ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εφαρμογή Αλγορίθμων Μηχανικής Μάθησης για την Πρόβλεψη Πωλήσεων

Του

ΔΗΜΗΤΡΙΟΥ ΣΑΡΑΝΤΗ ΤΟΥ ΑΘΑΝΑΣΙΟΥ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΓΕΩΡΓΙΟΣ ΛΥΜΠΕΡΟΠΟΥΛΟΣ

Η παρούσα διπλωματική εργασία υπεβλήθη για την εκπλήρωση μέρους των απαιτήσεων για την απόκτηση του Διπλώματος Μεταπτυχιακών Σπουδών «Διοίκηση Εφοδιαστικής Αλυσίδας και Logistics»

ΒΟΛΟΣ,ΝΟΕΜΒΡΙΟΣ 2023

ΕΥΧΑΡΙΣΤΙΕΣ

Για τη διεκπεραίωση της διπλωματικής μου εργασίας οφείλω να ευχαριστήσω θερμά την επιβλέπων καθηγητής μου, κύριο Λυμπερόπουλο Γεώργιο, ο οποίος με εμπιστεύτηκε και δέχτηκε να ξεκινήσει αυτή η συνεργασία με σκοπό την ολοκλήρωση της διπλωματικής μου εργασίας και των μεταπτυχιακών μου σπουδών. Η βοήθειά του ήταν πολύτιμη και την ευχαριστώ ιδιαιτέρως για την καθοδήγησή του καθ' όλη τη διάρκεια της διεκπεραίωσης της εργασίας μου.

Νιώθω την ανάγκη ακόμα να ευχαριστήσω το σύνολο των καθηγητών μου στο Μεταπτυχιακό Πρόγραμμα, για τις γνώσεις και τη βοήθεια που μου παρείχαν καθ' όλη τη διάρκειά του.

Επίσης, θα ήθελα να ευχαριστήσω μέσα από την καρδιά μου την οικογένεια μου που μου στάθηκε για άλλη μία φορά στην επιλογή μου για μάθηση και με στήριξε ψυχικά και έμπρακτα σε αυτό το δύσκολο μονοπάτι, δείχνοντας υπομονή και αγάπη.

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία έχει ως στόχο την πρόβλεψη πωλήσεων με τη μέθοδο των χρονοσειρών με τη χρήση πολλών αλγορίθμων μηχανικής μάθησης. Τα δεδομένα που θα παρουσιαστούν έχουν αποκτηθεί από το Kaggle, μια ιστοσελίδα αποκλειστικά για μηχανική μάθηση και ανάλυση δεδομένων, και έχουν να κάνουν με την πρόβλεψη πωλήσεων που αφορά μια συγκεκριμένη αλυσίδα super market. Οι αλγόριθμοι που χρησιμοποιήθηκαν στην πορεία της εργασίας είναι το δέντρο απόφασης, το τυχαίο δάσος, η μέθοδος των κ-πλησιέστερων γειτόνων, μηχανές διανύσματος υποστήριξης, οι αλγόριθμοι δέντρων ενισχυμένων με κλίση, επιπλέον ταξινομητές δέντρων, ο προσαρμοστικός ταξινομητής ενίσχυσης και η ταξινόμηση κατά Gauss. Αρχικά παρουσιάζεται η βιβλιογραφική ανασκόπηση και γίνεται επεξήγηση της βασικής ιδέας πίσω από τη μηχανική μάθηση και στη συνέχεια παρουσιάζεται η μεθοδολογία των αλγορίθμων που επιλέχθηκαν καθώς και αλγορίθμων που χρησιμοποιούνται πιο συχνά. Όλοι οι αλγόριθμοι έχουν προγραμματιστεί σε γλώσσα Python μέσω της οποίας χρησιμοποιήθηκαν και οι κατάλληλες βιβλιοθήκες για την υλοποίησή τους.

Λέξεις κλειδιά :Μηχανική μάθηση, χρονοσειρές, δέντρο απόφασης, τυχαίο δάσος, κ-πλησιέστεροι γείτονες, μηχανές διανύσματος υποστήριξης, δέντρα ενισχυμένα με κλίση, επιπλέον ταξινομητές δέντρων, προσαρμοστικός ταξινομητής ενίσχυσης, ταξινόμηση κατά Gauss.

ABSTRACT

This thesis has the objective to predict the sales through the time series method with the use of a lot of machine learning algorithms. The data have been collected by Kaggle, a site absolutely for machine learning and data science. The algorithms used through this thesis are Decision Tree, Random Forest, k-nearest neighbors, support vector machines, adaptive boost classifier, XGB Classifier, ExtraTree Classifier, Gaussian Naive Bayes. The literature review is presented to explain the basic idea of machine learning and after this is presented the methodology of the selected algorithms. The programming language used on this thesis is Python where appropriate libraries were used for the implementation.

Key words: Machine learning, time series, k-nearest neighbors, support vector machines(SVC), decision tree, random forest, Extreme Gradient Boosting, Ada Boost Classifier, Extra Tree Classifier, Gaussian Naïve Bayes

Περιεχόμενα

ΕΥΧΑΡΙΣΤΙΕΣ ii

ΠΕΡΙΛΗΨΗ iii

ABSTRACT iii

Εισαγωγή 1

ΚΕΦΑΛΑΙΟ 1. Μηχανική Μάθηση 3

1.1 Εισαγωγή 3

1.2 Είδη Μηχανικής Μάθησης 3

1.2.1 Επιβλεπόμενη (Εποπτευόμενη) Μάθηση (Supervised Learning) 4

1.2.2 Μη Επιβλεπόμενη Μάθηση (Μάθηση χωρίς επίβλεψη - UnSupervised Learning) 5

1.2.3 Ενισχυτική μάθηση 6

1.2.4 Νευρωνικά Δίκτυα & Βαθιά Μάθηση (Neural networks and deep learning) .. 7

1.3 Αλγόριθμοι ML 9

1.3.1 Δέντρο Απόφασης (Decision Tree) 9

1.3.1 Τυχαία Δάση (Random forests) 11

1.3.2 K-πλησιέστεροι Γείτονες (k nearest neighbor) 12

1.3.3 Μετρητές Απόστασης 13

1.3.4 Καθορισμός του αριθμού k 14

1.4 Μηχανές Διανυσμάτων Υποστήριξης 16

1.5 Αλγόριθμος Ακραίας Ενίσχυσης Βαθμίδας (XGBoost) 17

1.6 Extra Tree Classifier 18

1.7 AdaBoost Classifier 19

1.8 Gaussian Naive Bayes 21

1.9 Μέθοδοι αξιολόγησης μοντέλου 23
1.9.1 Πίνακας σύγχυσης (Confusion matrix) 23
1.9.2 K – Fold Cross Validation 25
1.9.3 Split 27
ΚΕΦΑΛΑΙΟ 2: ΠΕΡΙΒΑΛΛΟΝ ΥΛΟΠΟΙΗΣΗΣ ΚΑΙ ΔΕΔΟΜΕΝΑ 28
2.1 Περιβάλλον Υλοποίησης 28
2.2 Δεδομένα και Χαρακτηριστικά 29
2.3 Προ-επεξεργασία δεδομένων 30
2.4 Διαχείριση κατηγορικών μεταβλητών 31
2.5 Επιλογή χαρακτηριστικών 31
ΚΕΦΑΛΑΙΟ 3 ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ ΚΑΙ Η ΕΠΕΞΕΡΓΑΣΙΑ ΤΟΥ 33
3.1 Δεδομένα 33
3.1.1 Επεξεργασία δεδομένων 34
3.1.2 Διάσπαση των δεδομένων 38
3.2 Αξιολόγηση των μοντέλων 39
Κεφάλαιο 4 Πρακτικό Μέρος 42
4.1 Arima 42
4.2 MachineLearning 47
4.2.1 KNeighborsClassifier 49
4.2.2 SVC 50
4.2.3 Gaussian NB 54
4.2.4 ExtraTreesClassifier 55
4.2.5 Decision Tree Classifier 56
4.2.6 Random Forest Classifier 57
4.2.7 XGBClassifier 58
4.2.8 AdaBoostClassifier 60

Συμπεράσματα 62

Βιβλιογραφία 63

Εισαγωγή

Είναι βέβαιο ότι διανύουμε μια εποχή βαθιών μετασχηματισμών που σηματοδοτούνται από την ψηφιοποίηση, την τεχνολογία πληροφοριών και επικοινωνιών, τη μηχανική μάθηση, τη ρομποτική και την τεχνητή νοημοσύνη [1]. Δεν είναι λίγοι οι ερευνητές που κάνουν λόγο για μια νέα εποχή γεμάτη με αλλαγές που θα πραγματοποιηθούν στον τομέα της λήψης αποφάσεων [2]. Πιο συγκεκριμένα, η λήψη αποφάσεων ήταν ανθρώπινη ενώ η νέα αλλαγή θα είναι εμφανής στις αναδυόμενες τεχνολογίες που επιτρέπουν στους υπολογιστές να λαμβάνουν αξιόπιστα κατάλληλες αποφάσεις. Αυτή η αλλαγή ψηφιοποίησης έχει αρχίσει και θα έχει βαθιές επιπτώσεις στις προσωπικές λειτουργίες πωλήσεων και διαχείρισης πωλήσεων.

Η πρόβλεψη πωλήσεων παίζει έναν πολύ σημαντικό ρόλο στην ανάπτυξη μιας επιχείρησης. Δίνει στην εταιρεία την δυνατότητα να δει πως πρέπει να διαχειριστεί τους πόρους της, εργατικούς και χρηματικούς για την εύρυθμη λειτουργία της επίσης αποτελεί και πολύ σημαντικό παράγοντα στη διαμόρφωση του οικονομικού της σχεδιασμού, μακροπρόθεσμου και βραχυπρόθεσμου. Ακόμα βοηθά την επιχείρηση να κάνει προσεγμένες επιλογές που θα οδηγήσουν σε καλύτερη διαχείριση της εφοδιαστικής αλυσίδας, μεγαλύτερα κέρδη και καλύτερη παροχή στους πελάτες[3]. Επιπλέον με τη χρήση και την ανάλυση των δεδομένων είναι πιο εύκολο να κατανοηθεί η συμπεριφορά των καταναλωτών. Η χρήση αλγορίθμων μηχανικής μάθησης έχει εισχωρήσει τα τελευταία χρόνια σε πολλούς τομείς και ένας από αυτούς είναι ο τομέας των επιχειρήσεων. Μέσω αυτών παρέχεται η δυνατότητα εξαγωγής πληροφοριών από παλαιότερα δεδομένα με υψηλότερη ακρίβεια και ικανότητα σε σχέση με τους αλγορίθμους που μπορεί να προγραμματίσει κανείς χειροκίνητα. Η πρόοδος αυτή οφείλεται στον τεράστιο όγκο δεδομένων που υπάρχει τη σημερινή εποχή. Οι τεχνικές εξόρυξης δεδομένων είναι πολύ αποτελεσματικές στη διαμόρφωση ενός μεγάλου όγκου δεδομένων σε πληροφορίες χρήσιμες που έχουν να κάνουν με την πρόβλεψη κόστους και την πρόβλεψη πωλήσεων. Σε επίπεδο οργάνωσης η πρόβλεψη πωλήσεων αποτελεί σημαντική εισροή στην λήψη αποφάσεων σε πολλούς τομείς όπως είναι το marketing, οι πωλήσεις, η παραγωγή και τα οικονομικά.[4]

ΚΕΦΑΛΑΙΟ 1. Μηχανική Μάθηση

1.1 Εισαγωγή

Η μηχανική μάθηση έχει γίνει ένα από τα πιο σημαντικά θέματα στους αναπτυξιακούς οργανισμούς που αναζητούν καινοτόμους τρόπους για την αξιοποίηση των δεδομένων για να βοηθήσουν την επιχείρηση να αποκτήσει ένα νέο επίπεδο κατανόησης. Με τα κατάλληλα μοντέλα μηχανικής μάθησης, οι οργανισμοί έχουν τη δυνατότητα να προβλέπουν συνεχώς αλλαγές στην επιχείρηση, ώστε να είναι σε θέση να προβλέψουν καλύτερα το τι θα ακολουθήσει. Καθώς τα δεδομένα αυξάνονται συνεχώς, τα μοντέλα μηχανικής μάθησης διασφαλίζουν ότι η λύση ενημερώνεται συνεχώς και είναι εφικτή η πρόβλεψη μελλοντικών καταστάσεων.

Η μηχανική μάθηση δίνει τη δυνατότητα σε ένα σύστημα να εκπαιδεύεται από σετ δεδομένων και όχι μέσω ακολουθιακού (sequential) προγραμματισμού. Ωστόσο, πρέπει να σημειωθεί ότι η μηχανική μάθηση είναι μια αρκετά σύνθετη διαδικασία. Η μηχανική μάθηση με τη χρήση μιας σειράς αλγορίθμων που επαναλαμβανόμενα εκπαιδεύονται από τα δεδομένα για να βελτιώνουν, να περιγράφουν δεδομένα και να προβλέπουν αποτελέσματα. Καθώς οι αλγόριθμοι απορροφούν δεδομένα εκπαίδευσης, τότε είναι δυνατό να παραχθούν πιο ακριβή μοντέλα με βάση αυτά τα δεδομένα. Ένα μοντέλο μηχανικής μάθησης είναι το αποτέλεσμα που παράγεται από την εκπαίδευση του αλγόριθμου μηχανικής εκμάθησης με δεδομένα.. Για παράδειγμα, ένας προγνωστικός αλγόριθμος θα δημιουργήσει ένα προγνωστικό μοντέλο. Στη συνέχεια, όταν παρέχετε στο μοντέλο πρόβλεψης δεδομένα, θα λάβετε μια πρόβλεψη με βάση τα δεδομένα που εκπαιδεύουν το μοντέλο. Η μηχανική μάθηση είναι πλέον απαραίτητη για τη δημιουργία μοντέλων ανάλυσης.

1.2 Είδη Μηχανικής Μάθησης

Για την βελτίωση της ακρίβειας των προγνωστικών μοντέλων υπάρχουν διαφορετικά είδη τεχνικών μηχανικής μάθησης. Ανάλογα με τη φύση του επιχειρηματικού προβλήματος που αντιμετωπίζεται, υπάρχουν διαφορετικές προσεγγίσεις με βάση τον τύπο και τον όγκο των

δεδομένων. Σε αυτήν την ενότητα, παρουσιάζονται οι κυριότερες κατηγορίες της μηχανικής μάθησης.

1.2.1 Επιβλεπόμενη (Εποπτευόμενη) Μάθηση (Supervised Learning)

Η εποπτευόμενη μάθηση [5] συνήθως ξεκινά με ένα καθιερωμένο σύνολο δεδομένων και μια ορισμένη κατανόηση του τρόπου ταξινόμησης των δεδομένων αυτών. Η εποπτευόμενη μάθηση αποσκοπεί στην εύρεση προτύπων σε δεδομένα που μπορούν να εφαρμοστούν μετέπειτα σε μια διαδικασία ανάλυσης. Αυτά τα δεδομένα έχουν ετικέτα χαρακτηριστικών που καθορίζουν τον τύπο και την κατηγορία των δεδομένων.

Για παράδειγμα, θα μπορούσαν να υπάρχουν εκατομμύρια εικόνες ζώων και να περιλαμβάνουν μια εξήγηση για το τι είναι κάθε ζώο και στη συνέχεια την ανάπτυξη μιας εφαρμογής μηχανικής εκμάθησης που διακρίνει το ένα ζώο από το άλλο. Με την επισήμανση των δεδομένων για είδη ζώων, προκύπτουν εκατοντάδες κατηγορίες διαφορετικών ειδών. Επειδή υπάρχει αντιστοίχιση μεταξύ των χαρακτηριστικών και της έννοιας των δεδομένων, είναι ευκολότερη από τους χρήστες η εκπαίδευση των μοντέλων δεδομένων. Όταν η ετικέτα είναι συνεχής, θεωρείται ότι υπάρχει παλινδρόμηση. Όταν τα δεδομένα προέρχονται από ένα πεπερασμένο σύνολο τιμών, είναι γνωστό ως ταξινόμηση. Ουσιαστικά, η παλινδρόμηση που χρησιμοποιείται για την εποπτευόμενη μάθηση βοηθά στην εύρεση της συσχέτισης μεταξύ των μεταβλητών. Ένα παράδειγμα εποπτευόμενης μάθησης είναι η πρόγνωση του καιρού. Χρησιμοποιώντας ανάλυση παλινδρόμησης, η πρόγνωση του καιρού λαμβάνει υπόψη τα γνωστά ιστορικά πρότυπα καιρού και τις τρέχουσες συνθήκες για να παρέχει μια πρόβλεψη για τον καιρό.

Οι αλγόριθμοι εκπαιδεύονται χρησιμοποιώντας προ-επεξεργασμένα (pre-processed) παραδείγματα και η απόδοση των αλγορίθμων αξιολογείται με δεδομένα δοκιμής. Περιστασιακά, τα μοτίβα που προσδιορίζονται σε ένα υποσύνολο των δεδομένων δεν μπορούν να ανιχνευθούν στον μεγαλύτερο πληθυσμό δεδομένων. Εάν το μοντέλο είναι κατάλληλο να αντιπροσωπεύει μόνο τα πρότυπα που υπάρχουν στο υποσύνολο εκπαίδευσης, δημιουργείτε ένα πρόβλημα που ονομάζεται υπέρ-προσαρμογή (overfitting).

Η υπέρ-προσαρμογή σημαίνει ότι το μοντέλο έχει ρυθμιστεί με ακρίβεια για τα δεδομένα εκπαίδευσης, αλλά μπορεί να μην ισχύει για μεγάλα σύνολα άγνωστων δεδομένων. Για την αποφυγή της υπερβολικής προσαρμογής, πρέπει να γίνουν δοκιμές έναντι απρόβλεπτων ή άγνωστων δεδομένων με ετικέτα. Η χρήση απρόβλεπτων δεδομένων για το σύνολο δοκιμών μπορεί να βοηθήσει στην καλύτερη αξιολόγηση της ακρίβειας του μοντέλου στην πρόβλεψη αποτελεσμάτων. Τα εποπτευόμενα μοντέλα κατάρτισης έχουν ευρεία εφαρμογή σε μια ποικιλία επιχειρηματικών προβλημάτων, όπως ανίχνευση απάτης, ανάλυση κινδύνου κ.α..

1.2.2 Μη Επιβλεπόμενη Μάθηση (Μάθηση χωρίς επίβλεψη - UnSupervised Learning)

Η μάθηση χωρίς επίβλεψη [6] εφαρμόζεται όταν το πρόβλημα απαιτεί τεράστιο όγκο δεδομένων χωρίς ετικέτα. Για παράδειγμα, οι εφαρμογές κοινωνικών μέσων, όπως το Twitter, το Instagram, το Snapchat και ούτω καθεξής, έχουν όλες μεγάλες ποσότητες δεδομένων χωρίς ετικέτα. Η κατανόηση της σημασίας πίσω από αυτά τα δεδομένα απαιτεί αλγόριθμους που μπορούν να αρχίσουν να κατανοούν το νόημα με βάση την ικανότητα ταξινόμησης των δεδομένων με βάση τα πρότυπα ή τις ομάδες που βρίσκει. Ως εκ τούτου, η εποπτευόμενη μάθηση διεξάγει μια επαναληπτική διαδικασία ανάλυσης δεδομένων χωρίς ανθρώπινη παρέμβαση. Η μάθηση χωρίς επίβλεψη χρησιμοποιείται στην ανίχνευση ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου. Υπάρχουν πάρα πολλές μεταβλητές στα νόμιμα και ανεπιθύμητα μηνύματα ηλεκτρονικού ταχυδρομείου για έναν αναλυτή να επισημάνει ανεπιθύμητα μαζικά μηνύματα ηλεκτρονικού ταχυδρομείου. Αντ 'αυτού, εφαρμόζονται ταξινομητές μηχανικής μάθησης με βάση την ομαδοποίηση και τη συσχέτιση για τον εντοπισμό ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου.

Οι αλγόριθμοι μάθησης χωρίς επίβλεψη χωρίζουν τα δεδομένα σε σύνολα δεδομένων (συστάδες) ή ομάδες χαρακτηριστικών. Τα δεδομένα χωρίς ετικέτα δημιουργούν τις τιμές παραμέτρων και την ταξινόμηση των δεδομένων. Ουσιαστικά, αυτή η διαδικασία προσθέτει ετικέτες στα δεδομένα, έτσι ώστε να γίνονται υπό επίβλεψη.

Η χωρίς επίβλεψη μάθηση μπορεί να καθορίσει το αποτέλεσμα όταν υπάρχει τεράστια ποσότητα δεδομένων. Σε αυτήν την περίπτωση, ο χρήστης δεν γνωρίζει το πλαίσιο των δεδομένων που αναλύονται, επομένως η επισήμανση δεν είναι δυνατή σε αυτό το στάδιο. Επομένως, η χωρίς επίβλεψη μάθηση μπορεί να χρησιμοποιηθεί ως το πρώτο βήμα πριν από τη διαβίβαση των δεδομένων σε μια εποπτευόμενη διαδικασία μάθησης.

Οι αλγόριθμοι μάθησης χωρίς επίβλεψη μπορούν να βοηθήσουν τις επιχειρήσεις να κατανοήσουν μεγάλο όγκο νέων δεδομένων χωρίς ετικέτα. Ομοίως με την εποπτευόμενη μάθηση (δείτε την προηγούμενη ενότητα), αυτοί οι αλγόριθμοι αναζητούν μοτίβα στα δεδομένα. Ωστόσο, η διαφορά είναι ότι τα δεδομένα δεν είναι ήδη γνωστά. Για παράδειγμα, στην υγειονομική περίθαλψη, η συλλογή τεράστιων δεδομένων για μια συγκεκριμένη ασθένεια μπορεί να βοηθήσει τους ασκούμενους να αποκτήσουν πληροφορίες για τα πρότυπα των συμπτωμάτων και να τα συνδέσουν με τα αποτελέσματα των ασθενών. Απαιτείται πάρα πολύς χρόνος για την επισήμανση όλων των πηγών δεδομένων που σχετίζονται με μια ασθένεια όπως ο διαβήτης. Επομένως, μια προσέγγιση μάθησης χωρίς επίβλεψη μπορεί να βοηθήσει στον προσδιορισμό των αποτελεσμάτων πιο γρήγορα από μια προσέγγιση μάθησης με επίβλεψη.

1.2.3 Ενισχυτική μάθηση

Η ενισχυτική μάθηση [7] είναι ένα είδος «συμπεριφορικής» μάθησης. Περιλαμβάνει τη συνεχή ανατροφοδότηση από την επεξεργασία των δεδομένων, ώστε ο χρήστης να καθοδηγείται στο καλύτερο αποτέλεσμα. Σε σχέση με άλλους τύπους εποπτευόμενης μάθησης το σύστημα δεν μαθαίνει από το σύνολο δεδομένων δείγματος. Μάλλον, το σύστημα μαθαίνει μέσω δοκιμής και λάθους. Επομένως, μια ακολουθία επιτυχημένων αποφάσεων θα έχει ως αποτέλεσμα την "ενίσχυση" της διαδικασίας, διότι επιλύει καλύτερα το πρόβλημα.

Μία από τις πιο συνηθισμένες εφαρμογές της ενισχυτικής μάθησης είναι στη ρομποτική ή το παιχνίδι. Πάρτε το παράδειγμα της ανάγκης να εκπαιδεύσετε ένα ρομπότ για να πλοηγηθεί σε ένα σύνολο σκαλοπατιών. Το ρομπότ αλλάζει την προσέγγισή του στην

πλοήγηση στο έδαφος με βάση το αποτέλεσμα των ενεργειών του. Όταν το ρομπότ πέφτει, τα δεδομένα επαναβαθμονομούνται, έτσι ώστε τα βήματα να πλοηγούνται διαφορετικά μέχρι το ρομπότ να εκπαιδευτεί με δοκιμή και λάθος να κατανοήσει πώς να ανεβαίνει σκάλες. Με άλλα λόγια, το ρομπότ μαθαίνει με βάση μια επιτυχημένη ακολουθία ενεργειών. Ο αλγόριθμος εκμάθησης πρέπει να είναι σε θέση να ανακαλύψει μια συσχέτιση μεταξύ του στόχου της επιτυχούς ανόδου των σκαλοπατιών χωρίς πτώση και της ακολουθίας των γεγονότων που οδηγούν στο αποτέλεσμα.

Η ενισχυτική μάθηση είναι ο αλγόριθμος που χρησιμοποιείται για αυτόνομα αυτοκίνητα. Με πολλούς τρόπους, η εκπαίδευση ενός αυτόνομου αυτοκινήτου είναι απίστευτα περίπλοκη επειδή υπάρχουν τόσα πολλά πιθανά εμπόδια. Εάν όλα τα αυτοκίνητα στο δρόμο ήταν αυτόνομα, η δοκιμή και το λάθος θα ήταν πιο εύκολο να ξεπεραστούν. Ωστόσο, στον πραγματικό κόσμο, οι οδηγοί μπορεί συχνά να είναι απρόβλεπτοι. Ακόμη και με αυτό το περίπλοκο σενάριο, ο αλγόριθμος μπορεί να βελτιστοποιηθεί με την πάροδο του χρόνου για να βρει τρόπους προσαρμογής στην κατάσταση όπου οι πράξεις ανταμείβονται. Ένας από τους ευκολότερους τρόπους για να σκεφτείτε την ενίσχυση της μάθησης είναι ο τρόπος με τον οποίο ένα ζώο εκπαιδεύεται να λαμβάνει ενέργειες βάσει ανταμοιβών. Εάν ο σκύλος λατρεύει κάθε φορά που κάθεται στην εντολή, θα κάνει αυτήν την ενέργεια κάθε φορά.

1.2.4 Νευρωνικά Δίκτυα & Βαθιά Μάθηση (Neural networks and deep learning)

Η βαθιά μάθηση [7] είναι μια συγκεκριμένη μέθοδος μηχανικής μάθησης που ενσωματώνει νευρωνικά δίκτυα σε διαδοχικά στρώματα προκειμένου να μάθει από τα δεδομένα με επαναληπτικό τρόπο. Η βαθιά μάθηση είναι ιδιαίτερα χρήσιμη στην εύρεση μοτίβων από μη δομημένα δεδομένα. Η βαθιά μάθηση - σύνθετα νευρωνικά δίκτυα - έχουν σχεδιαστεί για να μιμούνται τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου, έτσι ώστε οι υπολογιστές να μπορούν να εκπαιδευτούν για την αντιμετώπιση προβλημάτων που δεν έχουν προσδιοριστεί σωστά. Ένα μέσο πεντάχρονο παιδί μπορεί εύκολα να αναγνωρίσει τη διαφορά μεταξύ του προσώπου του δασκάλου του και ενός περαστικού που διασχίζει τον δρόμο. Από την άλλη, ο υπολογιστής απαιτεί μια σειρά ενεργειών για να καταλάβει ποιος

είναι ποιος. Τα νευρωνικά δίκτυα και η βαθιά μάθηση χρησιμοποιούνται συχνά σε εφαρμογές αναγνώρισης εικόνας, ομιλίας και εφαρμογών όρασης υπολογιστή.

Ένα νευρωνικό δίκτυο αποτελείται από τρία ή περισσότερα στρώματα: ένα επίπεδο εισόδου, ένα ή πολλά κρυμμένα επίπεδα και ένα επίπεδο εξόδου. Τα δεδομένα φορτώνονται μέσω του επιπέδου εισόδου. Στη συνέχεια, τα δεδομένα τροποποιούνται στο κρυφό επίπεδο και τα επίπεδα εξόδου βασίζονται στα βάρη που εφαρμόζονται σε αυτούς τους κόμβους. Το τυπικό νευρωνικό δίκτυο μπορεί να αποτελείται από χιλιάδες ή και εκατομμύρια απλούς κόμβους επεξεργασίας που είναι πυκνά διασυνδεδεμένοι. Ο όρος βαθιά μάθηση χρησιμοποιείται όταν υπάρχουν πολλά κρυμμένα επίπεδα μέσα σε ένα νευρωνικό δίκτυο. Χρησιμοποιώντας μια επαναληπτική προσέγγιση, ένα νευρωνικό δίκτυο προσαρμόζεται συνεχώς και βγάζει συμπεράσματα μέχρι να επιτευχθεί ένα συγκεκριμένο σημείο στάσης. Τα νευρωνικά δίκτυα χρησιμοποιούνται συχνά για την αναγνώριση εικόνας και εφαρμογές όρασης υπολογιστή.

Η βαθιά μάθηση είναι μια τεχνική μηχανικής μάθησης που χρησιμοποιεί ιεραρχικά νευρωνικά δίκτυα μέσω ενός συνδυασμού αλγορίθμων χωρίς επίβλεψη και επίβλεψη. Η βαθιά μάθηση συχνά ονομάζεται υπό-κλάδος της μηχανικής μάθησης. Συνήθως, η βαθιά μάθηση μαθαίνει από δεδομένα χωρίς ετικέτα και χωρίς δομή. Η βαθιά μάθηση σε σχέση με ένα παραδοσιακό νευρωνικό δίκτυο, θα έχει πολλά περισσότερα κρυμμένα επίπεδα. Όσο πιο περίπλοκο είναι το πρόβλημα, τόσο πιο κρυμμένα επίπεδα θα υπάρχουν στο μοντέλο.

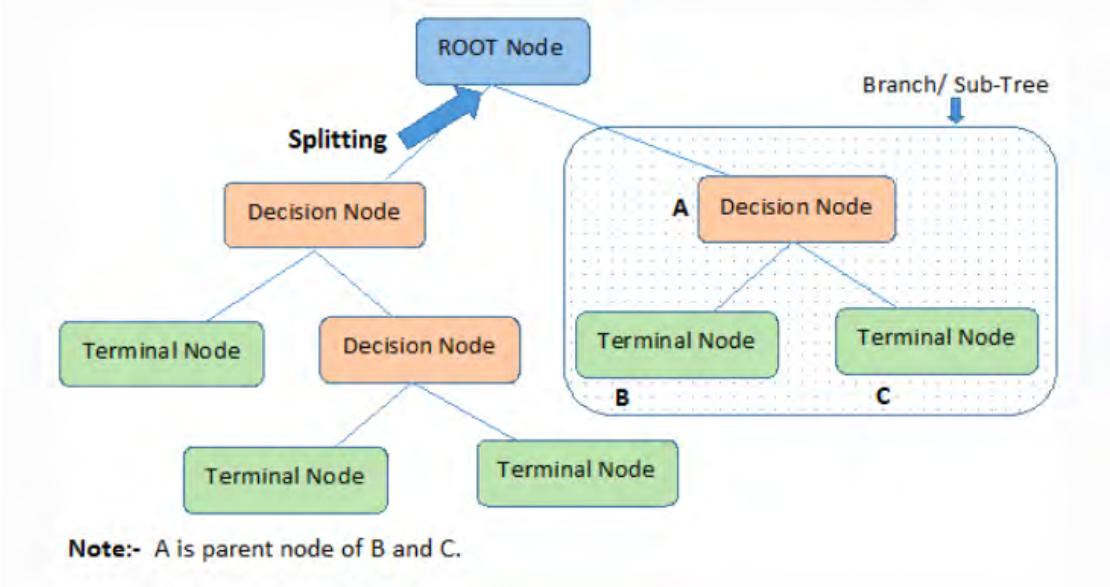
Υπάρχουν πολλοί τομείς στους οποίους η βαθιά μάθηση μπορεί να εφαρμοστεί. Για παράδειγμα, η αναγνώριση φωνής εφαρμόζεται σε αυτοκίνητα και στη διαχείριση πελατών. Στις εφαρμογές κατασκευής του Διαδικτύου των Πραγμάτων (IoT), η βαθιά μάθηση μπορεί να χρησιμοποιηθεί για να προβλέψει πότε μια μηχανή θα δυσλειτουργήσει. Οι αλγόριθμοι βαθιάς μάθησης μπορούν να βοηθήσουν το προσωπικό επιβολής του νόμου να παρακολουθεί τις κινήσεις ενός γνωστού υπόπτου.

1.3 Αλγόριθμοι ML

1.3.1 Δέντρο Απόφασης (Decision Tree)

Ο αλγόριθμος του δέντρου απόφασης ανήκει στην οικογένεια των αλγορίθμων επιβλεπόμενης μάθησης. Σε αντίθεση με άλλους αλγορίθμους μπορεί να χρησιμοποιηθεί για την επίλυση προβλημάτων τόσο ταξινόμησης όσο και παλινδρόμησης. Το πιο σημαντικό γνώρισμα του αλγορίθμου αυτού είναι η ικανότητά του να αναλύει ένα μεγάλο σύνολο δεδομένων, απλοποιώντας μια πολύπλοκη διαδικασία λήψης αποφάσεων και αποδίδοντας μια λύση που είναι ευκολότερο να ερμηνευθεί. Ο στόχος αυτού του αλγορίθμου είναι η δημιουργία ενός μοντέλου εκπαίδευσης (training model) που χρησιμοποιείται για την πρόβλεψη της κλάσης ή της αξίας της μεταβλητής στόχου [8]. Κάποια σημαντικά χαρακτηριστικά του αλγορίθμου αυτού είναι τα εξής:

- Ριζικός κόμβος, αντιπροσωπεύει ολόκληρο τον πληθυσμό ή δείγμα το οποίο αργότερα θα διασπαστεί σε δύο ή περισσότερα ομογενοποιημένα σετ.
- Διάσπαση, είναι η διαδικασία της διαίρεσης ενός κόμβου σε υπό-κόμβους.
- Κόμβος απόφασης, είναι όταν ένας υπό-κόμβος διασπάται και σε άλλους υπό-κόμβους.
- Φύλλο ή τερματικός κόμβος, είναι ο κόμβος ο οποίος δεν διασπάται περαιτέρω.
- Κλάδεμα είναι η αντίστροφη διαδικασία της διάσπασης.
- Κλάδος ή υπό-δέντρο , είναι ένα υποσύνολο ολόκληρου του δέντρου.
- Κόμβος γονέας, είναι ο κόμβος που διασπάται σε άλλους υπό-κόμβους.
- Κόμβος παιδί, είναι ο υπό-κόμβος.



Εικόνα 1 Ορολογία δέντρου απόφασης

Ο αλγόριθμος του δέντρου απόφασης για προβλήματα παλινδρόμησης παρατηρεί τα χαρακτηριστικά ενός αντικειμένου και εκπαιδεύει ένα μοντέλο στη δομή ενός δέντρου για να προβλέψει δεδομένα στο μέλλον με στόχο τη παραγωγή συνεχών αποτελεσμάτων [9]

Κάποια από τα πλεονεκτήματα του δέντρου απόφασης είναι:

- Συγκριτικά με άλλους αλγορίθμους χρειάζεται λιγότερη προσπάθεια για την προετοιμασία των δεδομένων.
- Δεν χρειάζεται κανονικοποίηση των δεδομένων.
- Αν υπάρχουν ελλιπείς τιμές στα δεδομένα δεν επηρεάζουν την διαδικασία χτισίματος του δέντρου απόφασης.
- Είναι εύκολη η κατανόηση και η εξήγηση στο ακροατήριο.

Κάποια από τα μειονεκτήματα είναι:

- Μια μικρή αλλαγή στα δεδομένα μπορεί να προκαλέσει μεγάλη αλλαγή στη δομή του δέντρου απόφασης προκαλώντας αστάθεια.

- Συχνά απαιτείται περισσότερος χρόνος για την εκπαίδευση του μοντέλου.[10]

1.3.1 Τυχαία Δάση (Random forests)

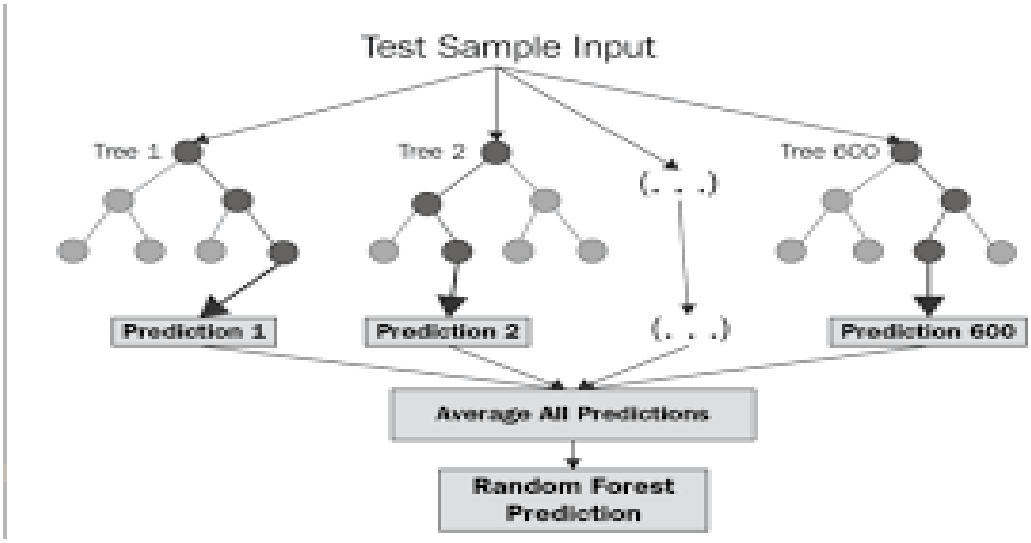
Ο αλγόριθμος του τυχαίου δάσους παρουσιάστηκε από τον L. Breiman το 2001 [11] και χρησιμοποιείται για την επίλυση προβλημάτων τόσο ταξινόμησης όσο και παλινδρόμησης. Αυτή η προσέγγιση, η οποία συνδυάζει πολλά τυχαία δέντρα απόφασης και αθροίζει τις προβλέψεις τους, έχει δείξει εξαιρετική απόδοση σε περιπτώσεις που ο αριθμός των μεταβλητών είναι πολύ μεγαλύτερος από τον αριθμό των παρατηρήσεων. Ο συνδυασμός ενός μεγάλου αριθμού δέντρων σε ένα τυχαίο δάσος μπορεί να οδηγήσει σε πιο αξιόπιστες προβλέψεις, ενώ ένα μεμονωμένο δέντρο απόφασης μπορεί να προσαρμόσει υπερβολικά τα δεδομένα [12]. Το τυχαίο δάσος χτίζει καινούργια υποσύνολα δεδομένων από το αρχικό. Ουσιαστικά επιλέγει τυχαία κάποια στοιχεία από τα αρχικά δεδομένα και δημιουργεί νέα. Κατά τη δημιουργία ενός νέου υποσυνόλου δεδομένων η επιλογή στοιχείων είναι τυχαία και επίσης το κάθε στοιχείο μπορεί να επιλεγεί παραπάνω από μια φορά στο υπό δημιουργία υποσύνολο. Η διαδικασία αυτή ονομάζεται bootstrapping και τα καινούρια υποσύνολα bootstrapped datasets. Ο αλγόριθμος αυτός είναι ιδιαίτερα δημοφιλής λόγω του γεγονότος ότι μπορεί να εφαρμοστεί σε μεγάλη ποικιλία προβλημάτων και έχει να ρυθμίσει λίγες παραμέτρους. Εκτός του ότι είναι απλός στην χρήση είναι γενικά αναγνωρισμένος για την ακρίβειά του και την ικανότητά του να αντιμετωπίζει μικρά μεγέθη δειγμάτων. Ταυτόχρονα μπορεί να αντιμετωπίζει μεγάλα συστήματα στην πραγματική ζωή [12]

Κάποια από τα πλεονεκτήματα του αλγορίθμου τυχαίου δάσους είναι:

- Η ευστροφία του, το γεγονός ότι μπορεί να χρησιμοποιηθεί για την επίλυση προβλημάτων ταξινόμησης και παλινδρόμησης.
- Χρησιμοποιείται για την επίλυση τόσο γραμμικών όσο και μη-γραμμικών.
- Είναι ένας αρκετά ακριβής αλγόριθμος.
- Το αποτέλεσμα που παράγει δεν επηρεάζεται από τις ακραίες τιμές.

Κάποια μειονεκτήματα του αλγορίθμου είναι:

- Μπορεί να είναι υπολογιστικά εκτενής για μεγάλες βάσεις δεδομένων.
- Είναι σαν ένας αλγόριθμος μαύρο κουτί, δηλαδή έχεις περιορισμένο έλεγχο πάνω σε αυτά που κάνει το μοντέλο.



Εικόνα 2 Παράδειγμα ένας τυχαίου δάσους

1.3.2 Κ-πλησιέστεροι Γείτονες (k nearest neighbor)

Ο αλγόριθμος των k-πλησιέστερων γειτόνων ή k-nearest neighbors είναι ένας αλγόριθμος που χρησιμοποιείται για την επίλυση τόσο προβλημάτων ταξινόμησης όσο και παλινδρόμησης. Παρά το γεγονός ότι είναι ένας πολύ απλός αλγόριθμος στην υλοποίησή του είναι τρομερά αποδοτικός σε πάρα πολλές περιπτώσεις [13]. Ο αλγόριθμος αυτός χρησιμοποιεί την εγγύτητα για να δημιουργήσει κλάσεις ή προβλέψεις σχετικά με την ομαδοποίηση κάθε μεμονωμένου σημείου.

Για προβλήματα ταξινόμησης μια ετικέτα τάξης αποδίδεται με βάση την ψήφο της πλειοψηφίας. Δηλαδή η ετικέτα που εμφανίζεται πιο συχνά γύρω από ένα δοσμένο σημείο χρησιμοποιείται. Ενώ αυτό τεχνικά θεωρείται ψήφος πολυφωνίας ο όρος ψήφος πλειοψηφίας χρησιμοποιείται πιο συχνά στη βιβλιογραφία. Σε προβλήματα παλινδρόμησης η ιδέα είναι παρόμοια με αυτά της ταξινόμησης αλλά σε αυτή την

περίπτωση η μέση τιμή των k πλησιέστερων γειτόνων λαμβάνεται υπόψιν για να γίνει η πρόβλεψη [14]. Η βασική διάκριση σε αυτή την περίπτωση είναι ότι η ταξινόμηση χρησιμοποιείται για διακριτές τιμές ενώ η παλινδρόμηση για συνεχείς. Είναι άξιο αναφοράς ότι ο αλγόριθμος αυτός ανήκει στην οικογένεια των τεμπέλικων αλγορίθμων, πράγμα που σημαίνει ότι αποθηκεύει μια εκπαιδευτική βάση δεδομένων περνώντας το στάδιο της εκπαίδευσης .

Η Evelyn Fix και ο Joseph Hodges ανέπτυξαν τις κεντρικές ιδέες γύρω από τον αλγόριθμο των κ πλησιέστερων γειτόνων το 1951 [15] ενώ ο Thomas Cover προχώρησε τις βασικές ιδέες με τη έρευνά του που φέρει τον τίτλο "Nearest Neighbor Pattern Classification" [16]. Παρά το γεγονός ότι δεν είναι τόσο ευρέως χρησιμοποιούμενος όσο ήταν στο παρελθόν είναι ακόμα ένας από τους πρώτους αλγορίθμους που κάποιος μαθαίνει στην επιστήμη δεδομένων λόγω της απλότητας και της ακρίβειάς του. Όμως καθώς μια βάση δεδομένων μεγαλώνει ο αλγόριθμος γίνεται όλο και πιο ανεπαρκής διακυβεύοντας την συνολική απόδοση του μοντέλου.

1.3.3 Μετρητές Απόστασης

Για να συνοψίσουμε ο στόχος του αλγορίθμου είναι να αναγνωρίσουμε τους πιο κοντινούς γείτονες ενός δεδομένου σημείου έτσι ώστε να δώσουμε μια ετικέτα σε αυτό το σημείο.

Για να καθοριστεί ποια σημεία είναι πιο κοντά στο ζητούμενο σημείο η απόσταση αυτή πρέπει να υπολογιστεί. Οι μετρητές απόστασης βοηθούν στη διαμόρφωση ορίων απόφασης [17]. Ενώ υπάρχουν πολλοί μετρητές απόστασης στο παρόν κείμενο θα παρουσιαστούν τρεις:

1. Ευκλείδεια απόσταση (p=2) αποτελεί τον πιο συχνά χρησιμοποιούμενο μετρητή απόστασης. Υπολογίζεται με βάση τον ακόλουθο τύπο:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

2. Απόσταση Manhattan (p=1) αποτελεί ένα πολύ δημοφιλή μετρητή που μετράει την απόλυτη τιμή μεταξύ δύο σημείων. Αναφέρεται επίσης ως απόσταση ταξί ή απόσταση οικοπέδου καθώς συχνά οπτικοποιείται ως ένα πλέγμα εικονογραφώντας πως μπορεί κάποιος να περιηγηθεί από μια διεύθυνση σε μια άλλη μέσω των δρόμων της πόλης. Ο τύπος είναι ο ακόλουθος:

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

3. Απόσταση Minkowski αυτό το μέτρο αποτελεί μια γενικευμένη μορφή της Ευκλείδειας απόστασης και της απόστασης Manhattan. Η παράμετρος p στον ακόλουθο τύπο επιτρέπει τη δημιουργία και άλλων μετρητών απόστασης.

   $(\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p}$

1.3.4 Καθορισμός του αριθμού k

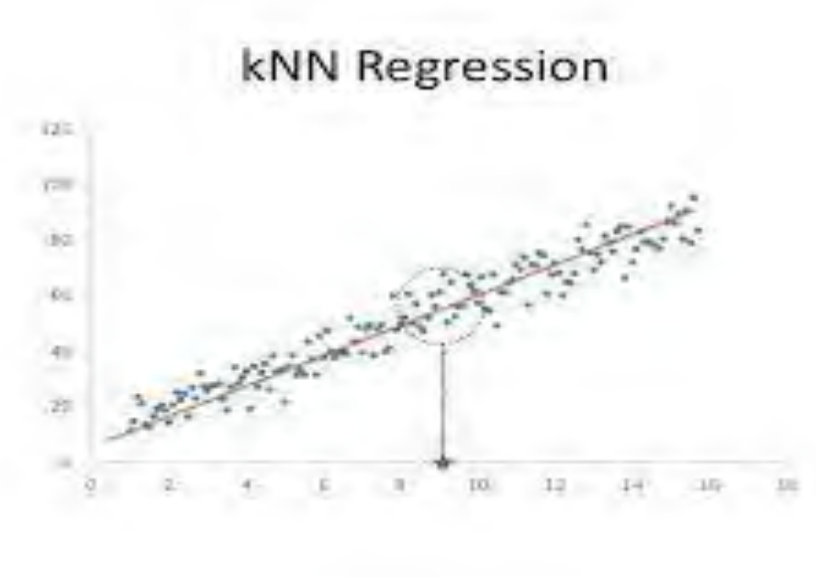
Η τιμή του αριθμού k δείχνει τον αριθμό των γειτόνων που θα ληφθεί υπόψιν για να ταξινομηθεί το ζητούμενο σημείο. Για παράδειγμα αν k=1 τότε θα ταξινομηθεί στην ίδια κλάση με τον πιο κοντινό γείτονα. Ο καθορισμός του k αποτελεί μια πράξη εξισορρόπησης καθώς διαφορετικές τιμές μπορούν να οδηγήσουν σε υπέρ-προσαρμογή ή υπό–προσαρμογή. Χαμηλότερες τιμές του μπορεί να έχουν υψηλή διακύμανση αλλά χαμηλή προκατάληψη ενώ υψηλότερες τιμές του μπορεί να οδηγήσουν σε χαμηλή διακύμανση αλλά υψηλή προκατάληψη. Η επιλογή του αριθμού k εξαρτάται σε μεγάλο βαθμό από τα δεδομένα εισόδου καθώς δεδομένα που έχουν πιο πολλά ακραία σημεία ή είναι θορυβώδη αποδίδουν καλύτερα με υψηλότερες τιμές του k .

Όπως όλοι οι αλγόριθμοι μηχανικής μάθησης έτσι και ο αλγόριθμος πλησιεστέρων γειτόνων έχει τα πλεονεκτήματα και τα μειονεκτήματά του. Κάποια θετικά είναι:

1. Ευκολία στην υλοποίηση, δεδομένης της απλότητας και της ακρίβειάς του είναι ένας από τους πρώτους αλγορίθμους που κάποιος νέος επιστήμονας δεδομένων μαθαίνει.
2. Ευκολία προσαρμογής, καθώς νέα δεδομένα εκπαίδευσης προστίθενται ο αλγόριθμος προσαρμόζεται για να λαμβάνει υπόψιν τυχόν νέα δεδομένα αφού όλα τα δεδομένα εκπαίδευσης αποθηκεύονται στη μνήμη.
3. Μικρός αριθμός παραμέτρων, για την υλοποίηση του αλγορίθμου απαιτείται μόνο ο καθαρισμός του k και ένας μετρητής απόστασης .

Κάποια αρνητικά είναι:

1. Δεν κλιμακώνεται καλά, καθώς είναι ένας αλγόριθμος τεμπέλικης μάθησης χρειάζεται πιο πολλή μνήμη σε σχέση με άλλους ταξινομητές.
2. Αδράνεια στις διαστάσεις, καθώς δεν αποδίδει καλά σε δεδομένα εισόδου πολλών διαστάσεων.
3. Επιρρεπής στην υπέρ-προσαρμογή, αυτό συμβαίνει λόγω της αδράνειας στις διαστάσεις [17].



Εικόνα 3 Αναπαράσταση αλγορίθμου k-nearest neighbors για πρόβλημα παλινδρόμησης

1.4 Μηχανές Διανυσμάτων Υποστήριξης

Οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines – SVM) αποτελεί έναν αλγόριθμο μηχανικής μάθησης που χρησιμοποιείται για την επίλυση προβλημάτων τόσο ταξινόμησης όσο και παλινδρόμησης, που προτάθηκε από τον Vapnik [18]. Η βασική ιδέα πίσω από αυτό τον αλγόριθμο είναι η κατασκευή ενός υπέρ-επιπέδου, το οποίο θα αποτελέσει την επιφάνεια απόφασης, με τέτοιο τρόπο ώστε να μεγιστοποιείται η απόσταση που διαχωρίζει τα πλησιέστερα δείγματα που ανήκουν σε διαφορετικές κλάσεις. Ο στόχος της ταξινόμησης με διανύσματα υποστήριξης είναι η σχεδίαση ενός υπολογιστικά αποδοτικού τρόπου για την εκμάθηση κατάλληλων διαχωριστικών υπέρ-επιπέδων, τα οποία βρίσκονται σε έναν υψηλών διαστάσεων διανυσματικό χώρο. Τα κατάλληλα υπέρ-επίπεδα είναι εκείνα που βελτιστοποιούν τα όρια στο σφάλμα γενίκευσης. Αυτή η διαδικασία διευκολύνει την εύρεση ενός γραμμικού διαχωρισμού ή την αποτελεσματικότερη ταξινόμηση του συνόλου δεδομένων. Η γενίκευση είναι η διαδικασία ταξινόμησης νέων δεδομένων, τα οποία δεν υπάρχουν στο σύνολο δεδομένων εκπαίδευσης.

Ο αλγόριθμος μπορεί να διαχειριστεί τόσο γραμμικά διαχωρίσιμα όσο και μη γραμμικά διαχωρίσιμα δεδομένα. Αυτό επιτυγχάνεται με πολλούς τύπους συναρτήσεων πυρήνα (kernel functions) όπως είναι του γραμμικού πυρήνα (linear kernel), του πολυωνυμικού πυρήνα (polynomial kernel), της συνάρτησης ακτινικής βάσης (radial basis function-RBF) και τον σιγμοειδή πυρήνα (sigmoid kernel). Όλες αυτές οι συναρτήσεις πυρήνα επιτρέπουν στις μηχανές διανυσμάτων υποστήριξης να καταγράφουν αποτελεσματικά σύνθετες σχέσεις και μοτίβα στα δεδομένα. Η επιλογή της συνάρτησης πυρήνα έχει να κάνει με την ακρίβεια και την πολυπλοκότητα. Η πιο δυνατή συνάρτηση, όπως είναι η συνάρτηση ακτινικής βάσης, μπορεί να πετύχει υψηλότερη ακρίβεια σε σχέση με άλλες πιο απλές, αλλά απαιτεί περισσότερα δεδομένα και χρόνο υπολογισμού για να εκπαιδεύσει τον αλγόριθμο.

Κάποια από τα πλεονεκτήματα του αλγορίθμου αυτού είναι:

1. **Είναι αποτελεσματικός σε χώρους υψηλών διαστάσεων**. Δεδομένα υψηλών διαστάσεων σημαίνει ότι ο αριθμός των χαρακτηριστικών είναι μεγαλύτερος από τον αριθμό των παρατηρήσεων.

2. **Ανθεκτικότητα σε υπέρ-προσαρμογή**. Είναι λιγότερο επιρρεπής σε υπέρ-προσαρμογή σε σχέση με άλλους αλγορίθμους όπως τα δέντρα απόφασης. Η χρήση της αρχής της μεγιστοποίησης του περιθωρίου βοηθά στην καλή γενίκευση σε μη ορατά δεδομένα.
3. **Εύστροφος / Πολύπλευρος**. Καθώς μπορεί να χρησιμοποιηθεί σε προβλήματα τόσο ταξινόμησης όσο και παλινδρόμησης.
4. **Αποτελεσματικός σε περιπτώσεις με περιορισμένα δεδομένα**.
5. **Ικανότητα επεξεργασίας μη-γραμμικών δεδομένων**. Ο αλγόριθμος μπορεί να χειριστεί μη γραμμικά διαχωρίσιμα δεδομένα με τη χρήση των συναρτήσεων πυρήνα.

Κάποια μειονεκτήματα είναι:

1. **Υπολογιστικά εκτεταμένος**. Ειδικά σε περιπτώσεις που έχουμε μεγάλο σύνολο δεδομένων. Ο χρόνος εκπαίδευσης και οι απαιτήσεις για μνήμη αυξάνονται σημαντικά.
2. **Ευαισθησία σε ρύθμιση παραμέτρων**. Έχει παραμέτρους όπως είναι η παράμετρος τακτοποίησης και η επιλογή της συνάρτησης πυρήνα. Η απόδοσή του επηρεάζεται από την ρύθμιση αυτών των παραμέτρων.
3. **Δυσκολία στην ερμηνεία περίπλοκων μοντέλων**. Μπορεί να δημιουργήσει περίπλοκα όρια απόφασης ειδικά όταν χρησιμοποιούνται μη γραμμικοί πυρήνες. Αυτή η πολυπλοκότητα κάνει πιο προκλητική την ερμηνεία του μοντέλου και την κατανόηση των υποκείμενων μοτίβων πίσω από τα δεδομένα.

1.5 Αλγόριθμος Ακραίας Ενίσχυσης Βαθμίδας (XGBoost)

Αποτελεί μια δημοφιλή και αποδοτική υλοποίηση των αλγορίθμων δέντρων ενισχυμένης βαθμίδας. Όταν χρησιμοποιείται για παλινδρόμηση πρόκειται για έναν αλγόριθμο επιβλεπόμενης μάθησης που χρησιμοποιείται για να κάνει προβλέψεις πάνω σε συνεχή αριθμητικά δεδομένα. Αποτελεί μια υλοποίηση αλγορίθμου μηχανικής μάθησης ενισχυμένης βαθμίδας που είναι ένας τύπος μεθόδου μάθησης ο οποίος συνδυάζει τις

προβλέψεις πολλαπλών πιο αδύναμων μοντέλων με στόχο τη δημιουργία ενός πιο ισχυρού και ακριβούς μοντέλου. Λειτουργεί δημιουργώντας ένα σύνολο δέντρων απόφασης, όπου το κάθε δέντρο εκπαιδεύεται για να κάνει προβλέψεις βασισμένο σε ένα υποσύνολο από τα διαθέσιμα δεδομένα. Τα δέντρα αναπτύσσονται διαδοχικά , με το κάθε δέντρο να μαθαίνει από τα λάθη του προηγούμενου. Η τελική πρόβλεψη προκύπτει από τη μέση τιμή όλων των προβλέψεων όλων των δέντρων στο σύνολο [19]

Κάποια από τα πλεονεκτήματα του αλγορίθμου αυτού είναι τα εξής:

1. Είναι ένας πολύ δυνατός αλγόριθμος με υψηλή ταχύτητα και απόδοση.
2. Μπορεί να αξιοποιήσει όλη την επεξεργαστική ισχύ των σύγχρονων υπολογιστών.
3. Είναι εφικτό να εκπαιδευτεί σε μεγάλα σύνολα δεδομένων.

Κάποια μειονεκτήματα του αλγορίθμου είναι:

1. Δεν αποδίδει τόσο καλά σε αραιά και αδόμητα δεδομένα.
2. Είναι πού ευαίσθητος σε ακραίες τιμές.
3. Ολόκληρη η μέθοδος είναι κλιμακούμενη

1.6 Extra Tree Classifier

Ο ExtraTreesClassifier [19] είναι μια μέθοδος εκμάθησης συνόλου που βασίζεται βασικά σε δέντρα αποφάσεων. Το ExtraTreesClassifier, όπως το RandomForest, τυχαιοποιεί ορισμένες αποφάσεις και υποσύνολα δεδομένων για να ελαχιστοποιήσει την υπερβολική μάθηση από τα δεδομένα και την υπέρ-προσαρμογή.

Το Extra Trees είναι σαν το Random Forest, καθώς δημιουργεί πολλά δέντρα και χωρίζει κόμβους χρησιμοποιώντας τυχαία υποσύνολα χαρακτηριστικών, αλλά με δύο βασικές διαφορές:

- δεν εκκινεί τις παρατηρήσεις (που σημαίνει ότι λαμβάνει δείγματα χωρίς αντικατάσταση) και

- οι κόμβοι χωρίζονται σε τυχαίες διαιρέσεις, όχι καλύτερα χωρίσματα. Έτσι, συνοπτικά, ExtraTrees:

    - δημιουργεί πολλά δέντρα με bootstrap = False από προεπιλογή, που σημαίνει ότι πραγματοποιεί δειγματοληψία χωρίς αντικατάσταση
    - οι κόμβοι χωρίζονται με βάση τυχαίους διαχωρισμούς μεταξύ ενός τυχαίου υποσυνόλου χαρακτηριστικών που επιλέγονται σε κάθε κόμβο

Στα Extra Trees, η τυχαιότητα δεν προέρχεται από την αρχική επιλογή δεδομένων, αλλά μάλλον από τους τυχαίους διαχωρισμούς όλων των παρατηρήσεων.

1.7 AdaBoost Classifier

Ο AdaBoost ή Adaptive Boosting [20] είναι ένας από τους ταξινομητές ενίσχυσης συνόλου που προτάθηκαν από τους Yoav Freund και Robert Schapire το 1996. Συνδυάζει πολλούς αδύναμους ταξινομητές για να αυξήσει την ακρίβεια των ταξινομητών. Το AdaBoost είναι μια επαναληπτική μέθοδος συνόλου. Ο ταξινομητής AdaBoost δημιουργεί έναν ισχυρό ταξινομητή συνδυάζοντας πολλούς ταξινομητές με κακή απόδοση, έτσι ώστε να προκύψει ένας ισχυρός ταξινομητής υψηλής ακρίβειας. Η βασική ιδέα πίσω από το Adaboost είναι ο ορισμός των βαρών των ταξινομητών και η εκπαίδευση του δείγματος δεδομένων σε κάθε επανάληψη έτσι ώστε να διασφαλίζει τις ακριβείς προβλέψεις ασυνήθιστων παρατηρήσεων.

Ο AdaBoost θα πρέπει να πληροί δύο προϋποθέσεις:

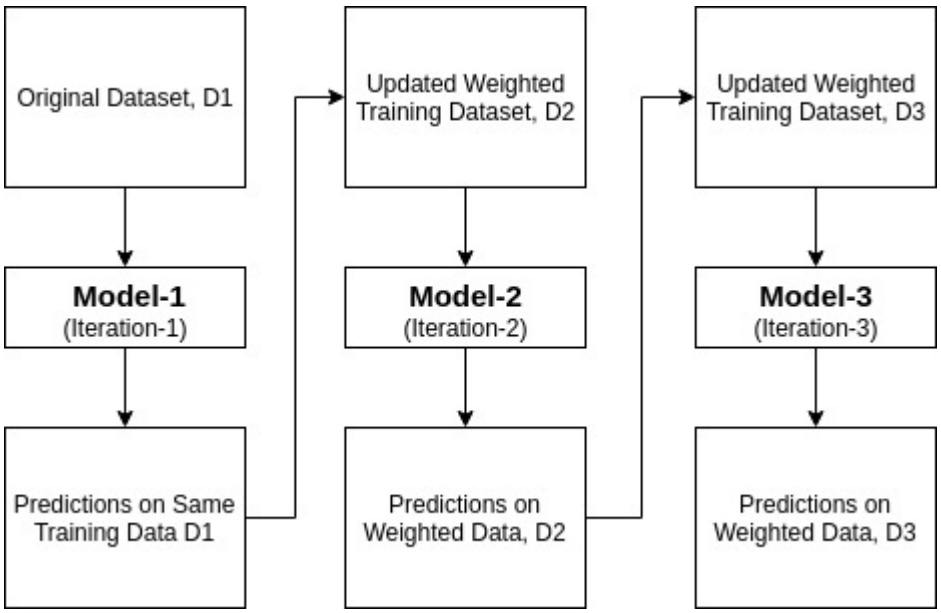
- Ο ταξινομητής θα πρέπει να εκπαιδεύεται διαδραστικά σε διάφορα σταθμισμένα παραδείγματα εκπαίδευσης.
- Σε κάθε επανάληψη, προσπαθεί να παρέχει εξαιρετική εφαρμογή σε αυτά τα παραδείγματα ελαχιστοποιώντας τα σφάλματα εκπαίδευσης.

Για να δημιουργήσετε έναν ταξινομητή AdaBoost, θεωρείστε ότι ως ταξινομητής πρώτης βάσης εκπαιδεύεται ένας αλγόριθμος Decision Tree για την πραγματοποίηση προβλέψεων για τα δεδομένα εκπαίδευσης. Τώρα, ακολουθώντας τη μεθοδολογία του AdaBoost, το

βάρος των λανθασμένων περιπτώσεων εκπαίδευσης αυξάνεται. Ο δεύτερος ταξινομητής εκπαιδεύεται και αναγνωρίζει τα ενημερωμένα βάρη και επαναλαμβάνει τη διαδικασία ξανά και ξανά. Στο τέλος κάθε πρόβλεψης μοντέλου ενισχύονται τα βάρη των εσφαλμένων ταξινομημένων περιπτώσεων, έτσι ώστε το επόμενο μοντέλο να κάνει καλύτερη δουλειά σε αυτά, και ούτω καθεξής. Το AdaBoost προσθέτει προγνωστικά στο σύνολο κάνοντάς το σταδιακά καλύτερο. Το μεγάλο μειονέκτημα αυτού του αλγορίθμου είναι ότι το μοντέλο δεν μπορεί να παραλληλιστεί αφού κάθε προγνωστικός παράγοντας μπορεί να εκπαιδευτεί μόνο αφού έχει εκπαιδευτεί και αξιολογηθεί ο προηγούμενος.

Ακολουθούν τα βήματα για την εκτέλεση του αλγόριθμου AdaBoost:

- Αρχικά, όλες οι παρατηρήσεις έχουν ίσα βάρη. Ένα μοντέλο βασίζεται σε ένα υποσύνολο δεδομένων.
- Χρησιμοποιώντας αυτό το μοντέλο, γίνονται προβλέψεις για ολόκληρο το σύνολο δεδομένων.
- Τα σφάλματα υπολογίζονται συγκρίνοντας τις προβλέψεις και τις πραγματικές τιμές. Κατά τη δημιουργία του επόμενου μοντέλου, δίνονται υψηλότερα βάρη στα σημεία δεδομένων που είχαν προβλεφθεί λανθασμένα.
- Τα βάρη μπορούν να προσδιοριστούν χρησιμοποιώντας την τιμή σφάλματος. Για παράδειγμα, όσο μεγαλύτερο είναι το σφάλμα τόσο μεγαλύτερο είναι το βάρος που αποδίδεται στην παρατήρηση.
- Αυτή η διαδικασία επαναλαμβάνεται μέχρι να μην αλλάξει η συνάρτηση σφάλματος ή να επιτευχθεί το μέγιστο όριο του αριθμού των εκτιμητών.



Εικόνα 4 AdaBoost Classifier

1.8 Gaussian Naive Bayes

Ο Naive Bayes είναι ένα βασικό αλλά αποτελεσματικός ταξινομητής στη μηχανική μάθηση που αντλεί επιρροή από το θεώρημα Bayes[21]. Το θεώρημα Bayes είναι ένας τύπος που προσφέρει μια υπό όρους πιθανότητα να συμβεί ένα γεγονός A, δεδομένου ότι ένα άλλο γεγονός B έχει προηγουμένως συμβεί. Ο μαθηματικός τύπος του είναι ο εξής:

$$P(A|B) = P(B|A)\,P(A) / P(B)$$

Όπου

Το A και το B είναι δύο γεγονότα

Το P(A|B) είναι η πιθανότητα του γεγονότος A με την προϋπόθεση ότι το γεγονός B έχει ήδη συμβεί.

P(B|A) είναι η πιθανότητα του γεγονότος B, εφόσον το γεγονός A έχει ήδη συμβεί.

Το P(A) είναι η ανεξάρτητη πιθανότητα του A

Το P(B) είναι η ανεξάρτητη πιθανότητα του B

Τώρα, το θεώρημα Bayes μπορεί να χρησιμοποιηθεί για τη δημιουργία του παρακάτω μοντέλου ταξινόμησης

$$P(y|X) = P(X|y).P(X) / P(y)$$

Όπου

$X = x_1,x_2,x_3,.. x_N$ είναι λίστα με ανεξάρτητες καταχωρήσεις

y είναι η ετικέτα της τάξης

P(y|X) είναι η πιθανότητα της ετικέτας y δεδομένου των προβλέψεων X

Η παραπάνω εξίσωση μπορεί να επεκταθεί ως εξής:

$$P(y|x1,x2,x3..xN) = P(x1|y)\ P(x2|y)\ P(x3|y)\ ...\ P(xN|y).P(y) / P(x1).P(x2).P(x3)..P(xN)$$

Χαρακτηριστικά του ταξινομητή Naive Bayes

Η μέθοδος Naive Bayes θεωρεί ότι οι προγνωστικοί παράγοντες συμβάλλουν εξίσου και ανεξάρτητα στην επιλογή της κλάσης εξόδου. Αν και η υπόθεση του μοντέλου Naive Bayes όλοι οι προγνωστικοί παράγοντες είναι ανεξάρτητοι ο ένας από τον άλλο δεν είναι εφικτή σε πραγματικές συνθήκες, αυτή η υπόθεση παράγει ένα ικανοποιητικό αποτέλεσμα στην πλειονότητα των περιπτώσεων. Ο Naive Bayes χρησιμοποιείται συχνά για την κατηγοριοποίηση κειμένου, καθώς η διάσταση των δεδομένων είναι συχνά αρκετά μεγάλη.

Τύποι ταξινομητών Naive Bayes

Οι ταξινομητές Naive Bayes ταξινομούνται σε τρεις κατηγορίες:

- Gaussian Naive Bayes - Αυτός ο ταξινομητής χρησιμοποιείται όταν οι τιμές πρόβλεψης είναι συνεχείς και αναμένεται να ακολουθήσουν μια κατανομή Gauss.
- Bernoulli Naive Bayes - Όταν οι προγνωστικοί παράγοντες είναι δυαδικού χαρακτήρα και υποτίθεται ότι ακολουθούν την κατανομή Bernoulli, αυτός ο ταξινομητής χρησιμοποιείται.
- Πολυωνυμικό Naive Bayes - Αυτός ο ταξινομητής χρησιμοποιεί μια πολυωνυμική κατανομή και χρησιμοποιείται συχνά για την επίλυση ζητημάτων που αφορούν την ταξινόμηση εγγράφων ή κειμένου.

1.9 Μέθοδοι αξιολόγησης μοντέλου

1.9.1 Πίνακας σύγχυσης (Confusion matrix)

Στον κλάδο της μηχανικής μάθησης ο όρος πίνακας σύγχυσης [22], ή διαφορετικά πίνακας σφάλματος, είναι μια συγκεκριμένη διάταξη πίνακα που δίνει τη δυνατότητα οπτικοποίησης της απόδοσης ενός αλγορίθμου, συνήθως μιας εποπτευόμενης μάθησης.

Ένας πίνακας σύγχυσης μεγέθους n x n που σχετίζεται με έναν ταξινομητή δείχνει την προβλεπόμενη και πραγματική ταξινόμηση, όπου n είναι ο αριθμός των διαφορετικών κλάσεων. Ο πίνακας 1 δείχνει έναν πίνακα σύγχυσης για n = 2, των οποίων οι καταχωρήσεις έχουν τις ακόλουθες έννοιες:

- α είναι ο αριθμός των σωστών αρνητικών προβλέψεων ·
- b είναι ο αριθμός των λανθασμένων θετικών προβλέψεων ·
- c είναι ο αριθμός των λανθασμένων αρνητικών προβλέψεων ·
- d είναι ο αριθμός των σωστών θετικών προβλέψεων.

	Αρνητική Πρόβλεψη	Θετική Πρόβλεψη
Αρνητικό	a	B

Θετικό	c	D

Πίνακας 1 Πίνακας σύγχυσης για N = 2

Η ακρίβεια πρόβλεψης και το σφάλμα ταξινόμησης μπορούν να ληφθούν από αυτόν τον πίνακα ως εξής:

$$\text{Accuracy} = (a + d) / (a + b + c + d)$$

$$\text{Error} = (b + c) / (a + b + c + d)$$

Ο ψευδοκώδικας που περιγράφεται παρακάτω δείχνει τα βήματα για την εκτέλεση επιλογής χαρακτηριστικών βάσει μήτρας σύγχυσης για πρόβλημα ταξινόμησης 2 κατηγοριών. Αυτή η μέθοδος βασικά κατασκευάζει υποσύνολα χαρακτηριστικών που:

(1) έχουν χαρακτηριστικά με καλή ατομική ταξινόμηση και

(2) έχουν χαρακτηριστικά που είναι συμπληρωματικά (δηλ. Διαφωνούν στις λανθασμένες ταξινομήσεις τους).

Σημειώστε ότι ο αλγόριθμος μπορεί να οδηγήσει σε διερεύνηση περισσότερων υποσυνόλων χαρακτηριστικών, δηλαδή μπορεί να επιλεγεί το υποσύνολο που αποδίδει υψηλότερη ακρίβεια ταξινόμησης. Επίσης, ο αλγόριθμος δεν λαμβάνει υπόψη την πιθανότητα δύο ξεχωριστά χαμηλότερης κατάταξης χαρακτηριστικά να συνδυαστούν σε ένα υποσύνολο υψηλής ακρίβειας ταξινόμησης λόγω της υψηλής συμπληρωματικότητάς τους.

Algorithm 1 Pseudocode for Confusion Matrix-based Attribute Selection Algorithm

Require: 2-class data of n attributes
Require: classification technique
Require: k - number of member subset
Ensure: Output k-attribute subset as tuple $S_k = (A_1, A_2, ..., A_k)$
Compute classifier C_i based on feature A_i, $i = 1..n$
Obtain: $Accuracy(C_i)$ and $ConfMatrix(C_i)$
Rank A_i according to $Accuracy(C_i) \Rightarrow R_A$
for $i = 1$ ... n **do**
  Compute disagreement based on $ConfMatrix(C_i)$ as: $D_i = \frac{|b-c|}{max\{b,c\}}$
end for
Rank A_i according to $D_i \Rightarrow R_D$
Select top k (according to R_A) attributes having large D (according to R_D) but in different classes: $\Rightarrow S_k = (A_1, A_2, ..., A_k)$

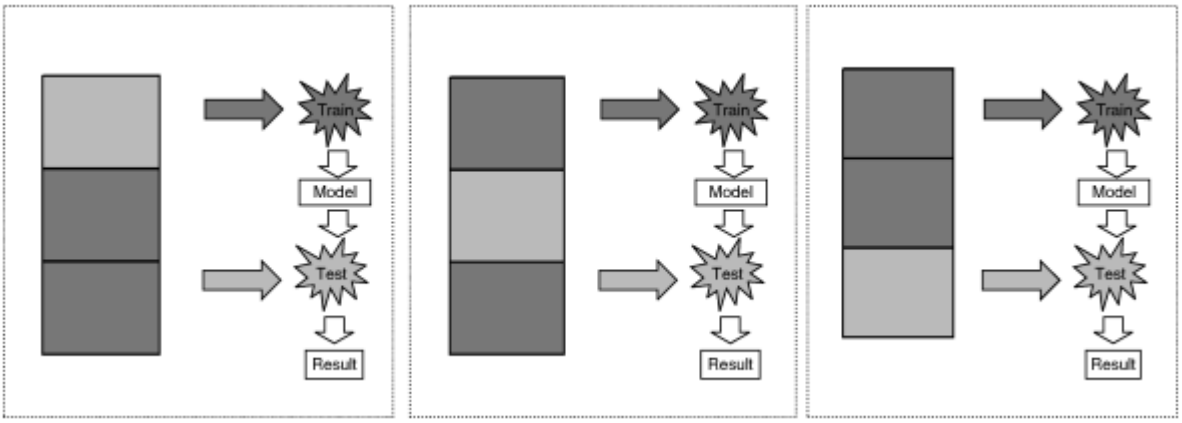
1.9.2 K – Fold Cross Validation

Η διασταυρούμενη επικύρωση (cross – validation) [23] είναι μια μέθοδος δειγματοληψίας δεδομένων για την αξιολόγηση της ικανότητας γενίκευσης των προγνωστικών μοντέλων και την πρόληψη της υπερβολικής προσαρμογής. Η διασταυρούμενη επικύρωση ανήκει στην οικογένεια των μεθόδων του Μόντε Κάρλο.

Πιο αναλυτικά, η διασταυρούμενη επικύρωση είναι μια στατιστική μέθοδος αξιολόγησης και σύγκρισης μαθησιακών αλγορίθμων διαιρώντας δεδομένα σε δύο τμήματα: το ένα χρησιμοποιείται για την εκμάθηση ή εκπαίδευση ενός μοντέλου και το άλλο που χρησιμοποιείται για την επικύρωση του μοντέλου. Σε τυπική διασταυρούμενη επικύρωση,

τα σύνολα εκπαίδευσης και επικύρωσης πρέπει να διασταυρώνονται σε διαδοχικούς γύρους, έτσι ώστε κάθε σημείο δεδομένων να έχει πιθανότητα επικύρωσης. Η βασική μορφή διασταυρούμενης επικύρωσης είναι η πολλαπλή διασταυρούμενη επικύρωση. Άλλες μορφές διασταυρούμενης επικύρωσης είναι ειδικές περιπτώσεις πολλαπλής διασταυρούμενης επικύρωσης ή περιλαμβάνουν επαναλαμβανόμενους γύρους διασταυρωμένης επικύρωσης k-fold.

Στην πολλαπλή διασταυρούμενη (k-fold) επικύρωση, τα δεδομένα χωρίζονται πρώτα σε τμήματα ή πτυχώσεις με ίσο (ή σχεδόν ίσο) μέγεθος. Στη συνέχεια, k επαναλήψεις εκπαίδευσης και επικύρωσης πραγματοποιούνται έτσι ώστε σε κάθε επανάληψη μια διαφορετική πτυχή των δεδομένων διατηρείται για επικύρωση ενώ οι υπόλοιπες πτυχές k x 1 χρησιμοποιούνται για μάθηση. Η παραπάνω εικόνα δείχνει ένα παράδειγμα με k = 3. Το πιο σκούρο τμήμα των δεδομένων χρησιμοποιείται για εκπαίδευση, ενώ το πιο ανοιχτόχρωμο τμήμα για επικύρωση. Στην εξόρυξη δεδομένων και τη μηχανική μάθηση, η 10-πλάσια διασταυρούμενη επικύρωση (k = 10) είναι η πιο κοινή.



Εικόνα 5 Procedure of three-fold cross-validation

Τα δεδομένα είναι συνήθως στρωματοποιημένα πριν χωριστούν σε k πτυχώσεις. Η διαστρωμάτωση είναι η διαδικασία αναδιάταξης των δεδομένων για να διασφαλιστεί ότι κάθε πτυχή είναι ένας καλός εκπρόσωπος του συνόλου. Για παράδειγμα, σε ένα πρόβλημα δυαδικής ταξινόμησης όπου κάθε κλάση περιλαμβάνει το 50% των δεδομένων, είναι καλύτερο να τακτοποιήσουμε τα δεδομένα έτσι ώστε σε κάθε πτυχή, κάθε κλάση να περιλαμβάνει περίπου τις μισές περιπτώσεις.

1.9.3 Split

Το πρόβλημα της κατάλληλης διάσπασης δεδομένων [24] μπορεί να αντιμετωπιστεί ως στατιστικό πρόβλημα δειγματοληψίας. Ως εκ τούτου, διάφορες κλασικές στατιστικές τεχνικές δειγματοληψίας μπορούν να χρησιμοποιηθούν για τον διαχωρισμό των δεδομένων. Αυτές οι μέθοδοι δειγματοληψίας μπορούν να χωριστούν στις ακόλουθες κατηγορίες με βάση τις αρχές, τους στόχους και την αλγοριθμική και υπολογιστική πολυπλοκότητά τους:

- Απλή τυχαία δειγματοληψία (SRS),
- Μέθοδοι δοκιμών και σφαλμάτων (Trial-and-error methods)
- Συστηματική δειγματοληψία (Systematic Sampling)
- Δειγματοληψία ευκολίας (Convenience sampling)
- CADEX, DUPLEX
- Στρωματοποιημένη δειγματοληψία (Stratified sampling)

Ορισμένες από τις μεθόδους είναι απλές και ευρέως χρησιμοποιούμενες, αν και υποφέρουν από μεγάλη διακύμανση της απόδοσης του μοντέλου (π.χ. SRS και μέθοδοι δοκιμής και σφάλματος). Άλλες τεχνικές είναι ντετερμινιστικές και αποτελεσματικές, ωστόσο περιορίζονται σε συγκεκριμένους τύπους συνόλων δεδομένων (π.χ. ευκολία και συστηματική δειγματοληψία). Οι πιο εξελιγμένες μέθοδοι (π.χ. CADEX, DUPLEX και στρωματοποιημένη δειγματοληψία) εκμεταλλεύονται τη δομή των δεδομένων για να επιτύχουν σίγουρα αποτελέσματα σε βάρος του υψηλότερου υπολογιστικού κόστους.

ΚΕΦΑΛΑΙΟ 2: ΠΕΡΙΒΑΛΛΟΝ ΥΛΟΠΟΙΗΣΗΣ ΚΑΙ ΔΕΔΟΜΕΝΑ

2.1 Περιβάλλον Υλοποίησης

Η παρούσα διπλωματική εργασία καθώς και οι αλγόριθμοι μηχανικής μάθησης έγιναν με τη γλώσσα προγραμματισμού Python στην έκδοση 3.8.6. Το υπολογιστικό περιβάλλον είναι το Jupyter Notebook, το οποίο επιτρέπει το συνδυασμό κειμένου, ενεργού κώδικα Python και οπτικοποιήσεων. Επίσης οι βιβλιοθήκες που χρησιμοποιήθηκαν είναι οι εξής:

- Pandas, μια βιβλιοθήκη που χρησιμοποιείται για επεξεργασία και αναδιαμόρφωση μεγάλου όγκου δεδομένων σε πολλούς τομείς όπως είναι τα χρηματοοικονομικά, η ανάλυση ιστού, η στατιστική και άλλα. Είναι μια open-source βιβλιοθήκη και χρησιμοποιείται ευρέως για την ανάλυση δεδομένων.
- NumPy, η οποία αποτελεί μια βιβλιοθήκη της Python που προσφέρει υποστήριξη στους μεγάλους πολυδιάστατους πίνακες με τη χρήση μαθηματικών συναρτήσεων υψηλού επιπέδου. Αποτελεί μια από τις πιο κοινά χρησιμοποιούμενες βιβλιοθήκες για επιστημονικούς υπολογισμούς. Πιο συγκεκριμένα δίνει τη δυνατότητα στον χρήστη να κάνει τα εξής: πολλαπλασιασμό διανυσμάτων, πολλαπλασιασμό πινάκων, πολλαπλασιασμό πίνακα με διάνυσμα, λειτουργίες σε διανύσματα πίνακες (πρόσθεση, αφαίρεση, πολλαπλασιασμό και διαίρεση με έναν αριθμό).
- Matplotlib, που χρησιμοποιείται για σχεδιασμό στην Python. Μέσω αυτής μπορεί κάποιος να κατασκευάσει γραφήματα, γραμμικά διαγράμματα, ιστογράμματα και άλλα. Στη μηχανικά μάθηση χρησιμοποιείται για την αναπαράσταση της βάσης δεδομένων με στόχο την ευκολότερη κατανόησή της και την εύρεση συσχετισμών μεταξύ των χαρακτηριστικών.
- Seaborn, είναι μια βιβλιοθήκη για τη δημιουργία στατιστικών γραφημάτων στην Python και ενσωματώνεται στενά με τις δομές δεδομένων της Pandas. Βοηθά στην εξερεύνηση και την κατανόηση των δεδομένων. Βασίζεται στην matplotlib και παρέχει μια υψηλού επιπέδου επιφάνεια για τη δημιουργία ελκυστικών και πληροφοριακών στατιστικών γραφημάτων.

- Scikit Learn, είναι η βασική βιβλιοθήκη που χρησιμοποιείται για την κλήση διάφορων τεχνικών μηχανικής μάθησης.
- Math, που εμπεριέχει βασικές συναρτήσεις μαθηματικών όπως είναι η ρίζα, ο λογάριθμος, οι τριγωνομετρικές συναρτήσεις και άλλα.

2.2 Δεδομένα και Χαρακτηριστικά

Τα δεδομένα αποτελούν βασικό συστατικό για τη σχεδίαση και την υλοποίηση μεθόδων επιβλεπόμενης μάθησης. Συλλέγονται με διάφορες τεχνικές από το περιβάλλον του προβλήματος, είτε παρέχονται ως γνώση από τους ειδικούς στο εκάστοτε επιστημονικό πεδίο. Στη συγκεκριμένη εργασία τα δεδομένα προέρχονται από το Kaggle, έναν ιστότοπο επιστημόνων ανάλυσης δεδομένων και μηχανικής μάθησης. Κάθε ξεχωριστό, ανεξάρτητο δείγμα που περιέχεται στο σετ δεδομένων και το οποίο χρησιμοποιείται ως είσοδος στα συστήματα μηχανικής μάθησης, χαρακτηρίζεται από τις τιμές του σε ένα συγκεκριμένο σύνολο χαρακτηριστικών (features, attributes).Τα χαρακτηριστικά αυτά μπορούν να θεωρηθούν ως μεταβλητές που λαμβάνουν συγκεκριμένες τιμές. Η τιμή ενός χαρακτηριστικού για ένα συγκεκριμένο δείγμα αποτελεί μια μέτρηση της πληροφορίας την οποία αναπαριστά το χαρακτηριστικό αυτό. Οι τιμές των χαρακτηριστικών μπορεί να είναι είτε αριθμητικές (numerical) είτε κατηγορικές (categorical), οπότε προκύπτουν οι εξής κατηγορίες χαρακτηριστικών:

- Αριθμητικά χαρακτηριστικά αντιπροσωπεύουν συνήθως δεδομένα με τη μορφή βαθμωτών τιμών που απεικονίζουν παρατηρήσεις, καταγραφές ή μετρήσεις. Με τον όρο αριθμητικά δεδομένα χαρακτηρίζονται τα συνεχή και όχι διακριτά δεδομένα. Οι ακέραιοι και οι δεκαδικοί αριθμοί είναι οι πιο συχνές χρησιμοποιούμενες μορφές αριθμητικών δεδομένων.
- Κατηγορικά χαρακτηριστικά , τα οποία παίρνουν τιμές από ένα πεπερασμένο και προκαθορισμένο σύνολο. Τα χαρακτηριστικά αυτά μπορούν να διασπαστούν και σε δύο περαιτέρω κατηγορίες σύμφωνα με τον τρόπο που μετρούν την πληροφορία την οποία αναπαριστούν ως εξής:

1. Ονομαστικά χαρακτηριστικά (nominal attributes) που λαμβάνουν ονομασίες οι οποίες λειτουργούν ως απλές ετικέτες ή ονόματα. Μια ειδική κατηγορία των ονομαστικών χαρακτηριστικών αποτελούν τα δυαδικά χαρακτηριστικά που έχουν μόνο δύο κατηγορίες όπως για παράδειγμα Ναι/Όχι, Αληθές/Ψευδές.
2. Τακτικά χαρακτηριστικά (ordinal attributes) που διαθέτουν συσχετίσεις μικρότερου/μεγαλύτερου. Παρόλο όμως που υπάρχει σχέση διάταξης δεν μπορεί να οριστεί η έννοια της απόστασης για αυτή τη κατηγορία τιμών. Παράδειγμα τέτοιου χαρακτηριστικού είναι το εισόδημα ενός ανθρώπου που μπορεί να έχει τις τιμές ΄χαμηλό΄,΄μεσαίο΄,΄υψηλό΄. Και δεν μας ενδιαφέρει αν παρουσιαστεί με την αντίστροφη διάταξη της αναγραφόμενης.

2.3 Προ-επεξεργασία δεδομένων

Αυτή η διαδικασία έχει ως στόχο την αντιμετώπιση προβλημάτων που προκύπτουν από τη συλλογή των δεδομένων, όπως είναι οι ακραίες τιμές και τιμές που λείπουν από κάποια δείγματα.

Στην περίπτωση των ελλιπών δεδομένων υπάρχουν πολλοί τρόποι για την αντιμετώπιση του προβλήματος όπως είναι:

- Η διαγραφή των δειγμάτων στα οποία λείπουν τιμές για ένα ή περισσότερα χαρακτηριστικά. Αν στα περισσότερα δείγματα λείπουν οι τιμές ενός συγκεκριμένου χαρακτηριστικού τότε η καλύτερη λύση είναι η αφαίρεση του συγκεκριμένου χαρακτηριστικού.
- Μια άλλη τεχνική για την αντιμετώπιση των ελλιπών τιμών είναι ο υπολογισμός των τιμών αυτών. Αυτό μπορεί να γίνει στην περίπτωση που η ελλιπής τιμή είναι αριθμητική ώστε να μπορεί να αντικατασταθεί με τη μέση τιμή ή την διάμεσο των υπολοίπων μετρήσεων του συγκεκριμένου χαρακτηριστικού. Αυτή η μέθοδος μπορεί να αποτρέψει την απώλεια δεδομένων σε σύγκριση με την προηγούμενη τεχνική.

- Στην περίπτωση που η ελλιπής τιμή είναι κατηγορική μπορεί να αντικατασταθεί με την κατηγορία που εμφανίζεται πιο συχνά σε κάθε χαρακτηριστικό.

2.4 Διαχείριση κατηγορικών μεταβλητών

Η ύπαρξη κατηγορικών μεταβλητών είναι σχεδόν σίγουρη σε κάθε σύνολο δεδομένων. Η πρόκληση είναι να καθοριστεί ο τρόπος χρήσης αυτών των δεδομένων στην ανάλυση και στην εκπαίδευση των αλγορίθμων μηχανικής μάθησης. Ο αναλυτής πρέπει να βρει μια λύση ώστε να μετατρέψει αυτά τα κατηγορικά στοιχεία σε αριθμητικά για την περαιτέρω επεξεργασία τους. Αξίζει να αναφερθεί ότι αν ο αλγόριθμος απαιτεί την μετατροπή αυτή και δεν πραγματοποιηθεί τότε κατά τη διάρκεια της εκτέλεσής του θα προκύψει κάποιο σφάλμα.

Κάποιες από τις πιο κοινές προσεγγίσεις της αντιμετώπισης αυτού του προβλήματος είναι:

- Η διαγραφή όλων των στηλών με κατηγορικά χαρακτηριστικά. Αυτή η τακτική καλό θα ήταν να χρησιμοποιείται μόνο όταν οι στήλες που διαγράφονται δεν περιέχουν χρήσιμη πληροφορία.
- Μία άλλη τακτική είναι η μετατροπή των κατηγορικών δεδομένων σε αριθμητικά . Αυτό μπορεί να γίνει με τη χρήση της εντολής LabelEncoder που υπάρχει στην Python.

2.5 Επιλογή χαρακτηριστικών

Κατά τη δημιουργία ενός μοντέλου είναι απαραίτητο να επιλεγούν κάποια χαρακτηριστικά που θα τροφοδοτήσουν τον αλγόριθμο για να υπάρξει κάποιο αποτέλεσμα. Συχνά σε ένα σύνολο δεδομένων υπάρχουν πολλά χαρακτηριστικά. Από αυτά κάποια δεν παρέχουν αρκετή πληροφορία και δεν είναι χρήσιμα για την κατασκευή του μοντέλου μηχανικής μάθησης που θα κάνει τις προβλέψεις. Χρησιμοποιώντας αυτά τα χαρακτηριστικά η ποιότητα των προβλέψεων θα χειροτερεύσει. Η επιλογή των κατάλληλων χαρακτηριστικών είναι πολύ σημαντική πριν από τη δημιουργία του μοντέλου. Ένας εύκολος και γρήγορος τρόπος για να εξεταστούν οι πιθανές συσχετίσεις και εξαρτήσεις μεταξύ όλων των

χαρακτηριστικών και της μεταβλητής εξόδου είναι ο πίνακας συσχετίσεων (correlation matrix). Ο πίνακας αυτός περιέχει ένα συντελεστή για κάθε σχέση μεταξύ δύο χαρακτηριστικών. Ο συντελεστής αυτός παίρνει τιμές μεταξύ -1 και 1. Αν η τιμή είναι αρνητική τότε τα χαρακτηριστικά αυτά αντιστρόφως ανάλογα, ενώ αν είναι θετική τότε τα χαρακτηριστικά είναι ανάλογα.

Με τη χρήση του πίνακα αυτού γίνεται εφικτή η επιλογή των χαρακτηριστικών που δίνουν τη μέγιστη πληροφορία στο μοντέλο, ενώ τα υπόλοιπα όχι τόσο χρήσιμα χαρακτηριστικά αγνοούνται.

ΚΕΦΑΛΑΙΟ 3 ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ ΚΑΙ Η ΕΠΕΞΕΡΓΑΣΙΑ ΤΟΥ

3.1 Δεδομένα

Στην παρούσα διπλωματική εργασία τα δεδομένα που χρησιμοποιήθηκαν για τις προβλέψεις έχουν αποκτηθεί δωρεάν από το Kaggle, μια ιστοσελίδα αποκλειστικά για ανάλυση δεδομένων και για μηχανική μάθηση, στην οποία γίνονται μάλιστα και διαγωνισμοί πάνω σε αυτά τα αντικείμενα. Το όνομα του αρχείου που επιλέχθηκε είναι ' Supermarket Sales' και αναφέρεται στο ιστορικό πωλήσεων σε τρία Supermarket τα οποία βρίσκονται σε μια χώρα της νοτιοανατολικής Ασίας, την Myanmar, και βρίσκονται σε τρεις διαφορετικές πόλεις της χώρας, την Yangon, την Mandalay και την Naypyitaw.

Περιέχει αρχεία πωλήσεων σε σούπερ μάρκετ (Supercentres), πληροφορίες προϊόντων από υποκαταστήματα σε τρεις πόλεις, συγκεκριμένα, Yangon, Naypyitaw και Mandalay. Λόγω ζητημάτων απορρήτου, δημιουργήθηκε η προβολή της αρχικής βάσης δεδομένων, εξαιρουμένων όλων των ευαίσθητων πληροφοριών όπως το όνομα, το επώνυμο και οι γραμμές διεύθυνσης. Λήφθηκε ένα δείγμα για μια τρίμηνη περίοδο πωλήσεων από τον Ιανουάριο έως τον Μάρτιο του 2019, που περιλαμβάνει 17 πεδία (χαρακτηριστικά) και 1001 σειρές δεδομένων (πριν από οποιαδήποτε προ-επεξεργασία).

Ποσοτικά δεδομένα

Αυτά τα δεδομένα έχουν νόημα ως μέτρηση, όπως το ύψος, το βάρος, το IQ ή η αρτηριακή πίεση ενός ατόμου. ή στο επιλεγμένο σύνολο δεδομένων αυτού του έργου, η ποσότητα αγοράς, ο φόρος (Chambers, 2017).

• Τα διακριτά δεδομένα αντιπροσωπεύουν στοιχεία που μπορούν να μετρηθούν παίρνουν πιθανές τιμές που μπορούν να απαριθμηθούν. Η λίστα των πιθανών τιμών μπορεί να είναι σταθερή (ονομάζεται επίσης πεπερασμένη).

• Τα συνεχή δεδομένα αντιπροσωπεύουν μετρήσεις. Οι πιθανές τιμές τους δεν μπορούν να μετρηθούν και μπορούν να περιγραφούν μόνο χρησιμοποιώντας διαστήματα στην πραγματική αριθμητική γραμμή.

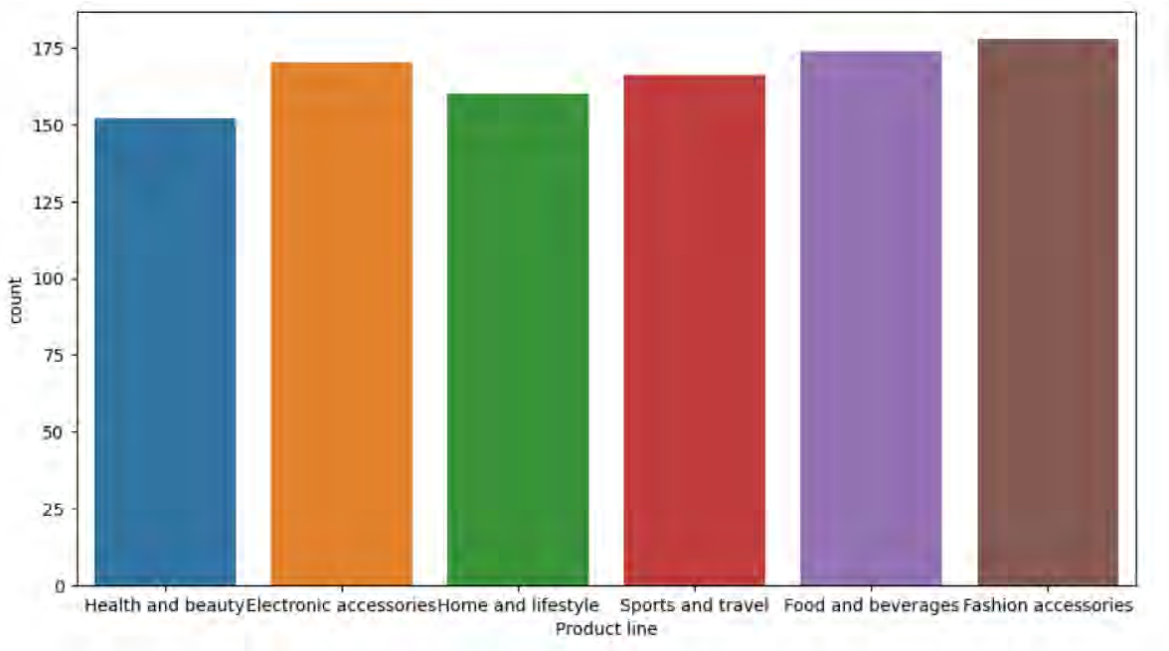
Ποιοτικά Δεδομένα

Τα κατηγορικά δεδομένα αντιπροσωπεύουν χαρακτηριστικά όπως το φύλο, η οικογενειακή κατάσταση, η πόλη καταγωγής ενός ατόμου ή τα είδη ταινιών που του αρέσουν (Chambers, 2017)

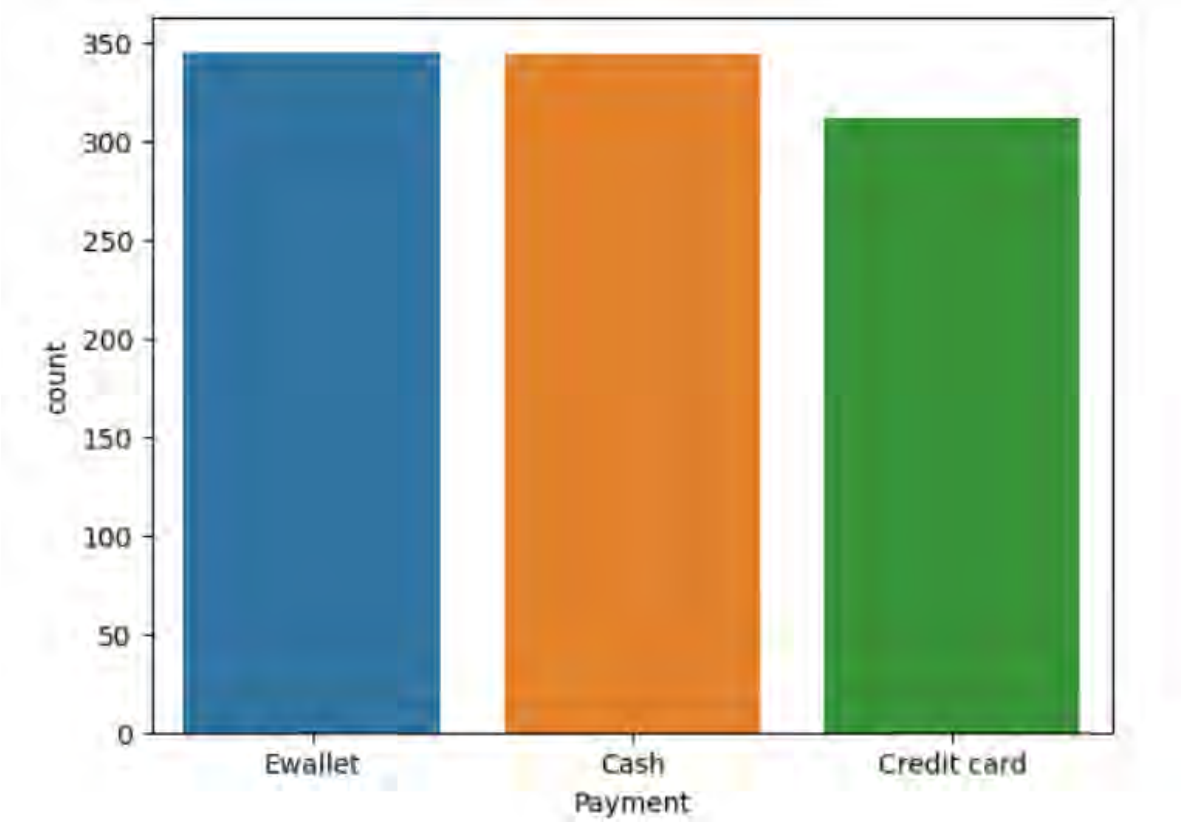
3.1.1 Επεξεργασία δεδομένων

Ανάλυση Μεταβλητών - Κατηγορικά Δεδομένα

Σύμφωνα με τον **Watson (2014)**, η ανάλυση κατηγορικών δεδομένων είναι όταν η μεταβλητή απόκρισης ομαδοποιείται σε ένα σύνολο αλληλοαποκλειόμενων κατηγοριών διατεταγμένων (όπως ηλικιακή ομάδα) ή μη ταξινομημένων (όπως το χρώμα των ματιών). Σε αυτό το σύνολο δεδομένων, τα κατηγορικά δεδομένα είναι ο "τύπος πελάτη", το "φύλο" και η "πόλη".



Εικόνα 6 Product Line attribute bar char



Εικόνα 7 Payment bar chart

Με τη χρήση της εντολής read και της εντολής head φαίνονται οι πέντε πρώτες εγγραφές του επιλεγμένου αρχείου.

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
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
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
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
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
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

Εικόνα 8 Πέντε πρώτες εγγραφές του αρχείου

Ανίχνευση Ανωμαλιών

Μια λεπτομερής επιθεώρηση που είχε ως αποτέλεσμα τον εντοπισμό σφαλμάτων ή ακραίων τιμών και ανωμαλιών, όπως τιμές που λείπουν, πολλαπλά μηδενικά στη στήλη ημερομηνίας, μηδενικές τιμές στην ποσότητα κ.λπ. εντός 1000 εγγραφών δεδομένων. Εντοπίστηκαν περίπου 90 ανωμαλίες και απαιτήθηκε καθαρισμός. Επιπλέον, από αυτές τις 90 ανωμαλίες που εντοπίστηκαν στο σύνολο δεδομένων, εντοπίστηκαν 7 ακραίες τιμές. Αυτά ήταν κυρίως αριθμητικά δεδομένα

Σε αυτή τη φάση, θεωρήθηκε ότι οι τιμές που λείπουν αντιπροσωπεύουν μηδενικές τιμές ή ακραίες τιμές εντός του συνόλου δεδομένων. Αυτό δικαιολογείται ώστε τα δεδομένα να είναι έτοιμα προς επεξεργασία. Ο θόρυβος στα δεδομένα κατέγραψε τιμές που οδήγησαν σε λύση λόγω της έλλειψης περιορισμών στα δεδομένα που εισήγαγε ένας ταμίας σε αυτά τα υποκαταστήματα.

Εικόνα 9 Missing value in the branch field

Μέσω της εντολής shape φαίνεται ότι το αρχείο αποτελείται από 1000 γραμμές και 17 στήλες χαρακτηριστικά τα οποία είναι :

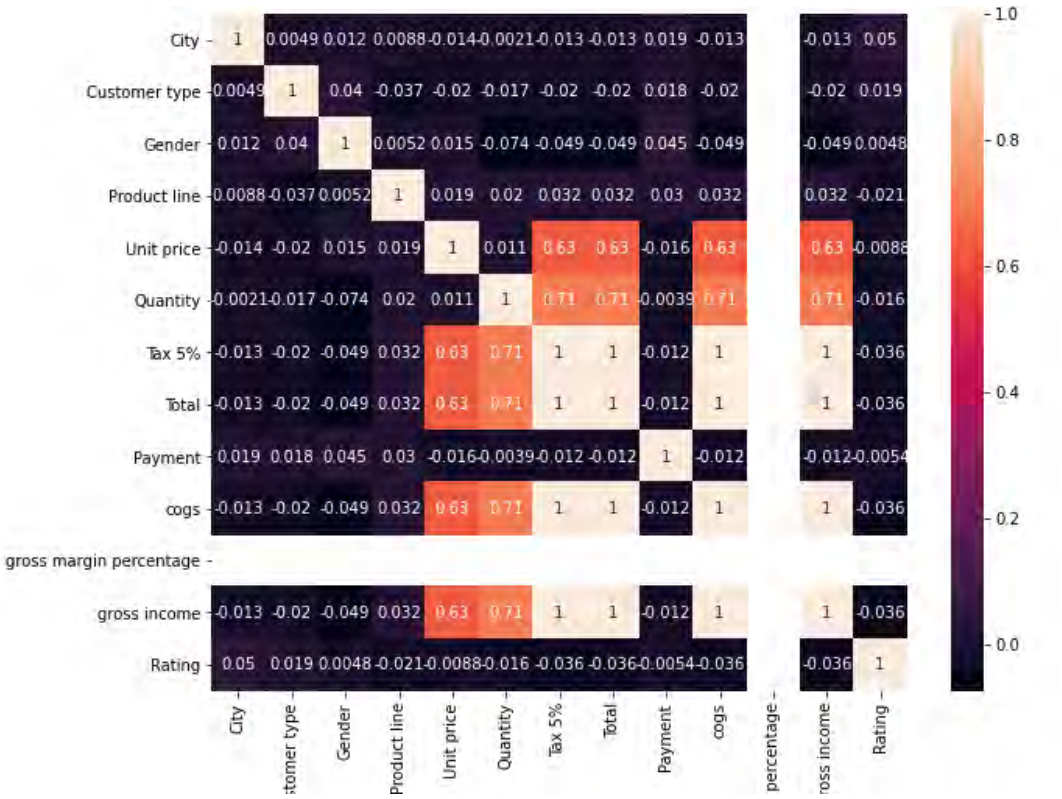
Invoice ID	**Παραστατικό πληρωμής**
Branch	Κλάδος
City	Πόλη που βρίσκεται το κατάστημα
Customer type	Τύπος πελάτη
Gender	Φύλο πελάτη
Product line	Κατηγοριοποίηση προϊόντων
Unit price	Τιμή μιας μονάδας του προϊόντος

Quantity	Ποσότητα αγορασμένη από τον πελάτη
Tax 5%	Φόρος
Total	Συνολική τιμή συμπεριλαμβανομένου του φόρου
Date	Ημερομηνία αγοράς
Time	Ώρα αγοράς
Payment	Τρόπος πληρωμής
Cogs	Κόστος πωληθέντων προϊόντων
Gross margin percentage	Περιθώριο κέρδους
Gross income	Ακαθάριστο εισόδημα
Rating	Αξιολόγηση της συνολικής αγοραστικής εμπειρίας

Το Invoice Id δεν προσφέρει κάποια πληροφορία έτσι λοιπόν αφαιρείται από το σύνολο δεδομένων. Το Branch συμπίπτει με τη City άρα κάποιο από τα δύο πρέπει να αφαιρεθεί, αυθαίρετα αφαιρείται το Branch. Τέλος το Date γίνεται η υπόθεση ότι δεν παίζει ρόλο και αφαιρείται και αυτό.

Ακόμα εντοπίστηκαν τα χαρακτηριστικά που έχουν κατηγορικές τιμές και μετατράπηκαν σε αριθμητικές προκειμένου να είναι έτοιμες για την εκπαίδευση του μοντέλου . Αυτή η διαδικασία έγινε με τη χρήση της κλήσης της εντολής LabelEncoder η οποία από αυτομάτως δίνει αριθμητικές τιμές σε χαρακτηριστικά με κατηγορικές.

Στη συνέχεια παρουσιάζεται ο πίνακας συσχέτισης (correlation matrix) μεταξύ των χαρακτηριστικών που δείχνει τις συσχετίσεις που έχουν όλα τα χαρακτηριστικά μεταξύ τους .



Εικόνα 10 Πίνακας συσχέτισης

Τα χαρακτηριστικά που θα μπούνε ως είσοδοι στο μοντέλο είναι τα: City, Customer type, Gender, Product line, Payment, cogs, gross income, Rating. Το gross margin percentage δεν λαμβάνεται υπόψιν καθώς είναι ίδιο για όλες τις παρατηρήσεις. Τα Quantity, Unit price, Tax 5% επίσης δεν θα μπούνε στο μοντέλο καθώς το Total που είναι το ζητούμενο της πρόβλεψης βγαίνει σε συνάρτηση με τα άλλα τρία (Total= Quantity * Unit price + Tax 5%).

3.1.2 Διάσπαση των δεδομένων

Τα δεδομένα διασπώνται σε δύο σύνολα το πρώτο το οποίο είναι για την εκπαίδευση του μοντέλου και το δεύτερο το οποίο χρησιμοποιείται για τον έλεγχο και την επαλήθευσή του. Αυτή η μέθοδος ονομάζεται train test split και είναι μια εύκολη και γρήγορη μέθοδος που χρησιμοποιείται για την αξιολόγηση της απόδοσης ενός αλγορίθμου μηχανικής μάθησης. Συνήθως το μέγεθος του πρώτου συνόλου είναι ίσο με το 80 % επί του συνολικού αριθμού των παρατηρήσεων και ονομάζεται training dataset και το δεύτερο σύνολο είναι ίσο με το υπόλοιπο 20% και ονομάζεται test dataset. Παρά το γεγονός ότι είναι απλή στη χρήση

υπάρχουν πολλές περιπτώσεις που δεν πρέπει να χρησιμοποιείται όπως όταν έχουμε ένα μικρό σύνολο δεδομένων.

3.2 Αξιολόγηση των μοντέλων

Ένα κοινό ερώτημα για κάποιον αρχάριο στα μοντέλα πρόβλεψης είναι 'Πως μπορώ να υπολογίσω την ακρίβεια σε ένα πρόβλημα παλινδρόμησης'?. Η απάντηση στην ερώτηση είναι ότι η ακρίβεια είναι ένα μέτρο που χρησιμοποιείται για προβλήματα ταξινόμησης. Η απόδοση σε αυτού του είδους τα προβλήματα μπορεί να αναφερθεί ως σφάλμα στις προβλέψεις. Αυτό βγάζει νόημα καθώς αν κάποιος θέλει να προβλέψει μια αριθμητική τιμή δεν θέλει να γνωρίζει αν το μοντέλο πρόβλεψε σωστά την τιμή , αλλά θέλει να γνωρίζει πόσο κοντά ήταν στις αναμενόμενες τιμές. Οι δείκτες που χρησιμοποιήθηκαν για την αξιολόγηση της απόδοσης είναι:

1. Μέσο απόλυτο σφάλμα (Mean Absolute Error, MAE). Αποτελεί την διαφορά σε απόλυτες τιμές μεταξύ της πραγματικής και της προβλεπόμενης τιμής. Αποτελεί ένα γραμμικό δείκτη που σημαίνει ότι σε κάθε μεμονωμένη διαφορά δίνει την ίδια βαρύτητα. Δεν είναι κατάλληλο για περιπτώσεις που κάποιος θέλει να δώσει προσοχή στις ακραίες τιμές . Ο τύπος του μέσου απόλυτου σφάλματος είναι :

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}\frac{|y_i \ - \ \hat{y}i|}{n}$$

2. Μέσο τετράγωνο σφάλμα ή μέση τετραγωνική απόκλιση (Mean Squared Error, MSE). Είναι έναν από τους προτιμότερους δείκτες για προβλήματα παλινδρόμησης. Μετρά την μέση τιμή των τετραγώνων των σφαλμάτων μεταξύ των πραγματικών τιμών και των εκτιμώμενων. Από το γεγονός ότι τετραγωνίζει τη διαφορά μεταξύ των τιμών τιμωρεί ακόμα και ένα μικρό σφάλμα πράγμα το οποίο οδηγεί σε υπέρ εκτίμηση του πόσο κακό είναι το μοντέλο. Προτιμάται έναντι άλλων δεικτών γιατί είναι εναλλακτικός και μπορεί να βελτιστοποιηθεί καλύτερα. Δεν υπάρχει μια συγκεκριμένη τιμή που να μας δείχνει ότι το μοντέλο μας είναι καλό. Απλά όσο πιο

κοντά στο 0 είναι τόσο πιο ακριβές είναι. Εάν ισούται με 0 σημαίνει ότι το μοντέλο είναι τέλειο. Ο τύπος του μέσου τετράγωνου σφάλματος είναι:

$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

MSE = μέσο τετραγωνικό σφάλμα

N = αριθμός σημείων δεδομένων

Y_i= τιμές παρατήρησης

$\hat{Y}_i$=τιμές πρόβλεψης

3. Ρίζα Μέσου Τετράγωνου Σφάλματος (Root Mean Squared Error, RMSE). Αποτελεί τον δείκτη που χρησιμοποιείται εκτενώς σε προβλήματα παλινδρόμησης και αποτελεί την τετραγωνική ρίζα του δείκτη MSE. Προτιμάται περισσότερο γιατί τα σφάλματα πρώτα τετραγωνίζονται πράγμα το οποίο επιβάλλει υψηλή ποινή σε μεγάλα σφάλματα. Όσο πιο μικρή είναι η τιμή του τόσο πιο εύκολα το μοντέλο κολλάει με ένα σύνολο δεδομένων. Η ιδανική περίπτωση είναι να ισούται με το 0 . Ο τύπος του RMSE είναι:

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\widehat{y_i} - y_i)^2}{n}}$$

4. Μέσο απόλυτο ποσοστό σφάλματος (Mean Absolute Percentage Error, MAPE). Μετράει την ακρίβεια της χρησιμοποιούμενης μεθόδου. Αντιπροσωπεύει το μέσο των απόλυτων ποσοστιαίων σφαλμάτων κάθε εισόδου στο σύνολο δεδομένων δείχνοντας κατά μέσο όρο πόσο ακριβείς ήταν οι προβλέψεις σε σχέση με τις πραγματικές τιμές. Αν το αποτέλεσμα είναι κάτω από το 0,2 η πρόβλεψη είναι καλή. Ο τύπος του MAPE είναι:

$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - Fi|}{A_i}$

Ai είναι η πραγματική τιμή

Fi είναι η προβλεπόμενη τιμή

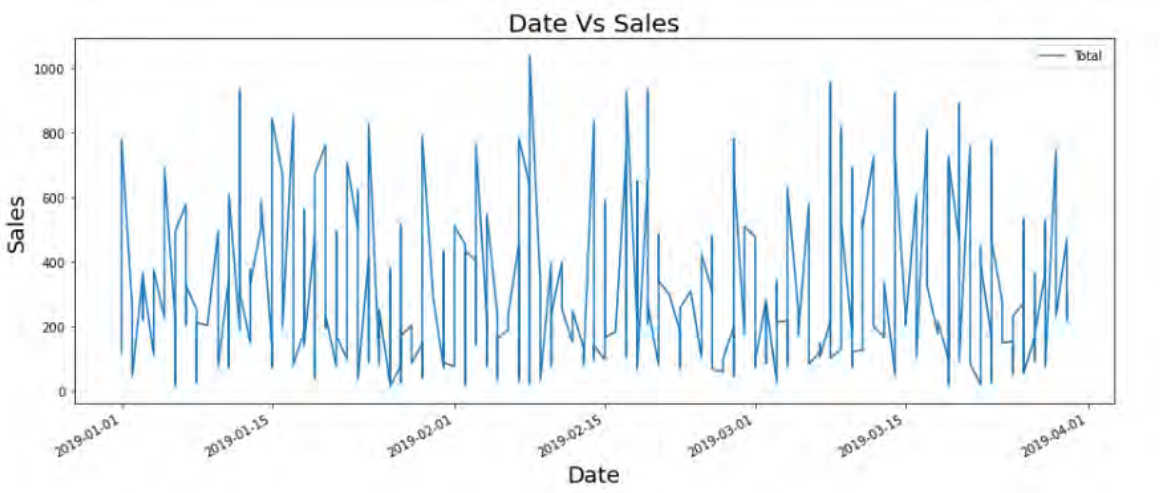
N είναι ο συνολικός αριθμός των παρατηρήσεων

Κεφάλαιο 4 Πρακτικό Μέρος

4.1 Arima

Ο αυτοπαλινδρομικός ολοκληρωμένος κινητός μέσος όρος (Autoregressive integrated moving average – ARIMA) αποτελεί ένα μοντέλο που χρησιμοποιείται στην στατιστική και στην οικονομετρία για να μετρήσει γεγονότα που θα συμβούν σε μια συγκεκριμένη περίοδο του χρόνου. Αποτελεί ένα κλασικό στατιστικό μοντέλο για πρόβλεψη και ανάλυση χρονοσειρών. Είναι επίσης γνωστό ως μέθοδος Box-Jenkins προς τιμήν των ανθρώπων που την πρότειναν[24]. Ένα μοντέλο ARIMA ορίζεται από τρεις παραμέτρους p,d and q. Πιο συγκεκριμένα το p καθορίζει τον αριθμό των αυτοπαλινδρομικών όρων στον μοντέλο, το d καθορίζει τον αριθμό των διαφοροποιήσεων που εφαρμόζονται στις τιμές των χρονοσειρών, το q καθορίζει τον αριθμό των όρων του κινητού μέσου όρου στο μοντέλο.

Prediction per city “Yangon”



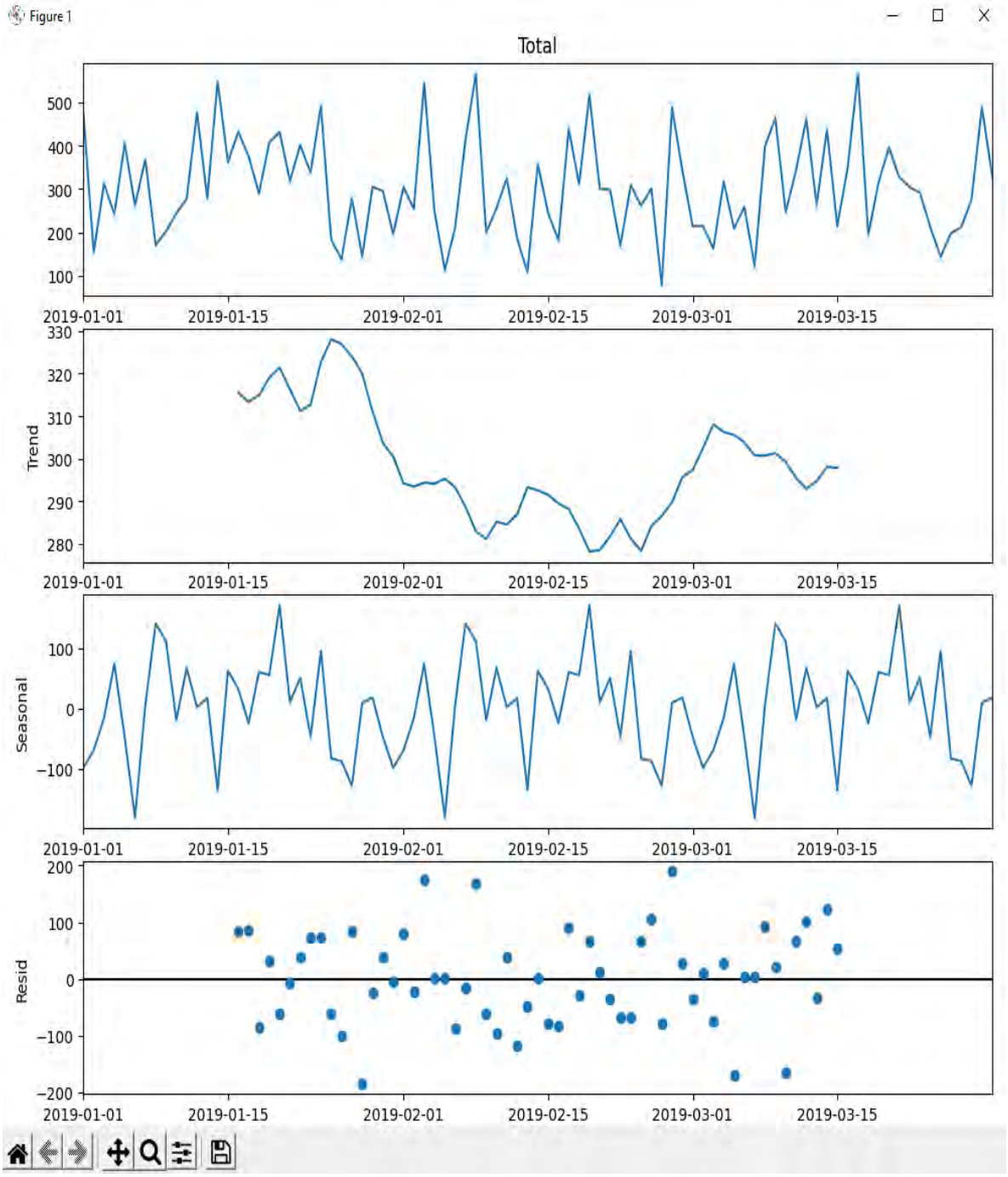
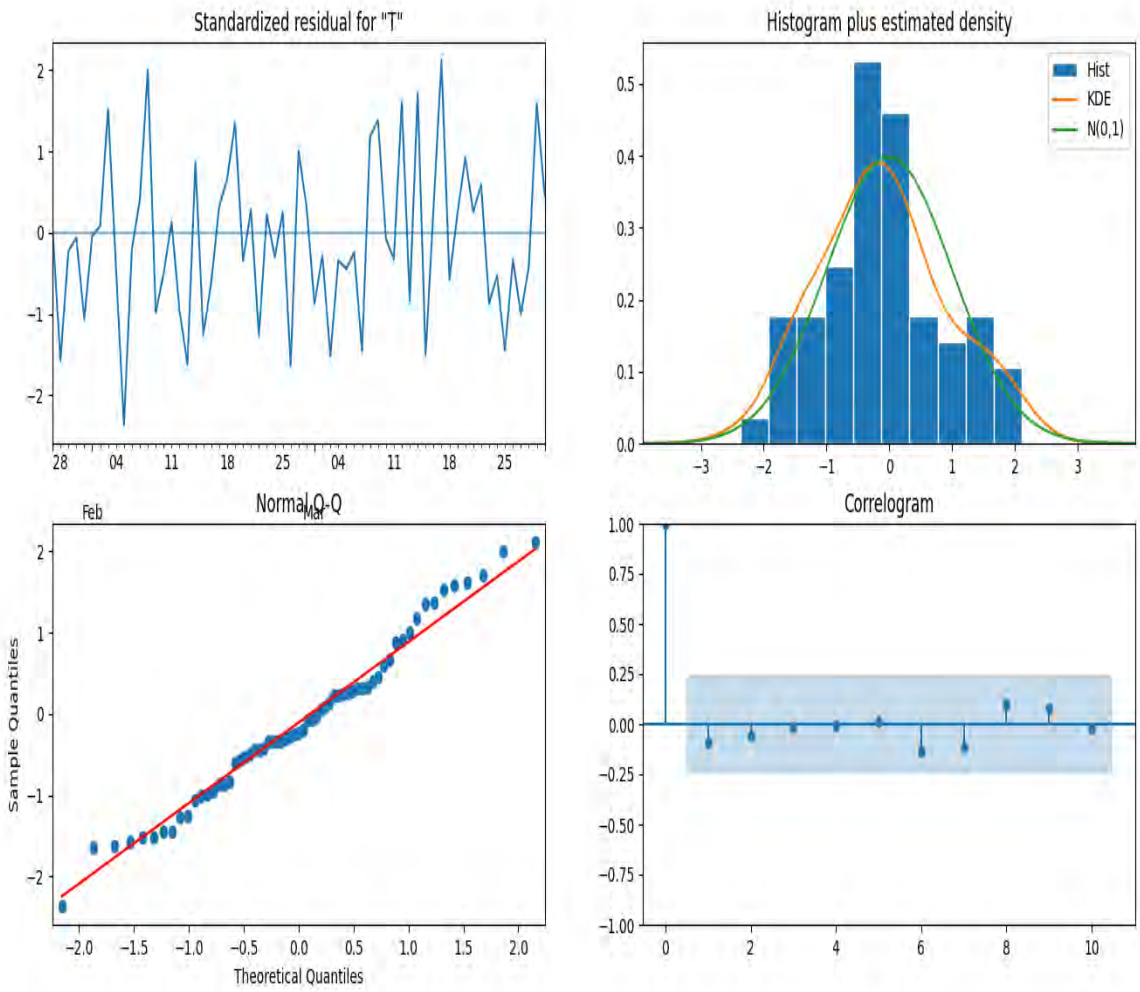


Figure 1
Total
Trend
Seasonal
Resid
2019-01-01
2019-01-15
2019-02-01
2019-02-15
2019-03-01
2019-03-15


Standardized residual for "T"
Histogram plus estimated density
Hist
KDE
N(0,1)
Normal Q-Q
Feb
Mar
Sample Quantiles
Theoretical Quantiles
Correlogram

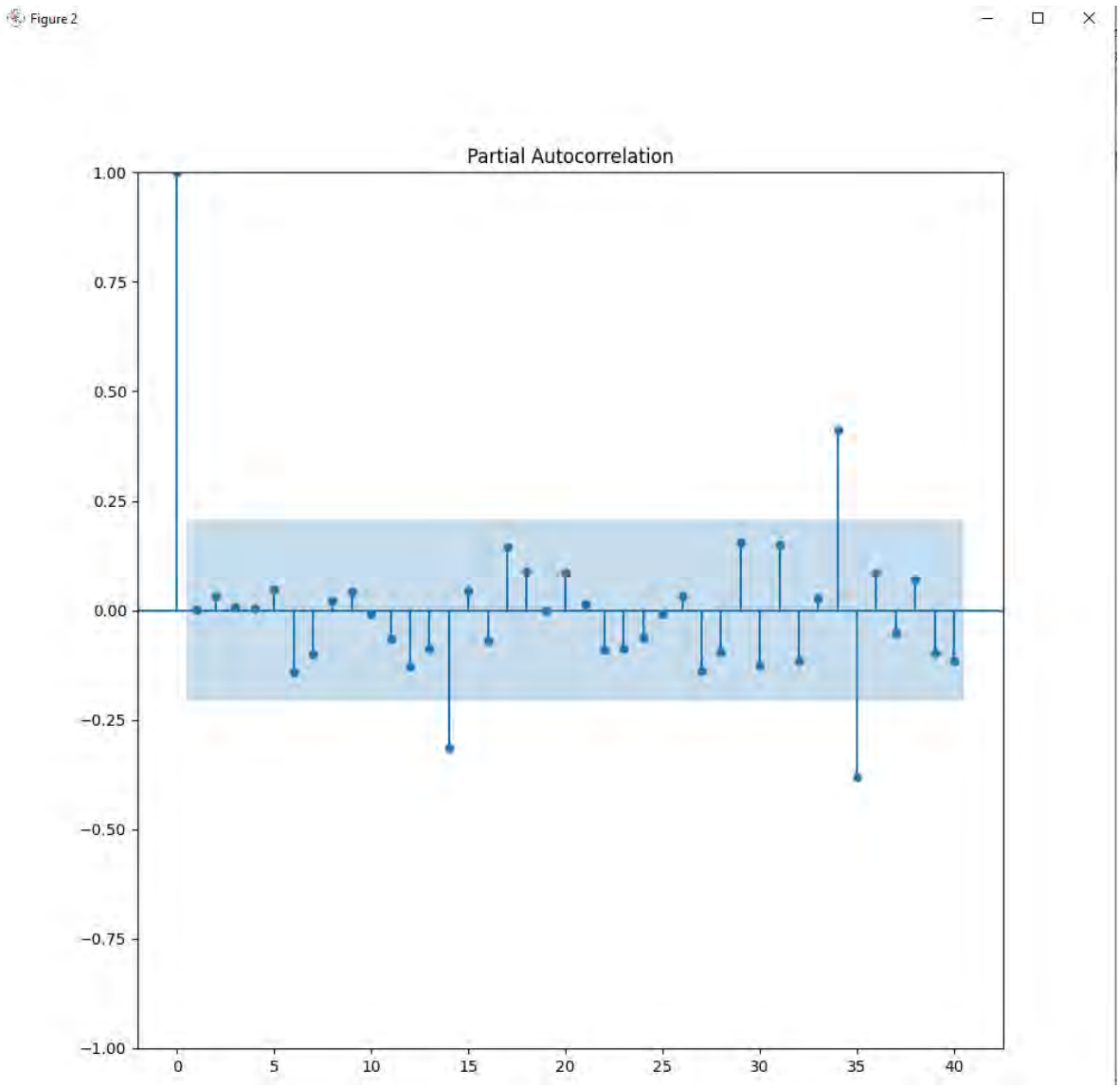
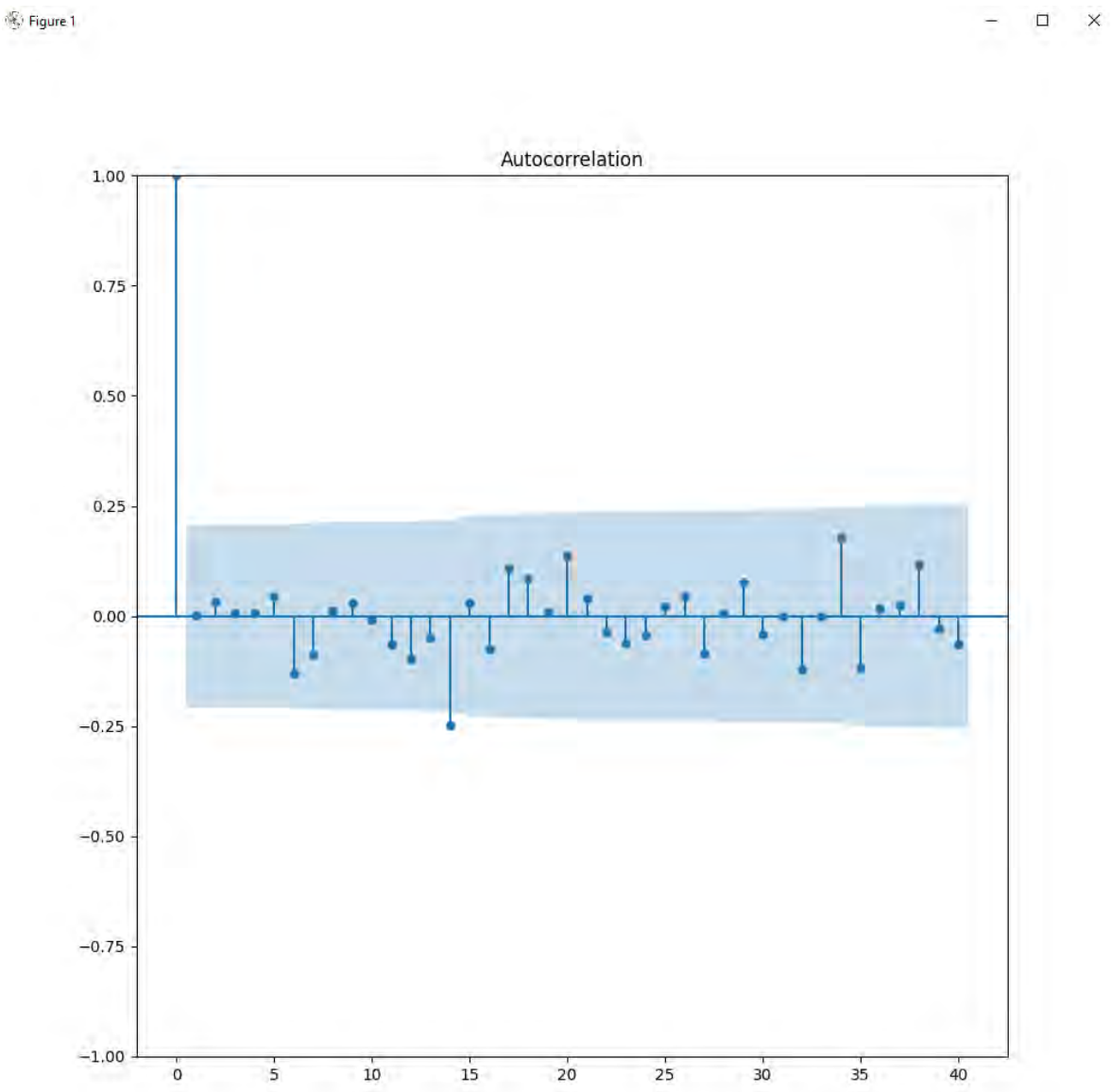


Figure 2
Partial Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20
25
30
35
40



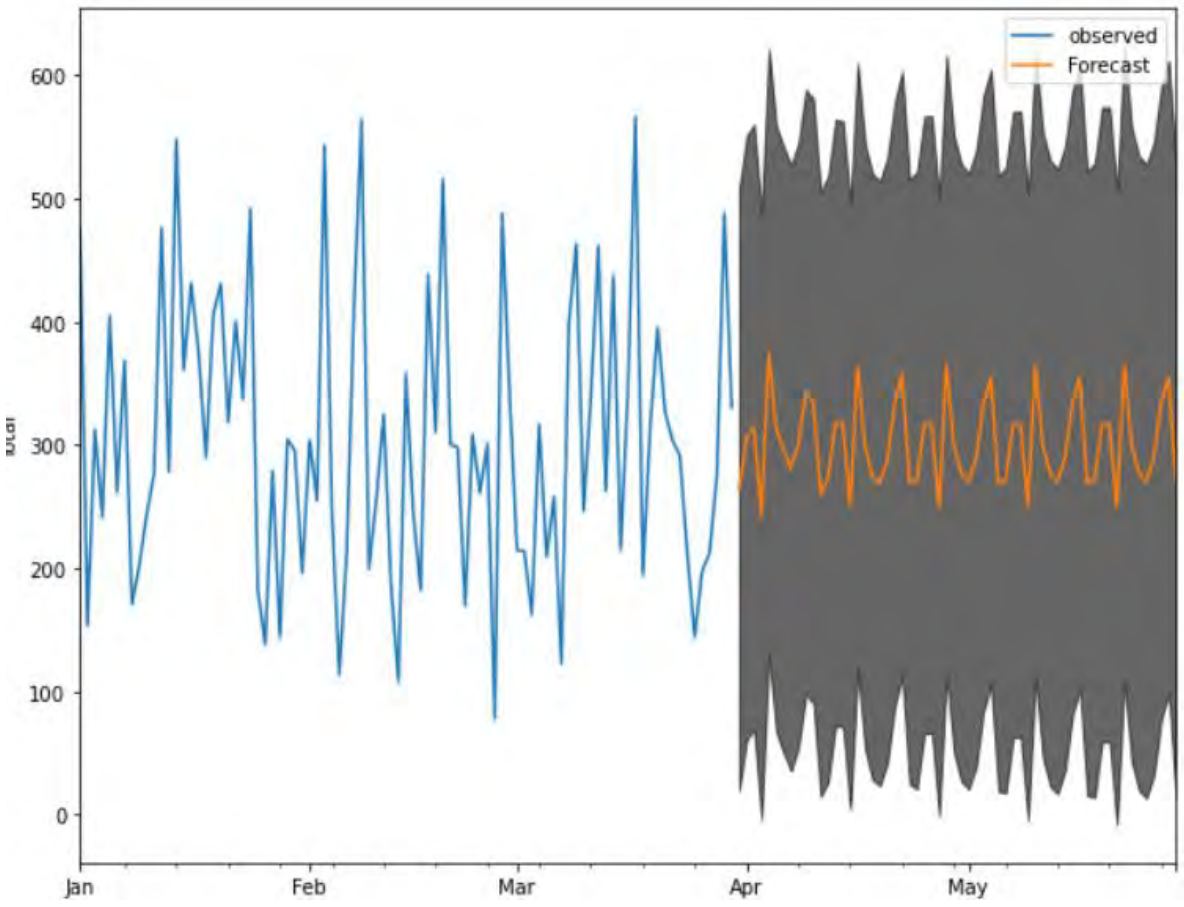

```
   N    Tit     Tnf  Tnint  Skip  Nact     Projg        F
   5     36      46      1     0     0   9.181D-06   4.455D+00
  F =   4.4545668725189369

CONVERGENCE: NORM_OF_PROJECTED_GRADIENT_<=_PGTOL
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
ar.L1         -0.7799      0.291     -2.678      0.007      -1.351      -0.209
ma.L1          0.8661      0.252      3.432      0.001       0.371       1.361
ar.S.L12      -0.1447      0.180     -0.804      0.421      -0.498       0.208
ma.S.L12      -0.7988      0.352     -2.271      0.023      -1.488      -0.109
sigma2       1.52e+04   4263.174      3.565      0.000    6841.862    2.36e+04
================================================================================
```

```
CONVERGENCE: REL_REDUCTION_OF_F_<=_FACTR*EPSMCH
ARIMA(1, 1, 1) x (1, 1, 1, 12) 12 -- AIC : 796.5081588034477
RUNNING THE L-BFGS-B CODE

           * * *

Machine precision = 2.220D-16
```



Εικόνα 4 Πρόβλεψη για 2 μήνες

Στο παραπάνω γράφημα, παρουσιάζεται μία πρόβλεψη για τους επόμενους 2 μήνες.

4.2 MachineLearning

Μια αναφορά ταξινόμησης χρησιμοποιείται για τη μέτρηση της ποιότητας των προβλέψεων από έναν αλγόριθμο ταξινόμησης. Πόσες προβλέψεις είναι αληθείς και πόσες ψευδείς. Πιο συγκεκριμένα, τα True Positives, False Positives, True negatives και False

Negatives χρησιμοποιούνται για την πρόβλεψη των μετρήσεων μιας αναφοράς ταξινόμησης όπως φαίνεται παρακάτω.

	precision	recall	f1-score	support
Iris-setosa	1.00	1.00	1.00	50
Iris-versicolor	0.77	0.96	0.86	50
Iris-virginica	0.95	0.72	0.82	50
avg / total	0.91	0.89	0.89	150

Η αναφορά δείχνει την ακρίβεια των βασικών μετρήσεων ταξινόμησης, την ανάκληση και τη βαθμολογία f1 σε βάση ανά τάξη. Οι μετρήσεις υπολογίζονται χρησιμοποιώντας αληθή και ψευδή θετικά, αληθή και ψευδή αρνητικά. Θετικά και αρνητικά σε αυτή την περίπτωση είναι τα γενικά ονόματα για τις προβλεπόμενες κλάσεις. Υπάρχουν τέσσερις τρόποι για τον έλεγχο οι προβλέψεις είναι σωστές ή λάθος:

Ακρίβεια (Precision) – Πόσο τοις εκατό των προβλέψεών ήταν σωστές;

Η ακρίβεια είναι η ικανότητα ενός ταξινομητή να μην επισημαίνει μια θετική παρουσία που είναι στην πραγματικότητα αρνητική. Για κάθε κλάση ορίζεται ως ο λόγος των αληθινών θετικών προς το άθροισμα των αληθών και των ψευδών θετικών.

TP – Αληθινά Θετικά

FP – Ψευδώς θετικά

Ακρίβεια – Ακρίβεια θετικών προβλέψεων.

Ακρίβεια = TP/(TP + FP)

```
from sklearn.metrics import precision_score
print("Precision score: {}".format(precision_score(y_true,y_pred)))
```

Recall (Ανάκληση) – Ποιο τοις εκατό των θετικών κρουσμάτων εντοπίσατε; Η ανάκληση είναι η ικανότητα ενός ταξινομητή να βρίσκει όλες τις θετικές περιπτώσεις. Για κάθε

κατηγορία ορίζεται ως ο λόγος των αληθινών θετικών προς το άθροισμα των αληθινών θετικών και των ψευδών αρνητικών.

FN – False Negatives

Ανάκληση: Κλάσμα θετικών που εντοπίστηκαν σωστά.

Ανάκληση = TP/(TP+FN)

```
from sklearn.metrics import recall_score
print("Recall score: {}".format(recall_score(y_true,y_pred)))
```

Βαθμολογία F1 – Ποιο ποσοστό των θετικών προβλέψεων ήταν σωστές;

Η βαθμολογία F1 είναι ένας σταθμισμένος αρμονικός μέσος ακρίβειας και ανάκλησης έτσι ώστε η καλύτερη βαθμολογία είναι 1,0 και η χειρότερη είναι 0,0. Σε γενικές γραμμές, οι βαθμολογίες F1 είναι χαμηλότερες από τα μέτρα ακρίβειας, καθώς ενσωματώνουν την ακρίβεια και την ανάκληση στον υπολογισμό τους. Ως εμπειρικός κανόνας, ο σταθμισμένος μέσος όρος του F1 θα πρέπει να χρησιμοποιείται για τη σύγκριση μοντέλων ταξινομητών και όχι συνολικής ακρίβειας.

Βαθμολογία F1 = 2*(Ανάκληση * Ακρίβεια) / (Ανάκληση + Ακρίβεια)

```
from sklearn.metrics import f1_score
print("F1 Score: {}".format(f1_score(y_true,y_pred)))
```

4.2.1 KNeighborsClassifier

Τα αποτελέσματα της έκθεσης ταξινόμησης, του KNeighborsClassifier για 7 κοντινούς γείτονες παρουσιάζονται παρακάτω. Η αναφορά ταξινόμησης χρησιμοποιείται συχνά στη μηχανική μάθηση για τον υπολογισμό της ακρίβειας ενός μοντέλου ταξινόμησης με βάση τις τιμές από τον πίνακα σύγχυσης. Από προεπιλογή, ο KNeighborsClassifier αναζητά τους 5 πλησιέστερους γείτονες. Πρέπει να οριστεί στον ταξινομητή η χρήση της Ευκλείδειας απόστασης για τον προσδιορισμό της εγγύτητας μεταξύ γειτονικών σημείων. Κατόπιν δοκιμών επιλέχθηκαν οι 7 κοντινότεροι γείτονες.

High Precision: Δεν προβλέφθηκαν πολλές True τιμές ως False

Το recall, που ονομάζεται επίσης ευαισθησία ή ρυθμός επιτυχίας, μας λέει το κλάσμα των σωστά προσδιορισμένων θετικών προβλέψεων.

Το f1-score, ή το μέτρο F, μετρά την ακρίβεια και την ανάκληση ταυτόχρονα βρίσκοντας τον αρμονικό μέσο όρο των δύο τιμών.

Το accuracy μετρά την ακρίβεια όλων των προβλέψεων (θετικών και αρνητικών).

```
from sklearn.neighbors import KNeighborsClassifier
knn=KNeighborsClassifier(n_neighbors=7)

knn.fit(x_train,y_train)

Classification Report is:
               precision    recall  f1-score   support

           0       0.47      0.49      0.48       100
           1       0.47      0.45      0.46       100

    accuracy                           0.47       200
   macro avg       0.47      0.47      0.47       200
weighted avg       0.47      0.47      0.47       200

Confusion Matrix:
 [[49 51]
 [55 45]]
Training Score:
 64.75
```

Το σκορ εκπαίδευσης είναι περίπου 65, το οποίο ναι μεν είναι καλό αλλά όχι το καλύτερο. Ο αλγόριθμος k-πλησιέστερων γειτόνων (KNN) είναι μια μη παραμετρική μέθοδος που χρησιμοποιείται για ταξινόμηση και παλινδρόμηση. Και στις δύο περιπτώσεις, η είσοδος αποτελείται από τα k πιο κοντινά παραδείγματα εκπαίδευσης στον χώρο χαρακτηριστικών. Το αποτέλεσμα της εξόδου εξαρτάται από το εάν το k-NN χρησιμοποιείται για ταξινόμηση ή παλινδρόμηση»

4.2.2 SVC

Στη μηχανική μάθηση, οι μηχανές διανύσματος υποστήριξης (SVM, επίσης δίκτυα διανυσμάτων υποστήριξης) είναι εποπτευόμενα μοντέλα εκμάθησης με σχετικούς

αλγόριθμους μάθησης που αναλύουν δεδομένα για ανάλυση ταξινόμησης και παλινδρόμησης.

```
from sklearn.svm import SVC

svc = SVC()
svc.fit(x_train, y_train)
```

```
Classification Report is:
               precision    recall  f1-score   support

           0       0.45      0.49      0.47       100
           1       0.44      0.40      0.42       100

    accuracy                           0.45       200
   macro avg       0.44      0.45      0.44       200
weighted avg       0.44      0.45      0.44       200

Confusion Matrix:
 [[49 51]
 [60 40]]
Training Score:
 55.50000000000001
```

Το σκορ εκπαίδευσης είναι περίπου 55, το οποίο ναι μεν είναι καλό αλλά όχι το καλύτερο και είναι χειρότερο από τον εκτιμητή KNeighborsClassifier.

Στην εφαρμογή sklearn της python του μοντέλου Ταξινόμησης διανυσμάτων υποστήριξης, υπάρχει μια λίστα διαφορετικών υπερπαραμέτρων. Οι πιο κρίσιμες υπερπαράμετροι για το SVM είναι ο πυρήνας, το C και το γάμμα.

- Η συνάρτηση πυρήνα (**kernel**) μετατρέπει το σύνολο δεδομένων εκπαίδευσης σε υψηλότερες διαστάσεις για να το μετατρέψει σε γραμμικά διαχωρισμό. Η προεπιλεγμένη συνάρτηση πυρήνα για την εφαρμογή python του ταξινομητή διανυσμάτων υποστήριξης είναι η συνάρτηση Radial Basis, η οποία συνήθως αναφέρεται ως rbf. Η συνάρτηση πυρήνα μπορεί να λάβει τιμές όπως linear, poly, rbf, sigmoid, precomputed, or callable.
- **C** είναι η παράμετρος τακτοποίησης l2. Η τιμή του C είναι αντιστρόφως ανάλογη με την ισχύ της τακτοποίησης.

- Το **gamma** είναι ο συντελεστής πυρήνα για rbf, poly και sigmoid. Μπορεί να θεωρηθεί ως το αντίστροφο της ακτίνας επιρροής του διανύσματος υποστήριξης. Η παράμετρος γάμμα επηρεάζει σε μεγάλο βαθμό την απόδοση του μοντέλου. Το gamma μπορεί να πάρει την τιμή scale, auto ή μιας τιμής float.

Οι παράμετροι που χρησιμοποιήθηκαν είναι οι παρακάτω

C	1.0
break_ties	False
cache_size	200
class_weight	None
coef0	0.0
decision_function_shape	ovr
degree	3
gamma	scale
kernel	rbf
max_iter	-1
probability	False
random_state	None
shrinking	True
tol	0.001
verbose	False

4.2.3 Gaussian NB

Είναι μια τεχνική ταξινόμησης που βασίζεται στο θεώρημα του Bayes με μια υπόθεση ανεξαρτησίας μεταξύ των προγνωστικών. Με απλά λόγια, ένας ταξινομητής Naive Bayes υποθέτει ότι η παρουσία ενός συγκεκριμένου χαρακτηριστικού σε μια κλάση δεν σχετίζεται με την παρουσία οποιουδήποτε άλλου χαρακτηριστικού.

```
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
gnb.fit(x_train,y_train)
```

```
Classification Report is:
               precision    recall  f1-score   support

           0       0.51      0.35      0.41       100
           1       0.50      0.66      0.57       100

    accuracy                           0.51       200
   macro avg       0.51      0.51      0.49       200
weighted avg       0.51      0.51      0.49       200

Confusion Matrix:
 [[35 65]
 [34 66]]
Training Score:
 55.125
```

Το σκορ εκπαίδευσης είναι περίπου 55, το οποίο ναι μεν είναι καλό αλλά όχι το καλύτερο και είναι χειρότερο από τον εκτιμητή KNeighborsClassifier. Είναι παρόμοιο με τον εκτιμητή SVC. Το σκορ είναι χαμηλό λόγω ότι ο Gaussian NB

- Βασίζεται σε και συχνά μια εσφαλμένη υπόθεση ανεξάρτητων χαρακτηριστικών. Δύσκολα, το dataset αποτελείται από ανεξάρτητα χαρακτηριστικά. Για παράδειγμα, η ανάλυση καταλληλότητας δανείου θα εξαρτηθεί από το εισόδημα, την ηλικία, το προηγούμενο δάνειο, την τοποθεσία και το ιστορικό συναλλαγών του αιτούντος που μπορεί να είναι αλληλεξαρτώμενα.
- Δεν είναι ιδανικό για σύνολα δεδομένων με μεγάλο αριθμό αριθμητικών χαρακτηριστικών. Εάν ο αριθμός των χαρακτηριστικών είναι μεγαλύτερος, τότε θα

υπάρχει υψηλό κόστος υπολογισμού και θα υποφέρει από την κατάρα της διάστασης.

- Εάν μια κατηγορία δεν καταγράφεται στο σετ εκπαίδευσης και εμφανίζεται στο σύνολο δεδομένων δοκιμής, τότε στο μοντέλο εκχωρείται 0 (μηδέν) πιθανότητα που οδηγεί σε λανθασμένο υπολογισμό. Αυτό το φαινόμενο αναφέρεται ως «μηδενική συχνότητα» και για να ξεπεραστούν τα φαινόμενα «μηδενικής συχνότητας» θα πρέπει να εφαρμοστούν τεχνικές εξομάλυνσης.

4.2.4 ExtraTreesClassifier

Το Extremely Randomized Trees Classifier (Extra Trees Classifier) είναι ένας τύπος τεχνικής εκμάθησης συνόλου που συγκεντρώνει τα αποτελέσματα πολλαπλών από-συσχετισμένων δέντρων αποφάσεων που συλλέγονται σε ένα «δάσος» για να εξάγει το αποτέλεσμα της ταξινόμησής του. Σε αυτήν την περίπτωση ο συνδυασμός των παρακάτω παραμέτρων είναι ο πιο αποδοτικός. Ο συνδυασμός προέκυψε από την βιβλιογραφία[1].

```
from sklearn.ensemble import ExtraTreesClassifier
etc = ExtraTreesClassifier(n_estimators=100, random_state=0)
etc.fit(x_train,y_train)
```

```
Classification Report is:
               precision    recall  f1-score   support

           0       0.50      0.50      0.50       100
           1       0.50      0.50      0.50       100

    accuracy                           0.50       200
   macro avg       0.50      0.50      0.50       200
weighted avg       0.50      0.50      0.50       200

Confusion Matrix:
 [[50 50]
 [50 50]]
Training Score:
 100.0
```

Το σκορ εκπαίδευσης είναι 100, το οποίο είναι το καλύτερο. Υπερτερεί φυσικά των εκτιμητών KNeighborsClassifier, SVC & Gaussian NB.

4.2.5 Decision Tree Classifier

Τα δέντρα αποφάσεων χρησιμοποιούν πολλαπλούς αλγόριθμους για να αποφασίσουν να χωρίσουν έναν κόμβο σε δύο ή περισσότερους υπο-κόμβους. Η δημιουργία υπο-κόμβων αυξάνει την ομοιογένεια των υποκόμβων που προκύπτουν. Το δέντρο απόφασης χωρίζει τους κόμβους σε όλες τις διαθέσιμες μεταβλητές και στη συνέχεια επιλέγει τη διαίρεση που οδηγεί στους περισσότερους ομοιογενείς υπο-κόμβους.

Ο αλγόριθμος ταξινομητή Decision Tree του Sci-kit Learn έχει πολλές υπερπαραμέτρους όπως

- **criterion**: Αποφασίζει το μέτρο της ποιότητας ενός διαχωρισμού με βάση κριτήρια όπως το "gini" για το Gini impurity και η "entropy" για το κέρδος πληροφοριών (information gain).
- **max_depth**: Το μέγιστο βάθος του δέντρου, τόσο μεγαλύτερο βάθος δέντρου γενικά υπερκαλύπτει τα δεδομένα.
- **max_leaf_nodes**: Μεγαλώνει το δέντρο με έναν καθορισμένο αριθμό κόμβων φύλλων με βάση τη μείωση του impurity.
- **min_samples_split**: Ο ελάχιστος αριθμός δειγμάτων που απαιτούνται για τον διαχωρισμό ενός εσωτερικού κόμβου.
- **min_samples_leaf**: Ο ελάχιστος αριθμός δειγμάτων που απαιτείται να βρίσκονται σε έναν κόμβο φύλλου του δέντρου.

```
from sklearn.tree import DecisionTreeClassifier
dtree = DecisionTreeClassifier(max_depth=6,
random_state=123,criterion='entropy')

dtree.fit(x_train,y_train)
```

```
Classification Report is:
              precision    recall  f1-score   support

           0       0.52      0.79      0.63       100
           1       0.56      0.27      0.36       100

    accuracy                           0.53       200
   macro avg       0.54      0.53      0.50       200
weighted avg       0.54      0.53      0.50       200
```

```
Confusion Matrix:
 [[79 21]
 [73 27]]
Training Score:
 63.87500000000001
```

Το σκορ εκπαίδευσης είναι περίπου 64, το οποίο είναι το ανάλογο των εκτιμητών KNeighborsClassifier, SVC & Gaussian NB. Παρόλο, τη βελτιστοποίηση των παραμέτρων δεν ήταν εφικτή η εύρεση ενός αποδοτικού μοντέλου.

4.2.6 Random Forest Classifier

Ένα τυχαίο δάσος είναι ένας μετά-εκτιμητής που ταιριάζει σε έναν αριθμό ταξινομητών δέντρων αποφάσεων σε διάφορα υποδείγματα του συνόλου δεδομένων και χρησιμοποιεί τον μέσο όρο για να βελτιώσει την προγνωστική ακρίβεια και τον έλεγχο της υπερπροσαρμογής. Στην περίπτωση αυτή δε χρειάστηκε καμία βελτίωση των παραμέτρων και το μοντέλο το οποίο δημιουργήθηκε είναι άκρως αποδοτικό.

```
from sklearn.ensemble import RandomForestClassifier
rfc=RandomForestClassifier()
rfc.fit(x_train,y_train)
```

```
Classification Report is:
               precision    recall  f1-score   support

           0       0.50      0.53      0.52       100
           1       0.51      0.48      0.49       100

    accuracy                           0.51       200
   macro avg       0.51      0.51      0.50       200
weighted avg       0.51      0.51      0.50       200

Confusion Matrix:
 [[53 47]
 [52 48]]
Training Score:
 100.0
```

Το σκορ εκπαίδευσης είναι 100, το οποίο είναι το καλύτερο όπως και του ExtraTreesClassifier και σαφώς καλύτερο των εκτιμητών KNeighborsClassifier, SVC & Gaussian NB.

4.2.7 XGBClassifier

Το XGBoost είναι μια δημοφιλής και αποτελεσματική εφαρμογή ανοιχτού κώδικα του αλγόριθμου δέντρων ενισχυμένων με κλίση. Η ενίσχυση κλίσης είναι ένας εποπτευόμενος αλγόριθμος μάθησης, ο οποίος επιχειρεί να προβλέψει με ακρίβεια μια μεταβλητή στόχο συνδυάζοντας τις εκτιμήσεις ενός συνόλου απλούστερων, πιο αδύναμων μοντέλων.

Υπάρχουν πολλές διαφορετικές παράμετροι στο XGBoost και ταξινομούνται ευρέως σε 3 τύπους:

- Γενικές παράμετροι
- Ενίσχυση παραμέτρων
- Παράμετροι εργασιών

Οι πιο σημαντικοί παράμετροι παρουσιάζονται παρακάτω.

XGBoost Hyperparameters A cheatsheet			
Hyperparameter	**Role**	**Range**	**Default**
subsample	Subsample ratio of training instance (e.g.|0.5 means 50% data used prior to growing trees)	>0 – 1	1
colsample_bytree	Frac. of columns selected for each tree	>0 - 1	1
colsample_bylevel	Frac. of columns selected for training at each level	>0 - 1	1
gamma (alias: min_split_loss)	Regularization param for tree pruning. Specifies minimum loss reduction required to make a split (a node is split only if there's positive reduction in loss func)	0 – inf.	0
booster	Selection of a booster (gbtree: tree-based, gblinear: Ridge regression)	'gbtree', 'gblinear', 'dart'	'gbtree'
learning_rate (alias: eta)	Regularization param. Step size. Shrinks feature weights in each boosting step.	0-1	0.3 {0.01, 0.1, 0.2}
max_depth	Maximum tree depth	0 – inf.	6
monotone_constraints	Increasing constraints on predictors (e.g. non-linear, increasing likelihood of loan approval with higher credit score)	(1,0), (0, -1)	'( )'
n_estimators	Number of trees boosted (early_stopping_rounds can be activated)	1 – inf.	100
reg_alpha (alias: alpha)	L1 regularization (increasing its value makes model more conservative)		0
reg_lambda (alias: lambda)	L2 regularization (increasing its value makes model more conservative)		1
scale_pos_weight	Dealing with imbalance data	Tot neg. instance (no-fraud)/tot pos. instance (fraud)	1

```
from xgboost import XGBClassifier

xgb =XGBClassifier(objective ='reg:linear', colsample_bytree = 0.3,
learning_rate = 0.1,
                max_depth = 5, alpha = 10, n_estimators = 10)

xgb.fit(x_train, y_train)
 Classification Report is:
               precision    recall  f1-score   support

           0       0.48      0.49      0.48       100
           1       0.47      0.46      0.47       100

    accuracy                           0.48       200
   macro avg       0.47      0.47      0.47       200
weighted avg       0.47      0.47      0.47       200
```

```
Confusion Matrix:
 [[49 51]
 [54 46]]
Training Score:
 63.87500000000001
```

Το σκορ εκπαίδευσης είναι περίπου 64, το οποίο είναι το ανάλογο των εκτιμητών KNeighborsClassifier, SVC & Gaussian NB. Η επιλογή των παραμέτρων δεν απέδωσε. Το σκορ εκπαίδευσης είναι σχετικά μικρό.

4.2.8 AdaBoostClassifier

Ένας ταξινομητής AdaBoost είναι ένας μετά-εκτιμητής που ξεκινά με την τοποθέτηση ενός ταξινομητή στο αρχικό σύνολο δεδομένων και στη συνέχεια προσαρμόζει πρόσθετα αντίγραφα του ταξινομητή στο ίδιο σύνολο δεδομένων, αλλά όπου τα βάρη των εσφαλμένα ταξινομημένων παρουσιών προσαρμόζονται έτσι ώστε οι επόμενοι ταξινομητές να εστιάζουν περισσότερο σε δύσκολες περιπτώσεις.

Οι κύριες παράμετροι είναι

- **base_estimator**: Αυτή η παράμετρος χρησιμοποιείται για να υποδηλώσει τον τύπο των βασικών μαθητών που μπορούμε να εφαρμόσουμε ή τον τύπο του αδύναμου εκπαιδευόμενου που θέλουμε να χρησιμοποιήσουμε. Μπορεί το δέντρο αποφάσεων, το Logistic Regressor, το SVC οτιδήποτε. Δεν μπορεί να είναι Knn καθώς το βάρος δεν μπορεί να εκχωρηθεί σε αυτό το μοντέλο. Από προεπιλογή, ο βασικός εκτιμητής είναι DecisionTreeClassifier(max_depth=1).
- **n_estimators**: Ο αριθμός των βασικών εκτιμητών ή αδύναμων μαθητών που θέλουμε να χρησιμοποιήσουμε στο σύνολο δεδομένων μας. Από προεπιλογή, ο n_estimator είναι 50.
- **Learning_rate**: Αυτή η παράμετρος παρέχεται για να συρρικνωθεί η συμβολή κάθε ταξινομητή. Από προεπιλογή, παρέχεται μια τιμή 1.

```
from sklearn.ensemble import AdaBoostClassifier
adb = AdaBoostClassifier(base_estimator = None)
```

```
adb.fit(x_train,y_train)

Classification Report is:
              precision    recall  f1-score   support

           0       0.54      0.54      0.54       100
           1       0.54      0.54      0.54       100

    accuracy                           0.54       200
   macro avg       0.54      0.54      0.54       200
weighted avg       0.54      0.54      0.54       200

Confusion Matrix:
 [[54 46]
 [46 54]]
Training Score:
 67.0
```

Το σκορ εκπαίδευσης είναι περίπου 67, το οποίο είναι το ανάλογο των εκτιμητών KNeighborsClassifier, SVC & Gaussian NB. Η επιλογή των παραμέτρων δεν απέδωσε. Το σκορ εκπαίδευσης είναι σχετικά μικρό.

Συμπεράσματα

Τα αποτελέσματα της πρόβλεψης πωλήσεων με τη μηχανική μάθηση παρουσιάζονται στον παρακάτω πίνακα

	f1 score	Training
AdaBoostClassifier	0.54	67
XGBClassifier	0,48	64
Random Forest Classifier	0,51	100
Decision Tree Classifier	0,53	64
ExtraTreesClassifier	0,50	100
Gaussian NB	0.51	55
SVC	0,45	55
KNeighborsClassifier	0,47	65

Από τον παραπάνω πίνακα προκύπτει ότι οι αλγόριθμοι Random Forest Classifier & ExtraTreesClassifier είχαν την καλύτερη επίδοση. Ως μελλοντική εργασία προτείνω την περαιτέρω βελτιστοποίηση (hyperparameter tuning) των παραμέτρων των αλγορίθμων οι οποίοι δεν απέδωσαν τα αναμενόμενα.

Βιβλιογραφία

1. Gupta, M. S., Keen, M. M., Shah, M. A., & Verdier, M. G. (Eds.). (2017*). Digital revolutions in public finance.* International Monetary Fund.
2. Marr, B. (2016). Why everyone must get ready for the 4th industrial revolution. *Forbes Tech*, *5*.
3. Kohli, S., Godwin, G. T., & Urolagin, S. (2021). Sales prediction using linear and KNN regression. In *Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019* (pp. 321-329). Springer Singapore.
4. Cheriyan, S., Ibrahim, S., Mohanan, S., & Treesa, S. (2018, August). Intelligent sales prediction using machine learning techniques. In *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)* (pp. 53-58). IEEE.
5. El Naqa, I., & Murphy, M. J. (2015). *What is machine learning?* (pp. 3-11). Springer International Publishing.
6. https://www.techtarget.com/searchenterpriseai/definition/supervised-learning.
7. https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/.
8. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html
9. Rokach, L., & Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *35*(4), 476-487.
10. Somvanshi, M., Chavan, P., Tambade, S., & Shinde, S. V. (2016, August). A review of machine learning techniques using decision tree and support vector machine. In *2016 international conference on computing communication control and automation (ICCUBEA)* (pp. 1-7). IEEE.
11. Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.
12. Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*, 197-227.
13. Cunningham, P., & Delany, S. J. (2021). k-Nearest neighbour classifiers-A Tutorial. *ACM computing surveys (CSUR)*, *54*(6), 1-25.

14. Ertuğrul, Ö. F., & Tağluk, M. E. (2017). A novel version of k nearest neighbor: Dependent nearest neighbor. *Applied Soft Computing*, *55*, 480-490.
15. Fix, E., & Hodges, J. L. (1951). Nonparametric discrimination: consistency properties. *Randolph Field, Texas, Project*, 21-49.
16. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.
17. Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, (4), 325-327.
18. https://www.techtarget.com/whatis/definition/support-vector-machine-SVM
19. Geurts, Pierre, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees." Machine learning 63.1 (2006): 3-42.
20. Hastie, Trevor; Rosset, Saharon; Zhu, Ji; Zou, Hui (2009). "Multi-class AdaBoost". Statistics and Its Interface. 2 (3): 349–360. doi:10.4310/sii.2009.v2.n3.a8. ISSN 1938-7989.
21. Ontivero-Ortega, Marlis, et al. "Fast Gaussian Naïve Bayes for searchlight classification analysis." *Neuroimage* 163 (2017): 471-479.
22. Markham, K., 2014. Simple guide to confusion matrix terminology. [Ηλεκτρονικό] Available at: http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
23. Brownlee, J. (2020) A gentle introduction to k-fold cross-validation, MachineLearningMastery.com. Available at: https://machinelearningmastery.com/k-fold-cross-validation/ (Accessed: April 2, 2023).
24. Ho, Siu Lau, and Min Xie. "The use of ARIMA models for reliability forecasting and analysis." *Computers & industrial engineering* 35.1-2 (1998): 213-216.