει.



Εικόνα 5: Data Flow

ΚΕΦΑΛΑΙΟ 3

ΑΛΓΟΡΙΘΜΟΙ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ ΣΤΟ ΕΥΦΥΕΣ ΗΛΕΚΤΡΙΚΟ ΠΛΕΓΜΑ

3.1 ΑΝΑΓΝΩΣΗ ΤΙΜΗΣ

Όπως προαναφέραμε θέλουμε να βελτιώσουμε και δείξουμε μια πιο αποδοτική και εύχρηστη μέθοδο ως προς την ανάλυση δεδομένων.

Συνήθως από μια πληθώρα δεδομένων μπορεί να μας ενδιαφέρουν κάποιες συγκεκριμένες τιμές(values) η μια ομάδα τιμών.

Κάποια από τα ενδιαφέροντα ερωτήματα(queries)μπορεί να είναι : η μέγιστη τιμή (max), ελάχιστη (min), ο μέσος όρος (average), το μέτρημα(count) η να χρησιμοποιήσουμε μεθόδους σε περιπτώσεις που θέλουμε να ορίσουμε τιμές (values), για παράδειγμα να είναι ίσες με μία τιμή α = 10 , μικρότερες ή μεγαλύτερες α > 20, b = < 12.

3.2 ΑΛΓΟΡΙΘΜΟΙ ΔΕΔΟΜΕΝΩΝ

Αλγόριθμος είναι ένα πεπερασμένο σύνολο κανόνων, οι οποίοι περιγράφουν μία μέθοδο (που αποτελείται από μία σειρά υπολογιστικών διεργασιών) για να λυθεί ένα συγκεκριμένο πρόβλημα. Τα αντικείμενα πάνω στα οποία επενεργούν αυτές οι διεργασίες λέγονται δεδομένα (data) [11].

3.2.1 ΕΥΡΕΣΗ ΤΟΥ ΜΕΓΑΛΥΤΕΡΟΥ ΚΑΙ ΤΟΥ ΜΙΚΡΟΤΕΡΟΥ ΣΤΟΙΧΕΙΟΥ

Έστω ένας πίνακας Sn που τα στοιχεία του είναι ακέραιοι αριθμοί μη ταξινομημένοι. Ζητάμε να βρούμε το μεγαλύτερο και το μικρότερο στοιχείο του πίνακα [12] [13].

$$T(n) = \begin{cases} 0 & ,για\ n = 1 \\ 1 & ,για\ n = 2 \\ 2T\left(\frac{n}{2}\right) + 2, & για\ n > 2 \end{cases}$$

Για $n > 2$έχουμε:$T = 2T\left(\frac{n}{2}\right) + 2 =$

$$= 4T\left(\frac{n}{2}\right) + 4 + 2 =$$

$$= 4T\left(\frac{n}{2}\right) + 8 + 4 + 2 =$$

$$= \cdots =$$

$$= 2^k T\left(\frac{n}{2^k}\right) + 2\sum_{i=0}^{k-1} 2^i =$$

$$= 2^k T\left(\frac{n}{2^k}\right) + \ 2\frac{2^k - 1}{2 - 1} =$$

$$= 2^k T\left(\frac{n}{2^k}\right) + \ 2^{k+1} - 2$$

Όταν ο πίνακας έχει $n = \ 2^{k+1}$στοιχεία $\left(\frac{n}{2}\right) = 2^k$τότε έχουμε:

$$T(n) = \frac{n}{2}T(2) + 2^{k+1} - 2 = \frac{3}{2}n - 2$$

3.2.2 ΤΑΞΙΝΟΜΗΣΗ

Η ταξινόμηση (sorting) τοποθετεί ένα σύνολο κόμβων ή εγγραφών σε μία ιδιαίτερη σειρά (αύξουσα ή φθίνουσα) με βάση την τιμή (πρωτεύοντος) κλειδιού της εγγραφής. Στην ουσία όταν αναφερόμαστε στην έννοια της ταξινόμησης επρόκειτο για την αναζήτηση των στοιχείων του αντίστοιχου συνόλου [19] [20].

Έστω τα στοιχεία $\alpha_1, \alpha_2 \ldots . \alpha_n$η ταξινόμηση συνιστάται στη διάταξη (permutation) της θέσης των στοιχείων, ώστε να τοποθετηθούν σε μία σειρά $\alpha_{k_1}, \alpha_{k_2} \ldots . \alpha_{\kappa_n}$έτσι ώστε να πάρουμε μια συνάρτηση διάταξης (ordering function), να ισχύει:

$$f(a_{k_1}) \leq f(a_{k_2}) \leq \cdots \leq f(a_{k_n})$$

Η αναζήτηση ενός κλειδιού σε μη ταξινομημένο πίνακα με n εγγραφές γίνεται σειριακά με συγκρίσεις της τάξης O(n), ενώ σε ταξινομημένο πίνακα η δυαδική αναζήτηση απαιτεί κόστος της τάξης O(logn).

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

Πίνακας μεγέθους n στοιχείων είναι $\Theta(n)$,$\Theta(n^2)$,και $\Theta(n^2)$, σε κάθε περίπτωση.

Ψευδό κώδικας εύρεσης ελαχίστου

```
procedure select

for i <-- 1 to n - 1 do

minj <-- i; minx <-- A[i];

    for j <-- i + 1 to n do
      if A[j] < minx then
   minj <-- j
      minx <-- A[j]
A[minj] <-- A[i]
A[i] <-- minx
```

Παράδειγμα 3.3: Ταξινόμηση με επιλογή από το μικρότερο στοιχείο.

3.2.3 ΜΕΣΗ ΤΙΜΗ

Ο αλγόριθμός αναζητά το μικρότερο μεταξύ των $n-1+1$στοιχείων του πίνακα στα δεξιά της θέσης.

$$R(m) = S(m)/m!$$

Όπου R(m) συμβολίζουμε το πλήθος των καταχωρήσεων που απαιτούνται για την επιλογή του ελάχιστου μεταξύ των m στοιχείων. Ενώ S(m) είναι το πλήθος των καταχωρήσεων που αθροίζονται κατά την εκτέλεση των m! διαφορετικών διατάξεων των m στοιχείων.

Έστω (m-1) περιπτώσεις όπου το ελάχιστο στοιχείο εντοπίζεται στην τελευταία θέση του πίνακα. Στην περίπτωση αυτή το πλήθος των καταχωρήσεων είναι

$$S(m-1)+(m-1)$$

Καθώς ο αλγόριθμος συμπεριφέρεται σαν να υπήρχαν αρχικά (m-1) μη επαναλαμβανόμενα στοιχεία, οπότε συνέχεια θα εκτελούνταν άλλη μία καταχώρηση για κάθε μία από τις (m-1)! διαφορετικές διατάξεις. Τότε απομένουν (m-1)(m-1)! Περιπτώσεις όπου το ελάχιστο στοιχείο βρίσκεται από την πρώτη μέχρι την (m-1) – οστή θέση, οπότε για κάθε ομάδα(m-1)! διατάξεων απαιτούνται S(m-1) καταχωρήσεις. Έτσι συνολικά παίρνουμε:

$(m-1)S(m-1)$καταχωρήσεις. Και προκύπτει η αναδρομική εξίσωση:

$$S(m)=mS(m-1)+(m-1)$$

Με αρχική συνθήκη S(1) = 0 . Άρα γα την μέση τιμή του R(m) ισχύει:

$$R(m)=R(m-1)+\frac{1}{m} \qquad m>1$$

Και παίρνουμε την λύση:

$$R(m)=\sum_{k=2}^{m}\frac{1}{k}=Hm-1$$

Έτσι προκύπτει ότι η μέση τιμή του πλήθους των καταχωρήσεων λαμβάνοντας n – 1 το σύνολο των επαναλήψεων είναι :

$$T(n)=\sum_{i=1}^{n-1}(H_{n-i+1}-1)=\sum_{i=2}^{n}(H_i-1)$$

$$=\left(\frac{1}{2}\right)+\left(\frac{1}{2}+\frac{1}{3}\right)+\cdots+\left(\frac{1}{2}+\frac{1}{2}+\cdots+\frac{1}{n}\right)$$

$$=\sum_{k=2}^{n}(n-k+1)\frac{1}{k}=\cdots=(n+1)H_n-2_n$$

Άρα προκύπτει ότι στη μέση περίπτωση το πλήθος των καταχωρήσεων είναι της τάξης $\Theta(n \log n)$.

3.2.4 ΔΥΑΔΙΚΗΑΝΑΖΗΤΗΣΗ

Η δυαδική αναζήτηση λειτουργεί σε ταξινομημένες συστοιχίες. Η δυαδική αναζήτηση ξεκινά με τη σύγκριση ενός στοιχείου στη μέση του πίνακα με την τιμή στόχου. Εάν η τιμή στόχος ταιριάζει με το στοιχείο, η θέση του στον πίνακα επιστρέφεται. Εάν η τιμή στόχος είναι μικρότερη από το στοιχείο, η αναζήτηση συνεχίζεται στο κάτω μισό του πίνακα. Εάν η τιμή-στόχος είναι μεγαλύτερη από το στοιχείο, η αναζήτηση συνεχίζεται στο άνω μισό του πίνακα. Κάνοντας αυτό, ο αλγόριθμος εξαλείφει το μισό στο οποίο η τιμή στόχος δεν μπορεί να βρεθεί σε κάθε επανάληψη [16].

Δεδομένου ενός πίνακα A με nστοιχεία με τιμές $A_0, A_1, A_2, \ldots\ldots, A_{n-1}$ταξινομημένα έτσι $A_0 \leq A_1 \leq A_2 \leq \ \ldots \leq A_{n-1}$.και την τιμή στόχου T, η ακόλουθη υπό ρουτίνα χρησιμοποιεί δυαδική αναζήτηση για να βρει το ευρετήριο το T σε A.[16]

Σειρά L προς το 0 και R προς το n-1

Αν$L > R$,η αναζήτηση τερματίζεται ως ανεπιτυχείς

Σειρά m«θέση του μεσαίου στοιχείου» $\frac{L+R}{2}$, ο οποίος είναι ο μεγαλύτερος ή ίσος με $\frac{L+R}{2}$

Αν $A_m < T$, Lπρος$m + 1$.

Αν $A_m > T$R προς $m - 1$.

$A_m = T$, επιστροφή m.

ηλειτουργία binary_search (A, n, T) **είναι**
L: = 0
R: = η − 1
ενώ L ≤ R **κάνουμε**
m: = ((L + R) / 2)

εάν A [m] < T **τότε**
L: = m + 1
αλλιώς εάν A [m] > T **τότε**
R: = m − 1
else :
επιστροφή m
επιστροφή μηεπιτυχής

παράδειγμα 3.2 : Διαδικασία ψευδοκώδικα.

3.2.5 ΑΛΓΟΡΙΘΜΟΣ ΕΠΙΛΟΓΗΣ

Έστω ότι επιθυμούμε να βρούμε το κ- οστό στοιχείο ενός πίνακα ενός αταξινόμητου πίνακα. Αυτό μπορεί να επιτευχθεί τον πίνακα και λαμβάνοντας το περιεχόμενο της αντίστοιχης θέσης του πίνακα. Παρακάτω βλέπουμε τον αλγόριθμο που σχεδιάστηκε από τον C.A.R. Hoare [17].

Πρέπει : $left \leq k \leq right$

procedure find(left, right, k);
1. if left = right then return A[left]
2. else
3. i <−− left; j <−− right + 1; pivot <−− A[left];
4. repeat
5. repeat i <−− i + 1 until A[i] >= $pivot$;
6. repeat j <−− j − 1 until A[j] <= $pivot$;
7. if i < j $THEN$ $swap(A[i], A[j])$;
8. until j <= i;
9. swap(A[left], A[j]);
10. if k = j then return A[j]
11. else if (k < j) $THEN$ $find(left, j-1, k)$
12. else find(j + 1, right, k − j);

$$T(n) = n + \sum_{k=0}^{n-1} \frac{1}{n} T(k)$$

Όπου *T(0) = Tα(1)=0.*

$$T(n) = n + \frac{1}{n} \sum_{k=2}^{n-1} T(k)$$

Πολλαπλασιάζουμε επί nκαι στην συνέχεια στην ίδια σχέση αντικαθιστούμε με n – 1 .

$$(n-1)T(n-1) = (n-1)^2 + \sum_{k=2}^{n-2} T(k)$$

$nT(n) - (n-1)T(n-1) = 2n-1 + T(n-1)$όπου θα πάρουμε:

$$T(2) - T(1) = 2 - \frac{1}{2}$$

Αθροίζοντας αντίστοιχα

$$T(n) = 2(n-1) - \left(\frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right)$$

$$= 2(n-1) + 1 - \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right)$$

$$= 2n - 1 - H_n$$

Παρακάτω βλέπουμε τον αλγόριθμο select ο οποίος υπερτερεί του findδιότι το πρόβλημα ήταν ότι ο διαχωρισμός της αρχικής ακολουθίας δεν ήταν πάντα σωστός και μπορούσε να οδηγήσει σε μεγάλα μεγέθη υποπροβλημάτων.

Ο αλγόριθμος selectαναπτύχθηκε από τους Blumet. Al [18]. Στην ουσία find και select είναι ο ίδιος αλγόριθμος.

Select(M; i)
(∗ Εύρεση του i − οστο μεγαλύτερου στοιχείου του M ∗)
1. n ← |M|;
2. **if** n ≤ 100 **then**
3. Ταξινόμηση του M και απευθείας εύρεση του i οστούμεγαλύτερου;
4. **else**
5. Διαχωρισμός του M σε υποσύνολαM1; M2; ::
: ; $M_{[\frac{n}{5}]}$5 στοιχείωντοκαθένα
τοMενδέχεταιναπεριέχει ≤ 5 στοιχεία);
6. Ταξινόμησηm$_j$, $1 \le j \le [\frac{n}{5}]$
7. Απευθείας εύρεση του μέσου mj σε καθένα Mj ;
8. $\bar{m} \leftarrow \text{Select}\left(\left(m1; m2; \cdots ; m_{[\frac{n}{5}]}\right), \left[\frac{n}{10}\right]\right)$;

(Εύρεση του μέσου των μέσων, έστω m)
9. M1 {m ∈ M; m̄ < sg; } (Ταμικρότερα του m στοιχεία)
10. M2 {m∈M; m̄ > sg; } (Ταμεγαλύτερα του m στοιχεία)
11. **if** i_ ≤ |M1| **then**
12. Select(M1; i);
13. **else**
14. Select(M_2i − |M_1|);
15. **fi**
16. **fi**

3.3 ΥΠΟΛΟΓΙΣΜΟΣ ΕΚΤΙΜΩΜΕΝΟΥ ΦΟΡΤΟΥ ΕΝΕΡΓΕΙΑΣ (BASELINECALCULATION)

Για να συσχετίσουμε τα δεδομένα μεταξύ τους κατά την εκχώρηση κάποιων από αυτών θα πρέπει τα σχετικά δεδομένα να ταιριάζουν με μερικά εξ αυτών και τέλος να καθορίζεται η καλύτερη δυνατή συνάρτηση αυτού και να μοντελοποιεί τα δεδομένα .

Αυτό γίνεται μέσου του μοντέλου της παλινδρόμησης και του τύπου:

$$y = X\beta + \varepsilon$$

Όπου θα να αναλύσουμε και πιο λεπτομερή παρακάτω.

3.3.1 ΈΝΝΟΙΑ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ (REGRESSION)

Όταν αναφερόμαστε στην στατιστική με την έννοια της παλινδρόμησης, εννοούμε την τεχνική ενός μοντέλου με σκοπό την συσχέτιση μιας εξαρτώμενης μεταβλητής η ανεξάρτητων μεταβλητών η περισσοτέρων αυτών, σε μια πραγματική μεταβλητή πρόβλεψης κατά την εκχώρηση δεδομένων. Η παλινδρόμηση προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συνάρτησης και μετά καθορίζει την καλύτερη συνάρτηση αυτού του είδους που μοντελοποιεί τα δεδομένα που έχουν δοθεί (Κεχαγιά – Παρδαλή Ευθαλία 2006). Τέλος με μια απλή πρόταση θα μπορούσαμε να δώσουμε την εξής έννοια, ότι στην παλινδρόμηση γίνεται πρόβλεψη και κατηγοριοποίηση των νέων δεδομένων έτσι ώστε σε κάποιας εφαρμογές να μπορούμε να έχουμε την δυνατότητα πρόβλεψης, όπως για παράδειγμα ενός νέου προϊόντος ή μιας υπηρεσίας.

3.3.2 ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Όπως προαναφέραμε η αρχή της παλινδρόμησης είναι ένα μοντέλο για την έρευνα της συσχέτισης μεταξύ μιας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών.

Μια συνάρτηση παλινδρόμησης $\Upsilon \cong F(X, \beta)$ όπου το Υ συσχετίζεται με την συνάρτηση, όπου: το β (διάνυσμα) είναι οι άγνωστες παράμετροι συσχέτισης, το Χ (διάνυσμα) αναφέρεται στις ανεξάρτητες μεταβλητές , ενώ το Υ η εξαρτώμενη μεταβλητή. Έτσι η εξαρτώμενη μεταβλητή Υ μεταβάλλεται όταν μία από τις εξαρτώμενες μεταβλητές Χ μεταβάλλεται [32].

3.3.3 ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Η παλινδρόμηση είναι μια τεχνική που χρησιμοποιείται για την μοντελοποίηση και την ανάλυση αριθμητικών δεδομένων, μιας εξαρτημένης μεταβλητής και κάποιων ανεξάρτητων μεταβλητών. Το μοντέλο είναι μια συνάρτηση συσχέτισης της εξαρτημένης μεταβλητής από τις ανεξάρτητες. Η μοντελοποίηση μπορεί να γίνει χωρίς να είναι γνωστή από πριν και η γνώση για τον τρόπο με τον οποίο συνδέεται η εξαρτημένη μεταβλητή από τις ανεξάρτητες και τότε ονομάζεται εμπειρική μοντελοποίηση. Στην γραμμική παλινδρόμηση, η απαίτηση του μοντέλου που θα παραχθεί είναι: η εξαρτημένη μεταβλητή yi να είναι ένας γραμμικός συνδυασμός των ανεξαρτήτων μεταβλητών.

Ένα σύνολο δεδομένων {yi , xil , ……., xip} των n στατιστικών μονάδων σε ένα μοντέλο γραμμικής παλινδρόμησης υποθέτει ότι η σχέση μεταξύ της εξαρτημένης μεταβλητής yi και το p-φορέα της παλινδρόμησης xi είναι γραμμική. Η σχέση αυτή διαμορφώνεται μέσα από έναν όρο διαταραχής ή λάθος μεταβλητή ei μία απαρατήρητη τυχαία μεταβλητή που προσθέτει με τον όρο «θόρυβο» με τη γραμμική σχέση ανάμεσα στην εξαρτημένη μεταβλητή και παλινδρόμηση. Έτσι έχουμε την μορφή :

$y_i = \beta_{ixi} + \ldots\ldots + \beta_{p\,xip} + \varepsilon_i = x_i^T \beta + \varepsilon_i$.

Όπου Τ συμβολίζεται ο ανάστροφος, άρα $x_i^T \beta$ είναι το εσωτερικό γινόμενο μεταξύ των διανυσμάτων x_i και β. Έτσι η τελική μορφή είναι $y = X\beta + \varepsilon$.

Στη στατιστική, η γραμμική παλινδρόμηση είναι μια προσέγγιση για τη μοντελοποίηση της σχέσης μεταξύ μιας βαθμωτής εξαρτημένης μεταβλητής Y και μία ή περισσότερες επεξηγηματικές μεταβλητές (ή ανεξάρτητη μεταβλητή) συμβολίζεται X. Περίπτωση μιας επεξηγηματικής μεταβλητής ονομάζεται απλή γραμμική παλινδρόμηση. Για περισσότερες από μία επεξηγηματικές μεταβλητές, η διαδικασία ονομάζεται πολλαπλή γραμμική παλινδρόμηση, ο όρος αυτός θα πρέπει να διακρίνεται από πολυμεταβλητή γραμμική παλινδρόμηση, όπου πολλαπλά προβλέπουν συσχέτιση με εξαρτημένες μεταβλητές αντί για μία ενιαία βαθμωτή μεταβλητή) [33].

Στην γραμμική παλινδρόμηση, τα δεδομένα μοντελοποιούνται χρησιμοποιώντας γραμμικές λειτουργίες προγνωστικά, και οι άγνωστες παράμετροι μοντέλου υπολογίζονται από τα δεδομένα. Τέτοια μοντέλα καλούνται γραμμικά μοντέλα. Συνηθέστερα, η γραμμική παλινδρόμηση αναφέρεται σε ένα μοντέλο στο οποίο ο υποθετικός μέσος όρος του Y δεδομένης της αξίας του X είναι μια συνάρτηση αφινικών X λιγότερο συχνά, όπου η γραμμική παλινδρόμηση μπορεί να αναφέρεται σε ένα μοντέλο στο οποίο η διάμεσος, ή κάποιο άλλο ποσοστημόριο της υποθετικής διανομής y που δίνεται X εκφράζεται ως γραμμική συνάρτηση του X. Όπως όλες τις μορφές ανάλυσης παλινδρόμησης, η γραμμική παλινδρόμηση επικεντρώνεται στους όρους κατανομής πιθανότητας του y που δίνονται X αντί για την από κοινού πιθανότητα διανομής του Y και X, η οποία είναι η και η περιοχή της πολυμεταβλητής ανάλυσης.

Η γραμμική παλινδρόμηση ήταν ο πρώτος τύπος της ανάλυσης παλινδρόμησης που μελετήθηκε αυστηρά, και προορίζεται να χρησιμοποιηθεί εκτενώς σε πρακτικές εφαρμογές. Αυτό συμβαίνει επειδή τα μοντέλα που εξαρτώνται γραμμικά από άγνωστες παραμέτρους τους είναι πιο εύκολο να μελετηθούν από τα μοντέλα τα οποία είναι μη-γραμμικά με παραμέτρους τους , και επειδή οι στατιστικές ιδιότητες των προκυπτόντων εκτιμήσεων είναι εύκολο να προσδιοριστούν.

Αν ο στόχος είναι η πρόβλεψη, ή η μείωση, η γραμμική παλινδρόμηση μπορεί να χρησιμοποιηθεί για να χωρέσει ένα προγνωστικό μοντέλο σε ένα παρατηρούμενο δεδομένο με X και Y τιμές. Μετά από την ανάπτυξη ενός

τέτοιου μοντέλου, μια πρόσθετη τιμή του X είναι τότε χωρίς την συνοδευτική αξία του y, όπου το μοντέλο μπορεί να χρησιμοποιηθεί για να κάνει μια πρόβλεψη της τιμής του y.

Δεδομένης μια μεταβλητή y και ενός αριθμού μεταβλητών X1, ..., Xp που μπορεί να σχετίζονται με το y, η ανάλυση γραμμικής παλινδρόμησης μπορεί να εφαρμοστεί στην ποσοτικοποίηση της αντοχής της σχέσης μεταξύ Y και του χj , προκειμένου να αξιολογηθεί η οποία σχέση χj με y , και να προσδιορίσει ποιες υποκατηγορίες του χj περιέχουν περιττές πληροφορίες σχετικά με το y.

Τα μοντέλα γραμμικής παλινδρόμησης συχνά χρησιμοποιούνται κατά την προσέγγιση ελαχίστων τετραγώνων. Αντιστρόφως, η προσέγγιση με τα τετράγωνα μπορεί να χρησιμοποιηθεί για να χωρέσει τα μοντέλα που δεν είναι γραμμικά μοντέλα.

Πρότυπα μοντέλα γραμμικής παλινδρόμησης με τυποποιημένες τεχνικές εκτίμησης κάνουν μια σειρά από υποθέσεις σχετικά με τις μεταβλητές πρόβλεψης, τις μεταβλητές απόκρισης και τη σχέση τους. Πολλές επεκτάσεις έχουν αναπτυχθεί που επιτρέπουν κάθε μία από αυτές τις υποθέσεις να χαλαρώσουν (δηλαδή μειώνεται σε μια ηπιότερη μορφή), και σε ορισμένες περιπτώσεις θα εξαλειφθούν εντελώς. Μερικές μέθοδοι είναι αρκετά γενικές ώστε να μπορούν να χαλαρώσουν πολλαπλές υποθέσεις ταυτόχρονα, και σε άλλες περιπτώσεις, αυτό μπορεί να επιτευχθεί με συνδυασμό διαφορετικών επεκτάσεων. Γενικά αυτές οι επεκτάσεις καταστούν την διαδικασία εκτίμησης πιο σύνθετη και χρονοβόρα, και μπορεί επίσης να απαιτεί περισσότερα δεδομένα για να παράγει ένα εξίσου ακριβό μοντέλο.

Τα ακόλουθα είναι οι κύριες παραδοχές που έγιναν από τα συνήθη μοντέλα γραμμικής παλινδρόμησης με τυποποιημένες τεχνικές εκτίμησης (π.χ. η διαδικασία ελαχίστων τετραγώνων):

- Ασθενής εξωγενείς: Αυτό ουσιαστικά σημαίνει ότι μπορούν να αντιμετωπιστούν οι μεταβλητές πρόβλεψης X ως σταθερές τιμές, παρά τυχαία μεταβλητή s. Αυτό σημαίνει, για παράδειγμα, ότι οι μεταβλητές πρόβλεψης θεωρείται ότι είναι χωρίς λάθη με σφάλματα μέτρησης. Παρά το γεγονός ότι η υπόθεση αυτή δεν είναι ρεαλιστική σε πολλές

ρυθμίσεις, οδηγεί σε πολύ πιο δύσκολο μοντέλο για τα σφάλματα των μεταβλητών s.

- Γραμμικότητα: Αυτό σημαίνει ότι η μέση τιμή της μεταβλητής απόκρισης είναι ένας γραμμικός συνδυασμός των παραμέτρων (μεταβλητές πρόβλεψης). Σημειώστε ότι αυτή η υπόθεση είναι πολύ λιγότερο περιοριστική από, τι μπορεί αρχικά να φαίνεται. Επειδή οι μεταβλητές πρόβλεψης αντιμετωπίζονται ως σταθερές τιμές , κατά την γραμμικότητα είναι πραγματικά μόνο ένας περιορισμός σχετικά με τις παραμέτρους. Οι ίδιες οι μεταβλητές πρόβλεψης μπορεί να μετασχηματιστούν αυθαίρετα, και στην πραγματικότητα μπορούν να προστεθούν πολλαπλά αντίγραφα της ίδιας υποκείμενης μεταβλητής πρόβλεψης, το καθένα μετασχηματισμένο με διαφορετικό τρόπο. Αυτό το τέχνασμα έχει χρησιμοποιηθεί, για παράδειγμα, στη πολυωνυμική παλινδρόμηση, το οποίο χρησιμοποιεί γραμμική παλινδρόμηση για να χωρέσει τη μεταβλητή απόκριση ως μια αυθαίρετη λειτουργία μιας μεταβλητής πρόβλεψης. Αυτό κάνει την γραμμική παλινδρόμηση μία εξαιρετικά ισχυρή μέθοδο εξαγωγής συμπερασμάτων. Στην πραγματικότητα, τα μοντέλα όπως η πολυωνυμική παλινδρόμηση είναι συχνά αποτελεσματική. Ως αποτέλεσμα, κάποια είδη νομιμοποίησης θα πρέπει να χρησιμοποιούνται τυπικά για να αποφεύγονται οι υπερβολικές λύσεις που προέρχονται από τη διαδικασία εκτίμησης. Κοινά παραδείγματα είναι παλινδρόμηση κορυφογραμμής και λάσο παλινδρόμησης (Bayesian) γραμμικής παλινδρόμησης, που μπορεί επίσης να χρησιμοποιηθεί, η οποία από τη φύση της είναι περισσότερο ή λιγότερο προσαρμοστική στο πρόβλημα της υπέρ προσαρμογής. (Στην πραγματικότητα, κορυφογραμμή παλινδρόμησης και λάσο παλινδρόμησης μπορεί τόσο να θεωρηθούν ως ειδικές περιπτώσεις Bayesian γραμμικής παλινδρόμησης, με συγκεκριμένους τύπους πριν από τη διανομή s που τοποθετούνται με τους συντελεστές παλινδρόμησης).
- Συνεχής Μεταβλητότητα: Αυτό σημαίνει πως οι διαφορετικές μεταβλητές απόκρισης έχουν την ίδια μεταβλητότητα στα λάθη τους, ανεξάρτητα από τις τιμές των προβλεπόμενων μεταβλητών. Πρακτικά,

αυτή η υπόθεση δεν ισχύει, αν οι μεταβλητές απόκρισης μπορούν να ποικίλλουν σε μια ευρεία κλίμακα. Με σκοπό να αποφασίσουμε για τα ετερογενή λάθη μεταβλητότητας, ή πότε μια διάταξη υπολοίπων παραβιάζει τη δομή των υποθέσεων (το λάθος είναι εξ' ίσου μεταβλητό γύρω από την «καλύτερη - ταιριαστή γραμμή» για όλα τα σημεία του x), είναι συνετό να ψάξουμε για ένα αποτέλεσμα μεταξύ του υπόλοιπου λάθους και των προβλεπόμενων τιμών. Πρέπει να πούμε ότι θα υπάρχει μια συστηματική αλλαγή στο απόλυτο ή στο τετράγωνο των υπολοίπων όταν ενεργούν σε βάρος του προβλεπόμενου αποτελέσματος. Το λάθος δε θα είναι ομοιόμορφα κατανεμημένο στη γραμμή παλινδρόμησης .Η ετεροσκεδαστικότητα θα συνεισφέρει στο να ανεβάσει το μέσο όρο των διακριτών μεταβλητών γύρω από τα σημεία για να μια μοναδική μεταβλητή που ανακριβώς αναπαριστά όλες τις μεταβλητές της γραμμής. Συνεπώς, τα υπόλοιπα εμφανίζονται συσσωρευμένα και απλωμένα στις προβλεπόμενες παραστάσεις τους για μεγαλύτερες ή μικρότερες τιμές για σημεία πάνω στη γραμμή παλινδρόμησης και το νόημα της παράστασης θα είναι λάθος. Τυπικά μια μεταβλητή απόκρισης που η σύγκλισή της είναι μεγάλη θα έχει μεγαλύτερη διασπορά από μια που η σύγκλισή της είναι μικρή. Για παράδειγμα, ένας συγκεκριμένος άνθρωπος που το εισόδημά του προβλέπεται (μια τυπική απόκλιση), ενώ παράλληλα όταν ένας άλλος άνθρωπος με προβλεπόμενο εισόδημα είναι απίθανο να έχει την ίδια τυπική απόκλιση ,η οποία σημαίνει ότι το πραγματικό τους εισόδημα θα ποίκιλλε οπουδήποτε μεταξύ αυτών. Οι απλές μέθοδοι γραμμικής παλινδρόμησης και εστίασης δίνουν λιγότερη ακρίβεια στην εκτίμηση των παραμέτρων και χάνουν σημαντικές ποσότητες όπως σίγουρα λάθη όταν ουσιαστικά η ετεροσκεδαστικότητα υπάρχει. Ωστόσο, πολλές τεχνικές εστίασης (π.χ. η proportional στάθμιση ελαχίστων τετραγώνων και η ετεροσκεδαστικότητα με συνεχή σφάλματα μπορούν να χειριστούν την ετεροσκεδαστικότητα με έναν απλό γενικό τρόπο. Οι τεχνικές της Μπευζιανής γραμμικής παλινδρόμησης μπορούν ακόμα να χρησιμοποιηθούν όταν η διακύμανση φέρεται να είναι συνάρτηση της τιμής. Είναι ακόμα πιθανό σε μερικές περιπτώσεις να διορθώσεις ένα πρόβλημα με το να εφαρμόζεις μια αλλαγή στην διακριτή μεταβλητή

(π.χ. να ταιριάξεις ένα λογάριθμο της διακριτής μεταβλητής χρησιμοποιώντας μια μέθοδο γραμμικής παλινδρόμησης, η οποία σημαίνει ότι η διακριτή μεταβλητή έχει μια ευρεία κανονική κατανομή από ότι μια απλή κανονική κατανομή.

- Ανεξαρτησία λαθών: Αυτό προϋποθέτει ότι τα σφάλματα των μεταβλητών απόκρισης είναι ασύνδετα μεταξύ τους. (Πραγματική στατιστική ανεξαρτησία). Είναι μια ισχυρότερη κατάσταση από την απλή έλλειψη συσχέτισης και συχνά δεν είναι απαραίτητη, αν και μπορεί να αξιοποιηθεί, εάν είναι γνωστό για να κρατήσει. Ορισμένες μέθοδοι είναι σε θέση να χειρίζονται σφάλματα, αν και συνήθως απαιτούν πολύ περισσότερα δεδομένα, εκτός αν κάποιο είδος νομιμοποίησης χρησιμοποιείται για την πόλωση του μοντέλου για την παραδοχή λαθών. Η μπεϋζιανή γραμμική παλινδρόμηση είναι ένας γενικός τρόπος χειρισμού αυτού του θέματος.
- Έλλειψη πολυσυγγραμμικότητας στους προγνωστικούς παράγοντες: Για πρότυπο ελαχίστων τετραγώνων μέθοδοι εκτίμησης, η μήτρα σχεδιασμού X πρέπει να έχει πλήρη βαθμίδα στήλης p . Αλλιώς, έχουμε μια κατάσταση που είναι γνωστή ως πολυσυγγραμμικότητα στις μεταβλητές πρόβλεψης. Αυτό μπορεί να προκληθεί από την κατοχή δύο ή περισσότερων τέλεια συσχετισμένων μεταβλητών πρόβλεψης (π.χ. αν η ίδια μεταβλητή πρόβλεψης λανθασμένα δίνεται δύο φορές, είτε χωρίς μετατροπή ενός από τα αντίγραφα ή με τη μετατροπή ενός από τα αντίγραφα γραμμικά). Μπορεί επίσης να συμβεί εάν υπάρχουν πάρα πολύ λίγα διαθέσιμα δεδομένα σε σχέση με τον αριθμό των παραμέτρων που πρέπει να υπολογίζεται (π.χ. λιγότερα σημεία δεδομένων από συντελεστές παλινδρόμησης). Στην περίπτωση πολυσυγγραμμικότητας, το διάνυσμα παραμέτρων β θα είναι μη αναγνωρίσιμο και δεν έχει μοναδική λύση. Στην καλύτερη περίπτωση θα είναι σε θέση να εντοπίσει κάποιες από τις παραμέτρους, δηλαδή περιορίζετε η αξία του σε κάποιο γραμμικό του Rp . Μέθοδοι για την τοποθέτηση γραμμικών μοντέλων με πολυσυγγραμμικότητα έχουν αναπτυχθεί και μερικοί απαιτούν επιπλέον παραδοχές, όπως η «επίδραση αραιότητας» δηλαδή ένα μεγάλο μέρος από τα

αποτελέσματα είναι ακριβώς μηδέν. Σημειώστε ότι τα περισσότερα υπολογιστικά ακριβά επαναληπτικών αλγορίθμων για την εκτίμηση των παραμέτρων, όπως αυτές που χρησιμοποιούνται σε γενικευμένο γραμμικό μοντέλο s, δεν υποφέρουν από αυτό το πρόβλημα, και στην πραγματικότητα αυτό είναι απολύτως φυσιολογικό κατά τον χειρισμό. Πέρα από αυτές τις υποθέσεις, αρκετές άλλες στατιστικές ιδιότητες των δεδομένων επηρεάζει σημαντικά την απόδοση των διαφορετικών μεθόδων εκτίμησης. Η στατιστική σχέση μεταξύ των όρων σφάλματος και στις ερμηνευτικές μεταβλητές παίζει σημαντικό ρόλο στον καθορισμό του κατά πόσον η διαδικασία εκτίμησης έχει επιθυμητές ιδιότητες δειγματοληψίας, όπως είναι αμερόληπτοι και συνεπείς. Η ρύθμιση ή κατανομή πιθανότητας του «x» »έχει μια σημαντική επιρροή των μεταβλητών πρόβλεψης σχετικά με την ακρίβεια των εκτιμήσεων των β. Δειγματοληψία και σχεδιασμός των πειραμάτων είναι τα ανεπτυγμένα υπό πεδία των στατιστικών στοιχείων που παρέχουν καθοδήγηση για τη συλλογή δεδομένων με τέτοιο τρόπο για να επιτευχθεί μια ακριβής εκτίμηση της β .

Ένα μοντέλο εφοδιασμένο με γραμμική παλινδρόμηση μπορεί να χρησιμοποιηθεί για τον προσδιορισμό της σχέσης μεταξύ μιας μόνο μεταβλητής πρόβλεψης xj και μιας μεταβλητής απόκρισης y όταν όλες οι άλλες μεταβλητές πρόβλεψης στο μοντέλο είναι σταθερές. Συγκεκριμένα, η ερμηνεία του βj συνεπάγεται με την αλλαγή στη y για αλλαγή μιας μονάδας στο xj όταν οι άλλες συνμεταβλητές παραμένουν σταθερές, που είναι η αναμενόμενη τιμή της μερικής παραγώγου του y σε σχέση με xj. Αυτό μερικές φορές ονομάζεται το μοναδικό αποτέλεσμα της xj με θέμα y . Σε αντίθεση, η οριακή επίδραση του xj με θέμα y μπορεί να αξιολογηθεί χρησιμοποιώντας συντελεστή συσχέτισης ή απλή γραμμική παλινδρόμηση μοντέλων, αφορούν xj το y .Αυτή η επίδραση είναι το συνολικό παράγωγο του y σε σχέση με την xj. Πρέπει να δοθεί προσοχή κατά την ερμηνεία των αποτελεσμάτων παλινδρόμησης, καθώς κάποιες από τις παλινδρομήσεις δεν μπορούν να χρησιμοποιηθούν για οριακές αλλαγές (όπως ψευδό μεταβλητές, ή ο όρος τομής), ενώ άλλες δεν μπορούν να θεωρηθούν σταθερές. Είναι πιθανό ότι το μοναδικό αποτέλεσμα μπορεί να είναι σχεδόν μηδέν, ακόμα και όταν το

οριακό αποτέλεσμα είναι μεγάλο. Αυτό μπορεί να σημαίνει ότι κάποια άλλη συνμεταβλητή καταγράφει όλες τις πληροφορίες xj, έτσι ώστε, όταν η μεταβλητή είναι στο μοντέλο, δεν υπάρχει καμία συμβολή των X j στη μεταβολή των y. Αντίθετα, η μοναδική επίδραση της xj μπορεί να είναι μεγάλη, ενώ το οριακό αποτέλεσμα είναι σχεδόν μηδενική. Αυτό θα συνέβαινε αν οι άλλες συνμεταβλητές εξηγούν τη μεγάλη διακύμανση του y , αλλά κυρίως εξηγούν παραλλαγή με έναν τρόπο που είναι συμπληρωματικό με αυτό που συλλαμβάνεται από xj. Στην περίπτωση αυτή, συμπεριλαμβανομένων και των άλλων μεταβλητών το μοντέλο μειώνει το μέρος διακύμανσης των y που δεν έχει σχέση με xj, ενισχύοντας έτσι την προφανή σχέση με xj. Η έννοια της έκφρασης "κρατούμενη σταθερά" μπορεί να εξαρτάται από το πώς προκύπτουν οι τιμές των μεταβλητών πρόβλεψης. Εάν ο πειραματιστής καθορίζει άμεσα τις τιμές των μεταβλητών πρόβλεψης σύμφωνα με μια σχεδίαση μελέτης, όπου οι συγκρίσεις μπορεί κυριολεκτικά να αντιστοιχούν σε συγκρίσεις μεταξύ των οποίων οι μονάδες πρόβλεψης μεταβλητών έχουν μια συνεχή σταθερά θα λέγαμε από τον πειραματιστή. Εναλλακτικά, η έκφραση συνεχή σταθερά μπορεί να αναφέρεται σε μια επιλογή που πραγματοποιείται στο πλαίσιο της ανάλυσης δεδομένων. Σε αυτή την περίπτωση, θα κρατήσουμε μεταβλητή σταθερά περιορίζοντας την προσοχή μας στα υποσύνολα των δεδομένων που τυχαίνει να έχουν μια κοινή τιμή για τη συγκεκριμένη μεταβλητή πρόβλεψης. Αυτή είναι η μόνη ερμηνεία της συνεχής σταθεράς που μπορεί να χρησιμοποιηθεί σε μια μελέτη παρατήρησης. Η έννοια του «μοναδικού φαινομένου" είναι ελκυστική όταν μελετά ένα πολύπλοκο σύστημα όπου πολλαπλά αλληλένδετα συστατικά επηρεάζουν τη μεταβλητή απόκρισης. Σε ορισμένες περιπτώσεις, μπορεί κυριολεκτικά να ερμηνευθεί ως η αιτιώδης επίδραση μιας παρέμβασης που συνδέεται με την τιμή μιας μεταβλητής πρόβλεψης. Ωστόσο, έχει υποστηριχθεί ότι σε πολλές περιπτώσεις η ανάλυση πολλαπλής παλινδρόμησης δεν αποσαφηνίζει τις σχέσεις μεταξύ των μεταβλητών πρόβλεψης και της μεταβλητής απόκρισης όταν οι προγνωστικοί παράγοντες συσχετίζονται μεταξύ τους και δεν έχουν εκχωρηθεί μετά από ένα σχέδιο μελέτης [1]. Μια ανάλυση των κοινών χαρακτηριστικών μπορεί να είναι χρήσιμη για την απεμπλοκή των κοινών και μοναδικών επιπτώσεων των συσχετισμένων ανεξάρτητων μεταβλητών. Πέρα από την πολλαπλή

παλινδρόμηση, χρησιμοποιώντας την ανάλυση των κοινών χαρακτηριστικών κατανοούνται καλύτερα τα αποτελέσματα.

Πολυάριθμες επεκτάσεις γραμμικής παλινδρόμησης έχουν αναπτυχθεί, οι οποίες επιτρέπουν ορισμένες ή όλες από τις υποθέσεις να στηρίζονται στο βασικό μοντέλο για να υπολογιστούν [33].

3.3.4 ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ (SIMPLE LINEAR REGRESSION)

Σε ένα σύνολο δειγμάτων τιμών {xi, yi} ψάχνουμε να βρούμε με ένα απλό μαθηματικό μοντέλο την σχέση μεταξύ αυτών των δύο μεταβλητών x και y. Στην ουσία αναζητούμε μια ευθεία γραμμή της μορφής $f(x) = y = a + \beta x$, με αυτό το μοντέλο μπορούμε να προβλέψουμε τις τιμές του y για κάθε νέα τιμή του x. Ο μηχανισμός αυτού του μοντέλου χρησιμοποιείται στην μηχανική μάθηση (machine learning).

Παράδειγμα : Ας υποθέσουμε ότι θέλουμε να διαπιστώσουμε την επίδραση ενός φαρμάκου σε 7 διαφορετικούς ασθενείς ανάλογα με τα κιλά τους.

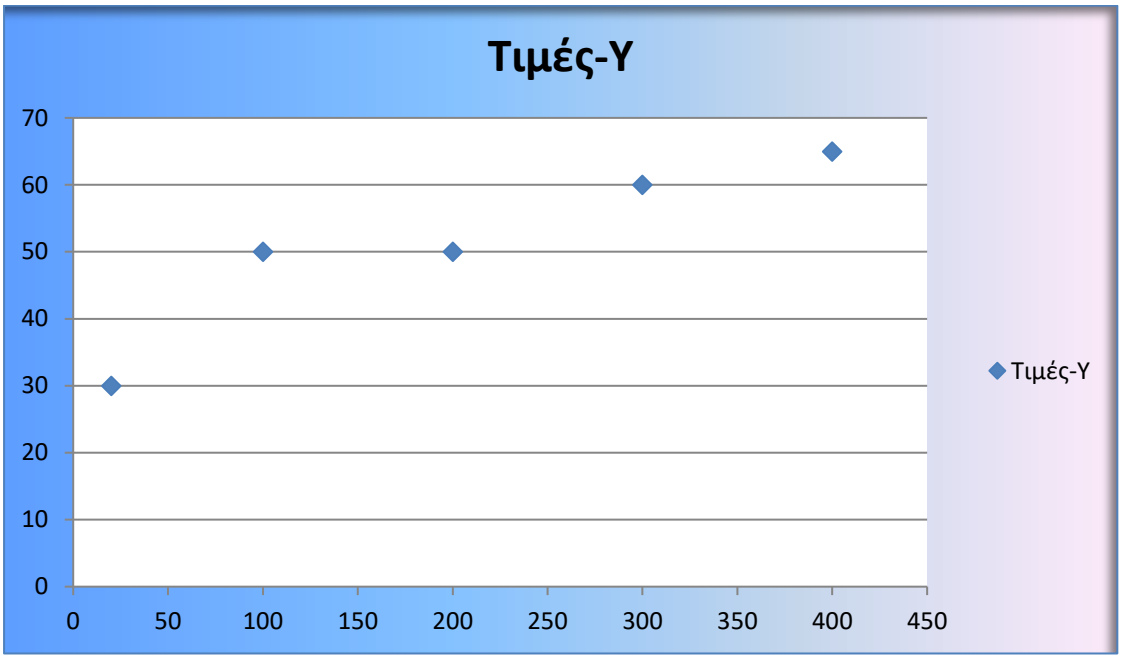
i	Ποσότητα Φαρμάκου(xi)	Κιλά(yi)
1	50	45
2	100	50
3	200	50
4	300	60
5	400	65
6	500	70
7	600	85

Εικόνα 6: Πίνακας Δεδομένων

Με βάση αυτή της μεθόδου μπορούμε να προβλέψουμε κάποιο σφάλμα στην κατάλληλη ποσότητα φαρμάκου που θα πρέπει να χορηγηθεί στον ασθενή.

Για τα σημεία n {xi, yi} , όπου i,…….,n για να καταλήξουμε σε μια ευθεία γραμμή.

Σε ένα τέτοιο παράδειγμα απλής γραμμικής παλινδρόμησης παίρνουμε το διάγραμμα διασποράς (scatter diagram, scater plot).



Εικόνα 7: Διάγραμμα Διασποράς

Έτσι προσεγγιστικά $f(x) = y = a + \beta x$, για i=1,2,……7 θεωρώντας μια ευθεία στο άνω διάγραμμα της μορφής.

3.3.5 ΜΕΘΟΔΟΣ ΤΩΝ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

Για να βρεθεί αυτή η ευθεία f (x), δηλαδή οι παράμετροι α και β μπορεί να χρησιμοποιηθεί η Μέθοδος των Ελαχίστων Τετραγώνων η οποία πρωτοεμφανίστηκε το 1805 σε μια εργασία του Γάλλου μαθηματικού Legendre (1752-1833) και στην συνέχεια στον Γερμανό μαθηματικό Gauss (177-1855) στην αστρονομική εργασία Theoria Motus όπου προσδιοριζόταν η τροχιά του μικρού πλανήτη Δήμητρα(Αδαμόπουλος Λεωνίδας, Δαμιανού Χαράλαμπος, Σβέρκος Ανδρέας – 2013).

Η απόσταση για κάθε σημείου {xi , yi} για να βρούμε την ευθεία θα είναι η ελάχιστη.

$\min_{\alpha,\beta} Q(\alpha, \beta)$, όπου$\sum_{i=1}^{n} \hat{e}i2 = \sum_{i=1}^{n} (y_i – \alpha - \beta x_i)^2$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_\iota - \bar{y})}{\sum_{i=1}^{n}}$$

Παραπάνω βλέπουμε χρησιμοποιούμε την μέθοδο με απειροστικό λογισμό.

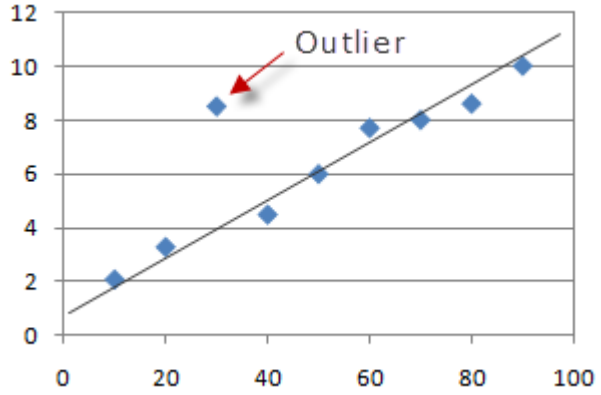
3.3.6 ΑΝΙΧΝΕΥΣΗ ΑΝΩΜΑΛΙΩΝ(OUTLIER DITECTION)

Ανίχνευση ανωμαλιών καλείται η αναγνώριση προτύπων από ένα σύνολο δεδομένων που εμφανίζουν διαφορετική συμπεριφορά από την προσδοκώμενη. Άλλες ονομασίες που συναντούνται στην βιβλιογραφία για τα πρότυπα είναι ακραίες τιμές, ασύμφωνες παρατηρήσεις, εξαιρέσεις, παρεκκλίσεις, ιδιαιτερότητες και προσμίξεις σε διαφορετικούς τομείς εφαρμογών [14]. Στόχος είναι η υψηλού επιπέδου ανίχνευση πιθανών ανωμαλιών, διατηρώντας όμως χαμηλά ποσοστά λανθασμένης προειδοποίησης. Ως εφαρμογή μπορούμε να αναφέρουμε τον προσδιορισμό απειλής στην έγκριση δανείων ή πιστωτικών καρτών από μια τράπεζα [15].

Υπάρχουν τρεις κατηγορίες τεχνικών ανίχνευσης ανωμαλιών [14]. Η επιβλεπόμενη ανίχνευση ανωμαλιών, όπου οι τεχνικές που έχουν εφαρμοστεί σε αυτήν την κατηγορία, λαμβάνουν την διαθεσιμότητα τόσο ενός συνόλου δεδομένων εκπαίδευσης που έχει χαρακτηριστεί ως κανονικό όσο και ενός χαρακτηρισμένου ως ακραίων τιμών. Σε τέτοιες περιπτώσεις, συνήθως, εφαρμόζονται μοντέλα πρόβλεψης για κανονικές έναντι ανώμαλων τάξεων. Η ημι-επιβλεπόμενη ανίχνευση ανωμαλιών, όπου οι τεχνικές σε αυτή την περίπτωση θεωρούν ότι ένα σύνολο δεδομένων εκπαίδευσης έχει χαρακτηριστεί μόνο κανονικό. Το θετικό αυτής της κατηγορίας είναι ότι εφαρμόζεται σε μεγαλύτερη κλίμακα από την προηγούμενη, για το λόγο ότι δεν απαιτείται επισήμανση για την τάξη ακραίων τιμών. Η συνήθης προσέγγιση σε αυτή την κατηγορία είναι η κατασκευή ενός μοντέλου για το σύνολο ώστε να ανταποκρίνεται στην κανονική συμπεριφορά, και ύστερα η εφαρμογή του για τον προσδιορισμό των ανωμαλιών στα δοκιμαστικά

δεδομένα. Τέλος, η μη-επιβλεπόμενη ανίχνευση ανωμαλιών, όπου δεν απαιτούνται δεδομένα εκπαίδευσης, και έτσι είναι οι πιο ευρέως εφαρμόσιμες τεχνικές. Οι τεχνικές σε αυτή τη περίπτωση υποθέτουν ότι τα κανονικά στιγμιότυπα είναι περισσότερα από τις ακραίες τιμές στα δοκιμαστικά δεδομένα. Αν αυτός ο συλλογισμός δεν ισχύει, τότε οι τεχνικές έχουν μεγάλο ποσοστό λανθασμένης εκτίμησης.

Στην ανίχνευση των ανωμαλιών συναντάμε δύο μεθόδους: την μονομερή μέθοδο (univariate methods) και τις πολυπαραγοντικές μεθόδους (multivariate methods), όπου η δεύτερη είναι και η επικρατέστερη κατά την χρησιμοποίηση της.



Εικόνα 8: Διάγραμμα ανίχνευσης ανωμαλιών

3.3.7 ΜΟΝΟΜΕΡΗΣ ΜΕΘΟΔΟΣ(UNIVARIATE METHODS)

Οι περισσότερες από τις πρώτες μονομερείς μεθόδων ανίχνευσης σφαλμάτων βασιζόταν στην παραδοχή μιας υποκείμενης κατανομής δεδομένων που όμως αυτές οι υποθετικές παραδοχές συνήθως έπεφταν έξω.

Μια κεντρική ιδέα ήταν αυτή της τυχαίας δειγματοληψίας όπου με την μη ταυτοποίηση χαρακτηριζόταν περιοχές σφαλμάτων – ανωμαλιών [21].

Όπου για α, $0 < \alpha < 1$, με α να ορίζεται το σφάλμα περιοχής και κανονική κατανομή $N(\mu, \sigma^2)$ έχουμε:

$$out(a, \mu, \sigma^2) = \left\{x: |x - \mu| > \zeta_{1-\frac{\alpha}{2}}\sigma\right\}$$

$\sigma^2 > 0$ = διακύμανση

μ = μέσος

χ = στήριγμα

Στην συνέχεια γίνανε κάποιες βελτιστοποιήσεις στην μονομερή μέθοδο

- ✓ Απλά Βήματα – Διαδοχικές Διαδικασίες.

  $out(a_n, \hat{\mu}_n, \hat{\sigma}_n^2) = \{x: |x - \hat{\mu}_n| > g(n, a_n)\hat{\sigma}_n\}$όπου ητο μέγεθος των δειγμάτων των $\hat{\mu}_n$, $\hat{\sigma}_n$ και $g(n, a_n)$ ορίζει τα όρια των αποκλεισμένων περιοχών [21].

- ✓ Εσωτερικές και Εξωτερικές Διαδικασίες (Inward and Outward Procedures). Όπου εδώ οι τιμές με τις μεγαλύτερες αποκλίσεις αποκλείονται και διαγράφονται από τα σετ των δεδομένων ,ενώ στην συνέχεια η διαδικασία συνεχίζεται [22].

- ✓ Εύρωστες Μετρήσεις (Univariate RobustMeasures) :εδώ το concept βασίζεται στο break down point σημείο καμπής θα μπορούσαμε να πούμε όπου ορίζεται ως το μικρότερο ποσοστό που μπορεί να λαμβάνει μεγάλες τιμές [23].

- ✓ Στατιστικός έλεγχος διαδικασίας (Statistical Process Control): Είναι συνδεδεμένοι με μονομερούς μεθόδους ανίχνευσης και αντιπροσωπεύουν μια στοχαστική διαδικασία και χρησιμοποιείται για περισσότερο από μισό αιώνα [24].

3.3.8 ΠΟΛΥΠΑΡΑΓΟΝΤΙΚΗ ΜΕΘΟΔΟΣ (MULTIVARIATE METHODS)

Πολλές φορές στην πολύ μεταβλητότητα δεν μπορεί να γίνει ανίχνευση ανωμαλιών για αυτό χρησιμοποιούμε την πολυπαραγοντική μέθοδο για να το επιτύχουμε [22].

- ✓ Στατιστική Μέθοδο Για Πολυπαραγοντική ανίχνευσή ανωμαλιών (Statistical Methods for Multivariate Outlier Detection): Εδώ γίνεται μια μέθοδος εξόρυξης δεδομένων και ανίχνευσή αυτών που είναι μακριά από το κέντρο κατανομής δεδομένων. Έτσι η απόσταση είναι ένα πολύ σημαντικό κριτήριο, η λεγόμενη Mahalanobis distance.[25] Όπου $\bar{x}_n$ ένα απλό διανυσματικό δείγμα και μήτρα V_n :

$$V_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)(x_i - \bar{x}_n)^T$$

Όπου απόσταση M_iγια iδεδομένα $i = 1, \ldots\ldots, n$

$$M_i = (\sum_{i=1}^{n}(x_i - \bar{x}_n)^T V_n^{-1}\,(x_i - \bar{x}_n))^{\frac{1}{2}}$$

Συνεπώς αυτές οι παρατηρήσεις θα μπορούσαμε να πούμε με μεγάλες αποστάσεις Mahalanobis υποδεικνύονται ως υπερβολικές τιμές.

- ✓ Πολυμεταβλητά εύρωστα μέτρα (Multivariate Robust Measures):Η χρήση αξιόπιστων εκτιμήσεων των πολυδιάστατων παραμέτρων διανομής μπορεί συχνά να βελτιώσει τις επιδόσεις του διαδικασίες ανίχνευσης σε περίπτωση υπερβολικών τιμών. Όπως και στις μονοδιάστατες διαδικασίες, ο μέσος όρος διανομής (μέτρηση της θέσης) και η μεταβλητότητα- διακύμανση (μέτρηση του σχήματος) είναι τα δύο πιο συχνά χρησιμοποιούμενα στατιστικά στοιχεία για την ανάλυση δεδομένων παρουσία υπερβολικών τιμών.

- ✓ Μέθοδοι εξόρυξης δεδομένων για ανίχνευση σφαλμάτων (Data-Mining Methods for Outlier Detection): Αυτές οι μέθοδοι έχουν σχεδιαστεί για τη διαχείριση μεγάλων βάσεων δεδομένων [27].

ΚΕΦΑΛΑΙΟ 4

ΥΛΟΠΟΙΗΣΗ ΚΑΙ ΕΠΙΔΕΙΞΗ

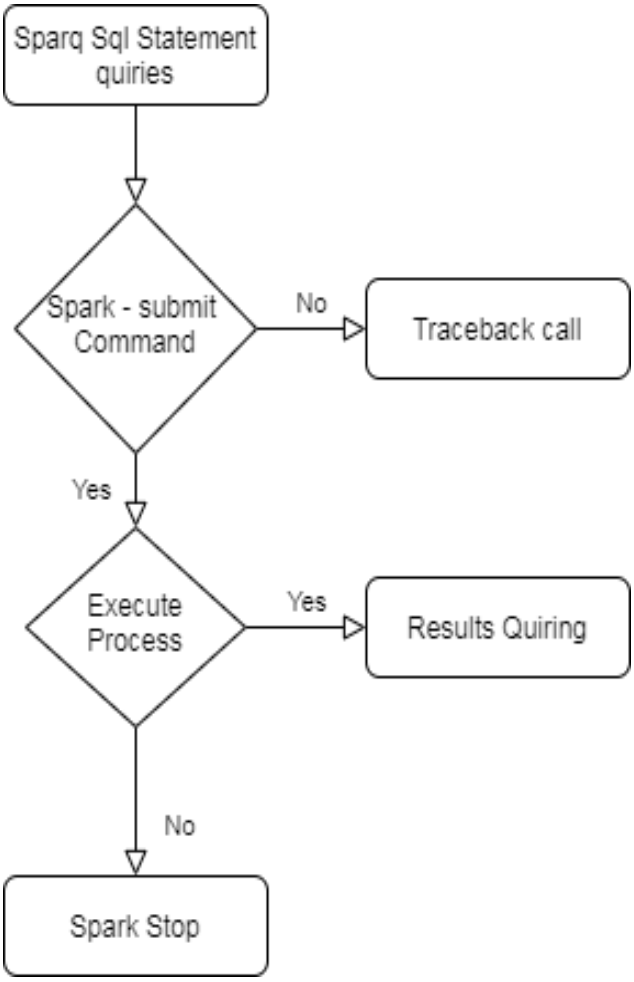
4.1 ΛΕΠΤΟΜΕΡΕΙΕΣ ΥΛΟΠΟΙΗΣΗΣ

Ο πειραματισμός έχει ως αρχή το framework Apache Sparkόπου σκοπός είναι να επιτύχουμε μία βέλτιστη λύση για πιο αποδοτική ανάλυση δεδομένων που μπορεί να αφορούν θέματα σχετικά με την ενέργεια ή για οποιαδήποτε άλλη διαδικασία.

Ένα από τα πλεονεκτήματα του Apache Spark είναι ότι έχει την δυνατότητα διασύνδεσης με αρκετές βάσεις δεδομένων, και δεύτερον ότι μπορεί να εκτελείται αρκετά πιο γρήγορα. Η κατασκευή μας θα στηριχθεί σε έναν απλό κώδικα ο οποίος θα είναι γραμμένος σε Pythonμέσω της βιβλιοθήκης του pyspark που χρησιμοποιείται στο Spark.

Το όλο εγχείρημα βασίζεται σε τρία προγράμματα (script) γραμμένα σε κώδικα Python, έτσι ώστε να εκτελούνται από τον μελετητή ή αναλυτή μεμονωμένα αν κρίνει αυτός σωστό η και τα τρία διαδοχικά. Και τέλος τα αποτελέσματα να οπτικοποιούνται σε μία οθόνη (monitor) ώστε να έχει μία πιο εμπεριστατωμένη εικόνα.

Όπως και προαναφέραμε θα βασιστούμε πάνω στο Apache Spark, και τα τρία προγράμματα και τα οποία θα τρέξουν είναι: σε θέμα με την γραμμική παλινδρόμηση όπου εξετάζει τη σχέση μεταξύ δύο μεταβλητών η και περισσοτέρων μεταβλητών με απώτερο σκοπό την πρόβλεψη μιας από αυτές μέσω των άλλων, την ανίχνευση ανωμαλιών όπου εξετάζουμε την διαφορετικότητα κάποιων δεδομένων από την προσδοκώμενη, και τέλος θα τρέξουμε κάποια βασικά ερωτήματα (queries) που χρειάζεται συνήθως να εξετάσει ο μελετητής.



Εικόνα 9: Flow chart διεργασίας και εκτέλεσης του προγράμματος

4.2 ΛΟΓΙΣΜΙΚΟ (SOFTWARECOMPONENT)

Η εργασία μας θα τρέξει σε λειτουργικό σύστημα Linux UBUNTU 18.04 στο οποίο είναι εγκαταστημένο το λογισμικό της Apache Spark μέσω της βιβλιοθήκης pyspark, ενώ η έκδοση του Sparkείναι 2.4.4 στο οποίο θα βασιστούμε για να τρέξουμε το project. Η γλώσσα προγραμματισμού στην οποία θα τρέξουμε είναι η Python έκδοσης 2.7.17.

4.3 ΕΚΤΕΛΕΣΗ

4.3.1 ΠΑΡΑΔΕΙΓΜΑ SPARK SQL ΜΕ ΕΝΕΡΓΕΙΑΚΑ ΔΕΔΟΜΕΝΑ (SPARK SQL QUERIES EXAMPLE)

Ο σκοπός ενός μελετητή είναι να λαμβάνει τα δεδομένα και να καταλήγει σε συμπεράσματα και αποτελέσματα τέτοια ώστε να βοηθούν στην σωστή λειτουργία ενός έξυπνου δικτύου και στην προκειμένη περίπτωση μιας έξυπνης κατοικίας.

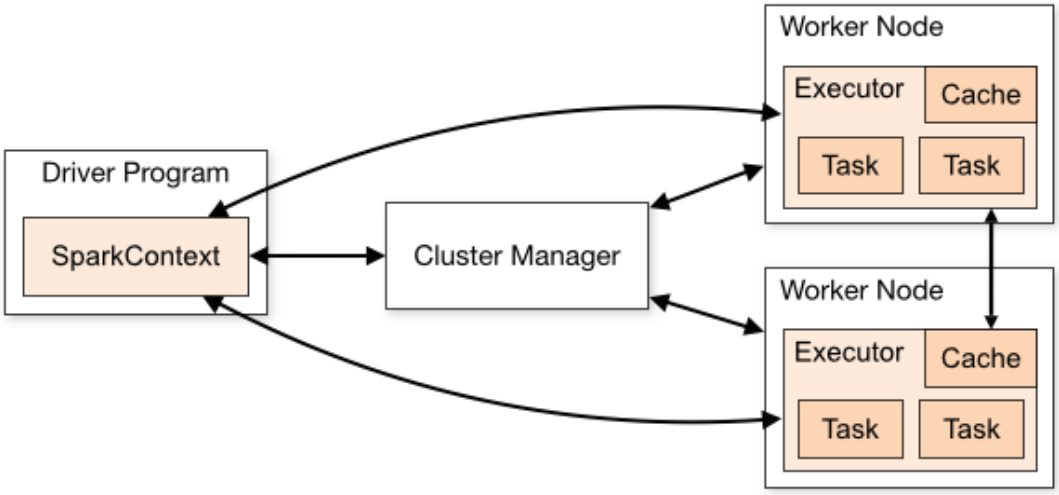
Έστω τώρα ότι έχουμε ένα έξυπνο σπίτι το οποίο αποτελείται από μετρητές και αισθητήρες όπου μετράνε κάθε μέρα εικοσιτέσσερις ώρες το εικοσιτετράωρο την τάση του ρεύματος. Γιατί όμως είναι απαραίτητη η μέτρηση της τάσης μιας οικίας ? γιατί πολύ απλά μπορεί να μας καθορίσει την κατανάλωση της ενέργειας κατά την διάρκεια της ημέρας αλλά και να μας οδηγήσει σε συμπεράσματα όπως προαναφέραμε σχετικά με την κατανάλωση ενέργειας που μπορεί να έχουμε από κάποια συσκευή αλλά και πως επηρεάζεται η κάθε συσκευή από τον ανθρώπινο παράγοντα που συνήθως τις πιο πολλές φορές είναι καθοριστικός.

Στο παρακάτω παράδειγμα θα χρησιμοποιήσουμε κάποια έτοιμα δεδομένα πειραματικά θα λέγαμε για να δείξουμε με ένα απλό προγραμματιστικό τρόπο πως μπορούμε να δώσουμε κάποια ερωτήματα στο πρόγραμμα μας το οποίο τρέχουμε και να πάρουμε κάποιες απαντήσεις.

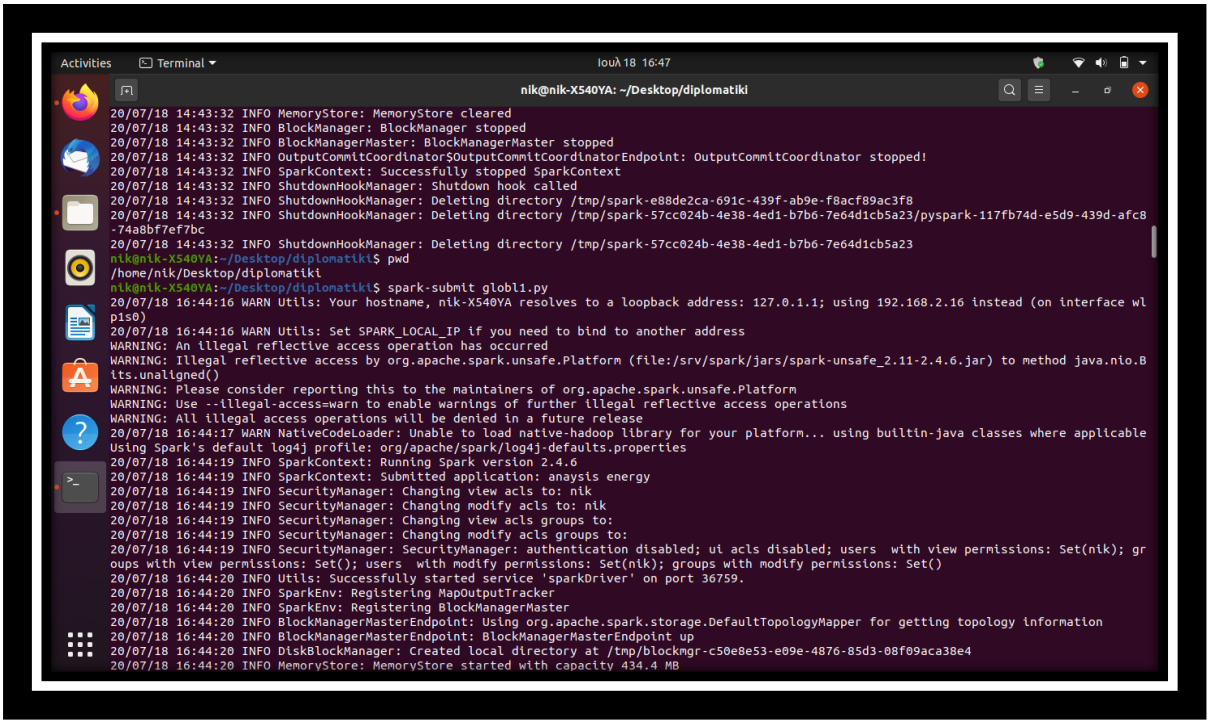
Τα δεδομένα τα οποία αναλύουμε στο παράδειγμα μας αναφέρονται στην τάση του ρεύματος που χρησιμοποιείται σε ένα έξυπνο σπίτι σε καθημερινή βάση και πραγματική χρονική συνέχεια κατά την διάρκεια της ημέρας.

Η μορφή τους δηλαδή το format το οποίο είναι αποθηκευμένα μπορεί να είναι σε μορφή .jason, .csv και μερικά ακόμα format. Στο δικό μας παράδειγμα θα χρησιμοποιήσουμε .csv format το οποίο μοιάζει πολύ με το πολύ γνώριμο xml .

Έχοντας λοιπόν τα δεδομένα μας μέσω του τερματικού μας και από τον τρέχοντα φάκελο εκτελούμε την εντολή spark-submit, απαραίτητη για να τρέξουμε και να πάρουμε τα αποτελέσματα μας. Χρησιμοποιούμε το Spark Stand Alone Cluster μέσω του SparkContex που καθορίζουμε στο πρόγραμμα μας.



Εικόνα 10: Spark Stand Alone Cluster



Εικόνα 11: Εκτέλεση του Προγράμματος

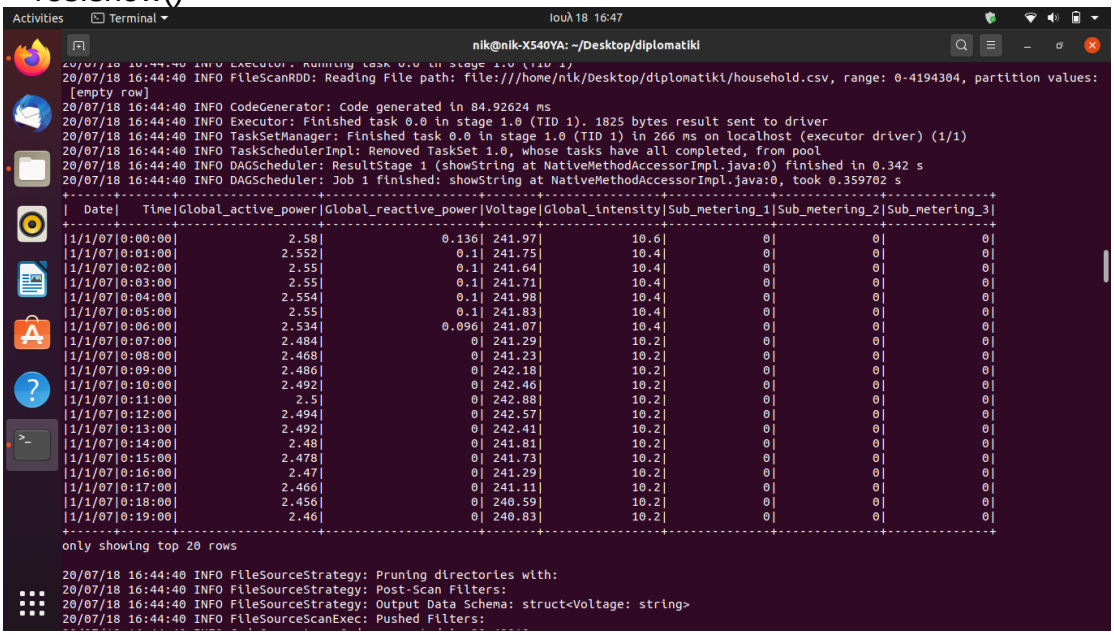
Τώρα όσο αφορά για τα ερωτήματα μας, αυτά τα θέτουμε κάθε φορά μέσα στο πρόγραμμα (script) το οποίο έχουμε δημιουργήσει. Παρακάτω βλέπουμε τον κώδικα τον οποίο χρησιμοποιούμε για το παράδειγμα μας.

```
from __future__ import print_function

from pyspark.sql import SparkSession

from pyspark.sql import Row

from pyspark.sql.functions import countDistinct, avg, stddev

from pyspark.sql.types import *
```

Οι βιβλιοθήκες τις οποίες τρέχουμε είναι :

- from__future__import print_function: η οποία χρησιμοποιείται και διατίθεται η συνάρτηση από το λογισμικό μας.
- From pyspark.sql import SparkSession: κύρια είσοδος για τα δεδομένα και την χρήση τους από την βάση δεδομένων (SQL).
- from pyspark.sql import Row: μια σειρά δεδομένων σε ένα πλαίσιο δεδομένων (data frame).
- from pyspark.sql.functions import countDistinct, avg, stddev: εισαγωγή συναρτήσεων countDistinct όπου η συνάρτηση επιστρέφει τον αριθμό των διακριτών στοιχείων σε στήλες. Η avg επιστρέφει τον μέσο όρο τιμών. Και η stddev επιστρέφει την τυπική απόκλιση τιμών των στηλών.
- from pyspark.sql.types import *: Περιέχει μια σειρά από βιβλιοθήκες όπως τον καθορισμό των δεκαδικών αριθμών και πολλά άλλα.

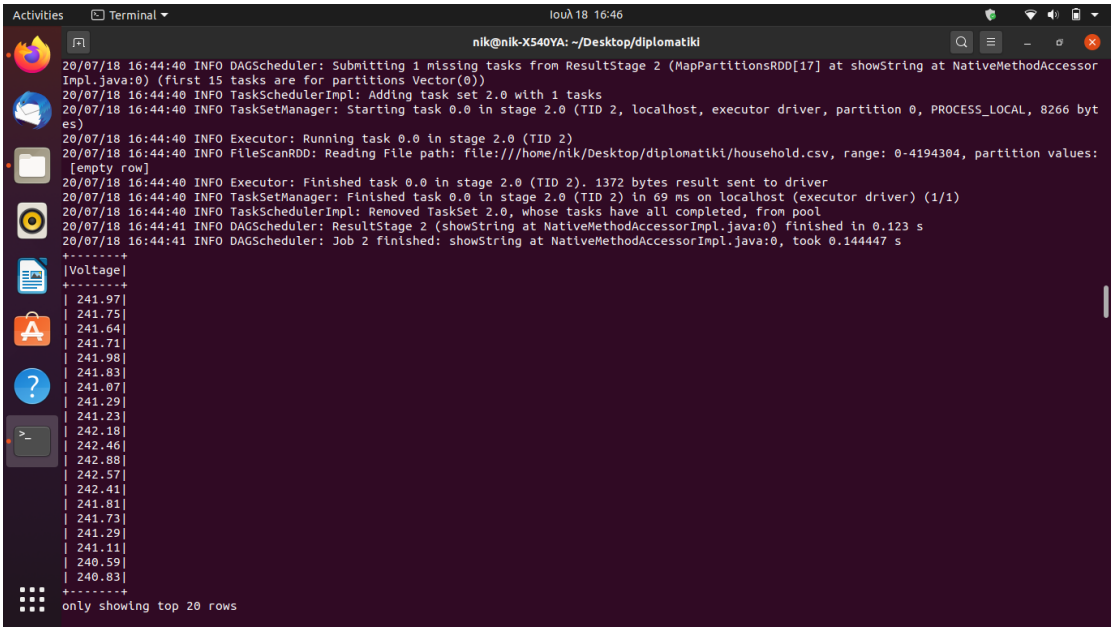
```
def thermo_results_v(spark):

res=spark.read.format('csv').options(header='true').load("file:////home/nik/Desktop/diplomatiki/household.csv")

  # shows matrix (Συνοπτική Εικόνα)

    res.show()
```



Εικόνα 12: Αποτελέσματα σε στήλες

```
    # shows column Voltage (Παίρνουμε Την Στήλη Της Τάσης)

    res.select("Voltage").show()
```

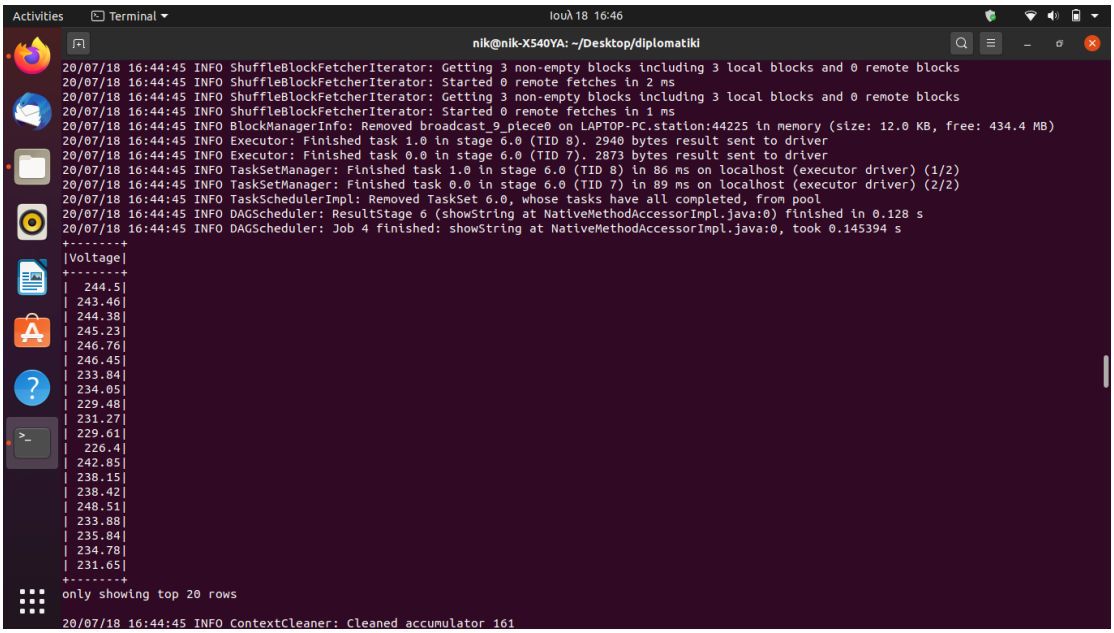
Η εντολή select είναι αρκετά χρήσιμη για τον παρατηρητή αφού μας δίνει την δυνατότητα σε περιπτώσεις που έχουμε μεγάλο αριθμό στηλών να αναζητάμε και να παίρνουμε αυτήν που θέλουμε.



Εικόνα 13: Στήλη Τάσης (Voltage)

Maximum Value (Μέγιστη Τιμή Της Στήλης που Επιθυμούμε)

```
res.groupBy("Voltage").max().show()
```



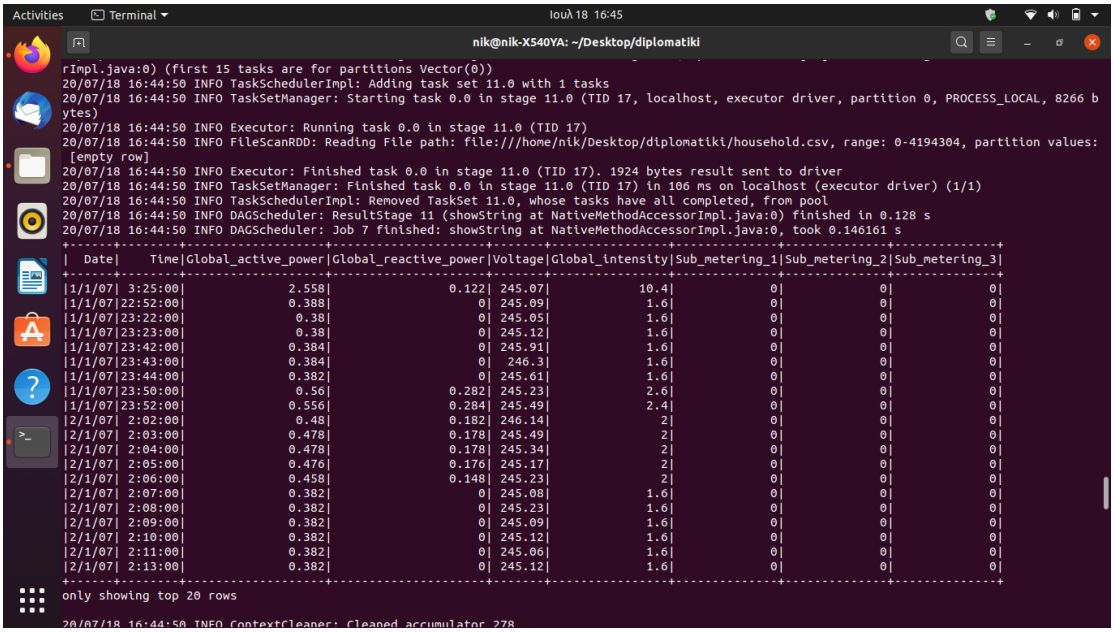
Εικόνα 14 : Στήλη μέγιστης τάσης

Minimum Value

```
res.groupBy("Time").min().show()
```

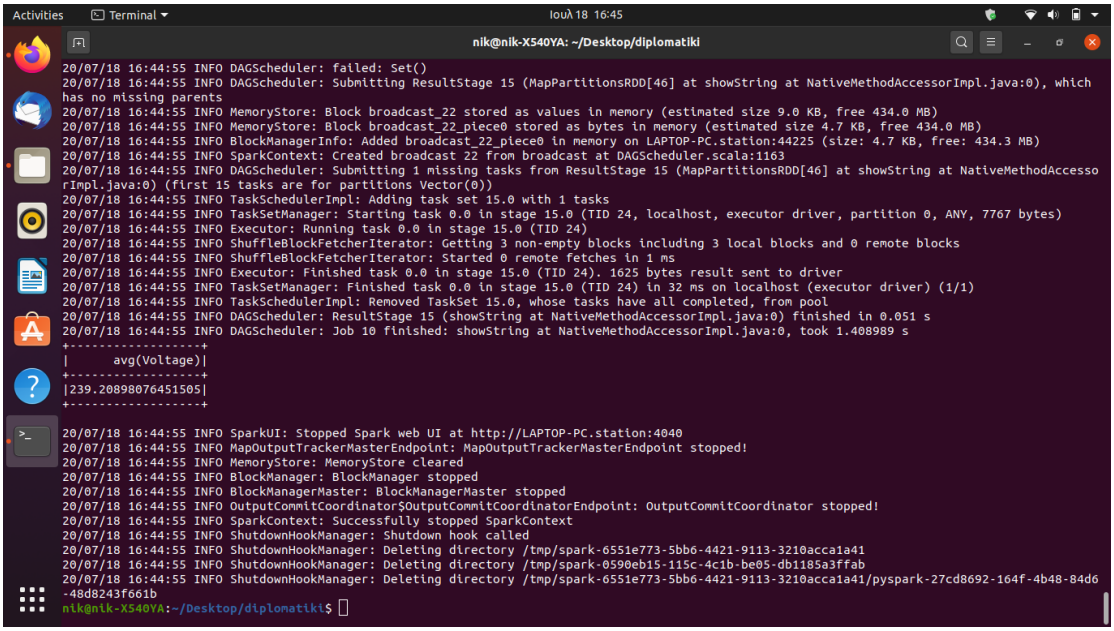
Take the number bigger than 245 (Εδώ επιλέγουμε την τάση μεγαλύτερη της τιμής 245)

```
res.filter(res["Voltage"] > 245.0).show()
```



Εικόνα 15: Απεικόνιση τάσης μεγαλύτερης της τιμής 245

```
# shows reuslt from 2007 (παίρνοντας αποτελέσματα μιας συγκεκριμένης χρονολογίας)

    res.filter(res["Date"] == 2007).show()

 # Results of average (Μέσος όρος)

    res.select(avg("Voltage")).show()
```



Εικόνα 16: Απεικόνιση μέσου όρου (average)

```
if __name__ == "__main__":

spark = SparkSession \
.builder \
.appName("anaysis energy") \
.config("spark.some.config.option",
"some-value") \
.getOrCreate()

thermo_results_v(spark)

spark.stop()
```

Το SparkSession είναι το σημείο εισόδου στο Spark SQL. Είναι ένα από τα πρώτα αντικείμενα που δημιουργείτε κατά την ανάπτυξη μιας εφαρμογής Spark SQL. Ως προγραμματιστής Spark, δημιουργείτε ένα SparkSession.

Τέλος κάθε φορά μπορούμε να τροποποιούμε το σκριπτάκι μας και να θέτουμε τα δικά μας ερωτήματα και να παίρνουμε της απαντήσεις που θέλουμε.

4.3.2 OUTLIER DETECTION

Στην συνέχεια εκτελούμε το πρόγραμμα μας για να ανιχνεύσουμε κάποια μη εμφανή σφάλματα κατά την επεξεργασία δεδομένων αλλά και τι θα μπορούσαν να μας δώσουνε αυτά τα δεδομένα, όπως για παράδειγμα κάποιος αισθητήρας θερμοκρασίας να έχει βγει εκτός λειτουργίας και να μην δίνει σωστές μετρήσεις. Εδώ σε αυτό το πρόγραμμα θα χρησιμοποιήσουμε την μέθοδο μιας κλάσης.

Στόχος της μεθόδου μιας κλάσης είναι σε περιπτώσεις που πρέπει να δοκιμαστούν νέα δεδομένα και να διαπιστώσουμε αν δεν είναι ίδια με τα είδη υπάρχοντα δεδομένα. Αυτή η μέθοδος τα τελευταία χρόνια έχει κερδίσει μεγάλη δημοτικότητα.

Φανταστείτε έναν τύπο εργοστασιακής ρύθμισης, βαριά μηχανήματα υπό συνεχή παρακολούθηση κάποιου προηγμένου συστήματος. Το καθήκον του συστήματος ελέγχου είναι να προσδιορίσει πότε κάτι πάει στραβά. τα προϊόντα είναι κάτω από την ποιότητα, το μηχάνημα παράγει παράξενες δονήσεις ή κάτι σαν μια θερμοκρασία που ανεβαίνει. Είναι σχετικά εύκολο να συγκεντρωθούν στοιχεία κατάρτισης για καταστάσεις που είναι εντάξει και είναι απλώς η συνήθης κατάσταση της παραγωγής. Αλλά από την άλλη πλευρά, τα δεδομένα από τη συλλογή ενός ελαττωματικού συστήματος ενδέχεται να είναι αρκετά εσφαλμένα. Αν μια κατάσταση ελαττωματικού συστήματος μπορεί να προσομοιωθεί, δεν υπάρχει τρόπος να διασφαλιστεί ότι όλες οι ελαττωματικές καταστάσεις είναι προσομοιωμένες και έτσι αναγνωρίζονται σε ένα παραδοσιακό πρόβλημα δύο κλάσεων.

Με την παροχή των κανονικών δεδομένων εκπαίδευσης, ένας αλγόριθμος δημιουργεί ένα (αντιπροσωπευτικό) μοντέλο αυτών των δεδομένων. Εάν τα πρόσφατα συναντηθέντα δεδομένα είναι πολύ διαφορετικά, σύμφωνα με ορισμένες μετρήσεις, από αυτό το μοντέλο, χαρακτηρίζονται ως εκτός τάξης. Θα εξετάσουμε την εφαρμογή των Μηχανισμών διεύθυνσης υποστήριξης σε αυτό το πρόβλημα μιας κατηγορίας [28].

Διαχωρίζει βασικά όλα τα σημεία δεδομένων από την προέλευση στο χώρο και μεγιστοποιεί την απόσταση από αυτό και ως την προέλευση. Αυτό έχει ως

αποτέλεσμα μια δυαδική συνάρτηση που καταγράφει περιοχές στον χώρο εισόδου όπου είναι η πιθανότητα πυκνότητας των δεδομένων [28].

Μια πολύ ωραία ιδιότητα των διανυσμάτων υποστήριξης (SVM) είναι ότι μπορεί να δημιουργήσει ένα μη γραμμικό όριο απόφασης προβάλλοντας τα δεδομένα μέσω μιας μη γραμμικής συνάρτησης φ σε ένα χώρο με υψηλότερη διάσταση. Αυτό σημαίνει ότι τα σημεία δεδομένων που δεν μπορούν να χωριστούν με ευθεία γραμμή στον αρχικό τους χώρο, "ανυψώνονται" σε ένα χώρο χαρακτηριστικών F [28].

Ο τύπος μιας κλάσης διανυσμάτων υποστήριξης (SVM) είναι:

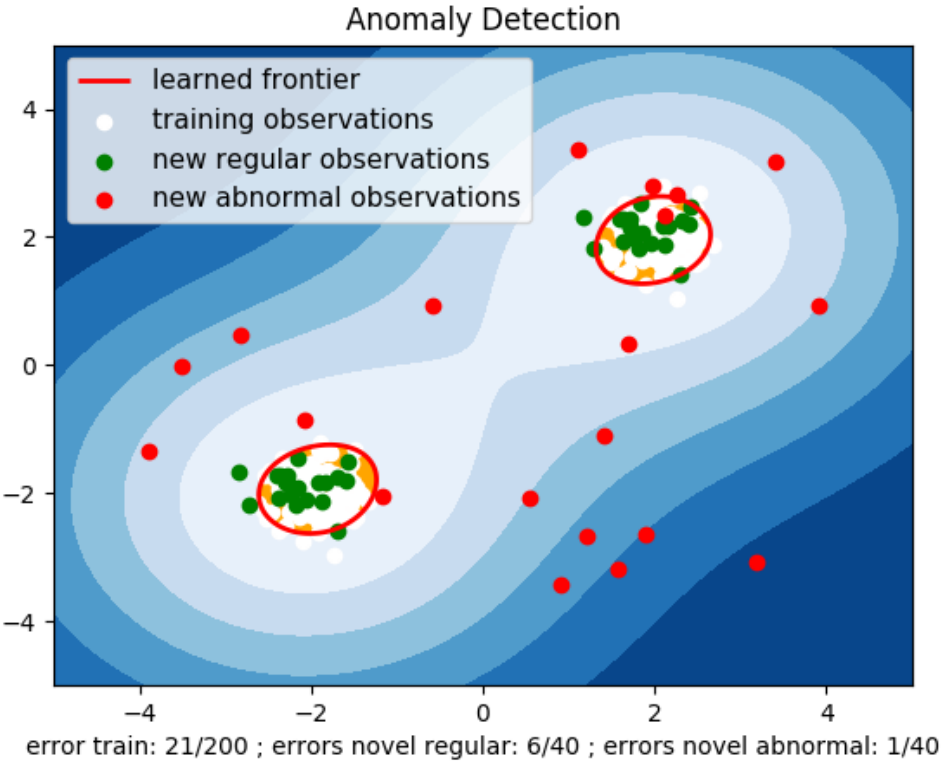
$$\min_{w,\xi_{i,\rho}} \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{\iota=1}^{n}\xi_i - \rho$$

$$\left(w \cdot \varphi(x_i)\right) \geq \rho - \xi_i for\ all\ i = 1, \dots, n$$

$$\xi_i \geq 0 \qquad\qquad for\ all\ i = 1, \dots, n$$

Όπου ξ_i επρόκειτο για τις slack variables,και το w και το ρ που έχει μέγιστη απόσταση από την προέλευση του χώρου χαρακτηριστικών F και διαχωρίζει όλα τα σημεία δεδομένων από την προέλευση.

Αυτό έχει ως αποτέλεσμα μια δυαδική συνάρτηση που καταγράφει περιοχές στον χώρο εισόδου όπου είναι η πυκνότητα πιθανότητας των δεδομένων.



Εικόνα 17: Ανίχνευση ανωμαλιών

Σε αυτό το πρόγραμμα έχει γραφτεί και κώδικας βασισμένος πάνω στο τομέα του matplot με χρησιμοποιήσει των βιβλιοθηκών που χρησιμοποιούνται στη python import matplotlib. pyplotasplt, και import matplotlib.font_manager για γραφική σχεδίαση και απεικόνιση των αποτελεσμάτων.

Επίσης χρησιμοποιούνται οι βιβλιοθήκες NumPy που χρησιμοποιούνται για την επιστημονική υπολογιστική και την εκτέλεση διαφόρων λειτουργιών [30]. Ενώ η βιβλιοθήκη sklearn προωθείτε για την εκμάθηση αλγορίθμων .[29]

Όπως βλέπουμε στην εικόνα 4.2 αλλά και στην συνέχεια στον κώδικα, σε ένα πλήθος δεδομένων παίρνουμε έναν αριθμό όμοιων δεδομένων (observations) όπως αναφέρεται και από τον μελετητή, καθώς και ένας αριθμός μη όμοιων δεδομένων (abnormal observations). Συμπερασματικά αυτό είναι μια πολύ καλή μέθοδος έτσι ώστε να μπορούμε να ελέγχουμε τα σφάλματα αριθμητικά μέσα σε ένα πλήθος δεδομένων αλλά και πόσο μεγάλη είναι η απόκλιση τους από τα συνήθη δεδομένα.

Παρακάτω βλέπουμε ένα κομμάτι του κώδικα:

```
print(__doc__)

import numpy as np
import matplotlib.pyplot as plt
import matplotlib.font_manager
from sklearn import svm

xx, yy = np.meshgrid(np.linspace(-5, 5, 500), np.linspace(-5, 5, 500))
# Generate train data
X = 0.3 * np.random.randn(100, 2)
X_train = np.r_[X + 2, X - 2]
# Generate some regular novel observations
X = 0.3 * np.random.randn(20, 2)
X_test = np.r_[X + 2, X - 2]
# Generate some abnormal novel observations
X_outliers = np.random.uniform(low=-4, high=4, size=(20, 2))

# fit the model
clf = svm.OneClassSVM(nu=0.1, kernel="rbf", gamma=0.1)
clf.fit(X_train)
y_pred_train = clf.predict(X_train)
y_pred_test = clf.predict(X_test)
y_pred_outliers = clf.predict(X_outliers)
n_error_train = y_pred_train[y_pred_train == -1].size
n_error_test = y_pred_test[y_pred_test == -1].size
n_error_outliers = y_pred_outliers[y_pred_outliers == 1].size

# plot the line, the points, and the nearest vectors to the plane
Z = clf.decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)

plt.title("Anomaly Detection")
plt.contourf(xx, yy, Z, levels=np.linspace(Z.min(), 0, 7), cmap=plt.cm.Blues_r)
a = plt.contour(xx, yy, Z, levels=[0], linewidths=2, colors='red')
plt.contourf(xx, yy, Z, levels=[0, Z.max()], colors='orange')
```

Εικόνα 18: Κώδικας για ανίχνευση ανωμαλιών σε γλώσσα Python

Μια μικρή αναφορά για τον κώδικα μας:

Αρχικά χρησιμοποιούμε την λειτουργία (function) np.meshgrid έτσι ώστε να καθορίσουμε το ορθογώνιο πλέγμα με τις καρτεσιανές συντεταγμένες με τον τετμημένο και τεταγμένο άξονα. Στην ουσία η meshgrid η οποία είναι και εμπνευσμένη από το λογισμικό MatLab επιστρέφει δύο δισδιάστατους πίνακες που αντιπροσωπεύουν τις συντεταγμένες όλων των σημείων x και y.

np.linspace(-5, 5, 500)

Επρόκειτο για μια λειτουργία των βιβλιοθηκών της Python που σκοπός της είναι η δημιουργία μις ομοιόμορφης αλληλουχίας σε ένα καθορισμένο διάστημα ορίζοντας και την αριθμητική ποσότητα των δειγμάτων.

np.random.rand Μία συνάρτηση με την οποία δημιουργούμε έναν πίνακα με τις γραμμές και της στήλες που επιθυμούμε.

Numpy.random.uniform: εδώ θα καθορίσουμε τα όρια στην ουσία οπού θα πάρουμε τα σφάλματα μας (outliers).

svmOneclassSVM(nu=0,1 kernel=’rbf’ gamma=0,1): Είναι ο κύριος αλγόριθμος μας τον οποίο θα τρέξουμε για να πάρουμε τα ταξινομημένα δεδομένα μας, τα οποία μπορεί να είναι όμοια η διαφορετικά στο σύνολο τους.

Kernel: καθορίζει τον τύπου του αλγορίθμου.

Nu=0,1 : Είναι το όριο σφάλματος

Gamma: συντελεστής συνήθως είναι ίσον με 0,1

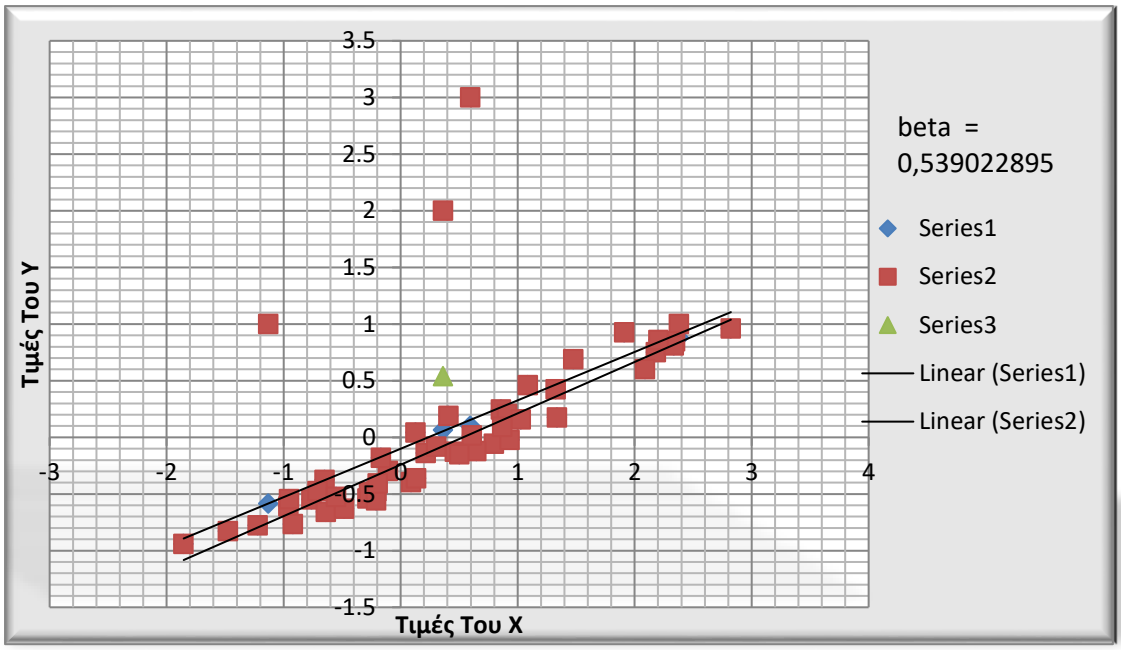
```
plt.title("Anomaly Detection")
plt.contourf(xx, yy, Z, levels=np.linspace(Z.min(), 0, 7), cmap=plt.cm.Blues_r)
a = plt.contour(xx, yy, Z, levels=[0], linewidths=2, colors='red')
plt.contourf(xx, yy, Z, levels=[0, Z.max()], colors='orange')

b1 = plt.scatter(X_train[:, 0], X_train[:, 1], c='white')
b2 = plt.scatter(X_test[:, 0], X_test[:, 1], c='green')
c = plt.scatter(X_outliers[:, 0], X_outliers[:, 1], c='red')
plt.axis('tight')
plt.xlim((-5, 5))
plt.ylim((-5, 5))
plt.legend([a.collections[0], b1, b2, c],
           ["learned frontier", "training observations",
            "new regular observations", "new abnormal observations"],
           loc="upper left",
           prop=matplotlib.font_manager.FontProperties(size=11))
plt.xlabel(
    "error train: %d/200 ; errors novel regular: %d/40 ; "
    "errors novel abnormal: %d/40"
```

Εικόνα 19: Δημιουργία γραφήματος μέσω κώδικα

Στην συνέχεια μέσω της βιβλιοθήκης μας matplotlib δημιουργούμε το γράφημα μας , ενώ μας δίνεται να καθορίσουμε τα χρώματα μας το πάχος της γραμμής και να τοποθετήσουμε τα όρια μας στον τετμημένο και τεταγμένο άξονα.

4.3.3 LINEAR REGRESSION

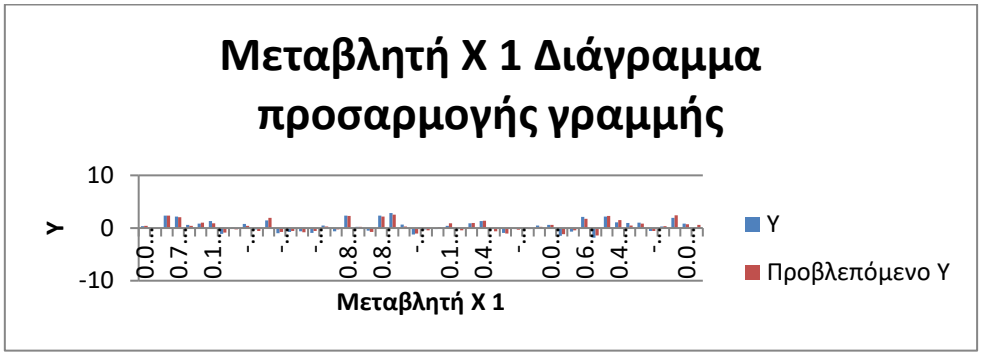
Στην συνέχεια θα τρέξουμε την γραμμική παλινδρόμηση όπου επανερχόμενοι στην γενικότερη περίπτωση, επιθυμούμε να διερευνήσουμε τη σχέση μεταξύ των μεταβλητών X, Y. Το γράφημα που απεικονίζουμε στην συνέχεια εξάγεται μέσω μιας βιβλιοθήκης της οποίας έχει γίνει εγκατάσταση στο λογισμικό του Microsoft Office 2007. Έτσι δίνοντας τα δεδομένα μας και τρέχοντας την f(x) της linear regression για το excel του Microsoft Office 2007 παίρνουμε την εκτιμώμενη ευθεία.



Εικόνα 20: Γράφημα απεικόνισης μέσω XML

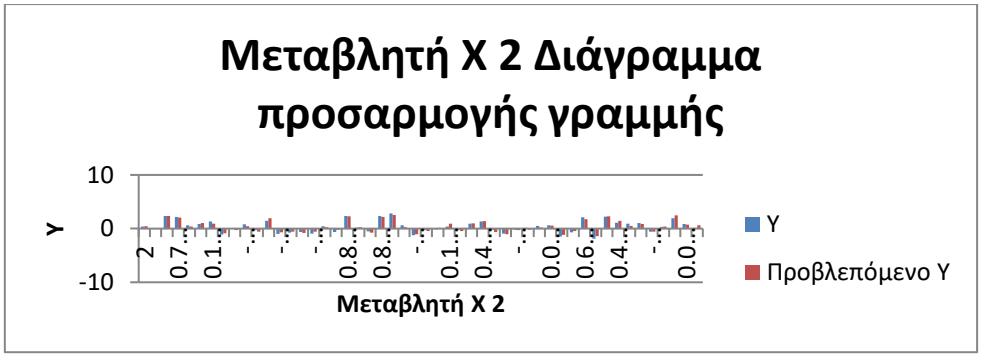
Υπογραμμίζεται ότι παραπάνω βλέπουμε την εκτιμημένη ευθεία.

Το standardized coefficients Beta είναι η εκτίμηση του b1 όταν εφαρμοστεί το μοντέλο Yi= b0 + b1Xi΄+ ei, όπου Xi΄ είναι οι τυποποιημένες τιμές των Xi (αυτό έχει μεγαλύτερη χρησιμότητα στο πολλαπλό μοντέλο όπου έχουμε πολλές ανεξάρτητες μεταβλητές και θέλουμε να δούμε τις εκτιμήσεις των b1 όταν οι μεταβλητές αυτές μετρώνται στην ίδια κλίμακα).



Εικόνα 21: Γράφημα με Τιμές του X1 και προβλεπόμενο Y

Επίσης η συνάρτηση μας δίνει ανάλογα με την μεταβλητή X το προβλεπόμενο Y, ονομαζόμενο και ως διάγραμμα προσαρμογής γραμμής.



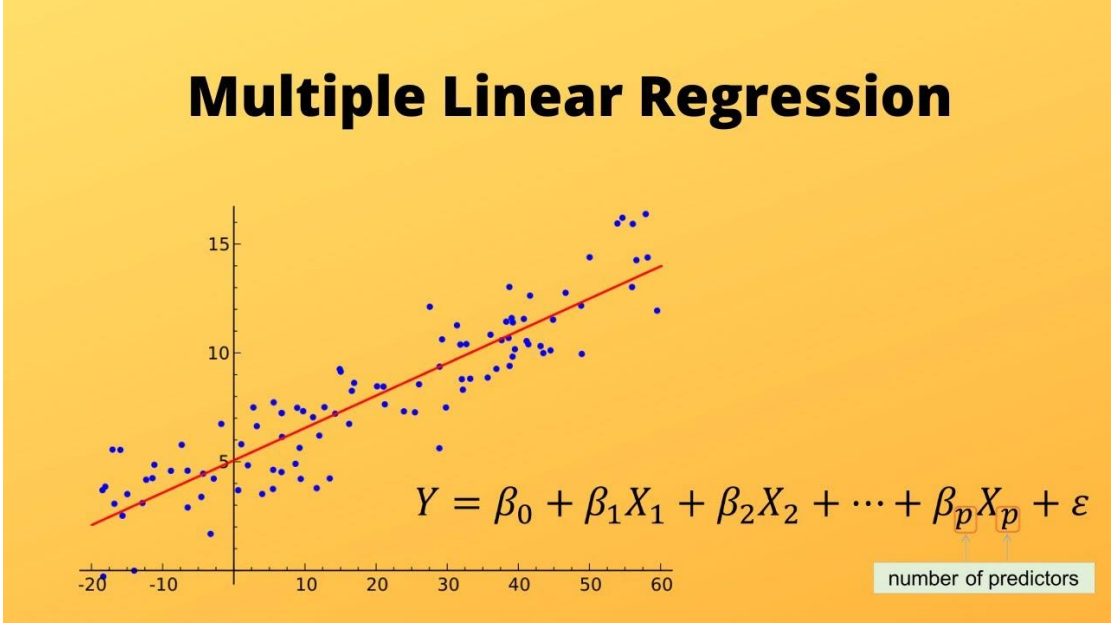
Εικόνα 22: Γράφημα με Τιμές του X2 και προβλεπόμενο Y

Θα λέγαμε είναι ένας εύκολος και γρήγορος τρόπος χωρίς χρειαστεί κάποιος μελετητής η μηχανικός να ανατρέξει σε κώδικα και να προγραμματίσει.

Η βιβλιοθήκη έχει την ονομασία LINEST και μπορεί να γίνει ελεύθερα η λήψη τους από το λογισμικό του Office και βασίζεται στην εξίσωση y = mx + b.

Στην συνέχεια μέσω κώδικα θα τρέξουμε μια πιο πολύπλοκη μέθοδο της γραμμικής παλινδρόμησης. Το πρόγραμμα μας είναι σε standalone γραμμένο σε κώδικα Python βασιζόμενοι στην λειτουργία lambda μέσω καθορισμού μιας συνάρτησης, καθώς και την χρήση βιβλιοθηκών Numpy με την χρήση του array όπου καθορίζουμε έναν πίνακα που συνήθως περιέχουν αριθμούς που όμως μπορεί να είναι αρκετά σημαντικοί για κάποιον μελετητή αφού μπορεί να δώσει άμεσα αποτελέσματα ώστε να τα επεξεργαστεί αλλά και να παρέμβει στον κώδικα και να τον διαμορφώσει όταν χρειάζεται σε περιπτώσεις όπου χρειάζεται για την εργασία του.

Παρακάτω βλέπουμε την εικόνα 4.11 και το παράδειγμα της πολλαπλής γραμμικής παλινδρόμησης.



Εικόνα 23: Πολλαπλή γραμμική παλινδρόμηση

Όπου το διαφορετικό πλήθος των μεταβλητών X_1 , X_2 ….X_p και οι παράμετροι β_0 , β_1 … β_p επηρεάζουν την τιμή της μεταβλητής Y.

Ενώ το **ε** είναι το τυχαίο σφάλμα.

Τέλος να πούμε ότι όταν ο χρήστης τρέχει κάποιες βιβλιοθήκες σε ένα προγραμματιστικό περιβάλλον όπως είναι η γλώσσα προγραμματισμού Python, τότε είναι έργο αυτών των βιβλιοθηκών να τρέχουν τις εξισώσεις όπως για παράδειγμα των παραμέτρων β και του τυχαίου σφάλματος ε πίσω στο background, έτσι διευκολύνετε το έργο του χρήστη.

Παρακάτω βλέπουμε ένα κομμάτι κώδικα

```
# Usage: spark-submit linreg.py <inputdatafile>
# Example usage: spark-submit linreg.py yxlin.csv

# import the necessary packages
from pyspark import SparkContext
import sys
import numpy as np
from numpy.linalg import inv

# function to get the y points from the file
def parsePointY(line):
        values = [float(x) for x in line.split(',')]
        return (values[0])

# function to get the x points from the file
def parsePointX(line):
        values = [float(x) for x in line.split(',')]
        return (values[1:])

def linearreg(x_lines,y_lines):
    Key_A = x_lines.map(lambda a:("KeyA",(np.array(a).reshape(len(a),1) * np.transpose(np.array(a).astype('float')).reshape(1,len(a)))))
    A=Key_A.reduceByKey(lambda x,y : x+y).values()
    Key_B = y_lines.map(lambda b:("KeyB",(np.array(b)[1:len(b)].reshape(len(b)-1,1) * np.array(b)[0].reshape(1,1))))
    B=Key_B.reduceByKey(lambda x,y : x+y).values()
    A= np.array(A.collect()).squeeze()
    B= np.array(B.collect())
    A_inverse = inv(A)
    weight = np.dot(A_inverse,B)
    return weight

# checking for the proper input arguments if two arguments are not passed exit the code
if len(sys.argv) !=2:
    print >> sys.stderr, "Usage: linreg <datafile>"
    exit(-1)
```

Εικόνα 24: Κώδικας γραμμικής παλινδρόμησης σε Python

Όπως βλέπουμε αρχικά καθορίζουμε της διάφορες βιβλιοθήκες μας ενώ στην συνέχεια με το line.split γραμμή γραμμή στον x,y κάνουμε διαχωρισμό των δεδομένων του αρχείου μας.

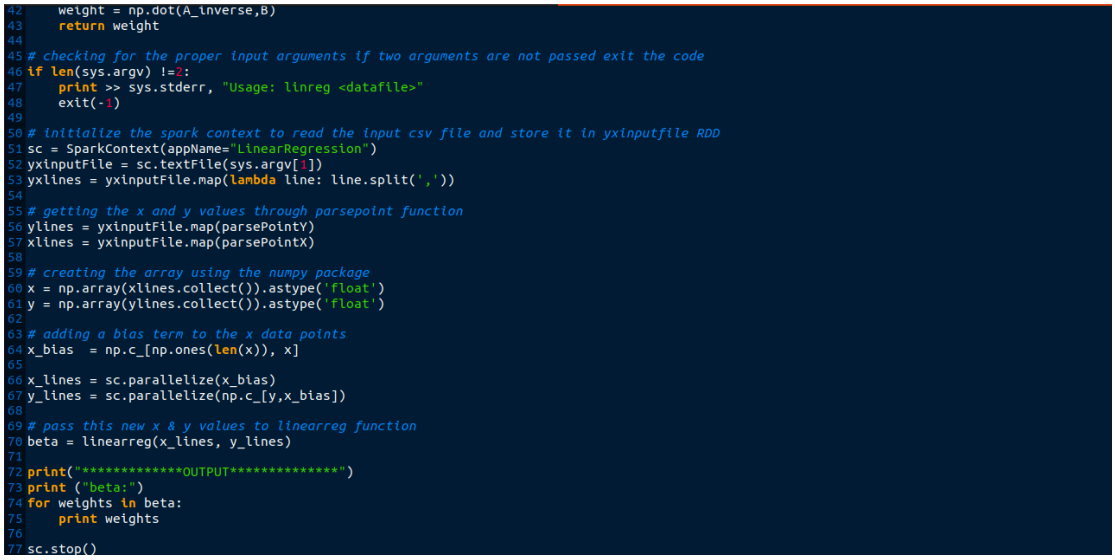
Στην συνέχεια χρησιμοποιούμε την λειτουργία (function), μια χρήσιμη λειτουργία θα λέγαμε αφού μας δίνει την δυνατότητα της δημιουργίας μιας ανώνυμης συνάρτησης την οποία θα τρέξουμε στην συνέχεια.

ReduceByKey: Η συνάρτηση η οποία συγχωνεύει κλειδί με τιμή σε ένα RDD σύστημα που χρησιμοποιεί το apache spark όπως έχουμε και προαναφέρει στην εργασία μας, δηλαδή για μια πλειάδα στοιχείων με ζεύγη κλειδιών και τιμών.

B= np.array.collect (): επρόκειτο για την συλλογή μιας πλειάδας τιμών μη αρνητικών ακέραιων αριθμών.

A_inverse = inv(A): υπολογισμός του αντίστροφου του matrix.

Np.dot: μια χρήσιμη λειτουργία για την δημιουργία δύο πινάκων.



```
    weight = np.dot(A_inverse,B)
    return weight

# checking for the proper input arguments if two arguments are not passed exit the code
if len(sys.argv) !=2:
    print >> sys.stderr, "Usage: linreg <datafile>"
    exit(-1)

# initialize the spark context to read the input csv file and store it in yxinputfile RDD
sc = SparkContext(appName="LinearRegression")
yxinputFile = sc.textFile(sys.argv[1])
yxlines = yxinputFile.map(lambda line: line.split(','))

# getting the x and y values through parsepoint function
ylines = yxinputFile.map(parsePointY)
xlines = yxinputFile.map(parsePointX)

# creating the array using the numpy package
x = np.array(xlines.collect()).astype('float')
y = np.array(ylines.collect()).astype('float')

# adding a bias term to the x data points
x_bias  = np.c_[np.ones(len(x)), x]

x_lines = sc.parallelize(x_bias)
y_lines = sc.parallelize(np.c_[y,x_bias])

# pass this new x & y values to linearreg function
beta = linearreg(x_lines, y_lines)

print("*************OUTPUT**************")
print ("beta:")
for weights in beta:
    print weights

sc.stop()
```

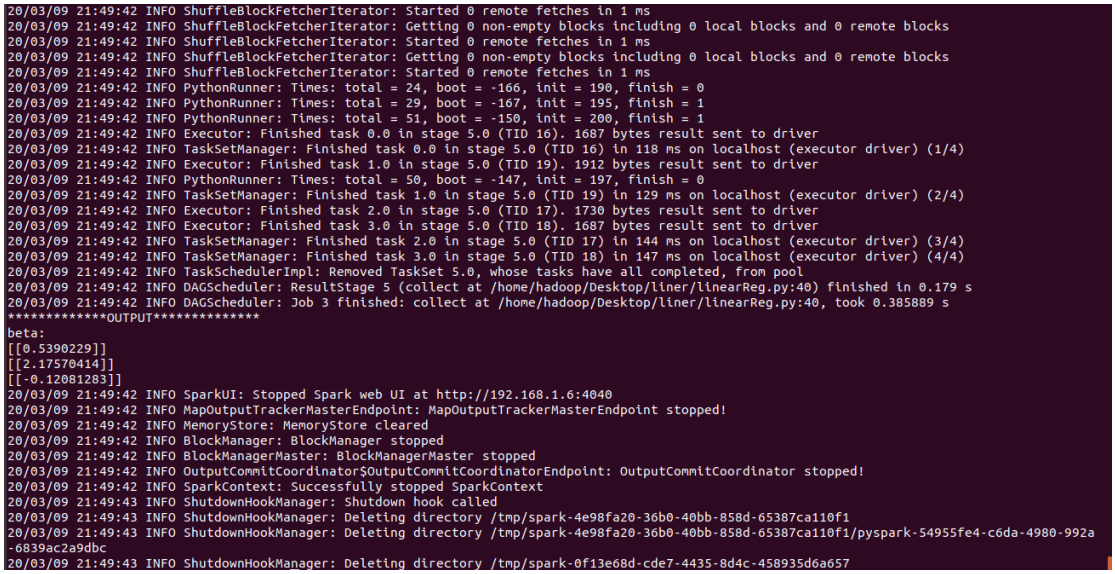
Εικόνα 25: Απεικόνιση για είσοδο δεδομένων μέσω κώδικα και επιπλέον απαραίτητων παραμέτρων

Spark Context: πρόκειται για την πιο χρήσιμη γραμμή κώδικα αφού επρόκειτο για ένα πρόγραμμα οδήγησης το οποίο τρέχει από το Spark πάντα και εκτελεί της λειτουργίες εντός των κόμβων.

Στην συνέχεια εισάγουμε τις μεταβλητές μας από το αρχείο δεδομένων που κάναμε πριν load στην συνάρτηση parsePointY και parsePointX την οποία καθορίσαμε στην αρχή του προγράμματος μας, και παράλληλα μέσω του np.array κατασκευάζονται οι δύο πίνακες μας που στην συνέχεια παραλληλίζονται και μεταφράζονται πάνω στον άξονα X και Y.

Θα μπορούσαμε να πούμε ότι ο κώδικας μας χωρίζεται σε δύο μέρη πρώτα στο να καθορίσουμε την συνάρτηση μας μέσω της lambda και στην συνέχεια μέσω της ReduceBykey να καθορίσουμε τα ζεύγη κλειδιών τιμών.

Τέλος μέσα από την λειτουργία (function) linearreg εκτυπώνουμε τα αποτελέσματα μας όπως βλέπουμε παρακάτω:



Εικόνα 26: Αποτελέσματα της εκτέλεσης

Συμπερασματικά θα λέγαμε η γλώσσα python είναι απλή προγραμματιστικά και με την δυνατότητα της πληθώρας βιβλιοθηκών σε συνδυασμό με το framework Apache Spark μπορεί κάποιος να πειραματιστεί και να βελτιώσει τον τομέα της εργασίας του η ακόμα και μιας επιχείρησης.

ΣΥΜΠΕΡΑΣΜΑΤΑ – ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

Όπως παρατηρούμε μέσω σύγχρονων εργαλείων μπορούμε να πάρουμε σημαντικά συμπεράσματα σχετικά με τιμές που εξάγουμε από τα δεδομένα που λαμβάνουμε, τα χαρακτηριστικά την μορφοποίηση τους ανάλογα με τις αλλαγές που μπορεί να γίνονται από εξωγενείς παράγοντες, τα διάφορα σφάλματα που μας δείχνουν σε πιο βαθμό μπορεί να έχουμε κάποια αποτυχία έτσι ώστε να μπορούμε να παρεμβαίνουμε και να διορθώνουμε οποιοδήποτε σφάλμα. Έτσι μας δίνεται η δυνατότητα στον τομέα της κατανάλωσης της ενέργειας να έχουμε αρκετά οφέλη, αφού σωστή χρήση των έξυπνων συσκευών μέσου της ανάλυσης των δεδομένων δίνει την δυνατότητα κέρδους προς τον καταναλωτή – σχέση κατανάλωση, ενώ με την εξοικονόμηση της ενέργειας επιτυγχάνουμε στην μείωση της ρύπανσης του περιβάλλοντος.

Μια άλλη παράμετρος που είναι επίσης σημαντική για τον καταναλωτή είναι η χρήση των συσκευών αν γίνεται με σωστό χειρισμό αφού πολλές φορές ο ανθρώπινος παράγοντας μπορεί να δημιουργεί σφάλματα, όπως για παράδειγμα η θερμοκρασία μιας κατοικίας μέσω ενός λέβητα για παράδειγμα. Αφού μέσω της σωστής ανάλυσης των δεδομένων μπορούμε να πάρουμε την μέση θερμοκρασία ενός σπιτιού και από εκεί να συμπεράνουμε αν όντως είναι η επιθυμητή θερμοκρασία, αν είναι η απαραίτητη θερμοκρασία για την καλή υγεία του ανθρώπου αλλά και πως να μειώσουμε την κατανάλωση ενέργειας ενός λέβητα με την σωστή χρήση του κατά την διάρκεια της ημέρας.

Ένας παράγοντας ακόμα της ανάλυσης δεδομένων που μπορεί να μας δώσει στοιχεία ως προς την κατανάλωση ενέργειας αλλά και την χρήση συσκευών είναι η ανίχνευση ανωμαλιών η σφαλμάτων αλλιώς. Για παράδειγμα μέσα από την αστοχία μιας συσκευής μπορούμε να δούμε ενδείξεις στα δεδομένα τα οποία δείχνουν την λανθασμένη λειτουργία της συσκευής η την αλλαγή συμπεριφοράς ως προς την κατανάλωση ενέργειας.

Οι μελλοντικές μας κατευθύνσεις θα ήταν να γίνει η ενοποίηση όλων αυτών των προγραμμάτων σε ένα το οποίο θα είναι εύχρηστο ως προς τον χρήστη

με ένα φιλικό περιβάλλον και με την δυνατότητα της γραφικής απεικόνισης, έτσι ώστε να δίνει την δυνατότητα σε νέους επαγγελματίες «Μηχανικός Ενέργειας» να παίρνουν γρήγορα και αποδοτικά αποτελέσματα.

ΑΝΑΦΟΡΕΣ

[1] .Benjamin Bengfort and Jenny Kim (Data Analytics with Hadoop) an Introduction for Data Scientist.

[2] https://www.whishworks.com/blog/big-data/understanding-the-3-vs-of-big-data-volume-velocity-and-variety

[3] https://en.wikipedia.org/wiki/Smart_grid

[4] http://www.wisegeek.net/what-is-data-efficiency.htm

[5] https://www.laserfiche.com/ecmblog/why-data-efficiency-next-big-phrase-it/

[6] https://en.wikipedia.org/wiki/Analytics

[7] https://www.guru99.com/big-data-analytics-tools.html

[8] https://spark.apache.org/

[9] https://www.edureka.co/blog/pyspark-tutorial/#pyspark

[10] https://www.tutorialspoint.com/pyspark/pyspark_sparkconf.htm

[11] https://en.wikipedia.org/wiki/Algorithm

[12] https://repository.kallipos.gr/

[13] Ευστάθιος Ζάχος Αριστείδης Παγουρτζής Δώρα Σούλιου (Θεμελίωση Επιστήμης Υπολογιστών) Εθνικό Μετσόβιο Πολυτεχνείο.

[14] Varun Chandola, Arindam Banerjee, and Vipin Kumar, Anomaly Detection:A Survey, 2009.

[15] Γεράσιμος Ε. Σταυλιώτης., Εξόρυξη Δεδομένων (DATA MINING) και Αναγνώριση προτύπων σε κατηγορικά δεδομένα μέσω συσταδοποίηση, 2009.

[16] Knuth 1998 , §6.2.1 ("Αναζήτηση πίνακα με παραγγελίες"), υποδιαίρεση "Αλγόριθμος B".

[17] C.A.R. Hoare. Algorithm 63 - Partition and algorithm 65 - Find.Communications of the ACM, 4:321–322, 1961.

[18 M. Blum, R.W. Floyd, V. Pratt, R.L. Lewis, and R.E. Tarjan. Time bounds forselection. Journal of Computer and System Sciences, 7:448–461, 1973.

[19] J. Boothroyd. Algorithm201 - Shellsort. Communications of the ACM, 6(8):445, 1963.

[20] C. Bron. Algorithm 426 - Merge sort algorithm. Communications of the ACM,15(5):357–358, 1972.

[21] Davies L., Gather U., ”The identification of multiple outliers,” Journal of the American Statistical Association, 88(423), 782-792, 1993.

[22] Hawkins D., Identification of Outliers, Chapman and Hall, 1980.

[23] Hampel F. R., ”A general qualitative definition of robustness,” Annals of Mathematics Statistics, 42, 1887–1896, 1971.

[24] Ben-Gal I., Morag G., Shmilovici A., ”CSPC: A Monitoring Procedure for State Dependent Processes,” Technometrics, 45(4), 293-311, 2003.

[25] Penny K. I., Jolliffe I. T., ”A comparison of multivariate outlier detection methods for clinical laboratory safety data,” The Statistician 50(3), 295-308, 2001.

[26] Rousseeuw P., Leory A., Robust Regression and Outlier Detection, Wiley Series in Probability and Statistics, 1987.

[27] Knorr E., Ng R., ”A unified approach for mining outliers,” In Proceedings Knowledge Discovery KDD, 219-222, 1997.

[28] http://rvlasveld.github.io/blog/2013/07/12/introduction-to-one-class-support-vector-machines/.

[29] https://www.codecademy.com/articles/scikit-learn

[30] https://www.quora.com/In-Python-what-is-NumPy-How-is-it-used

[31] http://www.unipi.gr/faculty/mbouts/statprog/SPSS_lesson9-10.pdf

[32] Cohen, J., Cohen P., West, S.G., & Aiken, L.S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences. (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates

[33] *Draper, N.R.; Smith, H. (1998). Applied Regression Analysis (3rd έκδοση). John Wiley.* ISBN 0-471-17082-8*.*

[34] https://en.wikipedia.org/wiki/Smart_meter

[35] https://en.wikipedia.org/wiki/Smart_home_technology

[36] https://www.ni.com/en-us/innovations/white-papers/06/labview-for-measurement-and-data-analysis.html

[37] https://www.mathworks.com/solutions/data-analysis.html

[38] https://www.coursera.org/learn/excel-data-analysis