Usporedba algoritama rudarenja podataka u društvenim znanostima

Faletar, Dominik

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Dominik Faletar

USPOREDBA ALGORITAMA RUDARENJA PODATAKA U DRUŠTVENIM ZNANOSTIMA

ZAVRŠNI RAD

Varaždin, 2024.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Dominik Faletar

JMBAG: 0016139082

Studij: Informacijski sustavi

USPOREDBA ALGORITAMA RUDARENJA PODATAKA U DRUŠTVENIM ZNANOSTIMA

ZAVRŠNI RAD

Mentorica:

Izv. prof. dr. sc. Dijana Oreški

Varaždin, rujan 2024.

Dominik Faletar

Izjava o izvornosti

Izjavljujem da je moj završni rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autor potvrdio prihvaćanjem odredbi u sustavu FOI-radovi

Sažetak

Ovaj rad istražuje primjenu i usporedbu dvaju algoritama strojnog učenja – stabla odlučivanja i neuronskih mreža u analizi stresa u društvenim mrežama. Istraživanje je provedeno na skupu podataka s platforme Reddit, koju obuhvaća tekstualne objave o različitim temama povezanim sa stresnim situacijama. U radu je korišten alat BigML, gdje su analizirani različiti pristupi obrezivanju stabla odlučivanja te konfiguracija neuronskih mreža kako bi se postigla optimalna predikcija.

Metodološki, analiza se oslanjala na testiranje modela sa smanjenim brojem atributa i optimizacijom neurona, čime se poboljšala točnost predikcije i smanjivala složenost modela. Cilj istraživanja bio je utvrditi koji algoritam pruža bolju ravnotežu točnosti i interpretabilnosti rezultata. Zaključci ukazuju na to da je stablo odlučivanja efikasnije u pogledu interpretabilnosti i brzine primjene, dok su neuronske mreže bolje u prepoznavanju složenih zadataka. Oba modela imaju svoje specifične prednosti, a odabir ovisi o kontekstu analize.

Ključne riječi: rudarenje podataka; stablo odlučivanja; neuronske mreže, analiza, stres, društvene mreže

Sadržaj

1. Uvod

Rudarenje podataka, također poznato kao otkrivanje znanja u bazama podataka (KDD), predstavlja primjenu strojnog učenja i statističke analize za otkrivanje vrijednih informacija iz velikih skupova podataka. S razvojem tehnologija kao što su strojno učenje i skladištenje podataka, rudarenje podataka postalo je ključan alat za analizu i predviđanje u različitim industrijama. Danas, u eri velikih podataka, gdje je sve veća količina informacija dostupna iz različitih izvora, uključujući društvene mreže, zahtijeva efikasne algoritme za analizu i automatizaciju procesa, čime se ubrzava donošenje informiranih odluka[1].

Glavni cilj rada je usporediti performanse dvaju popularnih algoritama rudarenja podataka, stabla odlučivanja i neuronskih mreža, primijenjenih na skupu podataka o analizi stresa na društvenim mrežama. Analiza podataka će se provesti korištenjem alata za strojno učenje - BigML, koji je jednostavno dostupan i prilagođen svim korisnicama.

Stablo odlučivanja predstavlja popularan i često korišten alat u strojnom učenju (machine learning), rudarenju podataka (data mining) i statistici. Koristi se za donošenje odluka ili previđanje. Sastoji se od čvorova za odluke ili testove na atributima, grana koje prikazuju ishode odluka, te listove koji predstavljaju konačne ishode ili predikcije[2].

Neuronske mreže su računalni modeli inspirirani funkcijama ljudskog mozga. Sastoje se od povezanih čvorova, odnosno neurona, koji obrađuju podatke i uče iz njih, omogućujući zadatke poput prepoznavanja uzoraka i donošenja odluka u kontekstu strojnog učenja[3].

Struktura rada sastoji se od nekoliko dijelova. U prvom dijelu pruža se teorijski pregled rudarenja podataka i algoritama koji se koristi u analizi. Drugi dio posvećen je detaljnom opisu skupa podataka i pripremi za analizu. Treći dio sadrži usporedbu algoritama, a četvrti dio donosi evaluaciju i interpretaciju dobivenih rezultata. Konačno, u zaključku rada iznose se ključni nalazi.

2. Rudarenje podatka

Rudarenje podataka uključuje pretraživanje i analizu velikih skupova podataka s ciljem otkrivanja uzoraka i dobivanje vrijednih uvida. U današnje vrijeme mnoge tvrtke koriste softvere za rudarenje podataka kako bi dobili što više informacija o svojim kupcima. Takva strategija im pomaže da maksimiziraju svoje prihode i smanje troškove.

Rudarenje podataka funkcionira na način da uključuje istraživanje i analizu velikih količina informacija kako bi se otkrili značajni obrasci i trendovi. Koristi se u upravljanju kreditnim rizikom, otkrivanju prijevarom i filtriranju neželjene pošte. Također je alat za istraživanje tržišta koji pomaže otkriti mišljenja ili stavove određene skupine ljudi. Proces rudarenja podataka sastoji se od četiri koraka:

1. Podaci se prikupljaju i učitavaju u skladišta podataka na licu mjesta ili na cloud-u
2. Poslovni analitičari, upravljački timovi i IT stručnjaci pristupaju podacima i određuju kako ih žele organizirati
3. Prilagođeni softver za aplikacije sortira i organizira podatke
4. Krajnji korisnik predstavlja podatke u formatu koji je jednostavan za dijeljenje, poput grafikona ili tablica [4]

Primjerice, društvene mreže poput Facebooka, Instagrama i TikToka koriste rudarenje podataka kako bi analizirali interakcije korisnika, prepoznali trendove i mišljenja o određenim temama. Na temelju tih analiza, društvene mreže mogu prilagoditi sadržaj koji korisnicima prikazuju, uključujući reklame i preporuke za praćenje određenih profila. Rudarenje podacima je doslovno svuda okolo nas.

2.1. CRISP DM

CRISP-DM (križnoindustrijski standardni proces za rudarenje podataka) je jedan od najpopularnijih standarda za rudarenje podataka. Razvijen od 1996. do 1999. godine uz doprinos više od 300 organizacija, CRISP-DM pruža strukturirani okvir za rudarenje podataka kroz šest faza: razumijevanje domene, razumijevanje podataka, priprema podataka, modeliranje, vrednovanje i korištenje. Ovaj standard je neovisan o industriji i alatu, te se temelji na iskustvu i fokusira na rješavanje poslovnih problema, što ga čini pouzdanim i ponovljivim[6].

1) Razumijevanje domene – definiranje poslovnih ciljeva, procjena situacije, određivanje ciljeva rudarenja podataka i izrada plana projekta
2) Razumijevanje podataka – prikupljanje, opisivanje, istraživanje i provjera kvalitete podataka
3) Priprema podataka – odabir, čišćenje, konstrukcija, integracija i formatiranje podataka za modeliranje
4) Modeliranje – odabir tehnika, generiranje testnih dizajna, izrada i procjena modela
5) Vrednovanje – ocjena rezultata, pregled procesa i određivanje daljnjih koraka
6) Korištenje – planiranje implementacije modela, praćenje, održavanje, završni izvještaj i pregled projekta[7].

2.2. Tehnike rudarenja podataka

Rudarenje podataka također koristi algoritme i tehnike za pretvaranje velikih količina podataka u korisne informacije. Stabla odlučivanja i neuronske mreže su među najpopularnijim tehnikama rudarenja podataka. Ostale najpopularnije tehnike su:

- Pravila asocijacije – Traže veze između varijabli, poput analiza košarice koja otkriva koje proizvode kupci često kupuju zajedno, npr. kruh i mlijeko
- Klasifikacija – Dodjeljuje objekte unaprijed definiranim klasama na temelju zajedničkih karakteristika, npr. banke klasificiraju korisnike na skupine niskog rizika, srednjeg rizika i visokog rizika na temelju podataka o povijesti plaćanja, prihoda i kreditnoj povijesti
- Klasterizacija – Grupira slične objekte na temelju njihovih razlika od drugih skupina, npr. umjesto klasifikacije proizvoda pojedinačno (npr. „hlače“, „majica“), klasterizacija stvara skupine poput „modna odjeća“
- K-najbliži susjed (KNN) – Klasificira podatke na temelju njihove blizine s drugim podacima, npr. ako kupac kupi određeni proizvod, KNN analizira najbliže susjede u podacima (kupce sličnog profila) i preporuča proizvode koje su oni kupili.
- Prediktivna analiza – Koristi povijesne podatke za predviđanje budućih ishoda, npr. tvrtka za proizvodnju koristi prediktivnu analizu kako bi predvidjela buduću potražnju za proizvodima. Na temelju podataka o prošlim prodajama, sezonskim trendovima i ekonomskim uvjetima, model predviđa koliko će proizvoda biti potrebno u sljedećem kvartalu[4].

2.3. Procesi rudarenja podataka

Kako bi analiza podataka bila što učinkovitija, analitičari obično slijede definirani slijed koraka tijekom procesa rudarenja podataka. Bez ove strukture, analitičar bi se mogao suočiti s problemima tijekom analize koje je mogao izbjeći pravilnom pripremom. Proces rudarenja podataka obično se sastoji od sljedećih faza:

1) **Razumijevanje poslovanja**

Prije nego što započne rad s podacima, važno je razumjeti poslovni kontekst i ciljeve projekta. Koje ciljeve tvrtka želi postići rudarenjem podataka? Kakva je trenutna poslovna situacija? Što je SWOT analiza pokazala? Proces rudarenja podataka započinje razumijevanjem što će na kraju definirati uspjeh.

2) **Razumijevanje podataka**

Nakon što je poslovni problem jasno definiran, potrebno je razmisliti o dostupnim podacima. Koji izvori podataka su dostupni? Kako će podaci biti osigurani i pohranjeni? Kako će se prikupljati? Ova faza također uključuje procjenu ograničenja vezanih uz podatke, kao što su pohrana, sigurnost i prikupljanje, te kako ta ograničenja mogu utjecati na proces rudarenja podataka.

3) **Priprema podataka**

Podaci se prikupljaju, učitavaju, izvlače ili izračunavaju, zatim se čiste i standardiziraju. U ovoj fazi se također uklanjaju izvanredne vrijednosti, ispravljaju greške i procjenjuje razumnost podataka. Provjerava se i veličina skupa podataka, jer preveliki set može usporiti izračune i analizu.

4) **Izgradnja modela**

S pripremljenim i očišćenim podacima, analitičari započinju s analizom podataka kako bi identificirali odnose, trendove, asocijacije ili sekvencijalne uzorke. Podaci se mogu unijeti u prediktivne modele kako bi se procijenilo kako prethodni podaci mogu utjecati na buduće ishode.

5) **Procjena rezultata**

Nakon izgradnje modela, potrebno je procijeniti rezultate analize. Nalazi se interpretiraju i prezentiraju donositeljima odluka, koji do sada nisu bili izravno uključeni u proces rudarenja podataka. U ovoj fazi organizacija može odlučiti kako će koristiti rezultate analize za donošenje poslovnih odluka.

6) **Implementacija i praćenje promjena**

Posljednja faza procesa rudarenja podataka uključuje donošenje poslovnih odluka na temelju rezultata analize. Tvrtka može odlučiti da su nalazi dovoljno relevantni za strateško usmjeravanje ili može odlučiti da informacije nisu dovoljno jake. U svakom slučaju, uprava prati utjecaj tih odluka na poslovanje i ponovno pokreće proces rudarenja podataka identificiranjem novih poslovnih izazova ili prilika[4].

2.4. Prednosti rudarenja podataka

Razumijevanje rudarenja podataka je ključno, kao i spoznaja njegovih prednosti i primjena u industriji. U današnjem svijetu orijentiranom na podatke, rudarenje podataka omogućuje učinkovito rješavanje problema. Prednosti uključuju:

- Pomaže tvrtkama u prikupljanju pouzdanih informacija i donošenju odluka
- Ekonomičnije je i učinkovitije od drugih aplikacija za obradu podataka
- Omogućuje profitabilne prilagodbe u poslovanju
- Otkriva kreditne rizike i prijevare
- Omogućuje brzu analizu velikih količina podataka i automatizirano predviđanje trendova te otkrivanje skrivenih uzoraka[5]

2.5. Budućnost rudarenja podataka

S rastom količine podataka, budućnost rudarenja podataka izgleda obećavajuće. Tehnologije poput strojnog učenja, umjetne inteligencije i dubokog učenja ubrzavaju procese analize podataka u oblačnim okruženjima. Internet stvari i nosive tehnologije omogućuju prikupljanje ogromnih količina podataka o pojedincima i organizacijama. Rješenja za analitiku temeljena na oblaku omogućuju tvrtkama brži i isplativiji pristup podacima i obradi, što im pomaže u donošenju boljih poslovnih odluka[5].

2.6. Stabla odlučivanja

Stablo odlučivanje je specifična vrsta dijagrama toka koja se koristi za vizualizaciju procesa donošenja odluka mapiranjem različitih akcija i njihovih mogućih ishoda. Stabla odlučivanja koriste se u raznim područjima, od financijske zaštite i zdravstvene zaštite do marketinga i računalnih znanosti. Posebno su korisna kod složenih odluka s više opcija i neizvjesnim ishodima. Stablo odlučivanja pomaže nam korak po korak kako doći do odluke postavljanjem manjih pitanja, jedno po jedno, poput kontrolnog popisa koji osigurava da ništa ne propustimo[8].

Stabla odlučivanja obično se sastoje od tri glavna različita elementa:

I. **Korijenski čvor** – Najgornji čvor koji predstavlja krajnji cilj ili glavnu odluku koju pokušavamo donijeti
II. **Grane** – Grane koje proizlaze iz korijena predstavljaju različite opcije ili radnje koje su dostupne tijekom donošenja određene odluke. Najčešće su označene strelicama i često uključuju pridružene troškove te vjerojatnost ostvarenja.
III. **Listovi** - Listovi, koji se nalaze na kraju grana, predstavljaju moguće ishode za svaku akciju. Obično postoje dvije vrste listova: kvadratni listovi, koji označavaju još jednu odluku, i kružni listovi, koji označavaju slučajni događaj ili nepoznati ishod[8].
IV.

Algoritam stabla odlučivanja zapravo radi na način kroz par jednostavnih koraka:

1) **Početak u korijenu:** Algoritam počinje na vrhu, zvanom "korijenski čvor," koji predstavlja cijeli skup podataka.
2) **Postavljanje najboljih pitanja:** Algoritam traži najvažniju značajku ili pitanje koje najbolje dijeli podatke u najrazličitije grupe, poput pitanja na raskrižju stabla.
3) **Grananje:** Na temelju odgovora, podaci se dijele u manje podskupove, stvarajući nove grane. Svaka grana predstavlja mogući put kroz stablo.
4) **Ponavljanje procesa:** Algoritam nastavlja postavljati pitanja i dijeliti podatke sve dok ne dođe do završnih "listova," koji predstavljaju predviđene ishode ili klasifikacije[9].

2.6.1 Pretpostavke kod stabla odlučivanja

Kod izrade stabla odlučivanja koriste se određene pretpostavke koje vode njegovu konstrukciju i utječu na performanse:

- **Binarne podjele**: Svaki čvor dijeli podatke na dva podskupa temeljem jedne značajke.
- **Rekurzivno particioniranje**: Podjela podataka u sve manje podskupove dok se ne zadovolji kriterij zaustavljanja.
- **Nezavisnost značajki**: Pretpostavlja se da su značajke nezavisne, iako to nije uvijek slučaj.
- **Homogenost**: Stabla teže stvaranju homogenih podskupova unutar čvorova.
- **Prevelika složenost i preprilagođavanje**: Stabla mogu biti sklona preprilagođavanju, stoga se koriste tehnike kao što su rezidba[9].

2.6.2. Prednosti

U današnje vrijeme strojevi postaju sve pametniji, a algoritmi stabla odlučivanja koriste se za donošenje važnih odluka, poput odobravanja kredita. Ovi algoritmi imaju brojne prednosti: jednostavni su za razumijevanje, mogu raditi s različitim vrstama podataka i nedostajućim vrijednostima, te su pouzdani i učinkoviti. Stabla odlučivanja omogućuju vizualizaciju procesa, olakšavajući komunikaciju s poslovnim subjektima. Osim toga, prilagođavaju se različitim situacijama, mogu generirati više izlaza i relativno su jeftina za implementaciju[10].

2.6.3. Primjena

Stabla odlučivanja, moćan su i popularan alat. Često ih koriste analitičari podataka za prediktivnu analizu, poput razvoja poslovnih strategija. Također su popularan alat u strojnome učenju i umjetnoj inteligenciji, gdje se koriste kao trenirajući algoritmi za nadzirano učenje. Zbog svoje fleksibilnosti, stabla odlučivanja primjenjuju se u širokom rasponu industrija, od tehnologije i zdravstva do financijskog planiranja. Primjeri uključuju procjenu mogućnosti širenja tehnološke tvrtke ili predviđanje kreditnog rizika u bankama[11].

2.7. Neuronske mreže

Neuronska mreža je niz algoritama koji nastoje prepoznati skrivene odnose u skupu podataka kroz proces koji oponaša rad ljudskog mozga. Neuronske mreže mogu se prilagoditi promjenama ulaznih podataka, čime generiraju najbolje moguće rezultate bez potrebe za redizajniranjem kriterija izlaza. Koncept neuronskih mreža, s korijenima u umjetnoj inteligenciji, sve više dobiva na popularnosti u razvoju trgovačkih sustava. Neuronska mreža ima tri glavne komponente: ulazni sloj, sloj za obradu i izlazni sloj. Ulazi mogu biti ponderirani prema različitim kriterijima, a sloj za obradu, koji je skriven, sadrži čvorove i veze između tih čvorova, slično neuronima i sinapsama u mozgu[12].

Postoji nekoliko vrsta neuronskih mreža:

- **Prosljeđujuće Neuronske Mreže**: Jednostavnije mreže koje prenose informacije u jednom smjeru, od ulaznih čvorova do izlaza. Često se koriste u tehnologijama prepoznavanja lica.
- **Rekurentne Neuronske Mreže**: Složenije mreže koje vraćaju izlaz natrag u mrežu, omogućujući "učenje" i poboljšanje. Koriste se u aplikacijama poput pretvaranja teksta u govor.
- **Konvolucijske Neuronske Mreže**: Imaju više slojeva za sortiranje podataka, često korištene za prepoznavanje slika.
- **Dekonvolucijske Neuronske Mreže**: Obrnuto od konvolucijskih mreža, koriste se za analizu slika.
- **Modularne Neuronske Mreže**: Sastoje se od nekoliko mreža koje rade neovisno kako bi omogućile složene računske procese[12].

2.7.1 Prednosti i mane neuronskih mreža

Neke od prednosti neuronskih mreža je to što mogu raditi kontinuirano i učinkovitije su od ljudi ili jednostavnijih analitičkih modela. Mogu učiti iz prethodnih rezultata kako bi predvidjele buduće ishode. Također, korištenje mreža u oblaku smanjuje rizike povezane s lokalnom hardverskom opremom, a mreže mogu istovremeno obavljati više zadataka. Primjena neuronskih mreža se proširila na medicinu, znanost, financije i druge sektore. Dok s druge strane za razvoj neuronskih mreža potreban je složen hardver, što nosi fizičke rizike i zahtijeva održavanje. Njihova složenost može produžiti vrijeme razvoja specifičnih algoritama, a također ih je teško revidirati zbog složenih procesa i nedostatka transparentnosti[12].

2.7.2. Primjena

Neuronske mreže široko se koriste u financijskim operacijama, planiranju poduzeća, trgovanju, poslovnoj analitici i održavanju proizvoda. Također su našle primjenu u poslovnim aplikacijama poput predviđanja i istraživanja tržišta, otkrivanja prijevara i procjene rizika. U financijama, neuronske mreže mogu obraditi stotine tisuća transakcijskih podataka, što omogućuje bolje razumijevanje volumena trgovanja, raspona trgovanja, korelacije među imovinom ili postavljanja očekivanja volatilnosti. Neuronske mreže prepoznaju suptilne nelinearne odnose i obrasce koje druge metode tehničke analize ne mogu otkriti[12].

3. Opis problema i skupa podataka

Živimo u vremenu kada svakodnevno primamo tisuće informacija, koje nas dovode do nepotrebnog stresa. Stres jako negativno može utjecati na naše zdravlje te uzrokovati glavobolje, visoki krvni tlak, srčane probleme, dijabetes, depresiju, anksioznost i još mnoge druge bolesti.

U današnje vrijeme, društvene mreže postale su glavni izvor komunikacije i interakcije među ljudima, što ih čini važnim izvorom podataka za istraživanje različitih društvenih fenomena. Jedan od tih fenomena je stres, koji može biti uzrokovan različitim čimbenicima i manifestira se kroz različite oblike ponašanja i izraza u tekstualnim objavama na društvenim mrežama. Razumijevanje i analiza stresa na društvenim mrežama može pružiti vrijedan uvid u psihološko stanje pojedinca te pomoći u identificiranju i rješavanju problema vezanih uz mentalno zdravlje[13].

Za potrebe ovog rada koristi se skup podataka analiza stresa na društvenim mrežama, koji sadrži tekstualne objave prikupljene s društvene mreže Reddit. Sastoji se od 190.000 objava iz pet različitih kategorija Reddit zajednice, uključujući teme vezane uz interpersonalne konflikte, mentalne bolesti, financijske probleme, posttraumtski stresni poremećaj (PTSP) i društvene odnose. Od tih objava, 3.553 segmenata su dodatno anotirani pomoću Amazon Mechanical Turk usluge, gdje su označeni kao stresni ili nestresni prema ocjeni više anotatora[13].

Podaci iz skupa omogućuju analizu stresa kroz dulje tekstualne objave. Osim toga, skup podataka je raznovrstan, jer pokriva širok spektar tema i stilova izražavanja stresa, što omogućuje sveobuhvatniju analizu[13].

Cilj je istražiti i usporediti performanse stabla odlučivanja i neuronskih mreža u predviđanju stresa na temelju ovog skupa podataka.

3.1. Opis i priprema podataka

Ovaj rad odnositi će se na skup podataka pod nazivom „Analiza stresa na društvenim mrežama“, koji se odnosi na podatke prikupljene 2019. godine. Skup podataka preuzet je s platforme Kaggle.com, koji pruža raznovrsne i kvalitetne skupove podataka za istraživačke i analitičke svrhe. Kako bih proveo analizu, skup podataka sam učitao pomoću .csv datoteke u alat za strojno učenje BigML.

Skup podataka sadrži ukupno 715 instanci, od kojih se svaka sastoji od 116 atributa. Od tih atributa, 3 su kategoričke, 111 numeričke, dok su preostala 3 tekstualna. Dva atributa nisu preferirana za analizu zbog njihove neprikladnosti ili irelevantnosti u kontekstu ovog istraživanja.

Tablica 1. Ključni atributi skupa podataka s objašnjenjima njihovih uloga i značenja (vlastita izrada)

Atribut	Tip podataka	Opis
id	Numerički	Jedinstveni identifikator za svaku instancu u skupu podataka.
subreddit	Kategorički	Naziv subreddita iz kojeg je objava preuzeta.
post_id	Kategorički	Jedinstveni identifikator za svaku objavu na Redditu.
sentence_range	Tekstualni	Raspon rečenica koje čine segment objave.
text	Tekstualni	Sadržaj objave u tekstualnom obliku.
label	Kategorički	Oznaka koja pokazuje je li segment objave klasificiran kao "stresan" ili "nestresan".
confidence	Numerički	

		Razina pouzdanosti u klasifikaciju segmenta objave.
lex_liwc_Tone	Numerički	Mjeri emocionalni ton objave.
lex_liwc_Clout	Numerički	Mjeri autoritativnost izraza u tekstu.
lex_liwc_negemo	Numerički	Označava prisutnost negativnih emocija u tekstu, poput tuge, ljutnje ili straha.
sentiment	Numerički	Sentiment analiza objave, koji kvantificira pozitivnost ili negativnost izraza.

U Tablici 1 prikazani su odabrani ključni atributi skupa podataka korištenih za analizu. Ovi atributi su identificirani kao značajni zbog njihove važnosti u prepoznavanju i analizi stresa u tekstualnim objavama na društvenim mrežama. Svaki od tih atributa pruža specifične informacije koje su bitne za razumijevanje kako se stres manifestira kroz tekst i kako se može mjeriti i predvidjeti korištenjem algoritama strojnog učenja[13].

Numerički atributi, poput „confidence“, ključni su za ocjenjivanje pouzdanosti klasifikacija koje je model napravio tijekom analize. „Confidence“ omogućuje procjenu koliko je model siguran u svoje predikcije, što je važno za optimizaciju modela i evaluaciju njegove točnosti[13].

LIWC (eng. Linguistic Inquiry and Word Count) atributi, poput „lex_liwc_Tone“, „lex_liwc_Clout“ i „lex_liwc_negemo“, mjere različite emocionalne aspekte teksta. „Tone“ procjenjuje opći emocionalni ton, pomažući u razlikovanju između pozitivnih i negativnih emocija. „Clout“ mjeri autoritativnost i samopouzdanje u izražavanju, što može pružiti uvid u to koliko autor čvrsto stoji iza svog sadržaja. S druge strane, „negemo“ označava prisutnost negativnih emocija, poput tuge, ljutnje ili straha, koje su obično povezane s povećanim stupnjem stresa. Ovi atributi su ključni za prepoznavanje i interpretaciju emocionalnih signala u podacima[13].

Kategorički atributi, kao što su „subreddit“ i „label“, omogućuju kategorizaciju konteksta u kojem je tekst napisan. Atribut „subreddit“ označava izvor objave, dok „label“ prikazuje je li

objava klasificirana kao "stresna" ili "nestresna", čime služi kao ciljna varijabla za modele nadziranog učenja[13].

Tekstualni atributi, poput „text“ i „sentence_range“, daju izravan uvid u sadržaj i strukturu objave. Atribut „text“ sadrži sirove tekstualne podatke koji se koriste za ekstrakciju značajki, dok „sentence_range“ označava duljinu i složenost objave. Ovi atributi ključni su za zadatke obrade prirodnog jezika, koji imaju za cilj izdvojiti značenje i prepoznati obrasce povezane sa stresom[13].

Ovi atributi zajedno pružaju sveobuhvatan skup podataka koji omogućuje dubinsku analizu stresa u objavama na društvenim mrežama. Korištenjem ovog raznovrsnog skupa informacija, prediktivni modeli mogu se učinkovito trenirati i evaluirati, što vodi do točnijih i značajnijih uvida u to kako se stres izražava i prepoznaje na društvenim mrežama[13].

4. Usporedba stabla odlučivanja i neuronskih mreža

Ovo poglavlje predstavlja ključni dio rada u kojem se analizira i uspoređuje učinkovitost dva popularna algoritma rudarenja podataka – stablo odlučivanja i neuronskih mreža. S obzirom na složenost analize stresa iz tekstualnih podataka prikupljenih s društvenih mreža, cilj ove usporedbe je pružiti uvid u to kako različiti algoritmi pristupaju problemu klasifikacije, te kako se njihova učinkovitost razlikuje.

U ovoj usporedbi, oba algoritma primijenjena su na istom skupu podataka, koji sadrži tekstualne objave s društvenih mreža klasificirane kao stresne ili nestresne. Kroz evaluaciju rezultata analizirane su njihove performanse korištenjem standardnih metrika, čime se jasno prikazuje koja metoda daje bolje rezultate u prepoznavanju stresa.

Analizom dobivenih rezultata pružit će se dublje razumijevanje razlika u pristupu ovih algoritama, kao i njihove praktične primjene u kontekstu društvenih znanosti. Također, rezultati ove usporedbe omogućit će preporuke o tome koji algoritam je prikladniji za zadatak analize stresa na društvenim mrežama, uzimajući u obzir kako točnost tako i interpretabilnost modela.

4.1. Analiza skupa podataka pomoću stabla odlučivanja

Zadatak započinje odabirom opcije „Model“, gdje prvo biramo atribut koji želimo analizirati. U ovom slučaju, odabrani atribut je „label“, koji nam omogućuje da prepoznamo je li objava stresna ili nije. Nakon toga, prelazimo na opciju „Tree“ u BigML alatu, gdje se odlučujemo za vrstu obrezivanja stabla.

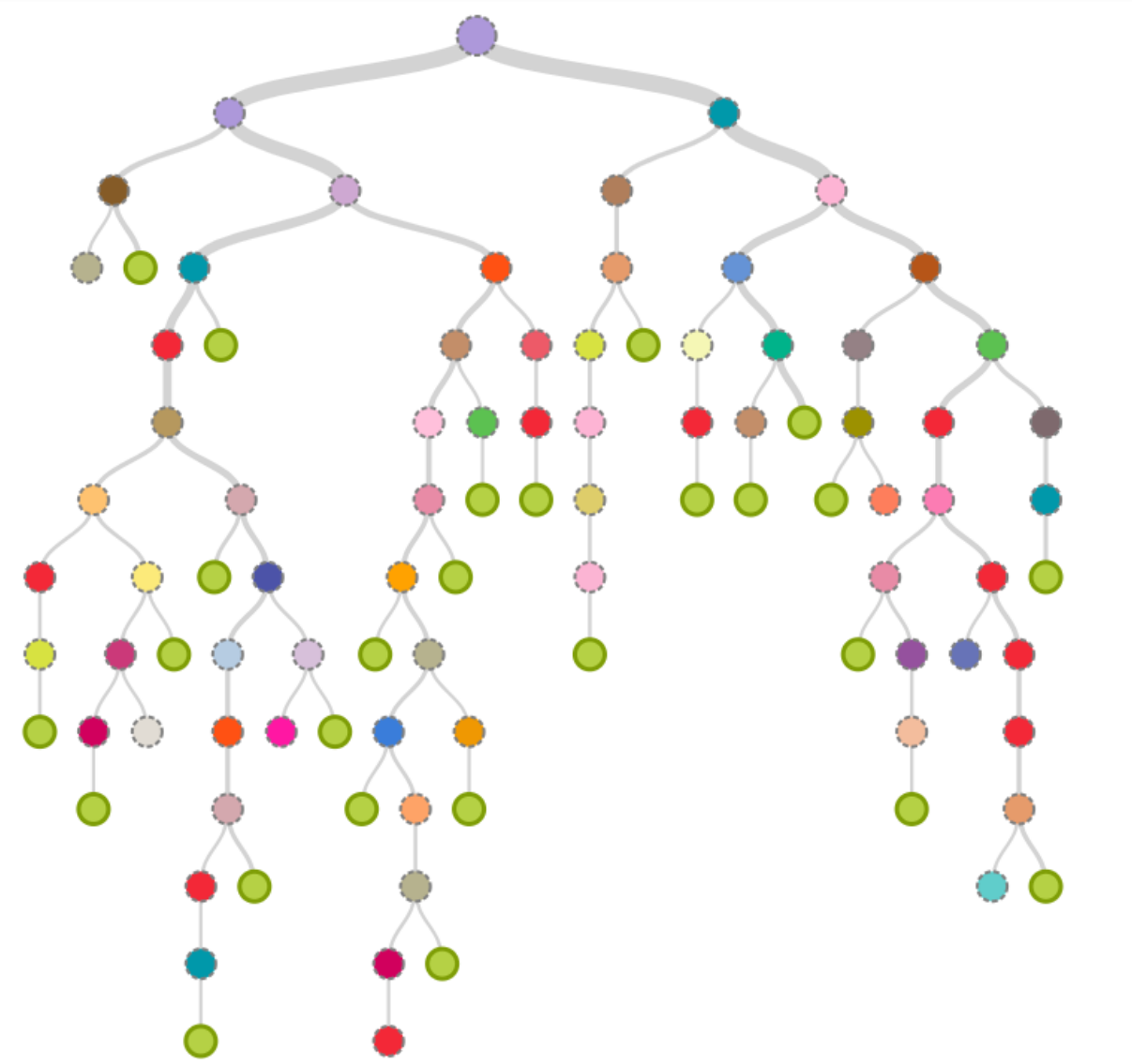
Obrezivanje (eng. pruning) je izrazito važno pri izradi prediktivnih modela jer omogućuje razvijanje modela koji je točan, pouzdan i dovoljno kvalitetan, ali bez prekomjerne složenosti. Jedna od najvećih prednosti stabla odlučivanja je to što razvija modele koji su jednostavni za interpretaciju. Kako bi zadržali jednostavnost, ključno je optimizirati veličinu stabla, a to se postiže pomoću obrezivanja[6].

Alat BigML nudi tri vrste obrezivanja, a to su: pametno obrezivanje (eng. smart pruning), aktivno statističko obrezivanje (eng. active statistical pruning) i nestatističko obrezivanje (eng. no statistical pruning). Pametno obrezivanje uklanja grane stabla koje ne doprinose značajno točnosti modela, aktivno statističko obrezivanje koristi statističke procjene za donošenje odluka o obrezivanju, dok nestatističko uopće ne koristi obrezivanje, dopuštajući modelu da zadrži sve grane, bez obzira na njihov doprinos[6].

Kako bismo dobili što bolji model, potrebno je isprobati sve tri vrste obrezivanja te analizirati rezultate svakog pristupa. U nastavku rada prikazat će se modeli razvijeni primjenom

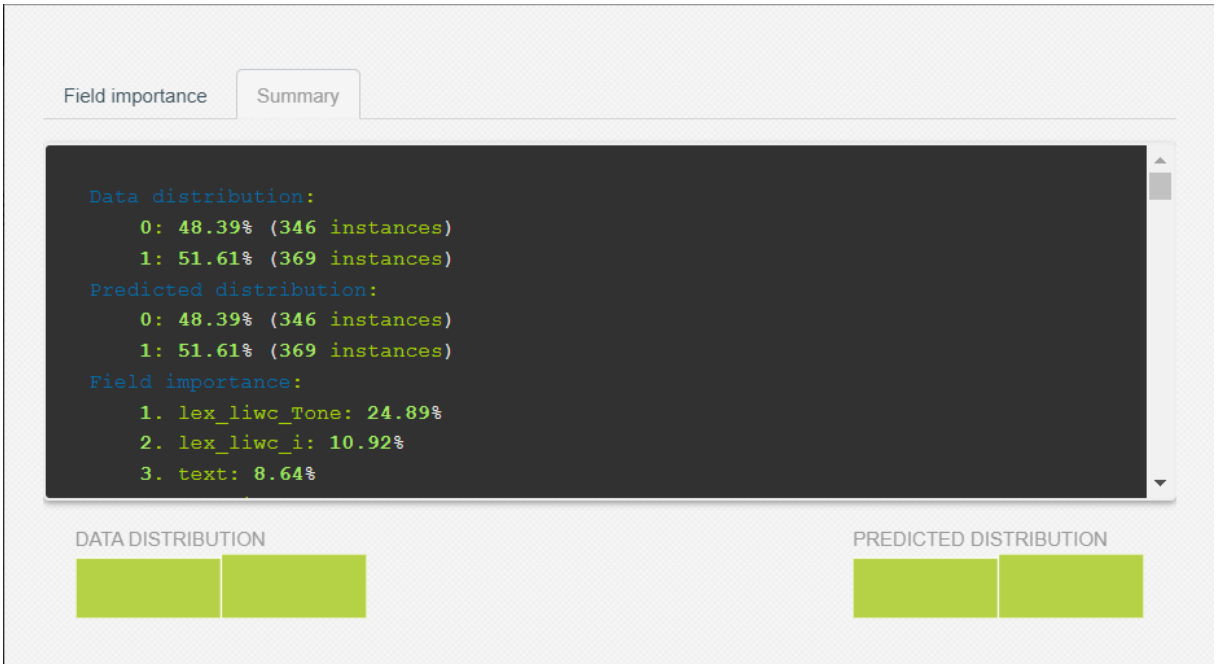
svake od ovih metoda obrezivanja. Ovakav pristup omogućuje izgradnju modela koji su točni, ali jednostavni, izbjegavajući prekomjerno grananje i složenost koja može otežati interpretaciju.

4.1.1. Pametno obrezivanje



Slika 1. Stablo odlučivanja - pametno obrezivanje (vlastita izrada)

Na slici 1 prikazan je prediktivni model stabla odlučivanja, razvijen korištenjem pametnog obrezivanja.



Slika 2. Točnost pametnog obrezivanja (vlastita izrada)

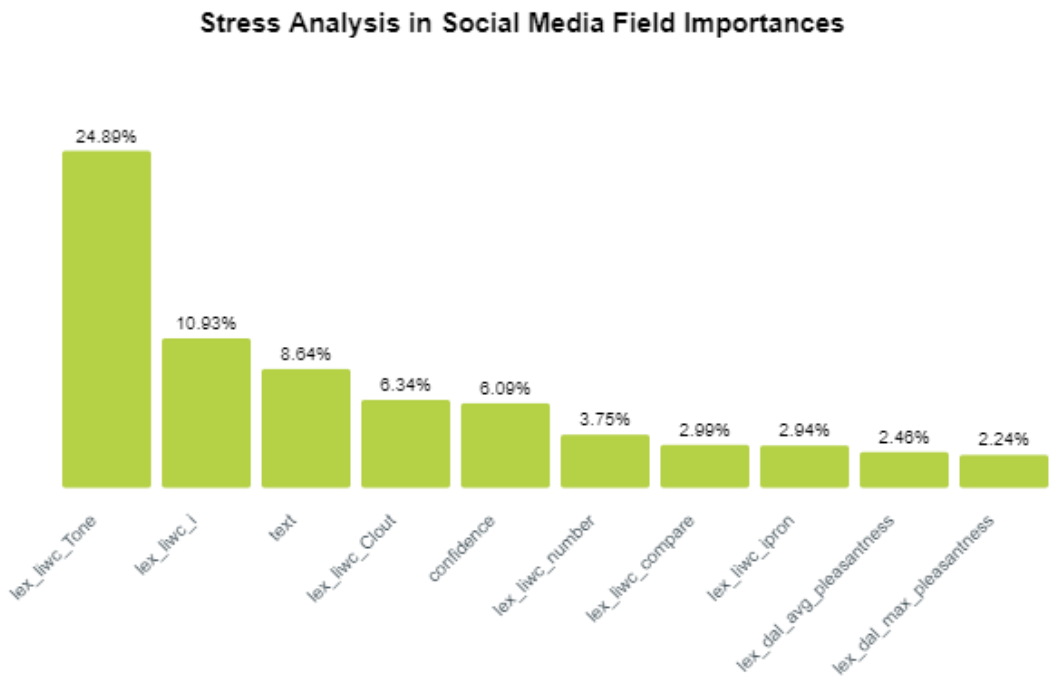
Slika 2 prikazuje izvještaj o distribuciji podataka u modelu. Skup podataka sadrži dvije klase: nestresan i stresan sadržaj.

Distribucija podataka je sljedeća:

● Nestresno (oznaka 0): 48.39% podataka, odnosno 346 instanci, označeno je kao nestresno.

● Stresno (oznaka 1): 51.61% podataka, odnosno 369 instanci, klasificirano je kao stresno.

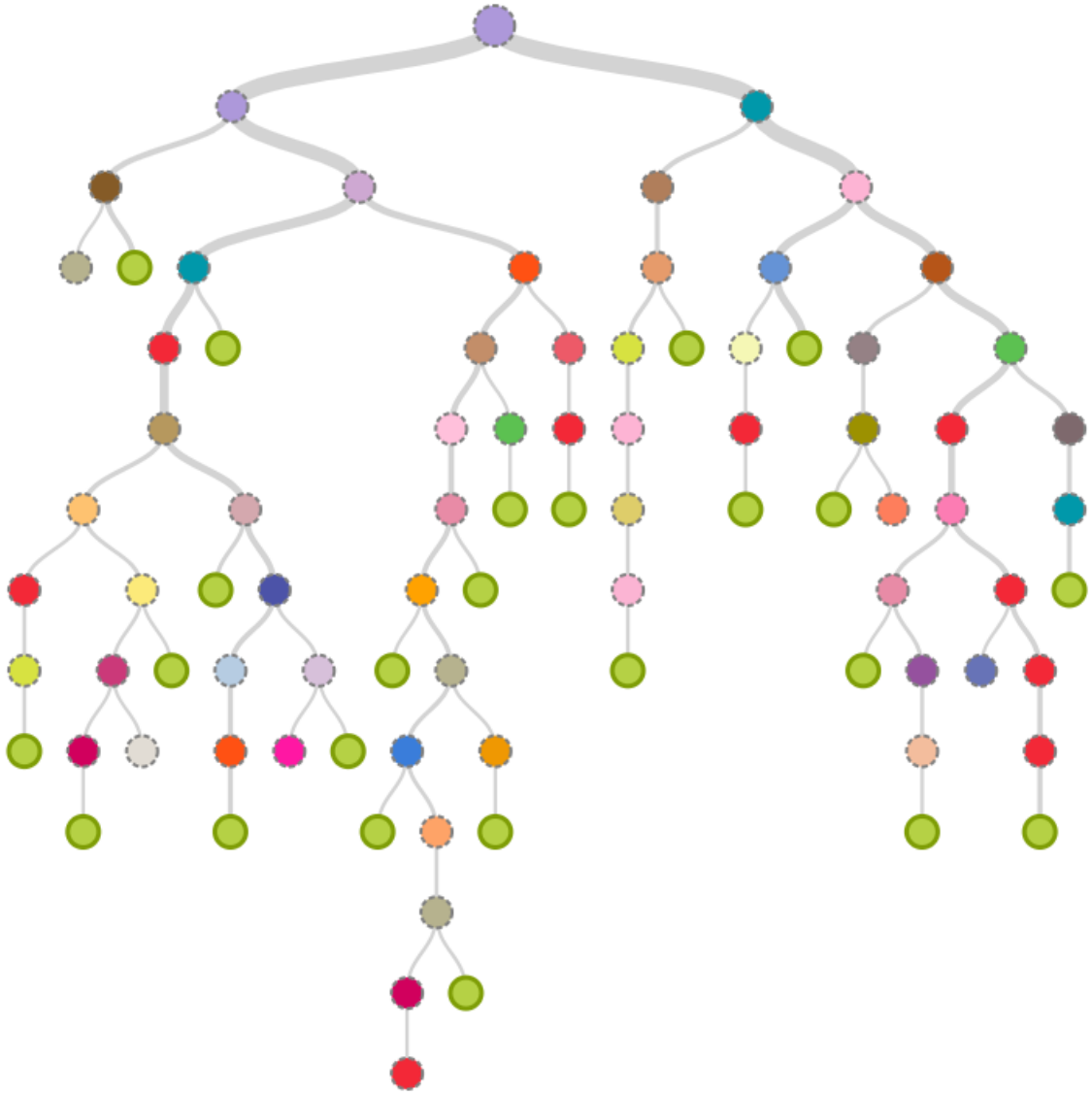
Ova raspodjela ukazuje na relativno izjednačenu podjelu između stresnog i nestresnog sadržaja, s blagom prevagom stresnih instanci. Što se tiče predviđene distribucije, model je postigao rezultate identične stvarnoj distribuciji podataka. Ovaj rezultat pokazuje da je model balansiran te uspješno prati distribuciju podataka, što ukazuje na dobru usklađenost modela sa stvarnim stanjem u skupu podataka.



Slika 3. Prikaz važnosti atributa - pametno obrezivanje (vlastita izrada)

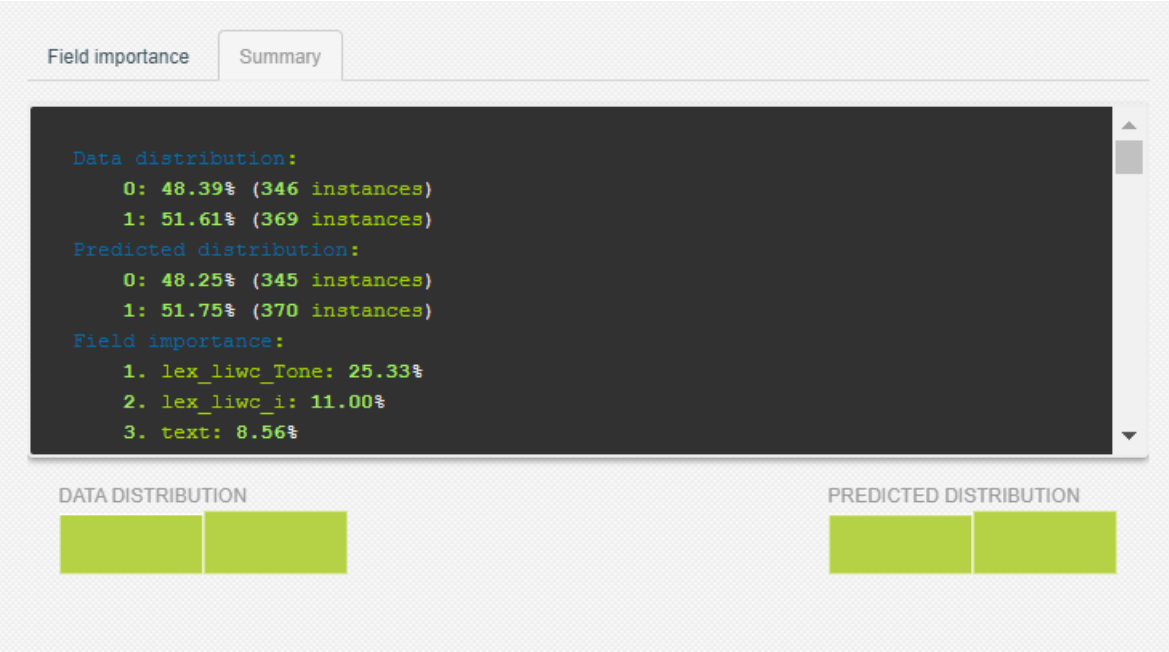
Na slici 3 prikazana je važnost različitih atributa korištenih u modelu. Svaki stupac predstavlja jedan atribut koji doprinosi modelu u donošenju odluka, a visina stupca označava koliki je njegov relativni značaj. Prema slici 3 lex_liwc_Tone ima najveću važnost u modelu, s najvećim udjelom doprinosa u predikciji s 24.89%. Iz toga možemo zaključiti da ton (vjerojatno emocionalni ton teksta) igra ključnu ulogu u određivanju da li je sadržaj stresan ili nije. Drugi po važnosti atribut je lex_liwc_i s 10.93%, s nešto manjim doprinosom u predikciji, ali i dalje značajnim. Ovaj atribut je povezan s određenom emocionalnom ili lingvističkom karakteristikom. Treći atribut text sam po sebi također ima značajnu uloga s 8.64%. Dalje su tu lex_liwc_CCog koji je povezan s kognitivnim procesima, ima srednju važnost i može ukazivati na to kako sadržaj potiče na razmišljanje ili angažman čitatelja. Confidence je također značajan i zauzima srednje mjesto po važnosti. Dok ostali niži atributi imaju manji, ali ipak relevantan doprinos u procesu donošenja odluka modela.

4.1.2. Aktivno statističko obrezivanje



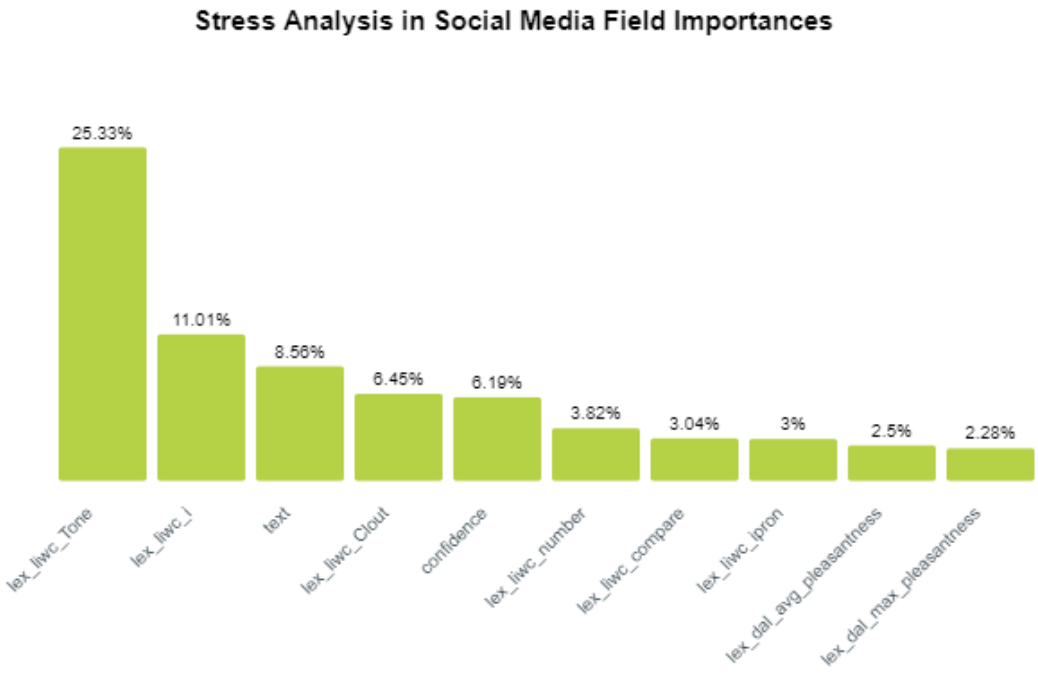
Slika 4. Stablo odlučivanja - aktivno statističko obrezivanje (vlastita izrada)

Na slici 4 je prikazan drugi model stabla odlučivanja dobivenog metodom aktivnog statističkog obrezivanja.



Slika 5. Točnost aktivnog statističkog obrezivanja (vlastita izrada)

Distribucija podataka gotovo je identična kao u prethodnom primjeru. Što se tiče predviđene distribucije u ovom slučaju nestresno je iznosilo 48.25% ili 345 instanci, što je vrlo blisko stvarnoj distribuciji podataka. Dok stresno iznosi 51.75% ili 370 instanci što je također gotovo identično originalnim podacima, što znači da je model zadržao uravnoteženost.

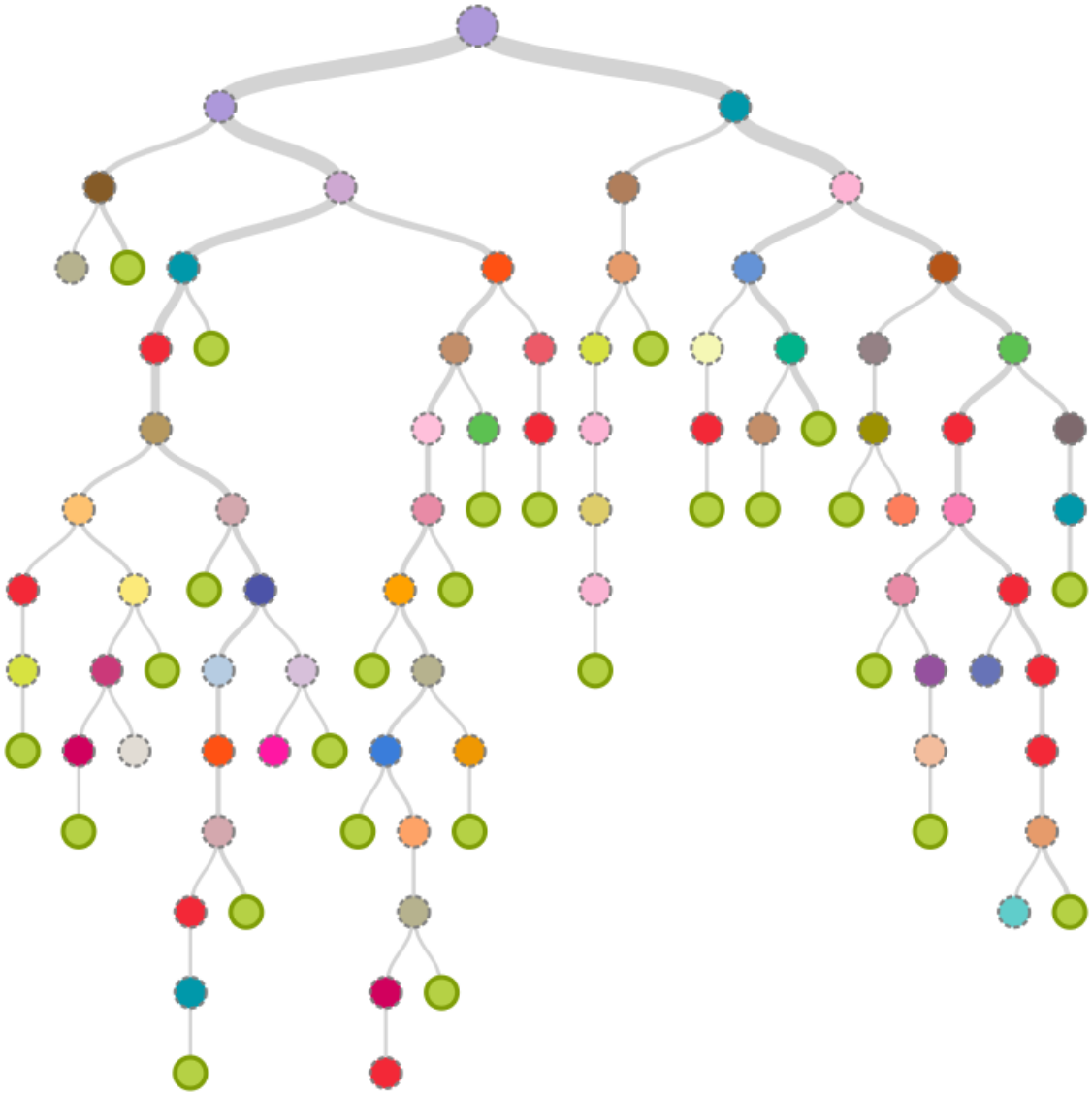


Slika 6. Prikaz važnosti atributa - aktivno statističko obrezivanje (vlastita izrada)

Kod aktivnog statističkog obrezivanja, atribut lex_liwc_Tone zadržao je svoju važnost s udjelom od 25.33%, što je vrlo slično s prethodnim rezultatom. Ovaj atribut ostaje najvažniji u donošenju odluka modela. Na drugom mjestu se nalazi lex_liwc_i koji ima važnost od 11.01% što je malo veći postotak u usporedbi s prethodnim prikazom, ali i dalje drži sličan rang. Također text ostaje treći po važnosti atribut s udjelom od 8.56%, što je blizu s prethodnom vrijednošću.

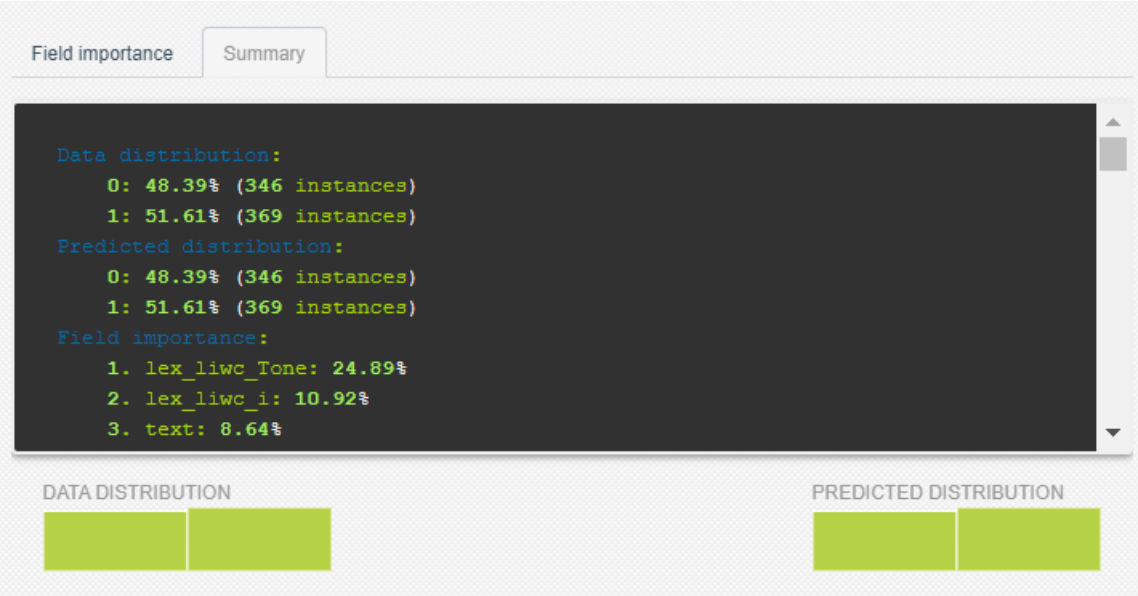
U oba slučaja, raspodjela podataka i predikcija je gotovo identična. Aktivno statističko obrezivanje nije značajno promijenilo distribuciju podataka u odnosu na rezultate dobivene pametnim obrezivanjem. Važnost atributa u oba pristupa ostaje slična, pri čemu lex_liwc_Tone ostaje najvažniji atribut, što znači da emocionalni ton teksta ostaje ključan faktor u predikcijama. Razlike u važnosti drugih atributa su minimalne, što ukazuje na to da aktivno statističko obrezivanje nije značajno promijenilo način na koji model vrednuje atribute.

4.1.3. Nestatističko obrezivanje

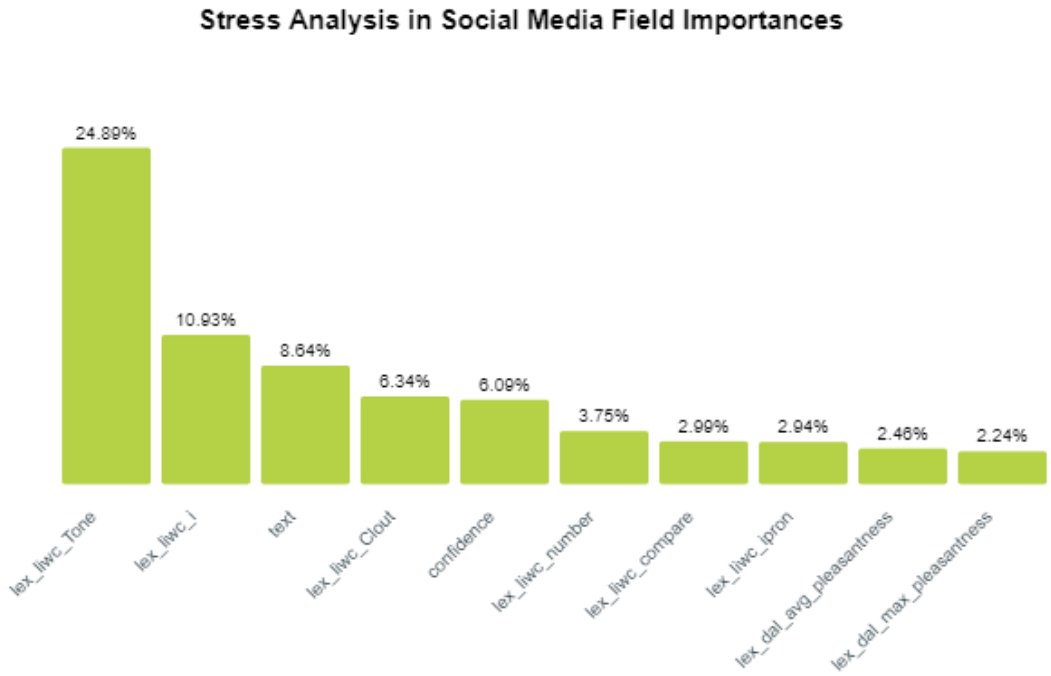


Slika 7. Stablo odlučivanja - nestatističko obrezivanje (vlastita izrada)

Slika 7 prikazuje treći model stabla odlučivanja dobivenog metodom nestatističkog obrezivanja.



Slika 8. Točnost nestatističkog obrezivanja (vlastita izrada)

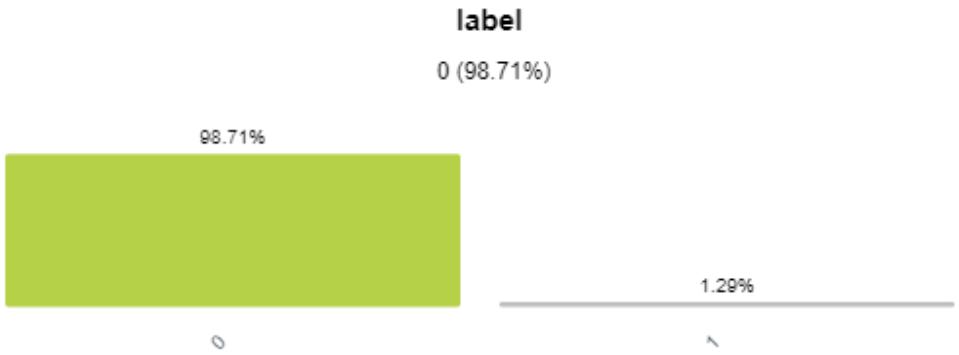


Slika 9. Prikaz važnosti atributa - nestatističko obrezivanje (vlastita izrada)

Kao i kod prethodnih metoda (pametno obrezivanje i aktivno statističko obrezivanje), distribucija podataka i predviđena distribucija identične su stvarnim podacima. Ovaj rezultat sugerira da model vrlo konzistentno prati distribuciju podataka, bez obzira na pristup obrezivanja. Atributi lex_liwc_Tone, lex_liwc_i i text zadržavaju svoje pozicije u sva tri pristupa, pri čemu lex_liwc_Tone ima najveći značaj u svim slučajevima. Promjene važnosti ostalih atributa su minimalne. Nestatističko obrezivanje nije značajno promijenilo performanse modela u usporedbi s prethodnim metodama.

4.1.4. Predikcija pomoću stabla odlučivanja

U ovom dijelu koristimo stablo odlučivanja razvijeno uz pomoć tehnike pametnog obrezivanja, budući da se ne razlikuje od aktivnog statističkog obrezivanja i nestatističkog obrezivanja. Tako ćemo predvidjeti je li sadržaj stresan ili nije.



Slika 10. Rezultat predikcije (vlastita izrada)

Na slici 10 je prikazan model koji je predvidio da je tekstualni unos nestresan (oznaka 0) s visokom točnošću od 98.71%. Ova predikcija pokazuje da model s visokim stupnjem sigurnosti klasificira sadržaj kao nestresan. Model stabla odlučivanja koristi nekoliko ključnih atributa kako bi donio odluku o stresnosti sadržaja. Neki od najvažnijih atributa su:

1. lex_liwc_Tone (24.89%): Emotivni ton teksta koji je najvažniji atribut u predikciji. U ovom slučaju, vrijednost 45.89 ukazuje na pozitivan ili neutralan ton, što podržava nestresnu klasifikaciju.
2. lex_liwc_Clout (6.34%): Osjećaj dominacije ili autoriteta u govoru također ima značajan utjecaj s vrijednošću 56.48. Viši lex_liwc_clout može ukazivati na samopouzdanje i jasnoću u komunikaciji, što smanjuje stresnost.
3. lex_liwc_AllPunc (2.14%): Korištenje interpunkcije u tekstu ima određeni utjecaj s vrijednošću 16.73, što može ukazivati na to da tekst sadrži dobro strukturirane rečenice, čime se umanjuje osjećaj stresa.

4. lex_liwc_negate (1.75%): Negacija u tekstu ima vrijednost 2.27, što sugerira da u tekstu nema mnogo riječi koje negiraju emocionalne izraze ili radnje, smanjujući potencijalnu stresnost.
5. lex_liwc_risk (1.45%): Ovaj atribut, koji ukazuje na spominjanje riječi povezanih s rizikom, ima nisku vrijednost 0.73, što dodatno podržava zaključak da tekst ne sugerira stresan sadržaj.

Na temelju ovog primjera, model stabla odlučivanja s pametnim obrezivanjem pokazao se kao učinkovit alat za klasifikaciju tekstualnog sadržaja. Ključni atributu poput emocionalnog tona, dominacije u govoru i upotrebe interpunkcije odigrali su glavnu ulogu u donošenju odluke. Ovakva visoka točnost predikcija (98.71%) ukazuje na uspješnost modela u prepoznavanju nestresnih tekstova.

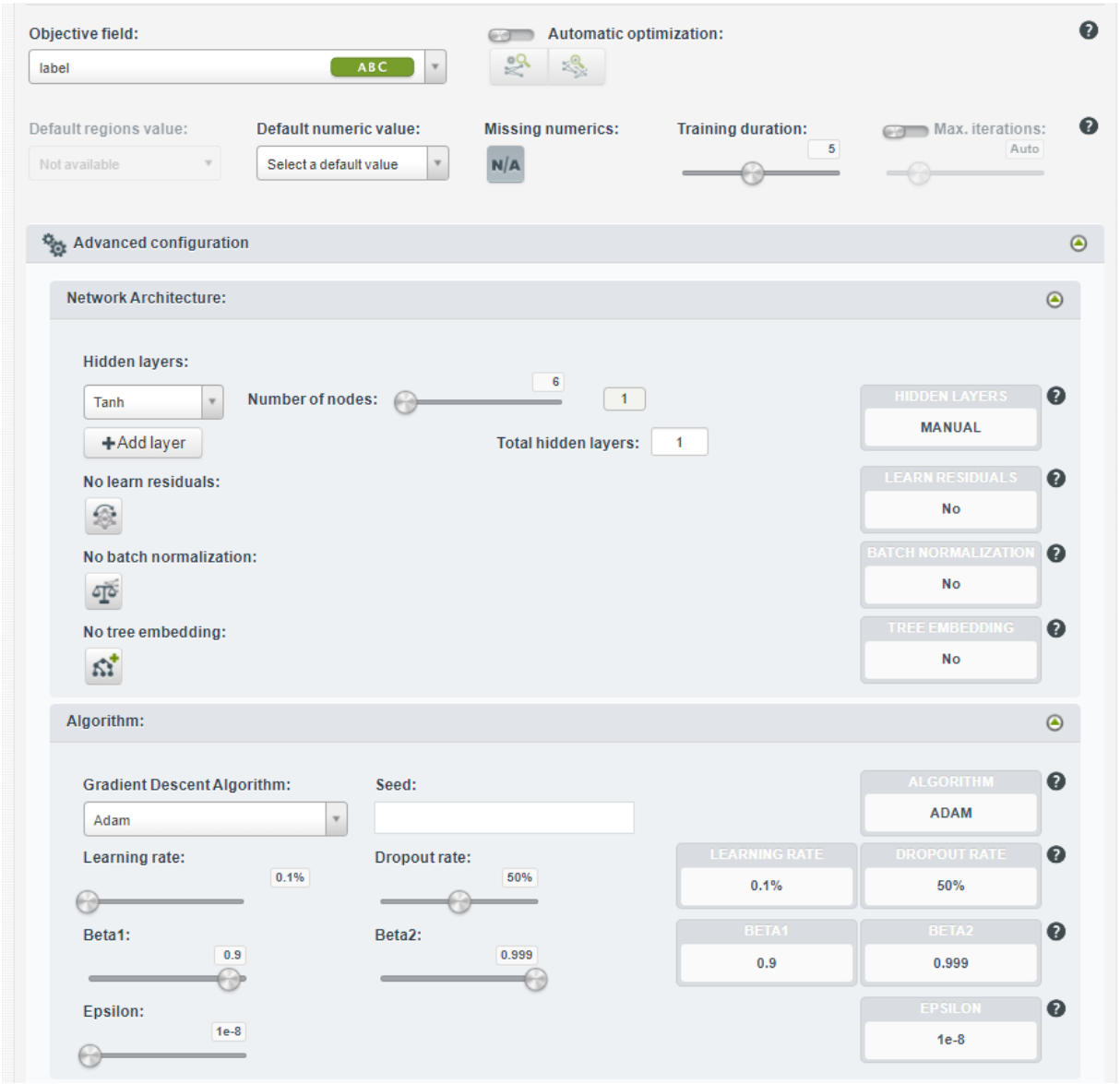
4.2. Analiza skupa podataka pomoću neuronskih mreža

U ovom poglavlju provest ćemo analizu skupa podataka koristeći neuronske mreže u alatu BigML. Skup podataka ostaje isti kao i u prethodnim analizama, te će se i dalje koristiti za predviđanje stresnih ili nestresnih objava na društvenim mrežama.

U odabranom skupu podataka potrebno je kliknuti na opciju „Configure“ i zatim odabrati „Deepnet“, odnosno neuronsku mrežu. Sljedeći korak je konfiguracija postavki za analizu. Umjesto korištenja automatskih postavki, ručno konfiguriramo model. Potrebno je odabrani zavisni atribut, onaj koji predviđamo. U ovom slučaju, zavisni atribut je „label“, koji označava je li sadržaj stresan (1) ili nestresan (0). Automatska optimizacija je isključena kako bismo mogli eksperimentirati s različitim brojem neurona.

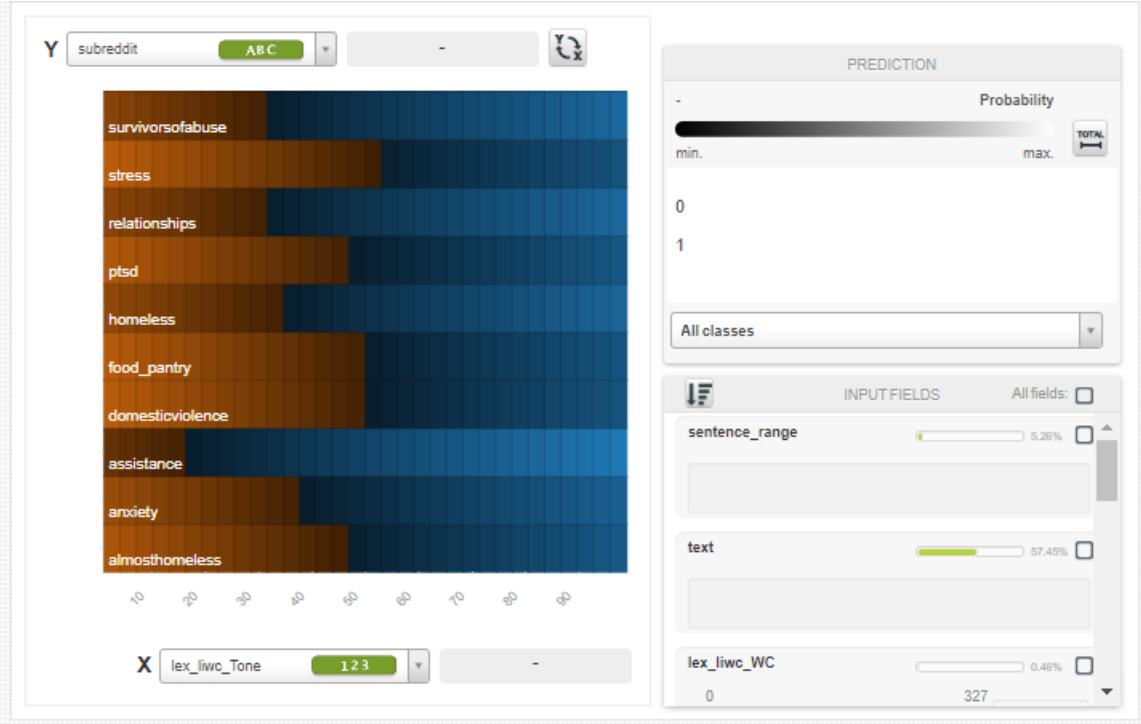
Zadana vrijednost u BigML alatu iznosi 64, ali u ovom slučaju koristit će se vrijednosti 6,7 i 8 jer smo izabrali 13 bitnijih atributa. Budući da su tri sloja neurona, jedan ulazni neuron predstavlja jednu ulaznu varijablu. Drugi je sloj u kojem se nalaze neuroni koji obrađuju podatke, te taj sloj može varirati. Treći sloj predstavlja sloj u kojem se nalazi ono što predviđamo. Broj neurona određuje se kao aritmetička sredina neurona na ulazu i izlazu, te je traženi broj neurona 7. Posljednji korak je varirati broj neurona za jedan manje ili za jedan više od traženog neurona [6].

U sljedećim koracima će biti pokrenut model s odabranim postavkama i usporedbom rezultata kako bismo odabrali optimalan broj neurona za ovaj slučaj.



Slika 11. n=6 (vlastita izrada)

U ovom prvom primjeru koristimo broj neurona jednak 6.

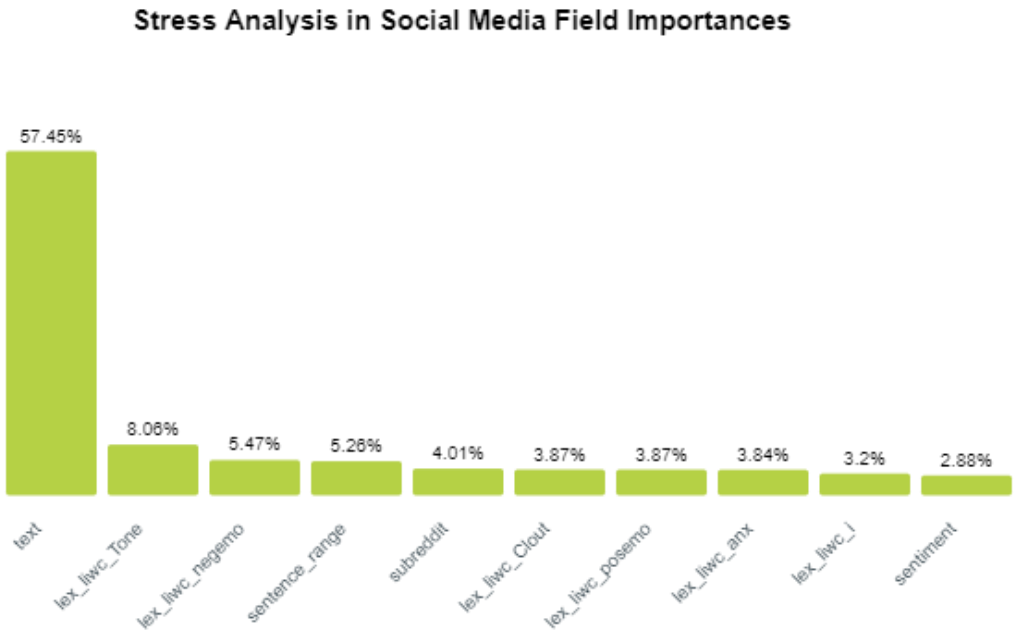


Slika 12. Grafikon n=6 (vlastita izrada)

Na ovoj slici vidimo toplinsku kartu (eng. heatmap) koja prikazuje subreddite na Y-osi, dok je na X-osi prikazana varijabla „lex_liwc_Tone**“**, koja mjeri emocionalni ton objave. Vrijednosti „lex_liwc_Tone**“** predstavljene su duž osi X, a boje na grafu pokazuju raspon od tamno smeđe (što označava niže vrijednosti tona) do plave (više vrijednosti tona).

S lijeve strane na Y-osi, prikazani su različiti subredditi (npr. „*survivorsofabuse“*, „*stress“*, „*relationships“*), dok su boje na grafikonu pokazatelji emocionalnog tona objava unutar tih subreddita. Na primjer, subredditi poput „*domesticviolence“* i „*survivorsofabuse“* imaju više tamno smeđih segmenata, što ukazuje na objave s nižim emocionalnim tonom (možda više negativnih ili stresnih objava).

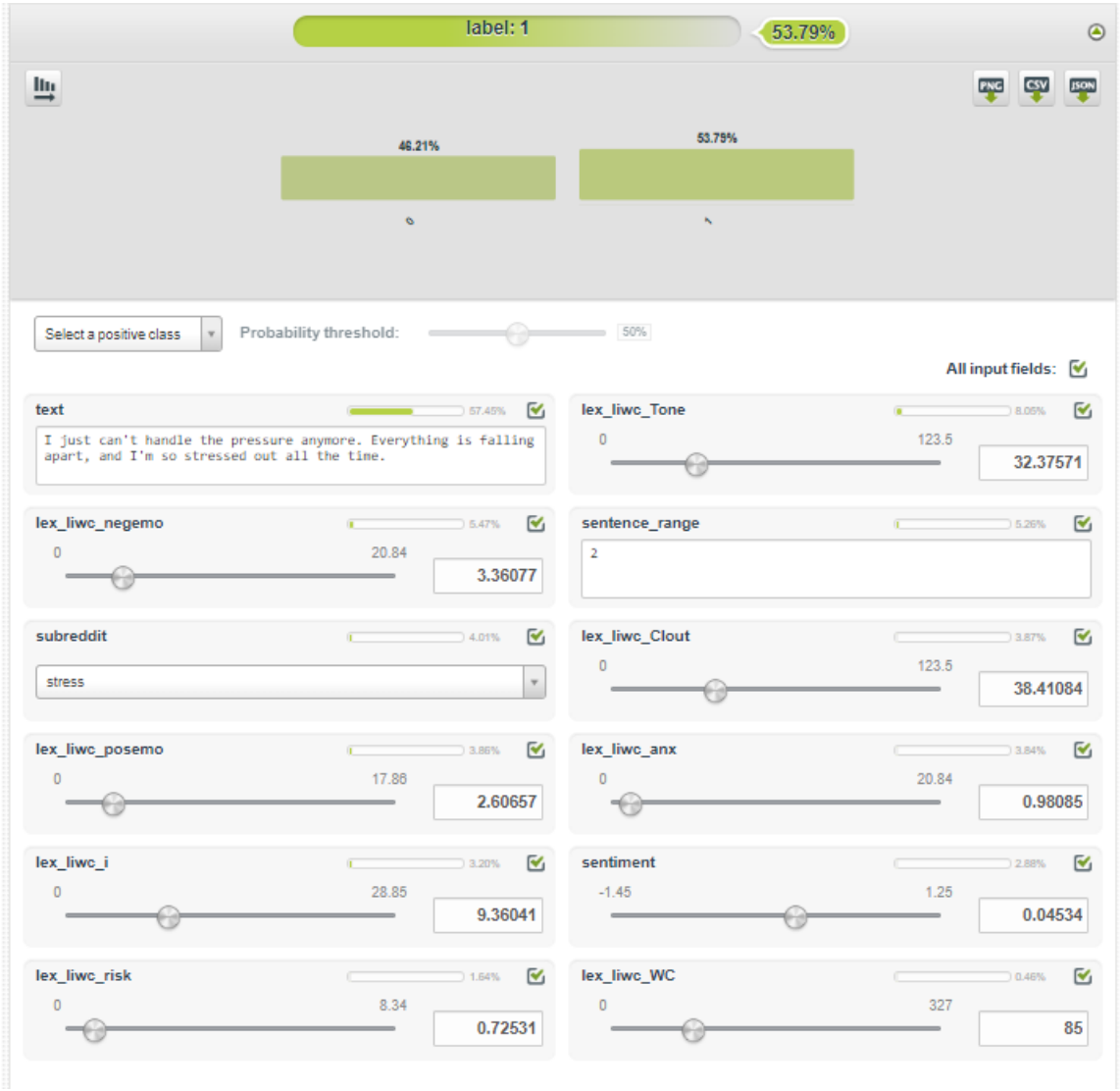
Desna strana slike prikazuje distribuciju vjerojatnosti, s prikazom ukupne vjerojatnosti (desno gore), koja je podijeljena između minimalnih i maksimalnih vrijednosti. Također su navedena dva atributa koja doprinose predikcijama: „sentence_range“ i „text“, s pripadajućim postotcima važnosti u ovom modelu.



Slika 13. Važnost atributa n=6 (vlastita izrada)

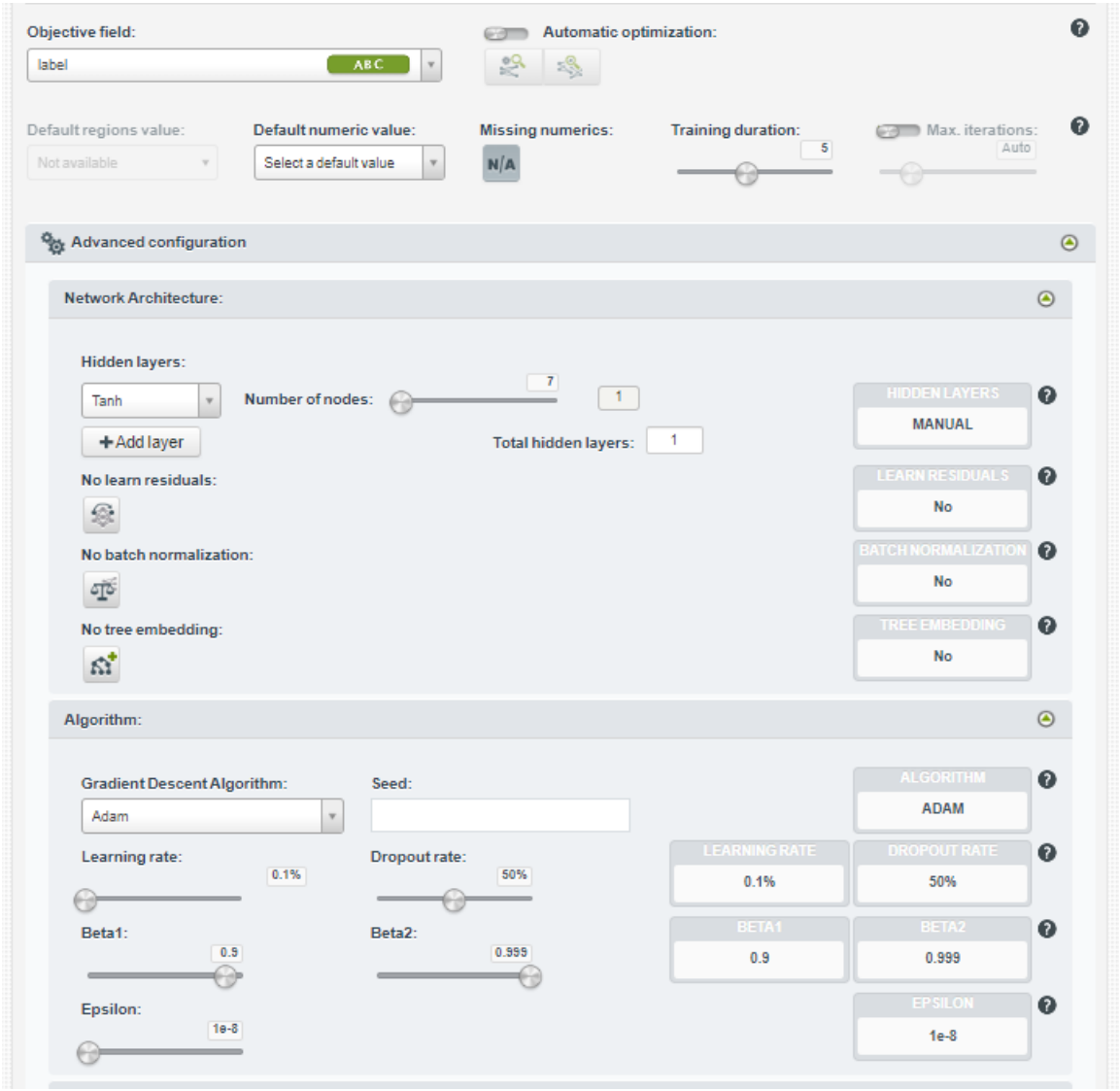
Na slici se prikazuju važnosti različitih atributa modela neuronske mreže korištene za predviđanje je li određena objava stresna ili nije. Svaki atribut pridonosi određenom postotku točnosti predikcije modela.

Možemo vidjeti da je „text“ najdominantniji atribut s udjelom 57.45%, koji sadrži stvarni tekst objave te najviše doprinosi modelu. To znači da se model u velikoj mjeri oslanja na sadržaj objave pri procjeni stresa. Emocionalni ton i jezične karakteristike ključni su za prepoznavanje stresnih situacija. Dalje imamo „lex_liwc_Tone koji je drugi najvažniji atribut s 8.06%. Ovaj atribut pokazuje razinu pozitivnih i negativnih emocija u tekstu, što pomaže u detekciji stresnih objava. Značajni su i atributi poput „lex_liwc_negemo“ (5.47%), koji mjeri negativne emocije, te „sentence_range“ s 5.26%, koji ukazuje na broj rečenica u objavi. Atribut „subreddit“ s 4.01% doprinosi klasifikaciji na temelju izvora objave, dok „lex_liwc_Clout“ i „lex_liwc_posemo“, s 3.87%, analiziraju autoritativnost i prisutnost pozitivnih emocija u tekstu.

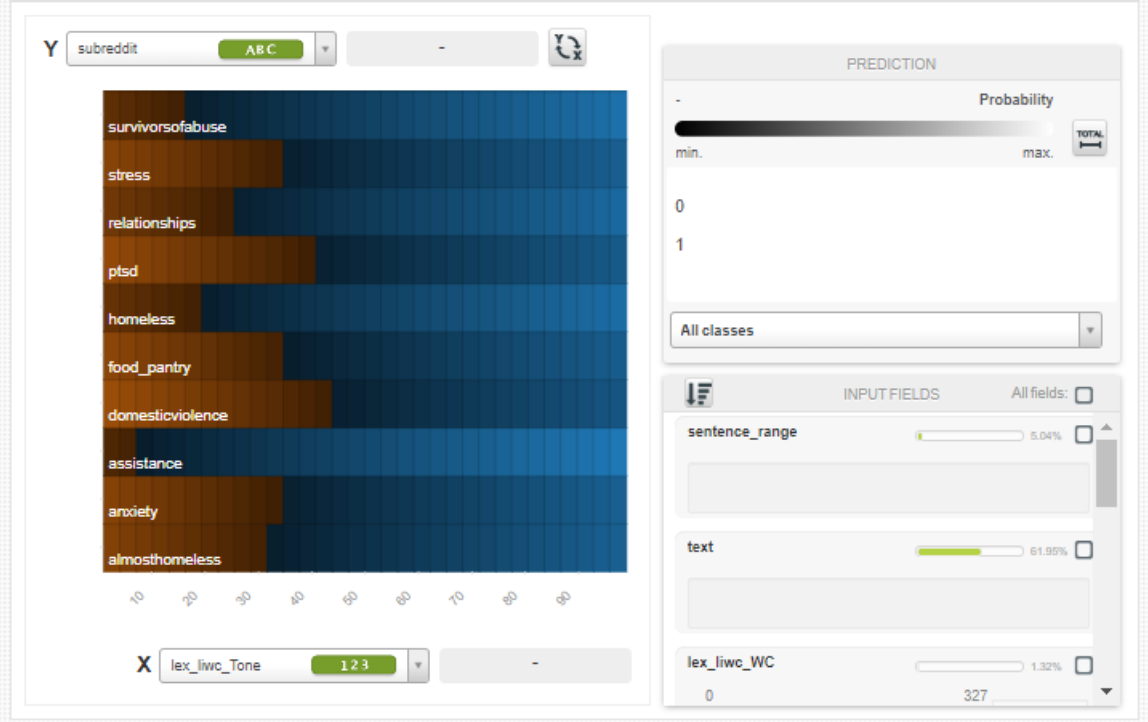


Slika 14. Predikcija n=6 (vlastita izrada)

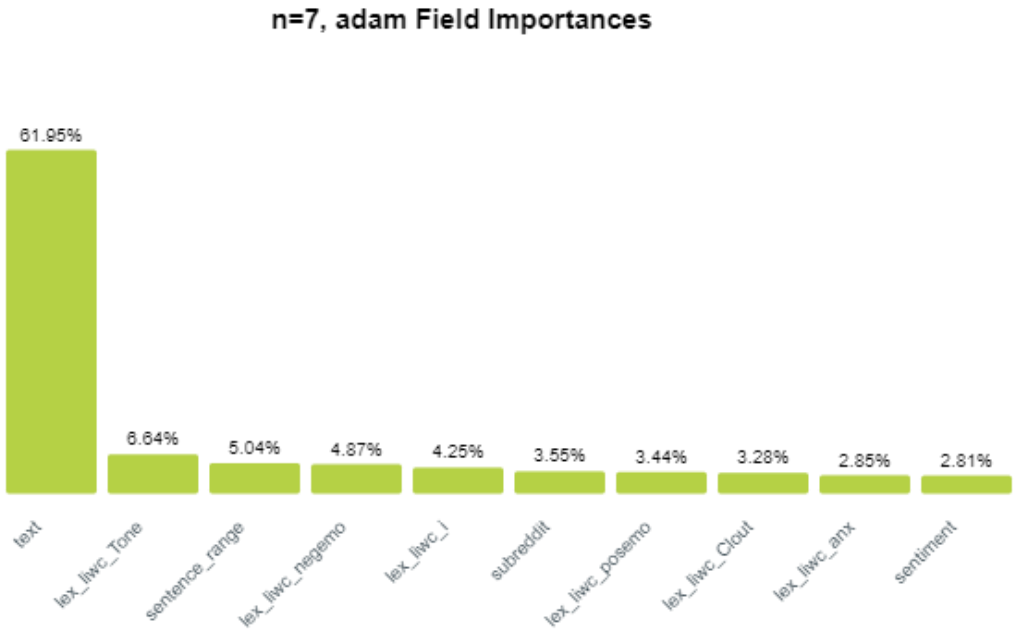
Na slici 14 su prikazane predikcije dvije labele, 0 i 1, koje predstavljaju predikcije modela neuronske mreže. Oznaka 1 iznosi 53.79% i ona ukazuje na predikciju da je objava stresna. Model je s relativno visokom vjerojatnošću klasificirao objavu kao stresnu, s naglaskom na atribute poput sadržaja teksta, emocionalnog tona i negativnih emocija. Oznaka 2 koja iznosi 46.21% predstavlja suprotnu klasifikaciju, odnosno da objava nije stresna. Iako je vjerojatnost manja, model još uvijek uzima u obzir atribute poput broja rečenica i sadržaja subbredita, kojimogu ukazivati na manje stresnu situaciju.



Slika 15. n=7 (vlastita izrada)

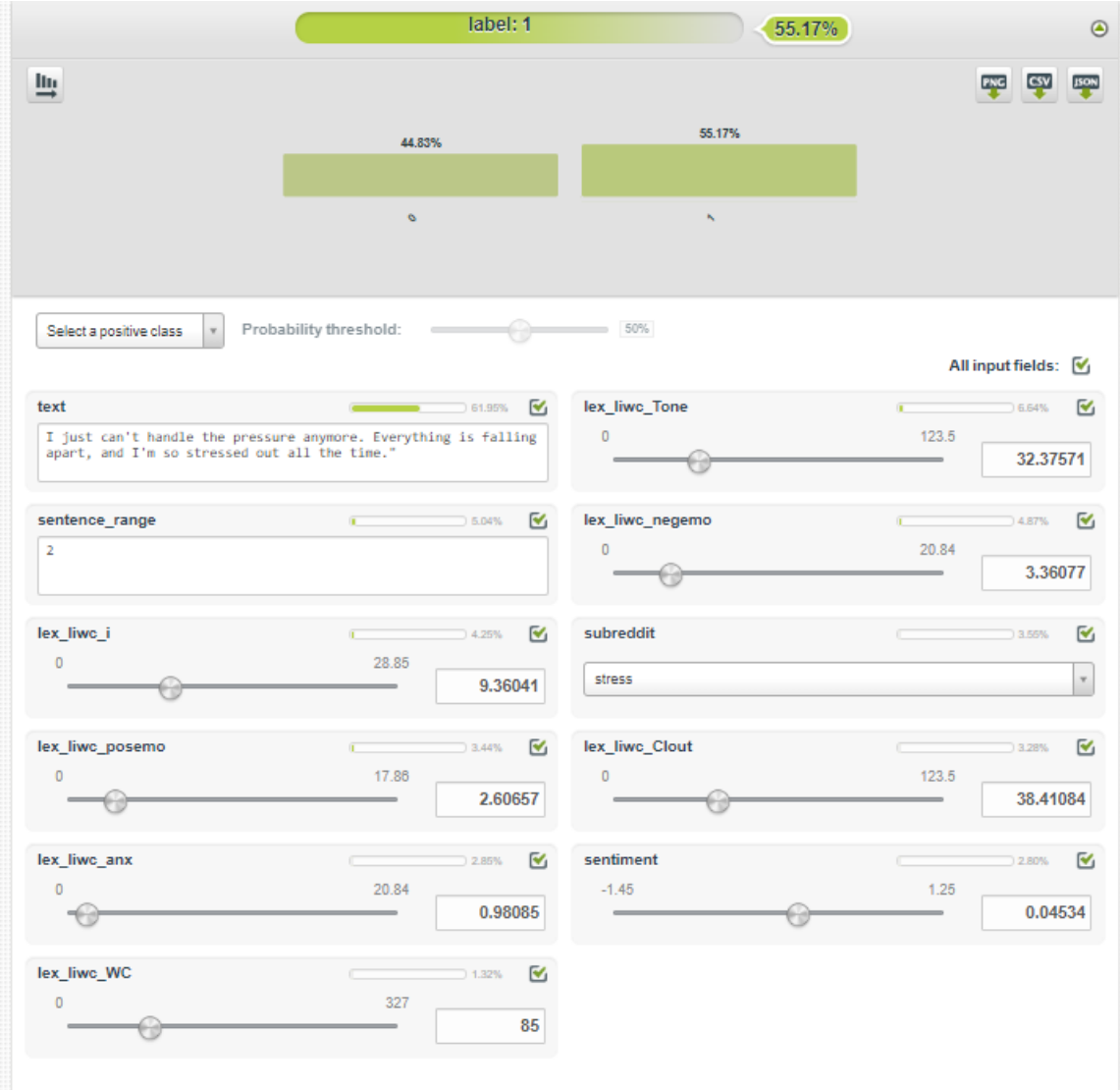


Slika 16. Grafikon n=7 (vlastita izrada)



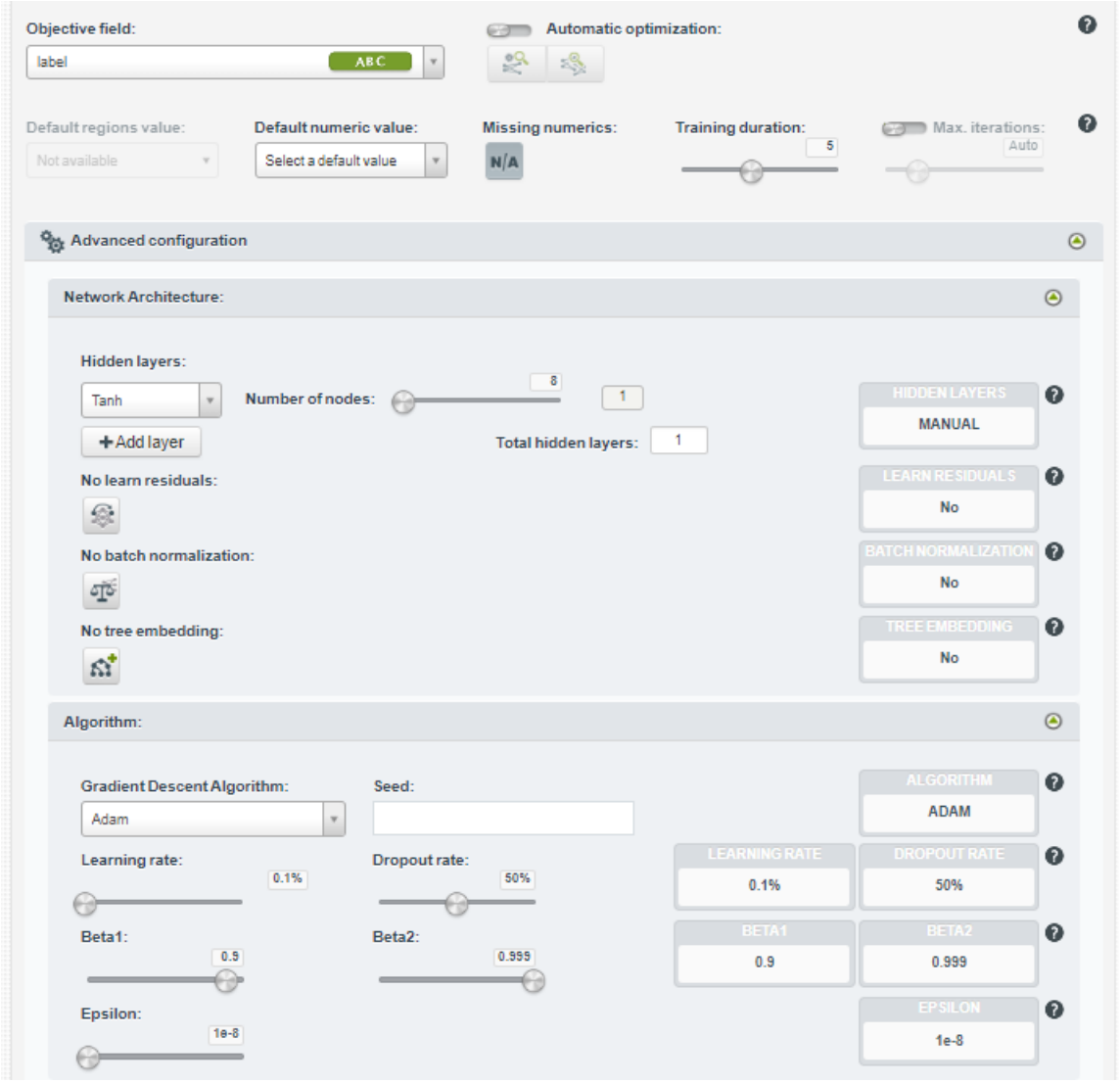
Slika 17. Važnost atributa n=7 (vlastita izrada)

Na slici 17 možemo vidjeti važnost atributa za predikciju modela neuronske mreže s 7 neurona. Najvažniji atribut je „text“ s udjelom od 61.95%, što jasno pokazuje da stvarni tekst objave najviše doprinosi predikciji. Zatim, „lex_liwc_Tone“ pridonosi s 6.64%, što se odnosi na emocionalni ton teksta, a „sentence_range“ s udjelom od 5.04% daje dodatni kontekst u vezi s brojem rečenica u objavi. Ostali atributi poput „lex_liwc_negemo“ (4.87%) i „lex_liwc_Clout „(3.44%) također pomažu u prepoznavanju emocionalnog stanja autora.

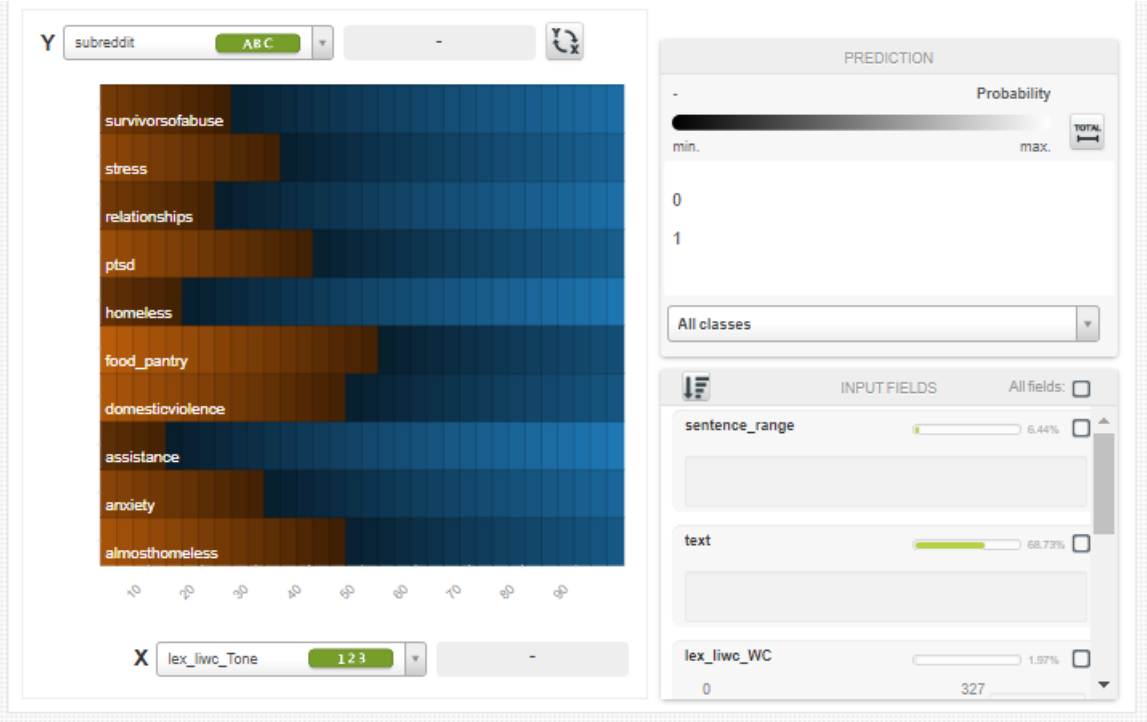


Slika 18. Predikcija n=7 (vlastita izrada)

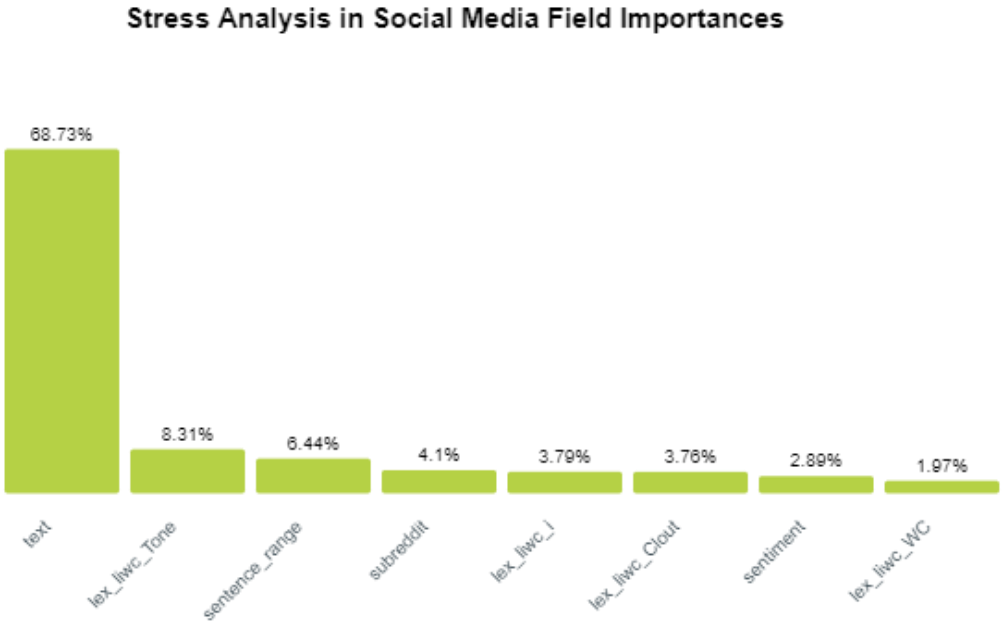
Slika 18 pokazuje predikciju stresa s vjerojatnošću od 55.17%, gdje je ključni atribut ponovno „text“ s visokim udjelom od 61.95%, što sugerira da se model oslanja prvenstveno na sadržaj objave pri donošenju zaključaka o stresu. Emocionalni ton i prisutnost negativnih emocija dodatno pomažu u procjeni, dok atributi poput „sentence_range“ i „subreddit“ pružaju dodatni kontekst u vezi s temom i složenošću objave.



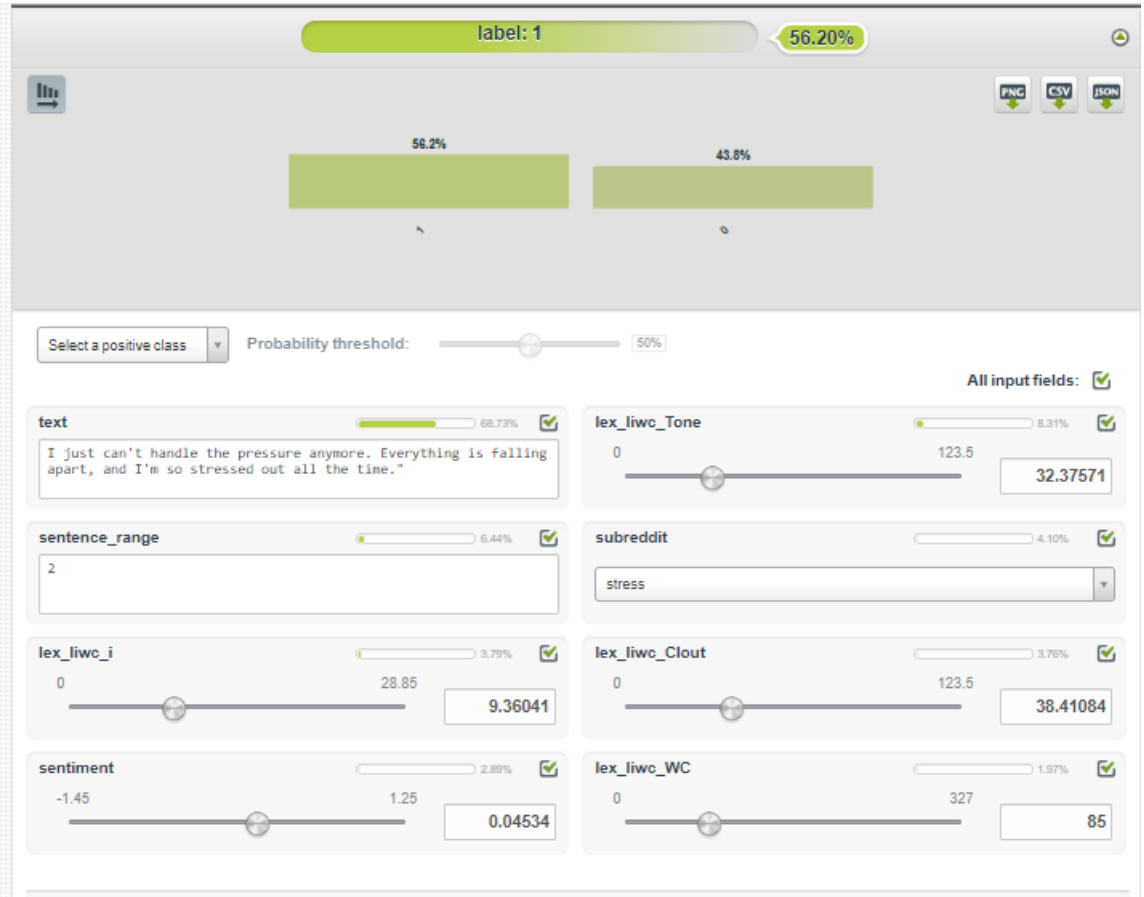
Slika 19. n=8 (vlastita izrada)



Slika 20. Grafikon n=8 (vlastita izrada)



Slika 21. Važnost atributa n=8 (vlastita izrada)



Slika 22. Predikcija n=8 (vlastita izrada)

Na temelju slike 17 i 18, gdje je korišteno 8 neurona za predikciju, možemo uočiti sljedeće da je „text“ i dalje najvažniji atribut s postotkom od 68.73%, drugi je i dalje „lex_liwc_Tone“ s 8.31%, te ostali atributi „sentence_range“, „lex_liwc_i i „subbredit“ imaju manji, ali značajan doprinos.

Na osnovi ovih atributa, model je dao predikciju s 56.20% vjerojatnošću da je objava stresna. Također, vidimo da je tekst objave: „I just can't handle the pressure anymore. Everything is falling apart, and I'm so stressed out all the time." snažno povezan s pristnošću stresa, što je logično s obzirom na to da je sadržaj teksta izravno povezan s izražavanjem osjećaja stresa.

5. Evaluacija i interpretacija modela

U analizi prediktivnih modela stabla odlučivanja i neuronskih mreža, važno je ne samo ocijeniti točnost modela, već i dublje razumjeti kako model dolazi do svojih predikcija, koji čimbenici najviše utječu na odluke te kolika je složenost modela.

Stabla odlučivanja pokazuju visoku točnost od 98.71%, uz relativno jednostavno interpretabilnost, što je njegova ključna prednost u odnosu na složenije modele poput neuronskih mreža. Ključni atributi, poput emocionalnog tona teksta („lex_liwc_Tone“), doprinose s 24.89% značaja u predikciji stresnih objava. Ovaj atribut jasno razlikuje stresne i nestresne objave na temelju emocionalnih signala, dok drugi važni atributi poput „lex_liwc_Clout“ i „lex_liwc_negemo“ doprinose s 6.34% i 2.14% važnosti. Pametno obrezivanje zadržalo je ravnotežu između točnosti i složenosti modela, uklanjajući nepotrebne grane koje ne doprinose značajno poboljšanju predikcije. Aktivno statističko obrezivanje donijelo je točnost od 98.71%, dok je nestatičko obrezivanje dalo slične rezultate, uz minimalne razlike u važnosti atributa. Ovi rezultati potvrđuju da pametno obrezivanje nudi najbolji omjer između jednostavnosti i točnosti, bez prekomjernog učenja. Što bi jasnije trebalo značiti da osigurava da model ostane što jednostavniji, a da pritom ne žrtvuje točnost, što ga čini idealnim izborom u situacijama gdje je jednostavnost modela jednako važna kao njegova točnost.

U analizi neuronskih mreža primijenjena je različita konfiguracija broja neurona, pri čemu je optimalan broj neurona utvrđen na 7, nakon eksperimentiranja s vrijednostima od 6 do 8. Preciznost modela kretala se od 53.79% za 6 neurona, do 55.17% za 7 neurona, dok je sa 8 neurona točnost bila 56.20%. S obzirom na složenost i 13 korištenih atributa, sedam neurona pružilo je najbolju ravnotežu između prekomjernog prilagođavanja i sposobnosti modela da prepozna složene uzorke. Iako se čini da je model s 8 neurona postigao nešto višu točnost (56.20%) u odnosu na model s 7 neurona (55.17%) kao optimalnog temelji se na ravnoteži između složenosti modela i njegove sposobnosti da generalizira nove podatke. Jednostavnije rečeno, model s više neurona postaje složeniji, što znači da bolje može „zapamtiti“ detalje u trenirajućem skupu podataka. Međutim, previše neurona može dovesti do prepilagođavanja, gdje model nauči previše specifičnosti i lošije se generalizira na nove podatke. Iako je točnost s 8 neurona veća, razlika između 7 i 8 neurona je vrlo mala (samo 1.03%). Ova mala razlika nije dovoljno značajna da opravda povećanu složenost koju donosi model s više neurona. Model s 7 neurona je nešto jednostavniji, a uz to nudi gotovo jednaku točnost kao model s 8 neurona. Jednostavniji model je poželjniji jer smanjuje šanse za preprilagođavanje, što znači da će bolje raditi na novim, nepoznatim podacima.

Neuronske mreže su pokazale snažnu sposobnost u prepoznavanju složenih obrazaca, posebno u tekstualnim podacima. Atribut „text“ je imao najveći doprinos s 61.95% u predikcijama kada je broj neurona bio sedam, dok je „lex_liwc_Tone*“* imao značaj od 8.06%. Ova sposobnost neuronskih mreža da detektiraju emocionalne nijanse u tekstu je vrlo korisna u kontekstima gdje su potrebne precizne analize emocionalnih stanja.

Evaluacijom oba modela, stablo odlučivanja pruža visoku točnost od 98.71% uz jednostavnu interpretabilnost, što ga čini boljim izborom u aplikacijama gdje je jasno tumačenje rezultata prioritet. Neuronske mreže, s druge strane, nude superiornost u prepoznavanju složenijih obrazaca, s maksimalnom točnošću od 56.20%, ali njihova složenost i manjak interpretabilnosti mogu biti prepreka u određenim aplikacijama. Optimalan broj neurona u ovoj analizi pokazao se kao sedam, ali uvijek treba uzeti u obzir specifičnosti skupa podataka pri konfiguraciji modela.

6. Zaključak

Na temelju provedene analize, stablo odlučivanja pokazalo se kao bolji izbor kada je u pitanju interpretabilnost i jednostavnost primjene u analizi stresa na društvenim mrežama. Njegova prednost leži u transparentnosti modela, što omogućuje lakše razumijevanje kako model dolazi do svojih predikcija. Uz korištenje, pametnog obrezivanja, model je postigao visoku točnost i zadržao balans između jednostavnosti i učinkovitosti.

Neuronske mreže, s druge strane, pokazale su superiornost u prepoznavanju dubljih obrazaca unutar podataka, no zbog svoje složenosti zahtijevaju više resursa i složeniju optimizaciju. Iako su predikcije bile precizne, složenost modela otežava njegovu interpretaciju.

U konačnici, rad sugerira da stablo odlučivanja može preporučiti kada se brza predikcija i jednostavno tumačenje rezultata prioriteti, dok su neuronske mreže prikladnije za dublje analize koje zahtijevaju prepoznavanje složenijih obrazaca u tekstualnim podacima. Ovisno o specifičnim potrebama analize, oba algoritma mogu pružiti korisne uvide, a njihova primjena treba biti prilagođena kontekstu istraživanja.

7. Popis Literature

[1] IBM. "Data mining" Dostupno: https://www.ibm.com/topics/data-mining Preuzeto 12. kolovoza 2024.)

[2] GeeksforGeeks. "Decision Tree" Dostupno: https://www.geeksforgeeks.org/decision-tree/. (Preuzeto: 12. kolovoza 2024.)

[3] GeeksforGeeks. "Neural Network" Dostupno: https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/ (Preuzeto: 12.kolovoza 2024.)

[4] A. Twin, "Data Mining," Investopedia. Dostupno: https://www.investopedia.com/terms/d/datamining.asp. (Preuzeto: 12. kolovoza 2024.)

[5] Simplilearn. "What is Data Mining?" Simplilearn. Dostupno: https://www.simplilearn.com/what-is-data-mining-article. (Preuzeto: 13. kolovoza 2024.)

[6] B. Kliček, D. Oreški, materijal s predavanja iz kolegija Otkrivanje znanja u podacima ak. god. 2021./2022., preuzeto s e-learning sustava Moodle (preuzeto: 13.kolovoza 2024.)

[7] N. Hotz, "CRISP-DM: A Standard Methodology to Ensure a Successful Project," Data Science Process Alliance. Dostupno: https://www.datascience-pm.com/crisp-dm-2/. (Preuzeto: 13. kolovoza 2024.)

[8] R. Cravit, "What is a Decision Tree?" Venngage. Dostupno: https://venngage.com/blog/what-is-a-decision-tree/#what. (Preuzeto: 13. kolovoza 2024.)

[9] A. Saini, "Decision Tree Algorithm," Analytics Vidhya. Dostupno: https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/. (Preuzeto: 13. kolovoza 2024.)

[10] N. Duggal, "Advantages of Decision Trees," Simplilearn. Dostupno: https://www.simplilearn.com/advantages-of-decision-tree-article. (Preuzeto: 13. kolovoza 2024.)

[11] W. Hillier, "What Is a Decision Tree?" CareerFoundry. Dostupno: https://careerfoundry.com/en/blog/data-analytics/what-is-a-decision-tree/. (Preuzeto: 13. kolovoza 2024.)

[12] J. Chen, "Neural Network," Investopedia. Dostupno: https://www.investopedia.com/terms/n/neuralnetwork.asp. (Preuzeto: 13. kolovoza 2024.)

[13] E. Turcan and K. McKeown. "Dreaddit: A Reddit Dataset for Stress Analysis in Social Media." Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), Hong Kong, 2019, pp. 97-107. Dostupno: https://aclanthology.org/D19-6213/. (Preuzeto: 20. kolovoza 2024.)

8. Popis slika

9. Popis tablica

10. Popis priloga

1. Kaggle (2024.). Dostupno 14.9.2024. na: https://www.kaggle.com

2. BigML (2024.). Dostupno 14.9.2024. na: https://bigml.com