ISBN: 978-99955-45-38-3
UDK: 005.96:004.42

Datum prijema rada: 08.06.2022.
Datum prihvatanja rada: 27.06.2022.
**Pregledni naučni rad**

# RUDARENJE PODATAKA U PREDVIĐANJU BROJA ZAPOSLENIH

# DATA MINING APPLICATION FOR PREDICTING NUMBER OF EMPLOYEES

**Rade Božić**
Univerzitet u Istočnom Sarajevu, Fakultet poslovne ekonomije Bijeljina,
Republika Srpska, Bosna i Hercegovina
rade.bozic94@hotmail.com
ORCID: 0000-0001-6956-8049

***Apstrakt:*** *Zaposlenost predstavlja jedan od ključnih makroekonomskih pokazatelja svake nacionalne privrede. U savremenim uslovima globalizacije, velike korporacije svoje poslovanje obavljaju na različitim nacionalnim tržištima zapošljavajući veliki broj radno sposobnog stanovništva. Kako rudarenje podataka pronalazi primenu u ekonomskoj sferi, odgovarajuće tehnike i metode se mogu primeniti i kod analize radnog kadra koji predstavlja pokretač svih ekonomskih aktivnosti. U ovom radu nastoji se odgovoriti na sledeće pitanje: da li je moguće predvideti broj stalno zaposlenih radnika u kompanijama na osnovu finansijskih i drugih podataka vezanih za njihovo poslovanje? Za analizu pomenutog problema formiran je skup od 150 kompanija koje pripadaju S&P500 berzanskom indeksu, zajedno sa osam atributa koji su se koristili za određivanje broja stalno zaposlenih radnika. Korištene metode u radu su linearna regresija, k-NN, višeslojni perceptron i stablo odlučivanja. Evaluacione mere modela su prikazane i protumačene.*

***Ključne riječi:*** *рударење података, предвиђање, број запослених*

***JEL класификација:*** *E47, O11*

***Abstract:*** *Employment is one of the main macroeconomic indicators of any national economy. In the modern conditions of globalization, large corporations conduct their business in various national markets, employing a large number of people. As data mining finds application in the economic sphere, appropriate techniques and methods can be applied in the analysis of the workforce that is the initiator of all economic activities. This paper seeks to answer the following question: is it possible to predict the number of full-time employees in companies based on financial and other data*

*related to their business? To analyze this problem, a data set of 150 companies from S&P500 stock exchange index was formed, together with eight attributes that were used to determine the number of full-time employees. The data mining methods used in the paper are linear regression, k-NN, multilayer perceptron and decision tree. The evaluation measures of the model are presented and interpreted.*

***Key Words:*** *data mining, forecasting, nuber of employees*

***JEL classification:*** *E47, O11*

## 1. UVOD

Rudarenje podataka se definiše kao proces prikupljanja, obrade, analize i ekstrahovanja korisnih informacija iz skupa podataka (Aggarwal, 2015). Koristi se za rešavanje različitih problema kroz postupak analize, a ujedno se može definisati i kao automatski ili poluautomatski proces koji otkriva šablone ponašanja u podacima. Ekonomisti, statističari, prognostičari i inženjeri komunikacija su dugotrajno radili sa idejom da se šabloni u podacima mogu pronaći, automatizovati, oceniti i koristiti za predviđanje. Oni su valjani samo ukoliko pružaju određenu (ekonomsku) prednost (Witten, Frank, Hall, & Cristopher, 2017). Savremeno društvo obiluje velikom količinom skladištenih podataka (eng. *big data*) što čini nesagledliv potencijal za njihovu obradu koja pruža korisne informacije. Upravo ovo predstavlja ključ razvoja rudarenja podataka i glavni uzrok velike zainteresovani istraživačke javnosti za njegovu primenu.

Rudarenje podataka kao disciplina pronalazi primenu u različitim sferama društva kao što su biologija, medicina, umetnost, telekomunikacije, transport, saobraćaj, sport, edukacija itd. Pored navedenih oblasti istaknuta je i primena u ekonomiji gde popularnost ovakve vrste obrade podataka raste razvojem informacionih tehnologija i skladištenja podataka. Kada govorimo o preduzeću kao ekonomskom subjektu, veliki broj istraživača nastoji primeniti rudarenje podataka nad ovim entitetom u cilju povećavanja performansi poslovanja. Predviđanje prodaje i troškova, klasifikacija i klasterisanje potrošača, procena neuspeha u poslovnim poduhvatima, automatsko odobravanje kredita, direktni marketing, procena odliva potrošača su samo neki od primera primene rudarenja podataka u preduzećima. Kako su zaposleni jedni od ključnih nosilaca proizvodnog i uslužnog procesa, u interesu menadžmenta je prikupljanje što veće količine podataka koji se odnose na njihove aktivnosti, performanse, mišljenja, finansijske i druge izdatke. Rudarenje podataka kroz svoje tehnike pronalazi primenu i u ovom aspektu, pružajući korisne informacije za olakšavanje procesa donošenja odluka koje se odnose na ljudske resurse zaposlene u poslovnim organizacijama.

U rudarenju podataka se posebno ističu četiri vrste njegove primene podeljene u dve grupe. Prvu grupu čine tehnike nadgledanog učenja – klasifikacija i regresija. Klasifikacija je najčešće primenjena tehnika i odnosi se na zadatak koji se javlja u svakodnevnom životu poput grupisanja pacijenata u klase prema riziku oboljenja od

određenih bolesti ili građana prema opredeljenosti za podršku odgovarajućim političkim partijama. Vrednost koja se predviđa ovom tehnikom je u formi labele (eng. *label*). Regresija ili numeričko predviđanje (eng. *numerical prediction*) se odnosi na predviđanje vrednosti koja ima numerički iskaz kao što je profit kompanije ili cena akcija. Neuronske mreže predstavljaju jedan od najčašće korišćenih metoda kada se posmatra ova tehnika. Drugu grupu primene čine nenadgledane tehnike – klasterisanje i pravila pridruživanja. Algoritmi klasterisanja kroz ispitivanje podataka nastoje pronaći grupe koje su slične. Npr. osiguravajuće kompanije mogu da grupišu potrošače prema prihodima, godinama, tipu polise ili prethodnom iskustvu. Pravila prudruživanja se odnose na pronalazak veza između vrednosti varijabli. Jedan od tipičnih primera primene ove tehnike naziva se analiza potrošačke korpe (eng. *market basket analysis*). Odnosi se na pronalaženje veza između proizvoda koje potrošači kupuju što dalje služi za povećanje efikasnosti prodaje (Bramer, 2015).

Cilj ovog rada je izvršiti predviđanje broja zaposlenih u kompaniji na osnovu prikupljenog skupa podataka. Prema prirodi rezultata koji predstavlja izlaz iz modela, tehnika koja će se primeniti je numeričko predviđanje ili regresija. Važno je napomenuti da sam pojam "predviđanje" nije povezan sa vremenskom serijom podataka i ne odnosi se na budućnost, nego na vrednost koja se dobija kao rezultat obrade algoritma. Za ulazne atribute (varijable) odabrani su sledeći finansijski pokazatelji: ukupni i operativni prihod, ukupni troškovi i profit, vrednost kompanije i poreske obaveze. Pored njih u skup je uključen još i privredni sektor kome preduzeće pripada, godina osnivanja i broj zaposlenih kao vrednost koja se predviđa. Skup se sastoji od 150 kompanija koja pripadaju S&P500 berzanskom indeksu.

Primenjeno je četiri metode: linearna regresija (eng. *linear regression*), metod najbližih suseda (eng. *k-Nearest Neighbors,* k-NN), stablo odlučivanja (eng. *decision tree*) i višeslojni perceptron (eng. *multi-layer perceptron*). Evaluacija performansi modela izvršena je sa četiri korištene mere: MAE (eng. *mean absolute error*), RMSE (eng. *root mean squared error*), RAE (eng. *relative absolute error*) i RRSE (eng. *root relative sqared error*). Rezultati modela su prikazani, protumačeni i međusobno upoređeni.

## 2. RUDARENJE PODATAKA I RADNA SNAGA, PREGLED LITERATURE

Veliki broj radova odnosi se na analizu radne snage u kompanijama. Različiti aspekti primene ukazuju na velike mogućnosti rudarenja podataka u ovom domenu. Autori koji su izučavali pomenutu oblast kreirali su različite modele sa raznovrsnim aspektima predviđanja.

Prepoznavajući ulogu i važnost radne snage u poslovnim organizacijama, najveći broj autora je pisao o predviđanju odlaska zaposlenih kao jednom od ključnih problema. YongKang (2022), Tanasescu i Bologa (2021), Choi i Choi (2021), Yadav, Jain i Singh (2018), Yigit i Shourabizdeh (2017), Wild Ali (2021), Ozdemir, Coskun, Gezer i Cagri (2020), Liu, Akkineni i Story (2020), Gao, Wen i Zhang (2019), Bindra, Sehgal i Jain (2019), Shankar, Rajanikanth, Sivaramaraju i Vssr Murthy (2018) su samo neki od autora koji su se posvetili izučavanju problema odlaska zaposlenih.

Wei, Lihond i Liu (2022), Mahmoud, Al Shawabkeh, Salameh i Al Amro (2019), Golestani i ostali (2019) pisali su o predviđanju performansi zaposlenih kao pokazatelja koji olakšava proces donošenja odluka od strane sektora ljudskih resursa i menadžmenta.

Pojedini autori kao što su Laijawala, Aachaliya i Jatta (2020), Soleimanfar, Navabakhsh i Sebt (2019), Abolfalz i ostali (2022) su uočili i važnost zdravstvenog stanja zaposlenih kao faktora koji utiče na njihove performanse i zadovoljstvom prema kompaniji, te primenili svoje modele za predviđanje istog.

Chuanzhu (2022) i Jantan, Hamdan i Othman (2011) su vršili predviđanje talenata među zaposlenima koji bi svojim kvalitetima mogli da unaprede poslovanje preduzeća. Al-rasheed (2021) i Rista, Ajdari i Zenuni (2020) su predviđali izostanke zaposlenih sa radnog mesta, dok su Dutta, Halder i Dasgupta (2018) i Ahn i ostali (2019) rudarenje podataka koristili u svrhu predviđanja adekvatne visine plate za zaposlene. O predviđanju nivoa zadovoljstva zaposlenih pisali su Dharani i Kamalakkannan (2022), dok su o predviđanju nivoa njihovog stresa prilikom obavljanja radnih aktivnosti pisali Anitha i Vanitha (2021).

Alola i Atsa'am (2020) su se bavili analizom prihološkog kapitala zaposlenih kojeg su definisali kao pozitivno stanje pojedinca koje se sastoji od samoefikasnosti, optimizma, nade i otpornosti. O otkrivanju insajdera u poslovnoj organizaciji pisali su Jiang i ostali autori (2018). U ovom radu se vrši predviđanje broja stalno zaposlenih radnika u kompaniji, što može ukazati na odstupanje broja angažovane radne snage u odnosu na pravila koja uoče algoritmi i modeli prezentuju kao rešenje.

## 3. SKUP PODATAKA

Za analizu pomenutog problema prikupljen je skup od 150 kompanija koje pripadaju S&P500 berzanskom indeksu, što ujedno govori i da sam skup ima toliko različitih instanci (vrsta). Standard and Poor kompanija objavljuje veliki broj indeksa među kojima je najpoznatiji upravo S&P500 kompozitni prosek akcija. U njegovom sastavu postoje određeni podindeksi: industrijski, transportni, uslužno-komunalni i finansijski. U njega su uključene akcija sa NYSE; AMEX i OTC (Šoškić, 2006). Svi podaci su prikupljeni preko *Yahoo.finance* veb platforme (Yahoo!, 2022).

Skup se sastoji od 9 različitih atributa (varijabli) navedenih u tabeli 1. Za svaku odabranu kompaniju uvršteni su neki od osnovnih finansijskih pokazatelja: ukupan prihod, ukupan profit, operativni prihod, poreske obaveze, ukupni troškovi i vrednost poslovnog subjekta. Vrednosti poslovnog subjekta iskazana je u milijardama američkih dolara (USD), dok su ostali finansijski pokazatelji iskazani u hiljadama USD.

Pored finansijskih pokazatelja u skup su još uključeni i sektor u kome kompanija obavlja poslovanje, godina osnivanja i broj zaposlenih koji ujedno predstavlja i atribut koji se predviđa. Uključeno je pet različitih privrednih sektora gde je za svaki odabrano po 30 kompanija: zdravstvo (eng. *healthcare*), informacione tehnologije

(eng. *information technology*), industrijski sektor (eng. *industrials*), komunalije (eng. *utilities*) i sektor svakodnevne finalne potrošnje (eng. *consumer staples*). Jedino ovaj atribut nema numeričku vrednost nego je dat u vidu klase. Izvršena je normalizacija podataka.

**Tabela 1**. odabrani atributi u skupu podataka

| R.br. | Naziv atributa | Jedinica |
|---|---|---|
| **1.** | Ukupan prihod | Hiljade USD |
| **2.** | Ukupan profit | Hiljade USD |
| **3.** | Operativni prihod | Hiljade USD |
| **4.** | Poreske obaveze | Hiljade USD |
| **5.** | Ukupni troškovi | Hiljade USD |
| **6.** | Sektor | Naziv |
| **7.** | Vrednost poslovnog subjekta | Milijarde dolara |
| **8.** | Godina osnivanja | Godina |
| **9.** | Broj zaposlenih | Broj |

***Izvor:*** *istraživanje autora, originalno prikupljeno iz Yahoo finance platforme (2022)*

U skupu nema nedostajućih podataka a svi finansijski pokazatelji prikupljeni su za prethodnu obračunsku godinu koja se može razlikovati u odnosu na kalendarsku (ne počinje kod svih kompanija prvog januara). U tabeli broj 2. prikazani su osnovni statistički podaci koji se odnose na minimalne, maksimalne i srednje vrednosti numeričkih pokazatelja:

**Tabela 2**. osnovni statistički podaci vezani za skup

| Naziv pokazatelja | Minimalna vrednost | Maksimalna vrednost | Prosečna vrednost |
|---|---|---|---|
| Ukupni prihod | 1.024.200,00 | 572.754.000,00 | 30.781.462,93 |
| Ukupan profit | 27.000,00 | 429.000.000,00 | 12.118.611,59 |
| Operativni prihod | -5.046.000,00 | 108.949.000,00 | 4.086.710,33 |
| Poreske obaveze | -743.000,00 | 14.527.000,00 | 610.785,95 |
| Ukupni troškovi | 1.059.700,00 | 546.812.000,00 | 26.699.857,29 |
| Vrednost kompanije | 9,10 | 2.450,00 | 87,17 |
| Ukupan broj stalno zaposlenih | 2.094 | 2.300.000 | 68.554 |

***Izvor:*** *istraživanje autora*

## 4. ALATI, METODE I EVALUACIONI PARAMETRI

U postupku analize korišten je *open source* WEKA alat sa verzijom 3.8.6. razvijen od strane Waikato univerziteta (Hamilton, Novi Zeland). Sastoji se od kolekcije algoritama za mašinsko učenje (eng. *machine learning*) koji se mogu koristiti u rudarenju podataka. Za više detalja o alatu pogledati "Data Mining: Practical Machine Learning Tools and Techniques" (Frank, Hall, & Witten, 2016). U radu je primenjeno

четири различите методе рударења података: линеарна регресија, метод најближих суседа, стабло одлучивања и вишеслојни перцептрон.

За поделу скупа на део за обуку и тестирање коришћен је метод *k-fold* унакрсне валидације (eng. *k-fold cross validation*). Најчешће се користи у случајевима када се скуп састоји од малог броја инстанци.

Ако се скуп састоји од *N* инстанци, оне се деле у *k* једнаких делова (најчешће од 5 до 10). Ако *N* није дељиво са *k*, онда последњи део ће имати мање инстанци него остали *k-1* делови. Сваки део се користи за тест, док се остатак *k-1* делова користи за обуку скупа (Bramer, 2015). У WEKA алату одабрано је 10 скупова што значи да се евалуација врши 11 пута (10 пута за сваки део и једном за цели скуп).

За евалуацију модела су употребљене најчешће коришћене мере: просечна апсолутна грешка (MAE), просечна квадратна грешка (RMSE), релативна апсолутна грешка (RAE), квадратни корен релативне квадратне грешке (RRSE). Формуле су приказане испод:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

$$RAE = \frac{\sum_{i=1}^{N}|\hat{y}_i - y_i|}{\sum_{i=1}^{N}|\overline{y_i} - y_i|}$$

$$RRSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\overline{y_i} - y_i)^2}}$$

У наведеним формулама $\hat{y}_i$ представља предвиђену вредност, $y_i$ представља стварну вредност за *N* број опсервација а $\overline{y_i}$ представља просечну вредност *y*. MAE представља просечну апсолутну разлику између вредности које је предвидио модел и посматраних историјских података, док MSE указује на просек квадрата њихове разлике.

RAE показује укупну апсолутну разлику између стварних и предвиђених вредности, док RRSE показује квадратни корен збира квадрираних грешака модела нормализованог збиром квадратних грешака простог модела.

## 5. REZULTATI ANALIZE

Prvo je primenjen algoritam linearne regresije koji je rezultatirao sljedećom funkcijom:

***Broj stalno zaposlenih =***

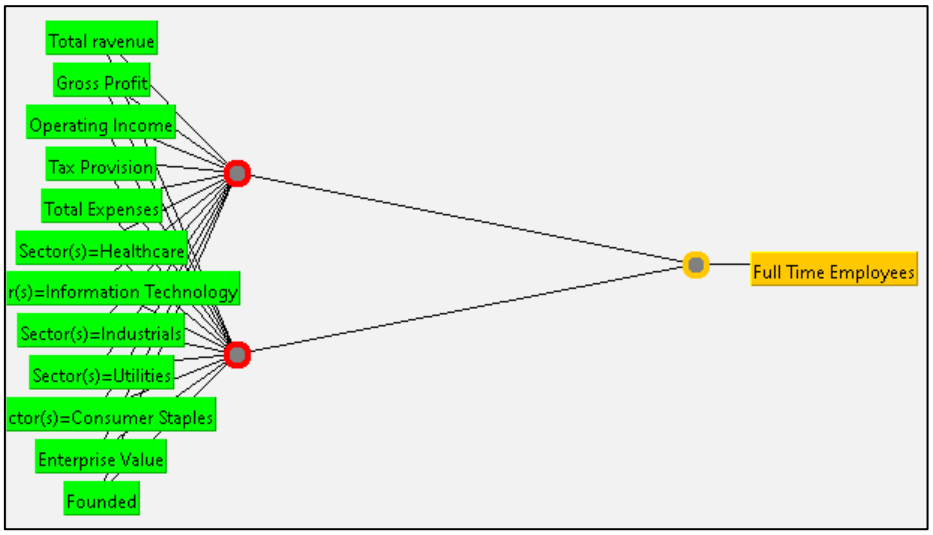
2116707.5384 * Gross Profit +
-1417508.2745 * Operating Income +
638668.1066 * Tax Provision +
321091.0046 * Total Expenses +
45122.5712 * Sector(s)=Industrials,Information Technology,Consumer Staples +
-22180.5617 * Sector(s)=Consumer Staples +
27802.4379

Kako je algoritam automatski selektovao atribute uz pomoć M5 metode i eliminisao one koji su kolinearni, iz jednačine su izuzeti ukupan prihod, vrednost poslovnog subjekta i godina osnivanja. Ovo ostavlja 5 ulaznih varijabli za pomenuti metod.

Kod metode najbližih suseda podešeno je da algoritam bira 5 susednih instanci (5-NN) pomoću Euklidove distance kao mere njihove udaljenosti. Uključeni su svi atributi kao ulaz u algoritam.

Višeslojni perceptron (grafikon 1) je dao najbolje rezultate sa dva skrivena sloja, stopom učenja od 0.2 i momentumom 0.2, dok je vreme obuke iznosilo 10 sekundi. Uključeni su svi atributi u model.

**Grafikon 1.** višeslojni perceptron




Stablo odlučivanja je implementiralo M5 model algoritma sa omogućenim orezivanjem stabla uz minimalne četiri instance neophodne za list. Jednačina glasi:

***Broj stalno zaposlenih =***
0.0049 * Gross Profit
- 0.0138 * Operating Income
+ 0.0515 * Tax Provision
+ 0.0006 * Total Expenses
+ 40395.2893 * Sector(s)=Information Technology,Consumer Staples
- 31830.7543 * Sector(s)=Consumer Staples
+ 8415.2804

Stablo odlučivanja je uzelo identične atribute prilikom kreiranja modela kao i kod linearne regresije eliminišući ukupan prihod, godine osnivanja i vrednost poslovnog subjekta. U tabeli 3. prilazani su rezultati analize kroz upotrebljene metode.

**Tabela 3**. evaluacioni parametri modela

| Metoda | MAE | RMSE | RAE | RRSE |
|---|---|---|---|---|
| Linearna regresija | 43947,88 | **86327,06** | 56,79% | **41,81%** |
| k-NN | 51904,09 | 193833,87 | 67,07% | 93,88% |
| Višeslojni perceptron | 54691,86 | 132774,41 | 70,67% | 64,31% |
| Stablo odlučivanja | **41032,04** | 87937,12 | **53,02%** | 42,59% |

**Izvor:** Kalkulacija autora

Rezultati modela pokazuju da je najmanju MAE imalo stablo odlučivanja u iznosu od **41032,04**, dok je najmanju RMSE imala linearna regresija u iznosu od **86327,06**. Što se tiče procentualno iskazanih pokazatelja najmanju RAE je imalo stablo odlučivanja (**53,02%**), dok je najmanju RRSE imala linearna regresija (**41,81%**). Analiza je pokazala da su najbolje rezultate pružili algoritmi linearne regresije i stabla odlučivanja. Nije bilo algoritma koji je imao najmanja sva četiri evaluaciona parametra. Takođe, ne može se izdvojiti ni algoritam koji je pružio najveće greške, iz razloga što su 5-NN i višeslojni perceptron kombinovano ostvarili najslabije rezultate. Posmatrajući iznose grešaka, može se utvrditi da ni jedan algoritam nije ostvario adekvatne rezultate koji bi ukazali na mogućnost predviđanja broja stalno zaposlenih radnika sa malim ostupanjem od tačnih vrednosti.

## ZAKLJUČAK

Rudarenje podataka pronalazi različitu primenu kada je reč o podacima koji se odnose na zaposlene u kompanijama. U ovom radu je cilje je bio predviđanje broja stalno zaposlenih na osnovu osam različitih atributa (finansijskih i nefinansijskih) koji sačinjavaju skup podataka. Analiza je sprovedena nad kompanijama koje pripadaju S&P500 berzanskom indeksu. Uz pomoć WEKA alata primenjeno je četiri različite metode rudarenja podataka: linearna regresija, k-NN, višeslojni perceptron i stablo odlučivanja. Za određivanje skupa na deo za obuku i test korištena je k-fold unakrsna validacija. Najbolje rezultate kombinovano su pružili stablo odlučivanja (MAE i RAE) i linearna regresija (RMSE i RRSE), međutim prema numeričkim vrednostima

grešaka može se zaključiti da modeli nisu imali velikog uspeha u predviđanju broja stalno zaposlenih radnika. Razlog tome mogu biti odabrani atributi, mali broj kompanija odabranih u skupu podataka ili korištene metode. Ujedno ovo predstavlja i osnovu za buduća istraživanja koja mogu biti poduzeta vezano za pomenuti problem.

## LITERATURA


[1] Abolfazl, M., Khan, I. R., Chakraborty, S., Karthik, M., Mehta, K., Liaqat, A., & Nuagah, S. J. (2022). Data Mining in Employee Healthcare Detection Using Intelligence Techniques for Industry Development. Journal of Healthcare Engineering.

[2] Aggarwal, C. C. (2015). Data Mining, The Textbook. Springer International Publishing Switzerland 2015.

[3] Ahn, S., Couture, S., Cuzzocrea, A., Dam, K., Grasso, G., Leung, C., . . . Wodi, B. (2019). A Fuzzy Logic Based Machine Learning Tool for Supporting Big Data Business Analytics in Complex Artificial Intelligence Environments. IEEE International Conference on Fuzzy Systems. New Orleans: Institute of Electrical and Electronics Engineers Inc.

[4] Alola, U. V., & Atsa'am, D. D. (2020). Measuring employees' psychological capital using data mining approach. JOURNAL OF PUBLIC AFFAIRS.

[5] Al-Rasheed, A. (2021). Identification of important features and data mining classification techniques in predicting employee absenteeism at work. International Journal of Electrical and Computer Engineering, 4587-4596.

[6] Anitha, S., & Vanitha, M. (2021). Optimal artificial neural network-based data mining technique for stress prediction in working employees. SOFT COMPUTING, 11523-11534.

[7] Bindra, H., Sehgal, K., & Jain, R. (2019). Optimisation of C5.0 using association rules and prediction of employee attrition. In Lecture Notes in Networks and Systems (pp. 21-29). Springer.

[8] Bramer, M. (2015). Principles of Data Mining - Third edition. Springer Nature.

[9] Choi, Y., & Choi, J. W. (2021). The prediction of workplace turnover using machine learning technique. International Journal of Business Analytics, 1-10.

[10] Chuanzhu, Z. (2022). Evaluation and analysis of human resource management mode and its talent screening factors based on decision tree algorithm. Journal of Supercomputing.

[11] Dharani, D., & Kamalakkannan, S. (2022). Prediction of Job Satisfaction from the Employee Using Ensemble Method. International Conference on Advanced Computing Technologies and Applications. Coimbatore: Institute of Electrical and Electronics Engineers Inc.

[12] Dutta, S., Halder, A., & Dasgupta, K. (2018). Design of a novel prediction engine for predicting suitable salary for a job. IEEE International Conference on Research in Computational Intelligence and Communication Networks,

ICRCICN 2018 (pp. 275-279). Kolkata: Institute of Electrical and Electronics Engineers Inc.

[13] Erkam, G., Kayakutlu, G., & Daim, T. (2011). Using artificial neural network models in stock market index prediction. Expert Systems with Applications, 10389-10397.

[14] Frank, E., Hall, M., & Witten, I. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

[15] Gao, X., Wen, J., & Zhang, C. (2019). An Improved Random Forest Algorithm for Predicting Employee Turnover. Mathematical Problems in Engineering.

[16] Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (The multilayer perceptron) - a review of applications in the atmospheric scenes. Atmospheric Environment, 2627-2636.

[17] Golestani, A., Masli, M., Shami, S. N., Jones, J., Menon, A., & Mondal, J. (2019). Real-Time Prediction of Employee Engagement Using Social Media and Text Mining. IEEE International Conference on Machine Learning and Applications (pp. 1383-1387). Orlando: Institute of Electrical and Electronics Engineers Inc.

[18] Jantan, H., Hamdan, A., & Orhaman, Z. (2011). Talent Knowledge Acquisition using Data Mining Classification Techniques. 3rd Conference on Data Mining and Optimization (DMO)/1st Multi Conference on Artificial Intelligence Technology (MCAIT), (pp. 32-37). Putrajaya, MALAYSIA.

[19] Jiang, J., Chen, J., Choo, K.-K., Liu, K., Liu, C., Yu, M., & Mohapatra, P. (2018). Prediction and Detection of Malicious Insiders' Motivation Based on Sentiment Profile on Webpages and Emails. IEEE Military Communications Conference (pp. 225-230). Los Angeles: Institute of Electrical and Electronics Engineers Inc.

[20] Kriesel, D. (2007). A Brief Introduction to Neural Networks. Retrieved from https://www.dkriesel.com/_media/science/neuronalenetze-en-zeta2-2col-dkrieselcom.pdf

[21] Laijawala, V., Aachaliya, A., & Jatta, H. (2020). Classification algorithms based mental health prediction using data mining. International Conference on Communication and Electronics Systems (pp. 1174-1178). Coimbatore: Institute of Electrical and Electronics Engineers Inc.

[22] Liu, L., Akkineni, S., & Story, P. (2020). Using HR analytics to support managerial decisions: A case study. ACM Southeast Conference (pp. 168-175). Tampa: Association for Computing Machinery, Inc.

[23] Mahmoud, A., Al Shawabkeh, T., Salameh, W., & Al Amro, I. (2019). Performance Predicting in Hiring Process and Performance Appraisals Using Machine Learning. International Conference on Information and Communication Systems (pp. 110-115). Irbid: Institute of Electrical and Electronics Engineers Inc.

[24] Nemes, M., & Butoi, A. (2013). Data Mining on Romanian Stock Market Using Neural Networks for Price. Informatica Economică, 125-136.

[25] Ozdemir, F., Coskun, M., Gezer, C., & Cagri, G. (2020). Assessing Employee Attrition Using Classifications Algorithms. International Conference on Information System and Data Mining (pp. 118-122). Hilo: Association for Computing Machinery.

[26] Rinkaj, G., Pravin, C., & Yogesh, S. (2014). Suitability of KNN Regression in the Development of. International Conference on Future Software Engineering and Multimedia, (pp. 15-21).

[27] Rista, A., Ajdari, J., & Zenuni, X. (2020). Predicting and analyzing absenteeism at workplace using machine learning algorithms. International Convention on Information, Communication and Electronic Technology (pp. 485-490). Opatija: Institute of Electrical and Electronics Engineers Inc.

[28] Santosh, R. S., & Sandeep, K. (2016). A Decision Tree Regression based Approach for the Number of Software Faults Prediction. ACM SIGSOFT Software Engineering Notes.

[29] Shankar, S., Rajanikanth, J., Sivaramaraju, V., & Vssr Murthy, K. (2018). PREDICTION of EMPLOYEE ATTRITION USING DATAMINING. IEEE International Conference on System, Computation, Automation and Networking, ICSCA 2018. Pondicherry: Institute of Electrical and Electronics Engineers Inc.

[30] Soleimanfar, E., Navabakhsh, M., & Sebt, M. V. (2019). Identification of Patterns and Factors Affecting the Health of Employees Based on Datamining of Occupational Examinations with the Purpose of Promoting Occupational Health. Iranian Journal of Health Education and Health Promotion, 295-305.

[31] Šoškić, D. (2006). Hartije od vrednosti: upravljanje portfolijom i investicioni fondovi. Beograd: Centar za izdavačku delatnost ekonomskog fakulteta u Beogradu.

[32] Tanasescu, L.-G., & Bologa, A.-R. (2021). Machine Learning and Data Mining Techniques for Human Resource Optimization Process—Employee Attrition. 20th International Conference on Informatics in Economy (pp. 259-269). Virtual, Online: Springer Science and Business Media Deutschland GmbH.

[33] Tsai, C.-F., & Wu, J.-W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Systems with Applications, 2639-2649.

[34] Wei, Z., Lihong, H., & Liu, T. (2022). Performance Prediction Model Based on K-Means Clustering Algorithm. Lecture Notes on Data Engineering and Communications Technologies (pp. 809-816). Springer Science and Business Media Deutschland GmbH.

[35] Wild Ali, A. (2021). Prediction of Employee Turn Over Using Random Forest Classifier with Intensive Optimized Pca Algorithm. Wireless Personal Communications, 3365-3382.

[36] Witten, I. H., Frank, E., Hall, M. A., & Cristopher, J. P. (2017). Data Mining Practical Machine Learning Tools and Techniques - Fourth Edition. Morgan Kaufmann is an imprint of Elsevier.

[37] Yadav, S., Jain, A., & Singh, D. (2018). Early Prediction of Employee Attrition using Data Mining Techniques. IEEE 8th International Advance Computing Conference (IACC), (pp. 349-354). Bennett Univ, Greater Noida, INDIA.

[38] Yahoo!, I. (2022). Yahoo finance. Retrieved from Yahoo finance: https://finance.yahoo.com

[39] Yigit, I. O., & Shourabizdeh, H. (2017). An Approach for Predicting Employee Churn by Using Data Mining. 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). Malatya, TURKEY.

[40] YongKang, D. (2022). Statistical Analysis and Prediction of Employee Turnover Propensity Based on Data Mining. International Conference on Big Data, Information and Computer Network (pp. 235-23). Sanya: Institute of Electrical and Electronics Engineers Inc.

**SUMMARY**

*Data mining finds various applications when it comes to data related to company employees. In this paper, the goal was to predict the number of permanent employees based on eight different attributes (financial and non-financial) from the data set. The analysis included companies belonging to the S&P500 stock index. With the help of the WEKA tool, four different data mining methods were applied: linear regression, k-NN, multilayer perceptron and decision tree. K-fold cross-validation was used to split the set into training and test parts. The decision tree (MAE and RAE) and linear regression (RMSE and RRSE) provided the best results combined, however, according to the values of the errors, it can be concluded that the models were not very successful in predicting the number of permanent employees. The reason for this may be the selected attributes, the small number of companies selected in the data set or the methods that were used. At the same time, this represents the basis for future research that can be undertaken related to the mentioned problem.*