ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
Τμήμα Ηλεκτρολόγων & Ηλεκτρονικών Μηχανικών

UNIVERSITY of WEST ATTICA
FACULTY OF ENGINEERING
Department of Electrical & Electronics Engineering

Πρόγραμμα Μεταπτυχιακών Σπουδών

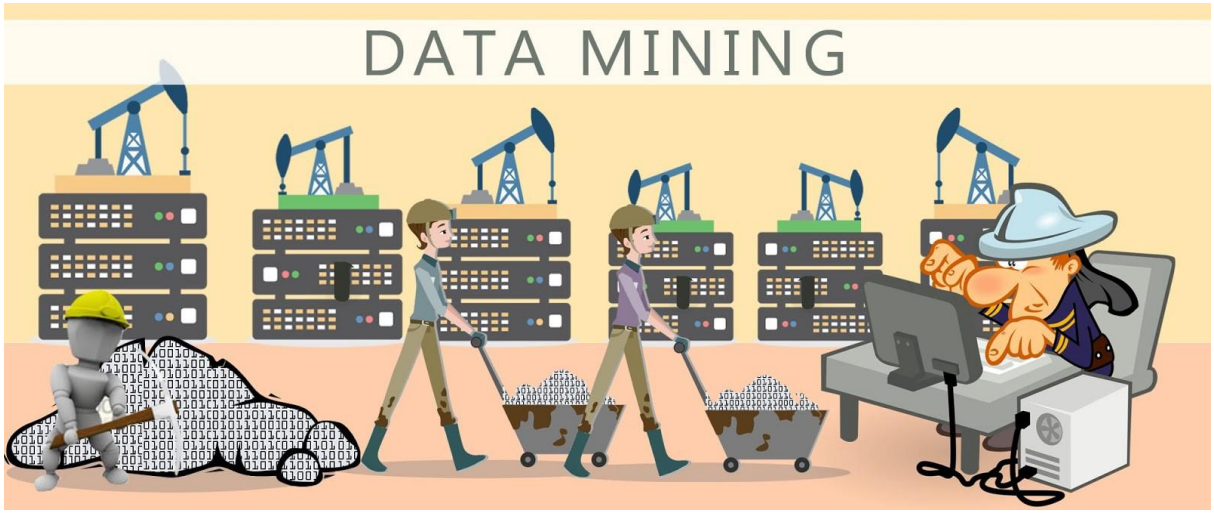
Δικτυωμένα Ηλεκτρονικά Συστήματα

Master of Science in

Internetworked Electronic Systems

# ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

## Ανάπτυξη αλγορίθμου εξόρυξης δεδομένων για ανίχνευση απάτης πιστωτικών καρτών και επίδειξη σε σύστημα μηχανικής μάθησης τεχνητής ευφυΐας.




Μεταπτυχιακός Φοιτητής: Θεόδωρος Ιωάννης Ζαφειρόπουλος , Αρ. Μητρ : 35
Επιβλέπων :Μετάφας Δημήτριος, Επίκουρος Καθηγητής

ΑΙΓΑΛΕΩ, ΟΚΤΩΒΡΙΟΣ 2019

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
Τμήμα Ηλεκτρολόγων & Ηλεκτρονικών Μηχανικών

UNIVERSITY of WEST ATTICA
FACULTY OF ENGINEERING
Department of Electrical & Electronics Engineering

Πρόγραμμα Μεταπτυχιακών Σπουδών
Δικτυωμένα Ηλεκτρονικά Συστήματα

Master of Science in
Internetworked Electronic Systems

# MSc THESIS

# Development of a data mining algorithm to detect fraud and demonstration of the algorithm in an artificial intelligence machine learning system - environment.


Student: Zafeiropoulos Theodoros Ioannis, Reg. Nr.: IES-0035
MSc Thesis Supervisor: Metafas Dimitrios, Assistant Professor

ATHENS-EGALEO, OCTOBER 2019

(ΥΠΟΓΡΑΦΗ)

**.....**

**ΖΑΦΕΙΡΟΠΟΥΛΟΣ ΘΕΟΔΩΡΟΣ ΙΩΑΝΝΗΣ**

**ΠΤΥΧΙΟΥΧΟΣ ΗΛΕΚΤΡΟΝΙΚΟΣ ΜΗΧΑΝΙΚΟΣ**

# Ευχαριστίες

Αρχικά, θα ήθελα εκφράσω την ιδιαίτερη εκτίμηση και τις θερμές ευχαριστίες μου στον καθηγητή μου κ. Μετάφα Δημήτρη, τόσο για την εποικοδομητική συνεργασία και την ουσιαστική του συνεισφορά στην ολοκλήρωση της παρούσας εργασίας, όσο και για την πολύτιμη καθοδήγησή του.

Πρόσθετα θα ήθελα να ευχαριστήσω το σύνολο των καθηγητών με την προσπάθεια και τη συμβολή των οποίων έχω το προνόμιο να βρίσκομαι σήμερα στην παρούσα θέση.

Φυσικά, ένα μεγάλο ευχαριστώ αναλογεί στην οικογένειά μου για την ηθική και οικονομική υποστήριξή τους όλα αυτά τα χρόνια, καθώς επίσης και στους καλούς μου φίλους που είναι πάντα δίπλα μου στις δύσκολες στιγμές.

# Περίληψη

Η παρούσα εργασία αποτελεί διπλωματική εργασία στα πλαίσια του μεταπτυχιακού προγράμματος «Δικτυωμένα Ηλεκτρονικά Συστήματα» του τμήματος Ηλεκτρολόγων και Ηλεκτρονικών Μηχανικών του Πανεπιστημίου Δυτικής Αττικής. Στόχος είναι να τεκμηριώσουμε ότι είναι αναγκαίο να διαθέτουμε ένα εργαλείο για την ανάλυση και ερμηνεία μιας σημαντικής ποσότητας πληροφορίας που είναι οργανωμένη σε βάσεις δεδομένων και επιπλέον να αποτελέσει κίνητρο για μελλοντικές ερευνητικές εργασίες στα ζητήματα αυτά. Δομείται σε τρία κεφάλαια εκ των οποίων στα δυο πρώτα εισάγονται οι προαπαιτούμενες γνώσεις που είναι αναγκαίες για την ανάλυση της μεθόδου που εφαρμόσαμε.

Πιο συγκεκριμένα στο 1$^{ο}$ κεφάλαιο δίνεται ο ορισμός της εξόρυξης δεδομένων, οι τύποι μοντέλων που παράγονται από αυτήν, οι τομείς με τους οποίους συνδυάζεται, οι τομείς εφαρμογής της, όπως και οι τεχνικές επεξεργασίας των δεδομένων.

Στο 2$^{ο}$ κεφάλαιο στο 1$^{ο}$ μέρος δίνεται η γενική ιδέα που κρύβεται πίσω από το Machine Learning και αναπτύσσονται οι θεωρίες μάθησης, ενώ στο 2$^{ο}$ μέρος γίνεται ανάλυση του προγράμματος Machine Learning που χρησιμοποιήθηκε για την ανάλυση των δεδομένων μας, όπως και περιγραφή της μεθόδου PCA που χρησιμοποιήσαμε ώστε να μετατρέψουμε τα δεδομένα σε αριθμούς.

Στο 3$^{ο}$ κεφάλαιο που αποτελεί και το ερευνητικό μέρος της παρούσας διπλωματικής, αφού αναλυθούν οι τεχνικές ανίχνευσης παρεκτροπών αναπτύσσονται οι βασικές ιδέες των δύο αλγορίθμων που επιλέξαμε για την έρευνά μας, του Local Outlier Factor και του Isolation Forest. Η ανάλυσή μας αναφέρεται σε δείγμα 284.807 συναλλαγών με πιστωτικές κάρτες που πραγματοποιήθηκαν τον Σεπτέμβριο του 2013 από Ευρωπαίους κατόχους καρτών. Αρχικά έγινε μία στατιστική ανάλυση σε ποσοστό 100% και στη συνέχεια έγινε εισαγωγή της μεταβλητής *class* όπου για περίπτωση απάτης παίρνει την τιμή 1 και για μη δόλια συναλλαγή παίρνει την τιμή 0. Έπειτα χρησιμοποιώντας μόνο το 10% των δεδομένων με τυχαία κατάσταση έγιναν δύο διαφορετικές αναλύσεις, μία με τον αλγόριθμο Local Outlier Factor και μία με τον Isolation Forest.

Τέλος συγκρίνονται τα αποτελέσματα των δύο μεθόδων και παρουσιάζονται τα συμπεράσματά μας ως προς το ποια μέθοδος πλεονεκτεί και γιατί, όπως και προτάσεις για βελτίωση της ακρίβειας της πρόβλεψης της μεθόδου.

## Σκοποί και στόχοι

### Σκοποί

- να μελετηθεί ένας πρόσφατος κλάδος της επιστήμης που βρίσκεται σε εξέλιξη,
- να παρέχουμε μια καλύτερη εικόνα του χώρου και να επιδείξουμε τα διάφορα εργαλεία που είναι διαθέσιμα καθώς επίσης και τη λογική χρήσης τους,
- να μελετήσουμε στην πράξη τις ειδικές συνθήκες που απαιτεί η Ανίχνευση Απάτης.

### Στόχοι :

- η ανακάλυψη γνώσης μέσα από μεγάλους όγκους δεδομένων,
- να αναπτύξουμε μηχανισμούς που εμποδίζουν ή ανακαλύπτουν ορισμένες δραστηριότητες του ανθρώπου που σαν αποκλειστικό σκοπό έχουν την παραπλάνηση ή την εξαπάτηση,
- να διευκολύνουμε με μια διαδικασία που διαθέτει αυτοματοποιημένο χαρακτήρα τους ειδικούς ερευνητές που δεν είναι έμπειροι αναλυτές, καλύπτοντας έτσι το κενό της γνώσης τους ολοκληρώνοντας και τα στατιστικά τους αποτελέσματα.

## Λέξεις - Κλειδιά

**Αλγόριθμος δασικής απομόνωσης, Ανίχνευση παρεκτροπών, Τοπικός συντελεστής απόκλισης, Μηχανική μάθηση, Εξόρυξη δεδομένων.**

# Abstract

This present thesis is a thesis within the framework of the postgraduate course «Internetworked Electronic Systems» of the department of electronic engineering University of West Attica. Our objective is to manifest the necessity for the existence of a tool for the analysis and the interpretation of a significant amount of information organized into databases, as well as to serve as a motive paving the way towards conducting further research. The thesis is divided into three units.

In the first two units there is a presentation of the prerequisite knowledge, which is indispensable for the analysis of the applied method. More specifically, the definition of the term ‘’data mining’’, the model types deriving from data mining, the sectors to which it can be applied, as well as the data processing techniques are presented in the first unit. In the first part of the second unit the underlying idea behind Machine Learning, as well as the existent learning theories are presented. The second part of unit -2 the program ‘’Machine Learning’’, which was used for our data analysis, is presented. Moreover, in this second part of unit 2 the PCA method, which was used for the conversion of our data into numbers, is described. Unit 3 constitutes the research of the thesis.

After the deviation detection techniques are analyzed in Unit 3, the two algorithms that were selected for our research are presented. These two algorithms are the Local Outlier Factor and the Isolation Forest. Our analysis is based on a sample of 284.807 credit card transactions that took place during September 2013 by European credit card holders. Initially a statistical analysis upon 100% of the sample took place. In continuation, the ‘’class’’ variable was introduced, which takes the value of 1 in the case of fraud and the value of 0 in the case of non-fraudulent transactions. Afterwards two different analyses were conducted on 10% of the data, which was selected randomly. The first analysis made use of the Local Outlier Factor and the second made use of the Isolation Factor. Finally, the results of the two methods are compared and our drawn conclusions are presented regarding which of the aforementioned methods is more advantageous and why.

Last but not least, suggestions are made concerning the improvement of the prediction accuracy of the method.

## Objectives and aims

### Objectives:

- The study of a scientific domain which has recently started existing and is continually evolving,
- Profiling the aforementioned domain more accurately and presenting the various available tools, as well as the rationale for their use,
- Studying in practice the special parameters.

**Aims:**

- The acquisition of knowledge through large data volume,
- Mechanisms that hinder or discover specific human activities, the only objective of which is fraud or deception,
- The development of mechanisms that hinder or discover specific human activities, the only objective of which is fraud or deception providing special researchers (who, however, are not experienced analysts) with a process which, thanks to its being automated, facilitates them. In this way, automated processed can compensate for researchers' lack of knowledge and consolidate their statistical conclusions.

## Key- Words

**Anomaly detection, Forest isolation algorithm, Local outlier factor, Machine learning, Data mining.**

# Περιεχόμενα

# Συντμήσεις - Λεξιλόγιο

Στο παράρτημα αυτό δίνεται ένας πίνακας με την αντιστοιχία όρων σε Αγγλική και Ελληνική γλώσσα με πρωτογενή και δευτερογενή ερμηνεία.

| Αριθμός | Αγγλική γλώσσα | Ελληνική γλώσσα |
|---|---|---|
| 1. | Accuracy Score | Τοπική βαθμολογία ανωμαλίας |
| 2. | Adaptive websites | Προσαρμοζόμενοι ιστότοποι |
| 3. | Anomaly Detection | Ανίχνευση παρεκτροπών |
| 4. | Artificial intelligence | Τεχνητή νοημοσύνη |
| 5. | Cheminformatics | Χημειοπληροφορική |
| 6. | Classifier | Κατηγοριοποίηση |
| 7. | Classifying DNA sequences | Ταξινόμηση ακολουθιών DNA |
| 8. | Computational finance | Υπολογιστική οικονομία |
| 9. | Computer vision | Ενόραση υπολογιστών |
| 10. | Contamination | Μόλυνση |
| 11. | Correlation rules | Κανόνες συσχέτισης |
| 12. | Clustering | Συσταδοποίηση |
| 13. | Clusters analysis | Ανάλυση συστάδων |
| 14. | Data bases | Βάσεις δεδομένων |
| 15. | Data Frames | Δυσδιάστατη |
| 16. | Data mining | Εξόρυξη δεδομένων |
| 17. | Data ownership and distribution | Κυριότητα και διανομή δεδομένων |
| 18. | Detecting credit card fraud | Ανίχνευση απάτης πιστωτικών καρτών |
| 19. | Decision support | Υποστήριξη απόφασης |
| 20. | Descriptive models | Περιγραφικά μοντέλα |
| 21. | Game playing | Παίξιμο Παιχνιδιών |
| 22. | Heterogeneous and complex data | Ετερογενή και πολύπλοκα δεδομένα |
| 23. | High dimensionality | Πολλές διαστάσεις |

| | | |
|---|---|---|
| 24. | Information retrieval | Ανάκτηση πληροφοριών |
| 25. | Input neurons | Νευρώνες εισόδου |
| 26. | Isolation Forest | Απομόνωση Δασών |
| 27. | Knowledge discovery in databases | Ανακαλυφθείσα γνώση από βάσεις δεδομένων |
| 28. | Local Outlier Factor | Τοπικός συντελεστής απόκλισης |
| 29. | Machine learning | Μηχανική μάθηση |
| 30. | Machine perception | Νόηση μηχανών |
| 31. | Neighbors | Γείτονες |
| 32. | Natural language processing | Επεξεργασία φυσικών γλωσσών |
| 33. | Neural networks | Νευρωνικά δίκτυα |
| 34. | Not-traditional analysis | Μη παραδοσιακή ανάλυση |
| 35. | Outliers | Υπερβάσεις |
| 36. | Outlier detection | Ανίχνευση εξωστρέφειας |
| 37. | Output neurons | Νευρώνες εξόδου |
| 38. | Outlier fraction | Κλάσμα |
| 39. | Pattern matching | Ταιριάσματος μοτίβων |
| 40. | Precision | Ακρίβεια |
| 41. | Pattern matching | Ταίριασμα μοτίβων |
| 42. | Predictive models | Μοντέλα πρόβλεψης |
| 43. | Recall | Ανάκληση |
| 44. | Recommender systems | Συνιστώμενα συστήματα |
| 45. | Regression | Παλινδρόμηση |
| 46. | Regression analysis | Ανάλυση παλινδρόμησης |
| 47. | Robot locomotion | Μετακινήσεις ρομπότ |
| 48. | Search engines | Μηχανές αναζήτησης |
| 49. | Scalability | Κλιμάκωση |
| 50. | Series | Μονοδιάστατη |
| 51. | Sequence patterns | Πρότυπα Ακολουθιών |
| 52. | Sequence mining | Εξόρυξη ακολουθιών |
| 53. | Statistic | Στατιστική |

| | | |
|---|---|---|
| 54. | Stock market analysis | Ανάλυση αγοράς μετοχών |
| 55. | Structural health monitoring | Δομημένος έλεγχος υγείας |
| 56. | Syntactic pattern recognition | Αναγνώριση σχεδίων |

# Ευρετήριο εικόνων

# Ευρετήριο σχημάτων

# Ευρετήριο διαγραμμάτων

# Εισαγωγή

Ο 21ος αιώνας έχει χαρακτηριστεί από πολλούς, ως ο αιώνας της πληροφορίας. Η πληροφορία αποτελεί τον πλέον πολύτιμο πόρο των σύγχρονων επιχειρήσεων. Η ικανότητα συλλογής πληροφοριών και δεδομένων συγκαταλέγεται ως πλεονέκτημα σε πολλούς τομείς. Στην σημερινή εποχή είναι απαραίτητο να διαθέτουμε εργαλείο για την ανάλυση και ερμηνεία μιας σημαντικής ποσότητας πληροφορίας που είναι οργανωμένη σε βάσεις δεδομένων. Ένα τέτοιο εργαλείο είναι η **τεχνική εξόρυξης δεδομένων (Data Mining)** η οποία δίνει την δυνατότητα κανόνων μέσω των υπολογιστών.

Ο συγκεκριμένος τομέας αποτελεί αντικείμενο μελέτης για ερευνητές και μηχανικούς. Συγκεκριμένα τα τελευταία χρόνια έχουμε μεγάλη αύξηση του όγκου πληροφορίας. Οι ερευνητές προχωρούν καθημερινά την έρευνα και έχουν γίνει τεράστιες προσπάθειες βελτίωσης.

Το μεγαλύτερο όμως πρόβλημα είναι το κενό που δημιουργείται μεταξύ της απόδοσης του υλικού και της ποσότητας των δεδομένων, το οποίο θα χρειαστεί να αναλύσουμε. Οι αλγόριθμοι που διαχειρίζονται πολύ λιγότερα δεδομένα, αντιμετωπίζουν προβλήματα απόδοσης από την στιγμή όπου το υλικό δε θα καταφέρει να καλύψει το κενό από τον όγκο δεδομένων. Ένας αλγόριθμος ταξινόμησης που λειτουργεί ορθά με λίγα megabyte δεδομένων ίσως παρουσίαζε προβλήματα απόδοσης αν εφαρμοστεί σε gigabytes δεδομένων.

# Κεφάλαιο 1ο : Εξόρυξη Δεδομένων

Στο παρόν κεφάλαιο θα εισαγάγουμε τον ορισμό της εξόρυξης δεδομένων, θα ορίσουμε τις βασικές αρχές και εισαγωγικές έννοιες και θα παρουσιάσουμε τους λόγους, για τους οποίους αναπτύχθηκε, εξελίχθηκε και διαδόθηκε το συγκεκριμένο επιστημονικό πεδίο. **Στόχος** είναι να αποκτήσουμε:

- ✓ μια σφαιρική εικόνα γύρω από την θεωρία της εξόρυξης δεδομένων (Data Mining) και
- ✓ όλα τα αναγκαία εφόδια που θα χρειαστούμε για την κατασκευή και προσομοίωση ενός προγράμματος εξόρυξης δεδομένων.

## 1.1 Ορισμός εξόρυξης Δεδομένων (Data Mining)

Η σύγκλιση της προόδου υπολογιστικών συστημάτων και της εξέλιξης στην επικοινωνία έχει οδηγήσει στην δημιουργία μια κοινωνίας ικανής να παρέχει διαρκώς νέες πληροφορίες. Το υλικό συγκεντρώνεται και καταγράφεται διαρκώς, με αποτέλεσμα να έχουμε τεράστιο όγκο βάσεων δεδομένων. Το γεγονός αυτό είναι ένα σύγχρονο φαινόμενο το οποίο παρατηρείται ως ανάγκη από τα απλούστερα ζητήματα της καθημερινής ζωής έως και τα πιο συνέθετα.

Ας σκεφτούμε λοιπόν για παράδειγμα ότι σε ένα σύστημα βιάσεων δεδομένων πρέπει να καταγράφονται οι συναλλαγές που γίνονται σε ένα κατάστημα ή η χρήση πιστωτικών καρτών και δανείων από τους πελάτες μιας τράπεζας ( σύστημα δοσοληψιών ). Από την άλλη πλευρά υπάρχουν και πολυπλοκότερα ζητήματα τα οποία χρειάζονται να οργανωθούν μέσω μια βάσης δεδομένων όπως θέματα ιατρικής, φωτογραφίες από δορυφόρους δηλαδή διαδικασίες συσσώρευσής ψηφιακών αρχείων.

Το ζήτημα επομένως που τίθεται είναι αν υπάρχει τρόπος να διαχειριστούμε τις πολύ μεγάλες αυτές βάσεις δεδομένων που ανανεώνονται διαρκώς από τους χρήστες. Επίσης θεωρείται αρκετά δύσκολη η άντληση του απαραίτητου υλικού από αυτές. Όλα αυτά τα θέματα προκάλεσαν το ενδιαφέρον και οδήγησαν στην διαδικασία της **εξόρυξης δεδομένων ( Data mining)** .

Με τον ορισμό **εξόρυξη δεδομένων (Data Mining)** εννοούμε μία διαδικασία που αυτόματα ανακαλύπτεις χρήσιμες πληροφορίες μέσα από μεγάλες δεξαμενές δεδομένων. Είναι μια τεχνολογία που συνδυάζονται οι παραδοσιακές μέθοδοι ανάλυσης δεδομένων με

τους σύγχρονους αλγόριθμους για επεξεργασία μεγάλου όγκου δεδομένων. Οι τεχνικές εξόρυξης δεδομένων εφαρμόζονται για να ερευνήσουν σε βάθος μεγάλες βάσεις δεδομένων και να βρεθούν νέα και χρήσιμα πρότυπα που σε δε διαφορετική περίπτωση θα παρέμεναν άγνωστα. Επίσης, παρέχουν δυνατότητες πρόβλεψης του αποτελέσματος μιας μελλοντικής παρατήρησης. Παρόλα αυτά οι τεχνικές εξόρυξής έχουν χρησιμοποιηθεί για να επεκτείνουν τα συστήματα ανάκτησης πληροφοριών. Η εξόρυξη δεδομένων είναι αναπόσπαστο κομμάτι της ανακάλυψης γνώσης από βάσεις δεδομένων (Knowledge discovery in databases –KDD).

Η σειρά αυτή αποτελείται από μία σειρά βημάτων μετασχηματισμών που αρχίζει από την προεπεξεργασία δεδομένων και συνεχίζει μέχρι την εκ των υστέρων επεξεργασία των αποτελεσμάτων της εξόρυξης δεδομένων. Τα δεδομένα της εισόδου μπορούν συχνά να αποθηκευτούν σε μια ποικιλία μορφών όπως είναι τα λογιστικά φύλλα και οι σχεσιακοί πίνακες.


Εικόνα 1: Data mining

Ο σκοπός της προ επεξεργασίας (pre-processing ) είναι για να μετατρέψει τα ακατέργαστα δεδομένα εισόδου σε μια μορφή κατάλληλη για την ανάλυση που θα ακολουθήσει. Επίσης, επειδή διαθέτουμε πολλά δεδομένα η προ επεξεργασία δεδομένων πιθανόν να είναι το ποιο χρονοβόρο βήμα σε όλη τη διαδικασία "ανακάλυψης γνώσης".

## 1.2 Τύποι Μοντέλων

Τα μοντέλα που παράγονται από το στάδιο της εξόρυξης δεδομένων διακρίνονται σε δυο βασικούς τύπους:

- τα μοντέλα πρόβλεψης (predictive models)
- και τα περιγραφικά μοντέλα (descriptive models).

**Στόχος ενός μοντέλου πρόβλεψης** είναι να προβλέψει τιμές για ένα συγκεκριμένο χαρακτηριστικό που παρουσιάζει ενδιαφέρον και που πιθανώς βασίζεται στη συμπεριφορά άλλων χαρακτηριστικών. Για παράδειγμα, η πρόβλεψη μπορεί να βασίζεται στη χρονολογική κατάταξη των δεδομένων.

**Ένα περιγραφικό μοντέλο** βρίσκει μοτίβα (patterns) ή σχέσεις (relations) που συνυπάρχουν στα δεδομένα και μελετά τις ιδιότητές τους, ώστε να δοθεί μια αιτιολόγηση της συμπεριφοράς τους.

## 1.3 Εξόρυξη δεδομένων και άλλοι επιστημονικοί τομείς


Σχήμα 1 :Η εξόρυξη δεδομένων στη συμβολή άλλων επιστημονικών πεδίων

Ο τομέας της εξόρυξης δεδομένων συνδυάζεται με πολλούς άλλους επιστημονικούς τομείς όπως:

- την στατιστική (statistic)
- την τεχνητή νοημοσύνη (artificial intelligence)
- την μηχανική μάθηση (machine learning)
- τις βάσεις δεδομένων (data bases)

- τις μηχανές αναζήτησης
- τα συστήματα υποστήριξης απόφασης (decision support systems)

- τα συστήματα άμεσης ανάλυσης δεδομένων (OLAP) και του ταιριάσματος μοτίβων ( pattern matching).

Παρακάτω θα αναλύσουμε τη σχέση που έχει η εξόρυξη δεδομένων με μερικούς από τους πιο βασικούς τομείς που μόλις αναφέραμε ανωτέρω.

### 1.3.1 Στατιστική

Είναι γνωστό πως ένα μεγάλο μέρος της ερευνητικής βάσης της εξόρυξης δεδομένων βασίζεται στην στατιστική. Αυτό είναι λογικό εφόσον και η στατιστική έχει ανάλογους σκοπούς με την εξόρυξη δεδομένων, αφού και οι δύο αποσκοπούν στην αναγνώριση χρήσιμων πληροφοριών και μοτίβων στα δεδομένα. Μέρος των διαδικασιών σε ένα μοντέλο εξόρυξης δεδομένων μπορεί να αποτελέσει η αναζήτηση των δεδομένων και η εξαγωγή συμπερασμάτων από τα αποτελέσματα μιας αναζήτησης. Μια συχνά χρησιμοποιούμενη τεχνική στην εξόρυξη δεδομένων είναι η τεχνική της δειγματοληψίας. Ο συγκεκριμένος τρόπος στη στατιστική λέγεται <<στατιστική εξαγωγή συμπεράσματος>> .Ακόμα και σήμερα ένα σημαντικό τμήμα των νεών υλοποιημένων αλγορίθμων εξόρυξης δεδομένων, αποτελούν στην ουσία στατιστικές τεχνικές που έχουν προσαρμοστεί στις απαιτήσεις των αλγορίθμων και των υπολογισμών. Όπως και με τις κλασσικές τεχνικές, στην εξόρυξη δεδομένων ακολουθούμε την ανάλυση παλινδρόμησης ( regression analysis ), ανάλυσης συστάδων (clusters analysis) κ.α. Ακόμα και όταν οι αλγόριθμοι εξόρυξης δεδομένων δεν χρησιμοποιούν άμεσα τις τεχνικές στατιστικής, πολλές φορές οι βασικές τους ιδέες έχουν ως αρχική επιρροή την στατιστική.

### 1.3.2 Τεχνητή Νοημοσύνη

Άλλος ένας τομέας που σχετίζεται με αυτόν της εξόρυξης δεδομένων είναι η τεχνητή νοημοσύνη. Μια βασική κατεύθυνση της επιστήμης της τεχνητής νοημοσύνης είναι η δημιουργία μηχανικών δομών που παραστάνουν τον τρόπο με τον οποίο ο ανθρώπινος εγκέφαλος αποθηκεύει δεδομένα, πληροφορία και γνώση και η αντιστοιχία των δομών αυτών με τον ανθρώπινο συλλογισμό. Ο βασικός σκοπός είναι να βγάζει λογικά συμπεράσματα από ανεπεξέργαστα δεδομένα, όπως συμβαίνει και στον τομέα της εξόρυξης δεδομένων. Επίσης ο τομέας της εξόρυξης δεδομένων κάνει εκτεταμένη χρήση εργαλείων τεχνητής νοημοσύνης και μηχανικής μάθησης. Ο τομέας της τεχνητής νοημοσύνης θεωρείται πιο γενικός και εμπεριέχει περιοχές εκτός των κλασικών μεθόδων εξόρυξης δεδομένων.

### 1.3.3 Βάσεις δεδομένων

Μια βάση δεδομένων είναι μια συλλογή από δεδομένα. Αντίθετα με ένα απλό σύνολο, τα δεδομένα σε μια βάση(σχεσιακή βάση) έχουν μια ορισμένη δομή ή σχήμα με το οποίο είναι σχετιζόμενα. Έχουμε εδώ σχεσιακή βάση δεδομένων οργανωμένη με συσχετισμένους πίνακες, στους οποίους έχουμε ένα μηχανισμό για ανάγνωση, εγγραφή, τροποποίηση. Αποθηκεύουμε την πληροφορία σε οργανωμένη μορφή και την εξάγουμε πάλι σε οργανωμένη μορφή. Ένα μεγάλο μέρος των σημερινών ερευνητών στην εξόρυξη δεδομένων είναι άτομα προερχόμενα από τον τομέα των βάσεων δεδομένων. Η σχέση των δύο αυτών τομέων είναι εμφανής μιας και πριν επεξεργαστούμε τα δεδομένα μας, πρέπει πρώτα να μπορούμε να τα διαχειριστούμε ορθά. Έτσι χωρίς καλά συστήματα διαχείρισης δεδομένων δεν μπορούμε να εφαρμόσουμε αλγόριθμους εξόρυξης δεδομένων. Οι δύο τομείς ακόμη μοιράζονται πολλά, όπως διαδικτυακές βάσεις δεδομένων (Web databases), προσωρινές η χωρικές βάσεις δεδομένων κ.α. Ένα αξιοσημείωτο παράδειγμα ενός πετυχημένου συνδυασμού εξόρυξής δεδομένων και βάσεων δεδομένων είναι η μηχανή αναζήτησης Google, η οποία εκτελεί εργασίες πολύ γρήγορα, πολύ αποδοτικά και με ακριβή αποτελέσματα σε οποιοδήποτε ερώτημα.

## 1.4 Τομείς εφαρμογής εξόρυξης δεδομένων

Η διαδικασία της ανεύρεσης γνώσης μέσα από βάσεις και η εξόρυξη δεδομένων έχει πολλές εφαρμογές. Μερικοί από τους τομείς εφαρμογής αναφέρονται παρακάτω :

### 1.4.1 Έρευνα

Δεδομένα συλλέγονται από αστρονόμους, ερευνητές του ανθρώπινου γονιδιώματος, βιοχημικούς που προσπαθούν να εξερευνήσουν τις ιατρικές ιδιότητες των πρωτεϊνών, και πολλούς άλλους ερευνητές.

### 1.4.2 Χρηματοοικονομικά

Πολύ μεγάλος αριθμός χρηματιστηριακών εταιριών χρησιμοποιούν τεχνικές εξόρυξης δεδομένων ώστε να γνωρίζουν πού να επενδύσουν. Στην πραγματικότητα μια μεγάλη μερίδα έρευνας στο τομέα εξόρυξης δεδομένων έχει γίνει έχοντας ως αφετηρία χρηματιστηριακές εφαρμογές. Η εξόρυξη δεδομένων γίνεται από κείμενα και τεχνικές αναφορές επιχειρήσεων, για την επίτευξη μια πρόβλεψης της τάσης των μετοχών.


Εικόνα 2: Επιχειρηματική Ευφυΐα

### 1.4.3 Διαδίκτυο

Ο τομέας της εξόρυξης δεδομένων είχε άμεση εφαρμογή με επιτυχία στο διαδίκτυο. Το πιο δημοφιλές παράδειγμα εξόρυξης δεδομένων είναι η Google. Είναι πρακτικά αδύναμη η ακριβής μέτρηση του όγκου δεδομένων που υπάρχει αυτή τη στιγμή στον παγκόσμιο ιστό, όμως κάθε ερώτημα στην μηχανή αναζήτησης δεν ξεπερνά σε χρόνο τα δύο δευτερόλεπτα.

### 1.4.4 Ασφάλεια συστημάτων

Από τις σημαντικότερες και πιο επιτυχημένες εφαρμογές της εξόρυξης δεδομένων αποτελεί η πρόσληψη και αποφυγή διαφόρων τύπων απάτης. Μπορεί να έχουμε απάτες διαδικτύου, τις οποίες κάποιος μπορεί να αντιληφθεί με εντοπισμό περίεργων συναλλαγών. Επίσης, συναλλαγές που μπορεί να σχετίζονται με οικονομικές παρανομίες η άλλου είδους απάτες, που μπορούν να προληφθούν με την χρήση συστημάτων αναγνώρισης ανωμαλιών.

## 1.5 Ενδιαφέρουσες Προκλήσεις

Οι παραδοσιακές μέθοδοι ανάλυσης δεδομένων συχνά αντιμετώπιζαν πρακτικές δυσκολίες στο να ανταποκριθούν στις προκλήσεις που δημιουργούνταν από νέα σύνολα δεδομένων. Παρακάτω θα αναφερθούμε σε ορισμένες προσκλήσεις που αποτέλεσαν κίνητρο για την εξόρυξη δεδομένων.

- Κλιμάκωση(Scalability)

Καθημερινά παρατηρούμε πρόοδο στην παραγωγή και συλλογή δεδομένων και για αυτό το λόγο οι αλγόριθμοι εξόρυξης δεδομένων που θα κληθούν να αναλύσουν τα ογκώδη σύνολα δεδομένων όπως είναι gigabyte,terabyte ακόμη και peta-byte(=$10^{15}$ byte) θα πρέπει να είναι κλιμακωτοί. Η κλιμάκωση μπορεί επίσης να βελτιωθεί κάνοντας δειγματοληψία είτε αναπτύσσοντας παράλληλους και κατανεμημένους αλγόριθμους.

- Πολλές Διαστάσεις ( High Dimensionality)

Μερικές δεκαετίες πριν συνήθιζαν να συναντούν μικρά σύνολα δεδομένων σε αντίθεση με τώρα που συναντάμε χιλιάδες η εκατοντάδες χαρακτηριστικά. Τα σύνολα δεδομένων με χρονικά η χωρικά στοιχεία τείνουν να έχουν πολλές διαστάσεις. Για παράδειγμα αν θεωρήσουμε ένα σύνολο δεδομένων , το οποίο διαθέτει μετρήσεις πιστωτικών καρτών σε διάφορες περιοχές, οι συγκεκριμένες μετρήσεις λαμβάνονται επαναληπτικά για μια εκτεταμένη χρονική περίοδο, ενώ ο αριθμός των χαρακτηριστικών αυξάνεται αναλογικά με το πλήθος των μετρήσεων που λαμβάνονται. Τις περισσότερες φορές παραδοσιακές τεχνικές ανάλυσης δεδομένων που είχαν αναπτυχθεί δεν είναι επιθυμητές για πολυδιάστατα δεδομένα. Σε ορισμένους αλγόριθμους η υπολογιστική πολυπλοκότητα αυξάνεται ραγδαία καθώς αυξάνεται και το πλήθος των διαστάσεων των δεδομένων.

- Μη παραδοσιακή ανάλυση (Not-traditional analysis)

Βασίζεται σε ένα πρότυπο υπόθεσης και ελέγχου. Με διαφορετικά λόγια , προτείνεται μια υπόθεση, σχεδιάζεται ένα πείραμα για τη συλλογή δεδομένων και τα δεδομένα αναλύονται σε σχέση με την υπόθεση. Το αρνητικό είναι ότι η συγκεκριμένη διαδικασία είναι χρονοβόρα. Τα σύνολα δεδομένων που αναλύονται στην εξόρυξη δεδομένων είναι συνήθως αποτέλεσμα ενός προσεκτικά σχεδιασμένου πειράματος, ενώ συχνά αναπαριστούν καιροσκοπικά δείγματα αντί για τυχαία δείγματα δεδομένων. Επίσης, συχνά περιλαμβάνουν μη παραδοσιακούς τύπους δεδομένων.

- Ετερογενή και πολύπλοκα δεδομένα (Heterogeneous and Complex Data)

Τα σύνολα δεδομένων πολλές φορές διαχειρίζονται παραδοσιακές μεθόδους οι οποίες περιέχουν χαρακτηριστικά ίδιου τύπου, είτε κατηγορηματικά είτε συνεχή. Επιπλέον υπάρχει

μεγάλη ανάγκη για να μπορούν να διαχειριστούν ετερογενή χαρακτηριστικά. Στις τεχνικές που αναπτύσσονται για την εξόρυξη τέτοιων πολύπλοκων αντικειμένων πρέπει να λαμβάνουμε υπόψιν τις διάφορες σχέσεις που υπάρχουν μέσα στα δεδομένα, όπως την χωρική και την χρονική αυτοσυσχέτιση και τη συνεκτικότητα των γραφημάτων.

- Κυριότητα και διανομή δεδομένων ( Data Ownership and Distribution)

Πολλές φορές τα δεδομένα δεν είναι αποθηκεμένα σε μια μόνο θέση ή δεν αποτελούν ιδιοκτησία κάποιου οργανισμού. Ορισμένες σημαντικές προκλήσεις που αντιμετωπίζουν οι αλγόριθμοι δεδομένων είναι οι παρακάτω:

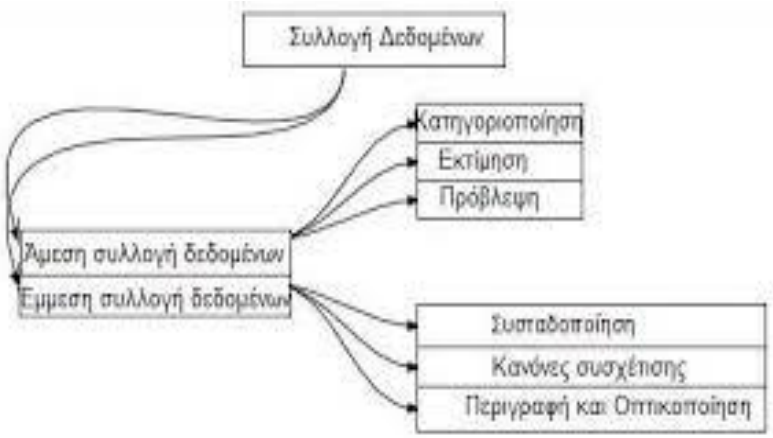
**i.** Υπάρχει περιορισμός της επικοινωνίας προκειμένου να εκτελεστεί ο κατανεμημένος υπολογισμός.

**ii.** Αποτελεσματική ενοποίηση των αποτελεσμάτων της εξόρυξης ,τα οποία λαμβάνονται από πολλαπλές πηγές.

**iii.** Χρειάζεται αντιμετώπιση για τα ζητήματα ασφάλειας δεδομένων.

## 1.6 Μέθοδοι Εξόρυξης Δεδομένων

Όπως η εξόρυξη γνώσης μπορεί να εφαρμοστεί σε διάφορους τύπους δεδομένων, αντίστοιχα μπορούμε να εφαρμόσουμε διαφορετικές μεθοδολογίες – τεχνικές και να ανακαλύψουμε αρκετούς τύπους μοτίβων (patterns) από τα δεδομένα μας. Να σημειώσουμε πως, η εξόρυξη δεδομένων ως αναπτυσσόμενος κλάδος, ανακαλύπτει διαρκώς νέες τεχνικές επεξεργασίας δεδομένων. Μερικές από τις σημαντικότερες μεθοδολογίες που θα αναλύσουμε κατατάσσονται ως εξής:




Εικόνα 3:Μεθοδολογίες εξόρυξης δεδομένων

και είναι οι παρακάτω:

- Κατηγοριοποίηση (Classifier)
- Συσταδοποίηση (Clustering)
- Παλινδρόμηση (Regression)
- Κανόνες συσχέτισης (Correlation rules)
- Πρότυπα Ακολουθιών ( Sequence patterns)

### 1.6.1 Κατηγοριοποίηση (Classifier)

Η κατηγοριοποίηση (classification) αποτελεί μια από τις βασικές μεθοδολογίες εξόρυξης δεδομένων. Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Τα αντικείμενα που πρόκειται να κατηγοριοποιηθούν αναπαριστάνονται γενικά από τις εγγραφές της βάσης δεδομένων και με τη διαδικασία της κατηγοριοποίησης γίνεται ανάθεση κάθε εγγραφής σε κάποιες από τις προκαθορισμένες κατηγορίες.

Για παράδειγμα, ένα κατάστημα ηλεκτρικών συσκευών μπορεί να κατηγοριοποιήσει τους πελάτες του σύμφωνα με τις προτιμήσεις τους για τα προϊόντα που εμπορεύεται έτσι ώστε οι πωλητές της να μεγιστοποιούν τις πιθανότητες επιτυχίας.

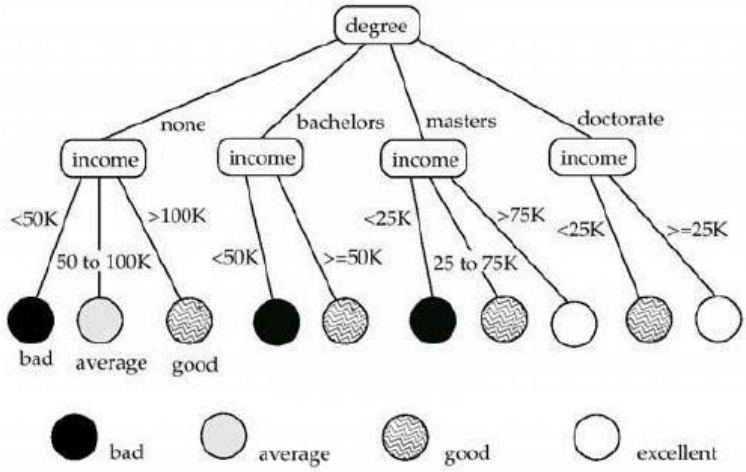
Στην μεθοδολογία αυτή, κυριότερο στόχο αποτελεί η παραγωγή ενός μοντέλου, το οποίο θα χρησιμοποιείται για να κατατάσσει τις νέες εγγραφές (δεδομένα), που δεν έχουν κατηγοριοποιηθεί σε άλλη κατηγορία. Η κατηγοριοποίηση, σαν διαδικασία, αποτελείται από 2 βήματα.

Το πρώτο βήμα αποτελείται από την εκμάθηση (learning) του μοντέλου. Αρχικά, γίνεται χρήση ενός μέλους δεδομένων, τα οποία ονομάζονται δεδομένα εκπαίδευσης (traininig data), προκειμένου να δημιουργηθεί το μοντέλο σύμφωνα με το οποίο θα κατηγοριοποιηθεί το σύνολο εκπαίδευσης (traininig set).

Η διαδικασία κατηγοριοποίησης (classification) αποτελεί το δεύτερο βήμα, όπου κατηγοριοποιούνται τα δεδομένα που δεν έχουν κατηγοριοποιηθεί. Στην περίπτωση που χρειάζεται έλεγχος αξιοπιστίας του μοντέλου, έχοντας το μοντέλο που προέκυψε από το προηγούμενο βήμα προσπαθούμε με χρήση δοκιμαστικών παραδειγμάτων (traininig samples) να επιβεβαιώσουμε την ακρίβεια του. Αν έχει τελικά μια αποδεκτή ακρίβεια τότε θα χρησιμοποιηθεί για την κατηγοριοποίηση νέων δεδομένων τα οποία δεν ανήκουν σε κάποια κατηγορία.

Στις περισσότερες περιπτώσεις υπάρχει ένα περιορισμένος αριθμός κατηγοριών και κάθε εγγραφή θα πρέπει να ανατεθεί στην κατάλληλη κατηγορία. Για αυτό το σκοπό χρησιμοποιούνται κάποιες τεχνικές, τις οποίες μπορούμε να κατατάξουμε σε δύο κατηγορίες. Η πρώτη χρησιμοποιεί δέντρα απόφασης (decision trees) και η δεύτερη, νευρωνικά δίκτυα (neural networks).

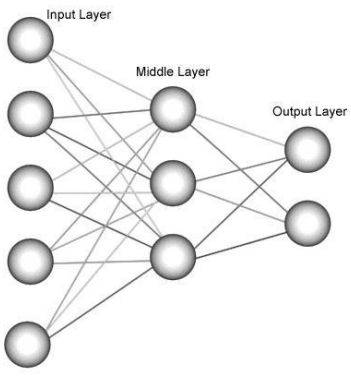
Τα δέντρα απόφασης είναι μοντέλα υποστήριξης αποφάσεων, τα οποία δημιουργούν κάποιους κανόνες, ώστε να ταξινομούν (κατηγοριοποιούν) ένα σύνολο δεδομένων. Το κάθε δέντρο αναπαριστά σύνολα από αποφάσεις (decisions). Κάθε κόμβος του δέντρου αναπαριστά ένα χαρακτηριστικό ενός αντικειμένου που πρόκειται να ταξινομηθεί, ενώ κάθε κλαδί που ξεκινά από τον κόμβο αυτό αντιστοιχεί σε μια από τις πιθανές τιμές του χαρακτηριστικού, τις οποίες ο κόμβος μπορεί να λάβει. Στο ακόλουθο σχήμα, βλέπουμε ένα δέντρο απόφασης που διακρίνει τους πελάτες μιας τράπεζας σε ‘’καλούς‘’ και ‘’κακούς‘’ ανάλογα με το επίπεδο εκπαίδευσής τους, αλλά και το εισόδημά τους.




Σχήμα 2: Δέντρο απόφασης

Τα νευρωνικά δίκτυα (neural networks) είναι επίσης μια διαδεδομένη μέθοδος ταξινόμησης (κατηγοριοποίησης). Συγκεκριμένα, είναι μια δομή που αποτελείται από ένα δίκτυο νευρώνων οι οποίοι συνδέονται μεταξύ τους. Οι νευρώνες χωρίζονται σε τρεις βασικές κατηγορίες:

- ✓ Τους νευρώνες εισόδου (Input neurons), οι οποίοι δέχονται πληροφορίες που θα υποστούν επεξεργασία,
- ✓ Τους νευρώνες εξόδου (Output neurons), στους οποίους καταλήγουν τα αποτελέσματα της επεξεργασίας και
- ✓ Τους ενδιάμεσους νευρώνες, οι οποίοι βρίσκονται μεταξύ των νευρώνων εισόδου και εξόδου.




Σχήμα 3: Νευρωνικό δίκτυο

### 1.6.2 Συσταδοποίηση (Clustering)

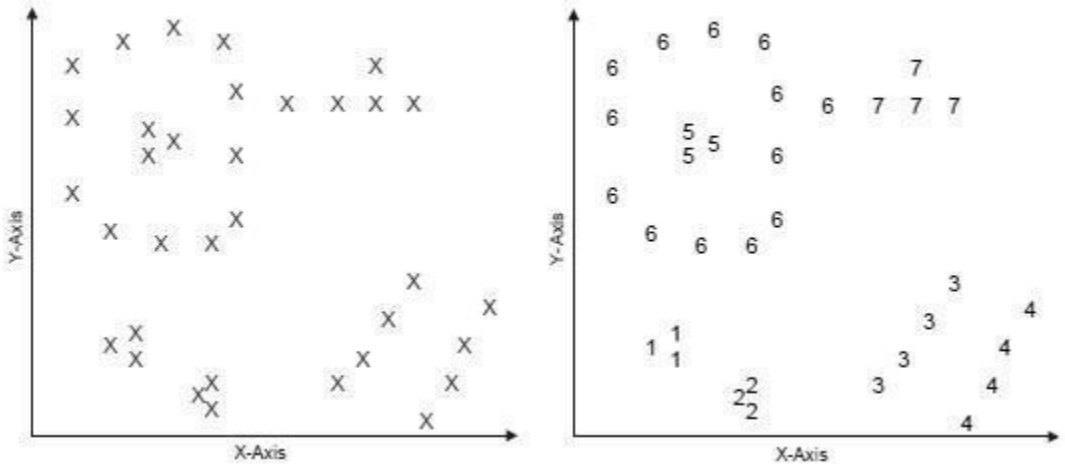
Η συσταδοποίηση (Clustering) είναι η εργασία του καταμερισμού ενός ετερογενούς πληθυσμού σε ένα σύνολο περισσότερων ετερογενών συστάδων (εύρεση συνόλων από αντικείμενα έτσι ώστε τα αντικείμενα ενός συνόλου να είναι περισσότερο όμοια ή να σχετίζονται μεταξύ τους και διαφορετικά με τα αντικείμενα των άλλων συνόλων. Αυτό που διαφοροποιεί τη συσταδοποίηση από την κατηγοριοποίηση είναι ότι η συσταδοποίηση δεν βασίζεται σε προκαθορισμένες κατηγορίες.

Επίσης ιδιαίτερα σημαντικό είναι για τις επιχειρήσεις να μπορούν να ομαδοποιούν τους πελάτες τους σε συγκεκριμένες κατηγορίες. Από την ανάλυση ενός πολύ μεγάλου συνόλου πελατών , μπορεί να μειωθεί το κόστος μια διαφημιστικής εκστρατείας που βασίζεται στην αποστολή μαζικών σύντομων μηνυμάτων sms. Αυτό γίνεται με τον περιορισμό του πλήθους των πελατών που απευθύνεται, επιλέγοντας αυτούς που έχουν μεγαλύτερη πιθανότητα να αντιδράσουν θετικά.

Οι αλγόριθμοι συσταδοποίησης χρειάζονται πλήθος δεδομένων και άρα χρειάζονται ένα μεγάλο πλήθος υπολογισμών. Έτσι, η πολυπλοκότητα εξαρτάται από το πλήθος των προς επεξεργασία δεδομένων.

Η διαδικασία συσταδοποίησης μπορεί να χωριστεί στα ακόλουθα βήματα:

- Επιλογή χαρακτηριστικών γνωρισμάτων, όπου επιλέγουμε τα γνωρίσματα (attributes) αυτά που θα μας συμπεριλάβουν τελικά την πληροφορία που χρειαζόμαστε,
- Αλγόριθμος συσταδοποίησης, όπου γίνεται επιλογή του καταλληλότερου αλγόριθμου συσταδοποίησης,
- Επικύρωση αποτελεσμάτων, όπου γίνεται έλεγχος της ακρίβειας των τελικών αποτελεσμάτων με χρήση διαφόρων μετρικών και κριτηρίων και
- Ερμηνεία και παρουσίαση αποτελεσμάτων. Γίνεται παρουσίαση και περαιτέρω ανάλυση της εξαχθείσας γνώσης με ειδικούς και από άλλους τομείς προκειμένου να χρησιμοποιηθεί κατά τον βέλτιστο τρόπο.

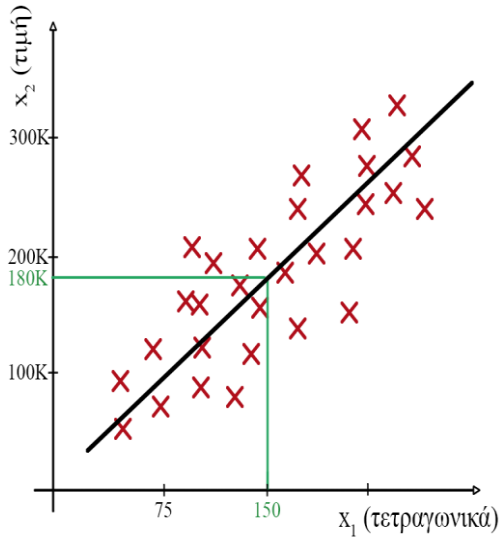



Σχήμα 4: Συσταδοποίηση ( Clustering)

### 1.6.3 Παλινδρόμηση (Regression)

Μια σχετική διαδικασία με την κατηγοριοποίηση είναι η παλινδρόμηση (regression), στόχος της οποίας είναι η μάθηση ή αλλιώς η εκπαίδευση (training) μιας συνάρτησης, η οποία απεικονίζει ένα αντικείμενο σε μία πραγματική μεταβλητή. Πρόκειται για μία προγνωστική μέθοδο. Στόχος είναι με βάση κάποιες ανεξάρτητες μεταβλητές (independent variables) να προβλεφθούν οι τιμές μιας εξαρτημένης μεταβλητής (dependent variable). Επίσης, η παλινδρόμηση είναι μια στατιστική ανάλυση που εμείς συχνά χρησιμοποιούμε για να κάνουμε πρόγνωση.

Ακριβώς παρακάτω παρουσιάζουμε ένα απλό παράδειγμα γραμμικής παλινδρόμησης. Οι μεταβλητές είναι τα τετραγωνικά ενός σπιτιού και η τιμή πώλησης του σε χιλιάδες Ευρώ. Η γραμμική παλινδρόμηση προσαρμόζει μια ευθεία στα δείγματα του συνόλου δεδομένων, τα οποία σηματοδοτούνται με κόκκινο Χ. Η προσαρμογή γίνεται με βάση μια συνάρτηση απόστασης ή συνάρτηση κόστους, την τιμή της οποία θέλουμε να ελαχιστοποιήσουμε. Έχοντας τη βέλτιστη ευθεία, δηλαδή την ευθεία που ελαχιστοποιεί την τιμή της συνάρτησης κόστους, μπορούμε να δώσουμε μια προσεγγιστικά καλή απάντηση σε ερωτήματα της μορφής: «Σε τι τιμές πωλούνται σπίτια των 150 τετραγωνικών;».




*Σχήμα 5: Παράδειγμα γραμμικής παλινδρόμησης*

### 1.6.4 Κανόνες συσχέτισης

Η εξαγωγή κανόνων συσχέτισης (association rules) θεωρείται μια από τις σημαντικότερες διεργασίες της εξόρυξης δεδομένων. Έχει προσελκύσει μεγάλο ενδιαφέρον επειδή:

i. παρέχουν ένα συνοπτικό τρόπο για να εκφραστούν οι ενδεχομένως χρήσιμες πληροφορίες που γίνονται εύκολα κατανοητές από τους τελικούς χρήστες,
ii. εφαρμόζονται ευρέως στην επιστήμη και την οικονομία,
iii. ανακαλύπτουν κρυμμένες «συσχετίσεις» μεταξύ των γνωρισμάτων ενός συνόλου δεδομένων.

Αυτοί οι συσχετισμοί (της 3$^{ης}$ περίπτωσης) παρουσιάζονται στην εξής μορφή: A→B όπου το A και το B αναφέρονται στα σύνολα γνωρισμάτων που υπάρχουν στα υπό ανάλυση δεδομένα. Ένα παράδειγμα ενός τέτοιου κανόνα συσχέτισης είναι το εξής:

Γάλα, Δημητριακά → Ψωμί για τοστ [sup = 5%, conf = 80%]

Ο κανόνας αυτός υποδεικνύει ότι οι πελάτες οι οποίοι αγοράζουν γάλα και δημητριακά μαζί , αγοράζουν και ψωμί για τοστ με ποσοστό υποστήριξης 5% και βεβαιότητας 80% για τον κανόνα. Με άλλα λόγια, το 80% των πελατών που αγοράζουν γάλα και δημητριακά, αγοράζουν επιπλέον και ψωμί για τοστ, και το 5% όλων των πελατών αγοράζουν όλα τα παραπάνω προϊόντα.

Η παραπάνω εφαρμογή των κανόνων συσχέτισης ονομάζεται ανάλυση του «καλαθιού της νοικοκυράς» (Market basket analysis), όπου σκοπός είναι να αναγνωριστούν τα αγαθά που αγοράζονται μαζί.

Στους κανόνες συσχέτισης το πρώτο μέρος ονομάζεται υπόθεση, ενώ το δεύτερο συμπέρασμα. Η ισχύς του κάθε κανόνα κρίνεται από :

- τα ποσοστά υποστήριξης (support)
- τα ποσοστά εμπιστοσύνης (confident) και
- τα ποσοστά κάλυψης (coverage)

Ο κανόνας X → Y ισχύει στο σύνολο των δεδομένων με εμπιστοσύνη (confidence) c, αν c% των εγγραφών που περιέχουν το X, περιέχουν επιπλέον και το Y. Εμπιστοσύνη λέγεται δηλαδή το ποσοστό των εγγραφών σε ένα σύνολο αντικειμένων που δεδομένου ότι περιλαμβάνουν τα γνωρίσματα του πρώτου μέρους του κανόνα, περιλαμβάνουν και το δεύτερο.Το πρόβλημα της εύρεσης κανόνων συσχέτισης εστιάζεται στην εύρεση όλων των κανόνων που έχουν μία καθορισμένη από το χρήστη ελάχιστη τιμή υποστήριξης και εμπιστοσύνης.

### 1.6.5 Πρότυπα ακολουθιών

Η εξόρυξη πρότυπων ακολουθιών (sequential patterns) είναι η εξόρυξη των συχνά εμφανιζόμενων προτύπων σχετικών με το χρόνο ή άλλες ακολουθίες. Οι περισσότερες μελέτες στα πρότυπα ακολουθιών επικεντρώνονται στα συμβολικά πρότυπα. Ο χρήστης εδώ μπορεί να προσδιορίσει περιορισμούς στα είδη των πρότυπων ακολουθιών που εξάγονται με την παροχή των προσχεδίων προτύπων (template patterns) υπό τη μορφή σειριακών επεισοδίων, παράλληλων επεισοδίων ή κανονικών εκφράσεων. Σημειωτέον ότι στην περίπτωση αυτή μας ενδιαφέρει η σειρά εμφάνισης των στοιχείων (γεγονότων). Παρακάτω αναφέρονται μερικά παραδείγματα:

- Ακολουθία από προσπελάσεις σελίδων στο διαδίκτυο,
- Ακολουθία στο δανεισμό βιβλίων από μια βιβλιοθήκη,
- Ακολουθία πακέτων που οδήγησαν σε επίθεση σε κάποιον υπολογιστή.

## 1.7 Ανακάλυψη Γνώσης

Επειδή δημιουργούνται όλο και περισσότερες μεγάλες βάσεις δεδομένων, είναι απαραίτητη η εύρεση τεχνικών και καλύτερων τρόπων για την εξαγωγή σχέσεων από αυτές τις βάσεις. Οι περισσότερες σύγχρονες επιχειρήσεις μπορούν να έχουν ηλεκτρονική πρόσβαση σε πλήθος δεδομένων. Η φράση ανακάλυψη γνώσης σε βάσεις δεδομένων αναφέρεται στη συνολική διαδικασία που απαιτείται για την εξαγωγή γνώσης από μεγάλες βάσεις δεδομένων. Η ανακάλυψη γνώσης απαιτεί ένα ακέραιο πλαίσιο με καλά ενοποιημένα

εργαλεία για τον καλύτερο χειρισμό των δεδομένων, για την δημιουργία μοντέλων, για τον έλεγχο και την εκτίμηση αυτών, για την οπτική παρουσίαση των δεδομένων και για τα αποτελέσματα της μοντελοποίησης. Η ανακάλυψη γνώσης θέτει νέες προκλήσεις για την τεχνολογία των βάσεων δεδομένων. Η διαδικασία της ανακάλυψης γνώσης (Knowledge Discovery in Databases) περιλαμβάνει:

i. την επιλογή δεδομένων από κάποια βάση δεδομένων,
ii. τον καθαρισμό - προεπεξεργασία,
iii. τον μετασχηματισμό,
iv. την εξόρυξη δεδομένων,
v. την επιλογή μοντέλου (ή συνδυασμού μοντέλων),
vi. την εκτίμηση και ερμηνεία των αποτελεσμάτων και
vii. την απομόνωση και χρήση της εξαχθείσας γνώσης (Fayyad, 1997).

Παρά το γεγονός ότι η εξόρυξη δεδομένων (Data mining) αναφέρεται σε ένα μόνο στάδιο της όλης διαδικασίας ανακάλυψης γνώσης, οι δύο όροι συχνά χρησιμοποιούνται αντικαθιστώντας ο ένας τον άλλο. Επιπλέον, αξίζει να σημειωθεί ότι η κύρια διαφορά μεταξύ ανακάλυψης γνώσης και μηχανικής μάθησης είναι ότι η πρώτη εφαρμόζεται σε πολύ μεγάλο όγκο δεδομένων τα οποία οργανώνονται σε βάσεις δεδομένων, ενώ στη μηχανική μάθηση τα δεδομένα είναι πολύ λιγότερα και προσεκτικά επιλεγμένα. Γενικά η διαδικασία εξόρυξης γνώσης όπως αναφέρεται στο βιβλίο Han & Kamber (2000) αποτελείται από μία διαδοχική ακολουθία βημάτων που είναι η παρακάτω:

i. Καθαρισμός των δεδομένων (απομάκρυνση θορύβων ή ασύμβατων δεδομένων),
ii. Ενοποίηση των δεδομένων (όπου μπορούν να συνδυαστούν πολλαπλές πηγές δεδομένων),
iii. Επιλογή δεδομένων (στο στάδιο αυτό επιλέγονται τα δεδομένα που θα χρησιμεύσουν στην έρευνα και ανακτώνται από τη βάση δεδομένων),
iv. Μετατροπή των δεδομένων (τα δεδομένα μετατρέπονται ή συγχωνεύονται σε μορφές που είναι κατάλληλες για εξόρυξη, μέσω διαδικασιών σύνοψης ή άθροισης),
v. Εξόρυξη δεδομένων (η διαδικασία κατά την οποία εφαρμόζονται ¨έξυπνες¨ μέθοδοι για να ανακαλυφθούν πρότυπα-patterns),
vi. Αξιολόγηση προτύπων (να αναγνωριστεί τα πραγματικά ενδιαφέροντα πρότυπα που οδηγούν στην δημιουργία γνώσης. Η αξιολόγηση γίνεται με βάση κάποιους δείκτες που δείχνουν την αποτελεσματικότητα, και αυτοί είναι οι confidence, support) και
vii. Παρουσίαση της γνώσης ( χρησιμοποιούνται τεχνικές παρουσίασης και τα δεδομένα οπτικοποιούνται, τα αποτελέσματα εμφανίζονται με τη μορφή διαγραμμάτων) .

Η διαδικασία εξόρυξης γνώσης από δεδομένα περιλαμβάνει τα ακόλουθα γενικά βήματα:

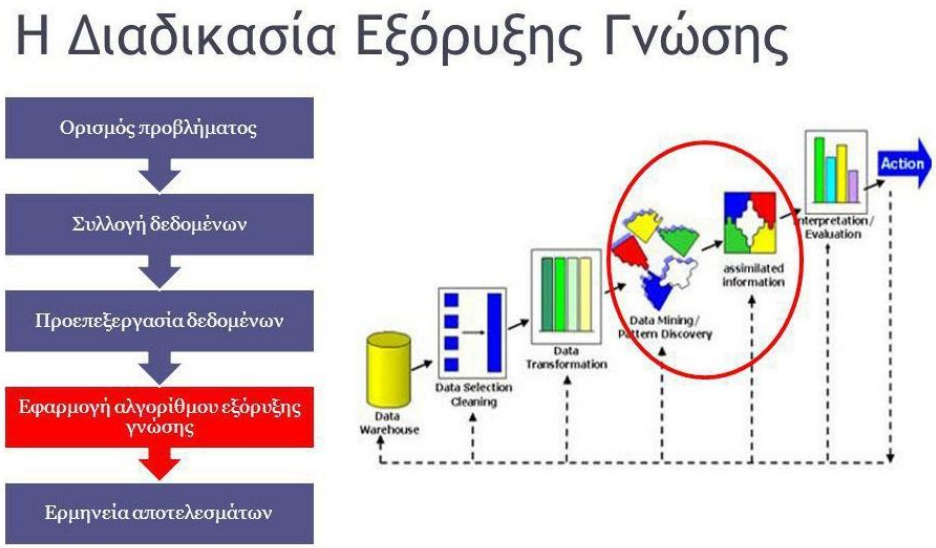
i. Ορισμός προβλήματος,

ii. Συλλογή δεδομένων,

iii. Προεπεξεργασία δεδομένων,

iv. Εφαρμογή αλγορίθμου εξόρυξης δεδομένων και

v. Ερμηνεία αποτελεσμάτων.




Σχήμα 6: Βασικά στάδια ανακάλυψης γνώσης από βάσεις δεδομένων.

## 1.8 Ρίζες Εξόρυξης Δεδομένων

Πολλοί ερευνητές ήρθαν σε συνεργασία και εστίασαν στην ανάπτυξη πιο αποτελεσματικών και κλιμακωτών εργαλείων τα οποία θα μπορούσαν να διαχειριστούν διαφόρους τύπους δεδομένων. Συγκεκριμένα η εξόρυξη δεδομένων βασίστηκε σε ιδέες όπως:

i. Η δειγματοληψία, η εκτίμηση, και έλεγχο από την επιστήμη της στατιστικής,

ii. Αλγόριθμοι αναζήτησης, τεχνικές μοντελοποίησης, και θεωρίες μάθησης από την τεχνητή νοημοσύνη, αναγνώριση προτύπων και μηχανική μάθηση.

Επίσης υιοθέτησε πολύ γρήγορα ιδέες και από άλλες περιοχές όπως βελτιστοποίηση, εξελικτικό υπολογισμό και ανάκτηση πληροφοριών. Ποιο συγκριμένα, τα συστήματα βάσεων δεδομένων είναι απαραίτητα για να παρέχουν υποστήριξη για την αποδοτική αποθήκευση, τη ευρετηριοποίηση και επεξεργασία ερωτημάτων. Οι τεχνικές των

παράλληλων συστημάτων είναι συχνά σημαντικές για την αντιμετώπιση μεγάλου μεγέθους συνόλων δεδομένων. Το θέμα του μεγέθους αντιμετωπίζεται με κατανεμημένες τεχνικές οι οποίες είναι απαραίτητες όταν τα δεδομένα δεν μπορούν να συγκεντρωθούν σε μια τοποθεσία.


Εικόνα 4: Ρίζες εξόρυξης δεδομένων

# Κεφάλαιο 2ο: Μηχανική Μάθηση

Στο παρόν κεφάλαιο θα παρουσιάσουμε την σημασία του Machine Learning. Στόχος είναι να αναπτύξουμε, αφενός μια σφαιρική εικόνα γύρω από τη θεωρία και το σκεπτικό που κρύβεται πίσω από το Machine Learning και αφετέρου την δεξιότητα εφαρμογής του. Επιπλέον παρακάτω θα γίνει ανάλυση του προγράμματος Μηχανικής Μάθησης (Machine Learning ) που χρησιμοποιήθηκε για την ανάλυση των δεδομένων μας και της μεθόδου PCA που χρειάστηκε για να μετατρέπουν τα δεδομένα μας σε αριθμούς.

## 2.1 Εισαγωγή στην Μηχανική Μάθηση (Machine Learning)

Η μηχανική μάθηση ή Γνωσιακές Μηχανές (Machine learning), ένας κλάδος της Τεχνητής Νοημοσύνης, είναι ένα επιστημονικό πεδίο που αναφέρεται στο σχεδιασμό και την ανάπτυξη αλγορίθμων που δέχονται ως είσοδο ( input ) εμπειρικά δεδομένα, όπως εκείνα που προέρχονται από αισθητήρες (Sensors) ή βάσεις δεδομένων, και δίνει σχέδια ή σχετικές προβλέψεις για τα χαρακτηριστικά των εμπλεκόμενων μηχανών που δημιούργησαν τα δεδομένα. Τα κύρια χαρακτηριστικά των άγνωστων βασικών κατανομών πιθανοτήτων μπορούν να γίνουν γνωστά έτσι ώστε τα δεδομένα να χρησιμοποιούν με αποδοτικό τα τρόπο από ένα εκπαιδευόμενο. Τέτοια δεδομένα μπορούν να θεωρηθούν ως περιπτώσεις πιθανών σχέσεων μεταξύ των παρατηρούμενων μεταβλητών.

Ένα από τα κύρια αντικείμενα της έρευνας της μάθησης μηχανών είναι ο σχεδιασμός αλγορίθμων που αναγνωρίζουν πολυσύνθετα σχέδια και να λάβουν νοήμονες αποφάσεις βασισμένες στα δεδομένα της εισόδου. Μια βασική δυσκολία είναι ότι η ομάδα όλων των δυνατών συμπεριφορών με όλα τα πιθανά δεδομένα εισόδου είναι πολύ μεγάλη, για να συμπεριληφθεί σε ένα σύνολο δεδομένων που έχουν παρατηρηθεί (επιλεγμένα δεδομένα [training data]). Με βάση τα προηγούμενα ο εκπαιδευόμενος (learner) πρέπει να γενικεύσει από τα δεδομένα παραδείγματα έτσι ώστε να μπορεί να παράγει χρήσιμα συμπεράσματα για νέα προβλήματα. ***Το 1959, ο Άρθουρ Σάμουελ ορίζει τη μηχανική μάθηση ως "Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί" για τέτοιο σκοπό.*** Είναι ένας Αμερικανός καθηγητής του πανεπιστημίου Stanford πρωτοπόρος στα πεδία τεχνητής νοημοσύνης και παιχνιδιών υπολογιστών. Το 1956 παρουσίασε ένα πρόγραμμα στον υπολογιστή για το παιχνίδι ντάμας (Checkers – playing program), που θεωρείται ως το πρώτο πρόγραμμα αυτομάθησης υπολογιστή.

Εικόνα 5: Παίζοντας πούλια στο 701. Στις 24 Φεβρουαρίου 1956, το πρόγραμμα Checkers του Arthur Samuel, το οποίο αναπτύχθηκε για να παίξει στο IBM 701, προβλήθηκε στο κοινό μέσω τηλεόρασης

Αναφέρεται ότι μια μηχανή μαθαίνει κάθε φορά που αλλάζει την δομή, το πρόγραμμα ή δεδομένα, που βασίζονται στα δεδομένα εισόδου ή σε ανταπόκριση εξωτερικών πληροφοριών, έτσι ώστε η αναμενόμενη απόδοση να βελτιωθεί. Τέτοιες αλλαγές όπως η πρόσθεση εγγράφων σε μια βάση δεδομένων, ανήκουν σε δικαιοδοσίες άλλων γνωστικών αντικειμένων και είναι γνωστές ως μάθηση . *Για παράδειγμα, όταν η απόδοση μίας μηχανής αναγνώρισης ομιλίας βελτιώνεται μετά από το άκουσμα διαφόρων δειγμάτων της ομιλίας ενός ατόμου, κατανοούμε και μπορούμε να πούμε ότι η μηχανή έχει μάθει*. Η μάθηση Μηχανών συνήθως συνδέεται με αλλαγές σε συστήματα που εκτελούν διαδικασίες (αναγνώριση, διάγνωση, σχεδιασμός, έλεγχος ρομπότ, προβλέψεις , κτλ.), που σχετίζονται με την τεχνητή νοημοσύνη. Διάφορα γνωστικά αντικείμενα έχουν χρησιμοποιηθεί στην Μάθηση Μηχανής, όπως:

- Στατιστική (Anderson 1958),
- Προσαρμοσμένη θεωρία ελέγχου ( Bolling et all. 1988, Sutton et all. 1987),
- Γενετικός προγραμματισμός ( Koza 1992-1994).

## 2.2 Διαδικασίες Μάθησης

Όπως οι άνθρωποι μαθαίνουν με ποικίλους τρόπους από το περιβάλλον τους έτσι και τα νευρωνικά δίκτυα προσαρμόζουν την λειτουργία τους και μέσω μιας αρκετά πολύπλοκης διαδικασίας μάθησης επιτυγχάνεται η εκπαίδευσή τους. Ο στόχος του τομέα της μηχανικής μάθησης είναι η δημιουργία των συστημάτων πληροφορικής που μαθαίνουν από την εμπειρία και που είναι ικανές να προσαρμοστούν στα περιβάλλοντά τους. Οι μαθησιακές τεχνικές και μέθοδοι που αναπτύχθηκαν από τους ερευνητές στον τομέα αυτό, έχουν εφαρμοστεί με επιτυχία σε ποικιλία μαθησιακών δραστηριοτήτων σε ένα ευρύ φάσμα τομέων, συμπεριλαμβανομένων για παράδειγμα την ταξινόμηση κειμένου, την ανακάλυψη γονιδίων, οικονομικές προβλέψεις, την ανίχνευση της απάτης σε πιστωτική κάρτα, συνεργατικό φιλτράρισμα, τον σχεδιασμό της προσαρμοστικής πρακτόρων και άλλων ιστοσελίδων. ils J. Nilsson, Kumagai Professor of Engineering (Emeritus) Stanford in 1958.

## 2.3 Εφαρμογές Μάθησης Μηχανών

Σύμφωνα με το έγκυρο διεθνές επιστημονικό σχήμα ταξινόμησης της ACM για τα γνωστικά αντικείμενα της πληροφορικής, η μάθηση μηχανών είναι ένας κλάδος της τεχνητής νοημοσύνης (Artificial Intelligence), που αποτελεί μια υποκατηγορία των μεθοδολογιών υπολογισμών , βασικής κατηγορίας της πληροφορικής. Το πολυδιάστατο φάσμα εφαρμογών της μάθησης μηχανών περιλαμβάνει τα ακόλουθα σημαντικά γνωστικά αντικείμενα:

**Εφαρμογές Προηγμένων Υπολογισμών**

- Νόηση Μηχανών ( Machine perception)
- Ενόραση Υπολογιστών ( Computer vision)
- Επεξεργασία Φυσικών Γλωσσών ( Natural language processing)
- Μηχανές Αναζήτησης (Search engines)
- Αναγνώριση Σχεδίων ( Syntactic pattern recognition)

**Προηγμένοι Υπολογισμοί και οικονομία**

- Ανάλυση Αγοράς Μετοχών (Stock market analysis)
- Ανίχνευση Απάτης Πιστωτικών Καρτών ( Detecting credit card fraud)
- Υπολογιστική Οικονομία ( Computational Finance)

**Γενικές Εφαρμογές Υπολογισμών**

- Παίξιμο Παιχνιδιών ( Game playing)
- Εξόρυξη ακολουθιών ( Sequence mining)
- Μετακινήσεις Ρομπότ ( Robot locomotion)
- Ανάκτηση Πληροφοριών ( Information retrieval)
- Προσαρμοζόμενοι Ιστότοποι ( Adaptive websites)
- Συνιστώμενα Συστήματα ( Recommender systems)

**Προηγμένοι Υπολογισμοί και Ιατρική**

- Δομημένος έλεγχος Υγείας ( Structural health monitoring)
- Ταξινόμηση Ακολουθιών DNA ( Classifying DNA sequences)
- Αλληλεπιδράσεις μεταξύ των εγκεφαλικών περιοχών (Neuroscience)
- Χημειοπληροφορική ( Cheminformatics)

## 2.4 Μορφές Μάθησης

Οι νοήμονες (ή ευφυείς) πράκτορες αποτελούν σύγχρονα συστήματα Τεχνητής Νοημοσύνης στα οποία δυνητικά μπορούν να χρησιμοποιηθούν επιλεκτικά και σε συνδυασμό μέθοδοι αναπαράστασης γνώσης και επίλυσης προβλημάτων. Με αυτό ως γνώμονα παρουσιάζονται παρακάτω οι διαφορετικές βασικές αφηρημένες αρχιτεκτονικές των νοημόνων πρακτόρων.

Οι πράκτορες μάθησης μπορεί να περιλαμβάνουν :

- ορισμένα στοιχεία εκτέλεσης τα οποία αποφασίζουν ποιες ενέργειες θα πραγματοποιηθούν και
- στοιχεία μάθησης που τροποποιούν τα στοιχεία εκτέλεσης έτσι ώστε να λαμβάνουν καλύτερες αποφάσεις.

Οι ερευνητές μηχανικής μάθησης χρησιμοποιούν μία μεγάλη ποικιλία στοιχείων μάθησης των οποίων η σχεδίαση επηρεάζεται από το περιβάλλον στο οποίο εφαρμόζονται. Η σχεδίαση αυτή επηρεάζεται από τους ακόλουθους παράγοντες:

- ποιες συνιστώσες στοιχείων εκτέλεσης πρέπει να κοινοποιηθούν,
- ποιες αναδράσεις πρόκειται να διατέθουν για την μάθηση των συνιστωσών αυτών και
- ποιες αναπαραστάσεις χρησιμοποιούνται για τις συνιστώσες.

Για την δημιουργία στοιχείων εκτέλεσης υπάρχουν διάφοροι τρόποι, ενώ οι συνιστώσες των πρακτόρων περιλαμβάνουν ορισμένες πληροφορίες και στόχους. Για κάθε μία από τις

συνιστώσες μπορεί να υπάρξει μάθηση μέσα από κατάλληλη ανάδραση. Ο τύπος ανάδρασης για τη μάθηση σχετίζεται με τον προσδιορισμό της φύσης μαθησιακών προβλημάτων που αντιμετωπίζουν οι πράκτορες. Η μηχανική μάθηση περιλαμβάνει τρεις διακεκριμένες κατηγορίες μάθησης.

- επιβλεπόμενη μάθηση ( supervised learning): μάθηση συνάρτησης από παραδείγματα εισόδων και εξόδων πρακτόρων,
- μη επιβλεπόμενη ( unsupervised learning): μάθηση προτύπων εισόδων χωρίς να δίνονται συγκεκριμένες τιμές εξόδων,
- ενισχυτική μάθηση: χρήση παρατημένων ειδών αναδράσεων ( ανταμοιβές ή ενισχύσεις) για μάθηση σχεδόν βέλτιστης πολιτικής για το περιβάλλον.

**Για την λειτουργία αλγορίθμων μάθησης** χρειάζονται αναπαραστάσεις γνωστών πληροφοριών, οι οποίες αντιστοιχούν στις συνιστώσες πρακτόρων, π.χ. αλγόριθμος μάθησης για πιθανοτικές περιγραφές ( δίκτυα Bayes) για συμπερασματικές συνιστώσες πρακτόρων θεωρίας αποφάσεων, αλγόριθμος μάθησης για γραμμικά σταθμισμένα πολυώνυμα για συναρτήσεις χρησιμότητας σε προγράμματα παιχνιδιών κ.α.

**Για το σχεδιασμό μαθησιακών συστημάτων** χρειάζεται η διαθεσιμότητα προηγούμενων γνώσεων. Το μεγαλύτερο μέρος της ανθρώπινης μάθησης προκύπτει μέσα στα πλαίσια του μεγάλου όγκου σχετικών γνώσεων και πληροφοριών, που βρίσκονται στο διαδίκτυο και σε σχετικές βάσεις δεδομένων.

### 2.4.1 Επιβλεπόμενη Μάθηση ( Supervised Learning )

Η διαδικασία μηχανικής μάθησης σύμφωνα με την οποία εξάγεται μια συνάρτηση από κατηγοριοποιημένα δεδομένα ονομάζεται “επιβλεπόμενη μάθηση” (ή όπως είναι γνωστή στον ευρύτερο κλάδο, supervised learning). Τα δεδομένα εκπαίδευσης αποτελούνται από ένα σύνολο από στιγμιότυπα κατάλληλα για την εκπαίδευση του αλγορίθμου. Στην επιβλεπόμενη μάθηση, κάθε στιγμιότυπο είναι ένα ζευγάρι το οποίο αποτελείται από ένα αντικείμενο εισόδου (συνήθως ένα διάνυσμα) και μία δεδηλωμένη τιμή εξόδου η οποία τυπικά μπορεί να ονομαστεί σήμα ελέγχου. Ένας αλγόριθμος επιβλεπόμενης μάθησης αναλύει τα δεδομένα εκπαίδευσης και δημιουργεί μια συνάρτηση ικανή να χρησιμοποιηθεί για να κατηγοριοποιήσει νέα δείγματα. Ένα βέλτιστο σενάριο θα επέτρεπε στον αλγόριθμο να καθορίζει σωστά τις ετικέτες κλάσης σε άγνωστα περιστατικά. Αυτή η ανάγκη απαιτεί τη γενικευμένη κατασκευή ενός γενικότερου αλγορίθμου έτσι ώστε από τα δεδομένα εισόδου με

λογικό τρόπο, να αποδίδει εξίσου καλά σε άγνωστα συμβάντα. Για να λύσουμε ένα πρόβλημα επιβλεπόμενης μάθησης, πρέπει να ακολουθήσουμε τα παρακάτω βήματα:

- Να καθορίσουμε τον τύπο των παραδειγμάτων εκπαίδευσης. Πριν κάνουμε οτιδήποτε άλλο, πρέπει να αποφασίσουμε τί είδους δεδομένα θα χρησιμοποιηθούν σαν σύνολο εκπαίδευσης (training set). Στην περίπτωση της ανάλυσης γραφικού χαρακτήρα για παράδειγμα, αυτό θα μπορούσε να αποτελείται από ένα απλό χαρακτήρα γραμμένο με το χέρι, μια ολόκληρη λέξη ή ακόμα και μια ολόκληρη χειρόγραφη πρόταση. Στην περίπτωση της ανάλυσης συναισθημάτων, αυτά αποτελούνται από δείγματα κειμένων με θετικές ή αρνητικές ετικέτες.
- Στην συνέχεια πρέπει να σκεφτούμε για το πώς θα συγκεντρώσουμε όλο το επιθυμητό σύνολο των διδομένων εκπαίδευσης. Αυτό το σύνολο θα πρέπει να είναι αντιπροσωπευτικό της χρήσης της συνάρτησης στον πραγματικό κόσμο. Συνεπώς, ένα σύνολο από αντικείμενα εισόδων συγκεντρώνεται μαζί με το αντίστοιχο σύνολο επιθυμητών και δεδηλωμένων αποτελεσμάτων, όπως έχει συλλεχθεί είτε από ειδικούς είτε από κατάλληλες μετρήσεις – αναζητήσεις.
- Έπειτα συνεχίζοντας θα πρέπει να καθοριστεί η αναπαράσταση των χαρακτηριστικών εισόδου της συνάρτησης εκπαίδευσης. Η ακρίβεια της συνάρτησης εξαρτάται σημαντικά από τη δομή των αντικειμένων εισόδου. Τυπικά ένα αντικείμενο εισόδου μετατρέπεται σε ένα διάνυσμα χαρακτηριστικών, το οποίο περιλαμβάνει έναν αριθμό από χαρακτηριστικά τα οποία περιγράφουν ικανοποιητικά όλο το αντικείμενο. Ο αριθμός των αντικειμένων δεν θα πρέπει να είναι πολύ μεγάλος, γιατί τότε διογκώνεται η διάσταση του παραγόμενου χώρου χαρακτηριστικών, όμως θα πρέπει να περιέχει τόση πληροφορία ώστε να μπορεί να προβλεφθεί με ακρίβεια η έξοδος του συστήματος.
- Το επόμενο βήμα είναι η επιλογή της δομής της συνάρτησης και του αλγορίθμου εκμάθησης, για παράδειγμα ο εκάστοτε μηχανικός μπορεί να επιλέξει να χρησιμοποιήσει είτε δέντρα εκμάθησης, είτε τον support vector machines, είτε τον Naïve Bayes αλγόριθμο. Συνήθως αυτή η επιλογή γίνεται μετά από επάλληλες δοκιμές και ανάλυση της ακρίβειας και προσαρμογής σε τυχαία δεδομένα. Ολοκληρώνοντας την διαδικασία, ο αλγόριθμος θα πρέπει να τρέξει με όλα τα συγκεντρωμένα δεδομένα εκπαίδευσης για να δημιουργήσουμε την συνάρτηση αντιστοίχισης. Μερικοί αλγόριθμοι επιβλεπόμενης μάθησης απαιτούν από τον χρήστη να καθορίσει ορισμένες παραμέτρους ελέγχου. Αυτές οι παράμετροι μπορούν να προσαρμοστούν είτε με βελτιστοποίηση της απόδοσης μετά από δοκιμές σε ένα

υποσύνολο δεδομένων (validation set), είτε με cross-validation που είναι μια τεχνική που θα αναλυθεί στην πορεία.

- Στο τέλος, θα πρέπει να αξιολογηθεί η απόδοση του μοντέλου εκμάθησης. Αφού προσαρμόσουμε όλες τις παραμέτρους και εκπαιδεύσουμε τον αλγόριθμο μας, θα πρέπει να υπολογίσουμε την απόδοση της συνάρτησης κατηγοριοποίησης εφαρμόζοντας την σε ένα δείγμα δεδομένων ελέγχου (test set), το οποίο μάλιστα θα πρέπει να είναι διαφορετικό από το training set.

Υπάρχει μια ευρεία ποικιλία διαθέσιμων αλγορίθμων επιβλεπόμενης μάθησης, ο καθένας με τα προτερήματα και τις αδυναμίες του φυσικά. Δεν υπάρχει ωστόσο κανένας από αυτούς ο οποίος να αποδίδει εξίσου καλά σε όλα τα προβλήματα επιβλεπόμενης μάθησης. Η γενικευμένη λογική που ακολουθεί αυτή η προσέγγιση αναλύεται ακριβώς παρακάτω, και μπορεί να προσαρμοστεί με την χρήση διαφόρων αλγορίθμων μετά από ειδική μελέτη και προσαρμογή στο εκάστοτε πρόβλημα. Πιο συγκεκριμένα, δεδομένου ενός συνόλου N από παραδείγματα εκπαίδευσης της μορφής $\{(x_1,y_1),\ldots( x_n, y_n)\}$ τέτοια ώστε το $x_i$ να είναι το διάνυσμα χαρακτηριστικών του i-στου παραδείγματος και το $y_i$ να είναι η ετικέτα κατηγορίας του, ένας αλγόριθμος μάθησης υπολογίζει μια συνάρτηση $g : X \rightarrow Y$ , όπου το X είναι ο χώρος των δεδομένων εισόδου (Input space) και Y είναι ο χώρος των προβλεπόμενων τιμών εξόδου (Output space). Η συνάρτηση g αποτελεί ένα αντικείμενο ενός χώρου από πιθανές συναρτήσεις G, που ονομάζεται συνήθως "υποθετικός χώρος". Μερικές φορές είναι πιο βολικό να παρουσιάζουμε την g ως μια συνάρτηση βαθμολόγησης $f :X \times Y \rightarrow \mathbb{R}$, έστω ότι F είναι ο χώρος όλων την συναρτήσεων βαθμολόγησης, έτσι ώστε η g να ορίζεται σαν την επιστρεφόμενη τιμή y που δίνει την υψηλότερη βαθμολογία στην συνάρτηση:

$$g(x) = \arg\max f(x ,y)$$

Στα μαθηματικά, τα επιχειρήματα των μεγίστων ή αλλιώς "the arguments of the maxima" (σε συντομογραφία arg max ή argmax), αντιπροσωπεύουν τα σημεία του πεδίου κάποιας συνάρτησης στην οποία μεγιστοποιούνται οι τιμές της συνάρτησης. Σε αντίθεση με τα μέγιστα όρια , αναφερόμενοι στις μεγαλύτερες *εξόδους* μιας συνάρτησης, το arg max αναφέρεται στις *εισόδους* ή στα επιχειρήματα στα οποία οι έξοδοι των λειτουργιών είναι όσο το δυνατόν μεγαλύτερες.

Αν και οι G και F μπορεί να αποτελούν οποιονδήποτε χώρο συναρτήσεων, οι περισσότεροι αλγόριθμοι μάθησης είναι πιθανολογικά μοντέλα όπου η g παίρνει την μορφή ενός μοντέλου δεσμευμένης πιθανότητας:

$$g(x)=P(y|x)$$

ή αντίστοιχα η f παίρνει τη μορφή ενός *μοντέλου με από κοινού πιθανότητες* που περιγράφεται από την σχέση που ακολουθεί:

$$f(x,y)=P(x,y)$$

Για παράδειγμα,

- ✓ ο Naïve Bayes και η γραμμική διαχωριστική ανάλυση (linear discriminant analysis) είναι **μοντέλα με από κοινού πιθανότητες**, ενώ
- ✓ η λογιστική παλινδρόμηση (logistic regression) είναι ένα **μοντέλο δεσμευμένης πιθανότητας.**

Γνωστότεροι αλγόριθμοι Επιβλεπόμενης Επαγωγικής Μάθησης είναι:

- Δένδρα Απόφασης (Decision Trees)
- Μάθηση βασισμένη σε Επεξηγήσεις (Explanation-Based Learning)
- Μάθηση βασισμένη σε Περιπτώσεις (Case-Based Learning)
- Μάθηση Νευρωνικών δικτύων (π.χ. για Backpropagation Neural Networks)
- Μάθηση Μέσω Στατιστικών Μεθόδων (π.χ. μάθηση κατά Bayes),
- Συλλογική Μάθηση από Ενδυνάμωση (Boosting) κ.ά.

### 2.4.2 Μη Επιβλεπόμενη Μάθηση

Η μη επιβλεπόμενη μάθηση (unsupervised learning) είναι μια διεργασία μηχανικής μάθησης κατά την οποία εξάγεται μία συνάρτηση για να περιγράψει κρυφές δομές από δεδομένα χωρίς ετικέτες. Εφόσον τα παραδείγματα που δίνονται στον μηχανισμό εκμάθησης είναι άγνωστης κατηγορίας, δεν υπάρχει λάθος, σφάλμα καθώς και ούτε κάποιο σήμα αξιολόγησης ώστε να γίνει εκτίμηση της πιθανής λύσης. Αυτό είναι άλλωστε και που ξεχωρίζει τις μη επιβλεπόμενες διαδικασίες από τις επιβλεπόμενες τεχνικές ανάλυσης.

Η μη επιβλεπόμενη μάθηση είναι στενά συνδεδεμένη με το πρόβλημα της εκτίμησης πυκνότητας στην στατιστική. Ωστόσο, ο συγκεκριμένος τρόπος μάθησης εμπλέκει και πολλές άλλες τεχνικές σύμφωνα με τις οποίες ψάχνει να συνοψίσει και να εξηγήσει χαρακτηριστικά κλειδιά των δεδομένων. Πολλές μέθοδοι που αναμειγνύονται σε αυτές τις τεχνικές βασίζονται σε μεθόδους εξόρυξης γνώσης που χρησιμοποιούνται κατά την προ επεξεργασία των δεδομένων. Την πιο σημαντική προσέγγιση της μη επιβλεπόμενης μάθησης αποτελεί η **συσταδοποίηση ή αλλιώς clustering** (π.χ. k-means, mixture models, hierarchical clustering, κτλ.)

Στην επιβλεπόμενη μάθηση (supervised learning) μας δίνεται ένα σύνολο δεδομένων με τις αντίστοιχες κλάσεις-ετικέτες κάθε εγγραφής. Στόχος είναι η δημιουργία ενός μοντέλου, το οποίο να μπορεί να κατηγοριοποιήσει νέα δεδομένα σε κάποια από τις προϋπάρχουσες κλάσεις. Αντίθετα, στη μη επιβλεπομένη μάθηση (unsupervised learning) μας δίνεται ένα σύνολο δεδομένων, χωρίς τις αντίστοιχες κλάσεις-ετικέτες κάθε εγγραφής και στόχος είναι η χρήση κάποιου αλγορίθμου, ώστε αυτόματα να ανακαλύψουμε κάποια ενδεχομένως ενδιαφέρουσα δομή των δεδομένων. Για παράδειγμα, η συσταδοποίηση είναι μια από τις τεχνικές μη επιβλεπόμενης μάθησης που χρησιμοποιείται ευρέως για αυτόν τον σκοπό. Δοθέντων κάποιων δεδομένων χωρίς κλάσεις, οι αλγόριθμοι συσταδοποίησης ομαδοποιούν τα δεδομένα σε συστάδες, έτσι ώστε εγγραφές, οι οποίες ανήκουν στην ίδια συστάδα, να έχουν όμοια ή παραπλήσια χαρακτηριστικά.

Στο πρόβλημα της συσταδοποίησης μας δίνεται ένα σύνολο δεδομένων χωρίς τις αντίστοιχες κλάσεις και χρειαζόμαστε κάποιον αλγόριθμο, ο οποίος θα ομαδοποιήσει αυτόματα τα δεδομένα σε συστάδες (clusters). Στις συστάδες που δημιουργούνται θέλουμε να διαχωρίζονται όσο το δυνατόν πιο ορθά τα δεδομένα. Αυτό πρακτικά σημαίνει ότι μια συστάδα θέλουμε να απαρτίζεται από αντικείμενα, όπου κάθε αντικείμενο είναι πιο κοντά σε κάθε άλλο αντικείμενο της ίδιας συστάδας απ’ ότι σε κάποιο άλλο αντικείμενο διαφορετικής συστάδας.

Ο αλγόριθμος k-means ξεκινάει με k τυχαία σημεία, τα οποία ονομάζονται κεντροειδή της συστάδας και δηλώνουν το κέντρο βάρους της συστάδας. Το k υποδηλώνει το πόσες συστάδες θέλουμε να δημιουργήσει ο αλγόριθμος καθώς εκτελεί επαναληπτικά δύο βήματα. Αρχικά, αφορά την ανάθεση σε κάποια συστάδα, ενώ έπειτα αφορά τον επαναπροσδιορισμό και τη μετατόπιση του κεντροειδούς κάθε συστάδας.

Πιο αναλυτικά, όσον αφορά στο πρώτο βήμα, δηλαδή την ανάθεση σε κάποια συστάδα, ο αλγόριθμος εξετάζει κάθε στιγμιότυπο σε σχέση με τα κεντροειδή των συστάδων. Με χρήση κάποιου μέτρου απόστασης, αναθέτει το εξεταζόμενο στιγμιότυπο στη συστάδα, της οποίας το κεντροειδές είναι το πλησιέστερο ως προς το συγκεκριμένο δείγμα. Στο δεύτερο βήμα, παίρνοντας τον μέσο όρο των δειγμάτων κάθε συστάδας, επαναϋπολογίζονται τα κεντροειδή της κάθε συστάδας, ώστε το κεντροειδές να είναι πιο αντιπροσωπευτικό στην πρόσφατα διαμορφωμένη συστάδα.

Ο αλγόριθμος εκτελεί επαναληπτικά αυτά τα δύο βήματα, μέχρι ότου τα κεντροειδή των συστάδων να μμετατοπίζονται ελάχιστα και σε απόσταση μικρότερη από κάποια δοθείσα τιμή κατωφλίου (threshold). Ως εναλλακτικό κριτήριο τερματισμού του αλγορίθμου μπορεί να χρησιμοποιηθεί και ο αριθμός επαναλήψεων του αλγορίθμου.

### 2.4.3 Ενισχυτική Μάθηση

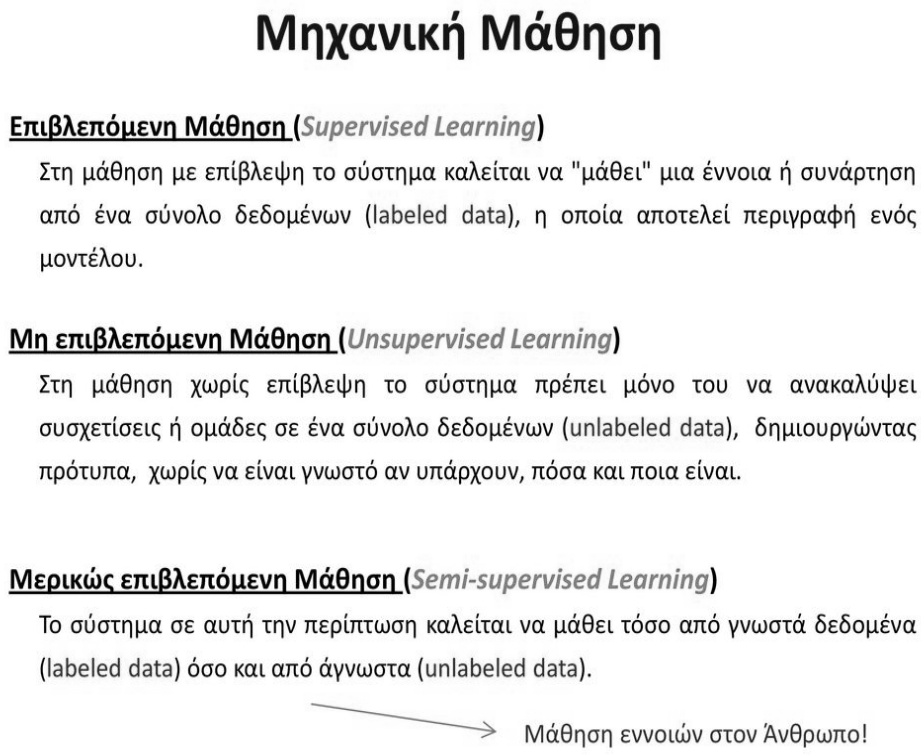
Η ενισχυτική μάθηση ( reinforcement learning) χρησιμοποιεί παρατημένες ειδικές αναδράσεις ( ανταμοιβές η ενισχύσεις ) για τη μάθηση σχεδόν βελτίστως πολιτικών για το περιβάλλον και σε πολλά πολύπλοκα πεδία θεωρείται ότι αποτελεί το μόνο εφικτό τρόπο για εκπαίδευση προγραμμάτων έτσι ώστε να επιτυγχάνονται υψηλά επίπεδα αποδόσεων. Βέλτιστη πολιτική είναι η πολιτική που μεγιστοποιεί την αναμενόμενη συνολική ανταμοιβή. Η ενισχυτική μάθηση θεωρείται ένας μικρόκοσμος του συνολικού προβλήματος της ΤΝ, που μελετάται σε διάφορα απλοποιημένα περιβάλλοντα για να διευκολύνεται η πρόοδος.

Η παθητική μάθηση ( passive learning) αναφέρεται στην μάθηση χρησιμότητας καταστάσεων ή ζευγών καταστάσεων ενεργειών, συμπεριλαμβανομένων και της μάθησης μοντέλων περιβάλλοντος. Η ενεργητική μάθηση ( active learning) αναφέρεται στην εξερεύνηση , όπου οι πράκτορες πρέπει να μάθουν τί κάνουν και να βιώσουν όσο το δυνατόν περισσότερα για το περιβάλλον με σκοπό να μάθουν πώς να συμπεριφέρονται μέσα σ αυτό.

Η εκμάθηση του τρόπου με τον οποίο συνδέονται οι καταστάσεις και η πληροφόρηση των περιορισμών μεταξύ των καταστάσεων μπορεί να χρησιμοποιηθεί ο *«προσαρμόσιμος δυναμικός προγραμματισμός»* ( adaptive dynamic programming), που μαθαίνει το μοντέλο μετάβασης περιβάλλοντος.

Η «*ιεραρχική ενισχυτική μάθηση»* (hierarchical reincorment learning) μπορεί να χρησιμοποιηθεί για επίλυση προβλημάτων ( σε πολλά αφαιρετικά επίπεδα ) με πολύπλοκες συμπεριφορές.

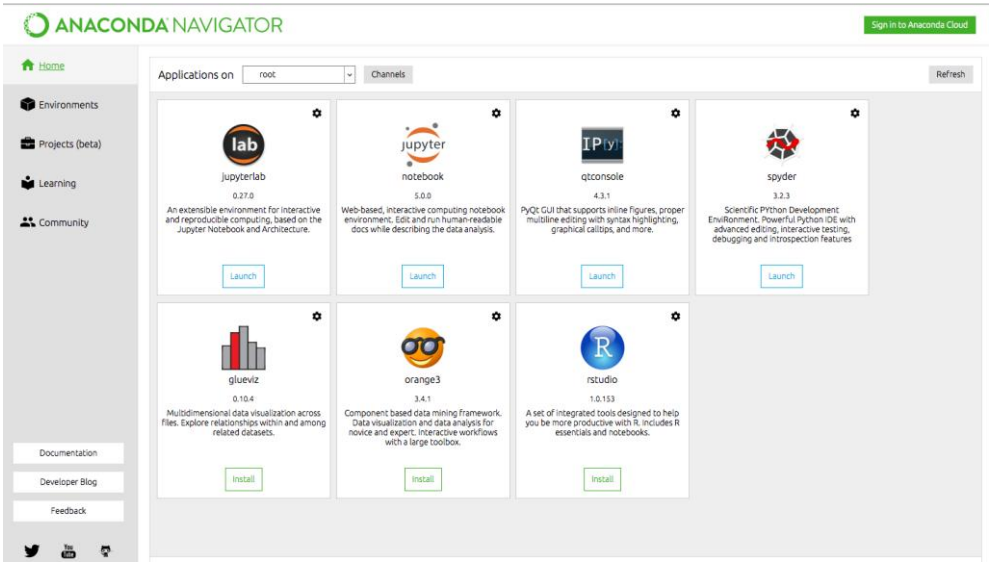
***Συνοψίζοντας, η μηχανική μάθηση διακρίνεται σε:***




Εικόνα 6:Μηχανική μάθηση

## 2.5 Anaconda

Το Anaconda είναι μια διανομή ελεύθερων και ανοιχτών πηγών των γλωσσών προγραμματισμού Python και R για εφαρμογές σχετικές με την επιστήμη των δεδομένων και την μηχανική μάθηση (επεξεργασία δεδομένων ευρείας κλίμακας, προγνωστική ανάλυση , επιστημονική πληροφορική ) με στόχο την απλοποίηση της διαχείρισης και της ανάπτυξης των πακέτων . Οι εκδόσεις των πακέτων διαχειρίζονται το *conda* σύστημα διαχείρισης




*Εικόνα 7: Anaconda cloud*

Στην παρούσα εργασία χρησιμοποιήθηκε Python 2 και το εργαλείο Jupyter της πλατφόρμας Anaconda cloud. Αρχικά ανοίγουμε το Jupyter και μας ανοίγει ένα command windows. Έπειτα με την εντολή Jupyter notebook ανοίγουμε ένα Cloud internet explorer. Στη συνέχεια κάνουμε upload τα δεδομένα μας και ξεκινάμε την ανάλυσή μας. Τα δεδομένα μας είναι σε αρχείο Excel.

## 2.6 Μέθοδος PCA

Η μέθοδος PCA (Ανάλυση Κύριων Συνιστωσών) αποτελεί μία γραμμική μέθοδο συμπίεσης Δεδομένων, η οποία συνίσταται από τον επαναπροσδιορισμό των συντεταγμένων ενός συνόλου δεδομένων σε ένα άλλο σύστημα συντεταγμένων το οποίο θα είναι καταλληλότερο στην επικείμενη ανάλυση δεδομένων. Αυτές οι νέες συντεταγμένες είναι το αποτέλεσμα ενός γραμμικού συνδυασμού προερχόμενου από τις αρχικές μεταβλητές και εκπροσωπούνται σε ορθογώνιο άξονα, ενώ τα επικείμενα σημεία διατηρούν μια φθίνουσα σειρά όσο αφορά στη τιμή της διακύμανσής τους**.** Για το λόγο αυτό, το πρώτο κύριο συστατικό (principal component) διατηρεί περισσότερες πληροφορίες δεδομένων σε σύγκριση με το δεύτερο το οποίο δεν διατηρεί πληροφορίες οι οποίες έχουν εισέλθει νωρίτερα (στο πρώτο συστατικό). Τα principal components δεν συσχετίζονται.

Η συνολική ποσότητα των principal components είναι ίση με τη ποσότητα των αρχικών μεταβλητών και παρουσιάζει τις ίδιες πληροφορίες στατιστικής. Εν τούτοις, η συγκεκριμένη μέθοδος επιτρέπει την μείωση του συνόλου των μεταβλητών, καθώς τα πρώτα συστατικά (principal components) διατηρούν περισσότερο από το 90% των στατιστικών δεδομένων από

τα αρχικά δεδομένα. Λόγω αυτών των σημαντικών πλεονεκτημάτων, η μέθοδος αυτή είναι ευρέως διαδεδομένη στην συμπίεση εικόνας.

### 2.6.1 Τα βήματα της μεθόδου

#### 1. Λήψη των Δεδομένων

Χρησιμοποιούμε τον δικό μας αυτοσχέδιο δισδιάστατο πίνακα δεδομένων που παρατίθεται στη συνέχεια, προκειμένου να δείξουμε τις μετατροπές που κάνουμε σε κάθε βήμα προκειμένου να επιτύχουμε τη συμπίεση κατά PCA.

Data =

| *X* | *Y* |
|---|---|
| 2.5 | 2.4 |
| 0.5 | 0.7 |
| 2.2 | 2.9 |
| 1.9 | 2.2 |
| 3.1 | 3.0 |
| 2.3 | 2.7 |
| 2 | 1.6 |
| 1 | 1.1 |
| 1.5 | 1.6 |
| 1.1 | 0.9 |

#### 2. Αφαίρεση του μέσου όρου

Προκειμένου να χρησιμοποιηθεί η μέθοδος της PCA κατάλληλα, θα πρέπει να αφαιρεθεί ο μέσος όρος από κάθε μία από τις διαστάσεις του πίνακα των στοιχείων. Ο μέσος όρος ο οποίος αφαιρείται είναι ο μέσος όρος των στοιχείων κάθε διάστασης. Επομένως όλες οι τιμές των ***x*** έχουν ως μέσο όρο το x=1.81 ο οποίος αφαιρείται από κάθε μία. Αναλόγως και όλες οι ***y*** τιμές έχουν μέσο όρο την τιμή y=1,91 η οποία και αυτή αφαιρείται από κάθε μία. Η διαδικασία αυτή παράγει ένα σύνολο δεδομένων με μέσο όρο ίσο με το μηδέν.

DataAdjust =

| *X* | *Y* |
|---|---|
| 0.69 | 0.49 |
| -1.31 | -1.21 |
| 0.39 | 0.99 |

0.09  0.29

1.29  1.09

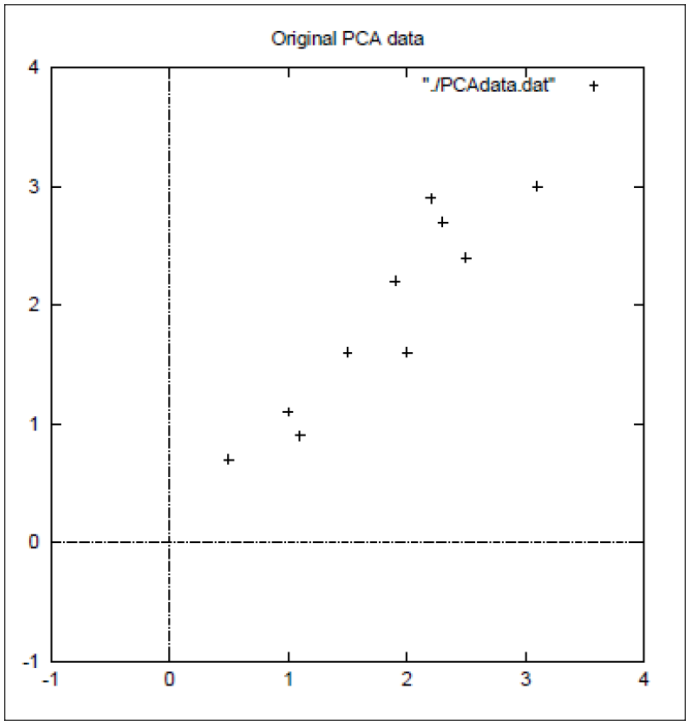
0.49  0.79

0.19  -0.31

-0.81  -0.81

-0.31  -0.31

-0.71  -1.01




Σχήμα 7 : Παράδειγμα δεδομένων με PCA, αρχικά δεδομένα αριστερά, τελικά (με αφαίρεση του μέσου όρου) δεξιά

### 3. Υπολογισμός του πίνακα Συνδιακύμανσης

Το παρακάτω βήμα γίνεται με τον τρόπο που περιγράφουμε αναλυτικά στην παράγραφο. Καθώς έχουμε στη διάθεσή μας δισδιάστατα δεδομένα, ο πίνακας συνδιακύμανσης θα έχει μέγεθος 2×2. Επομένως το αποτέλεσμα έχει την παρακάτω μορφή:

$$cov = \begin{pmatrix} 0.616555556 & 0.615444444 \\ 0.615444444 & 0.716555556 \end{pmatrix}$$

Δεδομένου ότι τα στοιχεία που δεν βρίσκονται στη διαγώνιο του πίνακα είναι θετικά, θα πρέπει να περιμένουμε πως και οι δύο μεταβλητές ***x, y*** θα αυξάνονται μαζί.

### 4. Υπολογισμός των ιδιοδιανυσμάτων και των ιδιοτιμών ενός πίνακα συνδιακύμανσης

Καθώς ο πίνακας συνδιακύμανσης είναι τετραγωνικός, είναι δυνατός ο υπολογισμός των ιδιοτιμών και των ιδιοδιανυσμάτων αυτού. Τα μεγέθη αυτά είναι πραγματικά σημαντικά, καθώς μέσω αυτών λαμβάνουμε χρήσιμες πληροφορίες για τα προς μελέτη στοιχεία.
Παρακάτω διαφαίνονται οι ιδιοτιμές (eigenvalues) και τα ιδιοδιανυσμάτων (eigenvectors) :

$$\text{Eigenvalues} = \begin{pmatrix} 0.04908339891 \\ 0.28402771 \end{pmatrix}$$

$$\text{eigenvectors} = \begin{pmatrix} -0.735178656 & -0.677873399 \\ -0.677873399 & -0.735178656 \end{pmatrix}$$

Είναι σημαντικό να αναφέρουμε πως τα ιδιοδιανύσματα αυτά είναι και τα δύο μοναδιαία, δηλαδή το μήκος τους είναι ίσο με τη μονάδα. Αυτό είναι πράγματι πολύ σημαντικό για την PCA, αλλά ευτυχώς σχεδόν τα περισσότερα μαθηματικά πακέτα προγραμμάτων υπολογίζουν τα ιδιοδιανύσματα όταν αυτά ζητούνται, ως μοναδιαία ιδιοδιανύσματα.

Εάν κοιτάξουμε την πλοκή των δεδομένων θα διαπιστώσουμε πως όπως προβλεπόταν από τον πίνακα συνδιακύμανσης, τα ιδιοδιανύσματα όντως αυξάνονται μαζί. Στην κορυφή των δεδομένων απεικονίζονται επίσης και δύο ιδιοδιανύσματα. Εμφανίζονται στο διάγραμμα ως διαγώνιες διακεκομμένες γραμμές. Όπως έχει ήδη επισημανθεί και παραπάνω τα ιδιοδιανύσματα είναι ορθογώνια μεταξύ τους.
Όμως εξίσου σημαντική είναι η δυνατότητα που μας δίνουν σχετικά με τα πρότυπα των στοιχείων, αφού όπως φαίνεται ένα από τα ιδιοδιανύσματα διέρχεται από τη μέση των σημείων, σαν να πρόκειται να σχεδιάζει μια γραμμή η οποία ταιριάζει με το βέλτιστο τρόπο ανάμεσα στα στοιχεία. Το επικείμενο ιδιοδιάνυσμα μας δείχνει πως αυτά τα δύο σύνολα δεδομένων σχετίζονται κατά μήκος αυτής της γραμμής. Το δεύτερο ιδιοδιάνυσμα μας δίνει το δεύτερο σημαντικό πρότυπο των δεδομένων, κατά το οποίο όλα τα στοιχεία που είναι ακολουθούν την κύρια γραμμή, αλλά και απέχουν από αυτή κατά ένα ποσό.

Επομένως μέσω της διαδικασίας υπολογισμού των ιδιοδιανυσμάτων ενός πίνακα συνδιακύμανσης, μας δίνεται η δυνατότητα απεικόνισης γραμμών στο χώρο οι οποίες φέρουν πληροφορίες σχετικά με τα στοιχεία μας. Το υπόλοιπο των βημάτων περιλαμβάνει τη μετατροπή των δεδομένων έτσι ώστε να είναι εκφρασμένα σε αυτές τις γραμμές.

### 5. Επιλογή των στοιχείων που θα αποτελέσουν το χαρακτηριστικό διάνυσμα

Σε αυτό το σημείο έρχεται η έννοια της συμπίεσης στοιχείων και της μείωσης των διαστάσεων. Αν λάβουμε υπόψη τα ιδιοδιανύσματα και τις ιδιοτιμές του προηγούμενου

παραδείγματος, θα δούμε ότι οι ιδιοτιμές είναι τελείως διαφορετικές τιμές. Στην πραγματικότητα αποδεικνύεται ότι το ιδιοδιάνυσμα με την υψηλότερη ιδιοτιμή είναι η κύρια συνιστώσα (principal component) του συνόλου των στοιχείων.

Γενικά λοιπόν, όταν υπολογιστούν τα ιδιοδιανύσματα από τον πίνακα συνδιακύμανσης, το επόμενο βήμα είναι η τοποθέτησή τους σε σειρά, σύμφωνα με τις αντίστοιχες τιμές των ιδιοτιμών από το μεγαλύτερο προς το μικρότερο. Αυτό μας δίνει όλα τα συστατικά αυτά στοιχεία σε σειρά σπουδαιότητας. Στο σημείο αυτό μπορούμε να αγνοήσουμε τα λιγότερο σημαντικά στοιχεία αν και μπορεί να χάνουμε κάποιες πληροφορίες. Όταν όμως οι ιδιοτιμές είναι μικρού μεγέθους, τότε δεν χάνουμε τόσα πολλά σε πληροφορίες-δεδομένα. Εάν λοιπόν όντως αγνοήσουμε κάποια δεδομένα τότε τα τελικά στοιχεία θα έχουν λιγότερες διαστάσεις από τα αρχικά δεδομένα.

Για να είμαστε ακριβείς, εάν αρχικά είχαμε ***n*** διαστάσεις στα δεδομένα μας και υπολογίσουμε ***n*** ιδιοτιμές και ***n*** ιδιοδιανύσματα και στη συνέχεια επιλέξουμε μόνο ***p*** ιδιοδιανύσματα από τα αρχικά (***p <n)***, τότε τα τελικά μας δεδομένα θα έχουν μόνο ***p*** διαστάσεις. Στη συνέχεια πρέπει να διαμορφωθεί ένα χαρακτηριστικό διάνυσμα (feature vector), το οποίο στην ουσία είναι ένα όνομα για έναν πίνακα διανυσμάτων. Η κατασκευή του γίνεται τοποθετώντας τα ιδιοδιανύσματα που τελικά αποφασίζουμε να κρατήσουμε από τη λίστα των ιδιοδιανυσμάτων, και τη διαμόρφωση ενός πίνακα με αυτά τα ιδιοδιανύσματα τοποθετημένα σε στήλες.

FeatureVector = (eig1 eig2 eig3 ⋯eign)

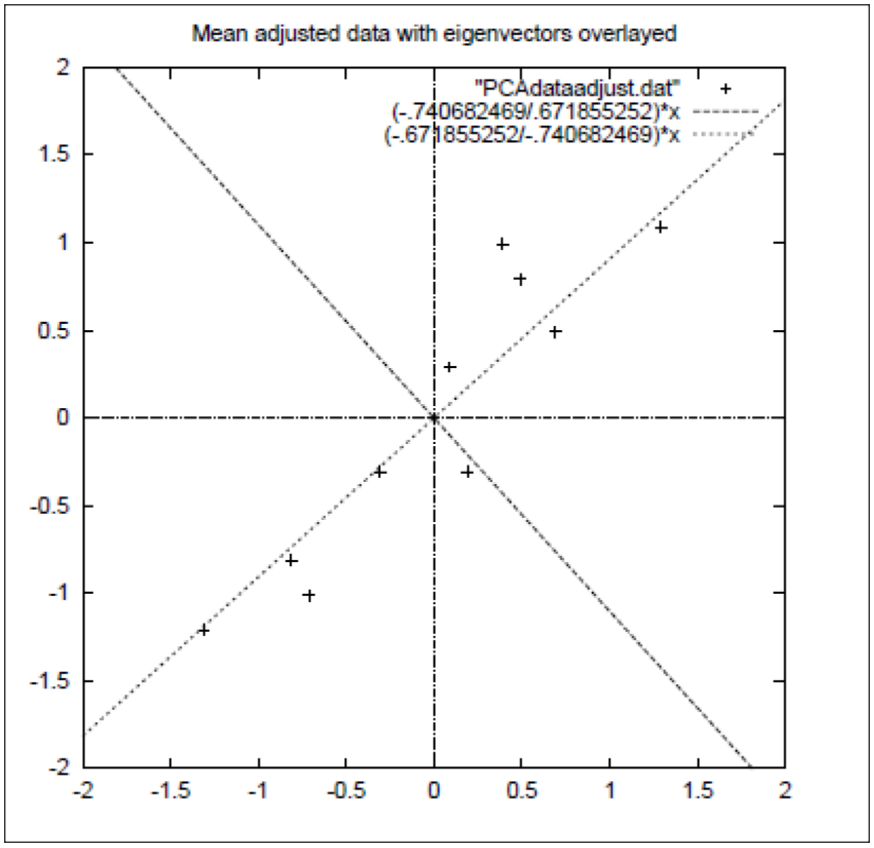
Δεδομένου του συνόλου των στοιχείων μας καθώς επίσης και του ότι έχουμε 2 ιδιοδιανύσματα, έχουμε δύο επιλογές: είτε να δημιουργήσουμε ένα χαρακτηριστικό διάνυσμα (feature vector) με την ταυτόχρονη χρησιμοποίηση και των δύο ιδιοδιανυσμάτων:

$$\begin{pmatrix} -0.677873399 & -0.735178656 \\ -0.735178656 & 0.677873399 \end{pmatrix}$$

είτε να επιλέξουμε να παραβλέψουμε το μικρότερο, το λιγότερο σημαντικό στοιχείο προκειμένου τελικά να έχουμε μόνο μία στήλη .

(-0.677873399 -0.735178656)

Το αποτέλεσμα των δύο παραπάνω επιλογών φαίνεται παρακάτω.




Σχήμα 8:Τα δεδομένα μετά την αφαίρεση του μέσου όρου με τα ιδιοδιανύσματα του πίνακα συνδιακύμανσης στην κορυφή

### 6. Συλλογή των νέων δεδομένων

Αυτό είναι το τελικό στάδιο της PCA, κατά το οποίο συλλέγουμε τα τελικά δεδομένα. Όταν θα έχουμε επιλέξει τα στοιχεία (eigenvectors), τα οποία επιθυμούμε να διατηρήσουμε στα τελικά δεδομένα μας και έχουμε σχηματίσει το χαρακτηριστικό διάνυσμα, μπορούμε ύστερα να αντιμεταθέσουμε το διάνυσμα, , *δηλαδή να πάρουμε το ανάστροφό του και στη συνέχεια να το πολλαπλασιάσουμε από την αριστερή του πλευρά του πρότυπου συνόλου δεδομένων* αντιμετατιθέμενο. (δηλ. Με το ανάστροφό του)

**FinalData** = **RowFeatureVector** X **RowDataAdjust**

Το RowFeatureVector, αντιστοιχεί σε έναν πίνακα με ιδιοδιανύσματα σε στήλες ανάστροφο, ούτως ώστε τα εν λόγω ιδιοδιανύσματα να βρίσκονται στις γραμμές του πίνακα, με τα περισσότερο σημαντικά ιδιοδιανύσματα να βρίσκονται στη κορυφή της. Ενώ ο όρος RowDataAdjust αντιστοιχεί στα μέσα δεδομένα αντιμετατιθέμενα, δηλαδή τα στοιχεία των δεδομένων βρίσκονται σε κάθε στήλη, με κάθε γραμμή να έχει ξεχωριστή διάσταση.

Τέλος ο όρος FinalData ισούται με το τελικό σύνολο δεδομένων, με στήλες τα στοιχεία των δεδομένων και γραμμές τις διαστάσεις τους.

Ο συγκεκριμένος πίνακας λοιπόν θα μας δώσει τα πρότυπα δεδομένα με αποκλειστική αντιστοιχία στα διανύσματα που τελικά επιλέξαμε. Τα πρότυπα δεδομένα μας έχουν δύο

άξονες συγκεκριμένα τους ***x, y,*** ούτως ώστε τα δεδομένα μας να βρίσκονται σε αντιστοιχία με τα σημεία των αξόνων. Είναι πιθανή οποιαδήποτε αντιστοίχιση με όποιους δύο άξονες προτιμούμε, στην περίπτωση όμως που οι επιλεγμένοι άξονες είναι μεταξύ τους κάθετοι τότε η έκφραση είναι περισσότερο αποδοτική. Για αυτό καθορίζεται ως τόσο σημαντική η καθετότητα των ιδιοδιανυσμάτων μεταξύ τους. Έχουμε μετατρέψει τα δεδομένα μας από σημεία των ***x, y*** αξόνων σε στοιχεία δισδιάστατων ιδιοδιανυσμάτων. Όπως έχουμε ήδη αναφέρει στη περίπτωση που έχουμε μειώσει τα δεδομένα μας –καθώς έχουμε επιλέξει μέρος αυτών- και επομένως και τις διαστάσεις τους, έχουμε αφήσει κάποια ιδιοδιανύσματα έξω. Τα νέα μας δεδομένα τότε είναι σε αντιστοιχία με τα διανύσματα τα οποία έχουμε αποφασίσει να κρατήσουμε. Προκειμένου να γίνει εμφανής η παραπάνω μείωση στα δεδομένα μας, πραγματοποιείται ακολούθως η τελική μεταμόρφωση με κάθε ένα από τα πιθανά χαρακτηριστικά διανύσματα. Έτσι λάβαμε το αντιμεταθετιμένο αποτέλεσμα σε κάθε πιθανή περίπτωση, προκειμένου να φέρουμε τα δεδομένα στην αρχική διαμορφωμένη μορφή του πίνακα. Επιπρόσθετα, έχει πραγματοποιηθεί και μία γραφική απεικόνιση προκειμένου να δειχθεί πώς σχετίζονται τα τελικά σημεία με τα στοιχεία- ιδιοδιανύσματα. Στη περίπτωση που επιλεχθούν και τα δύο ιδιοδιανύσματα για την επικείμενη μεταμόρφωση, τότε λαμβάνουμε τα δεδομένα και το γράφημα της Εικόνας 7. Το γράφημα αυτό απεικονίζει στην ουσία το πρότυπο των δεδομένων, το οποίο έχει περιστραφεί έτσι ώστε τα ιδιοδιανύσματα να αποτελούν τους άξονες. Το γεγονός αυτό είναι κατανοητό καθώς δεν έχουμε παραλείψει κανένα δεδομένο σε αυτήν την αποδόμηση-διάσπαση.

Η άλλη μεταμόρφωση η οποία δύναται να λάβει χώρα είναι αυτή η οποία έχει προκύψει κατόπιν επιλογής ιδιοδιανυσμάτων με τη μεγαλύτερη ιδιοτιμή. Ο πίνακας των δεδομένων ο οποίος έχει υπολογισθεί ύστερα από αυτή την επιλογή είναι ο πίνακας Transformed Data (Single eigenvector) και έχει μόνο μια διάσταση. Εάν αντιπαραβάλλουμε το σύνολο των δεδομένων με το σύνολο το οποίο προέκυψε από τη χρήση και των δύο ιδιοδιανυσμάτων, θα διαπιστώσουμε πως το επικείμενο σύνολο των δεδομένων είναι ακριβώς ταυτόσημο με τη πρώτη στήλη του άλλου. Επομένως εάν επρόκειτο να σχεδιάζαμε τα συγκεκριμένα δεδομένα, θα ήταν μονοδιάστατα και σημεία μιας γραμμής σε ακριβώς ***x*** θέσεις των σημείων απεικόνισης της Εικόνας 7. Συμπερασματικά λοιπόν έχουμε εξαλείψει έναν ολόκληρο άξονα, ο οποίος αντιπροσώπευε το δεύτερο ιδιοδιάνυσμα. Στην ουσία έχουμε μετασχηματίσει τα προς μελέτη δεδομένα μας ούτως ώστε να υπάρχει αντιστοιχία μεταξύ των προτύπων, όπου τα πρότυπα είναι οι γραμμές εκείνες οι οποίες αποδίδουν στο μέγιστο δυνατό τη σχέση αυτή μεταξύ των δεδομένων. Το γεγονός αυτό είναι χρήσιμο καθώς με τη διαδικασία αυτή έχουμε ταξινομήσει τα δεδομένα μας σε έναν χώρο ο οποίος καθορίζεται από το συνδυασμό

συνεισφορών κάθε γραμμής. Στην αρχή είχαμε απλώς τους άξονες ***x, y***. Το γεγονός αυτό ως αποτέλεσμα κρίνεται καλό, όμως οι ***x, y*** τιμές κάθε δεδομένου του χώρου που προκύπτει δεν μας δίνουν ικανοποιητικές πληροφορίες σχετικά με το πώς ο εν λόγω χώρος συνδέεται με το υπόλοιπο σύνολο των δεδομένων. Τώρα οι τιμές των δεδομένων του χώρου μας καθορίζουν επακριβώς προς τα πού τείνουν (δηλαδή κάτω ή πάνω) οι προσκοπτόμενες γραμμές του χώρου. Στην περίπτωση της μετατροπής με τη χρήση και των δύο ιδιοδιανυσμάτων, έχουμε στην ουσία τροποποιήσει τα δεδομένα, ούτως ώστε να πραγματοποιείται σε αντιστοιχία με τα ιδιοδιανύσματα σε αντίθεση με τους συνήθεις άξονες**.** Αντίθετα, το μονοδιάστατο ιδιοδιάνυσμα της αποσύνθεσης έχει αποσιωπηθεί από τη συνεισφορά εξαιτίας της ύπαρξης του μικρότερου ιδιοδιανύσματος, με αποτέλεσμα να μείνουμε με τα δεδομένα εκείνα τα οποία είναι σε αντιστοιχία με το προηγούμενο-μεγαλύτερο ιδιοδιάνυσμα.

**7. Επαναφορά των αρχικών δεδομένων**

Μετά από την συμπίεση των δεδομένων μας κατά PCA, θέλουμε να τα αποσυμπιέσουμε ώστε να πάρουμε τα αρχικά μας δεδομένα και να υπολογίσουμε το σφάλμα συμπίεσης.

X Y

-0.827970186 -0.175115307

1.77758033 0.142857227

-0.992197494 0.384374989

-0.274210416 0.130417207

-1.67580142 -0.209498461 = Transformed Data

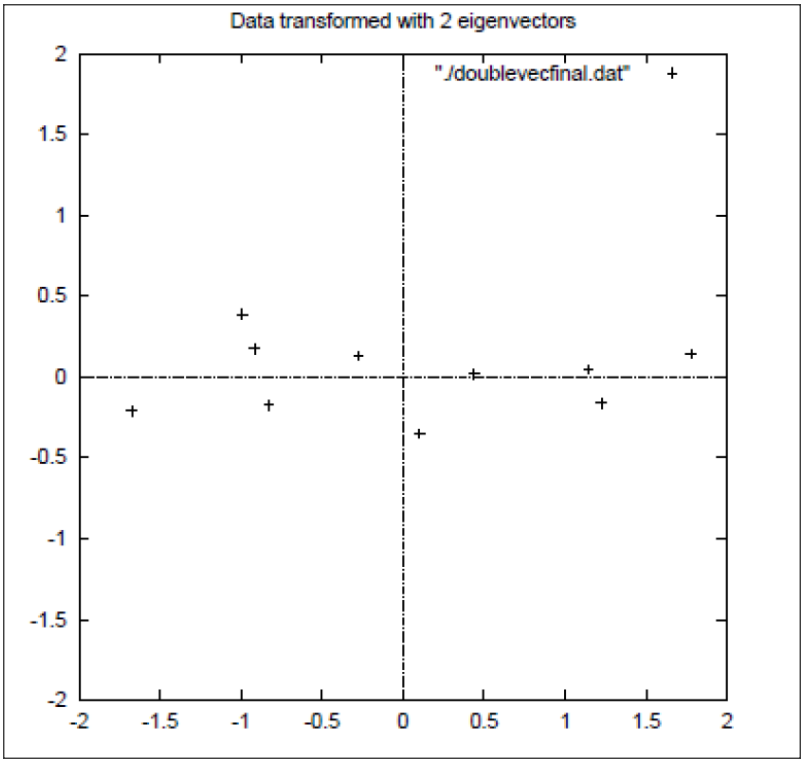
-0.912949103 0.175282444

0.0991094375 -0.349824698

1.14457216 0.0464172582

0.438046137 0.0177646297

1.22382056 -0.162675287




Σχήμα 9: Τα νέα δεδομένα μετά τη συμπίεση με PCA, με χρησιμοποίηση όλων των ιδιοδιανυσμάτων

Transformed Data (Single eigenvector)

-0.827970186

1.77758033

-0.992197494

-0.274210416

-1.67580142

-0.912949103

0.0991094375

1.14457216

0.438046137

1.22382056

Είναι προφανές πως μόνο εάν κάνουμε χρήση όλου του συνόλου των ιδιοδιανυσμάτων στην εν λόγω μετατροπή μας θα λάβουμε επακριβώς το σύνολο των αρχικών δεδομένων. Στη περίπτωση όμως που έχουμε μειώσει του αριθμού των ιδιοδιανυσμάτων, τότε τα δεδομένα που θα επαναφέρουμε θα έχουν προφανώς κάποια απώλεια πληροφοριών.

Ας υπενθυμίσουμε τον τύπο της τελικής μετατροπής :

**FinalData** = **RowFeatureVector** X **RowDataAdjust**

ο οποίος προκειμένου να λάβουμε την αρχική μορφή των δεδομένων, αντιστρέφεται και παίρνει τη μορφή:

**RowDataAdjust** = **RowFeatureVector-1** X **FinalData**

όπου RowFeatureVector -1 είναι ο αντεστραμμένος όρος RowFeatureVector.

Ωστόσο όταν λαμβάνουμε όλα τα ιδιοδιανύσματα στο χαρακτηριστικό μας διάνυσμα (feature vector), προκύπτει πως το αντίστροφο του χαρακτηριστικού διανύσματος είναι στην ουσία ίσο με το ανάστροφο χαρακτηριστικό διάνυσμα. Αυτό ισχύει γιατί τα στοιχεία του πίνακα είναι όλα μοναδιαία ιδιοδιανύσματα του συνόλου των δεδομένων μας. Αυτό καθιστά τη διαδικασία της επιστροφής του αρχικού συνόλου των δεδομένων ευκολότερη, καθώς η εξίσωση παίρνει τελικά τη μορφή:

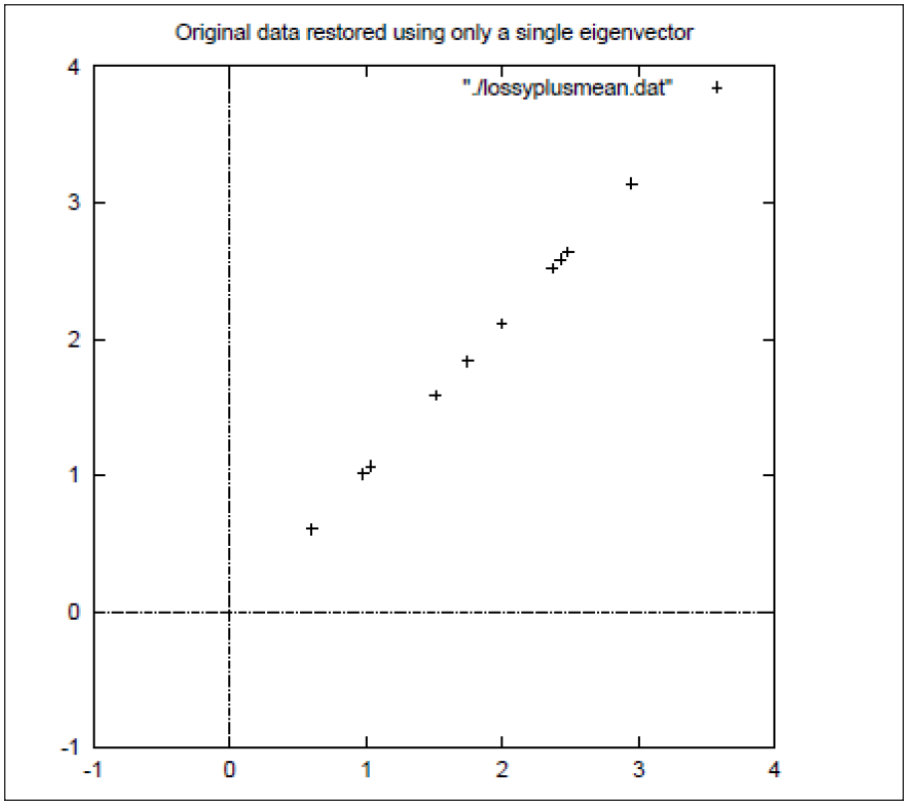
**RowDataAdjust** = **RowFeatureVectorT** X **FinalData**

Ωστόσο για να πάρουμε τα αρχικά δεδομένα, χρειαζόμαστε την πρόσθεση του μέσου όρου των πρωτογενών δεδομένων (καθώς όπως θυμόμαστε τον είχαμε αφαιρέσει προηγουμένως). Έτσι ο πλήρης τύπος δίνεται από την έκφραση που ακολουθεί:

**RowOriginalData** = (**RowFeatureVectorT** X **FinalData**) + **OriginalMean**

Ο τύπος αυτός ισχύει και για τις περιπτώσεις εκείνες που δεν έγινε χρήση όλων των ιδιοδιανυσμάτων στον επικείμενο υπολογισμό του χαρακτηριστικού διανύσματος (feature value).

Όσον αφορά στην επανάκτηση των δεδομένων με τη χρήση της πλήρους μορφής του χαρακτηριστικού διανύσματος, το αποτέλεσμα είναι ταυτόσημο με τα δεδομένα τα οποία παρατέθηκαν αρχικά. Ωστόσο, κάνοντας τον υπολογισμό με το μειωμένο χαρακτηριστικό διάνυσμα, είναι πιο εμφανής η απώλεια των πληροφοριών. Στο σχήμα 10 φαίνεται η εν λόγω απεικόνιση.




Σχήμα 10: Η επαναφορά των δεδομένων χρησιμοποιώντας ένα μόνο ιδιοδιάνυσμα

# Κεφάλαιο 3ο: Ανάλυση αλγόριθμου

Στο παρόν κεφάλαιο θα εισαγάγουμε τις τεχνικές Μηχανικής Μάθησης (Machine Learning ) που χρησιμοποιήθηκαν στην παρούσα εργασία και στη συνέχεια, θα γίνει ανάλυση και εξήγηση του αλγόριθμου που χρησιμοποιήθηκε για ανίχνευση απάτης πιστωτικών καρτών.

## 3.1 Ανίχνευση παρεκτροπών ( Anomaly Detection)

Στην εξόρυξη δεδομένων, ανίχνευση ανωμαλιών και επίσης στην ανίχνευση εξωστρέφειας ( outlier detection) αξιοποιείται η ταυτοποίηση σπάνιων στοιχείων, συμβάντων ή παρατηρήσεων που δημιουργούν υποψίες και διαφέρουν σημαντικά από την πλειοψηφία των δεδομένων. Συνήθως τα ανώμαλα στοιχεία θα μεταφραστούν σε κάποιο είδος προβλήματος, όπως ***τραπεζική απάτη*** (όπως και στην εργασία μας), δομικό ελάττωμα, ιατρικά προβλήματα ή σφάλματα σε ένα κείμενο. Οι διαφόρου είδους ανωμαλίες αναφέρονται επίσης ως υπερβάσεις (Outliers), καινοτομίες, θόρυβος, αποκλίσεις και εξαιρέσεις.

Ειδικότερα, στο πλαίσιο της κατάχρησης και της ανίχνευσης εισβολής σε δίκτυο, τα ενδιαφέροντα ανιχνευόμενα στοιχεία συχνά δεν είναι *σπάνια*, αλλά απροσδόκητες *εκρήξεις* στη δραστηριότητα. Αυτό το μοτίβο δεν συμμορφώνεται με τον κοινό στατιστικό ορισμό ενός εξωλέμβιου ως σπάνιο αντικείμενο και πολλές μέθοδοι ανίχνευσης (ιδίως μη επιτηρούμενες), θα αποτύχουν στα δεδομένα αυτά, εκτός αν έχουν συσσωρευτεί κατάλληλα. Αντ' αυτού, ένας αλγόριθμος ανάλυσης συμπλέγματος μπορεί να είναι σε θέση να ανιχνεύσει τα μικροσυστήματα που σχηματίζονται από αυτά τα πρότυπα. Υπάρχουν τρεις γενικές κατηγορίες τεχνικών ανίχνευσης ανωμαλιών που περιγράφονται παρακάτω:

- ✓ **Μη επιβλεπόμενη ανίχνευση παρεκτροπών** : είναι οι τεχνικές ανίχνευσης παρεκτροπών που δεν επιτηρούνται και ανιχνεύουν ανωμαλίες σε ένα μη επισημασμένο σύνολο δεδομένων δοκιμών και υπό την προϋπόθεση ότι η πλειοψηφία των περιπτώσεων στο σύνολο δεδομένων είναι φυσιολογικές, αναζητούν περιπτώσεις που φαίνονται να ταιριάζουν λιγότερο στο υπόλοιπο σύνολο δεδομένων. Σε αυτή την περίπτωση δεν απαιτούνται δεδομένα εκπαίδευσης, και έτσι είναι οι πιο ευρέως εφαρμόσιμες τεχνικές.

- ✓ **Επιβλεπόμενη ανίχνευση παρεκτροπών:** είναι εποπτευόμενες τεχνικές της ανίχνευσης παρεκτροπών και απαιτούν ένα σύνολο δεδομένων που έχει χαρακτηριστεί ως "κανονικό" και "μη φυσιολογικό" στις τιμές και περιλαμβάνει την κατάρτιση ενός ταξινομητή (η βασική διαφορά σε πολλά άλλα στατιστικά προβλήματα ταξινόμησης είναι ο εγγενής μη ισορροπημένος χαρακτήρας της ανίχνευσης των εξωστρεφών).
- ✓ **Ημι - επιβλεπόμενη ανίχνευση παρεκτροπών:** είναι οι τεχνικές ανίχνευσης παρεκτροπών με ημι - εποπτευόμενο τρόπο. Κατασκευάζουν ένα μοντέλο που αντιπροσωπεύει κανονική συμπεριφορά από ένα *φυσιολογικό* σύνολο δεδομένων και στη συνέχεια εφαρμόζεται για τον προσδιορισμό των ανωμαλιών στα δοκιμαστικά δεδομένα. Το θετικό αυτής της κατηγορίας είναι ότι εφαρμόζεται σε μεγαλύτερη κλίμακα από την **επιβλεπόμενη ανίχνευση ανωμαλιών**, για το λόγο ότι δεν απαιτείται επισήμανση για την τάξη ακραίων τιμών.

### 3.1.1 Εφαρμογές ανίχνευσης παρεκτροπών

Η ανίχνευση παρεκτροπών μπορεί να εφαρμοστεί σε μια ποικιλία τομέων, όπως η ανίχνευση εισβολέων , ***ανίχνευση της απάτης*** , ανίχνευση σφαλμάτων, την παρακολούθηση της υγείας συστήματος, ανίχνευσης συμβάντων σε δίκτυα αισθητήρων και την ανίχνευση διαταραχών του οικοσυστήματος. Συχνά χρησιμοποιείται στην προ-επεξεργασία για την απομάκρυνση ανώμαλων δεδομένων από το σύνολο δεδομένων. Στην εποπτευόμενη μάθηση , η αφαίρεση των ανώμαλων δεδομένων από το σύνολο δεδομένων συχνά οδηγεί σε στατιστικά με σημαντική αύξηση της ακρίβειας.

## 3.2 Τοπικός Συντελεστής Απόκλισης ( Local Outlier Factor)

Στην ανίχνευση παρεκτροπών , ***ο τοπικός συντελεστής απόκλισης ( Local Outlier Factor )*** είναι ένας αλγόριθμος που προτάθηκε από τους Markus M. Breunig, Hans-Peter Kriegel , Raymond T. Ng και Jörg Sander το 2000 για την ανεύρεση ανώμαλων σημείων δεδομένων με μέτρηση της τοπικής απόκλισης ενός δεδομένου σημείου δεδομένων σε σχέση με τους γείτονές του. Το LOF μοιράζεται κάποιες έννοιες με τα DBSCAN ( ένας από τους πιο συνηθισμένους αλγόριθμους ομαδοποίησης) και OPTICS (είναι ένας αλγόριθμος για την εύρεση συστοιχιών σε χωρικά δεδομένα) , όπως οι έννοιες της "βασικής απόστασης" και της "απόστασης απόδοσης", οι οποίες χρησιμοποιούνται για την εκτίμηση της τοπικής

πυκνότητας. ***Η βασική ιδέα του βασίζεται σε μια έννοια της τοπικής πυκνότητας****, στην οποία δίνεται η τοποθεσία πλησιέστεροι γείτονες, των οποίων η απόσταση χρησιμοποιείται για την εκτίμηση της πυκνότητας*. Συγκρίνοντας την τοπική πυκνότητα ενός αντικειμένου με τις τοπικές πυκνότητες των γειτόνων του, μπορεί κανείς να εντοπίσει περιοχές παρόμοιας πυκνότητας, όπως και σημεία που έχουν ουσιαστικά χαμηλότερη πυκνότητα από τους γείτονές τους. Αυτές θεωρούνται ως υπερβολικές τιμές.

Η τοπική πυκνότητα υπολογίζεται από την τυπική απόσταση στην οποία ένα σημείο μπορεί να «προσεγγιστεί» από τους γείτονές του. Ο ορισμός της "απόστασης απόδοσης" που χρησιμοποιείται στο LOF είναι ένα πρόσθετο μέτρο για την παραγωγή πιο σταθερών αποτελεσμάτων εντός των ομάδων.

### 3.2.1 Τυπική απόκλιση

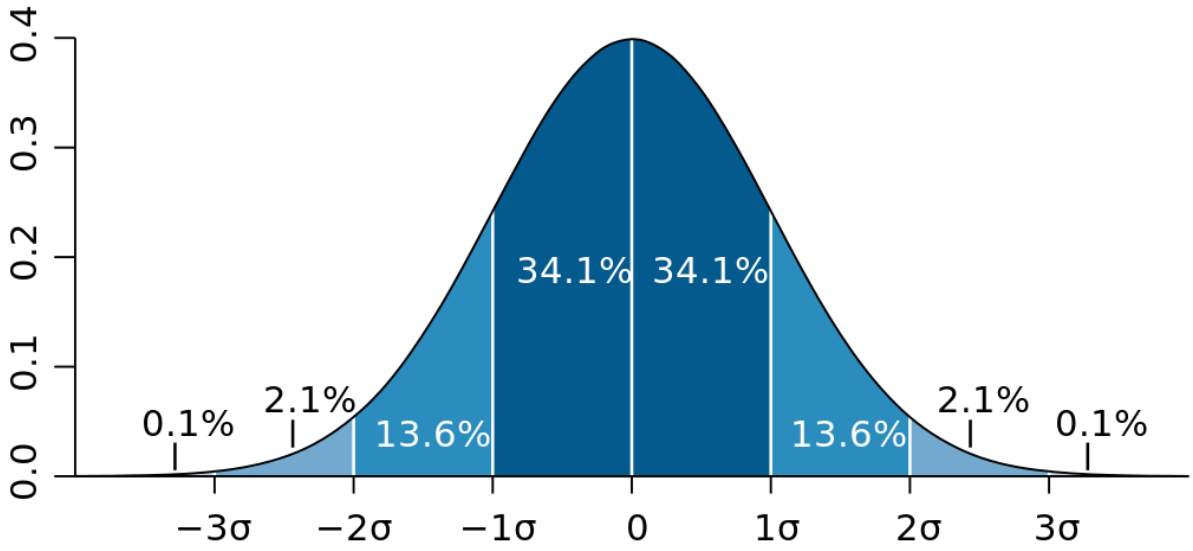
Στη στατιστική, η **τυπική απόκλιση** είναι ένα μέτρο που χρησιμοποιείται για να υπολογιστεί το ποσό της μεταβολής ή της διασποράς ενός συνόλου τιμών δεδομένων. Μια χαμηλή τυπική απόκλιση υποδηλώνει ότι τα σημεία των δεδομένων τείνουν να είναι κοντά στο μέσο όρο (που ονομάζεται επίσης η αναμενόμενη τιμή) του συνόλου, ενώ μία υψηλή τυπική απόκλιση υποδεικνύει ότι τα στοιχεία απλώνονται πάνω από ένα ευρύτερο φάσμα των τιμών.

Η τυπική απόκλιση μιας τυχαίας μεταβλητής, ενός στατιστικού πληθυσμού, ενός συνόλου δεδομένων, ή της κατανομής πιθανότητας είναι η τετραγωνική ρίζα της διακύμανσης της(ή αλλιώς διασποράς). Είναι αλγεβρικά απλούστερη, αν και στην πράξη λιγότερο ισχυρή από τη μέση απόλυτη απόκλιση. Μία χρήσιμη ιδιότητα της τυπικής απόκλισης είναι ότι, σε αντίθεση με την διακύμανση, εκφράζεται στις ίδιες μονάδες με τα δεδομένα. Υπάρχουν επίσης άλλα μέτρα απόκλισης από τον κανόνα, συμπεριλαμβανομένων της μέσης απόλυτης απόκλισης, η οποία παρέχει διαφορετικές μαθηματικές ιδιότητες από την τυπική απόκλιση.

Εκτός από την έκφραση της μεταβλητότητας του πληθυσμού, η τυπική απόκλιση συνήθως χρησιμοποιείται για τη μέτρηση της εμπιστοσύνης στα στατιστικά συμπεράσματα. Για παράδειγμα, το περιθώριο λάθους σε δεδομένα δημοσκοπήσεων προσδιορίζεται με τον υπολογισμό της αναμενόμενης τυπικής απόκλισης στα αποτελέσματα, αν η ίδια δημοσκόπηση έπρεπε να διεξαχθεί πολλές φορές. Αυτή η εξαγωγή της τυπικής απόκλισης συχνά αποκαλείται «τυπικό σφάλμα» της εκτίμησης ή «τυπικό σφάλμα της μέσης τιμής» όταν αναφέρεται σε μια μέση τιμή. Υπολογίζεται ως η τυπική απόκλιση όλων των μέσων τιμών που θα υπολογίζεται από τον εν λόγω πληθυσμό, εάν καταρτιστεί ένας άπειρος αριθμός

δειγμάτων και μια μέση τιμή για κάθε δείγμα που υπολογίζεται. Είναι πολύ σημαντικό να σημειωθεί ότι η τυπική απόκλιση ενός πληθυσμού και το τυπικό σφάλμα μιας στατιστικής που προέρχεται από τον εν λόγω πληθυσμό (όπως τη μέση τιμή) είναι αρκετά διαφορετικές αλλά σχετικές (σε σχέση με το αντίστροφο της τετραγωνικής ρίζας του αριθμού των παρατηρήσεων) . Το αναφερόμενο περιθώριο λάθους σε μια δημοσκόπηση υπολογίζεται από το τυπικό σφάλμα της μέσης τιμής (ή εναλλακτικά από το γινόμενο της τυπικής απόκλισης του πληθυσμού και του αντίστροφου της τετραγωνικής ρίζας του μεγέθους του δείγματος, το οποίο είναι το ίδιο πράγμα) και είναι τυπικά περίπου διπλάσια της τυπικής απόκλισης-του μισού πλάτους ενός διαστήματος εμπιστοσύνης 95 τοις εκατό. Στην επιστήμη, οι ερευνητές συνήθως αναφέρουν την τυπική απόκλιση των πειραματικών δεδομένων, και μόνο αποτελέσματα που πέφτουν πολύ μακρύτερα από δύο τυπικές αποκλίσεις μακριά από ό, τι θα αναμενόταν θεωρούνται στατιστικά σημαντικές- κανονικό τυχαίο σφάλμα ή διακύμανση των μετρήσεων με αυτό τον τρόπο διακρίνονται από τυχαίες μεταβολές. Η τυπική απόκλιση είναι επίσης σημαντική στα οικονομικά, όπου η τυπική απόκλιση στο ποσοστό απόδοσης της επένδυσης είναι ένα μέτρο της μεταβλητότητας της επένδυσης.

Όταν μόνο ένα δείγμα των δεδομένων από έναν πληθυσμό είναι διαθέσιμο, ο όρος τυπική απόκλιση του δείγματος ή δείγμα τυπικής απόκλισης μπορεί να αναφέρεται είτε στην ανωτέρω ποσότητα, όπως εφαρμόζεται στα εν λόγω δεδομένα είτε σε μία τροποποιημένη ποσότητα που είναι μια καλύτερη εκτίμηση του πληθυσμού της τυπικής απόκλισης (η τυπική απόκλιση του συνόλου του πληθυσμού).




Διάγραμμα 1: Τυπική απόκλιση

## 3.3 Isolation Forest ( Απομόνωσης Δασών)

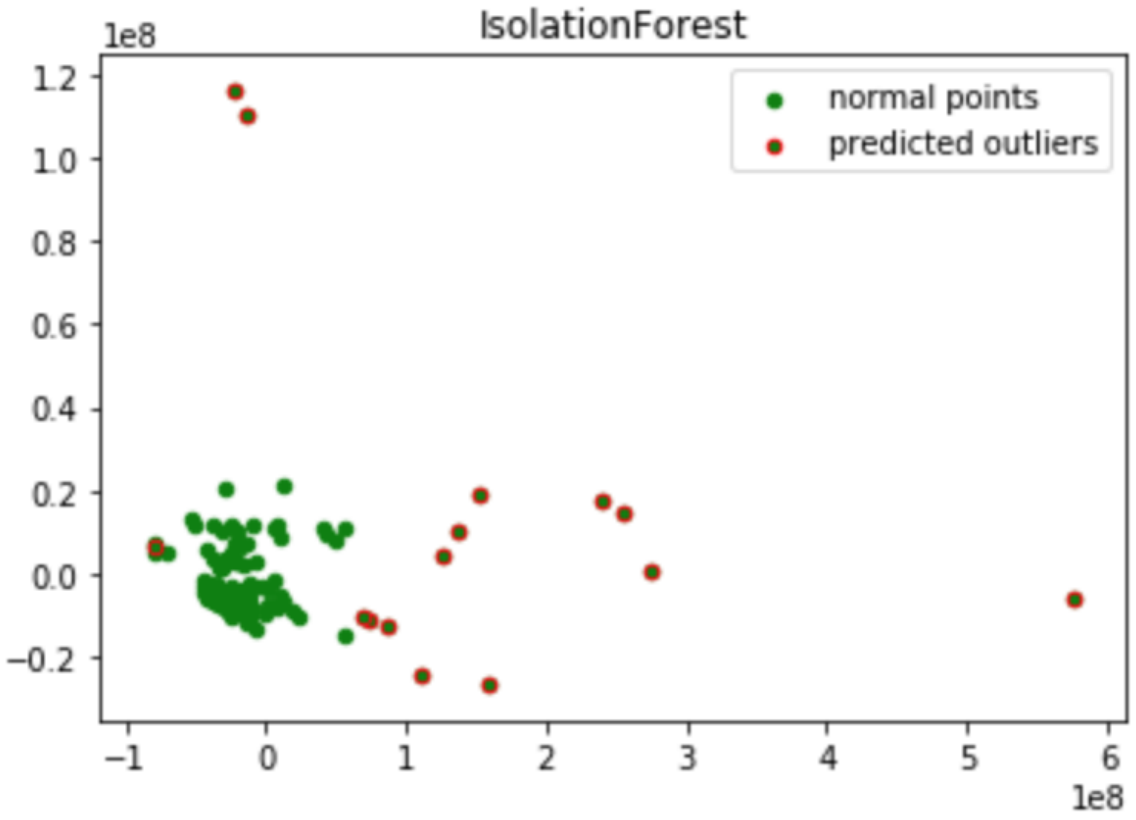
Μία από τις νεότερες τεχνικές ανίχνευσης παρεκτροπών ονομάζεται **Isolation Forest (Απομόνωσης Δασών)**. Ο αλγόριθμος βασίζεται στο γεγονός ότι οι ανωμαλίες είναι σημεία δεδομένων που είναι λίγα και διαφορετικά. Ως αποτέλεσμα αυτών των ιδιοτήτων, οι ανωμαλίες είναι ευαίσθητες σε ένα μηχανισμό που ονομάζεται απομόνωση. Αυτή η μέθοδος είναι εξαιρετικά χρήσιμη και βασικά διαφέρει από όλες τις υπάρχουσες μεθόδους. Εισάγει τη χρήση της απομόνωσης ως αποτελεσματικότερο μέσο για την ανίχνευση ανωμαλιών από τα χρησιμοποιούμενα βασικά μέτρα απόστασης και πυκνότητας. Επιπλέον, αυτή η μέθοδος είναι ένας αλγόριθμος με χαμηλή γραμμική πολυπλοκότητα χρόνου και μικρή απαίτηση μνήμης. Χτίζει ένα μοντέλο καλής απόδοσης με μικρό αριθμό δέντρων χρησιμοποιώντας μικρά υποδείγματα σταθερού μεγέθους, ανεξάρτητα από το μέγεθος ενός συνόλου δεδομένων. Οι τυπικές μέθοδοι μάθησης μηχανών τείνουν να λειτουργούν καλύτερα όταν τα μοτίβα που προσπαθούν να μάθουν να είναι ισορροπημένα, πράγμα που σημαίνει ότι στο σύνολο δεδομένων υπάρχει το ίδιο ποσό καλών και κακών συμπεριφορών.

### 3.3.1 Λειτουργία του Isolation Forest

Ο αλγόριθμος "απομόνωσης δασών " απομονώνει τις παρατηρήσεις επιλέγοντας τυχαία ένα χαρακτηριστικό και επιλέγοντας τυχαία μια διαχωρισμένη τιμή μεταξύ της μέγιστης και της ελάχιστης τιμής του επιλεγμένου χαρακτηριστικού. Το λογικό επιχείρημα πηγαίνει: η απομόνωση των παρατηρήσεων ανωμαλίας είναι ευκολότερη επειδή μόνο λίγες συνθήκες είναι απαραίτητες για να διαχωριστούν αυτές οι περιπτώσεις από τις κανονικές παρατηρήσεις.

Από την άλλη πλευρά, η απομόνωση κανονικών παρατηρήσεων απαιτεί περισσότερες συνθήκες. Επομένως, μια βαθμολογία ανωμαλίας μπορεί να υπολογιστεί ως ο αριθμός των συνθηκών που απαιτούνται για να διαχωριστεί μια δεδομένη παρατήρηση.

Ο τρόπος με τον οποίο ο αλγόριθμος κατασκευάζει τον διαχωρισμό είναι πρώτα δημιουργώντας δέντρα απομόνωσης ή δέντρα τυχαίων αποφάσεων. Στη συνέχεια, η βαθμολογία υπολογίζεται ως το μήκος διαδρομής για να απομονωθεί η παρατήρηση. Το παρακάτω παράδειγμα δείχνει ότι είναι εύκολο να διαχωρίσετε μια παρατήρηση ανωμαλίας:




Σχήμα 11 : Isolation forest

Προκειμένου να αποφευχθούν ζητήματα λόγω της τυχαιότητας του αλγόριθμου δέντρου, η διαδικασία επαναλαμβάνεται αρκετές φορές και το μέσο μήκος διαδρομής υπολογίζεται και κανονικοποιείται.

### 3.4 Εισαγωγή βιβλιοθηκών

Έπειτα θα κάνουμε μια ανάλυση στις βιβλιοθήκες που χρησιμοποιήθηκαν για την συγκεκριμένη εργασία.

- **NumPy:** Το πιο θεμελιώδες πακέτο, γύρω από το οποίο κατασκευάζεται η στοίβα επιστημονικών υπολογισμών, είναι βιβλιοθήκη NumPy η οποία σημαίνει Αριθμητική στην Python. Παρέχει μια πληθώρα χρήσιμων χαρακτηριστικών για λειτουργίες, σε συστοιχίες και πίνακες στη Python. Η βιβλιοθήκη παρέχει διανύσματα για μαθηματικές λειτουργίες, βελτιώνει την απόδοση και επομένως επιταχύνει την εκτέλεση.

- **Spicy:** Το SciPy είναι μια βιβλιοθήκη λογισμικού για την επιστήμη υπολογιστών (Data Science). Το SciPy περιέχει ενότητες για γραμμική άλγεβρα, βελτιστοποίηση, ολοκλήρωση και στατιστικά στοιχεία . Η κύρια λειτουργικότητα της βιβλιοθήκης SciPy βασίζεται στο NumPy και οι συστοιχίες της κάνουν ουσιαστική χρήση του NumPy. Παρέχει αποτελεσματικές αριθμητικές ρουτίνες όπως αριθμητική ολοκλήρωση, βελτιστοποίηση και πολλές άλλες μέσω των συγκεκριμένων υπομονάδων. Οι λειτουργίες σε όλες τις υπομονάδες του SciPy είναι καλά τεκμηριωμένες .
- Το Pandas είναι ένα πακέτο Python σχεδιασμένο για να δουλεύει με δεδομένα "ετικετοποιημένα" και "σχεσιακά" απλά και διαισθητικά. Το Pandas είναι ένα εργαλείο για την καταπολέμηση των δεδομένων. Σχεδιάστηκε για γρήγορο και εύκολο χειρισμό δεδομένων, συνάθροιση και οπτικοποίηση.

  Υπάρχουν δύο κύριες δομές δεδομένων στη βιβλιοθήκη:

  - Series : Μονοδιάστατη

| Series | |
|---|---|
| A | X0 |
| B | X1 |
| C | X2 |
| D | X3 |

Σχήμα 12:Μονοδιάστατος πίνακας

  - Data Frames: Δυσδιάστατη

| DataFrame | | | | |
|---|---|---|---|---|
| | A | B | C | D |
| 0 | A0 | B0 | C0 | D0 |
| 1 | A1 | B1 | C1 | D1 |
| 2 | A2 | B2 | C2 | D2 |
| 3 | A3 | B3 | C3 | D3 |

Σχήμα 13:Δυσδιάστατος πίνακας

- Matplotlib: είναι προσαρμοσμένη για την απλή και ισχυρή απεικόνιση Διαγραμμάτων με ευκολία. Χρειάζεται βοήθεια από τους NumPy, SciPy και Pandas.

- **Scikits**: Είναι σχεδιασμένα για συγκεκριμένες λειτουργίες όπως η επεξεργασία εικόνας και η διευκόλυνση της εκμάθησης μηχανών. Ένα από τα πιο σημαντικά από αυτά τα πακέτα είναι το scikit-learn. Το πακέτο κάνει μεγάλη χρήση των

μαθηματικών εργασιών του. Το scikit-learn εκθέτει μια συνοπτική και συνεπή διεπαφή στους κοινούς αλγόριθμους μηχανικής μάθησης, καθιστώντας απλό να φέρει μηχανική μάθηση ( Machine Learning) σε συστήματα παραγωγής. Η βιβλιοθήκη συνδυάζει τον ποιοτικό κώδικα, την καλή τεκμηρίωση, την ευκολία χρήσης και την υψηλή απόδοση και είναι βιομηχανικό πρότυπο για την εκμάθηση μηχανών με την Python.

- **Sys (Systems – Specific Parameters and Function) :** Η sys παρέχει πληροφορίες σχετικά με τις σταθερές, τις λειτουργίες και τις μεθόδους του διερμηνέα της Python.

## 3.5 Ανάλυση κώδικα

Τα σύνολα δεδομένων περιέχουν συναλλαγές που πραγματοποιούνται από πιστωτικές κάρτες. Αυτό το σύνολο δεδομένων παρουσιάζει συναλλαγές που συνέβησαν σε δύο ημέρες. Περιέχει μόνο αριθμητικές μεταβλητές εισόδου που είναι το αποτέλεσμα ενός μετασχηματισμού PCA.

Δυστυχώς, λόγω προβλημάτων εμπιστευτικότητας, δεν παρέχθηκαν οι αρχικές λειτουργίες και περισσότερες πληροφορίες σχετικά με τα δεδομένα. Χαρακτηριστικά V1, V2, ... V28 είναι τα κύρια συστατικά που λαμβάνονται με PCA ενώ τα μοναδικά χαρακτηριστικά που δεν έχουν μετατραπεί με PCA είναι 'Time' και 'Amount'. Η λειτουργία 'Χρόνος' περιέχει τα δευτερόλεπτα που έχουν περάσει μεταξύ κάθε συναλλαγής και της πρώτης συναλλαγής στο σύνολο δεδομένων. Το χαρακτηριστικό 'Ποσό' είναι η ποσότητα συναλλαγής,. ***Το χαρακτηριστικό 'Class' είναι η μεταβλητή απόκρισης και παίρνει αξία <u>ένα</u> σε περίπτωση απάτης και σε περίπτωση μη απάτης <u>μηδέν.</u>***

Αρχικά δηλώνουμε τις απαραίτητες βιβλιοθήκες που θα χρειαστούμε. Έπειτα ανεβάζουμε τα δεδομένα μας στο πρόγραμμά μας.

```
# edw dilwnoume tis aparaithtes vivliothikes mas
import sys
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import scipy as sp
import sklearn
```

```
# edw anevazoume ta dedomena mas
data = pd.read_csv("PredictionFraud.csv")
```

Εικόνα 8 : Δήλωση βιβλιοθηκών και ανέβασμα δεδομένων

Παρακάτω παρατηρούμε ότι έχουμε 284.807 συναλλαγές με πιστωτικές κάρτες σε 31 στήλες από σημαντικές πληροφορίες για κάθε μία συναλλαγή . Οφείλουμε επίσης να επισημαίνουμε ότι εμάς μας ενδιαφέρει *η μεταβλητή class* όπου για περίπτωση απάτης παίρνουμε το ένα και για δόλια συναλλαγή παίρνουμε το μηδέν.

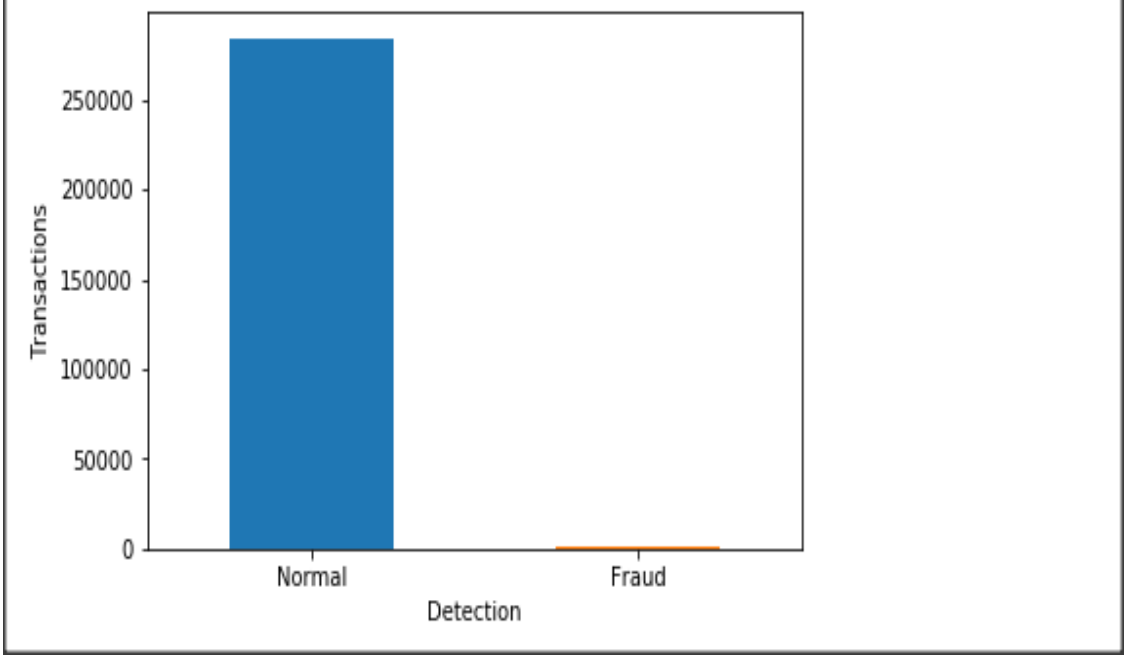
```
print(data.shape)
print(data.describe())
```

```
(284807, 31)
                Time             V1             V2             V3             V4  \
count  284807.000000  2.848070e+05  2.848070e+05  2.848070e+05  2.848070e+05
mean    94813.859575  3.919560e-15  5.688174e-16 -8.769071e-15  2.782312e-15
std     47488.145955  1.958696e+00  1.651309e+00  1.516255e+00  1.415869e+00
min         0.000000 -5.640751e+01 -7.271573e+01 -4.832559e+01 -5.683171e+00
25%     54201.500000 -9.203734e-01 -5.985499e-01 -8.903648e-01 -8.486401e-01
50%     84692.000000  1.810880e-02  6.548556e-02  1.798463e-01 -1.984653e-02
75%    139320.500000  1.315642e+00  8.037239e-01  1.027196e+00  7.433413e-01
max    172792.000000  2.454930e+00  2.205773e+01  9.382558e+00  1.687534e+01

                 V5             V6             V7             V8             V9  \
count  2.848070e+05  2.848070e+05  2.848070e+05  2.848070e+05  2.848070e+05
mean  -1.552563e-15  2.010663e-15 -1.694249e-15 -1.927028e-16 -3.137024e-15
std    1.380247e+00  1.332271e+00  1.237094e+00  1.194353e+00  1.098632e+00
min   -1.137433e+02 -2.616051e+01 -4.355724e+01 -7.321672e+01 -1.343407e+01
25%   -6.915971e-01 -7.682956e-01 -5.540759e-01 -2.086297e-01 -6.430976e-01
50%   -5.433583e-02 -2.741871e-01  4.010308e-02  2.235804e-02 -5.142873e-02
75%    6.119264e-01  3.985649e-01  5.704361e-01  3.273459e-01  5.971390e-01
max    3.480167e+01  7.330163e+01  1.205895e+02  2.000721e+01  1.559499e+01

            ...            V21            V22            V23            V24  \
count       ...   2.848070e+05  2.848070e+05  2.848070e+05  2.848070e+05
mean        ...   1.537294e-16  7.959909e-16  5.367590e-16  4.458112e-15
std         ...   7.345240e-01  7.257016e-01  6.244603e-01  6.056471e-01
min         ...  -3.483038e+01 -1.093314e+01 -4.480774e+01 -2.836627e+00
25%         ...  -2.283949e-01 -5.423504e-01 -1.618463e-01 -3.545861e-01
50%         ...  -2.945017e-02  6.781943e-03 -1.119293e-02  4.097606e-02
75%         ...   1.863772e-01  5.285536e-01  1.476421e-01  4.395266e-01
max         ...   2.720284e+01  1.050309e+01  2.252841e+01  4.584549e+00

                V25            V26            V27            V28         Amount  \
count  2.848070e+05  2.848070e+05  2.848070e+05  2.848070e+05  284807.000000
mean   1.453003e-15  1.699104e-15 -3.660161e-16 -1.206049e-16      88.349619
std    5.212781e-01  4.822270e-01  4.036325e-01  3.300833e-01     250.120109
min   -1.029540e+01 -2.604551e+00 -2.256568e+01 -1.543008e+01       0.000000
25%   -3.171451e-01 -3.269839e-01 -7.083953e-02 -5.295979e-02       5.600000
50%    1.659350e-02 -5.213911e-02  1.342146e-03  1.124383e-02      22.000000
75%    3.507156e-01  2.409522e-01  9.104512e-02  7.827995e-02      77.165000
max    7.519589e+00  3.517346e+00  3.161220e+01  3.384781e+01   25691.160000

              Class
count  284807.000000
mean        0.001727
std         0.041527
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max         1.000000

[8 rows x 31 columns]
```

Εικόνα 9 : Μεταβλητές δεδομένων

Μετέπειτα παίρνουμε το δείγμα πιστωτικών καρτών και το μελετάμε σε ποσοστό 100 %.Τα αποτελέσματα της μελέτης απεικονίζονται στο διάγραμμα 2 στο οποίο στον άξονα τον “χ” έχουμε την μεταβλητή detection ενώ στον άξονα τον “y” έχουμε τις transactions. Το ακιδωτό διάγραμμα των παραπάνω δεδομένων είναι το ακόλουθο:

```
count_classes = pd.value_counts(data['Class'] )
# rot=rotation
count_classes.plot(kind = 'bar',rot=0 )
# orizoume thn metavlhth prediction
prediction = ["Normal", "Fraud"]
plt.xticks(range(2),prediction)
# Ston aksona x exoume thn anixneysh
plt.xlabel("Detection")
# Ston aksona Y exoume tis sinallages
plt.ylabel("Transactions");
```

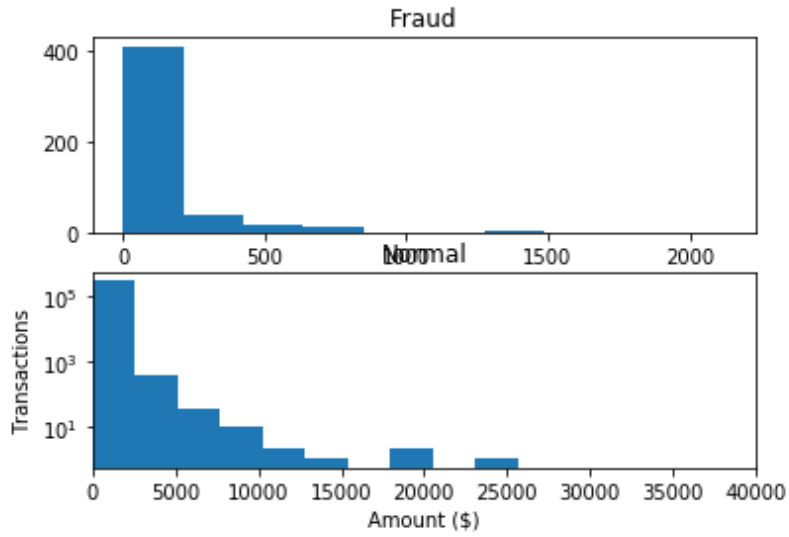



Διάγραμμα 2:Ανίχνευση σε σχέση με τις συναλλαγές

***Συμπεραίνουμε επομένως ότι ο αριθμός των δόλιων συναλλαγών είναι απειροελάχιστος.***

Στο παρακάτω **διάγραμμα 3** απεικονίζονται οι συναλλαγές σε σχέση με το ποσό. Στο πρώτο διάγραμμα έχουμε τις απάτες ενώ στο δεύτερο έχουμε τις μη δόλιες συναλλαγές.

```
# dilwnoume tis dolies sunnalages me 1
frauds = data[data.Class == 1]
# dilwnoume tis normal sunnalages me 0
normal = data[data.Class == 0]
f, (ax1, ax2) = plt.subplots(2,1 )
ax1.hist(frauds.Amount)
# To epanw diagramma mas deixnei ths apates
ax1.set_title('Fraud',)
ax2.hist(normal.Amount)
# To katw diagramma mas deixnei to Normal
ax2.set_title('Normal')
plt.xlabel('Amount ($)')
# Ston aksona y exoume tis sunnalages
plt.ylabel('Transactions')
plt.xlim((0, 40000))
plt.yscale('log')
plt.show();
```




Διάγραμμα 3:Ποσό σε σχέση με τις συναλλαγές

Παρατηρούμε ότι οι περισσότερες δόλιες συναλλαγές έχουν γίνει μέχρι το ποσό των 250 δολαρίων περίπου, ενώ σε μεγαλύτερα ποσά το ποσοστό τους μειώνεται. Από ποσό 1500 δολαρίων και πάνω θα μπορούσαμε να πούμε ότι το ποσοστό τους είναι σχεδόν μηδενικό.

Για τις μη δόλιες συναλλαγές συμπεραίνουμε ότι το μέγιστο ποσοστό τους είναι μέχρι 2500 χιλιάδες δολάρια. Σε ποσά μεγαλύτερα το 2500 χιλιάδων δολαρίων το ποσοστό τους μειώνεται κλιμακωτά και ελαχιστοποιείται κοντά στα 25000 χιλιάδες δολάρια.

Θα πάρω μόνο το 10 % των δεδομένων. Έπειτα πρόκειται να ορίσω μια τυχαία κατάσταση. Βέβαια οφείλω να επισημάνω ότι αν χρησιμοποιούσαμε όλα τα δεδομένα θα είχαμε καλύτερα αποτελέσματα.

```
# edw tha paroume mono to 10%
# orizoyme kai mia tixaia katastash
data =data.sample(frac = 0.1, random_state = 1 )
print(data.shape)
```

```
(28481, 31)
```

Εικόνα 10 : Χρησιμοποίηση του 10% των δεδομένων

Τώρα θα υπολογίσουμε τον αριθμό των δολίων περιπτώσεων ώστε να μπορούμε να πάρουμε ένα *κλάσμα (outlier fraction)* που πρόκειται να περάσει στις μελλοντικές μεθόδους ανίχνευσης παρεκτροπών (Anomaly Detection). Θα υπολογίσουμε το ποσοστό των ψευδών περιπτώσεων. Από τα αποτελέσματα που παρουσιάζουμε παρακάτω παρατηρούμε ότι έχουμε μόνο το 0,17 % της βάσης δεδομένων δόλιων υποθέσεων και συνολικά 49 δόλιες περιπτώσεις. Επίσης υπάρχει τεράστια ανισότητα μεταξύ των δόλιων υποθέσεων και των έγκυρων συναλλαγών.

```
# frauds have one
# normal have zero
fraud = data[data.Class == 1]
normal = data[data.Class == 0]
# vazoume to float gia na mhn mas gurisei sto mhden
outlier_fraction =  len (fraud) / float (len(normal))
print(outlier_fraction)
print('in case fraud:{}'.format(len(fraud)))
print('in case normal:{}'.format(len(normal)))
```

```
0.00172341024198
in case fraud:49
in case normal:28432
```

Εικόνα 11:Αποτελέσματα απάτης και μη απάτης

Πρέπει λοιπόν να έχουμε όλες τις στήλες μας από το πλαίσιο δεδομένων ( data frame). Για να γίνει αυτό θα κάνουμε τις στήλες να ισούνται με τις στήλες δεδομένων σε μια λίστα που θα δημιουργήσει μια λίστα με στήλες. Θέλουμε να φιλτράρουμε τις στήλες για την κατάργηση των δεδομένων. *Ο στόχος μας εδώ θα είναι η μεταβλητή "class".* Έπειτα προσπαθούμε να προβλέψουμε ώστε η χ μεταβλητή να είναι τα δεδομένα μας.

Οπότε τώρα είμαστε σε θέση να εκτυπώσουμε και μπορούμε να παρατηρήσουμε ότι έχουμε 30 στήλες στο χ που είναι όλα εκτός της ετικέτας της κλάσης και το y μας είναι ένας πίνακας διαστάσεων που έχει τις ετικέτες κλάσης για όλα τα δείγματα από το σύνολο δεδομένων μας. Έπειτα πρόκειται να χρησιμοποιήσουμε έναν αλγόριθμο απομόνωσης δασών (Isolation Forest) και έναν τοπικό αλγόριθμο εξωγενών παραγόντων ( Local outlier Factor ) για να προσπαθήσουμε να κάνουμε ανίχνευση παρεκτροπών (Anomaly Detection) σε αυτό το σύνολο δεδομένων.

```
# dinw oles ti steiles apo ta data mou
columns = data.columns.tolist()
# edw svinoume thn metavlhth class apo ta dedomena mas
columns =[c for c in columns if c not in["class"]]
target = "Class"
x = data[columns]
y = data[target]
print (x.shape)
print (y.shape)
```

```
(28481, 30)
(28481L,)
```

Εικόνα 12:Χρησιμοποίηση της μεταβλητής “Class”

Τώρα θα πάμε να προβλέψουμε ποιες είναι Outliers. Πρέπει να εισάγουμε πακέτα από την βιβλιοθήκη SK Learn. Εμείς θα κάνουμε Outlier Detection οπότε πρόκειται να κάνουμε μια Accuracy Score και μια ταξινόμηση Classification.

Επίσης θα χρησιμοποιήσουμε δύο μεθόδους

- Local Outlier Factor και
- Isolation Forests

Αυτές είναι οι δύο ποιο συνήθεις μέθοδοι για Anomaly detection.

Ο Local Outlier Factor είναι μια ανεξέλεγκτη μέθοδος ανίχνευσης παρεκτροπών και υπολογίζει την Anomaly score κάθε δείγματος. Ονομάζεται Local Outlier Factor επειδή μετρά την τυπική απόκλιση της πυκνότητας ενός δεδομένου δείγματος σε σχέση με τους γείτονές του, δηλαδή τοπική βαθμολογία ανωμαλίας Anomaly Score Επίσης εξαρτάται από το πόσο απομονωμένο είναι το αντικείμενο σε σχέση με τη γύρω γειτονιά, έτσι μιλάμε για Neighbors και αυτό θα καθοριστεί με τον ίδιο τρόπο όπως και η πλησιέστερη μέθοδος πλησιέστερων γειτόνων (K Nearest Neighbors Method) Έπειτα υπολογίζουμε ένα Anomaly score που βασίζεται σε εκείνους τους neighbors.

Ο αλγόριθμος Isolation Forest που πρόκειται να επιστρέψει το Anomaly score από κάθε δείγμα είναι λίγο διαφορετικός. Αυτή η μέθοδος Isolation Forest κάνει και απομονώνει τις παρατηρήσεις επιλέγοντας τυχαία ένα χαρακτηριστικό. Κατόπιν επιλέγει μια τυχαία διαχωρισμένη τιμή μεταξύ της μέγιστης και της ελάχιστης τιμής από την επιλεγμένη λειτουργία. Έτσι θα έχουμε όλες τις διαφορετικές στήλες εδώ, που θα μπορούσαν να θεωρηθούν σαν ένα χαρακτηριστικό δεδομένου.

Πρώτα από όλα θα ορίσουμε μια τυχαία κατάσταση και έπειτα πρόκειται να ορίσουμε τις μεθόδους ανίχνευσης. Τώρα θα βάλουμε το Classifiers όπου θα βάλουμε έναν Isolation Forest και έχουμε ορίσει ένα ζευγάρι από παραμέτρους, που στο πρώτο θα έχουμε τα μέγιστα δείγματα και στο δεύτερο θα βάλουμε το μήκος του χ έτσι ώστε τα μέγιστα δείγματα να είναι ο συνολικός αριθμός των δειγμάτων.

Επίσης, έχουμε και το Local outlier factor όπου είναι ένας τοπικός παράγοντας απόκλισης και ο συγκεκριμένος αποτελείται από κάποιες παραμέτρους καθώς ο αριθμός των neighbors είναι αυτό που πηγαίνει στην μέθοδο του πλησιέστερου K neighbors που χρησιμοποιείται.
Εμείς πρόκειται να ορίσουμε το 20 ως είδος προεπιλογής ή ως τυποποιημένο υψηλότερο ποσοστού του συνόλου. Αν το μειώσουμε θα διαπιστώσουμε ότι θα παραχθούν διαφορετικά αποτελέσματα. Επίσης ορίζουμε την μεταβλητή "contamination" ίση με το κλάσμα.

```
from sklearn.metrics import classification_report, accuracy_score
from sklearn.ensemble import IsolationForest
from sklearn.neighbors import LocalOutlierFactor

# orizw mia tixaia katastash
state=1
# kathorizw tis methodous anixneyshs
classifiers= {
    "Isolation Forest" : IsolationForest(max_samples=len(x),
                                         contamination=outlier_fraction,
                                        random_state = state ),
    "Local Outlier Factor": LocalOutlierFactor (n_neighbors = 20,
                                                contamination = outlier_fraction)
}

```

Εικόνα 13 : Μέθοδοι Ανίχνευσης

Πρόκειται να ορίσουμε μια μεταβλητή που ονομάζεται ο αριθμός των outliers και αυτό θα είναι ίσο με το μήκος της απάτης.
Επίσης θα θέλαμε να ξανά επισημάνουμε ότι η ετικέτα class είναι 0 για έγκυρες και 1 για δόλιες συναλλαγές.

```
n_outliers = len (fraud)

for i, (clf_name, clf) in enumerate (classifiers.items()):
    # edw prosarmozoume ta dedomena kai epishmanoume ta apotelesmata
    if clf_name == "Local Outlier Factor":
        Y_pred = clf.fit_predict(x)
        scores_pred = clf.negative_outlier_factor_
    else:
        clf.fit(x)
        scores_pred = clf.decision_function(x)
        Y_pred = clf.predict(x)

    # edw vazoume 0 gia kanonikh sunnalagh
    # 1 gia dolia sinalagh
        Y_pred[Y_pred == 1] = 0
        Y_pred[Y_pred == -1] = 1

        n_errors = (Y_pred !=y).sum()

        print('{}:{}'.format(clf_name, n_errors))
        print(accuracy_score(y, Y_pred))
        print(classification_report(y, Y_pred))

```

Εικόνα 14: Εκτύπωση αποτελεσμάτων

Στην ανάλυση του Local Outlier Factor με 97 σφάλματα έχουμε σχετικά προσεγγίσει σε υψηλό επίπεδο αλλά βρισκόμαστε στο 99% ακριβείς, ενώ άμα κοιτάξουμε την κλάση 1 και δούμε την ανάκληση (recall), το Score F1 όπως και την ακρίβεια (precision) 2 % σημαίνει ότι δεν είναι πραγματικά καλό και έχουμε πολύ λίγες πραγματικές δόλιες περιπτώσεις που παίρνουν την ετικέτα ψευδείς περιπτώσεις. Επίσης συμπεραίνουμε ότι μπορεί να προβλέψει απάτη μέχρι 2 %. Το υπόλοιπο ποσοστό παραμένει μη εντοπισμένο από το σύστημα. *Οι ετικέτες precision, recall, f1-score και το support είναι παράμετροι ανάλυσης*.

```
Local Outlier Factor: 97
0.996594220
             precision    recall  f1-score   support

          0       1.00      1.00      1.00     28432
          1       0.02      0.02      0.02        49

avg / total       1.00      1.00      1.00     28481
```

Εικόνα 15: Local outlier factor

Παρομοίως στο isolation forest βρισκόμαστε στο 99%.Το Isolation Forest είναι αρκετά καλύτερο από τον Local Outlier Factor.Στην κλάση 0 έχουμε precision και recall 100 %. Ενώ αντιθέτως στην κλάση 1 έχουμε precision και recall γύρω στο 30 %. Μπορεί επομένως να προβλέψει απάτη έως 28%, ενώ το υπόλοιπο ποσοστό παραμένει μη εντοπισμένο από το σύστημα

```
Isolation Forest:71
0.99
             precision    recall  f1-score   support

          0       1.00      1.00      1.00     28432
          1       0.28      0.29      0.28        49

avg / total       1.00      1.00      1.00     28481
```

Εικόνα 16:Isolation forest

# Βιβλιογραφία

**Ελληνική:**


[1] Μυλωνάς Γεώργιος (2011) , Αλγόριθμοι και Τεχνικές Εξόρυξης Δεδομένων για Παραγωγή Προτάσεων και Προσωποποίηση Χρηστών, Διπλωματική Εργασία για το ΠΜΣ «Εφαρμοσμένη Πληροφορική » , Πανεπιστήμιο Μακεδονίας.

[2] Κωνσταντίνου Π. Καρποδίνη (2016) , Ανάλυση συναισθημάτων σε δεδομένα από το Twitter χρησιμοποιώντας εργαλεία της R και μοντέλα μηχανικής μάθησης , Διπλωματική Εργασία για το τμήμα Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών της Πολυτεχνικής Σχολής , Πανεπιστημίου Πατρών.

[3] Σφακιανάκης Θεόδωρος ( 2013), Προσομοίωση Ανελίξεων Levy Με Εφαρμογές στην Αποτίμηση παράγωγων Χρηματοοικονομικών Προϊόντων, Διπλωματική Εργασία για το ΠΜΣ « Εφαρμοσμένη Στατιστική » , Πανεπιστήμιο Πειραιώς .

[4] Ματιάκη Άννα (2007) , Η Εξόρυξη Δεδομένων ( Data Mining) στη Λογιστική και Ελεγκτική, Διπλωματική Εργασία για το ΠΜΣ « Πληροφορική και Διοίκηση »

[5] Κούτρας, Μ (2007). Εφαρμοσμένη Πολυμεταβλητή Ανάλυση : Ανάλυση κατά συστάδες ( Σημειώσεις), ΠΜΣ στην Εφαρμοσμένη Στατιστική , Πανεπιστήμιο Πειραιώς.

[6] Φωκιανός, Κ και Χαραλάμπους, Χ (2008).Εισαγωγή στην R ( Σημειώσεις) ,Τμήμα Μαθηματικών και Στατιστικής, Πανεπιστήμιο Κύπρου.

[7] Αναστασία – Δήμητρα Λυπιτάκη (2014), Μηχανική Μάθηση σε Ανομοιογενή Δεδομένα, Μεταπτυχιακή διατριβή για το Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών.

[8] Χαλκίδη Μ, Βαρζιγιάννης Μ. (2006).Εξόρυξη γνώσης από βάσεις δεδομένων και παγκόσμιο ιστό.

[9] Φωκιανός, Κ. και Χαραλάμπους, Χ. (2008). Εισαγωγή στην R, (Σημειώσεις), Τμήμα Μαθηματικών και Στατιστικής, Πανεπιστήμιο Κύπρου.

[10] Βαρζιγιάννης, Μ. και Χαλκιδή, Μ. (2005). Εξόρυξη από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό, 2η εκδ, τυπωθήτω - Γιώργος Δαρδάνος, Αθήνα.

**Ξενόγλωσση:**

[11] Abonyi, J. and Feil, B. (2007).Cluster Analysis for Data Mining and System Identification, Birkhauser Verlag AG.

[12] Berkhin P.(2002).Survey of clustering Data Mining Techniques, Technical Report, Accrue Software.

[13] BATISTA G.E.A.P.A., PRATI R.C. and MONARD M.C. (2004): A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data, ACM SIGKDD Explorations Newsletter 6 (1), 20-29.

[14] BENDER E.A. (1996): Mathematical methods in Artificial Intelligence, IEEE Computer Society Press, Los Alamitos, California.

[15] Larose DT (2006). Data Mining: Methods and Models, New York.

[16] Cios K. J., Pedrycz W., Swiniarski R. W., Kurgan L. A. (2007). Data Mining: A Knowledge Discovery Approach

[17] Liu B. (2007). Web data mining: Exploring hyperlinks, contents, and usage data , Springer M. H. Dunham (2002). Data Mining: Introductory and Advanced Topics. Prentice Hall.

[18] J. Han and M. Kamber (2001), Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco.

**Διαδικτυακοί ισότοποι:**

[19]https://medium.com/activewizards-machine-learning-company/top-15-python-libraries-for-data-science-in-in-2017-ab61b4f9b4a7

[20]https://www.python-course.eu/sys_module.php

https://www.meccanismocomplesso.org/en/python-standard-library-sys-module-modulo-sys/

[21]https://pandas.pydata.org/pandasdocs/version/0.17.0/generated/pandas.DataFrame.shape.html

[22] https://scikit-learn.org/0.19/auto_examples/neighbors/plot_lof.html

https://el.wikipedia.org/wiki/%CE%A4%CF%85%CF%80%CE%B9%CE%BA%CE%AE_%CE%B1%CF%80%CF%8C%CE%BA%CE%BB%CE%B9%CF%83%CE%B7
[23]https://scikitlearn.org/stable/auto_examples/neighbors/plot_lof_outlier_detection.html
[24] http://jupyter.org/try
[25]https://www.oreilly.com/library/view/introduction-to-hine/9781449369880/ch01.html
[26]https://github.com/yazanobeidi/fraud-detection/blob/master/project.ipynb
[27]https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6
[28]https://machinelearningmastery.com/machine-learning-in-python-step-by-step/
https://www.data-blogger.com/2017/06/15/fraud-detection-a-simple-machine-learning-approach/
[29]http://nevonprojects.com/latest-data-mining-projects-topics-ideas/
[30]https://www.datacamp.com/community/tutorials/pandas-tutorial-dataframe-python#question1
[31]https://towardsdatascience.com/outlier-detection-with-extended-isolation-forest-1e248a3fe97b