Tehnike za razložljivost v strojnem učenju: LIME, SHAP in Grad-CAM

Dušan Todorović 89232094
Mentor: dr.Branko Kavšek
Fakulteta za matematiko, naravoslovje in informacijske tehnologije
Raziskovalni seminar 2026

Abstract

Razvoj sodobnih modelov strojnega učenja je omogočil visoke rezultate pri nalogah klasifikacije in napovedovanja, hkrati pa je odprl pomembno vprašanje njihove interpretabilnosti. Kompleksni modeli, kot so ansambelske metode in globoke nevronske mreže, se pogosto obravnavajo kot "črne škatle", saj uporabniku ni jasno, na podlagi katerih vhodnih informacij sprejemajo odločitve. V tem delu so analizirane tri pomembne metode za razložljivost modelov: LIME, SHAP in Grad-CAM. V teoretičnem delu so predstavljeni osnovni pojmi interpretabilnosti, matematična intuicija izbranih metod ter njihove prednosti in omejitve. V praktičnem delu sta izvedena dva eksperimenta. Prvi eksperiment uporablja podatkovni nabor Titanic in model Random Forest, pri čemer sta metodi LIME in SHAP uporabljeni za interpretacijo tabelaričnih podatkov. Drugi eksperiment uporablja podatkovni nabor Fashion-MNIST in konvolucijsko nevronsko mrežo, pri čemer sta primerjani razlagi SHAP in Grad-CAM na slikah. Rezultati kažejo, da LIME ponuja intuitivne lokalne razlage, SHAP omogoča stabilnejšo lokalno in globalno interpretacijo, Grad-CAM pa vizualni vpogled v regije slike, ki najbolj vplivajo na odločitev CNN modela.

Ključne besede: interpretabilnost, razložljivost, Explainable AI, LIME, SHAP, Grad-CAM, Random Forest, CNN, Titanic, Fashion-MNIST

1 Uvod

Razvoj metod strojnega učenja in umetne inteligence je omogočil gradnjo modelov, ki dosegajo zelo dobre rezultate pri nalogah, kot so klasifikacija, regresija in analiza kompleksnih podatkovnih množic. Vendar pa se z večanjem kompleksnosti modelov, posebej pri globokih nevronskih mrežah in ansambelskih metodah, povečuje tudi problem njihove interpretabilnosti. Takšni modeli se pogosto obravnavajo kot "črne škatle", saj ni jasno, kako pridejo do svojih odločitev.

V sodobnih uporabah, zlasti na področjih, kot so medicina, finance in avtonomni sistemi, ni dovolj, da model daje samo točne rezultate. Pomembno je razumeti tudi razloge, ki stojijo za njegovimi odločitvami. Ta problem je dodatno poudarjen v kontekstu etike in regulative, kjer obstaja potreba po transparentnosti in odgovornosti algoritmičnih sistemov [1, 9].

Zato se razvija področje razložljive umetne inteligence (angl. *Explainable Artificial Intelligence* - XAI), katerega cilj je omogočiti boljše razumevanje modelov strojnega učenja. Tehnike za razložljivost omogočajo analizo obnašanja modela, identifikacijo potencialnih napak in pristranskosti ter povečujejo zaupanje uporabnikov v takšne sisteme [11, 10].

V tem delu je poudarek na treh pomembnih metodah za interpretacijo modelov: LIME (*Local Interpretable Model-agnostic Explanations*), SHAP (*SHapley Additive exPlanations*) in Grad-CAM (*Gradient-weighted Class Activation Mapping*). LIME in SHAP se pogosto uporabljata za razlago napovedi pri tabelaričnih podatkih, medtem ko je Grad-CAM posebej pomemben za interpretacijo konvolucijskih nevronskih mrež pri nalogah obdelave slik.

Cilj tega dela je predstaviti teoretične osnove izbranih metod, izvesti njihovo primerjavo ter z dvema praktičnima eksperimentoma prikazati njihovo uporabo. Prvi eksperiment uporablja podatkovni nabor Titanic in metodi LIME ter SHAP, drugi eksperiment pa uporablja podatkovni nabor Fashion-MNIST in primerja SHAP ter Grad-CAM v kontekstu vizualne interpretacije CNN modela.

2 Motivacija in pomen interpretabilnosti

Z vedno širšo uporabo modelov strojnega učenja v realnih sistemih postaja vprašanje transparentnosti in zanesljivosti njihovih odločitev ključno. Čeprav sodobni modeli pogosto dosegajo visoko točnost, njihova kompleksnost otežuje razumevanje načina, na katerega pridejo do napovedi. V mnogih domenah, posebej tistih, ki neposredno vplivajo na ljudi, ima lahko pomanjkanje interpretabilnosti resne posledice [1].

Na področju medicine se modeli strojnega učenja uporabljajo za diagnostiko bolezni, ocenjevanje tveganja in podporo pri odločitvah o terapiji. V takšnih primerih zdravnik ne more zaupati samo rezultatu modela, temveč mora razumeti tudi razloge za odločitev.

Podobno se v finančnem sektorju modeli uporabljajo za ocenjevanje kreditne sposobnosti in odločanje o odobritvi posojil. Če model ni transparenten, obstaja tveganje za nepravične ali diskriminatorne odločitve.

Eden izmed ključnih izzivov pri uporabi strojnega učenja je problem pristranskosti (angl. *bias*). Modeli se učijo iz zgodovinskih podatkov, ki pogosto vsebujejo implicitne ali eksplicitne pristranskosti. Posledično lahko model reproducira ali celo okrepi obstoječe neenakosti. Brez ustreznih metod za interpretacijo lahko takšni problemi ostanejo skriti.

Interpretabilnost prispeva tudi k večjemu zaupanju uporabnikov v sisteme, ki temeljijo na umetni inteligenci. Kadar uporabniki razumejo, kako model deluje in zakaj sprejema določene odločitve, je večja verjetnost, da bodo takšne sisteme sprejeli in uporabljali. Nasprotno pa lahko popolnoma nerazložljivi sistemi povzročijo nezaupanje in odpor.

3 Osnovni pojmi interpretabilnosti

V strojnem učenju se pogosto uporabljata pojma interpretabilnost in razložljivost (angl. *explainability*). V literaturi se včasih uporabljata kot sopomenki, vendar med njima obstaja določena razlika. Interpretabilnost se nanaša na sposobnost modela, da je neposredno razumljiv človeku, razložljivost pa na uporabo dodatnih metod za pojasnjevanje obnašanja modela, posebej kadar je model sam po sebi kompleksen [1].

Modeli, ki so inherentno interpretabilni, kot so linearna regresija ali odločitvena drevesa, omogočajo neposredno razumevanje odnosa med vhodnimi podatki in izhodom. Nasprotno pa kompleksni modeli, kot so globoke nevronske mreže in ansambelske metode, na primer Random Forest, zahtevajo dodatne tehnike za razlago njihovega obnašanja. Te tehnike imenujemo *post-hoc* metode, saj se uporabijo po učenju modela.

Ena izmed osnovnih delitev metod za interpretacijo je delitev na globalno in lokalno interpretacijo. Globalna interpretacija poskuša razložiti celotno obnašanje modela, torej odgovoriti na vprašanje, kateri atributi so na splošno najpomembnejši in kako model sprejema odločitve na ravni celotnega podatkovnega nabora. Lokalna interpretacija se osredotoča na posamezne instance in pojasnjuje, zakaj je model za konkreten primer sprejel določeno napoved.

Metode za interpretacijo lahko delimo tudi na model-specifične in model-agnostične. Model-specifične metode uporabljajo notranjo strukturo modela in so omejene na določene vrste modelov. Primer takšne metode je Grad-CAM, ki se uporablja za interpretacijo konvolucijskih nevronskih mrež. Model-agnostične metode, kot sta LIME in SHAP, se lahko uporabijo na različnih modelih, saj model obravnavajo kot "črno škatlo" in analizirajo odnos med vhodom in izhodom.

Pomemben koncept je tudi zvestoba razlage (angl. *fidelity*), ki označuje, kako dobro razlaga predstavlja dejansko obnašanje izvirnega modela. Preprosta razlaga ni nujno tudi zvesta razlaga. Zato je pomembno uporabljati metode, ki dosegajo dobro ravnotežje med

razumljivostjo in točnostjo razlage.

4 Pregled metod za razložljivost

Za boljše razumevanje obnašanja modelov strojnega učenja je bilo razvitih veliko tehnik za razložljivost. Med seboj se razlikujejo po pristopu, kompleksnosti in področju uporabe. Klasificiramo jih lahko glede na raven interpretacije, odvisnost od modela in tip podatkov, na katere se uporabljajo.

Eden najpreprostejših pristopov k interpretaciji modela je analiza pomembnosti atributov (angl. *feature importance*), ki prikazuje, kolikšen vpliv imajo posamezne spremenljivke na odločitve modela. Ta metoda omogoča globalen vpogled v obnašanje modela, vendar pogosto ne poda dovolj informacij o tem, kako se atributi kombinirajo pri posameznih napovedih.

Poleg tega se uporabljajo tudi metode, kot so *Partial Dependence Plots* (PDP), ki prikazujejo odnos med eno ali več vhodnimi spremenljivkami in napovedjo modela. Te metode omogočajo vizualno razumevanje vpliva atributov, vendar so lahko omejene, kadar med spremenljivkami obstaja močna odvisnost.

V novejši literaturi posebno pozornost dobivajo post-hoc metode, ki omogočajo podrobnejšo interpretacijo kompleksnih modelov. Med najpomembnejšimi sta LIME in SHAP. LIME približa obnašanje modela v lokalni okolici z enostavnejšim modelom, medtem ko SHAP uporablja Shapleyjeve vrednosti iz teorije iger za kvantifikacijo prispevka vsakega atributa k napovedi [2, 3].

V nasprotju s tema metodama je Grad-CAM metoda, specifična za konvolucijske nevronske mreže. Omogoča vizualizacijo delov slike, ki najbolj prispevajo k določeni klasifikaciji, s čimer dobimo intuitiven vpogled v delovanje CNN modela [4].

5 LIME

LIME (*Local Interpretable Model-agnostic Explanations*) je ena najbolj znanih metod za lokalno interpretacijo modelov strojnega učenja. Predlagali so jo Ribeiro, Singh in Guestrin z namenom razlage posameznih napovedi kompleksnih modelov [2].

Osnovna ideja metode LIME je, da lahko obnašanje kompleksnega modela v lokalni okolici opazovane instance približamo z enostavnejšim, interpretabilnim modelom. Čeprav je globalno obnašanje modela lahko zelo kompleksno, ga je v majhni okolici posamezne instance pogosto mogoče dovolj dobro aproksimirati z linearnim modelom.

LIME obravnava izvirni model kot črno škatlo. Za izbrano instanco generira perturbacije, nato za vsako perturbacijo pridobi napoved izvirnega modela, vzorcem pa dodeli uteži glede na njihovo oddaljenost od izvirne instance. Nato se nauči enostaven interpretabilen model, najpogosteje linearni model, katerega koeficienti predstavljajo lokalno

razlago.

Matematično lahko LIME predstavimo kot optimizacijski problem:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g), \tag{1}$$

kjer je f izvirni model, g interpretabilni model, π_x funkcija bližine okoli instance x, L funkcija izgube, ki meri, kako dobro g lokalno aproksimira f, $\Omega(g)$ pa kazen za kompleksnost interpretabilnega modela.

Prednost metode LIME je njena intuitivnost in fleksibilnost, saj se lahko uporabi na različnih tipih modelov. Glavna slabost je nestabilnost: razlage so lahko odvisne od načina generiranja perturbacij, števila vzorcev in širine jedra.

6 SHAP

SHAP (*SHapley Additive exPlanations*) je ena najpomembnejših metod za interpretacijo modelov strojnega učenja, zasnovana na teoriji iger. Razvila sta jo Lundberg in Lee z namenom zagotoviti konsistenten in teoretično utemeljen način razlage napovedi modelov [3].

Osnovna ideja metode SHAP temelji na konceptu Shapleyjevih vrednosti iz kooperativne teorije iger. V tem kontekstu se vsak atribut obravnava kot igralec, ki prispeva h končni napovedi modela. Cilj je pošteno porazdeliti prispevek vsakega atributa.

Shapleyjeva vrednost za atribut i je definirana kot:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}) - f(S) \right], \tag{2}$$

kjer je F množica vseh atributov, S podmnožica atributov brez atributa i, $f(S)$ pa vrednost modela, kadar se uporabi samo množica atributov S.

SHAP razlage lahko predstavimo kot aditivni model:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i, \tag{3}$$

kjer je ϕ_0 pričakovana vrednost modela, ϕ_i pa prispevek atributa i.

V praksi obstaja več različic metode SHAP. TreeSHAP se uporablja za modele, zasnovane na drevesih, KernelSHAP je model-agnostična različica, za nevronske mreže pa se uporabljajo različice, kot sta DeepSHAP ali GradientExplainer. V tem delu je TreeSHAP uporabljen v prvem eksperimentu z modelom Random Forest, medtem ko je SHAP GradientExplainer uporabljen v drugem eksperimentu s CNN modelom.

7 Grad-CAM

Grad-CAM (*Gradient-weighted Class Activation Mapping*) je metoda za interpretacijo globokih nevronskih mrež, posebej konvolucijskih nevronskih mrež, ki omogoča vizualizacijo delov vhodne slike, ki najbolj vplivajo na odločitev modela [4].

Za ciljni razred c se uteži map značilk izračunajo kot povprečni gradient izhoda modela za ta razred glede na aktivacije zadnjega konvolucijskega sloja:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{4}$$

kjer je A^k k-ta mapa značilk, y^c izhod modela za razred c, Z pa število elementov v mapi.

Nato se Grad-CAM mapa izračuna kot:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right). \tag{5}$$

Rezultat je toplotna mapa, ki prikazuje regije slike, ki so najbolj pozitivno prispevale k odločitvi modela za opazovani razred. Toplejše barve običajno označujejo večji pomen regije, hladnejše barve pa manjši pomen.

8 Primerjava metod

Metode LIME, SHAP in Grad-CAM imajo isti splošni cilj - razložiti odločitve modela - vendar se pomembno razlikujejo po teoretični osnovi, tipu podatkov in načinu interpretacije. LIME in SHAP se lahko uporabljata kot model-agnostični metodi, medtem ko je Grad-CAM specifičen za konvolucijske nevronske mreže.

Tabela 1: Primerjava metod za razložljivost

Značilnost	**LIME**	**SHAP**	**Grad-CAM**
Tip metode	Model-agnostična	Model-agnostična / model-specifična različica	Model-specifična
Raven interpretacije	Lokalna	Lokalna in globalna	Lokalna, vizualna
Teoretična osnova	Lokalna aproksimacija	Shapleyjeve vrednosti	Gradienti v CNN modelu
Tip podatkov	Tabelarični podatki, besedilo, slike	Tabelarični podatki, besedilo, slike	Slike
Rezultat	Seznam atributov in prispevkov	Numerični prispevki in grafi	Toplotna mapa
Glavna prednost	Intuitivnost	Konsistentnost	Vizualna jasnost
Glavna omejitev	Nestabilnost	Večja računska zahtevnost	Omejenost na CNN

Posebej pomembna je primerjava med metodama SHAP in Grad-CAM pri delu s slikami. Grad-CAM odgovarja na vprašanje *kje* model usmerja pozornost, SHAP pa na vprašanje *koliko* posamezni deli vhoda prispevajo k napovedi. Zato ti metodi nista konkurenčni, temveč komplementarni.

9 Eksperiment 1: Titanic, LIME in SHAP

9.1 Podatkovni nabor in cilj eksperimenta

V prvem eksperimentu je bil uporabljen podatkovni nabor Titanic, dostopen prek platforme OpenML in naložen s funkcijo `fetch_openml('titanic', version=1)` [7]. Cilj naloge je napoved binarne ciljne spremenljivke `survived`, ki označuje, ali je potnik nesrečo preživel.

Pri pripravi podatkov so bili odstranjeni stolpci `boat`, `body`, `name` in `ticket`. Razlog za odstranitev teh stolpcev je preprečevanje pojava *data leakage*. Posebej sta stolpca `boat` in `body` neposredno povezana z izidom preživetja, zato bi njuna vključitev povzročila nerealno visoke rezultate modela.

Podatki so bili razdeljeni na učno in testno množico v razmerju 80:20, pri čemer je

bila uporabljena stratifikacija glede na ciljno spremenljivko. Numerične spremenljivke so bile obdelane z imputacijo mediane in standardizacijo, kategorične spremenljivke pa z imputacijo najpogostejše vrednosti in one-hot kodiranjem. Kot klasifikacijski model je bil uporabljen `RandomForestClassifier` s 500 drevesi in uravnoteženjem razredov.

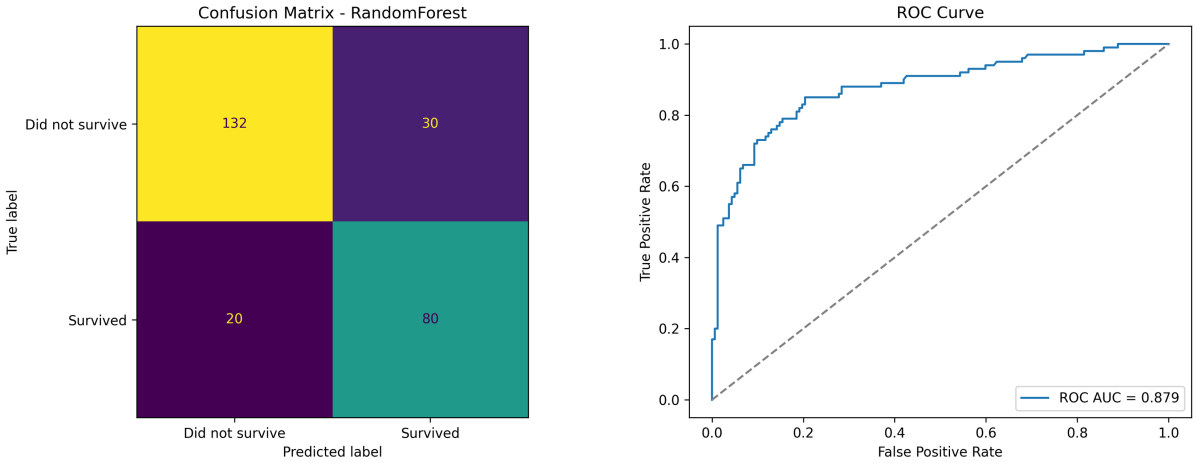
9.2 Rezultati klasifikacije

Model je na testni množici dosegel rezultate, prikazane v Tabeli 2.

Tabela 2: Rezultati modela Random Forest na podatkovnem naboru Titanic

Metrika	Vrednost
Accuracy	0.809
Precision	0.727
Recall	0.800
F1	0.762
ROC AUC	0.879
Average Precision	0.849

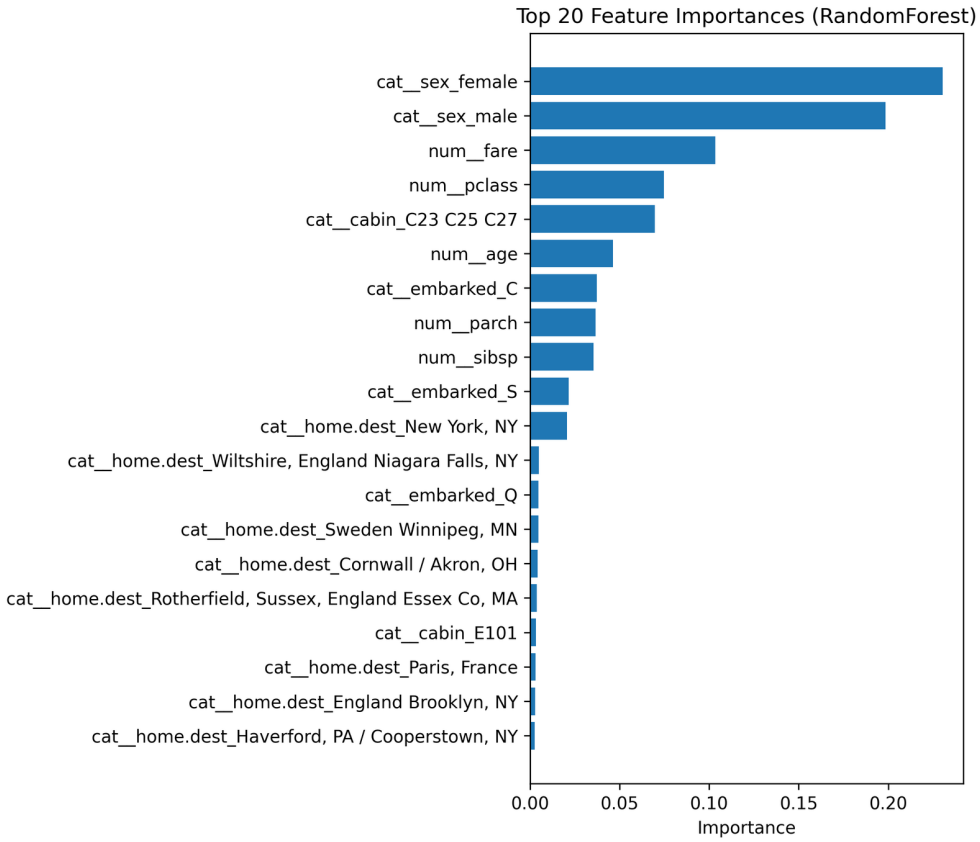
Dobljeni rezultati kažejo, da model uspešno identificira potnike, ki so preživeli, ob dobrem ravnotežju med natančnostjo in priklicem. Posebej pomembna je vrednost ROC AUC 0.879, ki kaže na dobro sposobnost modela za razlikovanje med razredoma.



(a) Matrika zmede **(b)** ROC krivulja

Slika 1: Evalvacija modela na podatkovnem naboru Titanic

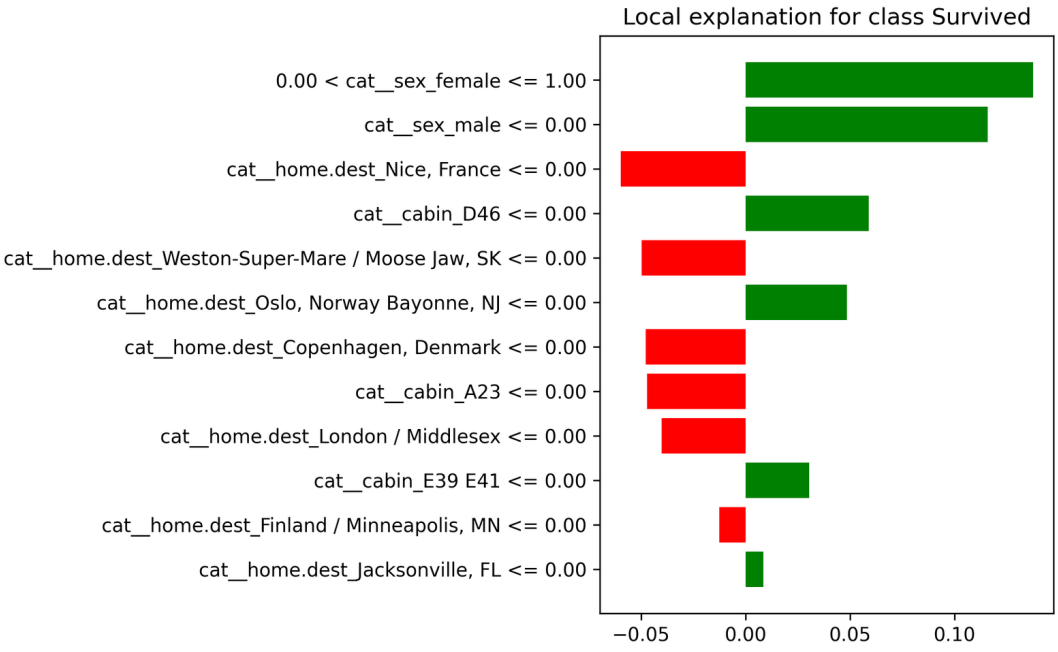
Analiza pomembnosti atributov kaže, da so med najpomembnejšimi značilnostmi spol, potovalni razred, cena vozovnice in starost. Ti rezultati so skladni z zgodovinskimi podatki o katastrofi Titanica, kjer so imele ženske in potniki višjih razredov večjo verjetnost preživetja.



Slika 2: Najpomembnejši atributi po modelu Random Forest

9.3 LIME interpretacija

Metoda LIME je bila uporabljena za lokalno razlago ene instance iz testne množice. Opazovana instanca predstavlja konkretnega potnika, razlaga LIME pa prikazuje, kateri atributi najbolj povečujejo ali zmanjšujejo verjetnost napovedanega razreda.

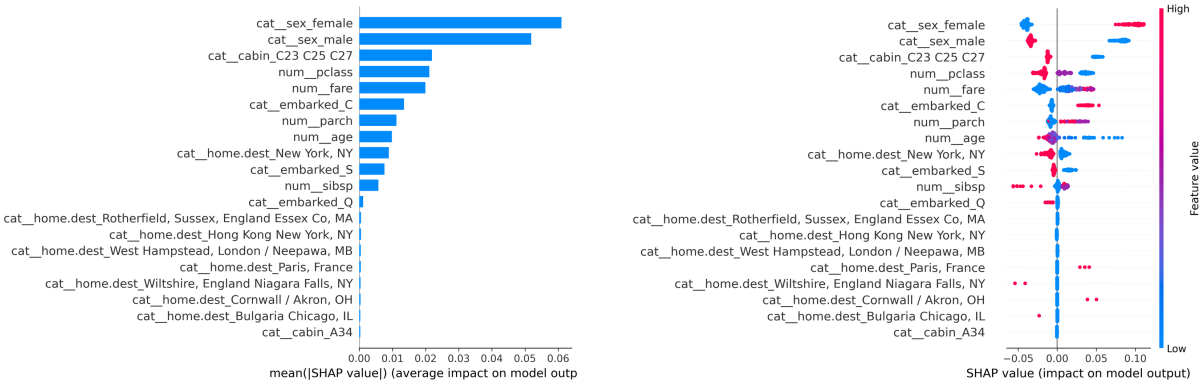


Slika 3: LIME lokalna razlaga za eno instanco iz nabora Titanic

Na Sliki 3 je razvidno, da LIME predstavi razlago kot lokalno linearno kombinacijo atributov. Pozitivni prispevki povečujejo verjetnost napovedanega razreda, negativni prispevki pa jo zmanjšujejo. Prednost takšnega prikaza je intuitivnost, saj se pri posamezni odločitvi jasno vidi, kateri atributi so imeli največji lokalni vpliv.

9.4 SHAP interpretacija

SHAP je bil uporabljen za globalno in lokalno interpretacijo modela. Globalni SHAP prikazi omogočajo vpogled v najpomembnejše atribute na ravni celotne testne množice, lokalni prikazi pa pojasnjujejo posamezno napoved.

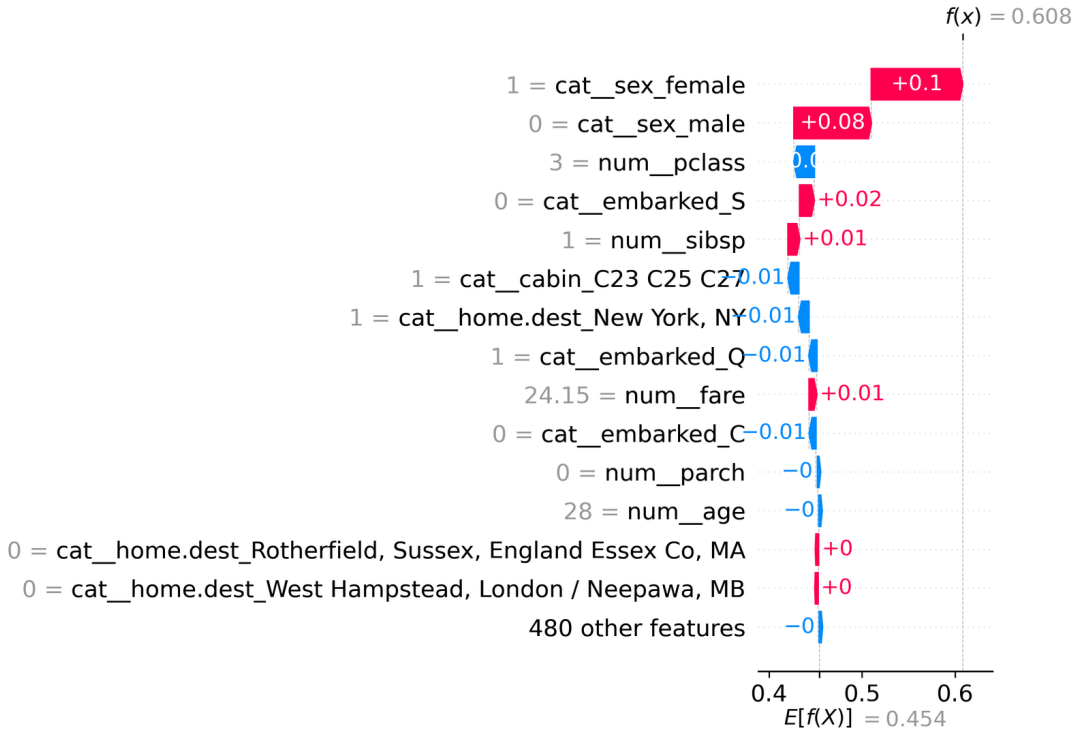


(a) SHAP bar prikaz **(b)** SHAP beeswarm prikaz

Slika 4: Globalna SHAP interpretacija modela na podatkih Titanic

Globalni SHAP prikazi na Sliki 4 potrjujejo, da so spol, cena vozovnice, potovalni

razred in starost med najpomembnejšimi atributi. Beeswarm prikaz dodatno kaže smer vpliva: določene vrednosti atributov premikajo napoved proti razredu preživetja, druge pa imajo nasproten učinek.



Slika 5: Lokalna SHAP razlaga za eno instanco

Waterfall prikaz na Sliki 5 kaže, kako se končna napoved oblikuje kot vsota prispevkov posameznih atributov, začenši z osnovno vrednostjo modela. S tem SHAP omogoča podrobnejšo in teoretično utemeljenejšo razlago kot metoda LIME.

9.5 Eksperiment z uhajanjem podatkov

Za prikaz pomena pravilne priprave podatkov je bil izveden tudi eksperiment, v katerem so bili vključeni stolpci, povezani z izidom. V tem primeru je model dosegel bistveno višje rezultate: accuracy 0.943, F1 0.925 in ROC AUC 0.983. Čeprav so ti rezultati na prvi pogled impresivni, metodološko niso pravilni, saj model uporablja informacije, ki v realnem scenariju napovedovanja ne bi bile na voljo. Zato se za veljavno interpretacijo uporablja izključno eksperiment brez atributov, ki povzročajo uhajanje podatkov.

10 Eksperiment 2: Fashion-MNIST, CNN, SHAP in Grad-CAM

10.1 Cilj eksperimenta

Drugi eksperiment je bil dodan, da bi interpretabilnost analizirali tudi v domeni slik. Medtem ko prvi eksperiment temelji na tabelaričnih podatkih ter metodah LIME in SHAP, drugi eksperiment uporablja konvolucijsko nevronsko mrežo in podatkovni nabor Fashion-MNIST. Cilj je primerjati metodi Grad-CAM in SHAP na istih slikah ter analizirati razliko med vizualno interpretacijo in kvantitativnim prispevkom pikslov.

10.2 Podatkovni nabor Fashion-MNIST

Fashion-MNIST je podatkovni nabor, ki vsebuje slike oblačil in modnih predmetov v 10 razredih, kot so majica, hlače, športni čevelj, torba in gležnjar. Vsaka slika je sivinska in velikosti 28×28 pikslov [5]. Ta nabor je bil izbran, ker je vizualno jasnejši in bolj interpretabilen od zelo majhnih slikovnih naborov, hkrati pa je dovolj preprost za učenje manjšega CNN modela.

V eksperimentu je bil uporabljen CNN model, implementiran v knjižnici PyTorch [6]. Model je sestavljen iz treh konvolucijskih slojev, max-pooling operacij, dropout regularizacije in popolnoma povezanih slojev za klasifikacijo v 10 razredov. Grad-CAM je bil uporabljen na konvolucijskem sloju modela, SHAP pa z gradientno zasnovano razlago.

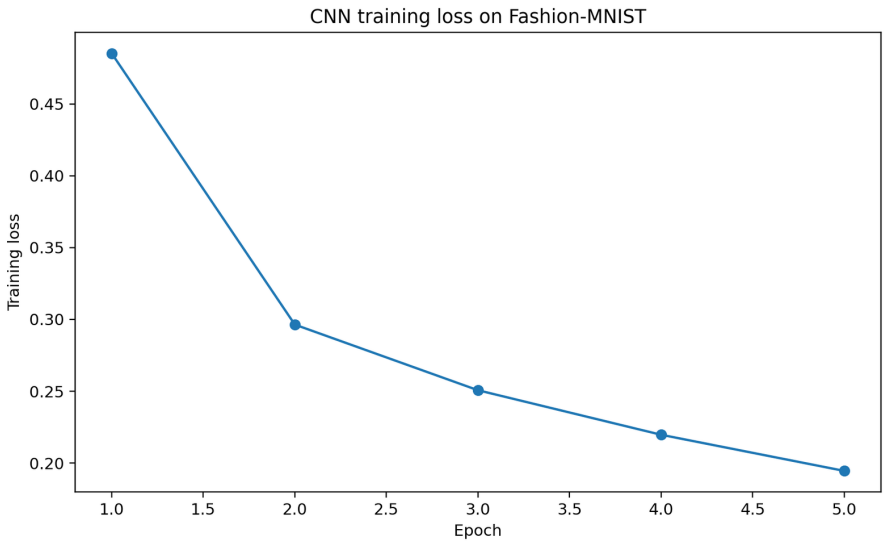
10.3 Rezultati klasifikacije

CNN model je na testni množici Fashion-MNIST dosegel rezultate, prikazane v Tabeli 3.

Tabela 3: Rezultati CNN modela na podatkovnem naboru Fashion-MNIST

Metrika	**Vrednost**
Accuracy	0.914
Macro precision	0.914
Macro recall	0.914
Macro F1	0.913

Ti rezultati kažejo, da model dosega stabilne rezultate in je dovolj zanesljiv za analizo razlag. Krivulja izgube med učenjem je prikazana na Sliki 6.



Slika 6: Krivulja izgube med učenjem CNN modela na naboru Fashion-MNIST

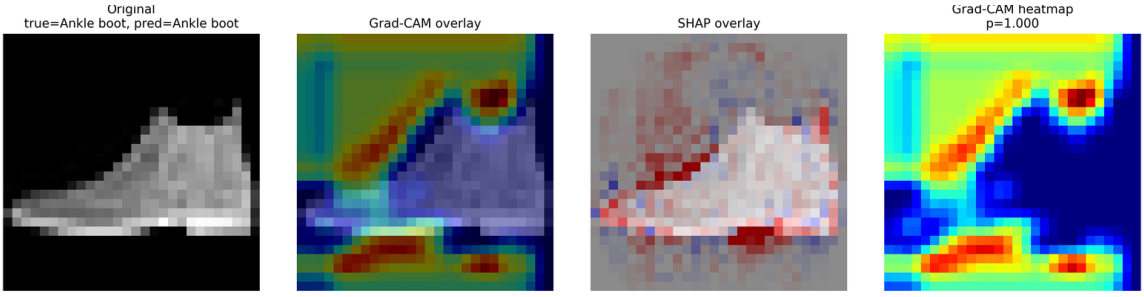
10.4 Razlaga barv pri prikazih Grad-CAM in SHAP

Pri metodi Grad-CAM toplejše barve, posebej rdeča in rumena, označujejo regije slike, ki jih model uporablja kot najpomembnejše za odločitev o napovedanem razredu. Hladnejše barve označujejo regije manjšega pomena. Grad-CAM ne prikazuje negativnih prispevkov, temveč predvsem regije, ki pozitivno podpirajo ciljni razred.

Pri metodi SHAP imajo barve drugačen pomen. Rdeča barva označuje dele slike, ki pozitivno prispevajo k napovedi opazovanega razreda, modra barva pa dele, ki zmanjšujejo verjetnost tega razreda. Zato SHAP omogoča podrobnejšo, vendar vizualno kompleksnejšo interpretacijo.

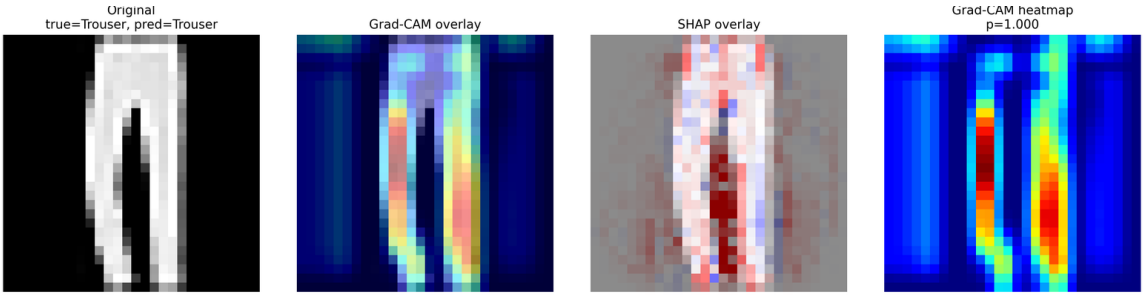
10.5 Vizualni rezultati

Na slikah 7, 8 in 9 so prikazani trije reprezentativni primeri: *Ankle boot*, *Trouser* in *T-shirt/top*. Za vsak primer so prikazani izvirna slika, Grad-CAM overlay, SHAP overlay in Grad-CAM toplotna mapa.



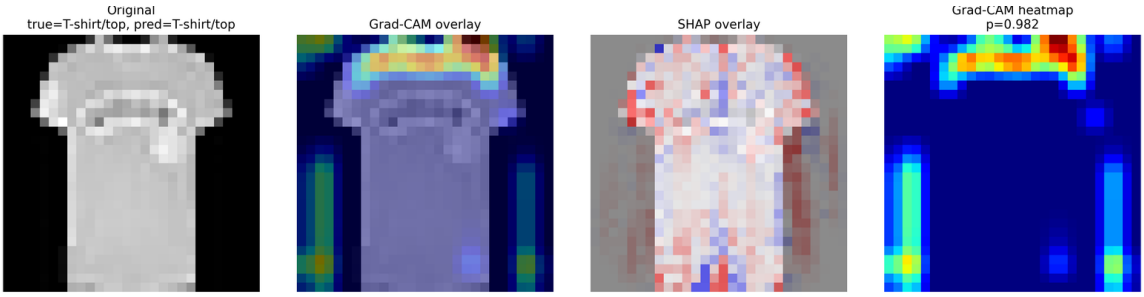
Slika 7: Primerjava razlag Grad-CAM in SHAP za razred Ankle boot

Pri razredu *Ankle boot* model pravilni razred napove z zelo visoko verjetnostjo. Grad-CAM izpostavi obliko obutve in dele slike, ki ustrezajo strukturi gležnjarja. SHAP dodatno pokaže, katere regije pozitivno prispevajo k napovedi in katere imajo negativen ali šibek prispevek.



Slika 8: Primerjava razlag Grad-CAM in SHAP za razred Trouser

Pri razredu *Trouser* Grad-CAM zelo jasno sledi navpični strukturi hlačnic, kar kaže, da model uporablja relevantne vizualne značilnosti. SHAP mapa ponuja podrobnejši vpogled v regije, ki podpirajo napoved razreda hlač.



Slika 9: Primerjava razlag Grad-CAM in SHAP za razred T-shirt/top

Pri razredu *T-shirt/top* Grad-CAM izpostavi zgornji del majice, vključno z območjem vratu in ramen, kar je smiselno, saj so ti deli značilni za ta razred. SHAP prikaz je podrobnejši in kaže, da nekateri deli slike pozitivno, drugi pa negativno vplivajo na končno napoved.

10.6 Primerjava SHAP in Grad-CAM v drugem eksperimentu

Rezultati eksperimenta Fashion-MNIST potrjujejo, da Grad-CAM in SHAP podajata različne, vendar komplementarne vrste razlag. Grad-CAM je bolj intuitiven za vizualno interpretacijo, saj jasno pokaže regije slike, na katere model usmerja pozornost. Vendar ne podaja informacij o negativnih prispevkih in ne kvantificira natančno vpliva posameznih pikslov.

SHAP na drugi strani podaja podrobnejšo interpretacijo, saj ločuje pozitivne in negativne prispevke. Zaradi tega je informativnejši, vendar tudi vizualno zahtevnejši za

uporabnike brez tehničnega predznanja. V tem eksperimentu je Grad-CAM uporaben za hitro preverjanje, ali CNN model opazuje relevantne regije slike, SHAP pa omogoča globljo analizo prispevkov posameznih regij.

11 Primerjava praktičnih eksperimentov

Izvedena eksperimenta prikazujeta uporabo metod za razložljivost na dveh različnih tipih podatkov. Eksperiment Titanic je osredotočen na tabelarične podatke, kjer so ključne razlage prispevkov atributov, kot so spol, potovalni razred in starost. Eksperiment Fashion-MNIST je osredotočen na slike, kjer je pomembno razumeti prostorske regije, ki vplivajo na odločitev modela.

Tabela 4: Primerjava praktičnih eksperimentov

Značilnost	**Eksperiment 1: Titanic**	**Eksperiment 2: Fashion-MNIST**
Tip podatkov	Tabelarični podatki	Slike
Model	Random Forest	CNN
Metode interpretacije	LIME in SHAP	SHAP in Grad-CAM
Glavni cilj	Razložiti vpliv atributov	Razložiti vizualne regije slike
Najpomembnejši rezultat	Spol, razred, cena vozovnice in starost so ključni dejavniki	Grad-CAM izpostavi relevantne dele objektov, SHAP poda pozitivne in negativne prispevke
Omejitev	Potencialno uhajanje podatkov in odvisnost od kakovosti atributov	Grad-CAM je odvisen od izbire sloja, SHAP je lahko vizualno kompleksen

Takšna kombinacija eksperimentov podaja širšo sliko o interpretabilnosti modelov. Prvi eksperiment kaže, kako metode za razložljivost delujejo na strukturiranih podatkih, drugi pa pokaže, kako se interpretacija prilagodi vizualnim podatkom in globokemu učenju.

12 Zaključek

V tem delu so bile analizirane tehnike za razložljivost v strojnem učenju, s posebnim poudarkom na metodah LIME, SHAP in Grad-CAM. Z naraščanjem kompleksnosti mod-

elov, posebej pri ansambelskih metodah in globokih nevronskih mrežah, postaja potreba po interpretaciji modelov vedno pomembnejša.

V teoretičnem delu so bili predstavljeni osnovni koncepti interpretabilnosti, razlika med lokalno in globalno interpretacijo ter razlika med model-agnostičnimi in model-specifičnimi metodami. LIME je bil predstavljen kot intuitivna metoda za lokalno aproksimacijo modela, SHAP kot teoretično utemeljena metoda na osnovi Shapleyjevih vrednosti, Grad-CAM pa kot vizualna metoda za interpretacijo CNN modelov.

V prvem eksperimentu je bil model Random Forest uporabljen na podatkovnem naboru Titanic. LIME je omogočil lokalno razlago posamezne napovedi, SHAP pa globalni in lokalni vpogled v obnašanje modela. Rezultati so pokazali, da so spol, potovalni razred, cena vozovnice in starost med najpomembnejšimi dejavniki, ki vplivajo na napoved preživetja.

V drugem eksperimentu je bil CNN model uporabljen na podatkovnem naboru Fashion-MNIST. Grad-CAM je prikazal regije slike, na katere model usmerja pozornost, SHAP pa je podal podrobnejši prikaz pozitivnih in negativnih prispevkov. Ta eksperiment kaže, da je treba interpretabilnost prilagoditi tipu podatkov in modelu, ki se uporablja.

Na podlagi izvedene analize lahko zaključimo, da ne obstaja ena univerzalna metoda za razlago vseh modelov. LIME je primeren za hitro in intuitivno lokalno interpretacijo, SHAP ponuja stabilnejše in teoretično utemeljene razlage, Grad-CAM pa omogoča vizualno razumevanje globokih modelov pri delu s slikami. Kombinirana uporaba teh metod lahko pomembno prispeva k odgovornejši in transparentnejši uporabi strojnega učenja.

Literatura

[1] Molnar, C. (2022). *Interpretable Machine Learning*. Dostopno na: `https://christophm.github.io/interpretable-ml-book/`

[2] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

[3] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*.

[4] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Conference on Computer Vision*.

[5] Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*.

[6] Paszke, A., Gross, S., Massa, F., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*.

[7] Vanschoren, J., van Rijn, J. N., Bischl, B., & Torgo, L. (2013). OpenML: Networked Science in Machine Learning. *SIGKDD Explorations*, 15(2), 49–60.

[8] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[9] Neptune.ai. (2023). Explainability, Auditability and Interpretability in Machine Learning. Dostopno na: `https://neptune.ai/blog/explainability-auditability-ml-definitions-techniques-tools`

[10] Analytics Vidhya. (2021). Model Explainability in Machine Learning. Dostopno na: `https://www.analyticsvidhya.com/blog/2021/11/model-explainability/`

[11] DataCamp. (2022). Explainable AI: Understanding and Trusting Machine Learning Models. Dostopno na: `https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models`

[12] Kaggle. Machine Learning Explainability Course. Dostopno na: `https://www.kaggle.com/learn/machine-learning-explainability`

[13] O'Reilly. Introduction to Local Interpretable Model-Agnostic Explanations (LIME). Dostopno na: `https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/`