XXVI Интернационални научни скуп
Стратегијски менаџмент и системи подршке одлучивању у стратегијском менаџменту

21. мај 2021, Суботица, Република Србија

Раде Божић
Факултет пословне економије Бијељина
Бијељина, Република Српска, Босна и Херцеговина
rade.bozic@fpe.ues.rs.ba

ТАЧНОСТ КЛАСИФИКАЦИОНЕ ТЕХНИКЕ РУДАРЕЊА ПОДАТАКА У ПОСЛОВНИМ ИНФОРМАЦИОНИМ СИСТЕМИМА – ПРЕГЛЕД ЛИТЕРАТУРЕ У ПОСЛЕДЊОЈ ДЕЦЕНИЈИ

Апстракт: Рударење података (енг. *Data mining*) представља концепт помоћу којег се настоје издвојити употребљиве информације из великог скупа података. Примену проналази у многобројним дисциплинама укључујући и област економије, нарочито на микро нивоу. Имплементирање рударења података у пословне информационе системе резултира информацијама које представљају кључ за доношење пословних одлука. На овај начин долази се до потенцијалних прилика које могу да обезбеде конкурентску предност. Колико су тачне технике које се примењују у пословним информационим системима јесте основно питање које уједно представља и истраживачко питање рада? До одговора на постављено истраживачко питање дошло се систематским прегледом литературе, радова који се односе на класификациону технику рударења података. Обухваћени су само модели који су примењиви у свим пословних субјектима без обзира на област пословања. Главни резултати радова из протекле деценије су приказани и протумачени.
Кључне речи: рударење података, пословни информациони системи, класификација, тачност, преглед литературе

ACCURACY OF CLASSIFICATION DATA MINING TECHNIQUES IN BUSINESS INFORMATION SYSTEMS - LITERATURE REVIEW IN THE LAST DECADE

Abstract: Data mining is a concept for extraction usable information from a large set of data. It is applied in many disciplines, including the field of economics, especially at the micro level. Data mining implementation into business information systems results in informations that are key for making business decisions. In this way, potential opportunities that can provide a competitive advantages are created. How accurate are the techniques used in business information systems is a research question of this paper? The answer to this research question was introduced by a systematic literature review of the papers related to the classification techniques of data mining. Only models that are applicable in all business entities, regardless of the area of business, are included. The main results of the papers from the past decade are presented and interpreted.
Key words: data mining, business information systems, classification, accuracy, literature review

1. УВОД

Савремено пословање одликује употреба информационих система и искоришћавање њиховог максималног потенцијала. Разлог томе је глобализација пословања која доводи до непрестане конкурентске борбе захтевајући константну флексибилност. Свака употребљива информација може да створи предност у односу на конкуренцију и допринесе предузећу у остваривању постављених циљева. Рударење података је техника која обезбеђује ову врсту информација, а њихов значај може да буде непроцењив при доношењу пословних одлука.
Рударење података (енгл. *data mining*) представља актуелан концепт у развоју информационих технологија. Дефинише се као издвајање имплицитних, раније непознатих и потенцијално корисних информација из података (Witten et al., 2011). Неки теоретичари поред наведеног користе и појам откривање знања (енгл. *knowledge discovery*). Употреба термина „рударење" почиње од 1980-их година где је ова техника обраде података служила као алат за обављање појединачних задатака. Односила се првенствено на методе класификације (енгл. *classification)* путем стабла одлучивања (енгл. *Decision Tree*, DT) или неуронских мрежа (енгл. *Neural Networks*), кластерисање и визуелизацију података (Piatetsky-Shapiro, 1999). Даљим развојем проналази примену у различитим областима као што су медицина, образовање, метеорологија, производни и софтверски инжењеринг, уметност, биоинформатика, енергетика, саобраћај, као и многе друге (Rahman, 2018). Поред наведених, посебно је актуелна примена у економији и њеним дисциплинама попут менаџмента, финансијске анализе, маркетинга, пословне организације, рачуноводства, итд. Неки од конкретних примера су: откривање превара са кредитним картицама, финансијско предвиђање, дизајнирање производа, процена вредности некретнина, маркетинг таргетирање и сл. (Brammer, 2016). Посебну пажњу привлачи имплементација на микро нивоу, односно у пословним системима и организацијама.
Рударење података у економској сфери предмет је истраживања великог броја научних радова како на микро, тако и на макро нивоу. Поставља се питање колика је тачност информација добијених поменутим техникама? Како би се одговорило на дато питање, спроведен је систематски преглед литературе ради сумирања индивидуалних истраживања и њихових резултата у једну целину. Фокус рада је на тачности класификационе технике примењиве у пословним субјектима без обзира на област пословања. Критеријумима укључивања биће детаљније објашњен начин селекције радова. Поред тачности обухватају се и коришћени модели, као и области пословања у којима се примењују. Наведено подручје прегледа литературе није обрађивано у већем броју случајева, па је услед актуелности самог истраживачког питања погодно за спровођење. На овај начин долази се до сазнања о модерним аспектима развоја пословних информационих система.

2. ТАЧНОСТ КЛАСИФИКАЦИОНЕ ТЕХНИКЕ РУДАРЕЊА ПОДАТАКА

Класификација представља једну од техника рударења података која припада категорији надзираног учења. На основу одређеног дела сета података који служе за тренинг модела (енгл. *training data*), класификатор уочава структуре међу подацима и распоређује их у групе које се називају *категорије* или *класе.* Поред класа често се користи термин лабеле (енгл. *labels*). Ово представља прву фазу у развоју класификатора, након чега следи фаза тестирања модела која се спроводи над остатком сета података који је намењен тестирању (енгл. *testing data*). Овде класификатор одређује класе (лабеле) за претходно изостављене инстанце (Aggrawal, 2019). Након завршетка касификовања извршава се евалуација перформанси која показује успешност модела. Класификација проналази практичну примену код сврставања потрошача у одговарајуће категорије, медицинској дијагностици, пословању банкарских и осталих финансијских институција, метеорологији, криминалистици, као и разним другим областима. Најчешће класификационе технике су стабла одлучивања, *Naïve Bayes* класификатор, неуронске мреже, метод најближих суседа (енгл. *k-Nearest Neighbors,* k-NN), методе потпорних вектора (енгл. *Support Vector Machines,* SVM), итд.
Тачност (енгл. *accuracy*) класификационе технике у рударењу података рачуна се на основу матрице конфузије (енгл. *confusion matrix*) тестног скупа и представља однос тачно класификованих инстанци и укупног броја инстанци. У табели бр. 1. дат је пример матрице конфузије за скуп који има само две класе, позитивну (+) и негативну (-). Матрица ће имати четири ћелије које су означене са тачни позитивни (енгл. *true positives*), тачни негативни (енгл. *true negatives*), нетачни позитивни (енгл. *false positives*) и нетачни негативни (енгл. *false negatives*) (Brammer, 2016).

Табела 1: Матрица конфузије

Стварне класе инстанци	**Предвиђене класе**	
	+	**-**
+	тачни позитивни (ТП)	нетачни негативни (ФН)
-	нетачни позитивни (ФП)	тачни негативни (ТН)

Извор: Brammer, 2016.

- **Тачни позитивни** представља број инстанци са позитивном класом које је модел класификовао као позитивне.

- **Тачни негативни** представља број инстанци са негативном класом које је модел класификовао као негативне.
- **Нетачни позитивни** представља број инстанци са негативном класом које је модел класификовао као позитивне.
- **Нетачни негативни** представља број инстанци са позитивном класом које је модел класификовао као негативне.

Да бисмо израчунали тачност потребно је да утврдимо колико је инстанци наш модел класификовао тачно. То добијемо тако што саберемо тачно позитивне и тачно негативне (позиционирани у дијагонали матрице) (1).

(1) Тачно класификовани = ТП + ТН

Затим је потребно израчунати укупан број тестних инстанци, односно збир свих тачно и нетачно класификованих (2).

(2) Укупан број тестних инстанци = ТП + ТН + ФП + ФН

На крају тачност рачунамо као количник тачно класификованих инстанци и укупног броја тестних инстанци (3) (Brammer, 2016).

(3) Тачност = тачно класификовани / укупан број тестних инстанци

Добијени резултат представља укупно учешће тачно класификованих инстанци у тестном скупу података. За класификатор кажемо да је перфектан уколико не постоје нетачни негативни и нетачни позитивни. Међутим, тачност није једини показатељ који се користи да би се оцениле перформансе модела. Примена одабраних модела захтева опрезност јер у сетовима где су класе неравномерно распоређене (нпр. 99% и 1%), тачност предвиђања може да буде непоуздан показатељ успешности класификатора (Brammer, 2016). Због тога постоје и друге мере доброте модела као прецизност, присећање модела, FPR (енгл. *False Positive Rate*, FPR), итд. У овом раду само се обухвата показатељ тачности као један од индикатора ваљаности модела.

3. МЕТОДОЛОГИЈА СПРОВОЂЕЊА ПРЕГЛЕДА ЛИТЕРАТУРЕ

Преглед литературе је спроведен према инструкцијама Barbare Kitchenham, првенствено намењених за област информационих технологија (Kitchenham et al., 2009). Према датим смерницама, преглед се обавља у три фазе:

- *Прва фаза* подразумева *планирање прегледа* где се указује на потребу за његовим спровођењем. У оквиру ње се дефинише протокол за његову реализацију који подразумева постављање истраживачких питања, извора претраге, критеријума за укључивање и искључивање радова, евалуацију квалитета, као и начине на које су екстраховани подаци из одабраних радова.
- *Друга фаза* се односи на процес *спровођење прегледа* и обухвата реализацију претходно дефинисаних корака у протоколу. Сваки поступак је детаљно описан уз приказане резултате.
- *Трећа фаза* обухвата *извештавање резултата*. Ова фаза представља суштину прегледа литературе и њој ће бити посвећена посебна пажња.

3.1. Дефинисање циља и протокола истраживања

Систематски преглед литературе се спровео са *циљем* да укаже на тачност класификационих метода рударења података која су потенцијално примењиве у пословним информационим системима. Сумирајући резултате појединачних емпиријских истраживања, поред метода и њихове тачности, приказани су и пословни задаци (подручја) у којима су оне пронашле примену.
Истраживачка питања - како бисмо остварили постављени циљ, неопходно је формулисати истраживачка питања:

- ***Истраживачко питање 1*** **(ИП1.):** Које пословне активности (подручја) су обухваћене класификационом техником рударења података?
- ***Истраживачко питање 2*** **(ИП2.):** Које класификационе методе су коришћене у предложеним моделима?
- ***Истраживачко питање 3*** **(ИП3.):** Колика је тачност класификационих метода потенцијално примењених у пословним информационим системима?

Одговором на прво питање стичемо увид у пословне задатке који могу да се унапреде применом рударења података. Друго питање је у директној вези са првим и служи да прикаже које су методе пружиле одговор на

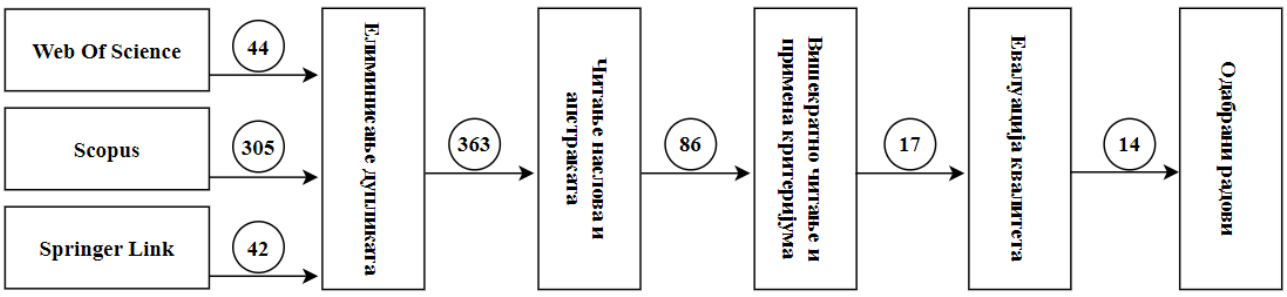
прво истраживачко питање. Одговором на треће истраживачко питање директно указујемо на тачност метода класификационе технике приказујући нумеричке резултате емпиријских истраживања.
Стратегија истраживања - извори који су служили за претрагу научних радова су познате и општеприхваћене електронске базе научних радова:

- Scopus,
- Springer Link,
- Web of Science.

Критеријуми укључивања - документи који су били укључени у претрагу су научни чланци (енгл. *scientific articles*) и излагања са научних скупова (енгл. *conference papers*). Временска димензија је ограничена на период од 10 година почевши од 15.01.2011. па до 15.01.2021. године. Обухваћени су само радови писани на енглеском језику. Критеријуми укључивања се подешавају приликом претраге електронских база и конфигуришу се након уношења кључних речи.
Критеријуми искључивања - документи који обухватају методе које не припадају класификационој техници рударења података, као и они у којима модел не може да буде примењен на све пословне субјекте независно од области пословања (елиминисане су нпр. методе примењене у медицини, банкарству, осигуравајућим друштвима, и сл.).
Критеријуми квалитета – према наведеном упутству за преглед литературе предложеном од стране Kitchenham, потребно је извршити евалуацију квалитета одабраног научног материјала. Примењени су критеријуми препоручени од стране аутора Dyba и Dingsoyr (2008).
Након филтрирања резултата, приликом мануелног читања апстраката, наслова и комплетних радова, примењују се критеријуми искључивања. Преузимају се само радови са класификационом техником рударења података, док се остале технике одбацују. Поред објашњеног критеријума, елиминишу се радови чије се методе примењују на специјализованим задацима за одређену област пословања. Пример томе су дијагностички поступци у медицини, процена ризика приликом одобравања кредита у банкарским институцијама, процена штете у осигуравајућим компанијама... Циљ је укључити методе које се односе на задатке примењиве у свим пословним организацијама, као што је нпр. однос са потрошачима.

3.2. Спровођење прегледа литературе

Након дефинисања протокола приступљено је фази спровођења прегледа литературе. Први корак представља претрага база научних радова на основу одабраних кључних речи. Пример коришћене претраге: *„data mining“ and „accuracy“ and „business“ and („application“ or „information system“).*У поменутој претрази није укључена реч „*classification*“ како би се филтрирао што већи број резултата. Разлог томе је то што поједини аутори у наслову и апстракту рада нису наводили примењене технике и методе рударења података. Укључени су временски, језички, као и критеријуми који се односе на тип документа. Примарна претрага резултовала је са 391 радом. Највећи број радова преузет је са Scopus базе (305 или 78,57%).
Следећи корак обухватао је елиминисање дупликата из примарне претраге којих је укупно било 28 (7,16%). Након елиминисања дупликата, над преосталим радовима спроведено је ручно читање апстраката и наслова како би се искључили они радови чији садржај није релевантан за постављени истраживачки циљ. Овај поступак искључио је укупно 277 радова (70,84%). Затим је уследило детаљно читање радова и примена критеријума за укључивање и искључивање. На основу овог поступка елиминисано је још 69 радова (17,65%). Након процене квалитета где је елиминисано 3 рада (0,77%), у преглед је укључено 14 радова (3,58%). На графикону бр. 1. визуелно је приказан описани поступак уз навођење преосталог броја радова након појединачних фаза инклузије.



Графикон 1: Приказ процеса инклузије радова
Извор: аутор

3.3. Приказ селектованих радова

У табели бр. 2. приказани су наслови одабраних радова који ће бити обухваћени прегледом литературе, као и година њиховог издања. У наставку рада користиће се идентификациони број уместо стандардног начина цитирања.

Табела 2: Наслови одабраних радова

Идентификациони број рада	Наслов рада	Година издања
1	"An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour" (Femina & Sudheep, 2015)	2015
2	"Analysis of Different Predicting Model for Online Shoppers' Purchase Intention from Empirical Data" (Kabir et al., 2019)	2019
3	"Bankruptcy Prediction using Data Mining Techniques" (Wagle et al., 2017)	2017
4	"Business Intelligence using the K-Nearest Neighbor Algorithm to Analyze Customer Behavior in Online Crowdfunding Systems" (Syadzali et al., 2020)	2020
5	"City digital pulse: a cloud based heterogeneous data analysis platform" (Li et al., 2017)	2017
6	"Classification of Customer Tweets Using Big Data Analytics" (Alharbi et al., 2018)	2018
7	"Decision Support System for Stock Prediction and Supplier Selection Using Least Square and C4.5 Algorithm" (Candra et al., 2018)	2018
8	"Development of IoT Mining Machine for Twitter Sentiment Analysis: Mining in the Cloud and Results on the Mirror" (Alzahrani, 2018)	2018
9	"Implementation of Data Mining Method for Classifying Company Application Data" (Setiawan & Subriadi, 2019)	2019
10	"Intelligent sentiment analysis approach using edge computing-based deep learning technique" (Sankar et al., 2020)	2019
11	"Predicting Startup Survival from Digital Traces: Towards a Procedure for Early Stage Investors" (Antretter et al., 2018)	2018
12	"Predictive Modeling For Telco Customer Churn Using Rough Set Theory" (Nafis et al., 2016)	2016
13	"Sentiment Analysis using Cosine Similarity Measure" (Bhattacharjee et al., 2015)	2015
14	"Viability prediction for retail business units using data mining techniques: a practical application in the Greek pharmaceutical sector" (Marinakos & Daskalaki, 2016)	2016

Извор: аутор на основу анализе

Највише радова написано је у 2018. години (4 или 28,57%). Радови који су написани прије 2015. године нису испунили критеријуме инклузије, као ни радови из 2021. године.

4. ИЗВЕШТАВАЊЕ О РЕЗУЛТАТИМА СПРОВЕДЕНЕ АНАЛИЗЕ

Након спроведене анализе, ради лакшег увида, одговори на постављена истраживачка питања сумирани су у табели бр. 3. За сваки рад понаособ (означен идентификационим бројем, ИБ) издвојене су коришћене класификационе технике, њихова тачност исказана у процентима, као и области (подручја) у којима су примењиване. У неким радовима приказано је само неколико техника које су дале најбоље резултате. Уместо стандардног начина цитирања, у наставку рада биће коришћен ИБ рада.

Табела 7: Сумирани резултати анализе

ИБ рада	Назив класификационе методе	Тачност исказана у %	Област (подручје) примене
1.	MLPNN Naïve Bayes	88,63 87,97	Понашање потрошача (куповина или одбијање производа)
2.	Random Forest Decision Tree Naïve Bayes SVM Stacking Voting Bagging (RF) Bagging (Extra Tree) Adaboosting Gradient boosting	89,55 85,9 84,17 83 89,65 88,58 90,25 89,88 89,2 90,34	Понашање потрошача (куповина или одбијање куповине након посете web сајта)
3.	Bayesian network Bayesian network bagging Bayesian network boosting	65,83 71,66 65,83	Предвиђање банкротства предузећа

	DT DT bagging DT boosting Logistic regression Logistic regression bagging Logistic regression boosting SVM Neural Network Neural Network boosting Neural Network (filter, bagging)	60,83 67,5 69,16 65,83 67,55 65,83 57,5 70,83 75,84 85,33	
4.	k-NN	94,37	Понашање потрошача (категорисање потрошача)
5.	LinearSVM (SemEval сет) KernelSVM (SemEval сет) LinearSVM (MVSA сет) KernelSVM (MVSA сет)	68 66,8 76 75,8	Сентимент анализа
6.	Naïve Bayes	99,39	Сентимент анализа
7.	C4.5 алгоритам	60	Анализа селекције добављача
8.	Naïve Bayes	99,2	Сентимент анализа
9.	Naïve Bayes Decision Tree Random Forest k-NN	81,7 99,92 95,83 69,54	Оптимизација софтвера
10.	CNN	85,3	Сентимент анализа
11.	Gradient boosting	55 91	Предвиђање банкротства предузећа
12.	Rough set theory класификатор Naïve Bayes	90,32 87,2	Предвиђање одлива потрошача
13.	Cosine similarity SVM MaxEnt Naïve Bayes	82,09 86,33 87,83 76,58	Сентимент анализа
14.	Fisher's Linear Discriminant k-NN C4.5 алгоритам	83,3 88,09 92,86	Предвиђање банкротства предузећа

Извор: аутор на основу анализе

4.1. ИП1. Које пословне активности (подручја) су обухваћене класификационом техником рударења података?

Пословна подручја или активности из одабраних радова у којима су класификационе технике пронашле примену су:

- понашање потрошача,
- предвиђање банкротства,
- сентимент анализа (анализа сентимента),
- анализа селекције добављача,
- оптимизација кориштених софтвера,
- предвиђање одлива потрошача.

Понашање потрошача (енгл. *consumer behavior*) представља област проучавања у три одабрана рада. Генерално посматрајући, односи се на активности које потрошач предузима када су у питању конкретни производи или услуге. У раду ИБ1 истраживачи су настојали да утврде да ли ће се клијенти определити за потписивање дугорочних депозита или не. У раду ИБ2 се на основу посете *web* сајта покушава установити да ли ће она резултовати куповином. ИБ4 указује на употребу класификационог алгоритма у разврставању потрошача на основу њиховог односа према компанији (куповној тенденцији). Формиране су четири класе: потенцијални, заинтересовани, активни и непожељни потрошачи. Рударење података у анализи понашања потрошача проналази широку примену и може да пружи помоћ у успостављању и контролисању међусобних односа између клијената и компаније.

Предвиђање банкротства предузећа (енгл. *bankrupcy prediction*) посебно је актуелно у условима савременог пословања јер оно не подразумева сигурност опстанка, нарочито новооснованих, малих и средњих компанија.

Три групе аутора су анализирале примену класификационих техника у овом подручју. Аутори ИБ11 рада фокус стављају на новостворена предузећа и предвиђање њиховог опстанка на основу дигиталних трагова. У ИБ14 раду предмет анализе је одрживост малопродајних предузећа из фармацеутског сектора на територији Грчке, док је у ИБ3 раду обухваћено 120 предузећа са одабраним атрибутима пословања кроз период од две године. Резултати класификационих техника у овом подручју примене могу да помогну наведеним групама предузећа у контроли и праћењу пословања. На основу њих се уочавају узроци проблема и предузимају корективне акције.

Сентимент анализа или анализа сентимента указује на мишљење потрошача о компанијама, њиховим производима или услугама, као и разним другим активностима, најчешће на основу њихових коментара (Farhadloo & Rolland, 2016). Као резултат рударења података јављају три класе које симболизују мишљење потрошача: позитивна, негативна или неутрална. Анализа сентимента се јавља у пет радова, где се у три (ИБ5, ИБ6 и ИБ8) рада анализа спроводи над коментарима преузетих са *twitter* платформе. У ИБ5 раду, мишљења се деле у три (позитивну, негативну и неутралну), док у ИБ6 и ИБ8 на две класе (позитивну и негативну). У ИБ10 раду аутори тестирају модел над сетом коментара који се односи на филмску базу (IMDb) и као излаз формирају две лабеле (позитивну и негативну). Аутори ИБ13 рада предлажу модел који резултира са пет излазних лабела градираних у интервалу од -2 до +2. Сет који су користили за обуку и тест односи се на коментаре корисника провајдерских услуга. Сврха наведених истраживања јесте да укажу на могућност примене рударења података у анализи сентимента, као и на погодности које оне пружају. Мишљење корисника о одређеним производима или услугама представља основу за доношење одлука које су везане за пословни асортиман. На основу прикупљених података менаџмент може да утврди предности и недостатке тренутног асортимана и да га измени уколико постоји потреба за тим.

Анализа селекције добављача (енгл. *supplier selection*) се обрађује у раду ИБ7**.** Аутори настоје одабрати оптималног добављача који се јавља у виду излазне лабеле. Обука и тест модела су спроведени над сетом података који обухвата период од 23 месеца и пет кључних атрибута. Резултати су указали да модел може пружити подршку менаџменту приликом доношења овакве врсте одлука.

Оптимизација софтвера постаје неминован процес узрокован развојем информационих технологија, нарочито софтверског инжењерства. Многобројне компаније располажу различитим апликацијама које обављају конкретне задатке или функције у пословном информационом систему. Међутим, често долази до појаве недостатака у њиховом функционисању, што је потребно исправити. Аутори у раду ИБ9 предлажу модел који на основу атрибута сврстава апликације у једну од четири раније поменуте класе. На основу резултата уочава се које је потребно модификовати, уклонити или задржати. Овај поступак доводи до смањења трошкова предузећа и повећања ефикасности пословања.

Предвиђање одлива потрошача (енгл. *customer churn*) је предмет анализе у раду ИБ12. На основу сета података телекомуникационих компанија, примењени модел настоји предвидети да ли ће потрошачи наставити да користе услуге компаније или ће се одлучити за конкуренте. Примена класификационих техника у ову сврху пружа могућност пословном субјекту да уочи и спречи одлив потрошача што директно утиче на финансијски резултат.

Највећи број радова бавио анализом сентимента, укупно њих пет. Понашање потрошача је била тема у три рада. Исти случај је и са предвиђањем банкротства. По један рад се бавио анализом селекције добављача, оптимизацијом софтверског портфолија и предвиђањем одлива потрошача. Наведена подручја примене могу да се имплементирају у пословне информационе системе свих компанија независно од области пословања. На тај начин би се стекле све предности рударења података у датим случајевима, а знатно би се олакшао процес доношења одлука од стране менаџмента.

4.2. ИП2. Које класификационе методе су коришћене у предложеним моделима?

Кроз 14 радова коришћена је укупно 41 класификациона метода. У 6 од 14 радова примењивана је само једна метода, док се у осталим радовима користило више различитих метода које су најчешће међусобно упоређиване.

Анализиран је највећи број метода који припадају категорији линеарних класификатора (укупно 9). Од тога је у чак 6 радова коришћен Naïve Bayes класификатор, што га уједно чини и најчешће коришћеном методом у овој анализи. Од наведених области употребе само није примењен у раду који се односи на анализу селекције добављача. У свим осталим областима је извршено његово тестирање.

Након линеарних класификатора најчешће коришћена метода је стабло одлучивања. Укупно се користила у 6 различитих радова. Два рада су користила C4.5 алгоритам за класификацију (ИБ7 и ИБ14), док је у ИБ13 примењена максимална ентропија. Једина област у којој није коришћена наведена метода односи се на предвиђање одлива потрошача.

SVM се као метода користила у 4 различита рада. У раду ИБ5 примењена су два различита модела наведене методе (LinearSVM и KernelSVM) тестирана над два сета података. Примењена је у сентимент анализи, предвиђању банкротства и понашању потрошача.

Класификација путем неуронских мрежа обрађена је у 3 рада. Коришћене су конволуцијске неуронске мреже и вишеслојни перцептрон. Примену проналазе у понашању потрошача, предвиђању банкротства предузећа и

сентимент анализи. K-NN метода се применила у 3 рада у областима предвиђања банкротства, оптимизацији софтвера и понашању потрошача.
Ансамбл метода за циљ има обуку више различитих класификатора, а затим комбинацију њихових предвиђања користи за нове инстанце путем неког облика гласања. Овакви методи се користе ради повећања тачности, међутим то није загарантовано (Brammer, 2016). У анализи је укупно извојено 15 различитих метода од којих су се по два пута користили случајна шума и Gradient Boosting. Укупно су се користили у 4 различита рада, а обухватали су области понашања потрошача, предвиђања банкротства и оптимизацију софтвера.
Поред наведених метода користиле су се још и косинусна сличност и теорија грубих скупова која припада *soft computing* техникама.

4.3. ИП3. Колика је тачност класификационих метода потенцијално примењених у пословним информационим системима?

Радови у којима је рударење података употребљено за анализу ***понашања потрошача*** обухватили су предвиђања њихове опредељености за куповину неке робе или услуге (ИБ1 и ИБ2), као и сврставање купаца у одговарајуће категорије (ИБ3). У раду ИБ1 највећу тачност пружила је MLPNN неуронска мрежа која је са 88,63% тачности предвиђала да ли ће се купац одлучити за куповину, док је у раду ИБ2 највећу тачност пружио Gradient Boosting алгоритам у износу од **90,34%**. Поред поменуте методе, у раду ИБ2 је кориштено још 9 метода од којих је најмању тачност имала SVM у износу од 83%. У раду ИБ4 аутори су формирали модел заснован на k-NN методи који са **94,37%** тачности сврстава потрошаче у одговарајуће категорије на основу њиховог односа према компанији. Висок ниво тачности наведених модела омогућава предузећима да предвиде понашање потрошача и да их сврстају у одговарајуће категорије што олакшава процес доношења одлука и прилагођавања маркетинг стратегије.
Модели који су кориштени за ***предвиђање банкротства*** (ИБ3, ИБ11, ИБ14) применили су различите методе од којих је највећу тачност имао C4.5 алгоритам у раду ИБ14 у износу од **92,86%**. У раду ИБ11 91% тачности је пружио Gradient Boosting метод за стопу преживљавања предузећа од 50%, док је за стопу од 10% пружио 55% тачности. Аутори ИБ3 рада су добили највећи процент тачности у износу од 85,33% код модела заснованог на неуронским мрежама унапређеног bagging техником и атрибутима одабраним филтер методом. Имплементирање наведених модела у информациони систем може да пружи помоћ у контроли и праћењу пословања упозоравајући када тренутно пословање води ка банкротству.
Највећи проценат тачности код ***сентимент анализе*** пружио је Naïve Bayes метод. У раду ИБ6 је резултовао са **99,39%**, док је у раду ИБ8 имао 99,20% тачности. Он се такође користи још и у раду ИБ13, али највећу тачности је пружио метод максималне ентропије у износу од 87,83% у случајевима када за резултат постоје три излазне класе. У наведеном раду метода косинусне сличности је имала највећи проценат тачности када је лабела имала нумеричку вредност у износу од 71,5%. CNN је кориштен у раду ИБ10 и пружио је 85,3% тачности, док је у раду ИБ5 кориштен SVM који је имао највећу стопу тачности у износу од 76%. Висока тачност класификације мишљења потрошача омогућава предузећима да прате њихова искуства и да на основу тога прилагођавају своје производе и услуге тржишним захтевима.
У раду ИБ7 C4.5 алгоритам који припада стаблу одлучивања имао је тачност у износу од **60%** приликом ***селекције добављача***. Иако проценат није висок, аутори рада сматрају да овај модел може пружити помоћ приликом избора добављача.
Оптимизација софтвера била је предмет истраживања аутора у раду ИБ9, где су применили четири различите методе у класификацији софтвера. Највећу тачност пружило је стабло одлучивања у износу од **99,92%**, док је најмању имао k-NN метод у износу од 69,54%. Резултати овог модела помажу у оптимизацији информационих система предузећа указујући на потребу за уклањањем, модификацијом и задржавањем тренутних апликација.
Предвиђање одлива потрошача је обрађено у раду ИБ12 где су кориштена два метода. Naïve Bayes је пружио 87,2% тачности, док је класификатор заснован на теорији тврдих скупова имао знатно боље резултате у износу од **90,32%**. Имплементирање овог модела у пословне информационе системе омогућава откривање и приступ потрошачима чији захтеви нису испуњени понудом предузећа. Као и код сентимент анализе оваква врста података је кључна за прилагођавање понуде.

5. ЗАКЉУЧАК

Улога рударења података у савременим информационим системима огледа се кроз пружање употребљивих информација за доношење пословних одлука. Како би се утврдила њихова релевантност за решавање одређених проблема, анализирана је тачност класификационих метода кроз преглед литературе у последњој деценији. Обухваћени су само они модели који проналазе примену у свим пословним субјектима без обзира на област пословања. Анализирајући одабране радове кроз одговоре на постављена истраживачка питања, долазимо до подручја и метода у којима су они пронашли примену. Код предвиђања куповних активности потрошача највећу тачност пружио је Gradient Boosting алгоритам у износу од 90,34%. У категорисању потрошача према њиховом односу са пословном организацијом највећу тачност има k-NN метод са 94,37%, као и код оптимизације софтверског портфолија са 99,92%. Предвиђање банкротства предузећа уз помоћ C4.5 алгоритма

остварује тачност од 92,86%, док исти алгоритам пружа и најбољи резултат код селекције добављача у износу од 60%. Сентимент анализа пружа висок ниво тачности од 99,39% употребом Naïve Bayes метода, а истим методом се долази до износа од 90% за предвиђање одлива потрошача. На основу високих стопа тачности анализираних класификационих модела закључује се да постоји потенцијална могућност њихове примене у пословним информационим системима. На тај начин се обезбеђују употребљиве информације доносиоцима одлука. Међутим, тачност је само један од показатеља доброте модела, па је због тога неопходно обратити пажњу и на остале уколико постоји несразмеран однос између класа у скупу података.

ЛИТЕРАТУРА

Aggrawal, C. C. (2019). Data Mining - The text book. In *Statistical Field Theor* (Vol. 53, Issue 9). https://doi.org/10.1007/978-3-319-14142-8

Alharbi, A. N., Alnnamlah, H., & Liyakathunisa. (2018). Classification of customer tweets using big data analytics. In *Advances in Intelligent Systems and Computing* (Vol. 753). Springer International Publishing. https://doi.org/10.1007/978-3-319-78753-4_13

Alzahrani, S. M. (2018). Development of IoT mining machine for Twitter sentiment analysis: Mining in the cloud and results on the mirror. *2018 15th Learning and Technology Conference, L and T 2018*, 86–95. https://doi.org/10.1109/LT.2018.8368490

Antretter, T., Blohm, I., & Grichnik, D. (2018). Predicting startup survival from digital traces: Towards a procedure for early stage investors. *International Conference on Information Systems 2018, ICIS 2018*, *Vc*, 1–9.

Bhattacharjee, S., Das, A., Bhattacharya, U., Parui, S. K., & Roy, S. (2015). Sentiment analysis using cosine similarity measure. *2015 IEEE 2nd International Conference on Recent Trends in Information Systems, ReTIS 2015 - Proceedings*, 27–32. https://doi.org/10.1109/ReTIS.2015.7232847

Brammer, M. (2016). Principles of data mining. In *Drug Safety* (Vol. 30, Issue 7). https://doi.org/10.2165/00002018-200730070-00010

Candra, B. P., Saputra, E. R. S. H., Ruhamah, Wicaksono, K., & Kusrini, K. (2018). Decision support system for stock prediction and supplier selection using least square and C4.5 algorithm. *Proceedings - 2018 3rd International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2018*, 241–246. https://doi.org/10.1109/ICITISEE.2018.8721001

Dybå, T., & Dingsøyr, T. (2008). Empirical studies of agile software development: A systematic review. In *Information and Software Technology* (Vol. 50, Issues 9–10). https://doi.org/10.1016/j.infsof.2008.01.006

Farhadloo, M., & Rolland, E. (2016). Fundamentals of sentiment analysis and its applications. *Studies in Computational Intelligence*, *639*(March), 1–24. https://doi.org/10.1007/978-3-319-30319-2_1

Femina, B. T., & Sudheep, E. M. (2015). An efficient CRM-data mining framework for the prediction of customer behaviour. *Procedia Computer Science*, *46*(Icict 2014), 725–731. https://doi.org/10.1016/j.procs.2015.02.136

Kabir, M. R., Ashraf, F. Bin, & Ajwad, R. (2019). Analysis of different predicting model for online shoppers' purchase intention from empirical data. *2019 22nd International Conference on Computer and Information Technology, ICCIT 2019*, *December*. https://doi.org/10.1109/ICCIT48885.2019.9038521

Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering - A systematic literature review. *Information and Software Technology*, *51*(1), 7–15. https://doi.org/10.1016/j.infsof.2008.09.009

Li, Z., Zhu, S., Hong, H., Li, Y., & El Saddik, A. (2017). City digital pulse: a cloud based heterogeneous data analysis platform. *Multimedia Tools and Applications*, *76*(8), 10893–10916. https://doi.org/10.1007/s11042-016-4038-2

Marinakos, G., & Daskalaki, S. (2016). Viability prediction for retail business units using data mining techniques: a practical application in the Greek pharmaceutical sector. *International Journal of Computational Economics and Econometrics*, *6*(1), 1. https://doi.org/10.1504/ijcee.2016.073310

Nafis, N. S. M., Makhtar, M., Awang, M. K., Rahman, M. N. A., & Deris, M. M. (2016). Predictive modeling for telco customer churn using rough set theory. *ARPN Journal of Engineering and Applied Sciences*, *11*(5), 3203–3207.

Piatetsky-Shapiro, G. (1999). The data-mining industry coming of age. *IEEE Intelligent Systems and Their Applications*, *14*(6), 32–34. https://doi.org/10.1109/5254.809566

Rahman, N. (2018). Data Mining Techniques and Applications. *International Journal of Strategic Information Technology and Applications*, *9*(1), 78–97. https://doi.org/10.4018/ijsita.2018010104

Sankar, H., Subramaniyaswamy, V., Vijayakumar, V., Arun Kumar, S., Logesh, R., & Umamakeswari, A. (2020). Intelligent sentiment analysis approach using edge computing-based deep learning technique. *Software - Practice and Experience*, *50*(5), 645–657. https://doi.org/10.1002/spe.2687

Setiawan, H., & Subriadi, A. P. (2019). Implementation of data mining method for classifying company application data. *Proceedings - 2019 5th International Conference on Science and Technology, ICST 2019*, 1–6. https://doi.org/10.1109/ICST47872.2019.9166222

Syadzali, C., Suryono, S., & Endro Suseno, J. (2020). Business Intelligence using the K-Nearest Neighbor Algorithm to Analyze Customer Behavior in Online Crowdfunding Systems. *E3S Web of Conferences*, *202*, 1–7. https://doi.org/10.1051/e3sconf/202020216005

Wagle, M., Yang, Z., & Benslimane, Y. (2017). Bankruptcy prediction using data mining techniques. *2017 8th International Conference on Information and Communication Technology for Embedded Systems, IC-ICTES 2017 - Proceedings*, 2–5. https://doi.org/10.1109/ICTEmSys.2017.7958771

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining. In *Data Mining*. https://doi.org/10.1016/C2009-0-19715-5