PD.TVZ.HR
POLYTECHNIC & DESIGN Vol. 8, No. 1, 2020.
DOI: 10.19279/TVZ.PD.2020-8-1-05

PRIMJENA ALGORITAMA DUBINSKE ANALIZE PODATAKA I STROJNOG UČENJA ZA KLASIFIKACIJU I PREDIKCIJU U DRUŠTVENOM PODRUČJU

APPLICATION OF DATA MINING AND MACHINE LEARNING ALGORITHMS FOR CLASSIFICATION AND PREDICTION IN THE SOCIAL AREA

Sanja Kraljević, Ognjen Staničić

Tehničko veleučilište u Zagrebu, Vrbik 8, Zagreb, Hrvatska

SAŽETAK

Ovaj rad se bavi otkrivanjem znanja primjenom postupaka strojnog učenja iz baze podataka web aplikacije iz društvenog područja. Izvršena je dubinska analiza podataka pomoću četiri različite metode na skupu podataka vezanih uz prijave i pohađanje poslovnih događaja organiziranih od strane jedne tvrtke u razdoblju od tri godine. Primarni cilj rada je otkrivanje znanja iz postojećeg skupa podataka. Pokazano je kako su se deskriptivnom dubinskom analizom otkrila nova znanja. Za analizu su korištene metode stabla odlučivanja, slučajne šume, indukcija pravila te metoda potpornih vektora. Proučene su razlike između navedenih metodologija, te prednosti i mane svake od njih s naglaskom na točnost i utrošak vremena. Osim toga, u radu je opisan i proces pripreme podataka koji je vremenski najzahtjevniji, te sami brojčani rezultati dobiveni prediktivnom analizom.

Ključne riječi: *otkrivanje znanja, dubinska analiza podataka, umjetna inteligencija, strojno učenje, klasifikacija, predikcija, stablo odlučivanja, slučajne šume, indukcija pravila, metoda potpornih vektora, događaj*

ABSTRACT

In this paper knowledge discovery is analyzed using machine learning methods from a database of a web application from the social field. Data mining is performed using four differend methods on a database linked to registrations and attendance of business events organized by a company in a three year span. The primary objective of the paper is to discover knowledge from an existing data set. New knowledge was discovered using descriptive deep analysis using the decision tree method, random forests, rule induction and the support vector machine. Differences between methods are studied, advantages and flaws with emphasis on precision and time consumption. The process of data preparation, which takes the most time, and numerical results yielded in the predictive analysis are also shown in this paper.

Keywords: *knowledge discovery, dana mining, artificial intelligence, machine learning, classification, prediction, decision tree, random forests, rule induction, support vector machine, event*

1. UVOD
1. INTRODUCTION

Marketinške kampanje usmjerene na određene profile korisnika ciljano povećavaju profit organizacije i uspješnost kampanje. Poslovni događaji čija je svrha predstavljanje tehnologija i proizvoda tvrtke, predstavljanje novih znanja, ideja, poslovnih strategija, međusobno upoznavanje poslovnih suradnika ili partnera ili samo poboljšanje odnosa unutar tvrtke, uspješni su ako ostvare zadani cilj. Tvrtke individualno, u međusobnoj suradnji ili uz potporu sponzora ulažu u organizaciju takvih događaja. Potencijalne sudionike pozivaju na događaje raznim kanalima, a u radu će fokus biti isključivo na internet kampanje te pozive e-mailom.

Detektirani su različiti profili korisnika, od onih koji nikada ne otvaraju takve e-mailove, korisnici koji pročitaju e-mail, ali se ne prijave na događaj te korisnici koji se prijave na događaj te u konačnici na njega dođu ili ne dođu. U radu će biti analizirani slučajevi kada potencijalni sudionici potvrde dolazak na događaj, te na njega ne dođu. Pravilima klasifikacije će se korisnici pridruživati jednoj kategoriji unutar konačnog skupa kategorija, te pravilima predikcije će se temeljem značajki prošlih slučajeva predviđati buduće vrijednosti odnosno slučaj dolaska ili nedolaska na događaj.

2. KORIŠTENA METODOLOGIJA
2. METHODOLOGY

Dubinskom analizom podataka (engl. *data mining*) naziva se primjena postupaka i alata čija je glavna namjena otkrivanje i izlučivanje znanja iz skupova podataka. Ona je usko vezana uz statističke pojmove i postupke. Ako je potrebno statistički provjeriti neku hipotezu koja je dobivena dubinskom analizom, tada je potrebno prikupiti nezavisni skup podataka primjeren za statističko testiranje te hipoteze. Analiza može biti **prediktivna** ili **deskriptivna**.

Proces dubinske analize podataka uključuje pripremu podataka, izvođenje postupaka analize te samu interpretaciju rezultata. Iznimno je bitno sudjelovanje čovjeka odnosno eksperta za određeno područje u fazi pripreme podataka i fazi analize rezultata. On mora izvrsno razumjeti ulazne podatake te tumačiti dobivene rezultate.

Glavno svojstvo koje analizu podataka čini dubinskom je primjena postupaka strojnog učenja. **Strojno učenje** (engl. *machine learning*) je dio područja računarstva poznatog kao **umjetna inteligencija** (engl. *artificial intelligence*). Umjetna inteligencija se bavi razvojem postupaka koji pomoću računala simuliraju ljudsko odnosno inteligentno ponašanje, a strojno se učenje bavi podskupom tih postupaka koje karakterizira to da ima mogućnost učenja na osnovi prethodnog iskustva. Uspješnost postupaka strojnog učenja ocjenjuje sa na osnovu prediktivne točnosti izrađenih modela nad skupom podataka koji nije korišten u procesu učenja.

Inicijalni skup podataka na početku se dijeli na skup za učenje te skup za testiranje.

Primjena postupaka strojnog učenja u analizi podataka može rezultirati u formiranju novog znanja. **Otkrivanje znanja** (engl. *knowledge discovery*) je proces koji nastaje integracijom postupaka strojnog učenja i ekspertne analize njegovih rezultata s ciljem otkrivanja i formuliranja novog znanja. Za razliku od strojnog učenja, kojemu je postizanje visoke prediktivne kvalitete primarni cilj, otkrivanje znanja prvenstveno se bavi razvojem i primjenom postupaka strojnog učenja koji sustavno pretražuju skup mogućih modela te rezultate prikazuju u obliku koji je pogodan za ljudsko razumijevanje. [7]

Plan istraživanja provedenog u sklopu rada prati korake ***Cross Industry Standard Process for Data Mining* (CRISP/DM) metodologije**. Ovo je jedna od najpopularnijih metoda dubinske analize podataka pri rješavanju određenih problema. Proces se odvija ciklički, te su promjene smjerova moguće i poželjne jer se će promatranjem rezultata međufaza poboljšati rezultati cjelokupne analize. S obzirom na to da je poslovna domena i struktura podataka bila unaprijed poznata, najveći dio vremena (70 – 80%) uložen je u pripremu podataka za analizirane slučajeve.

Za analizu podataka te primjenu metoda strojnog učenja korišten je **RapidMiner**, za analizu strukture i ključeva relacijske baze podataka te izvršavanje složenijih SQL upita korišten je **MySQL Workbench** te su Visual Studio Code i Excel korišteni kao pomoćni alati za regularne izraze i dodatnu provjeru formata podataka.

3. CILJEVI RADA I PRIKAZ POSLOVNE DOMENE
3. OBJECTIVES AND BUSINESS UNDERSTANDING

Podatci se nalaze u relacijskoj bazi podataka koja je dio web aplikacije, a prikupljeni u razdoblju od šest godina, 03/2013 – 10/2019, od čega su se radi verzioniranja aplikacije za ovo istraživanje smatrali relevantnima samo oni prikupljeni u posljednje tri godine.

Aplikacija je korištena za organiziranje događaja iz područja informacijskih i telekomunikacijskih tehnologija.

Relacijska baza podataka sastoji se od **41 entiteta** međusobno povezanih pomoću primarnih i stranih ključeva. Navedeni entiteti opisani su sa ukupno **372 atributa** i ukupno **82 ključa**, a svi entiteti sadrže ukupno **70634 retka (n-torke)**. Promatrana instanca aplikacije koristi se periodično, za nekoliko poslovnih događaja godišnje, a u radu će biti pokazano da je i to sasvim dovoljan skup podataka kako bi se detektirala ponašanja korisnika u predviđenom scenariju.

Cilj koji je postavljen u poslovnom kontekstu aplikacije je detekrirati profile korisnika koji se prijavljuju na događaje no ipak na njemu ne prisustvuju te koje su vrijednosti ključnih atributa koji će te profile opisati. Za navedeni slučaj korištene su tablice *leads* (26 atributa, 5207 n-torki), *events* (46 atributa, 25 n-torki) te *registrations* (36 atributa, 3898 n-torki), pri čemu je skup podataka dodatno reduciran potrebnim kriterijima u fazi pripreme podataka.

4. PRIPREMA PODATAKA
4. DATA PREPARATION

4.1. SELEKCIJA PODATAKA
4.1. DATA SELECTION

Iz baze podataka metodom SQL **selekcije** i **projekcije** preuzeti su podatci. Izvršena je **dimenzijska redukcija podataka**, odnosno **uklanjanje nepotrebnih i koreliranih atributa** izvršeno je odabirom kolona koje se upotrebljavaju. Analizirani su podatci vezani uz događaje koji su se održavali u Hrvatskoj, na nekoj fizičkoj lokaciji.

U nastavku su opisani entiteti iz relacijske baze podataka koji su korištene za analizu.

U postupku odabira značajki, iz početnog skupa od 108 atributa u ukupno 3 tablice, izbačeni su atributi koji su u svojoj koloni imali malo vrijednosti, atributi koji su predstavljali identifikator, atributi koji su predstavljali dupliranje drugog atributa te oni koji su korelirani međusobno, a nisu korelirani sa ciljnim atributom jer su smatrani irelevantnima. Također nisu razmatrane vremenske oznake koje se odnose na podrijetlo zapisa. Napravljena je usporedba rezultata sa tim skupom atributa i sa ciljano izdvojenih šest atributa pogodnih za definiranje korisnika i daljnjih marketinških kampanja, te su razlike u dobivenim rezultatima bile minimalne.

Prilikom odabira atributa, birani su ciljano oni atributi koji se u organizaciji ovakvih vrsta događaja od strane organizatora mogu prilagođavati, čime bi se dobilo na većoj posjećenosti događaja. Cilj dakle nije bio dobiti atribute koji pridonose učinkovitosti klasifikacije i predikcije već učinkovitosti daljnjih poslovnih strategija i marketinških poteza. Promatrano je na koji način dotični atributi utječu na klasifikaciju. Izbjegavani su koorelirani atributi koji bi definirali iste značajke, kombinacija više atributa za definiranje jedne značajke nije bila relevantna. Irelevantni atributi usporavali su postupke analize i nisu doprinosili deskriptivnoj analizi, te su rezultati analize u radu komentirani na idućih šest atributa:

- *gender* – spol korisnika
- *job_function_generated* – poslovna uloga korisnika (radno mjesto)
- *domestic_lead* – radi li se o korisniku koji stanuje u mjestu održavanja eventa
- *event_type* – vrsta događaja
- *length_in_days* - trajanje događaja u danima
- *event_attended_flag* – oznaka je li korisnik uistinu prisustvovao događaju nakon što se na njega prijavio

Tablica 1. *Kvantitativni opis analiziranih entiteta baze podataka*

Table 1. *Quantitative description of analized database entities*

Entitet	Broj atributa	Broj n-torki	Broj stranih ključeva
leads	26	5207	0
events	46	35	9
registrations	36	3898	7

Provjere radi, po završetku istraživanja dodatno je upotrebljen 1-R (engl. *One Rule*) algoritam za odabir značajki. Algoritam 1-R odabrao je 27 atributa, a među tih 27 atributa nalazi se svih šest koji odabrabi ručnim odabirom sa ciljem otkrivanja znanja, a prije upotrebe 1-R algoritma.

Zapisi su analizirani za vrijednost atributa *event_attended_flag,* čija je vrijednost 1 ako je korisnik došao, odnosno 0 ako korisnik nije došao na događaj.

4.2. ČIŠĆENJE PODATAKA
4.2. DATA CLEANING

Detektirani su **nepripadajući podatci** (engl. *outliers*) te su oni uklonjeni selekcijom u SQL upitu korištenjem WHERE klauzule. Primjer takvih podataka su događaji za koje nije postojala potpuna informacija o evidenciji dolazaka ili su otkazani. Iz takvih bi se podataka došlo do pogrešnog zaključka da se radi o nedolascima (primjerice stopa nedolaznosti od 100%).

4.3. KONSTRUIRANJE NOVIH PODATAKA I NEDOSTAJUĆE VRIJEDNOSTI
4.3. CONSTRUCTING NEW DATA AND MISSING VALUES

Vrijednosti koje su nedostajale nisu nadomještane radi brojnih nedostataka metoda koje se preporučaju za njihovo konstruiranje. Umjesto toga, vrijednosti koje nedostaju izuzete su iz analize ili su korištene kao kriterij.

Od ukupno 5207 korisnika, za njih 4212 je poznat spol. Od toga je 733 ženskih i 3478 muških korisnika, odnosno muškarci čine 83% ukupne promatrane populacije. Ovo nije neočekivana statistika upravo zbog podrijetla podataka, s obzirom da se radi o aplikaciji koja je korištena za organiziranje događaja iz područja ICT-a, pa su populacija većinom muški inženjeri.

Postupkom **normalizacije** podataka stvoren je stupac *job_function_generated* na temelju vrijednosti atributa *job_function* iz tablice *leads*.

Podatak se u aplikaciju unosi kroz web obrazac korištenjem polja za tekstualni unos `<input type='text'>`, što je rezultiralo sa 892 različite vrijednosti za atribut *job_function* te je kao takav bio neupotrijebiv. Vrijednosti su normalizirane ekspertnim pristupom na temelju ključnih riječi korištenjem regularnih izraza. Regularni izrazi su pisani dovoljno precizno sa ciljem da obuhvate sve vrijednosti gdje barem tri n-torke sadrže istu ključnu riječ. One ključne riječi koje su se pojavljivale u manje od tri n-torke, proglašene su skupinom „ostalo“. Normalizirane vrijednosti su „CEO“ (157), „voditelj“ (137), „inženjer“ (658), „prodaja“ (113), „ostalo“ (393) i a ostalo su nedostajuće vrijednosti.

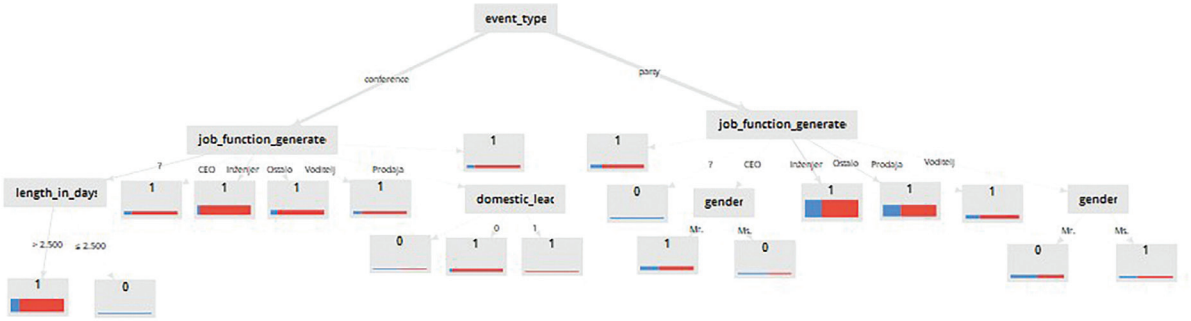
4.4. FORMATIRANJE PODATAKA
4.4. DATA FORMATTING

Podatci su uvezeni u RapidMiner u .csv formatu kao rezultat SQL upita. Prethodno su nad bazom postavljeni SQL upiti kako bi se dobile samo one n-torke koje se odnose na promatrane događaje opisane u poglavlju 4.1. Formatiranje podataka odrađeno je radi unificiranja *real* i *integer* vrijednosti koje su drugačije od očekivanih kada se dobiju pročišćeni skupovi podataka korištenjem modula RapidMiner alata, naziva Turbo Prep.

5. PROVOĐENJE ANALIZE
5. PERFORMING THE ANALYSIS

5.1. STABLA ODLUČIVANJA
5.1. DECISION TREE

Stabla odlučivanja (engl. *decision tree***)** su alat za aproksimaciju funkcija čija je osnovna ideja problem rekurzivno dijeliti u dva podproblema, od kojih će svaki biti rješiv metodom smanjene složenosti. Dioba problema izvršava se u unutarnjim čvorovima stabla na temelju binarnih testova za radni skup podataka. Listovi stabla sadrže jednostavne modele koji aproksimiraju traženu funkciju. Odabir testova u čvorovima i modela u listovima ovise o odabranom obliku stabla i njegovim parametrima.



Slika 1 *Stablo odlučivanja A – skup korisnika na događajima, bez obzira znamo li njihov spol ili ne*

Figure 1 *Decision tree A – set of users attenting events without regards to their gender*

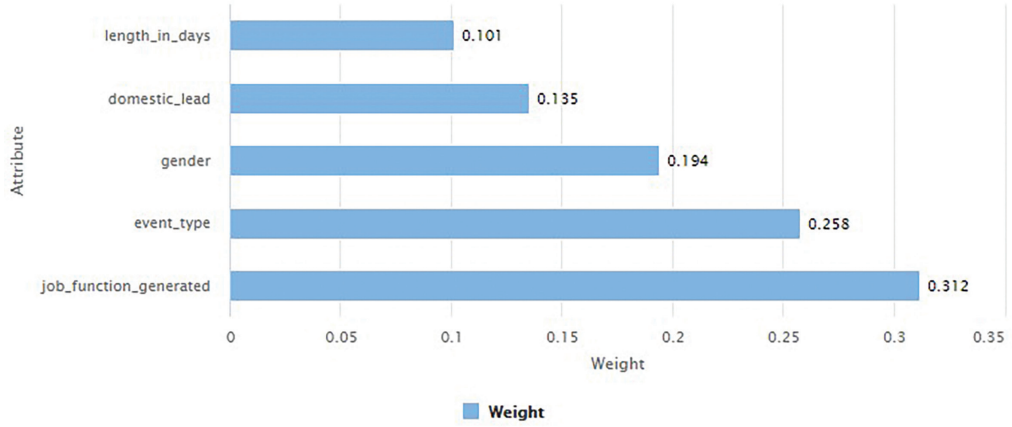
Grana pojedinog atributa koja završava samo sa jednom vrstom klasifikacije (0 ili 1) je poželjna jer na njoj nije potrebno dalje razvijati stablo. Postavljen je proces za model stabla odlučivanja sa potrebnim prosecima.

Skup podataka analiziran je metodom stabla odlučivanja, a kao mjera za razdvajanje korišten je **informacijski dobitak** (engl. *information gain*) i **indeks gini nečistoće** (engl. *gini impurity*). Informacijski dobitak radi razdvajanje na principu najvećeg otkrivanja informacija a razdvajanje po principu gini nečistoće razdvaja sa ciljem dobivanja što ujednačenijih vrijednosti u krajnjim čvorovima. Kako bi se spriječilo **pretreniranje**, na stablo je primijenjen postupak **podrezivanja** (engl. *pruning*). Prilikom podrezivanja, stablom se induciraju pravila za razdvajanje te se računaju uvjeti kojima se postiže najveća točnost klasifikacije, a na način da se izbacuju podstabla koja ne poboljšavaju točnost klasifikacije.

Rezultati su prikazani za vrijednost atributa *event_attended_flag*, čija je vrijednost 1 ako je korisnik došao, odnosno 0 ako korisnik nije došao na događaj.

Na slici **Slika 1**, prikazani su rezultati metode informacijskog dobitka, najveće dubine 10, pouzdanosti u pesimističnim slučajevima kod podrezivanja 0.1, najmanjeg dobitka 0.01, najmanje veličina čvora 2, najmanje veličina za dijeljenje 4, s brojem alternativa za podrezivanje 3.

```
event_type = conference
|   job_function_generated = : 1
{0=15, 1=87}
|   job_function_generated = ?
|   |   length_in_days > 2.500: 1
{0=91, 1=460}
|   |   length_in_days ≤ 2.500: 0
{0=4, 1=0}
|   job_function_generated = CEO: 1
{0=18, 1=110}
|   job_function_generated =
Inženjer: 1 {0=17, 1=367}
|   job_function_generated = Ostalo:
1 {0=22, 1=155}
|   job_function_generated = Prodaja
|   |   domestic_lead = : 0 {0=1,
1=1}
|   |   domestic_lead = 0: 1 {0=5,
1=90}
|   |   domestic_lead = 1: 1 {0=0,
1=3}
|   job_function_generated =
Voditelj: 1 {0=14, 1=78}
event_type = party
|   job_function_generated = : 1
{0=27, 1=99}
|   job_function_generated = ?: 0
{0=31, 1=0}
|   job_function_generated = CEO
|   |   gender = Mr.: 1 {0=47, 1=87}
|   |   gender = Ms.: 0 {0=13, 1=10}
|   job_function_generated =
Inženjer: 1 {0=230, 1=509}
|   job_function_generated = Ostalo:
1 {0=157, 1=319}
|   job_function_generated = Prodaja:
1 {0=34, 1=97}
|   job_function_generated = Voditelj
|   |   gender = Mr.: 0 {0=54, 1=54}
|   |   gender = Ms.: 1 {0=19, 1=38}
```



Slika 2 *Težinski koeficijenti atributa za stablo odlučivanja A*

Figure 2 *Weight coefficients of attributes for decision tree A*

Ostala stabla odlučivanja izvedena je u nekoliko verzija uz izmjene parametara sa ciljem dobivanja opisa profila korisnika iz ciljane skupine. Ona neće biti detaljno komentirana, već se ovdje iznosi dio rezultata. Obilježja nekih skupina korisnika koji koje neće doći na događaj - **broj korisnika koji neće doći, broj korisnika koji će doći, postotak od ukupne populacije**:

– ženski, CEO, party, 13, 10, 0.83%
– muški, voditelj, party, 54, 54, 3.88%
– party, osobe koje nisu navele radno mjesto, 7, 1, 0.25%

5.2. SLUČAJNE ŠUME
5.2. RANDOM FOREST

Slučajne šume (engl. *random forest*) je klasifikator koji se sastoji od kolekcije nezavisnih stabala odlučivanja, pri čemu svako stablo **predstavlja jedan glas** u većinskom donošenju odluke. Metoda radi tako da skup za učenje podijeli na podskupove s jednakim brojem primjera, nakon čega se od svakog podskupa napravi stablo odlučivanja, a cilj je smanjivanje varijance konačnog modela. Prednost ove metode za izradu stabla odlučivanja je u tome što nije potrebna međuvalidacija, slučajne šume se ne prilagođuju podacima za učenje. Prilikom podrezivanja grana, svako od pet stabala rezultiralo je sa različitim dubinama stabla, bez obzira što su sva imala iste postavke. Rezultat ovisi o informacijskom dobitku daljnjih grananja.

Slučajne šume dale su slučajna stabla koja su promatrana sa ciljem detektiranja vrijednosti atributa za otkrivanje znanja o tome koji korisnici neće doći na događaj.

Obzirom da pojedinačni rezultati nisu relevantni, ovdje neće biti detaljno analizirani. Dobivena stabla potvrđuju statistiku a to je da su najčešći uzorak osobe zaposlene kao inženjeri koji se prijavljuju na konferencije tipa party – oni će doći, čine 23.39% uzoraka, odnosno 220 neće doći a 432 hoće.

5.3. INDUKCIJA PRAVILA
5.3. RULE INDUCTION

Indukcija pravila (engl. *rule induction*) je metoda koja pronalazi opis pravilima koja uključuju što više primjera jednog razreda uz što manje primjera ostalih razreda. Korišten je algoritam RIPPER. Prilikom analize promatrano je više međurezultata sa ciljem dobivanja raznolikijih skupina i pravila sa više krajnjih grupa korisnika, uz dodatne filtere na vrijednostima značajki ili isključivanjem pojedinih.

U nastavku je prikazan međurezultat dobiven nad skupom n-torki sa poznatim atributom *job_function_generated* te isključenim atributom *domestic*.

1. if event_type = conference then 1 (92 / 891)
2. if job_function_generated = and gender = Ms. then 1 (5 / 24)
3. if job_function_generated = Inženjer and gender = Mr. then 1 (196 / 455)
4. if job_function_generated = Prodaja and gender = Ms. then 1 (11 / 43)

5. if job_function_generated = Ostalo and gender = Mr. then 1 (109 / 232)

6. if job_function_generated = and gender = Mr. then 1 (22 / 75)

7. if job_function_generated = Prodaja and gender = Mr. then 1 (23 / 54)

8. if job_function_generated = Ostalo and gender = Ms. then 1 (47 / 86)

9. if job_function_generated = CEO and gender = Mr. then 1 (47 / 87)

10. if gender = Ms. and job_function_generated = Voditelj then 1 (19 / 38)

11. if job_function_generated = Inženjer and gender = Ms. then 1 (34 / 54)

12. **if job_function_generated = CEO and gender = Ms. then 0 (13 / 10)**

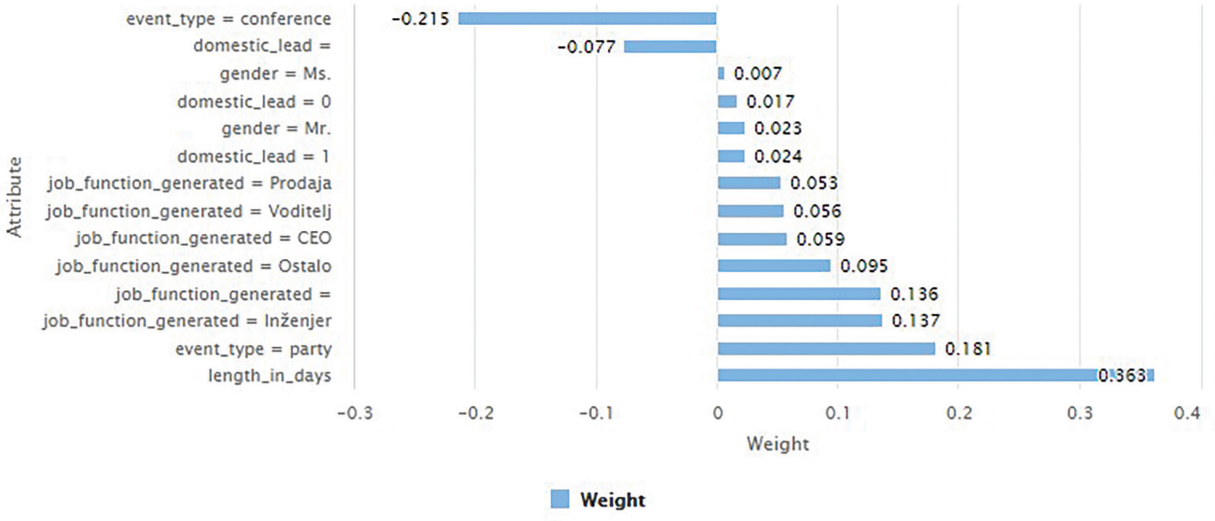
13. if gender = Mr. and job_function_generated = Voditelj then 1 (54 / 54)

14. **if job_function_generated = Ostalo then 0 (1 / 1)**

5.4. METODA POTPORNIH VEKTORA
5.4. SUPPORT VECTOR MACHINE

Metoda potpornih vektora (engl. *Support Vector Machine - SVM*) predstavlja algoritam čija je zadaća pronaći najveću marginu razdvajanja između klasa. Za marginu podrazumijevamo udaljenost koja obuhvaća rubne primjere, odnosno primjere koji bi bili krivo klasificirani da se ne koristi marginu već samo plohu razdvajanja. Metoda potpornih vektora poznata je i kao metoda maksimalno granične hiperravnine jer je njezin cilj nalaženje hiperravnine koja razdvaja hiperprostor primjera u dva poluprostora koji odgovaraju dvjema klasama podataka tako da je udaljenost te hiperravnine od najbližih točaka podataka maksimalna.

Za navedenu metodu je karakteristično da radi isključivo sa numeričkim vrijednostima, zbog čega su sve nominalne vrijednosti u skupu podataka pretvorene u numeričke. Zbog navedene transformacije, koeficijenti pojedinih značajki neće se odnositi samo na određeni atribut već na kombinaciju atribut i njegova vrijednost za nenumeričke vrijednosti, odnosno samo atribut za numeričke vrijednosti. Korišten je algoritam mySVM Stefana Ruperinga. Korištena je kernel jezgra sa cache parametrom podešenim na 200MB.



Slika 3 *Težinski koeficijenti za atribute - metoda potpornih vektora*

Figure 3 *Weight coefficients of attributes for support vector machine*

Tablica 2. *Usporedba rezultata predikcije*

Table 2. *Prediction results comparison*

	Apsolutna pogreška	**Standardna devijacija**	**Vrijeme [ms]**	**Vrijeme za treniranje na 1000 redaka [ms]**	**Vrijeme za predikciju na 1000 redaka [ms]**
Stabla odlučivanja	33.83%	0.01231	670.0	2.38	5.95
Slučajne šume	33.03%	0.01224	16547.0	39.84	89.22
Metode potpornih vektora	24.63%	0.01891	242778.0	3528.69	226.77

6. USPOREDBA I DISKUSIJA REZULTATA
6. RESULTS COMPARISON AND DISCUSSION

U tablici **Tablica 2** prikazana je usporedba rezultata za prediktivne modele. Iznos apsolutne pogreške kod stabla odlučivanja i slučajne šume je gotovo identičan, uz minimalna poboljšanja kod slučajnih šuma, no uz znatno veći utrošak vremena, i to od ~25 puta dulje vrijeme od čega je većina potrošena na predikciju. Općenito u odnosu na druge algoritme, stabla odlučivanja izuzetno brzo daju rezultat, no sa većom pogreškom. Druga krajnost su metode potpornih vektora koje su dale uz znatno veći utrošak vremena, ~362 puta dulje u odnosu na stablo odlučivanja. No zahvaljujući tome, uistinu su bolji rezultati gdje se iznos apsolutne pogreške sa 33.83% kod stabla odlučivanja spustila na 24.63%.

U istoj tablici se ne nalaze rezultati metode indukcije pravila jer se taj algoritam koristi za definiranje deskriptivnih pravila. Kod deskriptivnih pravila uočene su neočekivane povezanosti, a one su komentirane u poglavljima sa međurezultatima.

Obzirom na visoku apsolutnu pogrešku, dodatno je primijenjen algotiram **1-R** (engl. *One Rule*) za automatski odabir značajki umjesto ručnog ograničavanja na njih šest. Navedeno nije prednost kod deskriptivne analize jer su se promatrali ciljani atributi, ali jest kod prediktivne, no važnija razlika pokazala se samo u kombinaciji s metodom potpornih vektora, kada se apsolutna pogreška smanjila za 1.53% i iznosila 23.1%.

7. ZAKLJUČAK
7. CONCLUSION

Deskriptivna pravila dobivena metodama stabla odlučivanja, slučajnih šuma i metode indukcije pravila, rezultiraju grupama sličnih vrijednosti značajki i međusobno se potvrđuju, uz pravilo da osim što su se potvrdile, pojedine metode su dale pravila koja nisu bila vidljiva u ostalim promatranim metodama. Mnogo je podešavanja potrebno izvršiti nad odabranim metodama i njihovim parametrima da bi se postigli zadovoljavajući rezultati te otkrila nova znanja. Iako su ta podešavanja vremenski zahtjevna, gotovo su neusporediva sa vremenskom zahtjevnošću u fazi pripreme podataka.

Pravila i opisi grupa koji su dobiveni korištenim metodama stabla odlučivanja, slučajnih šuma i metode indukcije pravila su novi te se mogu koristiti u daljnjim marketinškim kampanjama. Iako se krenulo od 108 atributa u promatranim tablicama, samo njih šest je procjenjeno bilo relevantnim i zanimljivim za poslovne strategije, no u prilikom prediktivne analize preporuča se koristiti algoritme za automatiziran odabir značajki, primjerice 1-R. Dodatna pravila, koja u radu nisu navedena mogu se primijeniti iz ekspertne analize, a to je da je veći udio nedolazaka na događaju tipa party također i iz razloga što se na njima ne naplaćuje sudjelovanje.

Korištenjem metoda stabla odlučivanja, slučajnih šuma i metode potpornih vektora, za predikciju bi rezultati bili bolji kada bi se prikupljali dodatni atributi na temelju kojih bi se poboljšao skup za učenje i skup za testiranje. Navedeno jest očekivana problematika obzirom da se radi o društvenom području jer bi promatranjem sustava u tehničkom području skup značajki za razne sustave bio veći.

Navedeni atributi moraju ujedno biti korisni sa poslovnog aspekta odnosno analizu je potrebno raditi na temelju parametara na koje je moguće utjecati prilikom organizacije događaja.

8. REFERENCE
8. REFERENCES

[1.] Bogunović, N. Materijali za kolegij Otkrivanje znanja u skupovima podataka, Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, 2019. http://www.zemris.fer.hr/predmeti/kdisc/ (5.10.2019.)

[2.] CRISP-DM , CRISP-DM Overview http://www.datascience-pm.com/crisp-dm-2/ (4.11.2019)

[3.] Gabelica, H. Rudarenje podataka i CRISP metodologija, 2013., http://www.skladistenje.com/rudarenje-podataka-i-crisp-metodologija/ (4.11.2019.)

[4.] Gržinić, T. Hibridna metoda otkrivanja zlonamjernih programa, doktorska disertacija, Fakultet organizacije i informacije Varaždin, Sveučilište u Zagrebu, 2017. UDK 004.056.5(043.3)

[5.] Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009. ISBN: 978-0-387-84858-7

[6.] Witten I. H.; Frank E.; Hall M. A. Data Mining, 3rd. Ed., Morgan Kaufmann, 2011. ISBN 978-0-12-374856-0

[7.] Institut Ruđer Bošković, Otkrivanje znanja dubinskom analizom podataka, Priručnik za istraživače i studente, v 1.46, http://lis.irb.hr/Prirucnik/prirucnik-otkrivanje-znanja.pdf. (12.10.2019.)

[8.] Kramberger, T.; Duk S.; Kovačević R. Baze podataka, Manualia Polytechnici Studiorum Zagrebisensis, Tehničko veleučilište u Zagrebu, 2018. ISBN: 978-953-7048-70-9

[9.] RapidMiner documentation, https://docs.rapidminer.com/ (12.02.2020.)

AUTORI · *AUTHORS*

● **Sanja Kraljević -** nepromjenjena biografija nalazi se u časopisu Polytechnic & Design Vol. 7., No. 4, 2019.

Korespondencija · *Correspondence*
sanja@tvz.hr

● **Ognjen Staničić -** nepromjenjena biografija nalazi se u časopisu Polytechnic & Design Vol. 7, No. 4, 2019.

Korespondencija · *Correspondence*
ognjen.stanicic@tvz.hr