Primjena rudarenje podataka u predviđanje kretanja globalnih indeksa tržišta dionica

Cirković, Ivan

SVEUČILIŠTE U ZAGREBU
FAKULTET ORGANIZACIJE I INFORMATIKE
V A R A Ž D I N

Ivan Cirković

Primjena rudarenja podataka u predviđanju kretanja globalnih indeksa tržišta dionica

ZAVRŠNI RAD

Varaždin, 2019.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Ivan Cirković

Matični broj: 44064/15–R

Studij: Poslovni sustavi

Primjena rudarenja podataka u predviđanju kretanja globalnih indeksa tržišta dionica

ZAVRŠNI RAD

Mentor:

Prof. dr. sc. Božidar Kliček

Varaždin, rujan 2019.

Ivan Cirković

Izjava o izvornosti

Izjavljujem da je ovaj završni rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autor potvrdio prihvaćanjem odredbi u sustavu FOI-radovi

Sažetak

Ovaj rad je usmjeren na dva područja znanosti. U prvom području bit će obrađeni ekonomski pojmovi, poput financijskog tržišta, burze, dionica i globalnih indeksa, koji će služiti kao problemska domena u praktičnom dijelu rada. Drugo područje je fokusirano na proces rudarenja podataka, znanost koja se bavi obradom velikog skupa podataka te izvlačenje korisnih informacija iz tog skupa. Praktični dio rada će pokrivati istraživanje i predviđanje kretanja globalnih indeksa dionica preko procesa rudarenja podataka. Nakon predviđanja ti će rezultati biti interpretirani i analizirani.

Ključne riječi: dionice; globalni indeks; financijsko tržište; otkrivanje znanosti o podacima; rudarenje podataka; predviđanje; modeliranje

Sadržaj

1. Uvod

Svijet je prepun heterogenih podatka koji se godinama skupljaju, sortiraju i organiziraju u bazama podataka. Ti podaci mogu biti nizovi slova, brojevi, datumi, Booleove vrijednosti te ostali podtipovi navedenih oblika ili skupovi istih, ali ti podaci bez njihove točne interpretacije nemaju svrhu, osim navedenih podataka sve značajniji postaju multimedijski podaci, nad kojim se također provode analize.

Jačanjem tehnologija i znanja o podacima, krajem 90-tih godina, nastale su prve teorije o rudarenju podataka. Godinama taj proces je jačao i sada je to ključan čimbenik pri donošenju odluka.

U radu prvo opisujemo financijsko tržište u kojem ćemo definirati važne pojmove s kojima ćemo se susreti u praktičnom dijelu. Nadalje opisat ćemo proces rudarenja podataka, njegov model i metode kojima se možemo služiti i u kojim situacijama se pojedine metode koriste te će biti opisana i šira primjena rudarenja podataka u raznim industrijama.

Glavni problem u ovom radu bit će predviđanje kretanja globalnog indeksa dionica na temelju povijesnih podataka, a to ćemo postići procesom rudarenja podataka u software-u Rapid Miner 5.3.

2. Tržišta i dionice

Sljedeće poglavlje predstavljat će teoretski prvi dio ovog istraživačkog rada, odnosno bit će obrađeni pojmovi financijskog tržišta, burze, dionice i globalnog indeksa.

2.1. Financijsko tržište

Financijsko tržište kao pojam možemo najlakše opisati kao mjesto gdje se susreće ponuda i potražnja. Ponuda je proizvod ili usluga koju trgovac nudi za određenu naknadu odnosno cijenu, dok je potražnja želja i potreba klijenta za nekim proizvodom ili uslugom na tržištu. Primjer tržišta je svako mjesto gdje se odvijaju transakcije, odnosno razmjena usluge ili proizvoda za novac ili drugi oblik plaćanja. Financijska tržišta dijele se na tržište kapitala i novčano tržište, a u ovom poglavlju fokusirat ćemo se na tržište kapitala [1].

2.1.1. Tržište kapitala

Tržište kapitala bavi se ponudom i potražnjom dugoročnih vlasničkih udjela. Ovaj oblik tržišta ima 4 sudionika, a to su: korisnici kapitala, investitori, posrednici i država. Korisnici kapitala na ovom tržištu prikupljaju dugoročna sredstava za ispunjavanje dugoročnih ciljeva njihovog poduzeća, tako da izdaju vrijednosne papire. Emisijom vrijednosnih papira poduzeća gube prvobitnu vlasničku strukturu, a dobivaju potrebni temeljni kapital. Investitori su osobe koje imaju višak financijskih sredstava i žele kupiti vrijednosne papire tzv. dionice od korisnika kapitala, a njihov cilj je da dobro procijene poduzeće u koje će uložiti svoj novac odnosno kapital. Nakon kupovine određenog broja dionica investitor dobiva dividende koje su novčana potvrda dobrog poslovanja poduzeća. Ako investitor želi prodati svoje dionice, to radi preko posrednika, a dobiveni novac, ako je veći od uloženog, zove se kapitalni dohodak. Posrednici u kapitalnom tržištu su brokerska društva i brokerske kuće, odnosno brokeri. Broker je samostalan trgovac koji posreduje između korisnika i investitora, te slaže kupoprodajni ugovor, od kojeg nakon transakcije ima proviziju. Uloga države je da podupire razvoj tržišta kapitala te ga zakonom regulira.

U Republici Hrvatskoj, nadzor i regulativu provodi Hrvatska agencija za nadzor financijskih usluga, dok se za uspješnu provedbu transakcija brine Središnje klirinško društvo.

Kada vlasnik želi skupiti kapital za svoje strateške ciljeve, a da pritom ne podiže kredite u banci, izdaje dionice svojeg poduzeća na tržište. Izdavanje dionica je jako kompleksan proces, te se najčešće odvija preko investicijske banke, koja ima posredničku ulogu. Investicijska banka prvo savjetuje korisnika kapitala o količini i cijeni emisija dionica, nakon toga prikuplja novac i kupuje ih, te ih preprodaje preko javne ponude ili direktnih klijenata [1].

2.1.1.1. Primarno i sekundarno tržište

Kada banka stavi dionice poduzeća na inicijalnu javnu ponudu, tada je riječ o primarnom financijskom tržištu. Primarno financijsko tržište je tržište novih emisija vrijednosnih papira, a njihovom prodajom završava život takvog financijskog instrumenta na primarnom tržištu [2].

Osim primarnog financijskog tržišta postoji i sekundarno tržište. Na sekundarnom tržištu trenutni vlasnik može prodati svoje vrijednosne papire: primjer takvog tržišta je burza. Važno je spomenuti i drugi oblik sekundarnog tržišta koji se zove OTC tržište (eng. *Over the Counter*, što bi na hrvatskom značilo „preko šaltera“. Ovo tržište se naziva izvanburzovno tržište, što znači da nema fiksno mjesto postojanja, također nema brokere nego dilere koji rade transakcije putem telefona, e-mailova i društvenih mreža. Ovaj oblik trgovanja je legalan, ali zbog manjka regulacija i pravila zna biti i opasan u smislu broja prevara. Na OTC tržištima se trguje isto kao i na burzi, ali se u praksi obično ne trguje dionicama, osim dionica koje nisu ispunile burzovne uvjete za plasiranje na primarno tržište [3].

2.1.2. Burza

Burza je stalno i organizirano mjesto na kojem se sastaju kupci, prodavači i brokeri koji trguju prema reguliranim pravilima. Na burzi se trguje robom, vrijednosnim papirima, deviznim novcem i uslugama. Ta roba je fiktivna, to znači da se ne može fizički vidjeti, već je u obliku vrijednosnih papira. Na burzi kupci i prodavači ne mogu direktno vršiti transakciju nego ona se mora izvršiti preko brokera [4].

Postoji više vrsta burzi:

1. Robna
2. Devizna
3. Uslužna
4. Efektna

Robna burza je vrsta burze u kojoj se trguje sirovinama poput zlata, srebra, nafte, kave, čaja, šećera itd. Treba naglasiti da se ne kupuje pravo zlato ili kava nego se dobije vrijednosni papir kupljene robe, koji se može prodati ili zamijeniti s drugom robom. Ovaj oblik kupovanja sirovina najčešće se obavlja preko OTC tržišta, jer nema posrednika, što znači veći profit i brža razmjena, ali u svijetu postoje velike robne burze s fiksnim mjestom i radnim vremenom [4].

Devizna burza (eng. „Forex“ – *Foreign Exchange*) je najveće i najlikvidnije svjetsko financijsko tržište. Transakcije na ovom tržištu se obavljaju preko digitalnih platforma preko kojih, osim razmjena deviza, možemo pronaći i sirovine. Devize su u obliku valutnih parova, najčešći valutni par je EUR/USD, u kojem je euro bazna valuta, a dolar kotirana valuta. Forex je odličan za poduzetnike koji nemaju na raspolaganju za uložiti mnogo novca, zato što postoji poluga margine (eng. *Leverage*), koja može iznositi do 400 puta. To znači da bi trgovac trgovao s 400 puta uvećanim iznosom od uloženog što implicira veliki rast, ali može uzrokovati i veliki pad sredstava [5].

U uslužnim burzama trguje se uslugama koje su najčešće povezane s transportom: u tom obliku burze pojavljuje se ponuda i potražnja za brodski, kamionski i avionski prijevoz tereta. Ova burza je OTC tržište, koja funkcionira tako da se prikazuju sva vozila koja imaju već otprije određenu rutu puta, te u slučaju da pri povratku imaju slobodan prostor u vozilu,

trgovci mogu po jeftinijoj cijeni angažirati tog vozača [6].

Efektna burza je klasični oblik burze u kojoj se trguje vrijednosnim papirima. Na toj burzi osim dionica trguje se obveznicama i kratkoročnim kapitalima. Ova burza ima svoje fiksno mjesto i radno vrijeme. Najpoznatije svjetske burze su: Njujorška burza, NASDAQ, Tokijska burza i Šangajska burza, a od europskih Euronext sa sjedištem u Amsterdamu te Londonska burza [4].

2.1.3. Dionice

Dionice su vrijednosni papiri koji, nakon njihove primarne emisije, postaju dijelovi temeljnog kapitala, a vlasnici udjela, odnosno dionice, postaju dio strukture poduzeća koja ih je izdala. Dionice, kako je već bilo spominjano kroz prošla poglavlja, mogu se prodavati na sekundarnom tržištu, najčešće na burzi. Postoji nekoliko vrsta dionica, a te vrste se razlikuju u količini prava koje investitor ima kada kupi dionicu:

1. Redovne
2. Povlaštene
3. Povlaštene bez prava glasa
4. Osnivačke
5. Radničke dionice
6. Dionice s pravom otkupa

Investitori u redovnim dionicama imaju pravo na upravljanje, odnosno prava u odlučivanju u slučaju donošenja nekih odluka koje se tiču politike poduzeća, a uz to imaju prava na isplatu dividende.

U povlaštenim dionicama investitori imaju nešto veća prava nego vlasnici redovnih dionica, kao što su isplate odštete u slučaju stečaja, a u nekim slučajevima, investitori povlaštenih dionica imaju veća prava glasa dionica, no u većini zemalja to je zabranjeno. U povlaštenim dionicama bez prava glasa, investitor ima samo prava na isplatu dividendi.

U osnivačkim dionicama investitori imaju prava na dividende, tek u slučaju ako cijene redovnih dionica narastu do određene svote.

Kada vlasnici nemaju sredstava za plaćanje radnika postoji mogućnost da im za tu naknadu daju radničke dionice.

Dionice s pravom otkupa moraju biti unaprijed određene statutom, pri prvoj emisiji dionica na primarnom tržištu, jer u pravilu vlasnici ne mogu kupovati dionice natrag od investitora ako to investitori ne žele [7].

2.1.3.1. Globalni indeksi i cijene dionica

U svijetu dionica moramo znati razliku između cijene dionice i broja bodova globalnog indeksa. U ovom završnom radu zadatak je prognozirati globalni indeks. Globalni indeks je prikaz skupa cijena dionica različitih poduzeća izražen u realnom broju. Za primjer globalnog indeksa možemo uzeti hrvatski globalni indeks CROBEX koji je trenutno sastavljen od 16 poduzeća, a u vlasništvu je Zagrebačke burze [8]. Globalni indeks se računa preko ponderiranog prosjeka, što znači da svaka sastavnica u indeksu odnosno poduzeće ima svoju težinu, što je veća težina to više utječe na globalni indeks [9].

Izdavatelj	Broj dionica	Free float faktor	Težinski faktor	Zadnja cijena	▾ Tržišna kapitalizacija	Težina
Valamar Riviera d.d.	126.027.542	0,55	0,46067227	37,00	1.181.467.964,96	10,53%
ADRIS GRUPA d. d.	6.784.100	1,00	0,36464363	465,00	1.150.307.165,38	10,25%
PODRAVKA d.d.	7.120.003	0,85	0,46668737	401,00	1.132.581.654,47	10,10%
ATLANTIC GRUPA d.d.	3.334.300	0,45	0,59765656	1.210,00	1.085.061.232,93	9,67%
HT d.d.	81.670.064	0,45	0,18300154	159,00	1.069.368.232,47	9,53%
KONČAR, d.d.	2.572.119	1,00	0,68027413	595,00	1.041.098.878,91	9,28%
Arena Hospitality Group d.d.	5.128.721	0,50	1,00000000	362,00	928.298.501,00	8,28%
ERICSSON NIKOLA TESLA d.d.	1.331.650	0,55	1,00000000	1.120,00	820.296.400,00	7,31%
Zagrebačka banka d.d.	320.241.955	0,04	1,00000000	61,00	781.390.370,20	6,97%
AD PLASTIK d.d.	4.199.584	0,70	1,00000000	173,00	508.569.622,40	4,53%
ATLANTSKA PLOVIDBA d.d.	1.395.520	0,75	1,00000000	437,00	457.381.680,00	4,08%
JADRAN d.d.	27.971.463	1,00	1,00000000	15,40	430.760.530,20	3,84%
MAISTRA d. d.	10.944.339	0,11	1,00000000	302,00	363.570.941,58	3,24%
OT-OPTIMA TELEKOM d.d.	69.443.264	0,40	1,00000000	6,10	169.441.564,16	1,51%
Dalekovod, d.d.	24.719.305	0,40	1,00000000	6,12	60.512.858,64	0,54%
ĐURO ĐAKOVIĆ GRUPA d.d.	10.153.230	0,45	1,00000000	8,12	37.099.902,42	0,33%

Slika 1. Sastav CROBEX-a , izvor: zse.hr

Cijene dionice pojedinog poduzeća ovise o mnogo faktora. Temeljni čimbenici za cijenu dionice su EPS (eng. *Earnings per Share*) i P/E omjer (eng. *Price/Earnings*). EPS na hrvatskom znači „zarada po dionici“, izračunava se tako da podijelimo profit poduzeća s brojem izdanih dionica, što je veći taj broj, to je bolje za dioničare, jer tolika je zarada od jedne dionice. P/E omjer, odnosno omjer cijene i zarade, dobijemo tako da podijelimo trenutnu vrijednost dionice s EPS. Omjer koji dobijemo predstavlja koliko puta će cijena dionica narasti u budućnosti uzimajući u obzir da novac neće vrijediti isto kao danas.

Osim temeljnih čimbenika postoje tehnički faktori i osjetilni marketing, koji uvjetuju cijenu dionice. Tehnički faktori su: inflacija, druga tržišta, konkurencija, ponuda i potražnja, demografija, trendovi i likvidnost. U osjetilnom marketingu promatra se ponašanje investitora, odnosno postavlja se pitanje zašto je neki investitor kupio ili prodao dionice. Rastom promatranja nastala je znanost o bihevioralnim financijama u kojoj se promatraju nepravilnosti u donošenju poslovnih odluka u današnjoj ekonomiji [10].

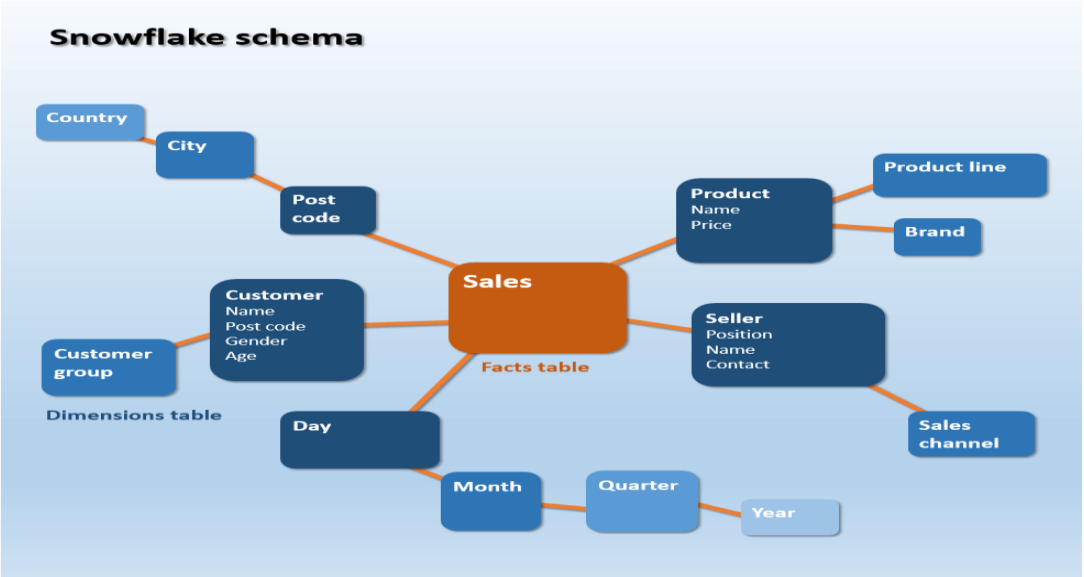
3. Rudarenje podataka

Za pojam rudarenje podataka postoji mnogo različitih definicija, iz kojih možemo izlučiti da je rudarenje podataka traženje korisnog znanja u hrpi skupova podataka, a služi za poboljšanje donošenja odluka.

U ovom poglavlju bit će objašnjeni pojmovi baza podataka i skladište podataka, nakon toga upoznat ćemo se sa znanosti o otkrivanju znanja iz baza podataka, zatim ćemo definirati pojam „rudarenje podataka“ kao kompleksni proces. Nadalje, bit će prikazan CRISP-DM model za rudarenje podataka i njegove faze i metode te kako se rudarenje podataka koristi u raznim industrijama [11].

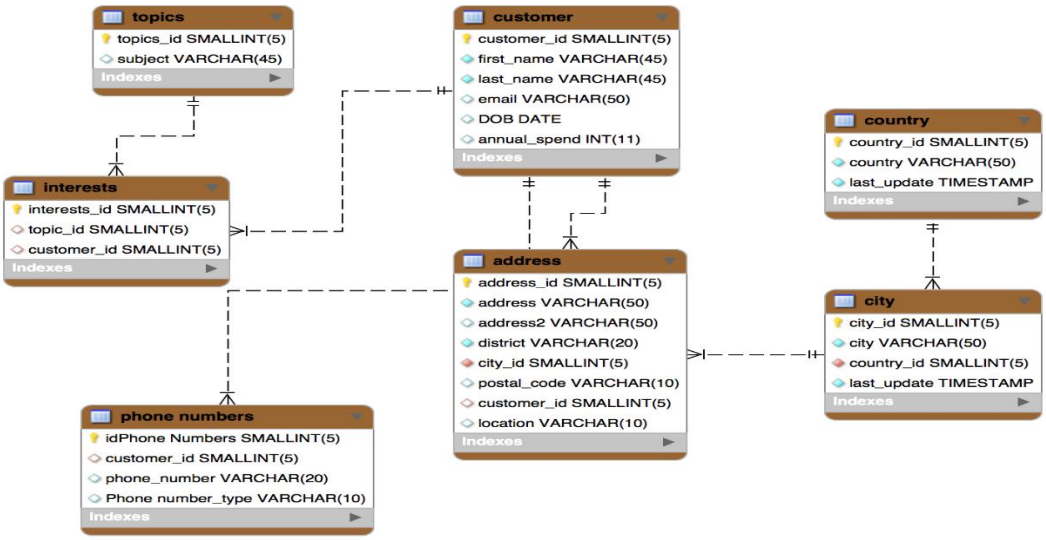
3.1. Spremanje podataka

Rastom tržišta, korisnika i broja transakcija nastaje veliki skup podataka koji se sprema u dva oblika sistema skladištenja. Prvi oblik spremišta podataka (eng. *Data Warehouse*) u kojima se spremaju povijesni podaci su tablice koje su višedimenzionalnih oblika, njihove veze su 1 naprama više, a ti podaci se pomoću procesa rudarenja podataka pretvaraju u informacije koje odgovaraju zadanom upitu. U procesu rudarenja podataka koristimo ovakav oblik sistema spremišta.



Slika 2. Struktura skladišta podataka,
izvor: https://www.ionos.com/digitalguide/online-marketing/web-analytics/dwh-what-is-a-data-warehouse/

Podaci o transakcijama spremaju se u relacijsku bazu podataka (eng. *Relation Database*). Ti podaci su normalizirani tj. organizirani u više manjih tablica koje su u relaciji. Podaci moraju biti dobro organizirani da bi korisnici u bilo koje vrijeme imali uvid u svoje transakcije. Podaci se mogu analizirati SQL upitima, ali budući da su tablice u vezi 1:1, sustav je pod velikim opterećenjem, jer se sve tablice moraju pretražiti. Spremljeni podaci se mogu manipulirati RDBMS softwareom (eng. *Relation Database Management System*), kojem su funkcije da definira, ažurira, daje informacije drugim podsustavima, nadgledava rad korisnika, štiti podatke te nadzire potencijalne prijetnje sustavu.



Slika 3. Struktura baze podataka,

izvor: https://www.mongodb.com/blog/post/mongodb-multi-document-acid-transactions-general-availability

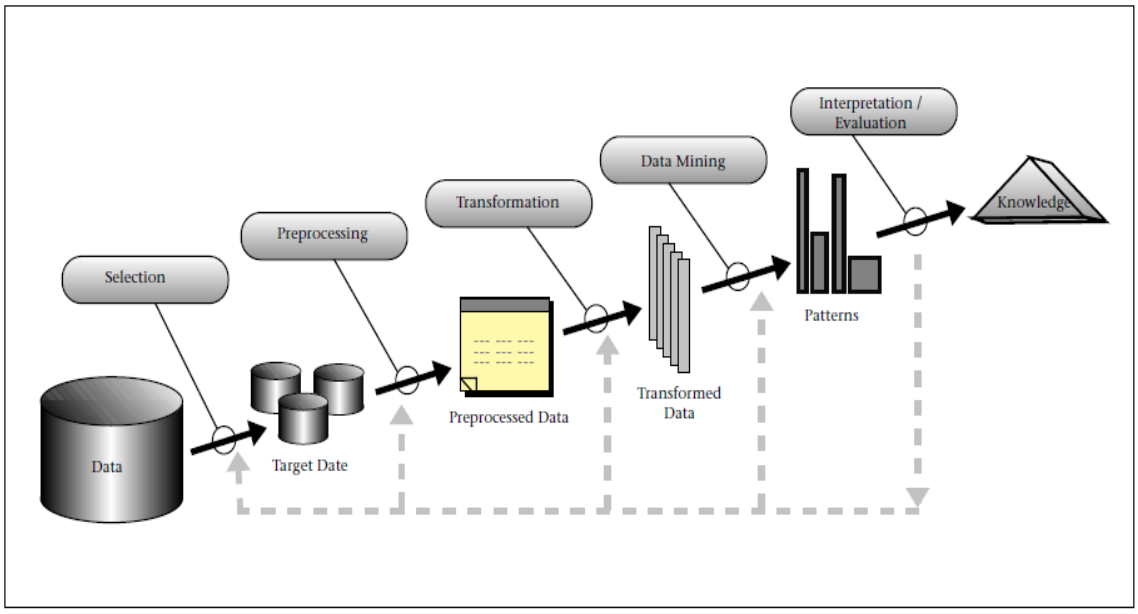
Iz definiranja ovih pojmova možemo zaključiti da skladište podataka i baza podataka imaju istu funkciju, a to je spremanje podataka, a glavne razlike su im uporaba i arhitektura spremanja podataka [12].

3.2. Otkrivanje znanja u bazama podataka – KDD

Krajem prošlog stoljeća u svim industrijskim granama broj novih podataka rastao je eksponencijalno pa se početkom 1990-tih pojavila teorija o otkrivanju znanja u bazama podataka (eng. *Knowledge Discovery in Databases*). U članku „From Data Mining to Knowledge Discovery in Databases“, autori navode: *„There is an urgent need for a new generation of computation-al theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data“.* Ovaj citat govori da postoji velika potreba za novim tehnologijama, znanjima i alatima koji će pomoći ljudima da lakše pronađu korisne informacije tj. znanje iz velikih skupova podatka koji rapidno rastu.
Iz bazičnog pogleda KDD je oblik znanosti koja se brine za napredak metoda i tehnika obrade informacija. Iz godine u godinu KDD raste i nastavit će rasti, a ključni cilj KDD-a je maksimalno izvlačenje visoko klasne informacije iz običnih podataka koji bez obrade nemaju svrhu. KDD se može definirati kao vrsta znanja u kojoj se spajaju sve teorije i alati za obradu podataka [13].

Mnogi ljudi griješe u razlikovanju pojmova KDD-a i rudarenja podataka. Naime opći pojam *„Data mining“* odnosno rudarenje podataka, može se poistovjetiti s KDD-om, ali iz slike 4. možemo zaključiti da je proces rudarenja podataka samo jedna faza procesa KDD-a, odnosno KDD ima faze pripreme, razumijevanja, selekcije podataka te ih završno interpretira i procjenjuje izvještaj .

3.2.1. Modeli procesa KDD-a



Slika 4. Model procesa KDD-a
izvor: https://www.researchgate.net/figure/The-Knowledge-Discovery-in-Databases-KDD-process-Fayyad-et-al-1996_fig1_220673065

U literaturama se spominju tri oblika modela procesa KDD-a. Osnovni KDD model sastoji se od 5 koraka koji su prikazani na slici iznad: selekcija, priprema, transformacija, rudarenje, interpretacija i procjena podataka.

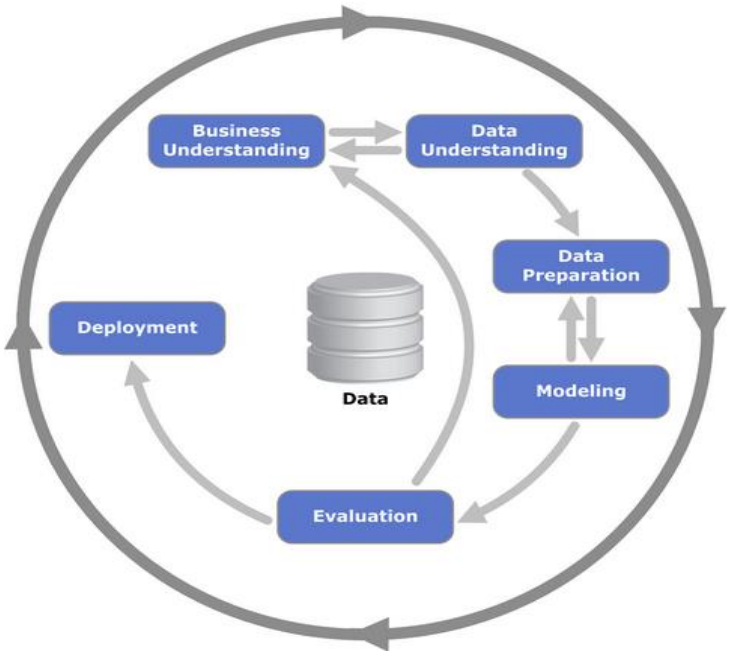
Drugi oblik modela procesa KDD-a zove se SEMMA model. SEMMA je akronim od 5 koraka od kojih se model sastoji: *Sample* (uzorak), *Explore* (istraživanje), *Modify* (izmjena), *Model* (model) i *Assess* (procjena). Prvi i drugi oblik model su identični, a njihova je razlika u nazivima korak, osim toga SEMMA model je djelo programske kompanije SAS instituta Zbog te činjenice možemo zaključiti da taj model najviše koriste njihovi korisnici - naime svih 5 koraka je integrirano u njihov software SAS Enterprise Miner software.

Treći oblik je „Cross-industry standard process for data mining“ tj. skraćeno zapisano CRISP-DM. CRISP-DM je prošireni model osnovnog KDD-a, te isto kao i SEMMA model za nazive koraka ima drugačije pojmove. Osim naziva koraka sadrži korak više od osnovnog KDD modela [14].

Godine 2015. kompanija IBM je objavila svoji novi model tj. prošireni CRISP-DM model, naziva „Analytic Solutions Unified Method“ koji je integriran u njihovom software-u „IBM Analytic Solutions“. U online anketi koja je provedena 2014. godine stranica „KDnuggets“, CRISP-DM je najpopularniji model kojim se služe analitičari podataka. U anketi u kojoj je glasalo 200 znanstvenika, CRISP-DM model je dobio 43% glasova, SEEMA model 8.5%, osnovni KDD model 7.5%, a ostali postotak glasova (41 %) koriste svoj sistem za proces KDD-a. Važno je napomenuti da su u ovom poglavlju korišteni vlastiti prijevodi za nazive koraka [15].

3.2.2. CRISP-DM model

U ovom poglavlju bit će navedeni i objašnjeni koraci modela CRISP-DM, koji je najpopularniji model i koji se pojavljuje u najviše svjetskih literatura, te se iz anketa može zaključiti da se taj model najviše koristi. Iz Slike 5. možemo zapaziti da CRISP-DM model ima 6 koraka, te da su veze između koraka dvosmjerne što znači da se možemo vraćati iz trenutnog u prethodni korak i obrnuto. Veze također kreiraju životni ciklus koji nam govori, da se proces nakon dobivanja traženih informacija vraća na prvu fazu, te se procjenjuje jesu li ti podaci točni, odnosno zadovoljavaju li kriteriji prve faze, pa u slučaju da rezultati nisu u skladu poslovnog cilja, postupak se ponavlja [16].



Slika 5. Model procesa CRISP-DM, izvor: itsalocke.com

izvor: https://itsalocke.com/blog/crisp-dm-and-why-you-should-know-about-it/

3.2.2.1. Poslovno razumijevanje

Prvi korak tj. proces u modelu CRISP-DM je poslovno razumijevanje (eng. *Business Understandig*). Primarni je cilj menadžera projekta da jasno razumije što klijent želi. Klijenti poduzeća koja se bave uslugom rudarenja podataka, mogu tražiti smanjenje odlaska trenutnih klijenata, stjecanje novih klijenata, poboljšanje prodaje i sprečavanje prevara. Nakon razumijevanja želja klijenta, menadžer i klijent dogovaraju kolika je zadovoljavajuća granica usluge, odnosno koliki je postotak točnosti i postotak sigurnosti u predviđanju. Kada se menadžer i klijent dogovore, menadžer mora sa svojim timom osmisliti strategiju projekta, analizirati poslovne ciljeve i procijeniti troškove i dobit, te zaključiti je li moguće zadovoljiti njegove potrebe. Zadnji korak u ovoj fazi je da menadžer jasno zna na koji način će napraviti ostale korake i kako razviti model [17].

3.2.2.2. Razumijevanje podataka

Drugi proces, nakon postavljanja i razumijevanja glavnog cilja projekta, je razumijevanje podataka (eng. *Data Understandig*). U ovom procesu modela projektni tim skuplja podatke, te kreira skladište podataka (eng. *Data Warehouse*). Ti podaci mogu biti jednostavne baze podataka, tekstovi, Excel dokumenti itd. U ovoj fazi modela ključno je iz tih podataka zaključiti jesu li oni potrebne kvalitete i kvantitete za rudarenje podataka. Specijalisti nakon pregleda podataka iste moraju istražiti detaljnije, a to rade preko traženja istih varijabli i uzorka u podacima te ih testiraju na jednostavnim hipotezama. Nakon testiranja specijalisti se moraju uvjeriti da su podaci realni i da nemaju anomalije npr. da im ne nedostaju neke vrijednosti [17].

3.2.2.3. Priprema podataka

Priprema podataka (eng. *Data Preparation*) prema mnogim specijalistima, vremenski najduža faza projekta, koja nekad zauzme i do 80% utrošenog vremena cijelog projekta. U ovom koraku specijalisti prvo moraju očistiti podatke od kojih neće imati koristi. Iz očišćenih podataka uzimaju se varijable koje su zavisne za poslovni cilj, potom se integriraju u skup podataka, koji se može urediti i proširiti novim vrijednostima preko trenutnih varijabli [17].

3.2.2.4. Modeliranje

Modeliranje (eng. *Modeling*), je faza u kojoj specijalisti izrađuju model za rudarenje podataka. Specijalisti prvo iz skupa podataka označuju ciljne varijable koje žele predvidjeti, te koje će koristiti kao inpute. Postoje dva stila biranja podatka, u prvom stilu specijalisti se oslanjaju da koriste varijable iz hipoteze koje su odredili u prvoj fazi poslovnog razumijevanje, drugi pristup podržava višak atributa u procesu rudarenja podataka, što specijalistima znači da nisu ništa propustili, ali postoji mogućnost da ima previše nepotrebnih podataka te rudarenje može biti sporo i dati krive informacije. Odabir ovog pristupa ovisi i o metodi strojnog učenja koje će biti objašnjeno u sljedećim poglavljima. Nakon odabira, ciljne varijable, specijalisti spajaju operatore algoritama koji obrađuju podatke i daju rezultate. Detaljnije o metodama rudarenja podataka u poglavlju 3.3. [17].

3.2.2.5. Procjena rezultata

Nakon napravljenog modela specijalisti moraju testirati točnost i preciznost tog modela. Procjena rezultata (eng. *Evaluation*) je korak u kojem specijalist razdvaja skup podataka u dvije particije najčešćeg omjera 30/70 %, te na obje provede iste metode rudarenja podataka, pa ako su rezultati identični, model je ispravan. Nakon toga uspoređuje se procjena točnosti s onom particijom koja je dogovorena u prvoj fazi CRISP-DM modela. Za kraj specijalisti se moraju uvjeriti da je taj model pokrio sve moguće situacije koje bi utjecale na točnost izvještaja [17].

3.2.2.6. Razvoj modela

Razvoj modela (eng. *Deployment*) može u ovom modelu biti najlakši, a ponekad i najteži korak od svih. Klijentima koji su npr. tražili popis imena kupaca koji će ponoviti kupnju dovoljan je samo popis s imenima i predviđanje, no drugi klijenti mogu tražiti puno kompleksniju uslugu, da se taj model razvije i integrira u software koji bi automatski ažurirao rezultate ako se dogodi nova kupnja, te bi im služio kao simulator za predviđanje potražnje. Ako se dogodi drugi tip suradnje, poduzeća mogu dogovoriti suradnju za nadziranje software i korisničku podršku. Za sam kraj ovog procesa menadžer pokazuje i objašnjava klijentu dokumentaciju te interpretira tj. prezentira rezultate klijentu [17].

3.3. Metode i njihov izbor za proces rudarenja podataka

Metode za rudarenje podataka primjenjuju se u 4. fazi CRISP-DM modela. Specijalist se prije izrade modela mora odlučiti koje će algoritme za rudarenje podataka koristiti. Postoji još jedan važan korak prije odabira metode, a to je kakve podatke imamo, odnosno moramo odrediti vrstu strojnog učenja. U znanosti o strojnom učenju (eng. *Machine Learning*) postoje dvije glavne vrste strojnog učenja, a to su „*Supervised Learning*“ i „*Unsupervised Learning*“ odnosno nadzirano i ne-nadzirano učenje, postoji još vrsta strojnog učenja, ali za ovu temu dosta su nam ova dva navedena [18].

3.3.1. Vrste strojnog učenja

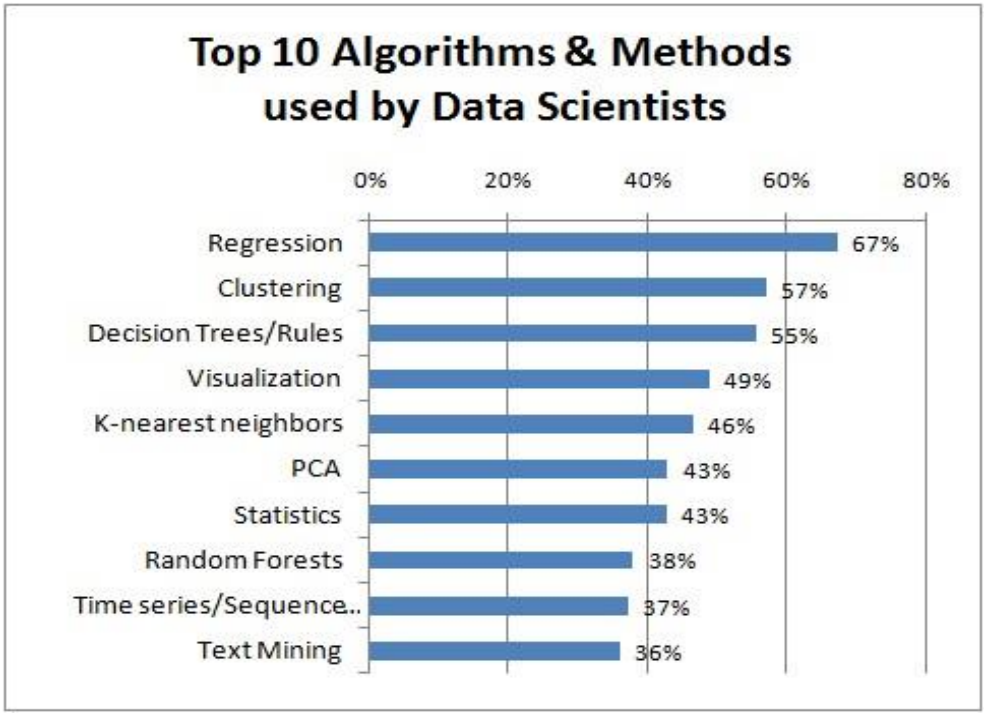
Kod nadziranog učenja cilj je software-u dati naše znanje, odnosno skup podataka koji ima svoju ciljnu varijablu koju želimo predvidjeti i ostale ulazne varijable preko kojih software traži zajedničke veze. Nadzirano učenje će se primjenjivati u ovom radu, gdje će ciljna varijabla biti zadnja cijena indeksa dionice, a prva, najviša i najniža cijena bit će ulazne varijable. Iz nadziranog učenja stvaraju se dvije vrste metoda: klasifikacija i regresija.

Kod ne-nadziranog učenja softwareu dajemo skupove podataka u kojima nema označenih varijabli, o tim podacima nemamo veliko znanje. Cilj software-a je da na temelju

tih podataka nalazi sličnosti u njima, odnosno uzorke, od kojih radi zasebne grupe podataka na temelju tog promatranja. Iz ne-nadziranog učenja isto nastaju dvije vrste metoda: asocijacija i klasteriranje [18].

3.3.2. Metode i algoritmi

Postoji preko tisuću algoritama za rudarenje podataka. Na slici 3. prikazani su rezultati ankete iz 2016. godine, koju je provela stranica KDnuggets i u kojoj je sudjelovalo više od 800 ljudi. Anketa pokazuje koje algoritme specijalisti za podatke najviše koriste. U idućim poglavljima bit će objašnjene 4 glavne metode i navedeni njihovi algoritmi.



Slika 6. Rezultati ankete o korištenju metoda,
izvor https://www.kdnuggets.com/2016/09/poll-algorithms-used-data-scientists.html

:

3.3.2.1. Klasifikacija

U ovoj metodi software-u dajemo dio skupa podataka iz skladišta podatka. Ti podaci se dijele na dva dijela: prvi dio su podaci za trening modela, odnosno učenje, a s ostatkom podatka se provodi testiranje točnosti i preciznosti modela. Podaci za trening su označeni s ciljnom varijablom i ostalim ulaznim varijablama. U klasifikaciji ta ciljna varijabla mora biti

u obliku kategorije. Najčešći primjer klasifikacije možemo pronaći u bankarstvu: kada klijenti žele podići kredit, bankar mora vidjeti je li klijent kreditno sposoban. U ovom primjeru glavna ciljna kategorija je kreditni rizik koji može biti visok, srednji ili nizak, te na temelju ulaznih varijabli, poput mjesečne plaće, dobi, spola i radnog iskustva, bankar dobiva kategoriziran odgovor je li klijent sposoban vraćati kredit.

Kada software pripremi podatke testira ih na jednom ili više algoritama. U praksi se uzima više algoritama te se iščitava njihova preciznost, a rezultati se uspoređuju. U klasifikaciji se koriste najviše sljedeći algoritmi:

1. Logička regresija
2. Naïve Bayes algoritam
3. Stohastički gradijent
4. K-Nearest Neighbours - K- najbliži susjedi
5. Decision tree - Stabla odlučivanja
6. Random forest - Slučajna šuma
7. Support vector machine (SVM) - Stroj s potpornim vektorima

Osim u klasifikaciji, ovi algoritmi se mogu koristiti i u drugim metodama [19].

3.3.2.2. Regresija

Ovoj vrsti metode isto kao i u prethodnoj, dajemo naše podatke za trening, koji isto sadrže ciljnu i ulazne varijable, ali primarna je razlika da je u slučaju regresije naša ciljna varijabla u brojčanom obliku, a ne kategorija. Regresija i klasifikacija mogu koristi iste algoritme za predviđanje. Ako se vratimo na poglavlje 3.1.2.5. Procjena rezultata za regresijski model možemo izračunati postotak točnosti i postotak sigurnosti predviđanja, koji su bili dogovoreni u prvoj fazi CRISP-DM modela. Postotak točnosti se računa pomoću srednje kvadratne pogreške koja nam služi da utvrdimo kvalitetu modela i algoritma. Postotak sigurnosti računa se tako da se usporede regresijski model i testni model (koji radi predviđanja na temelju prosjeka): što je veća razlika u postocima toliko je posto regresijski model sigurniji od testnog. Iz slike 6 rezultati ankete pokazuju da se metoda regresije najviše koristi u svijetu rudarenja podataka [20].

3.3.2.3. Asocijacija

Ovoj vrsti metode dajemo veliki skup najčešće transakcijskih podataka, a glavni je cilj da software nađe neku poveznicu između njih. Ova metoda koristi tehniku „Asocijativna pravila“, koja se najviše koristi u marketingu za analizu potrošačke košarice, odnosno za lakše promoviranje i procjenu cijene proizvoda. Ako se vratimo na poglavlje 3.1. možemo zaključiti da su ti podaci zapisani u više malih tablica koje su povezane vanjskim ključevima, stoga te zapise moramo transformirati, inače to uradi software automatski. Tehnika asocijativnih pravila analizira i predviđa ponašanje kupaca, a to se postiže s upitima poput: ako kupac kupi ovaj proizvod, hoće li kupiti i ovaj dugi?

Možemo zamisliti da jedna kompanija želi podići unakrsnu prodaju, odnosno da svojim klijetima uz jedan proizvod, odmah ponudi drugi, npr. kupac kupi miš za računalo, a prodavač će mu ponuditi podlogu za njega. Cilj ove metode je da u skupu zadanih podataka nađe takve parove proizvoda. Ovaj algoritam ima dva ključna parametra koji se bilježe u postocima, a to su „*Support* “ i „*Confidence*“, odnosno podršku i povjerenje, koji moraju biti što veći. Podršku možemo objasniti da je taj dan u tom dućanu bilo sveukupno 100 transakcija, a 20 kupaca je kupilo miš za računalo, kada podijelimo 20 sa 100, dobijemo 20% i to je podrška. Ako je od tih 20 kupaca, 8 kupaca kupilo i podlogu za miša, povjerenje u ovom slučaju dobivamo tako da 8 podijelimo s 20 i pomnožimo sa 100, dobijemo povjerenje u iznosu 40%. U praksi specijalisti odrede sami minimalni postotak ova 2 parametra pa software traži parove, te ih filtrira na temelju tih uvjeta.

Postoje više načina upita, to može biti jednostavan upit poput primjera s mišem i podlogom, a može biti i složeniji, npr. ako je kupac student i ima više od 20 godina, koliki je postotak sigurnosti kupovanja laptopa [21].

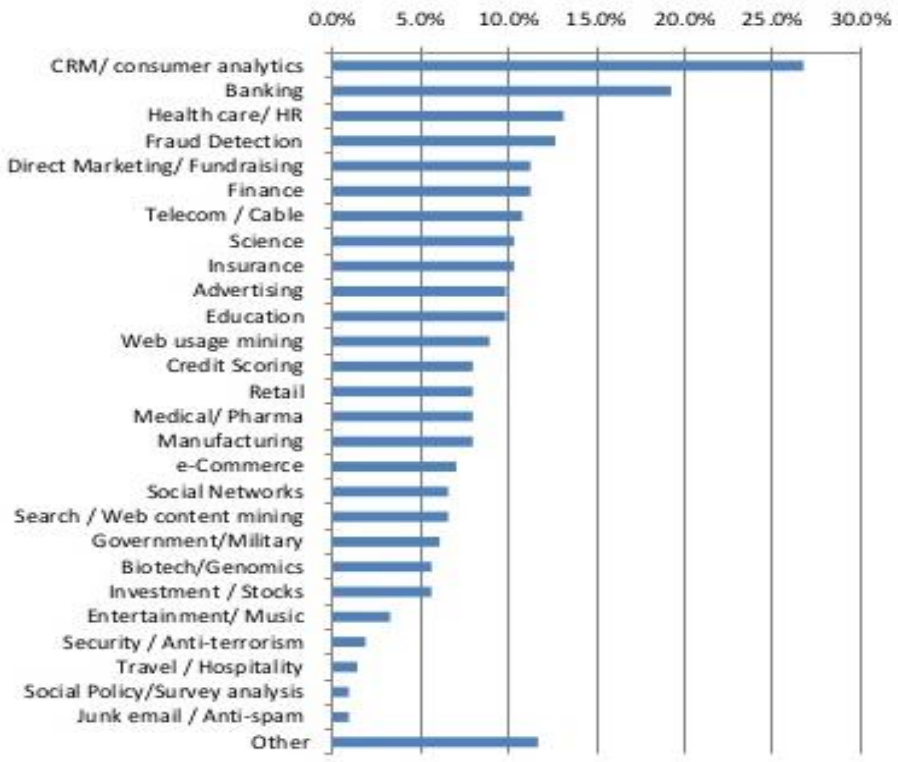
3.3.2.4. Klasteriranje

Klasteriranje je metoda slična metodi asocijacije, razlika između te dvije metode je u tome da se u asocijaciji procjenjuje postotak ponavljanja istog uzorka, dok je u klasteriranju glavni cilj pronaći podatke koji su homogeni iz heterogenog skupa podataka te ih razvrstati u zasebne particije. Primjer klasteriranja možemo pronaći svugdje u okolini, npr. knjižnica, koja sadrži sve vrste literature, odnosno knjiga koje su posložene u „klastere“ tj. zasebne police za svaku vrstu literature. Prema anketi (slika 6.) možemo primijetiti da je ovo metoda druga po korištenju. Klasteriranje se često koristi na internetu, za primjer možemo navesti web tražilice u kojima su podaci u oblaku (internetu) klasterirani za bržu pretragu podataka. Osim toga klasteriranje se primjenjuje i u digitalnom marketingu, pri analizi društvenih mreža, u kojoj se traže skupovi ljudi koje vole iste stvari; pomoću tih grupa reklame na društvenim mrežama izgledaju kao da su personalizirane.

Ova metoda ima mnogo algoritama za rudarenje podataka, pa se i s tom činjenicom javlja problem, koji algoritam primijeniti? Zato prije odabira moramo dobro pripaziti u fazi poslovnog razumijevanja, a i u fazi pripremanja podataka. Prema slici 6. na rezultatima možemo primijetiti dva algoritma koji se najviše koriste to su PCA („Principal Component Analysis“) i K-means clustering [22].

3.4. Raznovrsna primjena rudarenja podataka

U ovom smo radu već spominjali primjere iz prakse gdje se koristi rudarenje podataka, pa možemo zaključiti da se koristi u svakoj grani industrije, to možemo potvrditi činjenicom da nema grane industrije koja ne stvara bilo kakve podatke. Prema slici 7. možemo uočiti da se rudarenje podataka najviše koristi u: CRM, bankarstvu, zdravstvu, detekciji prevara u raznim područjima, marketingu i tako dalje. Iz ankete možemo vidjeti da nakon širokog popisa primjene postoji još opcija „*Other*“, odnosno ostala primjena koja je isto dobila značajni broj postotka.



Slika 7. Rudarenje podataka u industrijama,

izvor: https://www.kdnuggets.com/polls/2010/analytics-data-mining-industries-applications.html

U sljedećem dijelu ovog poglavlja bit će navedena 2 primjera primjene rudarenja podataka, gdje su specijalisti za podatke proveli istraživanja nad specifičnim slučajevima, kako bi otkrili neke nepoznate informacije iz skupova podataka.

2013. godine grupa profesora, napravilo je istraživanje u kojem su preko rudarenja podataka predviđali ocjene učenika različitih škola. Cilj ovog istraživanja je bio predvidjeti njihove konačne ocjene, te upozoriti na vrijeme one koji su bili u opasnosti od pada. Istraživanje su proveli nad 900 učenika, a iskoristili podatke od 500 njih, svaki učenik je ispunio upitnik od 50 pitanja. Nakon prikupljanja podataka i pripremanja, proveli su Chi-square test, koji se koristi kada imamo skup podataka koji ima puno različitih atributa te želimo saznati njihovu povezanost. Pomoću ovog testa razvili su nekoliko hipoteza, koje govore da vrsta škole ne utječe na predviđanje ocjene učenika, ali zato zanimanje roditelja jako utječe na predviđanje. Ovo istraživanje su proveli preko metoda klasifikacije, odnosno preko 5 algoritama, a najuspješniji je bio MLP algoritam („Multi Layer Perception“) koji je uspješno pogodio 72,38% [23].

Osim u edukaciji veliku primjenu rudarenja podataka možemo pronaći u zdravstvenom sektoru, u članku iz 2013. godine „Data Mining Applications In Healthcare Sector: A Study“, autori su napravili komparativno istraživanje u kojem su naveli sve grane

zdravstvenog sustava u kojima se rudarenje podataka koristi: u procjeni efikasnosti liječenja pacijenata, organiziranju i zaštiti zdravstvenog sustava, u sprečavanju prevara i iskorištavanja sustava, u farmaceutskom sektoru za organiziranje slanja lijekova, u bolničkom sustavu za organiziranje kapaciteta bolnica, te u genetici i neurologiji. Posebni dio istraživanja su posvetili predviđanju čestih i opasnih bolesti. U članku su naveli koju metodu i algoritme je najbolje koristi za predviđanje (otkrivanje) pojedine bolesti, te koliki je postotak točnosti predviđanja. Za većinu bolesti koristili su metodu klasifikacije i njene algoritme koji su prosjeku točni 80%, najbolji rezultat je dalo predviđanje raka koje je točno u čak 97,77%. Bolesti poput raka mozga i AIDS-a, za njihovu detekciju su koristili metodu klasteriranja, odnosno asocijativnih pravila, a točni su između 80-85% slučajeva [24].

4. Primjena rudarenja podataka za prognoziranje globalnih indeksa tržišta dionica

Ovo poglavlje predstavljat će istraživački dio rada, u kojem ćemo pomoću rudarenja podataka prikazati povezanosti između nekoliko globalnih aktivnosti svjetskog gospodarstva. U ovom radu koristit ćemo se softwareom Rapid Miner 5.3. pomoću kojeg ćemo izraditi model za zadani problem te interpretirati rezultate, trenutna aktualna verzija ovog software-a je 9.3., koja ima više mogućnosti i proširenja, ali s tim proširenjima dolazi do povećane potrošnje RAM memorije koja je bila prevelika za računalo na kojem je izrađen model. Važno je napomenuti da praktički nije moguće prognozirati točan iznos globalnog indeksa za pojedini dan, ali možemo predvidjeti trend, odnosno rast ili pad za idući dan.

4.1. Razumijevanje problema i priprema podataka

Odabirom ove teme pojavljuje se glavni problem: „Mnoga istraživanja potvrdila su empirijske zaključke, o povezanosti različitih pokazatelja globalnih aktivnosti svjetskog gospodarstva (indeksi tržišta dionica, valuta i zlata, te ključnih sirovina). U ovom radu je potrebno napraviti pregled tih istraživanja, te analizirati primijenjene metode, s naglaskom na metode rudarenja podataka. Nadalje, potrebno je napraviti opći model međuovisnosti tih veličina, te za njih osigurati podatke. Kritički se osvrnuti na prednosti i nedostatke takvog pristupa.“ [25].

4.1.1. Skupljanje podataka

Nakon razumijevanja ovog problema možemo krenuti sa skupljanjem podataka. Budući da u problemu nije točno definirano koje podatke treba koristiti, uzet ćemo ove ulazne varijable: cijena zlata (unca), srebra (unca), i nafte(barela), i pomoću njih prognozirati globalni indeks S&P 500. Svi podaci su preuzeti s internetske stranice Yahoo Finance.

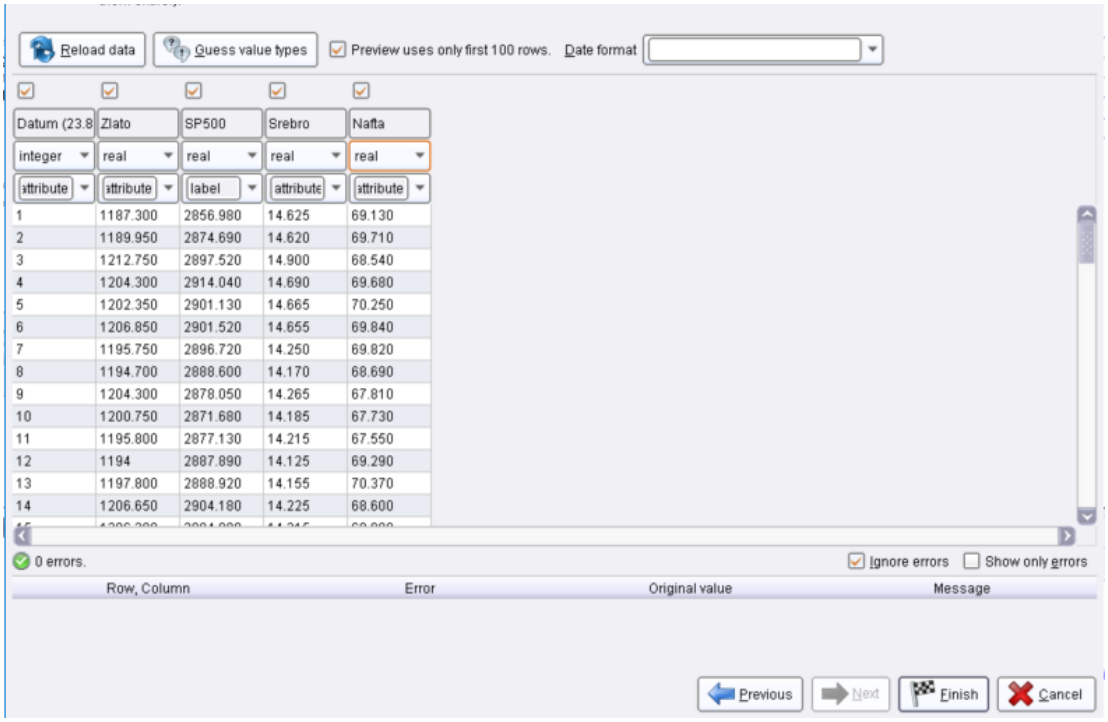
Proces rudarenja podataka koristi se godinama, te su provođena razna istraživanja na ovu temu rada, ali većinom su specijalisti predviđali kretanje globalnog indeksa na temelju atributa istog indeksa. Primjer takvog istraživanja su napravili brazilski znanstvenici 2009. godine, koji su na temelju historijskih podataka u razdoblju 1998. do 2007. godine, pokušali predvidjeti kretanje globalnog brazilskog indeksa Ibovespa, za razdoblje travnja 2007. do ožujka 2008. godine. U istraživanju su koristili algoritam Neuronskih mreža, a pokušali su predvidjeti 1 dan unaprijed sa što više kombinacija vremenskih modela. Kreiranje vremenskih modela će biti objašnjeno u fazi modeliranja našeg modela. Najbolja kombinacija ovog istraživanja imala je 60 % točnosti [26].

U našem istraživanju mi smo koristili 4 različite ekonomske varijable, te smo preko njih pokušali predvidjeti kretanje indeksa S&P500. Primjer takvog istraživanja napravljeno je 2015. godine gdje je grupa istraživača, preko algoritama regresije, SVM-a, Neuronskih mreža, prognoziralo kretanje indeksa S&P 500 pomoću 27 ekonomskih varijabli kao što su cijene unce zlata, barela nafte, cijena nekoliko važnih dionica, odnos nekoliko valutnih parova i brojčanog stanja za nekoliko globalnih indeksa. U radu su koristili 143 zapisa podataka od kojih je 100 bilo za treniranje a 43 za testiranje modela, Rezultati modela nisu iskazani brojčano, nego grafički kojih prikazuju da su pogodili svaki trend, a najbolje rezultate dao je SVM model [27].

Osim ovog istraživanja grupa znanstvenika iz Japana u istraživačkom radu 2005. godine je preko SVM algoritma je pokušalo predvidjeti glavni japanski globalni indeks NIKKEI 250, a ulazne varijable koristili su indeks S&P 500 i tečaj japanskog jena. Rezultati su dali 73% točnih predikcija [28].

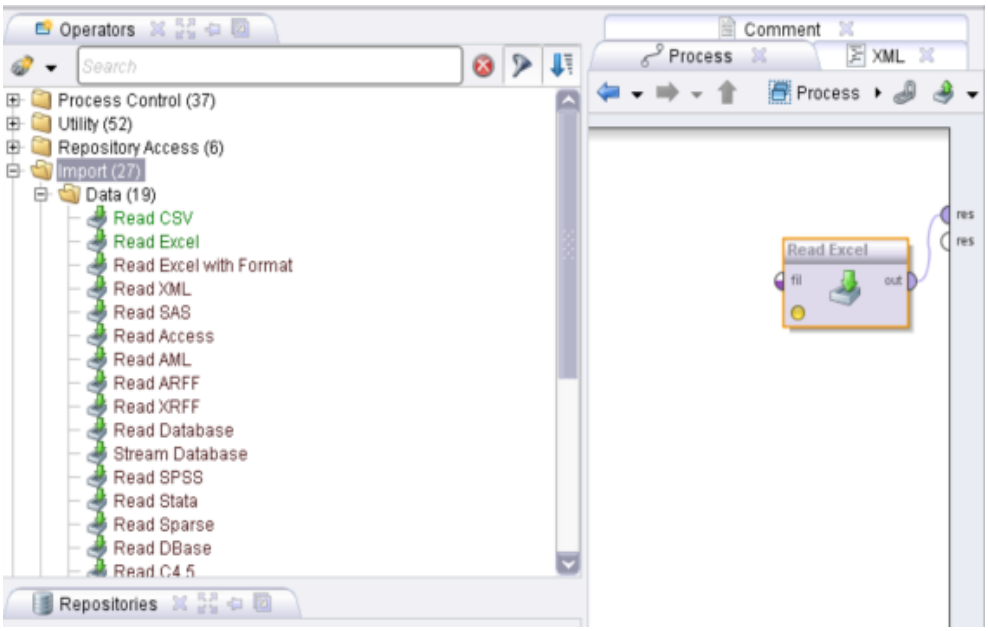
4.1.2. Priprema podataka

Nakon sakupljanja podataka, slijedi priprema podataka gdje ćemo prvo spojiti potrebne podatke u jednu Excel tablicu. U ovom radu koristit ćemo 2 skupa podataka odnosno dvije tablice, prvi skup podatak služi za učenje tj. trening modela, dok drugi skup podataka služi za testiranje. Prvi skup sadrži podatke u razdoblju od 23.08.2018 do 09.08.2019 godine, znači nešto manje od godine dana, a budući da vikendom burze ne rade, prvi skup se sastoji od 232 zapisa. Drugi skup se sastoji od podataka u razdoblju od 12.08.2019 do 30.08.2019 te se u njemu nalazi 15 zapisa. Zbog preskakanja datuma, datume ćemo pretvoriti u cijele brojeve radi lakšeg snalaženja. Uz pomoć čarobnjaka za uvoz podataka na slici 8., označit ćemo tipove podataka u prvom skupu, a atribut S&P 500 ćemo označiti kao ciljnu varijablu oznakom „label“.



Slika 8. Čarobnjak za uvoz podataka

Nakon uvoza podataka, podatke trebamo učitati u proces. Prvo ćemo u tražilicu operatera odabrati „*Read Excel*“ i zatim ga spojiti na izlaz.



Slika 9. Učitavanje podataka u proces

Nakon učitavanja podataka, pokrenut ćemo proces koji će nam dati izvještaj s našim podacima, kako su strukturirani i označeni.

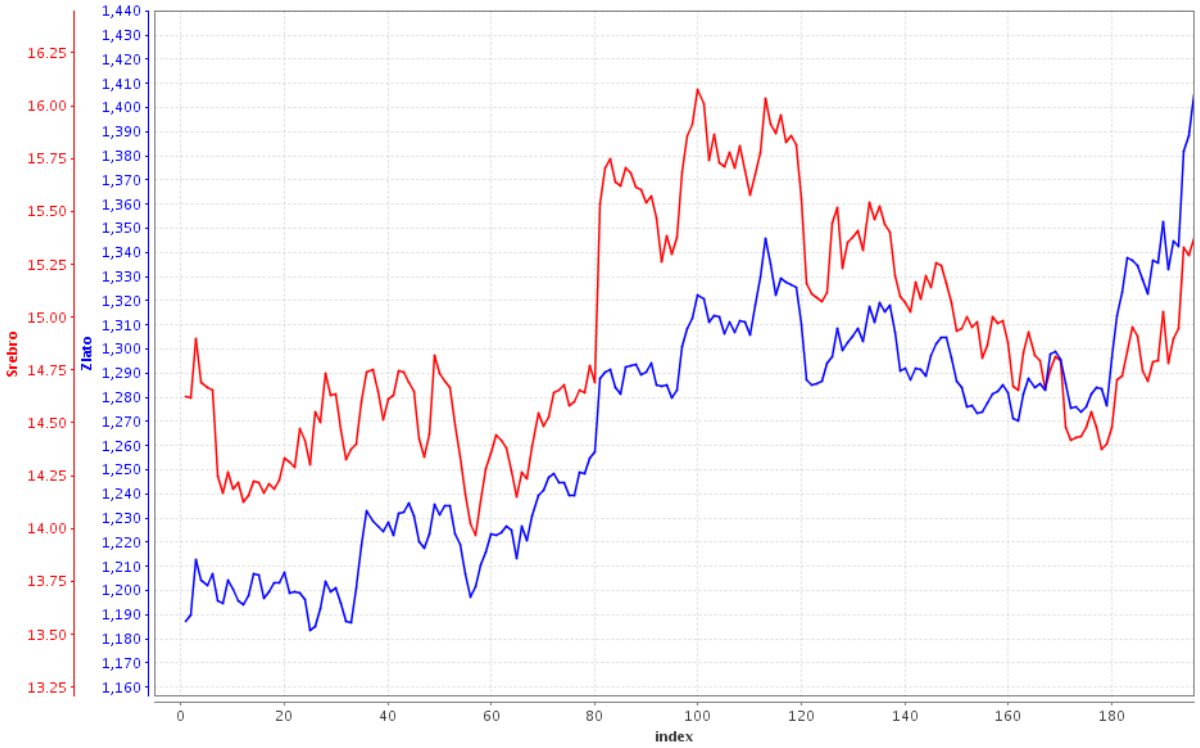
Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (232 examples, 1 special attribute, 3 regular attributes)

Row No.	SP500	Zlato	Srebro	Nafta
1	2856.980	1187.300	14.625	69.130
2	2874.690	1189.950	14.620	69.710
3	2897.520	1212.750	14.900	68.540
4	2914.040	1204.300	14.690	69.680
5	2901.130	1202.350	14.665	70.250
6	2901.520	1206.850	14.655	69.840
7	2896.720	1195.750	14.250	69.820
8	2888.600	1194.700	14.170	68.690
9	2878.050	1204.300	14.265	67.810
10	2871.680	1200.750	14.185	67.730
11	2877.130	1195.800	14.215	67.550
12	2887.890	1194	14.125	69.290
13	2888.920	1197.800	14.155	70.370
14	2904.180	1206.650	14.225	68.600
15	2904.980	1206.200	14.215	68.980
16	2888.800	1196.800	14.170	68.860
17	2904.310	1199.400	14.210	69.870
18	2907.950	1203	14.185	71.080
19	2930.750	1203	14.230	70.770
20	2929.670	1207.600	14.330	70.800

Slika 10. Prikaz podatka u izvještaju

Ako kliknemo na opciju „*Plot View*“, možemo vidjeti naše podatke u puno vizualnih oblika, ako izaberemo oblik grafa „*Series Multiple*“, možemo vidjeti međusoban odnos atributa u skaliranom obliku. Pomoću tog prikaza iz takvih grafova možemo primijetiti nekoliko trendova u njihovim odnosima kroz vrijeme, odnosno dane. Primjer takvog trenda je odnos zlata i srebra, na slici 11. gdje možemo vidjeti da zlato i srebro kad im se skaliraju vrijednosti, imaju gotovo sličan trend, što bi značilo da je njihova veza u većini slučajeva proporcionalna.

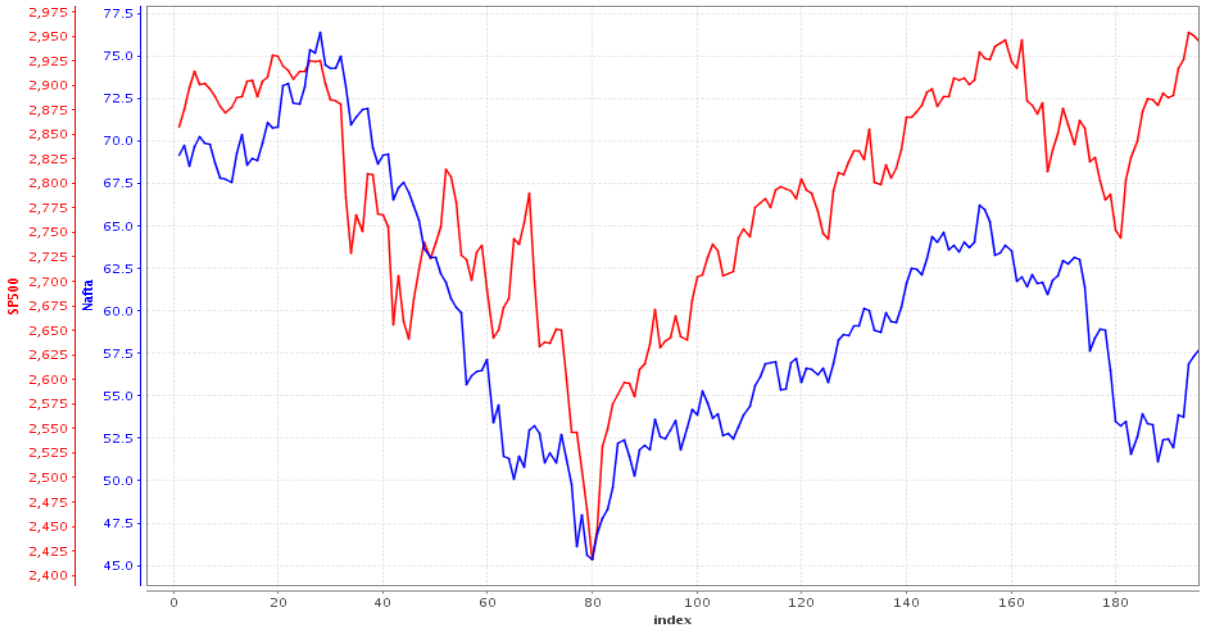


Slika 11. Vremenski prikaz vrijednosti atributa zlata i srebra

Nakon ovog trenda, na slici 12. možemo vidjeti i povezanost između globalnog indeksa S&P500 i zlata, koji je u najčešćem slučaju obrnuto proporcionalan s indeksom.

Slika 12. Vremenski prikaz vrijednosti atributa zlata i S&P500 indeksa

Osim ova 2 uočena trenda, postoji i povezanost između globalnog indeksa S&P500 i cijene barela nafte, a njihov je odnos sličan kao i odnos srebra i zlata.

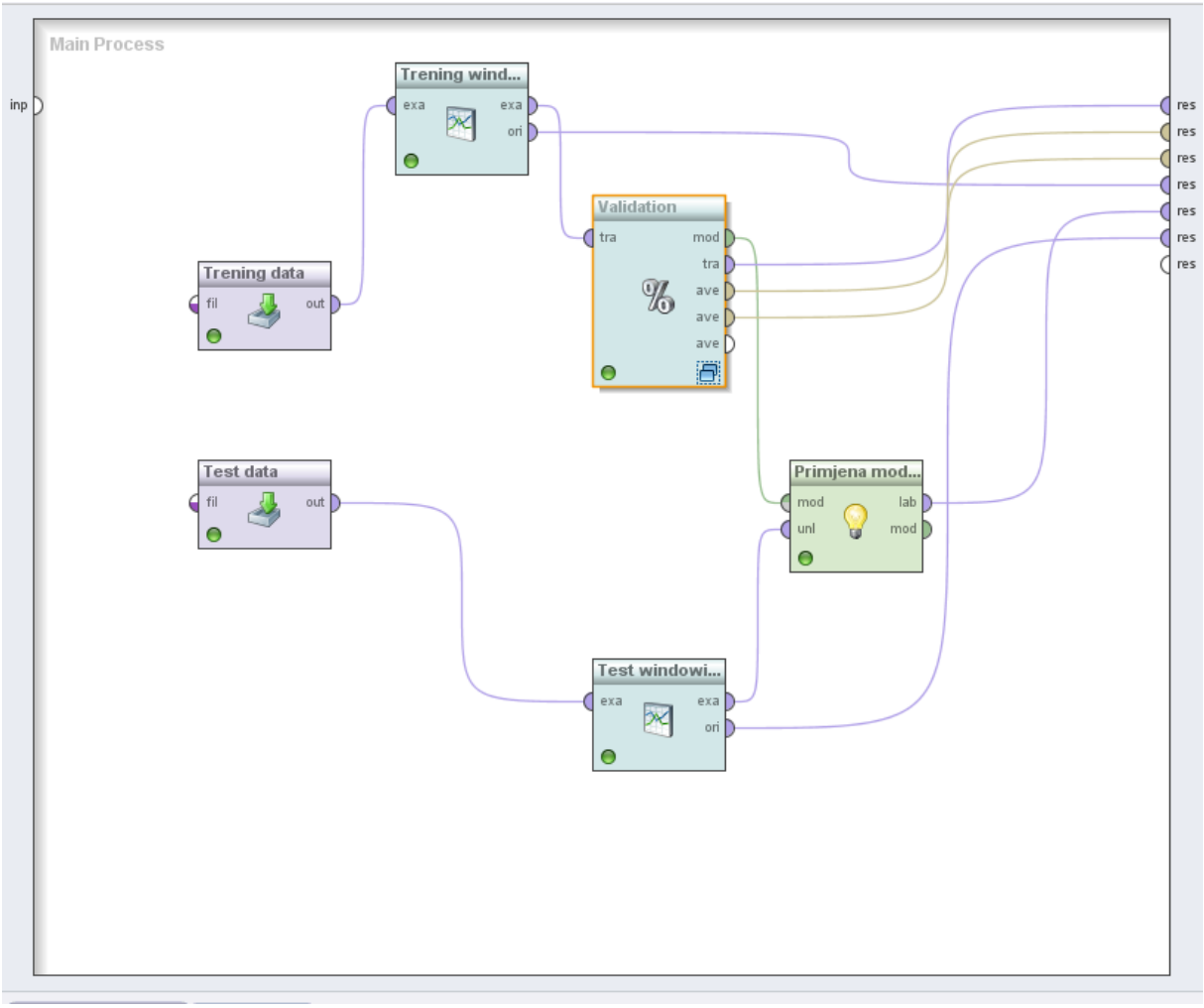


Slika 13. Vremenski prikaz vrijednosti atributa zlata i S&P500 indeksa

Uočavanjem svih ovih trendova možemo zaključiti da je istina da su kretanja različitih tržišta međuovisna, ovi grafovi nam impliciraju da će u slučaju porasta cijena zlata i srebra indeks S&P500 padati, a da će se padom indeksa cijena nafte će se smanjiti.

4.2. Modeliranje

Nakon uspješnog sakupljanja, pripremanja i razumijevanja podataka, možemo početi modelirati. Naš model se sastoji od nekoliko operatora od kojih za većinu treba postaviti parametre, da bi rezultati bili realni i točni.



Slika 14. Glavni proces modela

Na slici 14. možemo vidjeti operatore koje smo koristili pri modeliranju. Od operatora smo koristili:

1. *Read Excel* - ovaj operator nam služi za učitavanje skupova podataka: gornji operator služi za učitavanje podataka za trening, a donji za učitavanje podataka za testiranje.

2. *Windowing* - ovaj operator koristimo kada imamo više-varijabilne zapise u određenom periodu, te pomoću njega dobivamo novi vremenski model u kojem možemo predviđati određeni broj dana unaprijed. U ovom operatoru postoje važni parametri koje trebamo postaviti:

- Moramo odrediti na koji način operator učitava zapise podataka, mi ćemo postaviti da ih učitava preko redova.
- „*Horizon*“, odnosno koliko dana unaprijed želimo predvidjeti, ovu vrijednost ćemo postaviti na 1, odnosno želimo znati jedan dan unaprijed.
- Veličinu prozora, koliko će zapisa operator čitati paralelno, tu ćemo postaviti 1, odnosno operator će „snimati“ zapise dan po dan.
- Veličinu koraka, koja nam predstavlja koliko će redova pomaknuti u novom modelu, isto ćemo postaviti na 1.
- Moramo odrediti ciljnu varijablu koju želimo predvidjeti, u našem slučaju to će biti S&P 500 indeks

Ovaj operator uzima početni skup podataka i ciljnu varijablu, te radi novi skup podataka koji je umanjen za n predviđenih dana , a vrijednost ciljne varijable premješta za jedan red gore. Učinak ovog operatera najlakše je uočiti preko slike 15., na kojoj je novi skup podataka lijevo, a početni skup podataka desno. Na slici 16. vidimo da se taj operator pojavljuje 2 puta odnosno za svaki skup podataka, razlika je da u skupu podataka za trening, ne označujemo ciljnu varijablu, jer smo nju označili u prvom operatoru, ostale parametre postavimo jednako.

Data View | Meta Data View | Plot View | Advanced Char

ExampleSet (231 examples, 1 special attribute, 3 regular attributes)

Row No.	label	Zlato-0	Srebro-0	Nafta-0
1	2874.690	1187.300	14.625	69.130
2	2897.520	1189.950	14.620	69.710
3	2914.040	1212.750	14.900	68.540
4	2901.130	1204.300	14.690	69.680
5	2901.520	1202.350	14.665	70.250
6	2896.720	1206.850	14.655	69.840
7	2888.600	1195.750	14.250	69.820

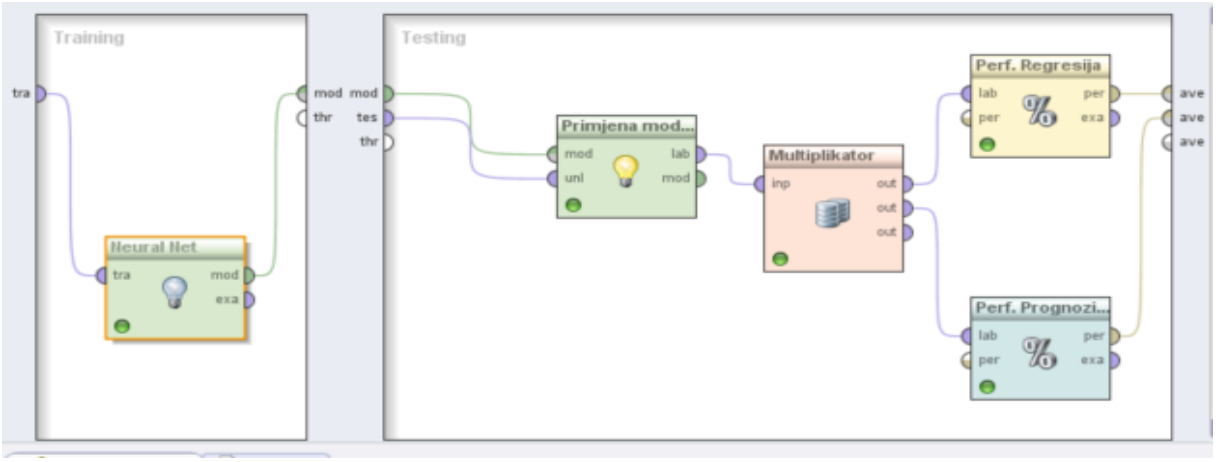
Data View | Meta Data View | Plot View | Advanced Charts

ExampleSet (232 examples, 1 special attribute, 3 regular attributes)

Row No.	SP500	Zlato	Srebro	Nafta
1	2856.980	1187.300	14.625	69.130
2	2874.690	1189.950	14.620	69.710
3	2897.520	1212.750	14.900	68.540
4	2914.040	1204.300	14.690	69.680
5	2901.130	1202.350	14.665	70.250
6	2901.520	1206.850	14.655	69.840
7	2896.720	1195.750	14.250	69.820

Slika 15. Razlike između skupova podataka

3. *Validation* – ovo je posebni oblik operatora koji služi za potvrđivanje modela, taj oblik se zove i ugniježđeni operator jer se u njemu nalazi više operatora.



Slika 16. Struktura operatora potvrđivanja modela

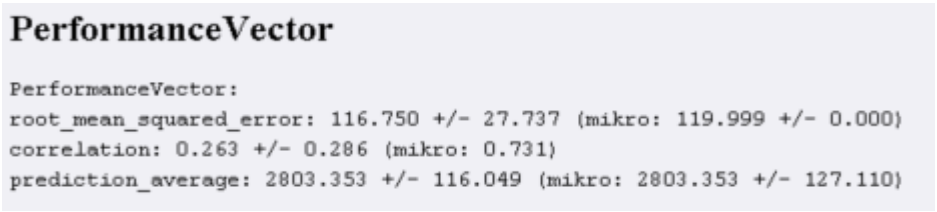
Na slici možemo vidjeti strukturu ovog operatora. On se sastoji od 2 dijela: prvi dio služi za treniranje modela, a drugi za testiranje, odnosno u prvom dijelu se nalazi algoritam preko kojeg ćemo rudariti podatke (u ovom modelu koristili smo algoritam Neuronske mreže), a u drugom dijelu nalaze se sljedeći operatori:

- Primjena modela (eng. *Apply Model*) je operator koji nam služi za učenje modela, on pamti informacije koje je „naučio“ na danom skupu podatka, te to „znanje“ može prenositi na druge operatore.
- Multiplikator služi za kopiranje ulazne informacije i šalje ih na 2 ili više izlaza.
- Operatori performansi: ovi operatori nakon izvršenog procesa rudarenja podataka, daju informacije o točnosti i preciznosti modela. Više o rezultatima ovog operatora bit će u sljedećem poglavlju.

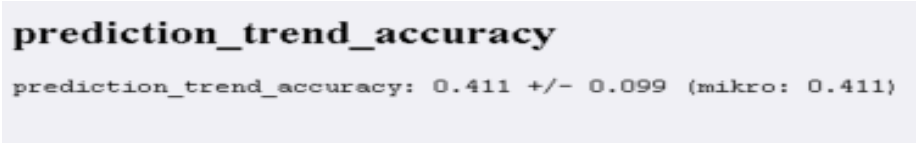
4.3. Rezultati i interpretacija

Nakon modeliranja slijedi najzanimljiviji dio cijelog procesa, a to je pregled rezultata. Pokretanjem programa dobijemo sve rezultate koje smo spojili na izlaz. Najzanimljiviji su nam izvještaji predikcija i performanse modela.

Prvo ćemo prokomentirati rezultate performansi. U modelu možemo vidjeti da imamo 2 operatora, prvi operator služi za performanse algoritma regresije u kojem smo označili da želimo dobiti podatke o: kvadratnoj pogreški, korelaciji i prosjeku predviđanja, dok nam drugi operator služi da nam ocijeni postotak predviđanja trenda. Važno je napomenuti da su ti rezultati za tzv. trening podatke, te da predikcija ne mora biti točna i precizna kako nam ove vrijednosti govore.



Slika 17. Performanse regresije



Slika 18. Performanse trenda predviđanja

Rezultati kvadratne pogreške govore nam koliko je moguća najveća greška u regresijskom pravcu, rezultat od 116.750+/- 27.737, nije velika pogreška budući da se naš indeks kreće oko 2500 do 3200, ali u donošenju odluka to može dosta utjecati. Rezultat korelacije između podataka je dosta mali, ali to je bilo i za očekivati, jer su atributi brojčano dosta različiti, npr. cijena srebra je prosječno 17 dolara po unci, a zlata 1500, a još imamo i dvije varijable koje se isto mnogo razlikuju. Prosjek predviđanja nam govori koliki je raspon mogućih vrijednosti predviđenog indeksa. Prognoziranje trenda ovdje je najvažniji podatak, koji nam govori da je model samo u 41 % slučajeva pogodio hoće li cijena rasti ili padati, što je dosta loš rezultat, ali opet ćemo naglasiti činjenicu da je to za skup podataka koji služi za učenje.

Nakon prikaza performansi, možemo pogledati i same rezultate predviđanja. Već smo naglasili da nećemo gledati brojčane rezultate nego ćemo se fokusirati na to je li model prognozirao pad ili rast indeksa.

ExampleSet (15 examples, 1 special attribute, 4 regular attributes)

Row No.	prediction(la...	SP500-0	Zlato-0	Srebro-0	Nafta-0
1	2920.057	2888.200	1522.800	17.200	54.720
2	2921.611	2926.320	1511.800	17.090	56.790
3	2918.426	2840.600	1527.900	17.350	54.900
4	2919.734	2847.600	1534.500	17.380	54.720
5	2919.773	2888.680	1523.600	17.230	54.850
6	2920.671	2923.650	1505.600	17	56
7	2918.570	2900.510	1516.700	17.250	56.100
8	2916.861	2924.430	1512.400	17.250	55.980
9	2916.883	2922.950	1507.200	17.140	55.540
10	2915.484	2847.410	1536.900	17.550	53.970
11	2910.442	2878.380	1538.100	17.770	53.790
12	2908.087	2869.160	1553.300	18.310	55.620
13	2904.439	2887.940	1548	18.440	55.850
14	2903.585	2924.580	1536.900	18.350	56.560
15	2892.915	2926.460	1529.200	18.480	55.160

Slika 19. Rezultati predviđanja

Slika 20. prikazuje izvještaj prognoziranja podataka, prvi stupac predstavlja broj dana, drugi stupac predstavlja predviđene vrijednosti za globalni indeks S&P 500 za 1 dan unaprijed, dok treći stupac prikazuje stvarne podatke kako su se indeksi kretali kroz dane.

Dan	Prognoziranje	Stvarnost
1	Rast	-
2	Pad	Rast
3	Rast	Pad
4	Rast	Rast
5	Rast	Rast
6	Pad	Rast
7	Rast	Pad
8	Pad	Rast
9	Pad	Pad
10	Rast	Pad
11	Pad	Rast
12	Rast	Pad
13	Rast	Rast
14	Pad	Rast
15		Rast

Tablica 1. Prognoza trendova

Tablica 1. prikazuje kako je naš model pogodio impresivnih 13 puta od 14 pokušaja, to je čak 93% točnosti u gađanju trendova, a rezultati performanse na temelju podatka učenja su nam dali samo 41% šanse točne procjene trenda.

Ako usporedimo naše rezultate s rezultatima iz istraživanja koji su navedeni u poglavlju 4.1.1., možemo zaključiti da naš model može konkurirati s modelima koji su opisani, a postotak točnosti je veći od većine modela. Možemo zaključiti da su rezultati slični istraživanju iz 2015. godine, ali to istraživanje je puno kompleksnije i složenije napravljeno.

5. Zaključak

U ovom završnom radu prikazali smo kako pomoću procesa rudarenja podataka možemo dobiti predikciju kretanja globalnog indeksa S&P 500, te smo dokazali njegovu međuovisnost s različitim tržištima, kao što su tržište nafte, zlata i srebra. Rezultati su pokazali da se ne može točno procijeniti kolika će biti točna vrijednost pojedinog indexa dionice, ali se može odrediti hoće li cijena rasti ili padati kroz određeno vrijeme. Rudarenje podataka je relativno nova disciplina, a u sljedećih nekoliko godina bi trebala još više jačati kao znanost o otkrivanju znanja iz podataka.

Praktični dio ovog rada izrađen je u alatu Rapid Miner 5.3., ovaj software ima mogućnosti da u njemu napravimo bilo kakav model za predikciju ili klasteriranje podataka, također ima karakteristike po kojima možemo vidjeti koliko je naš model točan i siguran u predviđanju, te karakteristike nam puno olakšavaju pri donošenju poslovnih odluka.

Rezultati u ovom istraživanju su ispali jako dobri. Postotak točnosti predviđanja trenda s podacima za trening nam nije dao obećavajuće šanse da će testni rezultati biti uspješni, ali na kraju s 93% točnosti možemo biti zadovoljni i zaključiti da s ovim modelom možemo previđati trendove.

S velikom sigurnošću možemo zaključiti da poduzetnici koji se bave kupnjom i prodajom dionica na burzama, s ovakvim i sličnim modelima mogu okvirno znati hoće li sljedeći dan cijena rasti ili padati. Ova hipoteza implicira da strojno učenje, odnosno umjetna inteligencija u ovom slučaju može zamijeniti čovjeka.

6. Popis literature

1. Klačmer Čalopa M., Cingula M.; *Financijske institucije i tržište kapitala*, FOI, Varaždin, 2009.

2. *„Primarno tržište“*, Poslovni dnevnik

- http://www.poslovni.hr/leksikon/primarno-trziste-1665
- Preuzeto 15.08.2019.

3. *„Izvanburzovno tržište (OTC)“*, Capital.com

- https://capital.com/hr/izvanburzovno-trziste-otc--definicija
- Preuzeto 15.08.2019.

4. *„Burza“*, Poslovni dnevnik

- http://www.poslovni.hr/leksikon/burza-48
- Preuzeto 15.08.2019.

5. C. Mitchell (2019.) *„Forex & Currencies Trading“*; Investopedia

- https://www.investopedia.com/terms/f/forex.aspdatum
- Preuzeto 17.08.2019.

6. *„Freight exchange“*, Haulage exchange

- https://haulageexchange.co.uk/freight-exchange
- Preuzeto 16.08.2019.

7. *„Dionice“*, Poslovni dnevnik

- http://www.poslovni.hr/leksikon/dionice-277
- Preuzeto 17.08.2019.

8. *„Indeks CROBEX“*; Zagrebačka burza

- https://www.zse.hr/default.aspx?id=44101&index=CROBEX
- Preuzeto 19.08.2019.

9. *„Burzovni indeks“*, Capital.com

- https://capital.com/hr/burzovni-indeks-definicija
- Preuzeto 19.08.2019.

10. Harper, D.R. (2019.) *„Forces That Move Stock Pices“*; Investopedia

- https://www.investopedia.com/articles/basics/04/100804.asp
- Preuzeto 20.08.2019.

11. Rouse, M. *„Data mining“*; SearchSQLServer

- https://searchsqlserver.techtarget.com/definition/data-mining
- Preuzeto 20.08.2019.

12. *„The Difference Between a Data Warehouse and a Database“*; Data Warehouse Guide

- https://panoply.io/data-warehouse-guide/the-difference-between-a-database-and-a-data-warehouse/
- Preuzeto 20.08.2019.

13. Fayyad, U., Piatetsky-Shapiro, G. i Smyth, P. (1996). *„From Data Mining to Knowledge Discovery in Databases“*; Calif.: American Association for Artificial Intelligence

14. Azevado, A. i Santos, M.F. (2008). *„KDD, SEMMA i CRISP-DM: A parallel overview“*. IADIS

15. Poll: *„What main methodology are you using for your analytics, data mining, or data science projects?“*; KDnuggets

- https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html
- Preuzeto 22.08.2019.

16. Fayyad, U., Piatetsky-Shapiro, G. i Smyth, P. (1996.) *„Six steps in CRISP-DM the standard data mining process" – „AI Magasine"*; Calif.: American Association for Artificial Intelligence

17. *„Introduction to the CRISP DM data mining methodology - webinar recording"*; Youtube

- https://www.youtube.com/watch?v=eAcxe3MzcVk&t=2124s
- Preuzeto 23.08.2019.

18. Al-Masri, A. (2019.) *„What Are Supervised and Unsupervised Learning in Machine Learning?"*; Towards data science

- https://towardsdatascience.com/what-are-supervised-and-unsupervised-learning-in-machine-learning-dc76bd67795d
- Preuzeto 24.08.2019.

19. Garg, R. (2018.) *„7 Types of Classification Algorithms"*;Analytics India

- https://www.analyticsindiamag.com/7-types-classification-algorithms/
- Preuzeto 24.08.2019.

20. *„Data Mining Concepts, Regression"*. (2008). Oracle

- https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/regress.htm#CHDBHBDI
- Preuzeto 24.08.2019.

21. *„Data Mining Concepts, Market Basket"*. (2008). Oracle

- https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/market_basket.htm#BABDIIFF
- Preuzeto 25.08.2019.

22. Seif G. *„The 5 Clustering Algorithms Data Scientsts Need to Know"*; Towards Data Science

- https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
- Preuzeto 24.08.2019

23. Ramesh V., Parkavi P. i Ramar.K (2013.) *„Predicting Student Performance:A Statistical and Data MiningApproach“*; International Journal of Computer Applications

24. Durairaj M. i Ranjani V. (2013.) *„Data Mining Applications In Healthcare Sector: A Study“*; International Journal of Scientific & Technology Research

25. „*Upute*“; FOI radovi

- https://radovi.foi.hr/
- Preuzeto 15.8.2019

26. E.L. de Fariaa, Marcelo P. Albuquerquea, J.L. Gonzalez, J.T.P. Cavalcantea i Marcio P. Albuquerque (2009.) *„Predicting the Brazilian stock market through neural networks and adaptiveexponential smoothing methods“*; Elsevier Ltd.

27. Sheta A., Ahmed S.E. i Faris H (2015.) *„A Comparison between Regression, ArtificialNeural Networks and Support Vector Machines forPredicting Stock Market Indeks“*; International Journal of Advanced Research in Artificial Intelligence.

28. Huang W., Nakamori Y. i Wang S.Y. (2005.) *„Forecasting stock market movement direction withsupport vector machine“*; Computers & Operations Research

7. Popis slika