Лабораторная работа №2

Тема: *Анализ и сравнение разных классификаторов*

Состав команды:

Найпак Дмитрий
Написание логистической регрессии, подготовка презентации и текста защиты

Усов Сергей
Teamlead, подготовка датасета, обучение стандартных моделей

Цели:

- Научиться обрабатывать датасет
- Реализовать разные классификаторы
- Понять, как интерпретировать важность признаков (feature importance)
- Понять, как увеличивать кол-во признаков (feature tuning)

Модели:

- Линейная регрессия
- Метод опорных векторов (SVM)
- K - ближайших соседей
- Дерево решений
- Случайный лес
- Градиентный бустинг

Постановка задачи

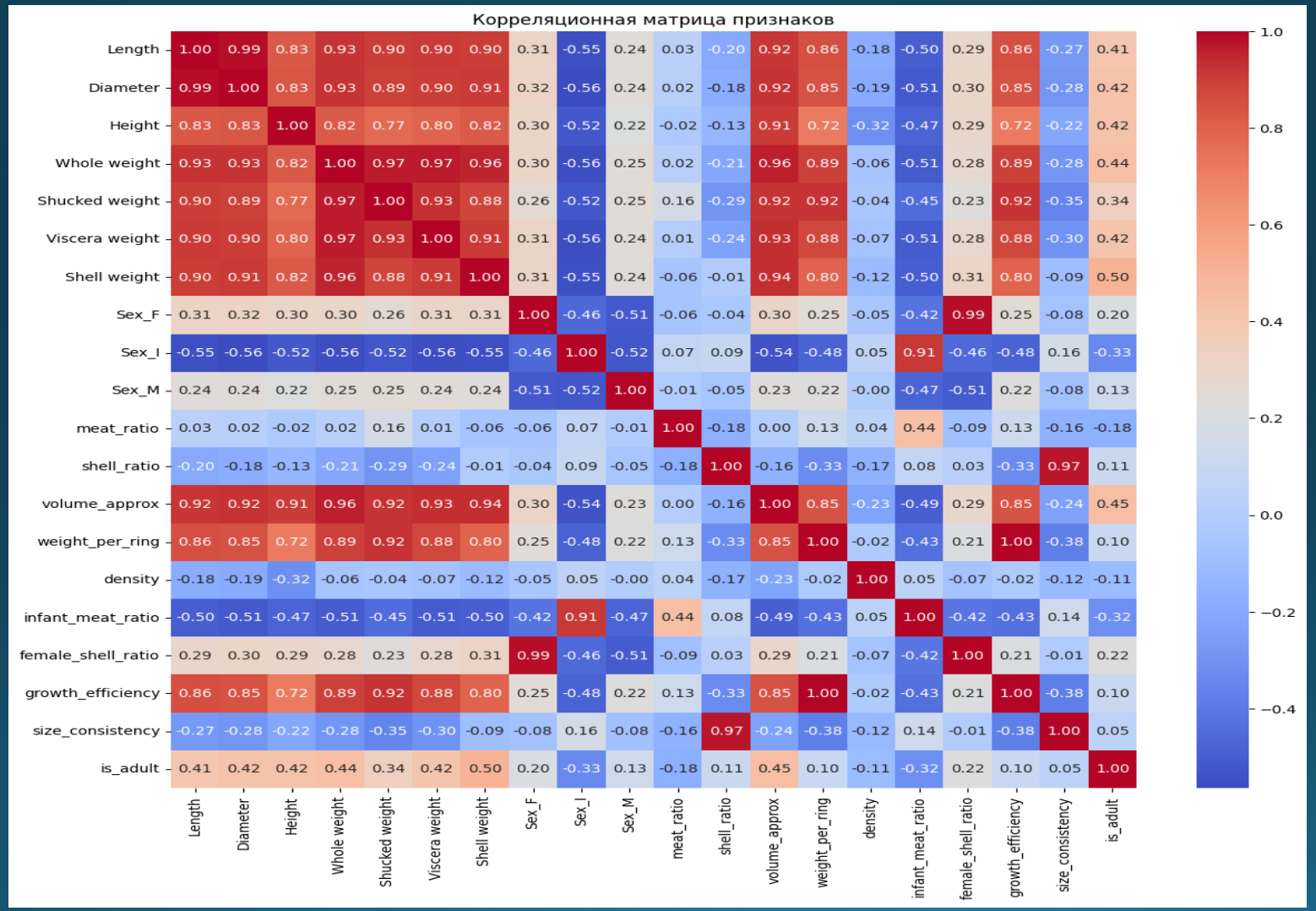
Целевая задача: задача классификации молюсков.

Мы можем решить её как:
Бинарная классификация: Молодые vs Взрослые. Rings <= 10 (возраст до 12 лет). Rings > 10 (возраст больше 12 лет).

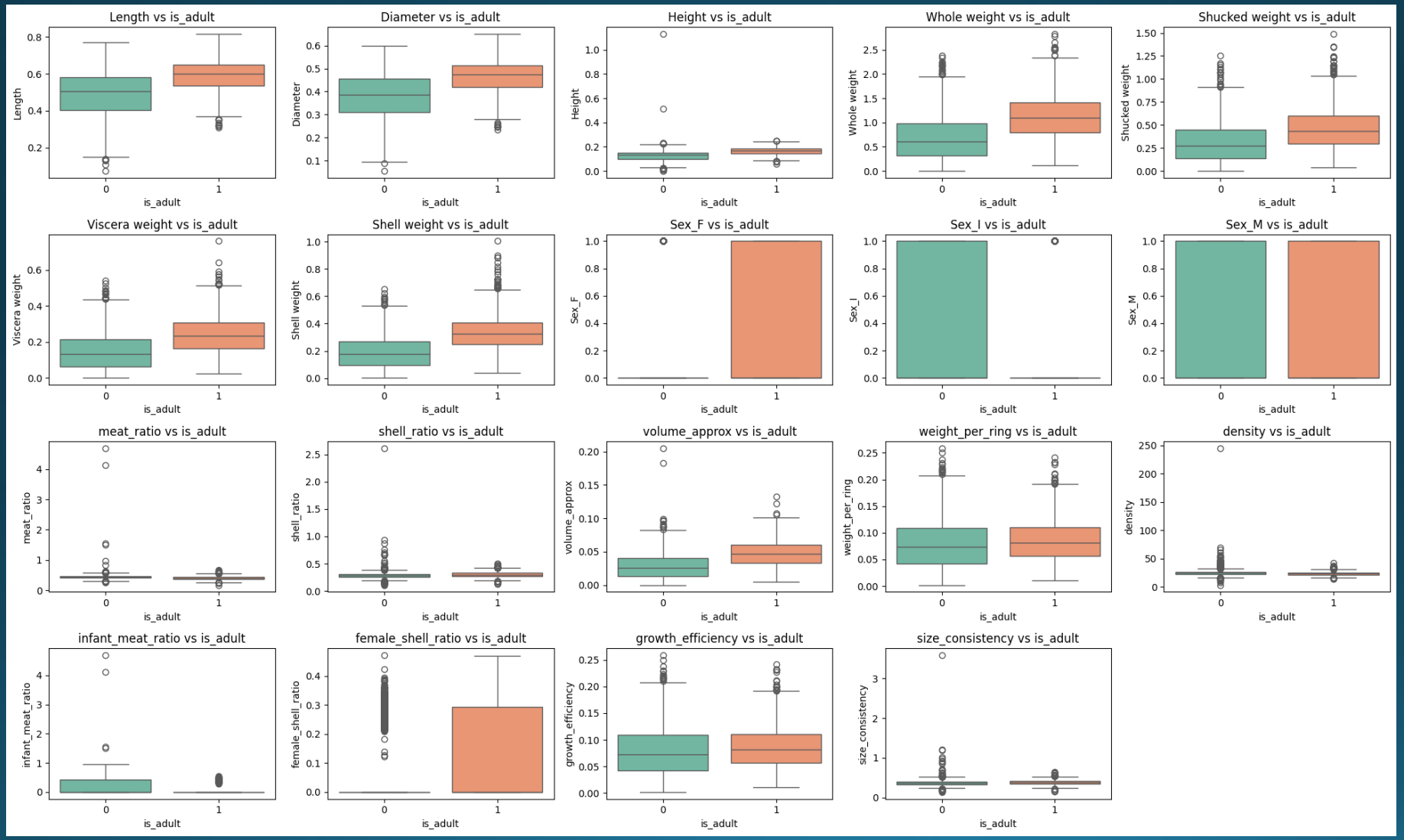
Базовая статистика в удобном виде:

	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings	Sex_F	Sex_I	Sex_M
count	4176.000000	4176.000000	4176.000000	4176.000000	4176.00000	4176.000000	4176.000000	4176.000000	4176.000000	4176.000000	4176.000000
mean	0.524009	0.407892	0.139527	0.828818	0.35940	0.180613	0.238852	9.932471	0.312979	0.321360	0.365661
std	0.120103	0.099250	0.041826	0.490424	0.22198	0.109620	0.139213	3.223601	0.463761	0.467055	0.481673
min	0.075000	0.055000	0.000000	0.002000	0.00100	0.000500	0.001500	1.000000	0.000000	0.000000	0.000000
25%	0.450000	0.350000	0.115000	0.441500	0.18600	0.093375	0.130000	8.000000	0.000000	0.000000	0.000000
50%	0.545000	0.425000	0.140000	0.799750	0.33600	0.171000	0.234000	9.000000	0.000000	0.000000	0.000000
75%	0.615000	0.480000	0.165000	1.153250	0.50200	0.253000	0.329000	11.000000	1.000000	1.000000	1.000000
max	0.815000	0.650000	1.130000	2.825500	1.48800	0.760000	1.005000	29.000000	1.000000	1.000000	1.000000

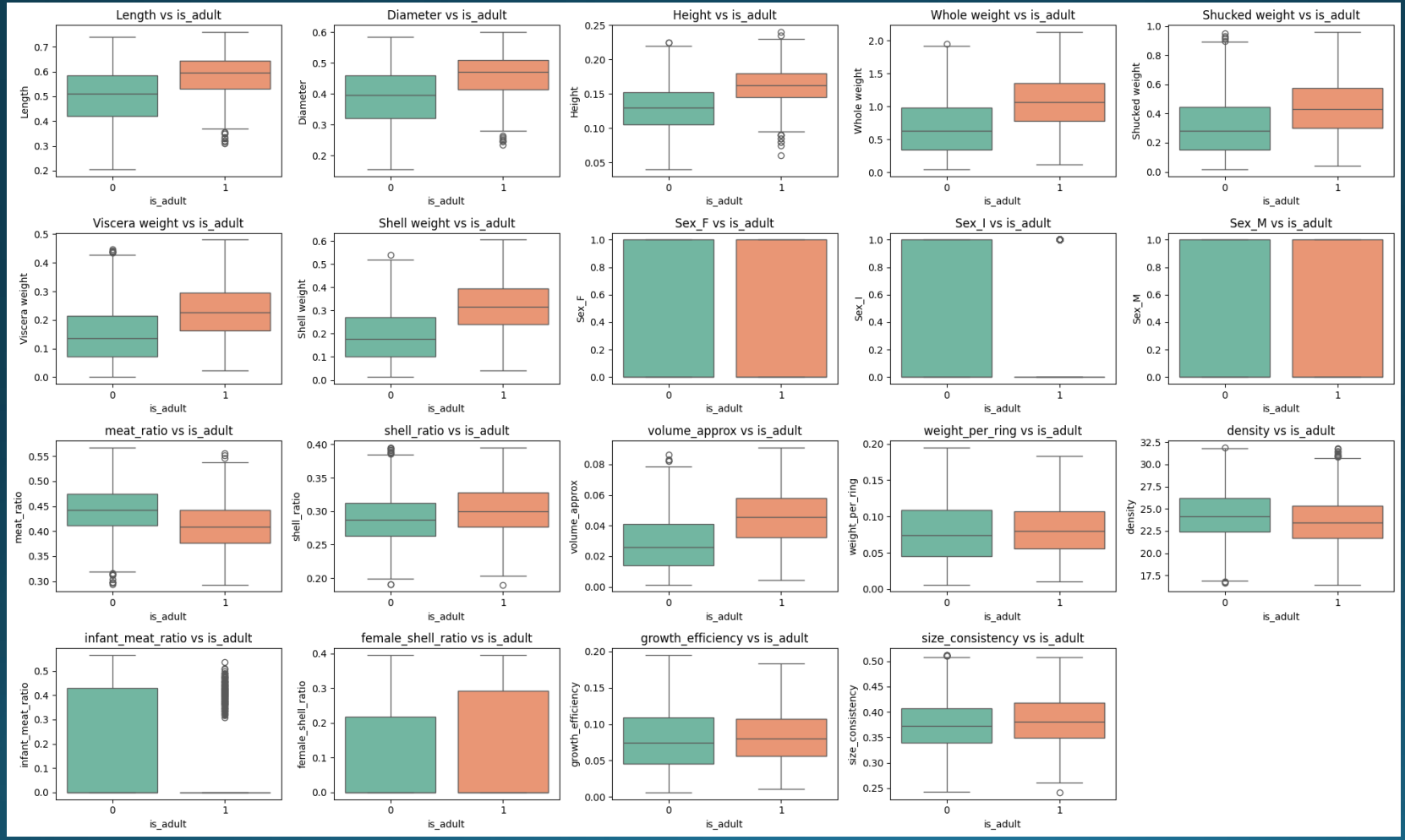
Корреляционная матрица



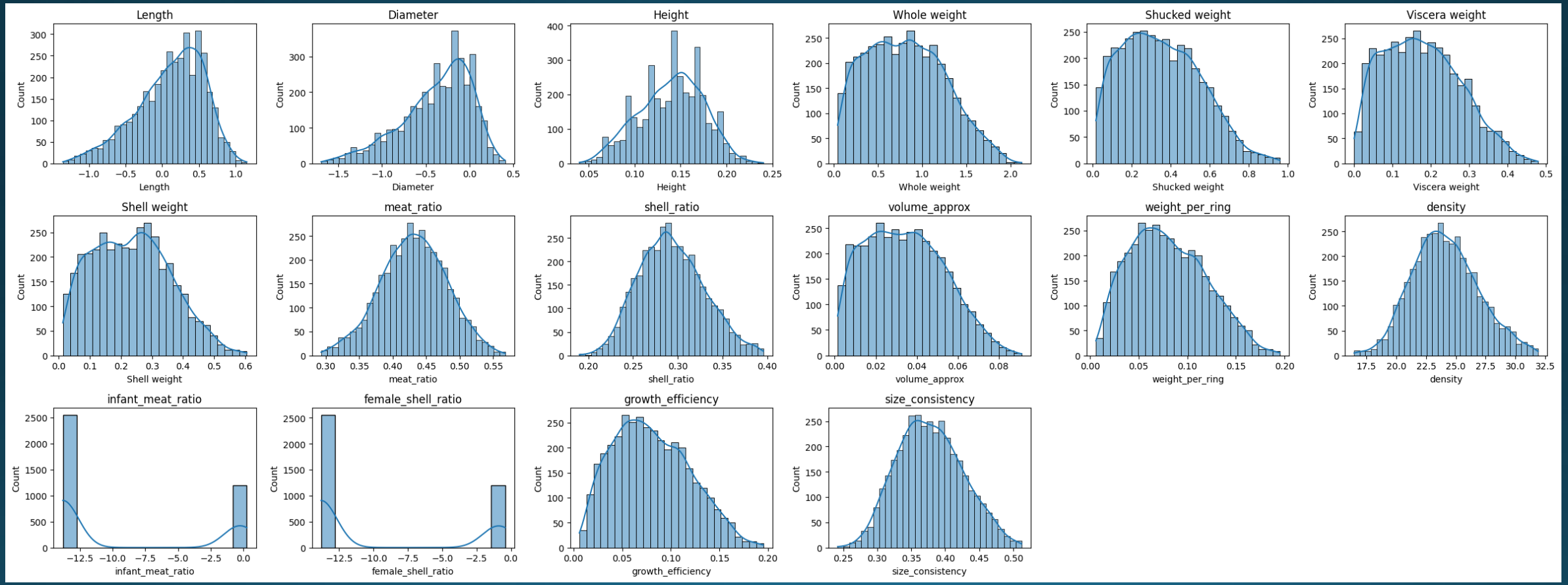
Выбросы переменных



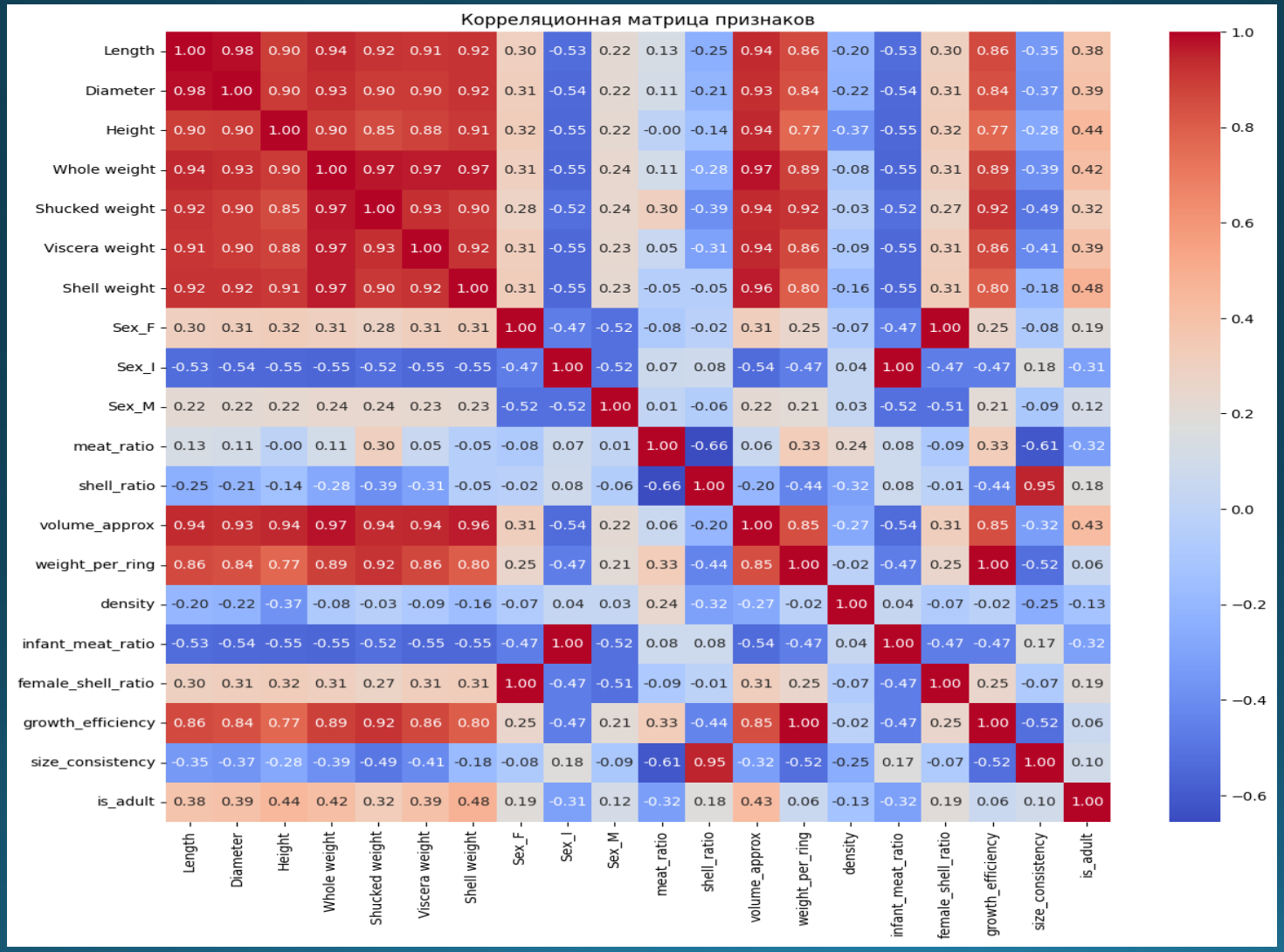
Удаление выбросов и выравнивание скосов



Гистограмма распределения после выравнивания

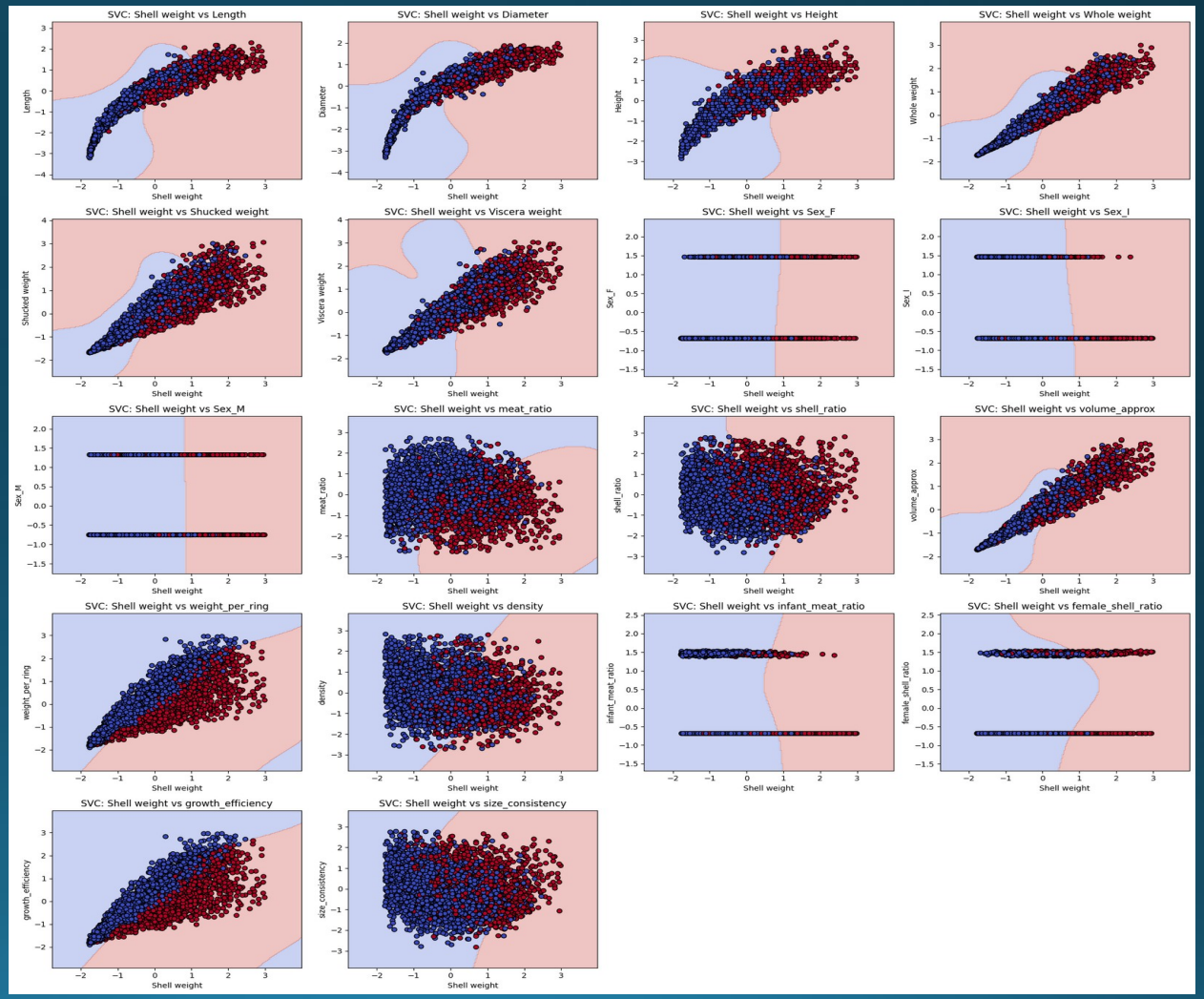


Результат подготовки данных



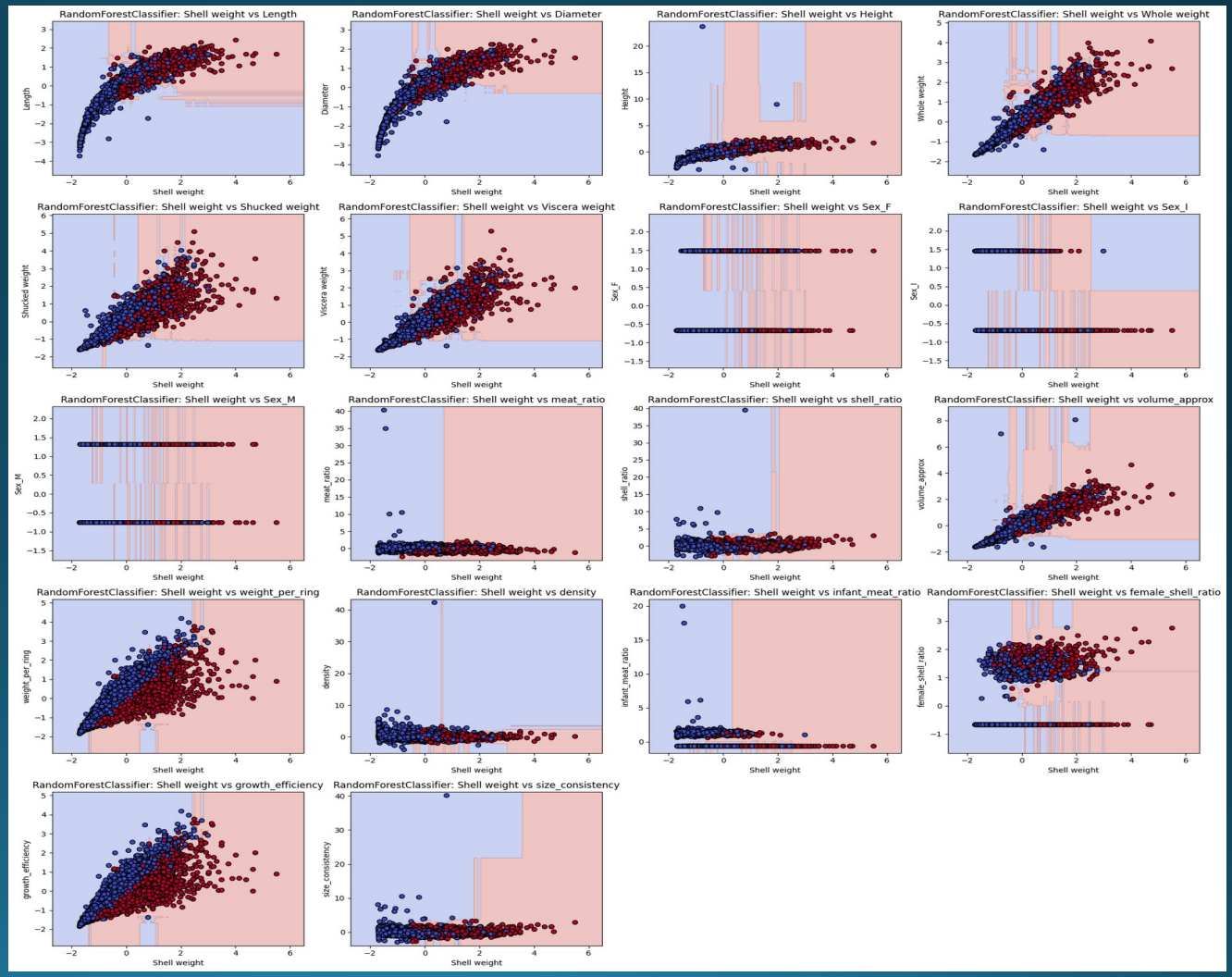
Обучение и тесты моделей {SVM, Logistic Regression, KNN}

```
=== SVM ===
Accuracy: 0.959 ± 0.006
F1-score: 0.936 ± 0.010
ROC-AUC: 0.996 ± 0.002
=== Logistic Regression ===
Accuracy: 0.954 ± 0.006
F1-score: 0.930 ± 0.010
ROC-AUC: 0.987 ± 0.003
=== KNN ===
Accuracy: 0.872 ± 0.004
F1-score: 0.799 ± 0.008
ROC-AUC: 0.926 ± 0.008
```



Обучение и тесты моделей {Desicion Tree, Random Forest}

```
=== DesicionTree ===
Accuracy: 0.949 ± 0.005
F1-score: 0.925 ± 0.008
ROC-AUC: 0.941 ± 0.008
=== RandomForest ===
Accuracy: 0.964 ± 0.003
F1-score: 0.947 ± 0.005
ROC-AUC: 0.993 ± 0.001
```



Реализация логистической регрессии и прогонка

```
=== LogisticRegressionCustom ===
Accuracy: 0.954 ± 0.007
F1-score: 0.931 ± 0.011
ROC-AUC: 0.988 ± 0.003
```