Usporedba algoritama za predviđanje cijena kriptovaluta

Brkić, Goran

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 502

USPOREDBA ALGORITAMA ZA PREDVIĐANJE CIJENA KRIPTOVALUTA

Goran Brkić

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 502

USPOREDBA ALGORITAMA ZA PREDVIĐANJE CIJENA KRIPTOVALUTA

Goran Brkić

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

DIPLOMSKI ZADATAK br. 502

Pristupnik:	**Goran Brkić (0036519085)**
Studij:	Računarstvo
Profil:	Računalno inženjerstvo
Mentorica:	doc. dr. sc. Martina Antonić

Zadatak: **Usporedba algoritama za predviđanje cijena kriptovaluta**

Opis zadatka:

Predviđanje vremenskih serija ili predviđanje budućih vrijednosti na temelju povijesnih podataka ključno je pri donošenju odluka u mnogim područjima, uključujući financije, ekonomiju i znanosti o okolišu. Rast računalnih alata rezultirao je razvojem brojnih modela predviđanja, od kojih je svaki specijaliziran za zadovoljavanje određenih svojstava vremenskih podataka pa je odabir prikladnog modela predviđanja postao ključan. Vaš je zadatak napraviti pregled različitih metoda koje se koriste za predviđanje vremenskih serija te osmisliti vlastiti algoritam baziran na stablima. Koristeći javno objavljene skupove podataka o kriptovalutama, potrebno je implementirati i usporediti sve predložene metode te pokazati njihovu efikasnost prilikom predviđanja promjene cijena kriptovaluta koristeći programski jezik Python.

Rok za predaju rada: 28. lipnja 2024.

SADRŽAJ

1. Uvod

Burza je organizirano tržište gdje se vrši kupovina i prodaja vrijednosnih papira, roba, deviza, ročnica, opcijskih ugovora, kriptovaluta i drugih instrumenata trgovanja. Trgovanje dionicama seže čak do 1531. u Nizozemskoj, a početkom burzovnog trgovanja može se smatrati osnivanje London Stock Exchangea (LSE), prve burze na svijetu, 1773. godine. Danas, više od 250 godina kasnije, ljudi i dalje pokušavaju pronaći načine predviđanja burze kako bi zaradili što više novaca. Može li se kretanje burze uspješno i konzistentno predviđati ostaje otvoreno pitanje.

Burza je (u svom najosnovnijem obliku, bez popratnih troškova) *zero-sum* igra, odnosno ima jednak broj gubitnika i pobjednika. U igri trgovanja na burzi, svaka razmjena dionice, kriptovalute ili nečeg trećeg ima pobjednika i gubitnika. U situaciji gdje vrijednost trgujućeg instrumenta poraste, oni koji su ga prethodno kupili su pobjednici. Padne li mu vrijednost, oni postaju gubitnici. Trgovac (engl. *trader*), ili investitor, može u danu više puta prijeći iz pozicije dobitnika u gubitnika i obrnuto no razlikuju se u tome što će investitor uobičajeno kupiti neki instrument te ga držati duži period vremena dok ne postane pobjednik ili veliki pobjednik, dok će trgovac u danu, ili kraćem periodu, više puta kupiti i prodati na temelju svoga predviđanja cijene i nadati se da će završiti dan s većim profitom nego gubitkom.

Proces predviđanja hoće li cijena ići gore ili dolje nečega što se kreće gotovo nasumično težak je zadatak, pogotovo uz snažne fluktuacije uzrokovane vijestima vezanim uz instrument (direktno ili indirektno).

Smisliti algoritamski pristup koji će uspješno i autonomno donositi odluke treba li neki instrument kupiti ili prodati zanimljiva je tema inženjerima, ekonomistima i brojnim drugima. Zadnjih godina, kako je došlo do snažnog razvoja strojnog učenja te umjetne inteligencije, ideja je stvoriti algoritam koji će na temelju podskupa podataka i raznih faktora, sasvim samostalno i automatizirano, donijeti odluku o kupnji i prodaji te na taj način ostvariti profit koji ljudi sami, barem u toj mjeri, ne mogu.

1.1. Pristup

Postoje brojne vrste algoritama za predviđanje burze i automatizirano trgovanje, primjerice, najjednostavniji jest onaj koji samo prati vijesti, kvartalna izvješća poduzeća, kamatnu stopu i sl. Takav pristup odličan je za predviđanje vrijednosti dolara, odnosno kretanja indeksa dolara (engl. *US Dollar Index*, DXY) – povišena kamatna stopa privući će strane investitore da investiraju u državnu imovinu SAD-a, kao što su obveznice i dionice, očekujući veće prinose što će povećati potražnju za dolarom zbog čega moraju zamijeniti svoju valutu za dolar što će rezultirati povećanjem vrijednosti dolara. I obratno, niža kamatna stopa će odbiti investitore što rezultira manjom potražnjom za dolarom što će onda potencijalno rezultirati smanjenjem vrijednosti samog dolara.

Druga vrsta bi bila algoritmi koji trguju na temelju indikatora, najjednostavniji od njih na temelju pomičnog prosjeka (engl. *moving average*, MA) - indikatora koji se računa uzimajući aritmetičku sredinu cijene dionice nad određenim periodom vremena. Koristeći MA nad dva različita perioda kao što su 20 dana i 50 dana, može se čekati trenutak u kojem će krivulja MA(20) prijeći preko krivulje MA(50) što obično signalizira moć kupaca nad prodavačima i prognozira pozitivno kretanje promatrane dionice. Ovakav algoritam mogao bi se izvršavati bez prestanka nad velikim brojem dionica, čekajući da se dogodi navedeni slučaj i kupiti dionicu, bez ikakve interakcije stvarne osobe.

Kompleksniji algoritmi autonomno prate kretanje cijene promatrajući takozvane svijeće u Candlestick grafu (ili drugim vrstama grafova) koji jednom svijećom prikazuju otvarajuću i zatvarajuću cijenu te maksimalnu i minimalnu cijenu postignutu u određenom periodu – primjerice jedna svijeća može prikazivati kretanje cijene unutar sat vremena, a onda 10 svijeća prikazuje kretanje cijene u razdoblju od 10 sati. Neki vjeruju da se izgled ovih grafova i način na koji se cijena kreće ponavlja te se zato može istrenirati algoritam tako da traži specifičan uzorak kretanja cijene i na temelju toga predvidi buduće kretanje cijene.

Ovaj rad uzima drugi navedeni pristup te ga proširuje na način da koristi 10 različitih indikatora koje koristi kao ulaz raznim modelima strojnog učenja, uspoređuje ih te pokušava pronaći model koji najbolje predviđa kretanje cijene kriptovaluta i ustanoviti je li to na ovaj način uopće moguće.

1.2. Cilj

Cilj ovog rada je razviti i usporediti nekoliko različitih modela strojnog učenja za predikciju kretanja cijene kriptovaluta. Od velikog je interesa vidjeti jesu li kriptovalute teže ili jednostavnije za predviđati od tradicionalnih instrumenata te funkcioniraju li s kriptovalutama

modeli koji su uspješni s dionicama.

U radu su uspoređeni različiti modeli strojnog učenja u zadatku predviđanja smjera kretanja cijene. Dakle, u prvom planu je točno predviđanje smjera budućeg kretanja kriptovalute, a ne točna cijena jer se zadatkom regresije mogu dobiti varljivo dobri rezultati samim time da model pogađa cijenu u blizini prethodne.

Promatrane kriptovalute su Bitcoin (BTC), Ethereum (ETH) te Chainlink (LINK) protiv kriptovalute Tether (USDT) koja je vezana uz dolar s idejom da je vrijednost jednog USDT uvijek jednaka jednom američkom dolaru. U radu se koriste samo podaci kriptovalutne burze Binance i to parovi BTC/USDT, ETH/USDT te LINK/USDT na promptnom tržištu (engl. *spot market*). Razlog korištenja ova tri valutna para je prvenstveno njihova popularnost, a zatim to što svaki reprezentira jednu od kategorija tržišne kapitalizacije – velika, srednja i mala, te je zanimljivo vidjeti utječe li količina novca u ovoj *zero-sum* igri na lakoću predviđanja.

1.3. Struktura rada

Rad je strukturiran na sljedeći način: u drugom poglavlju objašnjena je ideja burze i burzovnog poslovanja te su pojašnjena tehnička i fundamentalna analiza, a zatim i teorija slučajnog hoda. U trećem poglavlju pojašnjeno je pet modela strojnog učenja koji su se koristili za predviđanje cijene kriptovaluta. Četvrto poglavlje sadrži objašnjenja, matematičke izračune te vizualizacije 10 tehničkih indikatora koji su kasnije dovedeni na ulaze modela. U petom poglavlju objašnjeno je kako su podatci o cijeni tri valutna para (BTC/USDT, ETH/USDT i LINK/USDT) prikupljeni, organizirani i pripremljeni za dovođenje na ulaze modela. Naposljetku, u šestom poglavlju su dani rezultati za svaki od pet modela strojnog učenja, uspoređeni su njihovi rezultati te su dane smjernice za budući rad dok je u sedmom poglavlju dan zaključak.

2. Pozadina

Razumijevanje kako tržište kapitala funkcionira ključno je za analizu i predviđanje tržišnih kretanja. Burza predstavlja mjesto gdje se trguje financijskim instrumentima poput dionica i valuta, a znanje o njezinom radu pruža temelje za dublje analize koje slijede. U nastavku su pojašnjeni osnovni koncepti vezani uz burzovno poslovanje te metode za predviđanje koje se najčešće koriste.

2.1. Burza

Kako pri kupnji dionice osoba stvarno posjeduje dio kompanije, tako i na promptnom tržištu na burzi Binance osoba stvarno kupuje valutu te ju posjeduje u svome novčaniku. Kupuje ju na način da se transakcija dogodi između kupca i prodavača tako što će kupac navesti za koju cijenu želi kupiti valutu te koliko njih, a prodavač će navesti koliko prodaje i po kojoj cijeni. Skupina svih ljudi što kupuju i prodaju tvore knjigu naloga (engl. *order book*). Transakcija je zapravo uparivanje naloga s obje strane ove knjige, a trenutna cijena je cijena po kojoj je izvršeno zadnje uparivanje. Kako se cijena pomiče, tako se uparuju različiti nalozi unutar ove knjige.

Postoje i terminski ugovori (engl. *futures*) koji omogućuju trgovanje većim iznosima nego što osoba zapravo posjeduje, no ovakvo trgovanje uvodi veću volatilnost, manipulaciju te druge probleme koji otežavaju predviđanje te osim toga burza Binance u trenutku pisanja ovog rada ne nudi javno dostupne povijesne podatke ovakvog trgovanja.

Postoje brojne burze za razmjenu kriptovaluta, međutim Binance je, u vrijeme pisanja, najveća globalna kriptovalutna burza po dnevnom volumenu trgovanja [1]. Na sreću, sve burze prikazuju gotovo identično kretanje cijene jer postojanje bilo kakve neravnoteže u cijeni između različitih burzi iskorištavaju automatizirani programi za arbitražu i na taj način ponovno uvode ravnotežu između burzi. To znači da valjanost istreniranih modela ostaje čak i ako se navedena burza zatvori kao što se događalo s brojnim drugima.

2.2. Predviđanje burze

Dva najpoznatija pristupa analizi burzovnih instrumenata su fundamentalna analiza i tehnička analiza.

Fundamentalna analiza podrazumijeva pomno proučavanje financijskih izvještaja poduzeća kao što su bilanca, račun dobiti i gubitka i izvještaj o novčanom tijeku kako bi se dobilo što bolje saznanje o financijskom stanju poduzeća. Glavni promatrani elementi su: prihod, neto dohodak, imovina, dug i operativni novčani tok [2]. Kod kriptovaluta to može biti istraživanje o kvaliteti projekta, koliko je dobra ideja, koliko je kvalitetno novo rješenje koje kriptovaluta rješava, koliko je kvalitetan kod koji pokreće kriptovalutu, kako je riješeno rudarenje, koliko stvaratelji ulažu u marketing i sl.

Tehnička analiza se može definirati kao umjetnost i znanost predviđanja budućih cijena promatrajući prethodno kretanje cijene. Tehnička analiza temelji se na analiziranju trenutne ponude i potražnje promatranog instrumenta. Najčešće se na promatrani graf osim cijene postavlja volumen, otvoreni interes i drugi indikatori te se pokušava prepoznati poznata struktura kretanja cijene ili mjesta ponude i potražnje te tako predvidjeti kretanje cijene [3]. Često se koriste i razne statističke metode.

Postoji i teorija slučajnog hoda (engl. *random walk theory*) koja kaže da se cijena dionica i drugih tržišnih instrumenata kreće neovisno o prethodnoj kretnji i da se ne može predvidjeti tehničkom analizom pa čak ni fundamentalnom analizom. Popularizirao ju je Malkiel svojom knjigom iz 1973, *A Random Walk Down Wall Street* [4]. Postoje brojna istraživanja koja potvrđuju ovu teoriju, npr. Jiao i Jakubowicz su pokušali s modelima slučajne šume, stabala, umjetnih neuronskih mreža i logističkom regresijom predvidjeti kretanje 463 dionica indeksa The Standard and Poor's 500 (S&P 500) te došli do zaključka da tako nešto nije moguće [5]. Chitenderu, Maredza i Sibanda su pokušali pronaći pravilnost u kretanju glavne burze u Johannesburgu koristeći se testovima jediničnog korijena, testom autokorelacije te modelom autoregresivnog integriranog pomičnog prosjeka (ARIMA) te su pokazali kako čak ni Južnoafrička burza ne omogućuje predviđanje i podržava teoriju slučajnog hoda burze [6].

3. Strojno Učenje

Strojno učenje je metoda analize podataka koja na automatiziran način gradi analitički model za određeni problem na temelju seta podataka koji opisuje taj problem.

Algoritmi strojnog učenja mogu provoditi i regresiju i klasifikaciju. Algoritama je mnogo te se uvijek razvijaju novi, a načini na koje provode regresiju ili klasifikaciju su često vrlo različiti. No, svi modeli strojnog učenja dijele ideju "treniranja" modela nad skupom podataka tako da se stvori analitički model za specifičan problem. Jednom kad je model istreniran, predstavlja mu se novi skup podataka koji još nije koristio te se od njega traži da obavi zadatak regresije ili klasifikacije. Algoritmi korišteni u ovom radu te njihove osnovne ideje objašnjene su u nastavku.

3.1. Logistička regresija

Logistička regresija je probabilistički diskriminativni model. Ovaj model pronalazi hiperravninu koja maksimizira izglednost skupa podataka [7]. Iako naslov naslućuje da se provodi postupak regresije, riječ je o klasifikaciji. Model je diskriminativan, ali daje izlaz koji ima vjerojatnosno tumačenje. Logistička regresija izravno modelira aposteriornu vjerojatnost $P(C_j|x)$, dok generativni modeli tu vjerojatnost modeliraju posredno preko zajedničke gustoće $p(x, C_j)$ [7].

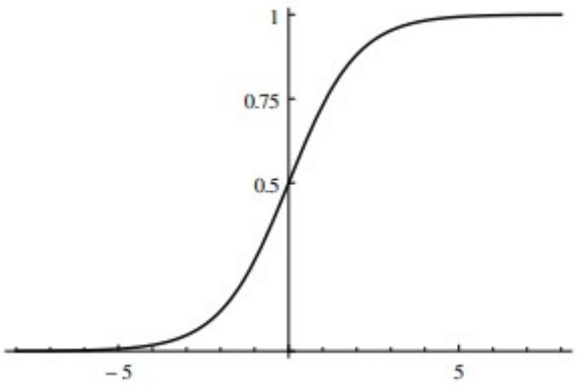
Funkcija $\sigma(\alpha)$ je logistička ili sigmoidalna funkcija definirana kao

$$\sigma(\alpha) = \frac{1}{1 + e^{-\alpha}} \tag{3.1}$$

gdje je (za primjer klasifikacije s dvije klase):

$$\alpha = \ln \frac{p(x|C_1)P(C_1)}{p(x|C_1)P(C_1) + p(x|C_2)P(C_2)}, \tag{3.2}$$

Logistička funkcija prikazana je na slici 3.1. Funkcija preslikava sve realne brojeve na konačan interval [0, 1]. Ova funkcija ima ulogu aktivacijske funkcije.



Slika 3.1: Logistička ili sigmoidalna funkcija

Logistička regresija je probabilistički model jer se njegov izlaz može tumačiti kao posteriorna vjerojatnost klase [7]. Važno je napomenuti da se ta vjerojatnosna interpretacija oslanja na pretpostavku normalne distribucije klasa i da klase imaju dijeljenu varijancu. Ovaj model ne kažnjava ispravno klasificirane primjere koji se nalaze daleko od granice; za sve takve primjere aposteriorna vjerojatnost bit će blizu jedinice [7].

U nastavku slijedi izraz za poopćeni linearni model [7]:

$$h(x) = P(C_1|x) = \sigma(\mathrm{w}^T\mathbf{x} + w_0) = \sigma(\tilde{\mathrm{w}}^T\tilde{\mathbf{x}}) \tag{3.3}$$

gdje je:

$$\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) \tag{3.4}$$

$$w_0 = -\frac{1}{2}\boldsymbol{\mu_1}^T\Sigma^{-1}\boldsymbol{\mu_1} + \frac{1}{2}\boldsymbol{\mu_2}^T\Sigma^{-1}\boldsymbol{\mu_2} + ln\frac{P(C_1)}{P(C_2)} \tag{3.5}$$

Ukoliko je cilj klasifikacija, a ne i modeliranje izglednosti pojedinih klasa niti generiranje primjera, onda nam diskriminativni modeli omogućavaju da to ostvarimo s mnogo manje parametara, a time i manjom složenosti [7]. Zato je ovaj model, makar jednostavan, često dobar izbor.

3.2. Stroj potpornih vektora (SVM)

Za razliku od logističke regresije, model stroja potpornih vektora (engl. *Support Vector Machine*, SVM) nema nikakvu probabilističku interpretaciju pa je u tom smislu bliže perceptronu, koji pronalazi proizvoljnu hipotezu konzistentnu s primjerima za učenje [7]. Međutim, SVM problemu proizvoljne hipoteze pristupa tako što primjenjuje kriterij maksimalne margine (engl. *maximum margin*). Kako bi model što bolje generalizirao, granica se postavlja

tako da razmak između pozitivnih i negativnih primjera bude što veći. Naziv SVM dolazi od toga što je hiperravninu koja klasificira ulazne podatke moguće prikazati kao kombinaciju odabranih vektora iz skupa za učenje koje nazivamo potporni vektori [7].

Treniranje modela svodi se na rješavanje problema kvadratnog programiranja koje je definirano sljedećim izrazom [7]:

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y^{(i)} y^{(j)} \phi(\mathbf{x}^{(i)})^T \phi(\mathbf{x}^{(j)}) \tag{3.6}$$

uz:

$$\alpha_i \geq 0, \quad i = 1, ..., N$$
$$\sum_{i=1}^{N} \alpha_i y^{(i)} = 0$$

Testiranjem se dobije N-dimenzijski vektor parametra α. Klasificiranje novog primjera provodi se izračunavanjem $sgn(h(x))$.

Dodatnu učinkovitost SVM postiže primjenom jezgrenih funkcija koje taj model svrstavaju u skupinu modela koji se nazivaju **jezgreni strojevi** (engl. *kernel machines*). Prema Coverovom teoremu većina je problema nelinearna dok je SVM uglavnom linearni model. Zato kada je ulazni prostor gusto napunjen i zbog toga teško linearno odvojiv nastaje problem. Za rješavanje ovakvog problema koristi se **jezgreni trik**: ako su linearni modeli dovoljno dobri za n $\gg$ N, onda se problem može preslikati u prostor više dimenzije u kojem je vjerojatnije pronaći odgovarajuće rješenje. Dakle, transformira se problem umjesto modela. Neke od osnovnih jezgrenih funkcija su: linearna jezgra, polinomijalna jezgra i radijalna bazna funkcija. Linearni model bez aktivacijske funkcije dan je u nastavku:

$$h(x) = \mathbf{w}^T \boldsymbol{\sigma}(\boldsymbol{x}) + w_0 \tag{3.7}$$

gdje je w vektor normalan na hiperravninu.

Optimizacijski se problem svodi na maksimizaciju

$$\underset{\mathbf{w},\, w_0}{\text{argmax}} \frac{1}{||\mathbf{w}||} \tag{3.8}$$

Optimalan model se odabire na način da se optimira hiperparametar koji određuje složenost modela. Manja vrijednost tog parametra znači dozvoljavanje više pogrešaka, a veća vrijednost manje pogrešaka. Parametar se tipično optimira unakrsnom provjerom.

3.3. Slučajna šuma

Algoritam slučajne šume (engl. *Random forest*) je vrsta *bootstrap* agregacije koja gradi veliki broj nekoreliranih stabala. *Bootstrap* agregacija, još poznata i kao *bagging*, je metoda ansambala često korištena s ciljem smanjivanja varijance pri treniranju skupa podataka s velikim šumom. Slučajna šuma je *bagging* sa stablima odluke kao osnovnim klasifikatorima, a dodatna diverzifikacija postiže se slučajnim odabirom podskupa značajki [7].

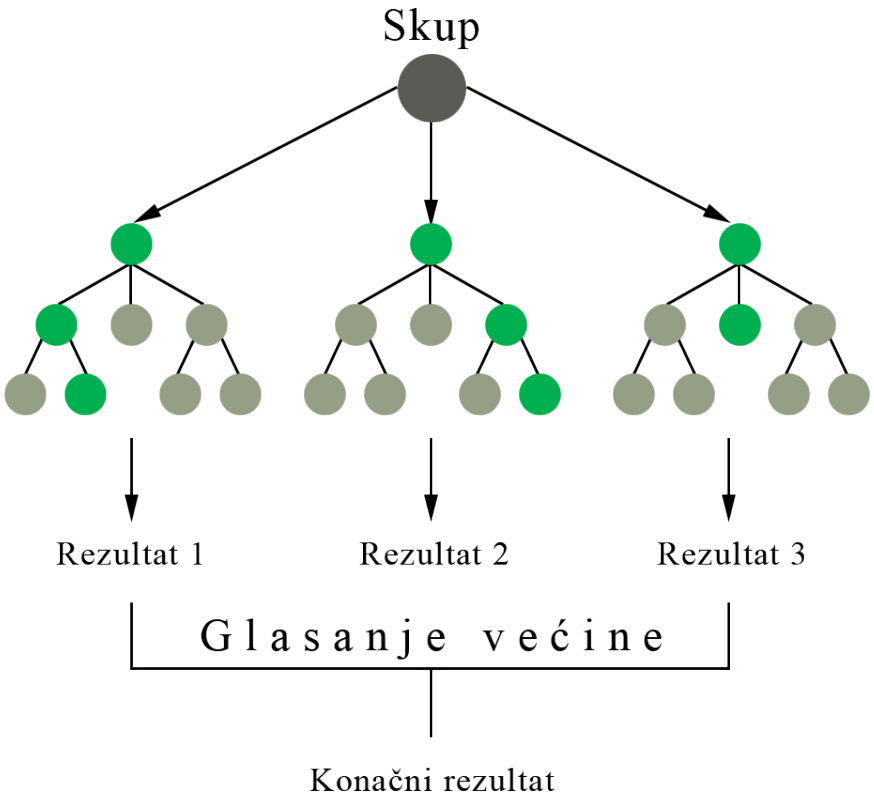
Treniranje se provodi na L poduzoraka skupa za učenje dobivenih uzorkovanjem s ponavljanjem. Vjerojatnost uključivanja primjera je

$$1 - \lim_{N\to\infty}(1 - \frac{1}{N})^N = 1 - \frac{1}{e} \approx 0.632$$

Cilj algoritma slučajnih šuma je reducirati varijancu tako što će smanjiti korelaciju između stabala [8]. Konačna predikcija postiže se većinskim glasanjem. Jedna od važnih značajki algoritma slučajne šume je "out of bag" (OOB) skup podataka. OOB sastoji se od 33% podataka koji su zanemareni tijekom uzorkovanja. OOB procjena greške gotovo je jednaka kao unakrsna validacija stabala te se stoga može koristiti za procjenu greške šume [8]. Dakle, trening se prekida kada se OOB greška stabilizira. Zbog jednostavnosti treniranja i podešavanja, slučajna šuma jedan je od najpopularnijih algoritama. Algoritam izgradnje slučajne šume dan je u nastavku te je model slučajne šume vizualiziran slikom 3.2.

Slučajna šuma

1. $forest \leftarrow \varnothing$
2. $for(j = 1; j <= L; j++)$
3. $\quad \mathcal{D}_j \leftarrow$ *bootstrap* uzorak
4. $\quad \mathcal{F}_j \leftarrow$ odabir n' značajki
5. $\quad t_j \leftarrow$ treniraj stablo odluke na $\mathcal{D}_j$ s $\mathcal{F}_j$
6. $\quad forest \rightarrow forest \cup t_j$



Slika 3.2: Slučajna šuma

3.4. XGBoost

XGBoost naziv je modela za strojno učenje koje dolazi od skraćenice engleskog naziva *Extreme Gradient Boosting*. Radi se o vrlo učinkovitom algoritmu strojnog učenja kojeg je razvio Tianqi Chen 2016. godine [9]. Model je stekao veliku popularnost zbog svojih performansi, fleksibilnosti i preciznosti pri rješavanju složenih problema modeliranja. Algoritam je implementacija *gradient boost* algoritma stabla odluke. Za razliku od *bagging* algoritama koji stvaraju raznolik skup nasumično odabranih podataka te uče sa svakim od njih, "boosting" algoritmi primjenjuju slijedno učenje na pogreškama prethodnih epoha - cilj je doći do jakog klasifikatora krenuvši od slabih.

Dva temeljna principa XGBoosta su **pojačavanje gradijenta (engl. *gradient boosting*)** te **korištenje stabala odluke kao osnovnih modifikatora**.

Pojačavanje gradijenta je tehnika ansambala koja sekvencijalno stvara modele s ciljem da svaki novi model poboljšava greške koje je napravio prošli. To se postiže minimiziranjem funkcije gubitka (engl. *loss function*) gradijentnim spustom gdje se novi model trenira na

rezidualnim greškama prethodnih modela. Kao funkcija gubitka najčešće se koristi diferencijabilna konveksna funkcija koja ujedno i vodi optimizacijski proces.

Za razliku od tradicionalnih stabala odluke koja rastu vrlo pohlepno i brzo, XGBoost implementira "Skicu težinskih kvantila" (engl. *Weighted Quantile Sketch*) koja omogućuje efikasnije pronalaženje razdiobe, tehniku "Pretraživanje razdjelnika uz svjesnost rijetkosti" (engl. *Sparsity-aware Split Finding*) koja omogućuje pronalazak nedostajućih vrijednosti u skupu podataka računanjem informacijske dobiti te koristi L1 (Laso) i L2 (Greben (engl. *Ridge*)) za regularizaciju u funkciji gubitka penalizirajući pretjerano kompleksne modele s čime smanjuje prenaučenost.

L1 regularizacija zbraja apsolutne vrijednosti koeficijenata modela te ih pridodaje funkciji gubitka. Na ovaj način "gura" koeficijente prema nuli, odnosno penalizira model ako ima koeficijente različite od nule. L1 regularizacija posebno je korisna pri treniranju visokodimenzionalnog skupa podataka jer se efektivno ignoriraju značajke manje vrijednosti. Matematički, L1 regularizacija može se zapisati sljedećom formulom:

$$L1 = \lambda \sum_i |w_i| \tag{3.9}$$

gdje je λ regularizacijski parametar koji kontrolira snagu regularizacije, a w_i predstavlja individualne koeficijente modela.

L2 regularizacija pridodaje zbroj kvadratnih vrijednosti težina funkciji gubitka. Za razliku od L1 regularizacije, L2 ne dovodi koeficijente do nule, ali potiče da su mali. Prednost L2 regularizacije dolazi na vidjelo kada postoji korelacija između značajki dovedenih na ulaz. Matematički, L2 regularizacija može se zapisati sljedećom formulom:

$$L2 = \lambda \sum_i (w_i)^2 \tag{3.10}$$

Funkcija koja opisuje XGBoost dana je u nastavku:

$$Obj(\Theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3.11}$$

gdje je:

$$\Theta = \text{Set svih parametara modela}$$
$$L(y_i, \hat{y}_i) = \text{Funkcija gubitka}$$
$$\Omega(f_k) = \text{Regularizator}$$

Regularizator obično poprima ovakvu formu:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3.12}$$

gdje je:

$T =$ Broj listova

$\gamma =$ Regularizacijski parametar za broj listova

$\lambda =$ Regularizacijski parametar za težinu listova

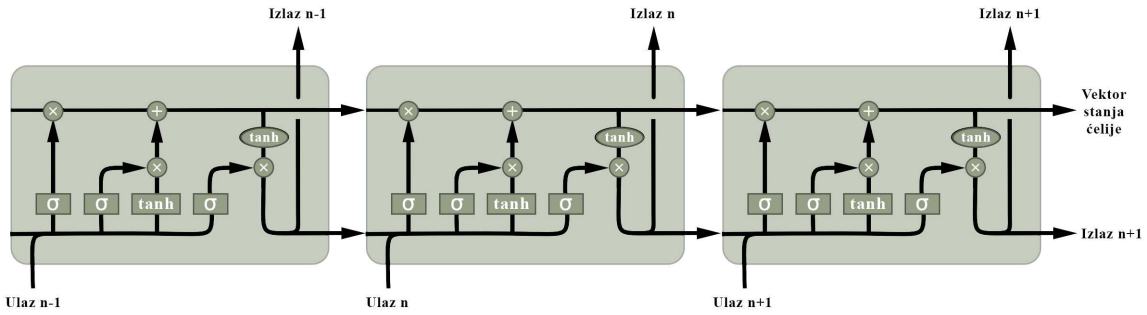
$w_j =$ Težina j-tog lista

3.5. Mreže dugotrajne kratkoročne memorije (LSTM)

Mreže dugotrajne kratkoročne memorije (engl. *Long Short-Term Memory*, LSTM) su specijalni oblik neuronskih mreža s povratnim prijenosom (RNN) osmišljene s idejom očuvanja ovisnosti između podataka kroz veći broj iteracija kod obrade sekvencijalnih podataka. Model su predstavili Hochreiter i Schmidhuber 1997. [10] kao rješenje problema nestajućeg i eksplodirajućeg gradijenta.

Problem s tradicionalnim RNN je da se fokusiraju samo na kratkoročne ovisnosti podataka upravo zbog eksponencijalnog raspadanja gradijenta. Do nestajanja gradijenta dolazi tijekom povratne propagacije, primjerice ako se gradijent računa pomoću pravila lanca (engl. *chain rule*), a hiperbolni tangens ima gradijente u rasponu [-1, 1]. Potrebno je pomnožiti n malih vrijednosti početnih slojeva u n-slojnoj mreži što rezultira eksponencijalnim smanjivanjem gradijenta te vrlo sporim treniranje prvih slojeva. Suprotno, može doći do eksplozije gradijenta ako se množe gradijenti kroz mrežu kojoj slojevi imaju vrijednosti veće od 1.0. Rezultat eksplodirajućeg gradijenta su nestabilne neuronske mreže ili u najgorem slučaju NaN vrijednosti težina koje se dalje ne mogu trenirati.

Arhitektura LSTM mreža rješava navedene probleme na način da uvodi memorijske ćelije koje čuvaju informaciju kroz duži period vremena odnosno kroz više iteracija. Svaka memorijska ćelija sastoji se od tri komponente: ulazna vrata, vrata zaboravljanja i izlazna vrata. Ulazna vrata određuju koliko će nove informacije ući u ćeliju. Vrata zaboravljanja kontroliraju do koje mjere će se informacija zaboraviti. Izlazna vrata reguliraju izlaz iz ćelije do skrivenog stanja (vektor koji čuva unutarnju memoriju mreže među koracima). Slikom 3.3. može se vidjeti tri ovakve ćelije te kako su međusobno povezane.



Slika 3.3: LSTM mreža

LSTM mreže imaju široku primjenu uključujući modeliranje jezika, prevođenje, prepoznavanje rukopisa i govora zbog svoje sposobnosti dugotrajnog pamćenja. Također su dobre u detekciji anomalija tako što uče što je uobičajeno ponašanje sustava te u prognozi vremenskih serija zbog čega se koriste i u ovome istraživanju kao dobar izbor predikcije kretanja cijene kroz vrijeme.

4. Tehnički indikatori

Tehnički indikatori su svi alati koji se koriste za analizu povijesti kretanja cijene i volumena trgovanja s ciljem predviđanja budućih kretanja cijene. Ovi alati su ključni alati tehničkih analitičara koji se oslanjaju na povijesne podatke kako bi donijeli informirane odluke o trgovanju. Tehnički indikatori koriste matematičke formule za obradu povijesnih podataka o cijeni i volumenu te najčešće generiraju signale za kupnju ili prodaju koji trgovcima pomažu u smanjenju rizika trgovanja i u prepoznavanju prilika na tržištu. Ovi indikatori mogu ukazivati na smjer trenda, snagu trenda, volatilnost tržišta i prekupljenost, odnosno potkupljenost promatranog trgujućeg instrumenta. U ovom poglavlju izneseni su osnovni principi i primjena 10 ovakvih indikatora kako bi bilo jasno zašto su odabrani kao ulaz modelima strojnog učenja.

4.1. Eksponencijalni pomični prosjek (EMA)

Eksponencijalni pomični prosjek (engl. *Exponential Moving Average*, EMA) je vrsta pomičnog prosjeka (MA) koja pridodaje veću težinu i značaj novijim cijenama. EMA se još naziva i eksponencijalno vagani pomični prosjek. U ovom istraživanju koriste se EMA(9) i EMA(25) što označava da se radi o EMA koji uzimaju 9, odnosno 25 perioda za svoj izračun. Uvijek kada se koriste dva MA, na jedan se referira kao brzi MA, a na drugi kao spori MA. Odabrani periodi su 9 i 25 jer su jedni od najčešće korištenih što im povećava značaj (puno ljudi reagira na isti prikaz na svojim ekranima) te su oni koji su postavljeni u Binance sučelju za trgovanje kada se otvori neki valutni par, a kao što je već navedeno, Binance je najznačajnija burza. EMA se računa navedenom formulom, a prikazan je crvenom linijom na slici 4.1.

$$EMA = (C * \frac{k}{1+N}) + (EMA_{prethodni} * (1 - \frac{k}{1+N})) \tag{4.1}$$

Slika 4.1: Eksponencijalni pomični prosjek

gdje je:

$C =$ Cijena zatvaranja trenutnog perioda

$k =$ Faktor zaglađenja

$N =$ Broj perioda

4.2. MACD indikator

Indikator MACD (engl. *Moving Average Convergence/Divergence*, MACD) indikator je koji investitorima pomaže u prepoznavanju trenda cijene te zamaha tog trenda. Prikazuje odnos brzog i sporog EMA linijama te često histogramom. Izmislio ga je Gerald Appel 1970-ih [11], a dan danas je jedan od najkorištenijih indikatora. Vrijednost MACD-a dobiva se sljedećom formulom:

$$MACD_{linija} = EMA_9 - EMA_{25} \tag{4.2}$$

$$Signal_{linija} = EMA_7(MACD_{linija}) \tag{4.3}$$

Inače se koriste EMA(12), EMA(26) te EMA 9 perioda MACD linije no ovdje se koriste 9, 25 i 7 zbog razloga navedenih u poglavlju 4.1. te veće responzivnosti indikatora kako se predviđanje radi nad petnaestominutnim svijećama što je relativno brz pregled kretanja cijene. Izgled MACD indikatora dan je slikom 4.2.

Slika 4.2: MACD indikator

4.3. Bollingerova ovojnica

Bollingerova ovojnica (engl. *Bollinger Bands*, BB), popularan je alat među investitorima i trgovcima, pomaže procijeniti volatilnost dionica i drugih trgujućih instrumenata kako bi se utvrdilo jesu li precijenjeni ili podcijenjeni. Razvio ju je 1980-ih godina financijski analitičar John Bollinger, a 2002. objavio knjigu *Bollinger on Bollinger Bands* gdje objašnjava kako ju koristiti [12]. Sastoji se od trake koju tvore tri linije koje se kreću s cijenom. Središnja linija je 20-periodni jednostavni pomični prosjek (engl. *Simple Moving Average*, SMA) cijene dok su gornja i donja linija postavljene na određeni broj standardnih devijacija, obično dvije, iznad i ispod srednje linije. Izračun Bollingerove ovojnice radi se na sljedeći način, a indikator je prikazan slikom 4.3.

$$\text{Gornja linija} = \text{SMA}_{20} + (\text{faktor} * \sigma) \tag{4.4}$$

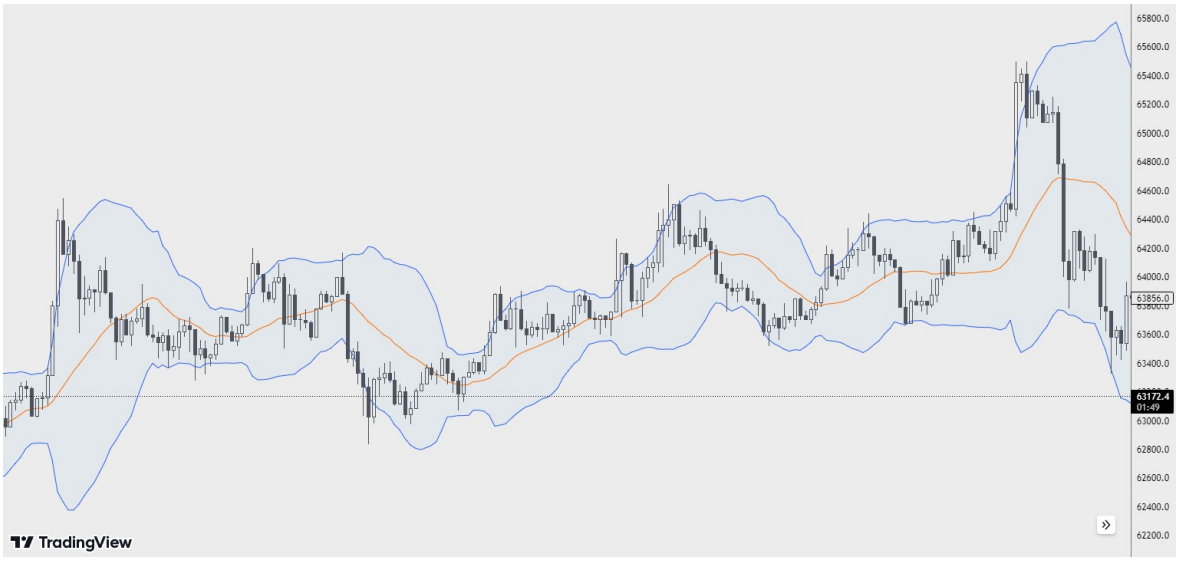
$$\text{Srednja linija} = \text{SMA}_{20} \tag{4.5}$$

$$\text{Donja linija} = \text{SMA}_{20} - (\text{faktor} * \sigma) \tag{4.6}$$

gdje je:

$$faktor = \text{Broj standardnih devijacija (najčešće 2)}$$

$$\sigma = \text{Standardna devijacija cijene kroz 20 perioda}$$



Slika 4.3: Bollingerova ovojnica

4.4. Indeks relativne snage (RSI)

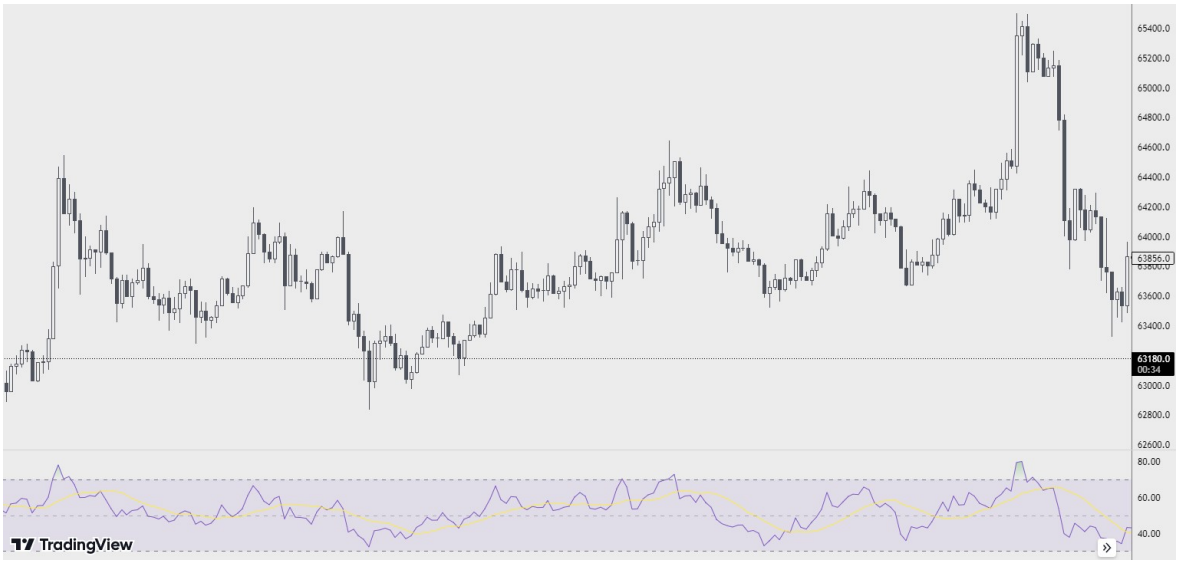
Indeks relativne snage (engl. *Relative Strength Index*, RSI) je indikator zamaha koji se koristi u tehničkoj analizi. RSI mjeri brzinu i veličinu nedavne promjene cijene kako bi se uočila precijenjenost ili podcijenjenost promatranog instrumenta. Izmislio ga je i predstavio J. Welles Wilder Jr. u svojoj knjizi iz 1978. *New Concepts in Technical Trading Systems* [13].

RSI je prikazan kao oscilator koji u odnos stavlja veličinu nedavnog rasta cijene i nedavne gubitke pri padu i taj odnos prikazuje vrijednostima na skali od 0 do 100 te se tradicionalno smatra da očitanje iznad 70 ukazuje na prekupljenost, a očitanje ispod 30 potkupljenost. Najčešće se koriste podaci o cijeni u roku od 14 perioda.

Nakon 14 perioda, drugim korakom RSI se zaglađuje kako bi prelazio 70 i 30 samo u snažnim trendovima i kako ne bi krivo reagirao pri naglim pokretima cijene. Izračun RSI-a dan je sljedećim jednadžbama, a prikaz indikatora slikom 4.4.

$$RSI_{\text{Prvi korak}} = 100 - \left[\frac{100}{1 + \frac{\text{Avg. uspon}}{\text{Avg. pad}}}\right] \tag{4.7}$$

$$RSI_{\text{Drugi korak}} = 100 - \left[\frac{100}{1 + \frac{(\text{Prethodni Avg. uspon} * 13) + \text{Trenutni rast}}{(\text{Prethodni Avg. pad} * 13) + \text{Trenutni pad}}}\right] \tag{4.8}$$



Slika 4.4: Indeks relativne snage

4.5. Prosječni stvarni raspon (ATR)

Prosječni stvarni raspon (engl. *Average True Range*, ATR) je indikator za tehničku analizu kojeg je osmislio J. Welles Wilder Jr. te predstavio u svojoj knjizi *New Concepts in Technical Trading Systems*, a mjeri volatilnost kretanja cijene [13].

Prvo se računa stvarni raspon (TR) koji se dobiva kao maksimum sljedećih vrijednosti: razlika najviše i najniže vrijednosti trenutne svijeće, apsolutne vrijednosti razlike najviše vrijednosti trenutne svijeće i cijene pri zatvaranju prošle svijeće i apsolutne vrijednosti razlike najniže vrijednosti trenutne svijeće i cijene pri zatvaranju prošle svijeće. ATR je zatim pomični prosjek, najčešće 14 perioda, stvarnog raspona. Proračun stvarnog raspona i ATR dani su sljedećim formulama, a prikaz indikatora može se vidjeti na slici 4.5.

$$TR = Max[(H - L), |H - C_p|, |L - C_p|] \tag{4.9}$$

gdje je:

$$H = \text{Najviša vrijednost trenutne svijeće}$$
$$L = \text{Najniža vrijednost trenutne svijeće}$$
$$C_p = \text{Cijena pri zatvaranju prošle svijeće}$$

Zatim:

$$ATR_{prvi} = \left(\frac{1}{n}\right) \sum_{i}^{n} TR_i \tag{4.10}$$

$$ATR = \frac{ATR_{prethodni} * (n-1) + TR}{n} \tag{4.11}$$

gdje je:

$$n = \text{Broj perioda (najčešće 12)}$$

Slika 4.5: Prosječni stvarni raspon

4.6. Prosječni indeks smjera (ADX)

Prosječni indeks smjera (engl. *Average Directional Index*, ADX) indikator je snage trenda. Izračun se temelji na pomičnom prosjeku raspona cijene kroz period vremena. ADX mjeri snagu trenda neovisno o smjeru kretanja cijene, a vrijednosti mu sežu od 0 do 100.

Obično se prikazuje zajedno s dva Indikatora smjera kretanja (engl. *Directional Movement Indicator*, DMI) +DI i -DI. Kada je +DI iznad -DI, cijena se kreće prema gore, a kada je -DI iznad +DI cijena se kreće prema dolje, ali ADX se svejedno može kretati prema gore u oba slučaja. Vrijednost ADX-a između 0 i 25 smatra se slabim trendom, dok vrijednosti između 75 i 100 vrlo snažnim trendom. ADX se dobiva sljedećim izračunom, a prikazan je slikom 4.6. gdje je narančasta linija -DI, plava linija +DI i crvena linija ADX.

$$+DI = \left(\frac{Zagladeni\ +DM}{ATR}\right) * 100 \tag{4.12}$$

$$-DI = \left(\frac{Zagladeni\ -DM}{ATR}\right) * 100 \tag{4.13}$$

$$DX = \left(\frac{|+DI - -DI|}{|+DI + -DI|}\right) * 100 \tag{4.14}$$

$$ADX = \frac{ADX_{prethodni} * 13 + DX}{14} \tag{4.15}$$

gdje je:

$$ADX_1 = \frac{\sum_1^{14} DX}{14}$$

$+ DM =$ Razlika najviše vrijednosti ove i prošle svijeće

$- DM =$ Razlika najniže vrijednosti ove i prošle svijeće

$$\text{Zaglađeni (+/-) DM} = \sum_1^{14} DM - \left(\frac{\sum_1^{14} DM}{14}\right) + DM$$

Slika 4.6: Prosječni indeks smjera

4.7. Stohastički oscilator

Stohastički oscilator je indikator zamaha na način da cijenu zatvaranja uspoređuje sa skupom cijena tog instrumenta kroz određeni period vremena. Poput RSI-a, funkcija mu je da ukaže na potkupljenost i prekupljenost promatranog instrumenta te mu vrijednosti variraju između 0 i 100. Očitanja iznad 80 ukazuju na prekupljenost, a ispod 20 potkupljenost.

Dakako, vrijednost oscilatora može ostati u ekstremima, a da pritom ne dođe do preokreta cijene, zato ovaj indikator, kao i sve ostale, treba upariti s drugima. Izmislio ga je 1950-ih George Lane [14] te se za izračun tipično koristi 14 perioda. Označava se s %K i obično dolazi u kombinaciji s 3-periodnim pomičnim prosjekom %K koji se označava s %D a naziva spori stohastički oscilator.

Formula je dana u nastavku izrazima (4.16) i (4.17), a izgled indikatora slikom 4.7. Plava linija je brzi stohastički oscilator (%K), a narančasta spori stohastički oscilator (%D).

$$\%K = \left(\frac{C - L_{14}}{H_{14} - L_{14}} \right) * 100 \tag{4.16}$$

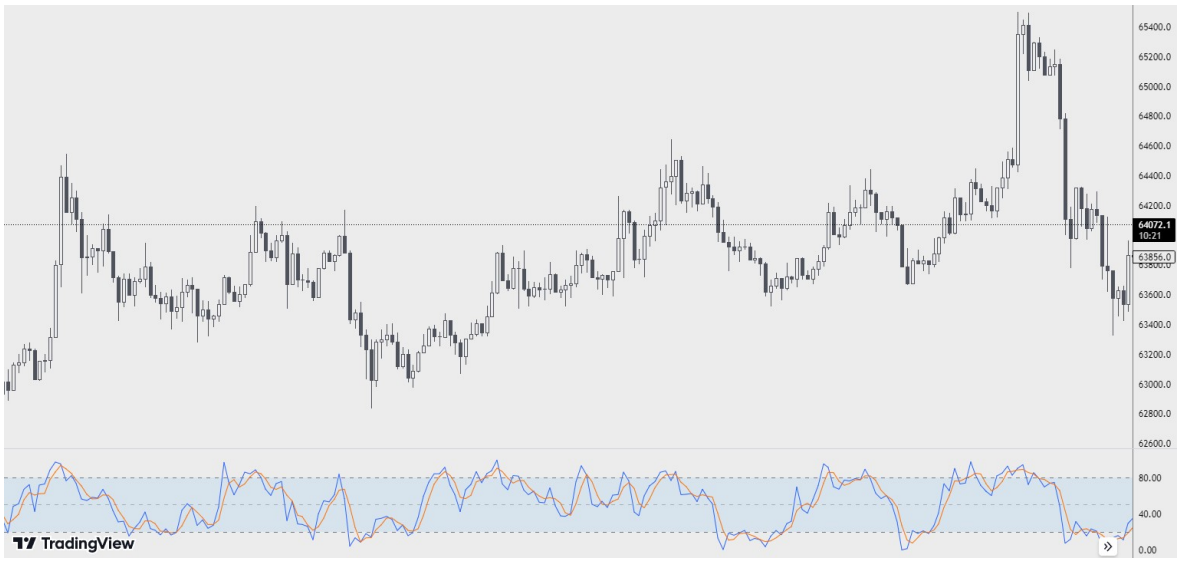
$$\%D = SMA_3(\%K) \tag{4.17}$$

gdje je:

$C =$ Cijena zatvaranja

$L_{14} =$ Najniža cijena u 14 perioda

$H_{14} =$ Najviša cijena u 14 perioda



Slika 4.7: Stohastički oscilator

4.8. Ravnotežni volumen (OBV)

Ravnotežni volumen (engl. *On-Balance Volume*, OBV) je indikator zamaha koji koristi tok volumena kako bi predvidio promjenu cijene promatranog instrumenta. Izmislio ga je Joseph Granville i predstavio u svojoj knjizi *Granville's New Key to Stock Market Profits* 1963. godine. Granville je smatrao da opaženi skok u volumenu bez značajne promjene cijene znači da će cijena s vremenom također napraviti veliki pomak [15].

Sama vrijednost OBV-a nije zanimljiva kako je indikator kumulativan i ovisi o početnoj točki, već je zanimljivo kretanje krivulje OBV-a te njen nagib [15]. Formula za izračun OBV-a dana je u nastavku, a izgled indikatora slikom 4.8.

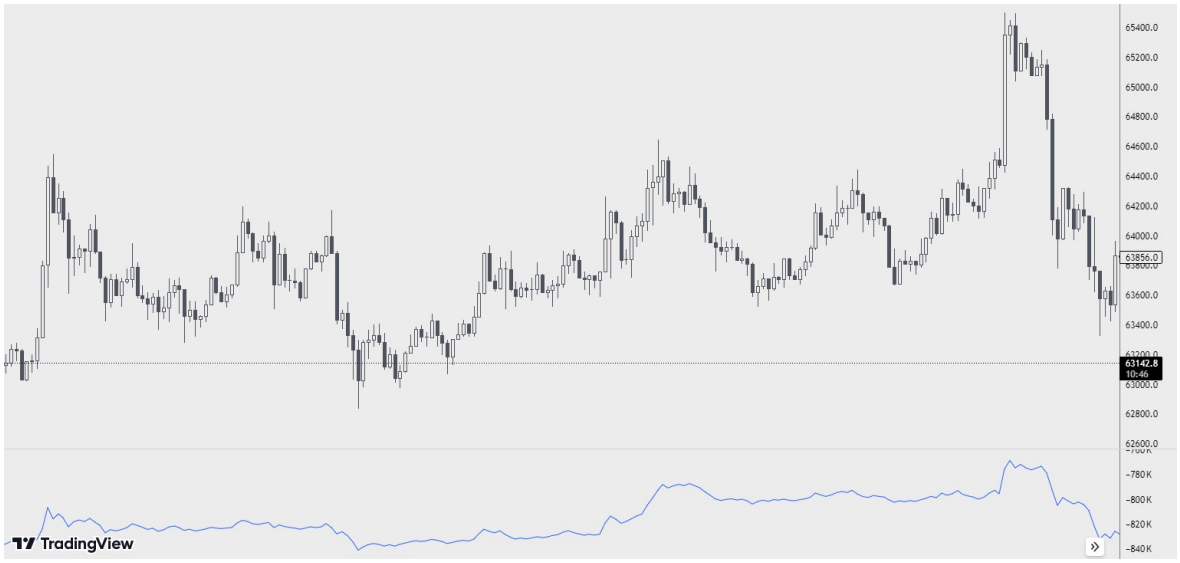
$$OBV = OBV_{prethodni} + \begin{cases} \text{volumen}, & \text{ako } C > C_p \\ 0, & \text{ako } C = C_p \\ \text{-volumen}, & \text{ako } C < C_p \end{cases} \tag{4.18}$$

gdje je:

$C =$ Cijena zatvaranja

$C_p =$ Cijena zatvaranja prošlog perioda

$OBV_1 =$ Volumen tog perioda



Slika 4.8: Ravnotežni volumen

4.9. Akumulacija/Distribucija (A/D)

Indikator akumulacije/distribucije (engl. *Accumulation/Distribution Indicator*, A/D) kumulativan je indikator koji koristi volumen i cijenu da ustanovi radi li se o akumulaciji ili distribuciji.

Ideja je prikazati koliko snažno se kupuje ili prodaje promatrani instrument. To čini na način da prvo provjeri je li se cijena zatvorila u gornjoj ili donjoj polovici njenog raspona. To se zatim množi s volumenom; ako se cijena zatvori u gornjoj polovici s visokim volumenom - A/D će skočiti, alternativno, ako se cijena zatvori u gornjoj polovici s niskim volumenom A/D se neće puno pomaknuti. Isti koncept vrijedi i za donju polovicu s razlikom da A/D pada umjesto raste.

Slično kao kod RSI, dođe li do divergencije kretanja cijene i kretanja A/D linije, npr. cijena raste dok A/D pada, to može signalizirati promjenu trenda. Za izračun A/D linije prvo treba izračunati volumen novčanog toka (engl. *Money Flow Volume*, MFV). Te su formule dane u nastavku. Prikaz indikatora dan je slikom 4.9.

$$MFM = \frac{(C - L) - (H - C)}{H - L} \tag{4.19}$$

$$MFV = MFM * Volumen \tag{4.20}$$

$$A/D_1 = MFV \tag{4.21}$$

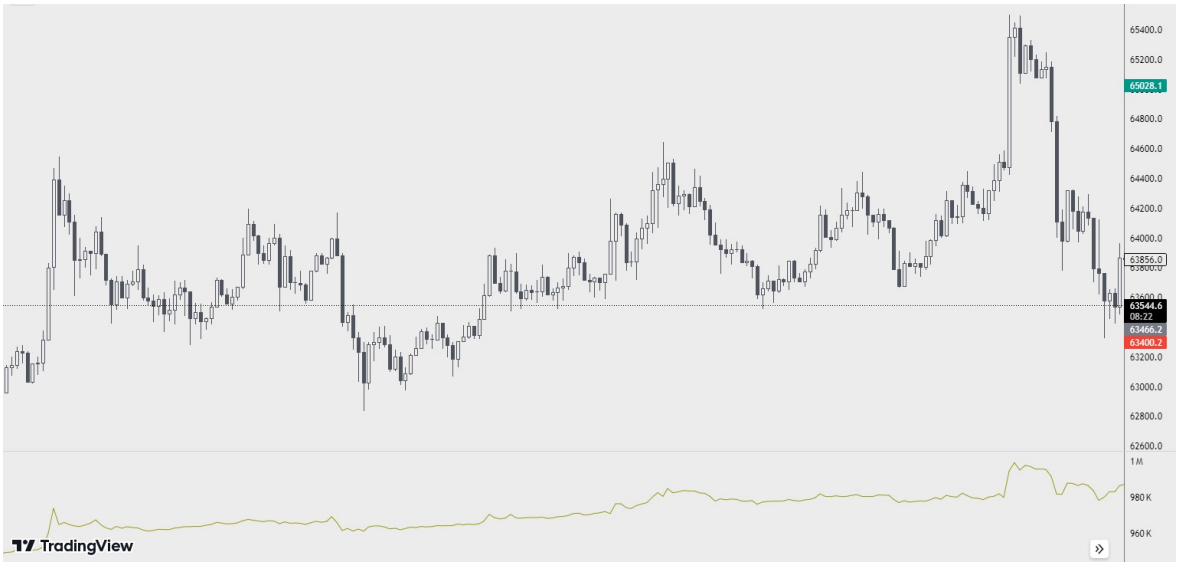
$$A/D = A/D_{prethodni} * MFV \tag{4.22}$$

gdje je:

C = Cijena zatvaranja trenutnog perioda

H = Najviša cijena trenutnog perioda

L = Najniža cijena trenutnog perioda



Slika 4.9: Akumulacija/Distribucija (A/D)

4.10. Aroon oscilator

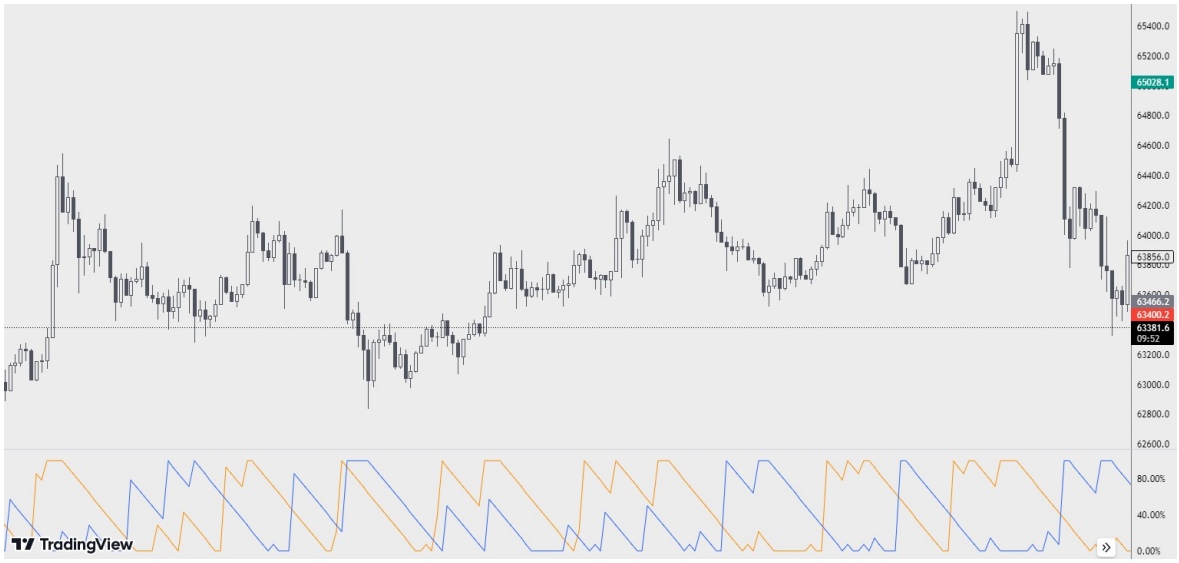
Aroon oscilator je indikator za praćenje trenda kretanja cijene promatranog trgujećeg instrumenta. Izmislio ga je Tushar Chande 1995. kao dio Aroon sistema [16]. Koristi Aroon Gore i Aroon Dolje da vizualizira snagu trenutnog trenda i vjerojatnost da se isti nastavi. Aroon Gore i Dolje mjere broj perioda od zadnje najviše i najniže vrijednosti u 25 perioda. Vrijednosti su između 0 i 100. Očitanje oscilatora iznad nule označava da je prisutan pozitivan trend, dok očitanje ispod nule da je trend negativan.

Trgovci obično čekaju prelazak preko nule kao signal promjene trenda te vrijednosti veće od 50 ili manje od -50 kao signal snažnog pomaka cijene. Izračuni Aroon Gore, Aroon Dolje i Aroon oscilatora prikazani su formulama u nastavku. Izgled Aroon oscilatora može se vidjeti na slici 4.10. Narančasta linija predstavlja Aroon Gore, a plava linija Aroon Dolje.

$$Aroon\ Oscilator = \text{Aroon Gore - Aroon Dolje} \tag{4.23}$$

$$Aroon\ Gore = 100 * \frac{25 - N_H}{25} \tag{4.24}$$

$$Aroon\ Dolje = 100 * \frac{25 - N_L}{25} \tag{4.25}$$



Slika 4.10: Aroon oscilator

5. Skup podataka

Ovo poglavlje usmjereno je na prikupljanje i pripremu podataka za daljnju analizu i modeliranje. Proces obuhvaća prikupljanje podataka s tržišta kriptovaluta te njihovo organiziranje za daljnju analizu i razvoj modela.

5.1. Python

Python je programski jezik visoke razine poznat i često korišten zbog svoje jednostavnosti i čitljivosti. Stvorio ga je Guido van Rossum 1980-ih, a prva verzija puštena u javnost je verzija 0.9.0. 1991. godine [17]. Danas je to drugi najzastupljeniji skriptni jezik u svijetu [18]. Filozofija iza dizajna Pythona je maksimizirati čitljivost koda te omogućiti što više u što manje linija koda.

Python je dinamički tipiziran, što znači da tip variable ne mora biti eksplicitno definiran i da se može mijenjati tijekom izvođenja koda što značajno olakšava i ubrzava razvoj koda. Također, Python izvršava linije koda jednu po jednu što olakšava pronalaženje grešaka, ali ga čini i jednim od sporijih programskih jezika. Diskutabilno najveća prednost Pythona je njegova široka standardna biblioteka te iznimno širok skup kvalitetnih vanjskih biblioteka koje programerima olakšavaju razvoj gotovo bilo čega pa tako i strojnog učenja. Python je postao *de facto* jezik znanosti o podatcima i strojnog učenja. Biblioteke kao što su NumPy, pandas i Matplotlib omogućuju manipulaciju podataka, analizu i vizualizaciju bez napora. Osim toga, scikit-learn, TensorFlow i PyTorch nude sve potrebno za razvoj modela strojnog učenja pa čak paralelizaciju i iskorištavanje grafičkog procesora za brže razvijanje modela pa su tako navedene biblioteke korištene i u ovom radu.

5.2. Dohvat i organizacija

Podatci su dohvaćeni koristeći se vanjskom bibliotekom CryptoCurrency eXchange Trading Library (CCXT). CCXT je biblioteka otvorenog koda za povezivanje s raznim kriptovalutnim burzama namijenjena olakšavanju dohvata povijesnih podataka, analize podataka, vizualizacije, algoritmičkog trgovanja i dr.

Dohvaćeni su podatci o 15-minutnim svijećama (engl. *candles*) te volumen svake svijeće, dakle cijena otvaranja, najviša cijena u tih 15 minuta, najniža cijena u tih 15 minuta, cijena pri zatvaranju svijeće te količina kriptovalute koja je razmijenjena u tih 15 minuta (engl. *Open (O), High (H), Low (L), Close (C), Volume*, OHLCV). Prikupljeni podatci su iz perioda od malo više od četiri godine od 26. prosinca 2019. 19:15:00 do 14. travnja 2024. 18:00:00 (GMT+2).

Nakon dohvata podataka u csv formatu, podatci su uvezeni u Microsoft Excel te su tamo formulama iskazanim u poglavlju 4. izračunati tehnički indikatori EMA, MACD, Bollingerova ovojnica, RSI, ATR, ADX, Stohastički oscilator, OBV, A/D, Aroon oscilator te svi međukoraci potrebni za izračun istih.

5.3. Obrada

Prije izgradnje modela strojnog učenja i provođenja istraživanja potrebno je prikupljene podatke obraditi, npr. obrisati NaN vrijednosti i slično. Prvi korak motiviran je istraživanjem Patel et al. koji su se bavili sličnom tematikom, ali s podatcima indijskih indeksa. Oni su pronašli da je prethodna obrada kontinuiranih vrijednosti indikatora na način da se svakom indikatoru pridodijeli vrijednost -1 ili 1 pridonijela značajno boljim rezultatima predviđanja [19]. U ovom radu je uzeta ova ideja te dodatno proširena vrijednošću 0 koja označava neutralnost (indikator ne daje signal da se kriptovaluta kupi niti proda).

Vrijednosti -1, 0 i 1 (prodaj, učini ništa, kupi) dodijeljene su svakom indikatoru na sljedeći način[1]:

$$EMA(9) = \begin{cases} -1, & \text{ako } C < EMA(9) \\ 1, & \text{ako } C > EMA(9) \end{cases} \tag{5.1}$$

$$EMA(25) = \begin{cases} -1, & \text{ako } C < EMA(25) \\ 1, & \text{ako } C > EMA(25) \end{cases} \tag{5.2}$$

$$MACD = \begin{cases} -1, & \text{ako } Histogram_n < 0 \text{ i } Historgram_{n-1} > 0 \\ 1, & \text{ako } Histogram_n > 0 \text{ i } Historgram_{n-1} < 0 \\ 0, & \text{inače} \end{cases} \tag{5.3}$$

[1]Način očitavanja signala svakog indikatora je određen po definiciji indikatora ili po onome što se tradicionalno smatra ispravnim te drugi sudionici na burzi najčešće koriste.

$$BB = \begin{cases} -1, & \text{ako } C \geq SMA \text{ i } C \leq BB_{gornja} \\ 1, & \text{ako } C < SMA \text{ i } C > BB_{donja} \\ 0, & \text{inače} \end{cases} \tag{5.4}$$

$$RSI = \begin{cases} -1, & \text{ako } RSI_{n-1} > 70 \text{ i } RSI_n < 70 \\ & \text{ili } \ RSI_{n-1} < 70 \text{ i } RSI_n < 70 \text{ i } RSI_n < RSI_{n-1} \\ 1, & \text{ako } RSI_{n-1} < 30 \text{ i } RSI_n > 30 \\ & \text{ili } \ RSI_{n-1} > 30 \text{ i } RSI_n > 30 \text{ i } RSI_n > RSI_{n-1} \\ 0, & \text{inače} \end{cases} \tag{5.5}$$

$$ATR = \begin{cases} -1, & \text{ako } C_n < (L_{n-1} - ATR_n) \\ 1, & \text{ako } C_n > (H_{n-1} + ATR_n) \\ 0, & \text{inače} \end{cases} \tag{5.6}$$

$$ADX = \begin{cases} -1, & \text{ako } ADX > 25 \text{ i } +DI < -DI \\ 1, & \text{ako } ADX > 25 \text{ i } +DI > -DI \\ 0, & \text{inače} \end{cases} \tag{5.7}$$

$$\text{Stohastički Osc.} = \begin{cases} -1, & \text{ako } \%D \text{ prelazi preko } \%K \text{ i } \%K > 80 \\ 1, & \text{ako } \%K \text{ prelazi preko } \%D \text{ i } \%D < 20 \\ 0, & \text{inače} \end{cases} \tag{5.8}$$

$$OBV_n = \begin{cases} -1, & \text{ako } OBV_n < \min(OBV_{n-1}...OBV_{n-21}) \\ & \text{i } \ OBV_{n-1} \geq \min(OBV_{n-2}...OBV_{n-22}) \\ 1, & \text{ako } OBV_n > \max(OBV_{n-1}...OBV_{n-21}) \\ & \text{i } \ OBV_{n-1} \leq \max(OBV_{n-2}...OBV_{n-22}) \\ 0, & \text{inače} \end{cases} \tag{5.9}$$

$$A/D = \begin{cases} -1, & \text{ako } A/D_n < A/D_{n-1} \text{ i } EMA(9)_n \leq EMA(9)_{n-1} \\ 1, & \text{ako } A/D_n > A/D_{n-1} \text{ i } EMA(9)_n \geq EMA(9)_{n-1} \\ 0, & \text{inače} \end{cases} \tag{5.10}$$

$$Aroon = \begin{cases} -1, & \text{ako } Aroon\ Down > 70 \\ 1, & \text{ako } Aroon\ Up > 70 \\ 0, & \text{inače} \end{cases} \tag{5.11}$$

Nakon toga, podatci su izvezeni u dvije csv datoteke, jednu s podatcima o cijeni i kontinuiranim vrijednostima indikatora te drugu s diskretnim vrijednostima indikatora dobivenih u prethodnom koraku. Zbog tog koraka i načina izračuna nekih od indikatora, potrebno je odrezati početnih 36 redova jer sadrže NaN vrijednosti.

Zatim se podatci učitavaju u Jupyter Notebook te se uz pomoć pandas biblioteke stvaraju dva podatkovna okvira (engl. *DataFrame*) iz te dvije datoteke i konkateniraju. Nakon toga dodan je novi stupac nazvan "%Change" koji mjeri promjenu u postotcima između trenutne i prošle cijene zatvaranja svijeće. Zatim je dodan stupac s točnim vrijednostima predviđanja nazvan "Prediction". On je popunjen na način da se kao vrijednost uzima predznak stupca %Change i zatim cijeli stupac pomakne za jedan red prema gore jer je vrijednost koja se želi predvidjeti vrijednost iduće svijeće, točnije hoće li iduća svijeća biti pozitivna (1) ili negativna (-1) promjena cijene. Naposljetku se odbacuju redovi okvira koji imaju NaN vrijednosti i oni redovi koji imaju vrijednost stupca Prediction 0 jer se radi o vrlo rijetkom trenutku gdje nije došlo do promjene u cijeni te bi to problem klasifikacije bespotrebno otežalo uvodeći treću klasu.

Drugi podatkovni okvir sastoji se od istih podataka, ali s kontinuiranim vrijednostima indikatora s tim da ovaj podatkovni okvir ima stupce za vrijednosti: gornje i donje Bollingerove ovojnice, %K i %D, te Aroon gore i Aroon dolje kako indikatori Bollingerova ovojnica, stohastički oscilator i Aroon oscilator imaju više od jedne vrijednosti.

Cjelokupan postupak te izgled podatkovnih okvira vrijedi i za parove ETH/USDT te LINK/USDT. Konačni izgled podatkovnih okvira (za par BTC/USDT) prikazan je tablicama u nastavku. Tablica 5.1. prikazuje OHLCV podatke s Unix vremenskom oznakom (engl. *timestamp*) i stupcima %Change i Prediction. Tablice 5.2. i 5.3. prikazuju kontinuirane vrijednosti indikatora dok tablica 5.4. diskretne vrijednosti indikatora nakon pretvorbe pojašnjene u potpoglavlju 5.3.

Tablica 5.1: Cjenovni podatci sa stupcima %Change i Prediction

Timestamp	Open	High	Low	Close	%Change	Prediction
1.577410e+12	7203.89	7214.07	7203.26	7212.00	0.001079	1.0
1.577410e+12	7212.35	7241.61	7210.00	7226.64	0.002030	1.0
1.577410e+12	7226.62	7232.61	7220.52	7228.15	0.000209	-1.0
1.577410e+12	7228.75	7230.97	7216.96	7225.68	-0.000342	-1.0
1.577410e+12	7225.11	7236.40	7220.50	7223.58	-0.000291	-1.0
...						
1.713280e+12	62692.41	62908.82	62459.68	62880.41	0.002999	-1.0
1.713280e+12	62880.41	63100.00	62397.52	62424.17	-0.007256	1.0
1.713280e+12	62424.18	62569.28	62020.01	62501.90	0.001245	-1.0
1.713280e+12	62501.89	62570.71	62082.46	62092.13	-0.006556	1.0
1.713280e+12	62092.12	62415.00	61888.00	62308.34	0.003482	-1.0

Tablica 5.2: Kontinuirane vrijednosti indikatora - I. dio

Fast EMA	Slow EMA	MACD	Upper BB	Lower BB	RSI	ATR
7207.67	7240.07	9.70	7236.27	7172.69	40.79	26.02
7211.47	7239.04	10.90	7238.31	7172.12	46.71	26.42
7214.80	7238.20	11.30	7237.44	7172.56	47.30	25.40
7216.98	7237.24	10.83	7235.07	7173.84	46.40	24.58
7218.30	7236.19	9.90	7237.27	7174.01	45.61	23.96
...						
62866.85	62922.65	-17.23	63680.96	62099.92	45.94	475.29
62869.56	62919.40	-8.45	63638.94	62100.98	49.20	473.42
62780.48	62881.30	-44.58	63552.69	62080.33	42.50	489.78
62724.76	62852.12	-53.33	63483.53	62066.58	43.90	494.03
62598.24	62793.66	-91.05	63451.27	61983.83	38.56	493.62

Tablica 5.3: Kontinuirane vrijednosti indikatora - II. dio

ADX	%K	%D	OBV	Acc Dist	Aroon Up	Aroon Down
14.97	83.82	75.61	-5171.16	76.24	44.44	22.22
17.75	78.79	78.20	-4549.33	32.85	33.33	11.11
20.42	76.11	79.72	-4350.74	52.07	88.89	0.00
22.88	69.13	75.15	-4522.58	42.07	77.78	55.56
25.42	65.06	70.28	-4741.14	-133.89	66.67	44.44
...						
64.94	59.92	66.18	499391.69	1040.96	44.44	88.89
65.69	57.83	56.52	500126.66	641.99	33.33	88.89
66.14	22.37	48.32	499302.96	-761.21	22.22	77.78
66.96	34.35	38.07	500200.41	677.27	11.11	66.67
67.72	5.14	20.57	499619.29	-558.10	0.00	88.89

Tablica 5.4: Tablica s diskretnim vrijednostima indikatora

Fast EMA	**Slow EMA**	**MACD**	**Bollinger Bands**	**RSI**	**ATR**	**ADX**	**Stochastic Oscillator**	**OBV**	**Acc/Dist**	**Aroon**
1	-1	0	1	1	0	0	0	0	0	0
1	-1	0	1	1	0	0	0	1	0	0
1	-1	0	1	1	0	0	0	0	0	1
1	-1	0	1	0	0	0	0	0	1	1
1	-1	0	1	0	0	1	0	0	0	0
...										
-1	-1	-1	-1	1	0	-1	0	0	-1	-1
1	-1	0	1	1	0	1	0	0	0	-1
-1	-1	0	-1	0	0	1	0	0	0	-1
-1	-1	0	-1	1	0	-1	0	0	-1	0
-1	-1	0	-1	0	0	1	0	0	0	-1

6. Rezultati

Podatci obrađeni na način opisan u poglavlju 5. dalje su razdvojeni u skup za učenje i skup za testiranje u omjeru 80:20 koristeći slučajno sjeme 42 (engl. *random seed*). Zatim su trenirani modeli strojnog učenja: logistička regresija, SVM, slučajna šuma, XGBoost te LSTM te je zatim obavljen test pomoću skupa testnih podataka. Za svaki od modela obavljen je proces optimizacije hiperparametara s ciljem pronalaženja parametara za koji svaki individualni model najbolje klasificira.

Također, u ovom istraživanju je osim dovođenja nemodificiranih podataka na ulaze modela dodatno proveden postupak inženjeringa podataka kako bi se stvorili skupovi ulaznih podataka koji uključuju vremensku ovisnost između podataka. Razlog provođenja ovog postupka je što navedeni modeli, osim LSTM-a, nemaju sposobnost prepoznavanja vremenskih uzroka te se htjelo vidjeti hoće li se poboljšati rezultati predviđanja. Navedeno se postiglo strukturiranjem podataka tako da se uzmu u obzir podatci iz određenog vremenskog okvira. Na primjer, za svaki korak, uključeni su podatci iz nekoliko prethodnih vremenskih koraka s njihovim vrijednostima iz stupca s točnim vrijednostima predviđanja (stupac Prediction) kao dodatne značajke kako bi se modelima pružio kontekst prošlosti. Kako je LSTM rekurentna neuronska mreža specifično dizajnirana za rad sa sekvencijalnim podatcima poput vremenskih nizova, smatrano je da nije bilo potrebe za dodatnim inženjeringom podataka. Umjesto toga, za taj su model podatci izravno korišteni kao vremenski nizovi, bez spomenutog oblikovanja. Svi drugi modeli ispitani su i s ovim oblikom podataka.

Za evaluaciju modela su nakon testiranja nad testnim skupom, izrađeni klasifikacijski izvještaji koji se sastoje od preciznosti (engl. *precision*), odziva (engl. *recall*), F1-ocjene (engl. *F1-score*), podrške (engl. *support*) i makro i težinskog prosjeka navedenih parametara.

Preciznost je udio točno klasificiranih primjera u skupu pozitivno klasificiranih primjera. Izračun je prikazan izrazom (6.1).

$$\text{Preciznost} = \frac{TP}{TP + FP} \tag{6.1}$$

gdje je:

TP broj stvarno pozitivnih rezultata

FP broj lažno pozitivnih rezultata

Odziv je udio točno klasificiranih primjera u skupu svih pozitivnih primjera. Izračun je prikazan izrazom (6.2).

$$\text{Odziv} = \frac{TP}{TP + FN} \tag{6.2}$$

gdje je:

TP broj stvarno pozitivnih rezultata

FN broj lažno negativnih rezultata

F1-ocjena je harmonijska sredina preciznosti i odziva. Ova mjera omogućuje ocjenjivanje modela s jednim brojem koji uravnotežuje obje vrijednosti. Posebno je korisna kada postoji neuravnoteženost između klasa ili kada je jednako važna i preciznost i odziv. Izračun je dan izrazom (6.3).

$$\text{F1-ocjena} = 2 \cdot \frac{\text{Preciznost} \cdot \text{Odziv}}{\text{Preciznost} + \text{Odziv}} \tag{6.3}$$

Podrška je broj stvarnih pojavljivanja klase u skupu podataka. Jednostavno rečeno, to je broj instanci svake klase.

U sljedećim potpoglavljima navedene tablice odnose se na kriptovalutni par BTC/USDT radi sažetosti. Tablice klasifikacijskih izvještaja za parove ETH/USDT i LINK/USDT mogu se pronaći u privitku A.

6.1. Logistička regresija

U ovom istraživanju je za optimizaciju parametara modela logističke regresije korištena metoda mrežnog pretraživanja (engl. *grid search*). To je proces koji podrazumijeva isprobavanje svih kombinacija parametara koje je korisnik naveo s ciljem pronalaženja kombinacije koja postiže najbolje rezultate klasifikacije. Kao metrika performansi odabrana je točnost modela. Osim toga, broj maksimalnih iteracija postavljen je na 1000 kako bi se modelu dao dovoljan broj iteracija za konvergenciju.

U mrežu pretraživanja postavljeni su parametri "C" i "solver". Parametar C je inverz snage regularizacije. Kontrolira prenaučenost penalizirajući velike koeficijente. Manja vrijednost C označava snažniju regularizaciju. Na ovaj način pokušava se postići ravnoteža

složenosti modela i sposobnosti generalizacije. Drugi parametar je parametar "solve". Ovaj parametar određuje algoritam koji će se koristiti za optimizaciju ciljne funkcije. "liblinear" je dobar izbor za manje skupove podataka, "lbfgs" je optimizacijski algoritam efikasan s velikim skupovima podataka, a "saga" brzo obrađuje velike skupove i dobro obrađuje raspršene skupove podataka.

```
param_grid = {
    'C': [0.01, 0.1, 1, 10, 100],
    'solver': ['liblinear', 'lbfgs', 'saga']
}
```

Odabrana je petodijelna unakrsna provjera (engl. *5-fold cross-validation*) kao razumna provjera te zbog ograničenja računalnih resursa. Unakrsna provjera je postupak provjere performansi modela gdje se skup podataka razdvaja na n (u ovom slučaju 5) dijelova te se model trenira na $n - 1$ dijelova, a zadnji dio se koristi za validaciju. Ovaj postupak se ponavlja n puta te se naposljetku izračuna prosjek performansi svih iteracija.

Kao rezultati mrežne pretrage najbolji hiperparametri su odabrani C:10 i solver:liblinear i s kontinuiranim podatcima i s diskretnim vrijednostima indikatora te su rezultati prikazani tablicama 6.1. i 6.2. za skupove podataka s kontinuiranim vrijednostima indikatora i diskretnim vrijednostima indikatora treniranim nad tim parametrima.

Tablica 6.1: Klasifikacijski izvještaj za kontinuirane vrijednosti (Logistička regresija)

	Precision	Recall	F1-Score	Support
-1	0.54	0.54	0.54	14981
1	0.54	0.54	0.54	15201
Accuracy		0.5397		30182
Macro avg	0.54	0.54	0.54	30182
Weighted avg	0.54	0.54	0.54	30182

Tablica 6.2: Klasifikacijski izvještaj za diskretne vrijednosti (Logistička regresija)

	Precision	Recall	F1-Score	Support
-1	0.53	0.53	0.53	14981
1	0.54	0.55	0.54	15201
Accuracy		0.5371		30182
Macro avg	0.54	0.54	0.54	30182
Weighted avg	0.54	0.54	0.54	30182

Klasifikacijski izvještaji impliciraju da modeli daju vrlo slične performanse za kontinuirane i diskretne podatke. Kontinuirane vrijednosti indikatora rezultiraju neznačajno većom preciznosti te manjim brojem lažno pozitivnih rezultata (engl. *false positives*) dok diskretne vrijednosti s malo većim brojem stvarno pozitivnih rezultata (engl. *true positives*). Sve u svemu, rezultati s oba skupa podataka su osrednji, s točnosti, preciznosti i specifičnosti oko 54%. Ovo ukazuje na to da model ima samo marginalno bolju sposobnost predviđanja od nasumičnog pogađanja (što bi rezultiralo točnošću od oko 50% za uravnotežen skup podataka).

Skaliranje skupa podataka niti provođenje inženjeringa podataka spomenutog na početku ovog poglavlja nije pomoglo u stvaranju boljeg modela logističke regresije. Skaliranje podataka je naprotiv, rezultiralo većim brojem istinitih pozitiva.

Parovi ETH/USDT i LINK/USDT imali su nešto lošije rezultate, Chainlink značajnije od Ethereuma. Par ETH/USDT je za kontinuirane vrijednosti imao točnost od 52.71%, a za diskretne 53.76%. Par LINK/USDT imao je za kontinuirane vrijednosti točnost od 51.50%, a za diskretne 52.34%. Dakle, kod ovog modela kriptovaluta manje tržišne kapitalizacije pokazala se težom za predvidjeti. Klasifikacijski izvještaji mogu se vidjeti u tablicama A.1., A.2., A.10. i A.11.

6.2. SVM

Za optimizaciju modela stroja potpornih vektora također je korištena metoda mrežnog pretraživanja. Ispitane su kombinacije sljedećih parametara:

```
param_grid = {
    'C': [0.1, 1, 10, 100],
    'gamma': [1, 0.1, 0.01, 0.001, 'scale'],
    'kernel': ['linear', 'rbf']
}
```

C je parametar regularizacije kao što je objašnjeno i u prethodnom poglavlju. "kernel" je parametar koji određuje jezgrenu funkciju koja će se koristiti. U ovom istraživanju isprobane su linearna i radijalna bazna funkcija (RBF). Linearna jezgrena funkcija definirana je kao skalarni produkt dva vektora značajki x i y: $K(x, y) = x \cdot y$. Linearna jezgrena funkcija preslikava ulazni prostor izravno u prostor značajki bez dodatne transformacije, što je čini prikladnom za probleme gdje je odnos između značajki i ciljne varijable linearan. RBF preslikava ulazni prostor u prostor više dimenzije što ga čini prikladnim za probleme gdje je veza između značajki nelinearna. RBF jezgra definirana je kao $K(x, y) = \exp(-\gamma||x - y||^2)$ gdje je γ parametar koji određuje širinu Gaussove funkcije, veći γ će rezultirati kompleksnijom granicom razdiobe. Ovdje su isprobane vrijednosti γ 1, 0.1, 0.01, 0.001 te "scale" gdje je $scale = 1/(\text{br. značajki} * var(\text{značajki}))$.

Za metodu usporedbe različitih kombinacija parametara unutar mreže korištena je točnost modela te je korištena petodijelna unakrsna provjera. Mrežnim pretraživanjem pronađeno je da je najbolji set parametara C=10, γ='scale', kernel='rbf'. Tablicama 6.3. i 6.4. prikazani su klasifikacijski izvještaji, prvo za kontinuirane podatke, a zatim za diskretne.

Tablica 6.3: Klasifikacijski izvještaj za kontinuirane vrijednosti (SVM)

	Precision	Recall	F1-Score	Support
-1	0.52	0.11	0.18	14981
1	0.51	0.90	0.65	15201
Accuracy		0.51		30182
Macro avg	0.51	0.50	0.41	30182
Weighted avg	0.51	0.51	0.42	30182

Tablica 6.4: Klasifikacijski izvještaj za diskretne vrijednosti (SVM)

	Precision	Recall	F1-Score	Support
-1	0.51	0.16	0.25	14981
1	0.51	0.85	0.63	15201
Accuracy		0.51		30182
Macro avg	0.51	0.50	0.44	30182
Weighted avg	0.51	0.51	0.44	30182

Rezultati ovog modela najlošiji su od svih modela promatranih u ovom istraživanju. Postignuta je točnost od 51% te vrlo loš odziv za klasu -1 od 0.18. Model velikom većinom kao izlaz daje klasu 1 što je i razlog vrlo loše F1-ocjene za skup -1.

Dovođenjem setova podataka modificiranih podatkovnim inženjeringom opisanim na početku poglavlja nisu postignuti bolji rezultati. Dovođenjem podatkovnih okvira stvorenih od podataka za par ETH/USDT postižu se rezultati isti kao za par BTC/USDT. Za skup podataka para LINK/USDT postiže se točnost od 50%, a odziv je za diskretne podatke za klasu -1 svega 0.02. Tablice A.3., A.4., A.12. i A.13. prikazuju klasifikacijske izvještaje za parove ETH/USDT i LINK/USDT. Pokazalo se da manja tržišna kapitalizacija negativno utječe na sposobnost stroja potpornih vektora u predviđanju kretanja cijene kriptovalute. Također, treba spomenuti da se treniranje i predviđanje ovog modela provodilo značajno duže od ostalih modela.

6.3. Slučajna šuma

Za potrebe optimizacije hiperparametara slučajne šume koristila se Slučajna pretraga s unakrsnom validacijom (engl. *Randomized Search Cross-Validation*). Ovakva pretraga slična je mrežnom pretraživanju, ali je efikasnija, pogotovo kada se radi s velikim skupom podataka, pod cijenu da se ne isprobaju sve kombinacije hiperparametara već određen broj slučajno odabranih kombinacija. Set parametara koji su se pretraživali može se vidjeti u nastavku:

```
param_dist = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'max_features': ['sqrt', 'log2', None]
}
```

Najbolji set parametara koji je pronađen je:

```
param_best = {
    'n_estimators': 100,
    'max_depth': 10,
    'min_samples_split': 5,
    'min_samples_leaf': 4,
    'max_features': None
}
```

Drugim riječima, stvarana je šuma od 100 stabala s maksimalnom dubinom stabla od 10 čvorova. Minimalan broj uzoraka potreban za dijeljenje čvora je bio postavljen na 5 dok je

minimalan broj uzoraka u listovima bio 4. Ne postavljanje "max_features" parametra znači da se nije specificiralo koliki broj značajki bi se trebao uzeti u obzir pri izvršavanju dijeljenja.

Za taj set parametara je unakrsnom validacijom postignuta najveća točnost od 53.98%. U nastavku slijedi tablica 6.5. koja sadrži klasifikacijski izvještaj za model treniran s kontinuiranim vrijednostima značajki, a zatim tablica 6.6. za rezultate dobivene s diskretnim vrijednostima. Rezultati su gotovo pa identični za oba tipa podataka uz malo veću točnost modela treniranog na kontinuiranim vrijednostima, 53.98% naspram 53.46%.

Tablica 6.5: Klasifikacijski izvještaj za kontinuirane vrijednosti (Slučajna šuma)

	Precision	Recall	F1-Score	Support
-1	0.54	0.51	0.52	14981
1	0.54	0.57	0.55	15201
Accuracy		0.5398		30182
Macro avg	0.54	0.54	0.54	30182
Weighted avg	0.54	0.54	0.54	30182

Tablica 6.6: Klasifikacijski izvještaj za diskretne vrijednosti (Slučajna šuma)

	Precision	Recall	F1-Score	Support
-1	0.53	0.51	0.52	14981
1	0.54	0.56	0.55	15201
Accuracy		0.5346		30182
Macro avg	0.53	0.53	0.53	30182
Weighted avg	0.53	0.53	0.53	30182

Dodatan inženjering podataka postigao je još slabije rezultate. Isprobani su koraci u kojima je spojeno 100, 30, 4 te 2 svijeće (vremenskih koraka). Rezultati svih slučajeva postigli su točnost od oko 52%, a najbolji rezultat imala je sekvenca od 4 vremenska koraka i to za diskretne vrijednosti značajki s 52.42% točnosti. Rezultati za parove ETH/USDT i LINK/USDT su nešto slabiji s točnošću između 52.67% i 53.25%. Klasifikacijski izvještaji mogu se vidjeti u tablicama A.5., A.6., A.14. i A.15. Najveću točnost postigao je ETH/USDT za kontinuirane podatke te je vrlo blizu rezultata za BTC/USDT što je očekivano s obzirom na visoku korelaciju Bitcoina i Ethereuma od oko 85% [20].

Rezultati ukazuju na to da klasifikacija modelom slučajne šume, čak i s optimizacijom hiperparametara, postiže točnost koja je samo malo bolja od pogađanja.

6.4. XGBoost

Pri stvaranju XGBoost klasifikatora, mjera za evaluaciju modela postavljena je na "logloss", odnosno funkciju gubitka koja kvantificira cijenu koju plaćamo za manjak točnosti pri klasificiranju, a odabrana je jer postiže dobre rezultate u radu s binarnim problemima. Kako bi se maksimizirali rezultati XGBoost modela korištena je Slučajna pretraga s unakrsnom validacijom (petodijelna validacija), kao i u prethodnom potpoglavlju. Za optimizaciju ovog modela korištena je mreža od sljedećih parametara:

```
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [3, 4, 5, 6, 7],
    'learning_rate': [0.01, 0.1, 0.2, 0.3],
    'subsample': [0.6, 0.7, 0.8, 0.9, 1.0],
    'colsample_bytree': [0.6, 0.7, 0.8, 0.9, 1.0],
    'gamma': [0, 0.1, 0.2, 0.3],
    'reg_alpha': [0, 0.01, 0.1, 1],
    'reg_lambda': [0, 0.01, 0.1, 1]
}
```

Broj mogućih stabala u ansamblu postavljen je na 100, 200 i 300. Maksimalna dubina stabla mogla je biti 3, 4, 5, 6 ili 7. Parametar "learning_rate" je parametar stope učenja, posebno bitan u metodama koje se temelje na gradijentom spustu, kao što je XGBoost. Ovaj parametar određuje veličinu koraka koje će algoritam uzimati pri ažuriranju težina. Ima funkciju sprječavanja prenaučenosti tako što kontrolira konvergenciju modela prema globalnom minimumu funkcije gubitka, smanjujući šanse zaglavljivanja u lokalnom minimumu. Manja stopa učenja znači sporije učenje modela, ali detaljnije prepoznavanje finijih struktura u podatcima. Ovdje su metodi optimizacije ponuđene stope 0.01, 0.1, 0.2 te 0.3. Parametar "subsample" kontrolira koliki postotak uzoraka će se koristiti za treniranje svakog stabla dok parametar "colsample_bytree" koliki postotak značajki će se koristiti za treniranje svakog stabla. Parametri "req_alpha" i "req_lambda" su regularizacijski parametri L1 i L2.

Rezultati optimizacije hiperparametara se za ovaj model strojnog učenja razlikuju s obzirom na odabrane podatkovne okvire (diskretne (D) i kontinuirane (K)) te su stoga prikazani tablicom 6.7.

Tablica 6.7: Najbolji parametri za XGBoost modele

	Subsample	Reg_L	Reg_A	N_Est	Max_Depth	LR	γ	Colsample
BTC (D)	0.8	0.01	1	200	6	0.01	0.3	1.0
BTC (K)	1.0	0	0.01	200	3	0.1	0.1	0.9
ETH (D)	0.6	0	0.01	100	3	0.01	0	1.0
ETH (K)	0.8	0.01	1	200	6	0.01	0.3	1.0
LINK (D)	0.8	0.01	0.01	300	3	0.01	0.3	1.0
LINK (K)	0.6	0	1	200	6	0.01	0	0.6

Ovaj model strojnog učenja postiže najveću točnost od testiranih modela s 55% za kontinuirane vrijednosti, no i dalje s prevelikim brojem lažno pozitivnih klasifikacija. Broj lažnih pozitiva je gotovo jednak broju istinitih negativa što ukazuje na to da se model muči s točnim predviđanjem klase -1 te je zamijećeno da i ovaj model u većem broju klasificira ulaze u skupinu 1. Klasifikacijski izvještaji mogu se vidjeti tablicama 6.8., za kontinuirane vrijednosti i 6.9., za diskretne vrijednosti.

Tablica 6.8: Klasifikacijski izvještaj za kontinuirane vrijednosti (XGBoost)

	Precision	Recall	F1-Score	Support
-1	0.54	0.51	0.53	14968
1	0.55	0.58	0.56	15214
Accuracy		0.5452		30182
Macro avg	0.55	0.54	0.54	30182
Weighted avg	0.55	0.55	0.54	30182

Tablica 6.9: Klasifikacijski izvještaj za diskretne vrijednosti (XGBoost)

	Precision	Recall	F1-Score	Support
-1	0.54	0.53	0.53	14968
1	0.55	0.56	0.55	15214
Accuracy		0.5432		30182
Macro avg	0.54	0.54	0.54	30182
Weighted avg	0.54	0.54	0.54	30182

Postupkom inženjeringa podataka nisu postignuti bolji rezultati od ovoga. Isprobani su setovi od 100, 30, 4 i 2 vremenska koraka. Najveća točnost od 54.4% postignuta je s 30 koraka, a F1-ocjene su za sve korake otprilike iste.

Valutni parovi ETH/USDT i LINK/USDT postigli su lošije rezultate od BTC/USDT. ETH/USDT samo marginalno uz točnost od 0.5417 za kontinuirane vrijednosti te 0.5409 za diskretne, a LINK/USDT nešto lošije s 0.5260 i 0.5251 s F1-ocjenom od 0.51 za klasu -1. Klasifikacijski izvještaji za ove kriptovalutne parove mogu se vidjeti u tablicama A.7., A.8., A.16. i A.17. Ovi rezultati ukazuju na opadanje mogućnosti predviđanja cijene kriptovalute s opadanjem tržišne kapitalizacije.

6.5. LSTM

Korištena LSTM mreža sastoji se od jednog ulaznog sloja, dva skrivena LSTM sloja i izlaznog sloja. Prvi LSTM sloj ima 50 čvorova s omogućenim povratnim sekvencama (engl. *return sequences*), dok drugi sloj ima 50 čvorova i ne vraća sekvence. Nakon svakog LSTM sloja, dodan je sloj ispada (engl. *dropout layer*) s koeficijentom od 0.2 kako bi se smanjila vjerojatnost prenaučenosti. Izlazni sloj je gusti sloj (engl. *dense layer*) jednog čvora te završava sa sigmoidnom aktivacijskom funkcijom, pogodnom za binarnu klasifikaciju zbog dvije moguće vrijednosti aktivacijske funkcije.

Za treniranje modela korišten je Adam optimizator s početnom stopom učenja od 0.001. Adam je optimizacijski algoritam koji spaja prednosti druga dva poznata optimizatora AdaGrad i RMSProp. Adam dinamički prilagođava stopu učenja svakog parametra na temelju veličine trenutnih i prošlih gradijenata. Računajući individualne stope učenja za različite parametre ubrzava konvergenciju i povećava efikasnost treniranja. Također, Adam koristi koncept zamaha, akcelerirajući vektore gradijenata u relevantnom smjeru te još i ublažava oscilacije što je vrlo korisno pri radu s podatcima koji sadrže puno šuma kao što je kretanje cijene. Kao funkcija gubitka korišten je binarni unakrsni entropijski gubitak (engl. *binary cross-entropy*), koji je prikladan za binarne klasifikacijske zadatke. Optimalnost modela određena je na temelju točnosti.

Podaci su normalizirani s pomoću MinMaxScaler-a kako bi se osiguralo da svi ulazi imaju sličan raspon vrijednosti, čime se ubrzava konvergencija modela i poboljšava stabilnost treniranja. Nakon normalizacije, podaci su pripremljeni na način da se ulaznim podatcima dodaju vremenski okviri te su na taj način stvorene sekvence. Svaki ulazni uzorak sastoji se od n uzastopnih vremenskih koraka. Nakon toga, podaci su podijeljeni u skupove za treniranje i testiranje u omjeru 80:20. Treniranje je trajalo 20 epoha s veličinom serije od 64. Isprobane su sekvence od 100, 30, 4 te 2 uzastopna vremenska koraka.

Kako nije bilo nikakve značajne razlike u rezultatima diskretnih i kontinuiranih podataka, u nastavku su tablicom 6.10. prikazani rezultati razdvojeni po vremenskim koracima. Mijenjanje broja koraka nije na značajan način utjecalo na točnost predviđanja već se čini da se za 30 koraka postiže bolji odziv za klasu 1, međutim potrebno je daljnje istraživanje. Najveća postignuta točnost je 53.44% za 30 vremenskih koraka. Međutim, za taj broj koraka, odziv za klasu -1 je 0.46, a za klasu 1 0.60 što pokazuje da model većinom predviđa klasu 1. Treba još jednom spomenuti da se na ovaj način postiže prividno veća točnost jer je globalni smjer kretanja kriptovalute prema gore.

Za parove ETH/USDT i LINK/USDT korištenje kontinuiranih ili diskretnih vrijednosti također se nije na značajan način razlikovalo pa su tablicama A.9. i A.18. također uspoređeni samo različiti vremenski koraci. Rezultati su i s ovim modelom malo lošiji za kriptovalute manje tržišne kapitalizacije. Par ETH/USDT snažno prati kretanje BTC/USDT pa su rezultati slični i najbolja točnost iznosi nešto manje nego ona za par BTC/USDT - 53.43% za 100 vremenskih koraka. Također, za razliku od para BTC/USDT, par ETH/USDT u većem broju predviđa klasu -1. Par LINK/USDT postiže točnost u prosjeku od 52.07% s F1-ocjenom koja varira od 0.48 do 0.55.

Tablica 6.10: Klasifikacijski izvještaj za LSTM model pri različitim brojem vremenskih koraka

	Precision	Recall	F1-Score	Support
Eksperiment 1 (100 vremenskih koraka)				
-1	0.53	0.51	0.52	14963
1	0.53	0.55	0.54	15199
Accuracy		0.5321		30162
Macro avg	0.53	0.53	0.53	30162
Weighted avg	0.53	0.53	0.53	30162
Eksperiment 2 (30 vremenskih koraka)				
-1	0.53	0.46	0.50	14929
1	0.53	0.60	0.57	15247
Accuracy		0.5344		30176
Macro avg	0.53	0.53	0.53	30176
Weighted avg	0.53	0.53	0.53	30176
Eksperiment 3 (4 vremenska koraka)				
-1	0.53	0.53	0.53	14960
1	0.54	0.53	0.53	15222
Accuracy		0.5322		30182
Macro avg	0.53	0.53	0.53	30182
Weighted avg	0.53	0.53	0.53	30182
Eksperiment 4 (2 vremenska koraka)				
-1	0.53	0.51	0.52	14970
1	0.53	0.55	0.54	15212
Accuracy		0.5240		30182
Macro avg	0.53	0.53	0.53	30182
Weighted avg	0.53	0.53	0.53	30182

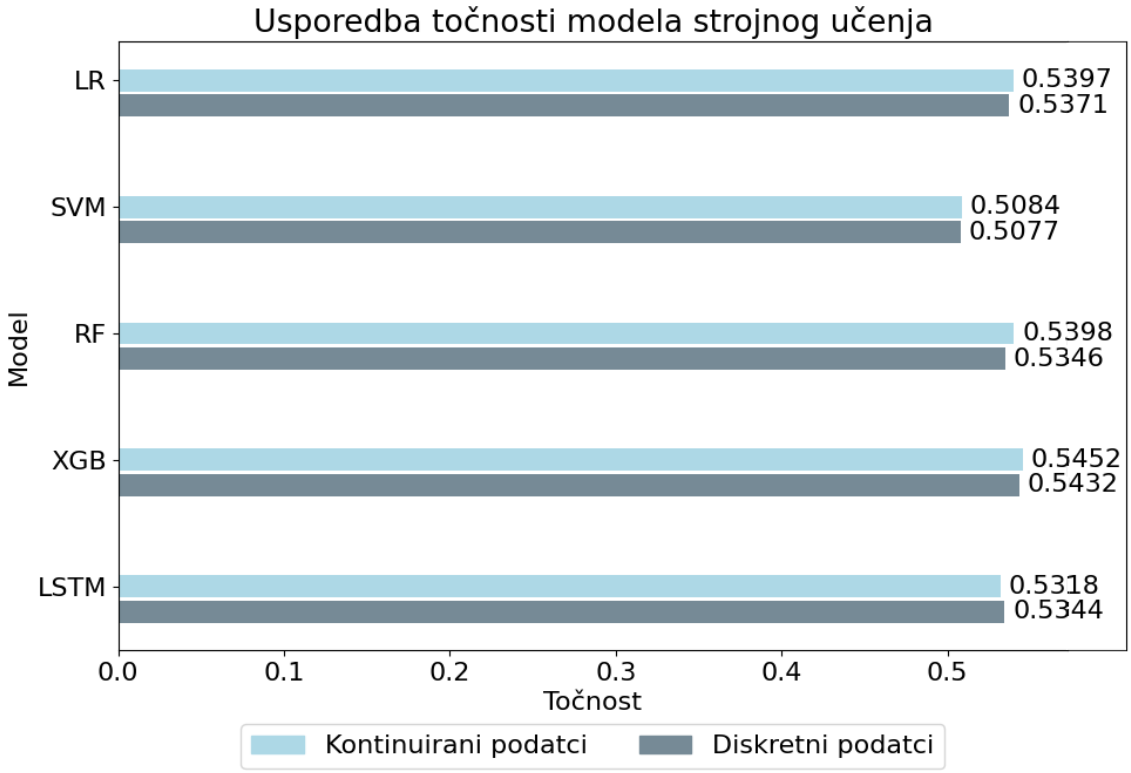
6.6. Diskusija

Korištenje modela logističke regresije (LR), SVM-a, slučajne šume (RF), XGBoosta (XGB) i LSTM-a nije se pokazalo kao valjana opcija za predviđanje kretanja cijene kriptovaluta. Svi modeli postigli su težinski prosjek F1-ocjene od oko 0.54 te su stoga na slici 6.1. u nastavku uspoređeni po postignutoj točnosti. Može se zamijetiti da se za diskretne vrijednosti indikatora za sve modele osim LSTM-a postižu malo slabiji rezultati nego za kontinuirane vrijednosti što je suprotno od pronalaska Patela et al. [19]. Svi modeli su postigli malo bolji odziv, a time i F1-ocjenu, za klasu 1. Ovo može značiti da je na temelju značajki dovedenih na ulaz lakše predvidjeti klasu 1 ili da je skup podataka neuravnotežen, međutim razlika u broju instanci klasa je oko 200 te je potrebno daljnje istraživanje kako bi se pronašao uzrok.

Modeli su po točnosti od najveće do najniže rangirani na sljedeći način: XGBoost, logistička regresija, slučajna šuma, LSTM, SVM. Najveću točnost, od 54.52%, postiže XGBoost za kontinuirane vrijednosti značajki. Treba istaknuti da je ovaj model postigao najveću točnost uz najbrže vrijeme treniranja od svih modela. Također treba istaknuti da model logističke regresije uz malo manju točnost postiže bolju ravnotežu između odziva klasa no što to čini XGBoost. Najlošiju točnost postiže SVM s točnosti od 50.84%, s velikom pristranosti klasi 1 i najdužim vremenom treniranja i testiranja.

Provedena je i usporedba rezultata predviđanja kriptovalutnog para visoke tržišne kapitalizacije (BTC/USDT) s parovima niže tržišne kapitalizacije (ETH/USDT, LINK/USDT). Svi modeli postigli su približno istu, ili malo nižu, točnost za par ETH/USDT što je objašnjeno visokom koreliranosti između ta dva para [20]. Također, svi modeli su za par LINK/USDT postigli nešto lošije rezultate što ukazuje na to da je predviđanje kretanja cijene kriptovaluta manje tržišne kapitalizacije teže no onih s većom.

Svi navedeni modeli postižu točnost malo veću od nasumičnog pogađanja smjera kretanja cijene te se iz tog razloga ne mogu koristiti za precizno predviđanje kretanja cijene kriptovaluta niti ostvarivanje profita na burzi. Rezultati ovog istraživanja potvrđuju teoriju slučajnog hoda u skladu s istraživanjima Jiaoa i Jakubowicza [5], Chitendera, Maredza i Sibanda [6] i drugih.



Slika 6.1: Usporedba točnosti modela strojnog učenja

U daljnjem istraživanju trebalo bi nastaviti raditi s modelom XGBoosta te pokušati još bolje optimizirati parametre modela. Također, treba pokušati trenirati modele s manjim i većim brojem indikatora te pronaći set indikatora koji postiže najbolje rezultate. Nadalje, treba pokušati s troklasnom klasifikacijom uvodeći klasu 0 koja označava neutralnost smjera kretanja. Na taj način postoji mogućnost da će se smjer moći predvidjeti rjeđe no točnije. Fokus bi se trebao prebaciti na predviđanje samo para BTC/USDT jer se predviđanje parova niže tržišne kapitalizacije pokazalo teže. Treba isprobati korištenje regresijskih modela koje se naposljetku može provući kroz signum funkciju i na taj način postići klasifikaciju okolnim putem uz moguću povećanu točnost. Također, trebalo bi se povećati vrijeme treniranja uz povećane računalne resurse te isprobati druge modele strojnog učenja kao što su mreže propusno povratnih ćelija (engl. *Gated Recurrent Unit*, GRU) i konvolucijske neuronske mreže (CNN).

7. Zaključak

U ovom istraživanju potaknutom problemom predviđanja kretanja burzovnih instrumenata i sve većom primjenom strojnog učenja u svakidašnjem životu, istraženi su različiti modeli strojnog učenja s ciljem predviđanja smjera kretanja cijene kriptovaluta. Problemu se pristupilo odabirom pet modela strojnog učenja koji su uspoređeni, a to su logistička regresija, SVM, slučajna šuma, XGBoost te LSTM. Podatci su prikupljeni za tri različita kriptovalutna para BTC/USDT, ETH/USDT i LINK/USDT kako bi se dodatno ustanovilo ima li tržišna kapitalizacija utjecaja na sposobnost predviđanja. Iz prikupljenih podataka izračunato je 11 različitih tehničkih indikatora kretanja cijene koji su se kasnije dovodili na ulaze modela strojnog učenja te je na taj način pokušano predvidjeti smjer kretanja cijene. Podatci su razdvojeni u dvije skupine, jedna s kontinuiranim vrijednostima indikatora te druga s diskretnim vrijednostima indikatora koja je ostvarena pretvorbom kontinuiranih vrijednosti u vrijednosti -1, 1 ili 0 koje simboliziraju da će se cijena kretati prema dolje, gore ili da se ne može odrediti. Za svaki od modela strojnog učenja proveden je postupak optimiziranja hiperparametara te je zatim na najboljim modelima obavljen postupak klasifikacije. Dodatno, za svaki od modela, osim LSTM-a, proveden je postupak podatkovnog inženjeringa spajanjem 2, 4, 30 i 100 vremenskih koraka koji sadrže točnu vrijednost predviđanja i time stvarajući vremenske sekvence kako bi se dao vremenski kontekst modelima koji ga inače sami nemaju. Modeli su ocijenjeni po preciznosti, odzivu i F1-ocjeni te kasnije rangirani i međusobno, zbog slične F1-ocjene, uspoređeni po točnosti. Najveću točnost, od 54.52% postigao je XGBoost za kontinuirane vrijednosti značajki te je najbrži od svih modela. Najlošije rezultate postigao je SVM uz točnost od 50.84% i velikim brojem lažno pozitivnih klasifikacija uz istovremeno najsporije vrijeme treniranja i testiranja. Za par ETH/USDT postignuta je približno ista, ili malo niža, točnost od one postignute za par BTC/USDT što je objašnjeno visokim postotkom koreliranosti s parom BTC/USDT. Za par LINK/USDT koji je još niže tržišne kapitalizacije postignuta je još niža točnost te je zaključeno kako manja tržišna kapitalizacija znači manja sposobnost predviđanja cijene kriptovalute. Svi navedeni modeli postigli su točnost malo veću od nasumičnog pogađanja smjera kretanja cijene te istraživanje u konačnici potvrđuje teoriju slučajnog hoda.

LITERATURA

[1] CoinMarketCap. Top cryptocurrency spot exchanges, 2024.

[2] Z. Zulkifli. Investigation into the fundamental analysis of stock investment instruments on the indonesian stock exchange. *Enigma in Economics*, 1(1):6–11, 2023.

[3] C. Boobalan. Technical analysis in select stocks of indian companies. *International Journal of Business and Administration Research Review*, 2(4):26–36, 2014.

[4] Burton G. Malkiel. *A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing*. W.W. Norton, United Kingdom, 2007.

[5] Yang Jiao and Jérémie Jakubowicz. Predicting stock movement direction with machine learning: An extensive study on sp 500 stocks. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4705–4713, 2017.

[6] T. T. Chitenderu, A. Maredza, and K. Sibanda. The random walk theory and stock prices: Evidence from johannesburg stock exchange. *International Business & Economics Research Journal (IBER)*, 13(6):1241–1250, 2014.

[7] Jan Šnajder and Bojana Dalbelo Bašić. *Strojno učenje*. Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, Zagreb, 3.7 edition, October 2014.

[8] Marko Čular. Modeli slučajnih šuma i primjene. Master's thesis / diplomski rad, University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, 2020.

[9] Tianqi Chen. Technical analysis in select stocks of indian companies. *International Journal of Business and Administration Research Review*, 2(4):26–36, 2014.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[11] CMT Association. Gerald appel, 2023.

[12] J. Bollinger. *Bollinger on Bollinger Bands*. McGraw-Hill Education, 2002.

[13] J. Welles Jr. Wilder. *New Concepts in Technical Trading Systems*. Trend Research, 1978. Page 6.

[14] Investopedia Staff. Stochastic oscillator. *Investopedia*. Pristupljeno: 2024-06-07.

[15] Adam Hayes. On-balance volume (obv): Definition, formula, and uses as indicator, 2024.

[16] Cory Mitchell. Aroon oscillator: Definition, calculation formula, trade signals, 2022.

[17] Guido Van Rossum. The history of python: A brief timeline of python. The History of Python, 2009. Pristupljeno: 2024-06-07.

[18] Stack Overflow. 2023 developer survey: Most popular technologies - programming, scripting, and markup languages, 2023. Pristupljeno: 2024-06-07.

[19] Jigar Patel, Sahil Shah, Priyank Thakkar, and Ketan Kotecha. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, 42:259–268, 01 2015.

[20] Macroaxis LLC. Correlation between bitcoin and ethereum, 2024. Pristupljeno: 2024-06-13.

SAŽETAK

Usporedba algoritama za predviđanje cijena kriptovaluta

Goran Brkić

Ovaj rad istražuje primjenu različitih modela strojnog učenja za predviđanje smjera kretanja cijene kriptovaluta. Analizirani su modeli logističke regresije, SVM-a, slučajne šume, XGBoosta i LSTM-a koristeći preciznost, odziv i F1-ocjenu kao mjeru uspješnosti. Analiza je provedena nad kriptovalutnim parovima BTC/USDT, ETH/USDT i LINK/USDT kako bi se dodatno saznalo je li predviđanje kriptovaluta niže tržišne kapitalizacije jednostavnije ili teže. Kroz pregled relevantne literature, tehničke analize cijene, implementaciju pet različitih modela strojnog učenja te analizu dobivenih rezultata pruža se pregled mogućnosti korištenja strojnog učenja za predviđanje kretanja kriptovaluta. Rezultati pokazuju da navedeni modeli pružaju ograničenu prednost u odnosu na nasumično predviđanje smjera te podržavaju teoriju slučajnog hoda. Također, rezultati pokazuju smanjenu sposobnost predviđanja kriptovalutnog para manje tržišne kapitalizacije. Istraživanje pruža korisne smjernice za implementaciju navedenih modela te naglašava potrebu za daljnjim optimiziranjem modela koji je postigao najveću točnost, XGBoost.

Ključne riječi: strojno učenje; kriptovalute; logistička regresija; SVM; slučajna šuma; XGBoost; LSTM; predviđanje cijene; burza; tehnički indikatori

SUMMARY

Comparison of cryptocurrency price prediction algorithms

Goran Brkić

This paper explores different machine learning models and their ability to predict the direction of price movement for different cryptocurrencies. Analysis was conducted on models of logistic regression, SVM, random forest, XGBoost and LSTM using precision, recall and F1-score as a method of comparison. Analysis was done on cryptocurrency pairs BTC/USDT, ETH/USDT and LINK/USDT to find out whether cryptocurrencies with lower market capitalization are harder to predict. Through a review of relevant literature, technical price analysis, the implementation of five different machine learning models, and the analysis of the obtained results, this paper provides an overview of the possibilities of using machine learning for predicting cryptocurrency movements. The results indicate that these models offer a limited advantage compared to random prediction and support the random walk theory. Additionally, the findings suggest a reduced ability to predict cryptocurrencies with lower market capitalizations. This research offers useful guidelines for implementing the mentioned models and emphasizes the need for further optimization of the model that achieved the highest accuracy, XGBoost.

Keywords: machine learning; cryptocurrency; logistic regression; SVM; random forest; XGBoost; LSTM; price prediction; exchange; technical indicators

Privitak A:
Klasifikacijski izvještaji za parove ETH/USDT i LINK/USDT

Tablica A.1: Klasifikacijski izvještaj za kontinuirane vrijednosti para ETH/USDT (Logistička regresija)

	Precision	Recall	F1-Score	Support
-1	0.54	0.52	0.53	15135
1	0.53	0.55	0.54	14988
Accuracy		0.5376		30123
Macro avg	0.54	0.54	0.54	30123
Weighted avg	0.54	0.54	0.54	30123

Tablica A.2: Klasifikacijski izvještaj za diskretne vrijednosti para ETH/USDT (Logistička regresija)

	Precision	Recall	F1-Score	Support
-1	0.53	0.52	0.52	15135
1	0.52	0.54	0.53	14988
Accuracy		0.5271		30123
Macro avg	0.53	0.53	0.53	30123
Weighted avg	0.53	0.53	0.53	30123

Tablica A.3: Klasifikacijski izvještaj za kontinuirane vrijednosti para ETH/USDT (SVM)

	Precision	Recall	F1-Score	Support
-1	0.49	0.28	0.36	14683
1	0.50	0.71	0.59	14898
Accuracy		0.50		29581
Macro avg	0.50	0.50	0.47	29581
Weighted avg	0.50	0.50	0.47	29581

Tablica A.4: Klasifikacijski izvještaj za diskretne vrijednosti para ETH/USDT (SVM)

	Precision	Recall	F1-Score	Support
-1	0.52	0.02	0.04	15135
1	0.50	0.98	0.66	14988
Accuracy		0.4984		30123
Macro avg	0.51	0.50	0.35	30123
Weighted avg	0.51	0.50	0.35	30123

Tablica A.5: Klasifikacijski izvještaj za kontinuirane vrijednosti para ETH/USDT (Slučajna šuma)

	Precision	Recall	F1-Score	Support
-1	0.5360	0.5174	0.5265	15135
1	0.5291	0.5476	0.5382	14988
Accuracy		0.5325		30123
Macro avg	0.5326	0.5325	0.5324	30123
Weighted avg	0.5326	0.5325	0.5323	30123

Tablica A.6: Klasifikacijski izvještaj za diskretne vrijednosti para ETH/USDT (Slučajna šuma)

	Precision	Recall	F1-Score	Support
-1	0.5340	0.5120	0.5228	15135
1	0.5269	0.5488	0.5376	14988
Accuracy		0.5303		30123
Macro avg	0.5305	0.5304	0.5302	30123
Weighted avg	0.5305	0.5303	0.5302	30123

Tablica A.7: Klasifikacijski izvještaj za kontinuirane vrijednosti para ETH/USDT (XGBoost)

	Precision	Recall	F1-Score	Support
0	0.54	0.51	0.52	14900
1	0.54	0.57	0.56	15223
Accuracy		0.5417		30123
Macro avg	0.54	0.54	0.54	30123
Weighted avg	0.54	0.54	0.54	30123

Tablica A.8: Klasifikacijski izvještaj za diskretne vrijednosti para ETH/USDT (XGBoost)

	Precision	Recall	F1-Score	Support
0	0.54	0.52	0.53	14900
1	0.54	0.56	0.55	15223
Accuracy		0.5409		30123
Macro avg	0.54	0.54	0.54	30123
Weighted avg	0.54	0.54	0.54	30123

Tablica A.9: Klasifikacijski izvještaj za LSTM model para ETH/USDT pri različitim brojem vremenskih koraka

	Precision	Recall	F1-Score	Support
Eksperiment 1 (2 vremenska koraka)				
0	0.52	0.59	0.55	14898
1	0.54	0.48	0.51	15224
Accuracy		0.5317		30122
Macro avg	0.53	0.53	0.53	30122
Weighted avg	0.53	0.53	0.53	30122
Eksperiment 2 (4 vremenska koraka)				
0	0.52	0.56	0.54	14850
1	0.53	0.50	0.51	15272
Accuracy		0.5259		30122
Macro avg	0.53	0.53	0.53	30122
Weighted avg	0.53	0.53	0.53	30122
Eksperiment 3 (30 vremenskih koraka)				
0	0.53	0.57	0.55	14959
1	0.54	0.49	0.51	15158
Accuracy		0.5308		30117
Macro avg	0.53	0.53	0.53	30117
Weighted avg	0.53	0.53	0.53	30117
Eksperiment 4 (100 vremenskih koraka)				
0	0.53	0.55	0.54	14880
1	0.54	0.52	0.53	15223
Accuracy		0.5343		30103
Macro avg	0.53	0.53	0.53	30103
Weighted avg	0.53	0.53	0.53	30103

Tablica A.10: Klasifikacijski izvještaj za kontinuirane vrijednosti para LINK/USDT (Logistička regresija)

	Precision	Recall	F1-Score	Support
-1	0.52	0.52	0.52	14683
1	0.53	0.52	0.53	14898
Accuracy		0.5234		29581
Macro avg	0.52	0.52	0.52	29581
Weighted avg	0.52	0.52	0.52	29581

Tablica A.11: Klasifikacijski izvještaj za diskretne vrijednosti para LINK/USDT (Logistička regresija)

	Precision	Recall	F1-Score	Support
-1	0.52	0.31	0.39	14683
1	0.51	0.72	0.60	14898
Accuracy		0.5150		29581
Macro avg	0.52	0.51	0.49	29581
Weighted avg	0.52	0.52	0.49	29581

Tablica A.12: Klasifikacijski izvještaj za kontinuirane vrijednosti para LINK/USDT (SVM)

	Precision	Recall	F1-Score	Support
-1	0.49	0.28	0.36	14683
1	0.50	0.71	0.59	14898
Accuracy		0.50		29581
Macro avg	0.50	0.50	0.47	29581
Weighted avg	0.50	0.50	0.47	29581

Tablica A.13: Klasifikacijski izvještaj za diskretne vrijednosti para LINK/USDT (SVM)

	Precision	Recall	F1-Score	Support
-1	0.52	0.02	0.04	15135
1	0.50	0.98	0.66	14988
Accuracy		0.50		30123
Macro avg	0.51	0.50	0.35	30123
Weighted avg	0.51	0.50	0.35	30123

Tablica A.14: Klasifikacijski izvještaj za kontinuirane vrijednosti para LINK/USDT (Slučajna šuma)

	Precision	Recall	F1-Score	Support
-1	0.5236	0.5169	0.5202	14683
1	0.5298	0.5364	0.5331	14898
Accuracy		0.5267		29581
Macro avg	0.5267	0.5267	0.5266	29581
Weighted avg	0.5267	0.5267	0.5267	29581

Tablica A.15: Klasifikacijski izvještaj za diskretne vrijednosti para LINK/USDT (Slučajna šuma)

	Precision	Recall	F1-Score	Support
-1	0.5340	0.5120	0.5228	15135
1	0.5269	0.5488	0.5376	14988
Accuracy		0.5303		30123
Macro avg	0.5305	0.5304	0.5302	30123
Weighted avg	0.5305	0.5303	0.5302	30123

Tablica A.16: Klasifikacijski izvještaj za kontinuirane vrijednosti para LINK/USDT (XGBoost)

	Precision	Recall	F1-Score	Support
0	0.52	0.51	0.51	14699
1	0.53	0.55	0.54	14882
Accuracy		0.5260		29581
Macro avg	0.53	0.53	0.53	29581
Weighted avg	0.53	0.53	0.53	29581

Tablica A.17: Klasifikacijski izvještaj za diskretne vrijednosti para LINK/USDT (XGBoost)

	Precision	Recall	F1-Score	Support
0	0.52	0.51	0.52	14699
1	0.53	0.54	0.53	14882
Accuracy		0.5251		29581
Macro avg	0.53	0.53	0.52	29581
Weighted avg	0.53	0.53	0.53	29581

Tablica A.18: Klasifikacijski izvještaj za LSTM model para LINK/USDT pri različitim brojem vremenskih koraka

	Precision	Recall	F1-Score	Support
Eksperiment 1 (2 vremenska koraka)				
0	0.51	0.52	0.52	14703
1	0.52	0.51	0.52	14878
Accuracy		0.5156		29581
Macro avg	0.52	0.52	0.52	29581
Weighted avg	0.52	0.52	0.52	29581
Eksperiment 2 (4 vremenska koraka)				
0	0.52	0.52	0.52	14828
1	0.52	0.52	0.52	14753
Accuracy		0.5204		29581
Macro avg	0.52	0.52	0.52	29581
Weighted avg	0.52	0.52	0.52	29581
Eksperiment 3 (30 vremenskih koraka)				
0	0.51	0.60	0.55	14541
1	0.53	0.44	0.48	15034
Accuracy		0.5210		29575
Macro avg	0.52	0.52	0.52	29575
Weighted avg	0.52	0.52	0.52	29575
Eksperiment 4 (100 vremenskih koraka)				
0	0.52	0.48	0.50	14698
1	0.52	0.56	0.54	14863
Accuracy		0.5184		29561
Macro avg	0.52	0.52	0.52	29561
Weighted avg	0.52	0.52	0.52	29561