UP FAMNIT | RAZISKOVALNI SEMINAR | 2026

Tehnike za razložljivost v strojnem učenju

Explainable AI (XAI): LIME, SHAP in Grad-CAM

Študent: Dušan Todorović, 89232094

Mentor: dr. Branko Kavšek

Pregled predstavitve

1 **Motivacija**

Zakaj razložljivost postaja ključna zahteva pri uporabi UI

2 **Temeljni pojmi**

Interpretabilnost, lokalna in globalna razlaga, post-hoc metode

3 **Tri metode**

LIME, SHAP in Grad-CAM — ideja, prednosti in omejitve

4 **Eksperiment 1**

Titanic + Random Forest: interpretacija z LIME in SHAP

5 **Eksperiment 2**

Fashion-MNIST + CNN: primerjava SHAP in Grad-CAM na slikah

6 **Zaključek**

Katera metoda za kateri problem — ključne ugotovitve

Zakaj je razložljivost pomembna?

Kompleksni modeli so „črne škatle“ — visoka točnost ne zadošča

Medicina

Zdravnik mora razumeti razloge za diagnozo — točnost sama po sebi ni dovolj za varno klinično odločanje.

Finance

Preglednost pri ocenjevanju kreditnega tveganja je zahteva regulatorjev; netransparentni modeli tvegajo diskriminatorne odločitve.

Etika in pravo

GDPR posamezniku zagotavlja pravico do razlage avtomatiziranih odločitev, ki nanj vplivajo.

Pristranskost in napake

Modeli se učijo iz zgodovinskih podatkov s pristranskostmi — razložljivost razkriva skrite sistemske napake.

Temeljni pojmi razložljivosti

Interpretabilnost

Model je neposredno razumljiv človeku (linearna regresija, odločitveno drevo).

Razložljivost

Dodatne metode za pojasnitev kompleksnih modelov, ki sami po sebi niso razumljivi.

Lokalna interpretacija

Zakaj je model sprejel konkretno odločitev za posamezno instanco.

Globalna interpretacija

Splošno vedenje modela prek celotnega podatkovnega nabora.

Model-agnostične metode

Delujejo s katerimkoli modelom — obravnavajo ga kot črno škatlo (LIME, SHAP).

Post-hoc metode + zvestoba

Uporabijo se po učenju modela; razlaga mora zvesto (fidelity) odražati njegovo dejansko vedenje.

LIME — lokalna razlaga modela

Local Interpretable Model-agnostic Explanations | Ribeiro, Singh & Guestrin (2016)

OSNOVNA IDEJA

Vedenje kompleksnega modela v lokalni okolici opazovane instance aproksimiramo z enostavnim, interpretabilnim (linearnim) modelom. Izvirni model obravnava kot črno škatlo.

POSTOPEK

1. Generiranje perturbacij v bližini instance
2. Napovedi izvirnega modela za perturbacije
3. Uteževanje vzorcev glede na oddaljenost (π_x)
4. Učenje lokalnega linearnega modela
5. Koeficienti = razlaga prispevkov atributov

$$\xi(x) = \operatorname{argmin}_{g \in G} \; L(f, g, \pi_x) + \Omega(g)$$

PREDNOSTI

+ Deluje z vsakim modelom (model-agnostična)
+ Intuitivna in hitro razumljiva razlaga
+ Hitra izvedba v praksi

OMEJITVE

– Nestabilnost rezultatov med ponovitvami
– Občutljivost na parametre (širina jedra, št. vzorcev)
– Brez globalnega vpogleda v model

SHAP — razlaga s Shapleyjevimi vrednostmi

SHapley Additive exPlanations | Lundberg & Lee (2017)

OSNOVNA IDEJA

Temelji na Shapleyjevih vrednostih iz kooperativne teorije iger: vsak atribut je „igralec“, ki prispeva h končni napovedi. Prispevek se porazdeli pošteno in konsistentno.

VARIANTE METODE

1. TreeSHAP — optimiziran za drevesne modele (Random Forest)
2. KernelSHAP — model-agnostična različica
3. DeepSHAP / GradientExplainer — za nevronske mreže

$$f(x) = \varphi_0 + \Sigma_i \varphi_i \quad \textit{(aditivni model prispevkov)}$$

PREDNOSTI

+ Teoretično utemeljena metoda (teorija iger)
+ Konsistentna in stabilna razlaga
+ Lokalna in globalna interpretacija
+ Bogate vizualizacije: bar, beeswarm, waterfall

OMEJITVE

– Računsko zahtevnejša od LIME
– Interpretacija zahteva več tehničnega znanja

Grad-CAM — vizualna razlaga CNN

Gradient-weighted Class Activation Mapping | Selvaraju et al. (2017)

OSNOVNA IDEJA

Razvita za konvolucijske nevronske mreže. Gradienti izhoda za ciljni razred glede na aktivacije zadnje konvolucijske plasti tvorijo toplotno karto, ki pokaže, kam model „gleda“.

POSTOPEK

1. CNN klasificira vhodno sliko
2. Izračun gradientov $\partial y^c / \partial A^k$ za ciljni razred
3. Prostorsko povprečenje gradientov → uteži α^k
4. Utežena kombinacija aktivacijskih map
5. ReLU + projekcija toplotne karte na sliko

$$L^c = ReLU(\Sigma_k \alpha^{kc} A^k)$$

PREDNOSTI

+ Intuitivna vizualna razlaga (toplotna karta)
+ Ne zahteva sprememb arhitekture modela
+ Uporabna v medicini in računalniškem vidu

OMEJITVE

– Omejena na CNN modele
– Manj natančna od metod na ravni pikslov
– Ne loči negativnih prispevkov (za razliko od SHAP)

Primerjava metod

LIME, SHAP in Grad-CAM niso konkurenčne, temveč komplementarne metode

Značilnost	LIME	SHAP	Grad-CAM
Vrsta metode	Model-agnostična	Model-agnostična / specifične različice	Model-specifična (CNN)
Raven interpretacije	Lokalna	Lokalna + globalna	Lokalna, vizualna
Teoretična osnova	Lokalna aproksimacija	Shapleyjeve vrednosti (teorija iger)	Gradienti v CNN
Tip podatkov	Tabele, besedilo, slike	Tabele, besedilo, slike	Slike
Rezultat	Seznam prispevkov atributov	Prispevki + bar / beeswarm / waterfall	Toplotna karta
Glavna prednost	Intuitivnost	Konsistentnost	Vizualna jasnost
Glavna omejitev	Nestabilnost	Računska zahtevnost	Omejenost na CNN

Eksperiment 1: Titanic + Random Forest

Metodologija, priprava podatkov in rezultati klasifikacije

OPIS IN PRIPRAVA

Nabor OpenML Titanic; binarna klasifikacija: ali je potnik preživel (survived).

- Odstranjeni stolpci boat, body, name, ticket (preprečitev uhajanja podatkov)
- Delitev 80 % : 20 % s stratifikacijo
- Numerične spr.: imputacija z mediano + standardizacija
- Kategorične spr.: imputacija z modusom + one-hot kodiranje
- Model: RandomForestClassifier, 500 dreves, uravnoteženje razredov

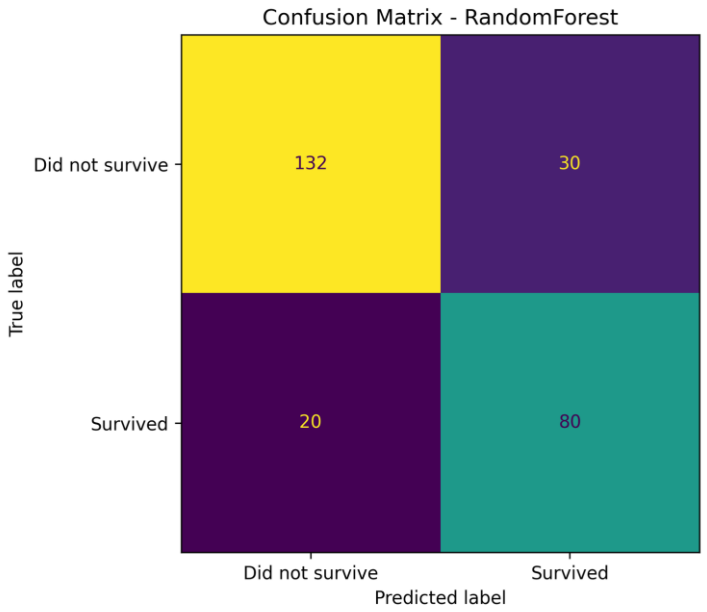
Test z uhajanjem podatkov: accuracy 0.943, AUC 0.983 — napihnjeno in metodološko neustrezno; model uporablja informacije o izidu.

REZULTATI EVALVACIJE

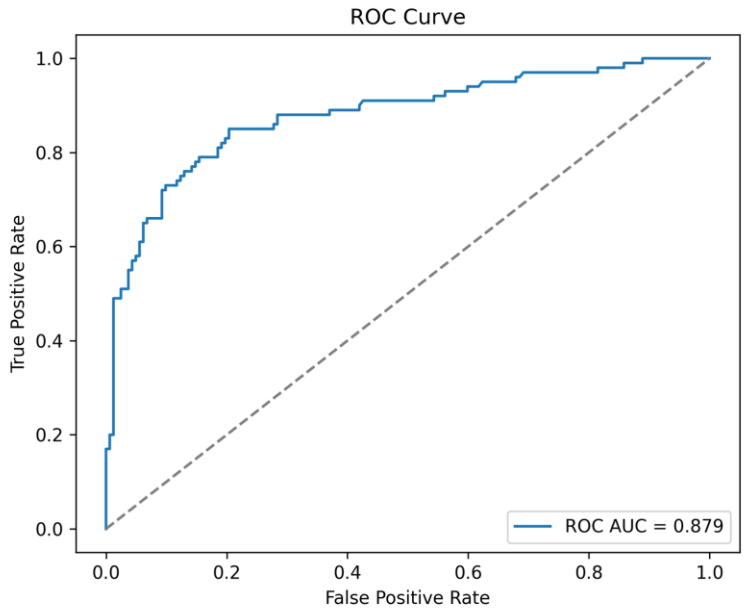
0.809 Accuracy	**0.727** Precision
0.800 Recall	**0.762** F1-mera
0.879 ROC AUC	**0.849** Avg. Precision

Evalvacija modela Random Forest

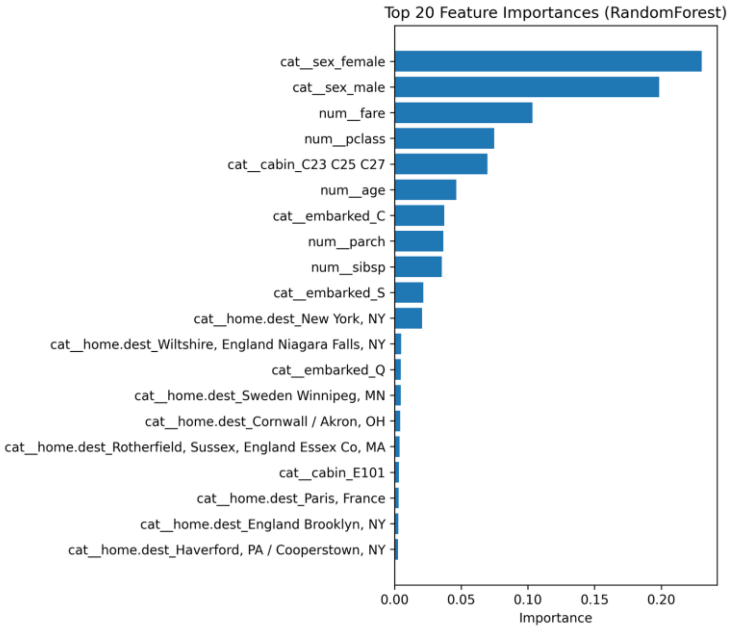
Matrika zmede, ROC krivulja in najpomembnejši atributi



Matrika zmede



ROC krivulja (AUC = 0.879)

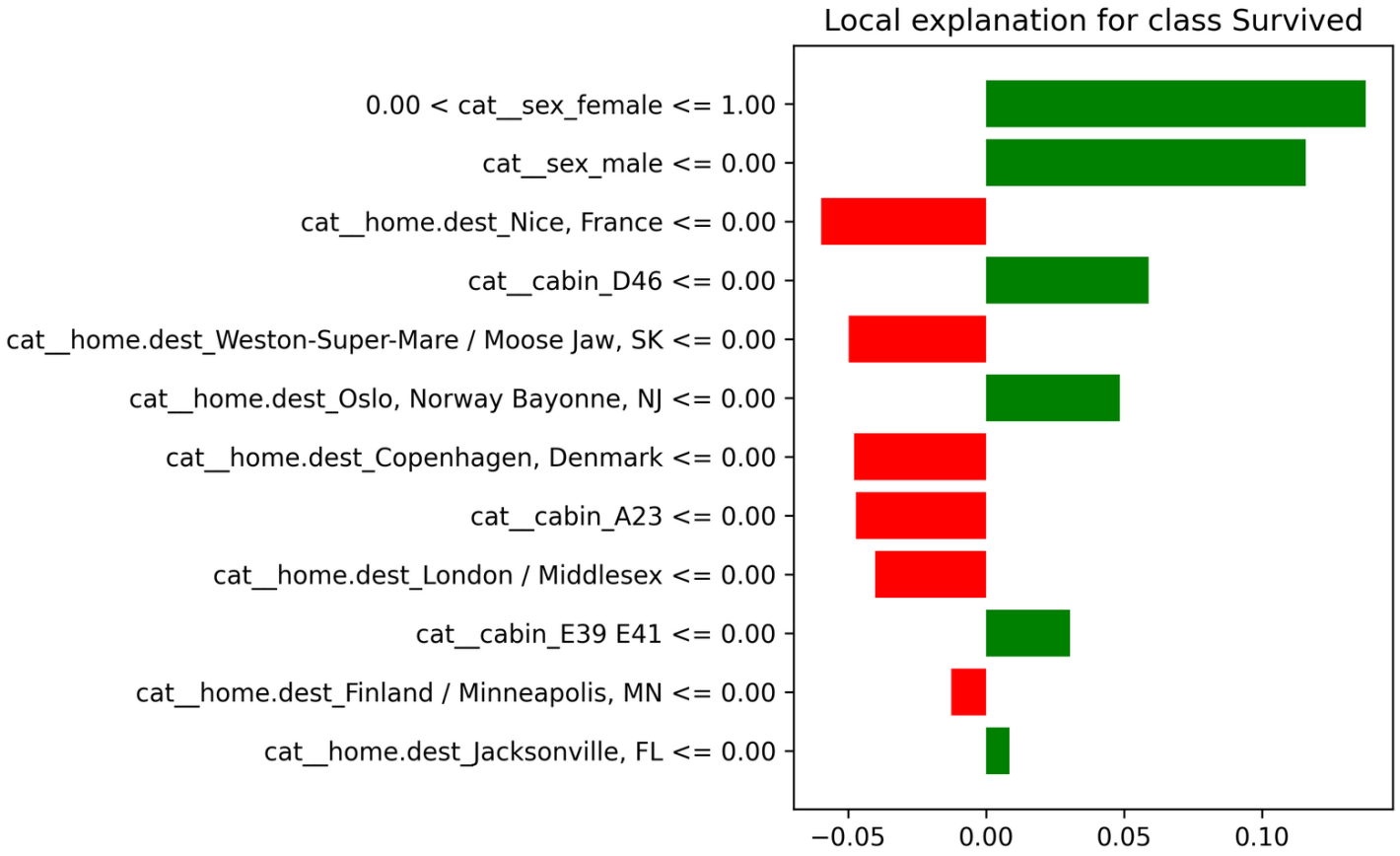


Pomembnost atributov

Najpomembnejši atributi: spol, potovalni razred, cena vozovnice in starost — skladno z zgodovinskimi podatki o katastrofi (prednost žensk in višjih razredov).

Interpretacija z LIME

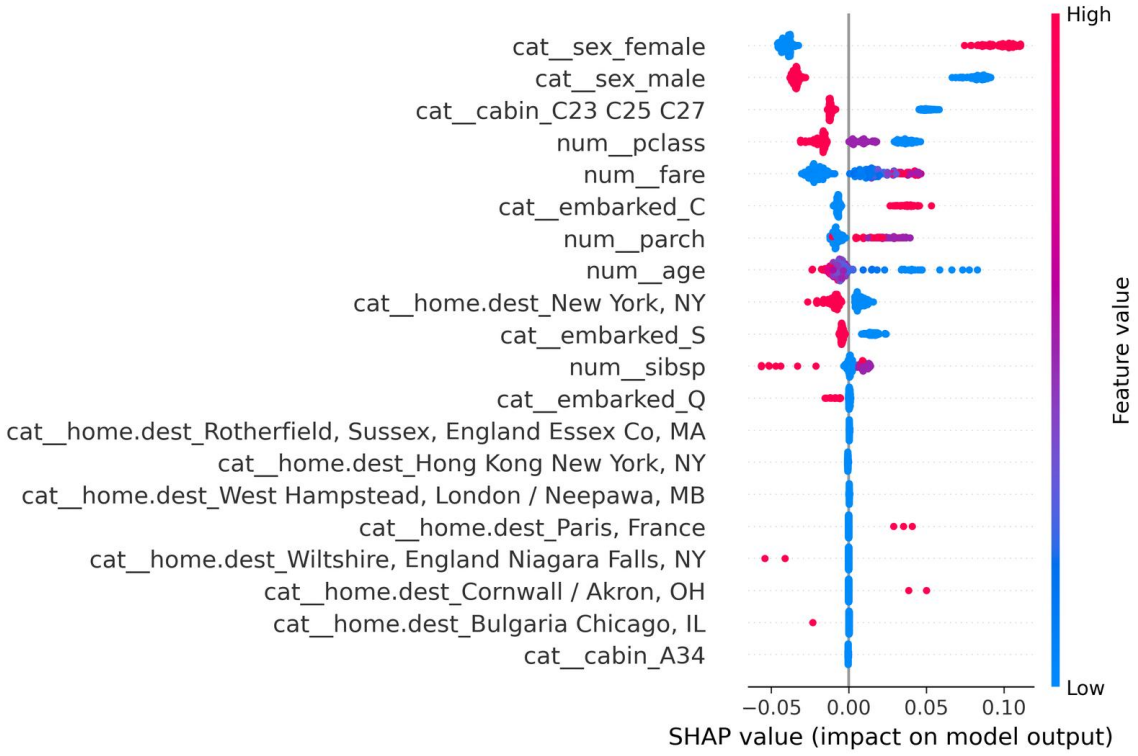
Lokalna razlaga ene napovedi iz testne množice



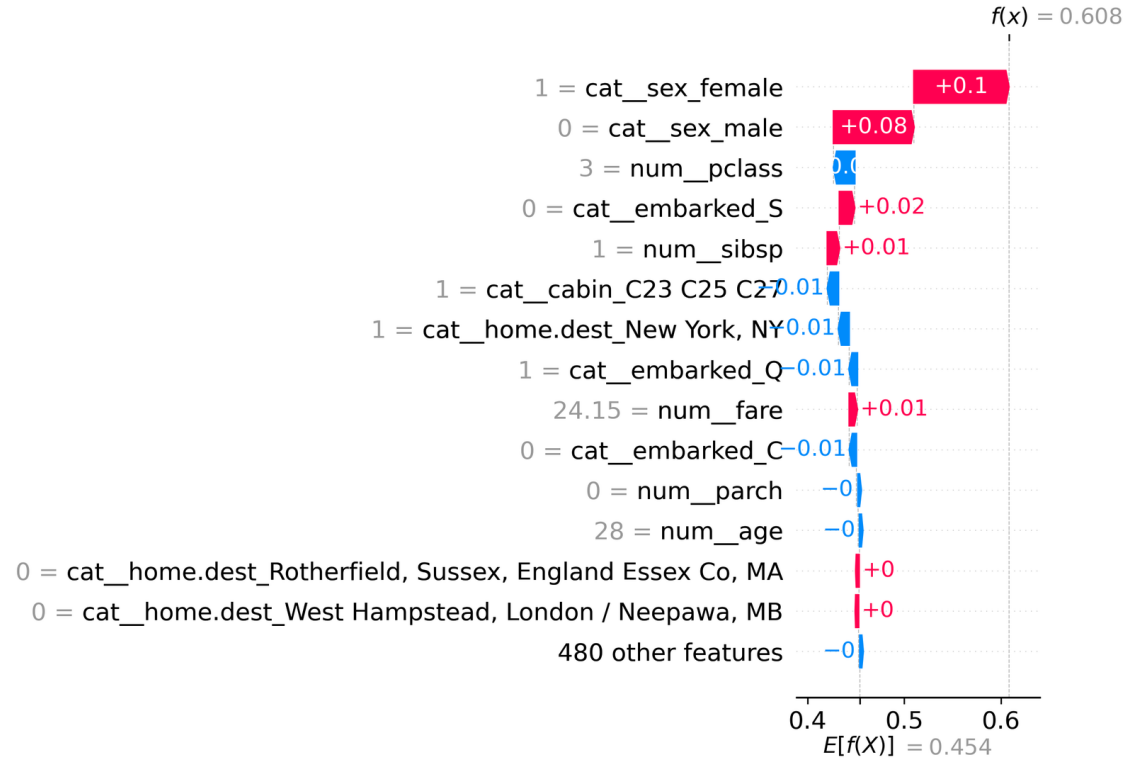
- Razlaga = lokalna linearna kombinacija atributov
- Zeleni (pozitivni) prispevki povečujejo verjetnost razreda „preživel“
- Rdeči (negativni) prispevki jo zmanjšujejo
- Spol (ženska) ima največji lokalni vpliv
- Prednost: intuitivnost — takoj vidimo, kaj je odločilo

Interpretacija s SHAP

Globalna razlaga (beeswarm) in lokalna razlaga (waterfall)



Globalno: beeswarm — pomembnost in smer vpliva



Lokalno: waterfall — napoved kot vsota prispevkov φ_i

SHAP potrjuje: spol, cena vozovnice, razred in starost so ključni — razlaga je stabilnejša in teoretično utemeljenejša kot pri LIME.

Eksperiment 2: Fashion-MNIST + CNN

Interpretabilnost v domeni slik — primerjava SHAP in Grad-CAM

OPIS IN MODEL

- 10 razredov oblačil, sivinske slike 28 × 28 px
- CNN v PyTorch: 3 konvolucijski sloji, max-pooling, dropout
- Grad-CAM na zadnjem konvolucijskem sloju
- SHAP z gradientno razlago (GradientExplainer)

Krivulja izgube med učenjem

REZULTATI EVALVACIJE

0.914
Accuracy

0.914
Macro precision

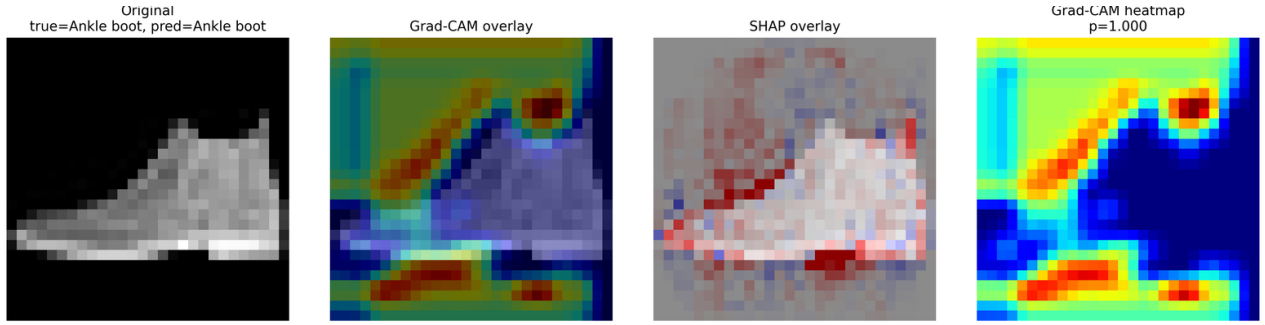
0.914
Macro recall

0.913
Macro F1

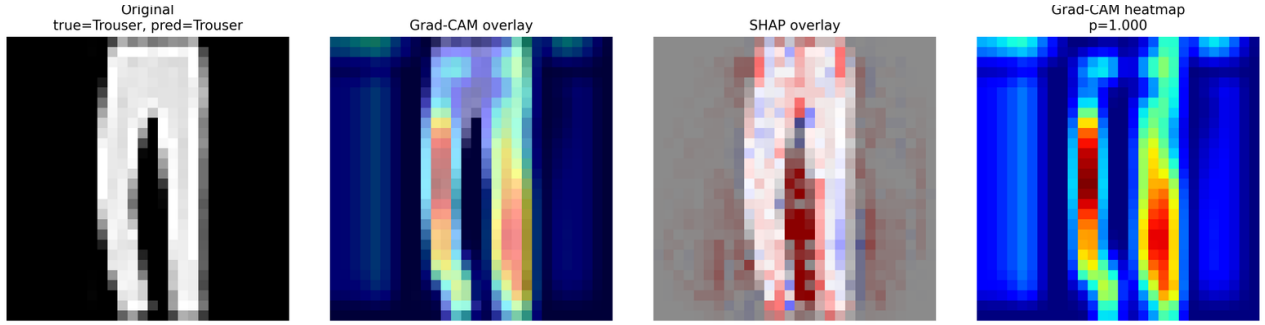
Barve razlag: Grad-CAM: tople barve = pomembne regije (le pozitivna podpora). SHAP: rdeča = pozitiven, modra = negativen prispevek.

Grad-CAM in SHAP na slikah

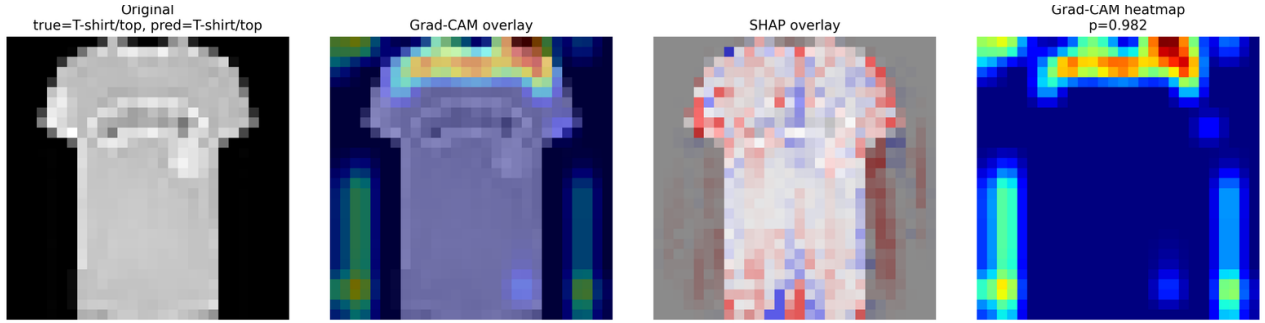
Original | Grad-CAM overlay | SHAP overlay | Grad-CAM toplotna karta



Ankle boot — Grad-CAM izpostavi obliko obutve; SHAP loči pozitivne in negativne regije



Trouser — Grad-CAM sledi navpični strukturi hlačnic



T-shirt/top — poudarek na vratu in ramenih, značilnih za razred

Zaključek

Ni univerzalne metode — izbira je odvisna od podatkov in modela

1 Razložljivost je nujnost

V medicini, financah in etični uporabi UI ni več tehnični dodatek, temveč zahteva.

2 LIME — hitro in intuitivno

Lokalna model-agnostična razlaga; primerna za hitre vpoglede, a manj stabilna.

3 SHAP — zanesljivo in celovito

Teoretično utemeljena, konsistentna lokalna in globalna razlaga.

4 Grad-CAM — vizualni vpogled v CNN

Toplotna karta pokaže relevantne regije slike; komplementaren s SHAP.

Kombinirana uporaba metod prispeva k odgovornejši in transparentnejši uporabi strojnega učenja.

Hvala za pozornost

Vprašanja?

KLJUČNA LITERATURA

- Ribeiro, Singh & Guestrin (2016). „Why Should I Trust You?“: Explaining the Predictions of Any Classifier. ACM SIGKDD.
- Lundberg & Lee (2017). A Unified Approach to Interpreting Model Predictions. NeurIPS.
- Selvaraju et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. ICCV.
- Molnar, C. (2022). Interpretable Machine Learning. christophm.github.io/interpretable-ml-book
- Xiao, Rasul & Vollgraf (2017). Fashion-MNIST: a Novel Image Dataset for Benchmarking ML Algorithms. arXiv:1708.07747.
- Paszke et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. NeurIPS.
- Vanschoren et al. (2013). OpenML: Networked Science in Machine Learning. SIGKDD Explorations.
- Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. JMLR.