ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Σύγκριση αλγορίθμων και ανάλυση τεχνικών μηχανικής μάθησης

Δημήτριος Χανιώτης

Επιβλέπων Καθηγητής:
Μιχαήλ Φιλιππάκης, Αναπληρωτής Καθηγητής

ΠΕΙΡΑΙΑΣ

ΣΕΠΤΕΜΒΡΙΟΣ 2022

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Σύγκριση αλγορίθμων και ανάλυση τεχνικών μηχανικής μάθησης

Δημήτριος Χανιώτης
Α.Μ.: ΜΕ2061

ΠΕΡΙΛΗΨΗ

Η ανάλυση δεδομένων στη σύγχρονης επιστήμης ασχολείται με τη διαχείριση και την ερμηνεία αξιοποιήσιμων πληροφοριών. Με τη χρήση αλγορίθμων εξάγεται πληθώρα συμπερασμάτων που βοηθούν στην ερμηνεία συμπεριφορών αλλά και στην πρόβλεψη μελλοντικών καταστάσεων. Αυτοί οι αλγόριθμοι βοηθούν τόσο στα επιστημονικά όσο και στα εμπορικά πεδία.

Η παρούσα διπλωματική εργασία έχει ως στόχο τη σύγκριση αλγορίθμων και την ανάλυση των τεχνικών της μηχανικής μάθησης. Με τη χρήση ενός συνόλου δεδομένων μιας αλυσίδας σούπερ μάρκετ εφαρμόζονται πέντε διαφορετικοί αλγόριθμοι και συγκρίνονται οι ικανότητές τους στην πρόβλεψη και την ανάλυση δεδομένων.

Στο τέλος της εργασίας συγκρίνονται τα αποτελέσματα των αλγορίθμων και παρουσιάζεται ο αλγόριθμος που σε αυτή την πρόβλεψη εμφανίζει την καλύτερη εκτίμηση.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Μηχανική μάθηση

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Μηχανική μάθηση, Ανάλυση δεδομένων, Αλγόριθμοι, Σύνολο δεδομένων

ABSTRACT

Data analysis in modern science deals with managing and interpreting actionable information. A lot of conclusions can derive from the use and application of algorithms, that help to analyze behaviors and predict future situations. These algorithms help in both scientific and commercial fields.

This thesis aims to compare algorithms and analyze machine learning techniques. Five different algorithms are applied, and their prediction and data analysis capabilities are compared using a data set of a supermarket chain.

At the end of this thesis, the results of the algorithms are compared and the algorithm that shows the best estimation in this prediction is presented.

SUBJECT AREA: Machine learning

KEYWORDS: Machine learning, Data analysis, Algorithms, Data set

Στη γιαγιά μου.

ΕΥΧΑΡΙΣΤΙΕΣ

Για τη διεκπεραίωση της παρούσας Πτυχιακής Εργασίας, θα ήθελα να ευχαριστήσω τον υπεύθυνο και επιβλέποντα καθηγητή κ. Μιχαήλ Φιλιππάκη, για τη συνεργασία και την πολύτιμη συμβολή του στην ολοκλήρωση της. Επιπλέον Θα ήθελα να ευχαριστήσω την Δρ Πούλου Μαριλένα για τη βοήθεια στην επίβλεψη της ανάλυσης των δεδομένων, τη συμβολή της στο πειραματικό μέρος της διατριβής και τα χρήσιμα σχόλιά της στην ανάλυση της έρευνας. Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια και τους φίλους μου που είναι πάντα δίπλα μου.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ 10
1. ΑΛΓΟΡΙΘΜΟΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ 12
1.1 ΟΡΙΣΜΟΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ 12
1.2 ΚΑΤΗΓΟΡΙΕΣ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ 13
1.2.1 ΕΠΙΒΛΕΠΟΜΕΝΗ ΜΑΘΗΣΗ 13
1.2.1.1 ΜΟΝΤΕΛΑ ΤΑΞΙΝΟΜΙΣΗΣ 14
1.2.1.2 ΠΑΛΙΝΔΡΟΜΙΣΗ 16
1.2.2 ΜΗ ΕΠΙΒΛΕΠΟΜΕΝΗ ΜΑΘΗΣΗ 16
1.2.2.1 ΜΟΝΤΕΛΑ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ 16
1.2.2.2 ΑΛΓΟΡΙΘΜΟΙ ΕΞΑΓΩΓΗΣ ΚΑΝΟΝΩΝ ΣΥΣΧΕΤΙΣΗΣ 18
1.2.3 Η ΕΠΙΡΡΟΗ ΤΟΥ ΘΟΡΥΒΟΥ ΣΗΝ ΕΚΠΑΙΔΕΥΣΗ ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ 18
1.2.3.1 ΜΕΡΟΛΗΨΙΑ ΠΑΡΑΛΕΙΠΟΜΕΝΩΝ ΜΕΤΑΒΛΗΤΩΝ 19
1.2.3.2 ΔΙΑΣΠΟΡΑ 20
1.2.3.3 BIAS – VARIANCE TRADE-OFF 20
1.2.4 ΥΠΕΡΑΠΛΟΥΣΤΕΥΣΗ & ΥΠΕΡΠΡΟΣΑΡΜΟΓΗ 21
1.2.5 ΕΦΑΡΜΟΓΕΣ ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ 21
2. ΥΛΟΠΟΙΗΣΗ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ 24
2.1 ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ 24
2.1.1 ΠΛΗΡΟΦΟΡΙΕΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ 24
2.1.2 ΑΝΑΛΥΣΗ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ 25
2.1.3 ΠΡΟΕΤΟΙΜΑΣΙΑ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΠΡΟΒΛΕΨΗ 29
3. ΕΦΑΡΜΟΓΗ ΠΡΟΒΛΕΠΤΙΚΩΝ ΜΟΝΤΕΛΩΝ 31
3.1 ΠΡΟΒΛΕΨΗ ΑΚΑΘΑΡΙΣΤΟΥ ΕΙΣΟΔΗΜΑΤΟΣ ΤΗΣ ΕΤΑΙΡΕΙΑΣ, ΜΕΜΟΝΩΜΕΝΑ ΚΑΙ ΑΝΑ ΠΕΡΙΟΔΟ 31
3.1.1 ΕΠΙΛΟΓΗ ΛΕΙΤΟΥΡΓΙΩΝ ΚΑΙ ΔΙΑΧΩΡΙΣΜΟΣ ΔΕΔΟΜΕΝΩΝ 31
3.1.1.1 ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ (LINEAR REGRESSION) 32
3.1.1.2 RANDOM FOREST 32
3.1.1.3 SUPPORT VECTOR MACHINE 32
3.1.1.4 kNN 32
3.1.1.5 GRADIENT BOOSTING 32
4. ΑΞΙΟΛΟΓΗΣΗ ΜΟΝΤΕΛΩΝ 33
ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ 34

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

Σχήμα 1 Κατηγοριοποίηση Μηχανικής Μάθησης.....12
Σχήμα 2 Κατηγορίες αλγορίθμων μηχανκής μάθησης 13
Σχήμα 3 Δέντρα απόφασης 15
Σχήμα 4 Νευρωνικά δίκτυα.....15
Σχήμα 5 Μεροληψία - Διασπορά 20

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 1Μέρος του συνόλου δεδομένων.....24
Πίνακας 2 Συγκεντρωτικά στοιχεία πωλήσεων ανά κατάστημα και ανά πόλη 28
Πίνακας 3Τα προβλεπτικά μοντέλα και οι καλύτεροι παράμετροι.....33
Πίνακας 4Τα προβλεπτικά μοντέλα και το MSE.....33

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ

Διάγραμμα 1Τύποι καταναλωτών 25
Διάγραμμα 2 Αριθμός συναλλαγών ανά φύλο 26
Διάγραμμα 3 Ακαθάριστο εισόδημα σε εκ. ανά πόλη 26
Διάγραμμα 4 Κατανομή ακαθάριστου εισοδήματος 27
Διάγραμμα 5 Κατανομή τιμής μονάδας.....27
Διάγραμμα 6 Ακαθάριστο εισόδημα ανά σειρά προϊόντων 28

ΠΡΟΛΟΓΟΣ

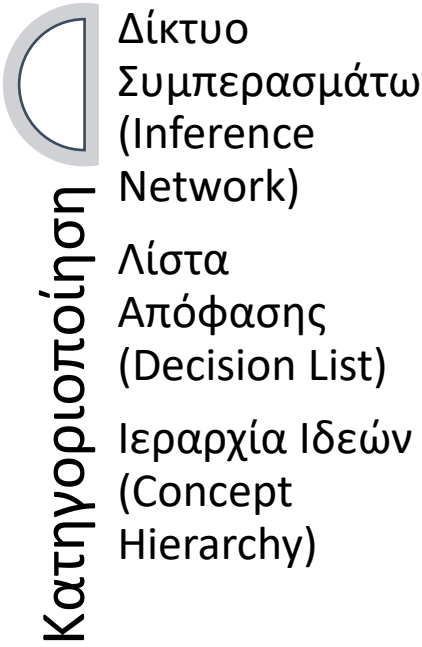
Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του Προγράμματος Μεταπτυχιακών Σπουδών «Πληροφοριακά Συστήματα και Υπηρεσίες» ειδίκευση «Προηγμένα Πληροφοριακά Συστήματα» του τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς.

1. ΑΛΓΟΡΙΘΜΟΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

1.1 ΟΡΙΣΜΟΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Η μηχανική μάθηση (ML: Machine Learning) ή αλλιώς προγνωστική αναλυτική στην επιστήμη των υπολογιστών είναι υπολογιστικές μέθοδοι, οι οποίες χρησιμοποιώντας δεδομένα μπορούν να παρέχουν προβλέψεις και εκτιμήσεις υψηλής ακρίβειας. Σύμφωνα με τον Mitchell (1997) ένα πρόγραμμα υπολογιστή μπορεί να μάθει από την εμπειρία E (Experience) όσον αφορά μια κατηγορία εργασιών T (Tasks) και ενός μέτρου απόδοσης P (Performance Measure), εάν η απόδοση στις εργασίες της T, όπως μετριούνται με την P, βελτιώνονται με την εμπειρία E.

Τις περισσότερες φορές ο συσχετισμός αλγορίθμων μηχανικής μάθησης απαιτείται, και έτσι κρίνεται απαραίτητη η οργάνωση της εξαγόμενης γνώσης.



Σχήμα 1 Κατηγοριοποίηση Μηχανικής Μάθησης

Η οργάνωση της γνώσης σύμφωνα με τον Langley (1996) μπορεί να κατηγοριοποιηθεί ακολουθώντας τα παρακάτω κριτήρια:

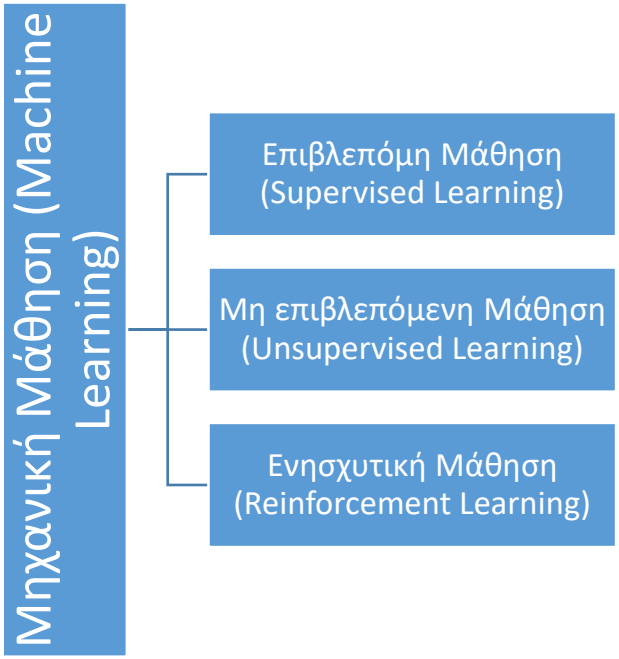
- Δίκτυο Συμπερασμάτων (Inference Network): Τα δεδομένα αυτής της δομής οργανώνονται σε μορφή κατευθυνόμενου γράφου ή δένδρου με συνέπεια η προέκταση του κάθε κόμβου επηρεάζεται άμεσα από τις συνδέσεις του με τους κόμβους κάτω από αυτό.
- Λίστα Απόφασης (Decision List): Οι πληροφορίες σε αυτή τη δομή ταξινομούνται με τέτοιο τρόπο ώστε οι πληροφορίες που βρίσκονται υψηλότερα στη λίστα να λαμβάνονται υπόψιν νωρίτερα κατά την εκτέλεση του συστήματος. Ορισμένες μεταβλητές υπάρχει η δυνατότητα να αποθηκευτούν σε αλληλοαποκλειόμενη σειρά έτσι ώστε η ύπαρξη μιας πληροφορίας να προϋποθέτει την

απουσία μιας άλλης. Αυτή η μέθοδος είναι καταλληλότερη για την οργάνωση ανταγωνιστικών μεταξύ τους στοιχείων.

- Ιεραρχία ιδεών (Concept Hierarchy): Σε αυτή τη δομή τα δεδομένα οργανώνονται με τη μορφή κατευθυνόμενου γράφου ή δένδρου. Σε αντίθεση με την οργάνωση σε δίκτυο συμπερασμάτων, σε αυτή τη δομή κάθε κόμβος αντιστοιχεί σε μία ιδέα που συνοδεύεται από τη σχετική περιγραφή της. Πιο ψηλά στην ιεραρχία κατατάσσονται οι κόμβοι με γενικές περιγραφές, ενώ στα πιο χαμηλά επίπεδα κατατάσσονται οι κόμβοι που έχουν πιο συγκεκριμένες περιγραφές.

1.2 ΚΑΤΗΓΟΡΙΕΣ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Οι αλγόριθμοι μηχανικής μάθησης διαχωρίζονται ανάλογα με τον βαθμό επίβλεψης τους. Έτσι λοιπόν τα μοντέλα μηχανικής μάθησης τα διαχωρίζουμε σε τρείς κατηγορίες: την Επιβλεπόμενη Μάθηση (Supervised Learning), τη Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning) και την Ενισχυτική Μάθηση (Reinforcement Learning).



Σχήμα 2 Κατηγορίες αλγορίθμων μηχανκής μάθησης

1.2.1 ΕΠΙΒΛΕΠΟΜΕΝΗ ΜΑΘΗΣΗ

Επιβλεπόμενη Μάθηση ορίζεται το είδος μηχανικής μάθησης στο οποίο εξετάζεται μια άγνωστη συνάρτηση χρησιμοποιώντας δεδομένα τα οποία αποδίδουν τα αποτελέσματα της συνάρτησης αυτής. Σε τέτοιου είδους αλγόριθμους ορίζονται πληροφορίες εισόδου αλλά και γνωστά αποτελέσματα και ανάλογα με τους μεταξύ τους συσχετισμούς παράγεται το μοντέλο που τα

περιγράφει. Ειδικότερα αυτό το προβλεπτικό μοντέλο δημιουργεί μια συνάρτηση που συνδέει τα δεδομένα εισόδου με τον στόχο. Όλη η ανωτέρω διαδικασία συνιστά την εκπαίδευση του μοντέλου αφού ολοκληρωθεί, ένα νέο σύνολο δεδομένων εισάγεται στο προβλεπτικό μοντέλο που παράχθηκε και με αυτό τον τρόπο ο αλγόριθμος παράγει νέες προβλέψεις. Έτσι λοιπόν στόχος σε αυτή την περίπτωση είναι η αξιολόγηση των συμπερασμάτων βάσει μετρήσεων σε έναν όγκο δεδομένων εισόδου.

1.2.1.1ΜΟΝΤΕΛΑ ΤΑΞΙΝΟΜΙΣΗΣ

Τα μοντέλα ταξινόμησης περιγράφουν τη σύνθεση μοντέλων τα οποία δύναται να κατηγοριοποιήσουν δεδομένα τα οποία είναι μη κατηγοριοποιημένα. Πιο συγκεκριμένα στόχος είναι η εύρεση των ορίων απόφασης που ορίζουν πως θα διαιρεθούν οι κλάσεις. Οι Tan et al.(2014) ορίζουν ως ταξινόμηση τη διαδικασία μάθησης μιας συνάρτησης f, η οποία αντιστοιχεί κάθε διάνυσμα μεταβλητών x σε ένα προκαθορισμένο πλήθος κλάσεων y.

Τα μοντέλα ταξινόμησης είναι αποτελεσματικότερα όταν εφαρμόζονται σε κατηγορικά (nominal) ή δυαδικά δεδομένα (binary) και λιγότερο αποτελεσματικά όταν υπάρχουν ιεραρχικά δεδομένα (ordinal) καθώς δε λαμβάνουν υπόψη την ύπαρξη σειράς ανάμεσα στις κατηγορίες.

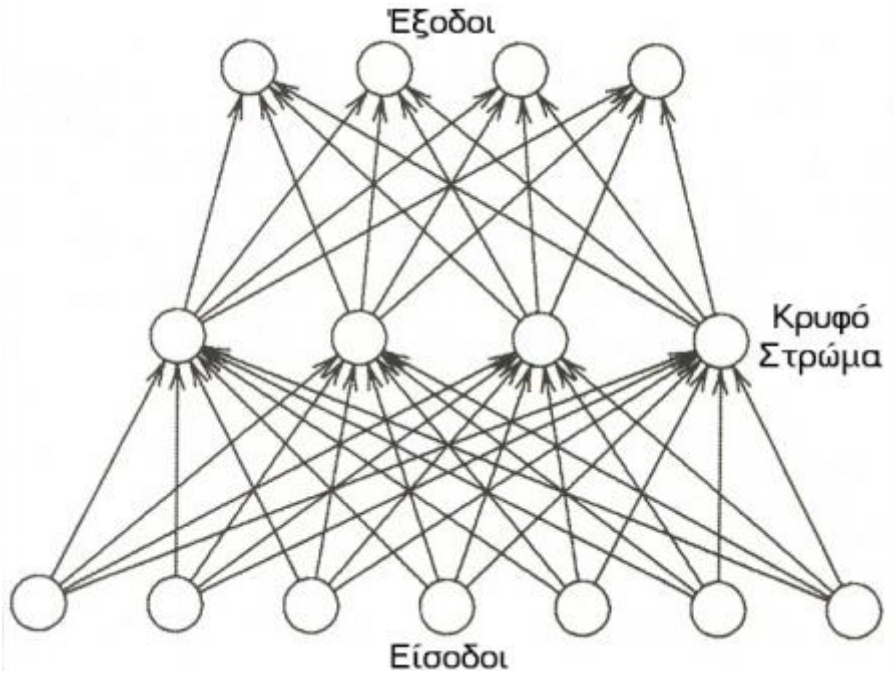
Ένα μοντέλο ταξινόμησης (Tan et al., 2014) μπορεί να χρησιμοποιηθεί ως διευκρινιστικό εργαλείο για το διαχωρισμό αντικειμένων που ανήκουν σε διαφορετικές κλάσεις αλλά και ως εργαλείο πρόβλεψης (descriptive modeling) κλάσεων νέων εγγραφών (predictive modeling).

Μερικές κατηγορίες αλγορίθμων ταξινόμησης είναι:

- Δέντρα Απόφασης (Decision Trees): Αποτελεί μια από τις δημοφιλέστερες μεθόδους ταξινόμησης και απεικονίζεται με μια δομή δέντρου. Στην περίπτωση που η επιλεγμένη για προσδιορισμό μεταβλητή λαμβάνει διακριτές τιμές το δέντρο απόφασης ονομάζεται δέντρο ταξινόμησης (classification tree), ενώ στην περίπτωση που λαμβάνει συνεχείς τιμές καλείται δέντρο παλινδρόμησης (regression tree). Στο δέντρο υφίσταται ο κόμβος – ρίζα που έχει μόνο ακμές που εξέρχονται και γίνεται σύνδεση με κόμβους χαμηλότερου επιπέδου, οι εσωτερικοί κόμβοι και οι εξωτερικοί κόμβοι – φύλλα. Για να μεταβούμε από ένα κόμβο ανώτερου επιπέδου σε κάποιον άλλο κατώτερου επιπέδου πραγματοποιείται έλεγχος μιας συνθήκης και στη συνέχεια διαχωρισμός των δεδομένων σε ένα ή περισσότερα υποσύνολα. Η διαδικασία επαναλαμβάνεται για τη μετάβαση σε χαμηλότερο επίπεδο μέχρι όλα τα χαρακτηριστικά να εισέλθουν στους κόμβους του δέντρου. Οι κόμβοι – φύλλα που βρίσκονται στο κατώτερο επίπεδο κάθε κλάδου αντιπροσωπεύουν τις προβλέψεις των τιμών του επιθυμητού χαρακτηριστικού οι οποίες έχουν προκύψει από το μονοπάτι με αφετηρία τη ρίζα και προορισμό το κάθε φύλλο.

Σχήμα 3 Δέντρα απόφασης

- Νευρωνικά δίκτυα (Neural Networks): Τα νευρωνικά δίκτυα μοιάζουν με το βιολογικό νευρικό σύστημα και ειδικότερα με το νευρικό σύστημα του εγκεφάλου. Ένα νευρωνικό δίκτυο μπορεί να περιγραφεί από έναν κατευθυνόμενο γράφο, οι κόμβοι του οποίου ονομάζονται νευρώνες και οι συνδέσεις αναπαριστούν τις σχέσεις μεταξύ των νευρώνων. Μέσω των συνδέσεων μεταφέρονται πληροφορίες από τις εισόδους του δικτύου προς τις εξόδους του. Οι συνδέσεις έχουν βάρη (Weight), τα οποία ρυθμίζουν το βαθμό αλληλεπίδρασης για κάθε ζεύγος νευρώνων. Ένα δίκτυο που περιλαμβάνει αμφίδρομες συνδέσεις λέγεται αναδρομικό (Recurrent), ενώ ένα δίκτυο χωρίς αμφίδρομες συνδέσεις ονομάζεται απλής εμπροσθοτροφοδοτούμενο (Feed Forward). Τέλος, τα νευρωνικά δίκτυα εφαρμόζονται τόσο για επιβλεπόμενη όσο και για μη επιβλεπόμενη μάθηση.



Σχήμα 4 Νευρωνικά δίκτυα

- Μπαϋσιανά δίκτυα (Bayesian Networks): Σε αυτή τη μέθοδο χρησιμοποιούνται πιθανοτικά μοντέλα τα οποία βασίζονται στο θεώρημα του Bayes. Βασική δυσκολία στη μέθοδο αυτή είναι η ανάγκη γνώσης πολλών τιμών πιθανοτήτων η οποία αρκετές φορές οδηγεί στην αντικατάστασή τους με εκτιμήσεις από παλαιότερες υποθέσεις ή από εμπειρική γνώση. Επιπλέον, για την αντιμετώπιση της παραπάνω δυσκολίας χρησιμοποιούνται συχνά τα απλά δίκτυα Bayes (Simple / Naive Bayes Classifier) τα οποία θεωρούν ανεξαρτησία ανάμεσα στα χαρακτηριστικά.

1.2.1.2 ΠΑΛΙΝΔΡΟΜΙΣΗ

Η παλινδρόμηση (Regression) αναφέρεται στη διαδικασία διερεύνησης των σχέσεων μεταξύ της μεταβλητής απόκρισης (Response Variable) ή εξαρτημένης μεταβλητή (Dependent Variable) με ένα σύνολο επεξηγηματικών μεταβλητών (Explanatory Variables) ή ανεξάρτητων μεταβλητών (Independent Variables).

Ανάλογα με τον αριθμό των επεξηγηματικών μεταβλητών με τις οποίες εξαρτάται η μεταβλητή απόκρισης και με την προϋπόθεση ότι η εξαρτημένη μεταβλητή λαμβάνει αριθμητικές τιμές η παλινδρόμηση διακρίνεται στις ακόλουθες κατηγορίες:

- Απλή γραμμική παλινδρόμηση όπου η εξαρτημένη μεταβλητή εξαρτάται μόνο από μια ανεξάρτητη μεταβλητή.
- Πολλαπλή γραμμική παλινδρόμηση όπου σε αυτή την περίπτωση η εξαρτημένη μεταβλητή εξαρτάται από περισσότερες μεταβλητές.
- Μη γραμμική παλινδρόμηση όπου η σχέση ανάμεσα στην εξαρτημένη μεταβλητή και στις ανεξάρτητες μεταβλητές είναι μη γραμμική.

Στην περίπτωση όπου η τιμή της εξαρτημένης μεταβλητής λαμβάνει ονομαστικές τιμές τότε γίνεται χρήση μοντέλων ταξινόμησης.

1.2.2 ΜΗ ΕΠΙΒΛΕΠΟΜΕΝΗ ΜΑΘΗΣΗ

Στη μη επιβλεπόμενη μάθηση (unsupervised models) είναι διαθέσιμα μόνο δεδομένα εισόδου χωρίς να καταγράφονται τα αποτελέσματα εξόδου. Έτσι στόχος είναι η αναγνώριση και η ομαδοποίηση των δεδομένων σε ομάδες χωρίς όμως να είναι εκ των προτέρων γνωστές.

1.2.2.1ΜΟΝΤΕΛΑ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ

Η συσταδοποίηση (Cluster Analysis) περιγράφει το διαχωρισμό των δεδομένων σε συστάδες με παρόμοια χαρακτηριστικά. Κατά τη διάρκεια της συσταδοποίησης πραγματοποιείται απόπειρα βελτιστοποίησης του κριτηρίου διαχωρισμού έτσι ώστε να πραγματοποιηθεί όσο το δυνατόν πιο σωστή ομαδοποίηση. Έτσι στόχος είναι η μεγαλύτερη ομοιογένεια μεταξύ των

δεδομένων μιας συστάδας και όσο το δυνατόν μεγαλύτερη διαφοροποίηση των συστάδων μεταξύ τους.

Μερικές από τις πιο γνωστές μεθόδους συσταδοποίησης είναι οι ακόλουθες:

- K-means: Ο αλγόριθμος αυτός χωρίζει το σύνολο των δεδομένων σε μη επικαλυπτόμενες συστάδες και είναι βασιζόμενο σε πρότυπα (prototype based). Πιο συγκεκριμένα, ένα αντικείμενο μπορεί να αναλογεί σε ακριβώς μια συστάδα και βρίσκεται κοντά με το πρότυπο που καθορίζει τη συστάδα αλλά διαφέρει με τα πρότυπα των άλλων συστάδων. Στην αρχή ο χρήστης επιλέγει τον επιθυμητό αριθμό των συστάδων K, δηλαδή τα αρχικά κεντροειδή (centroids). Στη συνέχεια γίνεται η ανάθεση των αντικειμένων στις συστάδες και υπολογίζονται τα νέα γεωμετρικά κέντρα κάθε συστάδας βάσει του μέσου όρου όλων των σημείων της κάθε συστάδας. Η διαδικασία επαναλαμβάνεται έως ότου η συσταδοποίηση των δεδομένων σταθεροποιηθεί.

- Συσσωρευτική ιεραρχική συσταδοποίηση (agglomerative hierarchical clustering): Στον αλγόριθμο αυτό η ομαδοποίηση είναι ιεραρχική, δηλαδή υπάρχουν ένθετες συστάδες οι οποίες διαμορφώνονται σε μορφή δέντρου (graph based). Στην αρχή όλα τα αντικείμενα θεωρούνται μεμονωμένες συστάδες οι οποίες στη συνέχεια σταδιακά συγχωνεύονται με κριτήριο την ομοιομορφία τους. Η διαδικασία προχωρά μέχρι όλα τα αντικείμενα να ομαδοποιηθούν σε μια συστάδα.

- DBSCAN: Σε αυτό τον αλγόριθμο υφίσταται ένα μοντέλο όπου ο διαχωρισμός σε συστάδες εξαρτάται από την πυκνότητα των αντικειμένων (density based). Μια συστάδα σχηματίζεται από αντικείμενα πολύ συσσωρευμένα μεταξύ τους, η οποία καλύπτεται από περιοχή με διασκορπισμένα αντικείμενα, δηλαδή περιβάλλεται από περιοχή χαμηλής πυκνότητας. Στην αρχή με τυχαίο τρόπο επιλέγεται ένα αρχικό σημείο εκκίνησης όπου στη συνέχεια δημιουργείται μια γειτονιά βάσει της ακτίνας που έχει ορισθεί θεωρώντας ως κέντρο το αντικείμενο εκκίνησης. Στην περίπτωση όπου η γειτονιά περιέχει αρκετά αντικείμενα δημιουργείται μια συστάδα, ενώ σε αντίθετη περίπτωση κρίνεται ως προσωρινός θόρυβος και επανεξετάζεται όταν ορισθεί νέο σημείο εκκίνησης. Αφού τηρείται η προϋπόθεση της ακτίνας, κάθε αντικείμενο μέσα στην συστάδα εισάγει αντικείμενα που ανήκουν σε γειτονιές του. Η διαδικασία επαναλαμβάνεται για κάθε αντικείμενο και ολοκληρώνεται όταν δεν μπορούν να προστεθούν άλλα αντικείμενα. Τέλος επιλέγεται ένα νέο τυχαίο αντικείμενο που δεν έχει ελεγχθεί προηγουμένως και γίνεται επανάληψη της διαδικασίας μέχρι όλα τα αντικείμενα να χωριστούν σε συστάδες.

1.2.2.2ΑΛΓΟΡΙΘΜΟΙ ΕΞΑΓΩΓΗΣ ΚΑΝΟΝΩΝ ΣΥΣΧΕΤΙΣΗΣ

Οι αλγόριθμοι εξαγωγής κανόνων συσχέτισης (association rules) επιδιώκουν την εύρεση κανόνων ή συσχετίσεων ανάμεσα στα δεδομένα μελετώντας την επανεμφάνιση γεγονότων στα δεδομένα αυτά. Βασική εφαρμογή των συγκεκριμένων αλγορίθμων είναι η ανάλυση της καταναλωτικής συμπεριφοράς των πελατών βάσει της ταυτόχρονης αγοράς προϊόντων ή υπηρεσιών. Οι αλγόριθμοι εξαγωγής κανόνων συσχέτισης είναι ικανοί να αναζητούν θετικές ή αρνητικές συσχετίσεις. Για παράδειγμα, η θετική συσχέτιση μπορεί να αφορά αναζήτηση προϊόντων ή υπηρεσιών που αγοράσθηκαν μαζί, ενώ αρνητική συσχέτιση μπορεί να αφορά αντικείμενα αλληλοαποκλειόμενης αγοράς.

Η διαδικασία ξεκινά με τον υπολογισμό των ποσοστών των συναλλαγών που περιέχουν δυο αντικείμενα μαζί, δηλαδή την υποστήριξη (support), καθώς και τον ορισμό της ελάχιστη αποδεκτής τιμής της για κάθε μια από τις συναλλαγές. Εκτός αυτού προσμετράται, για κάθε συναλλαγή που περιέχει ένα αντικείμενο, το ποσοστό εκείνων όπου εμφανίζεται ένα διαφορετικό αντικείμενο και αποτελούν την εμπιστοσύνη κάθε συνδυασμού (confidence). Στη συνέχεια γίνεται εντοπισμός των πιο συχνών συνόλων στοιχείων υπό την προϋπόθεση να έχουν υποστήριξη ίση ή μεγαλύτερη από την ελάχιστη αποδεκτή. Η διαδικασία ολοκληρώνεται με τη δημιουργία κανόνων συσχέτισης από τα σύνολα στοιχείων με τις περισσότερες επανεμφανίσεις υπό την προϋπόθεση τήρησης της εμπιστοσύνης του συνδυασμού. (Κύρκος, 2015)

Ο πιο γνωστός αλγόριθμος για την εξαγωγή κανόνων συσχέτισης είναι ο APRIORI ο οποίος ονομάστηκε έτσι λόγω της χρήσης προγενέστερης γνώσης (prior knowledge) των χαρακτηριστικών των συνόλων αντικειμένων με τις πιο συχνές επανεμφανίσεις.

Τα πιο συχνά μέτρα για την αξιολόγηση των κανόνων συσχέτισης είναι η υποστήριξη, η εμπιστοσύνη και το ποσοτικό μέτρο Lift. Με το μέτρο Lift (Chorianopoulos, 2016) γίνεται αποτίμηση της ικανότητας πρόβλεψη συγκρίνοντας πόσο καλός ή κακός είναι ο κανόνας που εξήχθη σε σχέση έναν άλλο τυχαίο κανόνα.

Σύμφωνα με τον Κύρκο (2015), η ύπαρξη μικρής υποστήριξης στους κανόνες μπορεί αποτελεί ένα τυχαίο γεγονός και αντίθετα η ύπαρξη μεγάλης υποστήριξης συνδυαστικά με μεγάλη εμπιστοσύνη μπορεί να μην ανταποκρίνεται σε πραγματική σχέση.

1.2.3 Η ΕΠΙΡΡΟΗ ΤΟΥ ΘΟΡΥΒΟΥ ΣΗΝ ΕΚΠΑΙΔΕΥΣΗ ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ

Συχνά στην εκπαίδευση των αλγορίθμων εμφανίζονται σφάλματα. Τα σφάλματα μπορεί να είναι είτε τυχαία, είτε συστηματικά. Το τυχαίο σφάλμα που προκύπτει (Tan et al., 2014) λέγεται θόρυβος και αφορά εσφαλμένες τιμές ή

ακραίες τιμές στα δεδομένα. Το πρόβλημα εσφαλμένης τιμής μπορεί να προκύπτει από τη λάθος καταχώρηση των δεδομένων. Απεναντίας τα σφάλματα ακραίων τιμών εμφανίζονται σε εξαιρετικές περιπτώσεις που δε συμβαδίζουν με τα συνήθη δεδομένα και δε προσφέρουν χρήσιμη πληροφορία για την εκπαίδευση των αλγορίθμων.

Στο στάδιο της προεπεξεργασίας των δεδομένων επιβάλλεται να γίνει προσπάθεια καταπολέμησης του θορύβου (Κύρκος, 2015). Μια μέθοδος αντιμετώπισης του είναι ο κατακερματισμός σε διαστήματα και η αντικατάσταση τιμών. Με αυτή τη μέθοδο τα δεδομένα διαχωρίζονται σε ίσα διαστήματα πλήθους ή συχνότητας, υπολογίζεται ο μέσος όρος του διαστήματος και στη συνέχεια γίνεται αντικατάσταση των τιμών του διαστήματος με τον μέσο όρο κάθε διαστήματος ή με τις τιμές των οριακών τιμών του κάθε διαστήματος. Προκειμένου να αντιμετωπισθεί η εμφάνιση ακραίων τιμών στα δεδομένα χρησιμοποιείται ο στατιστικός εντοπισμός υποθέσεων. Επιπλέον, μια μέθοδος για τον εντοπισμό αντικειμένων διάφορων με τα υπόλοιπα είναι η ανάλυση συστάδων. Με τον τρόπο αυτό τα όμοια αντικείμενα ομαδοποιούνται, ενώ τα αντικείμενα που δεν ταιριάζουν σε καμία ομάδα αποτελούν εξαιρέσεις. Τέλος, με την προσαρμογή των δεδομένων με χρήση μοντέλου είναι δυνατόν να γίνει πρόβλεψη των τιμών σε ένα συγκεκριμένο πεδίο με χρήση τιμών σε κάποιο άλλο πεδίο.

1.2.3.1ΜΕΡΟΛΗΨΙΑ ΠΑΡΑΛΕΙΠΟΜΕΝΩΝ ΜΕΤΑΒΛΗΤΩΝ

Κατά την εκπαίδευση των αλγορίθμων σε διαφορετικά δεδομένα προκύπτει ένα είδος σφάλματος το οποίο καλείται μεροληψία παραλειπόμενων μεταβλητών (bias). Στην στατιστική, η μεροληψία προκύπτει από τη διαφορά της μέσης τιμής των επαναλαμβανόμενων μετρήσεων από την πραγματική τιμή. Ορισμένα είδη μεροληψίας είναι (Suresh & Guttag, 2019):

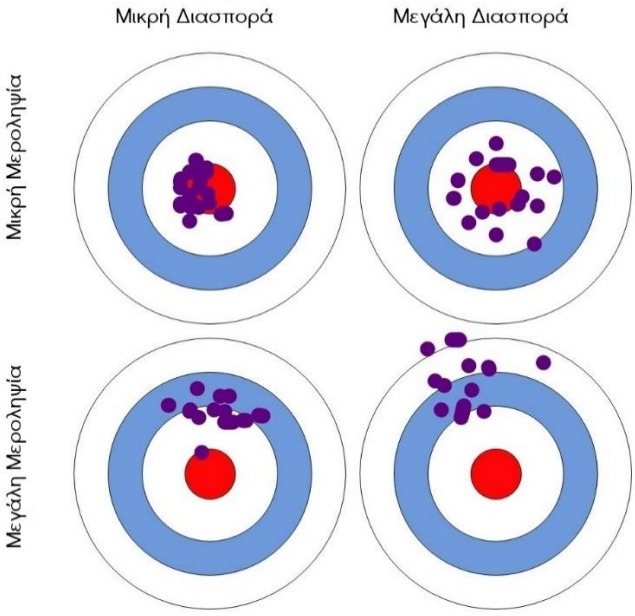
- Ιστορική μεροληψία (historical bias): Παρουσιάζεται ακόμα και στην περίπτωση τέλειας δειγματοληψίας και επιλογής χαρακτηριστικών.
- Μεροληψία αναπαράστασης (representation bias): Εμφανίζεται κατά την διαδικασία ορισμού των μεταβλητών και δειγματοληψίας από τον πληθυσμό.
- Μεροληψία μέτρησης (measurement bias): Προκύπτει από τον τρόπο μέτρηση ενός συγκεκριμένου χαρακτηριστικού.
- Μεροληψία συγκέντρωσης (aggregation bias): Εμφανίζεται όταν οι λάθος υποθέσεις για ένα πληθυσμό επηρεάζουν το αποτέλεσμα της εκπαίδευσης του αλγορίθμου.
- Μεροληψία αξιολόγησης (evaluation bias): Συμβαίνει κατά τη διαδικασία αξιολόγησης.
- Μεροληψία παράταξης (deployment bias): Γίνεται όταν υπάρχει ασυμφωνία ανάμεσα στο πρόβλημα που ήταν προγραμματισμένο να επιλύει και στο πρόβλημα που τελικά λύνει.

1.2.3.2 ΔΙΑΣΠΟΡΑ

Το σφάλμα λόγω διασποράς εμφανίζουν την απόσταση απέχουν οι προβλέψεις μεταξύ τους για ένα συγκεκριμένο πραγματικό σημείο. Για την εκπαίδευση του μοντέλου χρησιμοποιείται ένα συγκεκριμένο πλήθος δεδομένων. Η ευαισθησία των προβλέψεων αυτών, πιο συγκεκριμένα το πόσο η πρόβλεψη διαφέρει ανάμεσα σε διαφορετικά σύνολα δεδομένων εκπαίδευσης, ονομάζεται διασπορά (Hand et al., 2001).

1.2.3.3 BIAS – VARIANCE TRADE-OFF

Η μεροληψία μετράει το μέγεθος της απόκλισης των προβλέψεων από την πραγματική τιμή, ενώ η διασπορά δείχνει τη διαφορά μεταξύ των προβλέψεων. Το κέντρο του στόχου αναπαριστά το μοντέλο όπου περιγράφει τέλεια τις σωστές τιμές, ενώ η απομάκρυνση από το κέντρο του στόχου οδηγεί σε χειρότερες προβλέψεις. Στην περίπτωση όπου τα δεδομένα έχουν μικρή διασπορά και μικρή μεροληψία (Σχήμα 5, πάνω αριστερά) τα δεδομένα βρίσκονται κοντά στην πραγματική τιμή και κοντά μεταξύ τους, ενώ μικρή μεροληψία μαζί με μεγάλη διασπορά (Σχήμα 5, πάνω δεξιά) οδηγεί σε δεδομένα με κλίση το κέντρο της πραγματικής τιμής αλλά με μεγάλη απόκλιση μεταξύ τους. Από την άλλη, δεδομένα με μεγάλη μεροληψία και μικρή διασπορά (Σχήμα 5, κάτω αριστερά) οδηγούν σε αποκεντροποίηση από την πραγματική τιμή αλλά με μικρές αποκλίσεις μεταξύ τους. Τέλος, στην περίπτωση μεγάλης μεροληψίας και μεγάλης διασποράς (Σχήμα 5, κάτω δεξιά) τα δεδομένα είναι απομακρυσμένα από την πραγματική τιμή αλλά και μεταξύ τους.



Σχήμα 5 Μεροληψία - Διασπορά

Η πολυπλοκότητα εξαρτάται άμεσα με τη διασπορά και τη μεροληψία του μοντέλου που θα χρησιμοποιηθεί. Αύξηση της πολυπλοκότητας (Hastie et al.,

2017) της διαδικασίας οδηγεί σε αύξηση της διασποράς και μείωση του τετραγώνου της μεροληψίας και αντιστρόφως. Έτσι, πρέπει να γίνεται προσεκτικά η επιλογή της πολυπλοκότητας έτσι ώστε να περιορίζεται στο ελάχιστο το σφάλμα της δοκιμής. Πιο συγκεκριμένα θα πρέπει να είναι αρκετά σύνθετο προκειμένου να μπορεί να εκφράσει την δομή των δεδομένων αλλά συγχρόνως αρκετά απλό έτσι ώστε να είναι σε θέση να αποφύγει την προσαρμογή σε λανθασμένα μοτίβα.

1.2.4 ΥΠΕΡΑΠΛΟΥΣΤΕΥΣΗ & ΥΠΕΡΠΡΟΣΑΡΜΟΓΗ

Στους αλγορίθμους της μηχανικής μάθησης η βασική αιτία χαμηλής απόδοσης είναι η υπεραπλούστευση (underfitting) και η υπερπροσαρμογή (overfitting).

Υπεραπλούστευση παραπέμπει στην περίπτωση όπου το μοντέλο δεν έχει μάθει αρκετά από το σετ δεδομένα εκπαίδευσης με αποτέλεσμα να προκύπτουν μη αξιόπιστες προβλέψεις σε νέα δεδομένα. Σε αυτή την περίπτωση, το μοντέλο έχει υψηλή μεροληψία και έχει ως αποτέλεσμα να υπάρχει χαμηλή ακρίβεια στο σετ εκπαίδευσης και συνεπώς δε περιέχεται η λύση σε αυτό. Για αυτό το λόγο, ένα υπεραπλουστευμένο μοντέλο δεν είναι κατάλληλο για να πραγματοποιηθούν προβλέψεις, καθώς δεν είναι ικανό να κάνει γενίκευση σε νέα δεδομένα. Αντίθετα ένα μοντέλο που χαρακτηρίζεται από υπερπροσαρμογή έχει υψηλή διασπορά και έχει ως αποτέλεσμα να έχει μάθει πάρα πολλά από το σετ δεδομένων εκπαίδευσης, μεταξύ άλλων και τον θόρυβο, επηρεάζοντας έτσι αρνητικά την εφαρμογή του μοντέλου σε νέα δεδομένα. (Alpaydin, 2020)

Οι παραπάνω αιτίες οδηγούν σε μη αξιόπιστες προβλέψεις και είναι επιβεβλημένο να βελτιωθούν. Ορισμένες μέθοδοι για την αντιμετώπιση των μοντέλων υπερπροσαρμογής είναι η πρόωρη διακοπή της εκπαίδευσης του μοντέλου (early stopping), η έγχυση θορύβου (noise injection), η διάσπαση βάρους (weight decay) και οι προσεγγιστικοί αλγόριθμοι βελτιστοποίησης (optimized approximation algorithm). Μια επιπλέον μέθοδος για την αντιμετώπιση της υπερπροσαρμογής (Gupta, 2017) είναι η κανονικοποίηση (regularization), η οποία είναι ένα είδος παλινδρόμησης που αποθαρρύνει την εκπαίδευση ενός σύνθετου ή ευέλικτου μοντέλου. Με τη μέθοδο επιτυγχάνεται σημαντική μείωση της διασποράς του μοντέλου χωρίς όμως να αυξηθεί σημαντικά η μεροληψία.

1.2.5 ΕΦΑΡΜΟΓΕΣ ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Το φάσμα των εφαρμογών της μηχανικής μάθησης είναι πολυεπίπεδο και εφαρμόζεται σε πολλούς κλάδους της μοντέρνας καθημερινότητας. Εφαρμόζεται στην αυτοκίνηση, τη ρομποτική, την ιατρική διαγνωστική, τις τραπεζικές εφαρμογές και σε διάφορα μέρη της παραγωγής.

Στις βιομηχανικές μονάδες, με σκοπό να αυξηθεί η παραγωγικότητα του, γίνεται όλο και πιο πολύ αναγκαία η χρήση των βιομηχανικών ρομπότ. Το

γεγονός αυτό έχει προκαλέσει τη ζήτηση για πρόοδο, αλλά και στον αυτοματισμό των άκαμπτων δυνατοτήτων τους. Τρανταχτό παράδειγμα αποτελεί η Siemens, η οποία χρησιμοποιεί νευρωνικά δίκτυα για την παρακολούθηση των βιομηχανικών δραστηριοτήτων της. Με το συνδυασμό της χρήσης αισθητήρων και της ανάλυσης των μετρήσεων τους (μέσω MindShere) καταφέρνει να εντοπίσει αποτελεσματικά τις δυσλειτουργίες στις εγκαταστάσεις της και να πραγματοποιεί συντήρηση. Ακόμη, εφαρμόζει εργαλεία αυτοματοποίησης ηλεκτρονικού σχεδιασμού (Solido Design Automation) τα οποία μπορούν να μειώσουν σημαντικά το χρόνο σχεδιασμού κάποιου προϊόντος.

Μια ακόμη σημαντική εφαρμογή βιομηχανικού αυτοματισμού έχει αναπτυχθεί από την εταιρεία General Electric. Αυτή η εταιρεία έχει καταφέρει μέσω της συγκέντρωσης και ανάλυσης των δεδομένων από τους αισθητήρες, να παρακολουθεί και να ελέγχει την παραγωγή για πιθανά προβλήματα (Briliant Manufacturing Suite).

Σημαντικό ρόλο στην βιομηχανία της υγείας παίζει η μηχανική μάθηση λόγω της δυνατότητας των εφαρμογών τους να ελέγχουν μεγάλους όγκους δεδομένων και να εξάγουν με ακρίβεια και πολύ πιο γρήγορα αποτελέσματα σε σχέση με τις συμβατικές μεθόδους. Οι εφαρμογές των αλγορίθμων μηχανικής μάθησης για τη διάγνωση ασθενειών βασίζονται στις συσχετίσεις τους με διάφορα συμπτώματα των ασθενών αλλά και με διάφορες άλλες ασθένειες που πιθανόν εμφανίζονται από κοινού. (Ταβερνάκη, 2020)

Στην Ιατρική, συχνά γίνεται χρήση Κλινικών συστημάτων Υποστήριξης Αποφάσεων τα οποία εκτός των άλλων περιλαμβάνουν δραστηριότητες συνταγογράφισης φαρμάκων για την αποφυγή λαθών και κλινική παρακολούθηση των ασθενών. Για παράδειγμα, με ένα σύστημα όπως αυτό μπορεί να επιτευχθεί η ανίχνευση μοτίβων που να υποδεικνύουν την πιθανή υποτροπίαση κάποιου ασθενή.

Ακόμη, μέσω της αναγνώρισης εικόνων είναι δυνατόν να ερμηνευτούν ακτινογραφίες, αξονικές και μαγνητικές τομογραφίες και με τον τρόπο αυτό να συγκριθούν με τις αντίστοιχες εικόνες υγιών ατόμων. Στην περίπτωση σημαντικών διαφορών μεταξύ των εικόνων γίνεται αναφορά για περαιτέρω διερεύνηση από τους αρμόδιους ιατρούς.

Η χρήση αλγορίθμων μηχανικής μάθησης εφαρμόζεται επίσης και στην δημιουργία νέων φαρμάκων. Οι αλγόριθμοι αυτοί παίρνουν σαν είσοδο τα φάρμακα με δράσεις όπως αυτές της επιθυμητής και εξάγουν τις χημικές ενώσεις που είναι υπεύθυνες για τις δράσεις αυτές.

Η μηχανική μάθηση παίζει πολύ σημαντικό ρόλο στην πρόβλεψη της συμπεριφοράς των πελατών. Οι επιχειρήσεις είναι σε θέση να εντοπίζουν πιο εύκολα τους πελάτες που είναι πιθανότερο να προσελκυσθούν από μια δράση της επιχείρησης. Ακόμα, στο ηλεκτρονικό εμπόριο η εξυπηρέτηση των πελατών μπορεί να πραγματοποιείται μέσω των Chatbots η χρήση των οποίων επιτρέπει στους πελάτες να επικοινωνούν με την επιχείρηση μέσω ηλεκτρονικών μηνυμάτων οποιαδήποτε στιγμή θέλουν. Τα ερωτήματα των πελατών

απαντώνται από τα bot τα οποία προσομοιώνουν τον τρόπο με τον οποίο θα απαντούσε και ένας πραγματικός άνθρωπος. Από τις πλέον διαδεδομένες εφαρμογές chatboots χρησιμοποιούν εταιρείες όπως η Apple με την εφαρμογή Siri και η Amazon με την εφαρμογή Alexa.

Στον τραπεζικό τομέα, ένα σημαντικό πρόβλημα που επιλύεται με τη χρήση αλγορίθμων μηχανικής μάθησης είναι η ανίχνευση και η πρόληψη της απάτης. Οι τράπεζες καταγράφουν ένα μεγάλο όγκο δεδομένων των πελατών τους σε καθημερινή βάση και συνεπώς θα ήταν άκαρπο αν δε γινόταν η εκμετάλλευσή τους. Τα δεδομένα των συναλλαγών των πελατών τους υποδεικνύουν μοτίβα για την καταναλωτική τους συμπεριφορά. Επιπλέον, από την ανάλυση των δεδομένων μιας τράπεζας μπορούν να εξαχθούν συμπεράσματα σχετικά με το προφίλ των αφερέγγυων πελατών. Με αυτό τον τρόπο είναι δυνατόν να εκτιμηθεί η πιστοληπτική ικανότητα των πελατών και να αποφευχθεί κάποια συνεργασία της τράπεζας με τους εκάστοτε πελάτες.

Τέλος, όλες οι κινήσεις των επιχειρήσεων περιστρέφονται γύρω από τους πελάτες, από την προσέλκυση περνούν στην απόκτηση και τέλος στην διατήρησή τους. Οι επιχειρήσεις προκειμένου να ανταπεξέλθουν στο ανταγωνιστικό περιβάλλον αναζητούν συνεχώς λύσεις για την αύξηση της παραγωγικότητας και της αποτελεσματικότητας τους, την εύρεση των πιο επικερδών πελατών και στην προσέλκυση τους. Βασικός στόχος είναι η διατήρηση των πελατών τους που επιτυγχάνεται από την ικανοποίηση των αναγκών τους οι οποίες μπορεί να αφορούν κάποιο χαρακτηριστικό ενός προϊόντος ή κάποια υπηρεσία. Κύριος στόχος κάθε επιχείρησης είναι να ανακαλύψει πρώτη τις ανάγκες των πελατών και να τις εκτελέσει με τον πιο αποτελεσματικό τρόπο. Στο σύγχρονο αυτό περιβάλλον, η εκπλήρωση αυτή των αναγκών είναι αλληλένδετη με την εφαρμογή των αλγορίθμων μηχανικής μάθησης. (Ταβερνάκη, 2020)

2. ΥΛΟΠΟΙΗΣΗ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ

2.1 ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ

Η ανάπτυξη των σούπερ μάρκετ στις περισσότερες πυκνοκατοικημένες πόλεις αυξάνεται. Ανάλογη αύξηση ακολουθεί και ο ανταγωνισμός της αγοράς. Αυτό το σύνολο δεδομένων περιλαμβάνει ιστορικό πωλήσεων τριών μηνών για τρία καταστήματα μιας εταιρείας σούπερ μάρκετ.

	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
0	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.1415	9.1
1	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.8200	80.2200	3/8/2019	10:29	Cash	76.40	4.761905	3.8200	9.6
2	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	13:23	Credit card	324.31	4.761905	16.2155	7.4
3	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.2880	489.0480	1/27/2019	20:33	Ewallet	465.76	4.761905	23.2880	8.4
4	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37	Ewallet	604.17	4.761905	30.2085	5.3

Πίνακας 1Μέρος του συνόλου δεδομένων

2.1.1 ΠΛΗΡΟΦΟΡΙΕΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

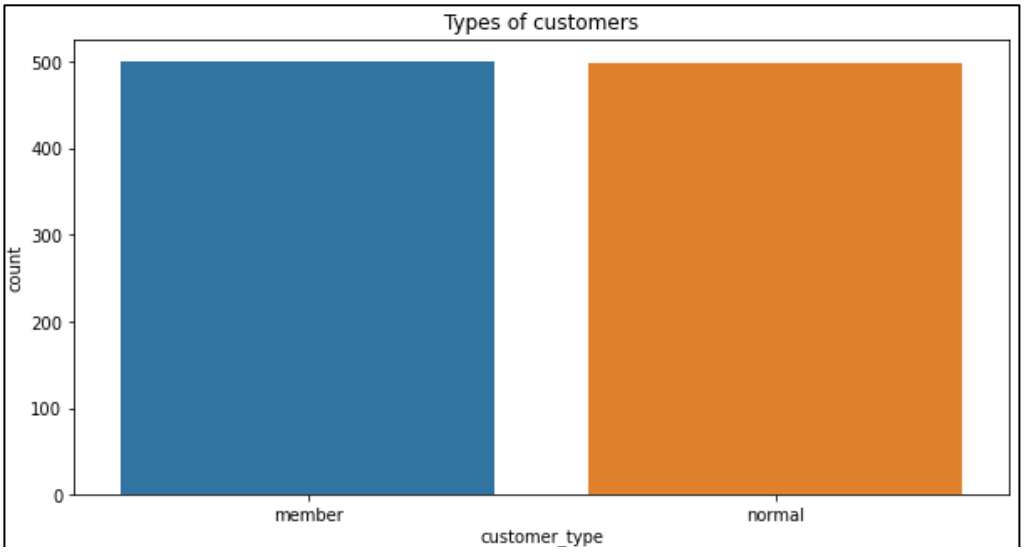
Οι μεταβλητές του συνόλου δεδομένων είναι οι παρακάτω:

- **Invoice ID**: Αυτόματα δημιουργούμενος αναγνωριστικός αριθμός παραστατικού
- **Branch**: Υποκατάστημα: υπάρχουν τρία υποκαταστήματα που χαρακτηρίζονται από το A, B και C και περιέχει ποιοτικά δεδομένα
- **City**: Τοποθεσία καταστήματος: τα καταστήματα βρίσκονται στις τρείς πόλεις Yangon, Mandalay και Naypyitaw και περιέχει ποιοτικά δεδομένα
- **Customer type**: Τύπος πελατών που διαχωρίζεται σε ‘μέλη’ για όσους χρησιμοποιούν κάρτα μέλους και ‘κανονικούς’ για όσους δεν έχουν κάρτες μέλους και περιέχει ποιοτικά δεδομένα
- **Gender**: Το φύλο πελάτη μπορεί να είναι ‘male’ ή ‘female’ και περιέχει ποιοτικά δεδομένα
- **Product line**: Γενικές ομάδες κατηγοριοποίησης προϊόντων όπως Ηλεκτρονικά αξεσουάρ, Αξεσουάρ μόδας, Τρόφιμα και ποτά, Υγεία και ομορφιά, Σπίτι και τρόπος ζωής, Αθλητισμός και ταξίδια και περιέχει ποιοτικά δεδομένα

- **Unit price**: Η τιμή μονάδας κάθε προϊόντος είναι σε $ και περιέχει ποσοτικά δεδομένα
- **Quantity**: Αριθμός προϊόντων που αγόρασε ένας πελάτηςω και περιέχει ποσοτικά δεδομένα
- **Tax 5%**: 5% φόρος για την αγορά πελατών και περιέχει ποσοτικά δεδομένα
- **Total**: Συνολική τιμή συμπεριλαμβανομένου του φόρου και περιέχει ποσοτικά δεδομένα
- **Date**: Ημερομηνία αγοράς (Το αρχείο είναι διαθέσιμο από τον Ιανουάριο 2019 έως τον Μάρτιο του 2019) και περιέχει ποιοτικά δεδομένα
- **Time**: Ώρα αγοράς (10 π.μ. έως 9 μ.μ.) και περιέχει ποιοτικά δεδομένα
- **Payment**: Πληρωμή που χρησιμοποιήθηκε από τον πελάτη για την αγορά (διατίθενται 3 μέθοδοι – Μετρητά, Πιστωτική κάρτα και Ewallet) και περιέχει ποιοτικά δεδομένα
- **Cogs**: Κόστος πωληθέντων αγαθών και περιέχει ποσοτικά δεδομένα
- **Gross margin percentage**: Ποσοστό μικτού περιθωρίου και περιέχει ποσοτικά δεδομένα
- **Gross income**: Ακαθάριστο εισόδημα και περιέχει ποσοτικά δεδομένα
- **Rating**: Βαθμολογία διαστρωμάτωσης πελατών στη συνολική εμπειρία αγορών τους (σε κλίμακα από 1 έως 10) και περιέχει ποσοτικά δεδομένα

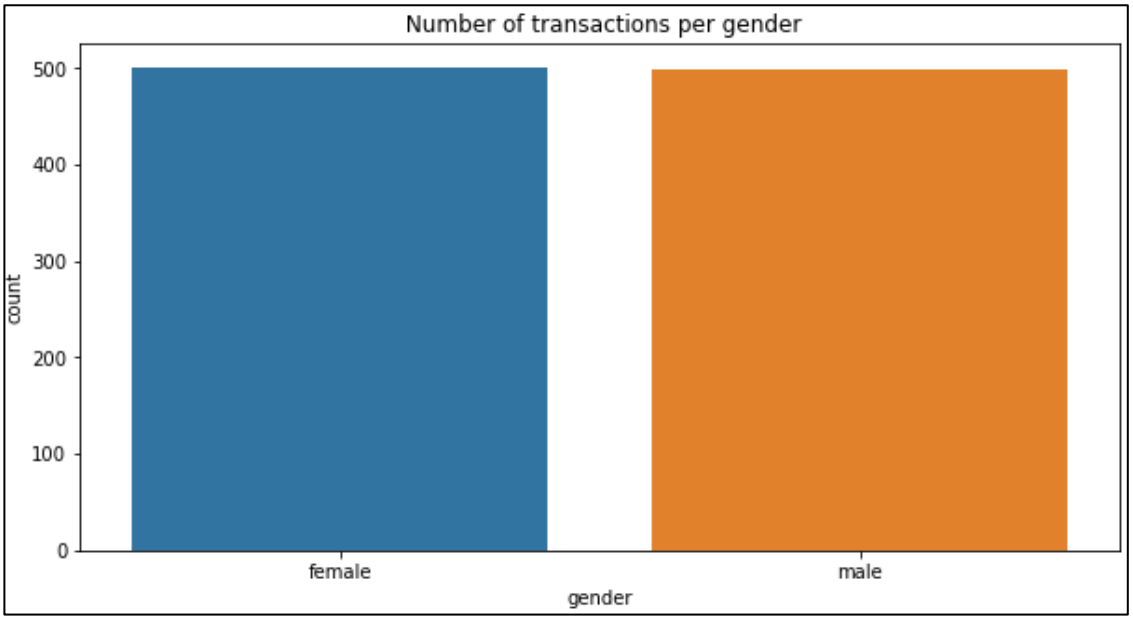
2.1.2 ΑΝΑΛΥΣΗ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

Αναλύοντας το σύνολο των δεδομένων παρατηρείται ότι ο τύπος των πελατών είναι ισόποσος άρα τα ποσοστά είναι 50% μέλη και 50% κανονικοί.



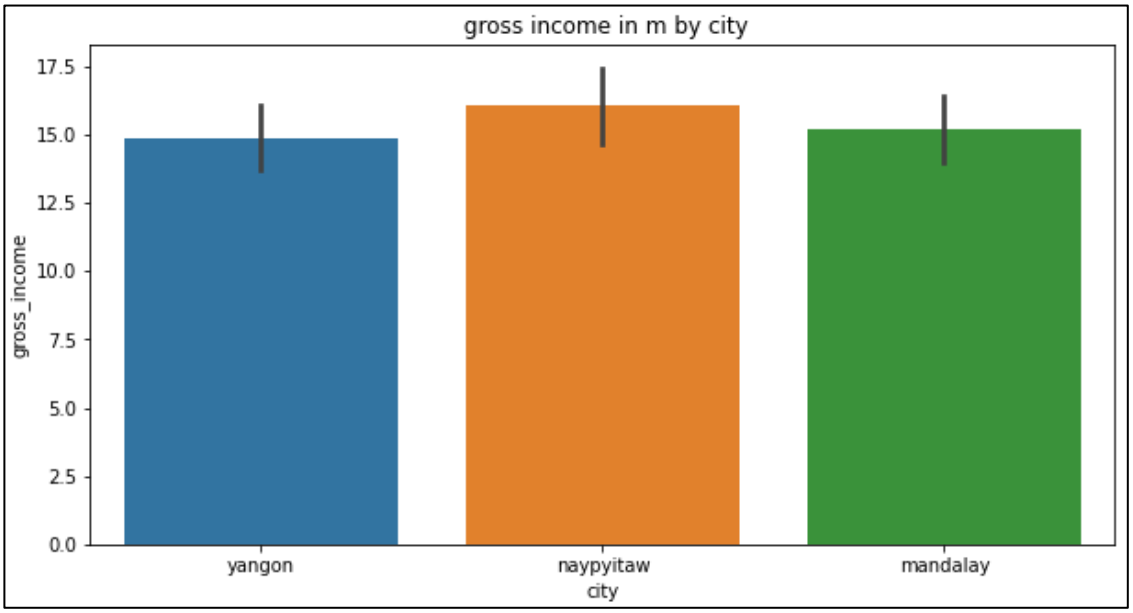
Διάγραμμα 1Τύποι καταναλωτών

Το φύλο των πελατών επίσης είναι μοιρασμένο 50% γυναίκες και 50% άνδρες.



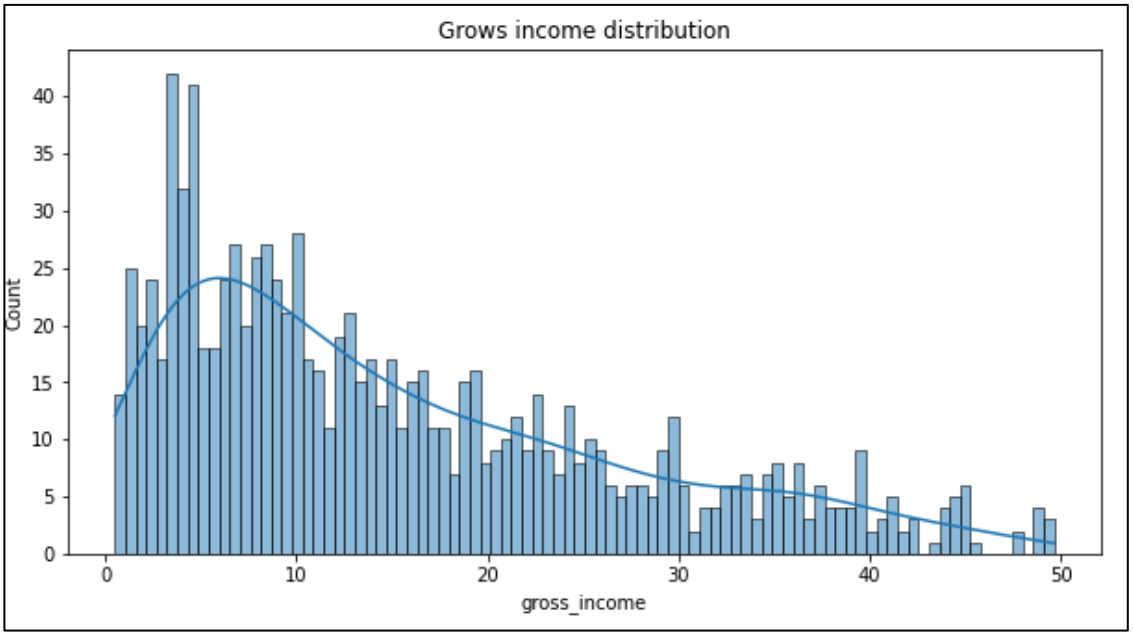
Διάγραμμα 2 Αριθμός συναλλαγών ανά φύλο

Στη συνέχεια, εξετάζοντας από ποια πόλη έρχεται το μεγαλύτερο ακαθάριστο εισόδημα παρατηρείται ότι πρώτη έρχεται η πόλη Naypyitaw με περίπου 16 εκατομμύρια δολάρια και ακολουθούν οι Yangon και η Mandalay με περίπου 15 εκατομμύρια δολάρια.



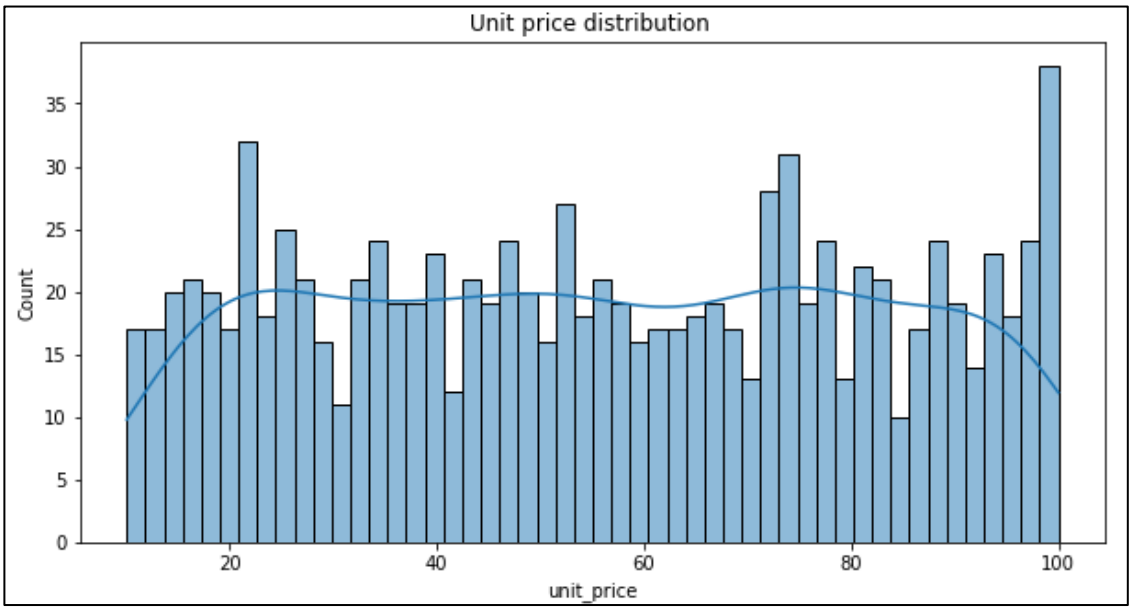
Διάγραμμα 3 Ακαθάριστο εισόδημα σε εκ. ανά πόλη

Στο ακόλουθο διάγραμμα παρουσιάζεται το μέγεθος της κατανομής του ακαθάριστου εισοδήματος.



Διάγραμμα 4 Κατανομή ακαθάριστου εισοδήματος

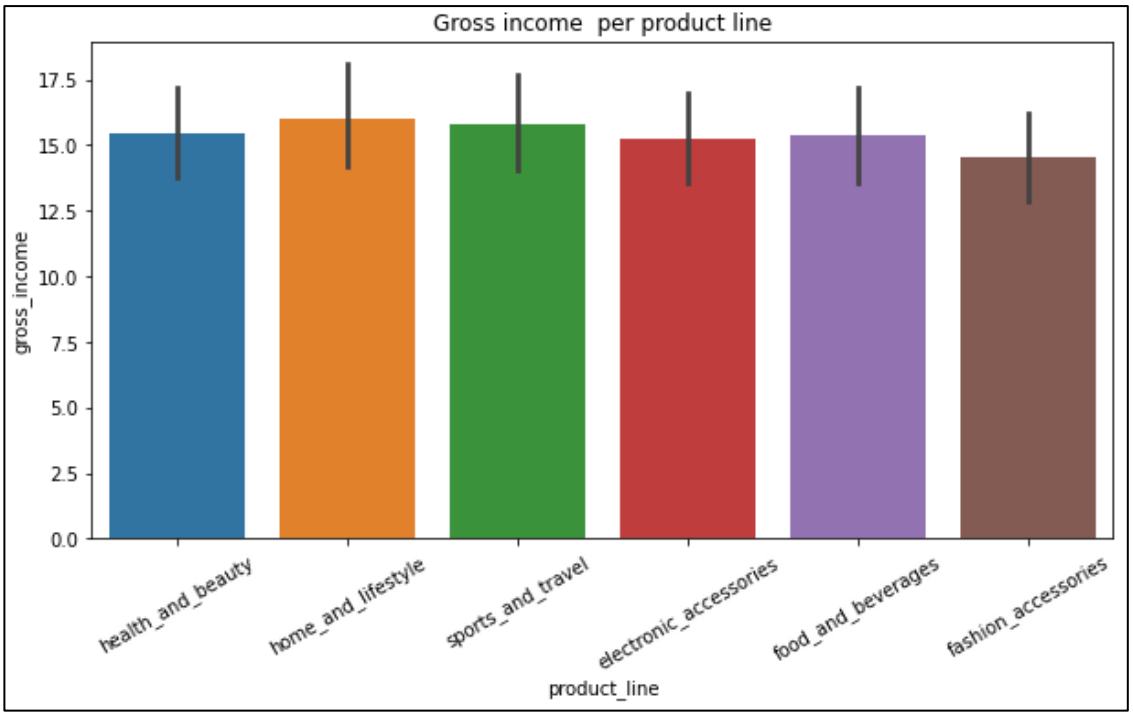
Στο διάγραμμα αυτό παρατηρείται η κατανομή ανά τιμή μονάδας



Διάγραμμα 5 Κατανομή τιμής μονάδας

Μια ακόμα ενδιαφέρουσα ανάλυση είναι το ακαθάριστο εισόδημα ανά γραμμή παραγωγής όπου πρώτη βρίσκεται η κατηγορία Home & Lifestyle,

δεύτερη η κατηγορία Sports & Travel, τρίτη η κατηγορία Food & beverages, τέταρτη η κατηγορία Health & Beauty. Τέλος, ακολουθούν η Electronic accessories και Fashion accessories.



Διάγραμμα 6 Ακαθάριστο εισόδημα ανά σειρά προϊόντων

Αθροίζοντας το ακαθάριστο εισόδημα ανά πόλη και κατάστημα παρατηρείται πως πρώτο σε πωλήσεις έρχεται το C κατάστημα της πόλη Naypytaw, ακολουθεί το A της πόλης Yangon και τελευταίο το B της πόλης Mandalay.

```
branch
c    5265.1765
a    5057.1605
b    5057.0320
Name: gross_income, dtype: float64
city
naypyitaw    5265.1765
yangon       5057.1605
mandalay     5057.0320
Name: gross_income, dtype: float64
```

Πίνακας 2 Συγκεντρωτικά στοιχεία πωλήσεων ανά κατάστημα και ανά πόλη

2.1.3 ΠΡΟΕΤΟΙΜΑΣΙΑ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΠΡΟΒΛΕΨΗ

Σε αυτό το μέρος εργασίας προκειμένου να γίνει η ανάλυση του συνόλου τον δεδομένων, είναι αναγκαία η μετατροπή των ποιοτικών δεδομένων σε ποσοτικά. Τα πεδία “gender”, ‘product_line’, 'health_and_beauty', 'customer_type', ‘city’, ‘branch’, ‘day’ όπου περιέχονται ποιοτικά στοιχεία μετατρέπονται στη συνέχεια σε ποσοτικά προκειμένου να γίνει εφαρμογή των αλγορίθμων.

Αρχικά, στο πεδίο “gender” το female αντικαθίσταται με τον αριθμό 1 και το male με τον αριθμό 2.

```
df['gender']=0
df.loc[df['gender']=="female",'gender']=1
df.loc[(df['gender']=="male"),'gender']=2
```

Στο πεδίο ‘product_line’ αντικαθίστανται οι περιγραφές σε αριθμούς με τον ακόλουθο κανόνα, όπου 'health_and_beauty' ο αριθμός 1, 'electronic_accessories' ο αριθμός 2, 'home_and_lifestyle' ο αριθμός 3, 'sports_and_travel' ο αριθμός 4, 'food_and_beverages' ο αριθμός 5 και 'fashion_accessories' ο αριθμός 6.

```
df['product_line']=0
df.loc[df['product_line'] == 'health_and_beauty', 'product_line']=1
df.loc[df['product_line'] == 'electronic_accessories', 'product_line'] = 2
df.loc[df['product_line']== 'home_and_lifestyle', 'product_line'] = 3
df.loc[df['product_line'] == 'sports_and_travel', 'product_line'] = 4
df.loc[df['product_line'] == 'food_and_beverages','product_line'] = 5
df.loc[df['product_line'] == 'fashion_accessories', 'product_line'] = 6
```

Στη συνέχεια, στο πεδίο 'customer_type' το member αντικαθίσταται με το 1 και το normal με το 2.

```
df.loc[df['customer_type'] =='member', 'customer_type']= 1
df.loc[df['customer_type'] == 'normal', 'customer_type'] = 2
```

Στο πεδίο ‘city’ ακολουθώντας τον ίδιο τρόπο σκέψης μετατρέπονται τα ονόματα των πόλεων σε αριθμούς με την ακόλουθη σειρά: Yangon ο αριθμός 1, Naypyitaw ο αριθμός 2 και Madalay ο αριθμός 3.

```
df.loc[df['city']== 'yangon', 'city'] = 1
df.loc[df['city']== 'naypyitaw', 'city'] =2
df.loc[df['city']== 'mandalay', 'city'] =3
```

Εφαρμόζοντας τον ίδιο κανόνα στο πεδίο ‘branch’ το a γίνεται 1, το b γίνεται 2 και το c γίνεται 3.

```
df.loc[df['branch'] == 'a', 'branch'] = 1
df.loc[df['branch'] == 'b', 'branch'] = 2
df.loc[df['branch'] == 'c', 'branch'] = 3
df.head()
```

Τέλος, στο πεδίο ‘day’ οι τιμές ανάλογα με την ημέρα (Δευτέρα – Κυριακή) θα αντικατασταθούν με τους αριθμούς 1 – 7 αντίστοιχα.

```
df.loc[df['day'] == 'Sunday', 'day'] = 1
df.loc[df['day'] == 'Monday', 'day'] = 2
df.loc[df['day'] == 'Tuesday', 'day'] = 3
df.loc[df['day'] == 'Wednesday', 'day'] = 4
df.loc[df['day'] == 'Thursday', 'day'] = 5
df.loc[df['day'] == 'Friday', 'day'] = 6
df.loc[df['day'] == 'Saturday', 'day'] = 7
```

3. ΕΦΑΡΜΟΓΗ ΠΡΟΒΛΕΠΤΙΚΩΝ ΜΟΝΤΕΛΩΝ

3.1 ΠΡΟΒΛΕΨΗ ΑΚΑΘΑΡΙΣΤΟΥ ΕΙΣΟΔΗΜΑΤΟΣ ΤΗΣ ΕΤΑΙΡΕΙΑΣ, ΜΕΜΟΝΩΜΕΝΑ ΚΑΙ ΑΝΑ ΠΕΡΙΟΔΟ

3.1.1 ΕΠΙΛΟΓΗ ΛΕΙΤΟΥΡΓΙΩΝ ΚΑΙ ΔΙΑΧΩΡΙΣΜΟΣ ΔΕΔΟΜΕΝΩΝ

Το πρώτο βήμα για την ανάπτυξη ενός μοντέλου μηχανικής μάθησης είναι η εκπαίδευση και η επικύρωση. Για να εκπαιδευτεί και να επικυρωθεί ένα μοντέλο, πρέπει πρώτα να χωριστεί το σύνολο δεδομένων, το οποίο περιλαμβάνει την επιλογή του ποσοστού των δεδομένων που θα χρησιμοποιηθεί για τα σύνολα εκπαίδευσης, επικύρωσης και δοκιμής.

Το σετ εκπαίδευσης είναι η υποενότητα ενός συνόλου δεδομένων από το οποίο ο αλγόριθμος μηχανικής μάθησης αποκαλύπτει ή «μαθαίνει» σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου. Στην εποπτευόμενη μηχανική μάθηση, τα δεδομένα εκπαίδευσης συμπληρώνονται με γνωστά αποτελέσματα.

Το σετ επικύρωσης είναι ένα άλλο υποσύνολο των δεδομένων στο οποίο εφαρμόζουμε τον αλγόριθμο μηχανικής μάθησης για να δούμε με πόση ακρίβεια προσδιορίζει τις σχέσεις μεταξύ των γνωστών αποτελεσμάτων για τη μεταβλητή στόχο και των άλλων χαρακτηριστικών του συνόλου δεδομένων.

Το σετ δοκιμής είναι ένα υποσύνολο που παρέχει μια τελική εκτίμηση της απόδοσης του μοντέλου μηχανικής μάθησης αφού έχει εκπαιδευτεί και επικυρωθεί. Τα σύνολα αυτά δεν πρέπει ποτέ να χρησιμοποιούνται για τη λήψη αποφάσεων σχετικά με το ποιοι αλγόριθμοι θα χρησιμοποιηθούν ή για τη βελτίωση ή τον συντονισμό των αλγορίθμων.

Διαχώρισα το αρχικό σύνολο δεδομένων 1000 δειγμάτων σε 3 μέρη:

- Ένα σετ εκπαίδευσης (train set) 800 στοιχείων
- Ένα σετ επικύρωσης (validation set) 100 στοιχείων
- Ένα σετ δοκιμής (hold-out test set) 100 στοιχείων

Στη συνέχεια, επιλέχθηκαν και εφαρμόστηκαν ορισμένοι γνωστοί προβλεπτικοί αλγόριθμοι, όπως οι παρακάτω:

- Linear Regression
- Random Forest
- Support Vector Machine
- kNN
- Gradient Boosting

3.1.1.1ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ (LINEAR REGRESSION)

Εκπαιδεύοντας τον αλγόριθμο της γραμμικής παλινδρόμησης στο σετ εκπαίδευσης έγιναν δοκιμές στο σετ δοκιμής.

3.1.1.2RANDOM FOREST

Κάνοντας εκπαίδευση στον αλγόριθμο στο σετ εκπαίδευσης και κάνοντας δοκιμές στις τιμές των παραμέτρων ‘depth’, ‘n_trees’, ‘min_samples_leaf’ στο σετ επικύρωσης το καλύτερο ήταν αυτό με τις παραμέτρους 'depth': 10, 'n_trees': 500, 'min_samples_leaf': 1.

3.1.1.3SUPPORT VECTOR MACHINE

Στη συνέχεια εκπαιδεύοντας τον αλγόριθμο στο σετ εκπαίδευσης και κάνοντας δοκιμές στις τιμές των παραμέτρων ‘kernel’, ‘C’, ‘degree’ στο σετ επικύρωσης το καλύτερο ήταν αυτό με τις παραμέτρους 'kernel': 'rbf', 'C': 10, 'degree': 3.

3.1.1.4kNN

Στον αλγόριθμο kNN έγινε εκπαίδευση στις τιμές της παραμέτρου στο σετ εκπαίδευσης και στη συνέχει δοκιμές στο σετ επικύρωσης, και το καλύτερο ήταν αυτό με την παράμετρο 'n_neighbors': 3.

3.1.1.5GRADIENT BOOSTING

Τέλος εκπαιδεύοντας τον αλγόριθμο στο σετ εκπαίδευσης και κάνοντας δοκιμές στις τιμές των παραμέτρων 'lr', 'n_trees' στο σετ επικύρωσης το καλύτερο ήταν αυτό με τις παραμέτρους 'lr': 0.01, 'n_trees': 1000.

4. ΑΞΙΟΛΟΓΗΣΗ ΜΟΝΤΕΛΩΝ

Τώρα συγκρίνουμε τα διαφορετικά μοντέλα στο σετ δοκιμής και θα επιλέξουμε τα καλύτερα:

Αλγόριθμος	Παράμετροι
Μοντέλο Γραμμικής Παλινδρόμησης	–
Τυχαίο δάσος	βάθος = 10, n_trees = 500, min_samples_leaf = 1
Support Vector Machine	πυρήνα RBF και C = 10
kNN	k = 3
Gradient Boosting	Learning_rate = 0,01 και n_estimators = 1000

Πίνακας 3Τα προβλεπτικά μοντέλα και οι καλύτεροι παράμετροι

Τα αποτελέσματα που πήραμε είναι τα παρακάτω:

Αλγόριθμος	MSE
Μοντέλο Γραμμικής Παλινδρόμησης	5.09630440136281e-29
Τυχαίο δάσος	0.0018496375702646277
Support Vector Machine	0.005406084452303256
kNN	0.0028627798148148503
Gradient Boosting	0.0014827099500266468

Πίνακας 4Τα προβλεπτικά μοντέλα και το MSE

Επομένως στη σύγκριση μεταξύ τον αλγορίθμων αυτός που είχε την μεγαλύτερη ακρίβεια ήταν ο Linear Regression.

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

[1] M. L. Agrawal, "Customer relationship management (CRM) & corporate renaissance," *Journal of Service Research,* pp. 150-171, 2003.

[2] E. Alpaydın, Introduction to Machine Learning, Fourth Edition, MIT Press, 2020.

[3] B. K. T. Stephen J. Smith, Building Data Mining Applications for CRM, New York: McGraw-Hill, 2000.

[4] F. Buttle, Customer Relationship Management: Concepts and Technologies, Burlington: Butterworth-Heinemann, 2009.

[5] R. Chalmeta, «Methodology for customer relationship management,» σε *Methodology for customer relationship management*, Casellon, Journal of Systems and Software, 2006, pp. 1015 - 1024.

[6] C. G. Tianqi Chen, «XGBoost: A Scalable Tree Boosting System,» σε *XGBoost: A Scalable Tree Boosting System*, New York, Association for Computer Machinery, 2016, pp. 785 - 794.

[7] A. Chorianopoulos, Effective CRM using Predictive Analytics, New Jersey : John Wiley & Sons Incorporated, 2015.

[8] D. J. John Fahy, ΑΡΧΕΣ ΜΑΡΚΕΤΙΝΓΚ, Αθήνα: Κριτική, 2014.

[9] R. E. S. Yoav Freund, «A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,» σε *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*, New Jersey, Journal of Computer and System Sciences, 1997, pp. 119-139.

[10] T. H. R. T. Jerome Friedman, «Additive logistic regression: a statistical view of boosting,» σε *Additive logistic regression: a statistical view of boosting*, Ohio, Institute of Mathematical Statistics, 2000, pp. 337-407.

[11] M. K. J. P. Jiawei Han, Data Mining: Concepts and Techniques, Massachusetts: Elsevier, 2012.

[12] S. L. R. X. S. David W. Hosmer Jr., Applied Logistic Regression, New Jersey: John Wiley & Sons, Inc., 2013.

[13] P. Gupta, «Towards Data Science,» 5 May 2017. [Ηλεκτρονικό]. Available: https://towardsdatascience.com/balancing-bias-and-variance-to-control-errors-inmachine-.

[14] H. M. P. S. David Hand, Principles of Data Mining, Cambridge: MIT Press, 2001.

[15] R. T. J. F. Trevor Hastie, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Berlin: Springer, 2017.

[16] V. W. J. S. G. A. PHILIP KOTLER, Principles of Marketing, London: Pearson , 2010.

[17] J. A. P. V. Kumar, Statistical Methods in Customer Relationship, New Jersey: John Willey & Sons, 2012.

[18] G. S. L. Michael J. A. Berry, Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, New Jersey: Wiley Publishing, 2011.

[19] T. M. Mitchell, Machine Learning, New York: McGraw Hill, 1997.

[20] A. R. A. T. Mehryar Mohri, Foundations of Machine Learning, Cambridge: MIT Press, 2012.

[21] A. Payne, Handbook of CRM: Achieving Excellence in Customer Management, Oxford: Butterworth-Heinemann, 2005.

[22] M. E. Porter, Competitive Strategy: Techniques for Analyzing Industries and Competitors, New York: Free Press, 1998.

[23] T. F. Foster Provost, Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking, California: O'Reilly Media, 2013.

[24] F. Rajola, Customer Relationship Management in the Financial Industry: Organizational Processes and Technology Innovation, Berlin: Springer Science+Business Media, 2013.

[25] B. D. Ripley, Pattern recognition and neural networks, Cambridge: Cambridge University Press, 1996.

[26] R. Mullin, Direct Marketing: A Step-by-Step Guide to Effective Planning and Targeting, London: Kogan Page, 2002.

[27] A. M. E. R. F. J. M. R. D. D. M. V. S. R. C. E. A. P. S. E. E. S. Cleiton dos Santos Garcia, «Process mining techniques and applications – A systematic mapping study,» σε *Process mining techniques and applications – A systematic mapping study*, Amsterdam, Elsevier BV, 2019, pp. 260-295.

[28] S. Strohmeier, «Employee relationship management — Realizing competitive advantage through information technology?,» σε *Employee relationship management — Realizing competitive advantage through information technology?*, Amsterdam, Elsevier BV, 2013, pp. 93-104.

[29] J. V. G. Harini Suresh, «Cornell University,» 28 January 2019. [Ηλεκτρονικό]. Available: https://arxiv.org/abs/1901.10002.

[30] M. S. V. K. Pang-Ning Tan, Introduction to Data Mining, London: Pearson Education, 2014.

[31] E. J. A. M. F. Terry M. Therneau, An Introduction to Recursive Partitioning Using the RPART Routines, Vienna: CRAN R-Project, 2022.

[32] A. Tiwana, The Essential Guide to Knowledge Management: E-Business and Crm Applications, New Jersey: Prentice Hall PTR, 2001.

[33] W. v. d. Aalst, Process Mining: Data Science in Action, Berlin: Springer Science+Business Media, 2016.

[34] Ε. Κύρκος και Π. Συμεωνίδης, Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων: Ανακάλυψη γνώσης για τη λήψη επιχειρηματικών αποφάσεων, Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015.

[35] Γ. Πανηγυράκης, Επικοινωνία και Δημόσιες Σχέσεις - Μελέτες Περιπτώσεων, Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015.

[36] Π. Φιτσιλής, Σύγχρονα πληροφοριακά συστήματα επιχειρήσεων, Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015.

[37] Μ. Ζ. Ταβερνάκη, Αλγόριθμοι μηχανικής μάθησης, Χανιά: Πολυτεχνείο Κρήτης, 2020.