

Przewidywanie funkcji troficznej gatunków z królestwa **Fungi** na podstawie adnotacji genomowej.



Filip Sobczak
Natalia Grodzka
Maksymilian Chmielewski
Zuzanna Wawrzyczek
Tomek Dzikowski
Maciej Hryc

Cel i kontekst biologiczny badań troficznych

- Zrozumienie roli ekologicznej grzybów
- Wykorzystanie rosnących zasobów genomowych
- Zastosowania praktyczne – rolnictwo, biotechnologia, ochrona środowiska
- Ochrona różnorodności biologicznej i monitoring środowiska



Tabela zawierająca informacje na temat liczby kopii genów

JGI ID	Phylogenetic	On Species	Ecology	Phylum	Class	Order	Family	Genus	GGGX ABH01	GGGX ABH03	GX ABH19	GX ABH23	GX ABH16
Chalo1	1	Chalara longipes	Saprotroph	Ascomycota	Leotiomyces	Helotiales		Chalara	13	23	1	2	0
Melva1	2	Meliniomyces va	Ericoid mycorrhiz	Ascomycota	Leotiomyces	Helotiales	Helotiaceae	Meliniomyces	12	17	1	3	0
Melbi2	3	Meliniomyces bi	Ericoid mycorrhiz	Ascomycota	Leotiomyces	Helotiales	Helotiaceae	Meliniomyces	11	13	1	1	0
Rhier1	4	Rhizoscyphus ei	Ericoid mycorrhiz	Ascomycota	Leotiomyces	Helotiales	Leotiaceae	Pezoloma	6	10	1	2	0
Acema1	5	Acephala macro	Ectomycorrhizae	Ascomycota	Leotiomyces	Helotiales		Acephala	9	21	1	3	0
Phisc1	6	Phialocephala si	Endophyte	Ascomycota	Leotiomyces	Helotiales		Phialocephala	10	24	1	2	0
Ascsa1	7	Ascocoryne sarc	Endophyte	Ascomycota	Leotiomyces	Helotiales	Helotiaceae	Ascocoryne	3	3	0	0	0
Glalo1	8	Glarea lozoyens	Saprotroph	Ascomycota	Leotiomyces	Helotiales	Helotiaceae	Glarea	6	10	1	2	0
Amore1	9	Amorphotheca ri	Saprotroph	Ascomycota	Leotiomyces		Myxotrichaceae	Amorphotheca	2	5	0	2	0
Oidma1	10	Oidiodendron m	Ericoid mycorrhiz	Ascomycota	Leotiomyces		Myxotrichaceae	Oidiodendron	17	16	0	4	0
Bissp1	11	Bisporella sp.	Saprotroph	Ascomycota	Leotiomyces	Helotiales	Helotiaceae	Bisporella	10	15	0	1	0
Scisc1	12	Sclerotinia scler	Pathogen	Ascomycota	Leotiomyces	Helotiales	Sclerotiniaceae	Sclerotinia	3	5	1	1	0
Botci1	13	Botrytis cinerea	Pathogen	Ascomycota	Leotiomyces	Helotiales	Sclerotiniaceae	Botrytis	7	12	1	0	0
Maggr1	14	Magnaporthe gri	Pathogen	Ascomycota	Sordariomycetes	Magnaporthales	Magnaporthaceae	Magnaporthe	6	7	1	3	0
Neucr_trp3_1	15	Neurospora cras	Saprotroph	Ascomycota	Sordariomycetes	Sordariales	Sordariaceae	Neurospora	1	5	1	2	0
Trire2	16	Trichoderma ree	Saprotroph	Ascomycota	Sordariomycetes	Hypocreales	Hypocreaceae	Trichoderma	1	9	0	2	0
Xylhyp1	17	Xylaria hypoxyla	Saprotroph	Ascomycota	Sordariomycetes	Xylariales	Xylariaceae	Xylaria	2	10	0	3	1
Cenge3	18	Cenococcum ge	Ectomycorrhizae	Ascomycota	Dothideomycetes		Gloniaceae	Cenococcum	2	8	1	2	0
Glost2	19	Glonium stellatu	Saprotroph	Ascomycota	Dothideomycetes		Gloniaceae	Glonium	6	10	0	3	0
Leppa1	20	Lepidopterella pi	Saprotroph	Ascomycota	Dothideomycete	Mytilinidiales	Argynnaceae	Lepidopterella	2	9	0	3	0
Stasp1	21	Stagonospora sp	Pathogen	Ascomycota	Dothideomycete	Pleosporales	Massarinaceae	Stagonospora	7	13	1	3	0
Altal1	22	Alternaria altern	Pathogen	Ascomycota	Dothideomycete	Pleosporales	Pleosporaceae	Alternaria	8	12	0	3	0
Aurpu_var_pul1	23	Aureobasidium p	Saprotroph	Ascomycota	Dothideomycete	Dothideales	Saccoltheciaceae	Aureobasidium	7	12	1	0	0
Clafu1	24	Cladosporium fu	Pathogen	Ascomycota	Dothideomycete	Capnodiales	Mycosphaerellae	Passalora	4	6	1	1	1
Aspcl1	25	Aspergillus clav	Saprotroph	Ascomycota	Eurotiomycetes	Eurotiales	Aspergillaceae	Aspergillus	2	4	1	2	0





Random Forest



1. Tworzy wiele drzew decyzyjnych:

(każde na innym losowym podzbiorze danych i cech)

2. Wprowadza losowość:

– wybiera losowe próbki (bootstrap) i losowe cechy przy każdym podziale drzewa

3. Uśrednia wyniki (regresja):

lub głosuje większością (klasyfikacja).

4. Minimalizuje przeuczenie (overfitting):

dzięki agregacji wielu słabych modeli

5. Działa dobrze na danych z wieloma cechami i szumem



Dobór parametrów

Zbiór danych

Zbiór uczący i testowy został losowo podzielony – 30% danych przeznaczono na testy, a pozostałe 70% wykorzystano do trenowania modelu.

Ocena modelu

Wydajność oceniano na podstawie metryki F1 score – im wyższy wynik, tym lepsza równowaga między precyzją a czułością.
Wygenerowano też heatmaps.

Sekretowane CAZyme

Secreted CAZymes

Total number of test samples: 34

Best Parameters: {'bootstrap': False, 'class_weight': {0: 1, 1: 1.7}, 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}

F1 score: 0.625

Classification Report:

precision recall f1-score support

0	0.82	0.96	0.88	24
1	0.83	0.50	0.62	10

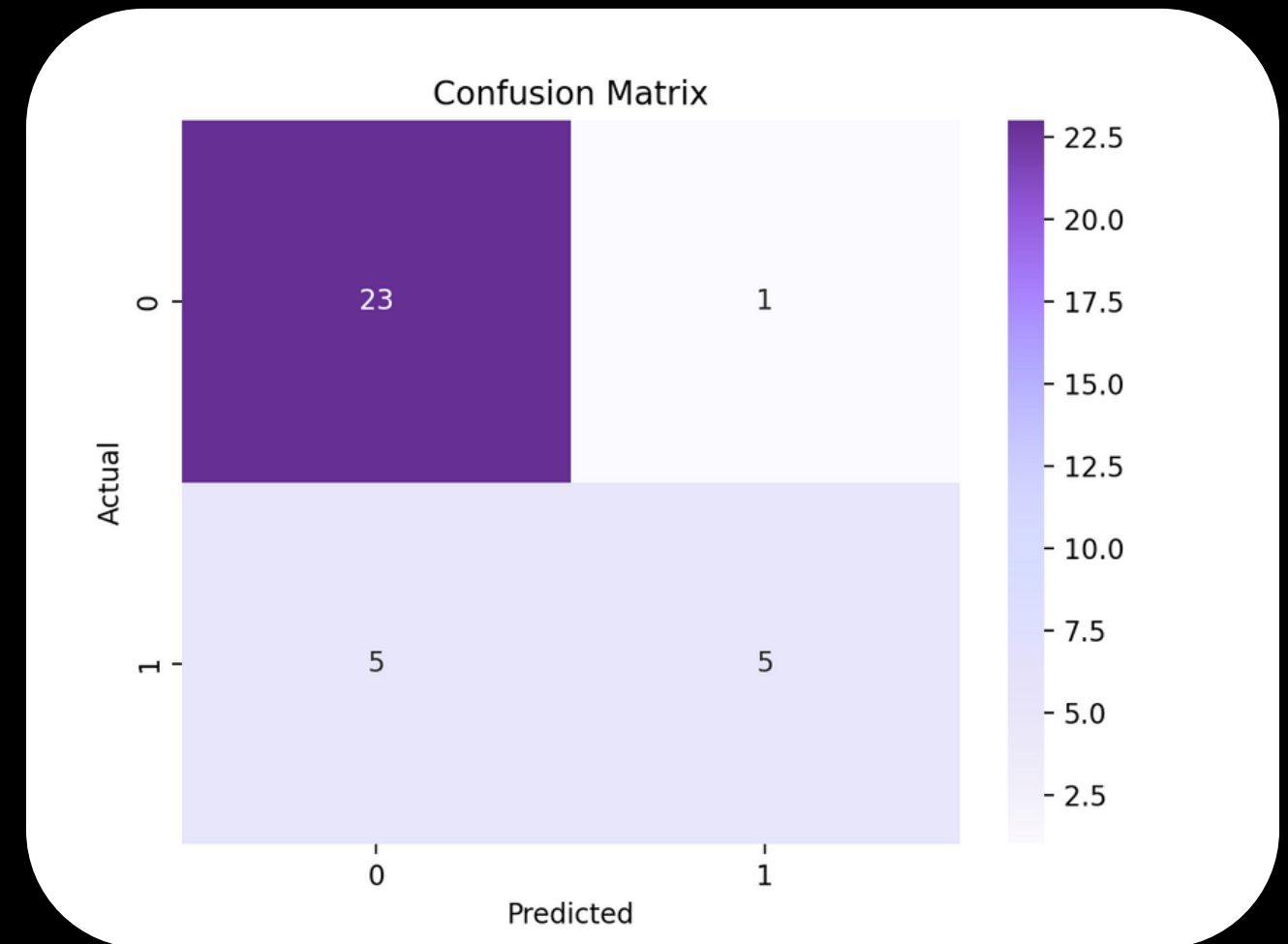
accuracy			0.82	34
macro avg	0.83	0.73	0.75	34
weighted avg	0.82	0.82	0.81	34

Top10 Important Features:

Feature Importance

20	CBM1	0.041582
50	CE15	0.040148
75	GH10	0.033276
14	AA9	0.032918
126	GH72	0.030554
10	AA5_1	0.026910
43	CE1	0.025717
95	GH28	0.023403
136	GH92	0.019269
130	GH78	0.018132

```
param_grids = [  
    {  
        'n_estimators': [40 + (5 * i) for i in range(5)] + [1000 + (500 * i) for i in range(5)],  
        'max_depth': [i for i in range(3, 7)],  
        'max_features': ['sqrt', 'log2', None],  
        'min_samples_split': [2, 3],  
        'min_samples_leaf': [1, 2, 3],  
        'bootstrap': [True, False],  
        'class_weight': [{0: 1, 1: 0.8}, {0: 1, 1: 1.6}, {0: 1, 1: 1.7}, {0: 1, 1: 1.8}]  
    }  
]
```



Sekretowane lipazy

Secreted lipase

Total number of test samples: 34
Best Parameters: {'bootstrap': False, 'class_weight': {0: 1, 1: 0.8}, 'max_depth': 2, 'max_features': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 30}
F1 score: 0.15384615384615385

Classification Report:
precision recall f1-score support

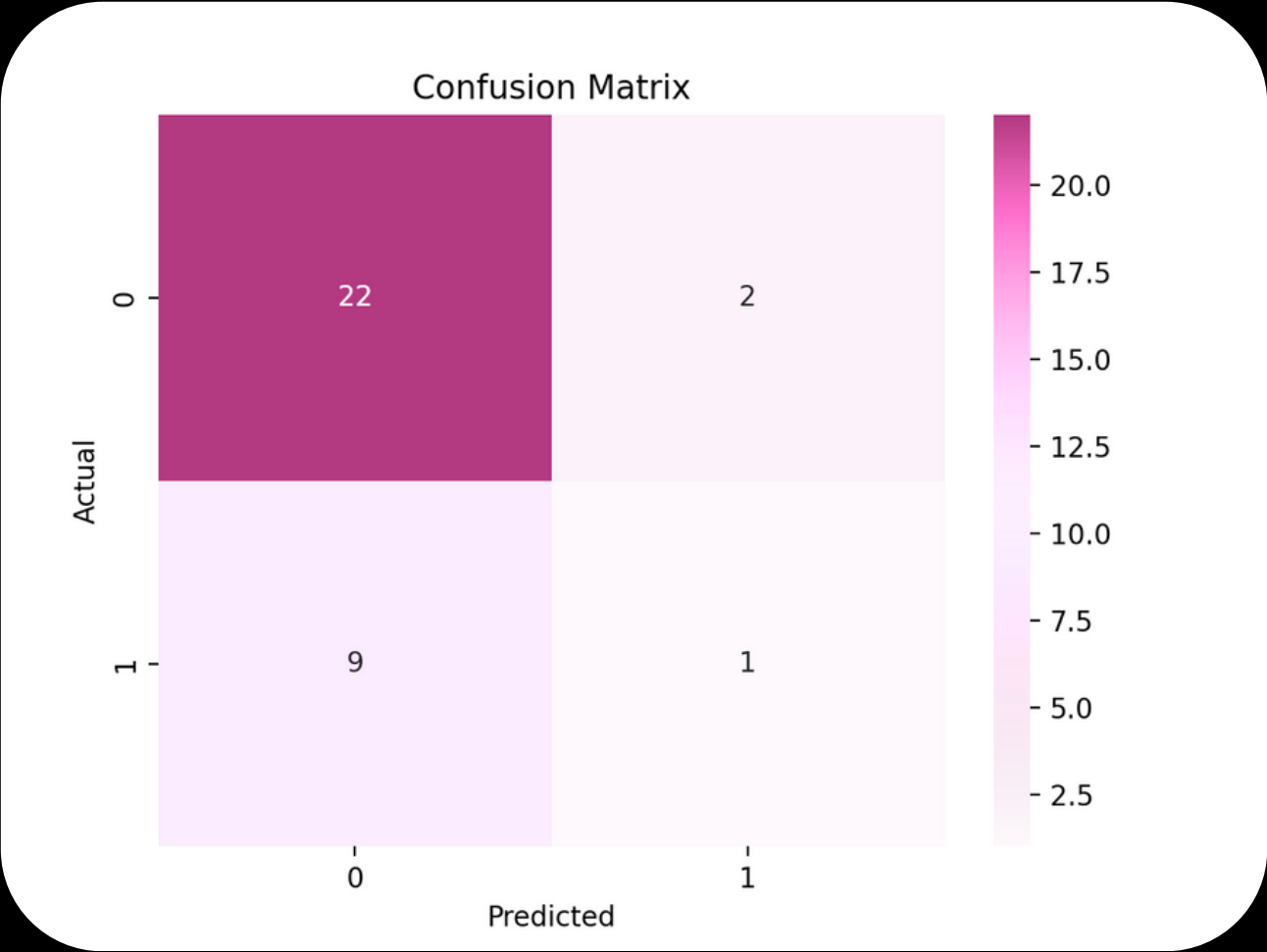
0	0.71	0.92	0.80	24
1	0.33	0.10	0.15	10

accuracy		0.68		34
macro avg	0.52	0.51	0.48	34
weighted avg	0.60	0.68	0.61	34

Top10 Important Features:

	Feature	Importance
1	GGGX ABH03	0.360012
0	GGGX ABH01	0.285908
3	GX ABH23	0.222285
2	GX ABH19	0.066717
6	GGGX ABH04	0.035538
8	GX ABH22	0.022019
4	GX ABH16	0.007521
5	GX ABH09	0.000000
7	GX ABH14	0.000000
9	Y ABH27	0.000000

```
param_grids = [  
    {  
        'n_estimators': [30 + (5 * i) for i in range(5)] + [1000 + (500 * i) for i in range(5)],  
        'max_depth': [i for i in range(1, 4)],  
        'max_features': ['sqrt', 'log2', None],  
        'min_samples_split': [2, 3, 4],  
        'min_samples_leaf': [1, 2],  
        'bootstrap': [True, False],  
        'class_weight': [{0: 1, 1: 0.5}, {0: 1, 1: 0.8}, {0: 1, 1: 1}]  
    }  
]
```



Sekretowane proteazy

Secreted Proteases

Total number of test samples: 34
Best Parameters: {'bootstrap': True, 'class_weight': {0: 1, 1: 2.6}, 'max_depth': 2, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 950}
F1 score: 0.3157894736842105

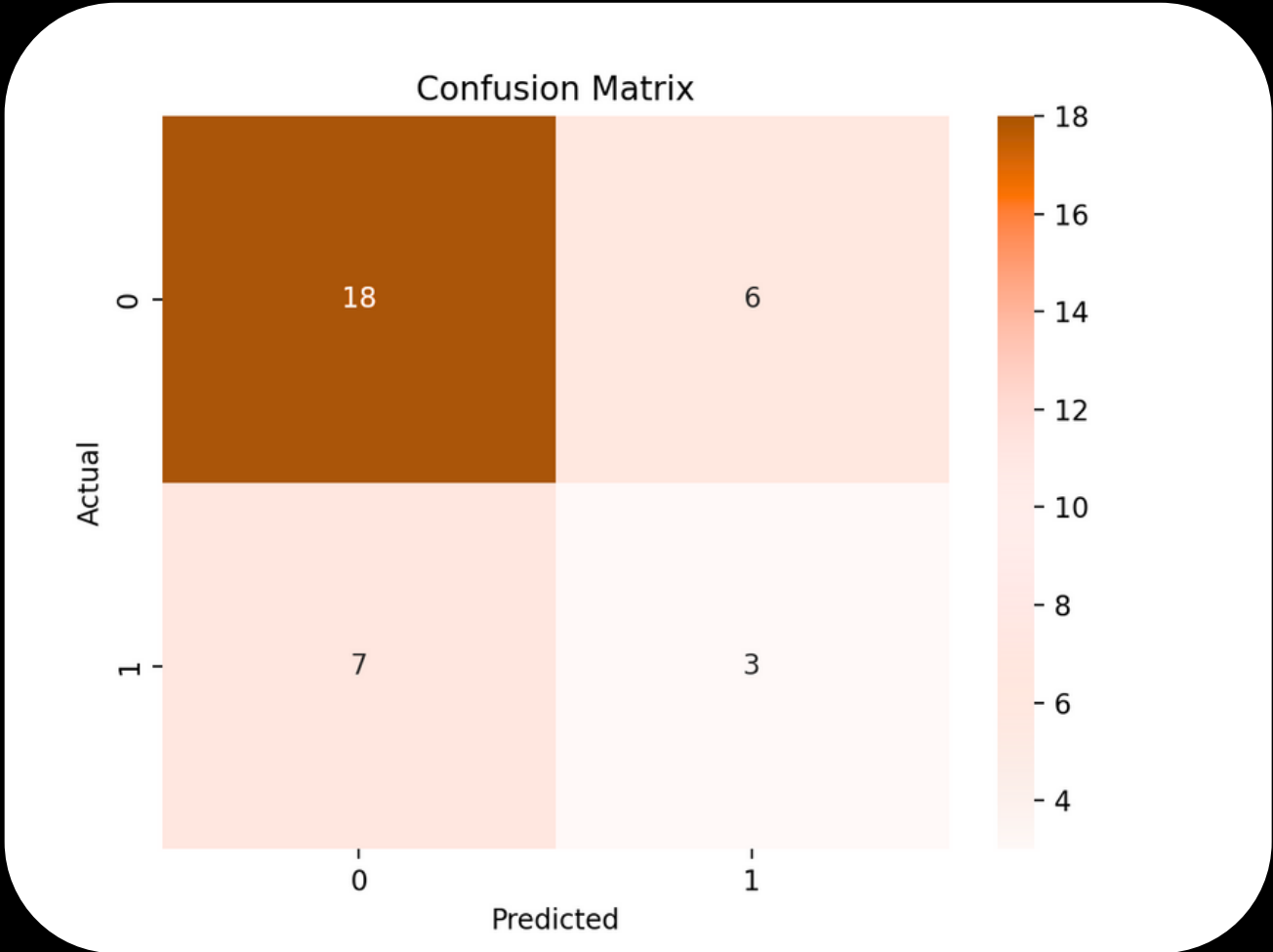
Classification Report:
precision recall f1-score support

0	0.72	0.75	0.73	24
1	0.33	0.30	0.32	10
accuracy			0.62	34
macro avg	0.53	0.53	0.53	34
weighted avg	0.61	0.62	0.61	34

Top10 Important Features:

	Feature	Importance
16	S53	0.063757
0	A01A	0.061653
15	S28	0.052438
14	S10	0.046521
11	M35	0.038084
21	M14A	0.036581
19	C56	0.036093
51	I87	0.035521
12	S08A	0.034937
3	I32	0.034487

```
param_grids = [  
    {  
        'n_estimators': [400 + (200 * i) for i in range(8)],  
        'max_depth': [i for i in range(1, 4)],  
        'max_features': ['sqrt', 'log2', None],  
        'min_samples_split': [2, 3],  
        'min_samples_leaf': [1, 2, 3],  
        'bootstrap': [True, False],  
        'class_weight': [{0: 1, 1: 0.8}, {0: 1, 1: 2.5}, {0: 1, 1: 2.6}, {0: 1, 1: 2.7}]  
    }  
]
```



Mate sekretowane białka

SSPs

Total number of test samples: 34
Best Parameters: {'bootstrap': True, 'class_weight': {0: 1, 1: 1.8}, 'max_depth': 4, 'max_features': None, 'min_samples_leaf': 5, 'min_samples_split': 2, 'n_estimators': 65}
F1 score: 0.6666666666666666

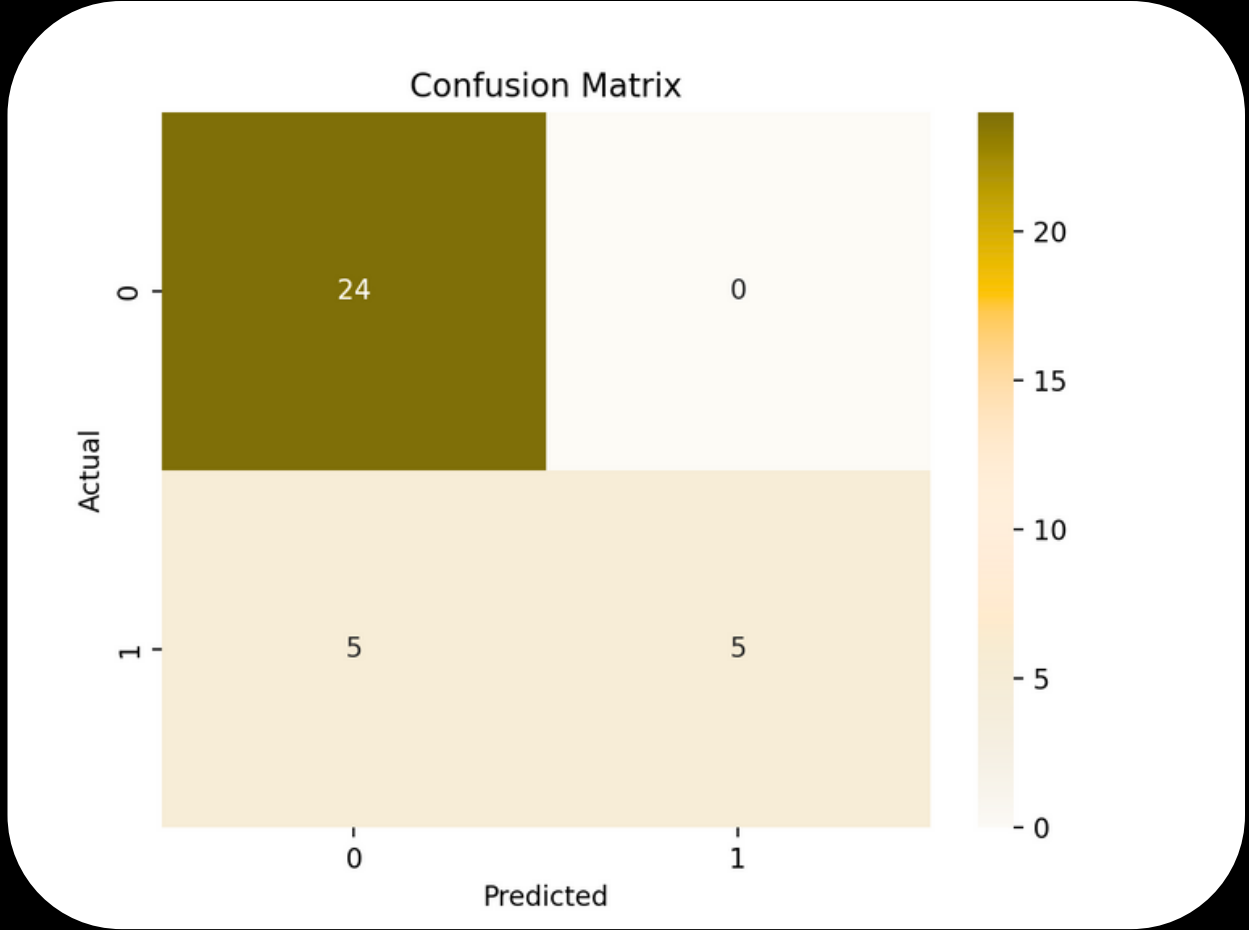
Classification Report:

	precision	recall	f1-score	support
0	0.83	1.00	0.91	24
1	1.00	0.50	0.67	10
accuracy		0.85	34	
macro avg	0.91	0.75	0.79	34
weighted avg	0.88	0.85	0.84	34

Top 10 Important Features:

	Feature	Importance
49	3443	0.039754
119	734	0.033594
105	unknown	0.023193
116	545	0.016969
162	1183	0.014960
35	1764	0.014785
13	445	0.014637
205	00457_00734	0.013437
159	314	0.012983
34	1670	0.012457

```
param_grids = [  
  {  
    'n_estimators': [60 + (5 * i) for i in range(5)] + [1000 + (500 * i) for i in range(5)],  
    'max_depth': [i for i in range(2, 5)],  
    'max_features': ['sqrt', 'log2', None],  
    'min_samples_split': [2, 3],  
    'min_samples_leaf': [i for i in range(2, 6)],  
    'bootstrap': [True, False],  
    'class_weight': [{0: 1, 1: 1.6}, {0: 1, 1: 1.7}, {0: 1, 1: 1.8}]  
  }  
]
```





Wnioski końcowe

- Wykorzystaliśmy algorytm Random Forest oraz różne zestawy cech: sekretowane CAZymes, lipazy, proteazy i małe sekretowane białka (SSPs).
- Skuteczność modeli oceniano na podstawie F1 score, uwzględniając równowagę między precyzją a czułością.
- Najlepsze wyniki uzyskano dla predykcji opartej na CAZymes i SSPs (F1 score: 0.625 i 0.667).
- Analiza istotności cech ujawniła geny potencjalnie kluczowe dla danego typu funkcji troficznej.
- Wyniki pokazują, że adnotacja genomowa może być skutecznym narzędziem do wnioskowania o ekologicznej roli grzybów.



Dziękujemy
za uwagę

