Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακή Διατριβή
Στην Ασφάλεια Υπολογιστών και Δικτύων

Επιλογή Χαρακτηριστικών Δικτυακής Κίνησης και Ανίχνευση Εισβολών με χρήση του Microsoft Azure Machine Learning Studio

Θωμάς Αθανασίου

Επιβλέπων Καθηγητής
Δρ. Ιωάννης Μαυρίδης

Ιανουάριος 2018

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Επιλογή Χαρακτηριστικών Δικτυακής Κίνησης και Ανίχνευση Εισβολών με χρήση του Microsoft Azure Machine Learning Studio

Θωμάς Αθανασίου

Επιβλέπων Καθηγητής
Δρ. Ιωάννης Μαυρίδης

Η παρούσα μεταπτυχιακή διατριβή υποβλήθηκε
προς μερική εκπλήρωση των απαιτήσεων για απόκτηση

μεταπτυχιακού τίτλου σπουδών
στην Ασφάλεια Υπολογιστών και Δικτύων

από τη Σχολή Θετικών και Εφαρμοσμένων Επιστημών
του Ανοικτού Πανεπιστημίου Κύπρου

Ιανουάριος 2018

Περίληψη

Σκοπός της έρευνας είναι η εισαγωγή της αυτοματοποιημένης πλατφόρμας μηχανικής μάθησης Microsoft Azure Machine Learning Studio στη Κυβερνοασφάλεια με στόχο την εκμετάλλευση των υπολογιστικών πόρων της, τη χρήση των προσφερόμενων και παραμετροποιήσιμων αλγόριθμων μηχανικής μάθησης, καθώς και των δυνατοτήτων οπτικοποίησης των αποτελεσμάτων. Το πλήθος των διαθέσιμων συνόλων δεδομένων δικτυακής κίνησης (Network Traffic Datasets) απαιτεί ισχυρή προεπεξεργασία ώστε να μπορεί ο ερευνητής ασφάλειας να τα χρησιμοποιήσει προς την κατεύθυνση της ανίχνευσης εισβολών σε δικτυακό περιβάλλον. Η πλατφόρμα Microsoft Azure Machine Learning Studio συγκεντρώνει χαρακτηριστικά βελτιστοποίησης και επιτάχυνσης της ερευνητικής μελέτης ακολουθώντας μια καλά ορισμένη μεθοδολογία. Στην παρούσα μεταπτυχιακή διατριβή εξερευνούμε τα διαθέσιμα σύνολα δεδομένων δικτυακής κίνησης, τα οποία προεπεξεργάζονται κατάλληλα επιλέγοντας τα σημαντικότερα χαρακτηριστικά τους(Feature Selection). Στη συνέχεια εισάγονται ως είσοδος σε αλγόριθμο μηχανικής μάθησης με σκοπό την ανίχνευση ανωμαλιών (Anomaly Detection). Ο αλγόριθμος PCA (Principal Component Analysis) χρησιμοποιείται στο στάδιο της προεπεξεργασίας των δεδομένων καθώς και στο τελικό στάδιο της εξαγωγής συμπερασμάτων. Η παρουσίαση των αποτελεσμάτων πραγματοποιείται με εκτενή χρήση διαγραμματικών εργαλείων απεικόνισης που προσφέρει η πλατφόρμα.

Summary

The aim of this research is to introduce Microsoft Azure Machine Learning studio into Cyber Security to exploit its computing resources, and to also use customized machine learning algorithms and visualization of results. Available network traffic datasets require strong pre-processing so that the security researcher can use them efficiently to detect intrusion efforts into a corporate network. Microsoft Azure Machine Learning Platform brings together optimization and accelerating features to the research study, by following a well-defined methodology and providing appropriate tools. In this thesis, we explore the available network traffic data sets, select and pre-process them by selecting their most important features (Feature Selection). They are then introduced as an input to a machine learning algorithm for the detection of network traffic anomalies (Anomaly Detection). The Principal Component Analysis (PCA) algorithm is used at the pre-processing stage and at the final stage of the outcome. The results are presented with extensive use of graphical imaging tools offered by the Microsoft Azure platform.

Ευχαριστίες

Ευχαριστώ τον επιβλέποντα καθηγητή μου Δρ Ιωάννη Μαυρίδη, για τις πολύτιμες παρατηρήσεις και συμβουλές του. Επίσης, ευχαριστώ τον Νικόλαο Τσίγγαννο για τη βοήθεια και συμβολή του. Ευχαριστώ τη σύζυγό μου Έλενα, για την υποστήριξη και συμπαράστασή της. Τέλος, ευχαριστώ την κόρη μου Ίριδα, που ήρθε στη ζωή μας.

Περιεχόμενα

Κεφάλαιο 1
Καταγραφή Δικτυακής Κίνησης

1.1 Γενικά

Στο κεφάλαιο αυτό, γίνεται αναφορά στις δύο μεγάλες φάσεις των δικτυακών εισβολών. Συγκεκριμένα γίνεται λόγος για τις δύο μεγάλες κατηγορίες συλλογής δεδομένων, την ιχνηλάτηση (foot printing) και τη σάρωση (scanning), που απαιτούνται στη φάση της προετοιμασίας επιθέσεων από επίδοξους εισβολείς (hacker) [13],[16]. Η προετοιμασία για μία επίθεση είναι ίσως το πιο σημαντικό κομμάτι στις δικτυακές επιθέσεις και αυτό διότι συγκεντρώνονται πολύτιμες πληροφορίες για το αντικείμενο της επίθεσης.

Επιπλέον, στο κεφάλαιο αυτό πραγματοποιείται μία έρευνα που αφορά στην απόκτηση δεδομένων για δικτυακές εισβολές. Από τη μία το πρώτο αντικείμενο είναι η αναζήτηση για τα διαθέσιμα σύνολα δεδομένων δικτυακής κίνησης (datasets). Από την άλλη το αντικείμενο είναι οι τρόποι που μπορούν να αξιοποιηθούν τα διάφορα αυτά δεδομένα ώστε να αποκτηθεί ένα αξιόπιστο σύνολο, που θα χρησιμοποιηθεί στο πείραμά μας και θα αφορά δεδομένα για ανίχνευση δικτυακών εισβολών.

1.2 Η διαδικασία συλλογής πληροφοριών επτά βημάτων

Το International Council of Electronic Commerce Consultants (EC- Council), το οποίο εξειδικεύεται σε διαδικασίες ελέγχου τρωτότητας (penetration testing), προσδιορίζει τη διαδικασία συλλογής πληροφοριών σε επτά βασικά βήματα [16],[12]. Οι πληροφορίες αυτές, αφορούν δεδομένα τα οποία είναι απαραίτητα σε δικτυακές επιθέσεις.

Συλλογή δεδομένων

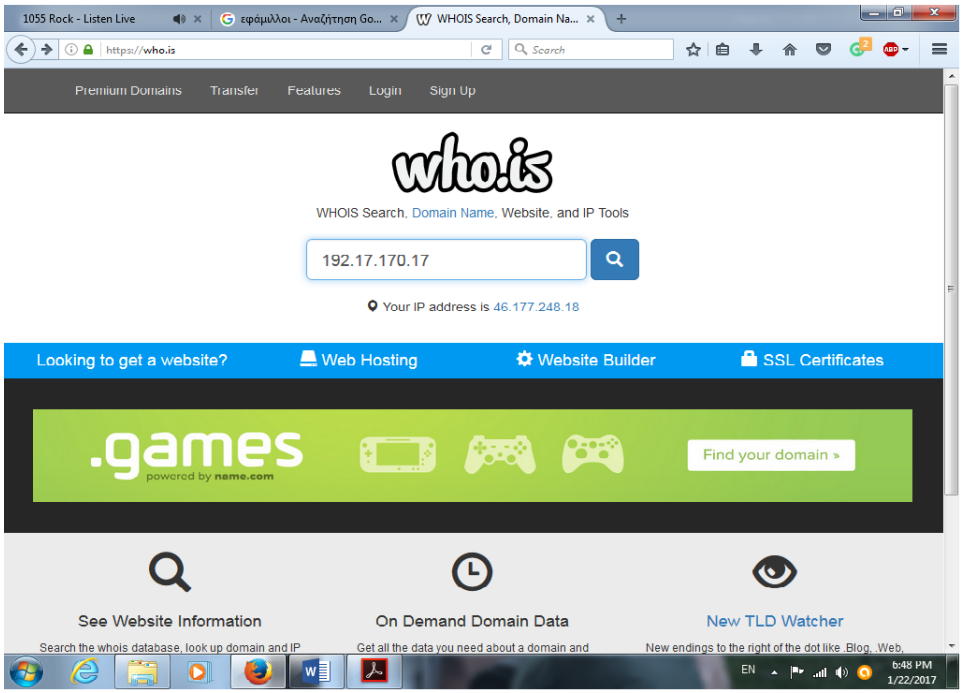
Το πρώτο και βασικό βήμα, είναι το να συλλεχθούν όλα τα απαραίτητα στοιχεία (information gathering), που θα προσδιορίσουν καλύτερα τον οργανισμό και θα καθορίσουν τον τρόπο για τις μετέπειτα επιθέσεις. Αυτό σημαίνει πως όσο περισσότερες και καλύτερες πληροφορίες μπορέσουν να εκμαιευθούν, τόσο πιθανότερο είναι να έχει αποδοτικά αποτελέσματα μία ηλεκτρονική επίθεση ή ένα τεστ διείσδυσης (penetration test).

Υπάρχει πλήθος πληροφοριών, που μπορούν να βρεθούν χωρίς κάποια ιδιαίτερη τεχνική, αλλά με σωστή οργάνωση. Οι πληροφορίες αυτές μπορούν να υπάρχουν σε ιστοσελίδες, σε νέα από εφημερίδες, γενικά σε συζητήσεις και σε αναρτήσεις στο Usenet, αλλά και από τους ίδιους τους υπαλλήλους.

Καθορισμός δικτυακής εμβέλειας

Προχωρώντας στο δεύτερο βήμα της διαδικασίας και θεωρώντας πως έχουν συλλεχθεί αρκετές πληροφορίες, προσπαθούμε να καθορίσουμε το εύρος του δικτύου που θα γίνει η επίθεση (determining the network range). Σε αυτό το στάδιο βρίσκουμε την αρχιτεκτονική του δικτύου σε ότι αφορά το πώς έχουν οριστεί οι διευθύνσεις. Συνεπώς έχουμε να κάνουμε με το εύρος των διευθύνσεων δικτύου (ip addresses) και την απαρίθμησή τους (enumeration). Μία απλή τεχνική για να έχουμε κάποιες βασικές πληροφορίες σχετικά με το στόχο της επίθεσης, είναι ένα ερώτημα (who is), σε ένα δικτυακό εργαλείο επίλυσης ονομάτων (online Dns Resolver). Παρακάτω φαίνονται τα

δεδομένα που έχουμε ως απάντηση σε ερώτημα (who is), για κάποιο όνομα δικτυακού τόπου (Domain Name) ή διεύθυνση δικτύου (Ip address).



Εικόνα 1.2 Αναζήτηση whois

Τα δεδομένα που λαμβάνουμε είναι του τύπου:

NetRange: **192.17.0.0 - 192.17.255.255**
CIDR: 192.17.0.0/16
NetName: UIUC-CLASS-C
NetHandle: NET-192-17-0-0-1
Parent: NET192 (NET-192-0-0-0-0)
NetType: Direct Allocation
OriginAS: AS38
Organization: University of Illinois (UIUC)
RegDate: 1995-01-06
Updated: 2014-12-02
Ref: https://whois.arin.net/rest/net/NET-192-17-0-0-1

OrgName: University of Illinois
OrgId: UIUC
Address: 1120 DCL, MC-256

Address:	1304 West Springfield Avenue
City:	Urbana
StateProv:	IL
PostalCode:	61801
Country:	US
RegDate:	
Updated:	2014-12-02
Ref:	https://whois.arin.net/rest/org/UIUC

OrgAbuseHandle:	UIUCS-ARIN
OrgAbuseName:	UIUC Security
OrgAbusePhone:	+1-217-265-0000
OrgAbuseEmail:	abuse@uiuc.edu
OrgAbuseRef:	https://whois.arin.net/rest/poc/UIUCS-ARIN

OrgTechHandle:	HOSTM159-ARIN
OrgTechName:	Hostmgr
OrgTechPhone:	+1-217-244-1000
OrgTechEmail:	dns-admin@illinois.edu
OrgTechRef:	https://whois.arin.net/rest/poc/HOSTM159-ARIN

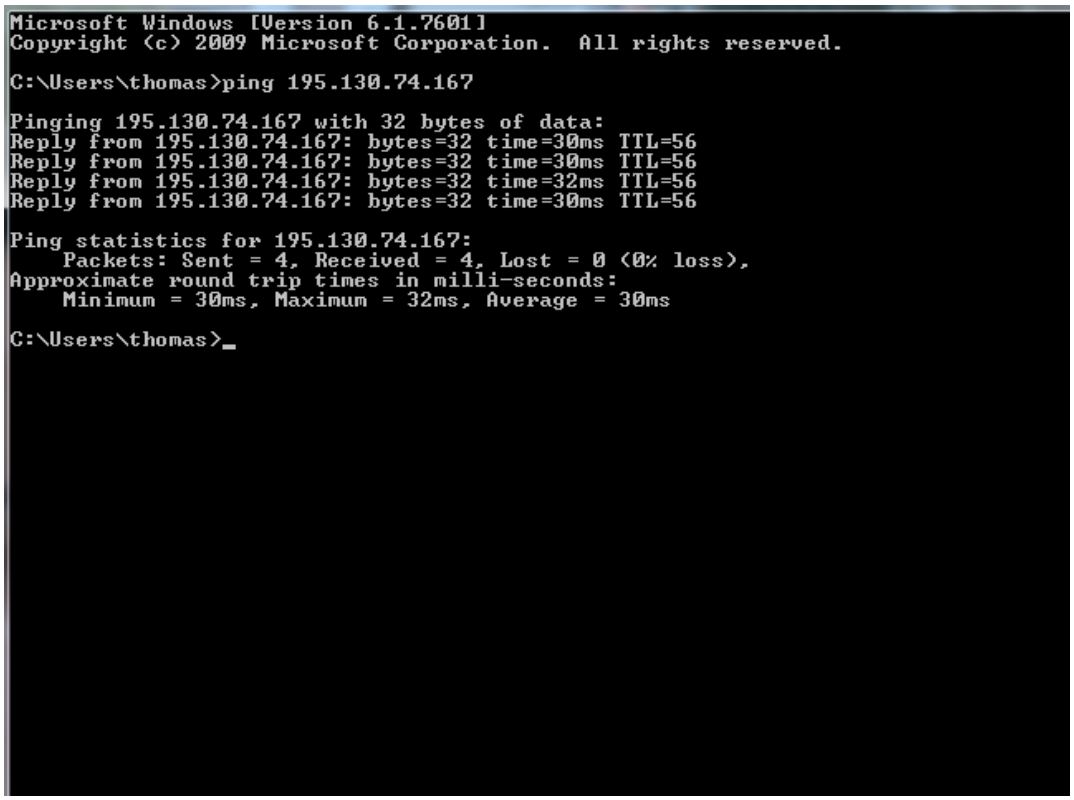
Πλέον γνωρίζουμε το εύρος των διευθύνσεων και μπορούμε να πειραματιστούμε με τα τερματικά.

Εντοπισμός ενεργών μηχανών

Αφού γίνει η 'χαρτογράφηση' του δικτύου και γνωρίζουμε πλέον τις διευθύνσεις των τερματικών, το επόμενο βήμα είναι να δούμε πόσες ή ποιες από αυτές είναι ενεργές (Identifying active machines). Η πληροφορία αυτή μας είναι διαθέσιμη μετά από την εξέταση των αιτημάτων που απαντώνται από τους διαθέσιμους hosts.

Κατά συνέπεια σε αυτή τη διαδικασία γίνονται διαδοχικά αιτήματα **echo request** και **echo reply**. Αυτό σημαίνει ότι αιτούμαστε επικοινωνία προς το συγκεκριμένο τερματικό, με τη διεύθυνση που το χαρακτηρίζει και περιμένουμε τα δεδομένα της απάντησής του.

Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα χρήσης του εργαλείου ping (Εικόνα 1.1).



Εικόνα 1.1 Χρήση εργαλείου Ping

Τα δεδομένα που αποκτούνται σε αυτό το βήμα τοποθετούν στο στόχαστρο συγκεκριμένες μηχανές(τερματικά), που είναι σε θέση να απαντήσουν στα αιτήματά μας. Από εκεί και έπειτα μπορούν με κάποιες τεχνικές να εισαχθούν δικά μας στοιχεία ώστε να δημιουργήσουμε μια πόρτα επικοινωνίας με το συγκεκριμένο τερματικό.

Εύρεση για ανοιχτές πόρτες και σημεία πρόσβασης

Έχοντας πλέον ορίσει το εύρος του δικτύου και τις ενεργές μηχανές από τα προηγούμενα βήματα, ο επόμενος στόχος είναι να βρεθούν ‘ανοιχτές πόρτες‘ και σημεία πρόσβασης (Finding open ports and access points). Αυτό στην ουσία είναι το λεγόμενο **Port scanning**, δηλαδή η σάρωση για το ποιες υπηρεσίες και διαδικασίες εκτελούνται και σε ποιες συγκεκριμένες θύρες (ports).

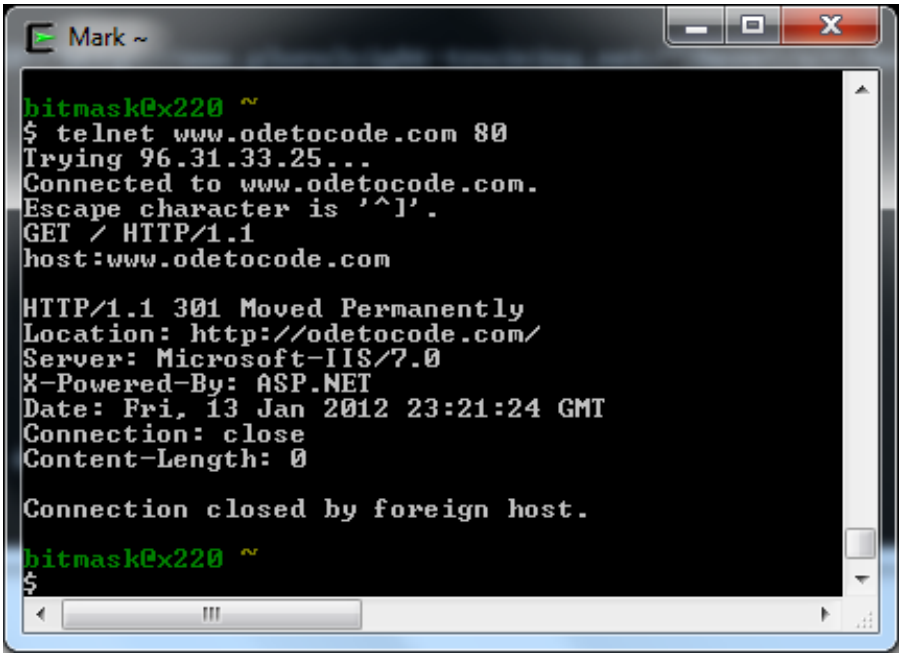
Αποτύπωμα λειτουργικού συστήματος

Στο σημείο αυτό και ενώ έχουν συλλεχθεί πληροφορίες που αφορούν ενεργές μηχανές, διευθύνσεις και πόρτες που εκτελούν συγκεκριμένες υπηρεσίες, η έρευνα επικεντρώνεται στην αναγνώριση των μηχανών που έχουν επιλεγεί (OS fingerprinting).

Για να προχωρήσουμε σε αυτή τη διαδικασία του fingerprinting, υπάρχουν δύο μέθοδοι. Η πρώτη μέθοδος είναι να γίνει παθητικά το λεγόμενο **passive fingerprinting**, ενώ η δεύτερη είναι να γίνει ενεργητικά **active fingerprinting** [12].

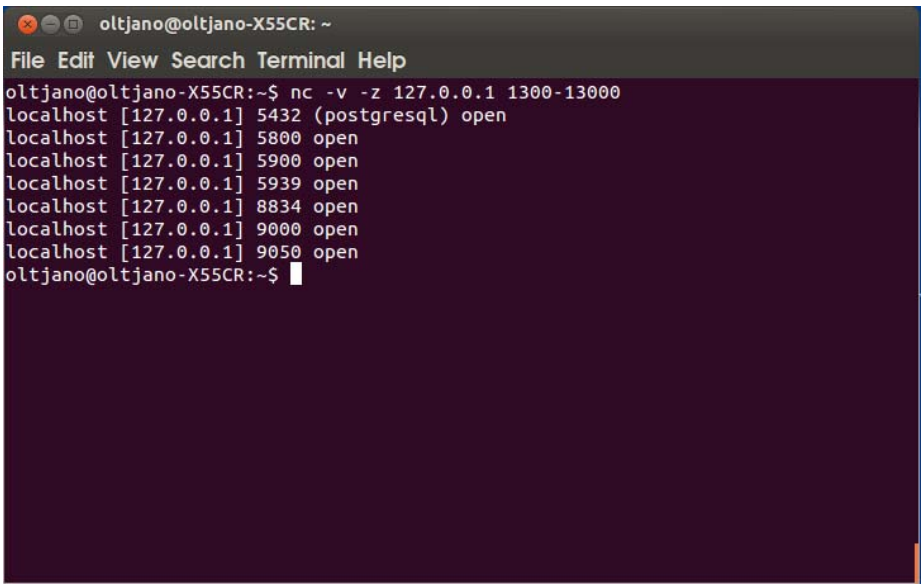
Αποτύπωμα υπηρεσιών

Για την περαιτέρω ενίσχυση των υποψιών και πληροφοριών σε ότι αφορά το λειτουργικό σύστημα που χρησιμοποιούν οι διακομιστές (hosts) που έχουν γίνει στόχος, χρησιμοποιούνται κάποια επιπλέον εργαλεία και υπηρεσίες (Fingerprinting services). Η χρήση υπηρεσιών όπως του telnet και εργαλείων όπως το NetCat , μπορούν να μας παρέχουν πολύτιμες πληροφορίες.



Εικόνα 1.4 Παράδειγμα χρήσης Telnet

Όπου λαμβάνουμε πληροφορία για το είδος http ή https του web server, Internet Information Services (IIS).



Εικόνα 1.5 Παράδειγμα χρήσης NetCat

Όπου λαμβάνουμε πληροφορίες για γνωστές υπηρεσίες, όπως VPN(Virtual Private Network) στην ανοιχτή θύρα 5900 .

Χαρτογράφηση δικτύου

Η συλλογή όλων των παραπάνω δεδομένων και η διασταύρωση των πληροφοριών οργανωμένα σε ένα είδος πίνακα αποτελεί τη χαρτογράφηση του δικτύου (Mapping the network). Αυτό σημαίνει ότι πλέον στη διάθεση του επίδοξου υποκλοπέα υπάρχουν στοιχεία που αφορούν σε: ip διευθύνσεις, εύρος δικτύου, access points, ανοιχτές ports, δεδομένα λειτουργικού συστήματος.

1.3 Σύνολα δικτυακής κίνησης

Μέχρι τώρα είδαμε τον τρόπο που μπορούν να συλλεχθούν δεδομένα από έναν επίδοξο υποκλοπέα. Αντίστοιχα οι ίδιοι τύποι δεδομένων μπορούν να χρησιμοποιηθούν για την εκπαίδευση μοντέλων που σκοπό έχουν να αντιμετωπίσουν τέτοιου είδους επιθέσεις. Η χρήση τέτοιων δεδομένων από πραγματικές επιθέσεις ή προσομοιώσεις τους, είναι απαραίτητη στη διαδικασία της εκπαίδευσης και τα σύνολα αυτά αναφέρονται στην βιβλιογραφία ως DataSets.

Για να μπορούν να εξαχθούν χρήσιμα συμπεράσματα και για να γίνει σωστός σχεδιασμός και εκπαίδευση ενός μοντέλου, θα χρειαστεί ένα καλά ορισμένο και επαρκές σύνολο δεδομένων.

Η πληροφόρηση επί του συνόλου που θα επιλέξουμε είναι πολύ σημαντική. Θα πρέπει λοιπόν το επιλεχθέν σύνολο, να διαθέτει χρήσιμες πληροφορίες ως ετικέτες στις στήλες, όπως επίσης και να συνοδεύεται από ένα έγγραφο περιγραφής των πληροφοριών αυτών. Η ύπαρξη εγγράφου περιγραφής του συνόλου είναι πολύ σημαντική, διότι θα έχουμε μία πρώτη εικόνα του λόγου που επιλέχθηκαν τα συγκεκριμένα χαρακτηριστικά στο σύνολο και για το αν εξυπηρετούν το σκοπό μας.

Το δεύτερο στοιχείο που πρέπει να χαρακτηρίζει ένα σύνολο δεδομένων για να θεωρηθεί αξιόπιστο είναι ο όγκος του. Αυτό σημαίνει ότι θα πρέπει να είναι αρκετά μεγάλο για να μπορεί να θεωρηθεί ότι η εκπαίδευση έγινε με αρκετούς συνδυασμούς στοιχείων και καλύφθηκαν αρκετά πιθανά σενάρια. Από την άλλη ο όγκος των δεδομένων δεν θα πρέπει να είναι τόσο μεγάλος που να μην είναι διαχειρίσιμος. Δύο ακόμη στοιχεία που θα πρέπει να συμπεριληφθούν στην αναζήτηση του κατάλληλου συνόλου, είναι αυτά της αξιοπιστίας και της εγκυρότητας.

Από την μία πλευρά, ιδανικά αξιόπιστο θα είναι ένα σύνολο το οποίο περιγράφει πραγματικά δεδομένα δικτυακής κίνησης που έχουν συλλεχθεί από οργανισμούς και εταιρίες. Από την άλλη πλευρά η εγκυρότητα του συνόλου θα επιβεβαιωνόταν εφόσον το σύνολο αυτό έχει ελεγχθεί και χρησιμοποιηθεί σε προηγούμενες έρευνες για ανάλογους σκοπούς.

Συνοψίζοντας η έρευνα επι συνόλων διαδικτυακών δεδομένων θα πρέπει να κινηθεί στις εξής κατευθύνσεις.

- **Καλά ορισμένο σύνολο** (περιλαμβάνει στήλες με δεδομένα που θα εξυπηρετούν το σκοπό μας).
- **Διαθέσιμη πληροφόρηση συνόλου** (περιγραφή των δεδομένων και του σκοπού για τον οποίο επιλέχθηκαν).
- **Επαρκής όγκος** (ιδανικός όγκος πληροφορίας για εκμάθηση του μοντέλου αλλά και διαχειρίσιμος).
- **Αξιοπιστία** (χρήση πραγματικών δεδομένων η τουλάχιστον καλή προσομοίωση).

- **Εγκυρότητα** (χρήση και έλεγχος του συνόλου σε προγενέστερες έρευνες).

1.4 Εξερεύνηση διαθέσιμων Συνόλων

Στα πλαίσια αυτής της διατριβής, ένα πολύ σημαντικό μέρος αποτελεί η εξερεύνηση διαθέσιμων συνόλων δικτυακής κίνησης(network datasets). Όπως έχουμε πει το σύνολο που θα επιλεχθεί θα πρέπει να τηρεί τις προϋποθέσεις που θέσαμε παραπάνω.

Η αναζήτηση συνόλου δεδομένων δικτυακής κίνησης, δεν είναι μία εύκολη υπόθεση. Υπάρχει ένας μεγάλος όγκος από δεδομένα που είναι διαθέσιμα στο δίκτυο. Κάποια είναι ελεύθερα προς χρήση και κάποια όχι (χρειάζονται έγκριση από τις εταιρίες και τους οργανισμούς στους οποίους ανήκουν). Όταν μιλάμε για δεδομένα δικτυακής κίνησης αναφερόμαστε σε αρχεία καταγραφής κίνησης από διάφορα Intrusion Detection Systems (IDS), τα γνωστά log files ή Pcap files. Τέτοιου είδους αρχεία δημιουργούνται για την παρακολούθηση των κινήσεων σε ένα διακομιστή (server). Αυτό σημαίνει ότι σε ένα υποψήφιο dataset μπορεί να περιέχονται αρχεία που έχουν να κάνουν για παράδειγμα μόνο με φυσιολογική λειτουργία ενός συστήματος. Επίσης τα διαθέσιμα σύνολα μπορούν να αφορούν δεδομένα από συγκεκριμένες επιθέσεις, όπως για παράδειγμα DDOS attacks ή UDP flood. Όπως είναι προφανές ανάλογα με τον τύπο της επίθεσης, τα χαρακτηριστικά που χρειάζεται να μελετηθούν στα datasets, αλλάζουν. Συνεπώς, σε περίπτωση που επιλέξουμε να χρησιμοποιήσουμε αρχεία από διαφορετικές καταγραφές για να συνδυάσουμε διαφορετικούς τύπους επιθέσεων, θα έχουμε διαφορετικά αρχεία ως προς τα διαθέσιμα χαρακτηριστικά τους. Με απλά λόγια δεν θα ταιριάζουν οι στήλες με τις ετικέτες τους (labels).

Ένα δεύτερο πρόβλημα είναι πως όταν εισάγουμε δεδομένα σε ένα μοντέλο πρόβλεψης, τότε η συχνότητα εμφάνισης συγκεκριμένων μοτίβων θα επηρεάσει την εκπαίδευση και κατά συνέπεια την αποδοτικότητά του. Για παράδειγμα, εάν από ένα αρχείο log/pcap με φυσιολογικού τύπου κίνηση κρατήσουμε 1000 στιγμιότυπα και από ένα αρχείο με επιθέσεις malware 100 στιγμιότυπα , τότε είναι σαν να δίνουμε στο μοντέλο εκπαίδευσης ένα ποσοστό 10% επι του συνόλου τα οποία είναι επιθέσεις. Αυτό στη συνέχεια θα

επηρεάσει τον τρόπο που θα γίνει η κατηγοριοποίηση, συνεπώς η κανονικοποίηση του συνόλου παίζει πολύ σημαντικό ρόλο στην απόδοση των μοντέλων.

Μελετήθηκαν διάφορες ιστοσελίδες και τα δεδομένα που παρέχουν για να υπάρξει μία πρώτη επαφή με τα logs και τα pcap αρχεία. Ενδεικτικά παραθέτουμε μερικά παρακάτω μαζί με τα αντίστοιχα links.

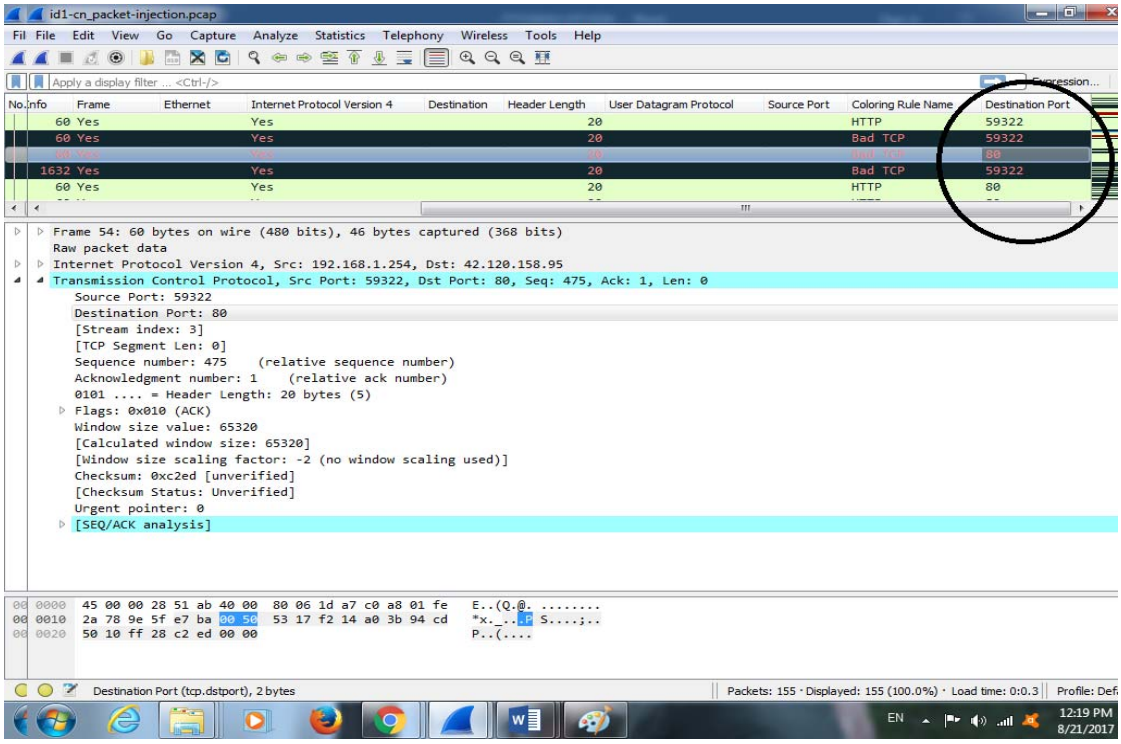
Εταιρία/Οργανισμός	Διεύθυνση
Shadow Server Malware Data site	*www.shadowserver.org*
Darpa CGC (known vulnerabilities)	*github.com/CyberGrandChallenge/samples*
SecRepo	*www.secrepo.com*
malware-traffic-analysis	*www.malware-traffic-analysis.net/*
NETRESEC Data	*www.netresec.com/?page=PcapFiles*
CTU Data	*stratosphereips.org/category/dataset.html*
Digital Corpora	*digitalcorpora.org*
Impact	*www.impactcybertrust.org*
Kyoto	*www.takakura.com/Kyoto_data/*
The Honeynet Project	*honeynet.org/challenges*
DARPA Intrusion Detection Data	*ll.mit.edu//ideval/data/*

Όπως μπορεί να δεί κάποιος που θα θελήσει να μελετήσει τα παραπάνω σύνολα, περιγράφονται δεδομένα που έχουν συγκεντρωθεί από συγκεκριμένους τύπους επιθέσεων και κυρίως malware. Επίσης η διαχείριση τέτοιων συνόλων δεδομένων είναι αρκετά δύσκολη. Ο όγκος τους είναι πολύ μεγάλος και τα περισσότερα δεν έχουν

προεπεξεργαστεί σε ότι αφορά τις απούσες (null) τιμές και δεν είναι και κανονικοποιημένα. Αυτό σημαίνει ότι το εύρος των τιμών για κάθε χαρακτηριστικό που αναπαριστούν δεν είναι σε τέτοια μορφή που να μπορεί να χρησιμοποιηθεί και να δοθεί σε ένα μοντέλο εκμάθησης ή να εφαρμοστούν στατιστικές μέθοδοι.

Επειδή παρόλα αυτά είναι σημαντικό να υπάρχει μία γενική άποψη επι των διαθέσιμων συνόλων, έτσι ώστε να μπορέσει να εκτιμηθεί ένα σύνολο που δεν θα έχει τα προβλήματα που προαναφέραμε, στην μελέτη κάποιων εξ αυτών έγιναν οι εξής ενέργειες.

- Σε ότι αφορά την ανάγνωση δεδομένων τύπου pcap υπάρχουν αρκετά διαθέσιμα εργαλεία. Το WireShark είναι ένα εργαλείο που μπορεί να καταγράφει αρχεία δεδομένων δικτυακής κίνησης με διάφορα φίλτρα διαθέσιμα αλλά και να τα διαβάζει. Επίσης μπορούμε από τα διαθέσιμα σύνολα να επιλέξουμε κάποια από τα χαρακτηριστικά τους, να προσθέσουμε δικά μας χαρακτηριστικά π.χ. κατηγοριοποίηση σε normal και μη, και τέλος να εξάγουμε επιλεκτικά κάποια ή όλα σε αρχεία διαχειρίσιμα, όπως τα csv,xls κτλ. Ένα στιγμιότυπο της ανάλυσης τέτοιου τύπου αρχείο Pcap από το WireShark φαίνεται παρακάτω.



Εικόνα 1.5 Παράδειγμα χρήσης WireShark

- Σε ότι αφορά την διαχείριση μεγάλου όγκου αρχείων, που δεν μπορούν να ‘διαβαστούν’ από τους συνήθεις editors ή το excel, η εισαγωγή τους σε μία βάση δεδομένων με χρήση κάποιου sql server μπορεί να αποτελέσει την λύση. Ένα χρήσιμο εργαλείο για τη διαχείριση βάσεων δεδομένων, είναι η phpmyadmin που συνδυάζεται συνήθως με κάποιο εξυπηρετητή ιστοσελίδων όπως ο ο xampp ή ο wampp. Χρησιμοποιώντας κώδικα php και ερωτήματα sql μπορούμε σταδιακά να εισάγουμε συγκεκριμένο αριθμό γραμμών στην βάση.

Συνοψίζοντας, η δημιουργία ενός συνόλου δεδομένων με συνδιασμό διαφόρων άλλων συνόλων,αποτελεί μία δύσκολη διαδικασία και ξεφεύγει από τα όρια αυτής της διατριβής.

1.5 Σύνοψη Κεφαλαίου

Στο κεφάλαιο αυτό έγινε λόγος για την διαδικασία συλλογής δικτυακών δεδομένων. Συγκεκριμένα έγινε αναφορά στην μεθοδολογία των επτά βημάτων όπως ορίζεται από το EC-COUNCIL[16]. Επίσης αναφέρωνται οι τρόποι αξιοποίησης των πληροφοριών αυτών στα στάδια των δικτυακών επιθέσεων. Έγινε αναφορά στα διαθέσιμα σύνολα δικτυακών δεδομένων που μπορούμε να αξιοποιήσουμε για να πειραματιστούμε στα επόμενα στάδια της εργασίας. Η επιλογή του κατάλληλου συνόλου θα αποτελέσει την βάση για το τελικό στάδιο της εργασίας, που είναι ο σχεδιασμός του πειράματος με τη πλατφόρμα της Microsoft: Azure Machine Learning Studio.

Κεφάλαιο 2

Προεπεξεργασία Δεδομένων και Επιλογή Χαρακτηριστικών

2.1 Γενικά

Όπως είδαμε στο προηγούμενο κεφάλαιο, για μία επίθεση σε ένα σύστημα χρειάζεται να συλλεχθούν πληροφορίες. Οι πληροφορίες αυτές είναι τόσο σημαντικές που θα καθορίσουν τον τύπο της μετέπειτα επίθεσης. Ανάλογα με τον τύπο των δεδομένων που θα ληφθούν, θα γίνουν και οι απαραίτητες ενέργειες για να καθοριστεί η στρατηγική της επίθεσης. Για τον λόγο αυτό, όσο καλύτερα γνωρίζουμε τα χαρακτηριστικά του συστήματος μας και τις αδυναμίες του, τόσο περισσότερο μπορούμε να προβλέψουμε τις κινήσεις αλλά και τη στρατηγική που θα έπρεπε να ακολουθήσει κάποιος προκειμένου να εισβάλει σε αυτό.

Για να μπορέσουν οι εταιρίες και οι οργανισμοί να προστατέψουν τα συστήματά τους και τα ευαίσθητα δεδομένα τους, έχουν καταφύγει στη χρήση συστημάτων ανίχνευσης εισβολών (Intrusion Detection Systems). Τα IDS συστήματα έχουν ως σκοπό να αναγνωρίσουν μία επικείμενη εισβολή. Ο τρόπος λειτουργίας τους ως προς την

αντιμετώπιση των επιθέσεων αλλά και οι μέθοδοι ανίχνευσης μπορεί να διαφέρουν. Το πιο σημαντικό ρόλο στην αξιοπιστία των συστημάτων ανίχνευσης παίζουν τα δεδομένα που τα συστήματα αυτά έχουν στην διάθεσή τους, και ο τρόπος με τον οποίο τα αξιοποιούν.

Η διαδικασία για την εξαγωγή γνώσης μέσα από βάσεις δεδομένων KDD (Knowledge Discovery in Databases) γενικά ορίζεται από τα εξής στάδια:

- Συλλογή
- Προεπεξεργασία
- Μετασχηματισμός
- Εξόρυξη δεδομένων
- Ερμηνεία/Αξιολόγηση

Στο κεφάλαιο αυτό θα αναφερθούμε στους τρόπους εκείνους που έχουμε στην διάθεσή μας για την κατηγοριοποίηση των δεδομένων.

2.2 Εξόρυξη Δεδομένων

Όπως έχουμε αναφέρει στο πρώτο κεφάλαιο, το βασικό βήμα για την ανάπτυξη της στρατηγικής επιθέσεων σε συστήματα, είναι η συλλογή πληροφοριών. Οι πληροφορίες και η χρήση τους καθορίζουν την έκταση και τη μεθοδολογία των εισβολών.

Από την άλλη πλευρά, αυτή των συστημάτων ανίχνευσης εισβολών, είναι πλέον δεδομένο πως τα πάντα έχουν να κάνουν με τη διαθέσιμη πληροφόρηση και την αξιοποίησή τους. Σε πολλά συστήματα το μέγεθος της βάσης δεδομένων που αξιοποιείται είναι και το μόνο που μετράει στην αξιοπιστία τους. Συνεπώς, ίσως το σημαντικότερο κομμάτι στην διαδικασία αναγνώρισης και αντιμετώπισης απειλών, είναι ο τρόπος που διαχειριζόμαστε τα δεδομένα. Παρόλα αυτά δεν αρκεί να έχουμε απλά ένα μεγάλο όγκο δεδομένων στην διάθεσή μας, αλλά θα πρέπει να είναι διαχειρίσιμα και αποδοτικά. Για να μπορέσουμε να καθορίσουμε και να φιλτράρουμε τα χρήσιμα και αποδοτικά στοιχεία των δεδομένων, χρειάζεται μία ισχυρή προεπεξεργασία. Η διαδικασία της

προεπεξεργασίας των δεδομένων θα καθορίσει σε πολύ μεγάλο βαθμό την αξιοπιστία του συστήματος ανίχνευσης και αντιμετώπισης των απειλών.

Ο όρος εξόρυξη δεδομένων (Data Mining), αναφέρεται στην διαδικασία εύρεσης γνωστών μοτίβων σε ένα σύνολο δεδομένων. Περιλαμβάνει τους τρόπους και τις μεθόδους εξαγωγής των στοιχείων εκείνων μέσα από ένα σύνολο δεδομένων με σκοπό την αξιοποίηση τους. Ο βασικός στόχος της εξόρυξης δεδομένων είναι να παράγει πληροφορίες για ένα σύνολο δεδομένων οι οποίες θα είναι σε τέτοιο βαθμό διαχειρίσιμες ώστε να μπορούν να ληφθούν οι κατάλληλες αποφάσεις.

Στην πράξη κάνουμε χρήση της εξόρυξης δεδομένων όταν:

- Χρειάζεται να διαχειριστούμε μεγάλο όγκο δεδομένων.
- Χρειάζεται να αποκαλύψουμε 'κρυφές' πληροφορίες επί των δεδομένων.
- Πρέπει να συσχετίσουμε και να ομαδοποιήσουμε δεδομένα.
- Χρειαζόμαστε πληροφορίες που είναι διαχειρίσιμες από μηχανές.
- Χρειάζεται ελαχιστοποίηση του κόστους ταχύτητας και επεξεργαστικής ισχύος.

Πριν την εφαρμογή των αλγορίθμων εξόρυξης,το ερευνώμενο σύνολο θα πρέπει να συναρμολογηθεί. Η προεπεξεργασία του συνόλου είναι απαραίτητο στάδιο πρίν την διαδικασία της εξόρυξης. Με την συναρμολόγηση το σύνολο 'καθαρίζεται', έτσι ώστε να διαγραφούν οι παρατηρήσεις που περιέχουν θόρυβο και αυτές με ελλιπή ή ελλείποντα δεδομένα. Με την διαδικασία της εξόρυξης αποκαλύπτονται μόνο τα πρότυπα που πράγματι εμφανίζονται στα δεδομένα μας, συνεπώς το φάσμα αυτών που ερευνούμε πρέπει να είναι αρκετά ευρύ για να περιέχει αυτά τα πρότυπα προκειμένου να προκύψει σε ένα αποδεκτό χρονικό διάστημα [8].

Παρακάτω αναφέρονται κάποιες από τις βασικές τεχνικές εξόρυξης δεδομένων [11],[18],[5].

- **Ανίχνευση ανωμαλιών (Anomaly detection).** Καλείται η αναγνώριση προτύπων από ένα σύνολο δεδομένων που εμφανίζουν διαφορετική συμπεριφορά από την προσδοκώμενη.

- **Κανόνες συσχέτισης (Μοντέλο αλληλεξάρτησης).** Με τη συσχέτιση εξετάζεται αν δύο ή περισσότερες μεταβλητές έχουν σχέση μεταξύ τους, πόσο ισχυρή είναι η σχέση αυτή και ποια κατεύθυνση έχει.

- **Συσταδιοποίηση (Clustering).** Αφορά τη διαδικασία εκείνη κατά την οποία ένα σύνολο από «αντικείμενα», διαχωρίζονται σε ένα σύνολο από λογικές ομάδες. Η καταχώρηση αντικειμένων σε ίδια ομάδα μεταφράζεται ως ομοιότητα των αντικειμένων αυτών και αντίστροφα (αντικείμενα που ανήκουν σε διαφορετικές ομάδες είναι ανόμοια).

- **Κατηγοριοποίηση (Classification).** Είναι η τεχνική κατά την οποία ένα στοιχείο ανατίθεται σε ένα προκαθορισμένο σύνολο κατηγοριών. Ο στόχος της διαδικασίας αυτής είναι η ανάπτυξη ενός μοντέλου, το οποίο αργότερα θα μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δεδομένων.

- **Παλινδρόμηση .** Αποτελεί μια ευρέως χρησιμοποιούμενη στατιστική τεχνική μοντελοποίησης για την έρευνα της συσχέτισης μεταξύ μίας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών.

2.3 Κατηγοριοποίηση

Η διαδικασία της κατηγοριοποίησης είναι στην ουσία η ταξινόμηση και οργάνωση των δεδομένων, σε μία σειρά από κλάσεις. Αυτό σημαίνει ότι τα αντικείμενα οργανώνονται σε κλάσεις βάσει κάποιων προκαθορισμένων κατηγοριών. Τα κατηγοριοποιημένα πλέον δεδομένα θα δοθούν σε κάποιον αλγόριθμο μάθησης ως training sets. Τα δεδομένα χωρισμένα σε κλάσεις που δίνονται ως είσοδοι για εκπαίδευση σε κάποιον αλγόριθμο, απαιτούν η διαδικασία της εκπαίδευσης να γίνεται με επίβλεψη.

Η κατηγοριοποίηση μπορεί να περιγραφεί ως μία διαδικασία δύο βημάτων[5]:

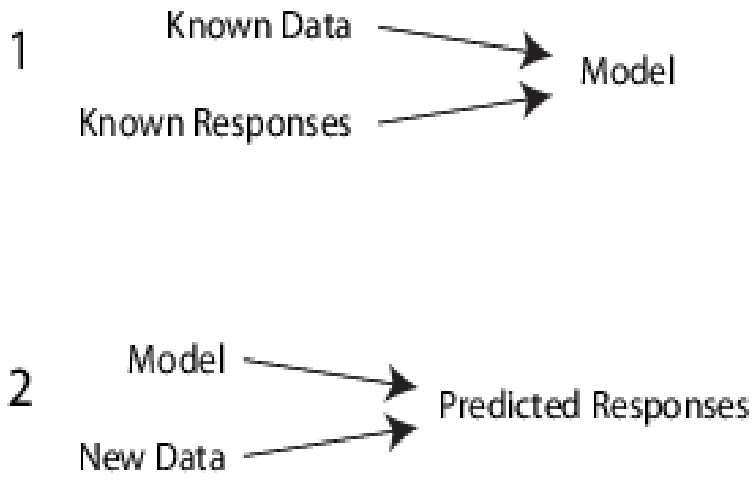
Εκμάθηση(Learning):Στο πρώτο βήμα της διαδικασίας δημιουργείται το μοντέλο με βάση ένα σύνολο προκατηγοριοποιημένων παραδειγμάτων, που ονομάζεται δεδομένα εκπαίδευσης(training data).Τα δεδομένα εκπαίδευσης αναλύονται από ένα αλγόριθμο κατηγοριοποίησης, προκειμένου να σχηματιστεί το μοντέλο. Το μοντέλο, αναπαρίσταται

με τη μορφή κανόνων κατηγοριοποίησης(classification rules), δέντρων απόφασης(decision trees) ή μαθηματικών τύπων.

Στη διαδικασία δημιουργίας ενός μοντέλου μηχανικής μάθησης, υπάρχουν κυρίως δύο τεχνικές:

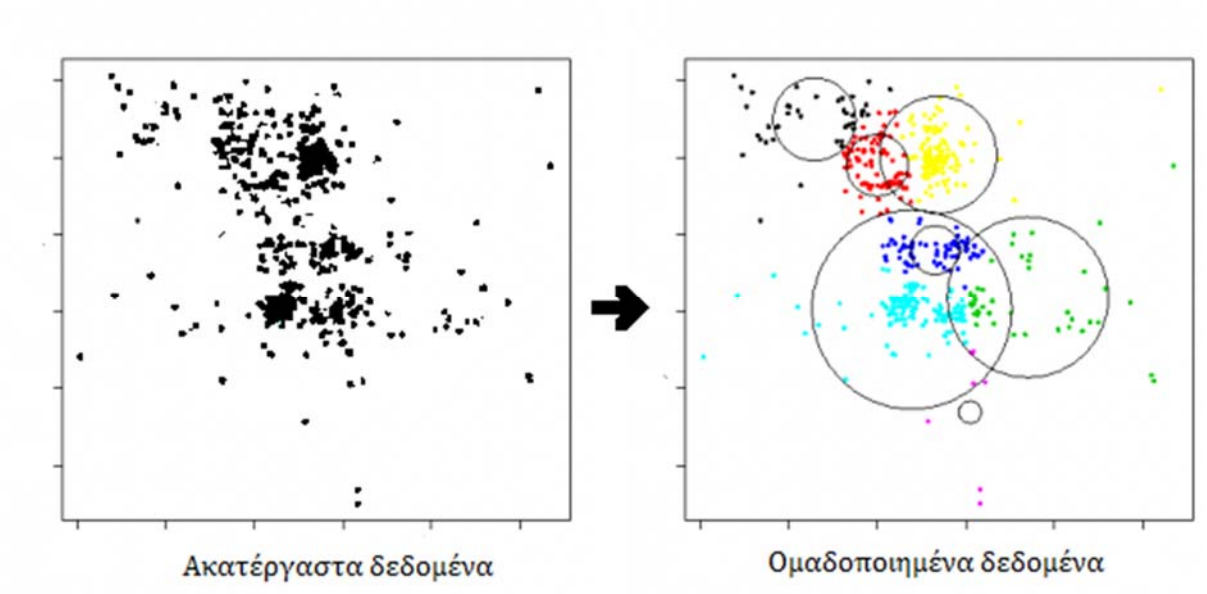
- Επιβλεπόμενη μάθηση (Supervised Learning).
- Μη επιβλεπόμενη μάθηση (Unsupervised Learning).

Στην επιβλεπόμενη μάθηση, προσπαθούμε να εκπαιδεύσουμε το μοντέλο βάσει κάποιων δεδομένων, τα οποία είναι στην ουσία κατηγοριοποιημένα. Κατά συνέπεια γνωρίζουμε εκ των προτέρων τα αποτελέσματα τα οποία εισάγονται στον αλγόριθμο. Εν συνεχεία μετά το τέλος της εκπαίδευσης, με ένα δεύτερο όμοιο set δεδομένων ή ένα υποσύνολο του πρώτου, το οποίο δεν είναι γνωστό στο μοντέλο εκπαίδευσης, ελέγχουμε το ποσοστό της επιτυχούς κατηγοριοποίησής τους.



Εκόνα 2.1 Επιβλεπόμενη(1) και μάθηση χωρίς επίβλεψη(2)

Στη διαδικασία της μάθησης χωρίς επίβλεψη, ο σκοπός είναι να ανακαλυφθεί η δομή πίσω από τα δεδομένα. Δηλαδή να βρεθούν οι συσχετίσεις εκείνες, που θα διαφοροποιήσουν και εν συνεχεία θα κατηγοριοποιήσουν τα δεδομένα μας. Στην ουσία εδώ ψάχνουμε να εξάγουμε την 'κρυφή γνώση' πίσω από τα δεδομένα προς ανάλυση.

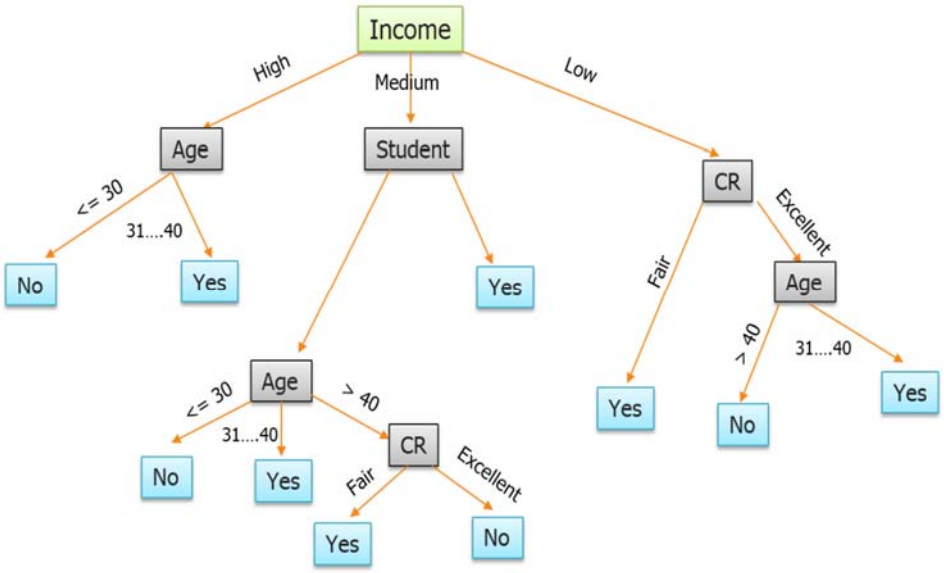


Εικόνα 2.2 Ακατέργαστα & Ομαδοποιημένα Δεδομένα

Κατηγοριοποίηση(Classification): Μετά την δημιουργία του μοντέλου, το επόμενο βήμα είναι η αξιολόγησή του. Για να επιτευχθεί αυτό, χρησιμοποιούμε τα δοκιμαστικά δεδομένα(test data) για να υπολογίσουν την ακρίβεια του μοντέλου. Το μοντέλο κατηγοριοποιεί τα δοκιμαστικά δεδομένα. Έπειτα, η κατηγορία που σχηματίστηκε με βάση τα δοκιμαστικά δεδομένα συγκρίνεται με την πρόβλεψη που έγινε για τα δεδομένα εκπαίδευσης, τα οποία είναι ανεξάρτητα από αυτά της δοκιμής. Η ακρίβεια του μοντέλου υπολογίζεται από το ποσοστό των δειγμάτων δοκιμής που κατηγοριοποιήθηκαν σωστά σε σχέση με το εκπαιδευόμενο μοντέλο.

Οι τεχνικές της κατηγοριοποίησης χρησιμοποιούνται για να εκπαιδεύσουν ένα μοντέλο μάθησης, με σκοπό την κατηγοριοποίηση των δεδομένων (data sets), σε γνωστά σύνολα. Τα σύνολα των δεδομένων σε ότι αφορά την δικτυακή κίνηση και τις πιθανές ευπάθειες μπορεί να χαρακτηρίζονται ως σύνολα φυσιολογικής συμπεριφοράς ή μη. Δηλαδή το σύστημα μπορεί να συγκρίνει τις κατηγορίες είτε ψάχνοντας για ομοιότητες με επιτρεπτές λειτουργίες, είτε ομοιότητες με μη επιτρεπτές λειτουργίες. Πολύ διαδεδομένες μέθοδοι για κατηγοριοποίηση δεδομένων είναι τα δένδρα απόφασης,τα νευρωνικά δίκτυα, η τεχνική Κ-πλησιέστερων γειτόνων, οι μηχανές υποστήριξης διανυσμάτων και οι Bayesian μέθοδοι[4].Παρακάτω γίνεται αναφορά σε δύο από τις βασικές τεχνικές, τα δένδρα απόφασης και τα νευρωνικά δίκτυα.

- **Δένδρα απόφασης:** Ένα δένδρο απόφασης, είναι μία δομή όπου οι αποφάσεις λαμβάνονται ξεκινώντας από τη ρίζα του δένδρου με μία σειρά από μη τερματικούς κόμβους και καταλήγουμε στα φύλλα. Κάθε μη τερματικός κόμβος παίρνει μία απόφαση για τη μετέπειτα διαδρομή έως ότου καταλήξουμε σε κάποιο τερματικό κόμβο- φύλλο του δένδρου.

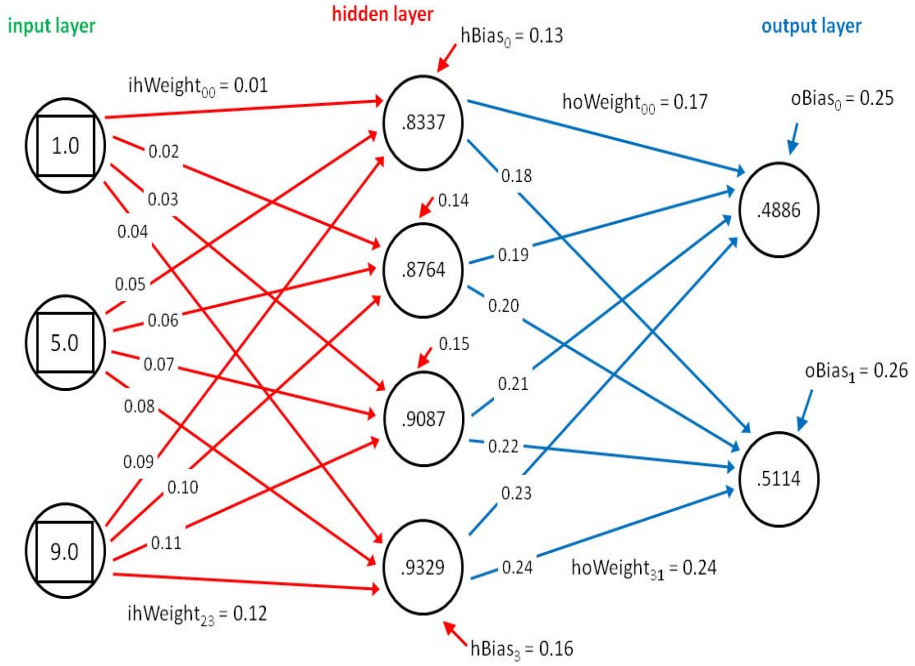


Εικόνα 2.3 Διάταξη δένδρου απόφασης

- **Νευρωνικά δίκτυα:** Τα νευρωνικά δίκτυα που είναι ευρέως διαδεδομένα σε πολλές εφαρμογές που έχουν να κάνουν με κατηγοριοποίηση και λήψη αποφάσεων, μιμούνται τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου. Η λογική είναι ότι αποτελούνται από ένα σύνολο συνάψεων, όπου σε κάθε σύναψη έχουμε την εισαγωγή δεδομένων και τη λήψη μιας απόφασης. Σε κάθε σύναψη μπορεί να έχουμε δεδομένα που έχουν προέλθει από κάποια προηγούμενη σύναψη ή και κατευθείαν ως είσοδος από το 'περιβάλλον'.

  Η έξοδος που παράγει μία σύναψη θα τροφοδοτήσει με την σειρά της κάποια άλλη. Η διαδικασία αυτή μπορεί να επαναληφθεί κυκλικά, μέχρι την τελική απόφαση. Η διαδικασία της επιλογής στους νευρώνες του δικτύου γίνεται πάνω στα βάρη που φέρουν τα δεδομένα. Οι αλλαγές επι αυτών των βαρών από τους νευρώνες γίνεται

στην προσπάθεια να προβλέψουν πώς θα επηρεάσουν τους επόμενους στη λήψη κάποιας απόφασης. Στην πραγματικότητα η αλλαγή των βαρών αυτών γίνεται σε ένα ενδιάμεσο επίπεδο νευρώνων το οποίο αποκαλούμε κρυφό.



Εικόνα 2.4 Διάταξη Νευρώνων

2.4 Συσταδιοποίηση

Σε ό,τι αφορά τη μέθοδο της συσταδιοποίησης ή αλλιώς ομαδοποίησης των αντικειμένων, υπάρχει μία βασική διαφορά σε σχέση με την προηγούμενη μέθοδο, αυτή της κατηγοριοποίησης. Εδώ στην ουσία ζητάμε από τον αλγόριθμο που χρησιμοποιούμε, να ανακαλύψει κάποια ενδιαφέρουσα δομή των δεδομένων. Βάσει της δομής αυτής καθορίζονται και οι ομάδες στο σύνολο[17].

Συνεπώς, στη συσταδιοποίηση αντικειμένων ο αλγόριθμος προσπαθεί να βρει όσο το δυνατόν περισσότερες ομοιότητες μεταξύτων δεδομένων και να δημιουργήσει τον κατάλληλο αριθμό ομάδων (clusters). Μία τέτοια διαδικασία γίνεται χωρίς επίβλεψη.

Κατά τη διαδικασία της συσταδιοποίησης προκύπτουν κάποια ζητήματα που πρέπει να έχουμε υπόψιν μας όταν χρησιμοποιούμε αυτή την μέθοδο όπως:

- **Τα ακραία σημεία:** Η πιθανότητα να βρεθούν αντικείμενα που δεν θα μπορούν να ‘ταιριάξουν’ σε καμία συστάδα είναι πάντοτε μεγάλη. Σε αυτές τις περιπτώσεις είναι προτιμότερο να αποτελέσουν από μόνα τους συστάδα.

- **Τα δυναμικά δεδομένα:** Τα δεδομένα που έχουμε στην διάθεσή μας, μπορεί να περιγράφονται από γνωρίσματα-χαρακτηριστικά που μεταβάλλονται στο χρόνο. Αυτό σημαίνει ότι οι κλάσεις που θα αποτελούν τις ομάδες στις οποίες ανήκουν τα δεδομένα θα πρέπει να επανακαθορίζονται όπου αυτό είναι απαραίτητο.

  Ο υπολογισμός της ομοιότητας στα αντικείμενα επιτυγχάνεται με βάση τα μέτρα ομοιότητας που χρησιμοποιούν μετρικές, δηλαδή την απόσταση δύο σημείων. Οι αποστάσεις Minkowski, όπως και η Ευκλείδια απόσταση χρησιμοποιούνται στα περισσότερα προβλήματα[16].

Ιδιαίτερη προσοχή πρέπει να δίνεται στα παρακάτω στοιχεία.

- **Απόσταση απλού συνδέσμου:** Η μικρότερη απόσταση μεταξύ δύο στοιχείων των δύο συστάδων.

  Με την Ευκλείδεια απόσταση να ορίζεται: $d_{j,k}=\sqrt{\sum (X_{ij}-X_{ik})^2}$

  Και την απόσταση Minkowski ως: $d_{j,k}=\sum (|X_{ij}-X_{ik}|^p)^{\frac{1}{p}}$

  Όπου

  Xij = Τιμή που αντιστοιχεί στο δείγμα j της μεταβλητής i.

  Xik = Τιμή που αντιστοιχεί στο δείγμα k της μεταβλητής i.

  i = Πλήθος των μεταβλητών.

  p= Παράμετρος που ορίζεται από τον ερευνητή.

- **Απόσταση πλήρους συνδέσμου:** Η μεγαλύτερη απόσταση μεταξύ δύο στοιχείων των δύο συστάδων.

- **Μέση απόσταση:** Η μέση απόσταση μεταξύ των στοιχείων των δύο συστάδων.

- **Απόσταση κέντρων βάρους:** Η απόσταση μεταξύ των κέντρων βάρους των δύο συστάδων. Κέντρο βάρους αποτελεί το κεντρικό σημείο μιάς ομάδας. Είναι δηλαδή το ενδιάμεσο σημείο που ορίζεται από το σύνολο των μεταβλητών που συμμετέχουν στην ομαδοποίηση.

2.5 Μείωση Διαστάσεων

Όπως αναφέραμε στο κεφάλαιο 1, τα δεδομένα που μπορεί να έχει στη διάθεσή του ένας επίδοξος υποκλοπέας είναι πολλά και διαφόρων τύπων. Ένας πίνακας συλλογής στοιχείων μπορεί να περιέχει τα χαρακτηριστικά του δικτύου για το οποίο σχεδιάζεται η επίθεση. Επίσης οι τεχνικές και οι τρόποι επίθεσης διαφέρουν ως προς τα δεδομένα που αξιοποιούνται κάθε φορά. Ένα IDS μπορεί να αναλύει κάθε φορά τα δεδομένα της διαδικτυακής κίνησης, συγκρίνοντάς τα στην ουσία με τις βάσεις δεδομένων που έχει στη διάθεσή του. Ο τελικός στόχος είναι να μπορεί να αποφανθεί για το αν υπάρχει ύποπτη συμπεριφορά. Η απόφαση αυτή θα πρέπει να είναι έγκυρη (not false alarm), άλλα και γρήγορη (time response).

Συνεπώς ένα μεγάλο πρόβλημα της διαδικασίας της ανάλυσης των δεδομένων, είναι η διαχείριση του όγκου των χαρακτηριστικών που αναλύουμε. Όταν σε ένα μοντέλο εκμάθησης, εφαρμόζουμε αλγόριθμους μείωσης διαστάσεων, προσπαθούμε να βρούμε τα χαρακτηριστικά εκείνα που παίζουν σημαντικότερο ρόλο στη διαδικασία κατηγοριοποίησής τους. Οι αλγόριθμοι αυτοί χρησιμοποιούνται για να βρούμε συσχετίσεις πάνω στα χαρακτηριστικά των δεδομένων

Ο μεγάλος αριθμός διαστάσεων στα δεδομένα μπορεί να οδηγήσει αλγόριθμους μη επιβλεπόμενης μάθησης, σε λιγότερη ακρίβεια στην κατηγοριοποίηση των δεδομένων ή σε δημιουργία κλάσεων χαμηλής ποιότητας[6]. Το σίγουρο είναι πως όσο περισσότερες είναι οι διαστάσεις, τόσο περισσότερος χρόνος απαιτείται για ανάλυση και περισσότερη επεξεργαστική δυνατότητα από πλευράς διαθέσιμων πόρων.

Η διαδικασία της εξαίρεσης κάποιου ή κάποιων χαρακτηριστικών σε ένα σύνολο, και η δημιουργία ενός νέου συνόλου με λιγότερα χαρακτηριστικά αποτελεί την μείωση των διαστάσεων(Dimensionality Reduction) επι του συνόλου αυτού.

Ένα απλό παράδειγμα

Ένα σύνολο δεδομένων dataset, διακρίνεται από τα χαρακτηριστικά που το απαρτίζουν. Στην περίπτωση μας, ένα σύνολο από διαδικτυακά δεδομένα μπορεί να εμπεριέχει χαρακτηριστικά όπως :

- Διευθύνσεις (ip) αποστολέα και παραλήπτη.
- Ημερομηνία (Date).
- Ώρα (Time).
- Μέγεθος (Size).

Ας πάρουμε την απλουστευμένη αυτή εκδοχή ενός συνόλου δεδομένων. Όπως μπορούμε να δούμε το σύνολο αποτελείται από τέσσερα χαρακτηριστικά. Μια πιθανή ταξινόμηση των δεδομένων, θα μπορούσε να γίνει με βάση το χαρακτηριστικό του μέγεθους(Size). Με βάση αυτό το χαρακτηριστικό θα μπορούσαμε να χωρίσουμε το σύνολο σε δύο κατηγορίες, μικρό (<10Kbytes) και μεγάλο (>10Kbytes). Από αυτήν την κατηγοριοποίηση, θα μπορούσε να προκύψει το συμπέρασμα πως τα χαρακτηριστικά Ώρα (Time) και οι διευθύνσεις αποστολέα και παραλήπτη μπορούν να εξαιρεθούν από την εκπαίδευση μιας και δεν παίζουν σημαντικό ρόλο στη πρόβλεψη.

Η διαδικασία της μείωσης διαστάσεων, στη πραγματικότητα αφορά αλγορίθμους και τεχνικές που χρησιμοποιούνται έτσι ώστε να δημιουργηθούν νέα χαρακτηριστικά (μεταβλητές), τα οποία αποτελούν συνδυασμό των αρχικών. Συνεχίζοντας στο παράδειγμά μας, θα μπορούσαμε να βρούμε μία πιθανή συχέτιση μεταξύ των χαρακτηριστικών του μεγέθους και της ώρας. Για παράδειγμα, έστω ότι στις βραδινές ώρες το μέγεθος των πακέτων που συναλλάσονται είναι στατιστικά μεγαλύτερο από ότι τις πρωινές. Ο συνδυασμός των δύο αυτών χαρακτηριστικών και η δημιουργία ενός νέου χαρακτηριστικού μεγέθους/ώρας αποτελεί και το ζητούμενο. Ο τελικός σκοπός της διαδικασίας αυτής είναι να δημιουργήσουμε ένα υποσύνολο χαρακτηριστικών το οποίο όμως διατηρεί την κύρια συνιστώσα μεταβολής του αρχικού. Το νέο αυτό αυτό σύνολο θα βελτιώσει την απόδοση των μοντέλων που θα εκπαιδεύσουμε.

2.6 Ανάλυση Σε Κύριες Συνιστώσες (PCA)

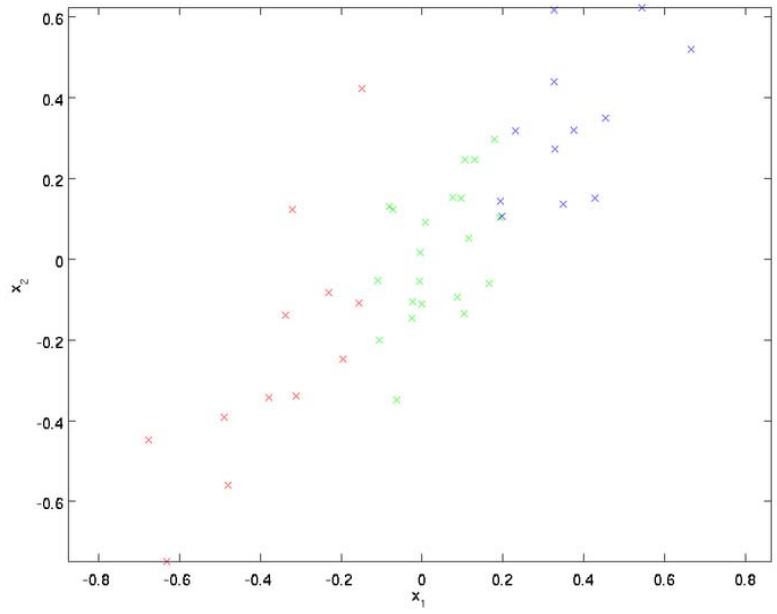
Η ανάλυση σε κύριες συνιστώσες αποτελεί μία διαδικασία η οποία με χρήση ορθογώνιων μετασχηματισμών(δηλαδή μετασχηματισμοί που αφήνουν αναλοίωτο το εσωτερικό γινόμενο στους μετασχηματισμούς πινάκων), μετατρέπει ένα σύνολο χαρακτηριστικών (μεταβλητών) πιθανώς σχετιζόμενων μεταξύ τους(Θυμίζουμε πως δύο διανύσματα ενός διανυσματικού χώρου m είναι γραμμικώς εξαρτημένα αν τουλάχιστον ένα από αυτά μπορεί να γραφτεί ως γραμμικός συνδυασμός των υπολοίπων), σε ένα νέο σύνολο γραμμικών ανεξάρτητων μεταξύ τους μεταβλητών. Οι νέες αυτές μεταβλητές ονομάζονται principal components.

Το νέο σύνολο των principal components παρουσιάζει τις ίδιες συνιστώσες μεταβολής με το αρχικό σύνολο. Παρόλα αυτά υπάρχει μείωση στο σύνολο των χαρακτηριστικών, οπου τα πρώτα principal components, να διατηρούν περισσότερο από το 90% των μεταβολών των αρχικών δεδομένων [3],[8]. Παρακάτω δίνεται ένα παράδειγμα απεικόνισης των βημάτων, εύρεσης των principal components και μείωσης των διαστάσεων.

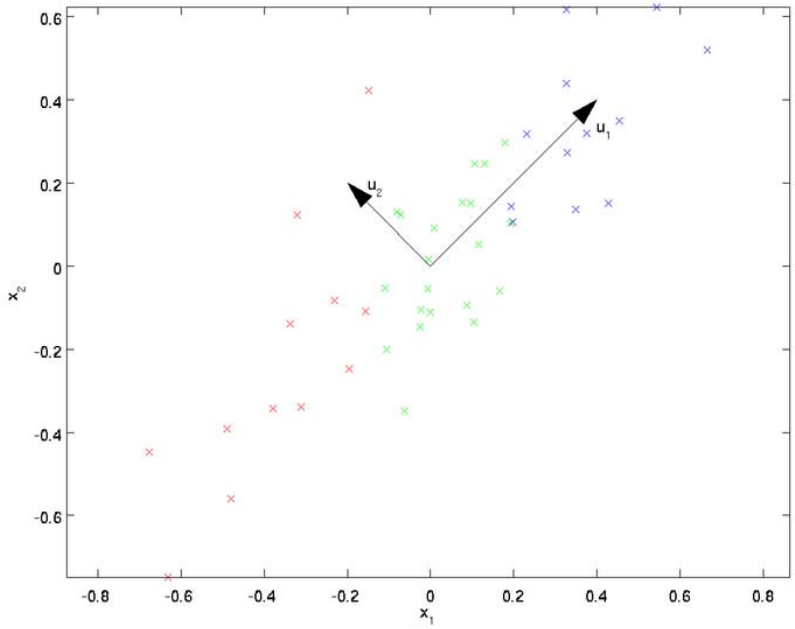
Για την καλύτερη κατανόηση του παραδείγματος δίνονται κάποιοι βασικοί ορισμοί στατιστικής.

> **Μέση Τιμή** : Ορίζεται ως το άθροισμα των παρατηρήσεων διά του πλήθους των παρατηρήσεων.
>
> **Διασπορά/Διακύμανση:** Ο μέσος όρος των τετραγώνων των αποκλίσεων των παρατηρήσεων από την μέση τιμή τους.

Έστω το διάγραμμα της παρακάτω εικόνας, που απεικονίζει ένα σύνολο δεδομένων.



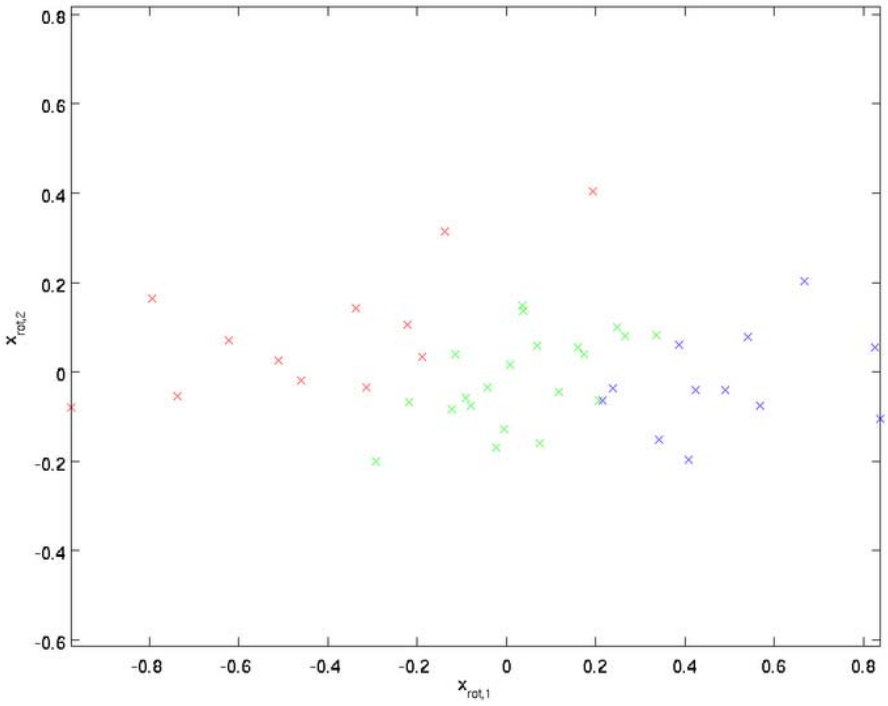
Εικόνα 2.5 Διάγραμμα αρχικών δεδομένων



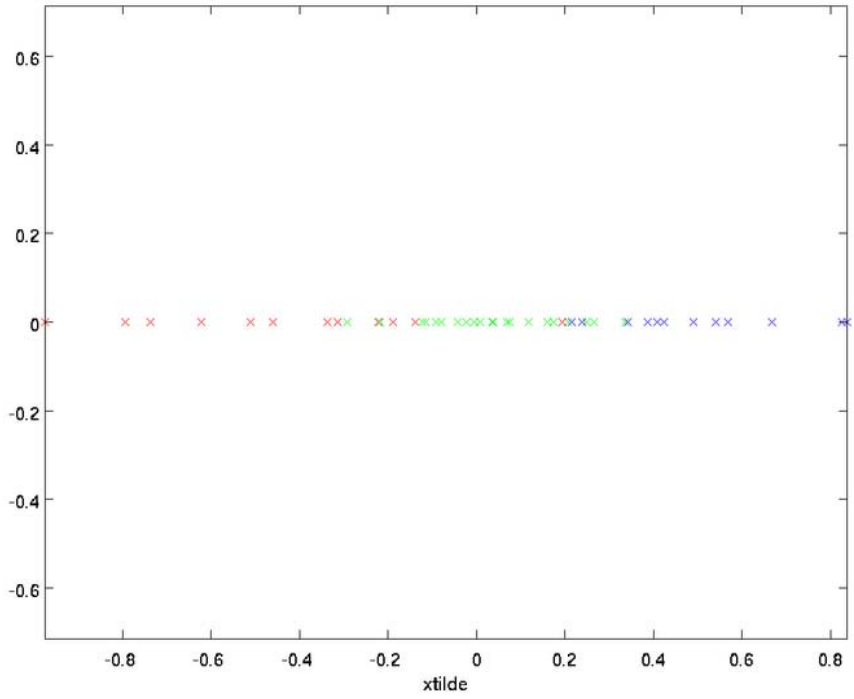
Εικόνα 2.6 Εύρεση των διανυσμάτων u_1 και u_2

Από το παραπάνω διάγραμμα οπτικά μπορούμε να πούμε ότι το διάνυσμα u_1 αποτελεί τη λεγόμενη principal κατεύθυνση της διασποράς των δεδομένων, και το u_2 την δευτερεύουσα κατεύθυνση της διασποράς (διακύμανσης). Με πολύ απλά λόγια στην ουσία αυτό που προσπαθούμε να δούμε εδώ, είναι ο διαχωρισμός.

Στη συνέχεια ο αλγόριθμος προχωρά σε περιστροφή των δεδομένων έτσι ώστε το διάνυσμα u1 να είναι παράλληλο στον άξονα X_1.

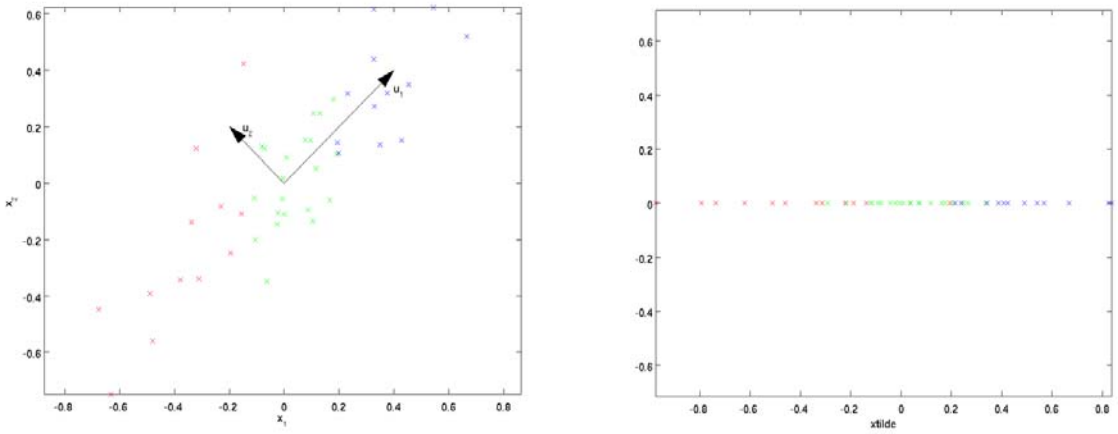


Εικόνα 2.7 Περιστροφή δεδομένων



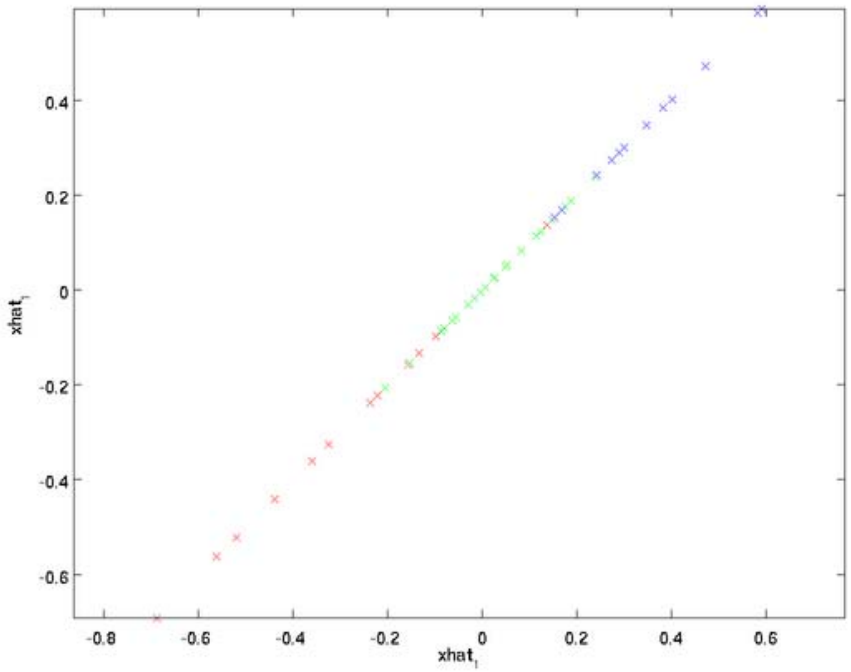
Εικόνα 2.8 Αναπαράσταση δεδομένων σε μία διάσταση

Αυτό που μπορούμε να παρατηρήσουμε είναι πως η απεικόνιση των αρχικών μας δεδομένων γίνεται πλέον σε μία διάσταση. Αυτό που βλέπουμε είναι στην ουσία σημεία πάνω σε μία ευθεία. Το σημαντικό εδώ είναι πως παρατηρώντας και συγκρίνοντας τις εικόνες 2.5 και 2.8, βλέπουμε πως έχει διατηρηθεί η βασική διασπορά των σημείων επί του διανύσματος u1. Ο λόγος χρωματισμού των δεδομένων είναι για να μπορέσουμε να συγκρίνουμε οπτικά τις θέσεις τους πάνω στους άξονες.



Εικόνα 2.9 Σύγκριση διαγραμμάτων πρίν και μετά τον μετασχηματισμό

Στο τελευταίο διάγραμμα, πλέον έχουμε την αναπαράσταση των δεδομένων πάνω σε μία ευθεία γραμμή. Σε αυτό το σημείο χρησιμοποιώντας το διάνυσμα u_2 επαναφέρουμε τα δεδομένα στους άξονες,



Εικόνα 2.10 Τελική απεικόνιση

Τα νέα δεδομένα μας μπορούν πλέον να δοθούν σε ένα μοντέλο μηχανικής μάθησης, απαιτώντας πολύ λιγότερο χρόνο επεξεργασίας.

2.8 Σύνοψη Κεφαλαίου

Στο κεφάλαιο αυτό αναλύσαμε κάποια στοιχεία, που θα χρησιμοποιήσουμε στο κεφάλαιο 4 της διατριβής. Κάναμε μία αναφορά στο κλάδο της εξόρυξης δεδομένων και σε κάποιες από τις βασικές τεχνικές που χρησιμοποιούνται , την κατηγοριοποίηση και συσταδιοποίηση. Αναφερθήκαμε επίσης στη μείωση διαστάσεων και συγκεκριμένα στην τεχνική ανάλυσης σε κύριες συνιστώσες. Δωσαμε ένα παράδειγμα μετασχηματισμού των δεδομένων με την τεχνική PCA κατά την οποία ορίσθηκαν τα νέα principal components. Στα επόμενα κεφάλαια γίνεται η επιλογή ενός συνόλου δεδομένων το οποίο θα δοθεί για εκπαίδευση ενός μοντέλου ανίχνευσης ανωμαλιών. Στο σύνολο αυτό θα γίνει η μείωση των διαστάσεων και θα προκύψουν τα νέα χαρακτηριστικά με τα οποία θα εκπαιδευτεί γίνει η εκπαίδευση.

Κεφάλαιο 3

Εισαγωγή στην πλατφόρμα Azure Machine Learning

3.1 Γενικά

Ο όγκος των διαθέσιμων συνόλων δεδομένων αυξάνεται συνεχώς. Οι επιχειρήσεις και οι οργανισμοί έχουν ανάγκη από νέα δεδομένα, ώστε να μπορέσουν να προσανατολίσουν καλύτερα τις υπηρεσίες τους. Ένας από τους βασικούς στόχους σήμερα, είναι να μπορέσουμε να κάνουμε προβλέψεις που θα μπορέσουν να μας δώσουν μία εικόνα για ένα νέο επιχειρηματικό πλάνο, για την προώθηση μίας υπηρεσίας ή ενός προϊόντος. Στην εποχή που διανύουμε κυρίαρχο ρόλο διαδραματίζει το διαδίκτυο.Τα δεδομένα που ανταλλάσσονται μέσα από τα δίκτυα αυξάνονται συνεχώς. Συνεπώς είναι επιτακτική η ανάγκη διαχείρισής τους.

Η διαχείριση δεδομένων σημαίνει οτι αυτά μπορούν να ερμηνευθούν, να βρεθούν συσχετίσεις μεταξύ τους, όπως επίσης και να τα φιλτράρουμε και να διατηρήσουμε μέσα από ένα μεγάλο όγκο τις χρήσιμες πληροφορίες. Όπως είδαμε και στο κεφάλαιο 2 η

εξόρυξη δεδομένων αποτελεί μια πλευρά διαχείρισης πληροφοριών που συνεχώς εξελίσσεται.

Η πλατφόρμα Azure studio μας προσφέρει ένα ολοκληρωμένο πακέτο εργαλείων που μας επιτρέπει να διαχειριστούμε μεγάλο όγκο δεδομένων και να δημιουργήσουμε τα δικά μας πειράματα. Τα πειράματα αυτά μας βοηθούν να βγάλουμε συμπεράσματα για ένα σύνολο που ερευνάται, ή ακόμη και να προβλέψουμε αποτελέσματα σε ένα μελλοντικό σύνολο δεδομένων.

Στα πλαίσια της διατριβής αυτής, χρησιμοποιούμε τα εργαλεία και τις δυνατότητες της πλατφόρμας, ώστε να δημιουργήσουμε ένα μοντέλο μηχανικής μάθησης. Σε ένα σύνολο δεδομένων διαδικτυακής κίνησης που έχουμε επιλέξει, θα προσπαθήσουμε να εκπαιδεύσουμε ένα μοντέλο έτσι ώστε να μπορεί να διαχωρίσει πιθανές ανωμαλίες στη διαδικτυακή κίνηση επίσης, με τη χρήση του αλγορίθμου ανάλυσης κύριων συνιστωσών (PCA) που είναι διαθέσιμος στην πλατφόρμα, θα προσπαθήσουμε να μειώσουμε τον όγκο των δεδομένων μειώνοντας τις διαστάσεις (μεταβλητές) τους. Εκτός των άλλων δυνατοτήτων της πλατφόρμας, χρησιμοποιούμε τους υπολογιστικούς πόρους των διακομιστών της Microsoft, αφού μιλάμε για μία υπηρεσία που τρέχει σε cloud. Τέλος, χρησιμοποιούμε διαγραμματικά εργαλεία απεικόνισης για να μπορούμε εύκολα να εξάγουμε συμπεράσματα οπτικά.

3.2 Χαρακτηριστικά της πλατφόρμας

Η πλατφόρμα Azure, προσφέρει ένα πλήθος εργαλείων που μπορούν να καλύψουν τη δημιουργία ενός απλού πειράματος με χρήση των modules που έχει ο χρήστης στη διάθεσή του. Συγκεκριμένα η πλατφόρμα επιτρέπει:

- Τη δημιουργία και εκπαίδευση μοντέλων μέσα από τους διαθέσιμους αλγορίθμους.

- Τη χρήση διαθέσιμων συνόλων δεδομένων(dataset) ως είσοδο στο μοντέλο εκπαίδευσης.

- Τη δυνατότητα εισαγωγής συνόλων δεδομένων από προσωπικά αρχεία σε τύπους όπως csv, arff, xml.

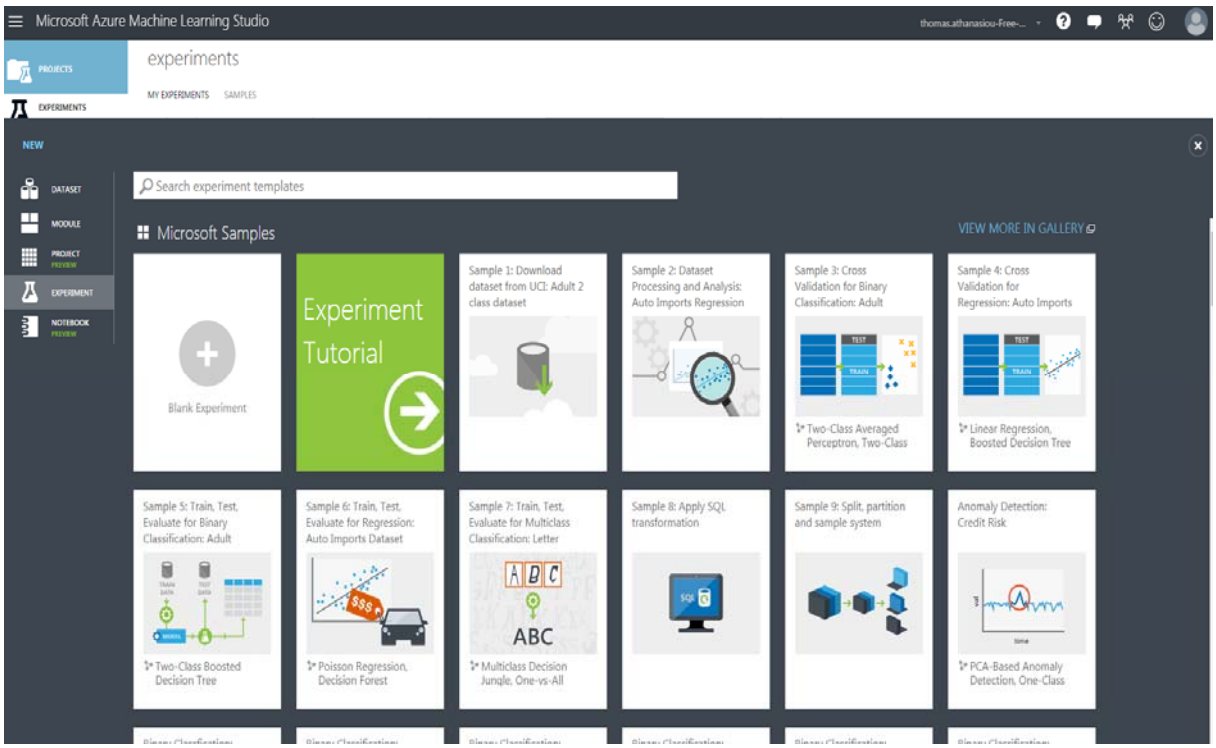
Εκτός όμως από αυτά, μπορεί να χρησιμοποιηθεί για πιο σύνθετα πειράματα με την μορφή project και τη δυνατότητα αυτά να χρησιμοποιηθούν ως έτοιμες εφαρμογές για χρήση στο ευρύ κοινό. Παρακάτω θα αναλύσουμε τα βασικά χαρακτηριστικά της πλατφόρμας, ώστε να είναι ευκολότερο στον αναγνώστη να καταλάβει τα βήματα για τη δημιουργία πειραμάτων[2].

ΤΟ ΒΑΣΙΚΟ INTERFACE

Με την είσοδο στην πλατφόρμα Azure, μετά από την απαραίτητη δημιουργία λογαριασμού, βρισκόμαστε στο τομέα δημιουργίας πειράματος, όπου μπορούμε να επιλέξουμε ανάμεσα σε έτοιμα templates (δομές πειραμάτων), τα οποία μπορεί να ταιριάζουν σε αυτό που επιχειρούμε να κάνουμε, ή να διαλέξουμε ένα κενό πείραμα και να ξεκινήσουμε τη δημιουργία του από την αρχή. Στην αρχική οθόνη στο menu αριστερά, έχουμε τις εξής επιλογές :

EXPERIMENTS

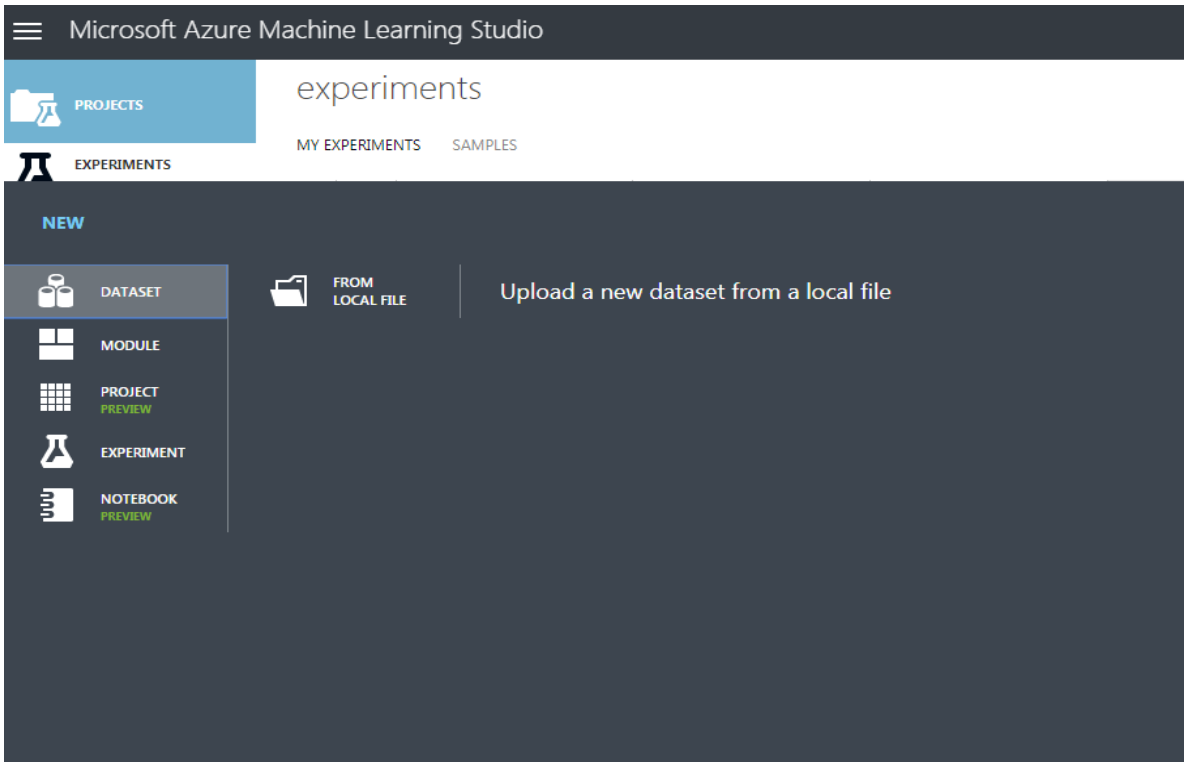
Εδώ επιλέγουμε από μία βάση πειραμάτων ή τη δημιουργία κενού πειράματος.



Εικόνα 3.1 Επιλογές δημιουργίας πειραμάτων

DATASETS

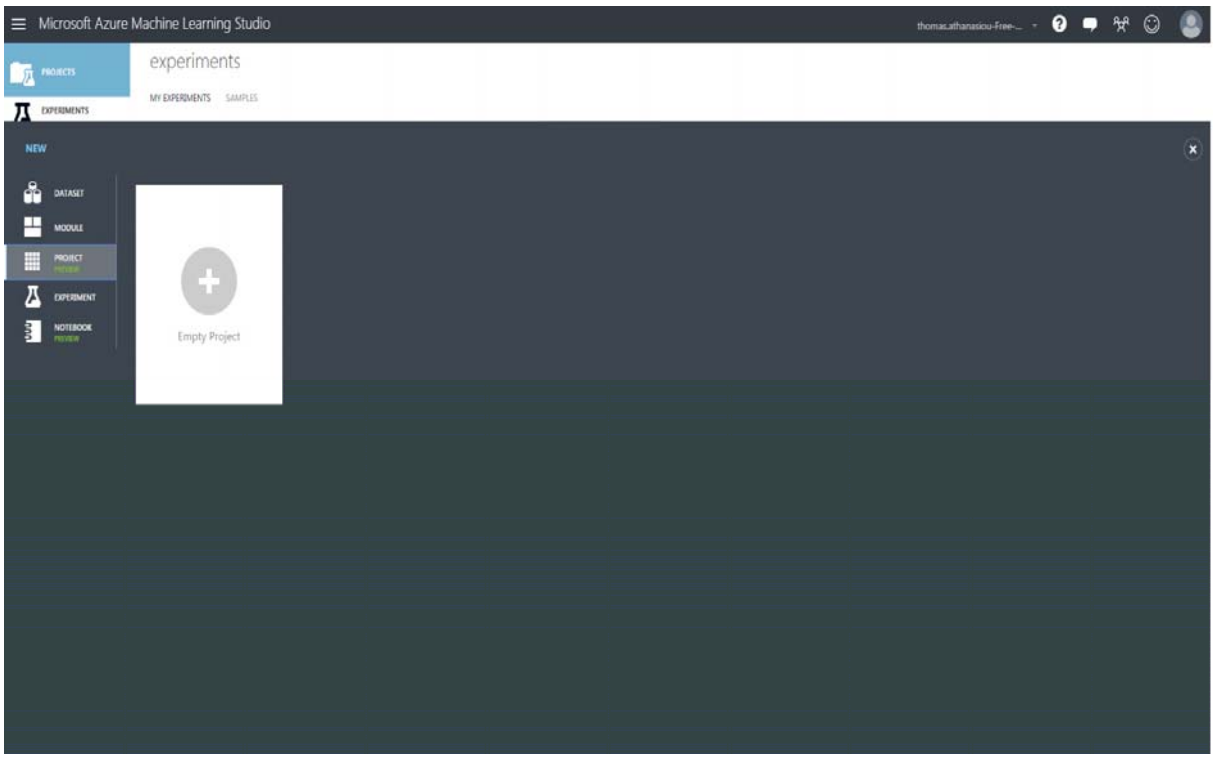
Σε αυτή τη καρτέλα μπορούμε να ανεβάσουμε το δικό μας σύνολο δεδομένων για να χρησιμοποιηθεί σε ένα ή σε περισσότερα πειράματα.



Εικόνα 3.2 Ανέβασμα συνόλου δεδομένων

PROJECTS

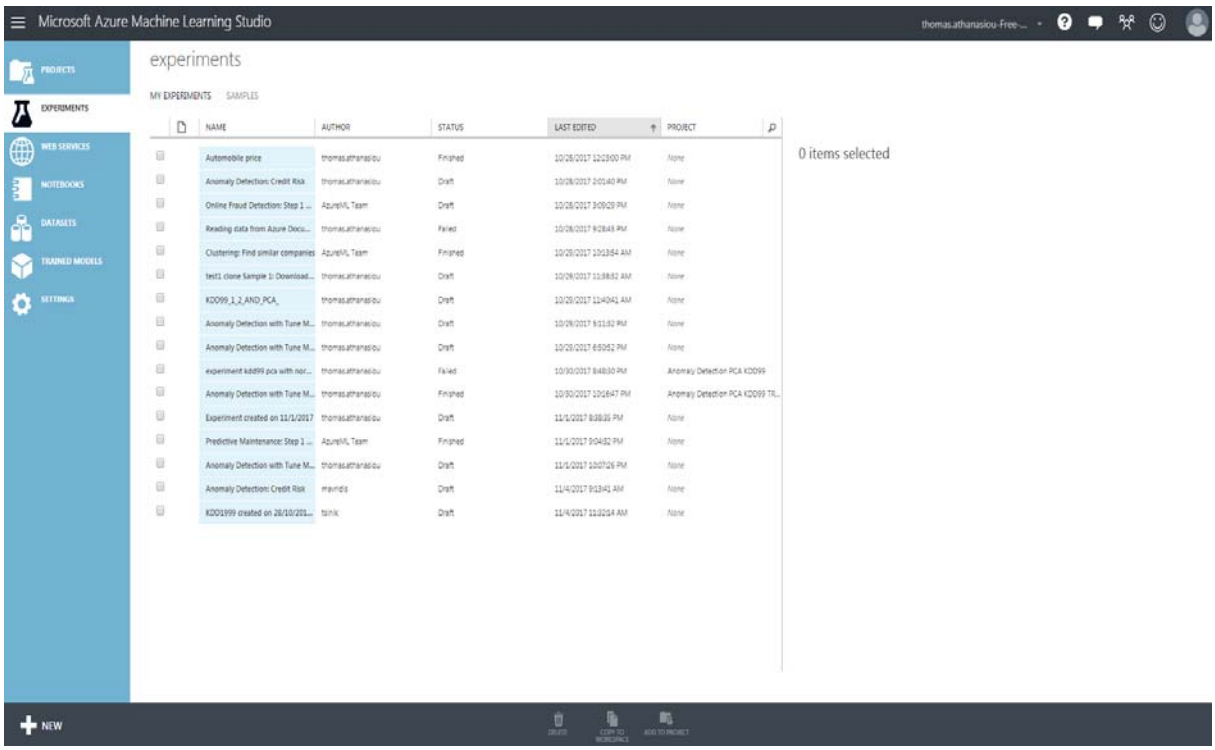
Εδώ γίνεται η δημιουργία νέου project το οποίο μπορεί να συμπεριλαμβάνει το πείραμα μας. Σημειώνεται πως ένα πείραμα δεν είναι απαραίτητο να συσχετίζεται με κάποιο project.



Εικόνα 3.3 Δημιουργία νέου Project

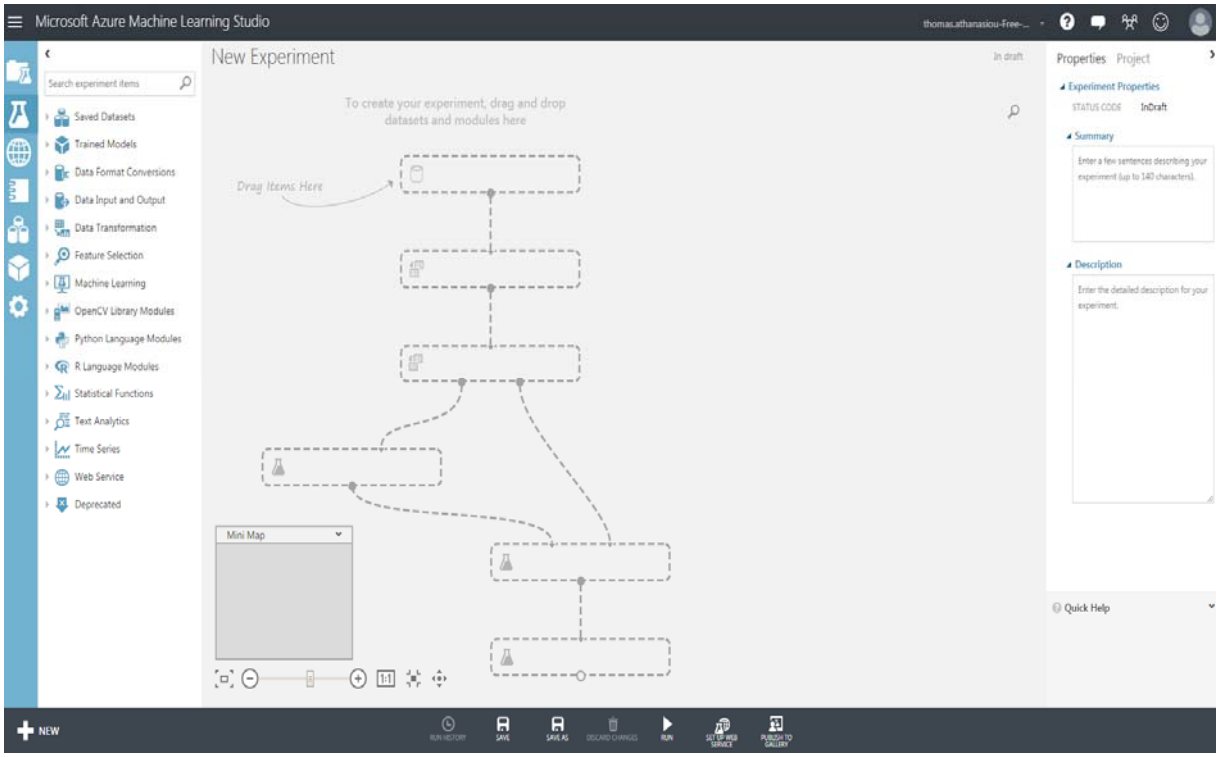
Το Βασικό menu επιλογών των πειραμάτων

Κλείνοντας με την καρτέλα του αρχικού Interface, περνάμε στο menu επιλογών που αφορά τα πειράματά μας. Αυτό που θα μας απασχολήσει κυρίως, είναι να δούμε τις επιλογές των διαθέσιμων modules που μπορούμε να εισάγουμε στο πείραμα, αλλά και τις κατηγορίες στις οποίες ανήκουν. Στην παρακάτω καρτέλα μπορεί να γίνει η επιλογή κάποιου από τα πειράματα που έχουμε δημιουργήσει ή να δημιουργηθεί νέο πείραμα.



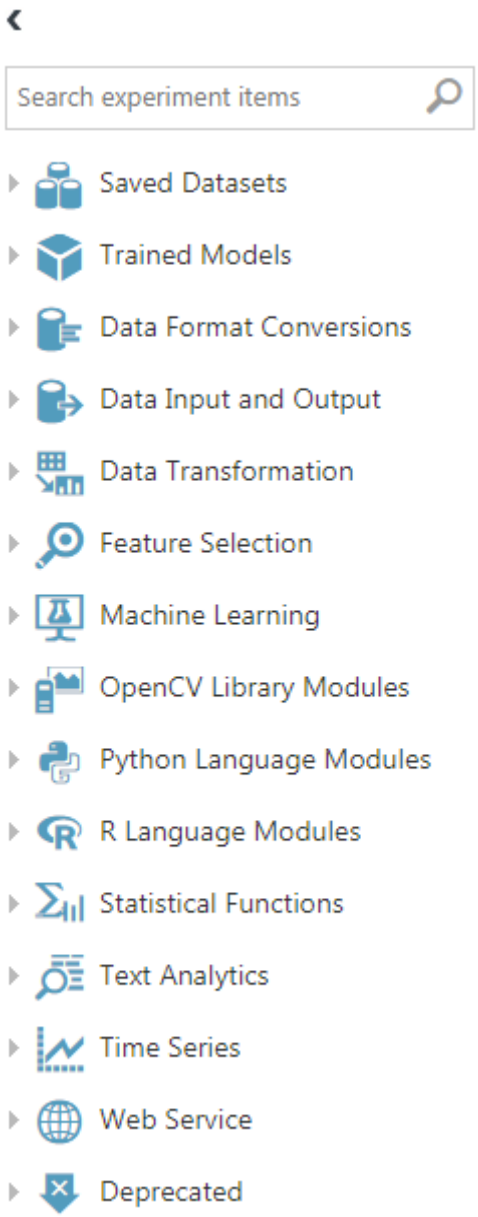
Εικόνα 3.4 Επιλογές διαθέσιμων Experiments

Έχοντας δημιουργήσει ένα νέο κενό πείραμα, μας δίνεται μαζί και ένας βασικός σκελετός με την λογική του σχεδιασμού των πειραμάτων.



Εικόνα 3.5 δημιουργία κενού πειράματος

Ενώ στην καρτέλα αριστερά, μας παρέχονται όλες οι κατηγορίες των διαθέσιμων module για να προσθέσουμε στο πείραμά μας.



Εικόνα 3.6 Επιλογές των κατηγοριών για τα module

Παρακάτω θα προχωρήσουμε σε ένα δοκιμαστικό πείραμα για την εισαγωγή των modules που θα χρειαστούμε, όπου θα περιγράψουμε και τις κατηγορίες στις οποίες ανήκουν.

3.3 Βήματα για τη δημιουργία πειραμάτων

Στο σημείο αυτό θα περιγράψουμε την σειρά των βημάτων για την δημιουργία ενός πειράματος. Το πείραμα που πρόκειται να εκτελέσουμε, θεωρείται ως ένα βασικό πρώτο πείραμα, και προτείνεται από την πλατφόρμα ως ένας καλός τρόπος για μία πρώτη επαφή με τις επιλογές και τα βήματα που χρειάζονται για την εξοικείωσή μας με αυτή.

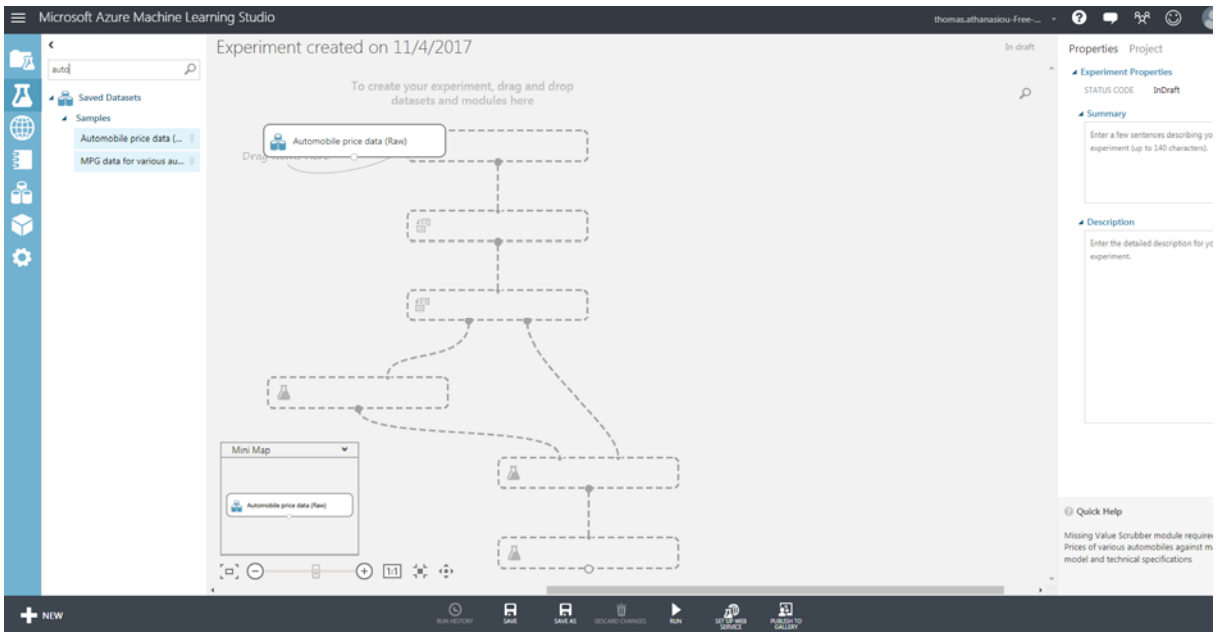
Η βασική λογική είναι να εκπαιδευτεί ένα μοντέλο με κάποιον αλγόριθμο, και στη συνέχεια να χρησιμοποιηθεί για να προβλέψει τις τιμές των αυτοκινήτων, βάσει κάποιων χαρακτηριστικών που παρέχονται. Το προτεινόμενο σύνολο δεδομένων βρίσκεται αποθηκευμένο στις βιβλιοθήκες της πλατφόρμας και μπορεί να χρησιμοποιηθεί ως έχει.

Βήμα 1ο Δημιουργία νέου πειράματος

Για το βήμα αυτό θα πατήσουμε στην επιλογή **New** όπως έχουμε δείξει και θα μας δοθεί ένα κενό template με αποτυπωμένο πάνω του το προτεινόμενο σκελετό για την εισαγωγή των modules (εικόνα 4.5). Να σημειωθεί πως ο σκελετός του πειράματος που εμφανίζεται, είναι μόνο για υπόδειξη και εξαφανίζεται μετά από την εισαγωγή του πρώτου module.

Βήμα 2ο Εισαγωγή του συνόλου δεδομένων(dataset)

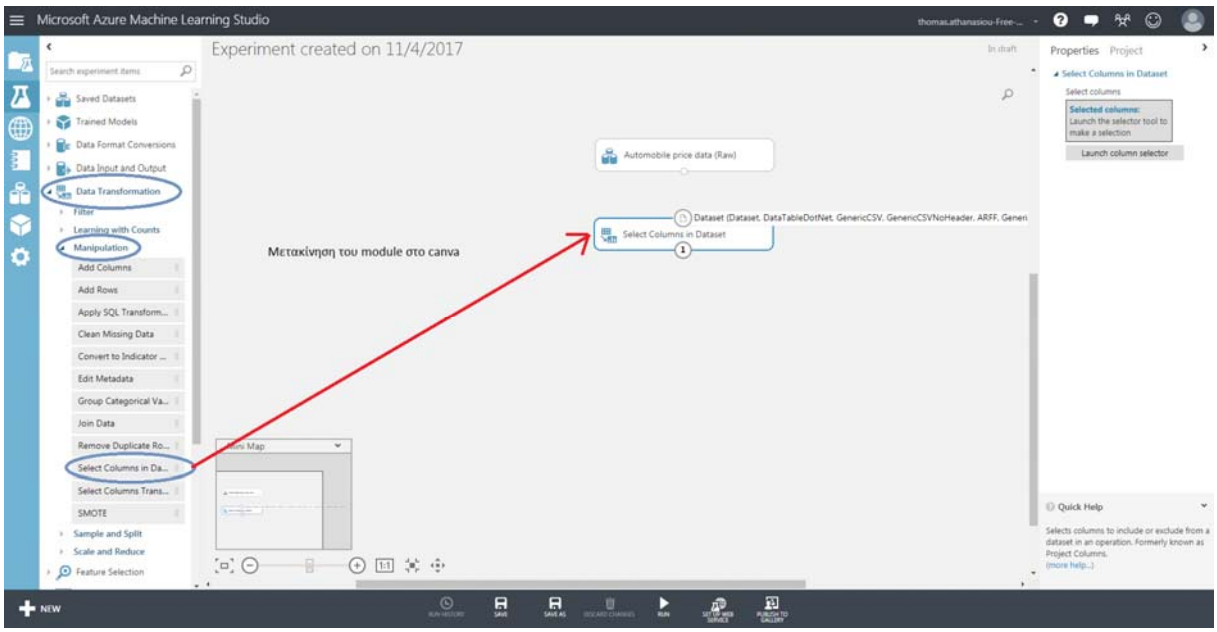
Από την καρτέλα αριστερά κάνουμε αναζήτηση για το σύνολο δεδομένων που μας ενδιαφέρει. Στην περίπτωσή μας το σύνολο ονομάζεται: **Automobile price data**. Εν συνεχεία με drag & drop περνάμε το σύνολο στο template ή αλλιώς καμβά όπως ονομάζεται η περιοχή σχεδιασμού.



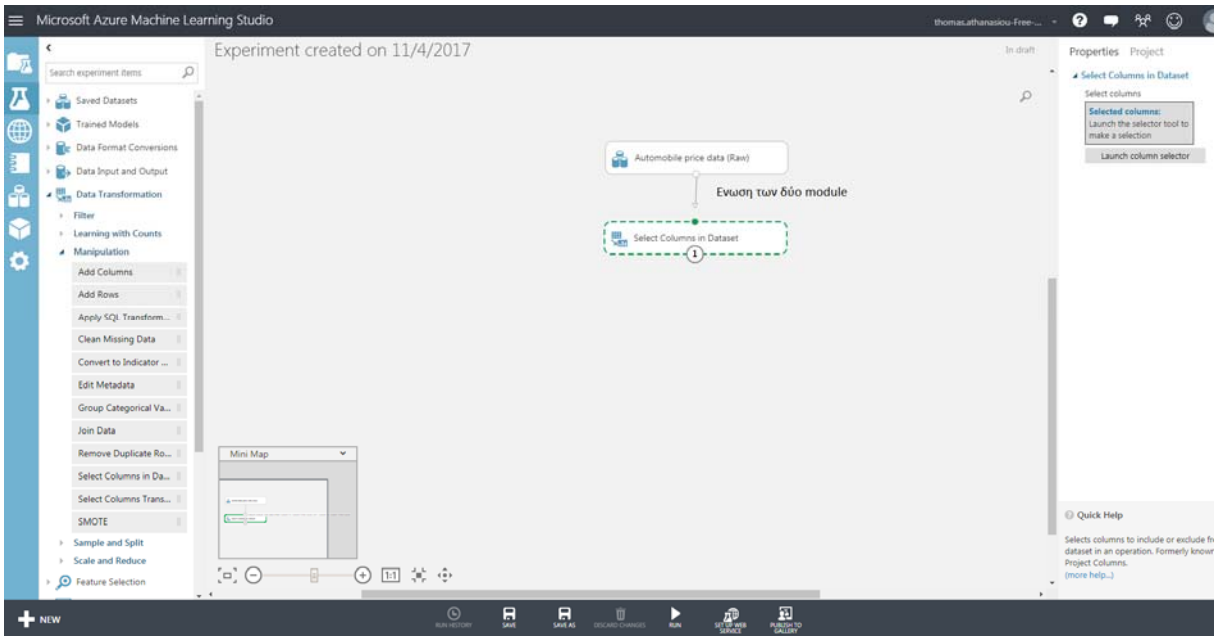
Εικόνα 3.7 Εισαγωγή dataset

Βήμα 3ο Επιλογή χαρακτηριστικών

Στο βήμα αυτό θα χρησιμοποιήσουμε το module **select columns in dataset** για να μπορέσουμε στην ουσία να επιλέξουμε τις στήλες εκείνες των χαρακτηριστικών που μας ενδιαφέρει να συμπεριλάβουμε στο πείραμά μας. Αφού βρούμε το συγκεκριμένο module, είτε γράφοντας στο πεδίο αναζήτησης, είτε με επιλογή από τις κατηγορίες της καρτέλας αριστερά, το εισάγουμε και το συσχετίζουμε με το προηγούμενο ενώνοντάς τα όπως φαίνεται παρακάτω.



Εικόνα 3.8 Εύρεση column selector και εισαγωγή στο καμβά

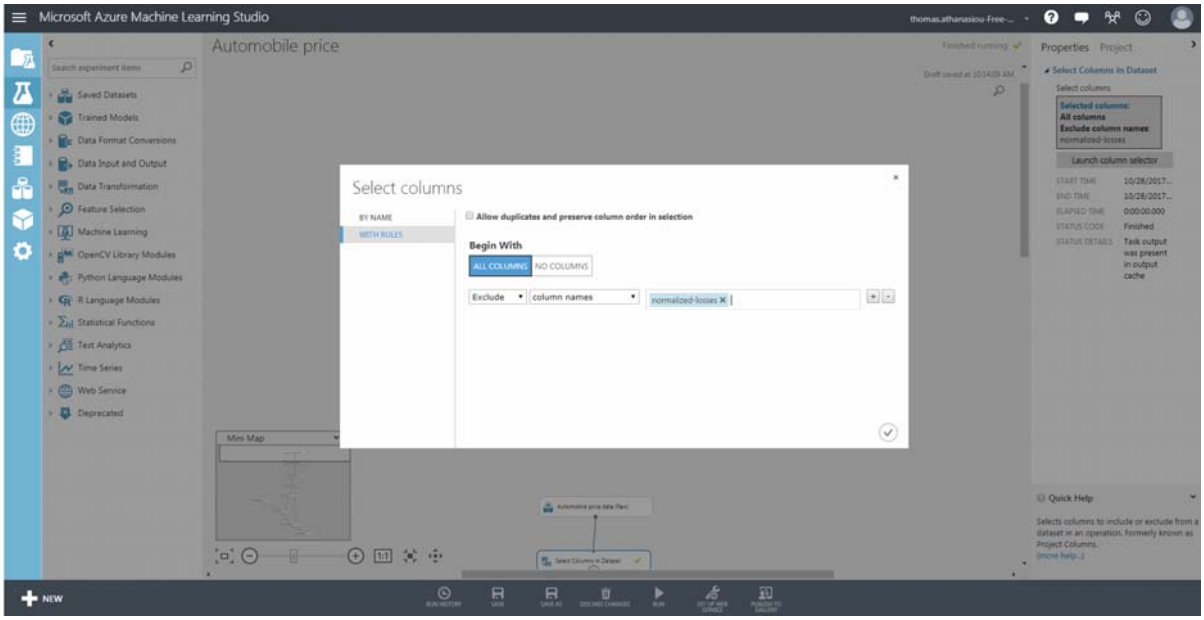


Εικόνα 3.9 Ένωση column selector με το dataset

Το κάθε module μετά την εισαγωγή του έχει κάποιες διαθέσιμες επιλογές. Για να επιλέξουμε στο συγκεκριμένο module **Select columns in dataset** τις στήλες που θέλουμε να συμπεριλάβουμε ή να εξαιρέσουμε πατάμε πάνω του.

Στη συνέχεια στη δεξιά στήλη του πειράματος πατάμε στις διαθέσιμες επιλογές του,όπως για παράδειγμα στην περίπτωση μας Select columns και στη συνέχεια exclude normalized losses, για να εξαιρέσουμε την συγκεκριμένη στήλη.

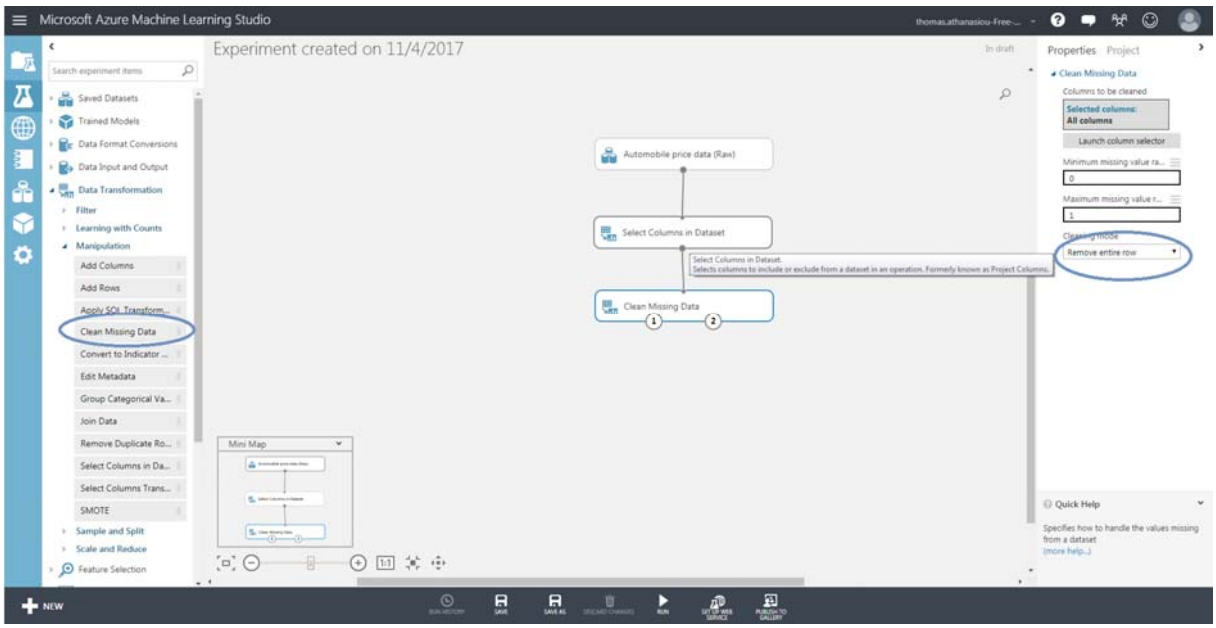
Για να μπορέσει κάθε νέο module να εκτελέσει τις λειτουργίες του με τις παραμέτρους που θέτουμε και να συσχετιστούν τα αποτελέσματα του με τα προηγούμενα, πατάμε Run στην κάτω μπάρα επιλογών.



Εικόνα 3.10 Επιλογή των στηλών που θέλουμε να εξαιρέσουμε

Βήμα 4ο Καθάρισμα του συνόλου δεδομένων

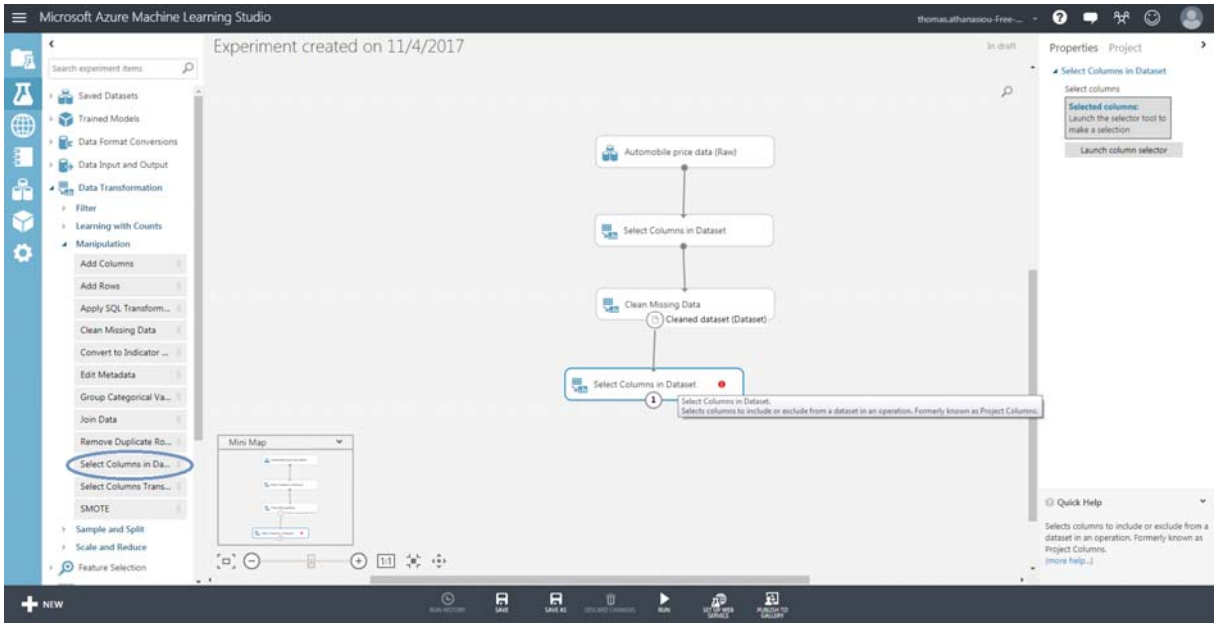
Ένα πολύ σημαντικό βήμα στα πειράματα είναι να μην υπάρχουν κενές εγγραφές, δηλαδή γραμμές στο σύνολο που να μην έχουν τιμές σε κάποια από τα χαρακτηριστικά. Αυτές οι εγγραφές θα πρέπει να αφαιρεθούν. Το module για την διαδικασία αυτή είναι το : **Clean missing data**.



Εικόνα 3.11 Αφαίρεση των εγγραφών στις οποίες λείπουν τιμές

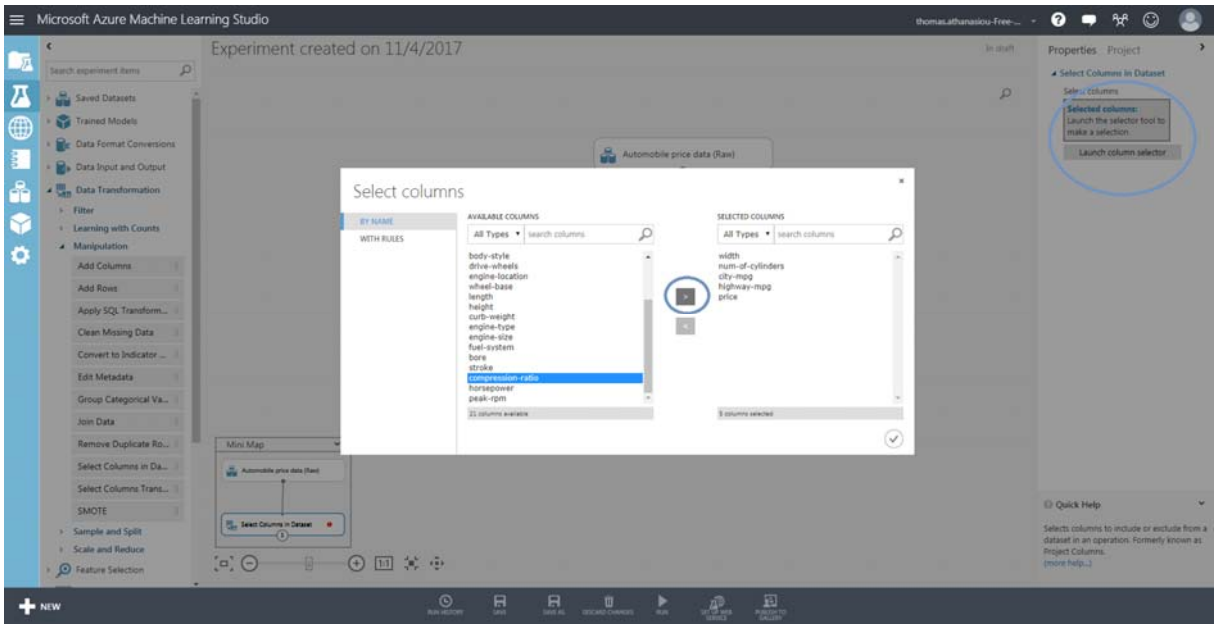
Βήμα 5ο Επιλογή των χαρακτηριστικών που θα συμπεριληφθούν

Με τον ίδιο τρόπο όπως στο βήμα 3, θα επιλέξουμε το module **Select columns in dataset,** αυτή τη φορά όχι για να εξαιρέσουμε κάποια στήλη αλλά για να επιλέξουμε όλες εκείνες τις στήλες που στη θα συμμετάσχουν στην εκπαίδευση του μοντέλου.



Εικόνα 3.12 Εισαγωγή Column Selector

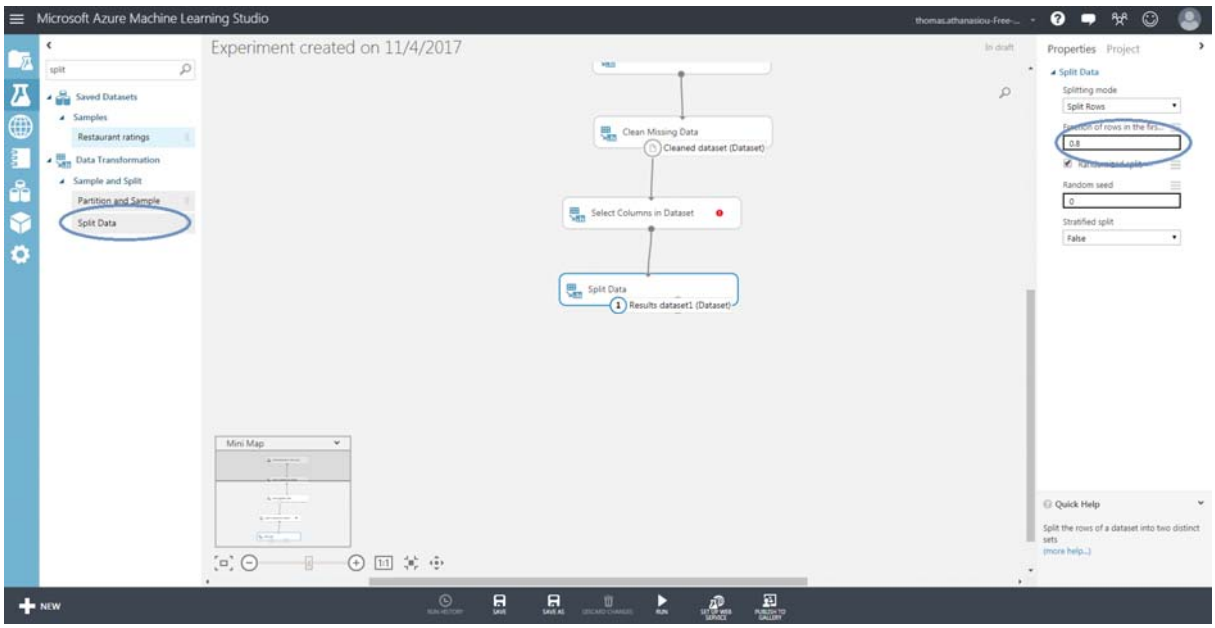
Η εισαγωγή γίνεται από τις διαθέσιμες στήλες αριστερά, προσθέτοντας τις επιλεγμένες στο δεξί μέρος.



Εικόνα 3.13 Επιλογή στηλών

Βήμα 6ο Χωρισμός των δεδομένων

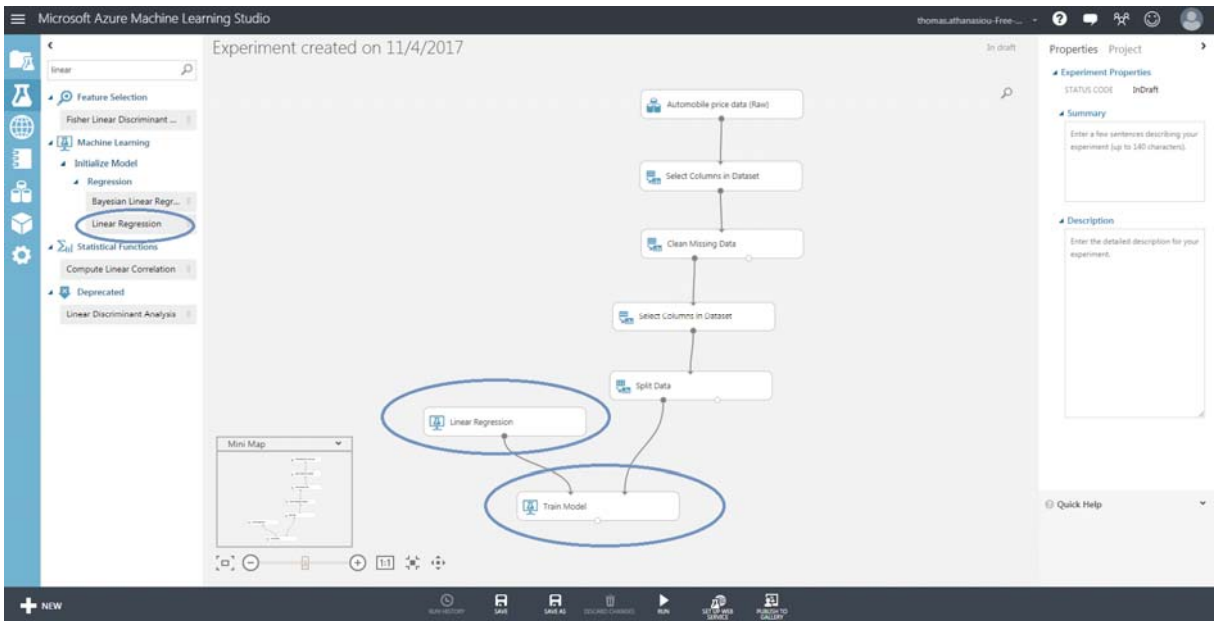
Στο βήμα αυτό θα χωρίσουμε το σύνολο των δεδομένων, έτσι ώστε να δοθεί ένα μέρος του για εκπαίδευση (training) του μοντέλου, και ένα μέρος του για τον έλεγχο της απόδοσης (testing) του μοντέλου. Ο διαχωρισμός αυτός θα γίνει με την εισαγωγή του module **Split data**. Στη δεξιά στήλη των επιλογών θα επιλέξουμε 0.8 για να δοθεί το 80% των δεδομένων για εκπαίδευση. Το ποσοστό αυτό του συνόλου θα μας δοθεί από την αριστερή έξοδο του module (έξοδος 1). Το υπόλοιπο 20% θα δοθεί για σύγκριση αργότερα και θα μας δοθεί από τη δεξιά έξοδο του module (έξοδος 2).



Εικόνα 3.14 Χωρισμός των δεδομένων

Βήμα 7ο Δημιουργία μοντέλου προς εκπαίδευση

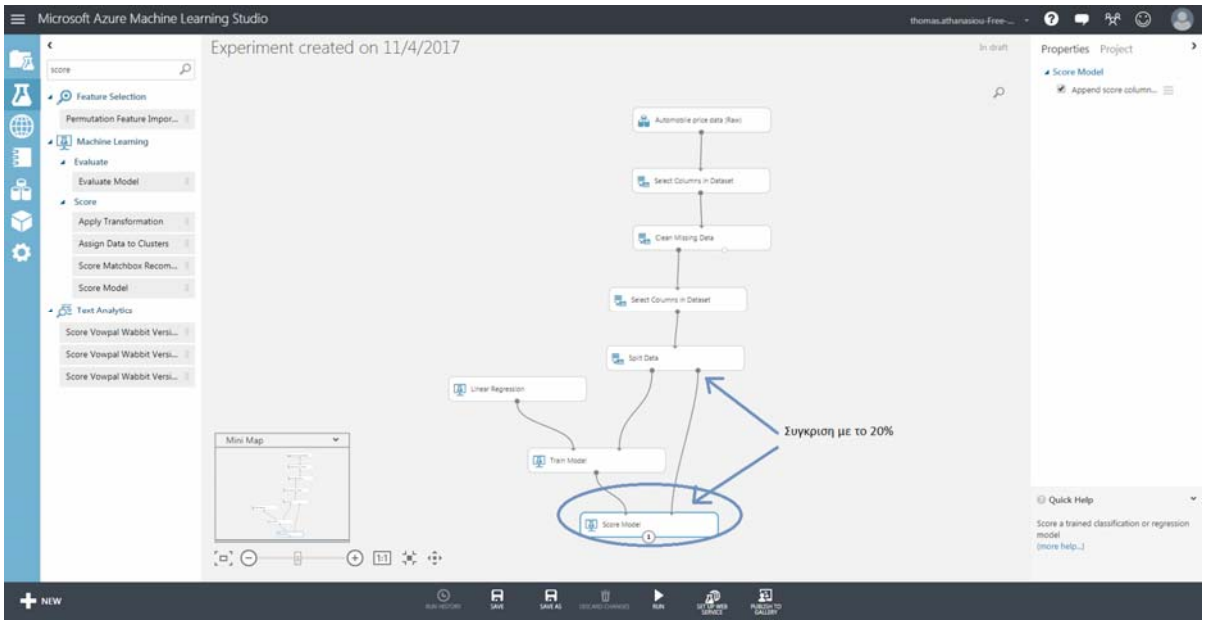
Εδώ θα εισάγουμε το μοντέλο προς εκπαίδευση με την εισαγωγή του module **Train model,** και θα γίνει επίσης η επιλογή του αλγορίθμου εκπαίδευσης, στην περίπτωσή μας **Linear Regression.** Μετά την εισαγωγή τους στο καμβά στη δεξιά είσοδο του **Train model** θα εισάγουμε τα δεδομένα που έχουμε χωρίσει από το προηγούμενο βήμα (80%), ενώ στην αριστερή είσοδό του θα ενώσουμε τον αλγόριθμο εκπαίδευσης.



Εικόνα 3.15 Δημιουργία μοντέλου προς εκπαίδευση

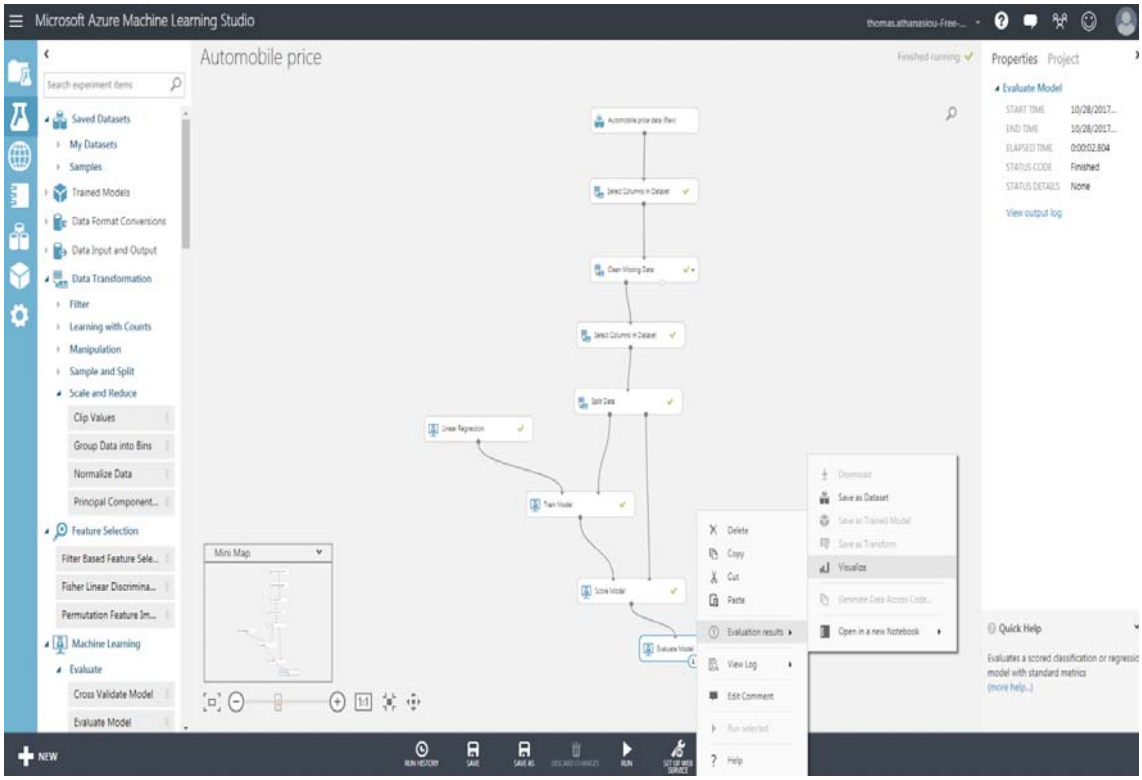
Βήμα 8ο Εκπαίδευση και αποτελέσματα

Το επόμενο βήμα για να αξιολογήσουμε τα αποτελέσματά μας είναι να προσθέσουμε ένα **Score module** το οποίο θα συνδεθεί με τη δεξιά έξοδο του module **Split data.** Στη συνέχεια το module θα προσπαθήσει να κατηγοριοποιήσει το υπόλοιπο 20% των δεδομένων, για να δούμε το ποσοστό επιτυχίας του μοντέλου που εκπαιδεύσαμε. Το τελικό στάδιο είναι να έχουμε την απεικόνιση των στατιστικών στοιχείων για το πείραμα μας και για αυτό το λόγο θα εισάγουμε το module **Evaluate model** .



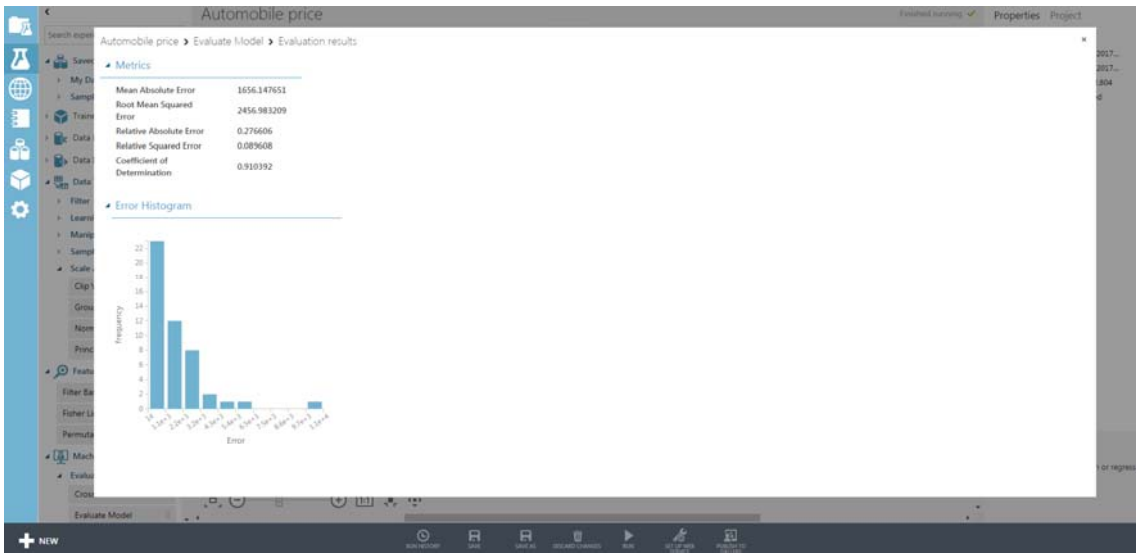
Εικόνα 3.16 Σύνδεση Score module

Για την αξιολόγηση του πειράματός μας μετά την είσοδο του **Evaluate model,** επιλέγουμε πάνω του και πατάμε evaluation results. Το τελικό πείραμα έχει την εξής μορφή:



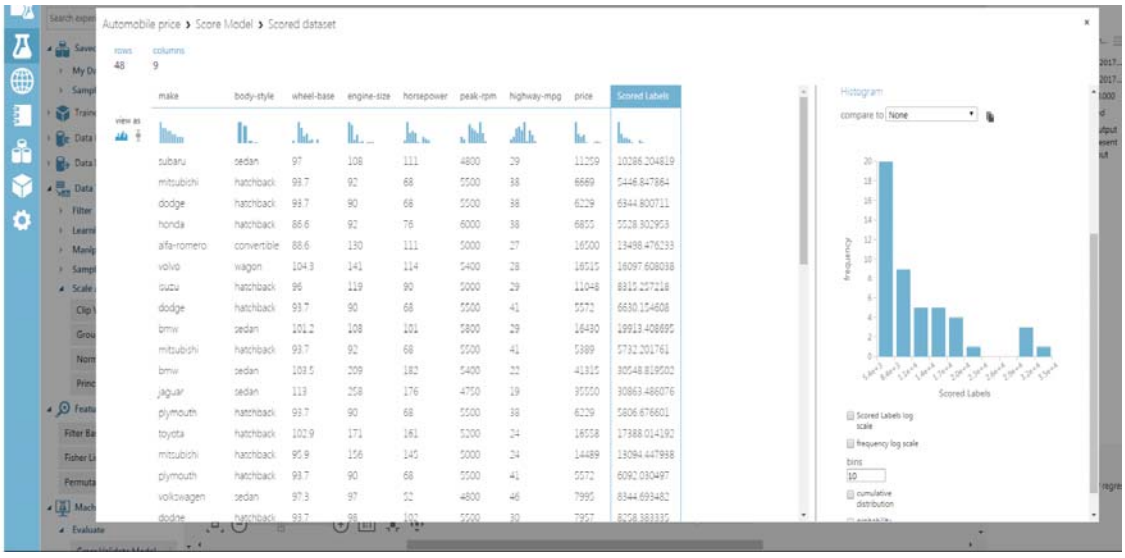
Εικόνα 3.17 Τελικό πείραμα και αξιολόγηση

Τα αποτελέσματα που έχουμε στην διάθεσή μας απεικονίζονται με error bar, δηλαδή ένα ιστόγραμμα που αποτυπώνει την μεταβλητότητα των δεδομένων και δηλώνει το σφάλμα ή την αβεβαιότητα. Έχουμε έτσι μία γενική ιδέα για το πόσο ακριβής είναι μια μέτρηση ή αντίθετα πόσο μακριά από την αναφερόμενη τιμή μπορεί να είναι η αληθινή τιμή.



Εικόνα 3.17 Evaluation results

Ενώ αντίστοιχα μπορούμε να δούμε τα score από την οπτικοποίηση του Score module, πατώντας δεξί κλικ πάνω του.



Εικόνα 3.18 Score visualize

Στη στήλη των score labels βλέπουμε τις τιμές που έχει προβλέψει το μοντέλο μας και μπορούμε να τις συγκρίνουμε με τις πραγματικές τιμές στην αριστερή στήλη(price).

3.4 Σύνοψη Κεφαλαίου

Στο κεφάλαιο αυτό, κάναμε μία εισαγωγή στο Azure studio και εξηγήσαμε κάποια βασικά βήματα για τη δημιουργία πειραμάτων. Επίσης παρουσιάσαμε ένα παράδειγμα πειράματος και τη χρήση κάποιων module . Στο επόμενο κεφάλαιο που αφορά το πειραματικό μέρος της διατριβής, θα δούμε τη χρήση και άλλων module μέσα από μία περισσότερο σύνθετη διαδικασία.

Αυτό που πρέπει να γνωρίζουμε γενικά για την πλατφόρμα είναι πως έχει πολλές επιλογές ακόμη, για παράδειγμα μπορούν να χρησιμοποιηθούν custom modules σε περισσότερο σύνθετες και πολύπλοκες ενέργειες. Τα custom module τρέχουν με κώδικα, σε γλώσσες όπως η Python και η R. Πολύ σημαντικό είναι επίσης πως σε συνδρομητικές εκδόσεις της πλατφόρμας είναι δυνατή και η χρήση βάσεων δεδομένων με Sql Servers για την αποθήκευση μεγάλου όγκου δεδομένων, αλλά και για την αξιοποίηση όλων των δυνατοτήτων της. Τέλος, ένα μεγάλο όφελος είναι πως υπάρχει η δυνατότητα για τη δημιουργία web application από ένα project.

Κεφάλαιο 4

Συνόλο Δεδομένων KDD99

4.1 Σύντομη Περιγραφή KDD99

Από τα διάφορα συγγράμματα που μελετήθηκαν και αφορούν τη χρήση ή την επεξεργασία δικτυακών δεδομένων, ένα σύνολο απαντάται πολύ συχνά. Το σύνολο αυτό είναι το KDD99[1] το οποίο μετά από μία πρώτη μελέτη, φαίνεται να φέρει όλα εκείνα τα στοιχεία που το καθιστούν ιδανικό για χρήση.

Τα τελευταία χρόνια, σχεδόν δύο δεκαετίες από την δημιουργία του συνόλου KDD99, η τεχνική αναγνώρισης ανωμαλιών έχει κερδίσει αρκετό ερευνητικό ενδιαφέρον σε ότι αφορά την ασφάλεια των δικτύων υπολογιστών. Οι έρευνες πλέον κινούνται στην κατεύθυνση της αξιολόγησης συστημάτων, τα οποία έχουν χαρακτήρα ανίχνευσης ανωμαλιών, σε αντίθεση με την παλαιότερη προσέγγιση των βασισμένων σε υπογραφές συστημάτων.

Τα IDS χρησιμοποιούν γενικά δύο μεθόδους. Η πρώτη μέθοδος είναι η Signature-based (στη βάση υπογραφών) δηλαδή το να ψάχνουν για επιθέσεις προσπαθώντας να αναγνωρίσουν συγκεκριμένα μοτίβα (patterns). Η προσπάθεια αυτή αναγνώρισης γνωστών προτύπων σε ακολουθίες εντολών που μπορούν για παράδειγμα να ταιριάζουν με ένα malware, είναι η λογική που δουλεύουν και τα γνωστά antivirus.

Η δεύτερη μέθοδος αφορά στην εκπαίδευση ενός μοντέλου με δεδομένα ομαλής λειτουργίας. Στη συνέχεια συγκρίνουμε τη συμπεριφορά του ως προς κάποια άλλη καινούργια αλληλουχία δεδομένων (πιθανή επίθεση). Με άλλα λόγια ορίζουμε κατά την εκπαίδευση το τί θεωρείται φυσιολογική λειτουργία και όταν δοθεί κάτι πέραν αυτής ορίζεται ως ανωμαλία. Το πρόβλημα με τη μέθοδο στη βάση ανωμαλιών (Anomaly-based), είναι πως η εισαγωγή κάποιας νέας φυσιολογικής αλληλουχίας δεδομένων άγνωστη στο σύστημα, αυτομάτως κατατάσσεται ως ανωμαλία, πράγμα που μειώνει την απόδοση του μοντέλου.

Το σύνολο δεδομένων KDD99, έδωσε την ευκαιρία να ερευνηθούν περαιτέρω και να αξιολογηθούν τα συστήματα που χρησιμοποιούν τη μέθοδο Anomaly-Based. Στη βιβλιογραφία αναφέρονται αρκετά προβλήματα που προκύπτουν από τη χρήση του συνόλου αυτού, υπάρχουν ωστόσο και εκτενείς αναφορές, συμπεράσματα όπως και κάποιες προτάσεις για την ενίσχυσή του, καθώς και κάποιες παραλλαγές του.

Το συγκεκριμένο dataset δημιουργήθηκε για λογαριασμό της DARPA (Defense Advanced Research Projects Agency) με σκοπό να δοκιμάσει το IDS της. Έτσι μία προσομοίωση προγραμματίστηκε στο εργαστήριο Lincoln του MIT το 1986. Το σενάριο αφορά μία επίθεση σε μία αεροπορική βάση. Η προσομοίωση αυτή επαναλήφθηκε ένα χρόνο μετά ενσωματώνοντας κάποιες βελτιώσεις. Τα δεδομένα που αποκτήθηκαν αφορούσαν Host και Network ανάλυση.

Στην ανάλυση δικτύου καταγράφηκαν δεδομένα που είχαν να κάνουν με φυσιολογικές λειτουργίες, δηλαδή χωρίς κάποια επίθεση, για διάστημα δύο εβδομάδων. Στη συνέχεια και για πέντε ακόμα εβδομάδες, συλλέχθηκαν δεδομένα από συγκεκριμένες επιθέσεις που επιδιώχθηκαν στο σύστημα. Οι συμμετέχοντες στις δύο ομάδες που ανέλαβαν την προσομοίωση (Lee, Stolfo), ανέλαβαν να εξάγουν συγκεκριμένα χαρακτηριστικά από την βάση της DARPA, τα οποία προεπεξεργάστηκαν κατάλληλα ώστε να δοθούν στον ετήσιο

διαγωνισμό του KDD (Knowledge Discovery and Data Mining) το 1999[1],[14]. Το σύνολο αυτό ονομάστηκε KDD99 και έχει τα εξής χαρακτηριστικά:

- Συνολικά επτά εβδομάδες καταγραφών σε δεδομένα, εκ των οποίων οι πέντε αφορούσαν δεδομένα επιθέσεων, και οι δύο σε δεδομένα φυσιολογικής κίνησης, καθιστώντας το κατάλληλο για ανίχνευση ανωμαλιών.

- Οι κλάσεις του αφορούν πέντε κατηγορίες: DOS (Denial of Service), Probe, R2L (Root 2 Local), U2R (User 2 Root) and Normal.

- Περιέχονται 24 τύποι επιθέσεων για εκπαίδευση, και 14 ακόμη για έλεγχο της ικανότητας του μοντέλου να τα κατατάσσει σε μη φυσιολογική δραστηριότητα χωρίς συγκεκριμένο τύπο επίθεσης.

- Διαθέτει συνολικά 4.898.430 περιπτώσεις, όπου το 80% αφορά επιθέσεις. Αυτό αμφισβητεί την αναπαράσταση πραγματικών δεδομένων αφού στατιστικά συνήθως μόνο το 0.01% αφορά επιθέσεις.

- Επίσης υπάρχουν πολλές διπλές εγγραφές όπως και σπάνιες εμφανίσεις συγκεκριμένων τύπων επιθέσεων που θα αναλύσουμε παρακάτω.

- Πρόκειται για ένα σύνολο δεδομένων μεγάλου όγκου, από το οποίο χρησιμοποιείται συνήθως μόνο ένα μέρος του για εκπαίδευση.

4.2 Ανάλυση Επιθέσεων KDD99

Όπως είδαμε, το KDD99 χωρίζει τις επιθέσεις σε 4 κλάσεις (κατηγορίες), που περιγράφονται ως εξής:

- **Denial of Service Attack (DoS)**: Αφορά σε τύπο επιθέσεων που σκοπό έχουν να απασχολήσουν τη μνήμη ή τον επεξεργαστή του διακομιστή, τόσο ώστε να μη μπορεί πλέον να απαντά σε αιτήματα χρηστών, και συνεπώς τον βγάζουν εκτός λειτουργίας.

- **User to Root Attack (U2R)**: Είναι μία κατηγορία επιθέσεων τύπου exploit (εκμετάλλευση ευπάθειας), κατά την οποία ο επιτιθέμενος εισάγεται κανονικά στο σύστημα ως πιστοποιημένος χρήστης (χρησιμοποιώντας δεδομένα που έχει αποκτήσει με τεχνικές sniffing, social engineering, dictionary attack). Μετά την είσοδο στο σύστημα προσπαθεί να αντλήσει πληροφορίες και να εκμεταλλευτεί συγκεκριμένες ευπάθειες.

- **Remote to Local Attack (R2L)**: Αφορά στην προσπάθεια απόκτησης πληροφοριών για ευπάθειες σε ένα σύστημα μέσα από την αποστολή πακέτων. Συνεπώς ο επιτιθέμενος βρίσκεται στο προηγούμενο βήμα των επιθέσεων U2R και προσπαθεί να αποκτήσει πληροφορίες όπως username και password για να εισαχθεί ως πιστοποιημένος χρήστης.

- **Probing Attack**: Είναι η προσπάθεια εύρεσης πληροφοριών σε ένα δίκτυο, ώστε να μπορούν να παρακαμφθούν τα συστήματα ελέγχου πρόσβασης.

Σε ό,τι αφορά τους τύπους των επιθέσεων που περιλαμβάνονται στο KDD99, θα πρέπει να σημειωθεί πως τα δεδομένα ελέγχου διαφέρουν με τα δεδομένα εκπαίδευσης, καθώς στα πρώτα περιλαμβάνονται κάποιοι επιπλέον τύποι επιθέσεων. Η λογική η οποία υποστηρίζεται από πολλούς ειδικούς στο τομέα των εισβολών, είναι πως διαφορετικοί τύποι επιθέσεων έχουν πολλά κοινά στοιχεία μεταξύ τους με αποτέλεσμα να μπορούν να συμπεριληφθούν στην ίδια κατηγορία.

4.3 Ανάλυση Χαρακτηριστικών

Τα διαθέσιμα χαρακτηριστικά του συνόλου KDD99, περιγράφονται από τις στήλες (Labels) και μπορούν να χωριστούν σε τρείς βασικές κατηγορίες.

1. **Βασικά χαρακτηριστικά (Basic features).** Εδώ περιλαμβάνονται όλα τα χαρακτηριστικά γνωρίσματα που μπορούν να εξαχθούν από μία σύνδεση τύπου TCP/IP. Σε αυτή τη κατηγορία ανήκουν τα χαρακτηριστικά εκείνα που η ανάλυση τους απαιτεί αρκετό χρόνο με αποτέλεσμα να επιβαρύνεται η διαδικασία ανίχνευσης των εισβολών.

2. **Χαρακτηριστικά δικτυακής κίνησης (Traffic features).** Σε αυτή τη κατηγορία περιλαμβάνονται οι ενέργειες που γίνονται μέσα σε ένα καθορισμένο χρονικό όριο και διακρίνονται σε :

   2.1 Χαρακτηριστικά του ίδιου διακομιστή (Same host features). Περιλαμβάνουν συνδέσεις που έγιναν τα τελευταία δύο δευτερόλεπτα και αφορούν στην ίδια διεύθυνση αποστολής με αυτή της τρέχουσας σύνδεσης. Περιλαμβάνουν στατιστικές τιμές από τις συμπεριφορές των αιτημάτων των πρωτοκόλλων.

   2.2 Χαρακτηριστικά ίδιας υπηρεσίας (Same service features). Περιλαμβάνουν συνδέσεις που έγιναν τα τελευταία δύο δευτερόλεπτα και χρησιμοποιούν την ίδια υπηρεσία με αυτή της σύνδεσης που επιτεύχθηκε. Οι δύο παραπάνω τύποι χαρακτηριστικών ονομάζονται time-based, λόγω του χρονικού περιορισμού των δύο δευτερολέπτων. Παρόλα αυτά επειδή υπάρχουν διάφορες παραλλαγές της συμπεριφοράς στα probe attacks που διαρκούν πολύ περισσότερο από δύο δευτερόλεπτα, προτιμήθηκε να δοθεί βάση στο σύνολο της σύνδεσης (connection window) συνυπολογίζοντας 100 συνδέσεις για κάθε υπηρεσία.

3. **Χαρακτηριστικά Περιεχομένων.** Εδώ περιλαμβάνονται χαρακτηριστικά όπως ο αριθμός των αποτυχημένων προσπαθειών εισόδου στο σύστημα. Ο λόγος για την χρήση τέτοιων δεδομένων, είναι πως στις επιθέσεις τύπου Dos και Probe, υπάρχουν συγκεκριμένα μοτίβα που μπορούν να αναγνωριστούν αφού

περιλαμβάνουν πλήθος συνδέσεων στον ίδιο host σε μικρό χρονικό διάστημα. Στις επιθέσεις τύπου R2L και U2R υπάρχει ενσωμάτωση σε τμήματα των δεδομένων των πακέτων και αφορούν συνήθως σε μία μόνο σύνδεση. Συνεπώς υπάρχει η ανάγκη χρήσης τέτοιου είδους δεδομένων για να μπορούν να αναγνωριστούν. Στους παρακάτω πίνακες παρουσιάζονται τα χαρακτηριστικά που περιγράφηκαν παραπάνω.

Ονομασία	**Περιγραφή**	**Τύπος**
duration	Αριθμός δευτερολέπτων σύνδεσης.	Συνεχής
protocol_type	Τύπος πρωτοκόλλου πχ udp.	Διακριτή
service	Υπηρεσία δικτύου προορισμού πχ http,telnet.	Διακριτή
src_bytes	Αριθμός δεδομένων σε bytes από την πηγή στο προορισμό.	Συνεχής
dst_bytes	Αριθμός δεδομένων σε bytes από τον προορισμό στη πηγή.	Συνεχής
flag	Τιμή normal ή error κατάσταση	Διακριτή
land	Κατάσταση 1 αν η σύνδεση αφορά στον ίδιο διακομιστή αλλιώς 0.	Διακριτή
wrong_fragment	Αριθμός λάθος fragments	Συνεχής
urgent	Αριθμός επειγόντων πακέτων.	Συνεχής

Πίνακας 4.1 Βασικά χαρακτηριστικά συνδέσεων TCP

Όνομα χαρακτηριστικού	**Περιγραφή**	**Τύπος**
hot	Αριθμός δεικτών hot	Συνεχής
num_failed_logins	Αριθμός αποτυχημένων προσπαθειών εισόδου.	Συνεχής
logged_in	Κατάσταση 1 σε περίπτωση επιτυχούς σύνδεσης αλλιώς 0.	Διακριτή
num_compromised	Αριθμός συμβιβαστικών συνθηκών.	Συνεχής
root_shell	Κατάσταση 1 σε περίπτωση απόκτησης δικαιώματος χρήσης root shell.	Διακριτή
su_attempted	Κατάσταση 1 αν δοκιμάστηκε χρήση κάποιας εντολής σε su root(superuser), 0 αλλιώς.	Διακριτή
num_root	Αριθμός προσβάσεων root	Συνεχής
num_file_creations	Αριθμός διαδικασιών/εντολών δημιουργίας αρχείων που επιτεύχθηκαν.	Συνεχής
num_shells	Αριθμός shell prompts που χρησιμοποιήθηκαν.	Συνεχής
num_access_files	Αριθμός διαδικασιών/εντολών που χρησιμοποιήθηκαν για πρόσβαση σε αρχεία τύπου -control files.	Συνεχής

num_outbound_cmds	Αριθμός εξερχομένων εντολών σε μία συνεδρία ftp.	Συνεχής
is_hot_login	Κατάσταση 1 αν ο συνδεδεμένος χρήστης ανήκει την hot λίστα, αλλιώς 0.	Διακριτή
is_guest_login	Κατάσταση 1 ο συνδεδεμένος χρήστης είναι guest, αλλιώς 0.	Διακριτή

Πίνακας 4.2 Χαρακτηριστικά συνδέσεων του ίδιου διακομιστή

Όνομα χαρακτηριστικού	**Περιγραφή**	**Τύπος**
count	Αριθμός συνδέσεων στον ίδιο host με αυτόν της τρέχουσας σύνδεσης τα τελευταία 2 δευτερόλεπτα. (Τύπος Ι)	Συνεχής
serror_rate	Ποσοστό συνδέσεων που έχουν 'SYN' error.(τύπου Ι)	Συνεχής
rerror_rate	Ποσοστό συνδέσεων που έχουν 'REJ' error. (τύπου Ι)	Συνεχής
same_srv_rate	Ποσοστό συνδέσεων στην ίδια υπηρεσία. (τύπου Ι)	Συνεχής
diff_srv_rate	Ποσοστό συνδέσεων σε διαφορετική υπηρεσία. (τύπου Ι)	Συνεχής
srv_count	Αριθμός συνδέσεων στην ίδια υπηρεσία με αυτή της τρέχουσας σύνδεσης τα	Συνεχής

	τελευταία 2 δευτερόλεπτα. (Τύπος II)	
srv_serror_rate	Ποσοστό συνδέσεων που έχουν 'SYN' error.(τύπου II)	Συνεχής
srv_rerror_rate	Ποσοστό συνδέσεων που έχουν 'REJ' error. (τύπου II)	Συνεχής
srv_diff_host_rate	Ποσοστό συνδέσεων σε διαφορετικό host	Συνεχής

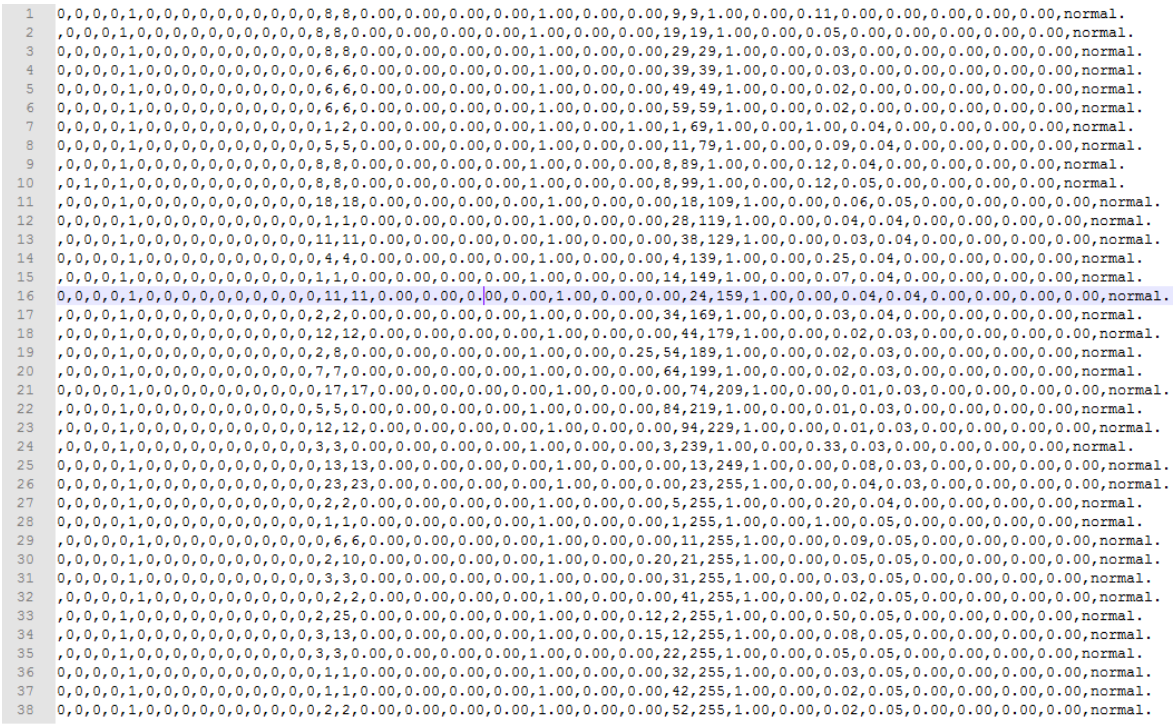
Πίνακας 4.3 Χαρακτηριστικά κίνησης στη βάση χρονικού περιορισμού 2 δευτερολέπτων

Στην παρακάτω λίστα αναφέρονται οι τύποι επιθέσεων που χρησιμοποιούνται για την εκπαίδευση των μοντέλων στο KDD99:

- back dos
- buffer_overflow u2r
- ftp_write r2l
- guess_passwd r2l
- imap r2l
- ipsweep probe
- land dos
- loadmodule u2r
- multihop r2l
- neptune dos
- nmap probe
- perl u2r
- phf r2l
- pod dos

- portsweep probe
- rootkit u2r
- satan probe
- smurf dos
- spy r2l
- teardrop dos
- warezclient r2l
- warezmaster r2l

Μία εικόνα στιγμιότυπου των δεδομένων όπως φαίνονται σε έναν editor δίνεται παρακάτω.



Εικόνα 4.4 Στιγμιότυπο των δεδομένων

Όπως είδαμε κατά την περιγραφή των δεδομένων του συνόλου KDD99, τα χαρακτηριστικά που περιλαμβάνονται, είναι προεπεξεργασμένα. Δηλαδή δεν χρησιμοποιούνται αυτούσια τα δεδομένα όπως εξήχθησαν κατά την ανάλυση τους, αλλά ένας συνδιασμός τους λαμβάνοντας υπόψην παράλληλα κάποια στατιστικά στοιχεία. Για παράδειγμα υπάρχουν χαρακτηριστικά που αναφέρονται σε ποσοστιαίες εμφανίσεις συγκεκριμένων ενεργειών και μοτίβων.

Ενώ το σύνολο αυτό με την προεπεξεργασία που έχει υποστεί είναι κατάλληλο για ερευνητικούς σκοπούς, έχει και αρκετά μειονεκτήματα που αναφέρονται στην επόμενη παράγραφο. Τα μειονεκτήματα αυτά έχουν μελετηθεί και δημοσιευτεί και έχουν προταθεί λύσεις για την βελτιστοποίηση του συνόλου.

4.4 Μειονεκτήματα

Όπως είπαμε κατά την περιγραφή του συνόλου, το KDD99 είναι το αποτέλεσμα της επεξεργασίας του συνόλου του πειράματος DARPA98. Πριν μιλήσουμε για τα μειονεκτήματα του KDD99 θα πρέπει να σημειώσουμε ότι κάποια από αυτά δημιουργήθηκαν κατά την δημιουργία του DARPA98. Συνεπώς τα προβλήματα αυτά μεταφέρονται και στο KDD99 που είναι απόρροια του πρώτου. Στην έρευνα που έκαναν οι Tavallaee, Mahbod; Bagheri, Ebrahim; Lu, Wei; Ghorbani, Ali-A, στο (A Detailed Analysis of the KDD CUP 99 Data Set). αναφέρουν για τα προβλήματα στο DARPA98 τα εξής:

1. Τα δεδομένα που συλλέχθηκαν είχαν σκοπό να προσομοιώσουν την κίνηση δεδομένων που παρατηρήθηκε σε διάφορες αεροπορικές βάσεις. Για λόγους προσωπικής ασφάλειας συνδυάστηκαν δεδομένα της κίνησης εντός δικτύου και δεδομένα από επιθέσεις .Παρόλα αυτά, αυτός ο συνδυασμός δεδομένων δεν φαίνεται να μοιάζει με αυτά που συναντάμε σε παρακολούθηση πραγματικών συνθηκών κίνησης. Αυτό έχει να κάνει με στατιστικές αναλύσεις στις επιθέσεις βάσει των συνολικών παρατηρήσεων και όχι μόνο.

2. Κατά την συλλογή στοιχείων κίνησης χρησιμοποιήθηκε το TCPdump, το οποίο σε πολλές περιπτώσεις δεν συλλέγει όλα τα πακέτα όταν υπάρχει μεγάλος φόρτος δικτύου. Συνεπώς είναι πιθανό να υπάρχει παράβλεψη επιπλέον πληροφορίας, πράγμα που αλλάζει τα σύνθετα δεδομένα και τα παραγόμενα αποτελέσματα.

3. Υπάρχουν ελλείψεις στους ορισμούς των επιθέσεων. Για παράδειγμα σε ότι αφορά το probe attack, υπάρχει συγκεκριμένο κατώφλι στον αριθμό των επαναλήψεων των ενεργειών για να θεωρηθεί επίθεση. Το ίδιο συμβαίνει και με πακέτα που προκαλούν υπερχείλιση στο buffer του διακομιστή και δεν είναι επιθέσεις. Παρόλα αυτά κατηγοριοποιούνται ως επιθέσεις.

4. Επίσης ένα άλλο ζήτημα αφορά την ταξινόμηση των επιθέσεων σε κατηγορίες. Τα περισσότερα συστήματα ανίχνευσης εισβολών αναφέρονται δυαδικά στις κατηγορίες και τις χωρίζουν σε normal και anomalous και όχι σε συγκεκριμένους τύπους επιθέσεων.

5. Κατά την έρευνα που έγινε από τους Mahoney και Chan, βρέθηκαν στοιχεία που καθιστούν το DARPA98 αναξιόπιστο σύνολο σε ότι αφορά την εκπαίδευση μοντέλων. Το αποτέλεσμα χρήσης του σε συγκεκριμένες μεθόδους μηχανικής μάθησης, οδηγεί σε υπερεκτίμηση της απόδοσής τους, αποδίδοντας έτσι πολύ υψηλά ποσοστά επιτυχούς κατηγοριοποίησης.

6. Η ανάλυση των τύπων επιθέσεων όπως περιγράφηκαν στο KDD, αφορά πέντε κατηγορίες. Παρόλα αυτά η ανάλυση του DARPA98, δείχνει τύπους επιθέσεων που δεν θα μπορούσαν να ενσωματωθούν σε κάποια από τις κατηγορίες που προτάθηκαν. Επίσης λόγω του ότι οι επιθέσεις ήταν προσομοιωμένες, υπάρχουν χαρακτηριστικά με συγκεκριμένες τιμές μόνο σε περιπτώσεις επίθεσης. Ένα παράδειγμα αφορά στις τιμές του χαρακτηριστικού TTL, στο οποίο εμφανίζονται οι τιμές 126 και 253 μόνο στα δεδομένα που αφορούν σε επιθέσεις.

Εκτός από τα προβλήματα στο DARPA98 που αναφέρονται στη συγκεκριμένη έρευνα, ιδιαίτερη σημασία έχει και η αναφορά που γίνεται σε προβλήματα που αφορούν το KDD99:

1. Κατά τον διαχωρισμό του DARPA98 σε 10 κομμάτια των 490.000 περιπτώσεων έκαστο, παρατηρήθηκε ανομοιογένεια στους τύπους επιθέσεων που περιλαμβάνονται. Συγκεκριμένα σε κάποια από τα επιμέρους κομμάτια περιλαμβάνεται μόνο ένας τύπος επίθεσης ο οποίος δεν εμφανίζεται καθόλου σε κάποιο άλλο.

2. Μία άλλη παρατήρηση, είναι πως σε ότι αφορά τις επιθέσεις τύπου neptune και smurf βρέθηκε πως απαρτίζουν το 71% του συνόλου ελέγχου των δεδομένων, πράγμα που επηρεάζει την αξιολόγηση του μοντέλου. Ακόμη, οι συγκεκριμένοι τύποι επιθέσεων δημιουργούν εξ ορισμού μεγάλο φόρτο κίνησης, πράγμα που τους κάνει εύκολα ανιχνεύσιμους χωρίς τη ανάγκη χρήσης αλγορίθμου ανίχνευσης ανωμαλιών.

3. Τέλος, ίσως το σημαντικότερο στοιχείο της συγκεκριμένης έρευνας είναι η πλεονάζουσα πληροφορία. Πολλές εγγραφές που αφορούν σε φυσιολογική κίνηση παρουσιάζονται με τους ίδιους συνδυασμούς τιμών πολλές φορές . Αυτό έχει ως αποτέλεσμα να επηρεάζονται τα μοντέλα μηχανικής μάθησης από τη συχνότητα εμφάνισης του ίδιου μοτίβου και να μην μπορούν να αναγνωρίσουν τύπους επιθέσεων που έχουν μικρότερη συχνότητα εμφάνισης, όπως οι U2R και R2L επιθέσεις.

Εξαιτίας της ύπαρξης των μειονεκτημάτων που αναφέρονται παραπάνω, έχουν γίνει προτάσεις για τη βελτίωση του dataset, οδηγώντας έτσι στη δημιουργία ενός νέου πλέον συνόλου που αναφέρεται παρακάτω.

4.5 Η δημιουργία του συνόλου KDD99+

Αν και αυτή η διατριβή δεν έχει στόχο την βελτιστοποίηση του συνόλου δεδομένων που εξετάζει, παρακάτω θα δώσουμε μία μικρή περιγραφή των μεθόδων που οδήγησαν στην δημιουργία του βελτιωμένου συνόλου KDD99+. Στην προσπάθεια βελτιστοποίησης του συνόλου δεδομένων KDD99, η έρευνα των συγγραφέων του [1] επικεντρώθηκε στο πειραματισμό με τους αλγορίθμους μηχανικής μάθησης. Για την διεξαγωγή της έρευνάς τους χρησιμοποίησαν επτά τεχνικές μηχανικής μάθησης.

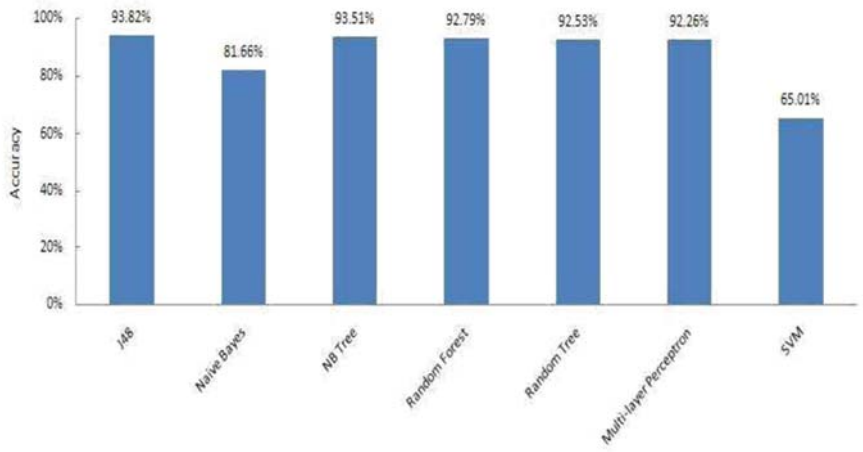
1) J48 decision tree learning
2) Naive Bayes
3) NBTree
4) Random Forest
5) Random Tree
6) Multilayer Perceptron
7) Support Vector Machine (SVM)

Στη συνέχεια έκαναν τις εξής ενέργειες:

- Χώρισαν το αρχικό set δεδομένων εκπαίδευσης σε τρία μικρότερα υποσύνολα των 50.000 εγγραφών το καθένα. Κατόπιν προχώρησαν σε εκπαίδευση μοντέλων με χρήση καθενός από τους αλγορίθμους για κάθε υποσύνολο.

- Τα συνολικά 21 μοντέλα εκπαίδευσης κλήθηκαν να συγκριθούν με τα αποτελέσματα που παρέχονται στο KDD99. Κάθε φορά που κάποιο από αυτά προέβλεπε σωστά η στήλη #successfulPrediction (από τις συνολικά 21) έπαιρνε την τιμή 1.

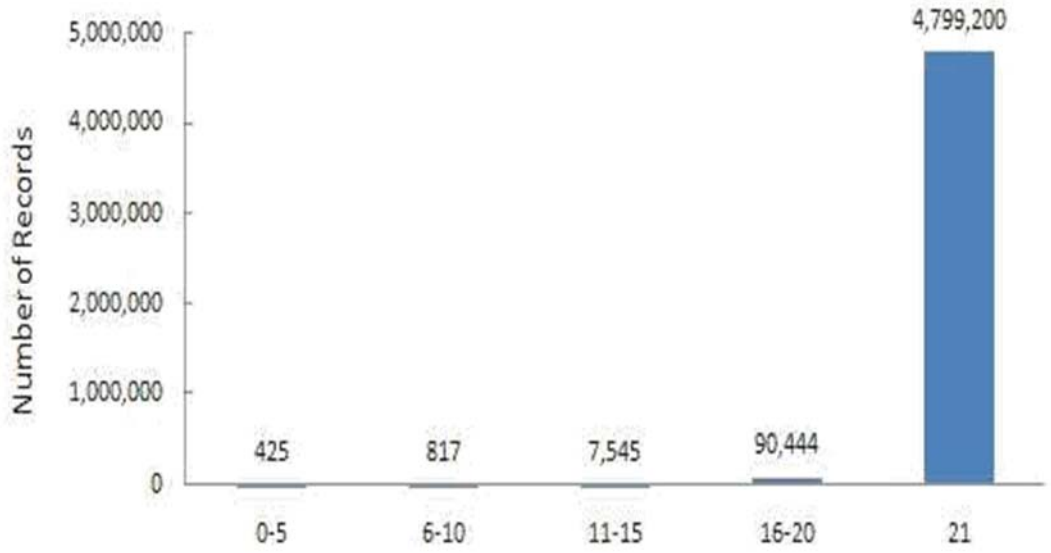
- Και τα 21 μοντέλα προέβλεψαν σωστά την κατηγορία με ποσοστά που άγγιξαν το 98% για τα train data και 87% για τα test data.



Πίνακας 4.5 Απεικόνιση της απόδοσης των αλγορίθμων επί των δεδομένων KDD Test

Όπως αναφέραμε και παραπάνω η αξιολόγηση της απόδοσης σε τυπικές μεθόδους μηχανικής μάθησης με το συγκεκριμένο σύνολο δεν μπορεί να αποφέρει χρήσιμα συμπεράσματα.

Η αξιολόγηση των 21 μοντέλων βάση των σωστών προβλέψεων στη στήλη #successfulPrediction φαίνεται στην παρακάτω απεικόνιση.



Πίνακας 4.6 Απόδοση μοντέλων επι του συνόλου KDD99 TEST

Η πρόταση της παραπάνω έρευνας είναι να επιλεχθούν συγκεκριμένες εγγραφές από το σύνολο KDD99 βάσει ενός κριτηρίου. Όπως βλέπουμε στην παραπάνω απεικόνιση το κάθε group αλγορίθμων έχει ένα συγκεκριμένο ποσοστό επιτυχούς αναγνώρισης των εγγραφών. Για παράδειγμα στο group 6-10 έχει κατηγοριοποιηθεί σωστά ένα ποσοστό 0.07% των αρχικών εγγραφών. Η πρόταση λοιπόν είναι να χρησιμοποιηθεί το ποσοστό των εγγραφών που ανήκουν στο συγκεκριμένο group 6-10 ,το οποίο είναι αντιστρόφως ανάλογο με αυτό της επιτυχούς κατηγοριοποίησης. Δηλαδή στην συγκεκριμένη ομάδα εγγραφών 6-10 θα χρησιμοποιηθεί ένα ποσοστό 99.03% (100-0.07)%. Όμοια θα επιλεχθούν και οι εγγραφές που ανήκουν στις υπόλοιπες ομάδες. Στον παρακάτω πίνακα φαίνεται η ποσοστιαία επιλογή των εγγραφών σε κάθε ομάδα στο train set.

Ομάδα	**Διακριτές εγγραφές**	**Ποσοστό**	**Επιλεγμένες εγγραφές**
0-5	407	0.04	407
6-10	768	0.07	767
11-15	6.525	0.61	6.485
16-20	58.995	5.49	55.757
21	1.008.297	93.80	62.557
Σύνολο	1.074.992	100.00	125.973

Πίνακας 4.7 Επιλογή εγγραφών για κάθε ομάδα

Μετα την επιλογή αυτή δημιουργήθηκε ένα νέο σύνολο που αποτελεί υποσύνολο του αρχικού KDD99 train set. Η ίδια λογική χρησιμοποιήθηκε και για την επιλογή εγγραφών στο σύνολο KDD99 test set. Τα δύο νέα αυτά σύνολα ονομάστηκαν KDD99$^+$ train set και KDD99$^+$ test set.

4.6 Σύνοψη Κεφαλαίου

Στο κεφάλαιο αυτό, αναλύσαμε το σύνολο KDD99 . Αναφερθήκαμε στα χαρακτηριστικά του και τους τύπους των επιθέσεων που περιλαμβάνει. Επίσης είδαμε τις αδυναμίες του συνόλου αυτού και τις προτάσεις για βελτίωση. Το συγκεκριμένο σύνολο όλα αυτά τα χρόνια έχει αποτελέσει αντικείμενο έρευνας , για το άν μπορεί να αποτελέσει αξιόπιστο dataset για την εκπαίδευση και αξιολόγηση μοντέλων ανίχνευσης ανωμαλιών. Παρόλα αυτά παραμένει μέσα στα χρόνια ένα αρκετά αξιόπιστο σύνολο, τόσο για τον όγκο του αλλά και τις συνθήκες για τις οποίες δημιουργήθηκε. Στο επόμενο κεφάλαιο, αυτό του σχεδιασμού και της εκτέλεσης του πειράματός μας, χρησιμοποιούμε το KDD99 και με μικρές αλλαγές που αφορούν στην μορφή της κατηγοριοποίησης των επιθέσεων, συνδυάζουμε τη μείωση διαστάσεων και την εκπαίδευση μοντέλου στη βάση ανίχνευσης ανωμαλιών.

Κεφάλαιο 5

Υλοποίηση Πειράματος

5.1 Εισαγωγή

Στα προηγούμενα κεφάλαια, αποκτήθηκαν όλες οι πληροφορίες που θα χρησιμοποιηθούν για τον τελικό σκοπό αυτής της διατριβής. Στο πείραμά μας θα χρησιμοποιήσουμε το σύνολο KDD99 το οποίο και αναλύσαμε σε προηγούμενο κεφάλαιο, γνωρίζοντας ότι αποτελεί ένα καλά ορισμένο και σχετικά αξιόπιστο dataset ανάλυσης επιθέσεων, παρά τις αδυναμίες που το διακρίνουν.

Σε ότι αφορά τη προεπεξεργασία του KDD99, θα χρησιμοποιηθεί ο αλγόριθμος κύριων συνιστωσών για να μειώσει τις διαστάσεις του αρχικού συνόλου. Έτσι θα καθοριστούν τα νέα principal components, μέσα από την εύρεση των λεγόμενων ακραίων τιμών (outliers) στα στοιχεία του KDD99.

Έχοντας εξηγήσει αναλυτικά τα χαρακτηριστικά και τις δυνατότητες που προσφέρει η πλατφόρμα Azure studio, θα σχεδιάσουμε και θα εκτελέσουμε το πείραμά μας το οποίο θα γίνει με τη χρήση και παραμετροποίηση των module που μας προσφέρονται.

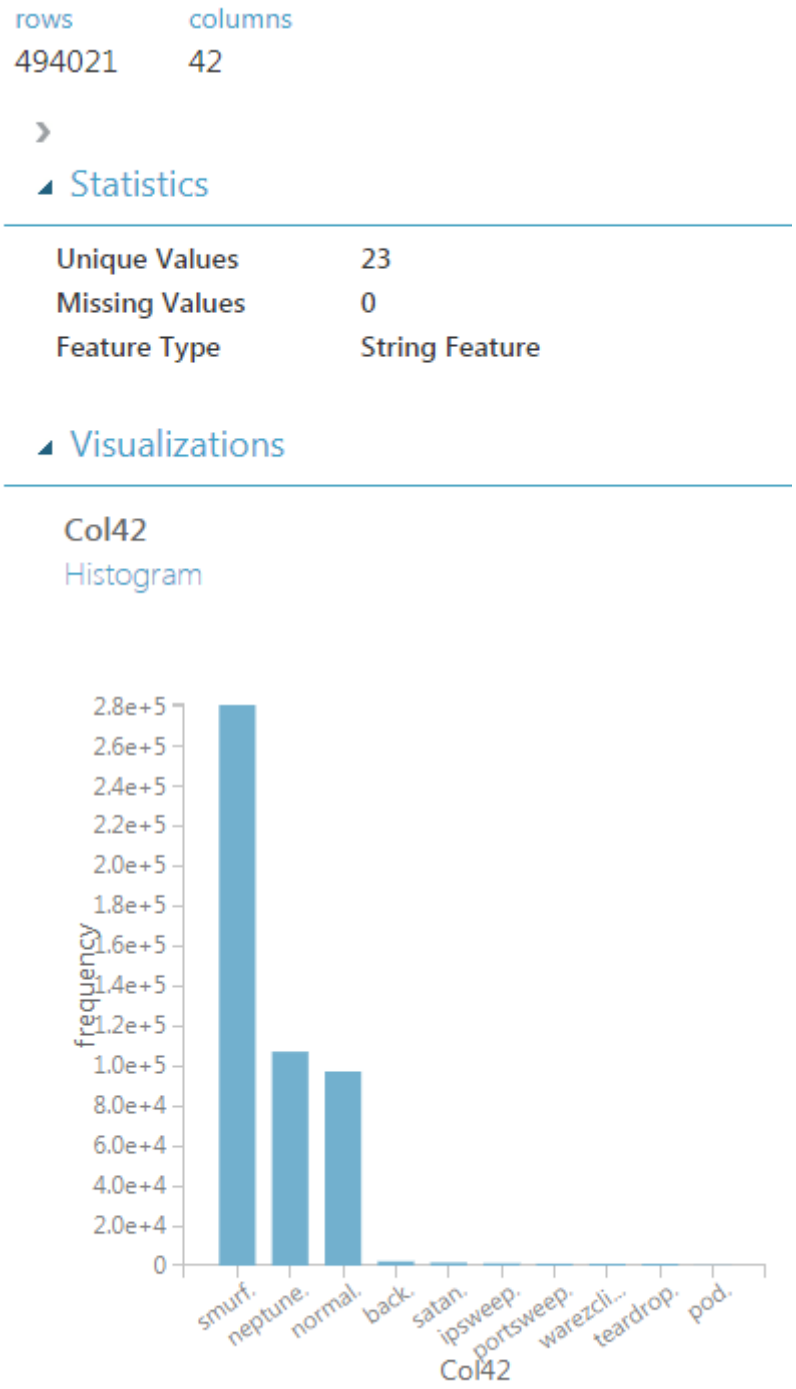
Το πειραματικό μέρος θα πραγματοποιηθεί σε δύο φάσεις για την καλύτερη κατανόησή του, αλλά και για να αναδείξουμε τις δύο χρήσεις που προσφέρει ο PCA. Στο πρώτο στάδιο θα γίνει η επιλογή χαρακτηριστικών και κατά συνέπεια η μείωση των διαστάσεων. Στο δεύτερο στάδιο θα χρησιμοποιηθεί ο PCA για Anomaly detection , δηλαδή θα εκπαιδεύσουμε ένα μοντέλο με δεδομένα φυσιολογικής (normal) κίνησης και στη συνέχεια θα παρακολουθήσουμε την συμπεριφορά του σε σύνολα που περιλαμβάνουν και δεδομένα μη φυσιολογικής (Anomaly) κίνησης.

Τα διαγραμματικά εργαλεία απεικόνισης που προσφέρονται μέσα από την πλατφόρμα, θα αναδείξουν όλα τα παραπάνω στοιχεία για την εξαγωγή χρήσιμων συμπερασμάτων.

5.2 Προεπεξεργασία Δεδομένων και Μείωση Διαστάσεων

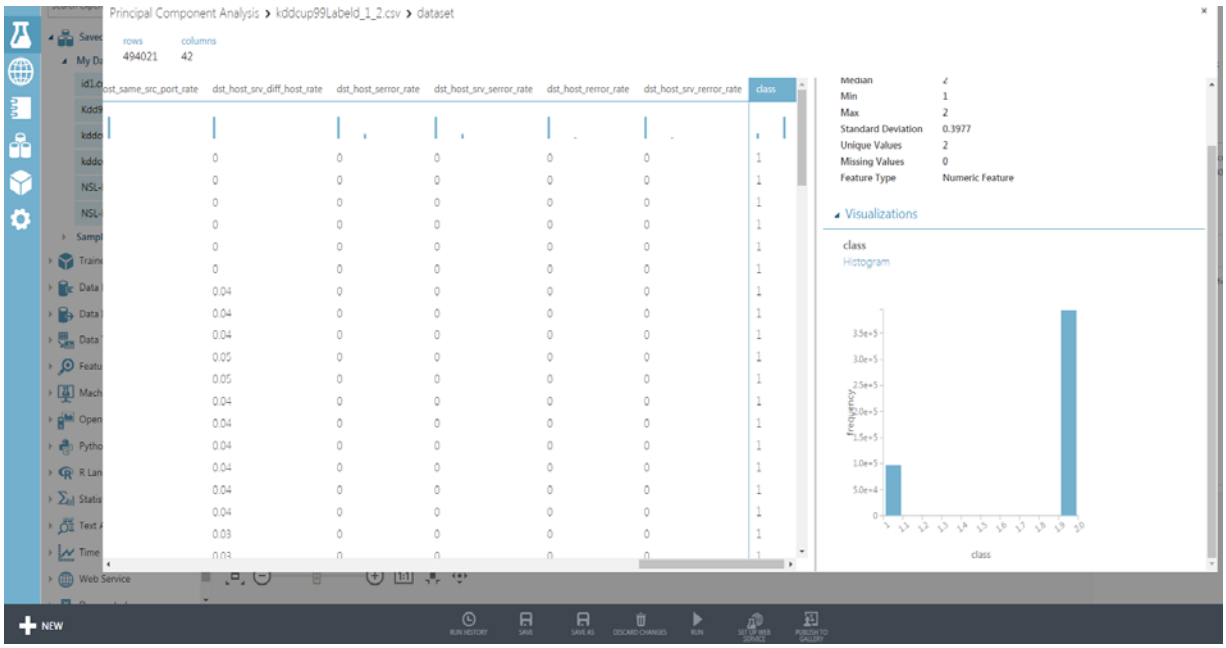
Το dataset KDD99 στη στήλη με τις κατηγορίες των επιθέσεων περιλαμβάνει και την κατηγορία με τα δεδομένα φυσιολογικής κίνησης με την τιμή normal. Αυτό που θα κάνουμε στο πείραμά μας θα είναι ανίχνευση ανωμαλιών στο συγκεκριμένο dataset και όχι κατηγοριοποίηση επιθέσεων. Με άλλα λόγια δεν μας ενδιαφέρει ο τύπος της επίθεσης , αλλά το να μπορεί να κατηγοριοποιηθεί ως επίθεση. Συνεπώς θα χρειαστεί να κάνουμε μία μικρή αλλαγή στο σύνολο και να συμπεριλάβουμε όλους τους τύπους των επιθέσεων σε μία κατηγορία.

Στο παρακάτω στιγμιότυπο μπορούμε να δούμε επιλέγοντας την στήλη Class, κάποια από τα ονόματα των επιθέσεων. Για τις ανάγκες του πειράματος μας θα συμπεριλάβουμε όλους τους τύπους επιθέσεων στην κλάση Anomaly.



Εικόνα 5.1 Ιστόγραμμα επιθέσεων

Ενώ στο επόμενο στιγμιότυπο μπορούμε να δούμε το ιστόγραμμα όπως έχουν διαμορφωθεί οι τελικές δύο κλάσεις, οπου η κλάση normal έχει τιμή 1 ενώ η anomaly 2.

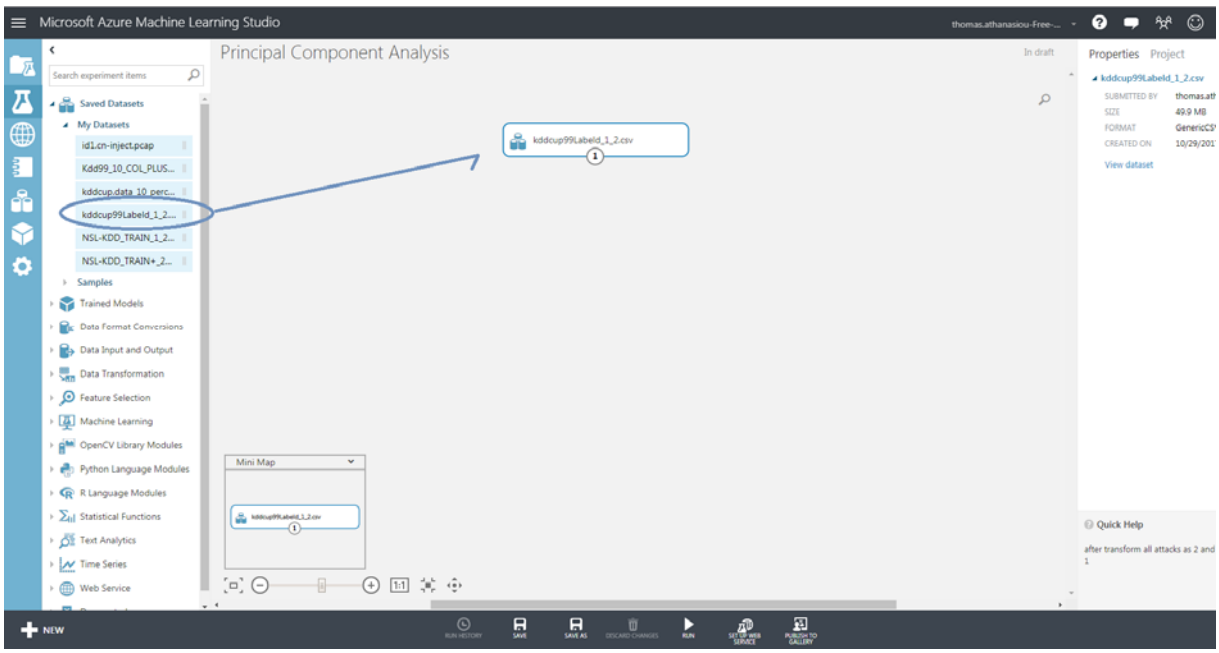


Εικόνα 5.2 Το KDD99 με δύο κατηγορίες

Έχοντας πλέον ορίσει το σύνολο έτσι ώστε να ταιριάζει στις προδιαγραφές του πειράματος μας, μπορούμε να ξεκινήσουμε το πρώτο μέρος το οποίο έχει να κάνει με την μείωση των διαστάσεων και την επιλογή των χαρακτηριστικών τους.

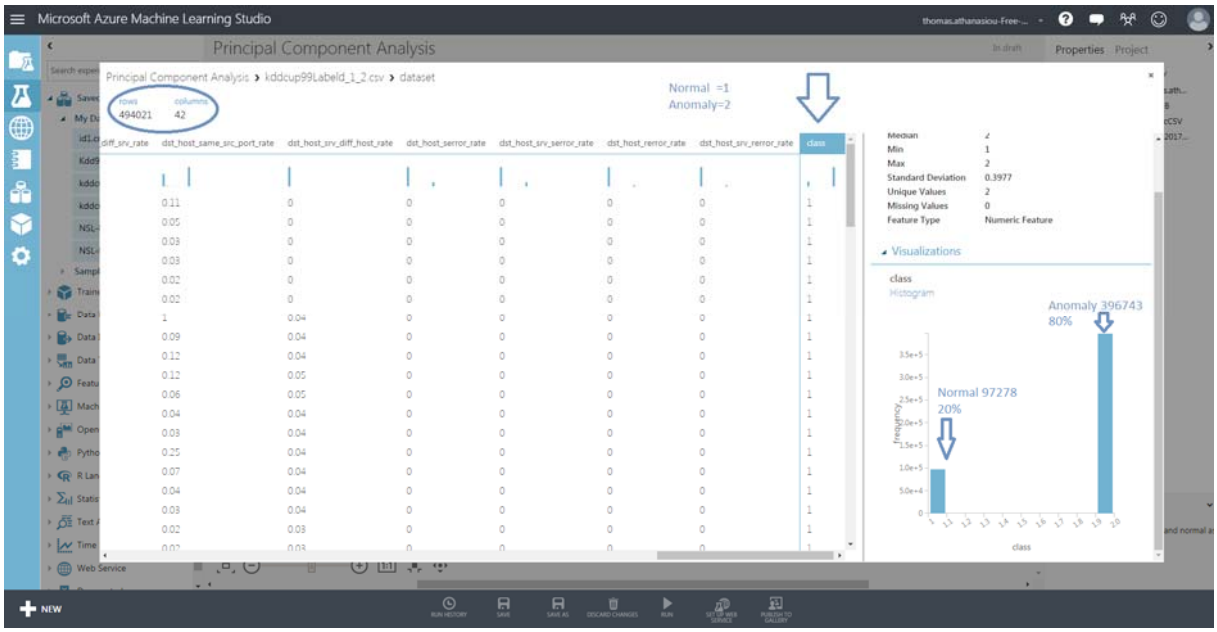
Βήμα 1ο Εισαγωγή του Dataset στο Πείραμα

Από το menu επιλογών αριστερά, βρίσκουμε τη καρτέλα my datasets και με drag & drop εισάγουμε το σύνολο στο καμβά. Το νέο αυτό σύνολο, το ονομάσαμε kddCup99Labeled_1_2 και είναι τύπου csv (comma separated values).



Εικόνα 5.3 Εισαγωγή του συνόλου KDD99

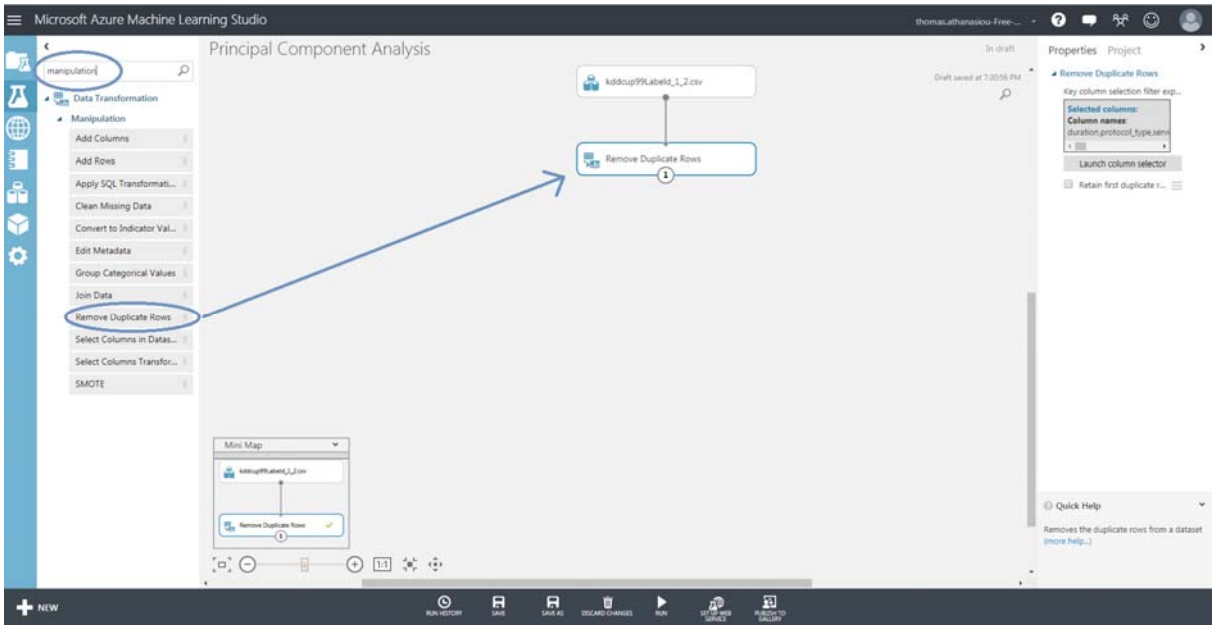
Με δεξί κλικ πάνω στο σύνολο και την επιλογή visualize, μπορούμε να δούμε τις στήλες με τα στοιχεία του όπως είναι πλέον. Πάνω στην εικόνα σημειώνονται και τα στατιστικά στοιχεία του συνόλου, με τα normal δεδομένα να αποτελούν το 20% και το υπόλοιπο 80% να αφορά το σύνολο των επιθέσεων.



Εικόνα 5.4 Οπτικοποίηση των χαρακτηριστικών του συνόλου

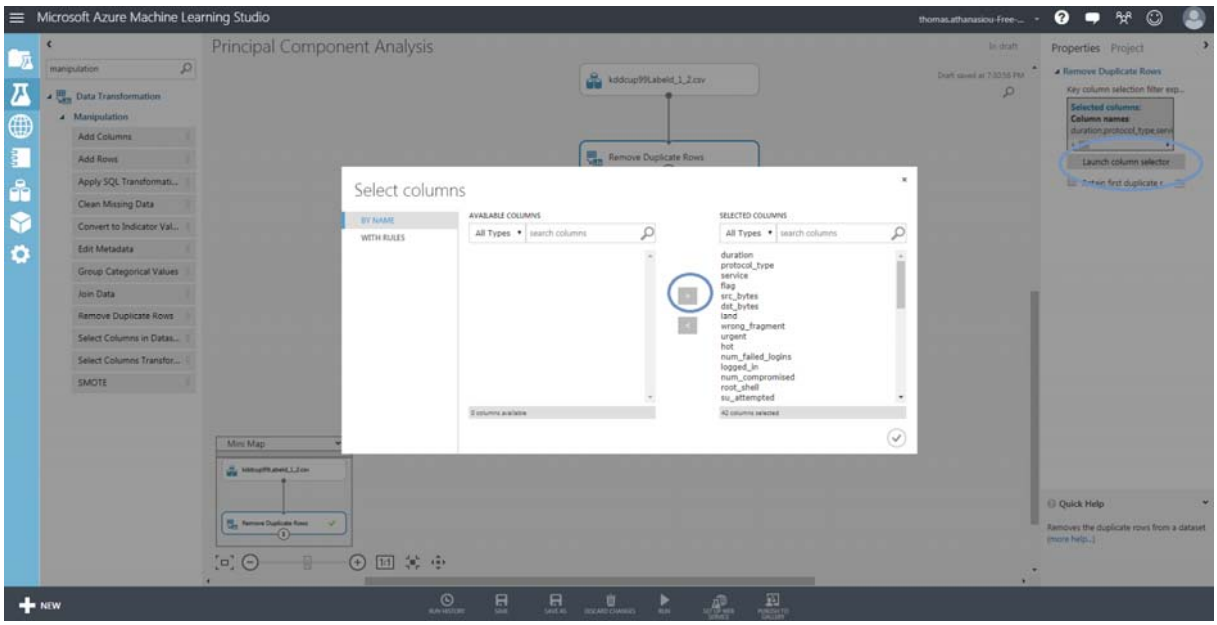
Βήμα 2ο Αφαίρεση διπλών εγγραφών

Όπως είδαμε από τα στατιστικά του συνόλου, οι εγγραφές των επιθέσεων είναι πολύ περισσότερες από αυτές της φυσιολογικής κίνησης(80%-20%). Για να σιγουρευτούμε πως δεν συμπεριλαμβάνουμε διπλότυπα δεδομένα στο πείραμα μας, θα εισάγουμε το module **Remove Duplicate Rows** από την καρτέλα **manipulation.**



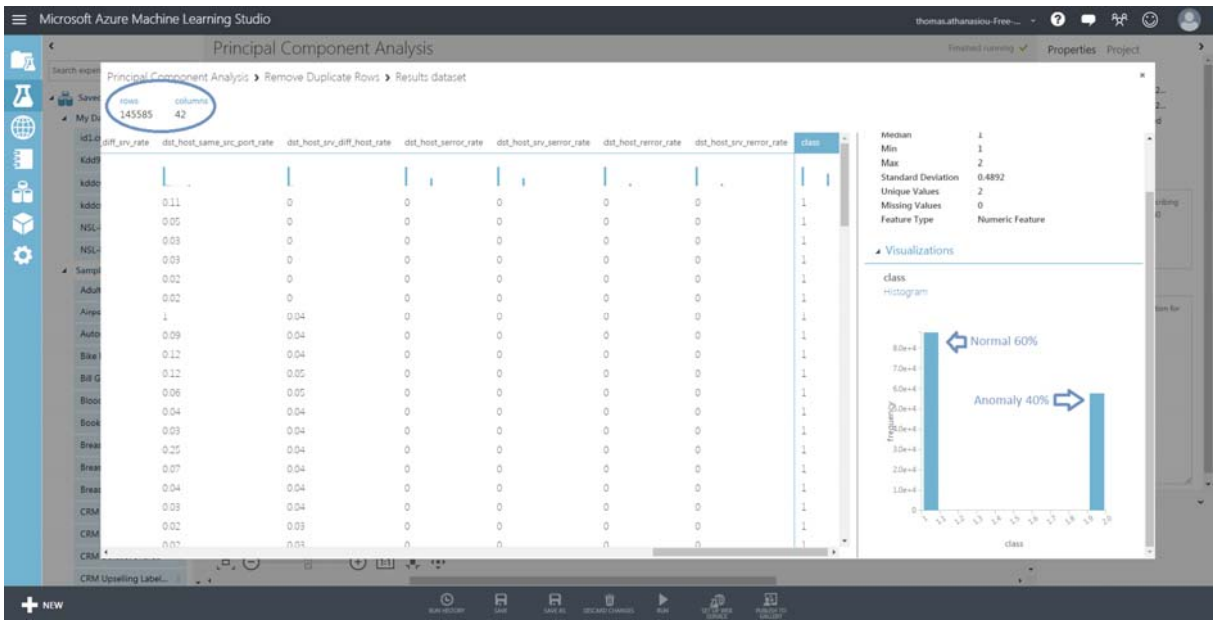
Εικόνα 5. 5 Αφαίρεση διπλών εγγραφών

Στη συνέχεια επιλέγουμε το module και πατάμε στις επιλογές των παραμέτρων του στο δεξί menu. Επιλέγουμε **Launch Column Selector** για να επιλέξουμε τα χαρακτηριστικά βάσει των οποίων θα ορίζεται μία διπλή εγγραφή. Οι επιλογές που θα κάνουμε θα είναι ο συνδυασμός όλων των στηλών, δηλαδή θα εξαιρέσουμε τις εγγραφές που έχουν όλες τις στήλες κοινές μεταξύ τους. Η εξαίρεση θα αφορά την αφαίρεση όλης της εγγραφής από το σύνολο.



Εικόνα 5.6 Επιλογή των στηλών που συμμετέχουν στον έλεγχο

Μετά από τις επιλογές, πατάμε Run στο πείραμα μας για να περάσουν οι αλλαγές από το σύνολο (πρώτο module)στην αφαίρεση των εγγραφών (δεύτερο module). Μετά το τέλος της εκτέλεσης μπορούμε να πατήσουμε πάνω του και να δούμε τις εγγραφές που απομένουν. Το αρχικό σύνολο περιλάμβανε 492021 εγγραφές με 20% normal δεδομένα και 80% anomaly (εικόνα 5.4), ενώ το νέο σύνολο περιλαμβάνει πλέον 145585 εγγραφές με 60% normal και 40% anomaly.

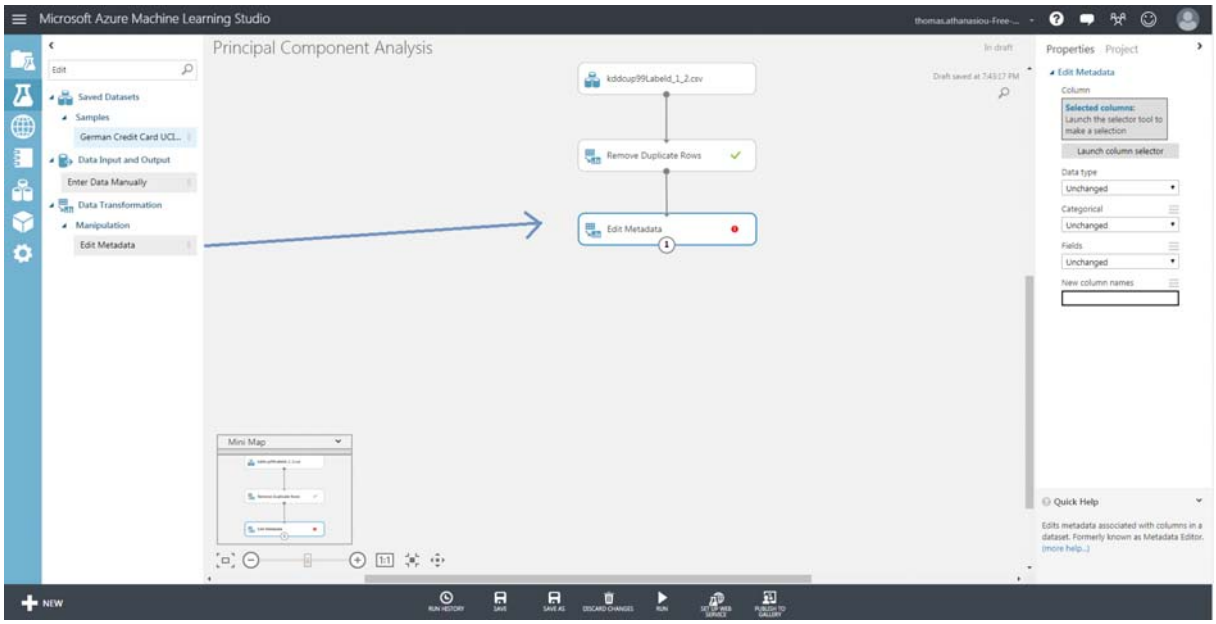


Εικόνα 5.7 Το νέο σύνολο μετά από την αφαίρεση διπλών εγγραφών

Βήμα 3ο Ορισμός της στήλης κατηγοριών ως ετικέτα

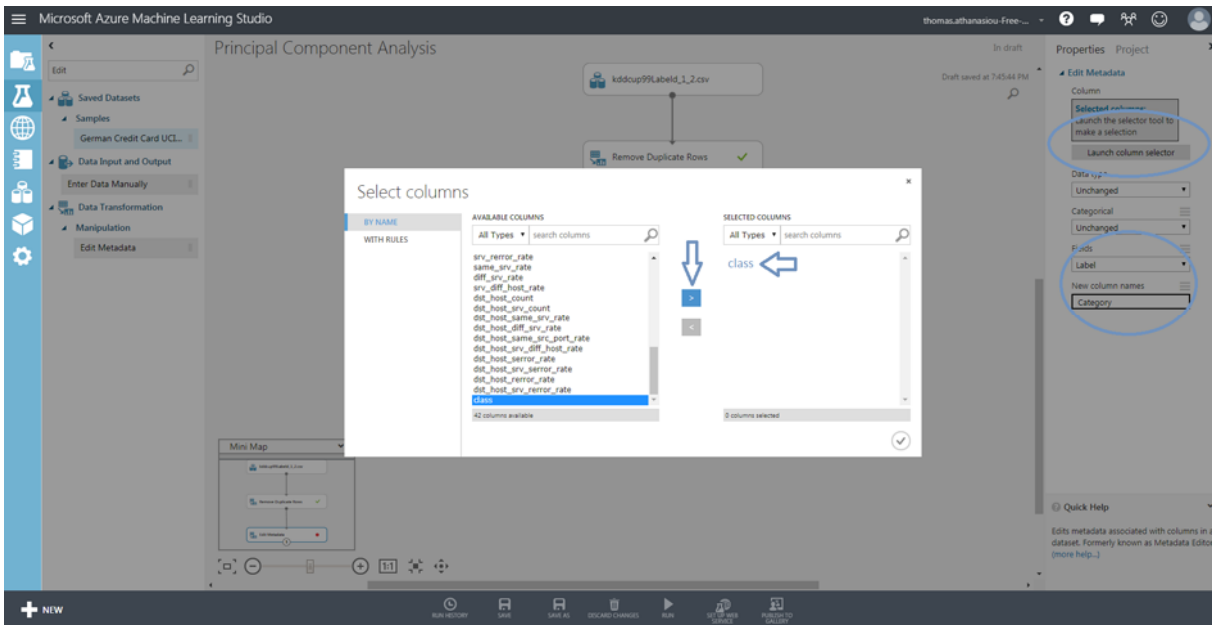
Ένα ακόμη βήμα, πριν προχωρήσουμε στην ανάλυση των κύριων συνιστωσών, είναι να χαρακτηρίσουμε τη στήλη που περιλαμβάνει την κατηγοριοποίηση των δεδομένων ως ετικέτα για να εξαιρεθεί από την επεξεργασία με τον αλγόριθμο κύριων συνιστωσών. Όπως έχουμε εξηγήσει ο PCA δέχεται ως είσοδο στήλες με χαρακτηριστικά. Στη συνέχεια αναλύει τις στήλες στις νέες διαστάσεις που απαιτούμε . Η στήλη με την κατηγοριοποίηση δεν αποτελεί χαρακτηριστικό για να συμπεριληφθεί σε κάποιο από τα παραγόμενα principal components και την χρειαζόμαστε αυτούσια για την εκπαίδευση του μοντέλου.

Έτσι σε αυτό το βήμα του πειράματος, θα χρησιμοποιήσουμε το module **Edit MetaData** για να ορίσουμε ως **Label** την στήλη **Class** και να μην συμμετέχει στην διαδικασία ανάλυσης.



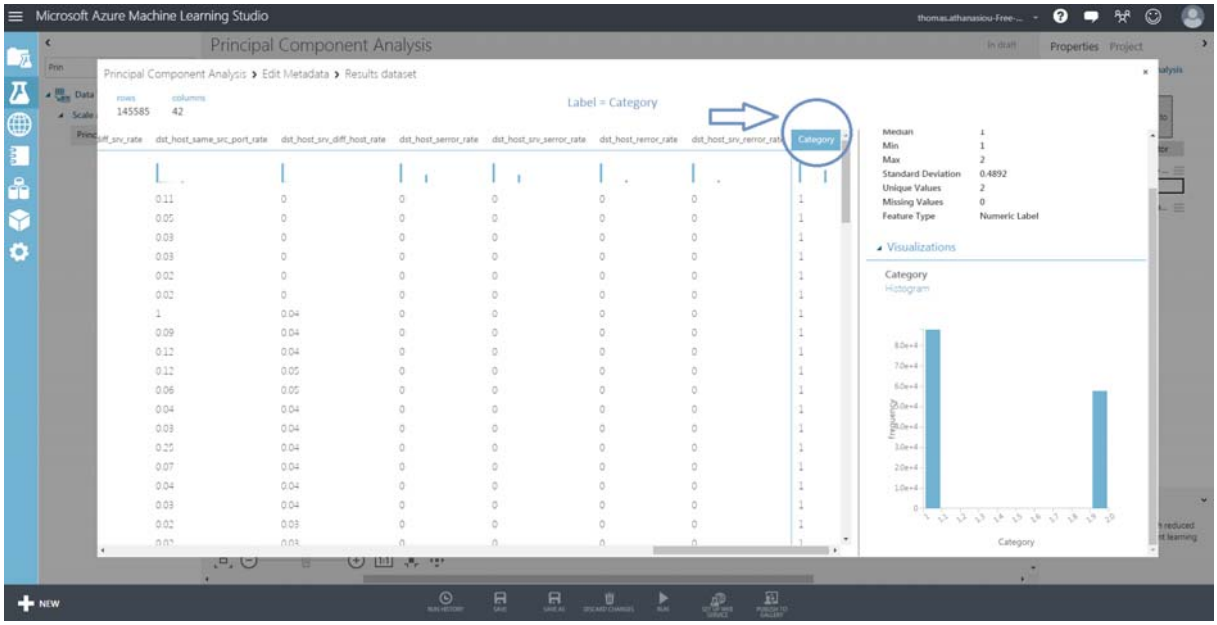
Εικόνα 5.8 Εισαγωγή του module Edit MetaData

Επιλέγουμε την ρύθμιση παραμέτρων του module και χαρακτηρίζουμε την στήλη Class ως Label.



Εικόνα 5.9 Ορισμός ως Label της στήλης Class με την ονομασία Category

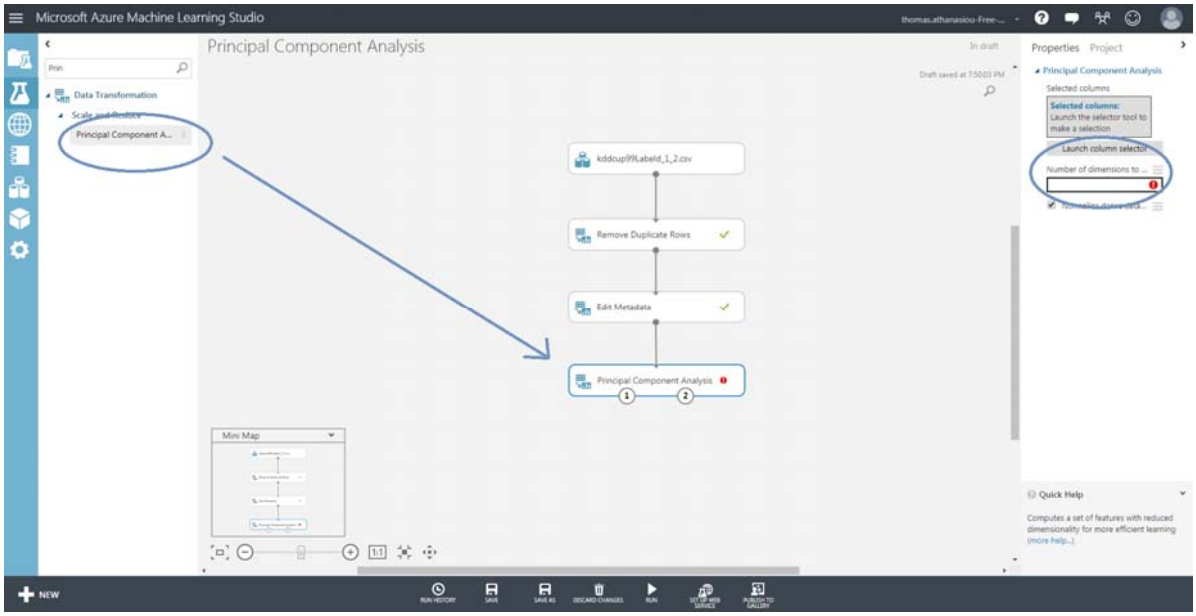
Αφού έχουμε ορίσει την στήλη **Class** ώς Label και την μετονομάζουμε σε **Category,** πατάμε **Run**. Παρακάτω μπορούμε να δούμε τα αποτελέσματα στη στήλη **Category**.



Εικόνα 5.10 Αποτελέσματα του module

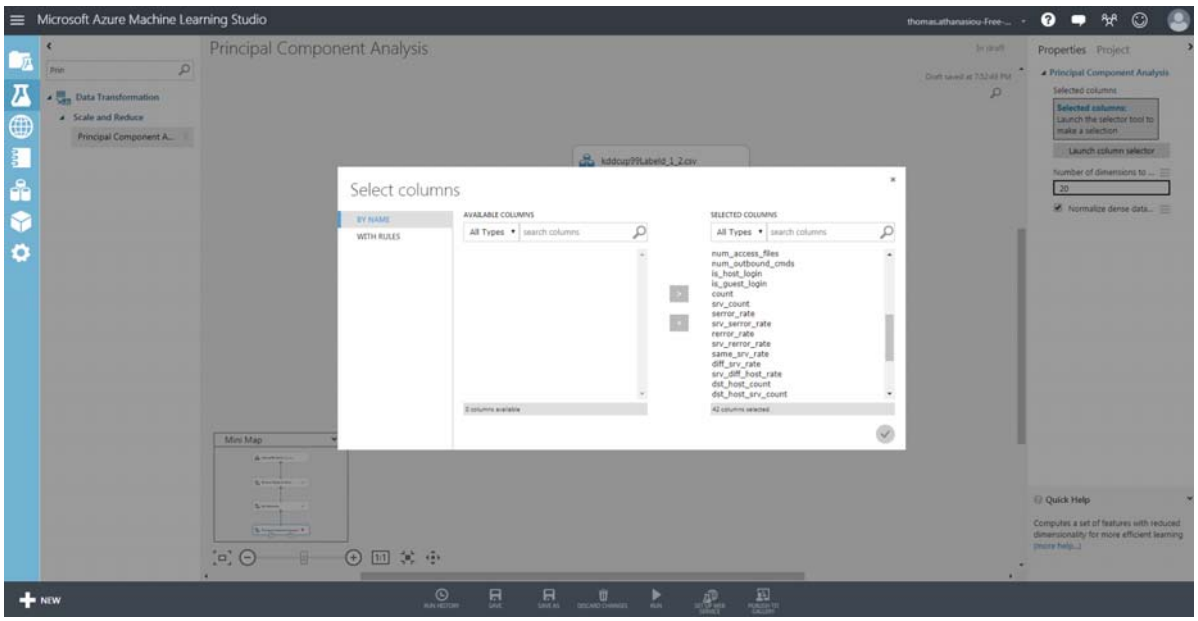
Βήμα 4ο Αλγόριθμος ανάλυσης κύριων συνιστωσών

Το επόμενο βήμα είναι η ανάλυση σε κύριες συνιστώσες. Αυτό που θα κάνουμε είναι να δώσουμε ως είσοδο στον αλγόριθμο PCA το dataset, όπως έχει διαμορφωθεί μετά από την επεξεργασία των προηγούμενων βημάτων καθώς και να ορίσουμε των αριθμό των διατσάσεων που θέλουμε ως έξοδο. Η επιλογή του module γίνεται από την καρτέλα **Data transformation** και την επιλογή **Scale and reduce Principal Component Analysis**.



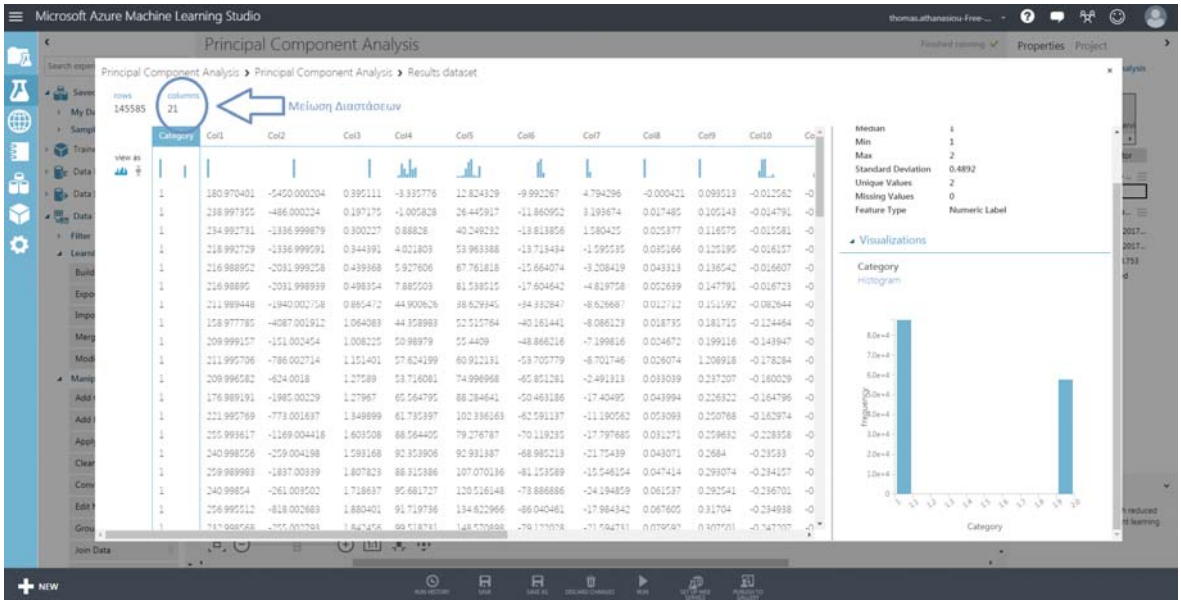
Εικόνα 5.11 Εισαγωγή Αλγορίθμου PCA

Στη συνέχεια γίνεται η επιλογή των στηλών , όπου στην περίπτωση μας θα τις επιλέξουμε όλες και στην επιλογή των παραμέτρων του module δεξιά θα ορίσουμε 20 τελικές στήλες που θα αποτελέσουν τα principal components της διαδικασίας.



Εικόνα 5.12 Επιλογή χαρακτηριστικών και ορισμός τελικών διαστάσεων

Όπως θα δούμε στα αποτελέσματα, έχουμε πλέον ένα νέο σύνολο 20 χαρακτηριστικών. Παρατηρώντας περισσότερο βλέπουμε, πως ενώ ορίσαμε 20 τελικές στήλες, το αποτέλεσμα μας θα είναι 21 , μαζί με την στήλη της κατηγοριοποίησης την οποία δώσαμε ως ετικέτα(Label) στον PCA.



Εικόνα 5.13 Το σύνολο των Principal Components

Το νέο πλέον σύνολο, περιέχει τις στήλες με τα αποτελέσματα από την επεξεργασία του αλγόριθμου ανάλυσης κύριων συνιστωσών και μπορεί να χρησιμοποιηθεί για εκπαίδευση κάποιου μοντέλου αναγνώρισης μοτίβων.

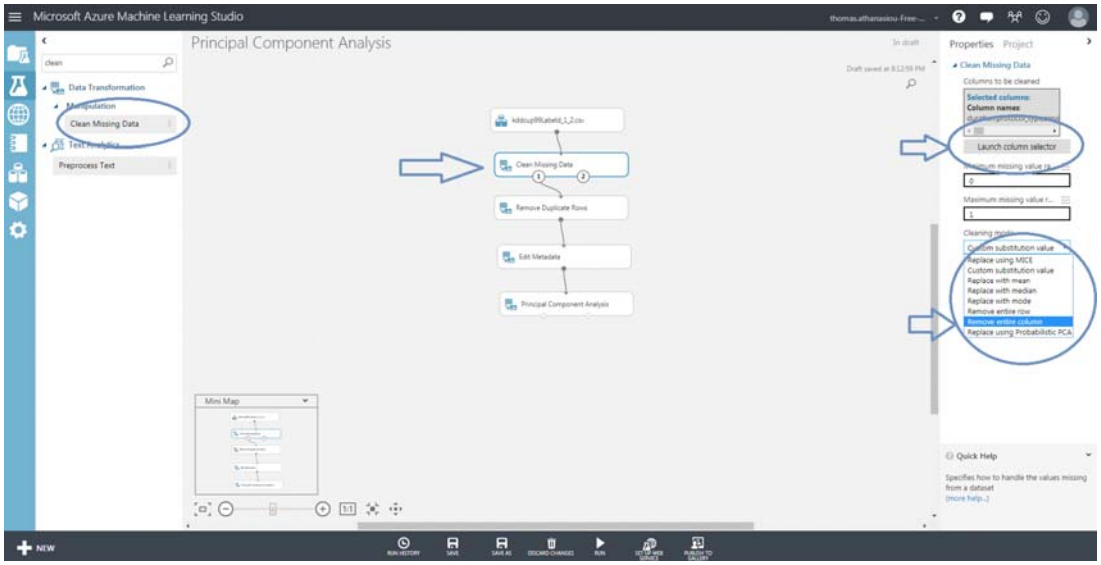
Αφού δείξαμε το μέρος του πειράματος που αφορά στη μείωση των διαστάσεων, μπορούμε να κάνουμε όσες παραλλαγές επιθυμούμε σε ότι αφορά τα principal components που θέτουμε ως έξοδο από τον αλγόριθμο.

5.3 Ανίχνευση Ανωμαλιών

Με βάση το σύνολο που έχουμε διαμορφώσει ήδη, θα εκπαιδεύσουμε ένα μοντέλο για να ανιχνεύει τα anomaly data στις εγγραφές. Η εκπαίδευση του μοντέλου θα γίνει εισάγοντας του τα normal δεδομένα. Τα αρχικά module που εισάγουμε δεν διαφοροποιούνται σε σχέση με το προηγούμενο πείραμα με τη διαφορά όμως ότι γίνεται και προσθήκη ενός ενδιάμεσου βήματος για να εξαιρεθούν οι γραμμές που έχουν ασυμπλήρωτα στοιχεία.

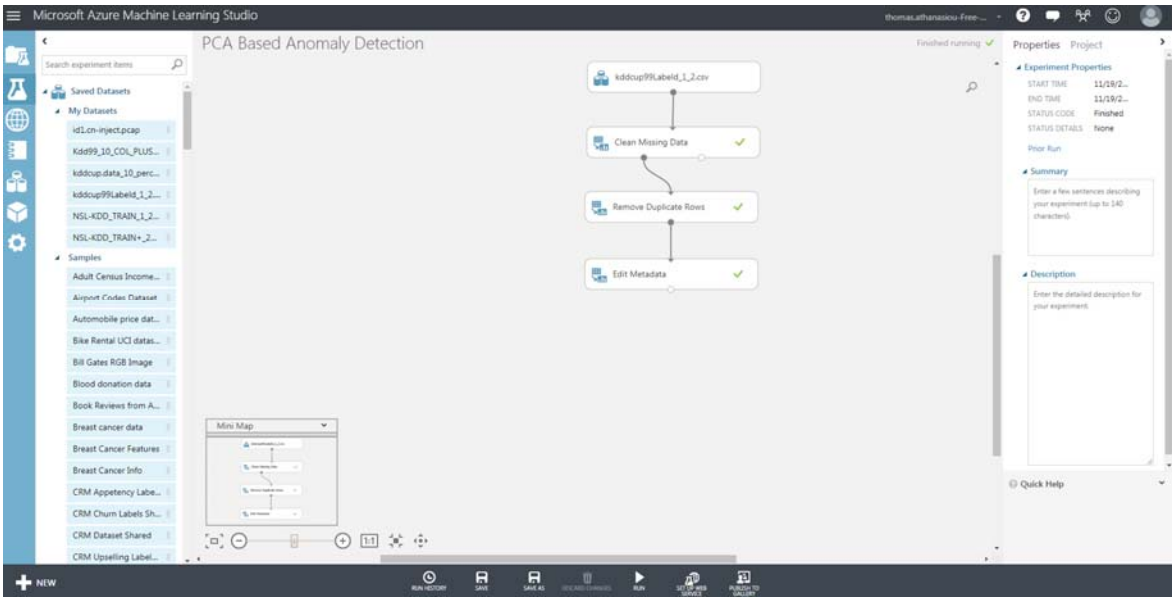
Βήμα 1ο Αφαίρεση εγγραφών με χαρακτηριστικά χωρίς τιμή

Το βήμα αυτό πραγματοποιείται με την χρήση του module **Clean missing Data** από την καρτέλα **Data Transformation** και την υποκατηγορία **Manipulation.**



Εικόνα 5.14 Αφαίρεση εγγραφών με κενά χαρακτηριστικά

Μετά την εισαγωγή των πρώτων module το πείραμά έχει την εξής μορφή:



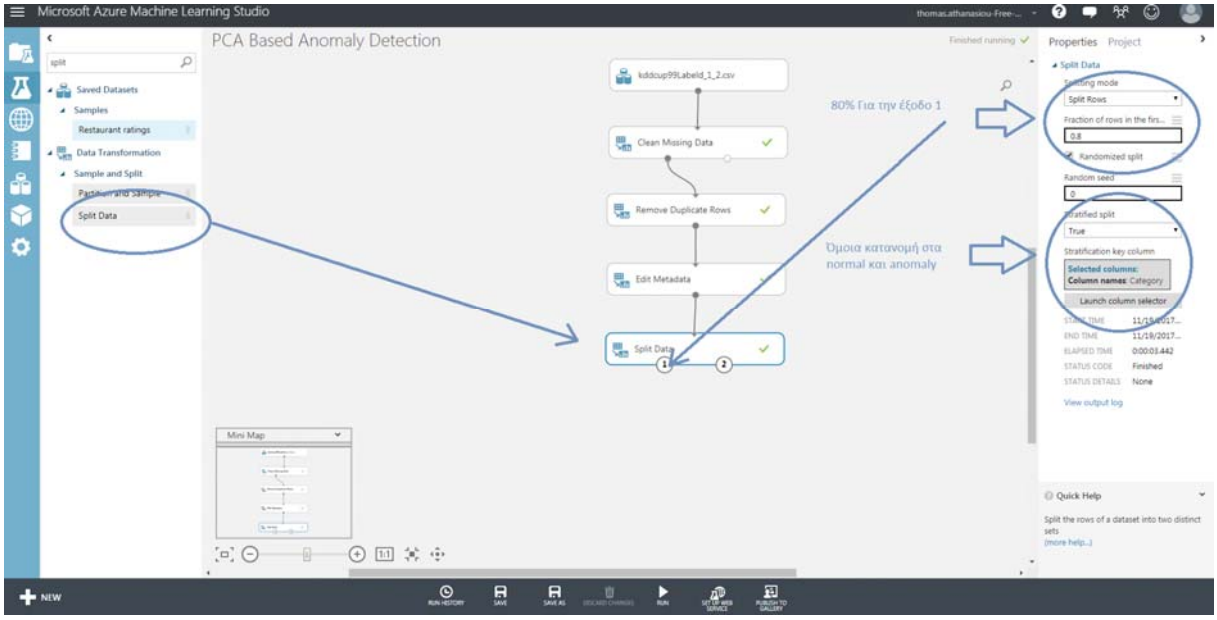
Εικόνα 5.15 Πείραμα PCA Based Anomaly Detection

Βήμα 2ο Διαχωρισμός συνόλου σε υποσύνολα εκπαίδευσης και ελέγχου

Σε αυτό το σημείο, που τα δεδομένα είναι έτοιμα για εισαγωγή σε κάποιο μοντέλο εκπαίδευσης, θα πρέπει να γίνει και ο διαχωρισμός τους. Αυτό σημαίνει πως θα χωρίσουμε το σύνολο των δεδομένων σε ένα υποσύνολο που θα χρησιμοποιηθεί για εκπαίδευση, και σε ένα υποσύνολο για να αξιολογήσουμε την συμπεριφορά του. Αυτή η μέθοδος χρησιμοποιείται κατά κόρον στα πειράματα που σχετίζονται με την μηχανική μάθηση. Για την εκτέλεση του βήματος αυτού θα χρησιμοποιήσουμε το module **Split Data,** από την καρτέλα **Data Transformation.** Οι παράμετροι που θα ρυθμίσουμε στο module από το menu επιλογών του δεξιά θα είναι οι εξής:

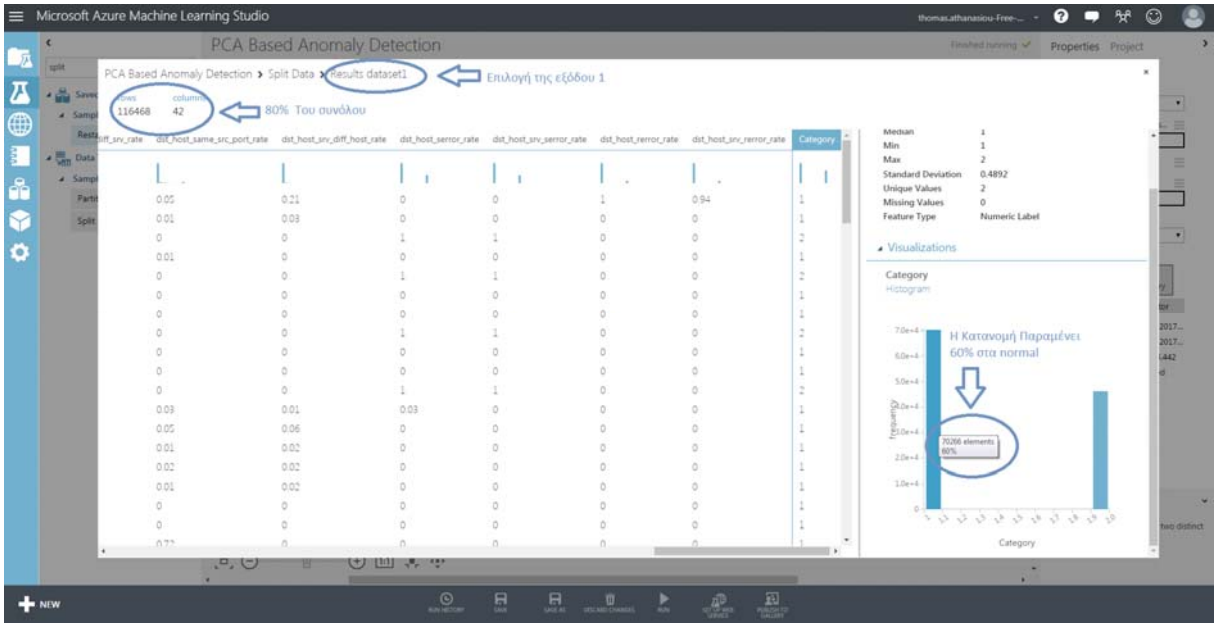
- **Fraction of rows in the first output dataset**: Εδώ θα ορίσουμε το ποσοστό των εγγραφών το οποίο θα μας δοθεί από την αριστερή έξοδο του module. Το ποσοστό αυτό θα χρησιμοποιηθεί για εκπαίδευση και αποτελεί το 80 % του συνόλου. Ενώ στη δεξιά του έξοδο(έξοδο 2), θα έχουμε το υπόλοιπο 20% του αρχικού συνόλου. Υπενθυμίζουμε πως αναφερόμαστε στο ποσοστό των εγγραφών που έχει προκύψει μετά από τον 'καθαρισμό' του από διπλές εγγραφές (τελικό σύνολο 145585 εγγραφές με 60% normal και 40% anomaly).

- **Stratified split**: Αυτή η επιλογή θα ρυθμίσει την ομοιογένεια του διαχωρισμού. Η ομοιογένεια στηρίζεται στην κατηγοριοποίηση(στήλη Category) και ο λόγος που θέτουμε ενεργή αυτή την επιλογή είναι για να διατηρήσουμε τα ποσοστά μεταξύ των normal και των anomaly (60% και 40%) εγγραφών.



Εικόνα 5.16 Διαχωρισμός των δεδομένων

Ενώ με δεξί κλικ πάνω στο module και visualize, βλέπουμε το σύνολο όπως έχει διαχωριστεί, το οποίο διατηρεί τα ποσοστά ανάμεσα στα normal και anomaly.



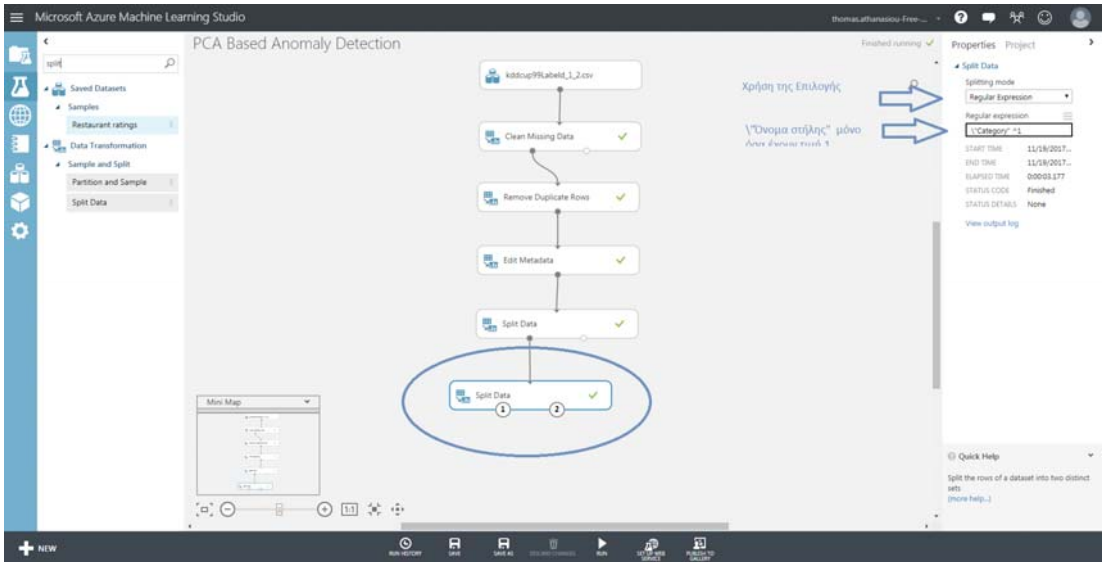
Εικόνα 5.17 Αποτελέσματα του διαχωρισμού

Βήμα 3ο Ορισμός των δεδομένων προς εκπαίδευση

Επειδή όπως έχουμε πει, στο πείραμα μας θα εκπαιδεύσουμε ένα μοντέλο να ανιχνεύει τις ανωμαλίες στο σύνολο των δεδομένων στη βάση του τί είναι φυσιολογικό, θα πρέπει να διαχωρίσουμε τα δεδομένα και να πάρουμε μόνο αυτά που ανήκουν στην κατηγορία normal. Η επιλογή αυτή των δεδομένων θα γίνει πάλι με τη χρήση του module **Split Data.** Αυτή τη φορά θα παραμετροποιήσουμε το module στα εξής στοιχεία του:

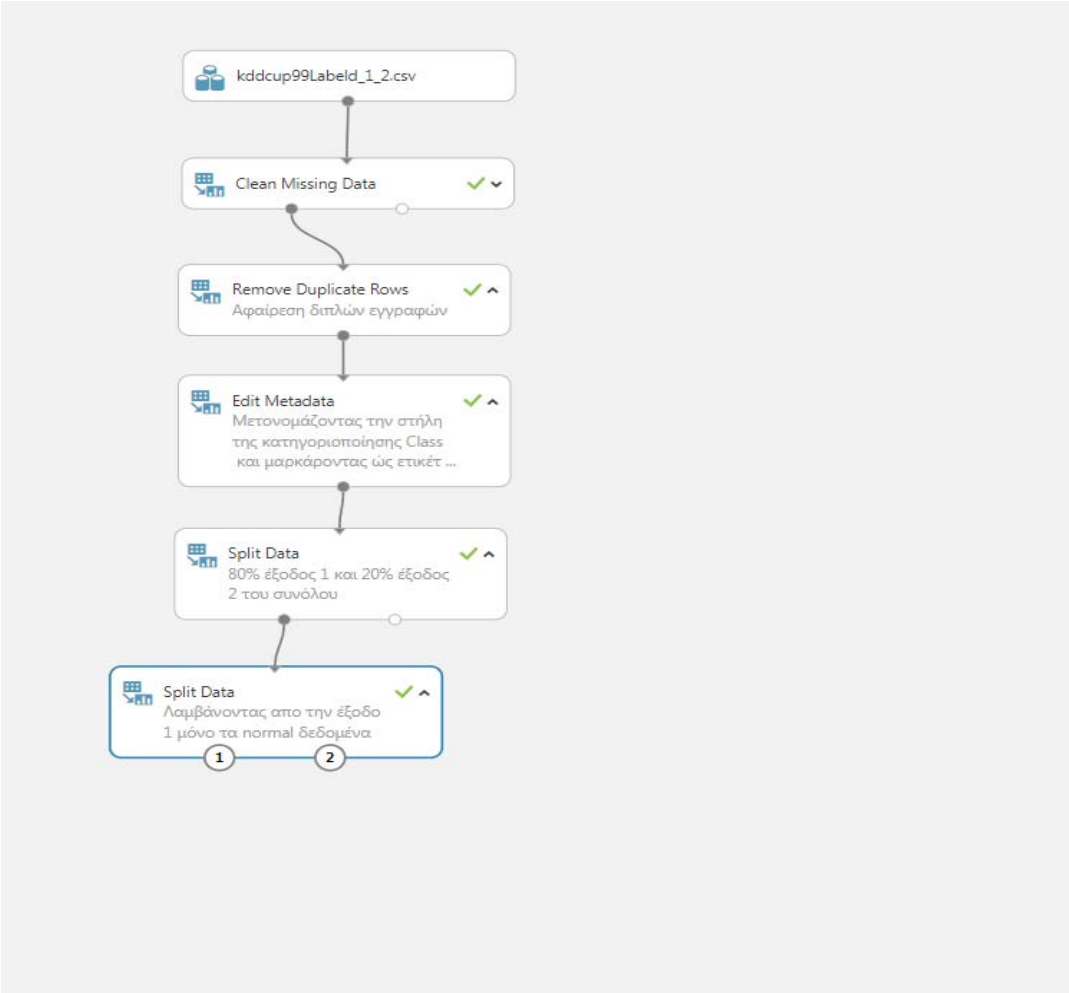
- **Splitting mode**: Ο διαχωρισμός αυτή τη φορά θα γίνει με βάση μία Regular Expression, μία έκφραση που χρησιμοποιείται συχνά π.χ. στους μεταγλωττιστές.

- **Regular expression:** Η έκφραση έχει την μορφή: \"Category" ^1, όπου (Category) το όνομα της στήλης βάσει της οποίας γίνεται ο διαχωρισμός, ενώ (^1) τα δεδομένα της στήλης που έχουν τιμή ίση με τη μονάδα (1).



Εικόνα 5.18 Διαχωρισμός των δεδομένων σε normal (τιμή 1)

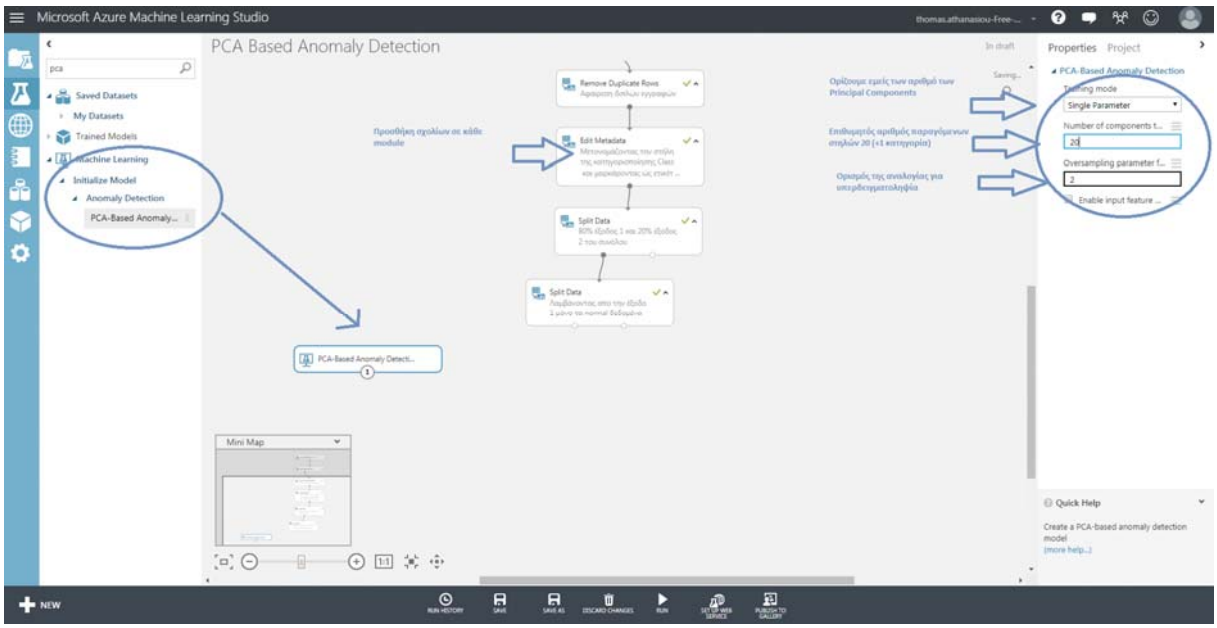
Μέχρι στιγμής το πείραμα έχει την εξής μορφή στην οποία συμπεριλαμβάνονται και σχόλια στην ετικέτα του κάθε module:



Εικόνα 5.19 Το Πείραμα μέχρι αυτό το βήμα με σχόλια στα module

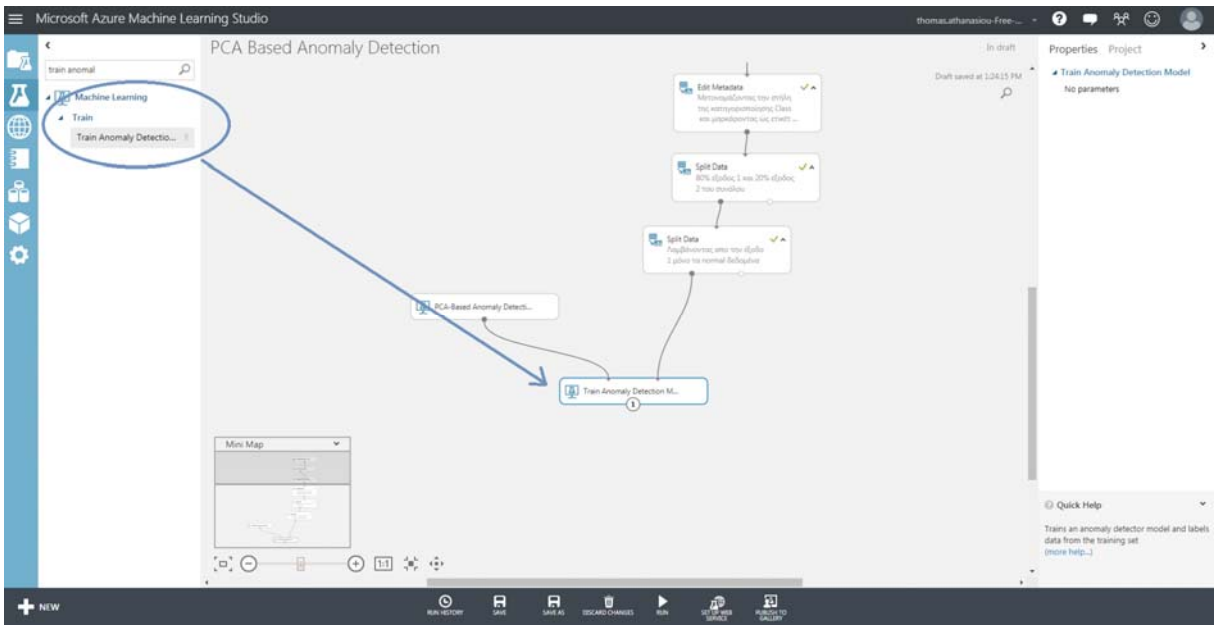
Βήμα 4ο Εισαγωγή αλγορίθμου και μοντέλου ανίχνευσης ανωμαλιών προς εκπαίδευση

Ενώ στο προηγούμενο πείραμα , χρησιμοποιήσαμε και δείξαμε την μείωση σε είκοσι διαστάσεις, με τη χρήση του module **Principal Component Analysis,** σε αυτή τη περίπτωση ο σκοπός είναι να γίνει η εκπαίδευση του μοντέλου ανίχνευσης ανωμαλιών βάσει της εύρεσης των κύριων συνιστωσών στο σύνολο. Για το σκοπό αυτό θα χρησιμοποιήσουμε το Module **PCA Based Anomaly Detection** από την καρτέλα **Machine Learning** . Για την παραμετροποίηση αυτού του module θα επιλέξουμε **Single Parameter**, δηλαδή θα ορίσουμε εμείς το σύνολο των διαστάσεων που θα έχουμε ως αποτέλεσμα. Έτσι θα επιλέξουμε είκοσι διαστάσεις περιμένοντας να πάρουμε ως έξοδο 21 μαζί με τη στήλη της κατηγοριοποίησης. Ακόμη θα επιλέξουμε στο **Oversampling parameter** τον αριθμό 2, όπου η επιλογή του αριθμού 1 θα ήταν να μην γίνει υπερδειγματοληψία. Ο ορισμός γίνεται για την αναλογία της δειγματοληψίας[7]. Ο λόγος που επιλέγουμε να εισάγουμε στην ουσία παραπάνω φορές κάποιες εγγραφές από ότι υπάρχουν, είναι για να δώσουμε μία καλύτερη αναλογία στα δεδομένα εκπαίδευσης που δίνονται στον αλγόριθμο.



Εικόνα 5.20 PCA ως αλγόριθμος εκπαίδευσης

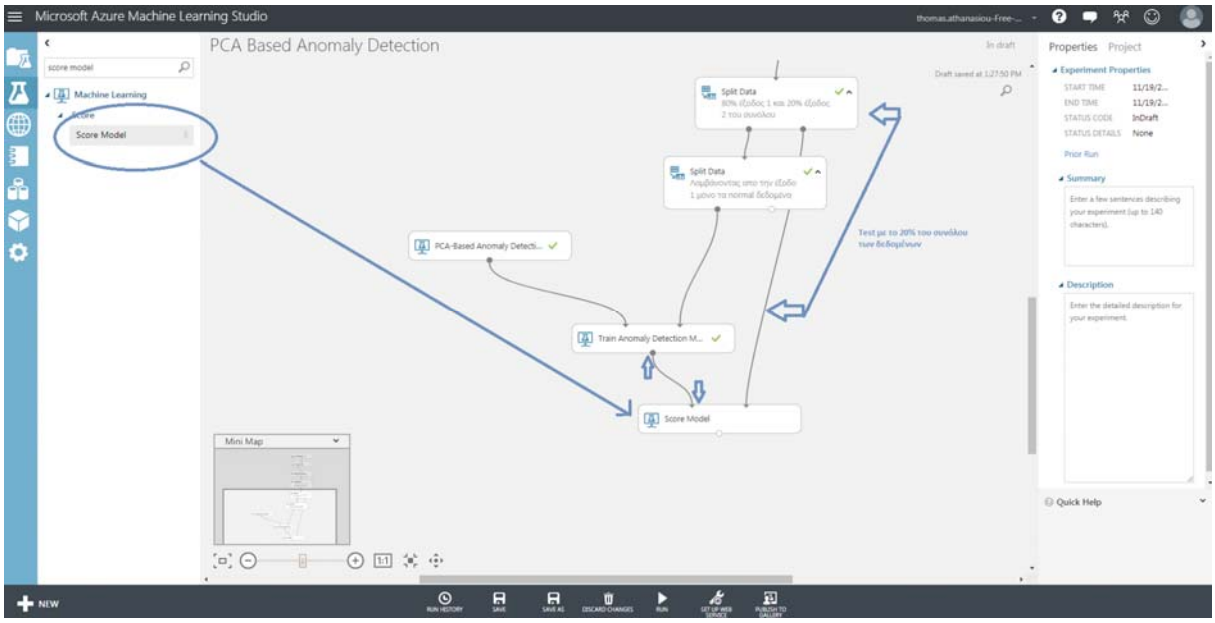
Στη συνέχεια το module **PCA Based Anomaly Detection**, θα πρέπει να συνδεθεί με ένα μοντέλο εκπαίδευσης(module τύπου train. Στη περίπτωση μας, θα κάνουμε χρήση του module **Train Anomaly Detection Model**. Από τα διαθέσιμα module επιλέγουμε **Machine Learning** και **Train.** Η σύνδεση των δύο module, γίνεται ενώνοντας το module του αλγορίθμου εκπαίδευσης PCA στην αριστερή είσοδο του Train model, ενώ το σύνολο δεδομένων που θα χρησιμοποιηθεί συνδέεται στην δεξιά του είσοδο. Υπενθυμίζουμε πως η εκπαίδευση γίνεται μόνο με τα normal δεδομένα.



Εικόνα 5.21 Εισαγωγή μοντέλου ανίχνευσης ανωμαλιών προς εκπαίδευση

Βήμα 5ο Εμφάνιση των κατηγοριοποιήσεων με χρήση του Score module

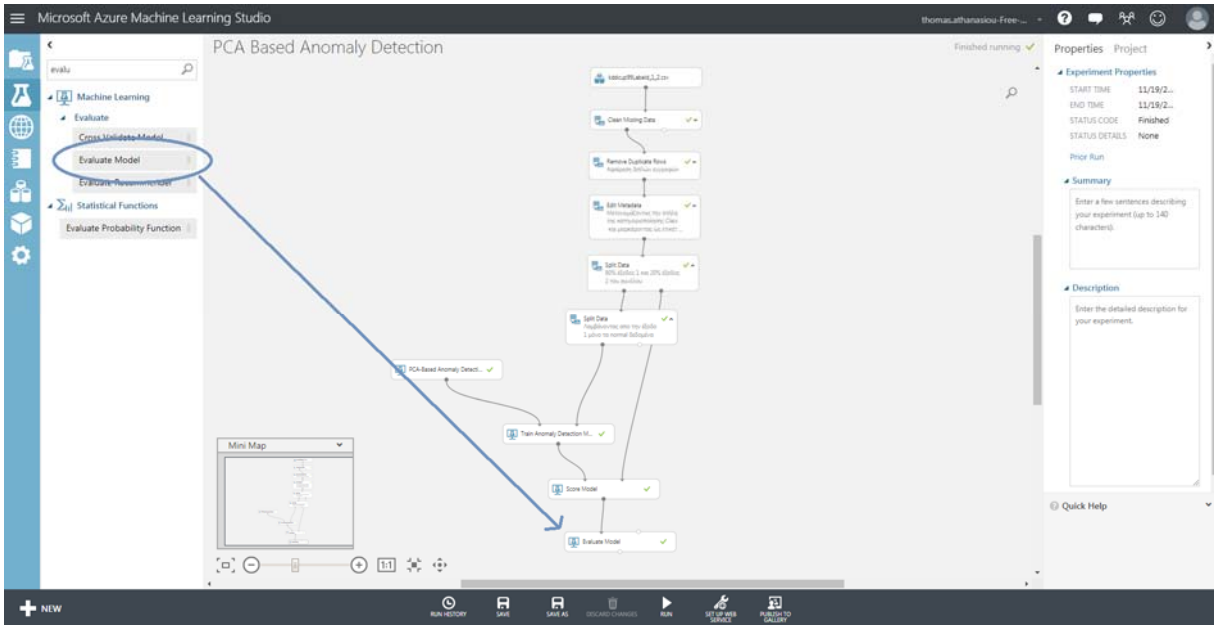
Το επόμενο βήμα είναι να δούμε τα αποτελέσματα της κατηγοριοποίησης του εκπαιδευμένου πλέον μοντέλου. Στο σημείο αυτό υπενθυμίζουμε πως χωρίσαμε το αρχικό σύνολο δεδομένων σε ποσοστό 80% για εκπαίδευση και 20% για έλεγχο. Στη συνέχεια επιλέγουμε **Score Model** από την καρτέλα **Machine Learning** και **Score.** Η σύνδεση του θα είναι μεταξύ της εξόδου του **Train model** και του αρχικού **Split Data**.



Εικόνα 5.22 Test του μοντέλου με το 20%

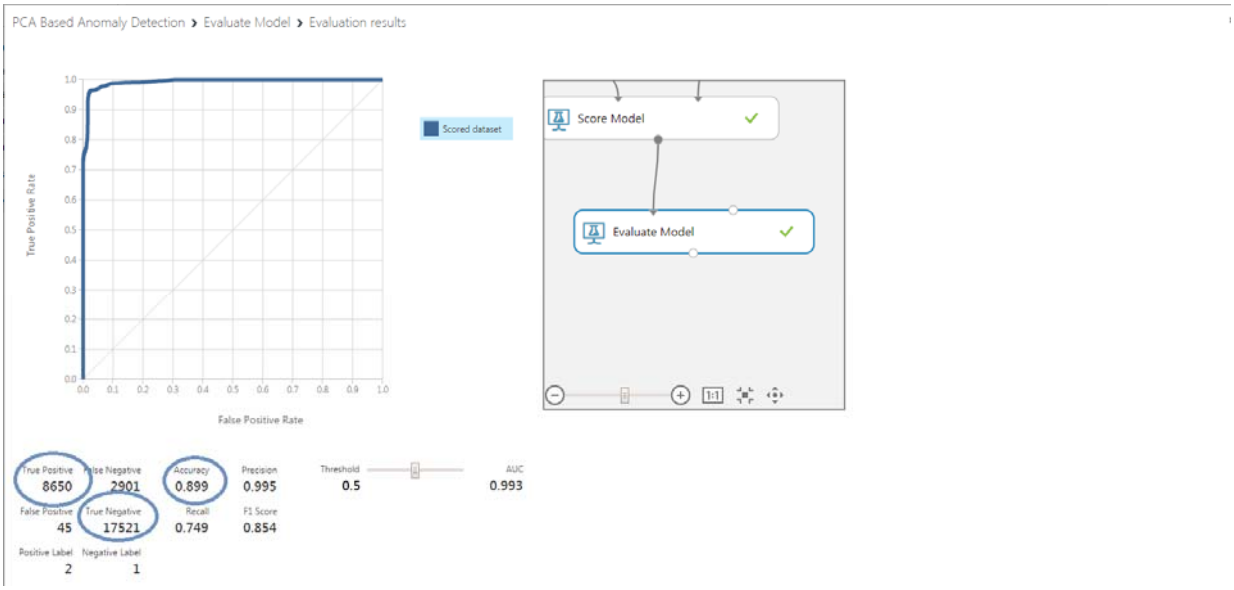
Βήμα 6ο Αξιολόγηση του μοντέλου

Έχοντας 'τρέξει' όλα τα προηγούμενα στάδια, περνάμε στη οπτικοποίηση των αποτελεσμάτων. Για να μπορέσουμε να δούμε τα διαγραμματικά εργαλεία και τα στατιστικά στοιχεία που μας προσφέρει η χρήση της πλατφόρμας, θα εισάγουμε ένα ακόμα module. Το module αυτό είναι το **Evaluate Model** από την καρτέλα **Machine Learning** και **Evaluate.**



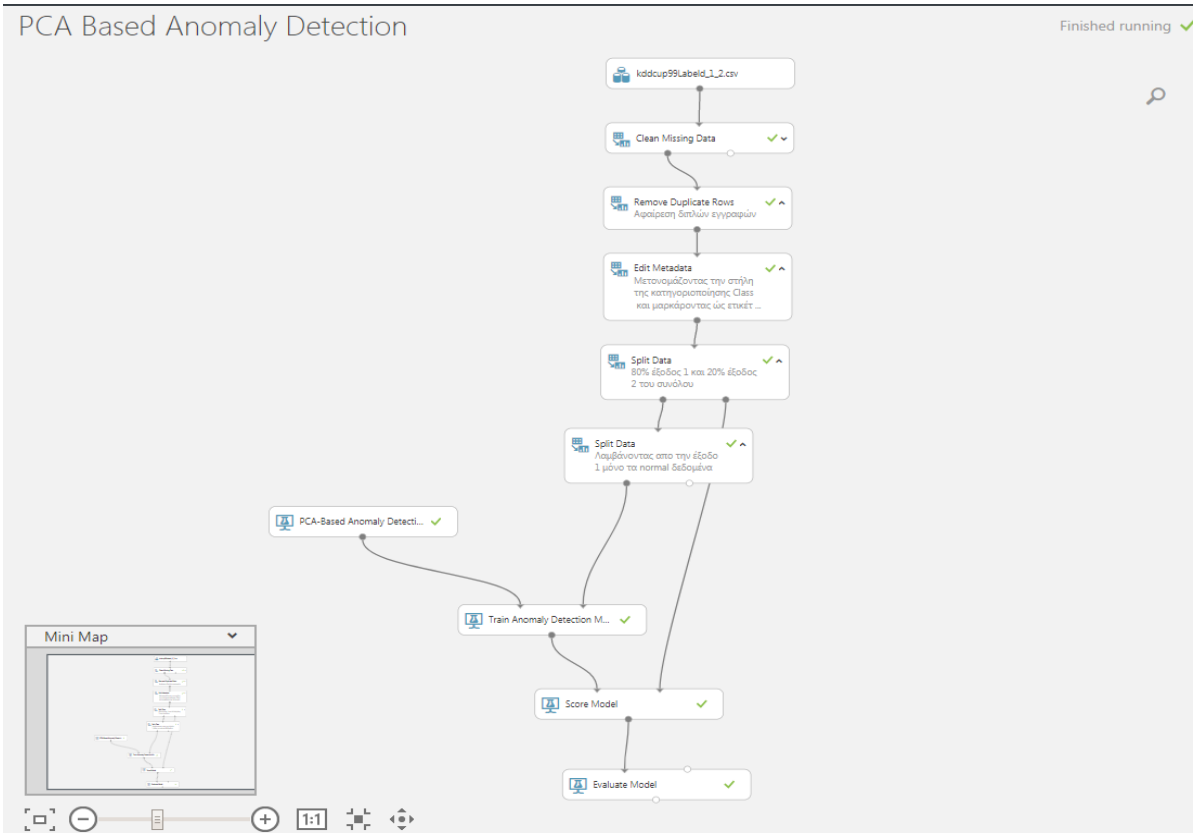
Εικόνα 5.23 Εισαγωγή module αξιολόγησης και οπτικοποίησης αποτελεσμάτων

Μετά από την επιλογή Run και σε αυτό το στάδιο, με δεξί κλικ και οπτικοποίηση μπορούμε να δούμε το παρακάτω γράφημα.



Εικόνα 5.24 Στατιστικά στοιχεία οπτικοποίησης αποτελεσμάτων

Ενώ η τελική μορφή του πειράματος μας είναι η εξής:



Εικόνα 5.25 Τελική δομή πειράματος PCA Based Anomaly Detection

5.4 Ερμηνεία και Αξιολόγηση Αποτελεσμάτων

Όπως είδαμε από τα αποτελέσματα που πήραμε στο γράφημα της εικόνας 5.24, το Evaluation Module κατηγοριοποίησε τα αποτελέσματα ως εξής:

Positive: Ως θετική ετικέτα, έχει ορίσει τα δεδομένα που ανήκουν στη κατηγορία Anomaly (κατηγορία 2), δηλαδή τις εγγραφές εκείνες που έχουμε συμπεριλάβει στις επιθέσεις, μιας και θυμίζουμε πως πλέον δεν ψάχνουμε για συγκεκριμένη κατηγορία επιθέσεων αλλά γενικά για επιθέσεις.

Negative: Ως Αρνητικά θεωρούνται τα δεδομένα που ανήκουν στη πρώτη κατηγορία, αυτά δηλαδή με ετικέτα Normal (κατηγορία 1).

Παρακάτω δίνεται η περιγραφή των αποτελεσμάτων, με τον τρόπο που παρουσιάζονται και με τις εγγραφές να ανήκουν στις κατηγορίες True Positive, True Negative, False Positive και False Negative. Τα δεδομένα αυτά μας δίνονται από το Evaluation Module.

True Positive: Αφορά τις εγγραφές που αποτελούν επιθέσεις και κατηγοριοποιήθηκαν ως τέτοιες. (**8650)**.

True Negative: Αφορά εκείνες τις εγγραφές που ανήκουν στη Normal κατηγορία και κατηγοριοποιήθηκαν από το μοντέλο ως Normal (**17521**).

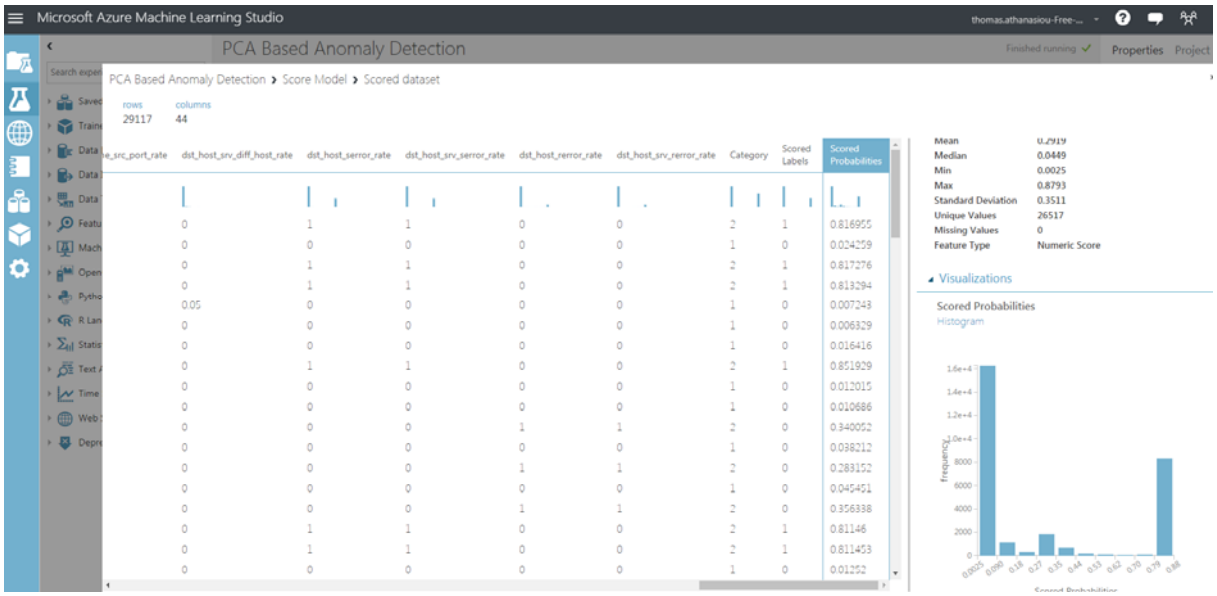
False Positive: Περιλαμβάνει τις εγγραφές εκείνες που κατηγοριοποιήθηκαν εσφαλμένα ως επιθέσεις, ενώ είναι τύπου Normal (**45**).

False Negative: Εδώ ανήκουν οι εγγραφές που κατηγοριοποιήθηκαν ως φυσιολογική δραστηριότητα (Normal), ενώ ανήκαν στην κατηγορία των επιθέσεων (**2901**).

Παρακάτω δίνεται μία μικρή επεξήγηση που αφορά στα στατιστικά μέτρα που μας παρέχονται από το module:

Accuracy: Μετρά την πιστότητα της κατηγοριοποίησης του μοντέλου που εκπαιδεύσαμε με βάση τις σωστές κατηγοριοποιήσεις των εγγραφών επι του συνόλου τους. Στην περίπτωσή μας έχουμε πιστότητα 0.899, δηλαδή 89.9%.

Precision: Αφορά στην ακρίβεια με την οποία βρέθηκαν τα αποτελέσματα, δηλαδή πόσο κοντά στις πραγματικές τιμές είναι η προβλέψεις. Η ακρίβεια δεν έχει να κάνει με το αν τελικά κατηγοριοποιήθηκαν με επιτυχία οι εγγραφές. Εδώ έχουμε ακρίβεια 0.995,δηλαδή 99.5%. Εύλογο είναι να αναρωτηθεί κανείς, γιατί γίνεται λόγος για ακρίβεια σε ένα μοντέλο που κατηγοριοποιεί βάσει σωστού και λάθους, δηλαδή επίθεσης και μη. Η απάντηση μπορεί να γίνει πιο σαφής αν επιλέξουμε visualize στο Score model.



Εικόνα 5.26 Scored Probabilities

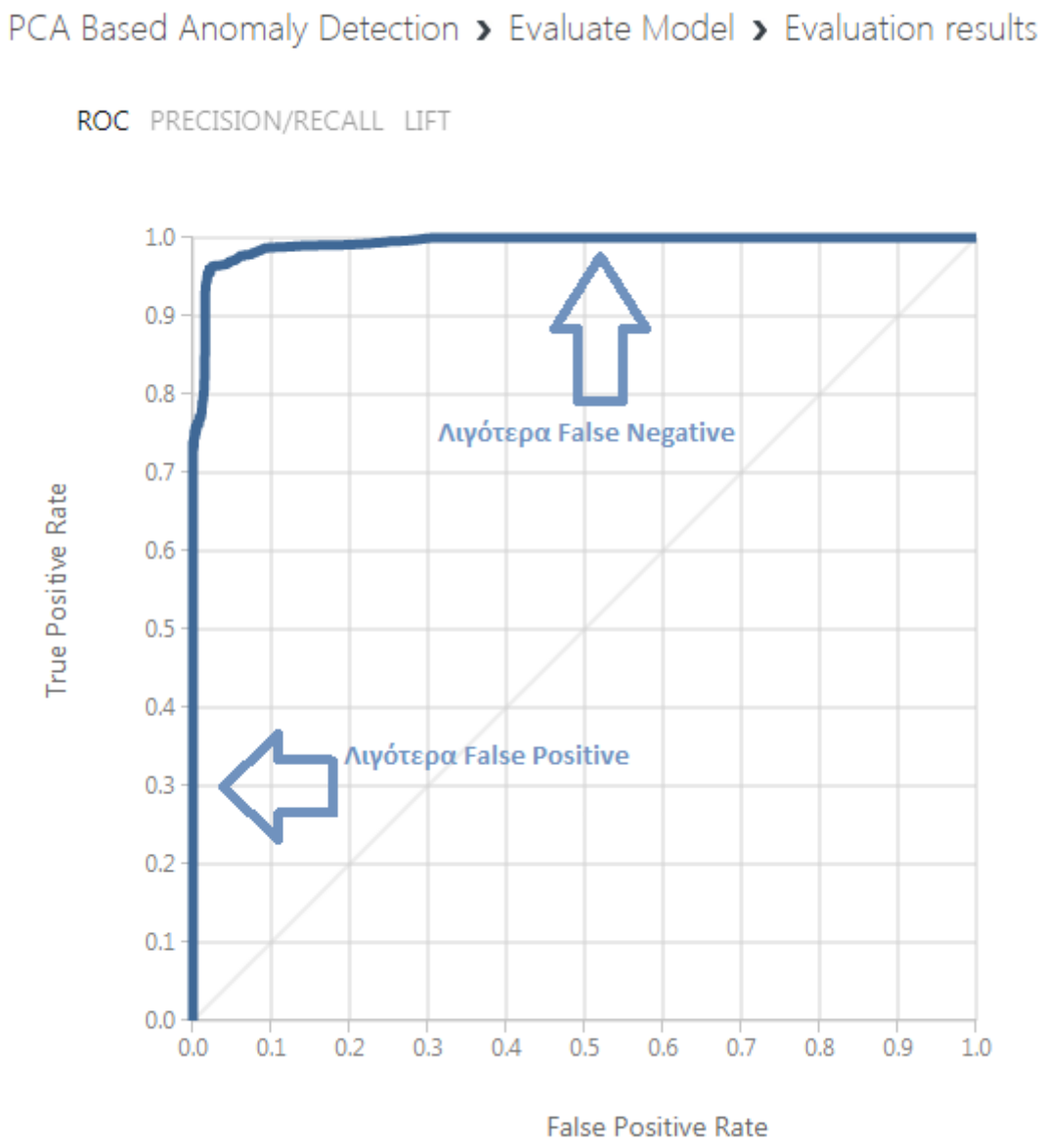
Όπως μπορούμε να δούμε στην στήλη Scored Probabilities, η απόφαση για την κατηγοριοποίηση, γίνεται με βάσει το σκορ της πιθανότητας να ανήκει στην κατηγορία 1 ή 2 και αφορά σε αριθμό με έξι δεκαδικά ψηφία όπως το 0,816955.

F1-score: Υπολογίζεται ως σταθμισμένος παράγοντας του μέσου όρου μεταξύ του precision και του accuracy. Ιδανική τιμή είναι η μονάδα. Η τιμή που έχουμε στο πείραμα μας είναι 0.854.

Recall: Αφορά το ποσοστό όλων των επιτυχών κατηγοριοποιήσεων από το μοντέλο. Η τιμή που έχουμε στο μοντέλο μας είναι 0.749.

ROC Καμπύλη: Η καμπύλη ROC που αφορά στην συμπεριφορά του εκπαιδευμένου μοντέλου ως προς την κατηγοριοποίηση των δεδομένων, έχει χρησιμότητα στο να μας δίνει το σημείο ισορροπίας μεταξύ των σωστών και λάθος προβλέψεων.

Στο πείραμα μας πρέπει με κάποιο τρόπο να μπορούμε να έχουμε το σημείο εκείνο της καμπύλης που είναι η χρυσή τομή μεταξύ των σωστών και των λανθασμένων προβλέψεων.

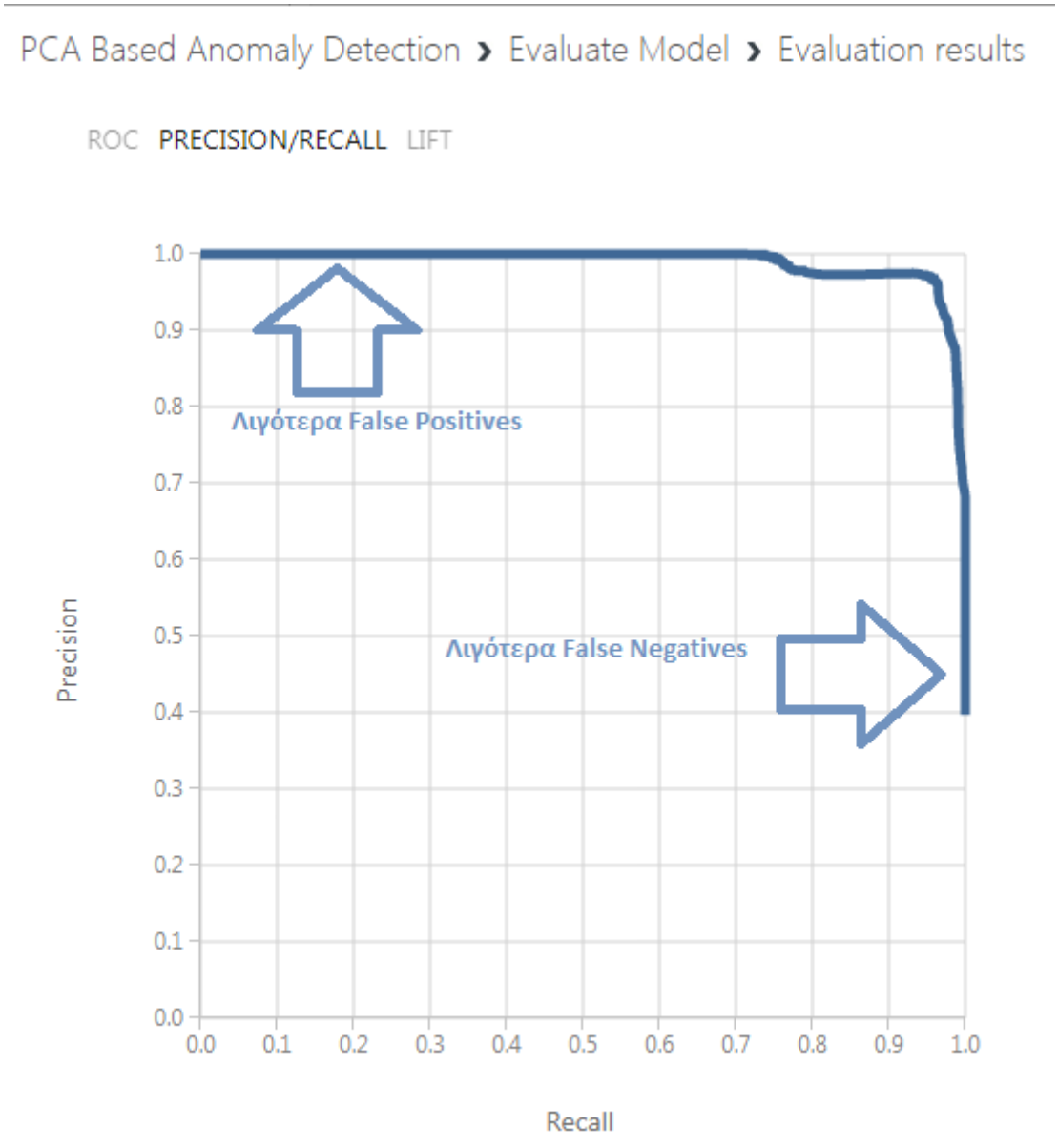


Εικόνα 5.27 Rock Καμπύλη

Όπως φαίνεται από το σχήμα 5.27 στον οριζόντιο άξονα έχουμε τις εγγραφές εκείνες του συνόλου που κατηγοριοποιήθηκαν ως απειλές (anomaly) ενώ ανήκαν στην κατηγορία normal . Όσο πιο κοντά στο 0 τόσο λιγότερες λάθος προβλέψεις και συνεπώς καλύτερα αποτελέσματα (λιγότερα False Positive). Στον κάθετο άξονα έχουμε τις προβλέψεις για τις εγγραφές που ανήκαν στις

επιθέσεις και κατηγοριοποιήθηκαν επιτυχώς όσο ψηλότερα, δηλαδή πιο κοντά στο 1, τόσο και καλύτερα αποτελέσματα.

Ακόμα μια χρήσιμη καμπύλη είναι η καμπύλη **Precision/Recall**.



Εικόνα 5.28 Καμπύλη Precision/Recall

Έχοντας εξηγήσει πιο πάνω τα Recall και Precision, θα πρέπει εδώ να σημειώσουμε πως τα Recall και Precision υπολογίζονται ώς εξής :

- **Recall =** Αληθώς Θετικά / (Αληθώς Θετικά + Ψευδώς Αρνητικά)

- **Precision =** Αληθώς Θετικά / (Αληθώς Θετικά + Ψευδώς Θετικά)

Συνοψίζοντας και έχοντας εξηγήσει τα παραπάνω , μπορούμε να πούμε πως η χρήση του PCA ως αλγόριθμος μείωσης διαστάσεων, αποτελεί μία ενδιαφέρουσα οπτική. Γενικά υπάρχει ενδιαφέρον ώστε να μελετηθεί περαιτέρω το ενδεχόμενο να γίνεται μείωση διαστάσεων σε δεδομένα που έχουν στην διάθεση τους κάποια IDS ή Antivirus συστήματα. Αυτό μειώνει από τη μία τον απαιτούμενο όγκο σε πληροφορία, και επιταχύνει από την άλλη τον χρόνο επεξεργασίας που απαιτείται για τη λήψη αποφάσεων. Το πιο σημαντικό όμως, είναι ότι βρίσκονται νέες συσχετίσεις επί των δεδομένων και αφαιρείται η περιττή πληροφορία.

Σε ό,τι αφορά τα αποτελέσματα του δικού μας πειράματος, μπορούμε κρίνοντας από το Accuracy, να χαρακτηρίσουμε τα αποτελέσματα 'υπερβολικά αισιόδοξα'. Κάποιοι από τους λόγους που συντελούν σε αυτό είναι η αμφισβητούμενη χρήση του Dataset KDD99, που αναλύσαμε και στο Κεφάλαιο 4 (4.4 Μειονεκτήματα). Ακόμη το μέρος του συνόλου που χρησιμοποιήθηκε στο πείραμα μας αποτελεί στην ουσία το 10% του KDD99 . Αυτό έγινε για δύο λόγους. Ο πρώτος είναι ότι το μέγεθος του συνόλου αγγίζει περίπου τα 800 Megabyte πράγμα που αυτομάτως δυσκολεύει την διαδικασία προεπεξεργασίας του(βλ. Κεφάλαιο 1.4). Ο δεύτερος λόγος είναι πως η πλατφόρμα Azure στις εκδόσεις με φοιτητική άδεια, μπορεί να μας προσφέρει την δυνατότητα χρήσης βάσεων δεδομένων με περιορισμό τα 30 Megabyte. Το 10% του συνόλου που χρησιμοποιήθηκε έχει μέγεθος κοντά στα 80 Megabyte.

6.5 Σύνοψη Κεφαλαίου

Στο κεφάλαιο αυτό ασχοληθήκαμε με το πειραματικό μέρος της διατριβής, όπου το πείραμα χωρίστηκε σε δύο μέρη. Το πρώτο μέρος αφορούσε τη μείωση διαστάσεων πάνω στα δεδομένα του συνόλου KDD99. Η μείωση των διαστάσεων έγινε με την χρήση του module Principal component analysis. Η δομή του πρώτου μέρους του πειράματος, μπορεί να συνδυαστεί στην συνέχεια με αλγόριθμους που χρησιμοποιούν διαφορετικές τεχνικές κατηγοριοποίησης. Για να μπορέσουμε να χρησιμοποιήσουμε την ίδια τεχνική και στο δεύτερο μέρος, δηλαδή στην ανίχνευση ανωμαλιών, χρησιμοποιήσαμε το module PCA Based Anomaly Detection. Το module αυτό χρησιμοποποιείται ως αλγόριθμος ανίχνευσης ανωμαλιών μέσω της μείωσης διαστάσεων και εύρεσης των principal components. Έτσι στο δεύτερο μέρος του πειράματος εκπαιδεύσαμε ένα μοντέλο να μπορεί να αναγνωρίζει πιθανές ανωμαλίες. Τέλος, με την χρήση των module Score και Evaluate model, εξάγαμε στατιστικές πληροφορίες για τη συμπεριφορά του μοντέλου.

Βιβλιογραφία

[1] Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). NRC Publications Archive (NPArC) Archives des publications du CNRC (NPArC) A Detailed Analysis of the KDD CUP 99 Data Set.

[2] Lukacs, M., & Bhadra, D. (2003). Table of of contents, (November), 2004.

[3] Smith, L. I. (2002). A tutorial on Principal Components Analysis Introduction. *Toturial*, 1–27.

[4] Janecek, A., Gansterer, W. N., Demel, M., & Ecker, G. (2008). On the Relationship Between Feature Selection and Classification Accuracy. *Fsdm*, 90–105.

[5] Schlesinger, M., & Hlavác, V. (2011). Supervised And Unsupervised learning. *Artificial Intelligence*, (April).

[6] Τεχνικές Μείωσης Διαστάσεων Ειδικά θέματα ψηφιακής εικόνας Εισαγωγή. (2008).

[7] Yeh, Y., Lee, Z., & Lee, Y. (n.d.). Anomaly Detection via Over-Sampling Principal Component Analysis, (43), 449–458. No Title. (2010).

[8] Analysis-ca, C., Averaging-ra, R., Roux, L., Analysis, D. C., & Braak, T. (2007).

[9] Chauhan, A., Mishra, G., & Kumar, G. (2011). Detection, *2*(7), 2–5.

[10] Knowledge and Data Extraction. Advantages and Disadvantages in. (2014).

[11] Joshi, M. (2012). Classification, Clustering And Intrusion Detection System. *International Journal of Engineering Research and Applications*, *2*(2), 961–964.

[12] (EC-Council), I. C. of E.-C. C. (2013). Footprinting and Scanning. *Certified Ethical Hacker V8.00*.

[13] Mr. Kamlesh Lahre Suresh Kumar Kashyap, Pooja Agrawal, M. T. dhar D. (2013). Analyze Different approaches for IDS using KDD 99 Data Set. *International Journal on Recent and Innovation Trends in Computing and Communication*, *1*(8), 7.

[14] Wang, Y., & Yu, S. Z. (2009). Supervised learning real-time traffic classifiers. *Journal of Networks*, *4*(7), 622–629.

[15] Fa, a F., & a, F. D. B. D. E. D. F. B. (2001). Interested in learning SANS Institute InfoSec Reading Room tu , A ho ll r igh ts. *Information Security*, 18.(2001).

[16] Amanpreet, C., Gaurav, M., & Kumar, G. (2011). Survey on Data Mining Techniques in Intrusion Detection. *International Journal of Scientific & Engineering Research*, *2*(7), 2–5.

[17] IOS, An Introduction to Classification: Feature Selection.

[18] Kumar, Steinbach, Tan (2004): Introduction to Data Mining, University of Stanford.

[19] Dunham M.H., (2004): Data Mining introductory and advanced topics", Prentice Hall.

[20] Trevor Hastie, Robert Tibshirani, Jerome Friedman, "The elements of statistical.

[21] Learning, Data Mining, Inference and Prediction", 2nd edition, Springer.