Задача классификации

Датасет: Abalone

Постановка задачи

Целевая задача: классическая задачи мультиклассовой классификации

Решение:

Разбить количество колец на классы по возрастам, где количесво < 9 — yong, между 9 и 10 — medium, >10 — old

Информация по датасету

Variable Name	Role	Type	Description	Units	Missing Values
Sex	Feature	Categorical	M, F, and I (infant)		no
Length	Feature	Continuous	Longest shell measurement	mm	no
Diameter	Feature	Continuous	perpendicular to length	mm	no
Height	Feature	Continuous	with meat in shell	mm	no
Whole_weight	Feature	Continuous	whole abalone	grams	no
Shucked_weight	Feature	Continuous	weight of meat	grams	no
Viscera_weight	Feature	Continuous	gut weight (after bleeding)	grams	no
Shell_weight	Feature	Continuous	after being dried	grams	no
Rings	Target	Integer	+1.5 gives the age in years		no

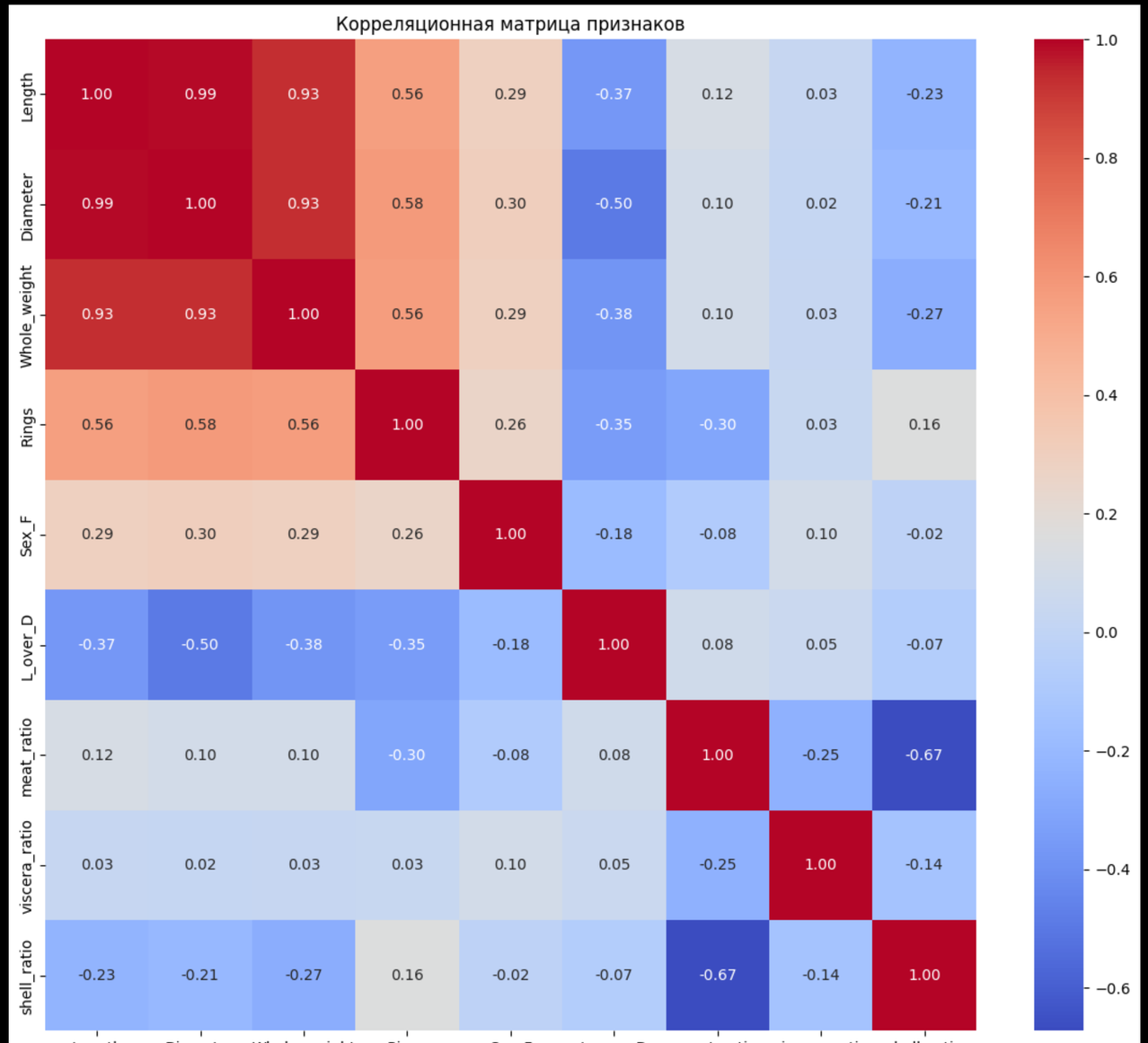
Базовая статистика до обработки

	count	mean	std	min	25%	50%	75%	max
Length	4177.0	0.523992	0.120093	0.0750	0.4500	0.5450	0.615	0.8150
Diameter	4177.0	0.407881	0.099240	0.0550	0.3500	0.4250	0.480	0.6500
Height	4177.0	0.139516	0.041827	0.0000	0.1150	0.1400	0.165	1.1300
Whole_weight	4177.0	0.828742	0.490389	0.0020	0.4415	0.7995	1.153	2.8255
Shucked_weight	4177.0	0.359367	0.221963	0.0010	0.1860	0.3360	0.502	1.4880
Viscera_weight	4177.0	0.180594	0.109614	0.0005	0.0935	0.1710	0.253	0.7600
Shell_weight	4177.0	0.238831	0.139203	0.0015	0.1300	0.2340	0.329	1.0050
Rings	4177.0	9.933684	3.224169	1.0000	8.0000	9.0000	11.000	29.0000

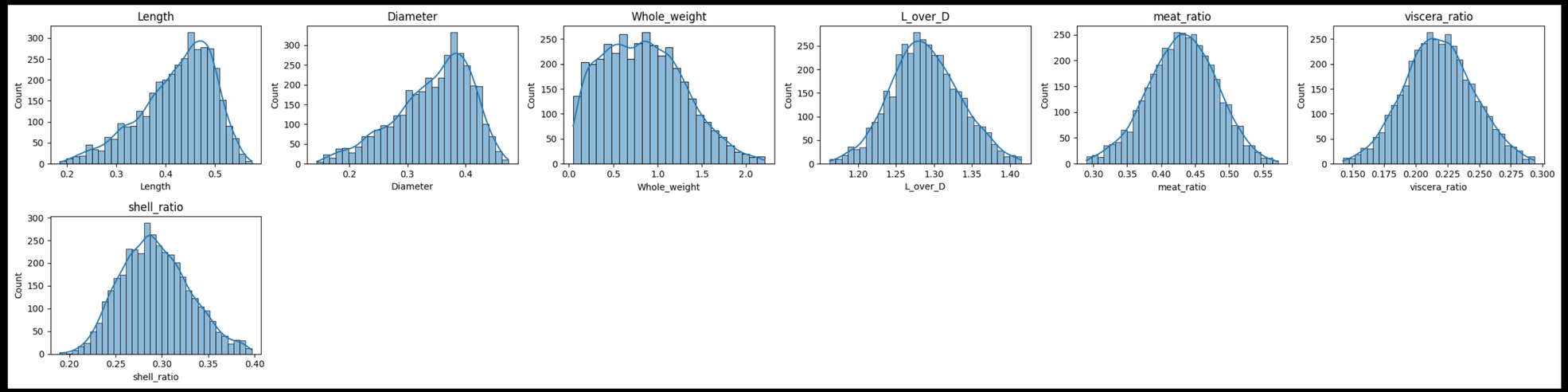
Базовая статистика после обработки

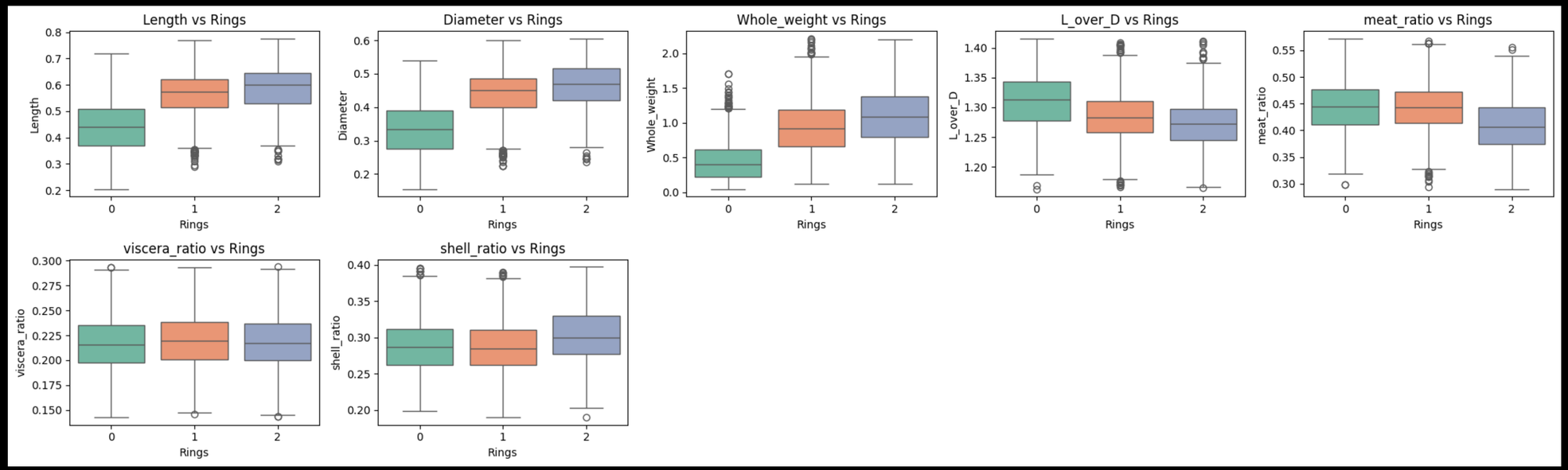
	count	mean	std	min	25%	50%	75%	max
Length	4177.0	0.523992	0.120093	0.075000	0.450000	0.545000	0.615000	0.815000
Diameter	4177.0	0.407881	0.099240	0.055000	0.350000	0.425000	0.480000	0.650000
Height	4177.0	0.139257	0.038359	0.010000	0.115000	0.140000	0.165000	0.250000
Whole_weight	4177.0	0.828742	0.490389	0.002000	0.441500	0.799500	1.153000	2.825500
Shucked_weight	4177.0	0.359367	0.221963	0.001000	0.186000	0.336000	0.502000	1.488000
Viscera_weight	4177.0	0.180594	0.109614	0.000500	0.093500	0.171000	0.253000	0.760000
Shell_weight	4177.0	0.238831	0.139203	0.001500	0.130000	0.234000	0.329000	1.005000
Rings	4177.0	9.933684	3.224169	1.000000	8.000000	9.000000	11.000000	29.000000
L_over_D	4177.0	1.291880	0.059065	0.493333	1.257732	1.288462	1.321839	2.333333
meat_ratio	4177.0	0.432414	0.105765	0.175258	0.395100	0.430592	0.466175	4.691943
viscera_ratio	4177.0	0.218537	0.034361	0.007634	0.198586	0.217252	0.236945	0.665399
shell_ratio	4177.0	0.295605	0.058785	0.109341	0.266097	0.290870	0.319410	2.615672

Корреляционная матрица



Анализ распределения





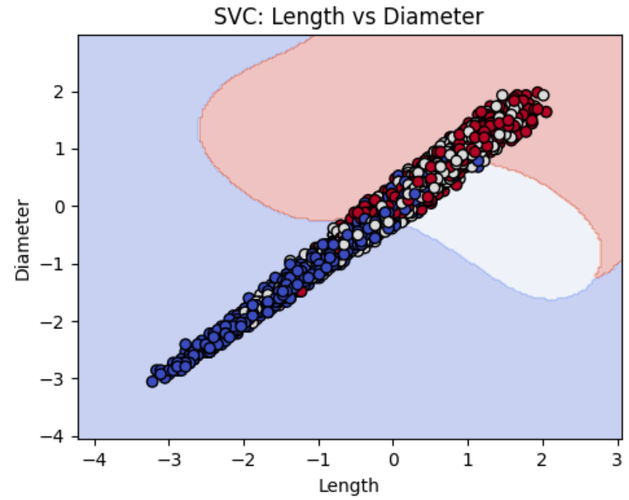
Результаты обучения

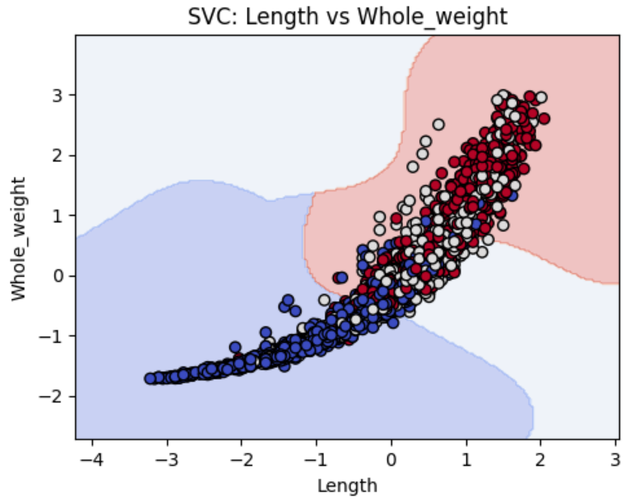
```
=== SVM ===
Accuracy: 0.646 ± 0.018
F1-macro: 0.647 ± 0.017
ROC-AUC: 0.823 ± 0.012
=== Logistic Regression ===
Accuracy: 0.651 ± 0.013
F1-macro: 0.650 ± 0.012
ROC-AUC: 0.830 ± 0.013
=== KNN ===
Accuracy: 0.608 ± 0.018
F1-macro: 0.608 ± 0.016
ROC-AUC: 0.774 ± 0.016
```

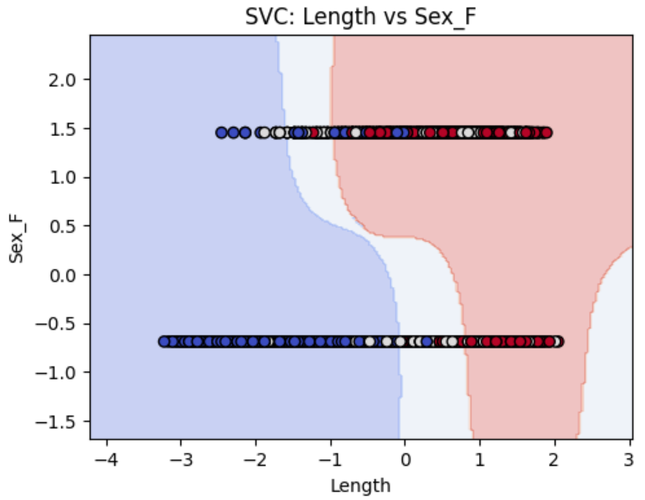
```
=== DesicionTree ===
Accuracy: 0.577 ± 0.009
F1-macro: 0.575 ± 0.009
ROC-AUC: 0.682 ± 0.007
=== RandomForest ===
Accuracy: 0.642 ± 0.017
F1-macro: 0.638 ± 0.016
ROC-AUC: 0.828 ± 0.010
```

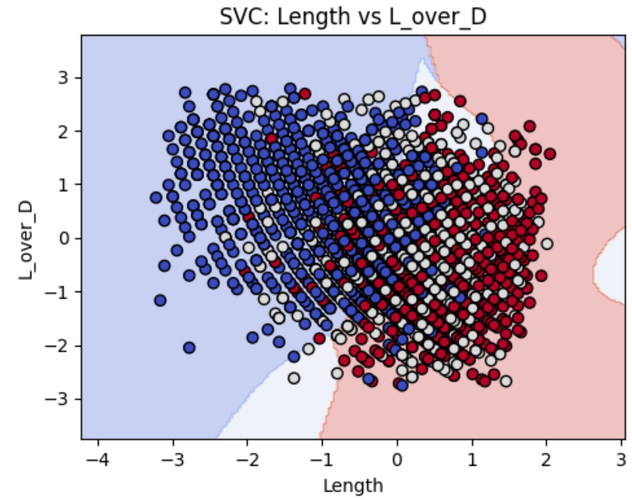
```
=== Boosting ===
Accuracy: 0.660 ± 0.016
F1-macro: 0.657 ± 0.016
ROC-AUC: 0.838 ± 0.013
=== XGBoost ===
Accuracy: 0.662 ± 0.015
F1-macro: 0.659 ± 0.014
ROC-AUC: 0.841 ± 0.013
=== CatBoost ===
Accuracy: 0.669 ± 0.016
F1-macro: 0.666 ± 0.015
ROC-AUC: 0.843 ± 0.011
```

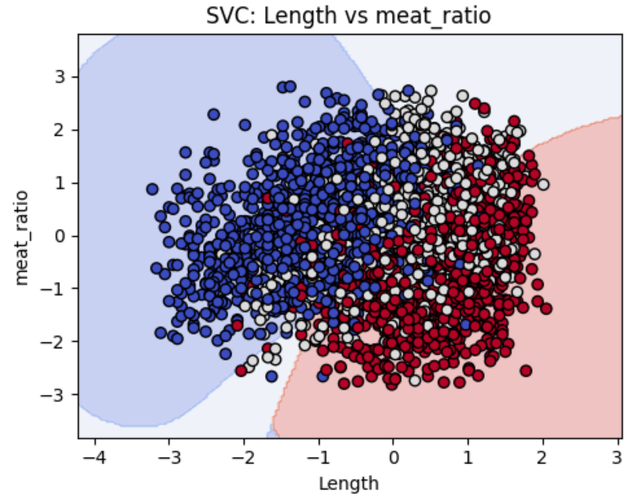
Графическое представление решения линейных моделей

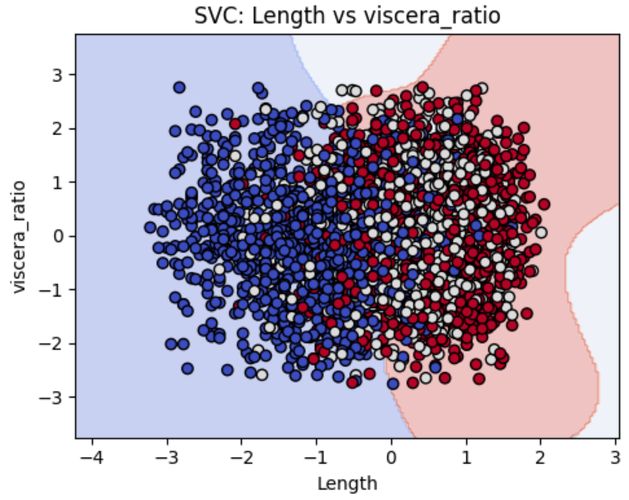


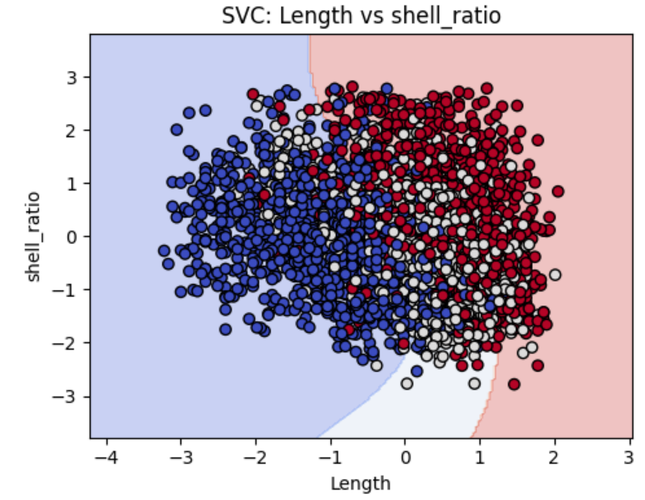




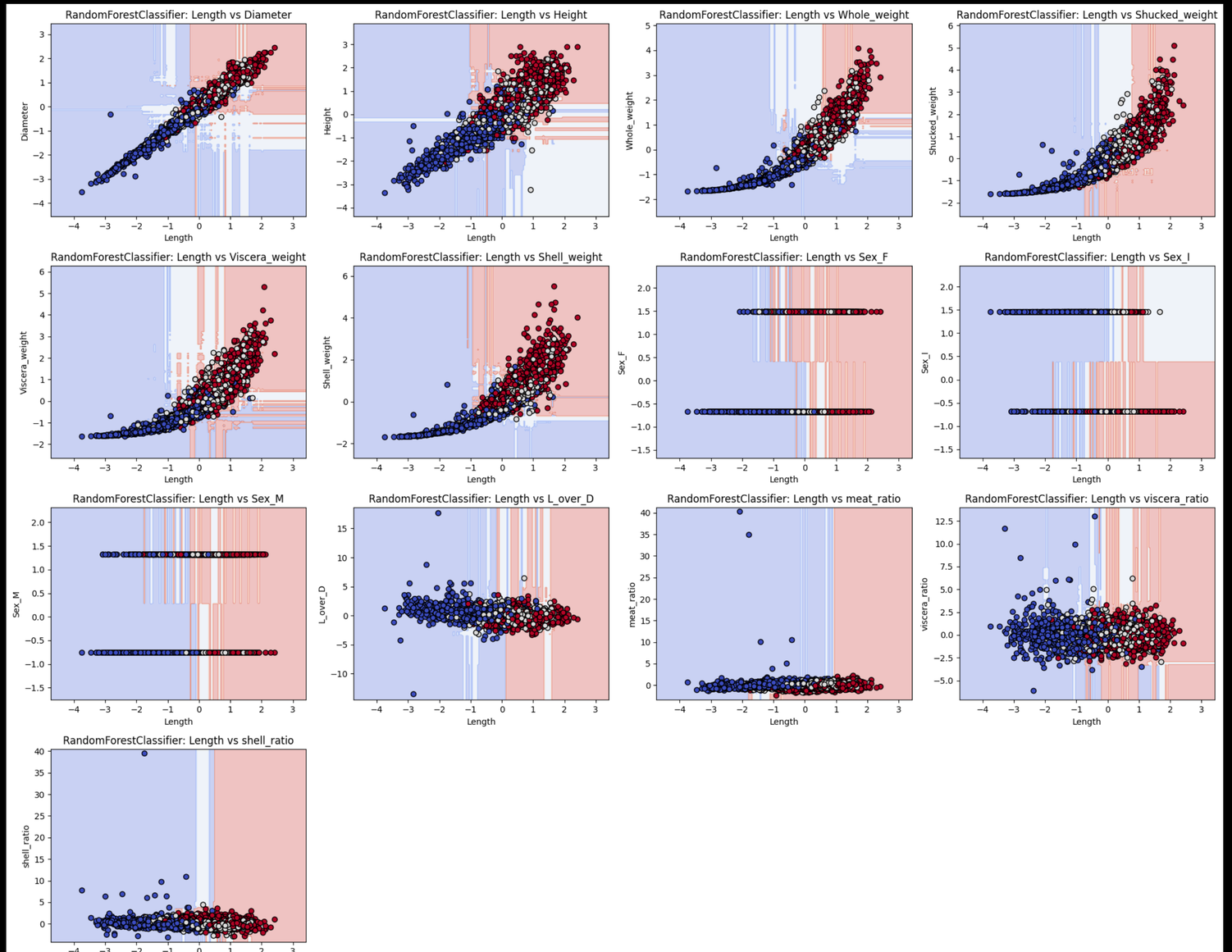








Графическое представление решения RF



Графическое представление решения CatBoost