ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Τμήμα Ψηφιακών Συστημάτων

Διπλωματική Εργασία

Τίτλος εργασίας	Νευρωνικά δίκτυα και Εφαρμογές αυτών για την πρόβλεψη τιμής Bitcoin
Όνομα:	Κοτταρίδης Αθανάσιος
Αρ. Μητρώου	me2016
Επιβλέπων καθηγητής	Φιλιππάκης Μιχαήλ
Ημερομηνία παράδοσης	Φεβρουάριος 2023

Περιεχόμενα

1. Abstract

Bitcoin price prediction has been an active area of research for a long time. The objective of this research project is to create a large-scale application that performs real-time forecasts to determine the short-term predictability of the Bitcoin in USD by machine learning techniques and sentiment analysis. Our goal is to apply sentiment analysis and supervised machine learning principles to the extracted reddit posts in order to analyze the correlation between bitcoin price movements and sentiments from reddit posts. We used latest technologies in order to create a scalable pipeline that retrieves, preprocess and stores Bitcoin prices every minute. Then we used Deep Learning and Neural networks in order to Predict Bitcoin price prediction across horizons ranging from 1 to 60 min. We analyzed the time series model prediction of bitcoin prices using univariate long short-term memory (LSTM) and multivariate LSTM, and compared both models against ARIMA model. The compression shows that LSTM with multi feature performs more accurate results.

2. Εισαγωγή

Η πρόβλεψη της τιμής του Bitcoin [1] αποτελεί ένα αρκετά διαδεδομένο πεδίο έρευνας, ειδικότερα μετά από την τεράστια άνοδο που παρουσιάστηκε στην τιμή του τα τελευταία χρόνια. Καθημερινά δημιουργούνται ολοένα και περισσότερα κρυπτονομίσματα εκ των οποίων κάποια καταφέρνουν να αναδειχτούν σε κορυφαία ενώ κάποια άλλα αποδεικνύονται απάτη. Όλα αυτά λοιπόν καθιστούν τα κρυπτονομίσματα ένα από τα δημοφιλέστεραθέματα συζήτησης με σχετικές αναφορές να παρουσιάζονται στις ειδήσεις στα social mediaαλλά και σε καθημερινές συζητήσεις που πραγματοποιούμε με τους γύρω μας. Παρόλα αυτά η αγορά των κρυπτονομισμάτων ποτέ δεν έπαψε να είναι μια ιδιαίτερα αβέβαιη αγορά, με αποτέλεσμα να προκαλεί μεγάλο ενδιαφέρον στους ερευνητές που θέλουν να προβλέψουν την τιμή του Bitcoin καθώς και να προσδιορίσουν ποια γεγονότα την επηρεάζουν άμεσα. Στόχος αυτής της έρευνας είναι η δημιουργία μιας κατανεμημένης εφαρμογής, που συλλέγει και επεξεργάζεται δεδομένα για την τιμή του Bitcoin καθώς και δημοσίευσης που έχουν αναρτηθεί σχετικά με αυτό στα social media, προκειμένου να πραγματοποιήσει βραχυπρόθεσμη πρόβλεψη της τιμής του σε USD (Αμερικάνικα Δολάρια), κάνοντας χρήση τεχνικών μηχανικής μάθησης (Machine Learning) και ανάλυσης συναισθήματος (Sentiment Analysis [2]).
Η ανάπτυξη της εργασίας πραγματοποιείται σε δυο σκέλη. Το πρώτο σκέλος αφορά την συλλογή και επεξεργασία των δεδομένων ενώ το δεύτερο σκέλος αφορά την εκπαίδευση μοντέλων μηχανικής μάθησης, πραγματοποίηση πρόβλεψης, καθώς και την αξιολόγηση της. Αρχικά κάθε λεπτό συλλέγουμε δεδομένα σχετικά με την τιμή του Bitcoin καθώς και διάφορα άλλα χαρακτηριστικά που την επηρεάζουν από πηγές δεδομένων που βρίσκονται στο διαδίκτυο. Μετατρέπουμε τα δεδομένα που συλλέξαμε σε μια μορφή εύκολη για αποθήκευση και τα εισάγουμε σε μια Mongo [3] βάση δεδομένων [4] που θα χρησιμοποιηθεί

ως Data Warehouse [4]. Έπειτα δημιουργούμε μια ροή δεδομένων μέσο της οποίας προωθούμε μετα-δεδομένα σχετικά με την τιμή του Bitcoin που μόλις συλλέξαμε σε επόμενα μηχανήματα μέσω ενός συστήματος KAFKA [5] που έχουμε εγκαταστήσει στον Server μας.Στη συνέχεια ένα δεύτερο μηχάνημα, καταναλώνει μηνύματα που παράγονται από την ροή και χρησιμοποιεί τα μετα-δεδομένα που αποστάλθηκαν προκειμένου να συλλέξει Redit posts σχετικά με το Bitcoin για ένα προκαθορισμένο χρονικό διάστημα και να πραγματοποιήσει Sentiment Analysis σε αυτά. Αφού ολοκληρωθεί η διαδικασία συλλογής και η συναισθηματική ανάλυση των δεδομένων, αποθηκεύουμε σε ένα νέο collection στην mongo τα δεδομένα που προέκυψαν από την ανάλυση φυσικής γλώσσας. Τέλος ενημερώνουμε μέσω της ροής δεδομένων ένα τρίτο μηχάνημα το οποίο επεξεργάζεται κατάλληλα ταδεδομένα που αφορούν την τιμή του bitcoin, εξάγει compound polarity για το συγκεκριμένοστιγμιότυπο της τιμής και τα αποθηκεύει στην βάση δεδομένων προκειμένου να χρησιμοποιηθούν για πρόβλεψη.
Αφού ολοκληρωθεί η συλλογή και η επεξεργασία των δεδομένων που περιγράψαμε παραπάνω, χρησιμοποιούμε τα δεδομένα που προέκυψαν προκειμένου να πραγματοποιήσουμε βραχυπρόθεσμη πρόβλεψη της τιμής του bitcoin. Σκοπός μας είναι να εξεταστεί η απόδοση των βαθιών νευρωνικών δικτύων (Deep Neural Networks[6]) στην πρόβλεψη χρονοσειρών της τιμής του Bitcoin. Για να πραγματοποιηθεί αυτό θα σχεδιαστούν και θα δοκιμαστούν ένα νευρωνικό δίκτυο LSTM[7] με ένα χαρακτηριστικό, καθώς και ένα νευρωνικό δίκτυο LSTM με πολλαπλά χαρακτηριστικά, τα οποία θα συγκριθούν με την απόδοση ενός μοντέλου ARIMA [8] οπού θα χρησιμοποιηθεί ως Benchmark για την αξιολόγηση τους.

3. Ανασκόπηση Πεδίου

3.1 Τι είναι το Bitcoin
Το Bitcoin [1] (BTC) είναι ένα κρυπτονόμισμα, δηλαδή ένα ψηφιακό νόμισμα το οποίο χρησιμοποιείται για την πραγματοποίηση συναλλαγών όπως οποιοδήποτε άλλο νόμισμα. Δημιουργήθηκε το 2009 από έναν Ανώνυμο developer ή μια ομάδα developers που ονομαζόταν Satoshi Nakamoto και από τότε έχει αποτελέσει πηγή έμπνευσης για πολλά αλλά κρυπτονομίσματα.

Η διαφορά του από τα τυπικά νομίσματα είναι ότι δεν έχει υλική υπόσταση για την έκδοση του και την πραγματοποίηση συναλλαγών δεν απαιτείται συμμέτοχη κάποιου τραπεζικού φορέα καθώς δεν υπόκειται στην νομισματική πολιτική των κρατών αλλά ούτε και στον έλεγχο κάποιου κεντρικού φορέα, αντιθέτως βασίζεται μόνο σε peer-to-peer λογισμικό και κρυπτογράφηση. Για αυτό τον λόγο ονομάζεται και Αποκεντρικοποιημένο (decentralized).

Τα κρυπτονομίσματα όπως και το bitcoin χρησιμοποιούν μια τεχνολογία που ονομάζεται blockchain. Το blockchain όπως λέει και το όνομα του είναι μια κατανεμημένη αλυσίδα από Μπλοκς. Δηλαδή μια κοινή βάση δεδομένων η οποία αποθηκεύει συναλλαγές με την μορφή blocks. Όταν ένα block συναλλαγής καταγράφεται στο σύστημα περιέχει πληροφορία απο το προηγούμενο μπλοκ. Το κάθε μπλοκ περιέχει τα δεδομένα και την

συναλλαγή η οποία έχει πιστοποιηθεί από τους validators ή αλλιώς miners, στη συνέχεια οι miners μπορούν να κρατήσουν το bitcoin να το πουλήσουν ή να το χρησιμοποιήσουν.

3.2 Τι είναι η μηχανική μάθηση και η πρόβλεψη τιμής Price Prediction:

Η πρόβλεψη της τιμής του Bitcoin αποτελεί ένα δύσκολο έργο, καθώς επηρεάζεται από ένα ευρύ φάσμα παραγόντων, όπως οι οικονομικές συνθήκες και η ζήτηση της αγοράς. Ως αποτέλεσμα, η ακρίβεια της πρόβλεψης των τιμών μπορεί να ποικίλλει σωματικά, για τον λόγο αυτό είναι σημαντικό να αξιολογούμε με προσοχή την επίδοση τον μοντέλων πρόβλεψης που χρησιμοποιούνται.
Ένας ακόμα πολύ σημαντικός παράγοντας που πρέπει να λάβουμε υπόψην μας κατά την διαδικασία της πρόβλεψης είναι ο χρονικός ορίζοντας κατά τον οποίον λαμβάνει χώρα η πρόβλεψη. Η μακροπρόθεσμη πρόβλεψη τιμών (Long-term price prediction) αναφέρεται στην πρόβλεψη τιμών για μεγάλο χρονικό διάστημα όπως αρκετές μέρες, μήνες ή χρόνια. Τέτοιου είδους προβλέψεις μπορούν να βυθίσουν στην λήψη στρατηγικών αποφάσεων όπως για παράδειγμα αν αξίζει να επενδύσω στο Bitcoin ή όχι. Από την άλλη η βραχυπρόθεσμη πρόβλεψη (Short-term price prediction) αναφέρεται στην πρόβλεψη τιμής για μικρότερο χρονικό διάστημα, όπως μερικά λεπτά, ώρες ή μέρες. Τέτοιου είδους προβλέψεις μπορούν να φανούν χρήσιμες στην λήψη αποφάσεων για το αν πρέπει να αγοράσω ή να πουλήσω Bitcoin.
Υπάρχει πληθώρα αλγορίθμων μηχανικής μάθησης, καθώς και μοντελοποιήσεων που μπορούν να χρησιμοποιηθούν τόσο για Long-term όσο και για Short-tem price prediction, όπως Νευρωνικά Δίκτυα, μοντέλα πρόβλεψης χρονοσειρών[9] καθώς και άλλα μοντέλα. Ωστόσο η επιλογή του καταλληλότερου μοντέλου θα εξαρτηθεί από τα χαρακτηριστικά των δεδομένων και τους στόχους της πρόβλεψης. Κατά την σύγκριση μεταξύ Long-term και Short-term price prediction είναι σημαντικό να λάβουμε υπόψην την αντιστάθμιση μεταξύ ακρίβειας και επικαιρότητας. Οι μακροπρόθεσμες προβλέψεις μπορεί να είναι πιο ακριβείς και να αποτυπώνουν καλύτερα τάσεις και μοτίβα καθώς διαθέτουν μεγαλύτερο πλήθος δεδομένων. Ωστόσο μπορεί να είναι λιγότερο επίκαιρες, καθώς βασίζονται σε ιστορικά δεδομένα και ενδέχεται να μην αντικατοπτρίζουν τις τρέχουσες συνθήκες της αγοράς.

Οι βραχυπρόθεσμες προβλέψεις, από την άλλη πλευρά, μπορεί να είναι λιγότερο ακριβείς και πιο ευαίσθητες, καθώς διαθέτουν λιγότερα δεδομένα για εργασία. Ωστόσο, μπορεί επίσης να είναι πιο επίκαιρες, καθώς βασίζονται σε πιο πρόσφατα δεδομένα και μπορούν να αντικατοπτρίζουν πιο στενά τις τρέχουσες συνθήκες της αγοράς. Συνολικά, η αξία της πρόβλεψης τιμής εξαρτάται από τις συγκεκριμένες ανάγκες και στόχους του χρήστη. Τόσο η μακροπρόθεσμη όσο και η βραχυπρόθεσμη πρόβλεψη μπορεί να είναι χρήσιμες σε διαφορετικά πλαίσια και είναι σημαντικό να αξιολογούνται προσεκτικά η απόδοση και οι περιορισμοί οποιουδήποτε μοντέλου πρόβλεψης.

3.3 Στατιστικές μέθοδοι πρόβλεψης χρονοσειρών:

Το ARIMA [8] είναι μια στατιστική μέθοδος που χρησιμοποιείται για πρόβλεψη χρονο-σειρών και αποτελείται από τρία επιμέρους μοντέλα.

AutoRegressive (AR): Το AutoRegration μοντέλο ή αλλιώς μοντέλο αυτοπαλινδρόμισης υποδεικνύει την συσχέτιση μιας παρατήρησης με ένα πλήθος προηγούμενων παρατηρήσεων που αποκαλούνται lagged observations. Στο AR μοντέλο η τρέχουσα τιμή της χρονοσειράς yt εκφράζετε ως γραμμικός συνδυασμός των παρελθοντικών τιμών και ενός όρου που αναπαριστά τα residuals e(t)

$$\boldsymbol{y_t = c + \varphi_1 * y_{t\#1} + \varphi_2 * y_{t\#2} + \cdots + \varphi_p * y_{t\#p}}$$

όπου:

- $\boldsymbol{y_t}$**:** είναι η τιμή της χρονοσειράς την χρονική στιγμή t
- $\boldsymbol{\mu}$**:** είναι συντελεστές του μοντέλου.
- $\boldsymbol{\varphi_1, \varphi_2, \dots, \varphi_p}$**:** είναι παράμετροι του μοντέλου που αποκαλούνται και συντελεστές αυτοπαλινδρόμισης.
- $\boldsymbol{y_{t\#1}, y_{t\#2}, \dots, y_{t\#p}}$**:** είναι οι lagged values της χρονοσειράς.
- $\mathrm{s}_{\boldsymbol{t}}$**:** είναι το σφάλμα ή αλλιώς residual την χρονική στιγμή t.

Integrated (I): Το Integration component του ARIMA αναφέρεται στην διαδικασία διαφοροποίησης της αρχικής χρονοσειράς προκειμένου να γίνει στάσιμη. Με τον όρο στάσιμη αναφερόμαστε σε μια χρονοσειρά που τα στατιστικά στοιχεία της όπως ο μέσος όρος (mean) και η διακύμανση (variance) είναι σταθερά ως προς τον χρόνο. Η διαφορά μιας χρονοσειράς συνήθως υπολογίζεται αφαιρώντας την προηγούμενη παρατήρηση από την τρέχουσα για παράδειγμα αν η αρχική σειρά συμβολίζεται με $\boldsymbol{y_t}$ η διαφορά πρώτης τάξεως συμβολίζεται με $\boldsymbol{y_t - y_{t\#1}}$**.**

Moving Average (MA): Το Moving Average Component του ARIMA περιγράφει την συσχέτιση μεταξύ της παρατήρησης που εξετάζεται και των σφαλμάτων που έχει υπολογίσει το μοντέλο MA από προηγούμενες περιόδους, τα σφάλματα αυτά συνήθως ονομάζονταιresiduals. Το μοντέλο MA εκφράζεται αλγεβρικά ως εξής:

$$\boldsymbol{y_t = \mu + \mathrm{s}_t + \theta_1 * \mathrm{s}_{t\#1} + \theta_2 * \mathrm{s}_{t\#2} + \cdots + \theta_q * \mathrm{s}_{t\#q}}$$

όπου:

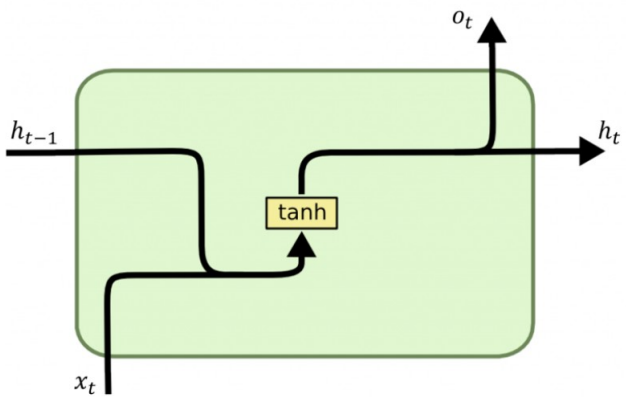
- $\boldsymbol{y_t}$**:** είναι η τιμή της χρονοσειράς την χρονική στιγμή t
- $\boldsymbol{\mu}$**:** είναι σταθερά.
- $\mathrm{s}_{\boldsymbol{t}}$**:** είναι το σφάλμα ή αλλιώς residual την χρονική στιγμή t.
- $\boldsymbol{\theta_1, \theta_2, \dots, \theta_q}$: είναι συντελεστές του μοντέλου.
- $\mathrm{s}_{\boldsymbol{t\#1}}, \mathrm{s}_{\boldsymbol{t\#2}}, \dots, \mathrm{s}_{\boldsymbol{t\#q}}$**:** είναι τα σφάλματα των παρελθοντικών στιγμών.

Η παραπάνω γραφική παράσταση εκφράζει την τιμή της χρονοσειράς την χρονική στιγμή t ως γραμμικό συνδυασμό των σφαλμάτων q περιόδων πίσω προσαυξημένα κατά μια σταθερά μ.

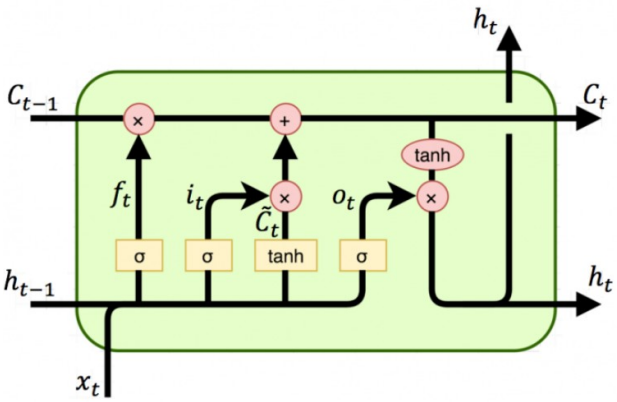
3.4 Τι είναι τα Νευρωνικά δίκτυα:

Τα νευρωνικά δίκτυα είναι ένας τύπος αλγόριθμου μηχανικής μάθησης που διαμορφώνονται σύμφωνα με τη δομή και τη λειτουργία του ανθρώπινου εγκεφάλου. Αποτελούνται από διασυνδεδεμένους «νευρώνες» που μπορούν να επεξεργάζονται και να μεταδίδουν πληροφορίες. Τα νευρωνικά δίκτυα είναι ιδιαίτερα χρήσιμα για εργασίες όπως η αναγνώριση εικόνας και ομιλίας, η επεξεργασία φυσικής γλώσσας, ακόμη και η πρόβλεψη χρονοσειρών. Τα νευρωνικά δίκτυα διαθέτουν πληθώρα αρχιτεκτονικών και παραλλαγών όπουη κάθε μια απο αυτές είναι σχεδιασμένη για συγκεκριμένους σκοπούς. Στην ανάλυση χρονοσειρών (η οποια θα μας απασχολήσει) είναι ιδιαίτερα διαδεδομένα τα RNN [10] (Recurrent Neural Networks) καθώς και οι παραλλαγές του LSTM [7] (Long short-term memory) και CRU[11] (Gated Recurrent Units).

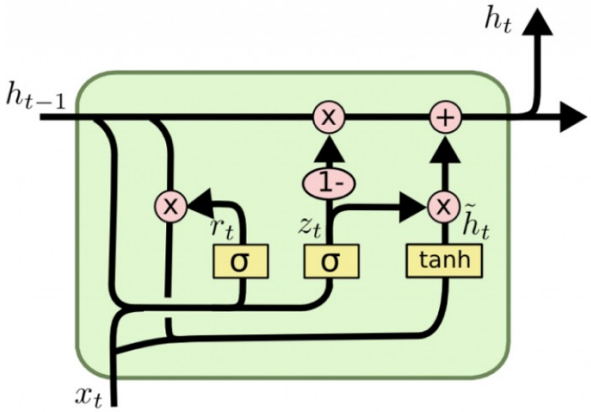
Τα RNN (επαναλαμβανόμενα νευρωνικά δίκτυα) είναι ένας τύπος νευρωνικών δικτύων που είναι ιδιαίτερα κατάλληλοι για την επεξεργασία διαδοχικών δεδομένων, όπως χρονοσειρές, φυσική γλώσσα και ομιλία. Τα RNN διαθέτουν ένα στοιχείο "μνήμης", το οποίοτους επιτρέπει να διατηρούν πληροφορίες από προηγούμενα βήματα της εκπαίδευσης και να το χρησιμοποιούν για να ενημερώνουν την επεξεργασία τους για το τρέχον χρονικό βήμα.Αυτό καθιστά τα RNN κατάλληλα για εργασίες πρόβλεψης τιμών όπως η τιμή Bitcoin, όπουη τρέχουσα τιμή παρουσιάζει κάποιου είδους εξάρτηση με τις προηγούμενες τιμές αλλά και με την τάση του.



Τα LSTM (Long-Short term memory) είναι ένας τύπος RNN που έχουν σχεδιαστεί ειδικά γιανα θυμούνται πληροφορίες για μεγαλύτερες χρονικές περιόδους. Το κάνουν αυτό χρησιμοποιώντας "πύλες" που ελέγχουν τη ροή πληροφοριών που εισέρχονται και εξέρχονται από το στοιχείο μνήμης του LSTM. Αυτό επιτρέπει στα LSTM να "ξεχνούν" άσχετες πληροφορίες και να "θυμούνται" σημαντικές πληροφορίες για μεγαλύτερα χρονικά διαστήματα. Τα LSTM χρησιμοποιούνται συνήθως για εργασίες όπως η μετάφραση γλώσσας , αναγνώριση ομιλίας και ανάλυση χρονοσειρών.



Τα GRU (Gated Recurrent Units) είναι ένας τύπος αρχιτεκτονικής νευρωνικών δικτύων που έχει σχεδιαστεί ειδικά για την επεξεργασία διαδοχικών δεδομένων με συνεχή τρόπο. Τα GRU είναι παρόμοια με τα LSTM καθώς διαθέτουν ένα στοιχείο μνήμης, αλλά δεν χρησιμοποιούν πύλες για τον έλεγχο της ροής πληροφοριών. Αντίθετα, χρησιμοποιούν μια συνεχή αναπαράσταση των δεδομένων, η οποία τους επιτρέπει να επεξεργάζονται τα δεδομένα με πιο ευέλικτο και αποτελεσματικό τρόπο. Οι μονάδες GRU είναι ιδιαίτερα κατάλληλες για εργασίες όπως η μετάφραση γλώσσας και η μοντελοποίηση γλώσσας, όπου μπορούν να αποτυπώσουν αποτελεσματικά τις πολύπλοκες σχέσεις μεταξύ λέξεων και φράσεων σε μια γλώσσα.



3.5 Τι είναι η Sentiment Analysis

Η ανάλυση συναισθήματος (Sentiment Analysis [2]) είναι μια περίπτωση χρήσης της Επεξεργασίας Φυσικής Γλώσσας (NLP) και εμπίπτει στην κατηγορία της ταξινόμησης κειμένου. Βασίζεται στην επεξεργασία φυσικής γλώσσας και αποσκοπεί στον εντοπισμό και

την εξαγωγή πληροφοριών από το κείμενο με στόχο να προσδιορίσει την στάση ή την γνώμη του συγγραφέα ή ομιλητή για ένα συγκεκριμένο θέμα. Η εφαρμογή τεχνικών συναισθηματικής ανάλυσης συναντάται σε διάφορα πεδία όπως:

- Social Media Monitoring for Brand Management: Οι εταιρίες μπορούν να χρησιμοποιούν Sentiment Analysis μέσω των Social Media για να προσδιορίζουν τηνδημοτικότητα και το αντίκτυπο που έχουν τα προϊόντα στους καταναλωτές. Αυτό τους βοηθάει στο να πάρουν αποφάσεις σχετικά με την προώθηση, την διαφήμιση και την ανάπτυξη του προϊόντος. Επίσης βοηθάει στον να αντιλαμβάνονται καλύτερα τις ανάγκες των καταναλωτών και να προσαρμόζουν καλύτερα τα προϊόντα τους.
- Product/Service Analysis: Αντίστοιχα με τα παραπάνω οι εταιρίες μπορούν να εφαρμόζουν Sentiment Analysis στις κριτικές πελατών για να δουν πόσο καλά τα πηγαίνει ένα προϊόν ή μια υπηρεσία στην αγορά και να λάβουν μελλοντικές αποφάσεις ανάλογα.
- Stock Price Prediction: Η πρόβλεψη εάν οι μετοχές μιας εταιρίας θα ανέβουν ή θα πέσουν είναι κρίσιμη για τους επενδυτές. Κάποιος μπορεί να προσδιορίσει το ίδιο πραγματοποιώντας ανάλυση συναισθήματος στους τίτλους ειδήσεων των άρθρων που περιέχουν το όνομα της εταιρίας. Εάν οι τίτλοι ειδήσεων που σχετίζονται με έναν συγκεκριμένο οργανισμό τυχαίνει να έχουν θετικό κλίμα οι τιμές των μετοχώντου θα πρέπει να ανέβουν και αντίστροφα.

Στο πλαίσιο των μέσων κοινωνικής δικτύωσης, η ανάλυση συναισθήματος μπορεί να χρησιμοποιηθεί για την κατανόηση του συναισθήματος των αναρτήσεων (posts) όπως τα tweets, τα reddit posts ή οι αναρτήσεις στο Facebook, και μπορεί να παρέχει πολύτιμες πληροφορίες για τις στάσεις και τις απόψεις των χρηστών.

Υπάρχει πληθώρα τεχνικών που μπορούν να χρησιμοποιηθούν προκειμένου να πραγματοποιήσουμε συναισθηματική ανάλυση σε αναρτήσεις μέσων κοινωνικής δικτύωσης. Μια από τις πιο διαδεδομένες είναι η εξαγωγή πολικότητας (Polarity Extraction) []. Η προσέγγιση αυτή χρησιμοποιεί αλγορίθμους προκειμένου να ταξινομήσει το κείμενο ως θετικό, αρνητικό ή ουδέτερο. Μια άλλη προσέγγιση είναι η χρήση συστημάτων βασισμένων σε κανόνες, τα οποία χρησιμοποιούν ένα σύνολο προκαθορισμένων κανόνων για να καθορίσουν το συναίσθημα του κειμένου.

Ανεξάρτητα από την τεχνική που χρησιμοποιείται, είναι σημαντικό να λαμβάνεται υπόψην το πλαίσιο στο οποίο γράφτηκε το κείμενο και το κοινό στο οποίο απευθύνεται η ερμηνεία των αποτελεσμάτων της ανάλυσης συναισθημάτων και της εξαγωγής πολικότητας. Οι αναρτήσεις στα μέσα κοινωνικής δικτύωσης συχνά περιέχουν αργκό, emoji και άλλη άτυπη γλώσσα, που μπορεί να κάνει πιο δύσκολη την ακριβή ανάλυσή τους. Επιπλέον, το συναίσθημα μιας ανάρτησης στα μέσα κοινωνικής δικτύωσης μπορεί να μην ευθυγραμμίζεται πάντα με το συναίσθημα του χρήστη, καθώς οι χρήστες μπορεί να χρησιμοποιούν ειρωνεία ή σαρκασμό στις αναρτήσεις τους.

3.6 Τι είναι Κατανεμημένα συστήματα:

Με τον όρο κατανεμημένα σύστημα (distributed System) αναφερόμαστε σε ένα δίκτυο υπολογιστών που συνεργάζονται για την επίτευξη ενός κοινού στόχου. Κάθε υπολογιστής στο σύστημα ονομάζεται κόμβος και κάθε κόμβος επικοινωνεί με άλλους για την ανταλλαγή πληροφοριών και τον συντονισμό των ενεργειών τους.

Τα κυριότερα χαρακτηριστικά που καθιστούν τα κατανεμημένα συστήματα πολύ σημαντικά και απαραίτητα είναι:

- **Decentralization:** Δεν υπάρχει κεντρική αρχή ή διακομιστής που να ελέγχει ολόκληρο το σύστημα. Αντίθετα, κάθε κόμβος είναι αυτόνομος και μπορεί να πάρει αποφάσεις μόνος του.
- **Κοινή χρήση πόρων:** Τα κατανεμημένα συστήματα επιτρέπουν στους κόμβους να μοιράζονται πόρους όπως δεδομένα, υπολογιστική ισχύ και αποθήκευση.
- **Ανοχή σφαλμάτων:** Τα κατανεμημένα συστήματα είναι σχεδιασμένα να συνεχίζουν να λειτουργούν ακόμη και αν ένας ή περισσότεροι κόμβοι αποτυγχάνουν. Αυτό επιτυγχάνεται μέσω περιττών αντιγράφων ασφαλείας και μηχανισμών ανίχνευσης σφαλμάτων και ανάκτησης.
- **Επεκτασιμότητα:** Τα κατανεμημένα συστήματα μπορούν εύκολα να κλιμακωθούν προσθέτοντας ή αφαιρώντας πόρους αντίστοιχα σύμφωνα με τις εκάστοτε απαιτήσεις της εφαρμογής.

Υπάρχει πληθώρα αρχιτεκτονικών κατανεμημένων συστημάτων όπως συστήματα Client-server όπου οι κόμβοι χωρίζονται σε πελάτες και διακομιστές. Οι πελάτες στέλνουν αιτήματα και οι διακομιστές τα εκτελούν. Peer-to-peer συστήματα όπου εκεί όλοι οι κόμβοι είναι ίσοι και μπορούν να στέλνουν και να λαμβάνουν αιτήματα. Υπολογιστικά συστήματα πλέγματος (Grid computing systems) όπου εκεί συνδυάζονται πολλαπλοί υπολογιστές προκειμένου να σχηματίσουν έναν υπολογιστικά ισχυρότερο υπολογιστή και τέλος συστήματα Νέφους (Cloud computing systems) όπου επιτρέπουν στους χρήστες να έχουν πρόσβαση και να χρησιμοποιούν απομακρυσμένους υπολογιστικούς πόρους μέσω του Διαδικτύου. Ωστόσο κατά την ανάπτυξη κατανεμημένων εφαρμογών υπάρχουν πολλές προκλήσεις στο σχεδιασμό και την εφαρμογή που κάποιος πρέπει να αντιμετωπίσει. Όπως για παράδειγμα: α) η διασφάλιση της συνέπειας του συστήματος, δηλαδή ότι όλοι οι κόμβοι έχουν την ίδια άποψη των δεδομένων η οποία μπορεί να είναι δύσκολη, ειδικά όταν οι κόμβοι μπορούν να ενημερώνουν τα δεδομένα ανεξάρτητα, β) η διασφάλιση ότι το σύστημα είναι πάντα διαθέσιμο και ικανό να επεξεργάζεται αιτήματα σαν διαδικασία μπορεί να είναι δύσκολη, ειδικά σε περίπτωση αστοχιών ή υψηλού φορτίου και γ) η διασφάλιση της ασφάλειας του συστήματος, δηλαδή ότι το σύστημα είναι ασφαλές και προστατευμένο από επιθέσεις.

Προκειμένου να κατασκευαστεί ένα κατανεμημένο σύστημα χρησιμοποιούνται διάφορα εργαλεία προκειμένου να πραγματοποιηθεί επικοινωνία μεταξύ των κόμβων για παράδειγμα:

- **Πρωτόκολλα δικτύωσης:** Πρωτόκολλα δικτύωσης όπως το TCP/IP και το HTTP χρησιμοποιούνται για να επιτρέπουν στους κόμβους να επικοινωνούν μεταξύ τους.
- **Κατανεμημένες βάσεις δεδομένων:** Οι κατανεμημένες βάσεις δεδομένων επιτρέπουν στους κόμβους να αποθηκεύουν και να ανακτούν δεδομένα από μια κεντρική τοποθεσία.
- **Κατανεμημένα συστήματα αρχείων:** Τα κατανεμημένα συστήματα αρχείων επιτρέπουν στους κόμβους να έχουν πρόσβαση και να μοιράζονται αρχεία μεταξύ τους.
- **Middleware:** Το Middleware είναι λογισμικό που βρίσκεται μεταξύ της εφαρμογής και του υποκείμενου υλικού και βοηθά στον συντονισμό της επικοινωνίας και της κοινής χρήσης πόρων μεταξύ των κόμβων.

3.7 Βιβλιογραφική ανασκόπηση:

Η πρόβλεψη της χρηματοοικονομικής αγοράς είναι ένας εξέχων κλάδος της χρηματοοικονομικής έρευνας και έχει μελετηθεί εκτενώς. Υπάρχουν αρκετές μελέτες σχετικά με την προβλεψιμότητα και την αποτελεσματικότητα των χρηματο-πιστωτικών αγορών, άλλες με θετικά και άλλες με αρνητικά αποτελέσματα. Το Bitcoin αποτελεί μια σχετικά νέα τεχνολογία και την δεδομένη χρονική στιγμή το ακριβότερο κρυπτονόμισμα. Βάση αυτού έχουν πραγματοποιηθεί αρκετές προσπάθειες προκειμένου να προσδιοριστεί η προβλε-

ψιμότητα του και να σχεδιαστούν μοντέλα που να μπορούν να προβλέψουν την τιμή του Bitcoin είτε μακροπρόθεσμα είτε βραχυπρόθεσμα.

Οι Amjad et al. χρησιμοποίησαν ιστορικά δεδομένα χρονοσειρών για την πρόβλεψη τιμής και την μεθοδολογία ανταλλαγής του Bitcoin [12] επίσης ο Garcia και οι συνεργάτες του παρατήρησαν ότι η αύξηση της πολικότητας της κοινής γνώμης και του όγκου των συναλλαγών προηγείται της αύξησης της τιμής του Bitcoin [13]. Οι Chen και Lazer [14] παρουσίασαν επενδυτικές τακτικές βασιζόμενες στην παρατήρηση και την ταξινόμηση των Tweeter Posts. Ο Go και οι συνεργάτες του εκπαίδευσαν ταξινομητές χρησιμοποιώντας Distant Supervice Learing και απέδειξε ότι υπάρχει υψηλή απόδοση ταξινόμησης [15], επίσης ο Go αναφέρεται και στο πολύ διαδεδομένο έγγραφο του Pang και των συνεργατών του [16] όπουδημιούργησαν ένα πρότυπο για ανάλυση αποφάσεων με χρήση μηχανικής μάθησης. Η ανάλυση του Pang θεωρείται μια απο τις πρωταρχικές προσπάθειες για εφαρμογή μηχανικής μάθησης στην ανάλυση αποφάσεων. Ορισμένες πιο πρόσφατες μελέτες επικεντρωθήκαν στην βραχυπρόθεσμη πρόβλεψη της τιμής του Bitcoin και πραγματοποίηση συναλλαγών εφαρμόζοντας τεχνικές βαθιάς μάθησης (deep-learning) όπως το RNN για την πρόβλεψη δεδομένων χρονοσειρών. Ο McNally [20] προβλέπει τη μεταβολή της τιμής του Bitcoin χρησιμοποιώντας τεχνικές μηχανικής μάθησης, όπως τα Recurrent Neural Networks (RNN) και Long-Short term memory (LSTM) και συγκρίνει τα αποτελέσματα με αυτά που λαμβάνονται χρησιμοποιώντας μοντέλα AutoRegressive integrated moving average (ARIMA).

Αντίστοιχα ο Jaquart και οι συνεργάτες του πραγματοποίησαν το 2019 μια μελέτη κατά την οποία αναλύουν τη βιβλιογραφία σχετικά με την πρόβλεψη της αγοράς bitcoin μέσω μηχανικής μάθησης, τα χαρακτηριστικά πρόβλεψης απόδοσης, τους ορίζοντες πρόβλεψης και τους τύπους πρόβλεψης. Απο την παραπάνω δημοσίευση παρατηρήθηκε ότι τα Recurrent neural networks παρουσίασαν αρκετά υποσχόμενα αποτελέσματα επίσης κατηγοριοποίησαν τα χαρακτηριστικά πρόβλεψης-απόδοσης στις παρακάτω: technical-based, blockchain-based, sentiment-/interest-based, and asset-based features. Τα τεχνικά χαρακτηριστικά περιγράφουν χαρακτηριστικά που σχετίζονται με ιστορικά δεδομένα της αγοράς bitcoin (π.χ. επιστροφές bitcoin). Τα χαρακτηριστικά που βασίζονται σε blockchain υποδηλώνουν χαρακτηριστικά που σχετίζονται με το blockchain bitcoin (π.χ., αριθμός συναλλαγών bitcoin). Οι λειτουργίες συναισθηματικής ανάλυσης και ενδιαφέροντος περιγράφουν χαρακτηριστικά που σχετίζονται με το συναίσθημα και τον όγκο αναζήτησης του bitcoin στο διαδίκτυο (π.χ., το συναίσθημα του bitcoin στο Twitter). Τα χαρακτηριστικά που βασίζονται σε περιουσιακά στοιχεία είναι χαρακτηριστικά που σχετίζονται με χρηματοπιστωτικές αγορές εκτός της αγοράς bitcoin (π.χ. επιστροφές χρυσού, αποδόσεις του δείκτη MSCI World) [19]. Σε μια άλλη ανάλυση του ο Jaquart και οι συνεργάτες του αναλύουν τη βραχυπρόθεσμη προβλεψιμότητα της αγοράς bitcoin. Τα αποτελέσματά τους τονίζουν τις δυνατότητες των επαναλαμβανόμενων νευρωνικών δικτύων για την πρόβλεψη της βραχυπρόθεσμης αγοράς bitcoin[20].

Τέλος ιδιαίτερο ενδιαφέρον παρουσιάζουν προσπάθειες που έχουν πραγματοποιηθεί και συνδυάζουν μεθόδους πρόβλεψης τιμής μέσω χρονοσειρών και τεχνικών βαθιάς μάθησης (deep-learing) μαζί με ανάλυση φυσικής γλώσσας (Natural Language Processing). Όπως παρουσιάζεται στην έρευνα του Raju και του Ali Mohammad Tarif πραγματοποιήθηκε απόπειρα δημιουργίας ενός πολλαπλών χαρακτηριστικών (multi feature) LSTM μοντέλου το οποίο μεταξύ άλλων χαρακτηριστικών λαμβάνει υπόψην και την ανάλυση φυσικής γλώσσας που έχει πραγματοποιηθεί σε Twitter posts προκειμένου να προβλέψει την τιμή του Bitcoin. Το παραπάνω μοντέλο δοκιμάστηκε συγκριτικά με single feature LSTM και ARIMA παρουσιάζοντας καλύτερα αποτελέσματα [17]. Αντίστοιχη μελέτη πραγματοποίησε και ο Jaquart με τους συνεργάτες του όπου χρησιμοποιώντας πολλαπλά χαρακτηριστικά όπως blockchain-based και sentiment/interest-based ανάλυσαν την προβλεψιμότητα του Bitcoin εφαρμόζοντας πληθώρα μοντέλων μηχανικής μάθησης για διαστήματα πρόβλεψης μεταξύ 1 και 60 λεπτών[18].

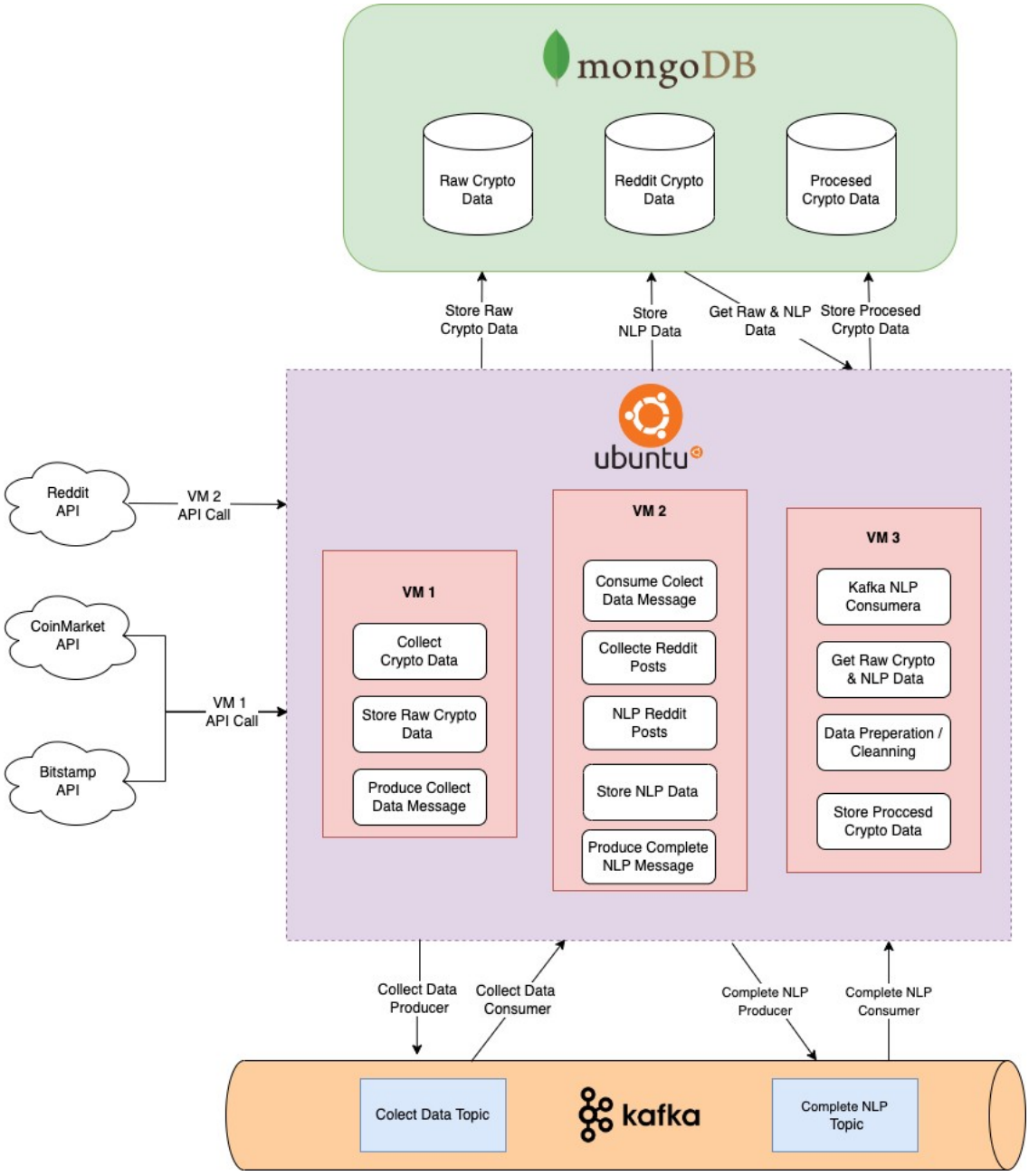
4. Συλλογή και Επεξεργασία Δεδομένων

Στην παρακάτω ενότητα θα περιγραφεί αναλυτικά η διαδικασία συλλογής και επεξεργασίας των δεδομένων που αφορούν την τιμή του Bitcoin μέσω ενός αυτοματοποιημένου και κατανεμημένου pipeline που δημιουργήσαμε. Πιο αναλυτικά, αρχικά για κάθε λεπτό συλλέγουμε δεδομένα σχετικά με την τιμή του Bitcoin καθώς και διάφορα άλλα χαρακτηριστικά που την επηρεάζουν από πηγές δεδομένων που βρίσκονται στο διαδίκτυο. Μετατρέπουμε τα δεδομένα που συλλέξαμε σε μια μορφή εύκολη για αποθήκευση και τα εισάγουμε σε μια Mongo βάση δεδομένων που θα χρησιμοποιηθεί ως Data Warehouse]. Έπειτα δημιουργούμε μια ροή δεδομένων μέσω της οποίας προωθούμε μετα-δεδομένα σε επόμενα μηχανήματα σχετικά με τα δεδομένα της τιμής του Bitcoin που μόλις συλλέξαμε μέσω ενός συστήματος KAFKA που έχουμε εγκαταστήσει στον Server μας. Στη συνέχεια ένα δεύτερο μηχάνημα που καταναλώνει μηνύματα που παράγονται από την ροή χρησιμοποιεί τα μετα-δεδομένα που αποστάλθηκαν προκειμένου να συλλέξει και να πραγματοποιήσει Sentiment Analysis σε Redit posts σχετικά με το Bitcoin για ένα προκαθορισμένο χρονικό διάστημα. Αφού ολοκληρωθεί η συλλογή και η συναισθηματική ανάλυση των δεδομένων εξάγουμε το polarity και αποθηκεύουμε σε ένα νέο collection στην mongo τα δεδομένα που προέκυψαν από την ανάλυση φυσικής γλώσσας. Τέλος ενημερώνουμε ένα τρίτο μηχάνημα μέσω της ροής δεδομένων το οποίο επεξεργάζεται κατάλληλα τα δεδομένα που αφορούν την τιμή του bitcoin, εξάγει compound polarity για το συγκεκριμένο στιγμιότυπο της τιμής και τα αποθηκεύει στην βάση δεδομένων προκειμένου να χρησιμοποιηθούν για πρόβλεψη.

4.1 Αρχιτεκτονική Συστήματος:

Το cluster αποτελείται από 3 μηχανήματα στα οποία είναι εγκατεστημένο λειτουργικό Linux Ubuntu Server LTS 14.04 τα οποία έχουν στηθεί στο οικοσύστημα του Cloud Okeanos[21], τα

μηχανήματα είναι συνδεδεμένα και επικοινωνούν μεταξύ τους μέσω private IP ενώ έχουμε στην διάθεση μας και μια public IP την οποία διαχειρίζεται ο Master Node προκειμένου να μπορεί να επικοινωνεί το cluster μας με το διαδίκτυο. Ποιο συγκεκριμένα το cluster μας αποτελείται:



Σχήμα 4.1 Αρχιτεκτονική κατανεμημένου συστήματος συλλογής δεδομένων

- **Master Node:** Είναι ο server ο οποίος διαθέτει την public ip και σε αυτόν είναι εγκατεστημένη η Βάση δεδομένων Mongo DB η οποία αποτελεί το Data warehouse του cluster μας. Επίσης σε αυτόν τον Node είναι εγκατεστημένος ο Zookeeper και αποτελεί τον μοναδικό Kafka Broker στην μέχρι τώρα αρχιτεκτονική μας. Εκεί αποθηκεύονται όλα τα δεδομένα που συλλέγονται από όλα τα στάδια της συλλογής και επεξεργασίας. Επίσης στο συγκεκριμένο μηχάνημα εκτελείται το service που είναι υπεύθυνο για την συλλογή δεδομένων σχετικά με την τιμή του Bitcoin από public APIs στο διαδίκτυο.
- **Node 1:** Σε αυτόν τον server εκτελείται το service που συλλέγει posts σχετικά με το Bitcoin τα οποία έχουν αναρτηθεί στο Reddit και πραγματοποιεί NLP σε αυτά προκειμένου να εξάγουμε την συναισθηματική βαρύτητα από αυτά τα posts και κατόπιν αποθηκεύει στο data warehouse τα αποτελέσματα της επεξεργασίας.
- **Node 2:** Ο server αυτός εκτελεί το service τελικής επεξεργασίας των δεδομένων. Το συγκεκριμένο service παρακολουθεί τις ροές δεδομένων που προέρχονται από τα προηγούμενα στάδια της συλλογής δεδομένων, και κατόπιν συνδυάζει την πληροφορία που συλλέχτηκε και την αποθηκεύει σε μορφή κατάλληλη για χρήση για εκπαίδευση και πρόβλεψη της τιμής του Bitcoin.

	Operating System	CPU	Ram	Storage
Master Node	Ubuntu Server LTS	8	8 GB	10 GB
Node 1	Ubuntu Server LTS	4	8 GB	5 GB
Node 2	Ubuntu Server LTS	4	8 GB	5GB

Πίνακας 4.1 πόροι υποδομής Okeanos

4.2 Τι είναι ο Apache Kafka.

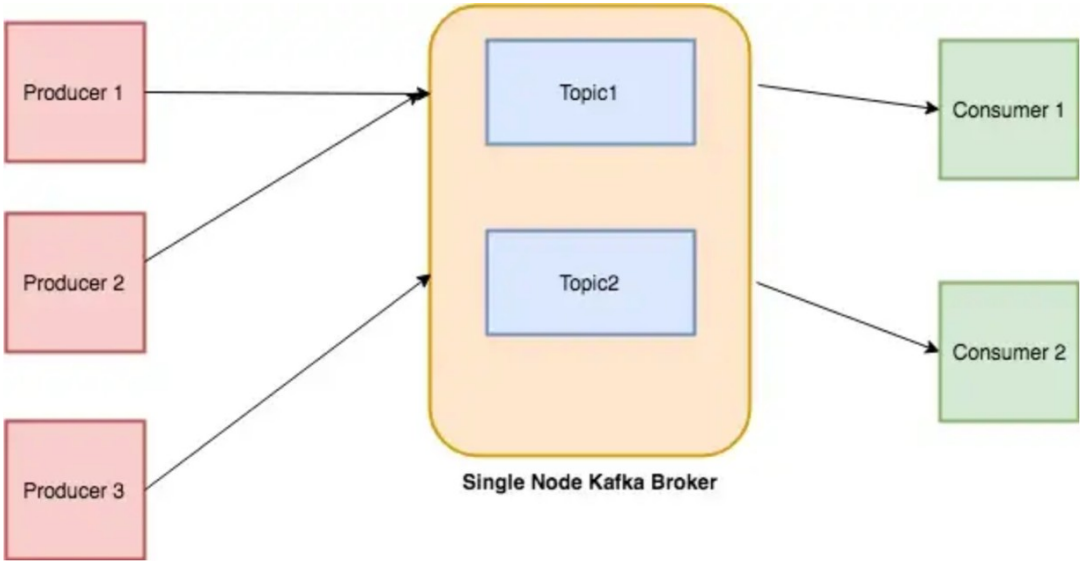
Ο Apache Kafka είναι μια open-source streaming πλατφόρμα η οποία αρχικά σχεδιάστηκε από το Linkedin και στην συνέχεια το 2011 αναπτύχθηκε από το Οργανισμό Apache ο οποίος την διέθεσε και σαν πλατφόρμα ανοιχτού κώδικα. Το Kafka είναι γραμμένο σε Java και Scala. Η εφαρμογή παρέχει μια πλατφόρμα υψηλής απόδοσης και χαμηλής καθυστέρησης (High-troughput low-latency) που χρησιμοποιείται για την διαχείριση ροών δεδομένων σε πραγματικό χρόνο. Διαθέτει επίσης ένα στρώμα αποθήκευσης (storage layer) λειτουργεί σαν μια επεκτάσιμη ουρά μηνυμάτων και έχει σχεδιαστεί σαν κατανεμημένο αρχείο καταγραφής, πράγμα που τον καθιστά ιδιαίτερα χρήσιμο για επεξεργασία ροών δεδομένων σε κατανεμημένα συστήματα.

Τα σύγχρονα δημοφιλή applications παράγουν ένα μεγάλο πλήθος από δεδομένα (logs) που συλλέγονται κατά την λειτουργία της εφαρμογής και για παράδειγμα αφορούν τον χρήστη όπως user activity, logins, clicks, αναζητήσεις, αντιδράσεις σε διαφημίσεις, αλλά και δεδομένα που αφορούν το σύστημα CPU, memory, disk utilization κ.α.

Το γεγονός που οδήγησε στην ανάπτυξη του Kafka απο το Linkedin ήταν η ανάγκη για επεξεργασία σε πραγματικό χρόνο των δεδομένων που παράγονται από τα logs της εφαρμογής προκειμένου να πραγματοποιήσουν πιο στοχευμένα recementation, διαφημίσεις και reporting καθώς και security operations όπως abuse detection και άλλα

4.2.1 Πως λειτουργεί ο Apache Kafka.

Ο Kafka αποτελείται από τις ακόλουθες βασικές έννοιες που απεικονίζονται στο σχήμα και θα αναλύσουμε και στη συνέχεια



Σχήμα 4.2 Αναπαράσταση Single node Kafka App

- **Topics:** Τα topics είναι στην πραγματικότητα ροές έγγραφων, κάθε μήνυμα που εισέρχεται στο σύστημα αντιστοιχεί σε ένα topic. Τα μηνύματα αποθηκεύονται στον Kafka σε μορφή key-value και εισέρχονται σε μια ακολουθία όπου ονομάζεται Offset. Το κάθε μήνυμα που αναρτάται σε ένα topic μπορεί να χρησιμοποιηθεί σαν είσοδος για επεξεργασία σε κάποιο άλλο σύστημα.
- **Producers:** Με τον όρο Producers αποκαλούμε τις εφαρμογές / services τα οποία είναι υπεύθυνα για την παραγωγή μηνυμάτων και την διάθεση τους σε κάποιο συγκεκριμένο topic της επιλογής τους.

- **Consumers:** Με τον όρο Consumer αναφερόμαστε στης εφαρμογές / services οι οποίες είναι συνδεδεμένες με κάποιο topic και παρακολουθούν για την εισαγωγή κάποιου μηνύματος σε αυτό προκειμένου να το καταναλώσουν. Σημείωση ένας Consumer αφού επεξεργαστεί ένα μήνυμα μπορεί με την σειρά του να εκτελέσει ρόλο producer προκειμένου να προωθήσει τα επεξεργασμένα δεδομένα σε κάποιο άλλο service μέσω κάποιου topic.
- **Brokers:** Οι Brokers είναι υπεύθυνοι για την ανταλλαγή μηνυμάτων σε ένα σύστημα Kafka. Ένα σύστημα Kafka μπορεί να αποτελείται από έναν ή περισσότερους Brokers. Στην περίπτωση που έχουμε μόνο έναν broker η υποδομή ονομάζεται stand alone kafka αλλιώς ονομάζεται kafka cluster.

4.2.2 Σενάρια Χρήσης Kafka.

Όπως αναφέρθηκε και παραπάνω ο Apache Kafka αποτελεί μια πλατφόρμα υψηλής απόδοσης και χαμηλής καθυστέρησης (High-troughput low-latency) που μπορεί να χρησιμοποιηθεί πολύ αποδοτικά στα παρακάτω business Use Cases.

- **Activity Monitoring:** μπορεί να χρησιμοποιεί δεδομένα που παράγονται από πραγματικούς senores ή δεδομένα που συλλέγονται από cloud εφαρμογές. Τα δεδομένα που συλλέγονται μπορούν μέσω του Kafka να αποθηκεύονται σε raw μορφή προκειμένου να χρησιμοποιηθούν και να επεξεργαστούν αργότερα.
- **Messaging:** Ένα πολύ διαδεδομένο σενάριο χρήσης του Kafka είναι για επικοινωνία μεταξύ services. Για παράδειγμα εξυπηρετεί σε μια εφαρμογή όπου έχουμε ενα service να συλλέγει δεδομένα και ένα να τα επεξεργάζεται.
- **Log Aggrication:** Συλλογή Logs από διάφορα συστήματα σε ένα κεντρικό σημείο προκειμένου να επεξεργαστούν σε δεύτερο χρόνο για error tracking ή για Analytics.
- **ETL:** Ο Kafka ταιριάζει ιδιαίτερα σε real time ETL processes. Δηλαδή στην συλλογή επεξεργασία και αποθήκευση δεδομένων σε πραγματικό χρόνο από διάφορες πήγες όπως internet APIs.
- **Database:** Ο Kafka μπορεί να χρησιμοποιηθεί και σαν Βάση δεδομένων (Database) προκειμένου να αποθηκεύσει κάποια δεδομένα σε raw μορφή λόγο της ταχύτατης εγγραφής δεδομένων σε αυτόν. Δεν είναι ιδιαίτερα ευνοϊκός για αναζήτηση τον δεδομένων.

4.2.3 Χρήση Kafka στην Εφαρμογή.

Στην εφαρμογή μας θα χρησιμοποιήσουμε τον Kafka προκειμένου να πραγματοποιήσουμε ETL διεργασίες συλλογής και επεξεργασίας δεδομένων σχετικά με το Bitcoin που

συλλέγονται από διάφορες online πηγές όπως Bitcoin Tracking Price APIs και το Reddit. Επίσης θα χρησιμοποιηθεί και ως messaging μηχανισμός προκειμένου να ενημερώνονται services που πραγματοποιούνε επιμέρους λειτουργίες κατά την συλλογή και επεξεργασία των δεδομένων με σκοπό να αποθηκευτούν στην Mongo warehouse σε μια μορφή ευνοϊκή για χρήση και πρόβλεψη.

Πιο αναλυτικά έχει αναπτυχθεί ένα single Broker Kafka application στο οποίο υπάρχουν δυο topics:

1. **Collect Crypto Data Topic:** Στο συγκεκριμένο topic τα μηνύματα παράγονται από το service που είναι υπεύθυνο για την συλλογή δεδομένων που αφορούν το Bitcoin (producer) και καταναλώνονται απο το service που είναι υπεύθυνο για να πραγματοποιεί NLP στα posts που έχουν αναρτηθεί στο Reddit σχετικά με το Bitcoin (consumer). Τα μηνύματα που αναρτώνται στο συγκεκριμένο topic περιέχουν το id της έγγραφης τον δεδομένων που παράχθηκαν από τον producer μαζί με το timestamp εισαγωγής στην βάση δεδομένων.
2. **Complete NLP Topic:** Στο συγκεκριμένο topic τα μηνύματα αναρτώνται από toservice που πραγματοποιεί NLP ανάλυση (producer) και καταναλώνονται απο το service που πραγματοποιεί την τελική επεξεργασία των δεδομένων προκειμένου να αποθηκευτούν σε μορφή έτοιμη για πρόβλεψη (consumer). Το payload των παραπάνω μηνυμάτων περιέχει το id της εγγραφής του στιγμιότυπου του bitcoin το timstamp κατά το οποίο αυτό εισαχθεί στην βάση καθώς και το reddit compound polarity που παράχθηκε από τον producer.

4.3 Τι είναι η Mongo DB.

Η MongoDB είναι μια δωρεάν και ανοιχτού κώδικα μη σχεσιακή (NoSQL) βάση δεδομένων στην οποία αποθηκεύονται δεδομένα σε μια μορφή παρόμοια με το JSON η οποία ονομάζετε BSON (Binary JSON). Είναι σχεδιασμένη προκειμένου να διαχειρίζεται μεγάλο όγκο δεδομένων ενώ παράλληλα να υποστηρίζει υψηλή απόδοση, οριζόντια κλιμάκωση (horizontal scaling) και συνέπεια των δεδομένων. Ένα από τα κυριότερα χαρακτηριστικά της MongoDB είναι η ευελιξία, καθώς είναι μια βάση δεδομένων σχεδιασμένη να χειρίζεται έγγραφα (Document Database) και δεν απαιτεί σταθερό σχήμα. Αυτό σημαίνει ότι κάθε έγγραφο σε μια συλλογή μπορεί να έχειδιαφορετικό σύνολο πεδίων και οι τύποι δεδομένων αυτών των πεδίων μπορεί να διαφέρουν από έγγραφο σε έγγραφο. Αυτό διευκολύνει την αποθήκευση και την εργασία με μη δομημένα δεδομένα.

4.3.1 Πως λειτουργεί η MongoDB.

Όπως αναφέραμε και στην ενότητα 4.2.1 η MongoDB είναι μια μη σχεσιακή βάση δεδομένων η οποία αποθηκεύει τα δεδομένα χρησιμοποιώντας JSON μορφής έγγραφα. Πιο αναλυτικά απαρτίζεται από τις συγκεκριμένες έννοιες:

1. **Database:** Με τον όρο Database στην MongoDB αναφερόμαστε σε μία δομή η οποία αποτελείται από collections. Ένας MongoDB server μπορεί να υποστηρίξει πολλαπλές databases.
2. **Collection:** Με τον όρο Collection στην MongoDB αναφερόμαστε σε μια συλλογή που αποτελείται από έγγραφα. Θα μπορούσε εννοιολογικά να παρομοιαστεί με την έννοια του πινάκα στις σχεσιακές βάσεις δεδομένων ωστόσο δεν είναι το ίδιο διότι τα collections δεν απαιτούν κάποιο schema. Τα Documents εντός ενός Collection μπορούν να έχουν διαφορετικά fields. Παρόλο που δεν είναι υποχρεωτικό θεωρείτε καλή πρακτική, όλα τα έγγραφα σε μια συλλογή έχουν παρόμοιο ή σχετικό σκοπό.
3. **Document:** Τα Document στην MongoDB έχουν δομή που μοιάζει με JSON αρχεία και μπορούν να παρομοιαστούν με τα rows (έγγραφες) στις σχεσιακές βάσεις δεδομένων. Ωστόσο τα Documents έχουν δυναμικό σχήμα, που σημαίνει ότι τα Documents σε ένα Collection μπορούν να έχουν διαφορετικά πεδία. Τα πεδία σε ένα έγγραφο μπορεί να διαφέρουν από το ένα έγγραφο στο άλλο.

4.3.2 Σενάρια Χρήσης MongoDB

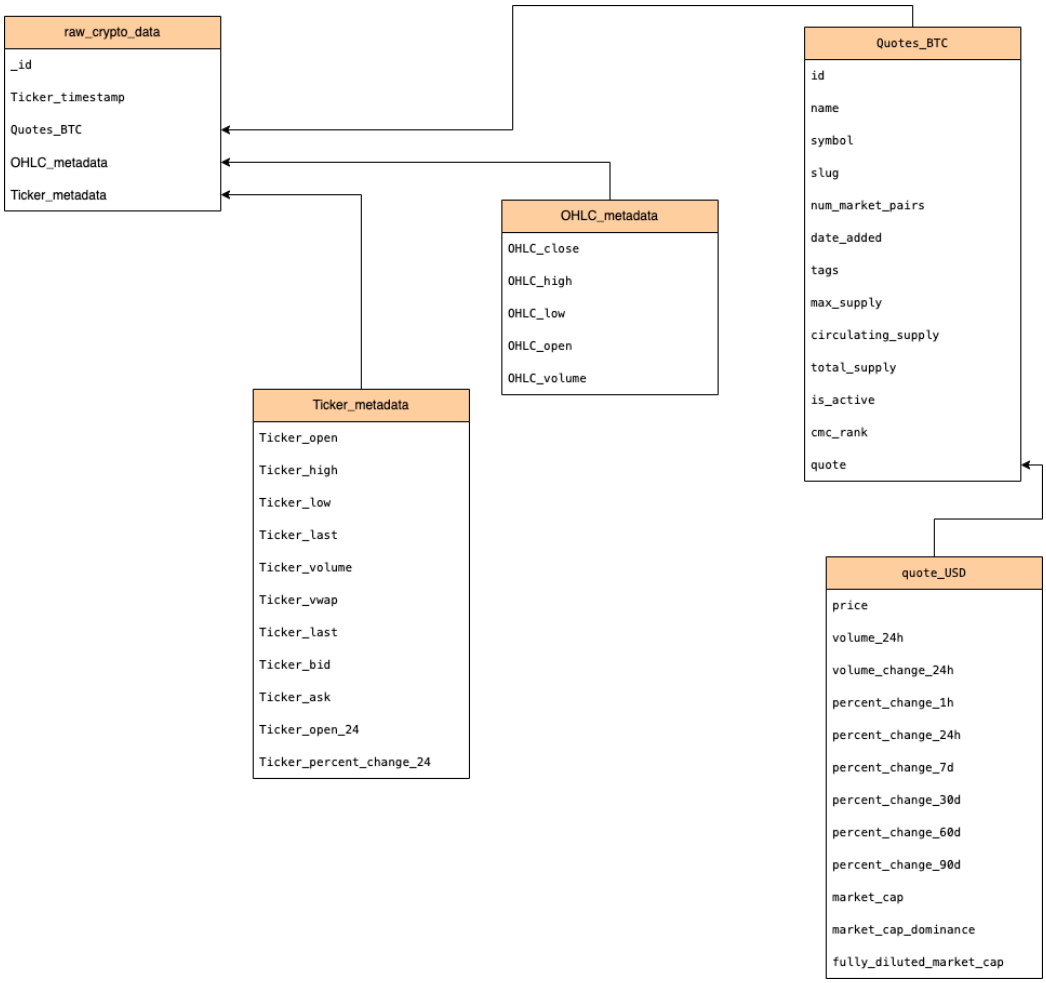
Ακολουθούν μερικά Σενάρια Χρήσης όπου ταιριάζει η χρήση της MongoDB:

- Αποθήκευση μεγάλου όγκου δεδομένων που δεν ταιριάζουν καλά σε μια δομή πίνακα, όπως δεδομένα συμπεριφοράς πελατών ή αναρτήσεις στα μέσα κοινωνικής δικτύωσης.
- Αναλύσεις και αναφορές σε πραγματικό χρόνο, καθώς το MongoDB μπορεί να χειριστεί μεγάλες ποσότητες δεδομένων και παρέχει γρήγορη απόδοσηανάγνωσης και εγγραφής.
- Συστήματα διαχείρισης περιεχομένου και πλατφόρμες blogging, καθώς η MongoDB μπορεί να αποθηκεύει και να οργανώνει μεγάλες ποσότητεςδεδομένων με ευέλικτο τρόπο.
- Εφαρμογές για κινητές συσκευές και web που απαιτούν συγχρονισμό δεδομένων σε πραγματικό χρόνο μεταξύ συσκευών.

4.3.3 Χρήση Mongo DB στην Εφαρμογή

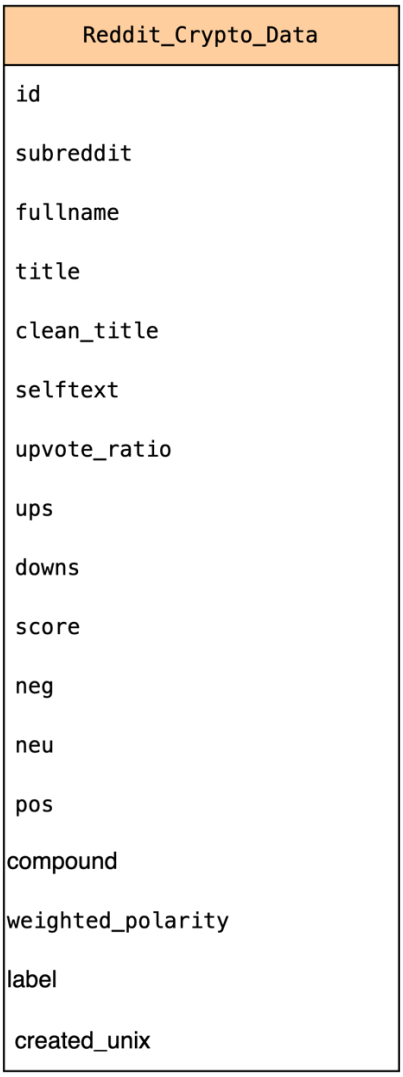
Η MongoDB στην εφαρμογή μας χρησιμοποιείται ως Data warehouse προκειμένου να αποθηκεύουμε τα δεδομένα που αφορούν το crypto πριν και μετά την επεξεργασία καθώς και τα δεδομένα που προκύπτουν από την ανάλυση συναισθηματικού περιεχομένου στα Reddit posts πιο αναλυτικά έχουν δημιουργηθεί τα παρακάτω collections:

1. **Raw Crypto Data Collection:** Αυτό το collection περιλαμβάνει όλα τα δεδομένα σε ανεπεξέργαστη μορφή που συλλέχθηκαν από τρία διαφορετικά API. Στο ακόλουθο σχήμα αναπαριστάται αναλυτικά η δομή των εγγράφων του Collection καθώς και των αμολημένων εγγράφων που περιέχονται σε αυτό.



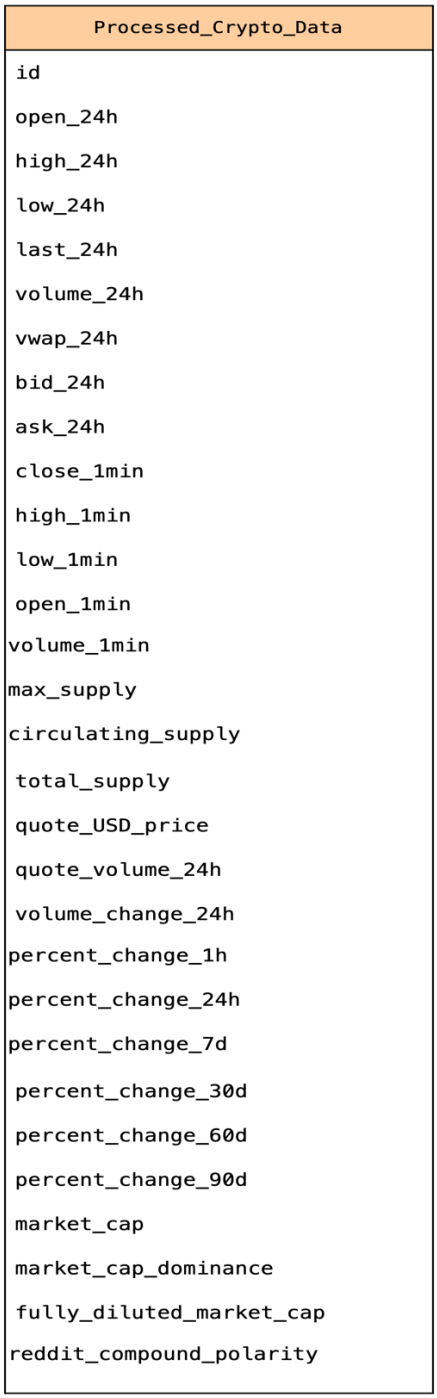
Σχήμα 4.3 Διάγραμμα μοντέλου Raw Crypto Data Collection

2. **Reddit Crypto Data Collection:** Το συγκεκριμένο collection περιέχει όλα τα Reddit posts που συλλέχτηκαν μαζί με τα δεδομένα που προέκυψαν από την ανάλυση συναισθηματικού περιεχομένου. Στο ακόλουθο σχήμα αναπαριστάται αναλυτικά η δομή των εγγράφων του Collection.



Σχήμα 4.4 Διάγραμμα μοντέλου Reddit Crypto Data Collection

3. **Processed Crypto Data:** Τέλος σε αυτό το Collection αποθηκεύονται τα επεξεργασμένα δεδομένα στην τελική τους μορφή και έτοιμα για πρόβλεψη. Στο ακόλουθο σχήμα αναπαριστάται αναλυτικά η δομή των εγγράφων του Collection.



Σχήμα 4.5 Διάγραμμα μοντέλου Processed Crypto Data Collection

4.4 Συλλογή Δεδομένων Bitcoin σε πραγματικό χρόνο

Όπως αναλύσαμε και σε προηγούμενα κεφάλαια η πρόβλεψη της τιμής του bitcoin αποτελεί μια εξαιρετικά δύσκολη και σύνθετη διαδικασία. Προκειμένου να κατα-φέρουμε να πραγματοποιήσουμε όσο το δυνατόν πιο ακριβείς προβλέψεις χρεια-ζόμαστε πληθώρα δεδομένων από έγκυρες πηγές, επίσης δεδομένου ότι στόχος της εργασίας είναι η βραχυπρόθεσμη πρόβλεψη της τιμής του bitcoin χρειαζόμαστε πληροφορία ανά αρκετά τακτικά χρονικά διαστήματα.

Για τον λόγο αυτό δημιουργήσαμε ένα αυτοματοποιημένο service στο cloud το οποίο είναι προγραμματισμένο να εκτελείται κάθε ένα λεπτό και να συλλέγει δεδομένα σχετικάμε την τιμή του bitcoin, καθώς και χαρακτηριστικών που το επηρεάζουν από διάφορες πήγες στο διαδίκτυο. Για την συλλογή δεδομένων σε πραγματικό χρόνο χρησιμο-ποιήθηκαν τρεις διαφορετικές διαδικτυακές πήγες δεδομένων όπου απο την κάθε μία από αυτές εξάγουμε διαφορετικές πληροφορίες σχετικά με την τιμή του bitcoin. Πιο αναλυτικά χρησιμοποιήθηκαν τα παρακάτω APIs.

- **CoinMarketCap Cryptocurrency V1 Quotes Latest:**
  Το CoinMarketCap[22] είναι μία σελίδα η οποία παρέχει πληροφορίες σχετικά με τις τιμές όλων των κρυπτονομισμάτων καθώς με χρήσιμες πληροφορίες σχετικά με αυτά, διαθέτει επίσης ένα δωρεάν API το οποίο επιστρέφει την τελευταία τιμή της αγοράς για ένα ή περισσότερα κρυπτονομίσματα. Στην υλοποίησή μας λαμβάνουμε προσφορές μόνο για το κρυπτονόμισμα Bitcoin. Το Quotes Latest API μας παρέχει τα ακόλουθα χρήσιμα δεδομένα:

Χαρακτηρίστηκα	**Ορισμός**
price	Τιμή στο ζητούμενο συνάλλαγμα (στην περί-πτωση μας USD)
circulating_supply	Ο κατά προσέγγιση αριθμός νομισμάτων που κυκλοφορούν για αυτό το κρυπτονόμισμα.
total_supply	Η κατά προσέγγιση συνολική ποσότητα κερμάτων που υπάρχουν αυτήν τη στιγμή (μείον τυχόν νομίσματα που έχουν καεί επαληθευμένα).
max_supply	Το αναμενόμενο μέγιστο όριο νομισμάτων που θα είναι ποτέ διαθέσιμο για αυτό το κρυπ-τονόμισμα.
volume_24h	Τον όγκο των συναλλαγών που πραγματοποιή-θηκαν τις τελευτές 24 ώρες .
volume_change_24h	24ωρη αλλαγή στον καθορισμένο όγκο νομισμά-των.
fully_diluted_market_cap	Πλήρως μειωμένη κεφαλαιοποίηση αγοράς στο καθορισμένο νόμισμα.
market_cap	Κεφάλαιο αγοράς στο καθορισμένο νόμισμα.
market_cap_dominance	Κυριαρχία κεφαλαίου αγοράς στο καθορισμένο νόμισμα.

percent_change_1h	1 ώρα αλλαγή στο καθορισμένο νόμισμα.
percent_change_24h	24ωρη αλλαγή στο καθορισμένο νόμισμα.
percent_change_7d	Αλλαγή 7 ημερών στο καθορισμένο νόμισμα.
percent_change_30d	Αλλαγή 30 ημερών στο καθορισμένο νόμισμα.
last_updated	Χρονική σήμανση (ISO 8601) του χρόνου αναφοράς της τρέχουσας αξίας του νομίσματος μετατροπής.

Πίνακας 4.2 Πίνακας χαρακτηριστικών CoinMarketCap
Cryptocurrency V1 Quotes Latest

- **Bitstamp API OHLC Data:** Το Bitstamp[23] αντίστοιχα είναι μια εφαρμογή μέσω της οποίας μπορείς να παρακολουθείς τις τιμές των κρυπτονομισμάτων καθώς επίσης και να πραγματοποιείς αγοραπωλησίες με αυτά. Το Bitstamp διαθέτει και αυτό δωρεάν API από το οποίο χρησιμοποιούμε δύο endpoints. Το Bitstamp OHLC API παρέχει δεδομένα όπως Open, High, Low και Close για καθορισμένο χρονικό διάστημα. Ένα γράφημα OHLC είναι ένας τύπος ραβδωτού γραφήματος που δείχνει τις τιμές ανοίγματος, κλεισίματος καθώς και την μέγιστη και την ελάχιστη τιμή για κάθε περίοδο. Τα γραφήματα OHLC είναι χρήσιμα καθώς δείχνουν τα τέσσερα κύρια σημεία μεταβολής της τιμής δεδομένου ενός χρονικού διαστήματος, με την τιμή κλεισίματος να θεωρείται η πιο σημαντική από πολλούς εμπόρους. Το συγκεκριμένο endpoint δέχεται ως παράμετρο την τιμή step (στην περίπτωση μας 60 seconds) προκειμένου να μας επιστρέψει τις ζητούμενες OHCL τιμές από την χρονική στιγμή που πραγματοποιήθηκε το αίτημα.

Features	**Definitions**
pair	Trading pair
high	Price high
timestamp	Unix timestamp date and time
volume	Volume
low	Price low
close	Closing price
open	Opening price

Πίνακας 4.3 Πίνακας χαρακτηριστικών
Bitstamp API OHLC Data

- **Bitstamp API Ticker Data:** Το Ticker endpoint είναι και αυτό ένα endpoint του Bitstamp API και μας παρέχει δεδομένα OHLC για τις τελευταίες 24 ώρες και μερικές επιπλέον πληροφορίες σχετικά με τη μέση σταθμισμένη τιμή προσφοράς, ζήτησης και όγκου. Πιο αναλυτικά στον επόμενο πίνακα παρουσιάζονται όλες οι τιμές που συλλέγουμε από το Ticker endpoint

Features	Definitions
last	Last price in counter currency.
high	Last 24 hours price high.
timestamp	Unix timestamp date and time
volume	Last 24 hours volume.
low	Last 24 hours price low.
open	First price of the day.
vwap	Last 24 hours volume weighted average price.
bid	Highest buy order.

Πίνακας 4.4 Πίνακας χαρακτηριστικών
Bitstamp API Ticker Data

Αφού ολοκληρώσουμε την συλλογή των δεδομένων που παρουσιάσαμε το service τροποποιεί τα δεδομένα κατάλληλα προκειμένου να είναι εύκολη η αποθήκευση τους και τα εισάγει στο Raw Crypto Data Collection που περιγράψαμε στην ενότητα 4.2.2 και αναπαριστάται στο σχήμα 4.3. Τέλος παράγετε ένα kafka event το οποίο αναρτάται στο Collect Crypto Data Topic (βλ. κεφάλαιο 4.1.4) προκειμένου να ξεκινήσει η διαδικασία εξαγωγής συναισθηματικού περιεχομένου από τα Reddit posts.

4.5 Ανάλυση συναισθήματος σε αναρτήσεις Bitcoin.

Τα μέσα κοινωνικής δικτύωσης έχουν πλέον σημαντική θέση στην καθημερινότητα μας, χρησιμοποιούνται από εκατομμύρια χρήστες οι οποίοι καθημερινά αναρτούν σκέψεις και σχολιάζουν διάφορα θέματα της επικαιρότητας, έχοντας άλλες φορές θετικό και άλλες φορές αρνητικό χαρακτήρα. Στόχος μας είναι να εξάγουμε δεδομένα σχετικά με τις αναρτήσεις που αφορούν το bitcoin και να εφαρμόσουμε Sentiment Analysis προκειμένουνα ελέγξουμε αν αυτές έχουν θετική ή αρνητική αναφορά για το bitcoin. Με τον τρόπο αυτόν μπορούμε να ελέγξουμε κατά πόσο η κοινή γνώμη επηρεάζει την τιμή του bitcoin και κατ' επέκταση να την χρησιμοποιήσουμε ως χαρακτηριστικό πρόβλεψης.

Για την εξαγωγή δεδομένων χρησιμοποιήθηκε η πλατφόρμα κοινωνικής δικτύωσης Reddit[24]η οποία χρησιμοποιείται σε μεγάλο βαθμό από κοινό που ασχολείται με θέματα τεχνολογίας και επιστήμης. Μέσω του Reddit εξάγαμε αναρτήσεις σχετικά με το bitcoin ωστόσο επειδή η γραφή στα μέσα κοινωνικής δικτύωσης δεν είναι πολύ τυπική χρειάζεται να εφαρμόσουμε κάποιο είδους καθαρισμό των δεδομένων πριν προχωρήσουμε σε ανάλυση συναισθηματικού περιεχομένου. Πιο αναλυτικά αφαιρούμε τα URL τα οποία μπορεί να περιέχονται μέσα στο κείμενο της ανάρτησης δεδομένου ότι δεν προσφέρουν κάποια χρήσιμη πληροφορία στην ανάλυση συναισθηματικού περιεχομένου. Έπειτα αφαιρούμε όλα τα emoji, τους single characters. Δηλαδή σκέτα γράμματα που δεν προσδίδουν κάποιο ιδιαίτερο νόημα στην ανάλυση και αντικαθίσταμαι τα πολλαπλά κενά με ένα κενό οπού

υπάρχουν. Τέλος αφαιρούμε όλα τα post που κατέληξαν να έχουν null τιμή μετά από την διαδικασία καθαρισμού.
Στην συνέχεια προχωράμε στην εφαρμογή αλγορίθμων μηχανικής μάθησης προκειμένου να πραγματοποιήσουμε συναισθηματική ανάλυση στα posts και να εξάγουμε το polarity τους χρησιμοποιώντας τον VADER[25] Analyzer που περιέχεται μέσα στην βιβλιοθήκη NLTK[26] της Python. Ο VADER (Valence Aware Dictionary for Sentiment Reasoning) Analyzer είναι ένα προ εκπαιδευμένο μοντέλο που χρησιμοποιείται για τον προσδιορισμό του συναισθήματος κειμένου, χρησιμοποιείται για τον προσδιορισμό του polarity (positive/negative) άλλα και της έντασης του συναισθήματος. Η συναισθηματική ανάλυση του VADER βασίζεται σε ένα λεξικό που αντιστοιχίζει ένταση συναισθημάτων σε χαρακτηριστικά λέξεων, με τον τρόπο αυτόν προσδιορίζει το sentiment score. Το sentiment score ενός κειμένου μπορεί να ληφθεί συνοψίζοντας την ένταση κάθε λέξης στο κείμενο. Για παράδειγμα λέξεις όπως «αγάπη», «απολαύστε», «ευτυχισμένος», «μου αρέσει» όλες μεταφέρουν ένα θετικό συναίσθημα. Επίσης ο VADER είναι αρκετά έξυπνος για να κατανοήσει το βασικό πλαίσιο αυτών των λέξεων, όπως το "δεν αγάπησα" ως αρνητική δήλωση. Κατανοεί επίσης την έμφαση των κεφαλαίων και των σημείων στίξης, όπως το "ENJOY". Ο SentimentIntensityAnalyzer() του VADER λαμβάνει ως είσοδο το κείμενο στο οποίο θέλουμε να πραγματοποιήσουμε ανάλυση συναισθηματικού περιεχομένου και μας επιστρέφει τέσσερα χαρακτηριστικά.

- **negative:** το ποσοστό αρνητικού περιεχομένου του κείμενου όπου λαμβάνει τιμές από 0 ως 1
- **positive:** Το ποσοστό θετικού περιεχομένου του κείμενου όπου λαμβάνει τιμές από 0 ως 1
- **neutral:** Το ποσοστό ουδέτερου περιεχομένου του κείμενου όπου λαμβάνει τιμές από 0 ως 1
- **compound:** Το οποίο υπολογίζεται κανονικοποιώντας τα παραπάνω αποτελέσματα καιλαμβάνει τιμές από -1 ως 1.

Τέλος εξάγουμε το weighted polarity χρησιμοποιώντας την σχέση 4.1. Αρχικά παίρνουμε τα reddit post μέσα σε ένα συγκεκριμένο χρονικό διάστημα (by default 15 minutes). Έπειτα αθροίζουμε το γινόμενο score με compound polarity όλον των posts και διαιρούμε το άθροισμα των score. Το score είναι μια παράμετρος που μας παρέχεται από το reddit API και σχηματίζεται συνδυάζοντας τα ups και τα downs που έχει δεχτεί το συγκεκριμένο post.

$$\frac{\sum score * compound\_polarity}{\sum score}$$

Σχέση 4.1 Σχέση υπολογισμού Weighted Polarity

Η διαδικασία εξαγωγής συναισθηματικού περιεχομένου πραγματοποιείτε από το 2ο service της εφαρμογής μας το οποίο είναι υπεύθυνο για την συγκεκριμένη εργασία. Η διαδικασία

ξεκινά λαμβάνοντας ένα kafka message στο topic στο "Collect Crypto Data Topic" (βλ. κεφάλαιο 4.1.4) προκειμένου να ξεκινήσει η διαδικασία εξαγωγής συναισθηματικού περιεχομένου από τα Reddit posts. Αφού ολοκληρωθεί η παραπάνω διαδικασία εισάγει τα δεδομένα στο 2. "Reddit Crypto Data Collection" που περιγράψαμε στην ενότητα 4.2.2 και αναπαριστάται στο σχήμα 4.4. Τέλος παράγεται ένα kafka event το οποίο αναρτάται στο Complete NLP Topic (βλ. κεφάλαιο 4.1.4) προκειμένου να ξεκινήσει η διαδικασία προετοιμασίας τον δεδομένων για πρόβλεψη.

4.6 Επεξεργασία και καθαρισμός.

Αφού ολοκληρώσουμε την συλλογή δεδομένων bitcoin και πραγματοποιήσουμε την διαδικασία συναισθηματικής ανάλυσης στα reddit posts ξεκινάμε την διαδικασία τελικής επεξεργασίας και καθαρισμού των δεδομένων προκειμένου να ετοιμαστούν για να πραγματοποιηθεί πρόβλεψη. Η διαδικασία τελικής επεξεργασίας πραγματοποιείται στο 3ο service της εφαρμογής και ξεκινάει λαμβάνοντας ένα kafka event στο topic "Complete NLP Topic" (βλ. κεφάλαιο 4.1.4) το message του οποίου λαμβάνει το id της εγγραφής με τα δεδομένα του bitcoin από το Raw Crypto Data Collection καθώς και το weighted polarity που υπολογίστηκε από το service 2 για το χρονικό διάστημα που έχει οριστεί (15 λεπτά by default).
Κατά την διαδικασία τελικής επεξεργασίας αρχικά διαλέγουμε τα σημαντικότερα από τα χαρακτηριστικά που συλλέξαμε και μετατρέπουμε το έγγραφο σε ενιαία μορφή (δηλαδή χωρίς να περιέχει εμφωλευμένα έγγραφα). Στην συνέχεια μετονομάζουμε τα properties προκειμένου να έχουν πιο αντιπροσωπευτικά και εύχρηστα ονόματα. Έπειτα ελέγχουμε για missing values, αν εντοπίσουμε κάποιο missing φορτώνουμε τις προηγούμενες 30έγγραφες από Raw Crypto Data Collection και γεμίζουμε τα null values χρησιμοποιώντας τηνinterpolate μέθοδο με feed for word direction.
Τέλος αποθηκεύουμε την εγγραφή στο Processed Crypto Data Collection προκειμένου να είναι έτοιμη για να χρησιμοποιεί για εκπαίδευση και πρόβλεψη των αλγορίθμων.

5. Σχεδιασμός και Εκπαίδευση Μοντέλων Μηχανικής Μάθησης

Στην παρακάτω ενότητα θα περιγραφεί αναλυτικά η διαδικασία εκπαίδευσης μοντέλων μηχανικής μάθησης προκειμένου να πραγματοποιηθεί πρόβλεψη της τιμής του Bitcoin. Αρχικά θα εξεταστούν τα δεδομένα που χρησιμοποιήσαμε κατά την διαδικασία της πρόβλεψης τα οποία έχουν προέλθει από την συλλογής δεδομένων που αποτυπώσαμε στο **κεφάλαιο 4**. Στη συνέχεια θα αναλυθούν οι λόγοι για τους οποίους επιλέχθηκαν αυτά τα μοντέλα καθώς επίσης και τεχνικές λεπτομέρειες σχετικά με την παραμετροποίηση τους. Τέλος θα περιγράψουμε αναλυτικά τη διαδικασία της εκπαίδευσης και εξαγωγής αποτελεσμάτων για κάθε έναν από τους αλγόριθμους που επιλέξαμε.
Στόχος μας είναι η εκπαίδευση μοντέλων μηχανικής μάθησης τα οποία να προβλέπουν την τιμή του Bitcoin με όσο το δυνατόν μεγαλύτερη ακρίβεια. Για να πραγματοποιηθεί αυτό θα σχεδιαστούν και θα δοκιμαστούν ένα νευρωνικό δίκτυο LSTM με ένα χαρακτηριστικό, καθώς και ένα νευρωνικό δίκτυο LSTM με πολλαπλά χαρακτηριστικά, τα οποία θα συγκριθούν με την απόδοση ενός μοντέλου ARIMA όπου θα χρησιμοποιηθεί ως

Benchmark για την αξιολόγηση τους. Η αξιολόγηση των μοντέλων θα πραγματοποιηθεί κάνοντας χρήση της τεχνικής RMSE (Root Mean Square Error).

5.1 Ανάλυση δεδομένων πρόβλεψης.

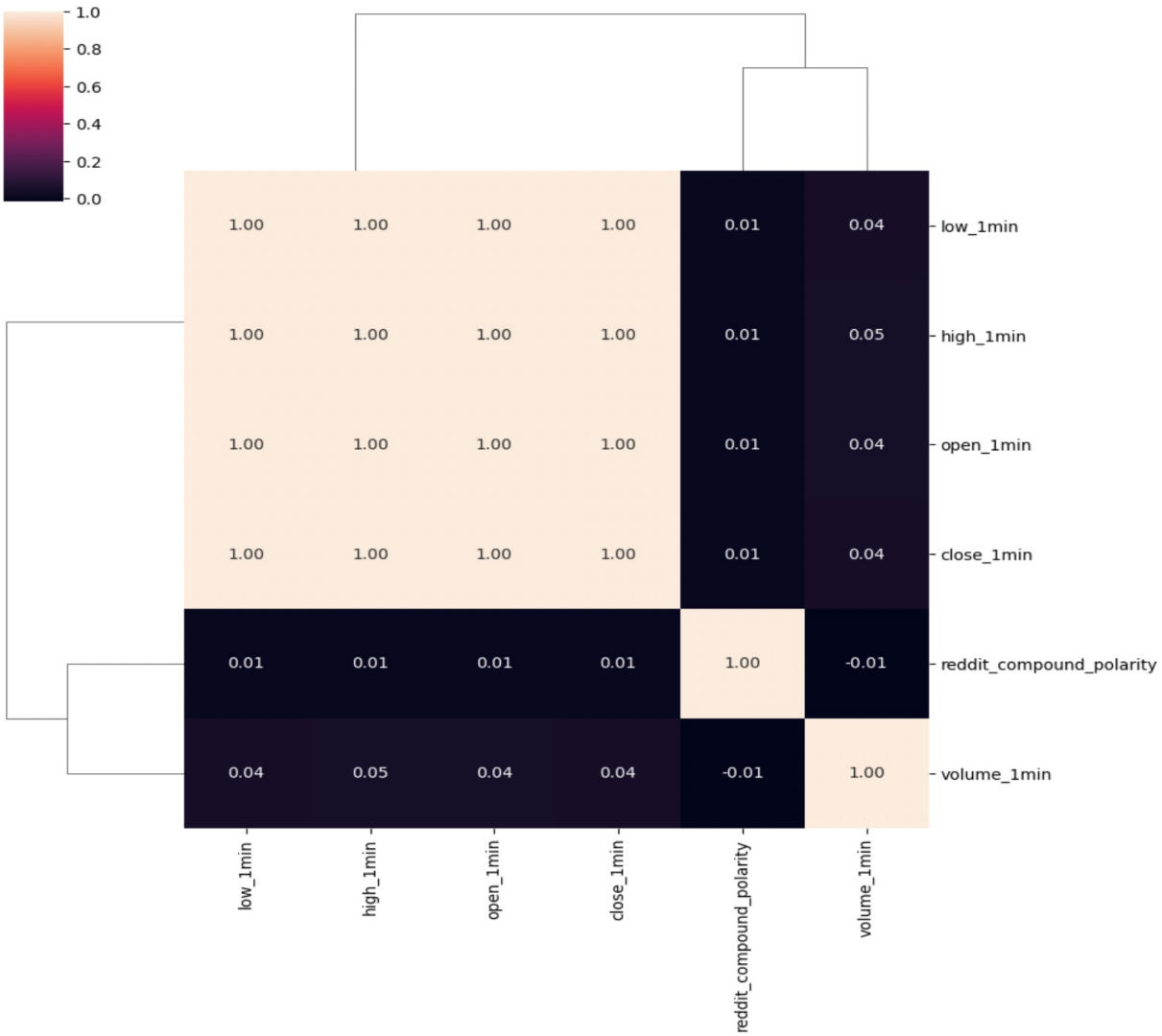
Πριν να προχωρήσουμε στον σχεδιασμό αποτελεσματικών μοντέλων είναι απαραίτητη η κατανόηση των δεδομένων. Για την διαδικασία ανάπτυξης των μοντέλων θα χρησιμοποιήσουμε δεδομένα σχετικά με την τιμή του Bitcoin που συλλέχτηκαν από 2021-10-29 έως και 2022-03-29 και αποτελούνται συνολικά από 206.567 εγγραφές. Κάθε εγγραφή του συνόλου αναπαριστά ένα στιγμιότυπο της τιμής του Bitcoin την δεδομένη χρονική στιγμή και διαθέτει έξι χαρακτηριστικά.

- **open_1min:** Το χαρακτηριστικό αυτό αποτελεί την τιμή που του Bitcoin που "άνοιξε" το λεπτό. Με τον όρο "open" στα χρηματιστηριακά αναφερόμαστε στην τιμή ενός asset στην αρχή ενός χρονικού διαστήματος ενδιαφέροντος. Στην περίπτωση μας το χρονικό διάστημα είναι 1 λεπτό.
- **close_1min:** Το χαρακτηριστικό αυτό αποτελεί την τιμή του Bitcoin που "έκλεισε" το λεπτό. Με τον όρο "close" στα χρηματιστηριακά αναφερόμαστε στην τιμή ενός assetστο τέλος ενός χρονικού διαστήματος ενδιαφέροντος. Στην περίπτωση μας το χρονικό διάστημα είναι 1 λεπτό.
- **high_1min:** Το χαρακτηριστικό αυτό αποτελεί την μέγιστη τιμή του Bitcoin για το λεπτό.
- **low_1min:** Το χαρακτηριστικό αυτό αποτελεί την ελάχιστη τιμή του Bitcoin για το λεπτό.
- **volume_1min:** Το χαρακτηριστικό αυτό αποτελεί τον "όγκο" των συναλλαγών που πραγματοποιήθηκαν για μια συγκεκριμένη χρονική περίοδο. Με τον όρο "όγκο" αναφερόμαστε στην συχνότητα αγοροπωλησιών που πραγματοποιήθηκαν σε ένα συγκεκριμένο διάστημα. Στην περίπτωση μας το χρονικό διάστημα είναι 1 λεπτό.
- **reddit_compound_polarity:** Το χαρακτηριστικό αυτό αποτελεί weighted polarity που εξάγουμε από τα reddit posts με τον τρόπο που περιγράψαμε στην **ενότητα 4.3**.

Datetime	High_1min	Low_1min	Open_1min	Close_1min	Volume_1min	Compound Polarity
2022-03-29 15:03:51	47841.78	47779.93	47787.41	47841.78	0.032706	-0.260219
2022-03-29 15:02:51	47839.85	47817.06	47823.15	47817.06	1.175399	-0.329272
2022-03-29 15:04:51	47842.74	47839.81	47842.74	47839.81	0.016080	-0.260219
2022-03-29 14:54:46	47852.52	47831.77	47831.77	47852.52	0.038551	-0.037502
2022-03-29 14:57:47	47881.91	47856.79	47879.44	47858.93	0.025290	-0.037749

Πίνακας 5.1: Οι πέντε τελευταίες εγγραφές του συνόλου δεδομένων

Τα παραπάνω χαρακτηριστικά θεωρούνται οι πιο σημαντικές πληροφορίες από αυτές που συλλέξαμε και επιλέχτηκαν για τους εξής λόγους. Το χαρακτηριστικό close_1min αποτελεί την τιμή που θα προσπαθήσουμε να προβλέψουμε, δεδομένου ότι σκοπός της πρόβλεψης είναι να προσδιορίσουμε την τιμή που θα κλείσει το χρονικό διάστημα ενδιαφέροντος. Τα OCHL (open, close, high και low) χαρακτηριστικά παρουσιάζουν μεγάλη συσχέτιση μεταξύ τους οπότε μπορούν να βυθίσουν στον προσδιορισμό της τιμής close_1min που αποτελεί και τον στόχο της πρόβλεψης. Το volume των συναλλαγών δηλώνει το πλήθος των αγοροπωλησιών ως προς τον χρόνο άρα μπορεί να μεταφραστεί και σε μια ένδειξη ενδιαφέροντος (είτε αρνητικού είτε θετικού) για το Bitcoin και τέλος το χαρακτηριστικό reddit_compound_polarity περιγράφει επίσης το ενδιαφέρον για το Bitcoin που επικρατεί στα social media. Στην συνέχεια ακολουθεί το διάγραμμα συσχετίσεων για τα χαρακτηριστικά που επιλέχθηκαν.

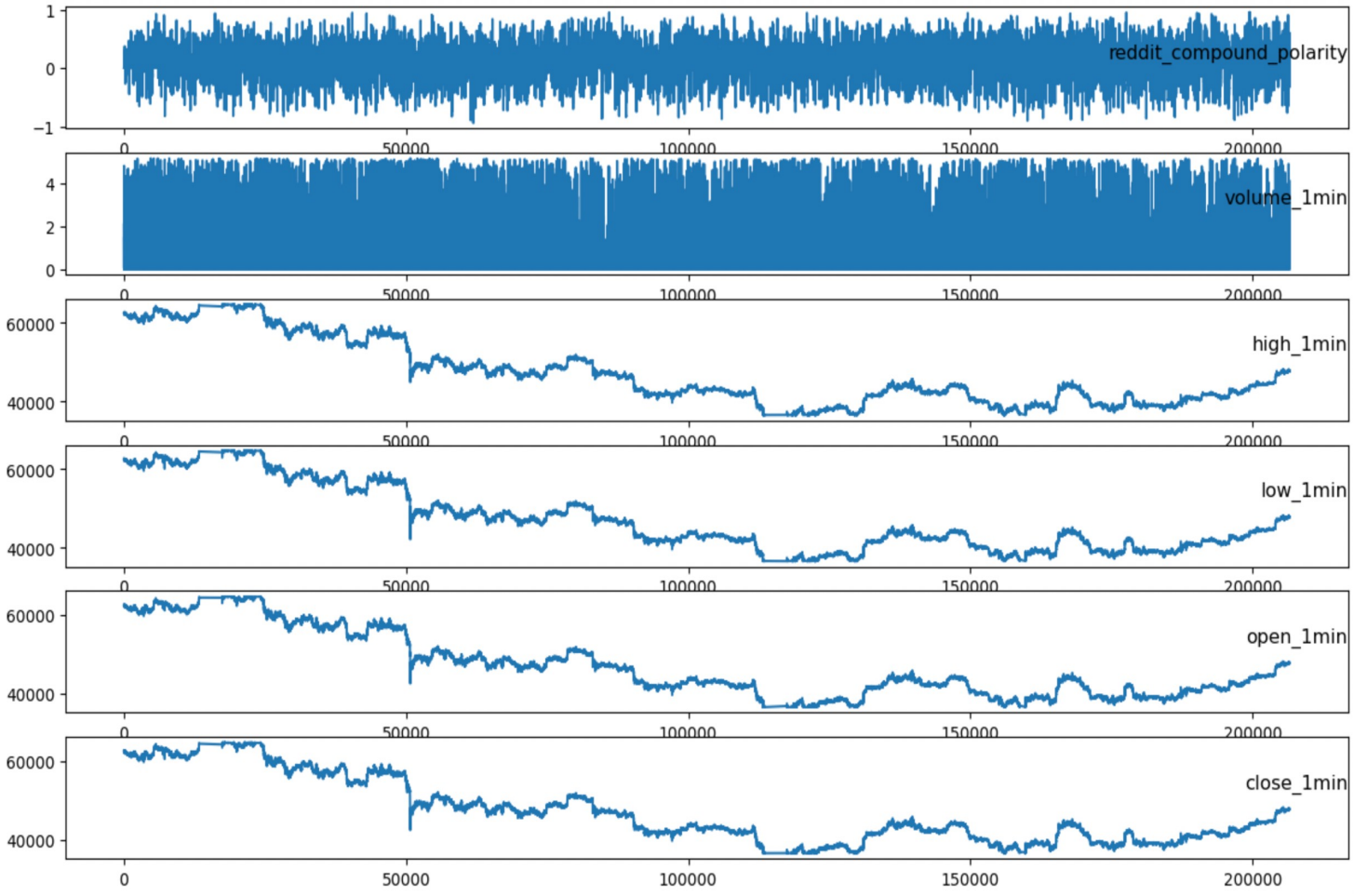


Διάγραμμα 5.1 Γράφημά συσχετίσεων των χαρακτηριστικών

Στον ακόλουθο πίνακα περιγράφονται ορισμένες βασικές πληροφορίες για κάθε μήνα δεδομένων που περιέχεται στο dataset καθώς και οι OCHL (open, close, high και low) τιμές που προκύπτουν για τον κάθε ένα από αυτούς. Επίσης ακολουθεί και γράφημα που αναπαριστά την τιμή κάθε χαρακτηριστικού ως προς τον χρόνο μέσα στο dataset.

Datetime	Monthly Count	Monthly High	Monthly Low	Monthly Open	Monthly Close	Monthly Volume	Compound Polarity
2021-10-31	3273	62741.72	60045.81	62519.06	61331.02	0.379439	0.088477
2021-11-30	42854	64671.60	53308.93	61346.30	57070.07	0.556870	0.067133
2021-12-31	37425	59090.05	42107.88	57027.28	46395.37	0.470280	0.064651
2022-01-31	42315	47960.98	36556.12	46214.37	38540.12	0.446390	0.071157
2022-02-28	39872	45783.31	36556.25	38435.71	43200.00	0.408108	0.078327
2022-03-31	40828	48218.28	37169.52	43221.71	47839.81	0.332203	0.065560

Πίνακας 5.2 Χαρακτηρίστηκα ανά μήνα του συνόλου δεδομένων



Διάγραμμα 5.2: Γραφικές παραστάσεις χαρακτηριστικών συνόλου δεδομένων

5.2 Ανάλυση Διαστημάτων πρόβλεψης και διαχωρισμός δεδομένων

Για την αντικειμενικότερη αξιολόγηση της προβλεψιμότητας των μοντέλων μας θα χρησιμοποιήσουμε τρία διαφορετικά διαστήματα πρόβλεψης. Σκοπός μας είναι να πραγματοποιήσουμε βραχυπρόθεσμη πρόβλεψη της τιμής του Bitcoin για τα χρονικά διαστήματα ενός λεπτού (1 min), δεκαπέντε λεπτών (15 min) και μίας ώρας (60 min). Προκειμένου να πραγματοποιηθεί αυτό θα τροποποιήσουμε το σύνολο δεδομένων που περιγράψαμε στην ενότητα 5.1 κατάλληλα χρησιμοποιώντας τις μεθόδους group by και aggregation [] που μας παρέχονται από την βιβλιοθήκη Pandas της Python. Στον πίνακα που ακολουθεί αναπαριστώνται οι εγγραφές ανά διάστημα πρόβλεψης που προέκυψαν απότην επεξεργασία.

Dateset	Count
1min Dataset	206.567
15min Dataset	13.771
60min Dataset	3343

Πίνακας 5.3 Εγγραφές ανά διάστημα πρόβλεψης

Επίσης προκειμένου να αξιολογηθεί σωστά το κάθε μοντέλο ανά διάστημα πρόβλεψης και να αποφύγουμε το overfiting θα χρησιμοποιήσουμε την μέθοδο train, validation, test. Η μέθοδος train, validation, test αποτελεί μια διαδικασία κατά την οποία διαχωρίζουμε το σύνολο δεδομένων μας σε τρία μέρη. Το πρώτο μέρος του dataset χρησιμοποιείται για την εκπαίδευση των αλγορίθμων, το δεύτερο για την αξιολόγηση τους κατά το στάδιο της παραμετροποίησης και των δοκιμών και το τρίτο χρησιμοποιείται αφού έχουμε ολοκληρώσει την διαδικασία εκπαίδευσης για να αξιολογήσουμε πόσο καλά αντεπεξέρχεται ο αλγόριθμος μας σε νέα δεδομένα. Ένας κλασικός τρόπος εφαρμογής αυτής της τεχνικής είναι ο διαχωρισμός των δεδομένων σε τρία μέρη χρησιμοποιώντας ποσοστά και πιο συγκεκριμένα 80% των δεδομένων για εκπαίδευση (training set) 10% για έλεγχο της εκπαίδευσης (validation set) και 10% για την τελική αξιολόγηση του μοντέλου (test set). Πιο αναλυτικά εφαρμόζοντας την παραπάνω τεχνική για κάθε ένα από τα διαστήματα πρόβλεψης προκύπτουν τα παρακάτω train, validation, test sets.

Dateset	Count
Train set 1min	165241
Valid set 1min	20655
Test set 1 min	20655
Train set 15min	11008
Valid set 15min	1376
Test set 15min	1377
Train set 60min	2746
Valid set 60min	343
Test set 60min	344

Πίνακας 5.4 Εγγραφές ανά train,valid,test set.

Διάγραμμα 5.3.1: Train, Valid, Test split
για χρονικό διάστημα ενός λεπτού

Διάγραμμα 5.3.2: Train, Valid, Test split
για χρονικό διάστημα 15 λεπτών

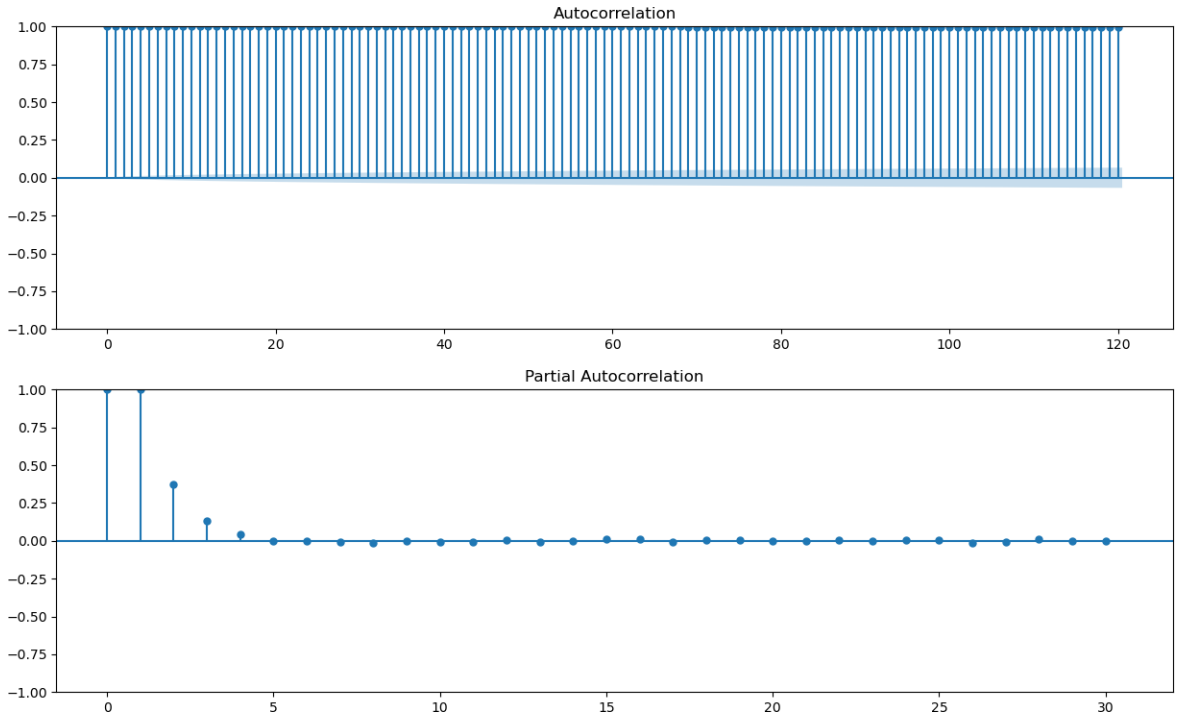
Διάγραμμα 5.3.3: Train, Valid, Test split
για χρονικό διάστημα 60 λεπτών

5.3 Σχεδιασμός μοντέλων ARIMA.

Δεδομένου ότι η τιμή του Bitcoin παρουσιάζει μεταβολή ως προς τον χρόνο η πρόβλεψη της μπορεί να χαρακτηριστεί ένα κατεξοχήν πρόβλημα εφαρμογής χρονοσειρών. Για τον λόγο αυτό το πρώτο μοντέλο που θα εκπαιδεύσουμε θα είναι το ARIMA (AutoRegressive Integrated Moving Average). Το ARIMA όπως αναφέραμε και σε προηγούμενη ενότητα είναι μια στατιστική μέθοδος που χρησιμοποιείται για πρόβλεψη χρονοσειρών και αποτελείται απότρία επιμέρους μοντέλα, AutoRegressive (AR), Integrated (I), Moving Average (MA). Στην συνέχεια θα εξετάσουμε πως πρέπει να σχεδιάσουμε το μοντέλο ARIMA προκειμένου να πετύχουμε όσο το δυνατό καλύτερη πρόβλεψη ανά διάστημα πρόβλεψης. Η τιμή στόχος για την πρόβλεψη θα είναι η τιμή close_1min του bitcoin που αναπαριστά την τιμή που έκλεισε η θέση της τιμής του bitcoin για το εκάστοτε χρονικό διάστημα.

5.3.1 Πρόβλεψη χρονικού διαστήματος 1 λεπτού:

Αρχικά προκειμένου να δημιουργήσουμε ένα αποδοτικό μοντέλο πρέπει να προσδιορίσουμε τις παραμέτρους του μοντέλου (p,b,c). Για να γίνει αυτό πρέπει αρχικά να ελέγξουμε τα γραφήματα ACF (Autocorelation Function)[] και PACF (Partial Autocorrelation Function)[] τα οποία μας βοηθούν να προσδιορίσουμε τους όρους (AR) και (MA) του μοντέλου ARIMA.



Διάγραμμα 5.4: Γραφικές παραστάσεις ACF και PACF
για χρονικό διάστημα 1 λεπτού

Από τα παραπάνω γραφήματα προκύπτει ότι υπάρχει μεγάλη αυτοσυσχέτιση μεταξύ των πρώτων 120 παρατηρήσεων της χρονοσειράς (lags) το οποίο συμπεραίνεται μέσω του γραφήματος ACF. Αντίστοιχα παρατηρούμε στο γράφημα PACF ότι τα residuals των παρατηρήσεων (lags) 1,2,3 είναι στατιστικά σημαντικά επομένως θα μπορούσαν να χρησιμοποιηθούν ως τάξεις του όρου MA. Αφού πραγματοποιήσαμε μια έρευνα για τις πιθανές τάξεις των όρων (AR) και (MA) πρέπει να ελέγξουμε την στασιμότητατης σειράς προκειμένου να προσδιορίσουμε τον όρο I. Από τον έλεγχο που πραγματοποιήσαμε παρατηρήσαμε ότι η αρχική σειρά δεν είναι στάσιμη αφού η τιμή της μετρικής ADF[] τάξεως 0 είναι στατιστικά σημαντική με p-value > 0.05, ωστόσο μετατρέπετε σε στάσιμη αν χρησιμοποιήσουμε διαφορά πρώτης τάξεως όπως φαίνεται και στον παρακάτω πίνακα.

Property	Order	Value
ADF Statistic	0	-1.848609
P-value	0	0.356609
ADF Statistic	1	-52.503307
P-value	1	0.000000

Πίνακας 5.4: Έλενχος μετρικής ADF
για χρονικό διάστημα 1 λεπτού

Λαμβάνοντας υπόψην τα παραπάνω δοκιμάστηκαν διάφοροι συνδυασμοί για του όρους (AR) και (MA) και οδηγηθήκαμε ότι τα καλύτερα αποτελέσματα προκύπτουν χρησιμοποιώντας το μοντέλο ARIMA (4,1,1) τάξεως το οποίο εκπαιδεύτηκε στα δεδομένα "Train set 1min" και παρουσίασε τα ακόλουθα αποτελέσματα στο Validation και στο Test set αξιολογώντας το με την μέθοδο RMSE [].

Dataset	RMSE
Valid set 1min	72.163
Test set 1 min	58.229

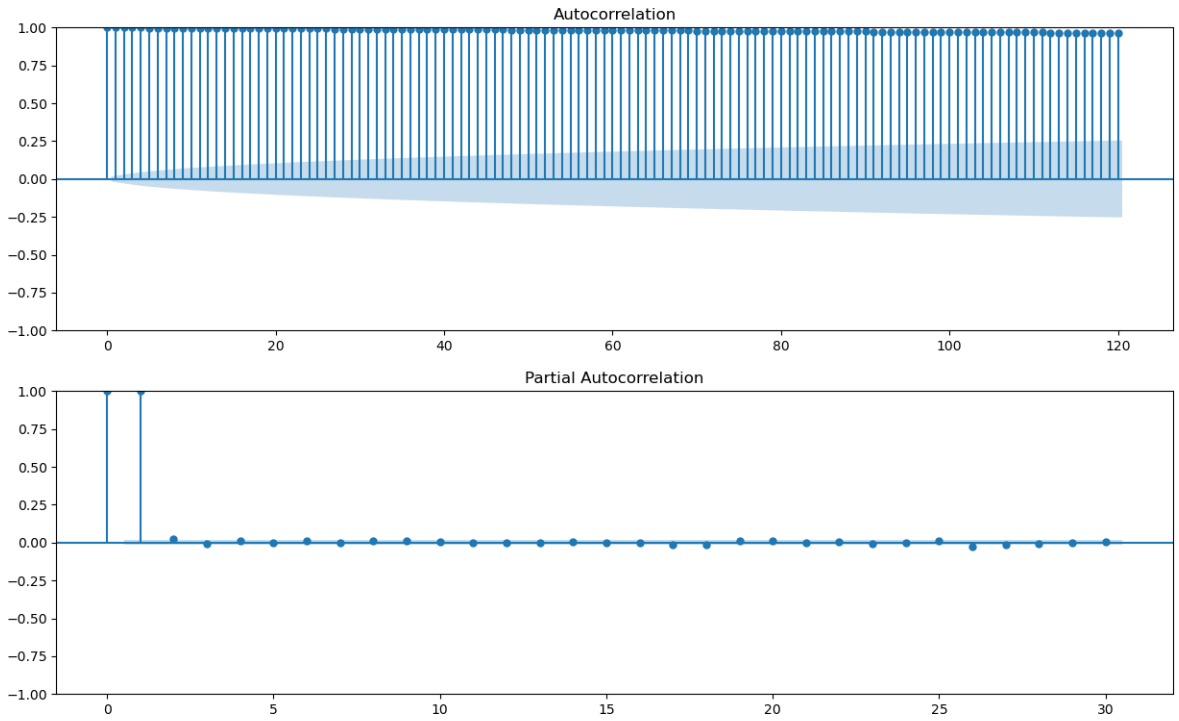
Πίνακας 5.5: Πίνακας αποτελεσμάτων ARIMA(4,1,1)
για χρονικό διάστημα 1 λεπτού

Διάγραμμα 5.5.1: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Valid set 1min για χρονικό διάστημα 1 λεπτού

Διάγραμμα 5.5.2: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Test set 1min για χρονικό διάστημα 1 λεπτού

5.3.2 Πρόβλεψη χρονικού διαστήματος 15 λεπτών:

Επαναλαμβάνουμε την ίδια διαδικασία για να προσδιορίσουμε της παραμέτρους του μοντέλου (p,b,c) για χρονικό διάστημα 15 λεπτών. Για να γίνει αυτό πρέπει αρχικά να ελέγξουμε τα γραφήματα ACF (Autocorelation Function) και PACF (Partial Autocorrelation Function) τα οποία μας βοηθούν να προσδιορίσουμε τους όρους (AR) και (MA) τουμοντέλου ARIMA.



Διάγραμμα 5.6: Γραφικές παραστάσεις ACF και PACF
για χρονικό διάστημα 15 λεπτών

Από τα παραπάνω γραφήματα προκύπτει ότι υπάρχει μεγάλη αυτοσυσχέτιση μεταξύ των πρώτων 120 παρατηρήσεων της χρονοσειράς (lags) το οποίο συμπεραίνεται μέσο του γραφήματος ACF. Αντίστοιχα παρατηρούμε στο γράφημα PACF ότι τα residuals της πρώτης παρατήρησης (lag = 1) είναι στατιστικά σημαντική επομένως θα μπορούσε να χρησιμοποιηθεί ως τάξη του όρου MA. Αφού πραγματοποιήσαμε μια έρευναγια τις πιθανές τάξεις των όρων (AR) και (MA) πρέπει να ελέγξουμε την στασιμότητα της σειράς προκειμένου να προσδιορίσουμε τον όρο I. Από τον έλεγχο που πραγματοποιήσαμε παρατηρήσαμε ότι η αρχική σειρά δεν είναι στάσιμη αφού η τιμή της μετρικής ADF[] τάξεως 0 είναι στατιστικά σημαντική με p-value > 0.05, ωστόσο μετατρέπετε σε στάσιμη αν χρησιμοποιήσουμε διαφορά πρώτης τάξεως όπως φαίνεται και στον παρακάτω πίνακα.

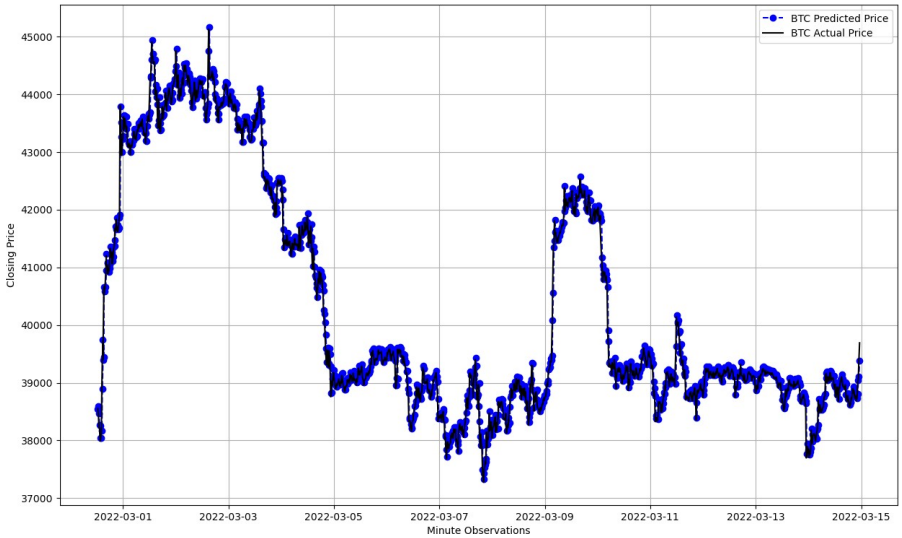
Property	Order	Value
ADF Statistic	0	-1.842163
P-value	0	0.359739
ADF Statistic	1	-22.284604
P-value	1	0.000000

Πίνακας 5.6: Έλενχος μετρικής ADF
για χρονικό διάστημα 15 λεπτών

Λαμβάνοντας υπόψην τα παραπάνω δοκιμάστηκαν διάφοροι συνδυασμοί για του όρους (AR) και (MA) και οδηγηθήκαμε ότι τα καλύτερα αποτελέσματα προκύπτουν χρησιμοποιώντας το μοντέλο ARIMA (4,1,1) τάξεως το οποίο εκπαιδεύτηκε στα δεδομένα "Train set 15min" και παρουσίασε τα ακόλουθα αποτελέσματα στο Validation και στο Test set αξιολογώντας το με την μέθοδο RMSE [].

Dataset	RMSE
Valid set 15 min	169.413
Test set 15 min	157.724

Πίνακας 5.7: Πίνακας αποτελεσμάτων ARIMA(4,1,1)
για χρονικό διάστημα 15 λεπτών

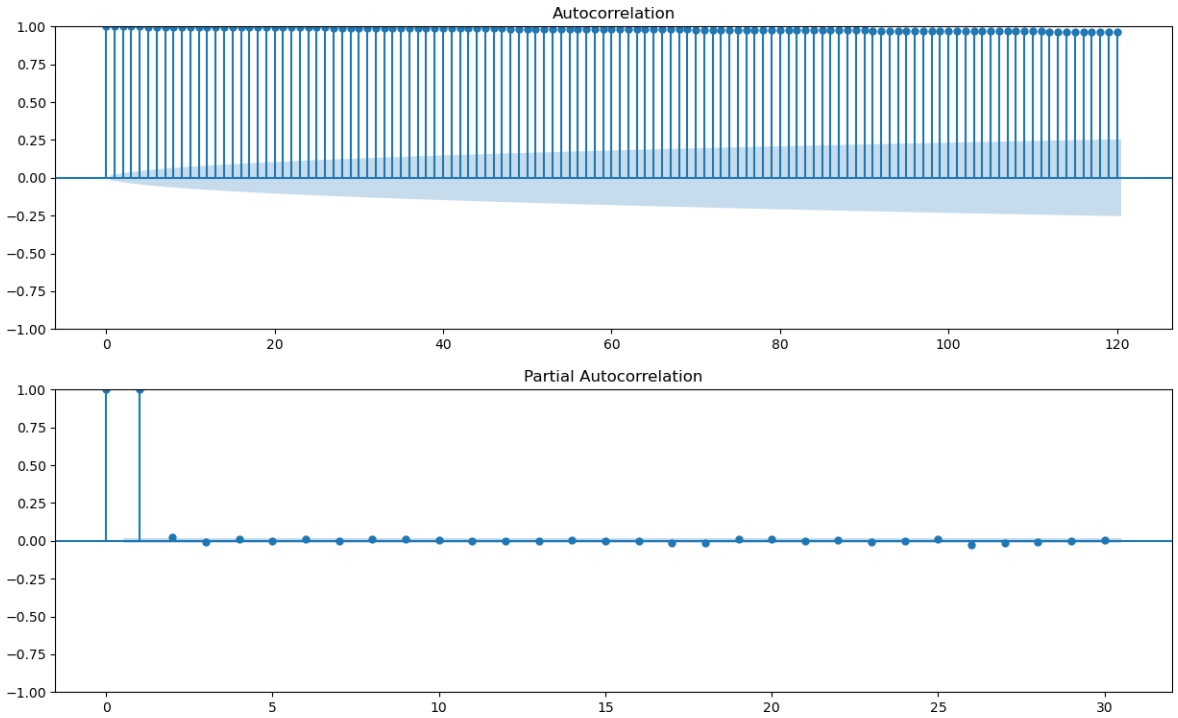


Διάγραμμα 5.7.1: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Valid set 15min για χρονικό διάστημα 15 λεπτών

Διάγραμμα 5.7.2: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Test set 15min για χρονικό διάστημα 15 λεπτών

5.3.3 Πρόβλεψη χρονικού διαστήματος 60 λεπτών:

Τέλος επαναλαμβάνουμε την ίδια διαδικασία για να προσδιορίσουμε τις παραμέτρους του μοντέλου (p,b,c) για χρονικό διάστημα 60 λεπτών. Για να γίνει αυτό πρέπει αρχικά να ελέγξουμε τα γραφήματα ACF (Autocorelation Function) και PACF (Partial Autocorrelation Function)[] τα οποία μας βοηθούν να προσδιορίσουμε τους όρους (AR) και (MA) του μοντέλου ARIMA.
Από τα παραπάνω γραφήματα προκύπτει ότι υπάρχει μεγάλη αυτοσυσχέτιση μεταξύ της των πρώτων 120 παρατηρήσεων της χρονοσειράς (lags) το οποίο συμπεραίνεται μέσω του γραφήματος ACF. Αντίστοιχα παρατηρούμε στο γράφημα PACF ότι τα residuals της πρώτης παρατήρησης (lag = 1) είναι στατιστικά σημαντικά επομένως θα μπορούσε να χρησιμοποιηθεί ως τάξη του όρου MA. Αφού πραγματοποιήσαμε μια έρευναγια τις πιθανές τάξεις των όρων (AR) και (MA) πρέπει να ελέγξουμε την στασιμότητα της σειράς προκειμένου να προσδιορίσουμε τον όρο I. Από τον έλεγχο που πραγματοποιήσαμε παρατηρήσαμε ότι η αρχική σειρά δεν είναι στάσιμη αφού η τιμή της μετρικής ADF[] τάξεως 0 είναι στατιστικά σημαντική με p-value > 0.05, ωστόσο μετατρέπετε σε στάσιμη αν χρησιμοποιήσουμε διαφορά πρώτης τάξεως όπως φαίνεται και στον παρακάτω πίνακα.



Διάγραμμα 5.8: Γραφικές παραστάσεις ACF και PACF
για χρονικό διάστημα 15 λεπτών

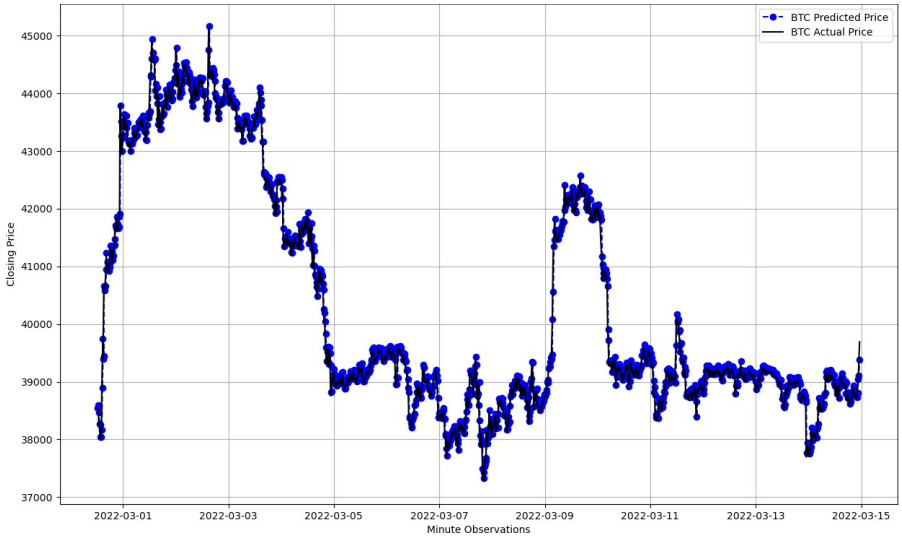
Property	Order	Value
ADF Statistic	0	-1.858118
P-value	0	0.352013
ADF Statistic	1	-61.387993
P-value	1	0.000000

Πίνακας 5.6: Έλενχος μετρικής ADF
για χρονικό διάστημα 60 λεπτών

Λαμβάνοντας υπόψιν τα παραπάνω δοκιμάστηκαν διάφοροι συνδυασμοί για του όρους (AR) και (MA) και οδηγηθήκαμε ότι τα καλύτερα αποτελέσματα προκύπτουν χρησιμοποιώντας το μοντέλο ARIMA (4,1,1) τάξεως το οποίο εκπαιδεύτηκε στα δεδομένα "Train set 60min" και παρουσίασε τα ακόλουθα αποτελέσματα στο Validation και στο Test set αξιολογώντας το με την μέθοδο RMSE.

Dataset	RMSE
Valid set 60 min	339.490
Test set 60 min	276.242

Πίνακας 5.7: Πίνακας αποτελεσμάτων ARIMA(4,1,1)
για χρονικό διάστημα 60 λεπτών



Διάγραμμα 5.9.1: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Valid set 60min για χρονικό διάστημα 60 λεπτών

Διάγραμμα 5.9.2: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Test set 60min για χρονικό διάστημα 60 λεπτών

5.4 Περιγραφή μοντέλων LSTM με χρήση ενός Χαρακτηριστικού.

Το επόμενο μοντέλο που θα χρησιμοποιήσουμε για πρόβλεψη της χρονοσειράς της τιμής του Bitcoin είναι το μοντέλο LSTM (Long-Short term memory). Όπως αναφέραμε και σε προηγούμενη ενότητα Τα LSTM είναι ένας τύπος RNN που έχουν σχεδιαστεί ειδικά για να θυμούνται πληροφορίες για μεγαλύτερες χρονικές περιόδους και χρησιμοποιούνται συνήθως για εργασίες όπως η μετάφραση γλώσσας αναγνώριση ομιλίας και ανάλυση χρονοσειρών. Στην εργασία μας θα χρησιμοποιήσουμε το LSTM προκειμένου να

πραγματοποιήσουμε πρόβλεψη για τα χρονικά διαστήματα 1 λεπτού, 15 λεπτών και 60 λεπτών και στην συνέχεια θα συγκρίνουμε τα αποτελέσματα της πρόβλεψης με αυτά που προέκυψαν από την χρήση του μοντέλου ARIMA στην προηγούμενη ενότητα. Η τιμή στόχος θα είναι η τιμή close_1min που χρησιμοποιήθηκε και κατά την πρόβλεψη με το μοντέλο ARIMA.

5.4.1 Πρόβλεψη χρονικού διαστήματος 1 λεπτού:

Αρχικά θα διαμορφώσουμε το μοντέλο μας κατάλληλα προκειμένου να προβλέψει αποδοτικά την τιμή του Bitcoin για χρονικό διάστημα 1ος λεπτού. Προκειμένου να γίνει αυτό θα διαμορφώσουμε ένα univariate LSTM, δηλαδή ένα μοντέλο LSTM που θα δέχεται μονό μια μεταβλητή υπό την μορφή χρονοσειράς που στην περίπτωση μας θα είναι η μεταβλητή στόχος. Για να είναι αποτελεσματικότερη η πρόβλεψη πρέπει να τροπο- ποιήσουμε τα δεδομένα μας κατάλληλα έτσι ώστε να κυμαίνονται σε τιμές μεταξύ 0 και 1. Αυτό το πραγματοποιούμε κάνοντας χρήση της μεθόδου MinMaxScaler που μας παρέχετεαπό τη βιβλιοθήκη sklearn της Python. Επίσης διαμορφώνουμε τα δεδομένα εισόδου του αλγορίθμου σε πολλαπλά ζευγάρια εισόδου εξόδου όπου τα 15 προηγούμενα time steps χρησιμοποιούνται σαν είσοδο προκειμένου να προβλέψουν την επόμενη τιμή.

Στη συνέχεια χρησιμοποιώντας την βιβλιοθήκη tensorflow θα διαμορφώσουμε το μοντέλο που θα χρησιμοποιήσουμε. Πιο αναλυτικά το μοντέλο μας θα αποτελείτε από ένα LSTM layer με 64 νευρώνες και ένα Dense layer που θα οδηγεί σε ένα νευρώνα όπως φαίνεται και στο πίνακα 5.8.1. Επίσης το μοντέλο μας χρησιμοποιεί και τις παρακάτω παραμέτρουςκατά το στάδιο της εκπαίδευσης οι οποίες απεικονίζονται στον πίνακα 5.8.2

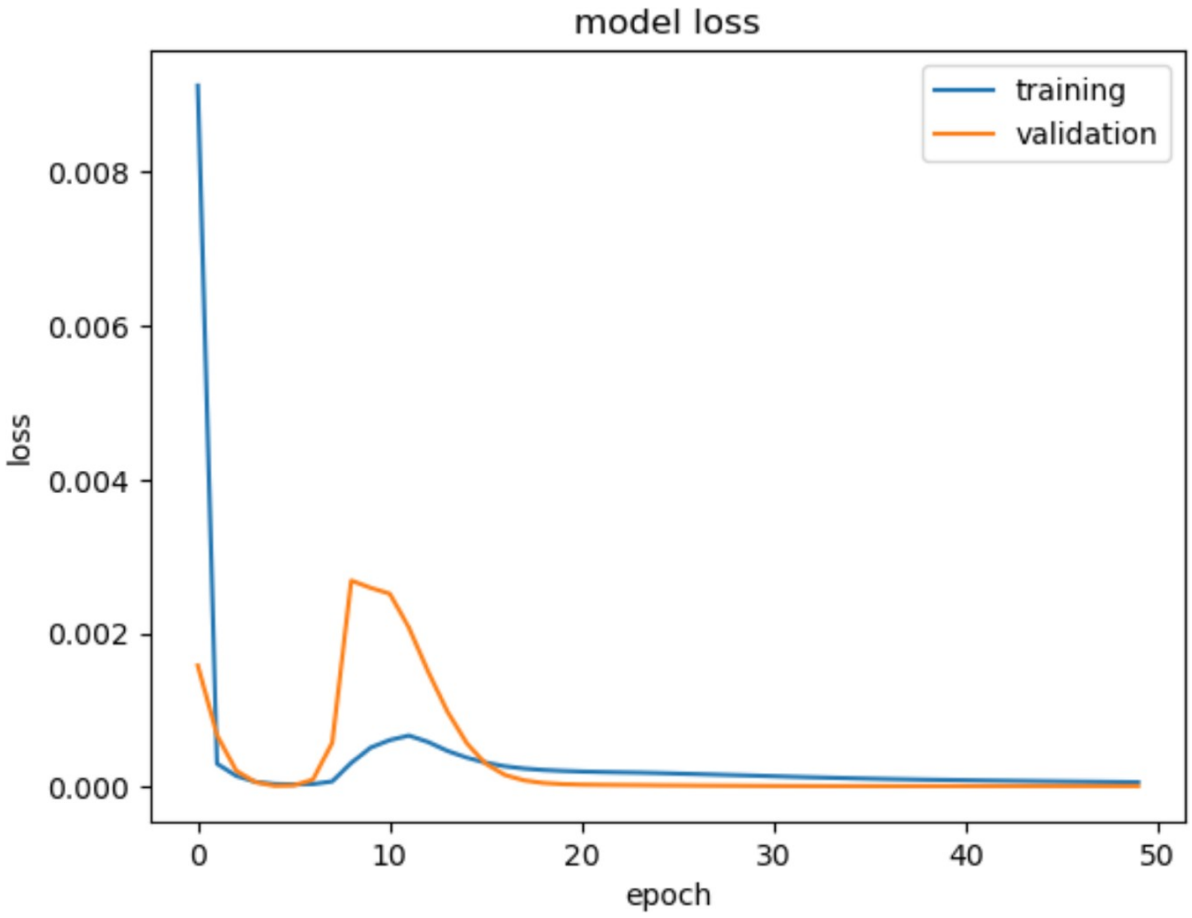
Layer	Neurons
LSTM	64
Dense	1

Πίνακας 5.8.1: Πίνακας αρχιτεκτονικής μοντέλου LSTM
για χρονικό διάστημα 1ος λεπτού

Layer	Neurons
Epochs	50
Batch_size	512
Learning_rate	0.0012

Πίνακας 5.8.2: Πίνακας υπερπαραμέτρων μοντέλου LSTM
για χρονικό διάστημα 1ος λεπτού

Για την εκπαίδευση του αλγορίθμου χρησιμοποιήθηκαν τα δεδομένα από το Train set 1min και το αποτέλεσμα του σφάλματος εκπαίδευσης και validation αναπαριστάται στο παρακάτω γράφημα.

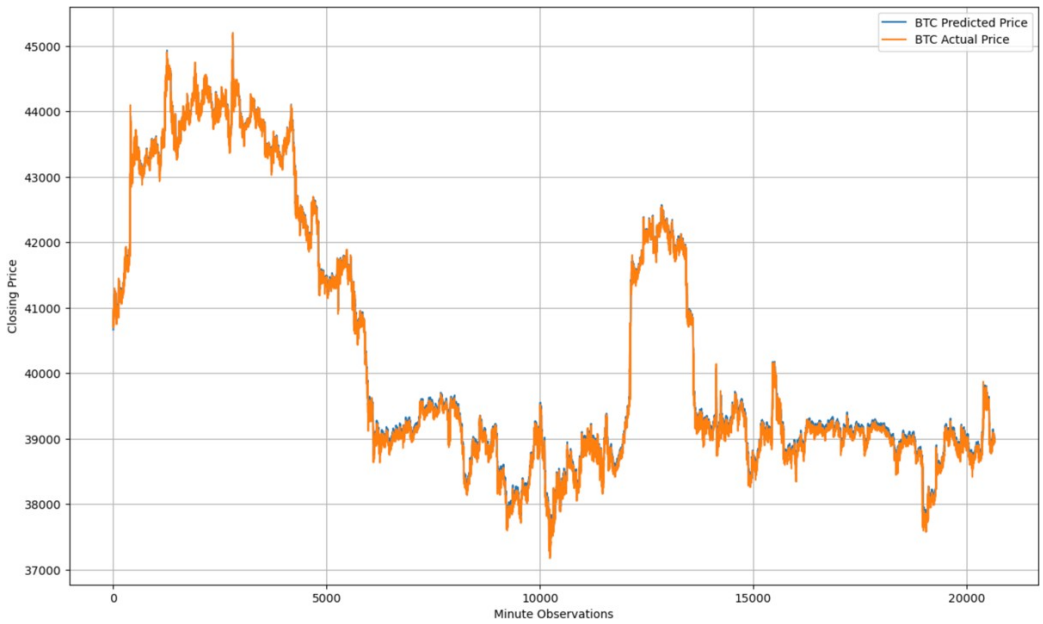


Διάγραμμα 5.10: Γραφική παράσταση train/validation loss της εκπαίδευσης το LSTM
Στο Test set 1min για χρονικό διάστημα 1 λεπτό

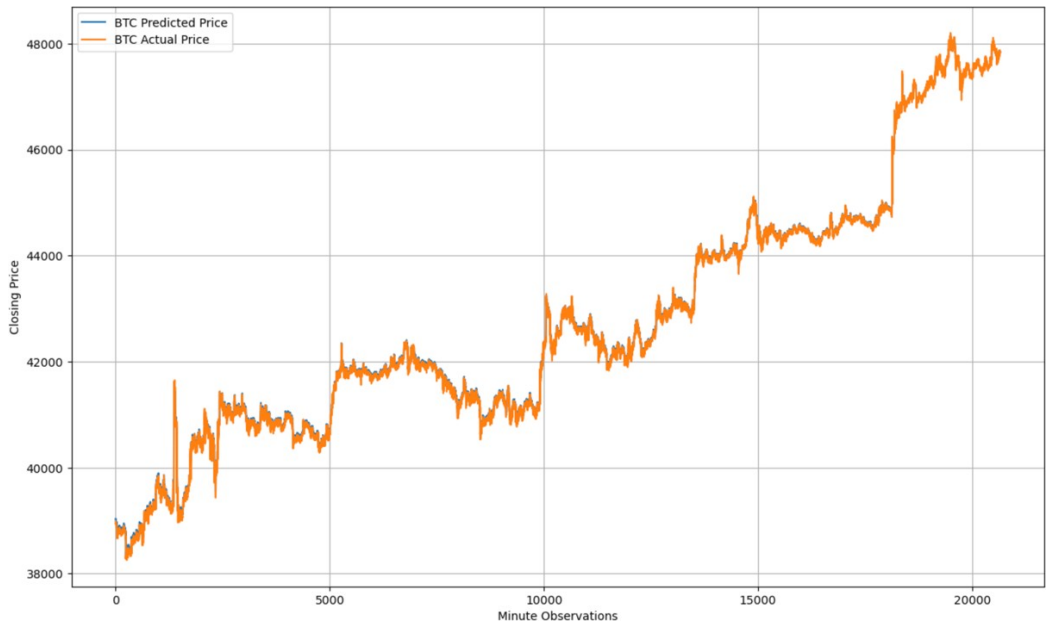
Τέλος πραγματοποιήθηκε πρόβλεψη των δεδομένων στα "Valid set 1 min" και " Test set 1 min" αντίστοιχα προκειμένου να αξιολογήσουμε το μοντέλο και να το συγκρίνουμε με τα αποτελέσματα που εξάγαμε από το ARIMA στην ενότητα 5.3. Η αξιολόγηση του μοντέλου έγινε χρησιμοποιώντας την μετρική RMSE []. Στη συνέχεια ακολουθεί ο πίνακας με τα αποτελέσματα καθώς και τα γραφήματα πρόβλεψης / πραγματικής τιμής.

Dataset	RMSE
Valid set 1 min	84.434
Test set 1 min	61.178

Πίνακας 5.9: Πίνακας αποτελεσμάτων LSTM
για χρονικό διάστημα 1 λεπτού



Διάγραμμα 5.11.1: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Valid set 1min για χρονικό διάστημα 1 λεπτού



Διάγραμμα 5.11.2: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Test set 1min για χρονικό διάστημα 1 λεπτού

5.4.2 Πρόβλεψη χρονικού διαστήματος 15 λεπτών:

Κατόπιν θα διαμορφώσουμε ένα αντίστοιχο μοντέλο προκειμένου να προβλέψει αποδοτικά την τιμή του Bitcoin για χρονικό διάστημα 15 λεπτών. Προκειμένου να γίνει αυτό θα διαμορφώσουμε ένα univariate LSTM με τον ίδιο τρόπο, όπως και στην ενότητα 5.4.1. Θα τροποποιήσουμε τα δεδομένα κατάλληλα έτσι ώστε να κυμαίνονται σε τιμές μεταξύ 0 και 1 και θα διαμορφώσουμε τα δεδομένα εισόδου του αλγορίθμου σε πολλαπλά ζευγάρια εισόδου εξόδου όπου τα 15 προηγούμενα time steps χρησιμοποιούνται σαν είσοδο προκειμένου να προβλέψουν την επόμενη τιμή.

Στη συνέχεια χρησιμοποιώντας την βιβλιοθήκη tensorflow[] θα διαμορφώσουμε το μοντέλο που θα χρησιμοποιήσουμε. Πιο αναλυτικά το μοντέλο μας θα αποτελείτε από ένα LSTM layer με 64 νευρώνες, δύο Dense layer με 32 νευρώνες και ένα Dense layer που θα οδηγεί σε ένα νευρώνα όπως φαίνεται και στο πίνακα 5.10.1. Επίσης το μοντέλο μας χρησιμοποιεί και τις παρακάτω παραμέτρους κατά το στάδιο της εκπαίδευσης οι οποίες απεικονίζονται στον πίνακα 5.10.2

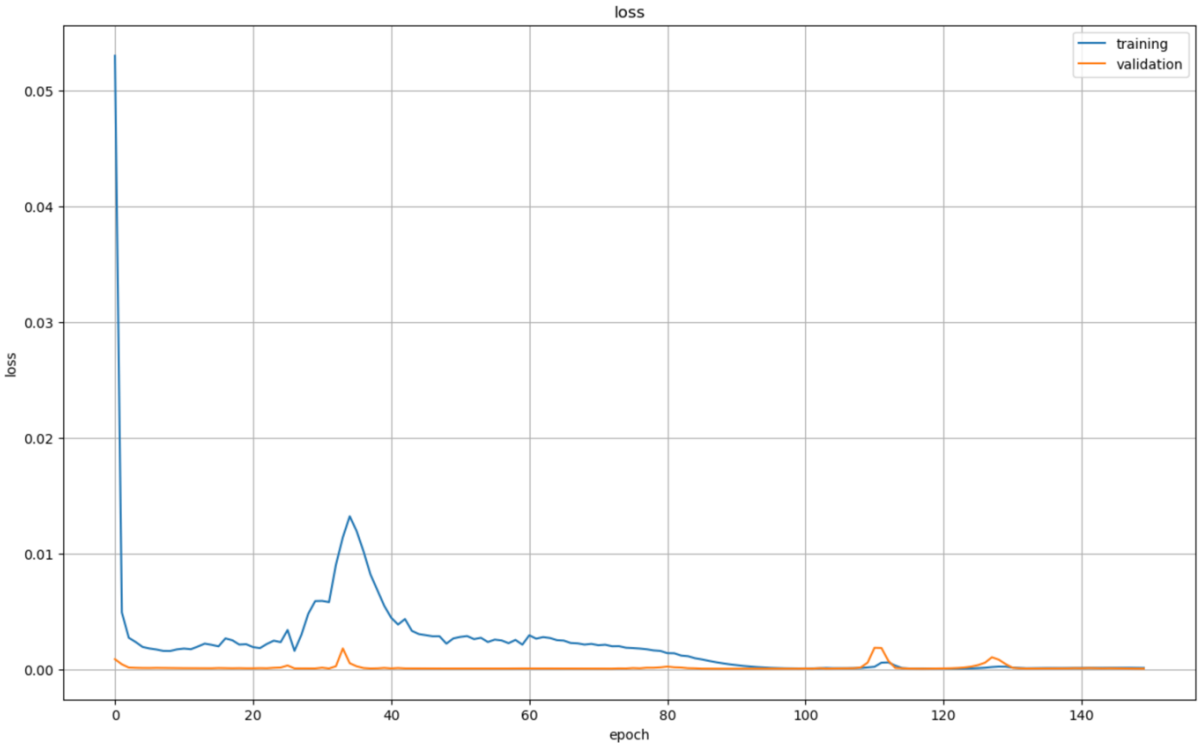
Layer	Neurons
LSTM	64
Dense	32
Dense	32
Dense	1

Πίνακας 5.10.1: Πίνακας αρχιτεκτονικής μοντέλου LSTM
για χρονικό διάστημα 15 λεπτών

Layer	Neurons
Epochs	150
Batch_size	128
Learning_rate	0.001

Πίνακας 5.10.2: Πίνακας υπερπαραμέτρων μοντέλου LSTM
για χρονικό διάστημα 15 λεπτών

Για την εκπαίδευση του αλγορίθμου χρησιμοποιήθηκαν τα δεδομένα από το Train set 15 min και το αποτέλεσμα του σφάλματος εκπαίδευσης και validation αναπαριστάται στο παρακάτω γράφημα.

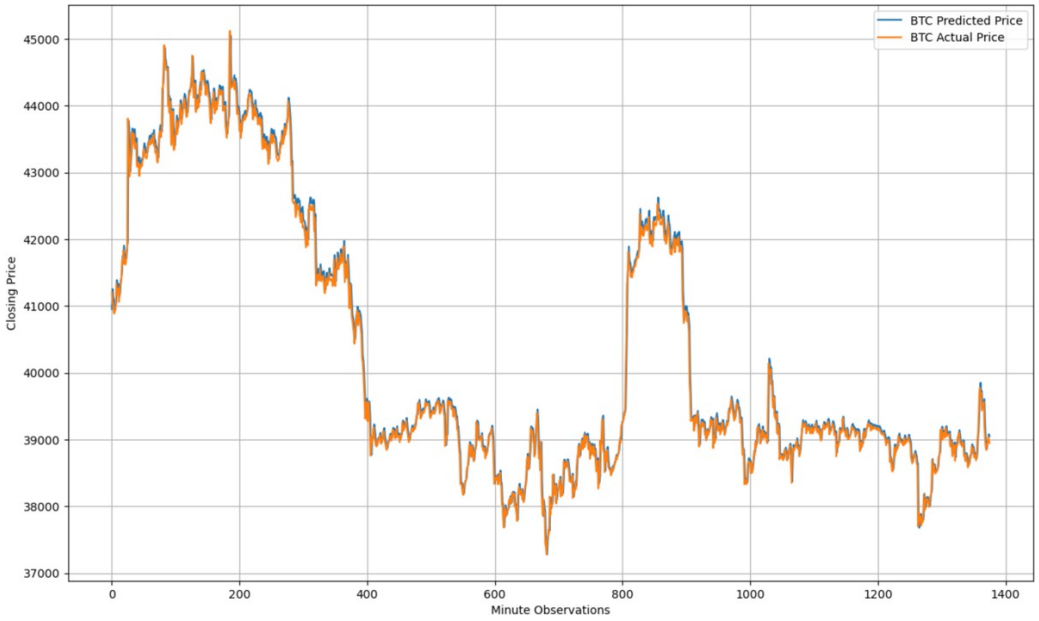


Διάγραμμα 5.12: Γραφική παραστάση train/validation loss της εκπαίδευσης το LSTM
Στο Test set 15min για χρονικό διάστημα 15 λεπτών

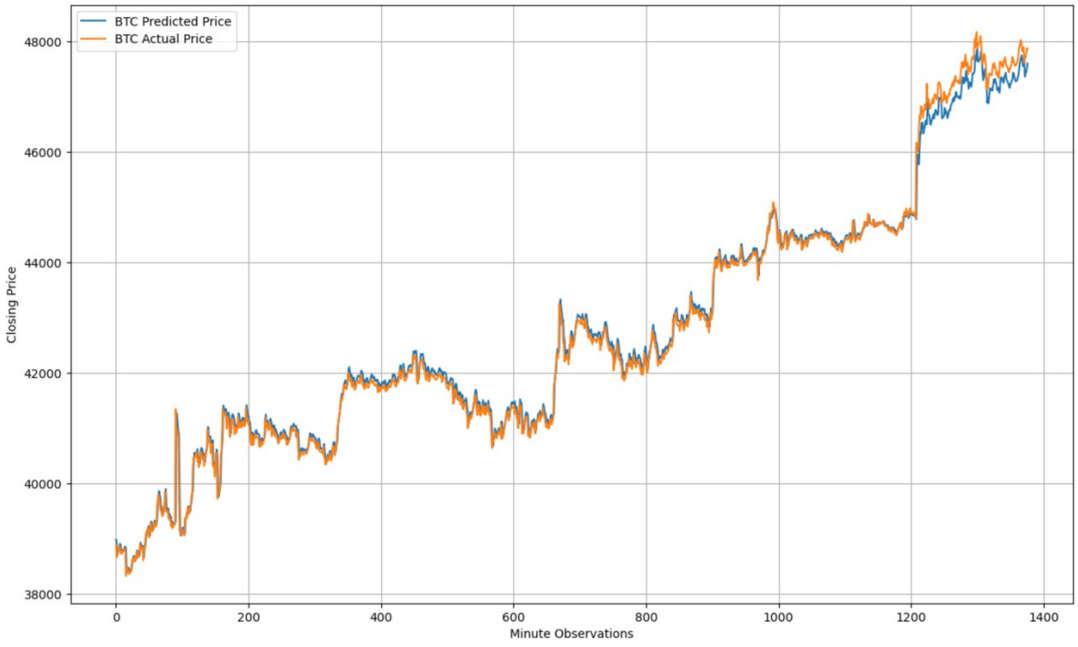
Τέλος πραγματοποιήθηκε πρόβλεψη των δεδομένων στα "Valid set 15 min" και " Test set 15 min" αντίστοιχα προκειμένου να αξιολογήσουμε το μοντέλο και να το συγκρίνουμε με τα αποτελέσματα που εξάγαμε από το ARIMA στην ενότητα 5.3. Η αξιολόγηση του μοντέλου έγινε χρησιμοποιώντας την μετρική RMSE. Στη συνέχεια ακολουθεί ο πίνακας με τα αποτελέσματα καθώς και τα γραφήματα πρόβλεψης / πραγματικής τιμής.

Dataset	RMSE
Valid set 15 min	168.261
Test set 15 min	177.040

Πίνακας 5.11: Πίνακας αποτελεσμάτων LSTM
για χρονικό διάστημα 15 λεπτών



Διάγραμμα 5.13.1: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Valid set 15min για χρονικό διάστημα 15 λεπτών



Διάγραμμα 5.13.2: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Test set 15min για χρονικό διάστημα 15 λεπτών

5.4.3 Πρόβλεψη χρονικού διαστήματος 60 λεπτών:

Τέλος θα δημιουργήσουμε ένα αντίστοιχο μοντέλο προκειμένου να προβλέψει αποδοτικά την τιμή του Bitcoin για χρονικό διάστημα 60 λεπτών. Προκειμένου να γίνει αυτό θα διαμορφώσουμε ένα univariate LSTM με τον ίδιο τρόπο, όπως και στις προηγούμενες ενότητες. Θα τροποποιήσουμε τα δεδομένα κατάλληλα έτσι ώστε να κυμαίνονται σε τιμές μεταξύ 0 και 1 και θα διαμορφώσουμε τα δεδομένα εισόδου του αλγορίθμου σε πολλαπλά ζευγάρια εισόδου εξόδου όπου τα 15 προηγούμενα time steps χρησιμοποιούνται σαν είσοδο προκειμένου να προβλέψουν την επόμενη τιμή.

Στη συνέχεια χρησιμοποιώντας την βιβλιοθήκη tensorflow θα διαμορφώσουμε το μοντέλο που θα χρησιμοποιήσουμε. Πιο αναλυτικά το μοντέλο μας θα αποτελείτε από ένα LSTM layer με 64 νευρώνες, ένα Dense layer με 32 νευρώνες και ένα Dense layer που θα οδηγεί σε ένα νευρώνα όπως φαίνεται και στο πίνακα 5.12.1. Επίσης το μοντέλο μας χρησιμοποιεί και τις παρακάτω παραμέτρους κατά το στάδιο της εκπαίδευσης οι οποίες απεικονίζονται στον πίνακα 5.12.2

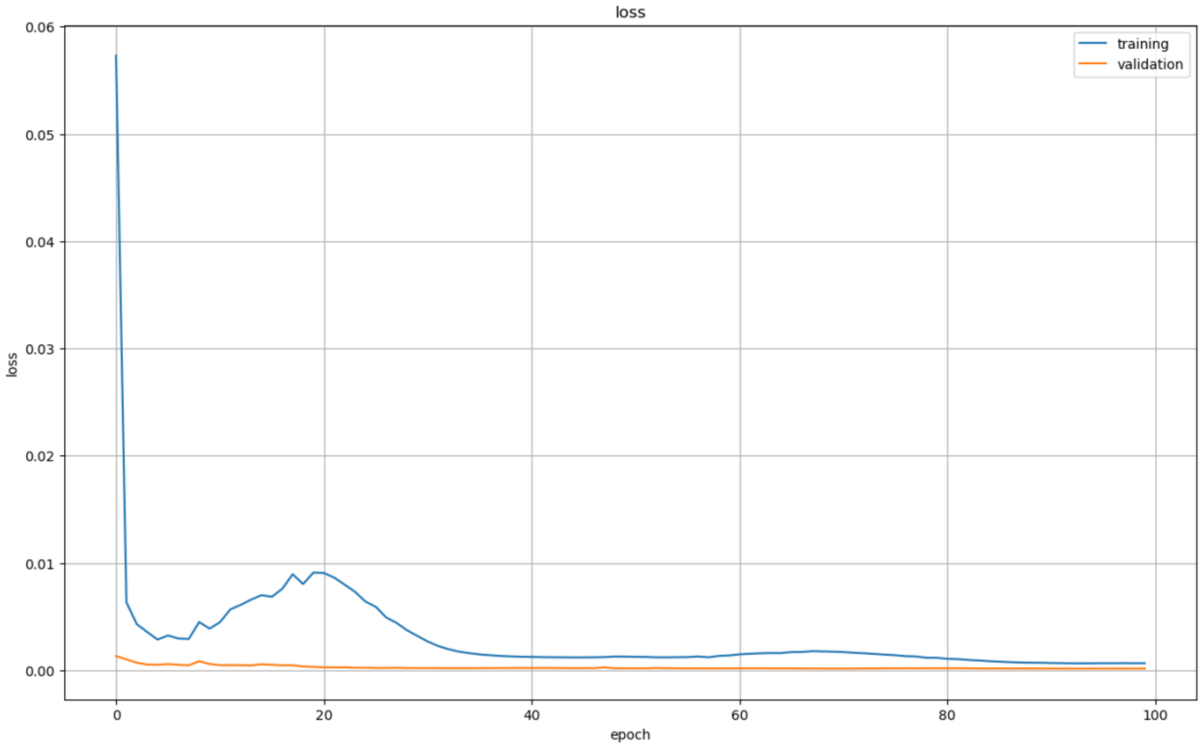
Layer	Neurons
LSTM	64
Dense	32
Dense	1

Πίνακας 5.12.1: Πίνακας αρχιτεκτονικής μοντέλου LSTM
για χρονικό διάστημα 60 λεπτών

Layer	Neurons
Epochs	100
Batch_size	32
Learning_rate	0.001

Πίνακας 5.12.2: Πίνακας υπερπαραμέτρων μοντέλου LSTM
για χρονικό διάστημα 60 λεπτών

Για την εκπαίδευση του αλγορίθμου χρησιμοποιήθηκαν τα δεδομένα από το Train set 60 min και το αποτέλεσμα του σφάλματος εκπαίδευσης και validation αναπαριστάται στο παρακάτω γράφημα.

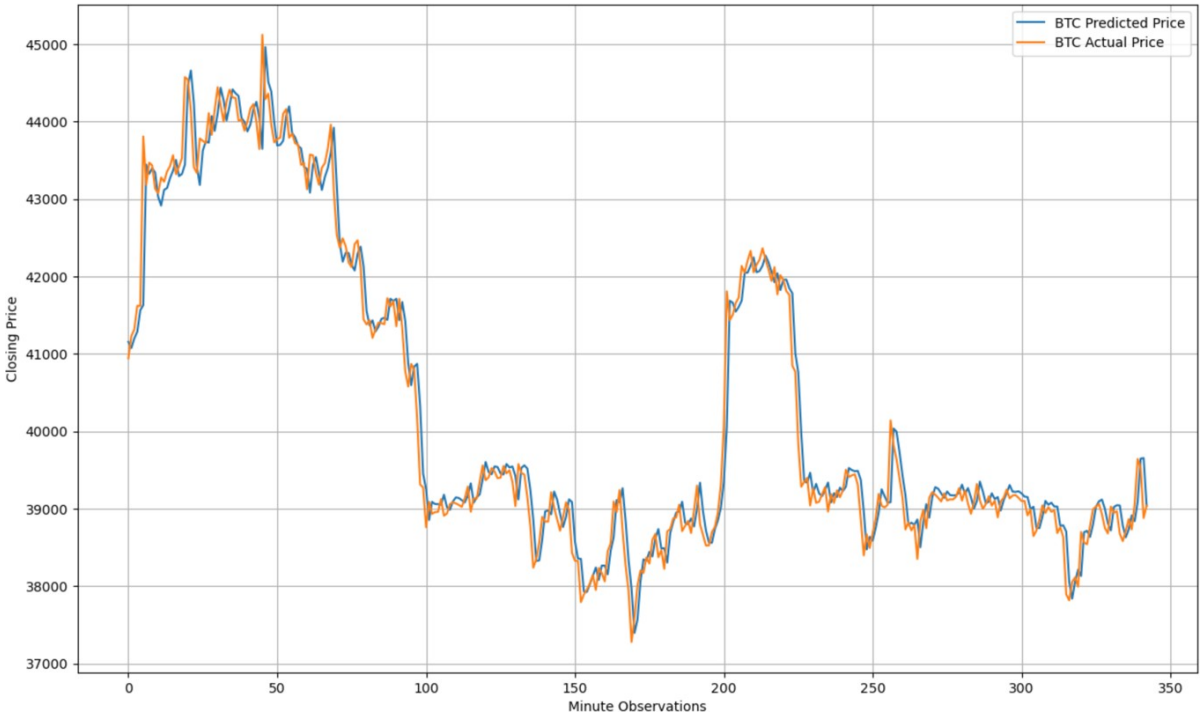


Διάγραμμα 5.14: Γραφική παραστάση train/validation loss της εκπαίδευσης το LSTM
Στο Test set 60min για χρονικό διάστημα 60 λεπτών

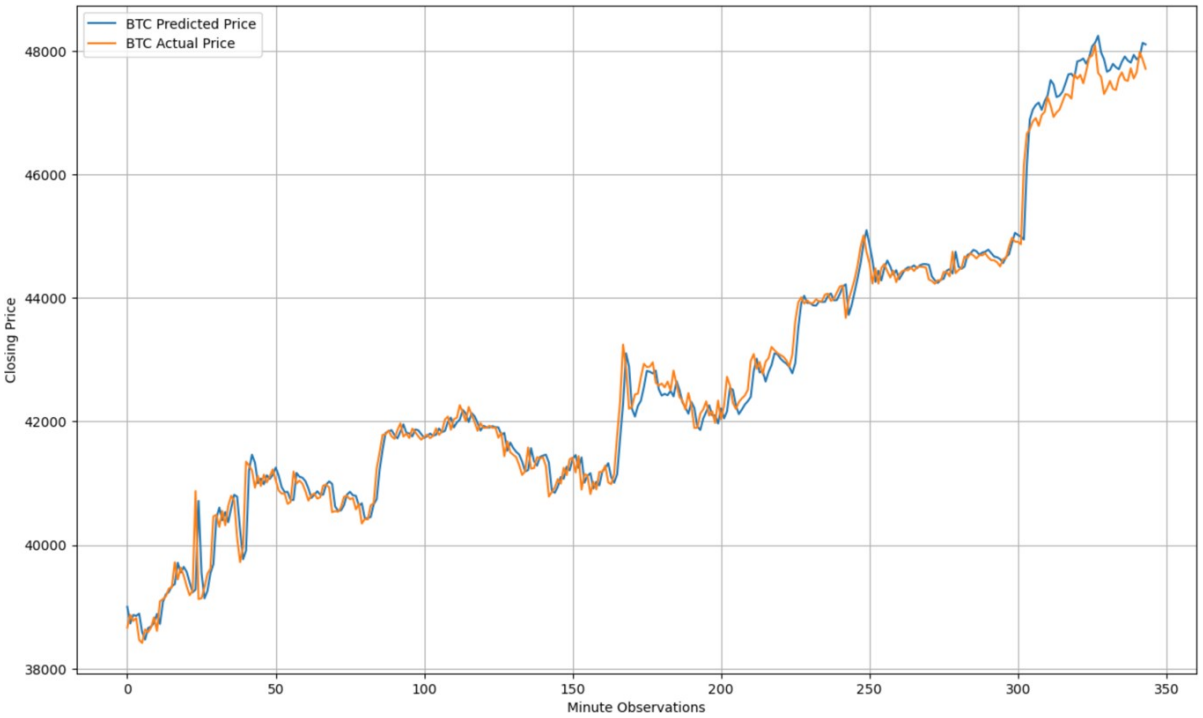
Τέλος πραγματοποιήθηκε πρόβλεψη των δεδομένων στα "Valid set 60 min" και " Test set 60 min" αντίστοιχα προκειμένου να αξιολογήσουμε το μοντέλο και να το συγκρίνουμε με τα αποτελέσματα που εξάγαμε από το ARIMA στην ενότητα 5.3. Η αξιολόγηση του μοντέλου έγινε χρησιμοποιώντας την μετρική RMSE. Στη συνέχεια ακολουθεί ο πίνακας με τα αποτελέσματα καθώς και τα γραφήματα πρόβλεψης / πραγματικής τιμής.

Dataset	RMSE
Valid set 60 min	332.667
Test set 60 min	277.256

Πίνακας 5.13: Πίνακας αποτελεσμάτων LSTM
για χρονικό διάστημα 60 λεπών



Διάγραμμα 5.15.1: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Valid set 60min για χρονικό διάστημα 60 λεπτών



Διάγραμμα 5.15.2: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Test set 60min για χρονικό διάστημα 60 λεπτών

5.5 Περιγραφή μοντέλων LSTM με χρήση πολλών Χαρακτηριστικών.

Στην προηγούμενη ενότητα εκπαιδεύσαμε μοντέλα LSTM που χρησιμοποιούσαν μόνο ένα χαρακτηριστικό "close_1min", επίσης μετασχηματίσαμε τα δεδομένα προκειμένου να πραγματοποιήσουμε πρόβλεψη της τιμής στόχου βάση των 15 προηγούμενων παρατηρήσεων. Στόχος μας σε αυτήν την ενότητα είναι να διαμορφώσουμε νευρωνικά μοντέλα LSTM τα οποία δέχονται πολλαπλά χαρακτηριστικά από προηγούμενες παρατηρήσεις και χρησιμοποιούνται προκειμένου να προβλέψουν την μεταβλητή στόχος. Η Αρχιτεκτονική αυτή ονομάζεται Multivariate LSTM. Για την ανάπτυξη των μοντέλων αυτών θα χρησιμοποιηθούν όλα τα χαρακτηριστικά που περιγράψαμε στην ενότητα 5.1 και θα πραγματοποιηθούν προβλέψεις για τα χρονικά διαστήματα 1 λεπτού, 15 λεπτών και 60 λεπτών. Τέλος θα αξιολογήσουμε τα αποτελέσματα των μοντέλων ανά διάστημαπρόβλεψης και θα τα συγκρίνουμε με τα αποτελέσματα που συλλέξαμε από προηγούμενεςενότητες.

5.5.1 Πρόβλεψη χρονικού διαστήματος 1 λεπτό:

Αρχικά προετοιμάζουμε τα δεδομένα μας προκειμένου να χρησιμοποιηθούν απο το Multivariate LSTM μοντέλο που θα αναπτύξουμε. Η διαδικασία περιλαμβάνει τον μετασχηματισμό του συνόλου δεδομένων σε supervised learning πρόβλημα καθώς και την κανονικοποίηση των δεδομένων. Ο μετασχηματισμός των δεδομένων θα γίνει με τέτοιο τρόπο έτσι ώστε να καταφέρουμε να προβλέψουμε την "close_1min" τιμή του Bitcoin για την χρονική στιγμή (t) συνδυάζοντας τα χαρακτηριστικά των προηγούμενων 15 παρατηρήσεων (n = 15). Πιο αναλυτικά στόχος του μοντέλου είναι να προβλέψει την τιμή κλεισίματος του διαστήματος πρόβλεψης δεδομένων των χαρακτηριστικών των προηγούμενων 15 λεπτών.

Αφού ολοκληρώσουμε την προετοιμασία, ορίζουμε το μοντέλο και προχωράμε στο στάδιο της εκπαίδευσης χρησιμοποιώντας το σύνολο δεδομένων. Πιο αναλυτικά το μοντέλο μας θα αποτελείτε από ένα LSTM layer με 128 νευρώνες και ένα Dense layer που θα οδηγεί σε ένα νευρώνα όπως φαίνεται και στο πίνακα 5.14.1. Επίσης το μοντέλο μας χρησιμοποιεί καιτις παρακάτω παραμέτρους κατά το στάδιο της εκπαίδευσης οι οποίες απεικονίζονται στον πίνακα 5.14.2

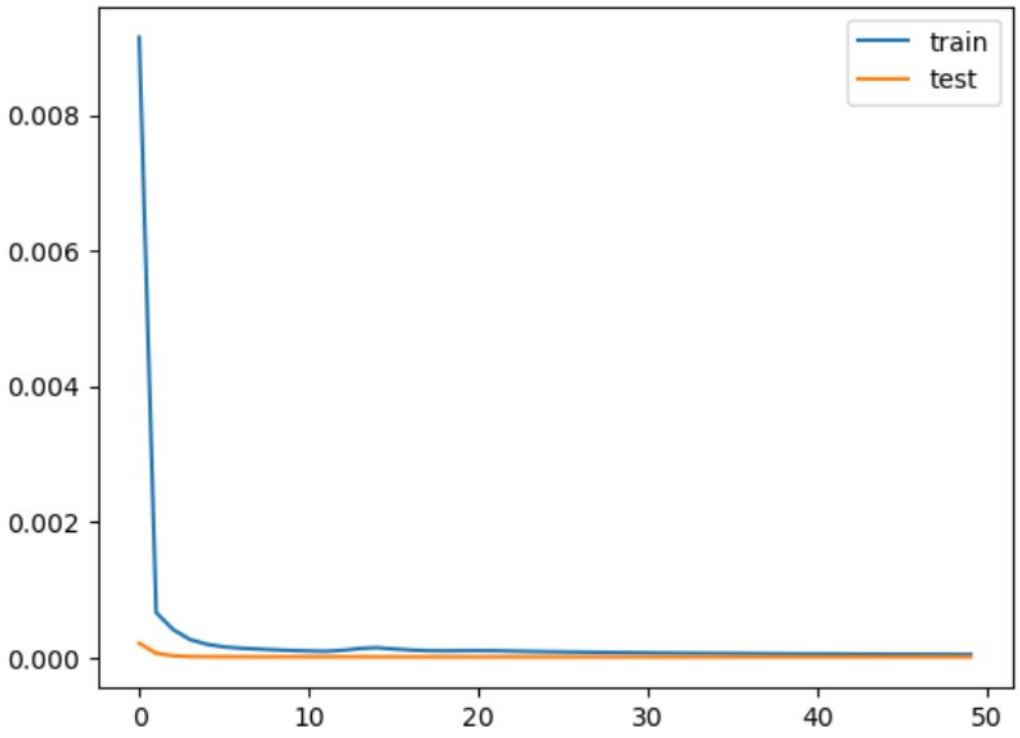
Layer	Neurons
LSTM	128
Dense	1

Πίνακας 5.14.1: Πίνακας αρχιτεκτονικής μοντέλου Multivariate LSTM
για χρονικό διάστημα 1 λεπτού

Layer	Neurons
Epochs	50
Batch_size	512
Learning_rate	0.001

Πίνακας 5.14.2: Πίνακας υπερπαραμέτρων μοντέλου Multivariate LSTM
για χρονικό διάστημα 1 λεπτού

Για την εκπαίδευση του αλγορίθμου χρησιμοποιήθηκαν τα δεδομένα από το Train set 1 min και το αποτέλεσμα του σφάλματος εκπαίδευσης και validation αναπαριστάται στο παρακάτω γράφημα.

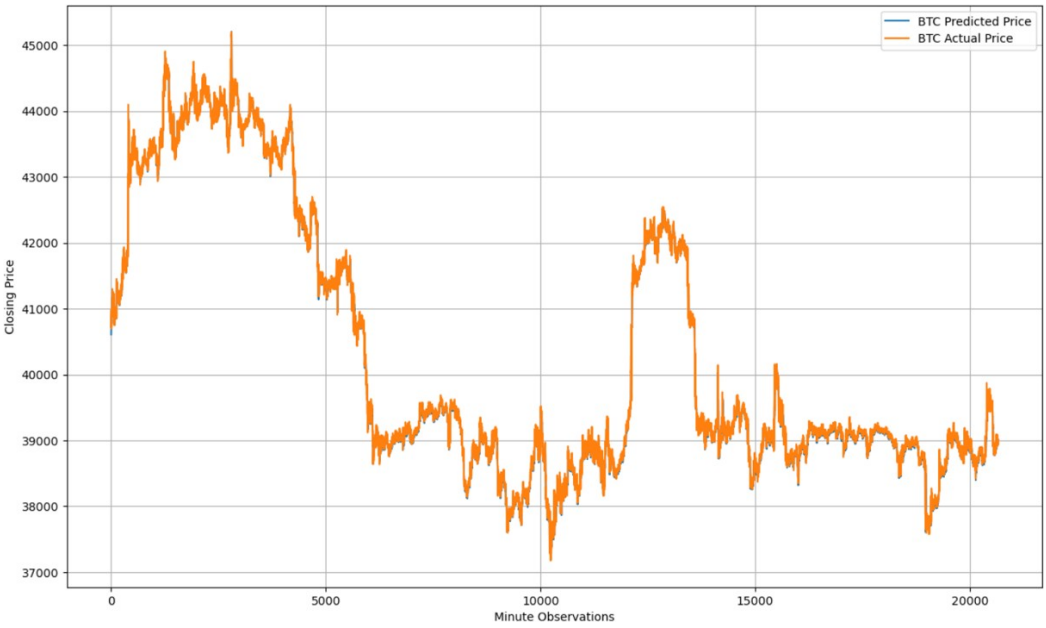


Διάγραμμα 5.16: Γραφική παραστάση train/validation loss της εκπαίδευσης του Multivariate LSTM
για χρονικό διάστημα 1 λεπτού

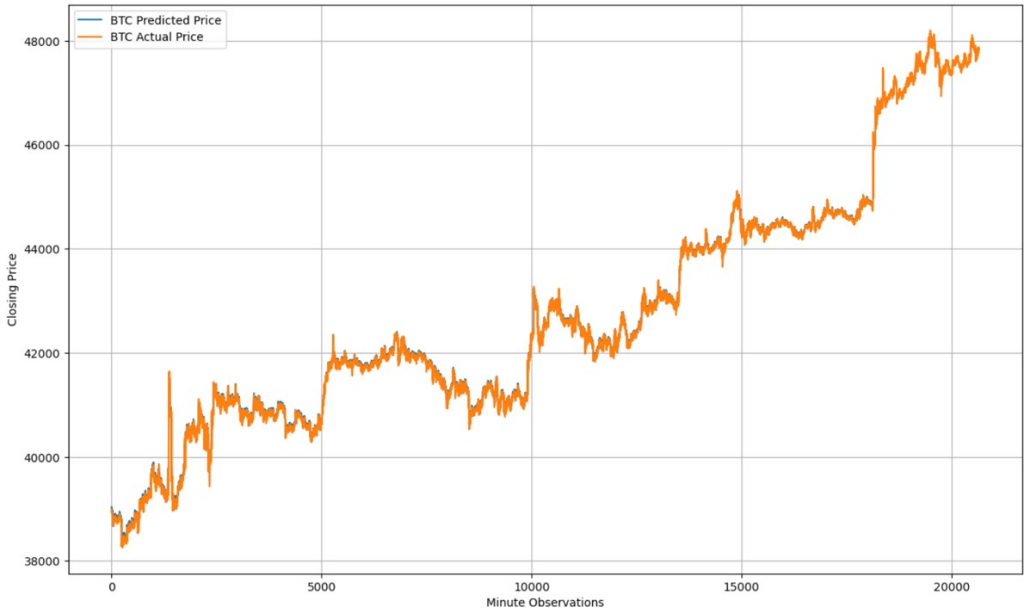
Τέλος πραγματοποιήθηκε πρόβλεψη των δεδομένων στα "Valid set 1 min" και " Test set 1 min" αντίστοιχα προκειμένου να αξιολογήσουμε το μοντέλο και να το συγκρίνουμε με τα αποτελέσματα που εξάγαμε από το ARIMA στην ενότητα 5.3 και το Univariate LSTM της ενότητας 5.4. Η αξιολόγηση του μοντέλου έγινε χρησιμοποιώντας την μετρική RMSE []. Στη συνέχεια ακολουθεί ο πίνακας με τα αποτελέσματα καθώς και τα γραφήματα πρόβλεψης / πραγματικής τιμής.

Dataset	RMSE
Valid set 1 min	70.832
Test set 1 min	53.803

Πίνακας 5.15: Πίνακας αποτελεσμάτων Multivariate LSTM
για χρονικό διάστημα 1 λεπτού



Διάγραμμα 5.17.1: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Valid set 1min για χρονικό διάστημα 1 λεπτού



Διάγραμμα 5.17.2: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Test set 1min για χρονικό διάστημα 1 λεπτού

5.5.2 Πρόβλεψη χρονικού διαστήματος 15 λεπτών:

Έπειτα επαναλαμβάνουμε την αντίστοιχη διαδικασία προετοιμασίας των δεδομένων του χρονικού διαστήματος 15 λεπτών προκειμένου να χρησιμοποιηθούν από το Multivariate LSTM μοντέλο που θα αναπτύξουμε. Η διαδικασία περιλαμβάνει τον μετασχηματισμό του συνόλου δεδομένων σε supervised learning πρόβλημα καθώς και την κανονικοποίηση των δεδομένων. Ο μετασχηματισμός των δεδομένων θα γίνει ξανά με τέτοιο τρόπο έτσι ώστε να προβλέψουμε την τιμή κλεισίματος του Bitcoin για την χρονική στιγμή (t) συνδυάζοντας τα χαρακτηριστικά των προηγούμενων 15 παρατηρήσεων (n = 15).

Αφού ολοκληρώσουμε την προετοιμασία, ορίζουμε το μοντέλο και προχωράμε στο στάδιο της εκπαίδευσης χρησιμοποιώντας το σύνολο δεδομένων. Πιο αναλυτικά το μοντέλο μας θα αποτελείτε από ένα LSTM layer με 128 νευρώνες και ένα Dense layer που θα οδηγεί σε ένα νευρώνα όπως φαίνεται και στο πίνακα 5.16.1. Επίσης το μοντέλο μας χρησιμοποιεί καιτις παρακάτω παραμέτρους κατά το στάδιο της εκπαίδευσης οι οποίες απεικονίζονται στον πίνακα 5.16.2

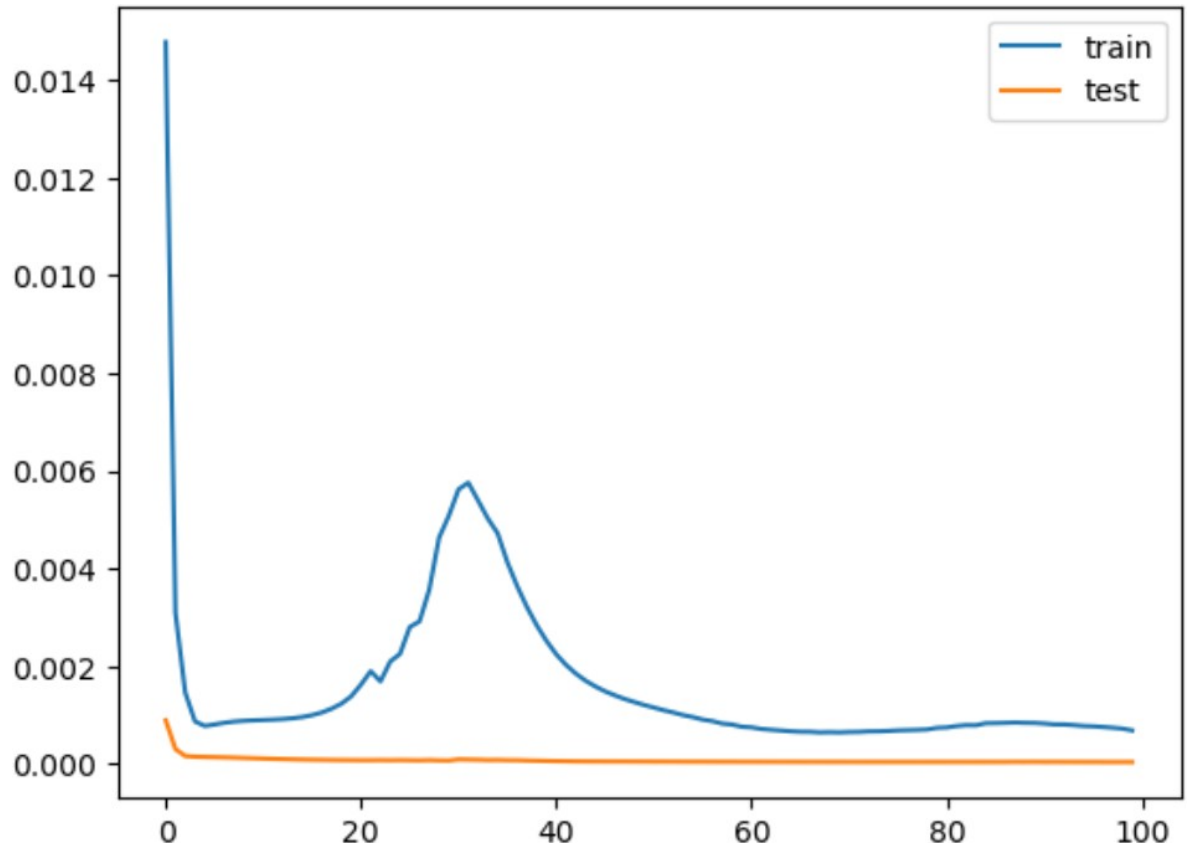
Layer	Neurons
LSTM	128
Dense	1

Πίνακας 5.16.1: Πίνακας αρχιτεκτονικής μοντέλου Multivariate LSTM
για χρονικό διάστημα 15 λεπτών

Layer	Neurons
Epochs	100
Batch_size	128
Learning_rate	0.001

Πίνακας 5.16.2: Πίνακας υπερπαραμέτρων μοντέλου Multivariate LSTM
για χρονικό διάστημα 15 λεπτών

Για την εκπαίδευση του αλγορίθμου χρησιμοποιήθηκαν τα δεδομένα από το Train set 15 min και το αποτέλεσμα του σφάλματος εκπαίδευσης και validation αναπαριστάται στο παρακάτω γράφημα.

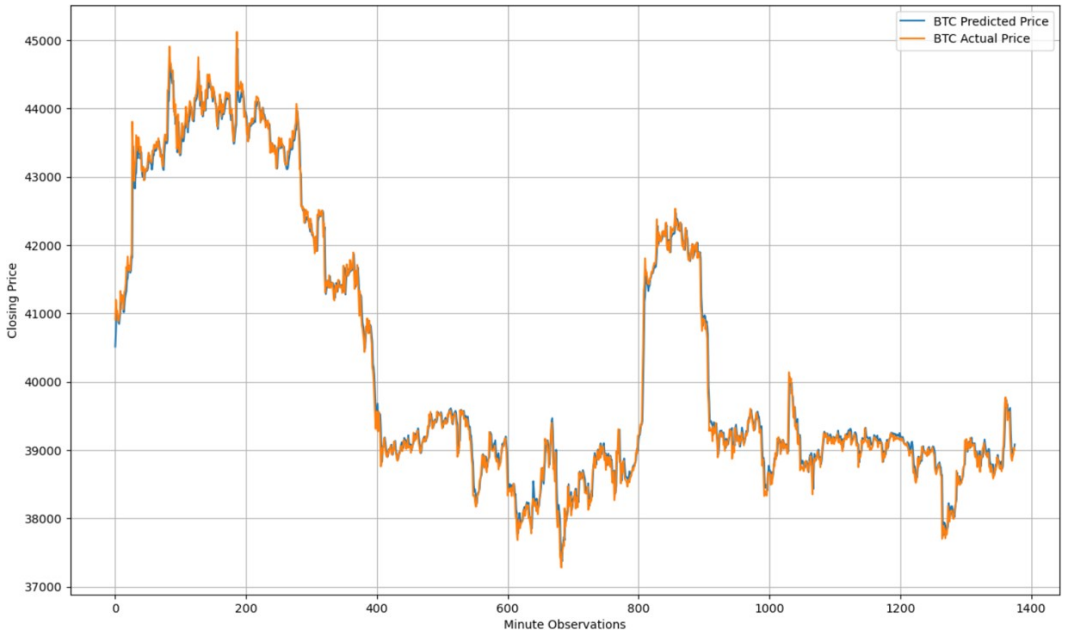


Διάγραμμα 5.18: Γραφική παραστάση train/validation loss της εκπαίδευσης του Multivariate LSTM για χρονικό διάστημα 15 λεπτών

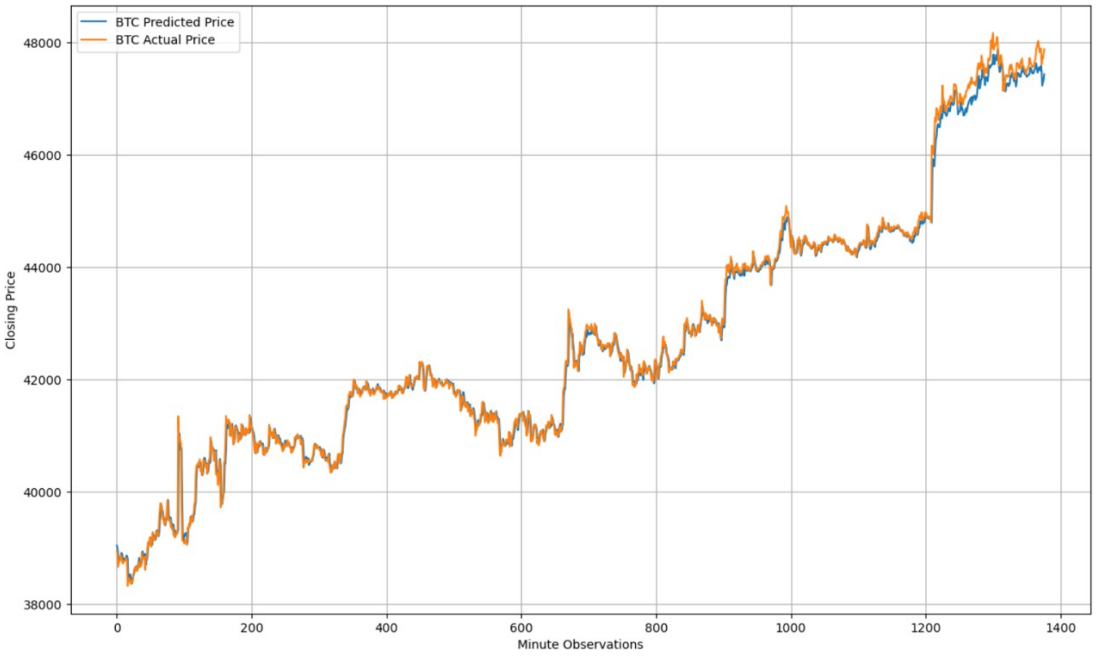
Τέλος πραγματοποιήθηκε πρόβλεψη των δεδομένων στα "Valid set 15 min" και " Test set 15 min" αντίστοιχα προκειμένου να αξιολογήσουμε το μοντέλο και να το συγκρίνουμε με τα αποτελέσματα που εξάγαμε από το ARIMA στην ενότητα 5.3 και το Univariate LSTM της ενότητας 5.4. Η αξιολόγηση του μοντέλου έγινε χρησιμοποιώντας την μετρική RMSE. Στη συνέχεια ακολουθεί ο πίνακας με τα αποτελέσματα καθώς και τα γραφήματα πρόβλεψης / πραγματικής τιμής.

Dataset	RMSE
Valid set 15 min	167.305
Test set 15 min	156.407

Πίνακας 5.17: Πίνακας αποτελεσμάτων Multivariate LSTM για χρονικό διάστημα 15 λεπτών



Διάγραμμα 5.19.1: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Valid set 1min για χρονικό διάστημα 15 λεπτών



Διάγραμμα 5.19.2: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Test set 1min για χρονικό διάστημα 15 λεπτών

5.5.3 Πρόβλεψη χρονικού διαστήματος 60 λεπτών:

Τέλος επαναλαμβάνουμε την αντίστοιχη διαδικασία προετοιμασίας των δεδομένων του χρονικού διαστήματος 60 λεπτών προκειμένου να χρησιμοποιηθούν από το Multivariate LSTM μοντέλο που θα αναπτύξουμε. Ο μετασχηματισμός των δεδομένων θα γίνει ξανά με τέτοιο τρόπο έτσι ώστε να προβλέψουμε την τιμή κλεισίματος του Bitcoin για την χρονική στιγμή (t) συνδυάζοντας τα χαρακτηριστικά των προηγούμενων 15 παρατηρήσεων (n = 15).

Αφού ολοκληρώσουμε την προετοιμασία, ορίζουμε το μοντέλο και προχωράμε στο στάδιο της εκπαίδευσης χρησιμοποιώντας το σύνολο δεδομένων. Πιο αναλυτικά το μοντέλο μας θα αποτελείτε από ένα LSTM layer με 128 νευρώνες, ένα Dense layer με 16 νευρώνες και ένα Dense layer που θα οδηγεί σε ένα νευρώνα όπως φαίνεται και στο πίνακα 5.18.1. Επίσηςτο μοντέλο μας χρησιμοποιεί και τις παρακάτω παραμέτρους κατά το στάδιο της εκπαίδευσης οι οποίες απεικονίζονται στον πίνακα 5.18.2

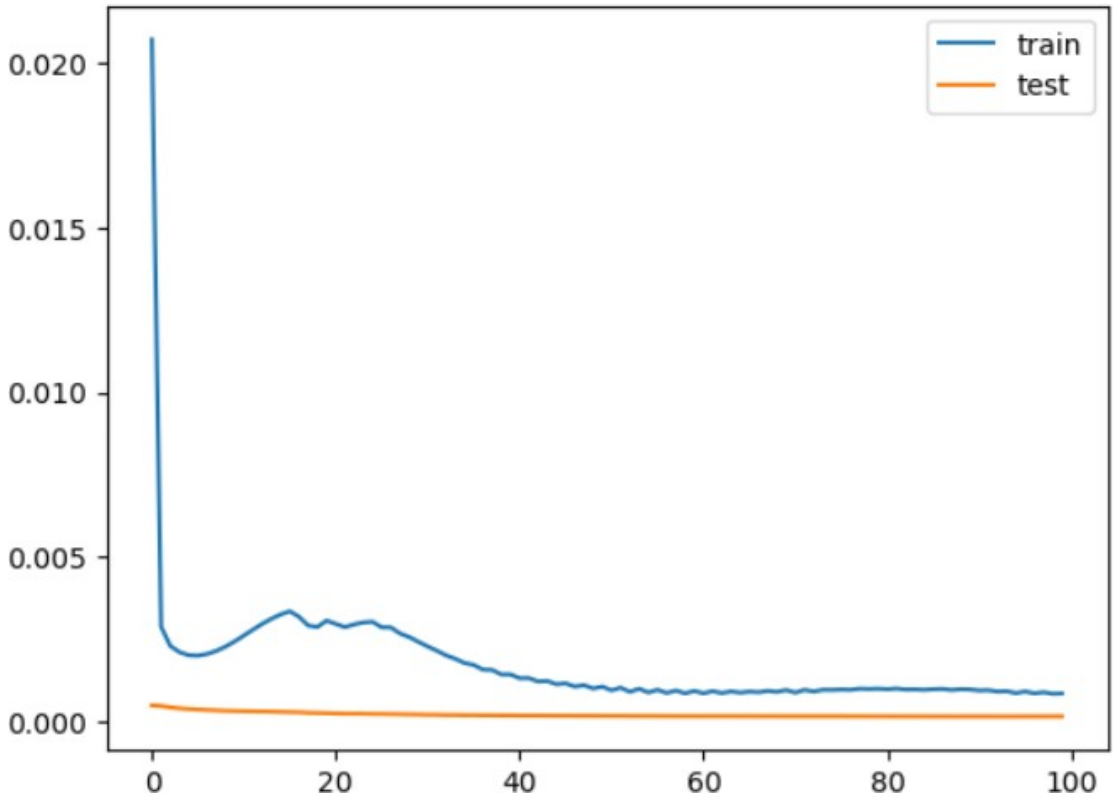
Layer	Neurons
LSTM	128
Dense	16
Dense	1

Πίνακας 5.18.1: Πίνακας αρχιτεκτονικής μοντέλου Multivariate LSTM
για χρονικό διάστημα 60 λεπτών

Layer	Neurons
Epochs	100
Batch_size	32
Learning_rate	0.001

Πίνακας 5.18.2: Πίνακας υπερπαραμέτρων μοντέλου Multivariate LSTM
για χρονικό διάστημα 60 λεπτών

Για την εκπαίδευση του αλγορίθμου χρησιμοποιήθηκαν τα δεδομένα από το Train set 60 min και το αποτέλεσμα του σφάλματος εκπαίδευσης και validation αναπαριστάται στο παρακάτω γράφημα.

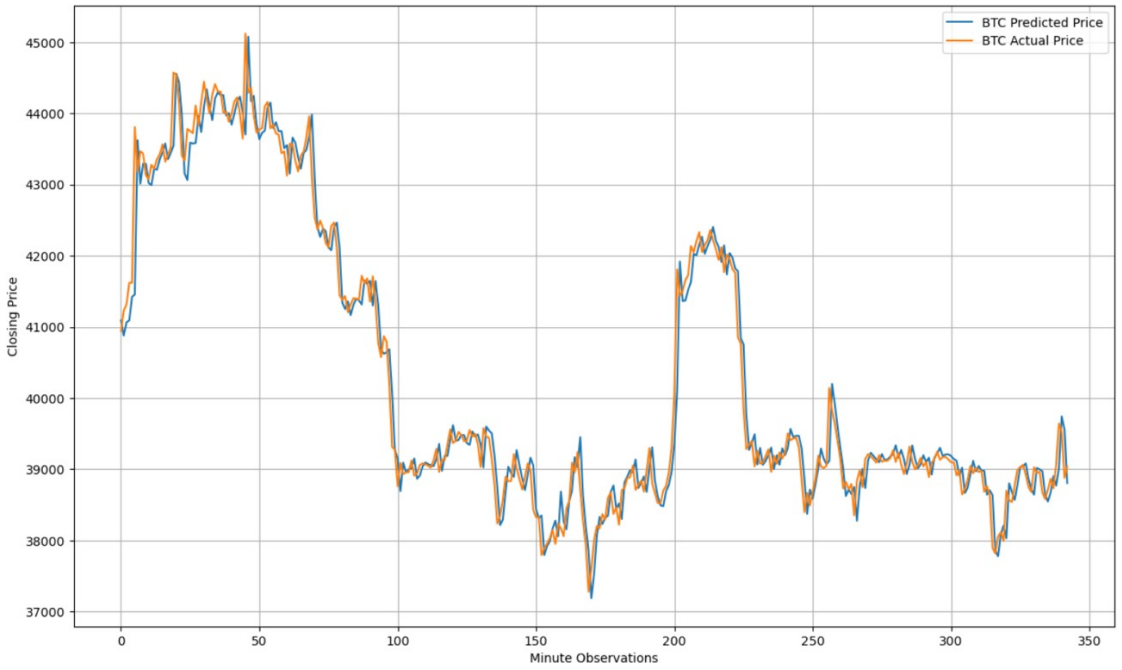


Διάγραμμα 5.20: Γραφική παραστάση train/validation loss της εκπαίδευσης του Multivariate LSTM για χρονικό διάστημα 60 λεπτών

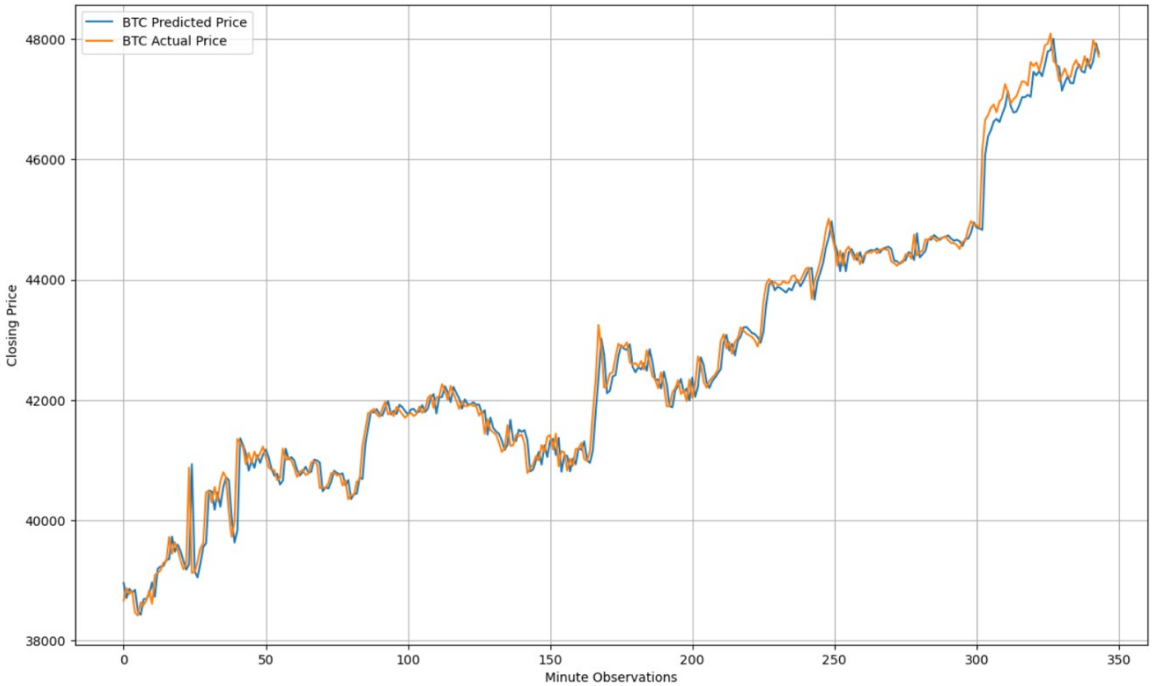
Τέλος πραγματοποιήθηκε πρόβλεψη των δεδομένων στα "Valid set 60 min" και " Test set 60 min" αντίστοιχα προκειμένου να αξιολογήσουμε το μοντέλο και να το συγκρίνουμε με τα αποτελέσματα που εξάγαμε από το ARIMA στην ενότητα 5.3 και το Univariate LSTM της ενότητας 5.4. Η αξιολόγηση του μοντέλου έγινε χρησιμοποιώντας την μετρική RMSE. Στη συνέχεια ακολουθεί ο πίνακας με τα αποτελέσματα καθώς και τα γραφήματα πρόβλεψης / πραγματικής τιμής.

Dataset	RMSE
Valid set 60 min	331.940
Test set 60 min	274.149

Πίνακας 5.19: Πίνακας αποτελεσμάτων Multivariate LSTM για χρονικό διάστημα 60 λεπτών



Διάγραμμα 5.21.1: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Valid set 1min για χρονικό διάστημα 60 λεπτών



Διάγραμμα 5.21.2: Γραφικές παραστάσεις Predicutuon / Actual BTC Price
Του Test set 1min για χρονικό διάστημα 60 λεπτών

6. Αξιολόγηση Αποτελεσμάτων

Σε αυτήν την ενότητα θα συγκρίνουμε τους αλγορίθμους που εκπαιδεύσαμε στο προηγούμενο κεφάλαιο και θα αξιολογήσουμε τα αποτελέσματα πρόβλεψης που προέκυψαν χρησιμοποιώντας κάθε έναν απο αυτούς.

Αρχικά θα χρησιμοποιήσουμε το Μοντέλο ARIMA ως δείκτη (Benchmark) για την αξιολόγηση της πρόβλεψης προκειμένου να εξετάσουμε την απόδοση των νευρωνικών δικτύων που εκπαιδεύσαμε. Στους παρακάτω πίνακες φαίνονται αναλυτικά τα αποτελέσματα όλων των μοντέλων στην πρόβλεψη τους τόσο στο validation dataset όσο και στο test dataset. Η αξιολόγηση όλων των μοντέλων πραγματοποιήθηκε χρησιμοποιώντας την μετρική RMSE [].

- **Αξιολόγηση Πρόβλεψης χρονικού διαστήματος 1 λεπτού:**

Στον παρακάτω πίνακα αναπαριστώνται τα αποτελέσματα πρόβλεψης για το χρονικό διάστημα 1 λεπτού που προέκυψαν και από το validation και από test dataset. Ελέγχονταστα αποτέλεσμα αυτά συμπεραίνουμε ότι την καλύτερη απόδοση την παρουσίασε το Multivariate LSTM τόσο στο validation όσο και στο test dataset, με επιδόσεις 70.832 και 53.803 αντίστοιχα. Την αμέσως καλύτερη επίδοση φαίνεται να παρουσίασε το ARIMA και με μικρή διαφορά το Univariate LSTM.

Model	Dataset	RMSE
ARIMA	Valid set 1 min	72.163
ARIMA	Test set 1 min	58.229
Univariate LSTM	Valid set 1 min	84.434
Univariate LSTM	Test set 1 min	61.178
Multivariate LSTM	Valid set 1 min	70.832
Multivariate LSTM	Test set 1 min	53.803

Πίνακας 6.1: Πίνακας αποτελεσμάτων
για χρονικό διάστημα 1 λεπτού

- **Αξιολόγηση Πρόβλεψης χρονικού διαστήματος 15 λεπτών:**
  Στον παρακάτω πίνακα αναπαριστώνται τα αποτελέσματα πρόβλεψης για το χρονικό διάστημα 1 λεπτού που προέκυψαν και από το validation και από test dataset. Ελέγχοντας τα αποτέλεσμα αυτά συμπεραίνουμε ότι την καλύτερη απόδοση την παρουσίασε το Multivariate LSTM τόσο στο validation όσο και στο test dataset, με επιδόσεις 167.305 και 156.407 αντίστοιχα. Την αμέσως καλύτερη επίδοση φαίνεται να παρουσίασε το ARIMA στο test set ενώ το Univariate LSTM στο validation.

Model	Dataset	RMSE
ARIMA	Valid set 15 min	169.413
ARIMA	Test set 15 min	157.724
Univariate LSTM	Valid set 15 min	168.261
Univariate LSTM	Test set 15 min	177.040
Multivariate LSTM	Valid set 15 min	167.305
Multivariate LSTM	Test set 15 min	156.407

Πίνακας 6.2: Πίνακας αποτελεσμάτων
για χρονικό διάστημα 15 λεπτών

- **Αξιολόγηση Πρόβλεψης χρονικού διαστήματος 60 λεπτών:**

Τέλος ακολουθεί ο πίνακας με τα αποτελέσματα πρόβλεψης για το χρονικό διάστημα 60 λεπτών που προέκυψαν και από το validation και από test dataset. Ελέγχοντας τα αποτέλεσμα αυτά συμπεραίνουμε ότι την καλύτερη απόδοση την παρουσίασε το Multivariate LSTM τόσο στο validation όσο και στο test dataset, με επιδόσεις 331.940 και 274.149 αντίστοιχα. Με το ARIMA και το Univariate LSTM να είναι πάρα πολύ κοντά στα αποτελέσματα πρόβλεψης.

Model	Dataset	RMSE
ARIMA	Valid set 60 min	339.490
ARIMA	Test set 60 min	276.242
Univariate LSTM	Valid set 60 min	332.667
Univariate LSTM	Test set 60 min	277.256
Multivariate LSTM	Valid set 60 min	331.940
Multivariate LSTM	Test set 60 min	274.149

Πίνακας 6.3: Πίνακας αποτελεσμάτων
για χρονικό διάστημα 60 λεπτών

Συνολικά από την παραπάνω ανάλυση προκύπτει ότι όλα τα μοντέλα παρουσίασαν πολύ υψηλές επιδόσεις με την καλύτερη να σημειώνεται στο Multivariate LSTM σε όλα τα διαστήματα πρόβλεψης με πολύ μικρές διαφορές από τα άλλα μοντέλα.

7. Συμπεράσματα και Μελλοντικές Επεκτάσεις

Στα πλαίσια της μελέτης που πραγματοποιήθηκε δοκιμάστηκαν διάφορα εργαλεία και τεχνικές τόσο στην δημιουργία αυτοματοποιημένων και κατανεμημένων συστημάτων όσο και στην ανάπτυξη μοντέλων μηχανικής μάθησης για την πρόβλεψη τιμών σε χρονοσειρές. Δημιουργήθηκε ένα ολοκληρωμένο σύστημα συλλογής και επεξεργασίας δεδομένων το οποίο συλλέγει πληροφορία από διάφορες πηγές την επεξεργάζεται και την αποθηκεύει σε πραγματικό χρόνο. Επίσης, πραγματοποιήθηκε και ανάλυση συναισθηματικού περιεχομένου σε redit posts. Όλες αυτές οι υλοποιήσεις συντέλεσαν στην συλλογή των δεδομένων και την ανάπτυξη διαφόρων μοντέλων που δοκιμάστηκαν από τα οποία την καλύτερη απόδοση την είχε το Multivariate LSTM. Συμπεραίνουμε λοιπόν ότι ένα επαναληπτικό νευρωνικό δίκτυο (RNN) που δέχεται και αξιολογεί πολλαπλά χαρακτηριστικά μπορεί να προβλέψει με σχετικά μικρή απόκλιση την τιμή του Bitcoin. Επίσης συμπεραίνουμε ότι όσο αυξάνεται το διάστημα πρόβλεψης / δειγματοληψίας επηρεάζονται αρνητικά τα αποτελέσματα. Ως μελλοντικές επεκτάσεις στοχεύουμε στην επέκταση του cluster μας προκειμένου να διαμορφωθεί ένα σύστημα end to end πρόβλεψης της τιμής του Bitcoin το οποίο να διαθέτει και αυτοματοποιημένη επανεκπαίδευση των μοντέλων σε πραγματικό χρόνο.

8. Ευχαριστίες

Κλίνοντας θα ήθελα να εκφράσω την ευγνωμοσύνη μου για την υποστήριξη που έλαβα κατά τη διάρκεια της φοίτησης μου στο Μεταπτυχιακό πρόγραμμα σπουδών του Πανεπιστημίου Πειραιώς. Αρχικά θα ήθελα να ευχαριστήσω από καρδιάς την κυρία Αλεξάνδρα Σατασινοπούλου καθώς και το κοινωφελές ίδρυμα Μιχαίλ Ν. Στασινόπουλος - Βιοχάλκο για την κάλυψη των διδάκτρων φοίτησης μου. Στην συνέχεια θα ήθελα να ευχαριστήσω τους παππούδες μου Αθανάσιο και Βασιλική Κοτταρίδη άλλα και την υπόλοιποι οικογένεια μου για όλη την στήριξη και την αγάπη που έχω λάβει από αυτούς σε όλα μου τα βήματα μέχρι σήμερα. Επίσης θα ήθελα να ευχαριστήσω τους καθηγητές μου και κυρίως των κύριο Φιλιππάκη Μιχαίλ που ήταν ο επιβλέπων της διπλωματικής μου εργασίας, καθώς και την Δρ Πούλου Μαριλένα για τη βοήθεια στην επίβλεψη της ανάλυσης των δεδομένων, τη συμβολή της στο πειραματικό μέρος της διατριβής και τα χρήσιμα σχόλιά της στην ανάλυση της έρευνας. Τέλος θα ήθελα να ευχαριστήσω όλους τους συνεργάτες μου που με την βοήθεια τους καταφέραμε να ολοκληρώσουμε αυτό το ταξίδι με επιτυχία.

9. Βιβλιογραφία

1. Satoshi Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash System”, https://bitcoin.org/bitcoin.pdf
2. M.D Devika, “Sentiment Analysis: A Comparative Study on Different Approaches” https://www.sciencedirect.com/science/article/pii/S187705091630463X
3. H.Krishnan, M.Sudheep Elayidom and T. Santhanakrishnan, “MongoDB – a comparison with NoSQL databases” https://www.researchgate.net/publication/327120267_MongoDB_-_a_comparison_with_NoSQL_databases
4. MongoDB official Documentation, “Databases vs. Data Warehouses vs. Data Lakes” https://www.mongodb.com/databases/data-lake-vs-data-warehouse-vs-database
5. Jay Kreps, Neha Narkhede, Jun Rao: Kafka: A Distributed Messaging System for Log Processing. Available online: http://notes.stephenholiday.com/Kafka.pdf
6. R.Saleem and Bo Yuan “Explaining deep neural networks: A survey on the global interpretation methods” https://www.sciencedirect.com/science/article/pii/S0925231222012218
7. Benjamin Lindemann “A survey on long short-term memory networks for time series prediction” https://www.sciencedirect.com/science/article/pii/S2212827121003796
8. Z. Asha Farhath, B. Arputhamary and Dr. L. Arockiam “A SURVEY ON ARIMA FORECASTING USING TIME SERIES MODEL” https://www.ijcsmc.com/docs/papers/August2016/V5I8201626.pdf
9. Wekipedia, “Time Series” https://en.wikipedia.org/wiki/Time_series
10. Avijeet Biswal “Recurrent Neural Network (RNN) Tutorial: Types, Examples, LSTM and More”: https://www.simplilearn.com/tutorials/deep-learning-tutorial/rnn
11. Simeon Kostadinov, “Understanding GRU Networks”: https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be
12. M.Amjad and D.Shah, "Trading Bitcoin and Online Time Series Prediction," in NIPS 2016 Time Series http://proceedings.mlr.press/v55/amjad16.pdf
13. David Garcia and Frank Schweitzer "Social signals and algorithmic trading of Bitcoin": https://royalsocietypublishing.org/doi/pdf/10.1098/rsos.150288

14. R. Chen and M. Lazer, "Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement,
http://cs229.stanford.edu/proj2011/ChenLazer-SentimentAnalysisOfTwitterFeedsForThePredictionOfStockMarketMovement.pdf

15. Alec Go, Richa Bhayani and Lei Huang "Twitter Sentiment Classification using Distant Supervision": https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf
16. Bo Pang, Lillian Lee and Shivakumar Vaithyanathan "Thumbs up? Sentiment Classification using Machine Learning Techniques"
https://www.cs.cornell.edu/home/llee/papers/sentiment.pdf
17. S M Raju and Ali Mohammad Tarif "Real-Time Prediction of BITCOIN Price using Machine Learning Techniques and Public Sentiment Analysis"
https://arxiv.org/pdf/2006.14473.pdf
18. Patrick Jaquart, David Dann and Christof Weinhardt "Short-term bitcoin market prediction via machine learning"
https://www.sciencedirect.com/science/article/pii/S2405918821000027
19. Jaquart P, Dann D, Martin C. "Machine learning for bitcoin pricing - a structured literature review."
https://library.gito.de/wp-content/uploads/2021/08/B4_manuscript_final.pdf
20. Patrick Jaquart, David Dann & Christof Weinhardt "Using Machine Learning to Predict Short-Term Movements of the Bitcoin Market":
https://link.springer.com/chapter/10.1007/978-3-030-64466-6_2
21. Okeanos Official Documentation:
https://okeanos.grnet.gr/support/user-guide/cyclades-how-can-i-access-all-my-vms-using-one-public-ip-nat/
22. CoinMarketCap API Documentation:
https://coinmarketcap.com/api/documentation/v1/
23. Bitstamp API Documentation:
https://www.bitstamp.net/api/
24. Pushshift Reddit API Documentation:
https://github.com/pushshift/api
25. Jason Brownlee, “Multivariate Time Series Forecasting with LSTMs in Keras”:
https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/
26. Jason Brownlee, “How to Develop LSTM Models for Time Series Forecasting”
https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/
27. Αρθρο με pyton NLP. on Tweets:
https://www.analyticsvidhya.com/blog/2022/07/sentiment-analysis-using-python/