ROC analiza

Zekić, Patrik

SVEUČILIŠTE U RIJECI

TEHNIČKI FAKULTET

Prijediplomski sveučilišni studij elektrotehnike

Završni rad

ROC ANALIZA

Rijeka, rujan 2023.

Patrik Zekić
0069087137

SVEUČILIŠTE U RIJECI

TEHNIČKI FAKULTET

Prijediplomski sveučilišni studij elektrotehnike

Završni rad

ROC ANALIZA

Mentor: izv. prof. dr. sc. Ivan Dražić

Komentor: doc. dr. sc. Angela Bašić-Šiško

Rijeka, rujan 2023.

Patrik Zekić
0069087137

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
POVJERENSTVO ZA ZAVRŠNE ISPITE

Rijeka, 6. prosinca 2022.

Zavod:	**Zavod za matematiku, fiziku i strane jezike**
Predmet:	**Inženjerska matematika ET**
Grana:	**1.01.07 primijenjena matematika i matematičko modeliranje**

ZADATAK ZA ZAVRŠNI RAD

Pristupnik:	**Patrik Zekić (0069087137)**
Studij:	Sveučilišni prijediplomski studij elektrotehnike
Zadatak:	**ROC analiza // ROC analysis**

Opis zadatka:

U radu je potrebno definirati temeljne pojmove vezane uz ROC analizu kao što su osjetljivost, specifičnost te pozitivna i negativna prediktivnost. Nadalje potrebno je objasniti konstrukciju ROC krivulje i objasniti AUC vrijednost.
ROC analizu potrebno je uvesti u kontekstu analize kvalitete dijagnostičkog testa, kao i kvalitete klasifikatora unutar algoritama umjetne inteligencije, dok je u uvodnom dijelu rada potrebno objasniti osnovne pojmove teorije vjerojatnosti nužne za definiranje pojmova.
U završnom dijelu rada sve uvedene pojmove potrebno je primijeniti na nekoliko praktičnih primjera te interpretirati dobivene rezultate.

Rad mora biti napisan prema Uputama za pisanje diplomskih / završnih radova koje su objavljene na mrežnim stranicama studija.

Zadatak uručen pristupniku: 20. ožujka 2023.

Mentor:

Izv. prof. dr. sc. Ivan Dražić

Doc. dr. sc. Angela Bašić-Šiško (komentor)

Predsjednik povjerenstva za završni ispit:

Prof. dr. sc. Dubravko Franković

IZJAVA

Sukladno članku 7. stavku 1. Pravilnika o završnom radu, završnom ispitu i završetku sveučilišnih prijediplomskih studija Tehničkog fakulteta Sveučilišta u Rijeci od 4. travnja 2023., izjavljujem da sam samostalno izradio završni rad prema zadatku preuzetom dana 20. ožujka 2023.

Rijeka, 4. rujna 2023.

Patrik Zekić

Ovom prilikom htio bi se zahvaliti mentorima izv. prof. dr. sc. Ivanu Dražiću i doc. dr. sc. Angeli Bašić-Šiško koji su mi uvelike pomogli i doprinijeli izradi ovog rada, a najviše se zahvaljujem svojoj obitelji koja mi je bila podrška kroz cijelo moje dosadašnje školovanje.

Sadržaj

1. Uvod

Teorija vjerojatnosti je pojam kojim su se kroz povijest bavili mnogi vrhunski matematičari. Tako su se razvile razne definicije vjerojatnosti, gdje se među prvim značajnijim pojavljuje Laplace koji opisuje klasičnu i geometrijsku definiciju vjerojatnosti. Nakon Laplacea pojavljuje se Kolmogorov koji daje aksiomatsku definiciju vjerojatnosti. Sa vremenom je također objašnjena i uvjetna vjerojatnost u sklopu koje su razvijeni matematički izrazi poput formule potpune vjerojatnosti i Bayesove formule. Uvjetna vjerojatnost je poslužila kao temelj za razvoj i primjenu ROC analize.

ROC (*Receiver Operating Characteristic*) analiza je statistička metoda, gdje se pomoću tablica kontingencije ili matrica konfuzije mogu odrediti parametri koji služe za procjenu performansi određenog testa ili modela. ROC analiza se prvotno krenula razvijati tokom Drugog svjetskog rata pri analizi radara i detekciji signala. Ova metoda se zatim krenula primjenjivati i u raznim drugim granama poput psihofizike i medicine. Danas je primjena ROC analize široko rasprostranjena od primjerice psihologije i medicine pa do računarstva, elektrotehnike, ekonomije i financijskog sektora. S razvojem računalne tehnologije dolazi i do digitalizacije ROC analize. ROC krivulja je grafički prikaz analize pomoću kojeg je moguć prikaz potpunih performansi sustava. Također prikazom više ROC krivulja moguće je usporediti više modela te odrediti koji model ima bolje performanse izračunom AUC vrijednosti odnosno izračunom površine ispod krivulje. Jedan od značajnijih parametara kod ROC analize je i Youdenov indeks. Pomoću njega je moguće odrediti prijelomnu vrijednost modela odnosno granicu pomoću koje je moguće odrediti klasifikacijsku vrstu primjera. ROC analiza se danas primjenjuje i u umjetnoj inteligenciji primjerice pri analizi umjetnih neuronskih mreže i strojnom učenju. Uzevši sve u obzir ROC analiza se pokazuje kao vrlo učinkovita metoda sa širokom primjenom.

Tako se u ovom radu uvodno poglavlje bavi povješću vjerojatnosti i osnovnim definicijama. Objašnjene su klasična i geometrijska definicija vjerojatnosti kao i aksiomatska definicija vjerojatnosti. Naredno poglavlje opisuje uvjetnu vjerojatnost u čijem kontekstu su objašnjeni matematički izrazi poput formule potpune vjerojatnosti i Bayesove formule. U četvrtom poglavlju rada opisana je ROC analiza, gdje su objašnjeni parametri za analizu dijagnostičkog testa dobiveni pomoću tablice kontingencije te su isti izračunati na primjeru brzog antigenskog testa za COVID-19. Sljedeće poglavlje bavi se grafičkim prikazom ROC analize odnosno ROC krivuljom kao i izračunom AUC vrijendosti. U šestom poglavlju se uvodi pojam Youdenovog indeksa i prijelomne vrijednosti te je dan primjer, gdje je pomoću Youdenovog indeksa potrebno odrediti prijelomnu vrijednosti modela čiji je cilj odrediti potreban broj bodova na kolokviju da bi se studentu svidio kolegij. U završnom poglavlju prikazana je primjena ROC analize u umjetnoj inteligenciji, gdje se na primjeru klasifikacijskog modela, čiji je zadatak prepoznati studente s velikom šansom da padnu kolegij u cilju sprječavanja takvog ishoda, izvrši analiza i odrede svi parametri pomoću matrice konfuzije.

2. Uvod u teoriju vjerojatnosti

2.1. Povijesni aspekt

Povijest teorije vjerojatnosti počinje krajem šesnaestog i početkom sedamnaestog stoljeća kada su i zabilježeni prvi spisi odnosno rasprave na tu temu. Kao jedni od prvih začetnika, očeva teorije vjerojatnosti spominju se Pierre Fermat[1] i Blaise Pascal[2] koji su djelovali sredinom sedamnaestog stoljeća. Također jedan od začetnika teorije vjerojatnosti je Girolamo Cardano[3], koji je 1553. godine napisao djelo pod naslovom *Liber de ludo aleae* no navedeno djelo je objavljeno tek 1663. godine, dok su Fermatova i Pascalova objavljena otprilike desetak godina ranije pa se zato u tom kontekstu začetnika teorije vjerojatnosti više spominju Pascal i Fermat nego Cardano.

Slika 2.1. Pierre Fermat i Blaise Pascal, izvor: [1]

[1] Pierre de Fermat francuski je matematičar rođen 17. kolovoza 1601. godine. Fermat je svoje prve naznake da će postati vrstan matematičar pokazao još kao student restauracijom Apolonijevog djela *Plane Ioci*. Uz to što je bio vrstan matematičar Fermat je također završio i pravo.

[2] Blaise Pascal francuski je matematičar rođen 19. lipnja 1623. godine. Osim u matematici Pascal je bio poznat i kao filozof i fizičar. Prvo važnije istraživanje provodi 1647. godine kada je započeo mjerenje tlaka u atmosferi te zaključio prisutnost vakuuma. Također je proučavao i tlak u tekućinama te temeljni zakon hidrostatike danas nosi njegovo ime, Pascalov zakon.

[3] Girolamo Cardano talijanski je matematičar rođen 24. rujna 1501. godine. Cardano je uz matematiku studirao i medicinu u Pavliji i Padovi te kasnije stekao i doktorat iz medicine. Jedno od njegovih glavnih postignuća je svakako jednadžba za izračunavanje kubne jednadžbe odnosno Cardanove jednadžbe.

Fermatova i Pascalova djela su bila okidač nakon čega su se mnogi drugi krenuli baviti vjerojatnošću. Tako se početkom osamnaestog stoljeća pojavljuje Jacob Bernoulli[4] koji inspiriran igrama na sreću izdaje djelo *Ars Conjectandi*. Abraham de Moivre[5] se promiče u Bernoullijevog nasljednika te postavlja temelje centralnog graničnog teorema, kojeg će u budućnosti Laplace uspjeti dokazati. Thomas Bayes[6] i Joseph Lagrange[7] također su matematičari vrijedni spomena zbog svojeg doprinosa vrlo važnim istraživanjima polja vjerojatnosti. Nešto kasnije, krajem osamnaestog te početkom devetnaestog stoljeća pojavljuje se Pierre-Simon Laplace[8].

Laplace 1812. godine objavljuje djelo *Théorie analytique des probabilites* odnosno *Analitička teorija vjerojatnosti*. To djelo je osnova na kojoj će teorija vjerojatnosti nastati te je u njemu po prvi puta definiran koncept vjerojatnosti te je izvedena metoda običnih najmanjih kvadrata koju je kao student prethodno uspio razviti njemački matematičar Carl Gauss[9]. Tako se Laplaceu, uz Gaussovo odobrenje, pripisuje dokaz i primjena normalne distribucije u teoriji vjerojatnosti. Nakon Laplacea teorija vjerojatnosti se nastavila razvijati, ali uz određene poteškoće koje su proizlazile iz mišljenja tadašnjih matematičara da joj nedostaje teoretska podloga koja bi bila prihvaćena u sklopu matematike.

[4]Daniel Bernoulli švicarski je matematičar rođen 8. veljače 1700.godine. Svakako jedno od njegovih najvažnijih postignuća je izvod osnovne jednadžbe za gibanje fluida odnosno Bernoullijeva jednadžba te se smatra začetnikom hidrodinamike. Bernoulli je također bio vrstan oceanograf te anatom pa se tako bavio i oceanskim strujama i proučavao rad srca.

[5]Abraham de Moivre francuski je matematičar rođen 26. svibnja 1667. godine. Uz svoj doprinos razvoju teorije vjerojatnosti Moivre je poznat po svom radu sa kompleksnim brojevima i trigonometrijom pa je tako izveo formulu koja povezuje kompleksne brojeve i trigonometriju. Moivre je bio poznat i po svojim poznanstvima sa Isaacom Newtonom.

[6]Thomas Bayes engleski je matematičar rođen 1701. godine. Bayes je poznat kao statističar te je formulirao teorem o uvjetnoj vjerojatnost, Bayesov teorem, kao i teoriju subjektivne vjerojatnosti, Bayesovu statistiku. Također je bio i prezbiterijanski svećenik u Londonu te član *Royal Society*.

[7]Joseph-Louis Lagrange talijanski je matematičar rođen 25. siječnja 1736. godine. Već sa 19 godina postaje profesor matematike u Torinu te se krenuo baviti metodom obrade problema iz računa varijacije te će kasnije dobiti nagradu Francuske akademije. Također je bio i vrstan astronom te je istraživao stabilnosti točaka u interakciji triju tijela i otkrio točke libracije odnosno Lagrangeove točke koje će kasnije postati značajne u astronomiji.

[8]Pierre-Simon Laplace francuski je matematičar rođen 23. ožujka 1749. godine. Bio je član Francuske akademije znanosti. Kao astronom utemeljio je hipotezu porijekla Sunčeva sustava te bio jedan od prvih koji je najavljivao postojanje crnih rupa. U matematici i fizici Laplace je postigao mnogo toga. Neki od značajnijih matematičkih radova se vežu uz diferencijalne jednadžbe, Laplaceove transformacije,vjerojatnost i statistiku.

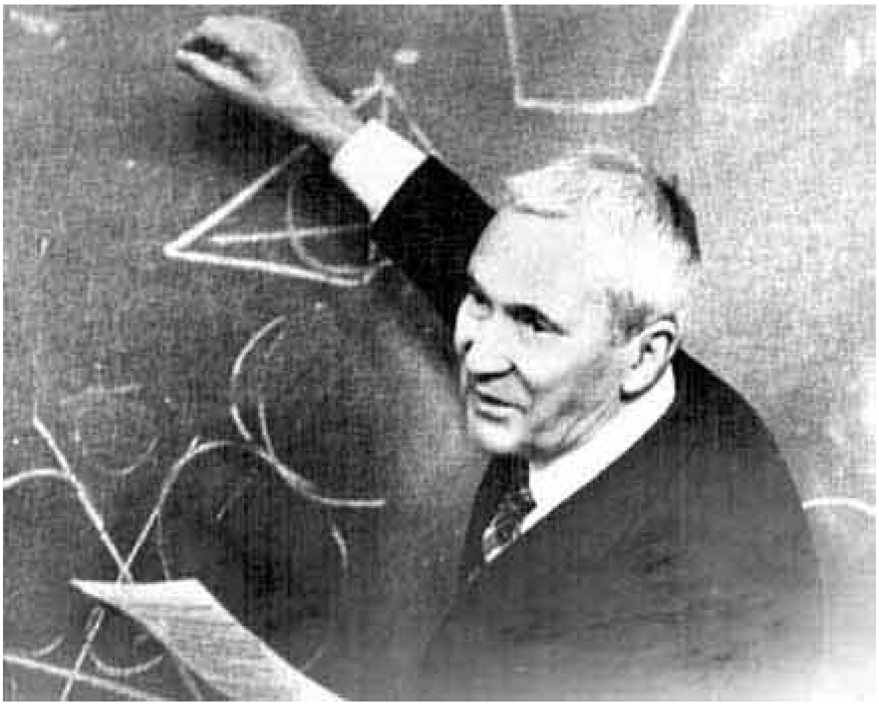
[9]Carl Friedrich Gauss njemački je matematičar rođen 30. travnja 1777. godine. Gauss je prve naznake genijalnosti krenuo pokazivati još u djetinjstvu, a prve rezultate postiže kao student kada je riješio problem konstrukcije pravilnih poligona ravnalom i šestarom. Kasnije je promoviran zahvaljujući doktorskoj disertaciji gdje je dokazao fundamentalni teorem algebre. Od velikog su značaja Gaussova istraživanja vezana za geometriju. Gauss se također bavio geodezijom i astronomijom.

Slika 2.2. Pierre-Simon Laplace, izvor: [2]

Nakon brojnih poteškoća i kritika kod teorije vjerojatnosti s kojima su se matematičari susretali teorija se počinje ponovno razvijati početkom dvadesetog stoljeća. Ruski matematičar Andrej Nikolajevič Kolmogorov[10], motiviran kritikama ostalih matematičara na račun teorije vjerojatnosti, objavljuje djelo *Temelji teorije vjerojatnosti* čime zaslužuje titulu eminencije vjerojatnosti. Motiviran Kolmogorovim djelom, Emile Borel[11] 1950. godine pridodaje svoj doprinos teoriji vjerojatnosti objavljivanjem knjige *Probabilite et Certitude*. Uz Kolmogorova i Borela, teorijom vjerojatnosti su se sredinom dvadesetog stoljeća bavili i mnogi drugi matematičari poput Paula Levyja, Mauricea Frecheta, Norberta Wienera. [4]

[10] Andrej Nikolajevič Kolmogorov ruski je matematičar rođen 25. travnja 1903. godine. Studirao je na Moskovskom državnom sveučilištu Lomonsov. Glavna Kolmogorovova istraživanja se vežu uz teoriju vjerojatnosti, topologiju i mehaniku. Za svog života je dobio mnoge prestižne nagrade za svoja djela.

[11] Félix Édouard Justin Émile Borel francuski je matematičar rođen 7. siječnja 1871. godine. Najpoznatiji radovi su u glavnom na temu teorije vjerojatnosti te teorije igara i strategije. Po Borelu se nazivaju mnogi matematički izrazi kao Borelova algebra, Borelov skup pa čak i krater na Mjesecu. Borel je također bio i političar te je tokom Drugog svjetskog rata bio i član francuskog pokreta otpora.



Slika 2.3. Andrej Nikolajevič Kolmogorov, izvor: [3]

2.2. Klasična i geometrijska definicija vjerojatnosti

Teorija vjerojatnosti je matematička disciplina čiji je zadatak formirati i proučavati matematički model slučajnog pokusa.

Pokus ili eksperiment je jedan od osnovnih pojmova koji pomaže u proučavanju i razumijevanju realnog svijeta. On je definiran pomoću odnosa uzroka i posljedica. Za samu realizaciju pokusa potrebno je ponavljati pokus proizvoljno konačno mnogo puta i poznavati sve moguće ishode. Ovisno o određenosti ishoda razlikuju se deterministički pokus te slučajni pokus.

Deterministički pokus je onaj u kojem je ishod jednoznačno određen uvjetima pokusa, dok u slučajnom pokusu ishod nije jednoznačno određen uvjetima pokusa. Slučajni pokus se temelji na tome da svako izvođenje pokusa mora rezultirati ishodom koji će odgovarati jednom i samo jednom elementarnom događaju.

Slučajni pokus je definiran osnovnim ishodima koji se međusobno isključuju. Ishodi se još nazivaju i elementarnim događajima. Elementarni događaji se označavaju malim grčkim slovom omega ω_1, ω_2, ... Skup Ω = $\{\omega_i$: ω_i = elementarni događaji, i = 1, ..., n$\}$ je neprazan skup i označava prostor elementarnih događaja.

Događaj je skup ishoda nekog pokusa odnosno podskup prostora elementarnih događaja te se označava sa velikim tiskanim slovom latinice A, B,... Oni se dijele na sigurne, nemoguće i povoljne događaje.

Siguran događaj je onaj događaj koji se mora dogoditi u svakom vršenju pokusa. Na primjer ako je pokus bacanje novčića, sigurni događaj je da će pasti glava ili pismo.

Nemoguć događaj je onaj događaj koji se nikada neće dogoditi. Na primjer ako je pokus bacanje kockice, nemoguć događaj je da će pasti broj 10 na kockici.

Povoljan događaj je onaj elementarni događaj čije pojavljivanje u pokusu povlači da se dogodio neki traženi događaj. Na primjer, neka je pokus bacanje kockice i neka je traženi događaj A da padne broj djeljiv s 3. Elementarni događaji odnosno mogući ishodi ovog pokusa su: pao je broj 1, pao je broj 2, pao je broj 3, pao je broj 4, pao je broj 5 i pao je broj 6, ali su samo ishodi pao je broj 3 i pao je broj 6 povoljni za događaj A.

Događaji se međusobno mogu isključivati te mogu činiti potpuni sistem. Za događaje A i B se kaže da se isključuju ukoliko je njihov presjek prazan skup odnosno ako se isti ne mogu dogoditi istovremeno. Na primjer neka je pokus bacanje kockice, a događaj A glasi pao je broj 1, a B pao je broj 4. U tom slučaju se događaji A i B međusobno isključuju.

Događaji čine potpun sistem događaja ukoliko se svi događaji međusobno isključuju i ako im je unija cijeli prostor elementarnih događaja. Na primjer neka je pokus bacanje kockice te događaj A glasi pao je paran broj, dok je događaj B pao je neparan broj. Događaji A i B čine potpuni sistem događaja zato što će se međusobno isključivati, ali njihova unija će biti svi mogući ishodi. [5,6]

2.2.1. Klasična definicija vjerojatnosti

U već navedenom djelu *Theorie analitique des probabillities* francuski matematičar Laplace uvodi pojmove klasične definicije vjerojatnosti *a priori* i *a posteriori*. Klasična definicija vjerojatnosti *a priori* se određuje pomoću omjera ovisno o prirodi događaja i definira se na sljedeći način.

Definicija 2.1. *Neka promatrani pokus ima n mogućih, jednako vjerojatnih elementarnih događaja koji se međusobno isključuju i neka odabrani događaj A ima m mogućih, povoljnih elementarnih događaja tada je vjerojatnost a priori događaja A jednaka:*

$$P(A) = \frac{m}{n}, \tag{2.1}$$

pri čemu je broj $P(A)$ vjerojatnost događaja A.

Primjer 2.1. *Korištenjem Definicije 1.1. Može se izračunati vjerojatnost da će pri bacanju kockice pasti broj 3. Svi mogući ishodi pokusa su pao je broj 1, pao je broj 2, pao je broj 3, pao je broj 4, pao je broj 5, pao je broj 6, dakle ima 6 mogućih ishoda. A povoljni događaj je samo 1 ukoliko padne broj 3. Vjerojatnost a priori za ovaj događaj iznosi:*

$$P(A) = \frac{m}{n} = \frac{1}{6} = 0.167. \tag{2.2}$$

Vjerojatnost događaja *a posteriori*, koja se također naziva i statističkom vjerojatnosti, definirana je i utvrđena nakon više ponavljanja pokusa te se računa kao omjer broja pojavljivanja događaja i ukupnog broja ponavljanja te se definira na sljedeći način.

Definicija 2.2. *Neka se promatrani pokus ponavlja n puta i neka se odabrani događaj A ponovi n_a puta. Tada je vjerojatnost a posteriori događaja A jednaka:*

$$P(A) = \lim_{n\to\infty} \frac{n_a}{n}, \tag{2.3}$$

pri čemu je broj $P(A)$ vjerojatnost događaja A.

Primjer 2.2. *Korištenjem Definicije 1.2. može se izračunati kolika je vjerojatnost da će paket isporučen poštom doći unutar 3 dana ukoliko je poznato da je isporučeno 900 paketa od kojih je 300 došlo unutar 3 dana. Vjerojatnost da će paket biti isporučen unutar 3 dana iznosi:*

$$P(A) = \lim_{n\to\infty} \frac{n_a}{n} = \frac{300}{900} = 0.33. \tag{2.4}$$

2.2.2. Geometrijska definicija vjerojatnosti

Uz klasične definicije vjerojatnosti *a priori* i *a posteriori* postoji i geometrijska definicija vjerojatnosti. Geometrijska vjerojatnost se definira na sljedeći način.

Definicija 2.3. *Neka promatrani pokus ima prostor ishoda Ω koji je jednak ograničenom izmjerivom podskupu $\boldsymbol{R}^n$ za koji vrijedi da je $m(\Omega) > 0$, gdje m predstavlja geometrijsku mjeru. Neka je događaj $A \subset \Omega$ tada je geometrijska vjerojatnost događaja A jednaka:*

$$P(A) = \frac{m(A)}{m(\Omega)}, \tag{2.5}$$

gdje je $m(A)$ mjera događaja A, a broj $P(A)$ vjerojatnost događaja A.

Geometrijska mjera je za $n = 1$ duljina, za $n = 2$ površina te za $n = 3$ volumen.

Primjer 2.3. *Hotelski bazen ima osvijetljenje koje mijenja boju svakih 30 sekundi. Osvijetljenje mijenja boju sljedećim rasporedom: prvih 30 sekundi bazen će biti osvijetljen zelenom bojom, zatim plavom bojom, zatim rozom bojom, pa zelenom i rozom bojom, zatim zelenom i plavom te crvenom bojom. Potrebno je izračunati kolika je vjerojatnost da će bazen biti osvijetljen zelenom bojom. Dakle bazen će kroz ciklus sve skupa izmijeniti 6 različitih kombinacija što znači da će trajati 180 sekundi, a ishodi povoljni za zadani događaj su kada je bazen osvijetljen zelenom, kada je osvijetljen zelenom i rozom te kada je osvijetljen zelenom i plavom. Dakle sve skupa 3 povoljna ishoda u ciklusu odnosno 90 sekundi. Tada je vjerojatnost da će bazen biti osvijetljen zelenom bojom:*

$$P(A) = \frac{m(A)}{m(\Omega)} = \frac{90}{180} = 0.5. \tag{2.6}$$

Primjer 2.4. *U kvadratu stranice duljine 4 centimetra je izabrana točka te je potrebno odrediti kolika je vjerojatnost da se točka nalazi unutar kruga upisanog kvadratu (događaj A). Dakle mjera* *traženog događaja će biti jednaka površini upisanog kruga, dok će mjera pokusa biti jednaka površini zadanog kvadrata. Vjerojatnost tada iznosi:*

$$P(A) = \frac{m(A)}{m(\Omega)} = \frac{\frac{4^2}{2^2} \cdot \pi}{4^2} = \frac{\pi}{4}. \tag{2.7}$$

2.3. Aksiomatska definicija vjerojatnosti

Vođen nedostatcima klasičnih definicija vjerojatnosti *a priori* i *a posteriori* ruski matematičar Kolmogorov razvija aksiomatsku definiciju vjerojatnosti te se zbog toga smatra ocem moderne teorije vjerojatnosti. Kolmogorov prvi put uvodi pojam aksiomatske definicije vjerojatnosti u svojoj monografiji *Osnovni pojmovi teorije vjerojatnosti*. Općenito, aksiom odnosno postulat je neka temeljna istina koja se ne dokazuje i služi kao osnova teorije, bilo matematičke, bilo logičke.

Neka je Ω prostor elementarnih događaja. Onda skup svih podskupova od Ω predstavlja skup svih mogućih događaja pokusa $P(\Omega)$. Podskup $F \subset P(\Omega)$ se naziva familija događaja.

Familija događaja koja ima sljedeća svojstva:

1. $\emptyset \in F$,
2. $A \subset F \Rightarrow A^c \subset F$,
3. $A_i \in F, i \in N \Rightarrow \bigcup_{i=1}^{\infty} A_i \in F$.

se naziva σ-algebra događaja.

Svakom događaju σ-algebre $A \in F$ može se pridružiti broj $P(A) \in [0,1]$ koji će predstavljati šansu da će se događaj A dogoditi. Aksiomatska vjerojatnost se definira na sljedeći način.

Definicija 2.4. *Preslikavanje $P : F \longrightarrow [0,1]$ sa sljedećim svojstvima:*

1. *$P(A) \geq 0$ za svaki $A \in F$,*
2. *$P(\Omega) = 1$,*
3. *$A_i \in F, i \in N, A_i \cap A_j = \emptyset, i \neq j \Rightarrow P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$,*

se naziva vjerojatnost.

Navedena svojstva su redom: svojstvo nenegativnosti, svojstvo normiranosti i svojstvo prebrojive aditivnosti.

Uređena trojka (Ω,F,P) naziva se prostorom vjerojatnosti. F predstavlja σ-algebru na Ω, dok P predstavlja vjerojatnost na F.

Prostor vjerojatnosti može biti diskretan odnosno n dimenzionalno diskretan. Ukoliko je Ω prebrojiv ili konačan skup elementarnih događaja, kažemo da je (Ω,F,P) diskretan prostor vjerojatnosti, a ako je Ω konačan skup elementarnih događaja u kojem je $|\Omega|$ = n, kažemo da je (Ω,$P(\Omega)$,P) n dimenzionalni diskretni prostor vjerojatnosti.

Korištenjem definicijskih svojstava vjerojatnosti, može se pokazati da vrijedi i sljedeće:

1. $P(A^c) = 1 - P(A)$,
2. $A \subseteq B \Rightarrow P(A) \leq P(B)$,
3. $A \cap B = \emptyset \Rightarrow P(A \cap B) = 0$,
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Navedena svojstva su redom: svojstvo suprotnog događaja, svojstvo monotonosti odnosno poddogađaja, svojstvo nemogućeg događaja te svojstvo unije dvaju događaja. Poznavanje ovih svojstava uvelike olakšava račun s vjerojatnosti. [7,8]

Primjer 2.5. *Bacaju se dvije kockice te je potrebno odrediti kolika je vjerojatnost da će pasti jednaki brojevi ili da će produkt brojeva koji su pali biti paran. Dakle 2 su događaja, događaj A koji označava da će pasti jednaki brojevi i događaj B odnosno da će produkt brojeva koji su pali na kockicama biti paran broj. Ovaj primjer moguće je riješiti primjenom svojstva vjerojatnosti unije dva događaja pa će tako rješenje glasiti:*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{6}{36} + \frac{27}{36} - \frac{3}{36} = \frac{30}{36} = 0.83. \tag{2.8}$$

3. Uvjetna vjerojatnost

3.1. Definicija

Tokom osamnaestog stoljeća engleski statističar Thomas Bayes uvodi pojam uvjetne vjerojatnosti te daje definiciju iste koja je objavljena posthumno.

Vjerojatnost događaja A ukoliko se dogodio događaj B naziva se uvjetna vjerojatnost događaja A uz ovisnost o događaju B te se označava kao $P(A|B)$. Ona je jednaka broju ishoda $A \cap B$ i broja ishoda događaja B.

Definicija 3.1. *Neka je* (Ω,F,P) *prostor vjerojatnosti te* $B \in F$ *događaj takav da je* $P(B) > 0$*, uvjetna vjerojatnost događaja* A *uz dano* B *je jednaka sljedećem izrazu:*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{3.1}$$

Svojstva uvjetne vjerojatnosti su:

1. $P(\emptyset|B) = 0$,
2. $P(A^c|B) = 1 - P(A|B)$,
3. $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B) - P(A_1 \cap A_2|B)$,
4. $A_1 \subseteq A_2 \Rightarrow P(A_1|B) \leq P(A_2|B)$.

Prethodna svojstva su redom: uvjetna vjerojatnost praznog skupa događaja uz dano B jednaka je 0, uvjetna vjerojatnost događaja suprotnog događaju A uz dano B jednaka je broju 1 umanjenom za uvjetnu vjerojatnost događaja A uz dano B, uvjetna vjerojatnost unije događaja A_1 i A_2 uz dano B jednaka je zbroju vjerojatnosti događaja A_1 i A_2 umanjenom za vjerojatnost presjeka događaja A_1 i A_2 uz dano B te ukoliko je događaj A_1 podskup ili jednak događaju A_2 onda će i uvjetna vjerojatnost događaja A_1 uz dano B biti manja ili jednaka od uvjetne vjerojatnosti događaja A_2. [9,10,11]

Primjer 3.1. *Na maratonu je prisustvovalo 1000 sudionika od kojih 450 muškaraca i 550 žena. Maraton je uspjelo završiti 350 žena i 300 muškaraca. Potrebno je odrediti kolika je vjerojatnost da je muška osoba uspješno istrčala maraton. Zadatak se sastoji od 2 događaja: događaj* A *odnosno osoba je istrčala maraton i događaj* B *odnosno osoba je muškog spola. Dakle vjerojatnost da je osoba istrčala maraton i da je muškarac jednaka je omjeru broja muškaraca koji su uspješno istrčali maraton i ukupnog broja sudionika, dok je vjerojatnost da je osoba koja je sudjelovala na* *maratonu muškarac jednaka omjeru ukupnog broja muškaraca koji su prisustvovali maratonu i ukupnom broju sudionika. Tako slijedi:*

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{300}{1000}}{\frac{450}{1000}} = \frac{300}{450} = 0.67. \tag{3.2}$$

3.2. Nezavisnost događaja

Događaji A i B čije su vjerojatnosti pozitivne ($P(A) > 0$ i $P(B) > 0$) su međusobno nezavisni ukoliko realizacija jednog događaja ne utječe na vjerojatnost realiziranja drugog događaja odnosno ako je:

$$P(A|B) = P(A), \tag{3.3}$$

$$P(B|A) = P(B). \tag{3.4}$$

Definicija 3.2. *Neka je (Ω,F,P) prostor vjerojatnosti, događaji $A, B \in F$ su međusobno nezavisni ako vrijedi:*

$$P(A \cap B) = P(A) \cdot P(B). \tag{3.5}$$

Primjer 3.2. *Ukoliko su A i B međusobno nezavisni događaji, gdje je $P(A) > 0$ i $P(B) > 0$, potrebno je dokazati da vrijedi izraz (2.3):*

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A). \tag{3.6}$$

Definicija 3.3. *Neka je $C \in F$, takav da je $P(C) > 0$. Tada su događaji A i B uvjetno nezavisni uz dani C ako vrijedi:*

$$P(A \cap B|C) = P(A|C) \cdot P(B|C). \tag{3.7}$$

Definicija 3.4. *Neka je (Ω,F,P) prostor vjerojatnosti i $A_i \in F$, $i \in I$ familija događaja. Familija događaja je nezavisna ako za svaki konačni podskup $F \subseteq I$ vrijedi:*

$$P\left(\bigcap_{i \in F} A_i\right) = \prod_{i \in F} P(A_i). \tag{3.8}$$

Definicija 3.5. *Familija događaja $A_i \in F$, $i \in I$ je po parovima nezavisna ako za sve $i, j \in I$, $i \neq j$ vrijedi:*

$$P(A_i \cap A_j) = P(A_i) \cdot P(A_j). \tag{3.9}$$

3.3. Formula potpune vjerojatnosti

Definicija 3.6. *Neka je Ω prostor događaja. Događaji $H_1, ..., H_n$ činit će potpuni sustav događaja ako vrijedi:*

1. $H_i \neq \emptyset$,

2. $H_i \cap H_j = \emptyset, i \neq j$,

3. $\bigcup_{i=1}^{n} H_i = \Omega$.

Prethodni izrazi predstavljaju svojstva potpunog sustava događaja. Događaj H_i mora biti moguć (1. svojstvo), događaji H_i se međusobno isključuju (2. svojstvo), unija događaja $H_1, ..., H_n$ je siguran događaj (3. svojstvo).

Događaji H_i se nazivaju hipotezama. Vjerojatnost svakog događaja A može se odrediti kao težinska sredina uvjetnih vjerojatnosti događaja A na hipoteze, s težinama jednakim vjerojatnosti hipoteza. [10,12]

Teorem 3.1. *Neka je (Ω,F,P) prostor vjerojatnosti i neka događaji $H_1, ..., H_n \in F$ čine potpuni sustav događaja. Tada formula potpune vjerojatnosti za svaki $A \in F$ glasi:*

$$P(A) = \sum_{i=1}^{n} P(H_i) \cdot P(A|H_i). \tag{3.10}$$

Pokusi u kojima se primijenjuje formula potpune vjerojatnosti se najčešće odvijaju kroz 2 faze: prva faza je ona u kojoj se izrealizira jedan od događaja H_i iz skupa događaja $H_1, ..., H_n$, a druga faza je faza u kojoj se promatra neki događaj A, čija vjerojatnost realizacije ovisi o ishodima iz prve faze.

Primjer 3.3. *Na stolu se nalaze 2 špila igraćih karata. U prvom špilu karti se nalazi 5 karata znaka tref, 3 karte znaka karo, 2 karte znaka pik i 5 karata znaka herc, dok se u drugom špilu karti nalaze 3 karte znaka tref, 5 karata znaka karo, 5 karata znaka pik i 2 karte znaka herc. Iz nasumičnog špila se izvlači jedna karta. Potrebno je odrediti kolika je vjerojatnost da će izvučena karta biti znaka tref.*

Navedena radnja se sastoji od 2 pokusa: prvi je odabir špila karti, a drugi je izvlačenje karte iz odabranog špila. Prvi pokus će imati dva ishoda: H_1 odnosno odabran je prvi špil i H_2 odnosno odabran je drugi špil. Tako će vjerojatnosti događaja H_1 i H_2 iznositi:

$$P(H_1) = \frac{1}{2}\,,\, P(H_2) = \frac{1}{2}. \tag{3.11}$$

Drugi pokus je izvlačenje karte iz špila. Ovaj pokus ovisi o prethodnom. Neka događaj A predstavlja da je izvučena karta tref. Tako izraz po formuli potpune vjerojatnosti da je izvučena karta tref glasi:

$$P(A) = P(H_1) \cdot P(A|H_1) + P(H_2) \cdot P(A|H_2). \tag{3.12}$$

Da bi se odredila vjerojatnost $P(A)$ iz prethodnog izraza potrebno je izračunati uvjetne vjerojatnosti $P(A|H_1)$ i $P(A|H_2)$. Uvjetne vjerojatnosti su jednake omjeru broja karti koje su tref i *ukupnog broja karti u špilu:*

$$P(A|H_1) = \frac{5}{5+3+2+5} = \frac{1}{3}\,,\; P(A|H_2) = \frac{3}{3+5+5+2} = \frac{1}{5}. \tag{3.13}$$

Sada je potrebno sve dobivene vrijednosti uvrstiti u izraz (2.12):

$$P(A) = P(H_1) \cdot P(A|H_1) + P(H_2) \cdot P(A|H_2) = \frac{1}{2} \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{5} = \frac{4}{15} = 0.267. \tag{3.14}$$

Dakle, vjerojatnost da će iz nasumičnog špila karti biti izvučena karta znaka tref je 0.267.

3.4. Bayesova formula

Svaka hipoteza H_k iz potpunog sustava događaja ima svoju vjerojatnost $P(H_k)$. Takva vjerojatnost se naziva originalna vjerojatnost odnosno apriorna vjerojatnost. To je vjerojatnost koja se pridružuje svakoj hipotezi H_k u nedostatku ostalih podataka odnosno ukoliko nije poznato da li se neki događaj A realizirao. Ako je poznato da se neki događaj A, na primjer u primjeru 2.2 da je izvučena karta tref, realizirao onda će se vjerojatnosti hipoteza modificirati. Tada je modificirana vjerojatnost hipoteza jednaka $P(H_k|A)$ te se naziva aposteriorna vjerojatnost. [12]

Teorem 3.2. *Neka događaji $H_1, ..., H_n$ čine potpun sustav događaja na prostoru vjerojatnosti (Ω,F,P) i neka je A promatran događaj. Tada za svaki $A \in F$, uz uvjet da je $P(A) > 0$, aposteriorna vjerojatnost hipoteze H_k iznosi:*

$$P(H_k|A) = \frac{P(A|H_k) \cdot P(H_k)}{\sum_{i=1}^{n} P(A|H_i) \cdot P(H_i)}. \tag{3.15}$$

Dobiveni izraz (2.15) naziva se Bayesovom formulom i dobiva se primjenom definicije uvjetne vjerojatnosti i formule potpune vjerojatnosti i to na sljedeći način:

$$\begin{aligned} P(H_k|A) &= \frac{P(H_k \cap A)}{P(A)} = \frac{P(H_k) \cdot P(A|H_k)}{P(A)} \\ &= \frac{P(A|H_k) \cdot P(H_k)}{\sum_{i=1}^{n} P(A|H_i) \cdot P(H_i)}. \end{aligned} \tag{3.16}$$

Primjer 3.4. *Laboratorijski test ima učinkovitost 96 % u otkrivanju bolesti kod osoba kod kojih je bolest prisutna. U 2 % slučajeva će test pokazati prisustvo bolesti iako je osoba zdrava. Pretpostavka je da 0.5 % populacije ima bolest. Potrebno je odrediti kolika je vjerojatnost da osoba koja je slučajno odabrana ima tu bolest ako je rezultat testa pozitivan.*

U pokusu su 2 bitna događaja: događaj H odnosno da je osoba bolesna i događaj A odnosno da je laboratorijski test pozitivan.

Iz dobivenih podataka može se zaključiti da uvjetna vjerojatnost da je test pozitivan i osoba je bolesna iznosi 0.96, dok je uvjetna vjerojatnost da je test pozitivan ako je osoba zdrava 0.02.

Vjerojatnost da je slučajno odabrana osoba bolesna iznosi 0.005, dok vjerojatnost da je slučajno izabrana osoba zdrava iznosi 0.995:

$$P(A|H) = 0.96\,,\ P(A|H^c) = 0.02\,,\ P(H) = 0.005\,,\ P(H^c) = 0.995. \tag{3.17}$$

Po Bayesovoj formuli, uvrštavanjem navedenih podataka, vjerojatnost da slučajno odabrana osoba ima bolest ako je rezultat testa pozitivan iznosi:

$$\begin{aligned} P(H|A) &= \frac{P(H)\cdot P(A|H)}{P(H)\cdot P(A|H) + P(H^c)\cdot P(A|H^c)} \\ &= \frac{0.005\cdot 0.96}{0.005\cdot 0.96 + 0.995\cdot 0.02} = 0.194. \end{aligned} \tag{3.18}$$

4. Analiza dijagnostičkog testa

4.1. Parametri za analizu dijagnostičkog testa

Dijagnostički testovi su osnovna mjerila prema čijim se rezultatima određuje odnosno postavlja dijagnoza. Nalazi dijagnostičkog testa mogu biti pozitivni ili negativni. Iako su dijagnostički testovi izuzetni važni pri postavljanju dijagnoze oni nisu savršeni te može doći do pogreške u rezultatima. Tako osobe koje su zdrave mogu biti klasificirane kao bolesne, a osobe koje su bolesne se mogu klasificirati kao zdrave. Upravo zbog takvih odstupanja i mogućih grešaka se nalazi vode kao pozitivni i negativni. Nalaz koji je pozitivan tako predstavlja veću vjerojatnost bolesti, dok nalaz koji je negativan predstavlja veću vjerojatnost da je osoba zdrava odnosno manja vjerojatnost bolesti.

Kako bi došlo do što manjih odstupanja i vjerodostojnosti testa, svaki dijagnostički test ima svoju valjanost. Valjanost testa je sposobnost testa da što pravilnije klasificira rezultate u bolesne i zdrave. Rezultati dijagnostičkog testa u odnosu na stvarno stanje se prikazuju tablicama kontingencije.

Tablice kontingencije su tablice dimenzija 2×2 pomoću kojih se određuju parametri dijagnostičkog testa kao i njegova pouzdanost. Tablica se sastoji od četiri ćelije. U gore lijevoj se zapisuje broj pozitivnih rezultata testova za osobe koje su bolesne (oznaka TP, eng. true positive), u gore desnoj se zapisuje broj pozitivnih rezultata testova za osobe koje su zdrave (oznaka FP, eng. false positive), u dole lijevoj se zapisuje broj negativnih rezultata testa za osobe koje su bolesne (oznaka FN, eng. false negative) te se u dole desnoj ćeliji zapisuje broj negativnih rezultata testa za osobe koje su zdrave (oznaka TN, eng. true negative).

Tablica 4.1. Tablica kontingencije

	Bolestan	Zdrav	
Pozitivan test	TP	FP	$TP+FP$
Negativan test	FN	TN	$FN+TN$
	$TP+FN$	$FP+TN$	

U Tablici 3.1. $TP+FN$ označava ukupan broj bolesnih ljudi, $FP+TN$ označava ukupan broj zdravih ljudi, $TP+FP$ označava ukupan broj pozitivnih rezultata testa te $TN+FN$ označava ukupan broj negativnih rezultata testa.

Osnovni parametri za analizu dijagnostičkog testa su: osjetljivost, specifičnost, preciznost, stopa pogreške, PPV odnosno pozitivna prediktivna vrijednost, NPV ili negativna prediktivna vrijednost i omjer vjerodostojnosti. [14]

Preciznost odnosno točnost dijagnostičkog testa se određuje kao omjer ukupnog broja ispravnih testova odnosno zbroj ispravno prepoznato bolesnih i zdravih te ukupni broj odrađenih testova. Ona predstavlja koliki je broj testova dao ispravne rezultate i definirana je izrazom:

$$PR = \frac{TP + TN}{TP + FP + TN + FN}, \tag{4.1}$$

gdje su TP, TN, FP, FN definirani u tablici 4.1 te je PR oznaka za preciznost.

Stopa pogreške dijagnostičkog testa se određuje kao omjer ukupnog broja neispravnih testova odnosno zbroj neispravno prepoznato bolesnih i zdravih te ukupni broj odrađenih testova. Ona predstavlja koliki je broj testova dao neispravne rezultate i definirana je izrazom:

$$ER = \frac{FP + FN}{TP + FP + TN + FN}, \tag{4.2}$$

gdje je ER oznaka za stopu pogreške.

Osjetljivost dijagnostičkog testa je vjerojatnost pozitivnog testa uz uvjet prisustva bolesti. Ona je jednaka omjeru ispravno prepoznato bolesnih i ukupno stvarno bolesnih i dana je izrazom:

Tablica 4.2. Izračun za osjetljivost

	Bolestan	Zdrav	
Pozitivan test	$\boxed{TP}$	FP	$TP + FP$
Negativan test	FN	TN	$FN + TN$
	$\boxed{TP + FN}$	$FP + TN$	

$$OS = \frac{TP}{TP + FN}, \tag{4.3}$$

gdje je OS oznaka za osjetljivost.

Pomoću osjetljivosti može se dobiti i vrijednost još jednog parametra koji se naziva proporcija lažno negativnih čiji zbroj skupa sa osjetljivosti daje broj jedan. Proporcija lažno negativnih definirana je sljedećim izrazom:

Tablica 4.3. Izračun za proporciju lažno negativnih

	Bolestan	Zdrav	
Pozitivan test	TP	FP	$TP + FP$
Negativan test	$\boxed{FN}$	TN	$FN + TN$
	$\boxed{TP + FN}$	$FP + TN$	

$$PLN = \frac{FN}{TP + FN} = 1 - OS, \tag{4.4}$$

gdje je PLN oznaka za proporciju lažno negativnih.

Specifičnost dijagnostičkog testa je vjerojatnost negativnog testa uz uvjet odsustva bolesti. Specifičnost je jednaka omjeru ispravno prepoznato zdravih odnosno negativnih i ukupnog broja stvarno zdravih te je definirana izrazom:

Tablica 4.4. Izračun za specifičnost

	Bolestan	Zdrav	
Pozitivan test	TP	FP	$TP + FP$
Negativan test	FN	$\boxed{TN}$	$FN + TN$
	$TP + FN$	$\boxed{FP + TN}$	

$$SP = \frac{TN}{FP + TN}, \tag{4.5}$$

gdje je SP oznaka za specifičnost.

Uz specifičnost se javlja proporcija lažno pozitivnih čija je ovisnost o specifičnosti analogna kao i proporcija lažno negativnih sa osjetljivosti. Dakle zbroj proporcije lažno pozitivnih skupa sa specifičnosti daje broj jedan prema izrazu:

Tablica 4.5. Izračun za proporciju lažno pozitivnih

	Bolestan	Zdrav	
Pozitivan test	TP	$\boxed{FP}$	$TP + FP$
Negativan test	FN	TN	$FN + TN$
	$TP + FN$	$\boxed{FP + TN}$	

$$PLP = \frac{FP}{FP + TN} = 1 - SP, \tag{4.6}$$

gdje je PLP oznaka za proporciju lažno pozitivnih.

Pomoću prethodna dva parametra, osjetljivosti i specifičnosti, može se odrediti i omjer vjerodostojnosti testa. Omjer vjerodostojnosti ukazuje na to koliko rezultat testa mijenja vjerojatnost bolesti. Omjer vjerodostojnosti se računa drugačije za one rezultate testa koji su pozitivni i one rezultate testa koji su negativni. Tako je omjer vjerodostojnosti za pozitivan test jednak kvocijentu senzitivnosti i proporcije lažno pozitivnih, dok je omjer vjerodostojnosti za negativan test jednak kvocijentu proporcije lažno negativnih i specifičnosti.

$$LR^{+} = \frac{OS}{1 - SP}, \tag{4.7}$$

gdje je LR^{+} oznaka za omjer vjerodostojnosti pozitivnog testa.

$$LR^{-} = \frac{1 - OS}{SP}, \tag{4.8}$$

gdje je LR^{-} oznaka za omjer vjerodostojnosti negativnog testa.

Omjer vjerodostojnosti se također može prikazati i kao omjer poslijetestne i prijetestne šanse da je osoba bolesna po izrazu:

$$LR = \frac{POTS}{PRTS}, \tag{4.9}$$

gdje je LR oznaka za omjer vjerodostojnosti, $POTS$ oznaka za poslijetestnu šansu, $PRTS$ oznaka za prijetestnu šansu.

Pa se tako preko omjera mogu dobiti i prijetestna i poslijetestna vjerojatnost. Prijetestna vjerojatnost je definirana sljedećim izrazom:

$$PRT = \frac{PRTS}{1 + PRTS}, \tag{4.10}$$

gdje je PRT oznaka za prijetestnu vjerojatnost.

Dok je poslijetestna vjerojatnost dana izrazom:

$$POT = \frac{POTS}{1 + POTS}, \tag{4.11}$$

gdje je POT oznaka za poslijetestnu vjerojatnost.

Formula (4.9) također vrijedi pri izračunu poslijetestne šanse za pozitivne rezultate testa, koja se označava sa $POTS^{+}$ i poslijetestne šanse za negativne rezultate testa, koja se označava sa $POTS^{-}$, dok formula (4.11) vrijedi i za izračun poslijetestne vjerojatnosti za pozitivne rezultate testa, koja se označava sa POT^{+} i poslijetestne vjerojatnosti za negativne rezultate testa, koja se označava sa POT^{-}.

Osjetljivost i specifičnost dijagnostičkog testa skupa čine mjere valjanosti testa. To su uvjetne vjerojatnosti koje se još nazivaju i nozološke vjerojatnosti zato što ovise isključivo o danoj bolesti, a ne o njezinoj prevalenciji u uzorku. Prevalnecija predstavlja omjer ukupno stvarno bolesnih u odnosu na ukupni broj testova. Pa se tako uvode novi parametri PPV, pozitivna prediktivna vrijednost i NPV, negativna prediktivna vrijednost. One su također dijagnostičke vjerojatnosti, ali za razliku od prethodnih ovise i o prevalenciji bolesti u uzorku.

Pozitivna prediktivna vrijednost se definira kao omjer ispravno prepoznato bolesnih i ukupnog broja prepoznato bolesnih prema izrazu:

Tablica 4.6. Izračun za pozitivnu prediktivnu vrijednost

	Bolestan	Zdrav	
Pozitivan test	$\boxed{TP}$	FP	$\boxed{TP + FP}$
Negativan test	FN	TN	$FN + TN$
	$TP + FN$	$FP + TN$	

$$PPV = \frac{TP}{TP + FP}, \tag{4.12}$$

gdje je PPV oznaka za pozitivnu prediktivnu vrijednost.

Negativna prediktivna vrijednost se definira kao omjer ispravno prepoznato zdravih i ukupnog broja prepoznato zdravih prema izrazu:

Tablica 4.7. Izračun za negativnu prediktivnu vrijednost

	Bolestan	Zdrav	
Pozitivan test	TP	FP	$TP + FP$
Negativan test	FN	$\boxed{TN}$	$\boxed{FN + TN}$
	$TP + FN$	$FP + TN$	

$$NPV = \frac{TN}{TN + FN}, \tag{4.13}$$

gdje je NPV oznaka za negativnu prediktivnu vrijednost.

4.2. Povezanost uvjetne vjerojatnosti i parametara za analizu dijagnostičkog testa

Parametri za analizu dijagnostičkog testa se mogu prikazati kao uvjetne vjerojatnosti događaja. Pri analizi dijagnostičkog testa moguća su 4 događaja: događaj A odnosno test je pozitivan, događaj A^c odnosno test je negativan, događaj B odnosno osoba je bolesna i događaj B^c odnosno osoba je zdrava.

Uz definirane moguće događaje pri analizi dijagnostičkog testa i poznavanje svojstva uvjetne vjerojatnosti moguće je prikazati osnovne parametre dijagnostičkog testa na sljedeći način.

Osjetljivost dijagnostičkog testa tako predstvalja vjerojatnost točnog predviđanja kod bolesnih osoba te se može zapisati kao uvjetna vjerojatnost događaja A uz uvjet da se dogodio događaj B:

$$OS = P(A|B). \tag{4.14}$$

Specifičnost dijagnostičkog testa predstavlja vjerojatnost točnog predviđanja kod zdravih osoba te se može zapisati kao uvjetna vjerojatnost događaja A^c uz uvjet da se dogodio događaj B^c:

$$SP = P(A^c|B^c). \tag{4.15}$$

Pozitivna prediktivna vrijednost predstavlja vjerojatnost točnog predviđanja pozitivnog testa odnosno kolika je vjerojatnost da je osoba sa pozitivnim testom stvarno bolesna te se može zapisati kao uvjetna vjerojatnost događaja B uz uvjet da se dogodio događaj A:

$$PPV = P(B|A). \tag{4.16}$$

Dok će negativna prediktivna vrijednost predstavljati vjerojatnost točnog predviđanja negativnog testa odnosno vjerojatnost da je osoba sa negativnim testom stvarno zdrava te se može zapisati kao uvjetna vjerojatnost događaja B^c uz uvjet da se dogodio događaj A^c:

$$NPV = P(B^c|A^c). \tag{4.17}$$

4.3. Primjer analize dijagnostičkog testa

Primjer 4.1. *Pojavom bolesti COVID-19 došlo je do potrebe korištenja antigenskih testova. Pri masovnom testiranju uzet je bris 1000 ljudi kod kojih su unutar tjedan dana prepoznati simptomi bolesti od kojih 50 osoba stvarno ima bolest COVID-19. Brzim antigenskim testom BAT utvrđeno je kako je 45 osoba pozitivno na COVID-19 od kojih 5 osoba ima lažno pozitivan nalaz, odnosno nema bolest COVID-19 te je 955 osoba negativno na COVID-19 od kojih 10 osoba ima lažno negativan nalaz, odnosno ima bolest COVID-19.*

Kako bi se izračunali parametri potrebni za analizu dijagnostičkog testa potrebno je navedene podatke unijeti u tablicu kontingencije.

Tablica 4.8. Tablica kontingencije BAT testa

	Bolestan	Zdrav	
Pozitivan test	40	5	45
Negativan test	10	945	955
	50	950	

Iz prethodne tablice moguće je iščitati vrijednosti varijabli TP, TN, FP, FN. Tako je broj stvarno pozitivnih odnosno TP jednak 40. Broj lažno pozitivnih odnosno FP iznosi 5. Broj stvarno negativnih odnosno TN iznosi 945, dok broj lažno negativnih odnosno FN iznosi 10. Pomoću navedenih podataka i tablice kontingencije za brzi antigenski test moguće je odrediti vrijednosti potrebnih parametara za analizu.

Prvo je potrebno izračunati vrijednosti preciznosti i stope pogreške testa. Preciznost BAT, brzog antigenskog testa se određuje prema formuli (4.1) i dobiva se:

$$PR = \frac{TP + TN}{TP + FP + TN + FN} = \frac{40 + 945}{40 + 5 + 945 + 10} = 0.985. \tag{4.18}$$

Stopa pogreške BAT, brzog antigenskog testa se određuje prema formuli (4.2) i dobiva se:

$$ER = \frac{FP + FN}{TP + FP + TN + FN} = \frac{5 + 10}{40 + 5 + 945 + 10} = 0.015. \tag{4.19}$$

Iz prethodnih izračuna vidljivo je da preciznost odnosno točnost brzog antigenskog testa iznosi 98.5%, dok stopa pogreške iznosi 1.5%. Nadalje, potrebno je odrediti osjetljivost i specifičnost antigenskog testa. Osjetljivost brzog antigenskog testa određuje se prema formuli (4.3) na sljedeći način:

$$OS = \frac{TP}{TP + FN} = \frac{40}{50} = 0.8. \tag{4.20}$$

Osjetljivost brzog antigenskog testa iznosi 0.8 što znači da je šansa da osoba koja stvarno ima COVID-19 bude prepoznata kao pozitivna na bolest 80%. Uz poznavanje vrijednosti osjetljivosti moguće je i odrediti proporciju lažno negativnih prema formuli (4.4):

$$PLN = \frac{FN}{TP + FN} = 1 - OS = 1 - 0.8 = 0.2. \tag{4.21}$$

Specifičnost brzog antigenskog testa određuje se prema formuli (4.5):

$$SP = \frac{TN}{FP + TN} = \frac{945}{950} = 0.9947. \tag{4.22}$$

Dakle, specifičnost brzog antigenskog testa iznosi 0.9947 što znači da je šansa da osoba koja nema COVID-19 bude na testu prepoznata kao takva, odnosno negativna 99.47%. Pomoću specifičnosti moguće je odrediti i proporciju lažno pozitivnih prema formuli (4.6):

$$PLP = \frac{FP}{FP + TN} = 1 - SP = 1 - 0.9947 = 0.0053. \tag{4.23}$$

Uz poznate vrijednosti osjetljivosti i specifičnosti moguće je odrediti omjere vjerodostojnosti kako za pozitivne tako i za negativne rezultate testa. Omjer vjerodostojnosti za pozitivne rezultate testa izračunava se pomoću formule (4.7):

$$LR^{+} = \frac{OS}{1 - SP} = \frac{0.8}{1 - 0.9947} = 150.94. \tag{4.24}$$

Omjer vjerodostojnosti za pozitivne rezultate BAT brzog antigenskog testa iznosi 150.94. Poznavajući omjer vjerodostojnosti za pozitivne rezultate testa moguće je odrediti iznos poslijetestne šanse za COVID-19 za pozitivne rezultate preko izraza:

$$LR = \frac{POTS}{PRTS}. \tag{4.25}$$

Da bi se odredila poslijetestna šansa iz prethodno navedenog izraza potrebno je izračunati prijetestnu šansu. Znajući da prijetestna vjerojatnost za COVID-19 iznosi 0.005 pomoću formule (4.10) moguće je odrediti prijetestnu šansu za COVID-19:

$$PRTS = \frac{PRT}{1 - PRT} = \frac{0.005}{1 - 0.005} = 0.005. \tag{4.26}$$

Prijetestna šansa za COVID-19 iznosi 0.005. Dobivenu prijetestnu šansu potrebno je uvrstiti u sljedeći izraz kako bi se dobila poslijetestna šansa za pozitivne rezultate:

$$POTS^{+} = LR^{+} \cdot PRTS = 150.94 \cdot 0.005 = 0.7547. \tag{4.27}$$

Dobivenu poslijetestnu šansu koja iznosi 0.7547 moguće je pretvoriti u posijetestnu vjerojatnost pomoću formule (4.11):

$$POT^{+} = \frac{POTS^{+}}{1 + POTS^{+}} = \frac{0.7547}{1 + 0.7547} = 0.43. \tag{4.28}$$

Na isti način se određuju i vrijednosti parametara za negativne rezultate testa pa se tako do omjera vjerodostojnosti za negativne rezultate testa dolazi preko formule (4.8):

$$LR^{-} = \frac{1 - OS}{SP} = \frac{1 - 0.8}{0.9947} = 0.2. \tag{4.29}$$

Omjer vjerodostojnosti negativnog testa iznosi 0.2 te je uz poznatu vrijednost prijetestne šanse moguće odrediti poslijetestnu šansu za negativne testove:

$$POTS^{-} = LR^{-} \cdot PRTS = 0.2 \cdot 0.005 = 0.001. \tag{4.30}$$

Dobivenu poslijetestnu šansu koja iznosi 0.001 moguće je pretvoriti u posijetestnu vjerojatnost pomoću formule (4.11):

$$POT^{-} = \frac{POTS^{-}}{1 + POTS^{-}} = \frac{0.001}{1 + 0.001} = 0.00099. \tag{4.31}$$

Nakon izračunatih mjera valjanosti testa koje ovise isključivo o samoj bolesti potrebno je odrediti i one koje ovise o prevalenciji bolesti u uzorku. To su, kao što je i prethodno navedeno PPV, pozitivna prediktivna vrijednost i NPV, negativna prediktivna vrijednost. Pozitivna prediktivna vrijednost će se uz pomoć tablice kontingecije BAT testa izračunati prema formuli (4.12):

$$PPV = \frac{TP}{TP + FP} = \frac{40}{45} = 0.89. \tag{4.32}$$

Dok će se negativna prediktivna vrijednost BAT testa izračunati prema formuli (4.13) i biti jednaka:

$$NPV = \frac{TN}{TN + FN} = \frac{945}{955} = 0.99. \tag{4.33}$$

Iz prethodnih izračuna pozitivne prediktivne vrijednosti i negativne prediktivne vrijednosti može se zaključiti da je brzi antigenski test ispravno prepoznao 89% osoba koje su pozitivne kao one koje imaju COVID-19 te da je ispravno prepoznao 99% osoba koje su negativne kao zdrave. Svi izračunati parametri brzog antigenskog testa prikazani su u sljedećoj tablici (Tablica 4.9.).

Tablica 4.9. Parametri BAT testa

PARAMETAR	VRIJEDNOST
PR	0.985
ER	0.015
OS	0.8
PLN	0.2
SP	0.9947
PLP	0.0053
$PRTS$	0.005
LR^+	150.94
$POTS^+$	0.7547
POT^+	0.43
LR^-	0.2
$POTS^-$	0.001
POT^-	0.00099
PPV	0.89
NPV	0.99

Primjer 4.2. *Testiran je klasifikacijski model koji predviđa je li e-mail spam ili nije na temelju određenih karakteristika. Model je testiran na uzorku od 500 e-mailova od kojih je 250 spam e-mailova i 250 legitimnih e-mailova. Testiranjem klasifikacijskog modela dobiveni su sljedeći rezultati: 225 e-mailova model je prepoznao kao pozitivne odnosno spam e-mailove od kojih je 25 lažno pozitivnih odnosno nisu spam e-mailovi već obični te 275 e-mailova kao negativne odnosno legitimne e-mailove od kojih 50 lažno negativnih odnosno spam e-mailova.*

Na temelju dobivenih podataka moguće je napraviti tablicu kontingencije klasifikacijskog modela kako bi se mogli odrediti njegovi parametri.

Tablica 4.10. Tablica kontingencije klasifikacijskog modela

	Spam	Legitiman	
Pozitivan test	200	25	225
Negativan test	50	225	275
	250	250	

Iz prethodne tablice moguće je iščitati vrijednosti varijabli TP, TN, FP, FN. Tako je broj stvarno pozitivnih odnosno TP jednak 200. Broj lažno pozitivnih odnosno FP iznosi 25. Broj stvarno negativnih odnosno TN iznosi 225, dok broj lažno negativnih odnosno FN iznosi 50. Pomoću navedenih podataka i tablice kontingencije za klasifikacijski model moguće je odrediti vrijednosti potrebnih parametara za analizu.

Prvo je potrebno izračunati vrijednosti preciznosti i stope pogreške modela. Preciznost klasifikacijskog modela se određuje prema formuli (4.1):

$$PR = \frac{TP + TN}{TP + FP + TN + FN} = \frac{200 + 225}{200 + 25 + 225 + 50} = 0.85. \tag{4.34}$$

Stopa pogreške klasifikacijskog modela za predviđanje spam e-mailova određuje se prema formuli (4.2) i iznosi:

$$ER = \frac{FP + FN}{TP + FP + TN + FN} = \frac{25 + 50}{200 + 25 + 225 + 50} = 0.15. \tag{4.35}$$

Iz prethodnih izračuna vidljivo je da preciznost odnosno točnost klasifikacijskog modela iznosi 85%, dok stopa pogreške iznosi 15%. Nadalje potrebno je odrediti osjetljivost i specifičnost modela. Osjetljivost se određuje prema formuli (4.3) na sljedeći način:

$$OS = \frac{TP}{TP + FN} = \frac{200}{250} = 0.8. \tag{4.36}$$

Osjetljivost klasifikacijskog modela iznosi 0.8 što znači da je šansa da model prepozna ispravno e-mail koji je stvarno spam 80%. Uz poznavanje vrijednosti osjetljivosti moguće je i odrediti proporciju lažno negativnih prema formuli (4.4):

$$PLN = \frac{FN}{TP + FN} = 1 - OS = 1 - 0.8 = 0.2. \tag{4.37}$$

Specifičnost klasifikacijskog modela određuje se prem aformuli (4.5) i jednaka je:

$$SP = \frac{TN}{FP + TN} = \frac{225}{25 + 225} = 0.9. \tag{4.38}$$

Dakle specifičnost klasifikacijskog modela iznosi 0.9 što znači da je šansa da model prepozna legitiman e-mail kao takav 90%. Pomoću specifičnosti moguće je odrediti i proporciju lažno pozitivnih prema formuli (4.6):

$$PLP = \frac{FP}{FP + TN} = 1 - SP = 1 - 0.9 = 0.1. \tag{4.39}$$

Uz poznate vrijednosti osjetljivosti i specifičnosti moguće je odrediti omjere vjerodostojnosti kako za pozitivne odnosno spam e-mailove tako i za negativne rezultate testa odnosno legitimne e-mailove. Omjer vjerodostojnosti za pozitivne rezultate testa izračunava se prema formuli (4.7):

$$LR^{+} = \frac{OS}{1 - SP} = \frac{0.8}{1 - 0.9} = 8. \tag{4.40}$$

Omjer vjerodostojnosti za pozitivne rezultate odnosno prepoznavanje spam e-mailova klasifikacijskim modelom iznosi 8. Poznavajući omjer vjerodostojnosti za pozitivne rezultate testa moguće je odrediti iznos poslijetestne šanse za pozitivne rezultate preko izraza:

$$LR = \frac{POTS}{PRTS}. \tag{4.41}$$

Da bi se odredila poslijetestna šansa iz prethodno navedenog izraza potrebno je izračunati prijetestnu šansu. Znajući da prijetestna vjerojatnost iznosi 0.05 pomoću formule (4.10) moguće je odrediti prijetestnu šansu:

$$PRTS = \frac{PRT}{1 - PRT} = \frac{0.05}{1 - 0.05} = 0.0526. \tag{4.42}$$

Prijetestna šansa iznosi 0.0526. Dobivenu prijetestnu šansu potrebno je uvrstiti u sljedeći izraz kako bi se dobila poslijetestna šansa za pozitivne rezultate:

$$POTS^{+} = LR^{+} \cdot PRTS = 8 \cdot 0.0526 = 0.421. \tag{4.43}$$

Dobivenu poslijetestnu šansu koja iznosi 0.421 moguće je pretvoriti u posijetestnu vjerojatnost prema formuli (4.11):

$$POT^{+} = \frac{POTS^{+}}{1 + POTS^{+}} = \frac{0.421}{1 + 0.421} = 0.3. \tag{4.44}$$

Na isti način se određuju i vrijednosti parametara za negativne rezultate testa klasifikacijskog modela pa se tako do omjera vjerodostojnosti za negativne rezultate testa dolazi preko formule (4.8):

$$LR^{-} = \frac{1 - OS}{SP} = \frac{1 - 0.8}{0.9} = 0.22. \tag{4.45}$$

Omjer vjerodostojnosti negativnog testa iznosi 0.22 te je uz poznatu vrijednost prijetestne šanse moguće odrediti poslijetestnu šansu za negativne testove:

$$POTS^{-} = LR^{-} \cdot PRTS = 0.22 \cdot 0.0526 = 0.012. \tag{4.46}$$

Dobivenu poslijetestnu šansu koja iznosi 0.012 moguće je pretvoriti u posijetestnu vjerojatnost prema formuli (4.11):

$$POT^{-} = \frac{POTS^{-}}{1 + POTS^{-}} = \frac{0.012}{1 + 0.012} = 0.0119. \tag{4.47}$$

Nakon izračunatih mjera valjanosti modela potrebno je odrediti i prediktivne vrijednosti klasifikacijskog modela. To su kao što je i prethodno navedeno PPV, pozitivna prediktivna vrijednost i NPV, negativna prediktivna vrijednost. Pozitivna prediktivna vrijednost će se uz pomoć tablice kontingecije klasifikacijskog modela izračunati prema formuli (4.12):

$$PPV = \frac{TP}{TP + FP} = \frac{200}{200 + 25} = 0.89. \tag{4.48}$$

Dok će se negativna prediktivna vrijednost klasifikacijskog modela izračunati prema formuli (4.13) i biti jednaka:

$$NPV = \frac{TN}{TN + FN} = \frac{225}{225 + 50} = 0.82. \tag{4.49}$$

Iz prethodnih izračuna pozitivne prediktivne vrijednosti i negativne prediktivne vrijednosti može se zaključiti da je klasifikacijski model ispravno prepoznao 89% spam e-mailova kao takve te da je ispravno prepoznao 82% legitimnih e-mailova kao takve. Svi izračunati parametri klasifikacijskog modela prikazani su u sljedećoj tablici (Tablica 4.11.).

Tablica 4.11. Parametri klasifikacijskog modela

PARAMETAR	VRIJEDNOST
PR	0.85
ER	0.15
OS	0.8
PLN	0.2
SP	0.9
PLP	0.1
$PRTS$	0.05
LR^+	8
$POTS^+$	0.421
POT^+	0.3
LR^-	0.22
$POTS^-$	0.012
POT^-	0.0119
PPV	0.89
NPV	0.82

5. ROC krivulja i AUC vrijednost

5.1. Povijest ROC analize

ROC (*Receiver Operating Characteristic*) analiza je statistička metoda kojoj je svrha procjena kvalitete nekog sustava u kojem postoji više vrijednosti ili klasa koje je potrebno razlikovati kao što su u prethodnim primjerima bili dijagnostički test sa osobama pozitivnim na COVID-19 i onim zdravima te klasifikacijski model čija je zadaća bila razlikovati spam e-mailove od običnih, legitimnih e-mailova. Danas je primjena ove metode široko rasprostranjena od medicine i ekonomije pa do računarstva, elektrotehnike i mnogih drugih grana. Kao i svaka druga metoda ROC analiza se kroz povijest postepeno razvijala te je rezultat mnogih istraživanja i pokusa.

Prva primjena ROC analize zabilježena je za vrijeme Drugog svjetskog rata za analizu radara i detekciju signala. Razvili su je američki inženjeri kao posljedicu napada na Pearl Harbour 1941. godine. Ideja razvitka ROC analize je bila izmjeriti performanse radara kako bi na vrijeme mogli predvidjeti i detektirati prisutnost japanskih zrakoplova.

U ranim 1950-ima ROC analiza se koristi u psihofizici kako bi se procijenila detekcija slabih signala kako kod ljudi tako i kod životinja. Tako je američki psiholog John A. Swets krenuo proučavati sposobnost ljudi da prepoznaju prisutnost signala uz prisutnost smetnji i šumova. Nakon mnogih istraživanja Swets je uspio postaviti temelje i doprinijeti razvoju signalnog detektora. Swets je također uveo ROC krivulju kako bi vizualizirao cijelu analizu te samim time postavio temelje za primjenu ROC analize.

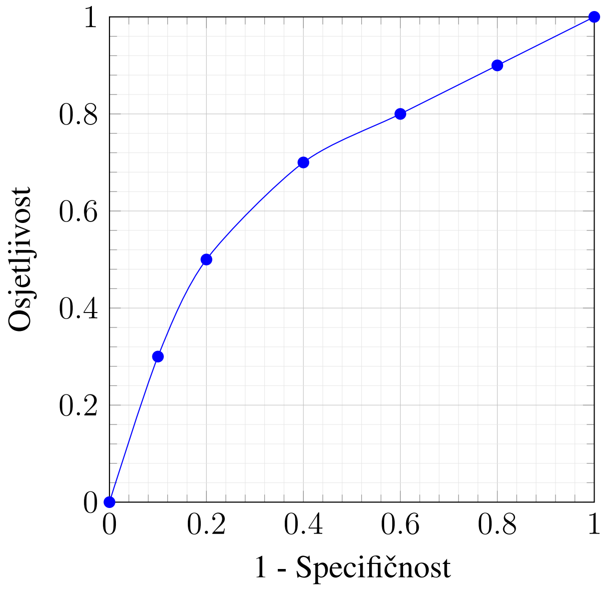
Krajem 1960-ih i početkom 1970-ih primjena ROC analize se kreće prakticirati i u medicini, posebno u radiologiji. Jedan od najznačajnijih i najzaslužnijih za razvoj ROC analize u medicini je američki radiolog William Douglas Lusted. On je u svojim istraživanjima primijenio ROC analizu na dijagnostičke testove. U svom djelu *Signal Detectability and Medical Decision-Making* objavljenom 1971. godine Lusted je objasnio koncept ROC analize u kontekstu dijagnostičkih testova na primjeru detekcije tumora na rendgenskim snimcima. Također je objasnio kako se ovom metodom mogu poboljšati performanse testova kao i kako da se njima ispravno koriste liječnici kako bi se pogrešne odluke pri dijagnostici svele na minimum. Tokom 1970-ih američki statističar Robert Dorfman razvija matematički okvir za analizu dijagnostičkog testa kako bi se omogućila preciznija procjena performansi istog.

Tokom 1980-ih i 1990-ih godina dolazi do značajnog razvoja računala pa se tako i ROC analiza postepeno implementirala u računalne tehnologije. Razvoj računalne tehnologije je doprinio boljem razumijevanju ROC krivulje i njenih karakteristika. Također ROC analiza je zahtijevala mnoge izračune koji su sada uz razvoj raznih softverskih alata postali znatno jednostavniji i precizniji.

Početkom 2000-ih godina ROC analiza se pokazala i kao vrlo korisna pri evaluaciji tehnika strojnog učenja. Tako omogućava bolju procjenu performansi klasifikacijskog modela, usporedbu više modela kako bi se utvrdio kvalitetniji, optimizaciju pragova kako bi se pronašao odgovarajući prag klasifikacije, procjenu kvalitete modela za primjerice prijevare ili kreditne ocjene u financijskom sektoru. Sve u svemu primjena ROC analize je danas široko rasprostranjena od vojne industrije, psihologije, medicine pa do računarstva, elektrotehnike, ekonomije i financija. [15]

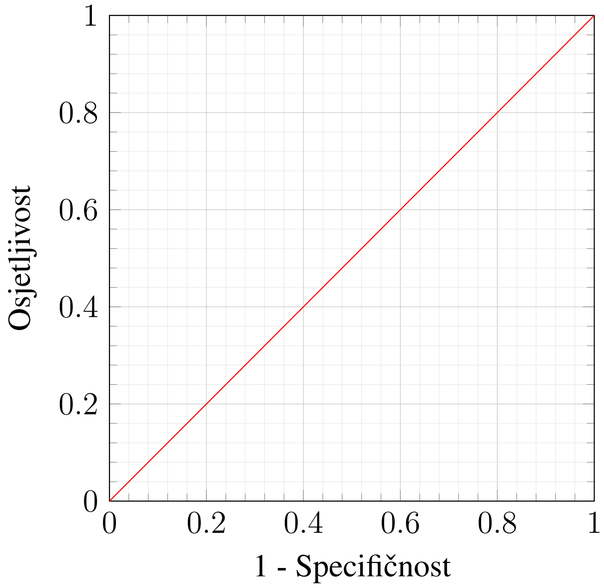
5.2. ROC krivulja

Grafički prikaz odnosa osjetljivosti odnosno stvarno pozitivne stope i lažno pozitivne stope za različite pragove klasifikacije naziva se ROC krivulja. Pomoću nje se mogu u potpunosti prikazati performanse raznih sustava, klasifikacijskih modela i dijagnostičkih testova. Dakle ROC krivulja je graf čija os apscisa predstavlja lažno pozitivnu stopu odnosno 1-specifičnost, dok os ordinata predstavlja osjetljivost odnosno stvarno pozitivnu stopu. Ona se dobiva tako što se za različite pragove određuju osjetljivost i lažno pozitivna stopa te sve te točke povezane čine krivulju kao primjerice na grafu prikazanom na slici 5.1.:



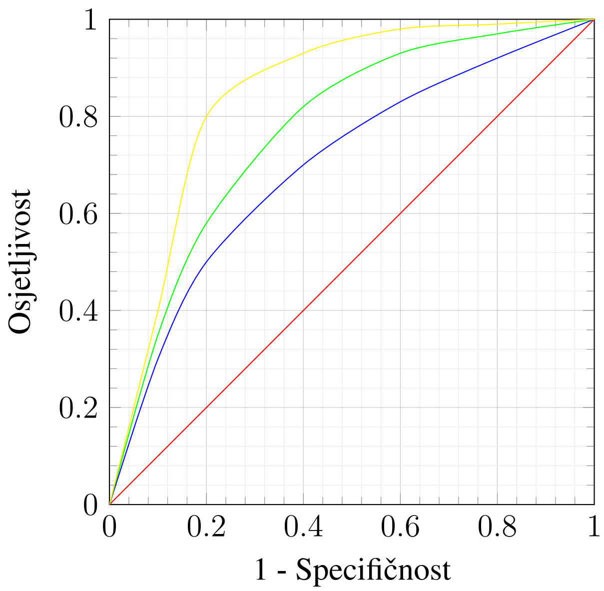
Slika 5.1. ROC krivulja

Linija koja odgovara testu koji sasvim slučajno može biti pozitivan ili negativan naziva se bazična linija. To je linija koja povezuje točke u kojima je osjetljivost jednaka 1-specifičnost i prikazana je na grafu na slici 5.2.:



Slika 5.2. Bazična linija

Pomoću ROC krivulje moguće je odrediti koliko je test kvalitetan i točan. Što je ROC krivulja strmija u početku odnosno bliža gornjem lijevom kutu to je test precizniji jer je samim time osjetljivost jednaka ili blizu 1 kao i specifičnost, a lažno pozitivna stopa je jednaka ili blizu 0 što znači da bi većina rezultata testa i pozitivnih i negativnih bili precizni. Tako je moguće usporediti različite testove prikazom pripadnih ROC krivulja na istom grafu (slika 5.3). Na tom grafu nalaze se ROC krivulje za 3 testa.



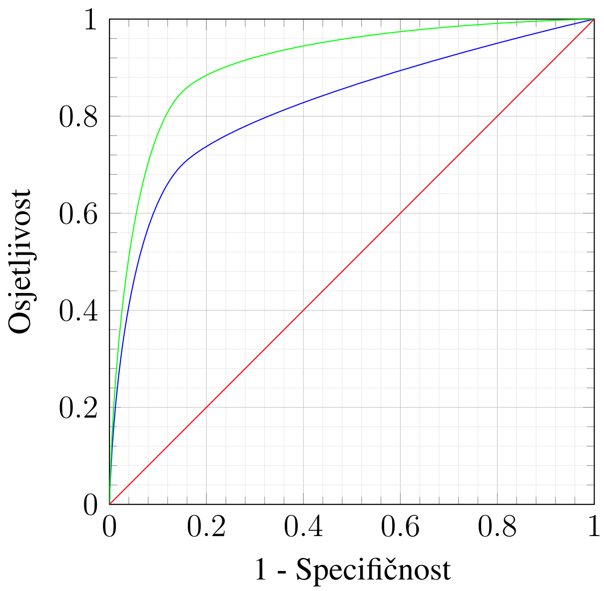
Slika 5.3. ROC krivulje različitih karakteristika

Iz grafa je vidljivo da se sve 3 krivulje nalaze iznad bazične linije što znači da se sva 3 testa mogu uzeti u obzir kao precizni. Najstrmija krivulja žuta što bi značilo da test sa žutom krivuljom ima najbolje karakteristike odnosno visoku osjetljivost i specifičnost te je najtočniji, zatim bi se nalazio test sa zelenom krivuljom, a kao najmanje precizan od ova 3 testa bi bio test sa plavom karakteristikom s obzirom na najniže vrijednosti osjetljivosti i specifičnosti.

5.3. AUC vrijednost

AUC vrijednost (*Area Under the Curve*) je numerička mjera za određivanje učinkovitosti i procjenu kvalitete nekog modela ili testa. Kao što samo ime kaže to je površina ispod ROC krivulje odnosno površina između ROC krivulje i bazične linije koja povezuje točku (0,0) u kojoj je osjetljivost jednaka 0, a specifičnost 1 i točku (1,1) u kojoj je osjetljivost jednaka 1, a specifičnost 0. [15]

AUC vrijednosti se kreću u intervalu od 0 do 1. Ukoliko je AUC vrijednost blizu ili jednaka 1 to znači da su karakteristike testa izvrsne te da je test ispravno prepoznao sve pozitivne i sve negativne primjere. Ako je AUC vrijednost blizu ili jednaka 0.5 to znači da test nema dobre karakteristike te da nije prepoznao pozitivne i negativne primjere već ih bira nasumično. Ako je AUC vrijednost manja od 0.5 i teži 0 to znači da test ne samo da ne razlikuje pozitivne od negativnih primjera već ih često zamijenjuje, one koji su pozitivni klasificira kao negativne, dok one koji su negativni klasificira kao pozitivne. Dakle što je veća AUC vrijednost to je veća sposobnost testa da razlikuje pozitivne od negativnih primjera.



Slika 5.4. ROC krivulje

Na grafu na slici 5.4. je vidljiv primjer 2 testa sa ROC krivuljama. Pomoću površine ispod krivulje odnosno AUC vrijednosti moguće je zaključiti kako će test sa zelenom karakteristikom imati veću učinkovitost pri raspoznavanju pozitivnih i negativnih primjera od testa sa plavom karakteristikom zato što je područje ispod zelene karakteristike veće nego ono ispod plave.

6. Određivanje prijelomnih vrijednosti

6.1. Youdenov indeks

Pri analizi određenog testa ili klasifikacijskog modela koriste se mnogi parametri. Jedan od značajnijih parametara za utvrđivanje kvalitete testa je svakako Youdenov indeks. Youdenov indeks je vrlo bitan parametar jer se opisuje pomoću osjetljivosti i specifičnosti testa. Vrijednosti Youdenovog indeksa tako se kreću od -1 do 1. Što je vrijednost Youdenovog indeksa manja to znači da su osjetljivost i specifičnost testa manji odnosno lošiji, dok vrijednosti Youdenovog indeksa koje se kreću bliže broju 1 označavaju veću osjetljivost i specifičnost testa te samim time kvalitetniji test. Youdenov indeks računa se prema sljedećem izrazu:

$$Yi = OS + SP - 1, \tag{6.1}$$

gdje je OS oznaka za osjetljivost, a SP oznaka za specifičnost te je Yi oznaka za Youdenov indeks.

Za svaki test postoje takve vrijednosti osjetljivosti i specifičnosti da Youdenov indeks testa postiže najveću moguću vrijednost. Kada je vrijednost Youdenovog indeksa određenog testa najveća tada je moguće odrediti prijelomnu vrijednost testa. Prijelomna vrijednost testa je kriterij testa koji određuje klasifikacijsku vrstu, primjerice da se student smatra uspješnim ukoliko je položio barem 3 predmeta tokom semestra.

6.2. Primjer određivanja prijelomne vrijednosti

Primjer 6.1. *Kolokviju iz kolegija Inženjerska metematika ET pristupilo je 35 studenata. Nakon kolokvija i objavljenih rezultata testa provedena je anonimna anketa gdje su studenti bili pitani da li im se kolegij Inženjerska metematika ET sviđa ili ne sviđa. Rezultati ankete su bili sljedeći: 28 studenata je izjavilo kako im se kolegij sviđa, dok je 7 studenata izjavilo kako im se kolegij ne sviđa. Uz pomoć rezultata kolokvija i anonimne ankete potrebno je odrediti koliki je broj bodova na kolokviju potreban kako bi se studentu sviđao kolegij Inženjerska metematika ET.*

Traženi broj bodova na kolokviju označava prijelomnu vrijednost odnosno svi studenti koji ostvare manje od traženog broja bodova će vrlo vjerojatno izjaviti kako im se kolegij ne sviđa, dok oni koji ostvare više će vrlo vjerojatno izjaviti kako im se kolegij sviđa. Kako bi se odredila prijelomna vrijednost potrebno je za različite brojeve bodova odrediti osjetljivost i specifičnost te pomoću njih Youdenov indeks. Programski su dobivene vrijednosti specifičnosti i osjetljivosti za sve bodovne pragove. U nastavku će se izračunati vrijednosti Youdenovog indeksa za par slučajeva.

Tako primjerice za 11.4 boda osjetljivost iznosi 92.86%, dok specifičnost iznosi 14.29%. Youdenov indeks za navedene vrijednosti iznosi:

$$Yi = OS + SP - 1 = 0.9286 + 0.1429 - 1 = 0.0715. \tag{6.2}$$

Za navedeni prag postignut je Youdenov indeks iznosa 0.0715. Kako bi se Youdenov indeks povećao potrebno je smanjiti osjetljivost odnosno povećati specifičnost što će se postići dizanjem bodovnog praga.

Za bodovni prag od 13.5 bodova osjetljivost iznosi 82.14%, dok specifičnost iznosi 28.57%, a Youdenov indeks iznosi:

$$Yi = OS + SP - 1 = 0.8214 + 0.2857 - 1 = 0.1071. \tag{6.3}$$

Povećanjem bodovnog praga se uistinu povisila vrijednost specifičnost te Youdenov indeks iznosi 0.1071.

Ukoliko je bodovni prag 18.7 bodova osjetljivost iznosi 71.43%, a specifičnost 42.86%. Youdenov indeks za prag od 18.7 bodova iznosi:

$$Yi = OS + SP - 1 = 0.7143 + 0.4286 - 1 = 0.1429. \tag{6.4}$$

Ako bodovni prag iznosi 20.65, osjetljivost će iznositi 67.86%, dok će specifičnost biti 71.43%. Youdenov indeks će iznositi:

$$Yi = OS + SP - 1 = 0.6786 + 0.7143 - 1 = 0.3929. \tag{6.5}$$

Youdenov indeks za prag od 20.65 bodova iznosi 0.3929, što je dosad ujedno i najveća vrijednost Youdenovog indeksa.

Povećanjem bodovnog praga na 23.3 boda osjetljivost iznosi 50%, dok specifičnost 71.43%. Za navedeni slučaj Youdenov indeks iznosi:

$$Yi = OS + SP - 1 = 0.5 + 0.7143 - 1 = 0.2143. \tag{6.6}$$

Daljnjim povećanjem praga na primjerice 29.5, osjetljivost iznosi 28.57%, a specifičnost 85.71%. Youdenov indeks jednak je:

$$Yi = OS + SP - 1 = 0.2857 + 0.8571 - 1 = 0.1428. \tag{6.7}$$

Može se zaključiti da će se za bodovni prag veći od 20.65 Youdenov indeks početi smanjivati kao i za bodovne pragove manje od 20.65. Dakle najveći Youdenov indeks iznosi 0.3929 odnosno prijelomna vrijednost, broj bodova na kolokviju, da bi se studentima svidio kolegij Inženjerska metematika ET iznosi 20.65, što je naznačeno i u sljedećoj tablici (Tablica 6.1.).

Tablica 6.1. Prijelomna vrijednost

BODOVNI PRAG	OSJETLJIVOST	SPECIFIČNOST	YOUDENOV INDEKS
11.4	0.9286	0.1429	0.0715
13.5	0.8214	0.2857	0.1071
18.7	0.7143	0.4286	0.1429
20.65	0.6786	0.7143	0.3929
23.3	0.5	0.7143	0.2143
29.5	0.2857	0.8571	0.1428

7. Primjena ROC analize u umjetnoj inteligenciji

7.1. Matrica konfuzije

U prethodnim poglavljima već je objašnjena i prikazana široka primjena ROC analize te kolika je njena korist. ROC analiza tako ima veliku važnost u današnjem svijetu od analize dijagnostičkog testa gdje uz procjenu performansi samog testa pomaže i pri donošenju pravilne dijagnoze pa do klasifikacijskih modela poput onog koji ima zadatak razlikovati spam e-mailove od legitimnih e-mailova. ROC analiza se također može primijeniti i u umjetnoj inteligenciji (*Artificial Intelligence*) kako bi se pomoću testova mogle predvidjeti određene situacije.

U svim prethodnim primjerima koristila se ROC analiza pomoću tablica kontingencije u kojima su se uspoređivali rezultati određenih testova i modela sa stvarnim stanjima te su se na temelju tih podataka određivali parametri i kvaliteta testova. ROC analiza se osim pomoću tablica kontingencije može provesti i pomoću takozvanih matrica konfuzije.

Matrica konfuzije je tablica koja se koristi kako bi se mogle precizno odrediti vrijednosti i karakteristike klasifikacijskog modela. Matrica konfuzije je tablica dimenzija 2×2, može biti i drugačijih dimenzija, ali se onda korištenjem svojstva matrice svaki od ishoda može svesti na dimenzije 2×2. Primjerice ako model podatke može klasificirati u 3 različite klase, onda će se od početne matrice konfuzije dimenzija 3×3 dobiti 3 odvojene matrice dimenzija 2×2. Matrica konfuzije se sastoji od stvarnih vrijednosti i predviđenih vrijednosti odnosno u tablicu se unose vrijednosti varijabli: TP koja predstavlja broj ispravno prepoznatih primjera kao pozitivnih, FP koja predstavlja broj neispravno prepoznatih primjera kao pozitivnih odnosno broj lažno pozitivnih, TN koja predstavlja broj ispravno prepoznatih primjera kao negativnih i FN koja predstavlja broj neispravno prepoznatih primjera kao negativnih odnosno broj lažno negativnih. Tablica 7.1. predstavlja shematski prikaz matrice konfuzije.

Tablica 7.1. Matrica konfuzije

	Pozitivan	Negativan	
Predviđen pozitivan	TP	FP	$TP + FP$
Predviđen negativan	FN	TN	$FN + TN$
	$TP + FN$	$FP + TN$	

Pomoću matrice konfuzije mogu se odrediti mnogi parametri koji opisuju karakteristike klasifikacijskog modela. Iz opisa i Tablice 7.1. jasno je vidljivo da su tablica kontigencije opisana u 4. poglavlju i matrica konfuzije vrlo bliski koncepti. Većina parametara dobivenih iz matrice konfuzije imaju svoje ekvivalente u parametrima dobivenim pomoću tablice kontingencije.

Tako se pomoću matrice konfuzije može dobiti osjetljivost, koja se još naziva i odzivom. Odziv se računa kao i osjetljivost preko tablice kontingencije prema formuli (4.3) te se označava sa RC. Pomoću odziva se može dobiti vrijednost stope lažno negativnih (*False Negative Rate*) primjera. Stopa lažno negativnih primjera je jednaka proporciji lažno negativnih rezultata testa te se računa prema formuli (4.4), označava se sa FNR.

Specifičnost dobivena preko matrica konfuzije je jednaka onoj dobivenoj preko tablica kontingencije, računa se prema formuli (4.5) i označava sa SP. Pomoću specifičnosti može se izračunati stopa lažno pozitivnih (*False Positive Rate*) primjera. Stopa lažno pozitivnih primjera je jednaka proporciji lažno pozitivnih rezultata testa, računa se prema formuli (4.6) i označava sa FPR.

Kao i kod tablica kontingencije i kod matrice konfuzije postoje prediktivne vrijednosti. Tako je preciznost jednaka pozitivnoj prediktivnoj vrijednosti dobivenoj pomoću tablice kontingencije i računa se prema formuli (4.12), označava se sa PR. Negativna prediktivna vrijednost dobivena pomoću matrice konfuzije jednaka je onoj dobivenoj pomoću tablica kontingencije i računa se prema formuli (4.13) te se označava sa NPV.

Točnost dobivena preko matrice konfuzije je jednaka preciznosti dobivenoj pomoću tablice kontingencije pa se tako računa prema formuli (4.1), ali se označava sa AC. Stopa pogreške je ekvivalentna onoj dobivenoj preko tablica kontingencije pa se tako računa prem aformuli (4.2) i označava sa ER. Pri ROC analizi u umjetnoj inteligenciji će se još koristiti i Youdenov indeks koji je objašnjen u prethodno poglavlju i računa se prema formuli (6.1), a označava se sa Yi.

Neki od važnijih parametara pri određivanju kvalitete modela koji nisu prethodno navedeni su: $F1$-mjera, MCC (*Matthews Correlation Coefficient*) te balansirana točnost.

$F1$- mjera je harmonijska sredina između preciznosti i osjetljivosti i računa se prema sljedećem izrazu:

$$F1 = \frac{2 \cdot PR \cdot RC}{PR + RC}, \tag{7.1}$$

gdje je $F1$ oznaka za $F1$-mjeru.

MCC (*Matthews Correlation Coefficient*) je parametar koji u obzir uzima sve vrijednosti iz matrice konfuzije te ocjenjuje cjelokupne karakteristike modela. Vrijednosti koeficijenta se kreću od -1 do 1. Ukoliko je vrijednost koeficijenta -1 to znači da model ima jako loše karakteristike. Ukoliko je vrijednost koeficijenta jednaka 1 to znači da model ima izvrsne karakteristike, dok za vrijednost koeficijenta 0 model radi praktički na sistemu slučajnog odabira. Definiran je izrazom:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}, \tag{7.2}$$

gdje je MCC oznaka za MCC koeficijent.

Srednja vrijednost osjetljivosti i specifičnosti naziva se balansirana točnost i jednaka je izrazu:

$$BAC = \frac{RC + SP}{2}, \tag{7.3}$$

gdje je BAC oznaka za balansiranu točnost. [16,17,18]

7.2. Primjer ROC analize u umjetnoj inteligenciji

Primjer 7.1. *Postotak studenata koji prođu odnosno padnu određeni kolegij jedno je od glavnih mjerila u današnjem obrazovnom sustavu. Svakako što je broj studenata koji su pali kolegij manji to je sustav učinkovitiji. Može se pretpostaviti kako bi neki od studenata koji imaju veliku mogućnost da padnu kolegij uz određenu pravovremenu intervenciju, primjerice u obliku individualnog rada, na kraju ipak savladali kolegij prije nego što uspiju doprinijeti broju onih koji su pali kolegij. Da bi takve situacije bile moguće, potrebno je na vrijeme prepoznati takve studente. Tako je osmišljen umjetna neuronska mreža koja će na temelju nekih ulaznih vrijednosti pokušati što preciznije predviditi koji studenti će uspješno položiti kolegij Matematika 2, a kojima prijeti pad. Cilj modela je da se uz ranu prognozu uspješnosti studenata uspije prepoznati skupina studenata koji su u opasnosti od pada kolegija. Navedeni model se pri procjeni i dobivenim prediktivnim vrijednostima služi podacima kao što su: aktivnosti pojedinca na kolegiju i kolegiju prethodniku odnosno Matematika 1, uspjehu pojedinca tokom srednjoškolskog obrazovanja, uspjehu na maturi i drugima te su dobiveni sljedeći rezultati: [19]*

Tablica 7.2. Matrica konfuzije umjetne neuronske mreže

	Prolaz	Pad	
Predviđen prolaz	22	2	24
Predviđen pad	1	13	14
	23	15	

U prethodnoj matrici konfuzije nalaze se dobiveni rezultati umjetne neuronske mreže. Tako je moguće iščitati vrijednosti varijabli TP, TN, FP, FN. Broj ispravno predviđenih studenata koji su prošli kolegij odnosno TP jednak je 22. Broj neispravno predviđenih studenata koji su prošli kolegij odnosno oni koji su pali, FP, iznosi 2. Broj ispravno predviđenih studenata koji su pali kolegij odnosno TN iznosi 13, dok broj neispravno predviđenih studenata koji su pali kolegij odnosno oni koji su prošli, FN, iznosi 1. Pomoću navedenih podataka iz matrice konfuzije zadane mreže moguće je odrediti parametre i ocijeniti karakteristike mreže.

Pri analizi mreže prvo je potrebno odrediti osjetljivost odnosno odziv i specifičnost mreže. Osjetljivost zadane umjetne neuronske mreže računa se prema formuli (4.3) i iznosi:

$$RC = \frac{TP}{TP + FN} = \frac{22}{23} = 0.957. \tag{7.4}$$

Uz poznatu vrijednost osjetljivosti moguće je i odrediti stopu lažno negativnih prema formuli (4.4):

$$FNR = \frac{FN}{TP + FN} = 1 - RC = 1 - 0.957 = 0.043. \tag{7.5}$$

Specifičnost umjetne neuronske mreže računa se prema formuli (4.5) i iznosi:

$$SP = \frac{TN}{FP + TN} = \frac{13}{15} = 0.867. \tag{7.6}$$

Uz poznatu vrijednost specifičnosti mreže moguće je odrediti i stopu lažno pozitivnih prema formuli (4.6):

$$FPR = \frac{FP}{FP + TN} = 1 - SP = 1 - 0.867 = 0.133. \tag{7.7}$$

Iz prethodno izračunatih parametara može se zaključiti kako je umjetna neuronska mreža precizno predvidjela prolaz studenata u 95.7% slučajeva te je precizno predvidjela pad studenata u 86.7% slučajeva.

Sada je potrebno odrediti prediktivne vrijednosti umjetne neuronske mreže. Preciznost mreže računa se prema formuli (4.12) i jednaka je:

$$PR = \frac{TP}{TP + FP} = \frac{22}{24} = 0.917, \tag{7.8}$$

dok se negativna prediktivna vrijednost NPV izračunava prema formuli (4.13) i iznosi:

$$NPV = \frac{TN}{TN + FN} = \frac{13}{14} = 0.929. \tag{7.9}$$

Dakle uz poznate prediktivne vrijednosti moguće je zaključiti kako je umjetna neuronska mreža imala u 91.7% slučajeva točne pozitivne predikcije, prolaz studenta, i u 92.9% slučajeva točne negativne predikcije odnosno pad studenta.

Pomoću parametara preciznosti i osjetljivosti moguće je odrediti $F1$-mjeru mreže prema formuli (7.1):

$$F1 = \frac{2 \cdot PR \cdot RC}{PR + RC} = \frac{2 \cdot 0.917 \cdot 0.957}{0.917 + 0.957} = 0.937. \tag{7.10}$$

Koristeći sve vrijednosti iz navedene matrice konfuzije umjetne neuronske mreže, pomoću formule (7.2), moguće je odrediti MCC koeficijent koji iznosi:

$$\begin{aligned} MCC &= \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \\ &= \frac{22 \cdot 13 - 2 \cdot 1}{\sqrt{(22 + 2) \cdot (22 + 1) \cdot (13 + 2) \cdot (13 + 1)}} = 0.834. \end{aligned} \tag{7.11}$$

Dakle MCC umjetne neuronske mreže iznosi 0.834 što označava da mreža ima izvrsne karakteristike.

Također još jedan od parametara koji koristi sve vrijednosti matrice konfuzije je točnost, koja se računa prema formuli (4.1) i iznosi:

$$AC = \frac{TP + TN}{TP + FP + TN + FN} = \frac{22 + 13}{22 + 2 + 13 + 1} = 0.921. \tag{7.12}$$

Točnost umjetne neuronske mreže iznosi 92.1%. Točnost je poput preciznosti i negativne prediktivne vrijednosti NPV prediktivna vrijednost pa tako ovaj podatak govori kako su u 92.1% slučajeva predikcije umjetne neuronske mreže bile ispravne.

Suprotno točnosti nalazi se stopa pogreške koja se računa prema formuli (4.2):

$$ER = \frac{FP + FN}{TP + FP + TN + FN} = \frac{2 + 1}{22 + 2 + 13 + 1} = 0.079, \tag{7.13}$$

odnosno umjetna neuronska mreža ima krive predikcije u 7.9% slučajeva.

Balansirana točnost mreže se računa prema formuli (7.3) i iznosi:

$$BAC = \frac{RC + SP}{2} = \frac{0.957 + 0.867}{2} = 0.911. \tag{7.14}$$

Za kraj je potrebno odrediti iznos Youdenovog indeksa umjetne neuronske mreže prema formuli (6.1):

$$Yi = RC + SP - 1 = 0.957 + 0.867 - 1 = 0.821. \tag{7.15}$$

Tako Youdenov indeks umjetne neuronske mreže iznosi 0.821 što još jednom kao i MCC koeficijent potvrđuje njene izvrsne karakteristike.

8. Zaključak

Teorija vjerojatnosti pojam je koji se proteže daleko kroz povijest. Tako su se kroz godine razvijale razne definicije vjerojatnosti. Od samih početaka i Laplacea koji je opisao klasičnu definiciju vjerojatnosti i geometrijsku definiciju vjerojatnosti pa do Kolmogorova i aksiomatske definicije vjerojatnosti. Sa vremenom je također objašnjena i uvjetna vjerojatnost te matematički izrazi poput formule potpune vjerojatnosti i Bayesove formule.

Upravo je ta uvjetna vjerojatnost poslužila kao osnova za razvitak ROC analize. ROC (*Receiver Operating Characteristic*) analiza je statistička metoda koja služi za procjenu kvalitete određenog testa ili klasifikacijskog modela. Razvoj ove metode krenuo je za vrijeme Drugog svjetskog rata pri analizi radara i detekcije signala. Tokom godina ROC analiza se krenula koristiti i u raznim drugim granama poput psihofizike i medicine. Danas je primjena ROC analize široko rasprostranjena od psihologije, medicine pa do računarstva, elektrotehnike, ekonomije i financija. Širina uporabe ROC analize vidljiva je i u ovom radu gdje su pomoću tablica kontingencije određene kvalitete BAT, brzog antigenskog testa za COVID-19 kao i karakterstike klasifikacijskog modela sa zadaćom prepoznavanja spam e-mailova.

S napretkom tehnologije i računala došlo je i do digitalizacije ROC analize. Tako je omogućen grafički prikaz ROC analize pomoću ROC krivulje i AUC vrijednosti. ROC krivulja je grafički prikaz analize koji omogućava prikaz potpunih performansi sustava. Uz to prikazom više ROC krivulja moguće je odrediti koji je model precizniji izračunom AUC vrijednosti odnosno izračunom površine ispod krivulje.

Jedan od važnijih parametara pri ROC analizi je svakako i Youdenov indeks pomoću kojeg je moguće odrediti prijelomnu vrijednost modela odnosno granicu koja može određivati klasifikacijsku vrstu primjera kao što je prikazano i u ovom radu na primjeru koliki je potreban broj bodova na kolokviju da bi se studentu svidio kolegij.

ROC analiza se može primijeniti i u umjetnoj inteligenciji kao što je navedeno u radu gdje je napravljena analiza klasifikacijskog modela pomoću matrica konfuzije čiji je zadatak prepoznati studente s velikom šansom da padnu kolegiji u cilju sprječavanja takvog ishoda. Sve u svemu ROC analiza ima široki spektar te je jako korisna metoda za vršenje analize testova i raznih modela.

Literatura

[1] "Blaise Pascal and Pierre de Fermat", s Interneta, https://www.linkedin.com/pulse/predicting-blaise-pascal-pierre-de-fermat-harmanjit-singh-bhogal, 26. srpnja 2019.

[2] "Pierre-Simon Laplace", s Interneta, https://hr.wikipedia.org/wiki/Pierre-Simon Laplace, 15. lipnja 2023.

[3] Proleksis Enciklopedija: "Kolmogorov, Andrej Nikolajevič", s Interneta, https://proleksis.lzmk.hr/31833/, 5. veljače 2015.

[4] "Povijest vjerojatnosti - što je to, definicija i pojam", s Interneta, https://hr.economy-pedia.com/11038374-history-of-probability, 2023.

[5] Dražić, I.: "Klasična i geometrijska definicija vjerojatnosti - interna skripta iz kolegija Numerička i stohastička matematika ", Sveučilište u Rijeci, Tehnički fakultet, 2023.

[6] "Poglavlje 2: KLASIČNA DEFINICIJA VJEROJATNOSTI", s Interneta, http://www.grad.hr/vera/webnastava/vjerojatnostistatistika/html/VISch2.html, 7. srpnja 2023.

[7] Dražić, I.: "Aksiomatska definicija vjerojatnosti - interna skripta iz kolegija Numerička i stohastička matematika", Sveučilište u Rijeci, Tehnički fakultet, 2023.

[8] "Poglavlje 3: AKSIOMATSKA DEFINICIJA VJEROJATNOSTI", s Interneta, http://www.grad.hr/vera/webnastava/vjerojatnostistatistika/html/VISch3.html, 7. srpnja 2023.

[9] Dražić, I.: "Uvjetna vjerojatnost i primjena na analizu dijagnostiˇckog testa - interna skripta iz kolegija Numerička i stohastička matematika", Sveučilište u Rijeci, Tehnički fakultet, 2023.

[10] Lubura Strunjak, S.; Vondraček, Z.: "Poglavlje 2: Uvjetna vjerojatnost i nezavisnost - interna skripta iz kolegija Vjerojatnost", Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Zagreb, 12. listopada 2020.

[11] "Poglavlje 4: UVJETNA VJEROJATNOST", s Interneta, http://www.grad.hr/vera/webnastava/vjerojatnostistatistika/html/VISch4.html, 7. srpnja 2023.

[12] Dražić, I.: "Formula potpune vjerojatnosti i Bayesova formula - interna skripta iz kolegija Numerička i stohastička matematika", Sveučilište u Rijeci, Tehnički fakultet, 2023.

[13] Roy D. Yates, David J. Goodman: "Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers", Wiley, 2nd edition, 2004.

[14] Božikov, J.: "Valjanost (dijagnostičkog) testa ROC analiza - interna skripta iz kolegija Osobitosti kliničkih medicinskih istraživanja", Sveučilište u Zagrebu, Medicinski fakultet, Zagreb, 4. ožujka 2008.

[15] National Library of Medicine: "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation", 14. srpnja 2023.

[16] "Matthews's correlation coefficient: Definition, Formula and advantages", s Interneta, https://www.voxco.com/blog/matthewss-correlation-coefficient-definition-formula-and-advantages/, 9. prosinca 2021.

[17] Đuričić, T.; Merćep, A.: "Strojno učenje: Bilješke s predavanja", Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, 2016.

[18] "Youden's J statistic", s Interneta, https://en.wikipedia.org/wiki/Youden%27s_J_statistic, 2023.

[19] Čotić Poturić, V.; Bašić-Šiško, A.; Lulić, I.: "Artificial neural network model for forecasting student failure in math course", ICERI2022 Proceedings, pp. 5872-5878, 2022.

Sažetak i ključne riječi

U ovom radu objašnjene su definicije vjerojatnosti, počevši od klasične i geometrijske definicije vjerojatnosti pa do aksiomatske definicije vjerojatnosti. Obrađena je i uvjetna vjerojatnost u čijem kontekstu se opisuju matematički izrazi poput formule potpune vjerojatnosti i Bayesove formule. Rad se pretežito bavi ROC analizom i njenom primjenom pa je tako objašnjena analiza testa pomoću tablice kontingencije i opisani su parametri potrebni za analizu testa. ROC analiza se nastoji grafički prikazati preko ROC krivulje i AUC vrijednosti. Uvodi se pojam prijelomne vrijednosti koja se određuje preko Youdenovog indeksa te je prikazana primjena ROC analize u umjetnoj inteligenciji pomoću matrica konfuzije.

Ključne riječi: klasična definicija vjerojatnosti, geometrijska definicija vjerojatnosti, aksiomatska definicija vjerojatnosti, uvjetna vjerojatnost, ROC analiza, tablica kontingencije, ROC krivulja, AUC vrijednost, Youdenov indeks, matrica konfuzije

Summary and key words

This paper explains the definitions of probability, starting with the classical and geometric definition of probability to the axiomatic definition of probability. The concept of conditional probability is also covered, which allowed us to introduce the law of total probability and Bayes' formula. The paper is mainly concerned with ROC analysis and its applications. In particular, the analysis of the test using the contingency table and the corresponding parameters is explained in detail and with examples. The graphical representation of the ROC analysis, namely the ROC curve and the AUC value, are also explained. The concept of breaking value is introduced which is determined by Youden's index, and the application of ROC analysis in artificial intelligence with the help of confusion matrices is presented.

Keywords: classical definition of probability, geometric definition of probability, axiomatic definition of probability, conditional probability, ROC analysis, contingency table, ROC curve, AUC value, Youden's index, confusion matrix