25 godina
13. SAVETOVANJE O ELEKTRODISTRIBUTIVNIM MREŽAMA sa regionalnim učešćem
13th CONFERENCE ON ELECTRICITY DISTRIBUTION with regional participation
12-16 / 09 / 2022, Kopaonik, Srbija
ZBORNIK RADOVA | PROCEEDINGS
CIRED
R-5.04

POUZDANOST DISTRIBUTIVNOG POSTROJENJA KORIŠĆENJEM KLASIFIKACIONIH ALGORITAMA MAŠINSKOG UČENJA

RELIABILITY OF DISTRIBUTION PLANTS USING CLASSIFICATION ALGORITHMS OF MACHINE LEARNING

Lazar Velimirović, Matematički institut SANU, Srbija

KRATAK SADRŽAJ

Cilj ovog rada je razvijanje modela koji je u stanju da predvidi kvar opreme na osnovu asimetričnog tipa podataka dobijenih od senzora. Nekoliko klasifikacionih algoritama mašinskog učenja je korišćeno za predviđanje otkaza elementa. Koristeći klasifikacione algoritme, bilo je moguće napraviti predviđanje budućih vrednosti jednostavnim unosom trenutnih vrednosti, kao i predviđanje verovatnoće svakog uzorka koji pripada svakoj klasi. Da bi se napravio model predviđanja, korišćen je skup podataka asimetričnog tipa koji sadrži gore pomenute varijable

Ključne reči: pouzdanost, mašinsko učenje, klasifikacioni algoritmi

ABSTRACT

The aim of this paper is to develop a model that is able to predict equipment failure based on the asymmetric type of data obtained from the sensor. Several machine learning classification algorithms have been used to predict element failure. Using classification algorithms, it was possible to predict future values by simply entering current values, as well as predicting the probability of each sample belonging to each class. To create a prediction model, an asymmetric data set containing the above-mentioned variables was used

Key words: reliabiity, machine learning, classification algorithm

Velimirovic.lazar@gmail.com

1. UVOD

Savremeni sistem donošenja odluka treba da optimizuje infrastrukturu sistema za distribuciju, njegovo funkcionisanje, praćenje, održavanje i upravljanje kroz: razvoj sistema pametnog predviđanja, praćenja i predviđanja kvara koristeći modele mašinskog učenja, pametnu analizu podataka za preciznije donošenje odluka u vezi sa pouzdanošću mreže, energetskom efikasnošću i smanjenjem troškova.

Održavanje i nabavka različitih fizičkih senzora može biti skupo, a vrlo mali broj instaliranih senzora radi u mreži [1]. Pošto se nekoliko atributa mreže ne može pratiti onlajn pomoću fizičkih senzora, “soft” senzor se definiše kao model koji je sposoban da predvidi varijable koje je teško izmeriti [2]. Soft senzori mogu pružiti informacije na mreži koje se ne mogu direktno dobiti korišćenjem modela izgrađenog od podataka o obuci dobijenih od fizičkih senzora. Njihov izlaz se može koristiti za onlajn predviđanje nekih varijabli, kontrolu procesa, strategije otkrivanja kvarova ili nadzor hardverskog senzora.

U ovom radu smo razvili model koji je u stanju da predvidi kvar opreme na osnovu podataka asimetričnog tipa dobijenih od senzora. Za predviđanje otkaza opreme korišćeno je nekoliko klasifikacionih algoritama mašinskog učenja, a dokazano je da se čak i sa smanjenim skupom podataka mogu predvideti mogući kvarovi. Dokaze u prilog tvrdnji pružili su eksperimenti koji su pokazali dobre performanse istraživanih modela mašinskog učenja.

Koristeći klasifikacione algoritame, moguće je napraviti predviđanje budućih vrednosti jednostavnim unosom trenutnih vrednosti, kao i predviđanje verovatnoće svakog uzorka koji pripada svakoj klasi. Da bi se napravio model predviđanja, formiran je skup podataka asimetričnog tipa koji sadrži promenljive. Skup podataka

u ovom radu obuhvata podatke jednog pumpnog postrojenja za period od oktobra 2014. do maja 2019. godine. Cilj je bio da se razvijena metodologija implementira u SCADA sistem i ponudi alat za donošenje odluka za održavanje opreme.

2. KLASIFIKACIONI ALGORITMI

U ovom radu su poređena četiri algoritma mašinskog učenja koji se odnose na klasifikaciju podataka: višeslojni perceptron (MLP), stabla za povećanje gradijenta (GBT), "K-nearest Neighbors" (KNN) i "random forest" (RF). Pošto modeli mašinskog učenja za obuku zahtevaju pažljiv izbor hiperparametara, ovaj rad je koristio tehnike procene parametara kako bi se pronašao najbolji skup parametara za naše modele.

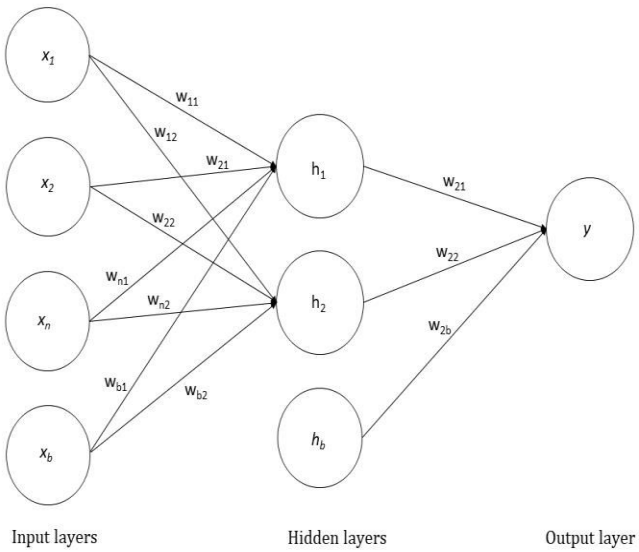
2.1. Multilayer Perceptron (MLP)

Veštačke neuronske mreže su zasnovane na biološkom nervnom sistemu. Biološki nervni sistem se sastoji od velikog broja neurona koji služe kao „jedinica za obradu", ako ga posmatramo iz perspektive računarskog sistema. Neuroni u mozgu su povezani preko sinapsi. Na isti način, veštačke neuronske mreže su sastavljene od slojeva gde svaki sloj sadrži određeni broj neurona koji su povezani sinapsama. Dakle, ANN se sastoji od ulaznog sloja, jednog ili više skrivenih slojeva i izlaznog sloja. Čvorovi iz prethodnog sloja su međusobno povezani sa čvorovima susednog sloja. Jačina veze je predstavljena kroz sinaptičke težine koje predstavljaju stepen korelacije između neurona. U klasifikaciji, izlazni sloj uvek sadrži onoliko čvorova koliko ima klasa (Slika 1).

MLP može da se predstavi kao [3]:

$$y = \sum_{j=1}^{d} w_n x_n + w_b \tag{1}$$

Na osnovu datih težinskih faktora w za ulaz k, može se izračunati izlaz i. Ulazni sloj, skriveni sloj i izlazni sloj su tri sloja koji čine MLP. Ulazni sloj ima zadatak da standardizuje vektor vrednosti prediktorske promenljive u opsegu od −1 do 1 i da ih distribuira svakom od neurona u skrivenom sloju. U skrivenom sloju, kombinovana vrednost se dobija sabiranjem rezultujućih ponderisanih vrednosti, dobijenih množenjem svakog ulaznog neurona sa težinskim faktorom.

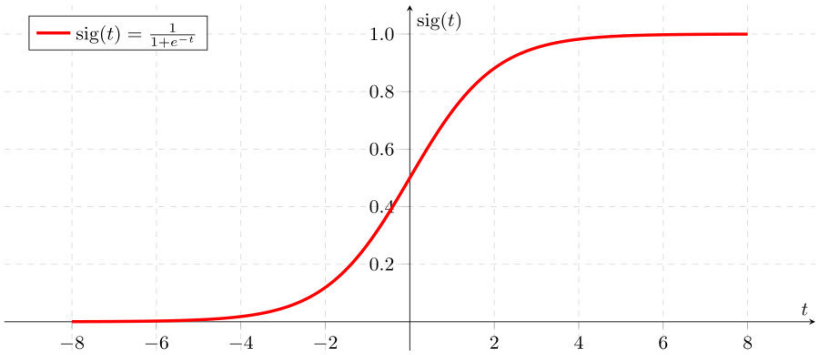


Slika 1. MLP mreža koja se sastoji od dva skrivena sloja. k1, k2, …, kn su ulazne jedinice; h1, h2 su skrivene jedinice; i je izlazna jedinica. vn su težine veza; kb i hb su jedinice pristrasnosti koje uvek imaju vrednost 1.

Ponderisani zbir se unosi u funkciju prenosa nakon čega se dobijaju izlazne vrednosti koje se dalje distribuiraju na izlazni sloj. Kombinovana vrednost u izlaznom sloju se dobija sabiranjem rezultujućih ponderisanih vrednosti dobijenih množenjem težine sa vrednošću svakog neurona skrivenog sloja. Unošenje ponderisane sume u funkciju prenosa daje vrednosti koje predstavljaju izlaze iz mreže.

Za slučaj regresione analize sa kontinualnom ciljnom promenljivom, izlazni sloj se sastoji od jednog neurona koji generiše jednu vrednost, dok se u slučaju klasifikacionih problema izvodi sa kategoričkim ciljnim

varijablama, a izlazni sloj se sastoji od N neurona koji generišu N vrednosti. Za izračunavanje izlaza ANN-a koristi se aktivaciona funkcija. Iz biološke perspektive, informacije prolaze unutar neurona preko akcionog potencijala, koji određuje da li će neuron biti aktiviran ili ne. Iz perspektive ANN-a, funkcija aktivacije određuje da li je neuron aktiviran ili ne. U stvari, funkcija aktivacije transformiše težine primljene od neurona u ulaznom sloju i šalje informacije izlaznom sloju. Postoji mnogo tipova aktivacionih funkcija, uključujući ispravljenu linearnu jedinicu (ReLU), logističku (sigmoidnu), funkciju aktivacije binarnog koraka, hiperbolički tangent (tanh) i druge. U ovom radu je korišćena logistička (sigmoidna) funkcija, kao što je predloženo algoritmom za procenu parametara.



Slika 2. Sigma funkcija.

Logistička (sigmoidna) aktivaciona funkcija (Slika 2) je nelinearna, monotona funkcija i može se predstaviti kao:

$$sig\left(t\right)=\frac{1}{1+e^{-x}} \tag{2}$$

Za mapiranje stvarnih vrednosti u predviđanja, neuronske mreže koriste funkcije optimizacije (gubitaka). Za ove svrhe, naša studija je koristila Adam (adaptivna procena momenta) optimizator. Prvi put je predstavljen u [4] i kombinuje impuls i srednje kvadratno širenje (RMSProp) radi bržeg približavanja. Momentum koristi gradijente prošlih i trenutnih koraka da odredi svoj pravac, dok RMSProp bira različitu brzinu učenja za svaki parametar.

2.2. Random Forest (RF)

RF spadaju u grupu metoda učenja gde se generiše višestruki broj klasifikatora i njihovi rezultati se agregiraju [5]. RF je algoritam stabla odlučivanja i zasnovan je na grupisanju gde se svako drvo konstruiše korišćenjem različitog uzorka za pokretanje, a čvor se deli korišćenjem najboljeg podskupa nasumično odabranih prediktora u tom čvoru. RF radi u nekoliko koraka. Prvo, izdvaja n uzoraka za pokretanje iz skupa podataka asimetričnog tipa. Zatim, za svaki od n uzoraka za pokretanje, razvija se neobrezano (potpuno izraslo) klasifikaciono drvo. Konačno, predviđanje novih tačaka podataka se vrši većinom glasova predviđanja svih stabala.

Stabla odlučivanja dele podatke na način da se dobije najveća informacija. Dobitak informacija identifikuje najvažniju karakteristiku, tj. osobinu koja je najkorisnija za pravljenje predviđanja. Ta funkcija se zatim koristi u osnovnom čvoru. Dobitak informacije je obrnuto povezan sa verovatnoćom pojave posmatranog događaja i može se predstaviti kao [6]:

$$\Delta I=\log\left(\frac{1}{p_i}\right)=-\log\left(p_i\right) \tag{3}$$

gde pi predstavlja verovatnoću klase i.
Dve uobičajene mere nečistoće su entropija i Gini indeks. Entropija se može predstaviti kao [6]:

$$Entropy=-\sum_{i=1}^{n}p_i\log_2\left(p_i\right) \tag{4}$$

Gini indeks se računa pomoću sledećeg izraza:

$$Gini=1-\sum_{i=1}^{n}p_i^2 \tag{5}$$

Ako svi uzorci čvora pripadaju istoj klasi, onda je entropija jednaka nuli. Za probleme binarne klasifikacije, maksimalna entropija je jedan, tj. [6]:

$$\sum_{i=1}^{n} p_i = 1, \quad 0 \le p_i \le 1 \tag{6}$$

Ova studija koristi entropiju kao kriterijum za dobijanje informacija, što će biti dalje objašnjeno u sledećem delu ovog rada.

2.3. Gradient Boosting Trees (GBT)

GBT su stabla odlučivanja koja se koriste za probleme klasifikacije i regresije. Glavna ideja GBT-a je da svako novo stablo treba da minimizira funkciju troškova sve dok ne dođe do poboljšanja. Za razliku od RF koji koristi tzv. "bagging", GBT koristi pojačavanje. Algoritam dodeljuje istu težinu svakom uzorku, a zatim koristi prvi uzorak za obuku prvog slabog klasifikatora [7]. Drugi slabi klasifikator se gradi tako što se pogrešno klasifikovanim uzorcima dodeljuju veće težine, dok se niže težine dodeljuju uzorcima koji su pravilno klasifikovani prvim slabim klasifikatorom [7]. Konačno, klasifikatori su kombinovani u jedan konačni model.
Kvalitet podele u ovoj studiji je meren korišćenjem srednje kvadratne greške (MSE) sa Fridmanovom ocenom poboljšanja. Štaviše, GBT model je imao za cilj da minimizira funkciju gubitka devijacije koja se može izraziti kao [8]:

$$deviance = \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp\left(-2\left(2y_i-1\right)f\left(x_i\right)\right)\right) \tag{7}$$

gde je ki ulaz, ii je izlaz i f(ki) je funkcija ki.

Utvrđeno je da algoritmi za pojačavanje postižu dobre performanse u mnogim različitim oblastima, ali treba napomenuti da veliki prostor karakteristika povećava vreme obuke [9].

2.4. K-Nearest Neighbors (KNN)

KNN je jednostavan algoritam koji se zasniva na meri sličnosti (udaljenosti). Algoritam je „lenji" jer koristi podatke obuke u fazi testiranja, što je računarski efikasnije. KNN prvo izračunava rastojanje između vektora d-dimenzionalnih karakteristika, a zatim vrši klasifikaciju [10]. Iako je KNN prilično jednostavan za korišćenje, na njegove performanse u velikoj meri utiče broj suseda k i izabrana mera udaljenosti [11]. Postoje različite mere sličnosti, uključujući euklidsku distancu, rastojanje Minkovskog i rastojanje Menhetna, ali ova studija je koristila rastojanje Menhetna, kao što je predloženo algoritmom za procenu parametara.
Udaljenost Menhetna (inače poznata kao rastojanje gradskog bloka) može se izračunati koristeći zbir apsolutne razlike između realnih vektora na sledeći način:

$$d\left(x,y\right) = \sum_{i=1}^{n} \left|x_i - y_i\right| \tag{8}$$

,

de je ki i-ti element vektora k, ii je i-ti element vektora i u dvodimenzionalnom vektorskom prostoru i n je broj elemenata u vektoru. Razdaljina "Menheten" tipa je pokazala dobre rezultate u literaturi u poređenju sa drugim merama udaljenosti.

3. PRIMENA KLASIFIKACIONOG ALGORITMA

3.1. Priprema podataka

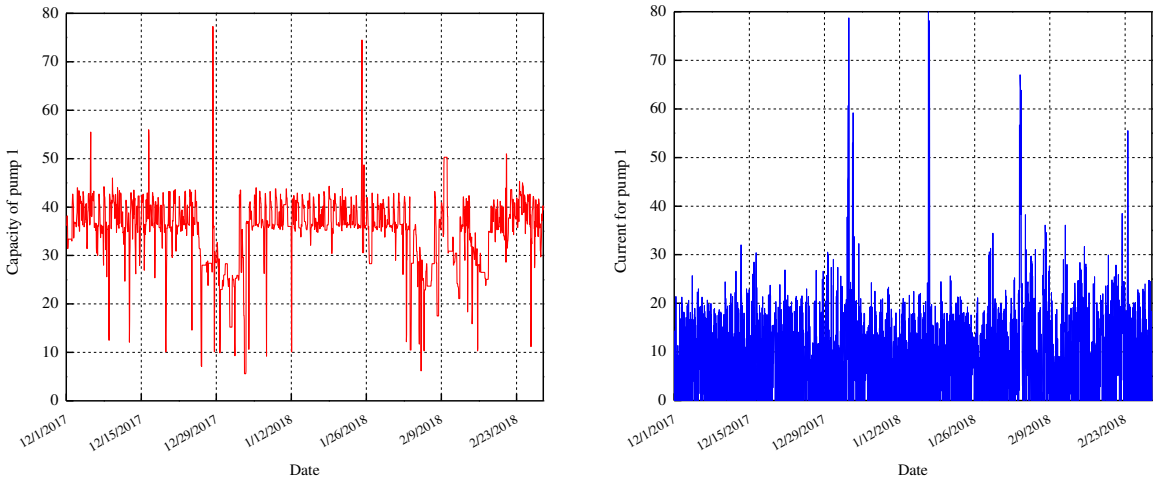
U ovom radu su analizirani podaci jednog složenijeg postrojenja u distributivnoj mreži - pumpne stanice. Skup podataka korišćen u ovoj studiji sastojao se od senzorskih asimetričnih podataka iz tri pumpe za vodu. Koristeći posmatrane podatke iz procesa, kao što su trenutni nivo vode u rezervoaru i dotok, parametri pumpe, kao što su vrednost struje i vreme rada, mogu se povezati i povezati sa neelektričnim podacima. Predviđanja se mogu napraviti korišćenjem tehnika klasifikacije ili regresije, u zavisnosti od tipa ciljne varijable. Da bi se napravio model predviđanja, formiran je skup podataka asimetričnog tipa koji sadrži promenljive. Zbog razlike između varijabli u pogledu vremena merenja, svi podaci su uprosečeni u satima. Redovi koji sadrže nedostajuće vrednosti su uklonjeni tokom faze predobrade, kao i kolona datuma i vremena.
Nedostajuće vrednosti su uklonjene pre modeliranja, pa se konačni skup podataka asimetričnog tipa sastojao od 25.625 merenja (70% podataka pripada skupu za obuku, 30% podataka pripada skupu testova). Podaci su standardizovani korišćenjem StandardScaler biblioteke scikit. StandardScaler uklanja srednju vrednost i skalira podatke na jediničnu varijansu tako što izračunava z rezultat kao:

$$z = \left(x-\mu\right)/s \tag{9}$$

gde μ predstavlja srednju vrednost uzoraka za obuku, a s je standardna devijacija.

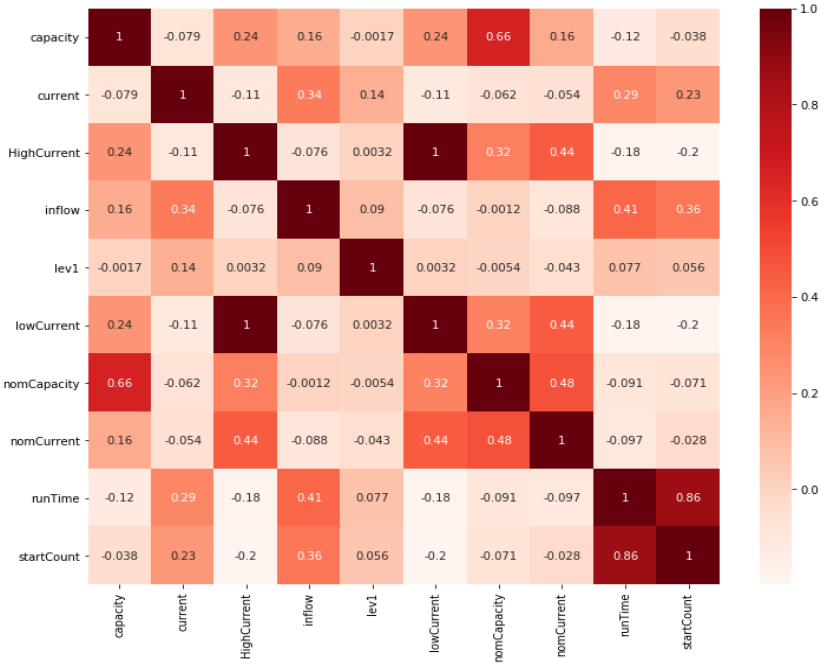
Ulazne varijable su uključivale kapacitet pumpe, struju, veliku struju, nisku struju, nominalnu struju, nominalni kapacitet, dotok, vrednost nivoa, vreme rada i broj pokretanja motora. Izlazna varijabla je bila binarna sa dve klase koje nisu bile simetrične — klasa nula koja predstavlja negativne slučajeve, tj. bez alarma, i klasa jedan koja predstavlja pozitivne slučajeve, tj. alarm). Ako postoji neravnoteža tokom faze predobrade, podaci o obuci treba da se transformišu korišćenjem kombinacije tehnika nedovoljno uzorkovanja i prekomernog uzorkovanja.

Koristeći posmatrane podatke iz procesa, kao što su trenutni nivo vode u rezervoaru i dotok, parametri pumpe, kao što su vrednost struje i vreme rada, mogu se povezati i povezati sa neelektričnim podacima. Primer ovih vrednosti uzetih iz realnog sistema dat je na slici 3.



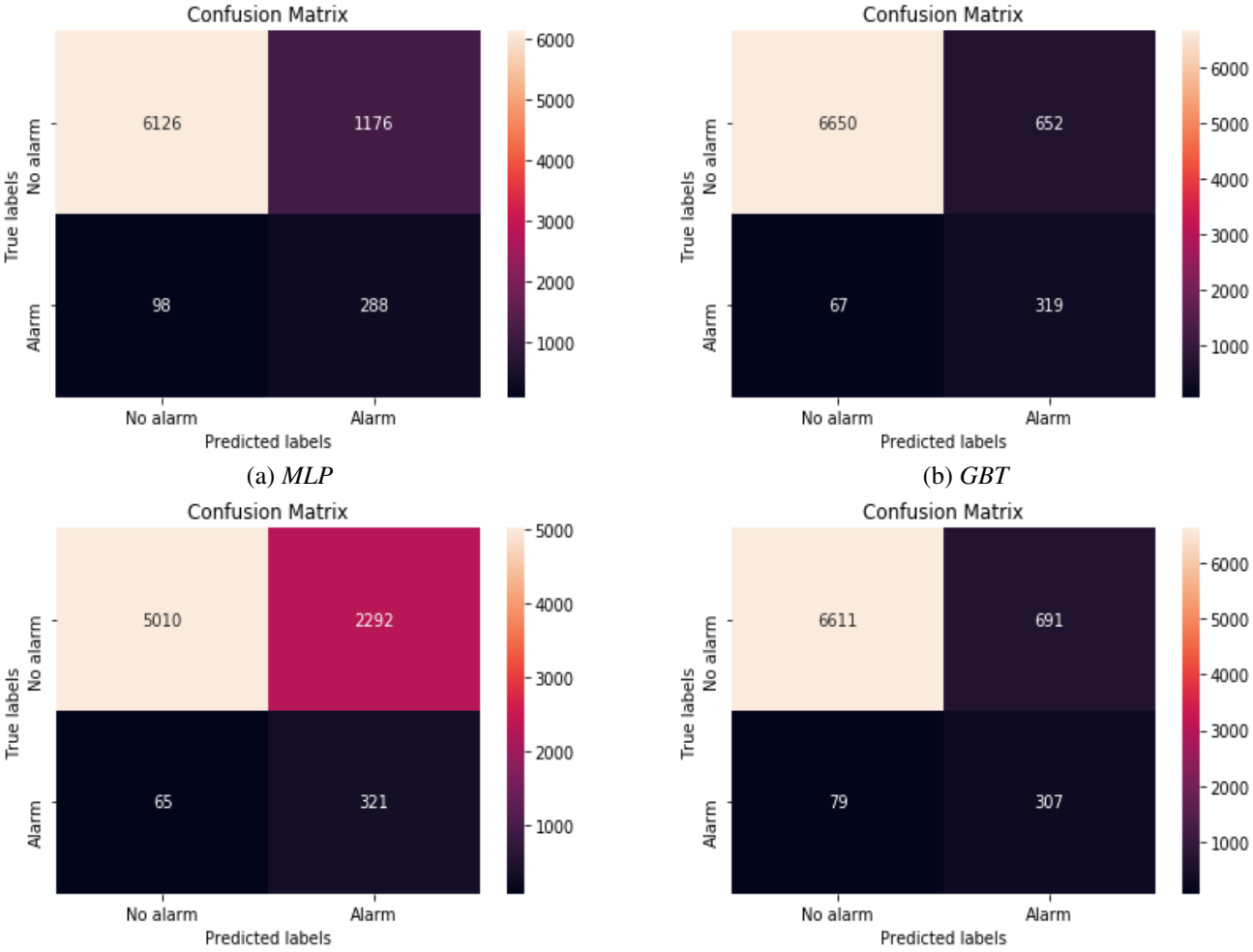
Slika 4 Vremenski dijagrami protoka i struje pumpe

Pre bilo kakve simulacije, potrebno je generisati matricu korelacije kako bi se posmatrao odnos između varijabli. Ako je izlazna promenljiva binarna, ona neće biti objekat korelacionog testa. Korelacija može pružiti neke vredne uvide u odnose između nezavisnih varijabli i može se izračunati korišćenjem Pirsonovog koeficijenta korelacije. Pirsonov koeficijent korelacije može biti pozitivan i negativan i kreće se od -1 do 1, gde vrednosti bliže nuli ukazuju na nepostojanje veze između posmatranih varijabli, a vrednosti bliže ±1 ukazuju na jaku vezu između varijabli (snaga, struja, struja reagovanja prekostrujne zaštite, protok, nivo, minimalna vrednost struje, nominalna snaga, nominalna struja, vreme rada, broj pokretanja (slika 5).



Slika 5 Korelaciona matrica

Da bismo bolje razumeli performanse ispitivanih klasifikatora i njihovu sposobnost da ispravno klasifikuju kvarove pumpi, generisana je matrica nedoumica ("confusion matrix") 2×2. Kolone pokazuju broj predviđenih test uzoraka za negativnu klasu (bez alarma) i pozitivnu klasu (alarm), dok redovi predstavljaju pravi broj test uzoraka koji pripadaju posmatranim klasama.



(a) *MLP* (b) *GBT*

(c) *KNN* (d) *RF*

Slika 6 Matrica nedoumica za svaki model

Na test setu, 6126 zaista negativnih uzoraka je tačno klasifikovano kao negativno (stvarno negativno) po MLP modelu, dok je 288 zaista pozitivnih uzoraka ispravno klasifikovano kao pozitivno (stvarno pozitivno). Bilo je 1176 lažno pozitivnih uzoraka (uzorci su predviđeni kao pozitivni, ali su zaista negativni) i 98 lažno negativnih uzoraka (predviđeni kao negativni, ali su zapravo pozitivni). S obzirom da je namera bila da se klasifikuju otkazi pumpe, bilo je važno imati niske vrednosti lažno negativnih uzoraka jer bi posledice u suprotnom mogle biti teške; ako se predviđa da uzorak nije kvar, ali je zaista došlo do kvara pumpe, postojali bi veliki troškovi koji bi se mogli izbeći blagovremenim predviđanjem i upravljanjem. Štaviše, povećanje veličine skupa podataka sa uzorcima koji pripadaju pozitivnoj klasi dalo bi više stvarnih informacija algoritmu što će značajno povećati performanse modela.

4 ZAKLJUČAK

Na osnovu dobijenih rezultata može se zaključiti da se ML algoritmi mogu uspešno koristiti za predviđanje kvara pumpe. Da bi se generisala detaljnija i tačnija predviđanja, važno je imati više podataka, posebno podataka o alarmnim slučajevima, što će povećati distribuciju ove klase i dati više informacija algoritmu, čime se dobija tačnije predviđanje.

Predviđanja kvarova pumpi sugerišu neke zanimljive nalaze. Na prvom mestu, RF tehnika može se uspešno koristiti za klasifikaciju senzorskih podataka. Model je prikladan za fino podešavanje, ali takođe vremenom poboljšava svoje performanse. Nadalje, razvijeni modeli sugerišu da se performanse značajno poboljšavaju sa većim brojem realnih vrednosti u manjinskoj klasi. Iako tehnike uzorkovanja (kao što je nedovoljno uzorkovanje

i prekomerno uzorkovanje) rešavaju problem neravnoteže klasa, model je u stanju da mnogo bolje uči na podacima iz stvarnog sveta, stoga sa dovoljno velikim brojem slučajeva u manjinskoj klasi, predviđanja su svakako veća.

Cilj budućeg rada je unapređenje razmatranih algoritama formiranjem što većih baza podataka svih podataka koji čine model kako bi se povećala tačnost predviđanja. Dodatno, cilj budućeg rada je da primenom razmatranih algoritama sistem identifikuje obrasce i pomogne u donošenju preciznijih odluka u vezi sa otpornošću mreže, energetskom efikasnošću, minimiziranjem otpadnih voda, smanjenjem troškova itd. Dalji razvoj ovog rešenja će optimizovati sisteme distribucije vode. , njihovu infrastrukturu, rad, praćenje, održavanje i upravljanje.

ZAHVALNICA

Ovaj rad je objavljen uz podršku Ministarstva prosvete, nauke i tehnološkog razvoja Matematičkom insitutu Srpske akademije nauka i umetnosti.

LITERATURA

[1] Vanrolleghem, P.; Lee, D. On-line monitoring equipment for wastewater treatment processes: State of the art. Water Sci. Technol. 2003, 47, 1–34, doi:10.2166/wst.2003.0074.
[2] alexandridis, A. Evolving RBF neural networks for adaptive soft-sensor design. Int. J. Neural Syst. 2013, 23, 1350029, doi:10.1142/s0129065713500299
[3] Alpaydin, E. Introduction to Machine Learning, 2nd ed; The MIT Press: Cambridge, MA, USA, 2014.
[4] Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. arXiv 2014, arXiv:1412.6980.
[5] Liaw, A.; Wiener, M. Classification and regression by randomForest. R News 2002, 2, 18–22.
[6] Shannon, C.E. A mathematical theory of communication. Bell Syst. Tech. J. 1948, 27, 379–423.
[7] Tian, Z.; Xiao, J.; Feng, H.; Wei, Y. Credit Risk Assessment based on Gradient Boosting Decision Tree. Procedia Comput. Sci. 2020, 174, 150–160, doi:10.1016/j.procs.2020.06.070.
[8] Truong, V.-H.; Vu, Q.-V.; Thai, H.-T.; Ha, M.-H. A robust method for safety evaluation of steel trusses using Gradient Tree Boosting algorithm. Adv. Eng. Softw. 2020, 147, 102825, doi:10.1016/j.advengsoft.2020.102825.
[9] Kocsis, L.; György, A.; Ban, A.N. BoostingTree: Parallel selection of weak learners in boosting, with application to ranking. Mach. Learn. 2013, 93, 293–320, doi:10.1007/s10994-013-5364-5.
[10]Zhang, S.; Li, X.; Zong, M.; Zhu, X.; Cheng, D. Learning k for kNN Classification. ACM Trans. Intell. Syst. Technol. 2017, 8, 1–19, doi:10.1145/2990508.
[11] Zhang, S.; Li, X.; Zong, M.; Zhu, X.; Wang, R. Efficient kNN Classification With Different Numbers of Nearest Neighbors. IEEE Trans. Neural Networks Learn. Syst. 2018, 29, 1774–1785, doi:10.1109/tnnls.2017.2673241.