Primjena strojnog učenja u predviđanju globalne prodaje video igara

Prpić, Nikola

SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA

Sveučilišni studij

PRIMJENA STROJNOG UČENJA U PREDVIĐANJU GLOBALNE PRODAJE VIDEO IGARA

Diplomski rad

Nikola Prpić

Osijek, 2024.

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA **OSIJEK**

Obrazac D1: Obrazac za ocjenu diplomskog rada na sveučilišnom diplomskom studiju

Ocjena diplomskog rada na sveučilišnom diplomskom studiju

Ime i prezime pristupnika:	Nikola Prpić
Studij, smjer:	Sveučilišni diplomski studij Računarstvo
Mat. br. pristupnika, god.	D1318R, 07.10.2022.
JMBAG:	0165082386
Mentor:	izv. prof. dr. sc. Časlav Livada
Sumentor:	doc. dr. sc. Ivana Hartmann Tolić
Sumentor iz tvrtke:	
Predsjednik Povjerenstva:	prof. dr. sc. Krešimir Nenadić
Član Povjerenstva 1:	izv. prof. dr. sc. Časlav Livada
Član Povjerenstva 2:	izv. prof. dr. sc. Alfonzo Baumgartner
Naslov diplomskog rada:	Primjena strojnog učenja u predviđanju globalne prodaje video igara
Znanstvena grana diplomskog rada:	**Umjetna inteligencija (zn. polje računarstvo)**
Zadatak diplomskog rada:	Zadatak je kreirati model za predviđanje prodaje igara koristeći strojno učenje. Predviđanje prodaje video igara može unaprijed prilagoditi prodajne strategije i razvojne planove pojedine video igre. Potrebno je implementirati model strojnog učenja koji može predvidjeti globalnu prodaju video igara ovisno o određenim značajkama podataka iz baze. Uspješnom primjenom tehnika strojnog učenja moguće je pomoći razvojnom programeru video igre da se fokusira na određenu vrstu video igre. Sumentor s FERIT-a: Ivana Hartmann Tolić Tema rezervirana za studenta:
Datum ocjene pismenog dijela diplomskog rada od strane mentora:	24.09.2024.
Ocjena pismenog dijela diplomskog rada od strane mentora:	Vrlo dobar (4)
Datum obrane diplomskog rada:	03.10.2024.
Ocjena usmenog dijela diplomskog rada (obrane):	Vrlo dobar (4)
Ukupna ocjena diplomskog rada:	Vrlo dobar (4)
Datum potvrde mentora o predaji konačne verzije diplomskog rada čime je pristupnik završio sveučilišni diplomski studij:	08.10.2024.

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA **OSIJEK**

IZJAVA O IZVORNOSTI RADA

Osijek, 08.10.2024.

Ime i prezime Pristupnika:	Nikola Prpić
Studij:	Sveučilišni diplomski studij Računarstvo
Mat. br. Pristupnika, godina upisa:	D1318R, 07.10.2022.
Turnitin podudaranje [%]:	6

Ovom izjavom izjavljujem da je rad pod nazivom**: Primjena strojnog učenja u predviđanju globalne prodaje video igara**

izrađen pod vodstvom mentora izv. prof. dr. sc. Časlav Livada

i sumentora doc. dr. sc. Ivana Hartmann Tolić

moj vlastiti rad i prema mom najboljem znanju ne sadrži prethodno objavljene ili neobjavljene pisane materijale drugih osoba, osim onih koji su izričito priznati navođenjem literature i drugih izvora informacija.
Izjavljujem da je intelektualni sadržaj navedenog rada proizvod mog vlastitog rada, osim u onom dijelu za koji mi je bila potrebna pomoć mentora, sumentora i drugih osoba, a što je izričito navedeno u radu.

Potpis pristupnika:

SADRŽAJ

1. UVOD

U današnjem digitalnom dobu, industrija video igara postala je jedna od najdinamičnijih i najprofitabilnijih oblika zabave diljem svijeta. Kako se tržište video igara nastavlja rasti i razvijati zbog tehnoloških napredaka, širokog pristupa internetu i proširenih opcija igranja, postaje važno za programere, izdavače i distributere stvoriti odgovarajuće platforme kako bi učinkovito predvidjeli globalnu prodaju video igara. To će im omogućiti povećanje profitabilnosti i pobjedu na tržištu. Strojno učenje, kao grana umjetne inteligencije, pruža moćne alate i tehnike za analizu podataka i predviđanje trendova. U slučaju video igara, upotreba strojnog učenja omogućuje dublje razumijevanje ponašanja potrošača, utjecaja različitih faktora na prodaju igara i ključnih čimbenika uspjeha na tržištu.

Ovaj rad istražuje upotrebu strojnog učenja u predviđanju globalne prodaje video igara. Ciljevi su analizirati različite tehnike strojnog učenja i njihove primjene u istraživanju tržišta video igara, identificirati varijable i relevantne faktore koji utječu na prodaju igara, te razviti referentne modele koji će pomoći tvrtkama da donose informirane strateške odluke u distribuciji, marketingu i proizvodnji svojih proizvoda. Očekuje se da će istraživanje u ovom području pridonijeti razvoju metoda i alata za predviđanje prodaje video igara, kao i korisne smjernice za tvrtke koje žele poboljšati učinkovitost i postići konkurentsku prednost na tržištu - potaknut će inovacije i nove strategije poslovnog razvoja.

1.1. Zadatak diplomskog rada

Zadatak je kreirati model za predviđanje prodaje igara koristeći strojno učenje. Predviđanje prodaje video igara može unaprijed prilagoditi prodajne strategije i razvojne planove pojedine video igre. Potrebno je implementirati model strojnog učenja koji može predvidjeti globalnu prodaju video igara ovisno o određenim značajkama podataka iz baze. Uspješnom primjenom tehnika strojnog učenja moguće je pomoći razvojnom programeru video igre da se fokusira na određenu vrstu video igre.

2. RAZRADA TEME

Strojno učenje predstavlja jezgru današnjih tehnoloških inovacija, pružajući moćne alate i tehnike za analizu podataka i prediktivno modeliranje [1]. Ova grana umjetne inteligencije omogućuje računalima da uče iz iskustva i prilagode svoje ponašanje kako bi bili učinkovitiji u zadacima koji nedostaju eksplicitnu strukturu [2]. Jedan od najvažnijih mehanizama učenja je neuronska aktivnost, koja je poticana strukturom ljudskog mozga [3]. Neuronske mreže omogućuju računalima da obrađuju kompleksne podatke koristeći slojeve neurona koji međusobno komuniciraju i izvode različite zadatke poput klasifikacije, regresije i prepoznavanja uzoraka [4]. Duboko učenje je posebna vrsta neuronskih mreža koja koristi višestruke slojeve kako bi naučila reprezentacije podataka na različitim apstraktnim razinama [3]. Ovaj pristup revolucionirao je mnoga područja, uključujući prepoznavanje slika, obradu prirodnog jezika, sustave preporuka i mnoge druge.

Osim neuronskih mreža i dubokog učenja, postoje mnogi drugi načini strojnog učenja za različite svrhe. To uključuje dizajne za nadzirano i nenadzirano učenje, reprezentativne modele poput potpornih vektorskih strojeva, k-srednjih vrijednosti klasteriranja i algoritme klasifikacije.

U ovom radu istražit će se principi i primjene tehnika strojnog učenja, uključujući neuronske mreže, duboko učenje te njihovu ulogu u detaljnoj prognozi globalne prodaje video igara. Istražit će se kako ove tehnike mogu pomoći razumijevanju ponašanja potrošača, identificiranju ključnih faktora uspjeha na tržištu i razvoju učinkovitijih strategija za proizvodnju, marketing i distribuciju video igara.

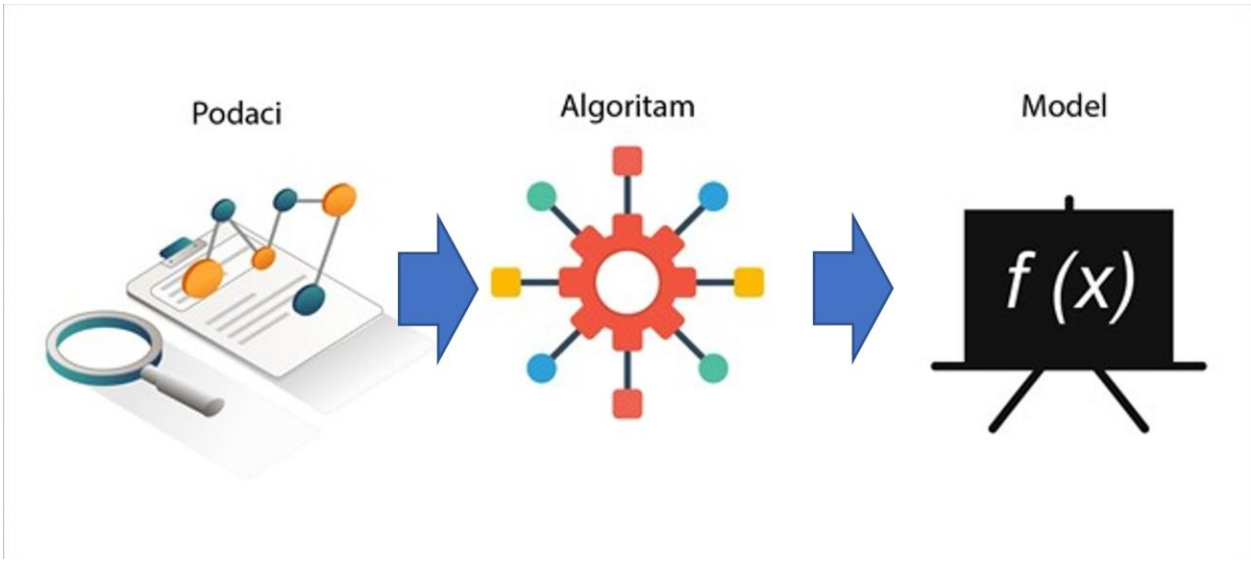
2.1. Strojno učenje

Strojno učenje (engl. *Machine Learning,* ML) je samo dio onoga što sustav treba kako bi postao umjetna inteligencija. Omogućuje umjetnoj inteligenciji obavljanje sljedećih zadataka: prilagođavanje novim okolnostima koje izvorni programer nije predvidio, detektiranje uzoraka u svim vrstama izvora podataka, stvaranje novih ponašanja na temelju prepoznatih uzoraka, donošenje odluka na temelju uspjeha ili neuspjeha tih ponašanja. Strojno učenje počiva na korištenju algoritama za manipulaciju podacima. Za postizanje uspjeha, sesija strojnog učenja zahtijeva primjenu odgovarajućeg algoritma kako bi se ostvarili željeni rezultati. Osim toga, podaci moraju biti prikladni za analizu odabranim algoritmom ili zahtijevaju pažljivu pripremu od strane stručnjaka [5].

Prema [6], strojno učenje je proces programiranja računala kako bi optimiziralo određeni kriterij uspješnosti putem analize podatkovnih primjera ili prethodnog iskustva. Koristi se model koji je

definiran do na neke parametre te se kroz učenje optimiziraju ti parametri temeljem dostupnih podataka. Na osnovu ovih podataka, model treba biti sposoban predvidjeti svojstva novih, još neviđenih podataka. Cilj strojnog učenja je razviti modele koji uspješno generaliziraju, odnosno pravilno funkcioniraju i na podacima koji nisu bili korišteni za njihovo treniranje.

Slikom 2.1. je prikazan temelj strojnog učenja koji se sastoji od tri glavna elementa: podataka, algoritma (postupka) te modela.



Slika 2.1. Temelj strojnog učenja. [6]

Skup podataka je zbirka zapisa. Svaki skup podataka sastoji se od nekoliko redaka i stupaca. Svaki stupac predstavlja nekoliko različitih aspekata skupa podataka [7].

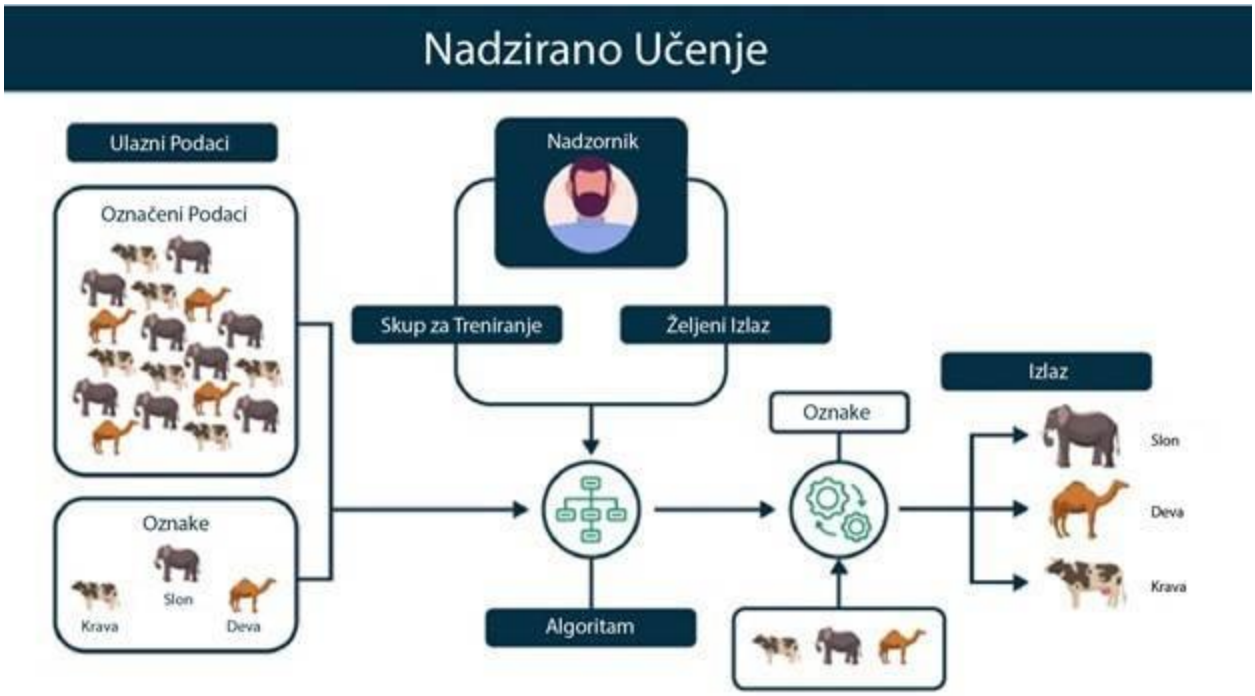
Sama bit strojnog učenja leži u algoritmima. Algoritam, točnije proces ili postupak, temelj je rješavanja problema. Iako priroda problema određuje odgovarajući algoritam, osnovne pretpostavke ostaju prilično konstantne - riješiti određeni problem. Možemo ih zamisliti kao kontejnere - opisuju dobro osmišljen pristup rješavanju problema. Algoritam obrađuje podatke kroz jasno definirane korake, s ciljem generiranja rješenja za problem koji je pred nama. Iako neki algoritmi uzimaju u obzir ulaze kako bi oblikovali izlaz, cilj ostaje na željenom izlazu. Kako bi olakšali ovaj proces, algoritmi izražavaju prijelaze između stanja koristeći formalni jezik koji računala razumiju. Dok algoritam obrađuje podatke i rješava problem, priprema i izvršava zadatak specifično dizajniran za rješavanje problema [5].

Model strojnog učenja može biti matematički izraz, jednadžba, složena struktura podataka iz teorije računarstva, ili kombinacija svega navedenog. On predstavlja spoj statistike, računarstva i programskog inženjerstva. Model može učiti iz ljudskih ili prirodnih postupaka te može simulirati

buduće ponašanje u nepoznatoj situaciji. Drugim riječima, model može predvidjeti buduće događaje. Oni mogu učiti iz povijesti postupaka koji su pohranjeni kao zapisi u bazi podataka [7].

Bez obzira na vrstu algoritma i modela, postoje nekoliko osnovnih pristupa strojnom učenju koji ovise o vrsti i količini podataka. To uključuje: nadzirano, nenadzirano i podržano/ojačano učenje.

Nadzirano učenje (engl. *Supervised learning*) uključuje obuku stroja na označenim podacima. Označeni podaci sastoje se od primjera s točnim odgovorom ili klasifikacijom. Stroj uči odnos između ulaznih i izlaznih podataka. Obučeni stroj tada može predviđati izlaz za nove, neoznačene podatke [8]. Primjer ovakvog učenja prikazan je slikom 2.2.. Nadzirano učenje klasificira se u dvije kategorije algoritama – regresija i klasifikacija.



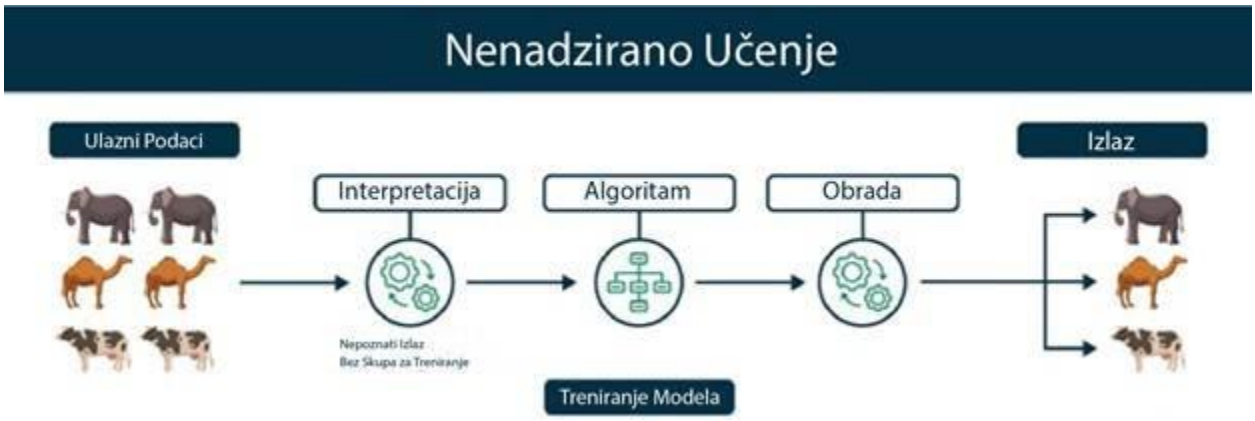
Slika 2.2. Nadzirano učenje. [8]

Regresija (engl. *Regression*) je vrsta nadziranog učenja koja se koristi za predviđanje kontinuiranih vrijednosti, poput cijena kuća, cijena dionica ili gubitka korisnika. Algoritmi regresije uče funkciju koja mapira ulazne značajke na izlaznu vrijednost, a uključuju: linearnu regresiju, polinomijalnu regresiju, regresiju potpornih vektora, regresiju odlučivanja stabla, regresiju slučajne šume [8].

Klasifikacija je vrsta nadziranog učenja koja se primjenjuje za predviđanje kategoričkih vrijednosti, kao što su korisnikova odjava, identifikacija neželjene pošte u e-pošti ili dijagnosticiranje tumora na medicinskoj slici. Algoritmi klasifikacije uče funkciju koja preslikava ulazne značajke na vjerojatnosnu distribuciju izlaznih klasa, a uključuju: logističku regresiju, potporne vektore, stabla odlučivanja, slučajne šume, naivni Bayes [8].

Primjene nadziranog učenja uključuju filtriranje neželjenih pošti, klasifikaciju slika, medicinsku dijagnostiku, otkrivanje prijevara i obradu prirodnog jezika. Nadalje, prednosti nadziranog učenja leže u mogućnosti optimizacije kriterija performansi uz pomoć iskustva, sposobnosti rješavanja različitih vrsta problema u stvarnom svijetu te omogućavanju procjene ili mapiranja rezultata na nove uzorke. Također, nadzirano učenje omogućuje kontrolu nad brojem klasa koje želimo u skupu podataka za obuku. S druge strane, nedostaci nadziranog učenja uključuju izazove u klasificiranju velikih skupova podataka, potrebu za dugim vremenom obuke modela, ograničenja u rješavanju svih složenih zadataka u strojnom učenju te potrebu za označenim skupovima podataka i složenim procesom obuke [8].

Nenadzirano učenje (engl. *Unsupervised learning*) je vrsta strojnog učenja koji se temelji na podacima koji nisu označeni. To znači da podaci ne posjeduju prethodno dodijeljene oznake ili kategorije. Glavni cilj nenadziranog učenja je otkrivanje uzoraka i veza u podacima bez bilo kakvog eksplicitnog vođenja. Kod nenadziranog učenja, stroj se trenira na osnovu informacija koje nisu klasificirane ili označene, omogućavajući algoritmu da istražuje te informacije bez vanjskog uplitanja. U ovom scenariju, stroj je zadužen za grupiranje nesortiranih podataka prema njihovim sličnostima, obrascima i razlikama, a sve to bez prethodne obuke na podacima. Za razliku od nadziranog učenja kod kojeg postoji učitelj koji usmjerava proces obuke, u nenadziranom učenju nema takve vanjske pomoći. Stoga, stroj samostalno istražuje skrivene strukture u podacima koji nisu označeni [8]. Slikom 2.3. se objašnjava što predstavlja takvo učenje. Dijeli se u dvije kategorije – grupiranje i asocijacija.



Slika 2.3. Nenadzirano učenje. [8]

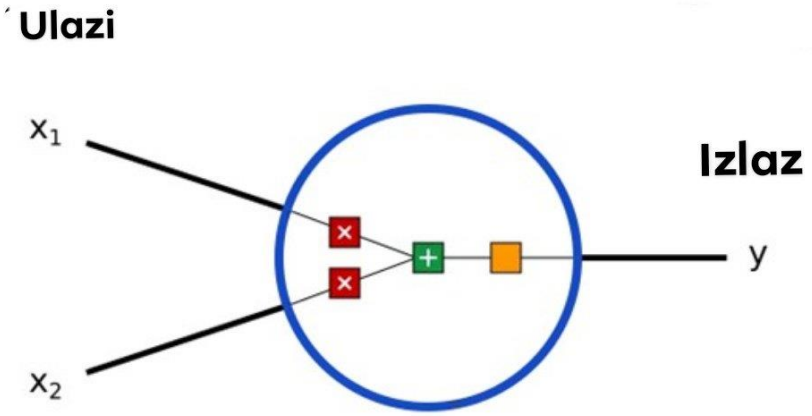
Grupiranje je vrsta nenadziranog učenja koja se koristi za otkrivanje prirođenih grupiranja u podacima, kao što je grupiranje kupaca prema njihovom obrascu kupovine. Asocijacija predstavlja problem učenja pravila koja opisuju velike dijelove vaših podataka, poput situacija kada ljudi koji kupuju proizvod X također često kupuju proizvod Y [8].

Neke od prednosti nenadziranog učenja: ne zahtijeva označene podatke za obuku, moguće je pronaći prethodno nepoznate uzorke u podacima, pomoć u dobivanju uvida iz neoznačenih podataka koje inače možda ne biste mogli dobiti, dobro je u pronalaženju uzoraka i odnosa u podacima bez navođenja što tražiti(može pomoći u učenju novih stvari o svojim podacima) [8]. S druge strane teško je mjeriti točnost ili učinkovitost zbog nedostatka prethodno definiranih odgovora tijekom obuke, rezultati često imaju manju točnost, može biti osjetljivo na kvalitetu podataka, uključujući nedostajuće vrijednosti, izvanredne vrijednosti i bučne podatke.

2.2. Neuronske mreže

Neuronske mreže predstavljaju ključan koncept u strojnome učenju, a njihova struktura i način funkcioniranja inspirirani su načinom na koji ljudski mozak obrađuje informacije. Ove mreže koriste analogiju s biološkim neuronima kako bi stvorile modele koji su sposobni učiti iz podataka i donositi složene odluke. Ljudski mozak sadrži milijarde neurona koji su povezani u mrežu, a svaka veza između neurona omogućuje prijenos informacija. Sličan princip koristi se i u neuronskim mrežama gdje umjetni neuroni međusobno komuniciraju kroz slojeve kako bi obradili podatke, učili iz primjera, i donosili predviđanja [3]. Neuronske mreže omogućuju računalima da obrađuju kompleksne podatke koristeći slojeve neurona koji međusobno komuniciraju i izvode različite zadatke poput klasifikacije, regresije i prepoznavanja uzoraka. Svaki neuron u mreži prima ulazne informacije, obrađuje ih kroz aktivacijske funkcije i prosljeđuje rezultate dalje kroz mrežu, omogućujući tako modelu da se prilagodi i nauči iz iskustva [4].

Prije nego što se započne detaljnije o neuronskim mrežama, potrebno je razumjeti neurone – osnovnu jedinicu neuronske mreže [9]. Neuron prima ulazne podatke od drugih neurona ili okoline, obrađuje ih pomoću matematičkih operacija te generira izlazni signal koji se onda predaje drugim neuronima ili dalje u sustav. Slika 2.4. prikazuje neuron s dva ulazna podatka. Takav proces prijenosa informacija od ulaznih podataka pa sve do izlaza čini osnovnu jedinicu obrade podataka kod neuronskih mreža.



Slika 2.4. Neuroni. [9]

Kod ovog jednostavnog primjera se događaju tri stvari. Prvo, svaki ulaz se množi s težinom. Matematički bi to izgledalo ovako:

$$x_1 \rightarrow x_1 * w_1$$

$$x_2 \rightarrow x_2 * w_2$$

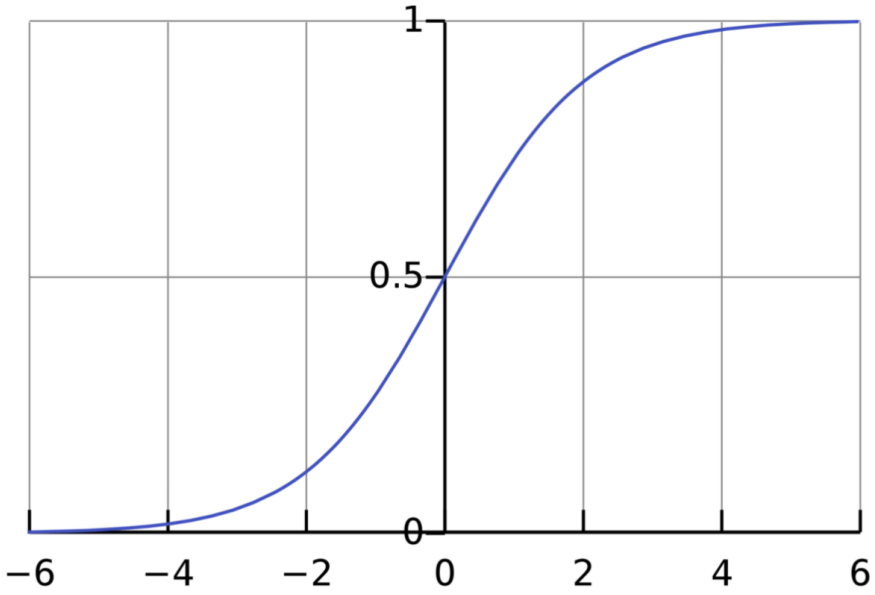
Zatim, svi umnoženi ulazi se zbroje s pomakom b koji omogućuje modelu da bolje prilagodi izlazne vrijednosti:

$$(x_1 * w_1) + (x_2 * w_2) + b$$

I konačno, taj zbroj se prosljeđuje aktivacijskoj funkciji koja ima izraz [5]:

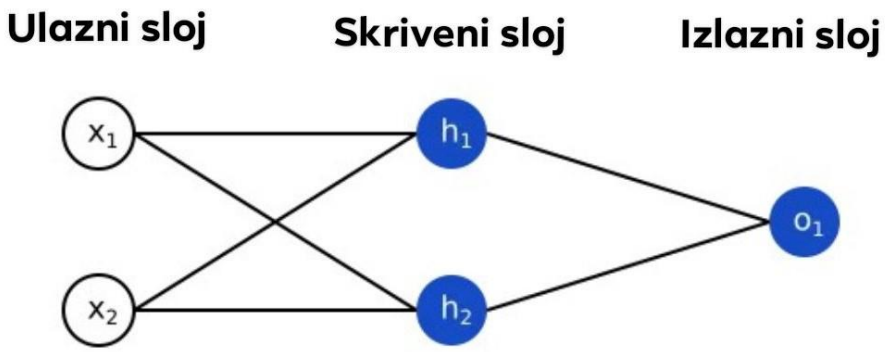
$$y = f(x_1 * w_1 + x_2 * w_2 + b)$$

Aktivacijska funkcija koristi se kako bi se neograničeni ulaz transformirao u izlaz s predvidivim oblikom. Često korištena aktivacijska funkcija je sigmoidna funkcija, koja ograničava izlaz u rasponu <0,1>. To znači da veliki negativni brojevi postaju blizu 0, dok veliki pozitivni brojevi postaju blizu 1 [9]. Primjer takve funkcije je prikazan slikom 2.5..



Slika 2.5. Sigmoidna funkcija. [9]

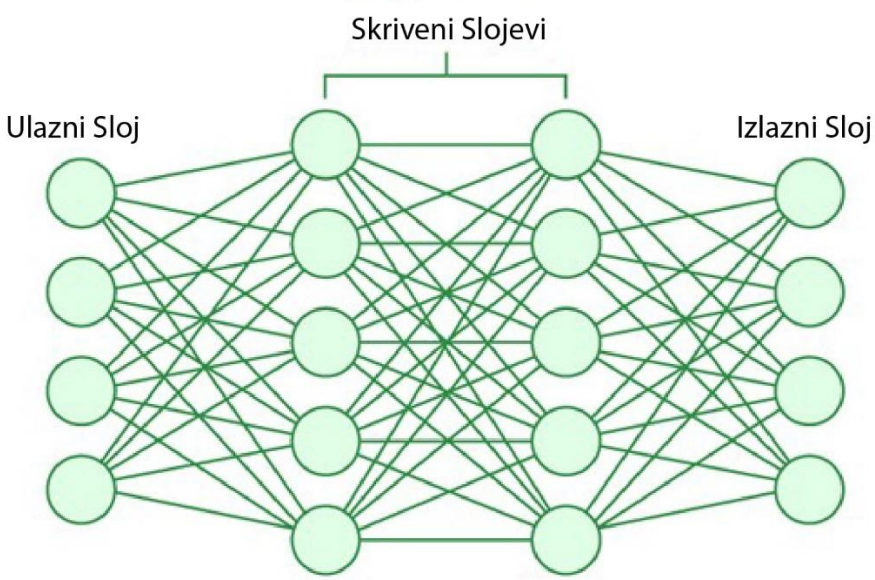
Neuronske mreže (engl. *Neural networks,* NN) su ništa drugo nego skup neurona povezanih zajedno. Primjer jedne neuronske mreže je dan slikom 2.6.. Sastoji se od dva ulaza, skrivenog sloja s dva neurona (h_1, h_2) te izlaznog sloja s jednim neuronom. Ovdje treba obratiti pažnju da su ulazi u o_1 zapravo izlazi h_1 i h_2. Upravo to čini ovo mrežom [9].



Slika 2.6. Primjer neuronske mreže. [9]

U kontekstu neuronskih mreža, važno je istaknuti ulogu umjetnih neuronskih mreža, koje su osmišljene prema biološkim neuronima, ali se primjenjuju u raznolikim područjima suvremenih računalnih tehnologija. Umjetne neuronske mreže (engl. *Artificial Neural Network,* ANN) koriste umjetne neurone – jedinice. One su organizirane u slojevitu strukturu koja čini umjetnu neuronsku mrežu u sustavu. Svaki sloj može sadržavati samo nekoliko desetaka ili milijune jedinica, ovisno o potrebama složenih neuronskih mreža za učenje skrivenih obrazaca u skupu podataka.

Uobičajeno, umjetna neuronska mreža sastoji se od ulaznog sloja, izlaznog sloja te skrivenih slojeva [10]. Slikom 2.7. je prikazana arhitektura umjetne neuronske mreže.



Slika 2.7. Umjetna neuronska mreža. [10]

Struktura i funkcioniranje ljudskih neurona služe kao temelj umjetnim neuronskim mrežama. Ulazni sloj umjetne neuronske mreže, prvi sloj, prima podatke iz vanjskih izvora te ih prosljeđuje skrivenom sloju, drugom sloju. U skrivenom sloju, svaki neuron prima ulaz od prethodnih neurona u istom sloju, izračunava ponderirani zbroj te ga prenosi neuronima u sljedeći sloj. Ponderirane veze optimiziraju se dodjeljivanjem različitih težina svakom ulazu, prilagođavajući se tijekom obuke radi poboljšanja performansi modela [10].

Umjetne neuronske mreže temelje se na biološkim neuronima pronađenima u životinjskim mozgovima, pa dijele mnoge sličnosti u strukturi i funkciji [10]. Struktura umjetnih neuronskih mreža modelirana je prema biološkim neuronima. Biološki neuron sastoji se od tijela (soma) za obradu impulsa, dendrita za primanje impulsa te aksona koji ih prenosi. Ulazni čvorovi umjetnih neuronskih mreža primaju ulazne signale, čvorovi skrivenog sloja obrađuju ove signale, a čvorovi izlaznog sloja računaju konačni izlaz koristeći aktivacijske funkcije. Sinapse omogućuju prijenos impulsa između bioloških neurona, dok u umjetnim neuronima povezuju čvorove slojeva s težinama koje određuju jačinu veze. Učenje u biološkim neuronima događa se putem sinaptičke plastificiranosti, dok u umjetnim neuronima koristi se algoritam propagacije pogreške unatrag (engl. *backpropagation*) za prilagodbu težina prema pogrešci. Aktivacija neurona u umjetnim

neuronskim mrežama ostvaruje se korištenjem matematičke aktivacijske funkcije koja preslikava ulaz u izlaz [10].

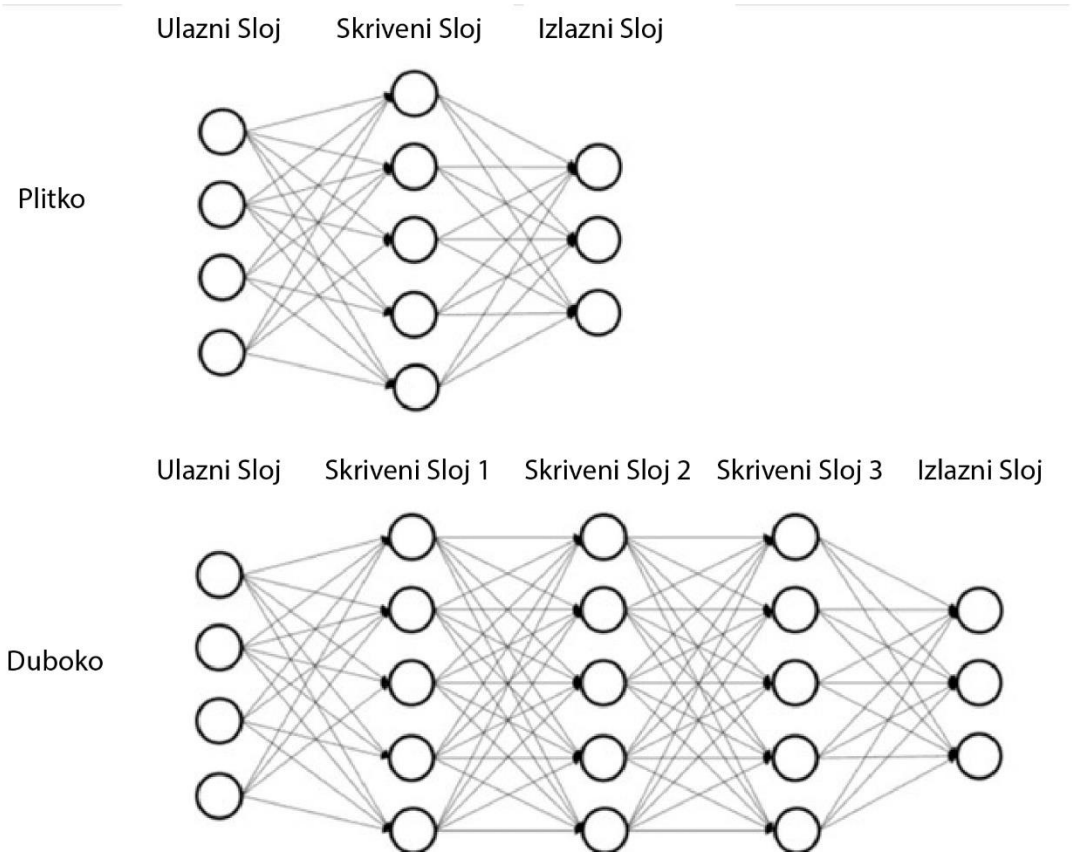
Neuronske mreže često se koriste za statističku analizu i modeliranje podataka, pružajući alternativu standardnim tehnikama nelinearne regresije ili analizi skupova podataka. Stoga se često primjenjuju u problemima koji se mogu formulirati kao klasifikacija ili prognoza. Primjeri uključuju prepoznavanje slika i govora, prepoznavanje tekstualnih znakova te područja ljudske ekspertize kao što su medicinska dijagnostika, geološko istraživanje za naftu i predviđanje financijskih tržišnih pokazatelja [11].

2.3. Duboko učenje

Duboko učenje (engl. Deep learning, DL) je grana umjetne inteligencije koja se fokusira na unapređenje metoda strojnog učenja kako bi se postigli ciljevi koji su prvotno zamišljeni za strojno učenje. Originalni ciljevi strojnog učenja uključuju sposobnost modeliranja i razumijevanja kompleksnih obrazaca u podacima kako bi se donosile precizne odluke i predviđanja. Duboko učenje nastoji postići ove ciljeve oponašanjem strukture i funkcionalnosti ljudskog mozga, posebice aktivnosti u slojevima neurona u neokorteksu, dijelu mozga koji je odgovoran za visoko složene kognitivne procese poput razmišljanja i odlučivanja [12]. Neokorteks, koji čini otprilike osamdeset posto ljudskog mozga, sadrži milijarde neurona i između sto i tisuću bilijuna sinapsi koje povezuju ove neurone. Duboko učenje koristi višeslojne neuronske mreže koje imitiraju ovu strukturu kako bi omogućile računalima da automatski uče iz velikih količina podataka, prepoznaju obrasce i donose odluke s visokim stupnjem preciznosti [12].

Naziva se dubokim učenjem jer koristi više od jednog sloja za nelinearnu transformaciju značajki. Svaki sloj u dubokoj neuronskoj mreži obrađuje podatke na drugačiji način i stvara nove, kompleksnije značajke. Ovo omogućuje sustavu da automatski uči i stvara prikaze podataka na različitim razinama apstrakcije, što znači da može razumjeti i modelirati složene obrasce bez potrebe za ručnim definiranjem značajki [12]. Na primjer, u zadatku prepoznavanja slika, prvi slojevi mreže mogu naučiti osnovne značajke poput rubova i boja, dok dublji slojevi mogu prepoznati složenije oblike poput lica ili objekata. Ovaj višerazinski pristup omogućuje mreži da automatski izvuče važne informacije iz podataka, čime se smanjuje potreba za ručnim odabirom značajki. Jedna od velikih prednosti dubokog učenja je mogućnost prethodnog učenja. Ovo znači da model može naučiti reprezentacije značajki iz velikih skupova podataka, a zatim primijeniti ta naučena znanja na druge, slične zadatke. Na primjer, model treniran za prepoznavanje objekata u slikama može se prilagoditi za prepoznavanje lica koristeći već naučene značajke.

Tradicionalni pristupi strojnog učenja često se nazivaju plitkim jer koriste samo jedan skriveni sloj i zahtijevaju značajno prethodno znanje od strane inženjera za ručno definiranje značajki. Nasuprot tome, duboko učenje, koje se fokusira na učenje reprezentacija kroz višestruke slojeve, pokazuje bolje performanse u prepoznavanju složenih i globalnih odnosa u podacima [12]. Slikom 2.8. prikazana je razlika u arhitekturi između plitkog i dubokog učenja.

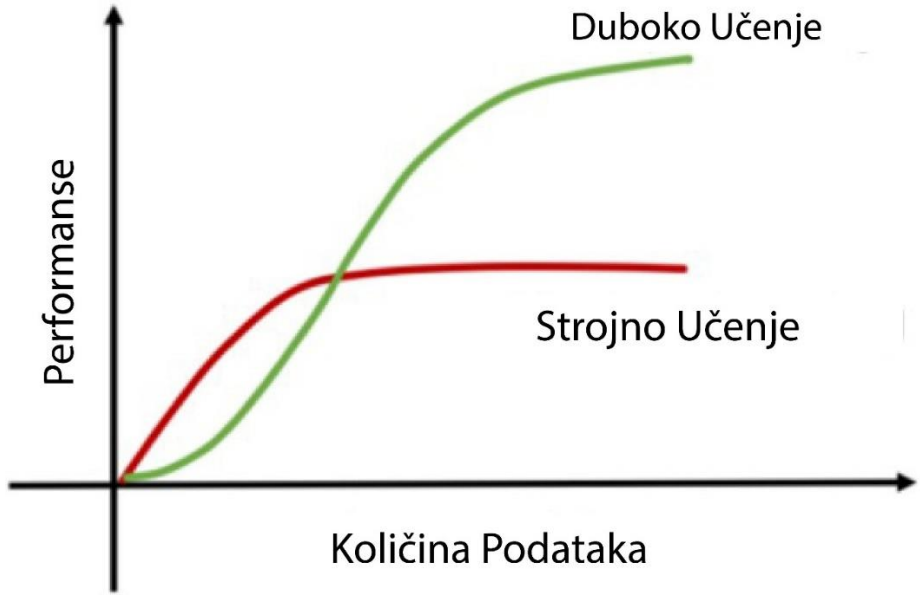


Slika 2.8. Usporedba plitke i duboke arhitekture. [12]

Algoritmi dubokog učenja pokazuju bolje performanse u izdvajanju nelinearnih i globalnih odnosa i uzoraka u podacima, ako ih usporedimo s relativno plitkim arhitekturama učenja. Neke od korisnih karakteristika naučenih apstraktnih reprezentacija dubokog učenja uključuju [12]:

- pokušava istražiti veći dio ogromnog volumena skupa podataka, čak i kada su podaci nenadzirani
- prednost nastavljanja poboljšanja kako se dodaju novi podaci za obuku
- automatsko izvlačenje reprezentacija podataka iz nenadziranih podataka ili nadziranih podataka, distribuirano i hijerarhijsko, obično najbolje kada je prostor ulaza lokalno strukturiran; prostorni ili vremenski - na primjer, slike, jezik, govor
- izvlačenje reprezentacija iz nenadziranih podataka omogućuje njihovu široku primjenu na različite vrste podataka poput tekstura slika, zvuka i slično

Razlika u performansama između tradicionalnog pristupa i dubokog učenja prikazana je slikom 2.9..



Slika 2.9. Performanse dva pristupa. [12]

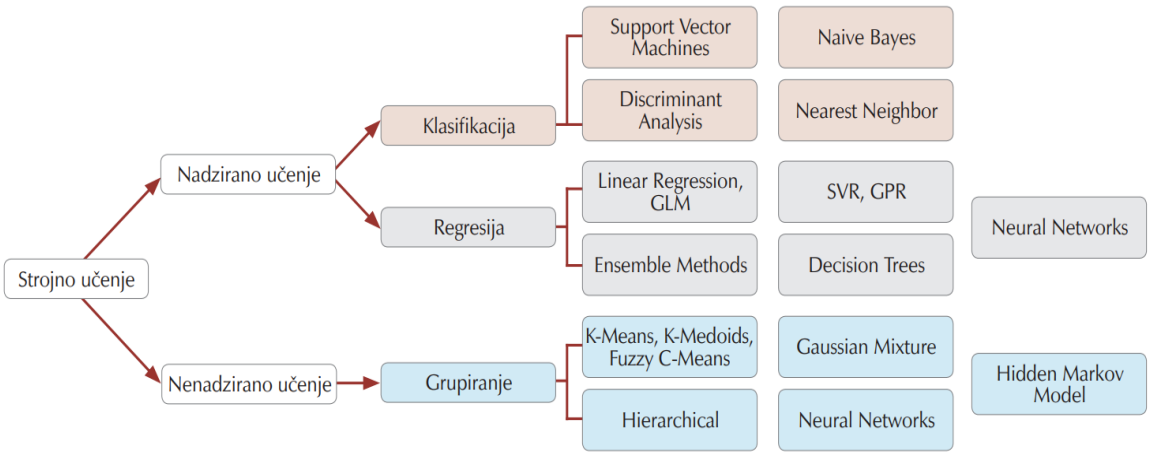
Duboko učenje ostvarilo je značajne napretke u različitim područjima, ali još uvijek postoje neki izazovi koje treba riješiti. Potrebne su velike količine podataka za učenje, što predstavlja izazov u prikupljanju dovoljno podataka za obuku. Osim toga, obuka modela dubokog učenja zahtijeva znatne računalne resurse, što može biti računalno skupo i zahtijevati specijaliziranu opremu poput GPU-a i TPU-a. Rad na sekvencijalnim podacima također može biti vremenski zahtjevan, ovisno o dostupnim računalnim resursima, često traje danima ili čak mjesecima. Modeli dubokog učenja su kompleksni i funkcioniraju kao crne kutije, što otežava interpretaciju rezultata. Pretreniranje je također izazov, jer se model može previše specijalizirati za podatke za obuku, što dovodi do loše performanse na novim podacima [13].

Prednosti dubokog učenja uključuju visoku točnost, automatsko inženjerstvo značajki, skalabilnost, fleksibilnost te kontinuirano poboljšanje performansi. Duboko učenje može postići najnovije rezultate u različitim zadacima, poput prepoznavanja slika i obrade prirodnog jezika. Algoritmi dubokog učenja mogu automatski otkriti i naučiti relevantne značajke iz podataka bez potrebe za ručnim inženjeringom značajki. Duboki modeli mogu rasti kako bi rukovali velikim i složenim skupovima podataka, te mogu učiti iz masivnih količina podataka. Također, mogu se primijeniti na različite zadatke i podatkovne tipove, poput slika, teksta i govora. Nedostaci dubokog učenja uključuju visoke zahtjeve za računalnim resursima, potrebu za velikim količinama označenih podataka, izazove u tumačenju rezultata, pretreniranje te tretiranje modela kao crne

kutije. Duboki modeli mogu se pretrenirati na skupu za učenje, što može rezultirati lošom izvedbom na novim i neviđenim podacima. Nadalje, interpretacija dubokih modela može biti izazovna, jer su često tretirani kao crne kutije, što otežava razumijevanje njihovog rada i donošenje predviđanja.

2.4. Algoritmi strojnog učenja

Danas postoji mnogo algoritama za odabir, svaki s jedinstvenim pristupom učenju. Nema univerzalne metode ili najboljeg algoritma. Niti iskusni stručnjaci ne mogu unaprijed predvidjeti učinkovitost algoritma. Stoga se često koristi metoda pokušaja i pogreške. Odabir algoritma ovisi o različitim faktorima, uključujući vrstu i obujam podataka, željene ciljeve te kontekst primjene rezultata [14]. Na slici 2.10. prikazana je podjela i popis tehnika strojnog učenja.



Slika 2.10. Podjela strojnog učenja. [14]

Kao što je već rečeno, postoje mnogi algoritmi strojnog učenja koji se koriste. Jedni od najpopularnijih i najčešće korištenih su:

1. linearna regresija
2. slučajna šuma
3. stablo odlučivanja
4. k – najbližih susjeda
5. potporni vektorski strojevi

Linearna regresija (engl. *Linear regression*) se koristi za procjenu stvarnih vrijednosti (kao što su cijene kuća, broj poziva, ukupne prodaje itd.) na temelju kontinuiranih varijabli. Ovdje se uspostavlja odnos između nezavisnih i ovisnih varijabli pronalaženjem najbolje prilagođene linije.

Ta najbolje prilagođena linija poznata je kao regresijska linija i predstavljena je linearnom jednadžbom $Y = a \cdot X + b$ [15].

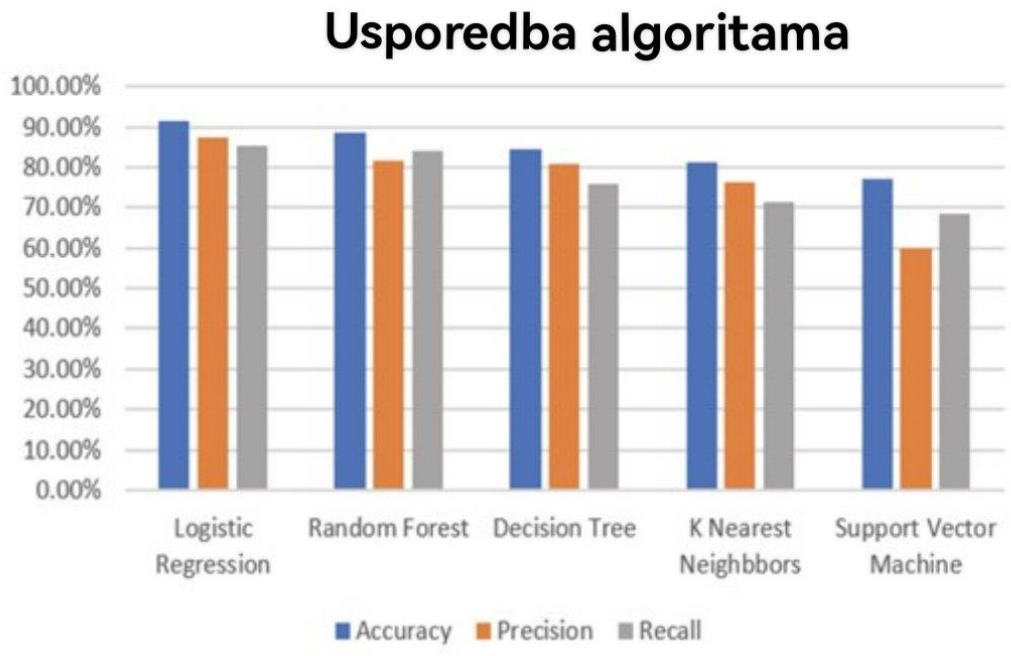
Slučajna šuma (engl. *Random Forest*) je zaštićeni naziv za učenje o stablima odluke. U ovom algoritmu, imamo kolekciju stabala odluke (također poznatih kao "šuma"). Da bi se klasificirao novi objekt na temelju atributa, svako stablo daje klasifikaciju, i kažemo da stablo glasuje za tu klasu. Šuma odabire klasifikaciju koja ima najviše glasova (preko svih stabala u šumi) [15].

Možda i najpoznatiji algoritam stablo odlučivanja (engl. *Decision tree*) se uglavnom koristi za probleme klasifikacije. Iznenađujuće, radi i za kategoričke i za kontinuirane zavisne varijable. U ovom algoritmu, populaciju dijelimo na dva ili više homogenih skupova. To se radi na temelju najvažnijih atributa/neovisnih varijabli kako bismo stvorili što različitije grupe [15].

Može se koristiti za klasifikaciju i regresiju, no češće se primjenjuje u klasifikacijskim problemima u industriji. K - najbližih susjeda (engl. *K – Nearest Neighbors,* kNN) je jednostavan algoritam koji pohranjuje sve dostupne slučajeve te klasificira nove slučajeve glasovanjem većine svojih k susjeda. Slučaj se dodjeljuje klasi koja je najčešća među k najbližih susjeda, a mjerenje se vrši funkcijom udaljenosti. Funkcije udaljenosti mogu biti euklidske, *Manhattan*, *Minkowski* i *Hammingove* udaljenosti. Prve tri koriste se za kontinuirane varijable, dok se četvrta (*Hammingova*) primjenjuje za kategoričke varijable. Ako je $K = 1$, slučaj se jednostavno dodjeljuje klasi najbližeg susjeda. Ponekad odabir vrijednosti K može biti izazovan prilikom modeliranja kNN-a. KNN se lako može povezati s našim stvarnim životima [15].

Potporni vektorski strojevi (engl. *Support Vector Machine*, SVM) je moćan algoritam strojnog učenja koji se koristi za linearnu ili nelinearnu klasifikaciju, regresiju, pa čak i za zadatke detekcije odstupanja. SVM-ovi se mogu koristiti za različite zadatke poput klasifikacije teksta, klasifikacije slika, detekcije neželjene pošte, identifikacije rukopisa, analize ekspresije gena, detekcije lica te otkrivanja anomalija. SVM-ovi su prilagodljivi i učinkoviti u različitim primjenama jer mogu upravljati visoko dimenzionalnim podacima i nelinearnim odnosima. SVM algoritmi su vrlo učinkoviti jer pokušavaju pronaći maksimalnu razdvajajuću hiperplohu između različitih klasa dostupnih u ciljnoj značajki [16].

Usporedba ovih algoritama, točnije njihove točnosti, preciznosti i odziva s jednog eksperimenta je prikazana slikom 2.11..



Slika 2.11. Usporedba algoritama. [17]

3. ANALIZA POSTOJEĆIH RJEŠENJA

Prethodni radovi na temu primjene strojnog učenja u predviđanju globalne prodaje video igara istražuju razne metode i algoritme kako bi precizno predvidjeli prodajne rezultate na temelju različitih čimbenika. Ti čimbenici mogu uključivati karakteristike igre, povijesne prodajne podatke, recenzije korisnika, marketinške kampanje i demografske podatke. Neki od najčešće korištenih algoritama u tim radovima već su spomenuti i objašnjeni u prethodnom poglavlju. Ovi algoritmi pomažu u prepoznavanju obrazaca i korelacija među podacima koji mogu utjecati na prodaju video igara. Radovi također često uključuju analizu značajki s ciljem utvrđivanja koji su čimbenici najvažniji za predviđanje prodaje. Na primjer, žanr igre, datum izlaska, recenzije i ocjene, kao i izdavač igre, mogu biti ključni pokazatelji buduće prodaje. Značajni radovi na tu temu mogu se pronaći u znanstvenim časopisima i na konferencijama vezanim uz strojno učenje, analizu podataka i industriju video igara. Pregledavanje baza podataka kao što su IEEE *Xplore, SpringerLink, arXiv* ili *Google Scholar* moglo bi pružiti priliku za uvid u relevantna istraživanja i studije slučaja.

Rad Julie Marcoux i Sid-Ahmed Selouani pod nazivom ***Hybrid Subspace-Connectionist Data Mining Approach for Sales Forecasting in the Video Game Industry*** istražuje novu tehniku za predviđanje prodaje video igara, kombinirajući metode povezničkih mreža s metodama dekompozicije pod prostora [18]. Ključni cilj rada je razviti alat koji podržava upravljanje tvrtkom u određivanju očekivanih prodajnih rezultata. U radu se koristi neuronska mreža, trenirana algoritmom povratnog širenja, za predviđanje tjedne prodaje video igara. Optimalna topologija i vremenski osjetljiva neuronska mreža su implementirani, uz procjenu relevantnosti ulaznih parametara kroz analizu glavnih komponenti (PCA). Slikom 3.1. su prikazani ključni faktori koji se koriste u radu.

Rezultati istraživanja pokazuju da je predloženi hibridni sustav PCA/AR-TDNN značajno poboljšao točnost predviđanja u usporedbi s referentnim sustavima. U testiranju s igricama razvijenim od strane Activision-a, hibridni sustav postigao je točnost predviđanja od 89.48%, dok su ostali modeli imali niže rezultate. Za igre razvijene od strane EA, hibridni sustav povećao je točnost predviđanja na 75.3%, u odnosu na 68.1% s jednostavnim modelom. Ova metoda pokazuje značajno poboljšanje u točnosti predviđanja prodaje i pruža prilagodljivost za različite razvojne studije, zahvaljujući modularnoj prirodi AR-TDNN-a.

	Opis
Igrača konzola	Wii, Xbox 360, itd.
Status za više igrača	jednoigrački, višeigrački ili oboje
Treća stranka	informacije s obzirom na izdavača
Pozicija u prvom tjednu	promjenjiva ovisno o tjednu izlaska
Prošli tjedni	broj tjedana proteklih od izlaska
Prodaja prošlog tjedna	broj prodanih jedinica tijekom prošlog tjedna
MSRP	maloprodajna cijena videoigre
Status za više igrača	izbor: jednoigrački, višeigrački, oboje
Online status	izbor: offline, unaprijeđeno, online
Status na više platformi	dostupno na više konzola ili ne
Status dodatne opreme	dodatna oprema: nije potrebna, potrebna
Status nastavka	opcije: nije nastavak, nastavak, ponovno izdanje
Žanr	vrsta videoigre, kako je dodijeljeno od strane IGN-a
Recenzije čitatelja	broj čitatelja koji su dali ocjenu
IGN ocjena	ocjena od IGN.com
Prosječna ocjena	ocjena kako su je dale novinarske organizacije
Prošli tjedni od objave	tjedni protekli od izlaska
Jedinice prodane u prvom tjednu	prodano u Sjevernoj Americi tijekom prvog tjedna

Slika 3.1. Ključni faktori. [18]

Rad ***Predicting Global Video-Game Sales*** autora Alice Yufa, Jonathan L. Yu, Henry Chan i Paul D. Berger bavi se predviđanjem globalne prodaje video igara na temelju prošlih prodajnih podataka [19]. Cilj rada bio je identificirati ključne varijable koje utječu na prodaju video igara, kao što su broj kritičara koji ocjenjuju igru, prosječna ocjena kritičara, ocjena korisnika, te broj korisnika koji su dali recenzije. Rad je koristio podatke s web stranice *Metacritic*, koja agregira recenzije video igara, kako bi analizirali utjecaj tih varijabli na globalnu prodaju igara.

Rezultati rada pokazuju da su visoke ocjene kritičara i korisnika značajan faktor koji doprinosi većoj prodaji igara, budući da utječu na interes korisnika i promociju igara na platformama za prodaju. Također, rad ističe da prodaja u Sjevernoj Americi ima ključan utjecaj na ukupnu globalnu prodaju, dok utjecaj žanra igre ostaje neodređen zbog složenosti klasifikacije video igara.

Istraživanje je imalo određena ograničenja, uključujući probleme s kodiranjem varijabli žanra i platforme te relativno nisku vrijednost R^2 (12.7%), što ukazuje na to da postoje dodatni faktori koji utječu na prodaju igara, ali nisu obuhvaćeni modelom. Rad zaključuje da je potrebno dodatno istražiti kako bi se utvrdio točan utjecaj drugih varijabli.

Treći rad ***Predicting Video Game Sales Based on Machine Learning and Hybrid Feature Selection Method*** predlaže novu hibridnu metodu odabira značajki, koja kombinira Pearsonov koeficijent korelacije i metodu odabira značajki putem slučajne šume (Random Forest Feature Selection - PCC-RFFS) [20]. Cilj rada je poboljšati točnost predviđanja prodaje video igara korištenjem devet metoda strojnog učenja u kombinaciji s hibridnom metodom odabira značajki. Kroz eksperimentiranje na stvarnim podacima prikupljenima od veljače 2006. do studenog 2016.,

autori su pokazali da kombinirana metoda PCC-RFFS nadmašuje metode koje koriste samo Pearsonov koeficijent korelacije ili samo Random Forest. Rad također istražuje različite pristupe odabira značajki, poput metoda temeljenih na filtriranju, omotačima i hibridnim metodama, te naglašava kako različiti pristupi mogu utjecati na točnost modela. Dok Pearsonov koeficijent korelacije uzima u obzir samo linearne odnose između značajki i cilja, metoda slučajne šume omogućuje procjenu važnosti značajki na temelju njihovog doprinosa konačnom modelu, uzimajući u obzir i nelinearne odnose. Kombinacija ovih metoda u hibridni pristup pokazala se učinkovitijom u otkrivanju složenih odnosa između značajki. Rad zaključuje da hibridna metoda PCC-RFFS pruža poboljšanu točnost u predviđanju prodaje video igara u usporedbi s tradicionalnim metodama odabira značajki, ali zahtijeva više vremena za izvođenje. Relevantnost ovog rada za istraživanje leži u primjeni naprednih tehnika strojnog učenja i odabira značajki kako bi se poboljšala točnost predviđanja, što može pomoći u razvoju preciznijih modela predviđanja prodaje u kontekstu industrije video igara.

Rad ***Video Game Sales Analysis*** autora V. Sarala i D. Akhile bavi se analizom prodaje videoigara na globalnom tržištu koristeći tehnike strojnog učenja [21]. Cilj istraživanja je identificirati igre s većom prodajom globalno u usporedbi s drugim zemljama. U tu svrhu, autori su koristili povijesne podatke o prodaji u obliku vremenskih serija, a skup podataka uključuje jedanaest varijabli i petsto uzoraka koji kombiniraju kategorijske i numeričke varijable.

Autori su proveli postupak obrade podataka kako bi uklonili neispravne ili nepotpune unose te identificirali ulazne i ciljne varijable za primjenu algoritama strojnog učenja. Korišteni algoritmi uključuju linearnu regresiju, podržanu vektorsku regresiju, slučajnu šumu i stablo odlučivanja. Nakon inicijalne analize podataka, dva su modela izrađena: jedan koji uklanja unose bez ocjena, i drugi koji koristi ponderirani faktor za ocjenjivanje preostalih unosa. Nakon primjene različitih algoritama strojnog učenja, autori su koristili optimizaciju hiperparametara putem nasumičnog pretraživanja kako bi dodatno smanjili pogrešku. Rezultati su pokazali da algoritam slučajne šume (Random Forest) daje najtočnije rezultate s najmanjom stopom pogreške.

Ovaj rad je značajan jer demonstrira kako različite tehnike obrade i analize podataka mogu utjecati na preciznost modela predviđanja prodaje videoigara. Rad je također važan jer pokazuje primjenu i usporedbu različitih algoritama strojnog učenja, što omogućava bolje razumijevanje koja metoda daje najbolje rezultate u kontekstu predviđanja prodaje na globalnom tržištu video igara.

U sljedećem radu ***Research on the Prediction on the Sales of Electronic Games*** autora Weiqi Huanga, glavni cilj je bio predvidjeti i analizirati karakteristike prodaje elektroničkih igara

koristeći različite metode strojnog učenja, uključujući neuronske mreže, XGBoost, i LightGBM [22]. Rad se fokusira na predviđanje prodaje igara temeljem različitih značajki kao što su vrijeme izdavanja, rangiranje, izdavač i žanr igre. Ciljevi rada se odnosne na predviđanje prodaje, analize performansi modela te usporedbi implementiranih modela.

Što se tiče predviđanja prodaje, bilo je potrebno razviti modele strojnog učenja koji mogu precizno predvidjeti prodaju video igara koristeći različite značajke. Zatim, korištenjem metrika kao što su Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), i Mean Squared Error (MSE), rad analizira i uspoređuje performanse triju modela (neuronske mreže, XGBoost, LightGBM). Na kraju, autori su detaljno usporedili prednosti i nedostatake svakog od modela u kontekstu predviđanja prodaje video igara, s posebnim naglaskom na to koji model daje najbolje rezultate za određene vrste podataka. Za rezultate su dobili kako su se neuronske mreže pokazale slabima u usporedbi s ostalim metodama, uz značajne varijacije u predikcijama. XGBoost model je pokazao najbolje rezultate, s najnižim RMSE i MSE, te je ocijenjen kao najučinkovitiji u predviđanju prodaje igara. LightGBM je također pokazao dobre performanse, ali nešto slabije od XGBoosta, iako brže konvergira.

Rad pod nazivom ***Automatic Machine Learning-Based Data Analysis for Video Game Industry*** autora Z. Zhou istražuje upotrebu automatiziranih metoda strojnog učenja za analizu podataka u industriji video igara [23]. Rad se fokusira na tri ključna cilja. Prvi cilj je automatizacija analize podataka, gdje se istražuje kako automatizirani alati mogu poboljšati preciznost i efikasnost u radu s velikim količinama podataka, uključujući recenzije, prodaju i korisničke ocjene video igara. Drugi cilj je procjena utjecaja recenzija i ocjena na ukupnu prodaju igara, s posebnim naglaskom na to koliko su korisničke recenzije i ocjene važni u predviđanju uspjeha igara na tržištu. Treći cilj odnosi se na optimizaciju marketinških strategija na temelju podataka o prodaji i recenzijama, pružajući preporuke izdavačima i developerima za bolje pozicioniranje proizvoda na tržištu.

Rezultati istraživanja pokazali su da automatizirane metode strojnog učenja značajno povećavaju preciznost predikcija prodaje, osobito kada se kombiniraju s podacima o korisničkim recenzijama. Analiza je otkrila da su ocjene korisnika i broj recenzija ključni indikatori buduće prodaje, što može značajno utjecati na marketinške odluke i strategije razvoja proizvoda. Rad također predlaže daljnju primjenu automatiziranih alata za bržu i efikasniju analizu podataka, što je važno za održavanje konkurentnosti u brzo rastućoj industriji video igara. Ovaj rad demonstrira kako automatizacija i napredne tehnike strojnog učenja mogu poboljšati analizu i predikciju, omogućujući developerima i izdavačima da bolje razumiju i prilagode tržišne trendove.

Rad pod nazivom ***Data Interpretation and Video Games Sales Prediction Using Machine Learning Algorithms - a Comparative Study*** autorstva Manimuthu, A., Udhayakumar, U., Cathrine, A., Gowri, T. D., Peter, J., Selvam, S., i Roseline, S., bavi se primjenom različitih algoritama strojnog učenja u svrhu predviđanja prodaje video igara [24]. Cilj istraživanja je analizirati kako različite metode, uključujući linearnu regresiju, višestruku regresiju, Random Forest, i Support Vector Machines, mogu biti korištene za točno predviđanje prodaje igara na temelju dostupnih podataka. Rad također uspoređuje performanse tih modela koristeći R^2 kao ključni mjerni kriterij. R^2 pomaže u ocjeni koliko dobro svaki model objašnjava varijaciju u podacima, što omogućuje razumijevanje koji algoritam najbolje odgovara stvarnim tržišnim uvjetima i pruža najpreciznija predviđanja.

Rezultati istraživanja pokazuju koji model nudi najbolju preciznost u predviđanju prodaje video igara. Usporedba performansi različitih modela otkriva koji od njih najučinkovitije koristi podatke za predviđanje buduće prodaje. Na temelju ovih rezultata, rad nudi preporuke za primjenu najučinkovitijih modela u industriji video igara, pružajući smjernice za optimizaciju marketinških i razvojnih strategija.

Ovaj rad nudi važne uvide u primjenu algoritama strojnog učenja za analizu tržišta video igara, omogućujući developerima i izdavačima da bolje razumiju tržišne trendove i unaprijede svoje strategije temeljem preciznijih predikcija.

Rad pod nazivom ***Sales Prediction on Video Games Using Machine Learning Approaches*** autora K. Saraswathi, N. T. Renukadevi, i S. Nandhinidevi, istražuje primjenu metoda strojnog učenja za predviđanje prodaje video igara [25]. Cilj rada je razviti i analizirati različite modele strojnog učenja kako bi se poboljšala točnost predviđanja prodaje video igara.

Autori su istražili nekoliko algoritama strojnog učenja, uključujući Linearnu regresiju, Random Forest i Support Vector Machines, kako bi ocijenili njihovu učinkovitost u predviđanju prodaje. Analizirali su kako različiti modeli obrađuju podatke o recenzijama, ocjenama i drugim faktorima koji utječu na prodaju video igara.

Rezultati istraživanja pokazali su koji modeli pružaju najbolje rezultate u predviđanju prodaje. Rad je otkrio da su neki algoritmi, poput Random Forest, postigli značajno bolje rezultate u točnosti predviđanja u usporedbi s drugim metodama. Ovi nalazi sugeriraju da se odabrani modeli strojnog učenja mogu koristiti za preciznije prognoze i unapređenje strategija marketinških i razvojnih odluka u industriji video igara. Rad pruža vrijedne uvide u primjenu različitih metoda strojnog

učenja za predviđanje prodaje, pomažući developerima i izdavačima da bolje razumiju i prilagode svoje strategije temeljem točnijih predviđanja.

Iz navedenih radova koji se bave sličnom temom kao i ovaj rad, može se zaključiti kako odabir algoritama strojnog učenja igra presudnu ulogu u točnosti predviđanja globalne prodaje video igara. Različiti radovi pokazuju da se algoritmi poput Random Forest-a, neuronskih mreža i linearne regresije mogu razlikovati u učinkovitosti ovisno o specifičnim značajkama podataka. Stoga je nužno testirati i usporediti više algoritama kako bi se pronašao onaj koji najbolje odgovara podacima. Važno je pažljivo odabrati značajke koje će se koristiti u modelu, jer radovi sugeriraju da elementi poput ocjena kritičara, broja recenzija, žanra igre i datuma izdavanja značajno utječu na prodaju. Analiza pokazuje da je jako bitno procijeniti učinkovitost modela na različitim skupovima podataka, kako bi se osigurala njegova sposobnost generalizacije na neviđene podatke. To uključuje evaluaciju modela pomoću metrika poput RMSE ili MAE, što će omogućiti razumijevanje stvarne preciznosti predviđanja.

4. BAZA PODATAKA

Baza podataka korištena u ovom istraživanju dostupna je na platformi *Kaggle* pod nazivom "*Video Game Sales with Ratings*". Ova baza podataka obuhvaća širok spektar informacija o prodaji videoigara diljem svijeta, uključujući detalje o pojedinačnim igrama, njihovim izdavačima, platformama na kojima su izdane te ocjenama koje su igre dobile od strane kritičara i korisnika. Podaci pokrivaju razdoblje od nekoliko desetljeća, pružajući sveobuhvatan pregled tržišta videoigara i omogućujući analizu povijesnih trendova u industriji.

Baza sadrži različite varijable koje omogućuju dubinsku analizu različitih faktora koji mogu utjecati na prodaju videoigara. To uključuje podatke o prodaji u različitim regijama (Sjeverna Amerika, Europa, Japan, ostatak svijeta), kao i globalnu prodaju, što istraživačima daje mogućnost da prouče geografske razlike u popularnosti igara. Osim toga, baza uključuje i meta podatke kao što su godina izdavanja, žanr igre te izdavač, koji mogu biti ključni faktori u predviđanju uspjeha igre na tržištu. Baza je kreirana s ciljem pružanja uvida u tržište videoigara, omogućavajući istraživačima, analitičarima i entuzijastima da analiziraju različite aspekte industrije, kao što su utjecaj platforme na prodaju, povezanost između ocjena i komercijalnog uspjeha, te evolucija tržišta kroz vrijeme. Njena sveobuhvatnost i raznolikost podataka čine je idealnim resursom za izgradnju modela strojnog učenja, koji mogu koristiti ove informacije za predviđanje budućih trendova u prodaji videoigara. Ova baza podataka nije samo koristan alat za akademska istraživanja, već i za poslovne analize, gdje se mogu identificirati ključni faktori uspjeha na tržištu videoigara. Korištenje ovakve baze podataka omogućuje donošenje informiranih odluka u razvoju i plasmanu novih naslova, kao i razumijevanje tržišnih dinamika koje mogu utjecati na prodajne rezultate.

Što se tiče strukture baze podataka, sastoji se od šesnaest atributa (značajki) koje su dobrim dijelom ispitivane u prethodnim poglavljima te 16 719 redaka. Iako su već spomenute i nabrojane, popis značajki uz objašnjenja što predstavljaju:

- ***Name***: naziv videoigre.
- ***Platform***: platforma na kojoj je igra izdana (npr. PS4, Xbox One).
- ***Year_of_Release***: godina izlaska igre.
- ***Genre***: žanr igre (npr. Action, Sports).
- ***Publisher***: izdavač igre.
- ***NA_Sales***: prodaja igre u Sjevernoj Americi (u milijunima).
- ***EU_Sales***: prodaja igre u Europi (u milijunima).

- ***JP_Sales***: prodaja igre u Japanu (u milijunima).
- ***Other_Sales***: prodaja igre u ostatku svijeta (u milijunima).
- ***Global_Sales***: ukupna globalna prodaja igre (u milijunima).
- ***Critic_Score***: prosječna ocjena kritičara (od 0 do 100).
- ***Critic_Count***: broj ocjena kritičara.
- ***User_Score***: prosječna ocjena korisnika (od 0 do 10).
- ***User_Count***: broj korisničkih ocjena.
- ***Developer***: razvijač igre.
- ***Rating***: ESRB ocjena igre (npr. E, T, M).

Baza podataka sadrži određeni broj nedostajućih vrijednosti, posebno u ključnim atributima kao što su *Year_of_Release* (godina izdavanja igre), *Critic_Score* (prosječna ocjena kritičara), *User_Score* (prosječna ocjena korisnika), *Developer* (razvijač igre) i *Rating* (ESRB ocjena igre). Ove praznine u podacima mogu biti rezultat nekompletnih izvora informacija, neusklađenih formata ili jednostavnog nedostatka dostupnih podataka za određene igre ili vremenska razdoblja. Nedostajući podaci predstavljaju značajan izazov u analizi jer mogu narušiti preciznost i pouzdanost rezultata. Na primjer, nedostatak informacija o godini izdavanja može otežati analizu povijesnih trendova, dok nedostatak ocjena kritičara ili korisnika može dovesti do nepreciznih modela predviđanja koji se oslanjaju na te varijable. Osim toga, nepostojanje informacija o razvijaču ili ESRB ocjeni može ograničiti razumijevanje uloge tih faktora u uspjehu videoigara na tržištu.

Podaci su prikupljeni iz različitih izvora, uključujući prodajne izvještaje, recenzije kritičara i ocjene korisnika. Iako baza pruža opsežan uvid u tržište videoigara, njezina pouzdanost može varirati ovisno o izvorima iz kojih su podaci preuzeti. Podaci o prodaji dolaze iz različitih geografskih regija, što omogućuje analizu regionalnih razlika u uspjehu igara.

Prije treniranja modela, podaci su prošli kroz nekoliko faza prije procesiranja:

- kodiranje oznaka: kategorizirani atributi kao što su platforma, žanr i izdavač pretvoreni su u numeričke vrijednosti.
- skaliranje podataka: atributi su standardizirani kako bi svi podaci imali jednak raspon, što je važno za rad s neuronskim mrežama.
- uklanjanje nedostajućih vrijednosti: redci s nedostajućim ključnim atributima su uklonjeni kako bi se osigurala točnost predviđanja.

4.1. Deskriptivna analiza podataka

Podaci korišteni u ovom istraživanju sastojali su se od 16 719 redaka i 16 atributa. Prije početka analize podataka, provedeno je temeljito čišćenje i priprema podataka. Proces čišćenja podataka uključivao je nekoliko koraka:

- uklanjanje nedostajućih vrijednosti: prvi korak u obradi podataka bio je identificiranje i uklanjanje redaka s nedostajućim ključnim vrijednostima. Nedostajuće vrijednosti mogu značajno utjecati na performanse modela, stoga su iz podataka isključeni svi zapisi koji nisu imali potpune informacije za ključne atribute.
- kodiranje kategorijskih podataka: kategorijski podaci, koji predstavljaju nesumjerljive varijable, pretvoreni su u numeričke vrijednosti koristeći tehniku kodiranja oznaka. Ova tehnika omogućava modelu da interpretira kategorijske vrijednosti kao brojeve, čime se olakšava proces učenja.
- isključivanje neželjenih podataka: osim uklanjanja nedostajućih vrijednosti, isključeni su i svi podaci koji su smatrani nerelevantnima za analizu. To je uključivalo zapise koji su imali irelevantne vrijednosti, što je moglo ometati model prilikom učenja.

Nakon čišćenja podataka, izvršena je podjela na trening i testni skup:

- trening skup: 80% podataka (13 375 redaka) korišteno je za treniranje modela. Ova podjela omogućava modelu da "nauči" obrazac iz većine dostupnih podataka, pružajući solidnu osnovu za točna predviđanja.
- testni skup: preostalih 20% podataka (3344 redaka) korišteno je za testiranje modela. Ovaj skup podataka nije korišten u fazi učenja, što omogućava evaluaciju modela na podacima koji su mu "nepoznati". Ova procjena je ključna za razumijevanje koliko dobro model generalizira na nove podatke.

Analiza podataka nakon čišćenja pokazala je da su podaci sada u homogenijem i konzistentnijem obliku. Nakon isključivanja nerelevantnih i nekvalitetnih podataka, preostali podaci pružaju čvrstu osnovu za modeliranje. Ovi optimizirani podaci omogućili su modelu da prepozna složene obrasce i značajno poboljšao preciznost predikcija.

Ova detaljna obrada i priprema podataka bila je ključna za osiguranje visokih performansi modela, jer su isključene sve potencijalne greške i šumovi koji bi mogli utjecati na konačne rezultate.

5. IMPLEMENTACIJA MODELA

U ovom poglavlju bavit će se implementacijom modela za predviđanje globalne prodaje video igara korištenjem općih tehnika strojnog učenja. Fokus će uglavnom biti na regresijske algoritme; ti će algoritmi uključivati linearne regresije i naprednije metode, poput neuronskih mreža i dubokog učenja. Međutim, prije prikaza detalja modela te raznih rezultata i grafova, predstavit će se nekoliko biblioteka i alata korištenih u izradi ovog projekta.

5.1. Biblioteke i alati

Korištena je *Community Edition* verzija *PyCharma*, opće primjenjivog integriranog razvojnog okruženja (IDE) za Python. PyCharm je IDE (engl. Integrated Development Environment) s dubokom podrškom za pisanje, testiranje i otklanjanje pogrešaka u Python kodu. Ovo uključuje inteligentno uređivanje koda s pametnim prijedlozima za kodiranje, automatskim dovršavanjem koda i naprednim funkcijama pretraživanja koje pomažu u radu s velikim kodnim bazama[26]. Osim toga, PyCharm nudi napredne alate za otklanjanje pogrešaka koji omogućuju praćenje i ispravljanje pogrešaka u kodu. *Debugger* omogućuje postavljanje točke prekida (engl. *breakpoint*), korak po korak izvođenje koda i pregled varijabli, što pomaže u detaljnom ispitivanju i ispravljanju problema u kodu [26]. Još jedna velika prednost PyCharma je njegova integracija s alatima poput Git-a. Ovo znatno olakšava upravljanje kodom i često omogućuje suradnju s drugim razvojnim timovima putem izvođenja radnji poput *commita*, *push*-a, *pull*-a i *merge*-a izravno iz IDE-a [26]. Također, PyCharm podržava virtualna okruženja koja su potrebna za izolaciju projekata i jednostavno upravljanje paketima. Ova funkcionalnost osigurava da ovisnosti ostanu odvojene i da projekti ne ometaju jedan drugoga [26]. Kako bi se izgradio što kvalitetniji model, iskorištene su mnoge Python biblioteke. Ključne biblioteke korištene u projektu:

- ***NumPy***: temeljna biblioteka za znanstveno računalstvo u Pythonu. To je Python biblioteka koja pruža višedimenzionalni niz objekata, različite izvedene objekte (kao što su maskirani nizovi i matrice), te niz rutina za brze operacije na nizovima, uključujući matematičke, logičke, manipulaciju oblikom, sortiranje, odabir, I/O, diskretne *Fourierove* transformacije, osnovnu linearnu algebru, osnovne statističke operacije, slučajnu simulaciju i još mnogo toga [27].
- ***Pandas***: *open-source* biblioteka s BSD licencom koja pruža visokoučinkovite, jednostavne za korištenje strukture podataka i alate za analizu podataka za programski jezik Python [28].
- ***Scikit-Learn*** (sklearn): Scikit-learn je open-source biblioteka za strojno učenje koja podržava nadgledano i nenadgledano učenje. Također pruža različite alate za prilagodbu

modela, predprocesiranje podataka, odabir modela, evaluaciju modela i mnoge druge korisne funkcionalnosti [29].

- ***Matplotlib*** i ***Seaborn***: ove biblioteke koriste se za vizualizaciju podataka. Matplotlib pruža svestrani okvir za crtanje, dok je Seaborn izgrađen na vrhu Matplotlib-a i nudi višu razinu sučelja za stvaranje informativnih i privlačnih statističkih grafika.
- ***TensorFlow*** i ***Keras***: TensorFlow je open-source biblioteka koju je razvila Google za zadatke strojnog učenja i dubokog učenja. Keras je visokorazinski API za izgradnju i treniranje neuronskih mreža koji radi na vrhu TensorFlow-a. Ove biblioteke će se koristiti za implementaciju modela neuronskih mreža koji predviđaju globalne rezultate prodaje video igara.
- ***Pickle***: Python modul koji se koristi za serializaciju i deserializaciju objekata. Omogućuje spremanje treniranih modela na disk i njihovo ponovno učitavanje, što je korisno za spremanje modela nakon treninga i njihovu kasniju upotrebu bez potrebe za ponovnim treniranjem.

5.2. Primjena algoritama strojnog učenja

Za predviđanje globalne prodaje video igara koristi će se regresijski algoritmi zbog njihove sposobnosti da modeliraju i predviđaju kontinuirane vrijednosti. Regresija je izuzetno korisna u situacijama kada želimo razumjeti odnose između varijabli i predviđati kvantitativne rezultate na temelju tih odnosa. U kontekstu globalne prodaje video igara, cilj je razviti model koji može precizno predvidjeti prihod od prodaje na temelju različitih čimbenika kao što su platforma, godina izdanja, žanr, i drugi atributi igre. Regresijski algoritmi omogućuju nam modeliranje odnosa između zavisnih i neovisnih varijabli, te predviđanje kontinuiranih vrijednosti na temelju tih odnosa. Različite metode će biti istražene kako bismo pronašli onaj koji najbolje odgovara specifičnostima naših podataka i uspješno ispunjava ciljeve predviđanja. Eksperimentiranjem s različitim algoritmima nastojimo identificirati najučinkovitiji model za predviđanje globalne prodaje video igara, što će nam omogućiti bolje razumijevanje tržišta i optimizaciju strateških odluka.

Prvi na redu od algoritama je linearna regresija. Za početak, kao i u svakom drugom algoritmu, učitavamo potrebne biblioteke i podatke. Slikom 5.1. je prikazano što je potrebno za ovaj algoritam. Preko varijable *data* učitavamo podatke. Naša baza podataka je pod nazivom *Video_Games_Sales_as_at_22_2016.csv* (što se može vidjeti i pod navodnicima), a javno je dostupna na https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings te su i mnogi drugi radovi napravljeni upravo prema njoj.

```
import pandas as pd
import numpy as np
import sklearn
from sklearn import linear_model
import matplotlib.pyplot as pyplot
import pickle
from matplotlib import style

data = pd.read_csv("Video_Games_Sales_as_at_22_Dec_2016.csv")
```

Slika 5.1. Učitavanje podataka.

Potrebno je uključiti i bitne značajke, a u ovom slučaju to su: *Name*, *Platform*, *Year_of_Release*, *Genre*, *Publisher*, *NA_Sales*, *EU_Sales*, *JP_Sales*, *Other_Sales*, *Global_Sales*, *Critic_Score*, *Critic_Count*, *User_Score*, *User_Count*, *Developer*, *Rating*. Ovdje je bitno napomenuti da mnogi algoritmi strojnog učenja ne mogu raditi direktno s tekstualnim podacima već zahtijevaju numeričke vrijednosti. Takvu pretvorbu omogućuje *get_dummies()* funkcija. Nakon toga, definiraju se ulazne i izlazne varijable te podjela podataka na trenirajući i testni skup. Treniranje modela za linearnu regresiju dan je slikom 5.2.:

```
linear = linear_model.LinearRegression()
linear.fit(x_train, y_train)
```

Slika 5.2. Model linearne regresije.

Ovdje se pomoću spomenute scikit-learn biblioteke poziva algoritam linearne regresije. Metoda fit trenira model na osnovu podataka gdje x_train predstavlja ulazne podatke, a y_train izlazne vrijednosti.

Kada je to sve napravljeno, moguće je vidjeti rezultate, odnosno predviđanje te stvarne rezultate. Ispis nekoliko predviđanja prikazan je slikom 5.3.

```
Predviđanje: 0.7831954485711314, Stvarna vrijednost: 0.8
Predviđanje: 0.08334020035667633, Stvarna vrijednost: 0.09
Predviđanje: 0.09949271236724622, Stvarna vrijednost: 0.1
Predviđanje: 0.5762669337313602, Stvarna vrijednost: 0.59
Predviđanje: 0.2757656955616663, Stvarna vrijednost: 0.28
Predviđanje: 0.5403497688649104, Stvarna vrijednost: 0.57
Predviđanje: 0.2592987270724074, Stvarna vrijednost: 0.27
Predviđanje: 1.5336363749350008, Stvarna vrijednost: 1.47
```

Slika 5.3. Predviđanja.

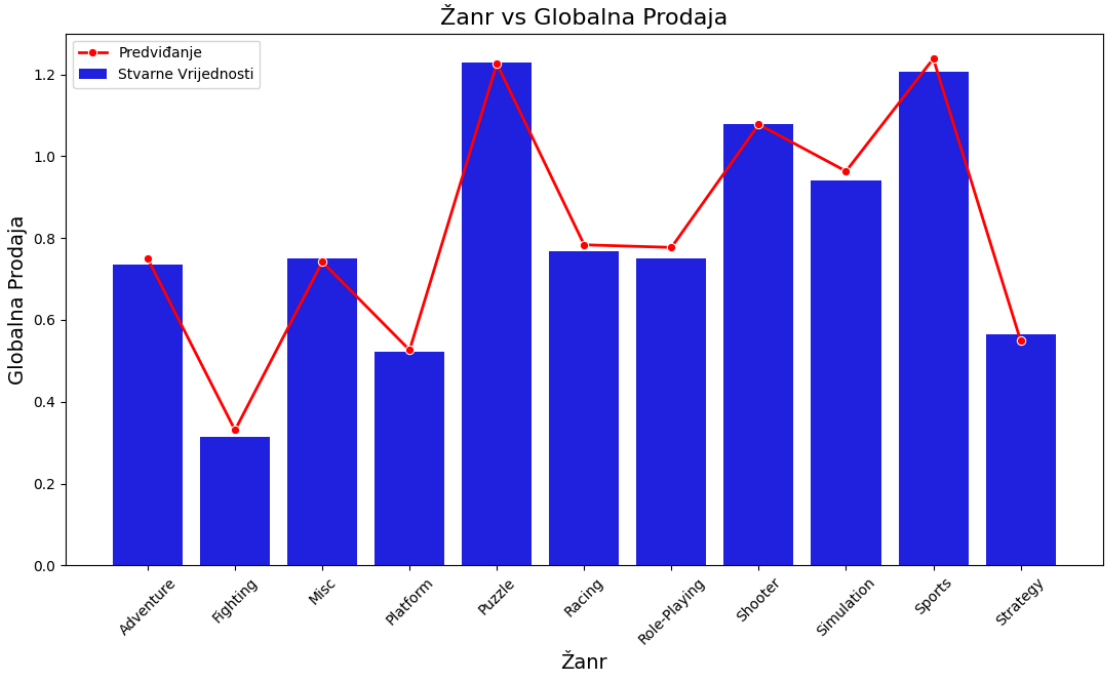
Ovaj ispis (promatra se zadnji) označava sljedeće:

- predviđanje: 1.5336363749350008 → Ovo je vrijednost koju je model linearne regresije predvidio za globalnu prodaju video igara na temelju ulaznih značajki. U ovom slučaju, model predviđa da će globalna prodaja biti približno 1.53 milijuna primjeraka.
- stvarna vrijednost: 1.47 → Ovo je stvarna (realna) vrijednost globalne prodaje za taj primjer. U ovom slučaju, stvarna prodaja bila je 1.47 milijuna primjeraka.

U okviru analize linearne regresije, nekoliko ključnih statističkih pokazatelja omogućava bolje razumijevanje utjecaja različitih faktora na globalnu prodaju video igara. Ovi pokazatelji uključuju koeficijente regresije, p-vrijednosti i koeficijent determinacije. Koeficijenti regresije za pozitivnu vrijednost varijable pokazuju kako ta varijabla utječe na globalnu prodaju. Kada je p-vrijednost za određenu varijablu ispod zadanog praga (u ovom slučaju 0.5), to ukazuje na to da ta varijabla doprinosi globalnoj prodaji. Koeficijent determinacije mjeri koliko dobro model objašnjava varijabilnost u podacima. Vrijednost bliža 1 označava bolji model, što implicira da su predviđanja preciznija i da model dobro odražava stvarne odnose između varijabli. S obzirom na te faktore, može se zaključiti kako:

- igre koje se dobro prodaju u Sjevernoj Americi i Europi imaju veći utjecaj na globalnu prodaju. Lokalizacija i marketinške strategije usmjerene prema tim tržištima mogu povećati globalni uspjeh.
- platforme PlayStation: igre za PlayStation konzole imaju značajan pozitivni utjecaj na prodaju, pa je preporučljivo razvijati igre za ove platforme.
- ocjene za široku publiku: Igre s ocjenama E i T omogućuju doseg široke publike i veće tržište, dok igre s ocjenom M imaju velik utjecaj među starijim igračima.
- visoko ocijenjeni razvojni timovi: rad s renomiranim studijima kao što su Rockstar North i Polyphony Digital može značajno povećati šanse za uspjeh na tržištu.

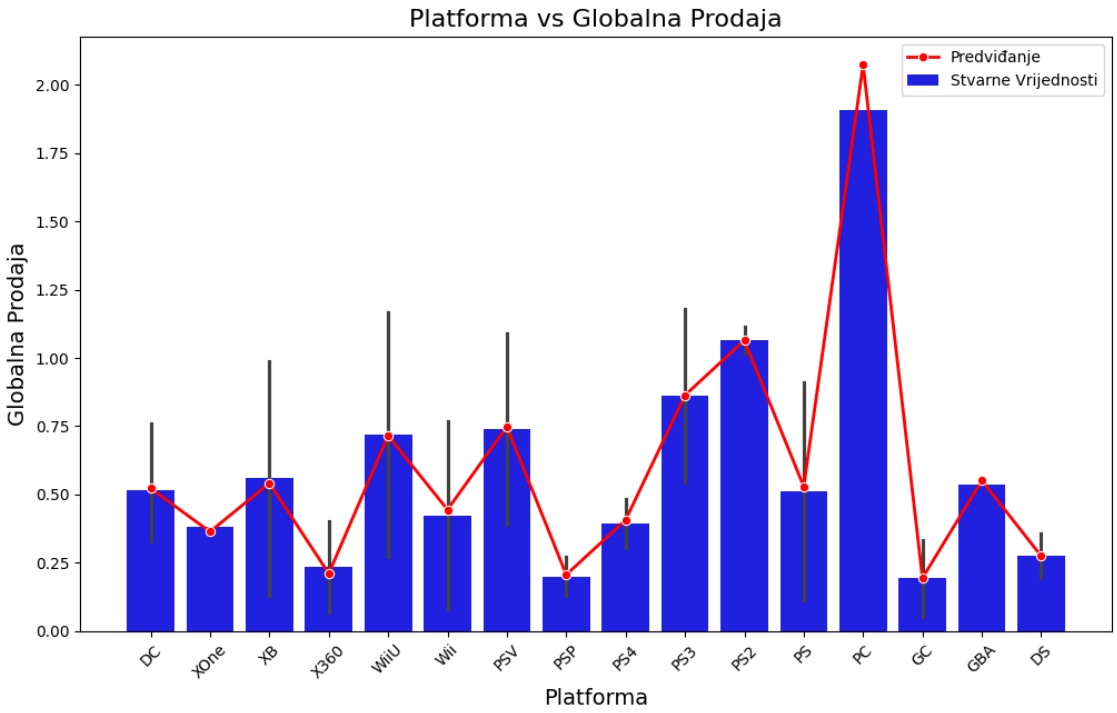
Kako bi vizualizirali podatke te ovisnost određenih značajki znatno pomaže biblioteka pyplot te su slikama 5.4., 5.5., 5.6. prikazane razlike između stvarnih vrijednosti i predviđanja. Za y-os je odabrana značajka *Global_Sales*, dok se x-os mijenja kako je prikazano. U primjeru x-osi su: *Genre, Platform, EU_Sales, NA_Sales*.



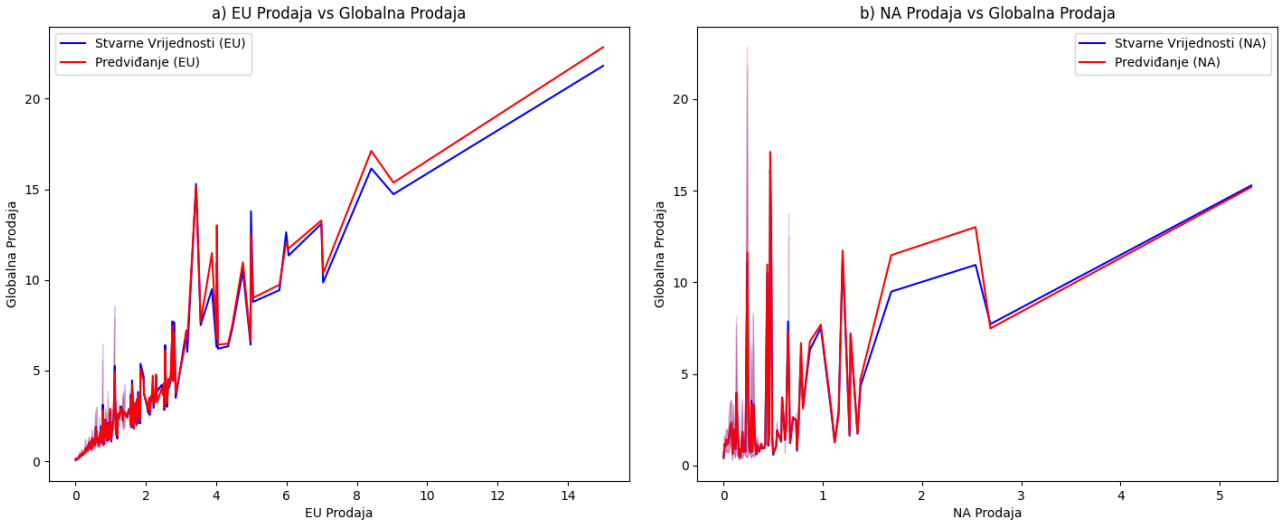
Slika 5.4. Predviđanja i stvarne vrijednosti globalne prodaje video igara ovisno o žanru.

Slika 5.4. prikazuje ovisnost globalne prodaje video igara o žanru igre. Plavi stupci predstavljaju prosječnu stvarnu globalnu prodaju videoigara za svaki žanr u testnom skupu. Na x-osi su prikazani različiti žanrovi, dok y-os pokazuje prosječnu globalnu prodaju u milijunima kopija. Crvena linija prikazuje prosječno predviđanje globalne prodaje za svaki žanr, koje je model linearne regresije izračunao. Kombinacijom toga, moguće je usporediti stvarne vrijednosti od onih predviđenih. Primjerice, prosječna stvarna globalna prodaja video igara žanra *Platform* je nešto manja od milijun primjeraka, dok je predviđena vrijednost nešto veća. Ovo vrijedi i za ostale grafove.

Slika 5.5. prikazuje ovisnost globalne prodaje video igara o platformi, dok slika 5.6. ovisnost globalne prodaje o prodaji u Europi i Sjevernoj Americi. Slika (a) prikazuje prodaju u Europi (označena kao EU), a slika (b) prodaju u Sjevernoj Americi (označena kao NA).

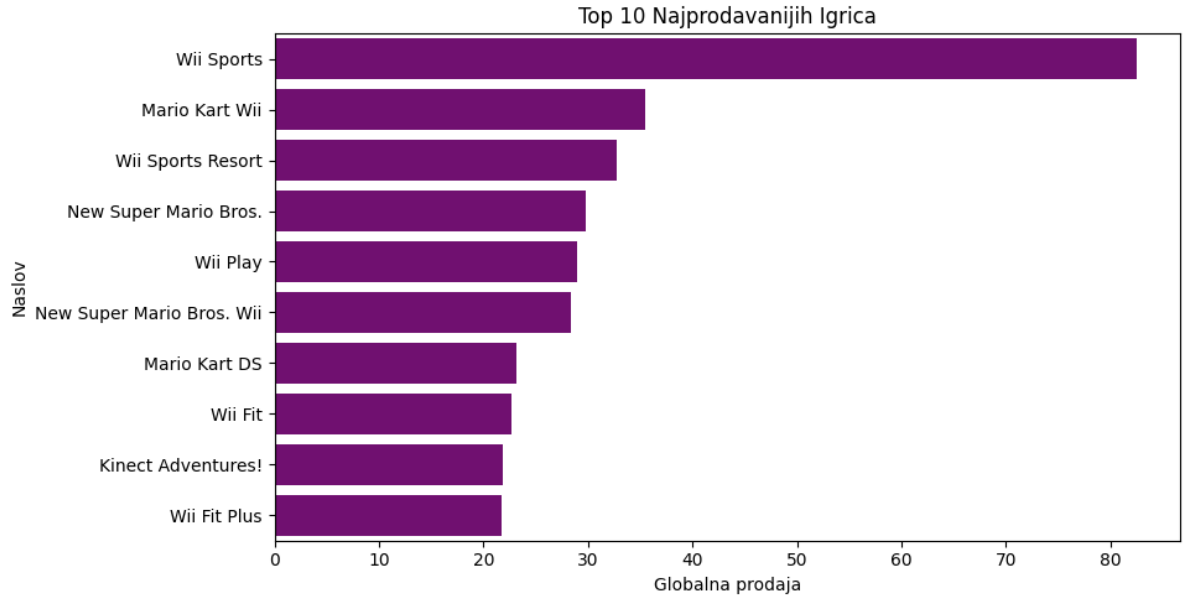


Slika 5.5. Predviđanja i stvarne vrijednosti globalne prodaje video igara ovisno o platformi.



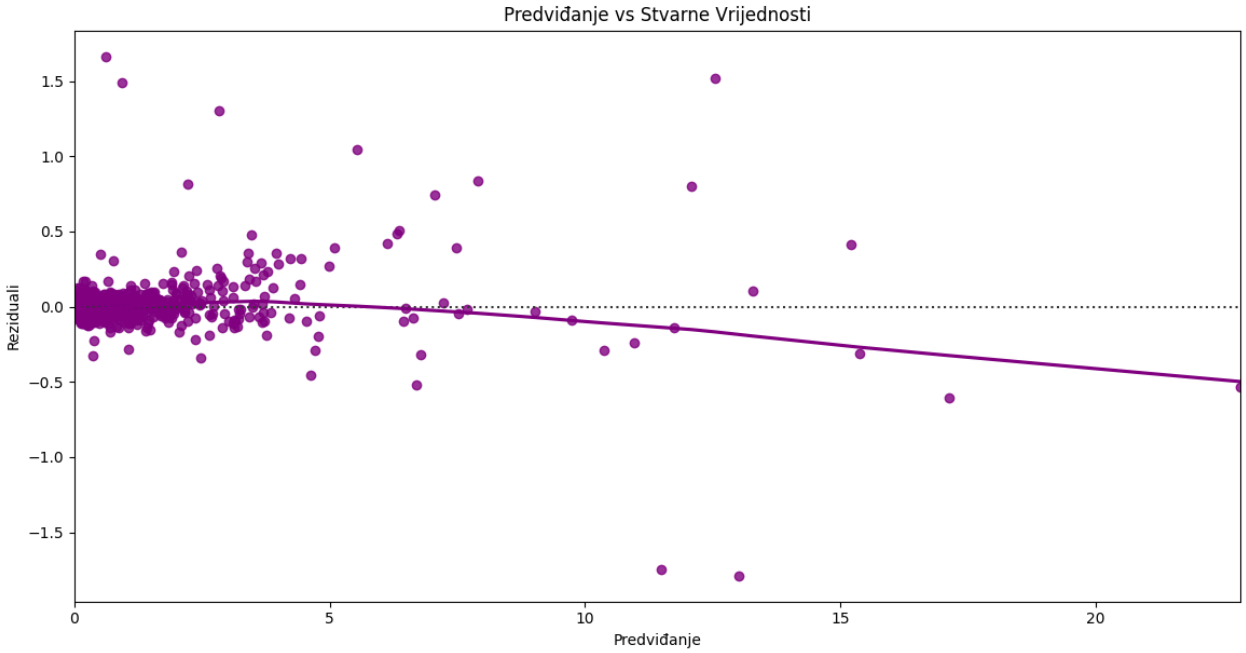
Slika 5.6. Usporedba prodaje u Europi i Sjevernoj Americi.

Kada se radi o samom naslovu igre, lista najprodavanijih igara prikazana je slikom 5.7.



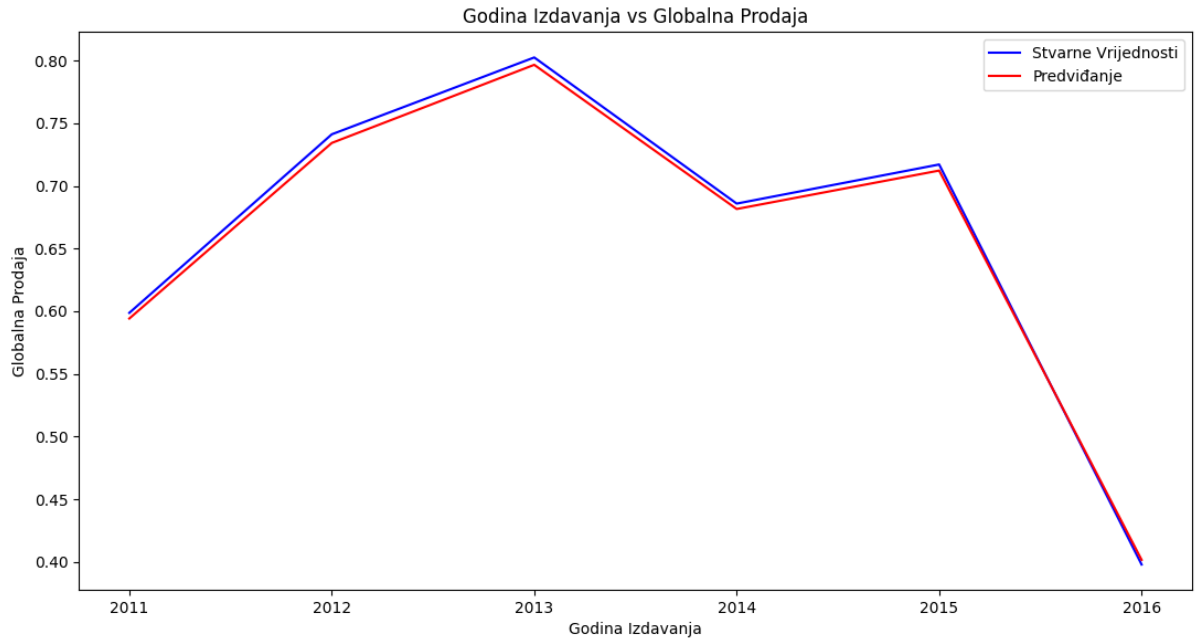
Slika 5.7. Najprodavanije video igre.

O uspješnosti i točnosti linearne regresije svjedoči slika 5.8. Rezidualni graf, često nazivan rezidualni dijagram, pomaže u vizualizaciji i procjeni kvalitete modela regresije. Cilj primjene linearne regresije u ovom kontekstu je predviđanje određene kvantitativne vrijednosti – u ovom slučaju, globalne prodaje videoigara na temelju različitih ulaznih značajki kao što su platforma, godina izdanja, žanr.



Slika 5.8. Evaluacija algoritma LR.

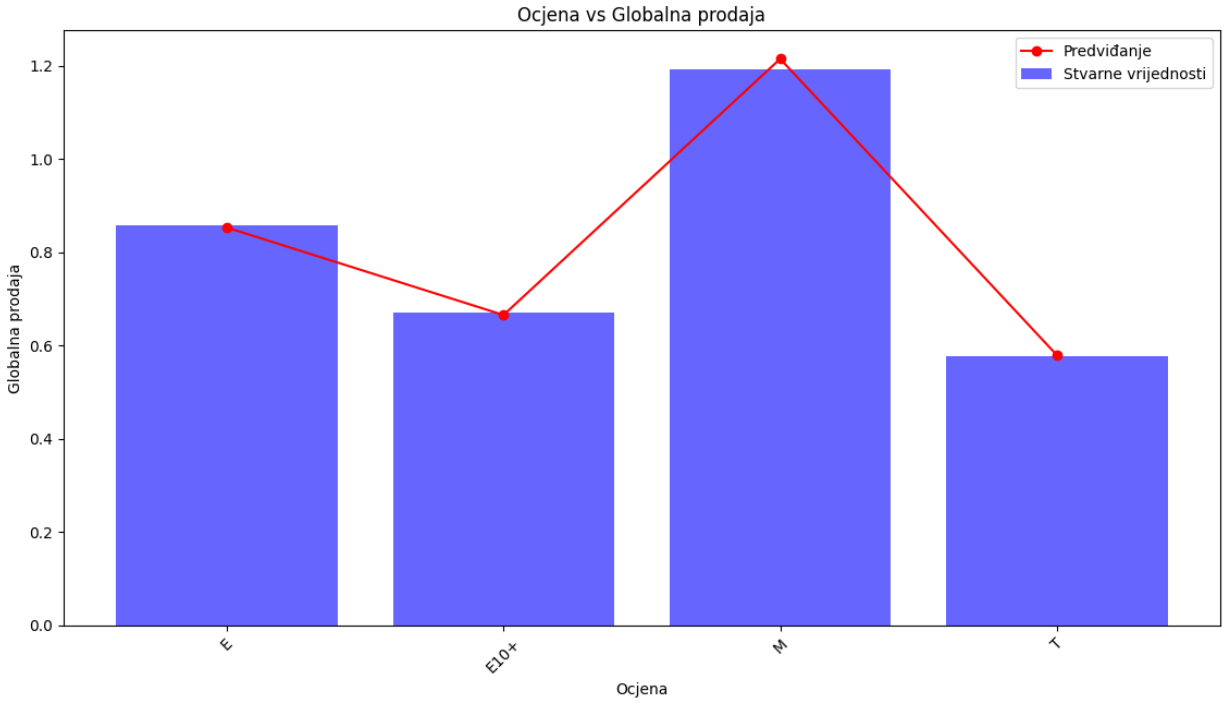
Drugi algoritam, *Random Forest* (RF), je na sličan način primijenjen kao i prošli, već spomenuti algoritam. Baza podataka ostaje ista te samim time i način učitavanja tih podataka, dok u samom kodu postoje određene razlike. Sami kod se neće detaljno komentirati, već samo rezultati dobiveni njime. Za razliku od prošlog algoritma, ovdje su odabrane neke druge značajke po kojima se gledala stvarna vrijednost i predviđanja. Za početak, slikom 5.9. je prikazan graf ovisnosti globalne prodaje video igara o godini izdavanja igre. Vrijeme je ključno prilikom podjele na trening i test skupove u ovom slučaju jer se podaci odnose na prodaju video igara, a ta prodaja zavisi od vremenskog toka (godina izdavanja igre). Podaci nisu nasumični jer su starije igre bile izdane prije novih igara, što znači da bi igre iz prošlih godina mogle imati drugačije tržišne uvjete u odnosu na novije igre. Pošto se radi o složenijim podacima, ovo se nije moglo izvesti linearnom regresijom, već je potreban složeniji algoritam te RF odgovora zahtjevima.



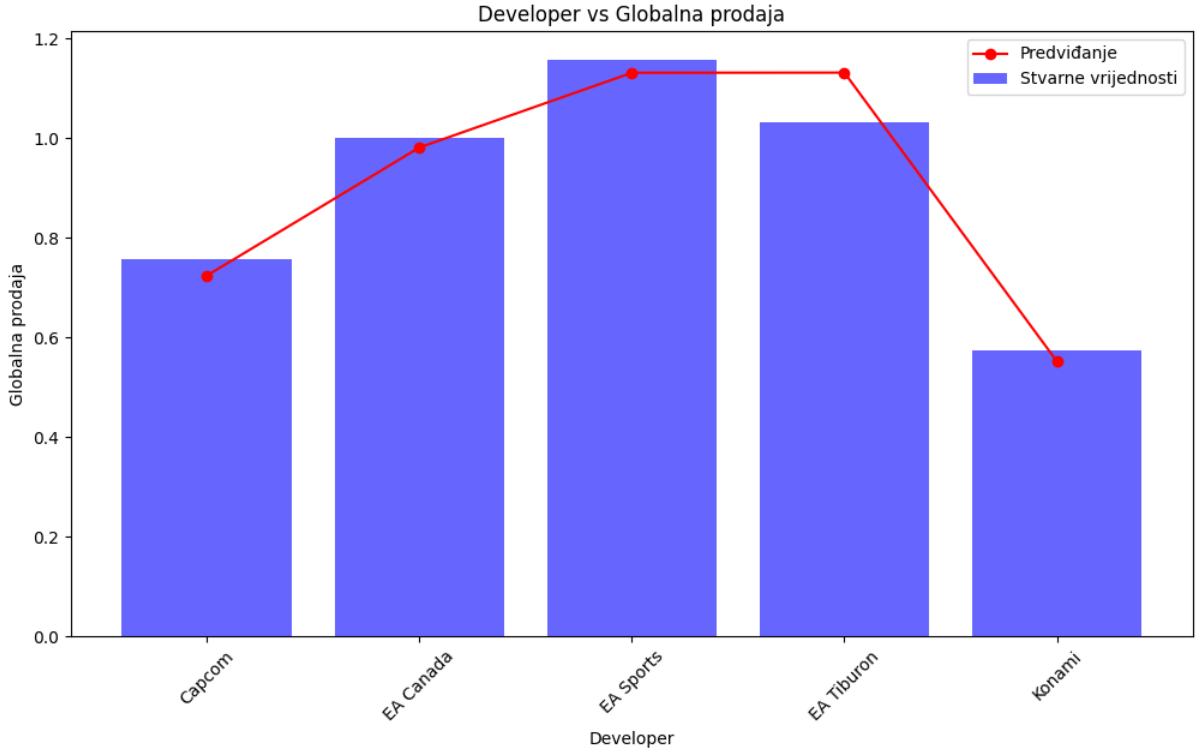
Slika 5.9. Predviđanja i stvarne vrijednosti globalne prodaje video igara ovisno o godini izdavanja igre.

Slikama 5.10., 5.11. te 5.12. su prikazane ovisnosti značajki *Rating*, *Developer* i *Critic Score* o značajki *Global Sales*, odnosno globalnoj prodaji video igara. Slika 5.10. predstavlja ovisnost globalne prodaje video igara o ocjeni. Izdvojene su neki od mogućih rejtinga te se na njima prikazuju prosječne stvarne prodaje i one predviđene. Što predstavljaju stupci, a što linija je već objašnjeno ranije. Slika 5.11. također prikazuje na x-osi samo neke od developera jer ih je izrazito puno i bilo bi nepregledno kada bi se postavili svi. Izdvojeni ih je pet te na njima primijenjen algoritam. Što se tiče ocjene kritičara, ocjena se kreće od nula do sto (kako i prikazuje graf na slici

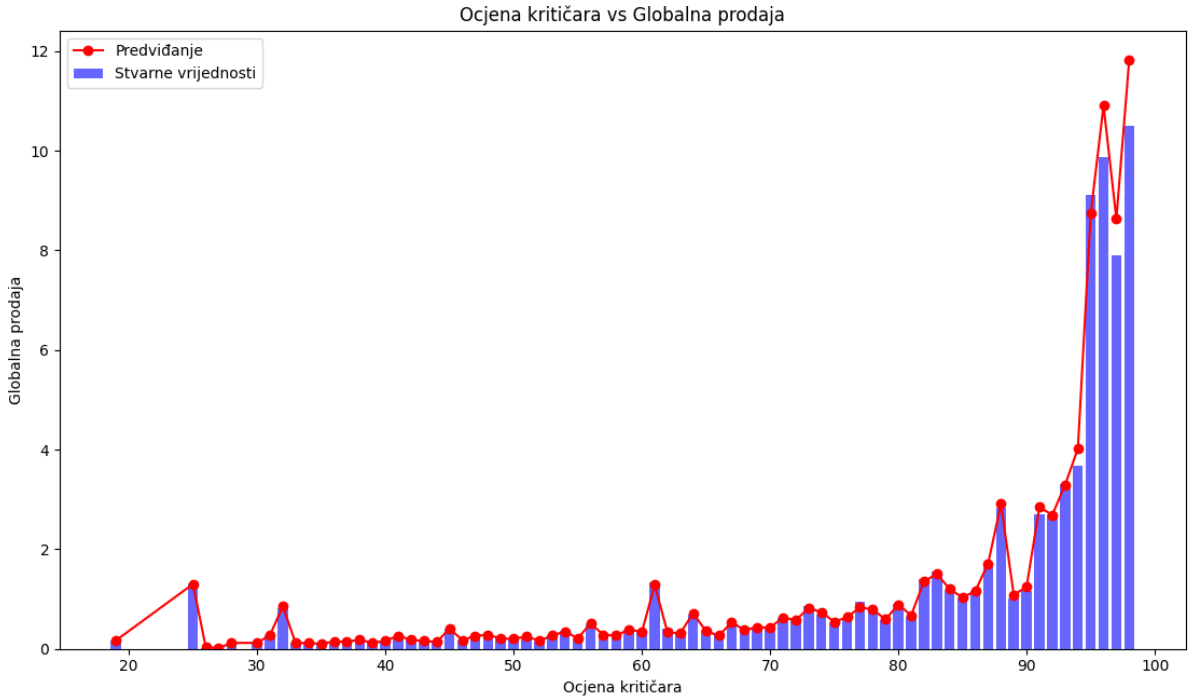
5.12.) te je za svaku ocjenu prikazana vrijednost koja se odnosi na prosječnu globalnu prodaju video igara u milijunima primjeraka.



Slika 5.10. Predviđanja i stvarne vrijednosti globalne prodaje video igara ovisno o ocjeni.

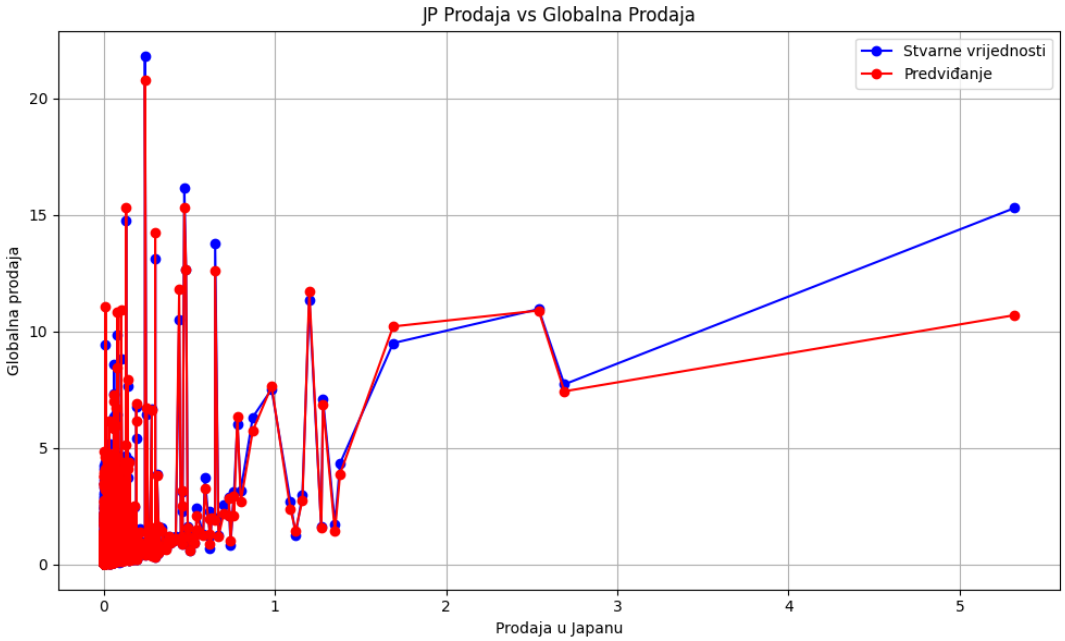


Slika 5.11. Predviđanja i stvarne vrijednosti globalne prodaje video igara ovisno o developeru.



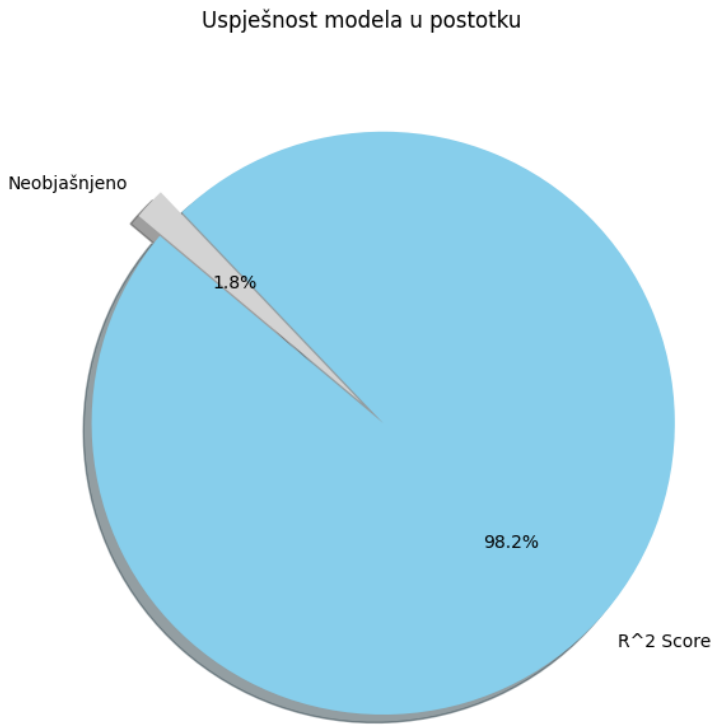
Slika 5.12. Predviđanja i stvarne vrijednosti globalne prodaje video igara ovisno o ocjeni kritičara.

S prošlim algoritmom se uspoređivao odnos prodaje u regijama kao što su Europa i Sjeverna Amerika. RF algoritmom se fokusiralo na prodaju video igara u Japanu, pošto je *JP_Sales* također jedna od značajki koju pruža baza podataka korištena u ovom projektu. Slikom 5.13. je prikazan odnos stvarnih vrijednosti te predviđanja vezanih za prodaju video igara.



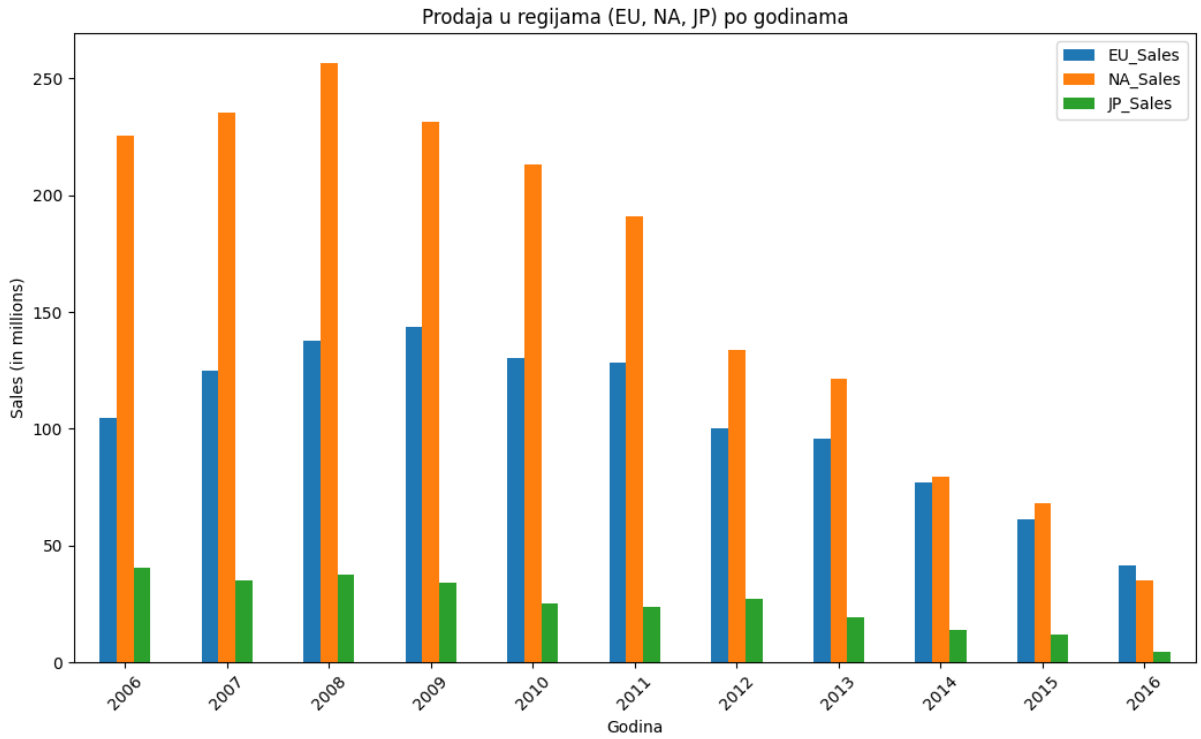
Slika 5.13. Prodaja video igara u Japanu – predviđanja i stvarne vrijednosti.

Uspješnost RF algoritma prikazana je slikom 5.14. gdje je prikazan koeficijent determinacije. Pokazuje koliko dobro neovisna varijabla (prediktor) objašnjava varijaciju u zavisnoj varijabli (cilj). U kontekstu modela regresije, to je mjera koliko dobro model objašnjava promjene u ciljnim vrijednostima. Iz grafa se može vidjeti velika učinkovitost i pokrivenost podataka.



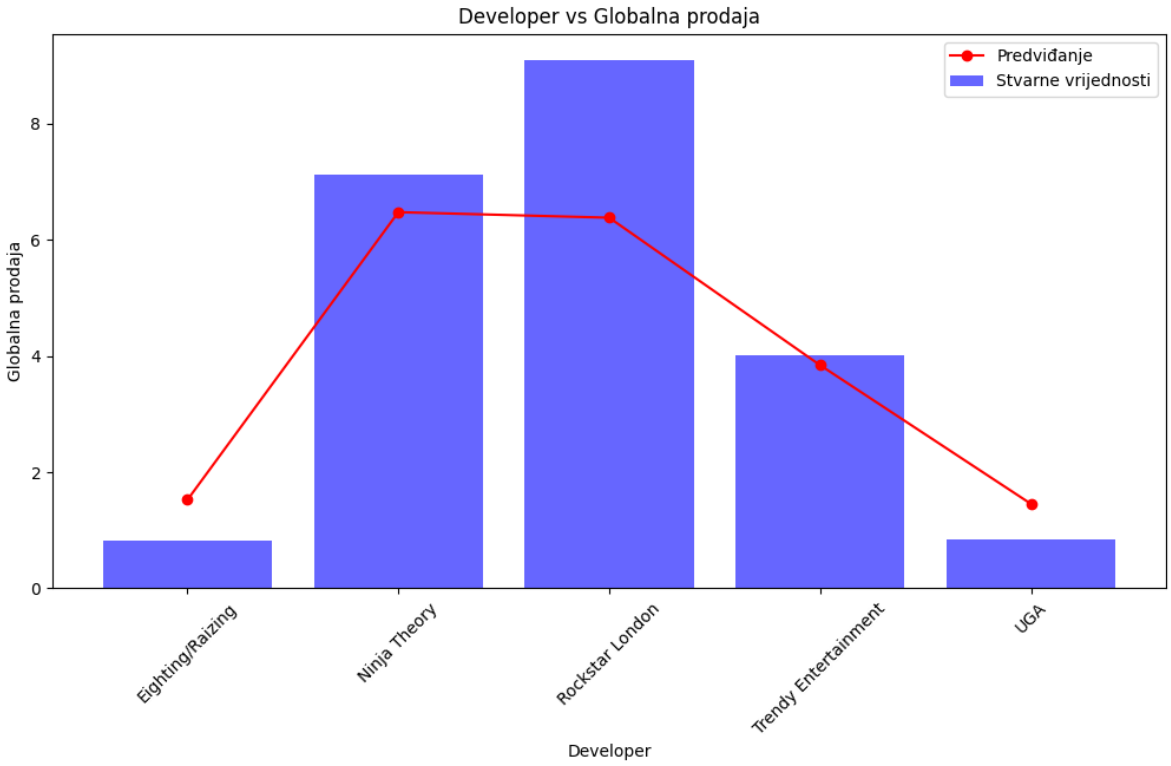
Slika 5.14. Koeficijent determinacije.

Nakon LR i RF, na redu je k – Nearest Neighbour algoritam. KNN se koristi za rješavanje problema i klasifikacije i regresije te je zbog toga ovdje korišten. Prethodnim algoritmima bavilo se analizama predviđanja i stvarnih vrijednosti prodaja video igara u različitim regijama. Slikom 5.15. je prikazana usporedba prodaje u rasponu od 2006. do 2016. godine.

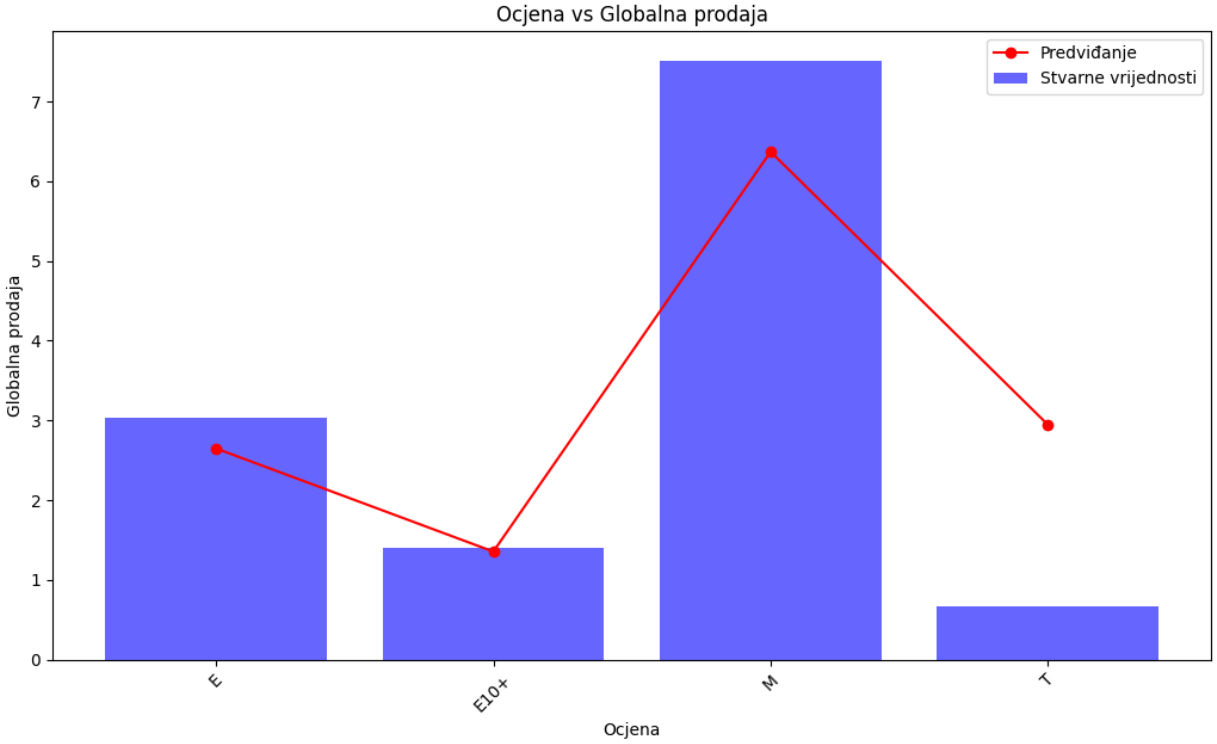


Slika 5.15. Prodaja u EU, NA, JP.

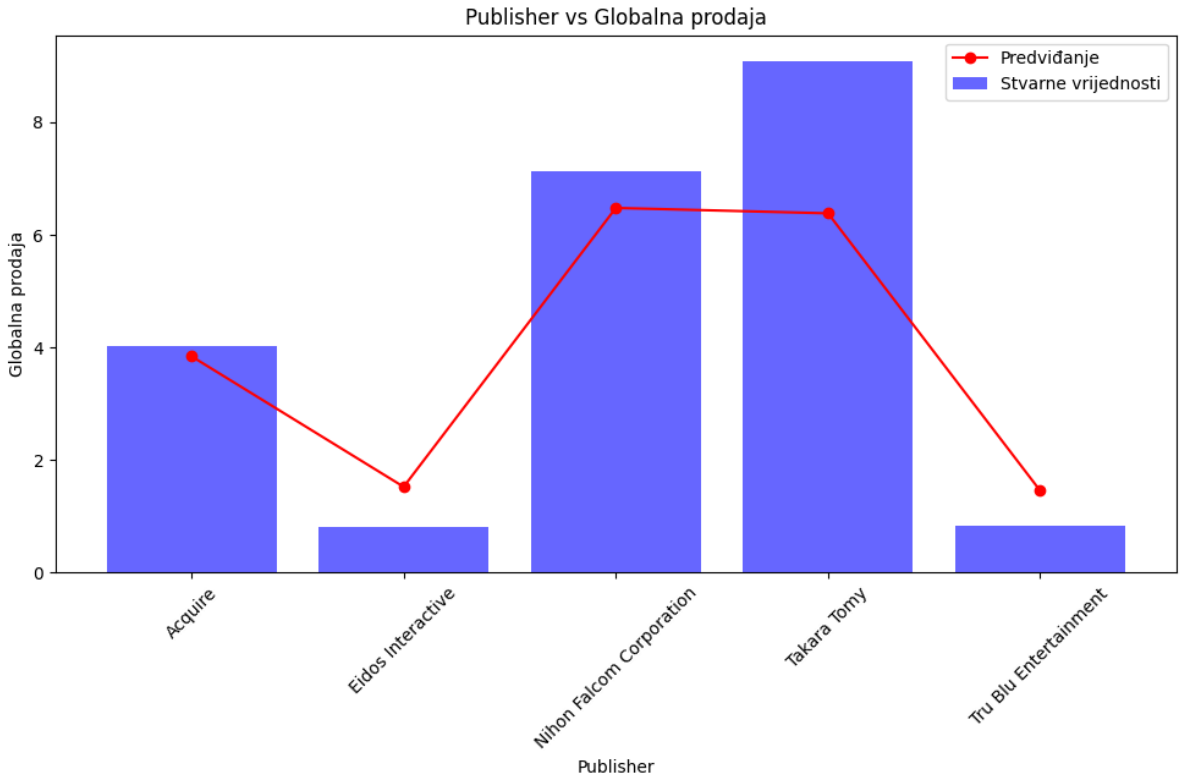
Ovdje su korištene *Publisher*, *Developer* i *Rating* značajke koje se uspoređuju s globalnom prodajom video igara kako bi se potencijalno popravila predviđanja za *developere*, *rating* i *publishere*. U analizi su korištene vizualizacije za usporedbu stvarnih vrijednosti i predviđanja za svaku od ovih značajki. Na slikama 5.16., 5.17., i 5.18. prikazane su usporedbe između stvarnih prodajnih rezultata i predviđanja generiranih modelom. Ove slike omogućuju uvid u to kako precizno model može predvidjeti globalnu prodaju temeljem različitih značajki. Posebna pažnja posvećena je značajci *Rating* jer se može dogoditi da neka video igra nema definiran rating, što može utjecati na točnost predviđanja. U takvim slučajevima, gdje je rating nedostajao, u analizama su označeni kao NaN (Not a Number). To je značajna informacija koja ukazuje na potencijalnu prazninu u podacima koja može utjecati na cjelokupnu kvalitetu modela te su oni uklonjeni prije računanja.



Slika 5.16. Predviđanja i stvarne vrijednosti globalne prodaje video igara ovisno o developeru.

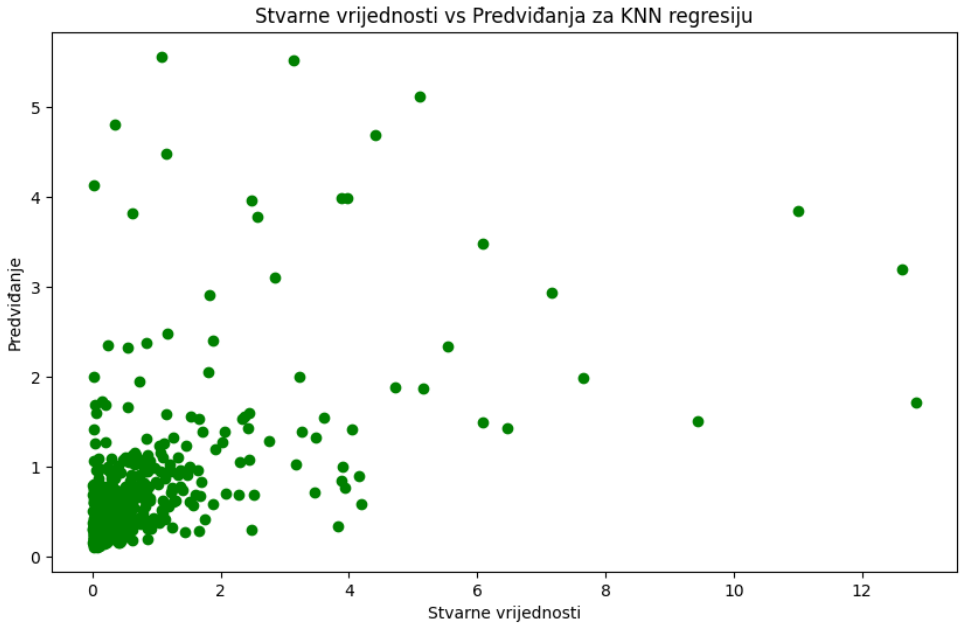


Slika 5.17. Predviđanja i stvarne vrijednosti globalne prodaje video igara ovisno o ocjeni.

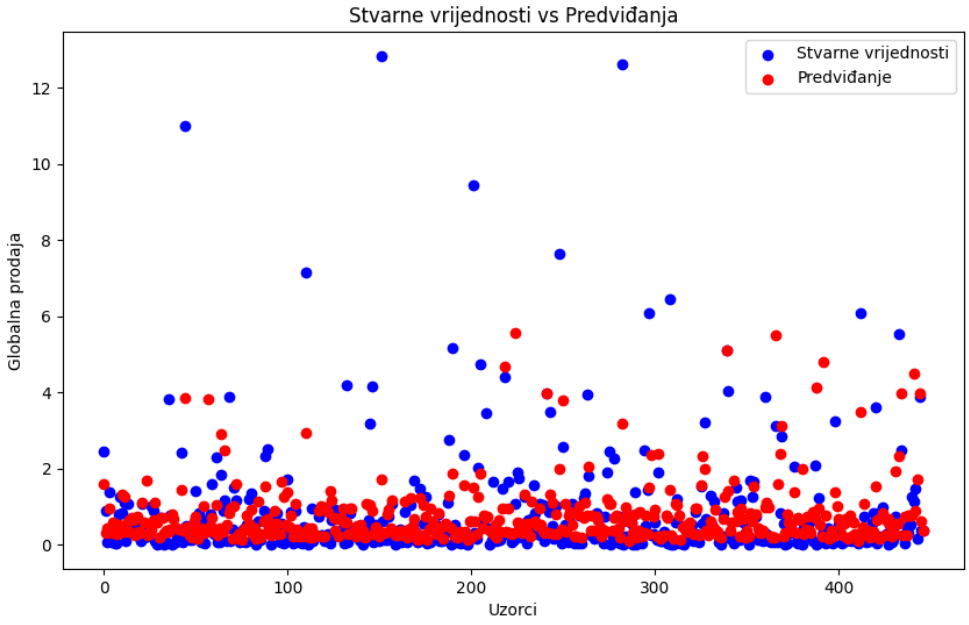


Slika 5.18. Predviđanja i stvarne vrijednosti globalne prodaje video igara ovisno o izdavaču.

Rezultati primjene KNN algoritma na skupu podataka o prodaji video igara pokazuju da trenutni model ne uspijeva postići zadovoljavajuće performanse. Konkretno, R^2 skor modela je 0.4035, što sugerira da model objašnjava samo oko 40.4% varijabilnosti u ciljnim varijablama, dok ostatak varijabilnosti ostaje neobjašnjen. Stoga, iz ovih rezultata se može zaključiti kao ovo nije najbolji algoritam koji se može primijeniti za ova predviđanja. Slikama 5.19. i 5.20. su prikazani ukratko rezultati dobiveni ovim algoritmom.



Slika 5.19. Stvarne vrijednosti i predviđanja.

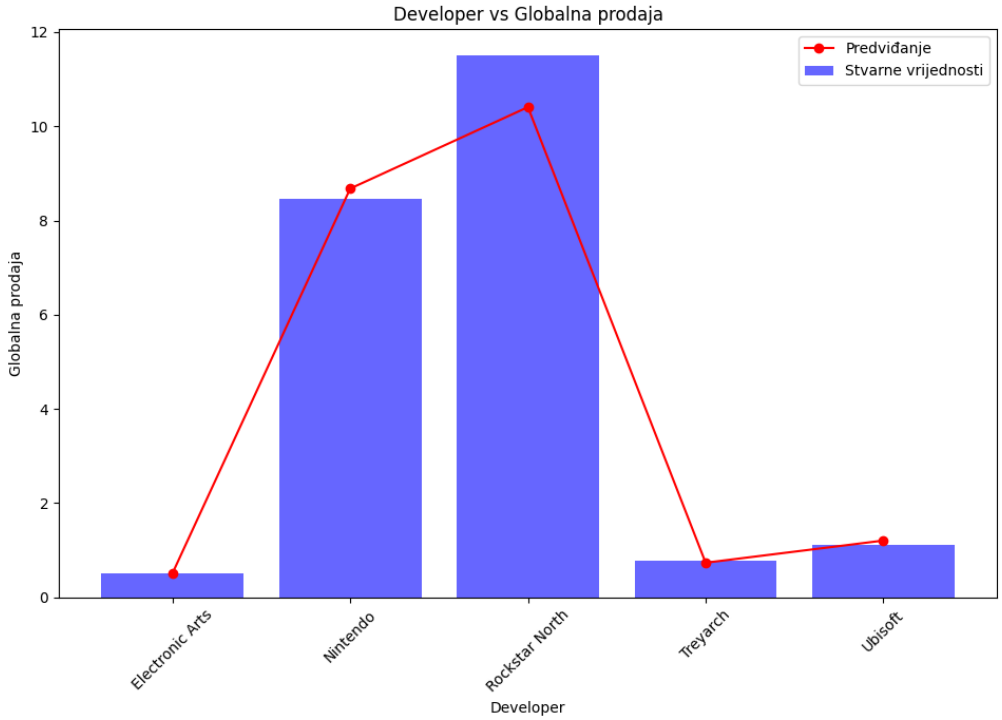


Slika 5.20. Uzorci.

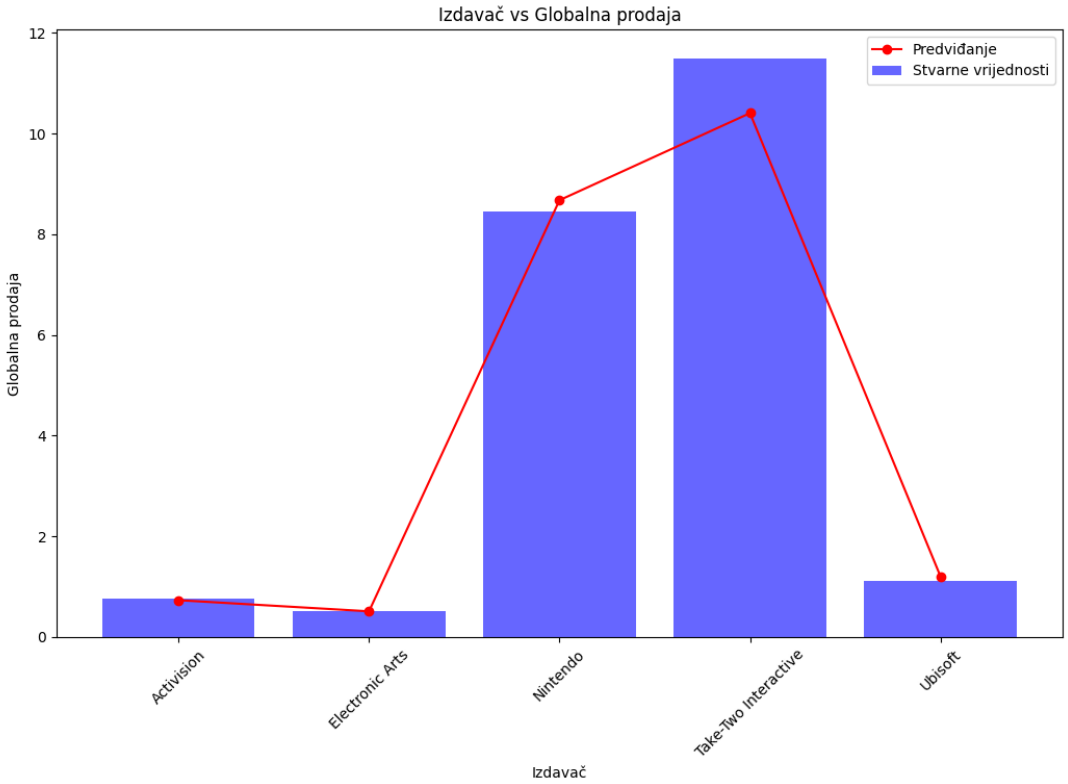
Zbog nedovoljno zadovoljavajućih rezultata dobivenih primjenom prethodnog algoritma, odlučeno je koristiti novi algoritam, *Gradient Boosting*, za iste značajke. G*radient Boosting Regressor* je vrlo popularan i moćan model koji se koristi za regresiju i klasifikaciju, poznat po svojoj sposobnosti da kombinira nekoliko slabih modela u jedan snažan model kroz *ensembling* metode. Ova tehnika omogućuje poboljšanje točnosti predviđanja, što često rezultira visokim R^2

koeficijentom ili koeficijentom determinacije, što je indikator dobrog prilagođavanja modela na podacima [30].

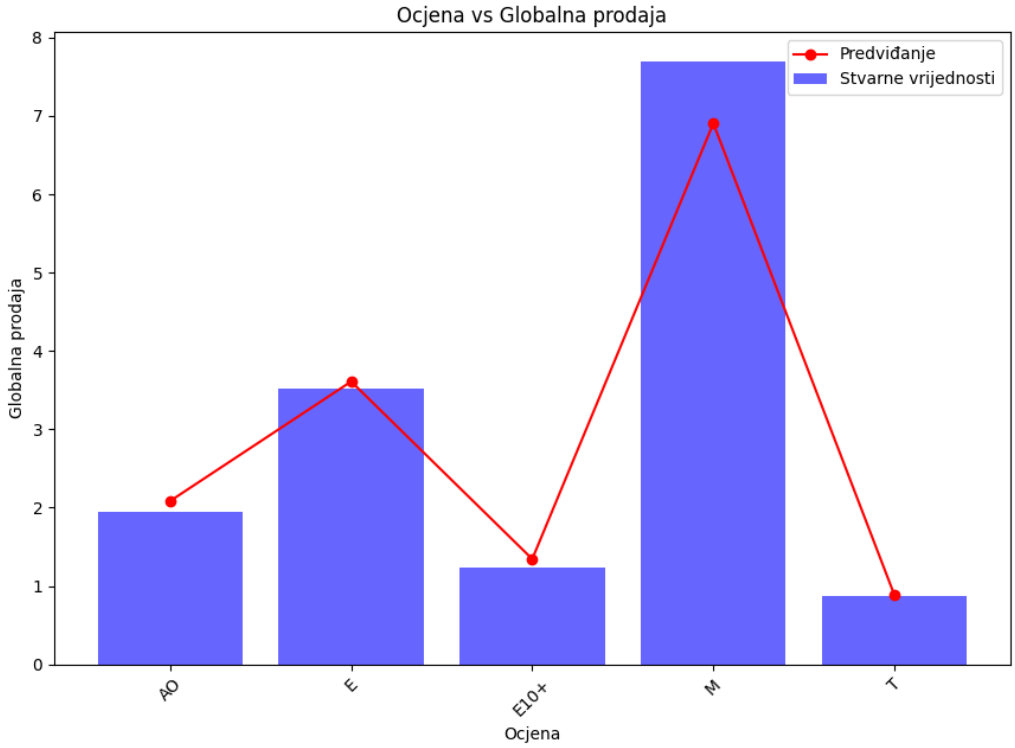
U ovoj analizi koristit će se tri ključne značajke koje utječu na globalnu prodaju, uz odabir relevantnih *developera* i *publishera*. Baš kao i kod prethodne primjene kNN algoritma, cilj je usporediti učinkovitost različitih algoritama na istom skupu podataka. Rezultati predviđanja i stvarnih vrijednosti prikazani su na slikama 5.21., 5.22., i 5.23. Usporedbom ovih grafova, jasno se može primijetiti kako su predviđanja dobivena *Gradient Boosting* algoritmom bliže stvarnim vrijednostima, što potvrđuje njegovu veću učinkovitost u odnosu na prethodno korišteni model.



Slika 5.21. Predviđanja i stvarne vrijednosti globalne prodaje video igara ovisno o developeru (GBR algoritam).

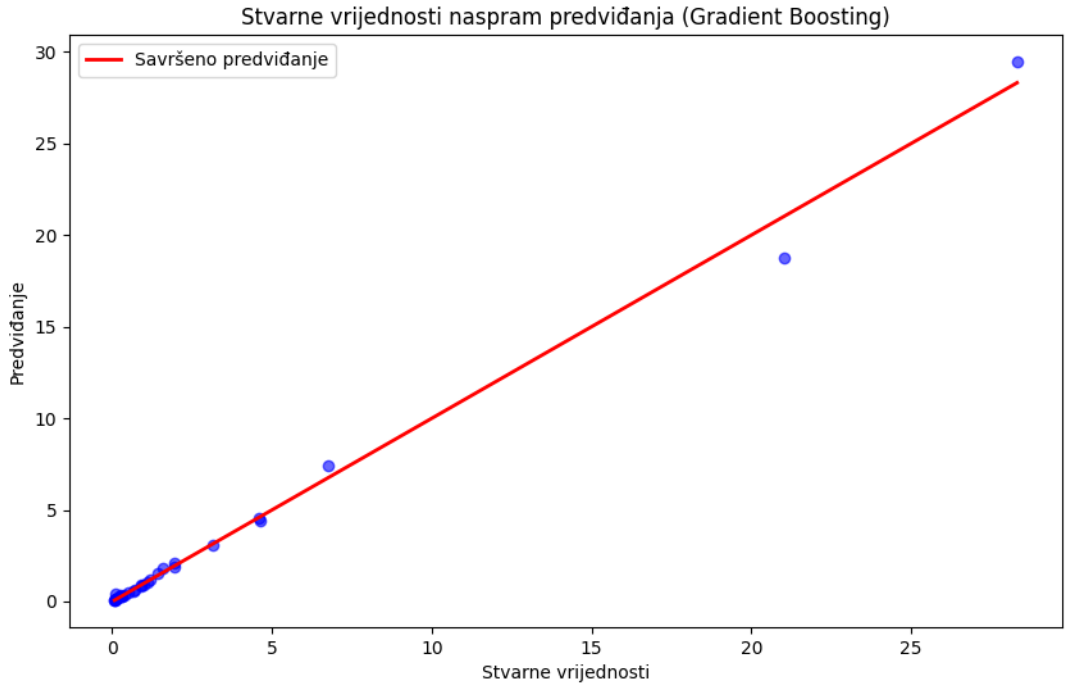


Slika 5.22. Predviđanja i stvarne vrijednosti globalne prodaje video igara ovisno o izdavaču (GBR algoritam).



Slika 5.23. Predviđanja i stvarne vrijednosti globalne prodaje video igara ovisno o ocjeni (GBR algoritam).

Na slici 5.24. prikazani graf ilustrira usporedbu stvarnih vrijednosti i predviđanja dobivenih korištenjem odabranog algoritma. Svaka točka na grafu predstavlja jedan podatak iz testnog skupa, gdje je x-osi prikazana stvarna vrijednost, a na y-osi predviđanje modela. Crvena linija na grafu označava savršena predviđanja, odnosno točke gdje bi stvarne vrijednosti i predviđanja bili identični. Što su točke bliže ovoj liniji, to su predviđanja modela točnija. Graf pruža vizualni uvid u uspješnost algoritma, gdje se može primijetiti koliko se predviđanja podudaraju sa stvarnim vrijednostima. Rezultati ukazuju na visoku točnost modela, posebno u onim područjima gdje su točke grupirane blizu crvene linije.



Slika 5.24. Evaluacija algoritma.

5.3. Primjena neuronskih mreža

Uz pomoć osnovnih algoritama strojnog učenja koji se koriste za regresiju, postignuti rezultati su pokazali značajnu razinu uspješnosti. Kroz proces istraživanja ispitivane su različite značajke skupa podataka, a potom su rezultati analizirani i prikazani kroz grafove koji uspoređuju predviđanja sa stvarnim vrijednostima. Ovi grafovi pružaju vizualni uvid u točnost modela, omogućujući lakše prepoznavanje obrazaca te identifikaciju područja u kojima su predviđanja najtočnija, kao i onih gdje postoji odstupanje. Pored toga, detaljno su analizirani različiti modeli kako bi se razumjela njihova sposobnost generalizacije na nove, nepoznate podatke. Uočeno je da su neki modeli pokazali veću robusnost i točnost u predviđanju globalne prodaje videoigara, dok su drugi bili osjetljiviji na određene vrste podataka.

Primjenom neuronskih mreža, koje su objašnjene u ranijim poglavljima, nastojalo se ispitati učinkovitost modela. Neuronske mreže, sa svojom sposobnošću da modeliraju složene nelinearne odnose, predstavljaju napredniji pristup koji je sposoban bolje obraditi složenost i heterogenost podataka. Kroz ovaj pristup, cilj je bio poboljšati točnost predviđanja te smanjiti pogreške koje su prisutne kod jednostavnijih regresijskih modela.

Jedan od ključnih koraka u primjeni neuronskih mreža je definicija samog modela. U ovom slučaju, model je definiran korištenjem *Keras* biblioteke koja omogućava jednostavno i intuitivno stvaranje slojevitih neuronskih mreža. Model je izgrađen kao sekvencijalna mreža, što znači da se slojevi dodaju jedan za drugim, u linearnom slijedu. Slikom 5.25. je definiran model za ovaj projekt:

```
model = keras.Sequential([
    keras.layers.Dense(128, activation="relu",
input_shape=(X_train.shape[1],)),
    keras.layers.Dense(64, activation="relu"),
    keras.layers.Dense(32, activation="relu"),
    keras.layers.Dense(1)
])
```

Slika 5.25. Model neuronske mreže.

Ovdje je model sastavljen od četiri sloja:

- prvi sloj je gusto povezani (*Dense*) sloj s 128 neurona i aktivacijskom funkcijom *relu* (engl. *Rectified Linear Unit*). Aktivacijska funkcija relu je često korištena jer uvodi nelinearnost u mrežu, čime omogućava modelu da uči složene obrasce u podacima. Ovaj sloj također definira *input_shape*, odnosno oblik ulaznih podataka, što je ovdje broj značajki (*feature*-a) u skupu za treniranje.
- drugi sloj je također gusto povezani sloj, ali s 64 neurona, i koristi istu relu aktivacijsku funkciju. Ovaj sloj dalje reducira dimenzionalnost podataka, ali zadržava dovoljnu količinu informacija potrebnih za učinkovitu predikciju.
- treći sloj ima 32 neurona, što dodatno smanjuje broj parametara i složenost mreže. I ovaj sloj koristi relu funkciju za aktivaciju, što omogućava nastavak modeliranja nelinearnih odnosa u podacima.
- izlazni sloj je završni sloj s jednim neuronom, što je prikladno za regresijske zadatke gdje je cilj predvidjeti jednu kontinuiranu vrijednost, u ovom slučaju globalnu prodaju videoigara. Ovdje nije korištena aktivacijska funkcija, jer regresijski modeli obično izlaze s kontinuiranom vrijednošću koja nije ograničena na neki specifičan raspon.

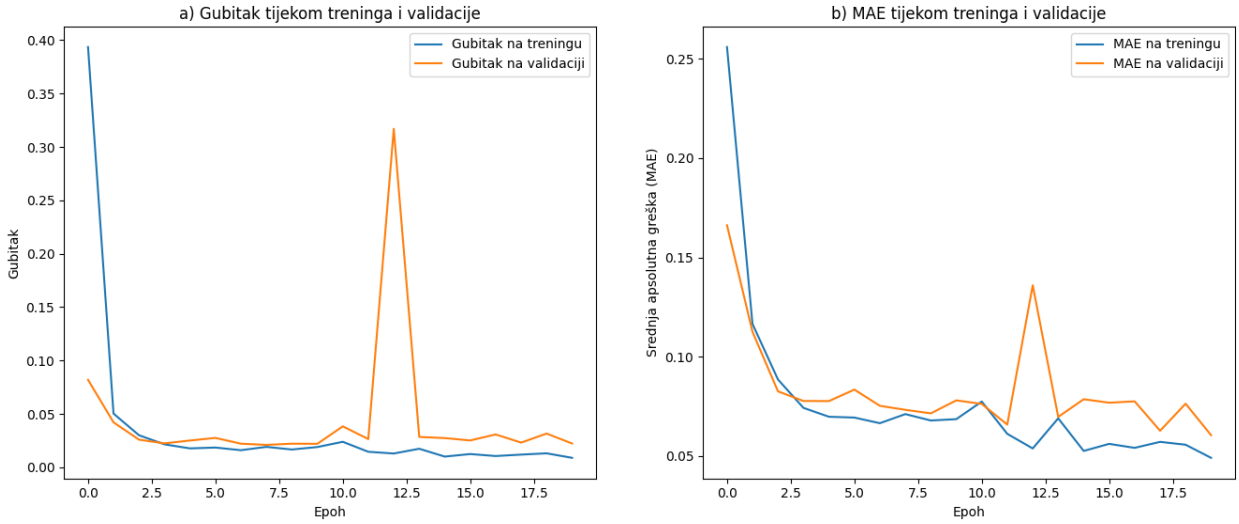
Nakon definiranja arhitekture modela, ključno je provesti treniranje modela kako bi on mogao naučiti obrasce u podacima. Proces treniranja modela kontrolira se pomoću određenih hiperparametara, a jedan od najvažnijih među njima je broj epoha. Takov treniranje modela je prikazano slikom 5.26. za ovaj slučaj:

```
history = model.fit(X_train, y_train, validation_data=(X_test, y_test),
epochs=20)
```

Slika 5.26. Treniranje modela

Treniranje modela sastoji se od višestrukih prolaza kroz skup podataka, gdje se svaki prolaz naziva epoha. U svakoj epohi model uči iz podataka, prilagođava svoje težine i poboljšava svoje predviđanje. Veći broj epoha omogućuje modelu da više puta prolazi kroz podatke i bolje uči obrasce. U ovom slučaju, odabrano je 20 epoha, što znači da će model proći kroz cijeli skup podataka 20 puta tijekom treniranja. Tijekom svakog prolaza (tj. epohe), model se trenira na skupu za treniranje (*X_train*, *y_train*) i evaluira na skupu za validaciju (*X_test*, *y_test*). Na taj način, *validation_data* se koristi kako bi se pratilo kako model radi na podacima koje nije vidio tijekom treniranja, što pomaže u detektiranju pretreniranja i optimizaciji hiperparametara. Nakon svake epohe, *history* objekt pohranjuje vrijednosti gubitka (engl. *loss*) i točnosti (engl. *accuracy*) za oba skupa podataka: onaj korišten za treniranje i onaj korišten za validaciju. Te informacije se koriste za generiranje grafova koji prikazuju trendove u učenju modela tijekom vremena.

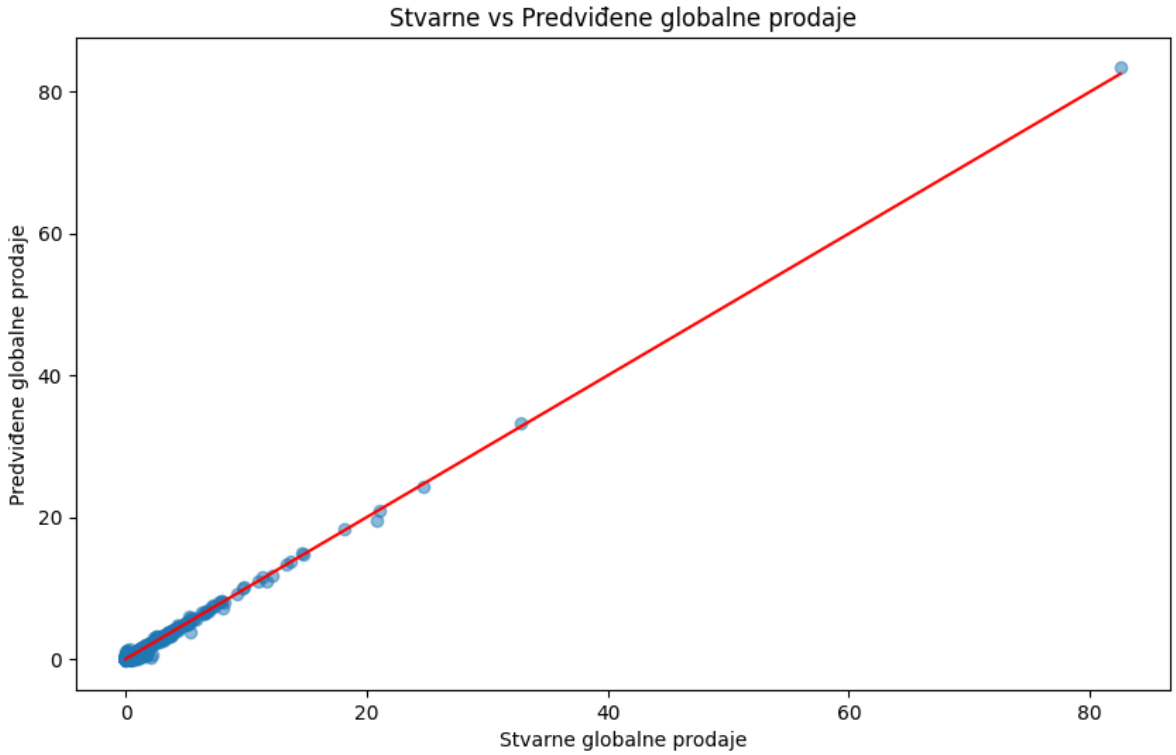
Grafovi povijesti treniranja prikazuju kako se gubitak (engl. *loss*) i srednja apsolutna greška (MAE) mijenjaju kroz epohe, kako za skup podataka za treniranje, tako i za validacijski skup te su ti rezultati prikazani slikom 5.27.



Slika 5.27. Gubitak i MAE.

Na slici 5.27., slika (a) prikazuje gubitak tijekom treninga i validacije. Ovaj graf omogućava praćenje kako se model poboljšava s vremenom. Idealno, gubitak na oba skupa (trening i validacija) trebao bi se smanjivati kako epohe napreduju. Ako primijetimo da se gubitak na validacijskom skupu prestane smanjivati ili počne povećavati dok se gubitak na skupu za treniranje nastavlja smanjivati, to može biti indikacija pretreniranja modela. Slika (b) prikazuje MAE kroz epohe za trening i validacijski skup. MAE je metrička vrijednost koja predstavlja prosječnu apsolutnu pogrešku između predviđenih i stvarnih vrijednosti. Smanjenje MAE tokom epoha ukazuje na poboljšanje preciznosti modela. Slično kao i kod gubitka, ako MAE na validacijskom skupu počne rasti, to može signalizirati da model previše uči iz podataka za treniranje i gubi sposobnost generalizacije.

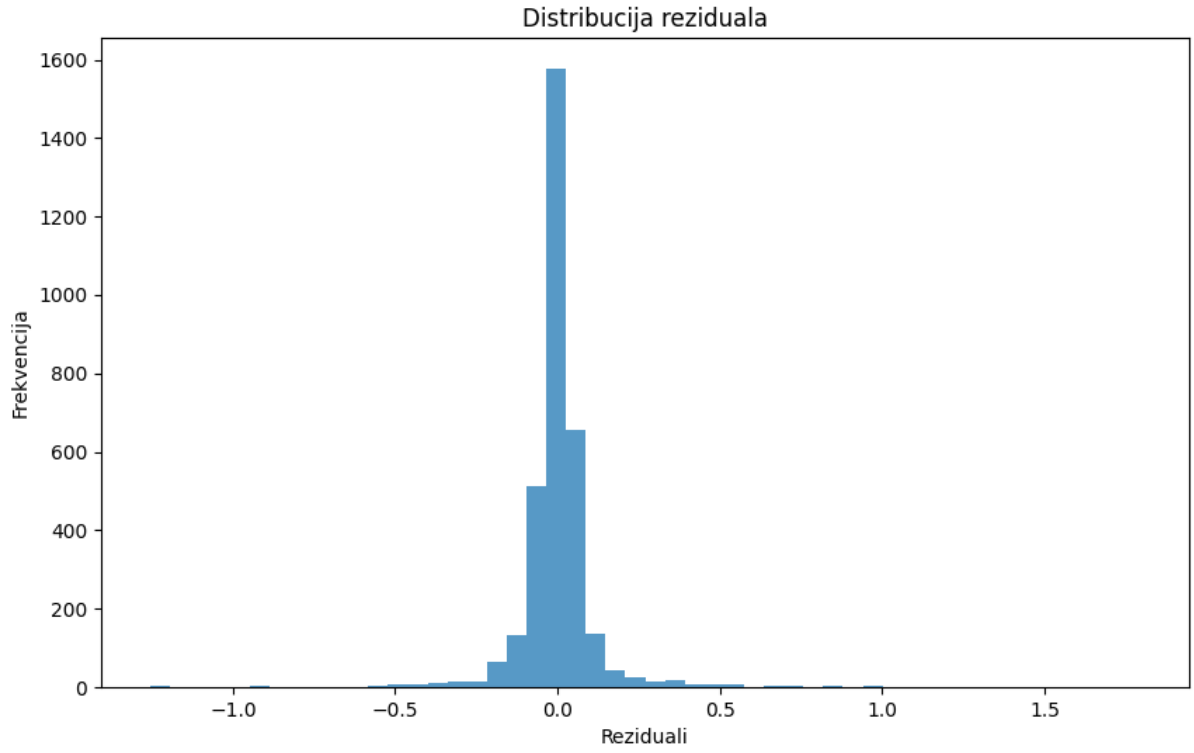
Slikom 5.28. su prikazane stvarne i predviđene vrijednosti globalne prodaje. Na ovom grafu, svaka točka predstavlja jednu igru iz skupa podataka za testiranje, pri čemu je pozicija točke određena stvarnom i predviđenom globalnom prodajom te igre. Crvena linija označava "liniju savršenstva", odnosno gdje bi sve točke bile smještene kada bi predviđanja bila savršena. Što su točke bliže crvenoj liniji, to su predviđanja modela preciznija.



Slika 5.28. Stvarne vrijednosti i predviđanja.

Rezidualna analiza omogućava dodatni uvid u razlike između predviđanja i stvarnih vrijednosti, što može pomoći u identifikaciji obrazaca ili anomaličnih podataka. Graf distribucije reziduala

prikazuje kako su pogreške raspodijeljene. Idealno, reziduali bi trebali biti raspoređeni simetrično oko nule, što bi značilo da model ne pokazuje sustavne pristranosti u predviđanjima. Ako se u grafu pojave neobične distribucije ili značajan broj velikih reziduala, to može ukazivati na specifične slučajeve gdje model ima poteškoća s predviđanjima. Distribuciju reziduala prikazuje slika 5.29.



Slika 5.29. Rezidualna analiza.

Evaluaciju algoritma moguće je detaljno ocijeniti analizom ispisanih metrika pri svakom pokretanju algoritma, kao što je prikazano tablicom 5.1. U ovom kontekstu, ključni statistički pokazatelji koji se koriste za ocjenjivanje performansi modela su *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE) i R^2*Score*.

Tablica 5.1. Performanse modela.

Srednja apsolutna pogreška	0.06275135433977055
Srednja kvadratna pogreška	0.017301050802903836
Korijen srednje kvadratne pogreške	0.13153345887227263
R^2 rezultat	0.99577670240722978

MAE je prosječna apsolutna greška između predviđanja modela i stvarnih vrijednosti. Ova metrika pokazuje koliko su u prosjeku predviđanja modela udaljena od stvarnih vrijednosti, bez obzira na

smjer (pozitivan ili negativan). U ovom slučaju, srednja apsolutna pogreška je vrlo mala (0.0628), što sugerira da model ima visoku točnost.

MSE je prosjek kvadrata grešaka, odnosno razlika između predviđanja i stvarnih vrijednosti. Ova metrika kažnjava veće greške više nego manje, zbog kvadriranja grešaka. Manji MSE ukazuje na bolju točnost modela. Ovdje je MSE vrlo nizak što ukazuje na dobar model.

RMSE je kvadratni korijen MSE-a i daje grešku u istim jedinicama kao i originalne podatke. U ovom slučaju, RMSE je nešto veći od MAE, što je očekivano jer RMSE više kažnjava velike greške. Vrijednost RMSE od 0.1315 i dalje ukazuje na visok nivo točnosti modela.

R^2 *score* pokazuje koliko dobro model objašnjava varijabilnost u podacima. Vrijednost R^2 kreće se od 0 do 1, pri čemu 1 znači savršeno objašnjenje podataka, a 0 znači da model ne objašnjava varijabilnost bolje od jednostavnog uzimanja srednje vrijednosti. R^2 rezultat od 0.9958 sugerira da model vrlo dobro objašnjava varijabilnost u podacima i da su predviđanja gotovo savršena.

5.4. Primjena dubokog učenja

Duboko učenje je podskup strojnog učenja koji se temelji na korištenju neuronskih mreža s velikim brojem slojeva i složenijih arhitektura. Ove duboke neuronske mreže (engl. DNN - *Deep Neural Networks*) karakterizira prisutnost više skrivenih slojeva između ulaznog i izlaznog sloja, što im omogućuje učenje složenih obrazaca u podacima. Zahvaljujući ovoj složenosti, DNN modeli mogu prepoznati i predstavljati podatke na višim apstraktnim razinama. U zadacima predviđanja kontinuiranih vrijednosti, poput globalne prodaje, često se primjenjuju potpuno povezane duboke mreže (engl. *fully connected*). Korištenjem većeg broja slojeva i neuronskih jedinica, ovi modeli postaju sposobniji za prepoznavanje i učenje kompleksnih uzoraka u podacima, što može značajno poboljšati njihovu učinkovitost i preciznost u predviđanju [3].

Prikazan tako dubljeg sloja prikazuje slika 5.30. odnosno dublje mreže:

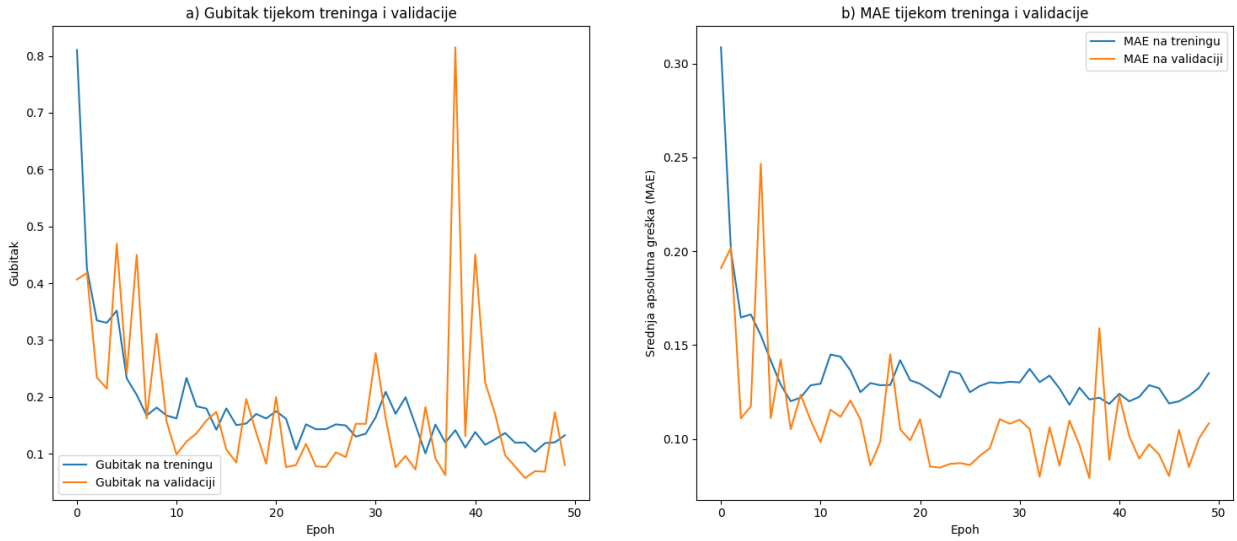
```
model = keras.Sequential([
    keras.layers.Dense(256, activation="relu",
input_shape=(X_train.shape[1],),
kernel_regularizer=keras.regularizers.l2(0.001)),
    keras.layers.Dropout(0.3),
    keras.layers.Dense(128, activation="relu",
kernel_regularizer=keras.regularizers.l2(0.001)),
    keras.layers.Dropout(0.3),
    keras.layers.Dense(64, activation="relu",
kernel_regularizer=keras.regularizers.l2(0.001)),
    keras.layers.Dense(32, activation="relu"),
    keras.layers.Dense(1)
])
```

Slika 5.30. Dublji sloj neuronske mreže.

Razlike u odnosnu na primjenu neuronske mreže:

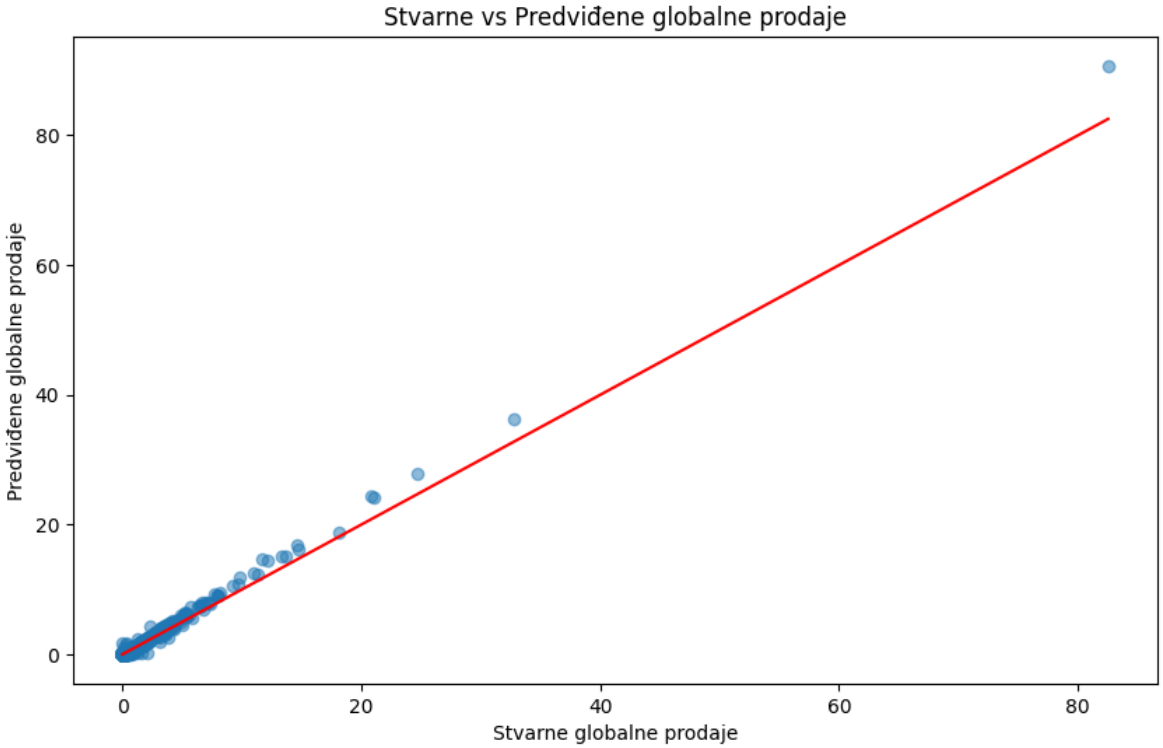
- Dodavanje više slojeva → Mreža sada ima 4 skrivena sloja s različitim brojem neurona, što povećava kapacitet za učenje složenih uzoraka u podacima.
- *Dropout* i L2 regularizacija → *Dropout* (0.3) nasumično isključuje 30% neurona tijekom treniranja kako bi se spriječila pretreniranost. L2 regularizacija kažnjava velika težinska rješenja, što dodatno pomaže u borbi protiv pretreniranosti.
- *Nadam* optimizator → Nadam je varijanta Adam optimizatora koji je prilagođen za duboke mreže i može poboljšati konvergenciju.
- Dulje treniranje (50 epoha) → Više epoha omogućava modelu da uči duže, no potrebno je pratiti rezultate kako bi se izbjeglo pretreniranost.

Slikama 5.31. i 5.32. su prikazani rezultati dobiveni primjenom dubokog učenja, kao i kod neuronskih mreža. Na slici 5.31., slika (a) prikazuje promjenu gubitka (engl. *loss*), dok slika (b) promjenu srednje apsolutne greške (MAE) tokom epoha, kako za skup podataka za treniranje, tako i za validacijski skup.



Slika 5.31. Gubitak i MAE.

Graf na slici 5.32. prikazuje odnos stvarnih i predviđenih vrijednosti, gdje crvena linija predstavlja savršenu liniju predviđanja.



Slika 5.32. Predviđanje.

Neuronske mreže i duboko učenje poznati su po svojoj visokoj točnosti u predviđanjima, posebno kada se primjenjuju na složene zadatke. Zahvaljujući sposobnosti da uče i prepoznaju obrasce u velikim količinama podataka, ovi algoritmi pružaju iznimno precizne rezultate te se u ovom slučaju to i pokazalo.

5.5. Usporedba algoritama

U prethodnim poglavljima, detaljnije se istraživala primjena različitih algoritama strojnog učenja na odabrane značajke iz baze podataka video igara. Svaki algoritam, uključujući K-*Nearest Neighbors* (KNN), *Random Forest Regressor* (RF), *Gradient Boosting Regressor* (GBR), i Neuronske Mreže (NN), bio je primijenjen pojedinačno kako bi se ocijenila njegova sposobnost predviđanja globalne prodaje video igara.

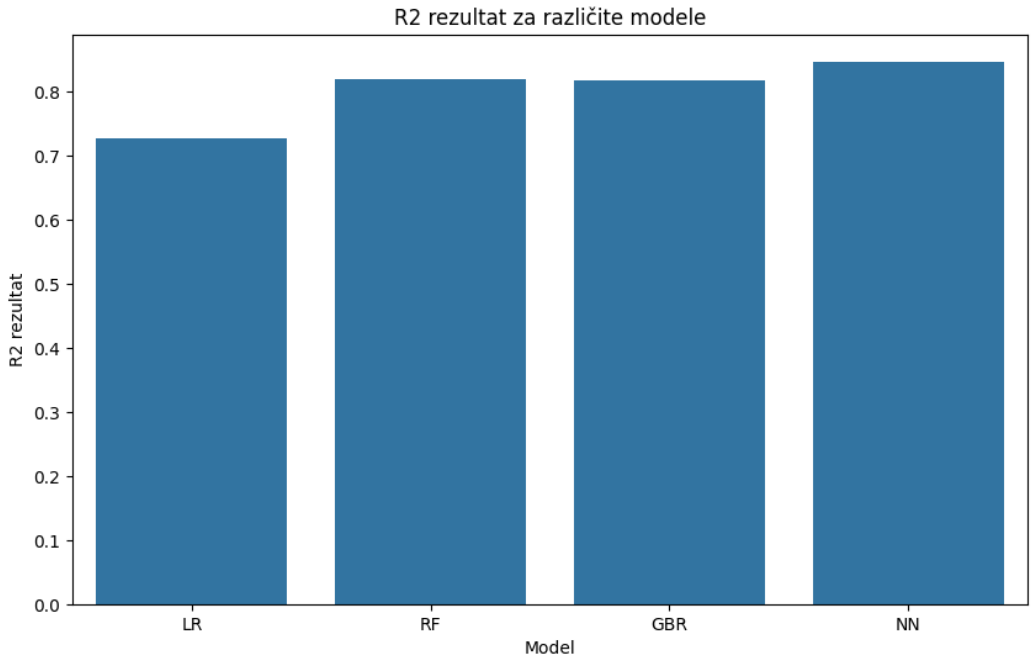
U ovom poglavlju, fokus će biti na usporedbu rezultata dobivenih primjenom različitih algoritama na iste značajke. Cilj je analizirati i usporediti performanse modela koristeći metrike poput R^2, *Mean Absolute Error* (MAE) , *Root Mean Squared Error* (RMSE), *Mean Squared Error* (MSE), F1 rezultat, točnost, preciznost i odziv. Ove metrike omogućavaju da se procijeni točnost i preciznost svakog modela te da se odredi koji algoritam pruža najpouzdanije rezultate za predviđanje globalne prodaje video igara. Na temelju dobivenih rezultata, bit će moguće identificirati prednosti i nedostatke svakog modela u kontekstu ovog zadatka te donijeti zaključke

o tome koji algoritam je najprikladniji za primjenu na sličnim problemima u budućnosti. Također, usporedba rezultata omogućit će bolje razumijevanje utjecaja različitih pristupa na učinkovitost modela i pružiti smjernice za buduća istraživanja i optimizaciju modela.

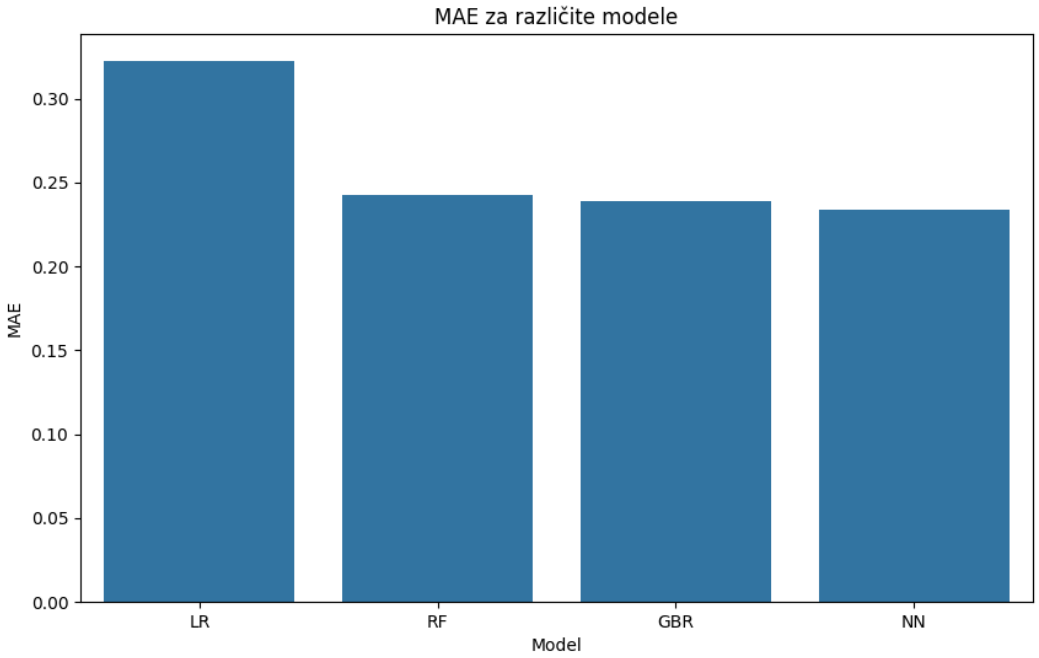
Algoritmi koji će se uspoređivati na istim značajkama su:

- linearna regresija (LR)
- slučajna šuma (RF)
- regresor gradijentnog pojačanja (GBR)
- neuronske mreže (NN)

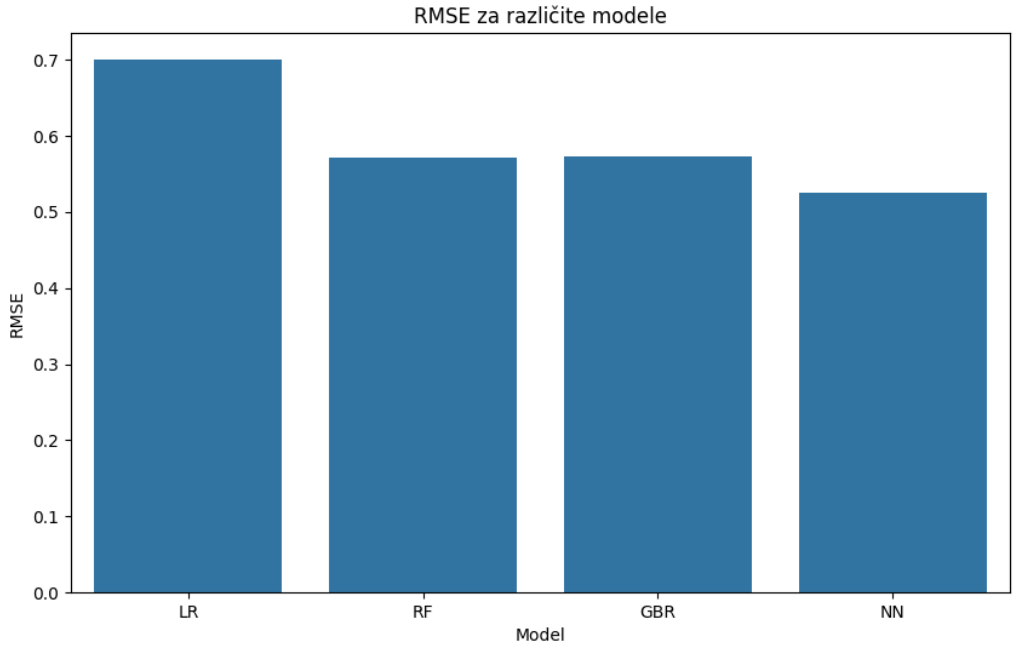
Algoritmi će se usporediti značajkama iz baze podataka, grupiranih po više značajki kao što se radilo u pojedinačnoj primjeni algoritama strojnog učenja, dok će ciljna varijabla uvijek biti **Global_Sales**. Značajke su odabrane na način koji uključuje numeričke podatke te kombinaciju numeričkih i kategorijskih podataka, kako bi se testirala učinkovitost različitih modela strojnog učenja u radu s raznovrsnim tipovima podataka. Odabir značajki nije vođen specifičnim kriterijima osim potrebe da se istraže različiti aspekti baze podataka. Cilj je bio vidjeti kako različiti algoritmi reagiraju na integer i string značajke te njihovu mješavinu te kako se nositi s različitim tipovima podataka u skupu. Ovakav pristup omogućava uvid u univerzalnost modela i njihovu sposobnost prilagodbe na raznolike informacije iz baze podataka. Za početak, značajke koje će se predviđati su: **JP_Sales**, **Other_Sales** i **Crtic_Score**. Sve značajke, uključujući i ove navedene, objašnjene su u prethodnim poglavljima. Slikama 5.33., 5.34., 5.35. i 5.36. prikazani su rezultati dobiveni nabrojanim algoritmima.



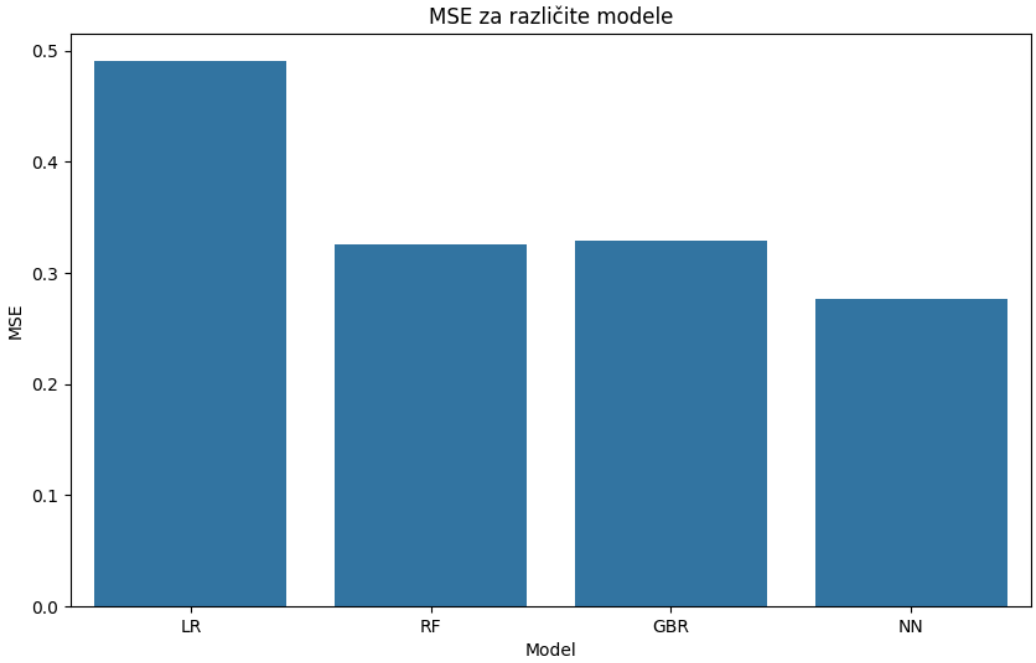
Slika 5.33. Usporedba R^2 rezultata.



Slika 5.34. Usporedba MAE.



Slika 5.35. Usporedba RMSE.



Slika 5.36. Usporedba MSE.

Tablica 5.2. R^2 rezultati.

Model	**R^2 rezultat**
LR	0.7272
RF	0.8189
GBR	0.8174

NN	0.8463

Slikom 5.33. graf prikazuje rezultate vezane za R^2 Score, što naslov grafa jasno i govori. Iz tablice 5.2. se može zaključiti kako neuronska mreža daje najbolji rezultat (0.8463). Ova analiza pokazuje da složeniji modeli poput neuronskih mreža Random Forest i Gradient Boosting imaju bolje performanse u odnosu na linearnu regresiju, što je očekivano zbog njihove sposobnosti da bolje modeliraju nelinearne odnose.

Tablica 5.3. MAE rezultati.

Model	**MAE**
LR	0.3224
RF	0.2429
GBR	0.2389
NN	0.2336

Neuronska mreža postiže najbolji rezultat s najnižom MAE vrijednošću, što ukazuje na najveću preciznost predikcija među svim testiranim modelima. Gradient Boosting Regressor i Random Forest također pokazuju solidne rezultate, s MAE vrijednostima koje su bolje od linearne regresije. Linearna regresija ima najvišu MAE vrijednost, što sugerira da jednostavan linearni model nije tako precizan kao složeniji modeli. Vrijednosti se mogu vidjeti u tablici 5.3.

Tablica 5.4. RMSE rezultati.

Model	**RMSE**
LR	0.7006
RF	0.5708
GBR	0.5732
NN	0.5260

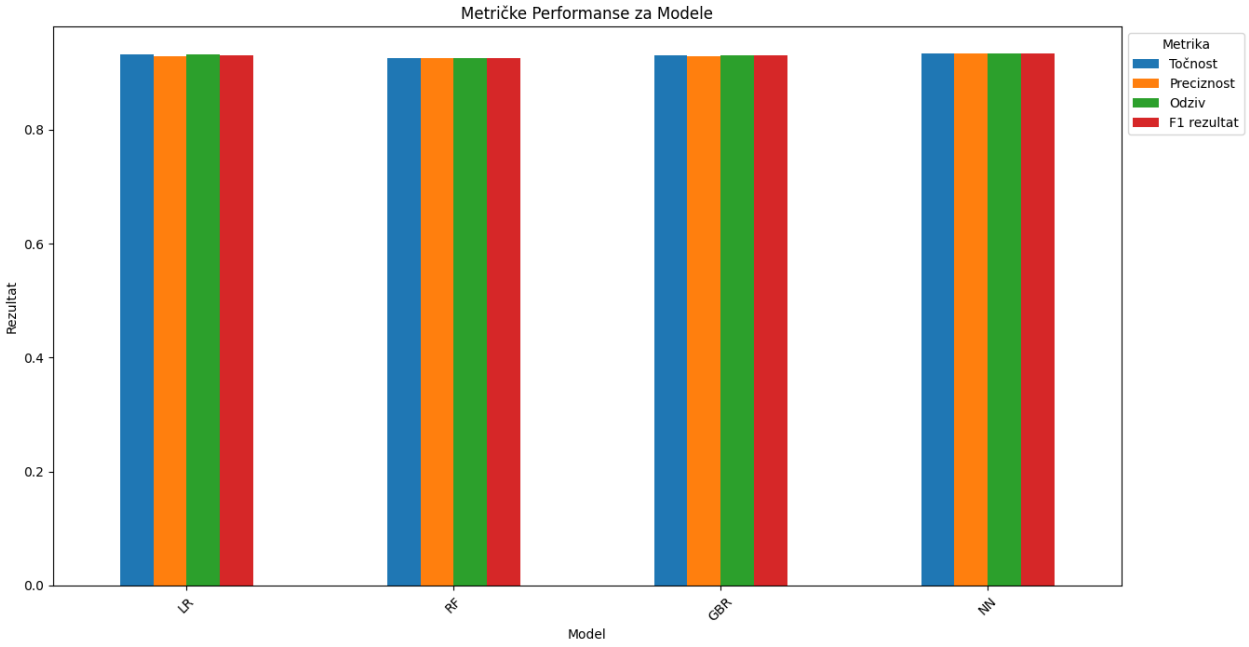
Iz tablice 5.4. se može zaključiti kako neuronska mreža ponovno postiže najbolji rezultat s najnižim RMSE-om, što ukazuje na najmanje prosječne pogreške u predikcijama među svim testiranim modelima. Random Forest i Gradient Boosting Regressor imaju slične RMSE vrijednosti koje su nešto više od NN-a, ali još uvijek bolje od linearne regresije.

Tablica 5.5. MSE rezultati.

Model	MSE
LR	0.4908
RF	0.3258
GBR	0.3286
NN	0.2766

Tablica 5.5. prikazuje rezultate MSE za različite modele. Neuronska mreža postavlja najbolji rezultat s najnižim MSE-om, RF i GBR su gotovo jednaki, dok linearna regresija odskače dosta od ostalih te imaju najveći MSE rezultat.

Metričke performanse za ova četiri modela i odabrane značajke prikazane su slikom 5.37., a njihove vrijednosti prikazane u tablici 5.6.



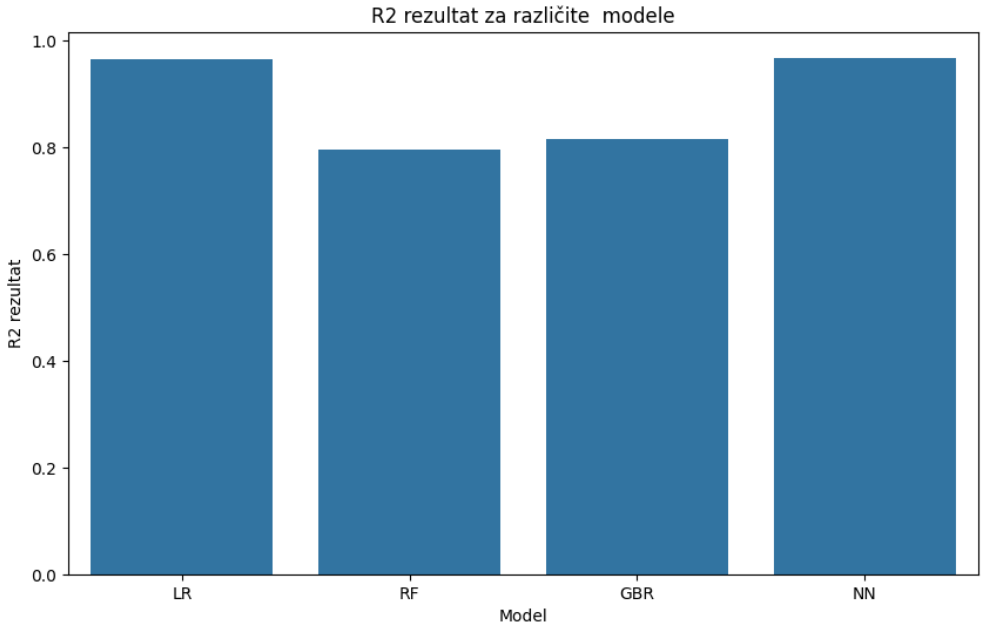
Slika 5.37. Metričke performanse za modele.

Tablica 5.6. Metričke performanse.

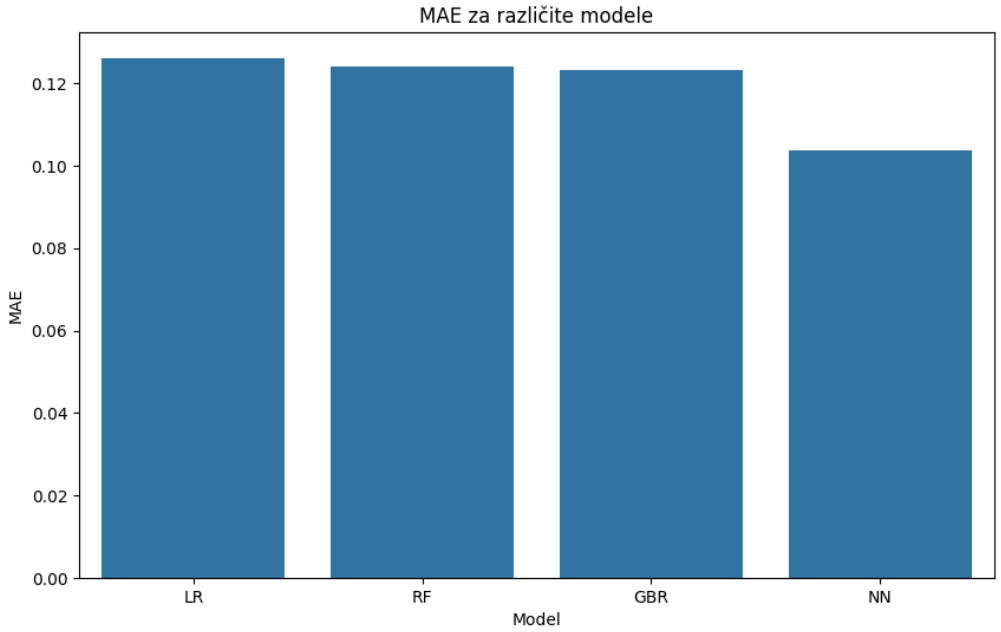
Model	Točnost	Preciznost	Odziv	F1 rezultat
LR	0.9318	0.9290	0.9318	0.9297
RF	0.9263	0.9263	0.9263	0.9261
GBR	0.9306	0.9292	0.93906	0.9298
NN	0.9330	0.9339	0.9330	0.9334

Neuronska mreža postavlja najbolji rezultat u svim metrima: točnosti, preciznosti, odzivu i F1 rezultatu. Svi ovi metrički pokazatelji su najbolji kod NN-a, što ukazuje na njegovu superiornost u usporedbi s ostalim modelima. Svi ostali modeli su također pokazali vrlo dobre rezultate.

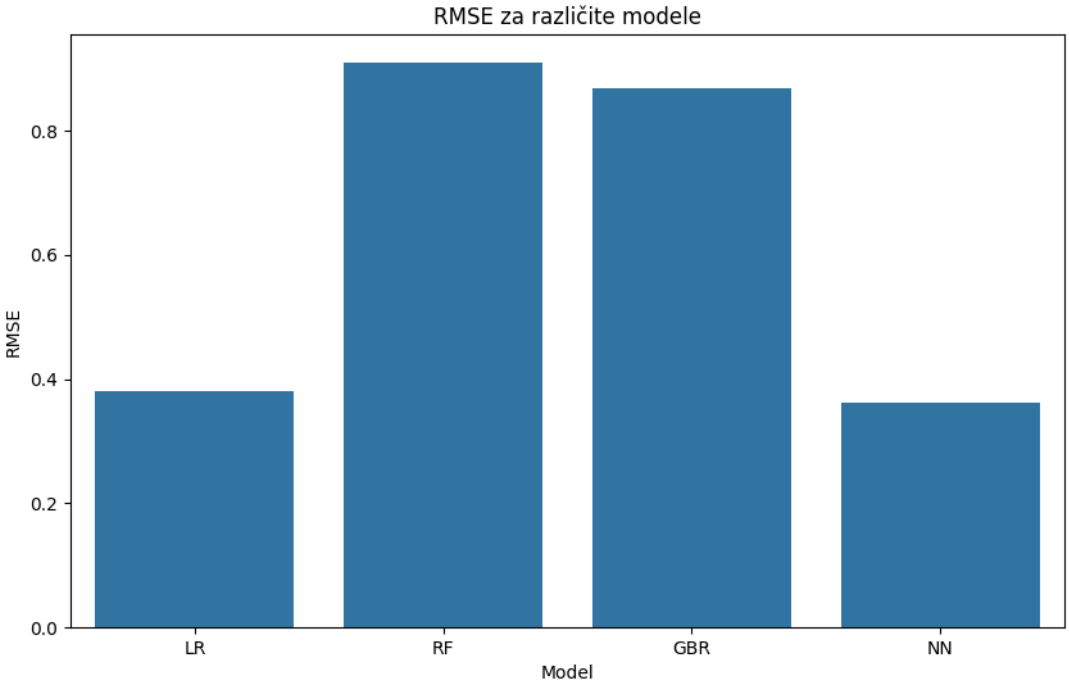
Sljedeće značajke koje će se koristiti za usporedbu su: **NA_Sales**, **EU_Sales**, **Year_of_Release**. Uspješnost algoritama će se ispitivati na jednak način kao i za prethodne značajke. Slikama 5.34., 5.35., 5.36. i 5.37. su prikazani rezultati R^2, MAE, RMSE i MSE.



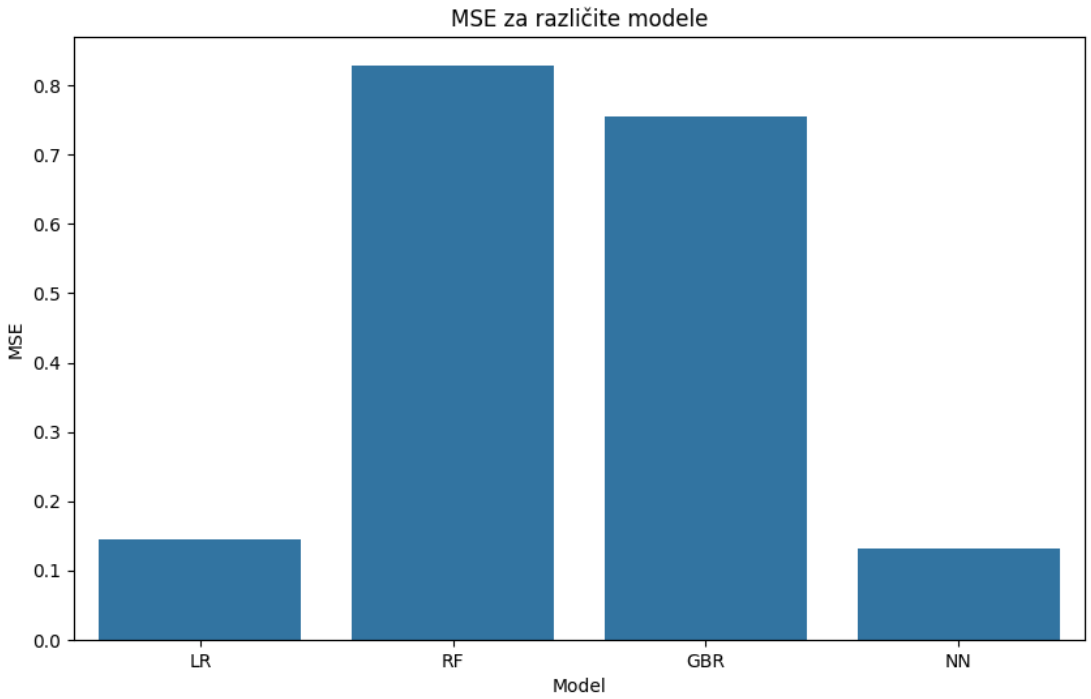
Slika 5.38. Usporedba R^2 rezultata.



Slika 5.39. Usporedba MAE rezultata.



Slika 5.40. Usporedba RMSE rezultata.



Slika 5.41. Usporedba MSE rezultata.

U tablicama 5.7., 5.8., 5.9. te 5.10. su prikazani rezultati vezani za R^2, MAE, RMSE i MSE i njihove modele.

Tablica 5.7. R^2 rezultati.

Model	$\boldsymbol{R^2}$
LR	0.9645
RF	0.7969
GBR	0.8150
NN	0.9677

Najbolji model prema R^2 rezultatu je neuronska mreža (0.9677). Neuronska mreža najbolje objašnjava varijaciju u podacima u usporedbi s ostalim modelima. Visoki R^2 rezultat sugerira da model daje vrlo precizna predviđanja i da su predviđene vrijednosti bliske stvarnim vrijednostima. Linearna regresija daje gotovo jednak rezultat kao i neuronska mreža.

Tablica 5.8. MAE rezultati.

Model	**MAE**
LR	0.1261
RF	0.1240

GBR	0.1232
NN	0.1037

Neuronska mreža pruža najbolju preciznost s najmanjim prosječnim apsolutnim pogreškama (MAE), odnosno njezina predviđanja su najbliža stvarnim vrijednostima. Ostali modeli poput Gradient Boosting Regressora i Random Foresta također pokazuju dobre rezultate, dok je linearna regresija, iako solidna, najmanje precizna među analiziranim modelima.

Tablica 5.9. RMSE rezultati.

Model	**RMSE**
LR	0.3808
RF	0.9102
GBR	0.8688
NN	0.3892

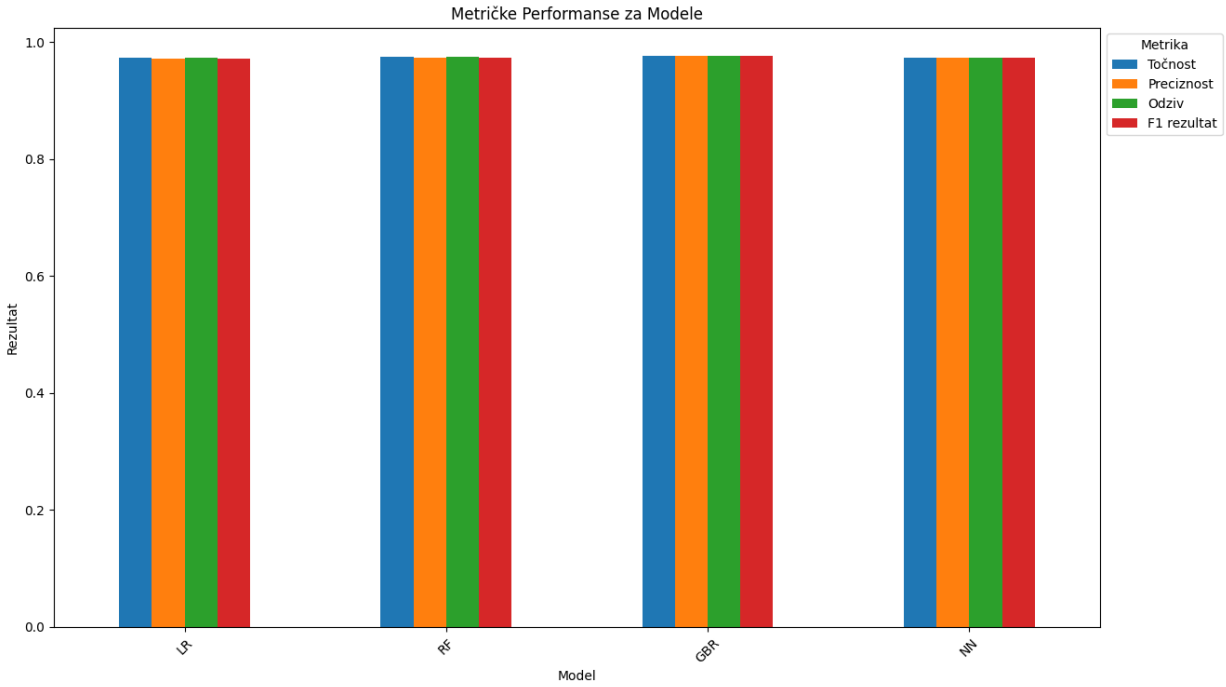
Linearna regresija pokazuje najbolji rezultat s najnižim RMSE, što znači da pruža najpreciznija predviđanja kada se uzmu u obzir kvadratne pogreške. Neuronska mreža je vrlo blizu, što također pokazuje odlične rezultate, ali s nešto većim RMSE. Gradient Boosting Regressor i Random Forest Regressor imaju veće RMSE, što sugerira da njihova predviđanja imaju veće pogreške u kvadratnom smislu.

Tablica 5.10. MSE rezultati.

Model	**MSE**
LR	0.1450
RF	0.8285
GBR	0.7548
NN	0.1316

Neuronska mreža pokazuje najbolji rezultat s najnižim MSE. Linearna regresija je također vrlo blizu i pokazuje dobre rezultate. Gradient Boosting Regressor i Random Forest Regressor imaju veće MSE, što ukazuje na to da njihova predviđanja imaju veće prosječne kvadratne pogreške. Kao i kod RMSE, važno je razmotriti sve metrike zajedno za cjelovitu evaluaciju modela.

Slikom 5.42. prikazan je graf koji pokazuje rezultate vezane za F1 rezultat, točnost, preciznost te odziv. Njihovi rezultati su dani tablicom 5.11.



Slika 5.42. Metričke performanse za modele.

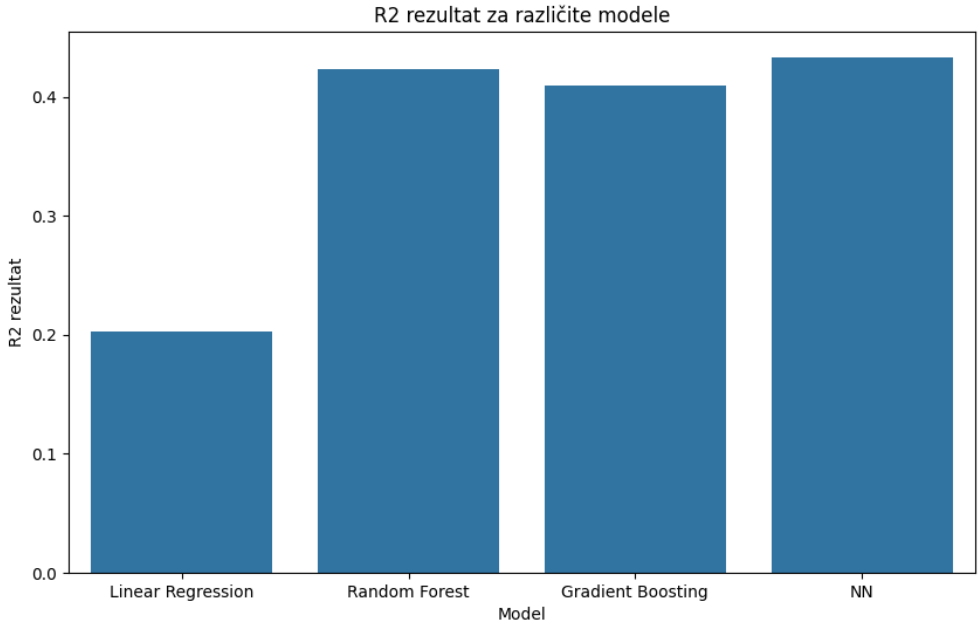
Tablica 5.11. Metričke performanse.

Model	Točnost	Preciznost	Odziv	F1 rezultat
LR	0.9733	0.9726	0.9733	0.9726
RF	0.9748	0.9743	0.9848	0.9744
GBR	0.9766	0.9761	0.9766	0.9763
NN	0.9778	0.9770	0.9775	0.9770

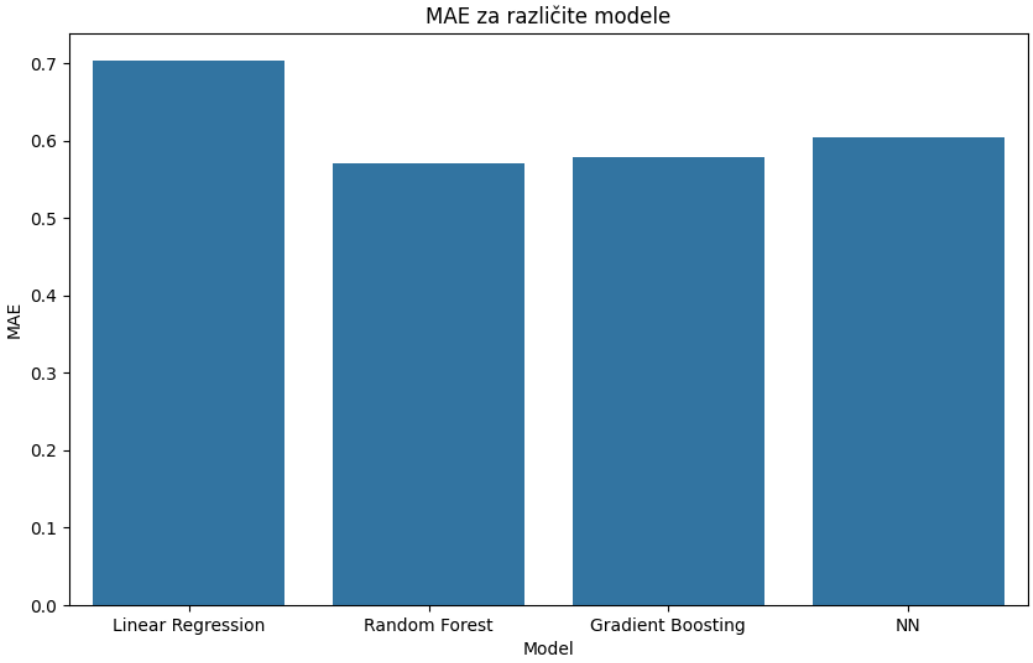
Iako svi modeli imaju dobre rezultate, neuronska mreža se pokazala kao najbolji model u svim metrikama, s najvišim vrijednostima za točnost, preciznost, odziv i F1 rezultat. Ovo ukazuje na njezinu iznimnu sposobnost da precizno klasificira podatke i upravlja s nebalansiranim ili kompleksnim skupovima podataka.

Za zadnje tri značajke koristit će se: **Publisher, Developer, Genre, Year_of_Release, User_Count**. U ovom slučaju, imamo pet značajki koje će se promatrati. To se radi jer kombinacija numeričkih i kategorijskih značajki omogućava bolje razumijevanje i predviđanje ciljne varijable. Korištenjem samo *string* značajki (Publisher, Developer, Genre) dobili bi se znatno lošiji rezultati te se zbog toga ubacuju značajke s numeričkim vrijednostima

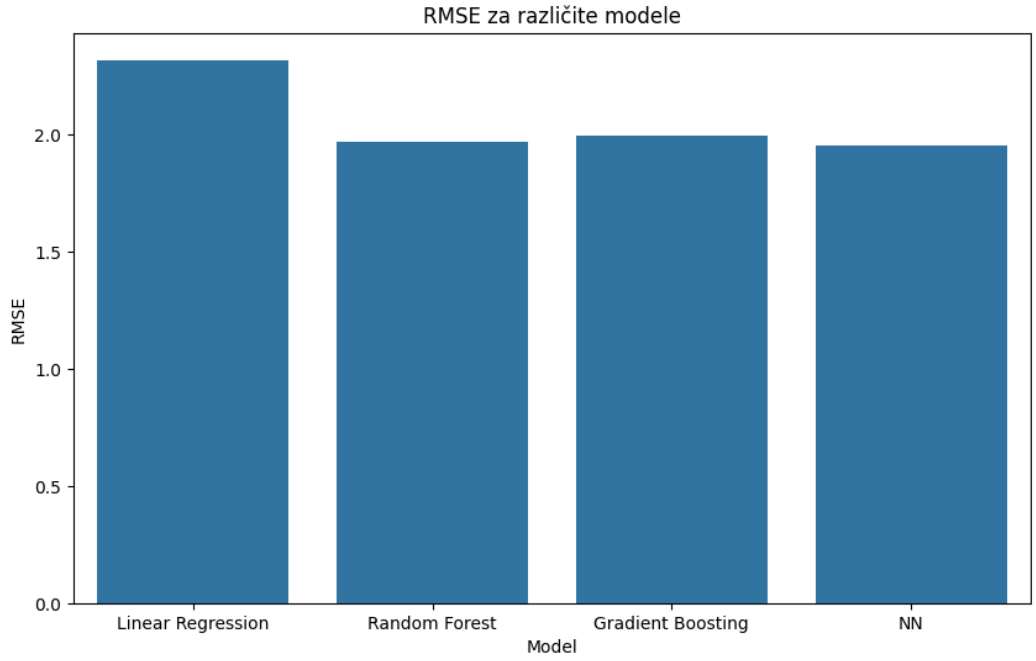
(Year_Of_Release, User_Count) kako bi model bio uspješniji. Slikama 5.43., 5.44., 5.45. i 5.46. su prikazani rezultati, a njihove vrijednosti tablicama 5.12., 5.13., 5.14. te 5.15.



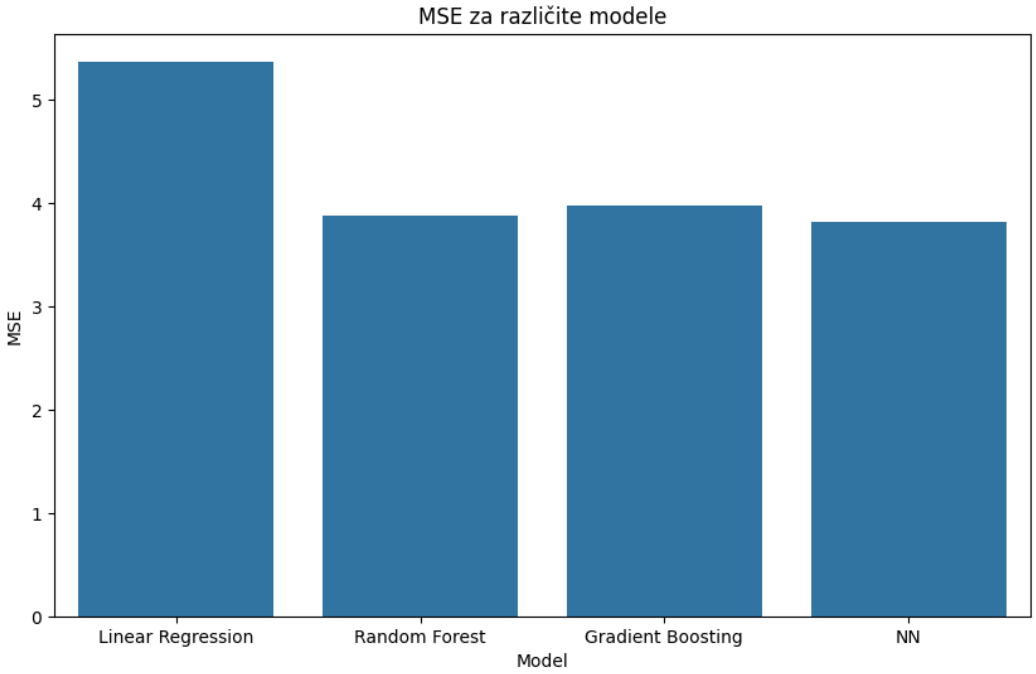
Slika 5.43. Usporedba R^2 rezultata.



Slika 5.44. Usporedba MAE.



Slika 5.45. Usporedba RMSE.



Slika 5.46. Usporedba MSE.

Tablica 5.12. R^2 rezultati.

Model	$\boldsymbol{R^2}$
LR	0.2027
RF	0.4238
GBR	0.4100

NN	0.4571

Neuronska mreža (NN) pokazuje najbolju sposobnost objašnjavanja varijance u ciljnim podacima, što ukazuje na njenu sposobnost da modelira kompleksne odnose između značajki i ciljne varijable. Random Forest (RF) i Gradient Boosting (GBR) su također vrlo učinkoviti, slični u rezultatima, ali nešto lošiji od neuronske mreže. Linearna regresija (LR) je najslabija među testiranim modelima.

Tablica 5.13. MAE rezultati.

Model	**MAE**
LR	0.7037
RF	0.5710
GBR	0.5782
NN	0.6283

Ovi rezultati pokazuju kako je Random Forest najpouzdaniji model s obzirom na minimiziranje prosječne apsolutne pogreške, dok neuronske mreže mogu zahtijevati dodatne prilagodbe za poboljšanje preciznosti predviđanja.

Tablica 5.14. RMSE rezultati.

Model	**RMSE**
LR	2.3166
RF	1.9693
GBR	1.9929
NN	1.9116

Neuronska mreža postiže najbolji rezultat s najnižim RMSE od 1.9116, što ukazuje na to da je najpreciznija u smislu minimiziranja kvadratnih pogrešaka predikcija. Nasuprot tome, linearna regresija ima najviši RMSE od 2.3166, što sugerira da su njegova predviđanja u prosjeku najdalja od stvarnih vrijednosti u usporedbi s ostalim modelima.

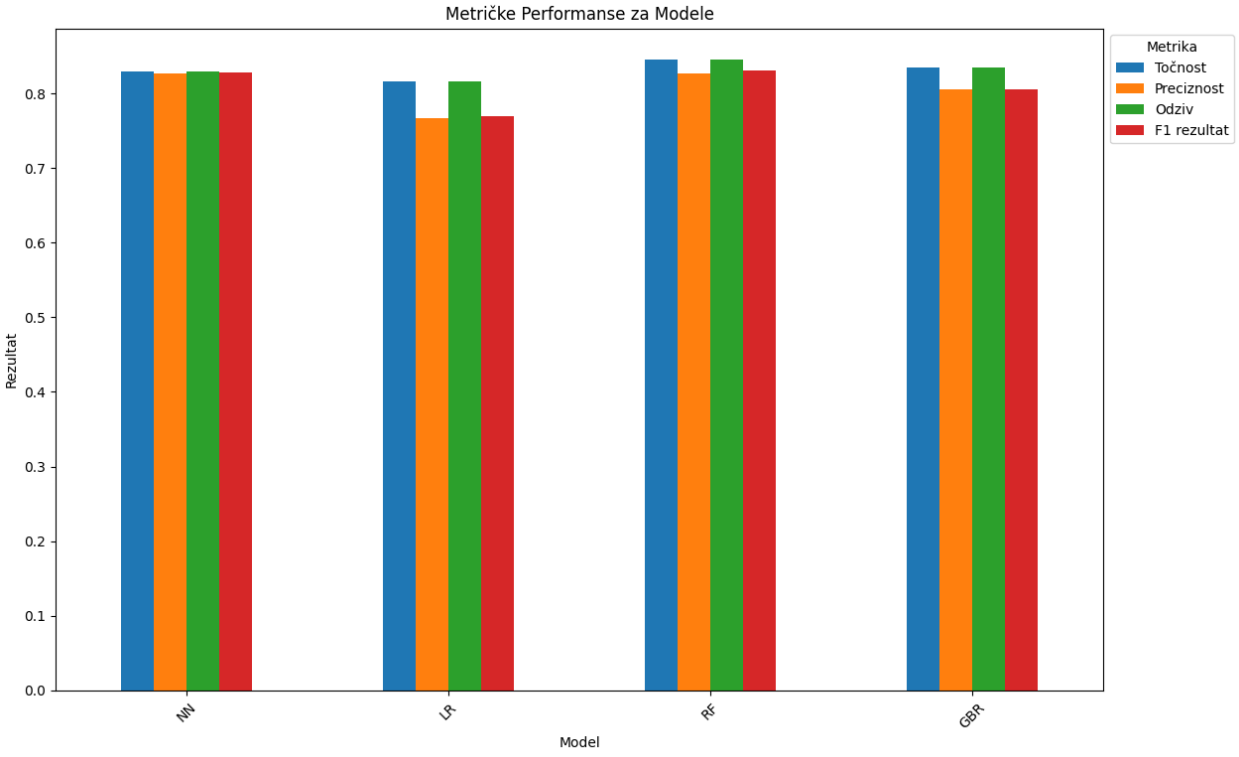
Tablica 5.15. MSE rezultati.

Model	**MSE**

LR	5.3666
RF	3.8783
GBR	3.9715
NN	3.6540

Neuronska mreža daje najbolje rezultate i za MSE, dok linearna regresija daje najlošije.

Kao i u prethodnim usporedbama, prikaz metričkih performansi za modele prikazan je slikom 5.47., a njegovi rezultati tablicom 5.16.



Slika 5.47. Metričke performanse za modele.

Tablica 5.16. Metričke performanse.

Model	Točnost	Preciznost	Odziv	F1 rezultat
NN	0.8168	0.8199	0.8168	0.8181
LR	0.8161	0.7668	0.8161	0.7691
RF	0.8459	0.8261	0.8450	0.8306
GBR	0.8342	0.8049	0.8342	0.8057

Random Forest (RF) je najbolji model za klasifikaciju, s najboljim rezultatima u većini metrika (točnost, preciznost, odziv i F1 rezultat). Gradient Boosting (GBR) i Neuronska Mreža (NN) su također vrlo dobri, s vrlo bliskim rezultatima. GBR je nešto bolji u točnosti i F1 rezultatu. Linearni Regresor (LR), iako koristan za regresiju, pokazuje slabije performanse u klasifikaciji u usporedbi s drugim modelima, s nižim rezultatima u preciznosti, odzivu i F1 rezultatu.

Neuronska mreža je postigla najbolje rezultate s najnižim MAE i RMSE vrijednostima te solidnim R^2 rezultatom. Njegova sposobnost prepoznavanja složenih obrazaca u podacima čini ga idealnim za skupove podataka s mnogim značajkama poput Developer i User_Count. Njegova prednost dolazi iz sposobnosti učenja složenih nelinearnih odnosa, što je ključno kada podaci sadrže mnogo varijabli koje međusobno utječu na izlaz. Model linearne regresije je pokazao loše performanse, s visokim MAE i RMSE vrijednostima, posebno kada se koristi s jednostavnijim značajkama. Razlog za ovo može biti pretpostavka linearnosti u modelu, koja ne može dovoljno dobro opisati složenost podataka o prodaji video igara, gdje su odnosi između ulaznih značajki i ciljne varijable često nelinearni. Iako bolji od linearne regresije, Random Forest također je imao ograničenja u predikciji kada su se koristile jednostavnije značajke. Prednost ovog modela je u njegovoj robusnosti i otpornosti na pretreniranje, ali mu je nedostajala preciznost u hvatanju složenih interakcija među značajkama kada su one bile ograničene. Gradient Boosting Regressor model je postigao bolje rezultate od Random Foresta u smislu RMSE, što ukazuje na njegovu sposobnost da bolje iskoristi kombinaciju značajki. Njegova sposobnost da uči i prilagođava model omogućuje bolju preciznost u predikcijama, ali je zahtijevao duže vrijeme treniranja i više računalnih resursa.

Neuronske mreže su bile najuspješnije zbog svoje sposobnosti da modeliraju složene nelinearne odnose između ulaznih značajki i ciljne varijable. Veličina skupa podataka i složenost značajki, kao što su Developer i User_Count, omogućili su ovom modelu da nauči obrasce koje drugi modeli nisu mogli prepoznati. Linearni modeli poput linearne regresije nisu bili u stanju dobro modelirati podatke zbog svoje jednostavne strukture koja ne može uhvatiti složene odnose. Ovo je bilo posebno vidljivo kada su korištene samo jednostavne značajke poput Year_of_Release i Publisher. Složeni algoritmi poput Gradient Boostinga i Random Foresta pokazali su se boljima u slučajevima kada su bili dostupni bogatiji skupovi značajki. Njihova sposobnost da rade s različitim kombinacijama značajki i prilagođavaju se nelinearnim odnosima omogućila im je bolje performanse u predviđanjima.

Algoritmi poput neuronske mreže najbolje odgovaraju skupu podataka s mnogim značajkama koje pružaju bogate informacije o proizvodima. U situacijama gdje su značajke ograničene, modeli

poput Random Foresta i Gradient Boostinga mogu pružiti solidne rezultate, ali će neuroni biti dominantni kada su dostupni bogati podaci koji omogućuju modeliranje složenih obrazaca.

6. ZAKLJUČAK

U ovom radu istražena je primjena i usporedba različitih tehnika strojnog učenja u svrhu predviđanja globalne prodaje video igara. Kroz analizu postojećih rješenja i implementaciju vlastitih modela, dokazano je da strojno učenje može značajno doprinijeti razumijevanju i predviđanju tržišnih trendova u industriji video igara. Primjena algoritama kao što su linearna regresija, random forest, neuronske mreže i drugi, pokazala je da su ovi modeli sposobni uhvatiti složene obrasce u podacima i pružiti precizna predviđanja.

Rezultati istraživanja pokazuju da duboke neuronske mreže, iako zahtjevnije za implementaciju i treniranje, pružaju najtočnija predviđanja, osobito kada se radi o kompleksnim skupovima podataka s velikim brojem značajki. Ovi modeli uspješno minimiziraju pogreške u predviđanjima i pokazuju visok R^2 rezultat, što ih čini izuzetno korisnim za detaljnu analizu i predviđanje. S druge strane, jednostavniji modeli poput linearne regresije i random foresta pokazali su solidne rezultate, ali s nešto većim pogreškama u predikcijama i nešto lošijim R^2 rezultatom. Linearna regresija, zbog svoje inherentne jednostavnosti, suočila se s najvećim pogreškama, dok je random forest postigao bolje rezultate, ali nije dostigao preciznost neuronskih mreža.

Ovaj rad naglašava da primjena strojnog učenja može značajno unaprijediti procese donošenja odluka u industriji video igara. Predloženi modeli pružaju čvrstu osnovu za daljnji razvoj alata za analizu tržišta i predviđanje prodaje, što može pomoći izdavačima i developerima u boljem razumijevanju tržišnih kretanja i optimizaciji njihovih poslovnih strategija. Nadalje, korištenje složenijih modela poput neuronskih mreža pokazuje potencijal za značajno unapređenje točnosti predikcija i dublje razumijevanje tržišta. Istovremeno, jednostavniji modeli mogu pružiti brza i učinkovitija rješenja za manje zahtjevne analize, čineći ih korisnim alatom u situacijama gdje su resursi ili vrijeme ograničeni.

Budući radovi mogli bi se fokusirati na optimizaciju odabira značajki i daljnju usporedbu modela kako bi se dodatno unaprijedila točnost i učinkovitost predikcija. Također, proširenje skupa podataka ili uključivanje novih varijabli moglo bi pružiti dublje uvide i poboljšati performanse modela, čime bi se još više povećala njihova korisnost u industriji video igara.

LITERATURA

[1] K.P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[2] S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 2010.

[3] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*. MIT Press, 2016.

[4] Y. LeCun, Y. Bengio, G. Hinton, *"Deep learning." Nature*, 521(7553), 436-444., 2015.

[5] J. P. Mueller, L. Massaron, *Machine learning for dummies*, Wiley, 31. svibnja 2016.

[6] B. D. Bašić, J. Šnajder, Strojno učenje: Uvod u strojno učenje, Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, 2016., dostupno na: https://www.fer.unizg.hr/_download/repository/SU-2016-01-Uvod.pdf

[7] A. Nag, *Pragmatic machine learning with Python*, Manish Jain for BPB Publications, New Delhi, 2020.

[8] Supervised and Unsupervised learning, 2024., dostupno na: https://www.geeksforgeeks.org/supervised-unsupervised-learning/

[9] V.Zhou, *Machine Learning for Beginners: An Introduction to Neural Networks*, Towards Data Science, 5. ožujka 2019., dostupno na: https://towardsdatascience.com/machine-learning-for-beginners-an-introduction-to-neural-networks-d49f22d238f9/

[10] https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/

[11] K. Gurney, *An Introduction to Neural Networks*, CRC Press, 5. kolovoza 1997.

[12] W.Di, A. Bhardwaj, J.Wei, *Deep learning essentials*, Packt Publishing, Birmingham, UK, 2018.

[13] *Introduction to Deep Learning*, 2023., dostupno na: https://www.geeksforgeeks.org/introduction-deep-learning/

[14] N. Bolf, Strojno učenje, Sveučilište u Zagrebu, Fakultet kemijskog inženjerstva i tehnologije, dostupno na: https://hrcak.srce.hr/file/382926

[15] S.Ray, *Top 10 Machine Learning Algorithms to Use in 2024*, Analytics Vidhya, 2024., dostupno na: https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

[16] *Support Vector Machine (SVM) Algorithm*, 2023., dostupno na: https://www.geeksforgeeks.org/support-vector-machine-algorithm/

[17] G. Bhumireddy, 2022., *Comparison of machine learning algorithms on detecting the confusion of students while watching MOOCs*. Faculty of Computing, Blekinge Institute of Technology, Dostupno na: https://www.diva-portal.org/smash/get/diva2:1641701/FULLTEXT02.pdf

[18] J. Marcoux, S.A. Selouani, *A Hybrid Subspace-Connectionist Data Mining Approach for Sales Forecasting in the Video Game Industry, 2009 World Congress on Computer Science and Information Engineering*, Canada

[19] A. Yufa, J. L. Yu, H. Chan, P. D. Berger, *Predicting Global Video-Game Sales, Quest Journals, Journal of Research in Business and Management*, 2019.

[20] J. Li, Y. Zheng, H. Hu, J. Lu, C. Zhan, *Predicting Video Game Sales Based on Machine Learning and Hybrid Feature Selection Method*, 2021 *IEEE International Conference on Intelligent Systems and Knowledge Engineering*, Guangzhou, China

[21] V.Sarala, D. Akhila, *Video Game Sales Analysis, Journal of Engineering Sciences*, n.d.

[22] W. Huang, *Research on the Prediction on the Sales of Electronic Games*, 2023., Computer Sciences, Wuhan University, Wuhan, 430072, China

[23] Z. Zhou, *Automatic Machine Learning-Based Data Analysis for Video Game Industry*, 2022., IEEE, Nanjing, China

[24] A. Manimuthu, U. Udhayakumar, A. Cathrine, T.D.Gowri, J. Peter, S. Selvam, & S. Roseline, *Data Interpretation and Video Games Sales Prediction Using Machine Learning Algorithms - a Comparative Study*. 2023. Electronic Journal

[25] K. Saraswathi, N.T. Renukadevi, & S. Nandhinidevi, *Sales Prediction on Video Games Using Machine Learning Approaches*, 2021., Conference Proceedings

[26] JetBrains. (n.d.). PyCharm: *Python IDE for Professional Developers*. Dostupno na: https://www.jetbrains.com/pycharm/

[27] Numpy Documentation, dostupno na: https://numpy.org/doc/stable/, pristupljeno: 1. kolovoza 2024.

[28] Pandas Documentation, dostupno na: https://pandas.pydata.org/docs/, pristupljeno: 1. kolovoza 2024.

[29] Scikit-learn Documentation, dostupno na: https://scikit-learn.org/stable/getting_started.html/ pristupljeno: 1. kolovoza 2024.

[30] Scikit-learn documentation. (n.d.). *Gradient Boosting for regression*. Dostupno na: https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting

POPIS SKRAĆENICA

ML – Strojno učenje (engl. *Machine Learning*)

DL – Duboko učenje (engl. *Deep Learning*)

EU – Europa (engl. Europe)

NA – Sjeverna Amerika (engl. *North America*)

LR – Linearna regresija (engl. *Linear Regression*)

RF – Slučajna šuma (engl. *Random Forest*)

JP – Japan (engl. *Japan*)

KNN – k Najbližih Susjeda (engl. *k – Nearest Neighbour*)

GBR - Regresor gradijentnog pojačanja (engl. *Gradient Boosting Regressor*)

R^2 – Koeficijent determinacije

MAE – Srednja apsolutna pogreška (engl. *Mean Absolute Error*)

RMSE – Korijen srednje kvadratne pogreške (engl. *Root Mean Squared Error*)

MSE – Srednja kvadratna pogreška (engl. *Mean Squared Error*)

SAŽETAK

U ovom diplomskom radu primijenjeni su razni algoritmi strojnog učenja, kao što su linearna regresija, slučajne šume (RF), neuronske mreže i drugi. Cilj je bio međusobno ih usporediti te prikazati rezultate dobivene primjenom svakog od njih. Projekt je izrađen u programskom okruženju PyCharm, a implementacija algoritama provedena je koristeći programski jezik Python. Uz Python, korištene su i razne biblioteke poput NumPy, Pandas, Scikit-Learn, TensorFlow i drugih. Rad također obuhvaća pregled drugih istraživanja koja se bave sličnom tematikom. Pomoću modela razvijenih ovim algoritmima predviđale su se globalne prodaje video igara te su ti rezultati uspoređeni sa stvarnim vrijednostima.

Ključne riječi: Globalna prodaja, neuronske mreže, predviđanje, strojno učenje

ABSTRACT

Application of machine learning in predicting global sales of video games

In this thesis, various machine learning algorithms, such as linear regression, random forests (RF), neural networks, and others, were applied. The goal was to compare these algorithms with one another and present the results obtained from each. The project was developed in the PyCharm environment, with the implementation of algorithms carried out using the Python programming language. In addition to Python, various libraries such as NumPy, Pandas, Scikit-Learn, TensorFlow, and others were used. The thesis also includes a review of other studies addressing similar topics. The models developed using these algorithms were employed to forecast global video game sales, and the predicted results were compared with actual values.

Keywords: Global sales, neural networks, prediction, machine learning