Analiza pristranosti u ocjenjivanju projekata - verzija s jednostavnim objašnjenjima Zašto je sustav sustav vrednovanja u okviru Projekta razvoja karijera mladih istraživača (DOK-2025-02) prijavljenih projekata s predloženim mentorima za znanstveno područje "Prirodne znanosti" dao besmislene i neupotrebljive rezultate?

Branimir K. Hackenberger

5. kolovoza 2025.

VAŽNO: Originalni dokument s detaljnim prikazom metodologije, rezultata i njihove interpretacije nalazi se u dokumentu: **Analiza_sustava_evaluacije_DOK_2025_02.pdf**

Najvažniji zaključak

Sustav ocjenjivanja je pokazao da samo 9% ocjena ovisi o kvaliteti projekata, dok 91% ovisi o tome koji recenzent ocjenjuje i čistoj sreći. To je kao da bacate kocku umjesto da ocjenjujete projekte!

1 O čemu se radi?

Zamislite da ste predali projekt s predloženim mentorom na natječaj za financiranje doktoranta. Tri različita stručnjaka (recenzenta) trebaju ocijeniti vaš projekt i mentora ocjenama od 1 do 5. Očekivali biste da će njihove ocjene biti slične ako su projekt i mentor dobri, zar ne?

Problem: Analizirali smo 95 projekata koje je ocijenilo ukupno 285 recenzenata (no vjerojatno ih je bilo manje) i otkrili kako sustav ocjenjivanja ima ozbiljne probleme. Recenzenti se ne slažu, ocjene su pristrane, a cijeli sustav nije bolji od bacanja kocke.

2 Što smo analizirali?

- 95 projekata iz područja prirodnih znanosti
- Svaki projekt ocijenila su 3 recenzenta
- Ukupno 285 ocjenjivanja
- 9 različitih kriterija po projektu raspoređeni u tri kategorije K1, K2 i K3
- Fokus je bio pored ukupne ocjene na kriterij K3 koji je bio posebno važan

3 Kako smo analizirali - Jednostavno objašnjenje metoda analiza

3.1 1. Koliko se recenzenti slažu? (ICC analiza)

Što je to ICC? ICC (*Intraclass Correlation Coefficient*) je neštro kao "mjerač slaganja". Ako svi recenzenti daju slične ocjene istom projektu (mentoru), ICC je visok. Ako daju potpuno različite ocjene, ICC je nizak.

Zašto je važno? Ako tri stručnjaka gleda isti projekt (mentora) i jedan kaže "odličan" (5), drugi "loš" (2), a treći "srednji" (3), kako znati koliko projekt stvarno vrijedi?

Što bi bilo dobro? ICC bi trebao biti najmanje 0.40 (40% slaganja) da bi sustav bio upotrebljiv. Izvrsno bi bilo preko 0.75.

3.2 2. Odakle dolaze razlike u ocjenama? (*Mixed Effects Model*, MEM)

Što je to? MEM MEM je metoda koja razdvaja ukupne razlike u ocjenama na tri dijela:

1. Razlike zbog kvalitete projekata (**to želimo!**)
2. Razlike zbog različitih recenzenata (**to ne želimo!**)
3. Nasumične razlike (**ovo također ne želimo!**)

Zašto je važno? Zamislite da analizirate zbog čega neki učenici imaju bolje ocjene, je li to zbog njihovog znanja, zbog toga što imaju blaže profesore, ili zbog biolje sreće?

3.3 3. Postoje li različiti "tipovi" recenzenata? (Klaster analiza)

Što je to klaster analiza (hrvatski, vjerovali ili ne, "rojna anliza")? Pomoću klaster analize grupiraju se recenzenti koji ocjenjuju na sličan način. Neki uvijek daju visoke ocjene ("blagi"), neki niske ("strogi"), neki variraju.

Zašto je važno? Ako vaš projekt (mentor) dobije tri "blaga" recenzenta, imat ćete prednost nad nekim čiji su recenzenti "strogi".

3.4 4. Koliko se recenzenti slažu o rangiranju? (Kendall-ov W)

Što je to Kendall-ov W)? To je mjera koliko se recenzenti slažu u tome koji su projekti (mentori) bolji, a koji lošiji. Vrijednost ide od 0 (potpuno neslaganje) do 1 (potpuno slaganje).

Zašto je važno? Čak i ako recenzenti daju različite ocjene, trebali bi se barem slagati koji projekti su na vrhu, a koji na dnu.

3.5 5. Je li sustav bolji od nasumičnog? (Monte Carlo simulacija)

Što je to Monte Carlo simulacija? To si najbolje možete predočiti na način da zamislite da bacite kocku s 5 brojeva 1000 puta (nam to srećom radi računalo) i dobijate nasumične ocjene. Nakon toga gledate jesu li prave ocjene bolje od nasumičnih.

Zašto je važno? Ako pravi sustav nije puno bolji od bacanja kocke, zašto uopće imati recenzente?

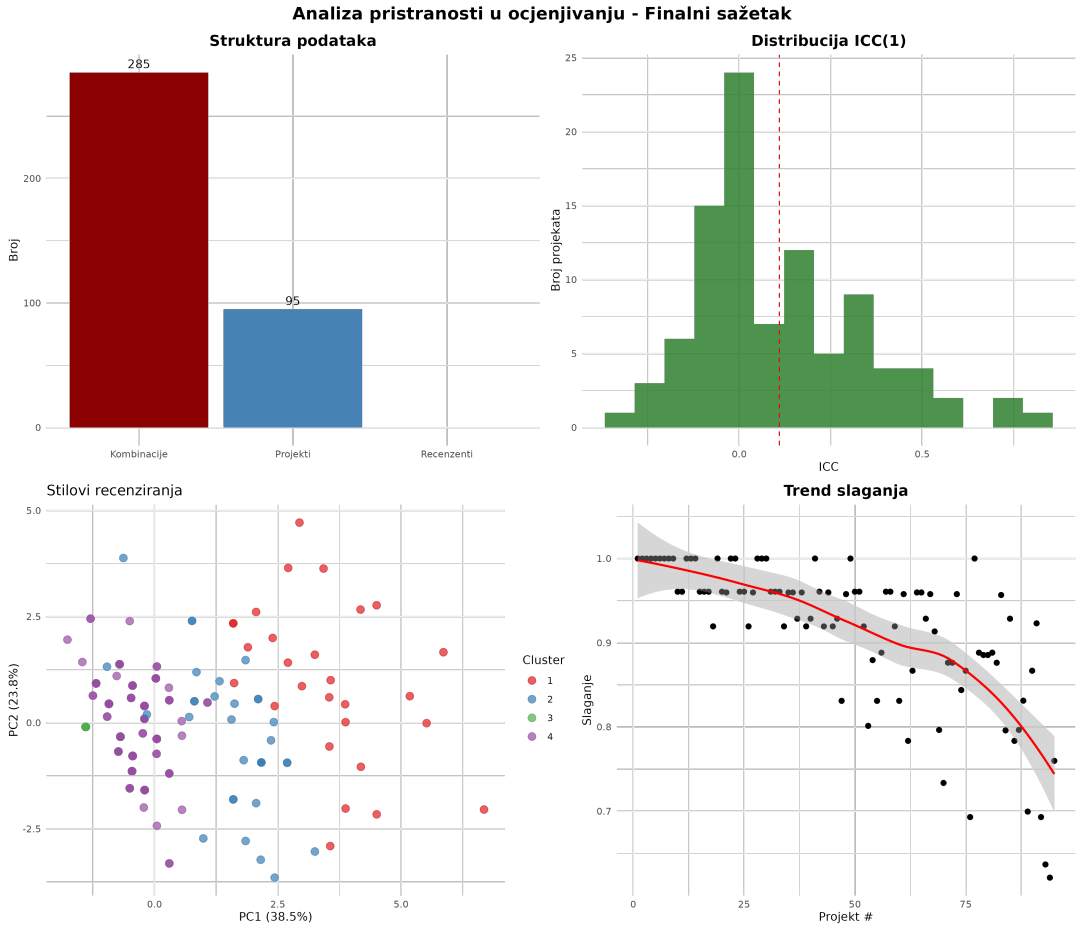
3.6 6. Koriste li recenzenti cijelu skalu? (Ceiling učinak)

Što je to? Na taj način možemo provjeriti daju li recenzenti uglavnom visoke ocjene (4 i 5) ili koriste cijelu skalu (1-5).

Zašto je važno? Ako svi dobiju 4 ili 5, kako razlikovati dobre od izvrsnih projekata (mentora)?

4 Što smo otkrili? - Rezultati

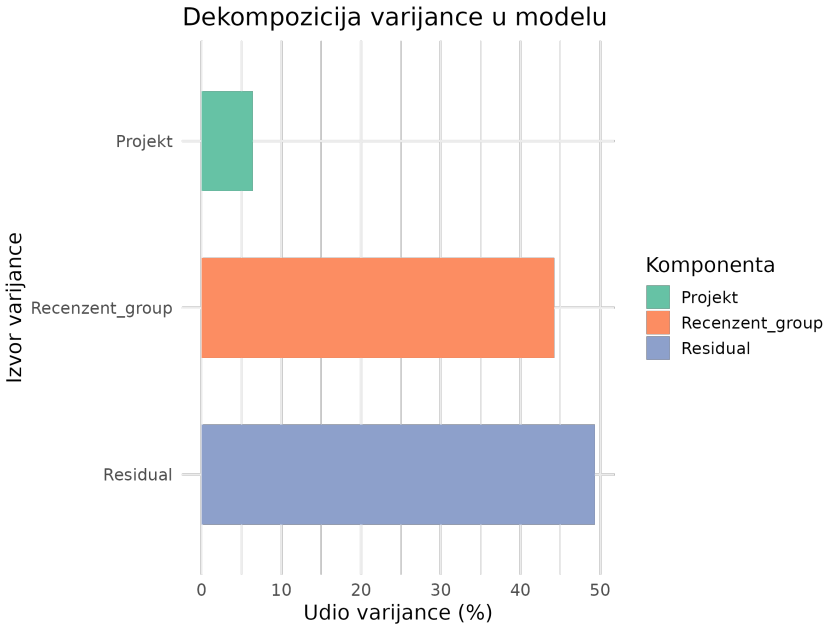
4.1 Recenzenti se ne slažu



Slika 1: Pregled rezultata. Gornji desni graf pokazuje da je prosječno slaganje recenzenata (ICC) samo 0.11 - daleko ispod minimalno prihvatljivih 0.40!

Jednostavno rečeno: Recenzenti se slažu u samo 11% slučajeva. To je kao da tri osobe gledaju iste slike i od 100 slika na njih 89 se ne mogu složiti je li na njima pas, mačka ili konj!

4.2 Ocjene ne ovise o projektima!



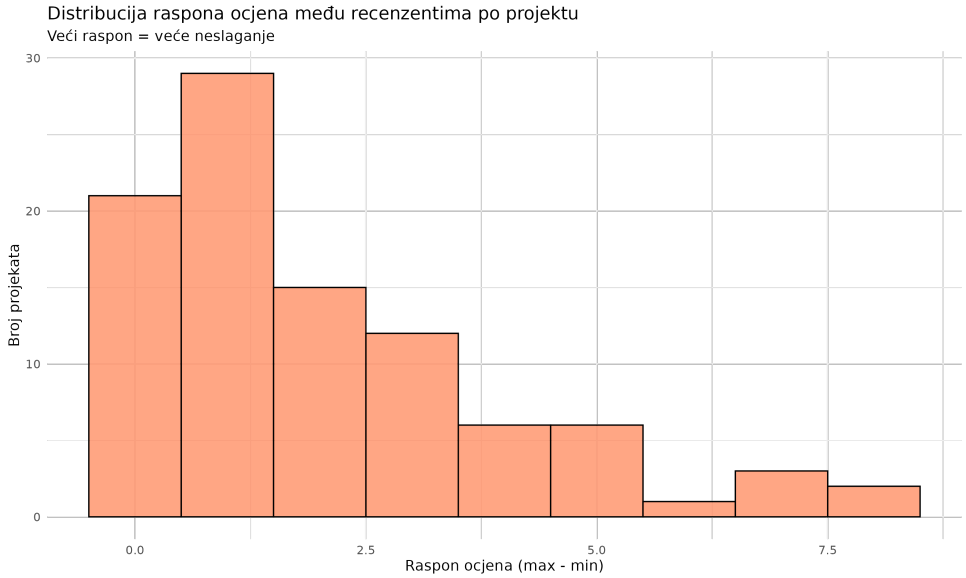
Slika 2: Ovaj graf pokazuje odakle dolaze razlike u ocjenama. Zeleni dio (Projekt (mentor)) je ono što želimo tj. razlike zbog kvalitete projekata. Nažalost, to je samo 6.4%!

Šokantno:

- Samo 6.4% varijance ocjena ovisi o kvaliteti projekta
- 44.2% ovisi o tome koji recenzent ocjenjuje
- 49.3% je čista nasumičnost

To znači da je velika većina ocjena više lutrija nego stvarna procjena kvalitete!

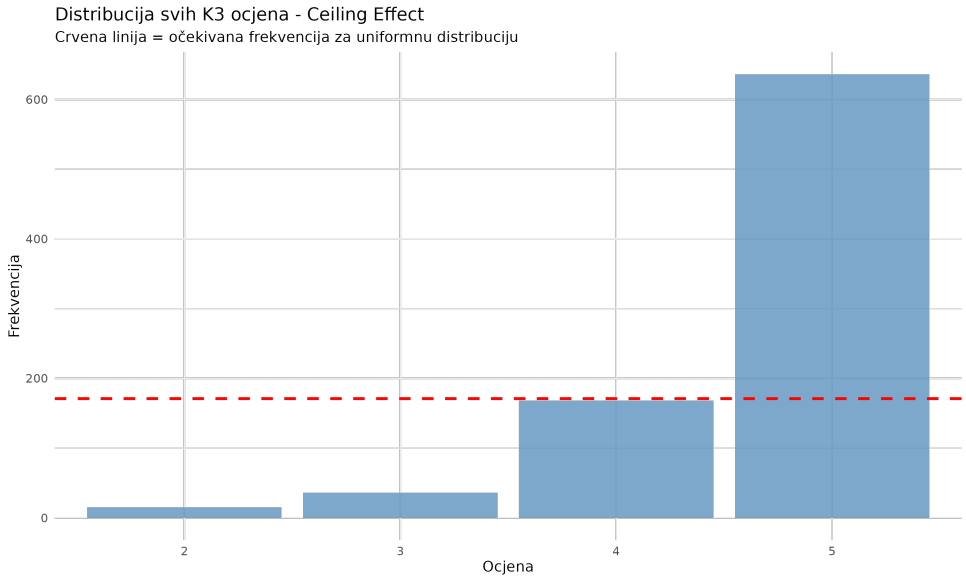
4.3 Neki projekti dobivaju potpuno različite ocjene



Slika 3: Ovaj graf pokazuje koliko se razlikuju ocjene istog projekta. Neki projekti (desna strana grafa) imaju razlike do 8 bodova između recenzenata!

Primjer: Projekt P94 je dobio ocjene 7, 10 i 15 od tri različita recenzenta. To je kao da jedan profesor da učeniku dvojku, drugi trojku, a treći peticu za isti ispit!

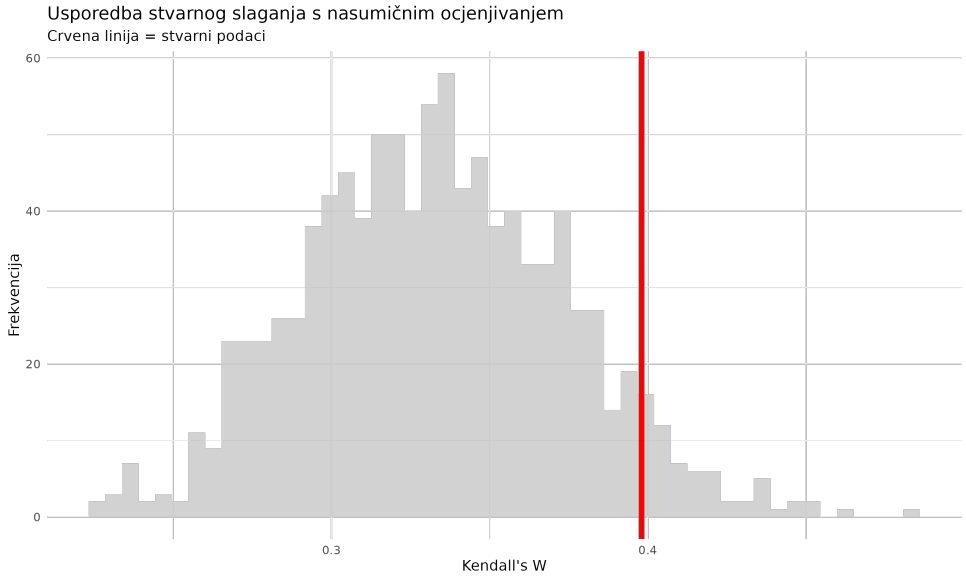
4.4 Svi daju visoke ocjene



Slika 4: Ovaj graf pokazuje koliko je koja ocjena česta. Vidite da su ocjene 4 i 5 daleko najčešće (94% svih ocjena), dok se ocjene 1, 2 i 3 gotovo ne koriste.

Problem: Ako svi dobiju 4 ili 5, kako znati tko je stvarno najbolji? To je kao da u školi svi učenici dobiju četvorke i petice, pa onda doista ne možete razlikovati prosječne od izvrsnih!

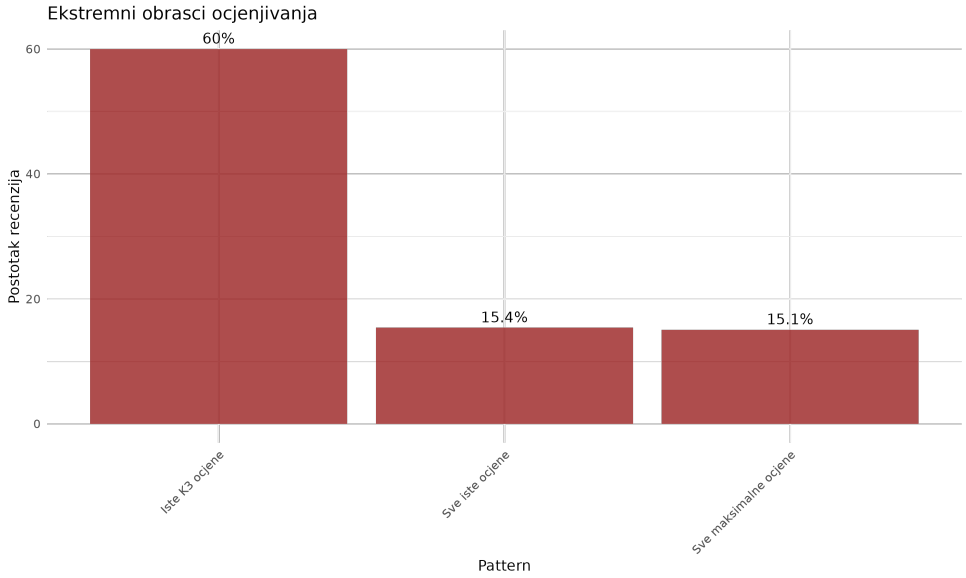
4.5 Sustav nije bolji od kocke



Slika 5: Crvena linija pokazuje koliko se slažu pravi recenzenti. Sivi stupci pokazuju koliko bi se slagali da bacaju kocku. Vidite da je crvena linija jedva nešto bolja!

Nevjerojatno: Samo 6% nasumičnih simulacija bilo je gore od pravog sustava. To znači da bi bacanje kocke dalo gotovo jednako dobre rezultate kao plaćanje (ili čekanje) recenzenata!

4.6 Recenzenti su lijeni i/ili nekritični



Slika 6: Ovaj graf pokazuje koliko recenzenata daje iste ocjene za sve. 60% ih daje iste ocjene za K3 komponente, a 15% daje maksimalne ocjene za sve!

***Copy-paste* ocjenjivanje**: Mnogi recenzenti očito ne čitaju projekte pažljivo već samo stavljaju iste ocjene za sve. To je kao da profesor ocijeni sve zadatke s 5 bez čitanja!

5 Dodatna (Bayesova) analiza - još gore!

Bayesova analiza je sofisticiranija metoda koja potvrđuje sve probleme:

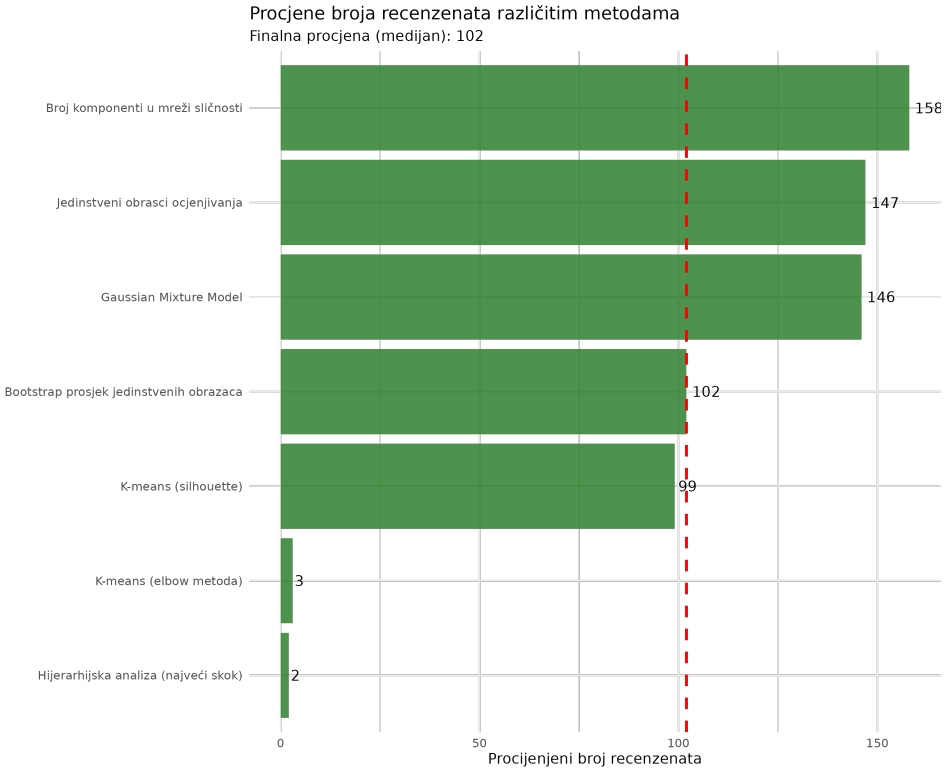
Tablica 1: Odakle dolaze razlike u ocjenama - Bayesova analiza

Izvor razlika	**Postotak**
Kvaliteta projekata	9%
Različiti recenzenti	32%
Slučajnost	59%

Još jedan dokaz: Čak i najsofisticiranija analiza pokazuje da samo 9% ocjena ovisi o projektima!

6 Koliko je bilo recenzenata?

Ne znamo koliko je doista bilo recenzenata. Je li svaki recenzent vrednovapo samo jedan projekt ili su neki recenzirali i više projekata?



Slika 7: Različite metode procjenjuju da je bilo oko 102 različita recenzenta za 95 projekata.

Što to znači? Svaki recenzent je u prosjeku ocijenio 2.8 projekata. To stvara dodatne probleme jer različiti projekti imaju različite skupine recenzenata.

7 Zaključak - sustav je potpuno nefunkcionalan

Glavni zaključci

1. **Ocjene ne mjere kvalitetu**: Samo 6-9% ocjena ovisi o projektima
2. **Svi su doibili visoke ocjene**: 94% ocjena je 4 ili 5
3. **Recenzenti se ne slažu**: Razlike do 8 bodova za isti projekt
4. **Sustav = bacanje kocke**: Statistički nije bolji od nasumičnog

8 Što to znači za vas?

Ako ste prijavili projekt:

- Vaša ocjena više ovisi o sreći nego o kvaliteti projekta
- Važnije je koji recenzenti vam se dodijele nego koliko je projekt dobar
- Sustav ne može razlikovati dobre od loših projekata

Ako ste samo porezni obveznik:

- Vaši novci se dijele u znanstvenoj zajednici po principu lutrije
- Raspodjela sredstava se izvodi temeljem recenzenata čiji rad nije bolji od bacanja kocke
- Najbolji znanstveni projekti možda neće dobiti financiranje jer ih sustav ne može prepoznati

9 Preporuke

Što treba učiniti?

1. **HITNO**: Poništiti rezultate ovog ocjenjivanja i dodijeliti doktorante svim podnositeljima zahtjeva
2. **KRATKOROČNO**: Potpuno redizajnirati sustav
3. **DUGOROČNO**: Razmotriti alternativne načine (npr. ponderirani ždrijeb gdje bolji projekti imaju veće šanse, ali bi svi imali priliku)

10 Završna poruka

Ovaj sustav ocjenjivanja je znanstvena katastrofa. Ne može razlikovati dobre od loših projekata i nije bolji od bacanja kocke. Mora se hitno promijeniti!

Jedino pravedno rješenje: dodijeliti sredstva SVIM prijaviteljima jer sustav nije utvrdio tko je bolji!

WWW verzija sažetka dostupna na poveznici https://branimir-k-hackenberger.github.io/