Binarna klasifikacija jestivih i nejestivih gljiva

uz odabir optimalnog klasifikatora i interpretaciju značajki

Projektni zadatak iz kolegija Skriptni jezici

Tin Brletić

6. veljače 2026.

Sadržaj

1 Uvod

Razlikovanje jestivih od otrovnih gljiva jedan je od praktički najvažnijih problema u biologiji. Pogrešna identifikacija može rezultirati teškim trovanjima, pa čak i smrtnim ishodom. S razvojem metoda strojnog učenja otvorila se mogućnost automatske klasifikacije na temelju morfoloških karakteristika gljiva.

Cilj ovog projekta je analizirati skup podataka o gljivama (UCI Mushroom Classification) i primjenom različitih algoritama binarne klasifikacije utvrditi koji klasifikator najuspješnije razlikuje jestive od nejestivih gljiva. Skup sadrži 8.124 uzorka s 22 kategorijske značajke koje opisuju fizičke karakteristike gljiva, pri čemu je svaki uzorak označen kao jestiv (`e`) ili otrovan (`p`).

Osim odabira najboljeg klasifikatora, važan dio projekta je i **selekcija značajki** (*feature selection*) – postupak kojim se utvrđuje koje karakteristike gljiva najviše utječu na njihovu jestivost. To je ključno za interpretabilnost modela i praktičnu primjenu u edukaciji, mobilnim aplikacijama za identifikaciju gljiva ili sustavima ranog upozorenja.

Rezultati ovog projekta mogu se primijeniti u razvoju mobilnih aplikacija za prepoznavanje gljiva, edukaciji neiskusnih skupljača (identificirane ključne značajke poput mirisa), farmaceutskim istraživanjima toksikologije gljiva, te kao metodološki okvir za druge probleme binarne klasifikacije.

2 Objašnjavanje projekta

2.1 Opis skupa podataka

Korišteni skup podataka je **UCI Mushroom Classification** preuzet s platforme Kaggle. Pohranjen je u datoteci `mushrooms.csv` i sadrži 8.124 uzorka s ukupno 23 stupca (1 ciljna varijabla + 22 značajke). Distribucija klasa je približno uravnotežena: oko 52% uzoraka pripada klasi `e` (jestiva), a oko 48% klasi `p` (otrovna).

Sve značajke su kategorijske i opisuju morfološke karakteristike gljiva:

Tablica 1: Značajke u skupu podataka mushrooms.csv

Značajka	**Opis**
class	Ciljna varijabla (e = jestiva, p = otrovna)
cap-shape	Oblik klobuka
cap-surface	Površina klobuka
cap-color	Boja klobuka
bruises	Prisutnost modrica
odor	Miris
gill-attachment	Način pričvršćivanja lamela
gill-spacing	Razmak između lamela
gill-size	Veličina lamela
gill-color	Boja lamela
stalk-shape	Oblik stabljike
stalk-root	Korijen stabljike
stalk-surface-above-ring	Površina stabljike iznad prstena
stalk-surface-below-ring	Površina stabljike ispod prstena
stalk-color-above-ring	Boja stabljike iznad prstena
stalk-color-below-ring	Boja stabljike ispod prstena
veil-type	Tip vela
veil-color	Boja vela
ring-number	Broj prstenova
ring-type	Tip prstena
spore-print-color	Boja otiska spora
population	Populacija
habitat	Stanište

Budući da su sve značajke kategorijske, u fazi predobrade primjenjuje se **one-hot encoding** (`pd.get_dummies`) čime se svaka kategorijska varijabla rastavlja na niz binarnih stupaca. Nakon kodiranja, broj značajki raste sa 22 na 117.

2.2 Tok izvršavanja skripte (`feature_selection.py`)

Skripta `feature_selection.py` implementira cjelokupni eksperimentalni tok projekta. U nastavku je opisan redoslijed koraka:

1. **Učitavanje i predobrada podataka** – Skripta učitava `mushrooms.csv`, odvaja ciljnu varijablu (`class`) od značajki te primjenjuje one-hot encoding na sve kategorijske stupce. Implementiran je i sustav cachiranja (Parquet format, fingerprint dataseta) za brže ponovljeno učitavanje.

2. **Statistička selekcija značajki** – Provodi se 5 statističkih testova (Mann-Whitney U, Wilcoxon rank-sum, Kruskal-Wallis, Chi-square, KS) za odabir top 5 značajki, uz Benjamini-Hochberg korekciju i minimalni prag veličine učinka ≥ 0.10.

3. **Definicija klasifikatora** – Ukupno 40+ konfiguracija modela: 6 baseline klasifikatora na svim značajkama, $6 \times 5 = 30$ kombinacija klasifikatora i statističkih testova (in-CV selekcija), te 10 konfiguracija s integriranim selektorima (`SelectFromModel` s L1-LR i RF, `RFECV` s L1-LR).

4. **Unakrsna validacija** – `RepeatedStratifiedKFold` s 10 preklopa i 5 ponavljanja (50 iteracija po modelu).

5. **Evaluacija modela** – Za svaki preklop računaju se: Accuracy, F1, AUC-ROC, Precision, Recall, MCC i matrica zabune. Također se prikupljaju važnosti značajki.

6. **Statistička usporedba modela** – Friedmanov test, post-hoc Nemenyi i Conover testovi, te Pairwise Wilcoxon test s Holm korekcijom za AUC-ROC i F1 metrike.

7. **Generiranje rezultata** – ROC krivulje (pojedinačne, grupne, kombinirane) te spremanje u TXT, CSV i PNG formatima.

Sljedeći isječak koda prikazuje kako izgleda evaluacijska petlja jednog preklopa:

```
for fold, (train_idx, test_idx) in enumerate(splits, 1):
    X_train, X_test = X.iloc[train_idx], X.iloc[test_idx]
    y_train, y_test = y.iloc[train_idx], y.iloc[test_idx]

    if clf_info.get('selector') == 'integrated':
        model_fitted = model.fit(X_train, y_train)
        y_pred = model_fitted.predict(X_test)
        proba = model_fitted.predict_proba(X_test)
        pos_idx = list(model_fitted.classes_).index('p')
        y_proba = proba[:, pos_idx]
    else:
        X_train_sel = X_train[selected_features]
        X_test_sel = X_test[selected_features]
        model_fitted = model.fit(X_train_sel, y_train)
        y_pred = model_fitted.predict(X_test_sel)
        ...

    fold_metrics = calculate_metrics(y_test, y_pred, y_proba)
```

Listing 1: Evaluacija modela unutar cross-validation petlje

2.3 Usporedba modela – Baseline (sve značajke)

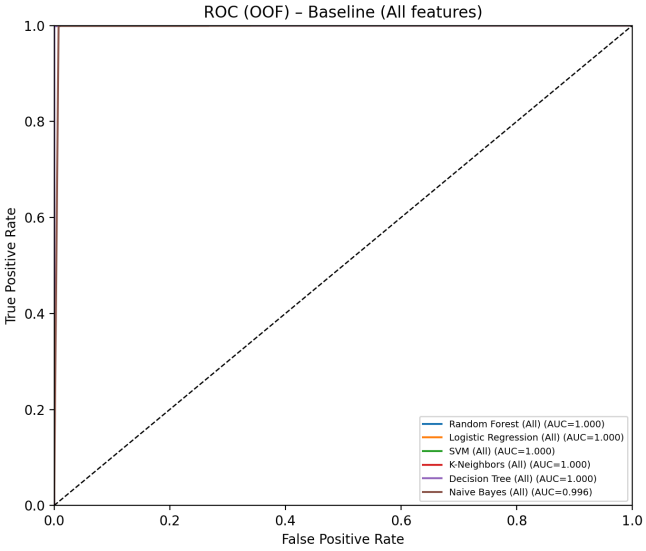
Tablica 2 prikazuje performanse svih 6 klasifikatora treniranih na kompletnom skupu značajki.

Tablica 2: Performanse baseline modela (sve značajke, 10-fold CV × 5 ponavljanja)

Klasifikator	**AUC-ROC**	**F1**	**Accuracy**	**Precision**	**Recall**	**MCC**
Random Forest	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Logistic Regression	1.0000	0.9995	0.9995	1.0000	0.9990	0.9990
SVM	1.0000	0.9991	0.9991	1.0000	0.9982	0.9983
K-Neighbors	1.0000	0.9990	0.9990	1.0000	0.9980	0.9981
Decision Tree	0.9990	0.9986	0.9987	0.9988	0.9985	0.9974
Naive Bayes	0.9945	0.9539	0.9534	0.9282	0.9810	0.9081

Random Forest postiže savršen rezultat (AUC-ROC = 1.0, F1 = 1.0, Accuracy = 1.0), dok Logistic Regression, SVM i K-Neighbors postižu gotovo savršene rezultate ($> 99.9\%$).

Decision Tree pokazuje minimalno niže performanse, dok Naive Bayes zaostaje s F1 od 0.954.



Slika 1: ROC krivulje za baseline modele trenirane na svim značajkama

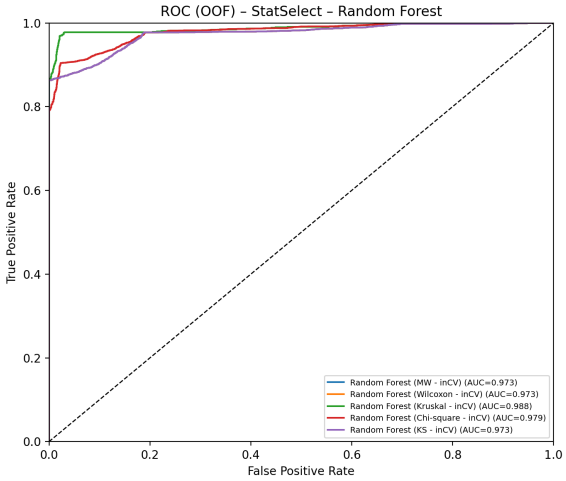
2.4 Usporedba modela – Statistička selekcija značajki (in-CV)

Tablica 3 prikazuje performanse klasifikatora u kombinaciji s različitim statističkim testovima za selekciju značajki. Prikazani su prosječni AUC-ROC, F1 i Accuracy. Selekcija se izvodi in-CV (unutar svakog preklopa) na top 5 značajki.

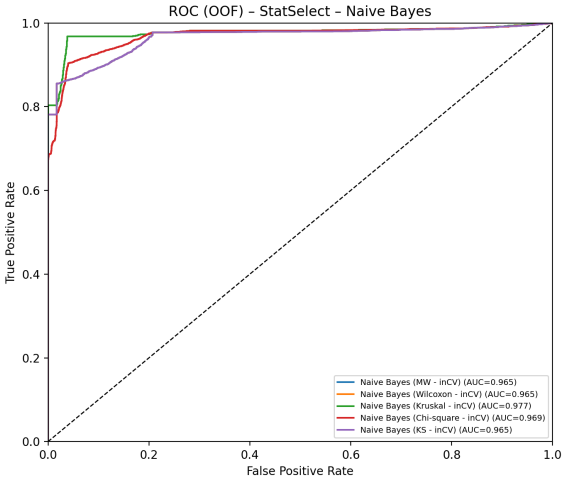
Tablica 3: Performanse modela sa statističkom selekcijom značajki (top 5, in-CV)

Klasifikator	**Stat. test**	**AUC-ROC**	**F1**	**Accuracy**
Random Forest	MW	0.9784	0.9265	0.9340
	Wilcoxon	0.9784	0.9265	0.9340
	Kruskal	0.9831	0.9456	0.9506
	Chi-sq.	0.9796	0.9374	0.9419
	KS	0.9784	0.9265	0.9340
Log. Regression	MW	0.9732	0.9218	0.9301
	Wilcoxon	0.9732	0.9218	0.9301
	Kruskal	0.9784	0.9437	0.9490
	Chi-sq.	0.9755	0.9374	0.9419
	KS	0.9732	0.9218	0.9301
SVM	MW	0.9504	0.9265	0.9340
	Wilcoxon	0.9504	0.9265	0.9340
	Kruskal	0.9658	0.9456	0.9506
	Chi-sq.	0.9623	0.9374	0.9419
	KS	0.9504	0.9265	0.9340
K-Neighbors	MW	0.9758	0.9265	0.9340
	Wilcoxon	0.9758	0.9265	0.9340
	Kruskal	0.9796	0.9435	0.9488
	Chi-sq.	0.9490	0.9335	0.9388
	KS	0.9758	0.9265	0.9340
Decision Tree	MW	0.9784	0.9265	0.9340
	Wilcoxon	0.9784	0.9265	0.9340
	Kruskal	0.9831	0.9456	0.9506
	Chi-sq.	0.9796	0.9374	0.9419
	KS	0.9784	0.9265	0.9340
Naive Bayes	MW	0.9685	0.8771	0.8946
	Wilcoxon	0.9685	0.8771	0.8946
	Kruskal	0.9742	0.8798	0.8967
	Chi-sq.	0.9726	0.9286	0.9330
	KS	0.9685	0.8771	0.8946

Iz tablice je vidljivo da redukcija na samo 5 značajki očekivano smanjuje performanse u usporedbi s baseline modelima. Ipak, modeli i dalje postižu visoke AUC-ROC vrijednosti (> 0.95), što potvrđuje da odabrane značajke nose velik dio diskriminacijske informacije. Kruskal-Wallis test konzistentno daje nešto bolje rezultate od ostalih testova. Mann-Whitney, Wilcoxon (rank-sum) i KS testovi odabiru identične značajke te stoga daju identične rezultate.



(a) Random Forest



(b) Naive Bayes

Slika 2: ROC krivulje za odabrane klasifikatore sa statističkom selekcijom značajki (top 5, in-CV). Prikazani su Random Forest (najbolji) i Naive Bayes (najlošiji) kao reprezentativni primjeri.

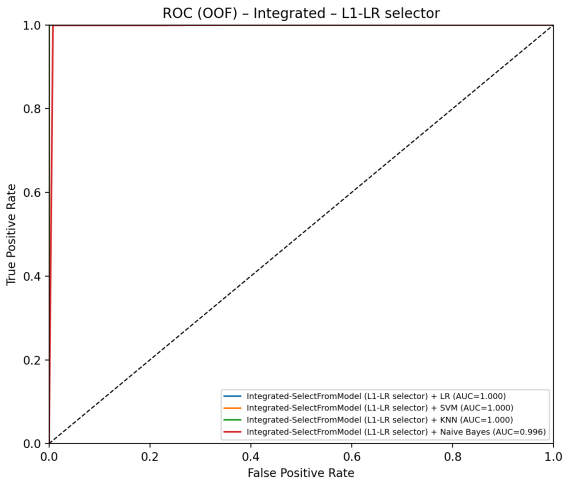
2.5 Usporedba modela – Integrirani selektori

Tablica 4 prikazuje performanse modela koji koriste integrirane metode selekcije značajki.

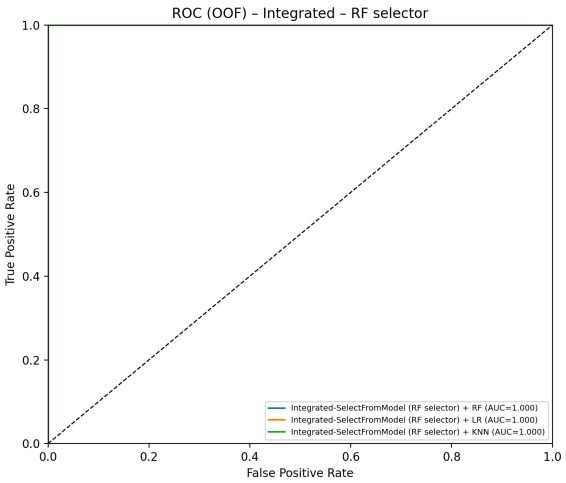
Tablica 4: Performanse modela s integriranim selektorima značajki

Konfiguracija	**AUC-ROC**	**F1**	**Accuracy**	**#Feat.**
SelectFromModel (L1-LR selektor):				
+ LR	1.0000	1.0000	1.0000	117
+ SVM	1.0000	0.9991	0.9991	117
+ KNN	1.0000	0.9990	0.9990	117
+ Naive Bayes	0.9945	0.9539	0.9534	117
SelectFromModel (RF selektor):				
+ RF	1.0000	1.0000	1.0000	59
+ LR	1.0000	1.0000	1.0000	59
+ KNN	1.0000	0.9990	0.9990	59
RFECV (L1-LR):				
+ LR	1.0000	0.9995	0.9995	29
+ KNN	1.0000	0.9990	0.9991	29
+ Naive Bayes	0.9969	0.9814	0.9811	29

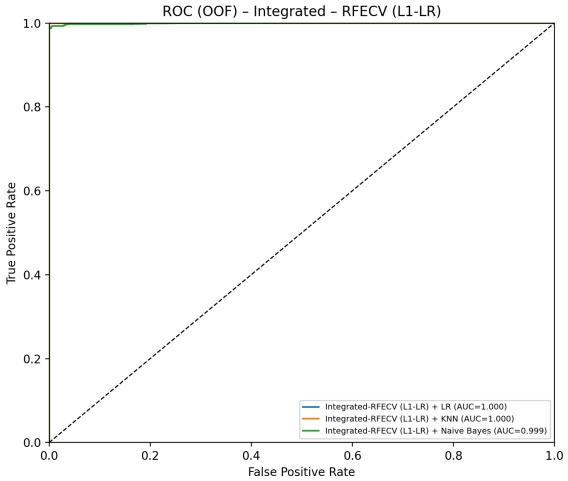
Ključno zapažanje je da **SelectFromModel (RF selektor)** postiže savršene performanse (AUC-ROC = 1.0, F1 = 1.0) koristeći samo 59 od 117 značajki, što predstavlja redukciju od gotovo 50%. Čak i **RFECV** s L1-LR, koji odabire svega 29 značajki (75% redukcija), postiže AUC-ROC od 1.0 s Logističkom regresijom.



(a) SelectFromModel (L1-LR)



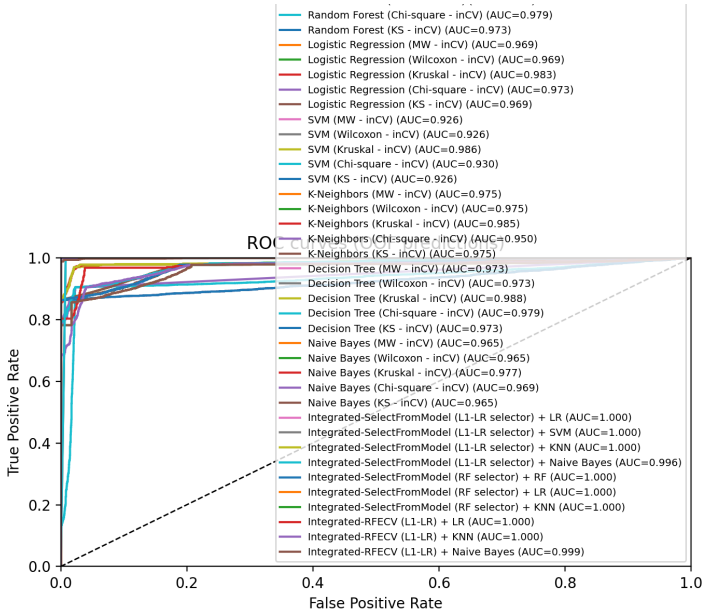
(b) SelectFromModel (RF)



(c) RFECV (L1-LR)

Slika 3: ROC krivulje za modele s integriranim selektorima značajki

2.6 Ukupna usporedba svih modela



Slika 4: ROC krivulje za sve evaluirane modele

2.7 Selekcija značajki – metode i rezultati

U projektu su korištene dvije kategorije metoda za selekciju značajki: **statističke metode** i **integrirane (model-bazirane) metode**.

2.7.1 Statističke metode selekcije

Statističke metode testiraju razlike u distribucijama svake značajke između dviju klasa (jestiva/otrovna):

1. **Mann-Whitney U test** – neparametarski test koji uspoređuje rangove dviju nezavisnih skupina. Veličina učinka je rank-biserijska korelacija: $r_{rb} = 1 - \frac{2U}{n_1 \cdot n_2}$. Ovaj test je robustan na ne-normalnost distribucija.

2. **Wilcoxon rank-sum test** – u ovoj implementaciji je alias za Mann-Whitney U test, što znači da koristi identičnu formulaciju za nezavisne uzorke i daje iste rezultate.

3. **Kruskal-Wallis test** – generalizacija Mann-Whitney testa za usporedbu dviju ili više skupina putem rangova. Veličina učinka se računa kao $\eta^2 = \frac{H-1}{N-1}$.

4. **Chi-square test (χ^2)** – testira nezavisnost kategorijskih varijabli putem kontingencijske tablice. Veličina učinka je Cramér-ov V: $V = \sqrt{\frac{\chi^2}{N \cdot (\min(r,c)-1)}}$.

5. **Kolmogorov-Smirnov test (KS)** – mjeri maksimalnu razliku između kumulativnih funkcija distribucije dviju skupina. KS statistika sama po sebi služi kao mjera veličine učinka.

Za sve testove primjenjuje se **Benjamini-Hochberg korekcija** za kontrolu stope lažnih otkrića (FDR) pri višestrukom testiranju, uz prag statističke značajnosti $\alpha = 0.05$ i minimalni zahtjev za veličinu učinka ≥ 0.10.

2.7.2 Integrirane metode selekcije

1. **SelectFromModel (L1-LR)** – koristi Logističku regresiju s L1 (Lasso) regularizacijom kao selektor. L1 penalizacija prisiljava koeficijente manje važnih značajki na nulu, čime se vrši implicitna selekcija. Prag je postavljen na medijan apsolutnih koeficijenata, što rezultira odabirom oko 117 značajki (50% originalnih).

2. **SelectFromModel (RF)** – koristi Random Forest klasifikator za procjenu važnosti značajki temeljem smanjenja nečistoće (Gini importance). S pragom na medijanu, odabire oko 59 značajki.

3. **RFECV (Recursive Feature Elimination with Cross-Validation)** – iterativno uklanja najmanje važne značajke i koristi unakrsnu validaciju za pronalaženje optimalnog broja značajki. U ovom projektu koristi L1-LR kao bazni estimator i automatski konvergira na oko 29 značajki.

2.7.3 Usporedba odabranih značajki

Tablica 5 prikazuje top 5 značajki odabranih svakim statističkim testom.

Tablica 5: Top 5 značajki odabranih statističkim testovima

Rang	Mann-Whitney	Wilcoxon	Kruskal	Chi-square	KS
1	odor_n	odor_n	odor_n	odor_n	odor_n
2	odor_f	odor_f	odor_f	stalk-surf.-a._k	odor_f
3	ring-type_p	ring-type_p	stalk-surf.-a._k	stalk-surf.-b._k	ring-type_p
4	stalk-surf.-a._k	stalk-surf.-a._k	stalk-surf.-b._k	ring-type_p	stalk-surf.-a._k
5	stalk-surf.-b._k	stalk-surf.-b._k	ring-type_p	gill-size_b	stalk-surf.-b._k

Ključna zapažanja: **odor_n** (miris = "none") je najvažnija značajka prema svim testovima, a **odor_f** (neugodan miris) druga prema 4 od 5 testova. Mann-Whitney, Wilcoxon i KS testovi odabiru identične značajke jer koriste ekvivalentne formulacije. Chi-square test jedini odabire `gill-size_b` umjesto `odor_f`.

Integrirani selektori zadržavaju više značajki (L1-LR: 117, RF: 59, RFECV: 29), a među najvažnijima prema koeficijentima i importancima dominiraju značajke mirisa (`odor_n/f/l/a`), boja otiska spora i površina stabljike. **Miris (odor)** se potvrđuje kao najdiskriminativnija grupa značajki neovisno o metodi selekcije.

2.8 Top 10 značajki po važnosti – Random Forest baseline

Tablica 6 prikazuje 10 najvažnijih značajki prema prosječnoj Gini importanci iz Random Forest klasifikatora treniranog na svim značajkama (prosjek preko 50 CV iteracija).

Tablica 6: Top 10 značajki prema Random Forest importanci (Gini, prosjek 50 iteracija)

Rang	Značajka	Prosječna importanca
1	odor_n	0.1393
2	odor_f	0.0773
3	gill-size_n	0.0652
4	gill-size_b	0.0535
5	stalk-surface-below-ring_k	0.0519
6	stalk-surface-above-ring_k	0.0455
7	gill-color_b	0.0430
8	spore-print-color_h	0.0376
9	ring-type_p	0.0302
10	bruises_f	0.0270

Primjećuje se da `odor_n` (miris = "none") ima gotovo dvostruko veću importancu od sljedeće značajke (`odor_f`), što potvrđuje dominantnu ulogu mirisa u klasifikaciji. Značajke vezane uz veličinu lamela (`gill-size`), površinu stabljike i boju otiska spora također igraju važnu ulogu.

2.9 Statistička usporedba modela – Friedmanov test i rangovi

Za rigoroznu usporedbu svih 46 konfiguracija modela provedeni su Friedmanov test i post-hoc Nemenyi analiza prosječnih rangova. Friedmanov test testira nultu hipotezu da su svi modeli jednako dobri.

2.9.1 Friedmanov test – AUC-ROC

Friedmanov test za AUC-ROC metriku dao je testnu statistiku $\chi_F^2 = 2065.67$ s p-vrijednošću $p \approx 0$ (numerički nula), što znači da se **svi modeli statistički značajno razlikuju** u performansama mjerenim AUC-ROC metrikom.

Tablica 7 prikazuje top 10 modela prema prosječnom Nemenyi rangu (niži rang = bolji model).

Tablica 7: Top 10 modela prema prosječnom Nemenyi rangu – AUC-ROC

Rang	Model	Prosj. rang
1	SVM (All)	6.42
1	Integrated-SFM (RF) + KNN	6.42
1	Integrated-SFM (L1-LR) + LR	6.42
1	Integrated-SFM (L1-LR) + SVM	6.42
5	Integrated-SFM (RF) + RF	6.76
6	Integrated-RFECV (L1-LR) + KNN	6.81
7	K-Neighbors (All)	6.94
8	Logistic Regression (All)	7.06
9	Integrated-RFECV (L1-LR) + LR	7.12
10	Random Forest (All)	7.52

SVM (All) dijeli prvi prosječni rang (6.42) s tri konfiguracije integriranih selektora. Svi baseline i integrirani modeli nalaze se u top 16, dok **svi statistički selektirani modeli**

(inCV) imaju rangove > 21.

2.9.2 Friedmanov test – F1

Friedmanov test za F1 metriku dao je testnu statistiku $\chi_F^2 = 1984.43$ s $p \approx 0$, potvrđujući statistički značajne razlike među modelima.

Tablica 8 prikazuje top 10 modela prema prosječnom Nemenyi rangu za F1.

Tablica 8: Top 10 modela prema prosječnom Nemenyi rangu – F1

Rang	Model	Prosj. rang
1	Random Forest (All)	4.92
1	Integrated-SFM (RF) + LR	4.92
1	Integrated-SFM (RF) + RF	4.92
1	Integrated-SFM (L1-LR) + LR	4.92
5	Integrated-RFECV (L1-LR) + LR	6.50
6	Logistic Regression (All)	6.71
7	Decision Tree (All)	8.11
8	Integrated-RFECV (L1-LR) + KNN	8.27
9	Integrated-SFM (L1-LR) + SVM	8.29
9	SVM (All)	8.29

Prema F1 rangu, **Random Forest (All)** dijeli prvi rang (4.92) s tri integrirane konfiguracije – sve postižu savršen F1 = 1.0. Svi statistički selektirani modeli (inCV) imaju rangove > 23.

2.9.3 Ključna zapažanja iz statističke analize

Pairwise Wilcoxon testovi s Holm korekcijom potvrđuju da su razlike između baseline/integriranih i statistički selektiranih modela **statistički značajne** ($p_{\text{adj}} < 10^{-6}$). Razlike između top baseline modela (RF, LR, SVM, KNN) nisu značajne za AUC-ROC ($p_{\text{adj}} > 0.4$), ali jesu za F1 ($p_{\text{adj}} < 0.05$ za RF vs. SVM/KNN). Naive Bayes konzistentno zauzima najniže rangove, dok integrirani selektori uspješno zadržavaju performanse bliske baseline modelima uz značajnu redukciju dimenzionalnosti.

3 Zaključak

Provedeni eksperimenti jasno pokazuju nekoliko ključnih zaključaka:

1. **Random Forest je najuspješniji klasifikator** za ovaj skup podataka, postižući savršenu klasifikacijsku točnost (AUC-ROC = 1.0, F1 = 1.0, Accuracy = 1.0) na svim 50 iteracija unakrsne validacije. Logistic Regression, SVM i K-Neighbors također postižu izuzetno visoke performanse ($> 99.9\%$), dok Decision Tree i Naive Bayes pokazuju nešto niže, ali još uvijek visoke rezultate.

2. **Selekcija značajki značajno reducira dimenzionalnost bez velikog gubitka performansi.** RFECV s L1-LR odabire samo 29 od 117 značajki (75% redukcija) uz zadržavanje AUC-ROC od 1.0. Čak i agresivna redukcija na samo 5 značajki zadržava AUC-ROC iznad 0.95, što govori o visokoj informativnosti ključnih značajki.

3. **Miris gljive je najvažniji prediktor jestivosti.** Značajke vezane uz miris (`odor_n`, `odor_f`) konzistentno su rangirane kao najvažnije od strane svih statističkih testova i integriranih selektora. Ovo je biološki potpuno razumno jer mirisni profil gljiva često izravno korelira s prisutnosti toksina.

4. **Statističke metode selekcije su konzistentne.** Četiri od pet testova (Mann-Whitney, Wilcoxon, KS, Kruskal-Wallis) odabiru gotovo identičan skup značajki, što potvrđuje robusnost odabira. Chi-square test pokazuje minimalna odstupanja koja proizlaze iz prirode testa.

3.1 Primjena u stvarnom svijetu

Rezultati imaju praktičnu primjenu: model s 5 ključnih značajki (AUC-ROC > 0.97) omogućuje razvoj **mobilnih aplikacija** za procjenu jestivosti gljiva temeljem nekoliko pitanja. Saznanje da su miris, tip prstena i površina stabljike najdiskriminativniji može služiti u **edukaciji** neiskusnih skupljača. Metodološki, usporedba 8 metoda selekcije i 6 klasifikatora u rigoroznom postavu (50 iteracija CV, Friedmanov test) primjenjiva je na druge probleme binarne klasifikacije.

Napomena: Iako su rezultati izuzetno visoki, valja napomenuti da je korišteni skup podataka relativno čist i dobro strukturiran. U realnim uvjetima, klasifikacija gljiva na temelju vizualnih i olfaktornih karakteristika može biti složenija zbog varijabilnosti unutar vrsta, utjecaja okolišnih čimbenika i subjektivnosti procjene pojedinih značajki.