UNIVERZITET U SARAJEVU

ELEKTROTEHNIČKI FAKULTET

ODSJEK ZA RAČUNARSTVO I INFORMATIKU

DIJAGNOZA BOLESTI PACIJENTA NA OSNOVU ZADANIH SIMPTOMA

Projekat iz predmeta

VJEŠTAČKA INTELIGENCIJA

Studenti:
Šahbaz Emina 19076,
Džinić Edna 19185,
Kovačević Nađa 19182

Sadržaj

Opis problema ... **3**
Definicija osnovnih pojmova i koristi ... 4
Pregled postojećih dataset-ova ... 5
Pregled stanja u oblasti ... **6**
Trenutno stanje i korištene metode ... 6
Postignuti rezultati ... 6
Identifikovani izazovi i pravci poboljšanja ... 6
Izbor, analiza i pretprocesiranje dataset-a ... **8**
Struktura dataset-a ... 8
Analiza i pretprocesiranje ... 8
Identifikovani rizici ... 9
Pripreme za modeliranje ... 9
Odabir, formiranje, treniranje i testiranje modela ... **11**
Rezultati treniranja modela ... 12
Testiranje modela na nepoznatim podacima ... 15
Cjelokupni osvrt na problem i dobijeno rješenje ... **16**

Opis problema

U savremenom zdravstvenom sistemu, pravovremena i tačna dijagnoza bolesti predstavlja jedan od najvažnijih koraka ka uspješnom liječenju pacijenata. Tačna identifikacija bolesti omogućava pravovremenu terapiju, smanjuje rizik od komplikacija i optimizuje korištenje medicinskih resursa. Ipak, uprkos napretku u medicini, proces dijagnostike i dalje predstavlja veliki izazov, posebno kada se uzme u obzir veliki broj potencijalnih bolesti i preklapanje simptoma među njima. Mnogi simptomi su nespecifični i mogu biti zajednički za više različitih oboljenja, što dodatno komplikuje proces donošenja medicinskih odluka.

U tom kontekstu, ovaj projektni zadatak ima za cilj razvoj inteligentnog sistema koji, na osnovu unosa simptoma od strane pacijenta, može predložiti vjerovatnu dijagnozu bolesti. Sistem se zasniva na primjeni metoda mašinskog učenja, koje omogućavaju računarima da „uče“ iz prethodnih podataka i identifikuju obrasce između simptoma i bolesti. Korištenjem relevantnih skupova podataka, kao što je *Disease Symptom Description Dataset*, model se može obučiti da automatski prepozna bolesti na osnovu kombinacije simptoma koje korisnik unese.

Cilj sistema nije da zamijeni ljekara, već da služi kao pomoćno sredstvo u procesu dijagnoze. Doktori mogu koristiti ovakav alat kao dodatnu podršku pri donošenju odluka, naročito u slučajevima kada postoji više mogućih dijagnoza. Sa druge strane, pacijenti mogu dobiti brzu preliminarnu analizu simptoma, što im može pomoći da odluče da li je potrebno hitno potražiti medicinsku pomoć ili preduzeti određene mjere opreza.

Primjena ovakvog sistema ima potencijal da: ubrzava dijagnostičke procese, smanji broj pogrešnih ili propuštenih dijagnoza, olakša rad medicinskog osoblja i poveća dostupnost zdravstvenih informacija za širu populaciju, posebno u sredinama sa ograničenim pristupom doktorima.

Uvođenje inteligentnih sistema u medicinu predstavlja važan korak ka digitalizaciji zdravstva i personalizovanoj medicini budućnosti.

Definicija osnovnih pojmova i koristi

U ovom projektnom zadatku koristi se dataset pod nazivom "*Disease Symptom Description Dataset*", preuzet sa Kaggle platforme, koji sadrži informacije o bolestima, njihovim simptomima, opisima i preporučenim mjerama opreza. Na osnovu ovog skupa podataka definišemo nekoliko ključnih pojmova koji čine osnovu zadatka. Simptomi su karakteristike koje pacijent osjeća ili izražava, a koje ukazuju na prisustvo određenog zdravstvenog problema.

U ovom dataset-u, simptomi su predstavljeni kao niz pojmova povezanih sa svakom bolešću (npr. glavobolja, mučnina, umor). Svaka bolest je povezana sa jednom ili više simptoma, što omogućava modelima mašinskog učenja da nauče obrasce između simptoma i bolesti. Bolest je oboljenje ili stanje koje remeti normalne fiziološke funkcije organizma. U dataset-u je svaka bolest zapisana kao posebna kategorija, uz odgovarajuće simptome, opis i mjere opreza. Bolesti predstavljaju ciljne vrijednosti koje model pokušava da predvidi na osnovu simptoma. Opis bolesti (*Description*) daje kraći pregled same bolesti – uključujući osnovne informacije o njenim uzrocima, simptomima i težini.

Ovi tekstualni podaci mogu se koristiti za dodatne analize ili informisanje korisnika o potencijalnim dijagnozama. Mašinsko učenje (eng. *machine learning*) je tehnologija koja omogućava računarima da uče iz podataka bez eksplicitnog programiranja. U ovom kontekstu, mašinsko učenje se koristi za prepoznavanje obrazaca između simptoma i bolesti, kako bi se omogućilo automatsko predlaganje dijagnoze.

Korištenjem ovog konkretno strukturiranog dataset-a moguće je izgraditi sistem koji može značajno pomoći u ranoj i brzinskoj dijagnostici. Takav sistem može imati višestruke koristi. Ljekarima može služiti kao alat za podršku pri odlučivanju, naročito u slučajevima kada je broj potencijalnih dijagnoza veliki. Za pacijente, sistem može omogućiti preliminarnu procjenu na osnovu unesenih simptoma, što može biti korisno u situacijama kada medicinska pomoć nije odmah dostupna. Pored toga, pružanje preporučenih mjera opreza doprinosi većoj informisanosti korisnika i boljem upravljanju zdravstvenim stanjem. Na ovaj način, uz pomoć pouzdanih podataka i algoritama, moguće je unaprijediti proces zdravstvene zaštite, smanjiti opterećenje na ljekare, te povećati dostupnost osnovnih dijagnostičkih informacija široj populaciji.

Pregled postojećih dataset-ova

Za razvoj sistema koji automatski predlaže dijagnozu bolesti na osnovu simptoma, od ključne je važnosti odabrati odgovarajući skup podataka. Na raspolaganju je više javno dostupnih dataset-ova koji se mogu koristiti u ovu svrhu, pri čemu svaki od njih nudi specifične prednosti u zavisnosti od ciljeva istraživanja i metoda koje se planiraju primijeniti.

Jedan od razmatranih dataset-ova je ***Disease Symptom Description Dataset***, dostupan na Kaggle platformi. Ovaj skup podataka sadrži informacije o bolestima, njihovim simptomima, kratkim tekstualnim opisima i preporučenim mjerama opreza. Njegova prednost ogleda se u jasno definisanoj vezi između simptoma i bolesti, ali i u dodatnim opisnim kolonama koje omogućavaju dublju analizu kroz tehnike obrade prirodnog jezika (NLP). Dataset je jednostavan za korištenje jer je predstavljen u CSV formatu, što omogućava brzo učitavanje i obradu podataka. također, pruža dodatnu vrijednost kroz preporuke koje mogu poslužiti korisnicima kao osnovne smjernice za ponašanje nakon što dobiju sugestiju o mogućoj bolesti. Ovaj dataset je jednostavan, često korišten u edukativne svrhe i pogodan za razvoj klasifikacionih modela sa manjom složenošću.

Drugi relevantan izvor podataka je ***SymCat Disease Symptoms Dataset***, koji se zasniva na stvarnim medicinskim zapisima. Ovaj dataset uključuje podatke o učestalosti simptoma kod različitih bolesti, što omogućava izradu statističkih modela za procjenu vjerovatnoće određenih dijagnoza. Njegova snaga leži u realnim kliničkim podacima i mogućnosti primjene u bayesovskim i drugim analizama zasnovanim na vjerovatnoći.

U okviru UCI Machine Learning Repository-ja dostupni su različiti medicinski dataset-ovi, kao što su oni za dijabetes, bolesti srca i druge hronične bolesti. Ovi dataset-ovi su dobro strukturirani, često korišteni u naučnim istraživanjima i pogodni za poređenje različitih mašinsko-učećih modela na specifičnim medicinskim stanjima.

Microsoft i Svjetska zdravstvena organizacija također su pokrenuli inicijativu pod nazivom *AI for Health*, koja uključuje brojne otvorene skupove podataka fokusirane na globalne zdravstvene izazove. Ovi podaci omogućavaju analize koje povezuju zdravstvene informacije sa populacijskim i prostornim faktorima, te služe za razvoj praktičnih i društveno korisnih rješenja u zdravstvu.

Iako svi navedeni dataset-ovi nude korisne informacije za rješavanje problema automatske dijagnoze, ***Disease Symptom Description Dataset*** se izdvaja kao posebno pogodan za ovaj projekat. Njegova kombinacija strukturiranih simptoma, tekstualnih opisa i preporuka pruža solidnu osnovu za razvoj inteligentnog sistema koji ne samo da predlaže potencijalnu dijagnozu, već i informiše korisnika o prirodi bolesti i daljim koracima koje može preduzeti.

Pregled stanja u oblasti

U posljednjih nekoliko godina, oblast primjene vještačke inteligencije (AI) u medicini doživjela je snažan rast. Jedan od najaktivnijih pravaca istraživanja jeste upravo automatska dijagnoza bolesti na osnovu simptoma, gdje se koriste različite metode mašinskog učenja (ML) i obrade prirodnog jezika (NLP) za analizu medicinskih podataka.

Trenutno stanje i korištene metode

Razna istraživanja su pokazala da se za predikciju bolesti koriste klasifikacioni algoritmi kao što su: *Decision Trees* (DT), *Random Forest* (RF), *Naive Bayes* (NB), *Support Vector Machines* (SVM), *K-Nearest Neighbors* (KNN), *Neural Networks* (NN), uključujući i duboko učenje.

Na primjer, istraživanje objavljeno na Google Scholar-u („*Disease Prediction using Machine Learning algorithms*“) pokazuje da *Random Forest* model može postići tačnost veću od 90% na dataset-ovima sličnim onom koji se koristi u ovom projektu, zahvaljujući njegovoj sposobnosti rada sa velikim brojem atributa (simptoma).

Također, u nekim radovima se koristi NLP za obradu opisa bolesti i simptoma kako bi se izvukle dodatne semantičke informacije koje mogu pomoći u klasifikaciji. Korištenjem vektorskih reprezentacija teksta (npr. TF-IDF, Word2Vec, BERT), moguće je unaprijediti performanse modela.

Postignuti rezultati

Na osnovu postojećih rješenja dostupnih na Kaggle-u i u naučnoj literaturi, modeli trenirani na simptom-bolest dataset-ovima mogu postići visoku preciznost (85–95%), zavisno od kvaliteta i obrade podataka. *Random Forest* i *Naive Bayes* se najčešće ističu kao najefikasniji algoritmi za ovu vrstu klasifikacije.

Jedan primjer sa Kaggle-a pokazuje implementaciju višeklasne klasifikacije koristeći Naive Bayes model, koji je postigao preciznost od 91%, uz jednostavnu obradu simptoma kao binarne vektore.

Identifikovani izazovi i pravci poboljšanja

Nejedinstveni zapisi simptoma: Simptomi se u tekstu često zapisuju različito (npr. "*headache*", "*severe headache*", "*head pain*"), što može otežati tačnu analizu. Primjena NLP tehnika za standardizaciju simptoma može pomoći.

Nedostatak hijerarhije simptoma: Većina modela tretira sve simptome kao jednake, iako neki imaju veću dijagnostičku vrijednost. Težinska obrada simptoma (*feature weighting*) je jedan pravac za istraživanje.

Ograničena pokrivenost dataset-a: Dataset sadrži ograničen broj bolesti i simptoma, pa generalizacija na šire populacije može biti upitna. Mogućnost proširenja dataset-a kombinovanjem sa drugim izvorima može povećati performanse modela.

Izbor, analiza i pretprocesiranje dataset-a

Izvor skupa podataka: Kaggle (https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset?select=dataset.csv)
Format u kojem je dostupan i način za preuzimanje: CSV fajlovi dostupni za preuzimanje
Broj instanci: 4920
Broj atributa (ukoliko ih ima): 18 (1 kolona za bolest + 17 kolona za simptome)
Broj klasa (ukoliko je u pitanju klasifikacija): 41 jedinstvenih bolesti
Broj instanci po klasama (ukoliko je u pitanju klasifikacija): 120 (ravnomjerno raspoređeno)
Količinu podataka (MB): 0.6
Podaci za treniranje/validaciju/testiranje: Podaci su podijeljeni u standardnom omjeru 80/20 za treniranje i testiranje

Struktura dataset-a

Dataset se sastoji od četiri glavna CSV fajla:
dataset.csv - glavni fajl sa simptomima i bolestima
Symptom-severity.csv - težine simptoma (rangovi od 1-7)
symptom_Description.csv - opisi bolesti
symptom_precaution.csv - preporučene mjere opreza za svaku bolest

Analiza i pretprocesiranje

Čišćenje podataka:

- Uklonjeni su nepotrebni razmaci iz naziva kolona i vrijednosti
- Svi nazivi su pretvoreni u mala slova radi konzistentnosti
- NaN vrijednosti su zamijenjene sa "none"

Analiza distribucije simptoma: Analiza je pokazala da postoji značajna neravnomjernost u frekvenciji pojavljivanja simptoma. Najčešći simptomi poput "*fatigue*" i "*vomiting*" se pojavljuju znatno češće od ostalih, što može uticati na performanse modela.
Analiza težina simptoma: Korištenjem Symptom-severity.csv fajla analizirane su težine simptoma, gdje se vrijednosti kreću od 1 do 7. Ova informacija je korisna za buduće poboljšanje modela kroz uvođenje težinskih faktora.
Distribucija simptoma po bolestima: Prosječan broj simptoma po bolesti varira, što ukazuje na različitu složenost dijagnostike pojedinih oboljenja. Neke bolesti imaju veliki broj specifičnih simptoma, dok druge dijele mnoge zajedničke simptome.

Identifikovani rizici

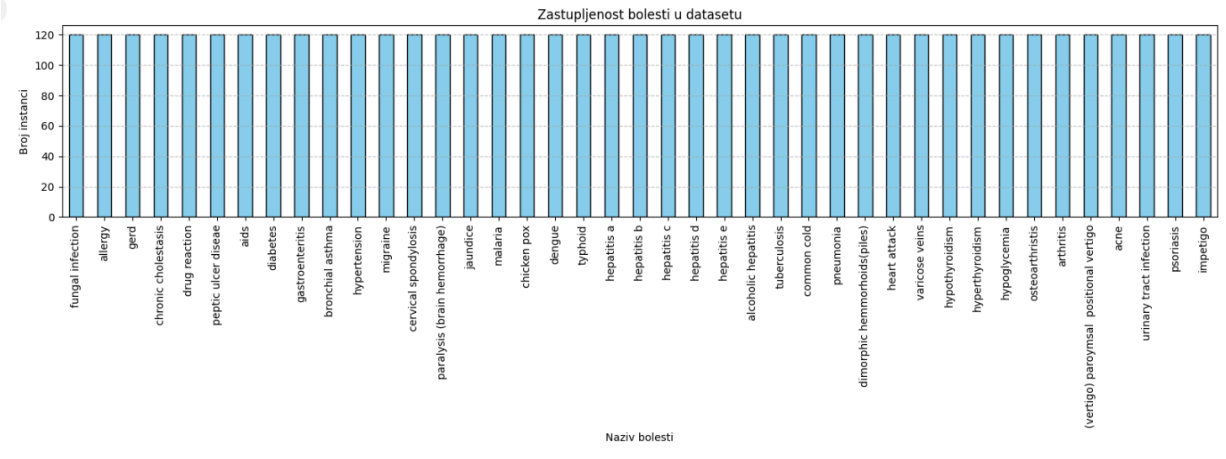
1. Preklapanje simptoma: Mnoge bolesti dijele iste simptome, što može otežati preciznu klasifikaciju
2. Neravnomjerna distribucija simptoma: Neki simptomi su znatno češći od drugih
3. Ograničenost dataset-a: Sadrži samo 41 bolest, što ne pokriva cijeli spektar mogućih oboljenja
4. Nedostatak konteksta: Simptomi se tretiraju nezavisno, bez uzimanja u obzir njihove međusobne povezanosti

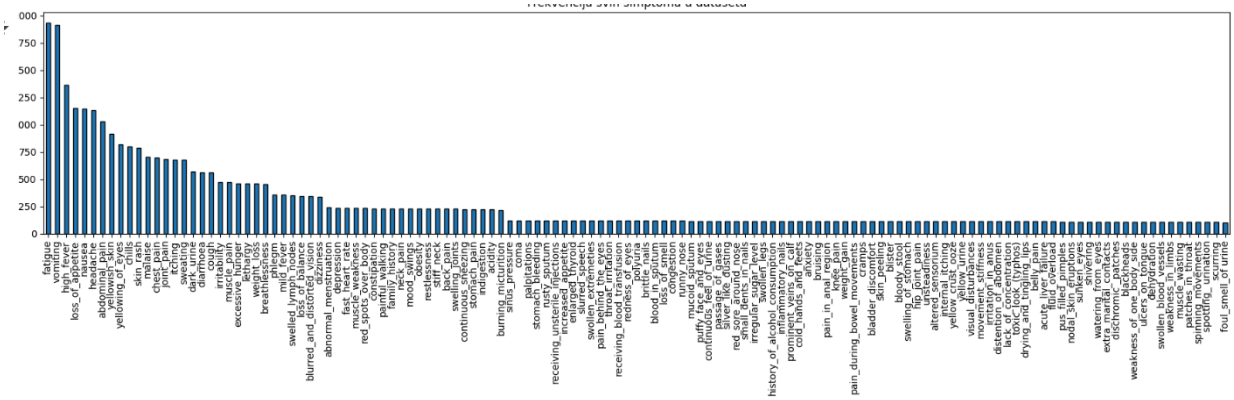
Pripreme za modeliranje

Podaci su transformisani u format pogodan za mašinsko učenje kroz:

- Vektorizaciju simptoma (binarna reprezentacija)
- Kodiranje bolesti u numeričke klase
- Kreiranje feature matrice za treniranje modela

Prije bilo kakve obrade modela, bilo je neophodno pretprocesirati podatke. Svaka kolona i vrijednost je očišćena od nepotrebnih razmaka, pretvorena u mala slova i zamijenjene su sve NaN vrijednosti sa "*none*" radi lakšeg rada. Simptomi su spojeni u jednu listu po instanci, a zatim vektorizovani tako sto je svakom unikatnom simptomu pridružena vrijednost u vektoru, čime je kreirana binarna matrica simptoma po instanci. Bolesti su zatim pretvorene u numeričke klase pomoću `*LabelEncoder*` klase. Ovaj korak pretprocesiranja je ključan jer bez normalizacije teksta, vektorizacije i transformacije klasa, model ne bi mogao efektivno učiti. Tokom analize, utvrđeno je da su sve bolesti u datasetu zastupljene ravnomjerno, što možemo vidjeti na sljedećoj slici:





Ovaj histogram prikazuje koliko se puta svaki simptom pojavljuje u cijelom datasetu, pri čemu su neki simptomi poput “*fatigue*” i “*vomiting*” znatno češći od drugih, što ukazuje na neravnomjernu raspodjelu simptoma među bolestima.

Odabir, formiranje, treniranje i testiranje modela

Za rješavanje problema dijagnoze bolesti na osnovu simptoma, izabrana je metoda višeklasne klasifikacije pomoću neuronske mreže. Odabrana tehnologija je Python sa TensorFlow i Keras bibliotekom za kreiranje, treniranje i evaluaciju modela. Korištene su i Pandas, NumPy, Matplotlib, Sklearn za manipulaciju podacima i vizualizaciju. Format podataka je pripremljen tako što su simptomi vektorizovani kao binarni niz (*one-hot encoding*), a bolesti su enkodirane pomoću *LabelEncoder* i konvertovane u kategorijske varijable korištenjem *to_categorical()*.

Model ima tri sloja:

- Prvi sloj sadrži 128 neurona i koristi ReLU aktivaciju, broj ulaznih parametara predstavlja broj siptoma

- Drugi sloj sadrži 64 neurona, također s ReLU funkcijom

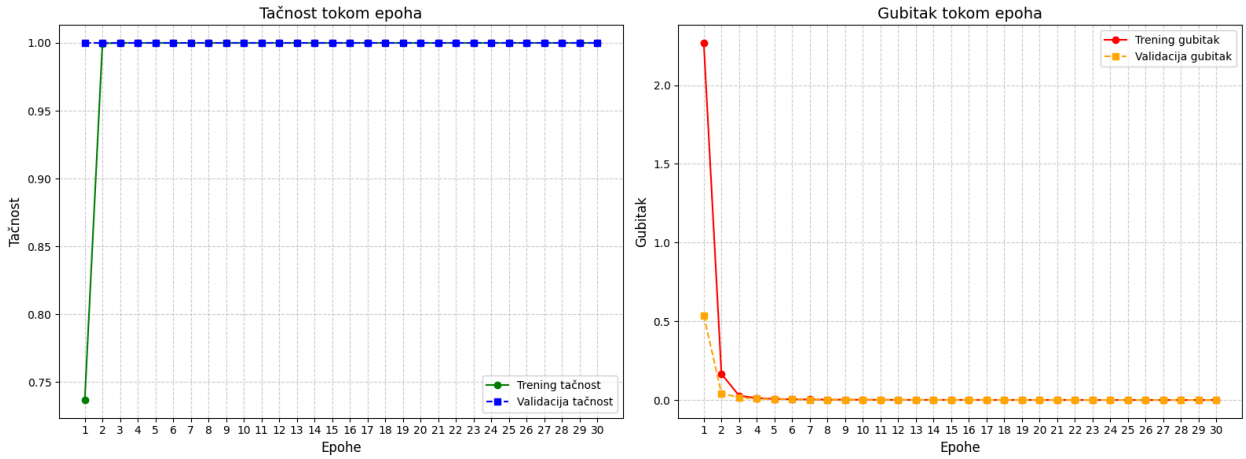
- Izlazni sloj ima 41 neuron (za 41 bolest) sa softmax aktivacijom, koja vraća vjerovatnoće za svaku klasu.

Model je kompajliran sa Adam optimizatorom i *categorical_crossentropy* funkcijom gubitka, a metrika za evaluaciju je tačnost (*accuracy*). Treniranje je izvršeno u trajanju od 30 epoha sa batch veličinom 32. Dataset je podijeljen na podatke za treniranje i testiranje u omjeru 80:20 respektivno. Rezultati treniranja su pokazali izuzetno visoku tačnost - preko 99% već od druge epohe. Na validacionom skupu model je postizao 100% tacnost, kao i na test skupu. Ovi rezultati upućuju na vrlo uspješno učenje, ali i na potencijalni *overfitting*, s obzirom da nema grešaka u predikciji. Također, do ovih rezultata dovodi i činjenica da su za svaku bolest pojedinačno svi simptomi slični, tako da je prilikom testiranja model na osnovu simptoma mogao tačno znati o kojoj se bolesti radi bez dvojbe. (primjer bolesti i simptoma na slici ispod)

Gastroenteritis	vomiting	sunken_eyes	dehydration	diarrhoea
Gastroenteritis	vomiting	sunken_eyes	dehydration	diarrhoea
Gastroenteritis	sunken_eyes	dehydration	diarrhoea	
Gastroenteritis	vomiting	dehydration	diarrhoea	
Gastroenteritis	vomiting	sunken_eyes	diarrhoea	
Gastroenteritis	vomiting	sunken_eyes	dehydration	
Gastroenteritis	vomiting	sunken_eyes	dehydration	diarrhoea
Gastroenteritis	sunken_eyes	dehydration	diarrhoea	
Gastroenteritis	vomiting	dehydration	diarrhoea	
Gastroenteritis	vomiting	sunken_eyes	diarrhoea	

Rezultati treniranja modela

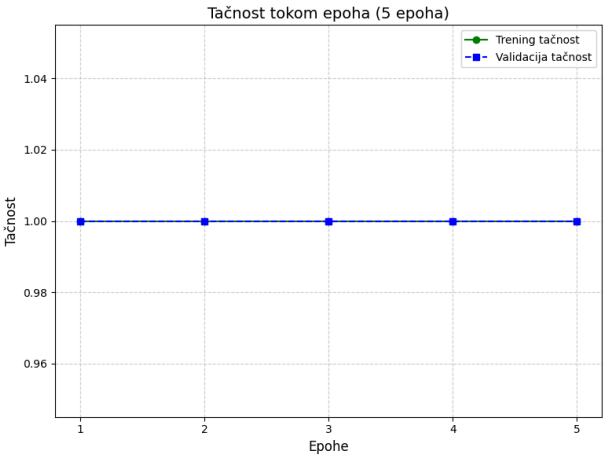
Grafik tačnosti i gubitka po epohama

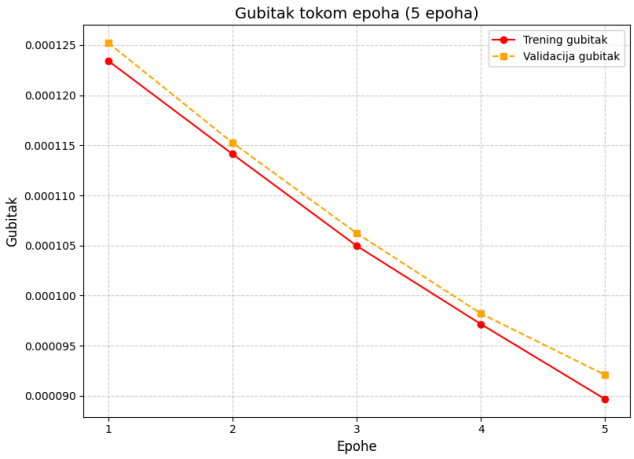


Na lijevom histogramu imamo prikazanu tačnost po epohama. Zeleni kružići predstavljaju tačnost na trening skupu, dok plavi kvadratići predstavljaju tačnost na skupu za validaciju. Već od 2. epohe, model postiže 100% tačnost na oba skupa i tu tačnost zadržava do kraja treniranja (30 epoha). Ovakav rezultat sugeriše da je model izuzetno dobro naučio uzorke iz podataka, ali može ukazivati i na moguće preučenje (*overfitting*) ako su podaci jednostavni ili nedovoljno raznovrsni što je vjerovatno slučaj u našem datasetu.

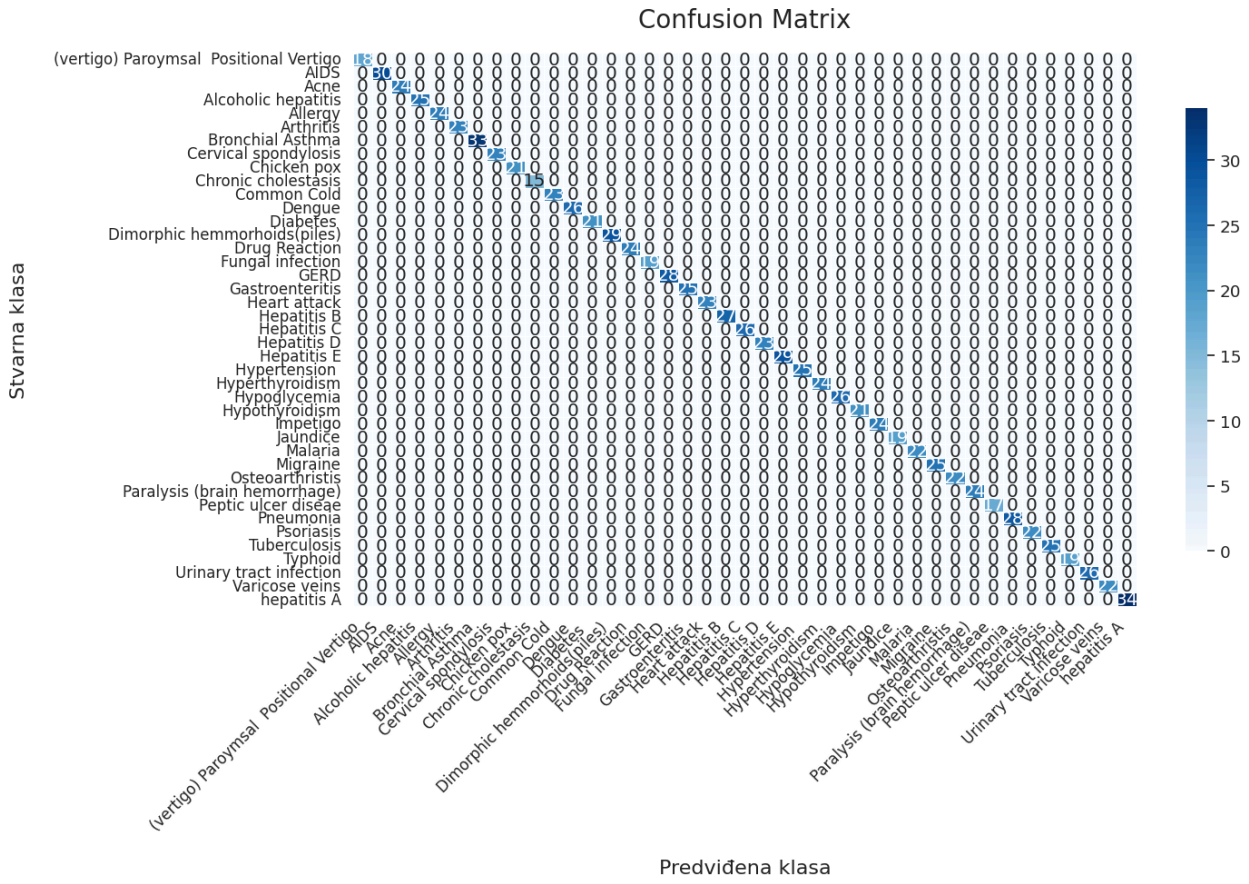
Desni histogram prikazuje gubitak tokom epoha. Crvena linija prikazuje gubitak na trening skupu, koji naglo opada nakon prve epohe i brzo se približava nuli. Narandžasta isprekidana linija prikazuje gubitak na validacionom skupu, koji također brzo opada i postaje gotovo zanemarljiv. Ovako nizak gubitak uz visoku tačnost dodatno potvrđuje da je model odlično naučio podatke.

Na osnovu prikazanih grafova, vidimo da se nakon 3-4 epohe vrijednosti tačnosti i gubitka više skoro ne mijenjaju. To znači da se model može trenirati na manjem broju epoha bez gubitka performansi.


Tačnost tokom epoha (5 epoha)
Trening tačnost
Validacija tačnost
Tačnost
1.04
1.02
1.00
0.98
0.96
1
2
3
4
5
Epohe


Gubitak tokom epoha (5 epoha)
Trening gubitak
Validacija gubitak
Gubitak
0.000125
0.000120
0.000115
0.000110
0.000105
0.000100
0.000095
0.000090
1
2
3
4
5
Epohe

Konfuzijska matrica



Na slici je prikazana konfuzijska matrica koja prikazuje koliko je model uspješno klasifikovao svaku klasu. Vrijednosti duž dijagonale (npr. 25, 26, 24...) označavaju broj ispravno klasifikovanih instanci za svaku klasu. Konfuzijska matrica je savršeno dijagonalna – nema ni jedne pogrešne klasifikacije – iz nekoliko međusobno povezanih razloga koji proizlaze iz prirode samog dataset-a i načina na koji je model građen. Ti razlozi su:

1. **Jedinstveni, nepreklapajući vektori simptoma**
   U dataset-u svaka bolest je predstavljena tačno određenom kombinacijom simptoma. Modelu je zato dovoljno da nauči precizan mapping iz tog unikatnog binarnog vektora u klasu.
2. **Visoka dimenzionalnost i rijetkost**
   Broj mogućih simptoma (dimenzija) je velik, a u svakom uzorku samo relativno mali broj njih bude “1”. U takvom prostoru su različite klase vrlo udaljene jedna od druge, što dodatno olakšava razdvajanje i sprječava preklapanje.
3. **Mali broj uzoraka po klasi bez šuma**
   Obzirom da za svaku bolest imamo dovoljno velik broj uzoraka i nijedan uzorak ne sadrži lažne ili izostavljene simptome, model se vrlo brzo “nauči” sve moguće kombinacije i ne stvara konfuziju između klasa.

4. **Treniranje do prekomjernog učenja (*overfitting*)**
   Model postiže 100 % tačnost već u drugoj epohi i zadržava je. To je tipično znak da je model "memorisao" trening i test primjere do te mjere da niti jedan test-primjer više ne bude pogrešno klasificiran.

Testiranje modela na nepoznatim podacima

U sklopu koda koji je korišten za pravljenje modela, treniranje i testiranje, također je napisan i dio kojim možemo sami testirati model.
Ako modelu unesemo sve nepoznate simptome on nam vraća poruku "Nijedan validan simptom nije prepoznat.". Ako modelu unesemo neke poznate simptome u kombinaciji sa nepoznatim, on vraća ispravne rezultate, što znači da model ispravno radi.

Cjelokupni osvrt na problem i dobijeno rješenje

Ukoliko uporedimo naše rezultate sa nekim drugim, npr.: *Symptoms Based Disease Prediction*, gdje je korišten sličan pristup-vektorski prikaz simptoma i neuronska mreža, vidimo da daju približno istu tačnost (99%-100%). Međutim, kao i u našem slučaju, trebalo bi se izvršiti testiranje na vanjskim podacima (koji nisu iz datasets) kako bi se provjerila sposobnost generalizacije modela.

Naš model daje stopostotnu tačnost što znači da je model dobar. To može značiti i da je model previše naučen, međutim prilikom ručnog testiranja sa podacima koji nisu iz dataset-a model je idalje na izlazu imao dobre rezultate čime demantujemo činjenicu da je došlo do overfittinga. Razlog zašto model ispravno radi leži u tome da model ima veliki opseg simptoma, koji se ne poklapaju s drugim bolestima tako da model nema dvojbe o kojoj se bolesti radi. U slučaju da se unesu svi nepoznati simptomi, model ne vraća neku bolest po svom nahođenju nego ispisuje poruku da unos nije validan.

Za buduće učenje modela trebao bi se formirati novi dataset u kojem bi se simptomi bolesti više preklapali, tako da bi model zaista morao da “razmisli” koju će vrijednost vratiti na osnovu trening podataka, a ne da mu izlaz odmah bude poznat. Također, moglo bi se dodati da u odnosu na bolest model ispiše i neke savjete poput pijte više vode, odmarajte se, hitno posjetite liječnika i slično.