ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μέθοδοι Νευρωνικών Δικτύων Γράφων για πρόβλεψη τιμής του Bitcoin βασισμένες στους γράφους συναλλαγών του Blockchain

ΧΑΡΑΛΑΜΠΟΣ ΕΠ. ΚΛΕΙΤΣΙΚΑΣ

Επιβλέπων : Νεκτάριος Κοζύρης
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2022

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μέθοδοι Νευρωνικών Δικτύων Γράφων για πρόβλεψη τιμής του Bitcoin βασισμένες στους γράφους συναλλαγών του Blockchain

ΧΑΡΑΛΑΜΠΟΣ ΕΠ. ΚΛΕΙΤΣΙΚΑΣ

Επιβλέπων : Νεκτάριος Κοζύρης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 17/10/2022.

.....	.....	.....
Νεκτάριος Κοζύρης Καθηγητής Ε.Μ.Π.	Γεώργιος Γκούμας Αναπληρωτής Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2022

.....

Χαράλαμπος Επ. Κλειτσίκας

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Υπάρχει πληθώρα δημοσιεύσεων σχετικά με τη πρόβλεψη του δημοφιλέστερου κρυπτονομίσματος αυτή τη στιγμή στην αγορά, του Bitcoin. Οι υπάρχουσες μελέτες χρησιμοποιούν διάφορα κοινωνικά, οικονομικά και τεχνικά χαρακτηριστικά μέσω συμβατικών μεθόδων μηχανικής μάθησης. Η κύρια ερευνητική συνεισφορά και ο σκοπός της παρούσας εργασίας είναι να εισάγει για πρώτη φορά στην βιβλιογραφία μια νέα μέθοδο πρόβλεψης της τιμής του Bitcoin χρησιμοποιώντας Νευρωνικά Δίκτυα Γράφων (GNN) με εισόδους τους γράφους συναλλαγών του Blockchain του Bitcoin. Αποδεικνύουμε ότι εκμεταλλευόμενα τη συνδεσιμότητα του δικτύου συναλλαγών καθώς και τα δομικά χαρακτηριστικά του, τα GNN είναι ιδανικά για να αναγνωρίζουν τοπολογικά μοτίβα τα οποία σε συνδυασμό με την παρελθοντική τιμή του BTC συμβάλλουν καθοριστικά στην πρόβλεψη της τιμής του. Κατασκευάζουμε διάφορες παραλλαγές του γράφου συναλλαγών του Bitcoin και προσδίδουμε επιπρόσθετες εξωτερικές πληροφορίες στους κόμβους του, έτσι ώστε να επεξεργάζονται με γραφοκεντρικό τρόπο, επιτυγχάνοντας ακόμα καλύτερες αποδόσεις για τις μακροχρόνιες προβλέψεις. Εξετάζουμε διάφορα σενάρια γράφων και βασικές αρχιτεκτονικές για τα GNN. Μελετούμε την απόδοση των μοντέλων στο πρόβλημα παλινδρόμησης για την ακριβή πρόβλεψη της τιμής του BTC για μία και έξι ώρες μπροστά. Γίνεται σύγκριση με τις μέχρι τώρα μεθόδους της βιβλιογραφίας και προκύπτει ότι η μέθοδος μας υπερτερεί έναντι των περισσότερων. Μέσω της δουλειάς μας ανοίγεται η δυνατότητα έρευνας σε ένα νέο συναρπαστικό πεδίο, τη πρόβλεψη τιμής με τη χρήση των state-of-the-art GNNs στους γράφους συναλλαγών, με πιο προχωρημένες τεχνικές που θα βγάλουν ακόμα καλύτερα αποτελέσματα και θα εφαρμοστούν όχι μόνο στο Bitcoin αλλά και σε άλλα διάσημα κρυπτονομίσματα.

Λέξεις Κλειδιά: Blockchain, Bitcoin, Πρόβλεψη Τιμής, Μηχανική Μάθηση, Νευρωνικά Δίκτυα Γράφων, Γράφοι Συναλλαγών, GraphSAGE, GAT, Κρυπτονομίσματα, Χρονοσειρές, Αλγόριθμος Κατασκευής Γράφων, Κυλιόμενο Παράθυρο, Παλινδρόμηση

Abstract

There is a plethora of publications about the price prediction of the most popular cryptocurrency currently on the market, Bitcoin. Existing studies use a variety of socio-economic and technical factors through conventional machine learning methods. The main research contribution and purpose of this paper is to introduce for the first time in the literature, a new method for Bitcoin price prediction, which uses Graph Neural Networks (GNN) that take as inputs the transaction graphs of the Bitcoins Blockchain. We demonstrate that by exploiting the connectivity of the transaction network as well as its structural features, GNNs are ideally suited to identify topological patterns which, combined with the past price of BTC, are instrumental in predicting its future price. We construct several variants of the Bitcoin transaction graphs and assign additional external information to its nodes so that they are processed in a graph-centric manner, achieving even better performance for long-term predictions. We consider several graph scenarios and basic architectures for GNNs. We study the performance of the models in the regression problem to accurately forecast the price of BTC for one and six hours ahead. We compare the existing methods in the literature with ours and find that our method outperforms most of them. Through our work we open up the possibility of research in an exciting new field, price prediction using state-of-the-art GNNs, with more advanced techniques that will yield even better results and will be applied not only to Bitcoin but also to other famous cryptocurrencies.

Keywords: Price Prediction, Graph Neural Networks, Bitcoin, Blockchain, Transaction Graph, GNN, Machine Learning, Graph Regression, Forecasting, GraphSAGE, GAT, cryptocurrencies, Time Series, Transaction Graph Construction Algorithm, Sliding Window

στους γονείς μου και την αδερφή μου

Ευχαριστίες

Πρωτίστως θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Νεκτάριο Κοζύρη, ο οποίος με εμπιστεύτηκε με την ανάθεση για την εκπόνηση της παρούσας διπλωματικής εργασίας.
Απευθύνω τις θερμότερες ευχαριστίες μου στην κα Αικατερίνη Δόκα, η οποία όχι μόνο μου πρότεινε ένα εξαιρετικά ενδιαφέρον θέμα αλλά ήταν δίπλα μου σε όλη τη διάρκεια της διπλωματικής μέσω των εποικοδομητικών συζητήσεών μας, των καίριων και πάντα βοηθητικών παρατηρήσεών της και της κατανόησης που έδειξε σε προσωπικά δύσκολους καιρούς για μένα.
Θερμά θα ήθελα να ευχαριστήσω τον Αγησίλαο Πολίτη, για την πολύτιμη καθοδήγηση και βοήθειά του, καθώς και το χρόνο που αφιέρωνε για να συζητάμε τα προβλήματα τα οποία προέκυπταν κατά τη διάρκεια της εργασίας αυτής.
Χωρίς την άνευ όρων και πηγαία αγάπη της οικογένειάς μου και την αμέριστη υποστήριξή της δεν θα μπορούσα να έχω πετύχει τίποτα και δεν θα ήμουν αυτός που είμαι τώρα. Οι λέξεις δεν είναι αρκετές για να εκφράσω την ευγνωμοσύνη και την αγάπη μου στον πατέρα μου Επαμεινώνδα, τη μητέρα μου Αρετή-Ελένη και την αδερφή μου Μαρία. Τους αφιερώνω απόλυτα, τη παρούσα εργασία.
Θα ήθελα να ευχαριστήσω επίσης τους ανθρώπους που κατά τη διάρκεια των σπουδών μου ήταν δίπλα μου καθημερινά, με στήριζαν ανιδιοτελώς και υπέμειναν τις ιδιοτροπίες μου. Να ξέρουν ότι τους αγαπώ πολύ.

Θα ήθελα να αφιερώσω τις τελευταίες αυτές γραμμές στον εκλιπόντα παππού μου Χαράλαμπο. Ακόμα θυμάμαι τη συγκίνηση του, όταν είχε μάθει ότι πέρασα στη σχολή μου. Στις πιο δύσκολες στιγμές, η μνήμη του με κράτησε όρθιο. Παππού, ελπίζω να είσαι περήφανος.

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κεφάλαιο 1: Εισαγωγή

1.1: Πρόβλεψη τιμής Bitcoin

Όπως ορίστηκε από τον δημιουργό του, το Blockchain του Bitcoin είναι ένα ηλεκτρονικό σύστημα πληρωμών το οποίο αφαιρεί την ανάγκη ύπαρξης κάποιου κεντρικού διαμεσολαβητή (π.χ. τράπεζες) για να εξασφαλιστεί η εγκυρότητα των συναλλαγών που πραγματοποιούνται εντός του συστήματος [17]. Το πρωτόκολλο του είναι κατανεμημένο και ψευδοανώνυμο και νόμισμά του αποτελεί το Bitcoin. Στη διάρκεια του χρόνου οι συναλλαγές και οι πληροφορίες τους μπαίνουν σε μπλοκς, τα οποία συνδέονται μεταξύ τους και σχηματίζουν μια αλυσίδα (blockchain). Οι πληροφορίες αυτές είναι δημόσια προσβάσιμες και έτσι υπάρχει πλήρης διαφάνεια αλλά παράλληλα και ανωνυμία καθώς, ενώ φαίνονται ποιοί λογαριασμοί πραγματοποιούν ποιες συναλλαγές, δεν υπάρχει γνώση της ταυτότητας των ατόμων στα οποία αντιστοιχούν οι εν λόγω λογαριασμοί. Λόγω των όσων πρεσβεύει το σύστημα του Bitcoin (αποκέντρωση, εξάλειψη τραπεζών από μεσάζοντες, διαφάνεια, πλήρη ασφάλεια συναλλαγών κλπ) απέκτησε εκθετικά γρήγορα τεράστια δημοτικότητα και μετέτρεψε το κρυπτονόμισμα που χρησιμοποιεί από ένα ψηφιακό μέσο συναλλαγής σε ένα επενδυτικό κεφάλαιο καθώς το Bitcoin μπορεί να ανταλλαχθεί και με ήδη υπάρχοντα νομίσματα. Μάλιστα η κεφαλαιοποίηση της αγοράς του από τα 1.3 δισεκατομμύρια δολάρια που είχε στις 1 Μαΐου του 2013, έφτασε τα 100.1 δισεκατομμύρια δολάρια στις 21 Οκτωβρίου του 2017 και μόλις τέσσερα χρόνια μετά έφτασε την υψηλότερη κεφαλαιοποίηση του, στις 9 Νοεμβρίου του 2021 ύψους 1.28 τρισεκατομμυρίων δολαρίων. Τη στιγμή συγγραφής της παρούσας διπλωματικής (10 Οκτωβρίου του 2022) η κεφαλαιοποίηση του Bitcoin αγγίζει τα 371 δισεκατομμύρια δολάρια. Είναι προφανές ότι το Bitcoin χαρακτηρίζεται κυρίως από τις τεράστιες διακυμάνσεις στις τιμές του, οι οποίες αφήνουν περιθώρια σημαντικού οικονομικού κέρδους αλλά και με μεγάλο ρίσκο. Κρίνεται θεμιτή έτσι η εύρεση μεθόδων που θα προβλέπουν σε ικανοποιητικό επίπεδο τη τιμή του Bitcoin τόσο βραχυπρόθεσμα όσο και μακροπρόθεσμα. Το εγχείρημα είναι ιδιαίτερα απαιτητικό, καθώς η τιμή του Bitcoin εξαρτάται από πολλούς παράγοντες και, σε αντίθεση με τα παραδοσιακά νομίσματα ή τον χρυσό, δεν αρκεί να χρησιμοποιήσουμε παραδοσιακούς γενικούς οικονομικούς δείκτες [45]. Η δυσκολία του προβλήματος σε συνδυασμό με τη δημοτικότητα του Bitcoin και τις ευκαιρίες που παρουσιάζει έχει πυροδοτήσει το ενδιαφέρον της ακαδημαϊκής κοινότητας, ειδικά των ερευνητών στο πεδίο της μηχανικής μάθησης.

1.2: Στόχος Εργασίας

Υπάρχει πληθώρα δημοσιεύσεων σχετικά με τη πρόβλεψη τιμής κρυπτονομισμάτων και συγκεκριμένα του δημοφιλέστερου κρυπτονομίσματος αυτή τη στιγμή στην αγορά, το Bitcoin [17]. Εξίσου πολλοί είναι και οι μέθοδοι πρόβλεψης του, από στατιστικά μοντέλα [35,36], μέχρι σύνθετα μοντέλα μηχανικής μάθησης [39]. Πολλές παράμετροι έχουν ληφθεί υπόψιν, από οικονομικούς δείκτες όπως τους δείκτες του χρηματιστηρίου και τη τιμή του χρυσού [37], μέχρι τα σχόλια χρηστών σε μέσα κοινωνικής δικτύωσης [40], και αρκετά μοντέλα έχουν βγάλει αρκετά ενθαρρυντικά αποτελέσματα [27,35,41].

Η κύρια ερευνητική συνεισφορά και ο σκοπός της παρούσας εργασίας είναι να εισάγει για πρώτη φορά στην βιβλιογραφία, μια νέα μέθοδο πρόβλεψης της τιμής του Bitcoin χρησιμοποιώντας Νευρωνικά Δίκτυα Γράφων (GNN) με εισόδους τους γράφους συναλλαγών του Blockchain του Bitcoin. Θέλουμε να αναδείξουμε έτσι, όχι μόνο την προβλεπτική ισχύ που παρουσιάζουν τα δομικά χαρακτηριστικά του δικτύου συναλλαγών του Blockchain καθώς και η ίδια η συνδεσιμότητα του η οποία αποτυπώνει την ροή του Bitcoin (και άρα του χρήματος) στο χρόνο, αλλά και να αποδείξουμε ότι εκμεταλλευόμενα αυτή τη δομή και τα χαρακτηριστικά, τα GNN είναι ιδανικά για να επεξεργαστούν μέσω των γράφων συναλλαγών τοπολογικές πληροφορίες του Blockchain οι οποίες, μαζί με την ίδια τη τιμή του BTC, συμβάλλουν καθοριστικά στην πρόβλεψη της τιμής του. Δείχνουμε ότι τα GNN μαθαίνουν αυτόματα μέσω των embeddings των δομικών χαρακτηριστικών των κόμβων του γράφου, δηλαδή των συναλλαγών, τις υποκείμενες δομές των γράφων που έχουν προβλεπτική ισχύ. Έτσι αποκτούν την ικανότητα να γενικεύουν και να αναγνωρίζουν μοτίβα και χαρακτηριστικές ιδιότητες συναλλαγών σε πραγματικούς μεταγενέστερους χρονικά γράφους του Bitcoin στους οποίους δεν έχουν εκπαιδευτεί.

Κατασκευάζουμε τις δικές μας παραλλαγές του γράφου συναλλαγών του Bitcoin, οι οποίες βασίζονται στον αλγόριθμο κατασκευής όπως παρουσιάζεται από τους Tharani κ.α. [26], τον οποίο επεκτείνουμε, τροποποιούμε και παραμετροποιούμε έτσι ώστε να είναι κατάλληλος για το είδος του προβλήματος το οποίο μελετάμε. Αποδίδουμε, από όσο γνωρίζουμε για πρώτη φορά στη βιβλιογραφία, επιπρόσθετες εξωτερικές πληροφορίες στους κόμβους του δικτύου των συναλλαγών, έτσι ώστε να επεξεργάζονται με έναν γραφοκεντρικό τρόπο, επιτυγχάνοντας ακόμα καλύτερες αποδόσεις για τις μακροχρόνιες προβλέψεις. Συγκεκριμένα, εξετάζουμε διάφορα σενάρια που αφορούν παραλλαγές κατασκευής των γράφων συναλλαγών, όπως τρίωρης και εξάωρης διάρκειας γράφους που περιλαμβάνουν είτε μόνο δομικά χαρακτηριστικά και τη τιμή του Bitcoin στις πληροφορίες των συναλλαγών είτε και οικονομικούς δείκτες όπως ο δεκαήμερος εκθετικός κυλιόμενος μέσος όρος της χρονοσειράς και ο όγκος δολαρίων που συναλλάσσεται με BTC σε ωριαία βάση. Μελετάμε βασικές αρχιτεκτονικές για τα GNN, συγκεκριμένα τις GraphSAGE και GATv2 και για την εύρεση της βέλτιστης αρχιτεκτονικής διευρευνάμε ένα ευρύ φάσμα των υπερπαραμέτρων τους.

Μελετούμε την απόδοση των μοντέλων στο πρόβλημα παλινδρόμησης, με σκοπό την ακριβή πρόβλεψη της τιμής του BTC για μία και έξι ώρες μπροστά. Γίνεται σύγκριση με τις μεθόδους της υπάρχουσας βιβλιογραφίας και προκύπτει ότι η μέθοδος μας παρουσιάζει μικρότερες τιμές RMSE και MAPE, υπερτερεί συνεπώς έναντί αρκετών μεθόδων [42, 43, 44] και είναι συγκρίσιμη με άλλες[41]. Μέσω της δουλειάς μας, ανοίγεται η δυνατότητα έρευνας σε ένα νέο συναρπαστικό πεδίο, τη πρόβλεψη τιμής με τη χρήση των state-of-the-art GNNs στους γράφους συναλλαγών, με πιο προχωρημένες τεχνικές που θα βγάλουν ακόμα καλύτερα αποτελέσματα και θα εφαρμοστούν τόσο στο Bitcoin όσο σε άλλα διάσημα κρυπτονομίσματα.

1.3: Βιβλιογραφική Ανασκόπηση

Όπως αναφέρθηκε, η δυσκολία του προβλήματος της πρόβλεψης της ακριβής τιμής του Bitcoin, σε συνδυασμό με τη δημοτικότητα του και τις ευκαιρίες για οικονομικό κέρδος που παρουσιάζει, έχει πυροδοτήσει το ενδιαφέρον της ακαδημαϊκής κοινότητας, ειδικά των ερευνητών στο πεδίο της μηχανικής μάθησης και της στατιστικής. Οι περισσότερες έρευνες όμως εστιάζουν στην πρόβλεψη για την τιμή της επόμενης μέρας, ενώ η δική μας εργασία μελετά τη πρόβλεψη για μία και έξι ώρες μετά. Για αυτό το λόγο θα αναφερθούμε στις έρευνες με τις οποίες μπορούμε να συγκρίνουμε τα αποτελέσματά μας.

Οι Schulte κ.α. [42] ανέπτυξαν στατιστικά μοντέλα όπως το VAR (Vector Autoregression) και το SARIMAX (Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors) και νευρωνικά δίκτυα τύπου LSTM (Long Short-term Memory) και BiLSTM (Bidirectional LSTM), για να προβλέψουν τη τιμή του Bitcoin μια ώρα μετά τα διαθέσιμα δεδομένα. Προέκυψε ότι το καλύτερο μοντέλο ήταν το LSTM το οποίο πέτυχε σφάλμα MAPE ίσο με 3.52%.

Οι Jiang κ.α. [43] ανέπτυξαν μοντέλα βαθιάς μηχανικής μάθησης, συγκεκριμένα MLP (Multi-Layer Perceptron), LSTM και GRU (Gated Recurrent Network), μεταξύ των οποίων τα καλύτερα αποτελέσματα προκύπτουν πάλι με LSTM και συγκεκριμένα υπολόγισαν το RMSE σφάλμα το οποίο ήταν ίσο με 125.387.

Οι Kilimci κ.α. [44] ανέπτυξαν και αυτοί μοντέλα μηχανικής μάθησης και συγκεκριμένα CNNs (convolutional neural networks), LSTMs, ConvLSTM, (convolutional LSTM) και CNN-LSTM. Μέσω εκτεταμένων πειραμάτων, προέκυψε ότι το υβριδικό μοντέλο ConvLSTM έκανε προβλέψεις για τη τιμή του Bitcoin στην επόμενη ώρα με σφάλμα MAPE 2.4076%.

Τέλος, οι Sridhar κ.α. [41] για τη πρόβλεψη της επόμενης ώρας χρησιμοποίησαν DB δίκτυα (Deep Belief Networks), τα οποία είχαν προεκπαιδευτεί με RBMs (Restricted Boltzmann Machines), πετυχαίνοντας σφάλμα RMSE ίσο με 66.560 και MAPE ίσο με 0.384%.

Για τη πρόβλεψη έξι ώρες μετά δεν βρέθηκε κάποιο ερευνητικό άρθρο το οποίο να εξετάζει αυτή την περίπτωση.

Στις παραπάνω έρευνες καθώς και στις περισσότερες στην βιβλιογραφία, δεν διερευνάται η επίδραση των γράφων συναλλαγών στη πρόβλεψη της τιμής του Bitcoin. Οι πρώτοι οι οποίοι το εισάγουν στην βιβλιογραφία είναι οι Ancora κ.α. [24]. Συγκεκριμένα, ορίζουν την έννοια των graphlets, τα οποία είναι ουσιαστικά υπογράφοι του γράφου συναλλαγών του Bitcoin για κάποια χρονικά διαστήματα, με συγκεκριμένες τοπικές για το εκάστοτε graphlet τοπολογικές ιδιότητες. Μέσω στατιστικών ελέγχων αποδεικνύουν ότι κάποια είδη graphlets με συγκεκριμένες ιδιότητες σε συνδυασμό με την συχνότητα που παρουσιάζονται στο δίκτυο, έχουν μεγάλη προβλεπτική ισχύ για τη τιμή του Bitcoin. Χρησιμοποιώντας τον πίνακα συχνοτήτων εμφάνισης και τον πίνακα πληθικότητας των chainlets σε Random Forest αλγορίθμους, δείχνουν πως βελτιώνονται οι RMSE μετρικές για προβλέψεις 1,5,10 και 20 ημερών αφού χρησιμοποιηθούν τα chainlets σε σχέση με πριν.
Οι Li κ.α. [25] ορίζουν και αυτοί, ωστόσο με διαφορετική αναπαράσταση από ότι οι Ancora κ.α. [24], τον γράφο του Bitcoin και στη συνέχεια, αντλώντας τα τοπολογικά χαρακτηριστικά διαφόρων υπογράφων του δικτύου, κατασκευάζουν τον πίνακα συχνοτήτων εμφάνισης τους. Αυτα τα χαρακτηριστικά τα τροφοδοτούν σε SVM (Support Vector Machines) και τα αποτελέσματα τους για τις προβλέψεις της τιμής του Bitcoin για την επόμενη μέρα έχουν πολύ χαμηλές τιμές MAPE, αναδεικνύοντας έτσι με τη σειρά τους την ισχυρή προβλεπτική ικανότητα των τοπολογικών μοτίβων στο γράφο συναλλαγών του Bitcoin.
Και οι δύο έρευνες όμως, χρησιμοποιούν τον γράφο συναλλαγών για να αντλήσουν μέσω graph mining τεχνικών στατιστικά χαρακτηριστικά για τα είδη των υπογράφων που απαντώνται σε αυτόν (συχνότητα εμφάνισης, πληθικότητα κλπ) και στη συνέχεια να τα τροφοδοτήσουν σε ML μοντέλα, τα οποία είχαν ήδη δοκιμαστεί για τη πρόβλεψη τιμής.
Σε σχέση με όλες τις παραπάνω έρευνες, αλλά και ακόμα περισσότερες οι οποίες αφορούσαν τη πρόβλεψη τιμής για άλλα χρονικά παράθυρα από αυτά που εξετάζουμε εμείς, από όσο γνωρίζουμε, είμαστε οι πρώτοι οι οποίοι χρησιμοποιούν Νευρωνικά Δίκτυα Γράφων σε γράφους συναλλαγών του Blockchain για τη πρόβλεψη της τιμής του Bitcoin.

1.4: Οργάνωση Εργασίας

Στο Κεφάλαιο 2 αναφέρεται το βασικό θεωρητικό υπόβαθρο για τη τεχνολογία του Blockchain και ειδικά του Bitcoin.
Στο Κεφάλαιο 3 αναλύονται οι βασικές έννοιες του τομέα της μηχανικής μάθησης και γίνεται μια εις βάθος παρουσίαση και περιγραφή των βασικών μαθηματικών εννοιών και λειτουργιών της, εστιάζοντας κυρίως στην επιβλεπόμενη μάθηση.
Στο Κεφάλαιο 4 παρουσιάζονται οι απολύτως απαραίτητες έννοιες σχετικά με τους γράφους, οι οποίες χρειάζονται για την κατανόηση εννοιών που χρησιμοποιούνται μετέπειτα.
Στο Κεφάλαιο 5 γίνεται εκτενής παρουσίαση και επεξήγηση εννοιών που αφορούν τα Νευρωνικά Δίκτυα Γράφων και μελετώνται βασικές αρχιτεκτονικές τους.

Στο Κεφάλαιο 6 αναφέρονται οι βασικές γραφοκεντρικές θεωρήσεις των γράφων συναλλαγών του Blockchain με βάση τη βιβλιογραφία.
Στο Κεφάλαιο 7 παρουσίαζεται όλη η διαδικασία που ακολουθήθηκε για να δημιουργηθούν τα διάφορα σενάρια γράφων συναλλαγών που αποτελούν τις εισόδους των GNNs που μελετήθηκαν, από την επιλογή και δημιουργία χαρακτηριστικών και τη κατάλληλη προεπεξεργασία τους μέχρι τον αλγόριθμο κατασκευής γράφων συναλλαγών που αναπτύξαμε. Επιπλέον παρουσίαστηκαν και οι υπερπαράμετροι που διερευνήθηκαν στο πειραματικό μέρος.
Στο Κεφάλαιο 8 και 9 παρατίθονται τα αποτελέσματα των πειραμάτων για τα σενάρια των γράφων συναλλαγών και τα διάφορα GNNs και γίνεται εκτενής ερμηνεία των αποτελεσμάτων. Τέλος, επιλέγεται το καλύτερο μοντέλο και εξετάζεται η απόδοση του στο σύνολο ελέγχου.
Στο Κεφάλαιο 10 γίνεται μια σύνοψη της διαδικασίας που ακολουθήθηκε, παρουσίαζονται τα συμπεράσματα που προέκυψαν από την έρευνά μας και τονίζεται η συνεισφορά της παρούσας εργασίας στην βιβλιογραφία. Επιπλέον προτείνονται πολλές πιθανές επεκτάσεις της δουλειάς μας για μελλοντική έρευνα.

ΘΕΩΡΗΤΙΚΟ ΜΕΡΟΣ

Κεφάλαιο 2: Blockchain και Bitcoin

Το Bitcoin και η τεχνολογία που το υποστηρίζει, το Blockchain, έχουν γίνει αρκετά δημοφιλή τα τελευταία χρόνια. Έχοντας σχεδιαστεί ως μια ασφαλή κατανεμημένη πλατφόρμα χωρίς κεντρικές αρχές, το Blockchain αποτελεί μια από τις πολλά υποσχόμενες τεχνολογίες και μπορεί να θεωρηθεί εξίσου σημαντική με το Cloud Computing, τη μηχανική μάθηση και τα Big Data [46]. Ενσωματώνει καινοτόμες ιδέες από διάφορους τομείς, όπως η κρυπτογράφηση δημόσιου κλειδιού και τα κατανεμημένα συστήματα. Ουσιαστικά, το Blockchain είναι μια κατανεμημένη βάση δεδομένων με τη φύση ενός δημόσιου λογιστικού βιβλίου, που αποτελεί ένα ηλεκτρονικό σύστημα πληρωμών το οποίο είναι ασφαλές εκ κατασκευής. Προτάθηκε από τον άγνωστο συγγραφέα Satoshi Nakamoto το 2008 [17]. Η δομή του είναι μια αλυσίδα από blocks, ψηφιακές δομές οι οποίες περιέχουν τα στοιχεία των συναλλαγών μεταξύ των χρηστών του δικτύου. Τα blocks μπορούν να επαληθευτούν και να επιβεβαιωθούν χωρίς κεντρική αρχή. Οι συναλλαγές των blocks έχουν ως ψηφιακό νόμισμα το Bitcoin.

2.1: Διευθύνσεις πορτοφολιών

Οι διευθύνσεις των ψηφιακών πορτοφολιών των χρηστών είναι αλφαριθμητικές συμβολοσειρές 26-35 χαρακτήρων. Σε κάθε διεύθυνση αντιστοιχεί ένα μοναδικό ιδιωτικό κλειδί, το οποίο το γνωρίζει μόνο ο κάτοχος της διεύθυνσης αυτής και μπορεί μέσω αυτού να πραγματοποιεί συναλλαγές και να ξοδεύει τα Bitcoins που έχει στο πορτοφόλι του.

2.2: Συναλλαγές

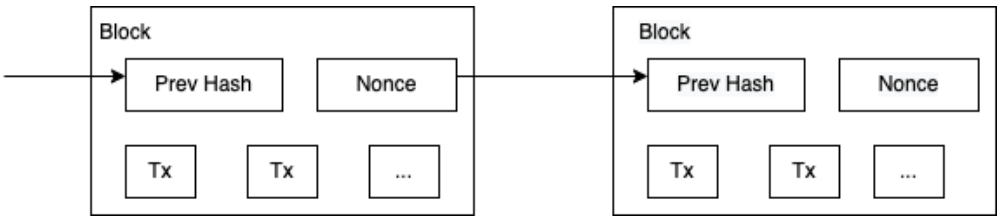
Μια συναλλαγή είναι μια μεταφορά περιουσιακών στοιχείων μεταξύ διευθύνσεων. Το Bitcoin επιτρέπει τη μεταφορά κεφαλαίων από πολλαπλές διευθύνσεις σε πολλαπλές διευθύνσεις. Τις περισσότερες φορές, όλες οι διευθύνσεις εισόδου ανήκουν στον ίδιο χρήστη, αλλά οι διευθύνσεις εξόδου μπορεί να ανήκουν σε διαφορετικούς χρήστες. Για κάθε είσοδο απαιτούνται τριών ειδών δεδομένα. Το μοναδικό αναγνωριστικό της προηγούμενης συναλλαγής που έφερε τα bitcoins της εισόδου, ο αριθμός δείκτη της εξόδου της προηγούμενης συναλλαγής και το μεταφερόμενο ποσό. Όταν μια συναλλαγή έχει περισσότερες από μία εισόδους, κάθε είσοδος υπογράφεται από το σχετικό ιδιωτικό κλειδί ξεχωριστά. Έτσι αποτρέπεται η αλλοίωση της συναλλαγής και αποδεικνύεται ότι ο χρήστης έχει εξουσιοδοτήσει τη μεταφορά των χρημάτων. Όταν μια συναλλαγή αποστέλλει χρήματα (bitcoins) σε μια διεύθυνση, το δημόσιο κλειδί της διεύθυνσης του παραλήπτη είναι άγνωστο για τους υπόλοιπους κόμβους (χρήστες) του δικτύου. Μόνο όταν τα

ληφθέντα χρήματα ξοδευτούν στο μέλλον, αποκαλείπτεται το δημόσιο κλειδί του παραλήπτη και οποιοσδήποτε χρήστης στο δίκτυο μπορεί να ελέγξει ότι τα χρήματα αυτά ανήκουν πράγματι στον παραλήπτη [46].
Στις συναλλαγές, οι αποστολείς δεν αναφέρουν ρητά το τέλος (fee) συναλλαγής. Η διαφορά των συνολικών bitcoin των εισόδων με τα συνολικά bitcoin των εξόδων θεωρείται ως το τέλος αυτό, και το οποίο αποστέλλεται στη διεύθυνση του miner που θα επιβεβαιώσει τη συναλλαγή. Οι συναλλαγές του Bitcoin μπορούν να είναι και χωρίς τέλη, αλλά χωρίς αυτά οι συναλλαγές έχουν μειωμένη πιθανότητα να συμπεριληφθούν στην αλυσίδα του Blockchain.

2.3: Επαλήθευση και επιβεβαίωση συναλλαγών

Ένα ψηφιακό νόμισμα, όπως το Bitcoin, έχει να αντιμετωπίσει δύο βασικά προβλήματα: την επαλήθευση και την επιβεβαίωση της πληρωμής. Η επαλήθευση πληρωμής σημαίνει τη δημιουργία ενός μηχανισμού στον οποίο θα ελέγχεται ότι αυτός που πραγματοποιεί την συναλλαγή έχει το απαραίτητο υπόλοιπο για τη πληρωμή, και ότι θέλει να πληρώσει το υποδεικνυόμενο ποσό. Ένας χρήστης παρουσιάζει το δημόσιο κλειδί του για να δείξει ότι η διεύθυνση του ανήκει και υπογράφει μια υπογραφή με το ιδιωτικό του κλειδί για να επιβεβαιώσει το ποσό. Στις πληροφορίες της συναλλαγής υπάρχει και ένα πεδίο το οποίο φέρει τα μοναδικά αναγνωριστικά των προηγούμενων συναλλαγών που λειτούργησαν ως είσοδοι για τη παρούσα, με τα αντίστοιχα ποσά. Μέσω αυτών των πληροφοριών επαληθεύεται ότι ο χρήστης έχει το υπόλοιπο για τη πληρωμή.
Υπάρχει το ενδεχόμενο ο αποστολέας μια συναλλαγής να προσπαθήσει να χρησιμοποιήσει τα ίδια bitcoin για να πραγματοποιήσει πολλαπλές πληρωμές. Το ζήτημα αυτό ονομάζεται Double Spending Problem. Στην πραγματικότητα, υπάρχει η πιθανότητα η γνωστοποίηση μιας συναλλαγής να μην φτάσει ποτέ σε ορισμένους χρήστες. Επιπλέον, οι χρήστες θα μπορούσαν να είναι κακόβουλοι και να λένε ψέματα σχετικά με τα υπόλοιπα των συναλλαγών. Λόγω αυτών των ενδεχομένων, οι χρήστες ενδέχεται να μην καταλήξουν ποτέ σε συναίνεση σχετικά με το ποιος κατέχει πόσα bitcoins και οποιαδήποτε πληρωμή θα αποτελούσε κίνδυνο απάτης. Ο Nakamoto πρότεινε μια καινοτόμα λύση για το Double Spending Problem. Ως πρώτη θα θεωρηθεί η συναλλαγή η οποία κατέφθασε πρώτη στο blockchain, αλλά καθώς το πότε φτάνει η πληροφορία αυτή σε κάθε κόμβο μπορεί να διαφέρει, τοποθετείται χρονοσφραγίδα σε κάθε συναλλαγή που εισέρχεται στα blocks. Τα block του κατανεμημένου συστήματος έχουν μοναδική χρονοσφραγίδα και αναγνωριστικό (ID) και περιέχουν ένα συγκεκριμένο πλήθος συναλλαγών. Κάθε μπλοκ φέρει το hash του προηγούμενο block, ώστε οι χρήστες να μπορούν να παρακολουθούν πώς αναπτύσσεται η αλυσίδα των blocks με την πάροδο του χρόνου. Ένα block δημιουργείται κάθε δέκα λεπτά μέσω ενός μηχανισμού γνωστού ως Proof-of-Work, ο οποίος απαιτεί την εύρεση ενός αριθμού 32 bit, γνωστού ως nonce, μέσω επαναληπτικών δοκιμών. Η χρήση μεγαλύτερης υπολογιστικής ισχύος αυξάνει την πιθανότητα εύρεσης του αριθμού, αλλά δεν την εγγυάται. Οι χρήστες που εργάζονται για την εύρεση επιβεβαιωμένων μπλοκ ονομάζονται

miners. Το Proof-of-Work περιλαμβάνει τα εξής βήματα: Κάθε miner λαμβάνει μια λίστα συναλλαγών από τους χρήστες, οι οποίες περιμένουν να επιβεβαιωθούν. Από αυτές επιλέγει κάποιες για να τις συμπεριλάβει σε ένα μπλοκ, έχοντας τη δυνατότητα να προτιμά συναλλαγές που πληρώνουν υψηλότερο τέλος.



Εικόνα 1: Αλληλουχία αλυσίδας Blocks

Αρχικά, για τη δημιουργία του block, συγκεντρώνονται οι απαραίτητες πληροφορίες, όπως η χρονοσφραγίδα, το hash του προηγούμενου μπλοκ και μια ειδική τιμή hash των περιεχόμενων συναλλαγών. Μόλις ο miner, μετά το απαραίτητο πλήθος δοκιμών, βρει το σωστό nonce, δημοσιεύει το block και λαμβάνει ως αμοιβή από το σύστημα κάποια bitcoin. Καθώς οι πληροφορίες για το block διαδίδονται στο δίκτυο, οι χρήστες ενημερώνουν τις αλυσίδες τους και ορίζουν το block αυτό ως τελευταίο. Μόλις ενημερωθούν και οι υπόλοιποι miners για αυτό το νέο μπλοκ, σταματούν τους υπολογισμούς που πραγματοποιούσαν μέχρι τότε και αλλάζουν το hash του προηγούμενου μπλοκ με το hash του block που διαδόθηκε στο δίκτυο και συνεχίζουν τις προσπάθειες εξόρυξης για το επόμενο block.

Κεφάλαιο 3: Μηχανική Μάθηση

Το πεδίο της Μηχανικής Μάθησης είναι υποκλάδος της επιστήμης των υπολογιστών,που αφορά την δυνατότητα των υπολογιστών να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά, χωρίς να έχουν προγραμματιστεί ρητά για αυτό [3]. Αυτό επιτυγχάνεται με τη βοήθεια των νευρωνικών δικτύων, τα οποία είναι μηχανές οι οποίες είναι σχεδιασμένες έτσι ώστε να μοντελοποιούν τον τρόπο με τον οποίο λειτουργούν οι νευρώνες και κατ επέκταση ο ίδιος ο εγκέφαλος, ώστε να εκτελέσουν μια συγκεκριμένη εργασία [4].

Στην πραγματικότητα ο εγκέφαλος μπορεί να θεωρηθεί σας ένας πολύπλοκος, μη γραμμικός, παράλληλος υπολογιστής , ο οποίος έχει την ικανότητα να μαθαίνει δικούς του τρόπους λειτουργίας μέσω της εμπειρίας και της ιδιότητας της πλαστικότητας, δηλαδή την ικανότητα του νευρικού συστήματος να αλλάζει και να προσαρμόζεται στην ποικιλομορφία

του περιβάλλοντος. Είτε μέσω ηλεκτρονικών κυκλωμάτων, είτε μέσω προσομοίωσης με λογισμικό, τα νευρωνικά δίκτυα πετυχαίνουν το ίδιο.

Συγκεκριμένα, το νευρωνικό δίκτυο, μέσω μιας διαδικασίας (αλγορίθμου) μάθησης, προσλαμβάνει γνώση από το περιβάλλον του, η οποία αποθηκεύεται μέσω βαρών μεταξύ των συνδέσεων των νευρώνων και έτσι αποκτά την ικανότητα να γενικεύει. Μέσω της γενίκευσης, το νευρωνικό δίκτυο μπορεί να παράγει λογικές εξόδους, για εισόδους τις οποίες δεν έχει συναντήσει καθόλη τη διάρκεια της μάθησής του [4].

3.1: Είδη Διαδικασιών Μάθησης.

Όπως ο εγκέφαλος ενός ανθρώπου μαθαίνει με διαφορετικούς τρόπους, έτσι και τα νευρωνικά δίκτυα έχουν διάφορες διαδικασίες τρόπων μάθησης ανάλογα με το επιθυμητό αποτέλεσμα. Οι βασικότερες κατηγορίες είναι η μάθηση με επίβλεψη (supervised learning), η μάθηση χωρίς επίβλεψη (unsupervised learning), η ημι-επιβλεπόμενη μάθηση (semi-supervised learning) και η ενισχυτική μάθηση (reinforcement learning) [5].

Οι αλγόριθμοι μάθησης με επίβλεψη μαθαίνουν να συσχετίζουν κάποια είσοδο με κάποια έξοδο, δεδομένου ενός συνόλου εκπαίδευσης, το οποίο τροφοδοτείται στο νευρωνικό δίκτυο με παραδείγματα εισόδων x και εξόδων y. Σε πολλές περιπτώσεις οι έξοδοι y μπορεί να είναι δύσκολο να συλλεχθούν αυτόματα και πρέπει να παρέχονται από έναν άνθρωπο "επόπτη", αλλά ο όρος εξακολουθεί να ισχύει ακόμη και όταν οι έξοδοι του συνόλου εκπαίδευσης συλλέγονται αυτόματα.

Οι αλγόριθμοι χωρίς επίβλεψη είναι εκείνοι που τους παρέχονται μόνο χαρακτηριστικά των δεδομένων, αλλά όχι κάποια συγκεκριμένη επιθυμητή έξοδος y. Η διάκριση μεταξύ αλγορίθμων με και χωρίς επίβλεψη δεν είναι αυστηρή, επειδή δεν υπάρχει αντικειμενικός έλεγχος για τη διάκριση μιας τιμής ως χαρακτηριστικό ή επιθυμητή έξοδο που παρέχεται από έναν επόπτη. Η μάθηση χωρίς επίβλεψη αναφέρεται στις περισσότερες προσπάθειες εξαγωγής πληροφοριών από μια κατανομή που δεν απαιτούν ανθρώπινη επέμβαση για τον χαρακτηρισμό των δεδομένων εισόδου. Ο όρος συνδέεται συνήθως με την εκτίμηση πυκνότητας, τη μάθηση για την άντληση δειγμάτων από μια κατανομή, τη μάθηση για την αποθορυβοποίηση δεδομένων από κάποια κατανομή ή την κατηγοριοποίηση των δεδομένων σε ομάδες με βάση αναγνωρισμένα κοινά τους χαρακτηριστικά [6].

Οι αλγόριθμοι ημι-επιβλεπόμενης μάθησης είναι εκείνοι που για ένα μεγάλο πλήθος δεδομένων εισόδων x, τους παρέχονται μερικές επιθυμητές έξοδοι y, για κάποια από αυτά τα δεδομένα. Στην πραγματικότητα τα περισσότερα προβλήματα μηχανικής μάθησης ανήκουν σε αυτή τη κατηγορία, καθώς το να επισημαίνει ένας επόπτης τις επιθυμητές

εξόδους y μπορεί να είναι μια εξαιρετικά χρονοβόρα και πολλές φορές και κοστοβόρα διαδικασία για συγκεκριμένα είδη προβλημάτων.

Η επιβλεπόμενη μάθηση, η μάθηση χωρίς επίβλεψη και η ημι-επιβλεπόμενη μάθηση φαίνεται να καλύπτουν το σύνολο των ειδών της μηχανικής μάθησης, αλλά δεν είναι έτσι. Οι αλγόριθμοι ενισχυτικής μάθησης είναι υπεύθυνοι στο να μαθαίνουν στο νευρωνικό δίκτυο πως να αντιστοιχίζει καταστάσεις σε ενέργειες, με τέτοιο τρόπο έτσι ώστε να μεγιστοποιείται η τιμή μιας συνάρτησης ανταμοιβής. Επιπλέον, ο αλγόριθμος ενισχυτικής μάθησης δεν λαμβάνει οδηγίες για το ποιες ενέργειες πρέπει να κάνει, όπως στην ενισχυτική μάθηση, αλλά πρέπει να ανακαλύψει ο ίδιος ποιες ενέργειες αποφέρουν τη μεγαλύτερη ανταμοιβή δοκιμάζοντάς τες. Αν και μπορεί ίσως να θεωρηθεί ότι η παραπάνω διαδικασία αποτελεί ένα είδος μάθησης χωρίς επίβλεψη, αυτό δεν είναι ορθό, καθώς η ενισχυτική μάθηση προσπαθεί να μεγιστοποιήσει τη τιμή μιας συνάρτησης ανταμοιβής αντί να προσπαθεί να εξάγει πληροφορία μέσω κάποιας κρυφής δομής των δεδομένων, κάτι το οποίο μπορεί να βοηθήσει στο πρόβλημα που αντιμετωπίζει η ενισχυτική μάθηση, αλλά δεν αποτελεί το στόχο της [7].

3.2: Είδη Προβλημάτων Μηχανικής Μάθησης και Μετρικές Απόδοσής τους.

Σε αυτή την υποενότητα θα παρουσιαστούν τα δύο βασικότερα είδη προβλημάτων μηχανικής μάθησης μαζί με μερικές από τις κυριότερες μετρικές αξιολόγησης που χρησιμοποιούνται για αυτά.
Επειδή συνήθως μας ενδιαφέρει πόσο καλά αποδίδει ο αλγόριθμος μηχανικής μάθησης σε δεδομένα που δεν έχει ξαναδεί, καθώς αυτό καθορίζει πόσο καλά θα λειτουργήσει στον πραγματικό κόσμο, οι μετρικές απόδοσης χρησιμοποιούνται τελικώς, σε ένα σύνολο δεδομένων ελέγχου (test set), που είναι ξεχωριστό από τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του νευρωνικού δικτύου.

3.2.1: Ταξινόμηση (Classification)

Σε αυτό το είδος προβλήματος, το νευρωνικό δίκτυο καλείται να προσδιορίσει σε ποια από τις k κατηγορίες ανήκει κάποια είσοδος. Για την επίλυση αυτού του προβλήματος, ο αλγόριθμος μάθησης, αφού έχει εκπαιδευτεί στο σύνολο δεδομένων εκπαίδευσης, καλείται να παράγει μια συνάρτηση $f : R^n \rightarrow \{1, \ldots, k\}$. Όταν $y = f(x)$, το μοντέλο αναθέτει μια είσοδο που περιγράφεται από το διάνυσμα x σε μια κατηγορία που προσδιορίζεται από το y. Υπάρχουν και άλλες παραλλαγές της ταξινόμηση, για παράδειγμα, όταν η συνάρτηση f εξάγει μια κατανομή πιθανότητας για την είσοδο πάνω στις κατηγορίες. Ένα κλασικό παράδειγμα ενός προβλήματος ταξινόμησης είναι η αναγνώριση αντικειμένων, όπου η είσοδος είναι μια εικόνα (που συνήθως περιγράφεται ως ένα σύνολο τιμών

φωτεινότητας εικονοστοιχείων) και η έξοδος είναι ο προσδιορισμός του αντικειμένου στην εικόνα [6].

3.2.1.1: Μετρικές Απόδοσης Ταξινόμησης

Accuracy :

Διαισθητικά η ακρίβεια (accuracy) μετράει πόσες φορές προβλέπει σωστά ο ταξινομητής. Λόγω της ευκολίας κατανόησης των αποτελεσμάτων της, αποτελεί και την πιο συχνή μετρική απόδοσης σε μοντέλα ταξινόμησης, αλλά δεν πρέπει να είναι και ο μοναδικός τρόπος αξιολόγησης του μοντέλου, καθώς πρέπει να λαμβάνονται και άλλες παράμετροι υπ' όψιν, τις οποίες θα εξετάσουμε παρακάτω.

Αν το πρόβλημα που εξετάζεται είναι δυαδικής ταξινόμησης (binary classification problem), δηλαδή το μοντέλο μπορεί να ταξινομήσει ένα αποτέλεσμα σε ακριβώς δύο κλάσεις, τότε μπορούμε να θεωρήσουμε ότι ένα αποτέλεσμα είτε θα είναι θετικό (Positive) είτε αρνητικό (Negative) και έτσι μπορούμε να έχουμε τις εξής περιπτώσεις:

• True Positive (TP): Αποτέλεσμα που προβλέφθηκε να ανήκει στην κλάση P και πράγματι ανήκει σε αυτή τη κλάση.

• True Negative (TN): Αποτέλεσμα που προβλέφθηκε να ανήκει στην κλάση N και πράγματι ανήκει σε αυτή τη κλάση.

• False Positive (FP): Αποτέλεσμα που προβλέφθηκε να ανήκει στην κλάση P αλλά δεν ανήκει σε αυτή τη κλάση.

• False Negative (FN): Αποτέλεσμα που προβλέφθηκε να ανήκει στην κλάση N αλλά δεν ανήκει σε αυτή τη κλάση.

Συνοπτικά έχουμε:

	Actual value is positive	Actual value is negative
Predicted value is positive	TP	FP
Predicted value is negative	FN	TN

Πίνακας 1: Πιθανές περιπτώσεις αποτελέσματος σε πρόβλημα δυαδικής ταξινόμησης

Τότε η ακρίβεια υπολογίζεται από τον τύπο:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision:

Το precision εξηγεί πόσα από τα σωστά προβλεπόμενα αποτελέσματα αποδείχθηκαν πράγματι θετικά. Το precision είναι χρήσιμο στις περιπτώσεις όπου τα Ψευδώς Θετικά (FP) αποτελέσματα αποτελούν μεγαλύτερη ανησυχία από τα Ψευδώς Αρνητικά (FN). Το precision υπολογίζεται από τον τύπο:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall:

Το recall εξηγεί πόσα από τα πραγματικά θετικά αποτελέσματα μπορέσαμε να προβλέψουμε σωστά με το μοντέλο μας. Είναι μια χρήσιμη μετρική απόδοσης σε περιπτώσεις όπου το Ψευδώς Αρνητικό (FN) έχει μεγαλύτερη σημασία από το Ψευδώς Θετικό (FP). Το Recall υπολογίζεται από τον τύπο:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1 Score:

Το F1 Score δίνει μια συνδυαστική μετρική απόδοσης λαμβάνοντας υπ' όψιν το Precision και το Recall. Γίνεται μέγιστο όταν το Precision είναι ίσο με το Recall. Στην πραγματικότητα είναι ο αρμονικός μέσος όρος του Precision και του Recall και δίνεται από τον τύπο:

$$F1 = \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Μέσω του αρμονικού μέσου όρου έναντι του απλού αριθμητικού, δίνεται μεγαλύτερη βαρύτητα στην χειρότερη εκ των δύο μετρικών.

Φυσικά, υπάρχουν πολλές αλλες μετρικές απόδοσης για το πρόβλημα της ταξινόμησης, πλην αυτών που αποτελούν τις πιο βασικές, και εστιάζουν κάθε φορά εκεί που έχει περισσότερη σημασία για το εκάστοτε πρόβλημα.

Όλα όσα αναφέρθηκαν παραπάνω, μπορούν να γενικευτούν για προβλήματα ταξινόμησης περισσότερων των δύο κλάσεων, πραγματοποιώντας τους κατάλληλους μετασχηματισμούς στις παραπάνω εξισώσεις.

3.2.2: Παλινδρόμηση (Regression)

Σε αυτό το είδος προβλήματος, το νευρωνικό δίκτυο καλείται να προβλέψει μία αριθμητική τιμή δεδομένης μιας εισόδου. Για την επίλυση αυτού του καθήκοντος, ζητείται από τον αλγόριθμο μάθησης να εξάγει μια συνάρτηση $f : R^n \rightarrow R$. Ένα κλασικό παράδειγμα ενός προβλήματος παλινδρόμησης είναι η πρόβλεψη της τιμής μιας μετοχής μιας εταιρείας στη διάρκεια του χρόνου.

3.2.2.1: Μετρικές Απόδοσης Ταξινόμησης

Μέσο Τετραγωνικό Σφάλμα - Mean Squared Error (MSE):

Εκφράζει το μέσο τετραγωνικό σφάλμα των προβλέψεων $\hat{y}$ του μοντέλου σε σχέση με τις πραγματικές τιμές y. Συγκεκριμένα αν $\hat{y}_i$ είναι η προβλεπόμενη τιμή του i-οστού δείγματος και y_i η πραγματική τιμή που του αντιστοιχεί, τότε το MSE για n δείγματα δίνεται από τον τύπο:

$$MSE(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2 \quad (5)$$

Επειδή χρησιμοποιείται το τετράγωνο των διαφορών των τιμών, η ερμηνεία μπορεί να φανεί λιγότερο διαισθητική. Όμως ο τετραγωνισμός των διαφορών εξυπηρετεί διάφορους σκοπούς. Συγκεκριμένα, εξαλείφει τις αρνητικές τιμές των διαφορών και έτσι εξασφαλίζει ότι και αυτές θα συμπεριληφθούν στον υπολογισμό, με αποτέλεσμα το μέσο τετραγωνικό σφάλμα είναι πάντα μεγαλύτερο ή ίσο με μηδέν. Επιπλέον, ο τετραγωνισμός αυξάνει το αντίκτυπο των μεγαλύτερων σφαλμάτων με αποτέλεσμα το συνολικό σφάλμα να αυξάνεται λόγω αυτών δυσανάλογα σε σχέση με τα μικρότερα σφάλματα. Έτσι, αν υπάρχουν στα δεδομένα μας ακραίες τιμές που δεν θα έπρεπε να υπάρχουν (outliers), και δεν έχει προηγηθεί διαδικασία προεπεξεργασίας για την αφαίρεσή τους, αυτή η μετρική δεν θα πρέπει να επιλέγεται.

Μέσο Τετραγωνικό Ριζικό Σφάλμα - Root Mean Squared Error (RMSE):

Εάν στον παραπάνω τύπο εφαρμόσουμε τετραγωνική ρίζα, προκύπτει το μέσο τετραγωνικό ριζικό σφάλμα (RMSE), το οποίο δίνεται από τον τύπο:

$$RMSE = \sqrt{MSE(y, \hat{y})} = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2} \quad (6)$$

Το RMSE αποτελεί μία από τις βασικότερες μετρικές σε προβλήματα παλινδρόμησης και, λόγω της ρίζας, παρέχει πιο διαισθητικά αποτελέσματα σε σχέση με το MSE.

Μέσο Απόλυτο Ποσοστιαίο Σφάλμα - Mean Absolute Percentage Error (MAPE)

Το μέσο απόλυτο ποσοστιαίο σφάλμα δίνεται από τον τύπο:

$$MAPE = \frac{100\%}{n}\sum_{i=0}^{n-1}\frac{|y_i - \hat{y}_i|}{|y_i|} \quad (7)$$

Το MAPE είναι αρκετά αποτελεσματικό για την ανάλυση μεγάλων συνόλων δεδομένων, αλλά δεν πρέπει να χρησιμοποιείται όταν το συνόλων δεδομένων έχει μηδενικές τιμές, καθώς ο υπολογισμός θα απαιτούσε διαίρεση με το μηδέν, το οποίο δεν ορίζεται μαθηματικά.
Αποτελεί μια αρκετά διαισθητική μετρική απόδοσης, αφού για παράδειγμα ένα 10% MAPE αντιπροσωπεύει ότι η μέση απόκλιση μεταξύ των προβλεπόμενων τιμών και των πραγματικών είναι 10%, ανεξάρτητα από το αν η απόκλιση ήταν θετική ή αρνητική.

3.3: Μηχανική Μάθηση ως Πρόβλημα Βελτιστοποίησης.

Στην παρούσα ενότητα, θα παρουσιαστούν οι βασικές μαθηματικές έννοιες και οι μαθηματικοί κανόνες που διέπουν τα νευρωνικά δίκτυα, θεωρώντας τη μηχανική μάθηση ως ένα πρόβλημα βελτιστοποίησης. Συγκεκριμένα, θα εστιάσουμε στην επιβλεπόμενη μάθηση, καθώς τέτοιου είδους είναι και το πρόβλημα που μελετάται στη παρούσα εργασία.

3.3.1: Μαθηματική Διατύπωση Προβλημάτων Επιβλεπόμενης Μάθησης.

Όπως έχει αναφερθεί, στα προβλήματα επιβλεπόμενης μάθησης δίνονται δεδομένα εισόδου x και ο στόχος είναι να προβλέψουμε τις ετικέτες (labels) y που τους αντιστοιχούν. Τα δεδομένα εισόδου μπορούν να είναι διανύσματα πραγματικών αριθμών, ακολουθίες λέξεων, πίνακες, εικόνες, ακόμα και γράφοι (πλήρης επεξήγηση της έννοιας παρατίθεται στο κεφάλαιο 4) οι οποίοι έχουν στις ακμές και τις κορυφές τους κάποια χαρακτηριστικά (features). Οι ετικέτες μπορούν να είναι πραγματικοί αριθμοί για προβλήματα παλινδρόμησης, κλάσεις για προβλήματα ταξινόμησης κ.λ.π.

Μοντελοποιούμε το έργο αυτό, ως ένα πρόβλημα βελτιστοποίησης στο οποίο έχουμε την εξίσωση:

$$min_{\theta} \boldsymbol{L}\big(y, f(x)\big) \ (8)$$

Μας ενδιαφέρει να παραμετροποιήσουμε την εξίσωση (8), η οποία έχει ως είσοδο τα δεδομένα x, μέσω του συνόλου των παραμέτρων του νευρωνικού Θ, με τέτοιο τρόπο ώστε να ελαχιστοποιείται η συνάρτηση απώλειας "$\mathcal{L}$" μεταξύ της πραγματικής τιμής y και της προβλεπόμενης τιμής που της αντιστοιχεί, f(x).

Οι παράμετροι Θ του μοντέλου μπορούν να περιέχουν διανύσματα, πίνακες κ.ο.κ. Με το σύμβολο Θ, αναφερόμαστε γενικά στις παραμέτρους του κατ' εξέταση μοντέλου.
Η συνάρτηση κόστους ποσοτικοποιεί την απόκλιση μεταξύ της πραγματικής εξόδου και της προβλεπόμενης. Υπάρχουν πολλές συναρτήσεις κόστους που χρησιμοποιούνται. Για παράδειγμα η L2 συνάρτηση κόστους δίνεται από τον τύπο:

$$\boldsymbol{L}\big(y, f(x)\big) = \|y - f(x)\|_2 \ (9)$$

Έτσι αν μελετάμε ένα πρόβλημα παλινδρόμησης, το y αντιπροσωπεύει την πραγματική αριθμητική τιμή της εξόδου, το f(x) την προβλεπόμενη από το μοντέλο τιμή και η συνάρτηση L2 αντιπροσωπεύει την ευκλείδεια νόρμα της διαφοράς αυτών των δύο τιμών. Έτσι, η εξίσωση (8) μεταφράζεται στο ότι ψάχνουμε τις παραμέτρους Θ, έτσι ώστε η ευκλείδεια απόσταση των πραγματικών με των προβλεπόμενων τιμών να ελαχιστοποιείται. Το είδος της συνάρτησης κόστους που θα επιλέξουμε, εξαρτάται από το είδος του προβλήματος που μελετάμε.

3.3.1.1: Ελαχιστοποίηση Συνάρτησης Κόστους Μέσω Αλγορίθμου Gradient Descent.

Βάση όσων αναφέρθηκαν στην προηγούμενη υποενότητα, το πρόβλημα βελτιστοποίησης έχει αναχθεί στην ελαχιστοποίηση της συνάρτησης κόστους, η οποία σχετίζεται με τις έννοιες σχετικά με το Gradient Descent.

Το διανυσμα κλισης (gradient vector) σε ένα δοσμένο σημείο/δεδομένο για μια διανυσματική διαφοροποιήσιμη συνάρτηση πολλών μεταβλητών αντικατοπτρίζει τη κατεύθυνση και το ρυθμό της γρηγορότερης αύξησης της. Θεωρούμε το διάνυσμα κλίσης της συνάρτησης κόστους με τέτοιο τρόπο ώστε να μπορούμε να την αξιολογήσουμε ως προς τις παραμέτρους Θ του μοντελου, οι οποίες αποτελούν και τις μεταβλητές της συνάρτησης κόστους. Συγκεκριμένα το διάνυσμα κλίσης δίνεται από την εξίσωση:

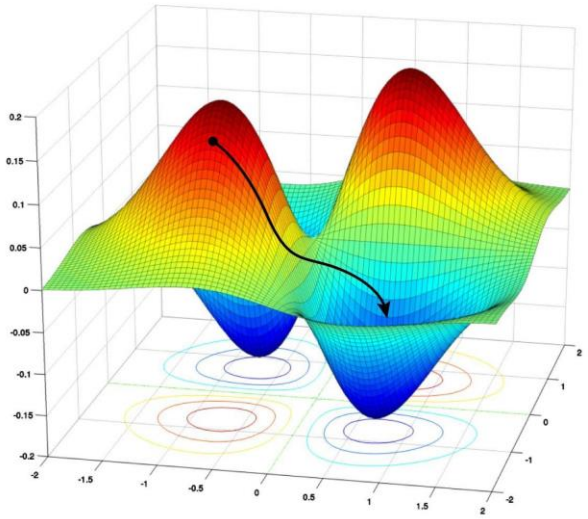
$$\nabla_\theta \boldsymbol{L} = \left(\frac{\partial \boldsymbol{L}}{\partial \theta_1}, \frac{\partial \boldsymbol{L}}{\partial \theta_2}, \ldots.\right) \ (10)$$

Μέσω της εξίσωσης (10) καταλαβαίνουμε ότι για δεδομένες τιμές των παραμέτρων σε ένα σημείο, στο διάνυσμα που αντιπροσωπεύει τον στιγμιαίο ρυθμό μεταβολής της συνάρτησης κόστους, μπορούμε να υπολογίσουμε προς ποια κατεύθυνση θα πρέπει να αλλάξουν οι παράμετροι έτσι ώστε η συνάρτηση κόστους να μειωθεί περισσότερο. Έχουμε δηλαδή μια τρέχουσα εκτίμηση των παραμέτρων μας, και υπολογίζουμε τις μερικές παραγώγους της συνάρτησης απωλειών μας μέχρι το σημείο που βρισκόμαστε, και στη συνέχεια κινούμαστε προς την κατεύθυνση, της ταχύτερης μείωσης της απώλειας και ελπίζουμε να φτάσουμε σε ένα ικανοποιητικό τοπικό ελάχιστο ή στο ολικό ελάχιστο της συνάρτησης. Έτσι, η κλίση που αξιολογείται για το συγκεκριμένο σημείο είναι η κατεύθυνση της παραγώγου που μου δίνει την μεγαλύτερη αύξηση. Αφού ενδιαφερόμαστε για την ταχύτερη μείωση, θα κινηθούμε προς την αντίθετη κατεύθυνση της κλίσης που μελετάμε, μέσω ενός αλγορίθμου που ονομάζεται Gradient Descent.
Συγκεκριμένα, σε κάθε βήμα του βρόχου επανάληψης της εκπαίδευσης του μοντέλου, ανανεώνουμε τα βάρη (παραμετρους Θ) προς την αντίθετη κατεύθυνση της κλίσης μέχρι να έχουμε σύγκλιση της συνάρτησης. Το βήμα δίνεται από τον παρακάτω κανόνα:

$$\Theta \leftarrow \Theta - \eta \nabla_\theta \boldsymbol{L} \ (11)$$

όπου **η**, ο ρυθμός μάθησης (learning rate) του μοντέλου, ο οποίος αποτελεί μια από τις υπερπαραμέτρους του και ελέγχει το μέγεθος του βήματος του gradient descent. Ιδανικά θέλουμε ο αλγόριθμος να τερματίζει όταν έχουμε μηδενικό gradient, που σημαίνει ότι ο αλγόριθμος έχει βρει κάποιο τοπικό ελάχιστο της συνάρτησης κόστους. Στη πράξη σταματάμε τον αλγόριθμο μάθησης όταν η απόδοση του μοντέλου παύει να βελτιώνεται στο σύνολο επικύρωσης (validation set), το οποίο αποτελεί ένα ξεχωριστό σύνολο

δεδομένων από το σύνολο εκπαίδευσης και στο οποίο εξετάζουμε την επίδραση των υπερπαραμέτρων του κατ εξέταση μοντέλου στις προβλέψεις μας, χωρίς να τις υπολογίζουμε βάση αυτού.



Εικόνα 2: Γραφική αναπαράσταση εύρεσης τοπικού ελαχίστου συνάρτησης μέσω αλγορίθμου gradient descent

3.3.1.2: Stochastic Gradient Descent

Το μειονέκτημα του Gradient Descent είναι ότι χρειάζεται να υπολογίσουμε τον όρο $\nabla_\theta \boldsymbol{L}(y, f(x))$, για όλο το σύνολο δεδομένων εκπαίδευσης, το οποίο πολλές φορές μπορεί να είναι τεράστιο σε όγκο, με αποτέλεσμα ο υπολογισμός αυτός να είναι πολύ ακριβός υπολογιστικά και χρονοβόρος. Έτσι χρειάζεται να εισάγουμε την έννοια του Stochastic Gradient Descent**,** η οποία αντί να υπολογίζει το κόστος (loss) για όλα τα δεδομένα εκπαίδευσης, θεωρεί τα minibatches B τα οποία περιέχουν ένα υποσύνολο των δεδομένων και σε κάθε βήμα του αλγορίθμου GD διαλέγει και διαφορετικό minibatch των δεδομένων και υπολογίζει σε αυτό το κόστος. Παραθέτουμε μερικούς βασικούς ορισμούς για έννοιες τις οποίες θα συναντήσουμε εκτενέστερα στο υπόλοιπο της παρούσας εργασίας.

Μέγεθος Batch (Batch size): Το πλήθος των δεδομένων σε ένα minibatch.

Επανάληψη (Iteration): Ένα βήμα του SGD στο minibatch.

Εποχή (Epoch): Ένα πλήρες πέρασμα σε όλο το σύνολο δεδομένων (Δηλαδή όταν το πλήθος των επαναλήψεων είναι ίσο με το λόγο του μεγέθους του συνόλου εκπαίδευσης και του μεγέθους του batch).

Στην πράξη, όταν χρησιμοποιούμε το SGD δεν υπάρχει εγγύηση για το ρυθμό της σύγκλισης και χρειάζεται να πειραματιστούμε με διάφορες τιμές του ρυθμού μάθησης. Υπάρχουν αρκετοί υλοποιημένοι βελτιστοποιητές για τον SGD όπως ο Adam, Adagrad, Adadelta, RMSprop κ.λ.π.

3.3.1.3: Απλό Παράδειγμα Συνάρτησης Νευρωνικού Δικτύου.

Όπως έχουμε αναφέρει στην μηχανική μάθηση η συνάρτηση f(x) που χρησιμοποιούμε για τις προβλέψεις μας μπορεί να είναι πολύ σύνθετη, όπως για παράδειγμα ένα πολυεπίπεδο νευρωνικό δίκτυο βαθιάς μάθησης [47]. Στην παρούσα ενότητα θα παρουσιάσουμε τους υπολογισμούς πίσω από ένα απλό παράδειγμα συνάρτησης νευρωνικού δικτύου και συγκεκριμένα μιας γραμμικής συνάρτησης που δίνεται από τον τύπο:

$$f(x) = W * x, \Theta = \{W\} \ (12)$$

Όπου x, η είσοδος του νευρωνικού δικτύου (τα στοιχεία του συνόλου των δεδομένων εκπαίδευσης), και W το αντικείμενο που περιγράφει τις παραμέτρους Θ.

Όπως έχουμε ήδη αναφέρει ο στόχος μας είναι να προσδιορίσουμε τις παραμέτρους Θ έτσι ώστε να ικανοποιήσουμε την εξίσωση (8)

Αν η f επιστρέφει αριθμό και το W είναι ένα διάνυσμα μάθησης (learnable vector) τότε η κλίση της f συναρτήση του διανύσματος μάθησης δίνεται από τον τύπο:

$$\nabla_W \boldsymbol{f} = \left(\frac{\partial \boldsymbol{f}}{\partial w_1}, \frac{\partial \boldsymbol{f}}{\partial w_2}, \frac{\partial \boldsymbol{f}}{\partial w_3}, \ldots.\right) \ (13)$$

όπου στην ουσία παίρνουμε τις μερικές παραγώγους της f για να υπολογίσουμε τη κλίση της, συναρτήσει των στοιχείων του διανύσματος μάθησης.

Ενώ αν η f επιστρέφει διάνυσμα, τότε W είναι ο ιακωβιανός πίνακας βαρών της f:

$$\nabla_W \boldsymbol{f} = \begin{bmatrix} \frac{\partial \boldsymbol{f}}{\partial w_{11}} & \frac{\partial \boldsymbol{f}}{\partial w_{12}} \\ \frac{\partial \boldsymbol{f}}{\partial w_{21}} & \frac{\partial \boldsymbol{f}}{\partial w_{22}} \end{bmatrix} \ (14)$$

3.3.1.3.1: Οπίσθια τροφοδότηση (Back Propagation)

Αν θεωρήσουμε πιο σύνθετες συναρτήσεις, όπως για παράδειγμα την:

$$f(x) = W_2(W_1 x), \theta = \{W_1, W_2\} \ (15)$$

όπου W_1, W_2 οι πίνακες βαρών που είναι το σύνολο των παραμέτρων Θ του μοντέλου.
Τότε, για να υπολογίσουμε τη κλίση μέσω των παραγώγων, χρειάζεται να χρησιμοποιήσουμε τον κανόνα αλυσίδας που δίνεται από τον τύπο:

$$\frac{\partial \boldsymbol{f}}{\partial x} = \frac{\partial \boldsymbol{g}}{\partial h} \cdot \frac{\partial \boldsymbol{h}}{\partial x} \ (16)$$

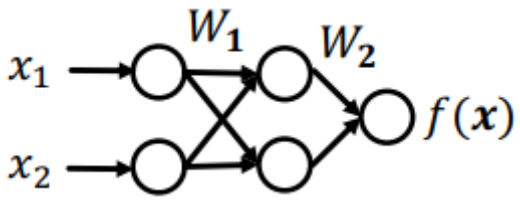
Όπου $f(x) = W_2(W_1 x)$, $h(x) = W_1 x$, $g(z) = W_2 z$

Η έννοια του back propagation, χρησιμοποιεί τον κανόνα της αλυσίδας για να διαδώσει τις κλίσεις των ενδιάμεσων βημάτων και τελικά να υπολογίσει τη κλίση της συνάρτησης κόστους συναρτήσει των παραμέτρων Θ.

Για το παράδειγμα που αναφέραμε, όπως δείξαμε:

$$f(x) = g\big(h(x)\big) = W_2(W_1 x)$$

μπορούμε να το αναπαραστήσουμε με το κάτωθι σχήμα:



Εικόνα 3 : Ένα απλό διεπίπεδο γραμμικό νευρωνικό δίκτυο.

Η εικόνα 3 δείχνει ότι έχουμε δισδιάστατη είσοδο x, η οποία πολλαπλασιάζεται με τον πίνακα βαρών W_1 και στην συνέχεια με τον πίνακα βαρών W_2 για να λάβουμε τελικά την έξοδό μας f(x).
Θεωρούμε τη συνάρτηση κόστους ως:

$$\boldsymbol{L} = \sum_{(\boldsymbol{x},\boldsymbol{y})\in \boldsymbol{B}} \|y, -f(x)\|_2 \ (17)$$

δηλαδή το άθροισμα των L2 απωλειών σε ένα Minibatch B.
Τέλος, θεωρούμε την ενδιάμεση αναπαράσταση των δεδομένων εισόδων x, $h(x) = W_1 x$ ως κρυφό επίπεδο και τότε έχουμε:

$$f(x) = W_2 h(x)$$

Για τον υπολογισμό του κόστους ξεκινώντας από την είσοδο έχουμε την πρόσθια τροφοδότηση (forward propagation):

$$x_{\times W_1} \rightarrow h_{\times W_2} \rightarrow g_{\times W_3} \rightarrow \boldsymbol{L}$$

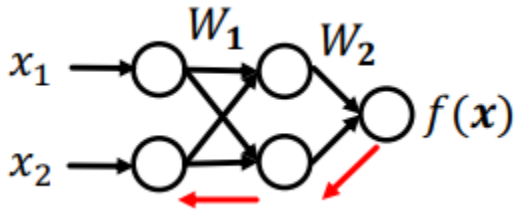
Για να εκτελέσουμε την οπίσθια τροφοδότηση έτσι ώστε να υπολογίσουμε τη κλίση του Θ = {W_1, W_2} και να βρούμε τις βέλτιστες τιμές των παραμέτρων Θ, χρησιμοποιούμε τον κανόνα της αλυσίδας:

$$\frac{\partial \boldsymbol{L}}{\partial W_2} = \frac{\partial \boldsymbol{L}}{\partial f} \cdot \frac{\partial \boldsymbol{f}}{\partial W_2} \quad (18)\,,$$

και

$$\frac{\partial \boldsymbol{L}}{\partial W_1} = \frac{\partial \boldsymbol{L}}{\partial f} \cdot \frac{\partial \boldsymbol{f}}{\partial W_2} \cdot \frac{\partial W_2}{\partial W_1} \quad (19)$$

όπου για τον δεύτερο όρο έχουμε έτοιμο τον υπολογισμό από την (18). Όπως φαίνεται από τις (18), (19), ο λόγος που καλείται back propagation οφείλεται στο γεγονός ότι ξεκινάμε τους υπολογισμούς από την έξοδο και δουλεύουμε προς τα πίσω, δηλαδή προς την είσοδο, το οποίο σχηματικά φαίνεται στην παρακάτω εικόνα:

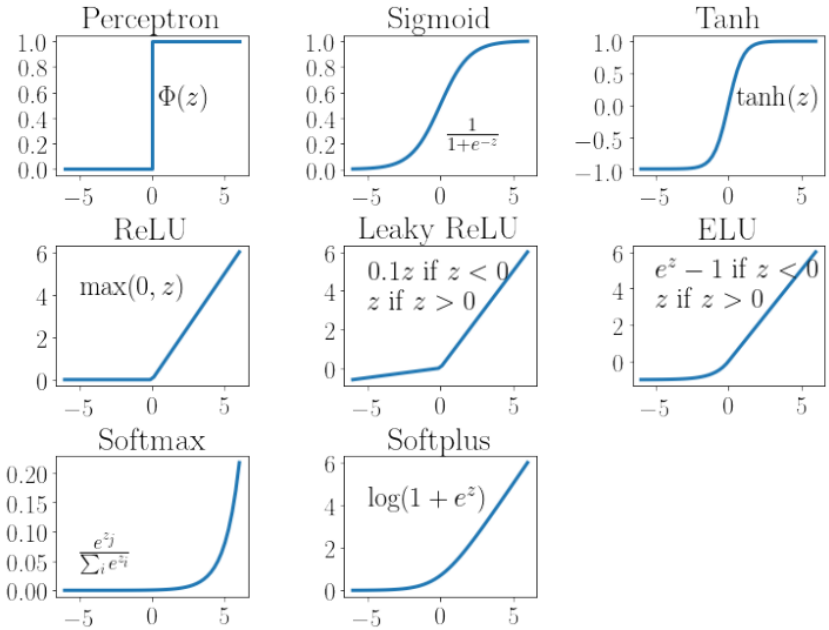


Εικόνα 4: Σχηματική αναπαράσταση του back propagation ενός διεπίπεδου γραμμικού νευρωνικού δικτύου.

3.3.1.3.2: Συναρτήσεις Ενεργοποίησης και Μη Γραμμικότητα

Στην εξίσωση (15) παρατηρούμε ότι το γινόμενο W_2W_1 είναι πάλι πίνακας και άρα όσους πίνακες βαρών και αν προσθέσουμε αλυσιδωτά, η f(x) παραμένει γραμμική συνάρτηση του x και έτσι δεν αυξάνεται η εκφραστική δύναμη στην f(x). Αν στο μοντέλο όμως εισάγουμε κάποια συνάρτηση ενεργοποίησης, η οποία καθορίζει αν ένας νευρώνας πρέπει να ενεργοποιηθεί ή όχι, τότε το μοντέλο γίνεται μη γραμμικό με αποτέλεσμα να διευκολύνεται

στο να γενικεύει και να προσαρμόζεται σε ποικίλα δεδομένα, διαφοροποιώντας την εκάστοτε έξοδο. Οι βασικές συναρτήσεις ενεργοποίησης που χρησιμοποιούνται ευρέως σε προβλήματα νευρωνικών δικτύων φαίνονται στην εικόνα 5.



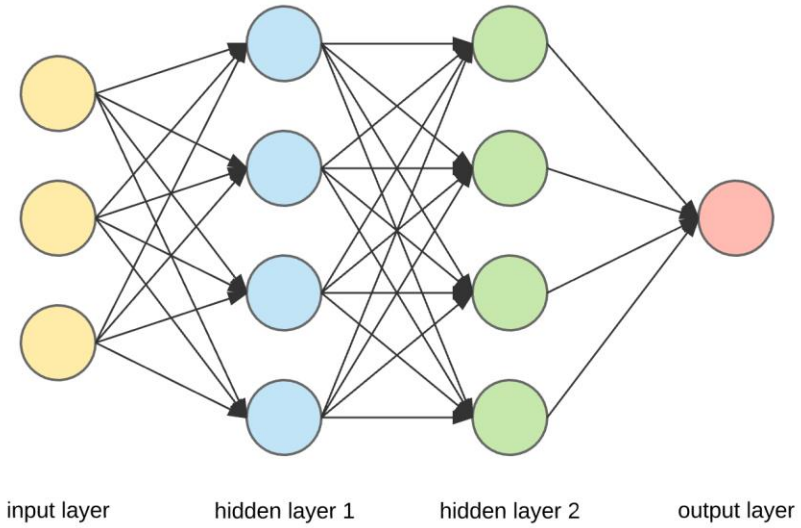
Εικόνα 5: Βασικές συναρτήσεις ενεργοποίησης. Πηγή από [48]

3.3.1.3.3: Multi-layer Perceptron (MLP)

Το MLP πρόκειται για το πιο κλασικό νευρωνικό δίκτυο, το οποίο συνδυάζει όλες από τις έννοιες που αναφέρθηκαν στις προηγούμενες ενότητες. Συγκεκριμένα αποτελείται από τουλάχιστον τρία επίπεδα νευρώνων, το επίπεδο εισόδου, το κρυφό επίπεδο και το επίπεδο εξόδου. Σε κάθε επίπεδο η είσοδος των νευρώνων δίνεται από τον τύπο:

$$x^{(l+1)} = \sigma\left(W_l x^{(l)} + b^l\right) (20)$$

όπου το W_l είναι ο πίνακας βαρών που μετασχηματίζει τη κρυφή αναπαράσταση από το layer l στο layer l+1, b^l το bias στο επίπεδο l, το οποίο προστίθεται στον γραμμικό μετασχηματισμό του x και σ μια οποιαδήποτε συνάρτηση ενεργοποίησης μέσω της οποίας το επόμενο επίπεδο παίρνει τη μη γραμμικά μετασχηματισμένη έξοδο του προηγούμενου.



Εικόνα 6: Σχηματική αναπαράσταση ενός MLP με τρισδιάστατη είσοδο, δύο κρυφών επιπέδων και μονοδιάστατη έξοδο.Κάθε επίπεδο έχει έναν γραμμικό και μη γραμμικό μετασχηματισμό όπως δίνεται από την εξίσωση (20)

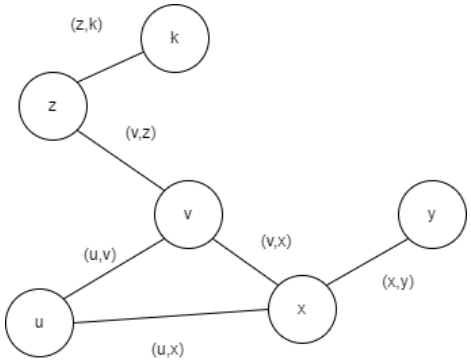
Κεφάλαιο 4: Γράφοι

4.1: Βασικοί Ορισμοί και Έννοιες Γράφων

Οι γράφοι (ή αλλιώς γραφήματα) είναι μαθηματικά αντικείμενα τα οποία αποτελούν την πιο γενική και αφηρημένη μορφή αναπαράστασης δομής δεδομένων. Συγκεκριμένα:

"Ένα γράφημα G είναι ένα ζεύγος συνόλων (V, E) , όπου V είναι ένα πεπερασμένο μη-κενό σύνολο n στοιχείων και E ένα πεπερασμένο σύνολο μη-διατεταγμένων ζευγών με στοιχεία του συνόλου V." [1]

Χρησιμοποιούμε τον συμβολισμό G = (V,E) όπου τα στοιχεία του συνόλου V ονομάζονται κόμβοι (vertices) και τα στοιχεία του συνόλου E ονομάζονται ακμές (edges). Μια ακμή είναι ένα ζεύγος (u,v) από κόμβους το οποίο αντιπροσωπεύει τη σχέση μεταξύ των δύο αυτών κόμβων.



Εικόνα 7: Παράδειγμα Γράφου

Οι γράφοι, με άλλα λόγια, αποτελούν μια γενική γλώσσα για την περιγραφή και την ανάλυση οντοτήτων και των μεταξύ τους σχέσεων και αλληλοεπιδράσεων. Αντί ο κόσμος να θεωρείται ως ένα σύνολο απομονωμένων σημειακών δεδομένων, ερμηνεύεται από την άποψη των δικτύων και των σύνθετων σχέσεων μεταξύ αυτών των οντοτήτων στον υποκείμενο γράφο. Υπάρχουν πολλά πραγματικά προβλήματα που μπορούν να αναπαρασταθούν ως γράφοι και η μοντελοποίηση των δομικών σχέσεων του εκάστοτε υπό μελέτη προβλήματος, μας επιτρέπει να κατασκευάσουμε ακριβή μοντέλα για την ανάλυσή τους.

Παρακάτω παρατίθενται μερικοί βασικοί ορισμοί γράφων καθώς και η αντίστοιχη ορολογία, που θα χρησιμοποιηθεί εκτενώς στα επόμενα κεφάλαια της παρούσας εργασίας:

Ένας γράφος ονομάζεται κατευθυνόμενος (directed graph, digraph) αν κάθε μια από τις ακμές του είναι προσανατολισμένη προς μία κατεύθυνση. Το πλήθος των ακμών που εκκινούν από ένα κόμβο λέγεται βαθμός εξόδου του κόμβου. Αντίστοιχα το πλήθος των βελών που καταλήγουν στον κόμβο λέγεται βαθμός εισόδου του κόμβου.

Ένας γράφος ονομάζεται μη-κατευθυνόμενος (undirected) αν οι ακμές του δεν είναι προσανατολισμένες.

Αν (u,v) είναι ακμή τότε λέμε ότι οι κορυφές u και v είναι γειτονικές (adjacent) ή ότι γειτνιάζουν.

Μονοπάτι ή διαδρομή (path) ενός γράφου μήκους n, είναι μια ακολουθία κόμβων v_0, v_1, …, v_n, όπου για κάθε i , $0 \leq i < n$, (v_i , v_{i+1}) είναι ακμή του γράφου. Μήκος ενός μονοπατιού είναι ο αριθμός των ακμών που περιέχει.

Κύκλος (cycle) ονομάζεται μια διαδρομή με μήκος >1 που ικανοποιεί $v_0 = v_n$.

Ένας γράφος που δεν περιέχει κύκλους ονομάζεται άκυκλος (acyclic).

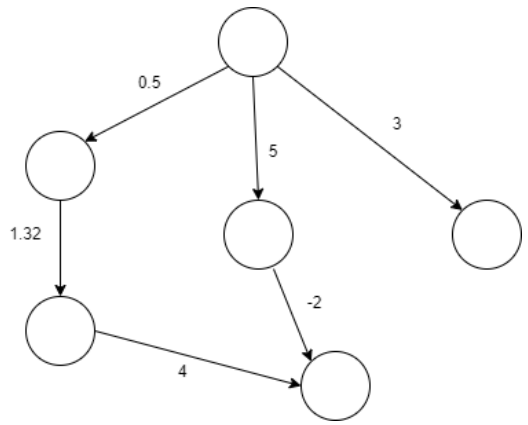
Έστω G=(V,E) και G΄= (V΄, E΄) γράφοι, όπου V΄⊆V και E΄⊆ E. Τότε ο γράφος G΄ είναι υπογράφος (subgraph) του γράφου G.

Η απόσταση δύο κορυφών είναι το μήκος της συντομότερης διαδρομής που οδηγεί από τη μια κορυφή στην άλλη.

Ένας μη κατευθυνόμενος γράφος λέγεται συνεκτικός (connected) αν για κάθε ζευγάρι κορυφών υπάρχει διαδρομή που τις συνδέει.

Ένας κατευθυνόμενος γράφος που ικανοποιεί την ίδια ιδιότητα ονομάζεται ισχυρά συνεκτικός (strongly connected). Αν ο μη κατευθυνόμενος γράφος στον οποίο αντιστοιχεί είναι συνεκτικός, τότε ο γράφος ονομάζεται ελαφρά συνεκτικός (weakly connected).

Συχνά συσχετίζουμε κάθε ακμή ενός γράφου με κάποιο βάρος (weight). Τότε ο γράφος ονομάζεται γράφος με βάρη (weighted graph).



Εικόνα 8: Παράδειγμα Κατευθυνόμενου Ακυκλικού Γράφου με Βάρη

4.2: Αναπαράσταση Γράφων

Οι γράφοι μπορούν να αναπαρασταθούν με πολλούς τρόπους και ανάλογα το εκάστοτε πρόβλημα χρησιμοποιείται ο πιο αποδοτικός. Παρακάτω παρατίθενται οι δύο πιο συνήθεις τρόποι αναπαράστασης γράφων, εκ των οποίων ο δεύτερος είναι και αυτός που χρησιμοποιείται στα νευρωνικά δίκτυα γράφων, τα οποία θα εξετάσουμε σε επόμενο κεφάλαιο.

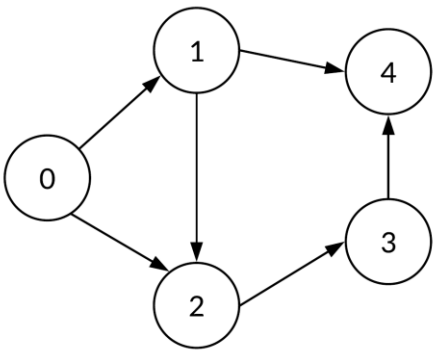
4.2.1: Πίνακες Γειτνίασης

Κάθε γράφος G = (G,E) όπου |V| = n (n πλήθος κόμβων) μπορεί να αναπαρασταθεί με έναν πίνακα γειτνίασης A διαστάσεων nxn. Έστω οι κορυφές v_1, v_2,..., v_n του γράφου G τότε:

A(G) = $[a_{ij}]$, όπου a_{ij} = 1, αν $\{v_i,v_j\}$ είναι ακμή του G
a_{ij} = 0, διαφορετικά

Αν ο γράφος G έχει βάρη τότε τα a_{ij} αντικαθίστανται από τα αντίστοιχα βάρη. Σε μη κατευθυνόμενο γράφο ισχύει: $a_{ij}=a_{ji}$.
Η απαιτούμενη μνήμη για την αποθήκευση του πίνακα A είναι $\Theta(n^2)$. Ως αποτέλεσμα όταν ο αριθμός των κόμβων σε έναν γράφο είναι μεγάλος και ο αριθμός των ακμών ανά κόμβο μεταβλητός, ο αντίστοιχος πίνακας γειτνίασης είναι αραιός και έτσι μη αποδοτικός ως προς το χώρο.

Ένα άλλο πρόβλημα με την αναπαράσταση αυτή είναι ότι μπορούν να υπάρχουν πολλοί πίνακες γειτνίασης που μπορούν να κωδικοποιήσουν τον ίδιο γράφο και ως εκ τούτου δεν υπάρχει καμία εγγύηση ότι αυτοί οι διαφορετικοί πίνακες θα παρήγαγαν το ίδιο αποτέλεσμα σε ένα νευρωνικό δίκτυο, όπως το MLP που παρουσιάστηκε στο 3.3.1.3.3, δηλαδή δεν είναι αμετάβλητοι ως προς τις μεταθέσεις (permutation invariant). Μάλιστα μπορεί να υπάρχουν μέχρι και n! διαφορετικοί πίνακες γειτνίασης για ένα γράφο με n κόμβους, αφού υπάρχουν n! διατάξεις για n κόμβους [2].



Adjacency Matrix

	0	1	2	3	4
0	0	1	1	0	0
1	0	0	1	0	1
2	0	0	0	1	0
3	0	0	0	0	1
4	0	0	0	0	0

Εικόνα 9: Κατευθυνόμενος γράφος με τον αντίστοιχο πίνακα γειτνίασης του.

4.2.2: Λίστες Γειτνίασης

Ένας κομψός και αποδοτικός για τη μνήμη τρόπος αναπαράστασης γράφων οι οποίοι είναι αραιοί (μικρό πλήθος ακμών), είναι οι λίστες γειτνίασης. Αυτές περιγράφουν τη συνδεσιμότητα μιας ακμής e_k μεταξύ των κόμβων n_i και n_j ως μια πλειάδα (i,j) στην k-οστή εγγραφή μιας λίστας γειτνίασης.

Το μειονέκτημα αυτής της αναπαράστασης έγκειται στο γεγονός ότι για να διαπιστώσουμε για παράδειγμα αν μια ακμή είναι παρούσα στο γράφημα θα έπρεπε να αναζητήσουμε τη λίστα των κόμβων που είναι γειτονικές είτε στην μια κορυφή της ακμής είτε στην άλλη. Αυτό μπορεί να απαιτήσει μέχρι και Θ(|V|) συγκρίσεις όταν είναι παρούσες πολλές ακμές. Αντίθετα με τη αναπαράσταση με πίνακα γειτνίασης, η χρονική πολυπλοκότητα είναι O(1) καθώς θα εξετάζαμε την καταχώρηση (i,j) του πίνακα.

Κόμβος	Γειτονικές Κορυφές
0	1,2
1	2,4
2	3
3	4
4	

Πίνακας 2: Λίστα Γειτνίασης του γραφήματος της εικόνας 9

Συνοπτικά: [[0,1], [0,2], [1,2], [1,4], [2,3], [3,4]]

Κεφάλαιο 5: Νευρωνικά Δίκτυα Γράφων (GNN)

Όπως έχει αναφερθεί ήδη, πολλά είδη δεδομένων με εφαρμογές στον πραγματικό κόσμο μπορούν να αναπαρασταθούν από γράφους, όπως η ανάλυση εικόνας [8], η πρόβλεψη της ροής της κίνησης στα οδικά δίκτυα [9] και οι χημικές δομές των μορίων [10]. Οι γράφοι όμως για τέτοιου είδους δεδομένα είναι αρκετά πολύπλοκοι για να μπορέσουν οι υπάρχοντες αλγόριθμοι μηχανικής μάθησης να τους επεξεργαστούν, αφού εξειδικεύονται σε απλούς τύπους δεδομένων. Μάλιστα, προϋποθέτουν και την ανεξαρτησία της κάθε περίπτωσης, το οποίο για δεδομένα γράφων δεν ισχύει καθώς ο κάθε κόμβος μοιράζεται μέσω των ακμών του πληροφορίες με τους γειτονικούς του σε αυτόν κόμβους. Επιπλέον δεν υπάρχει η έννοια της διάταξης και του σημείου αναφοράς στους γράφους, όπως υπάρχει για παράδειγμα σε μια εικόνα ή σε μια ακολουθία λέξεων. Για να καλυφθεί η αναγκαιότητα επεξεργασίας τέτοιων σύνθετων δομών δεδομένων, δημιουργήθηκαν τα Νευρωνικά Δίκτυα Γράφων (Graph Neural Networks), τα οποία αποτελούν μια κατηγορία μεθόδων βαθιάς μάθησης που έχουν σχεδιαστεί για την εξαγωγή συμπερασμάτων και προβλέψεων σε δεδομένα που περιγράφονται από γραφήματα [11].

5.1: Αμεταβλητότητα ως προς τις μεταθέσεις (Permutation Invariance)

Όπως αναφέρθηκε, οι γράφοι δεν έχουν μοναδική κανονική (canonical) διάταξη των κόμβων τους και μάλιστα ένας γράφος μπορεί να έχει μέχρι και n! διαφορετικές αναπαραστάσεις μέσω των πινάκων γειτνίασής του, , όπου n το πλήθος των κόμβων του γράφου. Έτσι αν διαισθητικά κανείς θεωρούσε ότι θα μπορούσαμε να εκπαιδεύσουμε κάποια από τις υπάρχουσες αρχιτεκτονικές νευρωνικών δικτύων με είσοδο τον πίνακα γειτνίασης A ενός γράφου, αυτό θα ήταν εσφαλμένο καθώς τέτοιες αρχιτεκτονικές δεν είναι αμετάβλητες ως προς τις μεταθέσεις και έτσι ο ίδιος γράφος με διαφορετικό πίνακα γειτνίασης A θα έδινε διαφορετικό αποτέλεσμα. Επιπλέον προφανώς αν εκπαιδεύαμε το νευρωνικό με γράφους, οι οποίοι έχουν διαφορετικού πλήθος κόμβων θα έπρεπε κάθε φορά να θεωρούμε ξεχωριστό νευρωνικό με διαστάσεις εισόδου τόσες όσες και το πλήθος κόμβων του εκάστοτε γράφου. Τα νευρωνικά δίκτυα γράφων είναι σχεδιασμένα ώστε να έχουν συναρτήσεις οι οποίες να είναι αμετάβλητες ως προς τις μεταθέσεις και να μπορούν να διαχειριστούν τα προβλήματα που αντιμετωπίζουν τα κλασικά νευρωνικά δίκτυα, με τρόπους που θα δούμε στις επόμενες ενότητες.

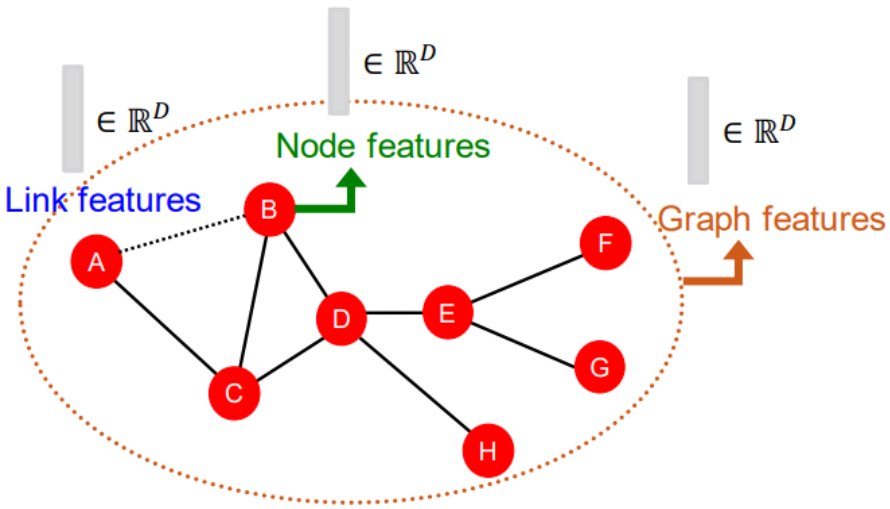
5.2: Είδη προβλέψεων μέσω των GNN

Τα είδη των βασικών προβλέψεων που μπορούν να επιτύχουν τα GNN είναι τρία: Πρόβλεψη αναφορικά με τους κόμβους (Node-level prediction), η οποία μπορεί να αφορά το νευρωνικό να ανακαλύψει κάποια ιδιότητα σε συγκεκριμένους κόμβους ή να βρει τον

ρόλο του κάθε κόμβου στο γράφο (ταξινόμηση), πρόβλεψη αναφορικά με τις ακμές (Link-level prediction) η οποία μπορεί να εξετάζει την ύπαρξη συσχέτισης μεταξύ κόμβων του γράφου ή ακόμα και πρόβλεψη ιδιοτήτων των συσχετίσεων μεταξύ κάποιων κόμβων, και πρόβλεψη αναφορικά με ολόκληρο το γράφο (Graph-level prediction), η οποία προβλέπει μια ιδιότητα ολόκληρου του γράφου ή σε προβλήματα ταξινόμησης, σε ποια κατηγορία ανήκει ο γράφος.

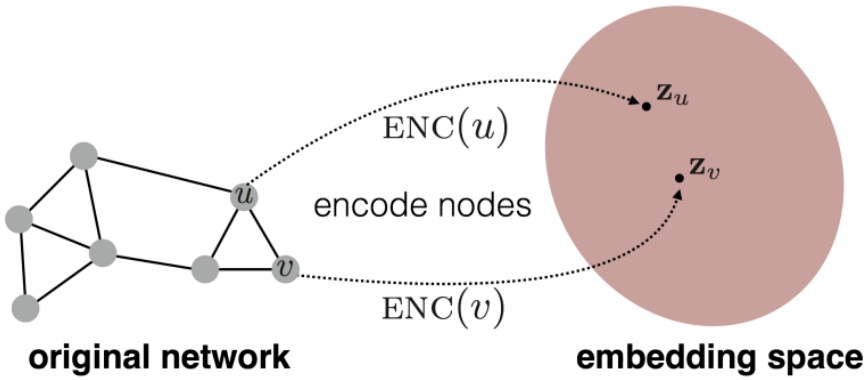
5.3: Αναπαράσταση και Κωδικοποίηση Χαρακτηριστικών (features) Κόμβων, Ακμών και Γράφων στο Embedding Space

Στα προβλήματα μηχανικής μάθησης σημαντικό ρόλο για την καλή απόδοση του μοντέλου παίζει η κατασκευή και η επιλογή των κατάλληλων χαρακτηριστικών (features) για τα δεδομένα μας. Στα GNN αντίστοιχα, η χρήση και η κατασκευή αποτελεσματικών χαρακτηριστικών στους γράφους είναι καίρια για να επιτύχουμε καλή απόδοση στο μοντέλο μας. Όπως και στο είδος των προβλέψεων διακρίνουμε τρία είδη χαρακτηριστικών: στους κόμβους, στις ακμές και στο επίπεδο του γράφου. Τα χαρακτηριστικά μπορούν να κατηγοριοποιηθούν σε δομικά, τα οποία αφορούν την τοπολογία του δικτύου, για παράδειγμα για κάποιο κόμβο μπορούν να είναι ο βαθμός εισόδου/εξόδου του, η σημαντικότητα του στον γράφο (node centrality) κλπ, και επιπρόσθετα χαρακτηριστικά που αφορούν συγκεκριμένες ιδιότητες του κόμβου, ανάλογα το τι αντιπροσωπεύει στο εκάστοτε πρόβλημα.
Τα χαρακτηριστικά αυτά είτε αφορούν ακμές, είτε κόμβους είτε το γράφο αναπαρίστανται μέσω d-διαστάσεων διανύσματα, όπου d το πλήθος των ξεχωριστών χαρακτηριστικών του εκάστοτε αντικειμένου.



Εικόνα 10: Είδη χαρακτηριστικών για γράφους. Πηγή από [49]

Αυτά τα διανύσματα αντιστοιχίζονται όπως αναφέρθηκε σε έναν d-διαστάσεων χώρο για τους κόμβους και ονομάζονται node embeddings. Ο χώρος αυτός είναι πολύ μικρότερος σε διαστάσεις από ότι η συνολική διάσταση του γράφου. Ο στόχος αυτής της διαδικασίας είναι οι κόμβοι που έχουν παρόμοιες ιδιότητες στον γράφο να μπορέσουν να κωδικοποιηθούν με τέτοιο τρόπο έτσι ώστε η ομοιότητά τους να αντικατοπτρίζεται και στον χώρο αυτόν. Έτσι με τρόπο αυτόματο η δομική πληροφορία του γράφου μπορεί να κωδικοποιηθεί κατάλληλα και στη συνέχεια να χρησιμοποιηθεί για οποιοδήποτε είδος πρόβλεψης μας ενδιαφέρει.



Εικόνα 11: Κωδικοποίηση των διανυσμάτων των χαρακτηριστικών στο embedding space. Πηγή από [49]

Θεωρώντας ως u, v δύο κόμβους ενός γράφου όπου x_u και x_v τα διανύσματα των χαρακτηριστικών τους, ψάχνουμε συνάρτηση κωδικοποίησης (encoder function) τέτοια ώστε ENC(u) και ENC(v) να μετατρέπει τα x_u, x_v σε z_u και z_v αντίστοιχα των οποίων οι συντεταγμένες στον καινούργιο χώρο να αντικατοπτρίζουν την ομοιότητα ή μη των κόμβων αυτών στον γράφο. Για την μέτρηση της ομοιότητας χρειάζεται μια συνάρτηση αποκωδικοποίησης. Συνήθως χρησιμοποιείται το εσωτερικό γινόμενο $z_v^T z_u$, το οποίο είναι το συνημίτονο της γωνίας των δύο διανυσμάτων. Τότε όσο πιο όμοια είναι δύο διανύσματα τόσο μεγαλύτερο εσωτερικό γινόμενο έχουν, και το αντίστροφο. Χρειάζεται οπότε να οριστεί η ομοιότητα similarity(u,v) του αρχικού δικτύου έτσι ώστε να ισχύει:

$$similarity = z_y{}^T \cdot z_u \quad (21)$$

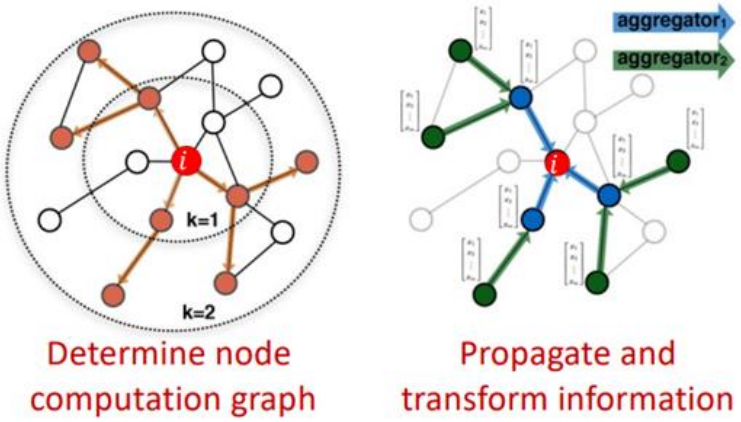
και η συνάρτηση κωδικοποίησης ENC, όπου για έναν οποιοδήποτε κόμβο v θα ισχύει ENC(v) = z_v, με παραμέτρους οι οποίες, μέσω μιας διαδικασίας μάθησης, θα βελτιστοποιούν τα αποτελέσματα της εξίσωσης (21).

Αυτή η συνάρτηση για να παράξει τα embeddings στη τελική τους μορφή z_v πρέπει να λαμβάνει υπ' όψιν τη θέση του κόμβου στο γράφο, δηλαδή τη τοπικότητα του κόμβου, και άρα να μπορεί να επεξεργάζεται την πληροφορία των γειτονικών του κόμβων ανταλλάσσοντας μηνύματα με αυτούς (neural message passing), να συγκεντρώνει με

κάποια συνάρτηση (aggregate) τις εν λόγω πληροφορίες και να εκτελεί σε πολλά επίπεδα μη γραμμικούς μετασχηματισμούς. Παράλληλα, πρέπει να βελτιστοποιούνται οι παράμετροι της συνάρτησης αυτής μέσω μιας διαδικασίας μάθησης. Η αρχιτεκτονική η οποία πληρή όλες τις παραπάνω προδιαγραφές είναι αυτή των Νευρωνικών Δικτύων Γράφων.

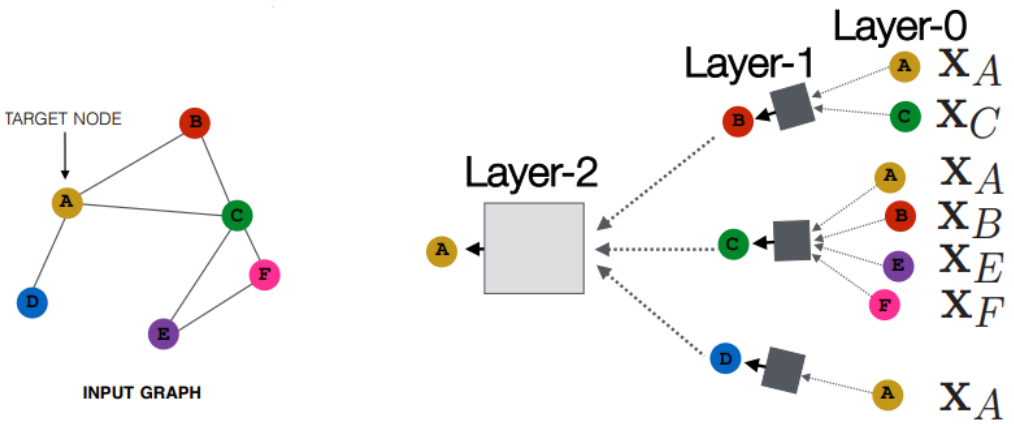
5.4: Γράφος Υπολογισμού (Computational Graph) κόμβου.

Κεντρική έννοια στα GNN είναι οι γράφοι υπολογισμού των κόμβων. Βασίζεται στην ιδέα ότι η γειτονιά του εκάστοτε κόμβου καθορίζει την αρχιτεκτονική του νευρωνικού δικτύου. Όπως φαίνεται στην εικόνα (12), για τον κόκκινο κόμβο i παρατηρούμε την συνδεσιμότητά του με τους γείτονες του σε απόσταση 1 και με τους γείτονες των γειτόνων του σε απόσταση 2. Για τον γράφο υπολογισμού του σχηματίζονται οι συνδυασμοί αυτοί και μέσω νευρωνικών δικτύων γίνεται η συγκέντρωσή τους, η οποία και αποτυπώνει με αυτόν τον τρόπο τη δομική πληροφορία για τον κόμβο αυτό χρησιμοποιώντας και τα χαρακτηριστικά των γειτονικών κόμβων ταυτόχρονα. Έτσι σε ένα node classification πρόβλημα για τον κόμβο i η διαδικασία είναι η εξής: Ο i θα πάρει πληροφορίες από τους γείτονές του και οι γείτονες θα πάρουν πληροφορίες από τους γείτονες των γειτόνων. Το νευρωνικό δίκτυο με τις παραμέτρους του Θ, θα μάθει πώς να διαδίδει αυτές τις πληροφορίες, πώς να τις συγκεντρώνει και να τις μετασχηματίζει κατά μήκος των ακμών του δικτύου και πώς να δημιουργεί ένα νέο μήνυμα που στη συνέχεια ο επόμενος κόμβος στο γράφο υπολογισμού μπορεί και πάλι να συγκεντρώνει, να μετασχηματίζει και να υπολογίζει.
Ο τρόπος με τον οποίο μπορούμε να σκεφτούμε τα νευρωνικά δίκτυα γράφων είναι μια διαδικασία δύο βημάτων. Στη διαδικασία του πρώτου βήματος, καθορίζουμε τον γράφο υπολογισμού του κόμβου και στη δεύτερη διαδικασία, διαδίδουμε και μετασχηματίζουμε τις πληροφορίες σε αυτόν. Αυτός ο γράφος υπολογισμού καθορίζει την αρχιτεκτονική του υποκείμενου νευρωνικού δικτύου, το οποίο μαθαίνει πώς να διαδίδει πληροφορίες σε όλη τη δομή του γράφου υπολογισμού για τον υπολογισμό των embeddings του εκάστοτε κόμβου.



Εικόνα 12: Γράφος υπολογισμού για τον κόμβο i. Πηγή από [49]

Τα GNN μπορούν να έχουν πολλά επίπεδα και άρα ο κάθε κόμβος στο γράφο υπολογισμού του να παίρνει πληροφορίες από κόμβους σε απόσταση τόση όση και το πλήθος των επιπέδων που ορίζει η αρχιτεκτονική του μοντέλου. Στο γράφο υπολογισμού έχουμε τα αρχικά διανύσματα x_v των κόμβων όπου στη συνέχεια μετασχηματίζονται και διαδίδονται με τον τρόπο που περιγράφηκε για να παράξουν το τελικό embedding z για τον κόμβο-στόχο. Σχηματικά αυτό φαίνεται στην εικόνα 13, όπου τα γκρίζα κουτιά είναι στην πραγματικότητα νευρωνικά δίκτυα υπεύθυνα για τις παραπάνω διαδικασίες. Σημαντικό είναι οι συναρτήσεις συγκέντρωσης που χρησιμοποιούν να είναι αμετάβλητες ως προς τη διάταξη, όπως οι sum, average, maximum καθώς είναι κατ επέκταση και αμετάβλητα ως προς τις μεταθέσεις.



Εικόνα 13: Γράφος υπολογισμού με 2 επίπεδα για τον κόμβο A. Πηγή από [49]

Στα GNN, ανεξάρτητα το είδος πρόβλεψης που θέλουμε να κάνουμε (graph, node, edge prediction) ο υπολογισμός του τελικού embedding z συμβαίνει για κάθε κόμβο κάθε γράφου στα οποία εκπαιδεύεται το GNN. Αυτά τα embedding χρησιμοποιούνται στη συνέχεια κατάλληλα ανάλογα το είδος της πρόβλεψης που μας ενδιαφέρει. Έτσι για τον γράφο της εικόνας 13 το GNN θα παράξει τα embeddings μέσω γράφων υπολογισμού και

για τους έξι κόμβους, μέσω της ξεχωριστής αρχιτεκτονικής νευρωνικών δικτύων που έχει ο καθένας λόγω της διαφορετικής τοπολογικής δομής του στο γράφο.

5.5: Αρχιτεκτονικές Επιπέδων Νευρωνικών Δικτύων Γράφων

Υπάρχουν πολλές αρχιτεκτονικές επιπέδων για τα GNN και οι διαφοροποιήσεις τους εντοπίζονται στο τι είδος συνάρτηση συγκέντρωσης χρησιμοποιούν και πως μετασχηματίζουν και μεταβιβάζουν τα μηνύματα των γειτόνων, στον γράφο υπολογισμού των κόμβων. Στη συνέχεια θα παρουσιαστούν οι τρεις βασικότερες αρχιτεκτονικές επιπέδων που χρησιμοποιούνται στα GNN, συγκεκριμένα τα Συνελικτικά Δίκτυα Γράφων (Graph Convolutional Networks - GCN), τα GraphSAGE και τα Graph Attention Networks (GAT), εκ των οποίων τα SAGE και GATv2 (μια τροποποιημένη πιο εκφραστική αρχιτεκτονική από τα GAT [16]) είναι οι αρχιτεκτονικές που χρησιμοποιήθηκαν στη παρούσα εργασία.

5.5.1: Συνελικτικά Δίκτυα Γράφων (Graph Convolutional Networks - GCN)

Όπως προκύπτει και από το όνομά τους, τα συνελικτικά δίκτυα γράφων [12] αποτελούν την γενίκευση των συνελικτικών νευρωνικών δικτύων (CNN) [13] για δεδομένα γράφων.

Συγκεκριμένα για το κάθε embedding η διαδικασία της πρόσθιας τροφοδότησης στον γράφο υπολογισμού στα GCN γίνεται ως εξής:

1. Αρχικοποιούνται τα κρυφά διανύσματα:

$$h_v^0 = X_v, (feature\ vector) \quad (22)$$

όπου X_v το διάνυσμα με τα αρχικά χαρακτηριστικά του κόμβου v.

2. Σε κάθε επίπεδο του δικτύου πραγματοποιείται ο μετασχηματισμός:

$$h_v^k = \sigma\left(W_\kappa \sum \frac{h_u^{k-1}}{N(v)} + B_k h_v^{k-1}\right), όπου\ k = 1, \dots, k-1 \quad (23)$$

ο οποίος παράγει τα υψηλότερης τάξης ενδιάμεσα κρυφά embeddings του κόμβου.

όπου ο όρος:

$$W_\kappa \sum \frac{h_u^{k-1}}{N(v)}$$

απλά υπολογίζει τον μέσο όρο των κρυφών διανυσμάτων των γειτόνων του v. Προφανώς θα μπορούσε να μην είναι ο μέσος όρος αλλά κάποια άλλη συνάρτηση, αλλά η σχεδιαστική επιλογή για τα GCN είναι αυτή. Στη συνέχεια το αποτέλεσμα αυτό πολλαπλασιάζεται για να μετασχηματιστεί με έναν πίνακα βαρών εκπαίδευσης, υπεύθυνο για τα βάρη που σχετίζονται με τη συνάθροιση των κρυφών διανυσμάτων των γειτόνων.

Ο όρος:

$$B_k h_v^{k-1}$$

είναι το embedding του κόμβου v στο προηγούμενο επίπεδο πολλαπλασιασμένο με το bias B_k, το οποίο είναι και άλλος ένας πίνακας βαρών εκπαίδευσης που σχετίζεται με τον μετασχηματισμό του ίδιου του κρυφού διανύσματος του κόμβου v.

Ο όρος "σ" είναι η μη γραμμική συνάρτηση ενεργοποίησης που εφαρμόζεται (π.χ. ReLu, tanh κλπ) .

3. Εφαρμόζεται η τελευταία εξίσωση στο τελευταίο επίπεδο:

$$z_v = h_v^K \ (24)$$

όπου z_v είναι το τελικό embedding του κόμβου z μετά από k επίπεδα συγκέντρωσης των γειτόνων του v.
Σημαντικό είναι το ότι οι πίνακες W_k, B_k δεν είναι ξεχωριστοί για κάθε κόμβο στο γράφο αλλά όλοι οι κόμβοι χρησιμοποιούν τους ίδιους.

Για να εκπαιδευτεί το μοντέλο χρησιμοποιούμε μια κατάλληλη συνάρτηση κόστους (όπως την L2 που αναφέρθηκε στο 3.3.1) για τα embeddings, με βάση την εξίσωση:

$$\min_{\theta} \boldsymbol{L}(y, f(z_v))$$

αν πρόκειται για πρόβλημα επιβλεπόμενης μάθησης, και στη συνέχεια τρέχουμε τον αλγόριθμο SGD για να εκπαιδεύσουμε τις παραμέτρους των βαρών.
Μετά την εκπαίδευση, αφού έχουν καθοριστεί τα βάρη, μπορούμε να τα χρησιμοποιήσουμε και να τα εφαρμόσουμε σε γράφους που δεν έχει χρησιμοποιήσει για εκπαίδευση το μοντέλο και έτσι να γενικεύσουμε τη διαδικασία παραγωγής των embeddings καθώς και να το χρησιμοποιήσουμε για προβλέψεις.

5.5.2: GraphSAGE

Ο μετασχηματισμός που πραγματοποιεί στα διανύσματα/embeddings σε κάθε επίπεδο του δικτύου η αρχιτεκτονική του GraphSAGE [14] δίνεται από τον τύπο:

$$h_v^{(k)} = \sigma\left(W^{(k)} \cdot CONCAT\left(h_v^{(k-1)}, AGG\left(\left\{h_u^{(k-1)}, \forall\, u \in N(v)\right\}\right)\right)\right) (25)$$

Το GraphSAGE βασίζεται στο GCN, αλλά το επεκτείνει σε αρκετές πτυχές του.
Όπως φαίνεται και από την εξίσωση (25), η συνάρτηση συγκέντρωσης (AGG()) δεν είναι αναγκαστικά όπως στο GCN η μέση τιμή αλλά επιτρέπονται πολλές διαφορετικές επιλογές συναρτήσεων οι οποίες μπορούν να συνδυαστούν. Για παράδειγμα, μπορούμε να χρησιμοποιήσουμε MLP για τον μετασχηματισμό των ενδιάμεσων κρυφών διανυσμάτων των κόμβων, καθώς δεν είναι απαραίτητο να χρησιμοποιήσουμε κάποια γραμμική συνάρτηση, και στη συνέχεια να πάρουμε τη μέση τιμής τους ή το άθροισμα τους, όπως φαίνεται από την παρακάτω εξίσωση:

$$AGG = Mean(\left\{MLP\left(h_u^{(k-1)}\right), \forall\, u \in N(v)\right\}) (26)$$

Επιπλέον το κρυφό διάνυσμα του ίδιου του κόμβου συνενώνεται με αυτά των γειτόνων του ((CONCAT()), προσφέροντας περισσότερη εκφραστική δύναμη στο μοντέλο. Στη συνέχεια το αποτέλεσμα αυτό πολλαπλασιάζεται με τον πίνακα βαρών των παραμέτρων του μοντέλου και εφαρμόζεται η μη γραμμική συνάρτηση.

Τέλος, το GraphSAGE προσθέτει προαιρετικά και την έννοια της κανονικοποίησης L2 για τα embeddings σε κάθε επίπεδο., όπως δίνεται κάτωθι:

$$h_v^{(k)} \leftarrow \frac{h_u^{(k)}}{\left\|h_u^{(k)}\right\|_2}, \forall\, v \in V \text{ όπου } \|u\|_2 = \sqrt{\sum_i u_i^2}\ (l2 - Norm)\ (27)$$

Έτσι μετά το βήμα κανονικοποίησης L2, όλα τα διανύσματα θα έχουν την ίδια L2 νόρμα.
Χωρίς την κανονικοποίηση τα embeddings είναι πιθανό να έχουν αρκετές διαφορετικές κλίμακες, γεγονός το οποίο οδηγεί σε μερικές περιπτώσεις σε πιο αργή σύγκλιση του αλγορίθμου SGD. Έτσι, με την κανονικοποίηση μπορεί να επιτύχουμε καλύτερη απόδοση του μοντέλου.

5.5.3: Graph Attention Networks (GAT)

Ο μετασχηματισμός που πραγματοποιεί στα διανύσματα/embeddings σε κάθε επίπεδο του δικτύου η αρχιτεκτονική του GAT [15], δίνεται από τον τύπο:

$$h_v^{(k)} = \sigma(\sum_{u \in N(v)} a_{vu} W^{(k)} h_u^{(k-1)}) \quad (28)$$

όπου ο όρος α_{vu} καλείται attention weight και αποδίδει την σημαντικότητα που παίζει το embedding κάθε κόμβου u στο embedding του κόμβου v, η οποία δεν είναι η ίδια για διάφορους κόμβους.

Τα GCN θα μπορούσαν να θεωρήσουμε σαν έμμεση ειδική περίπτωση αυτού του όρου όπου το $\frac{1}{N(v)}$, ισούται με το α_{vu}, και το οποίο ορίζεται ρητά με βάση το βαθμό του κόμβου v, και στη πραγματικότητα σημαίνει ότι δίνεται η ίδια σημαντικότητα στους γειτονικούς κόμβους u, οπότε ουσιαστικά δεν υπάρχει κάποια αξιοποίηση του ποιοί κόμβοι είναι πιο σημαντικοί για τον κόμβο v.

Η ιδέα του attention weight α_{vu} είναι εμπνευσμένη από τη γνωστική προσοχή (cognitive attention) και δίνει έμφαση στα σημαντικότερα δεδομένα της εισόδου θεωρώντας ότι η υπολογιστική ισχύς πρέπει να αφιερώνεται κυρίως σε αυτά. Το ποια δεδομένα εισόδου θεωρούνται σημαντικότερα καθορίζεται από το είδος του προβλήματος και μαθαίνεται μέσω της εκπαίδευσης του μοντέλου. Οι κόμβοι παρακολουθούν τα μηνύματα των γειτόνων τους και έμμεσα τους αναθέτουν διαφορετικά μεταξύ τους βάρη. Μαθηματικά αυτό γίνεται ως εξής:
Θα υπολογίσουμε τα βάρη α_{vu} ως παράγωγα ενός μηχανισμού προσοχής (attention mechanism) α, μέσω του οποίου υπολογίζουμε τους συντελεστές προσοχής (attention coefficients) e_{uv} μεταξύ των κόμβων u,v ως εξής:

$$e_{vu} = a(W^{(k)} h_u^{(k-1)}, W^{(l)} h_v^{(k-1)}) \quad (29)$$

όπου ο συντελεστής e_{vu} υποδεικνύει την σημαντικότητα του embedding του u στον κόμβο v.

Στη συνέχεια κανονικοποιούμε το e_{vu} στο τελικό βάρος α_{vu} έτσι ώστε:

$$\sum_{u \in N(v)} a_{vu} = 1$$

μέσω της εξίσωσης:

$$a_{vu} = \frac{\exp\left(e_{vu}\right)}{\sum_{k \in N(v)} \exp\left(e_{vk}\right)}$$

και τελικά χρησιμοποιείται στην εξίσωση (28).

Αυτό που δεν έχει οριστεί ρητώς στις παραπάνω εξισώσεις, είναι η μορφή του μηχανισμού προσοχής α. Γενικά, μπορεί να είναι το οτιδήποτε, όπως για παράδειγμα ένα απλό νευρωνικό δίκτυο ενός επιπέδου, του οποίου τα βάρη αποτελούν τις παραμέτρους προς μάθηση του α. Π..χ. το α μπορεί να δίνεται από την εξίσωση:

$$e_{vu} = a(W^{(k)} h_v^{(k-1)}, W^{(k)} h_u^{(k-1)}) = Linear(Concat(W^{(k)} h_v^{(k-1)}, W^{(k)} h_u^{(k-1)})$$

Οι παράμετροι της συνάρτησης α μαθαίνονται ταυτόχρονα με τους πίνακες βαρών W του GNN.
Όλη η λειτουργία των GAT ορισμένες φορές μπορεί να αποδειχθεί δύσκολη ως προς τη σύγκλιση και για αυτό χρησιμοποιείται μια διαδικασία που ονομάζεται multi-head attention, η οποία σταθεροποιεί την διαδικασία μάθησης. Συγκεκριμένα δημιουργούνται πολλές ρέπλικες των embeddings, οι οποίες χρησιμοποιούν διαφορετικό σύνολο παραμέτρων για των μηχανισμό προσοχής α, οι οποίες στη συνέχεια συναθροίζονται όπως φαίνεται παρακάτω:

$$h_v^{(k)}[1] = \sigma\Big(\sum_{u \in N(v)} a_{vu}^1 W^{(k)} h_u^{(k-1)}\Big) \quad (30)$$

$$h_v^{(k)}[2] = \sigma\Big(\sum_{u \in N(v)} a_{vu}^2 W^{(k)} h_u^{(k-1)}\Big) \quad (31)$$

$$h_v^{(k)}[3] = \sigma\Big(\sum_{u \in N(v)} a_{vu}^3 W^{(k)} h_u^{(k-1)}\Big) \quad (32)$$

Και τελικά:

$$h_v^{(k)} = AGG\Big(h_u^{(k)}[1], h_u^{(k)}[2], h_u^{(k)}[3]\Big) \quad (33)$$

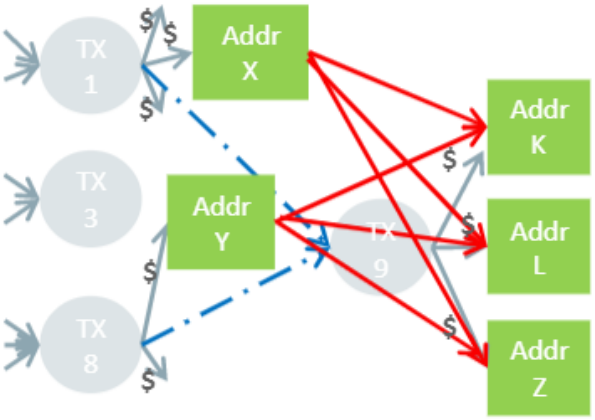
ΠΡΑΚΤΙΚΟ ΜΕΡΟΣ

Κεφαλαιο 6: Το blockchain του Bitcoin ως Γράφος Συναλλαγών.

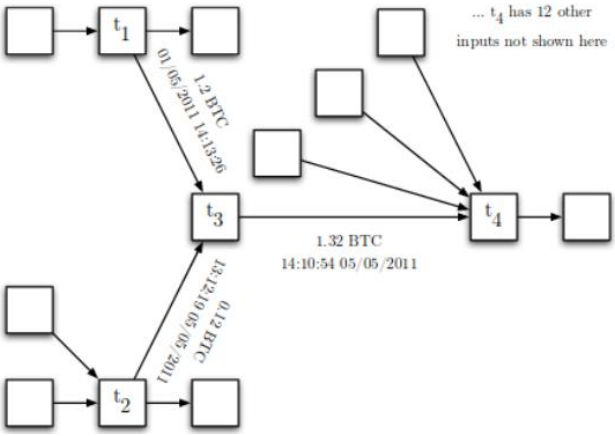
Όπως ορίστηκε από τον δημιουργό του, το Blockchain του Bitcoin είναι ένα ηλεκτρονικό σύστημα πληρωμών το οποίο αφαιρεί την ανάγκη ύπαρξης κάποιου κεντρικού διαμεσολαβητή (π.χ. τράπεζες) για να εξασφαλιστεί η εγκυρότητα της συναλλαγής και τη βασίζει σε κρυπτογραφικές μεθόδους απόδειξης [17]. Έτσι, επιτρέπει στους συμμετέχοντες να πραγματοποιούν συναλλαγές απευθείας μεταξύ τους. Το πρωτόκολλο είναι κατανεμημένο και ψευδοανώνυμο, καθώς κάθε χρήστης μπορεί να έχει πορτοφόλια (wallets) στα οποία αντιστοιχούν αλφαριθμητικές διευθύνσεις (addresses) χωρίς να φαίνεται η ταυτότητα τους. Τελικά, στη διάρκεια του χρόνου οι συναλλαγές και οι πληροφορίες τους μπαίνουν σε blocks, τα οποία συνδέονται μεταξύ τους και σχηματίζουν μια αλυσίδα (blockchain). Αυτή η αλυσίδα είναι δημόσια και κάθε χρήστης του blockchain έχει πρόσβαση στο ιστορικό όσων συναλλαγών έχουν πραγματοποιηθεί από την αρχή του Bitcoin μέχρι και τώρα, αλλά, λόγω της ψευδοανωνυμίας, δεν ξέρει τους συμμετέχοντες των συναλλαγών.

Αυτές οι συναλλαγές μπορούν να αναπαρασταθούν ως ανταλλαγή αξίας μεταξύ δύο ή περισσότερων οντοτήτων. Προκύπτει φυσικά, ότι εφόσον έχουμε πρόσβαση σε μια μορφή δεδομένων με οντότητες και τις μεταξύ τους σχέσεις, μπορούμε να την αναπαραστήσουμε σε μορφή γράφου. Πράγματι, υπάρχουν έρευνες στη βιβλιογραφία για τον τρόπο αναπαράστασης του blockchain σε μορφή γράφου και έχουν προκύψει διάφορες παραλλαγές. Οι Whu κ.α. [18] χωρίζει τις δυνατές αναπαραστάσεις σε τρεις κατηγορίες.

1) Γράφοι Διευθύνσεων (Address Graphs): Κόμβοι θεωρούνται οι διευθύνσεις των πορτοφολιών των χρηστών. Οι συναλλαγές μεταξύ διευθύνσεων αποτυπώνονται μέσω κατευθυνόμενων ακμών οι οποίες συνδέουν τις διευθύνσεις που στέλνουν BTC με αυτές στις οποίες καταλήγουν τα ποσά αυτά. Έτσι αναπαρίσταται η ροή των Bitcoins μεταξύ των διευθύνσεων.
2) Γράφοι Συναλλαγών (Transaction Graphs): Σε αντίθεση με το γράφο διευθύνσεων, κόμβοι θεωρούνται οι ίδιες οι συναλλαγές και συνδέονται μέσω κατευθυνόμενων ακμών οι οποίες αποτυπώνουν τη ροή των bitcoins που χρησιμοποιούνται από κάποιες συναλλαγές σε άλλες.
3) Γράφοι Συστάδων (Cluster Graphs): Παρόμοιοι με τους γράφους διευθύνσεων με την διαφορά ότι με βάση κάποια χειριστική συνάρτηση, ομαδοποιούνται κατάλληλα διευθύνσεις μεταξύ τους.



Εικόνα 14: Γράφος διευθύνσεων του Blockchain. Μια συναλλαγή μπορεί να έχεις πολλές διευθύνσεις εισόδου (input addresses) που συνεισφέρουν σε αυτή και πολλές διευθύνσεις εξόδου (output addresses) στις οποίες καταλήγουν τα BTC. Αυτό αποτυπώνεται με την δημιουργία ακμών μεταξύ όλων των διευθύνσεων που εμπλέκονται σε αυτή τη συναλλαγή. Πηγή από [18]



Εικόνα 15: Γράφος συναλλαγών του Blockchain. Κάθε ακμή αντιπροσωπεύει ένα μέρος του ποσού μιας προηγούμενης χρονικά συναλλαγής που χρησιμοποιήθηκε ως είσοδος για την επόμενη. Πηγή από [22]

Οι διαφορετικές αναπαραστάσεις χρησιμοποιούνται για να διαχειριστούν διαφορετικά είδους προβλήματα. Για παράδειγμα, οι Fleder κ.α. [20] χρησιμοποιούν το γράφο διευθύνσεων, τον οποίο χαρακτηρίζουν ως γράφο συναλλαγών καθώς δεν κάνουν τον διαχωρισμό του Whu [18], μαζί με άλλες παραμέτρους μέσω του πλαισίου ανάλυσης γραφημάτων που ανέπτυξαν για να αντιστοιχίσουν τις διευθύνσεις κάποιων πορτοφολιών με τους πραγματικούς ανθρώπους που τις χρησιμοποιούν. Οι Sharma [21] χρησιμοποίησαν πάλι των γράφων διευθύνσεων μιας χρονικής περιόδου του Bitcoin, για μια συγκεκριμένη διεύθυνση, σαν είσοδο για το Temporal Graph Convolutional Network (T-GCN), ένα είδος GNN, για να προβλέψουν το ποσό των Bitcoins που θα λάβει ο χρήστης της διεύθυνσης αυτής σε μια συγκεκριμένη χρονική στιγμή. Οι Sharma κ.α. [22] χρησιμοποιούν τον γράφο συναλλαγών του Bitcoin για να αντλήσουν χρήσιμα στατιστικά και να αναγνωρίσουν

τοπολογικά μοτίβα που μπορούν να χρησιμεύσουν σε οργανισμούς, οικονομολόγους κ.λ.π. Αντίστοιχα οι Ron κ.α. [23] εξήγαγαν μέσω ποσοτικής ανάλυσης πολλές στατιστικές ιδιότητες του γράφου συναλλαγών και κατέληξαν σε ενδιαφέροντα συμπεράσματα, όπως το ότι πολλές από τις μεγάλες συναλλαγές στο blockchain έχουν προέλθει από μια μεγάλη συναλλαγή που έλαβε χώρα τον Νοέμβριο του 2010.

Τα παραπάνω παραδείγματα δείχνουν τα πλεονεκτήματα και τις δυνατότητες που έχουν διάφορες προσεγγίσεις ανάλυσης του γράφου συναλλαγών για πολλές περιπτώσεις χρήσης. Θα εστιάσουμε τη προσοχή μας σε δύο έρευνες [24], [25] που αναδεικνύουν την προβλεπτική ισχύ των τοπολογικών χαρακτηριστικών του γράφου συναλλαγών του Bitcoin, στον οποίο καταφεύγουν μέσω graph mining μεθόδων για την εξαγωγή μοτίβων και ιδιοτήτων του, τα οποία στη συνέχεια χρησιμοποιούν για την πρόβλεψη της τιμής του Bitcoin.
Συγκεκριμένα οι Ancora κ.α. [24] εισάγουν στη βιβλιογραφία την έννοια των graphlets, τα οποία είναι ουσιαστικά υπογράφοι του γράφου συναλλαγών του Bitcoin για κάποια χρονικά διαστήματα, με συγκεκριμένες τοπικές για το εκάστοτε graphlet τοπολογικές ιδιότητες. Χρησιμοποιώντας στατιστικούς ελέγχους**,** αποδεικνύουν ότι κάποια είδη graphlets με συγκεκριμένες τοπολογικές ιδιότητες σε συνδυασμό με την συχνότητα που παρουσιάζονται στο δίκτυο, έχουν μεγάλη προβλεπτική ισχύ για τη τιμή του Bitcoin. Χρησιμοποιώντας το πίνακα συχνοτήτων εμφάνισης και τον πίνακα πληθικότητας των chainlets σε Random Forest αλγορίθμους δείχνουν πως βελτιώνονται οι RMSE μετρικές αφού χρησιμοποιηθούν τα chainlets σε σχέση με πριν.
Οι Li κ.α. [25] ορίζουν και αυτοί, ωστόσο με διαφορετική αναπαράσταση από ότι οι Ancora κ.α. [24], τον γράφο του Bitcoin και στη συνέχεια, αντλώντας τα τοπολογικά χαρακτηριστικά διαφόρων υπογράφων του δικτύου, κατασκευάζουν τον πίνακα συχνοτήτων εμφάνισης τους. Αυτα τα χαρακτηριστικά τα τροφοδοτούν σε SVM (Support Vector Machines) και τα αποτελέσματα τους για τις προβλέψεις της τιμής του Bitcoin έχουν πολύ χαμηλές τιμές MAPE, αναδεικνύοντας έτσι με τη σειρά τους την ισχυρή προβλεπτική ικανότητα των τοπολογικών μοτίβων στο γράφο συναλλαγών του Bitcoin.

Και οι δύο έρευνες όμως, χρησιμοποιούν τον γράφο συναλλαγών για να αντλήσουν, μέσω graph mining τεχνικών, στατιστικά χαρακτηριστικά για τα είδη των υπογράφων που απαντώνται σε αυτόν (συχνότητα εμφάνισης, πληθικότητα κλπ) και στη συνέχεια να τα τροφοδοτήσουν σε ML μοντέλα, τα οποία έχουν ήδη δοκιμαστεί για τη πρόβλεψη τιμής. Με έναυσμα αυτές τις έρευνες, έχοντας πειστεί ότι ο γράφος του Blockchain του BTC έχει πληροφορία σημαντική για τη πρόβλεψη της τιμής του, η έρευνά μας προβαίνει στις εξής συνεισφορές:

- Κατασκευάζουμε τις δικές μας παραλλαγές του γράφου συναλλαγών του Bitcoin, οι οποίες βασίζονται στον αλγόριθμο κατασκευής όπως παρουσιάζεται από τους

Tharani κ.α. [26], τον οποίο επεκτείνουμε, τροποποιούμε και παραμετροποιούμε έτσι ώστε να είναι κατάλληλος για το είδος του task το οποίο μελετάμε.

- Είμαστε οι πρώτοι οι οποίοι εφαρμόζουν Graph Neural Networks για το πρόβλημα πρόβλεψης της τιμής του Bitcoin για μελλοντικές χρονικές στιγμές, το οποίο έχει ως είσοδο τους γράφους συναλλαγών του Bitcoin. Τα GNN αυτόματα μαθαίνουν μέσω των embeddings των δομικών χαρακτηριστικών των κόμβων του γράφου, δηλαδή των συναλλαγών, τις υποκείμενες δομές των γράφων που έχουν προβλεπτική ισχύ. Έτσι αποκτούν την ικανότητα να γενικεύουν και να αναγνωρίζουν τοπολογικά μοτίβα και χαρακτηριστικές ιδιότητες συναλλαγών σε πραγματικούς μεταγενέστερους χρονικά γράφους του Bitcoin στους οποίους δεν έχουν εκπαιδευτεί.
- Εξετάζουμε σενάρια για τις προβλέψεις, στα οποία προσδίδουμε στις συναλλαγές πέρα από τα δομικά τους χαρακτηριστικά και εξωτερικά οικονομικά χαρακτηριστικά όπως τον όγκο των δολαρίων που συναλλάσσεται ωριαία με Bitcoins (volume usd) και τον κυλιόμενο εκθετικό μέσος όρο δέκα ημερών της τιμής του Bitcoin με τις τιμές που έχουν όταν συμβαίνουν χρονικά οι συναλλαγές αυτές. Είμαστε έτσι οι πρώτοι οι οποίοι εξετάζουν να προσδίδουν στους γράφους των συναλλαγών του Bitcoin εξωτερικά χαρακτηριστικά με έναν γραφοκεντρικό τρόπο και να μελετούν την επίδραση που έχει αυτή η αναπαράσταση στις προβλέψεις της τιμής του Bitcoin.

Κεφάλαιο 7: Κατασκευή Γράφων Συναλλαγών του Bitcoin, Σενάριά τους και Εξεταζόμενες Υπερπαράμετροι των GNN

Στο παρόν κεφάλαιο, θα παρουσιάσουμε τη διαδικασία που ακολουθήσαμε για να κατασκευάσουμε τους γράφους συναλλαγών του Blockchain του Bitcoin. Πρώτα, θα αναφερθούμε στη διαδικασία συλλογής των δεδομένων για τους κόμβους και τις ακμές των γράφων. Στη συνέχεια, θα παρουσιάσουμε την προεπεξεργασία και την κατασκευή των οικονομικών χαρακτηριστικών για τους κόμβους, την τεχνική του κυλιόμενου παραθύρου για την κατασκευή των λιστών συναλλαγών, τις οποίες τελικά θα χρησιμοποιήσουμε στον αλγόριθμο κατασκευής του γράφου συναλλαγών. Τέλος, θα παρουσιάσουμε τη μέθοδο κανονικοποίησης των χαρακτηριστικών, τα διάφορα σενάρια γράφων στα οποία θα εξετάσουμε τα GNN μοντέλα μας, τις αρχιτεκτονικές τους και το σύνολο των υπερπαραμέτρων τους οι οποίες θα εξεταστούν.

7.1 : Συλλογή Δεδομένων

Όπως έχουμε αναφέρει τα δεδομένα των συναλλαγών είναι δημόσια στο Blockchain του. Αν κάποιος θέλει να έχει πρόσβαση σε αυτά αρκεί να κατεβάσει το επίσημο λογισμικό του Bitcoin[1], να φτιάξει ένα πορτοφόλι και να κατεβάσει όλα τα ιστορικά δεδομένα του Blockchain από όταν δημιουργήθηκε μέχρι τώρα. Ύστερα, με τη χρήση κάποιου λογισμικού (parser), να εξάγει τις πληροφορίες σχετικά με τα blocks, τις συναλλαγές και τις διευθύνσεις. Για να αποφύγουμε αυτή τη διαδικασία, καθώς ξεφεύγει του σκοπού της παρούσας εργασίας, πήραμε αυτά τα αρχικά ανεπεξέργαστα δεδομένα των συναλλαγών από το Google Bigquery public dataset crypto_bitcoin[2] για τις ημερομηνίες από 01/01/2018 μέχρι και 26/04/2018. Τα ιστορικά δεδομένα του Bitcoin που αφορούν τη τιμή που είχε εκείνη τη χρονική περίοδο (close price) σε ωριαία βάση και το volume usd, τα πήραμε από το ανταλλακτήριο HitBTC[3].

Ο πίνακας των δεδομένων των συναλλαγών του dataset crypto_bitcoin, έχει συνολικά 34 διαφορετικά πεδία με πληροφορίες σχετικά τόσο με τις συναλλαγές όσο και με το είδος του πρωτοκόλλου που χρησιμοποιείται στο blockchain, τη συμβολική αναπαράσταση του opcode του bitcoin κ.α. για τη κάθε συναλλαγή. Για την κατασκευή των γράφων συναλλαγών, δεν χρειαζόμαστε όλα αυτά τα δεδομένα. Στον πίνακα 3 παρουσιάζονται τα 12 πεδία που χρειάστηκε να χρησιμοποιήσουμε:

[1] https://bitcoin.org/en/

[2] Το ID των δεδομένων είναι: bigquery-public-data.crypto_bitcoin στο σύνδεσμο https://cloud.google.com/bigquery

[3] Μέσω του συνδέσμου: https://www.cryptodatadownload.com/data/hitbtc/

hash	Μοναδική συμβολοσειρά που αντιπροσωπεύει τη συναλλαγή.
size	Το μέγεθος της συναλλαγής σε bytes.
block_hash	Μοναδική συμβολοσειρά που αντιπροσωπεύει το Block στο οποίο ανήκει η συναλλαγή.
block_timestamp	Χρονοσφραγίδα που έλαβε το block στο οποίο ανήκει η συναλλαγή.
input_count	Το πλήθος των πορτοφολιών που συμμετέχουν ως είσοδοι στη συναλλαγή, δηλαδή αποστέλλουν τα απαραίτητα bitcoin για να πραγματοποιηθεί η συναλλαγή.
output_count	Το πλήθος των πορτοφολιών που συμμετέχουν ως έξοδοι στη συναλλαγή, δηλαδή οι παραλήπτες των bitcoin της συναλλαγής.
input_value	Το άθροισμα (σύνολο) των bitcoins όλων των αποστολέων της συναλλαγής (σε satoshi).
output_value	Το άθροισμα (σύνολο) των bitcoins που έλαβαν όλοι οι παραλήπτες της συναλλαγής (σε satoshi).
is_coinbase	Αληθές αν η συναλλαγή είναι coinbase. Η πρώτη συναλλαγή σε ένα block είναι coinbase και είναι ένας ειδικός τύπος συναλλαγής που δημιουργεί ο miner του block για να λάβει την αμοιβή που του αντιστοιχεί αφού έλυσε το κρυπτογραφικό πρόβλημα του proof-of-work μαζί με τις χρεώσεις των συναλλαγών που επικύρωσε.
fee	Η χρέωση που πληρώνεται για να πραγματοποιηθεί η συναλλαγή.
inputs	Εμφωλευμένες πληροφορίες που αφορούν τη κάθε είσοδο της συναλλαγής.

inputs.Spent_transaction_hash	Το αναγνωριστικό της συναλλαγής από την οποία προήλθαν τα bitcoin τα οποία ξοδεύονται στη παρούσα συναλλαγή.
inputs.value	Τα bitcoin τα οποία ξοδεύονται στη παρούσα συναλλαγή και προήλθαν από τη συναλλαγή με το αναγνωριστικό spent_transaction_hash.

Πίνακας 3: Πεδία των συναλλαγών από το σύνολο δεδομένων μας.

Η χρονοσφραγίδα που έλαβε η κάθε συναλλαγή είναι η χρονοσφραγίδα του block στο οποίο ανήκει. Έτσι, συναλλαγές που πραγματοποιήθηκαν με κάποια διαφορά χρονικά, αλλά παρόλα αυτά μπήκαν στο ίδιο μπλοκ, θα έχουν ίδια χρονοσφραγίδα. Αυτό έχει σημασία για όταν κατασκευάζουμε κάποιον γράφο συναλλαγών, καθώς για να εντοπίζουμε για τη κάθε συναλλαγή σε ποιες μεταγενέστερες συναλλαγές χρησιμοποιήθηκαν τα bitcoin της προκειμένου να τις ενώσουμε με ακμές, θα πρέπει να εξετάζουμε όχι μόνο συναλλαγές σε μεταγενέστερα block αλλά και σε αυτό στο οποίο ανήκει.

Για την εξαγωγή αυτού του συνόλου δεδομένων, χρησιμοποιήσαμε SQL ερωτήματα στο Google BigQuery και ο τύπος αρχείου στον οποίον αποθηκεύτηκαν ήταν JSON, έτσι ώστε να είναι πιο εύκολη η διαχείριση των εμφωλευμένων πληροφοριών.

7.2: Χαρακτηριστικά για Κόμβους και Ακμές

Όπως θα δούμε στην ενότητα 7.4 στον αλγόριθμο κατασκευής γράφων συναλλαγών, όσο πιο μεγάλο είναι το χρονικό διάστημα για το οποίο θέλουμε να κατασκευάσουμε κάποιον γράφο, τόσο πιο αργή είναι και η κατασκευή. Για αυτό το λόγο θέλουμε η διαδικασία να είναι όσο το δυνατόν πιο παραμετροποιημένη. Στην παρούσα εργασία, θα εξετάσουμε προβλέψεις για μια και έξι ώρες μετά, για τις οποίες οι γράφοι των συναλλαγών θα είναι εξάωρης και τρίωρης διάρκειας, δηλαδή οι συναλλαγές οι οποίες ανήκουν στον ίδιο γράφο θα έχουν χρονική απόκλιση η μία από την άλλη το μέγιστο έξι και τρεις ώρες αντίστοιχα. Επιπλέον για τις προβλέψεις της τρίωρης διάρκειας γράφων, οι κόμβοι τους θα έχουν είτε μόνο δομικά χαρακτηριστικά των συναλλαγών και τη τιμή του Bitcoin είτε και επιπλέον δύο οικονομικά εξωτερικά χαρακτηριστικά. Για να μην κατασκευάζουμε όλους του γράφους training, validation και test έξι φορές (όσα και τα πιθανά σενάρια τα οποία εξετάζουμε), θα αποδώσουμε στους κόμβους όλα τα πιθανά χαρακτηριστικά που θα χρησιμοποιήσουμε, και για την περίπτωση που εξετάζουμε μόνο τα δομικά, θα αφαιρούμε τα οικονομικά χαρακτηριστικά από τους κόμβους. Θα κατασκευάσουμε εξ αρχής εξάωρους γράφους, από τους οποίους, για την περίπτωση των τρίωρων γράφων, θα αφαιρούμε τους κόμβους που ανήκουν στις τρεις πρώτες ώρες και θα βάλουμε σε κάθε γράφο δύο ετικέτες, της τιμής του

bitcoin για την επόμενη ώρα και για έξι ώρες μετά, όπου ανάλογα τη περίπτωση που εξετάζουμε, θα επιλέγουμε τη μία από τις δύο.

Επομένως συγκεντρωτικά τα χαρακτηριστικά που θα χρησιμοποιηθούν στα πειράματά μας είναι τα εξής:

Από τα δομικά χαρακτηριστικά των συναλλαγών που εμφανίζονται στο blockchain:

- size
- block timestamp
- input count
- output count
- input value
- fee
- value (το εμφωλευμένο από το πεδίο inputs)

Από τα εξωτερικά του blockchain χαρακτηριστικά:

- Bitcoin Close Price
- Volume USD
- 10-day EMA

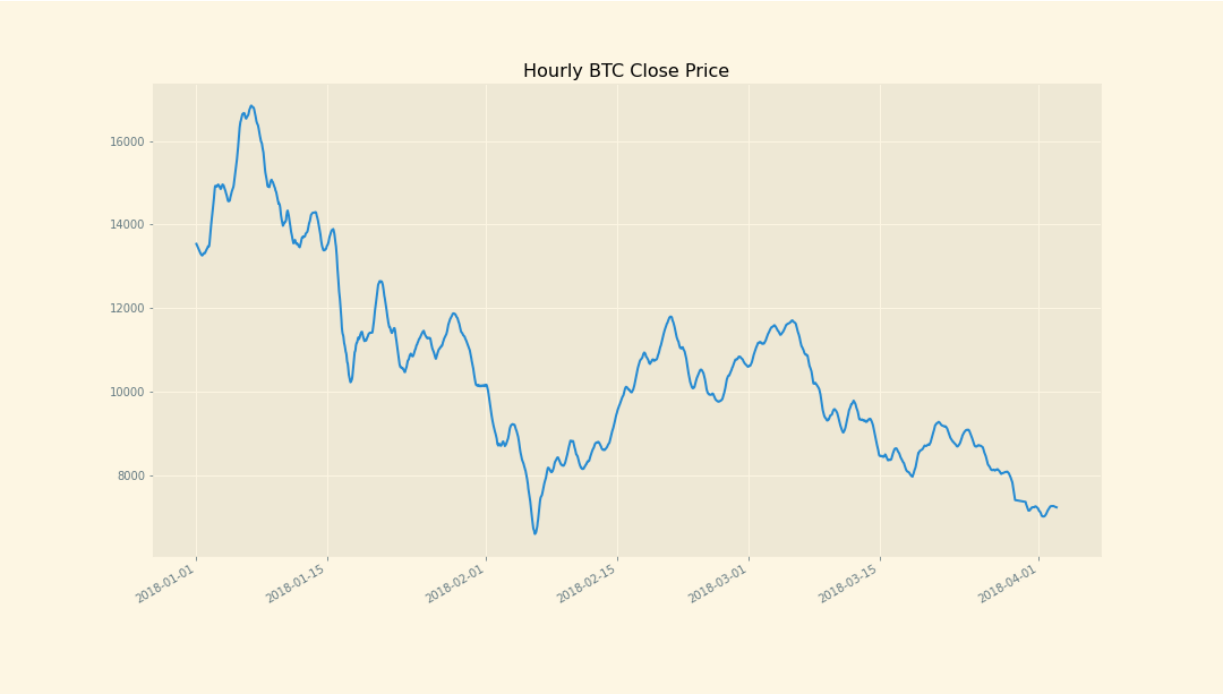
Το input count και το output count αποτελούν στην πραγματικότητα το βαθμό εισόδου και το βαθμό εξόδου της κάθε συναλλαγής (κόμβου). Το value το οποίο είναι το ποσό που συνεισέφερε η εκάστοτε συναλλαγή εισόδου στη παρούσα, μπαίνει ως βάρος στις ακμές μεταξύ αυτών των συναλλαγών. Το άθροισμα των βαρών των ακμών είναι το χαρακτηριστικό input value του κόμβου. Το block timestamp είναι η χρονοσφραγίδα της συναλλαγής και χρησιμοποιείται για δύο λόγους. Ο ένας είναι για να αποδώσουμε ως χαρακτηριστικά της κάθε συναλλαγής ανάλογα το πότε πραγματοποιήθηκε χρονικά τη τιμή του Bitcoin εκείνη τη χρονική στιγμή (Bitcoin Close Price) καθώς και το Volume USD, το οποίο είναι το συνολικό ποσό που ανταλλάχθηκε εκείνη τη χρονική στιγμή για Bitcoins. Ο δεύτερος λόγος, είναι για να ορίσουμε το χρονικό διάστημα διαρκείας του κάθε γράφου. Αφού επιτευχθούν τα παραπάνω, η χρονοσφραγίδα αφαιρείται από τον κάθε κόμβο.

Όπως έχει αναφερθεί, τα Bitcoin Close Price και Volume USD δίνονται ανά ώρα, οπότε οι συναλλαγές που είναι μέχρι και μια ώρα μετά πιο κοντά χρονικά σε κάποια συγκεκριμένη ώρα θα έχουν ίδιες τιμές σε αυτά τα χαρακτηριστικά. Το Bitcoin Close Price χρησιμοποιείται τόσο για τις ετικέτες για τις προβλέψεις στους γράφους για τις μελλοντικές τιμές του BTC, και οι παρελθοντικές τιμές του σαν χαρακτηριστικό για τις συναλλαγές των γράφων που έχουν πραγματοποιηθεί. Για τη δεύτερη περίπτωση, στα δεδομένα εκπαίδευσης, πριν αποδοθεί ως χαρακτηριστικό στους κόμβους, πρέπει να προεπεξεργαστεί κατάλληλα με τον τρόπο που θα δούμε παρακάτω.

7.2.1: Μείωση Θορύβου Χρονοσειράς Bitcoin

Η χρονοσειρά της τιμής του Bitcoin, όπως και των υπόλοιπων κρυπτονομισμάτων, χαρακτηρίζεται από ακραίες διακυμάνσεις στις τιμές της σε μικρά χρονικά διαστήματα. Αυτού του είδους οι διακυμάνσεις, δεν χαρακτηρίζουν την γενική πορεία της τιμής και πολλές φορές δυσκολεύουν τους αλγορίθμους μηχανικής μάθησης στο να εκπαιδευτούν και να αναγνωρίσουν μοτίβα στη πρόβλεψη της τιμής. Για αυτό, χρειάζονται να χρησιμοποιηθούν μέθοδοι οι οποίες εξομαλύνουν τη χρονοσειρά έτσι ώστε να αποτυπώνεται με ευκολία η πληροφορία που είναι πραγματικά απαραίτητη για τη πρόβλεψη [28]. Βέβαια, τέτοιες μέθοδοι έχουν χρησιμοποιηθεί σε μοντέλα μηχανικής μάθησης τα οποία παίρνουν ως είσοδο τα δεδομένα της χρονοσειράς αυτούσια. Στα GNN που χρησιμοποιούμε εμείς, αυτή η πληροφορία δίνεται έμμεσα, ως ένα από τα χαρακτηριστικά των κόμβων. Εν τούτοις θεωρούμε πως και στη δική μας περίπτωση αυτό είναι κάτι το οποίο μόνο να βελτιώσει μπορεί τη συμπεριφορά του μοντέλου. Για αυτό το λόγο, χρησιμοποιήσαμε τη μέθοδο Savitzky – Golay [29], η οποία χρησιμοποιήθηκε στους Πολίτης κ.α. [27] για την εξομάλυνση της χρονοσειράς του Ether σε πρόβλημα πρόβλεψης της τιμής του. Η μέθοδος Savitzky – Golay χρησιμοποιεί τη συνέλιξη για την προσαρμογή διαδοχικών υποσυνόλων γειτονικών σημείων δεδομένων σε ένα πολυώνυμο χαμηλού βαθμού με τη μέθοδο των γραμμικών ελαχίστων τετραγώνων. Στη δική μας περίπτωση θεωρήσαμε το βαθμό της πολυωνυμικής συνάρτησης ίσο με ένα, δηλαδή μια συνάρτηση γραμμικής παλινδρόμησης με το πλήθος των δεδομένων εισόδου κάθε φορά σε αυτήν ίσο με έντεκα. Η χρονοσειρά του Bitcoin πριν και μετά τη χρήση του φίλτρου αυτού φαίνεται στις παρακάτω εικόνες.

Εικόνα 16: Χρονοσειρά του Bitcoin (σε USD) από 01/01/2018 έως και 02/04/2018



Εικόνα 17: Χρονοσειρά του Bitcoin (σε USD) από 01/01/2018 έως και 02/04/2018 μετά την εξομάλυνση με χρήση του φίλτρου Savitzky – Golay.

Τελικά, οι κόμβοι στους γράφους συναλλαγών εκπαίδευσης θα πάρουν την κοντινότερη χρονικά εξομαλυμένη τιμή της τιμής του Bitcoin σε αυτούς, ως ένα από τα χαρακτηριστικά τους, ενώ οι κόμβοι στο σύνολο επικύρωσης (validation) και ελέγχου (test) την μη εξομαλυμένη.

7.2.2: Κυλιόμενος Εκθετικός Μέσος Όρος Δέκα Ημερών της Τιμής του Bitcoin (10-day Exponential Moving Average-EMA)

Ο κυλιόμενος εκθετικός μέσος όρος (EMA) είναι ένας τύπος κυλιόμενου μέσου όρου (ΚΜ) που δίνει μεγαλύτερη βαρύτητα στα πιο πρόσφατα σημεία δεδομένων και άρα αντιδρά πιο σημαντικά και ταχύτερα στις πρόσφατες μεταβολές των τιμών από ότι ένας κυλιόμενος απλός μέσος όρος (SMA), ο οποίος αποδίδει ίδια βαρύτητα σε όλες τις παρατηρήσεις της περιόδου [30]. Όπως όλοι οι κυλιόμενοι μέσοι όροι, αυτός ο τεχνικός δείκτης χρησιμοποιείται για την παραγωγή σημάτων αγοράς και πώλησης με βάση τις διασταυρώσεις και τις αποκλίσεις από τον ιστορικό μέσο όρο. Ο υπολογισμός του βασίζεται μόνο σε παρελθοντικές τιμές της χρονοσειράς, και το πόσο μακροχρόνια ή βραχυπρόθεσμα συμπεράσματα θέλουμε να εξάγουμε με βάση τον EMA εξαρτάται από το πόσες παρελθοντικές τιμές θα συμπεριλάβουμε στον υπολογισμό του. Για μακροχρόνιες

προβλέψεις συνήθως χρησιμοποιούνται δεδομένα από πενήντα μέχρι και διακοσίων ημερών. Για πιο βραχυπρόθεσμες προβλέψεις (π.χ. της επόμενης ημέρας) συνήθως χρησιμοποιούνται δεδομένα από δέκα μέχρι και είκοσι ημέρες. Καθώς έχει διαπιστωθεί η χρησιμότητα του για προβλέψεις τιμών τόσο σε μετοχές [31] όσο και κρυπτονομίσματα [27], και οι προβλέψεις μας είναι βραχυπρόθεσμες (μία και έξι ώρες μετά), επιλέξαμε να χρησιμοποιήσουμε των 10-ημερών (240 ωρών) EMA, ο οποίος φαίνεται στην παρακάτω εικόνα:

Εικόνα 18: Ο 240 ωρών (10 ημερών) EMA. Έχουν συμπεριληφθεί στη γραφική παράσταση της τιμής του BTC και οι δέκα προηγούμενες μέρες της 01/01/2018 που χρησιμοποιούνται για τον υπολογισμό του EMA.

Εφόσον ο κυλιόμενος εκθετικός μέσος όρος αποτελεί δημοφιλή δείκτη, επιλέξαμε να προσθέσουμε, ως εξωτερικό χαρακτηριστικό των κόμβων στους γράφους συναλλαγών, την κοντινότερη χρονικά σε αυτούς τιμή του.

7.3: Χωρισμός Δεδομένων σε Σύνολα Εκπαίδευσης (train), Επικύρωσης (validation) και Ελέγχου (test)

Προσεγγίζουμε το πρόβλημα παλινδρόμησης το οποίο μελετάμε επαγωγικά. Θα χωρίσουμε τα δεδομένα μας, όπως σε κάθε πρόβλημα μηχανικής μάθησης σε δεδομένα εκπαίδευσης, επικύρωσης και ελέγχου, από τα οποία θα προκύψουν οι γράφοι εκπαίδευσης, επικύρωσης και ελέγχου, οι οποίοι είναι ανεξάρτητοι και διαφορετικοί μεταξύ τους. Τα GNN θα εκπαιδευτούν στους γράφους εκπαίδευσης, στους γράφους επικύρωσης θα εντοπιστούν οι

κατάλληλοι υπερπαράμετροι τους και στους γράφους ελέγχου θα καταγραφεί η τελική απόδοση του καλύτερου μοντέλου και η ικανότητα του να γενικεύει τις προβλέψεις του σε γράφους που δεν έχει ξανασυναντήσει. Εφόσον ασχολούμαστε με τη πρόβλεψη τιμών χρονοσειράς, ο χωρισμός στα τρία αυτά σύνολα πρέπει να διατηρεί χρονολογική σειρά.

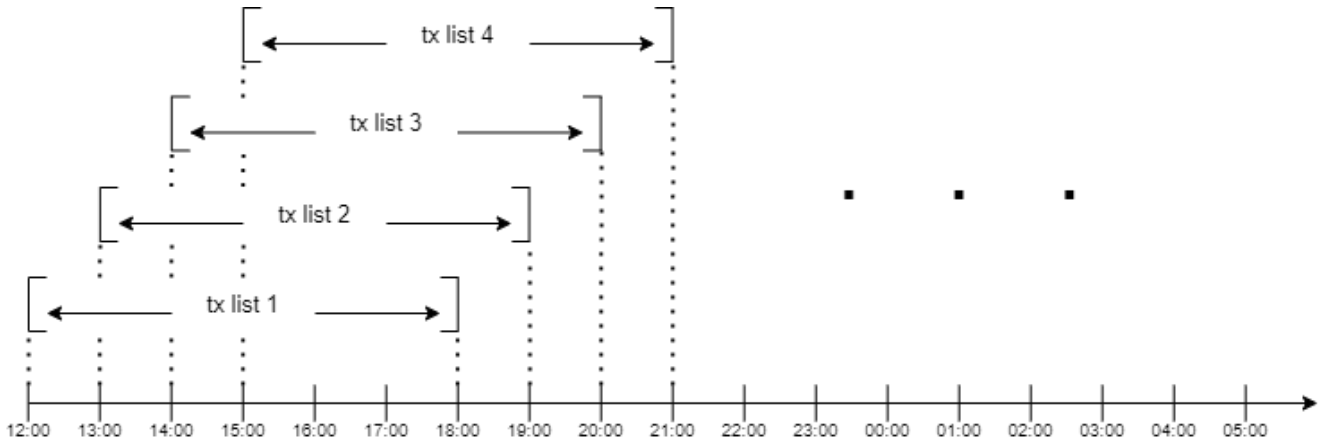
Τα δεδομένα εκπαίδευσης (80% των συνολικών δεδομένων) είναι από 01/01/2018 μέχρι και 02/04/2018 από τα οποία προέκυψαν 2203 γράφοι.
Τα δεδομένα επικύρωσης (10% των συνολικών δεδομένων) είναι από 03/04/2018 μέχρι και 14/04/2018 από τα οποία προέκυψαν 283 γράφοι.
Τα δεδομένα ελέγχου (10% των συνολικών δεδομένων) είναι από 15/04/2018 μέχρι και 26/04/2018 από τα οποία προέκυψαν 283 γράφοι.

7.4: Κατασκευή Λιστών Συναλλαγών και Κυλιόμενου Παραθύρου (Sliding Window)

Ο αλγόριθμος κατασκευής των γράφων συναλλαγών, που παρουσιάζεται στο 7.5, παίρνει ως είσοδο λίστα συναλλαγών την οποία χρησιμοποιεί για να δημιουργήσει τον γράφο. Σε αυτή τη λίστα, η κάθε συναλλαγή περιέχει τις πληροφορίες που αναφέραμε στην ενότητα 7.1. Στην παρούσα ενότητα θα περιγράψουμε τη διαδικασία κατασκευής των λιστών αυτών, οι οποίες χρησιμοποιούνται για τη δημιουργία των γράφων.

Τα απαραίτητα στοιχεία για τις συναλλαγές, όπως αναφέρθηκε, δίνονται σε μορφή JSON. Καθώς θέλουμε να δημιουργήσουμε γράφους n ωρών (συγκεκριμένα n = 6), είναι απαραίτητο για τις ακόλουθες διαδικασίες οι συναλλαγές να είναι κατά αύξουσα σειρά ταξινομημένες, ως προς την ημερομηνία που πραγματοποιήθηκαν.
Επιπλέον, λόγω του μεγάλου πλήθους των συναλλαγών που πραγματοποιούνται στο Bitcoin, υπάρχει πληροφορία που δεν χρειάζεται να συμπεριληφθεί στους γράφους. Θεωρούμε ότι συναλλαγές οι οποίες έχουν πολύ μικρή ποσότητα Bitcoin να ανταλλάσσεται, δεν βοηθούν στην πρόβλεψη της τιμής του, καθώς θα αποτελέσουν κόμβους με μικρή “βαρύτητα” στο δίκτυο, δηλαδή το αν κάποιος θα αγοράσει πολύ μικρή ποσότητα Bitcoin ή θα πουλήσει δεν θα προκαλέσει την ίδια αυξομείωση στη τιμή του σε σχέση με κάποιον που έκανε κάποια συναλλαγή πολλές τάξεις μεγέθους παραπάνω από τον πρώτο. Ωστόσο, κάποιος θα μπορούσε να ισχυριστεί ότι πολλές μικρές συναλλαγές σε κοντινό χρονικό διάστημα μπορεί να έχουν επίδραση στη τιμή του Bitcoin. Για αυτό το λόγο, επιλέξαμε να αφαιρέσουμε μόνο το 25% των συνολικών συναλλαγών, οι οποίες είχαν τις μικρότερες τιμές του στοιχείου input_value. Αυτό επιλέχθηκε να συμβαίνει στη διαδικασία διαβάσματος του αρχείου JSON με τις πληροφορίες των συναλλαγών για την δημιουργία των λιστών, πριν τον αλγόριθμο κατασκευής των γράφων, έτσι ώστε για τη κάθε συναλλαγή να υπάρχει μικρότερο πλήθος συναλλαγών που χρειάζεται να ελεγχθεί για την ύπαρξη ακμών μεταξύ τους.

Όπως αναφέρθηκε, θέλουμε να δημιουργήσουμε εξάωρης διάρκειας γράφους (οι τρίωροι θα προκύψουν εκ των υστέρων με αφαίρεση των συναλλαγών των τριών πρώτων ωρών των εξάωρων γράφων). Αν για παράδειγμα θέλουμε να κάνουμε μια πρόβλεψη τιμής του Bitcoin για την επόμενη ώρα, π.χ. για σήμερα στις 19:00, αυτό σημαίνει ότι θα κατασκευαστεί μια λίστα με τις συναλλαγές από τις 12:00 μέχρι και τις 18:00, θα περάσει ως είσοδος στη συνάρτηση κατασκευής γράφων έτσι ώστε να δημιουργήσει τον γράφο του Blockchain του Bitcoin με τις συναλλαγές μεταξύ 12:00 και 18:00, και στη συνέχεια θα περαστεί ως είσοδος στο GNN για να μας βγάλει το αποτέλεσμα της πρόβλεψης του για το ποια θα είναι η τιμή στις 19:00. Για να συνεχιστούν οι προβλέψεις για τις επόμενες ώρες, θα επαναληφθεί η ίδια διαδικασία. Δηλαδή, για πρόβλεψη στις 20:00, θα κατασκευαστεί η λίστα με τις συναλλαγές του Bitcoin από τις 13:00 μέχρι τις 19:00 κ.ο.κ. Το ίδιο ισχύει και για τη πρόβλεψη π.χ. έξι ωρών μετά. Πιο συγκεκριμένα, αν έχουμε τη λίστα συναλλαγών (και κατ' επέκταση το γράφο, όπως εξηγήθηκε) από 12:00 μέχρι 18:00 και θέλουμε να προβλέψουμε για την ώρα 00:00 της επόμενης μέρας, για να κάνουμε πρόβλεψη για την επόμενη ώρα μετά (δηλαδή για τη 01:00 της επόμενης μέρας) θα χρησιμοποιήσουμε τον γράφο από τις 13:00 μέχρι τις 19:00 κ.ο.κ. (αφού έχουμε δεδομένα έξι ωρών πριν τα οποία ανανεώνονται ανά ώρα και χρησιμοποιούμε εξάωρους γράφους δεδομένων συναλλαγών). Αυτομάτως καταλαβαίνουμε ότι χρειάζεται να δημιουργούμε τις λίστες με έναν κυλιόμενο χρονικά παράθυρο (Sliding Window) σε επίπεδο ώρας, για να έχουμε κάθε φορά τους πιο σύγχρονους n-ωρών γράφους. Έτσι, για τα ιστορικά δεδομένα, οι λίστες με τις συναλλαγές κατασκευάστηκαν με τέτοιο τρόπο, ώστε n-ωρών δεδομένων διαδοχικές λίστες να διαφέρουν κατά μια ώρα μεταξύ τους. Σχηματικά αυτό φαίνεται στην εικόνα 19.



Εικόνα 19: Δημιουργία λιστών συναλλαγών με την τεχνική του κυλιόμενου παραθύρου (sliding window).

7.5: Αλγόριθμος Κατασκευής Γράφων Συναλλαγών

Ο αλγόριθμος τον οποίο επιλέξαμε να υλοποιήσουμε για την κατασκευή των n-ωρών γράφων συναλλαγών στηρίχθηκε στον αλγόριθμο των Tharani κ.α. [26]. Ωστόσο έχει παραμετροποιηθεί και τροποποιηθεί αρκετά, έτσι ώστε να δημιουργεί τους γράφους συναλλαγών που θέλουμε εμείς να κατασκευάσουμε για να είναι κατάλληλοι για το πρόβλημα που μελετάμε. Ο αλγόριθμος αυτός παρουσιάζεται παρακάτω και ακολουθεί επεξήγησή του.

Αλγόριθμος 1: Κατασκευάζοντας τους γράφους συναλλαγών

Είσοδος: Λίστα συναλλαγών
Έξοδος: Γράφος Συναλλαγών

graph_construction(txList)
create Directional Graph “tx_graph”
add $V_{virtual}$ to tx_graph
for i ∋ range(1,len(txList)) **do**
  tx_i ← txList[i]
  $hash_i$ ← hash of the tx_i
  $blockhash_i$ ← hash of the block tx_i belongs to
  $size_i$ ← size of tx_i
  $timestamp_i$ ← timestamp of the block tx_i belongs to
  $input\_count_i$ ← number of inputs that contribute BTC to tx_i
  $output\_count_i$ ← number of outputs that receive BTC from tx_i
  fee_i ← fee paid for tx_i
  volume_usd ← volume of bitcoin in usd transacted globally at the time tx_i took place
  ewm ←exponential moving average of bitcoin timeseries at the time tx_i took place
  usd_value ← value in usd of the Bitcoin at the time tx_i took place
 if tx_i not a coinbase tx and input_value of tx_i >= minimum_btc **then**
      total_input_value of tx_i ← input_value of txi (sum of the BTCs that input addresses contribute to tx_i)
      Add V_i to tx_graph
 else if tx_i is a coinbase tx and input_value of txi >= minimum_btc **then**
      total_input_value of tx_i ← BTC received by the miner that validates the block
      Add V_i to tx_graph
 else
      continue to tx_{i+1}
 end

```
    k ← i
    while  block_hash_i equals to block_hash_k-1 and k >=  1
        tx_k ← txList[k]
        inputs_k ← the nested informations of inputs
        hash_k ← hash of the tx_k
        for input ∋ inputs_k do:
            source_hash_k ← spent_transaction_hash of input
            source_value_k ← value of BTC of the input transaction of tx_k
            if source_hash_k equals to hash_i then:
                add V_k to tx_graph
                add Edge from V_i  to V_k with weight equal to source_value_k
            end
        end
        k ← k-1
     end
    for j ∋ range(i+1,len(tx_list)) do
        tx_j ← txList[j]
        inputs_j ← the nested informations of inputs
        hash_j ← hash of the tx_j
        for input ∋ inputs_j do:
            source_hash_j ← spent_transaction_hash of input
            source_value_j ← value of BTC of the input transaction of tx_j
            if source_hash_j equals to hash_i then
                add V_j to tx_graph
                add Edge from V_i  to V_j with weight equal to source_value_j
            end
        end
    end
    if out degree of tx_i equals to 0 then:
        add Edge from V_i to V_virtual with 0 weight
    end
end
set_label(tx_graph)
return tx_graph
```

Για κάθε μία από τις λίστες συναλλαγών όπως παρουσιάστηκαν στην ενότητα 7.4, εφαρμόζεται ο παραπάνω αλγόριθμος και προκύπτει ο παραμετροποιημένος γράφος όπως έχει περιγραφεί στην ενότητα 7.2. Θα περιγράψουμε τον τρόπο κατασκευής του. Για κάθε συναλλαγή μέσα στη λίστα, δημιουργείται ένας κόμβος με τα χαρακτηριστικά της. Το μοναδικό αναγνωριστικό της συναλλαγής είναι το hash της.

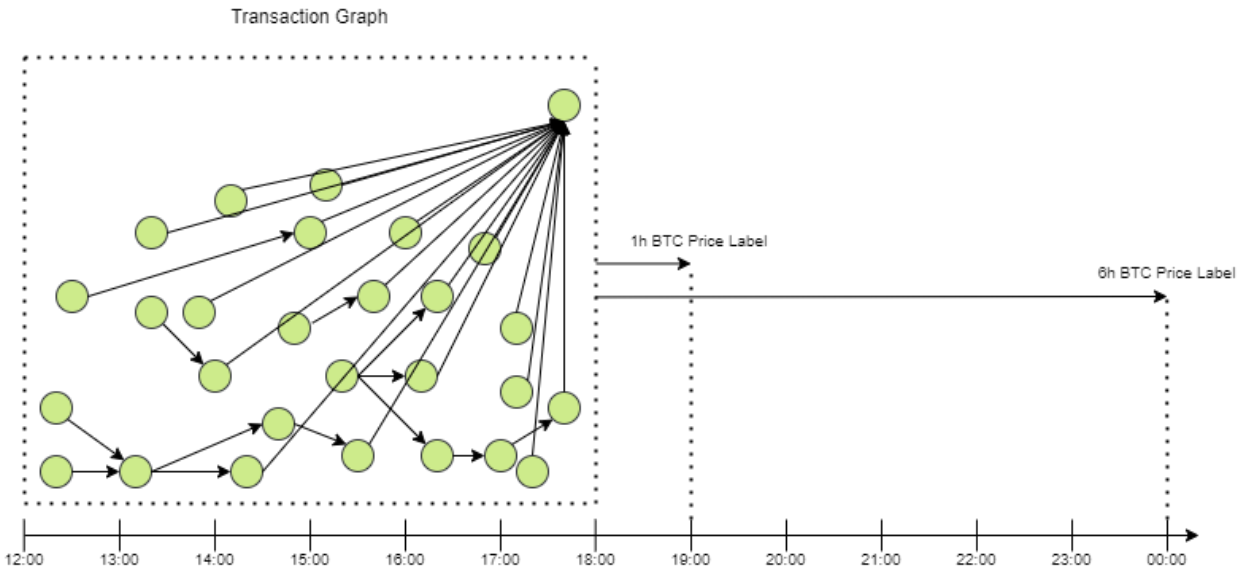
Πριν αποδώσουμε στον κόμβο το χαρακτηριστικό total_input_value, που είναι τα συνολικά Bitcoins που ανταλλάχθηκαν, πραγματοποιούμε τον έλεγχο για το αν μια συναλλαγή είναι coinbase ή όχι. Η πρώτη συναλλαγή ενός block είναι coinbase και είναι ένας ειδικός τύπος συναλλαγής που δημιουργεί ο miner του block για να λάβει την αμοιβή που του αντιστοιχεί, αφού έλυσε το κρυπτογραφικό πρόβλημα του proof-of-work μαζί με τις χρεώσεις των συναλλαγών που επικύρωσε. Θεωρούμε ότι τέτοιου είδους συναλλαγές περιέχουν χρήσιμη πληροφορία για τη πρόβλεψη μας και πρέπει να συμπεριληφθούν στο γράφο, καθώς και εισέρχονται στο δίκτυο καινούργια νομίσματα μεγάλης ποσότητας, και οι ίδιες οι συναλλαγές αυτές αποτελούν δομικό επαναλαμβανόμενο μοτίβο του blockchain. Όμως, το πεδίο του input τους είναι κενό, καθώς περιέχουν μόνο output το οποίο στέλνεται κατευθείαν στον miner του block. Καθώς για την κατασκευή του γράφου και τη προσθήκη ακμών μεταξύ των κόμβων, σχεδιαστικά επιλέξαμε να κοιτάμε τα hashes των inputs και τα value των inputs, μπορούμε εύκολα να θεωρήσουμε, χωρίς κάποιο πρόβλημα, ότι και τα coinbase transactions έχουν σαν input value το ίδιο το output τους από έναν κόμβο που δεν υπάρχει αλλά έχει την αμοιβή της coinbase συναλλαγής. Έτσι αν η συναλλαγή είναι όντως coinbase, απλά έχει ίδιο input και output. Αν δεν είναι, παίρνει το total_input_value που της αντιστοιχεί ούτως ή άλλως. Στον παραπάνω έλεγχο υπάρχει επίσης και η συνθήκη το value της συναλλαγής να είναι μεγαλύτερο από ένα κάτω όριο. Στην δική μας περίπτωση, αυτό είναι το 25% και το φιλτράρισμα αυτό προηγήθηκε στη δημιουργία των λιστών, ωστόσο, μπορεί να πραγματοποιηθεί και εδώ. Στη συνέχεια, ο κόμβος της συναλλαγής προστίθεται στον γράφο.

Η χρονοσφραγίδα που έλαβε η κάθε συναλλαγή, είναι η χρονοσφραγίδα του block στο οποίο ανήκει. Έτσι, συναλλαγές που πραγματοποιήθηκαν με κάποια διαφορά χρονικά αλλά παρόλα αυτά μπήκαν στο ίδιο block, θα έχουν ίδια χρονοσφραγίδα. Για αυτό το λόγο, πέρα από τις μεταγενέστερες χρονικά συναλλαγές, πρέπει να εξεταστούν και οι συναλλαγές που ανήκουν στο ίδιο block για το αν συνδέονται με τη συναλλαγή που εξετάζεται. Η διαδικασία είναι η εξής. Για κάθε συναλλαγή που ανήκει στο ίδιο block, κοιτάμε τα εμφωλευμένα πεδία inputs και συγκεκριμένα τα spet_transaction_hash που έχει, τα οποία είναι τα hashes των συναλλαγών που συμμετέχουν ως είσοδοι στη συναλλαγή. Αν εκεί βρεθεί το hash της αρχικής συναλλαγής που εξετάζουμε, σημαίνει ότι αυτή έχει συνεισφέρει στη συναλλαγή που βρίσκεται στο ίδιο block. Τότε, από την αρχική συναλλαγή δημιουργούμε μια κατευθυνόμενη ακμή προς τη συναλλαγή που είχε στα inputs της το hash της αρχικής. Αυτή η διαδικασία επαναλαμβάνεται για όλες τις συναλλαγές του ίδιου block της αρχικής και για όλες τις μεταγενέστερες χρονικά συναλλαγές.

Συνολικά για κάθε συναλλαγή δηλαδή, ελέγχουμε που εμφανίζεται ως είσοδος σε επόμενες (ή χρονικά ταυτόσημες) συναλλαγές. Όπου το hash της αποτελεί ένα από τα spent_transaction_hash άλλων συναλλαγών, σήμαινει ότι έχει συμμετέχει σε αυτές και άρα υπάρχει κατευθυνόμενη ακμή προς αυτές.

Όταν προστίθεται όμως ο κόμβος στον οποίο καταλήγει η ακμή, δεν περιέχει τα χαρακτηριστικά που θα έπρεπε, παρά μόνο το hash του. Αυτό δεν αποτελεί πρόβλημα, καθώς, όταν θα έρθει η επανάληψη όπου θα εξετάζεται αν αυτός αποτελεί είσοδο σε άλλες

συναλλαγές, αφού θα υπάρχει ήδη στον tx_graph μόνο με το hash του, αντί να προστεθεί, θα συμπληρωθούν τα επιπρόσθετα πεδία του με τα αντίστοιχα χαρακτηριστικά του.
Εφόσον εξετάζουμε συναλλαγές εντός κάποιου εξαώρου, πολλές από αυτές δεν θα αποτελούν είσοδο σε άλλες εντός του εξαώρου αυτού και έτσι θα υπάρχουν πολλοί κόμβοι οι οποίοι δεν θα είναι συνδεδεμένοι με το υπόλοιπο γράφημα. Για αυτό το λόγο, επιλέξαμε σχεδιαστικά όσοι κόμβοι δεν έχουν βαθμό εξόδου, δηλαδή τα “φύλλα” του γραφήματος, να συνδέονται προς έναν εικονικό κόμβο (virtual node). Αυτό έχει ως αποτέλεσμα ο γράφος να μετατρέπεται σε (χαλαρά) συνεκτικό.
Αυτή η διαδικασία επαναλαμβάνεται για κάθε συναλλαγή της λίστας, οπότε στο τέλος έχει δημιουργηθεί ένας κατευθυνόμενος ακυκλικός γράφος με βάρη όπου στη πορεία του χρόνου οι προηγούμενες συναλλαγές συνδέονται με κάποιες από τις επόμενες χρονικά. Δεν γίνεται μια μεταγενέστερη συναλλαγή να συμμετάσχει ως είσοδος σε μια χρονικά προηγούμενη της και ούτε να έχει ακμή στον εαυτό της. Το γεγονός ότι οι γράφοι που δημιουργούμε είναι ακυκλικοί είναι ένα σημαντικό πλεονέκτημα για τη χρήση τους στα GNNs, καθώς οι κύκλοι πολλές φορές μπορούν να δημιουργήσουν προβλήματα στο τρόπο επεξεργασίας της πληροφορίας από αυτά [49].
Αφού τελειώσει η προσθήκη κόμβων και ακμών στον γράφο, μέσω της συνάρτησης set_label(), βάζουμε δύο ετικέτες στον γράφο, τη τιμή του Bitcoin για την επόμενη ώρα και για έξι ώρες μετά.
Αφού ολοκληρωθεί ο αλγόριθμος, ένας γράφος συναλλαγών θα μοιάζει σχηματικά όπως στην εικόνα 20.

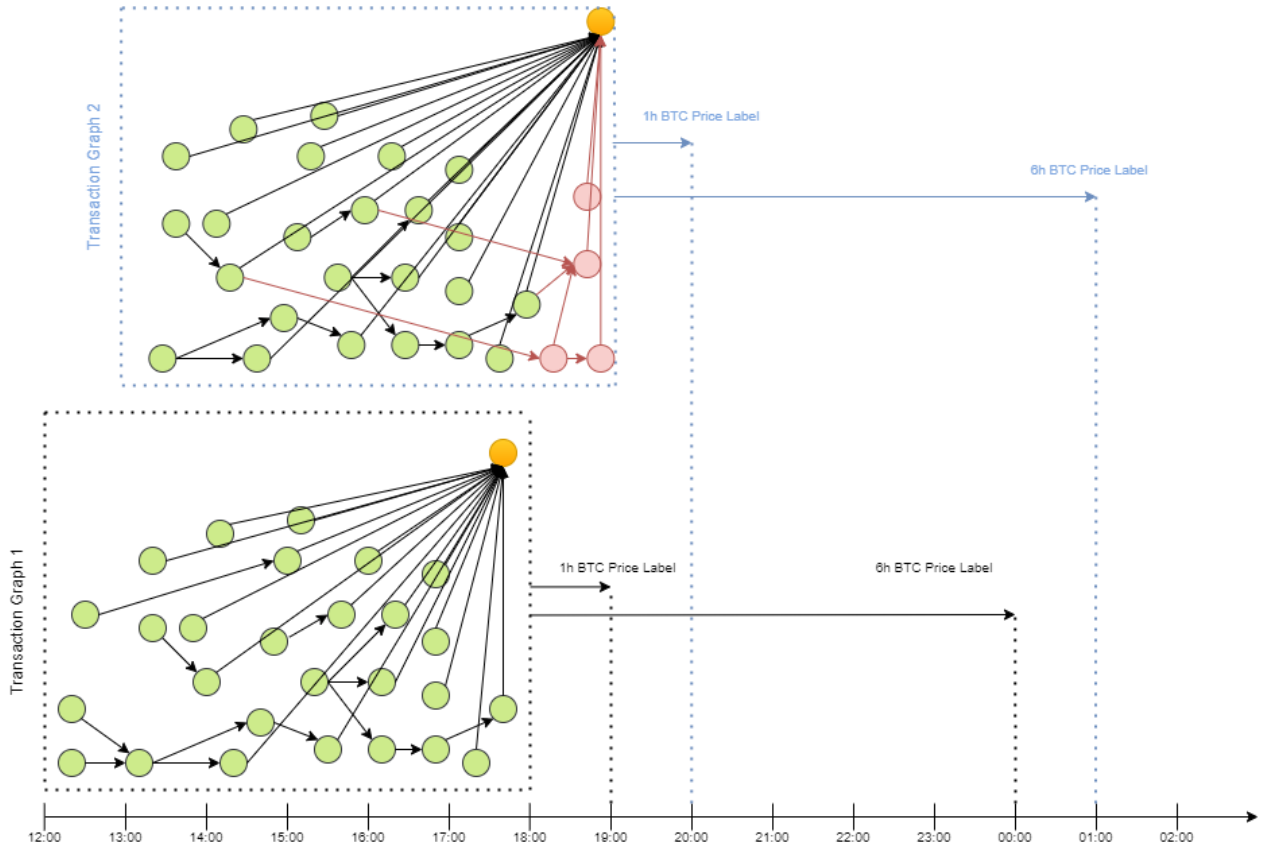


Εικόνα 20: Ενδεικτική σχηματική αναπαράσταση ενός γράφου συναλλαγών όπως προκύπτει από τον αλγόριθμο κατασκευής.

Όταν ο αλγόριθμος αυτός εφαρμοστεί στην επόμενη λίστα συναλλαγών, θα προκύψει ο κατά μια ώρα κυλιόμενος στο χρόνο καινούργιος γράφος. Αυτός ο γράφος θα περιέχει όχι μόνο καινούργιους κόμβους, αυτούς της επόμενης ώρας και τις μεταξύ τους συνδέσεις, αλλά και συνδέσεις των κόμβων που αντιστοιχούσαν στον προηγούμενο γράφο με τους καινούργιους, οι οποίες δεν υπήρχαν πριν αφού δεν είχαν συμπεριληφθεί. Σχηματικά αυτό φαίνεται στην εικόνα 21.

Εφαρμόζοντας επαναληπτικά τον αλγόριθμο κατασκευής στις λίστες συναλλαγών, προκύπτουν οι παραμετροποιημένοι γράφοι που αποτελούν το αρχικό σύνολο δεδομένων μας από το οποίο θα προκύψουν τα διάφορα σενάρια συνόλων δεδομένων για τις προβλέψεις της μίας και των έξι ωρών αντίστοιχα.

Στην εικόνα 22 φαίνονται για έναν τυχαίο γράφο εκπαίδευσης (πριν την επιλογή των παραμέτρων που θα καθορίσουν αν θα χρησιμοποιηθεί σε πρόβλεψη μίας ή έξι ωρών μετά, αν οι συναλλαγές θα έχουν όλα τα χαρακτηριστικά ή μόνο τα δομικά κ.λ.π.) μερικές πληροφορίες του.



Εικόνα 21: Δημιουργία γράφων συναλλαγών του Bitcoin χρησιμοποιώντας τις λίστες συναλλαγών που προέκυψαν μέσω της τεχνικής του κυλιόμενου παραθύρου.

```
train_graph_2018-02-16 22_00_00 is a DiGraph with 28070 nodes and 34681 edges

Graph labels of train_graph_2018-02-16 22_00_00:  {'Label_1': 10596.4, 'Label_6': 10687.84}

Some random transactions of the graph:
[('7dad481ffc890cac46c84ff0858cd2ef805f8fdd8d252c294648438f72a6b5fc', {'size': 510, 'no_of_inputs': 1, 'no_of_output
s': 5, 'total_btc_input': 0.07460366, 'fee': 2.236e-05, 'tx_timestamp': 1518818501, 'usd_worth': 10234.29, 'usd_volum
e': 3007372.07, 'ewm': 9108.599910737674})]
[('d4756351bef248cc520e77462c268f7154af42ca8ae55e00aa779e46f7345611', {'size': 226, 'no_of_inputs': 1, 'no_of_output
s': 2, 'total_btc_input': 0.5207641000000001, 'fee': 9.04e-05, 'tx_timestamp': 1518818501, 'usd_worth': 10234.29, 'us
d_volume': 3007372.07, 'ewm': 9108.599910737674})]
[('33a932487035a774d4324e17545ea4d8f43f3932346d999584b02ed9f2e6412e', {'size': 225, 'no_of_inputs': 1, 'no_of_output
s': 2, 'total_btc_input': 0.8905124400000001, 'fee': 0.00025297, 'tx_timestamp': 1518837119, 'usd_worth': 10473.24866
6666668, 'usd_volume': 2220277.18, 'ewm': 9158.672897695693})]

Some random edges of the graph:
[('e6a5b57d4c2b08c671ff1de36ebc147055ea11ad2f1d92331b476ea087db9311', 'e453529d530b57bea3d0dac4445f91ca531f7e2700c618
a0f54afa8bc8c2ea72', {'weight': 0.209})]
[('d12e732cff1d5024efe4c95ce47d2b97dc8a1ddb1a7ff35d00a9a9946e0681c3', '5277fea42d3b76b2e6cd4991a512201ed608b0778e408a
0d77caebc0dd73dc82', {'weight': 0.99998418})]
[('2f2fd9502294ffd8e270f5eb401d05a60f93521959f2e8705876e09d3128de53', 'virtual', {'weight': 0})]
```

Εικόνα 22: Μερικές πληροφορίες για έναν τυχαίο εξάωρο γράφο συναλλαγών που περιλαμβάνει τις συναλλαγές από 2018-02-16 22_00_00 μέχρι και 2018-02-17 04_00_00.

7.6: Κανονικοποίηση Χαρακτηριστικών των Γράφων

Στα πεδία της μηχανικής μάθησης και της εξόρυξης δεδομένων, η κανονικοποίηση των χαρακτηριστικών αποτελεί διαδικασία προεπεξεργασίας δεδομένων, της οποίας στόχος είναι να τα μετασχηματίσει και να τα μεταφέρει σε μικρότερα διαστήματα τιμών έτσι ώστε να έχουν κατάλληλη μορφή για τους αλγορίθμους μηχανικής μάθησης [32]. Μέσω αυτής της τεχνικής, η διαδικασία μάθησης γίνεται πιο σταθερή και συγκλίνει πολύ πιο γρήγορα [33], με αποτέλεσμα τα μοντέλα να έχουν καλύτερες επιδόσεις σε σχέση με τους αλγορίθμους που δεν λαμβάνουν προεπεξεργασμένα δεδομένα.
Υπάρχουν δύο βασικές τεχνικές κανονικοποίησης των δεδομένων, ο min-max scaler και ο standard scaler που δίνονται από τους τύπους (34) και (35) αντίστοιχα.

$$x_{new} = \frac{x_{min}}{x_{max} - x_{min}} \quad (34)$$

$$x_{new} = \frac{x - \mu}{\sigma} \quad (35)$$

Όπου οι τιμές των x_{min}, x_{max}, σ και μ προκύπτουν από τα δεδομένα εκπαίδευσης. Όπως φαίνεται και από τον τύπο του min-max scaler, τα δεδομένα μετασχηματίζονται στο διάστημα [0,1], ενώ στον standard scaler τα δεδομένα μετασχηματίζονται με τέτοιο τρόπο ώστε να έχουν μηδενική μέση τιμή και διακύμανση ίση με ένα. Το μειονέκτημα του min-max scaler είναι ότι είναι εξαιρετικά ευαίσθητος σε ακραίες τιμές. Ο standard scaler είναι και αυτός ευαίσθητος σε ακραίες τιμές και επίσης δεν αποτελεί καλή επιλογή για δεδομένα που δεν ακολουθούν κανονική κατανομή [34].

Με βάση τη στατιστική ανάλυση που πραγματοποιήσαμε στα χαρακτηριστικά των δεδομένων εκπαίδευσης, διαπιστώθηκε ότι και παρουσιάζουν ακραίες τιμές και δεν ακολουθούν κανονική κατανομή. Επιπλέον, δεν πρέπει να ακολουθηθεί κάποια διαδικασία αφαίρεσης ή ειδικής μεταχείρισης των ακραίων τιμών στα δεδομένα μας καθώς δεν είναι αποτέλεσμα λαθών στη διαχείριση των δεδομένων αλλά πρόκεινται για φυσικές ακραίες τιμές (natural outliers). Αυτό επαληθεύει την διαίσθησή μας καθώς ειδικά για τις ακραίες μεγάλες τιμές χαρακτηριστικών των συναλλαγών όπως τα bitcoins που ανταλλάχθηκαν, ο βαθμός εισόδου και εξόδου των συναλλαγών κ.λπ. θεωρούμε ότι θα καθορίζουν περισσότερο την τιμή του Bitcoin από ότι συναλλαγές με μικρότερες τιμές σε αυτά τα χαρακτηριστικά.

Κρίνεται έτσι απαραίτητη η χρήση κάποιου scaler που να έχει μεγαλύτερη ανοχή στις ακραίες τιμές και να μην επηρεάζεται από το τι κατανομή ακολουθούν τα δεδομένα. Αυτές τις προϋποθέσεις τις πληρεί ο robust scaler, του ο οποίου ο μετασχηματισμός δίνεται από την εξίσωση (36):

$$x_{new} = \frac{x - Q_2}{Q_3 - Q_1} \quad (36)$$

όπου Q_2 = μ, δηλαδή είναι η μέση τιμή, Q_1 είναι το 25^o ποσοστιαίο σημείο (percentile), δηλαδή η τιμή από την οποία και κάτω βρίσκεται το 25% των δεδομένων, Q_3 είναι το 75^o ποσοστιαίο σημείο, δηλαδή η τιμή από την οποία και κάτω βρίσκεται το 75% των δεδομένων (από εκείνη και πάνω βρίσκεται το 25% των δεδομένων) και η διαφορά τους $Q_3 - Q_1$ ονομάζεται interquartile range (IQR).

Μέσω αυτού του μετασχηματισμού εύκολα συμπεραίνουμε ότι τα δεδομένα δεν επηρεάζονται από τις ακραίες τιμές σε σχέση με τις προηγούμενες τεχνικές, καθώς η μέση τιμή και τα ποσοστιαία σημεία δεν παίρνουν τιμές από ακραίες τιμές όπως στον min-max scaler ούτε επηρεάζονται από το πως αυτοί επηρεάζουν την κατανομή και τη διακύμανση της. Έτσι κανονικοποιούνται και οι ακραίες τιμές που περιλαμβάνονται στα δεδομένα μικραίνοντας την απόσταση τους σε σχέση με τα υπόλοιπα δεδομένα.

Στα δεδομένα εκπαίδευσης, για κάθε χαρακτηριστικό που χρησιμοποιήθηκε συλλέχθηκαν οι τιμές των μέσων τιμών και του 25^{ου} και 75^{ου} ποσοστιαίου σημείου και στη συνέχεια, μέσω του τύπου (34), μετασχηματίστηκαν τα δεδομένα κάθε γράφου. Έτσι για παράδειγμα οι πληροφορίες της εικόνας (22) μετά την κανονικοποίηση φαίνονται στην εικόνα (23). Προφανώς η χρονοσφραγίδα δεν κανονικοποιείται.

```
train_graph_2018-02-16 22_00_00 is a DiGraph with 28070 nodes and 34681 edges

Graph labels of train_graph_2018-02-16 22_00_00:  {'Label_1': 0.11398403529230931, 'Label_6': 0.14704887948724885}

Some random transactions of the graph:
[('7dad481ffc890cac46c84ff0858cd2ef805f8fdd8d252c294648438f72a6b5fc', {'size': 1.9319727891156462, 'no_of_inputs': 0.
0, 'no_of_outputs': 3.0, 'total_btc_input': -0.05294813164519074, 'fee': -0.3089548986854528, 'tx_timestamp': 1518818
501, 'usd_worth': -0.016955495891302356, 'usd_volume': 0.5479431047307841, 'ewm': -0.4329948452813385})]
[('d4756351bef248cc520e77462c268f7154af42ca8ae55e00aa779e46f7345611', {'size': 0.0, 'no_of_inputs': 0.0, 'no_of_outpu
ts': 0.0, 'total_btc_input': 0.4669238931078524, 'fee': -0.2383615367855327, 'tx_timestamp': 1518818501, 'usd_worth':
-0.016955495891302356, 'usd_volume': 0.5479431047307841, 'ewm': -0.4329948452813385})]
[('33a932487035a774d4324e17545ea4d8f43f3932346d999584b02ed9f2e6412e', {'size': -0.006802721088435374, 'no_of_inputs':
0.0, 'no_of_outputs': 0.0, 'total_btc_input': 0.8977595318858592, 'fee': -0.06969071309255784, 'tx_timestamp': 151883
7119, 'usd_worth': 0.06945232434452228, 'usd_volume': -0.20968094464386938, 'ewm': -0.4161943254730869})]

Some random edges of the graph:
[('e6a5b57d4c2b08c671ff1de36ebc147055ea11ad2f1d92331b476ea087db9311', 'e453529d530b57bea3d0dac4445f91ca531f7e2700c618
a0f54afa8bc8c2ea72', {'weight': 0.10365225496729867})]
[('d12e732cff1d5024efe4c95ce47d2b97dc8a1ddb1a7ff35d00a9a9946e0681c3', '5277fea42d3b76b2e6cd4991a512201ed608b0778e408a
0d77caebc0dd73dc82', {'weight': 1.025317446090531})]
[('2f2fd9502294ffd8e270f5eb401d05a60f93521959f2e8705876e09d3128de53', 'virtual', {'weight': -0.1398772994983236})]
```

Εικόνα 23: Οι πληροφορίες της εικόνας (22) μετά την κανονικοποίηση των χαρακτηριστικών του γράφου.

7.7: Εκδοχές Γράφων Συναλλαγών (Σεναρίων) για τις Προβλέψεις.

Μέχρι στιγμής έχουν κατασκευαστεί μέσω του αλγορίθμου 1, εξάωροι γράφοι με οκτώ χαρακτηριστικά (και την χρονοσφραγίδα) στους κόμβους τους και δύο ετικέτες, οι οποίες είναι οι τιμές του Bitcoin για μία και έξι ώρες μετά τον εκάστοτε γράφο. Όπως έχει αναφερθεί, παραμετροποιήσαμε τους γράφους με τέτοιο ώστε να έχουν όλα τα δυνατά δεδομένα εξ αρχής για να μη χρειαστεί όταν εξετάζουμε διάφορα σενάρια να κατασκευάζουμε ξανά γράφους μέσω του αλγορίθμου 1.

Θεωρούμε ότι το είδος των γράφων συναλλαγών, το οποίο καθορίζεται από τη διάρκεια τους, και το τι είδος χαρακτηριστικών έχουν οι κόμβοι τους, επηρεάζει τα αποτελέσματα των προβλέψεών μας. Για αυτό το λόγο, μέσω των αρχικών μας γράφων, δημιουργούμε καινούργια σύνολα δεδομένων γράφων (εφόσον προκύπτουν από τους αρχικούς ο χωρισμός σε σύνολα εκπαίδευσης επικύρωσης και ελέγχου είναι ο ίδιος για κάθε περίπτωση) τα οποία μπορούν να ταξινομηθούν στα σενάρια όπως φαίνονται από τον πίνακα 4:

Σενάριο	Διάστημα Διάρκειας Γράφων (ώρες)	Actual value is negative	Χρονικός Ορίζοντας Πρόβλεψης
1A	6	Δομικά + Τιμή BTC	1
2A	3	Δομικά + Τιμή BTC	1
3A	3	Δομικά + Τιμή BTC + Οικονομικά	1

1B	6	Δομικά + Τιμή BTC	6
2B	3	Δομικά + Τιμή BTC	6
3B	3	Δομικά + Τιμή BTC + Οικονομικά	6

Πίνακας 4: Σενάρια Γράφων Συναλλαγών

7.8: Κατ' Εξέτασην Αρχιτεκτονικές GNN και Υπερπαράμετροι για τις Προβλέψεις

Τα αποτελέσματα που θα προκύψουν από τα GNNs που θα χρησιμοποιηθούν δεν εξαρτώνται μόνο από τα δεδομένα (σενάρια γράφων) που μπαίνουν ως είσοδοι σε αυτά καθώς και τη προεπεξεργασία τους, αλλά φυσικά και σε σημαντικό βαθμό από την ίδια την αρχιτεκτονική των μοντέλων και τις υπερπαραμέτρους τους. Για την εύρεση του βέλτιστου μοντέλου για τη πρόβλεψη της μίας και των έξι ωρών αντίστοιχα, θα εξετάσουμε για κάθε σενάριο κάθε πιθανό συνδυασμό των τιμών που επιλέχθηκαν να αποτελούν το σύνολο αναζήτησης εξαντλητικά, τόσο για τις αρχιτεκτονικές επιλογές των GNN όσο και των υπερπαραμέτρων τους (grid search). Για συντομία όταν χρησιμοποιούμε τη λέξη υπερπαράμετροι, θα αναφερόμαστε σε όλες τις παραμέτρους που λήφθηκαν υπ' όψιν μαζί και με τα διαφορετικά είδη συνόλων δεδομένων (σεναρίων). Κάθε μοντέλο που εξετάζεται, καθορίζεται από τις τιμές των υπερπαραμέτρων που αντιστοιχούν σε αυτό, και ουσιαστικά αποτελεί έναν από του πιθανούς συνδυασμούς του grid search. Οι υπερπαράμετροι καθώς και οι τιμές που εξετάζονται για αυτές φαίνονται στον πίνακα 5:

Υπερπαράμετροι	Σύνολο Αναζήτησης
Αρχιτεκτονική Επιπέδων	[SAGE, GATv2(head attention=2)]
Πλήθος Επιπέδων	[4, 5, 6]
Συνάρτηση Ενεργοποίησης	tanh
Hidden Dimensions ανά επίπεδο	[32, **64***]
Global Pooling Mechanism	CONCAT(global_mean_pool,global_max_pool)
Πλήθος Γραμμικών Επιπέδων	[1, 2]

Ρυθμός Εκπαίδευσης (Learning Rate)	[0.001, 0.0001]
Batch Size	[16, 32]
Optimizer	Adam
Dropout στα γραμμικά επίπεδα	0.3
Εποχές	50

Πίνακας 5: Σύνολο Αναζήτησης Υπερπαραμέτρων Μοντέλων

***** Η τιμή 64 για το πλήθος των κρυφών διαστάσεων (hidden dimensions) των επιπέδων εξετάζεται μόνο για τον τύπο επιπέδων SAGE, καθώς για το GATv2 η μνήμη RAM της GPU που χρησιμοποιήθηκε δεν ήταν αρκετή.

Με βάση τον παραπάνω πίνακα προκύπτει ότι οι συνολικοί συνδυασμοί του grid search που εξετάζονται για όλα τα σενάρια είναι για τα GNN τύπου SAGE: 6*3*1*2*1*2*2*2*1*1*1 = 288, ενώ για τα GNN τύπου GATv2 εφόσον δεν λαμβάνονται υπόψη τα μοντέλα με hidden dimension 64, είναι: 288/2 = 144. Οπότε συνολικά για την εύρεση του καλύτερου μοντέλου για τη πρόβλεψη της μίας ώρας και των έξι ωρών εκπαιδεύτηκαν συνολικά 432 μοντέλα και δοκιμάστηκαν στο σύνολο επικύρωσης, στο οποίο μετρήθηκε η απόδοση τους μέσω των μετρικών RMSE και MAPE και βρέθηκε σε κάθε περίπτωση το βέλτιστο μοντέλο. Αυτή η διαδικασία θα παρουσιαστεί αναλυτικά στις επόμενες ενότητες.

Αξίζει να σημειωθεί ότι στα νευρωνικά δίκτυα γράφων όταν χρησιμοποιούμε batches ουσιαστικά ενώνουμε τους γράφους που χρησιμοποιούνται στο batch σε έναν μεγάλο γράφο. Αυτό επιτυγχάνεται στοιβάζοντας διαγώνια τους πίνακες γειτνίασης των γράφων του batch και συνενώνοντας τα χαρακτηριστικά των κόμβων και τις ετικέτες του κάθε γράφου αντίστοιχα μεταξύ τους. Έτσι επιτυγχάνεται παραλληλοποίηση στη διαδικασία μάθησης στο εύρος των γράφων που εισάγονται στο batch και δεν επηρεάζει ο ένας γράφος τον άλλον. Επιπλέον, εξοικονομείται μνήμη καθώς οι πίνακες γειτνίασης των γράφων αποθηκεύονται σε αραιό πίνακα.

Επιπλέον, εφόσον μελετάμε πρόβλημα παλινδρόμησης σε επίπεδο γράφου, πρέπει τα embeddings των κόμβων ενός γράφου να καταλήξουν σε μια ενιαία αναπαράσταση. Αυτό γίνεται μέσω global pooling μηχανισμών. Συγκεκριμένα εδώ χρησιμοποιούμε το global_mean_pool το οποίο υπολογίζει τον μέσο όρο των embeddings των κόμβων και για να αυξήσουμε την εκφραστικότητα του pooling μηχανισμού, συνενώνεται με το global_max_pool το οποίο υπολογίζει το μέγιστο των embeddings των κόμβων. Τελικά όλη η εξαγώμενη πληροφορία του γράφου έχει γίνει embedded στο τελικό διάνυσμα που προέκυψε από αυτή τη διαδικασία. Το διάνυσμα αυτό στη συνέχεια τροφοδοτείται στα

γραμμικά επίπεδα, τα οποία το μετασχηματίζουν και μειώνουν τις διαστάσεις του σε μία, της οποίας η τιμή αποτελεί τη πρόβλεψη μας. Αναλυτική αναπαράσταση όλης της διαδικασίας του μοντέλου GNN για να υπολογίσει κάποια πρόβλεψη θα παρουσιαστεί στην ενότητα με τα τελικά αποτελέσματα στο σύνολο ελέγχου για τα καλύτερα μοντέλα για το κάθε είδος πρόβλεψης.

Κεφάλαιο 8: Πειραματικά αποτελέσματα για ωριαίες προβλέψεις

Στο κεφάλαιο αυτό, θα παρουσιαστούν πρώτα τα πειραματικά αποτελέσματα των διάφορων σεναρίων για την ωριαία πρόβλεψη στο validation set, θα επιλεχθεί το καλύτερο και θα δοκιμαστεί στο test set έτσι ώστε να λάβουμε την απόδοση του τελικού μοντέλου.

8.1: Αποτελέσματα Σεναρίου 1Α στους Γράφους Επικύρωσης

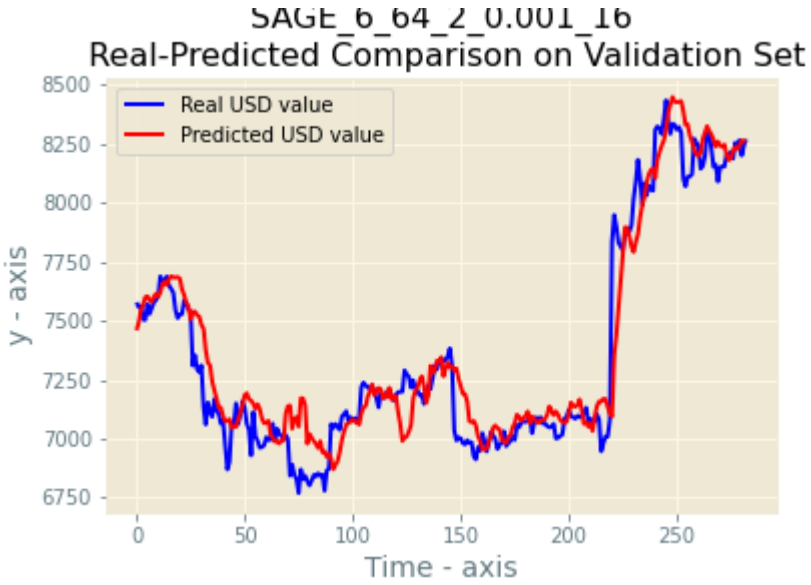
Το σενάριο 1Α έχει για σύνολο δεδομένων εξάωρης διάρκειας γράφους των οποίων οι κόμβοι περιέχουν μόνο τα δομικά χαρακτηριστικά των συναλλαγών (size, no_of_inputs, no_of_outputs, total_btc_input, fee) και τη τιμή του bitcoin (usd_worth).
Στον πίνακα 6 παρουσιάζεται ο συνδυασμός των υπερπαραμέτρων του μοντέλου με την καλύτερη απόδοση για το σενάριο 1Α (από τον πίνακα έχουν αφαιρεθεί σε σχέση με τον πίνακα 5 οι υπερπαράμετροι που είχαν μόνο μια τιμή αφού προφανώς αυτές οι επιλογές είναι σταθερές για όλα τα μοντέλα):

Υπερπαράμετροι	Σύνολο Αναζήτησης	Βέλτιστος Συνδυασμός
Αρχιτεκτονική Επιπέδων	[SAGE, GATv2(head attention=2)]	SAGE
Πλήθος Επιπέδων	[4, 5, 6]	6
Hidden Dimensions ανά επίπεδο	[32, 64]	64
Πλήθος Γραμμικών Επιπέδων	[1, 2]	2
Ρυθμός Εκπαίδευσης (Learning Rate)	[0.001, 0.0001]	0.001
Batch Size	[16, 32]	16

Πίνακας 6: Υπερπαράμετροι μοντέλων και βέλτιστος συνδυασμός για σενάριο 1Α

Συνοπτικά γράφουμε το μοντέλο με τις καλύτερες παραμέτρους ως SAGE_6_64_2_0.001_16 (οι καλύτερες υπερπαράμετροι με τη σειρά από πάνω προς τα κάτω).

Τα αποτελέσματα του μοντέλου στους γράφους επικύρωσης φαίνονται στην παρακάτω εικόνα:



Εικόνα 24: Προβλεπόμενη από το μοντέλο χρονοσειρά του BTC σε σχέση με την πραγματική για το σενάριο 1Α στο σύνολο επικύρωσης.

Για την απόδοση του μοντέλου χρησιμοποιήθηκαν οι μετρικές του μέσου τετραγωνικού ριζικού σφάλματος (RMSE) και του μέσου απολύτου ποσοστιαίου σφάλματος (MAPE) και τα αποτελέσματα τους παρουσιάζονται στον παρακάτω πίνακα:

Σενάριο	Μοντέλο	RMSE	MAPE
1Α	SAGE_6_64_2_0.001_16	134.057	1.274%

Πίνακας 7: Αποτελέσματα μετρικών RMSE και MAPE βέλτιστου μοντέλου στους γράφους επικύρωσης για το σενάριο 1Α

Το μοντέλο από όλα όσα εξετάστηκαν παρουσίασε τα καλύτερα αποτελέσματα σε σχέση με τα υπόλοιπα και στις δύο μετρικές.

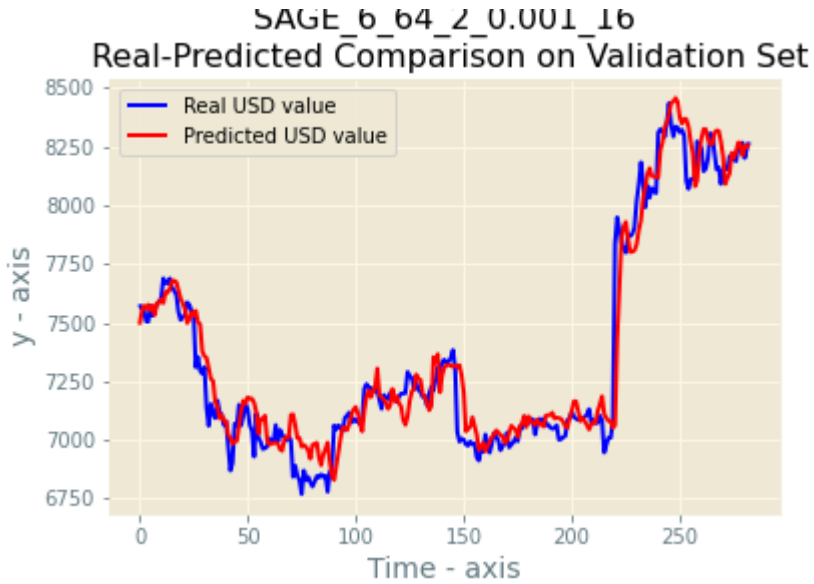
8.2: Αποτελέσματα Σεναρίου 2Α στους Γράφους Επικύρωσης

Το σενάριο 2Α έχει για σύνολο δεδομένων τρίωρης διάρκειας γράφους των οποίων οι κόμβοι περιέχουν μόνο τα δομικά χαρακτηριστικά των συναλλαγών (size, no_of_inputs, no_of_outputs, total_btc_input, fee) και τη τιμή του bitcoin (usd_worth).
Αντίστοιχα με το 8.1, στον πίνακα 8 παρουσιάζεται ο συνδυασμός των υπερπαραμέτρων του μοντέλου με την καλύτερη απόδοση για το σενάριο 2Α:

Υπερπαράμετροι	Σύνολο Αναζήτησης	Βέλτιστος Συνδυασμός
Αρχιτεκτονική Επιπέδων	[SAGE, GATv2(head attention=2)]	SAGE
Πλήθος Επιπέδων	[4, 5, 6]	6
Hidden Dimensions ανά επίπεδο	[32, 64]	64
Πλήθος Γραμμικών Επιπέδων	[1, 2]	2
Ρυθμός Εκπαίδευσης (Learning Rate)	[0.001, 0.0001]	0.001
Batch Size	[16, 32]	16

Πίνακας 8: Υπερπαράμετροι μοντέλων και βέλτιστος συνδυασμός για σενάριο 2A

Τα αποτελέσματα του μοντέλου SAGE_6_64_2_0.001_16 στους γράφους επικύρωσης φαίνονται στην παρακάτω εικόνα:



Εικόνα 25: Προβλεπόμενη από το μοντέλο χρονοσειρά του BTC σε σχέση με την πραγματική για το σενάριο 2A στο σύνολο επικύρωσης.

Τα αποτελέσματα των μετρικών RMSE και MAPE παρουσιάζονται στον παρακάτω πίνακα:

Σενάριο	Μοντέλο	RMSE	MAPE
2A	SAGE_6_64_2_0.001_16	110.195	1.021%

Πίνακας 9: Αποτελέσματα μετρικών RMSE και MAPE βέλτιστου μοντέλου στους γράφους επικύρωσης για το σενάριο 2A

Παρατηρούμε ότι προέκυψε το ίδιο μοντέλο με το σενάριο 1A και πάλι από όλα όσα εξετάστηκαν παρουσίασε τα καλύτερα αποτελέσματα σε σχέση με τα υπόλοιπα και στις δύο μετρικές.

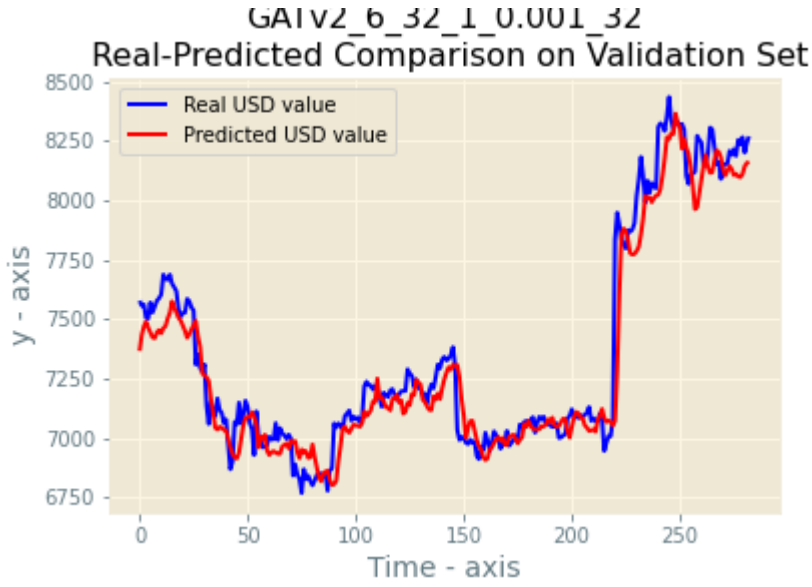
8.3: Αποτελέσματα Σεναρίου 3A στους Γράφους Επικύρωσης

Το σενάριο 3A έχει για σύνολο δεδομένων τρίωρης διάρκειας γράφους των οποίων οι κόμβοι περιέχουν τα δομικά χαρακτηριστικά των συναλλαγών (size, no_of_inputs, no_of_outputs, total_btc_input, fee), τη τιμή του bitcoin (usd_worth) και τα δύο οικονομικά χαρακτηριστικά που έχουν αναφερθεί σε προηγούμενες ενότητες (volume_usd, 10day-EMA)
Στον πίνακα 10 παρουσιάζεται ο συνδυασμός των υπερπαραμέτρων του μοντέλου με την καλύτερη απόδοση για το σενάριο 3A:

Υπερπαράμετροι	Σύνολο Αναζήτησης	Βέλτιστος Συνδυασμός
Αρχιτεκτονική Επιπέδων	[SAGE, GATv2(head attention=2)]	GATv2
Πλήθος Επιπέδων	[4, 5, 6]	6
Hidden Dimensions ανά επίπεδο	[32, 64]	32
Πλήθος Γραμμικών Επιπέδων	[1, 2]	1
Ρυθμός Εκπαίδευσης (Learning Rate)	[0.001, 0.0001]	0.001
Batch Size	[16, 32]	32

Πίνακας 10: Υπερπαράμετροι μοντέλων και βέλτιστος συνδυασμός για σενάριο 3A

Τα αποτελέσματα του μοντέλου GATv2_6_32_1_0.001_32 στους γράφους επικύρωσης φαίνονται στην παρακάτω εικόνα:



Εικόνα 26: Προβλεπόμενη από το μοντέλο χρονοσειρά του BTC σε σχέση με την πραγματική για το σενάριο 3Α στο σύνολο επικύρωσης.

Τα αποτελέσματα των μετρικών RMSE και MAPE παρουσιάζονται στον παρακάτω πίνακα:

Σενάριο	Μοντέλο	RMSE	MAPE
3Α	SAGE_6_32_1_0.001_32	116.064	1.083%

Πίνακας 11: Αποτελέσματα μετρικών RMSE και MAPE βέλτιστου μοντέλου στους γράφους επικύρωσης για το σενάριο 3Α

Σε αυτό το σενάριο προέκυψαν από το grid search δύο βέλτιστα μοντέλα. Το SAGE_4_64_1_0.001_32 με μετρικές RMSE = 113.167 και MAPE = 1.125% και το GATv2_6_32_1_0.001_32 με τα αποτελέσματα των μετρικών που φαίνονται στον παραπάνω πίνακα. Επιλέξαμε να κρατήσουμε το μοντέλο με το μικρότερο MAPE καθώς και στις δύο μετρικές οι διαφορές μεταξύ των μοντέλων είναι πολύ μικρές.

8.4: Συγκεντρωτικά Αποτελέσματα και Ερμηνεία τους

Συγκεντρωτικά τα αποτελέσματα των καλύτερων μοντέλων των διάφορων σεναρίων παρουσιάζονται στον παρακάτω πίνακα. Με bold επισημαίνεται το αποδοτικότερο μοντέλο το οποίο και επιλέγουμε ως τελικό μαζί με το σενάριο στο οποίο εξετάστηκε, για τις ωριαίες προβλέψεις που θα πραγματοποιήσουμε στο σύνολο ελέγχου.

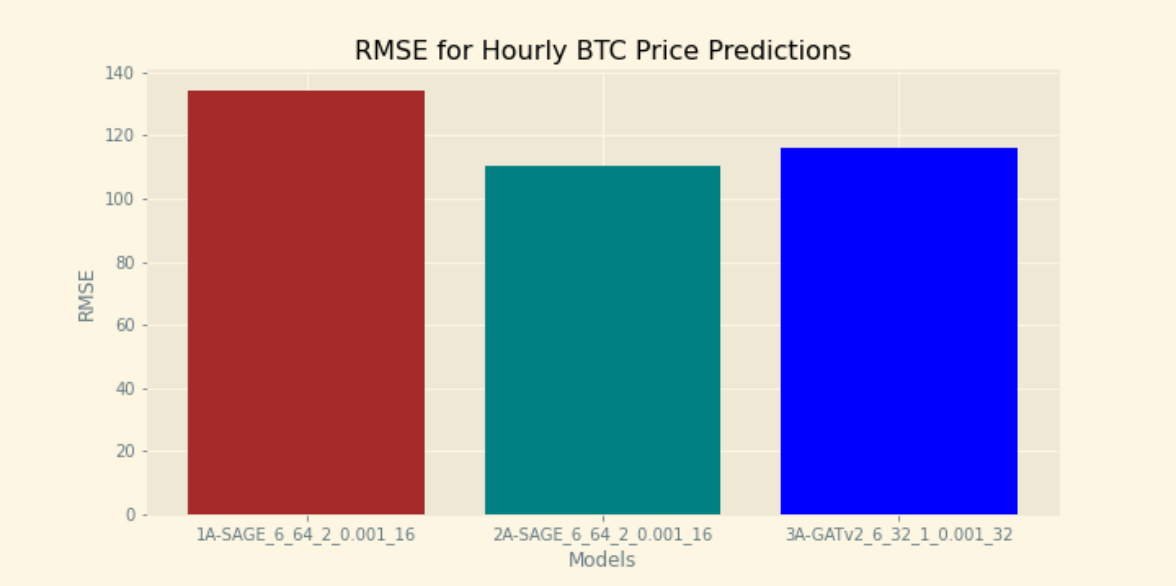
Σενάριο	Μοντέλο	RMSE	MAPE
1A	SAGE_6_64_2_0.001_16	134.057	1.274%
2A	**SAGE_6_64_2_0.001_16**	**110.195**	**1.021%**
3A	SAGE_6_32_1_0.001_32	116.064	1.083%

Πίνακας 12: Συγκεντρωτικά τα αποτελέσματα των καλύτερων μοντέλων των σεναρίων 1A, 2A, 3A

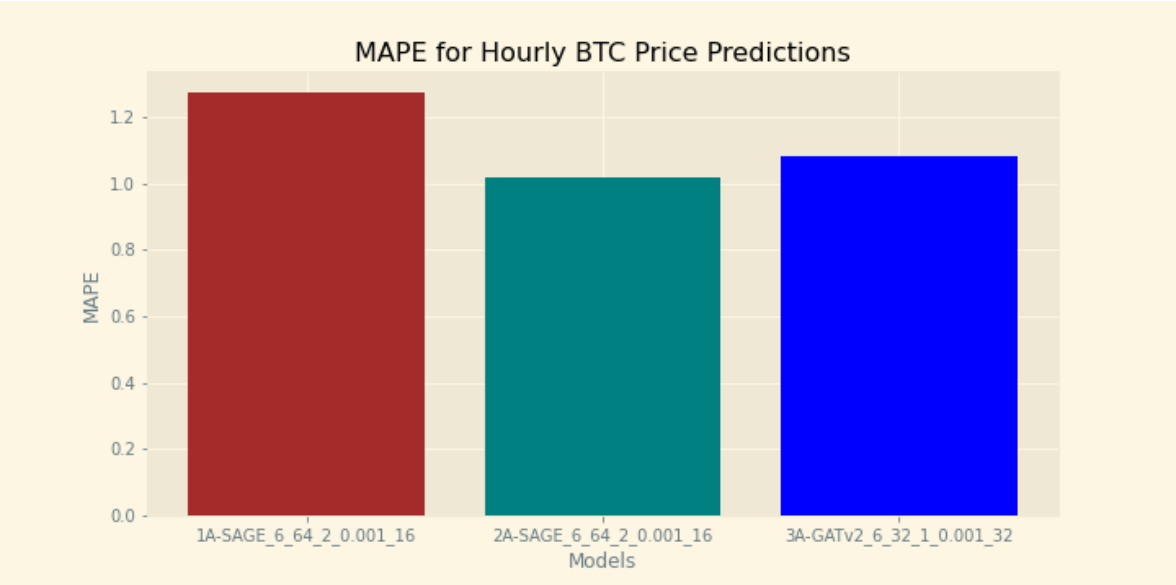
Με βάση όσα παρουσιάστηκαν στο παρόν κεφάλαιο μπορούμε να εξάγουμε αρκετά χρήσιμα συμπεράσματα. Αρχικά ως προς τα σενάρια, παρατηρούμε ότι οι τρίωροι γράφοι, δηλαδή στα σενάρια 2A, 3A, τα βέλτιστα μοντέλα τους είχαν και τα δύο καλύτερα αποτελέσματα και στις δύο μετρικές από το καλύτερο μοντέλο του σεναρίου 1A. Αυτό αναδεικνύεται ειδίκα με την σύγκριση των σεναρίων 1A και 2A καθώς η μόνη διαφορά μεταξύ των γράφων που χρησιμοποιήθηκαν, ήταν η διάρκεια τους και επιπλέον και τα δύο σενάρια είχαν το ίδιο καλύτερο μοντέλο, οπότε η διαφορά τους ήταν αποκλειστικά η διάρκεια των γράφων. Το συμπέρασμα αυτό είναι λογικό, καθώς, όταν τα δεδομένα είναι πιο κοντά στη πρόβλεψη χωρίς μεγάλο παρελθοντικό πλαίσιο, δηλαδή έχουμε γράφους με συνδέσεις των κοντινών στη πρόβλεψη συναλλαγών, είναι πιο πιθανό τα πρόσφατα δομικά μοτίβα, και κατ' επέκταση οι συνθήκες που τα δημιούργησαν, να επηρεάζουν περισσότερο τις μελλοντικές τιμές του BTC, παρά παλιότερα μοτίβα. Με τη χρήση των οικονομικών χαρακτηριστικών, παρατηρούμε ότι τα αποτελέσματα είναι αρκετά παρεμφερή χωρίς τη χρήση τους (και μάλιστα ελαφρώς χειρότερα), γεγονός που μπορεί να υποδεικνύει ότι, ειδικά για βραχυπρόθεσμες προβλέψεις, όπως αυτή της μίας ώρας, η δομική πληροφορία του γράφου συναλλαγών είναι αρκετή για τα καλύτερα αποτελέσματα.

Ως προς τα μοντέλα, παρατηρούμε ότι σε όλα τα σενάρια τα καλύτερα μοντέλα είχαν τον μεγαλύτερο αριθμό επιπέδων (6). Αυτό συμβαίνει διότι η αύξηση των επιπέδων συνεπάγεται το από πόσο μακριά παίρνει πληροφορία ο κάθε κόμβος, και άρα πόσοι γείτονες έχουν συμπεριληφθεί για τον υπολογισμό του τελικού embedding του, το οποίο συνεισφέρει στο αποτέλεσμα της πρόβλεψης. Εφόσον ο γράφος του Bitcoin τόσο για τις έξι όσο και για τις τρεις ώρες έχει διάμετρο πολύ μεγαλύτερη από το πλήθος των επιπέδων τα οποία χρησιμοποιούμε, δεν έχουμε τον κίνδυνο over-smoothing, δηλαδή οι τιμές των embeddings όλων των κόμβων να συγκλίνουν στην ίδια τιμή. Κάθε κόμβος έχει μικρό πλήθος κοινών γειτόνων με τους υπόλοιπους και άρα το embedding του συνεισφέρει με μοναδικό τρόπο (και άρα ουσιαστικά) στη πρόβλεψη. Παρατηρούμε ότι η αρχιτεκτονική

τύπου GraphSAGE έβγαλε καλύτερα αποτελέσματα από την αρχιτεκτονική GATv2 γεγονός το οποίο μπορεί να υποδηλώνει ότι τουλάχιστον για βραχυπρόθεσμες προβλέψεις ο attention μηχανισμός δεν ανιχνεύει μοτίβα που παίζουν μεγαλύτερο ρόλο από ότι άλλα στον καθορισμό της τιμής και η εκφραστικότητα που προσφέρει το GraphSAGE είναι αρκετή.



Εικόνα 27: RMSE για ωριαίες προβλέψεις

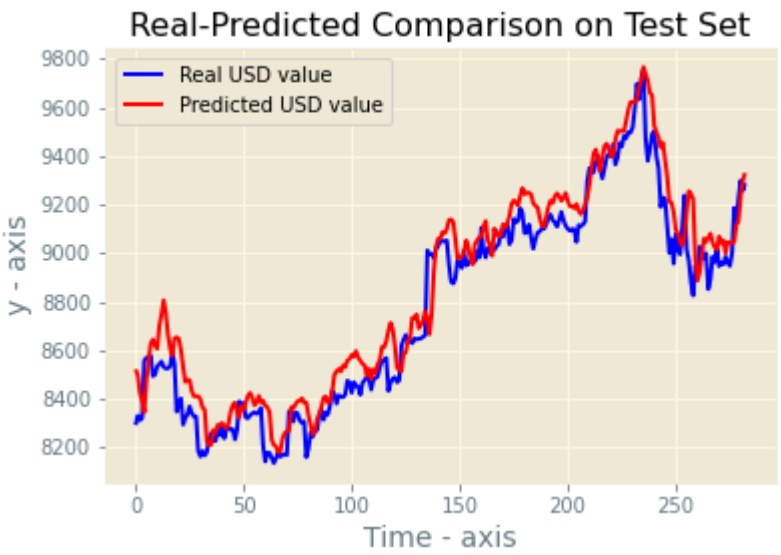


Εικόνα 28: MAPE για ωριαίες προβλέψεις

8.5: Επιλογή Καλύτερου Μοντέλου και Τελική Απόδοσή του στους Γράφους Ελέγχου

Με βάση όσα αναφέρθηκαν στην ενότητα 8.4, ο καλύτερος τρόπος να κατασκευάζουμε γράφους συναλλαγών για τη πρόβλεψη των ωριαίων τιμών του Bitcoin είναι το σενάριο 2A, δηλαδή τρίωροι γράφοι με μόνο τα δομικά χαρακτηριστικά των συναλλαγών και τη τιμή του Bitcoin στους κόμβους του γράφου. Επιπλέον, το καλύτερο μοντέλο GNN είναι το SAGE_6_64_2_0.001_16. Για τα παραπάνω θα εξετάσουμε την τελική απόδοση του μοντέλου με είσοδο αυτών των ειδών γράφων συναλλαγών στο τελικό σύνολο δεδομένων, τους γράφους ελέγχου.

Τα τελικά αποτελέσματα για την ωριαία πρόβλεψη φαίνονται στην παρακάτω εικόνα:



Εικόνα 29: Ωριαίες προβλέψεις της τιμής του BTC του βέλτιστου συνδυασμού (Σενάριο 1A - SAGE_6_64_2_0.001_16) στο τελικό σύνολο δεδομένων

Τα τελικά αποτελέσματα των μετρικών RMSE και MAPE παρουσιάζονται στον πίνακα 13:

Σενάριο	Μοντέλο	RMSE	MAPE
2A	SAGE_6_64_2_0.001_16	119.667	1.069%

Πίνακας 13: Αποτελέσματα μετρικών RMSE και MAPE για το τελικό μοντέλο, στο σύνολο ελέγχου για τη πρόβλεψη μίας ώρας μετά

Παρατηρούμε ότι το GNN μπορεί μέσω των γράφων συναλλαγών του Blockchain του Bitcoin να προβλέψει σε αρκετά καλό επίπεδο την τιμή του Bitcoin. Αυτό σημαίνει ότι είναι ικανό να εξάγει από τα τοπολογικά μοτίβα των υπογράφων των συναλλαγών καθώς και τα ίδια τα δομικά τους χαρακτηριστικά πληροφορίες οι οποίες παίζουν καθοριστικό ρόλο, μαζί με την ίδια τη τιμή του Bitcoin ως πληροφορία η οποία αποδίδεται με γραφοκεντρικό τρόπο στους κόμβους των συναλλαγών, την επιτυχή πρόβλεψη της τιμής του Bitcoin. Είναι σε θέση δηλαδή, όταν του δίνεται κάποιος γράφος συναλλαγών μια ώρα πριν την επιθυμητή πρόβλεψη, να αναγνωρίζει και να επεξεργάζεται την πληροφορία του και να προβλέπει μια τιμή αρκετά κοντά στη πραγματική.
Από την εικόνα 29 παρατηρούμε μάλιστα ότι παρόλες τις απότομες διακυμάνσεις της τιμής του Bitcoin, οι συναλλαγές που έχουν πραγματοποιηθεί καθώς και οι συνδεσιμότητα τους, είναι αρκετές για το GNN έτσι ώστε να προβλέψει τις διακυμάνσεις αυτές.
Το μοντέλο μπορεί να γραφτεί και ως:

```
GNN(
(convs): ModuleList(
      (0): SAGEConv (6, 64, aggr = mean)
      (1): SAGEConv (64, 64, aggr = mean)
      (2): SAGEConv (64, 64, aggr = mean)
      (3): SAGEConv (64, 64, aggr = mean)
      (4): SAGEConv (64, 64, aggr = mean)
      (5): SAGEConv (64, 64, aggr = mean) )
(acts): ModuleList(
      (0): Tahn()
      (1): Tahn()
      (2): Tahn()
      (3): Tahn()
      (4): Tahn()
      (5): Tahn() )
(out_1): Linear(in_feature = 128, out_features = 64, bias = True)
(out_2): Linear(in_feature = 64, out_features = 1, bias = True))
```

Κεφάλαιο 9: Πειραματικά αποτελέσματα για Εξάωρες Προβλέψεις

Πέρα από τις προβλέψεις της μίας ώρας, οι οποίες είναι σχετικά κοντά χρονικά με τα δεδομένα μας, εξετάζουμε και τη πρόβλεψη της τιμής του Bitcoin έξι ώρες μετά τη δημιουργία του εκάστοτε γράφου. Έτσι μελετάται η πιο μακροχρόνια προβλεπτική ισχύ που μπορούν να έχουν οι γράφοι του bitcoin, αλλά και η ικανότητα του GNN να τις αναγνωρίζει και να είναι σε θέση να τις επεξεργαστεί ώστε να υπολογίσει μια αρκετά πιο δύσκολη πρόβλεψη η οποία απαιτεί μεγάλο βαθμό γενίκευσης και εξαγωγής χρήσιμης πληροφορίας. Φυσικά, υπάρχει μεγάλο χρονικό πλαίσιο μεταξύ των δεδομένων και της ζητούμενης τιμής, στο οποίο πολλοί ξαφνικοί παράγοντες μπορούν να επηρεάσουν τη πραγματική τιμή, χωρίς να έχουν ληφθεί υπόψη από το μοντέλο. Για αυτό άλλωστε και οι πιο μακροχρόνιες προβλέψεις αποτελούν ένα αρκετά απαιτητικό εγχείρημα.

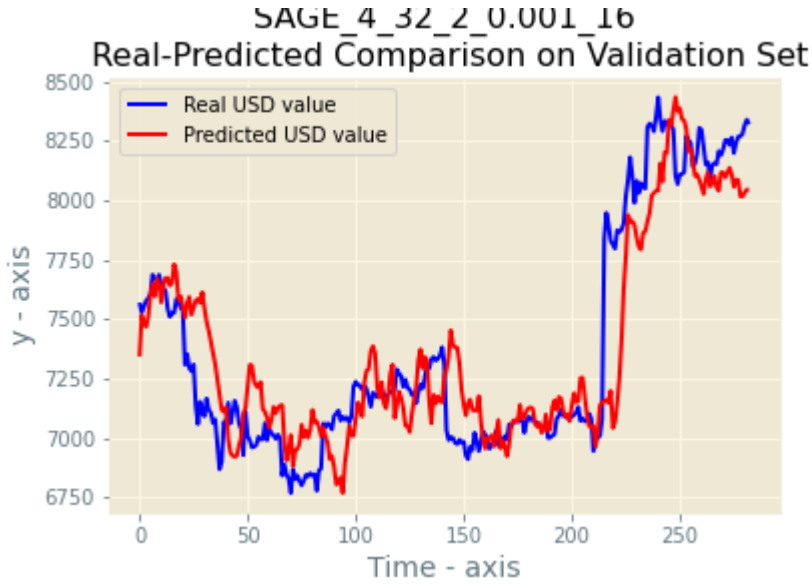
9.1: Αποτελέσματα Σεναρίου 1Β στους Γράφους Επικύρωσης

Στον πίνακα 14 παρουσιάζεται ο συνδυασμός των υπερπαραμέτρων του μοντέλου με την καλύτερη απόδοση για το σενάριο 1Β:

Υπερπαράμετροι	Σύνολο Αναζήτησης	Βέλτιστος Συνδυασμός
Αρχιτεκτονική Επιπέδων	[SAGE, GATv2(head attention=2)]	SAGE
Πλήθος Επιπέδων	[4, 5, 6]	4
Hidden Dimensions ανά επίπεδο	[32, 64]	32
Πλήθος Γραμμικών Επιπέδων	[1, 2]	2
Ρυθμός Εκπαίδευσης (Learning Rate)	[0.001, 0.0001]	0.001
Batch Size	[16, 32]	16

Πίνακας 14: Υπερπαράμετροι μοντέλων και βέλτιστος συνδυασμός για σενάριο 1Β

Τα αποτελέσματα του μοντέλου SAGE_4_32_2_0.001_16 στους γράφους επικύρωσης φαίνονται στην παρακάτω εικόνα:



Εικόνα 30: Προβλεπόμενη από το μοντέλο χρονοσειρά του BTC σε σχέση με την πραγματική για το σενάριο 1Β στο σύνολο επικύρωσης.

Τα αποτελέσματα των μετρικών RMSE και MAPE παρουσιάζονται στον παρακάτω πίνακα:

Σενάριο	Μοντέλο	RMSE	MAPE
1Β	SAGE_4_32_2_0.001_16	216.689	2.158%

Πίνακας 15: Αποτελέσματα μετρικών RMSE και MAPE βέλτιστου μοντέλου στους γράφους επικύρωσης για το σενάριο 1Β.

Σε αυτό το σενάριο προέκυψαν από το grid search δύο βέλτιστα μοντέλα. Το SAGE_4_64_1_0.0001_32 με μετρικές RMSE = 214.214 και MAPE = 2.200% και το SAGE_4_32_2_0.001_16 με τα αποτελέσματα των μετρικών που φαίνονται στον παραπάνω πίνακα. Επιλέξαμε να κρατήσουμε το μοντέλο με το μικρότερο MAPE, καθώς και στις δύο μετρικές οι διαφορές μεταξύ των μοντέλων είναι πολύ μικρές.

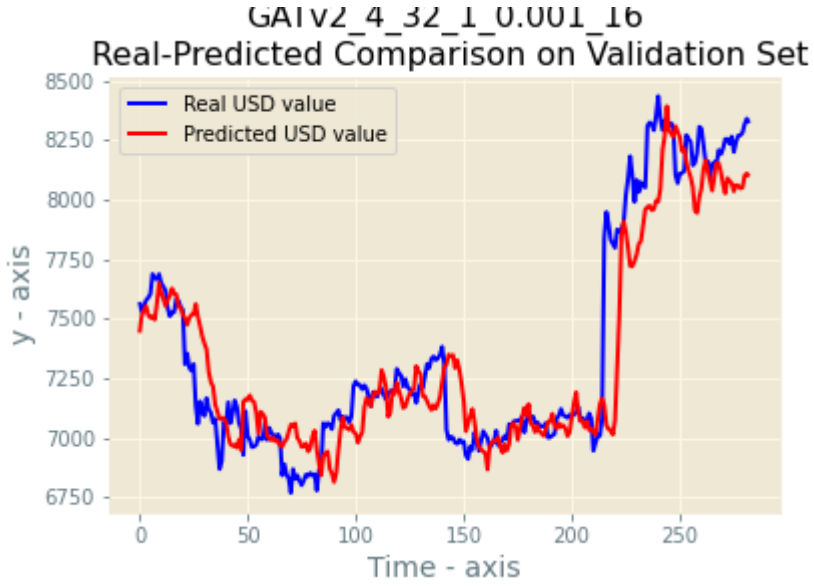
9.2: Αποτελέσματα Σεναρίου 2Β στους Γράφους Επικύρωσης

Στον πίνακα 16 παρουσιάζεται ο συνδυασμός των υπερπαραμέτρων του μοντέλου με την καλύτερη απόδοση για το σενάριο 2Β:

Υπερπαράμετροι	Σύνολο Αναζήτησης	Βέλτιστος Συνδυασμός
Αρχιτεκτονική Επιπέδων	[SAGE, GATv2(head attention=2)]	GATv2
Πλήθος Επιπέδων	[4, 5, 6]	4
Hidden Dimensions ανά επίπεδο	[32, 64]	32
Πλήθος Γραμμικών Επιπέδων	[1, 2]	1
Ρυθμός Εκπαίδευσης (Learning Rate)	[0.001, 0.0001]	0.001
Batch Size	[16, 32]	16

Πίνακας 16: Υπερπαράμετροι μοντέλων και βέλτιστος συνδυασμός για σενάριο 2Β

Τα αποτελέσματα του μοντέλου GATv2_4_32_1_0.001_16 στους γράφους επικύρωσεις φαίνονται στην παρακάτω εικόνα:



Εικόνα 31: Προβλεπόμενη από το μοντέλο χρονοσειρά του BTC σε σχέση με την πραγματική για το σενάριο 2Β στο σύνολο επικύρωσης.

Τα αποτελέσματα των μετρικών RMSE και MAPE παρουσιάζονται στον παρακάτω πίνακα:

Σενάριο	Μοντέλο	RMSE	MAPE
2B	GATv2_4_32_1_0.001_16	197.223	1.820%

Πίνακας 17: Αποτελέσματα μετρικών RMSE και MAPE βέλτιστου μοντέλου στους γράφους επικύρωσης για το σενάριο 2B.

Σε αυτό το σενάριο προέκυψαν από το grid search δύο βέλτιστα μοντέλα. Το SAGE_5_32_1_0.001_16 με μετρικές RMSE = 193.806 και MAPE = 2.005% και το GATv2_4_32_1_0.001_16 με τα αποτελέσματα των μετρικών που φαίνονται στον παραπάνω πίνακα. Επιλέξαμε να κρατήσουμε το μοντέλο με το μικρότερο MAPE καθώς σε αυτή τη μετρική βασίζεται κυρίως η ανάλυσή μας.

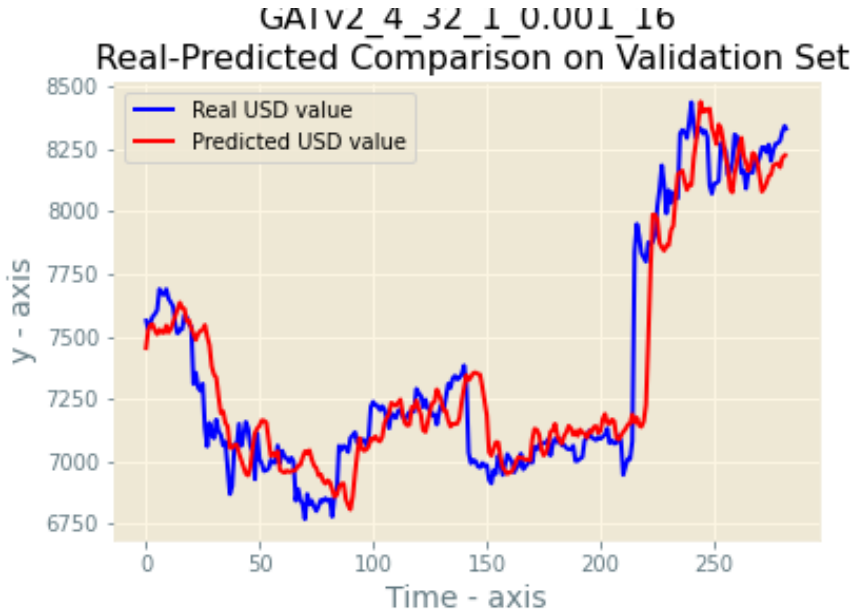
9.3: Αποτελέσματα Σεναρίου 3B στους Γράφους Επικύρωσης

Στον πίνακα 18 παρουσιάζεται ο συνδυασμός των υπερπαραμέτρων του μοντέλου με την καλύτερη απόδοση για το σενάριο 3B:

Υπερπαράμετροι	Σύνολο Αναζήτησης	Βέλτιστος Συνδυασμός
Αρχιτεκτονική Επιπέδων	[SAGE, GATv2(head attention=2)]	GATv2
Πλήθος Επιπέδων	[4, 5, 6]	4
Hidden Dimensions ανά επίπεδο	[32, 64]	32
Πλήθος Γραμμικών Επιπέδων	[1, 2]	1
Ρυθμός Εκπαίδευσης (Learning Rate)	[0.001, 0.0001]	0.001
Batch Size	[16, 32]	16

Πίνακας 18: Υπερπαράμετροι μοντέλων και βέλτιστος συνδυασμός για σενάριο 3B

Τα αποτελέσματα του μοντέλου GATv2_4_32_1_0.001_16 στους γράφους επικύρωσεις φαίνονται στην παρακάτω εικόνα:



Εικόνα 32: Προβλεπόμενη από το μοντέλο χρονοσειρά του BTC σε σχέση με την πραγματική για το σενάριο 3Β στο σύνολο επικύρωσης.

Τα αποτελέσματα των μετρικών RMSE και MAPE παρουσιάζονται στον παρακάτω πίνακα:

Σενάριο	Μοντέλο	RMSE	MAPE
3Β	GATv2_4_32_1_0.001_16	168.869	1.596%

Πίνακας 19: Αποτελέσματα μετρικών RMSE και MAPE βέλτιστου μοντέλου στους γράφους επικύρωσης για το σενάριο 3Β.

Παρατηρούμε ότι προέκυψε το ίδιο μοντέλο με το σενάριο 2Β και από όλα όσα εξετάστηκαν παρουσίασε τα καλύτερα αποτελέσματα και στις δύο μετρικές.

9.4: Συγκεντρωτικά Αποτελέσματα και Ερμηνεία τους

Συγκεντρωτικά τα αποτελέσματα των καλύτερων μοντέλων των διάφορων σεναρίων παρουσιάζονται στον παρακάτω πίνακα. Με bold επισημαίνεται το αποδοτικότερο μοντέλο το οποίο και επιλέγουμε ως τελικό μαζί με το σενάριο στο οποίο εξετάστηκε, για τις εξάωρες προβλέψεις που θα πραγματοποιήσουμε στο σύνολο ελέγχου.

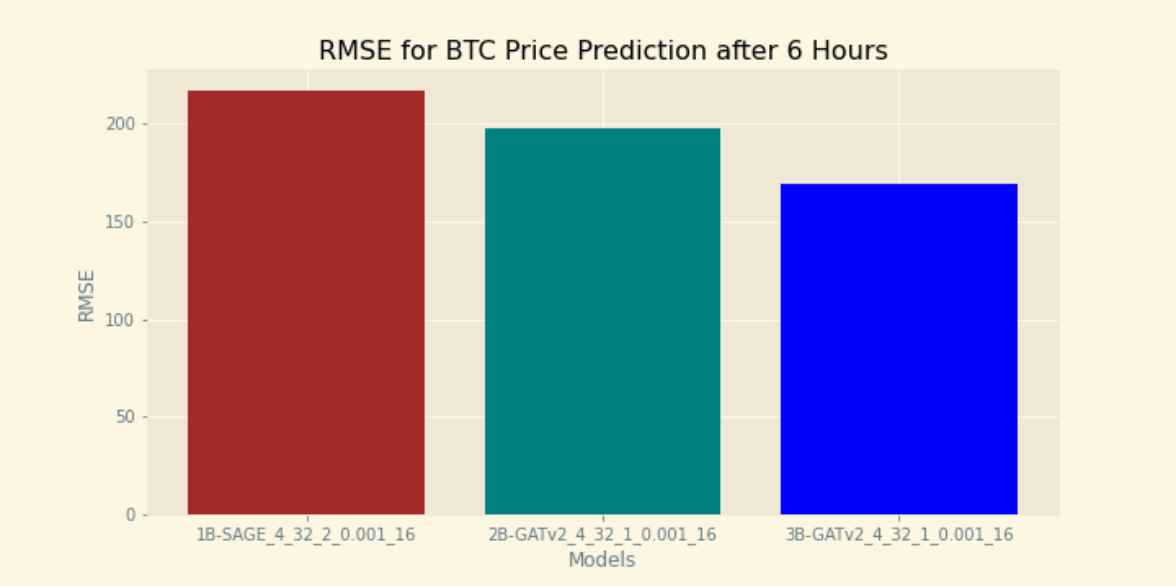
Σενάριο	Μοντέλο	RMSE	MAPE
1B	SAGE_4_32_2_0.001_16	216.689	2.158%
2B	GATv2_4_32_1_0.001_16	197.223	1.820%
3B	**GATv2_4_32_1_0.001_16**	**168.869**	**1.596%**

Πίνακας 20: Συγκεντρωτικά τα αποτελέσματα των καλύτερων μοντέλων των σεναρίων 1B, 2B, 3B

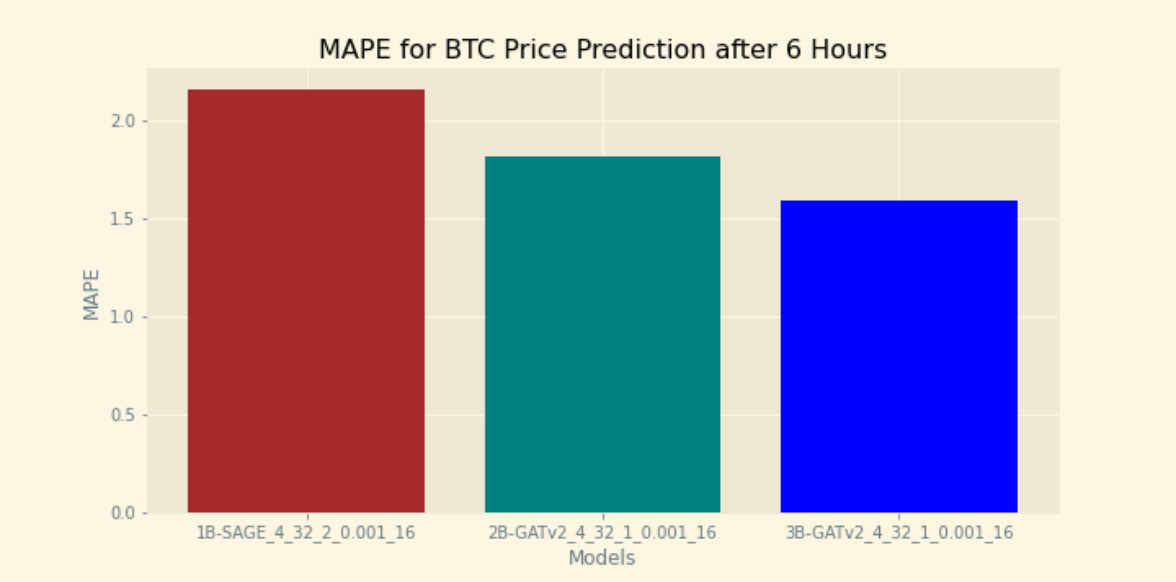
Ως προς τα σενάρια, παρατηρούμε για άλλη μια φορά ότι για τους τρίωρους γράφους (σενάρια 2B, 3B), τα αντίστοιχα μοντέλα τους είναι και ως προς τις δύο μετρικές καλύτερα από το μοντέλο για τους εξάωρους γράφους, το οποίο και για την εξάωρη πρόβλεψη είναι λογικό καθώς, όσο πιο πρόσφατα στη πρόβλεψη δεδομένα έχουμε, τόσο πιο πιθανό τα πρόσφατα δομικά μοτίβα να επηρεάζουν περισσότερο τις μελλοντικές τιμές του Bitcoin από ότι παλιότερα μοτίβα. Σε αντίθεση με το πρόβλημα πρόβλεψης της μίας ώρας, παρατηρούμε ότι το σενάριο με τα οικονομικά χαρακτηριστικά έχει επιτύχει χαμηλότερες τιμές και στις δύο μετρικές σφαλμάτων. Αυτό φαίνεται ακόμα εντονότερα από το γεγονός ότι τα σενάρια 2B, 3B έχουν αποκλειστική διαφορά μεταξύ των αποτελεσμάτων των καλύτερων μοντέλων μόνο το αν ή όχι χρησιμοποιούν στις συναλλαγές τα οικονομικά χαρακτηριστικά, καθώς από τα grid search τους προέκυψε το ίδιο καλύτερο μοντέλο GNN. Φαίνεται ότι για μακροπρόθεσμες προβλέψεις το να αποδοθούν με γραφοκεντρικό τρόπο εξωτερικά χαρακτηριστικά που θεωρούμε ότι συμβάλλουν στην διαμόρφωση της τιμής έχει αποτέλεσμα. Μια πιθανή ερμηνεία για αυτό θα μπορούσε να είναι το γεγονός ότι ενώ στις βραχυπρόθεσμες προβλέψεις η ροή του χρήματος μέσω των γράφων συναλλαγών μαζί με τα δομικά χαρακτηριστικά και τα τοπολογικά μοτίβα αρκεί να αποτυπώσει την τάση, για μακροπρόθεσμες προβλέψεις, όπου το πρόβλημα γίνεται αρκετά πιο πολυμεταβλητό, σύνθετο και εξαρτώμενο από ξαφνικούς παράγοντες, απαιτούνται εξωτερικά χαρακτηριστικά που αποδίδονται με γραφοκεντρικό τρόπο έτσι ώστε να “επεκτείνουν” το φάσμα παραγόντων που λαμβάνει υπόψη του το GNN για να πραγματοποιήσει μια πιο δύσκολη πρόβλεψη.

Ως προς τα μοντέλα, τα συμπεράσματα είναι εξίσου ενδιαφέροντα. Παρατηρούμε ότι για την εξάωρη πρόβλεψη η αρχιτεκτονική επιπέδων GNN που ξεχώρισε ήταν το GATv2 και μάλιστα με το μικρότερο πλήθος επιπέδων (4). Ο λόγος που βγήκε το GATv2 ως καλύτερο έναντι του GraphSAGE στη προκειμένη περίπτωση πάλι μπορεί να εξηγηθεί από το γεγονός της δυσκολίας του εγχειρήματος. Τα attention mechanisms σε αυτή τη περίπτωση έχουν αποτέλεσμα καθώς το GNN ξεχωρίζει σημαντικούς υπογράφους (τοπολογικά μοτίβα), των

οποίων η ύπαρξη μπορεί να μην έπαιζε ρόλο για τις βραχυπρόθεσμες προβλέψεις, η οποία καθορίζεται από τη γενικότερη τάση και ροή του Bitcoin στο Blockchain, αλλά η παρουσία τους στους γράφους να υποδηλώνει την σημασία τους για την πιο μακροχρόνια πορεία του Bitcoin. Επιπροσθέτως,το πλήθος τον επιπέδων δείχνει ότι το receptive field του κάθε κόμβου για την μακροχρόνια πρόβλεψη πρέπει να είναι μικρότερο, έτσι ώστε να έχει όσο το δυνατόν πιο πρόσφατους γείτονες που θα επηρεάζουν το τελικό embedding του για τη πρόβλεψη.



Εικόνα 33: RMSE για εξάωρες προβλέψεις

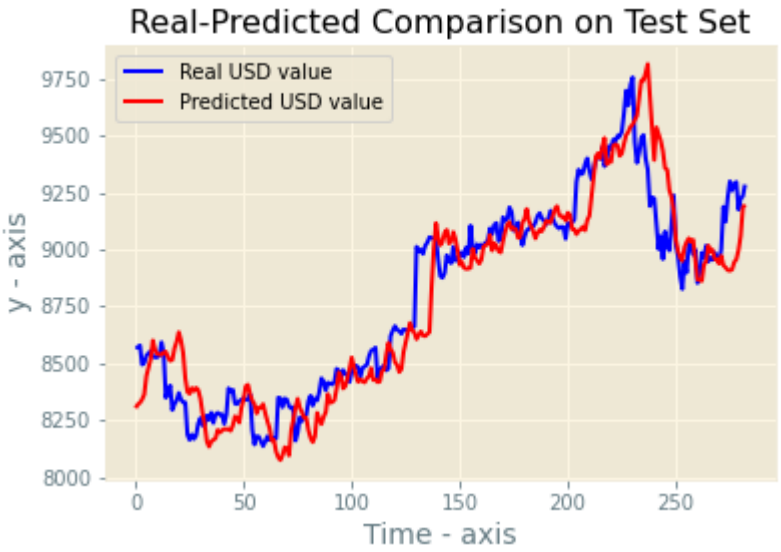


Εικόνα 34: MAPE για εξάωρες προβλέψεις

9.5: Επιλογή Καλύτερου Μοντέλου και Τελική Απόδοσή του στους Γράφους Ελέγχου

Με βάση όσα αναφέρθηκαν στην ενότητα 9.4, ο καλύτερος τρόπος να κατασκευάζουμε γράφους συναλλαγών για τη πρόβλεψη της τιμής του BTC έξι ώρες μετά τον εκάστοτε γράφο είναι το σενάριο 3Β, δηλαδή τρίωροι γράφοι και με τα δομικά χαρακτηριστικά των συναλλαγών και με τη τιμή του BTC αλλά και τα οικονομικά χαρακτηριστικά στους κόμβους του γράφου. Επιπλέον, το καλύτερο μοντέλο GNN είναι το GATv2_4_32_1_0.001_16. Για τα παραπάνω θα εξετάσουμε την τελική απόδοση του μοντέλου με είσοδο αυτών των ειδών γράφων συναλλαγών στο τελικό σύνολο δεδομένων ελέγχου.

Τα τελικά αποτελέσματα για την πρόβλεψη έξι ωρών μετά, φαίνονται στην παρακάτω εικόνα:



Εικόνα 35: Εξάωρες προβλέψεις της τιμής του BTC του βέλτιστου συνδυασμού (Σενάριο 3Β-GATv2_4_32_1_0.001_16) στο τελικό σύνολο δεδομένων.

Τα τελικά αποτελέσματα των μετρικών RMSE και MAPE παρουσιάζονται στον παρακάτω πίνακα:

Σενάριο	Μοντέλο	RMSE	MAPE
3Β	GATv2_4_32_1_0.001_16	158.887	1.331%

Πίνακας 21: Αποτελέσματα μετρικών RMSE και MAPE για το τελικό μοντέλο, στο σύνολο ελέγχου για τη πρόβλεψη έξι ώρες μετά.

Σε σχέση με την ωριαία πρόβλεψη παρατηρούμε προφανώς ότι τα σφάλματα των μετρικών εδώ είναι μεγαλύτερα. Αυτό είναι αναμενόμενο με βάση όσα εξηγήσαμε. Βέβαια το μέγεθος της αύξησης είναι μικρό, που αναδεικνύει πάλι την προβλεπτική ισχύ όλης της μεθόδου που ακολουθήσαμε στη παρούσα εργασία. Παρατηρούμε όμως ότι στις απότομες αλλαγές της τιμής του Bitcoin, το μοντέλο μας αντιδρά με αρκετή καθυστέρηση. Αυτό μπορεί να οφείλεται σε πολλούς λόγους, όπως το να επηρεάστηκε η τιμή από παράγοντες που δεν μπορούσαν να αποτυπωθούν στη δομή του γράφου από πριν, ή να υπάρχει κάποιο σχετικό μοτίβο για αυτούς (π.χ. ανακοίνωση σημαντικών πολιτικών νέων, απαγορεύσεων κλπ). Στις πιο ήπιες αυξομειώσεις παρατηρούμε ότι το μοντέλο σε μεγάλο μέρος πάλι έχει κάποια καθυστέρηση, μικρότερη όμως από ότι στις απότομες αυξομειώσεις. Σε πιο επαναλαμβανόμενα μοτίβα της τιμής, οι προβλέψεις συμβαδίζουν με τις πραγματικές τιμές, το οποίο δείχνει ότι το GNN εξάγει συμπεράσματα από τους γράφους ακόμα και για πιο μακροχρόνιες προβλέψεις όταν επαναλαμβάνονται γνωστά μοτίβα συναλλαγών. Αυτό αποδεικνύει ότι υπάρχουν τοπολογίες στους γράφους, ανεξάρτητα από τους παράγοντες οι οποίοι μπορεί να μεσολαβήσουν από τον γράφο μέχρι την πραγματική τιμή του Bitcoin, οι οποίες καθορίζουν σε κάποιο βαθμό την τιμή του Bitcoin και μάλιστα τα GNN είναι ικανά να τις ανιχνεύουν και να τις επεξεργάζονται με τέτοιο τρόπο ώστε να βγάζουν μια πρόβλεψη αρκετά κοντά στη πραγματική τιμή. Κρίνεται αναγκαία όμως περισσότερη διερεύνηση για τις πιο μακροχρόνιες προβλέψεις, πιο προηγμένων μοντέλων GNN για καλύτερες αποδόσεις με μικρότερη καθυστέρηση στη πρόβλεψη της τιμής.
Η αρχιτεκτονική του μοντέλου μπορεί να γραφτεί συνοπτικά και ως εξής:

```
GNN(
(convs): ModuleList(
(0): GATv2Conv (8, 32, heads = 2)
(1): GATv2Conv (32, 32, heads = 2)
(2): GATv2Conv (32, 32, heads = 2)
(3): GATv2Conv (32, 32, heads = 2) )
(acts): ModuleList(
(0): Tahn()
(1): Tahn()
(2): Tahn()
(3): Tahn()
)
(out): Linear(in_feature = 64, out_features = 1, bias = True))
```

Κεφάλαιο 10: Επίλογος

10.1: Σύνοψη

Στη παρούσα εργασία μελετήθηκε το πρόβλημα ακριβούς πρόβλεψης της τιμής του Bitcoin για χρονικό παράθυρο μίας και έξι ωρών. Συγκεκριμένα, εκμεταλλευόμενοι το γεγονός ότι οι συναλλαγές που πραγματοποιούνται στο Blockchain του Bitcoin στη πορεία του χρόνου μπορούν να αναπαρασταθούν ως οντότητες οι οποίες σχετίζονται μεταξύ τους, δημιουργήσαμε έναν αλγόριθμο, ο οποίος κατασκευάζει με κατάλληλο τρόπο τους κατευθυνόμενους ακυκλικούς γράφους των συναλλαγών αυτών. Ωστόσο μπορούν να προκύψουν πολλές παραλλαγές των γράφων αυτών, οι οποίες ενδεχομένως επηρεάζουν τη ποιότητα των προβλέψεών μας. Έτσι, κατασκευάσαμε διάφορα είδη γράφων συναλλαγών, όπως τρίωρης και εξάωρης διάρκειας γράφους, οι οποίοι περιλαμβάνουν είτε μόνο δομικά χαρακτηριστικά και τη τιμή του Bitcoin στις πληροφορίες των συναλλαγών, είτε και οικονομικούς δείκτες. Αφού έγινε η κατάλληλη προεπεξεργασία των χαρακτηριστικών, αναπτύξαμε διάφορα μοντέλα Νευρωνικών Δικτύων Γράφων στα οποία τροφοδοτήσαμε τους γράφων συναλλαγών. Εξετάστηκαν διάφοροι υπερπαράμετροι για τα GNN και προέκυψε ότι το καλύτερο μοντέλο για τη πρόβλεψη της μίας ώρας ήταν το GraphSAGE με RMSE ίσο με 119.667 και MAPE ίσο με 1.069% το οποίο είχε ως είσοδο τρίωρης διάρκειας γράφους με μόνο τα δομικά χαρακτηριστικά των συναλλαγών και τη τιμή του BTC. Για την εξάωρη πρόβλεψη, το καλύτερο μοντέλο ήταν τύπου αρχιτεκτονικής επιπέδων GATv2 με είσοδο τρίωρους γράφους συναλλαγών οι οποίοι είχαν και οικονομικούς δείκτες ως χαρακτηριστικά. Η απόδοση του μοντέλου στην RMSE μετρική ήταν ίση με 158.887 και MAPE 1.331%.

10.2: Συμπεράσματα - Συνεισφορά Εργασίας:

Οι υπάρχουσες έρευνες που αφορούν την πρόβλεψη τιμής κρυπτονομισμάτων χρησιμοποιούν ένα ευρύ φάσμα μοντέλων, από απλά στατιστικά μέχρι σύνθετα μοντέλα βαθιών νευρωνικών δικτύων. Εξίσου πολλές είναι και οι παράμετροι που εισάγουν σε αυτά για να βελτιώσουν τις προβλέψεις τους, από πολύπλοκους οικονομικούς δείκτες μέχρι και τα σχόλια χρηστών σε μέσα κοινωνικής δικτύωσης, προσπαθώντας έτσι να καταλάβουν ποιοι παράγοντες τελικά επηρεάζουν τη τιμή του Bitcoin. Ειδικά για τις βραχυπρόθεσμες προβλέψεις όμως, όλοι αυτοί οι εξωτερικοί παράγοντες, ανεξάρτητα του ποιοι εν τέλει είναι, θα αποτυπωθούν τελικά στις αλλαγές που συμβαίνουν στην ίδια τη δομή του Blockchain. Αυτό είναι λογική απόρροια του γεγονότος ότι οι εξωτερικοί παράγοντες που επηρεάζουν τη τιμή του Bitcoin θα έχουν επίδραση στις αποφάσεις των χρηστών, των οποίων οι αποφάσεις θα αποτυπωθούν μέσω συναλλαγών στο Blockchain. Το αποτέλεσμα είναι οι ίδιες οι πληροφορίες των συναλλαγών, μαζί με τον τρόπο που συνδέονται και

εξαρτώνται η μια από την άλλη, να δημιουργούν μοτίβα τα οποία αποτυπώνονται τοπολογικά στους υποκείμενους γράφους συναλλαγών. Αυτά με τη σειρά τους, περιλαμβάνουν αρκετή κωδικοποιημένη πληροφορία η οποία αποτυπώνει την πραγματική οικονομική κατάσταση της αγοράς του Bitcoin και άρα προσφέρει προβλεπτική ισχύ για την πρόβλεψη της τιμής μελλοντικά. Οι έρευνες που εστιάζουν σε αυτή τη προσέγγιση μέχρι στιγμής στη βιβλιογραφία είναι λίγες και εξάγουν στατιστικές πληροφορίες όπως τη συχνότητα εμφάνισης συγκεκριμένων υπογράφων συναλλαγών μέσω graph mining τεχνικών, τις οποίες χρησιμοποιούν ύστερα σε μοντέλα που έχουν ήδη δοκιμαστεί για το πρόβλημα της πρόβλεψης της τιμής κρυπτονομισμάτων.

Η κύρια ερευνητική συνεισφορά της παρούσας εργασίας ήταν να εισάγει για πρώτη φορά στην βιβλιογραφία, μια νέα μέθοδο πρόβλεψης της τιμής του Bitcoin χρησιμοποιώντας Νευρωνικά Δίκτυα Γράφων (GNN) με εισόδους τους γράφους συναλλαγών του Blockchain του Bitcoin και να προσφέρει έναν εντελώς διαφορετικό τρόπο αντιμετώπισης του προβλήματος. Αποδείξαμε ότι εκμεταλλευόμενα τη συνδεσιμότητα του δικτύου συναλλαγών, τα GNN μπορούν να αναγνωρίζουν από μόνα τους τα τοπολογικά μοτίβα των συναλλαγών, να συγκεντρώνουν και να μετασχηματίζουν κατάλληλα τα δομικά τους χαρακτηριστικά, καθως και τη τιμή που είχε το Bitcoin όταν έλαβαν χώρα οι συναλλαγές αυτές, και να υπολογίζουν μια αρκετά κοντά στη πραγματική τιμή πρόβλεψη.

Για να επιτευχθεί αυτό, κατασκευάσαμε εξειδικευμένους γράφους συναλλαγών που είναι κατάλληλοι για είσοδοι σε Νευρωνικά Δίκτυα Γράφων, μέσω ειδικού αλγορίθμου κατασκευής που αναπτύξαμε. Μάλιστα δημιουργήσαμε διάφορες παραλλαγές και δοκιμάσαμε για πρώτη φορά στη βιβλιογραφία, να αποδώσουμε με γραφοκεντρικό τρόπο εξωτερικά του δικτύου χαρακτηριστικά, για να εξετάσουμε τις επιδράσεις μιας τέτοιας αναπαράστασης χαρακτηριστικών.

Παρατηρήσαμε μέσω των πειραμάτων, ότι ενώ στην ωριαία πρόβλεψη οι κόμβοι των γράφων συναλλαγών αρκούν να έχουν τα δομικά χαρακτηριστικά, στην πρόβλεψη έξι ωρών μετά, η γραφοκεντρική αναπαράσταση των εξωτερικών οικονομικών χαρακτηριστικών μαζί με τα δομικά, βελτίωσε τη πρόβλεψη. Στις βραχυπρόθεσμες προβλέψεις, η ροή του χρήματος μέσω των γράφων συναλλαγών και τα τοπολογικά μοτίβα αρκεί να αποτυπώσει την τάση, ενώ για μακροπρόθεσμες προβλέψεις όπου το πρόβλημα γίνεται αρκετά πιο πολυμεταβλητό και εξαρτώμενο από σύνθετους παράγοντες, τα εξωτερικά χαρακτηριστικά προσθέτουν εκφραστικότητα στο φάσμα παραγόντων που λαμβάνει υπόψη του το GNN για να πραγματοποιήσει μια πιο δύσκολη πρόβλεψη.
Επιπλέον, παρατηρήθηκε ότι η αρχιτεκτονική GraphSAGE ήταν κατάλληλη για τη βραχυχρόνια πρόβλεψη, ενώ για τη πιο μακροχρόνια πρόβλεψη τα attention mechanisms του GATv2 βοήθησαν στο να βγάλει καλύτερα αποτελέσματα έναντι της GraphSAGE, καθώς λογικά κατάφεραν να εντοπίσουν πιο περίπλοκα μοτίβα που συμβάλλουν στο καθορισμό της μακροχρόνιας τάσης της τιμής.

Τα αποτελέσματα που λάβαμε από τα πειράματά μας είναι πέραν πάσης προσδοκίας. Λαμβάνοντας υπόψιν ότι η εργασία μας είναι η πρώτη στο είδος της και ως εκ τούτου προφανώς επιδέχεται βελτιώσεων, εν τούτοις παρήγαγε τα καλύτερα αποτελέσματα από όλες τις υπόλοιπες μεθόδους βραχυχρόνιων προβλέψεων της βιβλιογραφίας πλην μιάς.

10.3: Μελλοντικές Επεκτάσεις

Είμαστε στην ευχάριστη θέση να θεωρούμε ότι μέσω της δουλειάς μας ανοίγεται η δυνατότητα έρευνας σε ένα νέο συναρπαστικό πεδίο, τη πρόβλεψη τιμής με τη χρήση των state-of-the-art GNNs στους γράφους συναλλαγών, με πιο προχωρημένες τεχνικές που θα βγάλουν ακόμα καλύτερα αποτελέσματα και θα εφαρμοστούν όχι μόνο στο Bitcoin αλλά και σε άλλα διάσημα κρυπτονομίσματα. Οι μελλοντικές κατευθύνσεις έρευνας που μπορούμε να δώσουμε είναι πολλές. Αρχικά, για να βελτιωθούν οι αποδόσεις των μοντέλων, μπορούν να δοκιμαστούν πιο προχωρημένες τεχνικές στα GNN, για παράδειγμα αντί για global pooling, το οποίο, ειδικά για μεγάλους γράφους όπως αυτοί που μελετήσαμε, χάνει μέρος της πληροφορίας, μπορεί να δοκιμαστεί hierarchical pooling. Επιπλέον, λόγω των τεχνικών περιορισμών που είχαμε ως προς το μέγεθος μνήμης της GPU RAM (χρησιμοποιήθηκε NVIDIA Tesla T4 του Google Colab) δεν ήμασταν σε θέση να τρέξουμε μεγαλύτερο grid search και πιο σύνθετες σχεδιαστικές και αρχιτεκτονικές επιλογές για τα GNN. Θεωρούμε ότι πιο εκτεταμένο grid search και διαφορετικές σχεδιαστικές επιλογές, όπως η χρήση MLP επιπέδων αντί των γραμμικών, τα οποία προτείνονται για graph regression προβλήματα, θα συντελέσουν σε αποδοτικότερα μοντέλα. Επιπροσθέτως θα μπορούσε να μελετηθεί η χρήση περισσότερων εξωτερικών χαρακτηριστικών (οικονομικών και μη) στους κόμβους των συναλλαγών για να εξεταστεί εκτενέστερα ο βαθμός με τον οποίο επιδράνε στις προβλέψεις.
Σημαντικό θα ήταν επίσης, με τη μέθοδο που παρουσιάσαμε, να μελετηθεί και το πρόβλημα αύξησης ή μείωσης της τιμής του Bitcoin από τη σκοπιά της ταξινόμησης, για να γίνει σύγκριση της μεθόδου και ως προς τις μετρικές αυτού του είδους προβλήματος σε σχέση με την υπάρχουσα βιβλιογραφία.
Έντονου επιχειρηματικού ενδιαφέροντος θα ήταν η προσπάθεια χρήσης της μεθόδου της παρούσας δουλειάς σε δεδομένα πραγματικού χρόνου για να λαμβάνουμε προβλέψεις που θα αφορούν τωρινές τιμές του Bitcoin.
Τέλος, ενδιαφέρον θα παρουσίαζε η μελέτη περίπτωσης όπου το GNN θα λειτουργούσε με έναν ημιεπιβλεπόμενο τρόπο στους γράφους συναλλαγών, έτσι ώστε να παράξει embeddings τα οποία θα έμπαιναν ως είσοδοι σε άλλου τύπου μοντέλα νευρωνικών δικτύων.

Βιβλιογραφία

[1] "Αλγοριθμική Θεωρία Γραφημάτων" Σ. Νικολόπουλος, Λ. Γεωργιάδης, Λ. Παληός, Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα και Βοηθήματα - Αποθετήριο "Κάλλιπος"
[2] Rosen Kenneth H. Discrete Mathematics and Its Applications. Eighth ed. McGraw-Hill 2019.
[3] Phil Simon. Too Big to Ignore: The Business Case for Big Data. Wiley. pp. 89.
[4] Simon Haykin. Neural Networks and Learning Machines. Prentice Hall, 3rd Edition 2008.
[5] Russell, Stuart J. Artificial intelligence: a modern approach. New Delhi: Prentice-Hall of India. 2002.
[6] Deep Learning (Ian J. Goodfellow, Yoshua Bengio and Aaron Courville), MIT Press, 2016.
[7] Richard S. Sutton and Andrew G. Barto. 2018. Reinforcement Learning: An Introduction. A Bradford Book, Cambridge, MA, USA.
[8] E. Francesconi, P. Frasconi, M. Gori, S. Marinai, J. Sheng, G. Soda, and A. Sperduti, “Logo recognition by recursive neural networks,” in Lecture Notes in Computer Science — Graphics Recognition, K. Tombre and A. K. Chhabra, Eds. Berlin, Germany: Springer-Verlag, 1997.
[9] Weiwei Jiang, Jiayun Luo, Graph neural network for traffic forecasting: A survey, Expert Systems with Applications, Volume 207, 2022, 117921, ISSN 0957-4174.
[10] Fung, V., Zhang, J., Juarez, E. et al. Benchmarking graph neural networks for materials chemistry. npj Comput Mater 7, 84 (2021).
[11] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner and G. Monfardini, "The Graph Neural Network Model," in IEEE Transactions on Neural Networks, vol. 20, no. 1, pp. 61-80, Jan. 2009, doi: 10.1109/TNN.2008.2005605.
[12] T. N. Kipf and M. Welling, “Semi-Supervised Classification with Graph Convolutional Networks,” Sep. 2016
[13] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.
[14] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 1025–1035.
[15] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P. & Bengio, Y. (2017). Graph Attention Networks. 6th International Conference on Learning Representations, .
[16] Brody, Shaked and Alon, Uri and Yahav, Eran. (2021). How Attentive are Graph Attention Networks?
[17] S. Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash System,” 2008.

[18] Zhe (Alan) Wu. "Analyzing Blockchain and Bitcoin Transaction Data as Graph" Oracle Code (2018). FunkhausBerlin.
[19] Akcora, Cuneyt & Gel, Yulia & Kantarcioglu, Murat. (2017). Blockchain: A Graph Primer.
[20] Fleder, Michael & Kester, Michael & Pillai, Sudeep. (2015). Bitcoin Transaction Graph Analysis.
[21] S. Sharma and R. Sharma, "Forecasting Transactional Amount in Bitcoin Network Using Temporal GNN Approach," 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2020, pp. 478-485, doi: 10.1109/ASONAM49781.2020.9381363.
[22] Sharma, Aman & Bhatia, Ashutosh. (2020). Bitcoin's Blockchain Data Analytics: A Graph Theoretic Perspective.
[23] Ron, Dorit & Shamir, Adi. (2012). Quantitative Analysis of the Full Bitcoin Transaction Graph.
[24] C. G. Akcora, A. K. Dey, Y. R. Gel, and M. Kantarcioglu, “Forecasting bitcoin price with graph chainlets,” in Advances in Knowledge Discovery and Data Mining - 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings,Part III, ser. Lecture Notes in Computer Science, D. Q. Phung, V. S. Tseng, G. I. Webb, B. Ho, M. Ganji, and L. Rashidi, Eds., vol. 10939. Springer, 2018, pp. 765–776. [Online]. Available: https://doi.org/10.1007/978-3-319-93040-4 60
[25] Li, Xiao & Wu, Weili. (2020). A Blockchain Transaction Graph based Machine Learning Method for Bitcoin Price Prediction.
[26] J. S. Tharani, E. Y. A. Charles, Z. Hóu, M. Palaniswami and V. Muthukkumarasamy, "Graph Based Visualisation Techniques for Analysis of Blockchain Transactions," 2021 IEEE 46th Conference on Local Computer Networks (LCN), 2021, pp. 427-430, doi: 10.1109/LCN52139.2021.9524878.
[27] A. Politis, K. Doka and N. Koziris, "Ether Price Prediction Using Advanced Deep Learning Models," 2021 IEEE International Conference on Blockchain and Cryptocurrency (ICBC), 2021, pp. 1-3, doi: 10.1109/ICBC51069.2021.9461061.
[28] Livieris, I.E., Stavroyiannis, S., Iliadis, L. et al. Smoothing and stationarity enforcement framework for deep learning time-series forecasting. Neural Comput & Applic 33, 14021–14035 (2021). https://doi.org/10.1007/s00521-021-06043-1
[29] J. Luo, K. Ying, and J. Bai, “Savitzky-Golay smoothing and differentiation filter for even number data,” Signal Processing, vol. 85, no. 7, pp. 1429–1434, 2005, doi: 10.1016/j.sigpro.2005.02.002.
[30] Klinker, F. Exponential moving average versus moving exponential average. Math Semesterber 58, 97–107 (2011). https://doi.org/10.1007/s00591-010-0080-8
[31] Y. Zhai, A. Hsu, and S. K. Halgamuge, “Combining news and technical indicators in daily stock price trends prediction,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 4493 LNCS, no. PART 3, pp. 1087–1096, 2007, doi: 10.1007/978-3-540-72395-0_132.

[32] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques: Concepts And Techniques. Massachusetts: Morgan Kaufmann Publishers; 2011.
[33] Haykin SS. Neural Networks and Learning Machines. New Jersey: PearsonEducation Upper Saddle River; 2009
[34] Cao, Xi Hang & Stojkovic, Ivan & Obradovic, Zoran. (2016). A robust data scaling algorithm to improve classification accuracies in biomedical data. BMC bioinformatics. 17. 359. 10.1186/s12859-016-1236-x.
[35] I. M. Wirawan, T. Widiyaningtyas and M. M. Hasan, "Short Term Prediction on Bitcoin Price Using ARIMA Method," 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), 2019, pp. 260-265, doi: 10.1109/ISEMANTIC.2019.8884257.
[36] S. Roy, S. Nanjiba and A. Chakrabarty, "Bitcoin Price Forecasting Using Time Series Analysis," 2018 21st International Conference of Computer and Information Technology (ICCIT), 2018, pp. 1-5, doi: 10.1109/ICCITECHN.2018.8631923.
[37] A. Aggarwal, I. Gupta, N. Garg and A. Goel, "Deep Learning Approach to Determine the Impact of Socio-Economic Factors on Bitcoin Price Prediction," 2019 Twelfth International Conference on Contemporary Computing (IC3), 2019, pp. 1-5, doi: 10.1109/IC3.2019.8844928.
[38] L. Li, A. Arab, J. Liu, J. Liu and Z. Han, "Bitcoin Options Pricing Using LSTM-Based Prediction Model and Blockchain Statistics," 2019 IEEE International Conference on Blockchain (Blockchain), 2019, pp. 67-74, doi: 10.1109/Blockchain.2019.00018.
[39] Khedr, A, Arif, I, P V, PR, El-Bannany, M, Alhashmi, S, S, M. Cryptocurrency price prediction using traditional statistical and machine learning techniques: A survey. Intell Sys Acc Fin Mgmt. 2021; 28: 3– 34. https://doi.org/10.1002/isaf.1488
[40] D. R. Pant, P. Neupane, A. Poudel, A. K. Pokhrel and B. K. Lama, "Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis," 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), 2018, pp. 128-132, doi: 10.1109/CCCS.2018.8586824.
[41] S. Sridhar and S. Sanagavarapu, "Analysis and Prediction of Bitcoin Price using Bernoulli RBM-based Deep Belief Networks," 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), 2021, pp. 1-6, doi: 10.1109/INISTA52262.2021.9548422.
[42] Schulte, Maximilian & Eggert, Mathias. (2021). Predicting Hourly Bitcoin Prices Based on Long Short-Term Memory Neural Networks. 10.1007/978-3-030-86797-3_50.
[43] Jiang, X. (2020) Bitcoin Price Prediction Based on Deep Learning Methods. Journal of Mathematical Finance, 10, 132-139. doi: 10.4236/jmf.2020.101009.
[44] H. Kilimci, M. Yıldırım and Z. H. Kilimci, "The Prediction of Short-Term Bitcoin Dollar Rate (BTC/USDT) using Deep and Hybrid Deep Learning Techniques," 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2021, pp. 633-637, doi: 10.1109/ISMSIT52890.2021.9604741.

[45] Ciaian, P., Rajcaniova, M., & Kancs, D. A. (2016). The economics of BitCoin price formation. Applied Economics, 48(19), 1799–1815. https://doi.org/10.1080/00036846.2015.1109038
[46] Akcora, Cuneyt & Gel, Yulia & Kantarcioglu, Murat. (2017). Blockchain: A Graph Primer.
[47] M.W Gardner, S.R Dorling, Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences, Atmospheric Environment, Volume 32, Issues 14–15, 1998, Pages 2627-2636, ISSN 1352-2310, https://doi.org/10.1016/S1352-2310(97)00447-0.
[48] Johnson, N. & Vulimiri, P. & To, A. & Zhang, X. & Brice, C. & Kappes, B. & Stebner, Aaron. (2020). Machine Learning for Materials Developments in Metals Additive Manufacturing.
[49] Jure Leskovec, Stanford CS224W: Machine Learning with Graphs, http://cs224w.stanford.edu