ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

ΑΝΑΠΤΥΞΗ ΣΥΣΤΗΜΑΤΟΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΤΗ ΔΙΑΠΡΑΓΜΑΤΕΥΣΗ ΚΡΥΠΤΟΝΟΜΙΣΜΑΤΩΝ

Υπό
ΙΩΑΝΝΗΣ ΚΑΡΑΠΑΡΙΣΗΣ

Επίβλεψη

ΜΙΧΑΛΗΣ ΔΟΥΜΠΟΣ

Χανιά, 2021

Ευχαριστίες

Θα ήθελε να εκφράσω την ευγνωμοσύνη μου προς τον επιβλέποντα καθηγητή μου κ. Μιχάλη Δούμπο για την συνεχή βοήθεια και καθοδήγηση του. Στη συνέχεια, θα ήθελα να ευχαριστήσω την Ελένη Άννα Γιακουντή για την αδιάκοπη υποστήριξη της. Τους συμφοιτητές μου και φίλους μου, Ιωάννη Τραπεζόντα, Άγγελο Ποθουλάκη, Τατιανά Ταράτσα και Αντωνία Καλαϊτζάκη. Τέλος ένα μεγάλο ένα ευχαριστώ στην οικογένεια μου.

Περίληψη

Τα τελευταία χρόνια, τα κρυπτονομίσματα έχουν γίνει ιδιαίτερα δημοφιλή και έχουν κεντρίσει το ενδιαφέρον των επενδυτών. Στην ουσία, πρόκειται για ψηφιακά νομίσματα, τα οποία δεν έχουν κάποια κεντρική εξουσία, είναι δηλαδή αποκεντρωμένα, και η αξία τους καθορίζεται από τους άμεσα εμπλεκόμενους- ενδιαφερόμενους. Το πρώτο κρυπτονόμισμα ήταν το Bitcoin, το οποίο δημιουργήθηκε το 2009 και είναι το πιο ευρέως διαδεδομένο κρυπτονόμισμα. Η τιμή του Bitcoin, παρουσιάζει ισχυρές διακυμάνσεις με αποτέλεσμα τη δημιουργία πολλών επενδυτικών ευκαιριών ακόμα και εντός μίας ημέρας. Σκοπός της παρούσας διπλωματικής εργασίας είναι η υλοποίηση ενός συστήματος το οποίο θα ανακαλύπτει τις ευκαιρίες αυτές και θα επενδύει αυτόματα, με απώτερο σκοπό την μεγιστοποιήση του κέρδους. Συγκεκριμένα, το σύστημα σε πρώτη φάση θα προβλέπει εάν η τιμή του Bitcoin θα είναι ανοδική ή καθοδική χρησιμοποιώντας αλγορίθμους μηχανικής μάθησης όπως νευρωνίκα δίκτυα. Στην δεύτερη φάση δοκιμάζονται διάφορες στρατηγικές, σε συνδυασμό με τα αποτελέσματα της πρώτης φάσης, με τελικό σκοπό την μεγιστοποίηση του αρχικού κεφαλαίου.

Περιεχόμενα

Κεφαλαίο 1ο

Εισαγωγή

1.1 Γενικά

Περίπου στις αρχές του 1950, πρωτοπόροι ερευνητές της επιστήμης υπολογιστών, άρχισαν να διερωτώνται εάν θα μπορούσαν να προγραμματίσουν υπολογιστές οι οποίοι θα κατάφερναν να «σκέφτονται». Έτσι, γεννήθηκε η ιδέα της τεχνητής νοημοσύνης. Τα επόμενα χρόνια ακολούθησε μεγάλος ενθουσιασμός γι' αυτό το νέο, πολλά υποσχόμενο πεδίο. Ωστόσο, στα τέλη της δεκαετίας του 1980, αυτό το τεράστιο ενδιαφέρον έφτασε στο αποκορύφωμα του και ξεκίνησε να φθίνει παροδικά. Τα τελευταία χρόνια, με την τρομερή ανάπτυξη της υπολογιστικής δύναμης των ηλεκτρονικών υπολογιστών, η τεχνητή νοημοσύνη γνώρισε σημαντική πρόοδο και ουσιαστικά αναδύθηκε από τις στάχτες της, καθώς κατάφερε να τραβήξει πάλι τα βλέμματα πάνω της και να κερδίσει την προσοχή αρκετών ερευνητών και μεγάλων πολυεθνικών εταιριών. Ένα από τα πιο δημοφιλή και επιτυχημένα υπό-πεδία της τεχνητής νοημοσύνης (βλ. σχήμα 1.1) είναι η μηχανική μάθηση (machine learning).

Σχήμα 1.1: Απεικόνιση γνωστικού πεδίου τεχνητής νοημοσύνης και μηχανικής μάθησης

Η μηχανική μάθηση είναι στενά συνδεμένη με τα μαθηματικά και την επιστήμη των υπολογιστών και απώτερος της σκοπός είναι μέσα από τα διαθέσιμα δεδομένα να εξάγει συμπεράσματα ή «κανόνες», να ανακαλύψει, όχι τόσο προφανή, επαναλαμβανόμενα μοτίβα ή και ακόμα να κάνει προβλέψεις για το μέλλον. Τα είδη των κατηγοριών της μηχανικής μάθησης εμπίπτουν στις 4 παρακάτω κατηγορίες:

1. **Μάθηση με επίβλεψη** (Supervised Learning)

2. **Μάθηση χωρίς επίβλεψη** (Unsupervised Learning)
3. **Μάθηση με ημι-επίβλεψη** (Semi-Supervised Learning)
4. **Ενισχυτική μάθηση** (Reinforcement Learning)

Η παρούσα διπλωματική χρησιμοποιεί αλγορίθμους μηχανικής μάθησης με επίβλεψη. Στη μάθηση υπό επίβλεψη το σύστημα καλείται να «μάθει» μια έννοια ή συνάρτηση από ένα σύνολο δεδομένων, η οποία αποτελεί περιγραφή ενός μοντέλου. Ονομάζεται έτσι επειδή θεωρείται ότι υπάρχει κάποιος «επιβλέπων» ο οποίος παρέχει τη σωστή τιμή εξόδου της συνάρτησης, για τα δεδομένα που εξετάζονται. Στην μάθηση με επίβλεψη διακρίνονται δυο είδη προβλημάτων, τα προβλήματα ταξινόμησης (classification) και τα προβλήματα παλινδρόμηση (regression). Στην ταξινόμηση η έκβαση των αποτελεσμάτων αφορά διακριτά ενδεχόμενα, δηλαδή αποτελέσματα του τύπου ναι/όχι (binary), ανήκει (/δεν ανήκει) σε μια ομάδα κλάσεων. Στην άλλη περίπτωση, της παλινδρόμησης το αποτέλεσμα είναι η ανάλυση ενός ποσοτικού μεγέθους που λαμβάνει συνεχείς τιμές. Για παράδειγμα, ένα πρόβλημα ταξινόμησης είναι εάν ένα email που παραλαμβάνει κάποιος χρήστης είναι κακόβουλο ή όχι, τύπου ναι/όχι (binary classification), δηλαδή το διάστημα των αποτελεσμάτων αποτελείται από το σύνολο {0,1}. Στον αντίποδα ένα πρόβλημα παλινδρόμησης μπορεί να είναι η εκτίμηση του μισθού ενός εργαζομένου με βάση την ηλικία του, το φύλο του και αλλά δημογραφικά χαρακτηριστικά, που ο μισθός, δηλαδή το αποτέλεσμα της εκτίμησης, μπορεί να πάρει οποιαδήποτε τιμή στο διάστημα $[0, +\infty)$.

Η μηχανική μάθηση, πλέον, έχει βρει πληθώρα εφαρμογών, όπως η χρήση της σε δραστηριότητες που εντάσσονται στην καθημερινή ζωή του ανθρώπου έως και σε επιστημονικές έρευνες που γίνονται σήμερα. Για παράδειγμα, πολλές ιστοσελίδες όπως το Youtube, το Netflix, το Facebook «εκμεταλλεύονται» τους αλγορίθμους αυτούς για να παρέχουν καλύτερες υπηρεσίες με σκοπό οι χρήστες να μεγιστοποιούν τον χρόνο που αφιερώνουν στις ιστοσελίδες αυτές. Πέρα από τις εμπορικές εφαρμογές χρησιμοποιείται για την ανακάλυψη εξωπλανήτων, νέων σωματιδίων, για την ανάλυση της ακολουθίας του DNA, για την ανίχνευση συγκεκριμένων τύπων καρκίνου, για αυτό-οδηγούμενα οχήματα και για τον τομέα των οικονομικών.

Αναλυτικότερα, στον τομέα των οικονομικών, οι καινοτόμοι αλγόριθμοι της μηχανικής μάθησης που άρχισαν να δημοσιεύονται ανά διαστήματα, προσέλκυσαν μεγάλο ενδιαφέρον στην ακαδημαϊκή κοινότητα καθώς και πληθώρα ερευνητικών ομάδων να πειραματιστούν με αυτούς έναντι των κλασικών μεθόδων καθώς σε άλλα επιστημονικά πεδία είχαν πετύχει σημαντικά αποτελέσματα, κάνοντας τους έτσι αρκετά υποσχόμενους όσο αφορά τις δυνατότητες τους. Συγκεκριμένα έχουν εφαρμοστεί στις παρακάτω έξι κατηγορίες:

- Αλγοριθμική διαπραγμάτευση (Algorithmic trading)
- Κατανομή χαρτοφυλακίου (Portfolio allocation)
- Εκτίμηση πιστωτικού κινδύνου (Credit risk assessment)

- Αποτίμηση μετοχών και παράγωγων (Asset pricing and derivatives market)
- Ανίχνευση απάτης (Fraud detection)
- Ανάλυση χρονοσειρών χρηματοοικονομικών στοιχείων (financial time series forecasting)

Ειδικότερα, η ανάλυση χρονοσειρών είναι αδιαμφισβήτητα ένα από τα πιο μελετημένα πεδία στο χώρο της χρηματοοικονομικής βιομηχανίας, σύμφωνα με την έρευνα των Sezer et al. [1]. Η ανάλυση χρονοσειρών αφορά την πρόβλεψη μίας μελλοντικής τιμής ενός χρηματοοικονομικού στοιχείου σε ένα χρονικό ορίζοντα. Για παράδειγμα, η πρόβλεψη πως θα κυμανθεί η τιμή μιας μετοχής, ενός ομόλογου ή μια συναλλαγματική ισοτιμία. Τα τελευταία χρόνια, με την έλευση των κρυπτονομισμάτων δημιουργήθηκε ένα νέο πεδίο μελέτης και έρευνας.

Το πρώτο κρυπτονόμισμα που δημιουργήθηκε ήταν το Bitcoin το 2008 από τον Nakamato [2]. Το Bitcoin ξεκίνησε σταδιακά να γίνεται γνωστό, το ενδιαφέρον για αυτό εκτινάχθηκε τον Ιανουάριο του 2017 καθώς η τιμή ενός Bitcoin από $1000 έφτασε στα $20000, τραβώντας όχι μόνο την προσοχή των ερευνητών και των επενδυτών αλλά και του ευρύτερου κοινού. Με το πέρασμα των χρόνων, εμφανίστηκαν και άλλα κρυπτονομίσματα με τον αριθμό τους να ξεπερνάει τα 1600 κρυπτονομίσματα, με τα ποιο γνωστά να είναι το Etherium, από τον Vitalik Buterin, το Lite coin και το Ripple. Οι τεράστιες αποδόσεις των κρυπτονομισμάτων και η μεγάλη διακύμανση στις τιμές τους μέσα σε σύντομα χρονικά διάστημα δημιουργεί καθημερινά πολλές ευκαιρίες για μεγάλα οφέλη. Συνεπώς, η σωστή πρόβλεψη της απόδοσης των κρυπτονομισμάτων μπορεί να αποφέρει τεράστια κέρδη.

Οι τρόποι με τους οποίους αντιμετωπίζονται τα προβλήματα των χρονοσειρών είναι δύο. Στην πρώτη κατηγορία οι προβλέψεις γίνονται με βάση την απόλυτη (πραγματική) τιμή ενός περιουσιακού στοιχείου και ανάγεται σε πρόβλημα παλινδρόμησης. Αντίθετα, στην δεύτερη κατηγορία προσπαθεί να προβλεφθεί η τάση της τιμής, δηλαδή εάν θα επέλθει ανοδική ή καθοδική πορεία και το πρόβλημα θεωρείται τύπου ταξινόμησης.

Στο πλαίσιο αυτής της διπλωματικής εργασίας εξετάζεται πόσο αποτελεσματικοί είναι οι αλγόριθμοι ταξινόμησης μηχανικής μάθησης υπό επίβλεψη στην πρόβλεψη της τιμής ενός κρυπτονομίσματος, με στόχο τη δημιουργία ενός συστήματος το οποίο με τις κατάλληλες προγραμματισμένες στρατηγικές θα εκτελεί αυτόματα ενέργειες με στόχο τη μεγιστοποίηση του κέρδους.

1.2 Διάρθρωση κείμενου

Στο δεύτερο κεφάλαιο, αναλύεται αρχικά η έννοια του χρήματος, ποιες μορφές χρήματος υπήρξαν και πως εξελίχθηκαν με την πάροδο του χρόνου. Μετά αναλύονται τι είναι τα κρυπτονομίσματα, οι βάσεις λειτουργίας τους και πως μπορούν να χρησιμοποιηθούν. Ύστερα, γίνεται μια γενική αναφορά στην τεχνολογία που κρύβεται πίσω από τα κρυπτονομίσματα, δηλαδή την αλυσίδα ομάδων συναλλαγών (Blockchain). Έπειτα, θα αναλυθεί η διαδικασία της εξόρυξης η οποία συντηρεί όλο το δίκτυο του Blockchain. Και τέλος, γίνεται λόγος για τον τρόπο διαφύλαξη και αποθήκευση τους.

Στο τρίτο κεφάλαιο, αναλύεται το πεδίο της μηχανική μάθηση και δίνεται έμφαση στους αλγορίθμους μηχανικής μάθησης δυαδικής ταξινόμησης. Έπειτα, παρουσιάζονται αλγόριθμοι που χρησιμοποιήθηκαν. Τέλος, αναφέρονται οι διάφοροι τρόποι αξιολόγησης των μοντέλων που έχουν δημιουργηθεί.

Στο τέταρτο κεφάλαιο, παρουσιάζεται το πρακτικό μέρος της εργασίας. Αναλύονται τα βήματα που εκτελέστηκαν για την υλοποίηση ενός συστήματος το οποίο θα διαπραγματεύεται αυτόματα τις συναλλαγές με κρυπτονομίσματα, ξεκινώντας από την επεξεργασία των δεδομένων, την εκπαίδευση των αλγορίθμων, καταλήγοντας στην αξιολόγηση της λειτουργίας του συστήματος. Τέλος, παρουσιάζονται τα αποτελέσματα του κάθε αλγόριθμου και συγκρίνονται μεταξύ τους.

Το πέμπτο και τελευταίο κεφάλαιο είναι ο επίλογος της εργασίας και αναφέρονται μελλοντικές προτάσεις και ιδέες για περαιτέρω έρευνα.

Κεφάλαιο 2ο

Κρυπτονομίσματα και Blockchain

2.1. Τι είναι το χρήμα

Το χρήμα ως υλικό προϊόν είναι σχεδόν σε όλους προσβάσιμο και αποτελεί τον τρόπο διευκόλυνσης της απόκτησης αγαθών. Ο ορισμός που είναι αποδεκτός από την ακαδημαϊκή κοινότητα αναφέρει πως το χρήμα πρέπει να ικανοποιεί τρεις λειτουργίες: Αποτελεί **μέσο συναλλαγής**, **αποθήκευση αξίας** και **μονάδα μέτρησης**.

Το **μέσο συναλλαγής** είναι ο μηχανισμός με τον οποίο μπορείς κάποιος να αγοράσει κινητή, ακίνητη ή άυλη περιουσία. Ένα καλό μέσο συναλλαγής, πρέπει να είναι κοινώς αποδεκτό, όχι απαραίτητα παγκοσμίως, καθώς κάτι τέτοιο θα ήταν ανέφικτο, αλλά έστω για την αγορά στην οποία γίνεται χρήση.

Η **αποθήκευση αξίας** σημαίνει ότι σε βραχυπρόθεσμη βάση τα χρήματα θα πρέπει να έχουν την ίδια αξία με την σημερινή. Για να θεωρηθεί κάτι ως καλό μέσο αποθηκευτικής αξίας θα πρέπει στο μέλλον να έχει την ίδια αξία, ώστε να αγοράζει τα ίδια αγαθά και υπηρεσίες με αυτά που θα αγόραζε και σήμερα.

Ο όρος **μονάδα μέτρησης** νοείται ως μονάδα συναλλαγής, δηλαδή για να συγκριθούν ως αξίες δύο διαφορετικά αντικείμενα. Μια καλή μονάδα μέτρησης πρέπει να είναι αποδεκτή ή εύκολα κατανοήσιμη για την αξιολόγηση περιουσιακών στοιχείων.

Συνοπτικά για να έχουν τα χρήματα κάποια αξία πρέπει:

- Να βρίσκονται στην κατοχή πολλών ανθρώπων
- Να γίνεται δεκτό ως μορφή πληρωμής από προμηθευτές
- Να χαίρει εμπιστοσύνης από την κοινωνία και η αξία να γίνεται αποδεκτή.

2.2 Μορφές χρήματος

Πριν εμφανιστούν τα «σύγχρονα» χρήματα οι άνθρωποι αποκτούσαν τα αγαθά τους με άλλους τρόπους. Από τους παλαιότερους τρόπους αποτελεί το παζάρεμα (barter), με το οποίο κάποιος μπορούσε να αποκτήσει ένα αντικείμενο, με λίγα λόγια η ανταλλαγή ενός αγαθού έναντι κάποιου άλλου (ανταλλακτική οικονομία). Η διαδικασία αυτή ήταν αρκετά επίπονη, καθώς ήταν εξαιρετικά δύσκολο να βρεθεί κάποιος ο οποίος προσέφερε ένα συγκεκριμένο αγαθό για την απόκτηση ενός άλλου. Οι δυσκολίες αυτές άρχισαν να λύνονται με το εμπορευματικό χρήμα (commodity money). Τα σιτηρά και τα μέταλλα για παράδειγμα, ήταν ένα είδος εμπορευματικού χρήματος, όπου αφενός είχε εγγενή αξία,

ενώ τα μέταλλα είχαν εξωγενή αξία. Το εμπορευματικό χρήμα είχε το πλεονέκτημα της σταθερής και γνώστης αξίας και ήταν σχετικά εύκολο να διατηρηθεί ή να ξοδευθεί.

Το εμπορευματικό χρήμα άρχισε να αντικαθιστάται από το αντιπροσωπευτικό χρήμα. Στο αντιπροσωπευτικό χρήμα η αξία του πηγάζει από την κατοχύρωση ενός αντικείμενου σε κάποιο τρίτο πρόσωπο, όπως για παράδειγμα, η έκδοση χειρόγραφης απόδειξης από έναν χρυσοχόο ότι διατηρεί κάποια ποσότητα χρυσού. Αυτή η απόδειξη είχε ως χαρακτηριστικό ότι η αξία της ήταν εύκολη να μεταφερθεί σε κάποιον άλλο ενδιαφερόμενο. Ωστόσο υπήρχε το ρίσκο κατά πόσο έμπιστο είναι το τρίτο άτομο που είχε στην κατοχή του το αναφερόμενο αγαθό.

Η επόμενη μορφή και η επικρατούσα στην σημερινή εποχή είναι το παραστατικό χρήμα. Το παραστατικό χρήμα έχει αξία επειδή ορίζεται από ένα αποδεκτό ρυθμιστικό/κανονιστικό πλαίσιο και όχι επειδή διαθέτει κάποια εγγενή αξία. Εναλλακτικά, θα μπορούσε κανείς να πει ότι το παραστατικό χρήμα έχει κάποια αξία για τους δύο παρακάτω λόγους:

1. Έχει οριστεί ως νόμιμο χρήμα, το οποίο συνεπάγεται ότι σε νόμιμες συναλλαγές πρέπει να είναι αποδεκτό ως πληρωμή σε χρέη.
2. Οι κυβερνήσεις δέχονται ως πληρωμή μόνο το δικό τους παραστατικό χρήμα για την πληρωμή των φόρων.

Το παραστατικό χρήμα στη σύγχρονη εποχή για την πραγματοποίηση των συναλλαγών έχει διάφορες μορφές οι οποίες αναφέρονται παρακάτω:

- Τα νομίσματα, τα οποία ήταν η πιο διαδεδομένη μορφή στην ιστορία του παραστατικού χρήματος
- Τα χαρτονομίσματα
- Το πιστωτικό χρήμα
- Οι χρεωστικές κάρτες
- Κρυπτονομίσματα

Τα κρυπτονομίσματα, εάν και δεν έχουν γίνει ακόμα αποδεκτά ως νόμιμο μέσω πραγματοποίησης συναλλαγών, έχουν αρχίσει να γίνονται ιδιαίτερα αποδεκτά από την κοινότητα και μεγάλες πολυεθνικές έχουν αρχίσει να τα δέχονται ως μέσο πληρωμής.

2.3 Κρυπτονομίσματα

Τα κρυπτονομίσματα είναι ένα νέο είδος ψηφιακού χρήματος. Αυτά θα μπορούσαν να θεωρηθούν περισσότερο ως άυλα ηλεκτρονικά περιουσιακά στοιχεία παρά ως ψηφιακά νομίσματα. Στο παρελθόν, είχαν γίνει προσπάθειες για την δημιουργία ψηφιακών χρημάτων, όμως το πρώτο που κατάφερε να εδραιωθεί είναι το Bitcoin. Το Bitcoin μέχρι και το 2011 ήταν το μοναδικό κρυπτονόμισμα. Αρκετοί υποστηρικτές του Bitcoin άρχισαν

να παρατηρούν ατέλειες σε αυτό και έτσι αποφάσισαν να δημιουργήσουν τα δικά τους εναλλακτικά νομίσματα (altcoins), με σκοπό να βελτιώσουν τα ελαττώματα αυτά, όπως την ταχύτητα των συναλλαγών, την ανωνυμία, την ασφάλεια στο σύστημα και αλλά τεχνικά ζητήματα. Τα Bitcoins και εν γένει τα κρυπτο-νομίσματα, στηρίζονται στην τεχνολογία Blockchain και ήταν από τις πρώτες εφαρμογές που αναπτύχθηκαν πάνω σε αυτήν. Επειδή υπάρχουν πολλά διαφορετικά κρυπτονομίσματα και το καθένα στηρίζεται σε διαφορετικούς κανόνες και βάσεις λειτουργίας, δεν είναι εύκολο να γίνουν ακριβείς γενικεύσεις για το κάθε κρυπτονόμισμα. Για το λόγο, θα γίνεται περισσότερο αναφορά στα χαρακτηριστικά και τη λειτουργία του Bitcoin στο οποίο βασίστηκαν τα περισσότερα κρυπτονομίσματα.

Η ιδιοκτησία των Bitcoins καταγράφεται σε ένα ηλεκτρονικό «λογιστικό» (ledger) βιβλίο το οποίο ανανεώνεται στιγμιαία, καθώς είναι συνδεδεμένο σε ένα τεράστιο δίκτυο ανεξάρτητων υπολογιστών, οι οποίοι είναι διασπαρμένοι σε όλο τον πλανήτη. Η λειτουργία του βασίζεται σε ένα πρωτόκολλο, δηλαδή μία λίστα από κανόνες η οποία καθορίζει πως πραγματοποιούνται οι συναλλαγές, πως θα ανανεώνονται οι εγγραφές στο λογιστικό βιβλίο και συνολικά όλη τη λειτουργία του. Αυτή η διαδικασία στηρίζεται στην τεχνολογία του blockchain και το δίκτυο αυτό «κρατιέται» ζωντανό με την βοήθεια των miners.

Στην εργασία στην οποία παρουσιάστηκε για πρώτη φορά η λειτουργία του Bitcoin από τον Satoshi Nakamoto περιγράφεται ο σκοπός του Bitcoin. Στην εργασία αυτή περιγράφεται το Bitcoin ως:

«Μια μορφή διομότιμου δικτύου (peer-to-peer) ηλεκτρονικού νομίσματος που θα επιτρέπει την άμεση πραγματοποίηση συναλλαγών μεταξύ των εμπλεκόμενων, χωρίς την ανάγκη διαπραγμάτευσης κάποιου χρηματοοικονομικού ιδρύματος.»

Από το απόσπασμα αυτό, φανερώνεται από που πηγάζει η αξία και η χρησιμότητα των κρυπτονομισμάτων. Για πρώτη φορά στην ιστορία έχει δημιουργηθεί ένα σύστημα το οποίο μπορεί να αξία σε οποιοσδήποτε σημείο του πλανήτη χωρίς τη φυσική μετακίνηση του αντικειμένου και χωρίς τη χρήση κάποιου μεσάζοντα. Αυτό είναι ένα σημαντικό και δύσκολο κατόρθωμα στην εξέλιξη των πληρωμών. Σύμφωνα με τον Lewis [3], μερικά άλλα πλεονεκτήματα των κρυπτονομισμάτων είναι:

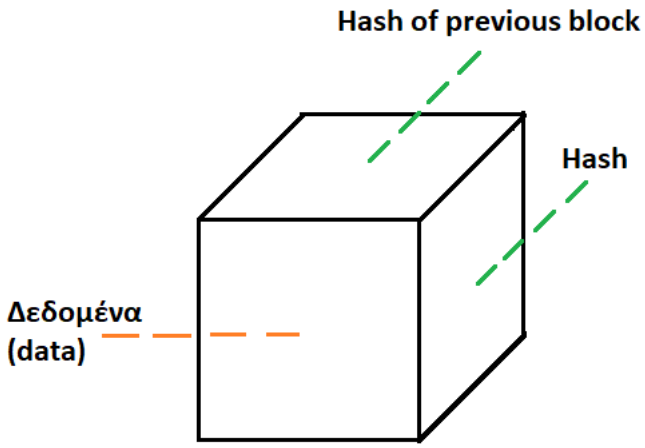
- Μείωση διαφθοράς
- Μείωση καταχρηστικής τύπωσης χρημάτων
- Ο κάτοχος είναι «τράπεζα» του εαυτού του, δηλαδή τον έλεγχο δεν τον έχει κάποια τράπεζα
- Διευκόλυνση των συναλλαγών μεταξύ μελών που δεν έχουν πρόσβαση σε παραδοσιακά συστήματα διαχείρισης συναλλαγών, όπως οι τράπεζες.

Ωστόσο, πέρα από τα θετικά στοιχεία που παρουσιάζουν, τα κρυπτονομίσματα έχουν συνδεθεί και με κάποια αρνητικά στοιχεία. Πρώτα απ’ όλα, λόγω της μεγάλης

ανωνυμίας που παρέχουν, έχουν συνδεθεί με την εγκληματικότητα καθώς στα πρώτα χρόνια τους χρησιμοποιήθηκαν για παράνομες ενέργειες. Επίσης, για πολλούς αποτέλεσε ένα μέσο φοροδιαφυγής. Τέλος, ένα από τα μεγαλύτερα μειονεκτήματά του είναι η επίπτωση στο περιβάλλον λόγω της ηλεκτρονικής εξόρυξης (mining), δηλαδή ο τρόπος με τον οποίο γίνεται να αποκτήσει κάποιος νέα κρυπτονομίσματα. Βέβαια, τα δύο πρώτα προβλήματα έχουν αντιμετωπισθεί σε μεγάλο βαθμό στα περισσότερα κράτη, μέσω της αλλαγής της νομοθεσίας. Όσο για τα περιβαλλοντικά ζητήματα που προκύπτουν από την εξόρυξη των κρυπτονομισμάτων, έχουν αρχίσει να αντιμετωπίζονται σταδιακά καθώς η νομοθεσία σε διάφορα κράτη έχει αρχίσει να προσαρμόζεται κατάλληλα, μειώνοντας έτσι ή ακόμα και απαγορεύοντας εντελώς την εξόρυξη των κρυπτονομισμάτων.

2.4 Η τεχνολογία του blockchain

Το blockchain θα μπορούσε να θεωρηθεί ως μια ειδική βάση δεδομένων όπου δεν βρίσκεται σε ένα κεντρικό σημείο αλλά κατανεμημένη σε όλο τον πλανήτη. Ειδικότερα, είναι μία αποκεντρωμένη τεχνολογία δικτύων, στην οποία τα ίδια δεδομένα καταγράφονται και διατηρούνται σε πολλαπλούς κόμβους (blocks). Η αλυσίδα των κόμβων αναφέρεται ως ένα ηλεκτρονικό λογιστικό βιβλίο (ledger) το οποίο μπορεί να καταγράψει οποιαδήποτε πληροφορία με τρόπο τέτοιο, που η αποθηκευόμενη πληροφορία δεν μπορεί να διαφθαρεί. Συγκεκριμένα, ο κάθε κόμβος θα μπορούσε να θεωρηθεί ως μια σελίδα του λογιστικού βιβλίου. Οι κόμβοι είναι συνδεδεμένοι μεταξύ τους με έναν σειριακό τρόπο σχηματίζοντας μια συνεχή ευθεία γραμμή, δηλαδή μια αλυσίδα από κόμβους. Ο κάθε κόμβος έχει τρία χαρακτηριστικά όπως φαίνεται στο σχήμα 2.1.



Σχήμα 2.1: Αναπαράσταση ενός κόμβου (block) με τα χαρακτηριστικά του.

Δεδομένα (data): Η μορφή των δεδομένων εξαρτάται από τον σκοπό της χρήσης του Blockchain. Στα κρυπτονομίσματα χρησιμοποιείται για να αποθηκεύσει πληροφορίες σχετικά με τις επιβεβαιωμένες συναλλαγές, δηλαδή τον αποστολέα, τον παραλήπτη, το ποσό που στάλθηκε και άλλες σχετικές πληροφορίες

Hash: Το hash στην αλυσίδα των κόμβων είναι το δακτυλικό αποτύπωμα του κόμβου, δηλαδή είναι κάτι μοναδικό το οποίο χρησιμοποιείται για να ξεχωρίζεται ο κάθε κόμβος από τον άλλον. Η μοναδικότητα αυτή διασφαλίζεται από την εφαρμογή των κρυπτογραφικών Hash συναρτήσεων.

Το Hash του προηγούμενου κόμβου: Κάθε κόμβος περιέχει το hash του προηγούμενου κόμβου, το οποίο βοηθάει στην ασφάλεια του δικτυού και είναι ένας από τους λόγος που θεωρείται αμετάτρεπτο και «απαραβίαστο».

Παρακάτω αναφέρεται ενδεικτικά ένα απλοποιημένο παράδειγμα της λειτουργίας ενός Blockchain τριών κόμβων.

Κόμβος #1:

Δεδομένα: Αποστολή 10 Bitcoins από την Αλίκη στον Μάκη

Hash: 12A

Προηγούμενο Hash: 000

Κόμβος #2:

Δεδομένα: Αποστολή 158 Bitcoins από Τάκη στην Μαίρη

Hash: 3B4

Προηγούμενο Hash: 12A

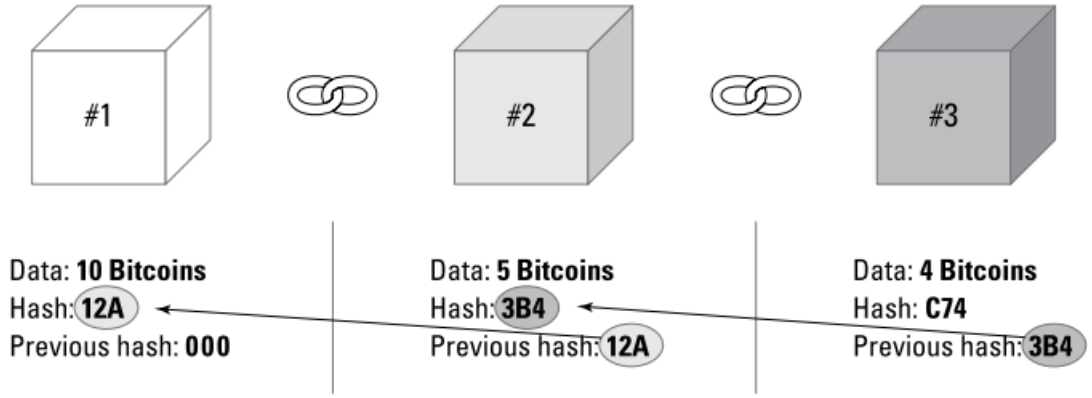
Κόμβος #3:

Δεδομένα: Αποστολή 40 Bitcoins από την Αφροδίτη στον Λάκη

Hash: C74

Προηγούμενο Hash: 3B4

Όπως φαίνεται και στο σχήμα 2.2 ο κάθε κόμβος έχει το δικό του μοναδικό hash και το hash του προηγούμενου κόμβου. Συνεπώς, ο κόμβος 3 δείχνει στον κόμβο 2 και ο κόμβος 2 δείχνει στον κόμβο 1, δημιουργώντας έτσι την αλυσίδα των κόμβων.



Σχήμα 2.2: Απλοποιημένη μορφή λειτουργίας ενός Blockchain.

2.5 Εξόρυξη (mining)

Η εξόρυξη είναι η διαδικασία με την οποία ελέγχονται, κατοχυρώνονται, αποθηκεύονται οι συναλλαγές και δημιουργούνται νέα κρυπτονομίσματα. Αυτοί που συμμετέχουν στην εξόρυξη, επειδή συνεισφέρουν τους υπολογιστικούς πόρους τους για την ύπαρξη του δικτύου, συλλέγουν ένα μικρό μερίδιο από τις συναλλαγές ως επιβράβευση για την συνεισφορά αυτή.

Στην ουσία η διαδικασία της εξόρυξης είναι η λύση ενός κρυπτογραφικού πάζλ (puzzle) με την βοήθεια του ηλεκτρονικού υπολογιστή. Ο υπολογιστής πρέπει να «μαντέψει» έναν τυχαίο αριθμό, που ο αριθμός αυτός λύνει μια εξίσωση. Η εξίσωση αυτή παράγεται από τον κώδικα του Blockchain. Όποιος καταφέρει να βρει πρώτος την λύση επιβραβεύεται από το Blockchain με νέα κρυπτονομίσματα. Όσες περισσότερες προβλέψεις μπορεί να κάνει το δευτερόλεπτο ο ηλεκτρονικός υπολογιστής, αυξάνονται οι πιθανότητες να βρεθεί η λύση της εξίσωσης. Συνεπώς, αυτός που έχει στην κατοχή του μεγάλη υπολογιστική ισχύ έχει πλεονέκτημα, καθώς αυξάνονται οι πιθανότητες επιτυχίας. Για το λόγο αυτό, μεγάλες πολυεθνικές έχουν φτιάξει ειδικές συσκευές για την αποτελεσματικότερη εξόρυξη κρυπτονομισμάτων. Ωστόσο είναι στατιστικά σχεδόν απίθανο να κερδίζει συνέχεια κάθε φόρα αυτός που έχει την μεγαλύτερη υπολογιστική δύναμη σε ένα παιχνίδι «μαντεψιάς». Συγκεκριμένα, αυτό το είδος εξόρυξης ονομάζεται «απόδειξη της δουλειάς» (Proof-of-work) και πάνω σε αυτό στηρίζεται το Bitcoin. Η διαδικασία αυτή, όπως έχει αναφερθεί ήδη, έχει και κάποια μειονεκτήματα. Ένα από αυτά είναι πως γίνεται μεγάλη σπατάλη υπολογιστικών πόρων και ηλεκτρικής ενέργειας απλά και μόνο για την παραγωγή τυχαίων αριθμών. Για τον προαναφερόμενο λόγο άλλα κρυπτονομίσματα έχουν στραφεί σε διαφορετικές μεθόδους εξόρυξης κρυπτονομισμάτων, με το πιο διαδεδομένο να είναι η «απόδειξη μεριδίου:» (Proof-of-stake).

Το Proof-of-stake απαιτεί την κατοχή ενός συγκεκριμένου αριθμού νομισμάτων για την εξόρυξη. Αυτό σημαίνει ότι όσα περισσότερα κρυπτονομίσματα διαθέτει κάποιος, τόσο περισσότερη δυναμική θα έχει στην εξόρυξη. Αυτή η προσέγγιση αποβάλει την ανάγκη για ακριβό εξοπλισμό σε αντίθεση με το Proof-of-work. Ωστόσο, το Proof-of-stake έχει και αυτό τις αδυναμίες του. Αυτοί που συμμετέχουν στην εξόρυξη ενός κόμβου, μπορούν να αποκτήσουν μόνο ένα ποσοστό των συναλλαγών το οποίο αντιστοιχεί στο ποσοστό του μεριδίου τους. Για παράδειγμα, κάποιος που διαθέτει το 10% ενός νομίσματος θα είναι ικανός να εξορύξει μόνο το 10% των κόμβων του δικτύου. Αυτός ο περιορισμός δίνει στους χρήστες ένα λόγο στο να κρατήσουν τα νομίσματα παρά να τα ξοδέψουν, οδηγώντας έτσι σε ένα σενάριο όπου ο «πλούσιος» θα γίνει «πλουσιότερος» λόγω της ικανότητας του να εξορύσσει το μεγαλύτερο ποσοστού του δικτύου.

Τέλος κλείνοντας αυτή την ενότητα πρέπει να αναφερθεί ότι κάποια κρυπτονομίσματα δεν χρησιμοποιούν διαδικασίες εξόρυξης όπως το Ripple (XPR) για να αποφύγουν την άσκοπη κατανάλωση ενέργειας όπως γίνεται σε άλλα δίκτυα.

2.6 Κρυπτο-πορτοφόλια

Ένα κρυπτο-πορτοφόλι είναι ένα πρόγραμμα λογισμικού το οποίο βοηθάει στην διαχείριση των κρυπτο-νομισμάτων. Χωρίς την κατοχή ενός τέτοιου πορτοφολιού δεν είναι δυνατή η απόκτηση κρυπτονομισμάτων. Τα κρυπτο-πορτοφόλια δεν αποθηκεύουν το ποσό των νομισμάτων που έχει ο κάτοχος του αλλά το δημόσιο κλειδί (Public key) και το ιδιωτικό κλειδί (private key).

Δημόσιο κλειδί: Ένας κωδικός ο οποίος επιτρέπει την λήψη κρυπτο-νομισμάτων από άλλους στο λογαριασμό του κάτοχου

Ιδιωτικό κλειδί: Ένας κωδικός ο οποίος συνδέεται με το δημόσιο κλειδί και χρησιμοποιείται για την διασφάλιση της ασφάλειας. Θα μπορούσε να παρομοιαστεί με τον κωδικό που χρησιμοποιείται για την πρόσβαση στον λογαριασμό μίας τράπεζες.

Ένα παράδειγμα ενός δημόσιου κλειδιού είναι ο αλφαριθμητικός κωδικός: **1A1zP1eP5QGaek2DMPTfTL5SLmv7DivfNa**, ενώ ο κωδικός ενός ιδιωτικού κλειδιού είναι λίγο περισσότερο περίπλοκος για την διασφάλιση της ασφάλειας: **03bf350d2821375158a608b51oly3898e507fe47f2d2e8c774de4a9a7paocf74eda**

Όπως γίνεται φανερό και στις δύο περιπτώσεις οι κωδικοί αυτοί αποτελούνται από έναν συνδυασμό αριθμών και χαρακτήρων, κεφαλαίων και μη. Αν και αυτές οι δύο διευθύνσεις μπορεί να φαίνονται εντελώς διαφορετικές και ανεξάρτητες στο μάτι, το λογισμικό που χρησιμοποιείται μπορεί να αναγνωρίσει ότι αυτοί οι δύο κωδικοί είναι στενά συνδεδεμένοι μεταξύ τους, το οποίο αποδεικνύει την κυριότητα του λογαριασμού.

Τα κρυπτο-πορτοφόλια έρχονται σε διάφορα είδη, όπως και στην πραγματικότητα υπάρχουν μεγάλα, μικρά, δερμάτινα έτσι και τα ηλεκτρονικά πορτοφόλια έχουν διαφορετικές ιδιότητες ανάλογα την χρήση τους. Παρακάτω θα αναλυθούν τα πιο γνωστά είδη κρυπτο-πορτοφολιών με τις δυνατότητες και τις αδυναμίες τους.

Διαδικτυακά Πορτοφόλια (Online wallet)

Τα διαδικτυακά πορτοφόλια δεν θεωρούνται τα πιο ασφαλή πορτοφόλια αλλά έχουν αρκετά πλεονεκτήματα για μικρούς λογαριασμούς. Αυτά επιτρέπουν την πρόσβαση στα κρυπτονομίσματα μέσω του διαδικτύου. Επομένως, όσο υπάρχει σύνδεση στο διαδίκτυο είναι εφικτή η αποθήκευση των νομισμάτων αυτών και γενικά οποιαδήποτε άλλη ενέργεια που σχετίζεται με πληρωμές. Τα πλεονεκτήματα τους περιλαμβάνουν τα παρακάτω:

- Είναι γρήγορα στις συναλλαγές
- Μπορούν να δεχτούν διαφορετικά είδη κρυπτονομισμάτων
- Είναι βολικά στην χρήση και εύκολα για τις συχνές συναλλαγές (active trading).

Τα μειονεκτήματα τους περιλαμβάνουν:

- Είναι εκτεθειμένα στον κίνδυνο ηλεκτρονικής απάτης (hacking) καθώς είναι συνδεδεμένα στο διαδίκτυο
- Η αποθήκευση τους πραγματοποιείται από $3^{ο}$ πρόσωπο

Τηλεφωνικό πορτοφόλι (Mobile wallet)

Αυτό το είδος πορτοφολιού λειτουργεί με το κινητό τηλέφωνο μέσω μιας εφαρμογής. Είναι παρόμοια σαν τα διαδικτυακά πορτοφόλια αλλά είναι φτιαγμένα μόνο για τα κινητά και προσφέρουν κάποια επιπλέον χαρακτηριστικά.

- Είναι ασφαλέστερα από τα διαδικτυακά πορτοφόλια (συνήθως)
- Είναι βολικά για χρήση «εν κίνηση».
- Προσφέρουν κωδικούς quick respone (QR) code scanning

Ενώ τα μειονεκτήματα είναι:

- Εάν καταστραφεί το κινητό υπάρχει μεγάλη πιθανότητα να χαθούν τα κρυπτονομίσματα
- Είναι ευάλωτα σε κακόβουλα λογισμικά

Offline πορτοφόλια

Στα desktop πορτοφόλια υπάρχει η δυνατότητα, κάποιος να τα κατεβάσει από το διαδίκτυο και να τα εγκαταστήσει στον ηλεκτρονικό του υπολογιστή.

Μερικά πλεονεκτήματα των desktop πορτοφολιών αναφέρονται παρακάτω:

- Η αποθήκευση του ιδιωτικού κλειδιού δεν γίνεται σε τρίτο πρόσωπο
- Εάν ο υπολογιστής δεν είναι συνδεδεμένος στο διαδίκτυο μπορεί να θεωρηθεί ασφαλέστερο από τα αλλά πορτοφόλια

Ενώ τα μειονεκτήματα :

- Δεν είναι βολικά στην χρήση για συναλλαγές στην καθημερινότητα
- Εάν χαλάσει ο υπολογιστής χάνονται τα κρυπτονομίσματα σε περίπτωση που δεν έχει αποθηκευτεί κάποιο αντίγραφο ασφαλείας.

Hardware wallet

Ένα τέτοιο πορτοφόλι αδιαμφισβήτητα είναι από τα ασφαλέστερα μέσα για την διαφύλαξη των κρυπτονομισμάτων. Τέτοιου είδους πορτοφόλια διαφυλάσσουν τα ιδιωτικά κλειδιά σε ένα φορητό αποθηκευτικό μέσο. Το πλεονέκτημα τους είναι ότι είναι κατάλληλα για την αποθήκευση μεγάλων ποσών κρυπτονομισμάτων λόγω της ασφάλειας που προσφέρουν, ενώ τα μειονεκτήματα είναι πως είναι τα πιο ακριβά πορτοφόλια και δεν είναι ιδιαίτερα φιλικά στην χρήση τους, ειδικά σε αρχάριους.

Paper wallet

Σε αυτό το είδος πορτοφολιού το δημόσιο και ιδιωτικό κλειδί εκτυπώνεται σε ένα χαρτί. Είναι το πιο ασφαλές πορτοφόλι καθώς τα κλειδιά του δεν είναι αποθηκευμένα σε κάποια ηλεκτρονική συσκευή (υπολογιστής, κινητό τηλέφωνο ή κάποιο τρίτο πρόσωπο). Μερικά από τα μειονεκτήματα είναι πως:

- Είναι αρκετά δύσκολο να χρησιμοποιηθούν στην καθημερινότητα και όχι βολικά για καθημερινές συναλλαγές
- Εάν χαθεί ή καταστραφεί το χαρτί δεν γίνεται ανάκτηση του λογαριασμού

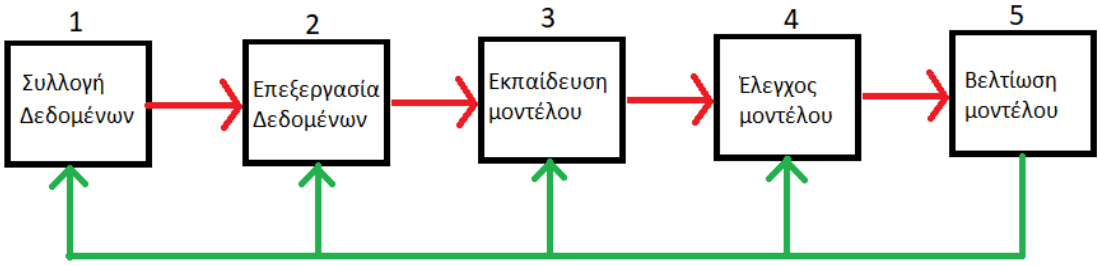
Κεφάλαιο 3ο

Μηχανική μάθηση με επίβλεψη

3.1 Εισαγωγή

Θα πρέπει να αποσαφηνιστεί ότι εν λόγω εργασία , όπως αναφέρθηκε στο πρώτο κεφάλαιο, εξετάζει μόνο αλγορίθμους δυαδικής ταξινόμησης με μηχανική μάθηση υπό επίβλεψη οπότε οποιαδήποτε αναφορά σε αλγορίθμους μηχανικής μάθησης θα αναφέρεται αποκλειστικά και μόνο σε αυτούς.

Η μηχανική μάθηση είναι η ικανότητα των υπολογιστών να μαθαίνουν από τα δεδομένα. Στην πράξη η μηχανική μάθηση πρόκειται για υπολογιστικούς αλγορίθμους οι οποίοι τροποποιούν τα βάρη μιας παραμετρικής εξίσωσης ανάλογα με τα δεδομένα εισόδου. Στο σχήμα 3.1 φαίνονται τα βήματα που πραγματοποιούνται στην μοντελοποίηση ενός προβλήματος μηχανικής μάθησης.



Σχήμα 3.1: Τυπική Διαδικασία ενός προβλήματος μηχανικής μάθησης

Το πρώτο στάδιο είναι η συλλογή των δεδομένων. Αυτό το στάδιο θα μπορούσε να θεωρηθεί ως ο ακρογωνιαίος λίθος για την επιτυχία ενός προβλήματος μηχανικής μάθησης καθώς, το σημαντικότερο συστατικό για να είναι αποτελεσματικός ένας αλγόριθμος είναι η ποιότητα των δεδομένων. Επίσης, πέρα από την ποιότητα των δεδομένων, αρκετά σημαντικό βαρόμετρο είναι και η πολυπλοκότητα των μοντέλων την οποία μπορεί να αναπαραστήσει ο αλγόριθμος που χρησιμοποιείται. Για να διευκρινιστεί πως τα δεδομένα και η πολυπλοκότητα ενός μοντέλου επηρεάζουν διαφορετικά την αποτελεσματικότητα του αλγορίθμου, θα χρησιμοποιηθεί μια αναλογία. Η ποιότητα των δεδομένων είναι σαν τα καύσιμα ενός αυτοκινήτου, όπου στην προκειμένη περίπτωση το αυτοκίνητο παρομοιάζει την πολυπλοκότητα του μοντέλου. Όσο φανταχτερό και μεγάλη ιπποδύναμη και να διαθέτει το αυτοκίνητο, εάν τα καύσιμα είναι χαμηλής ποιότητας τότε, παρά τις τεράστιες δυνατότητες του οχήματος, αυτό θα υπολειτουργεί και δεν θα αξιοποιείται σωστά. Αντίθετα, στην περίπτωση όπου τα καύσιμα είναι εξαίρετης ποιότητας (τα δεδομένα) αλλά το διαθέσιμο αυτοκίνητο είναι χαμηλής ιπποδύναμης τότε, τα καύσιμα δεν θα

εκμεταλλευτούν πλήρως. Συνεπώς, η σωστή συλλογή δεδομένων και η κατάλληλη επιλογή μοντέλου είναι ο κυρίαρχος παράγοντας για να δημιουργηθεί ένα αποτελεσματικό σύστημα.

Το δεύτερο στάδιο είναι η επεξεργασία των δεδομένων. Σε πολλές περιπτώσεις τα δεδομένα χρειάζονται κάποια επεξεργασία καθώς πολλές φορές η αποθηκευμένη τους μορφή δεν είναι εύχρηστη. Συγκεκριμένα, τα δεδομένα πρέπει να είναι της μορφής {**X,y**}, δηλαδή ένα διατεταγμένο ζεύγος από δεδομένα εισόδου **X** και από το διάνυσμα εξόδου **y** όπως φαίνεται στον πίνακα 3.1. Το **X** μπορεί να είναι από διάνυσμα έως τένσορας ανάλογα το είδος του προβλήματος, αλλά το πιο συνηθισμένο είναι πίνακας.

Πίνακας 3.1: Αναπαράσταση ενός συνόλου δεδομένων

x_1	x_2	x_3	x_4	x_5	**y (targets)**
2327	21879	89	78.43	4.9	**0**
4536	20430	809	76.5	3.4	**1**
1709	19756	89	76.5	2.5	**1**
3999.4	19236	87	77	45.6	**1**
3982.2	18733	100.5	80.4	47.8	**0**
8905.5	22839	86	81.3	39.8	**0**

Στον πίνακα 3.1 οι στήλες του πίνακα **X** ονομάζονται χαρακτηριστικά (attributes) και η κάθε στήλη (χαρακτηριστικό) είναι η μεταβλητή εισόδου (ανεξάρτητη μεταβλητή). Για παράδειγμα, το x_1 θα μπορούσε να συμβολίζει την τιμή κλεισίματος του Bitcoin, το x_2 την μέγιστη τιμή του στο χρονικό διάστημα που εξετάζεται, δηλαδή τα χαρακτηριστικά είναι διάφορες μετρήσεις οι οποίες αφορούν το υπό εξέταση στοιχείο. Αντίθετα, η στήλη **y**, η εξαρτημένη μεταβλητή, ονομάζεται στόχος (target) και οι τιμές οι οποίες παίρνει είναι 0 ή 1, οι οποίες μπορούν να κωδικοποιηθούν ως όχι – 0 ή ναι – 1. Στο πρόβλημα της εργασίας ερμηνεύεται ως «πώληση» (0) ή «αγορά» (1). Η κάθε γραμμή του πίνακα 3.1 ονομάζεται παράδειγμα (instance), επειδή μέσω αυτών των παραδειγμάτων ο αλγόριθμος εκπαιδεύεται.

Το τρίτο στάδιο αφορά τη διαδικασίας της εκπαίδευσης. Κατά τη διαδικασία της εκπαίδευσης χρησιμοποιούνται δεδομένα όπου η έκβαση των αποτελεσμάτων είναι γνωστή. Αυτό επιτρέπει στον αλγόριθμο να μάθει από ολοκληρωμένα παραδείγματα και να προσαρμοστεί κατάλληλα στο πρόβλημα. Με την εμπειρία που θα αποκτήσει ο αλγόριθμος μετά την εκπαίδευση του, θα μπορέσει να δώσει ικανοποιητικές εκτιμήσεις σε νέα παραδείγματα, που η έκβαση τους δεν είναι γνωστή. Ουσιαστικά, η διαδικασία της εκπαίδευσης είναι καθαρά μία υπολογιστική διαδικασία. Στους περισσότερους αλγορίθμους ο σκοπός είναι η ελαχιστοποίηση μια συνάρτησης, συχνά αποκαλούμενη συνάρτηση κόστους. Η συνάρτηση κόστους ελαχιστοποιείται χρησιμοποιώντας αλγόριθμους βελτιστοποίησης με τον πιο γνωστό της επικλινούς καθόδου (gradient descent). Συνεπώς, η εκπαίδευση ενός αλγορίθμου πρόκειται για την αναζήτηση των

παραμέτρων του μοντέλου πρόβλεψης, τα οποία ελαχιστοποιούν τη συνάρτηση κόστους, δηλαδή το σφάλμα μεταξύ της πραγματικής τιμής **y** και της πρόβλεψης $\hat{\boldsymbol{y}}$.

Το τέταρτο στάδιο σχετίζεται με την μέτρηση της αποτελεσματικότητας του μοντέλου που αναπτύχθηκε ύστερα από την εκπαίδευσή του. Υπάρχουν αρκετοί διαφορετικοί τρόποι για τη μέτρηση της αποτελεσματικότητας, η οποία κυρίως αναφέρεται στην ακρίβεια των προβλέψεων του μοντέλου.

Το πέμπτο και τελευταίο στάδιο είναι η βελτίωση συνολικά του μοντέλου. Είναι καθαρά μια διαδικασία που στηρίζεται στην εμπειρία και στη γνώση του προβλήματος που προσπαθεί να αντιμετωπιστεί.

Στις επόμενες ενότητες, που ακολουθούν, θα αναλυθεί με ποιον τρόπο ο κάθε αλγόριθμος αποδίδει εκτιμήσεις, πως εκπαιδεύεται και με ποιους τρόπους μετράται η αποτελεσματικότητα ενός μοντέλου.

3.2 Λογιστική παλινδρόμηση

Η λογιστική παλινδρόμηση ανήκει στην κατηγορία των γραμμικών μοντέλων. Τα γραμμικά μοντέλα έχουν χρησιμοποιηθεί πολύ στην πράξη και έχουν μελετηθεί εντατικά τις τελευταίες δεκαετίες, με τις ρίζες τους να ξεκινούν πριν από περίπου 100 χρόνια. Οι εκτιμήσεις της λογιστικής παλινδρόμησης δίνονται από τον τύπο (3.1).

$$\hat{\boldsymbol{y}} = \begin{cases} 0 & \alpha\nu\ \hat{p} < 0.5 \\ 1 & \alpha\nu\ \hat{p} \geq 0.5 \end{cases} \qquad (3.1)$$

Όπου το $\hat{\boldsymbol{p}}$ περιγράφεται από την παρακάτω εξίσωση (2.2):

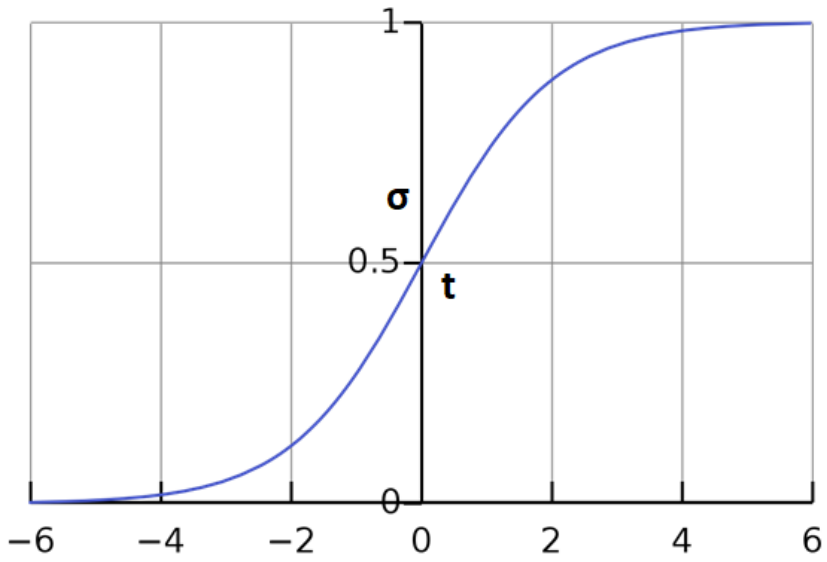
$$\hat{\boldsymbol{p}} = \sigma\left(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_j x_i + \cdots + \theta_n x_n\right) \qquad (3.2)$$

Στην εξίσωση (3.2):

- το $\hat{p}$ είναι η πιθανότητα σε ποια κλάση ανήκει το **x**
- το $\sigma(\cdot)$ είναι η σιγμοειδής συνάρτηση που περιγράφεται από την εξίσωση (3.3)
- n είναι ο αριθμός των χαρακτηριστικών
- x_i είναι η τιμή του ith χαρακτηριστικού
- θ_0 ο σταθερός όρος
- θ_j από 1 έως n οι συντελεστές παλινδρόμησης

Παρακάτω περιγράφεται η σιγμοειδής συνάρτηση και το γράφημα της φαίνεται στο σχήμα 3.2

$$\sigma(t) = \frac{1}{1+\exp(-t)} \qquad (3.3)$$



Σχήμα 3.2: Απεικόνιση της σιγμοειδής συνάρτηση

Η εξίσωση (3.2) μπορεί να γραφτεί πιο απλά σε διανυσματική μορφή:

$$\hat{p} = \sigma(\boldsymbol{\theta}^{\boldsymbol{T}}\boldsymbol{X}) \tag{3.4}$$

Στην εξίσωση (3.4) οι παράμετροι της εξίσωσης, δηλαδή ο σταθερός όρος και τα οι συντελεστές παλινδρόμησης βρίσκονται ελαχιστοποιώντας την συνάρτηση κόστους (3.5).

$$J(\boldsymbol{\theta}) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(\hat{p}^{(i)}\right) + (1-y^{(i)})\log(1-\hat{p}^{(i)})\right] \tag{3.5}$$

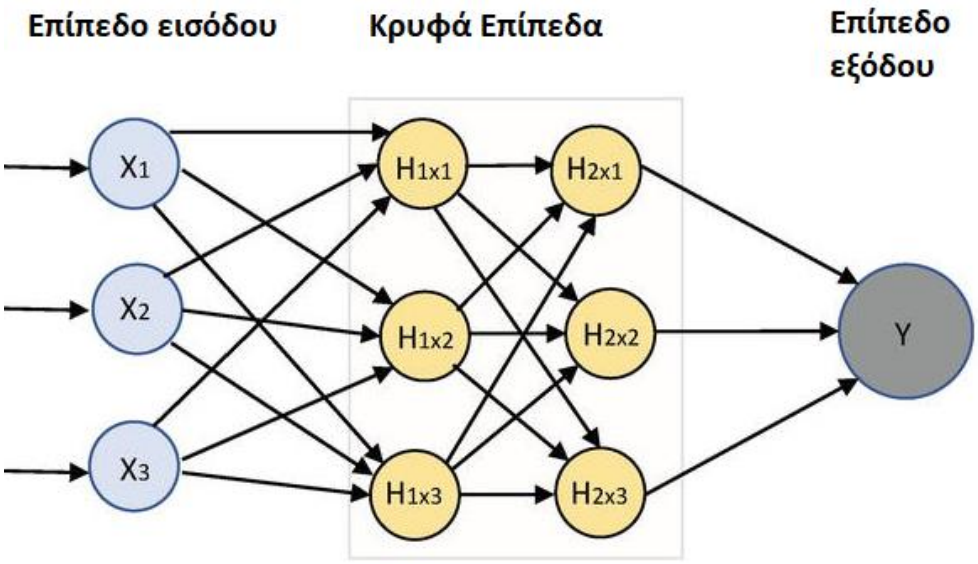
Η εξίσωση (3.5) δεν έχει κάποια κλειστού τύπου μορφή για την εύρεση των στοιχείων του διανύσματος $\boldsymbol{\theta}$ που την ελαχιστοποιεί. Ωστόσο, η συνάρτηση κόστους είναι κυρτή, οπότε αλγόριθμοι βελτιστοποίησης συγκλίνουν και μπορούν να εντοπίσουν το ολικό ελάχιστο.

3.3 Νευρωνικά δίκτυα

Μεγάλος αριθμός υπολογιστικών μεθοδολογιών είναι εμπνευσμένες από τη φύση. Ένα τέτοιο παράδειγμα είναι τα νευρωνικά δίκτυα, τα οποία είναι εμπνευσμένα από τον εγκέφαλο του ανθρώπου και την ικανότητα του να σκέφτεται, να πράττει και να λύνει προβλήματα χρησιμοποιώντας την λογική. Η πρώτη αναφορά, έγινε το 1943 από τον νεύρο-ψυχολόγο Warren McCulloch και τον μαθηματικό Walter Pitts στην δημοσίευση τους “A Logical Calculus of Ideas Immanent in Nervous Activity” [4].

Πλέον, υπάρχουν πολλά διαφορετικά είδη τεχνητών νευρωνικών δικτύων με διαφορετικές ιδιότητες και ικανότητες το καθένα. Ένα από τα πιο διαδεδομένα είδη είναι τα τεχνητά νευρωνικά δίκτυα με πρόσθια τροφοδότηση, τα οποία είναι η βάση για πιο σύνθετα μοντέλα βαθιάς μηχανικής μάθησης (deep learning). Το κύριο χαρακτηριστικό των

τεχνητών νευρωνικών δικτύων (ΤΝΔ) είναι οι τεχνητοί νευρώνες οι οποίοι πραγματοποιούν τους υπολογισμούς. Τα ΤΝΔ χωρίζονται σε 3 μέρη όπως φαίνεται στην εικόνα 3.3

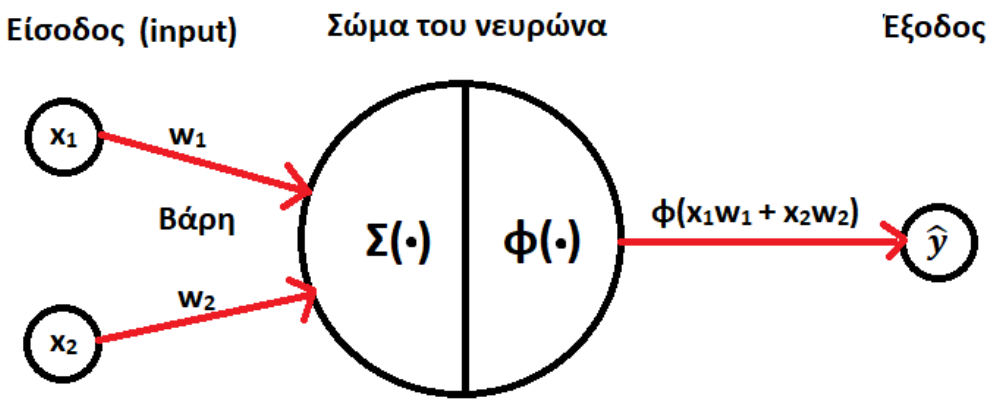


Σχήμα 3.3: Απεικόνιση ενός νευρωνικού δικτύου

Το πρώτο μέρος ονομάζεται «επίπεδο εισόδου» (input layer), καθώς αποτελείται μόνο από τα δεδομένα. Τα στοιχεία του δηλαδή, δεν είναι ουσιαστικά νευρώνες, γιατί δεν εκτελούν κάποιο υπολογισμό, απλώς χρησιμοποιούνται για την εισαγωγή των δεδομένων.

Το δεύτερο μέρος αποτελείται από τα κρυφά επίπεδα (hidden layers), των οποίων ο αριθμός καθορίζεται ανάλογα την αρχιτεκτονική του δικτύου. Το κάθε κρυφό επίπεδο αποτελείται από τεχνητούς νευρώνες που δεν υπάρχει κάποιος περιορισμός στην επιλογή του αριθμού των νευρώνων. Βέβαια, όσο μεγαλύτερος είναι ο αριθμός των κρυφών επιπέδων και των τεχνητών νευρώνων, η πολυπλοκότητα του μοντέλου αυξάνεται. Το τρίτο μέρος, το επίπεδο εξόδου, αποτελείται επίσης από τεχνητούς νευρώνες. Αυτό είναι το τελευταίο στάδιο των υπολογισμών, το οποίο δίνει την επιθυμητή απάντηση. Ωστόσο, η αρχιτεκτονική του επιπέδου εισόδου και του επιπέδου εξόδου εξαρτώνται από την δομή του προβλήματος, σε αντίθεση με την αρχιτεκτονική των κρυφών επιπέδων και την επιλογή των νευρώνων.

Όπως αναφέρθηκε προηγουμένως, οι τεχνητοί νευρώνες είναι το βασικό χαρακτηριστικό ενός τεχνητού νευρωνικού δικτύου. Ο τεχνητός νευρώνας (perceptron) είναι από τις πρώτες αρχιτεκτονικές ενός τεχνητού νευρωνικού δικτυού που προτάθηκε το 1957 από τον Frank Rosenblatt. Το σώμα του νευρώνα χωρίζεται σε δύο μέρη, την αθροιστική συνάρτηση και την συνάρτηση ενεργοποίησης, όπως απεικονίζεται στην εικόνα 3.4. Σκοπός της αθροιστικής συνάρτησης είναι απλώς, να αθροίζει τα σήματα εισόδου που δέχεται ο νευρώνας, αφού πολλαπλασιαστούν με τα αντίστοιχα βάρη. Συνεπώς, ο υπολογισμός που πραγματοποιείται είναι η εκτέλεση ενός εσωτερικού γινομένου.



Σχήμα 3.4: Σώμα ενός τεχνητού νευρώνα (Perceptron)

Η συνάρτηση ενεργοποίησης είναι συνήθως μία συνάρτηση μη-γραμμική, όπου μέσω της οποίας ο νευρώνας είναι ικανός να «μάθει» μέσα από τα δεδομένα μη-γραμμικές σχέσεις και περίπλοκα πρότυπα. Ουσιαστικά, ο τεχνητός νευρώνας είναι ένα απλό υπολογιστικό μοντέλο και περιγράφεται από την παρακάτω σχέση (3.6):

$$Output = \varphi(\sum_{i=1}^{n}(input_i \times w_i) + b) \quad (3.6)$$

Όπου:

- το $\varphi(\cdot)$ είναι η συνάρτηση ενεργοποίησης
- το *input* είναι τα δεδομένα εισόδου του νευρώνα
- το *b* ο σταθερός όρος
- το *w* τα βάρη των συνδέσεων

Οι πρώτες συναρτήσεις ενεργοποιήσεις, που άρχισαν να χρησιμοποιούνται στα τεχνητά νευρωνικά δίκτυα, είναι η λογιστική συνάρτηση (logistic), και η υπερβολική εφαπτομένη (tanh), οι οποίες περιγράφονται από τις σχέσεις (3.7) και (3.8) αντίστοιχα.

$$\varphi(x) = \frac{1}{1+e^{-\alpha x}} \quad (3.7)$$

$$\varphi(x) = \frac{e^{\alpha x}-e^{-\alpha x}}{e^{\alpha x}+e^{-\alpha x}} \quad (3.8)$$

Από τις πιο διαδεδομένες συναρτήσεις ενεργοποίησης είναι η rectified linear unit (ReLU), η οποία αντικατέστησε σε πολλές περιπτώσεις την χρήση της σιγμοειδούς συνάρτησης και δίνεται από τον τύπο (3.9)

$$\varphi(x) = \max(x, 0) \quad (3.9)$$

Επιπλέον, τα επόμενα χρόνια προτάθηκαν ορισμένες παραλλαγές της ReLU οι οποίες επίλυσαν μερικές αδυναμίες της [5]. Δύο από αυτές είναι η leaky ReLU [6] και η exponential linear unit (ELU) [7] οι οποία δίνονται από τις σχέσεις (3.10-3.11).

$$\varphi_a(x) = \max(ax, x) \tag{3.10}$$

$$\varphi_\alpha(x) = \begin{cases} x & \acute{\epsilon}\alpha\nu\ x \geq 0 \\ \alpha(e^x - 1) & \acute{\epsilon}\alpha\nu\ x < 0 \end{cases} \tag{3.11}$$

Τα ΤΝΔ θεωρούνται αρκετά ισχυρά εργαλεία και μπορούν να αποτελούνται από πολλούς τεχνητούς νευρώνες, οι οποίοι κατανέμονται σε διαφορετικά επίπεδα (layers), ανάλογα την αρχιτεκτονική του δικτύου. Μαθηματικά τα νευρωνικά δίκτυα μπορούν να περιγραφτούν από την σχέση (3.12):

$$\hat{y} = \varphi_k(\sum_{i=1}^{n}(input_k \times w_k) + b_k) \tag{3.12}$$

Όπου:

- ο δείκτης αναφέρεται στο τελευταίο επίπεδο ενώ οι υπολογισμοί στα υπόλοιποα επίπεδα γίνεται βάσει της ακόλουθης σχέσης:

$$input_k = \varphi_{k-1}(\sum_{i=1}^{n} input_{k-1} \times w_{k-1} + b_{k-1}) \tag{3.13}$$

Για πολλά χρόνια δεν υπήρχε ένας αποτελεσματικό αλγόριθμο για την εκπαίδευση ενός ΤΝΔ. Η λύση δόθηκε το 1986 όταν προτάθηκε από τους Rumelhart et al [8]. ο αλγόριθμος ανάστροφη μετάδοση σφάλματος (back-propagation training algorithm). Ο αλγόριθμος αυτός είναι μία διαδικασία βελτιστοποίησης επικλινούς καθόδου η οποία ελαχιστοποιεί συνήθως το μέσο τετραγωνικό σφάλμα, ανάλογα τη δομή του προβλήματος, όπως περιγράφεται από την σχέση (3.14):

$$J(w) = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 \tag{3.14}$$

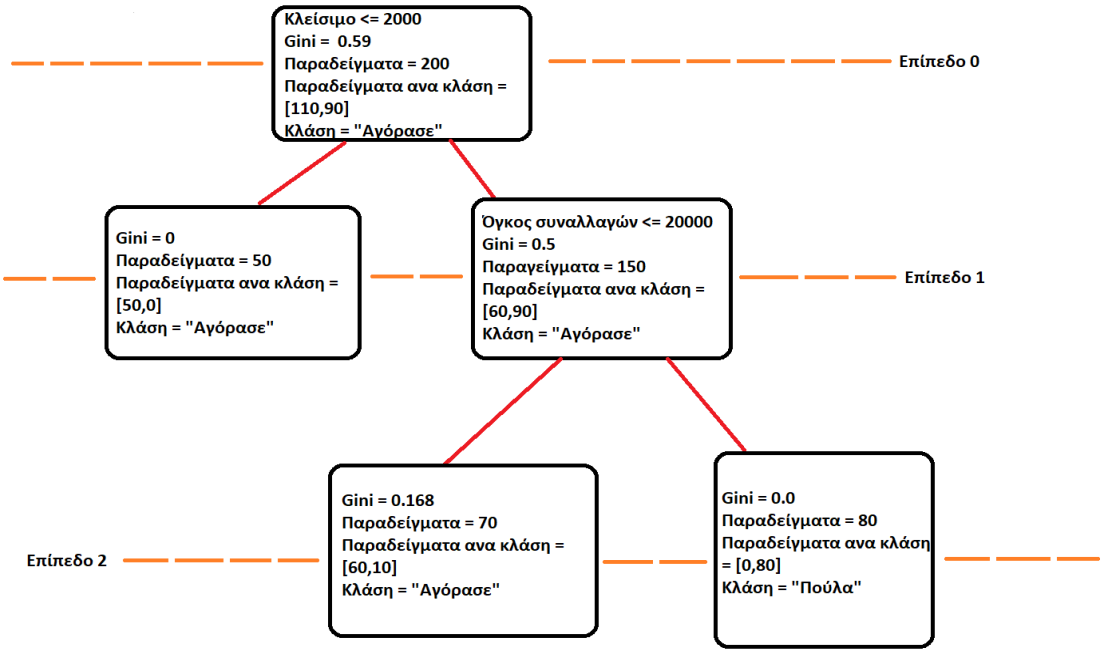
3.4 Δέντρα αποφάσεων

Τα δέντρα αποφάσεων είναι απλά μοντέλα και τα αποτελέσματα τα οποία παράγουν είναι εύκολο να ερμηνευτούν. Παράλληλα, παρά την απλότητα τους θεωρούνται ισχυροί αλγόριθμοι, καθώς προσαρμόζονται σε πολύπλοκα δεδομένα. Τα δέντρα αποφάσεων χρησιμοποιούνται αρκετά και σε προβλήματα ταξινόμησης και σε προβλήματα παλινδρόμησης. Ο τρόπος ο οποίος μαθαίνουν είναι ιεραρχικός με ερωτήσεις της μορφής «εάν», «αλλιώς», δηλαδή, χωρίζουν τα δεδομένα κάνοντας ερωτήσεις για τα χαρακτηριστικά των δεδομένων.

Πίνακας 3.2: Χαρακτηριστικά ενός κρυπτονομίσματος

Δείκτης	Κλείσιμο	Όγκος συναλλαγών	Υ
1^{ο}	3267.78	25000	0 (πώληση)
·	1902.65	70000	1 (αγορά)
·	3888.45	38229	0 (πώληση)
200^{ο}	1032.21	8000	-

Έστω ο πίνακας 3.2 που αναγράφει το κλείσιμο και τον όγκο συναλλαγών για κάποιο κρυπτο-νόμισμα και το δέντρο αποφάσεων της εικόνα 3.5. Στη ρίζα του δέντρου (επίπεδο μηδέν), η ερώτηση που τίθεται είναι: «Εάν η τιμή του κλεισίματος είναι μικρότερη από 2000 τότε "*αγορά*", αλλιώς "*πώληση*"».



Σχήμα 3.5: Δέντρο απόφασης

Η ίδια διαδικασία με τις ερωτήσεις συνεχίζεται και στα αλλά επίπεδα όταν ένα κλαδί δεν είναι φύλλο. Φύλλο, ονομάζεται όταν ένα κλαδί δεν έχει παιδιά, δηλαδή όταν το κλαδί δεν αναπτύσσεται άλλο. Για παράδειγμα, στο επίπεδο 1 το αριστερό κλαδί θεωρείται φύλλο καθώς δεν έχει παιδιά, ενώ το δεξί κλαδί συνεχίζει να διακλαδώνεται χωρίζοντας τα δεδομένα περαιτέρω. Η ένδειξη «παραδείγματα» αναφέρεται στο συνολικό αριθμό των δεδομένων ανά επίπεδο, ενώ η ένδειξη «παραδείγματα ανά κλάση» περιγράφει τον αριθμό των παραδειγμάτων που αντιστοιχεί σε κάθε κλάση. Στο επίπεδο μηδέν υπάρχουν συνολικά 200 παραδείγματα, όπου τα 110 ανήκουν στην κλάση «αγόρασε» ενώ τα υπόλοιπα στην κλάση «πούλα».

Η διαδικασία της πρόβλεψης της τιμής **y** είναι αρκετά απλή υπόθεση. Ξεκινώντας από το επίπεδο 0, πρέπει το κάθε διάνυσμα εισόδου να καταλήξει σε ένα φύλλο, όπου η εκτιμώμενη τιμή $\widehat{\boldsymbol{y}}$, είναι η κλάση του φύλλου. Για παράδειγμα, για την πρώτη τιμή του πίνακα 3.2 τίθεται η ερώτηση εάν η τιμή του κλεισίματος είναι μικρότερη από 2000: Όχι, άρα προχώρα στο επίπεδο ένα, στο δεξί κλαδί. Ύστερα, γίνεται η ερώτηση εάν ο όγκος συναλλαγών είναι μικρότερος από 20000: Η απάντηση είναι πάλι όχι, άρα προχώρα στο επόμενο επίπεδο, στο δεξί κλαδί. Τώρα το κλαδί είναι φύλλο (επίπεδο δύο, δεξί κλαδί), άρα το αποτέλεσμα του δέντρου είναι η σύσταση πώλησης.

Στη βιβλιογραφία έχουν προταθεί διάφορες υλοποιήσεις για την ανάπτυξη δέντρων αποφάσεων στο πεδίο της μηχανικής μάθησης. Ένας από τους πρώτους και πιο διαδεδομένους αλγορίθμους είναι ο CART (Classification and Regression Tree) [9]. Στον αλγόριθμο CART κατασκευή του δέντρου ξεκινώντας από τη ρίζα. Συγκεκριμένα, ο αλγόριθμος χωρίζει τα δεδομένα σε δύο υποσύνολα ανά επίπεδο, ελαχιστοποιώντας την παρακάτω συνάρτηση κόστους:

$$J(k, t_k) = \frac{m_a}{m} G_a + \frac{m_\delta}{m} G_\delta \qquad (3.15)$$

Όπου,

- m_a/m_δ είναι ο αριθμός των παραδειγμάτων στο αριστερό/δεξί κλαδί
- m είναι ο συνολικός αριθμός παραδειγμάτων στο επίπεδο
- G_α/G_δ μετράει την ομοιογένεια (impurity) του αριστερού/δεξιού κλαδιού και περιγράφεται από την παρακάτω σχέση.

$$G_i = 1 - \sum_{k=1}^{c} (p_{i,k})^2 \qquad (3.16)$$

Όπου:

- το i αναφέται στο i^{o} επίπεδο
- το c είναι το συνολικό πλήθος των κλάσεων, στο συγκεκριμένο πρόβλημα είναι δύο
- το k δείχνει για ποια κλάση γίνονται οι υπολογισμοί

- το $p_{i,k}$ είναι η αναλογία των σωστών προβλέψεων προς το συνολικό πλήθος παραδειγμάτων του i επιπέδου της k κλάσης

Υπάρχουν δύο διαφορετικοί μαθηματικοί τύποι οι οποίοι μετράνε την ομοιογένεια (impurity) ενός κλαδιού. Ο πρώτος είναι ο δείκτης Gini, ενώ ο δεύτερος τύπος είναι γνωστός ως εντροπία της πληροφορίας και περιγράφεται από την σχέση (3.17):

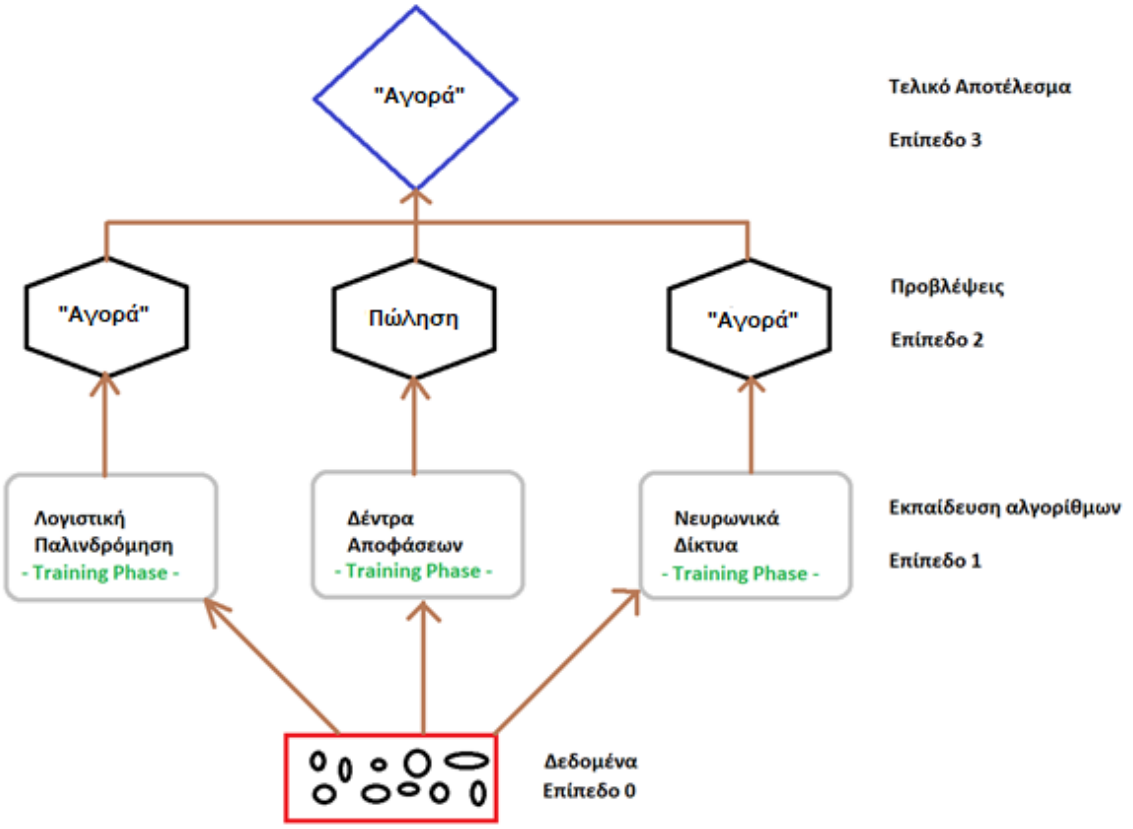
$$H_i = -\sum_{k=1}^{c} p_{i,k}\ log_2\ (p_{i,k}) \qquad (3.17)$$

3.5 Τυχαία δέντρα αποφάσεων

Τα τυχαία δέντρα θεωρούνται αρκετά ισχυρά εργαλεία και προτάθηκαν από τον Ho [10]. Είχαν μεγάλη επιτυχία σε αρκετά προβλήματα ταξινόμησης και είναι από τα πρώτα συνδυασμένα μοντέλα (ensemble models) που έγιναν ευρέως γνωστά. Τα τυχαία δέντρα έχουν ως βάση τρία στοιχεία:

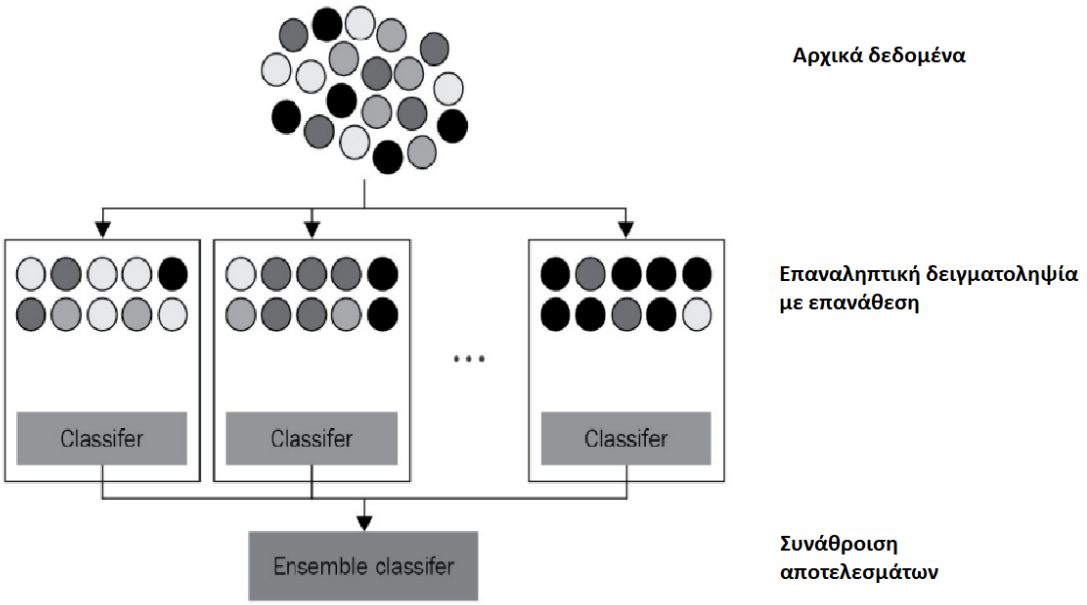
1. Συνδυασμένα μοντέλα
2. Bootstrap Aggregation
3. Δέντρα αποφάσεων

Τα συνδυασμένα μοντέλα είναι μια μεθοδολογία όπου συνδυάζει πολλούς αλγορίθμους μηχανικής μάθησης με σκοπό την δημιουργία ενός πιο ισχυρού μοντέλου. Τα τυχαία δέντρα αποφάσεων χρησιμοποιούν τον κανόνα της πλειοψηφίας (majority vote) για την εξαγωγή των αποτελεσμάτων τους. Στο σχήμα 3.6, φαίνεται γραφικά η διαδικασία πρόβλεψης συνδυάζοντας τρεις διαφορετικούς αλγορίθμους, την λογιστική παλινδρόμηση, τα δέντρα αποφάσεων και τα νευρωνικά δίκτυα. Οι αλγόριθμοι του επιπέδου 1 εκπαιδεύονται από τα ίδια δεδομένα του επιπέδου 0. Ύστερα, αφού τελειώσει η φάση της εκπαίδευσης, το κάθε μοντέλο κάνει ξεχωριστά την πρόβλεψη του για το ίδιο διάνυσμα εισόδου (επίπεδο 2). Το τελικό αποτέλεσμα του επιπέδου 3 εξαρτάται από τις προβλέψεις του επιπέδου 2. Τα αποτελέσματα στο επίπεδο 2 είναι {«Αγορά», «Πώληση», «Αγορά»}, συνεπώς δύο μοντέλα προτείνουν «αγορά», και ένα μοντέλο «πώληση». Άρα, το τελικό αποτέλεσμα είναι η πρόταση αγοράς, η οποία έχει την μεγαλύτερη συχνότητα.



Σχήμα 3.6: Διαδικασία της ψηφοφορίας της πλειοψηφίας

Η επαναληπτική δειγματοληψία είναι η επαναλαμβανόμενη επιλογή ενός δείγματος από τα ήδη παρατηρούμενα δεδομένα. Στην μηχανική μάθηση η επαναληπτική δειγματοληψία χρησιμοποιείται για να βελτιώσει τους αλγορίθμους, ώστε να επιτύχουν μεγαλύτερη ακρίβεια στις προβλέψεις τους. Η εκπαίδευση των τυχαίων δέντρων αποφάσεων στηρίζεται σε δεδομένα στα οποία πραγματοποιείται επαναληπτική δειγματοληψία με επανατοποθέτηση (bootstrap). Το bootstrap aggregation (bagging) προτάθηκε από τον Breiman [11] το 1994. Είναι μία μέθοδος η οποία στηρίζεται στην επαναληπτική δειγματοληψία με επανατοποθέτηση και στον συνδυασμό μοντέλων μηχανικής μάθησης. Στον αλγόριθμο bagging, εκπαιδεύονται πολλοί διαφορετικοί αλγόριθμοι μηχανικής μάθησης χρησιμοποιώντας το ίδιο μοντέλο πολλές φορές, σε δεδομένα στα οποία γίνεται επαναληπτική δειγματοληψία με επανατοποθέτηση. Η διαδικασία απεικονίζεται γραφικά στο Σχήμα 3.7.



Σχήμα 3.7: Διαδικασία Bagging *πηγή:* Siakorn, Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Ensemble_Bagging.svg

Τα τυχαία δέντρα αποφάσεων βασίζονται πάνω στην λογική του bagging. Σαν βάση έχουν τα δέντρα αποφάσεων τα οποία εκπαιδεύονται, όπως έχει αναφερθεί σε ένα δείγμα το οποίο έχει επιλεχθεί από τα δεδομένα με επανατοποθέτηση. Τέλος, αξίζει να σημειωθεί ότι σε ορισμένες υλοποιήσεις των τυχαίων δέντρων αποφάσεων, πραγματοποιείται δειγματοληψία με επανατοποθέτηση και στα χαρακτηριστικά (attributes), όπως στον αλγόριθμο Random Forest [12].

3.6 Boosting

Το boosting είναι επίσης μια τεχνική δημιουργίας συνδυασμένων μοντέλων. Αναπτύχθηκε περίπου την ίδια χρονική περίοδο όπως το bagging. Όπως το bagging, έτσι και το boosting χρησιμοποιείται συνήθως με δέντρα αποφάσεων. Ωστόσο, παρά τις ομοιότητες τους, το boosting ακολουθεί μια διαφορετική προσέγγιση. Σε ένα μοντέλο bagging το κάθε δέντρο αποφάσεων που δημιουργείται δεν δίνει σημασία στα λάθη τα οποία έκαναν τα προηγούμενα δέντρα που είχαν κατασκευαστεί. Σε αντίθεση με το bagging , τα μοντέλα με boosting επικεντρώνονται στο να βελτιώσουν τα λάθη, οπότε το κάθε νέο μοντέλο χτίζεται πάνω στο προηγούμενο. Η γενική ιδέα του boosting είναι η προσαρμογή των μοντέλων στα λάθη που κάνουν τα προηγούμενα εκπαιδευμένα μοντέλα. Υπάρχουν πολλές μεθοδολογίες boosting, αλλά δύο είναι οι πιο γνωστές: Το Adaptive boosting, (AdaBoost) και το Gradient Boosting.

Το AdaBoost, το οποίο προτάθηκε από τον Freund [13], είναι από τα πρώτα γνωστά Boosting μοντέλα. Στο adaptive boosting δίνεται έμφαση περισσότερο στις λανθασμένες

προβλέψεις, οι οποίες επηρεάζουν περισσότερο την αποτελεσματικότητα του μοντέλου. Η διαδικασία όπου ο αλγόριθμος του Adaboost μαθαίνει από τα λάθη του, μπορεί να βελτιώσει έναν αδύναμο αλγόριθμο σε έναν αρκετά ισχυρό. Ο αλγόριθμος περιγράφεται στα παρακάτω βήματα:

1. Όρισε τον αριθμό *P,* όπου *P* είναι ο μέγιστος αριθμός των μοντέλων των οποίων θα εκπαιδευτούν.
2. Αρχικοποίησε τον αριθμό $j = 1$, όπου $j = 1,\cdots, P$
3. Υπολόγισε την τιμή $r_j = \frac{\sum_{i=1}^{m}{}_{\hat{y}_j^{(i)} \neq y^{(i)}} w^{(i)}}{\sum_{i=1}^{m} w^{(i)}}$, όπου για $j = 1$ τότε αρχικοποίησε $w^{(i)} = \frac{1}{m}$
4. Υπολόγισε την τιμή $\alpha_j = \eta log \frac{1-r_j}{r_j}$ όπου η είναι ο ρυθμός εκμάθησης (learning rate)
5. Για i $= 1, 2,\cdots,$ m υπολόγισε τα νέα βάρη $w^{(i)} \leftarrow \begin{cases} w^{(i)} & \text{εάν } \hat{y}_j^{(i)} = y^{(i)} \\ w^{(i)} \exp(a_j) & \text{εάν } \hat{y}_j^{(i)} \neq y^{(i)} \end{cases}$
6. Διαίρεσε τα βάρη με το άθροισμα τους $\sum_{i=1}^{m} w^{(i)}$
7. Εάν $j < P$ τότε $j = j + 1$ και πήγαινε στο βήμα 3, αλλιώς τερμάτισε την διαδικασία

Οι εκτιμήσεις δίνονται από τον παρακάτω τύπο (3.18):

$$\hat{y} = argmax_k \sum_{j=1}^{P}{}_{\hat{y}_j(x)=k} a_j \qquad (3.18)$$

Στην ουσία η προβλεπόμενη κλάση είναι αυτή που συγκέντρωσε τις περισσότερες ψήφους με το αντίστοιχο βάρος, όπως στην διαδικασία της ψηφοφορίας της πλειοψηφίας.

Η άλλη γνώστη και η πλέον πιο διαδεδομένη μεθοδολογία Boosting είναι το Gradient Boosting. Προτάθηκε πρώτα από τον Breiman [14] το 1997 και αναπτύχθηκε αργότερα από τον Friedman [15]. Η διαδικασία είναι παρόμοια με αυτή του AdaBoost, απλά αντί να ρυθμίζει τα βάρη σε κάθε επανάληψη, εκπαιδεύει το κάθε μοντέλο με το σφάλμα υπολοίπου, το οποίο έκανε το προηγούμενο μοντέλο. Το σφάλμα υπολοίπου είναι η διαφορά μεταξύ της εκτιμώμενης τιμής και της πραγματικής τιμής, δηλαδή $\hat{y} - y$. Για παράδειγμα, εάν το $\hat{y} = 200$ και $y = 190$, τότε το σφάλμα υπολοίπου είναι $\hat{y} - y = 200 - 190 = 10$. Ο αλγόριθμος του Gradient Boosting μπορεί να περιγραφτεί από τα παρακάτω βήματα:

1. Όρισε τον αριθμό *P*, όπου $P > 1$, των μοντέλων που θα εκπαιδευτούν στο σφάλμα υπολοίπου του i^{στου}-1 μοντέλου, όπου $i = 1, \cdots, P$
2. Για $i = 1$ εκπαίδευσε το μοντέλο στα αρχικά δεδομένα
3. Κάνε προβλέψεις για το $i = 1$ μοντέλο
4. Υπολόγισε το σφάλμα υπολοίπου, $\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i$.
5. Όρισε το σφάλμα υπολοίπου ως τα νέα δεδομένα
6. Αύξησε το βήμα κατά ένα, $i = i + 1$

7. Εκπαίδευσε το μοντέλο i στα νέα δεδομένα.
8. Εάν $j < P$ τότε πήγαινε στο βήμα 4 αλλιώς συνέχισε στο επόμενο βήμα 9
9. Υπολόγισε το $\hat{\boldsymbol{y}}_{total} = \sum_{i=1}^{P}(\hat{\boldsymbol{y}}_i)$, όπου $\hat{\boldsymbol{y}}_{total}$ είναι η τελική εκτίμηση.

3.7 Μέτρα Αξιολόγησης

Η αξιολόγηση ενός αλγορίθμου μηχανικής μάθησης ταξινόμησης είναι δύσκολη διαδικασία και διατρέχει ορισμένους, κρυμμένους κινδύνους, όπου μπορεί να αποπροσανατολίσει τον αναλυτή πιστεύοντας ότι διαθέτει ένα αρκετά αξιόπιστο μοντέλο. Τα μέτρα αξιολόγησης είναι ο τρόπος με τον οποίο αναλύεται η αποτελεσματικότητα ενός μοντέλου μετά την εκπαίδευση του. Πολλά από τα μέτρα αξιολόγησης έχουν προέλθει από τον πίνακα ταξινόμησης (confusion matrix), ο οποίος θα μπορούσε να θεωρηθεί η καρδιά των μέτρων αξιολόγησης.

Ο πίνακας ταξινόμησης είναι ένας πίνακας ο οποίος δείχνει τον αριθμό των σωστών και λανθασμένων προβλέψεων, κατηγοριοποιημένες ανά κλάση. Η γενική ιδέα είναι να μετρήσει τη συχνότητα όπου η κλάση «πώληση» ταξινομείται ως την κλάση «Αγορά» και αντίστροφα. Στο σχήμα 3.8 απεικονίζεται ένας πίνακας ταξινόμησης.

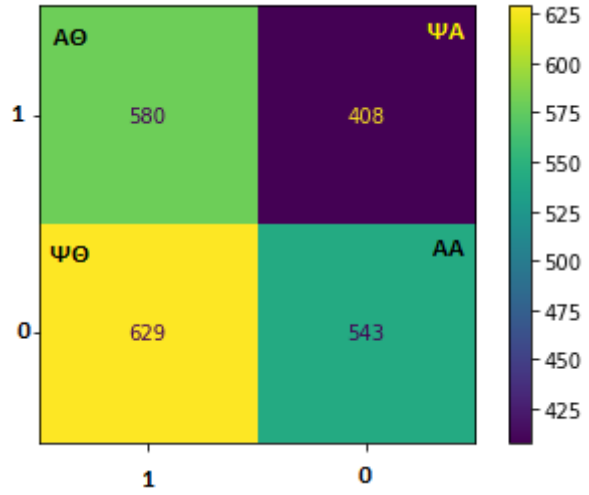
	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	1,1 True Positive	1,2 False Negative
$y = 0$	2,1 False Positive	2,2 True Negative

Σχήμα 3.8: Αναπαράσταση ενός πίνακα ταξινόμησης

Οι γραμμές του πίνακα παρουσιάζουν τις πραγματικές τιμές των κλάσεων, ενώ οι στήλες τις εκτιμώμενες τιμές των κλάσεων. Η ένδειξη $y = 1$ συμβολίζει την θετική κλάση η οποία κωδικοποιείται ως «αγορά», ενώ το $y = 0$ συμβολίζει την αρνητική κλάση, δηλαδή «πώληση». Αντίστοιχα, το $\hat{y} = 1$ είναι το αποτέλεσμα-σύσταση του αλγορίθμου, ότι το παράδειγμα ανήκει στην θετική κλάση, ενώ το $\hat{y} = 0$ είναι η εκτίμηση ότι το παράδειγμα ανήκει στην αρνητική κλάση. Το στοιχείο του πίνακα [1,1] έχει την ένδειξη αληθές θετικό (ΑΘ) (True positive), το οποίο δείχνει τον αριθμό των σωστών αποτελεσμάτων για την θετική κλάση, δηλαδή η πραγματική τιμή συμπίπτει με τη σύσταση του αλγορίθμου ($\hat{y} = y = 1$). Το στοιχείο του πίνακα με δείκτη [1,2], μετράει τον αριθμό των αποτελεσμάτων που υποδεικνύουν αρνητική κλάση, ενώ στην πραγματικότητα πρόκειται για θετικά παραδείγματα ($\hat{y} = 0, y = 1$). Αντίστοιχα, το στοιχείο [2,1] του πίνακα, δείχνει τον αριθμό των περιπτώσεων που έχουν ταξινομηθεί λάθος, ως θετική κλάση, ενώ ανήκουν στην αρνητική ($\hat{y} = 1, y = 0$). Τέλος, το στοιχείο του πίνακα [2,2], μετράει τον αριθμό των

σωστών αποτελεσμάτων που ανήκουν στην αρνητική κλάση ($\hat{y} = y = 0$). Το άθροισμα της κύριας διαγώνιας του πίνακα είναι ο αριθμός των συνολικών περιπτώσεων για τις οποίες ο αλγόριθμος έχει δώσει σωστές εκτιμήσεις.

Για παράδειγμα, στο σχήμα 3.9, το μοντέλο έχει κάνει 580+543=1123 σωστές εκτιμήσεις από τις 580 + 408 + 629 + 543 = 2160 περιπτώσεις στο σύνολο. Επίσης, το μοντέλο έχει προβλέψει ότι 408 παραδείγματα ανήκουν στην μηδενική κλάση ενώ ανήκουν στην θετική, δηλαδή σφάλμα τύπου ψευδώς αρνητικό (ΨΑ). Επιπλέον, 629 παραδείγματα έχουν ταξινομηθεί εσφαλμένα στη θετική κλάση, ενώ ανήκουν στην αρνητική, δηλαδή σφάλμα τύπου ψευδώς θετικό (ΨΘ). Τα υπόλοιπα παραδείγματα, που παραμένουν με την ένδειξη ΑΘ και ΑΑ, είναι τα παραδείγματα τα οποία έχουν ταξινομηθεί σωστά.



Σχήμα 3.9: Παράδειγμα ενός πίνακα σύγχυσης

Το πιο ευρέως χρησιμοποιούμενο μέτρο αξιολόγησης είναι η ορθότητα (accuracy). το οποίο μπορεί εύκολα να υπολογισθεί από τον πίνακα ταξινόμησης. Είναι το άθροισμα της κύριας διαγωνίου διά το άθροισμα όλων των στοιχείων του πίνακα.

$$Accuracy = \frac{A\Theta + AA}{A\Theta + \Psi A + \Psi\Theta + AA} \qquad (3.19)$$

Στην ουσία η ορθότητα είναι το κλάσμα των σωστών προβλέψεων προς το πλήθος των προβλέψεων. Η ορθότητα, ως μέτρο, δεν είναι συνήθως αξιόπιστη, καθώς δεν περιέχει όλες τις απαραίτητες πληροφορίες που χρειάζονται για την αποτελεσματικότητα του μοντέλου που εξετάζεται. Για το λόγο αυτό, συνήθως, χρησιμοποιείται μαζί με άλλα μέτρα όπως την ακρίβεια (precision), ανάκληση (recall), εξειδίκευση (specifity), και η περιοχή κάτω από την καμπύλη ROC (receiver operating characteristic curve).

Το precision είναι ένα βοηθητικό μέτρο της ορθότητας όπου περιγράφεται από την παρακάτω σχέση:

$$Precision = \frac{A\Theta}{A\Theta + \Psi\Theta} \qquad (3.20)$$

Αυτή η σχέση φανερώνει, από τις προβλέψεις που διαγνώστηκαν ως θετικής κλάσης, σε τι ποσοστό οι προβλέψεις ήταν σωστές, δηλαδή ανήκουν στην θετική κλάση. Για παράδειγμα το precision της εικόνας είναι $\frac{580}{580+629} = 0.4793$.

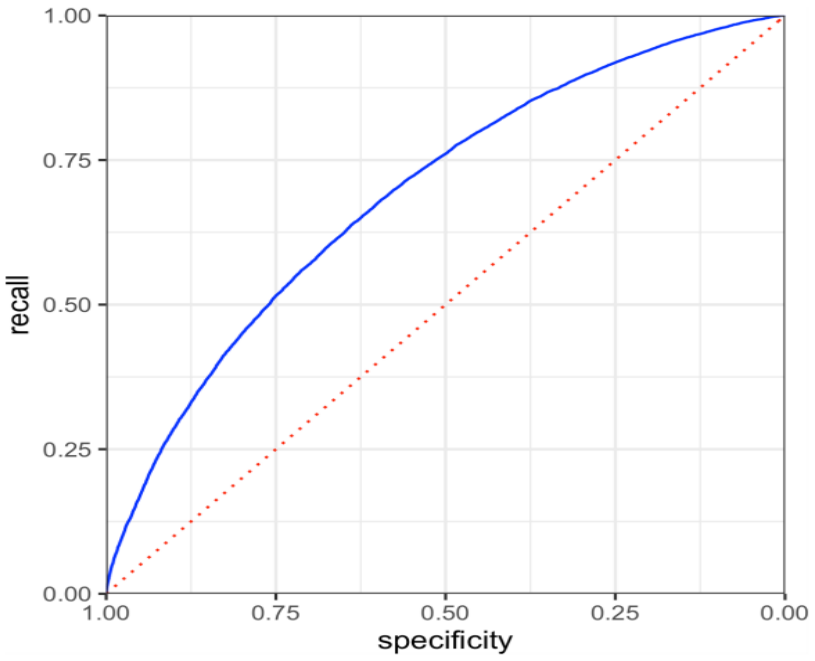
Ένα άλλο μέτρο αξιολόγησης το οποίο μοιάζει με το precision είναι το recall (sensitivity). Το recall μετράει την δυναμικότητα ενός μοντέλου να προβλέπει σωστά τη θετική κλάση. Η σχέση περιγράφεται από την σχέση (3.21)

$$Recall = \frac{A\Theta}{A\Theta+\Psi A} \tag{3.21}$$

Το specifity, σε αντίθεση με τα άλλα μέτρα, μετράει την ικανότητα ενός μοντέλου να προβλέπει τις αρνητικές εκβάσεις. Δίνεται από την παρακάτω σχέση:

$$Specifity = \frac{AA}{AA+\Psi\Theta} \tag{3.22}$$

Επίσης, ένα αρκετά χρησιμοποιούμενο μέτρο, το οποίο αναπαράγεται γραφικά (βλ. σχήμα 3.10), είναι το ROC curve. Το διάγραμμα της καμπύλης ROC έχει στον κάθετο άξονα το recall και στον οριζόντιο άξονα το μέτρο specifity.



Σχήμα 3.10: Παράδειγμα καμπύλης ROC

Οι διακεκομμένες τελείες στο διάγραμμα απεικονίζουν ένα μοντέλο ταξινόμησης το οποίο προβλέπει όχι καλύτερά από κάποιο μοντέλο που επιλέγει τυχαία. Τα μοντέλα τα οποία είναι πάρα πολύ αποτελεσματικά αγκαλιάζουν την πάνω αριστερή γωνία.

Η περιοχή κάτω από την καμπύλη ROC (Area Under the Curve, AUC) είναι ένα ποσοτικό μέτρο αξιολόγησης της διακριτικής/προβλεπτικής ικανότητας ενός μοντέλου ταξινόμησης, το οποίο βασίζεται στην καμπύλη ROC. Όσο μεγαλύτερο και πιο κοντά στην

μονάδα είναι το AUC, τόσο πιο αποτελεσματικό είναι το μοντέλο, με την μονάδα να είναι το ιδανικό μοντέλο.

Κεφάλαιο 4ο

Εφαρμογή

4.1 Εισαγωγή

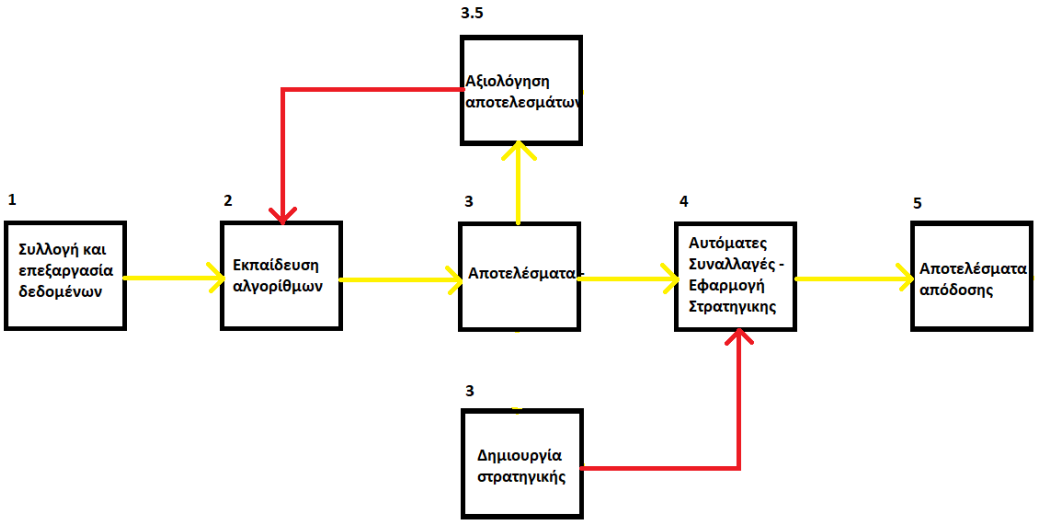
Σε αυτό το κεφάλαιο, θα αναφερθούν τα βήματα για την υλοποίηση ενός συστήματος το οποίο με τη βοήθεια αλγορίθμων μηχανικής μάθησης θα πραγματοποιεί συναλλαγές κρυπτονομισμάτων με εφαρμογή κάποιων απλών στρατηγικών με σκοπό να μεγιστοποιηθεί το κεφάλαιο που επενδύεται.

Το πρώτο βήμα είναι η συλλογή και η επεξεργασία των δεδομένων. Αφού διαμορφωθούν τα δεδομένα στην κατάλληλη μορφή, μπορούν να χρησιμοποιηθούν για την εκπαίδευση των αλγορίθμων. Οι αλγόριθμοι μηχανικής μάθησης οι οποίοι θα χρησιμοποιηθούν είναι οι εξής:

- Λογιστική παλινδρόμηση
- Νευρωνικά δίκτυα
- Τυχαία δέντρα αποφάσεων
- XGBoost (Extreme Gradient Boosting)

Με βάση αυτές τις εκτιμήσεις και προβλέψεις των μοντέλων εφαρμόζονται κάποιες επενδυτικές στρατηγικές, οι οποίες καθορίζουν το πότε και με ποια κριτήρια θα πραγματοποιούνται συναλλαγές κρυπτ-νομισμάτων (ενέργειες αγοράς - πώλησης κρυπτονομισμάτων). Τέλος, τα αποτελέσματα των στρατηγικών αυτών συγκρίνονται με μία απλή παθητική στρατηγική, η οποία ονομάζεται buy & hold («αγορά και διακράτηση»).

Τα βήματα αυτά συνοψίζονται στο σχήμα 4.1 και θα αναλυθούν με περισσότερες λεπτομέρειες στις επόμενες ενότητες, καθώς και τα αποτελέσματα τους.



Σχήμα 4.1: Συνοπτική παρουσίαση των βημάτων που ακολουθήθηκαν για την υλοποίηση του ενός συστήματος αυτόματης διαπραγμάτευσης κρυπτονομισμάτων.

4.2 Δεδομένα και η επεξεργασίας τους

Τα δεδομένα τα οποία χρησιμοποιήθηκαν για την εκπαίδευση των αλγορίθμων μηχανικής μάθησης αφορούν το Bitcoin και συλλέχθηκαν από το γνωστό ανταλλακτήριο (exchanger broker) Gemini[1]. Οι εγγραφές των δεδομένων ξεκινούν από την περίοδο του Ιανουαρίου του 2017 μέχρι τον Ιούνιο του 2021. Η κάθε εγγραφή αναφέρεται στις συναλλαγές που πραγματοποιήθηκαν ανά μία ώρα, δηλαδή 39614 περίπου παραδείγματα (instances). Τα χαρακτηριστικά (attributes), τα οποία περιλαμβάνονται στα δεδομένα, είναι η ημερομηνία, οι τιμές ανοίγματος και κλεισίματος, ο όγκος συναλλαγών, η μέγιστη και ελάχιστη τιμή του Bitcoin ανά ώρα.

Σε πρώτη φάση, ελέγχθηκε εάν στα δεδομένα υπάρχουν διπλές εγγραφές με βάση τις ημερομηνίες. Ύστερα, καθώς πρόκειται για πρόβλημα χρονοσειρών, η σειρά των δεδομένων έχει σημασία, οπότε τα δεδομένα ταξινομήθηκαν με την ημερομηνία από την παλιότερη στην νεότερη. Στο επόμενο βήμα, τα δεδομένα ελέγχθηκαν για ελλιπείς τιμές. Στο συγκεκριμένο σύνολο δεδομένων βρέθηκαν 120 ημερομηνίες χωρίς στοιχεία. Οι τιμές αυτές συμπληρώθηκαν με γραμμική παρεμβολή.

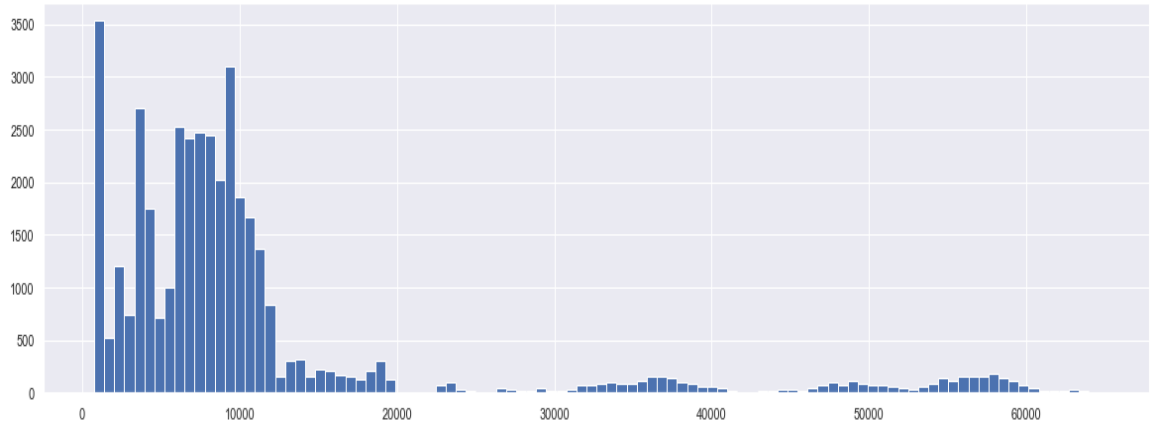
Στην επόμενη φάση για την αύξηση της απόδοσης των αλγορίθμων, προστέθηκαν ορισμένα, επιπλέον χαρακτηριστικά στα δεδομένα. Αυτά τα επιπλέον χαρακτηριστικά, είναι ποσοτικοί δείκτες, οι οποίοι χρησιμοποιούνται στην τεχνική ανάλυση για να βοηθήσουν τους επενδυτές να πάρουν καλύτερες αποφάσεις, σχετικά με επενδύσεις στο χρηματιστήριο. Οι δείκτες που δοκιμάστηκαν, έχουν χρησιμοποιηθεί και σε άλλες έρευνες [16-22] και σύμφωνα με τα αποτελέσματα τους μπορούν να βελτιώσουν την αποτελεσματικότητα των μοντέλων. Έπειτα από αρκετές δοκιμές, οι τεχνικοί δείκτες οι οποίοι εν τέλει έδειξαν ότι μπορεί να βελτιώνουν την απόδοση των αλγορίθμων είναι:

- *Exponential Moving Average (EMA)*
- *Relative Strength Index (RSI)*
- *Moving Average Convergence/Divergence (MACD)*
- *Stochastic Oscillator (SO)*
- *On Balance Volume (OBV)*

Στα βασικά χαρακτηριστικά, δηλαδή το άνοιγμα, το κλείσιμο, ο όγκος συναλλαγών, η μέγιστη και ελάχιστη τιμή του Bitcoin, παρουσιάζεται σταδιακή αύξηση με την πάροδο του χρόνου, όπως αποτυπώνεται στο σχήμα 4.2. Από το ιστόγραμμα του σχήματος 4.3 φαίνεται ότι η τιμή κλεισίματος δεν ακολουθεί κάποια συγκεκριμένη κατανομή.

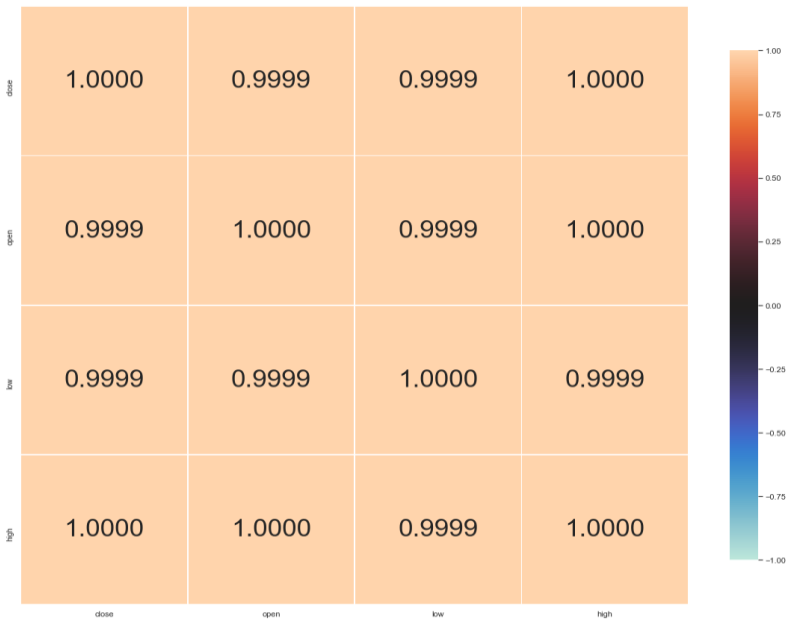
[1] https://www.gemini.com/

Σχήμα 4.2: Η εξέλιξη της τιμή του Bitcoin την περίοδο 2017 – 2021



Σχήμα 4.3:Ιστόγραμμα της τιμής κλεισίματος του Bitcoin

Επίσης, τα χαρακτηριστικά είναι σχεδόν τέλεια συσχετισμένα μεταξύ τους, με συντελεστή συσχέτισης ίσο με ένα, σχεδόν σε όλες τις περιπτώσεις, όπως φαίνεται και στο σχήμα 4.4. Αυτή η μη-στασιμότητα και αυτοσυσχέτιση δεν διευκολύνει ιδιαίτερα την διαδικασία της εκπαίδευσης, καθώς τα χαρακτηριστικά δεν είναι τόσο εύχρηστα. Συνεπώς, οι αλγόριθμοι δεν μπορούν να «ενσαρκώσουν» με ευκολία όλες τις χρήσιμες πληροφορίες. Οπότε, αυτά τα χαρακτηριστικά δεν μπορούν να χρησιμοποιήθηκαν στη διαδικασία εκπαίδευσης.



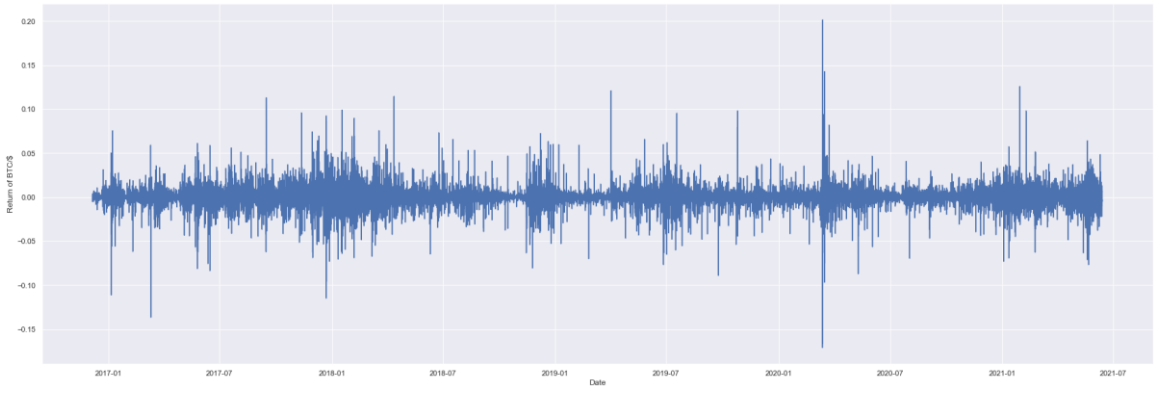
Σχήμα 4.4: Γραφική αναπαράσταση της αυτοσυσχέτισης των χαρακτηριστικών

Μία λύση για να αντιμετωπιστεί αυτό το πρόβλημα, είναι να υπολογιστεί η ποσοστιαία μεταβολή μεταξύ μίας ώρα της τιμής του Bitcoin, δηλαδή η απόδοση του Bitcoin ανά ώρα. Ο υπολογισμός αυτός γίνεται με τον παρακάτω τύπο (4.1).

$$R_t = \frac{C_t - C_{t-1}}{C_{t-1}} \qquad (4.1)$$

- όπου C_t είναι η τιμή κλεισίματος για την t περίοδο.

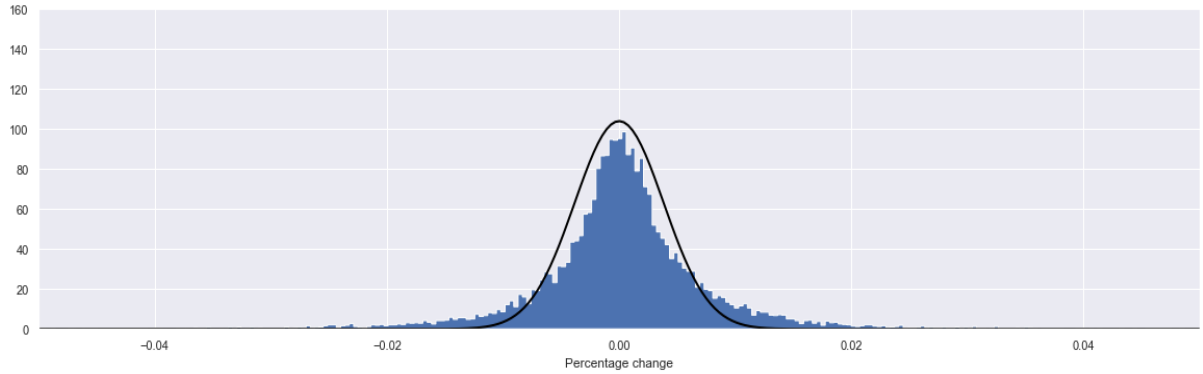
Χρησιμοποιώντας τις αποδόσεις, οι αλγόριθμοι είναι πιο εύκολο να απορροφήσουν πληροφορίες από τα δεδομένα, καθώς τα δεδομένα έχουν αποκτήσει μία στασιμότητα όπως φαίνεται στο διάγραμμα του σχήματος 4.5.



Σχήμα 4.5: Ποσοστιαία μεταβολή της τιμής του Bitcoin με την πάροδο του χρόνου

Επιπλέον, η στασιμότητα αυτή μπορεί να γίνει πιο φανερή, καθώς στο σχήμα της εικόνα 4.6 όπου απεικονίζεται το ιστόγραμμα της ποσοστιαίας μεταβολής της τιμής του Bitcoin, φαίνεται πως η κατανομή προσεγγίζει κανονική κατανομή. Παρόλο που η

κατανομή του ιστογράμματος θυμίζει κανονική κατανομή, τα δεδομένα δεν ακολουθούν αυτήν, καθώς στο γράφημα του σχήματος 4.6 η καμπύλη απεικονίζει το πως θα ήταν η κατανομή των δεδομένων εάν ακολουθούσαν κανονική κατανομή. Επίσης, τα άκρα της κατανομής που ακολουθεί η απόδοση του Bitcoin είναι αρκετά μακρύτερες και πυκνότερες από αυτές της κανονικής κατανομής, κάνοντας πιο εμφανές ότι η απόδοση δεν περιγράφεται από κανονική κατανομή.



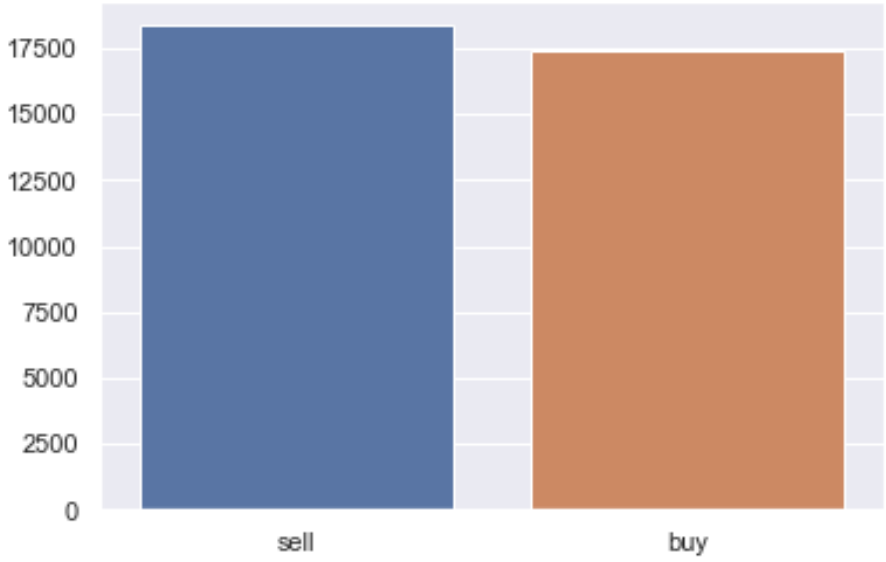
Σχήμα 4.6: Κατανομή των ωριαίων αποδόσεων του Bitcoin

Στο γράφημα του σχήματος 4.7 είναι αποτυπωμένες δύο διακεκομμένες γραμμές, μία μαύρη (άνω) και μία κόκκινη (κάτω), πάνω στο γράφημα της απόδοσης του Bitcoin. Η μαύρη διακεκομμένη γραμμή είναι το 97.5% ποσοστημόριο των αποδόσεων, ενώ η κόκκινη διακεκομμένη γραμμή αντιπροσωπεύει το 2.5% ποσοστημόριο των αποδόσεων. Από το γράφημα αυτό, επιβεβαιώνεται η ύπαρξη πολλών περιπτώσεων όπου οι αποδόσεις εμφανίζουν ακραίες τιμές που από το 95% διάστημα εμπιστοσύνης που ορίζεται από τα δύο παραπάνω ποσοστημόρια.

Σχήμα 4.7: Απεικόνιση ποσοστημορίων για το 2.5% και το 97.5% της απόδοσης του Bitcoin

Ένα ακόμα στάδιο που πρέπει να πραγματοποιηθεί είναι η δημιουργία των στόχων (targets), δηλαδή οι κλάσεις. Οι κλάσεις δημιουργούνται με το εξής κριτήριο: Εάν το κλείσιμο του Bitcoin είναι μεγαλύτερο από το κλείσιμο της προηγούμενης ώρας καταχωρείται ως «Αγορά», διαφορετικά θεωρείται ως «Πώληση». Η κλάση «Αγορά» συμβολίζεται με τον αριθμό ένα, ενώ η «Πώληση» με τον αριθμό μηδέν. Στο σχήμα 4.8

φαίνεται πως δεν υπάρχει μεγάλη διαφορά στις συχνότητες των κλάσεων, πράγμα το οποίο διευκολύνει την εκπαίδευση των αλγορίθμων και την αξιολόγηση τους.



Σχήμα 4.8: Ραβδόγραμμα της αρνητικής και θετικής κλάσης («πούλα» /«Αγόρασε» αντίστοιχα)

Τέλος, να αναφερθεί ότι στα χαρακτηριστικά πριν την έναρξη της εκπαίδευσης πρέπει να γίνει μια κανονικοποίηση των δεδομένων, ώστε τα χαρακτηριστικά να εκφραστούν σε μια κοινή κλίμακα (feature scaling). Αυτό συμβαίνει γιατί οι περισσότεροι αλγόριθμοι μηχανικής μάθησης δεν αποδίδουν καλά όταν οι τιμές των χαρακτηριστικών έχουν μεγάλες διαφορές μεταξύ τους. Ο μετασχηματισμός που χρησιμοποιήθηκε περιγράφεται από τον τύπο 4.2, ώστε όλα τα χαρακτηριστικά να έχουν μηδενική μέση τιμή και τυπική απόκλιση ίση με 1:

$$z = \frac{x - \bar{x}}{\sigma} \qquad (4.2)$$

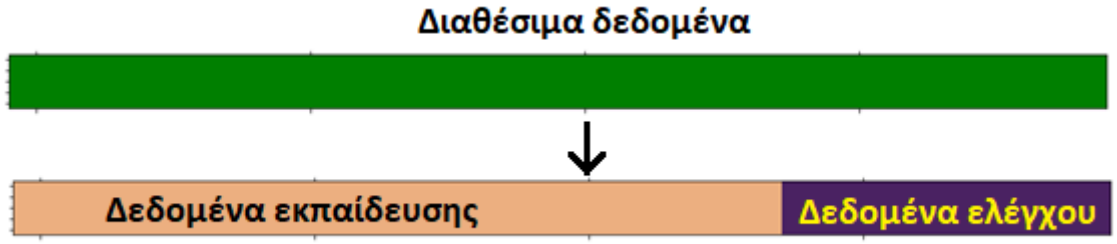
Τα χαρακτηριστικά που τελικά επιλέχθηκαν για την εκπαίδευση είναι 14 και συνοψίζονται στον πίνακα 4.1.

Πίνακας 4.1: Τα χαρακτηριστικά που χρησιμοποιήθηκαν για την εκπαίδευση

Χαρακτηριστικά	Περίοδοι (ώρες)
Απόδοση τιμής του Bitcoin (R_t)	
EMA	2, 4, 12, 24
RSI	12, 24, 48
MACD	{ (2 ,12 ,9), (2, 24, 9), (4, 24, 9) }
Stochastic Oscillator	{ (12, 12, 12), (24, 24, 12), (48, 48, 12) }
On Balance Volume	-

4.3 Εκπαίδευση

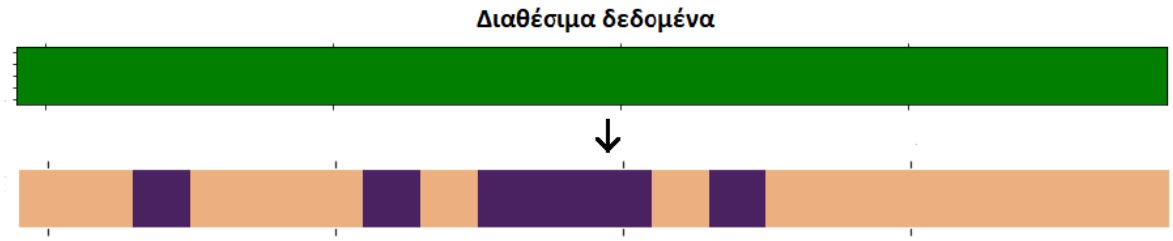
Στο στάδιο της εκπαίδευσης ενός αλγορίθμου μηχανικής μάθησης είναι συνηθισμένο τα δεδομένα να χωρίζονται σε δύο υποσύνολα. Το ένα υποσύνολο περιέχει τα δεδομένα εκπαίδευσης, ενώ το άλλο τα δεδομένα ελέγχου. Συνήθως, τα δεδομένα εκπαίδευσης αποτελούν το 70% των συνολικών δεδομένων, ενώ το υπόλοιπο ποσοστό, αποτελεί το δείγμα ελέγχου όπως φαίνεται στο σχήμα 4.9.



Σχήμα 4.9: Διαχωρισμός δεδομένων σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου.

Το ποσοστό που αποτελείται το κάθε υποσύνολο εξαρτάται από τον σχεδιασμό της λύσης του προβλήματος. Τα δεδομένα εκπαίδευσης είναι το υποσύνολο στο οποίο εκπαιδεύονται οι αλγόριθμοι και γίνεται η αναζήτηση των παραμέτρων που δίνουν το αποτελεσματικότερο μοντέλο. Αφού έχουν επιλεχθεί οι παράμετροι, τότε δοκιμάζεται το μοντέλο στα δεδομένα ελέγχου, τα οποία είναι παραδείγματα που το μοντέλο δεν έχει ξανά συναντήσει κατά τη διάρκεια της εκπαίδευσης του.

Μία άλλη συνηθισμένη πρακτική είναι η διαμόρφωση των δειγμάτων εκπαίδευσης και ελέγχου να γίνεται τυχαία όπως παρουσιάζεται παραστατικά στο σχήμα 4.10.



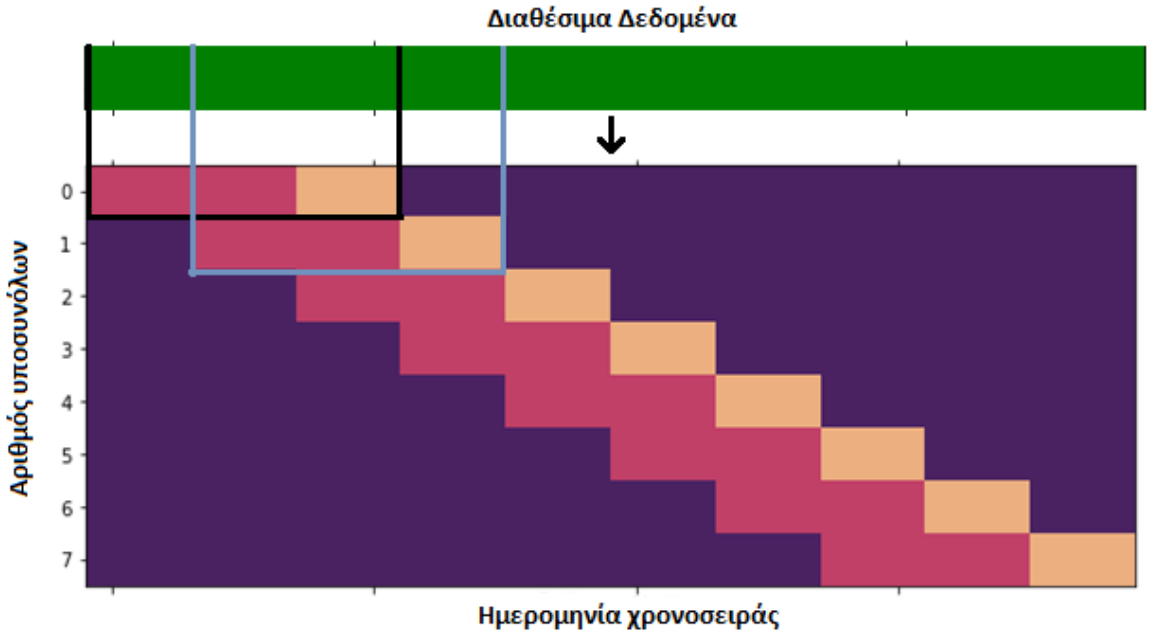
Σχήμα 4.10: Διαχωρισμός δεδομένων με τυχαία σειρά σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου.

Ωστόσο, στα προβλήματα των χρονοσειρών δεν είναι καλή πρακτική τα παραδείγματα να χωρίζονται με τυχαία σειρά, καθώς υπάρχει «διαρροή δεδομένων» (data leakage). Η διαρροή δεδομένων αναφέρεται σε περιπτώσεις όπου τα δεδομένα εκπαίδευσης συνίστανται από πληροφορίες, οι οποίες περιέχονται και στα δεδομένα ελέγχου. Αυτό συμβαίνει διότι ο τυχαίος διαχωρισμός των δεδομένων δεν λαμβάνει υπόψη

τη χρονολογική σειρά των δεδομένων σε προβλήματα χρονοσειρών. Αυτό έχει ως αποτέλεσμα την δυσκολία της σωστής αξιολόγησης των μοντέλων, καθώς οι προβλέψεις είναι πιθανό να είναι περισσότερο «αισιόδοξες» για τα δεδομένα ελέγχου, ενώ στην πραγματικότητα η ικανότητα πρόβλεψης του μοντέλου σε νέα δεδομένα να είναι ασθενής.

Για να αντιμετωπιστεί αυτό το πρόβλημα, τα δεδομένα χωρίζονται σε περισσότερα μικρά τμήματα, όπως φαίνεται στο σχήμα 4.11, και η εκπαίδευση πραγματοποιείται σε κύκλους. Η κάθε γραμμή του αναφερόμενου σχήματος αποτελεί ένα υποσύνολο των δεδομένων, όπου το κόκκινο χρώμα είναι τα δεδομένα εκπαίδευσης και το πορτοκαλί τα δεδομένα ελέγχου. Συγκεκριμένα, παράδειγμα του σχήματος 4.11, τα δεδομένα έχουν χωριστεί σε οκτώ υποσύνολα (η αρίθμηση ξεκινάει από το μηδέν).

Στον πρώτο κύκλο, ο αλγόριθμος εκπαιδεύεται και αξιολογείται στα δεδομένα που αποτελούνται από το πρώτο (μαύρο) πλαίσιο. Συγκεκριμένα, τα δεδομένα που αποτελούνται από το μαύρο πλαίσιο χωρίζονται σε δεδομένα εκπαίδευσης (κόκκινο χρώμα) και δεδομένα ελέγχου (πορτοκαλί χρώμα). Συνεπώς, ο αλγόριθμος εκπαιδεύεται στα δεδομένα με το κόκκινα χρώμα και αξιολογείται στα δεδομένα με το πορτοκαλί χρώμα. Στον δεύτερο κύκλο, η λογική είναι η ίδια, με την διαφορά ότι τα νέα δεδομένα εκπαίδευσης και ελέγχου, τα οποία ανήκουν στο μπλε πλαίσιο, έχουν μετακινηθεί δεξιά, κατά ένα βήμα, όσο είναι το μέγεθος των δεδομένων ελέγχου του πρώτου κύκλου. Έπειτα, η διαδικασία συνεχίζεται με τον ίδιο τρόπο. Επίσης, μπορεί να παρατηρηθεί από το σχήμα 4.11 ότι τα δεδομένα εκπαίδευσης του δεύτερου κύκλου, όπως και στους επόμενους κύκλους, περιέχουν αρκετά δεδομένα από τον πρώτο κύκλο, καθώς και τα δεδομένα ελέγχού του. Η διαδικασία αυτή ονομάζεται «κινούμενο παράθυρο» (rolling window).



Σχήμα 4.11: Διαχωρισμός δεδομένων σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου μέσω ενός κινούμενου παραθύρου.

Αναλυτικότερα, η διαδικασία του κινούμενου παραθύρου μπορεί να περιγραφεί ως εξής: Έστω ότι τα συνολικά διαθέσιμα παραδείγματα είναι N και για λόγους σχεδίασης είναι επιθυμητό το κάθε υποσύνολο να έχει n παραδείγματα για την εκπαίδευση και να ελέγχεται σε p παραδείγματα. Τότε, το πρώτο υποσύνολο που θα εκπαιδευτεί ο αλγόριθμος αποτελείται από τα πρώτα $[1, n+p]$ παραδείγματα, το δεύτερο υποσύνολο στο οποίο θα εκπαιδευτεί θα αποτελείται από τα $[p, n+2\times p]$, το τρίτο από τα $[2\times p, n+3\times p]$ και το ν^{οστο} από τα $[(\nu-1)\times p, n+\nu\times p]$. Επίσης, να σημειωθεί ότι το κάθε υποσύνολο έχει το ίδιο μέγεθος. Το μέγεθος αυτό εξαρτάται από το πρόβλημα και το μέγεθος των διαθέσιμων δεδομένων.

Βέβαια, για να εφαρμοστεί αποτελεσματικά αυτή η διαδικασία πρέπει να υπάρχουν αρκετά διαθέσιμα δεδομένα, ώστε η εκπαίδευση να γίνεται με ικανοποιητικό αριθμό παραδειγμάτων. Επιπλέον, με αυτό τον τρόπο δεν διαταράσσεται η χρονολογική σειρά των δεδομένων. Τέλος, η αξιολόγηση δεν γίνεται συνέχεια στα ίδια δεδομένα, με κίνδυνο το μοντέλο να υπερ-προσαρμοστεί στα δεδομένα, καθώς με αυτό το τέχνασμα δίνεται η δυνατότητα το μοντέλο να δοκιμαστεί σε περισσότερα διαφορετικά δεδομένα ελέγχου.

Όπως προαναφέρθηκε στην ενότητα 4.2, τα διαθέσιμα παραδείγματα είναι περίπου 39614. Από αυτά, τα 25281 χρησιμοποιήθηκαν για την εκπαίδευση των αλγορίθμων και την αναζήτηση των καλύτερων υπερ-παραμέτρων (βλ. ενότητα 4.4). Τα υπόλοιπα από αυτά χρησιμοποιήθηκαν την εξαγωγή των αποτελεσμάτων (ενότητα 4.5). Σε αυτά τα παραδείγματα εφαρμόστηκε η διαδικασία του κινούμενου παραθύρου. Συγκεκριμένα, το δείγμα χωρίστηκε σε 15 υποσύνολα. Τα οκτώ πρώτα υποσύνολα χρησιμοποιήθηκαν για την διαδικασία της εκπαίδευσης και αναζήτησης των κατάλληλων υπερ-παραμέτρων και τα υπόλοιπα εφτά για την εξαγωγή των αποτελεσμάτων. Το κάθε υποσύνολο δεδομένων περιελάμβανε 8000 παραδείγματα, (περίπου 333 ημέρες συναλλαγών), ενώ 2160 παραδείγματα (τρίμηνο) χρησιμοποιήθηκαν για τον έλεγχο.

4.4 Επιλογή Υπερ-παραμέτρων (hyper parameters)

Η λειτουργία κάθε αλγορίθμου μηχανικής μάθησης καθορίζεται από διάφορες παραμέτρους που αφορούν την αρχιτεκτονική δομή του τελικού μοντέλου και τη διαδικασία εκπαίδευσής του. Οι παράμετροι αυτές αναφέρονται ως «υπερ-παράμετροι» γιατί παραμετροποιούν τη διαδικασία ανάπτυξης ενός μοντέλου σε ένα ανώτερο επίπεδο. Δηλαδή, δεν προσδιορίζουν άμεσα τη τελική σύνθεση των χαρακτηριστικών σε ένα μοντέλο πρόβλεψης, αλλά αφορούν τον τρόπο με τον οποίο θα διαμορφωθεί και εκπαιδευτεί το μοντέλο. Επιλογή των υπερ-παραμέτρων είναι ιδιαίτερα σημαντική σε ορισμένους αλγόριθμους, καθώς επηρεάζει σε μεγάλο βαθμό την αποτελεσματικότητα των αλγορίθμων. Για το λόγο αυτό, τα μοντέλα δοκιμάζονται σε πολλές διαφορετικές υπερ-παραμέτρους και αξιολογούνται. Για την επιλογή των υπερ-παραμέτρων χρησιμοποιήθηκε το AUC, ενώ στα αποτελέσματα που παρουσιάζονται στην ενότητα 4.5 καταγράφεται και η ορθότητα (accuracy). Οι υπερ-παράμετροι, οι οποίοι σημείωσαν την μεγαλύτερη απόδοση

στον δείκτη AUC, είναι και αυτοί οι οποίοι επιλέχθηκαν για την ανάπτυξη των τελικών μοντέλων. Οι υπερ-παράμετροι αυτοί αναφέρονται για κάθε μοντέλο ξεχωριστά παρακάτω.

Λογιστική παλινδρόμηση:

Ο αλγόριθμος που χρησιμοποιήθηκε έχει υλοποιηθεί από το scikit-learn[2] και οι υπερ-παράμετροι που επιλέχθηκαν είναι οι εξής:

- *C*: 0.6
- *penalty*: l1
- *solver*: liblinear
- *intercept_scaling:* 1.5
- *max_iter: 1000*

Νευρωνικά δίκτυα:

Για την υλοποίηση των νευρωνικών δικτύων χρησιμοποιήθηκε η βιβλιοθήκη Keras[3] η οποία βασίζεται πάνω στο TensorFlow[4] το οποίο έχει αναπτυχθεί από την ομάδα της Google.

Η αρχιτεκτονική του νευρωνικού δικτύου η οποία επιλέχθηκε, περιέχει οκτώ κρυφά επίπεδα και το κάθε επίπεδο δεκαπέντε τεχνητούς νευρώνες. Ως συνάρτηση ενεργοποίησης μεταξύ των κρυφών επιπέδων χρησιμοποιήθηκε η ELU και τα βάρη στην έναρξη της εκπαίδευσης αρχικοποιούνται χρησιμοποιώντας “He normal initializer” το οποίο συστήθηκε από τους He et al. [23]. Όπως έχει αναφερθεί στο επίπεδο εξόδου, η συνάρτηση ενεργοποίησης η οποία χρησιμοποιείται είναι η σιγμοειδής.

Τέλος, η συνάρτηση κόστους που χρησιμοποιήθηκε είναι η ίδια με αυτή της λογιστικής παλινδρόμησης, δηλαδή η “binary_crossentropy”. Ο αλγόριθμος που χρησιμοποιήθηκε για την ελαχιστοποίηση της συνάρτησης κόστους, είναι ο “RMSprop”[5] και οι τιμές των παραμέτρων που επιλέχθηκαν για το learning rate είναι 0.1 και για το momentum είναι 0.9, ενώ οι υπόλοιπές τιμές έμειναν ως έχει.

Τυχαία δέντρα αποφάσεων:

Ο αλγόριθμος των τυχαίων δασών που χρησιμοποιήθηκε έχει υλοποιηθεί από το scikit-learn. Οι υπερ-παράμετροι οι οποίοι επιλέχθηκαν είναι οι εξής:

- *n_estimators*: 100
- *max_depth*: 10

[2] https://scikit-learn.org/stable/ version: **0.24.1**
[3] https://keras.io/
[4] https://www.tensorflow.org/ version: **2.3.0**
[5] https://keras.io/api/optimizers/rmsprop/

- *min_samples*: 2
- *min_samples_leaf*: 8
- *max_leaf_nodes*: None
- *criterion*: gini
- *random_state*: 13
- *n_jobs*: -1

XGboost: Extreme Gradient Boosting:

Το XGboost[6] είναι μια βελτιστοποιημένη υλοποίηση του αλγορίθμου Extreme Gradient Boosting που υλοποιήθηκε από τους Chen & Guestrin [24]. Πλέον, είναι από τις πιο γνωστές και χρησιμοποιούμενες βιβλιοθήκες στο χώρο της μηχανικής μάθησης. Οι υπερ-παράμετροι οι οποίοι επιλέχθηκαν είναι:

- *n_estimators*: 50
- *learning_rate*: 0.5
- *max_depth*: 5
- min_child_weight: 1
- *reg_alpha*: 5
- *reg_lambda*: 6
- *gamma*: 0
- *booster*: gbtree
- *objective*: binary:logistic
- *eval_metric*: logloss
- *use_label_encoder*: False
- *n_jobs*: -1

4.5 Αποτελέσματα

Η εξαγωγή των αποτελεσμάτων γίνεται χρησιμοποιώντας ξανά τη διαδικασία του κινούμενου παραθύρου, στις υπόλοιπες ωριαίες εγγραφές των συνολικών δεδομένων. Συγκεκριμένα, τα δεδομένα χωρίστηκαν σε εφτά υποσύνολα. Το κάθε υποσύνολο είχε 8000 παραδείγματα για την εκπαίδευση και το βήμα του παραθύρου ήταν 2160 παραδείγματα. Τα αποτελέσματα που σημείωσε ο κάθε αλγόριθμος στα μέτρα αξιολόγησης, ορθότητα (accuracy) και AUC, στο κάθε παράθυρο του παραπάνω δείγματος ελέγχου παρουσιάζονται στους πίνακες 4.2-4.3.

[6] https://xgboost.readthedocs.io/en/latest/# version: **1.3.3**

Πίνακας 4.2: Η ορθότητα των μοντέλων στο δείγμα ελέγχου

Παράθυρο	Λογιστική Παλινδρόμηση	Νευρωνικό Δίκτυο	Τυχαία Δέντρα αποφάσεων	XGBoost
1	0,5394	0,5718	0,5630	0,5431
2	0,5264	0,5426	0,5426	0,5282
3	0,5458	0,5718	0,5542	0,5315
4	0,5551	0,5667	0,5639	0,5273
5	0,5106	0,5352	0,5176	0,5338
6	0,5111	0,5157	0,5352	0,5111
7	0,5488	0,5524	0,5524	0,5415
Μέση τιμή	**0,5339**	**0,5509**	**0,5470**	**0,5309**
Τυπ. Απόκλ	**0,0167**	**0,0195**	**0,0153**	**0,0099**

Πίνακας 4.3: Αποτελέσματα για τον δείκτη AUC

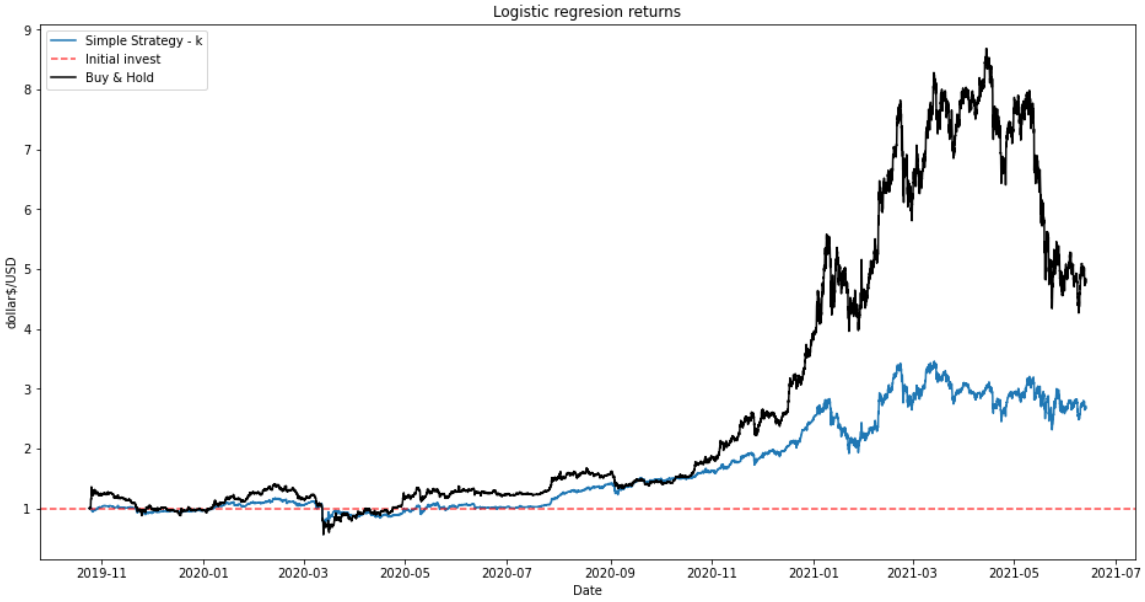
Παράθυρο	Λογιστική Παλινδρόμηση	Νευρωνικό Δίκτυο	Τυχαία Δέντρα αποφάσεων	XGBoost
1	0,5849	0,5912	0,5888	0,5593
2	0,5368	0,5518	0,5513	0,5394
3	0,5679	0,5908	0,5716	0,5470
4	0,5690	0,5859	0,5802	0,5457
5	0,5401	0,5340	0,5354	0,5383
6	0,5232	0,5164	0,5359	0,5250
7	0,5712	0,5667	0,5706	0,5627
Μέση τιμή	**0,5562**	**0,5624**	**0,5620**	**0,5454**
Τυπ. Απόκλ	**0,0209**	**0,0274**	**0,0197**	**0,0120**

Την υψηλότερη επίδοση στον δείκτη της ορθότητας και στο AUC την πέτυχε το νευρωνικό δίκτυο με μέσο όρο 0,5509 και 0,5624 αντίστοιχα. Επιπλέον, την χειρότερη επίδοση όπως φαίνεται και από τους πίνακες, την είχε το XGBoost με ορθότητα 0,5309 και AUC 0,5454.

Ωστόσο, για να φανεί καλύτερα η αποτελεσματικότητα των αλγορίθμων και να απαντηθεί το ερώτημα εάν θα μπορούσαν να χρησιμοποιηθούν για την εκμετάλλευση τους και αν αποφέρουν κέρδος για ένα επενδυόμενο κεφάλαιο, δοκιμάζεται μια απλή στρατηγική. Η στρατηγική ονόματι «k», συγκρίνεται με μία παθητική στρατηγική, αγοράς και διακράτησης (buy and hold) που ο επενδυτής επιλέγει να κρατήσει τα κρυπτονομίσματα του. Η στρατηγική «k» ακολουθεί τους δύο εξής κανόνες:

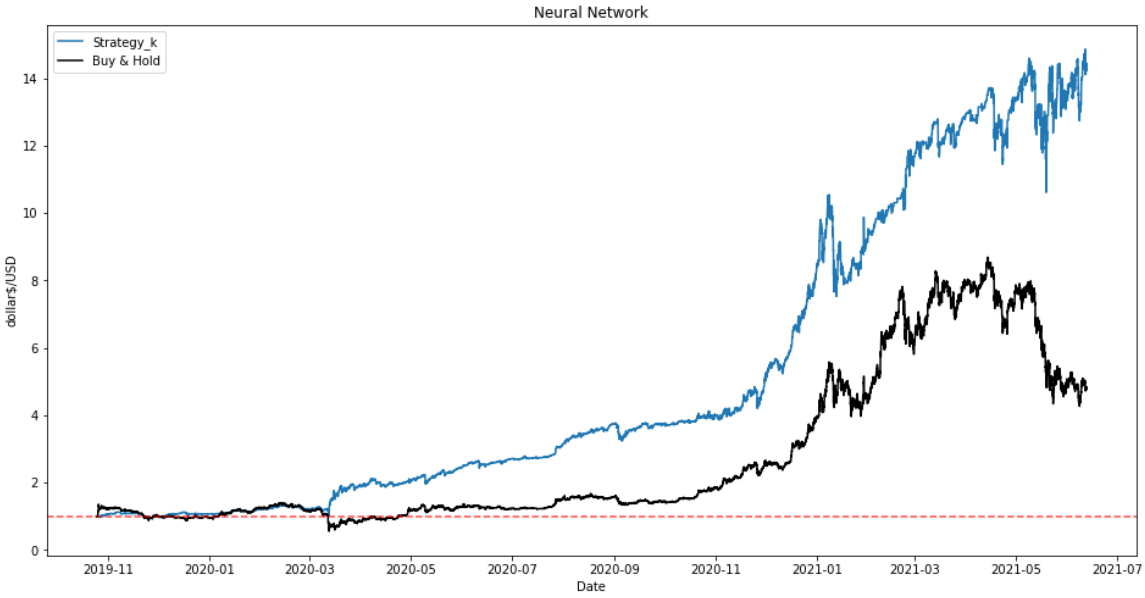
1. Εάν η πρόβλεψη του αλγορίθμου είναι θετική («Αγορά») τότε πραγματοποιείται αγορά του κρυπτονομίσματος με όλο το διαθέσιμο κεφάλαιο ή κρατείται η επένδυση ως έχει (εάν ήδη υπάρχει).
2. Εάν η πρόβλεψη του αλγορίθμου είναι αρνητική («πώληση») τότε ρευστοποιείται η επένδυση και δεν πραγματοποιείται αγορά.

Στα σχήματα 4.12 έως 4.15 απεικονίζεται η απόδοση της επένδυσης εφαρμόζοντας την στρατηγική «k» και την «Buy & Hold» για τον κάθε αλγόριθμο αντίστοιχα. Όπως φαίνεται στο διάγραμμα του σχήματος 4.12, τα αποτελέσματα του αλγορίθμου της λογιστικής παλινδρόμησης δεν είναι τόσο αποτελεσματικά/ικανοποιητικά. Η απλή στρατηγική «k» που εφαρμόστηκε, σε συνδυασμό με το μοντέλο της λογιστικής παλινδρόμηση, δεν έχει καλύτερη απόδοση από την στρατηγική Buy and hold. Ωστόσο, το μοντέλο παρά την κακή του απόδοση σε σύγκριση με την στρατηγική Buy and Hold, η απόδοση της επένδυσης παραμένει θετική και ελάχιστα ανοδική, δηλαδή δεν επιφέρει ζημίες, φτάνοντας απόδοση μέχρι 400%.



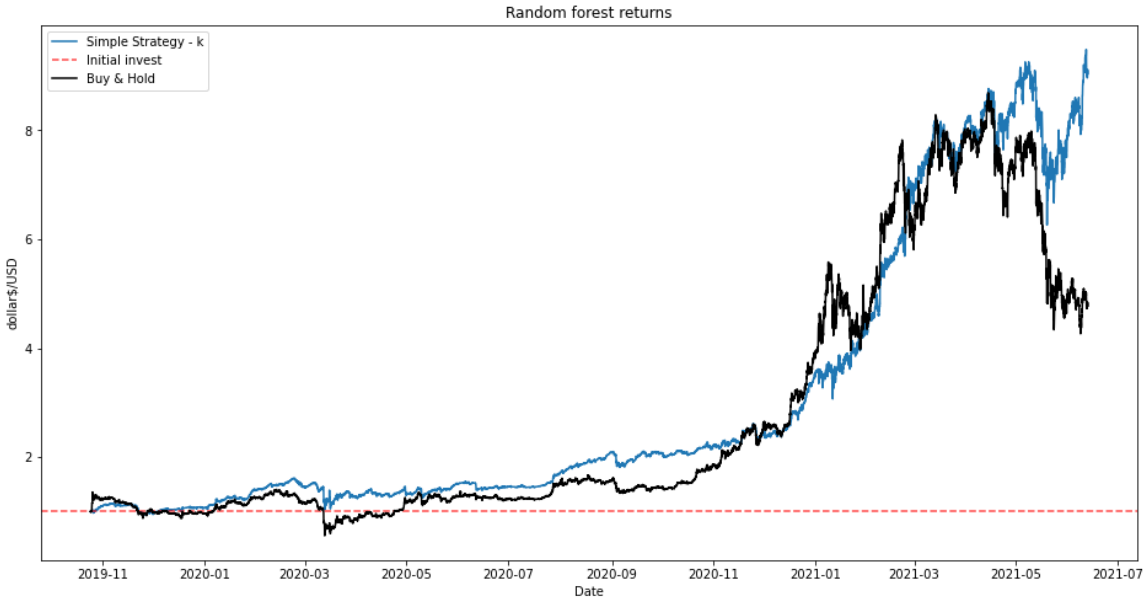
Σχήμα 4.12: Αποτελέσματα επένδυσης βάσει της λογιστικής παλινδρόμησης συγκριτικά με την στρατηγική Buy and hold

Σε αντίθεση με το μοντέλο της λογιστικής παλινδρόμησης, το νευρωνικό δίκτυο είναι αρκετά αποτελεσματικότερο. Η απόδοση του μοντέλου διατηρείται σε χαμηλά επίπεδα και παραμένει έτσι μέχρι το πρώτο τρίμηνο του 2020. Ύστερα, η απόδοση του μοντέλου αρχίζει να αυξάνεται σταδιακά μέχρι το Νοέμβριο του 2020, που μετά ξεσπάει μια απότομη ανοδική πορεία, με μία μικρή πτώση στις αρχές του 2021, όπως φαίνεται στο σχήμα 4.13. Η στρατηγική «k» νικάει κατά κράτος την στρατηγική Buy and hold με μέγιστη απόδοση να φτάνει μέχρι και 1400%.

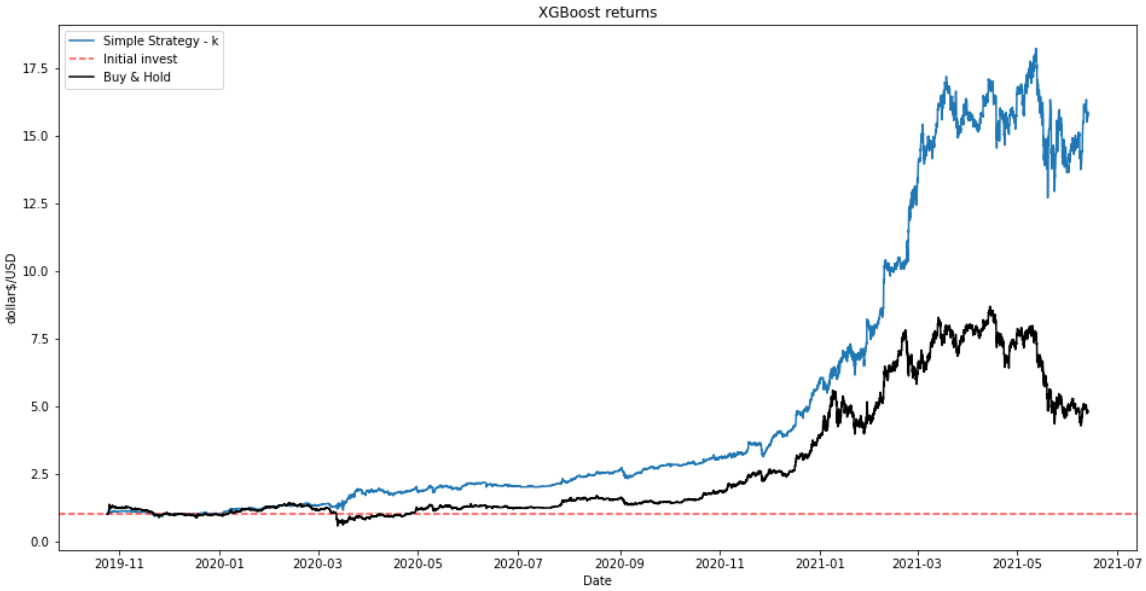


Σχήμα 4.13: Αποτελέσματα επένδυσης βάσει του νευρωνικού δικτύου συγκριτικά με την στρατηγική Buy and hold

Το μοντέλο των τυχαίων δέντρων αποφάσεων όπως διακρίνεται από το σχήμα 4.14 διατηρεί μια ανοδική πορεία καθ' όλη την διάρκεια των δοκιμών. Η στρατηγική, η οποία έχει την καλύτερη απόδοση, εναλλάσσεται συχνά μέχρι τις αρχές του 2021, οπότε και η στρατηγική «k» ξεπροβάλει ξαφνικά. Τέλος, αξιοσημείωτο είναι, πως την περίοδο του Μαΐου το 2021, ενώ η τιμή του Bitcoin πέφτει κατακόρυφα, το μοντέλο ακολουθεί ακριβώς αντίθετη πορεία, δηλαδή ανοδική και σημειώνει και την μεγίστη απόδοση με 900% περίπου.



Σχήμα 4.14: Αποτελέσματα επένδυσης βάσει των τυχαίων δέντρων αποφάσεων συγκριτικά με την στρατηγική Buy and hold



Σχήμα 4.15: Αποτελέσματα επένδυσης βάσει του αλγορίθμου XGBoost συγκριτικά με την στρατηγική Buy and hold

Όπως προαναφέρθηκε, ο αλγόριθμος XGBoost σημείωσε τις χειρότερες τιμές στα μέτρα αξιολόγησης, όμως αυτός αποτέλεσε μεγάλη έκπληξη, καθώς η απόδοση του μοντέλου ήταν η επικρατέστερη σε όλο το διάστημα των δοκιμών, σε σύγκριση με την Buy and hold στρατηγική (βλ. σχήμα 4.15). Η απόδοση του είναι όμοια με αυτή του νευρωνικού δικτύου μέχρι τις αρχές του 2021. Ύστερα, η απόδοση του εκρήγνυται φτάνοντας μέχρι 1700.5%, σημειώνοντας την μεγαλύτερη απόδοση απ' όλους τους άλλους αλγόριθμους. Αυτή η εντυπωσιακή απόδοση, σε σύγκριση με τα άλλα μοντέλα, μπορεί να αποδοθεί στο γεγονός ότι ο αλγόριθμος XGBoost δημιούργησε ένα εύρωστο μοντέλο, καθώς στα μέτρα αξιολόγησης σημείωσε την μικρότερη τυπική απόκλιση, άρα, για το κάθε παράθυρο δοκιμών είχε μια σταθερή καλή απόδοση σε σύγκριση με τους υπόλοιπους αλγορίθμους, οι οποίοι έχουν μεγαλύτερη τυπική απόκλιση και είναι λιγότερο συνεπείς στις προβλέψεις τους.

Επίσης, για την αξιολόγηση των στρατηγικών χρησιμοποιούνται τρεις δείκτες. Ο πρώτος δείκτης είναι ο δείκτης Sharpe και πρόκειται για τη μέση απόδοση της στρατηγικής διαιρεμένης με την τυπική απόκλιση όπως φαίνεται από τη σχέση (4.1).

$$Δείκτης\ Sharpe = \frac{Μέσος\ όρος\ απόδοσης}{Τυπική\ απόκλιση\ της\ απόδοση} \qquad (4.1)$$

Ο δεύτερος δείκτης είναι η μέση μηναία απόδοση θετικού σήματος. Με την έννοια θετικού σήματος νοείται ότι για τον υπολογισμό του μέσου όρου συμπεριλαμβάνονται τιμές μόνο όπου ο αλγόριθμος έχει προβλέψει άνοδο του κρυπτονομίσματος, δηλαδή για $\hat{y} = 1$ (αγορά). Όσο μεγαλύτερη είναι η τιμή αυτού του δείκτη, τόσο αποδοτικότερο θεωρείται το μοντέλο που χρησιμοποιήθηκε.

Ο τρίτος δείκτης είναι η μέση μηναία απόδοση αρνητικού σήματος. Αυτός είναι παρόμοιος με τον προηγούμενο δείκτη με μία μικρή διαφορά. Στον υπολογισμό του μέσου όρου συμπεριλαμβάνονται μόνο οι τιμές της απόδοσης, όπου ο αλγόριθμος έχει προβλέψει την τιμή του Bitcoin ως καθοδική, δηλαδή $\hat{y} = 0$ (πώληση). Προφανώς, αυτός ο δείκτης όσο πιο μικρός είναι, τόσο αποτελεσματικότερος θεωρείται ο αλγόριθμος.

Στους παρακάτω πίνακες 4.4-4.6 παρουσιάζονται οι μετρήσεις σε αυτούς τους δείκτες για τον κάθε αλγόριθμο.

Πίνακας 4.4: Αποτελέσματα απόδοσης θετικού σήματος για την στρατηγική «k»

Παράθυρο	Λογιστική Παλινδρόμηση	Νευρωνικό Δίκτυο	Τυχαία Δέντρα αποφάσεων	XGBoost	Buy & hold
1	9,8878	11,9826	17,7617	14,1738	6,6221
2	7,4044	34,4134	7,5713	32,1284	-2,0355
3	12,7838	30,0847	11,9105	11,5257	10,9505
4	23,6401	24,4120	22,8433	24,8300	8,2784
5	42,0795	40,3955	42,8791	65,8444	42,008
6	14,1432	78,0620	47,9493	51,2991	20,6618
7	4,0499	10,6998	9,6834	1,8288	-23,97
Μέση τιμή	**13,0115**	**32,8643**	**22,9426**	**28,8043**	**8,9308**
Τυπ. Απόκλ	**15,4776**	**21,0728**	**15,0402**	**21,1967**	**18,6864**

Πίνακας 4.5: Αποτελέσματα απόδοσης αρνητικού σήματος για την στρατηγική «k»

Παράθυρο	Λογιστική Παλινδρόμηση	Νευρωνικό Δίκτυο	Τυχαία Δέντρα αποφάσεων	XGBoost	Buy & hold
1	5,4318	1,7451	-3,2711	-0,5612	0
2	2,2919	-50,5162	-11,0726	-34,5932	0
3	8,8902	-2,7950	10,1811	10,4852	0
4	-12,0386	-5,4496	-7,0333	-6,4212	0
5	41,9623	45,6411	41,2287	20,9564	0
6	40,7597	3,8643	-29,5028	-26,2819	0
7	64,0725	-69,0931	-80,4527	-62,8566	0
Μέση τιμή	**3,3178**	**-10,9434**	**-11,4175**	**-14,1818**	**0**
Τυπ. Απόκλ	**33,1859**	**35,1040**	**34,6756**	**26,7827**	**0**

Πίνακας 4.6: Αποτελέσματα δείκτη Sharpe για τις στρατηγικές Buy and Hold και «k»

Παράθυρο	Λογιστική Παλινδρόμηση	Νευρωνικό Δίκτυο	Τυχαία Δέντρα αποφάσεων	XGBoost	B&H Sharpe ratio
1	0,0210	0,0287	0,0401	0,0309	0,0139
2	-0,0068	0,0403	0,0081	0,0378	-0,0023
3	0,0272	0,0621	0,0258	0,0240	0,0247
4	0,0675	0,0670	0,0620	0,0647	0,0229
5	0,0596	0,0593	0,0619	0,1022	0,0633
6	0,0192	0,0900	0,0654	0,0681	0,0289
7	0,0047	0,0120	0,0113	0,0021	-0,0285
Μέση τιμή	**0,0262**	**0,0513**	**0,0392**	**0,0471**	**0,0176**
Τυπ. Απόκλ	**0,0266**	**0,0242**	**0,0228**	**0,0309**	**0,0263**

Το νευρωνικό δίκτυο και ο αλγόριθμος XGBoost, όπως ήταν αναμενόμενο λόγω της ανοδικής πορείας της απόδοσης τους, σημείωσαν τις μεγαλύτερες τιμές στον δείκτη Sharpe με τιμές 0,0513 και 0,0471 αντίστοιχα. Επίσης, την μεγαλύτερη τιμή στην απόδοση του θετικού σήματος την πέτυχε το νευρωνικό δίκτυο με τιμή ίση με 32,8643, ενώ την μικρότερη απόδοση αρνητικού σφάλματος την πέτυχε ο αλγόριθμος XGBoost με τιμή ίση με -14,1818. Επιπλέον, η λογιστική παλινδρόμηση στο πέμπτο, έκτο και έβδομο «παράθυρο» σημείωσε μεγάλες τιμές στην απόδοση αρνητικού σφάλματος, μάλιστα μεγαλύτερες τιμές από ότι στα αντίστοιχα «παράθυρα» στην απόδοση θετικού σφάλματος. Αυτό διατυπώνεται και στο γράφημα που απεικονίζεται η συνολική απόδοση του μοντέλου (σχήμα 4.12), όπου χάνονται μεγάλες ευκαιρίες για την εκμετάλλευση του κέρδους. Τέλος, αν και δεν είναι ξεκάθαρο από το σχήμα (4.13), εάν τα τυχαία δέντρα αποφάσεων σε συνδυασμό με τη στρατηγική «k» είναι αποτελεσματικότερα από την στρατηγική Buy and hold, εξετάζοντας τα αποτελέσματα των τυχαίων δέντρων αποφάσεων στους δείκτες που εξετάστηκαν θα μπορούσε κανείς να βγάλει το συμπέρασμα ότι τα τυχαία δέντρα αποφάσεων έχουν καλύτερη απόδοση.

Κεφάλαιο 5ο

Επίλογος

Στο πλαίσιο αυτής της διπλωματικής εργασίας εξετάστηκε η αποτελεσματικότητα τεσσάρων αλγορίθμων, (η λογιστική παλινδρόμηση, νευρωνικά δίκτυα, τυχαία δέντρα αποφάσεων, XGBoost) για το σχεδιασμό επενδυτικών στρατηγικών στην αγορά κρυπτονομισμάτων του Bitcoin. Με βάση αυτή την ανάλυση που πραγματοποιήθηκε, εφαρμόστηκε μια απλή στρατηγική για την αυτοματοποιημένη πραγματοποίηση συναλλαγών και τα αποτελέσματα συγκρίθηκαν με την απόδοση της επένδυσης, εάν ο επενδυτής ακολουθούσε μια παθητική στρατηγική. Όλα τα μοντέλα είχαν θετική απόδοση κέρδους, ωστόσο ορισμένα από αυτά δεν κατάφεραν να ξεπεράσουν την απόδοση της παθητικής επένδυσης.

Σε μια περίοδο που η τιμή του Bitcoin βίωσε άνοδο, από τις αρχές του 2020 μέχρι τα μέσα του 2021, οι αλγόριθμοι οι οποίοι ξεχώρισαν και φάνηκαν να ανταποκρίνονται καλύτερα σύμφωνα με τα αποτελέσματα είναι τα νευρωνικά δίκτυά και ο αλγόριθμος XGboost, σημειώνοντας, μάλιστα, μέγιστη απόδοση 1450% και 1700% αντίστοιχα. Συγκριτικά αναφέρεται ότι η τιμή του Bitcoin αυξήθηκε 800% στο προαναφερόμενο διάστημα. Το μοντέλο της λογιστικής παλινδρόμησης ήταν το λιγότερο ικανοποιητικό, καθώς δεν κατάφερε να ξεπεράσει σε καμία χρονική στιγμή την απόδοση της παθητικής στρατηγικής.

Συνεπώς, σύμφωνα με τα αποτελέσματα της εργασίας, τα νευρωνικά δίκτυα και ο αλγόριθμος XGboost φαίνεται πως θα μπορούσαν να χρησιμοποιηθούν για αυτό το πρόβλημα και να μελετηθούν περαιτέρω. Συγκεκριμένα, θα μπορούσαν να δοκιμαστούν πιο πολύπλοκες αρχιτεκτονικές νευρωνικών δικτύων, όπως τα Recurrent neural networks (RNN). Μία άλλη πρόταση είναι η χρήση Convolutional neural networks (CNN), τα οποία έχουν πετύχει σημαντικά αποτελέσματα στην ταξινόμηση εικόνων. Συγκεκριμένα, τα CNN στις έρευνες που έχουν χρησιμοποιηθεί για την ανάλυση χρονοσειρων, είχαν ικανοποιητικά αποτελέσματα [25-27]. Επιπλέον, έχουν πολλά περιθώρια να ερευνηθούν, καθώς δεν έχουν μελετηθεί αρκετά σε προβλήματα χρονοσειρών συμφώνα με την έρευνα των Ozbayoglu et al. [28]. Επίσης, θα μπορούσαν να χρησιμοποιηθούν επιπλέον χαρακτηριστικά (attributes) στα δεδομένα εκπαίδευσης, όπως πληροφορίες από το blockchain ή πληροφορίες που γίνονται γνωστές στο διαδίκτυο, σχετικές με το εξεταζόμενο κρυπτονόμισμα. Τέλος, στην έρευνα μπορεί να συμπεριληφθεί και το κόστος συναλλαγών (transaction costs), καθώς αυτό σε μεγάλο όγκο συναλλαγών μπορεί να επηρεάσει σε σημαντικό βαθμό τα τελικά αποτελέσματα.

Βιβλιογραφία

[1] Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, *90*, 106181.

[2] Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review*, 21260.

[3] Lewis, A. (2018). The Basics of Bitcoins and Blockchains_ An Introduction to Cryptocurrencies and the Technology that Powers Them

[4] McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133

[5] Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013, June). Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml* (Vol. 30, No. 1, p. 3).

[6] Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical Evaluation of Rectified Activations in Convolutional Network. *ArXiv, abs/1505.00853*.

[7] Clevert, D., Unterthiner, T., & Hochreiter, S. (2016). Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *arXiv: Learning*.

[8] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation*. California Univ San Diego La Jolla Inst for Cognitive Science.

[9] Breiman, L., Freidman, J.H., Olshen, R.A., & Stone, C.J. (1984). CART: Classification and Regression Trees.

[10] Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, *20*(8), 832-844.

[11] Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

[12] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[13] Freund, Y., & Schapire, R.E. (1996). Experiments with a New Boosting Algorithm. *ICML*.

[14] Breiman, L. (1997). *Arcing the edge* (Vol. 7). Technical Report 486, Statistics Department, University of California at Berkeley

[15] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

[16] Ibrahim, A., Kashef, R., & Corrigan, L. (2021). Predicting market movement direction for bitcoin: A comparison of time series modeling methods. *Computers & Electrical Engineering*, *89*, 106905.

[17] Hsu, M. W., Lessmann, S., Sung, M. C., Ma, T., & Johnson, J. E. (2016). Bridging the divide in financial market forecasting: machine learners vs. financial economists. *Expert Systems with Applications*, *61*, 215-234.

[18] Kara, Y., Boyacioglu, M. A., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert systems with Applications*, *38*(5), 5311-5319.

[19] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert systems with applications*, *42*(1), 259-268.

[20] Nakano, M., Takahashi, A., & Takahashi, S. (2018). Bitcoin technical trading with artificial neural network. *Physica A: Statistical Mechanics and its Applications*, *510*, 587-609.

[21] Thakur, M., & Kumar, D. (2018). A hybrid financial trading support system using multi-category classifiers and random forest. *Applied Soft Computing*, *67*, 337-349.

[22] Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PloS one*, *12*(7), e0180944.

[23] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1026-1034.

[24] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[25] Sezer, O. B., & Ozbayoglu, A. M. (2018). Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach. *Applied Soft Computing*, *70*, 525-538.

[26] Alonso-Monsalve, S., Suárez-Cetrulo, A. L., Cervantes, A., & Quintana, D. (2020). Convolution on neural networks for high-frequency trend prediction of cryptocurrency exchange rates using technical indicators. *Expert Systems with Applications*, *149*, 113250.

[27] Gunduz, H., Yaslan, Y., & Cataltepe, Z. (2017). Intraday prediction of Borsa Istanbul using convolutional neural networks and feature correlations. *Knowledge-Based Systems*, *137*, 138-148.

[28] Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications: A survey. *Applied Soft Computing*, *93*, 106384.