SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. xxxx

MODEL STROJNOG UČENJA ZA DETEKCIJU KOMPROMITIRANIH WEB-STRANICA U HRVATSKOM DOMENSKOM PROSTORU~~DA~~

Bruno Skendrović

Zagreb, rujan 2022.

Sadržaj

Uvod

Broj aktivnih web stranica neprestano raste te se njihov broj popeo do dvjesto milijuna [1]. S porastom broja web stranica, raste i kriminalna aktivnost na Internetu. Napadači putem weba pokušavaju raznim metodama prevariti korisnika u davanju osjetljivih informacija ili ga navesti na skidanje i pokretanje zlonamjernog programa u svrhu krađe podataka i ucjenjivanja, najčešće radi financijske dobiti. Prema Googleovom istraživanju iz 2020. [2] broj phishing web sjedišta se popeo na skoro 2 milijuna što je porast od 20 puta od 2015. godine kada ih je bilo oko sto tisuća, a web sjedišta iz kojih se širi zlonamjerni softver također postaje sve veći problem za kibernetičku sigurnost. Zbog važnosti interneta i web stranica u današnjem svijetu, nužno je što prije otkriti zlonamjerne web stranice te ih ukloniti ili upozoriti korisnika na moguću opasnost pri pristupanju takvih web stranica. Ručne metode provjere web stranica postaju nedostatne i nailaze na problem skalabilnosti dok strojno učenje pruža zanimljive i obećavajuće tehnike detekcije zlonamjernih web stranica.

Prvo poglavlje ovog rada opisuje usredotočeni pretraživač weba (engl. *Focused Web Crawler*) *MalCrawler* koji je razvijen tako da efikasno pronalazi zlonamjerne web stranice na Internetu. Pomoću *MalCrawlera* napravljen je skup podataka sa zlonamjernim i benignim stranicama i odabranim značajkama. Značajke i njihov mogući doprinos u detekciji zlonamjerne stranice opisuje se kroz nekoliko idućih pot poglavlja.

U drugom poglavlju opisani su rezultati algoritama strojnog učenja nad *MalCrawler* skupom podataka te njihova uspješnost u klasifikaciji zlonamjernih stranica.

U trećem poglavlju raspisane su metode kojima se dolazi do podataka koji se nalaze u *MalCrawler* skupu podataka. Te metode se kasnije koriste i na stranicama u hrvatskom domenskom prostoru kako bi se stvorio novi skup podataka.

U četvrtom poglavlju predstavlja se statistika značajki za stranice u Hrvatskom domenskom prostoru te rezultati strojnog učenja pri klasifikaciji zlonamjernih stranica.

Rad završava sa zaključivanjem uspješnosti algoritama strojnog učenja u klasifikaciji zlonamjernih stranica u stvorenom skupu podataka.

1. MalCrawler

Zbog sve većeg broja zlonamjernih web stranica [3] razvijeni su usredotočeni programi za indeksiranje i pretraživanje weba (engl. *Focused Web Crawler*) kako bi se zlonamjerne stranice što lakše i brže mogle detektirati. Primjeri takvih usredotočenih web pretraživači su *MalCrawler*-a[4], *EvilSeed*[5], *Monkey-Spider*[6].

MalCrawler koristi procjenu zlonamjernosti stranice u stvarnom vremenu kako bi retroaktivno *Web Crawler*-u dao veći značaj sumnjivim stranicama te tako ga vodio prema pretraživanju stranica koje vode na nedavno otkrivenu sumnjivu stranicu i stranice na koje otkrivena sumnjiva stranica vodi. Princip rada *MalCrawler*-a će biti dodatno objašnjeni jer ovaj rad koristi skup podataka koje je stvorio *MalCrawler* i neke od metoda koje *MalCrawler* koristi za procjenu zlonamjernosti stranice.

EvilSeed [5] koristi početni *seed* poznatih zlonamjernih web stranica kako bi program automatski generirao upite prema stranicama koje su povezane s tim zlonamjernim stranicama.

MalCrawler je usredotočeni program za indeksiranje i pretraživanje weba (engl. *Focused Web Crawler*)[4]. U usporedbi s običnim *Web Crawler*-om pronalazi više zlonamjernih web stranica. Također, razvijen je tako da rješava problem tehnike prikrivanja web stranice. Prikrivanje web stranice je tehnika promjene učitavanja i prikaza stranice s obzirom na klijentov preglednik i postavke preglednika[7]. *MalCrawler* šalje više HTTP zahtjeva s različitim korisničkim agentom u parametrima zahtjeva na tu stranicu te uspoređuje dobivene odgovore. Kako ne bi zapeo u pretraživanju malog skupa web stranica koji su međusobno jako povezani, konstantno mijenja način pretraživanja iz pretraživanja u dubinu u pretraživanje u širinu i obrnuto, ovisno o dobivenim rezultatima o zlonamjernosti. Posljednja značajka *MalCrawler*-a vrijedna spomena jest višestruko pretraživanje stranica s dinamičnim sadržajem, odnosno stranice koje koriste asinkroni JavaScript i XML (engl. *Asynchronous JavaScript and XML*, skr. *AJAX*), što mnogi tradicionalni web pretraživači izbjegavaju.

MalCrawler unutar *Crawler* modula koristi modul za procjenu zlonamjernosti netom pretraženog URL-a. Atributi s kojima određuje zlonamjernost stranice su sljedeći:

- Preusmjeravanje (engl. *Redirection*) – preusmjeravanje web stranice na drugu web stranicu pri učitavanju

- Maskiranje (engl. *Cloacking*) – prikazivanje različitog web sadržaja s obzirom na klijentov preglednik
- Detekcija obfuskacije – traženje obfusciranog JavaScripta na stranici
- Broj izvršavanja dinamičkog koda
- Duljina dinamički evaluiranog koda
- Broj bajtova dodijeljeni u memorijskom prostoru za traženu stranicu
- Broj pokrenutih komponenti preglednika
- Atributi i vrijednost parametara pri pozivanju metoda
- Broj poziva i izvršavanja metoda

Važno je napomenuti da su vrijednosti svih atributa dobiveni dinamičkim analizom, to jest putem višestruke komunikacije sa samom web stranicom pomoću sljedećih alata: *Rhino JavaScript Emulation Library*, *HTML Unit Browser Emulation Library*, *JSoup Library*, *WEKA Data Mining Library*. Neki od načina dinamičke analiza web stranice su pokretanje JavaScript funkcija na stranici s različitim parametrima te proučavanje rezultata koje funkcija vrati, odvajanje statičkog i dinamičkog sadržaja stranice korištenjem emulatora koji ima direktnu interakciju s elementima web stranice poput gumba, unosa teksta, prelaska mišem preko elementa,... Statička analiza web stranice podrazumijeva da nakon poslanog zahtjeva na stranicu prestaje komunikacija s istom te se sav pohranjen sadržaj i kod stranice analizira. U ovom radu svi su atributi dobiveni statičkom analizom.

1.1. MalCrawler skup podataka

Autori *MalCrawlera* su postavili skup značajki koji u stvarnom vremenu javlja web pretraživaču radi li se o zlonamjernoj ili benignoj stranici i time *MalCrawler* stvara svoj idući skup URL-a koje će obići. Vrijednosti značajki za pojedini URL su se zasebno pohranjivale te je tako stvoren *MalCrawler* skup podataka.

MalCrawler skup podataka može se pronaći na stranicama tvrtke *Mendeley* [8] i napravljen je 2017. godine. Napravljen je od podskupa značajki kojeg su autori *MalCrawler*-a predložili u članku [9] gdje su koristili stranice dobivene usredotočenim web pretraživačem

MalCrawler. U istom članku, autori AK Singh i Navneet Goyal predložili su 25 atributa koji bi se mogli koristiti u modelima strojnog učenja.

MalCrawler skup podataka sastoji se od dvije csv datoteke: skup podataka za treniranje koji ima 1.2 milijuna zapisa i skup podataka za testiranje koji se sastoji od 0.354 milijuna zapisa.

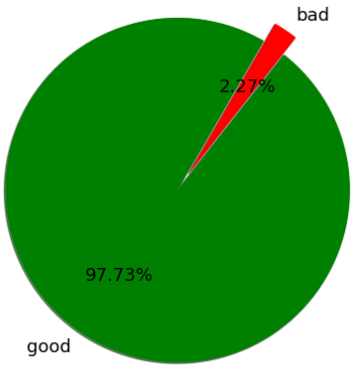
Značajke koji se mogu pronaći u navedenom skupu podataka su sljedeće:

- URL web stranice
- Duljina URL-a
- Vršna domena web stranice
- IP adresa web stranice
- Zemljopisna lokacija na kojoj se web stranica nalazi
- Veličina JavaScript koda na web stranici
- Veličina obfusciranog JavaScript koda na web stranici
- Koristi li stranica HTTP ili HTTPS protokol
- WHOIS informacije o domeni, te jesu li one potpune
- Sadržaj web stranice, uključujući i JavaScript kod
- Klasifikacijska oznaka za web stranicu, je li zlonamjerna ili benigna

1.1.1. Klasifikacija

Stranice su klasificirane Googleovim servisom za sigurno pretraživanje (engl. *Google Safe Browsing*). Servis *Safe Browsing* bit će detaljnije obrađen u idućem poglavlju.

U skupu za treniranje je 1 172 747 benignih i 27 253 zlonamjernih stranica dok je u skupu za testiranje 353 872 benignih i 8 062 zlonamjernih stranica. Sveukupno, to je 1 526 619 benignih stranica i 35 315 zlonamjernih stranica što čini omjer zlonamjernih prema benignim 0.0231. Odmah je jasno da postoji velika neuravnoteženost u podatcima što će za mnoge klasične algoritme strojnog učenja predstavljati problem kod klasifikacije jer će model naginjati prema tome da svaki ulaz označi klasom većinske klase, u ovom slučaju benigne. U kasnijem poglavljima prikazani su rezultati raznih algoritama strojnog učenja s namještenim hiperparametrima kako bi se što više smanjio ovaj problem nebalansiranih podataka. Na Slika 1.1 prikazan je kružni graf koji prikazuje postotke udjela svakog od klasa gdje zeleni dio grafa (*good*) predstavlja benigne a crvena (*bad*) zlonamjerne web stranice.

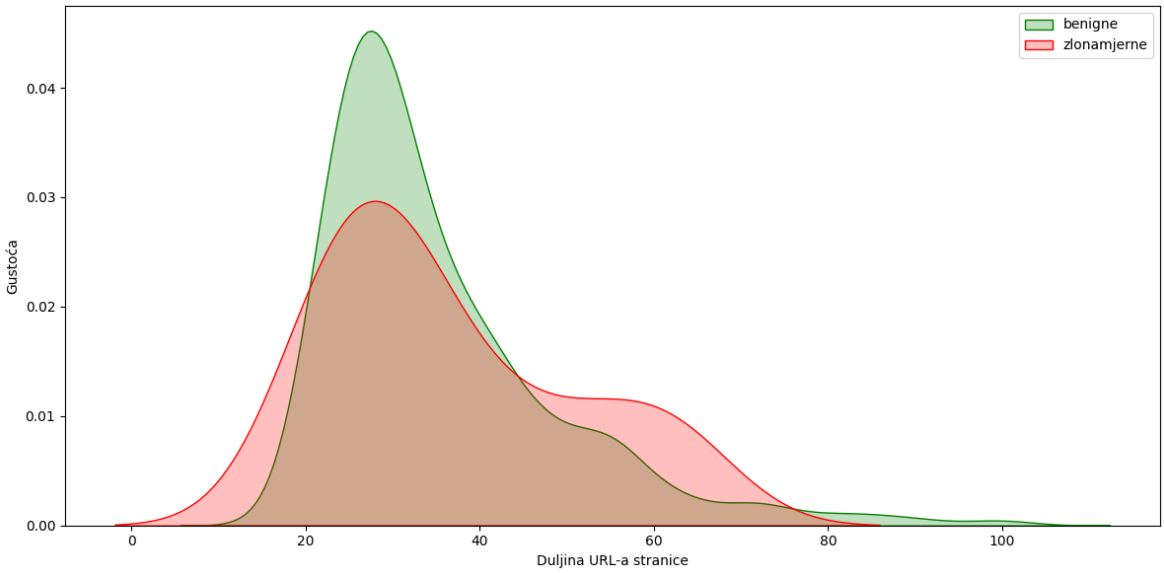


Slika 1.1: Kružni graf udjela benignih i zlonamjernih stranica

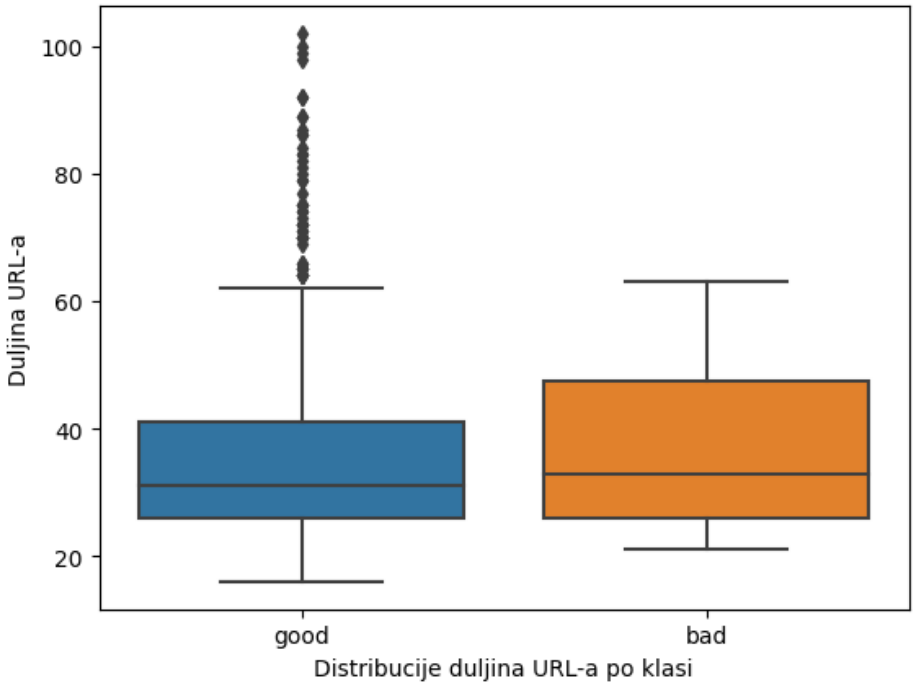
1.1.2. Duljina URL-a

Prema [10] URL-ovi se mogu skraćivati kako bi mogli prikrivati originalni URL. Neki od servisa koji nude takvo skraćivanje su *Tiny URL* i *Bitly*. Zbog ovakve mogućnosti prikrivanja, vrlo kratki URL-ovi mogu upućivati da se u stvarnosti radi o zlonamjernom URL-u. Druga pretpostavka je ta da se zlonamjerni URL može nalaziti na dugačkom putu popularnih benignih URL-ova i od tamo širiti zlonamjerni sadržaj [11][12].

Ispod su prikazani graf gustoće za benigne (zeleno) i zlonamjerne stranice (creveno) Slika 1.2 Gustoća vjerojatnosti duljina URL-a benignih i zlonamjernih stranica. Na Slika 1.3 prikazan je boxplot. Boxplot služi za prikaz distribucije vrijednosti gdje „kućica“ obilježava vrijednosti od nižeg do višeg kvartila s crtom unutar te „kućice“ koja označava medijan. Od gornjeg kvartila do gornjeg ekstrema te od donjeg kvartila do donjeg ekstrema nalaze se „dlake“ (engl. *Whiskers*) [13].



Slika 1.2 Gustoća vjerojatnosti duljina URL-a benignih i zlonamjernih stranica



Slika 1.3: Boxplot graf distribucija duljina URL-a po klasi

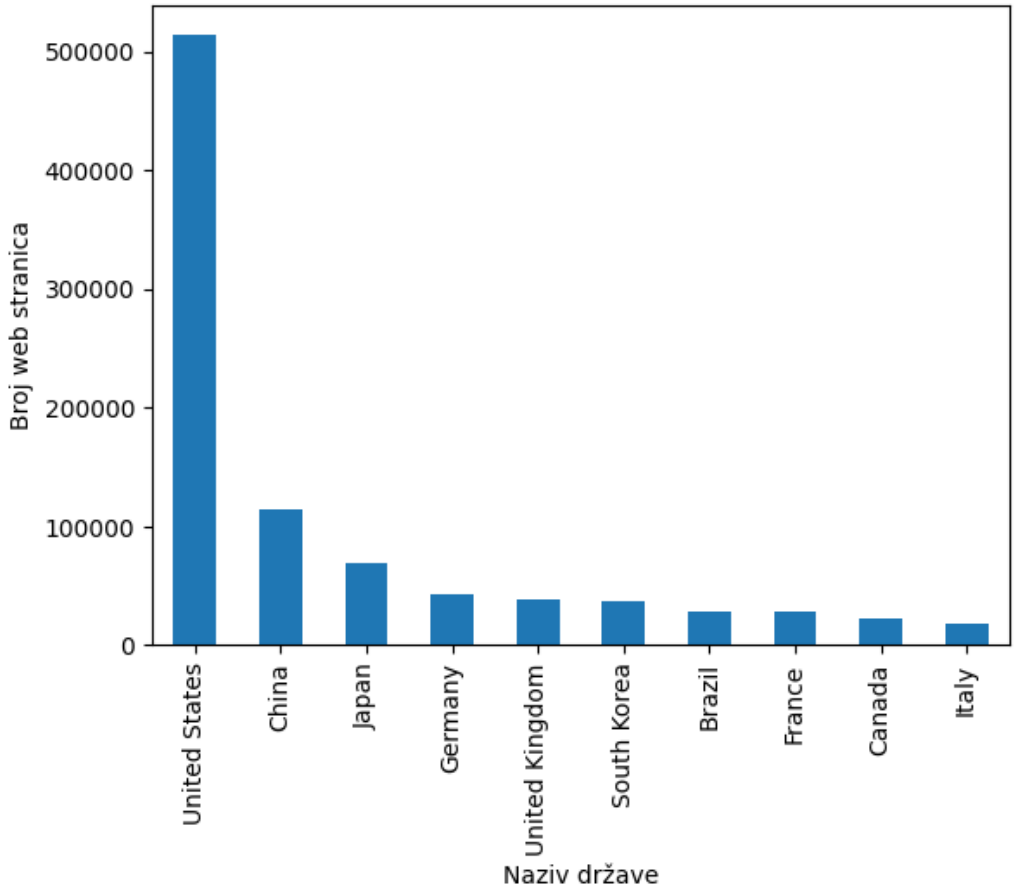
Iz grafova na Slika 1.2 raspoznaje se da su vrijednosti duljine URL-ova benignih i zlonamjernih stranica neznatno različite. Duljina zlonamjernih URL-ova u prosjeku nešto veća od duljine benignih URL-ova, konkretno prosječna duljina benignog URL-a jest 35.82, a zlonamjernog 37.15. Gustoća vjerojatnosti zlonamjernog grafa je veća u području od duljine 50 do 60 znakova dok benigni URL-ovi imaju veći raspon duljine znakova te URL-ove s jako dugačkim nizom znakova (od 80 nadalje). To potvrđuje i boxplot graf na Slika 1.3 gdje je gornji kvartil zlonamjernih URL-ova veći od benignih, a znatan broj benigni URL-ova preskače granicu gornjeg ekstrema. No, zbog relativne sličnosti distribucija duljine URL-ova, početna pretpostavka jest da ova značajka neće imati veliki utjecaj na klasifikacijski algoritam.

1.1.3. Zemljopisna lokacija na kojoj se nalazi domaćin (host) web stranice

Prema izvješću iz travnja 2022. o trendovima u web prijetnjama [14] pokazuje se da je broj zlonamjernih te omjer benignih i zlonamjernih web stranica vrlo različit od kontinenta do

kontinenta te od zemlje do zemlje. Također, pokazuje se da zemlje s većim omjerom zlonamjernih i benignih stranica su češće pod napadom i drugih vrsta kibernetičkih napada. Po istom navedenom izvješću, zemlje s najviše zlonamjernih URL-ova su redom: Sjedinjene Američke Države (65.1%), Njemačka (3.8%), Rusija (3.3%), Francuska (1.4%), Nizozemska (1.3%), Ujedinjeno Kraljevstvo (1.2%), Kina (1.2%), Brazil (1.2%) te ostali (21.5%). Shodno tome, kao značajka u klasifikaciji stranice, uzeto je u obzir i zemljopisna lokacija domaćina web stranice.

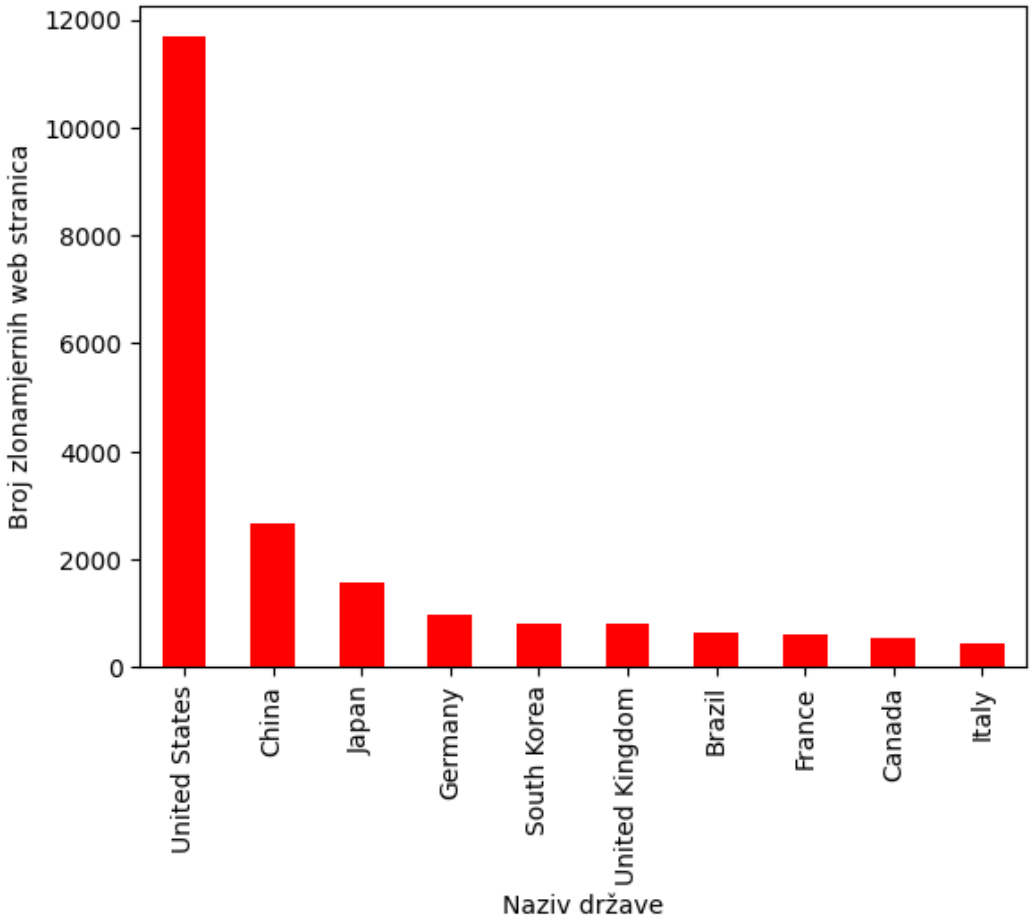
Na Slika 1.4 prikazano je prvih 10 zemalja s najvećim brojem web stranica koji se nalaze u skupu podataka za treniranje (1.2 milijuna zapisa).



Slika 1.4: Prvih 10 zemalja s najvećim brojem web stranica u MalCrawler skupu za treniranje

Na slici se vidi da je najveći broj stranica, koje se proučavaju u ovom skupu podataka, proizlazi iz Sjedinjenih Američkih Država (513 396) i to skoro pet puta više od sljedećeg najvećeg broja proučenih stranica iz neke zemlje što je u ovom slučaju Kina (113 659). Sljedeće zemlje su Japan (68 786), Njemačka (42 121) te Ujedinjeno Kraljevstvo (37 869). Iz Hrvatske je obrađeno 1 328 web stranica.

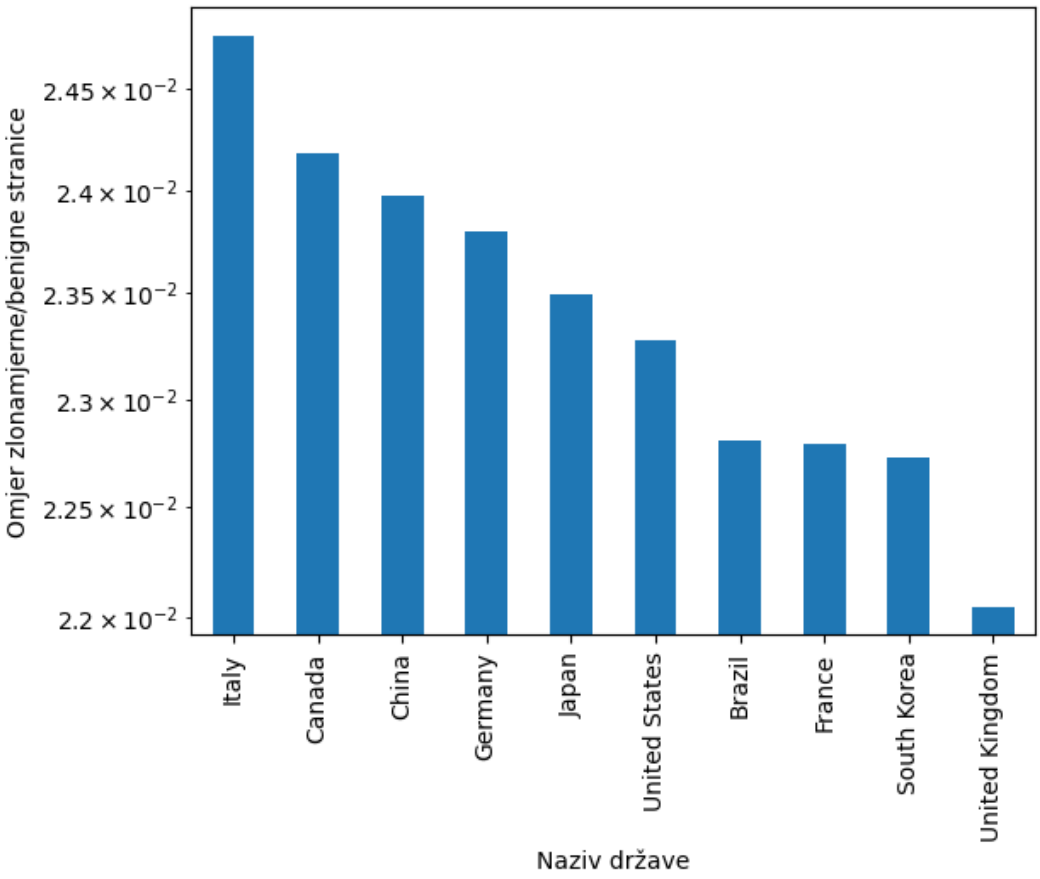
Na Slika 1.5 prikazan je broj zlonamjernih stranica za prvih 10 zemalja s najvećim brojem zlonamjernih stranica.



Slika 1.5: Prvih 10 zemalja s najvećim brojem zlonamjernih stranica u MalCrawler skupu podataka

Najveći broj zlonamjernih stranica locirani su u Sjedinjenim Američkim Državama (11 678), što je i očekivano s obzirom na sveukupni broj stranica lociranih u SAD-u i statistikama raznih kibernetičkih izvješća. Iza SAD-a slijede Kina (2 661), Japan (1 579), Njemačka (979) i Južna Koreja (823). Broj sveukupnih stranica proporcionalna je s brojem zlonamjernih stranica po zemlji. Hrvatskoj pripada 30 zlonamjernih stranica.

Na Slika 1.6 prikazan je omjer zlonamjernih prema benignim stranicama u logaritamskoj skali za prvih 10 zemalja s najvećim omjerom čiji je broj benignih stranica prelazio tisuću.



Slika 1.6: Prvih 10 zemalja s najvećim omjerom zlonamjernih i benignih stranica

Veći omjer je nepovoljniji jer to znači da na manji broj benignih stranica ima veći broj zlonamjernih. Zemlja s najvećim omjerom zlonamjernih prema benignim stranicama jest Italija (0.024767), zatim Kanada (0.024181), Kina (0.023973), Njemačka (0.023796) itd. Omjer Sjedinjenih Američkih Država jest 0.023276. U usporedbi s ukupnim broj zlonamjernih stranica, SAD se ovaj puta ne nalazi na prvom mjestu iako ima pet puta veći broj zlonamjernih stranica od Kine, sljedeće zemlje s najvećim brojem zlonamjernih stranica. Naime, ova statistika ne mora značiti da je internetski prostor određene zemlje nesigurniji od ostalih zbog načina rada *Web Crawler*-a. *Web Crawler* može ući u internetski potprostor određene zemlje gdje se nalazi veći broj zlonamjernih web stranica, a u drugim slučajevima, zbog veza između stranica, ostati većinom u benignom prostoru internetske domene drugih zemalja.

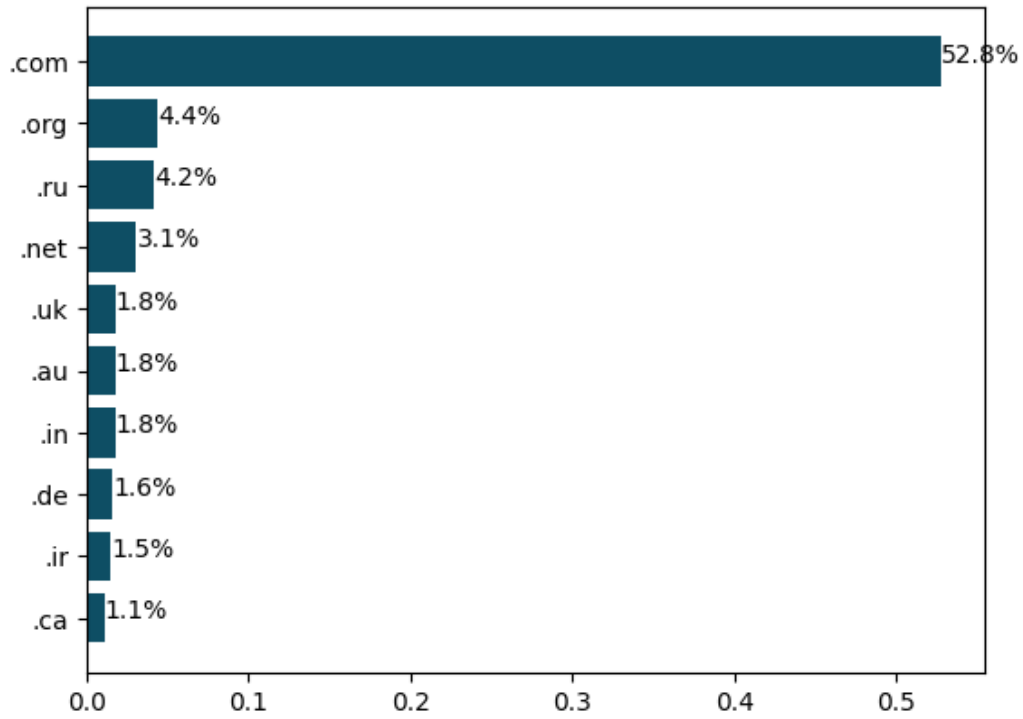
1.1.4. Vršna domena

Vršna domena (engl. *Top Level Domain*, skr. *TLD*) je najviša razina u internetskom hijerarhijskom sustavu naziva domena (engl. *Domain Name System*, skr. *DNS*) [15]. Vršne domene URL-a u Ispisu 1.1 su redom: com, hr, com.hr.

Ispis 1.1: Tri URL.a s navedenim vršnim domenama

```
https://www.google.com
https://www.virtuabit.hr
https://jolies.com.hr
```

Najpopularnije vršne domene u svijetu od lipnja 2022. prikazane su na Slika 1.7 [16]. Domena .com čini pola svih registriranih vršnih domena u svijetu (52.8%). Iza nje slijedi još jedna neutralna domena .org s 4.4%, zatim se pojavljuje prva domena vezana uz domenski prostor neke zemlje što je u ovom slučaju Rusija s .ru (4.2%). .net domena čini 3.1% svih svjetskih vršnih domena. Od najpopularnijih vršnih domena vezane uz domenski prostor neke zemlje su redom: Rusija (4.2%), Ujedinjeno Kraljevstvo (1.8%), Australija (1.8%), Indija (1.7%), Njemačka (1.6%), Iran (1.5%), Kanada (1.1%).



Slika 1.7: Najrasprostranjenije vršne domene u lipnju 2022. godine

Vršne domene imaju ulogu kategorizacije web stranice po tipu, lokaciji ili poslovnom modelu[17]. .com je najrasprostranjenija domena i prema tome ima najveću vrijednost za sve profesionalne web stranice te obuhvaća sve vrste web stranica. .org nosi značenje „organizacija“ i primarno se koristi za neprofitne web stranice poput dobrotvornih organizacija, obrazovne platforme i slično. .net domena označava mrežu (engl. *Network*) i koristi se za tvrtke koje pružaju usluge interneta, web hostinga, pohranjivanje i održavanje baze podataka ili kolaboracijskih alata [17].

Istraživači iz tvrtke *Palo Alto Networks* napravili su istraživanje o vršnim domenama s najvećom količinom distribucije zlonamjernog sadržaja [18]. Istraživači su proučavali stopu zloupotrebljavanja (engl. *Abuse*) najrasprostranjenijih vršnih domena te kumulativne distribucije [19] (engl. *Cumulative Distribution, skr. CD*) vršnih domena po određenim kategorijama zloupotrebljavanja. U Tablica 1.1 prikazane su najveće vršne domene i njihove kumulativne distribucije za razne kategorije zloupotrebe. Kumulativna distribucija izračunata je pomoću kumulativne distributivne funkcije [19]. Kumulativna distributivna funkcija definira je jednadžbom (1) gdje je $F_X(x)$ funkcija od X, X varijabla u skupu realnih brojeva, te P vjerojatnost da će X biti manji ili jednak x.

$$F_X(x) = P(X \leq x), \forall\, x\ \in \mathbb{R} \tag{1}$$

Drugim riječima, u Tablica 1.1 kumulativna distribucija vršne domene .com za phishing stranice jest 0.49, a vršne domene .top 0.66. Iako je phishing stranica s vršnom domenom .com mnogo više od phishing stranica s vršnom domenom .top, udio phishing stranica s .top domenom naspram svih stranica s .top domenom je veći od udjela phishing stranica s .com domenom naspram svih stranica s .com domenom.

Tablica 1.1: CD raznih kategorije kibernetičke zlouporabe za najrasprostranjenije vršne domene

Total		Malicious		Phishing		Malware		Grayware	
TLD	CD	TLD	CD	TLD	CD	TLD	CD	TLD	CD
com	0.47	com	0.49	com	0.41	com	0.51	xyz	0.38
net	0.51	icu	0.53	xyz	0.48	icu	0.56	com	0.68
de	0.55	xyz	0.58	tk	0.52	cn	0.61	tokyo	0.71
org	0.58	cn	0.62	ml	0.55	net	0.66	club	0.73
tk	0.61	net	0.66	cf	0.59	ml	0.70	net	0.75
uk	0.63	ml	0.70	icu	0.61	org	0.72	work	0.76
cn	0.65	tk	0.73	ga	0.64	tk	0.75	ru	0.77
ru	0.67	org	0.75	top	0.66	cf	0.77	co	0.79
icu	0.68	cf	0.77	pw	0.68	xyz	0.79	info	0.80
xyz	0.70	ga	0.79	net	0.70	ga	0.80	org	0.81

Prema navedenom istraživanju vršna domena .com je odgovorna za skoro pola svih registriranih domena te posljedično i pola svih zlonamjernih domena što i dalje ne čini .com domenu nesigurnom. Također, iz Tablica 1.1 može se iščitati da su domene poput .pw, .ml, .club, .cf i .top u prvih 10 po kumulativnoj distribuciji u nekih od kategorija, a nisu ni blizu prvih 10 po veličini određene vršne domene što je jasan pokazatelj sklonosti kriminalnih aktivnosti pod tom domenom. Autori još naznačuju da se velike domene poput .de i .uk ne pojavljuju u prvih 10 domena niti u jednoj od kategorija što ukazuje na redovito kontroliranje i strogo provjeravanje operatora navedenih domena.

Tablica 1.2 prikazuje vršne domene i s najvišim stopama zlonamjernih domena koja se računa kao medijan apsolutne devijacije (engl. *Median Absolute Deviation*, skr. *MAD*) [20]. U ovom slučaju se MAD računa kao omjer zlonamjernih stranica prema svim domenama

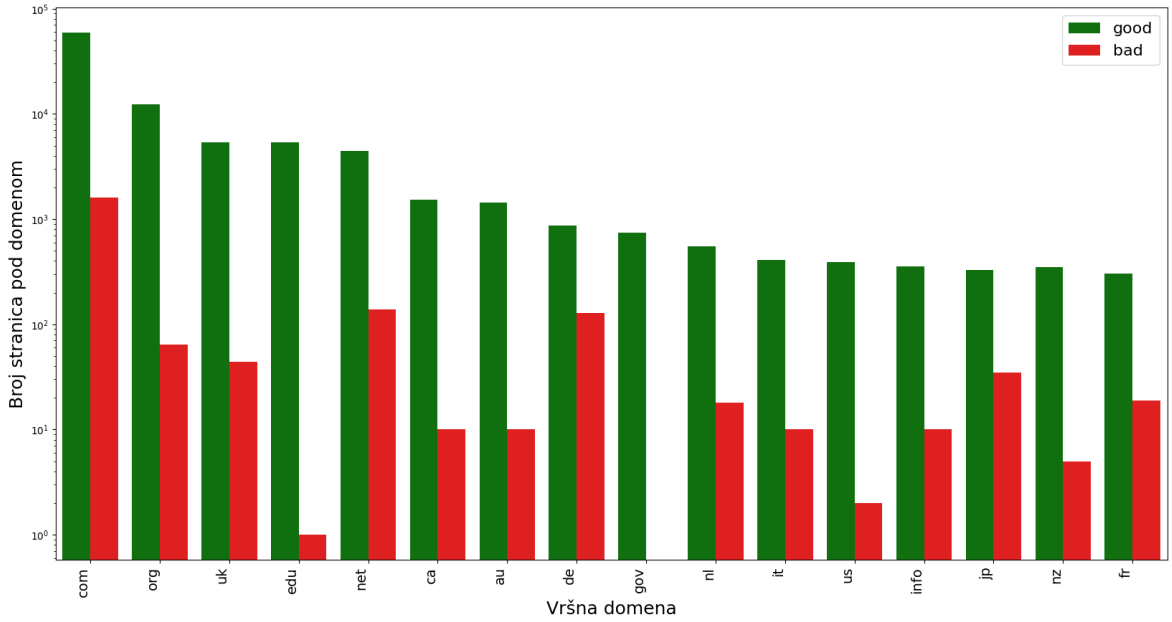
neke vršne domene. Veća stopa označava veću kriminalnu aktivnost u toj kategoriji za neku vršnu domenu.

Tablica 1.2: Vršne domene s najvišim stopama zlonamjernih domena

Malicious		Phishing		Malware		Grayware		C2	
TLD	MAD	TLD	MAD	TLD	MAD	TLD	MAD	TLD	MAD
zw	30.37	pw	43.48	zw	38.05	sbs	89.66	cyou	7.95
bd	26.18	quest	32.00	bd	30.98	tokyo	66.08	pw	6.72
ke	25.38	ke	17.28	ke	28.69	xyz	40.94	ws	4.25
am	18.48	date	15.47	am	24.46	cam	21.21	gq	4.03
sbs	17.58	cyou	13.80	cd	16.07	date	18.56	cf	3.84
date	15.38	support	11.38	date	13.12	cm	16.21	ml	3.81
pw	13.35	win	8.55	bid	12.81	casa	15.78	ga	3.36
quest	11.92	rest	7.14	ml	12.00	uno	11.17	info	2.93
cd	11.88	casa	6.45	ws	10.68	email	8.39	su	2.74
bid	10.96	help	5.47	icu	9.08	stream	7.38	best	2.44

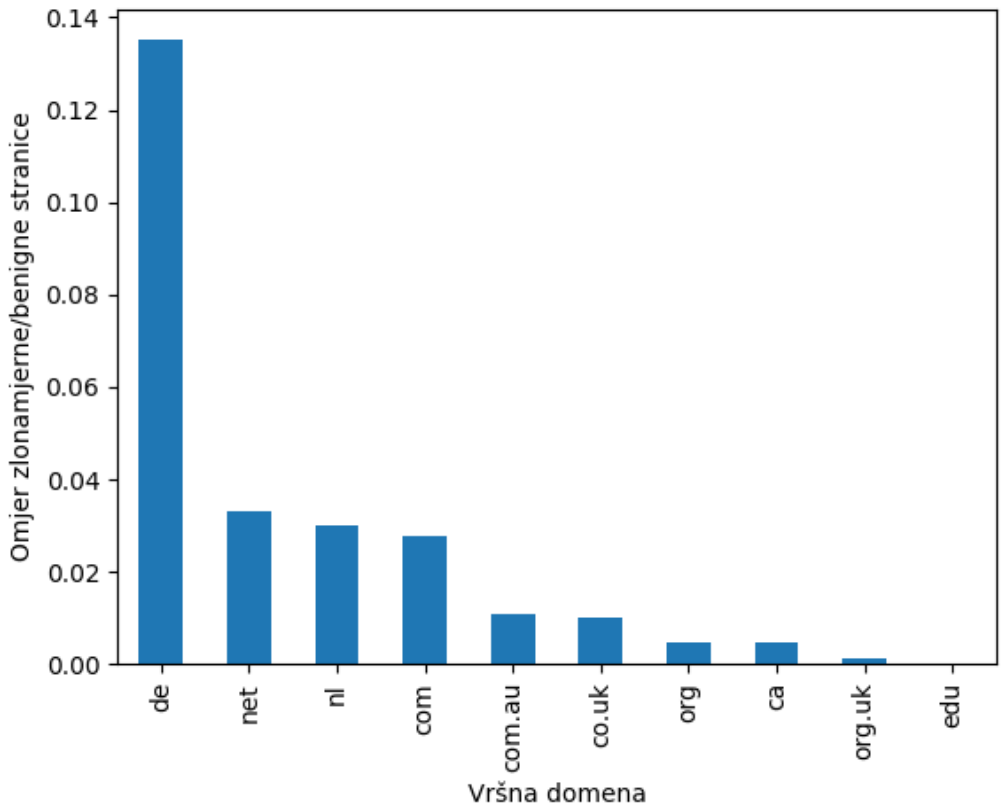
Vršne domene s velikim udjelom zlonamjernih stranica su .pw, .zw, .ke, i bd. što se može očitati iz Tablica 1.2.

Odnosi zlonamjernih i benignih stranica u *MalCrawler* skupu podataka prikazane su vizualno na Slika 1.8 i Slika 1.9. Na Slika 1.8 nalaze se stupčasti prikazi omjera zlonamjernih i benignih stranica za neke od odabranih domena u logaritamskoj skali.



Slika 1.8: Stupčasti graf omjera benignih i zlonamjernih stranica za vršne domene

Na slici se može očitati nepovoljan omjer klasa stranica za neutralne domene .com, .net i .org te domena vezanih uz domenski prostor neke zemlje kao što su .de i .jp. Ovo ne znači da je internetski prostor tih zemalja nesigurniji nego da je web pretraživač bolje pratio put međusobno povezanih zlonamjernih stranica te ih detektirao većom stopom nego kod drugih zemalja.



Slika 1.9: Prvih 10 vršnih domena s najvećim omjerom zlonamjernih i benignih stranica

Slika 1.9 prikazuje prvih 10 zemalja s najvećim omjerom zlonamjernih i benignih stranica. Prva je domena .de s vrijednošću od 0.135067, zatim .net 0.033094, .nl 0.030056, .com 0.027596, com.au 0.010816 itd.

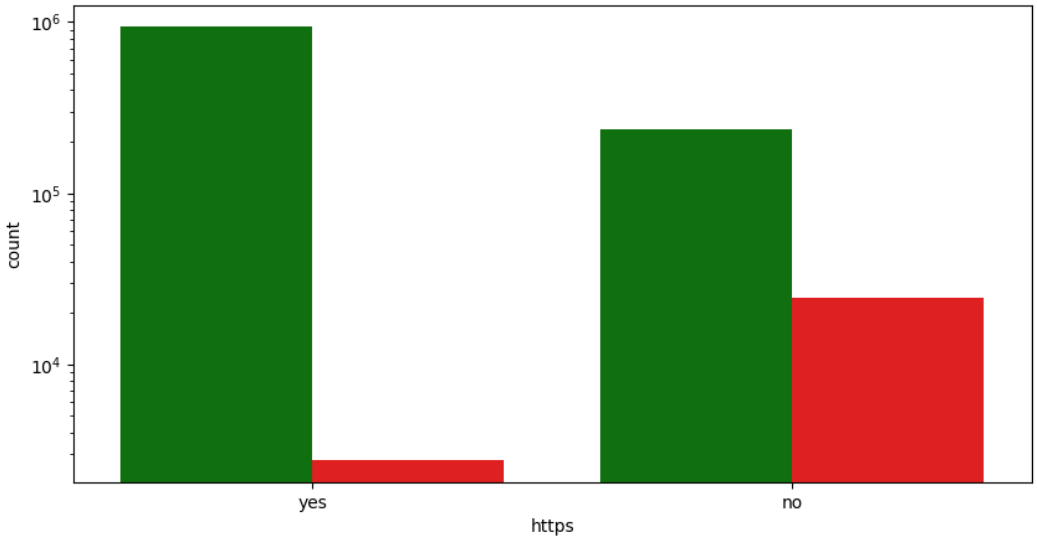
1.1.5. HTTP i HTTPS protokol

HTTPS koristi TLS za enkripciju HTTP zahtjeva i odgovora koje ujedno i digitalno potpiše, stoga je protokol HTTPS sigurniji od protokola HTTP [21]. HTTPS također implicira da web stranica ima legitimni SSL certifikat, dok stranice koji koriste HTTP nemaju SSL certifikat ili im je certifikat istekao. Iako je poželjno, nije nužno da sve stranice koriste HTTPS protokol, no ukoliko stranica obrađuje korisnikove podatke poput osobnih podataka, podataka o kreditnoj kartici i sl., neophodno je da koristi HTTPS protokol [22].

U *MalCrawler* skupu podatka, ova značajka predstavljena je s nazivom HTTPS koja poprima dvije vrijednosti: „yes“, ako stranica koristi HTTPS, te „no“ ako stranica koristi HTTP. U skupu za treniranje 937 824 benignih stranica i 2 752 zlonamjernih stranica koristi

HTTPS protokol što je omjer zlonamjernih prema benignim od 0.0029. HTTP protokol koristi 234 923 benignih i 24 501 zlonamjernih stranica što je omjer od 0.1043.

Na Slika 1.10 prikazana su dva višestruka stupčasta dijagrama (engl. *Multi-Bar Graph*). Dijagram predočava omjere benignih i zlonamjernih stranica koji koriste HTTPS protokol (lijevi stupac) i omjer onih koji ne koriste HTTPS protokol, odnosno koriste HTTP protokol (desni stupac). Stupci crvene boje su zlonamjerne stranice dok su zelenom bojom označene benigne stranice. Broj stranica koje koriste, tj. ne koriste HTTPS protokol su prikazani u logaritamskoj skali.



Slika 1.10: Višestruki stupčasti dijagram omjera benignih i zlonamjernih stranica koje koriste i ne koriste HTTPS protokol

Iz grafa se može zaključiti da korištenje, odnosno ne korištenje HTTPS protokola pridonosi točnijoj klasifikaciji stranice. Mali broj zlonamjernih stranica (795) koristi HTTPS protokol. Omjeri su vrlo različiti te su jasna indikacija da ako stranica koristi HTTP protokol pridonosi predikciji u stranu zlonamjerne stranice, bez obzira na to što velik broj benignih stranica koristi HTTP protokol.

1.1.6. DNS WHOIS informacije

WHOIS je popis internetskih zapisa koji identificiraju vlasnika određene domene te osnovne informacije o tom vlasniku [23]. Internetska korporacija za dodjeljivanje imena i brojeva

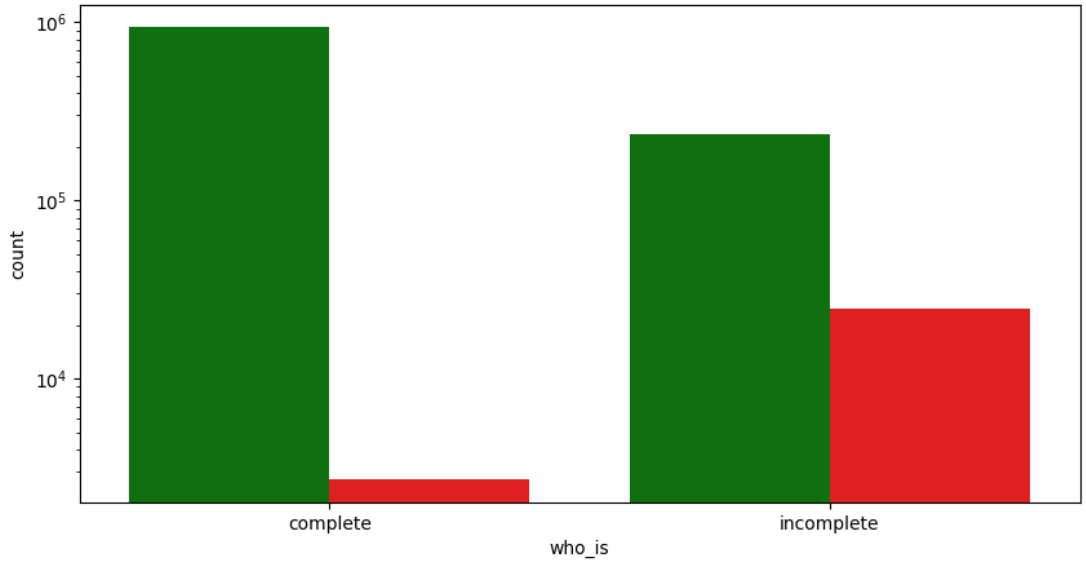
(engl. *The Internet Corporation for Assigned Names and Numbers*, skr. *ICANN*) regulira registraciju naziva i vlasništva domene [24]. WHOIS popis se koristi kao resurs za održavanje integriteta procesa registracije naziva i vlasništva domene [25]. Svaki WHOIS zapis sadržava ime i kontakt informacije registranta [26]. Registrant je vlasnik domene i podnositelj zahtjeva određenom registraru za registraciju domene. Također, WHOIS sadržava ime i kontakt informacije registrara. Registrar je organizacija ili tvrtka koja obrađuje zahtjev za registraciju domene od strane registranta. Registrar postavlja registrantu uvjete na koje on mora pristati prije nego li je domena službeno registrirana. Uloga registrara je vođenje evidencije tehničkih podataka o registrantima u središnjem imeniku zvanom „registrar“. Neki od tehničkih podataka su datum registracije, poslužitelj imena (engl. *Name Server*), datum ažuriranja zapisa te isteka ugovora.

Bez registriranja domene nije moguće dobiti simboličko ime za web stranicu. Bez registracije, jedini način pristupanju web stranice, koja je javno dostupna, je putem IP adrese.

Nekoliko radova [27] [28] je obradilo temu detekcije zlonamjernih URL-ova pomoću informacija koje se mogu pronaći u WHOIS zapisima. U ovom *MalCrawler* skupu podataka koristi se značajka potpunosti odnosno nepotpunosti informacija u WHOIS zapisu tako da se provjerava postoji li ime registrara u zapisu. Stoga ova značajka ima dvije vrijednosti: *complete* i *incomplete*.

U skupu za treniranje, 938 404 benignih stranica i 2 722 zlonamjernih stranica ima potpune WHOIS podatke što je omjer zlonamjernih prema benignim od 0.0029. 234 343 benignih i 24 531 zlonamjernih nema potpune WHOIS podatke što je omjer od 0.1047. Omjeri su skoro pa i jednaki onima za značajku HTTPS protokola.

Na Slika 1.11 prikazana su dva višestruka stupčasta dijagrama (engl. *Multi-Bar Graph*). Graf predočava omjere benignih i zlonamjernih stranica koji imaju potpune WHOIS podatke (lijevi stupac) i omjer onih koji nemaju potpune WHOIS podatke (desni stupac). Stupci crvene boje su zlonamjerne stranice dok su zelenom bojom označene benigne stranice. Broj stranica koji imaju, odnosno nemaju potpune podatke, prikazani su u logaritamskoj skali.



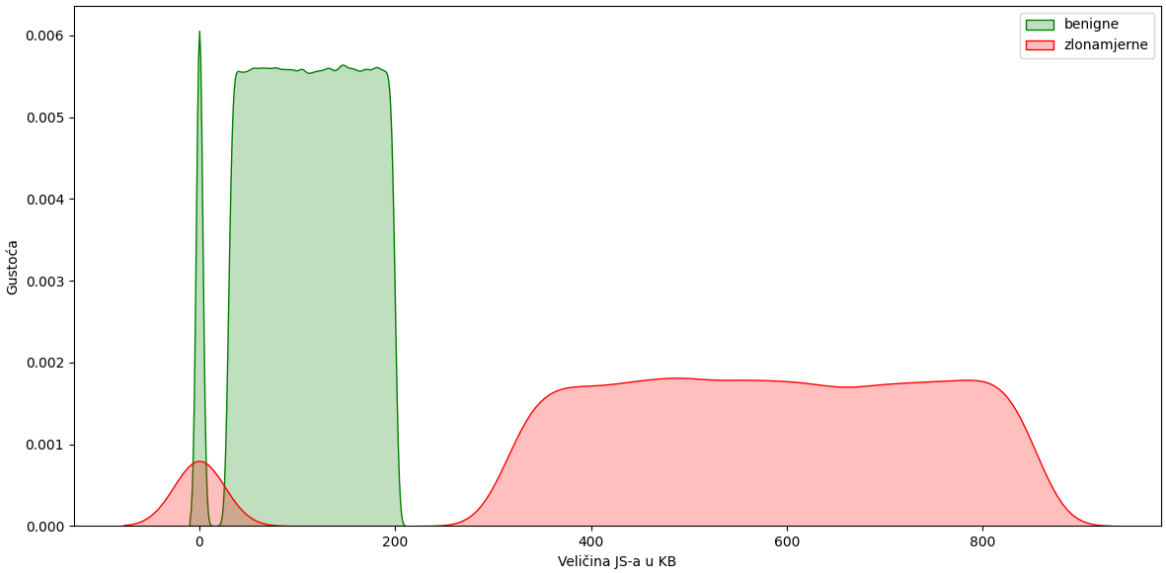
Slika 1.11: Višestruki stupčasti dijagram omjera benignih i zlonamjernih stranica koje imaju potpune i nepotpune WHOIS zapise

Kao i kod HTTPS protokola, pretpostavka je da će potpunost WHOIS informacija pridonijeti točnijoj klasifikaciji stranice. Omjeri su opet vrlo različiti te su indikacija da nepotpunost podataka indikacija zlonamjernosti stranice.

1.1.7. Veličina JavaScript koda

Prema istraživanju [29] zlonamjerni JavaScript (skr. JS) kod se pojavljuje u većim količinama stoga je kao značajka u *MalCrawler* skupu podataka uzeta veličina JavaScript koda na stranicai. U navedenom skupu podataka, prosječna veličina JS-a u zlonamjernim stranicama je 555.98, a u benignim 108.99, što je velika razlika i čini veličinu JS-a odličnim indikatorom zlonamjerne stranice za *MalCrawler* skup podataka.

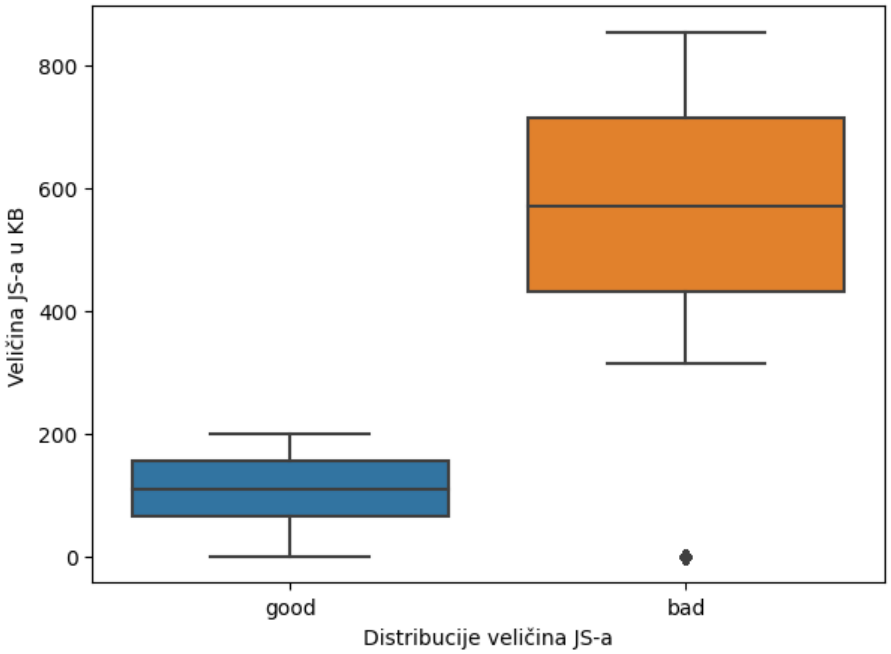
Na Slika 1.12 prikazane su gustoće vjerojatnosti veličine JS-a u KB za benigne i zlonamjerne stranice.



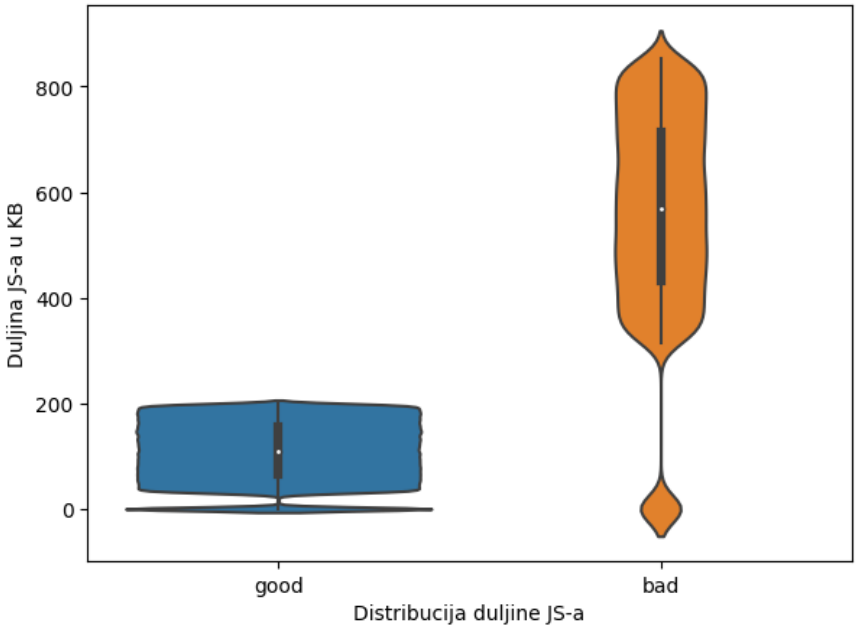
Slika 1.12: Gustoća vjerojatnosti veličine JS-a u KB za benigne i zlonamjerne stranice

Na slici se vrlo lako uočava razlika vrijednosti za navedene klase. Benigne stranice najviše poprimaju vrijednosti od 30 KB do 200 KB veličine JS, dok zlonamjerne stranice poprimaju vrijednosti od 300 KB do 900 KB veličine. Prema prikazanom grafu, čini se kao da je veličina JS-a veća od 300 KB sigurni pokazatelj da se radi o zlonamjernoj stranici. Koliko je to istina bit će odgovoreno u poglavlju s rezultatima algoritama strojnog učenja.

Razlike u distribucijama se još jasnije mogu vidjeti na sljedećim slikama: slika na kojoj je prikazan graf boxplot (Slika 1.13) te slika na kojoj je prikazan graf violine (engl. *Violin Graph*) (Slika 1.14). Graf violine je isto jedan od načina vizualizacije distribucije vrijednosti podataka te se lakše može uočiti frekvencija neke vrijednosti za razliku od boxplota [30].



Slika 1.13: Boxplot graf distribucija veličina JS-a za benigne i zlonamjerne stranice



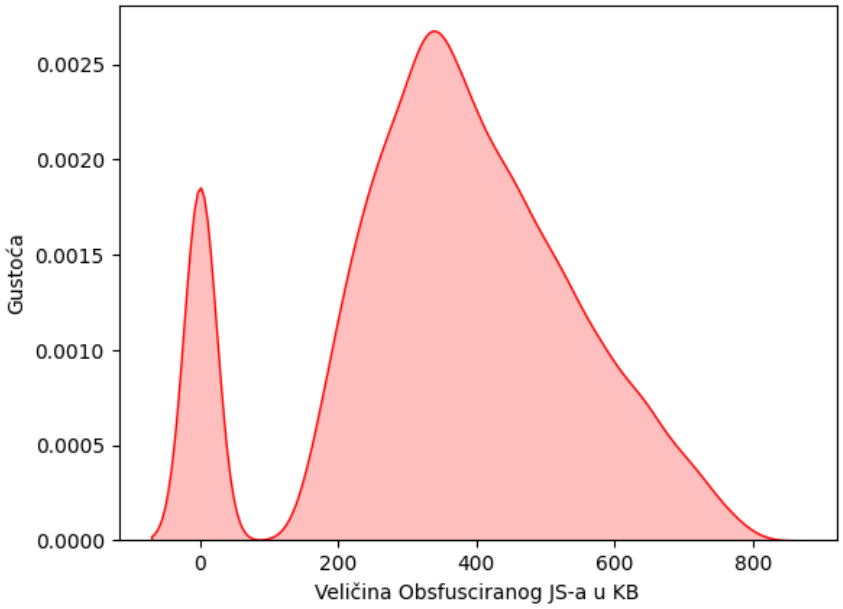
Slika 1.14: Graf zvonca za distribucije veličina JS-a za benigne i zlonamjerne stranice

1.1.8. Veličina obfusciranog JavaScript koda

Obfuskacija koda je proces preoblikovanja koda kako bi se teže razumio, a da funkcionalnost tog koda ostane ista [31]. Obfuskacija koda se koristi kako bi se zaštitilo intelektualno vlasništvo, ali ga koriste i kibernetički kriminalci kako bi prikrili zlonamjerne aktivnosti, odnosno kako bi takve aktivnosti bile teže otkrivene i blokirane [32]. Prema istraživanju [33], 26% od 10 000 zlonamjernih JavaScript kodova su bili obfuscirani, a od 20 000 najpopularnijih web stranica prema *Alexi.com* otkriveno je 0.05% obfusciranih JavaScript kodova. Stoga, može se reći da se obfuskacija JavaScript koda puno češće koristi kako bi se prikrila zlonamjerna aktivnost napadača.

U *MalCrawler* skupu podataka, čak 24 284 zlonamjernih stranica sadržava obfuscirani JavaScript kod. Ne postoji niti jedna benigna stranica koja sadržava obfuscirani JS kod, stoga će u idućih nekoliko grafova biti prikazana gustoća vjerojatnosti te distribucija veličine obfusciranog JS koda samo zlonamjernih stranica. Iz tog podatka lako je zaključiti da su jako male šanse da stranica bude zlonamjerna ako nema obfusciranog JS koda, a s druge strane bilo kakva količina obfusciranog JS koda je veliki indikator zlonamjernosti stranice. Srednja vrijednost veličine obfusciranog JS koda u zlonamjernim stranicama jest 359.01 KB.

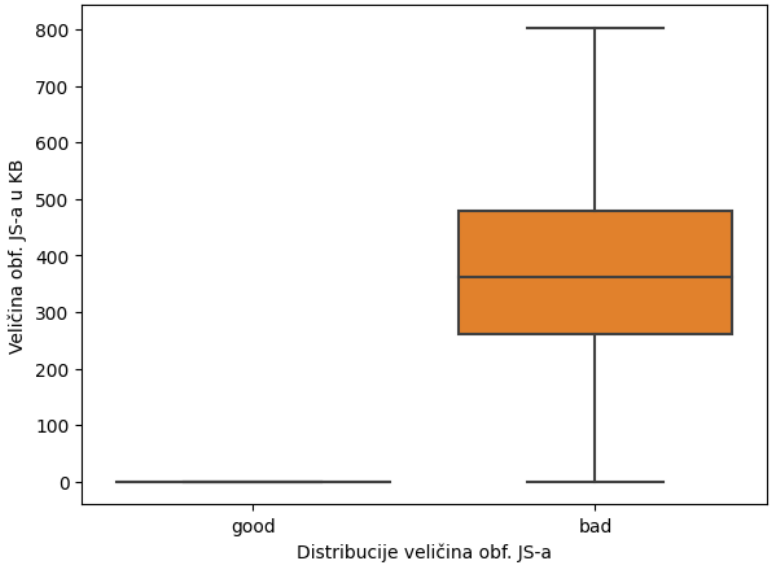
Iz Slika 1.15 vidljiv je graf gustoće vjerojatnosti veličine obfusciranog JS koda za zlonamjerne stranice.



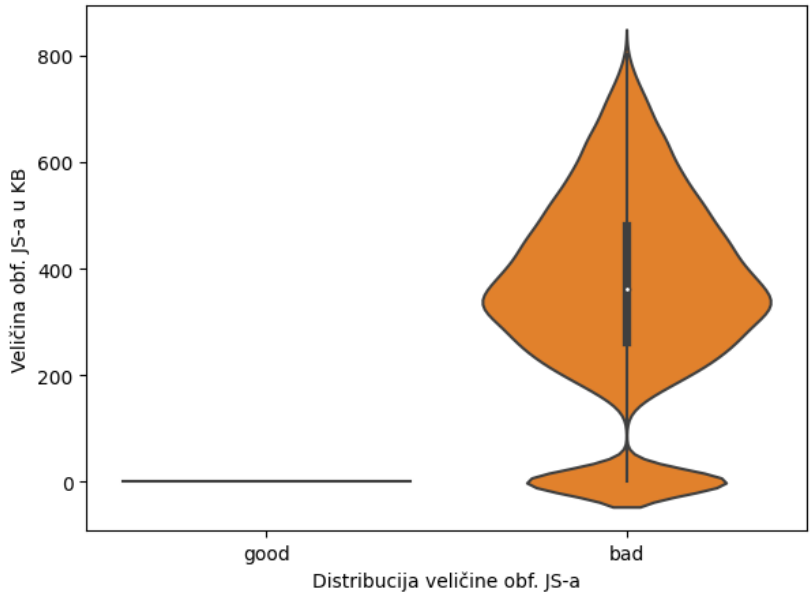
Slika 1.15: Graf gustoće vjerojatnosti za veličinu obfusciranog JS-a u KB zlonamjernih stranica

Iz slike uočavamo da je najveća gustoća oko 380 KB obfusciranog koda. Veličina JS koda u zlonamjernim stranicama te veličina obfusciranog JS koda u zlonamjernih stranicama jest slična po rasponu, od 200 do 800 KB, no razlikuju se u distribuciji. Dok je veličina JS koda podjednako distribuirana, distribucija obfusciranog koda ima piramidalnu strukturu gdje je vrh oko 380 KB.

Distribuciju vrijednosti veličine koda još je jasnija na grafovima na Slika 1.16 i Slika 1.17.



Slika 1.16: Boxplot graf distribucija veličina obfusciranog JS-a za benigne i zlonamjerne stranice



Slika 1.17: Graf zvonca za distribucije veličina obfusciranog JS-a za benigne i zlonamjerne stranice

Time su pokrivene sve numeričke vrijednosti unutar *MalCrawler* skupa podataka: Duljina URL-a, veličina JS koda i veličina obfusciranog JS koda. Usporedba tih vrijednosti dati će uvid u međusobnu dinamiku tih vrijednosti prikazanim u Tablica 1.3.

Tablica 1.3: Statističke vrijednosti numeričkih značajki *MalCrawler* skupa podataka

	Statistika benignih stranica			Statistika zlonamjernih stranica		
	url_len	js_len	js_obf_len	url_len	js_len	js_obf_len
count	1172747.0	1172747.0	1172747.0	27253.0	27253.0	27253.0
mean	35.82	108.99	0.0	37.15	555.98	359.01
std	14.42	53.97	0.0	14.02	199.36	180.63
min	12.0	0.0	0.0	13.0	0.0	0.0
25%	26.0	65.5	0.0	27.0	431.1	261.86
50%	32.0	110.0	0.0	33.0	569.7	361.35
75%	42.0	155.0	0.0	45.0	714.6	478.86
max	721.0	199.5	0.0	416.0	854.1	802.85

Tablica 1.3 sumira sve zaključke već prethodno donesene. Duljina URL-a se ne razlikuje između benignih i zlonamjernih dok su vrijednosti veličina koda JS-a i obfusciranog JS-a puno veće kod zlonamjernih stranica. Odnose numeričkih atributa kvantificira se korelacijskom Tablica 1.4.

Tablica 1.4: Korelacijska matrica numeričkih vrijednosti *MalCrawler* skupa podataka

	url_len	js_len	js_obf_len
url_len	1.000000	0.010398	0.012223
js_len	0.010398	1.000000	0.778568
js_obf_len	0.012223	0.778568	1.000000

Rezultati korelacijske tablice daju visoku korelaciju između veličine JS-a i obfusciranog JS-a te vrlo malu korelaciju između duljine URL-a i ostale dvije značajke. Dubljim proučavanjem korelacije, takva korelacija može se objasniti prethodno referenciranim istraživanjima, no također se mora napomenuti da autori rada nigdje nisu konkretno naveli

da obfuscirani JS ne ulazi u ukupan broj veličine JS-a, stoga visoku korelaciju treba razmatrati s oprezom.

2. Rezultati strojnog učenja

Skup podataka provučen je kroz nekoliko algoritama strojnog učenja koji su pogodni za problem binarne klasifikacije. Odabrani algoritmi strojnog učenja su:

- Gaussov naivni Bayes (engl. *Gaussian Naive Bayes*)
- Multinomijalni naivni Bayes (engl. *Multinomial Naive Bayes*)
- Komplementarni naivni Bayes (engl. *Complement Naive Bayes*)
- Stablo odluke C4.5 (engl. *Decision Tree C4.5*)
- Nasumična šuma (engl. *Random Forest*)
- Stroj potpornih vektora (engl. *Support Vector Machine*, skr. *SVM*)

Implementacije algoritama se koriste iz *scikit-learn*[34] Python biblioteke.

Uspješnost algoritama izražava se sljedećim vrijednostima:

- točnost (engl. *Accuracy*)
- preciznost (engl. *Precision*)
- odziv (engl. *Recall*)
- F1 mjera (engl. *F1 Score*)
- ROC AUC (engl. *Area Under the Receiver Operating Characteristic Curve*)

Cilj učenja algoritama je minimizirati broj stranica koje su označene kao lažno benigne i maksimizirati broj točno zlonamjernih. Stoga će prethodno navedeni algoritmi strojnog učenja biti konfigurirani tako da daju što manje lažno benignih klasifikacija, odnosno njihovi hiperparametri će biti podešeni kako bi se takva klasifikacija minimizirala [35].

Kombinacija hiperparametara napravljena je pomoću *scikit-learn* alata *HalvingRandomSearchCV* [36] koji koristi nasumičnu pretragu najboljih vrijednosti za sve željene hiperparametre sa strategijom da na početku procjeni sve hiperparametre s malim brojem resursa te iterativno odabire najbolje hiperparametre koristeći sve više i više resursa.

Hiperparametri za stablo odluke (*Decision Tree C4.5*) su sljedeće:

- criterion: entropy
- splitter: best
- max_depth: None ili >= 60
- min_samples_split: 2
- min_samples_leaf: 1
- min_weight_fraction_leaf: 0.00001
- max_features: None
- random_state: 42

- max_leaf_nodes: None ili >= 100
- min_impurity_decrease: 1e-6
- class_weight: {0: 1, 1: 10}
- cpp_alpha: 0.0

Najveći utjecaj i poboljšanje algoritma imali su parametri maksimalne dubine stabla te preddefinirane težine razreda. Podešavanje težina razreda preporučena je metoda kod stabla odluke gdje se podatci nebalansirani po klasama poput *MalCrawler* skupa podataka.

Hiperparametri za nasumičnu šumu su sljedeći:

- n_estimators = 100
- criterion: gini
- max_depth: None ili >= 150
- min_samples_split: 20
- min_samples_leaf: 1
- min_weight_fraction_leaf: 0.0001
- max_features: sqrt
- max_leaf_nodes: None ili >= 200
- min_impurity_decrease: 0.0
- bootstrap: True
- oob_score: True
- n_jobs: 10
- ranom_state: 42
- warm_start: True
- class_weight: {0: 1, 1: 10}
- cpp_alpha: 0.0
- max_samples: None

Najveći utjecaj na poboljšanje klasifikacije nasumične šume imao je broj stabala. Za razliku od stabla odluke, postavljanje težina klasa u nasumičnoj šumi višestruko je pogoršalo klasifikaciju.

Hiperparametri za Gaussov naivni Bayes:

- var_smoothing: 1e-5

Hiperparametri za multinomijalni naivni Bayes:

- alpha: 1.0
- fit_prior: True
- class_prior: None

Postavljanje alfe nije imalo nikakvog utjecaja na multinomijalni naivni Bayes.

Hiperparametri za komplementarni naivni Bayes:

- alpha: 1.0
- fit_prior: True
- class_prior: None
- norm = True

Hiperparametri za stroj potpornih vektora:

- gamma: auto
- kernel: rbf

Algoritmi su računati i testirani na skupu od 500 000 podataka s prijelomom od 0.33 u korist treniranja. U Tablica 2.1 prikazani su uspješnosti svih algoritama. Lažno benigni su označeni s FP, dok su lažno zlonamjerni s FN.

Tablica 2.1: Mjere uspješnosti klasifikacije algoritama strojnog učenja

	FN	FP	točnost	preciznost	odziv	F1 mjera	ROC AUC
Stablo odluke	15	23	0.99884	0.97981	0.96937	0.97456	0.98445
Nasumična šuma	9	28	0.99887	0.98770	0.96271	0.97505	0.98121
Gaussov Bayes	2	36	0.99884	0.99721	0.95206	0.97411	0.97600
Multinomijalni Bayes	0	77	0.99766	1.0	0.89747	0.94596	0.94873
Komplementarni Bayes	0	77	0.99766	1.0	0.89747	0.94596	0.94873
SVC	0	181	0.99451	1.0	0.75898	0.86298	0.87949

Svi algoritmi imaju visoku točnost, no mjera kojoj treba posvetiti najviše pozornosti je odziv jer nam on govori koliko uspješno su zlonamjerne stranice bile klasificirane zlonamjerno. Kod algoritama multinomijalnog i komplementarnog Bayesa te SVM-a imaju relativno visok broj lažno benignih klasifikacija te niži odziv, F1 mjeru i ROC AUC od ostala tri algoritma. Gaussov Bayes, stablo odluke te nasumična šuma imaju visoki odziv. Stablo odluke daje najbolje rezultate, razlog tome jest iscrpno pretraživanje kombinacija hiperparametara pomoću funkcije *HalvingRandomSearchCV* gdje je parametar *scoring* postavljen na odziv kako bi se minimizirao broj lažno benignih.

3. Pretprocesiranje i obrada podataka

U ovom poglavlju bit će prikazani načini pomoću kojih se došlo do značajki navedenih u prethodnom poglavljima. Većina podataka dobiveno je iz *WebSecRadar* baze podataka koja se nalazi na Zavodu za elektroniku, mikroelektroniku, računalne i inteligentne sustave na Fakultetu Elektrotehnike i Računarstva Sveučilišta u Zagrebu. No, u ovom poglavlju su prikazani i načini dobivanja istih podataka javno dostupnim resursima i servisima.

IPv4 adresa za neku domenu se nalazi pod kolekcijom *hostnames* pod atributom *dnsdata.A*. Javno dostupni način konverzije URL-a u IPv4 i IPv6 adresu postiže se bibliotekom *socket* koja pripada standardnim Python bibliotekama. Poziva se naredba *socket.gethostbyname(hostname)* gdje je *hostname* ime domaćina (hosta) web stranice.

Za zemljopisnu lokaciju URL-a koristi se Python biblioteka *geo2ip* [37] te biblioteka *maxminddb* [38] tvrtke *MaxMind* te baze podataka *GeoLite2-City.mmdb* i *GeoLite2-Country.mmdb* [39] u kojem su IP dometi gradova odnosno zemalja. U Ispis 3.1 nalazi se isječak koda korišten za dobivanje grada u kojem je domaćin tražene stranice.

Ispis 3.1: Korištenja biblioteke maxmindb za dobivanje zemljopisne lokacije IP adrese

```
import maxminddb
with maxminddb.open_database('GeoLite2-City.mmdb') as reader:
    inf = reader.get('161.53.78.35')
    print(inf['city']['names']['en'])
import geoip2.database
with geoip2.database.Reader('GeoLite2-City.mmdb') as reader:
     response = reader.city('203.0.113.0')
     response.country.name
```

Drugi način dolaska do ove informacije je korištenje usluge *ipwhois.io* [40] te krajnje točke *http://ipwho.is/ip_add* . Primjer korištenja *ipwhois* usluge dana je u Ispis 3.2.

Ispis 3.2: Korištenja stranice *ipwhois.io* za dobivanje zemljopisne lokacije IP adrese

```
ipwho_url = 'http://ipwho.is/'+ '161.53.68.5'
city = requests.get(ipwho_url).json()['city']
```

Vršna domena URL-a dobiva se pomoću Python biblioteke *tld* [41] te funkcija *get_tld* i *get_fld*. Biblioteka *tld* izdvaja vršnu domenu iz navedenog URL-a pomoću popisa javnih sufiksa „https://publicsuffix.org/list/public_suffix_list.dat“. Ukoliko vršna domena ne postoji u popisu sufiksa, vraća se greška. Primjer korištenja biblioteke i njenih funkcija *get_tld* i *get_fld* dan je u Ispisu 3.3.

Ispis 3.3: Demonstracija funckija *get_tld* i *get_fld* Python biblioteke *tld*

```
from tld import get_tld, get_fld
get_tld(„http://www.google.co.uk“)
# 'co.uk'
get_tld(„http://www.google.idontexist“, fail_silently=True)
# None
get_fld(www.google.co.uk, fix_protocol=True)
# 'google.co.uk'
```

DNS WHOIS informacije nalaze se u kolekciji *domainnames* pod atributom *whois*. Javno dostupni način dobivanja WHOIS zapisa je pomoću Python biblioteke *python-whois*[42]. No treba biti oprezan pri korištenju paketa jer će nakon pet brzih upita za sve sljedeće upite dojavljivati grešku. Problem je u tome koliko određeni WHOIS server dopušta upita zaredom i u određenom vremenskom periodu. Primjer korištenja biblioteke *python-whois* dan je u Ispisu 3.4.

Ispis 3.4: Demonstracija biblioteke *python-whois*

```
import whois
w = whois.whois('test.hr')
```

Drugi način je korištenje *WhoisXML API* [43] usluge gdje će se pri registraciji dobiti API ključ s 500 poziva, što je vrlo limitirajuće jer se rade upiti nad skoro 100 000 URL-a te bi trebali 200 računa napraviti kako bi se mogle dobit informacije od svih domena. Treća opcija je korištenje WHOIS alata integriranog u Linux operacijske sustave koji dopušta 10 upita po minuti.

Drugi problem s dobivanjem WHOIS zapisa je nekonzistentnost rezultata, pa je prema tome parsiranje i traženje određenih informacija iz zapisa teško. Svaka domena ima nadležnu službu za upravljanje domenom svoje zemlje (za Hrvatsku je to CARNET) i svaka ustanova pohranjuje WHOIS informacije na drugačiji način. Pošto je većina proučavanih URL-a pod

hr domenom, ovaj problem nije toliko utjecao na dobivanje podataka, no sve ostale domene su se morale ručno pregledati kako ne bi došlo do netočnih informacija ili kako se u podatcima ne bi zapisalo za određeni URL da ima nepotpune WHOIS informacije, a u stvarnosti ima.

Kako je već spomenuto, većina URL-ova je pod hr domenom. Točnije, tri su vršne domene koje su pod nadležnosti CARNET-a: .hr, .com.hr, .from.hr. WHOIS zapis za te tri domene može se pronaći upisivanjem tražene domene na stranicu domene .hr [44] koja je prikazana na Slika 3.1.

Slika 3.1: Rezultati upita „test.hr“ na stranici domene.hr

Kako bi se ubrzao proces prikupljanja WHOIS zapisa, napravljena je *Selenium* skripta u programskom jeziku Python [45] koja upisuje domenu u kućicu za pretraživanje na *domene.hr* stranici te preuzima podatke iz iskočnog prozora. Broj upita sa *domene.hr* stranice je također 10 po minuti.

3.1. Detekcija obfuskacije JavaScript koda

Detekcija obfuskacije JS koda je najkompleksniji dio ovoga rada, ali i najkritičniji zbog njegovog utjecaja na uspješnost detekcije zlonamjernih stranica. Nakon višemjesečnog istraživanja nije pronađen niti jedan online alat koji jednoznačno određuje je li neki JavaScript kod obfsuciran. Mnogi radovi se bave pitanjem detekcije obfuskacije JavaScript

koda te se neki radovi čak bave i njegovom deobfuskacijom, no svi radovi se temelje na algoritmima strojnog učenja s kompleksnim parametrima i značajkama koji bi sami po sebi mogli biti zasebni diplomski rad. Stoga se koristi završni rad sveučilišnog prvostupnika računalnog inženjerstva Tonija Dujmovića za detekciju obfusciranog JavaScript koda.

Rješenje Tonija Dujmovića temelji se na regularnim izrazima (engl. *regex*) i računanju Shannonove entropije [46]. Regexi su podijeljeni tako da razlikuju minifikaciju i obfuskaciju. Svaki pronađeni regex povećava indeks sigurnosti za jedan te u kombinaciji s izračunatom entropijom danog JavaScript koda određuje se je li JS kod nepromijenjen, minificiran ili obfusciran. Popis regexa koji detektiraju obfuskaciju dan je u Ispis 3.5.

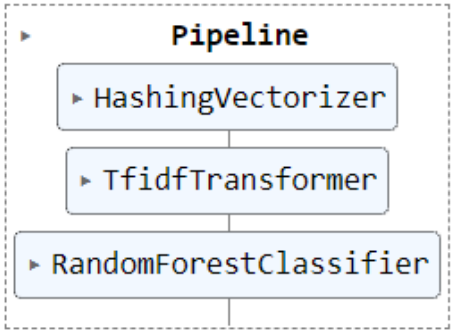
Ispis 3.5: Popisi regexa za detekciju obfuskacije JavaScripta

```
r"\(0x[1-9][0-9]+\)",
r"_0x([0-9]|[a-z])+=_0x([0-9]|[a-z])+",
r"\(0x[1-9]+\)\]\(",
r"\function\(_0x([0-9]|[a-z])+,_0([0-9]|[a-z])+",
r"while\(\!\!\[\]\)",
r"\'0x[0-9]+\'",
r"_0x([0-9]|[a-z])+\+\+",
r"_([a-z]+)\[[0-9]*\]\+_\1\[[0-9]+\]",
r"eval\(function\(p,a,c,k,e,d\)",
r"eval\(function\(p,a,c,k,e,r\)",
r"return !!\[\]",
r'\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|\|',
r"\[\]\+\[\]"
```

Osim toga je razvijen jednostavan model strojnog učenja za detekciju obfuskacije JavaScript koda. Problem ovog pristupa je prikupljanje obfusciranog JS-a kojeg ima vrlo malo javno dostupnog. Pronađen je jedan skup podataka s normalnim i obfusciranim JS kodom na platformi Kaggle [47]. U skupu podataka se nalazi 1 898 primjera normalnog JS koda te 1 477 primjera obfusciranog JS-a. Većina radova koji su obradili temu detekcije obfuskacije JS-a strojnim učenjem [48]–[51] se temelje na izvlačenju značajki iz samog koda JS-a, no način koji je istražen u ovom radu temelji se na vektorizatoru sažetka (engl. *Hashing Vectorizer*) [52] koji pretvara kolekciju tekstualnih dokumenata u matricu pojavljivanja tokena (engl. *Matrix of Token Occurrences*). *HashingVectorizer* je po svom principu rada vrlo sličan *CountVectorizer*-u, jedina razlika je ta da *HashingVectorizer* ne sprema vokabular unikatnih tokena [53]. Time je *HashingVectorizer* idealan za ovaj problem pošto je vrlo fleksibilan na raznovrsnost tekstova (u ovom slučaju JS koda) i za rad s obfuskacijom

pošto se kod obfuskacije najčešće radi o vrlo dugačkim nizovima znakova. Zatim *TfidfTransformer* [54] pretvara matricu pojavljivanja tokena u učestalost pojma (engl. *Term Frequency*, skr. *TF*) te učestalost pojma kroz više dokumenata (engl. *Term Frequency – Inverse Document Frequency*). *TfidfTransformer* pridodaje svakom tokenu izdvojenom putem *HashingVectorizer*-a određenu težinu koja se zatim koristi za klasifikaciju dokumenata. Za konačnu klasifikaciju JS koda odabrana je nasumična šuma (engl. *Random Forest*).

Shema konstruiranog cjevovoda za klasifikaciju JS-koda prikazan je na Slika 3.2.



Slika 3.2: Shema algoritma za klasifikaciju obfuskacije JavaScript koda

Treniranje i testiranje navedenog algoritma provedeno je na već spomenutim skupom podataka te je napravljena podjela od 0.67 : 0.33 u korist treniranja.

Točnost algoritma jest 0.96319, preciznost 0.94512, odziv 0.97077, F1 mjera 0.95777 te AUC ROC 0.96412. Ovo su vrlo dobri rezultati s obzirom na kompleksnost problema, ali tek će testiranje nad stranicama u hr domeni pokazati stvarnu korisnost ovog algoritma. Pri treniranju ovog algoritma nisu u obzir uzeti minificirani JS kod što će se kasnije ispostaviti kao veliki problem i izvor mnogih lažnih pozitivnih klasifikacija.

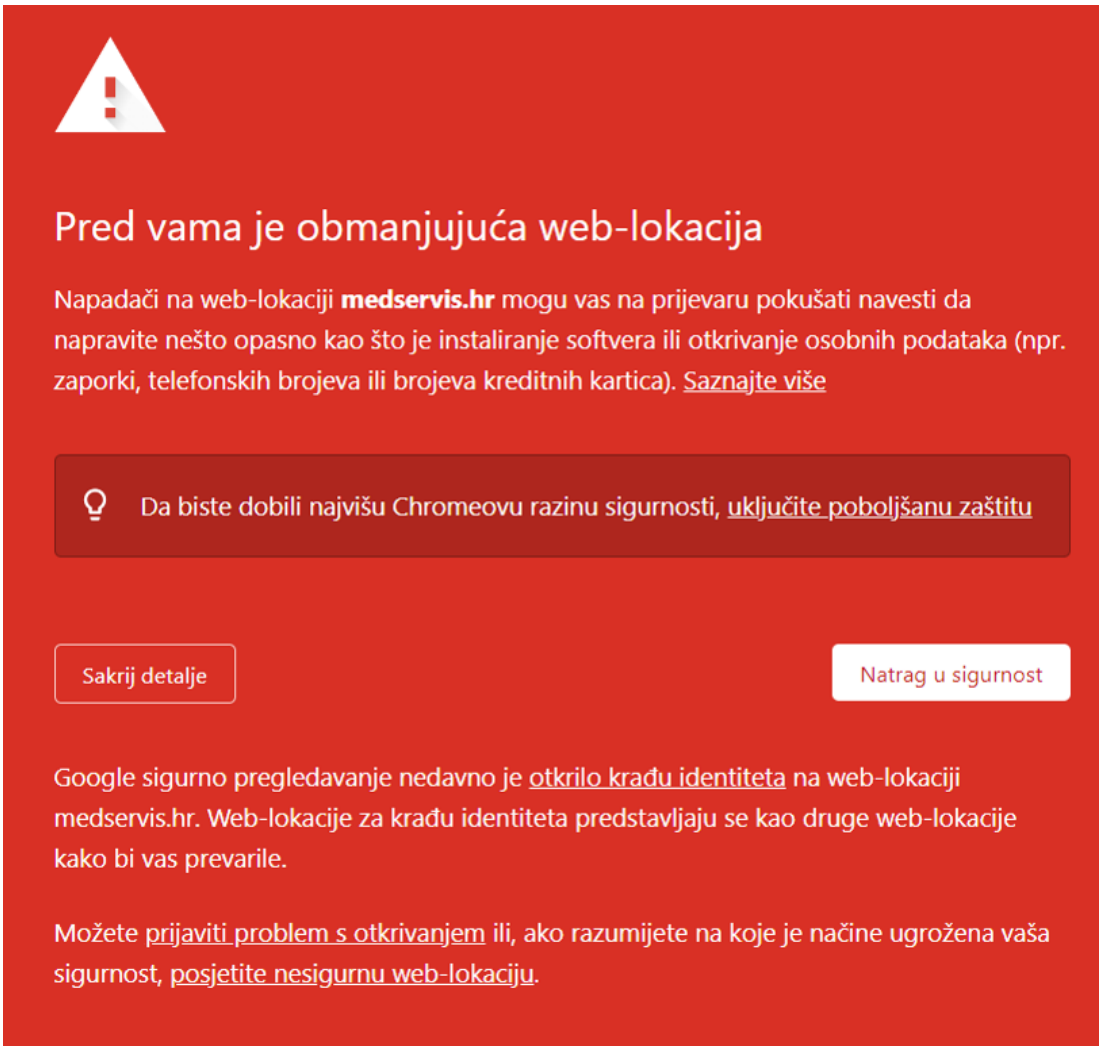
Treći način detekcije obfuskacije je Adobeov algoritam za detekciju obfuskacije [55]. Tvrtka Adobe razvila je duboku konvolucijsku neuronsku mrežu za detekciju obfusciranih komandnih linija, no može se koristiti za detekciju obfuskacije bilo kakvog tipa. No problem Adobeovog algoritma je vremenska složenost pri detekciji obfuskacije JS koda zbog same veličine JS koda naspram veličine jednolinijskih naredbi u sučelju naredbenog retka.

Sva tri algoritma uzeta su u obzir pri detekciji obfuskacije JS koda. Ako barem dva od tri algoritma detektiraju obfuskaciju, JS kod se smatra obfusciranim. Kasnije se Adobeov algoritam izbacio iz sustava glasanja zbog svoje vremenske složenosti i velike količine JS

datoteka koji su se testirali te je prioritetniji glas dan detekciji s regeximа, a razvijeni algoritam strojnog učenja s nasumičnom šumom se ručno provjeravao i ponovno trenirao s većim skupom podataka.

3.2. Googleov servis za sigurno pretraživanje

Googleov servis za sigurnog pretraživanje [56] (engl. *Google Safe Browsing*) je Googleov servis za analizu i pohranu zlonamjernih URL-ova i datoteka. On omogućava provjeru postojanja URL-a u Googleovoj bazi podataka nesigurnih web resursa te se ta baza podataka konstantno ažurira. Pri ulaska u stranicu *Google Safe Browsing* upozorava korisnika da se radi o stranici s zlonamjernih aktivnostima, a ujedno i sprječava korisnike da stavljaju zlonamjerne linkove na web stranicu na kojoj se nalaze (npr. putem komentara ili chat usluge). Na Slika 3.3 je primjer upozorenja *Safe Browsing* alata pri ulasku na zlonamjerni stranicu *medservis.hr*. Google upozorava da je stranica *medservis.hr* obmanjujuća stranica te da je moguće da stranica na prijevaru pokušava navesti korisnika na davanje osjetljivih podataka.



Slika 3.3: Upozorenje servisa za sigurno pretraživanje pri ulasku u stranicu medservis.hr

Kategorije upozorenja *Safe Browsing* alata su sljedeća: zlonamjerni program (engl. *Malware*), neželjeni softver (engl. *Unwanted Software*) i socijalni inženjering (engl. *Social Engineering*). Zlonamjerni program [57] podrazumijeva softver ili mobilnu aplikaciju dizajniranu da našteti uređaju poput instalacije softvera bez odobrenja korisnika, instalacije štetnog softvera i sl. Neželjeni softver je izvršna datoteka ili aplikacija koja se ponaša neočekivano ili negativno utječe na korisnikovo pregledavanje. Primjeri neželjenog softvera je softver koji mijenja postavke preglednika, aplikacije koje otkrivaju privatne i osobne podatke i slično. Socijalni inženjering se temelji na prevari korisnika, phishingu, obmanjujućem sadržaju te nedovoljno naznačenim uslugama trećih strana [58].

Google Safe Browsing se može koristiti na nekoliko načina. Najjednostavniji je putem stranice *transparencyreport.google.com/safe-browsing/search* [59] u kojem se ručno upiše

URL stranice te se dobije odgovor o zlonamjernosti na samoj stranici kao i kada je stranica zadnje pregledana.

Drugi način korištenja je putem API poziva kroz usluge *Safe Browsing Lookup API (v4)* [60]. Kako bi se API mogao koristiti potrebno je stvoriti ključ na *Google Cloud*-u unutar već stvorenog projekta. Nakon tog koraka postavlja se POST zahtjev na stranicu *https://safebrowsing.googleapis.com/v4/threatMatches:find* skupa s JSON objektom koji u sebi sadržava informacije o klijentu koji je poslao zahtjev i informacije o samoj prijetnji. Klijentski dio zahtjeva za ovaj rad je nepotreban i nebitan, on se koristi samo kada posebno registrirani klijenti imaju automatizirane upite za određene URL-ove u željenim vremenskim sekvencama. Informacije o prijetnji sadrže četiri dijela: tip prijetnje, platforma, ulazni tip prijetnje te sam sama prijetnja.

Prijetnja može biti u obliku URL-a, sažetka ili programa u binarnom ili heksadekadskom obliku. Platforme koje *Safe Browsing* podržava su *Windows, Linux, IOS, Android, OSXX, Chrome*.

Safe Browsing može se koristiti i putem Python biblioteke *pysafebrowsing* [61]. Primjer korištenja *pysafebeowsing* biblioteke dan je u Ispis 3.6.

Ispis 3.6: Korištenje biblioteke pysafebrowsing za dohvaćanje rezultata servisa za sigurno pretraživanje za određeni URL

```
from pysafebrowsing import SafeBrowsing
s = SafeBrowsing(KEY)
r =
s.lookup_urls(['http://malware.testing.google.test/testing/malware/'])
print(r)
> {'http://malware.testing.google.test/testing/malware/':
{'platforms': ['ANY_PLATFORM'], 'threats': ['MALWARE',
'SOCIAL_ENGINEERING'], 'malicious': True, 'cache': '300s'}}
```

Korištenje *Google Safe Browsing*-a je ograničena na 10 000 API zahtjeva po danu, 1 800 zahtjeva po minuti te 500 URL-ova u jednom API zahtjevu. Brzim izračunom, može se zaključiti da se dnevno može pregledati 5 milijuna URL-ova, no to nije tako. Svaki API zahtjev radi sažetak po tipu prijetnje, npr. ako se u 500 URL-ova nalazi 30 zlonamjernih URL-a koji pripadaju kategoriji *social engineering*, prikazat će se samo prvi pronađeni zlonamjerni URL! Stoga, kako bi se u velikom skupu podataka mogao pronaći svaki

zlonamjerni URL, mora se raditi posebni API poziv po svakom URL-u. No, u usporedbi s ostalim uslugama sličnim *Safe Browsing*-u, ovo ograničenje nije tako oštro kao što je npr. *Virus Total* koji dopušta samo 4 zahtjeva po minuti te 500 zahtjeva po danu.

Veliki problem s *Google Safe Browsing*-om jest neinterpretabilnost. Kada je stranica proglašena zlonamjernom nigdje nije naveden razlog zašto je to tako. Čak ni pregledom samih stranica koje je nije očit razlog zašto je *Safe Browsing* proglasio stranicu zlonamjernom. Koliko stranica je *Google Safe Browsing* označio zlonamjernim prikazano je u idućem poglavlju.

4. HR domena

U idućih nekoliko poglavlja bit će opisano parsiranje, statistika te rezultati strojnog učenja za klasifikaciju zlonamjernih stranica u hr domeni. Forma će biti ista kao i kod *MalCrawler* skupa podataka te će se proučavati i dohvaćati isti podatci te formirati iste značajke kao i kod *MalCrawler* skupa podataka.

Popis registriranih domena pod vršnom hr domenom omogućio je Nacionalni CERT preko projekta *eŠkole*. Preko tih domena pronađene su web stranice koje su se zatim dohvaćale i spremale u *MongoDB* bazu. Iz stranica su izdvojene sve skripte (uključujući i JS) te se i njihov sadržaj spremao u bazu podataka. Ovaj proces se odvija dvije godine otkada se piše ovaj rad stoga su podatci od web stranica dostupni od zadnjih dvije godine.

4.1. Podatci u *websecradar* bazi

Od 221 327 URL-ova dohvaćenih iz websecradar baze podataka, 94 999 URL-ova ima korisne podatke. Razlog tomu je što se u obzir uzimalo HTTP i HTTPS stranica, a najčešće samo jedan od ta dva protokola imalo je korisne podatke za web stranicu. Za ostale razloge se može samo pretpostavljati, ali još jedan od mogućih razloga je ta da je većina stranica je postala neaktivna, ali su i dalje u popisu registriranih domena pod vršnom hr domenom.

Od tih 94 999 URL-ova, može se izdvojiti 48 421 različita domena. Razlog tomu je taj da postoji veliki broj blog stranica poput *blogpost.com*, koji imaju različitu domenu za svakog korisnika.

4.2. Statistika za skup podataka hr domene

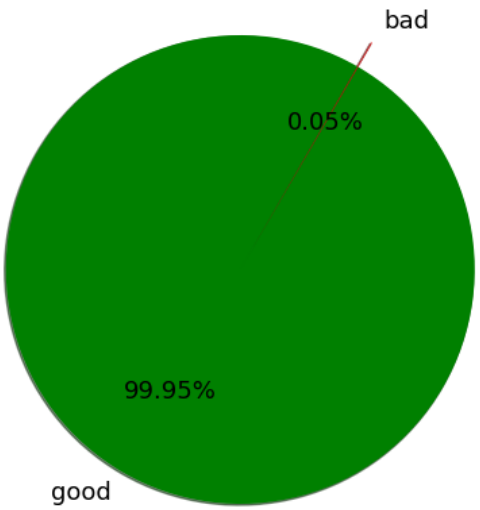
Od 94 999 skeniranih URL-ova Google-ovim alatom sigurnog pretraživanja, samo 43 URL-a bilo je označeno zlonamjernim. Kategorija svih označenih URL-ova pripada socijalnom inženjeringu. Dio popisa zlonamjernih URL-ova prikazan je u Ispis 4.1.

Ispis 4.1: Popis dijela zlonamjernih URL-a u hr domeni

```
https://www.uzrh-bb.hr
https://www.rubicon.hr/
https://medservis.hr
```

```
https://spread.hr
https://rosandaprojekt.hr
https://www.rubicon.hr
https://mctech.hr
https://adela.hr
https://solum.hr
https://edok.hr
https://kondenzatori.hr
https://www.zupa-duha-svetoga-sb.hr
https://t3m.hr
https://rewe.hr
```

Gledajući sve stranice, udio zlonamjernih stranica je 0.05% što je i prikazano kružnim grafom na Slika 4.1.

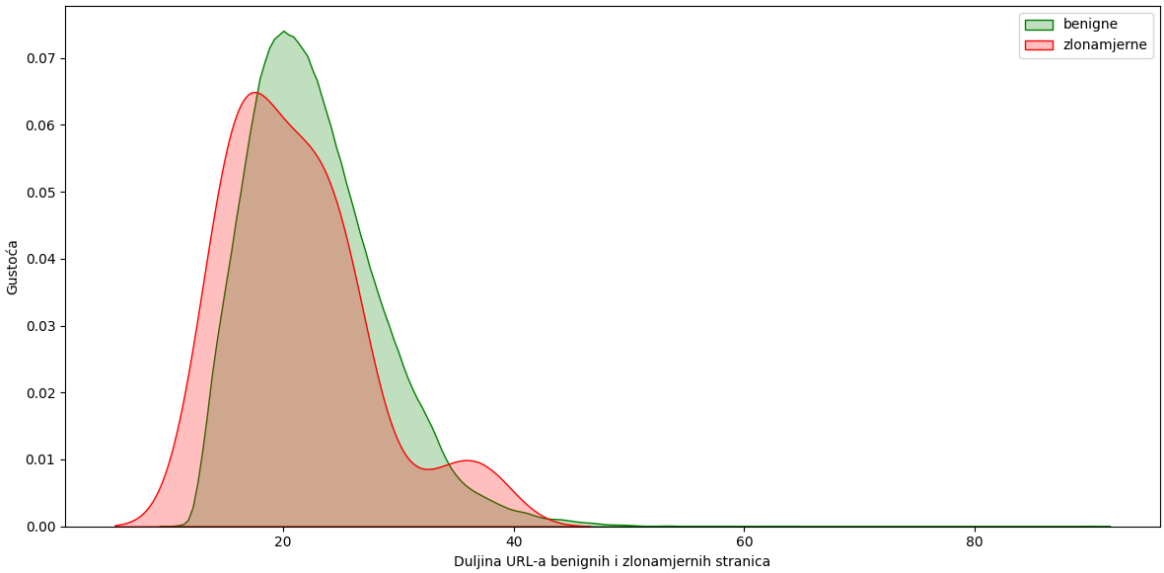


Slika 4.1: Kružni graf udjela benignih i zlonamjernih stranica

To je za 2.22% manji udio nego u *MalCrawler* skupu podataka. Ovako mali udio podataka premali je za testiranje algoritmima strojnog učenja, stoga će se trebati kombinirati s podatcima iz *MalCrawler*-a. Razlog ovako maloj količini zlonamjernih stranica u hr domeni može biti da postoji jako malo kriminalne aktivnosti u internetskom prostoru Hrvatske, no može značiti i da servis sigurnog pretraživanja rijetko obilazi stranice koje se nalaze u hr domeni.

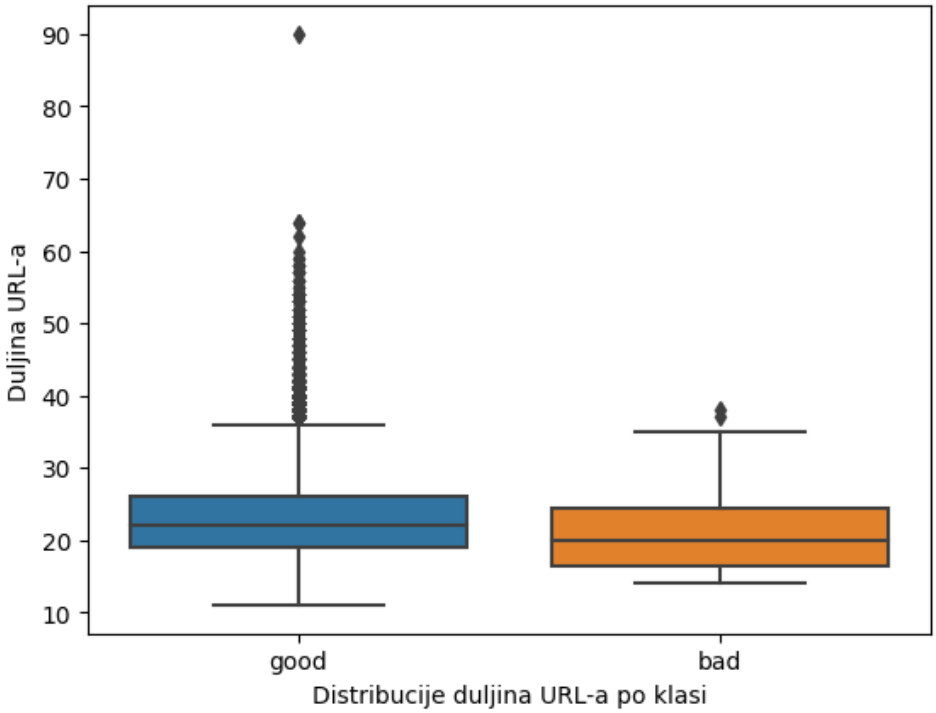
4.2.1. Duljina URL-a

Na Slika 4.2 prikazan suje grafovi gustoće duljine URL-a za benigne (zeleno) i zlonamjerne (crveno) stranice.



Slika 4.2: Gustoće vjerojatnosti duljine URL-a benignih i zlonamjernih stranica

Prema Slika 4.2 vidljivo je da su, kao i kod *MalCrawler* skupa podataka, duljina URL-a benignih i zlonamjernih stranica vrlo slične te da se kreću od raspona od 10 do 45 znakova. Isto se može zaključiti grafa sa Slika 4.3.



Slika 4.3: Boxplot graf distribucija duljina URL-a po klasi

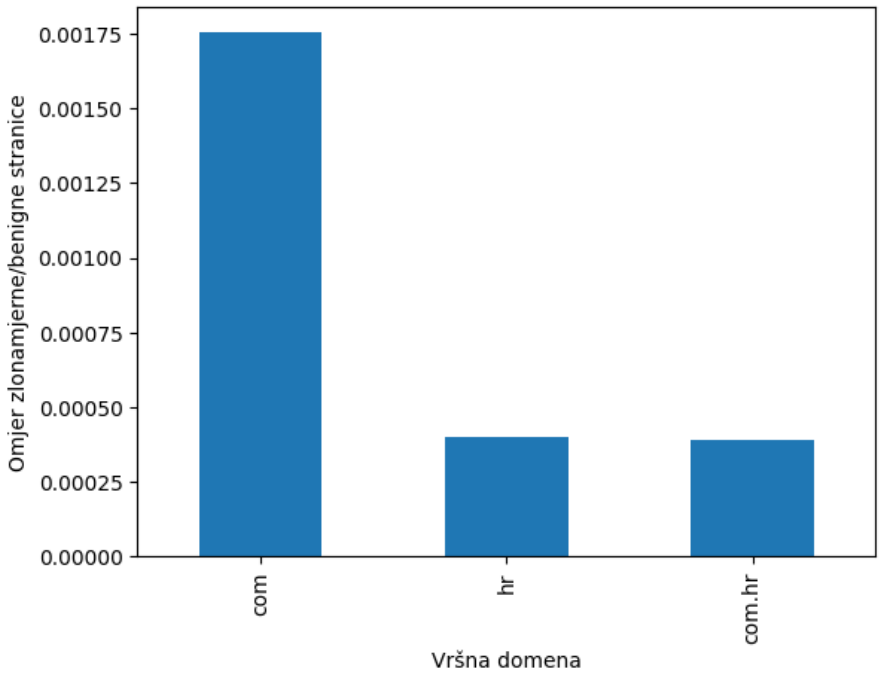
Iz slike vidimo vrlo sličnu distribucije s time da postoje neznatan broj URL-a benignih stranica koji iskaču iznad gornjeg ekstrema.

4.3. Vršna domena

Najpopularnijih 10 vršnih domena u hr domeni su: .hr (83 014), .com.hr (5 143), .com (4 561), .net (774), .info (262), .blogspot.com (239), .from.hr (188), .org (173), .eu (157), .biz (104).

Zlonamjerne stranice se nalaze na samo tri vršne domene: .hr (33), .com (8), .com.hr (2). Omjer zlonamjernih i benignih stranica za vršne domene vidi se na Slika 4.4.

Na slici vidi se omjer zlonamjernih i benignih stranica za vršne domene.



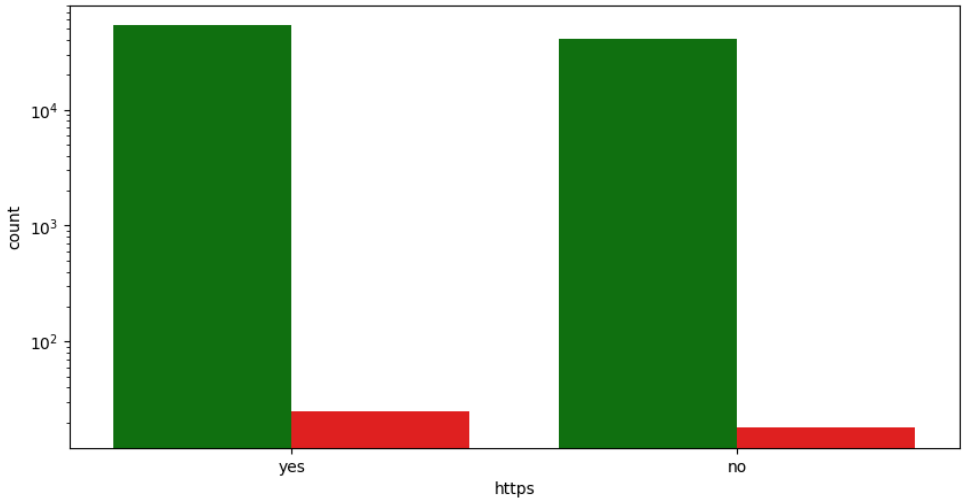
Slika 4.4 Omjeri zlonamjernih i benignih stranica po vršnoj domeni

Najveći omjer ima .com domena sa 0.001754, zatim .hr (0.000398) te .com.hr (0.000389). Omjeri su vrlo mali zbog malog broj zlonamjernih stranica u skupu podataka.

4.4. HTTP i HTTPS protokol

U kreiranom skupu podataka 53 668 benignih stranica koristi HTTPS protokol te 41 288 stranica koristi HTTP protokol. S druge strane 25 zlonamjernih stranica koristi HTTPS protokol te 18 stranica koristi HTTP protokol. Iako je malo podataka zlonamjernih stranica, omjer između benignih i zlonamjernih stranica koje koriste HTTPS protokol je vrlo različit od onoga u *MalCrawler* skupu podataka. Još jedna razlika je ta da mnogo više zlonamjernih stranica u *MalCrawler* skupu koristi HTTP protokol nego što koristi HTTPS, dok u skupu podataka hr domene više zlonamjernih stranica koristi HTTPS protokol nego HTTP. Također, razlika broja benignih stranica koje koriste HTTPS i HTTP je oko 10 000 u korist HTTPS dok u *MalCrawler* skupu podataka benignih stranica koji koriste HTTPS ima 700 000 više benignih stranica koji koriste HTTP.

Distribucije navedenih informacija za značajku *https* su prikazane na Slika 4.5.

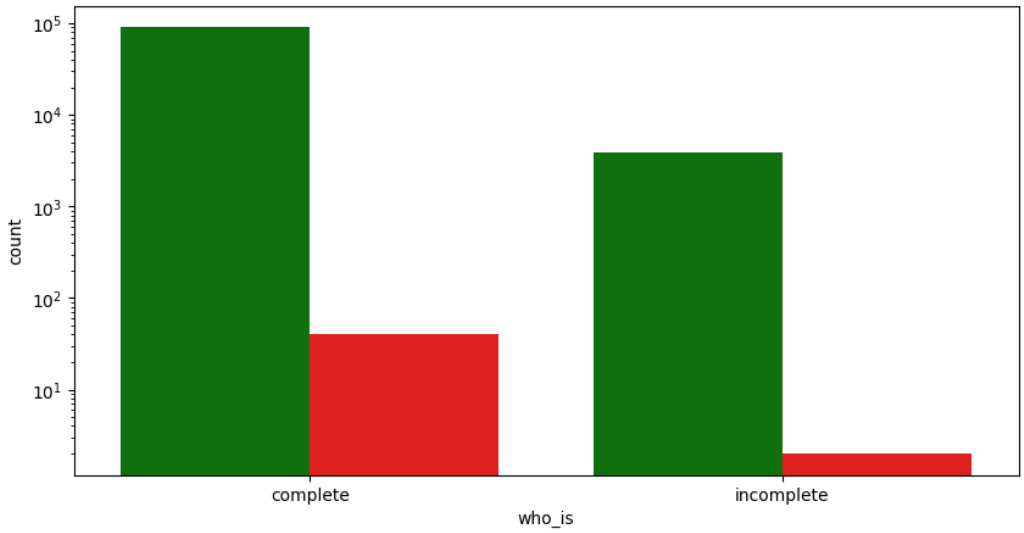


Slika 4.5: Višestruki stupčasti dijagram omjera benignih i zlonamjernih stranica koje koriste i ne koriste HTTPS protokol

4.5. DNS WHOIS zapisi

U skupu podataka hr domene 91 055 benignih stranica ima potpune WHOIS podatke te 3 901 nema potpune WHOIS podatke. 41 zlonamjerna stranica ima potpune WHOIS zapis te 2 nema potpune podatke. Za razliku od *MalCrawler* skupa podataka, više zlonamjernih stranica ima potpune WHOIS podatke nego što ih nema. Velika razlika između s *MalCrawler* podatcima je odnos koliko benignih stranica ima potpune, odnosno nepotpune WHOIS podatke. Dok u *MalCrawler* podatcima taj omjer 0.25, u podatcima hr domene taj je omjer 0.04. Većina stranica koje su pod nadležnosti CARNET-a (.hr, .com.hr, .from.hr) ima potpune WHOIS podatke. Prednost nadzora nad relativno manjim brojem stranica je lakša regulacija ispravnosti i potpunosti podataka za svaku domenu.

Distribucije stranica i potpunosti informacija za benigne i zlonamjerne stranice prikazane su na Slika 4.6.

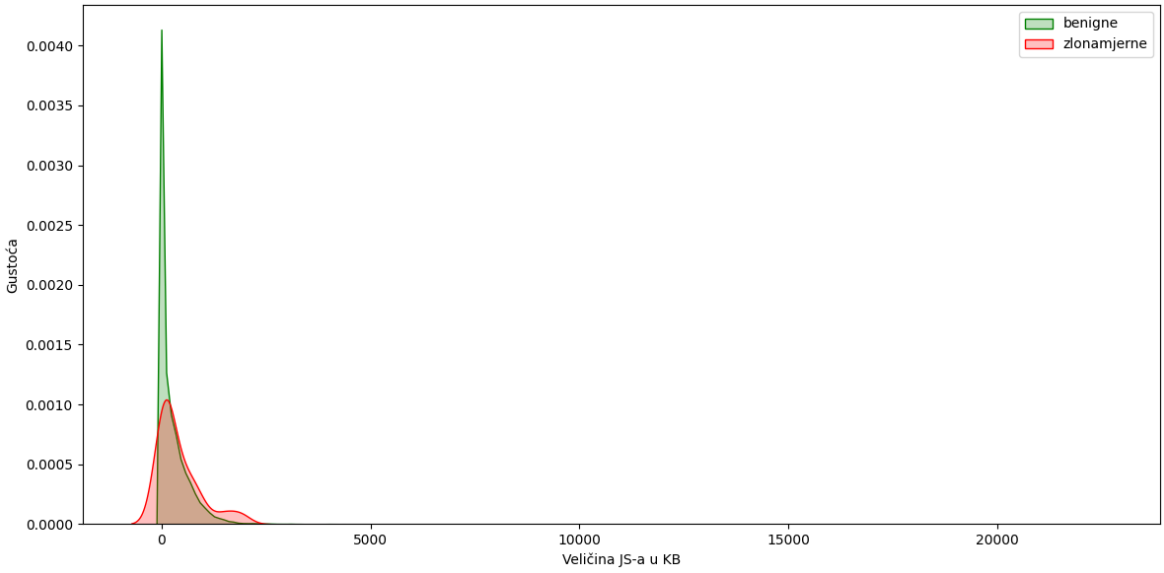


Slika 4.6: Višestruki stupčasti dijagram omjera benignih i zlonamjernih stranica koje imaju potpune i nepotpune WHOIS zapise

4.6. Veličina JavaScript koda

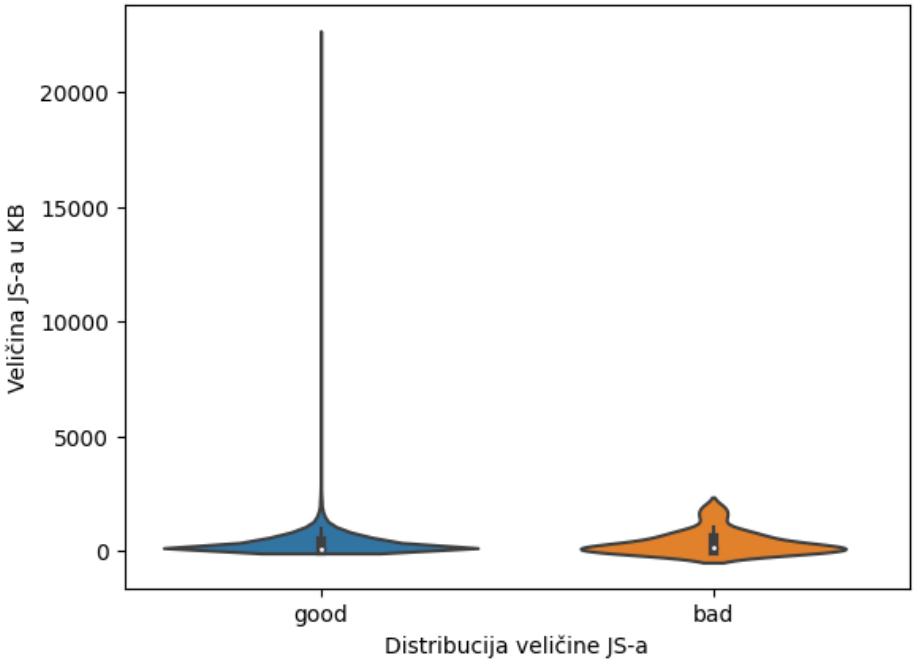
Prosječna veličina JavaScript koda benignih stranica jest 402.49 KB, a zlonamjernih 272.58 KB. To je također razlika od *MalCrawler* skupa podataka gdje je veličina JavaScripta u zlonamjernim stranicama bila veća nego kod benignih.

Slika 4.7 prikazuje gustoće vjerojatnosti veličine JavaScripta za benigne (zeleno) i zlonamjerne (crveno) stranice.



Slika 4.7: Graf gustoće vjerojatnosti veličine JS-a u KB za benigne i zlonamjerne stranice

Na slici se vidi da su veličine JavaScripta podjednako distribuirane i većinski su ispod 1 000 KB veličine. Postoji 4 921 stranica čija veličina JavaScripta prelazi 1 000 KB te šest stranica s preko 5 000 KB veličine. Distribucije veličine JavaScripta vide se na Slika 4.8.



Slika 4.8: Graf zvonca za distribucije veličina JS-a za benigne i zlonamjerne stranice

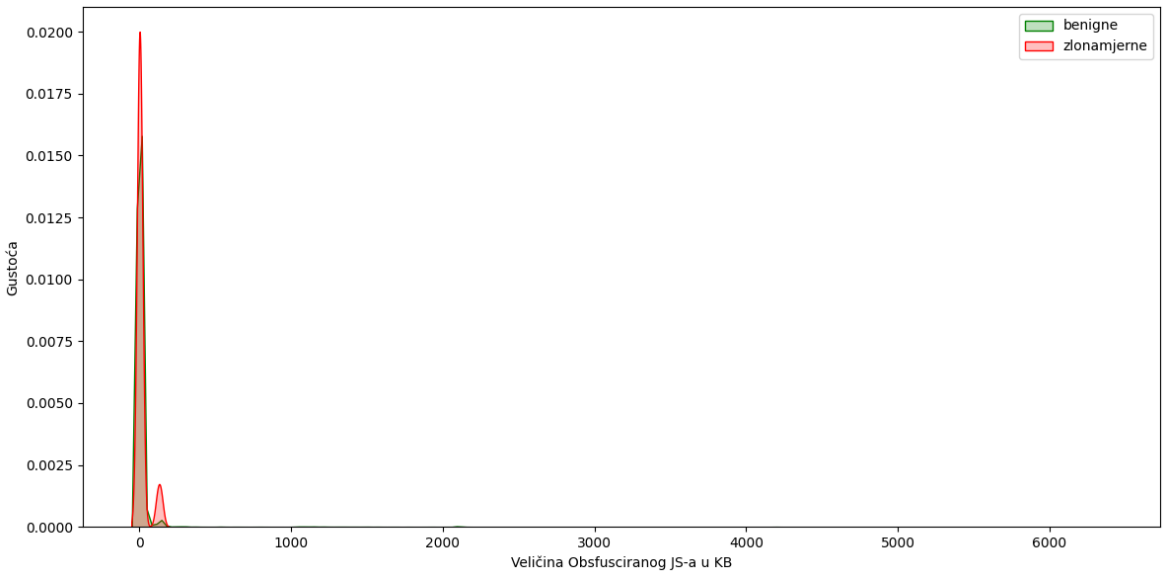
4.7. Veličina obfusciranog JavaScript koda

Adobeovu algoritmu je bilo potrebno za klasifikaciju većih JavaScript kodova do i po 20 minuta. Stoga je zbog svoje vremenske kompleksnosti zanemaren kod detekcije obfuskacije u ovom radu.

Razvijeni algoritam za klasifikaciju obfusciranih JavaScript kodova dao je na početku loše rezultate. Procijenjeno je oko 30 000 lažno pozitivno klasificiranih JavaScript datoteka, odnosno 30 000 JavaScripta krivo klasificiranih kao obfusciranim. Ručnim pregledom otkriveno je da je uzrok tog problema minificirani JavaScript, stoga je u proces učenja algoritma dodana i treća klasa koja predstavlja minificirani kod. Nakon toga, broj JavaScript koda koji su lažno klasificirani kao obfuscirani se smanjio za faktor od 10.

Prosječna veličina obfusciranog JavaScript koda benignih stranica jest 12.69 KB, a zlonamjernih 17.46 KB. Kao i u *MalCrawler* skupu podataka, prosječna veličina obfusciranog JavaScripta zlonamjernih stranica je veća od benignih, no ne za toliki faktor

veće kao kod *MalCrawler* podataka. Graf gustoće vjerojatnosti veličine obfusciranog JavaScripta benignih i zlonamjernih stranica prikazan je na Slika 4.9.



Slika 4.9: Graf gustoće vjerojatnosti veličine obfusciranog JS-a u KB za benigne i zlonamjerne stranice

Usporedba tri numeričke vrijednosti: duljine URL-a, veličina JavaScripta u KB, veličina obfusciranog JavaScripta u KB; prikazano je u Tablica 4.1.

Tablica 4.1: Statističke vrijednosti numeričkih značajki *hr* skupa podataka

	Statistika benignih stranica			**Statistika zlonamjernih stranica**		
	url_len	**js_len**	**js_obf_len**	**url_len**	**js_len**	**js_obf_len**
count	94956.00	94956.00	94956.00	43.00	43.00	43.00
mean	22.81	272.58	17.46	21.21	402.54	12.70
std	5.80	385.71	165.48	6.08	508.96	38.13
min	11.00	0.0	0.0	14.0	0.0	0.0
25%	19.00	2.29	0.0	16.50	41.35	0.00
50%	22.00	130.85	0.0	20.00	180.63	0.00
75%	26.00	411.89	0.0	24.50	601.93	0.00
max	90.00	22599.21	6355.64	38.00	1881.36	140.44

Iz Tablica 4.1 primjetno je da su brojevi veći kod benignih stranica i manji kod zlonamjernih stranica nego kod *MalCrawler* skupa podataka te da nema jasne korelacije između veličine JavaScripta i veličine obfusciranog JavaScripta.

Korelacijska matrica prikazana je Tablica 4.2.

Tablica 4.2: Korelacijska matrica numeričkih vrijednosti *hr* skupa podataka

	url_len	js_len	js_obf_len
url_len	1.000000	0.005079	-0.001251
js_len	0.005079	1.000000	0.070254
js_obf_len	-0.001251	0.070254	1.000000

Prema korelacijskog matrici, ovog puta nema nikakve korelacije između ijedne numeričke varijable.

4.8. Rezultati strojnog učenja za skup podataka hr domene

Korišteni algoritmi strojnog učenja su isti kao i za *MalCrawler* u poglavlju:

- Gaussov naivni Bayes (engl. *Gaussian Naive Bayes*)
- Multinomijalni naivni Bayes (engl. *Multinomial Naive Bayes*)
- Komplementarni naivni Bayes (engl. *Complement Naive Bayes*)
- Stablo odluke C4.5 (engl. *Decision Tree C4.5*)
- Nasumična šuma (engl. *Random Forest*)
- Stroj potpornih vektora (engl. *Support Vector Machine, SVM*)

Algoritmi su računati i testirani na skupu podataka hr domene od 94 499 stranica s prijelomom od 0.33 u korist treniranja. U skupu za testiranje je 31 333 benignih i 17 zlonamjernih primjeraka. U Tablica 4.3 prikazani su uspješnosti svih algoritama. Lažno benigni su označeni s FP, dok su lažno zlonamjerni s FN.

Tablica 4.3: Mjere uspješnosti klasifikacije algoritama strojnog učenja

	FN	FP	točnost	preciznost	odziv	F1 mjera	ROC AUC
Stablo odluke	5	14	0.99939	0.375	0.17647	0.24	0.58815
Nasumična šuma	2	17	0.99939	0.0	0.0	0.0	0.5
Gaussov Bayes	28490	0	0.09122	0.00059	1.0	0.00119	0.54536
Multinomijalni Bayes	8	77	0.99945	0.0	0.0	0.0	0.5
Komplementarni Bayes	0	77	0.98287	0.0	0.0	0.0	0.49170
SVC	0	181	0.99945	0.0	0.0	0.0	0.5

Rezultati svih algoritama strojnog učenja su jako loši što je i očekivano s obzirom na vrlo mali broj zlonamjernih stranica nad kojima algoritam može učiti. No, da je i bilo 100 puta više primjeraka zlonamjernih stranica sa sličnom distribucijom vrijednosti značajki kao i već postojeće zlonamjerne stranice, algoritmi bi i dalje imali lošiju klasifikaciju nego kod *MalCrawler* skupa podataka zbog distribucije korištenja HTTPS protokola, potpunosti WHOIS podataka te korelacije između količine JavaScript i obfusciranog JavaScript koda na benignim i zlonamjernih stranicama.

Drugi pristup testiranju algoritama strojnog učenja jest takav da se algoritam trenira nad *MalCrawler* skupom podataka te testira nad podatcima iz hr domene. Rezultati takvog pristupa pokazali su se također loši.

Prema dobivenim rezultatima, jasan je zaključak da uspješnost klasifikacije zlonamjernih stranica pomoću osnovnih značajki se ne prenosi na skup podataka hr domene. Potrebno dodati još naprednijih značajki koji se mogu dobiti statičkom analizom web stranice.

Zaključak

S obzirom na veličinu weba i udio benignih stranica naprema zlonamjernim, usredotočeni web pretraživači poput *MalCrawlera* pokazali su se uspješnima u većoj stopi detekcije zlonamjernih stranica na Internetu. Analizom značajki skupa podataka *MalCrawlera* pokazalo je da podatci poput korištenja HTTPS protokola, potpunost WHOIS zapisa domene te veličine obfusciranog JavaScripta, mogu biti odlični indikatori o potencijalnoj zlonamjernosti web stranice. Dolaženje do informacija poput WHOIS zapisa i detekcije obfusciranog JavaScript koda pokazali su se težima nego prvotno očekivano. Algoritmi poput stabla odluke, nasumične šume i Gaussovog naivnog Bayesa davali su izvrsne rezultate kod klasifikaciji zlonamjernih stranica. Stablo odluke dalo je najbolje rezultate zbog detaljne pretrage svih kombinacija hiperparametara. Stablo odluke postiglo je točnost od 0.99884, odziva od 0.96037 te f1 mjere od 0.97456.

Googleov servis za sigurno pretraživanje detektirao je samo 43 zlonamjerne stranice od 94 999 koji pripadaju Hrvatskom domenskom prostoru. Razlozi se jedino mogu nagađati, no neki od njih su neprivlačnost Hrvatske kao mete kibernetičkog kriminala te rijetko obilaženje stranice u hr domeni od strane servisa za sigurno pretraživanje.

Stvoreni skup podataka iz stranica u hr domeni te *MalCrawler* skupa podataka se jako razlikuju po omjerima i vrijednosnim distribucijama benignih i zlonamjernih stranica u svim značajkama. Algoritmi strojnog učenja dali su jako loše rezultate kod klasifikacije zlonamjernih stranica u hr domeni zbog prevelike neuravnoteženi klasa te zbog distribucije vrijednosti benignih i zlonamjernih stranica stvorenog skupa podataka.

Kako bi se poboljšali rezultati detekcije zlonamjernih stranica u hr domeni, potrebno je dodati skup naprednijih značajki. Eksplicitno pisana pravila te naknadna ručna provjera bi mogla davati bolje rezultate zbog vrlo male količine zlonamjernih stranica označenih od strane servisa sigurnog pretraživanja.

Literatura

[1] "How Many Websites Are There in the World? (2022) - Siteefy." https://siteefy.com/how-many-websites-are-there/ (accessed Sep. 19, 2022).

[2] "What is a Malicious Website? - How to Spot a Malicious Website - Tessian." https://www.tessian.com/blog/how-to-identify-a-malicious-website/ (accessed Sep. 19, 2022).

[3] Symantec, "Symantec Internet Security Threat Report: Attack Trends for Q3 and Q4 2002," *Symantec internet Secur. Threat Rep.*, vol. 3, pp. 1–48, 2003, [Online]. Available: https://www.symantec.com/content/dam/symantec/docs/security-center/archives/istr-03-jan-en.pdf.

[4] N. G. A.K. Singh, "MalCrawler: A Crawler for Seeking and Crawling Malicious Websites," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8956, pp. 5–6, 2015, doi: 10.1007/978-3-319-50472-8.

[5] L. Invernizzi, S. Benvenuti, M. Cova, C. Kruegel, and G. Vigna, "EVILSEED : A Guided Approach to Finding Malicious Web Pages."

[6] A. Ikinci, T. Holz, and F. Freiling, "Monkey-spider: Detecting malicious websites with low-interaction honeyclients," *Lect. Notes Informatics (LNI), Proc. - Ser. Gesellschaft fur Inform.*, vol. P-128, no. May, pp. 407–421, 2008.

[7] "Cloaking | Google Search Central | Documentation | Google Developers." https://developers.google.com/search/docs/advanced/guidelines/cloaking (accessed Sep. 19, 2022).

[8] "Dataset of Malicious and Benign Webpages - Mendeley Data." https://data.mendeley.com/datasets/gdx3pkwp47/2 (accessed Sep. 12, 2022).

[9] A. K. Singh and N. Goyal, "A Comparison of Machine Learning Attributes for Detecting Malicious Websites," vol. 2061, pp. 352–358.

[10] "5 URL Warning Signs to Watch For | INFORMATION TECHNOLOGY." https://www.du.edu/it/services/security/5-url-warning-signs (accessed Sep. 12, 2022).

[11] "What Is SEO Poisoning (Search Engine Poisoning) - Security Boulevard." https://securityboulevard.com/2019/10/what-is-seo-poisoning-search-engine-poisoning/ (accessed Sep. 12, 2022).

[12] L. Lu, R. Perdisci, and W. Lee, "SURF: Detecting and measuring search poisoning,"

Proc. ACM Conf. Comput. Commun. Secur., pp. 467–476, 2011, doi: 10.1145/2046707.2046762.

[13] "Understanding Boxplots: How to Read and Interpret a Boxplot | Built In." https://builtin.com/data-science/boxplot (accessed Sep. 19, 2022).

[14] "Web Threats: Malicious Host URLs, Landing URLs and Trends." https://unit42.paloaltonetworks.com/web-threats-malicious-host-urls/ (accessed Sep. 12, 2022).

[15] "Top-level Domain: What it is and How to choose one." https://rockcontent.com/blog/top-level-domain/ (accessed Sep. 12, 2022).

[16] "• Most popular TLDs worldwide 2022 | Statista." https://www.statista.com/statistics/265677/number-of-internet-top-level-domains-worldwide/ (accessed Sep. 13, 2022).

[17]. "ORG vs .COM vs .NET - What is .ORG and Other Extensions." https://www.wix.com/blog/2020/06/org-vs-com-vs-net-domain-extensions/ (accessed Sep. 13, 2022).

[18] "A Peek into Top-Level Domains and Cybercrime." https://unit42.paloaltonetworks.com/top-level-domains-cybercrime/ (accessed Sep. 14, 2022).

[19] "Cumulative Distribution Function." https://www.probabilitycourse.com/chapter3/3_2_1_cdf.php (accessed Sep. 19, 2022).

[20] "Median absolute deviation - Wikipedia." https://en.wikipedia.org/wiki/Median_absolute_deviation (accessed Sep. 19, 2022).

[21] "Why is HTTP not secure? | HTTP vs. HTTPS | Cloudflare." https://www.cloudflare.com/learning/ssl/why-is-http-not-secure/ (accessed Sep. 16, 2022).

[22] "Should You Switch Your Site to HTTPS? Pros and Cons." https://www.quicksprout.com/switching-from-http-to-https/ (accessed Sep. 16, 2022).

[23] "WHOIS Data – Frequently Asked Questions." https://docs.umbrella.com/investigate/docs/investigate-whois-data-frequently-asked-questions (accessed Sep. 16, 2022).

[24] "About WHOIS | ICANN WHOIS." https://whois.icann.org/en/about-whois (accessed Sep. 16, 2022).

[25] “What is Whois Information and Why is it Valuable? - DomainTools | Start Here. Know Now.” https://www.domaintools.com/support/what-is-whois-information-and-why-is-it-valuable/ (accessed Sep. 16, 2022).

[26] “Key Terms & Definitions - DomainTools | Start Here. Know Now.” https://www.domaintools.com/support/key-terms-definitions/#registrar (accessed Sep. 16, 2022).

[27] M. Kuyama and T. S. of D. I. and W. Communication, “Method for Detecting a Malicious Domain by using only Well-known Information,” *Int. J. Cyber-Security Digit. Forensics*, vol. 5, no. 4, pp. 166–174, 2016, Accessed: Sep. 16, 2022. [Online]. Available: https://www.academia.edu/28293914/Method_for_Detecting_a_Malicious_Domain_by_using_WHOIS_and_DNS_features.

[28] Y. Cheng, T. Chai, Z. Zhang, K. Lu, and Y. Du, “Detecting Malicious Domain Names with Abnormal WHOIS Records Using Feature-Based Rules,” *Comput. J.*, May 2021, doi: 10.1093/COMJNL/BXAB062.

[29] H. Kikuchi, D. Yu, 686,288 A Chander - US Patent 9, and undefined 2017, “Method and apparatus for constructing security policies for web content instrumentation against browser-based attacks,” *Google Patents*, 2017, doi: 10.1007/s10207-004-0046-8.

[30] “A Complete Guide to Violin Plots | Tutorial by Chartio.” https://chartio.com/learn/charts/violin-plot-complete-guide/ (accessed Sep. 17, 2022).

[31] “Obfuscation - Definition.” https://www.trendmicro.com/vinfo/us/security/definition/obfuscation (accessed Sep. 17, 2022).

[32] “Attackers are Quicker to Change Code With Obfuscation.” https://www.darkreading.com/threat-intelligence/attackers-are-quicker-to-change-code-with-obfuscation (accessed Sep. 17, 2022).

[33] “Akamai Blog | Over 25% of Malicious JavaScript Is Being Obfuscated.” https://www.akamai.com/blog/security/over-25-percent-of-malicious-javascript-is-being-obfuscated (accessed Sep. 17, 2022).

[34] “API Reference — scikit-learn 1.1.2 documentation.” https://scikit-learn.org/stable/modules/classes.html (accessed Sep. 19, 2022).

[35] R. G. Mantovani, T. Horváth, R. Cerri, S. B. Junior, J. Vanschoren, and A. C. P. de

L. F. de Carvalho, “An empirical study on hyperparameter tuning of decision trees,” Dec. 2018, Accessed: Sep. 18, 2022. [Online]. Available: https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680.

[36] “sklearn.model_selection.HalvingRandomSearchCV — scikit-learn 1.1.2 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.HalvingRandomSearchCV.html (accessed Sep. 19, 2022).

[37] “geoip2 · PyPI.” https://pypi.org/project/geoip2/ (accessed Sep. 18, 2022).

[38] “maxminddb · PyPI.” https://pypi.org/project/maxminddb/ (accessed Sep. 18, 2022).

[39] “P3TERX/GeoLite.mmdb: MaxMind’s GeoIP2 GeoLite2 Country, City, and ASN databases.” https://github.com/P3TERX/GeoLite.mmdb (accessed Sep. 18, 2022).

[40] “IP Geolocation API and IP Location Lookup Tools.” https://ipwhois.io/ (accessed Sep. 18, 2022).

[41] “tld · PyPI.” https://pypi.org/project/tld/ (accessed Sep. 18, 2022).

[42] “python-whois · PyPI.” https://pypi.org/project/python-whois/#description (accessed Sep. 19, 2022).

[43] “WHOIS API | 565M+ active domains & 7,298 TLDs tracked | WhoisXML API.” https://whois.whoisxmlapi.com/ (accessed Sep. 19, 2022).

[44]. “hr domene.” https://www.domene.hr/portal/home (accessed Sep. 18, 2022).

[45] “selenium · PyPI.” https://pypi.org/project/selenium/ (accessed Sep. 18, 2022).

[46] “Entropy (information theory) - Wikipedia.” https://en.wikipedia.org/wiki/Entropy_(information_theory) (accessed Sep. 18, 2022).

[47] “javascript obfuscation detection | Kaggle.” https://www.kaggle.com/code/fanbyprinciple/javascript-obfuscation-detection/data (accessed Sep. 18, 2022).

[48] P. Likarish, E. Jung, and I. Jo, “Obfuscated malicious javascript detection using classification techniques,” *2009 4th Int. Conf. Malicious Unwanted Software, MALWARE 2009*, no. May, pp. 47–54, 2009, doi: 10.1109/MALWARE.2009.5403020.

[49] S. Aebersold, K. Kryszczuk, S. Paganoni, B. Tellenbach, and T. Trowbridge, “Detecting Obfuscated JavaScripts using Machine Learning,” *ICIMP 2016 Elev. Int. Conf. Internet Monit. Prot.*, no. c, pp. 11–16, 2016, [Online]. Available: https://pd.zhaw.ch/publikation/upload/211098.pdf.

[50] B. Tellenbach, S. Paganoni, and M. Rennhard, “Detecting Obfuscated JavaScripts from Known and Unknown Obfuscators using Machine Learning,” *Int. J. Adv. Secur.*, vol. 9, no. 3, pp. 196–206, 2016.

[51] S. M. de S. Lima, “Automatic Identification of Obfuscated JavaScript using Machine Learning,” 2021.

[52] “sklearn.feature_extraction.text.HashingVectorizer — scikit-learn 1.1.2 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.HashingVectorizer.html (accessed Sep. 18, 2022).

[53] “HashingVectorizer vs. CountVectorizer - Kavita Ganesan, PhD.” https://kavita-ganesan.com/hashingvectorizer-vs-countvectorizer/ (accessed Sep. 18, 2022).

[54] “sklearn.feature_extraction.text.TfidfTransformer — scikit-learn 1.1.2 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html (accessed Sep. 18, 2022).

[55] “adobe/obfuscation-detection.” https://github.com/adobe/obfuscation-detection (accessed Sep. 18, 2022).

[56] “Safe Browsing – Google Safe Browsing.” https://safebrowsing.google.com/ (accessed Sep. 18, 2022).

[57] “Malware and Unwanted Software Overview | Google Search Central | Documentation | Google Developers.” https://developers.google.com/search/docs/monitor-debug/security/malware?visit_id=637991133108428556-1875394520&rd=1 (accessed Sep. 18, 2022).

[58] “Social Engineering (Phishing and Deceptive Sites) | Google Search Central | Documentation | Google Developers.” https://developers.google.com/search/docs/monitor-debug/security/social-engineering?visit_id=637991135142812912-769684795&rd=1 (accessed Sep. 18, 2022).

[59] “Google sigurno pregledavanje – Transparentnost na Googleu.” https://transparencyreport.google.com/safe-browsing/search (accessed Sep. 18, 2022).

[60] “Overview | Safe Browsing APIs (v4) | Google Developers.” https://developers.google.com/safe-browsing/v4/ (accessed Sep. 18, 2022).

[61] “pysafebrowsing · PyPI.” https://pypi.org/project/pysafebrowsing/ (accessed Sep. 18, 2022).

Sažetak

Sa sve većim brojem aktivnih web stranica rastu i kriminalne aktivnosti na Internetu. Napadači putem weba pokušavaju prevariti korisnika u davanje osjetljivih podataka ili skidanja i pokretanja zlonamjernih programa. Radi sprječavanja takvih aktivnosti ovaj rad se bavi detekcijom zlonamjernih web stranica strojnim učenjem. Proučavan je skup podataka *MalCrawler* te su napisane skripte za dobivanje njenih značajki u svrhu stvaranja novog skupa podataka nad stranicama u hr domeni. Rezultati strojnog učenja nad *MalCrawler* postigli su jako dobre rezultate gdje je najbolje rezultate postigao algoritam stabla odluke C4.5 s točnošću od 0.99884 te f1 mjere od 0.97456. Servis za sigurno pretraživanje pronašao je 43 zlonamjerna URL-a hr domeni. Rezultati strojnog učenja nad hr skupom podataka postigli su jako loše rezultate te je potreban ručni pregled stranica pomoću dodatnih značajki dobivenih statičkom analizom web stranice.

Summary

With the increasing number of active websites, criminal activities on the Internet are also increasing. Web-based attackers try to trick users into providing sensitive information or downloading and running malicious programs. In order to prevent such activities, this paper deals with the detection of malicious websites using machine learning. The MalCrawler dataset was studied and scripts were written to obtain its features in order to create a new dataset over pages in the hr domain. The results of machine learning and MalCrawler achieve very good results where the best result was achieved by the C4.5 decision tree algorithm with an accuracy of 0.99884 and an f1 measure of 0.97456. The Google Safe Browsing service found 43 malicious URLs in the hr domain. The results of machine learning on hr dataset give very poor results. A manual review of the web page is required by using additional features obtained from static analysis of the web page.