Modeli za predikciju ishoda političkih izbora korištenjem društvene mreže Facebook i algoritama strojnog učenja

Kišić, Alen

Sveučilište u Zagrebu

Fakultet organizacije i informatike

Alen Kišić

MODELI ZA PREDIKCIJU ISHODA POLITIČKIH IZBORA KORIŠTENJEM DRUŠTVENE MREŽE FACEBOOK I ALGORITAMA STROJNOG UČENJA

DOKTORSKI RAD

Mentor:
Prof.dr.sc. Božidar Kliček

Zagreb, 2023.

University of Zagreb

Faculty of Organization and Informatics

Alen Kišić

MODELS FOR PREDICTING THE OUTCOME OF POLITICAL ELECTIONS BY MEANS OF SOCIAL NETWORK FACEBOOK AND MACHINE LEARNING ALGORITHMS

DOCTORAL DISSERTATION

Supervisor:
Full Professor Božidar Kliček, Ph.D.

Zagreb, 2023

PODACI O DOKTORSKOM RADU

I. AUTOR

Ime i prezime	Alen Kišić
Datum i mjesto rođenja	18.04.1977.
Naziv fakulteta i datum diplomiranja	Fakultet političkih znanosti, 2004.
Sadašnje zaposlenje	Direktor Zone Sjever

II. DOKTORSKI RAD

Naslov	Modeli za predikciju ishoda političkih izbora korištenjem društvene mreže Facebook i algoritama strojnog učenja
Broj stranica, slika, tabela, priloga, bibliografskih podataka	106 stranica, 20 slika, 23 tabele, 1 prilog, 88 bibliografskih podataka.
Znanstveno područje i polje iz kojeg je postignut akademski stupanj	Informacijske znanosti, polje informacijske i komunikacijske znanosti
Mentor i voditelj rada	Prof.dr.sc. Božidar Kliček
Fakultet na kojem je rad obranjen	Fakultet organizacije i informatike
Oznaka i redni broj rada	167

III. OCJENA I OBRANA

Datum sjednice Fakultetskog vijeća na kojoj je prihvaćena tema	21.07.2020.
Datum predaje rada	01.07.2022.
Datum sjednice Fakultetskog vijeća na kojoj je prihvaćena pozitivna ocjena rada	20.04.2023.
Sastav Povjerenstva koje je rad ocijenilo	Prof.dr.sc. Markus Schatten Prof.dr.sc. Jasminka Dobša Izv.prof.dr.sc. Goran Klepac
Datum obrane	12.06.2023.
Sastav Povjerenstva pred kojim je rad obranjen	Prof.dr.sc. Markus Schatten Prof.dr.sc. Jasminka Dobša Izv.prof.dr.sc. Goran Klepac
Datum promocije	

SAŽETAK

Odlučivanje temeljeno na podacima nova je paradigma za rješavanje brojnih problema u širokom spektru domena. Politika i predviđanje ishoda političkih izbora nije iznimka te raste interes znanstvene i stručne zajednice za primjenom i iskorištavanjem prednosti podatkovne analitike i algoritama strojnog učenja. Ovaj rad nastoji dati nekoliko znanstvenih i stručnih doprinosa u toj domeni. Tradicionalno, najčešća i najtočnija metoda mjerenja javnog mnijenja bile su uzorkovane ankete kojima se postavljaju pažljivo izrađena pitanja precizno definiranim uzorcima populacije. No, takav pristup ima i visoku cijenu: velika ulaganja vremena, truda i novaca za istraživače koji dizajniraju istraživanje, anketari koji prikupljaju podatke i ispitanici koji dobrovoljno daju odgovore. Problem anketa je iskrenost ispitanika, kao i sam uzorak. Nadalje, niz prognoza ishoda izbora u zadnje vrijeme se razlikovalo od konačnih rezultata, što je znatno poljuljalo povjerenje u takve vrste prognoza. Nedavno se pojavila alternativa takvom pristupu s potencijalom za dopunu ili čak potpunu zamjenu dosad korištenih metoda istraživanja koja bi smanjila troškove za istraživače i uklonila napore za ispitanike. Naime, istraživači su počeli koristiti podatke društvenih mreža. U domeni političkih kampanja, taj je potencijal izrazito velik s obzirom na činjenicu da praktično svi politički kandidati i političke stranke u kampanjama koriste društvene mreže. Ovdje provedeno istraživanje primjenjuje algoritme strojnog učenja na podacima aktivnosti korisnika društvene mreže Facebook za razvoj prediktivnih modela ishoda izbora. Glavni je cilj istraživanja usporediti njihovu točnost i pouzdanost s modelima dobivenim tradicionalnim ispitivanjima javnog mnijenja. Istraživanje se provodi na francuskim lokalnim izborima. Za izradu prediktivnih modela koristit će se četiri različita pristupa strojnom učenju: pristup temeljen na pogrešci (algoritam umjetne neuronske mreže), pristup temeljen na informaciji (algoritam stabla odlučivanja), pristup temeljen na sličnosti (algoritam k-najbližih susjeda) i pristup temeljen na vjerojatnosti (naivni Bayesov klasifikator). U evaluaciji i usporedbi modela će se utvrditi koji je pristup strojnog učenja najefikasniji za predviđanje ishoda izbora temeljem podataka društvene mreže Facebook. Dobivanje jednako učinkovitih prediktivnih modela uz brži i jednostavniji pristup podacima daje značajan znanstveni i stručni doprinos istraživanjima u ovom području.

ABSTRACT

Data-driven decision-making is a new paradigm for solving various problems in a broad spectrum of domains. Politics and predicting the outcome of political elections is no exception, and interest is growing for both, scientific and professional communities in applying and exploiting the advantages of data analytics and machine learning algorithms. This paper strives to make several scientific and professional contributions in the domains. Traditionally, the most common approach to measure public opinion are sampling surveys that ask carefully defined questions to precisely selected samples of the population. However, such an approach has a high price: large investments of time, effort and money for researchers who design the survey, who collect the data, and the respondents who answers. The problem with surveys is the honesty of the respondents, as well as the sample. Furthermore, a number of elections outcomes forecasts recently differed from the final results, which shaken confidence in these kinds of forecasting. An alternative to such an approach has recently emerged with the potential to complement it or even to completely replace previously used research methods that would reduce costs for researchers and removed the effort for respondents by using social media data. In the domain of political campaigns, this potential is extremely large considering the fact that practically all political candidates and political parties use social networks in their campaigns. The research conducted here applies machine learning algorithms on data from social network Facebook user's activity to develop predictive models of election outcomes. Main goal of the research is to compare their accuracy and reliability with models obtained by traditional public opinion polls. The research is conducted on French local elections. Four different approaches to machine learning are used to develop predictive models: an approach based on errors (artificial neural network algorithm), information-based approach (decision tree algorithm), a similarity-based approach (k-nearest neighbours) and an approach based on probabilities (Naive Bayes classifier). The evaluation and comparison of the models will determine the best machine learning algorithm for predicting election outcomes based on social data network Facebook. Obtaining equally effective predictive models while being faster and simpler access to data provides a significant scientific and professional contribution to research in this area.

PROŠIRENI SAŽETAK

Odlučivanje temeljeno na podacima nova je paradigma za rješavanje brojnih problema u širokom spektru domena. Politika i predviđanje ishoda političkih izbora nije iznimka te raste interes znanstvene i stručne zajednice za primjenom i iskorištavanjem prednosti podatkovne analitike i algoritama strojnog učenja. Ovaj rad nastoji dati nekoliko znanstvenih i stručnih doprinosa u toj domeni. Tradicionalno, najčešća i najtočnija metoda mjerenja javnog mnijenja bile su uzorkovane ankete kojima se postavljaju pažljivo izrađena pitanja precizno definiranim uzorcima populacije. No, takav pristup ima i visoku cijenu: velika ulaganja vremena, truda i novaca za istraživače koji dizajniraju istraživanje, anketari koji prikupljaju podatke i ispitanici koji dobrovoljno daju odgovore. Problem anketa je iskrenost ispitanika, kao i sam uzorak. Nadalje, niz prognoza ishoda izbora u zadnje vrijeme se razlikovalo od konačnih rezultata, što je znatno poljuljalo povjerenje u takve vrste prognoza. Nedavno se pojavila alternativa takvom pristupu s potencijalom za dopunu ili čak potpunu zamjenu dosad korištenih metoda istraživanja koja bi smanjila troškove za istraživače i uklonila napore za ispitanike. Naime, istraživači su počeli koristiti podatke društvenih mreža. U domeni političkih kampanja, taj je potencijal izrazito velik s obzirom na činjenicu da praktično svi politički kandidati i političke stranke u kampanjama koriste društvene mreže.

Ovdje provedeno istraživanje primjenjuje algoritme strojnog učenja na podacima aktivnosti korisnika društvene mreže Facebook za razvoj prediktivnih modela ishoda izbora. Ciljevi istraživanja su: (i) utvrditi kolika je prediktivna moć modela temeljenih na društvenoj mreži Facebook i usporediti je s drugim vrstama istraživanja, (ii) utvrditi koje varijable su najznačajniji prediktori ishoda izbora, (iii) utvrditi značajnost temporalne komponente podataka društvene mreže Facebook, (iv) utvrditi koja od četiri metode strojnog učenja daje najtočnije prediktivne modele ishoda lokalnih izbora. U skladu s ciljevima, definirane su sljedeće hipoteze istraživanja:

H1:Točnost prediktivnih modela razvijenih na podacima aktivnosti korisnika društvene mreže Facebook veća je od modela temeljenih na anketama.

H2: Pristup strojnom učenju temeljen na pogrešci daje točnije prediktivne modele od ostala tri pristupa strojnom učenju.

Postavljena su i dva istraživačka pitanja:

IP1: Koje su varijable najznačajniji prediktori ishoda izbora?

IP2: U kojoj je mjeri temporalna komponenata podataka društvene mreže Facebook važna u predikciji ishoda izbora?

Hipoteze istraživanja će se testirati na način kako slijedi. Hipoteze H1 testirat će se tako da će se primjenom četiri algoritma strojnog učenja razviti prediktivni modeli ishoda izbora. Performanse tih modela usporedit će se s rezultatima ispitivanja javnog mnijenja prikupljenih sa stranica francuskih medija koji su radili predizborna ispitivanja javnog mnijenja. Za testiranje hipoteze koriste se testovi koji ne pretpostavljaju određenu distribuciju podataka, koja za podatke preuzete od francuskih medija nije poznata. Hipoteza H2 provjerava se primjenom mjere greške i testiranjem razlika u točnosti modela dobivenih različitim algoritmima. Kako bi se odgovorilo na istraživačko pitanje IP1 provest će se analiza osjetljivosti nad razvijenim modelima kako bi se utvrdilo kako promjena vrijednosti ulaznih varijabli utječe na vrijednosti izlazne varijable. Na taj će se način utvrditi prediktori ishoda izbora. Analiza osjetljivosti jedan je od pristupa objašnjivosti modela strojnog učenja. Kako bi se odgovorilo na istraživačko pitanje IP2 razvit će se prediktivni modeli na podacima tri razdoblja kampanje i usporedit će se njihovi rezultati.

Istraživanje prati korake i aktivnosti CRISP-DM metodologije koja se sastoji od šest faza. Prva faza je razumijevanje problema koja je uključivala određivanje ciljeva istraživanja i razvoj plana te definiranje tijeka istraživanja. Druga faza odnosi se na razumijevanje podataka. Fokus ove faze je prikupljanje podataka i početno opisivanje podataka. Za provođenje empirijskog istraživanja kojim će se ostvariti ciljevi rada i testirati hipoteze prikupljaju se podaci s društvene mreže Facebook, i to: podaci sa stranica kandidata na lokalnim izborima u Francuskoj 2020. godine. Društvena mreža Facebook je odabrana sukladno sugestijama prethodnih istraživanja koja tvrde da političke kampanje sve više „bježe“ na Facebook jer je to društvena mreža s najvećim brojem korisnika. Nadalje, istraživanja pokazuju da je generacijama tzv. Milenijalaca i generaciji X (dob od 18 do 51 godine) Facebook najčešće korišteni izvor političkih vijesti. U Francuskoj, koja se obrađuje u ovom radu, najveći dio stanovništva koristi društvenu mrežu, Facebook, i to 70% stanovnika od onih koji koriste društvene mreže. Odabir Facebooka dodatno je motiviran i recentnim radovima koji su istraživanje provodili samo na Twitteru, a u smjernicama za buduća istraživanja navode potrebu za uključivanjem drugih društvenih mreža. Uz podatke društvene mreže Facebook koristit će se i podaci o rezultatima predizbornih anketa.

Podaci su prikupljani na lokalnim izborima u Francuskoj, 2020. godine. Uključeni su svi francuski gradovi s više od 100 tisuća stanovnika (njih 41). Dodatni uvjet uključivanja grada

bio je da barem dva kandidata u tom gradu imaju otvorene stranice/profile. Podaci su kreirani od strane autora rada pojedinačnim preuzimanjem podataka sa stranica kandidata na izborima, u lipnju 2020. godine. Ime i prezime kandidata (njih 225) u gradovima preuzeti su sa službenih stranica izbornog povjerenstva. Svaki kandidat je pretraživan na društvenoj mreži Facebook s ciljem pronalaska stranica ili profila. Kod kandidata koji su imali otvorene stranice ili profile, pregledavan je svaki sadržaj i kategoriziran u jednu od sljedećih kategorija: događaj, fotografija, poveznica, video i status. Za svaki sadržaj ekstrahiran je podatak o broju sadržaja na stranici/profilu, broju likeova sadržaja, broju komentara sadržaja te broju dijeljenja sadržaja. Inicijalno je u istraživanje uključeno 25 varijabli: 24 ulazne varijable, a odnose se na: aktivnost kandidata i reakcije pratitelja stranice na aktivnost kandidata, dvije varijable koje se odnose na spol kandidata i pripadnost kandidata političkoj stranci. Jedna izlazna varijabla za prediktivne modele: rezultat na izborima mjeren postotkom glasova kandidata na izborima. Atributi su se pratili u tri vremenska perioda: (i) na početku kampanje, (ii) u sredini kampanje i (iii) zadnji dan kampanje. Francuski lokalni izbori su održani 15.03.2020. godine. Službena kampanja je započela 02.03.2020. Prvi skup podataka odnosio se na aktivnosti u samom početku kampanje (od 02.03. do 06.03.), drugi skup podataka na aktivnosti kandidata u sredini kampanje (07.03. do 11.03.), te treći skup podataka na sam završetak kampanje (12.03. do 14.03).

Nakon prikupljanja podataka, sastavni dio ove faze je: definiranje varijabli i tipova varijabli. Atributi koji će se uzimati u obzir su: grad i izlaznost birača u gradu, stranka, ukupan broj *likeova* stranice, broj fotografija, broj statusa (tekst), broj poveznica, broj videa, broj kreiranih događaja, broj dijeljenja fotografija, broj dijeljenja statusa, broj dijeljenja poveznica, broj dijeljenja videa, broj *likeova* fotografija, broj *likeova* statusa, broj *likeova* poveznica, broj *likeova* videa, broj komentara fotografije, broj komentara statusa, broj komentara poveznica, broj komentara videa, vrijeme objave. Za potrebe modeliranja kreirat će se nove, sintetičke varijable tako da se sve varijable koje se odnose na aktivnost na društvenoj mreži podijele s brojem birača u gradu u kojem glasaju. Većina varijabli ima eksponencijalnu distribuciju vrijednosti koju karakterizira velika vjerojatnost pojave manjih vrijednosti, a mala vjerojatnost pojave velikih vrijednosti. Karakteristika je takvih distribucija da su vrijednosti aritmetičke sredine veće od vrijednosti medijana. Korelacijskom analizom je utvrđen niz linearnih povezanosti između varijabli. Značajnu i visoku povezanost (vrijednost koeficijenta korelacije r iznad 0.7) imaju varijable koje se odnose na dijeljenje, lajkanje i komentiranje određene vrste sadržaja: najveća linearna povezanost utvrđena je između varijabli broj dijeljenja događaja i broj likeova događaja, Povezanost je pozitivna, što ukazuje da s porastom likeova događaja

raste i broj dijeljenja događaja, i obratno. Izrazito je mali broj negativnih koeficijenata korelacije.

Treća faza CRISP DM standarda je priprema podataka. Nakon prethodnog identificiranja izvora dostupnih podataka, ovdje slijedi čišćenje podataka, transformacija atributa i selekcija atributa. Cilj je pripreme podataka očistiti podatke, te odabrati atribute za modeliranje. Prva aktivnost bila je identifikacija stršila. Identificirana su stršila su jer su neki od primijenjenih algoritama strojnog učenja (umjetne neuronske mreže, k-najbližih susjeda) jako osjetljive na ekstremne vrijednosti. Interkvatrili su korišteni za identifikaciju stršila na sljedeći način. Vrijednost je označena kao stršilo ako se nalazi najmanje 1.5 interkvartila ispod prvog kvartila ili 1.5 interkvartila iznad trećeg kvartila. Micanjem stršila nastale su nedostajuće vrijednosti. Postupkom umetanja podataka umjesto vrijednosti koje nedostaju provedena je imputacija. Time su pridjeljivane vrijednosti na prazna mjesta temeljem sljedeće heuristike: nedostajuće vrijednosti zamijenjene su srednjim vrijednostima varijabli. Sljedeći korak unutar pripreme podataka bila je normalizacija podataka. Provedena je normalizacija s dva aspekta. Prvo su podaci normalizirani s obzirom na broj birača u pojedinom gradu. Originalna vrijednost svake varijable je podijeljena s brojem birača u gradu na koji se odnosi. U drugom koraku provedena je min-max normalizacija kojom su vrijednosti atributa svedene na skalu od 0 do 1. U aktivnosti selekcije atributa, odabrani su samo relevantni atributi koji predstavljaju ulaze u modele. U ovom se istraživanju koristi filter pristup koji radi rangiranje varijabli i odabire određeni podskup varijabli temeljem mjere vrednovanja. Kao mjera vrednovanja, korišten je ReliefF algoritam. Sukladno tome, 8 varijabli od početnog skupa varijabli koje se odnose na podatke društvene mreže Facebook su odabrane te će nam predstavljati ulaz u modeliranje: ukupan broj likeova stranice, Broj događaja, Broj likeova statusa, Broj fotografija, Broj poveznica, Broj likeova fotografija, Broj likeova poveznica, Broj statusa. Dodatno, uključeni su atributi spol i stranka kandidata.

Četvrta faza je modeliranje. Korak modeliranja je proveden nad pripremljenim i normaliziranim podacima. Podaci podijeljeni na skup za treniranje i skup za testiranje modela. Prediktivni modeli ishoda izbora izradit će se primjenom četiri vrste algoritama strojnog učenja: (i) strojno učenje temeljeno na informaciji (algoritam: stablo odlučivanja), (ii) strojno učenje temeljeno na sličnosti (algoritam: k-najbližih susjeda), (iii) strojno učenje temeljeno na vjerojatnosti (algoritam: naivni Bayesov klasifikator), (iv) trojno učenje temeljeno na pogrešci (algoritam: umjetne neuronske mreže). Primjenom svakog od četiri pristupa strojnom učenju razvijeni su prediktivni modeli na skupovima podataka prikupljenih u različitim vremenskim

periodima. U tom je koraku provedena i optimizacija hiperparametara svakog pojedinog algoritma kako bi se spriječila pretreniranost modela, a dobili kvalitetni pouzdani i točni prediktivni modeli.

Nakon izrade modela slijedi faza evaluacije modela. Kod evaluacije modela rezultate treba vrednovati u kontekstu ciljeva postavljenih u prvoj fazi. U sklopu ove faze radi se vrednovanje temeljem definiranih metrika kvalitete modela. Kao metoda validacije koristila se unakrsna validacija s *k*-preklapanja (engl. *k-fold cross validation*). Temeljem mjera točnosti (mjerena kroz RASE) i pouzdanosti modela (mjerena kroz RSquare) komparirali su se modeli kako bi se utvrdilo koji od njih daje najbolju predikciju. Model dobiven neuronskom mrežom ima najbolje parametre kvalitete: daje natočniju predikciju i model je najviše razine pouzdanosti. Dio objašnjenja ovih rezultata nalazi se u tipovima podataka. Ulazni atributi su većinom numerički kontinuirani atributi. Modeli s najnižim vrijednostima pouzdanosti i točnosti dobiveni su naivnim Bayesovim klasifikatorom. Objašnjenje rezultata također se može povezati s karakteristikama skupa podataka na kojem je model razvijen. Naivni Bayesov klasifikator radi s kategorijskim varijablama te je nužna transformacija kontinuiranog izlaza u kategorijski. Temeljem ovih rezultata potvrđuje se hipoteza H2. Testiranjem hipoteze H1, utvrđeno je da ne postoji statistička značajna razlika u rezultatima koji se dobivaju prediktivnim modelima temeljenim na podacima društvene mreže Facebook i algoritmu umjetne neuronske mreže u odnosu na predizborna ispitivanja javnog mnijenja pomoću ankete. Modeli dobiveni na podacima društvene mreže imaju veću točnost od modela dobivenih anketiranjem, ali ta razlika nije statistički značajna. U svrhu usporedbe prediktivnih modela dobivenih algoritmima strojnog učenja i rezultatima predizbornog anketiranja, korišteni su podaci francuskog instituta za ispitivanje javnog mnijenja, IFOP koji se odnose na 6 gradova: Pariz, Lyon, Marseille, Rennes, Nantes i Bordeaux, te su time uključeni podaci za ukupno 44 kandidata na izborima. Temeljem ovih rezultata djelomično se potvrđuje hipoteza H1.

U svrhu odgovaranja na istraživačko pitanje IP2, na svakom od tri skupa podataka iz tri različita vremenska perioda, razvijen je prediktivni model primjenom algoritma umjetne neuronske mreže koji se pokazao najboljim na cijelom skupu podataka. Tablica koja testira statistiku Friedman testa rezultati testa pokazuju da nema statistički značajne razlike (na razini značajnosti od 0.01) između prediktivnih modela dobivenih u prvom, drugom i trećem razdoblju.

U svrhu odgovaranja na istraživačko pitanje IP1, provedena je analiza osjetljivost na sva četiri prediktivna modela. Rezultati najznačajnijih prediktora rezultata četiri prediktivna modela dobivena primjenom četiri različitih pristupa strojnom učenju pokazuju stabilnost i konzistentnost modela. Ukupan broj *likeova* stranice kandidata najznačajniji je prediktor u sva četiri modela. Broj *likeova* poveznica drugi je najznačajniji prediktor također u sva četiri modela. Broj statusa treći je najjači prediktor izbornih rezultata u dva od četiri promatrana modela.

Posljednja, šesta faza CRISP DM metodologije je korištenje modela. U istraživanjima se ova zadnja faza odnosi na ekstrakciju znanja. Dobiveni prediktivni modeli služe kao alat za planiranje i upravljanje kampanjama te kao takvi služe kao instrument za potporu odlučivanje kroz profiliranje Najvažniji rezultati ovog istraživanja pružaju brojne implikacije za korištenje društvenih medija kao indikatora za predviđanje ishoda izbora te se mogu dati smjernice za učinkoviti pristup upravljanju podacima društvenih mreža. Rezultati istraživanja vode i boljem razumijevanju načina na koji društvene mreže prezentiraju stvavove birača, te ukazuju kako se može djelovati na biračko tijelo preko društvenih mreža. Prvo je otkriveno kako je apsolutni broj Facebook pratitelja jako dobar prediktor izbornih ishoda. Drugo je utvrđeno da je sadržaj koji kandidati plasiraju putem društvenih medija, kao i reakcije na specifične sadržaje koje dijele u određenoj mjeri indikator ishoda izbora. S obzirom na niske troškove povezane s povezivanjem na društvenim mrežama, ovi rezultati potencijalno vode smanjenju troškova u političkim kampanjama. Naravno, korisnici društvene mreže mogu biti prijatelji ili pratiti mnoge kandidate, ne samo njihovog preferiranog kandidata, uključujući kandidate u biračkim tijelima na koje nemaju pravo glasa. Stoga i broj pratitelja nije jedinstven indikator.

Znanstveni doprinos rada sadržan je u sljedećem: (i) sistematizacija znanja o prediktivnim modelima temeljenim na podacima ispitivanja javnog mnijenja i društvenih mreža, (ii) razvijeni i evaluirani prediktivni modeli ishoda izbora temeljem podataka društvene mreže prikupljenih na francuskim lokalnim izborima, (iii) utvrđena prediktivna moć varijabli društvene mreže Facebook, (iv) razvijene smjernice za korištenje algoritama strojnog učenja na podacima društvenih medija.

Stručni doprinos rada ogleda se u sljedećem: (i) povećana objašnjivost predviđanja ishoda izbora, (ii) smanjeno vrijeme i troškovi procesa ishoda izbora, (iii) utvrđivanje uloge društvene mreže Facebook na razini lokalnih izbora.

Dobivanje jednako učinkovitih prediktivnih modela uz brži i jednostavniji pristup podacima daje značajan znanstveni i stručni doprinos istraživanjima u ovom području.

Prilikom interpretacije rezultata treba biti oprezan i uzeti u obzir nekoliko ograničenja istraživanja, a koja s druge strane otvaraju smjernice za buduća istraživanja. Istraživanje je provedeno na uzorku jedne države, Republike Francuske, i na jednim provedenim izborima, i to za lokalnu vlast. Svaka država, kao i svaki izbori, imaju svoje specifičnosti koje treba uzeti u obzir prilikom provođenja istraživanja. U budućim istraživanjima će se dobiveni modeli testirati i u drugim okruženjima, i na drugim državama. Pitanje uključivanja jedne države vodi do problema malog uzorka. Naime, u skupu podataka je relativno mali broj instanci. Da bi se dobili kvalitetni modeli, u dizajnu istraživanja korištena je metoda k-unakrsne validacije, koja dobro radi s malim brojem instanci. U interpretaciji rezultata treba uzeti u obzir da se podaci odnose na specifično vrijeme, početak ožujka 2020., baš na samom početku širenja pandemije virusa COVID-19 i lockdowna koji je slijedio u drugoj polovici ožujka.

U analizi podataka korištena su četiri algoritma strojnog učenja, odabrana iz širokog broja dostupnih algoritama. Iako odabrani algoritmi predstavljaju različite pristupe načinu razvoja modela, postoji još niz algoritama koji bi se mogli primijeniti na podacima.

KLJUČNE RIJEČI: algoritmi strojnog učenja, društvena mreža Facebook, politički izbori, umjetne neuronske mreže, prediktivni modeli, predizborne ankete.

EXTENDED ABSTRACT

Data-driven decision-making is a new paradigm for solving various problems in a broad spectrum of domains. Politics and predicting the outcome of political elections is no exception, and interest is growing for both, scientific and professional communities in applying and exploiting the advantages of data analytics and machine learning algorithms. This paper strives to make several scientific and professional contributions in the domains. Traditionally, the most common approach to measure public opinion are sampling surveys that ask carefully defined questions to precisely selected samples of the population. However, such an approach has a high price: large investments of time, effort, and money for researchers who design the survey, who collect the data, and the respondents who answer. The problem with surveys is the honesty of the respondents, as well as the sample. Furthermore, a number of election outcomes forecasts recently differed from the final results, which has shaken confidence in these kinds of forecasting. An alternative to such an approach has recently emerged with the potential to complement it or even to completely replace previously used research methods that would reduce costs for researchers and remove the effort for respondents by using social media data. In the domain of political campaigns, this potential is extremely large because practically all political candidates and political parties use social networks in their campaigns. The research conducted here applies machine learning algorithms to data from social network Facebook users' activity to develop predictive models of election outcomes.

The objectives of the research are: (i) to determine the predictive power of models based on the Facebook social network and compare it with other types of research, (ii) to determine which variables are the most significant predictors of the election outcomes, (iii) to determine the significance of the temporal component of the Facebook social network data, (iv) to determine which of the four machine learning algorithms provide the most accurate predictive models of local election outcomes. Accordingly, the following research hypotheses were defined:

H1: The accuracy of predictive models developed on the activity data of the social network Facebook users is higher than models based on surveys.

H2: The error-based machine learning approach develops more accurate predictive models than the other three machine learning approaches.

Two research questions were set up:

IP1: Which variables are the most significant predictors of the election outcome?

IP2: To what extent is the temporal component of Facebook social network data important in predicting election outcomes?

The research hypotheses will be tested as follows. Hypotheses H1 will be tested by applying four machine learning algorithms to develop predictive models of election outcomes. The performance of these models will be compared with the results of public opinion polls collected from French media sites that conducted pre-election polls. To test the hypothesis, a statistical significance test is used that does not assume a specific data distribution, which is unknown for data taken from the French media. Hypothesis H2 is tested by applying the error measure and testing the differences in the accuracy of the models obtained by different algorithms. To answer the research question IP1, a sensitivity analysis is carried out on the developed models to determine how the change in the value of the input variables affects the values of the output variable. Thus, the predictors of the election outcomes are determined. Sensitivity analysis is one of the approaches to the explainability of machine learning models. In order to answer research question IP2, predictive models are developed on the data of three campaign periods and their results are compared.

The research follows the steps and activities of the CRISP-DM methodology, which consists of six phases. The first stage is the understanding of the problem, which includes setting up the research objectives and the development of the plan, as well as the definition of the research flow. The second phase is data understanding. The focus of this phase is data collection and initial data description. To conduct empirical research that will achieve the goals of the paper and test the hypotheses, data is collected from the social network Facebook, namely: data from the pages of the candidates in the local elections in France in 2020. The social network Facebook was chosen following the suggestions of previous research, which claim that political campaigns are increasingly "running away" to Facebook because it is the social network with the largest number of users. Furthermore, research shows that for generations of Millennials and Generation X (ages 18 to 51), Facebook is the most used source of political news. In France, which is explored here, the largest part of the population uses the social network, Facebook, and that is 70% of the population of those who use social networks. The choice of Facebook was additionally motivated by recent works that conducted the research only on Twitter, and the guidelines for future research stated the need to include other social networks.

In addition to the data from the social network Facebook, data on the results of pre-election polls are also used.

The data was collected on local elections in France in 2020. All French cities with more than 100 000 inhabitants are included (41 of them). An additional condition for the city's inclusion was that at least two candidates in that city had open pages/profiles. The data was created by the author of the paper by individually downloading data from the pages of the candidates in the elections in June 2020. The names and surnames of the candidates (225 of them) in the cities were taken from the official website of the electoral commission. Each candidate was searched on the social network Facebook to find pages or profiles. For candidates who had pages or profiles open, each piece of content was reviewed and categorized into one of the following categories: event, photo, link, video, and status. For each content, information on the number of contents on the page/profile, the number of content likes, the number of content comments, and the number of content shares was extracted. Initially, 25 variables were included in the research: 24 input variables, they refer to: the candidate's activity and the reactions of page followers to the candidate's activity, two variables related to the candidate's gender and the candidate's political party affiliation. One output variable for the predictive models: the election result measured by the candidate's vote percentage in the election. The attributes were monitored in three time periods: (i) at the beginning of the campaign, (ii) in the middle of the campaign and (iii) on the last day of the campaign. The French local elections were held on March 15, 2020. The official campaign started on March 2, 2020. The first set of data related to the activities at the very beginning of the campaign (from 02.03 to 06.03), the second set of data to the activities of the candidates in the middle of the campaign (07.03 to 11.03), and the third set of data to the very end of the campaign (12.03 to 14.03)). After data collection, an integral part of this phase is: defining variables and types of variables. Attributes that will be taken into account are city and voter turnout in the city, party, total number of page likes, number of photos, number of statuses (text), number of links, number of videos, number of created events, number of photo shares, number of status shares, number of link shares, number of video shares, number of photo likes, number of status likes, number of link likes, number of video likes, number of photo comments, number of status comments, number of link comments, number of video comments, post time. For modeling purposes, new, synthetic variables are created so that all variables related to social network activity are divided by the number of voters in the city where they vote. Most variables have an exponential distribution of values, characterized by a high probability of occurrence of smaller values, and a low

probability of occurrence of large values. A characteristic of such distributions is that the values of the arithmetic mean are higher than the values of the median. Correlation analysis revealed a series of linear relationships between variables. A significant and high correlation (correlation coefficient value r above 0.7) has variables related to sharing, liking, and commenting on a certain type of content: the highest linear correlation was found between the variables' number of event sharing and number of event likes. The correlation is positive, which indicates that with the increase in event likes, the number of event shares also increases, and vice versa. There is an extremely small number of negative correlation coefficients.

The third phase of the CRISP DM standard is data preparation. After previously identifying the sources of available data, data cleaning, attribute transformation, and attribute selection are applied here. The goal of data preparation is to clean the data and select attributes for modeling. The first activity was the identification of the outliers. Outliers were identified because some of the applied machine learning algorithms (artificial neural networks, k-nearest neighbors) are very sensitive to extreme values. Interquartiles were used to identify outliers as follows. A value is marked as an outlier if it is at least 1.5 interquartile's below the first quartile or 1.5 interquartile's above the third quartile. Missing values were created by moving the outliers. Imputation was performed by inserting data instead of missing values. This assigned values to empty spaces based on the following heuristic: missing values were replaced by the mean values of the variables. The next step in data preparation was data normalization. Normalization was performed with two aspects. First, the data were normalized concerning the number of voters in each city. The original value of each variable is divided by the number of voters in the city to which it refers. In the second step, min-max normalization was performed, which reduced the attribute values to a scale from 0 to 1. In the attribute selection activity, only relevant attributes that represent inputs to the models were selected. In this research, a filter approach is used, which ranks the variables and selects a certain subset of variables based on the evaluation measure. As an evaluation measure, the ReliefF algorithm was used. Accordingly, 8 variables from the initial set of variables related to the data of the social network Facebook were selected and will represent the input to the modeling: the total number of page likes, the number of events, the number of status likes, the number of photos, the number of links, the number of photo likes, the number like link, status number. Additionally, the gender and party affiliation attributes of the candidate are included.

The fourth phase is modeling. The modeling step was performed on the prepared and normalized data. The data is divided into a training set and a model testing set. Predictive

models of election outcomes are developed by applying four types of machine learning algorithms: (i) information-based machine learning (algorithm: decision tree), (ii) similarity-based machine learning (algorithm: k-nearest neighbors), (iii) machine learning probability-based learning (algorithm: naive Bayes classifier), (iv) error-based ternary learning (algorithm: artificial neural networks). By applying each of the four machine learning approaches, predictive models were developed on data sets collected in different periods. In this step, the optimization of the hyperparameters of each algorithm was carried out in order to prevent overtraining of the model and to obtain high-quality, reliable, and accurate predictive models.

After the model development, the model evaluation is carried out. When evaluating the model, the results should be evaluated in the context of the goals set up in the first phase. As part of this phase, evaluation is done based on the defined quality metrics of the model. The k-fold cross-validation was used as a validation method. Based on the measures of accuracy (measured through RASE) and model reliability (measured through RSquare), the models were compared to determine which of them gives the best prediction. The model obtained by the neural network has the best quality parameters: it gives a more accurate prediction and is a model of the highest level of reliability. Part of the explanation for these results is hidden in the attribute types. Input attributes are mostly numeric continuous attributes. The models with the lowest reliability and accuracy values were obtained by the naive Bayesian classifier. The explanation of the results can also be related to the characteristics of the data set on which the model was developed. Naive Bayes classifier works with categorical variables, and transformation of continuous output into categorical is necessary. Based on these results, hypothesis H2 is confirmed. By testing hypothesis H1, it was determined that there is no statistically significant difference in the results obtained by predictive models based on data from the social network Facebook and the artificial neural network algorithm compared to pre-election public opinion polls. Models obtained from social network data have higher accuracy than models obtained from surveys, but this difference is not statistically significant. To compare predictive models obtained by machine learning algorithms and the results of pre-election polls, data from the French Institute for Public Opinion Research, IFOP, related to 6 cities: Paris, Lyon, Marseille, Rennes, Nantes and Bordeaux, were used, thus including data for a total of 44 candidates in the elections. Based on these results, hypothesis H1 is partially confirmed.

In order to answer the research question IP2, on each of the three data sets from three different time periods, a predictive model was developed using an artificial neural network algorithm,

which proved to be the best on the entire data set. The table that tests the statistics of the Friedman test, shows that there is no statistically significant difference (at the significance level of 0.01) between the predictive models obtained in the first, second, and third periods.

In order to answer research question IP1, a sensitivity analysis was performed on all four predictive models. The results of the most significant predictors of the results of four predictive models obtained by applying four different approaches to machine learning show the stability and consistency of the model. The total number of likes on the candidate's page is the most significant predictor in all four models. The number of link likes is the second most significant predictor also in all four models. The number of statuses is the third strongest predictor of election results in two of the four observed models.

The last, sixth stage of the CRISP DM methodology is the use of models. In research, this last phase refers to knowledge extraction. The obtained predictive models serve as a tool for campaign planning and management and as such serve as an instrument for decision support through profiling. The most important results of this research provide numerous implications for the use of social media as an indicator for predicting election outcomes, and guidelines can be given for an effective approach to managing social media data. The results of the research also lead to a better understanding of how social networks present the opinions of voters and indicate how the voters can be influenced through social networks. First, it was discovered that the absolute number of Facebook followers is a very good predictor of election outcomes. Secondly, it was determined that the content that the candidates put out through social media, as well as the reactions to the specific content that they share, is to a certain extent an indicator of the outcome of the election. Given the low costs associated with social networking, these results potentially lead to cost reductions in political campaigns. Of course, users of a social network can be friends with or follow many candidates, not just their preferred candidate, including candidates in cities they do not have the right to vote for. Therefore, the number of followers is not a single indicator.

The scientific contribution of the work is contained in the following: (i) systematization of knowledge about predictive models based on data from public opinion polls and social networks, (ii) developed and evaluated predictive models of election outcomes based on social network data collected in French local elections, (iii) determine predictive power of Facebook social network variables, (iv) developed guidelines for using machine learning algorithms on social media data.

The professional contribution of the work is reflected in the following: (i) increased explainability of predicting the election outcome, (ii) reduced time and costs of the election outcome process, (iii) determination of the role of the Facebook social network at the level of local elections.

Obtaining equally effective predictive models with faster and simpler access to data provides a significant scientific and professional contribution to research in this area.

When interpreting the results, one should be careful and take into account several limitations of the research, which, on the other hand, open directions for future research. The research was conducted on a sample of one country, the Republic of France, and on one election held, namely for the local government. Each country, as well as each election, has its specificities that should be taken into account when conducting research. In future research, the resulting models will be tested in other environments and countries. The question of the inclusion of one country leads to the problem of a small sample. Namely, there is a relatively small number of instances in the data set. In order to obtain quality models, the k-cross-validation method was used in the research design, which works well with a small number of instances. When interpreting the results, it should be taken into account that the data refer to a specific time, the beginning of March 2020, right at the very beginning of the spread of the COVID-19 virus pandemic and the lockdown that followed in the second half of March.

Four machine learning algorithms were used in the data analysis, selected from a wide range of available algorithms. Although the selected algorithms represent different approaches to model development, there are still a number of algorithms that could be applied to the data.

SADRŽAJ

1. UVOD

1.1. Predmet istraživanja

Posljednjih godina većina aktivnosti u gospodarskom i javnom sektoru prolazi kroz digitalnu transformaciju. Političke kampanje pritom nisu iznimka. Korištenje novih tehnologija u političkim kampanjama je rezultiralo pojavom brojnih trendova kao što su digitalne kampanje koje generiraju velike količine podataka o aktivnostima kandidata i njihovih potencijalnih glasača. Ova vrsta tehnologije omogućila je praćenje aktivnosti, a algoritmi strojnog učenja počeli su se primjenjivati u svrhu identifikacije znanja o ponašanjima kandidata i potencijalnih birača, pretvarajući ga u važan alat za kandidate i voditelje kampanja. Kao rezultat toga, predviđanja rezultata izbora se više ne moraju raditi isključivo tradicionalnim anketiranjem birača. Usporedba točnosti predviđanja ishoda izbora tradicionalnim anketiranjem birača i korištenjem podataka društvenih mreža na kojima kandidati imaju kreirane stranice, usporedba performansi ra zličitih algoritama strojnog učenja koji analiziraju podatke društvenih mreža te identifikacija točnosti prediktivnih modela u različitim vremenskim periodima (Cameron et al., 2015) tri su smjera istraživanja najavljena kao važna za istražiti u prethodnim istraživanjima, a u fokusu su ovog rada.

1.2. Ciljevi istraživanja, hipoteze i istraživačka pitanja

Ciljevi istraživanja su:

- utvrditi kolika je prediktivna moć modela temeljenih na društvenoj mreži Facebook i usporediti je s drugim vrstama istraživanja,
- utvrditi koje varijable su najznačajniji prediktori ishoda izbora,
- utvrditi značajnost temporalne komponente podataka društvene mreže Facebook,
- utvrditi koja od četiri metode strojnog učenja daje najtočnije prediktivne modele ishoda lokalnih izbora.

Hipoteze istraživanja:

H1:Točnost prediktivnih modela razvijenih na podacima aktivnosti korisnika društvene mreže Facebook veća je od modela temeljenih na anketama.

H2: Pristup strojnom učenju temeljen na pogrešci daje točnije prediktivne modele od ostala tri pristupa strojnom učenju.

Postavljena su i dva istraživačka pitanja:

IP1: Koje su varijable najznačajniji prediktori ishoda izbora?

IP2: U kojoj je mjeri temporalna komponenata podataka društvene mreže Facebook važna u predikciji ishoda izbora?

Hipoteze istraživanja će se testirati na način kako slijedi. Primjenom četiri metode strojnog učenja razvit će se prediktivni modeli ishoda izbora. Performanse tih modela usporedit će se s rezultatima ispitivanja javnog mnijenja prikupljenih sa stranica francuskih medija koji su radili predizborna ispitivanja javnog mnijenja. Za testiranje hipoteze koristit će se testovi koji ne pretpostavljaju određenu distribuciju podataka, koja za podatke preuzete od francuskih medija nije poznata.

Hipoteza H2 provjeravat će se primjenom mjere greške i testiranjem razlika u točnosti modela dobivenih različitim algoritmima.

Kako bi se odgovorilo na istraživačka pitanja provest će se analiza osjetljivosti nad razvijenim modelima kako bi se utvrdilo kako promjena vrijednosti ulaznih varijabli utječe na vrijednosti izlazne varijable. Na taj će se način utvrditi prediktori ishoda izbora. Analiza osjetljivosti jedan je od pristupa objašnjivosti modela strojnog učenja.

Znanstveni doprinos rada bit će sadržan u sljedećem:

(i) sistematizacija znanja o prediktivnim modelima temeljenim na podacima ispitivanja javnog mnijenja i društvenih mreža,

(ii) razvijeni i evaluirani prediktivni modeli ishoda izbora temeljem podataka društvene mreže prikupljenih na francuskim lokalnim izborima,

(iii) utvrđena prediktivna moć varijabli društvene mreže Facebook,

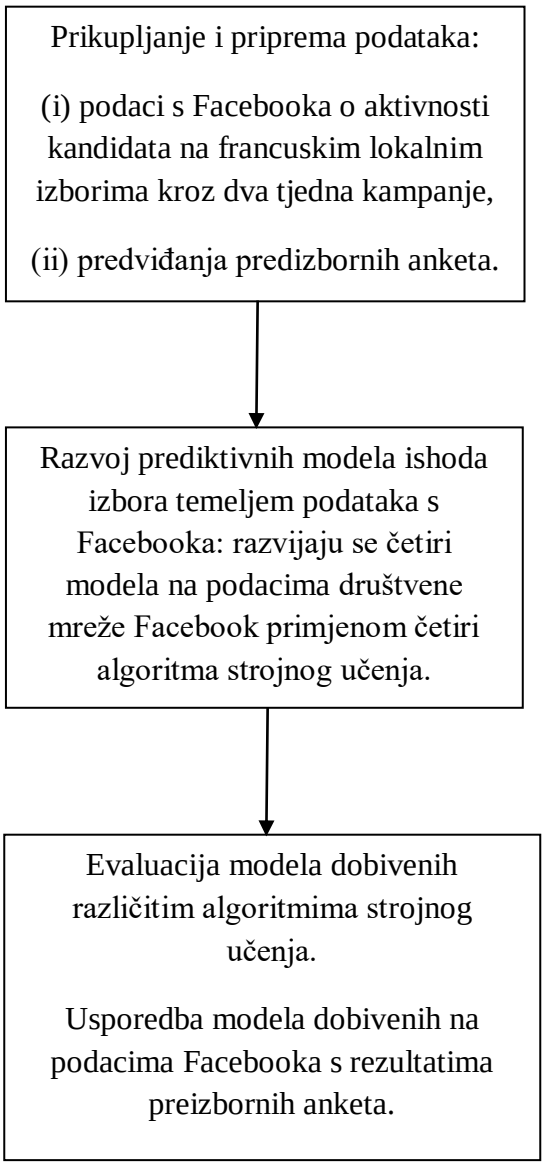
(iv) razvijene smjernice za korištenje algoritama strojnog učenja na podacima društvenih medija.

Stručni doprinos rada ogleda se u sljedećem:

(i) povećana objašnjivost predviđanja ishoda izbora,
(ii) smanjeno vrijeme i troškovi procesa ishoda izbora,
(iii) utvrđivanje uloge društvene mreže Facebook na razini lokalnih izbora.

1.3. Struktura rada

Tijek aktivnosti i proces istraživanja prikazan je kroz dijagram tijeka u nastavku ove sekcije.



Slika 1. Tijek istraživanja

Rad je organiziran kako slijedi. U poglavlju 2 daje se pregled relevantne i recentne literature temeljem kojeg se identificiraju nedostaci prethodnih istraživanja te definiraju ciljevi ovog rada. U poglavlju 3 opisuje se metodologija rada kao i podaci koji će se koristiti u istraživanju. Poglavlje 4 daje rezultate, a poglavlje 5 kroz diskusiju dobivenih rezultata daje odgovore na postavljena istraživačka pitanja i hipoteze istraživanja.

2. PREGLED PRETHODNIH ISTRAŽIVANJA

Internet je omogućio nove načine povezivanja, komunikacije i distribucije informacija. Sudionici političkih kampanja uvijek su koristili nove i inovativne načine komunikacije kako bi došli do birača. Ranije novine, radio, televizija, a danas Internet, u velikoj mjeri mijenjaju političku komunikaciju. Otkako je Internet postao „mainstream“, privlači pozornost istraživača i u domeni političke komunikacije. Novu dimenziju istraživanjima u ovom području nosi pojava platformi društvenih medija (Strömbäck & Esser, 2009). Mnogi autori navode da društveni mediji nose drastične promjene u način kreiranja, distribucije i mjerenja političke komunikacije. Dinamičke interakcije i složene međuovisnosti na različitim razinama i dimenzijama koje nose društveni mediji predstavljaju izazov tradicionalnom razumijevanju političke komunikacije (Strömbäck & Esser, 2009). Brzo širenje i primjena društvenih medija u političkim kampanjama diljem svijeta potaknulo je znanstvenike da istraže kako korištenje ove tehnologije utječe na političku orijentaciju, participaciju i stavove birača. Društveni mediji predstavljaju vrijedan izvor podataka koji omogućuju ispitivanje tog utjecaja. Društveni mediji pružaju velike baze svakodnevnih misli i osjećaja ljudi u mjeri koja je donedavno bila nezamisliva. Budući da je ponašanje korisnika na društvenim medijima odraz događanja u stvarnom svijetu, znanstvenici su prepoznali potencijal tih podataka za korištenje u svrhu predviđanja budućnosti. Prednost je tih podataka relativna jednostavnost stjecanja, njihova velika količina, i sposobnost za obuhvaćanja društveno relevantnih informacija, što može biti teško prikupiti iz drugih izvora podataka (Phillips et al., 2017). Velike količine podataka o korisnicima i njihovoj društvenoj interakciji pružaju znanstvenicima nove smjerove za istraživanja. Neki od najvažnijih radova predviđanja budućih događaja i razvoja temeljem podataka društvenih medija napravljeni su u domeni financija, zabave i zdravstva, a posebno u domeni političkih kampanja (Schoenfeld, 2020). Ovo potonje upravo i jeste tema ovog rada.

U nastavku će se rada prikazati pregled dosadašnjih istraživanja koja se bave navedenom tematikom.

2.1. Primjena društvenih medija u političkim kampanjama

Buettner (Buettner, 2016) definira društvene medije kao računalom posredovane alate koji omogućuju kreiranje, dijeljenje ili razmjenu informaciju, ideja, slika ili videa u virtualnim zajednicama i mrežama. Postoje razni oblici društvenih medija: blogovi (Blogspot, LiveJournal), forumi (Yahoo! answers, Epinions), mediji za dijeljenje sadržaja (Flickr, YouTube, Digg, Reddit), mikroblogovi (Twitter, foursquare) i društvene mreže (Facebook, Myspace, LinkedIn, Twitter, Tribe), kao najpoznatiji oblik društvenih medija (Wright et al., 2009). Ljudi su sve više integrirali društvene mreže u svoje živote. Društvene mreže imaju značajan utjecaj na ponašanje ljudi u odnosu s drugima, njihove kupovne navike, mišljenja i preferencije (Asur & Huberman, 2010) te su, kao takve, izrazito važne i korisne za poslovanje (Stefko et al., 2011). U različitim granama poslovanja, društvene mreže se koriste za održavanje odnosa s javnošću (Eyrich et al., 2008). I neprofitne organizacije sve više usvajaju korištenje društvenih mreža kao alata za odnose s javnošću (Curtis et al., 2010). Društvene mreže nude brojne mogućnosti za interakciju s javnošću, usvajanje novih oblika tehnologije i integriraju ih u svakodnevni život. Profesionalci odnosa s javnošću (PR) u velikoj mjeri koriste tehnološke napretke (Eyrich et al., 2008). Kako se pojavljuje sve više oblika društvenih mreža, PR profesionalcima je važno razumijeti te alate kako bi ih koristili u poslu. Kako se pozicija društvenih mreža mijenja iz statusa "*buzz word*" prema poziciji strateškog alata u komunikaciji, sve više PR profesionalaca razvija vještine vezane uz ovu tehnologiju *on-line* komunikacije (Eyrich et al., 2008). Prema istraživanju PRSA Wired for Change Survey, većina PR profesionalaca smatra da upotreba informacijske i komunikacije tehnologije olakšava njihov posao ubrzavanjem cirkulacije informacija i dosezanjem šire pub*like* (Gayo-Avello et al., 2011). Društveni mediji PR profesionalcima daju mogućnost uključivanja građana u dijalog i otvaraju put za jačanje odnosa s medijima. (Taylor & Kent, 2010) predlažu da se društvene mreže koriste kao alat za odnose s javnošću, a Mangold i Faulds (Mangold & Faulds, 2009) tvrde da društvene mreže predstavljaju hibridni element promocijskog miksa te su stoga izrazito zanimljiv i koristan alat za odnose s javnošću. Upotreba društvenih mreža za održavanje odnosa s javnošću nije ograničena samo na korporacije, već se komponenta društvenih mreža kao vrijednog alata za oglašavanje pokazala kao vrlo moćna platforma za izražavanje mišljenja širom svijeta. Prethodna istraživanja pokazuju da većina političkih kandidata koristi neki oblik društvenih mreža u svojoj kampanji ((Digrazia et al., 2013);(Gulati,Girish; Williams & Christine, 2013); (Hong &

Nadler, 2012); (Lilleker et al., 2011); (Strandberg, 2013); (Safiullah et al., 2017)). Štoviše, najnovija istraživanja dokazuju utjecaj društvenih medija na političko znanje i percepcije o politici (Lee & Xenos, 2019).

2.2. Uloga podataka društvenih mreža u političkim kampanjama

Difuzija društvenih mreža omogućila je nove načine mjerenja javnog mnijenja. Predviđanje ishoda izbora pomoću "velikih" podataka društvenih mreža novi je oblik političkog predviđanja i uglavnom se koristi za izradu predviđanja o rezultatima izbora prikupljanjem relevantnih podataka iz društvenih mreža. Neki autori sugeriraju da bi analiziranje podataka prikupljenih od društvenih mreža tijekom predizborne kampanje mogao biti koristan dodatak tradicionalnim metodama anketiranja (Tjong et al., 2012); (Schober et al., 2016). Glavne prednosti mjerenja javnog mnijenja putem društvenih medija su dostupnost i brzina (Schober et al., 2016). Pored toga, u usporedbi s tradicionalnim anketama, društveni mediji mogu omogućiti kontinuirano praćenje javnog mnijenja u stvarnom vremenu (Tjong et al., 2012). Prema tome, "pristup podacima" kroz društvene medije mogao bi ponuditi rješenje ograničenja tradicionalnih anketa.

Političari sve više koriste društvene medije u kampanjama i time kreiraju takve podatke. Bivši američki predsjednik Barack Obama okvalificiran je kao „prvi Internet predsjednik“, jer se smatra da je u njegovoj kampanji korišten Internet i informacijska tehnologija kao ni u jednoj političkoj kampanji do tada i da je ta kampanja promijenila način interakcije između političara i glasača (Greengard, 2009). Od tada pa do danas istraživanje važnosti društvenih medija za izbore dobiva sve više pozornosti. Ključna tema povezana s predviđanjem izbora pomoću društvenih medija je upotreba društvenih medija koje koriste političari. Prethodna istraživanja sugeriraju da su političari usvojili društvene medije i da postoje različiti čimbenici koji utječu na obrasce njihove upotrebe (Strandberg, 2013). Recentna istraživanja su provedena s ciljem utvrđivanja utjecaja društvenih medija na političke kampanje iz nekoliko perspektiva: političke participacije, političkog znanja i političke efikasnosti (Rahmawati, 2014). Ono što je karakteristika svima jest uvjeravanje, što je i ključno za političke kampanje. Svaki govor, svaki telefonski poziv, svako kucanje na vrata, svaka objava putem društvenih medija ima u cilju propagandu stvorenu da bi se utjecalo na birače. Pitanje je u kojoj mjeri to utječe na birače i je li ta propaganda efikasna? Vepsäläinen i Suomi (Vepsäläinen et al., 2017) tvrde da je nekoliko istraživanja provedeno o utjecaju društvenih medija na ishode izbora i da je temeljem njih teško donositi zaključke o tome mogu li se podaci društvenih medija,

npr. podaci s Facebooka, koristiti za predviđanje rezultata izbora. Predviđanje izbornog rezultata pomoću "velikih" podataka društvenih medija je nova tema istraživanja koja je tek u nastajanju zbog eksponencijalnog rasta društvenih medija. Naime, na izborima je tradicionalno anketiranje dominantno sredstvo za predviđanje rezultata političkih izbora. U tradicionalnim ispitivanjima predviđanje se vrši putem reprezentativnog uzorka. Prema Metaxasu, Mustafaraju i Gayo-Avellu, najvažniji aspekt uspješnog ili točnog ispitivanja je odabir uzorka koji dobro predstavlja potencijalne birače (Metaxas et al., 2011). Podaci društvenih medija se razlikuju od podataka prikupljenih u anketi (Gayo-Avello et al., 2011). Prethodna istraživanja koja koriste podatke društvenih medija uglavnom koriste statističke modele za predviđanje rezultata izbora. Manji broj istraživanja kombinira podatke društvenih medija s drugim relevantnim varijablama za predviđanje rezultata izbora, kao što su ekonomska uspješnost ili politički indeksi (Metaxas et al., 2011).

Revolucionarni uspon društvenih medija privukao je pozornost znanstvenika u različitim disciplinama. U nekoliko su prethodnih istraživanja korišteni podaci društvenih medija za predviđanje rezultata u različitim domenama u stvarnom svijetu. Rezultati tih istraživanja sugeriraju da su društveni mediji važan pokazatelj ponašanja korisnika i usmjeravanja njihovih preferenci u domenama kao što su predviđanja prodaje, promocije mrežnih stranica i političkih izbora (Cameron et al., 2015); (Gerodimos & Justinussen, 2014); (Kudeshia et al., 2016); (Petrocchi et al., 2015)). Većina istraživanja u području izbora usredotočena su na Twitter (npr. (Franch, 2013); (Gayo-Avello, 2012a); (Gayo-Avello, 2012b)), a puno manje na Facebook (npr. (Vitak et al., 2011)). Dosadašnja istraživanja rađena su na američkim predsjedničkim izborima 2008. (*The Internet's Role in Campaign 2008 | Pew Research Center*, n.d.), parlamentarnim izborima u Novom Zelandu 2011. godine (Cameron et al., 2015), i 2010. na švedskim izborima (Larsson & Moe, 2011).

Koji su učinci kampanja na društvenim mrežama i kako se mjere istražujemo u sljedećoj sekciji rada.

2.3. Metodološki pristupi mjerenju utjecaja društvenih mreža na ishode političkih izbora

Pregled literature je pokazao da postoji potreba da se prati sadržaj koji se objavljuje te da se analizira utjecaj tog sadržaja na ciljanu publiku. Danas, u eri velikih podataka, to je i moguće. Pristup podacima omogućio je pomak od pukog procjenjivanja prema odlukama i postupcima koji se temelje na podacima. Kako su se, i s kojom svrhom, u dosadašnjim istraživanjima analizirali podaci?

Housholder i LaMarre (Housholder & LaMarre, 2015) povezuju očekivanja od kampanje na društvenim medijima s podacima o participaciji na izborima koja je nastala kao rezultat aktivnosti na društvenim medijima. Rezultati daju indicije da angažman na društvenim medijima ima pozitivan utjecaj na kampanju. Autori ističu da angažman na društvenim mrežama može pomoći i u predviđanju rezultata izbora. Housholder i LaMarre (Housholder & LaMarre, 2015) smatraju da buduća istraživanja trebaju ispitati u kojoj mjeri određene vrste angažmana na društvenim medijima rezultiraju željenim ishodima.

Guleria i suradnici (Guleria et al., 2016) u radu iz 2016. debatiraju o utjecaju elektronske kampanje na ponašanje glasača, njihovu svijest i poimanje političkih stranaka. Ističu prednosti ovakvih kampanja u smislu šire pokrivenosti u odnosu na tradicionalne tehnike. Njihovo istraživanje je kvalitativno. U smjernicama za buduća istraživanja ukazuju na potrebu za provođenjem kvantitativnog istraživanja za dokazivanje odnosa između elektronskih medija i političke participacije (Guleria et al., 2016).

Chen i Chang ispituju povezanost korištenja Facebooka i bloga s motivacijom za informacijom i političkom diskusijom (Chen & Chang, 2017). Rezultati regresijske analize su pokazali da je želja za političkim diskusijama značajan prediktor korištenja blogova, dok je motivacija za informacijom povezana s upotrebom Facebooka.

I časopis Nature objavljuje istraživanja na ovu temu. Bond i suradnici su proveli istraživanje (Bond et al., 2012) političkih poruka korisnicima Facebooka za vrijeme američkih izbora za Kongres 2010. godine. Rezultati pokazuju da poruke koje političari šalju putem društvenih medija imaju direktan utjecaj na krajnji rezultat, ne samo osoba koje su poruke čitale, nego i osoba koje su s njima u interakciji. U interpretaciji rezultata naglašavaju važnost velikog broja sljedbenika na društvenim mrežama kao čimbenika uspjeha na izborima (Bond et al., 2012). Vepsäläinen, Li, i Suomi istražuju u kojoj mjeri se Facebook *likeovi* mogu koristiti za predviđanje izbornih rezultata

(Vepsäläinen et al., 2017) . Hong i Nadler su istraživali da li, i u kojoj mjeri, upotreba Twittera u političke svrhe ima potencijal da utječe na javno mnijenje (Hong & Nadler, 2012). U istraživanju autori analiziraju aktivnosti američkih predsjedničkih kandidata na Twitteru. Njihovi rezultati su pokazali da aktivnost kandidata na Twitteru nije statistički značajno povezana s brojem spominjanja njihovih imena na Twitteru (Hong & Nadler, 2012).

Pregledom recentnih članaka utvrđeno je kako je malo studija provedeno o utjecaju društvenih medija na same rezultate izbora. Većina se tih studija fokusira na opisivanje društvenih medija kao marketinškog alata političarima (Ahmad & Popa, 2014). Za razumijevanje utjecaja društvenih medija i realnu ocjenu učinkovitosti tog alata nužno je mjeriti upotrebu društvenih medija u svrhu političkih aktivnosti (Rahmawati, 2014).

Borah (Borah, 2016) govori o povećanju istraživanja o primjeni stranica društvenih medija u političke svrhe, ali i ističe da mali broj tih istraživanja ispituje sadržaj stranica kandidata. Kao nedostatke prethodnih istraživanja navodi izostanak ispitivanja povezanosti sadržaja stranica s ishodom izbora.

U istom smjeru idu i napuci Praudea i Skulmea (Praude & Skulme, 2015) koji ističu nužnost razvijanja metrika za mjerenje učinkovitosti duštvenih medija i poruka koje se šalju putem društvenih medija. Effing je svom doktoratu iz 2014. išao u tom smjeru te razvio okvir za mjerenje učinaka društvenih medija na neprofitne organizacije (Effing et al., 2012). U radu je kreiran okvir koji se temelji na izračunu Social Media Indexa, a testira se na tri studije slučaja u domenama crkvene zajednice, apostolske zajednice i gradskog vijeća. Rezultati testiranja su ukazali na određene nedostatke predloženog okvira i autor daje smjernice kako ga poboljšati i prilagoditi drugim domenama. Rao (Rao, 2016) također zaključuje da se ne može ista metrika primjenjivati u različitim domenama.

Khairuddin i Rao (Khairuddin & Rao, 2017) prepoznaju da sve više političkih kandidata koristi društvene medije za kampanju, te ističu važnost mjerenja učinkovitosti ovog medija. Tvrde da se većina dosadašnjih istraživanja, koncentrira samo na spominjanja na Twitteru u pokušajima da povežu prisutnost na društvenim medijima i pobjedu. U svom radu predlažu mjerenje pasivnih interakcija između postova kandidata na Facebooku i korisnika (kojima se sviđaju objave) te u tom kontekstu ističu potrebu za mjerenjem broja komentara i povezivanjem s rezultatima izbora.

Nadalje, što preciznije prognoziranje ishoda izbora u ranoj fazi kampanje izuzetno je važno jer marketinške aktivnosti odlučuju o konačnom uspjehu na izborima. Kandidati na izborima kao i

stručnjaci uključeni u kampanje mogu imati višestruke koristi od točnosti prognoze, što bi im omogućilo da donose odgovarajuće upravljačke odluke koje se tiču raspodjele proračuna za dodatne marketinške aktivnosti i fokus na segmente koji vode prema uspjehu.

Dos Santos Brito i suradnici (2021) nedavno su proveli sistematski pregled literature na temu predviđanja ishoda izbora temeljem podataka društvenih mreža. Jedno od istraživačkih pitanja odnosilo se na primjenjenu metodologiju. Pregled je pokazao da se nakon analize ekstrahiranih podataka, identificirani metodološki pristupi mogu grupirati u pet skupina: 1) analiza sentimenta; 2) regresija ili vremenska serija; 3) interakcije profila ili postova; 4) analiza teme; i 5) ostalo. Od toga je analiza sentimenta korištena u 50% slučajeva, a regresijska analiza u 20% slučajeva.

2.4. Nedostaci prethodnih istraživanja

Temeljem pregleda literature moguće je sumirati nedostatke dosadašnjih istraživanja:

- nije razvijen model za mjerenje učinkovitosti političkih kampanja na društvenim mrežama: varijable koje su prikladne za vrednovanje u jednoj specifičnoj domeni nisu prikladne za vrednovanje u drugoj domeni,
- u većini istraživanja su se podaci prikupljali samo na Twitteru,
- nisu razvijeni modeli koji bi utvrdili u kojoj mjeri aktivnost na društvenoj mreži Facebook doprinosi predikciji ishoda izbora,
- veliki broj modela se temelji isključivo na broju prijatelja ili sljedbenika na društvenim mrežama, što je premali broj varijabli da objasni aktivnost na društvenoj mreži: zaključci su generalizirani temeljem premalog broja varijabli koje ne uključuju interakciju između kandidata i potencijalnog birača,
- pojedina istraživanja su se temeljila na premalom uzorku ispitanika,
- većina modela se temelji na sentiment analizi koja nije dovoljna da utvrdi utjecaj na ishod izbora,
- recentna istraživanja su dokazala da dosadašnje metode nisu dovoljne i ukazala na nužnost uključivanja pristupa strojnog učenja,
- mnogi modeli nisu u dovoljnoj mjeri teorijski utemeljeni niti empirijski potvrđeni.

Društveni mediji danas su sveprisutni, a njihov utjecaj će se osjetiti na svim područjima ljudskog djelovanja u vremenu, pa tako i u sferi političkog djelovanja. Kako bi komunikacija kroz društvene medije polučila željene učinke, potrebno je identificirati ključne odrednice te komunikacije, oblikovati instrumente mjerenja te vrednovati učinkovitost komunikacije političkih kandidata kroz društvene mreže. Predviđanje rezultata izbora podacima društvenih medija plodonosno je područje, ali su svakako potrebna daljnja istraživanja. Identifikacija uočenih nedostataka vodila je postavljanju ciljeva u ovom radu. Dodatni podstrek ovom istraživanju daju Kalampokis i suradnici (Kalampokis et al., 2013) koji tvrde da je primjena podataka društvenih medija u domeni politike izrazito kompleksna jer uključuje međusobno povezane entitete stvarnog svijeta. Analiza takvih skupova podataka je izazovna i zahtijeva sofisticirane metode koje će se i koristiti u ovom istraživanju.

3. METODOLOGIJA ISTRAŽIVANJA

Dvije su primarne vrste zadaća rudarenja podataka, predikcija i deskripcija. Cilj je izrade prediktivnih modela predvidjeti na temelju povijesnih podataka, a najčešći su zadaci klasifikacija i regresija. Regresija se koristi za predviđanje numeričkih vrijednosti, dok se klasifikacija koristi za predviđanje kategorijskih vrijednosti. Opisno istraživanje podataka odnosno deskriptivni modeli se koriste za organiziranje podataka i njihovo bolje razumijevanje. Generalno se dijele na identifikaciju asocijacija i grupiranje.

Proces rudarenja podataka dinamičan je i izazovan te je razvijeno niz standarda koji propisuju korake procesa rudarenja podataka. Neki od njih su: CRISP-DM (Cross Industry Standard Process for Data Mining), KDD (Knowledge Discovery in Data), SEMMA (Sample, Explore, Modify, Model, Assess).

Metodologija ovog istraživanja temelji se na CRISP-DM standardu koji je posebno prikladan za istraživačke projekte temeljene na podacima kao što je to pokazamo u ranijim radovima od (Moro et al., 2014), a i nekoliko nedavnih studija na temeljenih na podacima koje su usvojili npr. (Almahadeen, Akkaya & Sari, 2017) Ovaj se pristup temelji na iterativnom slijedu od šest faza (razumijevanje poslovanja, razumijevanje podataka, priprema podataka, modeliranje, evaluacija, implementacija) s ciljem kreiranja podatkovnog modela koji će adekvatno prikazati promatrani problem prema metrikama procjene. CRISP-DM sugerira tok slijeda, a metodologija je fleksibilna. U istraživanju se obično faza implementacije zamjenjuje ekstrakcijom znanja kako bi se razumio zadani problem (Moro et al., 2014).

CRISP-DM pristup će se nadopuniti Kalampokisovim okvirom za analizu podataka društvenih medija, koji je specifično razvijen i prilagođen za podatke društvenih medija.

S ciljem utvrđivanja točnosti i pouzdanosti prediktivnih modela temeljenih na podacima društvene mreže Facebook u dva vremenska perioda i njihove usporedbe s prediktivnim modelima temeljenih na podacima ispitivanja javnog mnijenja provest će se istraživanje slijedeći dvije faze okvira za analizu podataka društvenih medija kojeg su razvili Kalampokis i suradnici (2013): faza pripreme podataka i faza prediktivnog modeliranja. Svaka od te dvije faze sastoji se od dva koraka. Faza pripreme podataka uključuje: (i) prikupljanje i filtiranje podataka, te (ii) izračun prediktorskih

varijabli. Faza prediktivnog modeliranja uključuje: (i) izradu prediktivnih modela te (ii) evaluaciju prediktivnih modela.

3.1. CRISP-DM standard

(Schröer et al., 2021) su proveli sistematski pregled literature na temu primjene CRISP-DM u istraživanjima. Rezultati pregleda koji se temelji na Web of science i Scopus rezultatima pokazuju da većina autora definira CRISP-DM kao *de-facto* standard za primjenu u projektima rudarenja podataka. Većina radova uspoređuje različite modele (npr. sa SEMMA i KDD) prije nego što su se odlučili za CRISP-DM. Niz autora opisuje CRISP-DM kao jednostavan i strukturiran, pouzdan, često korišten i industrijski neovisan model procesa. Temeljem toga, u ovom se istraživanju primjenjuje CRISP-DM standard, a u ovom poglavlju opisuju se glavne zadaće koje je potrebno provesti za rješavanje svake CRISP-DM faze.

CRISP-DM daje potpuni plan i prikazuje životni ciklus procesa rudarenja podataka u sljedećih šest faza (Wirth & Hipp, 2000):

1) Razumijevanje problema (engl. Domain/Business Understanding)
2) Razumijevanje podataka (engl. Data Understanding)
3) Priprema podataka (engl. Data Preparation)
4) Modeliranje (engl. Modelling)
5) Vrednovanje (engl. Evaluation)
6) Korištenje, stavljanje u produkciju (engl. Deployment)

U nastavku će biti prikazan pregled CRISP-DM metodologije koja se sastoji od ovih šest faza koje bi se trebale ciklično odvijati. Prva faza je **razumijevanje problema** koje uključuje određivanje ciljeva istraživanja, procjenu trenutne situacije i razvoj plana projekta. Rudarenje podataka se vrlo često predstavlja kao tehnički problem pronalaženja modela koji objašnjavaju odnos ciljne varijable i grupe ulaznih varijabli. Cilj je ove faze razumijevanje poslovne perspektive, a zatim prevođenje problema i ciljeva domene u problem rudarenja podataka. Cilj je ovog istraživanja utvrditi kolika je prediktivna moć modela ishoda izbora temeljenih na društvenoj mreži Facebook i usporediti je s drugim vrstama istraživanja te utvrditi koje su varijable najznačajniji prediktori ishoda izbora.

Druga faza odnosi se na **razumijevanje podataka**. Fokus ove faze razumijevanja podataka je prikupljanje podataka i početno opisivanje podataka. Stoga ova faza obuhvaća početno prikupljanje podataka, opis podataka, istraživanje i provjeru kvalitete podataka. Izvori podataka razlikuju se od problema do problema i od industrije do industrije. Primjeri korisnih izvora podataka su podaci web logova, podaci društvenih mreža, podaci koji su prikupljeni anketiranjem, demografski podaci i sl.

Za provođenje empirijskog istraživanja kojim će se ostvariti ciljevi ovog rada i testirati hipoteze prikupljaju se podaci s društvene mreže Facebook, i to: podaci sa stranica kandidata na lokalnim izborima u Francuskoj 2020. godine. Društvena mreža Facebook je odabrana sukladno sugestijama Chena i Changa (2017) koji tvrde da političke kampanje sve više „bježe“ na Facebook jer je to društvena mreža s najvećim brojem korisnika. Nadalje, istraživanja pokazuju da je generacijama tzv. Milenijalaca i generaciji X (dob od 18 do 51 godine) Facebook najčešće korišteni izvor političkih vijesti (Pew Research Center, 2015). U Francuskoj, koja se obrađuje u ovom radu, najveći dio stanovništva koristi društvenu mrežu, Facebook, i to 70% stanovnika od onih koji koriste društvene mreže (Chaffey, 2019). Odabir Facebooka dodatno je motiviran i recentnim radom Singha i suradnika (Singh et al., 2020) koji su istraživanje provodili samo na Twitteru, a u smjernicama za buduća istraživanja navode potrebu za uključivanjem drugih društvenih mreža. Uz podatke društvene mreže Facebook koristit će se i podaci o rezultatima predizbornih anketa.

Nakon prikupljanja podataka, sastavni dio ove faze je: definiranje varijabli i tipova varijabli. Atributi koji će se uzimati u obzir su: grad i izlaznost birača u gradu, stranka, ukupan broj *likeova* stranice, broj fotografija, broj statusa (tekst), broj poveznica, broj videa, broj kreiranih događaja, broj dijeljenja fotografija, broj dijeljenja statusa, broj dijeljenja poveznica, broj dijeljenja videa, broj *likeova* fotografija, broj *likeova* statusa, broj *likeova* poveznica, broj *likeova* videa, broj komentara fotografije, broj komentara statusa, broj komentara poveznica, broj komentara videa, vrijeme objave. Za potrebe modeliranja kreirat će se nove, sintetičke varijable tako da se sve varijable koje se odnose na aktivnost na društvenoj mreži podijele s brojem birača u gradu u kojem glasaju.

Zavisna varijabla, rezultat na izborima, prikazat će se kao postotak osvojenih glasova. Atributi su se pratili u tri vremenska perioda: (i) na početku kampanje, (ii) u sredini kampanje i (iii) zadnji dan kampanje. Francuski lokalni izbori su održani 15.03.2020. godine. Službena kampanja je započela

02.03.2020. Prvi skup podataka odnosio se na aktivnosti u samom početku kampanje (od 02.03. do 06.03.), drugi skup podataka na aktivnosti kandidata u sredini kampanje (07.03. do 11.03.), te treći skup podataka na sam završetak kampanje (12.03. do 14.03). Istraživanjem će se obuhvatiti svi gradovi u Francuskoj koji po službenom popisu stanovništva imaju više od 100 000 stanovnika (ukupno 41 grad).

Treća faza je **priprema podataka**. Nakon prethodnog identificiranja izvora dostupnih podataka, ovdje slijedi čišćenje podataka, deskriptivna analiza podataka (srednje vrijednosti atributa, mjere varijacije i asimetrije), ispitivanje distribucija vrijednosti atributa, transformacija atributa i selekcija atributa. Cilj je pripreme podataka očistiti podatke, generirati značajke, te odabrati značajke za modeliranje.

Prikupljeni podaci nisu uvijek prikladni za provođenje procesa strojnog učenja. Stoga je potrebna transformacija podataka. Priprema podataka uključuje mnoge različite zadatke, poput skaliranja, imputiranja nedostajućih vrijednosti, identifikacija vrijednosti koje strše te transformacije tipova podataka.

Tehnike skaliranja osiguravaju da su numerički atributi proporcionalno ponderirani po algoritmima. Izrazi koji se koriste za tu tehniku su standardizacija i normalizacija. U ovom će se istraživanju provesti normalizacija numeričkih vrijednosti prije provedbe algoritama k-sredina i neuronskih mreža jer su ti algoritmi osjetljivi na velike raspone vrijednosti.

Nadalje, provest će se identifikacija nedostajućih vrijednosti, a ovisno o rezultatima, i imputacija vrijednosti koje nedostaju ili će se maknuti atributi ukoliko će biti više od 50% nedostajućih vrijednosti za neki atribut.

Transformacija atributa predstavlja pretvaranje atributa iz numeričkih u kategorijske i obrnuto, ovisno o potrebama algoritma koji se koristi za razvoj prediktivnih modela. Npr. algoritam neuronske mreže i k-najbližih susjeda zahtijevaju numeričke atribute, dok naivni Bayesov klasifikator zahtijeva kategorijski izlazni atribut.

Četvrta faza je **modeliranje**. Podatke je u ovoj fazi potrebno podijeliti na one koji služe za treniranje i one koji služe za testiranje modela. U sklopu modeliranja odabiru se algoritmi koji će koristiti za razvoj modela nad podacima koji su pripremljeni te se podešavaju parametri algoritama

za dobivanje optimalnih modela. Prediktivni modeli ishoda izbora izradit će se primjenom četiri vrste algoritama strojnog učenja (Kelleher & Namee, 2015):

(i) Strojno učenje temeljeno na informaciji (metoda: stablo odlučivanja).

(ii) Strojno učenje temeljeno na sličnosti (metoda: k-najbližih susjeda),

(iii) Strojno učenje temeljeno na vjerojatnosti (metoda: naivni Bayesov klasifikator),

(iv) Strojno učenje temeljeno na pogrešci (metoda: neuronske mreže).

Primjenom svakog od četiri pristupa strojnom učenju razvit će se prediktivni modeli na skupovima podataka prikupljenih u različitim vremenskim periodima.
Optimizacija hiperparametara jedna je od najvažnijih aktivnosti ovog koraka procesa rudarenja podataka. Proces optimizacije specifičan za svaki algoritam strojnog učenja opisuje se kasnije. Nakon izrade modela slijedi faza **evaluacije modela**. Peta faza odnosi se na evaluaciju modela kod koje rezultate modela treba vrednovati u kontekstu ciljeva postavljenih u prvoj fazi. To dovodi do identifikacije drugih potreba najčešće kroz raspoznavanje uzoraka, te će se uputiti na potrebu vraćanja na prethodne faze. U sklopu ove faze radi se vrednovanje temeljem definiranih metrika kvalitete modela. Kao metoda validacije koristit će se unakrsna validacija s *k*-preklapanja (engl. *k-fold cross validation*). Temeljem mjera točnosti (mjerena kroz RASE) i pouzdanosti modela (mjerena kroz RSquare) komparirat će se modeli kako bi se utvrdilo koji od njih daje najbolju predikciju.

Prilikom razvoja svakog od četiri modela testirat će se odabir hiperparametara kako bi se postigli optimalni rezultati za svaki od predloženih metoda.

Prije konačnog uvođenja, ovaj će korak procijeniti jesu li ispunjeni poslovni ciljevi. Šesta faza koja je ujedno i posljednja u ovoj metodologiji je **korištenje** modela. Modeli se koriste kako bi se potvrdile ranije postavljene hipoteze ili za otkrivanje znanja i predikciju.

U zadnjoj se fazi radi planiranje implementacije, izrada konačnog izvješća i pregled projekta kako bi se procijenilo što je dobro napravljeno, a što nije. U istraživanjima se ova zadnja faza odnosi na ekstrakciju znanja.

Nadalje, kako bi se utvrdilo koji od pristupa strojnom učenju daje najbolje rezultate, komparirat će se rezultati koristeći mjere kvalitete modela te će se testirati razlika u točnosti i pouzdanosti prediktivnih modela dobivenih različitim pristupima strojnom učenju. Ovo istraživanje, temeljeno na podacima aktivnosti kandidata i pratitelja društvene mreže Facebook, usporedit će se s rezultatima temeljenih na ispitivanju javnog mnijenja. Pristup podacima društvenih mreža brži je i jednostavniji nego ispitivanje javnog mnijenja. Nadalje, pristup podacima društvenih mreža moguć je u stvarnom vremenu. Dobivanje jednako učinkovitih prediktivnih modela uz brži i jednostavniji pristup podacima svakako daje značajan znanstveni i stručni doprinos istraživanjima u ovom području.

3.2. Algoritmi strojnog učenja

U ovom poglavlju će se ukratko opisati četiri algoritma strojnog učenja koji će se koristiti u istraživanju:

(i) Strojno učenje temeljeno na sličnosti, a kao primjer ove grupe algoritama koristit će se algoritam k-najbližih susjeda,

(ii) Strojno učenje temeljeno na pogrešci, a kao primjer ove grupe algoritama koristit će se algoritam umjetne neuronske mreže,

(iii) Strojno učenje temeljeno na informaciji a kao primjer ove grupe algoritama koristit će se algoritam stabla odlučivanja,

(iv) Strojno učenje temeljeno na vjerojatnosti, a kao primjer ove grupe algoritama koristit će se algoritam naivni Bayesov klasifikator.

3.2.1. Algoritam k-najbližih susjeda

Strojno učenje temeljeno na sličnosti bazira se na ideji da instance sličnih karakteristika imaju slične izlaze. Sličnost se mjeri u okviru mjera udaljenosti. Jedan od najpoznatijih i najčešće korištenih algoritama je algoritam k-najbližih susjeda (engl k-nearest neigbours, k-NN). Algoritam k-NN koristi se za predviđanje kategorijske ili kontinuirane izlazne varijable za nove instance na temelju ishoda sličnih instanci (otud mu i naziv najbliži susjed). Ovaj algoritam pripada grupi algoritama tzv. lijenog učenika. Lijeni učenici jednostavno pohranjuju podatke za treniranje i čekaju dok se ne pojave podaci za testiranje. Klasifikacija se provodi na temelju najpovezanijih podataka u pohranjenim podacima za treniranje. U usporedbi s vrijednim učenikom, lijeni učenici trebaju manje vremena za treniranje, ali više vremena za predviđanje. Osim algoritma k-najbližih susjeda, drugi primjer lijenog učenika je zaključivanje temeljeno na slučajevima.

Temeljni pojmovi važni za izgradnju prediktivnog modela temeljenog na ideji strojnog učenja baziranoj na sličnosti su obilježja prostora i mjere sličnosti. Ovim se pojmovima omogućuje razumijevanje standardnog pristupa izgradnji modela temeljenih na sličnosti: algoritma najbližeg susjeda. Stoga je za ovaj algoritam posebno važno korištenje normalizacije podataka. Ove su tehnike općenito primjenjive na sve algoritme strojnog učenja, ali su osobito važne ako se koriste pristupi temeljeni na sličnosti.

Algoritam k-NN temelji se na određivanju blizine ili daljine, tj. sličnosti instance koju klasificiramo s poznatim instancama iz skupa za treniranje. Mjere blizine dviju instanci odnose se na mjere blizine njihovih atributa. Za kvantificiranje blizine dvije instance postoji niz pristupa kao što je računanje udaljenosti, računanje povezanosti, Jaccard i kosinusna sličnost (Chomboon et.al., 2015).

Vrijednost *k* u nazivu algoritma odnosi se na to koliko susjeda, koji okružuju instancu za koju radimo predviđanje, će se uzeti u obzir. Ne postoji formula koja bi točno rekla koliko susjeda treba koristiti. Mali broj najbližih susjeda može biti vrlo učinkovit u postizanju malih, suptilnih pomaka u podatkovnom prostoru, ali mogu biti i vrlo osjetljivi na šum u podacima. Stoga se često broj najbližih susjeda odabire kroz iterativnu pretragu, počevši od malog broja najbližih susjeda, poput 1, i nastavljajući povećavati broj najbližih susjeda dok se stopa pogreške ne počinje povećavati. Broj *k* najčešće se uzima u intervalu od 1 do 20.

3.2.2. Algoritam neuronske mreže

Algoritam umjetne neuronske mreže (engl. artificial neural network) temelji se na ideji rada biološke neuronske mreže. Neuronska mreža je masivno paralelni distribuirani procesor koji je dobar za pamćenje iskustvenog znanja. Slična je mozgu u dva aspekta:

(i) znanje se stječe kroz proces učenja,
(ii) međusobne veze između neurona se koriste za spremanje znanja.

Neuronske mreže u području računarstva započinju se intenzivno istraživati posljednjih desetak godina zbog sve bržeg računalnog sklopovlja u vidu računanja matematičkih operacija. Živčani sustav inteligentnijih živih bića (mozak ljudi i životinja) sastoji se od stanica koji se zovu neuroni. Neuron ili živčana stanica predstavlja osnovnu jedinicu živčanog sustava i sastoji se od 3 glavna elementa:

(i) Tijelo s jezgrom
(ii) Kratki ogranci
    - Dendriti
(iii) Dugi ogranak
    - Akson

Neuron prima signal kroz dendrite, obrađuje informacije u jezgri te ih šalje u sljedeći neuron pomoću aksona. Na sličan način funkcionira i umjetni neuron. Dendrit predstavlja ulaz u neuron do kojeg dolaze izlazi iz prethodnog sloja nakon čega se računa suma ulaza nad kojom se poziva aktivacijska funkcija čija se vrijednost potom prosljeđuje dalje na sljedeći sloj. Za uspješan rad i konstruiranje neuronske mreže od velikog su značaja aktivacijske funkcije. Svrha aktivacijske funkcije u umjetnoj neuronskoj mreži je postizanje nelinearnosti. To bi značilo da u slučaju kada bi postojao samo linearan izlaz iz svakog sloja neuronske mreže, sve slojeve bi se moglo zamijeniti jednim slojem zbog nedostatka nelinearnosti. Bez nelinearnosti nije moguće napraviti kompleksniji model klasifikacije, te bi se neuronska mreža mogla svesti na logističku regresiju. Postoji niz aktivacijskih funkcija koje se mogu primijeniti. Najčešće korištena aktivacijska funkcija je tangens

hiperbolni (Haykin, 1998; Rumelhart i McClelland, 1986). Dodatno, postoje i sljedeće aktivacijske funkcije:

(i) step funkcija
    - Nije dobra u praksi
    - Gotovo nemoguće izgraditi klasifikator koji radi za više klasa

(ii) Linearna funkcija
    - Postoji bolja funkcija
    - Nelinearnost je svojstvo koje omogućava naprednije zakonitosti

(iii) Sigmoidalna funkcija
    - Slična step funkciji, ali je glatka
    - Svojstvo nelinearnosti
        - Kompozicije su nelinearne
    - Izlaz uvijek između 0 i 1

(iv) ReLu funkcija
    - Jako korisna u praksi zbog svoje nelinearnosti i gradijenta koji je uvijek 1 ili 0

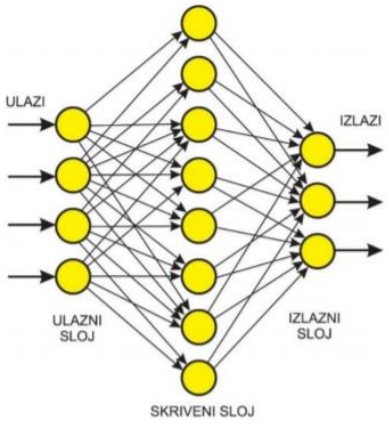
Umjetni neuron može se shematski prikazati na sljedeći način: $\sum = \langle w_j, x \rangle + b_j$. Na tu sumu se primjenjuje aktivacijska funkcija te se izračunata vrijednost prosljeđuje do sljedećeg neurona ukoliko to nije trenutna tražena vrijednost predviđanja.

Neuronska mreža sastoji se od više neurona koji su raspoređeni u više slojeva. Postoje tri vrste slojeva:

(i) Ulazni sloj:
    - Prvi sloj mreže
    - Obrađuje ulazne podatke iz sustava

(ii) Izlazni sloj:
    - Posljednji sloj mreže
    - Izlazi neurona predstavljaju rješenja sustava

(iii) Skriveni sloj:
    - Svi ostali slojevi su skriveni
    - Izvode nelinearne transformacije ulaza



Slika 2. Struktura neuronske mreže

Neuronsku mrežu moguće je prikazati kao kompleksni računski graf u kojem je jedinica računanja neuron. Mreže je moguće klasificirati na više različitih modela od kojih je jedna navedena pomoću prethodne slike (Slika 2). Iako ima više mogućnosti, svi rade na sličnom principu. Svi imaju ulaze iz kojih kroz slojeve trebaju kreirati smisleni izlaz koji su naučile postupkom učenja. Slojevi se sastoje od neurona koji su simulirani pomoću ranije navedenih aktivacijskih funkcija. Neuroni su međusbono povezani, a veze između neurona imaju određenu težinu (oznaka w, a dolazi od engl. *weight*) koja se prosljeđuje tj. propagira na sljedeći sloj kroz aktivacijsku funkciju ($\sum = \langle w_j, x \rangle + b_j$) i na taj način se dobivaju izlazi. Neuronska mreža će se najčešće sastojati od nekoliko nelinearnih skrivenih slojeva, a izlazni sloj će biti različit s obzirom na to što se od neuronske mreže očekuje da radi.

Učenje neuronske mreže izvodi se pomoću algoritma koji se zove „širenje unatrag“ (engl. back propagation). Za učenje neuronske mreže potrebno je odlučiti koji će se optimizacijski algoritam koristiti za učenje. Prilikom učenja neuronske mreže prvo je potrebno definirati ono što se uči, a to su ulazni podaci, i koji se podaci očekuju kao izlazni. Za to je potrebno poznavati dio matematike koja se bavi statističkim pogreškama. Funkcija koja računa pogrešku naziva se funkcija gubitka (greška između željenih rezultata i dobivenih rezultata). Nakon izračuna greške, računa se gradijent kako bi se znalo u kojem smjeru treba mijenjati vrijednosti i propagirati ih po neuronskoj mreži pomoću algoritma propagacije unazad.

3.2.3. Algoritam stabla odlučivanja

Algoritmom stabla odlučivanja (engl. decision tree algorithm) dobiva se grafički prikaz modela utjecaja ulaznih varijabli na izlaznu. Varijabla može biti izražena u obliku kategorija ili kao numerička vrijednost. Ovisno o tipu varijable, razlikujemo klasifikacijska i regresijska stabla odlučivanja (Loh, 2011). Svaki čvor u grafičkom stablu predstavlja jednu ulaznu varijablu na čijim su rubovima označena „djeca-čvorovi" za svaku moguću vrijednost neke ulazne varijable. Svaki list u stablu predstavlja vrijednost ciljne (izlazne) varijable ako su dane vrijednosti ulaznih varijabli predstavljene putem od korijena stabla do tog lista. Stablo se dobiva „učenjem" na podacima, na način da se vrši grananje izvornog skupa podataka u podskupove na temelju testiranja vrijednosti varijabli. Grananje se vrši temeljem izračuna relevantnosti ulaznih varijabli temljem neke od metrika (Loh, 2011). Proces se ponavlja na svakom izvedenom podskupu na rekurzivni način. Rekurzija je završena kada podskup određenog čvora ima sve iste vrijednosti izlazne varijable ili kada daljnje grananje više ne pridonosi poboljšanju rezultata. Prednosti algoritma stabla odlučivanja su: jednostavan za razumijevanje i objašnjavanje, traži jednostavnu pripremu podataka, može raditi s numeričkim i kategorijskim vrijednostima, brz je i koristi model bijele kutije.

3.2.4. Algoritam naivni Bayesov klasifikator

Naivni Bayesov klasifikator (engl. Naive Bayes classifier) je algoritam koji pripada grupi algoritama strojnog učenja koji se temelje na vjerojatnosti. Temelji se na primjeni Bayesova teorema uvjetne vjerojatnosti, i otud pojam Bayesov klasifikator. Drugi ključni pojam u nazivu algoritma, naivni, sugerira naivni pristup rješenju problema izračunavanja velikog broja uvjetnih vjerojatnosti. Da bi se to izbjeglo algoritam pretpostavlja neovisnost ulaznih varijabli i kreće s pretpostavkom da ulazne varijable nisu povezane jedna s drugom. Uz pretpostavku neovisnosti, ne računaju se uvjetne vjerojatnosti koje uključuju kombinacije ulaza. Ostaju samo uvjetne vjerojatnosti koje pokazuju odnos izlazne varijable i ulazne varijable, što predstavlja ogromno pojednostavljenje. Ova pretpostavka nije realistična, ali značajno smanjuje kompleksnost izračuna. Naivni Bayesov klasifikator je algoritam koji uvelike pojednostavljuje učenje pretpostavljajući da

su varijable neovisne s obzirom na zavisnu varijablu. Iako je neovisnost općenito loša pretpostavka, u praksi je naivni Bayes klasifikator izrazito konkurentan i sa sofisticiranijim klasifikatorima (Rish, 2001).

3.2.5. Optimizacija hiperparametara algoritama strojnog učenja

Odabir najbolje konfiguracije hiperparametara za modele strojnog učenja ima izravan utjecaj na izvedbu modela. Taj proces zahtijeva duboko poznavanje algoritama strojnog učenja kao i poznavanje odgovarajućih tehnika optimizacije hiperparametara.
Izgradnja efikasnog modela strojnog učenja je složen i dugotrajan proces koji uključuje najprije odabir odgovarajućeg algoritma, a potom i dobivanje optimalne arhitekture modela pomoću podešavanja njegovih hiperparametara (Shawi et al. 2019). Postoje različiti pristupi optimizaciji, te imaju različite prednosti i nedostatke kad se primjene na različite vrste modela.
U modelima strojnog učenja postoje dvije vrste parametara: jedni koji se mogu inicijalizirati i ažurirati kroz proces učenja podataka (npr. težine neurona u neuronskim mrežama), to su parametri modela. Drugi, nazvani hiperparametri, ne mogu se izravno procijeniti iz učenja podataka već se moraju postaviti prije treniranja modela strojnog učenja jer oni definiraju arhitekturu modela strojnog učenja (Kuhn & Johnson, 2013). Hiperparametri su parametri koji se koriste za konfiguriranje modela strojnog učenja (npr. stopa učenja za treniranje neuronske mreže) ili za specificiranje algoritma koji se koristi za minimiziranje funkcije gubitka (npr. aktivacijska funkcija u neuronskim mrežama) (Diaz et al., 2017).

Da bi se stvorio optimalan model, mora se istražiti niz mogućnosti. Postupak dizajniranja idealne arhitekture modela s optimalnom konfiguracijom hiperparametara naziva se podešavanje hiperparametara. Podešavanje hiperparametara se smatra ključnom komponentom izgradnje učinkovitog modela strojnog učenja, posebno za modele zasnovane na stablu i neuronske mreže, koji imaju mnogo hiperparametara (Hutter et al., 2019). Postupak podešavanja hiperparametara je različit između različitih algoritama strojnog učenja zbog njihovih različitih tipova hiperparametara, uključujući kategorijske, diskretne i kontinuirane hiperparametre (Decastro-García et al., 2019). Ručno testiranje je tradicionalan način podešavanja hiperparametara i još

uvijek prevladava u nekim istraživanjima, iako zahtijeva duboko razumijevanje korištenih algoritama strojnog učenja i njihove postavke vrijednosti hiperparametara (Abreu, 2019).

Očekuje se da će se optimalna arhitektura modela strojnog učenja dobiti nakon postupka optimizacije hiperparametara. U nastavku se navodi nekoliko snažnih argumenata za primjenu optimizacije hiperparametara na modele strojnog učenja (Hutter et al., 2019):

1. Smanjuje se potrebni ljudski napor jer se znatno vrijeme utroši za podešavanje hiperparametara, posebno za velike skupove podataka ili složene algoritme s velikim brojem hiperparametara.

2. Poboljšava performanse modela. Mnogi hiperparametri algoritama strojnog učenja imaju različite optimume za postizanje najboljih performansi u različitim skupovima podataka.

3. Čini modele i istraživanje ponovljivima. Tek kad je ista razina postupka podešavanja hiperparametara primijenjena može se dobro i pošteno usporediti modele; dakle, koristeći isti postupak optimizacije hiperparametara pomaže odrediti najprikladniji model za specifični problem.

U nastavku se opisuju parametri za četiri algoritma strojnog učenja koji se koriste u ovom istraživanju.

3.2.5.1. k-najbližih susjeda

U k-NN algoritmu je broj najbližih susjeda, *k,* presudni hiperparametar (Zuo et al., 2007). Ako je k premalen, model će biti podtreniran, ako je k prevelik, model će biti pretreniran i zahtijevat će više vremena. Uz to, ponderirana funkcija korištena u predviđanju također može biti odabrana iz "uniformne" (svi se ponderiraju jednako) ili "udaljenosti" (točke su ponderirane inverzom njihove udaljenosti), ovisno o specifičnosti problema koji se modelira. Metrika udaljenosti i parametar snage Minkowski metrike mogu također biti podešeni jer mogu rezultirati manjim poboljšanjima.

3.2.5.1. Naivni Bayesov klasifikator (NB)

Četiri glavne vrste NB modela su: Bernoullijev NB, Gaussov NB, multinomni NB i komplement NB (Sulzmann et al., 2007). Uobičajeno ne postoji nijedan hiperparametar koji treba biti podešen

za Gaussov NB. Izvedba Gaussova NB modela uglavnom ovisi o tome koliko dobro skup podataka slijedi Gaussovu raspodjelu.

3.2.5.2. Neuronske mreže

U usporedbi s drugim modelima strojnog učenja, neuronske mreže imaju više hiperparametara koji zahtijevaju podešavanje.

Prvi skup hiperparametara povezan je s razvojem modela neuronske mreže. Dakle, radi se o hiperparametrima dizajna modela. Budući da sve neuronske mreže imaju ulazni i izlazni sloj, složenost modela uglavnom ovisi o broju skrivenih slojeva i broju neurona u svakom sloju, što su dva glavna hiperparametra.

U izradi modela neuronske mreže ova su dva hiperparametra postavljena i prilagođena ovisno o složenosti skupa podataka na kojem se gradi model. Model neuronske mreže treba biti dovoljno složen za modeliranje objektivnih funkcija (ili zadataka predviđanja), dok s druge strane ne smije biti presložen da bi izbjegao pretreniranost.

U sljedećoj fazi potrebno je postaviti određene vrste funkcija. Prva funkcija koju treba konfigurirati je tip funkcije gubitka, koji se odabire uglavnom na temelju vrste problema (npr. binarna unakrsna entropija za binarnu klasifikaciju, višeklasna unakrsna entropija za višestruku klasifikaciju, i RMSE za probleme regresije) (Koutsoukas et al., 2017). Još jedan važan hiperparametar je tip aktivacijske funkcije koji se koristi za modeliranje nelinearnih funkcija. Neke od najpoznatijih aktivacijskih funkcija su tangens hiperbolni i sigmoida.

Na kraju, tip optimizatora može biti: stohastički gradijentni spust (engl. stochastic gradient descent, SGD), adaptivni moment procjene (engl. aadaptive moment estimation, ADAM), propagacija prosječne kvadratne pogreške (engl. root mean square propagation, RMSprop) itd. (Domhan et al., n.d.).

S druge strane, neki drugi hiperparametri povezani su s optimizacijom i procesom treniranja modela neuronske mreže, pa se nazivaju hiperparametri optimizacije. Stopa učenja jedan je od najvažnijih hiperparametara u modelima neuronske mreže (Ozaki et al., 2017). Određuje veličinu koraka u svakoj iteraciji, što omogućuje konvergiranje ciljne funkcije. Velika stopa učenja ubrzava proces učenja, ali gradijent može oscilirati oko lokalnog minimuma ili čak ne može konvergirati. S druge strane, mala stopa učenja omogućuje približavanje, ali uvelike će povećati vrijeme

treniranja modela. Odgovarajuća stopa učenja trebala bi omogućiti ciljnoj funkciji da može konvergirati na globalni minimum u razumnoj količini vremena.
Drugi uobičajeni hiperparametar je stopa napuštanja. Napuštanje je standardna metoda regularizacije modela predložena za smanjenje pretreniranosti. U slučaju napuštanja, dio neurona se nasumce uklanja, a postotak neurona koji se uklanjaju treba prilagoditi.

Veličina mini-serije i broj epoha preostala su dva hiperparametra koji predstavljaju broj obrađenih uzoraka prije ažuriranja modela i broj kompletnih prolaza kroz čitav skup za treniranje (Soon et al., 2018). Veličina mini-serije određuje se prema potrebama resursa procesa treniranja i broju iteracija. Broj epoha ovisi o veličini skupa za treniranje i treba ga prilagoditi polako povećavajući svoju vrijednost dok se točnost validacije ne počne smanjivati, što ukazuje na pretreniranost.

S druge strane, modeli se često konvergiraju unutar nekoliko epoha, a sljedeće epohe mogu dovesti do nepotrebnog dodatnog trošenja vremena i pretreniranosti, što se može izbjeći ranim zaustavljanjem. Rano zaustavljanje oblik je regularizacije kojim treniranje modela staje kad se točnost validacije ne poveća nakon određenog broj uzastopnih epoha. Broj epoha čekanja također se može prilagoditi kako bi se smanjilo vrijeme treniranja modela.

3.2.5.3. Stablo odlučivanja

Stablo odlučivanja uobičajena je metoda klasifikacije koja koristi strukturu stabla za modeliranje odluka i mogućih posljedica sažimanjem skup pravila iz podataka (Safavian & Landgrebe, 1991). Stablo odlučivanja ima tri glavne komponente: korijenski čvor koji predstavlja cjelokupne podatke; više čvorova odluke koji ukazuju na testove odluke i pod-čvorove koji se dijele na svaki atribut te nekoliko čvorova listova koji predstavljaju vrijednost klase (Manias et al., 2020). Algoritam stabla rekurzivno dijeli skup podataka za treniranje s boljim vrijednostima atributa za postizanje dobrih odluka o svakom podskupu.

Obrezivanje (engl. prunning), koje podrazumijeva uklanjanje nekih pod-čvorova čvorova odluke, koristi se kod stabla odlučivanja kako bi se izbjegla pretreniranost modela. Budući da dublje stablo ima više podstabala za donošenje preciznijih odluka, maksimalna dubina stabla, 'maksimalna dubina', je bitni hiperparametar koji kontrolira složenost algoritama (Yang et al., 2019).

Postoje mnogi drugi važni hiperparametri koji se trebaju prilagoditi za izgradnju učinkovitih modela stabla [56]. Prvo, kvaliteta razdvajanja može se izmjeriti postavljanjem mjerne funkcije. Gini indeks i informacijska dobit neke su od mogućih mjernih funkcija. Metoda grananja „splitter“ može se odabrati za izbor najboljeg grananja ili „random“ da se nasumično radi grananje.

Nadalje, postoji nekoliko diskretnih hiperparametara povezanih s postupkom grananja: minimalan broj instanci za razdvajanje čvora odluke ili za dobivanje čvora lista. (Pedregosa et al., 2011) (Sanders & Giraud-Carrier, 2017).

3.3. Opis podataka

Ovo se istraživanje temelji na podacima o aktivnosti na Facebooku političkih kandidata u 41 francuskom gradu: Paris, Marseille, Lyon, Toulouse, Nice, Nantes, Strasbourg, Montpellier, Bordeaux, Lille, Rennes, Reims, Le Havre, Saint-Étienne, Toulon, Grenoble, Dijon, Angers, Le Mans, Nîmes, Aix-en Provence, Brest, Saint-Denis, Clermont-Ferrand, Limoges, Villeurbanne, Tours, Amiens, Metz, Besançon, Perpignan, Orléans, Mulhouse, Boulogne-Billancourt, Caen, Rouen, Nancy, Argenteuil, Montreuil, Saint-Paul, Saint-Denis. Odabrani su najveći francuski gradovi, odnosno svi oni koji imaju više od 100 000 stanovnika. Dodatni uvjet uključivanja grada bio je da barem dva kandidata u tom gradu imaju otvorene stranice/profile. Podaci su kreirani od strane autora rada pojedinačnim preuzimanjem podataka sa stranica kandidata na izborima, u lipnju 2020. godine. Ime i prezime kandidata u gradovima preuzeti su sa službenih stranica izbornog povjerenstva. Svaki kandidat je pretraživan na društvenoj mreži Facebook s ciljem pronalaska stranica ili profila. Kod kandidata koji su imali otvorene stranice ili profile, pregledavan je svaki sadržaj i kategoriziran u jednu od sljedećih kategorija: događaj, fotografija, poveznica, video i status. Za svaki sadržaj ekstrahiran je podatak o broju sadržaja na stranici/profilu, broju *likeova* sadržaja, broju komentara sadržaja te broju dijeljenja sadržaja.

Tablica u nastavku daje nazive i opise varijabli koje su se koristile u istraživanju.

Tablica 1. Opis varijabli

Naziv varijable	**Opis varijable i vrijednosti**
Grad	Naziv grada u kojem je političar kandidat: svi gradovi u Francuskoj koji imaju više od 100 000 stanovnika. U analizu uključeno 38 gradova kod kojih su bar dva kandidata imala otvorene stranice/profile.
Ukupan broj *likeova* stranice	Ukupan broj *likeova* stranice
Broj događaja	Ukupan broj događaja podijeljenih tijekom kampanje na stranici kandidata.
Broj fotografija	Ukupan broj fotografija podijeljenih tijekom kampanje na stranici kandidata.
Broj poveznica	Ukupan broj poveznica podijeljenih tijekom kampanje na stranici kandidata.
Broj videa	Ukupan broj videa podijeljenih tijekom kampanje na stranici kandidata.
Broj statusa	Ukupan broj statusa podijeljenih tijekom kampanje na stranici kandidata.
Broj *likeova* događaja	Prosječan broj *likeova* jednog događaja.
Broj *likeova* fotografija	Prosječan broj *likeova* jedne fotografije.
Broj *likeova* poveznica	Prosječan broj *likeova* jedne poveznice.
Broj *likeova* videa	Prosječan broj *likeova* jednog videa.
Broj *likeova* statusa	Prosječan broj *likeova* jednog statusa.
Broj komentara događaja	Prosječan broj komentara jednog događaja.

Broj komentara fotografija	Prosječan broj komentara jedne fotografije.
Broj komentara poveznica	Prosječan broj komentara jedne poveznice.
Broj komentara videa	Prosječan broj komentara jednog videa.
Broj komentara statusa	Prosječan broj komentara jednog statusa.
Broj dijeljenja događaja	Prosječan broj dijeljenja jednog događaja.
Broj dijeljenja fotografija	Prosječan broj dijeljenja jedne fotografije.
Broj dijeljenja poveznica	Prosječan broj dijeljenja jedne poveznice.
Broj dijeljenja videa	Prosječan broj dijeljenja jednog videa.
Broj dijeljenja statusa	Prosječan broj dijeljenja jednog statusa.
Stranka	Pripadnost kandidata političkoj stranci: LVEC, LUG, LUD, LUC, LSOC, LRN, LREM, LREG, LRDG, LLR, LFI, LEXG, LEXD, LECO, LDVG, LDVD, LDVC, LDIV.
Spol	Spol kandidata; muško i žensko.
Rezultat na izborima	Postotak osvojenih glasova kandidata na izborima

Inicijalno je u istraživanje uključeno 25 varijabli: 24 ulazne varijable, a odnose se na: (i) aktivnost kandidata i reakcije pratitelja stranice na aktivnost kandidata, (ii) dvije varijable koje se odnose na spol kandidata i pripadnost kandidata političkoj stranci. Jedna je izlazna varijabla za prediktivne modele, rezultat na izborima mjeren postotkom glasova kandidata na izborima. Nakon koraka pripreme podataka i aktivnosti selekcije atributa, odabrane su samo relevantne značajke koji predstavljaju ulazne parametre u modele.

U sljedećem poglavlju daju se rezultati istraživanja koristeći ovdje opisane podatke.

4. REZULTATI ISTRAŽIVANJA

Rezultati istraživanja prezentiraju se unutar nekoliko sekcija. Najprije će se napraviti razumijevanje podataka kroz prikaz distribucija vrijednosti varijabli i ispitivanje međusobne povezanosti varijabli mjerene korelacijom. Najvažniji rezultati istraživanja bit će prikazani kroz drugu sekciju ovog poglavlja u kojoj će se detaljno opisati prediktivni modeli dobiveni algoritmima strojnog učenja i njihovi parametri kvalitete. Na kraju se provodi testiranje statističke značajnosti razlika među rezultatima s ciljem dobivanja odgovora na postavljene hipoteze istraživanja.

4.1. Razumijevanje podataka

U koraku razumijevanja podataka opisuju se varijable kroz deskriptivnu statistiku. U tablici 2 prikazane su minimalne, maksimalne i srednje vrijednosti (aritmetička sredina i medijan) za sve varijable koje se odnose na aktivnost kandidata na društvenoj mreži Facebook.

Tablica 2. Deskriptivna statistika varijabli

Naziv varijable	Opis vrijednosti
Ukupan broj *likeova* stranice	Minimum: 1 Maksimum: 114 740 Aritmetička sredina: 11 551 Medijan: 5026
Broj događaja	Minimum: 0 Maksimum: 27 Aritmetička sredina: 2.7 Medijan: 2
Broj fotografija	Minimum: 0 Maksimum: 136 Aritmetička sredina: 17

	Medijan:10
Broj poveznica	Minimum: 0 Maksimum: 24 Aritmetička sredina: 4.4 Medijan: 3
Broj videa	Minimum: 0 Maksimum: 70 Aritmetička sredina: 12 Medijan: 8
Broj statusa	Minimum: 0 Maksimum: 7 Aritmetička sredina: 1 Medijan: 1
Broj *likeova* događaja	Minimum: 0 Maksimum: 530 Aritmetička sredina: 47 Medijan: 2
Broj *likeova* fotografija	Minimum: 0 Maksimum: 667.5 Aritmetička sredina: 109.5 Medijan: 48.88
Broj *likeova* poveznica	Minimum: 0 Maksimum: 573

	Aritmetička sredina: 70 Medijan: 56.7
Broj *likeova* videa	Minimum: 0 Maksimum: 574.5 Aritmetička sredina: 103.4 Medijan: 56.7
Broj *likeova* statusa	Minimum: 0 Maksimum: 88 Aritmetička sredina: 8.67 Medijan: 0
Broj komentara događaja	Minimum: 0 Maksimum: 90.8 Aritmetička sredina: 5.65 Medijan: 0
Broj komentara fotografija	Minimum: 0 Maksimum: 215.86 Aritmetička sredina: 17.5 Medijan: 3.1
Broj komentara poveznica	Minimum: 0 Maksimum: 144 Aritmetička sredina: 12.85 Medijan: 2
Broj komentara videa	Minimum: 0

	Maksimum: 140 Aritmetička sredina: 17.9 Medijan: 4.7
Broj komentara statusa	Minimum: 0 Maksimum: 13.6 Aritmetička sredina: 1 Medijan:0
Broj dijeljenja događaja	Minimum: 0 Maksimum: 154 Aritmetička sredina: 8.5 Medijan: 0
Broj dijeljenja fotografija	Minimum: 0 Maksimum: 120.5 Aritmetička sredina: 18.1 Medijan:7.7
Broj dijeljenja poveznica	Minimum: 0 Maksimum: 308.2 Aritmetička sredina: 26.7 Medijan: 7
Broj dijeljenja videa	Minimum: 0 Maksimum: 232 Aritmetička sredina: 30.2 Medijan: 8.5

Broj dijeljenja statusa	Minimum: 0 Maksimum: 32 Aritmetička sredina: 2 Medijan: 0
Rezultat na izborima	Minimum: 0.68 Maksimum: 66.32 Aritmetička sredina: 15.65

Većina varijabli ima eksponencijalnu distribuciju vrijednosti koju karakterizira velika vjerojatnost pojave manjih vrijednosti, a mala vjerojatnost pojave velikih vrijednosti. Karakteristika je takvih distribucija da su vrijednosti aritmetičke sredine veće od vrijednosti medijana.

U koraku razumijevanja podataka ispitane su i povezanosti varijabli koje se odnose na aktivnost na društvenoj mreži Facebook. Za to je primijenjena korelacijska analiza. Rezultati su dani u tablici u dodatku rada, a sortirani po vrijednosti *r*, koeficijenta korelacije, od najveće vrijednosti prema najmanjoj. Uz vrijednost koeficijenta korelacija dana je i vrijednost *p* koja sugerira statističku značajnost rezultata. Interpretirati se mogu samo statistički značajni koeficijenti korelacije. Koeficijenti korelacije po apsolutnoj vrijednosti veći od 0.7 smatraju se jako velikom povezanosti, a koeficijenti korelacije po apsolutnoj vrijednosti veći od 0.4 velikom povezanosti (Petz, 1997.). Korelacijskom analizom je utvrđen niz linearnih povezanosti između varijabli. Značajnu i visoku povezanost (vrijednost koeficijenta korelacije *r* iznad 0.7) imaju varijable koje se odnose na dijeljenje, lajkanje i komentiranje određene vrste sadržaja. Npr. najveća linearna povezanost utvrđena je između varijabli broj dijeljenja događaja i broj *likeova* događaja. Povezanost je pozitivna, što ukazuje da s porastom *likeova* događaja raste i broj dijeljenja događaja, i obratno. Druga najveća povezanost je između varijabli broj *likeova* videa i broj *likeova* fotografija. Ako promatramo varijablu Rezultat, koja će predstavljati zavisnu varijablu u modeliranju, najveću linearnu povezanost ima s varijablom Ukupan broj *likeova* stranice. Zanimljivo je istaknuti izrazito mali broj negativnih koeficijenata korelacije, odnosno potpuni izostanak velike negativne povezanosti.

4.2. Priprema podataka

Priprema podataka oduzima barem 60% ukupnog vremena procesa otkrivanja znanja u podacima (Cabena i suradnici, 1998). U sklopu koraka pripreme podataka provedeno je niz aktivnosti. Inicijalno, istraživanje uključuje kandidate na izborima u svim gradovima u Francuskoj s više od 100 000 stanovnika. Pri tome su isključeni kanidati koji nemaju otvorene stranice i/ili profile. Nadalje, isključeni su gradovi koji nemaju više od dva kandidata s otvorenim stranicama i/ili profilima. Prva aktivnost čišćenja podataka bila je identifikacija stršila. Stršila su ekstremne vrijednosti koje idu van trenda preostalih podataka i nalaze se daleko od srednjih vrijednosti atributa. Identificirane su jer su neki od primijenjenih algoritama strojnog učenja (umjetne neuronske mreže, k-najbližih susjeda) jako osjetljive na ekstremne vrijednosti. Interkvatrili su korišteni za identifikaciju stršila na sljedeći način. Vrijednost je označena kao stršilo ako se nalazi najmanje 1.5 interkvartila ispod prvog kvartila ili 1.5 interkvartila iznad trećeg kvartila. Ovim pristupom identificirana su stršila koja su isključena iz daljnje analize podataka. Micanjem stršila nastale su nedostajuće vrijednosti. Postupkom umetanja podataka umjesto vrijednosti koje nedostaju provedena je imputacija. Time su pridjeljivane vrijednosti na prazna mjesta temeljem sljedeće heuristike: nedostajuće vrijednosti zamijenjene su srednjim vrijednostima varijabli.

Sljedeći korak unutar pripreme podataka bio je normalizacija podataka. Provedena je normalizacija s dva aspekta. Prvo su podaci normalizirani s obzirom na broj birača u pojedinom gradu. Originalna vrijednost svake varijable je podijeljena s brojem birača u gradu na koji se odnosi. U drugom koraku provedena je *min-max* normalizacija kojom su vrijednosti svih varijabli svedene na skalu od 0 do 1. Modeliranje je provedeno nad pripremljenim, normaliziranim podacima. Treća aktivnost provedena u pripremi podataka je selekcija atributa. Selekcija atributa je proces odabira važnih varijabli (značajki, atributa) koje će biti korisne za proces klasifikacije i/ili regresije. Selekcija atributa smanjuje dimenzionalnost podataka, a time smanjuje i vrijeme potrebno za modeliranje. Odabir relevantnih varijabli poboljšava točnost prediktivnih modela, dok uključivanje manje važnih varijabli smanjuje točnost. Uklanjanje suvišnih i irelevantnih varijabli štedi resurse, a istovremeno poboljšava model (Guyon i Elisseeff, 2003., Arauzo – Azofra, Aznarte i Benitez, 2011., Cadenas, Garrido, i Martinez, 2013.). Nekoliko je pristupa selekciji atributa. U ovom se istraživanju koristi filter pristup koji radi rangiranje varijabli i odabire određeni podskup varijabli

temeljem mjere vrednovanja. Kao mjera vrednovanja, koristit će se ReliefF algoritam. Ovaj algoritam je odabran iz razloga jer je računalno učinkovit, a ipak osjetljiv na složene obrasce povezivanja, tj. interakcije, tako da informativne varijable nisu greškom eliminirane (Urbanowicz i suradnici, 2018). Osim toga, Relief algoritam ima karakteristiku da se fleksibilno prilagođava različitim karakteristikama podataka (Urbanowicz i suradnici, 2018). U sljedećoj tablici prikazani su rezultati provedene selekcije atributa primjenom ReliefF algoritma.

Tablica 3. Rezultati selekcije atributa

Varijabla	Mjera vrednovanja
Ukupan broj *likeova* stranice	0.029702
Broj događaja	0.026505
Broj *likeova* statusa	0.011427
Broj fotografija	0.008699
Broj poveznica	0.004903
Broj *likeova* fotografija	0.002092
Broj *likeova* poveznica	0.001551
Broj statusa	0.001498
Broj *likeova* događaja	-0.00054
Broj komentara događaja	-0.001936
Broj *likeova* videa	-0.002369
Broj dijeljenja fotografija	-0.003654
Broj komentara statusa	-0.00369
Broj videa	-0.004196
Broj komentara videa	-0.006043
Broj dijeljenja poveznica	-0.006576
Broj dijeljenja statusa	-0.006611

Broj dijeljenja videa	-0.008027
Broj dijeljenja događaja	-0.008855
Broj komentara poveznica	-0.01112
Broj komentara fotografija	-0.011854

Sve varijable koje imaju pozitivnu vrijednost mjere vrednovanja su selektirane kao relevantne za predikciju rezultata te će predstavljati ulazne parametre u prediktivne modele. Sukladno tome, 8 varijabli od početnog skupa od 21 varijable koje se odnose na podatke društvene mreže Facebook su temeljem rezultata selekcije atributa primjenom algoritma ReliefF odabrane te će nam predstavljati ulaz u modeliranje.

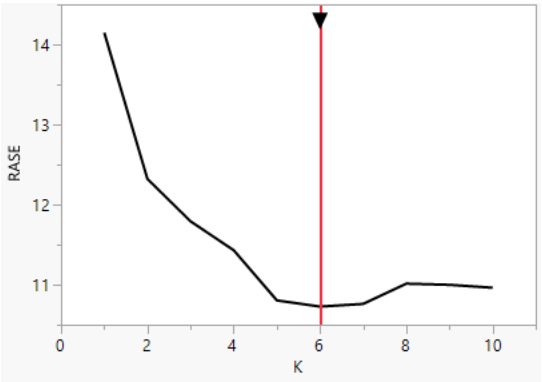
Cilj je pripreme podataka očistiti podatke, generirati značajke, te odabrati atribute za modeliranje. Sljedeća sekcija opisuje rezultate razvoja prediktivnih modela primjenom četiri različita algoritma strojnog učenja.

4.3. Prediktivni modeli

U četvrtom koraku CRISP-DM standarda razvijeni su prediktivni modeli primjenom četiri algoritma strojnog učenja. CRISP-DM je primijenjen kao iterativan proces. Podaci su za potrebe primjene svakog algoritma posebno pripremani (korak pripreme podataka je opisan ranije u ovom poglavlju). Kao jedna od najvažnijih aktivnosti u ovom koraku samog istraživanja, provedena je optimizacija hiperparametara svakog pojedinog algoritma kako bi se spriječila pretreniranost modela, a dobili kvalitetni pouzdani i točni prediktivni modeli. Optimizacija hiperparametara je provedena na skupu relevantnih podataka za treniranje.

4.2.1. Model k-najbližih susjeda

Na slici u nastavku vidljivo je kretanje vrijednosti RASE (Root Average Square Error) ovisno o broju odabranih susjeda. Kako je ranije navedeno, u kNN je broj promatranih najbližih susjeda, *k*, presudni hiperparametar. Ako je *k* premalen, model će biti podtreniran, ako je *k* prevelik, model će biti pretreniran.



Slika 3. Kretanje greške modela ovisno o parametru k

Najmanja mjera greške RASE je za *k=6*, te će se ovaj broj odabrati kao optimalan broj susjeda koji se uzima u obzir. Nakon *k=6* RASE opet kreće rasti. Taj je broj i u nekoliko ranijih istraživanja odabran kao optimalan. Vrijednosti su prikazane i u tablici 4.

Tablica 4. Kretanje parametara ovisno o k

K	RASE
1	14.148
2	12.322
3	11.792
4	11.431
5	10.805
6	**10.728**
7	10.761
8	11.014
9	10.999
10	10.963

Prema grafikonu i tablici, u slučaju *k=6*, postiže se najtočniji model. Stoga model koji promatra 6 susjeda nadalje evaluiramo i interpretiramo te se u nastavku prikazuju rezultati za *k=6*.

Parametri točnosti i pouzdanosti modela prikazani su u tablici u nastavku.

Tablica 5. Pouzdanost modela k-najbližih susjeda

Prediktor	Algoritam	RSquare	RASE
Predviđena Formula Rezultat	K najbližih susjeda	0.5044	8.9899

Tablica 5. daje informaciju i mjeri pouzdanost modela, RSquare, koji nam govori koliki udio varijance zavisne varijable je objašnjen prediktorskim varijablama u modelu. Vrijednost od 0.5044 sugerira da je malo više od polovice varijance objašnjeno s 10 prediktorskih varijabli u prediktivnom modelu.

S ciljem utvrđivanja najznačajnijih prediktora ishoda izbora temeljem modela k-najbližih susjeda, provedena je analiza osjetljivosti. Rezultati analize osjetljivosti govore koliko se promijenila vrijednost izlazne varijable, kad se mijenja ulazna, prediktorska varijabla. Rezultati su dani u tablici u nastavku. Vrijednost glavni utjecaj u stupcu 2 govori koliki je pojedinačni utjecaj varijable na izlaz, a vrijednost ukupni utjecaj uzima u obzir interakcije između ulaznih varijabli. Male su razlike u rangiranju varijabli s obzirom na ova dva parametra. U ovom istraživanju uzimat će se u obzir ukupni utjecaj varijable na izlaz.

Tablica 6. Najznačajniji prediktori rezultata u modelu k-najbližih susjeda

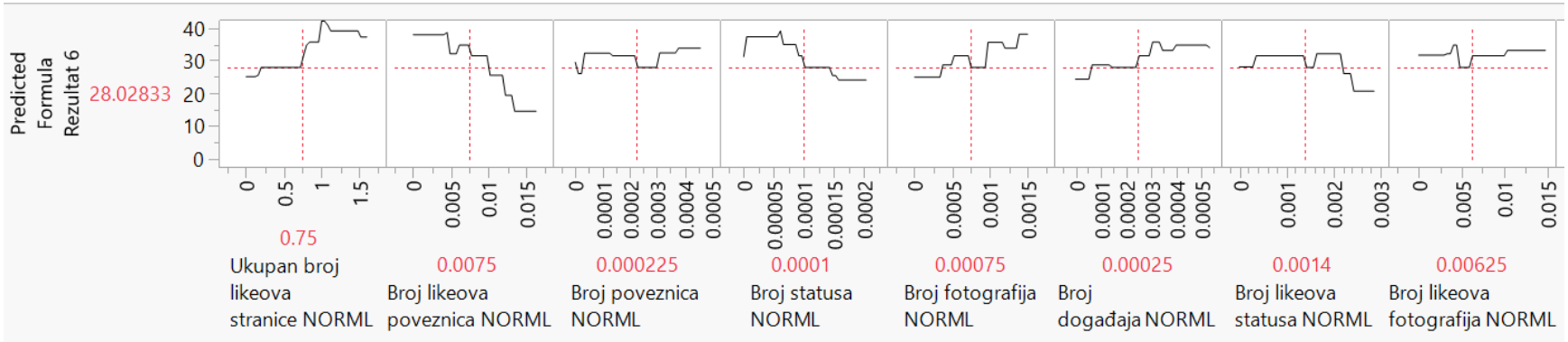
Varijabla	Glavni utjecaj	Ukupni utjecaj
Ukupan broj *likeova* stranice	0.2859237825	0.4400653187
Broj *likeova* poveznica	0.1509932834	0.3313436475
Broj poveznica	0.0927691955	0.2041950425
Broj statusa	0.0513987634	0.2029955118
Broj fotografija	0.0717890466	0.1760975209
Broj događaja	0.058148065	0.1707763427
Broj *likeova* statusa	0.0256769817	0.1278334651
Broj *likeova* fotografija	0.0215813862	0.1018842784
Spol	0.0057650638	0.0158457996

Varijabla	Glavni utjecaj	Ukupni utjecaj
Stranka	0.0059326611	0.0128171448

Ukupan broj *likeova* stranice najznačajniji je prediktor ishoda izbora temeljem modela k-najbližih susjeda. Temeljem broja osoba koje su lajkale stranicu kandidata u najvećoj mjeri se mogu predvidjeti izbori. Sljedeće tri varijable koje imaju najveću prediktorsku snagu odnose se na poveznice, i to redom: broj dijeljenja poveznica, broj *likeova* poveznica i broj poveznica. Model k-najbližih susjeda u velikoj mjeri prepoznaje poveznice, dijeljenje i lajkanje poveznica kao dobar prediktor rezultata kandidata na političkim izborima. I sljedeći po snazi prediktori ishoda se odnose na aspekte interakcije kandidata i potencijalnog birača, a referiraju se na komentiranje sadržaja, i to: broj komentara događaja, broj komentara fotografija i broj komentara poveznica. Na začelju tablice nalazi se politička stranka kojoj kandidat pripada te spol kandidata.

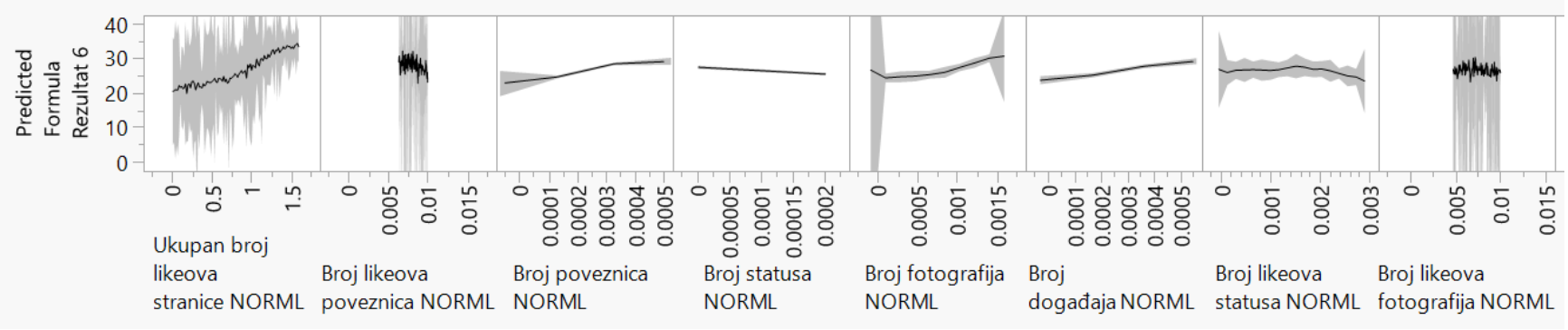
U nastavku analize rezultati analize osjetljivosti su se dodatno koristili za predviđanje rezultata. Sljedeće dvije slike prikazuju grafički kretanje vrijednosti izlazne varijable ovisno o kretanjima vrijednosti ulaznih varijabli. Ulazne varijable poredane su po važnosti tj. razini utjecaja na promjenu izlaza.

Model K-najbližih susjeda predviđa vrijednost izlaza za danu instancu koristeći odgovore instanci u lokalnom susjedstvu tog promatranja. U ovom istraživanju, koristi se za predviđanje kontinuiranog odgovora, postotka glasova koje će kandidat dobiti na izborima.

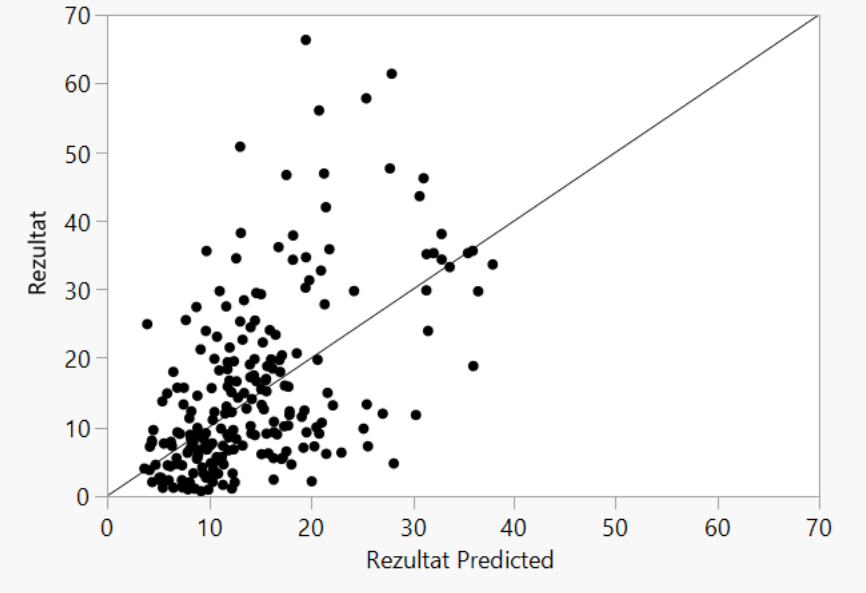


Slika 4. Primjer predviđanja primjenom modela k-najbližih susjeda

Prikazi graničnih modela temelje se na granično predviđenim vrijednostima i graničnim rezidualima. Varijable koje su tamnije osjenčane na grafu imaju veći utjecaj na promjenu izlaza od varijabli koje su svjetlije osjenčane na grafu.



Slika 5. Prikazi graničnih modela



Slika 6. Razlika stvarnih i predviđenih vrijednosti u modelu k-najbližih susjeda

Slika prikazuje grafikon kretanja stvarnih i predviđenih vrijednosti na skupu za validaciju modela. Slika prikazuje da točke djelomično padaju duž linije, što znači da su određene predviđene vrijednosti slične stvarnim vrijednostima. Nekoliko je instanci na gornjem kraju prikaza koje su udaljenije.

4.2.2. Model neuronske mreže

Prilikom razvoja modela neuronske mreže, najprije definiramo vrijednosti parametara temeljem kojih biramo strukturu i metodu učenja. U ovom istraživanju kao aktivacijska funkcija koristi se tangens hiperbolni koji transformira vrijednosti u interval -1 do 1 te predstavlja skaliranu verziju logističke funkcije.

Formula je:

$$\frac{e^{2x} - 1}{e^{2x} + 1}$$

Arhitektura neuronske mreže temelji se na tri sloja: ulaznom koji se sastoji od svih ulaznih varijabli, jednog skrivenog sloja koji obrađuje podatke te jednog izlaznog sloja u kojem je jedan neuron, varijabla ishod izbora mjerena ostvarenim postotkom glasova na izborima. Ovdje se koristi jedan skriveni sloj s obzirom na rezultat istraživanja Hornika i suradnika (Hornik et al., 1989) koji su empirijski dokazali da je jedan skriveni sloj dovoljan za obradu i najsloženijih problema.

Tablica 7. Parametri modela

Parametri modela	**Vrijednosti**
Metoda validacije	*k-struka* validacija (k=5)
Aktivacijska funkcija	*TanH* (Tangens hiperbolni)
Broj skrivenih slojeva	1
Broj neurona u skrivenom sloju	4

Broj neurona u skrivenom sloju jedan je od najvažnijih parametara neuronske mreže. Premali broj vodi do podtreniranosti, a preveliki do pretreniranosti. Odabir optimalnog broja neurona svodi se na metodu pokušaja i pogreški. Nekoliko prethodnih istraživanja u smjernicama za odabir broja

neurona navodi aritmetičku sredinu neurona na ulazu i izlaza kao broj oko kojeg se treba temeljiti odabir broja neurona u srednjem sloju.

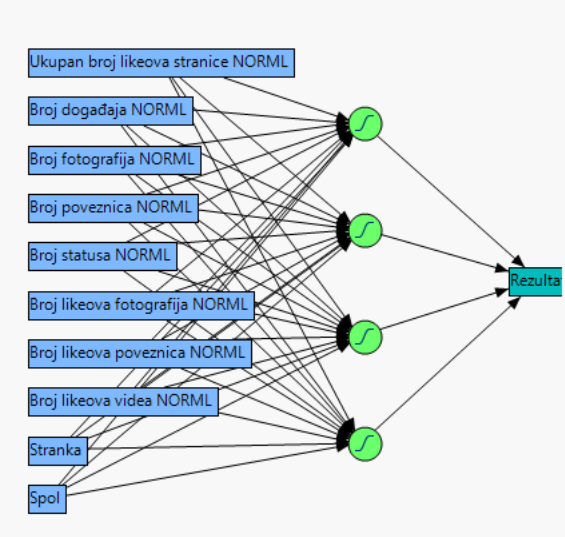
U ovom je istraživanju napravljeno 10-ak modela neuronske mreže u kojima je variran broj neurona u skrivenom sloju. Na kraju je odabran model sa četiri neurona u skrivenom sloju.

Tablica u nastavku prikazuje vrijednosti parametara kvalitete modela dobivenih ovom konfiguracijom.

Tablica 8. Parametri kvalitete modela neuronske mreže

Parametri	**Vrijednosti**
RSquare	0,8092956
RASE	5.9779712
Mean Abs Dev	3.7351163
Log*Like*lihood	141.10887059
SSE	1572.3901667

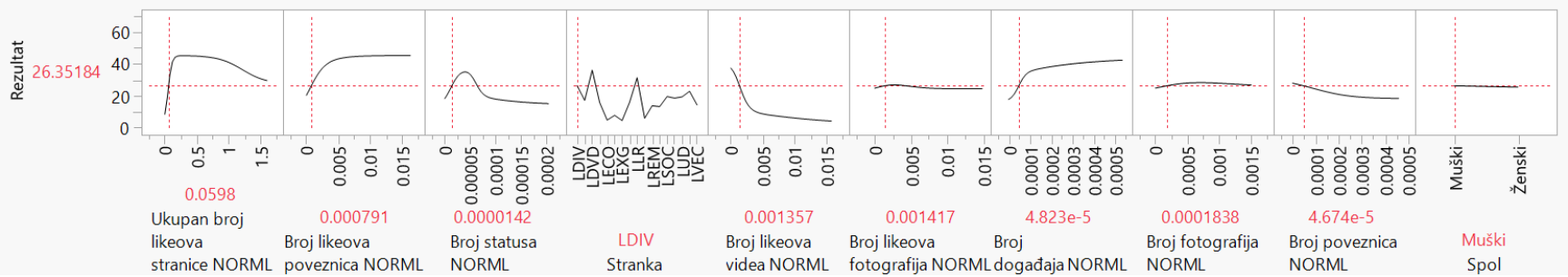
Pouzdanost modela mjerena kroz parametar Rsquare u iznosu od 0,809 pokazuje da se ovi rezultati mogu generalizirati izvan skupa podataka francuskih lokalnih izbora na kojima je model napravljen. Pouzdanost prediktivnog modela primjenom algoritma neuronske mreže je veća od pouzdanosti modela dobivenog algoritmom k-najbližih susjeda, a i greška modela je manja.



Slika 7. Arhitektura neuronske mreže

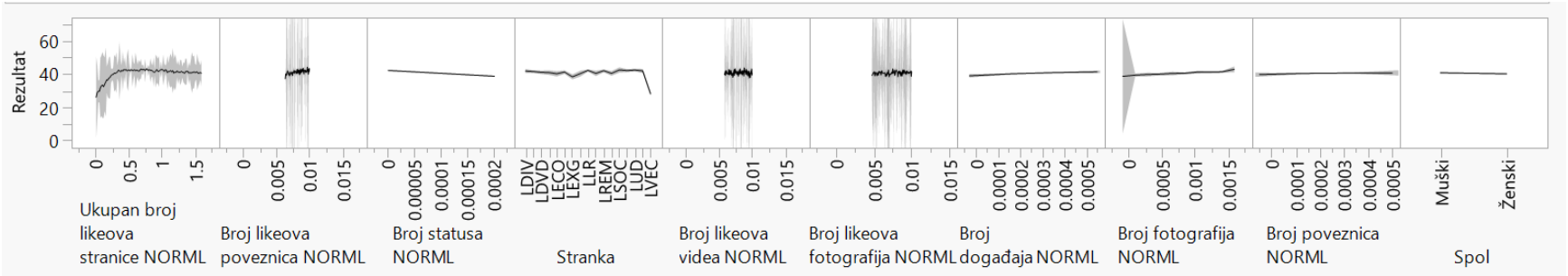
Slika 7. prikazuje arhitekturu neuronske mreže s tri sloja: jednim ulaznim slojem koji se sastoji od 10 neurona, jednim srednjim slojem koji se sastoji od 4 neurona te jednim izlaznim slojem koji se sastoji od jednog neurona – Rezultat izbora. Model neuronske mreže je jednostavan, nije pretreniran, a u velikoj mjeri prediktorske varijable objašnjavaju izlaznu.

Sljedeće dvije slike grafički prikazuju kretanje vrijednosti izlazne varijable ovisno o kretanjima vrijednosti ulaznih varijabli. Ulazne varijable poredane su po važnosti tj. razini utjecaja na promjenu izlaza.

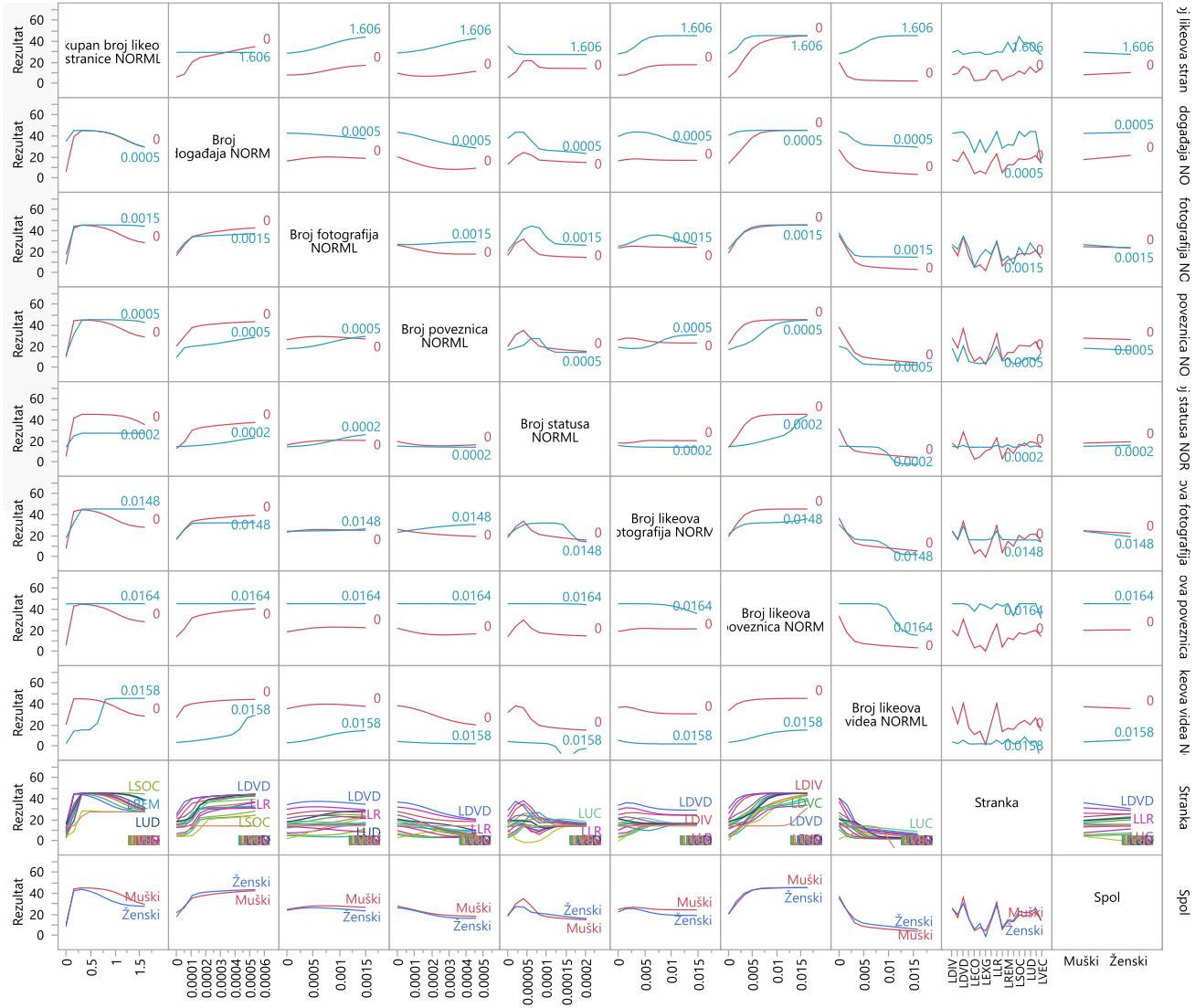


Slika 8. Predviđanje neuronskom mrežom

Prikazi graničnih modela temelje se na granično predviđenim vrijednostima i graničnim rezidualima. Varijable koje su tamnije osjenčane na grafu imaju veći utjecaj na promjenu izlaza od varijabli koje su svjetlije osjenčane na grafu.



Slika 9. Prikazi graničnih modela neuronske mreže



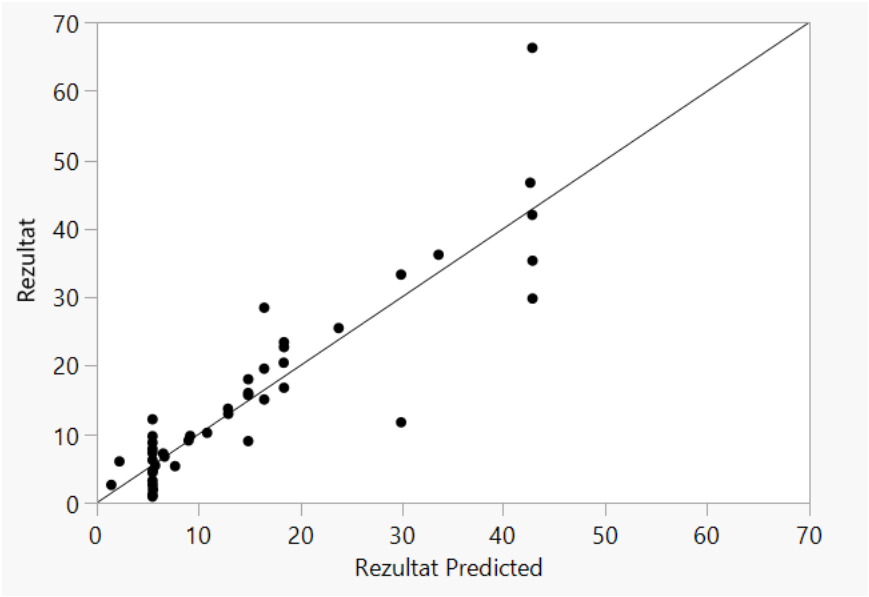
Slika 10. Profili interakcije

Slika 10. prikazuje profile interakcije koji daju dodatnu dimenziju objašnjivosti modela kroz prikaz interakcije varijabli.

Tablica 9. Najznačajniji prediktori rezultata u modelu neuronske mreže

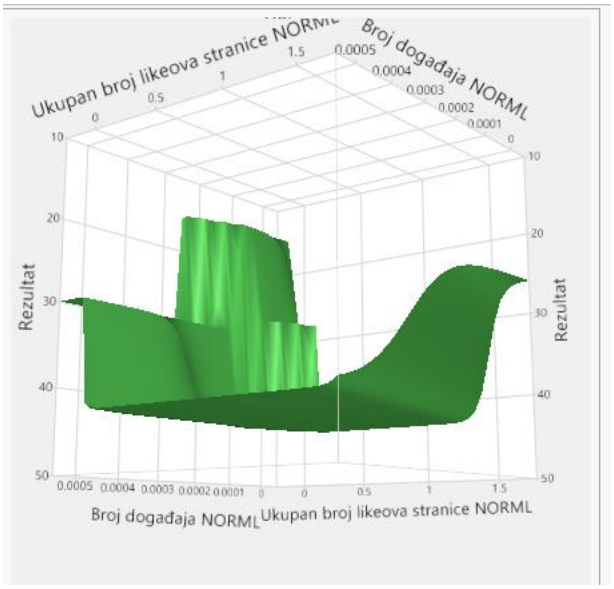
Varijabla	**Glavni utjecaj**	**Ukupan utjecaj**
Ukupan broj *likeova* stranice	0.1795001577	0.4190542356
Broj *likeova* poveznica	0.1312848765	0.3369768052
Broj statusa	0.0728433135	0.318096921
Stranka	0.1563168922	0.3016411502
Broj *likeova* videa	0.021411763	0.1876325292
Broj *likeova* fotografija	0.029100607	0.1452227427
Broj događaja	0.0213307034	0.0711208749
Broj fotografija	0.0150276427	0.0381945132
Broj poveznica	0.0094376789	0.0241733963
Spol	0.0079708049	0.0193208182

I za prediktivni model dobiven neuronskom mrežom, provedena je analiza osjetljivosti. Rezultati dani u tablici 9. pokazuju najveću prediktorsku snagu varijable Ukupan broj *likeova* stranice. Varijabla Spol ima najmanju prediktorsku snagu.



Slika 11. Razlika stvarnih i predviđenih vrijednosti neuronskom mrežom

Vizualizacija razlike između stvarnih vrijednosti rezultata i vrijednosti koje predviđa model dane su na slici 1. Vidljivo je da su instance mahom tik uz liniju, što sugerira točno predviđanje rezultata izbora.

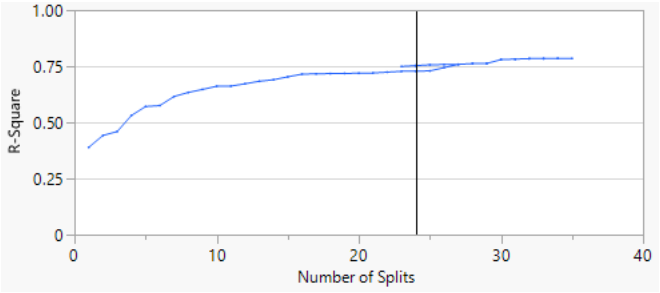


Slika 12.Prikaz 3D modela neuronske mreže

Slika 12. daje i 3D prikaz modela neuronske mreže.

4.2.3. Model stabla odlučivanja

Model stabla odlučivanja razvijen je temeljem smjernica za optimiranje hiperparametara predstavljenih u poglavlju 3.2.5. Broj grananja parametar je koji u velikoj mjeri određuje perfomanse modela te je razvijeno niz modela s različitim brojem grananja za koja je praćena pouzdanost modela. Ovisnost vrijednosti RSquare o broju grananja prikazana je na slici 13. u nastavku.



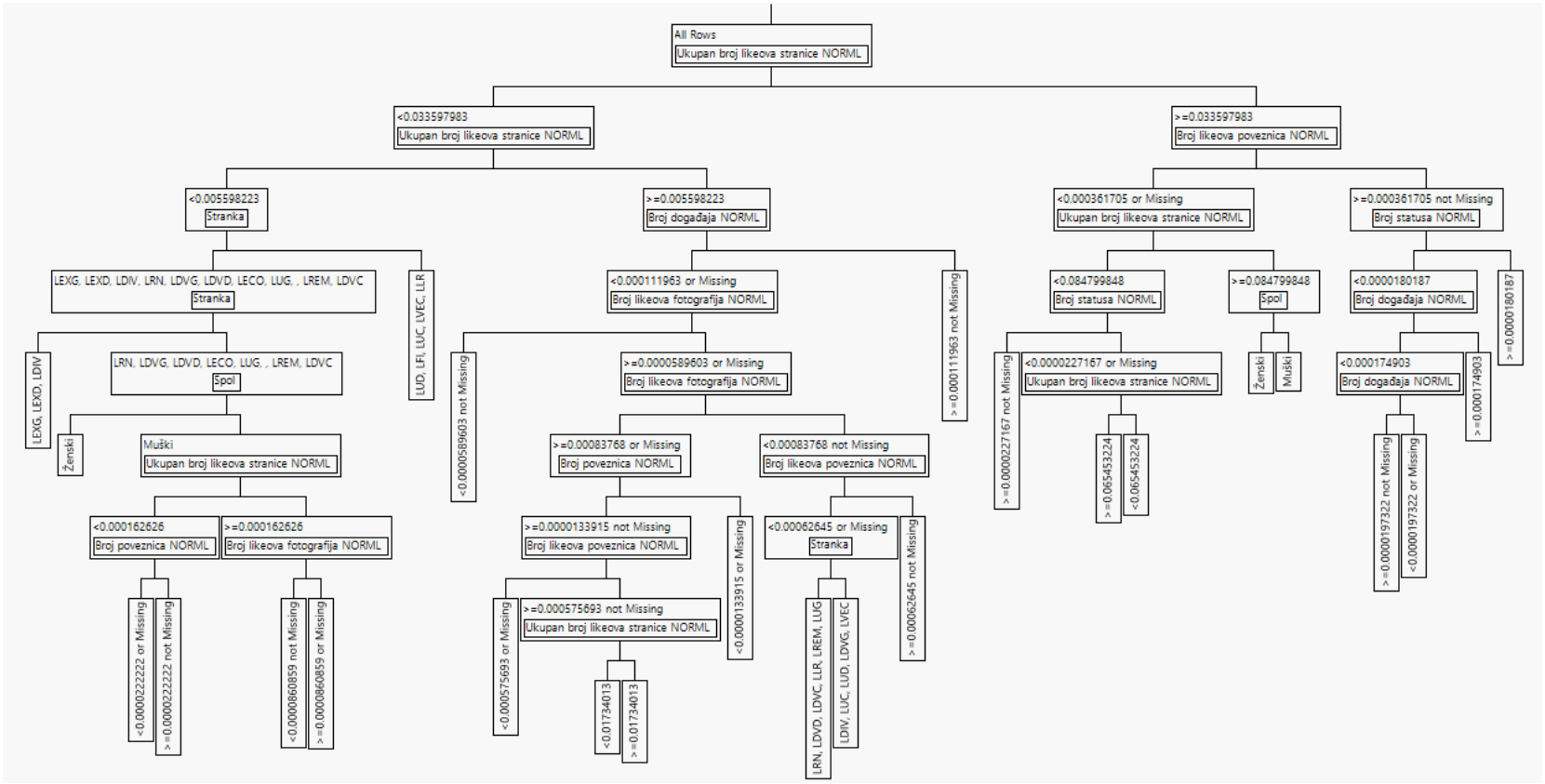
Slika 13. Kretanje parametra RSquare u ovisnosti od broja grananja

Rano zaustavljanje grananja modela stabla odlučivanja sprječava pretreniranost modela. Iako bi model s npr. 30 grananja postigao veću razinu pouzdanosti, grananje je zaustavljeno na 24. Metodom lakta utvrđeno je da se ovim brojem grananja postiže dovoljna razina pouzdanosti modela, a da se ne dolazi u pretreniranost.

Tablica 10. Parametri kvalitete modela stabla odlučivanja

RSquare	RASE	N	Broj grananja	AICc
0,753	6,3441513	223	24	1516,01

Parametri kvalitete stabla odlučivanja dani su u tablici 11. Model stabla odlučivanje postiže pouzdanost od 0,753, i grešku od 6,344.



Slika 14. Model stabla odlučivanja

Na slici 14. prikazano je cjelokupno stablo odlučivanja. Stablo se može pretvoriti u niz pravila oblika IF THEN. Pri tome svaka grana u stablu postaje jedno pravilo. Sva pravila generirana iz stabla prikazana su u tablici 11.

Tablica 11. Model stabla u obliku pravila

Pravilo	Prosječni rezultat
Ukupan broj *likeova* stranice <0.033597983 & Ukupan broj *likeova* stranice <0.005598223 & Stranka(LEXG, LEXD, LDIV)	2.473125
Ukupan broj *likeova* stranice <0.005598223 &Stranka(LRN, LDVG, LDVD, LECO, LUG, , LREM, LDVC) &Spol(Ženski)	4.4633333333
Ukupan broj *likeova* stranice NORML<0.000162626 & Stranka(LRN, LDVG, LDVD, LECO, LUG, , LREM, LDVC) &Spol(Muški) &Broj poveznica <0.0000222222	3.254
Ukupan broj *likeova* stranice <0.033597983 &Ukupan broj *likeova* stranice <0.005598223	5.8828571429

&Stranka(LRN, LDVG, LDVD, LECO, LUG, , LREM, LDVC) &Spol(Muški) &Ukupan broj *likeova* stranice <0.000162626 &Broj poveznica >=0.0000222222	
Ukupan broj *likeova* stranice <0.005598223 &Stranka(LRN, LDVG, LDVD, LECO, LUG, , LREM, LDVC) &Spol(Muški) &Ukupan broj *likeova* stranice >=0.000162626 &Broj *likeova* fotografija <0.0000860859	4.94
Ukupan broj *likeova* stranice NORML<0.005598223 &Stranka(LRN, LDVG, LDVD, LECO, LUG, , LREM, LDVC)& Spol(Muški) &Ukupan broj *likeova* stranice >=0.000162626 &Broj *likeova* fotografija >=0.0000860859	8.1505555556
Ukupan broj *likeova* stranice <0.005598223 &Stranka(LUD, LFI, LUC, LVEC, LLR)	11.6715
Ukupan broj *likeova* stranice <0.033597983 &Ukupan broj *likeova* stranice >=0.005598223 &Broj događaja <0.000111963	7.7566666667

&Broj *likeova* fotografija <0.0000589603	
Ukupan broj *likeova* stranice <0.033597983 &Ukupan broj *likeova* stranice >=0.005598223 &Broj događaja <0.000111963 &Broj *likeova* fotografija >=0.0000589603 &Broj *likeova* fotografija >=0.00083768 &Broj poveznica >=0.0000133915 &Broj *likeova* poveznica NORML<0.000575693	6.9771428571
Ukupan broj *likeova* stranice >=0.005598223 &Broj događaja <0.000111963 &Broj *likeova* fotografija >=0.00083768 &Broj poveznica >=0.0000133915 &Broj *likeova* poveznica >=0.000575693 &Ukupan broj *likeova* stranice <0.01734013	8.5466666667
Ukupan broj *likeova* stranice <0.033597983 &Broj događaja <0.000111963 &Broj *likeova* fotografija >=0.00083768 &Broj poveznica >=0.0000133915 &Broj *likeova* poveznica >=0.000575693 &Ukupan broj *likeova* stranice >=0.01734013	11.814
Ukupan broj *likeova* stranice <0.033597983 &Ukupan broj *likeova* stranice >=0.005598223	15.764285714

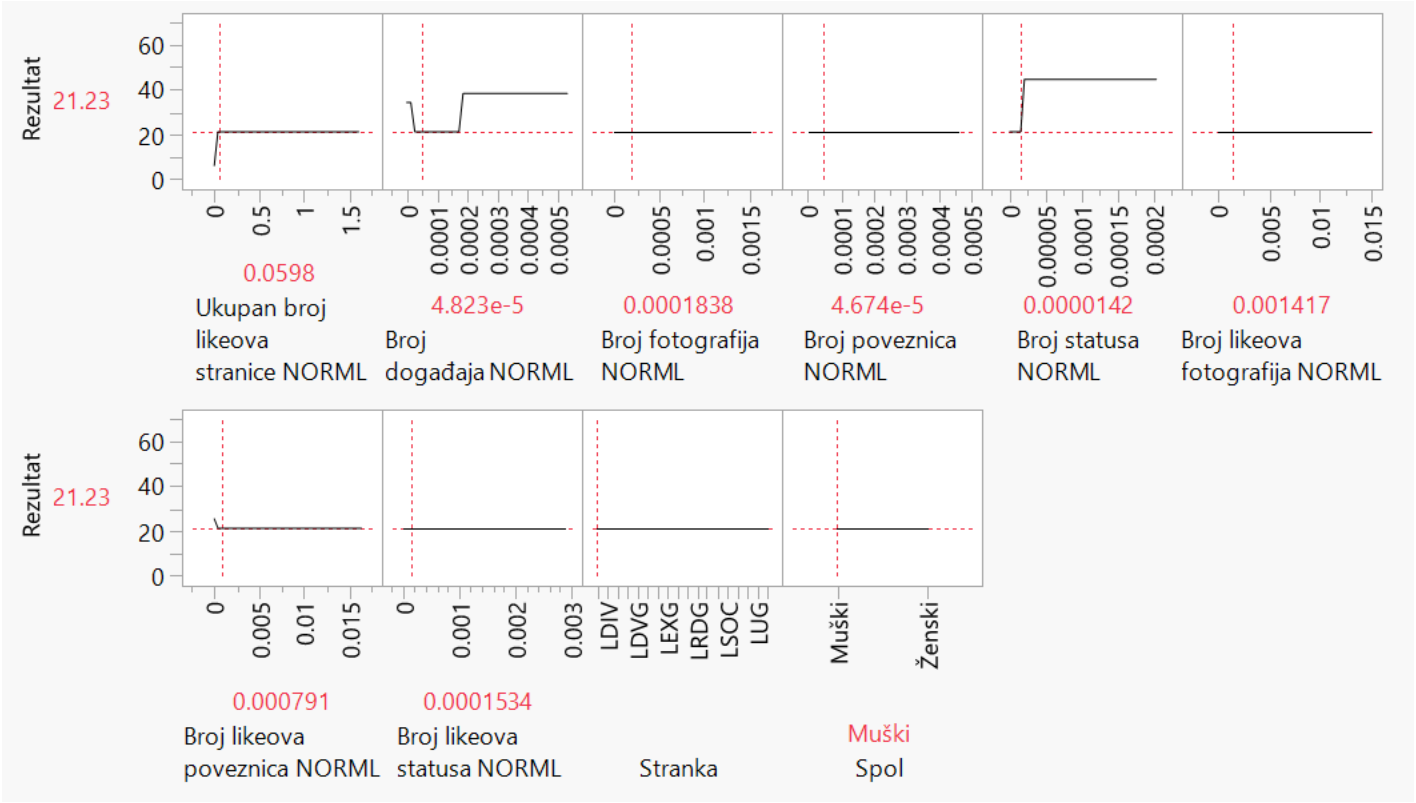
&Broj događaja <0.000111963 &Broj *likeova* fotografija >=0.00083768 &Broj poveznica <0.0000133915	
Ukupan broj *likeova* stranice <0.033597983 &Ukupan broj *likeova* stranice >=0.005598223 &Broj događaja <0.000111963 &Broj *likeova* fotografija >=0.0000589603 &Broj *likeova* fotografija <0.00083768 &Broj *likeova* poveznica <0.00062645 &Stranka(LRN, LDVD, LDVC, LLR, LREM, LUG)	11.322352941
Ukupan broj *likeova* stranice <0.033597983 &Ukupan broj *likeova* stranice >=0.005598223 &Broj događaja <0.000111963 &Broj *likeova* fotografija >=0.0000589603 &Broj *likeova* fotografija <0.00083768 &Broj *likeova* poveznica <0.00062645 &Stranka(LDIV, LUC, LUD, LDVG, LVEC)	19.203333333
Ukupan broj *likeova* stranice <0.033597983 &Ukupan broj *likeova* stranice >=0.005598223 &Broj događaja <0.000111963 &Broj *likeova* fotografija >=0.0000589603 &Broj *likeova* fotografija <0.00083768	25.084

&Broj *likeova* poveznica >=0.00062645	
Ukupan broj *likeova* stranice <0.033597983 &Ukupan broj *likeova* stranice >=0.005598223 &Broj događaja >=0.000111963	23.047142857
Ukupan broj *likeova* stranice >=0.033597983 &Broj *likeova* poveznica <0.000361705 o &Ukupan broj *likeova* stranice <0.084799848 &Broj statusa >=0.0000227167	11.03875
Ukupan broj *likeova* stranice >=0.033597983 &Broj *likeova* poveznica <0.000361705 &Ukupan broj *likeova* stranice <0.084799848 &Broj statusa <0.0000227167 &Ukupan broj *likeova* stranice >=0.065453224	13.585
Ukupan broj *likeova* stranice >=0.033597983 &Broj *likeova* poveznica <0.000361705 &Ukupan broj *likeova* stranice <0.084799848 &Broj statusa <0.0000227167	25.494444444
Broj *likeova* poveznica <0.000361705 &Ukupan broj *likeova* stranice >=0.084799848 &Spol(Ženski)	27.57625
Ukupan broj *likeova* stranice >=0.033597983 &Broj *likeova* poveznica <0.000361705	33.908333333

&Ukupan broj *likeova* stranice >=0.084799848 &Spol(Muški)	
Ukupan broj *likeova* stranice >=0.033597983 &Broj *likeova* poveznica >=0.000361705 &Broj statusa <0.0000180187 &Broj događaja >=0.0000197322	21.23
Ukupan broj *likeova* stranice >=0.033597983 &Broj *likeova* poveznica >=0.000361705 &Broj statusa NORML<0.0000180187 &Broj događaja NORML<0.0000197322	34.348
Ukupan broj *likeova* stranice >=0.033597983 &Broj *likeova* poveznica >=0.000361705 &Broj statusa <0.0000180187 &Broj događaja >=0.000174903	38.36

Unutar svakog pravila navedena je i predviđena izlazna vrijednost, tj. definirano je koliki se postotak osvojenih glasova predviđa za kandidata karakteristika navedenih s lijeve strane pravila. Treba napomenuti da su vrijednosti varijabli normalizirane. Pa prilikom interpretacije rezultata tu činjenicu treba uzeti u obzir i vrijednosti varijable ponderirati s veličinom biračkog tijela kako bi se postigla puna i točna interpretacija rezultata.

Sljedeća slika grafički prikazuje kretanje vrijednosti izlazne varijable ovisno o kretanjima vrijednosti ulaznih varijabli. Ulazne varijable poredane su po važnosti tj. razini utjecaja na promjenu izlaza. Prikazana situacija je za prosječne vrijednosti varijabli.



Slika 15. Predviđanje stablom odlučivanja

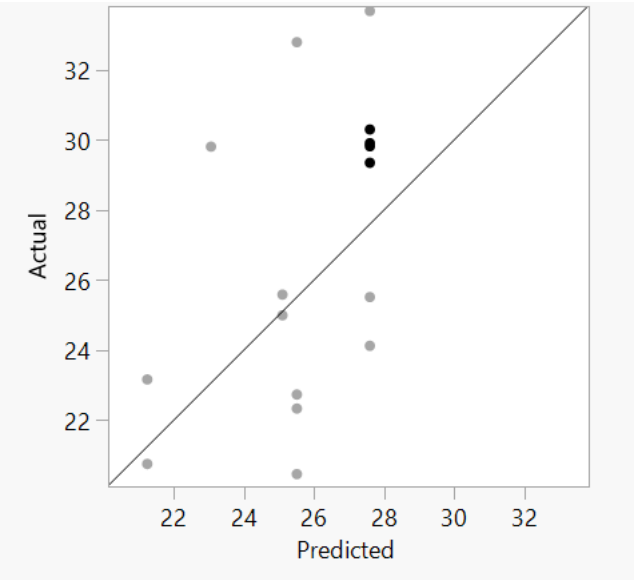
Tablica 12. prikazuje rezultate provedene analize osjetljivosti u obliku udjela kojeg izaziva promjena pojedine ulazne varijable na vrijednost izlazne varijable, postotak koji kandidat osvaja na izborima.

Tablica 12. Najznačajniji prediktori rezultata u modelu stabla odlučivanja

Varijabla	**Broj razdvajanja**	**SS**	**Udio**
Ukupan broj *likeova* stranice	6	18126.869424	0.6618
Broj *likeova* poveznica	3	3106.075239	0.1134
Broj statusa	2	1950.0035294	0.0712
Broj događaja	3	1721.9387355	0.0629
Stranka	3	1231.7289931	0.0450
Broj *likeova* fotografija	3	823.82962193	0.0301
Broj poveznica	2	261.53189894	0.0095
Spol	2	168.30945443	0.0061
Broj fotografija	0	0	0.0000
Broj *likeova* statusa	0	0	0.0000

Ukupan broj *likeova* stranice i u ovom je modelu najznačajniji prediktor ishoda izbora, a slijedi varijabla broj *likeova* poveznica. Ovi rezultati u skladu su s rezultatima ranije razvijenih modela, k-najbližih susjeda i neuronske mreže.

Slika 16. vizualizira odstupanja predviđenih vrijednosti rezultata izbora u odnosu na stvarno ostvarene rezultate izbora.



Slika 16. Razlika stvarnih i predviđenih vrijednosti u modelu stabla odlučivanja

4.2.4. Model naivnog Bayesova klasifikatora

Za potrebe izrade prediktivnog modela naivnim Bayesovim klasifikatorom zavisna varijabla nije bila kontinuirana numerička kao u prethodno opisanim modelima (postotak osvojen na izborima), već kategorijska (mjesto ostvareno na izborima). Točnost i pouzdanost dobivenog prediktivnog modela prikazana je u tablici 13.

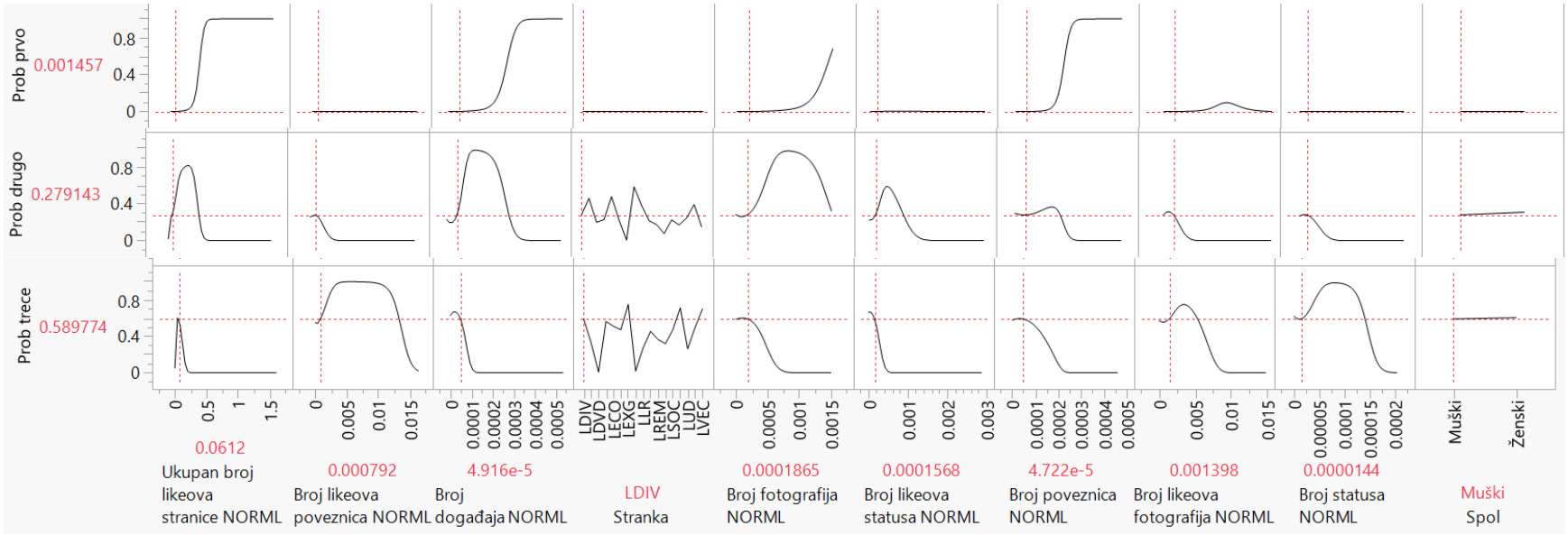
Tablica 13. Parametri kvalitete modela Naivnog Bayesa

Mjera	Vrijednost
RSquare	-0,243
RASE	0,547

Vrijednost prosječne greške je najveća u odnosu na prethodna tri modela. Mjera prilagođene pouzdanosti modela je blago negativna, što nam govori da ulazi modela ne doprinose objašnjavanju izlaza.

Sljedeće dvije slike prikazuju grafički kretanje vrijednosti izlazne varijable ovisno o kretanjima vrijednosti ulaznih varijabli. Ulazne varijable poredane su po važnosti tj. razini utjecaja na

promjenu izlaza. Prikazana situacija je za prosječne vrijednosti varijabli za svaku kategoriju izlazne varijable.



Slika 17. Predviđanje modelom Naivnog Bayesa

Analiza osjetljivosti provedena je za svaku kategoriju (prvo, drugo, treće, četvrto, peto, šesto, sedmo i osmo mjesto). Ovdje su prikazani samo rezultati za cijeli model. Kad se uzmu u obzir sve kategorije, najjači prediktor rezultata je varijabla ukupan broj *likeova* stranice.

Tablica 14. Najznačajniji prediktori rezultata naivnim Bayesom za cjelokupni model

Varijabla	**Glavni utjecaj**	**Ukupni utjecaj**
Ukupan broj *likeova* stranice	0.1398407403	0.3401802236
Broj *likeova* poveznica	0.082436651	0.2208578911
Broj događaja	0.0927907327	0.2108167548
Stranka	0.0635672272	0.2011485921
Broj fotografija	0.0699747023	0.1837435109
Broj *likeova* statusa	0.0620748859	0.1333705057
Broj poveznica	0.0509488554	0.1223962581
Broj *likeova* fotografija	0.0254032896	0.0858654447
Broj statusa	0.0237154426	0.0564038757
Spol	0.0020741181	0.0046525387

4.4. Evaluacija modela

4.4.1. Usporedba prediktivnih modela temeljenih na podacima društvene mreže

Usporedba metoda strojnog učenja i odabir najboljeg modela uobičajena je praksa u primijenjenom strojnom učenju. Modeli se obično procjenjuju pomoću metoda ponovnog uzorkovanja, poput k-struke unakrsne validacije (engl. *k-fold cross validation*), iz koje se izračunavaju prosječne vrijednosti parametara kvalitete modela i izravno uspoređuju prosječni rezultati. U ovom je istraživanju primijenjena 10-struka unakrsna validacija prilikom razvoja svaka od četiri modela strojnog učenja. Prosječni rezultati pogreške modela su prikazani u tablici u nastavku.

Tablica 15 Usporedba greški modela

Model	Greška modela
k-NN	8.9899
Neuronska mreža	5.9779712
Stablo odlučivanja	6.3441513
Naivni Bayes	12.4532

U prethodno prikazanim modelima vrijednost RASE za modele k-NN, neuronske mreže i stabla odlučivanja bila je prikazana kao apsolutna vrijednost, a naivni Bayes kao relativna vrijednost. U tablici 19. sve su vrijednosti svedene na apsolutnu kako bi bile usporedive.

Po istom principu računata je i pouzdanost modela mjerena parametrom *Rsquare,* a prikazana je u tablici u nastavku.

Tablica 16. Rezultati pouzdanosti modela

Model	Pouzdanost modela
k-NN	0.5044
Neuronska mreža	0.8093
Stablo odlučivanja	0,753
Naivni Bayes	-0,243

Model dobiven neuronskom mrežom ima najbolje parametre kvalitete: daje natočniju predikciju i model je najviše razine pouzdanosti, što znači da se rezultati mogu generalizirati i izvan skupa podataka na kojima je razvijen.

4.4.2. Usporedba prediktivnih modela temeljenih na podacima društvene mreže i podacima ankete

Kako bi se usporedili prediktivni modeli dobiveni algoritmima strojnog učenja i rezultati predizbornog anketiranja, korišteni su podaci ustupljeni od strane francuskog instituta za ispitivanje javnog mnijenja, IFOP. Podaci se odnose na 6 gradova: Pariz, Lyon, Marseille, Rennes, Nantes i Bordeaux, te su time uključeni podaci za ukupno 44 kandidata na izborima.

Temeljem samih rezultata ne možemo utvrditi je li razlika u pogrešci prediktivnih modela dobivenih algoritmima strojnog učenja i anketiranjem stvarna ili slučajna. Kako bi se to utvrdilo potrebno je provesti test statističke značajnosti koji rješava ovaj problem. Općenito, test statističke hipoteze za usporedbu uzoraka kvantificira koliko je vjerojatno promatranje dva uzorka podataka s obzirom na pretpostavku da uzorci imaju istu distribuciju. Pretpostavka statističkog testa naziva se nulta hipoteza te primjenom testa računamo statističke mjere i interpretiramo ih kako bismo odlučili hoćemo li prihvatiti ili odbaciti hipotezu.

Usporedba modela strojnog učenja putem testova statističke značajnosti nameće neka očekivanja koja utječu na vrstu statističkih testova koji se mogu koristiti:

(i) Procjena kvalitete. Mora se izabrati određena mjera kvalitete modela. To može biti točnost klasifikacije ili srednja apsolutna pogreška koja će ograničiti vrstu ispitivanja koja se može koristiti.

(ii) Ponovljene procjene. Za izračun testa potreban je uzorak ocjena kvalitete. Ponovljena testiranja datog modela na istim ili različitim podacima utjecat će na vrstu testa koji se može koristiti.

(iii) Distribucija procjena. Uzorak procjena ocjena kvalitete imat će možda Gaussovu distribuciju, a možda ne. Distribucija određuje mogu li se koristiti parametrijski ili neparametrijski testovi.

(iv) Središnja tendencija. Kvaliteta modela često će se opisivati i uspoređivati pomoću zbirne statistike, kao što je srednja vrijednost ili medijan.

Najčešće korišteni test statističke značajnosti je t-test (Witten et al., 2016). No, ovdje ne može biti primijenjen iz nekoliko razloga. Prvo, uzorci nisu nezavisni. Naime, kao dio k-strukog postupka unakrsne validacije, određena se instanca koristi u skupu podataka za treniranje (k-1) puta. To znači da su procijenjeni rezultati ovisni, a ne neovisni, te će izračun t-statistike u testu biti pogrešan zajedno s bilo kakvim tumačenjima statistike i p-vrijednosti. Sukladno sugestijama Demšara koristit će se neparametrijski Wilcoxon test (Demšar, 2006).

Rezultati statističkog testa su statistika testa i p-vrijednost, a oboje se tumači i koristi u interpretaciji rezultata kako bi se kvantificirala razina značajnosti u razlici među rezultatima.

Testiranjem značajnosti razlika traži se statistički značajna razlika u pogrešci modela na razini $p < ,0001$

U tablici 17. dan je dio podataka (za 9 kandidata) koji je korišten u testiranju statističke značajnosti razlika u rezultatima.

Tablica 17. Dio podataka za testiranje razlika

Rezultat stvarni	Predviđanje anketa	Predviđanje ANN	Greška anketa	Greška ANN model
4.57	5	5.008123666	0.43	0.438123666
29.33	25.5	25.51043416	3.83	3.819565838
10.79	11.5	11.35966427	0.71	0.569664266
7.1	7	7.134700585	0.1	0.034700585
18.04	19	17.4004099	0.96	0.639590098
22.72	24	20.78283435	1.28	1.937165652
4.57	5	5.008123666	0.43	0.438123666
29.33	25.5	25.51043416	3.83	3.819565838
10.79	11.5	11.35966427	0.71	0.569664266

U tablici 18. dani su rezultati provedenog testiranja, kao i statistička značajnost dobivenih rezultata.

Tablica 18. Rezultati testiranja statističke značajnosti razlika u modelima temeljenih na anketi i modelima temeljenih na podacima Facebooka

	Greška anketa- Greška ANN model
ChiSquare	44.955
Prob>|ChiSquare|	0.4317

Razina značajnosti od 0,0001 ukazuje na rizik manji od 0,1% zaključka da razlika postoji kad nema stvarne razlike. Pošto je p-vrijednost veća od razine značajnosti, utvrđuje se da ne postoji statistička značajna razlika u rezultatima koji se dobivaju prediktivnim modelima temeljenim na podacima društvene mreže Facebook i algoritmu umjetne neuronske mreže u odnosu na predizborna ispitivanja javnog mnijenja pomoću ankete. Modeli dobiveni na podacima društvene mreže imaju veću točnost od modela dobivenih anketiranjem, ali ta razlika nije statistički značajna.

4.4.3.Utvrđivanje značajnosti temporalne komponente

S ciljem utvrđivanja značajnosti ispitivanja temporalne komponente, promatrani podaci su se pratili u tri vremenska perioda: (i) na početku kampanje, (ii) u sredini kampanje i (iii) zadnji dan kampanje. Periodi su kreirani tako da se prvi period odnosi na aktivnosti u početku kampanje (od 02.03. do 06.03.), drugi skup podataka na aktivnosti kandidata u sredini kampanje (07.03. do 11.03.) te treći skup podataka na završetak kampanje (12.03. do 14.03). Na svakom od ta tri skupa podataka razvijen je prediktivni model primjenom algoritma umjetne neuronske mreže koji se pokazao najboljim na cijelom skupu podataka. Rezultati usporedbe prosječne progreške tri modela dana su u tablici u nastavku.

Tablica 19. Usporedba greški modela u tri vremenska perioda

	Prvo razdoblje	Drugo razdoblje	Treće razdoblje
Greška modela ANN	7.2145234	7.5467111	7.4428194

Friedman test koristi se za utvrđivanje postoji li statistički značajna razlika u pogrešci modela između triju razdoblja. Za ovu vrstu evaluacije predlažu ga Japkowicz i Shah (Japkowicz i Shah,

2011), a koristili su ga i Čehovin i Bosnić (Čehovin i Bosnić, 2010) te Demšar (Demšar, 2007) u istraživanjima koja su se također odnosila na situaciju komparacije više tehnika/modela. Friedman test je neparametrijski test koji pod nul hipotezom pretpostavlja da su rezultati svih modela jednaki, dok odbijanje nul hipoteze sugerira postojanje razlika između razvijenih modela u različitim vremenskim periodima.

Tablica koja testira statistiku Friedman testa nalazi se u nastavku, a govori da li postoji statistički signifikantna razlika u modelima. Vrijednosti koje daju su: vrijednost statistike testa (test statistics, X^2), broj stupnjeva slobode (df) i razinu statističke značajnosti (p).

Rezultati su dani u tablici u nastavku.

Test statistics (X^2)	5.2
Df	2
p	0.07427

Tablica 20. Rezultati testiranja statističke značajnosti razlika u modelima u različitim vremenskim periodima

S obzirom na rezultate Friedman testa i vrijednost p veću od 0.01 (nije signifikantna vrijednost), zaključujem da su medijalne vrijednosti za sva tri promatrana razdoblja jednaka. Stoga rezultati testa pokazuju da nema statistički značajne razlike (na razini značajnosti od 0.01) između prediktivnih modela dobivenih u prvom, drugom i trećem razdoblju.

4.5. Korištenje modela

U znanstvenim istraživanjima zadnji korak CRISP-DM standarda, korištenje modela, provodi se kao ekstrakcija znanja. U kontekstu ove teme, ekstrakcija znanja se odnosi na utvrđivanje najznačajnih prediktora ishoda izbora temeljem dobivenih prediktivnih modela. S tom je svrhom

provedena analiza osjetljivosti svakog prediktivnog modela. Rezultati u obliku ranga svake varijable prikazani su u tablici u sljedećem poglavlju.

4.5.1. Utvrđivanje najznačajnijih prediktora rezultata

Tablica 21. Usporedba najznačajnijih prediktora rezultata

Varijabla	Rang u k-NN	Rang u neuronskoj mreži	Rang u Naivnom Bayesu	Rang u stablu odlučivanja	Prosječan rang
Ukupan broj *likeova* stranice	1	1	1	1	**1**
Broj događaja	6	7	3	4	**5**
Broj fotografija	5	8	5	9	**6**
Broj poveznica	3	9	7	7	**6,5**
Broj statusa	4	3	9	3	**4,75**
Broj *likeova* fotografija	8	6	8	6	**7**
Broj *likeova* poveznica	2	2	2	2	**2**
Broj *likeova* statusa	7	5	6	10	**7**
Spol	9	10	10	8	**9,25**
Stranka	10	4	4	5	**5,75**

Rezultati najznačajnijih prediktora rezultata četiri prediktivna modela dobivena primjenom četiri različitih pristupa strojnom učenju pokazuju stabilnost i konzistentnost modela. Ukupan broj

likeova stranice kandidata najznačajniji je prediktor u sva četiri modela. Broj *likeova* poveznica drugi je najznačajniji prediktor također u sva četiri modela. Broj statusa treći je najjači prediktor izbornih rezultata u dva od četiri promatrana modela.

4.5.2. Profiliranje birača

Dobiveni prediktivni modeli služe kao alat za planiranje i upravljanje kampanjama te kao takvi služe kao instrument za potporu odlučivanje kroz profiliranje (Klepac, Mrsic & Kopal, 2016). Profiliranje će se provesti temeljem razvijenih prediktivnih modela dobivenih neuronskim mrežama iz razloga jer su ti modeli najkvalitetniji. Neuronske mreže igraju značajnu ulogu u prediktivnom modeliranju, pa tako i za profiliranje. Uzimanjem u obzir arhitekture neuronskih mreža i konačnog izlaza, modeliranje neuronskom mrežom i profiliranje iz prediktivnih modela nije lak zadatak (Larose, 2005).

Profiliranje služi kao temelj za potporu odlučivanju kod vođenja kampanja jer podaci o profilima glasača koji preferiraju određenog kandidata služe kao okvir za planiranje kampanja i upravljanje kampanjama. Profili birača će se identificirati temeljem provedene analize osjetljivosti i dobivenih vrijednosti relevantnosti varijabli za predikciju ishoda izbora. Za svakog promatranog kandidata, ciljna varijabla ishod izbora treba postići maksimalnu vjerojatnost odabira kandidata upotrebom analize osjetljivosti. Takvav pristup, između ostalog, može dati i odgovor na pitanje: koje su tipične karakteristike birača određenog kandidata.

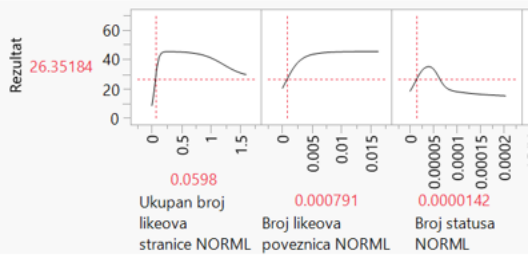
Planiranje kampanje u vezi s upravljanjem odnosima s biračima usko je povezano s profilima birača. Nemaju svi birači iste navike prilikom korištenja društvene mreže Facebook. Prepoznavanje profila birača ključno je za uspješno vođenje kampanje na društvenoj mreži.

Analiza relevantnosti atributa kroz analizu osjetljivosti ima važnu ulogu u razumijevanju ključnih čimbenika u vezi s upravljanjem odnosima s biračima. Identificirana relevantnost varijabli i znanje o odnosima varijabli i utjecaju jedne varijable na drugu predstavljaju moćan alat za upravaljanje političkim kampanjama, posebno na vođenje kampanja na društvenim mrežama.

Analiza osjetljivosti ima nekoliko uloga: (i) prepoznavanje najvažnijih varijabli koje imaju najveći utjecaj na zavisnu varijablu, (ii) razumijevanje odnosa i logike između važnih prediktora i zavisne varijable kao i razumijevanje odnosa i logike između važnih prediktora iz perspektive zavisne varijable (Klepac, 2017).

Obje funkcije su u skladu s prepoznavanjem profila birača (Klepac, 2015).

Kako bi se dobilo razumijevanje utjecaja najznačajnijih varijabli na ciljnu varijablu provedena je analiza osjetljivosti, te će se interpretirati kako najutjecanije varijable u predikitvnom modelu dobivenom algoritmom neuronskih mreža (broj *like*-ova stranice, broj statusa i broj *like*-ova poveznica) utječu na ciljnu varijablu. U tu svrhu ispitana je zavisnost tri najutjecajnije prediktivne varijable na ciljnu varijablu, a vizualizacija je prikazana na slici u nastavku.



Slika 18. Zavisnosti najutjecajnijih prediktivnih varijabli i ciljne varijable

Rezultati prikazani na slici pokazali su da neke od varijabli imaju profile s pozitivnim nagibima, a neke s negativnim. Na primjer, varijabla broj *like*-ova poveznica ima pozitivan nagib. To znači da što više je više *like*-ova poveznica kandidat ostvario na društvenoj mreži Facebook, veća je predviđena srednja vrijednost postotka glasova dobivenih na izborima. Varijable ukupan broj *like*-ova stranice i broj statusa u nekom intervalu vrijednosti imaju pozitivan nagib, a u nekom intervalu negativan nagib. Obje varijable kod manjih vrijednosti imaju pozitivan nagib, a kod većih vrijednosti imaju negativan nagib. Ovakva ovisnost prediktivnih varijabli na ciljnu varijablu ukazuje nelinearne odnose između prediktorskih varijabli i ciljne varijable.

Temeljem tri najznačajnije varijable provedeno je profiliranje potencijalnih birača primjenom algoritma klasteriranja k-sredina. Rezultati prikazani u tablici 22. ukazuje da se najbolje grupe profila dobivaju s 4 klastera.

Tablica 22. Evaluacija modela klastera

Broj klastera	CCC
2	-2.14378
3	-4.4026
4	1.195353
5	1.126908

Kao kriterij evaluacije korišten je Cubic Cluster Criterion (CCC) pri čemu veća vrijednost CCC ukazuje na bolje grupiranje instanci. Varirani su različiti brojevi klastera (od 2 do 5) i model s 4 klastera ima najveću vrijednost CCC. Za taj su model identificirane karakteristike mjerene srednjim vrijednostima varijabli temeljem kojih su napravljene. Rezultati su prikazani u tablici u nastavku.

Tablica 23. Karakteristike klastera

Klaster	Ukupan broj likeova stranice	Broj statusa	Broj likeova poveznica
1	0.035719	0.103282	0.053128
2	0.768205	0	0.30987
3	0.038391	0.571429	0.072261
4	0.077642	0.040816	0.564132

Prvu grupu čini profil s najmanjim prosječnim brojem *like*-ova stranica i poveznica, ali umjerenim brojem statusa kandidata. Drugu grupu čine kandidati s najvećim brojem *like*-ova stranice, umjerenim brojem *like*-ova poveznice i bez statusa. Treću grupu karakterizira najveći broj statusa kandidata, a četvrtu grupu najveći broj *like*-ova poveznica kandidata.

4.5.3. Diskusija rezultata

Najkvalitetniji modeli dobiveni su primjenom algoritma umjetne neuronske mreže u razvoju prediktivnih modela. Dio objašnjenja nalazi se u tipovima podataka. Ulazne varijable su većinom numerički kontinuirani atributi. Modeli s najnižim vrijednostima pouzdanosti i točnosti dobiveni su naivnim Bayesovim klasifikatorom. Objašnjenje rezultata također se može povezati s

karakteristikama skupa podataka na kojem je model razvijen. Naivni Bayesov klasifikator radi s kategorijskim varijablama te je nužna transformacija kontinuiranog izlaza u kategorijski.

Najvažniji rezultati ovog istraživanja pružaju brojne implikacije za korištenje društvenih medija kao indikatora za predviđanje ishoda izbora te se mogu dati smjernice za učinkoviti pristup upravljanju podacima društvenih mreža. Prvo je otkriveno kako je apsolutni broj Facebook pratitelja jako dobar prediktor izbornih ishoda. Drugo je utvrđeno da je sadržaj koji kandidati plasiraju putem društvenih medija, kao i reakcije na specifične sadržaje koje dijele u određenoj mjeri indikator ishoda izbora.

S obzirom na niske troškove povezane s povezivanjem na društvenim mrežama, ovi rezultati potencijalno vode smanjenju troškova u političkim kampanjama. Naravno, korisnici društvene mreže mogu biti prijatelji ili pratiti mnoge kandidate, ne samo njihovog preferiranog kandidata, uključujući kandidate u biračkim tijelima na koje nemaju pravo glasa. Stoga i broj pratitelja nije jedinstven indikator.

U diskusiji rezultata treba uzeti u obzir kako znanstvenici sugeriraju da korisnici društvenih medija mogu biti manje osjetljivi na društvenu pristranost nego sudionici ankete kada raspravljaju o svojim političkim preferencijama (Payne, 2010.). Općenito je anketa pristrana jer sadrži samo određeni uzorak ispitanika iz određene regije ili izborne jedinice. Koristeći o podatke društvenih mreža, može se dobiti šira slika jer se podaci društvenih medija lakše prikupljaju iz različitih gradova, regija i/ili izbornih jedinica. Nadalje, možemo poboljšati preciznost predviđanja kombiniranjem predloženog pristupa s pristupom tradicionalnog anketiranja.

Naravno, ograničenje istraživanja je da mišljenje baze korisnika društvene mreže Facebook nije reprezentativno za cjelokupnu populaciju, ali svakako pridonosi fenomenu objašnjavanja i predviđanja izbornih rezultata.

Rezultati istraživanja vode i boljem razumijevanje načina na koji društvene mreže prezentiraju stvavove birača, te ukazuju kako se može djelovati na biračko tijelo preko društvenih mreža.

Rezultati istraživanja provedenog u ovom radu prvenstveno pomažu u definiranju smjernica kandidatu kako voditi kampanju na društvenoj mreži Facebook s kojom bi utjecao na bolji rezultat na političkim izborima te bi sugestije išle u nekoliko smjerova. Prije svega, to je postizanje

vidljivosti na Facebooku, budući da se iz nje generira ono što je najbitnije, a to je da se dobije velik broj *likeova* stranice odnosno pratitelja. Pritom se mogu koristiti razne metode povećavanja vidljivosti od plaćanja oglašavanja (što u omjeru s klasičnim medijskim oglašavanjem daje bolji *impact)*, pa do korištenja tzv. stranačkih mašinerija (tu su velike stranke ili kandidati koji su već na postojećim funkcijama u prednosti) koje generiraju iz svojih baza članstva velik broj *likeova* stranice ili dijeljenja sadržaja, a što onda generira visoko mjesto kod prosječnih korisnika Facebooka u tzv. *news feedu* i samim time dolazi do većeg prenošenja poruka i indirektnog utjecaja na birače. Pritom je zanimljivo da sadržaj i nije toliko presudan koliko vidljivost kandidata, pa bih primjerice kandidatu bilo uputnije prikazati sadržaj iz njegovog uobičajenog života (primjerice, bavljenje rekreacijom ili nekim hobijem) nego neke političke poruke. Takva poruka bi siguran utjecala na veću vidljivost, a koja može utjecati na rezultat izbora za tog kandidata, budući da se glavne političke poruke uglavnom znaju i komuniciraju putem drugih medija, a kod toga se Facebook sudeći po ovom istraživanju treba manje koristiti, a glavni naglasak staviti na aktivnosti na povećanje vidljivosti kandidata kombiniranom s manjim brojem poruka političkog sadržaja.

Odgovori na istraživačka pitanja, rezultati testiranja hipoteza, doprinos rada, ograničenja istraživanja i smjernice za nastavak istraživanja u ovom području dana su u zaključku rada.

5. ZAKLJUČAK

Definirani ciljevi istraživanja su ispunjeni:

- utvrđena je prediktivna moć modela temeljenih na društvenoj mreži Facebook i uspoređena je s drugim vrstama istraživanja,
- utvrđeno je koje varijable su najznačajniji prediktori ishoda izbora,
- diskutirana je uloga temporalne komponente podataka društvene mreže Facebook,
- dokazano je koja od četiri metode strojnog učenja daje najtočnije prediktivne modele ishoda lokalnih izbora.

U zaključku rada daju se odgovori na rezultate testiranja hipoteza te na postavljena istraživačka pitanja. Na početku istraživanja postavljene su dvije hipoteze istraživanja:

H1:Točnost prediktivnih modela razvijenih na podacima aktivnosti korisnika društvene mreže Facebook veća je od modela temeljenih na anketama.
Temeljem rezultata usporedbe prediktivnih modela dobivenih na podacima društvene mreže i podacima anketem prikazanih u sekciji 4.4.2. **ova se hipoteza djelomično potvrđuje!** Naime, točnost prediktivnog modela temeljenog na neuronskoj mreži je veća od modela temeljenih na anketama, ali ta razlika nije statistički značajna.

H2: Pristup strojnom učenju temeljen na pogrešci daje točnije prediktivne modele od ostala tri pristupa strojnom učenju.
Temeljem rezultata usporedbe prediktivnih modela dobivenih primjenom četiri algoritma strojnog učenja prikazanih u sekciji 4.4.1. **ova se hipoteza potvrđuje!** Neuronska mreža dala je točnije prediktivne modele od preostala tri pristupa strojnom učenju. Temeljem ovih rezultata sugerira se primjena algoritma umjetnih neuronskih mreža na analizu podataka društvene mreže Facebook kada su podaci inherentno kontinuirani numerički. Ako je najvažniji kriterij odabira algoritma, objašnjivost modela, tada treba uzeti u obzir i stablo odlučivanja, koji daje malo manje točne

prediktivne modele, ali su rezultati takvih modela lako objašnjivi. Ovo su temeljne smjernice za odabir algoritama strojnog učenja za primjenu na podacima društvene mreže Facebook.

Na početku istraživanja postavljena su dva istraživačka pitanja:

IP1: Koje su varijable najznačajniji prediktori ishoda izbora?
Temeljem rezultata prikazanih u sekciji 4.5.1. Ukupan broj *likeova* stranice kandidata i Broj *likeova* poveznica najznačajniji su prediktori u ishoda izbora u sva četiri prediktivna modela ishoda izbora. Pošto su isti rezultati dobiveni u sva četiri modela, možemo tvrditi da su modeli konzistentni.

IP2: U kojoj je mjeri temporalna komponenata podataka društvene mreže Facebook važna u predikciji ishoda izbora?
Usporedbom prediktivnih modela dobivenih na podacima triju vremenskih perioda utvrđeno je da nema razlike u parametrima aktivnosti na društvenoj mreži Facebook unutar promatrana dva tjedna kampanje. Da bi se dao potpun odgovor na ovo pitanje trebalo bi uzeti u ozbir širi vremenski raspon i identificirati vrijeme od trenutka kreiranja stranice kandidata do trenutka početka kampanje te do vremena održavanja izbora. Broj *likeova* stranice, varijabla koja se pokazala najvažnijim prediktorom rezultata, eksponencijalno raste od dana kreiranja stranice kandidata, a u ovdje promatranom vremenu, u dva tjedna kampanje, se malo mijenja. Najveći broj pratitelja dobiva se na početku, a taj broj kroz kampanju kod većine kandidata eksponencijalno pada. To implicitno vodi i do dinamike interakcija na stranici kandidata.

Rezultati ovog istraživanja daju nekoliko **znanstvenih doprinosa:**

(i) sistematizirano je znanje o modelima temeljenim na podacima ispitivanja javnog mnijenja i društvenih mreža,

(ii) razvijeni i evaluirani su prediktivni modeli ishoda izbora temeljem podataka društvene mreže te su uspoređeni s podacima dobivenih anketiranjem,

(iii) utvrđena je prediktivna moć varijabli društvene mreže Facebook,

(iv) razvijene su smjernice za korištenje algoritama strojnog učenja na podacima društvenih medija.

U sklopu ovog istraživanja je prikupljen jedinstven skup podataka društvene mreže Facebook koji će biti javno dostupan te su temeljem tih podataka razvijeni modeli koji doprinose razumijevanju pojavu utjecaja društvenih mreža na ishode izbora.

Prilikom interpretacije rezultata treba biti oprezan i uzeti u obzir nekoliko ograničenja istraživanja, a koja s druge strane otvaraju smjernice za buduća istraživanja. Istraživanje je provedeno na uzorku jedne države, Republike Francuske, i na jednim provedenim izborima. Svaka država ima svoje specifičnosti koje treba uzeti u obzir prilikom provođenja istraživanja. U budućim istraživanjima će se dobiveni modeli testirati i u drugim okruženjima, i na drugim državama. Pitanje uključivanja jedne države vodi do problema malog uzorka. Naime, u skupu podataka je relativno mali broj instanci. Da bi se dobili kvalitetni modeli, u dizajnu istraživanja korištena je metoda k-unakrsne validacije, koja dobro radi s malim brojem instanci. U interpretaciji rezultata treba uzeti u obzir da se podaci odnose na specifično vrijeme, početak ožujka 2020., baš na samom početku širenja pandemije virusa COVID-19 i lockdowna koji je slijedio u drugoj polovici ožujka.

U analizi podataka korištena su četiri algoritma strojnog učenja, odabrana iz širokog broja dostupnih algoritama. Iako odabrani algoritmi predstavljaju različite pristupe načinu razvoja modela, postoji još niz algoritama koji bi se mogli primijeniti na podacima. U budućim istraživanjima prvenstveno će se testirati metode ansambla i usporediti s ovdje dobivenim rezultatima.

POPIS LITERATURE

Abreu, S. (2019). *Automated Architecture Design for Deep Neural Networks*. https://arxiv.org/abs/1908.10714v1

Ahmad, N., & Popa, I.-L. (2014). The Social Media Usage and the Transformation of Political Marketing and Campaigning of the Emerging Democracy in Indonesia. *Public Administration and Information Technology*, *13*, 97–125. https://doi.org/10.1007/978-3-319-04666-2_7

Almahadeen, L, Akkaya, M., & Sari, A. (2017). Mining student data using CRISP-DM model. *International Journal of Computer Science and Information Security*, *15*(2), 305–316.

Arauzo-Azofra, A., Aznarte, J.L., Benitez, J.M., Emprical study of feature selection

methods based on individual feature evaluation for classification problems, Expert

systems with applications, 38, 2011., str. 8170-8177.

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. *Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010*, *1*, 492–499. https://doi.org/10.1109/WI-IAT.2010.63

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature 2012 489:7415*, *489*(7415), 295–298. https://doi.org/10.1038/nature11421

Borah, P. (2016). Political Facebook use: Campaign strategies used in 2008 and 2012 presidential elections. *Http://Dx.Doi.Org/10.1080/19331681.2016.1163519*, *13*(4), 326–338. https://doi.org/10.1080/19331681.2016.1163519

Buettner, R. (2016). PERSONALITY AS A PREDICTOR OF BUSINESS SOCIAL MEDIA USAGE: AN EMPIRICAL INVESTIGATION OF XING USAGE PATTERNS. *PACIS*.

Cabena, P, Hadjinian, P, Stadler, R, Verhees, J and Zanasi, A, 1998, Discovering Data Mining: From Concepts to Implementation. Prentice Hall.

Cadenas, J.M., Garrido, C.M., Martinez, R., Feature subset selection Filter-Wrapper

based on low quality data, Expert systems with applications, 40, 2013, str. 6241-6252.

Cameron, M. P., Barrett, P., & Stewardson, B. (2015). Can Social Media Predict Election Results? Evidence From New Zealand. *Http://Dx.Doi.Org/10.1080/15377857.2014.959690*, *15*(4), 416–432. https://doi.org/10.1080/15377857.2014.959690

Chaffey, D. (2019). *Global social media statistics research summary [updated 2021]*. https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

Chang, K. C., Chiang, C. F., & Lin, M. J. (2021). Using Facebook data to predict the 2016 U.S. presidential election. *PloS One*, *16*(12). https://doi.org/10.1371/JOURNAL.PONE.0253560

Chen, C. Y., & Chang, S. L. (2017). User-orientated perspective of social media used by campaigns. *Telematics and Informatics*, *34*(3), 811–820. https://doi.org/10.1016/J.TELE.2016.05.016

Chomboon, K., Chujai, P., Teerarassamee, P., Kerdprasop, K., & Kerdprasop, N. (2015, March). An empirical study of distance metrics for k-nearest neighbor algorithm. In Proceedings of the 3rd international conference on industrial application engineering (Vol. 2).

Curtis, L., Edwards, C., Fraser, K. L., Gudelsky, S., Holmquist, J., Thornton, K., & Sweetser, K. D. (2010). Adoption of social media for public relations by nonprofit organizations. *Public Relations Review*, *36*(1), 90–92. https://doi.org/10.1016/J.PUBREV.2009.10.003

Cehovin, L., Bosnic, Z., Empirical evaluation of feature selection methods in classification, Intelligent data analysis, 14, 2010., str. 265-281.

Decastro-García, N., Muñoz Castañeda, Á. L., Escudero García, D., & Carriegos, M. V. (2019). Effect of the Sampling of a Dataset in the Hyperparameter Optimization Phase over the Efficiency of a Machine Learning Algorithm. *Complexity*, *2019*.

https://doi.org/10.1155/2019/6278908

Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, *7*, 1–30.

Diaz, G., Fokoue, A., Nannicini, G., & Samulowitz, H. (2017). An effective algorithm for hyperparameter optimization of neural networks. *IBM Journal of Research and Development*, *61*(4). https://arxiv.org/abs/1705.08520v1

Digrazia, J., Mckelvey, K., Bollen, J., & Rojas, F. (2013). More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior. *PLoS ONE, 8 (11)*. http://ssrn.com/abstract=2235423

Domhan, T., Springenberg, T., & Hutter, F. (n.d.). *Speeding up Automatic Hyperparameter Optimization of Deep Neural Networks by Extrapolation of Learning Curves*.

dos Santos Brito, K., Silva Filho, R. L. C., & Adeodato, P. J. L. (2021). A systematic review of predicting elections based on social media data: research challenges and future directions. IEEE Transactions on Computational Social Systems.

Effing, R., Hillegersberg, J. van, & Huibers, T. (2012). Measuring the effects of social media participation on political party communities. *Public Administration and Information Technology*, *1*, 201–217. https://doi.org/10.1007/978-1-4614-1448-3_13

Eyrich, N., Padman, M. L., & Sweetser, K. D. (2008). PR practitioners' use of social media tools and communication technology. *Public Relations Review*, *34*(4), 412–414. https://doi.org/10.1016/J.PUBREV.2008.09.010

Fagni, T., & Cresci, S. (2022). Fine-Grained Prediction of Political Leaning on Social Media with Unsupervised Deep Learning. Journal of Artificial Intelligence Research, 73, 633-672.

Franch, F. (2013). (Wisdom of the Crowds)2: 2010 UK Election Prediction with Social Media. *Http://Dx.Doi.Org/10.1080/19331681.2012.705080*, *10*(1), 57–71. https://doi.org/10.1080/19331681.2012.705080

Gayo-Avello, D. (2012a). *" I Wanted to Predict Elections with Twitter and all I got was this*

Lousy Paper" -- A Balanced Survey on Election Prediction using Twitter Data. https://arxiv.org/abs/1204.6441v1

Gayo-Avello, D. (2012b). No, You Cannot Predict Elections with Twitter. *IEEE Internet Computing*, *16*(06), 91–94. https://doi.org/10.1109/MIC.2012.137

Gayo-Avello, D., Metaxas, P., & Mustafaraj, E. (2011). Limits of Electoral Predictions Using Twitter. *In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence*.

Gerodimos, R., & Justinussen, J. (2014). Obama's 2012 Facebook Campaign: Political Communication in the Age of the *Like* Button. *Http://Dx.Doi.Org/10.1080/19331681.2014.982266*, *12*(2), 113–132. https://doi.org/10.1080/19331681.2014.982266

Greengard, S. (2009). The first internet president. *Communications of the ACM*, *52*(2), 16–18. https://doi.org/10.1145/1461928.1461935

Gulati,Girish; Williams, & Christine. (2013). Social Media and Campaign 2012. *Social Science Computer Review*, *31*(5), 577–588. https://doi.org/10.1177/0894439313489258

Guleria, A., Sharma, A., Bansal, D., & Sharma, G. (2016). The Impact of the Electronic Campaign by the Political Parties on the Voters. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.2750224

Guyon, I., Elisseeff, A., An introduction to variable and feature selection. Journal of

Machine Learning Research, 3, str. 1157–1182, 2003.

Haykin, S. (1998). Neural networks: a comprehensive foundation. Prentice Hall PTR.

Harshvardhan, G., Gourisaria, M. K., Pandey, M., & Rautaray, S. S. (2020). A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, *38*. https://doi.org/10.1016/J.COSREV.2020.100285

Hong, S., & Nadler, D. (2012). Which candidates do the public discuss online in an election campaign?: The use of social media by 2012 presidential candidates and its impact on

candidate salience. *Government Information Quarterly*, *29*(4), 455–461. https://doi.org/10.1016/J.GIQ.2012.06.004

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Housholder, E., & LaMarre, H. L. (2015). Political social media engagement: Comparing campaign goals with voter behavior. *Public Relations Review*, *41*(1), 138–140. https://doi.org/10.1016/J.PUBREV.2014.10.007

Hutter, F., Kotthof, L., & Vanschoren, J. (2019). *Automated Machine Learning - Methods, Systems, Challenges | Frank Hutter | Springer*. https://www.springer.com/gp/book/9783030053178

Jaidka, K., Ahmed, S., Skoric, M., & Hilbert, M. (2018). Predicting elections from social media: a three-country, three-method comparative study. *Https://Doi.Org/10.1080/01292986.2018.1453849*, *29*(3), 252–273. https://doi.org/10.1080/01292986.2018.1453849

Japkowicz, N., Shah, M., Evaluating learning algorithms: A classification perspective,

Cambridge University Press, New York, 2011.

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*, *23*(5), 544–559. https://doi.org/10.1108/INTR-06-2012-0114

Kelleher, J. D., & Namee, B. Mac. (2015). *[PDF] Fundamentals Of Machine Learning For Predictive Data Analytics: Algorithms, Worked Examples, And Case Studies (MIT Press)* (1st ed.).

Khairuddin, M. A., & Rao, A. (2017). Significance of *like*s: Analysing passive interactions on Facebook during campaigning. *PLOS ONE*, *12*(6), e0179435. https://doi.org/10.1371/JOURNAL.PONE.0179435

Klepac, G. (2015). Particle Swarm Optimization Algorithm as a Tool for Profiling from Predictive Data Mining Models. In S. Bhattacharyya & P. Dutta (Eds.), *Handbook of Research on Swarm Intelligence in Engineering* (pp. 406–434). Hershey, PA, USA: IGI Global. doi:10.4018/978-1-4666-8291-7.ch013

Klepac, G., Mrsic, L., & Kopal, R. (2016). Efficient Risk Profiling Using Bayesian Networks and Particle Swarm Optimization Algorithm. In D. Jakóbczak (Ed.), *Analyzing Risk through Probabilistic Modeling in Operations Research* (pp. 91–124). Hershey, PA, USA: IGI Global. doi:10.4018/978-1-4666-9458-3.ch004

Klepac, G. (2017). Customer Profiling in Complex Analytical Environments Using Swarm Intelligence Algorithms // Nature-Inspired Computing: Concepts, Methodologies, Tools, and Applications (3 Volumes) / Information Resources Management Association (USA) (ur.). USA: IGI-Global, str. 1391-1422

Koutsoukas, A., Monaghan, K. J., Li, X., & Huan, J. (2017). Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *Journal of Cheminformatics 2017 9:1*, *9*(1), 1–13. https://doi.org/10.1186/S13321-017-0226-Y

Kudeshia, C., Sikdar, P., & Mittal, A. (2016). Spreading love through fan page liking: A perspective on small scale entrepreneurs. *Computers in Human Behavior*, *54*, 257–270. https://doi.org/10.1016/J.CHB.2015.08.003

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. *Applied Predictive Modeling*, 1–600. https://doi.org/10.1007/978-1-4614-6849-3

Larsson, A. O., & Moe, H. (2011). Studying political microblogging: Twitter users in the 2010 Swedish election campaign: *Http://Dx.Doi.Org/10.1177/1461444811422894*, *14*(5), 729–747. https://doi.org/10.1177/1461444811422894

Lee, S., & Xenos, M. (2019). Social distraction? Social media use and political knowledge in two U.S. Presidential elections. *Computers in Human Behavior*, *90*, 18–25. https://doi.org/10.1016/J.CHB.2018.08.006

Lilleker, D. G., Koc-Michalska, K., Schweitzer, E. J., Jacunski, M., Jackson, N., & Vedel, T.

(2011). Informing, engaging, mobilizing or interacting: Searching for a European model of web campaigning: *Http://Dx.Doi.Org/10.1177/0267323111416182*, *26*(3), 195–213. https://doi.org/10.1177/0267323111416182

Loh, W. Y. (2011). Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(1), 14-23.

Mangold, W. G., & Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. *Business Horizons*, *52*(4), 357–365. https://doi.org/10.1016/J.BUSHOR.2009.03.002

Manias, D. M., Jammal, M., Hawilo, H., Shami, A., Heidari, P., Larabi, A., & Brunner, R. (2020). Machine Learning for Performance-Aware Virtual Network Function Placement. *2019 IEEE Global Communications Conference, GLOBECOM 2019 - Proceedings*. https://arxiv.org/abs/2001.07787v1

Metaxas, P. T., Mustafaraj, E., & Gayo-Avello, D. (2011). How (Not) to predict elections. *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, 165–171. https://doi.org/10.1109/PASSAT/SOCIALCOM.2011.98

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. https://doi.org/10.1016/J.DSS.2014.03.001

Ozaki, Y., Yano, M., & Onishi, M. (2017). Effective hyperparameter optimization using Nelder-Mead method in deep learning. *IPSJ Transactions on Computer Vision and Applications 2017 9:1*, *9*(1), 1–12. https://doi.org/10.1186/S41074-017-0030-7

Pedregosa FABIANPEDREGOSA, F., Michel, V., Grisel OLIVIERGRISEL, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., Duchesnay, andÉdouard, & Duchesnay EDOUARDDUCHESNAY, Fré. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. *Journal of Machine Learning Research*, *12*, 2825–2830. http://scikit-

learn.sourceforge.net.

Petrocchi, N., Asnaani, A., Martinez, A. P., Nadkarni, A., & Hofmann, S. G. (2015). Differences Between People Who Use Only Facebook and Those Who Use Facebook Plus Twitter. *Http://Dx.Doi.Org/10.1080/10447318.2014.986640*, *31*(2), 157–165. https://doi.org/10.1080/10447318.2014.986640

Petz B. Osnovne statističke metode za nematematičare. III. dopunjeno izdanje. Jastrebarsko: Naklada Slap, 1997.

Phillips, L., Dowling, C., Shaffer, K., Hodas, N., & Volkova, S. (2017). *Using Social Media to Predict the Future: A Systematic Literature Review*. https://arxiv.org/abs/1706.06134v1

Praude, V., & Skulme, R. (2015). Social Media Campaign Metrics in Latvia. *Procedia - Social and Behavioral Sciences*, *213*, 628–634. https://doi.org/10.1016/J.SBSPRO.2015.11.462

Rahmawati, I. (2014). *Social media, politics, and young adults : the impact of social media use on young adults' political efficacy, political knowledge, and political participation towards 2014 Indonesia general election*.

Rao, N. R. (2016). Social media listening and monitoring for business applications. *Social Media Listening and Monitoring for Business Applications*, 1–469. https://doi.org/10.4018/978-1-5225-0846-5

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 41–46.

Rumelhart, D. E., & McClelland, J. L. (1986). Parallel distributed processing, vol. 1: Foundations. Explorations in the Microstructure of Cognition.

Safavian, S. R., & Landgrebe, D. (1991). A Survey of Decision Tree Classifier Methodology. *IEEE Transactions on Systems, Man and Cybernetics*, *21*(3), 660–674. https://doi.org/10.1109/21.97458

Safiullah, M., Pathak, P., Singh, S., & Anshul, A. (2017). Social media as an upcoming tool for political marketing effectiveness. *Asia Pacific Management Review*, *22*(1), 10–15.

https://doi.org/10.1016/J.APMRV.2016.10.007

Sanders, S., & Giraud-Carrier, C. (2017). Informing the use of hyperparameter optimization through metalearning. *Proceedings - IEEE International Conference on Data Mining, ICDM*, *2017-November*, 1051–1056. https://doi.org/10.1109/ICDM.2017.137

Saritas, M. M., & Yasar, A. (2019). Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. *International Journal of Intelligent Systems and Applications in Engineering*, *7*(2), 88–91. https://doi.org/10.18201//IJISAE.2019252786

Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016). Social Media Analyses for Social Measurement. *Public Opinion Quarterly*, *80*(1), 180–211. https://doi.org/10.1093/POQ/NFV048

Schoenfeld, B. (2020). *Metalearning by Exploiting Granular Machine Learning Pipeline Metadata.pdf*. Brigham Young University.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/J.PROCS.2021.01.199

Singh, P., Dwivedi, Y. K., Kahlon, K. S., Pathania, A., & Sawhney, R. S. (2020). Can twitter analytics predict election outcome? An insight from 2017 Punjab assembly elections. *Government Information Quarterly*, *37*(2), 101444. https://doi.org/10.1016/J.GIQ.2019.101444

Soon, F. C., Khaw, H. Y., Chuah, J. H., & Kanesan, J. (2018). Hyper-parameters optimisation of deep CNN architecture for vehicle logo recognition. *IET Intelligent Transport Systems*, *12*(8), 939–946. https://doi.org/10.1049/IET-ITS.2018.5127

Stefko, R., Dorcak, P., & Pollak, F. (2011). Virtual Social Networks And Their Utilization For Promotion. *Advanced Logistic Systems*, *4*(1), 1–238. https://ideas.repec.org/a/pcz/alspcz/v4y2011i1p126-134.html

Strandberg, K. (2013). A social media revolution or just a case of history repeating itself? The use of social media in the 2011 Finnish parliamentary elections:

Http://Dx.Doi.Org/10.1177/1461444812470612, *15*(8), 1329–1347. https://doi.org/10.1177/1461444812470612

Strömbäck, J., & Esser, F. (2009). *Shaping politics: mediatization and media interventionism - Zurich Open Repository and Archive*. https://www.zora.uzh.ch/id/eprint/29325/

Sulzmann, J.-N., Fürnkranz, J., & Hüllermeier, E. (2007). On Pairwise Naive Bayes Classifiers. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4701 LNAI*, 371–381. https://doi.org/10.1007/978-3-540-74958-5_35

T.k., B., Annavarapu, C. S. R., & Bablani, A. (2021). Machine learning algorithms for social media analysis: A survey. *Computer Science Review*, *40*, 100395. https://doi.org/10.1016/J.COSREV.2021.100395

Taylor, M., & Kent, M. L. (2010). Anticipatory socialization in the use of social media in public relations: A content analysis of PRSA's Public Relations Tactics. *Public Relations Review*, *36*(3), 207–214. https://doi.org/10.1016/J.PUBREV.2010.04.012

The Internet's Role in Campaign 2008 | Pew Research Center. (n.d.). Retrieved August 9, 2021, from https://www.pewresearch.org/internet/2009/04/15/the-internets-role-in-campaign-2008/

Tjong, E., Sang, K., & Bos, J. (2012). *Predicting the 2011 Dutch Senate Election Results with Twitter*. 53–60. https://doi.org/10.5555/2389969

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018).

Vepsäläinen, T., Li, H., & Suomi, R. (2017). Facebook *like*s and public opinion: Predicting the 2015 Finnish parliamentary elections. *Government Information Quarterly*, *34*(3), 524–532. https://doi.org/10.1016/J.GIQ.2017.05.004

Vitak, J., Zube, P., Smock, A., Carr, C. T., Ellison, N., & Lampe, C. (2011). It's Complicated: Facebook Users' Political Participation in the 2008 Election. *Https://Home.Liebertpub.Com/Cyber*, *14*(3), 107–114. https://doi.org/10.1089/CYBER.2009.0226

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.*, 1–11.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. In *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier Inc. https://doi.org/10.1016/C2009-0-19715-5

Wright, D., Hinson, M., Wright, D. K., & Hinson, M. D. (2009). *An Analysis of the Increasing Impact of Social and Other New Media on Public Relations Practice*. www.instituteforpr.org

Yang, L., Moubayed, A., Hamieh, I., & Shami, A. (2019). Tree-based intelligent intrusion detection system in internet of vehicles. *2019 IEEE Global Communications Conference, GLOBECOM 2019 - Proceedings*. https://doi.org/10.1109/GLOBECOM38437.2019.9013892

Zhang, X. (2018). Social media popularity and election results: A study of the 2016 Taiwanese general election. *PLOS ONE*, *13*(11), e0208190. https://doi.org/10.1371/JOURNAL.PONE.0208190

Zuo, W., Wang, K., Zhang, H., & Zhang, D. (2007). Kernel Difference-Weighted k-Nearest Neighbors Classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4682 LNAI*, 861–870. https://doi.org/10.1007/978-3-540-74205-0_89

POPIS SLIKA

POPIS TABLICA

PRILOZI

Prilog 1.

Varijabla	**Varijabla**	**Koeficijent korelacije**	***P***
Broj dijeljenja događaja	Broj *likeova* događaja	0.962628963	8.2392E-123
Broj *likeova* videa	Broj *likeova* fotografija	0.958789601	6.3496E-119
Broj dijeljenja fotografija	Broj *likeova* fotografija	0.9265007	2.27583E-92
Broj komentara videa	Broj *likeova* videa	0.925558287	8.39899E-92
Broj komentara videa	Broj *likeova* fotografija	0.915366258	4.15397E-86
Broj komentara videa	Broj komentara fotografija	0.910702669	9.74571E-84
Broj komentara poveznica	Broj komentara fotografija	0.901577715	1.87014E-79
Broj komentara događaja	Broj *likeova* videa	0.895802202	5.90778E-77
Broj komentara statusa	Broj *likeova* statusa	0.894656623	1.77726E-76
Broj komentara fotografija	Broj *likeova* videa	0.887679082	1.11905E-73
Broj dijeljenja videa	Broj dijeljenja fotografija	0.886232376	4.03665E-73
Broj komentara događaja	Broj *likeova* događaja	0.877916723	4.6623E-70
Broj *likeova* videa	Broj *likeova* događaja	0.872333571	1.95041E-68
Broj komentara fotografija	Broj *likeova* fotografija	0.868703589	6.46693E-67
Broj komentara fotografija	Broj komentara događaja	0.866858182	2.57914E-66
Broj dijeljenja fotografija	Broj *likeova* videa	0.862585412	5.86974E-65
Broj dijeljenja poveznica	Broj *likeova* poveznica	0.854131175	2.1083E-62

Broj dijeljenja događaja	Broj komentara događaja	0.852326561	7.05507E-62
Broj dijeljenja videa	Broj *likeova* videa	0.845349694	6.49589E-60
Broj dijeljenja fotografija	Broj komentara videa	0.842305948	4.35818E-59
Broj dijeljenja statusa	Broj *likeova* statusa	0.832741144	1.33984E-56
Broj komentara videa	Broj komentara poveznica	0.825353247	8.7741E-55
Broj dijeljenja videa	Broj komentara videa	0.823340873	2.64927E-54
Broj komentara videa	Broj komentara događaja	0.817410539	6.34913E-53
Broj dijeljenja videa	Broj *likeova* fotografija	0.810515989	2.21238E-51
Broj dijeljenja događaja	Broj *likeova* videa	0.807424346	1.03757E-50
Broj komentara događaja	Broj *likeova* fotografija	0.806870474	1.3645E-50
Broj *likeova* fotografija	Broj *likeova* događaja	0.795820459	1.62815E-48
Broj komentara poveznica	Broj *likeova* poveznica	0.779930701	3.1218E-45
Broj komentara poveznica	Broj *likeova* fotografija	0.774402515	3.1702E-44
Broj dijeljenja videa	Broj *likeova* događaja	0.770945343	1.30682E-43
Broj komentara poveznica	Broj *likeova* videa	0.752188201	1.88243E-40
Broj dijeljenja poveznica	Broj komentara poveznica	0.752089186	1.95271E-40
Broj dijeljenja fotografija	Broj *likeova* događaja	0.74678594	1.35655E-39
Broj komentara videa	Broj *likeova* događaja	0.744937383	2.63606E-39
Broj dijeljenja fotografija	Broj komentara fotografija	0.735809962	6.4545E-38
Broj dijeljenja statusa	Broj komentara statusa	0.735371702	7.50069E-38

Broj dijeljenja događaja	Broj *likeova* fotografija	0.73149774	2.79371E-37
Broj dijeljenja videa	Broj komentara poveznica	0.709155369	3.59469E-34
Broj dijeljenja fotografija	Broj komentara poveznica	0.7076354	5.70838E-34
Broj dijeljenja videa	Broj dijeljenja poveznica	0.705705128	1.02266E-33
Broj dijeljenja videa	Broj komentara fotografija	0.697142848	1.28424E-32
Broj komentara poveznica	Broj komentara događaja	0.676066721	4.52299E-30
Broj dijeljenja fotografija	Broj dijeljenja događaja	0.669550623	2.515E-29
Broj komentara fotografija	Broj *likeova* događaja	0.666447537	5.60757E-29
Broj dijeljenja događaja	Broj komentara videa	0.664108745	1.01974E-28
Broj dijeljenja videa	Broj *likeova* poveznica	0.656313239	7.20296E-28
Broj dijeljenja fotografija	Broj komentara događaja	0.651004061	2.63936E-27
Broj dijeljenja videa	Broj dijeljenja događaja	0.64558144	9.6833E-27
Broj dijeljenja poveznica	Broj dijeljenja fotografija	0.640850975	2.94611E-26
Broj dijeljenja fotografija	Broj *likeova* poveznica	0.637691397	6.12804E-26
Broj dijeljenja videa	Broj komentara događaja	0.637085673	7.04494E-26
Broj dijeljenja događaja	Broj komentara fotografija	0.563883161	1.9155E-19
Broj *likeova* poveznica	Broj *likeova* fotografija	0.556333664	5.97788E-19

Broj komentara poveznica	Broj *likeova* događaja	0.551312177	1.70948E-18
Broj komentara videa	Broj *likeova* poveznica	0.520645509	2.46238E-16
Broj dijeljenja poveznica	Broj *likeova* fotografija	0.513185019	7.66034E-16
Broj dijeljenja poveznica	Broj komentara fotografija	0.507538373	1.77572E-15
Broj dijeljenja poveznica	Broj komentara videa	0.507350989	1.82547E-15
Broj komentara fotografija	Broj *likeova* poveznica	0.489112328	2.48348E-14
Broj dijeljenja poveznica	Broj *likeova* videa	0.473448006	2.07199E-13
Broj *likeova* videa	Broj *likeova* poveznica	0.466804203	4.3536E-13
Broj *likeova* poveznica	Broj videa	0.439206421	9.64024E-12
Broj dijeljenja poveznica	Broj *likeova* događaja	0.432068208	3.43533E-11
Broj dijeljenja poveznica	Broj videa	0.376741207	1.17407E-08
Broj *likeova* poveznica	Broj *likeova* događaja	0.37033178	1.47627E-08
Broj dijeljenja događaja	Broj komentara poveznica	0.367132276	2.91034E-08
Broj dijeljenja fotografija	Broj videa	0.365342096	3.43544E-08
Broj dijeljenja videa	Broj videa	0.364843942	3.59708E-08
Ukupan broj *likeova* stranice	Broj biraca	0.359012069	1.91679E-07
Broj komentara poveznica	Broj videa	0.354809222	8.93318E-08
Broj *likeova* fotografija	Broj videa	0.351869545	1.08273E-07
Broj poveznica	Broj fotografij	0.340959862	1.78661E-07
Broj komentara statusa	Broj statusa	0.339305912	3.42804E-07
Broj događaja	Rezultat	0.32525507	6.85338E-07
Broj poveznica	Broj događaja	0.319469364	1.10383E-06

Broj dijeljenja poveznica	Broj komentara događaja	0.31627838	2.21698E-06
Broj *likeova* događaja	Broj videa	0.31542007	1.90517E-06
Ukupan broj *likeova* stranice	Rezultat	0.311136099	7.71332E-06
Broj *likeova* videa	Broj videa	0.31034552	3.31949E-06
Broj komentara videa	Broj videa	0.307503951	4.33955E-06
Broj *likeova* statusa	Broj statusa	0.288325832	1.6722E-05
Izlaznost	Broj biraca	0.278451774	2.35682E-05
Broj komentara fotografija	Broj videa	0.257395662	0.000135266
Broj fotografija	Broj događaja	0.255224246	0.000112168
Broj komentara događaja	Broj *likeova* poveznica	0.251282256	0.000196904
Broj poveznica	Rezultat	0.242150468	0.000270587
Broj videa	Broj fotografija	0.238796467	0.000320494
Broj komentara poveznica	Broj događaja	0.235239062	0.000504961
Broj dijeljenja događaja	Broj videa	0.23245164	0.000590973
Broj videa	Broj događaja	0.231115783	0.000503062
Broj poveznica	Ukupan broj *likeova* stranice	0.229689981	0.001133152
Broj dijeljenja poveznica	Broj događaja	0.229531682	0.000695447
Broj *likeova* poveznica	Broj događaja	0.221606188	0.000934724
Broj komentara događaja	Broj videa	0.211724598	0.001797106
Broj dijeljenja poveznica	Broj dijeljenja događaja	0.207288847	0.002250728
Broj fotografij	Rezultat	0.202425507	0.002386473
Broj dijeljenja statusa	Broj statusa	0.198295292	0.003503925
Broj komentara fotografija	Broj događaja	0.186159878	0.006187053
Broj *likeova* fotografija	Broj događaja	0.16530617	0.015009337

Broj dijeljenja videa	Broj događaja	0.156943885	0.021328481
Broj dijeljenja događaja	Broj *likeova* poveznica	0.155317292	0.022728739
Broj videa	Broj poveznica	0.153610291	0.02175405
Broj *likeova* videa	Broj događaja	0.150246307	0.027251523
Broj komentara videa	Broj događaja	0.150070187	0.027799264
Broj komentara događaja	Broj događaja	0.147313192	0.030831749
Broj dijeljenja poveznica	Broj poveznica	0.142404034	0.036930994
Broj komentara događaja	Izlaznost	0.138012843	0.04322259
Broj događaja	Ukupan broj *likeova* stranice	0.131555947	0.063998234
Broj *likeova* događaja	Broj događaja	0.131453974	0.05151941
Broj *likeova* poveznica	Broj poveznica	0.13133884	0.052266743
Broj dijeljenja fotografija	Broj događaja	0.129935409	0.057146024
Broj dijeljenja događaja	Izlaznost	0.123663217	0.070351549
Ukupan broj *likeova* stranice	Izlaznost	0.118536687	0.095411863
Broj dijeljenja statusa	Broj dijeljenja videa	0.11430966	0.094559419
Broj komentara poveznica	Ukupan broj *likeova* stranice	0.111154159	0.125807751
Broj komentara fotografija	Ukupan broj *likeova* stranice	0.107842897	0.137551997
Broj dijeljenja videa	Ukupan broj *likeova* stranice	0.107458328	0.138968599
Broj komentara videa	Izlaznost	0.10735649	0.116527311
Broj komentara videa	Ukupan broj *likeova* stranice	0.104313762	0.150973219
Broj *likeova* događaja	Izlaznost	0.101314036	0.134126359
Broj *likeova* videa	Broj biraca	0.100029711	0.142845192

Broj *likeova* videa	Ukupan broj *likeova* stranice	0.099841311	0.169375605
Broj *likeova* fotografija	Ukupan broj *likeova* stranice	0.098365418	0.17580056
Broj *likeova* događaja	Broj biraca	0.096920345	0.151930203
Broj dijeljenja fotografija	Broj komentara statusa	0.095543064	0.162725037
Broj komentara događaja	Broj biraca	0.094874808	0.165699045
Broj komentara fotografija	Izlaznost	0.093832421	0.170418873
Broj *likeova* poveznica	Ukupan broj *likeova* stranice	0.092748365	0.197184328
Broj dijeljenja događaja	Broj biraca	0.092345123	0.177325365
Broj *likeova* događaja	Ukupan broj *likeova* stranice	0.090413528	0.208748175
Broj dijeljenja videa	Broj komentara statusa	0.089954166	0.188858568
Broj *likeova* fotografija	Broj biraca	0.0898063	0.188551066
Broj *likeova* videa	Izlaznost	0.089750356	0.188827655
Broj videa	Ukupan broj *likeova* stranice	0.087832136	0.218524165
Broj komentara fotografija	Broj biraca	0.086704962	0.20539888
Broj dijeljenja statusa	Broj dijeljenja fotografija	0.086172034	0.208208641
Broj komentara videa	Broj biraca	0.084421061	0.217634868
Broj dijeljenja fotografija	Ukupan broj *likeova* stranice	0.082928716	0.254061375
Broj komentara događaja	Ukupan broj *likeova* stranice	0.082369805	0.257286963
Broj dijeljenja događaja	Broj događaja	0.080848086	0.237806034

Broj *likeova* poveznica	Rezultat	0.079409696	0.241888306
Broj *likeova* poveznica	Broj fotografij	0.078987213	0.243323713
Broj dijeljenja fotografija	Broj *likeova* statusa	0.076399427	0.264706493
Broj dijeljenja događaja	Ukupan broj *likeova* stranice	0.072591037	0.318298815
Broj dijeljenja fotografija	Broj biraca	0.072416734	0.290495072
Broj videa	Broj biraca	0.071851574	0.285374269
Broj dijeljenja videa	Broj *likeova* statusa	0.070130014	0.306036478
Broj dijeljenja poveznica	Ukupan broj *likeova* stranice	0.067970544	0.350156221
Broj komentara poveznica	Broj biraca	0.066622713	0.330917923
Broj dijeljenja videa	Broj biraca	0.066498987	0.331818749
Broj *likeova* fotografija	Izlaznost	0.063549052	0.352631576
Broj komentara poveznica	Izlaznost	0.057902065	0.398238942
Broj komentara statusa	Broj komentara videa	0.045790023	0.504229404
Broj *likeova* poveznica	Broj biraca	0.041260782	0.542677404
Broj dijeljenja poveznica	Broj fotografija	0.039659441	0.563021668
Broj *likeova* statusa	Rezultat	0.0371747	0.587751059
Broj događaja	Broj biraca	0.033427398	0.618740935
Broj dijeljenja statusa	Broj komentara videa	0.031762634	0.643268678
Broj *likeova* događaja	Broj poveznica	0.026073813	0.701190246
Broj dijeljenja videa	Izlaznost	0.022128199	0.746989775
Broj komentara poveznica	Broj poveznica	0.019492969	0.776269524
Broj dijeljenja fotografija	Izlaznost	0.018746375	0.784624596

Broj komentara statusa	Ukupan broj *likeova* stranice	0.01417944	0.845637846
Broj *likeova* poveznica	Izlaznost	0.011633426	0.863772118
Broj komentara videa	Broj *likeova* statusa	0.011482938	0.867057768
Broj komentara statusa	Broj *likeova* fotografija	0.007296308	0.915294648
Broj dijeljenja poveznica	Broj biraca	0.006391666	0.925764673
Broj događaja	Izlaznost	0.005897713	0.930054861
Broj dijeljenja događaja	Broj poveznica	0.003943471	0.954158231
Broj komentara statusa	Broj komentara fotografija	0.003768681	0.956188049
Broj komentara statusa	Broj fotografija	0.0011751	0.986332974
Broj dijeljenja fotografija	Broj fotografija	-0.000930213	0.989180945
Broj dijeljenja fotografija	Broj poveznica	-0.001168854	0.986405612
Broj dijeljenja videa	Broj poveznica	-0.004049926	0.952922132
Broj dijeljenja videa	Broj fotografija	-0.004896368	0.943097921
Broj biraca	Rezultat	-0.00759319	0.910224878
Broj videa	Rezultat	-0.007600523	0.910341687
Broj *likeova* statusa	Ukupan broj *likeova* stranice	-0.008289053	0.909390826
Broj *likeova* statusa	Broj fotografija	-0.008605898	0.899930794
Broj statusa	Broj poveznica	-0.015932447	0.812967045
Broj dijeljenja poveznica	Rezultat	-0.016399521	0.811481485
Broj komentara statusa	Rezultat	-0.017570249	0.798302253
Broj dijeljenja statusa	Broj komentara fotografija	-0.019839628	0.772398749
Broj *likeova* fotografija	Broj poveznica	-0.020234288	0.767469277
Broj dijeljenja statusa	Broj *likeova* videa	-0.020326809	0.766968418

Broj komentara poveznica	Broj fotografij	-0.020545648	0.764532827
Izlaznost	Rezultat	-0.021209995	0.752770605
Broj poveznica	Izlaznost	-0.021300571	0.751749275
Broj *likeova* fotografija	Broj fotografij	-0.021515199	0.753210305
Broj *likeova* statusa	Broj *likeova* fotografija	-0.022164227	0.746016821
Broj dijeljenja statusa	Broj *likeova* fotografija	-0.02238691	0.744133956
Broj statusa	Rezultat	-0.024784899	0.71342629
Broj dijeljenja statusa	Broj biraca	-0.031521999	0.645786819
Broj komentara fotografija	Broj *likeova* statusa	-0.032397599	0.63664359
Broj komentara događaja	Broj poveznica	-0.032587436	0.634668425
Broj fotografija	Ukupan broj *likeova* stranice	-0.032788568	0.645696037
Broj komentara statusa	Broj *likeova* videa	-0.033938055	0.620691041
Broj *likeova* statusa	Broj poveznica	-0.03459537	0.613104368
Broj dijeljenja statusa	Broj fotografij	-0.035158851	0.608172955
Broj komentara statusa	Broj događaja	-0.03662219	0.593316962
Broj statusa	Broj događaja	-0.042065121	0.532028377
Broj dijeljenja statusa	Ukupan broj *likeova* stranice	-0.043767267	0.547711272
Broj *likeova* videa	Broj poveznica	-0.044487905	0.51545699
Broj dijeljenja poveznica	Broj komentara statusa	-0.048035994	0.483524896
Broj fotografija	Broj biraca	-0.048896328	0.466518945
Broj dijeljenja fotografija	Rezultat	-0.049533674	0.471024073
Broj *likeova* videa	Broj fotografija	-0.050468145	0.460579029
Broj *likeova* statusa	Broj biraca	-0.052035395	0.446753177
Broj komentara poveznica	Rezultat	-0.055293759	0.420960295

Broj *likeova* fotografija	Rezultat	-0.056720835	0.407948759
Broj *likeova* statusa	Broj *likeova* videa	-0.05735138	0.401643628
Broj poveznica	Broj biraca	-0.057702065	0.391142197
Broj komentara fotografija	Broj poveznica	-0.058027072	0.397219524
Broj dijeljenja statusa	Broj videa	-0.058571977	0.392794177
Broj komentara statusa	Izlaznost	-0.061451616	0.369904343
Broj komentara fotografija	Broj fotografija	-0.061891591	0.366480987
Broj *likeova* događaja	Broj fotografija	-0.062256099	0.358075374
Broj komentara fotografija	Broj statusa	-0.063316735	0.355527541
Broj komentara videa	Broj fotografija	-0.063396646	0.354919496
Broj komentara statusa	Broj biraca	-0.064557155	0.346162644
Broj *likeova* statusa	Broj događaja	-0.064699176	0.343967923
Broj dijeljenja događaja	Broj fotografija	-0.067098668	0.327467249
Broj komentara događaja	Broj statusa	-0.067408852	0.325230942
Broj komentara videa	Broj poveznica	-0.06778984	0.322497704
Broj statusa	Broj fotografija	-0.067860663	0.313049367
Broj dijeljenja poveznica	Broj *likeova* statusa	-0.067902872	0.321689669
Broj dijeljenja statusa	Rezultat	-0.072139331	0.29349015
Broj *likeova* statusa	Izlaznost	-0.076040912	0.265839141
Broj dijeljenja događaja	Broj statusa	-0.077312785	0.259020336
Broj komentara videa	Broj statusa	-0.080976611	0.237058452
Broj dijeljenja statusa	Broj poveznica	-0.08203348	0.23097353
Broj statusa	Izlaznost	-0.082178853	0.221570042
Broj komentara događaja	Broj fotografija	-0.082336969	0.229246745
Broj dijeljenja fotografija	Broj statusa	-0.085281371	0.212966063
Broj komentara statusa	Broj *likeova* poveznica	-0.085459286	0.212009565

Broj komentara poveznica	Broj statusa	-0.090804788	0.184694104
Broj komentara statusa	Broj komentara poveznica	-0.091474736	0.181461982
Broj videa	Izlaznost	-0.092035973	0.170819666
Broj dijeljenja statusa	Broj dijeljenja događaja	-0.092138688	0.178300106
Broj dijeljenja statusa	Izlaznost	-0.093695172	0.171047701
Broj dijeljenja statusa	Broj događaja	-0.09370691	0.170993854
Broj *likeova* fotografija	Broj statusa	-0.094939047	0.164419744
Broj dijeljenja videa	Broj statusa	-0.096122122	0.160180488
Broj dijeljenja statusa	Broj komentara događaja	-0.096378238	0.159064612
Broj dijeljenja videa	Rezultat	-0.097133513	0.1567846
Broj fotografija	Izlaznost	-0.09715671	0.147224722
Broj dijeljenja poveznica	Izlaznost	-0.099500421	0.145927391
Broj dijeljenja događaja	Broj komentara statusa	-0.100361531	0.142453361
Broj komentara statusa	Broj komentara događaja	-0.103795148	0.12922524
Broj komentara statusa	Broj poveznica	-0.10477245	0.12563893
Broj *likeova* statusa	Broj *likeova* poveznica	-0.105103079	0.123560649
Broj komentara videa	Rezultat	-0.105227766	0.124882779
Broj *likeova* događaja	Broj statusa	-0.105994511	0.117815721
Broj *likeova* poveznica	Broj statusa	-0.107119213	0.113946848
Broj dijeljenja statusa	Broj komentara poveznica	-0.108686179	0.112043955
Broj dijeljenja poveznica	Broj statusa	-0.109505189	0.109350427
Broj dijeljenja statusa	Broj dijeljenja poveznica	-0.111329887	0.103532303
Broj *likeova* videa	Broj statusa	-0.113049629	0.097487271

Broj statusa	Ukupan broj *likeova* stranice	-0.114682684	0.107654931
Broj *likeova* statusa	Broj videa	-0.114798437	0.092382044
Broj komentara fotografija	Rezultat	-0.116692438	0.088589084
Broj komentara poveznica	Broj *likeova* statusa	-0.118290495	0.083550172
Broj komentara statusa	Broj videa	-0.118384358	0.083303593
Broj dijeljenja statusa	Broj *likeova* događaja	-0.118833749	0.082131141
Broj dijeljenja događaja	Broj *likeova* statusa	-0.12216603	0.073846279
Broj *likeova* videa	Rezultat	-0.123391954	0.070974541
Broj komentara događaja	Broj *likeova* statusa	-0.128442464	0.060087324
Broj dijeljenja statusa	Broj *likeova* poveznica	-0.129659924	0.057679636
Broj komentara statusa	Broj *likeova* događaja	-0.132017765	0.053243039
Broj *likeova* statusa	Broj *likeova* događaja	-0.15346745	0.024083537
Broj dijeljenja događaja	Rezultat	-0.17286461	0.011306547
Broj *likeova* događaja	Rezultat	-0.180927518	0.0072671
Broj statusa	Broj videa	-0.181277884	0.006640227
Broj komentara događaja	Rezultat	-0.187082833	0.006050438
Broj statusa	Broj biraca	-0.193548089	0.003713416

ŽIVOTOPIS AUTORA

Alen Kišić je rođen 1977. godine u Varaždinu gdje i živi. Djetinjstvo, školske i studentske dane proveo je u Svetom Đurđu i Ludbregu. Osnovnu školu je završio u Ludbregu, gimnaziju u Varaždinu (dva razreda) i Koprivnici. Diplomirao je na dva fakulteta u Zagrebu – ekonomiju poduzetništva na Veleučilištu Vern (bacc.oec.), te novinarstvo na Fakultetu političkih znanosti (dipl.nov.). Poslijediplomski specijalistički studij završio je 2012. godine na Sveučilištu u Zagrebu, Fakultetu organizacije i informatike u Varaždinu, smjer Menadžment poslovnih sustava, stekavši titulu sveučilišnog specijalista ekonomije (univ.spec.oec.). Na istom je fakultetu upisao 2016. godine i doktorat iz informacijskih znanosti.

Poslovnu karijeru započeo je u području odnosa s javnošću – bio je, među ostalim, pomoćnik direktora Sektora za korporativne komunikacije u Podravki d.d., te direktor Korporativnih komunikacija za jugoistočnu Europu u trgovačkom lancu METRO Cash&Carry. Četiri godine proveo je kao izabrani zamjenik župana Varaždinske županije, nakon čega je radio kao direktor Centra kompetencije za obnovljive izvore energije d.o.o. u Varaždinu. Trenutačno obnaša dužnost direktora Zone Sjever d.o.o. u Trnovcu kraj Varaždina. Završio je brojna stručna usavršavanja poput programa Londonske škole za odnose za odnose s javnošću (LSPR) i programa Političke akademije njemačke zaklade Friedrich Ebert Stiftung i Zaklade Novo društvo. Govori engleski jezik. Autor je 7 znanstvenih radova objavljenih u međunarodnim znanstvenim časopisima i konferencijama.

POPIS OBJAVLJENIH RADOVA

Kišić, Alen

Information and Communications Technologies as a Driver of Effective Internal Communication. // OPEN JOURNAL FOR INFORMATION TECHNOLOGY (OJIT), 3 (2020), 2; 39-52 doi:10.32591/coas.ojit.0302.01039k

Kišić, Alen

Pregled primjene blockchain tehnologije: perspektiva organizacije i menadžmenta. // Zbornik radova Međimurskog veleučilišta u Čakovcu, 9 (2018), 1; 41-45

Kišić, Alen

SWOT I TOWS ANALIZA DRUŠTVENE MREŽE FACEBOOK KAO ALATA ZA ODNOSE S JAVNOŠĆU U POLITIČKIM KAMPANJAMA. // South Eastern European journal of communication, 2 (2020), 2; 95-102

Kišić, Alen; Kliček, Božidar

Machine learning based prediction of Croatian 2017. local elections. // MIPRO 2021, 44th International Convention Proceedings / Skala, Karolj (ur.).
Opatija, Hrvatska: Croatian Society for Information, Communication and Electronic Technology – MIPRO, 2021. str. 1577-1581

Kišić, Alen

The Use of Social Media in Political Campaigns: The Case of Croatian Local Elections 2017. // Proceedings of the Central European Conference on Information and Intelligent Systems / Strahonja, V. ; Kirinić, V. (ur.).
Varaždin: Faculty of Organization and Informatics, 2018. str. 133-139

Huđek, Miroslav; **Kišić, Alen**; Kelemen, Robert

Criteria of excellence in primary and secondary education at the level of regional self-government. // EDULEARN15 Proceedings / Gomez Chova, L ; Lopez Martinez, A. ; Candel Torres, I. (ur.). Barcelona: IATED Academy, 2015. str. 1814-1823

Kišić, Alen; Huđek, Miroslav

Centres of excellence in the county of Varaždin. // ICERI Proceedings. In 7th International Conference of Education, Research and Innovation Seville, SPAIN: IATED-INT ASSOC TECHNOLOGY EDUCATION A& DEVELOPMENT, 2014. str. 289-298