UNIVERZA V LJUBLJANI

FAKULTETA ZA MATEMATIKO IN FIZIKO

Interdisciplinarni doktorski študij statistike

Matematična statistika - 3. stopnja

Marija Gorenc Novak

IZGRADNJA IN UPORABA KLASIFIKATORJEV V FINANČNI MATEMATIKI

Doktorska disertacija

MENTOR: dr. Dejan Velušček

SOMENTOR: prof. dr. Matjaž Omladič

Ljubljana, 2015

Izjava

Podpisana Marija Gorenc Novak izjavljam:

- da sem doktorsko disertacijo z naslovom *Izgradnja in uporaba klasifikatorjev v finančni matematiki* izdelala samostojno pod mentorstvom dr. Dejana Veluščka in somentorstvom prof. dr. Matjaža Omladiča
- da Fakulteti za matematiko in fiziko Univerze v Ljubljani dovoljujem objavo elektronske oblike svojega dela na spletnih straneh.

Ljubljana, september 2015 Podpis: . . . . . .

Zahvala

Ob zaključku pisanja doktorske disertacije bi se rada zahvalila veliko ljudem, ki so mi stali ob strani in prispevali na moji poti do doktorata. Seznam je obsežen in na tem mestu ne bi vseh omenjala. Izpostavila bi pa nekaj ključnih posameznikov.

Na prvem mestu velika iskrena zahvala mojemu mentorju, prof. dr. Dejanu Veluščku za potrpežljivost in vztrajnost tekom raziskovanja. Hvala za strokovne nasvete in vodstvo pri iskanju rešitev in rezultatov zaradi katerih je delo postalo še kvalitetnejše.

Hvala tudi mojemu somentorju, prof. dr. Matjažu Omladiču, ki mi je omogočil okolje za študij.

Zahvaljujem se podjetju XLAB d.o.o. in direktorju dr. Gregorju Pipanu, ki sta mi omogočila študij v prijetnem in vzpodbujajočem okolju ter vsem sodelavcem za podporo. Še posebna zahvala gre dr. Danielu Vladušiču ter dr. Dragu Bokalu za vodenje in koristne nasvete. Hvala vsem mladim raziskovalcem, s katerimi smo se tekom mojega študija skupaj prebijali na poti do cilja.

Hvala dr. Borisu Cergolu, ki me je s svojimi izkušnjami usmerjal in tako veliko prispeval k nastanku tega dela.

Zahvala gre tudi moji družini, ki me je podpirala in se veselila ob mojih uspehih.

Še posebej velika hvala pa možu, Gregorju Novaku, ki je budno spremljal moje napredke in me bodril ob težjih trenutkih ter veliko prostega časa namenil mojemu raziskovanju. Vsi ti trenutki so imeli zame neprecenljivo vrednost. Velik del te poti bi bil veliko težji, če ne bi bilo tvojega razumevanja.

Hvala Evropskemu socialnemu skladu za delno financiranje.

Povzetek

V doktorski disertaciji je predstavljen postopek izgradnje klasifikacijskih modelov ter njihova uporaba na finančnih podatkih. Za razliko od večine raziskovalnih del na tem področju, ki so se trudili napovedati smer gibanja delnic na zaključnih dnevnih tečajih, smo v tem delu napovedovali smer gibanja delnic na najvišjih tečajih, torej napovedovali smo, ali bo smer najvišjih dnevnih tečajev v naslednjem dnevu višja ali nižja. Razlog za to odločitev je precej manjša volatilnost na najvišjih in ravno tako na najnižjih dnevnih tečajih. V disertaciji smo z izgradnjo klasifikatorjev želeli preveriti, ali z zgrajenimi modeli dobimo dovolj zanesljive napovedi gibanja delniškega trga (smeri najvišjih dnevnih tečajev) na 370 delnicah, ki so članice indeksa $S\&P500$ in te napovedi želeli uporabiti kot podporo pri odločanju pri trgovanju z delnicami. Za vhodne podatke smo uporabili tehnične indikatorje (98 tehničnih indikatorjev), ki so izračunani iz preteklih podatkov na časovnih vrstah in pri tem bili pozorni, kateri so tisti indikatorji, ki največ prispevajo k uspešnosti napovedi. V tem delu so predstavljene multivariatne filtrirne metode za izbor relevantnih tehničnih indikatorjev ter nova filtrirna metoda 'FSuC–ward–comb', kateri dobljeni tehnični indikatorji največ prispevajo k napovedi klasifikacijskega modela. Dobljeni tehnični indikatorji sugerirajo, da so pri sestavi modela najbolj pomembni tisti indikatorji, ki ne vključujejo daljnih podatkov. Klasifikacijski rezultati presežejo 60% klasifikacijske točnosti na testnih podatkih, kar je dovolj za izgradnjo profitabilnih trgovalnih strategij, zato smo klasifikacijske napovedi vključili v posebej konstruirane trgovalne strategije in jih primerjali s trgovalnimi strategijami brez vključitve napovedi. Izkazalo se je, da je uspešnost trgovalnih strategij, kjer smo vključevali napovedi, v splošnem višja, kot če ne vključujemo napovedi. Predlagane strategije presežejo tudi indeks $S\&P500$.

Math. Subj. Class.(2010): 97K80, 97M40, 97M10
JEL: C10, C38, C45, C53, G11

Ključne besede: klasifikacijski modeli, metode za izbor atributov, filtrirne metode, trgovalna strategija, strojno učenje

Abstract

In the thesis, the design and application of classifiers on financial data is presented. We focus on a prediction based on daily high prices instead of commonly used daily close prices. The reason for that decision is that daily close prices are more volatile than daily high/low prices. Due to this reason, we decided to classify the daily high prices in order to forecast whether the daily high price will rise or fall. We examined if the predictions of daily high returns movement using statistical classifiers give good performance on a large part of stocks from the $S\&P500$ index. We include only technical indicators as an input data, i.e. 98 technical indicators, which were calculated on a lagged time series (volume, open, high, low, close and adjusted values). As there are many algorithms for feature selection, we focused on multivariate filter methods and proposed novel 'FSuC–ward–comb' filter method, which chosen attributes gave the highest classification results. We also analyze, which are the most relevant attributes that contain the most useful information for prediction of the future daily high movement. We obtained technical indicators with shorter indicator lengths as the more relevant attributes, which are also more suitable for short–term trading (e.g. daily). The classification results on testing data set exceed 60%, which is enough to make economically profitable strategies. Comparing the strategies in which we use classification predictions with strategies without any use of classifiers showed that classifiers can significantly increase performance, which demonstrates the benefits of integrating classifiers in the strategies. The trading experiments show that the proposed strategies also outperform $S\&P500$ index.

Math. Subj. Class.(2010): 97K80, 97M40, 97M10
JEL: C10, C38, C45, C53, G11

Keywords: classification models, feature selection, filter methods, trading strategy, machine learning

Kazalo

1 UVOD

1.1 Klasifikacijski modeli in izbor atributov

V sodobnem času so IKT (informacijsko–komunikacijske tehnologije) naplavile množico podatkov. V tej kopici podatkov lahko izgubimo pregled. Veliko metod, ki odkrivajo in opisujejo vzorce v podatkih je bilo odkritih znotraj področja, ki se mu reče strojno učenje. Znotraj področja strojnega učenja obstajajo številne metode, v tem delu pa se bomo ukvarjali s klasifikacijskimi metodami, s katerimi bomo poskušali napovedati smer dnevnega gibanja delnic. Smer gibanja delnic v tem kontekstu pomeni, ali bo vrednost delnice v dnevu $t + 1$ presegla vrednost delnice v dnevu t ali ne.

Klasifikacijski modeli se dobro obnesejo pri napovedih gibanja delniškega trga, kjer se napoveduje le pozitiven in negativen donos [16, 40, 41, 45, 48, 60], ne obnesejo pa se pri napovedih relativnih donosov, saj je napoved gibanja delniškega trga preveč odvisna od slučajnih komponent, kot so na primer: situacija v podjetjih, politika, globalna ekonomija, pričakovanja trgovalcev itd.

Osnovni namen doktorske disertacije je zgraditi in poiskati primerne klasifikacijske modele na finančnih podatkih. Na podlagi pridobljenih zgodovinskih podatkih o gibanju cen delnic, ki so ali so bile članice ameriškega indeksa Standard & Poors 500 ($S\&P500$), smo želeli čim natančneje določiti smer njihovega gibanja. S tem delom smo želeli preveriti, ali se da delniški trg napovedati zgolj s statističnega vidika, torej ali med gibanjem delniških tečajev v preteklosti in med tistim, kar se bo zgodilo v prihodnosti obstaja direktna povezava.

V literaturi se omenjajo pomembni faktorji, ki vplivajo na vrednosti delnic, ni pa jasnega odgovora na vprašanje, kateri faktorji imajo največjo napovedno moč. Z uporabo različnih atributov se napovedni modeli različno obnašajo, zato je izgradnja optimalnega napovednega modela težka naloga. Za dober model je potrebno podati relevantne atribute, to je atribute, ki vsebujejo kar se da največ informacij o prihodnjem gibanju delnice. Atribute v našem delu predstavljajo tehnični indikatorji, ki so skupek formul na zajetih preteklih podatkih. V tem delu nas zanimajo predvsem tisti tehnični indikatorji, ki povedo največ informacij o gibanju delniškega trga in te atribute uporabimo kot osnovo za izgradnjo finančnih modelov.

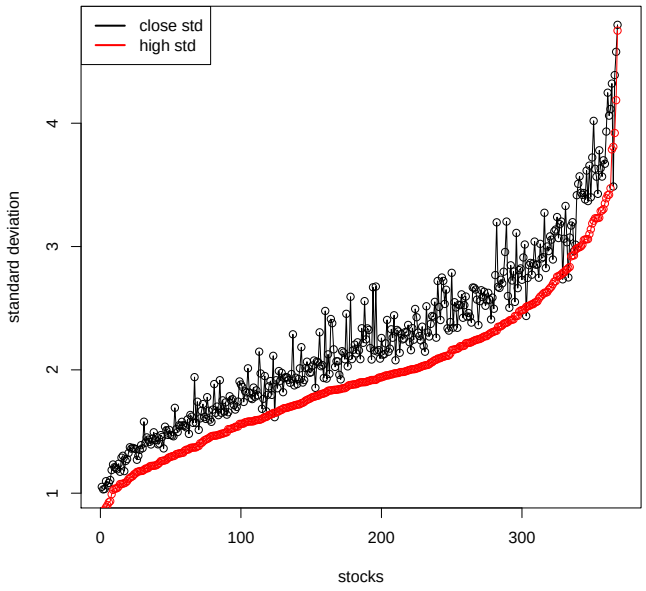
V dosedanjih raziskavah so raziskovalci s pomočjo metod za izbor atributov raziskovali uporabnost izbranih atributov za uspešno napoved gibanja cen delnic. Z izborom atributov so ohranili le tiste, ki so dali visoko napovedno moč, kar zmanjša razsežnost prostora in posledično izboljša dejanski čas izračunov algoritmov, izognili so se prekomernemu prilagajanju podatkov (ang. 'overfitting') in tako dosegli boljše posploševanje in lažjo interpretacijo rezultatov [33, 89]. Dosedanja dela na finančnem področju, ki vključujejo metode za izbor atributov, običajno obsegajo le nekaj atributov. Avtorja Atsalakis in Valavanis [8] sta naredila pregled nad 100 znanstvenimi članki, ki vključujejo analizo delniških trgov. Število indikatorjev, ki so jih avtorji vključili v raziskavo se giblje med 4 in 10. Večina jih je vključila le po en tip indikatorjev (bodisi tehnični tip indikatorjev bodisi fundamentalni tip indikatorjev) s katerimi so poskušali napovedati dnevno ali tedensko gibanje indeksa [24, 40, 43, 48, 51, 59, 60, 99]. V našem

raziskovalnem delu smo se ravno tako osredotočili na en tip: na tehnične indikatorje. Znanstveni članki pričajo o tem, da so za dnevne napovedi primerni tehnični indikatorji [90].
Za ocenjevanje informativnosti atributov obstaja več metod. Najenostavnejše so filtrirne metode, ki so računsko enostavne in hitre metode [77], ki z ocenjevalnimi metrikami ocenijo vsak atribut neodvisno od klasifikacijske metode napovedovanja [98]. V našem delu smo se zato omejili le na te metode. Filtrirne metode lahko razdelimo v 2 skupini: v skupino atributnega razvrščanja (univariatne filtrirne metode), ki ocenjuje atribute individualno in multivariatno izbiranje atributov (ang. 'subset feature selection'), ki ocenjuje podmnožico atributov. Ocenjevanje na podmnožici atributov prinese veliko prednosti v primerjavi z atributnim razvrščanjem, predvsem pri doseganju višje napovedne moči. Pri velikem deležu multivariatnih filtrirnih metod je vedno potrebno vnaprej določiti število atributov, ki naj jih vrne algoritem za izbor atributov.
V doktorski disertaciji smo izbirali podmnožico atributov s pomočjo multivariatnih filtrirnih metod. Predlagali smo svoj algoritem 'FSuC' (ang. 'Feature Selection using Clustering') [70], katerega ideja je izbor podmnožice atributov s pomočjo metod za razvrščanje v skupine. Ta razvrsti enote v skupine ne glede na to, kakšne so vrednosti razredov (katera informacija le teh nam je znana). Število skupin k je enako, kot je število klasifikacijskih razredov, zato ni potrebno določati optimalnega števila skupin za razvrščanje v skupine. Algoritem rekurzivno gradi množico relevantnih atributov, dokler ne zadosti zaustavitvenem pogoju. Na vsakem koraku na trenutni podmnožici izbranih atributov poskuša razvrstiti enote v k skupin tako, da so razvrščene skupine kar se da podobne skupinam, ki jih inducirajo vrednosti razredov (dan kot apriori podatek). Metoda je intuitivno preprosta in razumljiva.

1.2 Otvoritveni, zaključni, najvišji in najnižji dnevni tečaj

Pokazano je bilo, da indeks $S\&P500$ dosega nenavadno visoko realizirano volatilnost trgovanja v zadnjih 15 minutah pred zaprtjem borze. Razlog je velik promet dnevnih trgovanj z delnicami v zadnjih 15 minutah ([86]). Posledično so najvišji dnevni tečaji manj volatilni kot zaključni dnevni tečaji (glej sliko 1), saj najvišji tečaj v veliki večini primerov ne doseže vrha v zadnjih 15 minutah znotraj trgovalnega dneva.
Pri strojnem modeliranju je na volatilnih podatkih prisotnega veliko več šuma kot na manj volatilnih podatkih. V primerih, kjer se izvaja učenje na šumnih podatkih, se učni algoritem prilagodi naključnim atributom na učni množici, ki pa nimajo nobene relacije s ciljno funkcijo (z vrednostjo razredov).
Dejstvo, da so najvišji tečaji delnic (ang. 'high') manj volatilni kot zaključni tečaji delnic (ang. 'close'), naredi gibanje tečajev delnic na najvišjih dnevnih tečajih bolj predvidljivo. V naši raziskavi dobimo dovolj informacij o obnašanju najvišjih tečajev, ki so koristne pri dnevnih napovedih. Kljub dejstvu, da je volatilnost na najvišjih dnevnih tečajih signifikantno nižja kot na zaključnih dnevnih tečajih in da dnevne napovedi na najvišjih tečajih lahko vključimo v npr. avtomatsko trgovalno strategijo ali v ekspertne sisteme za upravljanje portfelja, se je večina povezanih raziskav ukvarjala le z napovedmi gibanja tečaja delnice pri zadnjem sklenjenem poslu v trgovalnem dnevu ([16, 31, 41, 40, 48, 53, 45, 47, 60, 63, 84]). Razlog, zakaj so raziskovalna dela usmerjena v 'close–to–close' gibanje cen, je, da

je čas zaključnih dnevnih tečajev natančno definiran z zaprtjem borze, medtem ko pa je čas najvišjih in najnižjih dnevnih tečajev zagotovo znan šele ob koncu trgovalnega dneva. Točen čas, ko je najvišji tečaj dosežen, je slučajna spremenljivka, ki ni niti čas ustavljanja glede na naravno filtracijo cen. Kljub slednjemu trdimo, da lahko konstruiramo trgovalno strategijo, ki bo vrnila pozitiven donos s pomočjo dobljenih napovedi. Akademske literature, ki raziskujejo napoved najvišjih in najnižjih tečajev, je razmeroma malo (npr: [65, 66]) v primerjavi z velikim številom literature, ki vključujejo zaključne dnevne tečaje: avtorji Martinez et al. [65] pokažejo, da z uporabo umetnih nevronskih mrež lahko napovejo najnižjo in najvišjo ceno delnic trenutnega trgovalnega dneva na dveh glavnih delnicah na Brazilskem delniškem trgu. Avtorji Mettenheim et al. [66] pokažejo, da je možno uspešno zmodelirati dinamičnost 5 likvidnih ameriških delnic z uporabo umetnih nevronskih mrež. Za vsak dan napovedujejo dnevni najnižji ali najvišji tečaj delnic in predlagajo trgovalni sistem. V dosegljivi literaturi nikjer nismo zasledili raziskav, kjer bi vključevali napovedi na najvišjih dnevnih tečajih (ali inverzno na dnevnih najnižjih tečajih) s pomočjo klasifikacijskih modelov in ali ti vrnejo perspektivne rezultate na velikem deležu delnic, ki so članice indeksa $S\&P500$. V naši raziskavi se bomo osredotočili na najvišje dnevne tečaje, saj v predlaganih strategijah nismo vključevali kratkih pozicij (ang. 'short positions').



Slika 1: Primerjava standardnih deviacij na najvišjih in zaključnih dnevnih tečajih na celotnem časovnem intervalu. Delnice so urejene po naraščajoči standardni deviaciji na najvišjih dnevnih tečajih. S slike je razvidno, da ima večina delnic nižjo standardno deviacijo izmerjeno na najvišjih dnevnih tečajih.

1.3 Finančni podatki, portfelj in napovedi delniškega trga

Portfelj je finančni pojem, ki označuje nabor naložb, ki jih imajo investicijske družbe, finančni skladi ali pa posamezniki. Sprejemanje odločitev, kot na primer, katere vrednostne papirje bomo vključili v portfelj, kakšen delež le teh bomo vključili, kdaj spremeniti naložbeni portfelj (kdaj kupiti oziroma prodati vrednostne papirje), itd. imenujemo upravljanje s portfeljem (ang. 'portfolio management'). Uvrščanje vrednostnih papirjev v razrede ima neposredno uporabno vrednost pri podpori odločanja pri samem trgovanju. Pri sestavi portfelja in pri vlaganjih nas najbolj zanima donosnost in pri tem lahko ključno odigrajo dobro izbrane delnice. S statističnimi metodami za uvrščanje lahko dobimo množico delnic, ki lahko koristijo pri sestavi portfelja.
Uporabnost dobljenih klasifikacijskih rezultatov smo želeli preveriti tako, da napovedi vključimo v sistem za trgovanje. S strategijami upravljanja, nadgrajenimi z informacijami o napovedih gibanja delnic, želimo preveriti uspešnost klasifikacijskih modelov. Pri tem nas zanima, ali takšne strategije prekašajo samo golo strategijo (brez vključitve napovedi gibanja delnic) ter ali med njimi obstaja signifikantna statistična razlika.

1.4 Vsebina disertacije

Doktorsko delo je zasnovano v 9 poglavjih, katerih struktura je sledeča. V uvodnem poglavju 1 predstavimo motivacijo in definiramo problem raziskovanja.

V naslednjem poglavju 2 je narejen pregled algoritmov strojnega učenja, ki se deli na nadzorovano učenje ali na nenadzorovano učenje. Obe vrsti algoritmov v raziskovalnem delu tudi uporabimo; nenadzorovano učenje uporabimo pri predlagani metodi za izbor atributov, nadzorovano učenje pa pri izgradnji klasifikacijskih modelov.

V poglavju 3 opišemo metode za izbor atributov, kjer je fokus predvsem na filtrirnih multivariatnih metodah. Tukaj predstavimo predlagan algoritem in njegove različice. V tem poglavju smo opisali tudi podatke, na katerih smo izvedli eksperimentalno delo.

V naslednjih dveh poglavjih 4 in 5 smo predstavili trgovalne strategije, obravnavane v eksperimentalnem delu in kazalce upsešnosti le teh, s katerimi primerjamo uspešnosti trgovalnih strategij.

Poglavji 6 in 7 predstavljata jedro doktorske disertacije. Na tem mestu so zbrani postopki, analize in rezultati eksperimentalnega dela.

Disertacijo končamo s poglavjem 9, ki vsebuje zaključni komentar predstavljenih prispevkov in oriše nekaj možnosti za nadaljnje delo.

2 STROJNO UČENJE IN UČNI ALGORITMI

2.1 Uvod

Strojno učenje (ang. 'machine learning') je področje umetne inteligence (ang. 'artificial intelligence'), ki se ukvarja z razvojem tehnik, ki omogočajo računalnikom oz. strojem, da se lahko učijo. Strojno učenje lahko opredelimo kot opisovanje ali **modeliranje** podatkov. Vhod v sistem za strojno učenje sta množica podatkov ter predznanje, izhod pa opis (model, hipoteza, teorija), ki te podatke skupaj s predznanjem opisuje in pojasnjuje. Predznanje je ponavadi kar prostor možnih modelov, v katerem bo algoritem iskal tistega, ki čim bolj ustreza vhodnim podatkom, ter kriterij optimalnosti, ki ga bo sistem med iskanjem poskušal izpolniti. Izvajalnim algoritmom, ki avtomatsko naučeno znanje uporabljajo za reševanje novih problemov, rečemo **model**. Za model zahtevamo, da čim bolj ustreza vhodnim podatkom in predznanju. Problem strojnega učenja lahko predstavimo tudi kot optimizacijski problem. Pri danem prostoru možnih rešitev (modelov) in pri danem kriteriju optimalnosti ali kriterijski funkciji, je treba poiskati tisto rešitev (model), ki zadošča kriteriju optimalnosti oziroma minimizira vrednost kriterijske funkcije. Pri tem je seveda vrednost kriterijske funkcije odvisna od trenutnega modela, predznanja in vhodnih podatkov, ki jih modeliramo. Ker je prostor možnih rešitev ponavadi zelo velik, je iskanje optimalne rešitve prezahtevno in se moramo zadovoljiti s čim boljšimi suboptimalnimi rešitvami [55]. Strojno učenje se močno opira na statistiko, saj se tudi statistika ukvarja s podatki, vendar v nasprotju z njo se strojno učenje bolj ukvarja s samimi algoritmi in računskimi operacijami. Ima širok spekter uporabnosti, in se uporablja pri spletnih iskalnikih, medicinskih diagnozah, detekciji ponarejenih dokumentov, analizi gibanja tečajev na borzah, razpoznavanju DNA sekvenc, razpoznavanju govora in pisave, razpoznavanju objektov pri strojnem vidu, računalniških igrah, robotiki itd. Nekateri sistemi strojnega učenja poskušajo eliminirati potrebo po človeški intuiciji pri analizi podatkov, medtem ko drugi sistemi temeljijo na sodelovanju med človekom in strojem. Algoritme strojnega učenja delimo na več vrst, glede na to, kaj je njihov cilj oz. rezultat učenja.

2.2 Učenje brez učitelja ali nenadzorovano učenje

Učenje brez učitelja ali nenadzorovano učenje je metoda strojnega učenja, ki temelji na podatkih, kjer imamo podane samo atribute, nimamo pa podanih razredov, katerim pripadajo. Naloga učnega algoritma je določiti te razrede. Eno izmed takih orodij je razvrščanje v skupine. Poleg besede razvrščanje se včasih uporabljajo sinonimi rojenje, grupiranje ali grozdenje (ang. 'clustering'). Razvrščanje se uporablja pri analizi naravnih in tehnoloških procesov, pri analizi ekonomskih trendov, pri preverjanju konsistentnosti in odvisnosti podatkov itd. Učenje brez učitelja se povezuje z razvrščanjem, kar je nasprotno od klasifikacije (učenje z učiteljem). Število želenih razredov je lahko podano vnaprej kot predznanje, ali pa mora primerno število razredov določiti sam učni algoritem. Naloga učnega algoritma je torej določiti relativno majhno število koherentnih razredov; t.j. skupin vzorcev, ki so si med seboj čimbolj podobni. Podobnost med vzorci je odvisna od izbrane mere podobnosti, ki je odločilnega

pomena za rezultat razvrščanja. Primerna mera podobnosti je lahko del predznanja [1, 55].

2.2.1 Razvrščanje v skupine

Naslednji zapisi so povzeti iz [26].
Klasifikacijski modeli so odvisni od vhodnih podatkov oziroma množice podatkov s pripadajočimi oznakami razredov. Kadar ni vednosti o pripadnosti podatkov danim razredom, poskušamo kategorije odkriti implicitno iz podatkov. Tem kategorijam rečemo skupine (ang. clusters). Dane podatke razvrstimo v nekaj skupin med seboj (znotraj skupin) podobnih podatkov. Poiskati skupine iz danih podatkov je naloga algoritmov, ki podatke razvrščajo v skupine.
Najpogostejši razlogi za razvrščanje v skupine:

- **pregledovanje podatkov:** pregled podatkov, predvsem otipati strukturo v podatkih, relacije, podobnosti in različnosti med enotami. Lahko se tudi poišče osamelce (ang. 'outliers') v podatkih.
- **zgoščanje podatkov:** kjer je potrebno analizirati velike količine podatkov, lahko namesto vseh enot analiziramo skupine enot, ki jih dobimo z razvrščanjem v skupine.
- **določitev tipologije:** empirična določitev tipologije pojavov v konkretnem področju raziskovanja in preverjanje domnev o tipologiji, ki jo raziskovalec postavi na osnovi teorije ali že opravljenih analiz podatkov.

Kasneje, v poglavju 3, predstavimo metodo za izbor informativnih atributov. V predstavljeni metodi uporabimo metode za razvrščanje v skupine. Pregledati želimo strukturo podatkov, tako da jih razvrstimo v skupine in tako na naraven način dobiti pripadnost enot danim razredom, kljub temu, da imamo te pripadnosti že vnaprej podane klasifikacijskim razredom. Skrbna uporaba metod razvrščanja v skupine lahko razkrije neznane strukture v podatkih, ki jih kasneje izkoristimo za izbor atributov pri sestavi modela. Torej ob dobljenih razvrstitvah nas bo predvsem zanimalo, kateri upoštevani atributi najbolj ločijo skupine med seboj in hkrati, da so v teh skupinah podatki, ki so si najbolj podobni med seboj glede na določeno mero razdalj.
Pri razvrščanju enot je najpogosteje uporabljena evklidska razdalja. Za enoti X in Y, opisani s številskimi spremenljivkami $X = (x_1, x_2, \ldots, x_m)$ in $Y = (y_1, y_2, \ldots, y_m)$, je evklidska razdalja med njima definirana takole: $d(X, Y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$. Pogosto so uporabljene tudi druge razdalje (npr. razdalja Manhattan, razdalja Minkowskega, itd.).
Množico enot označimo z $E = \{X_i\}_{i=1}^{n}$. Skupina enot je neprazna podmnožica množice enot, ki jo označimo s $C \subset E$, razvrstitev pa je množica skupin enot $\mathcal{C} = \{C_j\}_{j=1}^{k}$. Razvrstitev je popolna, če je vsaka enota natanko v eni skupini (ni prekrivanja med enotami). Ustreznost razvrstitve ponavadi izrazimo s kriterijsko funkcijo P, ki vsaki razvrstitvi $\mathcal{C}$ iz množice dopustnih rešitev priredi neko nenegativno realno število:

$$P : \mathcal{C} \to \mathbb{R}_0^+ .$$

Z vpeljanimi pojmi lahko zastavimo problem razvrščanja v skupine kot optimizacijski problem na naslednji način:
Določi razvrstitev $\mathcal{C}^*$ tako, da bo:

$$P(\mathcal{C}^*) = \min_{\mathcal{C} \in \phi} P(\mathcal{C}),$$

kjer je ϕ množica (dopustnih) razvrstitev; kar pomeni, če imamo množico razvrstitev ϕ in izračunamo za vsako razvrstitev $\mathcal{C} \in \phi$ vrednost kriterijske funkcije, je najboljša (najprimernejša) razvrstitev $(\mathcal{C}^*)$ tista, ki ima najmanjšo vrednost kriterijske funkcije.
Obstaja več metod razvrščanja v skupine, zato se moramo odločiti, katera je najprimernejša za reševanje postavljenega problema. Večino metod lahko razvrstimo v tri osnovne skupine: hierahične, nehierarhične in geometrijske metode. Glavne metode so:

- **Hierarhične metode** so najbrž največkrat uporabljene metode za razvrščanje v skupine. Te metode je mogoče deliti na **metode združevanja**, kjer iterativno združujemo najbolj podobne skupine med seboj, dokler ne ostaneta samo dve skupini. Zatem uporabnik ali pa sistem izbere najustreznejše število skupin; in **metode cepitve**, kjer na vsakem koraku izbrano skupino razcepimo na dve ali več skupin. Najobsežnejši razred metod hierarhičnega razvrščanja v skupine predstavljajo metode, ki temeljijo na zaporednem združevanju (zlivanju) dveh skupin v novo skupino. Postopek je sledeč (glej algoritem 1):

Algorithm 1 Hierarhične metode

1.) na začetku je vsaka enota skupina: $C_i = \{X_i\}, i = 1, 2, \ldots, n$
2.)
while ostane več kot ena skupina:
do
2.1.) določi najbližji si skupini C_p in C_q:
2.2.) $d(C_p, C_q) = \min_{u,v} d(C_u, C_v)$;
2.3.) združi skupini C_p in C_q v skupino $C_r = C_p \cup C_q$;
2.4.) zamenjaj skupini C_p in C_q s skupino C_r;
2.5.) določi mere različnosti d med novo skupino C_r in ostalimi.
end while

Postopek metod združevanja začnemo z razvrstitvijo z n skupinami (vsaka enota je v svoji skupini) in združujemo najbolj podobne skupine med seboj. Končamo z razvrstitvijo z eno samo skupino (po $n - 1$ korakih). Pri strukturi podatkov, kjer so skupine ločene med seboj, dobimo pravo razvrstitev v skupine z vsako metodo hierarhičnega združevanja. Kadar pa imamo prekrivajoče ali zelo specifično oblikovane skupine, tedaj se razvrstitve, dobljene z različnimi metodami, razlikujejo med seboj, in sicer toliko bolj, kolikor bolj je naravna struktura podatkov slaba, neizrazita. V raziskovalnem delu imamo podatke tesno med seboj prepletene, zato smo vključili več različnih metod razvrščanja v skupine in opazovali, kako se z različnimi metodami rezultati spreminjajo.

V literaturi so najpogosteje omenjene minimalna, maksimalna in Wardova metoda. Najbrž zato, ker ima vsaka od teh metod zanimive specifične lastnosti. Te metode bomo uporabili pri eksperimentalnem delu. Minimalna metoda se imenuje tudi enojna povezanost (ang. 'single linkage'), ker v vsakem koraku postopka združuje tisti skupini, med katerima obstaja največja povezanost izmerjena med najbližjima enotama ene in druge skupine, je pa neuporabna pri neizrazito ločenih skupinah. Minimalna metoda se zelo obnese pri razkrivanju dolgih 'klobasastih', tudi neeliptičnih struktur. Pri neizrazito ločenih skupinah pa se kaže 'verižni' učinek metode, ko v vsakem koraku združevanja skupini dodaja le posamezno enoto. Minimalna metoda išče skupine, ki so izrazito ločene med seboj in se ne zmeni za homogenost znotraj njih. Maksimalna metoda pa je osredotočena na razkrivanje znotraj homogenih skupin. Več avtorjev ([25, 27, 68]) je empirično primerjalo različne metode hierarhičnega združevanja v skupine na več slučajno generiranih skupinah podatkov in pri tem ugotavljalo primernost posameznih metod. Njihove empirične primerjave so pokazale, da je Wardova metoda najprimernejša za eliptično strukturirane podatke, medtem ko je minimalna metoda primernejša za odkrivanje verižno strukturiranih podatkov. Maksimalna metoda pa dobro razkriva okrogle skupine.

Mere različnosti d med novo skupino in ostalimi v postopku združevanja v skupine določamo na več načinov in ti določajo različne metode hierarhičnega združevanja v skupine. Vzemimo, da imamo v nekem koraku postopka tri skupine C_i, C_j in C_k ter podane mere različnosti. Denimo, da sta skupini C_i in C_j najbližji, zato ju združimo v novo skupino $C_i \cup C_j$. Mero različnosti med novo skupino in skupino C_k določimo na naslednje načine:

- **Minimalna metoda ali enojna povezanost**([82])

$$d_{\text{min}}(C_i \cup C_j, C_k) = \min(d(C_i, C_k), d(C_j, C_k))$$

- **Maksimalna metoda ali polna povezanost**([74])

$$d_{\text{maks}}(C_i \cup C_j, C_k) = \max(d(C_i, C_k), d(C_j, C_k))$$

- **Wardova metoda**([95])

$$d_{\text{w}}(C_i \cup C_j, C_k) = \frac{(n_i + n_j)n_k}{n_i + n_j + n_k} d^2(T_{ij}, T_k),$$

  kjer s T_{ij} označimo težišče združene skupine $C_i \cup C_j$ in s T_k težišče skupine C_k ter z n_i, n_j in n_k pa označimo število enot, ki se nahajajo v skupinah C_i, C_j in C_k.

Za naš problem predvidevamo, da minimalna metoda ne bo prinesla zadovoljivih rezultatov, saj so skupine na podatkih prekrivajoče in niso strogo ločene med seboj.

- **Pri nehierarhičnih metodah** je potrebno vnaprej podati število skupin iskane razvrstitve. Te metode razvrščajo enote tako, da z izbranim optimizacijskim kriterijem izboljšujejo vnaprej podano

začetno razvrstitev: začnejo z začetno razvrstitvijo s podanim številom skupin in prestavljajo enote iz ene skupine v druge skupine z namenom, da s temi prestavitvami dosežejo zmanjšanje (ali povečanje) vrednosti izbrane kriterijske funkcije razvrščanja. Proces se iteracijsko nadaljuje, dokler nobena prestavitev enote ne izboljša vrednosti kriterijske funkcije. Zaradi nevarnosti lokalnih optimalnih razvrstitev in v želji po čim boljši rešitvi je priporočljivo, da razvrščanje s temi metodami ponovimo z več različnimi začetnimi razvrstitvami, po možnosti dobljenimi z različnimi metodami (npr. razvrstitev z metodami hierarhičnega združevanja v skupine). Metode, ki so primerne za razvrščanje v skupine tudi večjih količin podatkov (nekaj tisoč) so poznane pod imeni metoda voditeljev (npr. [36]) ali K-MEANS (npr. [29, 42, 64]) ali metoda (dinamičnih) oblakov (npr. [19, 20]). Obravnavali bomo metodo voditeljev (k–means), saj smo to tudi vključili v raziskovanje.

- **Metoda voditeljev**.

  Ta metoda je zelo popularna, ker zmore razvrščati v skupine večje število enot. Metoda voditeljev je iteracijska metoda, na vsakem koraku izračuna določeno kriterijsko funkcijo in poskuša z drugačno razvrstitvijo doseči, da se zmanjša vrednost kriterijske funkcije. Pri metodi voditeljev se je potrebno vnaprej odločiti o številu razvrstitev enot v skupine.

  Postopek se začne z vnaprej podano množico predstavnikov posameznih skupin (z začetnimi voditelji). Metoda doda enote najbližjim voditeljem in tako nastanejo novonastale skupine. Tem skupinam se izračuna težišča, ki so novi voditelji, nato metoda spet priredi enote najbližjim voditeljem, itd. Postopek se konča, ko se nova množica voditeljev ne razlikuje od množice voditeljev, dobljene korak pred njo. Za najboljšo razvrstitev se vzame tisto, ki ima najmanjšo vrednost kriterijske funkcije.

  Osnovna shema metode voditeljev je predstavljena v algoritmu 2:

Algorithm 2 Metoda voditeljev

1.) določi začetno množico voditeljev $L = \{L_i\}$
2.)
while voditelji niso ustaljeni **do**
  2.1.) določi razvrstitev $\mathcal{C}$ tako, da prirediš vsako enoto njej najbližjemu voditelju
  2.2.) za vsako skupino $C_i \in \mathcal{C}$ izračunaj njeno središče $\overline{C_i}$
  in ga določi za novega voditelja L_i skupine C_i
end while

Ker je množica enot, ki jih razvrščamo, končna, je končna tudi množica vseh razvrstitev. Zato zgornji postopek prej ali slej skonvergira v lokalno optimalno rešitev.

- **Geometrijske metode** omogočajo preslikavo podatkov iz originalnega večrazsežnega prostora v manj razsežni, pogosto kar v dvorazsežni prostor, v katerem je lahko grafično ali kako drugače

raziskati strukturo podatkov. Najbolj znani geometrijski metodi sta metoda glavnih komponent in večrazsežnostno lestvičenje ([57, 81]).

2.2.2 Izbor metod za razvrščanje v skupine

V primeru, ko nimamo jasne domneve o številu skupin, lahko izbiramo med hierarhičnimi metodami, če pa poznamo število skupin, so primernejše metode nehierarhičnega razvrščanja v skupine (npr. metoda voditeljev). Naslednji kriterij, ki se upošteva pri odločanju o ustrezni metodi, je tudi število enot. Najbolj znane metode razvrščanja v skupine, kot so hierarhične metode združevanja in metoda prestavljanj, so uporabne le za razvrščanje manjšega števila enot (nekaj sto), medtem ko pa za razvrščanje nekaj tisoč enot je primerna na primer metoda voditeljev ali nekatere druge metode, ki so razvite posebej za večje količine podatkov. Pri izbiranju ustrezne metode je koristno vedeti, kakšen tip skupin želimo razkriti v podatkih: ali gre za eliptične ali verižne skupine, ali za med seboj ločene skupine ali za prekrivajoče, itd. Vsaka metoda pri iskanju strukture v podatkih vsiljuje strukturo, ki je vgrajena v metodi. Nekatere metode na primer znajo razkriti le krogle, nekatere le dolge 'klobase', ne glede na to, ali te v naravni strukturi podatkov so ali niso.
Pri predstavitvi podatkov s tehničnimi indikatorji (glej poglavje 3.7.2) pri zastavljenem eksperimentalnem problemu (glej 6.1) dobimo izrazite eliptične skupine, ki pa so med seboj prekrivajoče. V našem primeru poznamo število skupin, število enot, ki nastopajo kot vhodni podatki za razvrščanje je nekaj 100 v večrazsežnem prostoru (učne množice so dolge 500 trgovalnih dni). V eksperimentalnem delu bomo obravnavane enote razvrščali z več različnimi metodami in primerjali dobljene razvrstitve: metodo voditeljev iz nehierarhične skupine metod, Wardovo metodo, minimalno ter maksimalno metodo iz hierarhične skupine metod.

2.3 Učenje z učiteljem ali nadzorovano učenje

Učenje z učiteljem ali nadzorovano učenje je princip strojnega učenja za modeliranje funkcije na podlagi učne množice vzorcev. **Učna množica** vzorcev je par množice vzorcev in množice njihovih **oznak razredov** oz. par množice vhodnih podatkov sistema in množice željenih izhodov sistema. Izhod sistema je lahko zvezno področje vrednosti–regresija ali pa enolična oznaka razreda kateremu recimo pripadajo vhodni podatki– klasifikacija. Naloga učečega sistema je, da generalizira znanje, ki ga dobi iz učne množice. Rezultat učenja z učiteljem je v večini primerov nek globalen model funkcije, ki preslika vhodne podatke v nek želen izhod. Postopek učenja z učiteljem je sestavljen iz naslednjih korakov:

- Najprej je treba določiti področje uporabe. S tem tudi določimo vrsto podatkov, ki jih bo sistem obdeloval.

- Naslednji korak je zbiranje podatkov, na katerih se bo sistem učil. Ti podatki morajo biti značilni za področje, v katerem bo v končni fazi sistem deloval.

- Definirati je treba tudi atribute vzorcev. Pravilnost modelirane funkcije je precej odvisna od izbire atributov.
- Določiti strukturo modela, na primer lahko se odločimo za nevronske mreže ali odločitvena drevesa.
- Izvedba algoritma učenja na učni množici vzorcev. Parametre algoritma se nastavi optimalno glede na testno množico vzorcev.

Ločimo med dvema vrstama modelov: klasifikacija ali uvrščanje in regresija, ki ju opišemo v naslednjih podpoglavjih.

2.3.1 Klasifikacija ali uvrščanje

Kadar želimo razdeliti podatke v diskretne kategorije, kjer so kategorije ponavadi vnaprej opredeljene z nekim logičnim ozadjem (npr. ali so celice karcinogene ali ne, kaj predstavlja slika: tigra, rožo ali ocean, kakšno besedilo obdelujemo: z versko vsebino ali politično vsebino). Podobnim problemom in problemom katerih ciljni model je diskretna funkcija, pravimo klasifikacijski problemi, kategorijam pa razredi. Kadar imamo učno množico vzorcev s pripadajočimi oznakami razredov, imamo nadzorovano učenje. Ko je funkcija naučena, jo uporabljamo za klasifikacijo (uvrščanje). Naloga klasifikatorja je na testni množici podatkov, opisan z množico atributov določiti, kateremu izmed možnih razredov pripada. Atributi so neodvisne zvezne ali diskretne spremenljivke, s katerimi opisujemo enote, razred pa je odvisna diskretna spremenljivka, ki ji določimo vrednost glede na vrednosti neodvisnih spremenljivk. Zato, da lahko klasifikator določi razred, mora imeti na nek način predstavljeno diskretno funkcijo, ki preslika prostor atributov v razred. Klasifikatorje ločimo glede na način predstavitve klasifikatorjeve funkcije. Najbolj pogosti klasifikatorji so: odločitvena drevesa, odločitvena pravila, naivni Bayesov klasifikator, Bayesove verjetnostne mreže, klasifikator z najbližjimi sosedi, linearna diskriminantna funkcija, logistična regresija, klasifikator po metodi podpornih vektorjev (SVM) ter usmerjene (večnivojske nevronske mreže). Mnogi odločitveni problemi, diagnostični problemi, problemi vodenja in problemi napovedovanja se lahko predstavijo kot klasifikacijski problemi. Tipični primeri so medicinska diagnostika in prognostika, napovedovanje vremena, diagnostika industrijskih procesov, klasifikacija izdelkov po kakovosti, vodenje dinamičnih sistemov ipd.

2.3.2 Regresija

Problemom, katerih ciljni model je zvezna funkcija, pravimo regresijski problemi. Tako kot pri klasifikaciji tudi tu uporabljamo avtomatsko zgrajeno funkcijo za ugotavljanje vrednosti funkcije pri danih vrednostih neodvisnih spremenljivk. Zaloga vrednosti je (potencialno) neskončna urejena množica. Odvisni spremenljivki pravimo regresijska spremenljivka ali zvezni razred. Naloga regresijskega prediktorja je za vzorec, ki je tako kot pri klasifikaciji opisan z množico atributov, določiti vrednost odvisne regresijske spremenljivke, ki pa je, za razliko od klasifikacijskega razreda, zvezna (številska). Podobno

kot pri klasifikaciji, mora tudi regresijski prediktor imeti na nek način predstavljeno zvezno funkcijo, ki preslika prostor atributov v napovedano vrednost. Naloga učnega algoritma je torej iz množice vzorcev z znanimi vrednostmi odvisne spremenljivke izračunati zvezno funkcijo, ki jo lahko uporabimo za določanje vrednosti regresijske spremenljivke novih primerov. Poleg vrednosti funkcije je pogosta zahteva pri regresijskih problemih tudi interval zaupanja. Regresijske predikatorje ločimo glede na način predstavitve regresijske funkcije. Najpogostejše strukture regresijskih modelov so: regresijska drevesa, linearna regresija, lokalno utežena regresija, regresija po metodi podpornih vektorjev ter usmerjene umetne nevronske mreže.
V našem raziskovalnem delu se bomo osredotočili na nadzorovano učenje in opisali bomo klasifikacijske modele, ki jih bomo skozi eksperimentalno delo uporabljali. V naslednjih podpoglavjih bomo predstavili sledeče klasifikacijske algoritme: linearna diskriminantna analiza, klasifikator po metodi podpornih vektorjev in naivni Bayesov klasifikator.

2.4 Linearna diskriminantna analiza

Z diskriminantno analizo poiščemo tako linearno kombinacijo merjenih spremenljivk, da bo maksimalno ločila vnaprej določene skupine in da bo napaka pri uvrščanju enot v skupine najmanjša. Pri diskriminantni analizi torej gre za iskanje tistih razsežnosti, ki kar najbolj obrazložijo razlike med skupinami (pojasnjevanje), in za kar se da dobro prirejanje enot vnaprej danim skupinam (napovedovanje). Diskriminantna analiza služi doseganju naslednjih ciljev:

- določiti tiste spremenljivke, ki kar najbolje ločujejo dve skupini;
- kreirati funkcije, ki predstavljajo razlike med skupinama in
- uporabiti izbrane spremenljivke in dano funkcijo pri napovedovanju pripadnosti določeni skupini.

Diskriminantna analiza želi torej poiskati pravilo, po katerem bi za posamezno enoto lahko napovedali, kateri skupini pripada. Diskriminantna analiza zato išče take linearne kombinacije merjenih spremenljivk, da bodo čim bolje ločile vnaprej dane skupine med seboj. Dobljene linearne kombinacije imenujemo diskriminanatne spremenljivke, vzorcu, na katerih jih iščemo, pa rečemo učni vzorec. Za enote izven učnega vzorca izračunamo vrednosti na diskriminantnih spremenljivkah in jih uvrstimo v tisto skupino, za katero so take vrednosti najbolj značilne. V primeru več skupin razlike med skupinami lahko popišemo z več diskriminantnimi spremenljivkami - največ jih je lahko $\min\{p, k-1\}$ (p je število atributov, k pa število skupin). Poiščemo jih tako, da maksimiziramo kvocient med variabilnostjo med skupinami (razlike povprečij) in variabilnostjo (varianco) znotraj skupin.

2.4.1 Predpostavke diskriminantne analize

1.) Število skupin $k \geq 2$.

2.) Vsaj 2 enoti v vsaki skupini.

3.) $p < n - 2$; p je število atributov in n število vseh enot v vzorcu.

4.) Nobena spremenljivka ne sme biti linearna kombinacija preostalih spremenljivk (multikolinearnost).

5.) Pri statističnemu ocenjevanju se predpostavlja, da so v vsaki skupini enot (vzorcu) enote slučajno izbrane iz populacije, kjer so spremenljivke porazdeljene večrazsežno normalno.

6.) Variančno-kovariančna matrika $p \times p$ je v vsaki populacijski skupini enaka. [10]

2.4.2 Diskriminantna analiza v primeru dveh skupin

Denimo, da imamo množico n enot dimenzije p: $x_1, \ldots x_n$, n_1 enot naj pripada skupini D_1, označimo jih z r_1 in n_2 enot naj pripada skupini D_2, označimo jih z r_2. Ko tvorimo linearno kombinacijo merjenih atributov x, dobimo skalarni produkt oblike $w^T x$ (diskriminantna funkcija), kjer je w vektor diskriminantnih koeficientov ali uteži in dobimo pripadajočo množico n enot $y_1, \ldots y_n$ ločenih v dveh skupinah Y_1 in Y_2.
Fisher je definiral diskriminantno funkcijo kot linearno kombinacijo $D = w^T x$ tako, da je kvocient razlik aritmetičnih sredin diskriminantne spremenljivke v obeh skupinah glede na varianco diskriminantne spremenljivke znotraj skupine maksimalen.
Želimo torej poiskati vektor uteži w, ki bodo najbolj pojasnile razlike med skupinama. Za mero, s katero merimo razlike med linearno kombinacijo enot, vzamemo razliko aritmetičnih sredin, definirane kot $m_i = \frac{1}{n_i} \sum_{x \in D_i} x$. Definirajmo variančno-kovariančni matriki S_i in S_W: $S_i = \sum_{x \in D_i} (x - m_i)(x - m_i)^T$ in $S_W = S_1 + S_2$. Definirajmo tudi $S_B = (m_1 - m_2) \cdot (m_1 - m_2)^T$. Kvocient, ki ga želimo maksimizirati je:

$$J(w) = \frac{w^T S_B w}{w^T S_W w},$$

ki je pogoj, na osnovi katerega izračunamo uteži w diskriminantne spremenljivke. Reševanje tega optimizacijskega problema privede do rešitve za w, ki je:

$$w = \alpha n S_W^{-1} (m_1 - m_2),$$

kjer je α konstanta. Skupine določimo tako, da enota x spada v skupino r_1, če je $w^T(x - m) \geq 0$ sicer pa v skupino r_2 [23].

2.5 Klasifikator po metodi podpornih vektorjev

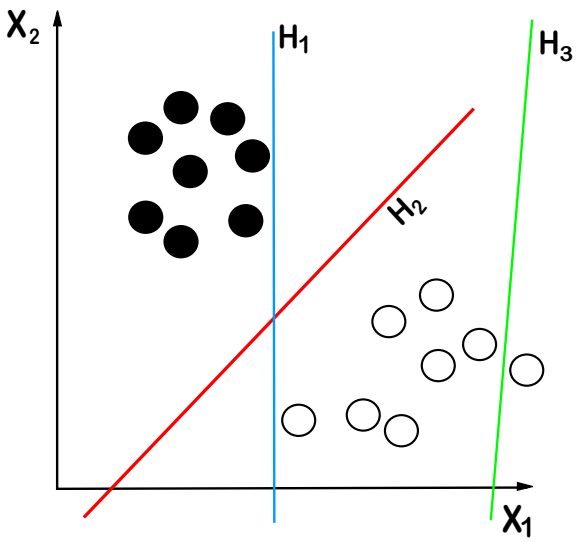
Metoda podpornih vektorjev je primerna za učenje na velikih množicah enot, opisanih z velikim številom atributov. Metoda podpornih vektorjev ali po angleško Support Vector Machines, s kratico SVM, dosega visoko točnost napovedi. Slaba stran metode je, da je interpretacija naučenega težavna. Metodo podpornih vektorjev je predlagal Vapnik in je bila najprej uporabljena za probleme v bioinformatiki in tekstovni kategorizaciji. Ko uporabljamo SVM, se soočamo z dvema problemoma, in sicer kako izbrati

optimalno jedro za SVM in kako določiti najboljše parametre. Ta dva problema sta odločilna, saj na podlagi jedra izberemo prave parametre. V doktorski disertaciji smo se omejili na problem uvrščanja v razrede, zato bomo metodo podpornih vektorjev predstavili v tem pogledu.
Osnovna različica SVM-ja se ukvarja s klasifikacijskimi problemi med dvema razredoma. Vsaka enota je predstavljena s točko v nekem večrazsežnem realnem prostoru, učenje pa postane optimizacijski problem. Iščemo tako hiperravnino, ki bo razmejila predstavnike enega razreda od predstavnikov drugega tako, da bodo učne enote ležale na pravi strani ravnine in še čim dlje od nje. Ta matematični problem lahko s pomočjo nekaj matematičnih prijemov preoblikujemo v dualno obliko, to pa rešujemo numerično. Metodo je mogoče razširiti tako, da namesto hiperravnine odkriva tudi drugačne razmejitvene ploskve. Obstajajo tudi razširitve na več razredov. Cilj te metode je narediti klasifikator, tako da bo deloval tudi na še nepoznanih primerih.

2.5.1 Izpeljava optimizacijskega problema:

- imamo množico enot, ki so predstavljene z vektorji $x_i \in \mathbb{R}^d$;
- imamo dva razreda, pozitivnega in negativnega. Vsak primer spada v enega od dveh razredov: $y_i \in \{-1, 1\}$ in
- podatki so predstavljeni kot

$$D = \{(x_i, y_i) | x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}\}_{i=1}^n.$$



Slika 2: Grafična predstavitev klasifikacijskega problema.

Enačbo ravnine lahko zapišemo kot $w^T x + b = 0$. Za w in b velja

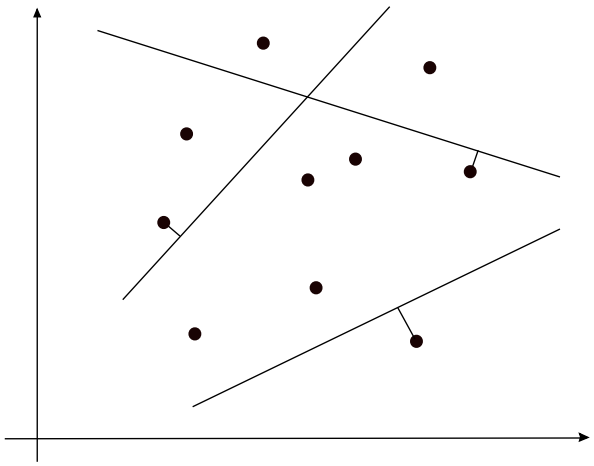
$$\min_i |\langle w, x_i \rangle + b| = 1$$

in za podporni vektor x_i velja:

- $w^T x_i + b = -1$, za negativni razred
- $w^T x_i + b = 1$, za pozitivni razred,

kjer je w normalni vektor ravnine, b pa konstanta. Hiperravnina je veljavna, kadar izpolnjuje naslednji pogoj:

$$y_i(w^T x_i + b) \geq 1, \forall i.$$



Slika 3: Hiperravnine.

Norma vektorja w mora biti enaka obratu razdalje od najbližje točke do hiperravnine. Ideja je ilustrirana na sliki 3, kjer je prikazana razdalja od najbljižje točke pa do hiperravnine. Hiperravnin, ki ustrezajo danemu pogoju, je lahko več (slika 2), samo ena je takšna, ki maksimizira razdaljo med mejo hiperravnine in najbljižjo točko do vsakega razreda. Potrebujemo dodatni kriterij, ki določa najboljšo izmed njih. Ko računamo $w^T x + b$ za različne primere x, se pri vsaki točki pozna le to, kako daleč je od razmejitvene hiperravnine in na kateri strani je. Ne želimo, da bi bili učni primeri preblizu hiperravnine, saj lahko že majhne perturbacije povzročijo napake v napovedi. Pričakujemo, da bo ta meja dobra posplošitev.
Funkcijski rob učnega primera (x_i, y_i) glede na hiperravnino (w, b), je število γ.

$$\gamma(w, b, i) = y_i(w^T x_i + b)$$

Funkcijski rob hiperravnine glede na učno množico je razdalja najbližje točke do hiperravnine

$$\gamma(w, b) = \min_{i=1,\ldots,l} \gamma(w, b, i)$$

Geometrijski rob učnega primera (x_i, y_i) definiramo kot:

$$\overline{\gamma}(w, b, i) = \gamma(\frac{w}{||w||}, \frac{b}{||w||}, i)$$

Med funkcijskim in geometrijskim robom velja naslednja zveza $\overline{\gamma} = \frac{\gamma}{||w||}$.
Najpogosteje iščemo hiperravnino z najširšim robom. Hiperravnina z najširšim robom ima največjo razdaljo do najbližje točke. Rob učne množice $S = \{(x_1, y_1), \ldots, (x_l, y_l)\}$ je definiran kot:

$$\gamma(S) = \max_{w,b} \overline{\gamma}(w, b)$$

Hiperravnini (w^*, b^*), ki izpolnjuje pogoj $\gamma(S) = \overline{\gamma}(w^*, b^*)$ pravimo hiperravnina z najširšim robom (ang. maximal margin hyperplane). Najbližjim enotam optimalne hiperravnine pravimo **podporni vektorji**.

2.5.2 Formulacija optimizacijskega problema

Naša želja je, da imamo hiperravnino z najširšim geometrijskim robom. Če uporabimo zvezo $\overline{\gamma}(w, b) = \frac{\gamma(w,b)}{||w||}$ in fiksiramo $\gamma(w, b) = 1$, dobimo sledeči optimizacijski problem:

- minimiziraj $\frac{1}{2}||w||^2$
- pri pogojih : $y_i(w^T x_i + b) \geq 1, \forall i = 1 \ldots l$

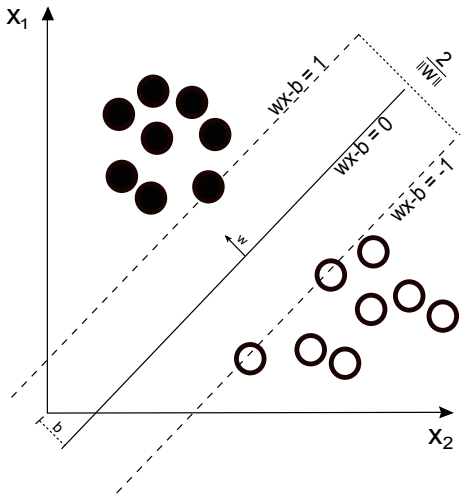
Razdalja med dvema roboma pa je enaka:

$$\frac{2}{||w||}.$$

Vektor w predstavlja normalo za dani **podporni vektor**. Razdalja med podpornimi vektorji je odvisna od w (normale), zato je naš cilj dobiti minimalen w.
Problem klasifikacije z metodo podpornih vektorjev se tako prevede v optimizacijski problem, ki je odvisen od w.
Ekstreme funkcije lahko izračunamo z Lagrangeejevimi multiplikatorji. Tako dobimo dokončno izpeljavo dualnega optimizacijskega problema:

- maksimiraj $\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$
- pri pogojih: $\sum_{i=1}^{n} y_i \alpha_i = 0$

Optimizacijski problem rešujemo s kvadratičnim programiranjem, kjer $k(x_i, x_j) = x_i^T x_j$ predstavlja notranji produkt dveh primerov.



Slika 4: Prikaz podpornih vektorjev in ločilne hiperravnine.

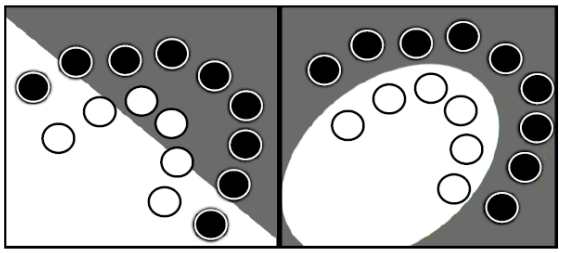
2.5.3 Trik z jedri

V tem poglavju bomo opisali tako imenovani trik jedra, kjer lahko notranji produkt zamenjamo z jedrno fukcijo K. Znotraj jedra lahko dve enoti preslikamo v poljubno dimenzijo in izračunamo njuno podobnost. Ideja je bila predlagana z namenom oblikovanja nelinearnih klasifikatorjev. Najbolj znana nelinearna jedra:

- RBF (radial basis function:) $k(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2)$, za$\gamma \geq 0$
- Polinomsko jedro: $k(x_i, x_j) = (x_i, x_j)^d$
- Gaussovo jedro: $k(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{2\sigma^2})$

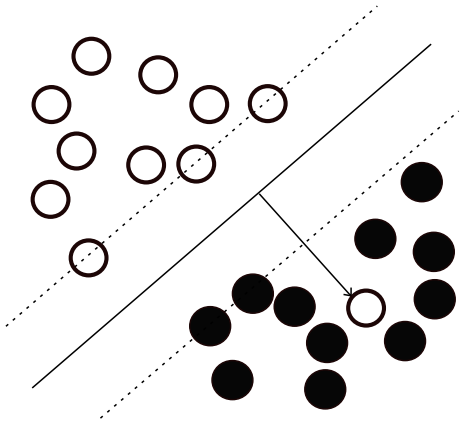
Najbolj učinkovito jedro v splošnem je RBF, saj ima v primerjavi z linearnim sposobnost ustvarjanja ukrivljenih hiperravnin in je tako lahko veliko učinkoviteje pri kompleksnih problemih in obravnavi šuma. Parameter γ določa, kako zelo lahko ukrivi hiperravnino pri ločevanju razredov.
S slike 5 je lepo razvidno, da v določenih primerih linearno jedro ne more najti optimalne hiperravnine, medtem ko jedro RBF pravilno najde ukrivljeno hiperravnino. Moramo pa biti pazljivi, saj se ob velikih vrednostih parametra γ hiperravnina preveč prilagodi obliki podatkov in posledično onemogoči učinkovito obravnavo šuma.
V primeru, da pozitivnih in negativnih primerov ne moremo razmejiti z ravnino, vpeljemo kazenske spremenljivke ξ. Optimizacijski problem razširimo v:

- minimiziraj $\frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i$
- pri pogojih: $y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \forall i = 1 \ldots l$



Slika 5: Na levi strani je prikazano linearno, na desni strani pa RBF jedro.

Konstanta C je parameter metode, s katero določamo obravnavo šumov pri učenju. Z njo določimo, kako zelo naj dopušča odstopanja od osnovne skupine enega razreda. Za dobro klasifikacijo je potrebno izbrati pravilno vrednost konstante, s katero bo algoritem znal dovolj dobro obravnavati šum. Višji kot je C, bolj je metoda tolerantna do odstopanja. Biti pa moramo pazljivi, saj od neke vrednosti dalje višji C-ji ne prinašajo boljše klasifikacije, ampak po nepotrebnem upočasnijo proces učenja.

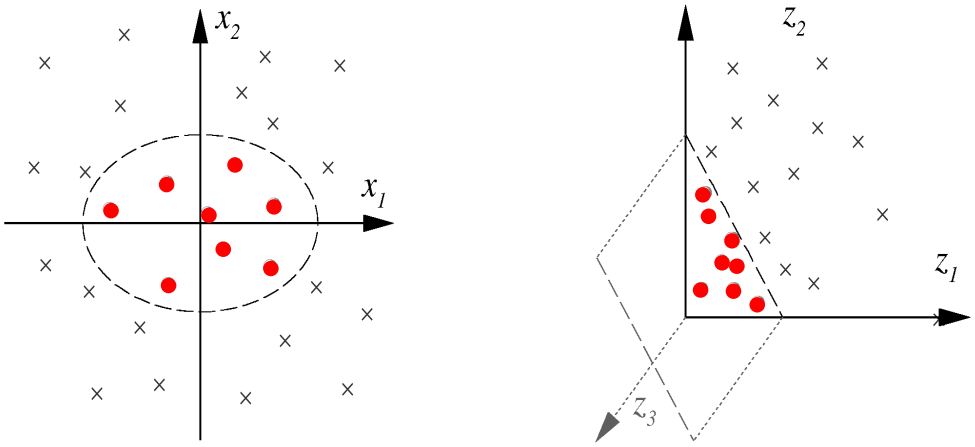


Slika 6: Kako s parametrom C obravnavamo šum.

2.5.4 Jedra

V tem razdelku obravnavamo nelinearen algoritem podpornih vektorjev. To naredimo s preslikavo učnih primerov x_i, $\phi : X \to F$ v nek prostor F in nato izvedemo standarden algoritem podpornih vektorjev. Poglejmo si algoritem na primeru.

Primer 1 (Kvadratne funkcije v $\mathbb{R}^2$) *Denimo, da imamo funkcijo $\phi : \mathbb{R}^2 \to \mathbb{R}^3$ podano s $\phi(x_1, x_2) =$*



Slika 7: Sumimo, da je meja med enotami (med križci in krožci) elipsa (leva slika). Ko preslikamo v drug funkcijski prostor z nelinearno preslikavo $\phi_2(x) = (z_1, z_2, z_3) = (|x|_1^2, |x|_2^2, \sqrt{2}|x|_1|x|_2)$ na desni sliki, elipsa postane hiperravnina, ki je vzporedna z_3 - osi, vse točke pa so preslikane na (z_1, z_2) ravnino.

$(x_1^2, \sqrt{2}x_1x_2, x_2^2)$. V tem primeru so indeksi mišljeni kot komponente vektorja $x \in \mathbb{R}^2$. Primeri postanejo računsko nezmogljivi, če imamo polinomske funkcije višjih redov ali višjih dimenzij, kot na primer število različnih monomov stopnje p, ki je $\binom{d+p-1}{p}$, kjer je $d = dim(X)$.
Ta način dela seveda ni možen, zato moramo poiskati računsko zmogljivejšo pot. Pri primeru 1 imamo:

$$\langle (x_1^2, \sqrt{2}x_1x_2, x_2^2), (x_1'^2, \sqrt{2}x_1'x_2', x_2'^2) \rangle = \langle x, x' \rangle^2. \tag{1}$$

Kot smo že v prejšnjem poglavju omenili, je algoritem podpornih vektorjev odvisen samo od skalarnega produkta med primeri x_i, zato zadošča poznati $k(x, x') := \langle \phi(x), \phi(x') \rangle$ (dejansko poznavanje funkcije ϕ ni potrebno). Zagotoviti moramo izračun skalarnega produkta s pomočjo jedrne funkcije, ki omogoča implicitno transformacijo. V nelinearnem primeru je optimizacijski problem poiskati najprimernejšo funkcijo v preslikanem prostoru in ne v vhodnem prostoru [80].

2.6 Naivni Bayesov klasifikator

Naloga Naivnega Bayesovega klasifikatorja je izračunati pogojne verjetnosti za vsak razred pri danih vrednostih (vseh) atributov za dani vzorec, ki ga želimo klasificirati. Bayesov klasifikator, ki izračuna pogojne verjetnosti razredov, je optimalen, saj minimizira pričakovano napako. Ker Bayesovega klasifikatorja, ki bi eksaktno izračunal pogojne verjetnosti razredov ne poznamo (razen v primerih, ko učna

množica pokriva celoten prostor vrednosti vseh atributov), je potrebno izračunati približke verjetnosti z vpeljavo določenih predpostavk. Naivni Bayesov klasifikator predpostavi pogojno neodvisnost atributov pri danem razredu. To omogoči, da učna množica zadošča za zanesljivo oceno vseh potrebnih verjetnosti za izračun končne pogojne verjetnosti vsakega razreda. Implementacije naivnega Bayesovega klasifikatorja navadno predpostavljajo samo diskretne atribute, zato je potrebno v takih primerih zvezne atribute vnaprej ustrezno diskretizirati.

2.6.1 Naivna Bayesova formula za 2 razreda

Denimo, da imamo vzorec podatkov s T enotami, kjer vsaka enota pripada enemu izmed razredov C_1 ali C_2 in je predstavljena kot n-razsežen vektor $X = \{x_1, x_2, \ldots, x_n\}$ z n atributi A_1, A_2, …, A_n. Dana enota X pripada tistemu razredu, ki ima najvišjo apriorno verjetnost pogojeno na X, kar pomeni, da enota X pripada razredu C_i natanko tedaj, ko je izpolnjen pogoj $P(C_i|X) > P(C_j|X)$, za $i, j = 1, 2$ in $i \neq j$. Poiščemo torej tak razred, ki maksimizira pogojno verjetnost $P(C_i|X)$. Po Bayesovem izreku velja:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}.$$

Verjetnost $P(X)$ je enaka za vse razrede, zato maksimiziramo le vrednost $P(X|C_i)P(C_i)$. Če apriorne verjetnosti razredov $P(C_i)$ niso znane, potem lahko predpostavimo, da so razredi enako verjetni in zatorej maksimiziramo le izraz $P(X|C_i)$. Apriorne verjetnosti razredov lahko ocenimo kot $P(C_i) =$ frekv$(C_i, T)/|T|$. Pri podatkih z veliko atributi je računsko potratno izračunavati vrednosti $P(X|C_i)$. Z namenom, da bi zmanjšali računsko zahtevnost pri izračunu vrednosti $P(X|C_i)P(C_i)$, privzamemo naivno predpostavko o pogojnih neodvisnosti atributov pri danem razredu, kar pomeni:

$$P(X|C_i) = \Pi_{k=1}^{n} P(x_k|C_i).$$

Verjetnosti $P(x_1|C_i), P(x_2|C_i), \ldots, P(x_n|C_i)$ lahko ocenimo s pomočjo učne množice, kjer je x_k vrednost atributa A_k za enoto X.

- Če je A_k kategorična spremenljivka, potem je vrednost $P(x_k|C_i)$ število enot, ki pripadajo razredu C_i v vzorcu T; x_k pomeni vrednost atributa A_k, deljeno z frekv(C_i, T), kar je število vrednosti razreda C_i v vzorcu T.

- Če je A_k zvezna spremenljivka, potem predpostavimo, da so vrednosti porazdeljene normalno s povprečjem μ in standardno deviacijo σ, definirano kot:

  $$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp - \frac{(x-\mu)^2}{2\sigma^2},$$

  in velja:

  $$p(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}).$$

Izračunati moramo μ_{C_i} in σ_{C_i}, ki sta povprečje in standardna deviacija vrednosti atributa A_k na učni množici razreda C_i. Da bi pravilno napovedali, kateremu razredu pripada enota X, izračunamo za vsak razred C_i vrednost $P(X|C_i)P(C_i)$. Klasifikator napove, da X pripada razredu C_i natanko tedaj, če razred C_i maksimizira $P(X|C_i)P(C_i)$ [35, 46].

3 METODE ZA IZBOR ATRIBUTOV

Na področju strojnega učenja in odkrivanja vzorcev iz podatkov imajo osrednjo vlogo atributi, zaradi česar se je smiselno spraševati tudi o njihovi kvaliteti oz. pomembnosti za dani problem [96]. Skrbno izbrani atributi iz nabora podatkov igrajo pomembno vlogo pri klasifikaciji, saj lahko vhodne primere učinkoviteje umestijo v razrede. Na veliko področjih nabor podatkov postaja vedno večji. Iz teh večdimenzionalnih podatkov želimo pridobiti kar se da veliko koristnih informacij, kar pa ni enostavna naloga. Z algoritmi za izbor atributov lahko izločimo odvečne (redundantne) in nerelevantne atribute iz originalnega nabora podatkov. Redundantni atributi lahko povzročijo neželene situacije pri tradicionalnem učnem procesu (brez izbire spremenljivk) kot na primer preveliko prileganje podatkom ('over-fitting'), slaba napovedna uspešnost, slabša učinkovitost (npr. hitrost izvedbe klasifikacijskega algoritma), itd. Iz praktičnih razlogov je potrebno zreducirati atribute v večdimenzionalnih podatkih.

3.1 Atributna predstavitev učnih primerov

Najpogosteje se pri klasifikacijskih in regresijskih problemih uporablja atributna predstavitev učnih primerov. Atribut je spremenljivka, ki ima določeno množico možnih vrednosti. Atributom pravimo včasih tudi značilke ('attribute', 'feature'). Vsak učni primer je opisan z vektorjem vrednosti atributov. Atributi so lahko zvezni ali diskretni in njihovo število je dano vnaprej. Atributno predstavitev definiramo z:

- množico atributov $A = \{A_i, i = 0, \ldots, a\}$;
- za vsak diskretni atribut A_i imamo množico možnih vrednosti $\mathcal{V}_i = \{V_{1_i}, \ldots, V_{n_i}\}$;
- za vsak zvezni atribut A_i imamo interval možnih vrednosti $\mathcal{V}_i = [\min_i, \max_i]$;
- razred je podan z atributom A_0: če rešujemo klasifikacijski problem, potem je A_0 diskretni atribut, če pa rešujemo regresijski problem, je A_0 zvezni atribut;
- učni primer je vektor vrednosti atributov $u_j = \langle r^{(j)}, v^{(1,j)}, \ldots, v^{(a,j)} \rangle$, pri tem je razred označen z $r^{(j)} = v^{(0,j)}$;
- množica učnih primerov je podana kot množica vektorjev $\mathcal{U} = \{u_j, j = 1, \ldots, n\}$.

3.2 Lastnosti atributov in njihove soodvisnosti

Atributi imajo različne lastnosti, ki so pomembne pri uporabi atributov za strojno učenje:

- Šumen atribut: je skoraj vsak atribut v realnih podatkih.
- Pomanjkljiv atribut: je atribut, ki ima pri nekaterih učnih primerih manjkajoče vrednosti, ki jih mora učni algoritem pravilno upoštevati.

Poleg lastnosti posameznih atributov so pomembne tudi odvisnosti med atributi in razredom (odvisno spremenljivko):

- Močno soodvisni atributi glede na razred: Pri močnih soodvisnosti glede na razred je ciljno funkcijo težko odkriti, saj se posamezni atributi šele v kontekstu z ostalimi atributi pokažejo za pomembne.

- Relevanten atribut. Kohavi in John [44] sta podala enostavno in intuitivno definicijo relevantnosti.

  Definicija 1 (relevantnost)**.** *Naj bo S množica vseh atributov in naj bo $S_i = S - \{X_i\}$. Y naj predstavlja vrednosti razredov. Atribut X_i je relevanten natanko tedaj, kadar obstajajo vrednosti x_i, y in s_i ter $P(X_i = x_i, S_i = s_i) > 0$, da velja*

  $$P(Y = y|X_i = x_i, S_i = s_i) \neq P(Y = y|S_i = s_i), \forall i.$$

  Definicija 1 govori, da je atribut statistično relevanten, če njegova odstranitev iz množice atributov zmanjša napovedno moč. Atribut je statistično relevanten zaradi dveh razlogov: (1) je močno koreliran z vrednostmi razredov; ali (2) skupaj s podmnožico atributov je močno koreliran z vrednostmi razredov.

- Redundanten atribut: je atribut, katerega informacijo vsebuje že nek drug atribut ali množica atributov. Trivialen primer je dodaten atribut, ki je kopija obstoječega atributa ali pa atribut, koreliran z ostalimi, že izbranimi atributi.

  Definicija 2 (redundantnost)**.** *Naj bo S množica vseh atributov in naj bo $S_i = S - \{X_i\}$. Y naj predstavlja vrednosti razredov. Atribut X_i je redundanten natanko takrat, kadar obstaja takšna podmnožica atributov S'_i množice S_i in vrednosti x_i, y, s_i in s'_i ter $P(X_i = x_i, S_i = s'_i) > 0$ tako da velja*

  $$P(Y = y|X_i = x_i, S_i = s_i) = P(Y = y|S_i = s_i),$$

  in pri tem obstaja $S'_i \subset S_i$, da $P(Y = y|X_i = x_i, S_i = s_i) \neq P(Y = y|S'_i = s'_i), \forall i$.

  Dober atribut naj ne bi bil redundanten. Pri klasifikacijskih problemih na realnih podatkih relevanten atribut ni nujno tudi optimalen atribut za algoritem in obratno; nerelevanten atribut je lahko tudi v optimalni podmnožici atributov.

3.3 Preiskovalne strategije

Ker je problemski prostor vseh možnih podmnožic atributov ogromen, se ga običajno preiskuje s požrešnim algoritmom, ki ga vodijo ocene oz. napake obiskanih stanj. Najpogostejše tehnike preiskovanja so:

- Iskanje naprej (ang. 'forward selection'), kjer začnemo s prazno množico in dodajamo v vsakem nadaljnjem koraku po en ali naključno izbran ali najboljše ocenjen atribut, vse dokler se napaka še zmanjšuje.
- Vzvratno iskanje (ang. 'backward search'), kjer začnemo z vsemi atributi in jih postopoma odstranjujemo dokler napaka ne začne naraščati.
- Kombinirani pristop, kjer začnemo z naključno podmnožico atributov in v nadaljnjih korakih ali dodajamo nove atribute kot pri iskanju naprej ali odstranjujemo obstoječe kot pri vzvratnem iskanju.

3.4 Izbor atributov in filtrirne metode

Izbor atributov je primarno usmerjen na izbor relevantnih in informativnih atributov. Pri tem želimo:

1.) zmanjšati nabor atributov, s katerim zmanjšamo količino uporabljenega pomnilnika in povišamo hitrost izvedbe algoritmov;

2.) zmanjšati nabor atributov, s katerim želimo privarčevati z viri pri naslednjem zbiranju atributov ali pa privarčevati z viri med samo uporabo atributov pri sestavljanju modelov;

3.) izboljšati izvedbo oziroma napovedno natančnost;

4.) bolje razumeti podatke in boljšo predstavitev podatkov.

Algoritme za izbor atributov lahko razdelimo na tri velike razrede: **filtrirne metode** (ang. 'filter methods'), **ovojne** (ang. 'wrapper methods') in **vgrajene** (ang. 'embedded methods') metode. Med filtrirne metode štejemo tiste, kjer je korak izbire vhodnih atributov ločen od gradnje končnega modela.
Ovojne in vgrajene metode neposredno vključujejo gradnjo modela za ovrednotenje napovednih sposobnosti posameznega nabora vhodnih atributov. Ovojne metode za razvrstitev podmnožic atributov glede na njihovo napovedno moč uporabijo strojno učenje kot črno škatlo. Sistemu za analizo podatkov (prediktorju) zagotovijo podmnožico atributov, sprejmejo pa povratne informacije. Skozi optimizacijo ovrednotimo in primerjamo vse ali zgolj del vseh potencialnih kombinacij atributov. Kot optimalno podmnožico atributov izberemo tisto kombinacijo, s katero dobljeni model dosega najboljšo kvaliteto napovedi. Vgrajene metode izvedejo izbor atributov med samim procesom učenja in so specifične glede na dano strojno učenje. Filtrirne in ovojne metode se med seboj razlikujejo po ocenjevalnem kriteriju (ang. 'evaluation criterion'). Filtrirne metode ponavadi uporabljajo kriterije, ki ne vključujejo strojnega učenja, npr. indeks relevantnosti, ki je izračunan s pomočjo korelacijskih koeficientov ali testnih statistik, medtem ko pa ovojne metode vključijo izvedbo samega strojnega učenja na dani podmnožici atributov.
V našem raziskovalnem delu se bomo osredotočili le na filtrirne metode, saj so računsko manj zahtevne in primerne za analizo obsežnejših podatkov.

Filtrirne metode so hitre metode, z ocenjevalnimi metrikami ocenijo vsak atribut neodvisno od klasifikacijske metode napovedovanja, ki kasneje za vhodne podatke dobi podatke z zreduciranimi atributi [98]. Izločijo atribute, ki imajo majhno verjetnost uporabnosti pri analizi podatkov. Filtrirni algoritmi so računsko manj zahtevni [77] kot ovojne in vgrajene metode. Lahko jih razdelimo v 2 skupini: v skupino atributnega razvrščanja (ang. 'feature ranking methods'), ki ocenjuje atribute individualno, in multivariatno izbiranje atributov (ang. 'subset feature selection'), ki ocenjuje podmnožico atributov. Ocenjevanje na podmnožici atributov prinese veliko prednosti v primerjavi z atributnim razvrščanjem, predvsem pri doseganju višje napovedne moči. Pri velikem številu multivariatnih filtrirnih metod je velikokrat potrebno vnaprej določiti število atributov, ki nam jih vrne algoritem za izbor atributov. Filtrirne metode so velikokrat enačene kar z metodami za atributno razvrščanje. Te metode razvrstijo atribute s pomočjo indeksa relevantnosti. Vključujejo korelacijski koeficient, ki ocenjuje stopnjo odvisnosti posameznega atributa z razredom. Vključenih je veliko statistik, na primer klasični statistični testi: t-test, F-test, Chi-kvadrat, itd. **Indeks relevantnosti** $J(\mathcal{S}|\mathcal{D})$ pri danih podatkih $\mathcal{D}$ ocenjuje relevantnost dane podmnožice atributov $\mathcal{S}$ za nalogo Y (ponavadi je ta naloga razvrščanje ali uvrščanje). Ker so podatki in naloga ponavadi enaki in se spreminja samo podmnožica atributov $\mathcal{S}$, indeks relevantnosti zapišemo kot $J(\mathcal{S})$. Teoretično je težko določiti, katere filtrirne metode so primerne za izbran model.

Indeksi relevantnosti so izračunani na posameznih atributih $X_i, i = 1, \ldots, N$. Uredimo jih v vrstnem redu: $J(X_{i_1}) \leq J(X_{i_2}) \ldots \leq J(X_{i_N})$. Atribute, ki imajo najnižjo razvrstitev, odstranimo. Pri neodvisnih atributih je tak pristop zadosten, v primeru, da so atributi medseboj korelirani, je lahko veliko pomembnih atributov redundantnih. Z razvrščanjem ne dobimo nujno najboljše podmnožice pomembnih atributov. Vrednost indeksa relevantnosti bi morala biti pozitivno korelirana z natančnostjo prediktorja, ki je naučen na problemu Y, na podatkih $\mathcal{D}$ in na podmnožici atributov $\mathcal{S}$. Obstaja nekaj eksperimentalnega dela, kjer se ugotavlja ujemanje filtrirnih metod s klasifikacijskimi modeli. Verjetno se različni tipi filtrirnih metod ujemajo z različnimi tipi prediktorjev, kar pa ni podprto s teoretičnimi argumenti.

Možen pristop za izbor atributov je razvrstiti atribute glede na lastno relevatnost (univariatne metode). Takšne metode so hitre in efektivne, še posebej, kadar je število vseh atributov veliko in je število učnih podatkov v primerjavi z njimi majhno. Obstaja pa kar nekaj omejitev pri individualnem atributnem razvrščanju:

- atributi, ki posamično niso relevantni, lahko postanejo relevantni v kontekstu z ostalimi atributi;
- atributi, ki so posamično relevantni niso nujno tudi uporabni, saj je možno, da so odvečni (redundantni) v kombinaciji z ostalimi.

Multivariatne metode vzamejo v vednost tudi odvisnost med atributi. Dosežejo boljše rezultate, saj ne predpostavijo, da so atributi med seboj neodvisni. Multivariatne metode odkrivajo redundantne atribute [33]. V našem raziskovalnem delu se bomo osredotočili le na multivariatne filtrirne metode.

3.5 Multivariatne filtrirne metode

Atribut je 'dober', če je relevanten glede na razred, hkrati pa ni redundanten glede na ostale atribute. Če vzamemo korelacijo med dvema atributoma, potem po zgornji definiciji dobimo, da je atribut 'dober', če je visoko koreliran z razredom in ni koreliran z ostalimi atributi. Z drugimi besedami, korelacija med atributom in razredom mora biti dovolj visoka in korelacija med tem atributom in ostalimi atributi ne preseže neke vnaprej predpisane meje. Pri izbiri atributov je pomembno poiskati primerno mero korelacije med atributi [100]. Najbolj znan pristop, ki meri korelacijo med dvema atributoma, je osnovan na **klasični linearni korelaciji**, obstajajo pa tudi ostale mere, ki so variacije le-te, kot npr. napaka po metodi najmanjših kvadratov, vzajemna informacija, informacijski prispevek itd.

V sledečih vrsticah so predstavljene različne multivariatne filtrirne metode. FOCUS [7] uporablja mero konsistentčnosti in ocenjuje vse možne podmnožice atributov. Mera konsistenčnosti [18] poskuša ohraniti reprezetativnost podatkov. Metoda išče najmanjšo podmnožico atributov, ki razlikuje med razredi enako dobro, kot če bi uporabili celotno množico atributov. FCBF [100] je hitra filtrirna metoda, ki ocenjuje relevantnost med atributi s pomočjo simetrične negotovosti (ang. 'Symmetric uncertainty') in odstrani nerelevantne atribute z uporabo aproksimativne Markovske ovojnice (ang. 'Markov blanket'). mRMR maksimizira relevantnost z vnaprej podanimi razredi in minimizira redundantnost izbrane podmnožice atributov z uporabo Pearsonovega korelacijskega koeficienta [21] ali različica metode, ki uporablja oceno vzajemne informacije (ang. 'mutual information') [72]. CFS optimalne atribute izbira s pomočjo simetrične negotovosti [34]. Podrobneje bomo predstavili metode FCBF, CFS, mRMR in CCCA, saj jih bomo uporabili v eksperimentalnem delu.

FCBF [103] (ang. 'Fast Correlation Based Filter') je algoritem za izbor atributov in meri korelacije med atributom in razredom ter med atributi. FCBF začne iskanje atributov na množici atributov $S^{'}$, ki so visoko korelirani z razredom in je vrednost $SU \geq \delta$, pri nekem danem δ, kjer je SU simetrična negotovost (ang. 'symmetrical uncertainty') (glej Enačbo 2). Atribut f_i s simetrično negotovostjo $SU_{i,c}$ se imenuje **predominanten** natanko tedaj, ko je $SU_{i,c} > \delta$ in ne obstaja f_j, da je $SU_{j,i} \geq SU_{i,c} \quad \forall f_j \in S^{'}$, kjer $(j \neq i)$. Če obstaja atribut f_j kjer velja $SU_{j,i} \geq SU_{i,c}$, potem rečemo, da je f_j redundanten atribut glede na atribut f_i. Množico redundantnih atributov označimo s S_{P_i}, ki ga lahko nadalje razdelimo na $S^{+}_{P_i}$ in na $S^{-}_{P_i}$, ti vsebujejo redundantne atribute glede na atribut f_i, kjer velja $SU_{j,c} > SU_{i,c}$ in $SU_{j,c} \leq SU_{i,c}$. FCBF odstrani redundantne atribute in obdrži atribute, ki so najbolj relevantni glede na razred. Simetrična negotovost je definirana kot:

$$SU(X,Y) = 2\frac{IG(X|Y)}{H(X) + H(Y)}, \tag{2}$$

kjer $IG(X|Y)$, $H(X)$ in $H(X|Y)$ pomenijo Informacijski prispevek atributa (ang. 'Information gain'), ter H entropijo. Metoda zagotovi efektiven način za obvladanje redundantnih atributov. Časovna zahtevnost algoritma je: $O(m \cdot n \cdot \log n)$, kjer je m število primerov in n število atributov.

CFS uporablja za oceno vrednosti spremenljivk korelacije:

$$\text{Merit}_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}. \tag{3}$$

Tukaj Merit_S predstavlja 'zaslužek' podmnožice atributov S, ki vsebujejo k atributov.
$\overline{r_{cf}} = \sum_{f_i \in S} \frac{1}{k} \sum(f_i, c)$ je povprečje korelacij atribut–razred in $\overline{r_{ff}}$ je povprečje med atributi. Povprečje korelacij atribut–razred (števec v enačbi(3)) je mera, ki nam pove, kako zlahka lahko napovemo razred na podlagi atributa. Povprečje korelacij atribut–atribut (imenovalec) nam pove korelacijo med atributi in kažejo na stopnjo redundantnosti med njimi. Korelacije med atributi so ocenjene z uporabo informacijskega prispevka, ki določi stopnjo povezanosti med atributi. Velikokrat informacijski prispevek nadomesti simetrična negotovost (enačba (2)). CFS izračuna atribut-razred in atribut-atribut korelacije z uporabo simetrične negotovosti in izbere podmnožico atributov z uporabo metode preiskovanja 'najprej najboljši' (ang. Best First search). Prednosti metode CFS so, da ta deluje dobro na manjših množicah podatkov in izbira relevantne atribute ter se izogiba izboru redundantnim atributom [103].
Minimum-Redundancy-Maximum-Relevance (mRMR) izbira atribute na tak način, da imajo ti veliko medsebojno razdaljo, vendar pa imajo še vedno 'visoko' korelacijo z vrednostmi razredov.
Minimizacija Redundantnosti

$$\text{Za diskretne atribute: } \min W_I, W_I = \frac{1}{|S|^2} \sum_{i,j \in S} I(i,j), \tag{4}$$

$$\text{za zvezne atribute: } \min W_c, W_c = \frac{1}{|S|^2} \sum_{i,j \in S} c(i,j). \tag{5}$$

Maximizacija relevantnosti

$$\text{Za diskretne atribute: } \max V_I, V_I = \frac{1}{|S|} \sum_{i \in S} I(i,h), \tag{6}$$

$$\text{za zvezne atribute: } \max V_F, V_F = \frac{1}{|S|} \sum_{i \in S} F(i,h). \tag{7}$$

kjer je S množica atributov,
$I(i,j)$ je vzajemna informacija med atributoma i in j,
$c(i,j)$ je korelacija med atributoma i in j,
h je vektor vrednosti razredov,
$F(i,h)$ je $F-$ statistika.
$I(S,h)$, predstavlja vzajemno informacijo med izbranimi atributi S in vrednostmi razredov h:

- če imamo dve univariatni spremenljivki x in y, je vzajemna informacija definirana kot:

$$I(x;y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy.$$

- če imamo multivariatne spremenljivke S_m in vektor vrednosti razredov h, je vzajemna informacija definirana kot:

$$I(S_m; h) = \iint p(S_m, h) \log \frac{p(S_m, h)}{p(S_m)p(h)} dS_m dh$$

mRMR vključi oba optimizacijska pogoja v eno samo kriterijsko funkcijo. Če obravnavamo oba pogoja kot da sta si enakovredna, potem je najlažja kombinacija:

$$\max(V_I - W_I) \tag{8}$$

$$\max(V_I / W_I). \tag{9}$$

Correlation Coefficient Clustering Algorithm ('CCCA')
Avtorja Hsu in Hsieh [38] sta opisala metodo, kjer sta uporabila tako imenovani 'feature clustering'. Svoj algoritem za izbor atributov sta poimenovala 'Correlation Coefficient Clustering Algorithm'. Avtorja pri metodi za razvrščanje uporabita korelacijski koeficient kot distančno mero namesto tipične standardne evklidske razdalje. Metoda za razvrščanje razvrsti atribute v skupine tako, da so atributi znotraj skupin čim bolj podobni med seboj glede na korelacijski koeficient in atributi različnih skupin kar čim bolj različni med seboj oziroma, imajo nizko korelacijo. Atributi znotraj skupine so najbolj korelirani med seboj, zato iz vsake skupine metoda izbere le po en atribut. Avtorja predlagata izbor atributov, ki so najbolj korelirani z vrednostmi razredov, saj s takšnimi atributi lahko najbolj izboljšajo klasifikacijsko natančnost.

3.6 Predlagan algoritem za izbiro atributov

Relevantnost atributov lahko ocenjujemo s filtrirnimi metodami, z ocenjevalnimi metrikami, ki jih dobimo iz samega podatkovja, kot z npr. Fisherjevo oceno (ang. 'Fisher score') [17, 30], χ^2-testom [32], vzajemno informacijo [21, 72, 102], simetrično negotovostjo ('Symmetric uncertainty') [34, 100], Hilbert-Schmidtovim operatorjem [83], z uteževanjem atributov na podlagi razdalj med enotami (ReliefF) [54], itd. Nekateri avtorji kot ocenjevalno metriko pri izboru atributov uporabljajo tudi razvrščanje v skupine. Avtorji Liu et al. [62] s pomočjo karakteristik metode razvrščanja v skupine izbirajo podmnožice atributov. Njihova razvrščevalna metoda deluje kot nadzorovano učenje, kjer je vektor podanih razredov obravnavan kot posebna skupina ('cluster') in je uporabljen kot vodilo pri procesu razvrščanja. Njihova evaluacijska metrika za izbiro atributov vključuje vzajemno informacijo (ang. 'mutual information' (MI)). Klasične metode razvrščanja zlahka pretransformiramo v razvrščanje atributov namesto razvrščanje enotv, zato veliko avtorjev poskuša izvesti izbor atributov s pomočjo algoritmov za razvrščanje (ang. 'clustering of features') [13, 37, 38, 56, 61, 83]. Avtorji najprej zgrupirajo atribute v različne skupine po nekem kriteriju in kasneje sestavijo podmnožico atributov z reprezentativnimi atributi iz vsake skupine.

Razvrščanje enot v skupine tako, da so si objekti znotraj skupin kar čim bolj podobni in objekti različnih skupin kar čim bolj različni med seboj, je zelo star, intuitivno preprost in razumljiv problem. Razvrščanje enot ponavadi izvedemo, kadar ni informacij o tem, v kateri razred spada dana enota. Zaradi slednjega spada razvrščanje v skupine med nenadzorovano učenje, pri katerem so velikokrat zahtevani dodatni parametri, potrebni za izvedbo algoritmov za razvrščanje. Na primer, določiti število skupin k v katere bomo razvrstili enote, je eden izmed najtežjih problemov na tem področju. Veliko avtorjev se je ukvarjalo z različnimi tehnikami za določitev optimalnega k [28, 39, 78, 91, 94].
Obstajajo različne metode razvrščanja. V našem raziskovalnem delu smo uporabili nekaj znanih najbolj reprezenatativnih metod za razvrščanje v skupine. V naslednjih poglavjih bomo predstavili novo multivariatno filtrirno metodo, ki smo jo poimenovali FSuC (ang. 'Feature Selection using Clustering'). Glavna ideja predlagane metode leži v iskalni proceduri, kjer uporabimo metodo za razvrščanje v skupine, pri čemer je število vnaprej določenih skupin k enako številu razredov pri uvrščanju (klasifikaciji). Metodo za razvrščanje v skupine bomo v tem delu določili eksperimentalno.

Učinkovitost izbranih atributov je izmerjena s klasifikacijskimi merami (tako da primerjamo skupine, dobljene z razvrščanjem z vrednostjo razredov, ki jim pripada posamezna enota), kar pomeni, da vrednosti razredov služijo kot vodilo pri procesu razvrščanja. Proces iskanja atributov ponavljamo tako dolgo, dokler se klasifikacijska natančnost izboljšuje.

V predlagani metodi za izbor atributov smo hoteli, tako kot pri metodah za razvrščanje v skupine, razvrstiti podatke v skupine ne glede na to, kakšne so vrednosti razredov (katera informacija le teh nam je znana). Obe vrednosti za posamezno enoto nato primerjamo (glej 3.6.1). Spodaj smo podali natančen opis predlagane metode (glej algoritem 3), ki dodaja atribute v seznam F tako dolgo, dokler ne zadosti zaustavitvenem pogoju.

Ker je število vseh možnih podmnožic atributov običajno preveliko ($2^{|\text{atributi}|}$), se prostor vseh možnih atributov običajno preiskuje s požrešnim algoritmom, ki ga vodijo ocene obiskanih stanj. Uporabljena preiskovalna strategija je 'Iskanje naprej' (ang. 'forward selection'), kjer začnemo s prazno množico in dodajamo v vsakem nadaljnem koraku najbolje ocenjen atribut vse dokler se natančnost multivariatnega razvrščanja v skupine zvišuje.

V prvem koraku algoritma izvedemo univariatno razvrščanje v k skupin na vsakem atributu posebej. Izberemo atribut f_1, na katerem so skupine univariatnega razvrščanja najbolj podobne razvrščanju, ki je definiran z vrednostmi klasifikacijskih razredov. Atribut shranimo v seznam F. Podobnost v našem primeru izmerimo z uspešnostjo uvrstitve A_1, ki jo izračunamo z uporabo nekaj mer za vrednotenje (glej poglavje 3.6.1). Na tem koraku izvedemo n krat metodo za razvrščanje v skupine, kjer je n število atributov. Naslednji korak je bivariatno razvrščanje v skupine, ki za vhodne podatke uporabi izbran atribut v prejšnjem koraku v kombinaciji z ostalimi atributi (v tem koraku izvedemo $n-1$ krat metodo voditeljev oz. razvrščanje v skupine). Izberemo atribut f_2, s katerim dosežemo najvišjo uspešnost razvrstitve A_2 v kombinaciji s prej izbranim atributom f_1. Atribut f_2 dodamo v seznam atributov F, če klasifikacijska natančnost presega A_1. Postopek ponavljamo, dokler se uspešnost uvrstitve zvišuje. Algoritem vrne podmnožico atributov, shranjeno v seznamu F. Časovna zahtevnost predlaganega algoritma je

Algorithm 3 FSuC algoritem

Inputs: Množica podatkov D z atributi $\mathcal{S}$

$\mathcal{M}(\in$ {OSR, povprečje senzitivnosti, F–mera$\})$ // uspešnost razvrstitve

Inicializacija: $F = \{\}$, $S = \mathcal{S}$, $C_L =$ skupine, ki jih določa vrednost razreda

1.) $f_1 \leftarrow \arg\max_{f \in S}\{\mathcal{M}$(skupine dobljene z univariatnim razvrščanjem z vhodnimi podatki $f, C_L)\}$

2.) $i \leftarrow 1$; $A_1 \leftarrow$ uspešnost uvrstitve, kadar uporabimo atribut $\{f_1\}$

3.)

repeat

3.1) $F \leftarrow F \cup \{f_i\}$

3.2) $S \leftarrow S - \{f_i\}$

3.3) $i \leftarrow i + 1$; če $i > |\mathcal{S}|$, končaj;

3.4) $f_i \leftarrow \arg\max_{f \in S}\{\mathcal{M}$(skupine dobljene z multivariatnim razvrščanjem z vhodnimi podatki $F \cup \{f\}, C_L)\}$

3.5) $A_i \leftarrow$ uspešnost uvrstitve dosežena, kadar uporabimo atribute $F \cup \{f_i\}$

until $A_i < A_{i-1}$ **or** $i > |\mathcal{S}|$

Output: Izbrana podmnožica atributov F.

polinomska.

3.6.1 Metodologija vrednotenja

Pri procesu izbiranja atributov potrebujemo mero, s katero ovrednotimo doprinos atributa f_i k seznamu atributov F (glej algoritem 3). Matrika razvrstitev (ang. 'confusion matrix') (glej tabelo 1) vsebuje informacijo o pravilnih in napačnih uvrstitvah. Za mero uspešnosti uvrstitve vključimo več različnih mer, predvem zaradi različne narave ocenjevanja. Uporabimo 3 različne mere, ki so primerne tudi za večrazredne podatke (ne samo za 2 razreda) [58].

	C_1	$\dots$	C_k
$\hat{C}_1$	p_{11}	$\cdots$	p_{1k}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\hat{C}_k$	p_{k1}	$\cdots$	p_{kk}

Tabela 1: Matrika razvrstitev za splošni primer. Oznaka p_{ij} pomeni verjetnost, da smo enoto uvrstili v razred i ($\hat{C}_i$), medtem ko v resnici pripada razredu j (C_j). Vsoto elementov matrike razvrstitev po vrstici i ali stolpcu j označimo z p_{i+} ali p_{+j}.

V implementaciji algoritma FSuC se pojavijo 3 različne možne mere za merjenje uspešnosti uvrstitve. Spodaj podajamo opis za vsako od teh mer. Vrednost k naj v nadaljnjem pomeni oznako za število skupin/razredov.

1.) Splošna stopnja uspeha

Najbolj znana in enostavna mera pri merjenju uspešnosti uvrstitve je splošna stopnja uspeha (ang. 'overall success rate' (OSR)), ki je definirana kot sled matrike razvrstitev (glej tabelo 1):

$$\text{OSR} = \sum_{i=1}^{k} p_{ii}.$$

2.) Povprečje senzitivnosti:

$$\frac{1}{k}\sum_{i=1}^{k}\frac{p_{ii}}{p_{+i}}.$$

3.) F–mera

Naslednja mera je F–mera, ki pripada harmoničnemu povprečju PPV (pozitivne napovedne vrednosti ali preciznosti) in TPR (senzitivnosti),

$$F_i = \frac{2PPV_i \times TPR_i}{PPV_i + TPR_i},$$

kjer je $TPR_i = \frac{p_{ii}}{p_{+i}}$ in $PPV_i = \frac{p_{ii}}{p_{i+}}$. V našem eksperimentalnem delu uporabimo $F = \frac{1}{k}\sum_{i=1}^{k} F_i$.

Klasifikacijski rezultati so odvisni od izbire evaluacijskih mer. V nadaljnjem oznake 'FSuC–1', 'FSuC–2', 'FSuC–3' pomenijo 3 različne FSuC metode z vključeno splošno stopnjo uspeha, povprečje senzitivnosti in F-mero.

V članku [70], kjer smo predlagali metodo 'FSuC', ta dobro deluje tudi na 15 izbranih podatkih z različnih področij (glej tabelo 2), kjer smo klasifikacijsko točnost testirali z LDA, NB in 1–kNN klasifikatorji. V članku smo uporabili le metodo voditeljev, saj se je izkazala za najučinkovitejšo metodo v splošnem za izbran nabor podatkov. Zavedamo se, da ima vsako podatkovje svoje značilnosti, kar pomeni tudi za vsako podatkovje posebej določiti metodo za razvrščanje v skupine, zato bi bilo potrebno članek izboljšati tako, da bi te metode prilagodili glede na izbrane podatke. Hevristično algoritem 'FSuC' smo primerjali z nekaj širše uporabljanimi filtrirnimi metodami na 15ih večrazsežnih podatkih, ki jih raziskovalci uporabljajo za empirične analize algoritmov na področju strojnega učenja in so prosto dostopni iz repozitorijev [2, 9]. Podatki vsebujejo različno število enot, razredov in atributov. Atributna predstavitev je bodisi diskretna ali zvezna in vsi atributi so skalirani s povprečjem 0 in standardnim odklonom z vrednostjo 1. Tabela 2 prikazuje lastnosti izbranih podatkov. V članku smo predlagano metodo primerjali s tremi različnimi tipi reprezentativnih filtrirnih algoritmov, ki so: FCBF, CFS in mRMR. Eksperimentalni rezultati so pokazali, da so 'FSuC' metode superiorne v primerjavi z ostalimi reprezentativnimi metodami. Delež izbranih atributov drastično zmanjša velikost dimenzij podatkov, kar prispeva k hitrejši izvedbi klasifikatorjev.

št	podatki	# enot	# atributov	# razredov
1	Image Segmentation	2310	19	7
2	Breast Cancer Wisconsin (wdbc)	569	32	2
3	Ionosphere	351	34	2
4	Soybean (Small)	47	35	4
5	Statlog (Landsat Sattelite)	6435	36	6
6	Tennis Major Tournament Match Statistics	951	42	2
7	QSAR biodegradation	1055	41	2
8	Musk (Version 1)	476	168	2
9	Musk (Version 2)	6598	168	2
10	Mfeat–factors	2000	216	10
11	LSVT Voice Rehabilitation	126	309	2
12	Madelon	2000	500	2
13	Human Activity Recognition (HAR)	10299	561	6
14	Colon	62	2000	2
15	Leukemia	72	7130	2

Tabela 2: Opis podatkov v eksperimentalnem delu (v članku).

3.7 Opis podatkov

3.7.1 Tehnična in temeljna analiza

Tehnični analitiki ne verjamejo, da se tečaji delnic gibajo naključno, ampak da med gibanjem tečajev v preteklosti in med tistim, kar se bo zgodilo v prihodnosti, obstaja direktna povezava. Njihov cilj je, da določijo, kakšna je ta medsebojna povezava z željo pravilno napovedati, v katero smer se bo gibal trg ali gibala cena posamezne delnice.
Pri tehnični analizi so pomembni naslednji podatki:

- otvoritveni dnevni tečaj - tečaj delnice pri prvem sklenjenem poslu v trgovalnem dnevu (open),
- zaključni dnevni tečaj - tečaj delnice pri zadnjem sklenjenem poslu v trgovalnem dnevu (close),
- najvišji dnevni tečaj - najvišji tečaj v trgovalnem dnevu, ki ga je dosegla delnica (high),
- najnižji dnevni tečaj - najnižji tečaj v trgovalnem dnevu, ki ga je dosegla delnica (low),
- dnevni promet - dosežen promet z delnico v trgovalnem dnevu (volume).

Delniški trg lahko analiziramo s pomočjo tehnične analize z uporabo vzorcev tečajev/cen iz preteklosti in temeljne (fundamentalne) analize, ki jemlje svoje temelje v klasični ekonomski teoriji. Temeljna analiza si prizadeva napovedati dejansko vrednost naložbe. Izhaja iz teorije, da ima tržna cena delnice tendenco, da se vrne k njeni dejanski vrednosti. S temeljno analizo preučujemo poslovanje podjetja. S pomočjo letnih poročil lahko ugotovimo pomembne kazalnike zadolženosti ter uspešnosti poslovanja.

Temelji na informacijah o finančni moči podjetja, podatkih o preteklih izplačanih dividendah, stopnji rasti celotnega prihodka v prihodnosti, značilnosti panoge, v kateri podjetje posluje, značilnosti njegove konkurence in splošnih ekonomskih in političnih razmerah v državi in svetu [69]. Torej ocenjujemo vrednost delnice, ki jo izračunamo na podlagi (predvsem bilančnih) podatkov o podjetju [87]. Temeljna analiza ima tudi svoje pomanjkljivosti. Skupni problem poročil podjetja je uporaba zadnjih objavljenih podatkov, ki so lahko stari tudi mesec ali več, zato prikazujejo preteklo stanje podjetja in ne prihodnje. Ker so podatki stari, torej javnosti znani, temeljni analitik zaostaja za trgom, kar ga na trgu postavi v podrejen položaj. Naslednja slabost temeljne analize je, da se izkazi poslovnega izida lahko prirejajo po lastnih željah, torej dobički in drugi kazalci uspešnosti niso nujno kakovostna slika uspešnosti [104]. Temeljna analiza je primerna predvsem za dolgoročne investicije.
Tehnična analiza ali branje grafov je način analiziranja finančnih instrumentov z uporabo vzorcev cen iz preteklosti [14]. Za razliko od temeljne analize lahko s pomočjo tehnične analize razložimo dogodke, kot so padec cene podcenjene delnice ali porast cene delnice podjetja, ki je tik pred stečajem. V večini primerov imamo pri tehnični analizi veliko opravka s človeško psihologijo. Grafični vzorci prikazujejo gibanje cene skozi čas in kažejo naraščajočo ('bullish') ali padajočo ('bearish') psihološko razpoloženje na trgu. Sklepamo lahko, da tehnična analiza temelji na študiji človeškega obnašanja, za katerega je značilno, da se s časom ne spreminja veliko. Zato tehnični analitiki verjamejo, da je prihodnost zgolj ponavljajoča se zgodovina, oziroma, da ključ do poznavanja prihodnjega gibanja cen delnic leži v preučevanju preteklega gibanja cen delnic. Delo temeljnih analitikov zahteva večjo specializacijo na določena podjetja in mnogo več podatkov, ki jih morajo zajeti v analizi, kar vzame kar nekaj časa. Tehnični analitik mnogo lažje vidi celotno sliko. Tehnični analitiki poudarjajo, da temeljni analitiki lahko dosegajo nadpovprečne donose, če imajo nove informacije pred ostalimi investitorji in če je proces dosleden in hiter. Trdijo, da je prednost njihove metode v tem, da analiza ni odvisna od računovodskih izkazov, ki so lahko velikokrat prirejeni. Za razvoj sistema za avtomatsko trgovanje z delnicami je najenostavneje vključiti tehnične indikatorje, ki opisujejo stanje delnice na delniškem trgu.

3.7.2 Tehnični indikatorji

Tehnični indikatorji so serije podatkov, izpeljani iz vrednostnega papirja s pomočjo formule. Podatki vrednostnih papirjev, ki so uporabljeni pri izračunu, so lahko tečaj ob odprtju, zaprtju, najvišji ali najnižji tečaj v določenem časovnem obdobju. Njihova funkcija je opozarjanje, potrjevanje in predvidevanje cenovnih sprememb. Pomembna je njihova uporaba hkrati z ostalimi orodji tehnične analize.
Zaradi velikega števila indikatorjev, se moramo vprašati, kateri so dovolj kvalitetni za uporabo [12]. Pri vključevanju indikatorjev v sam sistem za avtomatsko trgovanje z delnicami se je potrebno odločiti o nekaterih parametrih, ki določajo posamezen tehničen indikator. Ti parametri so: **kategorija/vrsta tehničnega indikatorja**, **število časovnih enot, zajetih v izračun** oziroma koliko preteklih dni bomo vključili v sam izračun tehničnega indikatorja ter **število indikatorjev**. Kadar izbiramo kategorijo tehničnega indikatorja, se moramo zavedati problema multikolinearnosti, ki pomeni preveliko korelacijo med posameznimi indikatorji in se zgodi, kadar indikatorje iste kategorije vključimo v trgovalno

strategijo/model. Rezultat multikolinearnosti so redundantni signali, ki so lahko zavajajoči. Nekateri trgovalci namensko uporabijo več indikatorjev iste kategorije v upanju, da bodo tako poiskali potrditev pri gibanju cene delnice. V resnici pa multikolinearnost lahko povzroči, da nekateri indikatorji v prisotnosti drugih izpadejo manj pomembni in si lahko otežimo analizo delniškega trga [40]. Število časovnih enot, zajetih v izračun tehničnega indikatorja, je eden izmed glavnih parametrov indikatorja in služi kot vhodni podatek pri večini indikatorjev [67]. Velik nabor indikatorjev lahko razloži gibanje cene delnice, zato različne študije/raziskave izbirajo različne tehnične indikatorje kot vhodni podatek za konstrukcijo modela. Obstaja ogromno vseh možnih kombinacij pri izboru parametrov. Atsalakis in Valavanis [8] sta v svojo raziskavo vključila 100 znanstvenih člankov, ki napovedujejo gibanje delniških trgov. Ugotovila sta, da je povprečno število indikatorjev, ki se pojavlja v raziskavah med 4 in 10. Velika večina raziskovalcev vključuje le tehnične indikatorje, s katerimi poskušajo napovedati dnevna ali tedenska gibanja delniških trgov [24, 40, 43, 48, 59, 60, 99]. Po pregledu literature smo ugotovili, da je objavljenih le malo raziskav o tem, kateri so najprimernejši parametri tehničnih indikatorjev, ki se bodo uporabili pri napovednih modelih in trgovalnih strategijah ali kateri so najpomembnejši indikatorji, zato smo velik del raziskave namenili le tej problematiki.

Različne študije različno izbirajo vhodne podatke za konstrukcijo modela. Glavna ideja za uspešno napoved gibanja finančnih trgov je doseči čim višje rezultate s čim manj vhodnimi podatki in čim manj kompleksnim modelom. Veliko podobnih raziskovalnih del vsebuje le en tip indikatorjev (bodisi tehnični tip indikatorjev bodisi fundamentalni tip indikatorjev). Avtorji Vanstone et. al [90] v svojem delu sklenejo, da naj bi napovedni modeli, ki so zgrajeni za daljše časovno obdobje, uporabili fundamentalne indikatorje, medtem ko za modele, ki so zgrajeni za krajše časovno obdobje, kot za npr. dnevne napovedi, naj bi vključevali tehnične indikatorje. Pri izgradnji modelov za vhodne podatke uporabimo 98 tehničnih indikatorjev, ki so izračunani na časovnih vrstah (uporaba otvoritvenih, zaključnih, najvišjih, najnižjih dnevnih tečajev in prilagojenih zaključnih tečajih (ang. 'adjusted values')). S tem želimo tudi preveriti, ali časovne vrste na preteklih cenah delnic vsebujejo kakšne informacije, ki so uporabne za napoved prihodnjih gibanj najvišjih dnevnih tečajev. Za izračun tehničnih indikatorjev smo uporabili $n = 2, 3, 5, 10, 20, 40, 80, 150$ število časovnih enot (ang. 'indicator length' ali 'lag length'). Krajše obdobje n povzroči več signalov za trgovanje kot daljše časovno obdobje in je zato primerneje za trgovanje na krajše časovno obdobje. Domnevamo, da nam bodo metode za izbor indikatorjev vrnile manjši n iz posamezne kategorije.

Tabela 3: Opis in definicija tehničnih indikatorjev

Ime indikatorja	Formula
CCI (ang. 'Commodity Channel Index')	$\frac{M_t - SM_t}{0.015 D_t}$, kjer je $M_t : \frac{H_t + L_t + C_t}{3}$, $SM_t : \frac{\sum_{i=1}^{n} M_{t-i+1}}{n}$ in $D_t = (\sum_{i=1}^{n} \lvert M_{t-i+1} - SM_t \rvert)/n$
SMA (ang. 'Simple n-day moving average')	$\frac{C_t + C_{t-1} + \ldots + C_{t-n}}{n}$
EMA (ang. 'Exponential n-day moving average')	$\text{EMA}_{t-1} + \frac{2}{n+1} \cdot (C_t - \text{EMA}_{t-1})$
ZLEMA (ang. 'Zero lag exponential n-day moving average')	$\frac{2}{n+1} \cdot (2 \cdot C_t - C_{t - \frac{n-1}{2}}) + (1 - \frac{2}{n+1} \cdot ZLEMA_{t-1})$
WMA (ang. 'Weighted n-day moving average')	$\frac{n \cdot C_t + (n-1) \cdot C_{t-1} + \ldots + C_{t-n}}{n + (n-1) + \ldots + 1}$
RSI (ang. 'Relative Strength Index')	$100 - \frac{100}{1 + \frac{\sum_{i=0}^{n-1} Up_{t-i}}{n} / \frac{\sum_{i=0}^{n-1} Dw_{t-i}}{n}}$
MACD (ang 'moving average convergence divergence')	$\text{MACD}(n)_{t-1} + \frac{2}{n+1} \cdot (\text{DIFF}_t - \text{MACD}(n)_{t-1})$ kjer je $\text{DIFF}_t : \text{EMA(12)}_t - \text{EMA(26)}_t$
Momentum	$C_t - C_{t-n}$
VHF (ang. 'Vertical Horizontal Filter')	$\frac{HH_{t-n} - LL_{t-n}}{\sum_{i=1}^{n} \frac{C_i - C_{i-1}}{C_{i-1}}}$ kjer sta LL_t in HH_t najnižji najnižji dnevni tečaj in najvišji najvišji dnevni tečaj zadnjih t dnevih.
ROC (ang. 'Rate of change')	$\frac{C_t - C_{t-n}}{C_{t-n}} \cdot 100$
SAR (ang. 'Stop and reverse')	$SAR_{t-1} + 0.02 \cdot (EP - SAR_{t-1})$
CMO (Chande Momentum Oscilator)	$100 \cdot \frac{Up - Dw}{Up + Dw}$
Williams Accumulation/Distribution	if $C_t > C_{t-1} : AD_t = AD_{t-1} + (C_t - \min(L_t, C_{t-1}))$ if $C_t < C_{t-1} : AD_t = AD_{t-1} + \max(H_t, C_{t-1} - C_t)$

Continued on next page

Tabela 3 – *Continued from previous page*

Ime indikatorja	Formula
	if $C_t = C_{t-1} : AD_t = AD_{t-1}$
ATR (Average true range)	$\frac{TR_{t-1} \cdot (n-1) + TR_t}{n}$
KST (Know Sure Thing)	RCMA1 = 10-Period SMA of 10-Period Rate-of-Change RCMA2 = 10-Period SMA of 15-Period Rate-of-Change RCMA3 = 10-Period SMA of 20-Period Rate-of-Change RCMA4 = 15-Period SMA of 30-Period Rate-of-Change KST = (RCMA1· 1) + (RCMA2 · 2) + (RCMA3· 3) + (RCMA4 · 4)
Donchian Channel	Upper Band = $\max(H_t, n)$ Lower Band = $\min(L_t, n)$ Middle Band = (Upper Band + Lower Band) / 2
Bollinger Bands	Middle Band = 20-day simple moving average (SMA) Upper Band = 20-day SMA + (20-day standard deviation of price · 2) Lower Band = 20-day SMA - (20-day standard deviation of price · 2)

C_t je prilagojen zaključni dnevni tečaj, L_t je prilagojen najnižji dnevni tečaj, H_t je prilagojen najvišji dnevni tečaj v trgovalnem dnevu t in n je število časovnih enot (v našem primeru število trgovalnih dni). $Up = \sum_{i-n}^{i} Up_t$; $Dw = \sum_{i-n}^{i} Dw_t$, kjer je $Up_t = C_t - C_{t-1}$ in $Dw_t = 0$, če $C_t > C_{t-1}$ $Dw_t = C_{t-1} - C_t$ in $Up_t = 0$, če $C_t < C_{t-1}$; EP (ang. 'the extreme point') ekstremna točka je najvišja vrednost dosežena v pozitivnem trendu — ali najnižja vrednost dosežena v negativnem trendu. Na vsakem časovnem intervalu vsakič znova osvežimo vrednost EP.
TR_t : TrueHigh − TrueLow,
TrueHigh: $\max\{H_t, C_{t-1}\}$, TrueLow: $\min\{L_t, C_{t-1}\}$.
Za indikatorje CCI, SMA, EMA, ZLEMA, WMA, RSI, momentum, CMO, VHF, ROC in ATR uporabimo
$n = 2, 3, 5, 10, 20, 40, 80, 150$ število časovnih enot. Pri ostalih indikatorjih za izračun uporabimo privzeto število enot, zajetih v izračun. Kot dodaten tehničen indikator vzamemo še promet delnice.

Kratek opis uporabljenih tehničnih indikatorjev

- EMA, eksponentno drseče povprečje, je uteženo drseče povprečje, kjer imajo novejše vrednosti tečaja večjo težo od starejših. Teoretična podlaga takšnemu načinu je prepričanje, da novejše vrednosti boljše izražajo sedanje stanje kot starejše, še posebej kadar uporabljamo daljša obdobja.

- SMA, navadno drseče povprečje je najbolj pogosto med drsečimi povprečji. Dobimo ga tako, da izračunamo aritmetično sredino na preteklih n izbranih tečajev. Manjši kot je n (število časovnih enot zajetih v izračun), bolj odzivno je drseče povprečje. Tipično se za drseča povprečja uporabi $n = 5, 25$ preteklih tečajev.

- ZLEMA, brez zamika eksponentno drseče povprečje je tehničen indikator, podoben eksponentno drsečemu povprečju, ki pa daje večjo utež nedavnim tečajem. Podatkom dodatno odstrani kumulativni efekt tako, da od njih odšteje starejše podatke z $(n-1)/2$ zamikom.

- WMA, je uteženo drseče povprečje in je indikator, podoben eksponentnemu uteženemu povprečju, vendar uporablja linearno uteževanje.

- MACD, je zelo pomemben indikator, ki uporablja drseča povprečja (ang. 'Moving average convergence–divergence') in predstavlja razliko med 12–dnevnim in 26–dnevnim eksponentnim povprečjem.

- SAR, Paraboličen SAR (ang. 'The Parabolic Stop-and-Reverse') je 'stop-loss' sistem. SAR omogoča investitorjem, da relativno zgodaj ujamejo trende. Z njim lahko določimo signale za prodajo (kadar je tečaj nad vrednostjo SAR) ali za nakup (kadar tečaj pade pod vrednostjo SAR).

- mom, Moment je najosnovnejša oblika oscilatorja in meri hitrost spreminjanja cene. Izračuna se kot razlika med tečajem na koncu in tečajem na začetku določenega časovnega intervala. Za interpretacijo momenta je pomembna smer gibanja momenta. Če moment raste hkrati z naraščajočim tečajem, to pomeni, da pozitivni trend pridobiva na moči. Nasprotno padajoči moment ob naraščajočem tečaju pomeni, da pozitivni trend slabi. Trendi se večinoma ne obračajo nenadoma, zaradi česar je potrebno nedvomno spremljati, kaj se dogaja s trendom, ali ta slabi ali krepi. Oscilatorji so v splošnem namenjeni za generiranje nakupnih in prodajnih signalov, za kar je potrebno določiti prekupljeno in preprodano območje. Če se vrednost oscilatorja nahaja v prekupljenem območju, nam to signalizira, da je tečaj zrasel prehitro in je zrel za korekcijo. Obratno velja za preprodano območje.

- ROC, zelo pogosta variacija momenta nosi ime 'Rate of change' indikator, ki v nasprotju z momentom ne meri razlike, temveč zgolj razmerje med tečajem na začetku in na koncu časovnega intervala. Rezultati analize z ROC indikatorjem so zelo podobni kot pri momentu, tudi interpretacija ostaja enaka.

- RSI, je eden najpopularnejših oscilatorjev. RSI izračunamo kot $RSI = 1000 - 100/(1 + RS)$, kjer je RS razmerje med povprečjem porastov tečaja zadnjih n preteklih tečajev in povprečjem padcev tečaja zadnjih n preteklih tečajev. S pomočjo te enačbe dobimo oscilator, katerega vrednost se giba med 0 in 100. Na podlagi gibanja indikatorja lahko določimo prekupljeno ali preprodano območje. Ponavadi kot prekupljenost interpretiramo vrednosti nad 70, kot preprodano območje pa vrednosti pod 30. Ko vrednost RSI doseže prekupljeno ali preprodano območje, lahko pričakujemo, da je tečaj zrel za obrat [92].

- CCI, (ang. 'Commodity Channel Index') je indikator za odkritje trenda ali ekstremnih tržnih pogojev. Indikator domneva, da se tečaj delnice premika v cikličnih trendih z vrhi in najnižjimi vrednostmi. CCI valovi nad in pod ničlo in ga lahko uporabljamo za določitev preprodanosti ali prekupljenosti določenega vrednostnega papirja. Vrednostni papir je preprodan, če CCI pade pod -100, ter je prekupljen, če zraste nad 100 [73].

- VHF, (ang. 'Vertical Horizontal filter') indentificira začetek in konec trendov.

- CMO, (ang. 'Chande Momentum Oscilator') je Chande-ov oscilator zagona, ki meri hitrost spremembe cene. V formuli vključuje 'U_p' in 'D_w' vrednosti in tako določa precenjeno in podcenjeno območje. Ker je indikator osnovan na prejšnjih zaključnih dnevnih tečajih, bo njegova vrednost oscilirala med vrednostima +100 and -100. Tehnični analitiki uporabljajo splošno pravilo: ko CMO doseže vrednost vsaj 50, potem je vrednostni papir prekupljen, medtem ko vrednosti pod -50 pomenijo, da je vrednostni papir preprodan.

- Promet je število vrednostnih papirjev, ki so bili kupljeni/prodani znotraj trgovalnega dneva. Višji kot je promet, likvidnejši/aktivnejši je vrednostni papir. Promet je pomemben indikator v tehnični analizi saj z njim potrjujemo trende in iščemo vzorce. Vsako gibanje cene je veliko bolj podkrepljeno z višjim prometom kot pa z nižjim.

  Večina tehničnih indikatorjev upošteva ceno in količino, obstajajo pa tudi indikatorji, ki upoštevajo nihanje cen. Med najpogostejšimi sta Average True Range (ATR) in Bollingerjeva obroča.

- Bollinger Bands, Bolingerjeva obroča [88] omogočata uporabniku primerjati nihanje cene in relativno raven cen v določenem časovnem obdobju. Sestavljen je iz dveh obrob in ene sredinske črte, ki predstavlja drseče povprečje. Zgornja obroba predstavlja drseče povprečje s prištetima dvema standardnima odklonoma, spodnja pa drseče povprečje z odštetima dvema standardnima odklonoma. Smatra se, da so razmere precenjene, ko se cena dotakne zgornjega obroča in podcenjene, ko se dotakne spodnjega.

- ATR, (ang. 'Average True Range') je indikator, ki meri volatilnost, ki pa ni indikator za napovedovanje smeri trenda. Odraža zanimanje ali nezanimanje v gibanju cen oziroma nihanje cen.

- KST, (ang. 'Know Sure Thing') je oscilator in identificira glavne tržne cikle.

- Williams Accumulation/Distribution je indikator, ki poskuša oceniti ponudbe in povpraševanja z ugotavljanjem razlik med ceno in prometom in na podlagi teh interpretira signale za nakup/prodajo.

- Donchian Channel, je preprost kazalec, ki pravi, da kupujemo na trgu, ko cena preseže najvišjo vrednost zadnjih n trgovalnih dni (v tabeli označeno z 'Upper band') in prodajamo na trgu, ko cena pade pod najnižjo vrednostjo zadnjih n trgovalnih dni (v tabeli označeno z 'Lower band').

Tečaji so ponavadi zelo volatilni s precej nepomembnimi kratkoročnimi premiki, zaradi katerih je trend težko razviden. Najpreprostejša metoda za odpravo teh šumov so drseča povprečja, ki učinkovito zgladijo kratkoročne fluktuacije ([97]). Poznamo veliko vrst drsečih povprečij. V naši analizi bomo uporabili navadno (SMA), uteženo (WMA), eksponentno drseče povprečje (EMA) in ZLEMA. Poglobljeno analizo izbranih indikatorjev lahko najdemo v [6].

4 TRGOVALNE STRATEGIJE

4.1 Opis trgovalne strategije

Trgovalna strategija je skupek enostavnih pravil, ki določajo, kdaj določen vrednostni papir kupiti in kdaj prodati.
Uspešnost napovedi smeri gibanja posameznih delnic lahko preverimo tako, da z njimi trgujemo, zato smo v našem eksperimentalnem delu sestavili trgovalno strategijo, ki uporabi klasifikacijske rezultate. Dobljene rezultate primerjamo s strategijami, ki ne uporabijo napovedi gibanja posameznih delnic.
Pričakovano obdobje držanja posameznih pozicij nam podaja osnovna logika trgovalnega sistema, ki smo jo podali v tem poglavju.
Velikost provizije je pomembna predvsem v primeru, ko imamo trgovalni sistem z veliko nakupi in prodajami in zelo kratkim držanjem pozicij (lahko tudi samo nekaj minut). Z razvojem sodobnih tehnologij in interneta so se možnosti za trgovanje z vrednostnimi papirji tudi za manjše trgovce občutno izboljšale. Kljub temu, da je osnovni princip trgovanja ostal nespremenjen, je z novimi tehnologijami trgovanje postalo cenejše, hitrejše, enostavnejše, bolj pregledno in bolj varno. Obstajata dve vrsti trgovalnih provizij: tako imenovana fiksna provizija ter stopenjska provizija, ki se spremnija glede na število trgovanj v koledarskem mesecu. Današnje provizije za trgovanje preko elektronskega trgovanja lahko znašajo tudi manj kot $0,2\%$ [3]. Pri fiksni proviziji strošek nakupa/prodaje delnice običajno znaša $0,005$ ameriških dolarjev (pol centa), maksimalna provizija je omejena na $0,5\%$ vrednosti posla in na posel zaračunajo vsaj 1 ameriški dolar [4]. Stroški tekočih borznih podatkov znašajo približno 100 ameriških dolarjev mesečno, trgovalna platforma pa je običajno ob zadostnem številu transkacij brezplačna, kar tudi manjšim trgovcem omogoča aktivno trgovanje. V naši predlagani trgovalni strategiji ne vključimo trgovalnih stroškov, saj trgovalna strategija služi zgolj kot ilustracija, da lahko apliciramo klasifikacijske rezultate v eno izmed predlaganih strategij.
Predlagano trgovalno strategijo, ki je prilagojena glede na naše klasifikacijske napovedi gibanja najvišjih tečajev v trgovalnem dnevu bomo testirali na 1920 trgovalnih dnevih in dnevno spreminjali vstopne komponente trgovalnega sistema, saj smo napovedovali gibanje delnic le en dan vnaprej. Predlagana trgovalna strategija je sestavljena iz dveh korakov:

- izgradnja klasifikacijskih modelov;
- uporaba zgrajenih modelov za dnevne napovedi, ki vsebujejo informacijo s katerimi delnicami naj bi trgovali naslednji dan. Trgovanje je vedno izvedeno v aktualnem, trenutnem času in na tekočih cenah (tečajih).

S trgovalnimi strategijami bi radi pokazali, da z uporabo napovedi klasifikatorjev dobimo statistično boljše rezultate, kot pa če teh ne upoštevamo. V tem poglavju definiramo tako imenovane **Vodene D–trgovalne strategije**, v katerih so odločitve 'vodene' s pomočjo uporabe klasifikacijskih modelov.

Kot merilo za primerjavo Vodenim trgovalnim strategijam vzamemo: Naivne strategije, ki so trgovalne strategije, v katerih ni vključenih klasifikacijskih napovedi, uporabimo pa nekaj pravil za odločanje o trgovanju (podobna pravila kot pri Vodenih D–trgovalnih strategijah), $S\&P$ 500 indeks, ki je velikokrat obravnavan kot primerjalna strategija. Vodene D–trgovalne strategije primerjamo tudi s 'Primerjalno strategijo' (ang. 'Benchmark'), na kateri uporabimo vseh 370 delnic, ki so enakomerno utežene in s katerimi vsak dan trgujemo. V tem poglavju natančneje definiramo vse zgoraj naštete trgovalne strategije. Kot kazalec uspešnosti trgovalnih strategij definirajmo **skupno letno stopnjo rasti** (ang. 'compound annual growth rate', CAGR), ki predstavlja donosnost naložbe skozi celotno obdobje naložbe. Skupni letni stopnji rasti rečemo tudi 'zglajena' stopnja donosa, saj meri rast naložbe tako, kot bi rasla enakomerno na skupni letni osnovi. Skupno letno stopnjo rasti izračunamo tako, da vzamemo n-ti koren od skupne stopnje rasti (ang. 'total percentage growth rate'), kjer n predstavlja število let med začetnim in končnim opazovanim obdobjem:

$$\text{CAGR} = \left(\frac{\text{zadnja vrednost}}{\text{začetna vrednost}}\right)^{\frac{1}{n}} - 1.$$

4.2 Vodena D–trgovalna strategija

Investitor upošteva napovedane smeri delniških donosov, ki jih dobimo s pomočjo klasifikacijskih modelov (glej poglavje 6). Strategija izbere tiste delnice, ki imajo napovedano rast najvišjih tečajev v naslednjem dnevu.

Strategija za trgovalni dan t in delnico j:

Če model napove, da bo najvišji tečaj delnice j v naslednjem dnevu t zrasel in če bo relativna razlika med najvišjim dnevnim tečajem v dnevu $t-1$ ($high_{t-1,j}$) in otvoritvenem dnevnem tečaju v dnevu t ($open_{t,j}$) opisana v enačbi 10, višja kot neka vnaprej podana meja D (D smo eksperimentalno določili na učnih množicah), potem kupimo delnico j po otvoritvenem dnevnem tečaju v trgovalnem dnevu t ($open_{t,j}$), sicer s to delnico v trenutnem dnevu ne trgujemo. Če trenutna vrednost tečaja delnice znotraj dneva t doseže najvišji tečaj v prejšnjem dnevu ($high_{t-1,j}$), potem zapremo našo pozicijo (delnico j takoj prodamo po tečaju $high_{t-1,j}$). V tem primeru dobimo za dan t in delnico j donos

$$donos_{t,j} = (high_{t-1,j} - open_{t,j})/open_{t,j}.$$

Če trenutna vrednost tečaja delnice ne preseže najvišjega tečaja v prejšnjem dnevu (kar pomeni, da model ni podal prave napovedi), zapremo našo pozicijo ob koncu trgovalnega dneva t po zaključnem dnevnem tečaju ($close_{t,j}$). Dobimo donos

$$donos_{t,j} = (close_{t,j} - open_{t,j})/open_{t,j},$$

ki je lahko tudi negativen.

$$\text{relativna razlika}_{t,j} = \left(\frac{high_{t-1,j}}{open_{t,j}} - 1\right) \cdot 100 \tag{10}$$

Upravljanje portfelja

Investitor bo sledil naslednji strategiji, ki je financirana iz lastnih sredstev (ang. 'self–financing strategy'):

- v trgovalnem dnevu t bo celotno premoženje investirano v delnice, ki imajo predvideno rast in ki imajo relativno razliko višjo kot nek vnaprej predpisan D (glej enačbo (10)).
- V vsako od izbranih delnic bomo investirali enak delež celotnega premoženja.
- Za izbrano delnico j, ki bo v prisotna v novem portfelju, sledimo zgoraj opisani **strategiji za trgovalni dan t in delnico j**.
- Ob koncu trgovalnega dneva t v portfelju nimamo nobene delnice.

Izračun donosov in skupna letna stopnja rasti

$$\text{donos}_t = \frac{\left(\sum_{j=1}^{N} donos_{t,j}\right)}{N}, \tag{11}$$

$$\text{skupen donos} = \left(\Pi_{t=1}^{d}(1 + donos_t)\right) - 1, \tag{12}$$

$$\text{CAGR} = (\text{skupen donos} + 1)^{\frac{1}{\#\text{let}}} - 1. \tag{13}$$

N je število vseh delnic, ki imajo predvideno rast (informacijo o predvideni rasti nam vrne klasifikacijski model) za naslednji dan t in za katere velja, da je relativna razlika višja kot postavljena meja D. d predstavlja število vseh trgovalnih dni. Parameter D smo vpeljali v ta namen, da bi minimizirali število transakcij, ki nam dajo premalo donosa in z njim lahko odpravimo možnost izgube. Naša predlagana strategija je osnovana na napovedi, ali bo najvišji tečaj delnice v dnevu $t + 1$ dosegel višjo vrednost kot pa najvišji tečaj delnice v dnevu t. Pozicijo v trgovalnem dnevu $t + 1$ odpremo takoj ko se borza odpre. Relativna razlika med $high_{t,j}$ in $open_{t+1,j}$ nam pove, kolikšen donos dobimo v primeru, če je klasifikacijska napoved pravilna.

4.3 Naivne strategije

Z Naivnimi strategijami želimo primerjati, ali uporaba klasifikatorjev povzroči signifikantne izboljšave v sami izvedbi trgovalnih strategij (primerjava z Vodenimi D–trgovalnimi strategijami) v primerjavi z enakimi strategijami brez vključitve klasifikatorjev (Naivne strategije). Na tak način želimo pokazati, da vključitev klasifikatorjev v trgovalnih strategijah prinese dodano vrednost. V Naivnih strategijah uporabimo različne meje D, ki so enake kot v Vodenih D–trgovalnih strategijah. Določitev teh je odvisna od izbranega klasifikatorja. V Naivni strategiji ne vključujemo napovedi, ki jih predhodno dobimo s klasifikacijskimi modeli.

Če bo relativna razlika med $high_{t-1}$ in $open_t$, opisana v enačbi (10), višja kot neka vnaprej podana meja D, potem kupimo delnico j po otvoritvenem dnevnem tečaju v trgovalnem dnevu t ($open_{t,j}$). Če trenutna vrednost tečaja delnice znotraj dneva t doseže najvišji tečaj v prejšnjem dnevu ($high_{t-1,j}$), potem zapremo našo pozicijo (delnico j takoj prodamo). V tem primeru dobimo za dan t in delnico j donos

$$donos_{t,j} = (high_{t-1,j} - open_{t,j})/open_{t,j}.$$

Če trenutna vrednost tečaja delnice ne preseže najvišjega tečaja v prejšnjem dnevu, zapremo našo pozicijo ob koncu trgovalnega dneva t po zaključnem dnevnem tečaju. Dobimo donos

$$donos_{t,j} = (close_{t,j} - open_{t,j})/open_{t,j}.$$

donos$_t$ je podobno definiran kot v enačbi (11), skupen donos kot v enačbi (12) in CAGR kot v enačbi (13). N je število vseh delnic, za katere velja, da je relativna razlika, opisana v enačbi (10), višja kot postavljena meja D.

4.4 Primerjalna '(benchmark)' strategija

Kot primerjalno strategijo predlagamo naslednji portfelj: vsak dan investiramo enako razmerje premoženja v vseh 370 delnic (po otvoritvenem dnevnem tečaju) in te ob koncu dneva prodamo po zaključnem dnevnem tečaju. Denar, ki ga pridobimo s prodajo delnic je naslednji dan reinvestiran. Strategija je znana tudi pod angleškim imenom 'equal–weighted portfolio'. Primerjalno strategijo bomo označili z 'bench'.

Izračun donosov:

Za vsak trgovalni dan imamo za trgovanje na razpolago 370 delnic. Naš dnevni donos znaša:

$$\text{donos}_t = \frac{\left(\sum_{t=1}^{370} \frac{close_t - open_t}{open_t}\right)}{370}.$$

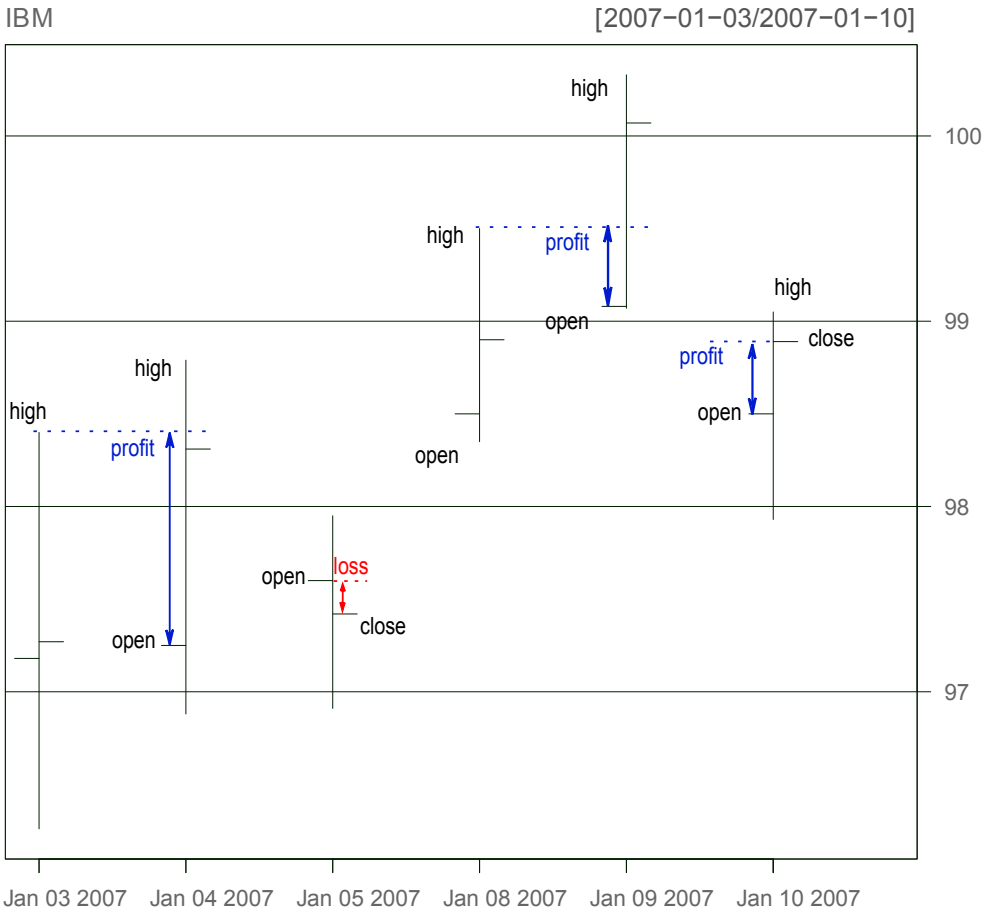
Po d trgovalnih dnevih, skupen donos znaša:

$$\text{skupen donos} = \left(\Pi_{t=1}^{d}(1 + donos_t)\right) - 1$$

4.5 Indeks S&P500

Kot primerjalno strategijo smo uporabili skupen donos indeksa $S\&P500$ (označili smo ga z oznako 'SPY'). Za ameriški trg je Standard & Poors 500 reprezentativen indeks.

4.5.1 Primer Vodene D–trgovalne strategije



Slika 8: Primer delovanja trgovalne strategije na IBM delnici. V tem primeru nismo vključili vrednosti meja D. Predpostavimo, da 4. januarja 2007 zjutraj uporabimo Vodeno D–trgovalno strategijo, ker ima otvoritveni dnevni tečaj nižjo vrednost kot pa najvišji tečaj v prejšnjem dnevu (3. januarja 2007) in ker klasifikacijski model vrne, da naj bi najvišji dnevni tečaj dosegal višje vrednosti kot pa najvišji tečaj v prejšnjem dnevu, IBM delnico kupimo po otvoritvenem dnevnem tečaju in jo obdržimo dokler trenutni tečaj ne preseže najvišjega tečaja v prejšnjem dnevu, nato jo prodamo. V primeru, da model ne napove, da bo najvišji tečaj presegel najvišji tečaj v prejšnjem dnevu, tisti dan ne trgujemo z IBM delnico. Na sliki zgoraj je na dne 4. januarja 2007 označen potencialen profit (pozitiven donos). Naslednji dan, 5. januarja 2007, je otvoritveni tečaj nižji kot najvišji tečaj v prejšnjem dnevu, vendar model lahko vrne napoved, da bo najvišji dnevni tečaj nižji, kot pa najvišji tečaj v prejšnjem dnevu–v tem dnevu delnice ne kupimo, saj se zanašamo na napoved modela. Tako se izognemo morebitni izgubi (glej sliko). Model lahko napove napačno napoved v tem dnevu (model predvidi rast najvšjega dnevnega tečaja). V tem primeru IBM delnico kupimo po otvoritvenem tečaju. Tečaj ni presegel najvišje vrednosti v prejšnjem dnevu, zato zapremo pozicijo ob zaprtju borze. Na sliki zgoraj smo označili potencialno izgubo. Naslednji trgovalni dan je tečaj ob odprtju višji kot pa najvišji dnevni tečaj 5. januarja 2007, zato tisti dan ne trgujemo z IBM delnico. Predpostavimo, da v naslednjih dveh dnevih model vrne, da bo najvišji dnevni tečaj višji kot pa dan prej, zato bi lahko dosegli pozitiven donos (profit) za oba trgovalna dneva. 10. januarja je cena IBM delnice ob zaprtju borze višja kot pa cena delnice ob otvoritvenem dnevnem tečaju, zato je kljub mogoči napačni napovedi zagotovljen profit.

5 KAZALCI USPEŠNOSTI TRGOVALNIH STRATEGIJ UPRAVLJANJA S PORTFELJEM

Na podlagi spodnjih kazalnikov uspešnosti upravljanja želimo primerjati različne trgovalne strategije.

5.1 Tveganje portfelja

Portfelj je kompozicija finančnih naložb (delnic, obveznic, depozitov in drugih naložb, enakovrednih denarju), s katerimi razpolaga posamezen vlagatelj ali upravljalec premoženja. Osnovni cilj vsakega vlagatelja (ali upravljalca portfelja) je poskrbeti, da bo donosnost portfelja čim višja ob zmerni ravni tveganja. Osnovno načelo obvladovanja tveganja je doseganje čim ugodnejša razmerja med tveganjem in donosnostjo [22, 101].

5.1.1 Donos in donosnost

Donos je absolutna količina, merimo jo v denarnih enotah in nam pove, koliko denarnih enot nam prinese lastništvo vrednostnega papirja v določenem obdobju nad številom enot, ki smo jih investirali. Donosnost pa je relativna količina, pomeni razmerje med donosom vrednostnega papirja in njegovo nabavno vrednostjo. Znana je tudi kot stopnja donosa, merimo pa jo v odstotkih [93]. Vsi kazalniki uspešnosti omenjeni spodaj, so bili povzeti s knjige [76].

5.1.2 Sharpeov koeficient

Sharpeov koeficient predstavlja mero uspešnosti in primerja donosnost premoženja z variabilnostjo njegove donosnosti v preučevanem obdobju. Sharpeov koeficient prikazuje uspešnost upravljanja premoženja. Pozitivna vrednost odraža učinkovito poslovanje, negativne vrednosti pa pomenijo slabo uspešnost upravljanja. Višja kot je vrednost Sharpeovega koeficienta, večja je stopnja preseženega donosa na enoto tveganja. Je eden najbolj uporabljenih kriterijev za merjenje dodatne donosnosti na enoto tveganja [93].

$$S_i = \frac{\overline{R_i} - \overline{RFR}}{\sigma_i},$$

kjer $\overline{R_i}$ predstavlja povprečje donosnosti naložbe, $\overline{RFR}$ predstavlja povprečje donosnosti netvegane naložbe, σ_i pa je standardni odklon donosnosti naložbe ([79]).
Za izračun Sharpeovega koeficienta za povprečen portfelj v individualnem uporavljanju premoženja uporabimo povprečno donosnost celotnega obdobja. V naši raziskavi smo vzeli $\overline{RFR} = 0$.

5.1.3 Kazalnik Sortino

Kazalnik Sortino pri izračunu upošteva le negativen odklon oziroma slabo tveganje, saj je odklon navzgor za vlagatelja pozitiven. Je variacija Sharpeovega koeficienta, ki loči škodljivo volatilnost od splo-

šne volatilnosti. Kazalnik omogoča investitorjem boljši pregled nad tveganjem, kot samo upoštevanje volatilnosti. Višja vrednost kazalnika Sortino je za vlagatelja ugodnejša. Definiran je kot:

$$ST_i = \frac{\overline{R_i} - \tau}{DR_i}.$$

V enačbi $\overline{R_i}$ predstavlja povprečno donosnost naložbe, τ predstavlja minimalno še sprejemljivo donosnost, DR_i pa negativno tveganje. Negativno tveganje v statistiki opredelimo z negativnim odklonom donosnosti. V naši raziskavi smo postavili $\tau = 0$.

5.1.4 Informacijski koeficient

Informacijski koeficient je relativna mera donosnosti. Je tveganju prilagojena mera uspešnosti upravljanja, ki vključuje preostalo (aktivno) donosnost glede na preostalo (aktivno) tveganje. Koeficient je zelo uporaben za merjenje uspešnosti med upravljavci, saj meri uspešnost upravljavca glede na informacijski stil oziroma naložbeno politiko. Informacijski koeficient za portfelj j izračunamo kot:

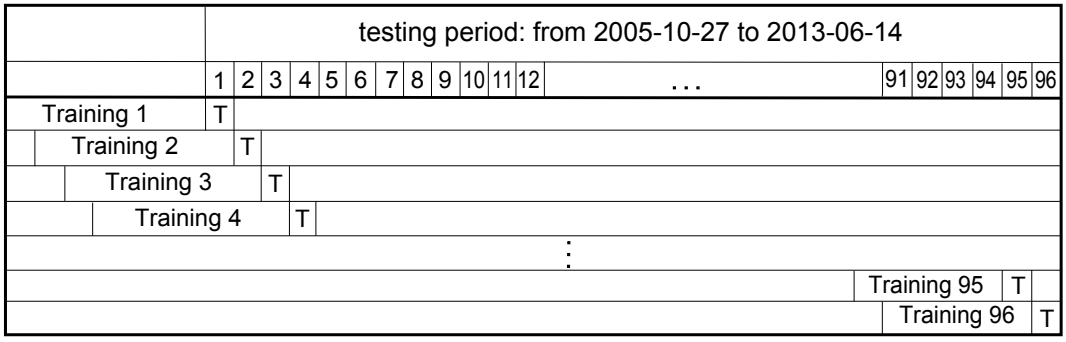
$$IR_j = \frac{\overline{R_j} - \overline{R_b}}{\sigma_{ER}},$$

kjer $\overline{R_j}$ predstavlja povprečno donosnost portfelja v preučevanem obdobju, $\overline{R_b}$ predstavlja povprečno donosnost prilagojenega kriterijskega indeksa in σ_{ER} predstavlja sledilno napako, ki je opredeljena kot standardna deviacija razlike med donosnostjo portfelja R_j in donosnostjo prilagojenega kriterijskega indeksa R_b. V našem raziskovalnem delu smo za kriterijski indeks vzeli indeks $S\&P500$.

6 REZULTATI NAPOVEDOVANJA RASTI IN PADCEV NAJVIŠJIH TEČAJEV

6.1 Eksperimentalno delo

Podatki, ki smo jih uporabili v eksperimentalnem delu, so zajeti s spletne strani 'Yahoo! Finance'[1]. Opazovali bomo delnice podjetij, ki sestavljajo indeks $S\&P500$. Za analizo smo uporabili 6 časovnih vrst delnic: tečaj delnic ob odprtju, zaprtju, najvišji, najnižji tečaj delnic, prilagojen tečaj ob zaprtju ('adjusted close') ter promet delnic. Te časovne vrste smo prilagodili glede na cepitev delnic in izplačane dividende ('adjusted close'). Podatke smo zajeli v časovnem razponu od 31. oktobra 2003 pa do 14. junija 2013, kar je skupno 2420 trgovalnih dni. V eksperimentalno delo smo vključili 370 delnic, ki so bile 1. avgusta 2013 članice indeksa $S\&P500$ in so članice indeksa v zgoraj omenjenem časovnem razponu. Podatke normaliziramo: za vsak atribut/časovno vrsto x s povprečjem 0 in standardno deviacijo z vrednostjo 1: $(x-\mu)/\sigma$, kjer μ predstavlja povprečje in σ standardno deviacijo za atribut x. Za analizo smo uporabili metodo **drsečih oken**, s pomočjo katere razdelimo podatke na učne in testne množice. Prvo učno obdobje, ki obsega 500 trgovalnih dni, je uporabljeno za izgradnjo napovednega modela in je označeno na sliki 9 z 'Training 1'. Izbrana dolžina učne množice, 500 trgovalnih dni, je nekakšen kompromis med majhno, stacionarno časovno vrsto, s katero ne dosežemo veliko uspeha pri učenju, ter med daljšo, manj stacionarno časovno vrsto [60]. Naslednje časovno obdobje, od 0 do h dni po 1. učni množici (na sliki 9 označeno z oznako 'T'), predstavlja testno obdobje in ga uporabimo za primerjavo z dobljenimi razvrščenimi rezultati ter z dejanskimi vrednostmi. Nato nadaljujemo s postopkom. Dolžino testnih množic, $h = 20$ smo določili eksperimentalno (glej [71] in prilogo 8.8).



Slika 9: Metoda drsečih oken z 1-mesečnimi testnimi podatki.

Nova učna množica se zamakne od začetne za h trgovalnih dni in proces učenja in testiranja se nadaljuje. Predlagano shemo smo učili na množicah, dolgih 500 trgovalnih dni in testirali na množicah, dolgih 20 trgovalnih dni. Kot rezultat dobimo 96 modelov za vsako delnico in s tem 96 množic napovedi, kako

[1] http://finance.yahoo.com/

pravi razred ↓ / klasificiran kot→	rast	padec
rast	a	b
padec	c	d

Klasifikacijska točnost = $\frac{a+d}{a+b+c+d}$
Senzitivnost = $\frac{a}{a+c}$
Specifičnost = $\frac{d}{b+d}$
Prec -1 = $\frac{d}{c+d}$
Prec 1 = $\frac{a}{a+b}$

Tabela 4: Klasifikacijska točnost, tabela klasifikacij.

se gibajo najvišji tečaji v danem trgovalnem dnevu. Če povzamemo na kratko, zgoraj opisani proces pomeni, da model osvežimo vsakih 20 dni. Uporabili smo 98 tehničnih indikatorjev (glej tabelo 3). Donosi, ki so označeni z 1 pomenijo, da je v naslednjem trgovalnem dnevu predvidena rast najvišjih tečajev in donosi, ki so označeni z -1 pomenijo delnice, katerih najvišji dnevni tečaji bodo v naslednjem trgovalnem dnevu padli. Raziskali bomo dnevne spremembe najvišjih dnevnih tečajev in, ali se da z izbranimi tehničnimi indikatorji napovedati rast ali padec v naslednjem trgovalnem dnevu.
Za dvorazredne probleme sta se v praksi uveljavili dve meri, senzitivnost (ang. 'sensitivity') in specifičnost (ang. 'specificity'), ki sta izpeljani iz štirih osnovnih količin, razvidnih iz tabele 4. Senzitivnost ocenjuje odstotek pravilno klasificiranih pozitivnih primerov (rasti), specifičnost pa ocenjuje odstotek pravilno klasificiranih negativnih primerov (padcev). Mnogokrat se raziskovalci trudijo maksimizirati senzitivnost testa, vendar je pri tem fiksirana spodnja meja še sprejemljive specifičnosti. Seveda je trivialno doseči 100% senzitivnost, tako da vse primere klasificiramo v pozitivni razred (razred rasti) in dobimo specifičnost 0%. Enako je trivialno doseči 100% specifičnost, tako da vse primere klasificiramo v negativni razred (in dobimo senzitivnost 0%). Preciznost za razred rasti ocenjuje odstotek pravilno klasificiranih primerov, ki so bili klasificirani kot pozitivni (v razredu rast) [55] in preciznost za razred padcev, simetrično. Za izmero pravilnih napovedi smo uporabili klasifikacijsko točnost na učni množici in testni množici, senzitivnost, specifičnost, preciznost za razred '1' in preciznost za razred '-1' s spremljajočimi standardnimi deviacijami. Za analizo in eksperimentalno delo smo uporabljali program R z nekaterimi paketi [75].

6.2 Rezultati-izbor ustreznega jedra in SVM parametrov

Uspešnost metode podpornih vektorjev pripisujemo predvsem uporabi jedrnih funkcij, ki primere preslika v višjo dimenzijo. Izbor jedrnih funkcij je pomemben, saj glede na izbor le-teh dobimo različne klasifikacijske rezultate. V splošnem sta linearno in RBF jedro najbolj primerni, zato smo v tej raziskavi kot jedrno funkcijo za SVM uporabili radialno bazno funkcijo (RBF) in linearno funkcijo. V primerjavi s polinomskimi jedri, RBF jedro dosega boljše rezultate pri klasifikacijski natančnosti in pri

času izvedbe ([11, 52, 85]). Pri linearni jedrni funkciji je C edini parameter, ki bi ga morali določiti in je tudi najenostavnejša funkcija za uporabo. S parametrom C obravnavamo šume pri učenju in z njim določamo, kako zelo naj dopušča odstopanja od osnovne skupine enega razreda. Za dobro klasifikacijo je potrebno izbrati pravilno vrednost konstante, s katero bo algoritem znal dovolj dobro obravnavati šum. Višji kot je C, bolj je metoda tolerantna do odstopanja. Od neke vrednosti dalje višji parametri C ne prinašajo boljše klasifikacije, le po nepotrebnem upočasnijo proces učenja. RBF jedro ima v primerjavi z linearnim jedrom sposobnost ustvarjanja ukrivljenih hiperravnin. Pri določitvi RBF jedra igrata pomembno vlogo parametra C in γ.
Premajhne vrednosti parametra C lahko povzročijo premajhno prileganje podatkom, pri izboru prevelike vrednosti parametra C, pa lahko pomeni prekomerno prileganje podatkom ([50, 52, 85]). Odločili smo se, kot tudi avtorja Tay in Cao [85], da bomo parameter C omejili med vrednostma 1 in 100. Za dosego čim višjih klasifikacijskih rezultatov na testnih množicah, je potrebno določiti primerne parametre C in γ. Parameter γ smo določili hevristično, kot je opisano v [49] in [15], pri parametru C pa smo izbirali med $C = 1, 10, 33, 55, 78, 100$. Zaradi velike časovne zahtevnosti smo parametre za RBF in linearno jedro določili le na prvi učni množici (ki sega od 31. oktobra 2003 pa do 26. oktobra 2005). Parametre smo določili za vsako delnico posebej in jih obdržimo skozi celotni eksperiment. Na nekaj delnicah smo tudi eksperimentalno preverili, kako se spreminjajo parametri skozi čas, vendar se na modelih, zgrajenih na dobljenih parametrih, klasifikacijski rezultati niso signifikantno razlikovali.
Napovedi smo naredili na 96 testnih množicah (za vsako podjetje/delnico) v obdobju od 27. oktobra 2005 do 14. junija 2013. Poskusili smo določiti optimalna parametra C in γ tako, da bomo dosegli najvišjo klasifikacijsko natančnost na testni množici.

6.3 Klasifikacijski rezultati

V tabelah 5, 6, 7 in 8 ter tabelah 9, 10, 11 in 12, učna točnost pomeni povprečje vseh zadetkov (s spremljajočo standardno deviacijo) na vseh 96 učnih množicah, medtem ko testna točnost pomeni povprečje vseh zadetkov na testni množici (s spremljajočo standardno deviacijo) na vseh 96 testnih množicah; senzit, specif, preciz 1 in preciz -1 pa pomenijo povprečje za senzitivnost, specifičnost, preciznost za razred 1 in preciznost za razred -1, na vseh 96 testnih množicah. Zelo pomembna mera je preciznost za razred 1 (razred rasti), ki predstavlja odstotek pravilno klasificiranih delnic, ki so bile klasificirane kot pozitivne (v razredu 1).
Pri SVM klasifikatorjih smo uporabili vnaprej izbrane hiperparametre C in γ (za RBF jedro) in vnaprej izbrane parametre C (za linearno jedro) za vse delnice na prvi učni množici. Najvišjo klasifikacijsko natančnost na testni množici ima LDA klasifikator. Izbrani klasifikatorji so po rezultatih sodeč zmožni uvrščanja v razrede v povprečju več kot 61% klasifikacijske natančnosti na testni množici. Omenjene klasifikatorje bomo uporabili na predlaganih trgovalnih strategijah, ki so opisane poglavju 4.
Primerjali smo tudi klasifikacijske točnosti po delnicah, zanimalo nas je namreč, ali so delnice bolj/manj uspešne v primerjavi z drugimi. Raziskava je pokazala, da se v povprečju vse delnice obnašajo približno enako, ni bilo videti nobenih velikih odstopanj ('outliers'). Razpon klasifikacijskih točnosti se giba med

55.21%–67.29% (glej prilogo 8.1, slike: 40, 41, 42 in 43).

6.4 Rezultati filtrirnih metod in analiza relevantnih atributov

V eksperimentalnem delu smo izbirali relevantne atribute z nekaj multivariatnimi filtrirnimi metodami, med katerim je tudi algoritem za izbor atributov, FSuC. V predlagan algoritem smo vključili metodo za razvrščanje v skupine, vprašanje, ki se pojavi pa je, katera izmed številnih obstoječih metod za razvrščanje, je za naše podatkovje najbolj primerna. Izbirali smo med različnimi reprezentativnimi metodami za razvrščanje, med metodo voditeljev, Wardovo metodo, minimalno ter maksimalno metodo. Kot smo že omenili, se minimalna metoda zelo obnese pri razkrivanju dolgih 'klobasastih', tudi neeliptičnih struktur in je neuporabna pri neizrazito ločenih skupinah, kar pa v naših podatkih ne zasledimo. Kadar uporabimo minimalno metodo, so klasifikacijski rezultati med FSuC različicami najslabši, kar pa ni presenetljivo. Glede na klasifikacijske rezultate, ki so prikazani v tabelah 5, 6, 7 ter 8, ugotovimo, da Wardova metoda vrne najvišje klasifikacijske rezultate na testnih množicah, saj je tudi najprimernejša za eliptično strukturirane podatke. Pri vseh metodah za razvrščanje smo uporabili evklidsko razdaljo. Pri metodi voditeljev so bili začetni voditelji naključno izbrani. V predlaganem algoritmu FSuC smo predstavili 3 obstoječe mere za klasifikacijo (glej poglavje 3.6.1). Katera mera vrne najbolj relevantne atribute in s tem najvišje klasifikacijske rezultate, je naslednje vprašanje, ki se pojavi, zato smo zgradili kombinirane klasifikatorje (v tabelah 5, 6, 7 in 8 ter tabelah 9, 10, 11 in 12, so vrstice s kombiniranimi klasifikatorji označene s 'FSuC–kmeans–comb', 'FSuC–ward–comb', 'FSuC–min–comb' ter 'FSuC–max–comb'). Kombinirane klasifikatorje smo definirali tako, da na učni množici uporabimo vse tri možne klasifikacijske mere FSuC algoritma. Na ostalih, testnih podatkih pa uporabimo tisto mero, ki vrne najvišji rezultat na učni množici, datumsko najbližji testni množici. Rezultati, ki jih dobimo s kombiniranimi klasifikatorji so precej podobni tistim, ki jih posamezna različica algoritma vrne na učni množici z najvišjim rezultatom (statistično značilno se rezultati na testni množici razlikujejo le pri NB klasifikatorju).

Sedaj, ko smo eksperimentalno potrdili, katero metodo bomo uporabili pri razvrščanju v skupine (Wardova metoda), želimo primerjati dobljene klasifikacijske rezultate z reprezentativnimi multivariatnimi filtrirnimi metodami, ki so CFS, FCBF, mRMR in CCCA. V tabelah 9, 10, 11 in 12 so predstavljeni dobljeni klasifikacijski rezultati, ki pričajo, da v vseh primerih kombinirane metode z FSuC algoritmi (FSuC–ward–comb) vrnejo najbolj relevantne atribute.

V prilogi 8.1 smo prikazali povprečne klasifikacijske točnosti po delnicah.

metoda	**uč toč**	std	**test toč**	std	**senzit**	std	**specif**	std	**preciz 1**	std	**preciz -1**	std
FSuC1–kmeans	60.93	3.44	59.45	11.09	59.43	24.29	57.96	23.46	59.49	16.80	59.95	17.06
FSuC2–kmeans	61.20	3.60	59.59	11.19	59.77	24.53	58.03	23.51	59.66	16.81	60.33	17.23
FSuC3–kmeans	60.67	3.46	59.08	11.07	59.08	24.77	57.43	23.94	59.03	16.88	59.55	17.30
FSuC–kmeans–comb	61.20	3.60	59.59	11.19	59.77	24.53	58.03	23.51	59.66	16.81	60.33	17.23
FSuC1–ward	63.04	2.82	61.92	10.54	61.74	19.00	59.49	18.11	61.04	15.25	61.35	15.61
FSuC2–ward	63.08	2.84	**61.92**	10.55	61.82	19.07	59.39	18.12	**60.96**	15.23	61.38	15.62
FSuC3–ward	62.85	2.90	61.71	10.55	61.57	19.19	59.10	18.37	60.75	15.26	61.11	15.69
FSuC–ward–comb	63.08	2.84	**61.92**	10.55	61.82	19.07	59.39	18.12	**60.96**	15.23	61.38	15.62
FSuC1–min	58.15	3.58	56.73	11.44	56.95	28.39	53.15	28.62	56.00	16.17	55.95	16.01
FSuC2–min	58.19	3.57	56.80	11.37	56.91	28.04	53.27	28.29	56.08	16.00	55.90	16.02
FSuC3–min	58.01	3.68	56.65	11.57	56.32	30.04	53.72	30.16	56.15	15.77	56.03	15.56
FSuC–min–comb	58.19	3.57	56.80	11.37	56.91	28.04	53.27	28.29	56.08	16.00	55.90	16.02
FSuC1–max	62.03	3.07	60.98	10.63	60.61	20.02	58.23	19.49	59.99	15.58	60.10	15.87
FSuC2–max	62.08	3.10	61.01	10.61	60.65	20.08	58.20	19.34	59.95	15.61	60.20	15.89
FSuC3–max	61.65	3.20	60.58	10.64	60.29	20.52	57.51	19.92	59.43	15.71	59.70	16.01
FSuC–max–comb	62.08	3.10	61.01	10.61	60.65	20.08	58.20	19.34	59.95	15.61	60.20	15.89

Tabela 5: Klasifikacijski rezultati, ki smo jih dobili z LDA klasifikatorjem, za vhodne podatke vzamemo tehnične indikatorje dobljene z različnimi FSuC metodami za izbor spremenljivk.

metoda	**uč toč**	std	**test toč**	std	**senzit**	std	**specif**	std	**preciz 1**	std	**preciz -1**	std
FSuC1–kmeans	60.41	3.34	58.70	11.07	63.19	25.02	52.38	24.69	58.00	16.21	60.29	18.73
FSuC2–kmeans	60.59	3.58	58.73	11.23	62.68	25.65	53.12	24.92	58.12	16.46	60.50	18.67
FSuC3–kmeans	60.14	3.33	58.43	11.06	62.70	25.37	52.12	25.01	57.68	16.23	59.84	18.76
FSuC–kmeans–comb	60.59	3.58	58.73	11.23	62.68	25.65	53.12	24.92	58.12	16.46	60.50	18.67
FSuC1–ward	62.04	2.81	**61.09**	10.38	68.76	19.23	49.75	19.53	**58.56**	14.06	62.76	18.42
FSuC2–ward	62.04	2.81	61.02	10.41	68.58	19.25	49.79	19.51	58.54	14.10	62.62	18.39
FSuC3–ward	61.78	2.82	60.81	10.38	68.16	19.41	49.61	19.78	58.34	14.15	62.16	18.29
FSuC–ward–comb	62.04	2.81	61.02	10.41	68.58	19.25	49.79	19.51	58.54	14.10	62.62	18.39
FSuC1–min	58.08	3.47	56.44	11.40	57.25	29.05	52.06	29.33	55.59	16.21	55.64	16.43
FSuC2–min	58.14	3.46	56.52	11.36	57.12	28.88	52.33	29.07	55.65	16.13	55.68	16.45
FSuC3–min	57.98	3.59	56.52	11.59	56.56	30.32	53.34	30.35	56.03	15.64	56.07	15.79
FSuC–min–comb	58.14	3.46	56.52	11.36	57.12	28.88	52.33	29.07	55.65	16.13	55.68	16.45
FSuC1–max	61.21	2.85	60.12	10.41	65.47	20.61	50.62	20.61	57.86	14.70	60.51	18.00
FSuC2–max	60.78	2.90	59.73	10.38	64.59	21.06	50.34	20.95	57.37	14.82	59.76	17.86
FSuC3–max	60.78	2.90	59.73	10.38	64.59	21.06	50.34	20.95	57.37	14.82	59.76	17.86
FSuC–max–comb	61.22	2.86	60.13	10.40	65.38	20.59	50.65	20.49	57.82	14.69	60.47	17.93

Tabela 6: Klasifikacijski rezultati, ki smo jih dobili z NB klasifikatorjem, za vhodne podatke vzamemo tehnične indikatorje dobljene z različnimi FSuC metodami za izbor spremenljivk.

metoda	**uč toč**	std	**test toč**	std	**senzit**	std	**specif**	std	**preciz 1**	std	**preciz -1**	std
FSuC1–kmeans	60.85	3.60	59.18	11.22	60.43	26.48	56.31	25.58	59.28	16.68	60.47	17.69
FSuC2–kmeans	61.12	3.75	59.30	11.31	60.74	26.76	56.33	25.59	59.39	16.61	60.85	17.90
FSuC3–kmeans	60.57	3.63	58.82	11.25	59.83	27.07	56.11	26.07	58.94	16.75	60.04	17.76
FSuC–kmeans–comb	61.12	3.75	59.30	11.31	60.74	26.76	56.33	25.59	59.39	16.61	60.85	17.90
FSuC1–ward	63.00	2.92	**61.88**	2.92	63.83	19.66	57.12	18.98	**60.63**	14.88	62.00	16.35
FSuC2–ward	63.03	2.93	61.86	2.93	63.89	19.73	57.00	18.95	60.57	14.88	62.02	16.39
FSuC3–ward	62.78	2.98	61.66	2.98	63.51	19.89	56.84	19.26	60.35	14.95	61.67	16.40
FSuC–ward–comb	63.03	2.93	61.86	2.93	63.89	19.73	57.00	18.95	60.57	14.88	62.02	16.39
FSuC1–min	57.94	3.63	56.83	11.43	59.20	28.83	50.92	29.50	56.04	14.98	56.26	15.28
FSuC2–min	58.01	3.62	56.95	11.36	58.85	28.62	51.41	29.18	56.12	14.86	56.23	15.34
FSuC3–min	57.71	3.67	56.65	11.57	58.20	30.49	51.77	31.00	56.14	14.64	56.28	14.89
FSuC–min–comb	58.01	3.62	56.95	11.36	58.85	28.62	51.41	29.18	56.12	14.86	56.23	15.34
FSuC1–max	61.94	3.18	60.84	10.65	62.46	20.82	55.96	20.44	59.57	15.17	60.58	16.50
FSuC2–max	61.98	3.21	60.88	10.63	62.53	20.86	55.91	20.27	59.53	15.15	60.67	16.57
FSuC3–max	61.52	3.31	60.47	10.68	62.07	21.29	55.33	20.86	59.10	15.10	60.12	16.52
FSuC–max–comb	61.98	3.21	60.88	10.63	62.53	20.86	55.91	20.27	59.53	15.15	60.67	16.57

Tabela 7: Klasifikacijski rezultati, ki smo jih dobili z SVM klasifikatorjem (z linearnim jedrom), za vhodne podatke vzamemo tehnične indikatorje dobljene z različnimi FSuC metodami za izbor spremenljivk.

metoda	**uč toč**	std	**test toč**	std	**senzit**	std	**specif**	std	**preciz 1**	std	**preciz -1**	std
FSuC1–kmeans	64.33	5.57	58.75	11.13	62.87	24.78	52.89	24.24	58.10	16.11	60.32	18.53
FSuC2–kmeans	65.11	6.47	58.75	11.21	62.76	25.10	53.10	24.40	58.18	16.23	60.45	18.61
FSuC3–kmeans	64.04	5.68	58.38	11.19	62.37	25.40	52.60	24.74	57.69	16.28	59.93	18.72
FSuC–kmeans–comb	65.11	6.47	58.75	11.21	62.76	25.10	53.10	24.40	58.18	16.23	60.45	18.61
FSuC1–ward	64.85	3.20	**61.50**	10.54	65.65	19.74	54.33	19.08	**59.72**	14.69	62.28	17.19
FSuC2–ward	64.99	3.37	61.46	10.57	65.53	19.86	54.40	19.09	59.72	14.71	62.31	17.20
FSuC3–ward	64.71	3.35	61.28	10.53	65.23	20.00	54.19	19.33	59.46	14.65	61.96	17.15
FSuC–ward–comb	64.99	3.37	61.46	10.57	65.53	19.86	54.40	19.09	59.72	14.71	62.31	17.20
FSuC1–min	59.61	3.36	56.55	11.53	59.02	29.24	51.01	29.51	55.83	16.24	56.46	17.12
FSuC2–min	59.68	3.33	56.63	11.49	58.91	28.97	51.22	29.23	55.93	16.14	56.44	17.17
FSuC3–min	59.25	3.50	56.67	11.62	58.01	30.56	52.44	30.56	56.28	15.68	56.75	16.28
FSuC–min–comb	59.68	3.33	56.63	11.49	58.91	28.97	51.22	29.23	55.93	16.14	56.44	17.17
FSuC1–max	63.54	3.22	60.48	10.68	64.07	21.11	53.57	20.69	58.85	15.11	60.86	17.41
FSuC2–max	63.61	3.30	60.49	10.68	64.10	21.31	53.50	20.68	58.76	15.20	60.96	17.54
FSuC3–max	63.14	3.39	60.09	10.67	63.51	21.68	53.11	21.16	58.37	15.23	60.36	17.43
FSuC–max–comb	63.61	3.30	60.49	10.68	64.10	21.31	53.50	20.68	58.76	15.20	60.96	17.54

Tabela 8: Klasifikacijski rezultati, ki smo jih dobili z SVM klasifikatorjem (z RBF jedrom), za vhodne podatke vzamemo tehnične indikatorje dobljene z različnimi FSuC metodami za izbor spremenljivk.

metoda	**uč toč**	std	**test toč**	std	**senzit**	std	**specif**	std	**preciz 1**	std	**preciz -1**	std
FSuC–ward–comb	63.08	2.84	**61.92**	10.55	61.82	19.07	59.39	18.12	**60.96**	15.23	61.38	15.62
CFS	62.08	2.26	60.10	10.42	57.51	19.79	59.73	18.92	59.20	15.84	59.17	15.01
FCBF	61.46	3.20	58.54	11.16	60.58	24.05	55.33	22.96	58.45	16.68	60.08	18.18
mRMR	56.40	3.74	53.92	11.57	56.73	34.29	48.52	34.04	53.76	18.02	54.54	20.19
CCCA	57.41	3.50	54.94	11.47	51.95	35.20	54.90	34.61	54.86	17.40	54.41	16.43

Tabela 9: Klasifikacijski rezultati, ki smo jih dobili z LDA klasifikatorjem, za vhodne podatke vzamemo tehnične indikatorje dobljene s kombiniranim klasifikacijskim algoritmom ter različnimi metodami za izbor spremenljivk.

metoda	**uč toč**	std	**test toč**	std	**senzit**	std	**specif**	std	**preciz 1**	std	**preciz -1**	std
FSuC–ward–comb	62.04	2.81	**61.02**	10.41	68.58	19.25	49.79	19.51	**58.54**	14.10	62.62	18.39
CFS	56.61	4.65	55.31	11.88	56.24	36.92	51.42	37.08	55.59	13.57	55.98	14.57
FCBF	60.70	3.09	57.45	11.25	62.95	25.98	50.11	25.20	56.78	16.50	60.06	19.93
mRMR	56.31	3.46	53.38	11.51	59.08	33.63	45.38	33.31	52.87	17.74	54.53	21.61
CCCA	56.83	3.10	54.21	11.28	53.28	36.39	51.68	36.06	53.68	16.70	53.74	17.21

Tabela 10: Klasifikacijski rezultati, ki smo jih dobili z NB klasifikatorjem, za vhodne podatke vzamemo tehnične indikatorje dobljene s kombiniranim klasifikacijskim algoritmom ter različnimi metodami za izbor spremenljivk.

metoda	**uč toč**	std	**test toč**	std	**senzit**	std	**specif**	std	**preciz 1**	std	**preciz -1**	std
FSuC–ward–comb	63.03	10.03	**61.86**	2.93	63.89	19.73	57.00	18.95	**60.57**	14.88	62.02	16.39
CFS	62.12	2.29	61.28	10.06	60.82	16.44	57.71	16.48	59.60	14.50	59.59	14.46
FCBF	61.29	3.27	58.95	10.98	63.40	22.96	52.54	22.50	58.09	15.51	60.55	18.36
mRMR	56.05	3.80	53.57	11.72	57.72	38.00	47.03	37.64	53.70	16.09	55.53	19.53
CCCA	57.14	3.57	55.16	11.46	53.71	35.84	53.14	35.59	55.06	15.95	54.74	15.97

Tabela 11: Klasifikacijski rezultati, ki smo jih dobili z SVM klasifikatorjem (z linearnim jedrom), za vhodne podatke vzamemo tehnične indikatorje dobljene s kombiniranim klasifikacijskim algoritmom ter različnimi metodami za izbor spremenljivk.

metoda	**uč toč**	std	**test toč**	std	**senzit**	std	**specif**	std	**preciz 1**	std	**preciz -1**	std
FSuC–ward–comb	64.99	3.37	**61.46**	10.57	65.53	19.86	54.40	19.09	**59.72**	14.71	62.31	17.20
CFS	65.20	5.62	60.50	10.36	60.84	19.36	56.39	18.90	59.04	15.08	59.40	15.51
FCBF	67.64	4.30	59.03	11.00	64.15	22.76	52.47	21.94	58.31	15.69	61.15	18.69
mRMR	60.90	6.72	53.48	11.51	56.53	33.07	48.44	32.78	53.33	18.03	54.18	20.32
CCCA	59.13	4.07	54.04	11.44	52.95	36.58	52.26	36.27	53.93	17.08	53.87	17.21

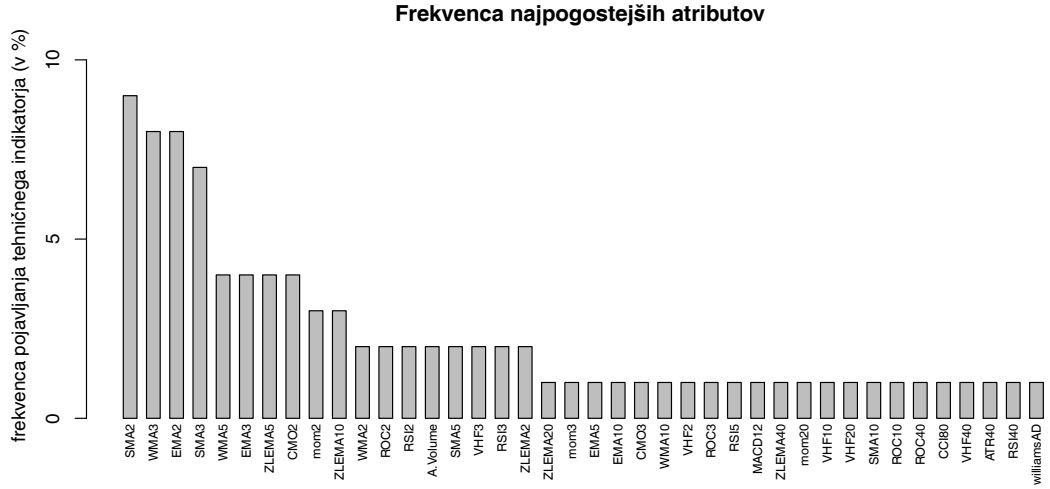
Tabela 12: Klasifikacijski rezultati, ki smo jih dobili z SVM klasifikatorjem (z RBF jedrom), za vhodne podatke vzamemo tehnične indikatorje dobljene s kombiniranim klasifikacijskim algoritmom ter različnimi metodami za izbor spremenljivk.

6.5 Uporabljeni atributi

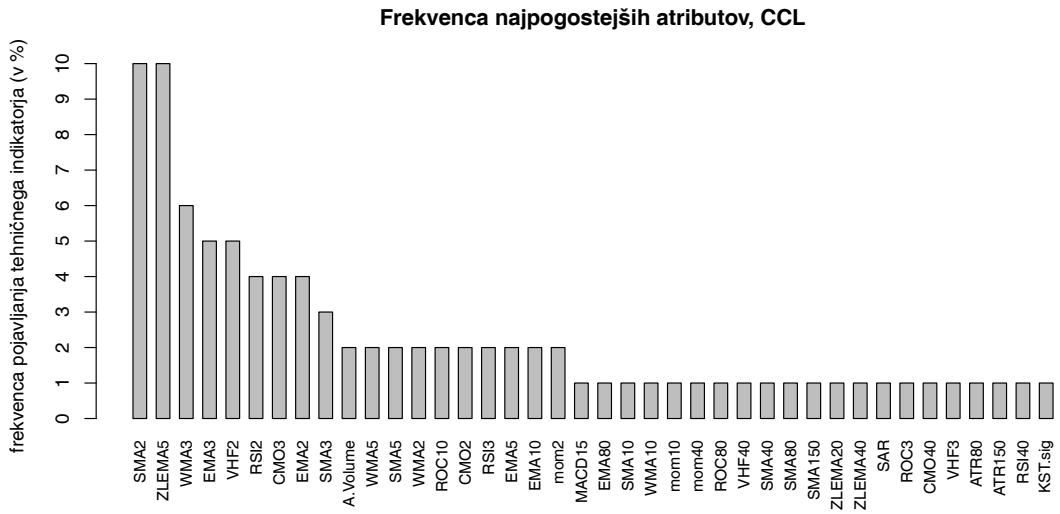
Iz histogramov 10, 14, 18, 22 in 26, lahko opazimo, da so relevantni atributi različno zastopani po različnih metodah. Vsak stolpec v histogramu predstavlja frekvenco pojavljanja posameznega tehničnega indikatorja za vse delnice in na vseh 96 časovnih oknih (modelov) v %. Prevladujejo atributi, ki imajo v izračun zajetih majhno število preteklih dni. Po metodah smo predstavili frekvenco pojavljanja tehničnih indikatorjev na vseh delnicah in na treh izbranih delnicah: CCL, AMZN ter AAPL. Omenjene delnice smo izbrali glede na dobljene klasifikacijske rezultate (glej poglavje 8.1): delnica CCL ima slabše klasifikacijske rezultate (in je v prvi četrtini po rezultatih), AMZN je nekje v sredini, ter AAPL, kot ena izmed najuspešnejših delnic po dobljenih klasifikacijskih rezultatih.
V prilogi 8.2 smo podali še natančnejše informacije o relevantnih tehničnih indikatorjih in sicer predstavitev relevantnih atributov skozi celotno časovno obdobje (najbolj frekventen relevanten atribut, ki se je pojavil v tem času), glej slike: 44, 47, 50, 53 in 56, ter predstavitev relevantnih atributov po delnicah (najbolj frekventen relevanten atribut na delnico), glej slike 45, 46, 48, 49, 51, 52, 54, 55, 57 in 58.

6.5.1 FSuC–ward-comb

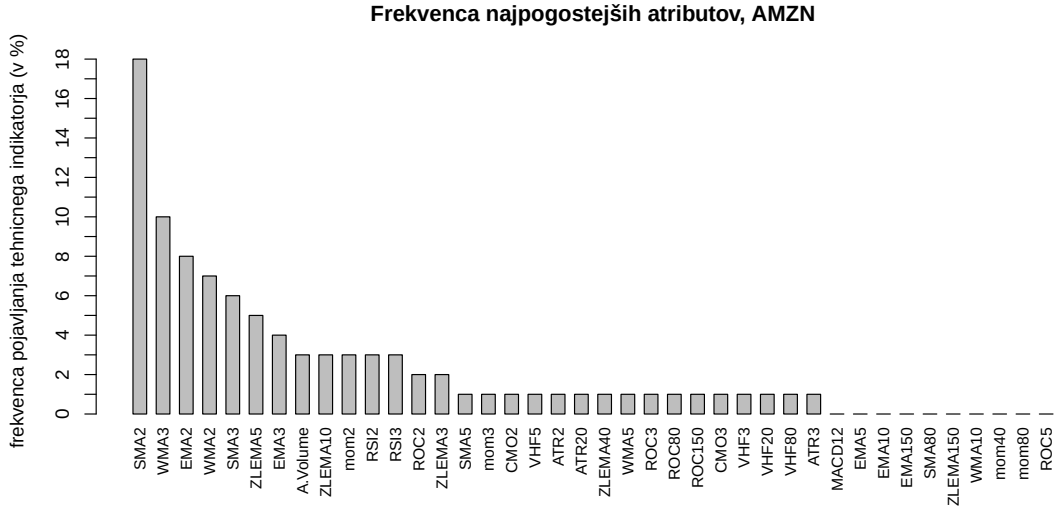
Na slikah 10, 11, 12 in 13, so prikazane frekvence pojavljanja (v %) relevantnih tehničnih indikatorjev, ki jih vrne predlagana metoda FSuC–ward-comb. Zanimivo je primerjati, kako se razlikujejo relevantni atributi glede na posamezne delnice. Pri vseh delnicah je frekvenca pojavljanja atributa SMA2 najpogostejša, z grafov je razvidno tudi, da je najvišja frekvenca pojavljanja pri drsečih sredinah s krajšim obdobjem časovnih enot.



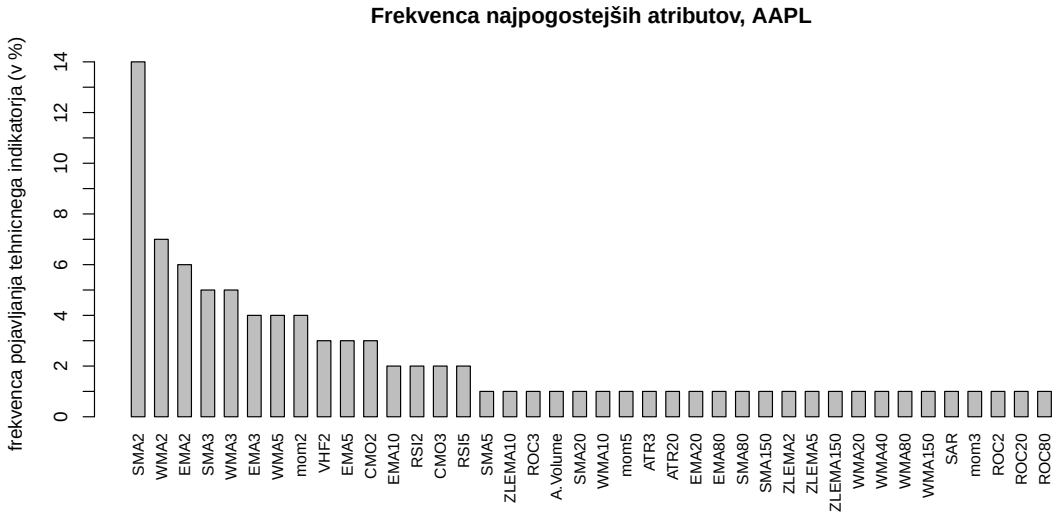
Slika 10: 40 najbolj frekventnih tehničnih indikatorjev na vseh 370 delnicah, dobljenih z metodo FSuC–ward–comb.



Slika 11: 40 najbolj frekventnih tehničnih indikatorjev za delnico CCL, dobljenih z metodo FSuC–ward–comb.



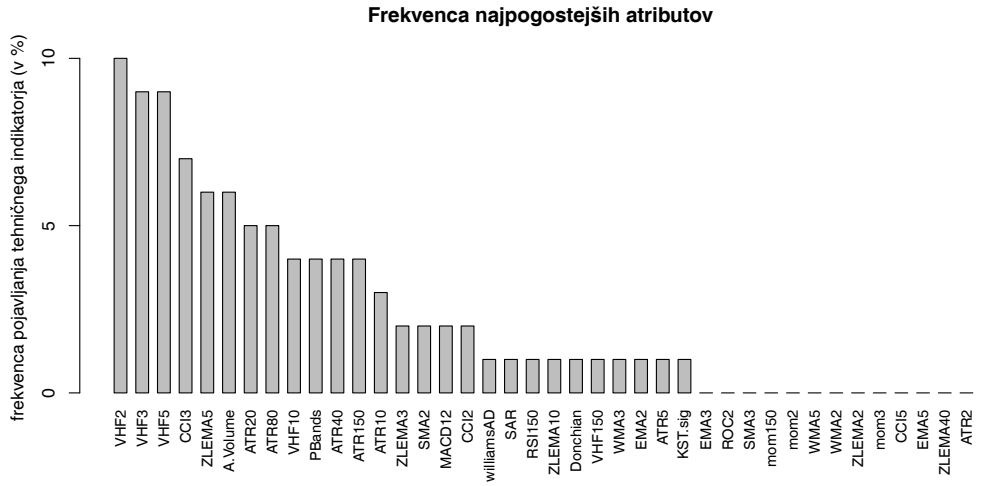
Slika 12: 40 najbolj frekventnih tehničnih indikatorjev za delnico AMZN, dobljenih z metodo FSuC–ward–comb.



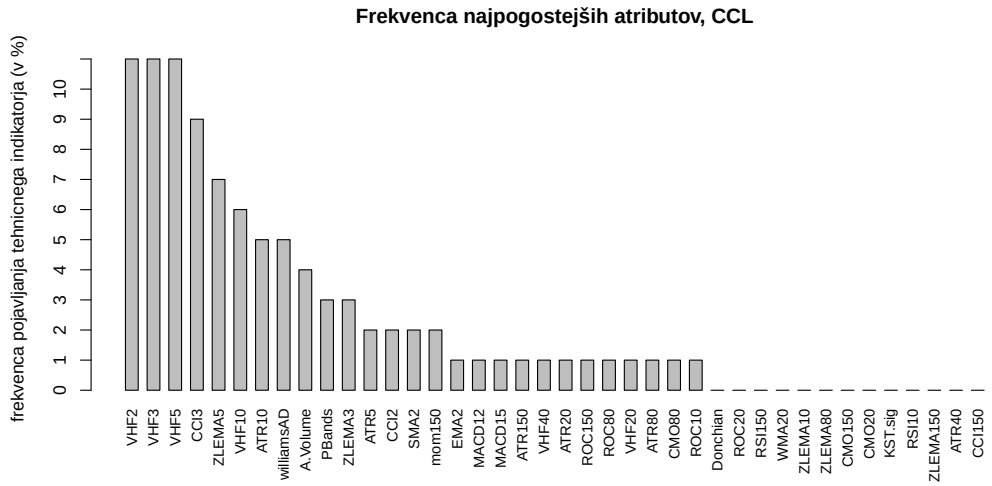
Slika 13: 40 najbolj frekventnih tehničnih indikatorjev za delnico AAPL, dobljenih z metodo FSuC–ward–comb.

6.5.2 FCBF

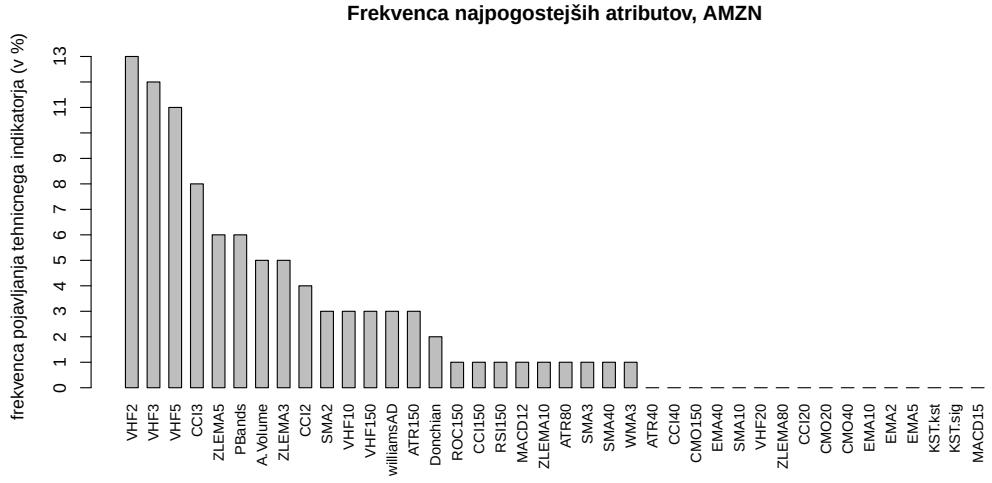
Na slikah 14, 15, 16 in 17 so prikazane frekvence pojavljanja (v %) relevantnih tehničnih indikatorjev, ki jih vrne metoda FCBF. Pri vseh delnicah in tudi pri posameznih delnicah (CCL, AMZN, AAPL) je frekvenca pojavljanja atributov VHF2, VHF3 in VHF5 najpogostejša.



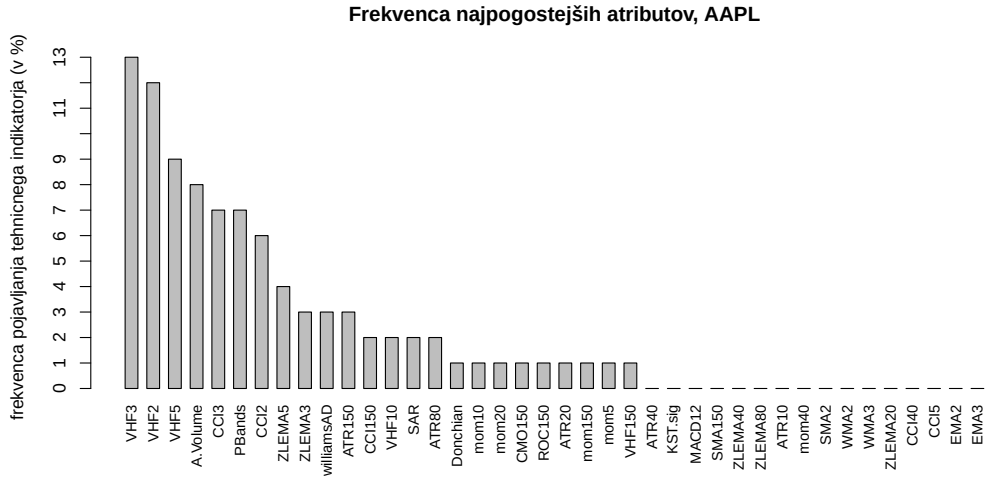
Slika 14: 40 najbolj frekventnih tehničnih indikatorjev na 370 delnicah, dobljenih z metodo FCBF.



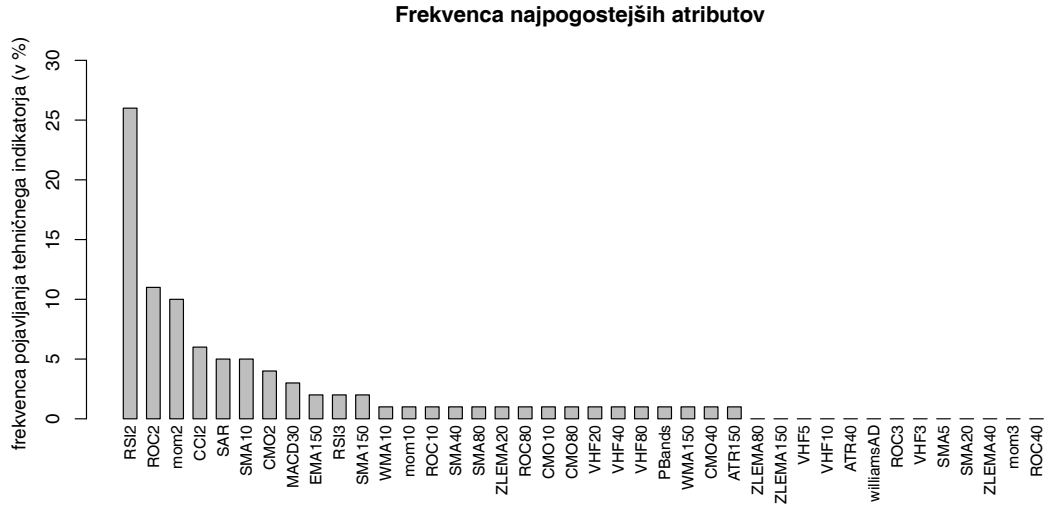
Slika 15: 40 najbolj frekventnih tehničnih indikatorjev za delnico CCL, dobljenih z metodo FCBF.



Slika 16: 40 najbolj frekventnih tehničnih indikatorjev za delnico AMZN, dobljenih z metodo FCBF.



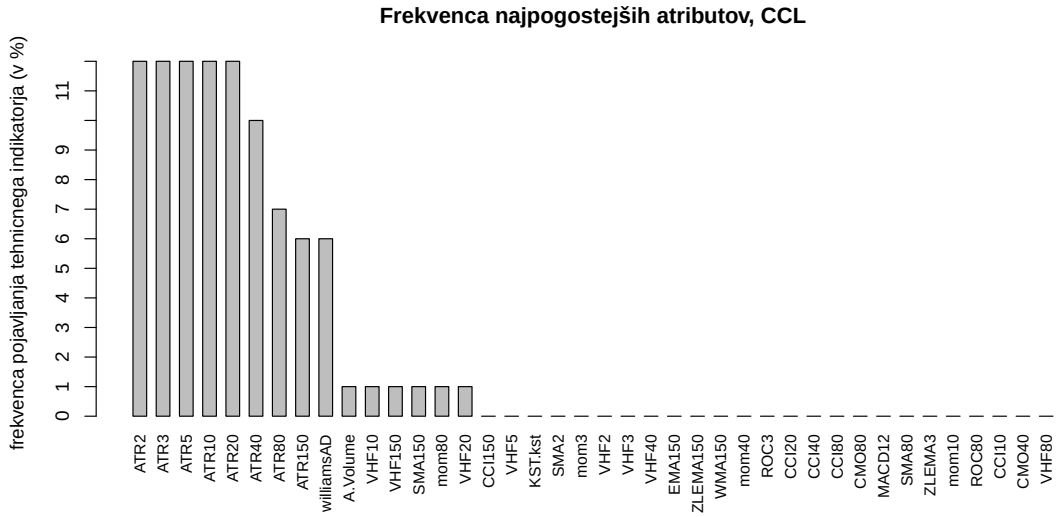
Slika 17: 40 najbolj frekventnih tehničnih indikatorjev za delnico AAPL, dobljenih z metodo FCBF.



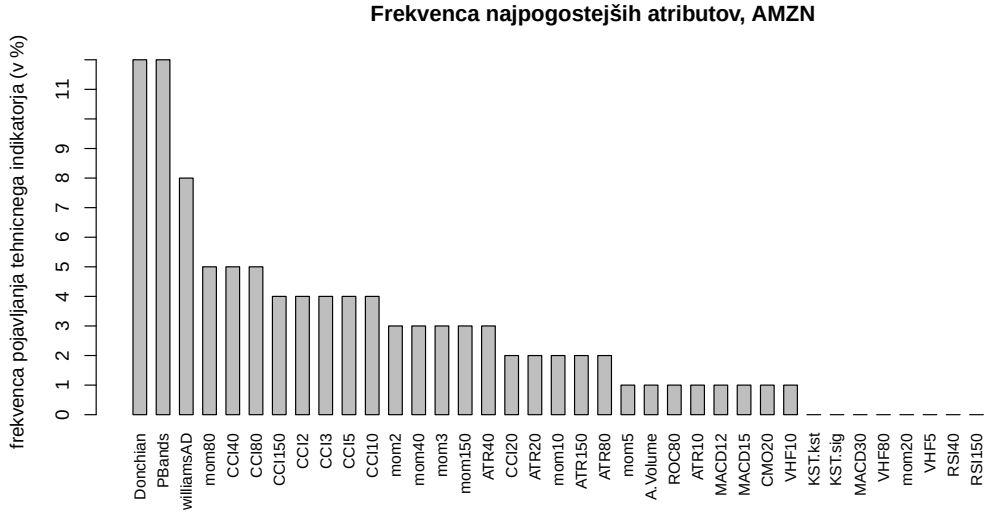
Slika 18: 40 najbolj frekventnih tehničnih indikatorjev na 370 delnicah, dobljenih z metodo CFS.

6.5.3 CFS

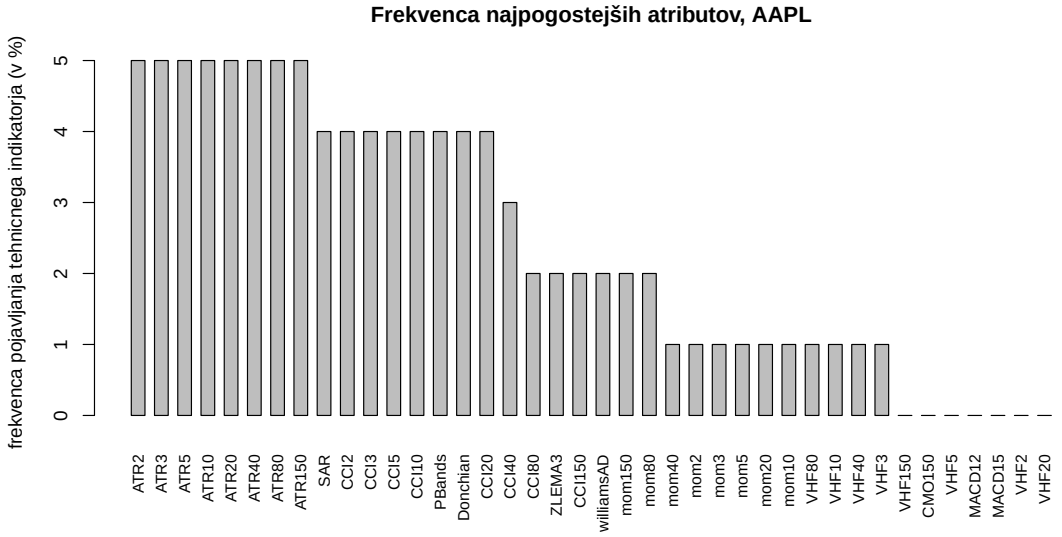
Na slikah 18, 19, 20 in 21 so prikazane frekvence pojavljanja (v %) relevantnih tehničnih indikatorjev, ki jih vrne metoda CFS. Pri vseh delnicah (slika 18) je frekvenca pojavljanja atributov RSI2, ROC2, mom2, CCI2 (momentov, oscilatorjev) najpogostejša. Zanimivo je, da se pri delnicah izrazijo različni relevantni atributi. Med bolj frekventnimi atributi pri delnicah CCL in AAPL je opazna tudi družina tehničnih indikatorjev ATR.



Slika 19: 40 najbolj frekventnih tehničnih indikatorjev za delnico CCL, dobljenih z metodo CFS.



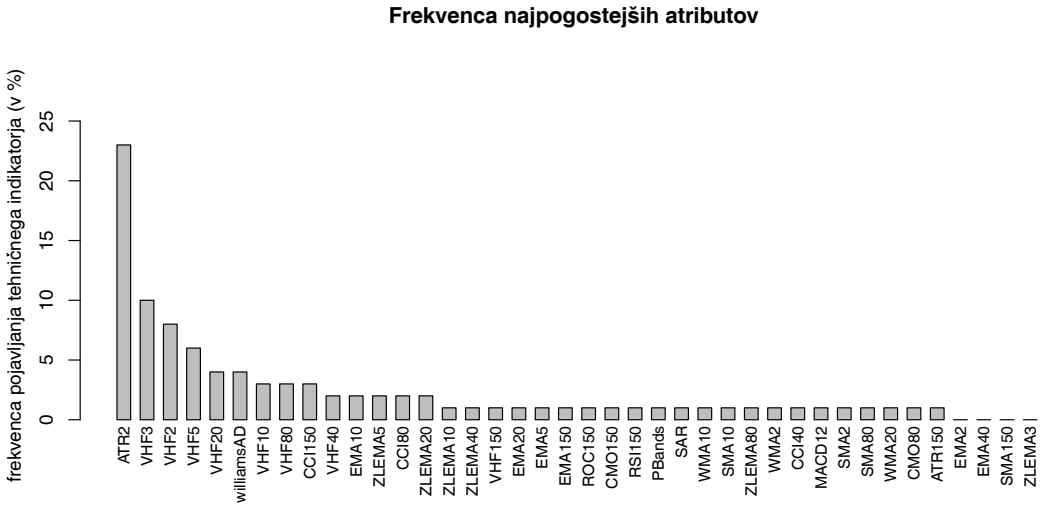
Slika 20: 40 najbolj frekventnih tehničnih indikatorjev za delnico AMZN, dobljenih z metodo CFS.



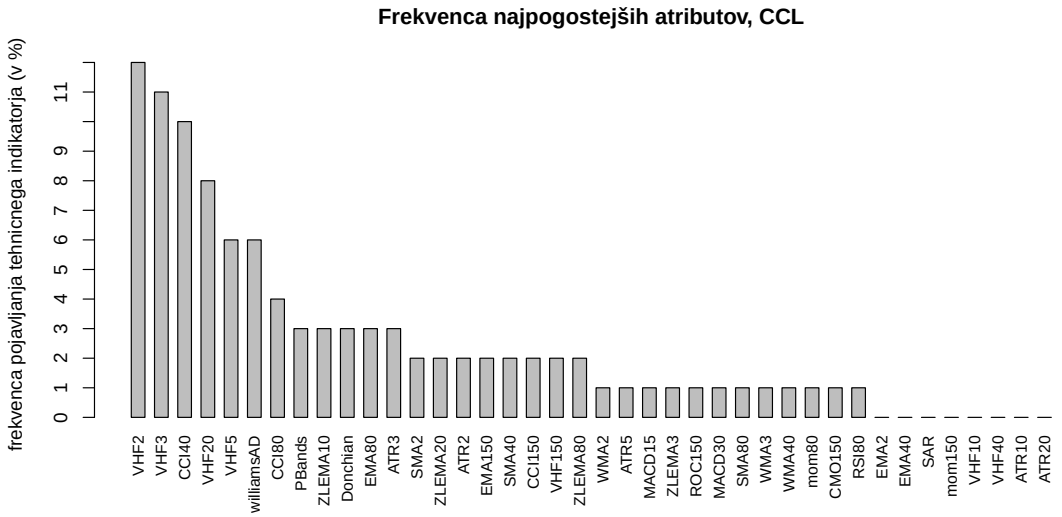
Slika 21: 40 najbolj frekventnih tehničnih indikatorjev za delnico AAPL, dobljenih z metodo CFS.

6.5.4 mRMR

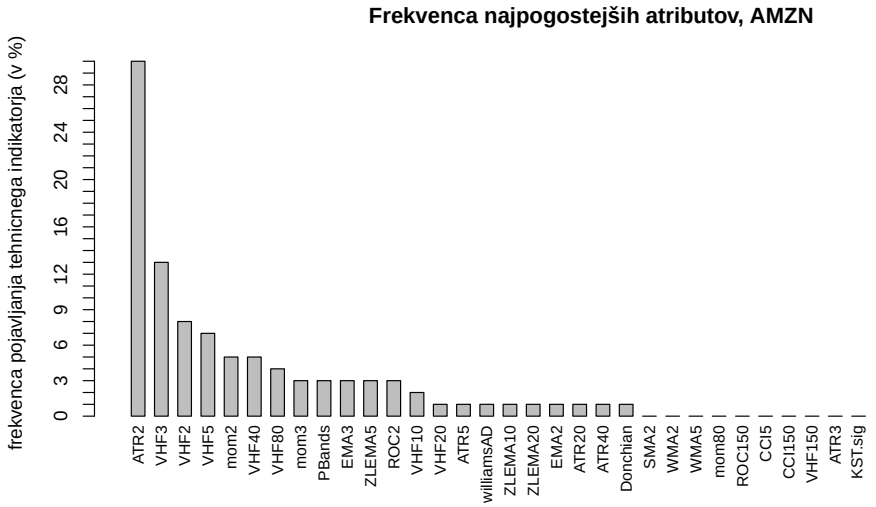
Na slikah 22, 23, 24 in 25, so prikazane frekvence pojavljanja (v %) relevantnih tehničnih indikatorjev, ki jih vrne metoda mRMR. Pri vseh delnicah je frekvenca pojavljanja atributov ATR2 ter družine tehničnih indikatorjev VHF najpogostejša (pri majhnem številu dni zajetih v izračun). Pri delnici CCL je odstotek pojavljanja atributa ATR2 zelo nizek v primerjavi z delnicama AMZN ter AAPL. Pri vseh treh omenjenih delnicah se pa družina tehničnih indikatorjev VHF pojavlja najbolj frekventno.



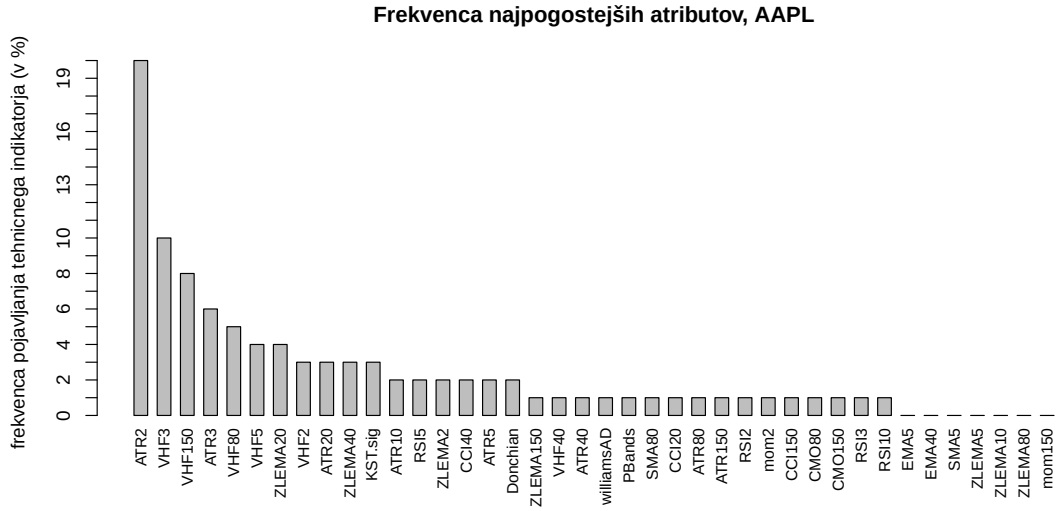
Slika 22: 40 najbolj frekventnih tehničnih indikatorjev na vseh 370 delnicah, dobljenih z metodo mRMR.



Slika 23: 40 najbolj frekventnih tehničnih indikatorjev za delnico CCL, dobljenih z metodo mRMR.



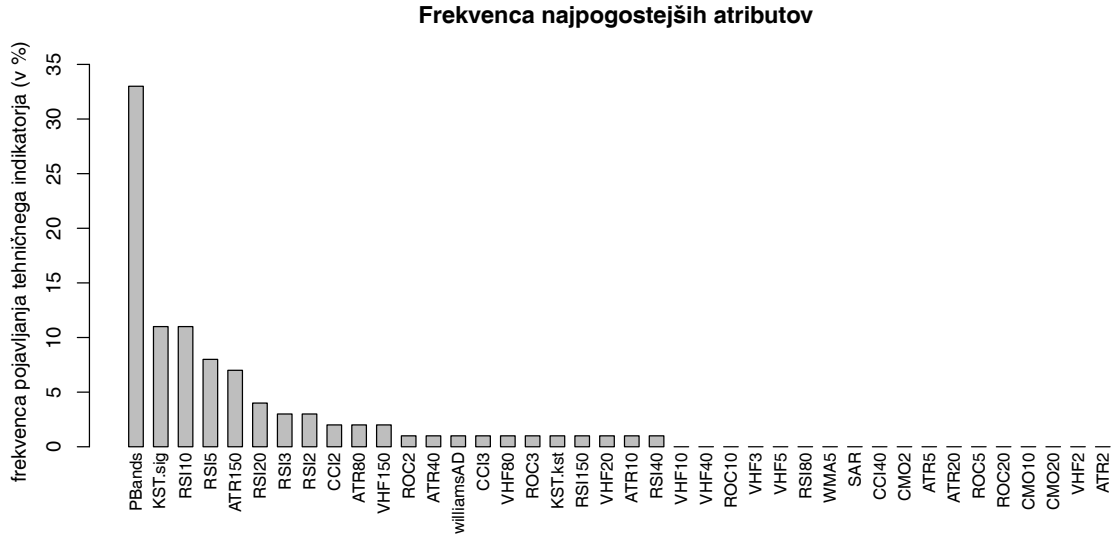
Slika 24: Najbolj frekventnih tehnični indikatorji za delnico AMZN (dobimo le 32 tehničnih indikatorjev), dobljenih z metodo mRMR.



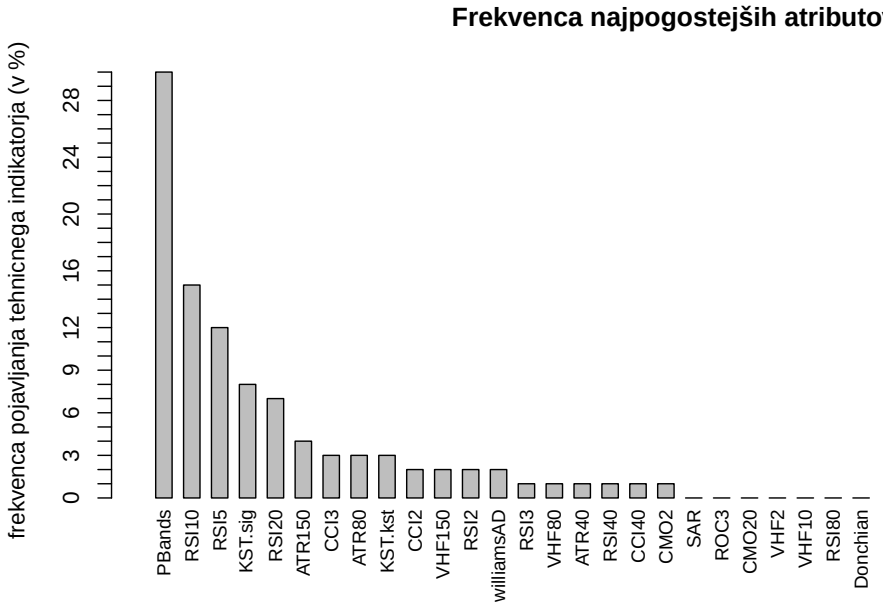
Slika 25: 40 najbolj frekventnih tehničnih indikatorjev za delnico AAPL, dobljenih z metodo mRMR.

6.5.5 CCCA

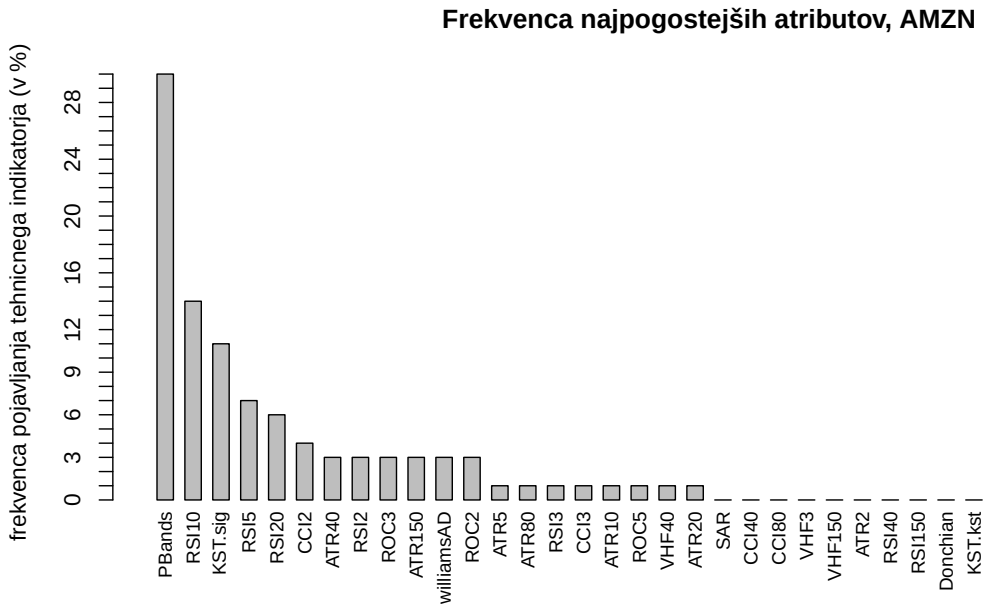
Na slikah 26, 27, 28 in 29 so prikazane frekvence pojavljanja (v %) relevantnih tehničnih indikatorjev, ki jih vrne predlagana metoda CCCA. Pri vseh delnicah je frekvenca pojavljanja atributa PBands najpogostejša, ti so tudi pri posameznih delnicah na vodilnem mestu.



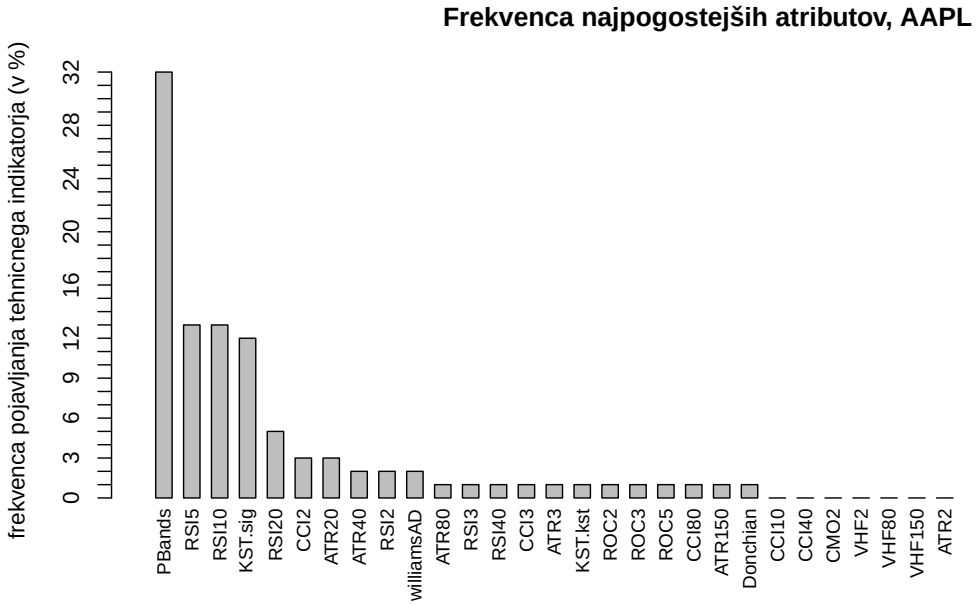
Slika 26: 40 najbolj frekventnih tehničnih indikatorjev na vseh 370 delnicah, dobljenih z metodo CCCA.



Slika 27: Najbolj frekventni tehnični indikatorji za delnico CCL (dobimo le 26 tehničnih indikatorjev), dobljenih z metodo CCCA.



Slika 28: Najbolj frekventni tehnični indikatorji za delnico AMZN (dobimo le 31 tehničnih indikatorjev), dobljenih z metodo CCCA.



Slika 29: Najbolj frekventni tehnični indikatorji za delnico AAPL (dobimo le 29 tehničnih indikatorjev), dobljenih z metodo CCCA.

7 ANALIZA USPEŠNOSTI TRGOVALNIH STRATEGIJ

7.1 Rezultati predlaganih trgovalnih strategij

Konsistentnost in uspešnost trgovalnih strategij nam kažejo rezultati testiranj na podlagi preteklih podatkov z uporabo različnih kazalcev uspešnosti trgovalnih strategij (glej poglavje 6). V tem poglavju podamo rezultate testiranja trgovalnih strategij na osnovi preteklih podatkov. V trgovalnih strategijah na podlagi vstopnih in izstopnih pravil simuliramo prodajo in nakup delnic na vnaprej določenem območju preteklih podatkov. Radi bi pokazali, da z uporabo napovedi SVM, LDA in NB klasifikacijskih modelov, ki jih uporabimo kot podporo pri odločanju v avtomatskem trgovalnem sistemu, lahko dobimo donosen trgovalni sistem. Formulirali smo množico trgovalnih pravil, katerim vodilo so napovedi gibanja delnic, ki jih vrnejo klasifikacijski modeli (glej poglavje 7.1). Predlagano trgovalno strategijo, ki je prilagojena glede na naše klasifikacijske napovedi gibanja najvišjih tečajev v trgovalnem dnevu, testiramo na 1920 trgovalnih dnevih in dnevno spreminjamo vstopne komponente trgovalnega sistema, saj smo napovedovali gibanje delnic le za en dan vnaprej.

Kvantitativni kazalci, ki smo jih upoštevali kot mero trgovalnih strategij so skupna letna stopnja rasti CAGR (enačba (13)), Sharpeov koeficient, Informacijski koeficient, kazalnik Sortino, povprečje donosov na vseh 1920 trgovalnih dnevih in spremljajoča standardna deviacija (glej poglavje 5). Rezultati v tabeli 9 kažejo, da bi najboljši rezultati pri predlaganih strategijah morali biti za LDA klasifikator (gledamo samo klasifikacijske točnosti na testni množici in preciznost za razred 1). Najslabše klasifikacijske rezultate pa poda NB klasifikator, glej tabelo 10. Rezultati so precej odvisni tudi od izbora delnic, saj jih vsak dan znova izberemo za vsak model posebej.

V Vodenih D–trgovalnih strategijah ter Naivnih strategijah vključimo prag D, ki je opisan z enačbo (10) in predstavlja najmanjši pozitiven donos v primeru, da so naše dobljene napovedi o gibanju delnic pravilne. Vsaka sprememba parametrov v Vodeni D–trgovalni strategiji močno vpliva na izid izvedbe trgovalne strategije, zato bomo z eksperimentalnim delom na učnih podatkih predhodno določili prag D, ki ga uporabimo na naslednji testni množici. Med različnimi vrednostmi pragov $D = 1\%, 2\%, 2.5\%, 3\%$ določimo tisti prag, ki ima najvišjo skupno letno stopnjo rasti CAGR na učni množici. V tabeli 15 smo prikazali izbor D pragov pri uporabi različnih klasifikacijskih modelov z uporabo metode FSuC–ward–comb. Ker imamo 96 učnih in testnih množic, dobimo 96 različnih D vrednosti za vsak klasifikator posebej. Ko predhodno določimo vrednosti pragov D, te uporabimo pri Vodenih D–trgovalnih strategijah in Naivnih strategijah. Na primer vrednosti D_{LDA}, ki jih vrne LDA klasifikator, vključimo tako pri Vodenih kot tudi pri Naivnih strategijah.

Izid trgovalnih strategij je podan v tabeli 14, kjer za klasifikacijske modele vzamemo: 'NB' Naivni Bayesov klasifikator (v tabeli 14 'Vodena D_{NB}–strat'), 'LDA' klasifikator linearne diskriminantne analize ('Vodena D_{LDA}–strat'), 'RBF' SVM klasifikator, pri katerem uporabimo RBF jedro ('Vodena D_{RBF}–strat') in 'linear', ki označuje SVM klasifikator, pri katerem uporabimo linearno jedro ('Vodena D_{lin}–strat'). Za primerjavo smo prikazali tudi rezultate za Naivne strategije, kjer uporabimo različne D

pragove: 'Naivna D_{LDA}–strat', 'Naivna D_{NB}–strat', 'Naivna D_{lin}–strat', 'Naivna D_{RBF}–strat', 'Naivna D_0–strat' ter indeks $S\&P500$ ('SPY') in Primerjalno strategijo ('bench'). D_0 predstavlja vektor samih ničel. Najvišje rezultate glede na CAGR vrednost, Sharpeov koeficient, informacijski koeficient, kazalnik Sortino in povprečje ima Vodena D_{NB}–strategija. V Vodenih D–trgovalnih strategijah smo s pomočjo klasifikacijskih modelov vključili predvidevanja o gibanju delnic. Zanimivo je, da uspešnosti izidov na izbranih kvantitativnih kazalnikih ne kažejo podobne slike kot pri klasifikacijskih rezultatih, vendar, kot bomo kasneje videli, se uspešnost izidov Vodenih D–trgovalnih strategij med seboj signifikantno ne razlikuje. Veliko je možnih razlogov, ki lahko vplivajo na izid trgovalnih strategij: izbor D pragov, izbor delnic, izbor klasifikacijskih modelov, izbor relevantnih tehničnih indikatorjev, itd.

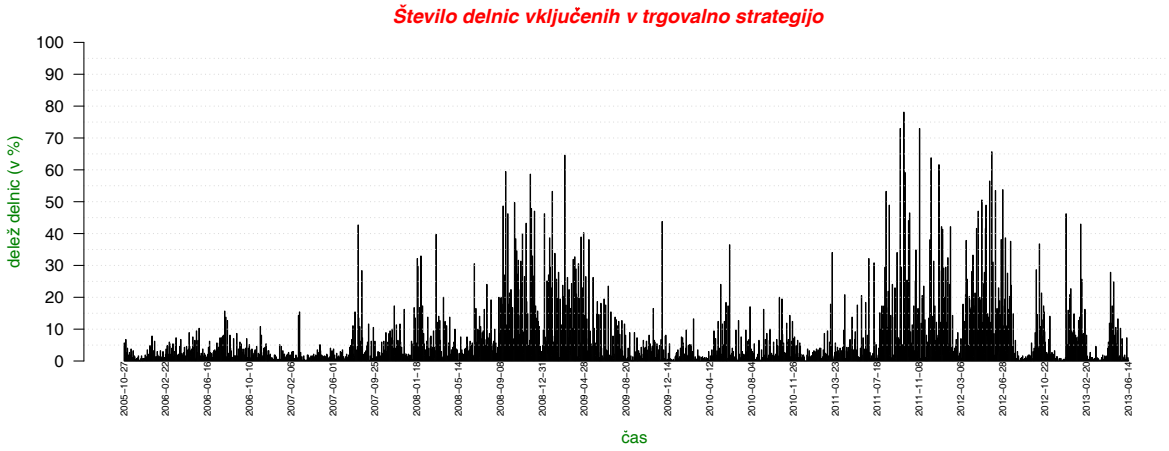
Vse Vodene D–trgovalne strategije vrnejo boljše rezultate kot Naivne D–trgovalne strategije, kar pomeni, da z vključitvijo napovedi v predlagane trgovalne strategije te dajo višje rezultate kot pa brez vključitve napovedi (glej tabelo 14 in sliko 34). S samo postavitvijo vrednosti D pragov pa presenetljivo Naivne strategije ne vrnejo samo pozitivne CAGR vrednosti, ampak je ta tudi višja od $S\&P500$ indeksa. Kot smo že omenili se pri Naivnih strategijah ne upošteva napovedi modelov, ampak samo vrednosti D: kadar je relativna razlika med $high_{t-1}$ in $open_t$ višja kot neka vnaprej podana meja, potem imamo signal za trgovanje. Te naivne napovedi rasti smo primerjali z dejanskimi izidi; klasifikacijski rezultati Naivnih strategij so prikazani v tabeli 13. Dobljeni 'naivni klasifikacijski rezultati' na testnih množicah imajo klasifikacijsko točnost nekoliko pod 40%, specifične vrednosti pa presegajo kar 60% točnosti, kjer pa ima senzitivnost na drugi strani zelo nizke vrednosti. Slednje pomeni, da pri naivnem modelu sama postavitev vrednosti pragov D ni dovolj. Kadar Naivna strategija dobi signale za trgovanje, so ti povečini napačni (zelo nizka senzitivnost), kar pomeni, da je odstotek pravilno klasificiranih primerov rasti zelo majhen. Kadar ne trgujemo, pa se izkaže za pravilno odločitev (visoka specifičnost), kar pomeni, da je odstotek pravilno klasificiranih primerov padcev visok. Razmerje predvidenih rasti/padcev je na strani negativnih donosov, torej zaradi postavljenih višjih vrednosti pragov D, Naivne trgovalne strategije pri $D > 0$ prejemajo seveda manj signalov za trgovanje. Medtem ko pa Naivna D_0 strategija, prejema več signalov za trgovanje, kar lahko vidimo tudi iz klasifikacijskih rezultatov v tabeli 13. Naivni model za prag D_0 uvrstil veliko vrednosti, kjer naj bi se zgodila rast, pravilno, padce pa uvrstil med rasti. Če povzamemo, večino gibanja je naivni model za D_0 uvrstil med rasti. Kljub temu, da je preciznost za razred 1 (odstotek pravilno klasificiranih primerov, ki so bili klasificirani kot '1') najvišja med vsemi Naivnimi strategijami, je pa donos toliko manjši zaradi manjše vrednosti postavljenega praga D_0, kar se vidi iz tabele 14.

V prilogi 8.3 prikažemo še rezultate Vodenih D–trgovalnih strategij ter Naivnih strategij, kjer za vhodne podatke pri grajenju klasifikacijskih modelov uporabimo relevantne atribute, ki smo jih dobili z nekaterimi drugimi metodami za izbor atributov. Iz rezultatov lahko vidimo (glej prilogo 8.4), da 'FSuC–ward–comb' vrne v splošnem najvišje rezultate trgovalnih strategij na podlagi izbranih kvantitativnih kazalcev.

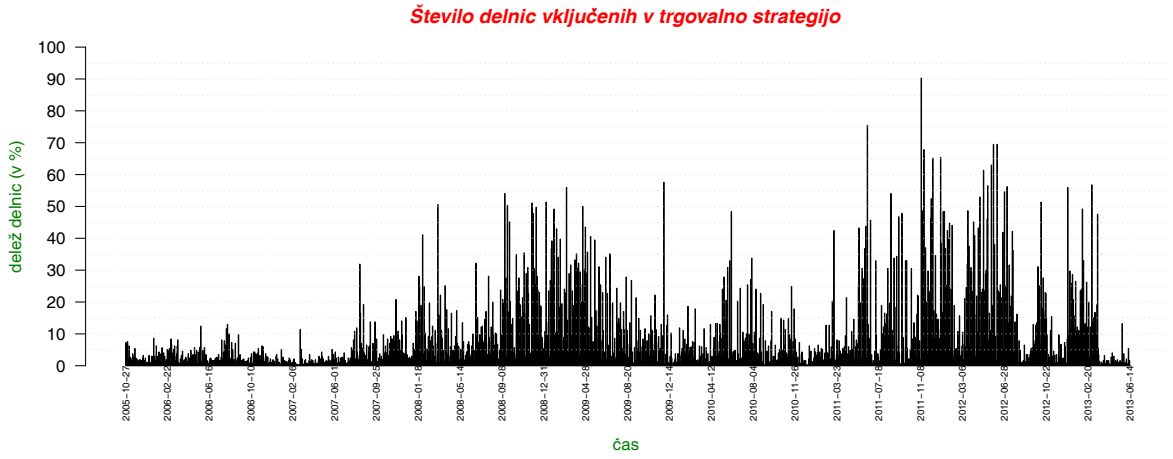
Uspešnost Vodenih D–trgovalnih strategij je odvisna tudi od izbranih delnic, s katerimi dnevno trgujemo, ki pa se skozi čas različno vključujejo v trgovalne strategije.

	test toč $\pm$ std	senzit $\pm$ std	specif $\pm$ std	preciz 1 $\pm$ std	preciz -1 $\pm$ std
Naivna D_{LDA}	34.63 $\pm$ 11.53	7.21 $\pm$ 8.34	63.34 $\pm$ 19.73	10.97 $\pm$ 15.80	38.00 $\pm$ 12.92
Naivna D_{NB}	34.60 $\pm$ 11.44	7.24 $\pm$ 8.24	63.55 $\pm$ 19.40	10.89 $\pm$ 16.02	37.91 $\pm$ 12.92
Naivna D_{RBF}	35.84 $\pm$ 11.58	6.01 $\pm$ 7.46	67.20 $\pm$ 19.28	9.83 $\pm$ 15.82	39.37 $\pm$ 12.73
Naivna D_{lin}	34.68 $\pm$ 11.46	7.22 $\pm$ 8.21	63.61 $\pm$ 19.34	10.85 $\pm$ 15.88	37.94 $\pm$ 12.81
Naivna D_0	35.33 $\pm$ 10.63	70.14 $\pm$ 14.42	0.04 $\pm$ 0.58	41.49 $\pm$ 11.57	0.10 $\pm$ 1.51

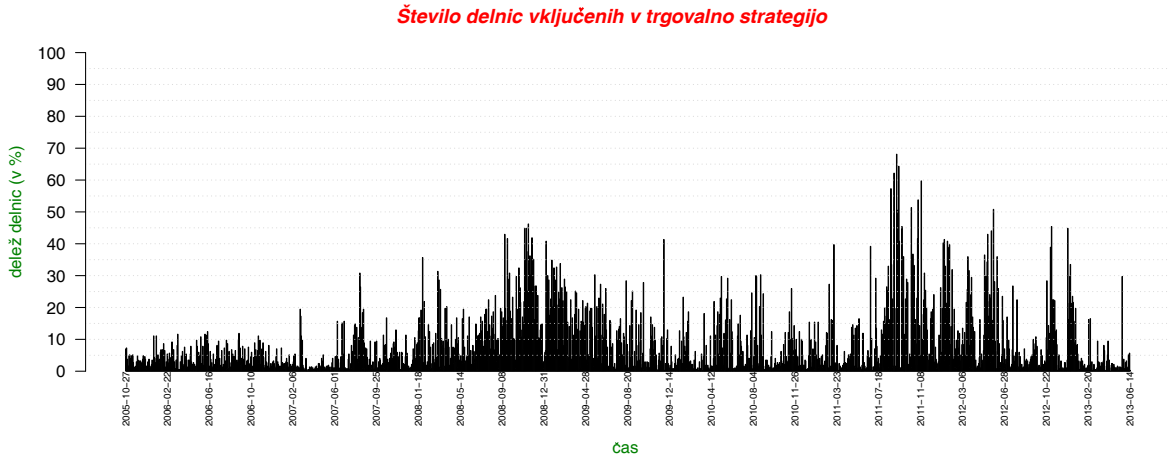
Tabela 13: Klasifikacijski rezultati 'naivnih klasifikatorjev', kjer smo za izbor atributov uporabili metodo 'FSuC–ward–comb'.



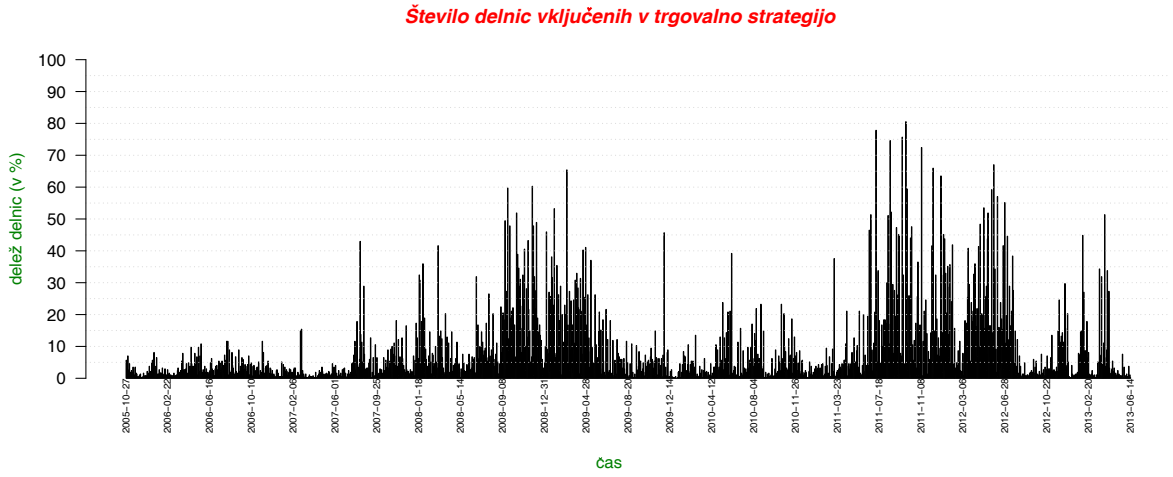
Slika 30: Delež delnic (v %) skozi testno časovno obdobje, ki jih vključimo v Vodeno D_{LDA}–trgovalno strategijo.



Slika 31: Delež delnic (v %) skozi testno časovno obdobje, ki jih vključimo v Vodeno D_{NB}–trgovalno strategijo.



Slika 32: Delež delnic (v %) skozi testno časovno obdobje, ki jih vključimo v Vodeno D_{RBF}–trgovalno strategijo.



Slika 33: Delež delnic (v %) skozi testno časovno obdobje, ki jih vključimo v Vodeno D_{lin}–trgovalno strategijo.

Na slikah 30, 31, 32 in 33 smo prikazali število vključenih delnic v Vodene D–trgovalne strategije z uporabo različnih klasifikacijskih modelov (LDA, NB, SVM z linearnim jedrom in SVM z RBF jedrom). Pri grajenju klasifikacijskih modelov smo za vhodne podatke uporabili atribute, ki smo jih dobili z 'FSuC–ward–comb' metodo. S slik je razvidno, da je intenziteta vključenih delnic v Vodene D–trgovalne strategije povečana v zadnjem četrtletju leta 2008 (okoli oktobra 2008) ter nadaljuje v leto 2009 in sredi leta 2011, 2012. V prilogi smo grafično prikazali tudi vključene delnice, kjer smo pri gradnji modelov uporabili tudi nekatere druge metode za izbor atributov (glej 8.6).

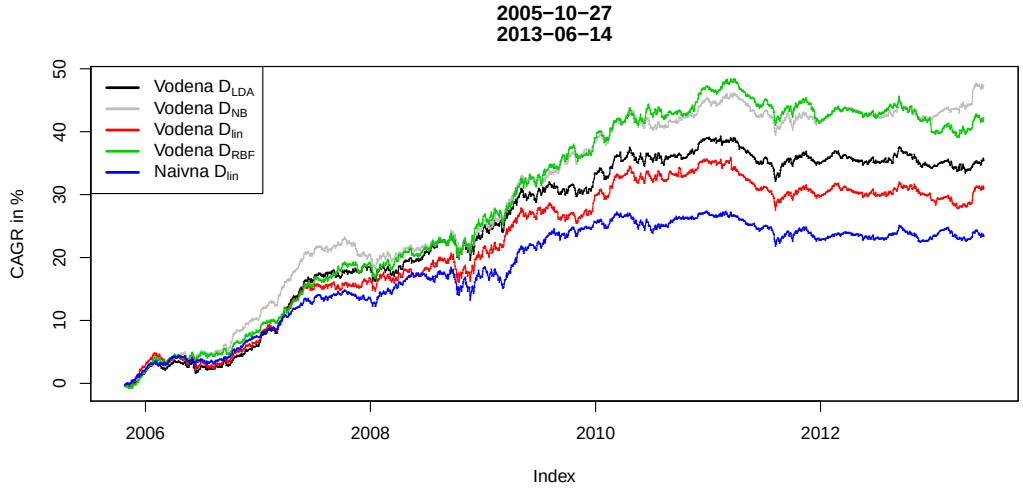
	povp(%)	**std(%)**	**Sharpe**	**Sortino**	**info koef**	**CAGR(%)**
Vodena D_{LDA}–strat	0.13	0.04	3.55	0.12	3.33	35.41
Vodena D_{NB}–strat	**0.17**	0.04	**4.56**	**0.15**	**4.46**	**47.31**
Vodena D_{lin}–strat	0.12	0.04	3.21	0.10	2.90	31.07
Vodena D_{RBF}–strat	0.15	0.04	4.01	0.13	3.92	42.16
Naivna D_{LDA}–strat	0.09	0.03	2.76	0.09	2.54	22.68
Naivna D_{NB}–strat	0.10	0.03	2.90	0.09	2.75	24.69
Naivna D_{lin}–strat	0.11	0.03	3.23	0.10	3.17	27.60
Naivna D_{RBF}–strat	0.09	0.03	2.83	0.09	2.64	23.39
Naivna D_0–strat	0.02	**0.02**	0.92	0.03	-0.27	4.03
SPY	0.03	0.03	0.81	0.03	/	4.17
Primerjalna strat (bench)	0.03	0.03	1.03	0.03	0.25	5.78

Tabela 14: Rezultati izvedbe Vodenih D–trgovalnih strategij in primerjava z Naivnimi strategijami, indeksom $S\&P500$ ter Primerjalno strategijo.

V eksperimentalnem delu smo primerjali izvedbo trgovalnih strategij med seboj s pomočjo Wilcoxonovega testa s predznačenimi rangi, ki ga uporabljamo za ugotavljanje razlik med dvema povprečnima

D	1%	2%	2.5%	3%
D_{LDA}	12	23	52	9
D_{NB}	13	13	62	8
D_{RBF}	13	18	53	12
D_{lin}	5	25	46	20

Tabela 15: Frekvence uporabljenih D pragov za vsak klasifikator posebej. Vrednosti D pragov smo izbirali med $D = 1\%, 2\%, 2.5\%, 3\%$. Uporabili smo FSuC–ward–comb metodo.



Slika 34: Z zgornje slike lahko vidimo, da so najvišji rezultati izvedbe Vodene D–trgovalne strategije, kadar uporabimo NB klasifikator, tesno mu sledi SVM klasifikator z RBF jedrom. Vse predlagane Vodene D– trgovalne strategije vrnejo boljše rezultate kot pa Naivne strategije, indeks $S\&P500$ in Primerjalna strategija. Po sedmih letih izvedbe predlaganih trgovalnih strategij presežemo vrednost CAGR čez 40%.

vrednostma za neodvisna vzorca, ko proučevana številska spremenljivka ni normalno porazdeljena ali za opisne spremenljivke, merjene na ordinalni skali. Test predstavlja neparametričen ekvivalent parametričnemu t–testu. Vrednosti številske spremenljivke se pretvorijo v range, tako da se najmanjši vrednosti pripiše rang 1, naslednji najmanjši rang 2, itd. Za izračun testne statistike se uporabijo vrednosti rangov. Pri Wilcoxonovem testu s predznačenimi rangi je testna statistika W_s, ki je pri enako velikih skupinah enaka manjši od obeh vsot rangov skupine oziroma vsoti rangov manjše skupine, ko skupini nista enako veliki. Vrednost statistike W_s je statistično značilna pri $p < 0.05$, če je njena absolutna standardizirana vrednost z večja od 1.96 [5]

Potek testa: naj bo N velikost meritev oz. število parov. Skupaj imamo $2N$ meritev. Naj bo $x_{1,i}$ in $x_{2,i}$ oznaka za meritve za $i = 1, \ldots, N$.

H_0 : mediana med razlikami parov je 0

H_1 : mediana razlik ni enaka 0.

- Za $i = 1, \ldots, N$ izračunaj $|x_{2,i} - x_{1,i}|$ in $\text{sgn}(x_{2,i} - x_{1,i})$, kjer sgn predstavlja funkcijo predznaka.
- izključi pare, ki imajo $|x_{2,i} - x_{1,i}| = 0$. Naj bo N_r število preostalih parov.
- Uredi pare N_r od najmanjše do najvišje vrednosti $|x_{2,i} - x_{1,i}|$.
- Rangiraj pare tako, da ima najnižja vrednost para številko 1. Naj R_i predstavlja rang.
- Izračunaj testno statistiko W

$$W = |\sum_{i=1}^{N_r} [\text{sgn}(x_{2,i} - x_{1,i}) \cdot R_i]|$$

- Ko narašča N_r, vzorčna porazdelitev W konvergira k normalni porazdelitvi. Če je $z > z_\alpha$, potem zavrnemo H_0 ali če je $W \geq W_{\alpha,N_r}$ potem zavrnemo H_0.

7.2 Statistična analiza z Wilcoxonovim testom s predznačenimi rangi

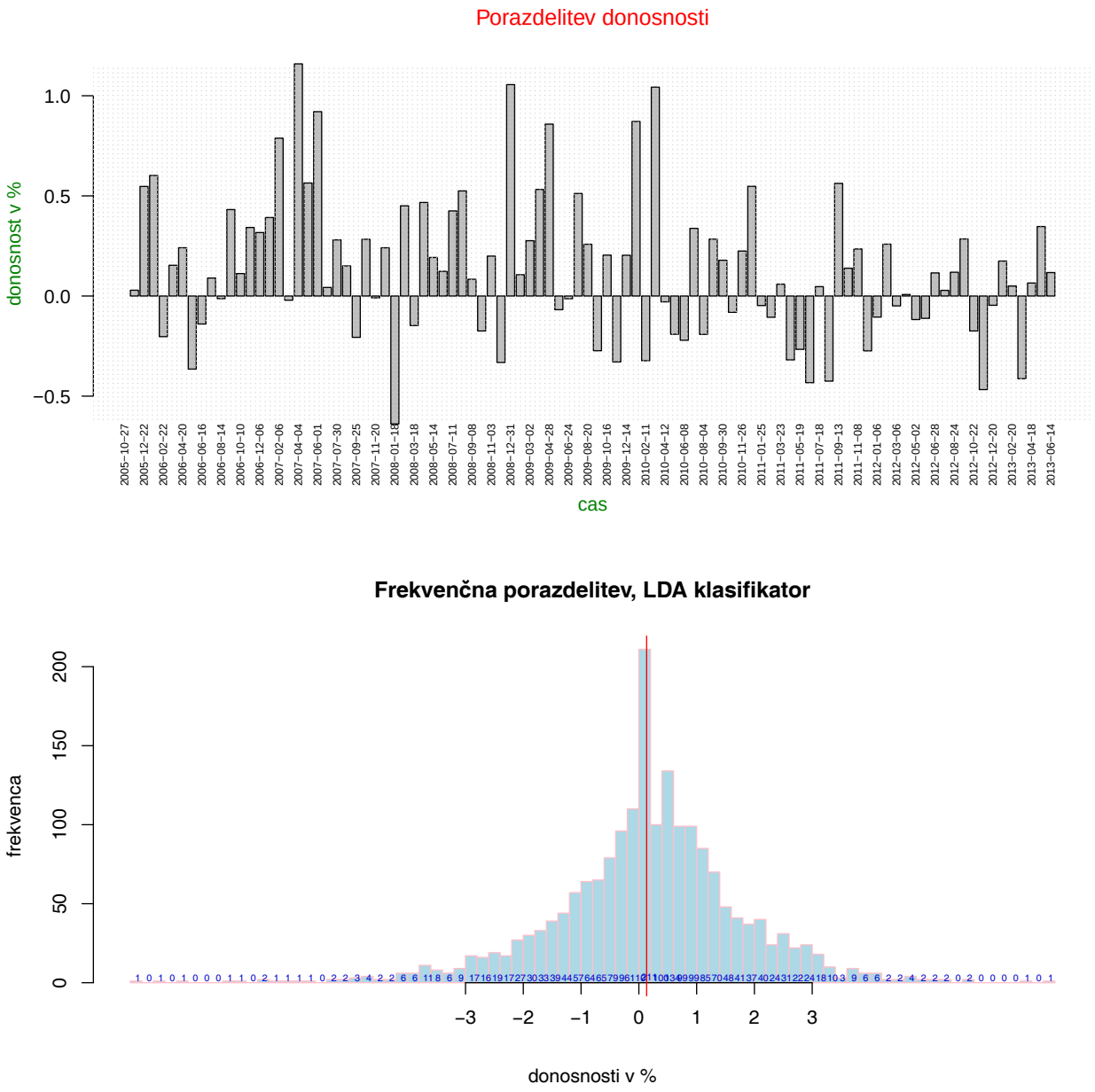
Oceniti želimo, ali se 2 povezana vzorca dnevnih donosov razlikujeta v rangih median. V ta namen uporabimo Wilcoxonov test s predznačenimi rangi. V tabeli 16 so prikazane p-vrednosti Wilcoxonovih testov s predznačenimi rangi dveh merjenih časovnih vrst dnevnih donosov (dobljenih z dvema različnima trgovalnima strategijama, ki ju želimo primerjati) in katerih porazdelitev njunih razlik je simetrična okoli median.
V tabeli 16 lahko opazimo, da se uspešnost Vodenih trgovalnih strategij med seboj ne razlikujejo signifikantno na danem vzorcu dnevnih donosov, tako da ne moremo zaključiti, katera izvedba trgovalnih strategij je boljša. Večina Vodenih D–trgovalnih strategij se signifikantno razlikuje od Naivne D_{lin} in Naivne D_0–trgovalnih strategije (izjema je Vodena D_{lin}–trgovalna strategija), kar lahko vidimo s slike 34 in tabele 14. Za indeks S&P500 in Primerjalno strategijo ne moremo zaključiti, katera od njiju je boljša (p-vrednost je 0.3131). Primerjalna strategija v bistvu predstavlja nam bolj predstavljiv indeks, ki vključuje vseh 370 delnic, ki jih enakomerno utežimo v trgovalni strategiji. Vse Vodene D–trgovalne strategije se statistično razlikujejo tudi od indeksa in Primerjalne strategije. V prilogi 8.5 smo podali rezultate Wilcoxonovih testov s predznačenimi rangi tudi za ostale metode FCBF, CFS, mRMR in CCCA.

	Naivna D_{lin}	Vodena D_{LDA}	Vodena D_{NB}	Vodena D_{RBF}	Vodena D_{lin}	SPY	bench
Naivna D_0	0.0000^{**}	0.0000^{**}	0.0000^{**}	0.0000^{**}	0.0000^{**}	0.6912	0.6753
Naivna D_{lin}		0.0405^{*}	0.0001^{**}	0.0047^{**}	0.1454	0.0000^{**}	0.0000^{**}
Vodena D_{LDA}			0.3425	0.6668	0.2387	0.0000^{**}	0.0000^{**}
Vodena D_{NB}				0.4671	0.0680	0.0000^{**}	0.0000^{**}
Vodena D_{RBF}					0.4643	0.0000^{**}	0.0000^{**}
Vodena D_{lin}						0.0000^{**}	0.0000^{**}
SPY							0.3131

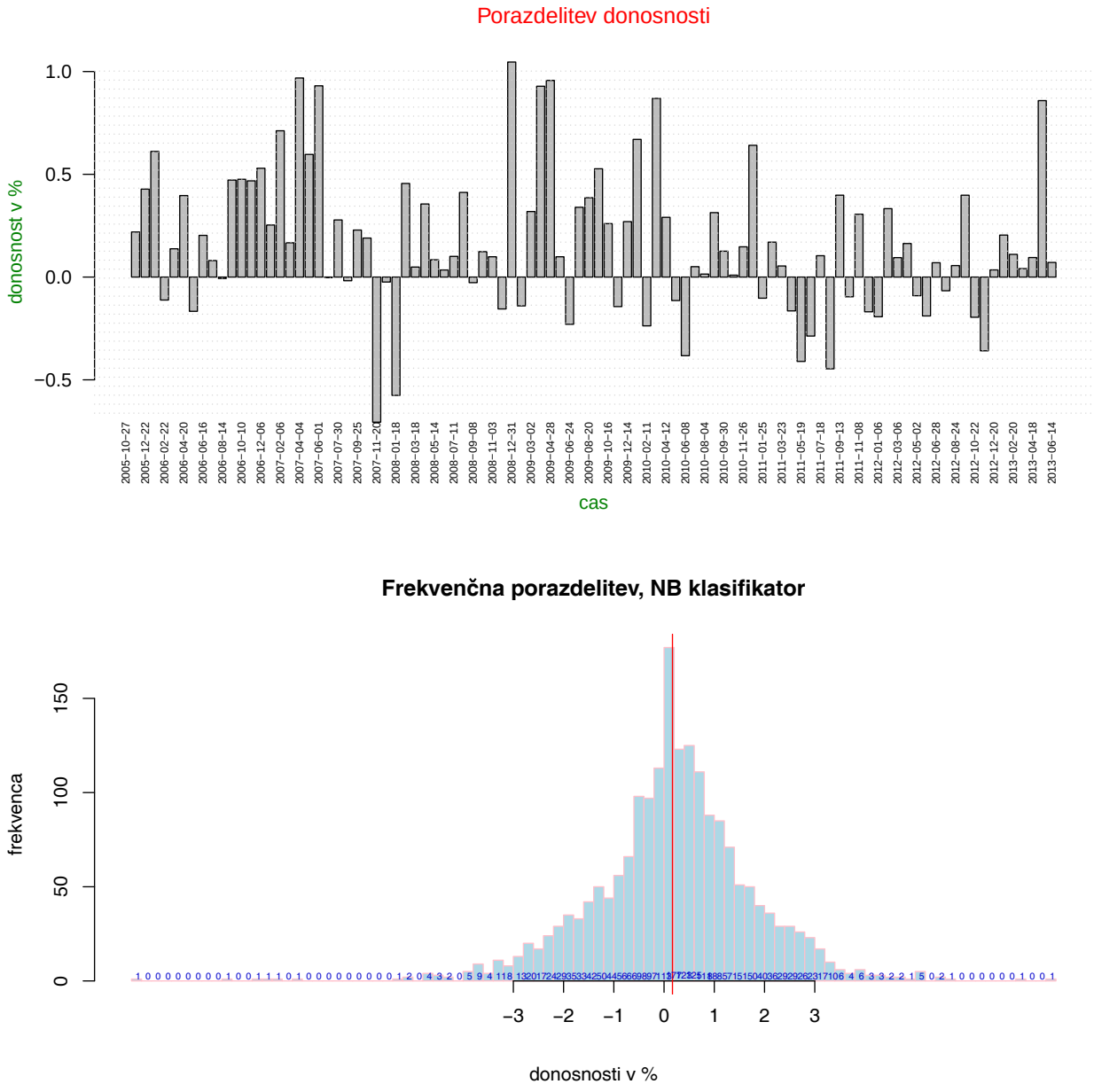
Tabela 16: Izveden Wilcoxonov test s predznačenimi rangi. Vhodna podatka za izvedbo testa sta vektorja dnevnih donosov, ki jih vrnejo predlagane trgovalne strategije ali pa indeks na celotnem testnem obdobju. Prikazane p-vrednosti med 0.01 in 0.05 smo označili z eno zvezdico $*$ ter p-vrednosti manjše kot 0.01 smo označili z dvema zvezdicama $**$.

7.3 Porazdelitev donosnosti skozi čas, FSuC–ward–comb metoda

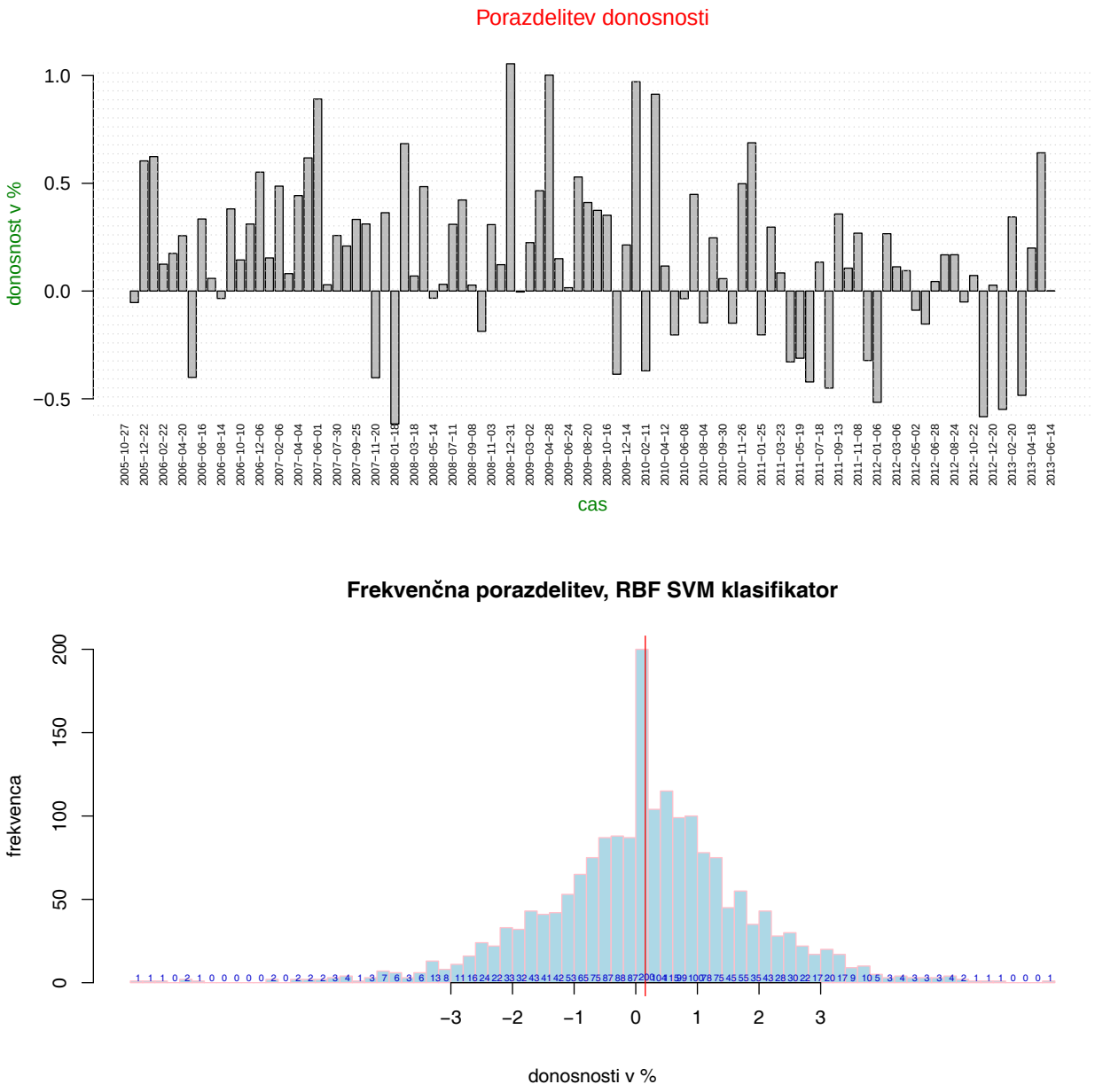
Zanima nas, kako se donosnosti Vodenih D–trgovalnih strategij spreminjajo skozi čas in kakšne so njihove vrednosti. Na slikah 35, 36, 37 in 38, smo prikazali po 2 grafa in sicer porazdelitev donosnosti v odvisnosti od časa, kjer vsak stolpec predstavlja povprečje vseh donosnosti v enem trgovalnem mesecu ter frekvenčno porazdelitev dnevnih donosnosti. Grafi ponazarjajo, da je več pozitivnih donosov in ti dosegajo višje vrednosti od negativnih donosov, ki imajo tudi manjši razpon. Skozi obdobje pri vseh različicah Vodenih D–trgovalnih strategij se pojavljajo podobne donosnosti. Odstotek vseh dni, ko smo trgovali s pozitivno donosnostjo, je 59.11% za LDA klasifikator (Vodeno D_{LDA}–strategijo), 58.75% za NB klasifikator (Vodeno D_{NB}–strategijo), 58.96% za SVM z RBF jedrom (Vodeno D_{RBF}–strategijo) ter 57.40% za SVM z linearnim jedrom (Vodeno D_{lin}–strategijo). Za primerjavo: z indeksom $S\&P500$ je 54.89% dni s pozitivno donosnostjo, porazdelitev donosnosti za indeks smo prikazali na sliki 39, kjer se vidi, da je negativnih donosov nekoliko več v primerjavi z Vodenimi D–trgovalnimi strategijami ter imajo višje negativne vrednosti. Posebej je to razvidno konec leta 2008 (v času krize). Za primerjavo donosnosti Vodenih D–trgovalnih strategij glej poglavje 8.7.



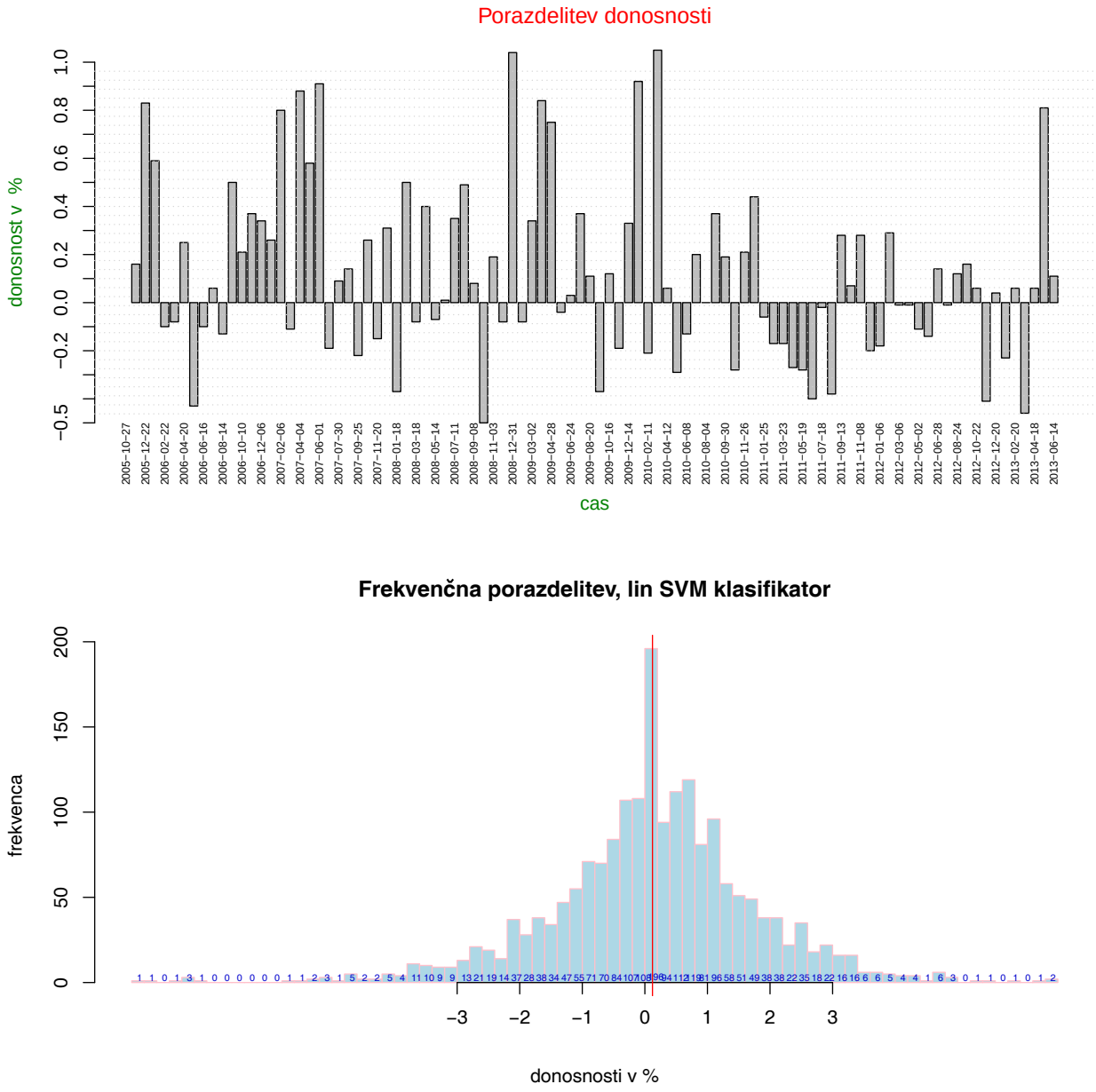
Slika 35: Porazdelitev donosnosti skozi čas ter frekvenčna porazdelitev, Vodena D_{LDA}–strategija. Povprečje donosnosti je 0.1346%.



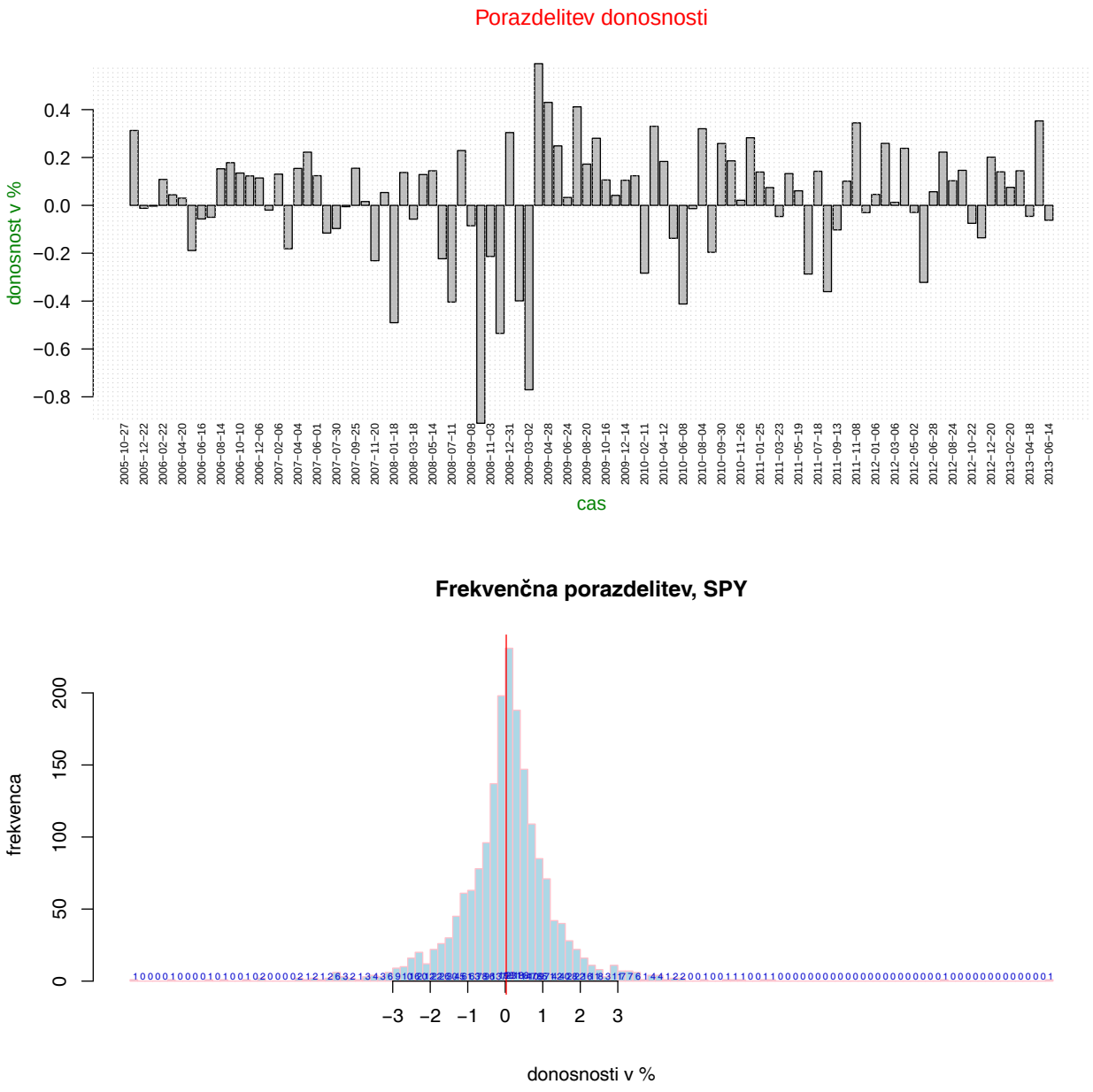
Slika 36: Porazdelitev donosnosti skozi čas ter frekvenčna porazdelitev, Vodena D_{NB}–strategija. Povprečje donosnosti je 0.167%.



Slika 37: Porazdelitev donosnosti skozi čas ter frekvenčna porazdelitev, Vodena D_{RBF}–strategija. Povprečje donosnosti je 0.154%.



Slika 38: Porazdelitev donosnosti skozi čas ter frekvenčna porazdelitev, Vodena D_{lin}–strategija. Povprečje donosnosti je 0.122%.



Slika 39: Porazdelitev donosnosti skozi čas ter frekvenčna porazdelitev za indeks $S\&P500$. Povprečje donosnosti je 0.03%.

8 PRILOGE

8.1 Prikaz klasifikacijskih rezultatov na testni množici

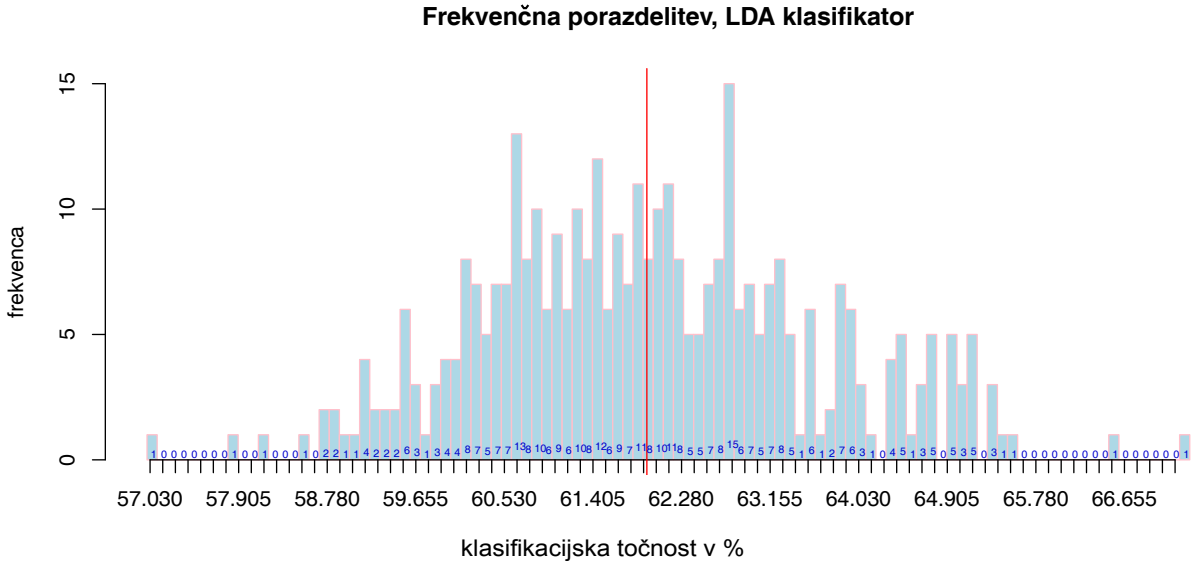
Za boljši vpogled v dobljene rezultate smo klasifikacijske točnosti na testnih množicah prikazali tudi grafično.

8.1.1 Prikaz klasifikacijskih rezultatov po delnicah za FSuC–ward-comb metodo

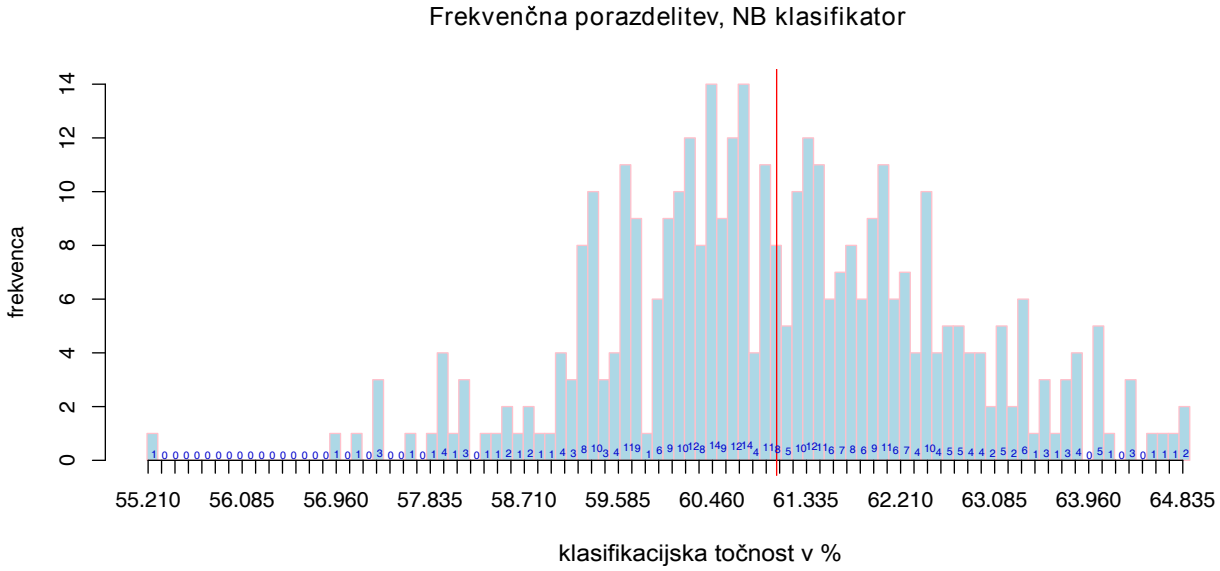
Na slikah 40, 41, 42 in 43, so 4 histogrami, kjer vsak histogram predstavlja frekvenco pojavljanja povprečne klasifikacijske točnosti delnic na vseh 1920 trgovalnih dnevih. Rezultati se gibljejo v razponu od 55.21%–67.29%; najslabše se odreže NB klasifikator, z razponom 55.21%–64.90% klasifikacijske točnosti, najbolje pa LDA klasifikator, z razponom 57.03%–67.24% klasifikacijske točnosti. Zanimivo je, da se med slabšimi delnicami (po klasifikacijski točnosti) pri vseh klasifikatorjih pojavljajo iste delnice, ravno tako med boljšimi delnicami.
Pri ostalih metodah so klasifikacijske točnosti nekoliko nižje. Pri CFS metodi se klasifikacijske točnosti na testni množici gibljejo v razponu od 46.77%–64.01%, kjer najvišje vrednosti doseže s klasifikatorjem SVM z linearnim jedrom. Tudi povprečne klasifikacijske točnosti na vseh 370 delnicah kažejo, da je SVM z linearnim jedrom klasifikator z najvišjimi rezultati (glej tabelo 11, poglavje 6.4).
Pri FCBF metodi je razpon klasifikacijske točnosti na testni množici od 53.39%–63.70%, kjer SVM z RBF jedrom vrne najvišje vrednosti, ravno tako pri vseh povprečnih klasifikacijskih točnostih (glej tabelo 12, poglavje 6.4).
Tudi pri mRMR metodi SVM z RBF jedrom ta vrne najvišje klasifikacijske točnosti na testni množici. Razpon točnosti je od 48.80%–60.94%.
Klasifikacijske točnosti v razponu od 49.90%–59.79% dosežemo s CCCA metodo, kjer SVM z linearnim jedrom dosega najvišje rezultate, vendar pa ne pri vseh povprečnih klasifikacijskih točnostih na testni množici (LDA klasifikator vrne višje povprečje).
V tabeli 17 smo prikazali vrstni red klasifikacijske točnosti na testni množici (od 1 do 370) za delnice AAPL, AMZN in CCL. Nižji vrstni red pomeni nižje klasifikacijske točnosti. Pri vseh klasifikacijskih modelih in metodah za izbor atributov se delnice z nekaj izjemami obnašajo podobno. Delnica CCL se nahaja nekje v prvi četrtini, AMZN v sredini ter AAPL v zadnji četrtini po rezultatih klasifikacijskih točnosti.

delnica/metoda	FSuC				FCBF				CFS				mRMR				CCCA			
	LDA	NB	RBF	lin	LDA	NB	RBF	lin	LDA	NB	RBF	lin	LDA	NB	RBF	lin	LDA	NB	RBF	lin
AAPL	370	370	368	370	281	299	318	292	360	368	196	367	324	360	308	328	369	370	369	368
AMZN	228	256	192	179	256	239	264	225	272	320	271	309	304	268	287	240	98	88	78	78
CCL	1	1	2	1	78	16	37	11	76	69	13	11	109	212	28	126	79	69	117	15

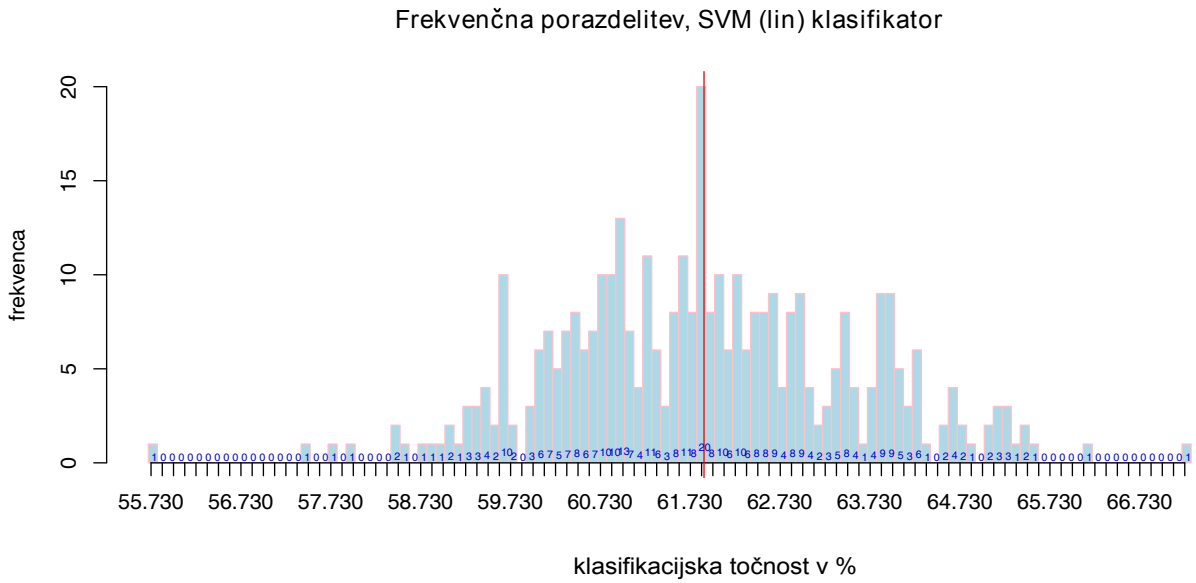
Tabela 17: Vrstni red izbranih delnic glede na klasifikacijske točnosti.



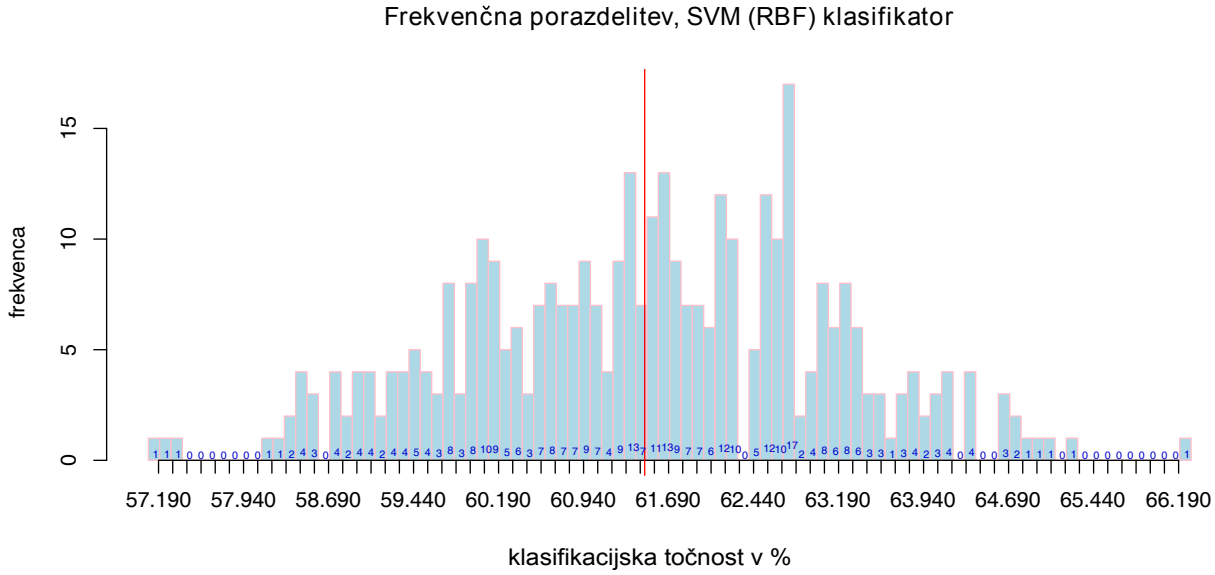
Slika 40: Predstavitev klasifikacijskih točnosti na testnih množicah. Uporabili smo LDA klasifikator ter FSuC–ward–comb metodo.



Slika 41: Predstavitev klasifikacijskih točnosti na testnih množicah. Uporabili smo NB klasifikator ter FSuC–ward–comb metodo.



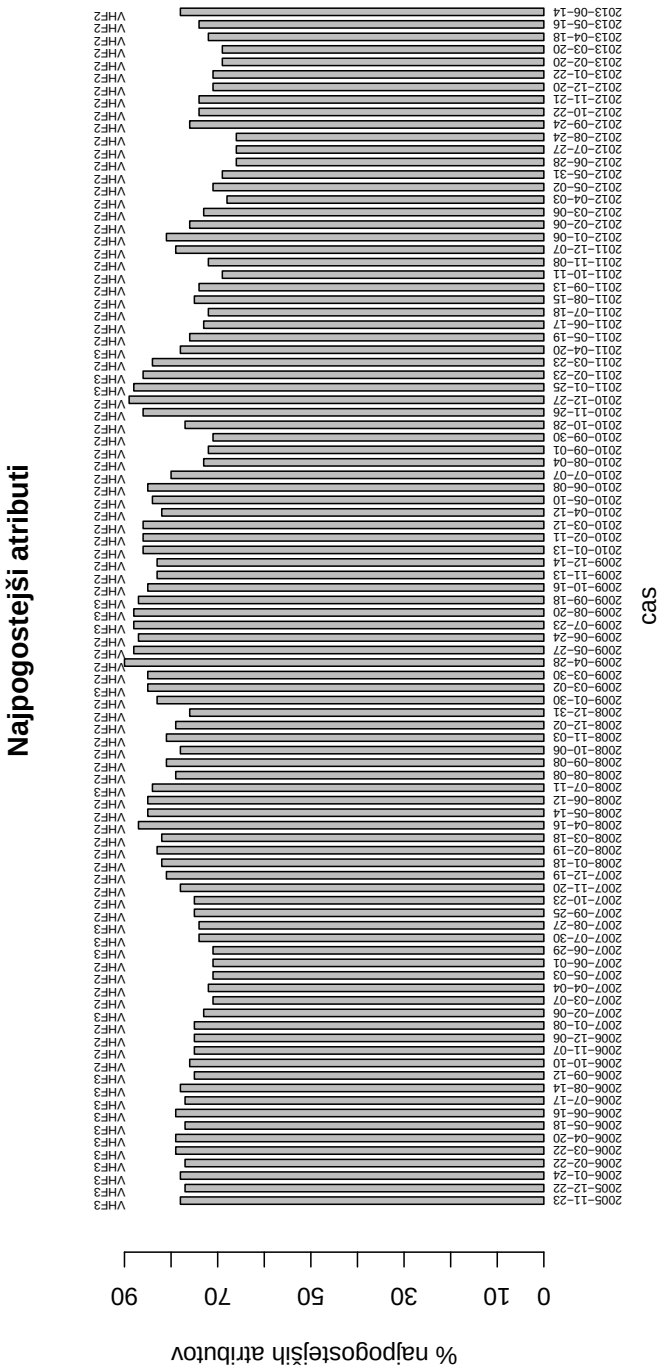
Slika 42: Predstavitev klasifikacijskih točnosti na testnih množicah. Uporabili smo linearen SVM klasifikator ter FSuC–ward–comb metodo.



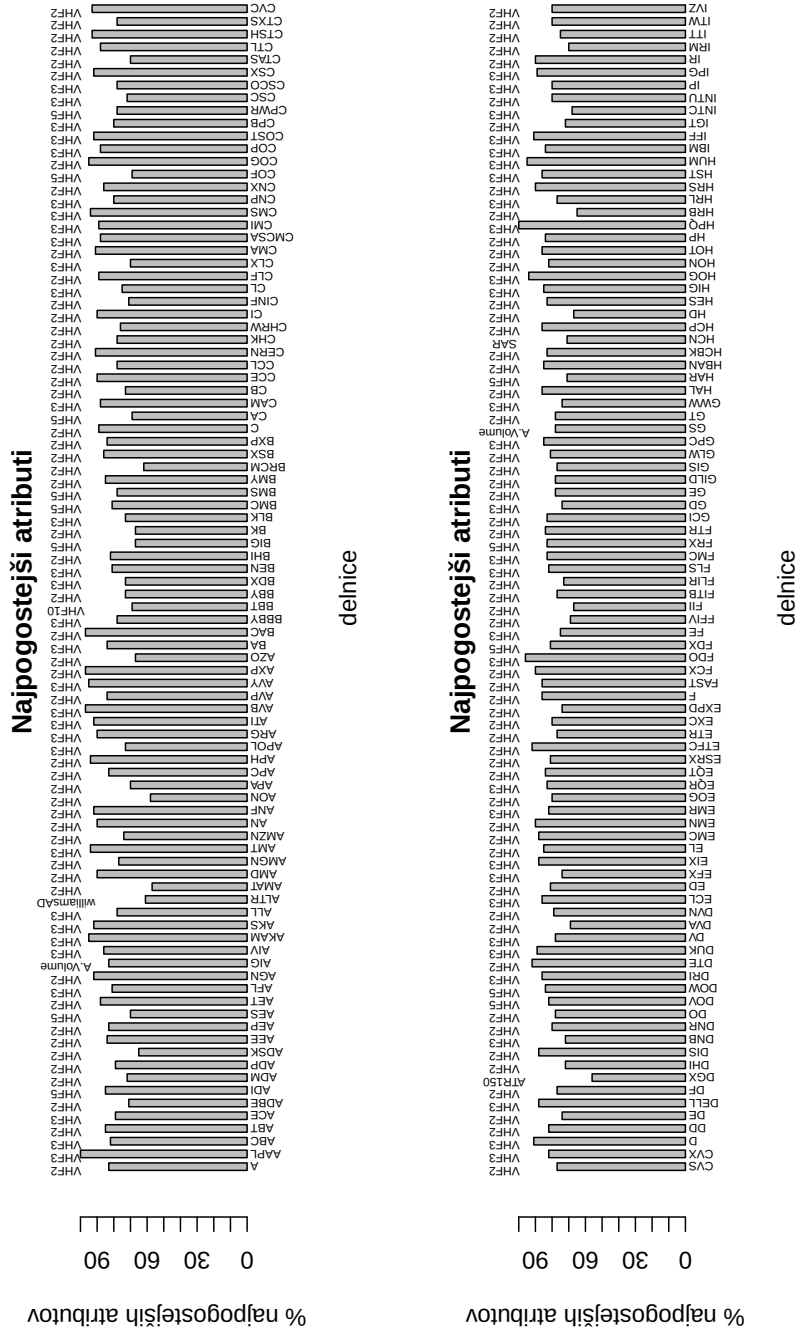
Slika 43: Predstavitev klasifikacijskih točnosti na testnih množicah. Uporabili smo RBF SVM klasifikator ter FSuC–ward–comb metodo.

8.2 Relevantni atributi oziroma najpogostejši tehnični indikatorji pri grajenju modelov

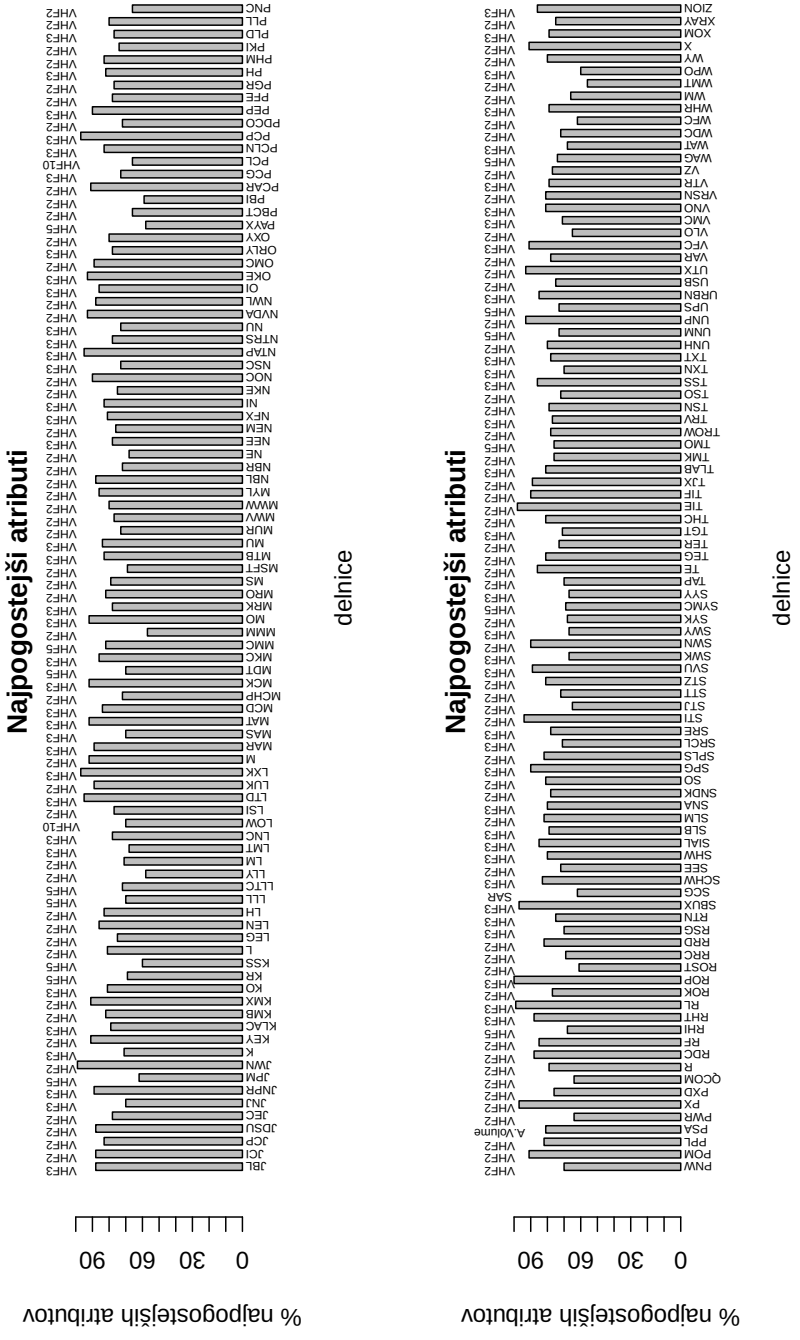
Na slikah 44, 47, 50, 53 in 56, smo grafično predstavili relevantne atribute, ki jih uporabimo kot vhodne podatke za grajenje modelov z uporabo različnih metod za izbor atributov: FCBF, CFS, FSuC–ward–comb, mRMR ter CCCA. Z različnimi metodami dobimo različne relevantne atribute. Stolpec v histogramu predstavlja na določen datum odstotek najbolj frekventnega relevantnega atributa, ki je prišel v poštev pri grajenju modelov. Na posamezen datum smo tako zgradili 370 različnih modelov, saj imamo 370 različnih delnic. Za vsak model pa lahko pride v poštev več relevantnih atributov.
Na slikah 45, 46, 48, 49, 51, 52, 54, 55, 57 in 58 je prikazana zastopanost najbolj pogosto uporabljenih relevantnih atributov po delnicah. Stolpec v prikazanih histogramih predstavlja za vsako delnico odstotek najbolj frekventnega relevantnega atributa. Za posamezno delnico smo zgradili 96 modelov, saj smo imeli 96 različnih časovnih oken.
Najbolj pogosto uporabljena relevantna indikatorja pri metodi FCBF se skozi čas kažeta VHF3 in VHF2 (glej sliko 44), kar je razvidno tudi iz histogramov na slikah 45 in 46.
Pri CFS metodi prevladuje atribut RSI2 (glej slike 47, 48 in 49). Skozi časovno obdobje se pri predlagani metodi (FSuC–ward–comb) za izbor atributov najpogosteje pojavlja SMA2 (glej histogram 50), vendar je odstotek pogostosti relativno nizek, kar lahko nakazuje na pestrost izbora različnih relevantnih atributov po delnicah. Če pogledamo histograme po delnicah 51 in 52, vidimo, da prevladuje atribut SMA2 skupaj z nekaterimi ostalimi drsečimi povprečji kot so npr. WMA3, EMA3, EMA2, SMA3, ZLEMA5, itd. Pri mRMR metodi je razvidno spreminjanje relevantnih atributov skozi čas. Prvi del časovnega obdobja je najbolj pogosto uporabljen atribut VHF3 (glej sliko 53), katerega frekventnost skozi čas prične upadati in nadomesti ga ATR2, katerega frekventnost skozi čas v splošnem raste. Pri metodi CCCA je atribut PBands prisoten pri izgradnji skoraj vsakega modela, kar je razvidno tudi iz slik 57 in 58, kjer metoda vrne kot relevanten atribut PBands in sicer pri vsaki delnici.



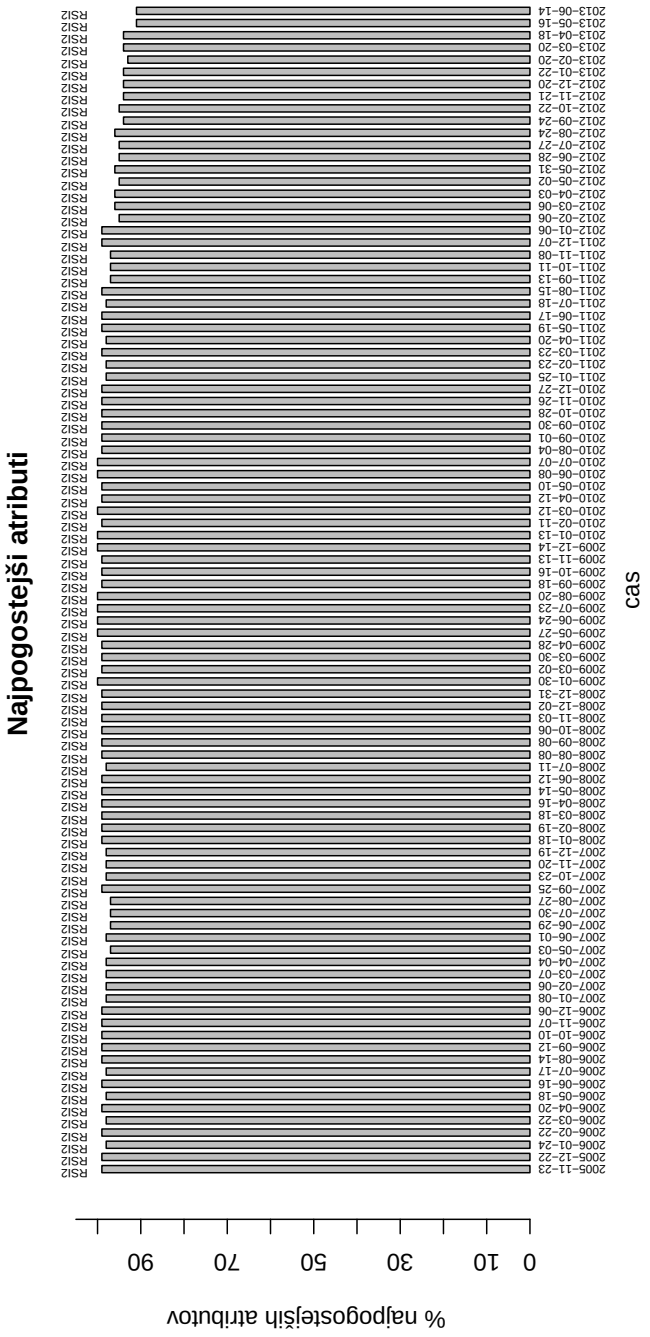
Slika 44: Odstotek najbolj frekventnega atributa, ki se pojavlja kot relevanten atribut na posamezen trgovalni dan za vse delnice. Uporabljena metoda za izbor atributov je FCBF metoda. Najpogostejša atributa sta VHF3 ali VHF2.



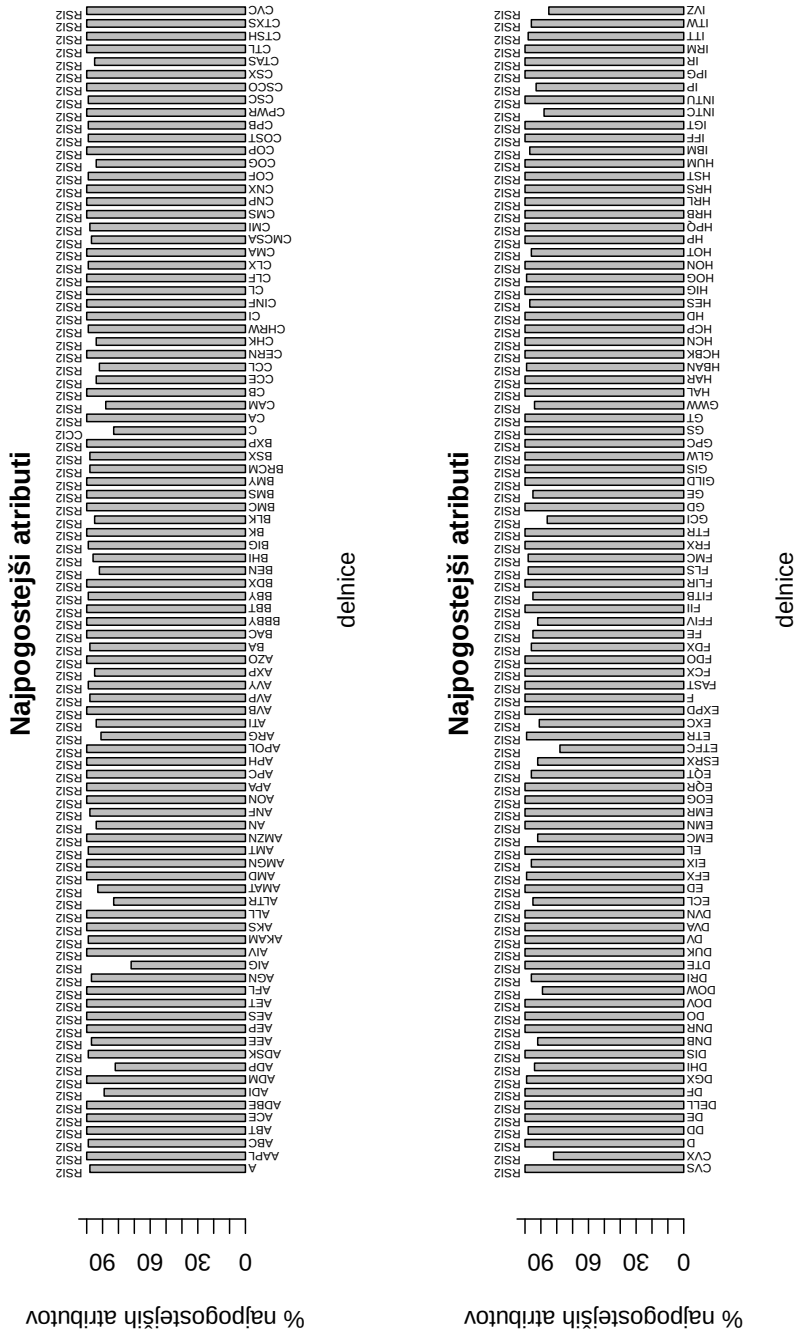
Slika 45: Najpogostejši atributi posamično po delnicah (prva dva seta delnic). Uporabljena metoda za izbor atributov je FCBF metoda. S slike je razvidno, da sta najpogostejša atributa VHF3 ali VHF2.



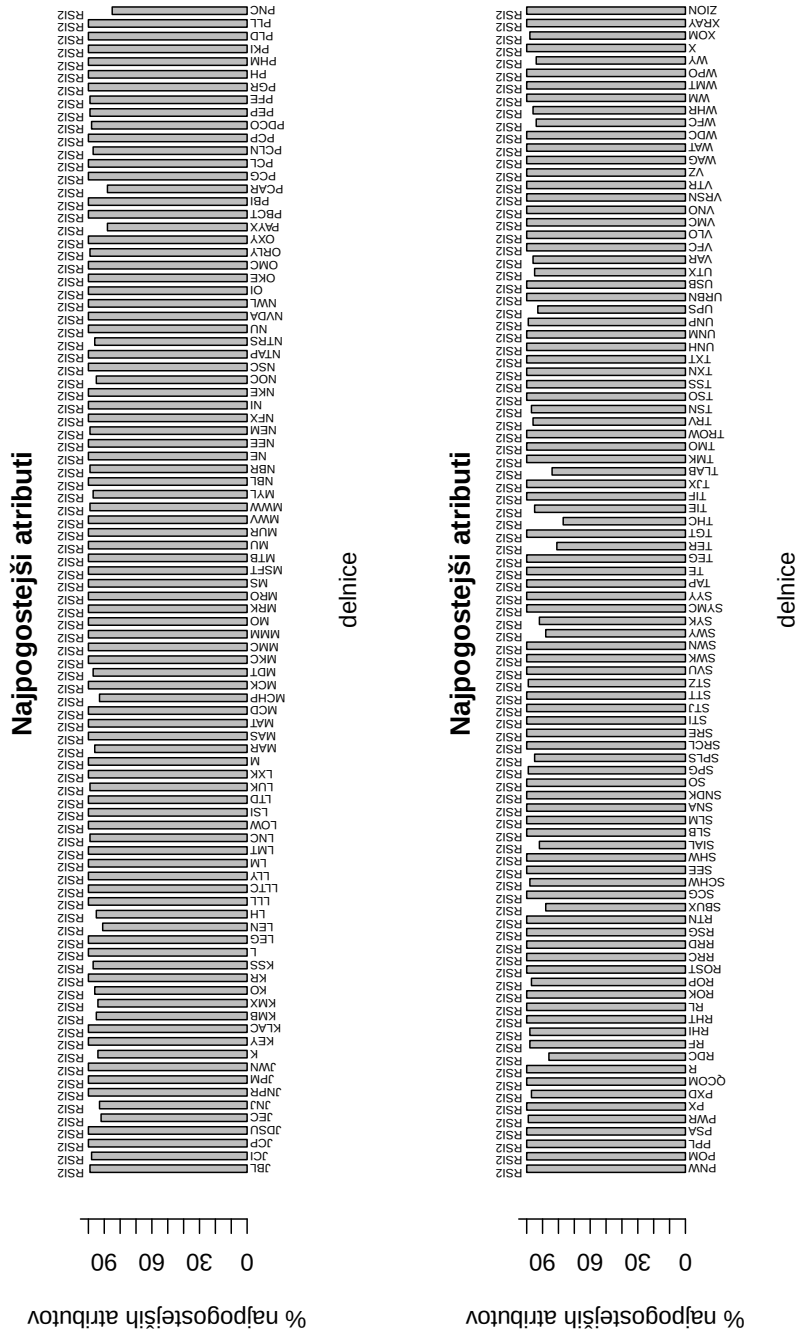
Slika 46: Najpogostejši atributi posamično po delnicah (druga dva seta delnic). Uporabljena metoda za izbor atributov je FCBF metoda. Najpogostejša atributa sta VHF3 ali VHF2.



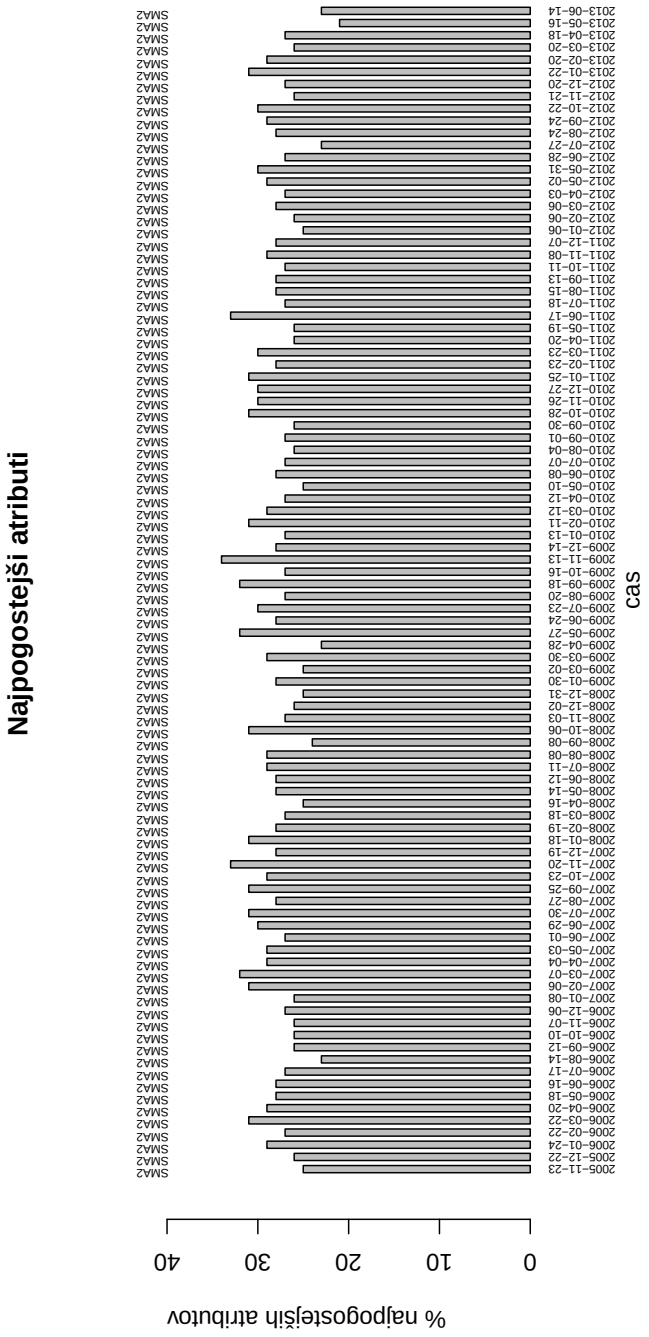
Slika 47: Odstotek najbolj frekventnega atributa, ki se pojavlja kot relevanten atribut na posamezen trgovalni dan za vse delnice. Uporabljena metoda za izbor atributov je CFS metoda. S slike je razvidno, da je najbolj pogosto uporabljen atribut RSI2.



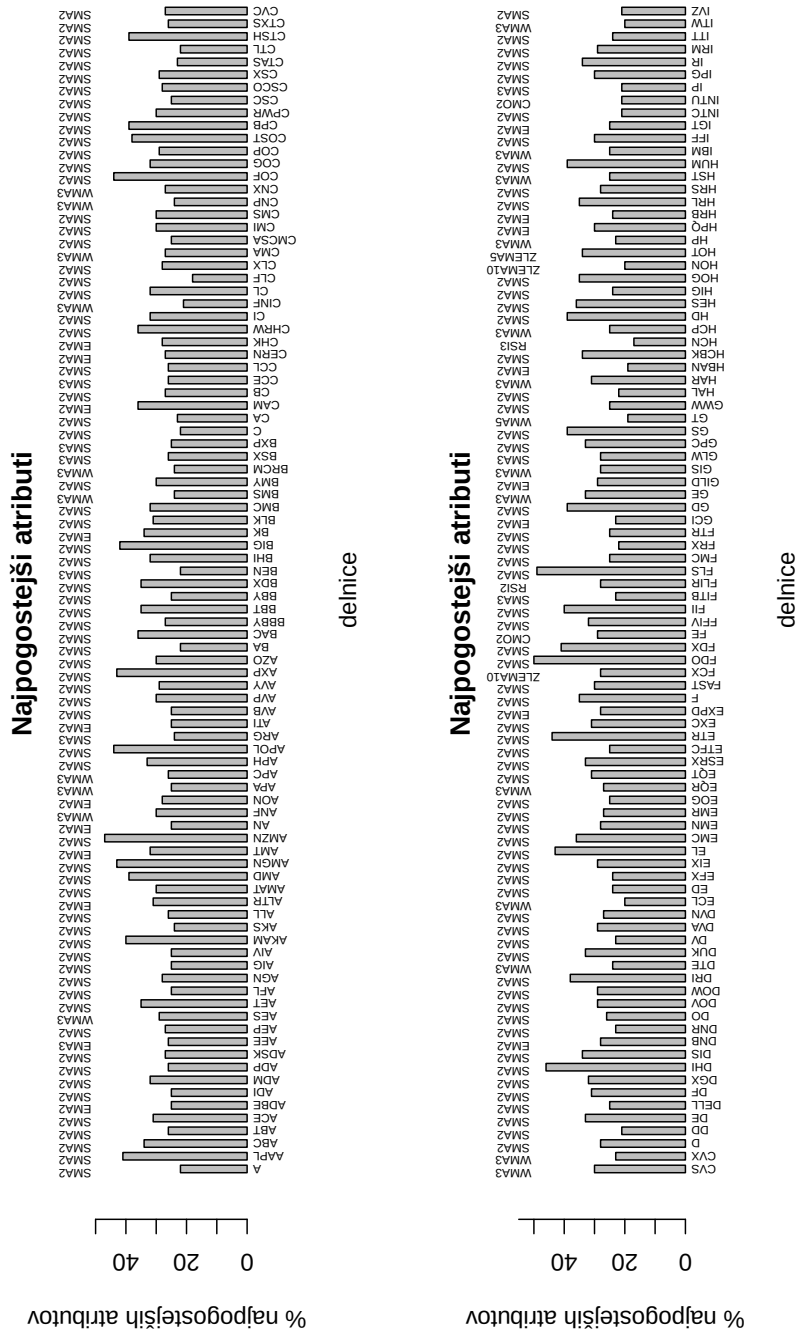
Slika 48: Odstotek najbolj frekventnih atributov posamično po delnicah (prva dva seta delnic). Uporabljena metoda za izbor atributov je CFS metoda.



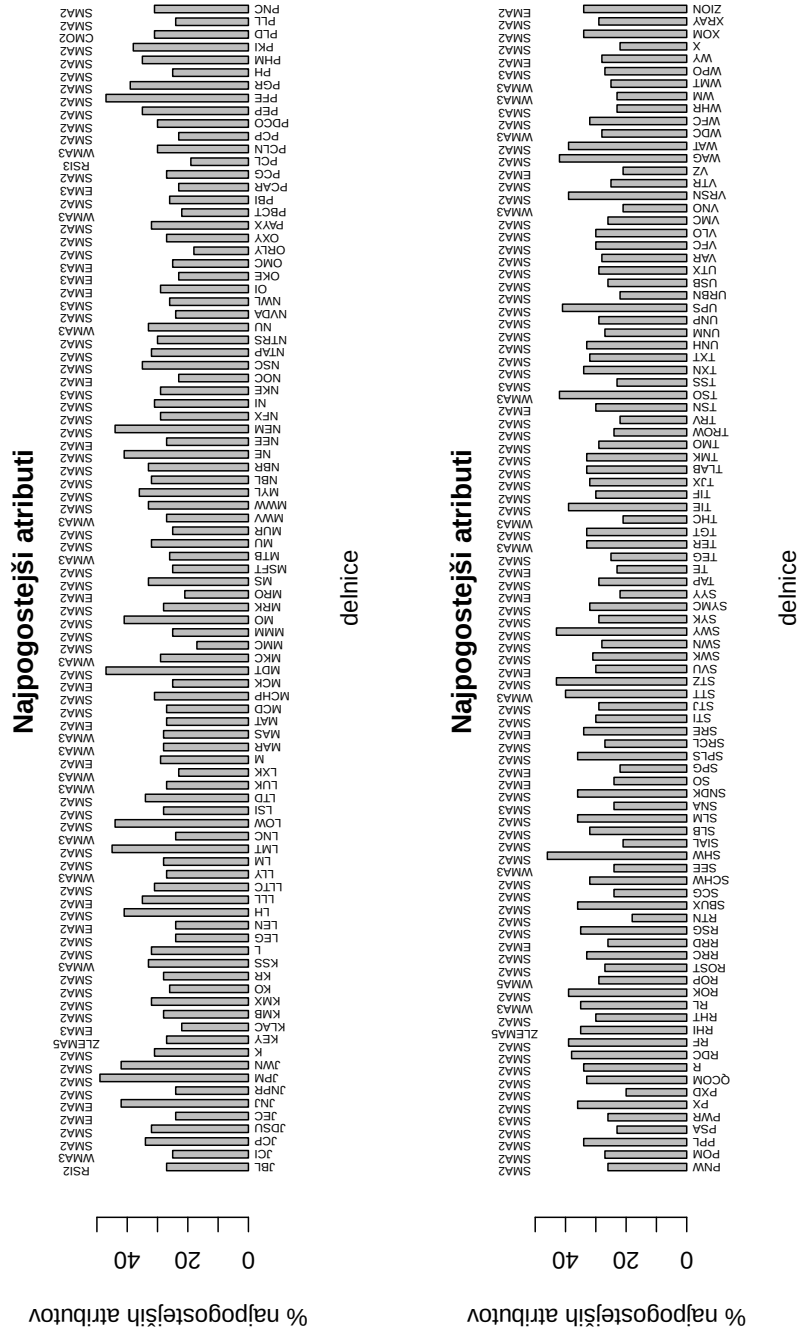
Slika 49: Odstotek najbolj frekventnih atributov posamično po delnicah (druga dva seta delnic). Uporabljena metoda za izbor atributov je CFS metoda.



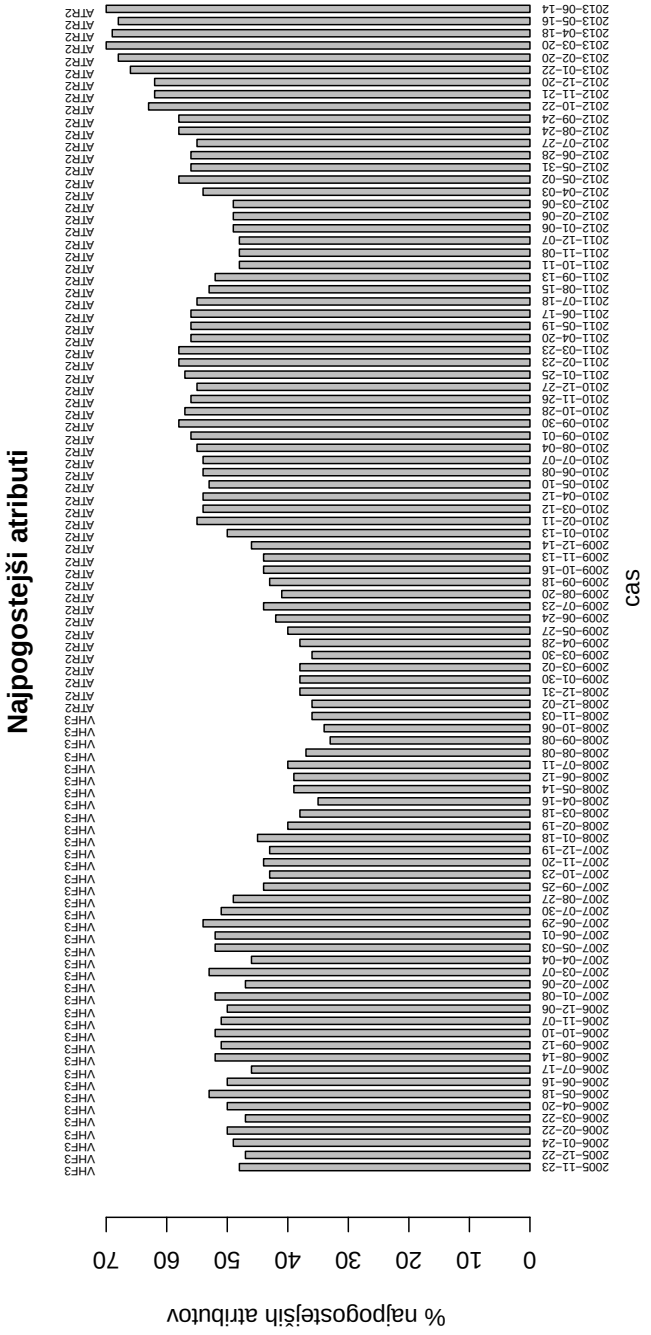
Slika 50: Odstotek najbolj frekventnega atributa, ki se pojavlja kot relevanten atribut na posamezen trgovalni dan za vse delnice. Uporabljena metoda za izbor atributov je FSuC–ward–comb metoda. S slike je razvidno, da je SMA2 najbolj pogosto uporabljen relevanten atribut.



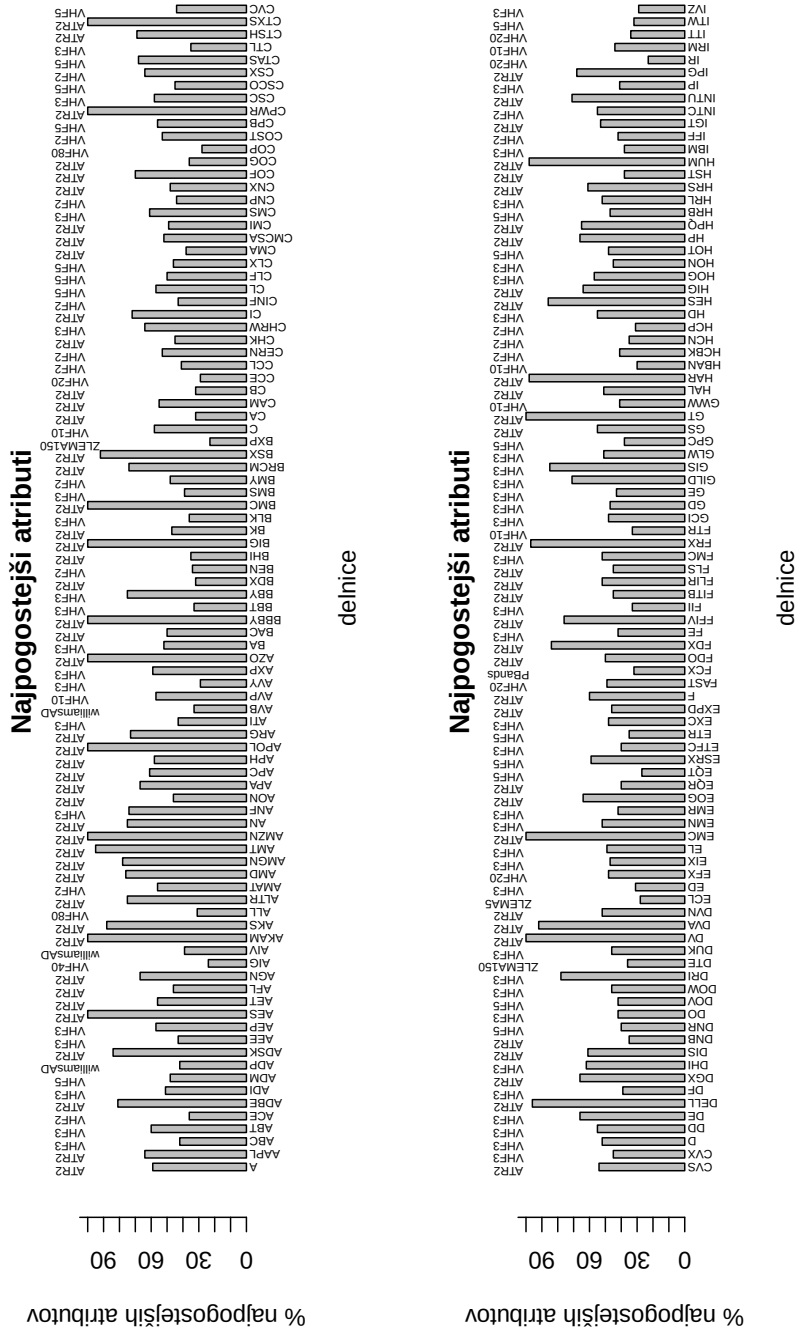
Slika 51: Odstotek najbolj frekventnih atributov posamično po delnicah (prva dva seta delnic). Uporabljena metoda za izbor atributov je FSuC–ward–comb metoda. S slike je razvidno, da sta najpogostejša atributa SMA2 ali SMA3, tudi nekatera druga drseča povprečja, kot npr. EMA, WMA, ZLEMA.



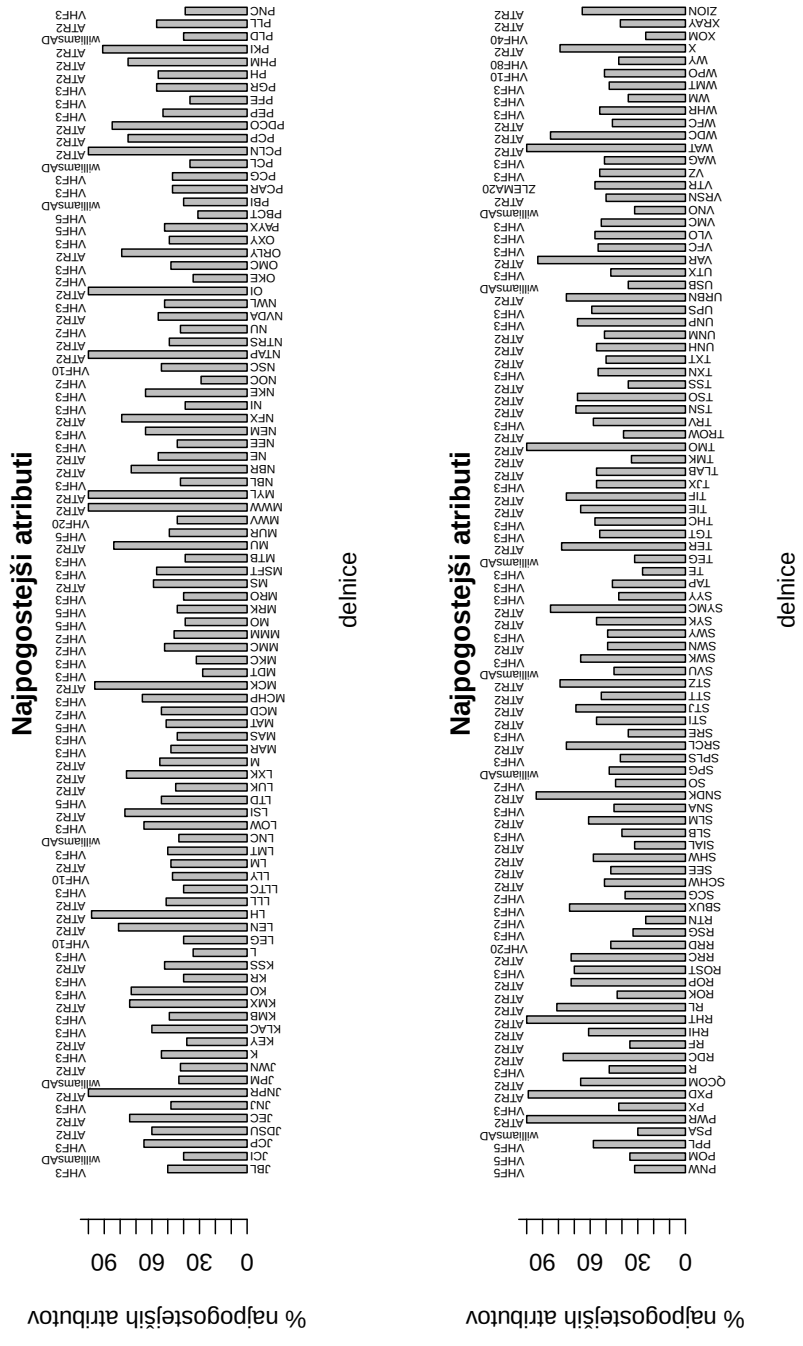
Slika 52: Odstotek najbolj frekventnih atributov posamično po delnicah (druga dva seta delnic). Uporabljena metoda za izbor atributov je FSuC–ward–comb metoda. Najpogostejša atributa sta SMA2 ali SMA3, tudi nekatera druga drseča povprečja, kot npr. EMA, WMA, ZLEMA.



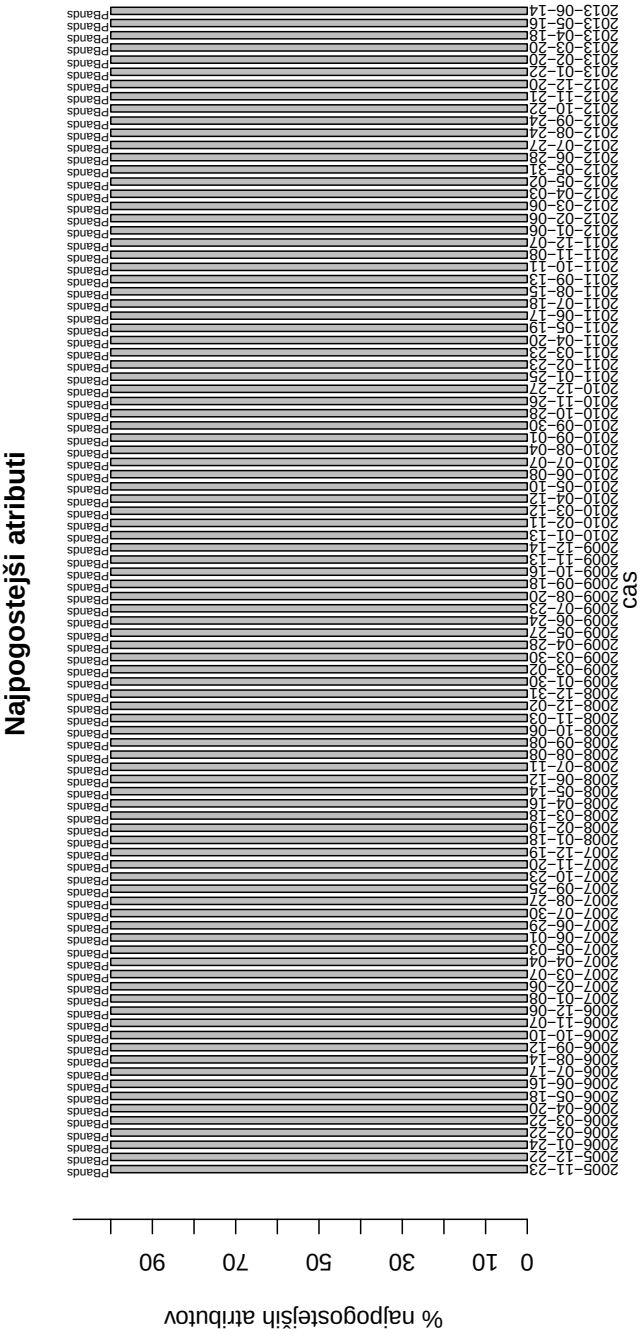
Slika 53: Odstotek najbolj frekventnega atributa, ki se pojavlja kot relevanten atribut na posamezen trgovalni dan za vse delnice. Uporabljena metoda za izbor atributov je mRMR metoda. S slike je razvidno, da sta atributa VHF3 in ATR2 najbolj pogosto uporabljena atributa.



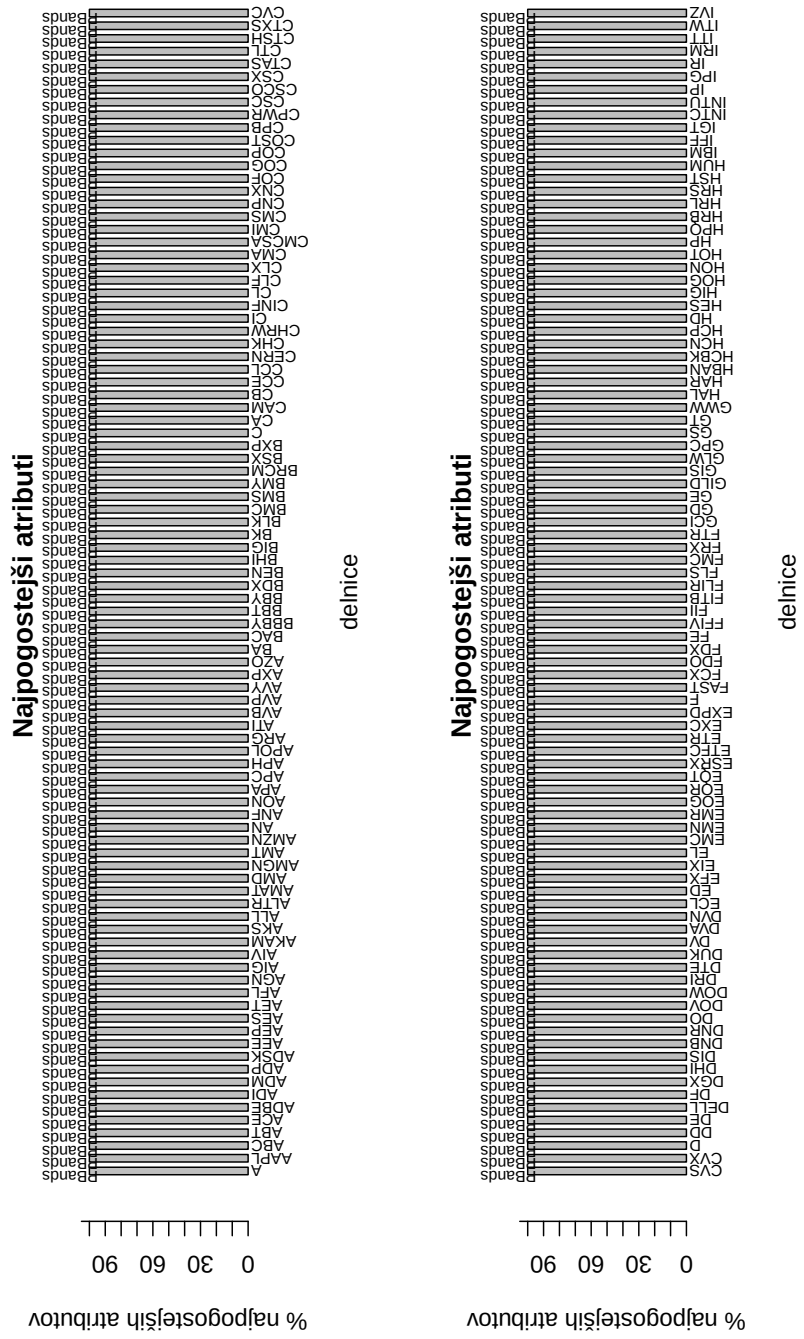
Slika 54: Odstotek najbolj frekventnih atributov posamično po delnicah (druga dva seta delnic). Uporabljena metoda za izbor atributov je mRMR metoda. Najpogostejši atributi so VHF3, VHF2 ali ATR2, ATR3. Tudi drugi tehnični indikatorji se pojavljajo kot npr. williams AD, VHF10.



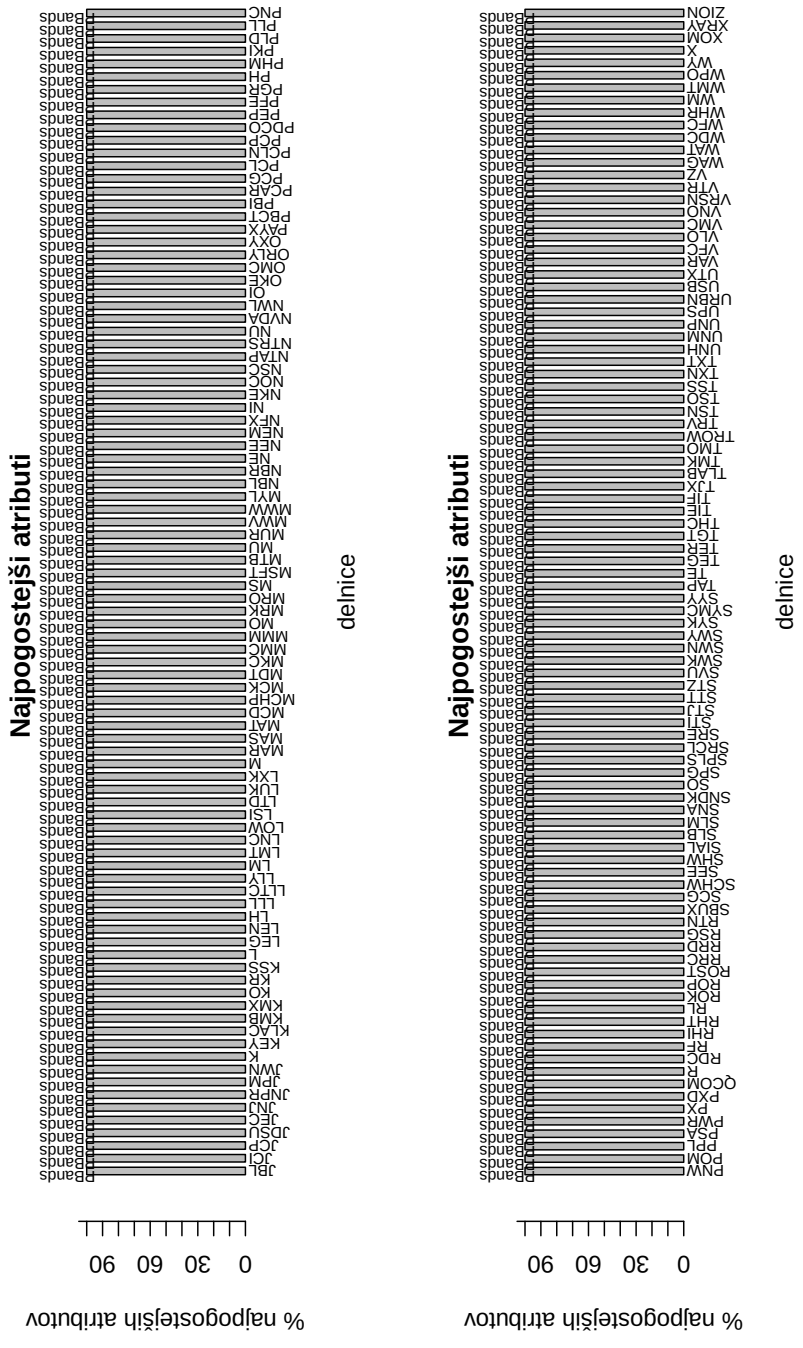
Slika 55: Odstotek najbolj frekventnih atributov posamično po delnicah (druga dva seta delnic). Uporabljena metoda za izbor atributov je mRMR metoda. S slike je razvidno, da so najpogostejši atributi VHF3, VHF2 ali ATR2, ATR3. Tudi drugi tehnični indikatorji se pojavljajo kot npr. williams AD, VHF10.



Slika 56: Odstotek najbolj frekventnega atributa, ki se pojavlja kot relevanten atribut na posamezen trgovalni dan za vse delnice. Uporabljena metoda za izbor atributov je CCCA metoda. S slike je razvidno, da je najpogosteje uporabljen atribut PBands.



Slika 57: Odstotek najbolj frekventnih atributov posamično po delnicah (prva dva seta delnic). Uporabljena metoda za izbor atributov je CCCA metoda. S slike je razvidno, da je najpogostejši atribut PBands.



Slika 58: Odstotek najbolj frekventnih atributov posamično po delnicah (druga dva seta delnic). Uporabljena metoda za izbor atributov je CCCA metoda. S slike je razvidno, da je najpogostejši atribut PBands.

8.3 Rezultati Vodenih D–trgovalnih strategij

V tem poglavju smo prikazali rezultate trgovalnih strategij pri uporabi različnih modelov. Za vhodne podatke pri grajenju klasifikacijskih modelov uporabimo relevantne atribute, ki smo jih dobili z nekaterimi metodami za izbor atributov. Spodnji rezultati, predstavljeni v tabelah 24, 25, 26, 27, 28, in rezultati prikazani na slikah 59, 60, 61, 62, 63, kažejo, da so najboljši rezultati pri izbranih kazalnikih pri FSuC–ward–comb izboru relevantnih atributov, kar pa ni presenetljivo glede na klasifikacijske rezultate (glej poglavje 6). V vsak graf smo dodali tudi Naivno strategijo, ki je dosegla najvišji CAGR. V nekaterih grafih lahko vidimo, da so Naivne strategije precej blizu Vodenim D–trgovalnim strategijam, vendar jih pa nikoli ne presežejo. V tabeli 18 je prikazan izbor D vrednosti po različnih metodah za izbor atributov. Pri vseh metodah je največkrat uporabljen prag pri $D = 2.5\%$.

Tabela 18: Frekvence D vrednosti pragov po različnih metodah.

Tabela 19: FSuC metoda.

D	1%	2%	2.5%	3%
D_{LDA}	12	23	52	9
D_{NB}	13	13	62	8
D_{RBF}	13	18	53	12
D_{lin}	5	25	46	20

Tabela 20: FCBF metoda.

D	1%	2%	2.5%	3%
D_{LDA}	15	24	45	12
D_{NB}	10	22	51	13
D_{RBF}	4	14	55	23
D_{lin}	13	19	56	8

Tabela 21: CFS metoda.

D	1%	2%	2.5%	3%
D_{LDA}	5	21	63	7
D_{NB}	3	17	61	15
D_{RBF}	5	15	73	3
D_{lin}	1	25	57	13

Tabela 22: mRMR metoda.

D	1%	2%	2.5%	3%
D_{LDA}	3	14	46	33
D_{NB}	10	15	35	36
D_{RBF}	1	21	36	38
D_{lin}	7	17	40	32

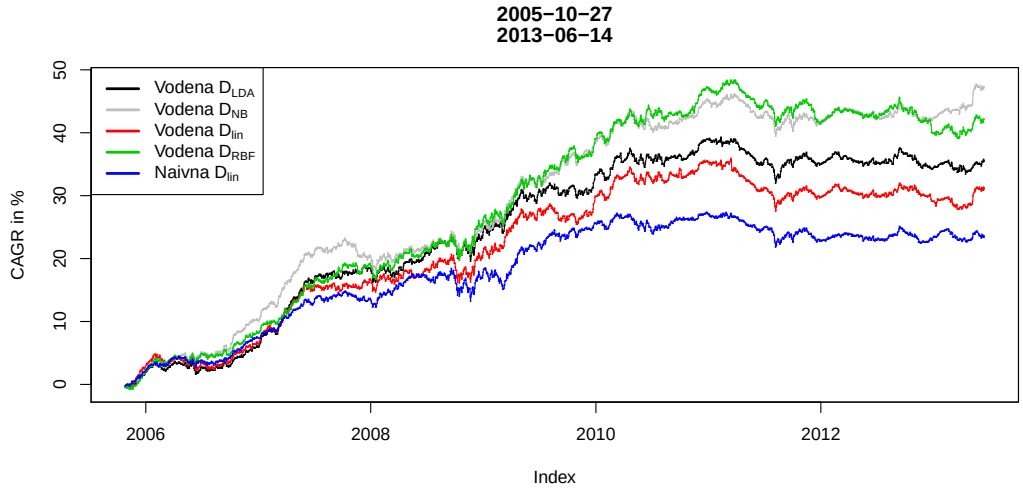
Tabela 23: CCCA metoda.

D	1%	2%	2.5%	3%
D_{LDA}	5	15	48	28
D_{NB}	11	14	51	20
D_{RBF}	4	11	52	29
D_{lin}	5	18	49	24

8.3.1 FSuC–ward–comb

	povp(%)	**std(%)**	**Sharpe**	**Sortino**	**info koef**	**CAGR(%)**
Vodena D_{LDA}	0.13	0.04	3.55	0.12	3.33	35.41
Vodena D_{NB}	**0.17**	0.04	**4.56**	**0.15**	**4.46**	**47.31**
Vodena D_{lin}	0.12	0.04	3.21	0.10	2.90	31.07
Vodena D_{RBF}	0.15	0.04	4.01	0.13	3.92	42.16
Naivna D_{LDA}	0.09	0.03	2.76	0.09	2.54	22.68
Naivna D_{NB}	0.10	0.03	2.90	0.09	2.75	24.69
Naivna D_{lin}	0.11	0.03	3.23	0.10	3.17	27.60
Naivna D_{RBF}	0.09	0.03	2.83	0.09	2.64	23.39

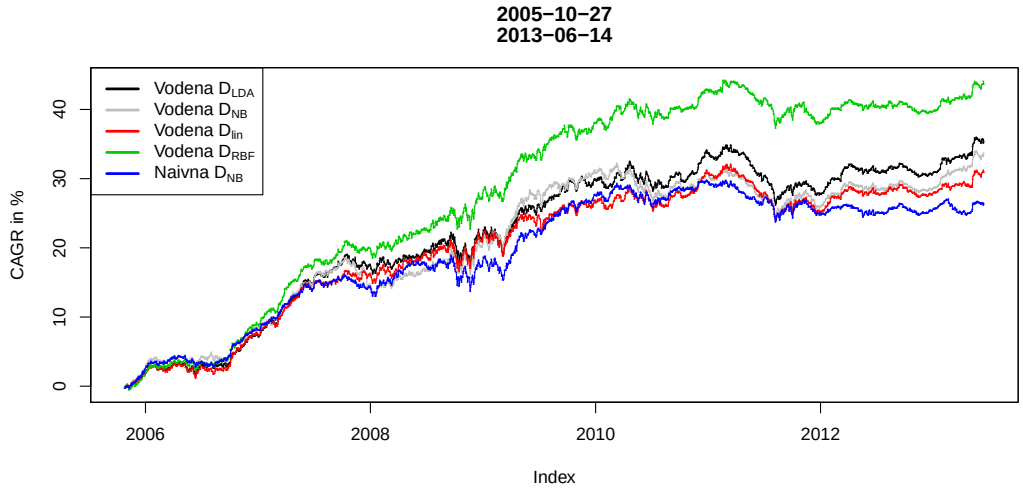
Tabela 24: Rezultati izvedbe Vodenih D–trgovalnih strategij in primerjava z Naivnimi strategijami, indeksom $S\&P500$ ter Primerjalno strategijo. Metoda za izbor atributov je FSuC–ward–comb.



Slika 59: Grafični prikaz Vodenih D–trgovalnih strategij, kjer smo pri grajenju klasifikacijskih modelov uporabili metodo za izbor atributov FSuC–ward–comb.

8.3.2 FCBF

Najvišje rezultate med strategijami doseže Vodena D_{RBF}–trgovalna strategija, kar pomeni uporabo SVM klasifikatorja z RBF jedrom. Ta je za kar slabih 10 odstotnih točk nad Vodeno D_{LDA}–trgovalno strategijo. Vse predlagane Vodene D– trgovalne strategije vrnejo boljše rezultate kot pa Naivne strategije, indeks $S\&P500$ in Primerjalna strategija (glej sliko 60).



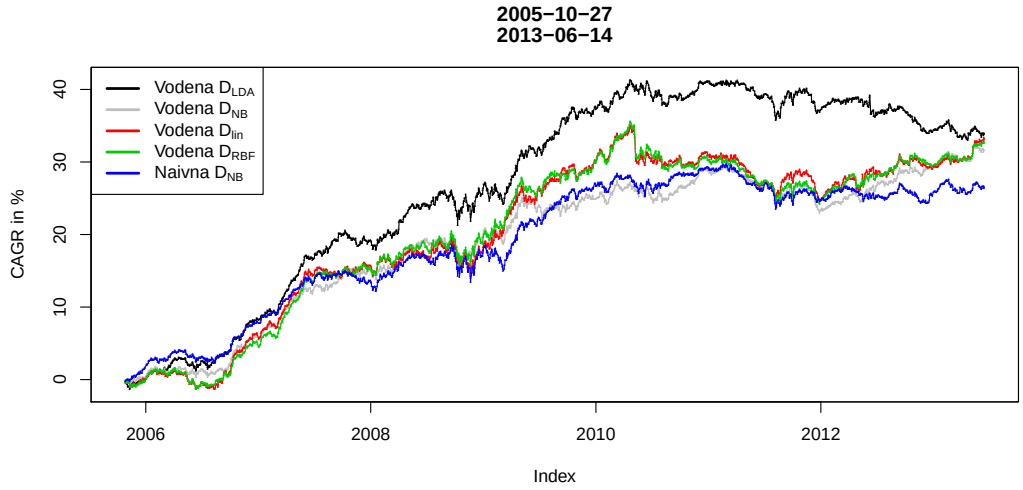
Slika 60: Grafični prikaz Vodenih D–trgovalnih strategij, kjer smo pri grajenju klasifikacijskih modelov uporabili metodo za izbor atributov FCBF.

	povp(%)	**std(%)**	**Sharpe**	**Sortino**	**info koef**	**CAGR(%)**
Vodena D_{LDA}	0.13	0.04	3.67	0.12	3.52	35.31
Vodena D_{NB}	0.13	0.04	3.50	0.11	3.36	33.51
Vodena D_{lin}	0.12	0.04	3.35	0.11	3.17	30.93
Vodena D_{RBF}	0.16	0.04	4.43	0.15	4.49	43.71
Naivna D_{LDA}	0.10	0.03	2.87	0.09	2.74	23.93
Naivna D_{NB}	0.10	0.03	3.11	0.10	3.02	26.36
Naivna D_{lin}	0.10	0.03	2.91	0.09	2.77	24.36
Naivna D_{RBF}	0.10	0.03	2.76	0.09	2.59	23.67

Tabela 25: Rezultati izvedbe Vodenih D–trgovalnih strategij in primerjava z indeksom $S\&P500$ ter Primerjalno strategijo. Metoda za izbor atributov je FCBF.

8.3.3 CFS

Najvišje rezultate med strategijami doseže Vodena D_{LDA}–trgovalna strategija. Vse predlagane Vodene D– trgovalne strategije vrnejo boljše rezultate kot pa Naivne strategije, indeks $S\&P500$ in Primerjalna strategija (glej sliko 61).



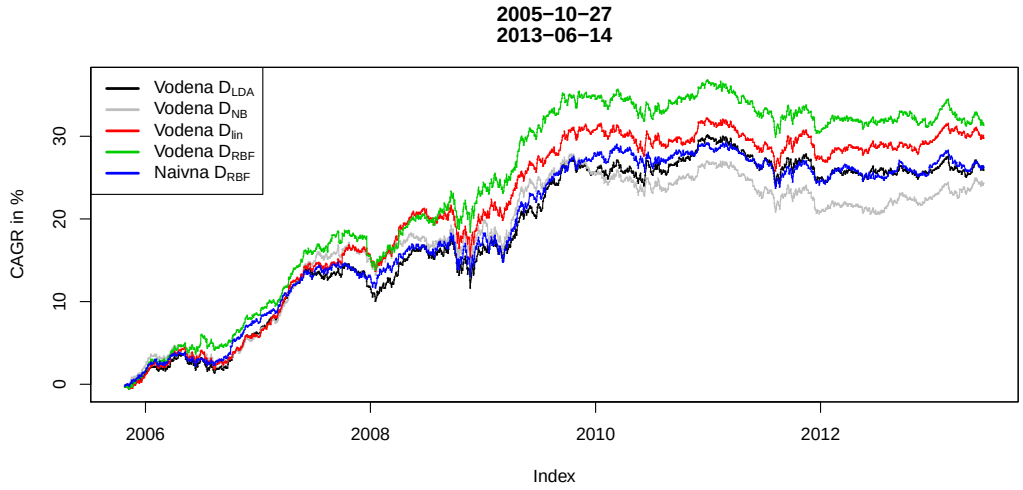
Slika 61: Grafični prikaz Vodenih D–trgovalnih strategij, kjer smo pri grajenju klasifikacijskih modelov uporabili metodo za izbor atributov CFS.

	povp(%)	**std**(%)	**Sharpe**	**Sortino**	**info koef**	**CAGR**(%)
Vodena D_{LDA}	0.13	0.04	3.56	0.12	3.40	33.88
Vodena D_{NB}	0.12	0.04	3.38	0.11	3.24	31.63
Vodena D_{lin}	0.13	0.04	3.26	0.10	2.86	33.07
Vodena D_{RBF}	0.13	0.04	3.32	0.11	2.97	32.60
Naivna D_{LDA}	0.10	0.03	2.90	0.09	2.77	24.51
Naivna D_{NB}	0.10	0.03	3.06	0.10	2.97	26.42
Naivna D_{lin}	0.09	0.03	2.68	0.08	2.47	22.44
Naivna D_{RBF}	0.09	0.04	2.51	0.08	2.25	20.97

Tabela 26: Rezultati izvedbe Vodenih D–trgovalnih strategij. Metoda za izbor atributov je CFS.

8.3.4 mRMR

Med Vodenimi D–trgovalnimi strategijami doseže najvišje rezultate Vodena D_{RBF}–trgovalna strategija. Vse predlagane Vodene D– trgovalne strategije vrnejo boljše rezultate kot pa Naivne strategije, indeks $S\&P500$ in Primerjalna strategija (glej sliko 62).



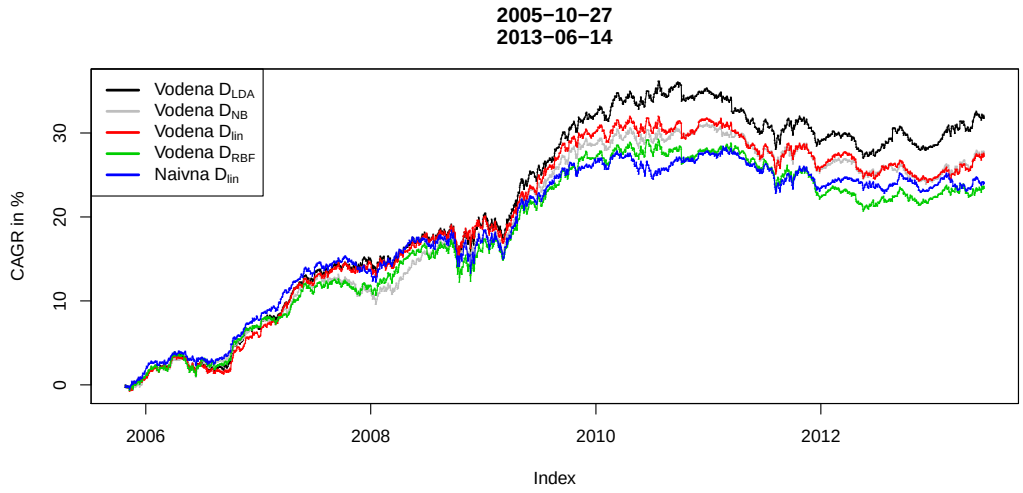
Slika 62: Grafični prikaz Vodenih D–trgovalnih strategij, kjer smo pri grajenju klasifikacijskih modelov uporabili metodo za izbor atributov mRMR.

	povp(%)	**std(%)**	**Sharpe**	**Sortino**	**info koef**	**CAGR(%)**
Vodena D_{LDA}	0.11	0.04	2.84	0.09	2.59	25.92
Vodena D_{NB}	0.10	0.04	2.67	0.09	2.39	24.32
Vodena D_{lin}	0.12	0.04	3.18	0.10	2.98	29.78
Vodena D_{RBF}	0.12	0.04	3.22	0.10	3.00	31.31
Naivna D_{LDA}	0.09	0.04	2.54	0.08	2.30	21.67
Naivna D_{NB}	0.10	0.03	2.79	0.09	2.61	23.93
Naivna D_{lin}	0.09	0.04	2.62	0.08	2.40	22.36
Naivna D_{RBF}	0.10	0.04	2.96	0.10	2.89	26.21

Tabela 27: Rezultati izvedbe Vodenih D–trgovalnih strategij in primerjava z indeksom $S\&P500$ ter Primerjalno strategijo. Metoda za izbor atributov je mRMR.

8.3.5 CCCA

S slike 63 lahko vidimo, da so najvišji rezultati pri Vodeni D_{LDA}–trgovalni strategiji. Pri metodi CCCA rezultati kažejo, da Naivna D_{lin}–strategija preseže Vodeno D_{RBF}–strategijo.



Slika 63: Grafični prikaz Vodenih D–trgovalnih strategij, kjer smo pri grajenju klasifikacijskih modelov uporabili metodo za izbor atributov CCCA.

	povp(%)	**std(%)**	**Sharpe**	**Sortino**	**info koef**	**CAGR(%)**
Vodena D_{LDA}	0.12	0.04	3.27	0.11	3.08	31.76
Vodena D_{NB}	0.11	0.04	3.01	0.10	2.79	27.55
Vodena D_{lin}	0.11	0.04	2.98	0.10	2.72	27.25
Vodena D_{RBF}	0.10	0.04	2.64	0.08	2.29	23.40
Naivna D_{LDA}	0.10	0.03	2.76	0.09	2.59	23.74
Naivna D_{NB}	0.09	0.03	2.69	0.09	2.49	22.60
Naivna D_{lin}	0.10	0.03	2.81	0.09	2.68	23.97
Naivna D_{RBF}	0.09	0.04	2.50	0.08	2.25	21.45

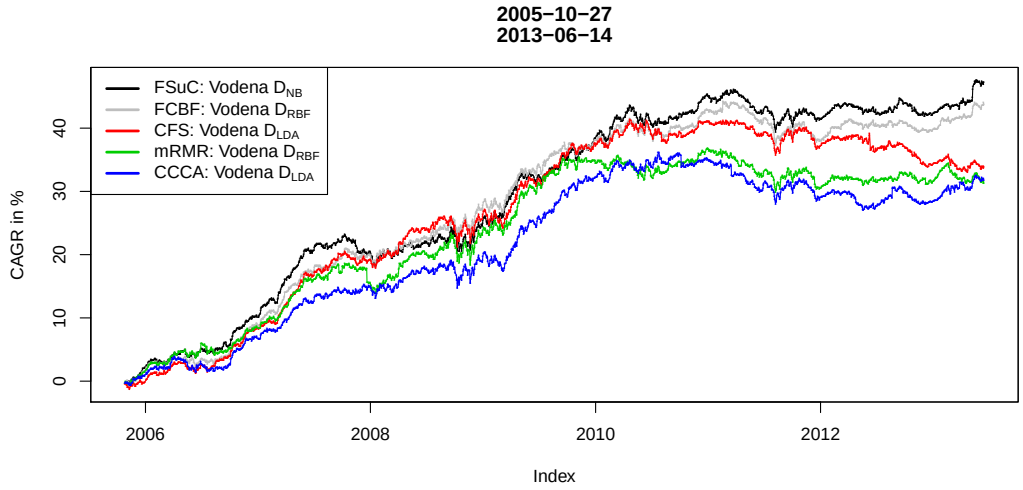
Tabela 28: Rezultati izvedbe Vodenih D–trgovalnih strategij in primerjava z indeksom $S\&P500$ ter Primerjalno strategijo. Metoda za izbor atributov je CCCA.

8.4 Primerjava izvedbe Vodenih D–trgovalnih strategij po metodah

V tem poglavju smo zbrali najboljše izvedbe trgovalnih strategij po metodah za izbor atributov in jih med seboj primerjali.

metoda	**ime strategije**	**povp**(%)	**std**(%)	**Sharpe**	**Sortino**	**info koef**	**CAGR**(%)
FSuC–ward–comb	Vodena D_{NB}	**0.17**	0.04	**4.56**	**0.15**	**4.46**	**47.31**
FCBF	Vodena D_{RBF}	0.16	0.04	4.43	0.15	4.49	43.71
CFS	Vodena D_{LDA}	0.13	0.04	3.56	0.12	3.40	33.88
mRMR	Vodena D_{RBF}	0.12	0.04	3.22	0.10	3.00	31.31
CCCA	Vodena D_{LDA}	0.12	0.04	3.27	0.11	3.08	31.76

Tabela 29: Primerjava rezultatov izvedbe Vodenih D–trgovalnih strategij po metodah.



Slika 64: Grafični prikaz najboljših Vodenih D–trgovalnih strategij po CAGR izvedbi po različnih metodah.

Iz slike 64 lahko vidimo, da so si Vodene D–trgovalne strategije precej podobne med seboj. Po opravljenem Wilcoxonovem testu s predznačenimi rangi pa lahko zaključimo, da se Vodena D_{NB}–strategija pri FSuC–comb–ward metodi statistično razlikuje le od CFS Vodene D_{NB}–strategije in CCCA Vodene D_{NB}–strategije (glej tabelo 30). Zanimivo je, da kljub temu, da ima Vodena D_{RBF}–trgovalna strategija pri metodi mRMR večji odklon od Vodene D_{LDA}–trgovalne strategije pri metodi CFS, pa ne moremo govoriti o statistično značilnih razlikah (glede na izveden test).

	FCBF	CFS	mRMR	CCCA
FSuC: Vodena D_{NB}	0.4882	0.0260*	0.2203	0.0385*
FCBF: Vodena D_{RBF}		0.0338*	0.2612	0.0469*
CFS: Vodena D_{LDA}			0.5651	0.6749
mRMR: Vodena D_{RBF}				0.3989
CCCA: Vodena D_{LDA}				

Tabela 30: Prikazane p-vrednosti pri Wilcoxonovem testu s predznačenimi rangi po različnih metodah.

8.5 Wilcoxonovi testi s predznačenimi rangi

Naivne D–strategije se med seboj ločijo glede na različno postavljene vrednosti D pragov, ki jih predhodno dobimo s pomočjo izvedbe Vodenih D–trgovalnih strategij na učni množici. Ker so v Vodenih D–trgovalnih strategijah vključeni tudi klasifikacijski modeli in tako različni vhodni atributi, dobimo za vsako metodo za izbor atributov in vsak klasifikacijski model različne vrednosti D pragov (glej tabelo 18, poglavje 8.3) in s tem različna obnašanja Naivnih–D strategij. Nekatere Naivne strategije so precej blizu Vodenim D–trgovalnim strategijam, kar lahko pomeni dobro izbiro D–pragov in pa slab klasifikacijski model. V poglavju 7.3 smo že predstavili izide Wilcoxonovih testov s predznačenimi rangi za metodo FSuC–ward–comb. Poglejmo si, kako se z Wilcoxonovimi testi s predznačenimi rangi trgovalne strategije razlikujejo med seboj. V tabelah 31, 32, 33 in 34, so izvedeni Wilcoxonovi testi s predznačenimi rangi pri različnih metodah za izbor atributov (FCBF, CFS, mRMR in CCCA). Vhodna podatka za izvedbo testa sta vektorja dnevnih donosov, ki jih vrnejo predlagane trgovalne strategije ali pa indeks na celotnem testnem obdobju. Prikazane p-vrednosti med 0.01 in 0.05 smo označili z eno zvezdico $*$ ter p-vrednosti manjše kot 0.01 smo označili z dvema zvezdicama $**$.

8.5.1 FCBF

V tabeli 31 je razvidno, da se uspešnost Vodenih trgovalnih strategij med seboj signifikantno ne razlikuje na danem vzorcu dnevnih donosov, tako da ne moremo zaključiti, katera izvedba trgovalnih strategij je boljša. Večina Vodenih D–trgovalnih strategij se signifikantno razlikuje od Naivne D_{NB} in Naivne D_0–trgovalnih strategij (izjemi sta Vodena D_{NB} in Vodena D_{lin}–trgovalni strategiji, zanju ne moremo zaključiti, da se signifikantno razlikujeta od Naivne D_{NB}–strategije). Vse Vodene D–trgovalne strategije se statistično razlikujejo tudi od indeksa in Primerjalne strategije (glej sliko 60 in tabelo 25 v poglavju 8.3.2.)

	Naivna D_{NB}	Vodena D_{LDA}	Vodena D_{NB}	Vodena D_{RBF}	Vodena D_{lin}	SPY	bench
Naivna D_0	0.0000^{**}	0.0000^{**}	0.0000^{**}	0.0000^{**}	0.0000^{**}	0.6912	0.6753
Naivna D_{NB}		0.0149^{*}	0.0514	0.0014^{*}	0.0992	0.0001^{**}	0.0000^{**}
Vodena D_{LDA}			0.4902	0.5167	0.1123	0.0000^{**}	0.0000^{**}
Vodena D_{NB}				0.1434	0.5920	0.0000^{**}	0.0000^{**}
Vodena D_{RBF}					0.0092^{**}	0.0000^{**}	0.0000^{**}
Vodena D_{lin}						0.0000^{**}	0.0000^{**}
SPY							0.3131

Tabela 31: Prikazane p-vrednosti pri Wilcoxonovih testih s predznačenimi rangi, kjer je FCBF izbrana metoda za izbor relevantnih atributov.

8.5.2 CFS

Iz tabele 32 lahko sklenemo, da se uspešnost Vodenih trgovalnih strategij signifikantno razlikuje od Naivne D_0–strategije. Na podlagi dobljenih p–vrednosti pa ne moremo pa zaključiti, ali se Vodene D–trgovalne strategije signifikantno razlikujejo od Naivne D_{NB}–strategije z izjemo VodeneD_{LDA}–trgovalne strategije, ki pa se tudi signifikantno razlikuje od preostalih Vodenih D–trgovalnih strategij, kot je razvidno s slike 61 v poglavju 8.3.3, ki po CAGR vrednosti presega ostale. Za ostale Vodene D–trgovalne strategije ne moremo reči, ali se signifikantno razlikujejo med seboj.

	Naivna D_{NB}	Vodena D_{LDA}	Vodena D_{NB}	Vodena D_{RBF}	Vodena D_{lin}	SPY	bench
Naivna D_0	0.0000^{**}	0.0000^{**}	0.0000^{**}	0.0000^{**}	0.0000^{**}	0.6912	0.6753
Naivna D_{NB}		0.002^{**}	0.2396	0.1025	0.0539	0.0001^{**}	0.0000^{**}
Vodena D_{LDA}			0.0002^{**}	0.0002^{**}	0.0004^{**}	0.3397	0.3384
Vodena D_{NB}				0.4204	0.2205	0.0000^{**}	0.0000^{**}
Vodena D_{RBF}					0.8112	0.0000^{**}	0.0000^{**}
Vodena D_{lin}						0.0000^{**}	0.0000^{**}
SPY							0.3131

Tabela 32: Prikazane p-vrednosti pri Wilcoxonovih testih s predznačenimi rangi, kjer je CFS izbrana metoda za izbor relevantnih atributov.

8.5.3 mRMR

Iz prikazanih p–vrednosti v tabeli 33 je razvidno, da se uspešnost Vodenih trgovalnih strategij na danem vzorcu dnevnih donosov med seboj signifikantno ne razlikuje, tako da ne moremo zaključiti, katera izvedba trgovalnih strategij je boljša. Večina Vodenih D–trgovalnih strategij se signifikantno razlikuje od Naivne D_0–trgovalne strategije, vendar pa se te ne razlikujejo od Naivne D_{RBF}–trgovalne strategije, ki jim je po uspešnosti kvantitativnih kazalcev zelo blizu (oz. celo presega CAGR vrednosti nekaterih Vodenih D–trgovalnih strategij, glej sliko 62 v poglavju 8.3.4). Vse Vodene D–trgovalne strategije se statistično razlikujejo od indeksa in Primerjalne strategije (glej sliko 62 in tabelo 27.)

	Naivna D_{NB}	Vodena D_{LDA}	Vodena D_{RBF}	Vodena D_{RBF}	Vodena D_{lin}	SPY	bench
Naivna D_0	0.0000^{**}	0.0000^{**}	0.0000^{**}	0.0000^{**}	0.0000^{**}	0.6912	0.6753
Naivna D_{RBF}		0.4925	0.2975	0.1835	0.2723	0.0001^{**}	0.0000^{**}
Vodena D_{LDA}			0.5885	0.2630	0.3460	0.0001^{**}	0.0000^{**}
Vodena D_{NB}				0.2597	0.7498	0.0002^{**}	0.0001^{**}
Vodena D_{RBF}					0.6277	0.0000^{**}	0.0000^{**}
Vodena D_{lin}						0.0000^{**}	0.0000^{**}
SPY							0.3131

Tabela 33: Prikazane p-vrednosti pri Wilcoxonovih testi s predznačenimi rangi, kjer je mRMR izbrana metoda za izbor relevantnih atributov.

8.5.4 CCCA

V tabeli 34 lahko vidimo, da se uspešnost Vodenih trgovalnih strategij na danem vzorcu dnevnih donosov med seboj signifikantno ne razlikuje, tako da ne moremo zaključiti, katera izvedba trgovalnih strategij je boljša. Večina Vodenih D–trgovalnih strategij se signifikantno razlikuje od Naivne D_0–trgovalne strategije. Po prikazanih p–vrednostih se Vodene D–trgovalne strategije in Naivna D_{lin}–strategija ne razlikujejo signifikantno (glej sliko 63 v poglavju 8.3.5), se pa vse Vodene D–trgovalne strategije statistično razlikujejo od indeksa in Primerjalne strategije (glej sliko 63 in tabelo 28).

	Naivna D_{lin}	Vodena D_{LDA}	Vodena D_{RBF}	Vodena D_{RBF}	Vodena D_{lin}	SPY	bench
Naivna D_0	0.0000^{**}	0.0000^{**}	0.0000^{**}	0.0000^{**}	0.0000^{**}	0.6912	0.6753
Naivna D_{lin}		0.1565	0.3144	0.9233	0.7449	0.0002^{**}	0.0000^{**}
Vodena D_{LDA}			0.6544^{**}	0.5175	0.829	0.0000^{**}	0.0000^{**}
Vodena D_{NB}				0.3341	0.4331	0.0000^{**}	0.0000^{**}
Vodena D_{RBF}					0.4872	0.0001^{**}	0.0001^{**}
Vodena D_{lin}						0.0001^{**}	0.0001^{**}
SPY							0.3131

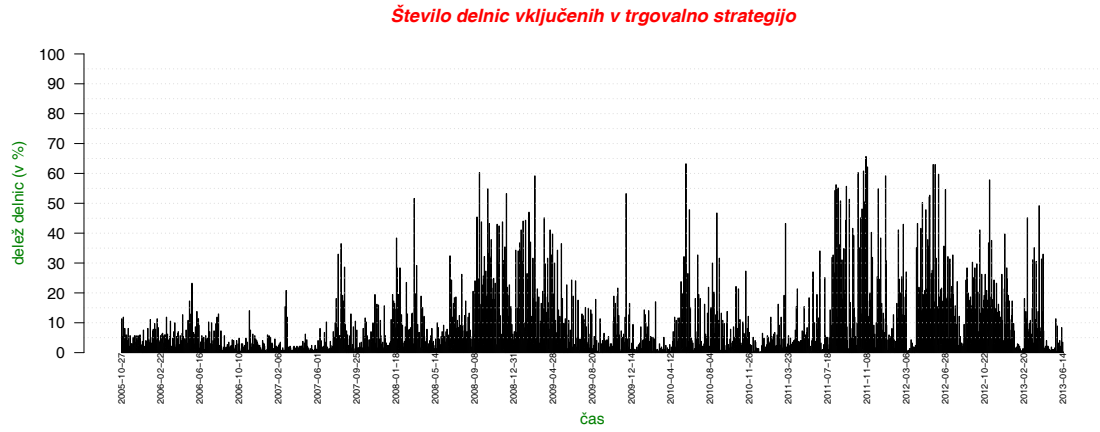
Tabela 34: Prikazane p-vrednosti pri Wilcoxonovih testih s predznačenimi rangi, kjer je CCCA izbrana metoda za izbor relevantnih atributov.

Na podlagi rezultatov Wilcoxonovih testov s predznačenimi rangi lahko sklepamo, da ima izbira D pragov pomembno funkcijo pri trgovalnih strategijah. Naivne D–trgovalne strategije so pri nekaterih izborih D alternativa Vodenim D–trgovalnim strategijam.

8.6 Število vključenih delnic v trgovalne strategije

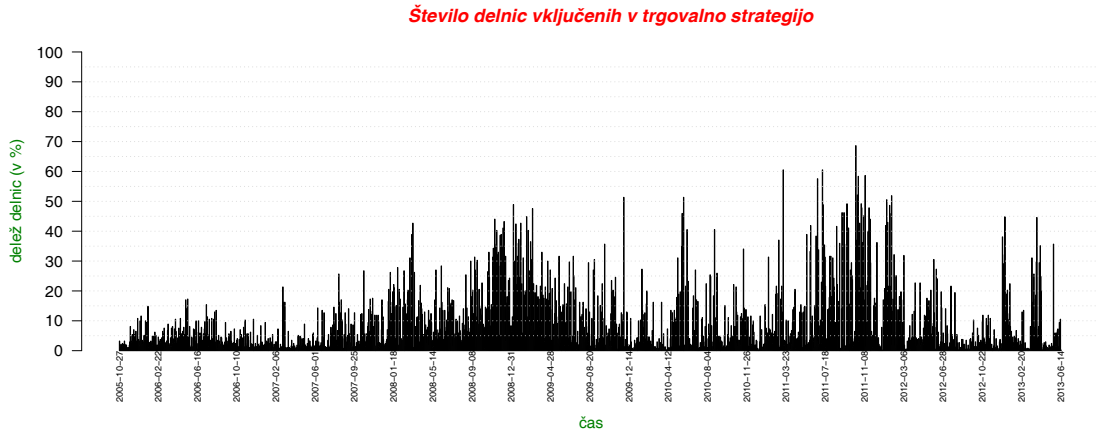
V tem poglavju smo grafično prikazali število izbranih delnic pri Vodenih D_{LDA}–trgovalnih strategijah (grafični prikaz števila vključenih delnic v odvisnosti od časa pri ostalih klasifikatorjih ne podajo bistvenih signifikantnih informacij). V grafih 65, 66, 67 in 68, smo prikazali, kako se delnice vključujejo v strategije skozi čas. Vsak stolpec v grafu predstavlja delež delnic od vseh 370 možnih v enem trgovalnem dnevu. Opazimo lahko, da je kdaj delež res visok, dosega tudi okoli 70%, pa tudi relativno nizek in se giblje okoli nekaj %. Število delnic različno niha skozi celotno testno časovno obdobje, v splošnem pa je vidna porast deleža delnic konec leta 2008 in ob koncu leta 2011 ter začetku leta 2012 pri vseh metodah in izbranih klasifikatorjih. Z grafov lahko opazimo, da se znotraj metod za izbor atributov število delnic ne spreminja veliko glede na različno uporabljene klasifikatorje.

8.6.1 FCBF



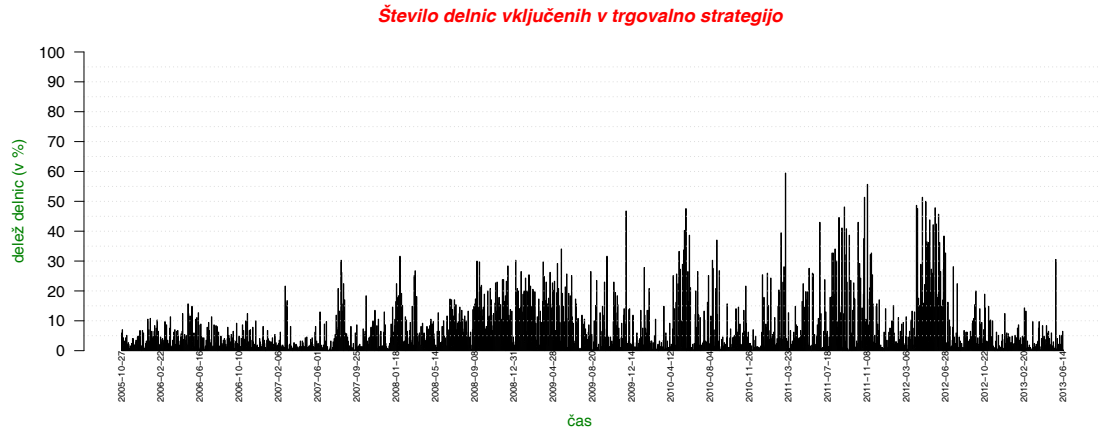
Slika 65: Prikazan delež delnic (v %) v odvisnosti od časa. Uporabili smo FCBF metodo za izbor relevantnih atributov. Prikazana je Vodena D_{LDA}–trgovalna strategija.

8.6.2 CFS



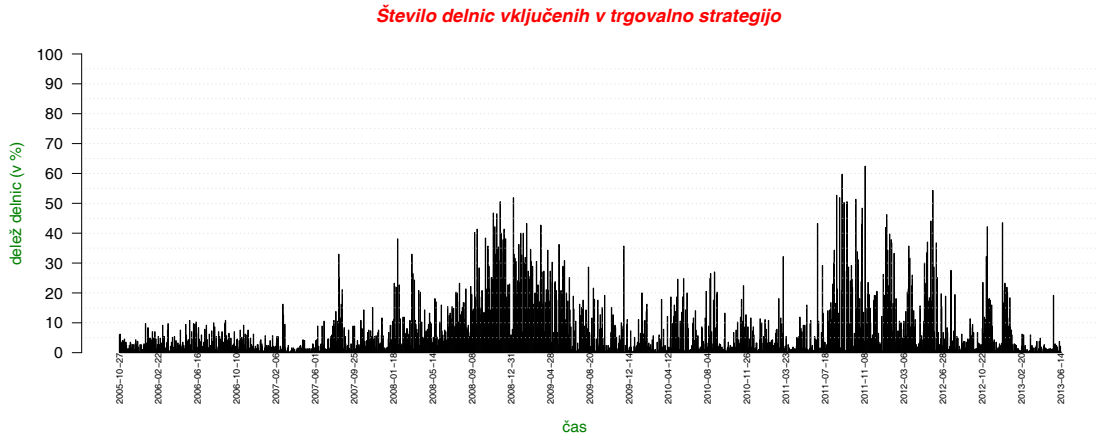
Slika 66: Delež delnic (v %) skozi testno časovno obdobje, ki jih vključimo v Vodeno D_{LDA}–trgovalno strategijo. Uporabili smo CFS metodo za izbor relevantnih atributov.

8.6.3 mRMR



Slika 67: Delež delnic (v $\%$) skozi testno časovno obdobje, ki jih vključimo v Vodeno D_{LDA}–trgovalno strategijo. Uporabili smo mRMR metodo za izbor relevantnih atributov.

8.6.4 CCCA

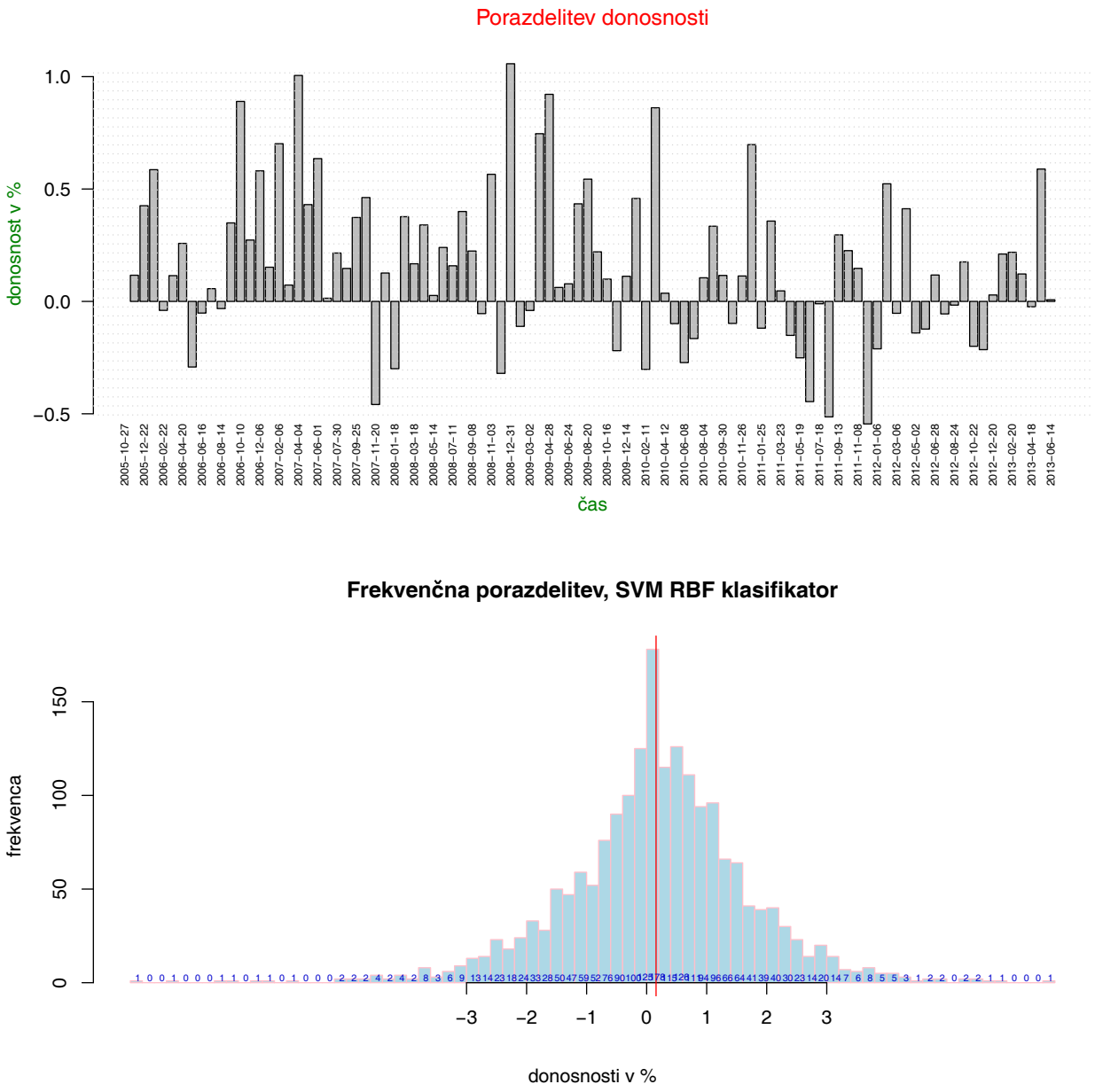


Slika 68: Delež delnic (v %) skozi testno časovno obdobje, ki jih vključimo v Vodeno D_{LDA}–trgovalno strategijo. Uporabili smo CCCA metodo za izbor relevantnih atributov.

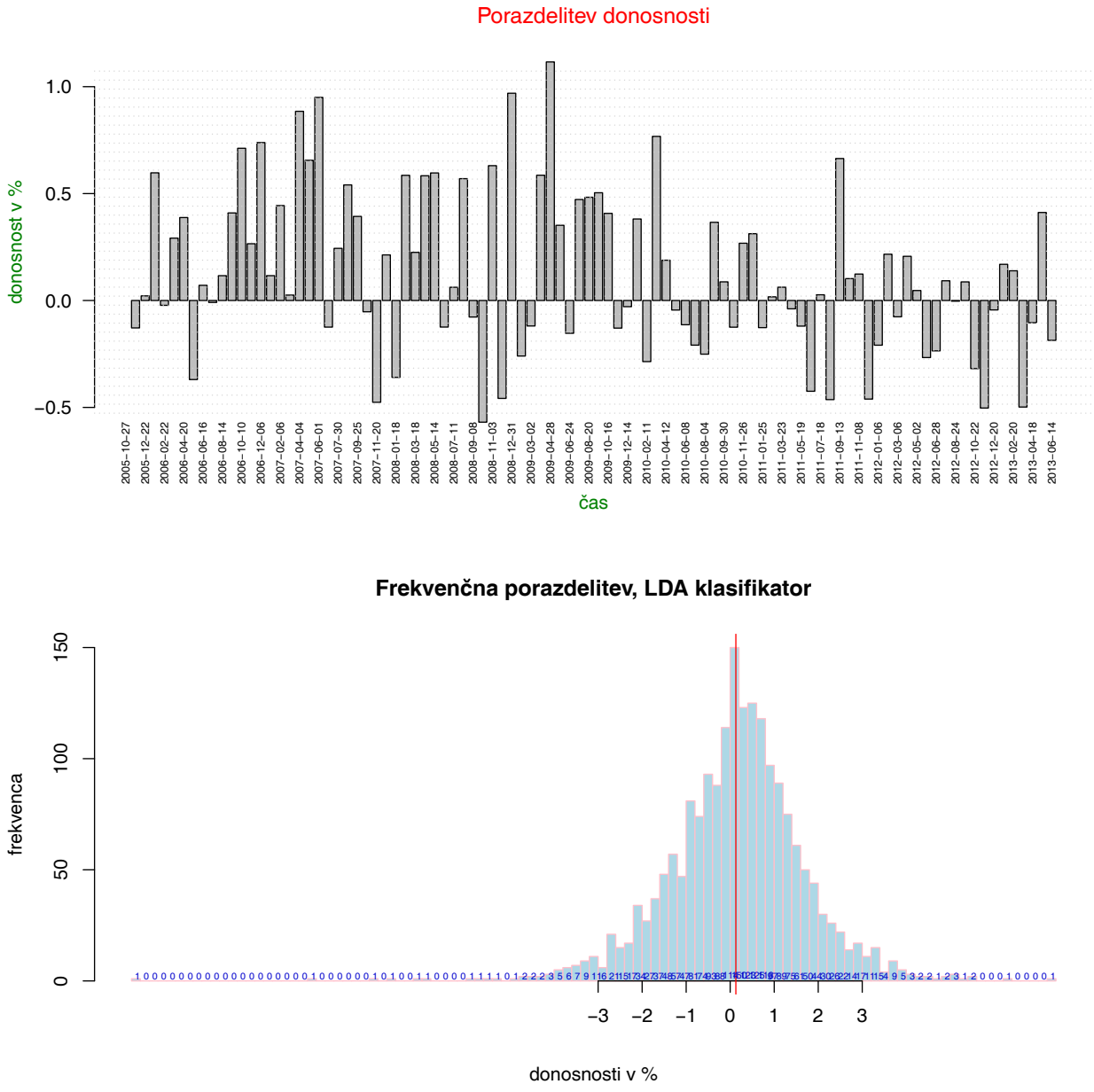
8.7 Porazdelitev donosnosti skozi čas, primerjava metod

V tem poglavju smo donosnost skozi čas primerjali med nekaj izbranimi Vodenimi D–trgovalnimi strategijami pri različnih metodah za izbor atributov (glej poglavje 8.4). Donosi so si po različnih metodah in Vodenih D–trgovalnih strategijah precej podobni: pri vseh je opazna razlika med pozitivnimi in negativnimi donosi, tako v frekvenci le teh, kot tudi po razponu, kar je razvidno iz grafov 69, 70, 71 in 72. mRMR metoda ima v začetku leta 2008 presenetljivo nizko donosnost, ta sega celo pod -1% donosnosti.

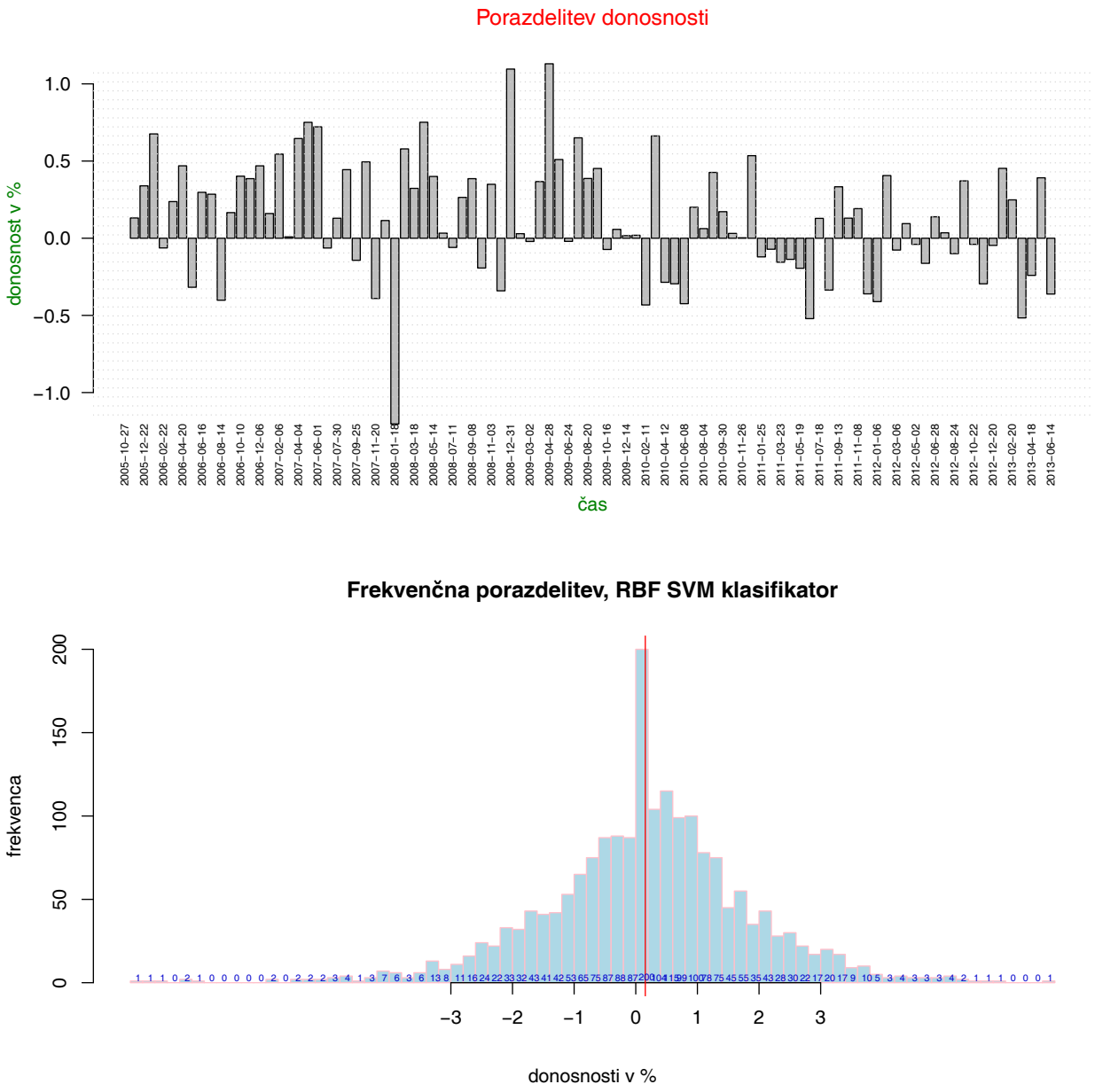
Odstotek vseh dni, ko smo trgovali s pozitivno donosnostjo je 58.17% za FCBF metodo (Vodena D_{RBF}–strategija), 57.39% za CFS metodo (Vodena D_{LDA}–strategija), 58.13% za mRMR metodo (Vodena D_{RBF}–strategija) ter 57.39% za CCCA metodo (Vodena D_{LDA}–strategija). V primerjavi z FSuC–comb–ward metodo je odstotek vseh dni trgovanja s pozitivno donosnostjo nižji (glej poglavje 7.3).



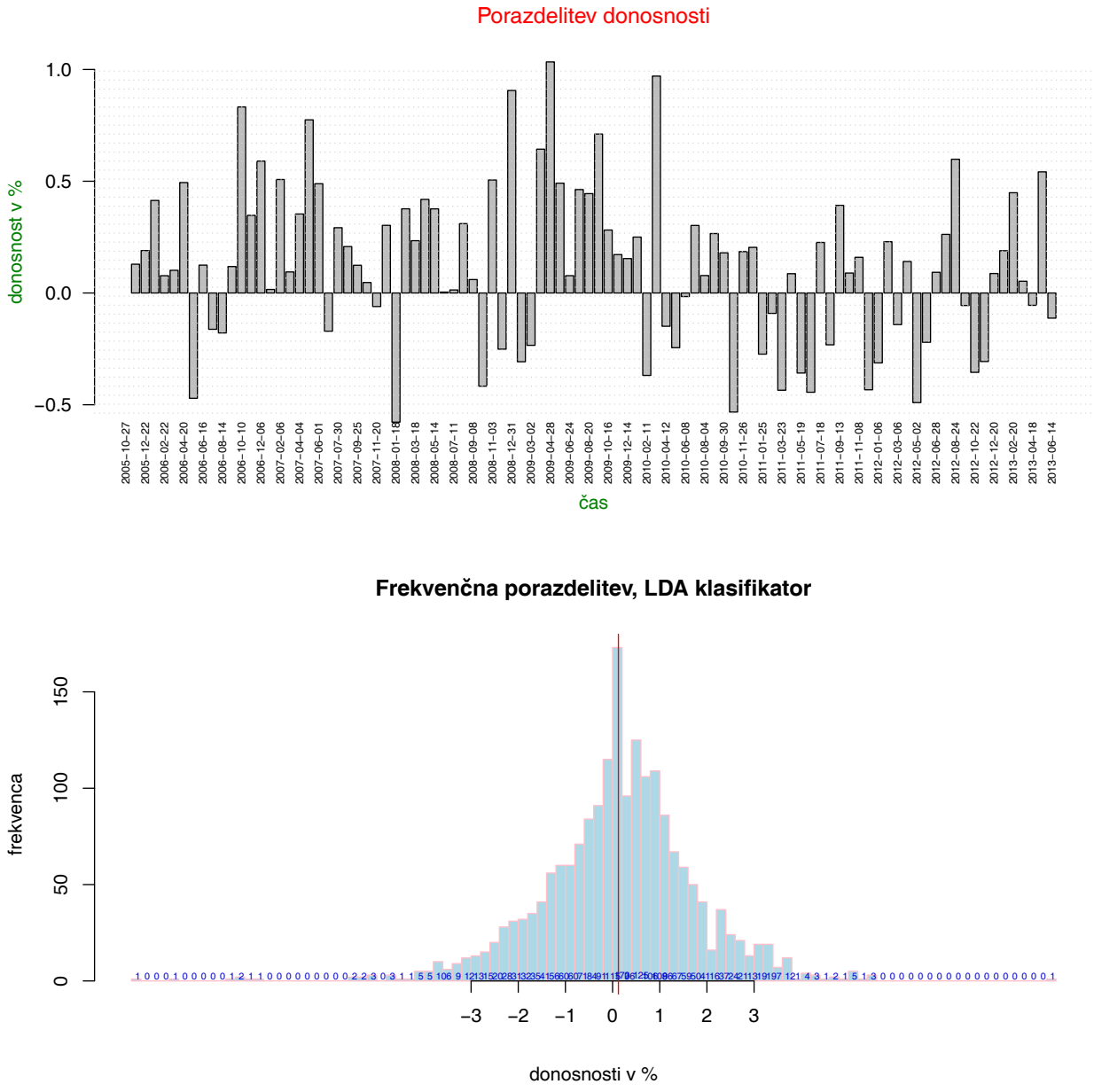
Slika 69: Porazdelitev donosnosti skozi čas ter frekvenčna porazdelitev, Vodena D_{RBF}–strategija, FCBF metoda. Povprečje donosnosti je 0.156%.



Slika 70: Porazdelitev donosnosti skozi čas ter frekvenčna porazdelitev, Vodena D_{LDA}–strategija, CFS metoda. Povprečje donosnosti je 0.128%.



Slika 71: Porazdelitev donosnosti skozi čas ter frekvenčna porazdelitev, Vodena D_{RBF}–strategija, mRMR metoda. Povprečje donosnosti je 0.123%.



Slika 72: Porazdelitev donosnosti skozi čas ter frekvenčna porazdelitev, Vodena D_{LDA}–strategija, CCCA metoda. Povprečje donosnosti je 0.124%.

8.8 Dolžina testnih množic

V članku [71] smo eksperimentalno določili dolžino testnih množic, vendar na nekoliko drugačnih podatkih: predhodno smo izbrali 11 atributov[2] in na enako dolgem časovnem obdobju, kot je omenjeno disertaciji, poizkušali določiti dolžino testnih množic, ki imajo ravno tako vpliv na klasifikacijske napovedi. Testirali smo različne dolžine testnih množic (označili s h). V izogib visoki časovni izvedbi smo pričeli z dolžino testnih množic pri $h = 5$ in končali pri $h = 160$. V tabelah 35, 36, 37 in 38, so podani rezultati eksperimenta za SVM z linearnim jedrom, SVM z RBF jedrom, LDA in NB klasifikatorji, kjer smo uporabili različne dolžine testnih množic h. Najvišja povprečna klasifikacijska točnost je pri $h = 20$ (in sicer pri NB klasifikatorju, linearnem SVM ter SVM klasifikatorju z RBF jedrom) ter pri $h = 10$ (LDA klasifikator). Za izgradnjo klasifikatorjev smo v disertaciji upoštevali dolžino testnih množic pri $h = 20$.

h	učna toč	**test toč ± std**	**senzit ± std**	**specif ± std**	**preciz 1 ± std**	**preciz -1 ± std**
160	63.54 ± 2.33	59.93 ± 4.32	60.42 ± 14.01	59.11 ± 12.92	60.34 ± 6.29	60.64 ± 6.28
80	63.53 ± 2.30	60.26 ± 5.65	61.32 ± 15.20	58.68 ± 14.21	60.55 ± 8.09	61.08 ± 8.26
40	63.53 ± 2.29	60.40 ± 7.64	61.30 ± 17.80	58.40 ± 16.97	60.55 ±11.12	61.13 ± 11.59
20	63.95 ± 2.47	60.66 ± 10.69	57.92 ± 21.74	62.18 ± 20.10	60.41 ± 17.08	61.71 ± 15.58
10	64.10 ± 2.62	**61.07 ± 15.03**	59.11 ±28.23	59.24 ± 27.28	60.53 ± 23.97	60.41 ± 23.68
5	64.20± 2.69	60.99 ± 21.37	56.33 ± 36.96	58.29 ± 36.55	59.43 ± 32.54	58.39 ± 31.84

Tabela 35: Klasifikacijski rezultati pri različnih dolžinah testnih množic. Uporabili smo LDA klasifikator.

h	učna toč	**test toč ± std**	**senzit ± std**	**specif ± std**	**preciz 1 ± std**	**preciz -1 ± std**
160	57.62 ± 2.09	56.94 ± 6.92	68.03 ± 17.39	41.82 ± 18.37	54.86 ± 9.72	57.77 ± 14.12
80	57.62 ± 2.09	56.95 ± 6.95	68.05 ± 17.39	41.80 ± 18.40	54.86 ± 9.75	57.77 ± 14.18
40	57.63 ± 2.09	56.92 ± 6.98	67.83 ±17.48	41.95 ± 18.46	54.81 ± 9.81	57.72± 14.26
20	57.87 ± 2.27	**57.28 ± 9.84**	60.92 ± 23.24	47.84 ± 23.27	54.01 ± 15.86	56.98 ± 18.77
10	58.15 ± 2.29	57.13 ± 14.24	63.32 ± 30.93	42.18 ± 30.93	53.85 ± 22.34	55.39 ± 26.98
5	58.25 ± 2.42	57.05 ± 20.71	62.09 ± 38.36	40.71 ± 38.23	52.90 ± 28.80	54.36 ± 32.85

Tabela 36: Klasifikacijski rezultati pri različnih dolžinah testnih množic. Uporabili smo NB klasifikator.

[2] V disertaciji smo atribute izbirali med 98 tehničnimi indikatorji s pomočjo metod za izbor atributov, v članku pa smo uporabili konstantnih 11 tehničnih indikatorjev, ki smo jih izbrali glede na predhodna sorodna raziskovalna dela. Ti izbrani indikatorji so: promet, SMA10, EMA10, EMA20, ZLEMA10, WMA10, RSI10, mom10, ROC10, SAR, CCI10.

h	učna toč	**test toč ± std**	**senzit ± std**	**specif ± std**	**preciz 1 ± std**	**preciz -1 ± std**
160	63.52 ± 2.33	60.09 ± 4.35	62.72 ± 15.09	56.83 ± 14.03	60.02 ± 6.38	61.45 ± 6.56
80	63.49 ± 2.33	60.36 ± 5.64	63.51 ± 16.04	56.34 ± 15.15	60.13 ± 8.06	61.80 ± 8.58
40	63.49 ± 2.32	60.49 ± 7.60	63.47 ± 18.41	56.01 ± 17.69	60.11 ± 10.96	61.82 ± 12.09
20	63.49 ± 2.32	**61.16 ± 10.68**	63.36 ± 22.04	56.44 ± 21.37	60.60 ± 15.67	62.08 ± 17.34
10	63.51 ± 2.31	60.57 ± 14.81	62.32 ± 27.55	54.53 ± 27.47	59.13 ± 22.52	61.10 ± 24.28
5	63.51 ± 2.31	60.57 ± 20.98	60.21 ± 35.51	52.75 ± 35.66	57.66 ± 30.06	59.26 ± 31.71

Tabela 37: Klasifikacijski rezultati pri različnih dolžinah testnih množic. Uporabili smo SVM klasifikator z linearnim jedrom.

h	učna toč	**test toč ± std**	**senzit ± std**	**specif ± std**	**preciz 1 ± std**	**preciz -1 ± std**
160	70.80 ± 5.78	58.82 ±4.37	60.20 ± 15.30	56.72 ± 14.30	58.93 ± 6.29	59.73 ± 6.53
80	70.80 ± 5.79	59.10 ± 5.69	61.16 ± 16.04	56.16 ± 15.22	59.06 ± 8.19	60.09 ± 8.52
40	70.78 ± 5.78	59.30 ± 7.76	61.34 ± 18.10	55.97 ± 17.52	59.16 ± 11.20	60.24 ± 11.86
20	70.79 ± 5.78	**59.39 ±10.71**	61.30 ± 21.83	55.54 ± 21.46	59.05 ± 15.98	60.15 ± 17.28
10	70.78 ± 5.77	59.34 ± 7.73	61.48 ± 18.08	55.87 ± 17.46	59.16 ± 11.19	60.31 ± 11.91
5	70.79 ± 5.77	59.33 ± 21.16	58.79 ± 36.21	53.10± 36.21	56.97 ± 30.47	58.22 ± 31.82

Tabela 38: Klasifikacijski rezultati pri različnih dolžinah testnih množic. Uporabili smo SVM klasifikator z RBF jedrom.

8.9 Primerjava klasifikacijskih rezultatov

Volatilnost na najvišjih dnevnih tečajih je signifikantno nižja v primerjavi z zaključnimi dnevnimi tečaji, zato smo se v disertaciji ukvarjali z napovedovanjem najvišjih dnevnih tečajev.
Za ilustracijo smo primerjali klasifikacijske točnosti, kjer smo napovedovali zaključne dnevne tečaje ter najvišje dnevne tečaje. Rezultati so podani na izbranih enajstih tehničnih indikatorjih [3] in prikazani v tabelah 39 in 40, v katerih je razvidno, da z izbranimi klasifikacijskimi metodami dobimo v vseh primerih višje točnosti, kadar napovedujemo gibanje najvišjih dnevnih tečajev. Ti rezultati so bili tudi motivacija, zakaj opustiti napovedi zaključnih dnevnih tečajev.

metoda	učna toč	**test toč ± std**	**senzit ± std**	**specif ± std**	**preciz 1 ± std**	**preciz -1 ± std**
LDA	63.95 ± 2.47	60.66 ± 10.69	57.92 ± 21.74	62.18 ± 20.10	60.41 ± 17.08	61.71 ± 15.58
NB	57.87 ± 2.27	57.28 ± 9.84	60.92 ± 23.24	47.84 ± 23.27	54.01 ± 15.86	56.98 ± 18.77
linear SVM	63.49 ± 2.32	61.16 ± 10.68	63.36 ± 22.04	56.44 ± 21.37	60.60 ± 15.67	62.08 ± 17.34
RBF SVM	70.79 ± 5.78	59.39 ±10.71	61.30 ± 21.83	55.54 ± 21.46	59.05 ± 15.98	60.15 ± 17.28

Tabela 39: Klasifikacijski rezultati z uporabo različnih klasifikatorjev, napovedovanje najvišjih tečajev.

metoda	učna toč	**test toč ± std**	**senzit ± std**	**specif ± std**	**preciz 1 ± std**	**preciz -1 ± std**
LDA	56.18 ± 1.93	50.77 ± 10.83	53.81 ± 28.81	49.17 ± 28.96	52.78 ± 18.20	51.55 ± 19.69
NB	53.37 ± 1.97	50.71 ± 10.87	49.35 ± 30.53	55.14 ± 30.54	54.67 ± 20.01	52.55 ± 18.92
linear SVM	55.71 ± 1.87	50.76 ± 10.82	54.74 ± 33.29	47.94 ± 33.52	52.99 ± 18.39	51.72 ± 19.76
RBF SVM	62.03 ± 4.23	50.87 ± 10.91	54.82 ± 29.31	48.24 ± 29.53	52.91 ± 18.23	51.79 ± 20.21

Tabela 40: Klasifikacijski rezultati z uporabo različnih klasifikatorjev, napovedovanje zaključnih tečajev.

[3] Izbrani indikatorji so: promet, SMA10, EMA10, EMA20, ZLEMA10, WMA10, RSI10, mom10, ROC10, SAR, CCI10.

8.10 Primerjava standardnih deviacij med najvišjimi in zaključnimi tečaji

Izračunane volatilnosti (standardne deviacije) na 2604 trgovalnih dnevih za vsako delnico posebej so prikazane v tabeli 41. Iz tabele je razvidno, da so volatilnosti (standardne deviacije) na najvišjih dnevnih tečajih nižje kot volatilnosti na zaključnih dnevnih tečajih pri večini delnic. Rezultati so urejeni glede na volatilnost pri najvišjih dnevnih tečajih.

	SO	ED	JNJ	KMB	GIS	K	PEP	KO	WMT	D	MKC	SCG	CL	CPB	DUK	DTE	CLX	PNW	ABT	NU
std close	1.05	1.03	1.03	1.10	1.06	1.08	1.11	1.19	1.23	1.21	1.20	1.19	1.24	1.17	1.29	1.30	1.18	1.26	1.28	1.33
std high	0.83	0.85	0.87	0.89	0.90	0.93	0.93	0.99	1.03	1.04	1.04	1.04	1.05	1.07	1.07	1.08	1.08	1.09	1.10	1.11
	ADP	AEE	PCG	IBM	AEP	HRL	BDX	MO	RTN	MCD	XOM	VZ	LMT	MMM	UPS	SYY	NI	ECL	NEE	NOC
std close	1.37	1.36	1.37	1.36	1.36	1.27	1.30	1.35	1.39	1.36	1.58	1.41	1.45	1.43	1.42	1.40	1.44	1.49	1.43	1.46
std high	1.12	1.13	1.14	1.15	1.16	1.17	1.17	1.18	1.18	1.18	1.19	1.20	1.20	1.20	1.22	1.22	1.22	1.22	1.22	1.23
	ETR	GPC	LLY	TEG	LH	UTX	LLL	PPL	EIX	SRE	WM	POM	CVX	PFE	DNB	GD	COST	IFF	PAYX	FE
std close	1.40	1.40	1.47	1.46	1.36	1.54	1.51	1.47	1.51	1.47	1.47	1.46	1.69	1.52	1.50	1.55	1.55	1.58	1.55	1.54
std high	1.24	1.25	1.26	1.26	1.26	1.27	1.27	1.28	1.29	1.29	1.30	1.30	1.31	1.32	1.32	1.32	1.32	1.33	1.33	1.34
	MDT	TE	BMY	SIAL	EXC	BMS	HCN	KR	CB	DGX	CNP	AZO	PX	OMC	TAP	PBCT	ITW	XRAY	ABC	AMGN
std close	1.53	1.60	1.48	1.63	1.62	1.57	1.94	1.57	1.74	1.51	1.61	1.67	1.72	1.62	1.60	1.78	1.67	1.60	1.58	1.68
std high	1.35	1.35	1.36	1.37	1.37	1.37	1.37	1.37	1.38	1.38	1.40	1.41	1.42	1.43	1.43	1.44	1.45	1.45	1.46	1.47
	CINF	SYK	MSFT	SRCL	TRV	GWW	DD	EFX	NKE	RSG	WAG	DIS	SHW	HON	CVS	CTL	CTAS	AGN	COP	INTU
std close	1.88	1.65	1.70	1.63	1.92	1.65	1.75	1.66	1.74	1.64	1.66	1.79	1.75	1.76	1.72	1.68	1.70	1.74	1.91	1.88
std high	1.47	1.47	1.47	1.47	1.48	1.48	1.48	1.49	1.49	1.49	1.52	1.52	1.52	1.53	1.53	1.54	1.54	1.54	1.56	1.56
	PGR	TJX	PDCO	VFC	HCBK	EMR	HD	PBI	LLTC	MCK	BA	MRK	PCL	ACE	AON	TMO	INTC	DVA	WPO	UNP
std close	1.86	1.74	1.83	1.82	2.01	1.82	1.77	1.76	1.86	1.78	1.83	1.79	2.15	1.97	1.68	1.77	1.95	1.66	1.80	1.87
std high	1.56	1.56	1.58	1.58	1.58	1.58	1.58	1.58	1.58	1.59	1.59	1.59	1.59	1.60	1.60	1.61	1.62	1.62	1.62	1.63
	FDX	MMC	TMK	CMS	DOV	ORLY	L	TGT	LEG	MAT	ROST	FTR	GE	SNA	CCE	PLL	NTRS	TSS	ROP	FRX
std close	1.91	1.80	2.11	1.62	1.92	1.85	1.99	1.91	1.98	1.82	1.93	1.94	1.94	1.92	1.90	1.97	2.29	1.87	1.94	1.88
std high	1.64	1.64	1.65	1.65	1.65	1.66	1.66	1.67	1.68	1.69	1.69	1.69	1.69	1.70	1.70	1.70	1.71	1.71	1.71	1.72
	AVY	NSC	ALL	STJ	CA	SWK	QCOM	CSCO	CMCSA	LOW	PH	ESRX	OKE	GILD	BBBY	PSA	ADI	TXN	SWY	BXP
std close	1.93	2.01	2.18	1.90	1.92	2.01	2.07	2.01	1.98	1.98	2.04	2.08	1.85	2.07	2.06	2.30	2.03	2.04	1.93	2.48
std high	1.72	1.72	1.73	1.73	1.74	1.75	1.76	1.76	1.77	1.77	1.78	1.78	1.79	1.79	1.79	1.80	1.80	1.81	1.81	1.82
	EL	SBUX	VAR	AVB	USB	ADM	SEE	AMT	MCHP	WAT	IRM	FAST	EMN	MWV	VTR	BMC	EMC	VNO	KSS	MAR
std close	1.93	2.13	1.97	2.41	2.38	2.17	2.02	2.07	2.07	1.96	1.92	2.15	2.14	2.09	2.45	2.03	2.11	2.59	2.09	2.16
std high	1.82	1.83	1.83	1.83	1.84	1.84	1.84	1.84	1.84	1.84	1.84	1.85	1.85	1.85	1.86	1.87	1.88	1.88	1.88	1.88
	UNH	CSX	ARG	SPLS	EQT	BEN	AMAT	HCP	MTB	FMC	OXY	CCL	HRB	EQR	APH	SPG	DELL	PCP	HPQ	SYMC
std close	2.20	2.13	2.22	2.16	2.09	2.34	2.22	2.56	2.25	2.34	2.32	2.18	2.08	2.67	2.16	2.68	2.15	2.16	2.09	2.26
std high	1.88	1.88	1.89	1.90	1.90	1.90	1.90	1.90	1.90	1.91	1.91	1.91	1.92	1.92	1.92	1.92	1.93	1.94	1.94	1.94
	STZ	PKI	EXPD	MRO	CSC	WY	PCAR	JCI	AXP	HRS	DVN	APA	MYL	ROK	DE	FII	KLAC	MUR	SLB	RRD
std close	2.12	2.13	2.21	2.40	2.15	2.17	2.33	2.26	2.44	2.08	2.32	2.31	2.14	2.29	2.25	2.25	2.31	2.30	2.36	2.22
std high	1.95	1.95	1.95	1.96	1.96	1.96	1.96	1.96	1.97	1.97	1.97	1.97	1.98	1.98	1.98	1.98	1.99	1.99	2.00	2.00
	AVP	ALTR	A	BBT	NBL	DOW	ADBE	DRI	R	NWL	FDO	TROW	AET	IR	AAPL	NEM	DO	LUK	TSN	JPM
std close	2.16	2.34	2.24	2.49	2.43	2.30	2.27	2.24	2.35	2.19	2.15	2.52	2.30	2.27	2.37	2.44	2.43	2.55	2.26	2.72
std high	2.00	2.01	2.01	2.01	2.01	2.01	2.02	2.02	2.03	2.03	2.04	2.04	2.04	2.05	2.06	2.06	2.07	2.08	2.08	2.09
	GS	BLK	PNC	AFL	SCHW	BK	LTD	RL	BSX	UNM	DF	BHI	CI	RHI	IP	EOG	NE	LXK	VRSN	HAL
std close	2.51	2.40	2.75	2.72	2.53	2.65	2.34	2.32	2.39	2.79	2.34	2.55	2.51	2.34	2.53	2.52	2.61	2.42	2.52	2.60
std high	2.09	2.10	2.10	2.11	2.11	2.11	2.11	2.12	2.13	2.16	2.16	2.16	2.17	2.17	2.18	2.19	2.19	2.19	2.20	2.22
	HUM	TIF	FLIR	CERN	AES	HES	CAM	APC	ITT	ADSK	FLS	CTSH	JEC	WHR	CTXS	JWN	GLW	VMC	IGT	BBY
std close	2.43	2.46	2.43	2.38	2.67	2.67	2.65	2.57	2.36	2.56	2.65	2.59	2.63	2.52	2.56	2.62	2.57	2.41	2.59	2.50
std high	2.22	2.22	2.22	2.22	2.23	2.24	2.24	2.24	2.25	2.25	2.26	2.26	2.27	2.27	2.27	2.29	2.29	2.30	2.31	2.31
	CMA	HST	MAS	HOT	VLO	NTAP	HP	WFC	PLD	HOG	CVC	RDC	M	PXD	CPWR	STT	APOL	BRCM	TLAB	URBN
std close	2.77	3.20	2.67	2.66	2.73	2.70	2.80	2.96	3.20	2.60	2.50	2.85	2.72	2.80	2.55	3.11	2.66	2.81	2.82	2.73
std high	2.32	2.32	2.35	2.35	2.35	2.36	2.37	2.38	2.38	2.39	2.40	2.40	2.41	2.41	2.44	2.44	2.44	2.47	2.48	2.49
	NBR	AIV	AN	BIG	NFX	COG	AMZN	RRC	JBL	CMI	JNPR	JCP	IVZ	ANF	TER	STI	OI	CHK	SVU	LM
std close	2.91	3.02	2.44	2.75	2.85	2.87	2.77	2.86	3.04	2.85	2.85	2.75	3.02	2.90	2.91	3.27	2.83	3.00	2.96	3.08
std high	2.49	2.49	2.51	2.52	2.52	2.53	2.53	2.53	2.55	2.55	2.56	2.58	2.59	2.60	2.62	2.63	2.63	2.63	2.65	2.66
	PWR	DV	LSI	SWN	COF	TXT	DNR	FCX	IPG	WDC	TSO	HAR	KMX	RHT	FFIV	PCLN	GCI	F	PHM	X
std close	3.06	2.89	3.12	3.14	3.24	3.07	3.18	3.20	2.73	3.06	3.33	3.03	2.75	3.07	3.17	3.20	2.96	3.02	3.42	3.51
std high	2.67	2.70	2.72	2.72	2.76	2.76	2.77	2.77	2.79	2.79	2.81	2.81	2.83	2.83	2.92	2.93	2.93	2.97	2.98	2.99
	BAC	ZION	CNX	MU	NVDA	ATI	DHI	SLM	GT	MS	LNC	JDSU	AMD	MWW	C	LEN	THC	AKAM	SNDK	RF
std close	3.57	3.44	3.42	3.43	3.38	3.62	3.37	3.66	3.40	3.72	4.02	3.63	3.57	3.43	3.78	3.63	3.57	3.70	3.67	3.93
std high	2.99	3.00	3.02	3.05	3.06	3.06	3.06	3.10	3.14	3.19	3.21	3.23	3.23	3.23	3.23	3.29	3.29	3.30	3.35	3.39
	HBAN	CLF	FITB	ETFC	KEY	AKS	HIG	AIG												
std close	4.25	4.06	4.11	4.32	3.49	4.39	4.58	4.80												
std high	3.42	3.42	3.47	3.79	3.81	3.92	4.19	4.75												

Tabela 41: Primerjava standardnih deviacij.

9 ZAKLJUČEK

V doktorski disertaciji smo predstavili postopek učenja klasifikacijskih modelov na finančnih podatkih. Preveriti smo želeli, ali se na preteklih finančnih podatkih klasifikacijski modeli znajo naučiti tako, da bodo znali napovedati smeri gibanja delniškega trga. Pri sestavi klasifikacijskih modelov je ključnega pomena, kakšni so vhodni podatki. S pomočjo multivariatnih filtrirnih metod smo poiskali relevantne tehnične indikatorje, ki so nato služili kot vhodni podatki pri sestavi modelov. Predlagali smo tudi novo multivariatno filtrirno metodo 'FSuC–ward–comb', s katero dosegamo najvišje klasifikacijske točnosti na testnih množicah v primerjavi z ostalimi metodami. Pri tem nas je zanimalo, kateri so najbolj relevantni tehnični indikatorji, ki največ prispevajo h klasifikacijski točnosti. Zanimalo nas je tudi, kakšno je razmerje med učno in testno množico, tako da je model vračal čim manjše napake. Dolžino obeh množic smo določili eksperimentalno v članku [71]. Rezultate dnevnih napovedi uvrščanja smo uporabili kot podporo odločanja pri samem trgovanju. Na primeru enostavnih trgovalnih strategij smo ugotovili, da vključitev napovedi gibanja posameznih delnic pripomore k doseganju višjih rezultatov.

9.1 Klasifikacijski modeli in izbor atributov

Podatki, na katerih smo izvedli eksperimentalno delo so obsežni, saj vključimo velik nabor tehničnih indikatorjev (98 tehničnih indikatorjev) za vsako delnico posebej. Skozi analizo so dobljeni rezultati vedno bolj kazali dejansko sliko o zakonitostih na izbranih podatkih. Ker smo odkrili, da je volatilnost na najvišjih dnevnih tečajih (ang. 'daily high prices') signifikantno nižja kot na zaključnih dnevnih tečajih (ang. 'daily close prices'), smo za razliko od večine raziskovalnih del na tem področju napovedovali gibanje delnic glede na najvišje dnevne tečaje. Napovedi dajo tudi višje klasifikacijske točnosti v primerjavi z zaključnimi dnevnimi tečaji. V disertaciji zgradimo nekaj klasifikacijskih modelov, ki za vhodne podatke vzamejo najbolj relevantne tehnične indikatorje. Klasifikacijski rezultati so seveda različni glede na uporabljene metode za izbor atributov (glej poglavje 6.4). Najvišji klasifikacijski rezultati so doseženi pri predlagani 'FSuC–ward–comb' metodi in sicer z LDA klasifikatorjem, če gledamo klasifikacijsko točnost na testnih množicah (ravno tako je najvišja vrednost pri preciznosti za razred 1). Zanimalo nas je tudi, kateri so najbolj informativni tehnični indikatorji, ki lahko največ povedo o prihodnjih gibanjih posameznih delnic. V našem delu smo se osredotočili le na multivariatne filtrirne metode, ki so razmeroma hitre in vzamejo v obzir tudi odvisnosti med ostalimi tehničnimi indikatorji, kar pomeni, da pri sestavi klasifikacijskega modela nismo vključevali odvečnih atributov. Skupno vsem metodam za izbor atributov je, da so metode vrnile tehnične indikatorje, ki v izračun zajamejo le majhno število preteklih dni, kar pomeni, da so dnevne napovedi odvisne od informacij, ki so se zgodile le nekaj dni nazaj. Predlagana metoda 'FSuC–ward–comb' v povprečju na vseh delnicah vrne kot najbolj relevantne tehnične indikatorje drseče sredine s krajšim obdobjem časovnih enot zajetih v izračun (npr. SMA2, WMA3, EMA2); metoda FCBF vrne tehnične indikatorje VHF2, VHF3, VHF5; metoda CFS kot najbolj relevantne tehnične indikatorje vrne celotno družino ATR; mRMR vrne ATR2 in družino

tehničnih indikatorjev VHF ter metoda CCCA kot najbolj relevanten tehničen indikator vrne PBands. Metodi mRMR in CCCA odstopata po klasifikacijskih rezultatih in so vrednosti nekoliko nižje, zato se pri njima lahko vprašamo o smiselnosti izbora tehničnih indikatorjev. Opisani rezultati so povprečeni glede na vseh 370 delnic in zato lahko pri posameznih delnicah izstopajo tudi drugi relevantni atributi. Za primer smo izbrali 3 delnice, CCL, AMZN in AAPL in opazovali, kako se relevantnost atributov pokriva s tehničnimi indikatorji, ki so povprečeni po vseh delnicah; največ odstopanj po delnicah je razvidno pri metodi CFS, sicer pa ni razvidnih bistvenih odstopanj (glede na te 3 izbrane delnice).
Zanimali so nas tudi razponi (porazdelitve) klasifikacijskih točnostih na testnih množicah. Rezultati za metodo 'FSuC–ward–comb' se gibljejo v razponu od 55.21%–67.29%, ki so tudi najvišje doseženi rezultati, pri CFS metodi se klasifikacijske točnosti na testni množici gibljejo v razponu od 46.77%–64.01%, pri FCBF metodi je razpon klasifikacijske točnosti od 53.39%–63.70%, pri mRMR metodi je razpon točnosti od 48.80%–60.94% ter pri CCCA metodi dosežemo najnižje klasifikacijske točnosti v razponu od 49.90%–59.79%.

9.2 Analiza uspešnosti trgovalnih strategij

Klasifikacijska natančnost na eksperimentih je okoli 60%, ki je dovolj visoka za izgradnjo profitnih trgovalnih strategij. Napovedi gibanja delnic na najvišjih dnevnih tečajih smo vključili v predlagane trgovalne strategije (Vodene D–trgovalne strategije) in uspešnost le teh primerjali s strategijami, ki ne vključujejo napovedi gibanja. Teh strategij je toliko kot je različnih klasifikacijskih modelov, torej za 4 klasifikatorje (LDA, NB, lin SVM in RBF SVM), kjer za vsak klasifikator dobimo različne vhodne podatke, torej različne relevantne atribute, prinese 20 različnih Vodenih D–trgovalnih strategij. V Vodenih D–trgovalnih strategijah smo s pomočjo klasifikacijskih modelov vključili predvidevanja o gibanju delnic.
Najvišje rezultate glede na CAGR vrednost, Sharpeov koeficient, informacijski koeficient, kazalnik Sortino in povprečje ima Vodena D_{NB}–strategija, ki jo dobimo tako, da upoštevamo klasifikacijske napovedi NB klasifikatorja, za vhodne podatke pa uporabimo tehnične indikatorje, dobljene s predlagano 'FSuC–ward–comb' metodo. Zanimivo je, da uspešnost izidov na izbranih kvantitativnih kazalnikih ne kaže podobne slike kot pri klasifikacijskih rezultatih (pri klasifikacijskih rezultatih najvišjo točnost dosežemo z LDA klasifikatorjem, pri Vodenih D–trgovalnih strategijah pa to dosežemo z NB klasifikatorjem, glej poglavje 6.4 in poglavje 7). Če primerjamo uspešnost Vodenih D–trgovalnih strategij med seboj, lahko opazimo, da med njimi ne moremo govoriti o signifikantnih razlikah, je pa prisotnih nekoliko izjem (glej poglavje 8.4 in 8.5). Vse Vodene D–trgovalne strategije vrnejo boljše rezultate kot Naivne D–trgovalne strategije (izjeme pri metodah mRMR in CCCA), indeks $S\&P500$ in Primerjalne strategije, kar pomeni, da z vključitvijo napovedi v predlagane trgovalne strategije te dajo višje rezultate kot pa če ne bi vključili napovedi (glej tabele 14, 24, 25, 26, 27 in 28).
Veliko je možnih razlogov, ki lahko vplivajo na izid trgovalnih strategij: izbor D pragov, izbor delnic, izbor klasifikacijskih modelov, izbor relevantnih tehničnih indikatorjev, relativna vrednost izgube/profita itd., zato smo si natančneje pogledali nekaj izmed naštetih razlogov.

Zanimalo nas je, s koliko delnicami trgujejo izbrane strategije. Rezultati so si med seboj precej podobni (primerjava po strategijah); od vseh 370 delnic, ki jih imamo dnevno na razpolago, delež vključenih delnic variira od nekaj procentov pa do 90% (po dnevih). Opazili smo, da je v nekaterih obdobjih intenziteta vključenih delnic v Vodene D–trgovalne strategije povečana, npr. v zadnjem četrtletju leta 2008 (okoli oktobra 2008) in sredi leta 2011, 2012. Za ta obdobja bi lahko rekli, da so čas krize.
V disertaciji smo predstavili tudi, kako se donosnosti spreminjajo skozi čas. Z grafov v poglavju 7.3 je razvidno, da je več pozitivnih donosov kot negativnih in ti dosegajo višje razpone. Odstotek pozitivnih donosov pri FSuC–ward–comb metodi pri vseh klasifikatorjih je nekoliko nižji od 60%.
Če primejamo dobljene rezultate v članku [71], kjer smo eksperimentalno delo izvedli na istem časovnem okvirju, vendar izbrali fiksne tehnične indikatorje pri sestavi modela (vzeli smo 11 tehničnih indikatorjev in v njihov izračun zajeli 10 preteklih trgovalnih dni oziroma dolžina 'lag'-a je 10), klasifikacijski rezultati niso veliko nižji, isto velja za uspešnost trgovalnih strategij. Ker z izborom indikatorjev ne dobimo drastičnega izboljšanja, domnevamo:

- obstajajo še veliko bolj informativni indikatorji, ki jih morda nismo zajeli v ta okvir raziskovanja, npr. fundamentalni indikatorji;
- lahko z drugimi uporabljenimi metodami za izbor indikatorjev dosežemo še višjo napovedno moč;
- je za tak tip podatkov primernejša povsem drugačna metoda nadzorovanega učenja;
- podatki vsebujejo preveč šuma, če se jih napoveduje na dnevni ravni.

9.3 Prispevki k znanosti

Znanstvene prispevke disertacije lahko strnemo v nekaj glavnih točk:

- V podakih smo odkrili nižjo volatilnost na najvišjih in najnižjih dnevnih tečajih (kot pa na zaključnih dnevnih tečajih, glej sliko 1 v poglavju 1). Klasifikacijska točnost na testni množici na teh podatkih presega 60%.
- Predlagali smo metodo za izbor atributov ('FSuC–ward–comb'), ki v primerjavi z ostalimi reprezentativnimi metodami vrača najvišje klasifikacijske rezultate na testnih množicah, kar pomeni, da atributi, ki jih vrne metoda, največ prispevajo k napovedni moči ([70]).
- Raziskali smo, kolikšen del vzeti za učno in kolikšen del za testno množico (za validacijo), tako da bodo naučeni modeli vrnili čim manjše napake.
- V disertaciji smo predlagali niz trgovalnih strategij (Vodene D–trgovalne strategije), ki so prilagojene glede na napovedi gibanja za najvišje dnevne tečaje. Te trgovalne strategije (v katerih smo vključili klasifikacijske napovedi gibanja delnic), so po uspešnosti višje kot trgovalne strategije, v katerih ne vključujemo klasifikacijskih napovedi, vendar pa Wilcoxonovi testi s predznačenimi

rangi kažejo na to, da se nekatere kombinacije primerjav med Vodenimi D–trgovalnimi strategijami z Naivnimi D– trgovalnimi strategijami, ne razlikujejo signifikantno. Slednje pomeni, da sama vednost, kako se bodo gibali najvišji dnevni tečaji, ne vpliva signifikantno na uspešnost trgovalnih strategij.

9.4 Odprti problemi

Vsako obsežnejše delo pušča mnoga neodgovorjena vprašanja, hkrati pa zastavlja veliko novih odprtih problemov. Sledi opis nekaterih izmed njih.

- V tem delu smo dnevno napovedovali smer gibanja delnic. Zanima nas, kako se obnese trgovanje na dolgi rok, npr. za vsak kvartal bi lahko napovedali smer gibanja delnic za naslednji kvartal. Pri tem bi lahko vključili tudi fundamentlane indikatorje, ki jih lahko dobimo v poročilih podjetij vsako četrtletje.

- V podatkih smo odkrili nižjo volatilnost na najvišjih in najnižjih dnevnih tečajih, v našem delu pa smo uporabili le najvišje dnevne napovedi gibanj tečajev. Kot razširitev te študije, podobno raziskavo lahko naredimo na dnevnih najnižjih tečajih. S kombiniranjem obeh nizov napovedi lahko konstruiramo globalno strategijo, ki se sestoji iz dolgih pozicij, kot smo opisali v tej disertaciji (ang. 'buy/long position') in kratkih pozicij (ang. 'sell/short position'), kjer uporabimo napovedi na najnižjih dnevnih tečajev. Ker ni enolične kombinacije, kako skombinirati ta dva pristopa v globalno strategijo, se poraja vprašanje, ali lahko s kakšno kombinacijo konstruiramo tržno nevtralni portfelj.

- Predlagana metoda 'FSuC–ward–comb' ima tudi svoje pomanjkljivosti, omejili smo se namreč le na nekaj metod za razvrščanje v skupine. Obstajajo številne druge metode, naj kot primer navedemo 'model based clustering', ki pa je pa časovno veliko potratnejši.

- Portfelj izbranih delnic bi bilo zanimivo nadgraditi tako, da vanj vključimo različne deleže delnic in upoštevati različne mere tveganja. V članku [71] smo poizkusili s tremi različnimi pristopi, ki so upoštevali klasifikacijske rezultate, pa niso prinesli boljših rezultatov kot standardno uporabljen portfelj z enakomernimi utežmi.

- Predlagane trgovalne strategije so zgolj oris strategij, ki jih uporabljajo upravljavci premoženja. Kot smo že opisali v poglavju 4.1, trgovalne strategije lahko približamo realnim sistemom za trgovanje tako, da vključimo tudi stroške trgovanja (npr. provizijo).

- Pri nehierarhičnem razvrščanju v skupine smo predstavnike posameznih skupin (voditelje) določili slučajno, kar je najenostavnejši način, lahko pa bi jih določili tudi tako, da voditelje maksimalno razpršimo po prostoru. Na ta način lahko zmanjšamo število iteracijskih korakov.

- Z razvrstitvami v skupine dobimo v splošnem le lokalni minimum kriterijske funkcije za dano sosedstveno strukturo. Da bi dobili čim boljšo razvrstitev (po možnosti globalni minimum funkcije), postopek ponovimo z različnimi začetnimi razvrstitvami.
- Razvrstitve v skupine lahko izboljšamo tudi z uporabo drugih distančnih mer. Pri eksperimentih smo se odločlili za najenostavnejšo: evklidsko razdaljo.
- V poglavju 6.4 smo sestavili nov klasifikator tako, da smo kombinirali med tremi različnimi klasifikatorji glede na rezultate na učnih množicah. Na testnih množicah smo uporabili klasifikator, ki je vračal najvišje klasifikacijske točnosti na učnih množicah. Obstajajo pa tudi drugi pristopi, kako skombinirati napovedi različnih klasifikatorjev, kot npr. glasovanje (ang. voting), uteženo glasovanje, kombiniranje po metodi naivnega Bayesa [55], itd. ki bi jih lahko preizkusili.

References

[1] http://dat.si/publikacije/Article/Strojno-u--269-enje/66. (20. 06. 2015).

[2] http://code.google.com/p/ml-dolev-amit/source/browse/trunk/weka/required_datasets/?r=35.

[3] http://www.poems.com.hk/en-us/customer-service/fees-and-charges/basic/. (20. 6. 2015).

[4] https://www.interactivebrokers.com/en/index.php?f=commission&p=stocks1. (20. 6. 2015).

[5] http://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test. (20. 6. 2015).

[6] S. B. Achelis. *Technical Analysis from A to Z.* McGraw–Hill, New York, 2001.

[7] H. Almuallim and T. G. Dietterich. Learning boolean concepts in the presence of many irrelevant features. *Artificial Intelligence*, 69(1):279–305, 1994.

[8] George S. Atsalakis and Kimon P. Valavanis. Surveying stock market forecasting techniques—Part II: Soft computing methods. *Expert Systems with Applications*, 36:5932–5941, 2009.

[9] K. Bache and M. Lichman. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml, 2013. (20. 06. 2015).

[10] Vladimir Batagelj. Diskriminantna analiza. http://vlado.fmf.uni-lj.si/vlado/podstat/Mva/DA.pdf, 2015. (20. 06. 2015).

[11] Asa Ben-Hur and Jason Weston. A User's Guide to Support Vector Machines. *Data Mining Techniques for the Life Sciences Methods in Molecular Biology*, 609:223–239, 2010.

[12] T. Bohinc. Tehnična analiza delnice. Master's thesis, Univerza na Primorskem, Fakulteta za management Koper, 2008.

[13] R. Butterworth, G. Piatetsky-Shapiro, and D. A. Simovici. On Feature Selection through Clustering. In *In Proceedings of the Fifth IEEE international Conference on Data Mining*, pages 581–584, 2005.

[14] G. Caginalp and D. Balenovich. A Theoretical Foundation for Technical Analysis. *JOURNAL of Technical Analysis*, 59:5–22, 2003.

[15] B. Caputo, K. Sim, F. Furesjo, and A. Smola. Appearance-based object recognition using SVMs: which kernel should I use? In Whistler, editor, *Proceedings of the Proc of NIPS workshop on*

Statistical methods for com- putational experiments in visual processing and computer vision, 2002.

[16] An-Sing Chen, Mark T. Leung, and Hazem Daouk. Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index. *Computers & Operations Research*, 30:901–923, 2003.

[17] Y-W. Chen and C.-J. Lin. Combining svms with various feature selection strategies. In *Feature extraction*, pages 315–324. Springer, 2006.

[18] M. Dash, H. Liu, and H. Motoda. Consistency Based Feature Selection. *Knowledge Discovery and Data Mining*, 1805:98–109, 2000.

[19] E. Diday. Optimization in nonhierarchical clustering. *Pattern Recognition*, 6:17–33, 1974.

[20] E. Diday. *Optimisation en classification automatique*. Rocquencourt: INRIA, 1979.

[21] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. In *Proceedings of the IEEE Conference on Computational Systems Bioinformatics*, pages 523–528, 2003.

[22] S. Dragonja. Obvladovanje tržnega tveganja delniškega portfelja. Technical report, Univerza v Ljubljani Ekonomska fakulteta, 2008.

[23] Richard O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2nd edition edition, 2000.

[24] A. Esfahanipour and W. Aghamiri. Adapted Neuro-Fuzzy Inference System on direct approach TSK fuzzy rule base for stock market analysis. *Expert Systems with Applications*, 37(7):4742–4748, 2010.

[25] B. S. Everitt. *Cluster analysis*. London: Heinemann Educational Books, 1974.

[26] A. Ferligoj. Razvrščanje v skupine. *Zbirka metodološki zvezki*, 4:18–19; 25–28, 2003.

[27] A. Ferligoj and V. Batagelj. *Taksonomske metode v družboslovnem raziskovanju*. RSS, 1980.

[28] M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3):381–396, 2002.

[29] E. W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classification. *Biometrics*, 21:768–769., 1965.

[30] T. S. Furey, N. Christianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *BIOINFORMATICS*, 16(102000):906–914, 2000.

[31] T. Gestel, J. A. K. Suykens, D.-E. Baestaend, A. Lambrechts, G. Lanckriet, B. Vandaele, B. Moor, and J. Vandewalle. Financial Time Series Prediction Using Least Squares Support Vector Machines Within the Evidence Framework. *IEEE TRANSACTIONS ON NEURAL NETWORKS*, 12:809–820, 2001.

[32] I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[33] I. Guyon, S. Gunn, Masoud Nikravesh, and Lotfi A. Zadeh. *Feature Extraction Foundations and Applications*. Springer, 2006.

[34] M. A. Hall. *Correlation-based feature selection for machine learning*. PhD thesis, Department of Computer Science, University of Waikato, 1999.

[35] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Elsevier, 2006.

[36] J. A. Hartigan. *Cluster algorithms*. New York: Wiley., 1975.

[37] T.-P. Hong, P.-C. Wang, and Y.-C. Lee. An effective attribute clustering approach for feature selection and replacement. *Cybernetics and Systems: An International Journal*, 40(8):657–669, 2009.

[38] H.-H. Hsu and C.-W. Hsieh. Feature Selection via Correlation Coefficient Clustering. *JOURNAL OF SOFTWARE*, 5(12):1371–1377, December 2010.

[39] P. Hu, C. Vens, B. Verstrynge, and H. Blockeel. Generalizing from Example Clusters. *Computer Science*, 8140:64–78, 2013.

[40] C.-L. Huang and C.-Y. Tsai. A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting. *Expert Systems with Applications*, 36:1529–1539, 2009.

[41] W. Huang, Y. Nakamori, and S.-Y. Wang. Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, 32:2513–2522, 2005.

[42] R. C. Jancey. Multidimensional group analysis. *Austral. J. Botany*, 14:127–130., 1966.

[43] S.-M. Jhou and C.B. Yang. Taiwan stock forecasting with the genetic programming. Master's thesis, Department of Computer Science and Engineering, National Sun Yat-sen University, 2011.

[44] G.H. John, R. Kohavi, and K. Pfleger. Irrelevant feature and the subset selection problem. In W.W. Cohen and H. Hirsh, editors, *Machine Learning: Proceedings of the Eleventh International Conference*, pages 121–129, New Brunswick, N.J., 1994. Rutgers University.

[45] C. Jung and R. Bond. Forecasting UK stock prices. *Applied Financial Economics*, 6:279–286, 1996.

[46] M. Kantardzic. *Data Mining - Concepts, Models, Methods, and Algorithms*. IEEE Press, Wiley-Interscience, 2003.

[47] L. J. Kao, C. C. Chiu, C. J. Lu, and C. H. Chang. A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting. *Decision Support Systems*, 54:1228–1244, 2013.

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

[49] A. Karatzoglou, A. Smola, and K. Hornik. Kernlab An S4 Package for Kernel Methods in R. http://cran.r-project.org/web/packages/kernlab/vignettes/kernlab.pdf, 2010. (20.06.2015).

[50] S. S. Keerthi and C.-J. Lin. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation*, 15:1667–1689, 2003.

[51] K. Kim and I. Han. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications*, 19:125–132, 2000.

[52] K. J. Kim. Financial time series forecasting using support vector machines. *Neurocomputing*, 55:307–319, 2003.

[53] K. J. Kim. Artificial neural networks with evolutionary instance selection for financial forecasting. *Expert Systems with Applications*, 30:519–526, 2006.

[54] I. Kononenko. Estimating attributes: analysis and extensions of Relief. In *Machine Learning: ECML-94*, pages 171–182. Springer, 1994.

[55] Igor Kononenko. *Strojno učenje*. Fakulteta za računalništvo in informatiko, 2005.

[56] C. Krier, D. Francois, F. Rossi, and M. Verleysen. Feature clustering and mutual information for the selection of variables in spectral data. In *European Symposium on Artificial Neural Networks*, pages 157–162, Bruges (Belgium), April 2007.

[57] J. B. Kruskal. Multidimensional scaling. *Psychometrika*, 2:1–27 and 115–129., 1964.

[58] V. Labatut and H. Cherifi. Evaluation of performance measure for classifiers comparison. In *Special Issue of ICIT 2011*, pages 21–34, 2011.

[59] M.-C. Lee. Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications*, 36:10896–10904, 2009.

[60] A. Lendasse, E. de Bodt, Wertz V, and M. Verleysen. Non-linear financial time series forecasting —Application to the Bel 20 stock market index. *European Journal of Economic and Social Systems*, 14:81–91, 2000.

[61] J. Li and H. Zha. Simultaneous Classification and Feature Clustering Using Discriminant Vector Quantization with Applications to Microarray Data Analysis. In *Bioinformatics Conference, 2002. Proceedings. IEEE Computer Society*, pages 246–255. IEEE, 2002.

[62] H. Liu, X. Wu, and S. Zhang. Feature Selection using Hierarchical Feature Clustering. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 979–984. ACM, 2011.

[63] C. J. Lu, T. S. Lee, and C.C. Chiu. Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems*, 47:115–125, 2009.

[64] J. B. MacQueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of 5th Berkley Symposium*, volume 1, pages 281–297, 1967.

[65] L. C. Martinez, D. N. da Hora, J. R. de M. Palotti, M. Jr. Wagner, and G. L. Pappa. From an Artificial Neural Network to a Stock Market Day-Trading System: A Case Study on the BM&F BOVESPA. In *In Proceedings of the Proceedings of International Joint Conference on Neural Networks*, pages 2006–2013, 2008.

[66] H. J. Mettenheim and M. H. Breitner. Forecasting Daily Highs and Lows of Liquid Assets with Neural Networks. *In Proceedings of the Operations Research Proceedings 2012*, 2012.

[67] A. Milton. Adjusting the Time Frame. http://daytrading.about.com/od/tradingsystems/a/AdjustTimeFrame.htm, 2014. (20. 06. 2015).

[68] R. Mojena. Hierarchical grouping methods and stopping rules: An evaluation. *Computer Journal*, 20(4):359–363, 1977.

[69] D. Mramor. *Uvod v poslovne finance*. Gospodarski Vestnik, 1993.

[70] M. G. Novak and D. Velušček. Feature selection using k–means. 2015. še neobjavljen članek.

[71] M.G. Novak and Dejan Velušček. Prediction of stock price movement based on daily high prices. *Quantitative Finance*, 2015.

[72] H. Peng and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1226–1238, 2005.

[73] R. Pičulin. Tehnična analiza sekundarnega trga vrednostnih papirjev. Technical report, Fakulteta za organizacijske vede, Univerza v Mariboru, 2006.

[74] L. L. Mc Quitty. Hierarchical linkage analysis for the isolation of types. *Educ. Psychol. Measur.*, 20:55–67, 1960.

[75] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[76] Frank K. Reilly and Keith C. Brown. *Analysis of Investments & Manegements of Portfolios*. South-Western, 2012.

[77] Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 19:2507–2517, 2007.

[78] S. Salvador and P. Chan. Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pages 576 – 584, 2004.

[79] William F. Sharpe. The Sharpe Ratio. *The Journal of Portfolio Management*, 21:49–58, 1994.

[80] John Shawe-Taylor and N. Christianini. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2010.

[81] R. N. Shepard. The analysis of proximities multidimensional scaling with an unknown distance function. *Psychometrika*, pages 125–139 and 219–246, 1962.

[82] P. H. A. Sneath. The application of computers to taxonomy. *J. Gen. Microbiol.*, 17:201–226, 1957.

[83] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature Selection via Dependence Maximization. *Journal of Machine Learning Research*, 13:1393–1434, 2012.

[84] M. V. Subha and S. T. Nambi. Classification of Stock Index movement using k–Nearest Neighbours (k–NN) algorithm. *WSEAS TRANSACTIONS on INFORMATION SCIENCE and APPLICATIONS*, 9:261–270, 2012.

[85] F. E. H. Tay and L. Cao. Application of support Vector machines in Financial time series forecasting. *Omega*, 29:309–317, 2001.

[86] E. ter Horst, A. Rodriguez, H. Gzyl, and G. Molina. Stochastic volatility model indcluding open, close, high and low prices. *Quantitative Finance*, 12:199–212, 2012. http://arxiv.org/abs/0901.1315 (20.06.2015).

[87] M. Tomanič and D. Zbašnik. Tehnična in temeljna analiza vrednostnih papirjev (microsoft corporation). Master's thesis, Ekonomska poslovna fakulteta, 2005.

[88] B. Tominc. Tehnična analiza delnice. Technical report, Univerza na primorskem, Fakulteta za management KOper, 2008.

[89] Chih-Fong Tsai and Yu-Chieh Hsiao. Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems*, 50:258–269, 2010.

[90] B. Vanstone and G. Finnie. An Emperical Methodology for Developing Stockmarket Trading Systems using Artificial Neural Networks. *Expert Systems with Applications*, 36:6668–6680, 2009.

[91] C. Vens, B. Verstrynge, and H. Blockeel. Semi-supervised clustering with example clusters. In *In Proceedings of the 5th International Conference on Knowledge Discovery and Information Retrieval*, 2013.

[92] Gregor Vollmaier. Tehnična anliza delnic s poudarkom na teoriji Elliotovih valov in Fibonaccijevih številih ter njena praktična uporaba na Frankfurtski borzi. Technical report, Univerza v Mariboru, Ekonomsko poslovna fakulteta Maribor, 2004.

[93] A. Vrh. Primerjava uspešnosti upravljanja vzajemnih skladov z uspešnostjo upravljanja portfeljev v individualnem upravljanju premoženja. Master's thesis, Univerza v Ljubljani Ekonomska fakulteta , 2008.

[94] K. Wagstaff and C. Cardie. Clustering with Instance-level Constraints. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 1103–1110, 2000.

[95] J. H. Ward. Hierarchical grouping to optimize an objective function. *JASA*, 58:236–244., 1963.

[96] G. Weiss. Ocenjevanje pomembnosti atributov iz uspešnosti učnih algoritmov na vzorcih atributnega prostora. Master's thesis, Univerza v Ljubljani; Fakulteta za računalništvo in informatiko; Fakulteta za matematiko in fiziko, 2011.

[97] J. Yao and C. L. Tan. A case study on using neural networks to perform technical forecasting of forex. *Neurocomputing*, 34:79–98, 2000.

[98] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.

[99] Lean Yu, Shouyang Wang, and Kin Keung Lai. Mining Stock Market Tendency Using GA-Based Support Vector Machines. *Lecture Notes in Computer Science*, 3828:336–345, 2005.

[100] Lei Yu and Huan Liu. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In *Proceedings of The Twentieth International Conference on Machine Learning (ICML-03)*, pages 856–863, Washington, D.C., August 21-24 2003.

[101] A. Kane Z. Bodie and A. Marcus. *Investments*. New York: McGraw Hill/Irwin, 2005.

[102] C. K. Zhang and H. Hu. An effective feature selection scheme via genetic algorithm using mutual information. In *Fuzzy Systems and Knowledge Discovery*, pages 73–80. Springer, 2005.

[103] Zheng Zhao, Fred Morstatter, Shashvata Sharma, Salem Alelyani, Aneeth Anand, and Huan Liu. Advancing feature Selection Research - ASU Feature Selection Repository. Technical report, 2010.

[104] B. Zvi and R. C. Merton. *Finance*. Prentice Hall, 2000.