# Predikcija raka pluća primjenom metoda strojnog učenja

**Šereg, Matija**

**SVEUČILIŠTE U ZAGREBU**

**FAKULTET ORGANIZACIJE I INFORMATIKE**

**V A R A Ž D I N**

**Matija Šereg**

# PREDIKCIJA RAKA PLUĆA PRIMJENOM METODA STROJNOG UČENJA

**DIPLOMSKI RAD**

**Varaždin, 2022.**

**SVEUČILIŠTE U ZAGREBU**

**FAKULTET ORGANIZACIJE I INFORMATIKE**

**V A R A Ž D I N**

**Matija Šereg**

**Matični broj: 46226/17-R**

**Studij: Organizacija poslovnih sustava**

# PREDIKCIJA RAKA PLUĆA PRIMJENOM METODA STROJNOG UČENJA

**DIPLOMSKI RAD**

**Mentorica:**

Izv. prof. dr. sc. Dijana Oreški

**Varaždin, Rujan 2022.**

*Matija Šereg*

**Izjava o izvornosti**

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

*Autor potvrdio prihvaćanjem odredbi u sustavu FOI-radovi*

_____

**Sažetak**

Tema ovog rada je predikcija raka pluća primjenom metoda strojnog učenja. Ovaj rad se bavi izradom modela za predviđanje raka pluća pomoću tri različite metode strojnog učenja.

Modeli predviđanja omogućavaju predviđanje različitih stvari u ovom slučaju raka pluća. Također kod izrade modela predviđanja potrebno je odabrati odgovarajuću metodu strojnog učenja iz razloga jer nisu sve metode jednako dobre za svaku vrstu podataka. U ovom radu će se koristiti metode stablo odlučivanja, neuronska mreža te bayesova mreža.

Modeli predviđanja su jako korisni jer mogu pomoći kod na primjer predviđanja potražnje za nečim i slično. U ovom radu konkretno modeli mogu pomoći brže otkriti rak pluća ili povećani rizik za dobivanje istog tako da se može prije početi s liječenjem pa time automatski raste vjerojatnost za izlječenje.

Ovaj rad doprinosi razumijevanju modela predikcije, njihovih prednosti i nedostataka te različitih metoda strojnog učenja.

**Ključne riječi:** Sustav; Predviđanje; Stablo odlučivanja; Neuronska mreža; Bayesova mreža; Strojno učenje;

# Sadržaj

# 1. Uvod

Učinkovitost sustava predviđanja raka pomaže ljudima da znaju svoj rizik od raka uz nisku cijenu, a također pomaže ljudima da donesu odgovarajuću odluku na temelju njihovog statusa rizika od raka. Podaci se prikupljaju s web stranice online sustava za predviđanje raka pluća. Rak pluća je vodeći uzrok smrti od zloćudnih tumora kod muškaraca i žena. Od svih bolesnika koji imaju rak pluća, samo 15% preživi 5 godina od postavljanja dijagnoze. Najčešće se bolest otkrije kada je već uznapredovala pa je šansa za preživljavanje manja. Bitna je prevencija.

Cilj ovog rada je na temelju spomenutog skupa podataka identificirati i predvidjeti određeni rizik koji osoba ima, te navesti zašto je potencijalno došlo do raka pluća. Model predikcije na temelju karakteristika tj. simptoma koje čovjek ima, predviđa vjerojatnost da li čovjek ima ili nema rak pluća. Sustav se može koristiti za brzu dijagnozu raka pluća te se može prije početi s liječenjem, a time se povećava šansa za preživljavanjem. Korist od ovog sustava imaju liječnici jer im sustav pomaže kod dijagnoze, a također i bolesnici jer im može produžiti život ako se rak rano otkrije. U ovom radu će se putem web aplikacije BigML kreirati i koristiti stablo odlučivanja generirano uz pomoć baze podataka preuzeto s Kaggle platforme te uz pomoć neuronske i Bayesove mreže izraditi prediktivni model uz primjenu CRISP DM standarda.

Skup podataka koji će se upotrijebiti sastoji se od 16 atributa putem kojih ćemo provesti analizu i na kraju moći dati konkretniji zaključak. Ovaj rad će se raditi prema CRISP DM standardu koji se sastoji od sljedećih 6 koraka: razumijevanje domene, razumijevanje podataka, priprema podataka, modeliranje, vrednovanje i korištenje.

Za pisanje rada korišteni su izvori na internetu, a metoda prikupljanja podataka koja se koristila je Desk metoda. To je metoda u kojoj se koriste već dostupni podaci i informacije.

Diplomski rad se sastoji od trinaest poglavlja. Na početku rada su opisane metode strojnog učenja koje će se koristiti za izradu prediktivnih modela te je opisan CRISP DM standard prema kojem će rad biti izrađen. Zatim je napravljena analiza prethodnih istraživanja na ovu temu dok je ostatak rada fokusiran na izradu modela predikcije pomoću različitih metoda strojnog učenja. Na kraju se nalazi zaključak gdje su izneseni zaključci na temelju izrađenih modela predikcije.

Ova tema je značajna iz razloga što modeli predikcije predstavljaju jeftin, a i dosta učinkovit način da se predvide neke vrijednosti. U ovom slučaju želimo predvidjeti mogućnost raka pluća. Modeli predikcije mogu pomoći u spašavanju ljudskih života.

Motivacija za ovu temu je zanimljivost teme tj. Izrada modela predikcije mi se činila zanimljivom te sam htio još malo produbiti znanje o metodama strojnog učenja.

# 2. Metode strojnog učenja

U ovom poglavlju će biti navedene i objašnjene metode strojnog učenja koje će se koristiti za izradu modela predikcije raka pluća.

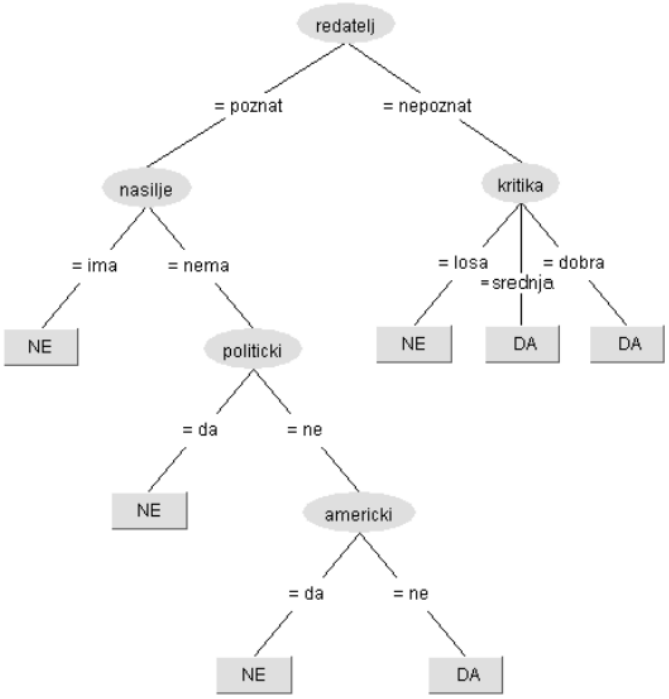
## 2.1 Stablo odlučivanja

Stablo odlučivanja je specifična vrsta dijagrama toka koji se koristi za vizualizaciju procesa donošenja odluka mapiranjem različitih tijekova djelovanja, kao i njihovih potencijalnih ishoda (Cravit, 2021.).

Druga definicija kaže da je stablo odlučivanja alat za podršku sa strukturom nalik stablu koji modelira vjerojatne ishode, troškove resursa i moguće posljedice. Uključuje grane koje predstavljaju korake donošenja odluka koji mogu dovesti do povoljnog rezultata (CFI, 2022.).

Stablo odlučivanja obično sadrži tri različita elementa (Cravit, 2021.):

- Korijenski čvor – čvor najviše razine predstavlja krajnji cilj ili veliku odluku koja se pokušava donijeti.
- Grana – grane koje proizlaze iz korijena predstavljaju različite opcije ili smjerove djelovanja koje su dostupne prilikom donošenja određene odluke. Najčešće su označene strelicom i često uključuju povezane troškove kao i vjerojatnost da će se pojaviti.
- Čvor list – pričvršćeni su na kraju grane te predstavljaju moguće ishode za svaku radnju. Obično postoje dvije vrste čvora lista:
  Četvrtasti čvor list koji označava drugu odluku koju treba donijeti.
  Okrugli čvor list koji ukazuje na slučajan događaj ili nepoznati ishod.




Slika 1: Stablo odlučivanja (Žitko, 2003.)

Stablo odlučivanja je nelinearno što znači da postoji puno više fleksibilnosti za istraživanje, planiranje i predviđanje nekoliko mogućih ishoda odluka bez obzira na to kada se one stvarno dogode. Također vizualno pokazuje uzročno-posljedične odnose pružajući pojednostavljeni pogled na potencijalno kompliciran proces. Stablo odlučivanja su jednostavna i lako razumljiva čak i ako ih nikad niste prije napravili. Usredotočena su na vjerojatnost i podatke, a ne na emocije i pristranost. Pruža uravnotežen pogled na proces donošenja odluka dok izračunava rizik. Isto tako stablo odlučivanja pojašnjava izbore, rizike, ciljeve i dobitke. Velika prednost stabla je njihov prediktivni okvir koji omogućava iscrtavanje različitih mogućnosti i u konačnici određivanje smjera djelovanja koji ima najveću vjerojatnost uspjeha. Ova prednost pomaže u zaštiti odluka od nepotrebnih rizika ili nepoželjnih ishoda (Cravit, 2021.).

Stablo odlučivanja je jedan od najboljih algoritama učenja koji se temelje na različitim metodama učenja. Pojačava prediktivne modele s točnošću, lakoćom interpretacije i stabilnošću. Postoje dvije glavne vrste stabla odlučivanja koje se temelje na ciljnoj varijabli, a to su stablo odlučivanja kategoričkih varijabli i stablo odlučivanja kontinuiranih varijabli (CFI, 2022.):

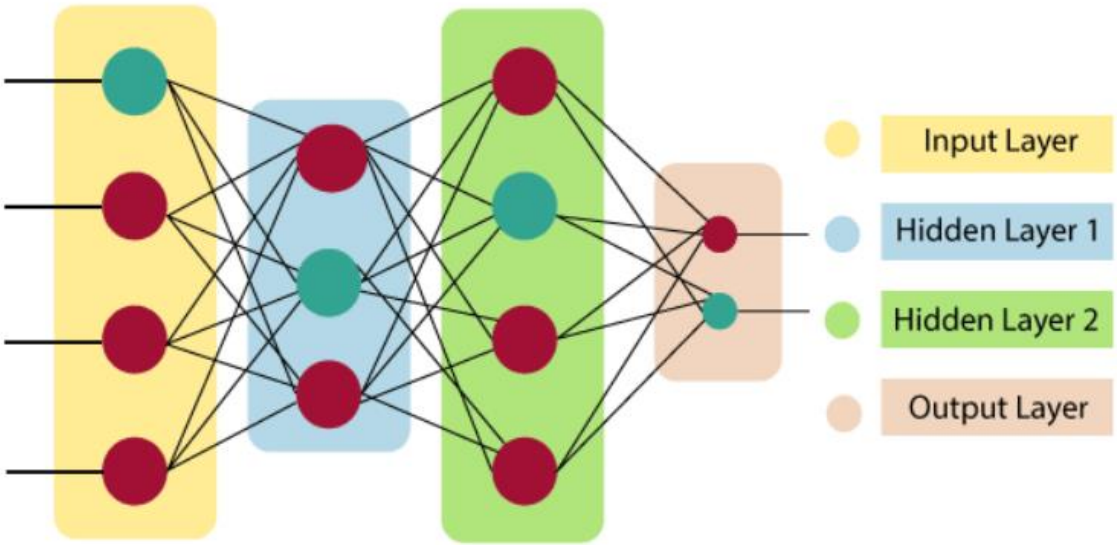
- Stablo odlučivanja o kategorijskim varijablama uključuje kategoričke ciljne varijable koje su podijeljene u kategorije npr. Kategorije mogu biti da ili ne. Kategorije znače da svaka faza procesa odlučivanja spada u jednu kategoriju i da nema između.

- Stablo odlučivanja kontinuirane varijable je stablo s kontinuiranom ciljnom varijablom npr. Prihod pojedinca čiji je prihod nepoznat može se predvidjeti na temelju dostupnih informacija kao što su zanimanje, dob i druge kontinuirane varijable.

Također stablo odlučivanja ima i neke nedostatke, a jedan od njih je da je stablo nestabilno u usporedbi s drugim prediktorima odluka. Mala promjena u podacima može rezultirati velikom promjenom u strukturi stabla odlučivanja što može prenijeti drugačiji rezultat od onoga što će korisnici dobiti u normalnom događaju. Isto tako stablo odlučivanja je manje učinkovito u predviđanju kada je glavni cilj predvidjeti ishod kontinuirane varijable. To je zato jer stablo gubi informacije pri kategorizaciji varijabli u više kategorija (CFI, 2022.).

## 2.2 Neuronska mreža

Izraz „umjetna neuronska mreža“ izveden je iz bioloških neuronskih mreža koje razvijaju strukturu ljudskog mozga. Slično ljudskom mozgu koji ima neurone međusobno povezane, umjetne neuronske mreže također imaju neurone koji su međusobno povezani u različitim slojevima mreža. Ovi neuroni su poznati kao čvorovi. Umjetna neuronska mreža u području umjetne inteligencije pokušava oponašati mrežu neurona koja čini ljudski mozak tako da će računala imati mogućnost razumjeti stvari i donositi odluke na način kako ih čovjek donosi. Dizajnirana je programiranjem računala da se ponaša kao međusobno povezane moždane stanice („Artificial neural networks“, bez dat.).




Slika 2: Arhitektura neuronske mreže („Artificial neural networks“, bez dat.)

Neuronska mreža se sastoji od ulaznog sloja, skrivenih slojeva te izlaznog sloja („Artificial neural networks“, bez dat.):

- Ulazni sloj prihvaća unose u nekoliko različitih formata.
- Skriveni sloj se nalazi između ulaznog i izlaznog sloja te izvodi sve izračune kako bi pronašao skrivene značajke i uzorke.
- Izlazni sloj - ulaz prolazi kroz niz transformacija pomoću skrivenog sloja što rezultira izlazom koji se prenosi pomoću ovog sloja. Neuronska mreža uzima ulaz i izračunava ponderirani zbroj ulaza i uključuje pristranost. Izračunavanje je predstavljeno u obliku prijenosne funkcije. Dostupne su različite prijenosne funkcije koje se mogu primijeniti na vrstu zadatka koji se obavlja.

Prednost neuronske mreže je što može obavljati više od jednog zadatka istovremeno. Također podaci koji se koriste pohranjuju se na cijeloj mreži, a ne u bazi podataka tako da nestanak par podataka na jednom mjestu ne sprječava rad mreže. Nakon treninga neuronske mreže, informacije mogu proizvesti izlaz i s čak neadekvatnim podacima. Gubitak izvedbe oslanja se na značaj podataka koji nedostaju. Da bi se neuronska mreža mogla prilagoditi važno je odrediti primjere i poticati mrežu prema željenom rezultatu. Uspješnost mreže izravno je proporcionalan odabranim instancama, a ako se događaj ne može prikazati mreži u svim svojim aspektima može doći do lažnog izlaza. Iznuđivanje jednog ili više čvora neuronske mreže ne zabranjuje generiranje izlaza što mrežu čini tolerantnom na greške.
Isto tako neuronska mreža ima i neke nedostatke. Ne postoje posebne smjernice za određivanje strukture umjetnih neuronskih mreža. Odgovarajuća mrežna struktura postiže se iskustvom, pokušajima i pogreškama. Jedan od nedostataka je i neprepoznato ponašanje mreže. To je najznačajniji problem neuronske mreže. Kada neuronska mreža izvede rješenje za testiranje ne daje uvid u to zašto i kako što smanjuje povjerenje u mrežu. Neuronske mreže trebaju procesore s paralelnom procesorskom snagom, prema njihovoj strukturi. Može raditi s brojčanim vrijednostima pa se problemi moraju pretvoriti u numeričke vrijednosti prije nego se uvedu u neuronsku mrežu. Nedostatak je što je mreža svedena na određenu vrijednost pogreške, a ta vrijednost ne daje optimalne rezultate („Artificial neural networks“, bez dat.).

## 2.3 Bayesova mreža

Bayesova mreža pruža jednostavan način primjene Bayesovog teorema na složene probleme. Mreže nisu baš Bayesove po definiciji, iako s obzirom na to da su obje distribucije vjerojatnosti za slučajne varijable (čvorove) i odnosi između slučajnih varijabli specificirani subjektivno može se reći da model obuhvaća „vjerovanje“ o kompleksnoj domeni. Bayesova

vjerojatnost je proučavanje subjektivnih vjerojatnosti ili vjerovanja u ishod u usporedbi s frekventističkim pristupom gdje se vjerojatnosti temelje isključivo na prošlom događanju događaja. Bayesova mreža opisuje zajedničku distribuciju vjerojatnosti za skup varijabli. Središnji dio Bayesove mreže je pojam uvjetne neovisnosti. Ovisnost se odnosi na slučajnu varijablu na koju druge varijable ne utječu. Zavisna varijabla je slučajna varijabla čija je vjerojatnost uvjetovana jednom ili više drugih slučajnih varijabli. Uvjetna neovisnost opisuje odnos između više slučajnih varijabli gdje određena varijabla može biti uvjetno neovisna o jednoj ili više drugih slučajnih varijabli. To ne znači da je varijabla neovisna sama po sebi, umjesto toga to je definicija da je varijabla neovisna o određenim drugim poznatim slučajnim varijablama. Pruža način definiranja vjerojatnosnog modela za složeni problem navodeći sve uvjetne pretpostavke neovisnosti za poznate varijable dok dopušta prisutnost nepoznatih varijabli (Brownlee, 2019.).

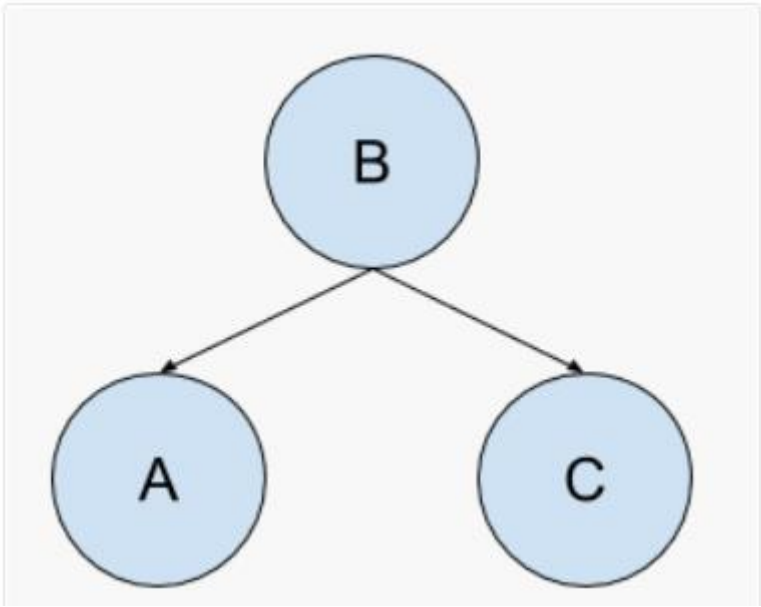
Bayesova mreža pruža korisne prednosti kao probabilistički model (Brownlee, 2019.):

- **Vizualizacija** – pruža izravan način vizualizacije strukture modela i motivira dizajn novih modela.
- **Odnosi** – pruža uvid u prisutnost i odsutnost odnosa između slučajnih varijabli.
- **Izračuni** – pruža način strukturiranja složenih izračuna vjerojatnosti.

Za izradu Bayesove mreže potrebno je definirati najmanje tri stvari (Brownlee, 2019.):

- **Slučajne varijable** – koje su slučajne varijable u problemu.
- **Uvjetni odnosi** – koji su uvjetni odnosi između varijabli.
- **Distribucije vjerojatnosti** – koje su distribucije vjerojatnosti za svaku varijablu.

Stručnjak u domeni problema može specificirati neke ili sve ove stvari u dizajnu modela. U mnogim slučajevima arhitekturu grafičkog modela može specificirati stručnjak, ali distribucije vjerojatnosti moraju se procijeniti iz podataka domene. Distribucija vjerojatnosti i struktura grafikona mogu se procijeniti iz podataka iako to može biti zahtjevan proces. Uobičajeno je koristiti algoritme za učenje u tu svrhu. Nakon što je mreža pripremljena za domenu, može se koristiti za razmišljanje npr. Donošenje odluka. Rezoniranje se postiže zaključivanjem s modelom za danu situaciju. Npr. Ishod nekih događaja je poznat i uključen u slučajne varijable. Model se može koristiti za procjenu vjerojatnosti uzroka događaja ili mogućih daljnjih ishoda (Brownlee, 2019.).




Slika 3: Jednostavna Bayesova mreža (Brownlee, 2019.)

# 3. CRISP DM metodologija

CRISP DM je kratica za Cross Industry Standard Process for Data Mining. To je model procesa koji služi kao baza za proces znanosti o podacima. Sastoji se od sljedećih šest koraka (Hotz, 2022.):

1. Poslovno razumijevanje – što je potrebno poduzeću?
2. Razumijevanje podataka – koje podatke trebamo/imamo?
3. Priprema podataka – kako organiziramo podatke za modeliranje?
4. Modeliranje – koje tehnike modeliranja trebamo primijeniti?
5. Evaluacija – koji model najbolje ispunjava poslovne ciljeve?
6. Implementacija – kako zainteresirane strane pristupaju rezultatima?

Ova metodologija je postala najčešća metodologija za rudarenje podataka, analitiku i projekte znanosti o podacima.

## 3.1 Poslovno razumijevanje

Ova faza se fokusira na razumijevanje ciljeva i potreba projekta. Osim trećeg zadatka, ostala tri zadatka u ovoj fazi su temeljne aktivnosti upravljanja projektima koje su univerzalne za većinu projekata (Hotz, 2022.):

- **Odrediti poslovne ciljeve**: prvo treba temeljito razumjeti iz poslovne perspektive što korisnik želi postići, a zatim definirati kriterije poslovnog uspjeha.
- **Procijeniti situaciju**: odrediti dostupnost resursa, projektne zahtjeve, procijeniti rizike i nepredviđene situacije te provesti analizu troškova i koristi.
- **Odrediti ciljeve rudarenja podataka**: treba odrediti kako uspjeh izgleda iz perspektive tehničkog rudarenja podataka.
- **Izraditi plan projekta**: odabrati tehnologije i alate te definirati detaljne planove za svaku fazu projekta.

## 3.2 Razumijevanje podataka

Dodavanje temelja poslovnog razumijevanja, usmjerava fokus na prepoznavanje, prikupljanje i analizu skupova podataka koji mogu pomoći ostvariti ciljeve projekta. Ova faza ima četiri zadatka (Hotz, 2022.):

- **Prikupiti inicijalne podatke**: prikupiti potrebne podatke i ako je potrebno učitati u alat za analizu.

- **Opisati podatke**: ispitati podatke i dokumentirati njihova površinska svojstva poput formata podataka, broja zapisa ili identiteta polja.
- **Istražiti podatke**: istražiti podatke dublje. Vizualizirati ih i identificirati odnose među podacima.
- **Provjeriti kvalitetu podataka**: koliko su podaci čisti, dokumentirati sve probleme s kvalitetom.

## 3.3 Priprema podataka

Ova faza koja se često naziva „provjera podataka“ priprema konačne skupove podataka za modeliranje. Ima pet zadataka (Hotz, 2022.):

- **Odabrati podatke**: odrediti koji će se skupovi podataka koristiti i dokumentirati razloge za uključivanje/isključivanje.
- **Očistiti podatke**: često najduži zadatak. Ispravljanje, imputiranje ili uklanjanje pogrešnih vrijednosti.
- **Konstruirati podatke**: izvesti nove atribute koji će biti od pomoći.
- **Integrirati podatke**: stvoriti nove skupove podataka kombiniranjem podataka iz više izvora.
- **Formatirati podatke**: po potrebi ponovno formatirati podatke.

## 3.4 Modeliranje

U ovoj fazi se izgrađuju i procjenjuju različiti modeli na temelju nekoliko različitih tehnika modeliranja. Sastoji se od četiri zadatka (Hotz, 2022.):

- **Odabrati tehnike modeliranja**: odrediti koje algoritme isprobati.
- **Generirati dizajn testa**: tijekom modeliranja možda će biti potrebno podijeliti podatke u skupove za trening, testiranje i validaciju.
- **Izraditi model**
- **Procijeniti model**: potrebno je interpretirati rezultate modela na temelju poznavanja domene, definiranih kriterija uspjeha i dizajna testa.

## 3.5 Evaluacija

Dok se zadatak procjene modela u fazi modeliranja fokusira na tehničku procjenu modela, faza evaluacije gleda šire na to koji model najbolje odgovara poslovanju i što učiniti sljedeće. Faza se sastoji od tri zadatka (Hotz, 2022.):

- **Ocijeniti rezultate**: ispunjavanju li modeli kriterije poslovnog uspjeha.
- **Proces pregleda**: pregledati obavljeni posao, ispraviti nešto ako je potrebno.
- **Odrediti sljedeće korake**: na temelju prethodna tri zadatka odrediti hoće li se nastaviti s implementacijom, ponavljati dalje ili pokrenuti nove projekte.

## 3.6 Implementacija

Ovisno o zahtjevima ova faza može biti jednostavna poput generiranja izvješća ili složena poput implementacije ponovljivog procesa rudarenja podataka u cijeloj tvrtki. Složenost ove faze varira, a sastoji se od četiri zadatka (Hotz, 2022.):

- **Planiranje implementacije**: razvoj i dokumentiranje plana za implementaciju modela.
- **Planiranje praćenja i održavanja**: kako bi se izbjegli problemi tijekom operativne faze modela.
- **Izrada konačnog izvješća**: sažetak projekta koji može uključivati konačnu prezentaciju rezultata.
- **Pregled projekta**: napraviti retrospektivu projekta o tome što je prošlo dobro, što je moglo bolje i kako poboljšati.


Slika 4: CRISP DM dijagram („Cross - industry standard process for dana mining“, 2022.)

# 4. Razumijevanje domene

CRISP-DM standard kaže kako je u ovoj fazi potrebno definirati poslovne ciljeve, procijeniti situaciju, definirati ciljeve rudarenja podataka te izraditi plan. Ovaj rad se bavi predviđanjem raka pluća pa bi domena bila predviđanje raka pluća. Poslovni cilj rada je izrada prediktivnog modela za predviđanje raka pluća. Podaci koji se koriste za izradu projekta su javno dostupni te su preuzeti sa stranice Kaggle. Cilj rudarenja podataka je izmjeriti pouzdanost kreiranog sustava.

Kako bi bolje razumjeli domenu potrebno je istražiti slična istraživanja te sustave. Slijedi pregled sličnih istraživanja koja su vezana za ovu temu.

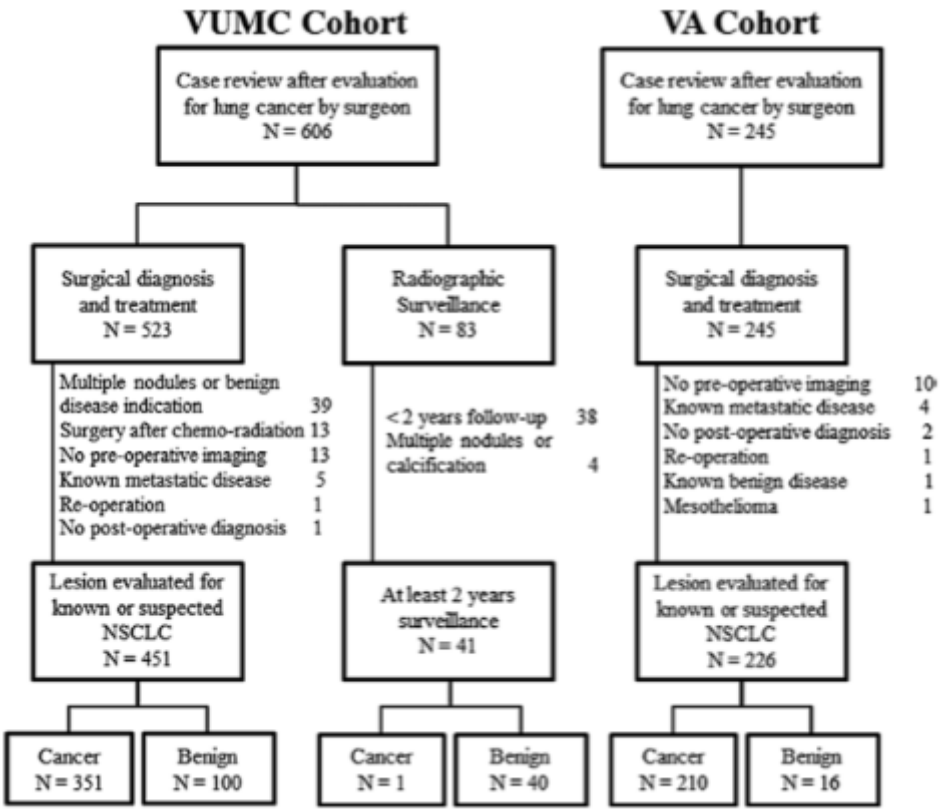
## 4.1 Metodologija

### 4.1.1 Predviđanje karcinoma pluća prije kirurške resekcije u bolesnika s plućnim čvorovima

Prema (A. Deppen I sur., 2014.), razvio se i interno potvrdio model kliničkog predviđanja za karcinom pluća u prospektivnoj skupini procijenjenog u ustanovi. Najbolje statističke prakse korištene su za konstruiranje, evaluaciju i validaciju modela logističke regresije u prisutnosti nedostajućih podataka koji se mljenjaju korištenjem tehnika pokretanja i korigiranja optimizma. Korišten je model tretmana (eng. treat).

Ova metoda razvijena je u organizaciji za rak pluća pod nazivom „Vanderbilt University Medical Center“ (VUMC) zajedno s skupinom iz Tennessee Valleya (VA) kako bi se ispitala sama generalizacija modela “treat”. Vanderbiltova skupina bila je sastavljena od pacijenata identificiranih iz dva odvojena izvora. Koristeći VUMC-ovu bazu podataka o poboljšanju kvalitete kirurgije i kliničku evidenciju, identificirano je 606 pacijenata koji su od siječnja 2005. do listopada 2010. godine dobili potvrdu da imaju plućni čvorić od strane kirurga zbog poznatog ili sumnjivog karcinoma pluća. Demografski i klinički podaci za svaki zahvat izvučeni su korištenjem Nacionalne baze podataka Društva za opću torakalnu kirurgiju. Podaci o slikama izvučeni su iz izvješća radiologa ili iz originalnih skeniranja najnovijih CT skeniranja. Na kraju, eliminacijom pacijenata koji nisu odgovarali za validaciju modela “treat” dobili u broj od 226 pacijenata (A. Deppen I sur., 2014).

S druge strane VA organizacija sastojala se od 245 pacijenata koji su sudjelovali u operaciji zbog poznatog ili sumnjivog karcinoma pluća. Pojedinci koji nisu odgovarali za “treat” model bili su isključeni. Preostalih 226 pacijenata korišteno je za validaciju modela.




Slika 5: Međunarodne udruge za proučavanje raka pluća (A. Deppen I sur., 2014)

### 4.1.2 Individualizirani model predviđanja rizika za rak pluća kod korejskih muškaraca

U ovom istraživanju su koristili Coxov model proporcionalnih opasnosti. Coxov model je u biti regresijski model koji se obično koristi statistički u medicinskim istraživanjima za istraživanje povezanosti između vremena preživljavanja pacijenata i jedne ili više prediktorskih varijabli. Ovaj model radi i za kvantitativne i za kategorijske varijable. Također proširuje metode analize preživljavanja kako bi se istovremeno procijenio učinak nekoliko čimbenika rizika na vrijeme preživljavanja. Model je jedna od najvažnijih metoda korištenih za modeliranje podataka analize preživljavanja. Omogućuje nam da ispitamo kako određeni čimbenici utječu na stopu određenog događaja u određenom trenutku. Valjanost modela su provjerili koristeći C statistiku i Hosmer – Lemeshow hi-kvadrat test na vanjskom skupu podataka. U kliničkim studijama C statistika daje vjerojatnost da je slučajno odabrani pacijent koji je doživio neki događaj (npr. smrt) imao višu ocjenu rizika od pacijenta koji nije doživio taj događaj. Hosmer – Lemeshow test se često koristi u modelima predviđanja rizika. Test ocjenjuje odgovaraju li opažene stope događaja očekivanim stopama događaja u podskupinama populacije modela (Park I sur., 2013.).

### 4.1.3 Model rizika za predviđanje raka pluća

U ovom istraživanju korišteni su epidemmiološki podaci od 1851 pacijenta koji su imali rak pluća te podaci od 2001 podudarnih kontrolnih subjekata. Ti podaci su slučajnim odabirom raspoređeni u odvojene setove. 75% podataka u skup za trening, a 25% podataka u skup za validaciju i to za bivše, sadašnje pušače i one koji nisu nikad pušili. Multivarijabilni modeli su izrađeni na temelju skupova za trening. Diskriminirajuća sposobnost modela je procijenjena u skupovima za validaciju ispitivanjem područja ispod operativnih karakterističnih krivulja i statistikom podudarnosti. Apsolutni jednogodišnji rizici od raka pluća su izračunati pomoću nacionalnih podataka o incidenciji i smrtnosti. Indeks rizika konstruiran je za svaku kategoriju pušača zbrajanjem omjera izgleda iz multivarijabilnih regresijskih analiza za svaki faktor rizika (R. Spitz i sur., 2007.).

### 4.1.4 Model predviđanja rizika od raka pluća - funkcija: Razvoj i validacija u UK Biobank

UK Biobank je kohortna studija u kojoj je sudjelovalo 502.321 sudionika bez prethodne dijagnoze raka pluća, uglavnom u dobi između 40 i 70 godina. Koristili su fleksibilne parametarske modele preživljavanja za procjenu dvogodišnje vjerojatnosti raka pluća, uzimajući u obzir konkurentni rizik od smrti.

Na početku procjene, prema (C. Muller i sur., 2017.) sudionici su ispunili upitnik koji je uključivao ove stavke koje procjenjuju opće zdravlje i povijest bolesti, način života i prehrane te obiteljsku povijest bolesti. Sudionici su također prošli fizičku procjenu koja je uključivala mjerenje krvnog tlaka, pulsa, visine, težine, tjelesne masti, vida, kondicije, snage hvata, gustoće kostiju i funkcije pluća. Praćenje incidencije i smrti od raka provedeno je uz pomoć povezivanja s dokumentacijskim registrima raka i smrti. Sudionici su praćeni od datuma pohađanja do datuma dijagnoze karcinoma pluća, datuma smrti ili do 1. siječnja 2012. Sam protokol studije je odobrio NWMREC (eng. North West Multicenter Research Ethics Committee) u Ujedinjenom Kraljevstvu.

Dodatni opis razvoja modela i procjena

Fokus je bio na ograničavanju analiza istrage na čimbenike koji su bili dosljedni povezivanja s rizikom od raka pluća. Pod te čimbenike podrazumijevalo se na one koji su rutinski dostupni ili ih liječnici opće prakse mogu lako utvrditi tijekom standardnih konzultacija. Posebno se istraživao spol, varijable vezane uz povijest pušenja i ovisnost o nikotinu, osobnu anamneza, obiteljska anamnezu raka pluća i funkciju pluća procijenjenu spirometrijom. Kako bi izračunali vjerojatnost raka pluća u prisutnosti konkurentnog rizika od smrti, zasebno se modelirala opasnost od raka pluća i opasnost od smrti od svih uzroka (C. Muller i sur., 2017.).

Koristili su se fleksibilne parametarske modele preživljavanja na kumulativnoj skali opasnosti za procjenu osnovnih opasnosti i omjera opasnosti. 33 godine starosti uzete su i korištene kao vremenska skala i modelirana je ograničenim kubičnim splineom s graničnim čvorovima na 40 i 70 godina i unutarnjim čvorovima na 60 i 65 godina. Odvojeni modeli za rizik od raka pluća postavljeni su za nikad, bivše i sadašnje sudionike pušača. Model rizika od smrti je uključivala spol, karcinom prije početka i pušenje kao kovarijate (vrijednosti koje se skupa mijenjaju). Sudionici s nedostajućim podacima o nekoj od kovarijata isključeni su iz primarnih analiza, a provedene su analize osjetljivosti na temelju više podataka. Neproporcionalne opasnosti istraživane su prilagodbom interakcija između kovarijata i vremenske skale. Dvogodišnji apsolutni rizik od raka pluća procijenjen je kumulativnom funkcijom incidencije koja je uvjetovana s dobi na početku pohađanja. Kako navodi [5], sve statističke analize provedene su korištenjem alata Stata 12.1 (Stata Corporation, College Station, Texas) i R verzija 3.2.1.35 (C. Muller i sur., 2017.).

### 4.1.5 Razvoj i validacija modela predviđanja rizika od raka pluća za Afroamerikance

Ova multirasna studija kontrole slučajeva raka pluća, od 1995. do 2005. uključila je sudionike studija sa Sveučilišta Texas i  Medicinskog centra., oba iz Houstona. Svi slučajevi s novim dijagnosticiranim, potvrđenim i neliječeni karcinomom pluća bili su podobni za istraživanje. Kriteriji za isključenje slučajeva iz studije uključivali su prethodnu kemoterapiju, radioterapiju ili nedavnu transfuziju krvi. Uključili su kontrolnu populaciju iz društvenih centara u području Houstona i iz Kelsey-Seybold klinike, najveće multi specijalističke grupe liječnika u Houstonu (L. Etzel i sur., 2008.).

Potencijalne kontrole prvo su ispitane kratkim upitnikom o njihovoj spremnosti da sudjeluju u istraživačkim studijama i daju preliminarne podatke koji su im pomogli u usklađivanju demografskih karakteristika. Kontrole su usklađene sa slučajevima na temelju dobi (±5 godina), spola i etničke pripadnosti. U ovoj analizi fokus je bio na podskup slučajeva

i kontrola ljudi koji su se sami prijavili kao Afroamerikanci. Ove osobe su predstavljale ~14% ukupne populacije istraživanja (L. Etzel i sur., 2008.).

Broj godina konzumiranja za bivše i sadašnje pušače izračunate su s godinama kada su pušili cigarete s prosječnim brojem cigareta koje su pušili dnevno te su to podijelili s 20. Pušači su također zamoljeni da prijave svoju upotrebu cigareta s mentolom, a bivši pušači da prijave dob u kojoj su prestali pušiti i broj godina koliko je prošlo od prestanka pušenja. Sudionike se također klasificiralo kao pozitivne ukoliko su bili izloženi azbestu najmanje 8 sati tjedno tijekom godine ili ako su prijavili da su radili u industriji povezanoj s azbestom. Ostala izloženost okolišu uključivala je izloženost tijekom rada ili hobija tijekom više od 8 sati tjedno tijekom jedne godine na piljevinu, vlakna (tekstilna vlakna ili pamuk), toluenu i/ili ksilenu (boja, otapala, razrjeđivači za boje i tinte za pisače) i pesticid. Također se klasificiralo sudionike prema medicinskim problemima koje su sami dijagnosticirali liječnici, uključujući astmu, kroničnu plućnu bolest i peludnu groznicu.Sudionici su sami prijavili povijest raka svojih najbližih. Konkretno, za svakog rođaka, zamolili su sudionike da navedu godinu rođenja, dob u vrijeme istraživanja, pušački status (nikad ili ikad), prisutnost ili odsutnost raka (da ili ne), vrstu raka, dob dijagnoze i godine smrti (L. Etzel i sur., 2008.).

U ovoj analizi usredotočili su se na obiteljsku povijest raka pluća ili drugih karcinoma povezanih s pušenjem (pluća, mjehura, bubrega, glave i vrata ili gušterače).

Statistička analiza

Koristila se deskriptivna statistička analiza za karakterizaciju populacije istraživanja, uključujući Pearsonov $\chi 2$ test za testiranje razlika u distribuciji između slučajeva i kontrola za kategoričke varijable i Studentov t test za određivanje razlika u kontinuiranim varijablama. Rabila se analiza logističke regresije za izračunavanje omjera vjerojatnosti te je pouzdanost intervala iznosila 95% kako bi se identificirao potencijalni čimbenik rizika za rak pluća prilagođene dobi, spola i povijesti pušenja.

Većina analiza, izvršene su korištenjem softvera “Statistical Analysis System” te kada je bilo potrebno, LogXact korišten je za točnu analizu logističke regresije za procjenu OR i izračunavanje pouzdanosti od 95% (L. Etzel i sur., 2008.).

Razvoj modela rizika

Čimbenici rizika uključeni su u multivarijabilni model rizika za rak pluća. Vršila se procjena interakcije između varijabli pušenja i izloženosti okolišu (azbest i drvena prašina) te su upotrijebili test omjera vjerojatnosti za testiranje značajnih interakcija. Kako bi identificirali druge moguće interakcije koje nisu naveli kao prioritet, izgradili su stablo klasifikacije koristeći tehniku. Kreirali su stablo odlučivanja tako da je svaki sljedeća podjela dala dva čvora s najmanje 10 sudionika po čvoru. Svaka grana koja se nije smatrala statistički značajnom na razini od 5% bila je odklonjena sa stabla (L. Etzel i sur., 2008.).

U procesu razvoja konačnog modela, najprije su se uključilo statistički značajni glavni učinci pušačkog statusa, pušačkih godina pušenja i dobi pri prestanku pušenja u multivarijabilni model. Međutim, također se uzelo u obzir i složenu varijablu koja je kombinirala ove tri varijable s “nikad” ne pušačima kao referencom. Na kraju, analitičari su usporedili informacijski sadržaj (AIC) između tri modela i model s najnižim AIC-om odabran je kao konačni model (L. Etzel i sur., 2008.).

Validacija modela

Konačni model znanstvenici, doktori i analitičari potvrdili su koristeći dva neovisna, interna i eksterna, skupa podataka.

Interna validacija temeljila se na dodatnim slučajevima i kontrolama iz iste multirasne studije kontrole slučajeva raka pluća u M. D. Anderson centru za rak pluća iz koje su dobiveni podaci za izgradnju modela. Ukupan broj sudionika analiziranih za internu provjeru bio je 156. Podaci za vanjsku validaciju bili su iz dvije različite studije o procjeni rizika od raka pluća među Afroamerikancima u gradu Detroitu. Prva studija bila je studija slučaj-kontrola o riziku od raka pluća ne-malih stanica među ženama. Druga studija bila je studija slučaja i kontrole rizika od raka pluća među rođacima ranih slučajeva raka pluća i kontrolne skupine. Zbog nedostajućih podataka na konačnom popisu varijabli modela, uključili su samo 325 sudionika za analizu vanjske validacije (L. Etzel i sur., 2008.).

Za svaki skup za validaciju izračunali su  površinu ispod krivulje (AUC) kako bi procijenili sposobnost modela da razlikuju afroameričke pacijente s rakom pluća i kontrolne skupine. Približno 95% pouzdanosti za AUC izračunato je korištenjem STATA statističkog softvera. Također su usporedili AUC afroamerički model s objavljenim i interno potvrđenim Spitz modelom (L. Etzel i sur., 2008.).

## 4.2 Podaci

### 4.2.1 Razvoj i validacija modela predviđanja rizika od raka pluća za Afroamerikance

Cilj ove analize bio je razviti i potvrditi model predviđanja rizika od raka pluća koji je specifičan za Afroamerikance.

U ovom modelu istraživanja analiziraju se podaci od 491 afroamerikanca sa dijagnoziranim rakom pluća i 497 odgovarajućih afroameričkih kontrola kako bi se identificirali specifični rizici i samim time odredili rizik od raka pluća. U navedenom modelu provedeni su interne i vanjske provjere modela rizika. Tu su se koristili podaci o dodatnim slučajevima (L. Etzel i sur., 2008.).

Neke od varijabli odnosno atributa koji su se koristili prilikom ovog istraživanja jesu (L. Etzel i sur., 2008.):

- Spol
- Godište
- Status o pušenju
- Trenutačni broj pušača
- Broj odviknutih pušača
- Izloženost okolišu
- Komorbiditet
- Povijesna bolest obitelji
- Razina stadija raka

Važno je napomenuti kako su u ovom modelu, osobe koje su definirane kao nepušači, osobe koje su popušile manje od 100 cigareta. Za bivše pušače se smatraju oni koji su prestali više od godinu dana, a sadašnju pušači uključuju nedavne prestanke pušenja koji su prestali pušiti u posljednjih 12 mjeseci (L. Etzel i sur., 2008.).

### 4.2.2 Predviđanje karcinoma pluća prije kirurške resekcije u bolesnika s plućnim čvorovima

U ovom istraživanju korišteno je 18 atributa odnosno varijabli. Model je razvijen korištenjem unaprijed određenog skupa varijabli izvedenih iz prethodno objavljenih i validiranih modela za određivanje odgovarajuće populacije za probir niske doze CT skeniranja ili za procjenu

vjerovatnosti da li je plućni nodul maligni nakon svog otkrića. Također u modelu je prikazani i rast lezija . Sve ostale varijable odabrane na savjet kirurga s obzirom na ostale čimbenike koji se obično susreću u njihovoj praksi (A. Deppen i sur., 2014.).

Cilj ovog modela je utvrditi procjenu rizika od raka pluća među zdravom populacijom. Razlog tome je kako bi se saznalo tko bi točno trebao obaviti pregled te ustanoviti da li ima rak pluća.

Ključne riječi korištene u ovom istraživanju jesu „Rak pluća“, „Dijagnoza“ i „Modeli predviđanja“.

### 4.2.3 Model predviđanja rizika od raka pluća

Rak pluća je odgovoran za više od 1,5 milijuna smrtnih slučajeva tokom svake godine. Kako je svima poznato, stopa preživljavanja raka pluća općenito je iznimno niska, ali isto tako i varira ovisno o stadiju bolesti odnosno kada je dijagnoza uspostavljena. U Ujedinjenom Kraljevstvu, jednogodišnja relativna stopa preživljavanja iznosi približno 15% za bolesnika sa stadijem četvrtim bolesti, ali približno 70% za pacijente s prvim stadijem bolesti (C. Muller i sur., 2017.).

Ovaj model istraživanja uključuje i funkciju pluća s time da koristi podatke iz prospektivne skupine UK Biobank.  U ovoj analizi je obuhvaćeno 502.321 sudionika koji nisu ranije bili dijagnozirani s rakom pluća. Razmatran dob je uglavnom između 50 do 70 godina. Pretežno se koriste atributi koji su uglavnom povezani kao indikatori s rakom pluća. Dakle spol, varijable vezane uz povijest pušenja i ovisnosti o nikotinu, povijest bolestim obiteljsku povijest raka pluća i druge (C. Muller i sur., 2017.).

Čimbenici su ograničeni na one koji su lako dostupni i/ili mogu biti lako dijagnozirani od strane liječnika. Starost je korištena ko vremenska skala u iznosu od 40 – 70 godina. Svi sudionici za koje nisu bili utvrđeni navedeni atributi su isključeni iz istraživanja kako bi analiza dala što točnije rezultate (C. Muller i sur., 2017.).

### 4.2.4 Individualizirani model predviđanja rizika za rak pluća u korejskih muškaraca

Rak pluća vodeći je uzrok smrti od raka u Koreji. Cilj ove studije bio je razviti individualizirani model predviđanja rizika za rak pluća u korejskih muškaraca koristeći kohortne podatke temeljene na populaciji (Park i sur., 2013.).

Ova analiza je provedena nad podacima koji se dotiču 1 324 804 korejskih muškaraca bez raka na početku.

Razmatrana dob za ovo istraživanje jesu svi muškarci u dobi od 30 – 80 godina koji su bili podvrgnuti zdravstvenim pregledima između 1996. i 1997. godine. Nakon zdravenih pregleda,

učesnici su morali ispuniti anketu. Anketa je sadržavala pitanja o navikama pušenja, konzumiranju alkohola, fizičkim aktivnostima, prethodnoj povijesti bolesti , prehrani itd. Također, poslije tih zdravstvenih pregleda mjereni su visina, težina i krvni tlak. Također je ustanovljen status o pušenju prema uvjetima navedenima ranije. Također, velika je pozornost obraćena na tjelesnu aktivnost. Ona je procijenjena na temelju intenziteta i trajanja te je klasificirana na tri skupine (Park i sur., 2013.):

- niska, ≤4 puta tjedno na < 30 minuta po sesiji
- umjereno, 2-4 puta tjedno na ≥30 minuta po sesiji ili ≥5 puta tjedno na <30 minuta po sesiji
- visoka, ≥5 puta tjedno na ≥ 30 minuta po sesiji

Slučajevi incidencije raka među sudionicima studije identificirani su putem baze podataka Središnjeg registra raka Koreje. Korejski središnji registar raka kombinacija je bolničkog i populacijskog sustava registra raka koji pokriva više od 95% svih novodijagnosticiranih slučajeva raka u Koreji (Park i sur., 2013.).

### 4.2.5. Model rizika za predviđanje raka pluća

Na ovom istraživanju testirano je 1851 bolesnik s rakom pluća. Podijeljeni su nasumično u zasebne skupine za obuku (75% podataka) i validaciju (25% podataka) za nikad, bivše i sadašnje pušače. Apsolutni jednogodišnji rizici od raka pluća izračunati su pomoću nacionalnih podataka o incidenciji i smrtnosti. Redni indeks rizika konstruiran je za svaku kategoriju pušačkog statusa zbrajanjem omjera izgleda iz multivarijabilnih regresijskih analiza za svaki faktor rizika (R. Spitz i sur., 2007.).

Svi pacijenti koji su testirani su novodijagnosticirani pacijenti s histološki potvrđenim rakom pluća i upisani su prije početka kemoterapije ili terapije zračenjem. Na ovom modelu se ne obazire pozornost na dob, spol, etničku pripadnost ili stadij bolesti. Svi ispitanici su trebali ispuniti upitnik. Podaci prikupljeni tijekom upitnika uključuju demografske karakteristike, povijest pušenja, zanimanje, informacije o specifičnim izloženostima na poslu ili iz hobija, medicinsku povijest i obiteljsku povijest raka kod srodnika prvog stupnja [3]. Pojedinac koji nikada nije pušio ili je popušio manje od 100 cigareta u svom životu definira se kao nepušač. Osoba koja je popušila najmanje 100 cigareta tijekom svog života, ali je prestala pušiti više od 12 mjeseci prije dijagnoze raka pluća (za bolesnike) ili prije intervjua (za kontrolne subjekte) smatra se bivšim pušačima. Trenutni pušači uključuju one koji trenutno puše i „nedavne osobe koje su prestale pušenje“, tj. one koji su prestali pušiti manje od 12 mjeseci prije dijagnoze (za bolesnike) ili intervjua (za kontrolne subjekte) [3]. Podaci o povijesti pušenja uključuju trajanje pušenja, broj popušenih cigareta po danu, izračunato popušenu kutiju-godina i dob u početku

pušenja (za sve pušače) plus dob prestanka pušenja i izračunate godine od prestanka pušenja (za bivše pušače). Izloženost pasivnom pušenju (duvanski dim iz okoliša ili ETS) utvrđuje se za nikad i bivše pušače i definira se kao redovito izloženost tuđem dimu cigareta kod kuće ili na poslu, tj. dnevno ili tjedno, npr. kao i na godine izloženosti ETS-u (R. Spitz i sur., 2007.).

Isto tako, bitno je napomenuti kako su svi sudionici istraživanja stanovnici Teksasa.

## 4.3 Rezultati

### 4.3.1 Predviđanje karcinoma pluća prije kirurške resekcije u bolesnika s plućnim čvorovima

Rezultatima je ustanovljeno kako je u skupini VUMC (N = 492) prevalencija raka pluća bila 72% dok je u skupini VA bila (N = 226). Patološka dijagnoza nakon resekcije postavljena je u 451 (92%), a aktivnim nadzorom u 41 (8%), a aktivnim nadzorom u 41 (8%) osobe u VUMC kohorti. Sve dijagnoze postavljene u VA kohorti su utvrđene patološki (A. Deppen i sur., 2014.).

Osobe koje su bile dijagnozirane s rakom, vjerovatnije je da su imali potpune podatke, točnije 58% točnošću. Oni s benignom bolešću su imali 42% točnost podataka. Možemo primjetiti, da su podaci nepotpuni kod onih osoba koje su skenirane FDG-PET skeniranjem odnosno 22% u VUMC i 21% u VA), rastom na serijskim CT pretragama odnosno 13$ u VUMC i 1% u VA te simptomima bolesti prije operacije odnosno 7% u VUMC i 16% u VA. Preostale varijable od interesa imale su manje od 5% podataka koji nedostaju u skupu podataka za razvoj VUMC-a (A. Deppen i sur., 2014.).

Zanimljivo je kako je 25 kohorta VA bila sačinjena gotovo cijela od muške populacije, točnije 97%. Imala je i veću prevalenciju prije operativnih simptoma odnosno 62%. Ustanovljeno je da je i veća vjerojatnost da su osobe pušači odnosno čak 95% i to da su pušili preko 50 godina (A. Deppen i sur., 2014.).

### 4.3.2 Model rizika za predviđanje raka pluća

Što se tiče rizičnog modela za predviđanje raka pluća, dobiven je zaključak da sve varijable koje su imale statistički značajnu povezanost s rakom pluća (poput dima duhana iz okoliša, obiteljske povijesti raka, izloženosti prašini, prethodnih respiratornih bolesti) imaju jaku biološki vjerojatnu etiološku ulogu u bolesti. Statistike podudarnosti u skupovima validacije za modele nepušača, bivših i sadašnjih pušača bile su 0.57, 0.63 i 0.58 respektivno. Izračunati jednogodišnji apsolutni rizik od raka pluća za hipotetskog muškog trenutnog pušača s procijenjenim relativnim rizikom blizu 9 iznosi 8.68%. Redni indeks rizika se pokazao dobar u

stvarno pozitivnim stopama označenih kategorija visokog rizika te je iznosio 69% za sadašnje odnosno 70% za bivše pušače (R. Spitz i sur., 2007.).

### 4.3.3 Individualizirani model predviđanja rizika za rak pluća u korejskih muškaraca

Kao što je već opisano ranije u projektu, model predviđanja rizika za rak pluća u korejskih muškaraca najviše je uključivao izloženost pušenju i dob u kojoj su počeli pušiti. Također, pozornost je obrađena na tjelesnu masu i tjelesnu aktivnost (Park i sur., 2013.).

Ovaj model je pokazao izvrsne performanse. Dobiveno je sljedeće (Park i sur., 2013.) :

- C statistika = 0.871 (95%)
- CI = 0.867 – 0.876

Isto tako je ustanovljeno kako je pušenje bilo značajno povezano s rizikom od raka pluća kod korejskih muškaraca. Rizik je povećan čak četverostruko kod sadašnjih pušača koji konzumiraju više od jedne kutije dnevno u odnosu na nepušače. Važno je napomenuti kako je dob u početku pušenja bila značajan prediktor za razvoj raka pluća, pogotovo ukoliko su osobe počele pušiti u mlađoj dobi (Park i sur., 2013.).

### 4.3.4 Model predviđanja rizika od raka pluća koji uključuje pluća

S druge strane imamo model predviđanja od raka pluća koji uključuje pluća. Njegov zaključak navodi kako je tijekom akumuliranog praćenja od 1469518 ljudskih godina bilo ukupno 738 dijagnoza raka pluća. Model koji uključuje sve prediktore je imao izvrsnu diskriminaciju (konkordancija c-statistike [95% CI] = 0.85 [0.82 - 0.87]). Interna provjera sugerira da će model dobro diskriminirati kada je primijenjen na nove podatke (c-statistika ispravljena optimizmom = 0.84). Cijeli je model, uključujući FEV1, također imao skromno superiorniju diskriminaciju od one koja je osmišljena isključivo na temelju upitnih varijabli (c-statistika = 0.84 [0.82 - 0.86], c-statistika ispravljena optimizmom = 0.83, $p_{FEV1} = 3.4 \times 10^{-13}$). Potpuni model je imao bolju diskriminaciju od standardnog probira raka pluća (c-statistika = 0.66 [0.64 - 0.69]) (C. Muller i sur., 2017.).

## 4.4 Diskusija

### 4.4.1 Predviđanje karcinoma pluća prije kirurške resekcije u bolesnika s plućnim čvorovima

Kliničari koji procjenjuju plućne čvorove suočeni su s osnovnim pitanjem ravnoteže. Za razliku od biopsije raka dojke, prostate ili debelog crijeva, čvoru pluća teško je pristupiti, a biopsija pluća ima značajne rizike povezane sa zahvatom. Pregledi ishoda nakon operacije pluća otkrili su stope smrtnosti od 1 do 3% unutar 30 dana i stope čak 7% nakon 90 dana. Jedno od predloženih rješenja je da kliničar odgodi biopsiju i liječenje dok ne bude moguća preciznija dijagnoza. Učinci odgode dijagnoze i liječenja od trenutka otkrivanja lezije do faze progresije i metastaze nisu dobro poznati. Stoga se klinička praksa obično usredotočuje na pravodobnost skrbi čak i za mali i lokalizirani karcinom (A. Deppen i sur., 2014.).

Time se predlagao novi, validirani model kliničkog predviđanja raka pluća u populaciji pacijenata s plućnim čvorovima koji se procjenjuju za kiruršku biopsiju pluća. Model “treat” raka pluća pružio je visoko i dosljedno predviđanje raka pluća na temelju uobičajenih kliničkih karakteristika i bila je bolja od modela klinike Mayo (A. Deppen i sur., 2014.).

Uporaba visoke razine diskriminacije modela “treat” može biti vrijedna u pružanju kliničkih smjernica u procjeni individualne vjerojatnosti raka pluća. Na kraju, uspostavile su se superiorne performanse modela raka pluća “treat” u usporedbi s modelom klinike Mayo i potvrdili su model “treat” u zasebnoj skupini s većom prevalencijom bolesti. Kada se ispituje klinički model predviđanja raka pluća, doktori koji procjenjuju pacijenta neposredno prije operacije nemaju validirane modele za ovu populaciju visokog rizika. Model “treat” rješava tu potrebu. Budući rad će potvrditi model “treat” u vanjskim skupovima podataka s različitom prevalencijom maligniteta kako bi se izmjerile promjene negativne prediktivne vrijednosti modela (A. Deppen i sur., 2014.).

### 4.4.2 Individualizirani model predviđanja rizika za rak pluća kod korejskih muškaraca

Ovaj model je pokazao veću sposobnost razlikovanja (86%) od većine drugih modela rizika od raka pluća koji su pokazali sposobnost razlikovanja u rasponu od 57% do 75%. Postoji nekoliko ograničenja što se tiče ove studije. Nema procjene učinka okolišnih ili radnih čimbenika rizika na rak pluća kao što su pasivno pušenje, izloženost onečišćenju zraka ili azbestu. Nadalje informacije o obiteljskoj povijesti raka pluća nisu bile dostupne u podacima koji su korišteni za izradu modela predviđanja rizika u ovom istraživanju. Treće ograničenje je što model uključuje samo muškarce. Autori očekuju da će ovo istraživanje imati važnu ulogu u primjeni strategija prevencije raka u Koreji i može pružiti daljnju referencu za drugu azijsku populaciju (Park i sur., 2013.).

### 4.4.3 Model rizika za predviđanje raka pluća

Ova studija ima nekoliko ograničenja. Alat za predviđanje temelji se na procjenama relativnog rizika koje su izvedene iz jedne velike studije u kojoj su pacijenti regrutirani iz jednog centra za tercijarni karcinom, a kontrolna skupina nije bila bazirana na populaciji. Drugo ograničenje je da su za izradu modela koristili samo podatke od bijelaca koji nisu Hispanci pa model možda nije primjenjiv na druge etničke skupine. Ostala potencijalna ograničenja uključuju pristranost opoziva i izvješćivanja, posebno prethodnih zdravstvenih stanja. Autori planiraju uključiti podatke o varijacijama gena temeljene na putevima u model kako bi potvrdili važan doprinos genetske osjetljivosti na rizik od raka pluća. Dodavanje takvih podataka će vjerojatno dodatno poboljšati osjetljivost i specifičnost modela (R. Spitz i sur., 2007.).

### 4.4.4 Razvoj i validacija modela predviđanja rizika od raka pluća za Afroamerikance

Rezultati unutarnje i vanjske validacije pokazali su da model dobro razlikuje slučajeve i kontrole osim za podskupinu slučajeva raka pluća s ranim početkom. Ovaj model je koristan za pružanje preciznijih procjena u usporedbi s postojećim modelima predviđanja rizika i dodatno naglašava potrebu za etnički specifičnim modelima predviđanja rizika od raka pluća. Međutim, veličina uzorka za validaciju je bila mala. Ostala ograničenja uključuju činjenicu da je istraživanje bilo bazirano u bolnici i da su kontrole izvučene samo iz gradskog područja Houstona u Teksasu. Autori misle kako bi rezultati bili bolji s većim skupovima podataka ili skupovima podataka u kojima sudionici nisu selektivno upisani na temelju određenih kriterija (spol, dob). Dobar alat za predviđanje rizika od raka pluća trebao bi uključivati i druge čimbenike osim pušenja i dobi, a njihovi manjinski modeli rizika uključivali su profesionalnu

izloženost, komorbiditete i tipove pušenja pa njihovi rezultati pružaju temelj za modele rizika od raka pluća za manjinske populacije (L. Etzel i sur., 2008.).

### 4.4.5 Model predviđanja rizika od raka pluća koji uključuje funkciju pluća

Ova studija ima nekoliko prednosti. Prva je studija za izravnu procjenu sposobnosti plućne funkcije, mjerene spirometrijom da predvidi rak pluća u populaciji kohorta. Uspjeli su izravno modelirati stope incidencije raka pluća i ukupne stope smrtnosti unutar jedne od najvećih skupina na svijetu. Model zahtjeva samo informacije koje su lako dostupne liječnicima opće prakse pa bi ga bilo jednostavno implementirati u širokim razmjerima u okruženju primarne zdravstvene zaštite. Potencijalno ograničenje za provedbu modela u praksi je zahtjev za ocjenom FEV1 ( količina zraka koju čovjek može izbaciti iz pluća u jednoj sekundi). Glavna slabost studije je to što autori nisu eksterno potvrdili model predviđanja rizika. Rezultati interne provjere su dali obećavajuće dokaze da će model dobro funkcionirati kada se primjenjuje na nove pojedince. U praksi učinak može varirati kada se primijeni na nove podatke, osobito u populacijama s različitim prosječnim rizicima. Autori navode da bi se model mogao unaprijediti uključivanjem genetskih i molekularnih prediktora raka pluća u model (C. Muller i sur., 2017.).

# 5. Razumijevanje podataka

Druga faza CRISP-DM standarda se sastoji od sljedećih koraka: prikupljanje podataka, opisivanje podataka, istraživanje podataka i provjere kvalitete podataka. U ovom radu se koristi gotov i javno dostupan skup podataka pa ne treba raditi korak prikupljanja podataka. U ovom poglavlju će se opisati podaci tj. od kojih atributa se sastoji skup podataka nakon čega će se istražiti podaci te na kraju će se provjeriti kvaliteta podataka.

## 5.1 Opis podataka

Ovdje će se u tablici prikazati svi atributi skupa podataka te njihovi opisi. Skup podatak se sastoji od ukupno 16 atributa od kojih 15 atributa opisuju karakteristike i simptome bolesnika, a 1 atribut nam govori da li taj bolesnik ima ili nema rak pluća. Skup podataka se sastoji od ukupno 309 zapisa. Slijedi tablica s opisom atributa.

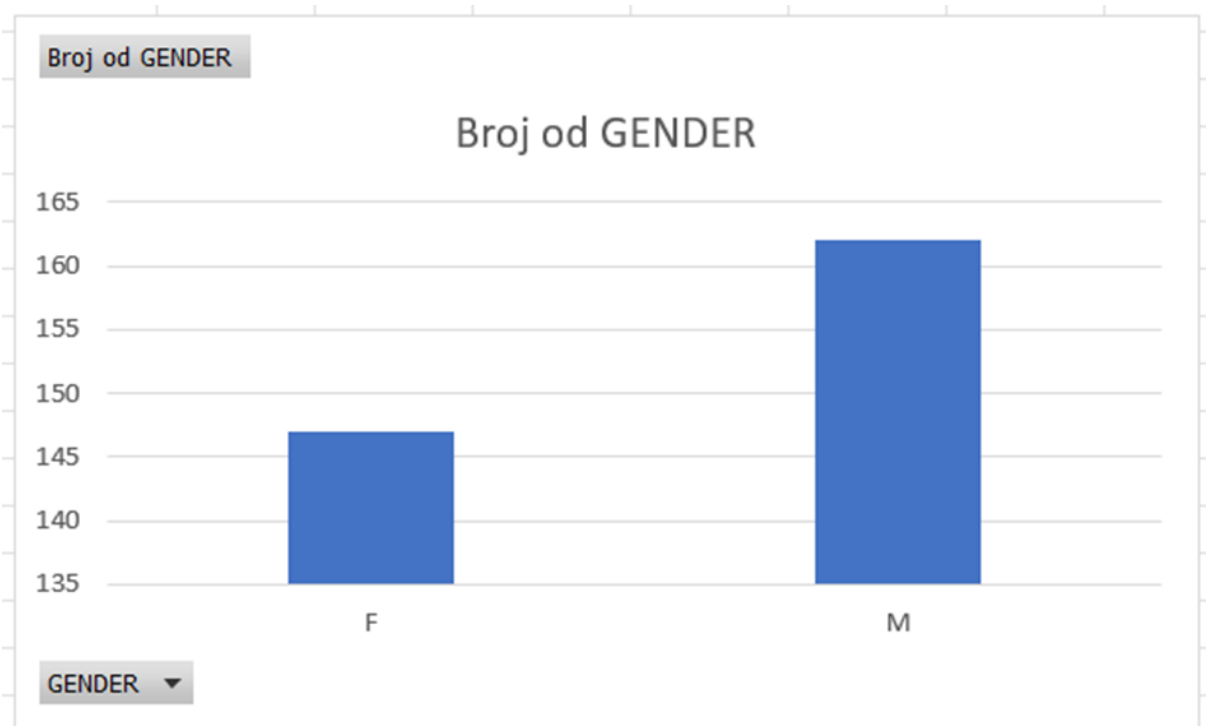
| Atribut | Opis | Vrijednosti |
|---|---|---|
| **Gender** | Spol bolesnika | M za muškarce<br>F za žene |
| **Age** | Godine bolesnika | Broj godina |
| **Smoking** | Puši li bolesnik | 2=Da 1=Ne |
| **Yellow fingers** | Ima li bolesnik žute prste | 2=Da 1=Ne |
| **Anxiety** | Da li je bolesnik anksiozan | 2=Da 1=Ne |
| **Peer pressure** | Pritisak vršnjaka | 2=Da 1=Ne |
| **Chronic disease** | Kronična bolest | 2=Da 1=Ne |

| | | |
|---|---|---|
| **Fatigue** | Umor | 2=Da 1=Ne |
| **Allergy** | Alergija | 2=Da 1=Ne |
| **Wheezing** | Teško disanje | 2=Da 1=Ne |
| **Alcohol consuming** | Konzumiranje alkohola | 2=Da 1=Ne |
| **Coughing** | Kašalj | 2=Da 1=Ne |
| **Shortness of breath** | Kratak dah | 2=Da 1=Ne |
| **Swallowing difficulty** | Otežano gutanje | 2=Da 1=Ne |
| **Chest pain** | Bol u prsima | 2=Da 1=Ne |
| **Lung cancer** | Rak pluća | 2=Da 1=Ne |

Tablica 1: Opis atributa

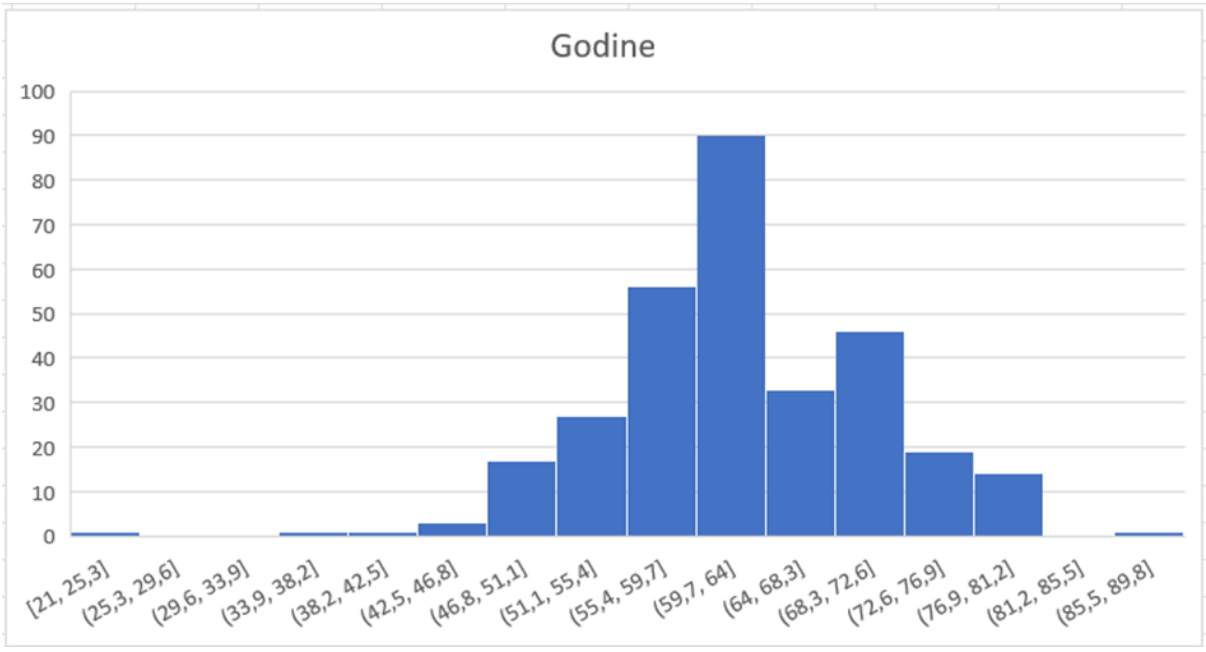
## 5.2 Istraživanje i kvaliteta podataka

Istraživanje podataka se radi nad podacima koji su prikupljeni za izradu rada tj. u ovom slučaju koji su odabrani sa stranice Kaggle. Što se tiče kvalitete podataka, ona je dobra. Podaci se sastoje od ukupno 16 atributa od kojih su 15 atributa kategorijski i samo jedan atribut je numerički.



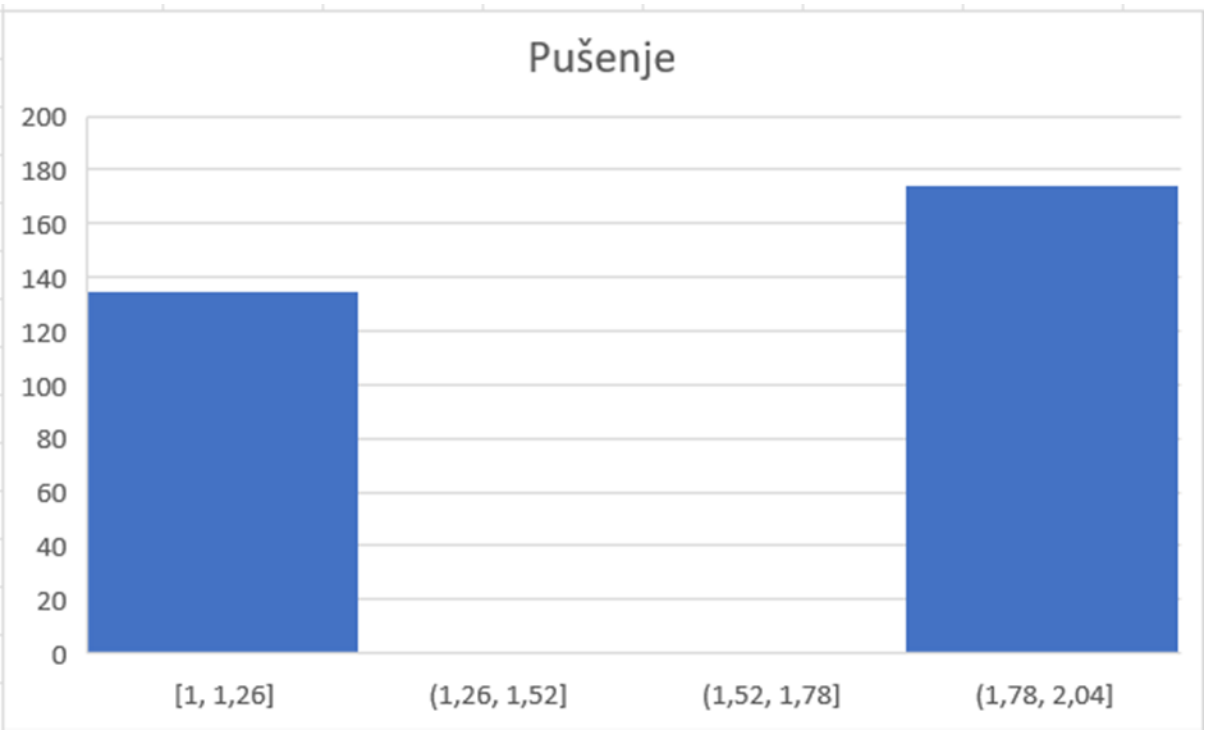

Slika 6: Odnos spolova u izvornim podacima

Na slici 6 se nalazi omjer muških i ženskih bolesnika. Može se vidjeti kako nije prevelika razlika u zastupljenosti spolova, muških ima 162 dok ženskih bolesnika ima 147. Može se zaključiti da podjednako oba spola razvijaju simptome raka pluća.




Slika 7: Broj bolesnika po godinama

Na slici 7 se nalazi histogram o godinama bolesnika. Može se vidjeti kako najviše bolesnika razvije rak pluća u dobi od 60 do 64 godine. Zatim slijede bolesnici u dobi od 55 do 60 godina.




Slika 8: Prebrojene vrijednosti atributa Smoking

Na slici 8 se nalazi histogram koji prikazuje broj bolesnika koji puše odnosno ne puše. Može se vidjeti kako 135 bolesnika ne puši dok 174 njih puši. Iz ovog histograma može se zaključiti kako se rak pluća može dobiti i u slučaju da čovjek ne puši tj. da pušenje nije jedini uzrok raka pluća.

# 6. Priprema podataka

U ovoj fazi rada potrebno je provesti nekoliko koraka koji su ključni za isti. Ti koraci jesu čišćenje, integracija i oblikovanje podataka te čišćenje. Ovo je potrebno izvesti kako bi na kraju modeliranja dobili što bolje i što točnije rezultate. Odabrani skup podataka se sastoji od 310 zapisa te 16 atributa.

Skup podataka je potrebno preuzeti sa interneta odnosno sa određenog servisa koji pruža podatke odnosno tzv. “dataset”. U ovom slučaju to je sustav “Kaggle”. Nakon preuzimanja podaci su uvezeni u Microsoft Office alat pod nazivom Excel. Excel sadrži opciju filtriranje pomoću koje je provjeriti valjanost podataka za svaki atribut. Nevaljan zapis bi predstavljala npr. praznina. U ovom slučaju sa ovim “datasetom”, excel nije pronašao niti jedan zapis koji nije valjan pa skup podataka može zadržati sve početne zapise, točnije njih 310.

# 7. Modeliranje podataka

U ovom poglavlju će biti opisani svi koraci koje je bilo potrebno provesti prilikom modeliranja. Koristit će se 3 tehnike modeliranja, a to su:

- Stablo odlučivanja
- Neuronska mreža
- Bayesova mreža

Samim time potrebno je koristiti i prikladne alate za provedbu navedenih tehnika, a to su programi BigML i Netica.

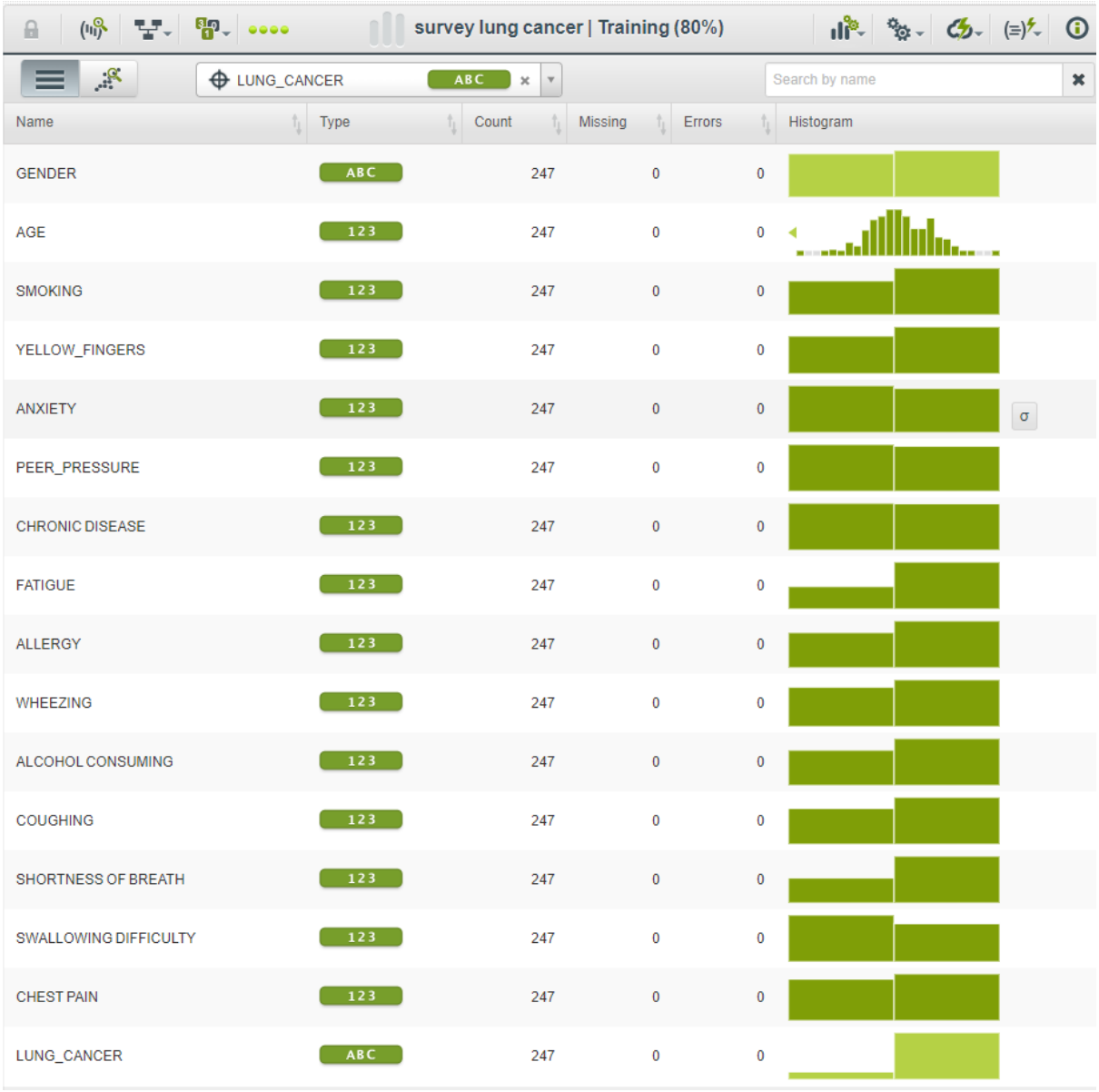
Prvi korak je tzv. Uvođenje odabranog skupa podataka u alat BigML.

| Name | Type | Instance 1 | Instance 2 | Instance 3 |
|---|---|---|---|---|
| GENDER | ABC | M | M | F |
| AGE | 123 | 69 | 74 | 59 |
| SMOKING | 123 | 1 | 2 | 1 |
| YELLOW_FINGERS | 123 | 2 | 1 | 1 |
| ANXIETY | 123 | 2 | 1 | 1 |
| PEER_PRESSURE | 123 | 1 | 1 | 2 |
| CHRONIC DISEASE | 123 | 1 | 2 | 1 |
| FATIGUE | 123 | 2 | 2 | 2 |
| ALLERGY | 123 | 1 | 2 | 1 |
| WHEEZING | 123 | 2 | 1 | 2 |
| ALCOHOL CONSUMING | 123 | 2 | 1 | 1 |
| COUGHING | 123 | 2 | 1 | 2 |
| SHORTNESS OF BREATH | 123 | 2 | 2 | 2 |
| SWALLOWING DIFFICULTY | 123 | 2 | 2 | 1 |
| CHEST PAIN | 123 | 2 | 2 | 2 |
| LUNG_CANCER | ABC | YES | YES | NO |

Slika 9: Skup podataka

## 7.1 Dizajn za potrebe testiranja

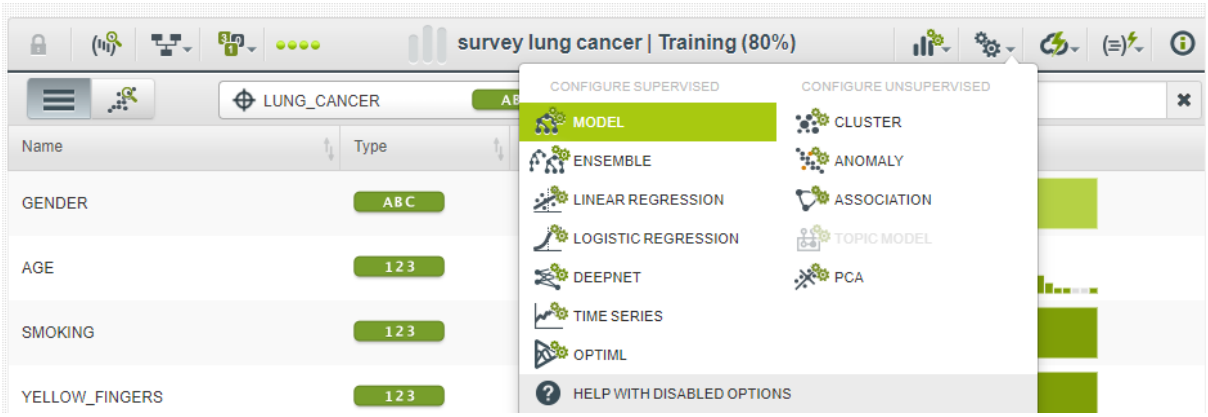
BigML sadrži jako korisnu funkciju koja je nužna za potrebe ovog poglavlja odnosno dizajn za potrebe testiranja. Funkcija se zove Random Split. Ova funkcija dijeli odabrani skup podataka na dva manja podskupa. 80% inicijalnih podataka ide na skup pod imenom “Training” dok ostalih 20% inicijalnih podataka ide u novi skup podataka pod nazivom “Testing”. Na slici ispod se nalazi skup podataka “Training”.




Slika 10: Skup podataka “Training”

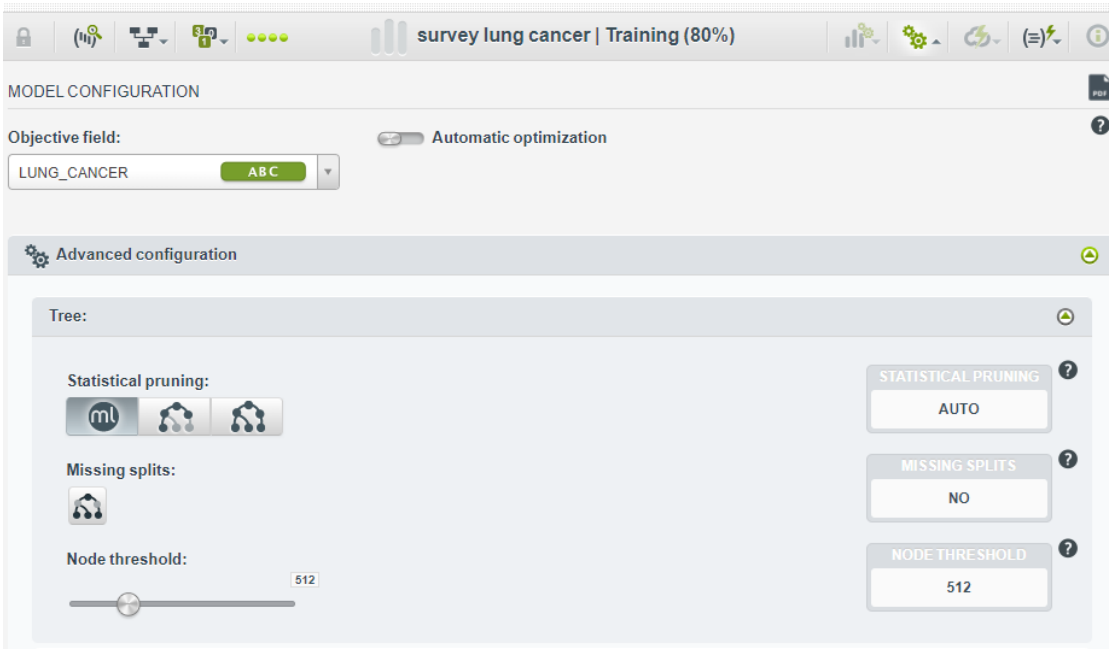
## 7.2 Izrada modela (stablo odlučivanja)

Za izradu modela potrebno je odabrati “Training” skup podataka i odabrati funkciju “MODEL”.



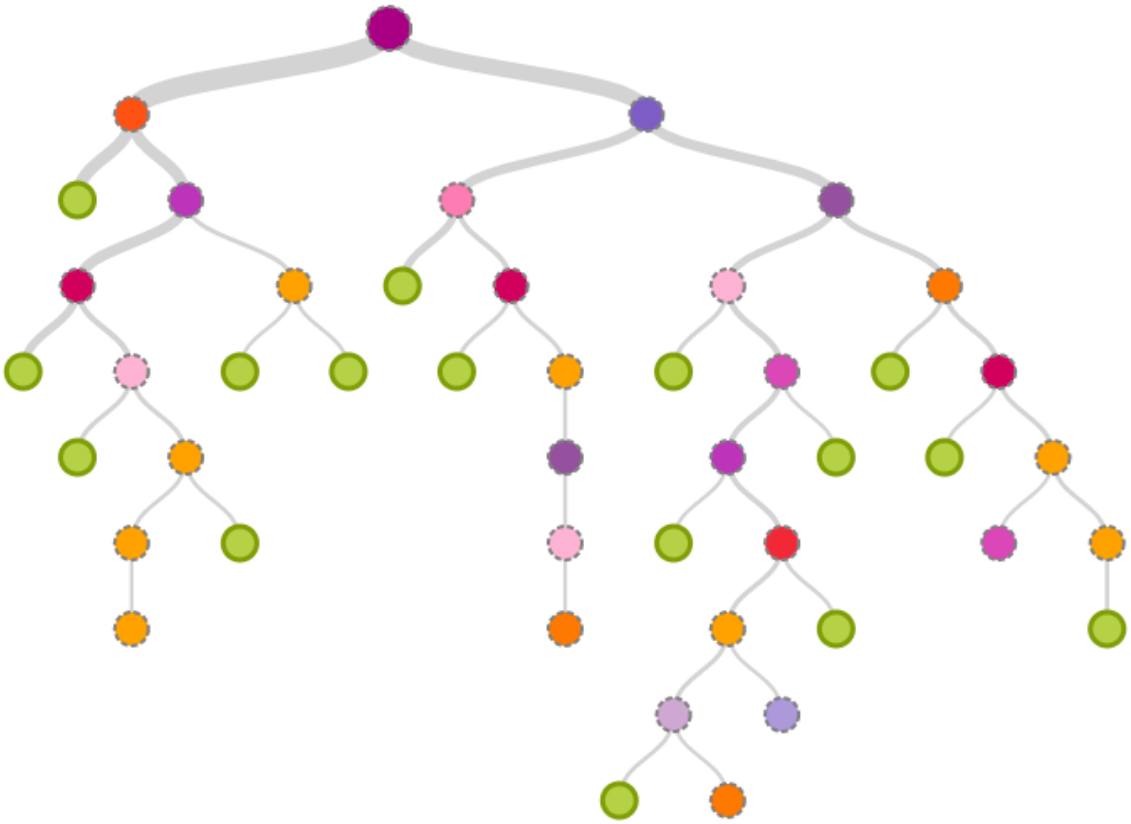

Slika 11: Skup podataka

Nakon odabira navedene funkcije iskače novi prozor sa dodatnim postavkama i konfiguracijama. Tu je moguće odabrati različite vrste obrezivanja. Odabrano je tzv. pametno obrezivanje odnosno smart pruning.



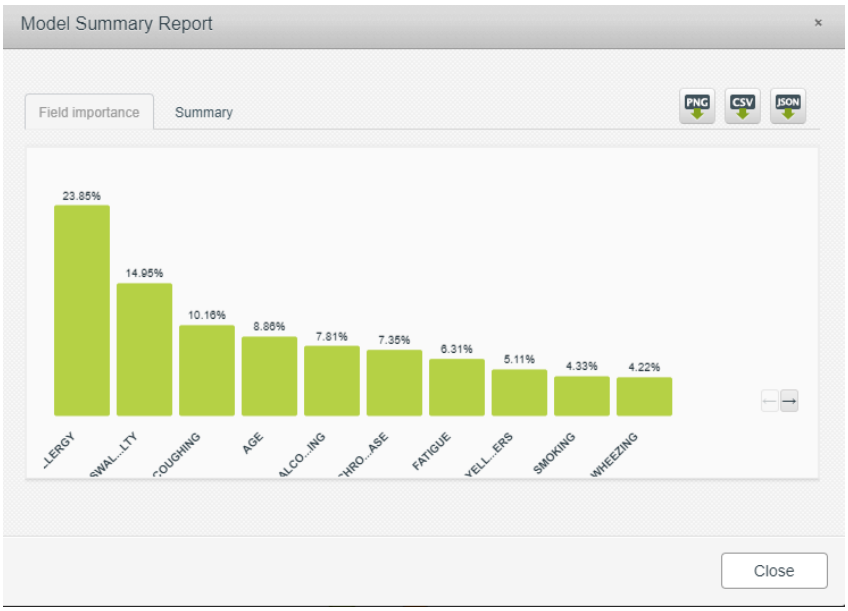

Slika 12: Konfiguracija - smart pruning

Nakon podešavanja svih postavki može se kreirati model klikom na gumb “CREATE MODEL”. Time se dobiva model u obliku stabla, čiji prikaz možemo vidjeti na slici 12.

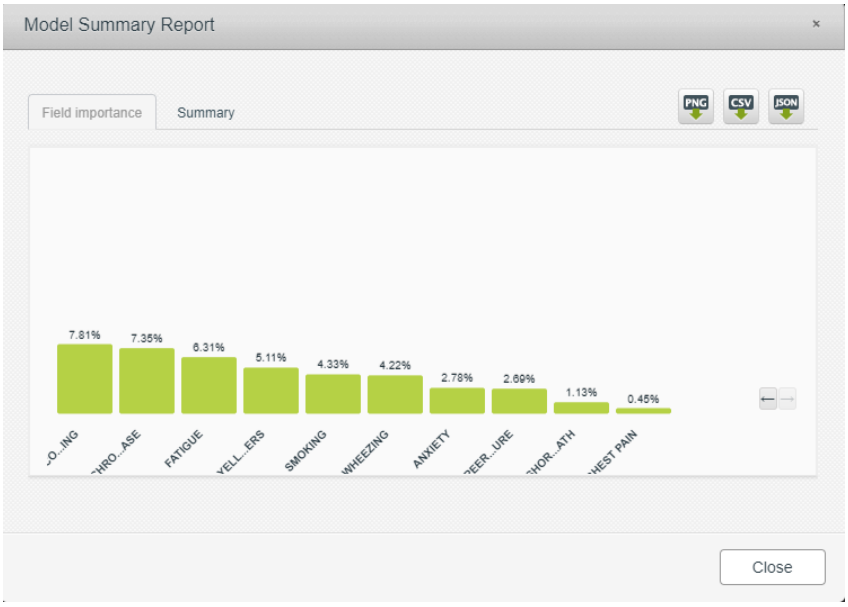


Slika 13: Stablo - model

Nakon izrade željenog modela možemo vidjeti i njegov sažetak klikom na gumb “Model summary report”. Na njemu možemo vidjeti važnosti pojedinih varijabli u obliku grafa. Na slici ispod možemo vidjeti kako najveći utjecaj ima varijabla “Allergy” zatim varijabla “Swallowing_difficulty”, onda “coughig” itd. Detaljniji prikaz možemo vidjeti na slikama 14 i 15. Također pouzdanost ovog modela je 83,65% dok vjerojatnost iznosi 88,26%.



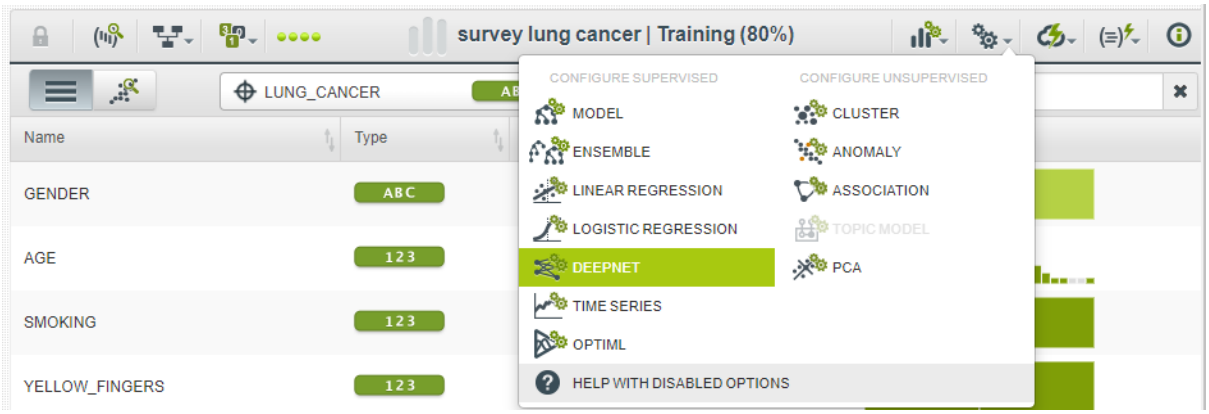

Slika 14: Model Summary Report




Slika 15: Model Summary Report

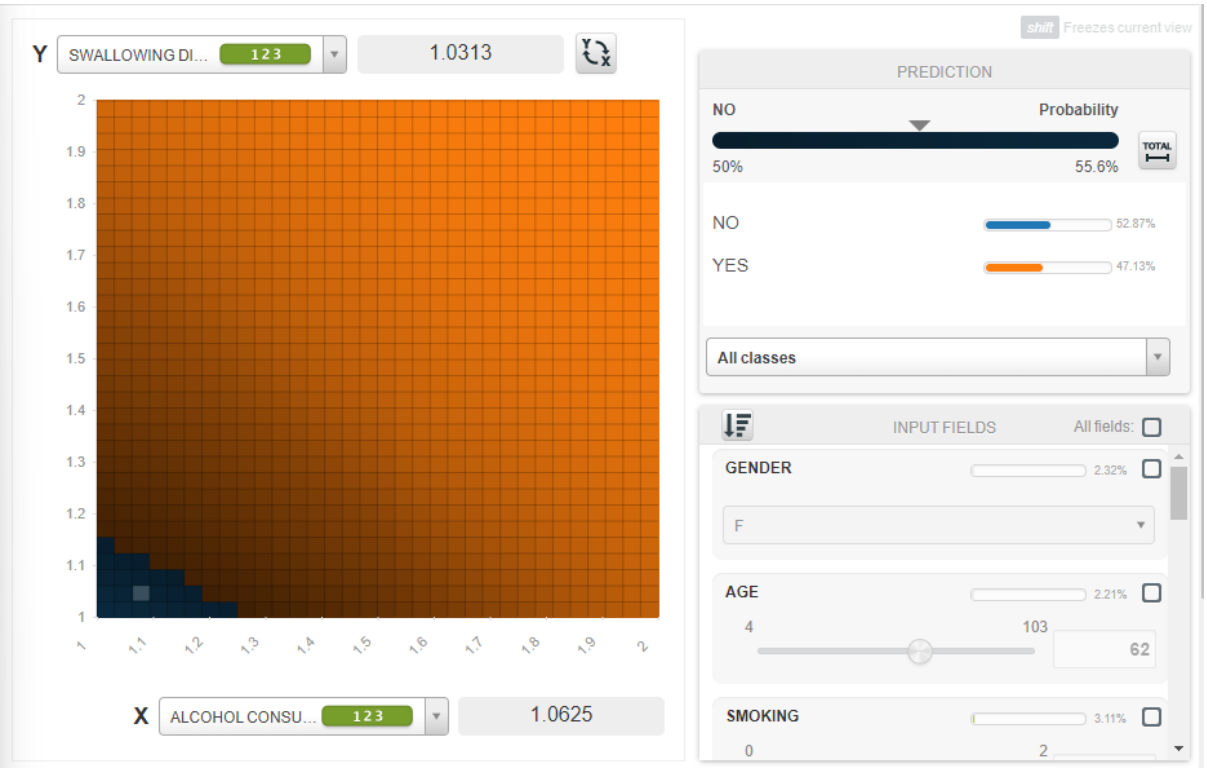
## 7.3 Izrada modela (neuronske mreže)

Kako bi izradili model neuronskih mreža potrebno je kliknuti na funkciju “DEEPNET” kod odabranog dataseta. Ovo će nam također otvoriti dodatne postavke i konfiguraciju. Bitno je napomenuti kako se i dalje koristi “Training” skup podataka.




Slika 16: Izrada neuronske mreže

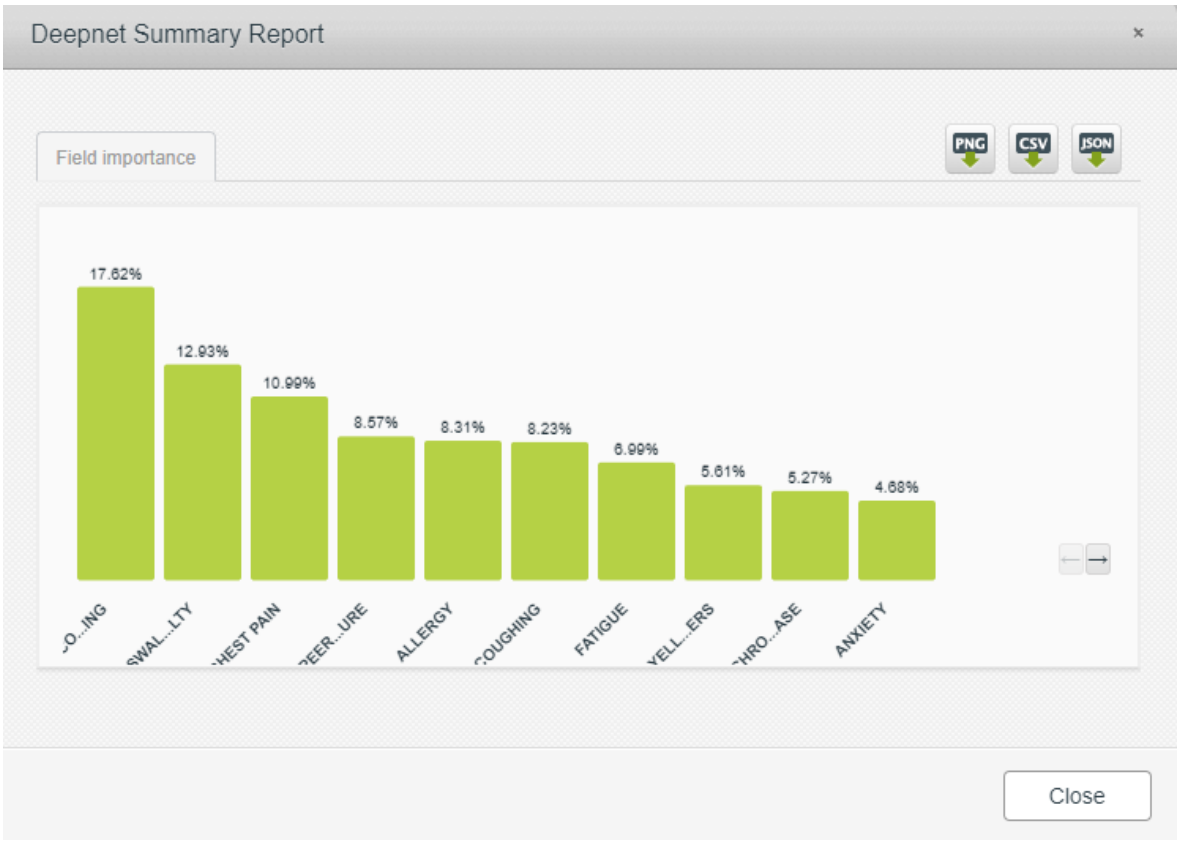
Prva stvar koju je potrebno napraviti je odabrati varijablu za koju se radi predikcija. U ovom slučaju to je varijabla “LUNG_CANCER” odnosno varijabla da li je ispitani imao rak pluća. Neuronsku mrežu kreiramo klikom na gumb “Create Deepnet”.



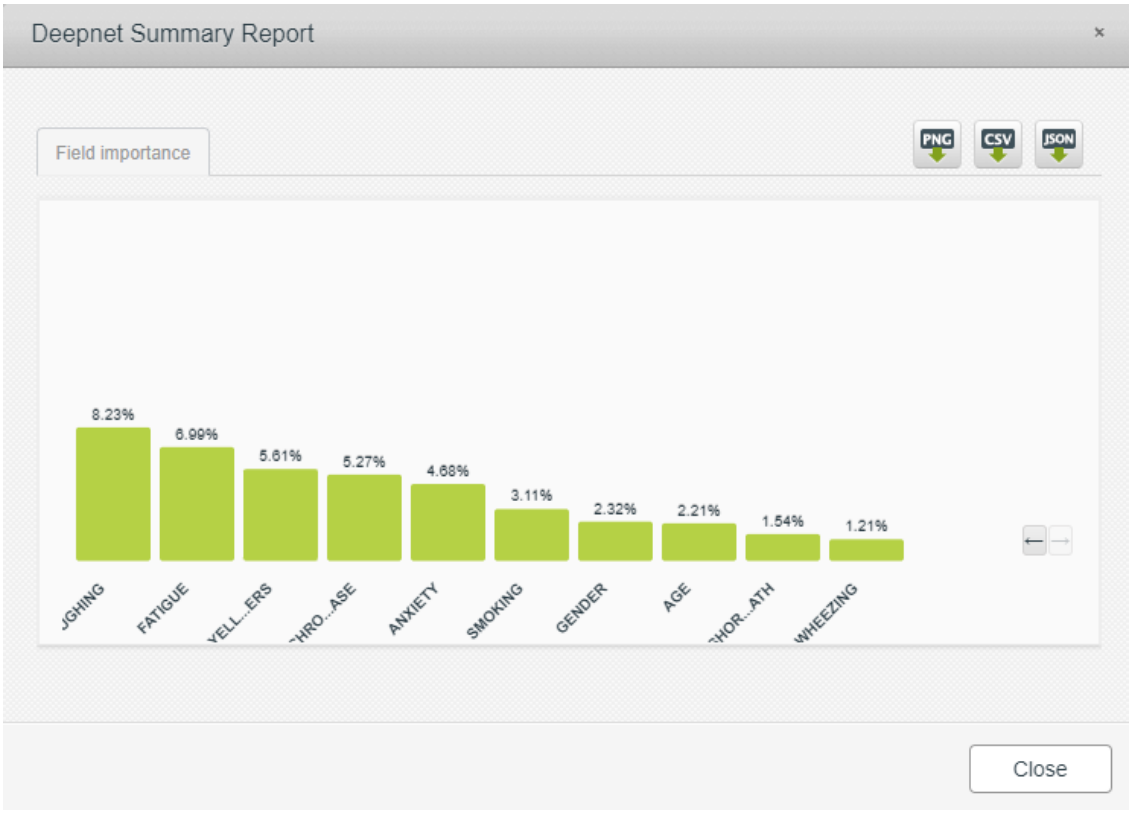

Slika 17: Neuronska mreža

Slika 17 prikazuje varijablu “SWALLOWING_DIFFICULTY” i varijablu “ALCOHOL_CONSUMING”. Narančasta boja prikazuje da je šansa za rak pluća veća od 50% dok plava radi obrnuto. Isto tako možemo zamijetiti što je boja tamnija to je i postotak za dobivanje raka pluća veći ili manji. Dakle pomoću boja možemo dosta toga zaključiti kada je u pitanju rak pluća.  Iz slike također možemo vidjeti ako je konzumacija alkohola veća od 1.3 time je veća mogućnost za rak pluća.

Klikom na gumb “Deepnet Summary Report” možemo vidjeti važnost pojedinih varijabla. Važnost pojedinih varijabla možemo vidjeti na slikama 18 i 19.



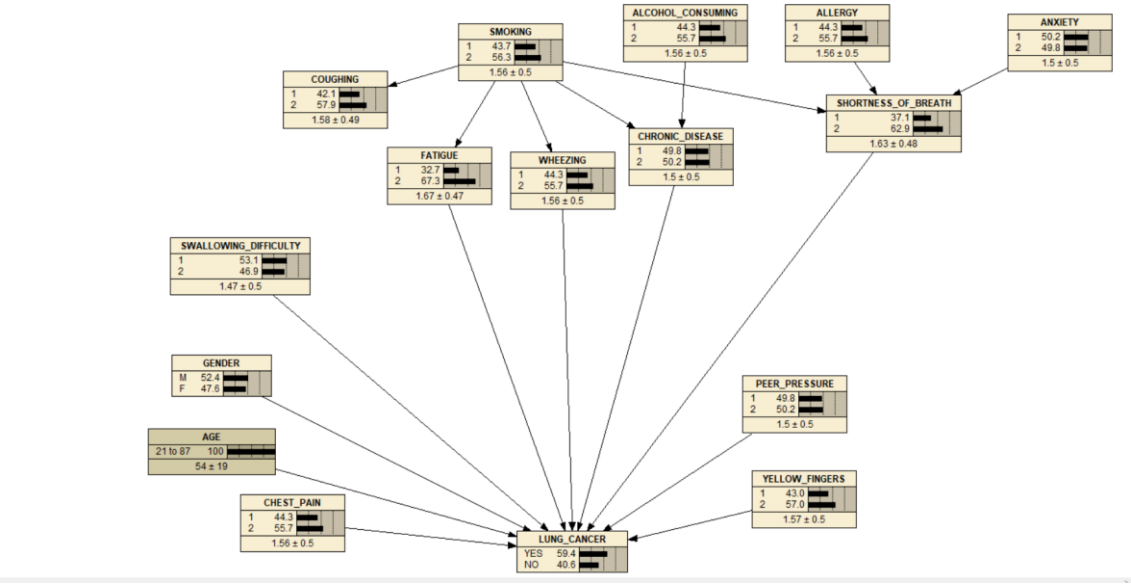

Slika 18: Izvještaj neuronske mreže




Slika 19: Izvještaj neuronske mreže

## 7.4 Izrada modela (Bayesova mreža)

Zadnji model za predviđanje je Bayesova mreža. Za njenu izradu koristit će se program Netica. Izraditi ćemo mrežu prema izvornim podacima. Mreža se sastoji od 16 čvorova, jedan čvor za svaki atribut. Nazivi čvorova moraju se podudarati s nazivima atributa u Excelu. Tu je došlo do jednog problema jer u Excelu su neki nazivi atributa bili s razmakom, a kako Netica ne podržava razmake u nazivima čvorova bilo je potrebno u Excelu umjesto razmaka staviti "_". Nakon kreiranja čvorova bilo je potrebno te čvorove spojiti u jednu mrežu. Dobivena Bayesova mreža se može vidjeti na slici 20.




Slika 20: Bayesova mreža

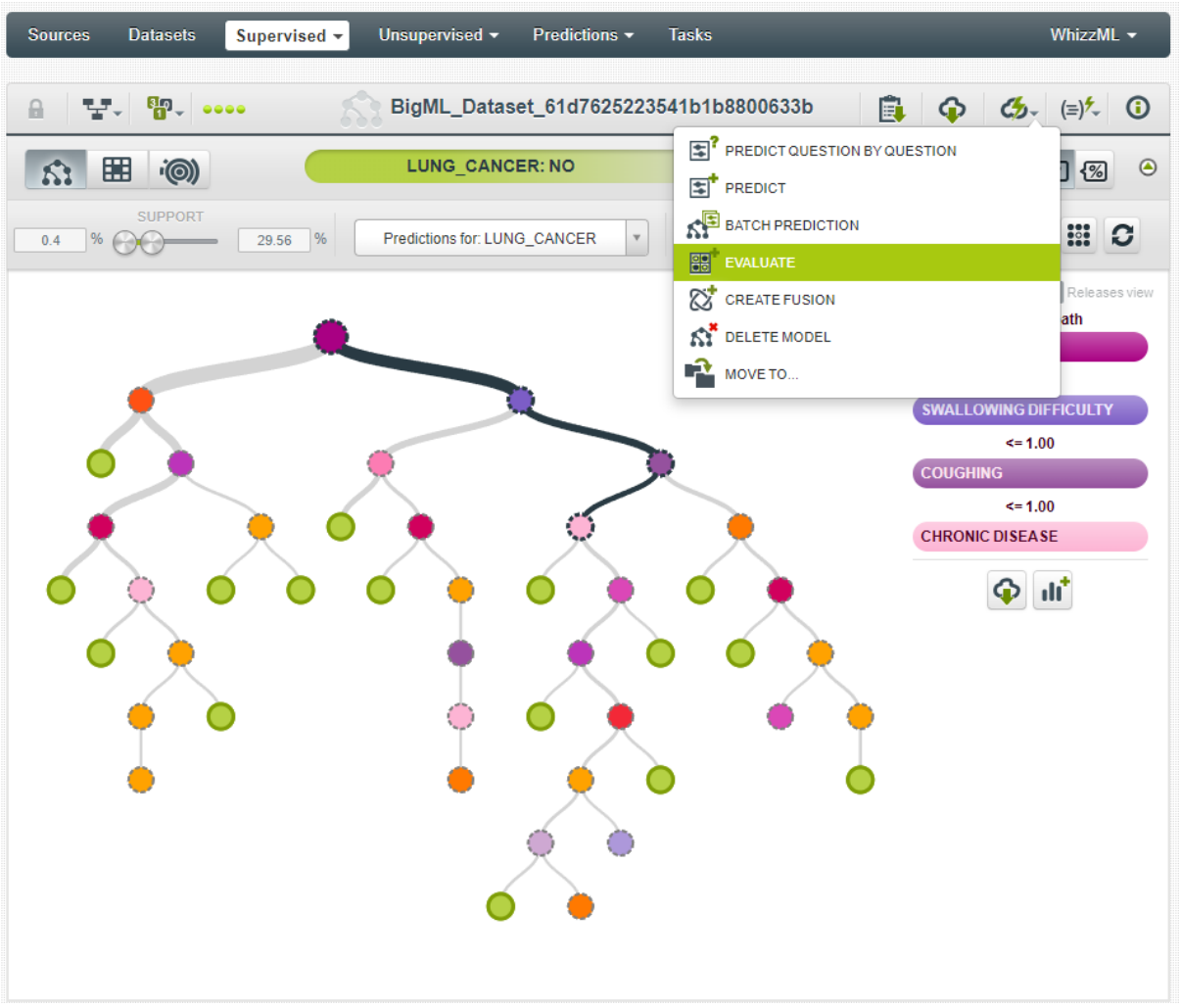
Za učenje mreže koristila se Excel datoteka koja je prethodno spremljena u formatu "Tekst(razdvojeno tabulatorom)". Nakon učitavanje datoteke upisan je stupanj od 1 000 000 što znači da će program učitati datoteku 1 000 000 puta te će se time dobiti precizniji rezultati. Nakon toga se mreža kompajlira te se dobiju rezultati. Nakon toga se mogu mijenjati vjerojatnosti pojedinih čvorova te vidjeti kako pojedini čvor utječe na izlazni čvor tj. čvor koji pokazuje vjerojatnost da bolesnik ima rak pluća.

Na slici 19 se može vidjeti kako postoje čvorovi koji utječu na druge čvorove. Tako imamo čvor "smoking" koji utječe čvorove "coughing", "fatigue", "wheezing", "chronic disease" i "shortness of breath" tj. ovisno da li bolesnik puši to utječe na simptome kašlja, umora, kratkoće daha, teškog disanja i kronične bolesti. Također čvor "alcohol consuming" utječe na čvor "chronic disease" jer pijenje alkohola utječe na razvoj kroničnih bolesti. Isto tako posljedica alergije ili anksioznosti može biti kratkoća daha.

Na temelju izvornih podataka dobiven je rezultat kako je vjerojatnost da bolesnik ima rak pluća 59.4% dok vjerojatnost da nema rak pluća iznosi 40.6%.

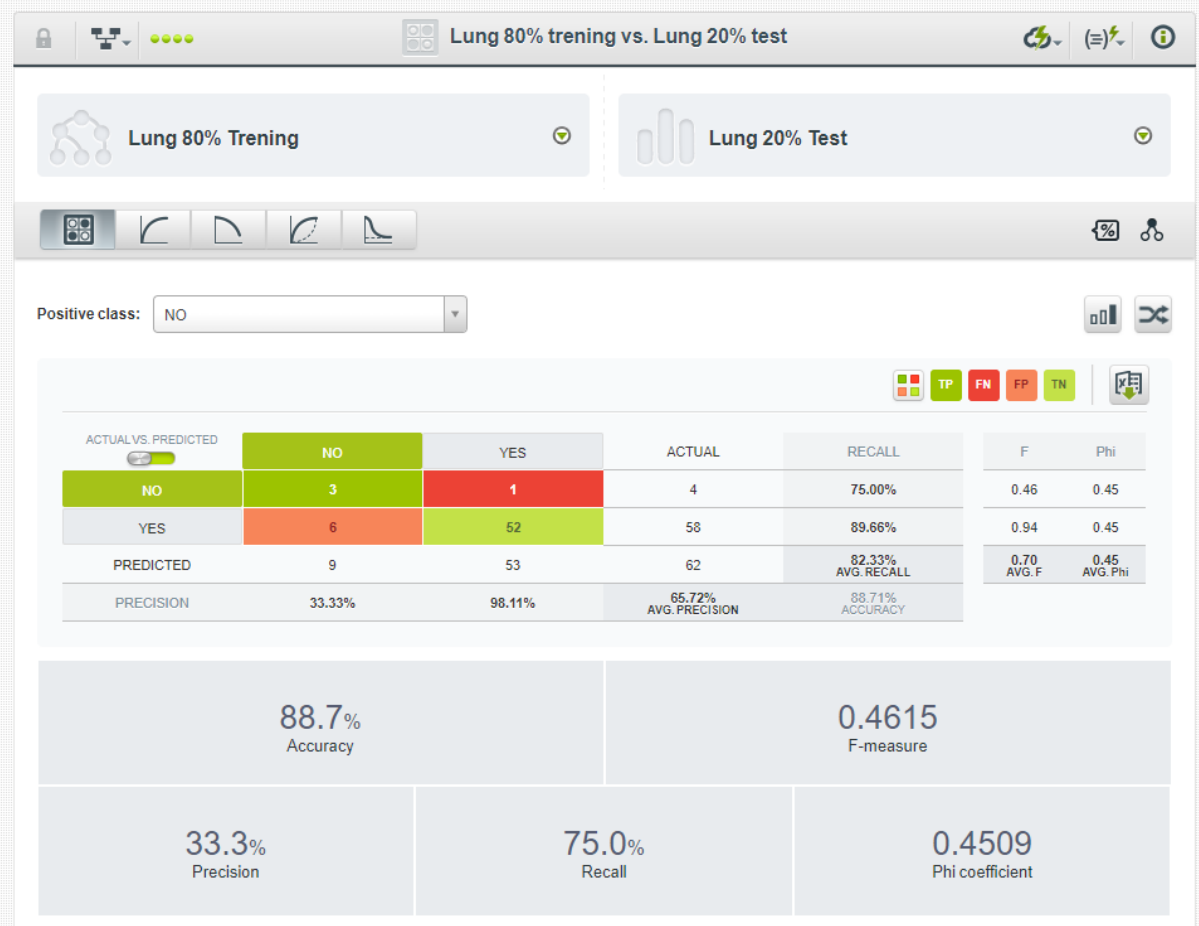
# 8. Evaluacija podataka

Nakon što otvorimo model koji želimo evaluirati unutar BigML-a iz izbornika koristimo funkciju evaluacije (*eng. Evaluate*) kako bi izvršili analizu stabla odlučivanja ili neuronske mreže. Odabiremo dakle skup podataka za treniranje i testiranje.



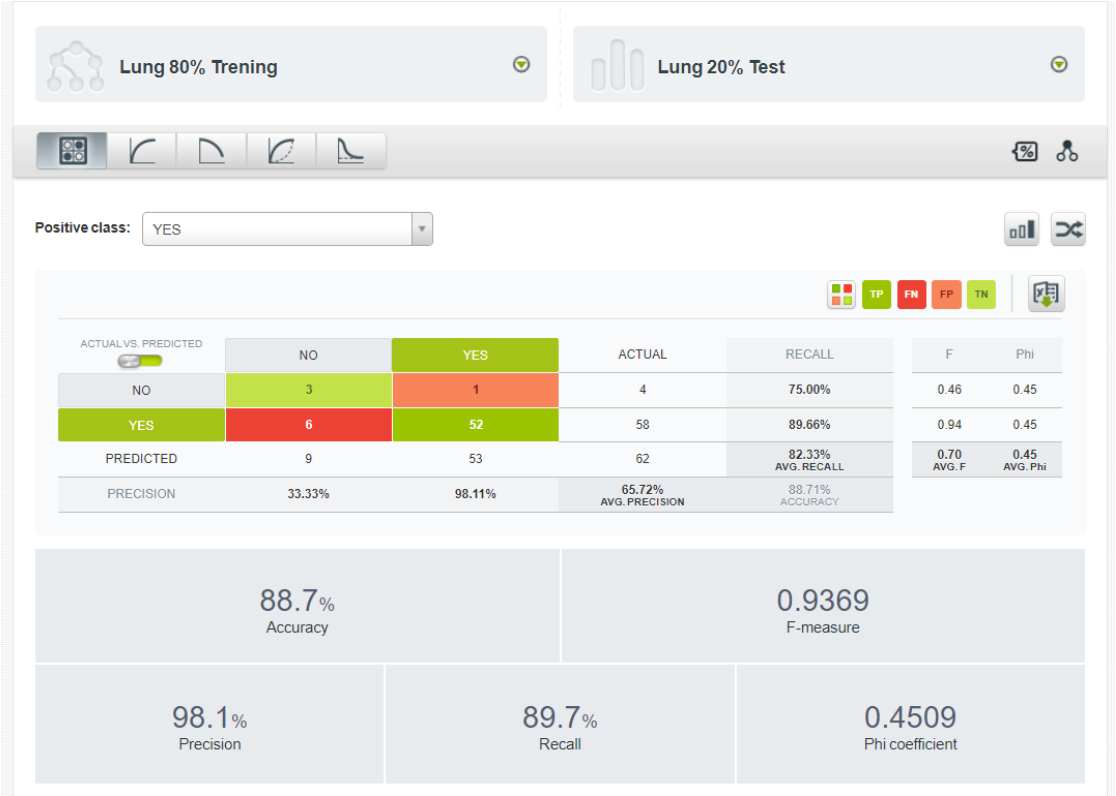

Slika 21: Uporaba evaluacije

Nakon što provedemo evaluaciju dobijemo podatke koji su prikazani na slici 22:



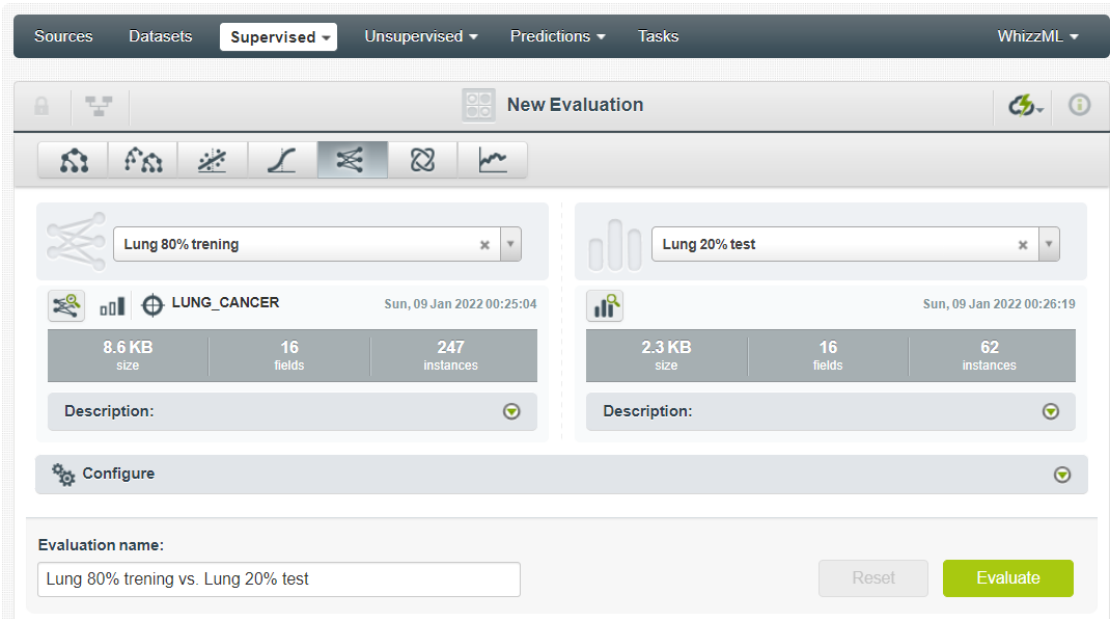

Slika 22: Rezultat evaluacije stabla

Kada postavimo “Positive class” na “NO” dobijemo da je točnost podataka vrlo visoka i ona iznosi čak 88.7% dok je preciznost podataka nešto znatno manja te iznosi 33.3% .Broj točno pozitivnih instanci iznosi 75%. Kada “Positive class” postavimo na “YES” preciznost iznositi također 88.7%, preciznost velikih 98.1% a broj pozitivnih instanci 89.7% :



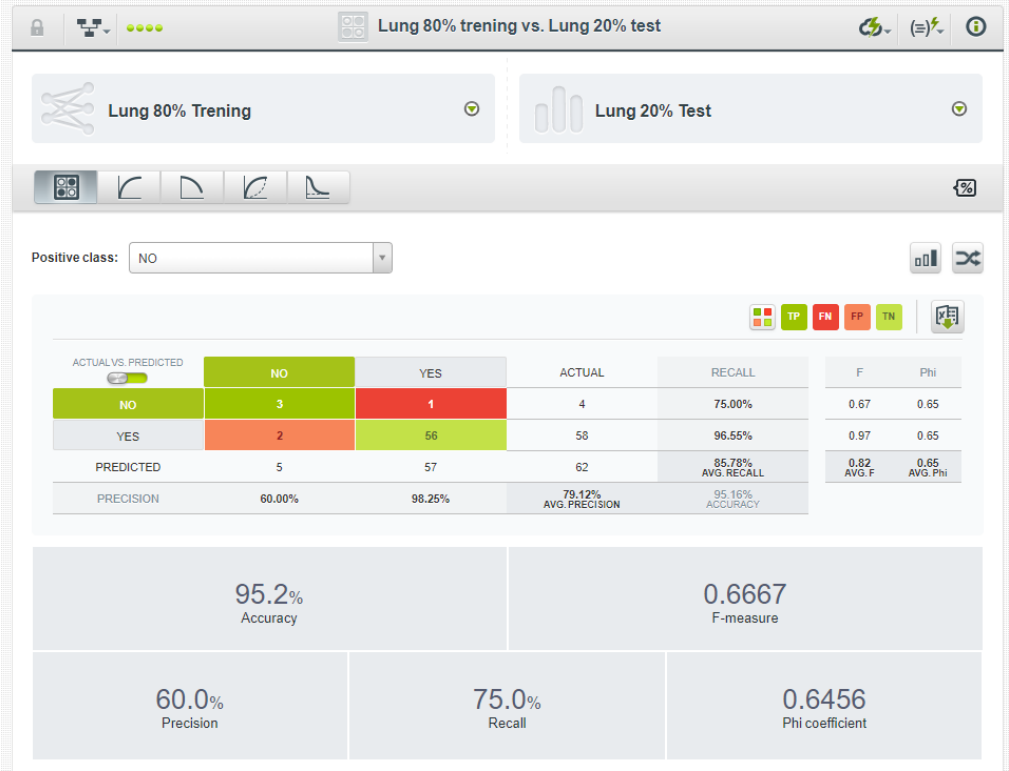

Slika 23: Rezultat evaluacije stabla

Dalje slijedi evaluacija na neuronskoj mreži gdje je postupak sličan. Prvo ćemo otići u podatke za treniranje i koristiti funkciju “Evaluate” koja se nalazi u traci s funkcijama:

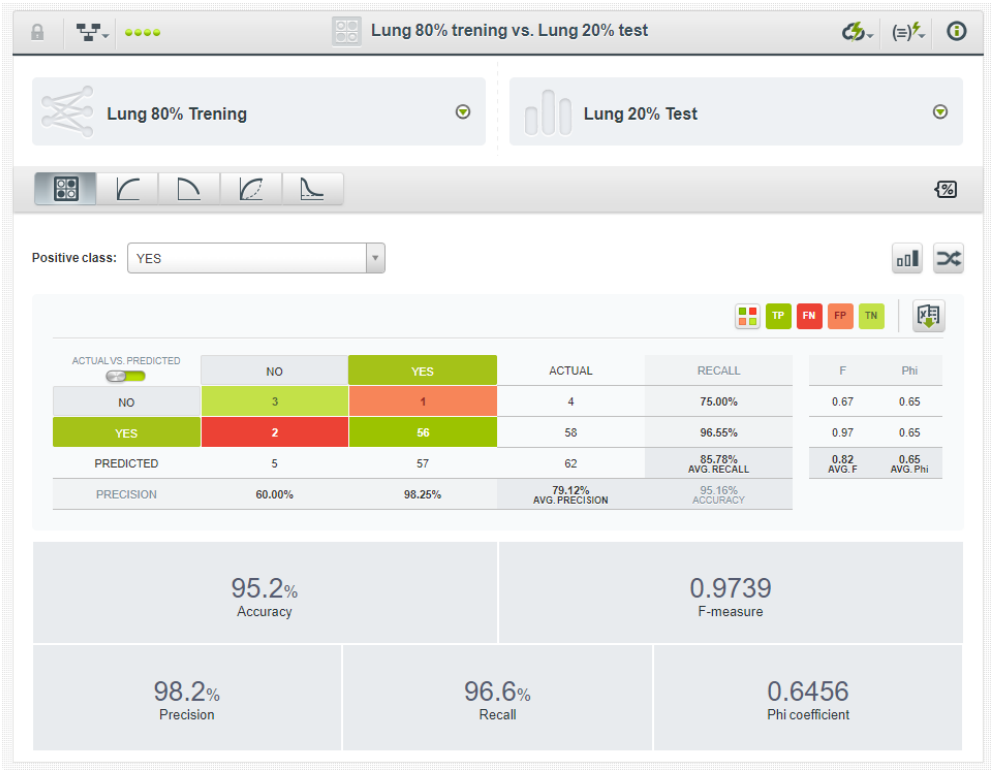



Slika 24: Evaluacija neuronske mreže

Nakon pokrenute evaluacije dobiju se sljedeći rezultati za neuronsku mrežu:




Slika 25: Rezultat evaluacije neuronske mreže




Slika 26: Rezultat evaluacije neuronske mreže

Kada postavimo “Positive class” na “NO” točnost podataka iznosi visokih 95.2%, preciznost solidnih 60% a broj točno pozitivnih instanci 75%. Kada “Positive class” postavimo

na “YES” dobijemo da je točnost podataka također 95.2% dok je preciznost podataka velikih 98.2% i broj točno pozitivnih instanci 96.6%.

### Analiza osjetljivosti Bayesove mreže

Sensitivity of 'SHORTNESS_OF_BREATH' to a finding at another node:

| Node | Variance Reduction | Percent | Mutual Info | Percent | Variance of Beliefs |
|---|---|---|---|---|---|
| SHORTNESS_OF_BREATH | 0.2333 | 100 | 0.95126 | 100 | 0.2332982 |
| ANXIETY | 0.00558 | 2.39 | 0.01734 | 1.82 | 0.0055799 |
| SMOKING | 0.002143 | 0.918 | 0.00661 | 0.695 | 0.0021426 |
| LUNG_CANCER | 0.0005943 | 0.255 | 0.00183 | 0.193 | 0.0005942 |
| ALLERGY | 0.0001142 | 0.049 | 0.00035 | 0.0372 | 0.0001143 |
| CHRONIC_DISEASE | 4.22e-05 | 0.0181 | 0.00013 | 0.0137 | 0.0000420 |
| WHEEZING | 3.597e-05 | 0.0154 | 0.00011 | 0.0117 | 0.0000359 |
| COUGHING | 3.587e-05 | 0.0154 | 0.00011 | 0.0117 | 0.0000359 |
| FATIGUE | 2.04e-06 | 0.000874 | 0.00001 | 0.000615 | 0.0000019 |
| PEER_PRESSURE | 0 | 0 | 0.00000 | 0 | 0.0000000 |
| SWALLOWING_DIFFICULTY | 0 | 0 | 0.00000 | 0 | 0.0000000 |
| GENDER | 0 | 0 | 0.00000 | 0 | 0.0000000 |
| AGE | 0 | 0 | 0.00000 | 0 | 0.0000000 |
| ALCOHOL_CONSUMING | 0 | 0 | 0.00000 | 0 | 0.0000000 |
| CHEST_PAIN | 0 | 0 | 0.00000 | 0 | 0.0000000 |
| YELLOW_FINGERS | 0 | 0 | 0.00000 | 0 | 0.0000000 |

Slika 27: analiza osjetljivosti

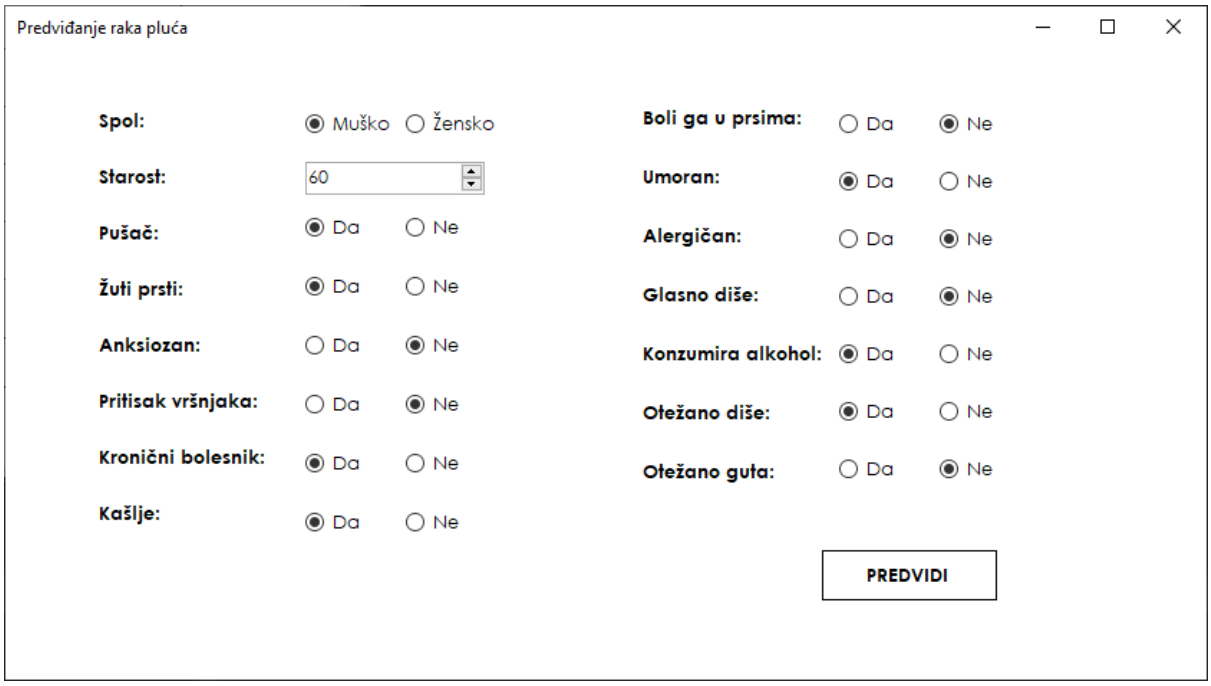
Na slici 26 se nalazi osjetljivost čvora “shortness_of_breath”. Iz ovog izvještaja se može vidjeti u kojem postotku čvor “shortness_of_breath” utječe na druge čvorove.

Ovaj čvor utječe na ukupno osam drugih čvorova i to ima utjecaj od 0.0137% na čvor “chronic_disease” tj. na postojanje kronične bolesti, 0.0117% na čvor “coughing” tj. na kašalj i na čvor “wheezing” tj. teško disanje. Nadalje utječe u postotku od 1.82% na čvor “anxiety” tj. anksioznost, 0.000615% na čvor “fatigue” odnosno umor, 0.0372% na čvor “allergy” tj. na postojanje alergije, 0.695% na čvor “smoking” tj. pušenje te na kraju ima utjecaj na izlazni čvor “lung_cancer” odnosno na postojanje raka pluća u postotku od 0.193%.

# 9. Korištenje podataka

Za potrebe korištenja podataka kreirana je aplikacija koja predviđa rak pluća ovisno o unesenim atributima u formu. Aplikacija se sastoji od jednog prozora u kojem je potrebno unijeti podatke o osobi za koju se želi predvidjeti da li ima rak pluća.

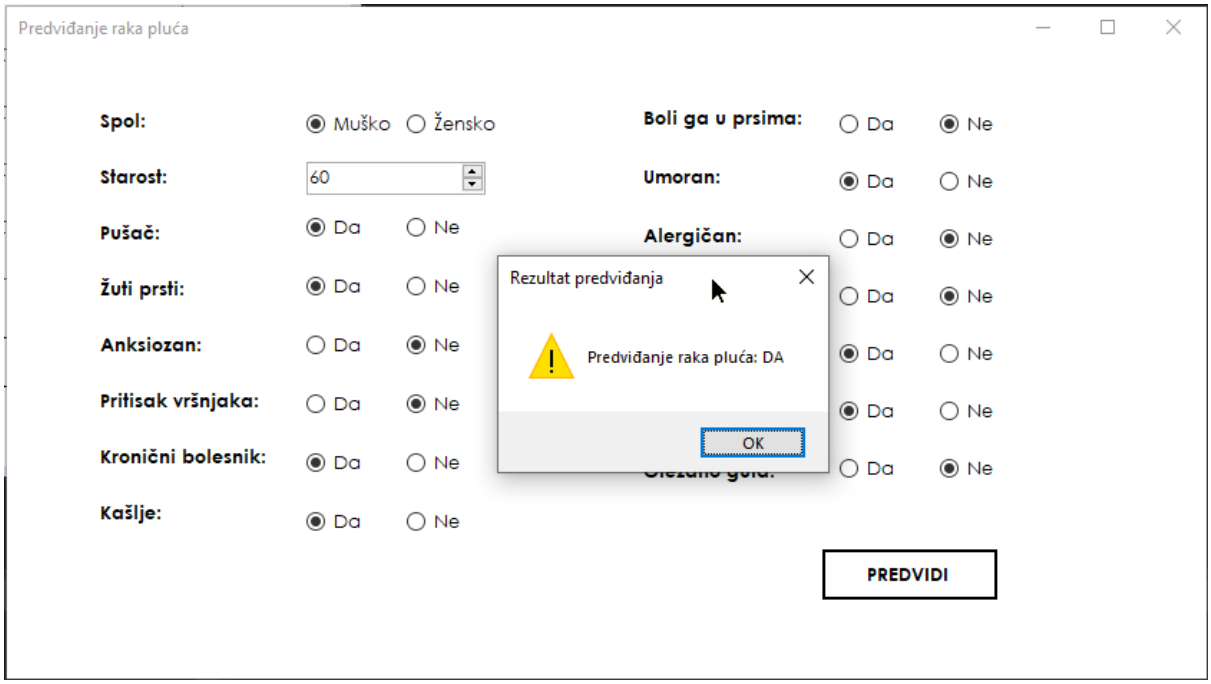
Prvo je prikazan primjer u kojem se unose podaci za osobu kojoj aplikacija predviđa da li ima rak pluća, a na slikama 30 i 31 je prikazan primjer kada aplikacija predviđa relativno zdravoj osobi da nema rak pluća.



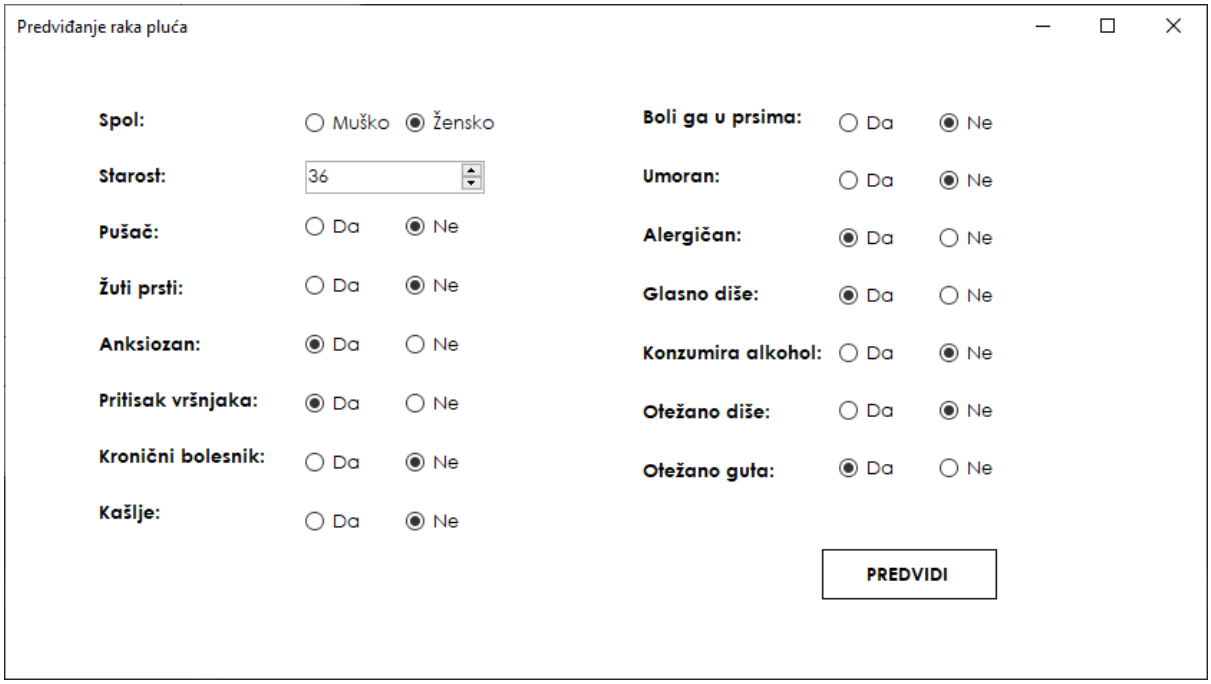

Slika 28: Unos podataka bolesne osobe u aplikaciju

Osoba muškog spola od 60 godina koja puši, ima žute prste, kronično je bolesna, kašlje, umorna je, konzumira alkohol i teško diše je reprezentativan primjer nekog tko je vrlo vjerojatan kandidat za bolesnika raka pluća. Tako na slici 28 vidimo rezultat predviđanja aplikacije da li osoba ima rak pluća ili ne.

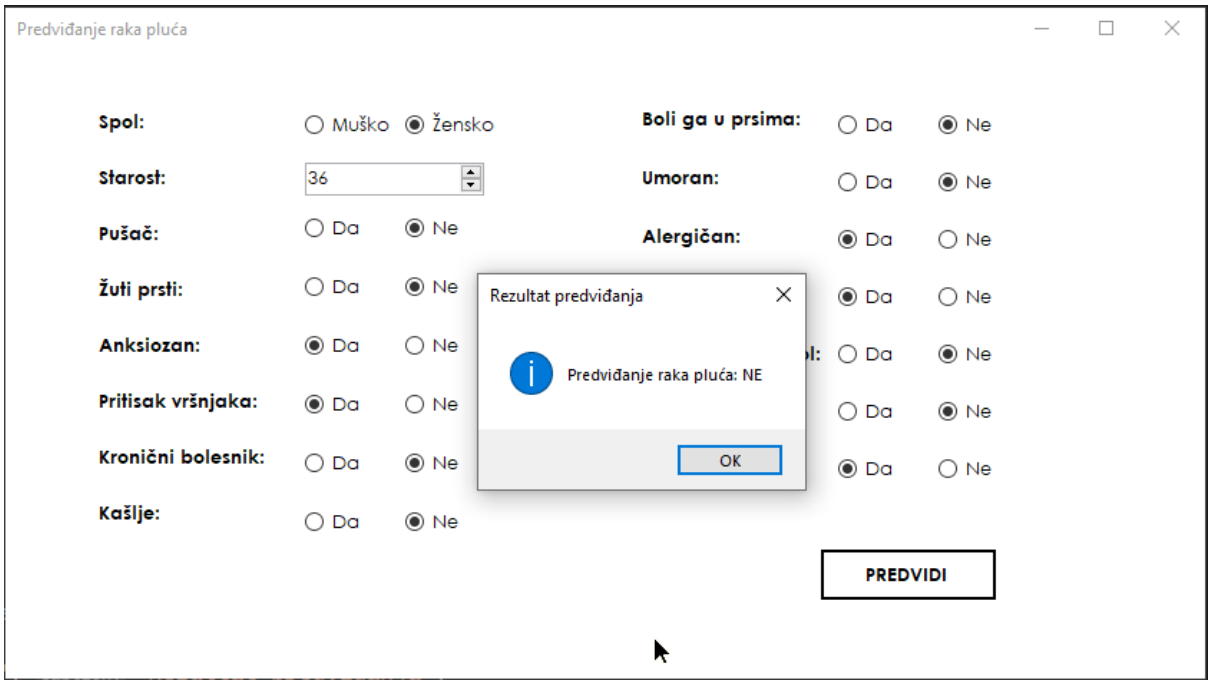
S druge strane na slici 29 se vide podaci osobe ženskog spola od 36 godina koja je anksiozna i osjeti pritisak vršnjaka, alergična je, glasno diše i otežano guta. Takvoj osobi aplikacija ne predviđa da ima rak pluća što je razumljivo.




Slika 29: Rezultat predviđanja aplikacije za bolesnu osobu




Slika 30: Unos podataka relativno zdrave osobe u aplikaciju




Slika 31: Rezultat predviđanja aplikacije za relativno zdravu osobu

# 10. Zaključak

Možemo zaključiti kako modeli za predviđanje raka pluća uvelike koriste ljudima da saznaju svoj rizik od raka. Također, valja zaključiti da pomoću tih istih modela osobe mogu uočiti najbitnije faktore koji mogu utjecati na rak pluća i samim time obratiti više pozornosti na njih.

Iz odabranih modela, njihovih podataka, metodologija i rezultata možemo zaključiti kako skoro svi obraćaju pozornost na spol, dob i status pušenju. Svi modeli su dali slične rezultate i to jako detaljno objašnjene.

U projektu su korištena tri različita algoritma strojnog učenja, a to su stablo odlučivanja, neuronske mreže i Bayesovu mrežu. U ovom slučaju neuronske mreže su bolje od stabla odlučivanja jer daju bolju točnost i preciznost podataka. Također se može vidjeti kako svaki algoritam daje različite rezultate nad istim skupom podataka pa za svaki skup podataka treba proanalizirati koji algoritam strojnog učenja više odgovara skupu podataka. Isto tako korištenjem malog uzorka podataka će se dobiti precizniji rezultati nego korištenjem velikog uzorka podataka, ali će rezultati dobiveni nad manjim uzorkom podataka biti manje relevantni. Također za točnije rezultate trebalo bi imati bolji skup podataka tj. skup podatak u kojem se nalazi podjednak broj zapisa o bolesnicima koji su imali rak pluća i koji nisu imali rak pluća dok u ovom skupu podataka koji je korišten za izradu ovog rada sadrži puno više zapisa o bolesnicima koji su imali rak pluća.

Na kraju se može zaključiti koliko je uistinu velika važnost izrade modela za predviđanje te bih završio s rečenicom “Modeli predviđanja spašavaju živote”!

# 11. Popis literature

A. Deppen, S., D. Blume, J., C. Aldrich, M., A. Fletcher, S., P. Massion, P., C. Walker, R., C. Chen, H., Speroff, T., A. Degesys, C., Pinkerman, R., S. Lambright, E., C. Nesbitt, J., B. Putman Jr., J., L. Grogan, E. (2014.):"Predicting lung cancer prior to surgical resection in patients with lung nodules". Journal of thoracic oncology, volume 9, issue 10, pages 1477-1484. Preuzeto 15.6.2022. na https://www.sciencedirect.com/science/article/pii/S1556086415307048

R. Spitz, M., Ki Hong, W., I. Amos, C., Wu, X., B. Schabath, M., Dong, Q., Shete, S., J. Etzel, C. (2007.):"A risk model for prediction of lung cancer". Journal of the National Cancer Institute, Volume 99, Issue 9, pages 715-726. Preuzeto 15.6.2022. na https://academic.oup.com/jnci/article/99/9/715/2544273?login=true

Park, S., Ho Nam, B., Ryung Yang, H., An Lee, J., Lim, H., Tae Han, J., Su Park, II, Shin, R., Soo Lee, J. (2013.):"Individualized risk prediction model for lung cancer in Korean men". Preuzeto 15.6.2022. na https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0054823

L. Etzel, C., Kachroo, S., Liu, M., D'Amelio, A., Dong, Q., L. Cote, M., S. Wenzlaff, A., Ki Hong, W., J. Greisinger, A., G. Schwarz, A., R. Spitz, M. (2008.):"Development and validation of lung cancer risk prediction model for African-Americans". Preuzeto 15.6.2022. na https://cancerpreventionresearch.aacrjournals.org/content/1/4/255.full

C. Muller, D., Johansson, M., Brennan, P. (2017.): "Lung cancer risk prediction model incorporating lung function: development and validation in the UK biobank prospective cohort study", Journal of clinical oncology. Preuzeto 15.6.2022. na https://web.archive.org/web/20180723084800id_/http://spiral.imperial.ac.uk/bitstream/10044/1/45909/6/jco.2016.69.2467.pdf

Cravit, R. (2021.) What is a decision tree and how to make one. Preuzeto 25.7.2022. s https://venngage.com/blog/what-is-a-decision-tree/

CFI (2022.) Decision tree. Preuzeto 25.7.2022. s https://corporatefinanceinstitute.com/resources/knowledge/other/decision-tree/

Žitko, B. (2003.) Upotreba stabla odlučivanja kod testiranja znanja metodom kviza. Preuzeto 25.7.2022. s

https://www.bib.irb.hr/145853/download/145853.Upotreba_stabla_odlucivanja_u_testiranju_znanja_pomocu_kviza.pdf

Artificial neural network (bez dat.). Preuzeto 27.7.2022. s https://www.javatpoint.com/artificial-neural-network

Brownlee, J. (2019.) A gentle introduction to Bayesian belief networks. Preuzeto 28.7.2022. s https://machinelearningmastery.com/introduction-to-bayesian-belief-networks/

Hotz, N. (2022.) What is CRISP DM? Preuzeto 15.8.2022. s https://www.datascience-pm.com/crisp-dm-2/

Cross - industry standard process for dana mining (2022.). Preuzeto 15.8.2022. s https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

# 12. Popis slika

# 13. Popis tablica