# Primjena strojnog učenja u analizi i procjeni toksičnosti kemijskih spojeva

**Buljeta, Luka**

SVEUČILIŠTE U ZAGREBU
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

ZAVRŠNI RAD br. 1604

# PRIMJENA STROJNOG UČENJA U ANALIZI I PROCJENI TOKSIČNOSTI KEMIJSKIH SPOJEVA

Luka Buljeta

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

ZAVRŠNI RAD br. 1604

# PRIMJENA STROJNOG UČENJA U ANALIZI I PROCJENI TOKSIČNOSTI KEMIJSKIH SPOJEVA

Luka Buljeta

Zagreb, lipanj 2024.

**SVEUČILIŠTE U ZAGREBU**
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

Zagreb, 4. ožujka 2024.

# ZAVRŠNI ZADATAK br. 1604

| | |
|---|---|
| Pristupnik: | **Luka Buljeta (0036539861)** |
| Studij: | Elektrotehnika i informacijska tehnologija i Računarstvo |
| Modul: | Računarstvo |
| Mentorica: | izv. prof. dr. sc. Mihaela Vranić |

Zadatak: **Primjena strojnog učenja u analizi i procjeni toksičnosti kemijskih spojeva**

Opis zadatka:

Toksičnost kemijskih spojeva predstavlja važan aspekt u različitim industrijama, uključujući farmaceutsku, kozmetičku, prehrambenu i poljoprivrednu industriju. Tradicionalne metode procjene toksičnosti često su skupe i vremenski zahtjevne te se stoga sve više istražuje primjena strojnog učenja za predviđanje toksičnosti kemijskih spojeva. Vaš je zadatak proučiti recentnu literaturu za ovo područje, identificirati ključne značajke povezane s toksičnošću i dati pregled metoda strojnog učenja opisanih u literaturi za ovo specifično područje. Za odabrani skup podataka, potrebno je razviti modele strojnog učenja, usporediti njihove performanse i odabrati najprikladniji model za predviđanje toksičnosti.

Rok za predaju rada: 14. lipnja 2024.

**Sadržaj**

# Uvod

U suvremenom svijetu, kemijski spojevi igraju ključnu ulogu u raznim industrijama, uključujući farmaceutsku, poljoprivrednu, kozmetičku i prehrambenu industriju. S obzirom na njihovu sveprisutnost, procjena toksičnosti kemijskih spojeva postala je od iznimne važnosti za zaštitu ljudskog zdravlja i okoliša. Tradicionalne metode procjene toksičnosti, koje često uključuju laboratorijske eksperimente na životinjama, mogu biti dugotrajne, skupe i etički sporne. Razvojem novih tehnologija i pristupa, strojno učenje nudi potencijalno brža, ekonomičnija i humanija rješenja za ovaj problem.

Strojno učenje, kao grana umjetne inteligencije, omogućuje računalima učenje iz podataka i donošenje odluka bez eksplicitnog programiranja. U kontekstu analize i procjene toksičnosti kemijskih spojeva, strojno učenje može koristiti podatke o poznatim kemijskim spojevima i njihovim toksičnim svojstvima za izgradnju modela koji predviđaju toksičnost novih spojeva. Ovi modeli mogu pomoći znanstvenicima i regulatornim tijelima donošenje informiranih odluka o sigurnosti kemijskih tvari prije nego što one dospiju na tržište ili u okoliš.

Primjena strojnog učenja u ovom području uključuje različite tehnike i pristupe, kao što su regresijski modeli, klasifikacijski algoritmi, neuronske mreže, i duboko učenje. Ove tehnike omogućuju analizu kompleksnih odnosa između kemijske strukture spojeva i njihovih bioloških učinaka, identificirajući ključne karakteristike koje utječu na toksičnost. Uz to, razvoj velikih baza podataka o kemijskim spojevima i njihova dostupnost znanstvenoj zajednici dodatno potiču istraživanje i primjenu strojnog učenja u procjeni toksičnosti.

Ovaj završni rad istražuje potencijal i izazove korištenja strojnog učenja za analizu i procjenu toksičnosti kemijskih spojeva. Razmatraju se različite metode i algoritmi, njihove prednosti i ograničenja, te se predstavljaju studije slučaja koje ilustriraju praktične primjene ovih tehnologija. Cilj rada je pružiti sveobuhvatan pregled trenutnog stanja na ovom području i ukazati na smjerove za buduća istraživanja i razvoj.

U daljnjim poglavljima bit će detaljno obrađeni osnovni pojmovi strojnog učenja, metode za procjenu toksičnosti, te konkretni primjeri primjene strojnog učenja u

analizi kemijskih spojeva. Na kraju, raspravit će se rezultati, izazovi i perspektive ove inovativne metode u kontekstu zaštite zdravlja i okoliša.

# 1. Korištenje strojnog učenja za analizu i procjenu toksičnosti kemijskih spojeva

## 1.1 Kemoinformatika

Kemoinformatika, kao interdisciplinarna znanost, predstavlja spoj kemije, računalnih znanosti i informacijske tehnologije. Njezin glavni cilj je korištenje računarskih metoda za prikupljanje, pohranu, analizu i interpretaciju kemijskih podataka. Razvoj kemoinformatike omogućio je znanstvenicima upravljanje ogromnim količinama podataka generiranih eksperimentalnim radom te izvlačenje korisnih informacija iz tih podataka, što potiče napredak u kemiji, farmakologiji i biotehnologiji. Ključnu ulogu u kemoinformatici ima modeliranje i predikcija svojstava molekula. Kvantitativni odnosi struktura-aktivnost (QSAR quantitative/qualitative structure activity relationships) modeli koriste matematičke i statističke metode za povezivanje kemijske strukture spoja s njegovim quantitative/qualitative structure activity relationshipsbiološkim ili kemijskim svojstvima. Ovi modeli mogu predvidjeti aktivnost novih spojeva temeljenih na poznatim podacima, što znatno smanjuje potrebu za skupim i dugotrajnim eksperimentalnim ispitivanjima. Osim QSAR modela, metode strojnog učenja, poput neuronskih mreža, potpornih vektorskih strojeva (SVM) i metoda najbližeg susjeda, postaju sve važnije u predikciji kemijskih svojstava i dizajnu novih molekula [1][2].

## 1.2 Toksikologija

Toksikologija se bavi identifikacijom, karakterizacijom i kvantifikacijom toksičnih tvari te razumijevanjem mehanizama njihova djelovanja. Tradicionalno, procjena toksičnosti kemijskih spojeva provodila se putem in vitro (laboratorijskih) i in vivo (na živim organizmima) eksperimenata. Takvi pristupi, iako temeljiti i pouzdani, često su skupi, dugotrajni i etički sporni, osobito kada uključuju testiranje na životinjama. S razvojem računalnih metoda, otvorila se nova era u toksikologiji koja uključuje primjenu strojnog učenja za analizu i predikciju toksičnosti kemijskih spojeva. Strojno učenje omogućava predviđanje kako će određene molekule, bazirano na njihovoj strukturi i lako mjerljivim svojstvima, utjecati na procese unutar organizama. Ovaj pristup može značajno smanjiti potrebu za laboratorijskim testiranjima te pružiti brze i točne procjene toksičnosti [3][4].

## 1.3 Strojno Učenje

Strojno učenje je grana umjetne inteligencije koja omogućava računalima učenje iz podataka i donošenje odluka bez eksplicitnog programiranja. Korištenjem algoritama, strojevi mogu analizirati ogromne količine podataka, prepoznati obrasce i generirati predviđanja. To čini strojno učenje izuzetno korisnim u različitim područjima, uključujući kemoinformatiku i toksikologiju. U kemoinformatici, strojno učenje može pomoći u predikciji kemijskih svojstava i dizajnu novih molekula, dok u toksikologiji može predvidjeti toksičnost spojeva na temelju njihove kemijske strukture. Strojno učenje koristi različite vrste algoritama, uključujući nadzirano učenje, nenadzirano učenje i učenje s pojačanjem. Nadzirano učenje uključuje treniranje modela na označenim podacima, gdje algoritam uči povezivati ulazne podatke s odgovarajućim izlazima, što je korisno za klasifikaciju i regresiju. Nenadzirano učenje analizira neoznačene podatke kako bi pronašao skrivene obrasce ili strukture, što je korisno za grupiranje i smanjenje dimenzionalnosti. Učenje s pojačanjem uključuje treniranje modela putem interakcije s okolinom, gdje algoritam uči donositi odluke koje maksimiziraju dugoročne nagrade [5][6][7].

## 1.4 Projekt Tox21

Projekt Tox21 (Toxicology in the 21st Century) je inicijativa osmišljena za proučavanje i razvoj metoda za predikciju utjecaja kemijskih spojeva na ljude i okoliš. Ovaj projekt predstavlja suradnju između nekoliko ključnih institucija:

- National Institute of Environmental Health Sciences (NIEHS)
- National Center for Advancing Translational Sciences (NCATS)
- U.S. Food and Drug Administration (FDA)
- National Center for Computational Toxicology pri U.S. Environmental Protection Agency (EPA)

Cilj Tox21 programa je unaprijediti toksikološka istraživanja kroz prikupljanje opsežnih podataka i razvoj novih tehnologija. Tradicionalni klinički testovi za procjenu toksičnosti su spori, skupi i često uključuju testiranje na životinjama, što je etički i praktično izazovno. Tox21 ima za cilj prevladati ove prepreke pomoću modernih tehnologija i metoda, uključujući testiranje na ljudskim stanicama (in vitro) i korištenje robotskih sustava za brzu i preciznu analizu.

Projekt se provodi u nekoliko faza:

**Faza 1:** Uvođenje robotskih sustava omogućilo je brzo i učinkovito testiranje velikog broja kemijskih spojeva u znatno kraćem vremenu nego što je to moguće s konvencionalnim metodama na životinjama.

**Faza 2:** Proširivanje baze podataka i povećanje broja testiranih spojeva iznad 10,000. Ova faza je omogućila detaljnije proučavanje i razumijevanje načina na koje kemijski spojevi djeluju na biološke sustave.

**Faza 3:** U tijeku za vrijeme pisanja ovoga završnog rada, ova faza fokusira se na razvoj testova za procjenu utjecaja različitih doza na ljude, pronalaženje ograničenja in vitro testiranja i opće poboljšanje metoda testiranja.

Projekt Tox21 prioritet stavlja na detaljno proučavanje i pronalazak mehanizama djelovanja kemijskih spojeva. Cilj je razviti prediktivne modele koji će omogućiti pouzdane procjene toksičnosti, smanjiti troškove i trud potrebne za testiranje te smanjiti potrebu za eksperimentalnim ispitivanjima na životinjama. Ovaj napredak ne samo da ubrzava proces toksikoloških istraživanja nego i značajno doprinosi sigurnosti ljudi i zaštiti okoliša [10][11][12].

# 2. Skup podataka

Skup podataka korišten u projektu Tox21, naveden na poveznici [11], obuhvaća 12,060 uzoraka za treniranje i 647 uzoraka za testiranje. Svaki uzorak se sastoji od približno 273,500 deskriptora raspodijeljenih u dva zapisa. Prvi zapis, nazvan “dense features” ili guste značajke sastoji se 801 značajke koje opisuju reaktivnost, topologiju, i različita druga svojstva molekula. Drugi zapis odnosi se na “sparse features” ili rijetke značajke koje opisuju unutarnju strukturu kemijskih spojeva i međusobnu povezanost atoma unutar navedenih spojeva.

## 2.1 Deskriptori podataka

**Guste značajke** (801 značajka) predstavljaju kemijske deskriptore kao što su molekulska težina, topljivost i površina. Evo nekoliko primjera:

**Atom Weights and Counts**

1. **AW, AWeight**: Ukupna atomska masa molekule.
2. **Arto**: Broj aromatskih atoma.

**Topološki Deskriptori**

3. **BertzCT**: Bertzov indeks složenosti, koji mjeri složenost molekule.
4. **Chi Indices (Chi0, Chi1, ..., Chi10)**: Kier-Hallovi topološki indeksi, koji opisuju strukturu molekule.
5. **Cluster and Chain Indices (Chi3c, Chi4c, ..., Chi6ch)**: Topološki deskriptori fokusirani na klastere i lance unutar molekule.
6. **Valence Connectivity Indices (Chiv0, Chiv1, ..., Chiv10)**: Indeksi povezivanja valencija.

**Elektro topološki Indeksi**

7. **EstateVSA (EstateVSA0, EstateVSA1, ..., EstateVSA9)**: Kombinacija elektro topoloških stanja i površinske dostupnosti molekule.

**Geometrijski Deskriptori**

8. **GATSe, GATSm, GATSp, GATSv (GATSe1, GATSe2, ..., GATSv8)**: GETAWAY deskriptori koji kombiniraju geometrijske, topološke i težinske informacije molekule.
9. **Geometric Topological Indices (GMTI, GMTIV)**: Deskriptori koji kombiniraju geometrijske i topološke informacije.
10. **Gravitational Index (Gravto)**: Mjera gravitacijskog efekta unutar molekule.

**Vodikove Vezne Karakteristike**

11. **Hy**: Broj donora vodikovih veza u molekuli.

**Informacijski Deskriptori**

12. **ISIZ**: Informacijski deskriptori koji mjere složenost informacija unutar molekule.

**Površinski Deskriptori**

13. **LabuteASA**: Procijenjena površina molekule prema Labudovoj metodi.
14. **LogP, LogP2**: Koeficijent raspodjele (Log P), koji pokazuje hidrofilnost ili lipofilnost molekule.

**Autokorelacijski Deskriptori**

15. **MATS (MATSe1, MATSe2, ..., MATSv8)**: 2D autokorelacijski deskriptori koji koriste 2D reprezentacije molekula za računanje autokorelacija.

**Molarna Refrakcija**

16. **MR, MRVSA (MRVSA0, ..., MRVSA9)**: Deskriptori koji se odnose na molarnu refrakciju.

**Petitjeanov Indeks**

17. **Petitjeant, petito**: Petitjeanov indeks koji mjeri simetriju molekule.

**Polarizabilnost**

18. **PEOEVSA (PEOEVSA0, PEOEVSA1, ..., PEOEVSA13)**: Deskriptori koji kombiniraju efekt polarizabilnosti i površinske dostupnosti.

**Radijalna Distribucija**

19. **RDF (RDFE1, RDFE2, ..., RDFV30)**: Deskriptori koji koriste funkciju radijalne distribucije za opisivanje molekularne strukture.

**Površinska Područja**

20. **SlogP (SlogPVSA0, SlogPVSA1, ..., SlogPVSA9)**: Deskriptori povezani s SlogP vrijednostima molekula.
21. **VSAEstate (VSAEstate0, VSAEstate1, ..., VSAEstate9)**: Deskriptori koji kombiniraju van der Waals površinsku površinu (VSA) i elektro topološko stanje.

**Masa i Težina**

22. **W, Weight**: Molekularna masa.

**Topološki Indeksi**

23. **Xu**: Xu indeks koji opisuje strukturu molekule.
24. **ZM1, ZM2**: Zagreb M indeksi, topološki indeksi bazirani na Zagreb indeksima.

**Ostali Indeksi**

25. **Diameter, Radius**: Deskriptori promjera i radijusa molekule.
26. **KnotP, KnotPV**: Deskriptori čvorova unutar molekule.
27. **Naccr, Naro**: Broj akceptora i donora.
28. **Ncarb, Nring**: Broj ugljikovih atoma i prstenova.
29. **Nrot, Nsb**: Broj rotacijskih veza i jednostrukih veza.
30. **Nhet, Nhal**: Broj heteroatoma i halogenih atoma.
31. **Noxy, Nphos**: Broj kisikovih i fosfornih atoma.
32. **Ndb, Ndonr**: Broj dvostrukih veza i donora.
33. **RASA, PSA**: Relativna i apsolutna površina molekule.
34. **TIAC**: Total Internal Atomic Charge, ukupni unutarnji atomski naboj.

**Rijetke značajke** (272,776 značajki) predstavljaju kemijske podstrukture, pohranjene u Matrix Market formatu. Ove značajke uključuju ECFP (Extended Connectivity Fingerprints) i DFS (Discriminative Fingerprint Signatures), koje opisuju topološku strukturu molekula:

1. **ECFP (Extended Connectivity Fingerprints):**
    - **ECFP2, ECFP4, ECFP6, ECFP8, ECFP10:** Otisci prsta koji koriste kružne potpise promjera od 2 do 10 veza, koriste se za opisivanje molekularne strukture.
2. **DFS (Discriminative Fingerprint Signatures):**
    - **DFS6, DFS8:** Deskriptori optimizirani za razlikovanje između različitih klasa molekula.

Strojno učenje može koristiti bilo gusto, rijetko ili kombinirano predočenje podataka za predikciju toksičnosti kemijskih spojeva. Za svaki uzorak u skupu podataka postoje 12 binarnih oznaka koje predstavljaju ishod (aktivno/neaktivno) 12 različitih toksikoloških eksperimenata. Ovi eksperimenti uključuju mjerenje aktivacije ili inhibicije različitih bioloških receptora i odgovora, kao što su:

1. **NR.AhR**: Aktivacija aril-hidrokarbonskog receptora. Aril-hidrokarbonski receptor (AhR) je nuklearni receptor koji reagira na različite endogene i egzogene spojeve, uključujući one iz onečišćenja okoliša.
2. **NR.AR**: Aktivacija androgenog receptora. Androgeni receptor (AR) je nuklearni receptor koji reagira na androgene hormone poput testosterona i dihidrotestosterona.
3. **NR.AR.LBD**: Aktivacija ligand-vezujućeg dijela androgenog receptora. Ovo je specifičnija podskupina testova koji se odnose na vezivanje liganda za ligand-vezujući dio androgenog receptora.
4. **NR.Aromatase**: Inhibicija aromataze. Aromataza je enzim koji katalizira konverziju androgena u estrogene.
5. **NR.ER**: Aktivacija estrogenog receptora. Estrogeni receptor (ER) je nuklearni receptor koji reagira na estrogene hormone poput estradiola.
6. **NR.ER.LBD**: Aktivacija ligand-vezujućeg dijela estrogenog receptora. Ovo je specifičnija podskupina testova koji se odnose na vezivanje liganda za ligand-vezujući dio estrogenog receptora.
7. **NR.PPAR.gamma**: Aktivacija peroksizomnog proliferator-aktiviranog receptora gamma. PPAR-gamma je nuklearni receptor koji ima ulogu u regulaciji metabolizma lipida i glukoze.
8. **SR.ARE**: Aktivacija odlaznog odgovora na antiestrogene. Ova oznaka se odnosi na testiranje učinka spojeva na stanične procese koji se aktiviraju putem antiestrogenskog odgovora.
9. **SR.ATAD5**: Aktivacija odlaznog odgovora na oksidacijski stres. Ova oznaka se odnosi na testiranje učinka spojeva na stanične procese koji su aktivirani oksidacijskim stresom.
10. **SR.HSE**: Aktivacija toplinske šoka odgovora. Ova oznaka se odnosi na testiranje učinka spojeva na stanične procese koji su aktivirani toplinskim stresom.

11. **SR.MMP**: Inhibicija proizvodnje matriksnih metaloproteinaza. Matriksne metaloproteinaze (MMP) su enzimi koji sudjeluju u razgradnji ekstracelularne matrice u tijelu.
12. **SR.p53**: Aktivacija odgovora na stres p53. Ova oznaka se odnosi na testiranje učinka spojeva na stanične procese koji su regulirani tumor-supresorskim proteinom p53.

Ovaj skup podataka pruža detaljan opis kemijskih spojeva i njihovih svojstava. Uz podatke o aktivaciji ili inhibiciji receptora, skup podataka nam omogućava treniranje različitih modela strojnog učenja s ciljem brže i jeftinije procjene određenih svojstava koja bi u protivnom zahtijevala dugotrajna i skupa testiranja u laboratoriju [13][14][15][16][17].

## 2.2 Receptori

Identificiranje receptora na koje određeni kemijski spojevi utječu ključno je u razumijevanju njihovih bioloških učinaka i potencijala za terapijske primjene. Receptori su specifični proteini ili molekule koji, kad se vežu s određenim ligandima (kao što su hormoni, neurotransmiteri, ili lijekovi), aktiviraju biološke signale koji pokreću određene stanične procese. Evo nekoliko razloga zašto je važno identificirati na koje receptore kemijski spojevi djeluju:

1. **Predviđanje Terapijskih Učinaka i Nuspojava**:
    - **Terapijski Učinci**: Spoznaja o tome na koje receptore kemijski spoj djeluje omogućuje predviđanje njegovih terapijskih učinaka. Na primjer, agonisti dopaminskih receptora mogu biti korisni u liječenju Parkinsonove bolesti.
    - **Nuspojave**: Identificiranjem svih receptora na koje spoj može djelovati, moguće je predvidjeti i potencijalne nuspojave. Na primjer, antipsihotici koji djeluju na dopaminske receptore mogu izazvati nuspojave povezane s pokretima.

2. **Razvoj Ciljanih Terapija**:
    - Razumijevanje specifičnosti veze između kemijskih spojeva i receptora omogućuje razvoj ciljane terapije koja djeluje samo na određene receptore, smanjujući time nuspojave i povećavajući učinkovitost liječenja.
3. **Personalizirana Medicina**:
    - Informacije o receptorima na koje kemijski spojevi djeluju mogu se koristiti za personalizaciju terapije, prilagođavajući liječenje individualnim potrebama pacijenata na temelju njihovog jedinstvenog receptorskog profila.
4. **Razumijevanje Bioloških Procesa**:
    - Identifikacija receptora pomaže u razumijevanju osnovnih bioloških procesa i puteva koji su uključeni u različite fiziološke i patološke uvjete.

Kako ne bi došlo do neočekivanih posljedica, važno je znati koje točno receptore svaki kemijski spoj aktivira ili inhibira. To nam omogućuje, prije samog testiranja, predviđanje utjecaja određenog kemijskog spoja na različite procese unutar organizma.

# 3. Modeli strojnog učenja

U zadanom skupu podataka, kao ulazne podatke, model strojnog učenja prima velik broj različitih deskriptora u obliku kontinuiranih podataka (npr. različiti indeksi, površine, fizičke veličine kemijskih spojeva) i diskretnih podataka (npr. broj određenih podstruktura unutar spoja, prisustvo ili odsustvo određenih elemenata). Kao izlaz modela strojnog učenja, očekuje se diskretni podatak, ili preciznije binarna klasifikacija koja se odnosi na aktivaciju / inhibiciju ili neaktivaciju / neihibiciju određenog receptora. Za navedeni kontekst, modeli strojnog učenja odabrani za treniranje su model logističke regresije, model slučajne šume (eng. *random forest model*) i *XGBoost model*.

## 3.1 Model logističke regresije

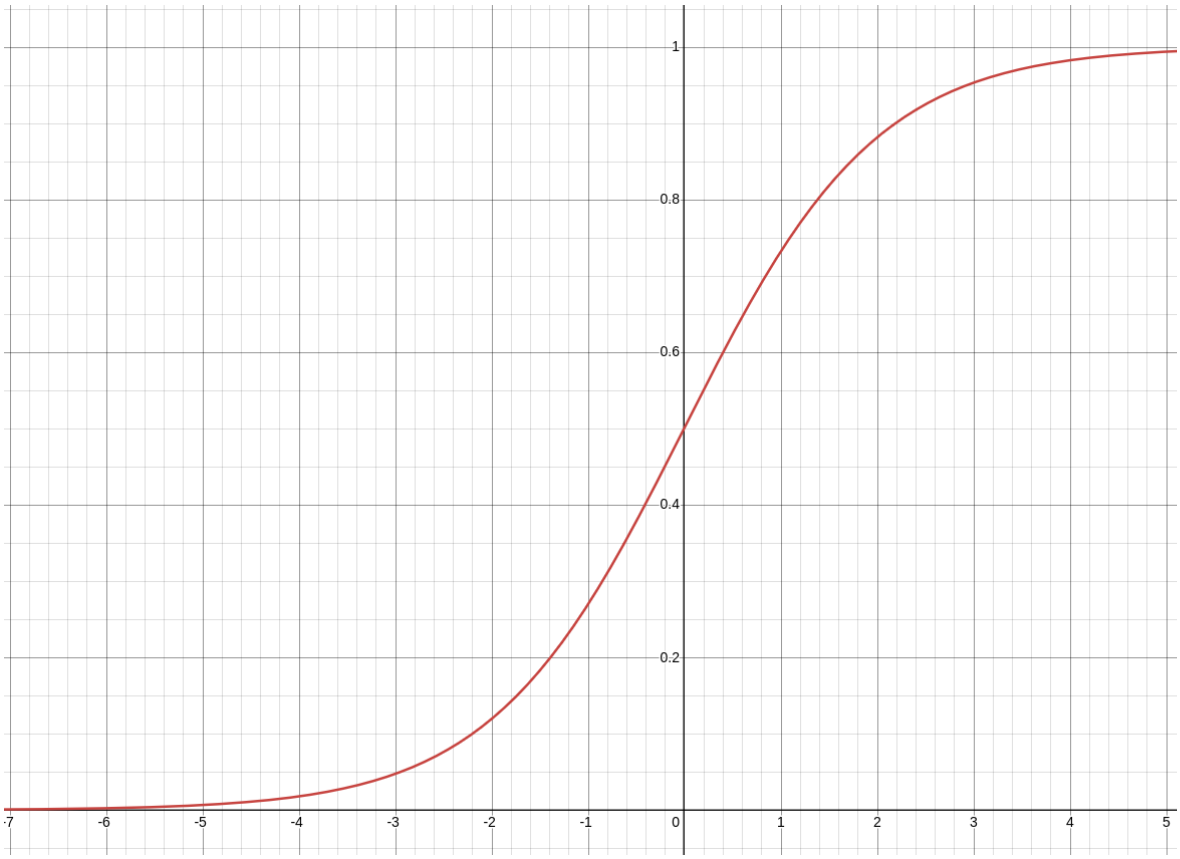
Model logističke regresije je metoda statističke analize koja se koristi za modeliranje binarnih ovisnih varijabli, odnosno varijabli koje imaju dva moguća ishoda. Ulazni podaci u model logističke regresije uključuju jednu ili više nezavisnih varijabli, koje mogu biti kontinuirane ili kategoričke. Cilj je analizirati kako te nezavisne varijable utječu na vjerojatnost ishoda binarne zavisne varijable. Izlazni podatak je binarna zavisna varijabla koja može imati dvije moguće vrijednosti. Funkcija logističke regresije koristi logističku funkciju (poznatu i kao sigmoidnu funkciju) za modeliranje vjerojatnosti kako bi zavisna varijabla poprimila određenu vrijednost na temelju nezavisnih varijabli. Matematički, logistička funkcija je definirana u formuli [Formula 3.1] i grafički prikaz funkcije nalazi se u slici [Slika 3.1].

$$P(Y = 1 \mid X) = \frac{1}{1+e^{-\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n}}$$ *Formula 3.1 Pirkaz formule logističke regresije*

Gdje je P(Y=1|X) vjerojatnost da varijabla Y bude 1 za zadani skup nezavisnih varijabli X, e baza prirodnog logaritma (približno 2.718), $\beta_0$ konstanta (intercept), $\beta_1$ ,$\beta_2$,...,$\beta_n$ koeficijenti regresije za nezavisne varijable $X_1$,$X_2$,...,$X_n$.

Način na koji model koristi logističku funkciju je prvo procjena koeficijenata, zatim Kalkulacija vjerojatnosti, nakon čega je moguća predikcija ishoda. Procjena

koeficijenata se izvršava na način gdje se koeficijenti β procjenjuju metodom maksimalne vjerojatnosti. Maksimalna vjerojatnost traži vrijednosti koeficijenata koje maksimiziraju vjerojatnost promatranog skupa podataka. Kalkulacija vjerojatnosti se izvršava nakon što se procijene koeficijenata. Model koristi logističku funkciju za izračunavanje vjerojatnosti kada zavisna varijabla poprima vrijednost 1 za dane vrijednosti nezavisnih varijabli. Predikcija ishoda omogućava na temelju izračunate vjerojatnosti, model može klasificirati ishod. Obično se koristi prag od 0.5: ako je vjerojatnost veća od 0.5, model predviđa Y=1, inače predviđa Y=0.

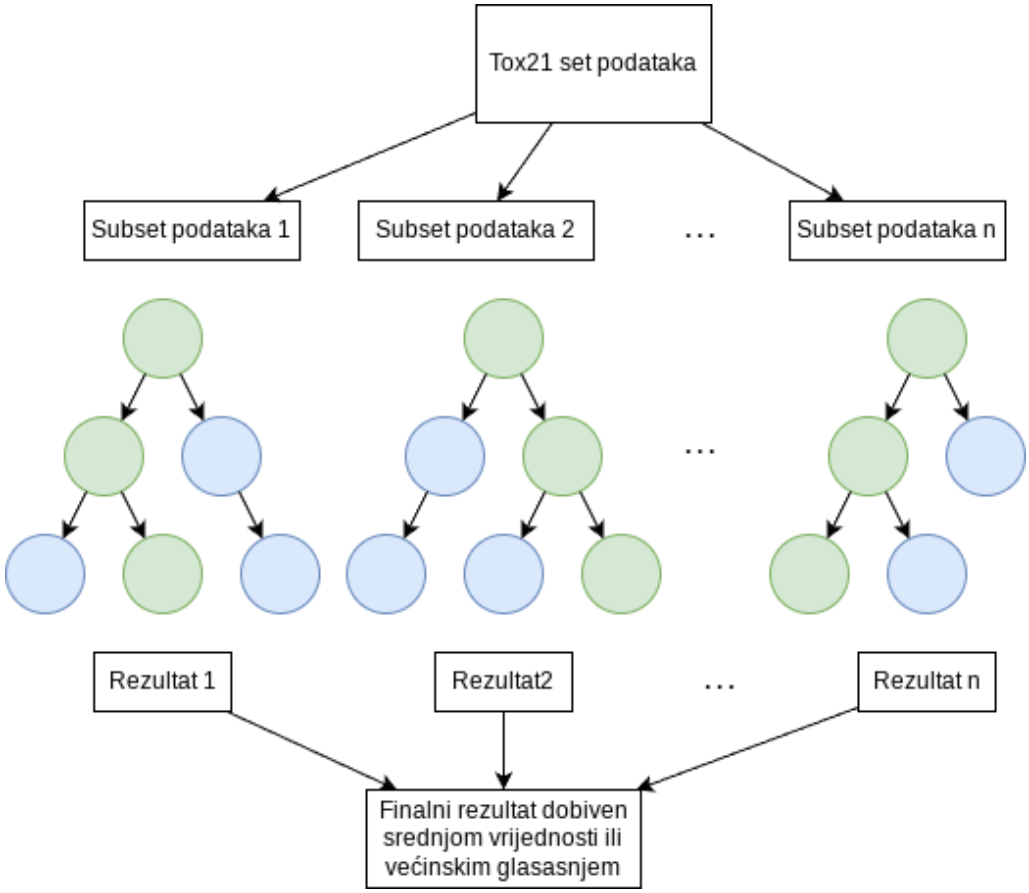



*Slika 3.1 Prikaz funkcije sigmoide*

## 3.2 Model slučajne šume

Model slučajne šume (eng. *random forest model*) sastoji se od mnogih pojedinačnih stabala odlučivanja (eng. *decision trees*). Svako stablo u šumi donosi svoju predikciju, a konačna predikcija modela temelji se na prosjeku (u slučaju regresije) ili većinskom glasanju (u slučaju klasifikacije) tih predikcija. Ključne komponente modela slučajne šume uključuju *bagging* i *random subspace* metodu. *Bagging* je tehnika koja koristi višestruke uzorke podataka s ponavljanjem kako bi stvorila

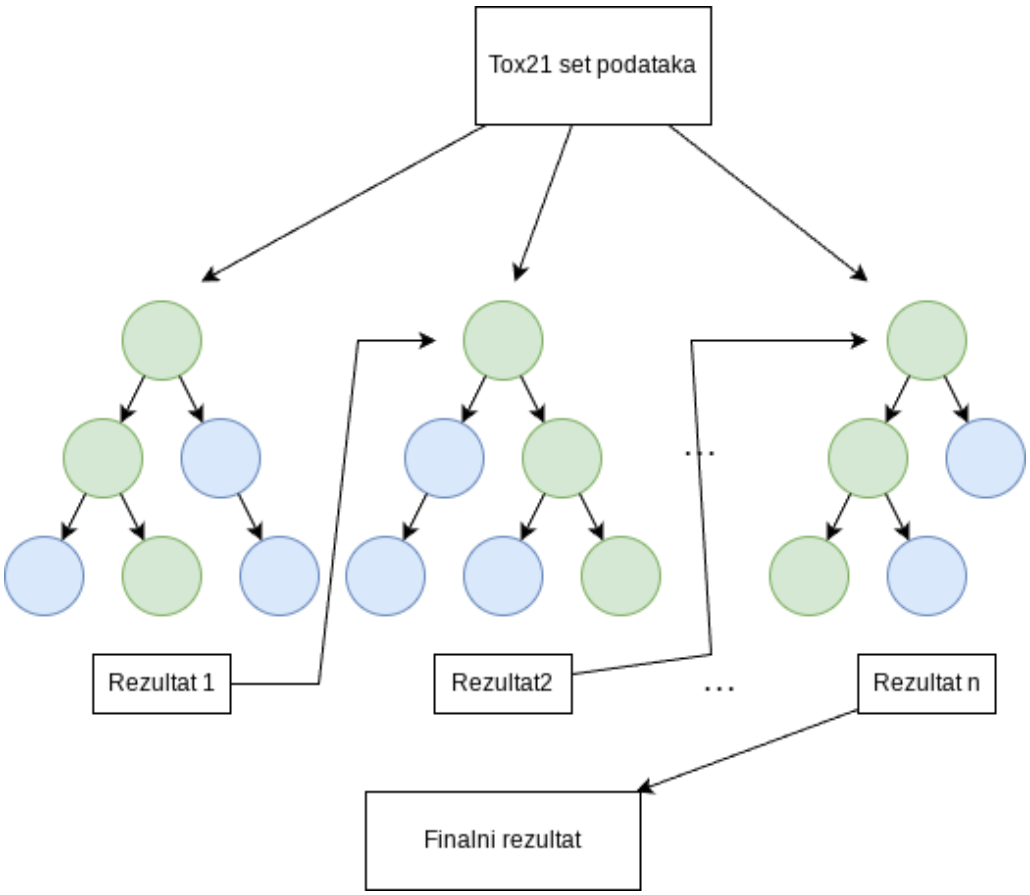
različite modele. Navedeni uzorci se također nazivaju "*bootstrap samples*". Za model slučajne šume, to znači da se svaki model (stablo odlučivanja) trenira na različitim podskupovima podataka iz originalnog skupa podataka. Prilikom izgradnje svakog stabla odlučivanja, model slučajne šume dodatno uvodi slučajnost odabirom nasumičnog podskupa značajki (feature subset) za razmatranje pri svakom podjelu čvora. Ovi načini treniranja omogućavaju modelu slučajne šume sposobnost otpornosti na preprilagođavanje, sposobnost rukovanja velikim brojem značajki i procjenu značajnosti značajki. Otpornost na preprilagođavanje se ostvaruje kombiniranjem više stabala odlučivanja, random forest smanjuje rizik od preprilagođavanja. Svako pojedinačno stablo može biti preprilagođeno na svom bootstrap uzorku, ali njihova kombinacija često rezultira boljom generalizacijom na nove podatke. Rukovanje velikim brojem značajki se ostvaruje zahvaljujući slučajnom odabiru podskupova značajki prilikom izgradnje stabala. To ga čini otpornim na nepotrebne ili manje važne značajke. Model slučajne šume može pružiti informacije o značajnosti značajki, što je korisno za razumijevanje koje značajke najviše doprinose predikcijama modela. Ovo se postiže mjerenjem koliko svaki podskup značajki poboljšava kvalitetu podjela čvorova kroz sva stabla. Shema modela slučajne šume strojnog učenja prikazana je na slici [Slika 3.2].




*Slika 3.2 Prikaz random forest modela strojnog učenja*

## 3.3 XGBoost model

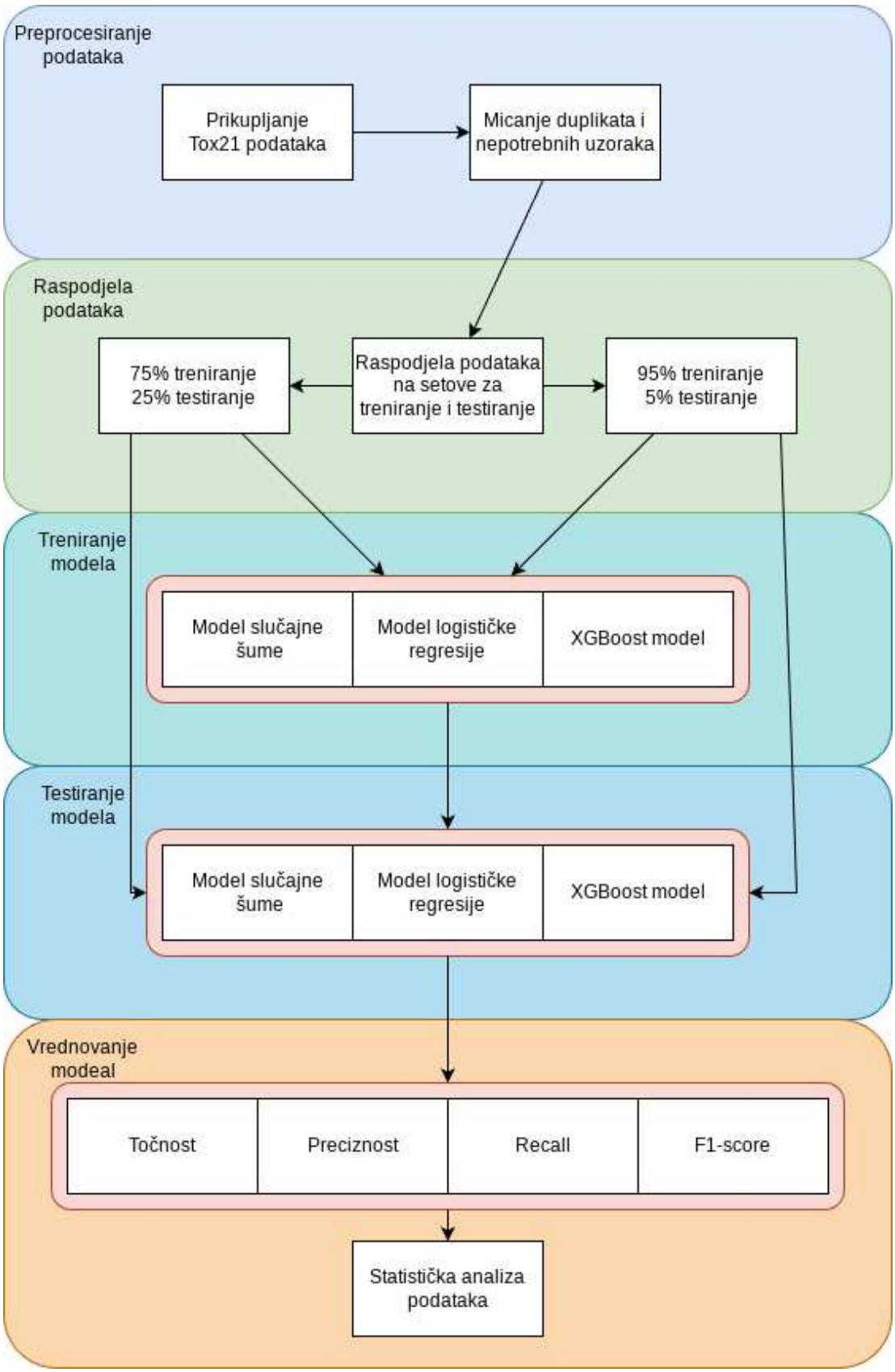
*XGBoost* (Extreme Gradient Boosting) je model strojnog učenja koji se temelji na metodi "*boosting*". *Boosting* metoda radi iterativnim treniranjem modela,pri čemu svaki novi model pokušava ispraviti pogreške svojih prethodnika. U *XGBoost*-u, modeli su obično stabla odlučivanja, a treniranje se provodi pomoću tehnike "*gradient boosting*". Glavne komponente *XGBoost* modela uključuju *gradient boosting*, *additive training* i regularizaciju. *Gradient boosting* je tehnika koja trenira nove modele kako bi ispravili pogreške prethodnih modela. To se postiže minimiziranjem funkcije gubitka koristeći gradijente. Vezano uz *gradient boosting*, metoda additive training se odnosi na dodavanje modela. Modeli se dodaju jedan po jedan, a svaki novi model trenira se kako bi korigirao rezidualne pogreške prethodnih modela. *XGBoost* uključuje regularizacijske tehnike koje pomažu spriječiti preprilagođavanje (eng. *overfitting*). Shema *XGBoost* modela strojnog učenja prikazana je na slici [Slika 3.3].




*Slika 3.3 Prikaz XGBoost modela strojnog učenja*

## 3.4 Proces učenja i vrednovanja modela

Proces obrade podataka u strojnim modelima učenja prolazi kroz nekoliko ključnih faza. Prvo se prikupljaju podaci, koji uključuju razne deskriptore u obliku kontinuiranih (npr. fizičke veličine kemijskih spojeva) i diskretnih podataka (npr. broj podstruktura unutar spoja). Prikupljeni podaci se čiste kako bi se uklonili duplikati i nepotrebni uzorci. Zatim se podaci dijele na trening i test skupove, postoje dvije podjele podataka. Prva podjela, već definirana unutar Tox21 skupa podataka u omjeru 95% za trening i 5% za testiranje. I druga podjela u omjeru 75% za trening i 25% za testiranje. Ovaj korak omogućuje treniranje modela na dovoljno podataka dok se performanse testiraju na neviđenim podacima. Treniranje uključuje korištenje algoritama Logističke regresije, Slučajne šume i *XGBoost*. Svaki model se trenira na trening skupu, a optimizacija se provodi za postizanje najboljih rezultata. Nakon treniranja, modeli se testiraju na test skupu podataka, a performanse se procjenjuju pomoću metričkih mjera: točnost, preciznost, odziv i F1-score. Na temelju rezultata vrednovanja, odabire se model s najboljim performansama za daljnju upotrebu. Tok podataka prikazan je u dijagramu [Dijagram 3.1]. Ovaj tok podataka omogućava učinkovitu pripremu i obradu podataka te osigurava optimalne performanse modela strojnog učenja u predviđanju binarnih klasifikacija za aktivaciju ili inhibiciju određenih receptora [7][18][19][20].




*Dijagram 3.1 Dijagram toka podataka*

# 4. Vrednovanje modela

Vrednovanje modela strojnog učenja je ključan korak u razumijevanju performansi modela i donošenju odluka o njegovoj korisnosti. Postoji nekoliko metoda vrednovanja koje se koriste u različitim situacijama, ovisno o karakteristikama skupa podataka i ciljevima analize. Za navedene modele, vrednovanje će se izvršiti koristeći mjere točnosti, preciznost, odziv i F1-score.

Točnost je mjera koja pokazuje postotak točno predviđenih instanci u odnosu na ukupan broj instanci. Točnost je definirana kao omjer broja točnih predikcija i ukupnog broja predikcija. Točnost modela definirana je u formuli [Formula 4.1].

$$Točnost = \frac{TP+TN}{Tp+TN+FP+FN} \quad \textit{Formula 4.1 Formula za vrednovanje točnosti modela}$$

U formuli, TP se odnosi na True Positive, odnosno vrijednosti koje je model ispravno procijenio kao točne. TN se odnosi na True Negative, odnosno vrijednosti koje je model ispravno procijenio kao netočne. FP se odnosi na False Positive, odnosno vrijednosti koje je model procijenio kao točne, iako su u stvarnosti netočne. FN se odnosi na False Negative, odnosno vrijednosti koje je model procijenio kao netočne, iako su u stvarnosti točne.

Preciznost mjeri koliko su točne pozitivne predikcije modela i prikazana je u formuli [Formula 4.2].

$$Preciznost = \frac{TP}{TP+FP} \quad \textit{Formula 4.2 Formula za vrednovanje preciznosti modela}$$

Preciznost kao mjera nam daje informaciju o tome koliko će često model, za određene spojeve, reći da aktivira ili inhibira više receptora nego što je u stvarnosti točno.

Odziv mjeri koliko dobro model identificira stvarne pozitivne primjere i prikazana je u formuli [Formula 4.3].

$$Recall = \frac{TP}{TP+FN}$$ *Formula 4.3 Formula za vrednovanje vrijednosti odziva modela*

Visok odziv upućuje na rijetko izostavljanje receptora koje kemijski spojevi aktiviraju ili inhibiraju.

F1-score je harmonijska sredina preciznosti i odziva i koristi se kada je potrebno balansirati između njih. Vrijednost F1-score vrijednosti definirana je formulom [Formula 4.4].

$$F1 - score = 2 \cdot \frac{Preciznost \cdot Recall}{Preciznost + Recall}$$ *Formula 4.4 Formula za vrednovanje F1-score vrijednosti modela*

F1-score pruža informaciju o tome ispisuje li model, kao izlaz, podataka o tome aktivira li ili inhibira kemijski spojevi previše ili premalo receptora [7].

# 5. Rezultati

Za treniranje različitih modela korišten je Tox21 skup podataka. Korišteni su i guste i rijetke značajke za opisivanje molekula. Za testiranje modela korištena je unakrsna provjera, tj. uzorak se podijelio na skup za treniranje i skup za testiranje. Originalno, unutar Tox21 skupa podataka postoji podjela podataka na skup za treniranje i skup za testiranje. Skup podataka je podijeljen tako da se u skupu za treniranje nalazi približno 95% uzoraka. Također, uzorci unutar skupa za testiranje nisu nasumično uzeti iz skupa podataka. Tox21 skup podataka nema potpune informacije o interakciji između svakog kemijskog spoja i svakog receptora, nego se mogu pojaviti podaci poput ovog navedenog u primjeru [Primjer 5.1]:

*Primjer 5.1 Primjer izlaznih vrijednosti za jedan uzorak*

| |
|---|
| "NCGC00178831-03",NA,NA,NA,NA,NA,NA,NA,NA,NA,0,NA,NA |

Kemijski spoj u primjeru se naziva Akriflavin hidroklorid, ime se može dobiti koristeći ID molekule (prva vrijednost). Preostale vrijednosti odnose se na interakciju s određenim receptorom. Vrijednost NA označava „*not available*" tj. nema informacija o interakciji između kemijskog spoja i receptora. Preostala vrijednost je 0, što označava nepostojanje interakcije između kemijskog spoja i receptora (vrijednost 1 bi označavala postojanje interakcije). Kemijski spojevi poput ovog u primjeru su stavljeni u skup za treniranje, dok skup za testiranje sadrži uzorke koji imaju ili sve interakcije definirane ili veći broj interakcija definiran. Uz originalnu raspodjelu podataka, modeli su bili trenirani i na raspodjeli podataka gdje je udio podataka u skupu za treniranje bio 75%. Važno je napomenuti da su se modeli trenirali zasebno za svaki receptor i prije raspodjele uzoraka na skup za treniranje i skup za testiranje, izbačeni su uzorci koji za navedeni receptor nemaju definiranu interakciju.

### 5.1 Tablični prikaz podataka

Ispitivanje treniranja medela izvedeno je na način da su za ulaz priložene sve značajke uzorka, dok za izlaz priložena vrijednost postoji li interakcija s određenim receptorom. Za svaku kombinaciju modela strojnog učenja i receptora, model je treniran 15 puta kako bi se smanjila vjerojatnost gdje će rezultati biti slučajno ekstremno loši ili optimalni zbog specifičnog izbora trening i test podataka, čime se osigurava pouzdanije i stabilnije vrednovanje performansi modela. Unutar tablica je navedena srednja vrijednost, RF se odnosi na model slučajne šume, LR na model logističke regresije i XGB je *XGBoost* model. [Tablica 5.1] prikazuje rezultate za slučaj gdje se skup za treniranje sastoji od 95% uzoraka, a [Tablica 5.2] prikazuje rezultate kada je skup za treniranje bio 75% uzoraka.

*Tablica 5.1 Prikaz podataka gdje je skup za treniranje 95%*

| Metoda vrednovanja | Model | NR. AhR | NR. AR | NR. AR. LBD | NR. Aromatase | NR. ER | NR. ER. LBD | NR. PPAR .gamma | SR. ARE | SR. ATAD 5 | SR. HSE | SR. MMP | SR. p53 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Točnost | RF | 0.9 | 0.98 | 0.98 | 0.93 | 0.91 | 0.97 | 0.95 | 0.85 | 0.94 | 0.97 | 0.91 | 0.94 |
| | LR | 0.88 | 0.97 | 0.97 | 0.9 | 0.87 | 0.95 | 0.94 | 0.81 | 0.92 | 0.94 | 0.89 | 0.91 |
| | XGB | 0.9 | 0.98 | 0.98 | 0.93 | 0.91 | 0.97 | 0.94 | 0.84 | 0.94 | 0.96 | 0.92 | 0.93 |
| Precizno | R | 0.71 | 0.67 | 0 | 1 | 0.69 | 0.5 | 0 | 0.63 | 0 | 0.6 | 0.68 | 0.67 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| st | F | | | | | | | | | | | | |
| | LR | 0.5 | 0 | 0 | 0.28 | 0.32 | 0.22 | 0.33 | 0.42 | 0.33 | 0.3 | 0.5 | 0.29 |
| | XGB | 0.6 | 0.67 | 0 | 0.8 | 0.63 | 0.67 | 0 | 0.55 | 0.71 | 0.17 | 0.67 | 0.38 |
| Odziv | RF | 0.33 | 0.17 | 0 | 0.03 | 0.18 | 0.15 | 0 | 0.18 | 0 | 0.14 | 0.38 | 0.05 |
| | LR | 0.55 | 0 | 0 | 0.21 | 0.29 | 0.2 | 0.19 | 0.42 | 0.26 | 0.41 | 0.68 | 0.29 |
| | XGB | 0.45 | 0.17 | 0 | 0.1 | 0.24 | 0.2 | 0 | 0.31 | 0.13 | 0.05 | 0.58 | 0.07 |
| F1-score | RF | 0.45 | 0.27 | 0 | 0.05 | 0.28 | 0.23 | 0 | 0.28 | 0 | 0.22 | 0.49 | 0.09 |
| | LR | 0.52 | 0 | 0 | 0.24 | 0.31 | 0.21 | 0.24 | 0.42 | 0.29 | 0.35 | 0.58 | 0.29 |
| | XGB | 0.52 | 0.27 | 0 | 0.18 | 0.34 | 0.31 | 0 | 0.4 | 0.22 | 0.07 | 0.63 | 0.12 |

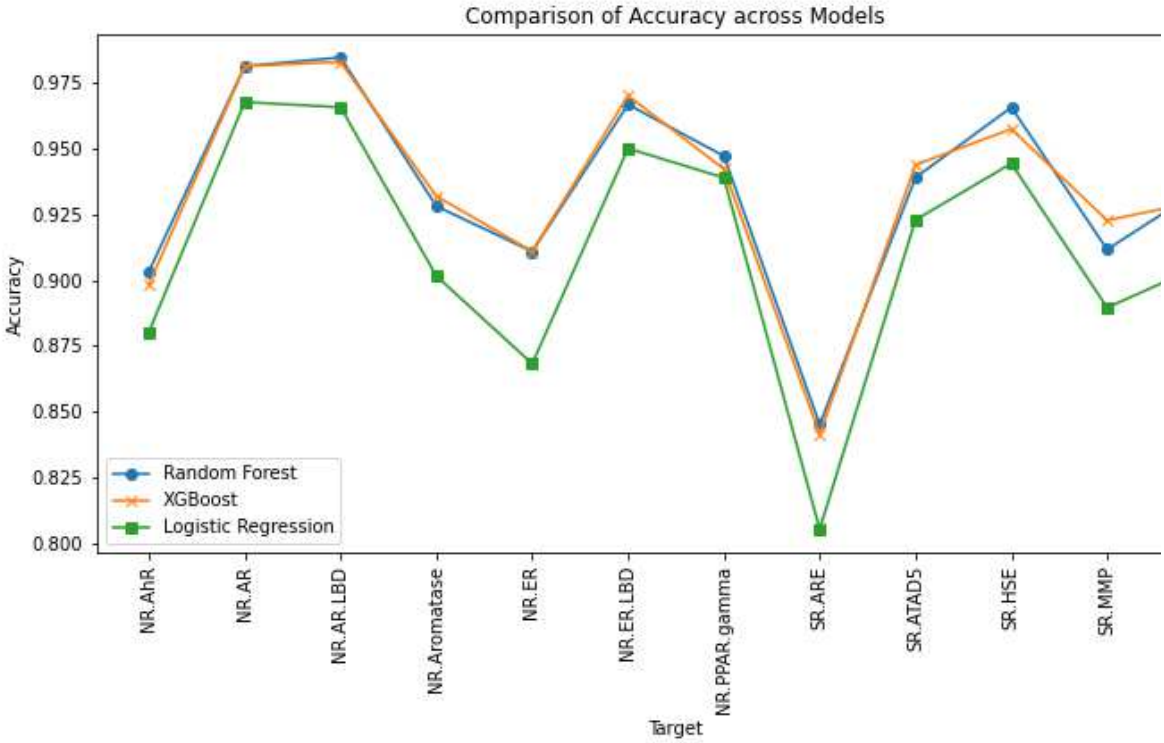
*Tablica 5.2 Prikaz podataka gdje je skup za treniranje 75%*

| Metoda vrednovanja | Model | NR.AhR | NR.AR | NR.AR.LBD | NR.Aromatase | NR.ER | NR.ER.LBD | NR.PPAR.gamma | SR.ARE | SR.ATAD5 | SR.HSE | SR.MMP | SR.p53 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Točnost | RF | 0.92 | 0.98 | 0.98 | 0.96 | 0.91 | 0.97 | 0.97 | 0.87 | 0.97 | 0.95 | 0.91 | 0.95 |
| | LR | 0.92 | 0.97 | 0.97 | 0.93 | 0.88 | 0.96 | 0.96 | 0.84 | 0.95 | 0.94 | 0.90 | 0.93 |
| | XGB | 0.94 | 0.98 | 0.98 | 0.95 | 0.91 | 0.97 | 0.98 | 0.88 | 0.97 | 0.95 | 0.92 | 0.95 |
| Preciznost | RF | 0.84 | 0.88 | 0.89 | 0.83 | 0.79 | 0.87 | 0.85 | 0.79 | 0.83 | 0.75 | 0.84 | 0.89 |
| | LR | 0.64 | 0.66 | 0.63 | 0.44 | 0.49 | 0.52 | 0.41 | 0.51 | 0.49 | 0.44 | 0.69 | 0.47 |
| | XGB | 0.81 | 0.89 | 0.88 | 0.80 | 0.74 | 0.88 | 0.94 | 0.77 | 0.90 | 0.71 | 0.82 | 0.80 |
| Odziv | RF | 0.40 | 0.52 | 0.54 | 0.24 | 0.32 | 0.39 | 0.15 | 0.27 | 0.23 | 0.18 | 0.54 | 0.25 |
| | LR | 0.51 | 0.49 | 0.62 | 0.44 | 0.39 | 0.49 | 0.36 | 0.42 | 0.37 | 0.33 | 0.64 | 0.40 |
| | XGB | 0.55 | 0.53 | 0.63 | 0.30 | 0.35 | 0.48 | 0.23 | 0.37 | 0.32 | 0.22 | 0.63 | 0.30 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F1-score | RF | 0.54 | 0.65 | 0.67 | 0.37 | 0.46 | 0.54 | 0.25 | 0.40 | 0.36 | 0.29 | 0.66 | 0.39 |
| | LR | 0.57 | 0.56 | 0.63 | 0.44 | 0.43 | 0.51 | 0.38 | 0.46 | 0.42 | 0.38 | 0.66 | 0.43 |
| | XGB | 0.66 | 0.66 | 0.73 | 0.44 | 0.48 | 0.62 | 0.37 | 0.50 | 0.47 | 0.33 | 0.71 | 0.43 |

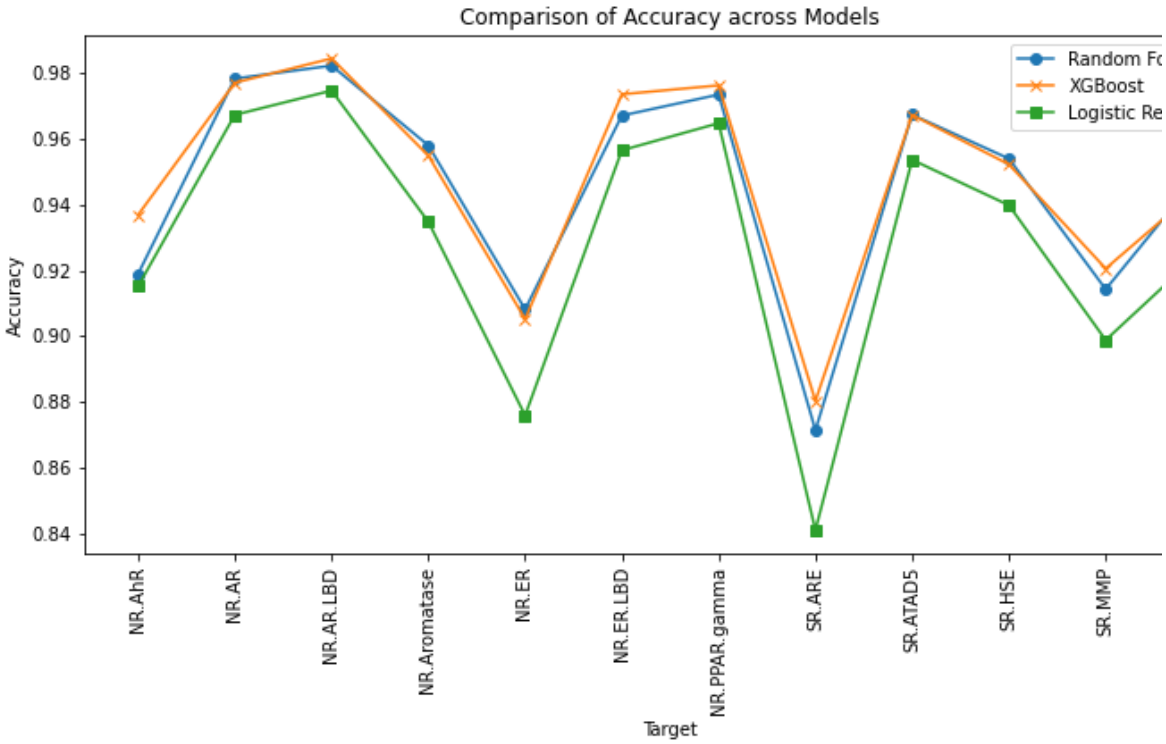
## 5.2 Točnost

Grafički prikaz točnosti modela na oba skupa podataka (95% i 75% treninga) jasno pokazuje sličnost sveukupnih performansi modela, uz izuzetak Logistic Regression (LR) modela koji je u nekoliko slučajeva pokazao nešto niže performanse.




*Graf 5.1 Usporedba točnosti gdje je skup za treniranje 95%*

Na grafu za 95% trening skup [Graf 5.1], *XGBoost* (XGB) model je postigao najvišu točnost za većinu receptora, dok je model slučajne šume (RF) također bio vrlo blizu. Model logističke regresije je općenito pokazao nižu točnost, posebno kod receptora NR.ER i SR.ARE gdje je razlika bila značajna.

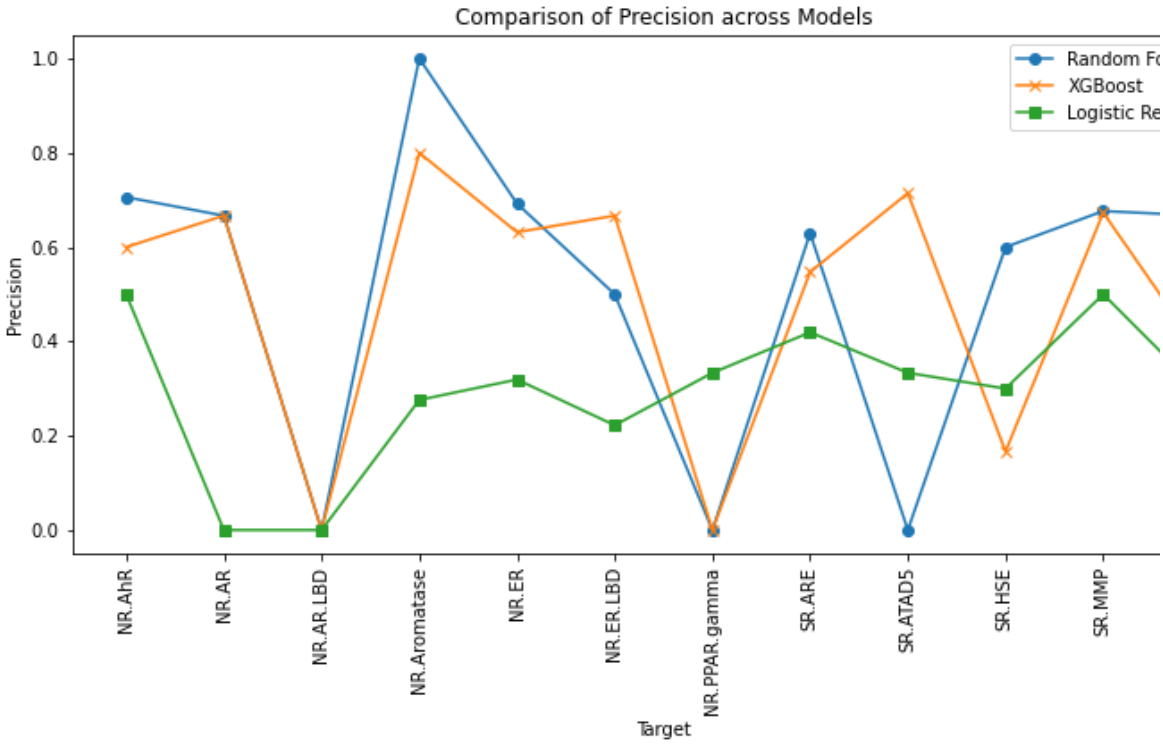



*Graf 5.2 Usporedba točnosti gdje je skup za treniranje 75%*

Na grafu za 75% trening skup [Graf 5.2], ponovno se vidi da su *XGBoost* model i model slučajne šume vrlo blizu u točnosti, dok model logističke regresije zaostaje, posebno kod receptora NR.ER i SR.ARE. Ova konzistentna prednost *XGBoost* modela može se pripisati njegovoj sposobnosti da bolje iskoristi dostupne podatke i složenije strukture stabala odlučivanja.

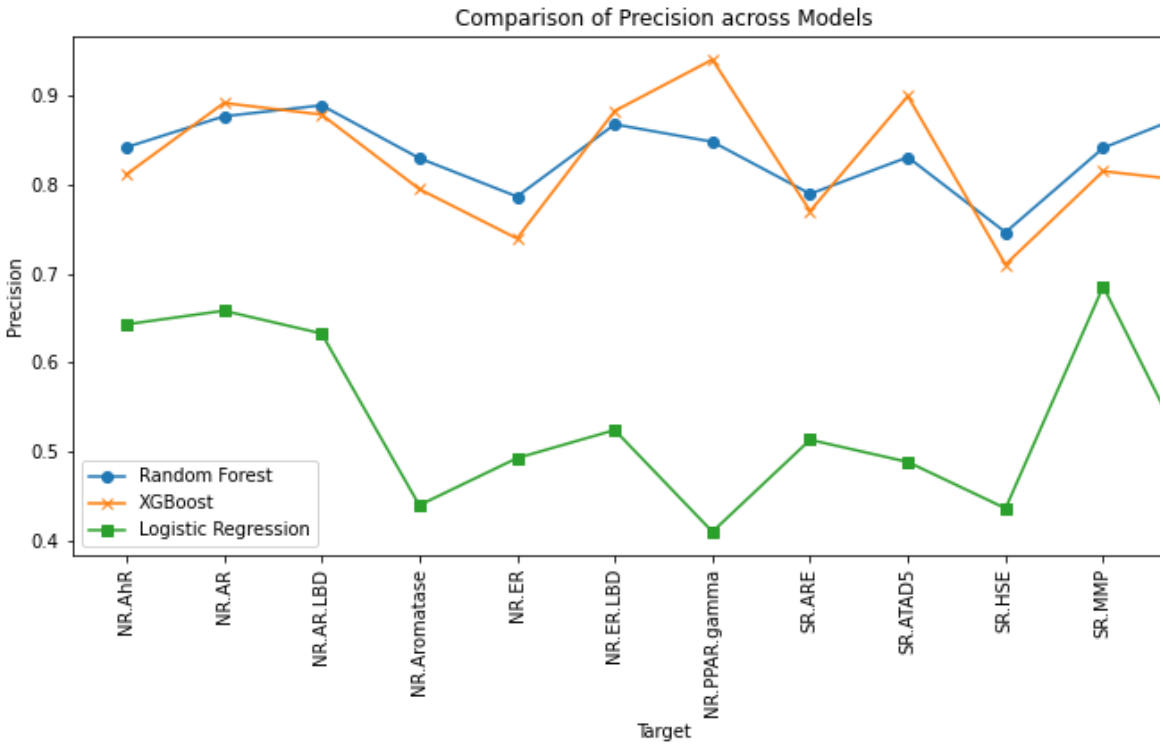
## 5.3 Preciznost

Grafički prikaz preciznosti modela na oba skupa podataka (95% i 75% treninga) jasno pokazuje razlike u performansama među modelima.




*Graf 5.3 Usporedba preciznosti gdje je skup za treniranje 95%*

Na grafu za 95% trening skup [Graf 5.3], preciznost za modele slučajne šume i *XGBoost* varira više u odnosu na 75% skup. Model slučajne šume pokazuje znatne padove kod receptora NR.AR i NR.PPAR.gamma, dok *XGBoost* model ima znatno nižu vrijednost kod SR.HSE. Model logističke regresije ostaje najlošiji po pitanju preciznosti za većinu receptora, osim kod NR.PPAR.gamma gdje pokazuje poboljšanje. Model slučajne šume također bilježi vrlo nisku vrijednost preciznosti kod receptora SR.ATAD5.

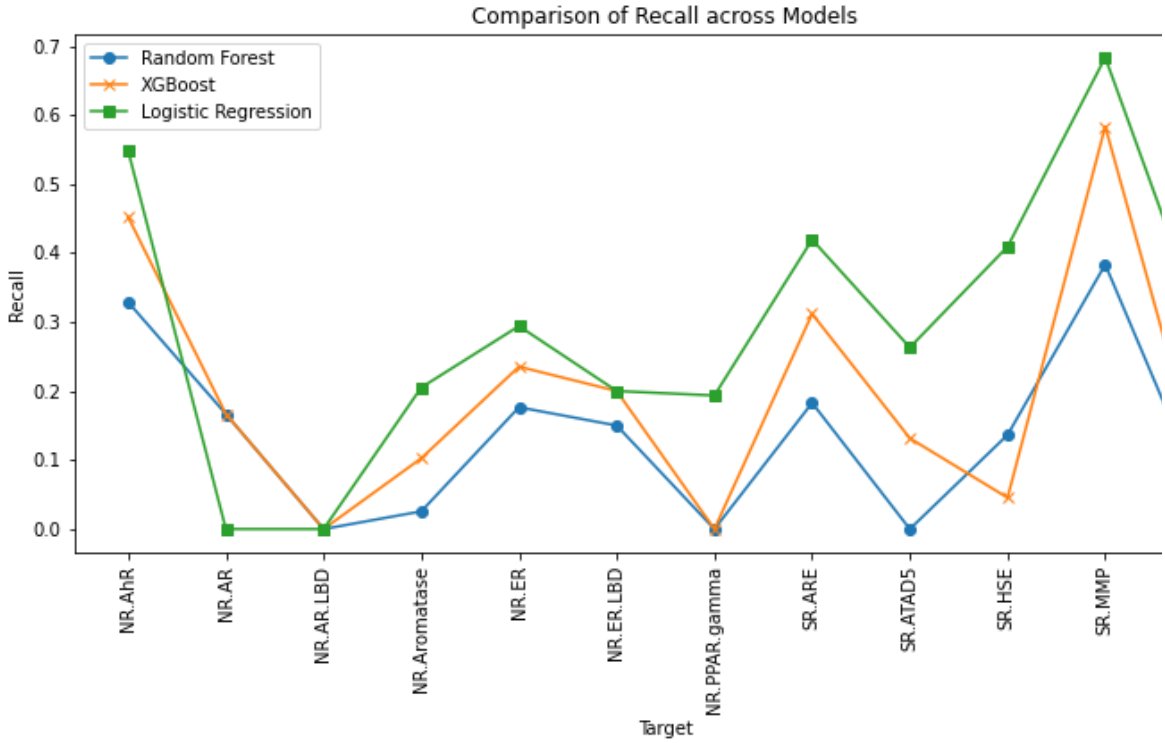



*Graf 5.4 Usporedba preciznosti gdje je skup za treniranje 75%*

Na grafu za 75% trening skup [Graf 5.4], preciznost za modele slučajne šume i *XGBoost* pokazuje konzistentne vrijednosti između 0.8 i 0.9 za većinu receptora. Ovo pokazuje stabilne performanse u prepoznavanju točnih pozitivnih rezultata. S druge strane, model logističke regresije ima značajno niže vrijednosti, krećući se u intervalu od 0.4 do 0.7, što ukazuje na veću učestalost lažno pozitivnih rezultata.

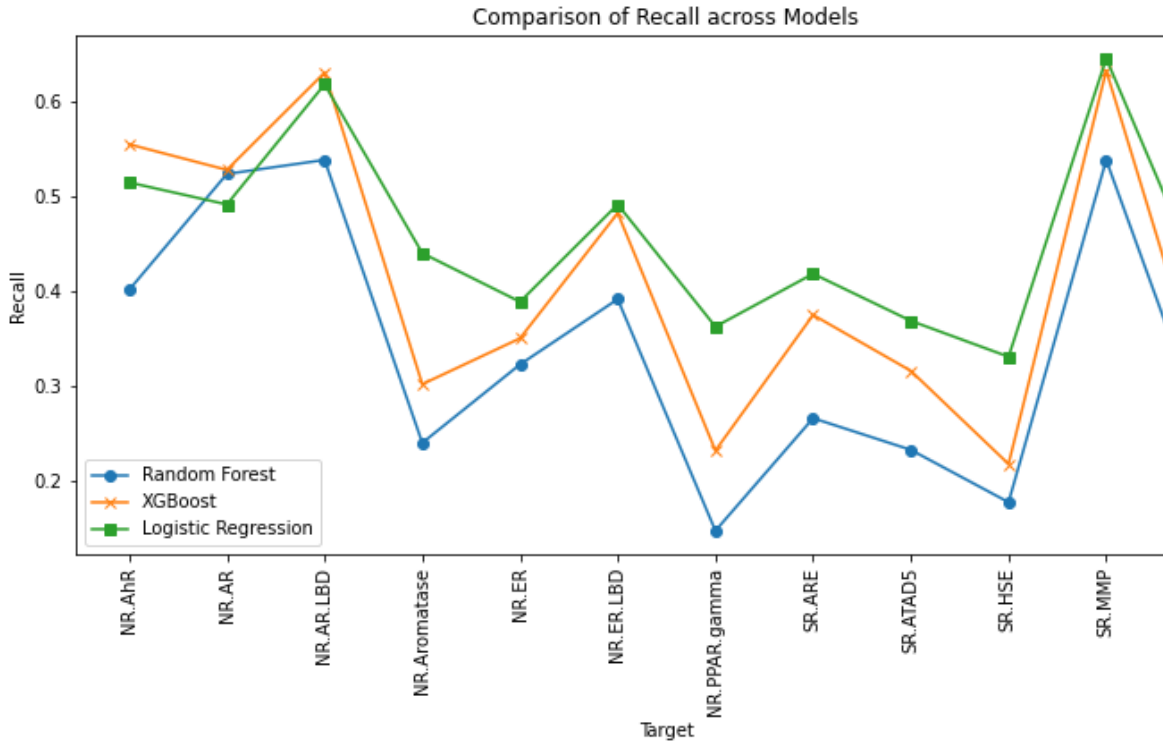
## 5.4 Odziv

Grafički prikaz odziva modela na oba skupa podataka (95% i 75% treninga) pokazuje konzistentne vrijednosti rezultata između modela. Odziv metrike pruža uvid u sposobnost identificiranja modela na sve stvarne pozitivne rezultate, što je ključno za razumijevanje učinkovitosti modela u detekciji ciljnih klasa.




*Graf 5.5 Usporedba vrijednsoti odziva gdje je skup za treniranje 95%*

Na grafu za 95% trening skup [Graf 5.5], odziv je niže, ne prelazeći 0.7. I ovdje model logističke regresije pokazuje najviše vrijednosti odziva, dok su modeli slučajne šume i *XGBoost* nešto niži, ali vrlo blizu. Ova konzistentnost među modelima sugerira da, iako postoje varijacije u točnosti i preciznosti, sposobnost prepoznavanja stvarnih pozitivnih primjera ostaje stabilna.

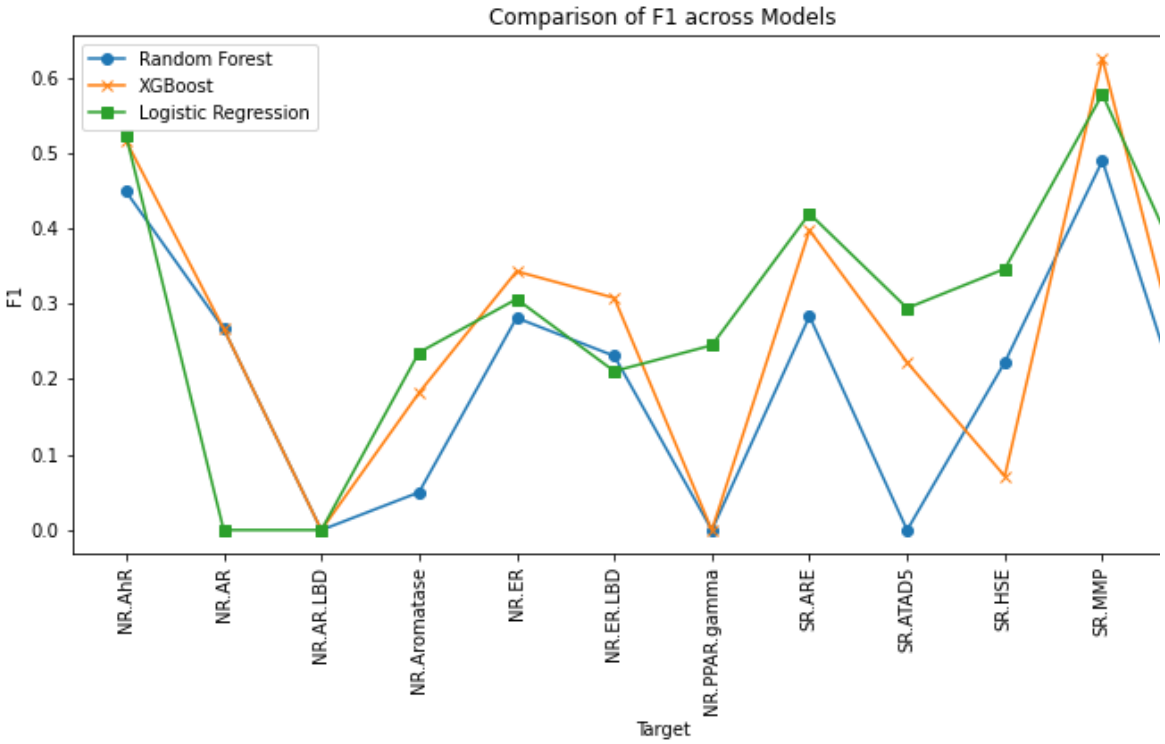



*Graf 5.6 Usporedba vrijednsoti odziva gdje je skup za treniranje 75%*

Na grafu za 75% trening skup [Graf 5.6], odziv je općenito viši u odnosu na 95% skup, krećući se između 0.3 i 0.7. Model logističke regresije pokazuje najviši odziv među modelima, dok su *XGBoost* model(XGB) i model slučajne šume (RF) nešto niži, ali i dalje relativno blizu. Ovo povećanje od 0.1 do 0.2 u odnosu na 95% skup ukazuje na bolju sposobnost prepoznavanja pozitivnih primjera kada se koristi veći udio podataka za trening.

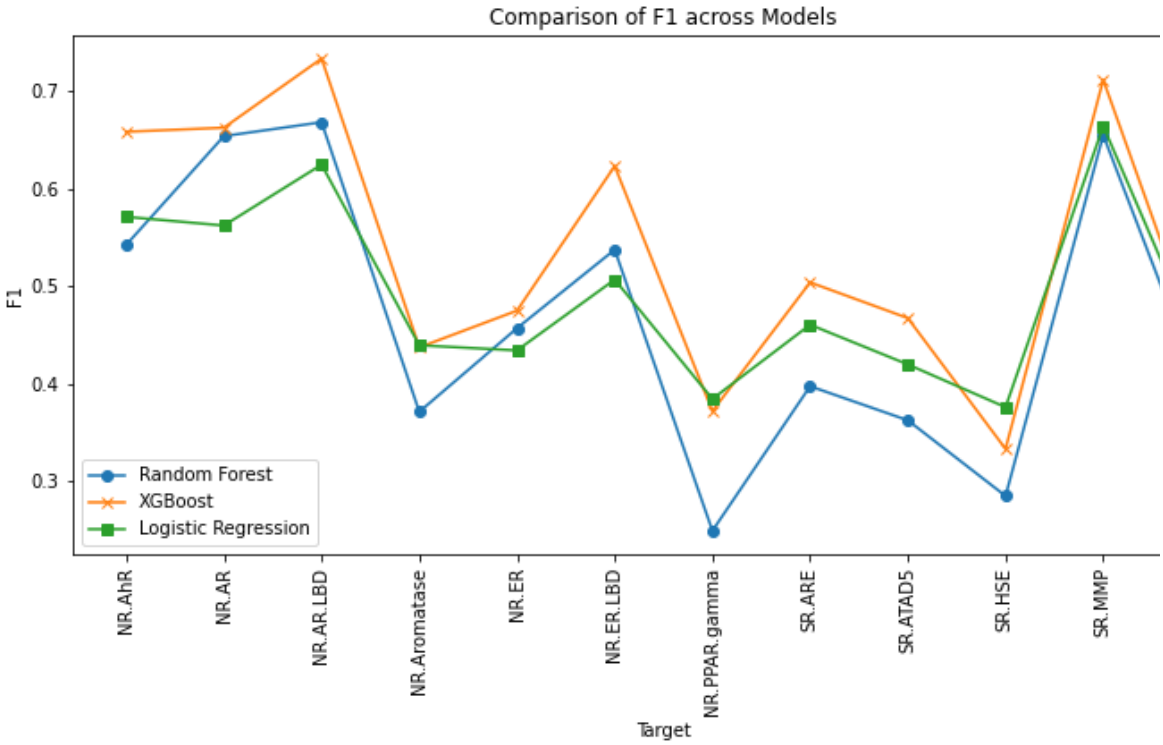
## 5.5 F1-score

F1-score je harmonijska sredina preciznosti i odziva, pružajući uravnotežen pogled na performanse modela. Grafički prikaz F1-score vrijednosti modela na oba skupa podataka (95% i 75% treninga) pokazuje značajno variranje rezultata ovisno o receptorima, ali su općenito dosljedni među modelima.




*Graf 5.7 Usporedba F1-score vrijednsoti gdje je skup za treniranje 95%*

Na grafu za 95% trening skup [Graf 5.7], F1-score vrijednosti su generalno jednake među modelima, ali postoji vidljiva varijacija od receptora do receptora. Model logističke regresije pokazuje malo bolje performanse u usporedbi s drugim modelima. Ova superiornost modela logističke regresije u F1-score sugerira sposobnost održavanja dobre ravnoteže između preciznosti i odziva, čak i uz niže preciznosti i odziv u nekim slučajevima.

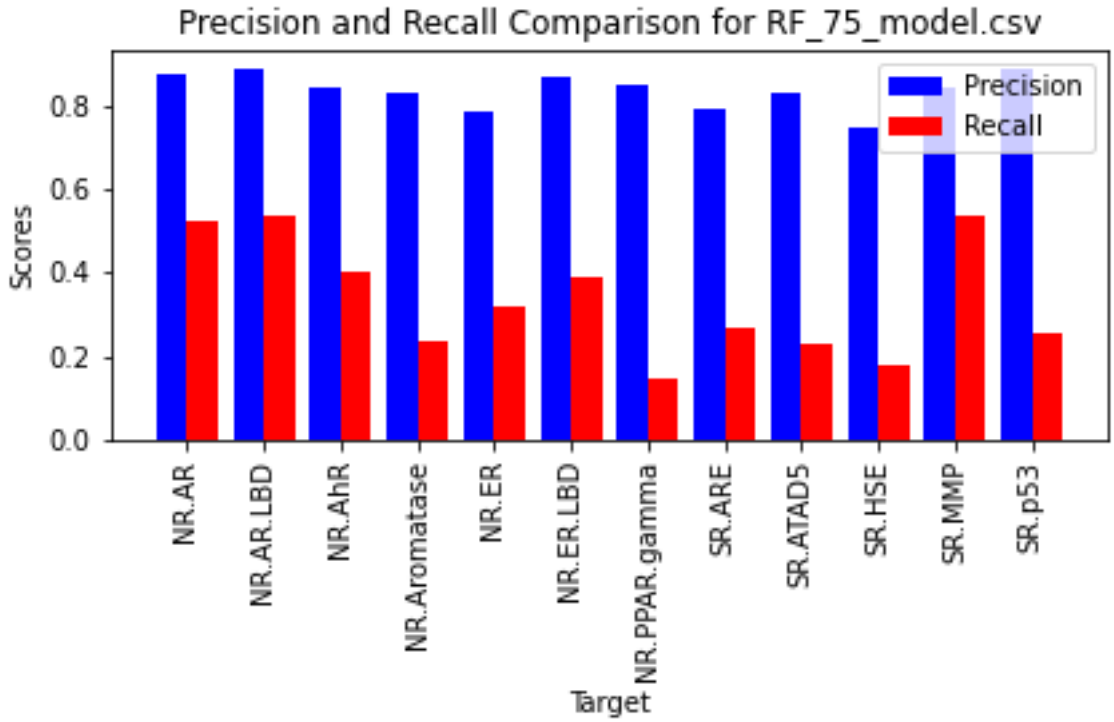



*Graf 5.8 Usporedba F1-score vrijednsoti gdje je skup za treniranje 75%*

Na grafu za 75% trening skup [Graf 5.8], *XGBoost* (XGB) model postiže najbolje F1-score vrijednosti, dok je model slučajne šume (RF) malo lošiji u usporedbi s drugim modelima. Ovdje također postoji značajna varijacija F1-score vrijednosti među receptorima, što ukazuje na bolje performanse modela nad nekim receptorima, u usporedbi s drugima.

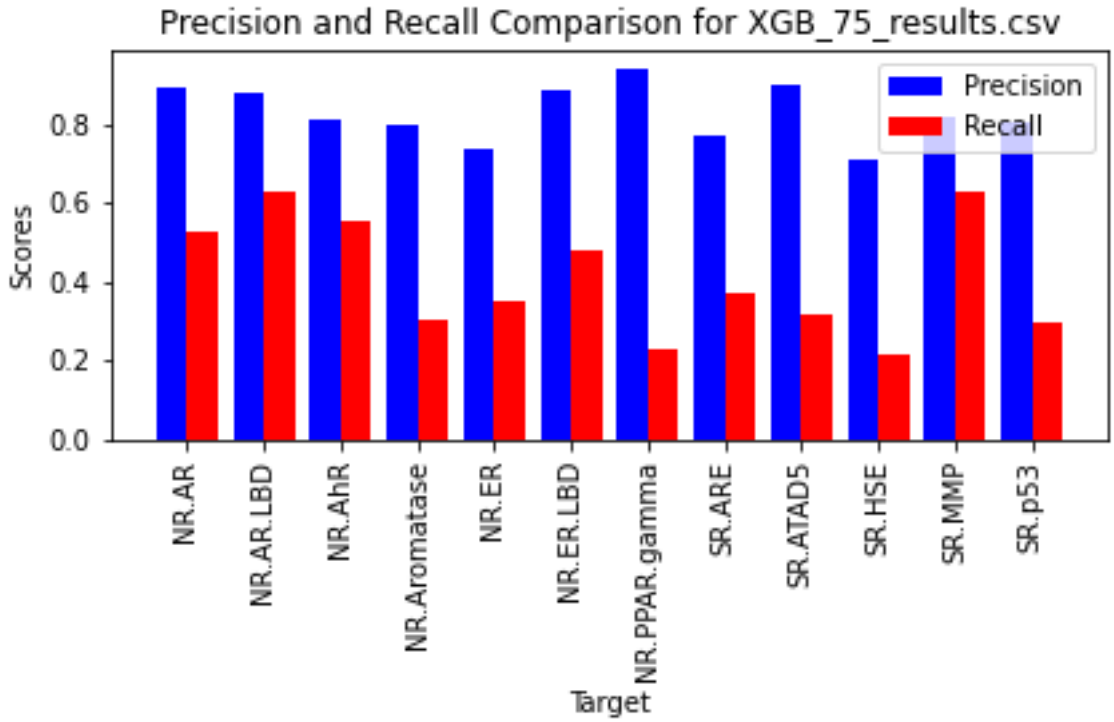
## 5.6 Dodatne Napomene

Stupčasti dijagrami koji uspoređuju vrijednosti preciznosti i odziva za svaki receptor kod modela slučajne šume (RF), *XGBoost* (XGB) i logističke regresije (LR) pružaju dodatne uvide u performanse ovih modela.



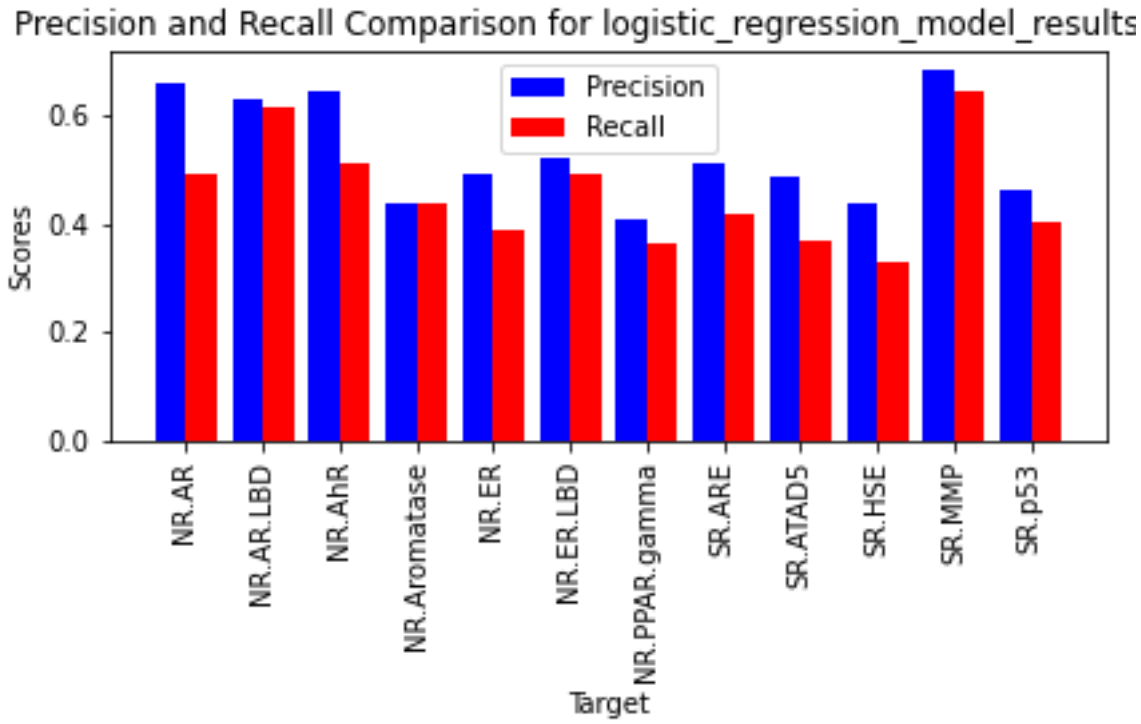

*Graf 5.9 Prikaz usporedbe vrijednosti preciznosti i odziva za model slučajne šume*

U grafu za model slučajne šume [Graf 5.9] jasno se vidi da je vrijednost preciznosti znatno veća od vrijednosti odziva za svaki receptor. Ova značajna razlika između preciznosti i odziva ukazuje na bolje performanse modela slučajne šume pri vrednovanju točnosti predviđanja pozitivnih ishoda, ali ima poteškoće u identificiranju svih stvarnih pozitivnih slučajeva. Na primjer, receptori poput NR.AhR i NR.ER pokazuju preciznost koja je znatno iznad odziva, što potvrđuje ovu analizu.




*Graf 5.10 Prikaz usporedbe vrijednosti preciznosti i odziva za XGBoost model*

Slično kao i kod modela slučajne šume, graf za *XGBoost* model [Graf 5.10] pokazuju znatno veće vrijednosti preciznosti u odnosu na vrijednosti odziva za svaki receptor. To ukazuje na sličan obrazac u kojem *XGBoost* model ima visoku točnost u predviđanju pozitivnih ishoda, ali se suočava s izazovima u potpunom identificiranju svih stvarnih pozitivnih slučajeva. Na primjer, kod receptora NR.AR i SR.ATAD5, preciznost je značajno viša od odziva, što odražava ovu razliku u performansama.




*Graf 5.11 Prikaz usporedbe vrijednosti preciznosti i odziva za model logističke regresije*

Za razliku od modela slučajne šume i *XGBoost*, graf za model logističke regresije [Graf 5.11] pokazuju podjednake vrijednosti preciznosti i odziva za svaki receptor. Ovaj obrazac ukazuje na uravnotežen pristup modela logističke regresije, gdje je sposobnost predviđanja pozitivnih ishoda i identifikacija stvarnih pozitivnih slučajeva gotovo jednaka. Na primjer, kod receptora NR.AR.LBD i NR.PPAR.gamma, preciznost i odziv su gotovo identične, što odražava konzistentnost performansi modela logističke regresije.

Ovi grafički prikazi dodatno potvrđuju opservacije iz prethodnih analiza. Modeli slučajne šume i *XGBoost* pokazuju visoku preciznost uz niži odziv, što ukazuje na njihovu sposobnost točnog predviđanja pozitivnih ishoda, ali s poteškoćama u potpunoj identifikaciji svih pozitivnih slučajeva. S druge strane, model logističke regresije prikazuje uravnotežene performanse, gdje su preciznost i odziv konzistentno jednaki, pružajući stabilan omjer između točnosti i sposobnosti identifikacije svih pozitivnih ishoda.

Dodatno, vrijednosti za izlaz su generalno raspoređene na način gdje se vjerojatnost 1 pojavljuje u prosjeku oko 400 puta za pojedini receptor, s par iznimaka koje dosežu

do 1000. Ostatak od 12.000 uzoraka raspodijeljen je po vrijednosti 0 i NA, pri čemu je vrijednost 0 obično dvostruko češća. Ova distribucija pokazuje značajnu neravnotežu između pozitivnih i negativnih ishoda, što dodatno objašnjava zašto su preciznost i odziv ključne metrike u ocjenjivanju performansi modela.

Ove dodatne napomene pomažu u razumijevanju specifičnih prednosti i ograničenja svakog modela, pružajući dublji uvid u njihove performanse i mogućnosti primjene na različitim skupovima podataka i za različite ciljeve.

# 6. Diskusija

U ovom radu uspoređena su  tri modela strojnog učenja – Slučajna šuma (RF), Logistička Regresija (LR), i *XGBoost* (XGB) – koristeći metrike vrednovanja točnost, preciznost, odziv, i F1-score na skupu podataka koji uključuje različite receptore. Performanse modela analizirane su na originalnom skupu podataka s 95% za treniranje i 75% za treniranje, te dodatno proučavana raspodjela preciznosti i odziva za svaki model i receptor. Također, važno je naglasiti da je podjela od 95% za treniranje bila definirana unutar skupa podataka, dok je podjela od 75% za treniranje bila napravljena tijekom ovog rada kako bi se dobila kvalitetnija očitanja vrednovanja modela.

Rezultati pokazuju da su svi modeli postigli približno jednake vrijednosti točnosti, s modelom logističke regresije koji je imao nešto niže performanse u usporedbi s modelima slučajne šume i *XGBoost*. Na grafovima koji prikazuju točnost za skupove od 95% i 75%, primjetno je da model logističke regresije ima konstantno niže vrijednosti, dok su modeli slučajne šume i *XGBoost* konzistentni u svojim performansama. Kod analize preciznosti, modeli slučajne šume i *XGBoost* pokazali su znatno veće vrijednosti preciznosti u usporedbi s odzivom, što ukazuje na njihovu sposobnost točnog predviđanja pozitivnih ishoda, ali s poteškoćama u identifikaciji svih stvarnih pozitivnih slučajeva. Nasuprot tome, model logističke regresije ima podjednake vrijednosti preciznosti i odziva za svaki receptor, što pokazuje uravnotežene performanse između točnosti i identifikacije pozitivnih ishoda.

Analize odziva upućuju na konzistentne vrijednosti između modela, s modelom logističke regresije koji je postigao najviše vrijednosti odziva, dok su modeli slučajne šume i *XGBoost* imali nešto niže vrijednosti. Iako vrijednosti odziva ne prelaze 0.7, model logističke regresije je u oba skupa pokazao bolje performanse u identifikaciji stvarnih pozitivnih slučajeva.

F1-score analize su pokazale varijacije ovisno o receptoru. U skupu s 95% za treniranje [Graf 5.1], model logističke regresije je imao nešto bolje performanse, dok je u skupu s 75% [Graf 5.2] *XGBoost* model bio najbolji.  Model slučajne šume je imao nešto niže vrijednosti F1-score-a u oba skupa, što ukazuje na potrebu za poboljšanjem uravnoteženosti između preciznosti i odziva.

Dodatne napomene ističu značajne razlike vrijednosti preciznosti i odziva za modele slučajne šume i *XGBoost*, dok su za model logističke regresije podjednake, što dodatno potvrđuje različite prednosti i ograničenja svakog modela. Distribucija izlaznih vrijednosti također pokazuje značajnu neravnotežu, s prosjekom od 400 pojavljivanja vjerojatnosti 1 po receptoru i nekoliko iznimaka koje dosežu do 1000, dok je ostatak od 12.000 uzoraka raspodijeljen po vrijednostima 0 i NA, pri čemu je vrijednost 0 dvostruko češće pojavljuje od vrijednosti NA. Ova distribucija pokazuje značajnu neravnotežu između pozitivnih i negativnih ishoda, što dodatno objašnjava zašto su preciznost i odziv ključne metrike u ocjenjivanju performansi modela.

Ovi rezultati imaju važne implikacije za primjenu ovih modela u praktičnim situacijama gdje je ravnoteža između preciznosti i odziva kritična. Na primjer, u medicinskoj dijagnostici, model logističke regresije mogao bi biti preferiran zbog svoje uravnoteženosti, dok bi modeli slučajne šume i XGBoost mogli biti korisniji u scenarijima gdje je visoka preciznost ključna, poput otkrivanja prijevara.

Za buduća istraživanja preporučuje se daljnja analiza ovih modela na balansiranijim skupovima podataka kako bi se bolje razumjele njihove performanse. Također, istraživanje dodatnih modela i tehnika, poput metoda za balansiranje podataka ili kombiniranja više modela, moglo bi poboljšati ukupne performanse.

# 7. Zaključak

Na temelju provedenih analiza, najbolji model za zadani skup podataka je *XGBoost* (XGB) model. Ovaj izbor temelji se na sljedećim razlozima:

1. **Najbolje performanse u F1-score**: *XGBoost* model je postigao najviše F1-score vrijednosti u skupu s 75% za treniranje, što upućuje na uravnotežene performanse između preciznosti i odziva [Graf 5.7] i [Graf 5.8].
2. **Visoka točnost**: *XGBoost* model je dosljedno pokazao visoku točnost za većinu receptora u oba skupa podataka, čime nadmašuje modele slučajne šume i logističke regresije u pogledu ukupne točnosti predviđanja [Graf 5.1 i Graf 5.2].
3. **Konzistentne performanse**: Iako model slučajne šume pokazuje visoku preciznost, *XGBoost* model je pokazao bolje uravnotežene performanse između preciznosti i odziva, što je ključno za aplikacije koje zahtijevaju visok stupanj točnosti u prepoznavanju svih pozitivnih ishoda.

Usporedba performansi modela ovisno o raspodjeli podataka pokazuje da su rezultati s većim skupom za treniranje (95%) bolji za uravnoteženost performansi među metrikama, dok su rezultati s manjim skupom za treniranje (75%) pokazali bolju sposobnost generalizacije, osobito za *XGBoost* model. To sugerira da *XGBoost* model bolje koristi raspoložive podatke za treniranje, posebno kada je raspoloživo manje podataka.

Ovi nalazi sugeriraju da izbor modela treba biti prilagođen specifičnim potrebama aplikacije, uzimajući u obzir neravnotežu u podacima i specifične zahtjeve za preciznost i odziv. *XGBoost* model se ističe kao najbolji izbor za zadani skup podataka zbog svoje sposobnosti pružanja uravnotežene performanse i visoke točnosti predviđanja, čime osigurava pouzdanost i učinkovitost u detekciji interakcija s različitim receptorima.

# Literatura


[1] Leach, A. R., & Gillet, V. J. **Introduction to Chemoinformatics**. Springer, 2007.

[2] Varnek, A. **Chemoinformatics: Theory, Practice, & Products**. Springer, 2008.

[3] Hodgson, E. **A Textbook of Modern Toxicology**. 4. izdanje. John Wiley & Sons, 2010.

[4] Zhang, C., Cheng, F., Li, W., Liu, G., & Tang, Y. **In silico Prediction of Drug-Induced Toxicity**. *Journal of Chemical Information and Modeling*, 2016. Poveznica: https://pubs.acs.org/doi/abs/10.1021/acs.jcim.5b00568; pristupljeno 9. lipnja 2024.

[5] Mitchell, T. M. **Machine Learning**. McGraw Hill, 1997.

[6] Bishop, C. M. **Pattern Recognition and Machine Learning**. Springer, 2006.

[7] Šnajder, J. **Strojno učenje**. Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva. Poveznica: https://www.fer.unizg.hr/predmet/struce1; pristupljeno 15. lipnja 2024.

[8] [Mayr2016] Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2016). DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science*, **3**:80.

[9] [Huang2016] Huang, R., Xia, M., Nguyen, D. T., Zhao, T., Sakamuru, S., Zhao, J., Shahane, S., Rossoshek, A., & Simeonov, A. (2016). Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science*, **3**:85.

[10] National Toxicology Program. **Tox21: Transforming Environmental Health Testing**. Poveznica: https://ntp.niehs.nih.gov/whatwestudy/tox21; pristupljeno 10. lipnja 2024.

[11] NIH Chemical Genomics Center. **Tox21 Challenge**. Poveznica: https://tripod.nih.gov/tox21/challenge/; pristupljeno 10. lipnja 2024.

[12] Institute of Bioinformatics, Johannes Kepler University. **DeepTox: Toxicity Prediction**. Poveznica: http://bioinf.jku.at/research/DeepTox/tox21.html; pristupljeno 11. lipnja 2024.

[13] Chemicalize.org. **Chemicalize**. Poveznica: https://chemicalize.com; pristupljeno 12. lipnja 2024.

[14] National Center for Biotechnology Information. **PubChem**. Poveznica: https://pubchem.ncbi.nlm.nih.gov; pristupljeno 8. lipnja 2024.

[15] OpenBabel. **Open Babel: The Open Source Chemistry Toolbox**. Poveznica: http://openbabel.org; pristupljeno 5. lipnja 2024.

[16] U.S. National Library of Medicine. **TOXNET**. Poveznica: https://toxnet.nlm.nih.gov; pristupljeno 17.travnja 2024.

[17] National Toxicology Program. **NTP Tox21**. Poveznica: https://ntp.niehs.nih.gov/whatwestudy/tox21; pristupljeno 9. lipnja 2024.

[18] Breiman, L. Random Forests. Machine Learning, 45,1 (2001).

[19] Chen, T., Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, (2016).

[20] XGBoost Documentation. Poveznica: https://xgboost.readthedocs.io; pristupljeno 5. lipnja 2024.

# Sažetak

Ovaj rad istražuje primjenu različitih modela strojnog učenja za binarnu klasifikaciju kemijskih spojeva u kontekstu njihove interakcije s određenim receptorima, koristeći Tox21 skup podataka. Istražena je učinkovitost triju modela: logistička regresija, slučajna šuma i XGBoost.

Logistička regresija procjenjuje vjerojatnost binarnih ishoda pomoću logističke funkcije. Model slučajne šume koristi mnoga stabla odlučivanja i kombinira njihove predikcije većinskim glasanjem. XGBoost iterativno trenira modele kako bi ispravio pogreške prethodnih, minimizirajući funkciju gubitka.

Podaci su očišćeni, podijeljeni na trening i test skupove, te su modeli trenirani i testirani koristeći metrike točnosti, preciznosti, odziva i F1-score-a. Eksperimenti su provedeni na dvije podjele podataka: 95% za treniranje i 5% za testiranje, te 75% prema 25%.

Rezultati pokazuju da je XGBoost model postigao najviše točnosti za većinu receptora u oba eksperimentalna postava, dok je model slučajne šume također pokazao visoke performanse, ali s većom varijabilnošću. Logistička regresija je općenito imala najniže performanse.

## Summary

This paper investigates the application of various machine learning models for binary classification of chemical compounds in the context of their interaction with specific receptors, using the Tox21 dataset. The effectiveness of three models was explored: logistic regression, random forest, and XGBoost.

Logistic regression estimates the probability of binary outcomes using the logistic function. The random forest model employs multiple decision trees and combines their predictions through majority voting. XGBoost iteratively trains models to correct the errors of previous ones, minimizing the loss function.

The data were cleaned and split into training and test sets, and the models were trained and tested using accuracy, precision, recall, and F1-score metrics. Experiments were conducted on two data splits: 95% for training and 5% for testing, and 75% for training and 25% for testing.

The results show that the XGBoost model achieved the highest accuracy for most receptors in both experimental setups, while the random forest model also demonstrated high performance but with greater variability. Logistic regression generally had the lowest performance.