**SVEUČILISTE U SPLITU**

**PRIRODOSLOVNO-MATEMATIČKI FAKULTET**

SEMINARSKI RAD

**PROJEKT IZ UVODA U UMJETNU INTELIGENCIJU**

**„KLASIFIKACIJA SLIKA PET BIOMA“**

Profesor: Student:

Saša Mladenović Ivona Ercegovac

Split, siječanj 2024.

Sadržaj

[1 UVOD 1](#_Toc157610251)

[1.1 Motivacija 1](#_Toc157610252)

[1.2 Općenito 1](#_Toc157610253)

[2 PRIPREMA 2](#_Toc157610254)

[2.1 Opis modela 2](#_Toc157610255)

[2.2 Orange 2](#_Toc157610256)

[3 PODATCI 3](#_Toc157610257)

[3.1 Učitavanje slika 3](#_Toc157610258)

[3.3 Ugradnja slika 4](#_Toc157610259)

[4 KLASIFIKACIJA 5](#_Toc157610260)

[4.1 Udaljenosti 5](#_Toc157610261)

[4.2 Hijerarhijsko klasteriranje 5](#_Toc157610262)

[4.3 Linearna projekcija 6](#_Toc157610263)

[5 EVALUACIJA MODELA 8](#_Toc157610264)

[5.1 Logistička regresija, Testiranje i ocjenjivanje 8](#_Toc157610265)

[5.2 Matrica konfuzije 9](#_Toc157610266)

[6 ZAKLJUČAK 11](#_Toc157610267)

[7 LITERATURA 12](#_Toc157610268)

[8 SAŽETAK 13](#_Toc157610269)

# UVOD

## Motivacija

Motivacija iza istraživanja strojnog učenja, točnije koncepta klasifikacije, leži u praktičnim prednostima za jednog razvojnog programera. Kategorizacija slika u različite klase omogućuje pojednostavljenu obradu kompleksnih zadataka, štedeći pritom vrijeme i resurse. Drugim riječima, ova tehnologija poboljšava organizaciju podataka, što je korisno za programere, posebice one koji rade na projektima s vizualnim elementima.

## Općenito

Biomi predstavljaju velika područja okarakterizirana vegetacijom, tlom, klimom i životinjskim svijetom. Možemo pojednostavniti opis te reći da svaki biom ima svoj specifični izgled utemeljen na određenim oblicima. Glavnih pet su vodeni, travnjački, šumski, pustinjski i tundra. Uzevši za primjer pustinjski biom već možemo zamisliti značajke slike koja bi ga predstavljale. Naime, usredotočili bi se na oblik deva, kaktusa i palmi, te na glatku površinu pijeska. Danas imamo vrlo pristupačan način izrade raznih modela, baziranih na umjetnoj inteligenciji, koji „razmišljaju“ na upravo naveden način i efikasno rade posao umjesto nas. U ovom seminaru bavit ćemo se pretežito konceptom klasifikacije – najčešće korištene nadzirane metode strojnog učenja koja iz svojstava dostupnih podataka pokušava naučiti optimalnu podjelu istih na zadane grupe. Poetično rečeno, svrha jest „pustiti podatke da govore sami za sebe“ (Šnajder, 2020./2021.).

Cilj je ovog projekta objasniti teoriju strojnog učenja, s indikatorom na klasifikaciju, na primjeru modela koji će prepoznavati i klasificirati slike zavičaja u njihove predviđene biome. Prije svega potrebno je osmisliti i realizirati prikladan model. Za potrebe izrade istog, koristit ćemo se softverom Orange, alatom za vizualizaciju i analizu podataka, kako bi štoviše intuitivnim sadržajem približili ideju čitatelju.

# PRIPREMA

## Opis modela

Ranije smo istakli kako je potrebno realizirati model koji mora razvrstavati određene podatke u odgovarajuće klase. Njegov zadatak nije pretjerano zahtjevan, ali se ipak sastoji od nekoliko procesa koje je neophodno razumjeti i primijeniti postepeno. Adekvatan model trebao bi okvirno moći: učitati slike, ugraditi slike, izračunati udaljenosti između ugrađenih slika, izvršiti hijerarhijsko klasteriranje, primijeniti logističku regresiju, testirati i ocijeniti svoju učinkovitost, te prikazati matricu konfuzije. Kroz sve navedene postupke proći ćemo promatrajući kontinuirano naš model u okruženju Orange sustava. Konstrukcija takvog modela temelji se na pažljivoj pripremi podataka, selekciji widget-a u programu, te preciznom podešavanju parametara kako bi postigao visoku točnost klasifikacije slika. Kroz procese evaluacije, uključujući analizu matrice konfuzije, model se ocjenjuje i prilagođava kako bi osigurao pouzdanu klasifikaciju.

## Orange

Orange je besplatan softver kojim se koristimo kao alatom za vizualizaciju i upravljanje podacima kod strojnog učenja. Svrha je korisnicima znatno poboljšati iskustvo tijekom kreiranja i analiziranja modela. Njegovo sučelje olakšava kompleksne zadatke ponudom raznolikog skupa widget-a, tj. opcija pomoću kojih kreiramo vizualni model na platnu. Prilagodljivost sučelja i opsežan raspon funkcionalnosti Orange softvera čine ga optimalnim izborom za naš projekt iz razloga što osiguravaju uspješan model, te vrlo intuitivan popratni materijal kojim ćemo se služiti kao podlogom za razumijevanje same teorije modela.

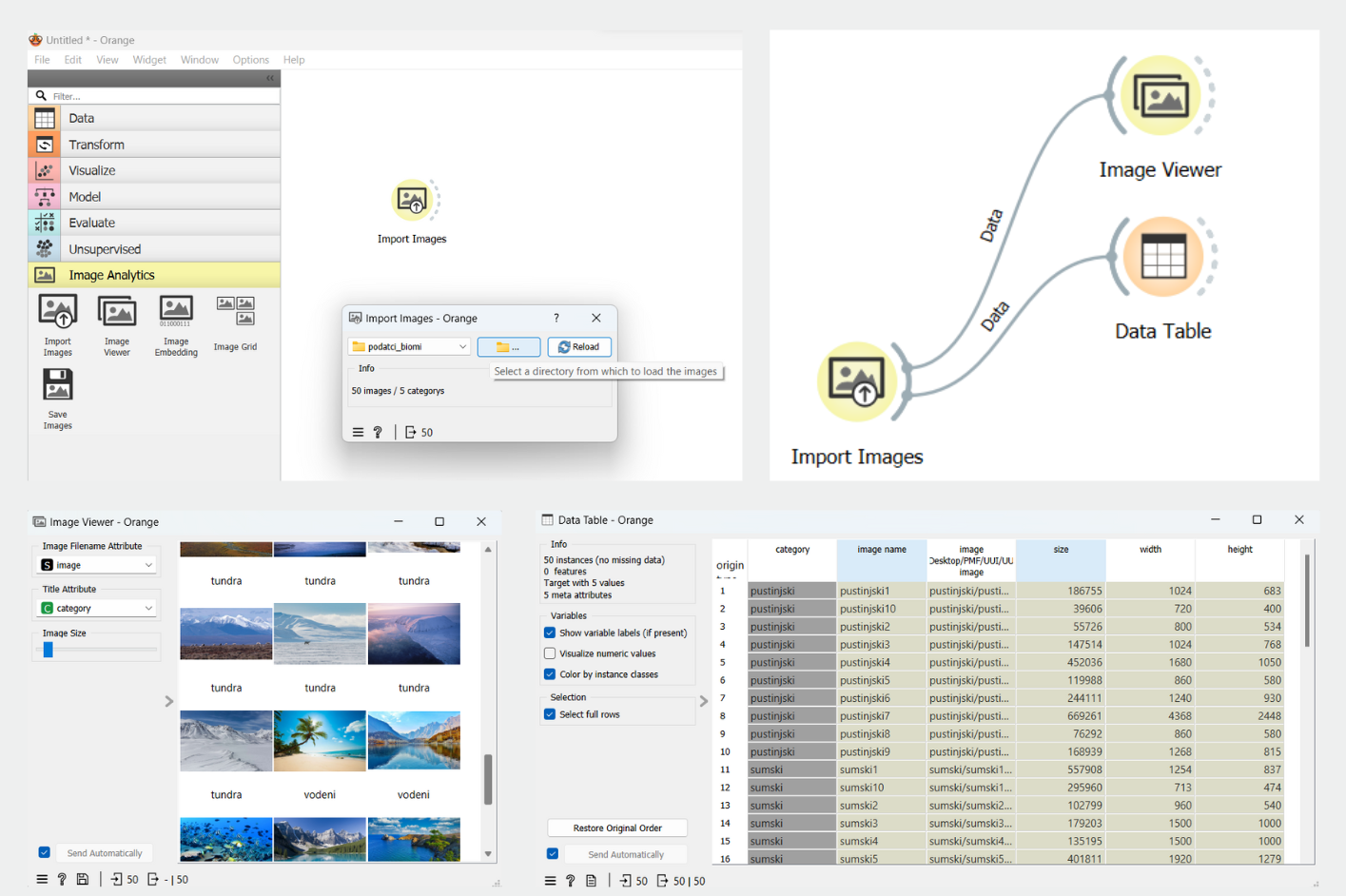
U nastavku su priloženi widget-i koje ćemo koristiti za naš model.

1. Učitavanje Slika (*Import Images*)
2. Pregled Slika (*Image Viewer*)
3. Tablica Podataka (*Data Table*)
4. Ugradnja Slika (*Image Embedding*)
5. Udaljenosti (*Distances*)
6. Hijerarhijsko Klasteriranje (*Hierarchial Clustering*)
7. Linearna Projekcija (*Linear Projection*)
8. Logistička Regresija (*Logistic Regression*)
9. Testiranje i Ocjenjivanje (*Test and Score*)
10. Matrica Konfuzije (*Confusion Matrix*)

# PODATCI

## Učitavanje slika

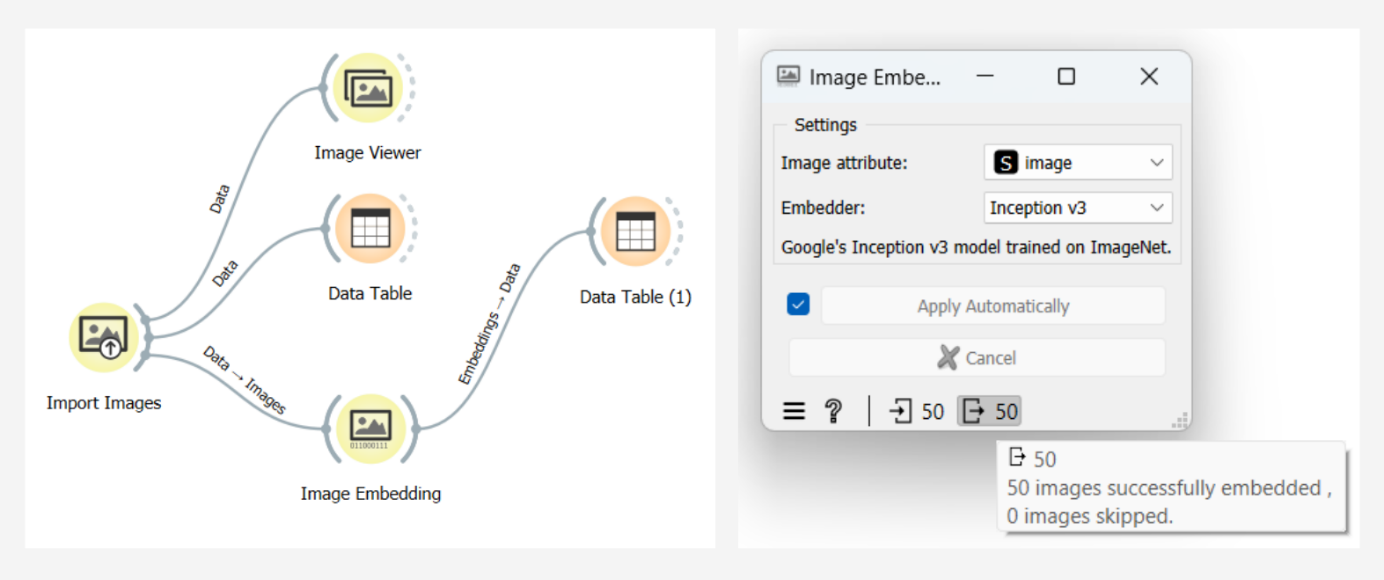
Učitajmo prije svega slike koje smo prethodno preuzeli s interneta. Bitno je napomenuti da mapa s podatcima treba obuhvaćati broj mapa koliko je u našem slučaju bioma. Ovim postupkom poručujemo programu da će morati klasificirati slike u pet zadanih klasa. Nadalje, kako bismo uopće mogli predstaviti slike kao podatke moramo instalirati dodatak programu (*add-on*) koji sadržava sve widget-e vezane isključivo za slikovni sadržaj. Jednom kada smo to učinili, lijevo u izborniku samo kliknemo na widget „Import Images“, te će se on pokazati na platnu. Dvoklikom na widget otvara nam se prozor gdje zatim učitamo mapu. Nakon toga možemo dodati widget-e „Image Viewer“ i „Data Table“ s kojima ćemo dobiti uvid u naše podatke.



Slika 1: Učitavanje i prikaz podataka

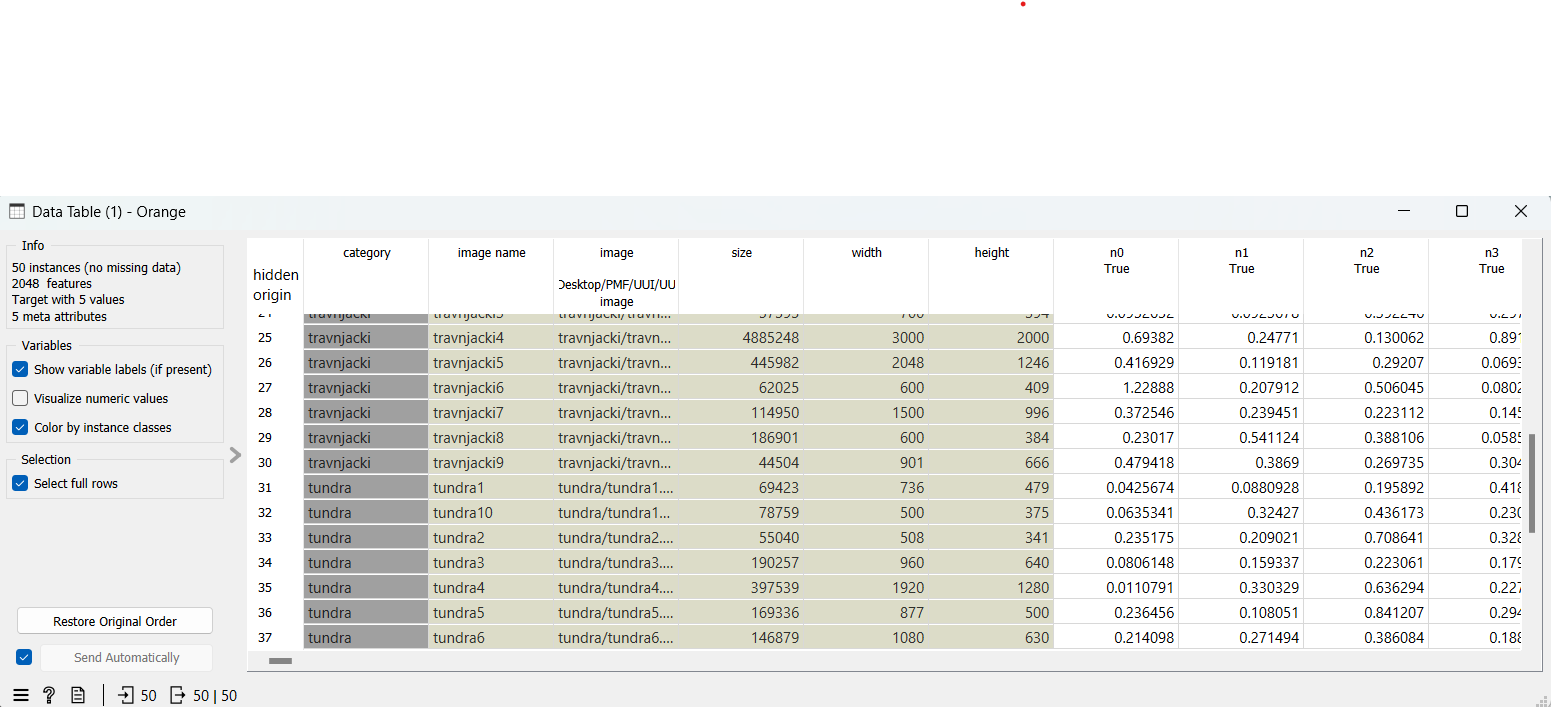
## Ugradnja slika

Znamo da model mora razlikovati slike po njihovim karakteristikama. Premda smo ranije utvrdili da je njegovo „razmišljanje“ identično našem, ipak moramo razumjeti da model nema oči, već se oslanja isključivo na podatke. Ugradnjom slike dobivamo upravo tražene informacije, odnosno izračunate vektore njenih obilježja. Ti su nam podatci ključan parametar pri mjerenju međusobne udaljenosti slika, što ćemo proučiti u idućem poglavlju. Implementirajmo widget „Image Embedding“ na importiranim slikama.



Slika 2: Ugradnja slika

Ono što zapravo ugradnja podrazumijeva u Orange softveru jest slanje slika na poslužitelj, ili procjenjivanje istih lokalno na računalu, gdje se izračunavaju prikazi vektora. Za ovaj je zadatak zaslužan tzv. „ugrađivač“ (*embedder*). U našem slučaju to je Inception v3, neuronska mreža trenirana na ImageNet skupu podataka koji sadrži preko 14 milijuna slika.



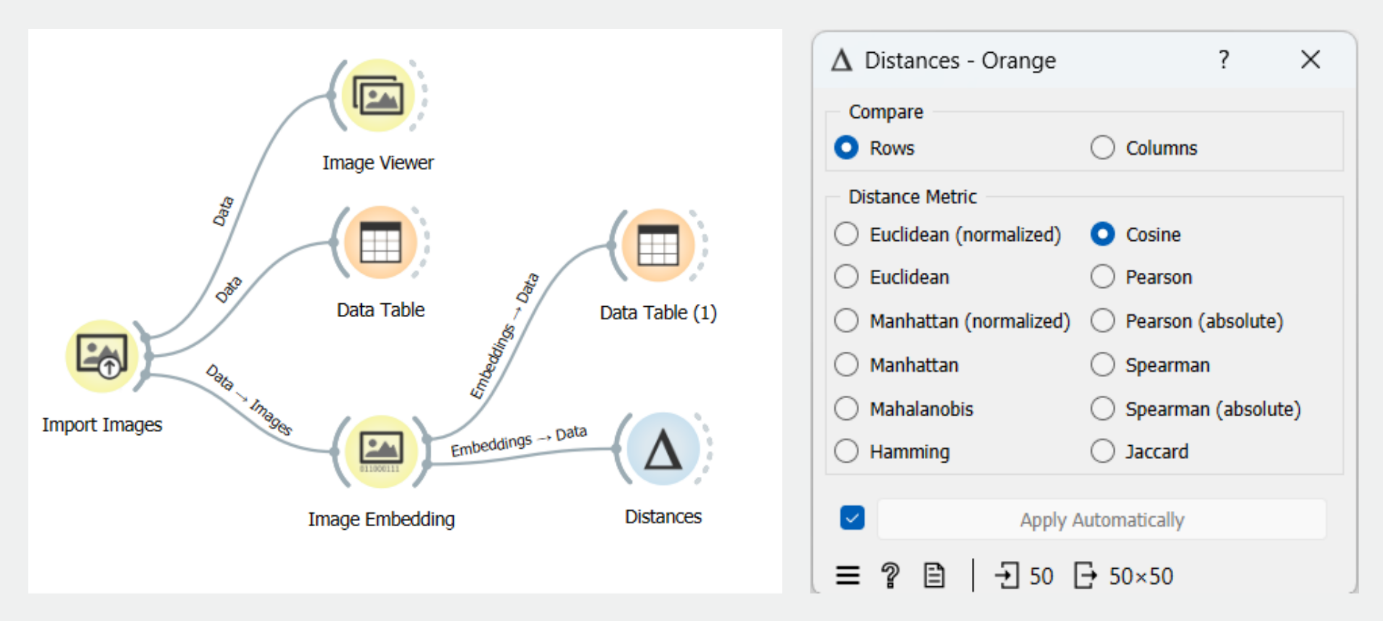
Slika 3: Prikaz podataka nakon ugradnje

Kreirajmo ponovno tablicu podataka nakon što smo izvršili ugradnju slika. Možemo vidjeti (slika 3) da za razliku od prethodne tablice (slika 1) imamo nove podatke koji predstavljaju vektorski zapis.

# KLASIFIKACIJA

## Udaljenosti

S obzirom na to da sada imamo spremne modificirane podatke, možemo računati udaljenosti. Ranije smo istakli kako je ugradnja slika ključna za izračun udaljenosti, a sada ćemo proučiti zašto je to tako. Naime, normalizacija podatka podrazumijeva skaliranje značajki što je bitan korak u procesiranju podataka za modele strojnog učenja. Trenutačni raspon podataka pretvara se u novi raspon koji poboljšava funkcionalnost prediktivnih modela na način da standardizira značajke unutar skupa podataka, što dovodi do stabilnijih i pouzdanijih rezultata. Dakle, koristit ćemo widget „Distances“ koji za ulaz uzima dobivene vektore, a vraća matricu izračunatih udaljenosti.

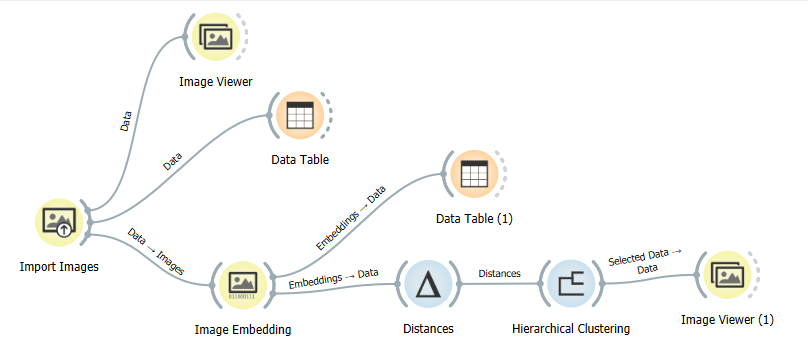


Slika 4: Izračunavanje udaljenosti

Dvoklikom na dodan widget otvara se prozor u kojem odabiremo „Cosine“ metriku udaljenosti. Ova mjera sličnosti odnosi se na udaljenost s dimenzijama koje predstavljaju značajke podatkovnog objekta u skupu podataka, drugim riječima, često je najbolja opcija za rad sa slikama. Ako je izračunata udaljenost između dviju slika manja, njihov stupanj sličnosti će biti veći.

## Hijerarhijsko klasteriranje

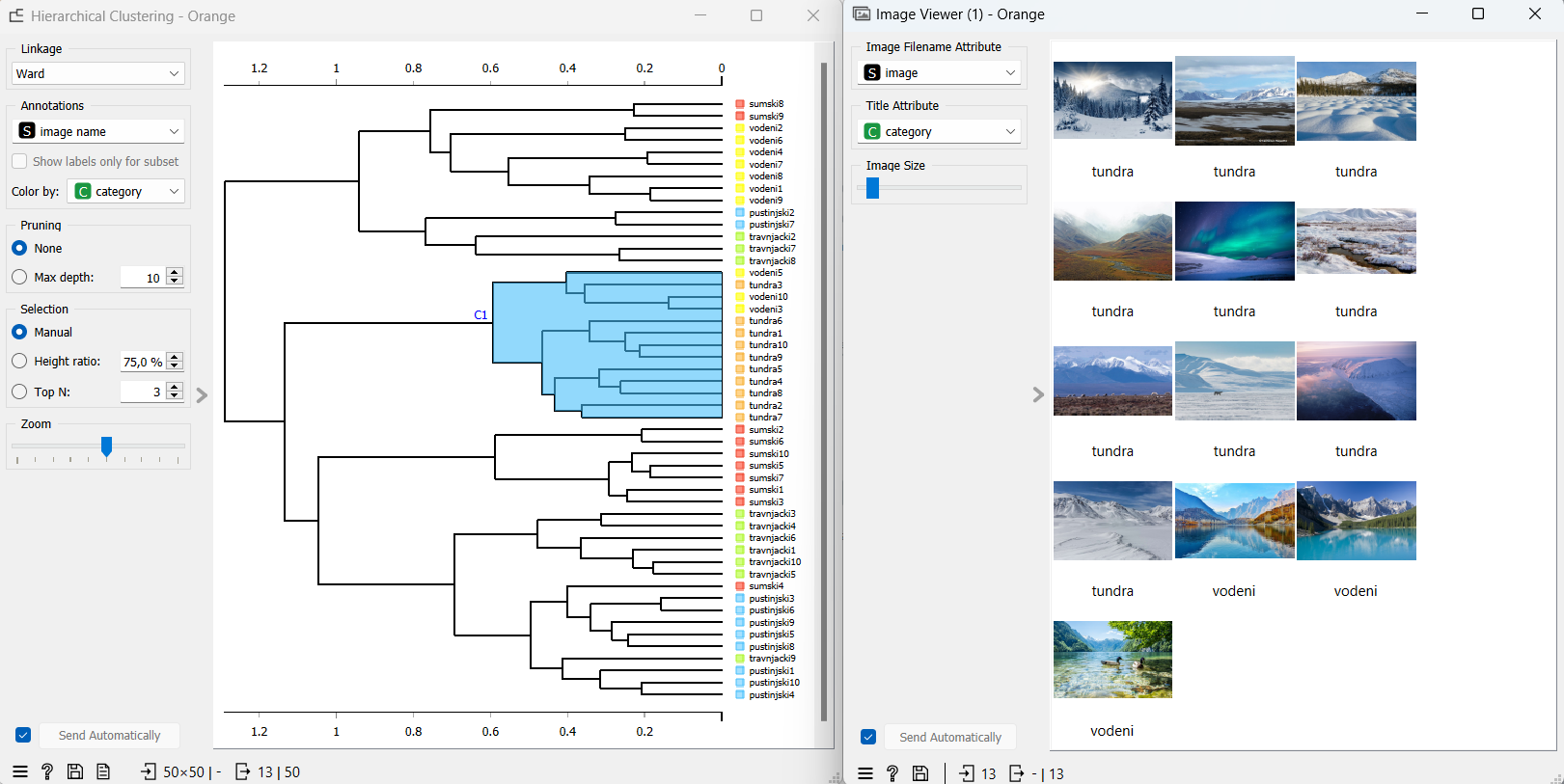
Nadalje, vrijeme je da u naš model implementiramo algoritam hijerarhijskog klasteriranja kojeg ćemo nadovezati na prethodno dodan widget. Izraz "hijerarhijsko" odnosi se na činjenicu da metoda stvara stablo klastera, često nazvano dendrogram. Ova struktura predstavlja odnose između naših podataka dobivene upravo pomoću njihovih ranije izračunatih udaljenosti.



Slika 5: Implementacija hijerarhijskog klasteriranja

Naziv hijerarhijskog *klasteriranja* dolazi od riječi klaster. Klasteri (eng. *clusters*) su grupe podataka formirane na temelju njihovih sličnosti, to jest raznolikosti s podatcima iz ostalih klastera. Takve grupe podataka formiraju se postepeno, točnije svaka slika zasebno počinje kao vlastiti klaster. Nakon toga, dva najbliža klastera spajaju se u jedan zajednički. Ovaj se proces ponavlja dok se sve slike ne svedu na samo jedan klaster.

Svaki čvor u dendogramu predstavlja jedan klaster, a grane predstavljaju spajanje ili razdvajanje klastera na različitim razinama sličnosti. Listovi na kraju stabla označavaju sve naše podatke, dakle u našem slučaju to je 50 listova.



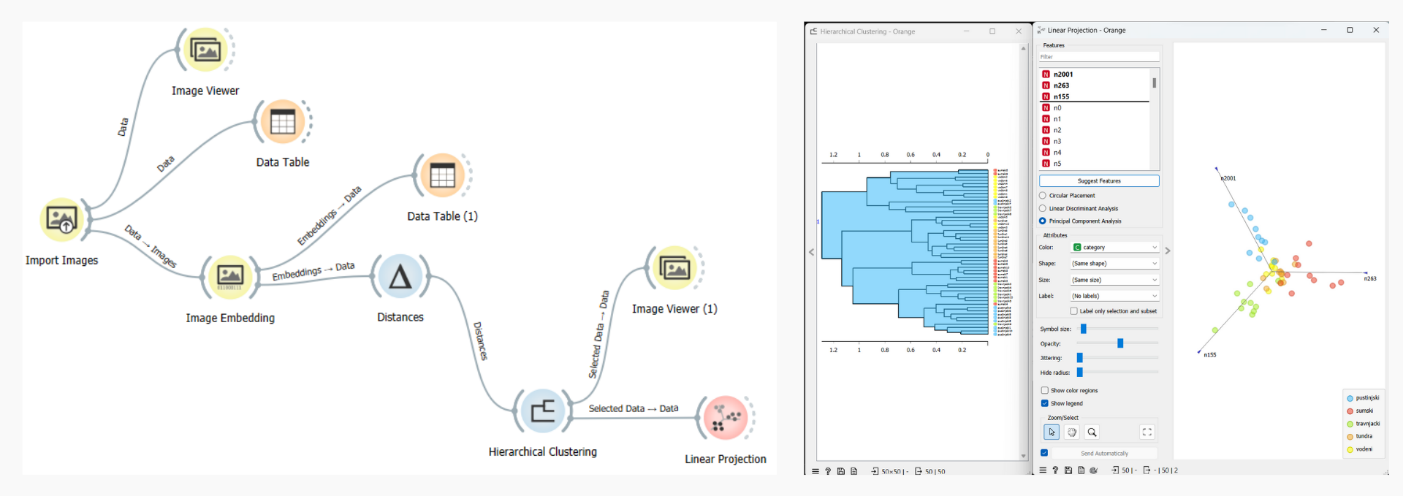
Slika 6: Prikaz stabla hijerarshijskog klasteriranja

Na primjeru (slika 6) vidimo po bojama klasa da su slike relativno dobro razvrstane, s ponekim izuzetcima koji su „zalutali“. Odaberimo neki dio stabla, te istovremeno otvorimo widget za prikaz slika koji smo dodali u nastavku modela. Primijetimo da su tri slike vodenog bioma svrstane u klaster sa slikama tundra bioma, jer unatoč različitoj pripadnosti klasama vrlo su im slična obilježja. Premda to nije poželjno, možemo prihvatiti neka razumna odsupanja poput ovog.

## Linearna projekcija

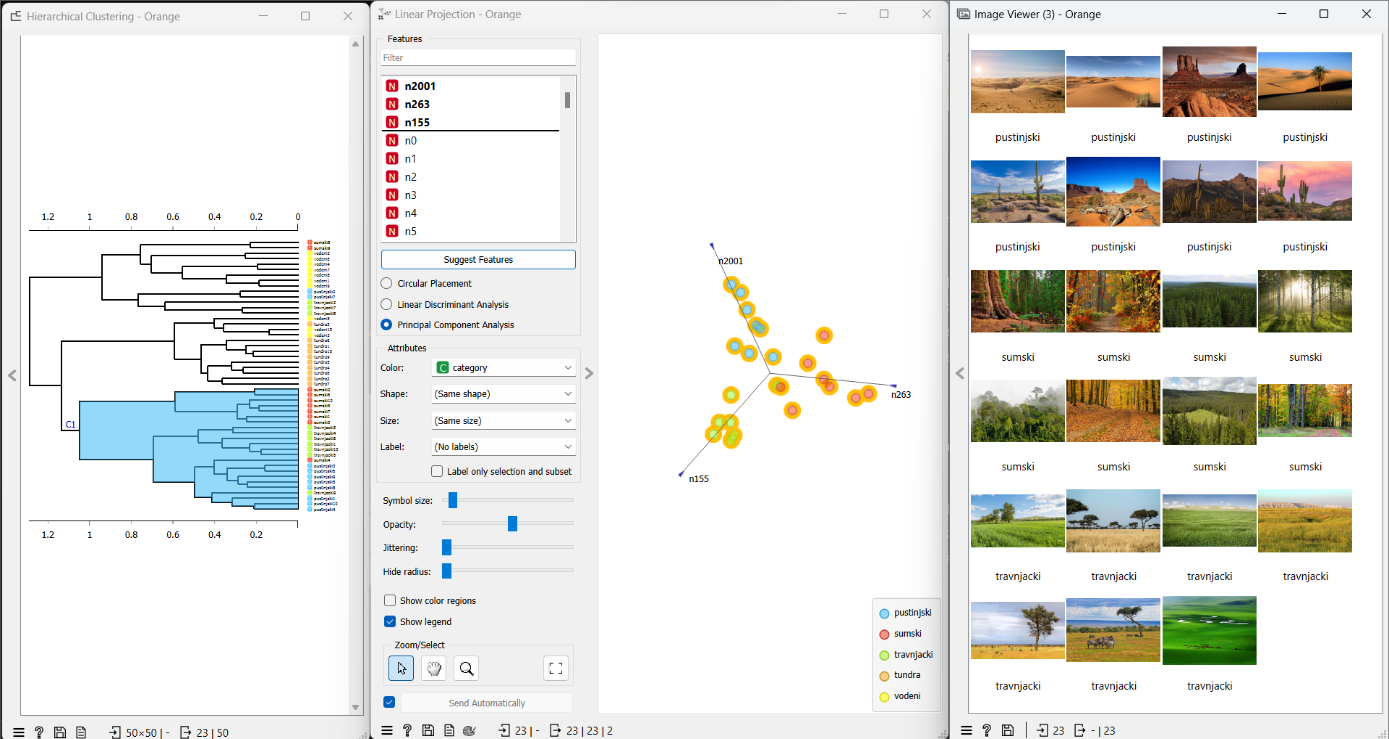
Kako bismo dobili još intuitivniji uvid u naše klasificirane podatke, možemo proširiti model s widget-om „Linear Projection“. Ovaj dodatak nam ostvaruje mogućnost da na grafu vidimo stvarnu udaljenost slika na temelju njihovih vektora.

Linearna projekcija je tehnika korištena za smanjenje dimenzionalnosti pretvarajući podatke visoke dimenzionalnosti u prostor niže dimenzionalnosti, zadržavajući pritom određene karakteristike originalnih podataka. Cilj joj je vizualizirati kompleksne odnose među podacima na pristupačniji način. U našem slučaju imamo visokodimenzionalne vektore koje projiciramo u 2D prostor.



Slika 7: Hijerarhijsko klasteriranje prikazano linearnom projekcijom, primjer 1

Nakon dodavanja novog widget-a pokrenimo istovremeno njega i prethodno izvršen algoritam hijerarhijskog klasteriranja. Držat ćemo se tri nasumično odabrana vektora radi jednostavnosti. Označimo li cijelo stablo, možemo po bojama uočiti iz slike (slika 7) da su biomi većinom odvojeni jedni od drugih. Najmanje udaljenosti se očituju kod vodenog i tundra bioma (narančaste i žute točke na grafu). Ranije smo istakli ovu situaciju, a ona se dešava upravo zbog sličnosti ovih zavičaja međusobno, a ponekad i s drugima biomima. S druge strane, one najudaljenije slike imaju najveću „originalnost“ kada su u pitanju njihove značajke.



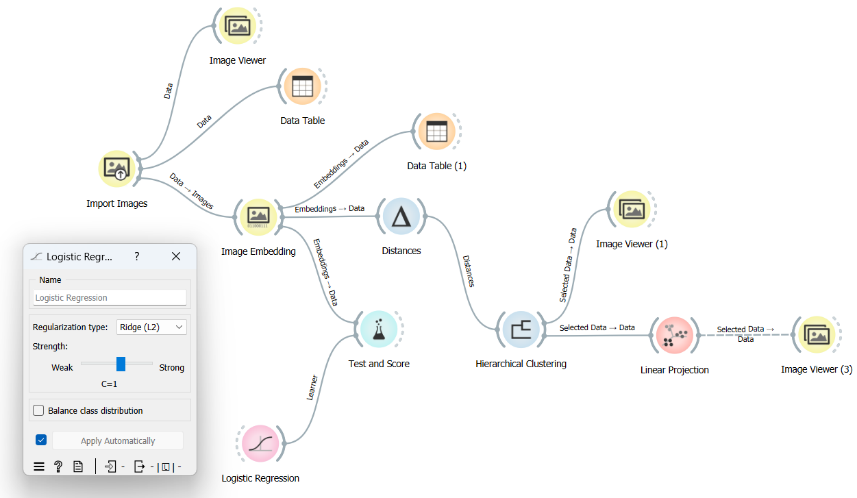
Slika 8: Hijerarhijsko klasteriranje prikazano linearnom projekcijom, primjer 2

Ako sada označimo manji dio stabla, dakle klaster slika koje bi trebale biti više međusobno slične, vidimo (slika 8) da podjela zaista izgleda bolje. Preklapanja gotovo pa nema, što znači da model odvaja biome kako bi trebao, ali je njihove slike prije stavio u isti klaster zato što su im obilježja slična kao što se da primijetiti na ponovno dodanom widget-u za prikaz slika. Očekivano je da će zavičaji ovakvih bioma prije biti kategorizirani jedni s drugima nego li s zavičajem npr. tundra bioma.

# EVALUACIJA MODELA

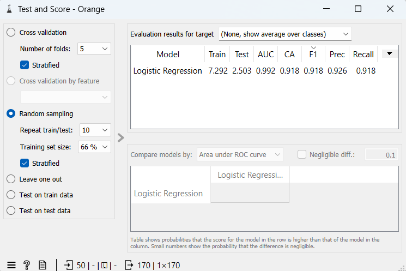
## Logistička regresija, Testiranje i ocjenjivanje

Logistička regresija je algoritam koji nam je prije svega ključan za evaluacijski proces jer se upravo pomoću njega model trenira kako bi rezultati kasnije tijekom testiranja bili valjani. Algoritam će se bazirati na ridge regularizaciji (L2), tj. tehnici koja se koristi za sprječavanje prenaučenosti modela. Jako je bitno paziti da se model ne prilagodi previše podatcima za treniranje, jer onda neće kvalitetno grupirati nove, testne podatke. Inače, logistička se regresija pretežito koristi za binarnu klasifikaciju, međutim može se proširiti i na klasifikaciju više klasa. U našem slučaju radi se o pet klasa, te algoritam koristi tehniku zvanu „*one-vs-all*“. Dakle, u kontekstu razvrstavanja slika u pet bioma, stvara se pet zasebnih binarnih klasifikatora, od kojih svaki razlikuje klasu od ostalih. Klasifikator za svaku klasu osposobljen je za razlikovanje instanci te klase od svih ostalih instanci, tretirajući ih kao jednu pozitivnu klasu, a preostale klase kao kombiniranu negativnu klasu. U priloženome (slika 9) možemo vidjeti implementaciju samog algoritma kao podlogu za testiranje.



Slika 9: Implementacija logističke regresije

Za potrebe evaluacije koristit ćemo se widget-om „Test and Score“ kako bi temeljito provjerili na novim podatcima koliko su predikcije modela pouzdane. Naših slika ima 50, od čega je u logističku regresiju poslano 66% podataka za treniranje (33 slike), što znači da će se model testirati na preostalih 34% (17 slika). Broj epoha je postavljen na 10, što znači da će model klasificirati sveukupno 170 slika. Svih deset puta podatci za treniranje/testiranje se odabiru nasumično (opcija „*random sampling*“).



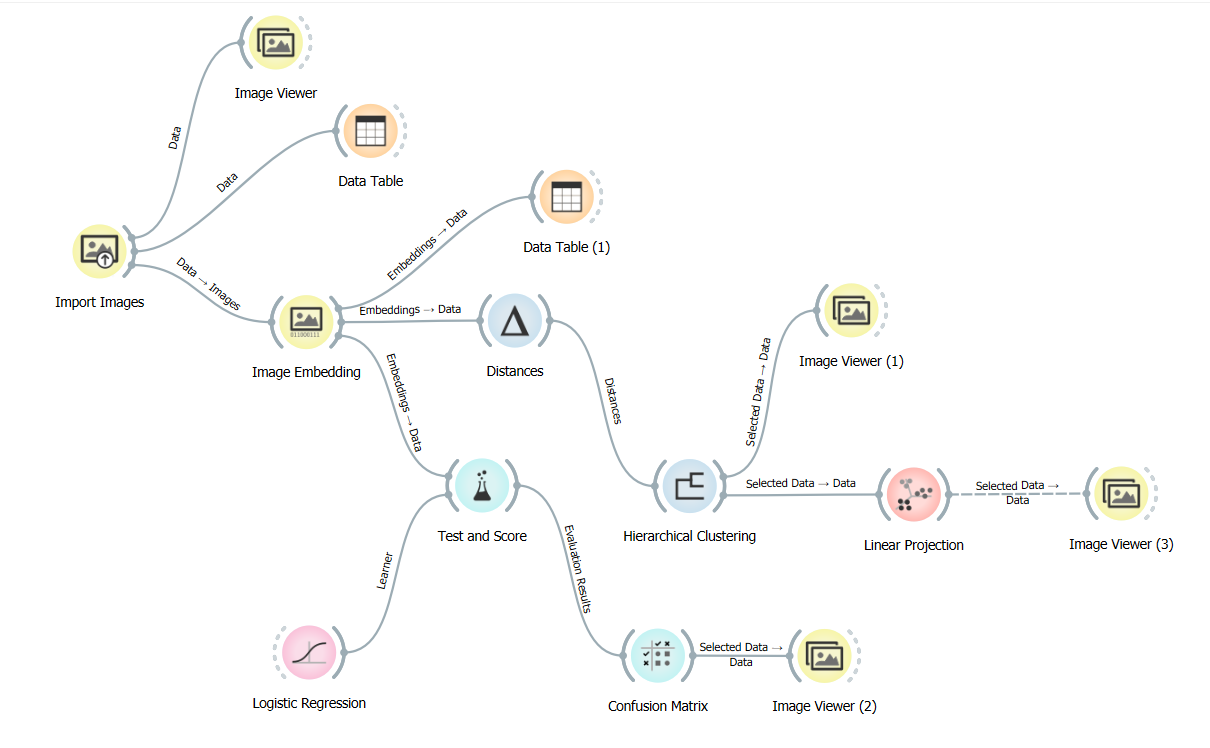
Slika 10: Evaluacija prediktivnog modela

Nakon odrađenog testiranja dobivamo uvid (slika 10) u validaciju promatrajući korištene metrike: AUC, CA, F1, preciznost, te odziv. CA (točnost klasifikacije) je izravna metrika koja predstavlja omjer ispravno klasificiranih instanci prema ukupnom broju instanci u skupu podataka. AUC (površina ispod „ROC“ krivulje) mjeri sposobnost modela da razlikuje između pozitivnih i negativnih klasa, dok se preciznost i odziv fokusiraju na specifične aspekte performansi modela u vezi s pozitivnim predviđanjima. F1 mjera je harmonična sredina preciznosti i odziva. Također imamo zabilježeno vrijeme koje je bilo potrebno za treniranje modela (7.292s), te zatim i testiranje istog (2.503s). Preciznost, odziv, F1 mjera i ROC AUC u svom su osnovnom obliku zamišljene kao mjere uspješnosti binarne klasifikacije, ali se mogu jednostavno proširiti na probleme klasifikacije s više od dvije klase na način da se izračunavaju posebno za svaku klasu pri čemu sve ostale tretiramo kao jednu zajedničku klasu (Ivandić, 2023.).

Primijetimo da sve naše mjere započinju s 0.9, što kada se pomnoži sa 100 rezultira s preko 90%. Preciznije, rezultati dokazuju da CA iznosi 91,8%, AUC 99,2%, preciznost 92,6%, odziv 91,8%, te F1 mjera 91,8%. Možemo reći da je model veoma uspješan u klasifikaciji naših slika zavičaja.

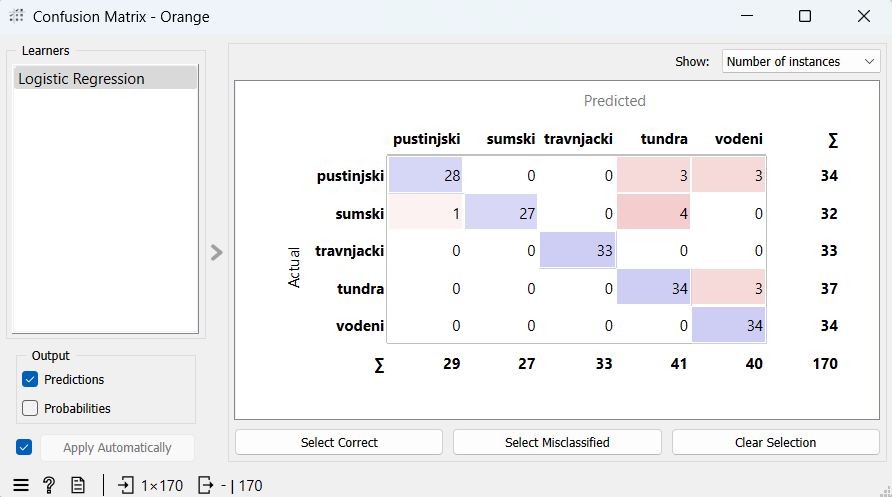
## Matrica konfuzije

Osim metrika, uključivanjem matrice konfuzije dodatno ćemo obogatiti evaluacijsku fazu. Matrica nam omogućuje detaljnu analizu performansi modela predstavljanjem instanci ispravnih i pogrešnih klasifikacija. Ova vizualna pomoć posebno je vrijedna za razumijevanje gdje je model zakazao i identificiranje potencijalnih područja za poboljšanje. Dodajmo widget „Confusion Matrix“ kao posljednji dodatak modelu, zajedno s još jednim widget-om za prikaz slika kako bi mogli popratiti matricu.



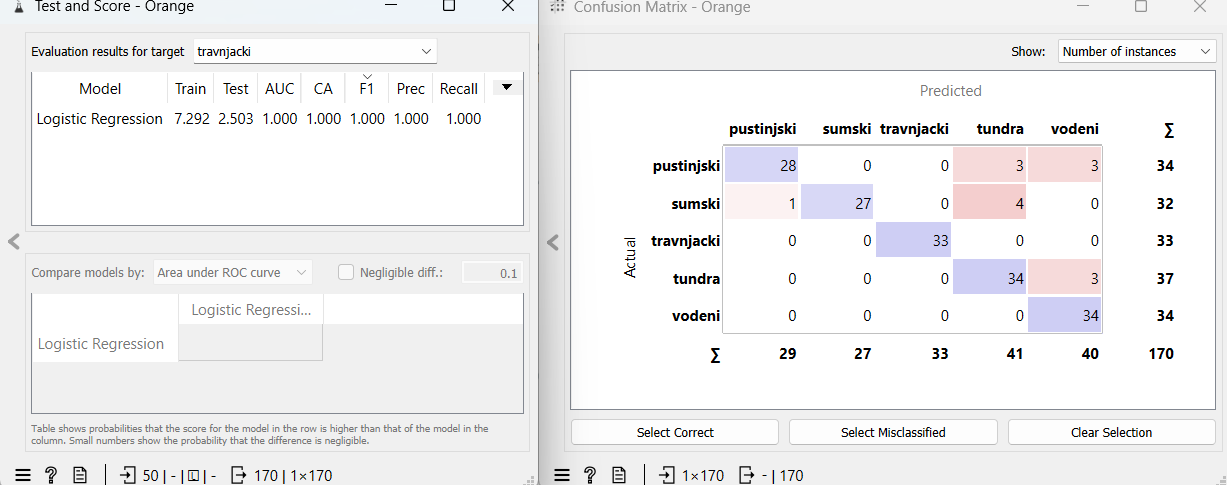
Slika 11: Model završen dodavanjem matrice konfuzije

U nastavku (slika 12) prikazana je matrica konfuzije u kojoj ljubičasta boja označava sve točno razvrstane instance, dok ostala polja ukazuju na pogrešnu klasifikaciju, ukoliko imaju vrijednost više od nula. U prethodnom potpoglavlju smo istakli uspješnost modela na temelju metrika, sada to možemo vidjeti i ovdje, budući da količina pogrešnih klasifikacija naspram ukupnih 170 zaista nije velika. Greške se javljaju najviše kod tundra i vodenog bioma, za koje smo već uočili da su „problematični“ iz razloga što njihova obilježja podosta naliče obilježjima drugih bioma.



Slika 12: Prikaz matrice konfuzije

Nadalje, uzmimo za primjer samo jednu od pet klasa, travnjački biom. Otvorimo istovremeno rezultate testiranja i matricu konfuzije, te postavimo filter na traženu klasu. Vidimo (slika 13) da je za ovu klasu model uspio predvidjeti sve instance točno, dakle one koje joj pripadaju i ne pripadaju. Paralelu povlačimo sa stopostotnih metrika na matricu koja pokazuje kako uistinu nije bilo krivih klasificiranja za ovaj biom. Ovo nam dokazuje kako je tijekom treniranja model naučio razlikovati travnjački biom bolje nego ikoji drugi, vjerojatno zato što ima više unikatan izgled u usporedbi s ostalim kategorijama koje sve međusobno dijele barem neke značajke (primjerice brda/planine).



Slika 13: Filtriranje rezultata evaluacije na samo jednu klasu

# ZAKLJUČAK

Povodom završetka kompletne realizacije i detaljne analize rezultata, možemo zaključiti da razvijeni model pruža snažan okvir za klasifikaciju slika krajolika u različite biome. Nakon treniranja, odrađenog uz pažljiv odabir presudnih stavki, visoka točnost klasificiranja testnih podataka potvrđuje izvrsne performanse modela, dok su eventualne manje pogreške primijećene samo u očekivanim slučajevima, gdje su slike bile zaista slične biomu u koji su svrstane.

Uvažavajući Orange okruženje, cijeli projekt bio je obogaćujuće iskustvo, pružajući pristupačan način izrade modela uz istovremeno učinkovito iščitavanje vizualnih prikaza klasificiranja instanci. Ovime se čitateljima pruža intimniji uvid u koncept klasifikacije u strojnom učenju, bez obzira na njihovo prethodno iskustvo s područjem umjetne inteligencije.

Impresionirani rezultatima ovog modela, otvaramo vrata budućim mogućnostima. Daljnja nadogradnja novim funkcionalnostima, optimizacija putem dubljih istraživanja i moguća integracija u veće projekte predstavljaju perspektivu koja seže izvan trenutačnog okvira. Ovaj prediktivni model jednoga dana može poslužiti kao temelj za inovativne primjene u području analize slika krajolika, te na koncu pridonijeti napretku u području umjetne inteligencije.

# LITERATURA

1. Saša Mladenović, vježbe kolegija UUI na portalu <https://edit.dalmacija.hr/>
2. Jan Šnajder, Strojno učenje: 19. Grupiranje, 2020./2021.
3. Ivandić Tomislav, Primjena strojnog učenja u klasifikaciji slika kao tipičnih primjera nestrukturiranih tipova podataka, 2023.
4. <https://orangedatamining.com/>
5. <https://www.wikipedia.org/>
6. <https://www.geeksforgeeks.org/>
   1. <https://www.geeksforgeeks.org/embedded-images/>
   2. <https://www.geeksforgeeks.org/ml-hierarchical-clustering-agglomerative-and-divisive-clustering/>
   3. <https://www.geeksforgeeks.org/projection-perspective-in-machine-learning/>
   4. <https://www.geeksforgeeks.org/understanding-logistic-regression/>
   5. <https://www.geeksforgeeks.org/confusion-matrix-machine-learning/>

Formirajući ovaj seminar, uz znanje stečeno na kolegiju, korišteni su navedeni izvori kako bi se potkrijepilo objašnjenje određenih pojmova u sklopu teme. Resursi Orange Data Mining stranice primijenjeni su isključivo kod predstavljanja samog Orange sustava, što naravno uključuje rad s njegovim alatima. Wikipedia i GeeksforGeeks portali poslužili su na opširnijem tumačenju određenih pojmova koji se dotiču klasifikacije u strojnom učenju.

# SAŽETAK

Ideja projekta se temelji na sastavljanju prediktivnog modela za klasifikaciju slika krajolika u pet različitih bioma koristeći alat za upravljanje podatcima zvan Orange. Projekt je motiviran praktičnim prednostima strojnog učenja, posebice kategorizacijom slika koja olakšava obradu kompleksnih zadataka. Opisali smo općenite karakteristike pet glavnih bioma - vodenog, travnjačkog, šumskog, pustinjskog i tundre, te istaknuli sklop razmišljanja koji dovodi do prepoznatljivosti svakog zasebno, što je upravo opis pozadinskog procesa modela umjetne inteligencije.

Priprema podataka uključivala je opis planiranog modela i predstavljanje Orange softvera, a zatim smo počeli s kreiranjem. Redom smo učitali slike, izvršili njihovu ugradnju, te implementirali algoritme klasifikacije – izračun vektorske udaljenosti, hijerarhijsko klasteriranje, linearnu projekciju, i logističku regresiju. Evaluacija modela obuhvatila je testiranje i ocjenjivanje rada logističke regresije, te zatim analizu matrice konfuzije.

Zaključak naglašava da razvijeni model pruža, za klasifikaciju slika, snažan okvir s visokom točnošću. Rezultati evaluacije dokazali su kvalitetu predikcija uz manje očekivane pogreške. Model se može dalje proširiti dodavanjem novih funkcionalnosti i optimizirati daljnjim istraživanjem.