I CHOOSE YOU...

Pokemon study

Dzmitry Kurch, DSE

Meeting Agenda

Discussion topics for today

01

Exploratory Data Analysis

02

Classification Logistic regression

03

Classification Random Forest

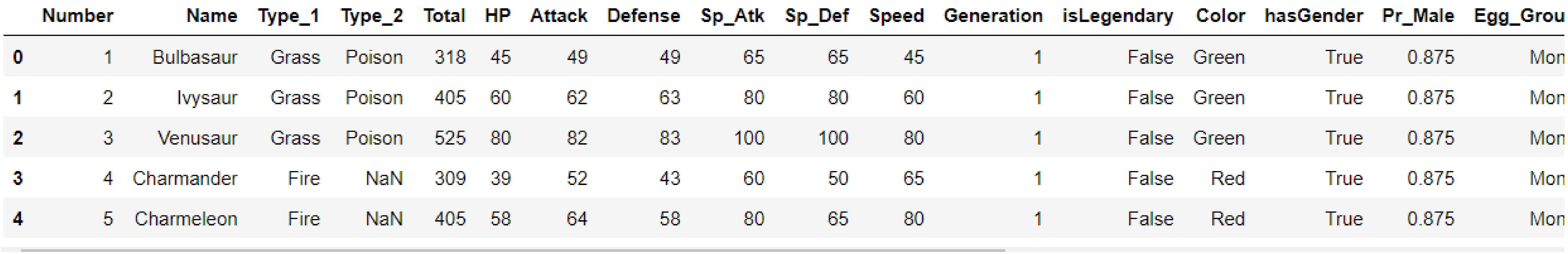
04

PCA and hierarchical clustering

Exploratory data analysis

Dataset

Dataset sample

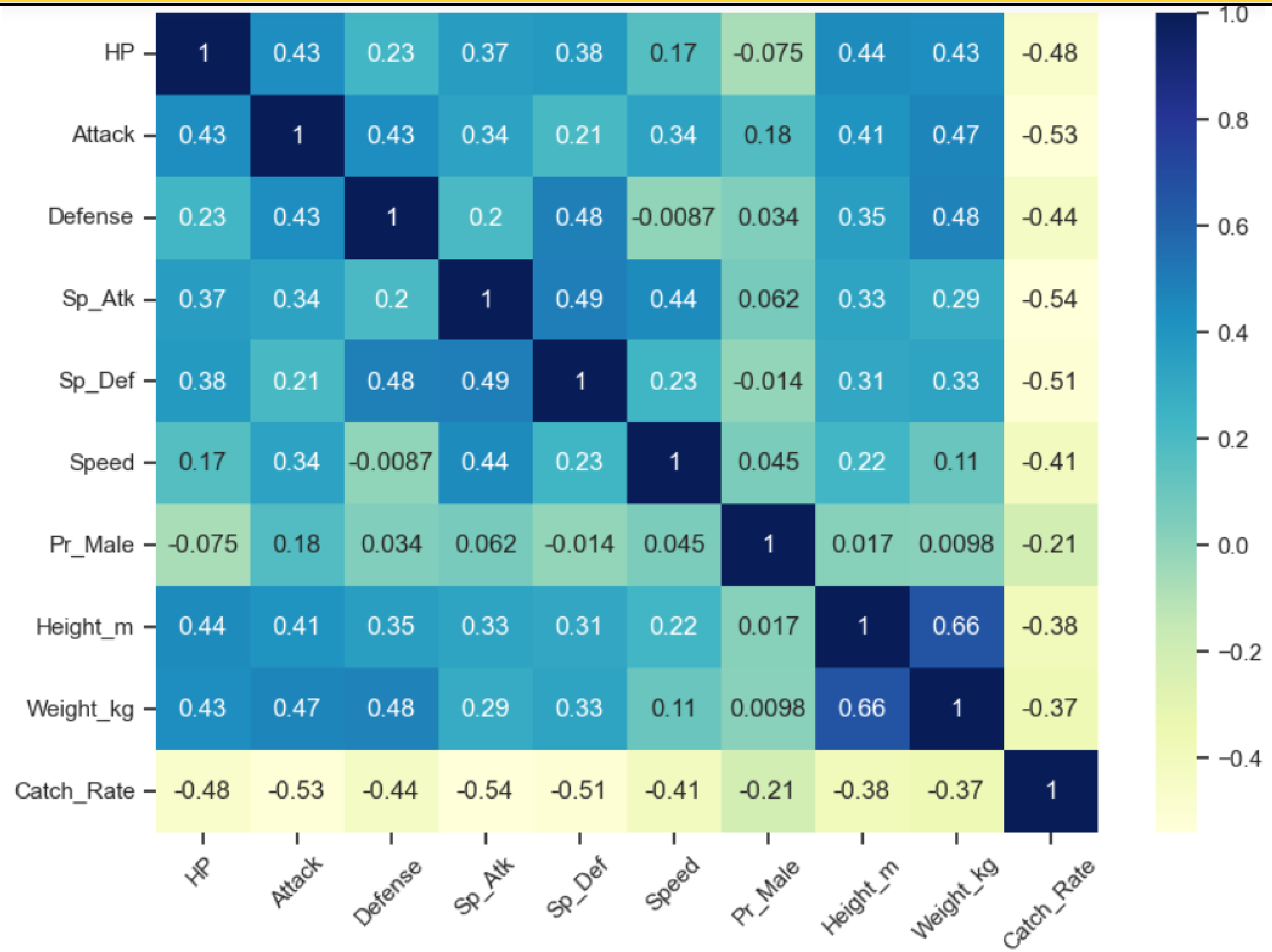


	Number	Name	Type_1	Type_2	Total	HP	Attack	Defense	Sp_Atk	Sp_Def	Speed	Generation	isLegendary	Color	hasGender	Pr_Male	Egg_Grou
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False	Green	True	0.875	Mon
1	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False	Green	True	0.875	Mon
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False	Green	True	0.875	Mon
3	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False	Red	True	0.875	Mon
4	5	Charmeleon	Fire	NaN	405	58	64	58	80	65	80	1	False	Red	True	0.875	Mon

Dataset consists of the Pokemons and their characteristics.

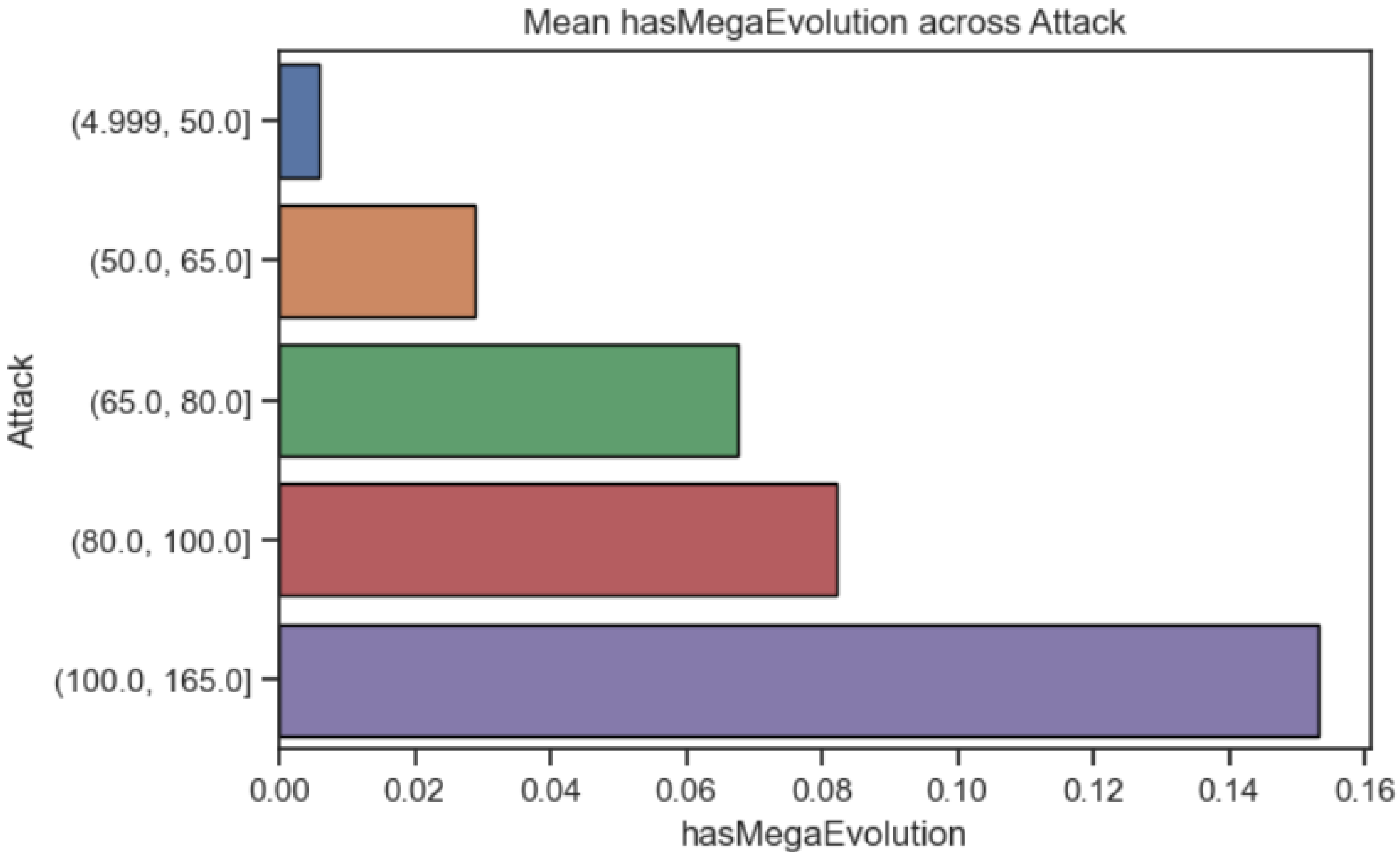
Binary field **hasMegaEvolution** is an important feature in the Pokemon world which indicates its ability to have a temporary *superpower*. We have only **6.4%** samples with that attribute activated.

Correlation matrix

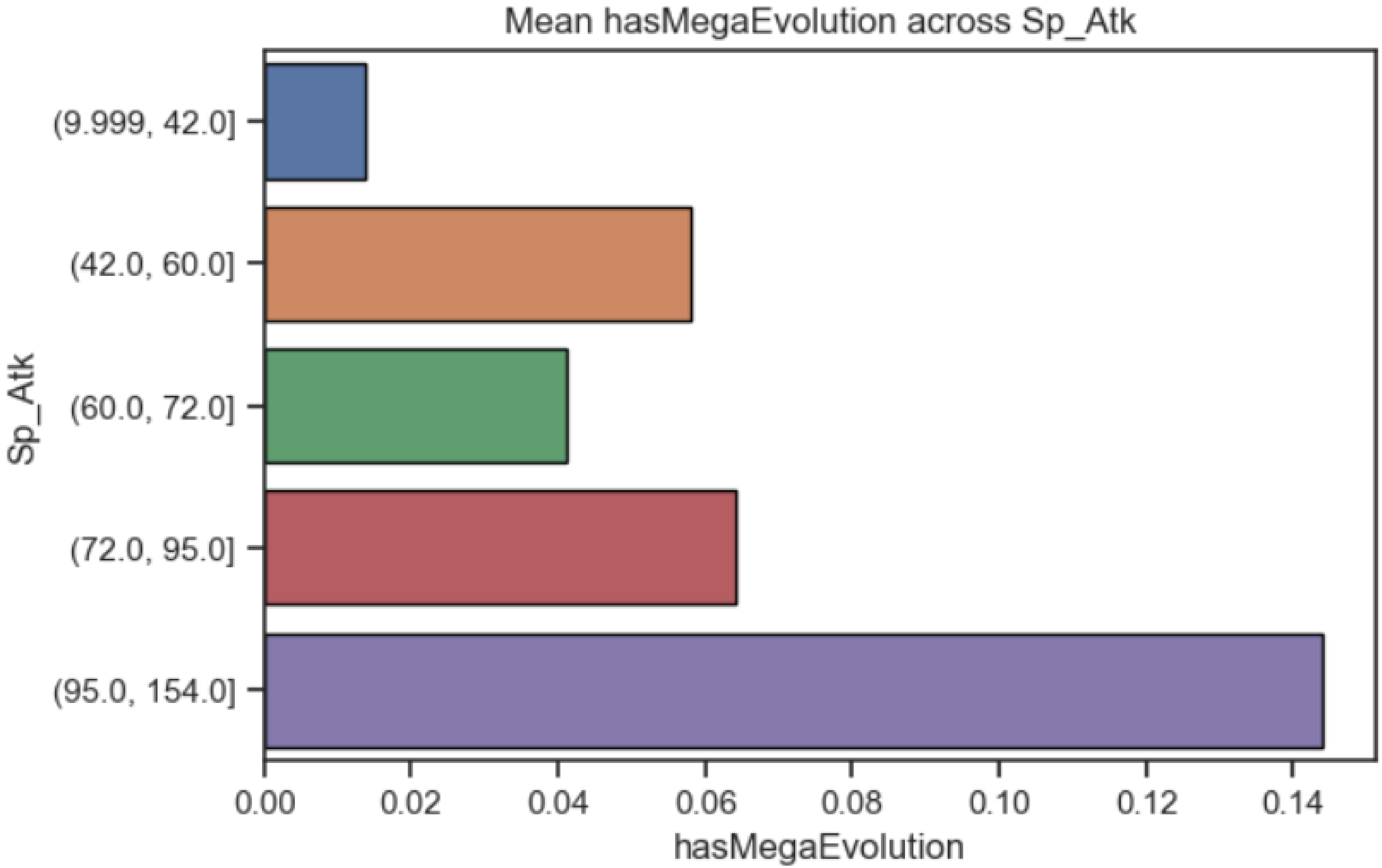


	HP	Attack	Defense	Sp_Atk	Sp_Def	Speed	Pr_Male	Height_m	Weight_kg	Catch_Rate
HP	1	0.43	0.23	0.37	0.38	0.17	-0.075	0.44	0.43	-0.48
Attack	0.43	1	0.43	0.34	0.21	0.34	0.18	0.41	0.47	-0.53
Defense	0.23	0.43	1	0.2	0.48	-0.0087	0.034	0.35	0.48	-0.44
Sp_Atk	0.37	0.34	0.2	1	0.49	0.44	0.062	0.33	0.29	-0.54
Sp_Def	0.38	0.21	0.48	0.49	1	0.23	-0.014	0.31	0.33	-0.51
Speed	0.17	0.34	-0.0087	0.44	0.23	1	0.045	0.22	0.11	-0.41
Pr_Male	-0.075	0.18	0.034	0.062	-0.014	0.045	1	0.017	0.0098	-0.21
Height_m	0.44	0.41	0.35	0.33	0.31	0.22	0.017	1	0.66	-0.38
Weight_kg	0.43	0.47	0.48	0.29	0.33	0.11	0.0098	0.66	1	-0.37
Catch_Rate	-0.48	-0.53	-0.44	-0.54	-0.51	-0.41	-0.21	-0.38	-0.37	1

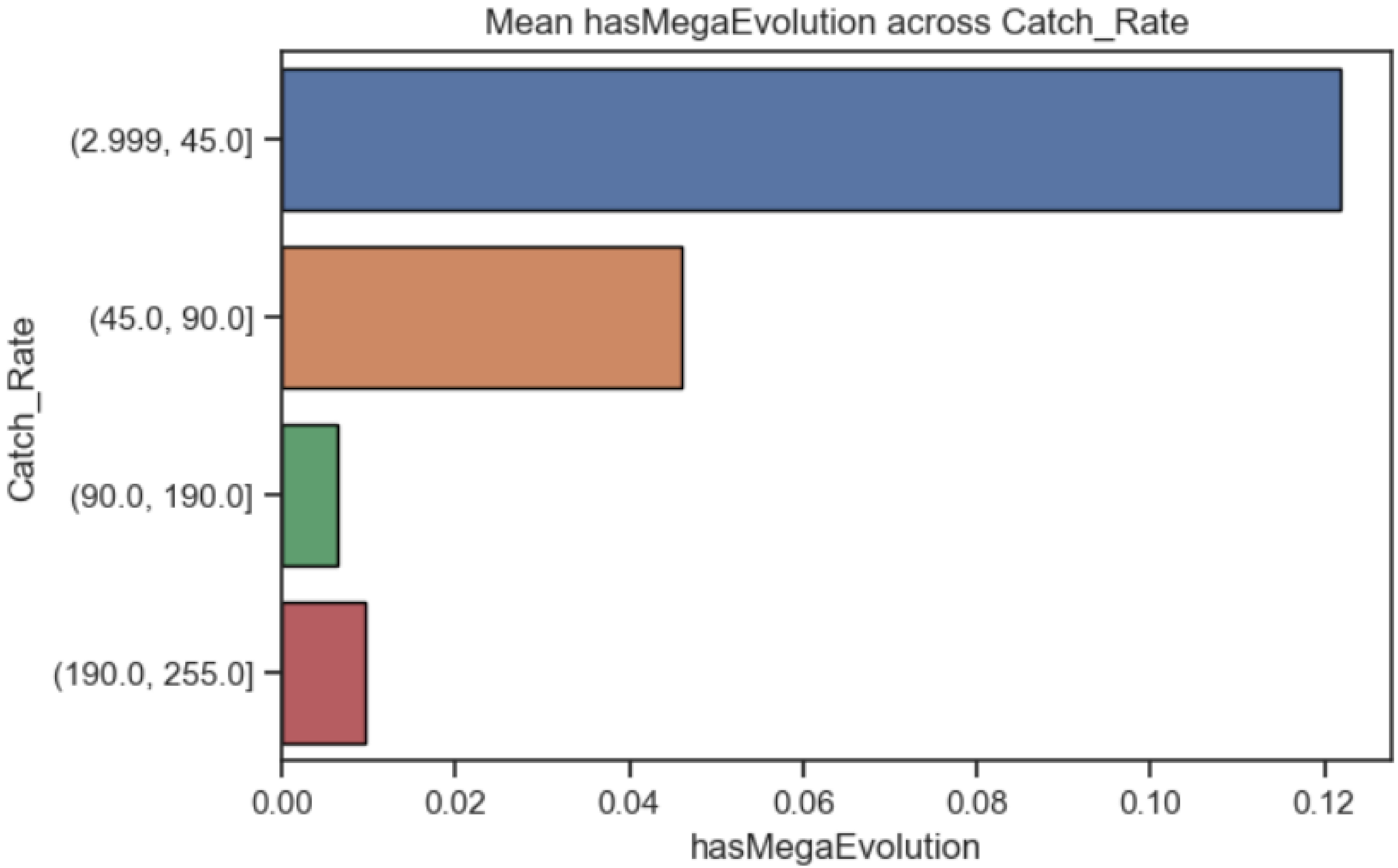
Target density across features



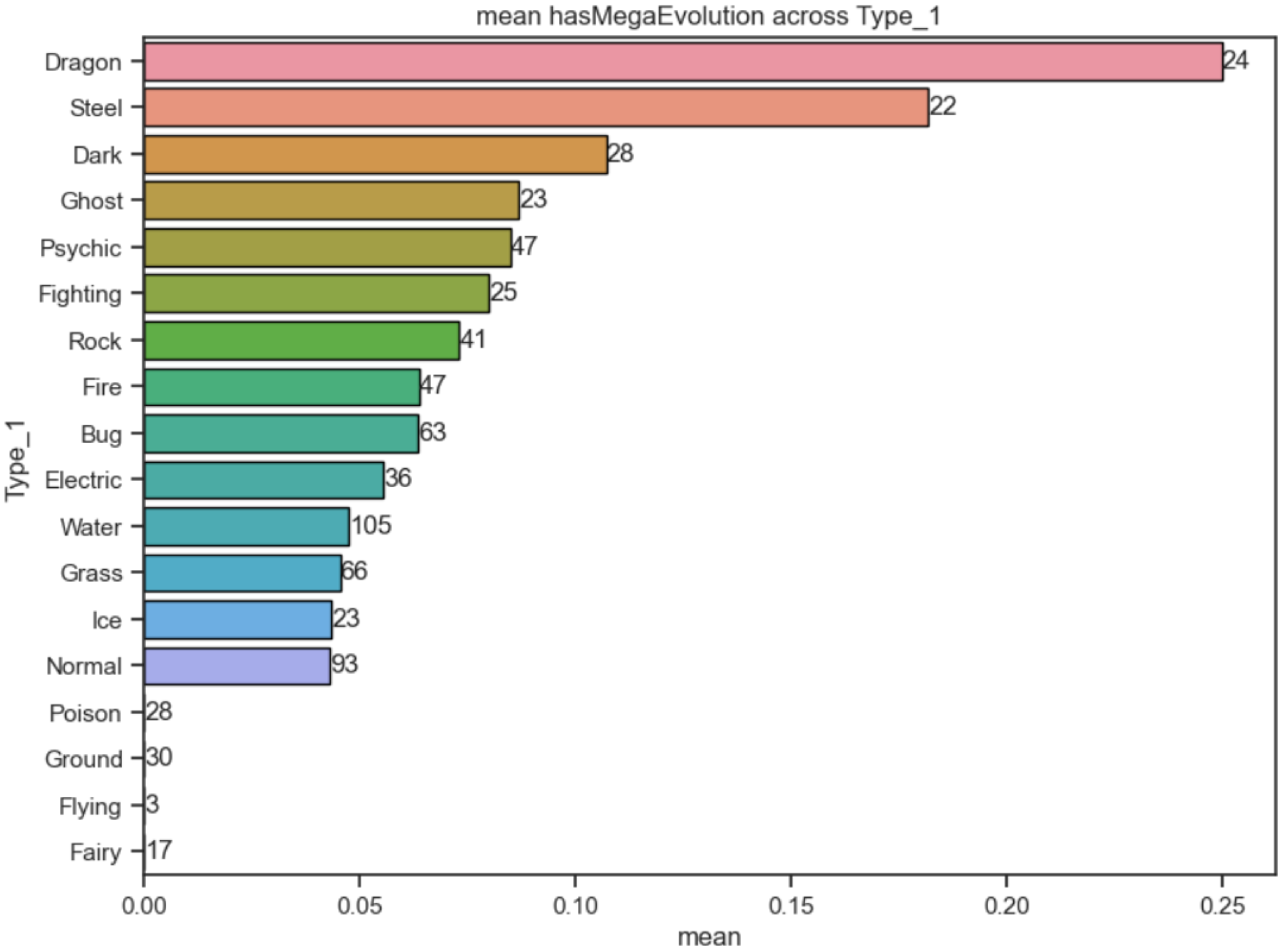
Target density across features



Target density across features



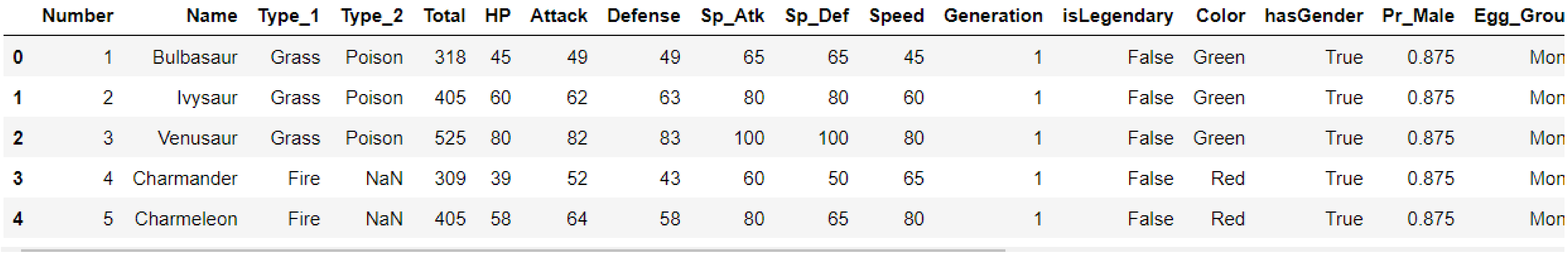
Target density across features



Logistic regression

Data preprocessing

Dataset sample

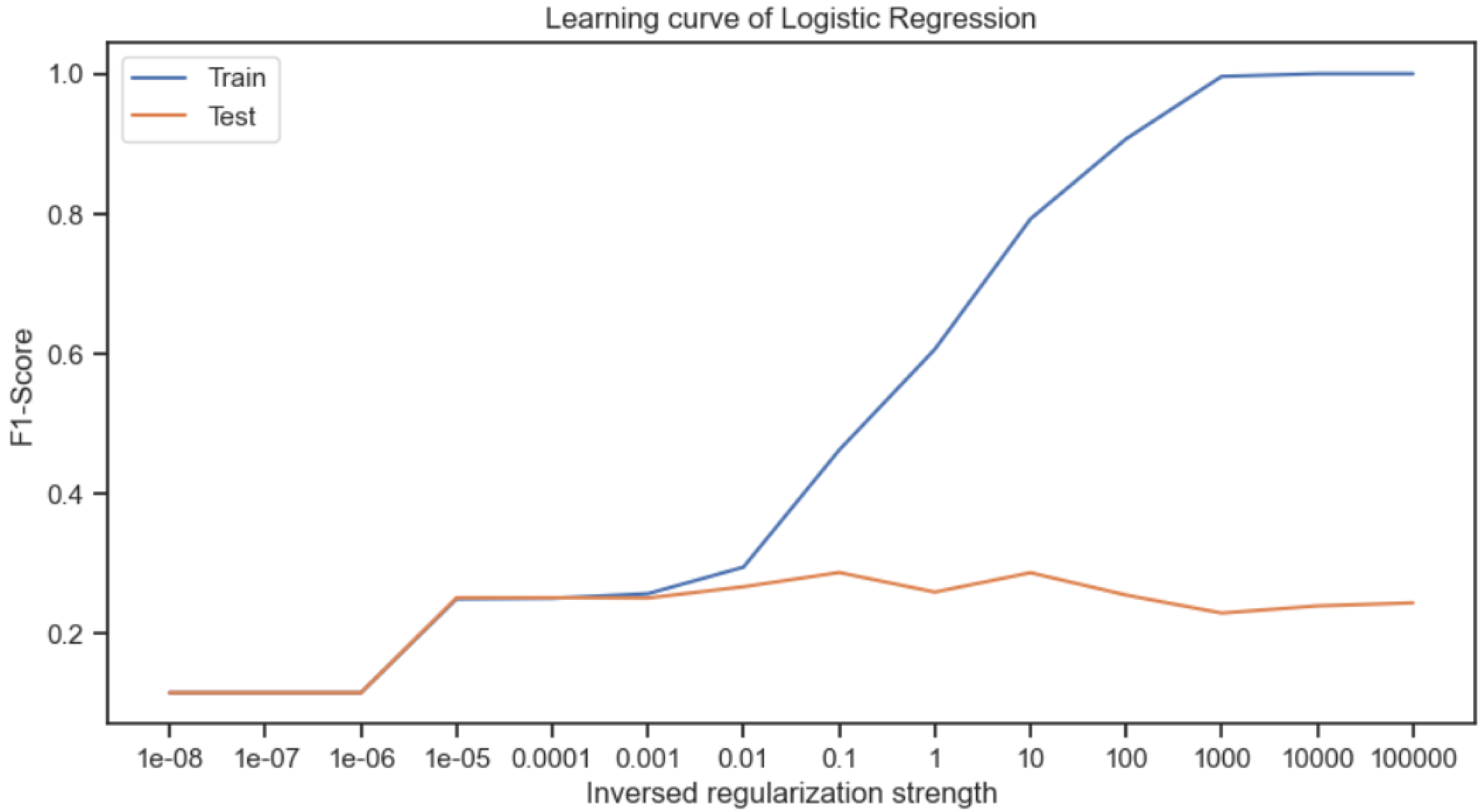


	Number	Name	Type_1	Type_2	Total	HP	Attack	Defense	Sp_Atk	Sp_Def	Speed	Generation	isLegendary	Color	hasGender	Pr_Male	Egg_Grou
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False	Green	True	0.875	Mon
1	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False	Green	True	0.875	Mon
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False	Green	True	0.875	Mon
3	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False	Red	True	0.875	Mon
4	5	Charmeleon	Fire	NaN	405	58	64	58	80	65	80	1	False	Red	True	0.875	Mon

Preprocessing steps

- Train-Validation-Test split
- OHE categorical features
- Scaling numerical features

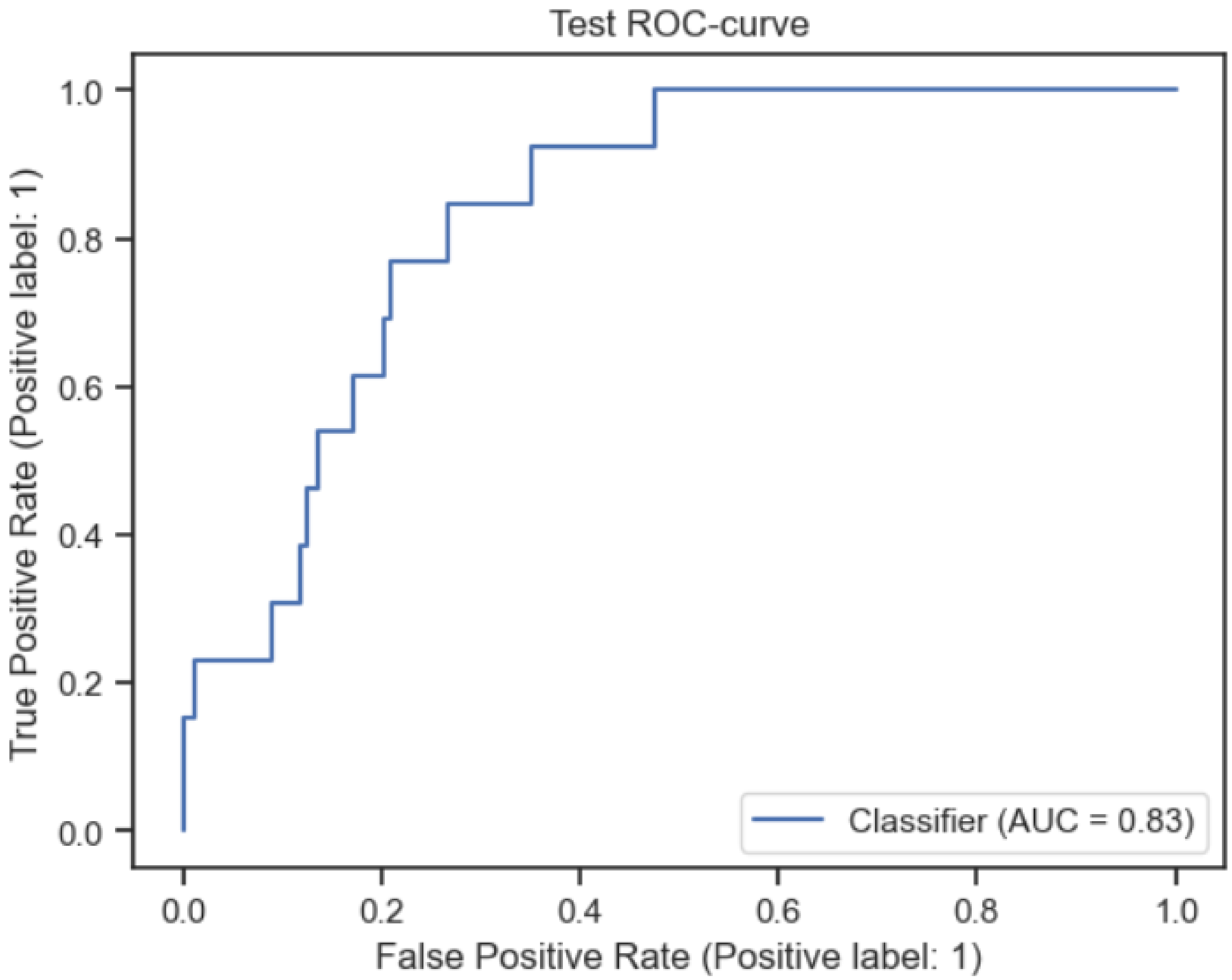
Hyperparameter tuning



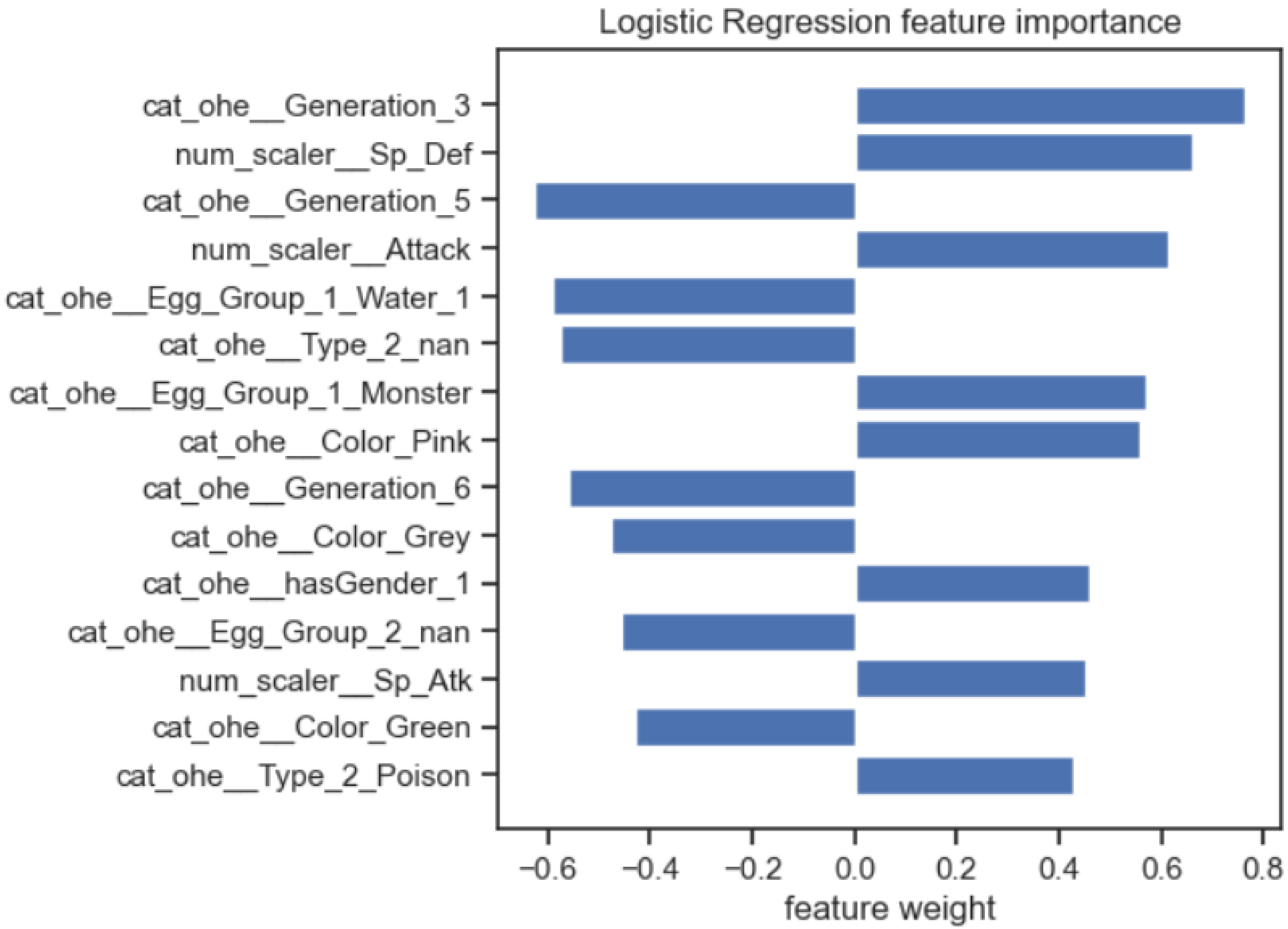
Test results

```
   precision    recall  f1-score   support

0       0.95      0.88      0.91       168
1       0.19      0.38      0.26        13
```

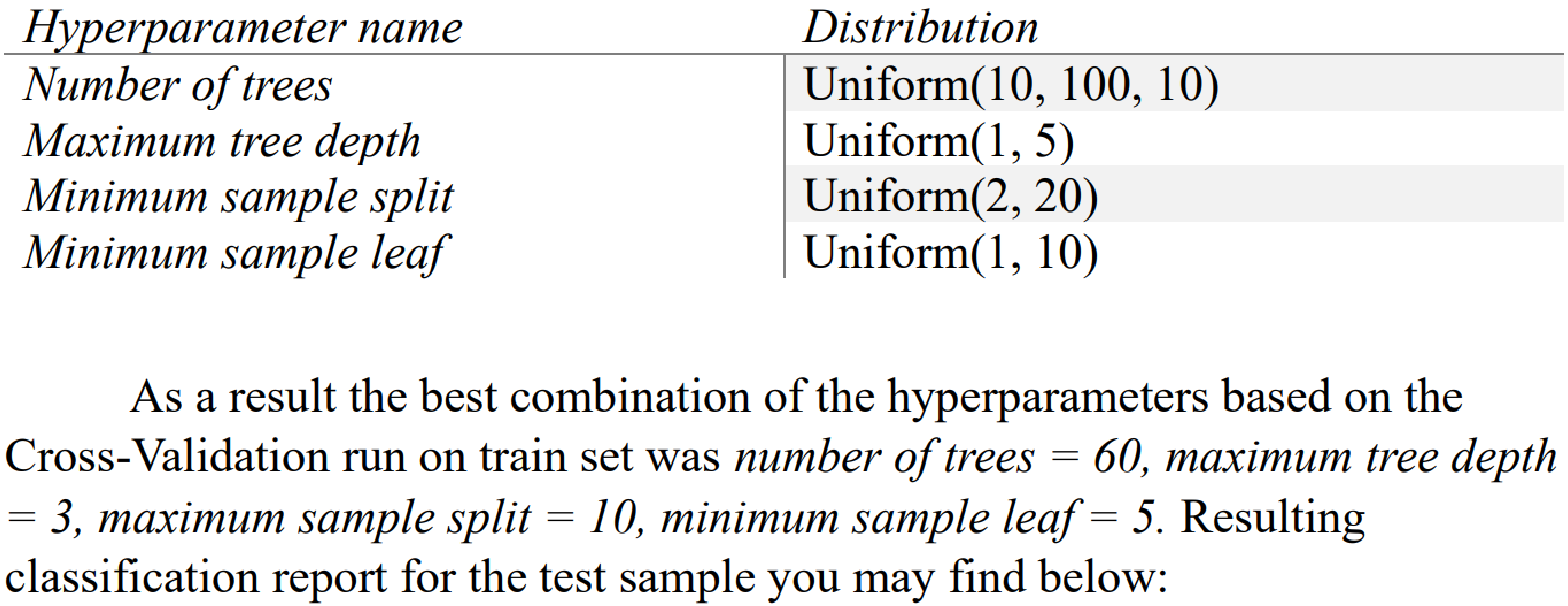


Feature importance



Random Forest Classifier

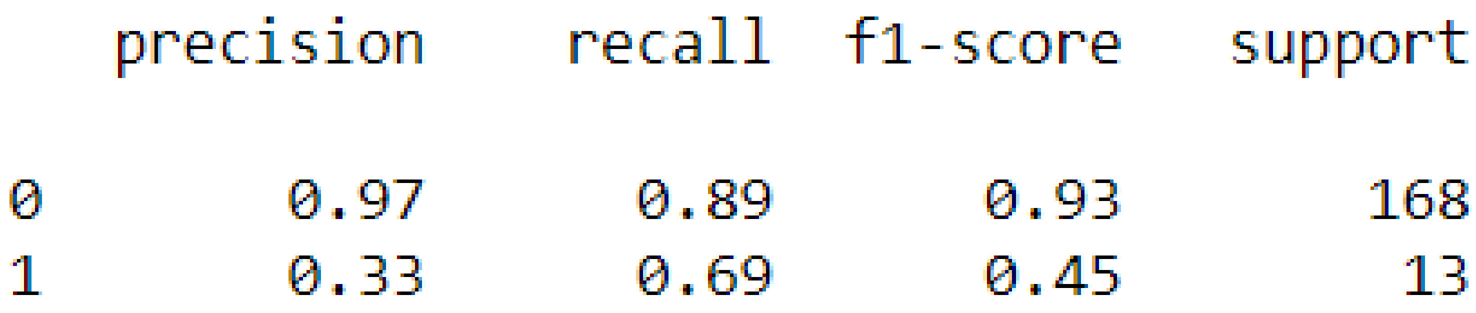
Hyperparameter tuning



Hyperparameter name	*Distribution*
Number of trees	Uniform(10, 100, 10)
Maximum tree depth	Uniform(1, 5)
Minimum sample split	Uniform(2, 20)
Minimum sample leaf	Uniform(1, 10)

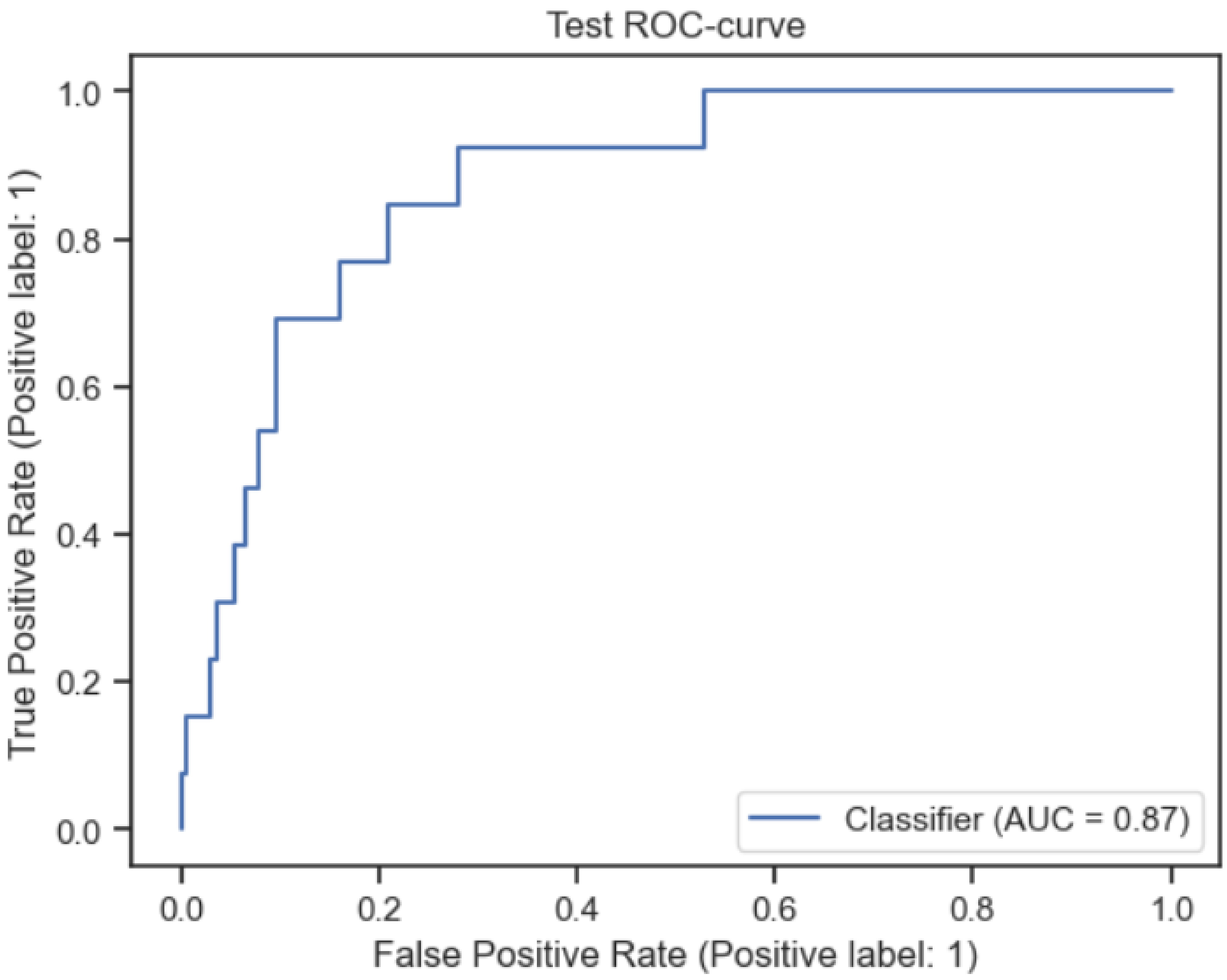
As a result the best combination of the hyperparameters based on the Cross-Validation run on train set was *number of trees = 60, maximum tree depth = 3, maximum sample split = 10, minimum sample leaf = 5*. Resulting classification report for the test sample you may find below:

Test results

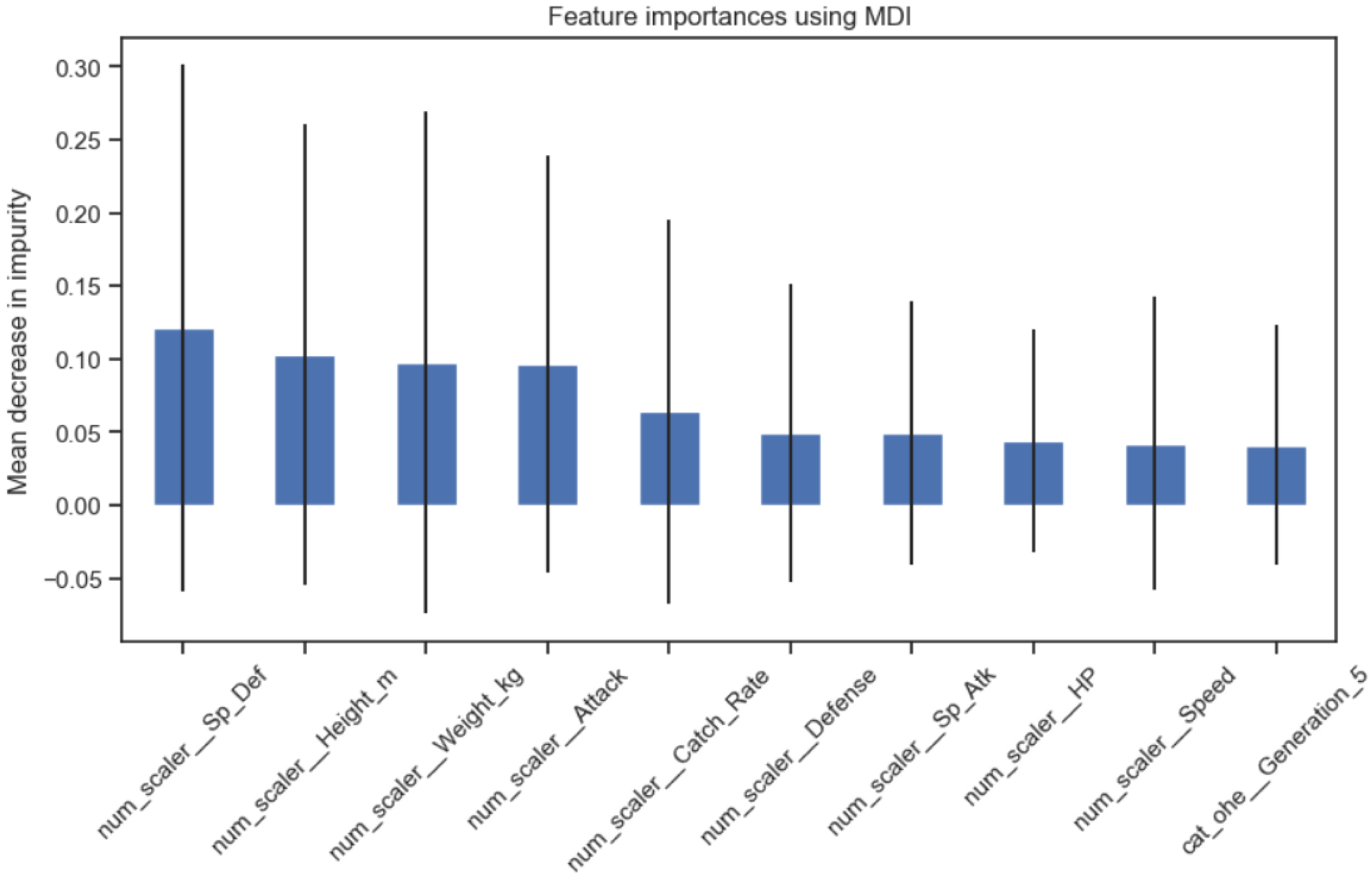


```
   precision    recall  f1-score   support

0       0.97      0.89      0.93       168
1       0.33      0.69      0.45        13
```

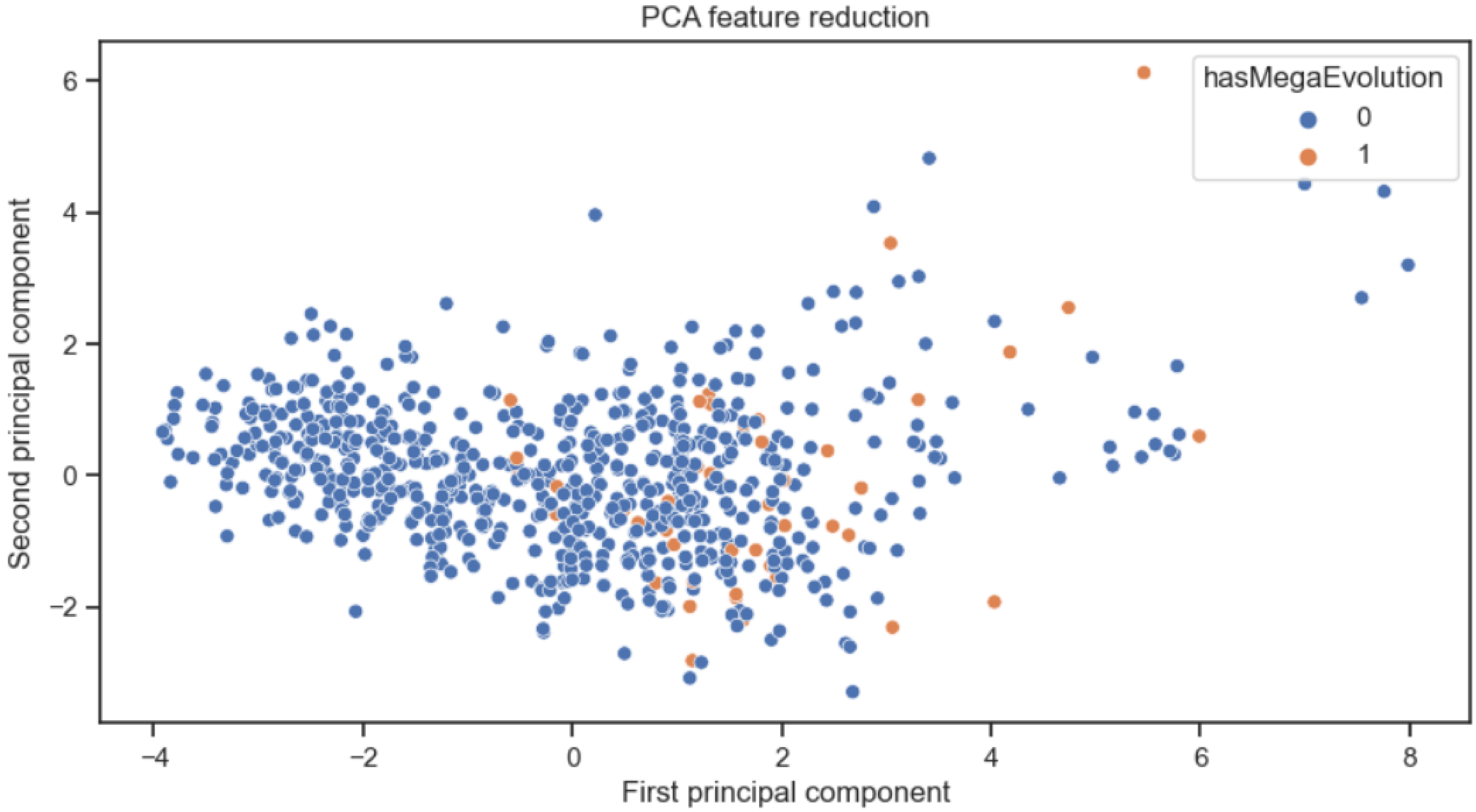


Feature importance

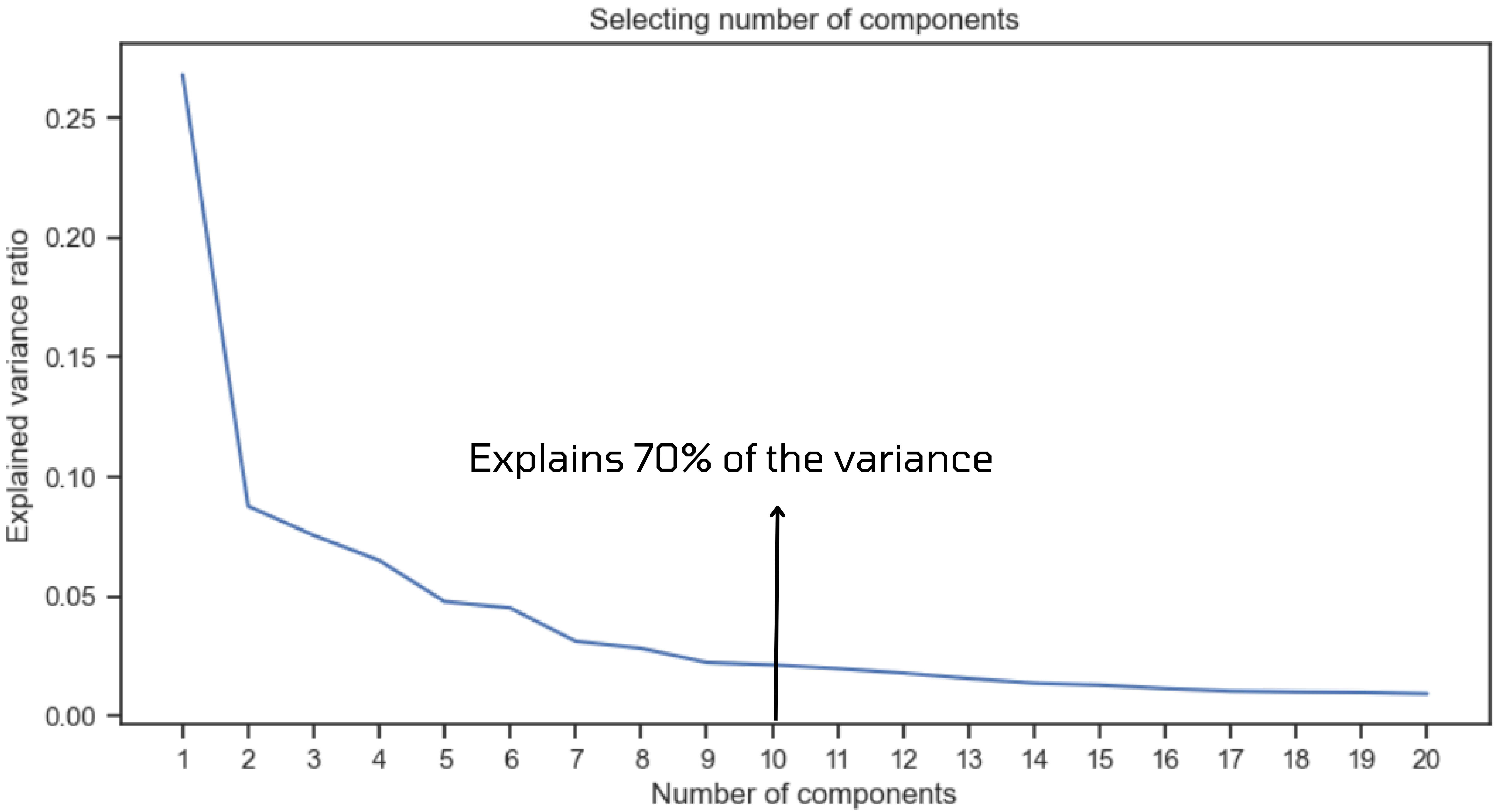


PCA and clustering

PCA target

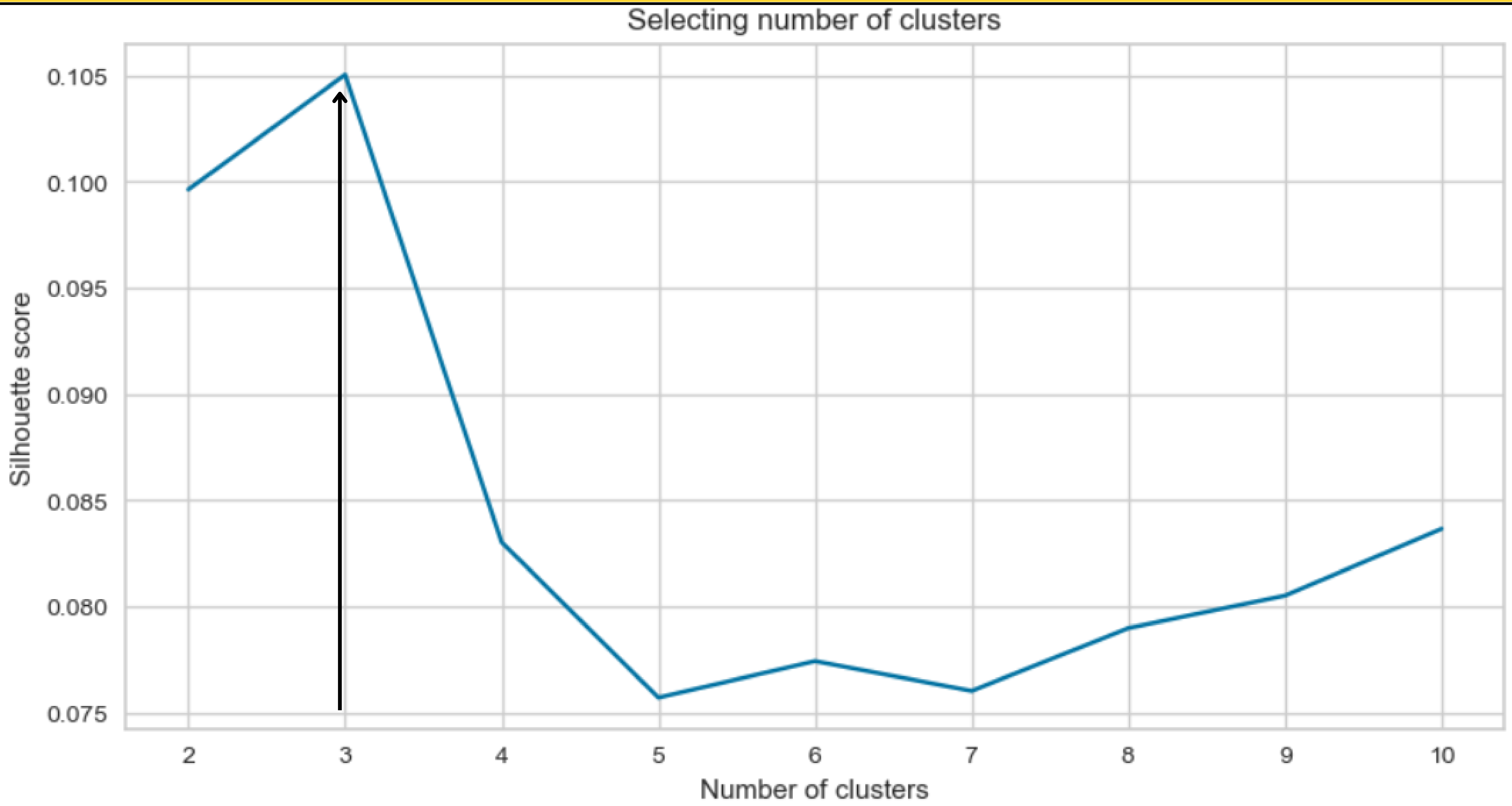

PCA feature reduction
Second principal component
First principal component
hasMegaEvolution
● 0
● 1
6
4
2
0
−2
−4
−2
0
2
4
6
8

Selecting number of comp

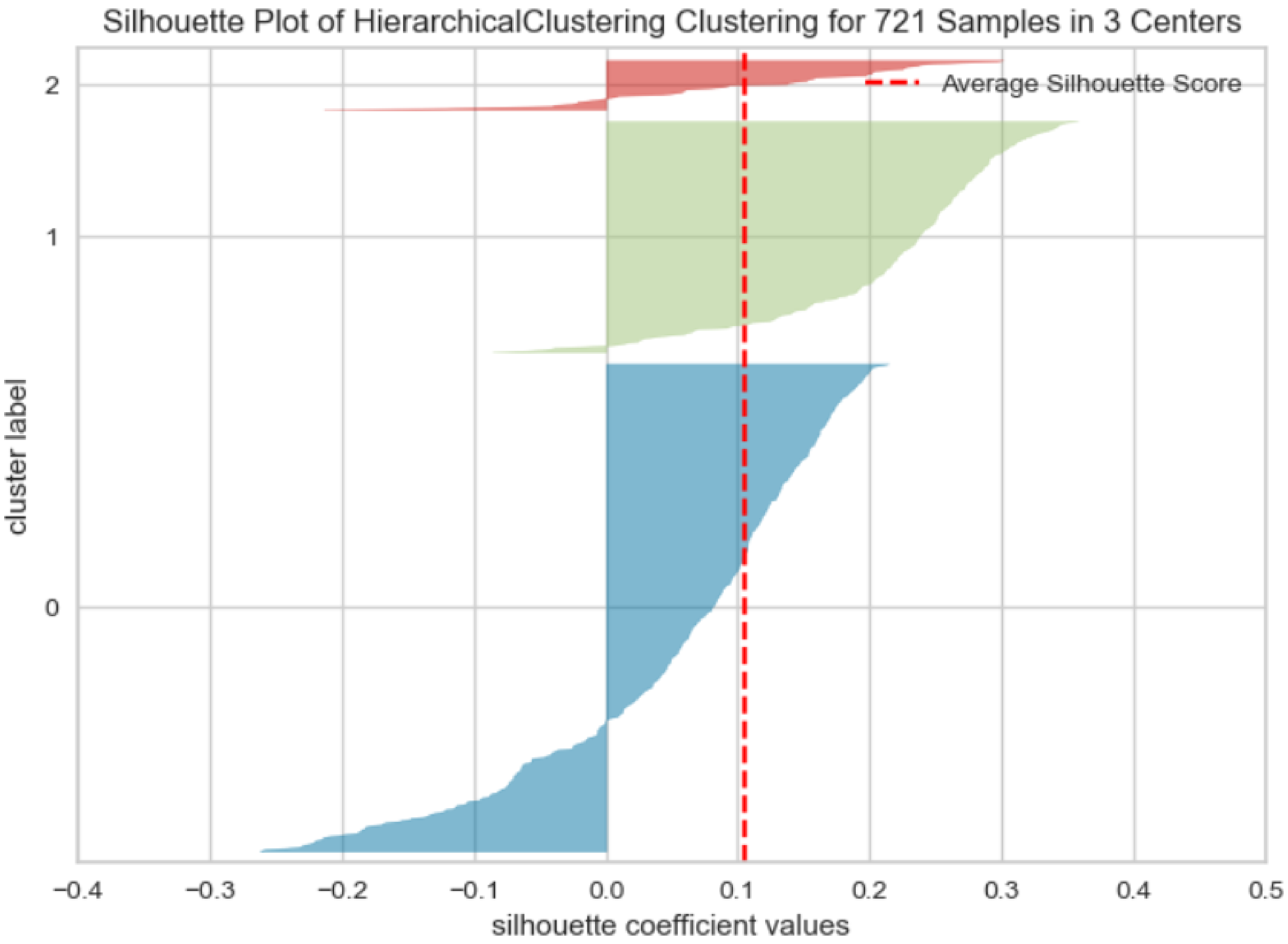


Explains 70% of the variance

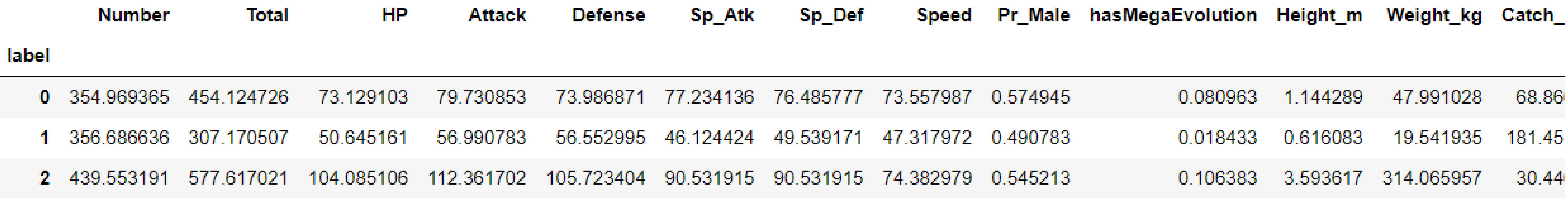
Hierarchical clustering



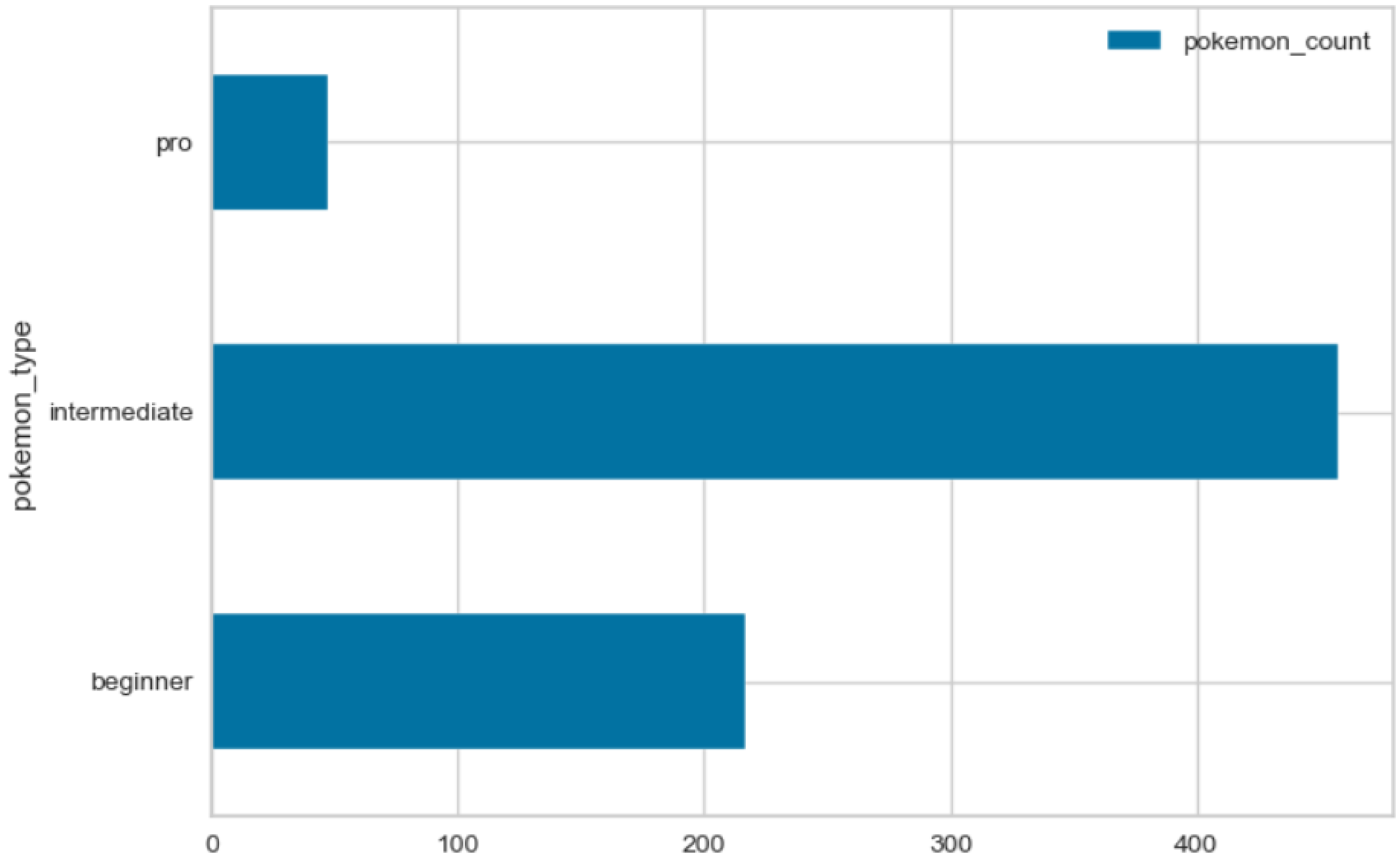
Silhouette plot



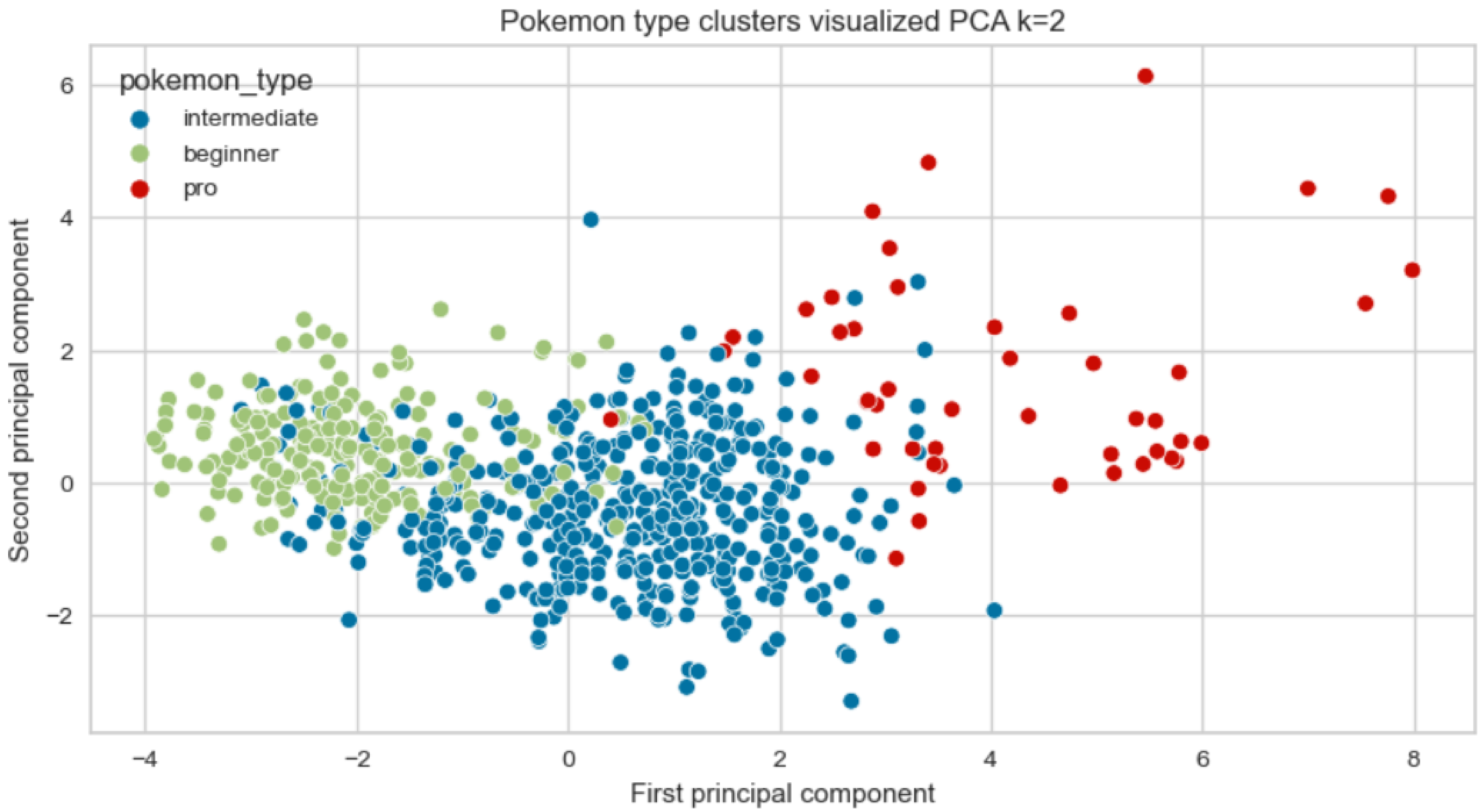
Clustering results



	Number	Total	HP	Attack	Defense	Sp_Atk	Sp_Def	Speed	Pr_Male	hasMegaEvolution	Height_m	Weight_kg	Catch_
label													
0	354.969365	454.124726	73.129103	79.730853	73.986871	77.234136	76.485777	73.557987	0.574945	0.080963	1.144289	47.991028	68.86
1	356.686636	307.170507	50.645161	56.990783	56.552995	46.124424	49.539171	47.317972	0.490783	0.018433	0.616083	19.541935	181.45
2	439.553191	577.617021	104.085106	112.361702	105.723404	90.531915	90.531915	74.382979	0.545213	0.106383	3.593617	314.065957	30.44



PCA clusters



Pokemon type clusters visualized PCA k=2
pokemon_type
● intermediate
● beginner
● pro
Second principal component
First principal component