Predikcija visine temperature i analiza uticaja vremenskih prilika na broj nezgoda u Beogradu

Ivana Tomić
Računarstvo i automatika
Fakultet tehničkih nauka, Univerzitet u Novom Sadu
ivana1996tomic@gmail.com

Mihajlo Levarski
Računarstvo i automatika
Fakultet tehničkih nauka, Univerzitet u Novom Sadu
mlevarski@gmail.com

Abstract— **Ovaj rad se bavi predikcijom visine temperature, uticaja vremenskih uslova na saobraćajne nezgode na teritoriji Beograda, kao i predikcijom vrste saobraćajne nezgode. Cilj ovog istraživanja je automatizacija predikcije visine temperature kao i smanjenje broja saobraćajnih nezgoda. Kreiran je za potrebe predmeta Sistemi za istraživanje i analizu podataka i Neuronske mreže. Uspešnost predikcije je predstavljena sa nekoliko različitih modela mašinskog učenja i to linearna regresija, DNN, MLP, RFR, LSTM, GRU, SVM, RFC, naivni Bajes, CNN. Rezultati za predikciju temperature su dobijeni korišćenjem istorijskih podataka vremenskih uslova sa teritorije Beograda i gradova iz regije. Nakon sređivanja prikupljenih podataka, trenirani su modeli, a zatim i izvršena evaluacija dobijenih rezultata. Na osnovu evaluacije je uočljivo da je za ovaj problem najpogodniji MLP model i model slučajne šume (RFR), kao i LSTM i CNN. Podaci vezani za vremenske uslove na teritoriji Beograda su spojeni na osnovu datuma i vremena sa podacima o saobraćajnim nezgodama za 2018. godinu. Nakon toga, atributi od interesa su grafički prikazani, analizirani, a potom je odrađena predikcija vrste saobraćajne nezgode. Analizom smo došli do zaključka koji vremenski uslovi i drugi faktori imaju najveći uticaj na broj saobraćajnih nezgoda. Na osnovu rezultata za predikciju vrste saobraćajne nezgode smo zaključili da su gotovo podjednako dobri RFC i SVM. Predstavljeno je osnovno rešenje i analiza, kao i neke mogućnosti proširenja i unapređenja.**

Keywords— vremenska prognoza, predikcija, analiza, saobraćajne nezgode

I. UVOD

Napredak tehnologije je vidljiv na gotovo svim poljima. Jedna od posledica toga je olakšavanje poslova. Većina tradicionalnih procesa je zamenjena procesima koje obavljaju računari. U ovom radu je opisano rešenje koje daje primenu tehnologije u predikciji visine temperature na osnovu istorijskih podataka na teritoriji Beograda i gradova iz regiona (Budimpešta, Sofija , Skoplje , Bukurešt , Podgorica, Sarajevo, Zagreb). Poznato je koliko je neophodno da se informišemo o vremenskim prilikama u budućnosti kako zbog nekih sitnih potreba, tako i zbog kompleksnijih stvari. Međutim, moramo biti svesni da je zbog velikog broja faktora koji utiču na vremenske prilike ovaj proces veoma složen. Zbog toga, da bi bio što tačniji, zahteva domensko znanje iz ove oblasti.

Ideja ovog rešenja je bila da korišćenjem raznih metoda i istorijskih podataka za ovaj rad, kao i istorijskih podataka gradova iz regiona dobijemo neke više ili manje dobre rezultate koji će biti opisani u nastavku rada.

Pored rešenja vezanog za predikciju visine temperature u Beogradu, u ovom radu je data i analiza uticaja vremenskih prilika na saobraćajne nezgode, kao i predikcija vrste saobraćajne nezgode. Analiza i predikcija su urađene na osnovu podataka o saobraćajnim nezgodama za 2018. godinu. Shodno tome, korišćeni su i podaci o vremenskim prilikama za ovu godinu na teritoriji Beograda. Ovo je još jedna od mogućnosti primene novih tehnologija. Analizom ovih podataka dolazimo do bitnih zaključaka koji nam mogu pomoći da se u budućnosti stvari poput saobraćajnih nezgoda manje dešavaju. Odnosno, iz ovih podataka možemo zaključiti kada su neophodne veće mere predostrožnosti, bolje vidljivost i prilagođavanje brzine. Pored toga, kombinacijom ovih obeležja uradili smo predikciju ishoda saobraćajnih nezgoda.

Kako na stanje na putevima i uslovima u saobraćaju visina temperature ne predstavlja jedini validan parametar, u analizi su korišćeni i opisi vremenskih prilika na primer „oblačno", „vedro" itd.

Nakon uvoda je, u drugom odeljku, dat pregled srodnih istraživanja, nakon toga sledi opis skupa podataka. Zatim, u četvrtom odeljku, je dato objašnjenje metodologije korišćenje u ovom istraživanju. U petom odeljku je data analiza podatka, a u šestom su opisani rezultati predikcije. Sedmi odeljak donosi zaključna razmatranja i pravce budućih istraživanja.

II. SRODNA ISTRAŽIVANJA

U nastavku će biti opisani radovi koji se bave predikcijom visine temperature, a zatim I radovi u kojima je izvršena analiza uticaja vremenskih prilika na saobraćajne nezgode.

Rad [1] rešava problem predviđanja vremena na osnovu istorijskih podataka meteoroloških stanica iz ciljanog područja kao i susednih područja . Podaci su prikupljeni za grad Nešvil i okolne gradove . Jedinstveni zapisi vremenskih prilika u određeno vreme iz odrđenih područja . Svaki zapis sadrži podatke o temperaturi, vlazi , smeru vetra, atmosferskom pritisku i atmosferskom stanju . Ukupno 68

zapisa za svaki dan od 1. jula 2018. do 7. septembra 2018. Ciljna varijabla zapisa je temperatura , tako da se predikcija za naredni dan vrši na osnovu posmatranja vremenskih prilika tekućeg dana u okviru istog vremenskog perioda . Metode korišćene u radu : Linearna regresija , Support Vector Regressor, Multi-Layer Perceptron Regressor (MLPR) , Random Forest Regressor (RFR) , Extra-Tree Regressor (ETR) . Pokazano je da ovi modeli mašinskog učenja mogu predvideti vremenske karakteristike dovoljno dobro . Zaključeno je da upotreba istorijskih podataka okolnih područja u odnosu na ciljano područje pokazuje efikasnije i preciznije predviđanje vremenske prognoze . Primećeno je da su tačke na grafu mnogo više rasute kad se koriste podaci vezani za jedan grad u odnosu na podatke iz 10 gradova , što govori da je mogućnost za greškom mnogo manja u drugom slučaju . Na osnovu ovih rezultata možemo primeniti istu vrstu metodologije na podatke našeg istraživanja . Trening skup sadrži podatke u periodu od dva meseca (počevši od 1. jula 2018.) dok test skup sadrzi podatke u periodu od sedam dana (od 1. do 7. septembra 2018.). Za procenu modela korišćen je koren srednje kvadratne greške.

U radu [2] , Holmstrom et al. primenjuju mašinsko učenje u prognoziranju vremena kako bi potencijalno stvorili tačnije vremenske prognoze za duže vremenske periode s obzirom da su te tehnike otpornije na smetnje u odnosu na fizičke modele atmosfere . Prikupljeni podaci na dnevnom nivou (maksimalna i minimalna temperatura , srednja vlaga , srednji atmosferski pritisak) u periodu od 2011. do 2015. za mesto Stanford u Kaliforniji . Što znači da sadrži 1826 zapisa , svaki se odnosi na pojedinačan dan. Podaci iz prve četiri godine korišćeni su za obučavanje algoritama, a podaci iz poslednje godine korišćeni su za test. Podaci klasifikovani na vedro , umereno oblačno, veoma oblačno i padavine . Koristili su metode linearne regresije kao i varijaciju funkcionalne regresije . Računali su srednju kvadratnu grešku za rezultate dobijene linearnom regresijom, varijacijom funkcionalne regresije, ali i za profesionalnu prognozu vremena . Rezultati pokazuju da linearna regresija kao model ne zavisi toliko od samih parametara, koliko od količine informacija u odnosu na funkcionalnu regresiju koja pokazuje bolje performanse sa pogodno odabranim parametrima.

Abrahamsen et al. u radu [3] razvijaju model za predviđanje temperature u periodu od jedan , tri , šest ili dvanaest sati. Izvršeni su dva eksperimenti u kojima su korišćeni AR – NN (auto-regressive neural network), ARX-NN (auto-regressive with exogenous input) kao i Multi Layer Perceptron (MLP). Podaci su prikupljeni u periodu od 2016. do 2017. sa meteorološke stanice koja se nalazi u Porsgrunu, Norveška. U prvom primeru se koristi samo temperatura kao podatak, a u drugom temperatura i padavine . Eksperimentima zaključeno da se greška povećava srazmerno u odnosu na broj sati za koji se predikcija temperature vrši . Ukoliko se u skup podataka uključe i dodatni podaci , kao što su informacije o padavinama , primećuju se poboljšanja u performansama. Podaci su podeljeni tako da 60% ide na obučavanje, 20% na validaciju i 20% na testiranje . Proračun greške pomoću srednje kvadratne greške. Dobijeni rezultati su predstavljeni grafički sa realno izmerenim temperaturama u određeno vreme. Posmatranjem prikazanih rezultata testiranja dolazimo do zaključka da se greška povećava srazmerno sa povećanjem broja sati za koje pravimo predikciju temperature.

U radu [4] je detaljno prikazan izveštaj koji istražuje veze između različitih faktora koji mogu biti povezani sa saobraćajnim nezgodama na teritoriji Ujedinjenog Kraljevstva. Pored vremenskih uslova, autor uzima u obzir i vrstu puta, uslove osvetljenja itd. Korišćeni su podaci za 2016. godinu koji su smešteni u tri tabele. Prva tabela je vezana za podatke o nesrećama i sadrži 32 kolone i ukupno 136621 zapisa . Neke od ovih kolona su: lokacija, vreme, datum, osvetljenje, vremenski uslovi i stanje na površini puta, broj učesnika, tip puta i druge promenljive. U drugoj tabeli je dat detaljniji opis žrtava. Ova tabela ima 16 kolona i 181384 redova. Svaki red u ovoj tabeli sadrži podatke o jednoj povređenoj osobi u saobraćajnoj nezgodi. Treća tabela daje detaljniji opis o vozilima koja su učestvovala u saobraćajnim nezgodama. Pored toga, koriste se i vladine procene pređenih razdaljina na različitim vrstama puta u Velikoj Britaniji, relativna gustina drumskog saobraćaja na svakih sat vremena svakog dana u nedelji i podaci o stanovništvu u Velikoj Britaniji na nivou lokalne uprave. Korelacija između obeležja izračunata je pomoću Pirsonovog i Spirmanovog koeficijenta. Koristili su metode višestruke linearne regresije kao i klasifikacije (K-najbližih suseda). Autor je imao očekivanja da će vožnja u mraku i po lošijim vremenskim uslovima biti prikazana kao opasna. Odnosno, da će rezultati pokazati da je npr. viša stopa sudara tokom zime i na područjima na kojima je uglavnom hladnije. Međutim, rezultati ne dokazuju ovu pretpostavku. Što se tiče vremenskih uslova, rezultati pokazuju suprotno. Autor je doneo zaključak da loše vreme obeshrabruje ljude da voze, što dovodi do manje vozača, manje zagušenih puteva i niže brzine, pa samim tim i manjeg broja nezgoda. Broj nezgoda se ne može vezati za neki mesec ili period u godini, jer je broj na mesečnom nivou sličan.

Rad [5] nam daje detalje istraživanja koje pruža različite analize podataka o saobraćajnim nezgodama koje su se dogodile na teritoriji države Katar . Korišćeni su podaci na mesečnoj bazi u periodu od 7 uzastopnih godina , u rasponu od 2010. do 2016. godine. Zahvaljujući tome, omogućeno je poređenje frekvencija u određenom vremenskom periodu. Metoda analize pogodna u ovom slučaju je analiza vremenskih serija , kao sto je dekompoziciona metoda. Uporedo sa analizom dekompozicije vremenskih serija, koristi se i multivarijantna analiza varijanse, kako bi se istražila veza između sezonskih varijacija vremena i broja saobraćajnih nezgoda. Rezultati analize su otkrili da se na vozače i pešake koji su učestvovali u sudaru sa teškim povredama ili smrtnim ishodom značajno utiču sezonske promene vremena i to najviše u zimskoj i jesenjoj sezoni. Letnja sezona se navodi kao mirnija jer malo pešaka izlazi i šeta ulicama zbog temperatura koje su nepodnošljive, za razliku od jesenskog i zimskog perioda kada je broj pešaka u porastu zbog prijatnijih i blažih temperatura. Navodi se da jesenji i zimski period karakteriše magla koja se javlja u zoru i rano jutro zbog severoistočnih ili jugoistočnih vetrova koji uzrokuju vlažnost, što takođe utiče na broj saobraćajnih nezgoda. U ovoj analizi saobraćajne nesreće su klasifikovane u četiri različite grupe i to: materijalna šteta, lakše povrede, teške povrede i smrtni ishod

U radu [14] je dat opis primene metoda mašinskog učenja kako bi se izvršila predikcija procene rizika u saobraćajnim nezgodama. Skup podataka se sastoji od evidencija saobraćajnih nezgoda koji su se dogodile u gradu Porto Alegre, Brazil tokom 2013. godine i koje su opisane sa 44 obeležja. U skupu podataka se nalazi 20798 zapisa. Zbog dupliranih i nevažećih podataka je bilo neophodno sređivanje skupa. Skup podataka je podeljen na trening i test skup. Trening podaci čine 60%, a test podaci 40% od ukupnog broja. U ovom rešenju su upotrebljeni sledeći algoritmi: logička regresija, metoda potpornih vektora (SVM - Support Vector Machines), naivni Bajes klasifikator, K-najbližih suseda, slučajna šuma (*Random Forests*). Korišećene mere kvaliteta modela su: površina ispod ROC (*Receiver Operating Characteristic*) krive odnosno AUC (*Area Under the Curve*), zatim preciznost, odziv i F1 meru (harmonijska sredina preciznosti i odziva). Rezultati su prikazani tabelarno i u vidu grafika. Njihovom analizom se dolazi do zaključka da se modeli predviđanja za procenu rizika od povreda mogu kreirati sa velikom preciznošću. Najbolje i gotovo iste rezultate su pružile logistička regresija i metoda potpornih vektora, a zatim slučajna šuma.

U radu [18] je opisano rešenje koje koristi rekurentne neuronske mreže za predviđanje temperature. Koriste se arhitektura od dva sloja od kojih je jedan LSTM. Drugi sloj je gusti sloj koji je vezan sa LSTM. Predviđanje se vrši na osnovu podataka koji su preuzeti sa *IoT* stanica koje imaju senzore za temperaturu, vlažnost i neke od gasova. Dobijeni podaci se porede sa podacima koji su stvarno izemereni. LSTM ulazni sloj se sastoji od osam neurona, sadrži tri skrivena sloja od po 100 neuroan i jedan izlazni neuron.

III. Skup podataka

U ovom poglavlju će biti opisan skup podataka koji je korišćen u predikciji i analizi.

Skupovi podataka vremenskih prilika preuzeti su sa [7] . Podaci su nastali u meteorološkoj stanici Beograd/Batajnica . Pored podataka koji su zabeleženi na ciljnoj lokaciji, skup sadrži i podatke zabeležene u sledećim gradovima: Budimpešta, Sofija, Skoplje, Bukurešt, Podgorica, Sarajevo, Zagreb. Istorijski podaci prikupljeni su tehnikama preuzimanja i parsiranja HTML koda pomoću *BeautifulSoup* API-ja . Za svaki od gradova povučene su informacije od 1. januara 2018. do 1. januara 2020. godine u intervalu od 30 minuta i pretvoreni u CSV datoteke radi dalje obrade . Skupovi sadrže sledeće atribute: datum, vreme , temperatura, osećaj temperature, vetar, udari vetra, relativna vlažnost, tačka rošenja, pritisak, ikonica i opis.

Analizom skupova odlučeno je da se zbog nekonzistentnosti zapisa određenih atributa upotrebe samo sledeći atributi: datum i vreme , temperatura, vlažnost vazduha i pritisak. Daljom analizom podataka uočeno je da nemaju svi gradovi zapise o vremenskim prilikama u određenom vremenskom intervalu ili trenutku . Kako bi uskladili podatke kao i vremenske trenutke u kojima su zapisi validni, višak podataka susednih gradova je izbačen i usklađen u odnosu na podatke za Beograd. Spajanjem i sortiranjem svih podataka u odnosu na datum i vreme dobijamo ukupno 25025 zapisa .

Skup podataka saobraćajnih nezgoda preuzet je sa [8] u okviru ODS datoteke. Podaci se odnose na teritoriju Beograda u toku 2018. godine. Sadrže sledeće atribute: identifikator, datum i vreme, tip, broj vozila kao i opis saobraćajne nezgode. Tip saobraćajne nezgode može biti sa materijalnom štetom, povređenim ili poginulim. Broj vozila se odnosi na vozilo u pokretu, parkirano vozilo ili pešaka. Analizom podataka uklonjeni su zapisi vezani za atribute opis i identifikator, s obzirom na nekonzistentnost i zanemarljiv doprinos u analizi uticaja vremenskih prilika na saobraćajne nezgode.

Kategoričke varijable su pretvorene u numeričke. Nakon sređivanja podataka, ukupan broj zapisa iznosi 18064.

Da bi izvršili analizu uticaja vremenskih prilika na saobraćajne nezgode, kao i predikciju vrste saobraćajne nezgode spojili smo podatke iz oba skupa. Koristili smo podatke o vremenskim prilikama na teritoriji Beograda koje uključuju sledeće atribute: datum, vreme, temperatura, pritisak, vlažnost i opis vremenskih prilika. Podaci su pridruženi po datumu i vremenu uz pomoć Pandas API funkcije koja spaja redove na osnovu najpribližnijih vrednosti za zajedničko obeležje.

Nakon obrade, skup podataka vremenskih prilika u Beogradu i okolnim gradovima podeljen je na podskupove. Trening skup sadrži 80% podataka, dok test skup sadrži 20%. Primenom skaliranja vrši se distribucija vrednosti promenljivih između -1 i 1, pri čemu je srednja vrednost blizu 0. Na ovaj način smanjujemo cenu obuke modela, utičemo na performanse algoritma.

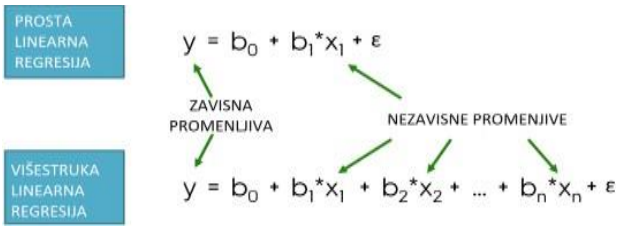
IV. Metodologija

A. *Korišćeni algoritmi :*

Dve osnovne vrste problema nadgledanog učenja , koje rešavamo , su regresija i klasifikacija. Regresija je problem predviđanja kontinualne ciljne promenljive. Klasifikacija je problem predviđanja kategoričke ciljne promenljive koja uzima konačan broj vrednosti [15] .

Linearna regresija je metod koji nam omogućava proučavanje veze između nezavisnih i zavisnih promenljivih. Ulaz nam predstavlja vektor obeležja *Xi* (nezavisne promenljive), odnosno podaci koji utiču na predviđanje ciljnog obeležja. *Y* (zavisna promenljiva) je varijabla ishoda, odnosno ciljna promenljiva [6]. U ovom rešenju ciljna promenljiva je visina temperature u Beogradu, dok ostali podaci predstavljaju vrednosti sadržane u vektoru *Xi*. Linearna regresija, u zavisnosti od broja nezavisnih promenljivih, može biti prosta ili složena. Prosta linearna regresija ima jednu nezavisnu promenljivu, a višestruka dve ili više. U ovom rešenju je korišćena višestruka linearna regresija. Na slici 1 su prikazane formule proste i višestruke linearne regresije sa obeleženim zavisnim i nezavisnim promenljivama. Parametri

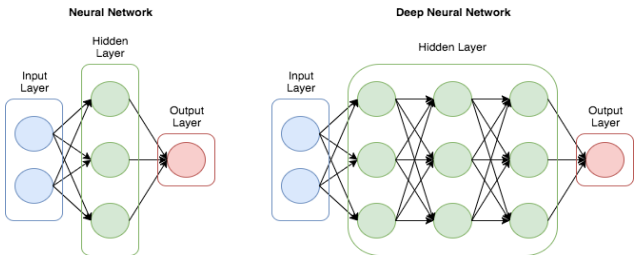
od b_0 do b_n, odnosno b_1 u slučaju proste linearne regresije, označavaju nepoznate parametre koje treba oceniti. Greška merenja, odnosno rezidual je ε.



Slika 1. Formule proste i složene linearne regresije

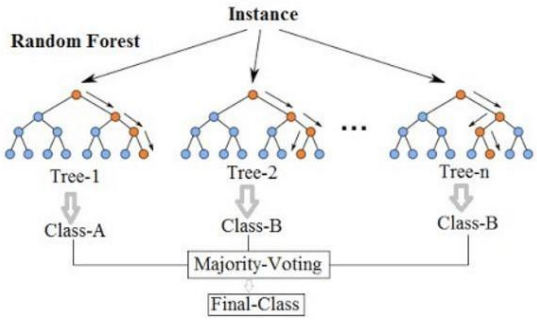
Duboke neuronske mreže (DNN- *Deep Neural Network*) se mogu posmatrati kao oblik veštačke neuronske mreže sa više skrivenih slojeva (slika 2). Neuronska mreža ima ulazni i izlazni sloj neurona. Slojevi između ulaznog i izlaznog se nazivaju skriveni slojevi i duboke neuronske mreže moraju imati više ovakvih slojeva. Broj neurona u slojevima može biti različit. U ovom rešenju su za DNN korišćene različite kombinacije broja neurona i slojeva. Kao najbolja od isprobanih se pokazala kombinacija četiri sloja sa sledećim brojem neurona: 256, 128, 64 i 32. Svaki sloj vrši određene vrste sortiranja i uređivanja podataka, pa su zbog toga pogodne za bavljenje sa neobeleženim ili nestrukturiranim podacima. U poređenju sa plitkim arhitekturama, duboke su efikasnije kada je u pitanju broj parametara i računarskih elemenata za predstavljanje funkcija [9]. Duboke neuronske mreže mogu koristiti petlje. Višeslojni perceptron (MLP - *Multilayer perceptron*) je veštačka duboka neuronska mreža koja na osnovu skupa ulaznih promenljivih generiše vrednosti izlaznih. MLP povezuje više slojeva u usmerenom grafu, što znači da put signala kroz čvorove ide samo u jednom pravcu, odnosno nema povratnih veza [11]. Koristi *Backpropagation* algoritam koji može da podešava težine u skrivenim slojevima. Svaki izlaz iz jednog sloja se koriste kao ulazi za naredni sloj (slika 2), ovo ne važi za izlazni sloj gde izlazi predstavljaju krajnje vrednosti. MLP je dao najbolje rezultate u kombinaciji sa tri sloja od 256, 128 i 64 neurona. U DNN i MLP modelima je korišćena ReLU (*Rectified Linear Unit*) aktivaciona funkcija.

ReLU uvek vraća vrednost koja je veća ili jednaka nuli. Ona je trenutno jedna od najpopularniji activation funkcija za neurone u skrivenom sloju jer se lako računa, a samim tim dobijamo i brže obučavanje. Izlaz iz neuronske mreže je jedan i predsavlja naše ciljno obeležje, odnosno temperaturu, a imamo 23 ulaza koji predstavljaju ostala obeležja iz opisanog skupa podataka.



Slika 2. Primer neuronske mreže i duboke neuronske mreže (Preuzeto sa [10])

Slučajna šuma (RF - *Random Forest*) je algoritam nadgledanog učenja koji koristi ansambl odlučujućih stabala (*Decision tree*) kao metode učenja za rešavanje klasifikacionih i regresionih problema . Ona koristi tehniku pakovanja (*Bootstrap aggregation*) kao jednu od ansambl metoda . Na taj način obuhvata više stabala odlučivanja koji su pokrenuti paralelno i ne interaguju međusobno . Svako od stabala vrši predikciju koje se na kraju objedinjuju za postizanje konačnog rezultata , prikazano na slici 3 . Ukoliko je u pitanju klasifikacija , izvodi se modalitet svih klasa ili srednja vrednost predviđanja u slučaju regresije . Broj osobina koje se mogu podeliti na svakom čvoru ograničen je na neki procenat od ukupnog broja (hiperparametar). Na taj način model se ne oslanja previše na bilo koju pojedinačnu karakteristiku i pošteno koristi sve potencijalno predviđajuće karakteristike. Svako stablo uzima nasumični uzorak iz izvornog skupa podataka prilikom podele čvorova , uvodeći dodatnu slučajnost koja sprečava preteranu prilagođenost podacima (*Overfitting*) [12].



Slika 3. Struktura klasifikacije RF(Preuzeto sa [13])

Naivni Bajes algoritam predstavlja verovatni model mašinskog učenja koji se koristi za rešavanje problema klasifikacije , korišćenjem Bajesove formule (slika 4)

$$p(y|X) = \frac{p(X|y)p(y)}{p(X)}$$

Slika 4. Formula za uslovnu verovatnoću

Predstavlja verovatnoću da se desi događaj y ukoliko se desio događaj X . X predstavlja dokaz, dok je y hipoteza. Pretpostavlja se da su atributi nezavisni , da njihove vrednosti ne utiču jedni na druge kao i da su svi podjednako važni . Otuda naziv "Naivni Bajes" . Promenljiva y se odnosi na konkretnu diskretnu vrednost (promenljiva klase) , dok promenljiva X predstavlja obeležja ($X=x_1....x_n$) . Na osnovu toga dobijamo formulu na slici 5.

$$p(y|x_1, x_2 \dots x_n) = \frac{p(x_1|y)p(x_2|y) \dots p(x_n|y)p(y}{p(x_1)p(x_2) \dots p(x_n)}$$

Slika 5. Formula za uslovnu verovatnoću sa vektorom x

Primećujemo da za svaki skup podataka p(x) je uvek isto i ne zavisi od y. Na taj način p(x) se ne računa , tako da ostaje samo vrednost u brojiocu i na taj način dobijamo proporcionalnost (slika 6).

$$p(y|x_1, x_2 \dots x_n) = \frac{p(x_1|y)p(x_2|y) \dots p(x_n|y)p(y)}{p(x_1)p(x_2) \dots p(x_n)}$$

Slika 6.Proporcionalnost

Zaključujemo da se , ukoliko je xi kategorička vrednost, relacija svodi na računanje frekvencija njenih različitih vrednosti (slika 7) [15] [16].

$$p(y|x) \sim p(y) \prod_{i=1}^{n} P(x_i|y)$$

Slika 7. Diskriminativni model

Metod potpornih vektora (SVM - *Support Vector Machine*) predstavlja model nadgledanog učenja koji se može koristiti za rešavanje problema i klasifikacije i regresije. U ovom rešenju je korišćen za rešavanje problema klasifikacije. Cilj SVM klasifikatora je da konstruiše hiperravan koja će sa najvećom mogućom marginom razdvajati klase podataka. Ova hiperravan se naziva optimalna hiperravan. Ukoliko bi kroz najbližu tačku sa obe strane hiperravni povukli paralelne hiperravni sa optimalnom, dobili bi zapravo dva potporna vektora po kojima je ova metoda i dobila ime. Rastojanje između optimalne hiperravni i potpornih vektora je jednako sa obe strane. U praksi veoma retko možemo očekivati ovakvu lineranu razdvojenost klasa, stoga moramo biti spremni da prihvatimo greške. Kod metode potpornih vektora možemo birati između broja grešaka, odnosno broja tačaka koje će završiti sa pogrešne strane hiperravni i veličine margine. Uvodimo parametar *C* koji daje određenu težinu svakoj tački koja se nalazi sa pogrešne strane. Mala vrednost parametra *C* znači da će margina biti velika, ali da će biti dozvoljeno da čak i veliki broj slučajeva bude pogrešno klasifikovan, dok velika vrednost parametra *C* znači da zanemarujemo veličinu margine i da ćemo imati mali broj tačaka koje su pogrešno klasifikovane [15]. U ovom rešenju smo probali sa različitim vrednostima za parametar C i opredelili se za vrednost hiljadu.

Rekurentne neuronske mreže imaju neku vrstu internog stanja (memorije) jer se izlaz neurona vraća na njegov ulaz. Metoda dugo kratkoročne memorije (*Long short term memory network*-LSTM) je veštačka rekurentna neuronska mreže. Razlikuje se od obične rekurentne neuronske mreže po tome što ima dodatne komponente (*input gate, forget gate, output gate*). Memorija, odnosno interno stanje je vidljivo samo kroz aktivacionu funkciju. GRU (*Gated Recurrent Unit*-GRU) je modifikacija LSTM-a. Umesto *input* i *forget gate*-a ima *update* i *reset gate*. Nema *output gate*. Još jedna razlike je u tome što je interno stanje vidljivo spolja jer nema aktivacionu funkciju.

Konvolutivne neuronske mreže (*Convolutional neural network*-CNN) se sastoje od ulaza, konvolutivnog sloja, aktivacione funkcije, agregacionog (*pooling*) sloja i MLP-a. U ovom rešenju je korišćen 1D CNN čiji kernel se kreće u jednom smeru. Najčešća primena konvolutivnih mreža je u obradi slika, gde su se pokazale izuzetno uspešnim [15].

Na osnovu istraživanja radova na ovu temu smo zaključili da bi bilo pogodno koristiti Linearnu regresiju , MLP i RFR. Posmatranjem rezultata koji su dobijeni u radovima, koje smo izabrali kao srodne sa ovim problemom, smo zaključili da su ovi modeli zadovoljavajući za predikciju visine temperature na osnovu istorijskih podataka. Takođe, na osnovu rezultata postignutih u radu [14] smo odlučili da za problem klasifikacije u slučaju predikcije vrste saobraćajne nezgode koristimo SVM, RFC i naivnog Bajesa.

B. *Metode evaluacije :*

Srednja kvadratna greška (MSE - Mean Squared Error) se računa po formuli prikazanoj na slici 8 (formula za MSE). Dimenzija ulaza je predstavljena kao n, Yi predstalja i-tu vrednost dobijenog izlaza, a Ŷi je i-ta vrednost očekivanog, odnosno ispravnog izlaza. Isto važi i za promenljive u formuli za srednju apsolutnu grešku (MAE – Mean Apsolute Error) koja se nalazi na slici 8 ispod formule za MSE. Pogodnost kvadratne greške je što je diferencijalna u svim tačkama, ali je osetljiva na ekstremne vrednosti. Za apsolutnu grešku važi suprotno, odnosno nije osetljiva na ekstremne vrednosti jer se sporije udaljava od nule u odnosu na kvadratnu grešku, ali nije diferencijalna u svim tačkama. MSA i MSE se koriste kod regresionih problema.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

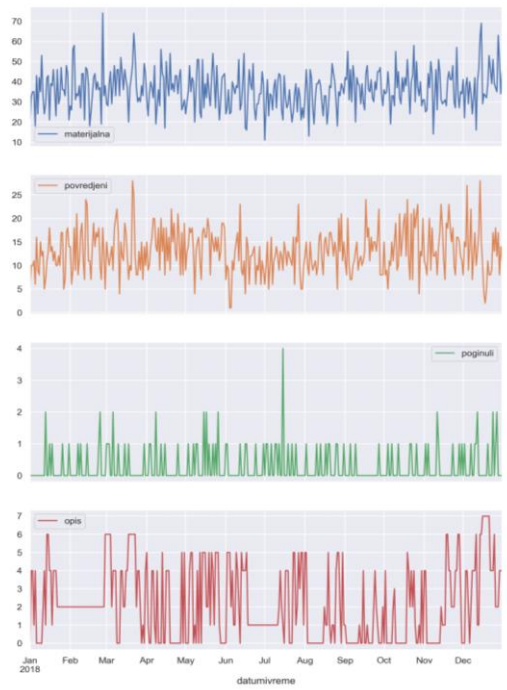
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

Slika 8. Formula za računanje MSE i MAE

Za mere uspešnosti kod problema klasifikacije, u ovom rešenju smo koristili tačnost (accuracy), preciznost (precision), odziv (recall) i F1 metodu. Tačnost predstavlja koliki deo instanci je tačno klasifikovan u odnosu na ukupan broj instanci. Ova metoda ne daje tačne rezultate ukoliko je broj podataka između klasa neizbalansiran. Preciznost predstavlja odnos između stvarno pozitivnih i zbira stvarno pozitivnih i lažnih pozitivnih, dok nam odziv daje odnos između stvarno pozitivnih i zbira stvarno pozitivnih i lažnih negativnih. F1 mera kombinuje preciznost i odziv. Ona predstavlja njihovu harmonijsku sredinu.

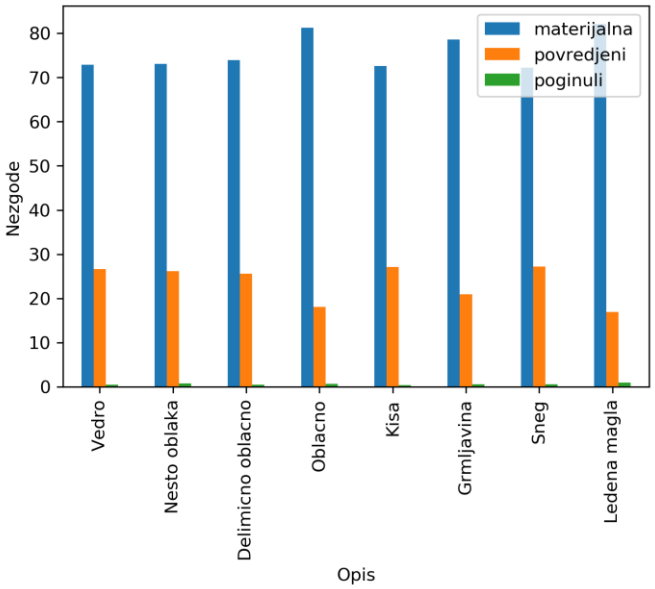
V. ANALIZA PODATAKA

Analizom grafova na slici 9. Možemo zaključiti da se najveći broj saobraćajnih nezgoda sa materijalnom štetom dogodio u februaru , iako su u tom periodu vremenski uslovi bili povoljni , odnosno bilo je pretežno oblačno bez padavina. Najveći broj povređenih u saobraćajnim nezgodama dogodio se u martu i decembru kada su uslovi za vožnju bili nepovoljni, odnosno kada je bilo snega i leda . Najveći broj tragedija dogodio se u letnjem periodu kada je bilo oblačno sa periodičnim padavinama.



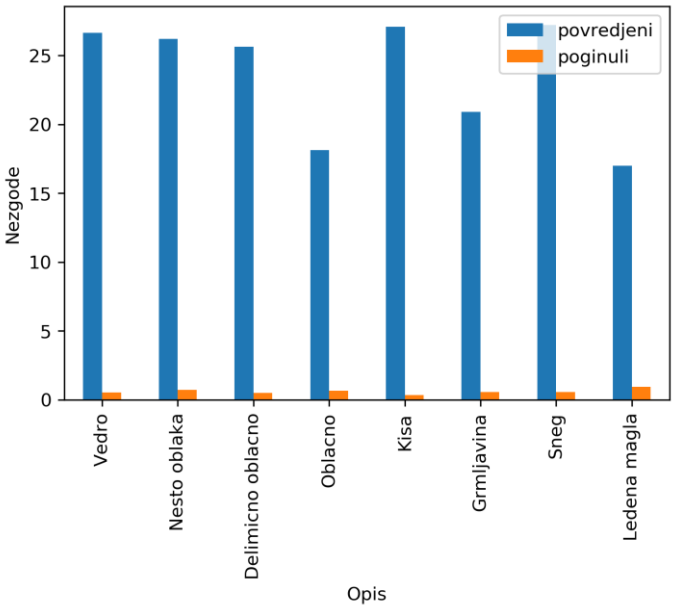
Slika 9. Broj saobraćajnih nezgoda po mesecima

Podaci na slici 10 predstavljaju procenat saobraćajnih nezgoda po opisima vremenskih uslova. Podaci su prikazani tako što je sumiran broj po grupama (materijalna , povređeni i poginuli) , za određeni vremenski uslov , podeljen sa ukupnim brojem zabeleženih zapisa za tu vremensku priliku. Na ovaj način prikazujemo uticaj opisa vremenskih uslova na saobraćajne nezgode bez obzira na učestalost vremenske pojave. Ovde vidimo da je najveći broj saobraćajnih nezgoda bio bez žrtava i to po oblačnom vremenu i ledenoj magli.



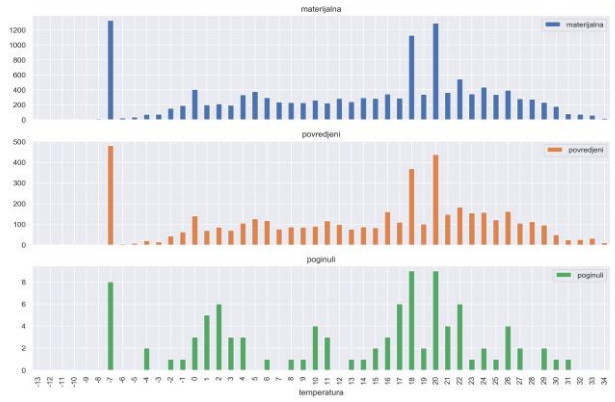
Slika 10. Broj saobraćajnih nezgoda podeljenih na tri klase na osnovu opisa vremenskih uslova

Na slici broj 11 imamo precizniji uvid u ozbiljnije saobraćajne nezgode. Možemo potvrditi zaključak do kojeg smo došli analizirajući sliku 9, ali i da postoji opasnost kad je vreme bez padavina. Pretpostavljamo da se ovo dešava zbog povećanog broja pešaka kada nema padavina. Podaci na slici broj 11 su predstavljeni na isti način kao i na slici broj 10.



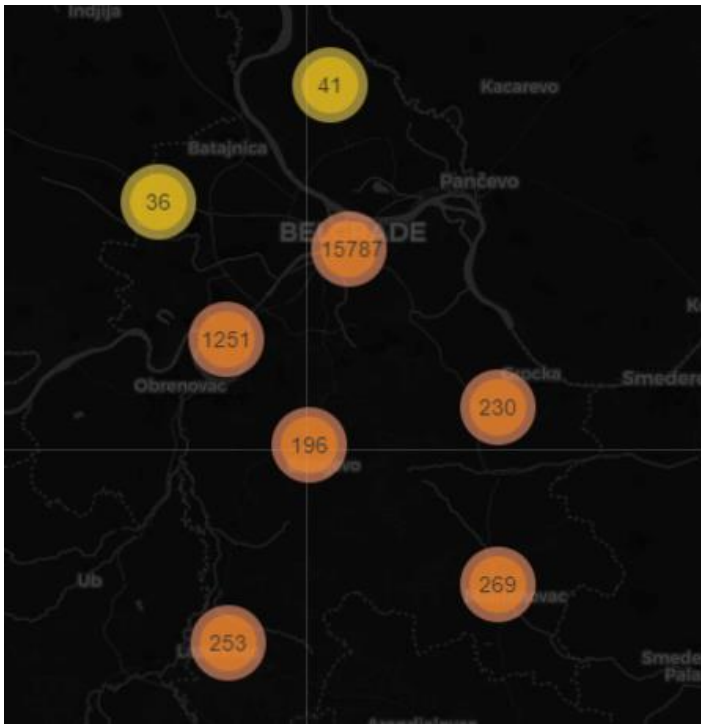
Slika 11. Broj saobraćajnih nezgoda sa isključenim nezgodama koje su rezultirale materijalnom štetom na osnovu opisa vremenskih uslova

Sa grafika prikazanih na slici 12 možemo zaključiti da su se nezgode (posmatrajući sve tri grupe) dešavale po ekstremnijem hladnom vremenu kao i prijatnijim temperaturama , što dodatno potvrđuje prethodni zaključak.



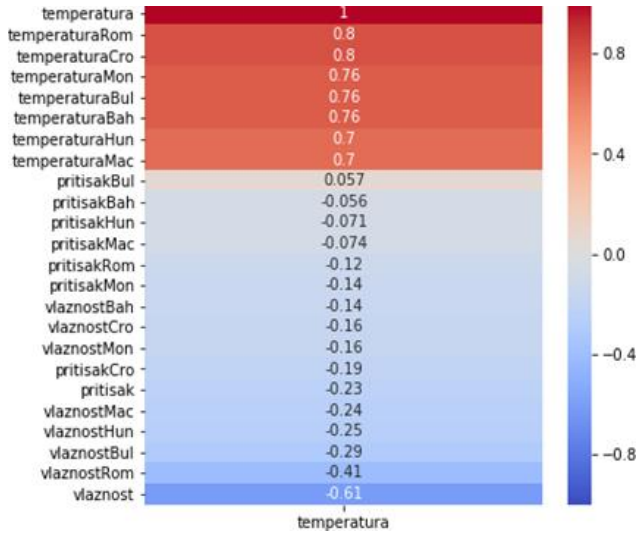
Slika 12. Ukupan broj saobraćajnih nezgoda u odnosu na temperaturu

Na slici 13 prikazan je ukupan broj saobraćajnih nezgoda po mestu zbivanja. Možemo zaključiti da se na putevima veće frekvencije saobraćaja češće dešavaju saobraćajne nezgode, odnosno u centru grada i bližoj okolini.



Slika 13. Prikaz broja saobraćajnih nezgoda na osnovu regije

Na slici 14 je prikazana zavisnost između temperature u Beogradu i drugih obeležja . Korelacija je izvršena nad skupom podataka koji je korišćen za predikciju temperature u Beogradu. Možemo videti da najveću korelaciju ima visina temperature u okolnim gradovima, a zatim pritisak i vlažnost.



Slika 14. Pirsonov koeficijent medjusobne korelacije temperature sa drugim obeležjima

U skupu podataka koji sadrži vremenske prilike, možemo primetiti da je najveći broj vedrih i oblačnih vremenskih uslova, što je i karakterisično za naše područje. Što se tiče skupa sa saobraćajnim nesrećama, možemo zaključiti da je neizbalansiran. Odnosno, imamo drastično veći broj materijalnih saobraćajnih nesreća u odnosu na ostale dve klase.

VI. Rezultati i analiza

U ovom delu će biti predstavljeni rezultati predikcije temperature i vrste saobraćajne nezgode koji su postignuti sa različitim modelima. Od modela se očekuje da dobro generalizuje, odnosno da prilikom predviđa vrednosti ciljne promenljive na osnovu vrednosti atributa retko pravi velike greške.

Provera će se vršiti poređenjem temperature koja je predviđena sa realno izmirenim temperaturom, pošto je u pitanju nadgledano učenje. Razlika između dobijene i tačne temperature se smatra greškom. Za proračun greške je korišćena metoda srednje kvadratne greške i metoda srednje apsolutne greške.

U tabeli 1 vidimo da MAE i MSE nisu veliki za linearnu regresiju, odnosno da je predikcija urađena zadovoljavajuće. Međutim, ovaj model nije u potpunosti pogodan za predikciju zato što zahteva linearnu zavisnost između zavisnih i nezavisnih promenljivih.

	MAE	MSE
Linearna regresija	1.75	5.11
Linearna regresija T	3.79	22.6
Linearna regresija V	5.38	44.46
Linearna regresija P	6.02	53.61
DNN	2.86	13.98
DNN T	2.73	13.41
DNN V	4.53	32.91
DNN P	6.4	59.67
MLP	0.66	0.73
MLP T	2.26	9.21
MLP V	3.43	21.13
MLP P	2.7	12.41
RFR	0.65	0.94
RFR T	1.62	5.73
RFR V	1.90	7.53
RFR P	0.97	2.16

Tabela 1. Prikaz srednje apsolutne i srednje kvadratne greške po modelima

U tabeli 1 oznaka T pored naziva modela označava da su se za njegov trening koristila samo obeležja koja prikazuju temperaturu. Isto važi za ostale slučajeve gde P označava pritisak, a V vlažnost. Za trening modela čiji su rezultati prikazani u tabeli 1 su korišćeni podaci iz Beograda i gradova iz regiona.

Na osnovu rezultata iz tabele 1 možemo zaključiti da se metoda MLP i RFR pokazale kao najbolje za treniranje modela kada koristimo obeležja vezana za temperaturu, pritisak i vlažnost. Možemo primetiti kako tačnost modela opada isključivanjem obeležja iz njega. Trenirani modeli potvrđuju rezultate dobijene korelacijom. Odnosno, vidimo kako je greška manja za modele koji su obučeni temperaturom, od modela za čiji trening smo koristili samo vlažnost ili samo pritisak. Pored ovih modifikacija u treningu modela su vršene i izmene sa brojem slojeva i neurona u njima. Na primer, MLP model obučen sa tri sloja od po sto neurona je imao MAE=1.98, a MSE=6.13, dok nam tri sloja sa 256, 128 i 64 neurona kao rezultat daju MAE=0.66, a MSE=0.73. DNN model je obučavan sa različitim brojem slojeva i neurona, ovim smo zaključili da dobijamo tačnije. rezultate ukoliko koristimo više slojeva. Rezultati dobijeni sa tri sloja po sto neurona su MAE=3.89 i MSE=20.69, a za dva sloja po 50 neurona su dobijene MAE=4.06 i MSE=25.94. Vođeni ovim rezultatima, povećali smo broj slojeva na četri i broj neurona po slojevima na 256, 128, 64 i 32 i dobili rezultat prikazan u tabeli. Eksperimente nad brojem slojeva i neurona smo izvodili nad kompletnim skupom obeležja.

Ukoliko se kao skup podataka koriste samo istorijski podaci vremenskih prilika na teritoriji Beograda , primećujemo da preciznost predikcije opada . Linearnom regresijom dobijamo MAE=5.8, a MSE=48.78. Ukoliko uporedimo ovaj rezultat sa rezultatom iz tabele koji su trenirani sa kompletnim skupom obeležja možemo zaključiti da podaci iz gradova iz regiona imaju veliki značaj na tačnost modela. Isto važi i za ostale metode. MLP je nad skupom obeležja koji obuhvata samo podatke iz Beograda dao sledeću tačnost: MAE=5.23 i MSE=40.43.

Korišćenjem metoda LSTM smo dobili zadovoljavajuće rezultate. MSE i MEA su bile približne za različite tipove modela. Model je obučavan nad istorijskim podacima vezanim samo za temperaturu u Begoradu, zatim nad podacima koji sadrže pored temperature i vlažnost i pritisak, kao i nad čitavim skupom podataka koji sadrži ove informacije i za okolne gradove. Greške tipa modela koji je dao najmanje greške nad skupom podataka koji sadrži temperaturu, vlažnost i pritisak na teritorji Beograda je MAE=0.65, a MSE=0.87. Model koji je obučavan nad skupom podataka koji sadrži samo temperaturu na teritoriji Beograda su korišćeni istorijski podaci od prethodna tri dana kako bi se izvršila predikcija temperature koje za trideset minuta. Najbolji tip ovog modela je dao rezultate MAE=0.61, a MSE=0.71. Nad kompletnim skupom podataka koji uključuje i podatke iz gradova iz regiona najbolji tip modela ima MSE=1.11, a MAE=0.79. Kod za ove primere je dostupan na [17], kao I za primere čiji rezultati su opisani u nastavku.

Gotovo svi tipovi LSTM-a su se pokazli bolje od GRU modela. GRU model je obučen nad skupom podataka koji uključuje i podatke o okolnim gradovima. Dobijeni rezultat je MAE=0.79, a MSE=1.26.

Model konvolucijske neuronske mreže je treniran nad skupom podataka koji se odnose samo na teritoriju Beograda. Podaci pored temperature uključuju i vrednosti za vlažnost vazduha i pritisak. Za ovaj model MAE iznosi 0.13.

	Preciznost	Odziv	F1-metod
Materijalna RFC	0.77	0.99	0.87
Povređeni RFC	0.89	0.22	0.35
Poginuli RFC	0.00	0.00	0.00
Tačnost RFC			0.78
Prosek RFC	0.80	0.78	0.72
Materijalna NB	0.74	0.94	0.83
Povređeni NB	0.47	0.13	0.20
Poginuli NB	0.00	0.00	0.00
Tačnost NB			0.71
Prosek NB	0.66	0.71	0.65
Materijalna SVM	0.76	0.98	0.86
Povređeni SVM	0.79	0.19	0.31
Poginuli SVM	0.00	0.00	0.00
Tačnost SVM			0.76
Prosek SVM	0.77	0.76	0.70

Tabela 2. Prikaz rezultata dobijenih za predikciju vrste saobraćajne nezgode

U tabeli 2 su prikazani rezultati koji su dobijeni za predikciju vrste saobraćajne nezgode. Korišćena su tri prikazana modela klasifikacije. Poređenjem rezultata vidimo da metoda potpornih vektora (SVM) i slučajnih šuma (RFC) imaju približne rezultate. Dobijeni rezultati su i očekivani. Ukoliko bi obogatili skup podataka sa više obeležja i proširili na neki veći vremenski interval, kvalitet ovih modela bi se povećao. Analizom podataka smo uvideli da u ovom skupu nedostaju obeležja koja imaju veću korelaciju sa vrstom odnosno ishodom saobraćajne nezgode. Metodom naivnog Bajesa dobili smo najslabije rezultate , što smo i očekivali na osnovu osobina algoritma koje su opisane u poglavlju 4.

Obeležje	**Važnost**
Broj vozila	0.69
Geografska širina	0.10
Geografska dužina	0.09
Pritisak	0.04
Vlažnost	0.03
Temperatura	0.03
Opis	0.01

Tabela 3. Važnost atributa u okviru RFC modela

Na osnovu tabele 3 možemo videti važnost ostalih atributa u odnosu na ciljni, konkretno vrste saobraćajne nesreće. Metodom slučajnih šuma dobijamo ovu tabelu važnosti atributa. Zaključujemo da najveći uticaj ima broj vozila.

VII. ZAKLJUČAK

U ovom radu je pokazano da je moguće doći do zadovoljavajućih rezultata predikcije visine temperature i predikcije ishoda saobraćajne nezgode primenom metoda mašinskog učenja. Da bi se rezultati poboljšali , neophodno je u ovo istraživanje uključiti što pouzdanije podatke i osobe sa domenskim znanjem iz oblasti meteorologije. Pored poboljšanja rezultata, ovaj rad bi mogao da se nadogradi i predikcijom dodatnih parametara vezanih za vremenske uslove, kao na primer količine padavina. Poteškoća na koju smo naišli prilikom izrade ovog projekta je bila nedostatak podataka za određene vremenske periode.

U današnje vreme, dok se saobraćajne nesreće stalno dešavaju, bezbednost u saobraćaju postala je glavni fokus savremenih društvenih pitanja. Mnogi faktori utiču na broj saobraćajnih nezgode i njihov ishod. Neki od njih su navedeni u ovom radu kao što su mesto nesreće, period u godini i vremenski uslovi. Ovaj rad ne rešava pitanje saobraćajnih nezgoda, ali ukazuje ne neke bitne faktore. Neophodno je podići svest svih učesnika u saobraćaju kako bi bili oprezniji u situacijama koje su opisane kao najkarakterističnije za saobraćajne nezgode. Pored toga, uvođenje povećane kontrole toka saobraćaja u urbanijim sredinama bi doprinelo povećanju bezbednosti.

Ovo istraživanje bi moglo da se proširi dodatnom analizom ljudskih faktora koji su bili učesnici u saobraćajnim nezgodama, kao i podaci o vozilu i kategoriji puta. Podaci koji bi bili od koristi u ovom istraživanju vezani za ljudske faktore su broj godina, pol, koliko dugo poseduje vozačku dozvolu, ali i dodatni parametri vezani za stanje vozača (uticaj alkohola, pospanost i slično). Karakteristike vozila od značaja bi bile starost, marka, model, kubikaža i slično. Dodatni atribut vezan za same saobraćajne nezgode bi mogao da bude uzrok. Zahvaljujući tome bi imali uvid da li je do saobraćajne nezgode došlo zbog nepažnje vozača, agresivne vožnje ili kršenje saobraćajnih propisa.

Literatura

[1] A H M Jakaria, Md M. Hossain, M. A. Rahman, "SmartWeather Forecasting Using Machine Learning: A Case Study in Tennessee", 2019., Dostupno na: https://www.researchgate.net/publication/330369173_Smart_Weather_Forecasting_Using_Machine_Learning_A_Case_Study_in_Tennessee [Pristupljeno 2. maja 2020].

[2] M. Holmstrom, D. Liu, C. Vo, "Machine Learning Applied to Weather Forecasting", Stanford University 2016., Dostupno na: http://cs229.stanford.edu/proj2016/report/HolmstromLiuVo-MachineLearningAppliedToWeatherForecasting-report.pdf [Pristupljeno 2. maja 2020].

[3] E. B. Abrahamsen, O. M. Brastein, B. Lie, "Machine Learning in Python for Weather Forecast based on Freely Available Weather Data", Proceedings of The 59th Conference on Simulation and Modelling (SIMS 59), 26-28 September 2018, Oslo Metropolitan University, Norway. Dostupno na: https://www.ep.liu.se/ecp/153/024/ecp18153024.pdf [Pristupljeno 2. maja 2020].

[4] Toby Staines, "Environmental Conditions and Road Traffic Collisions in the UK", 2018., Dostupno na:https://github.com/tobystaines/RoadAccidentsPODS/blob/master/Environmental%20Conditions%20and%20Road%20Traffic%20Collisions%20in%20the%20UK%20v1.2.pdf [Pristupljeno 2. maja 2020].

[5] Ch.Timmermans, W.Alhajyaseen, A. Al Mamun, T. Wakjira, M. Qasem, M. Almallah, H. Younis, "Analysis of road traffic crashes in the State of Qatar" , 2019. , Dostupno na: https://www.tandfonline.com/doi/full/10.1080/17457300.2019.1620289# [Pristupljeno 3. maja 2020].

[6] E. Marsden, "Regression analysis using Python", Dostupno na: https://risk-engineering.org/static/PDF/slides-linear-regression.pdf [Pristupljeno 4. maja 2020].

[7] FreeMeteo - Vremenske prognoze ., Dostupno na: https://freemeteo.rs/ [Pristupljeno 4. maja 2020].

[8] Republika Srbija Portal otvorenih podataka., Dostupno na: https://data.gov.rs/sr/datasets/podatsi-o-saobratshajnim-nezgodama-za-teritoriju-grada-beograda/ [Pristupljeno 4. maja 2020].

[9] A. N George, "Deep Neural Network Toolkit & Event Spotting in Videousing DNN features", Department of computer science and engineering Indian Institute of Tehnology Madras, .May 2015. Dostupno na: https://www.academia.edu/20434118/Deep_Neural_Network_Toolkit_And_Event_Spotting_in_Video_using_DNN_features [Pristupljeno 4. maja 2020].

[10] "Simple tutorial to write deep neural network by TensorFlow". Dostupno na: http://marubon-ds.blogspot.com/2017/09/simple-tutorial-to-write-deep-neural.html [Pristupljeno 4. maja 2020].

[11] N. K. Kain, "Understanding of Multilayer perceptron (MLP)". Dostupno na: https://medium.com/@AI_with_Kain/understanding-of-multilayer-perceptron-mlp-8f179c4a135f [Pristupljeno 4. maja 2020].

[12] Afroz Chakure, "Random Forest Regression". Dostupno na: https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f [Pristupljeno 5. maja 2020]

[13] Afroz Chakure, "Random Forest Classification". Dostupno na: https://towardsdatascience.com/random-forest-classification-and-its-implementation-d5d840dbead0 [Pristupljeno 5. maja 2020]

[14] Christian S. Perone, "Injury risk prediction for traffic accidents in Porto Alegre/RS, Brazil (draft)", Pontif˙ıcia Universidade Catolica do Rio Grande do Sul, februar 2015., Dostupno na: https://arxiv.org/pdf/1502.00245.pdf [pristupljeno 25. maja 2020]

[15] M. Nikolić ,A. Zečević., "Mašinsko učenje", Beograd, 2019., Dostupno na: http://ml.matf.bg.ac.rs/readings/ml.pdf [pristupljeno 30. Maja 2020]

[16] R. Gandhi., "Naive Bayes Classifier", 2018., Dostupno na: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c [pristupljeno 30. maja 2020]

[17] Github repozitorijum, Dostupno na: https://github.com/TomicI/NM

[18] Ben A. O. I., Boudhir A.A., Astito A., Bassam Z., Bouhorma M., „Deep Learning architecture for temperature forecasting in an IoT LoRa based system", NISS19, March 27–29, 2019, Rabat, Morocco 2019 Association for Computing Machinery, Dostupno na: https://www.researchgate.net/publication/333258384_Deep_Learning_architecture_for_temperature_forecasting_in_an_IoT_LoRa_based_system [pristupljeno 14. juna 2020]