

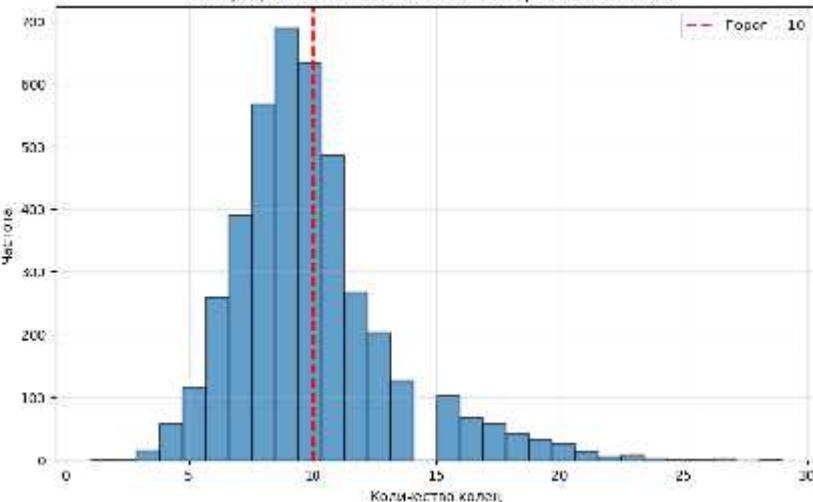


Название работы: "Классификация моллюсков Abalone"

Презентацию подготовила Глебова Мария

Группа: М8О-307Б-23

Распределение количества колец (возраста) моллюсков



Медиана колец: 9.0

Среднее значение колец: 9.93

Распределение по классам (порог = 10):

Незрелые ($\text{rings} < 10$): 2096 (50.2%)

Зрелые ($\text{rings} \geq 10$): 2081 (49.8%)



Преобразованная задача: бинарная классификация

Шаг 1: Класс 0: Незрелый моллюск ($\text{rings} < 10$)

Шаг 2

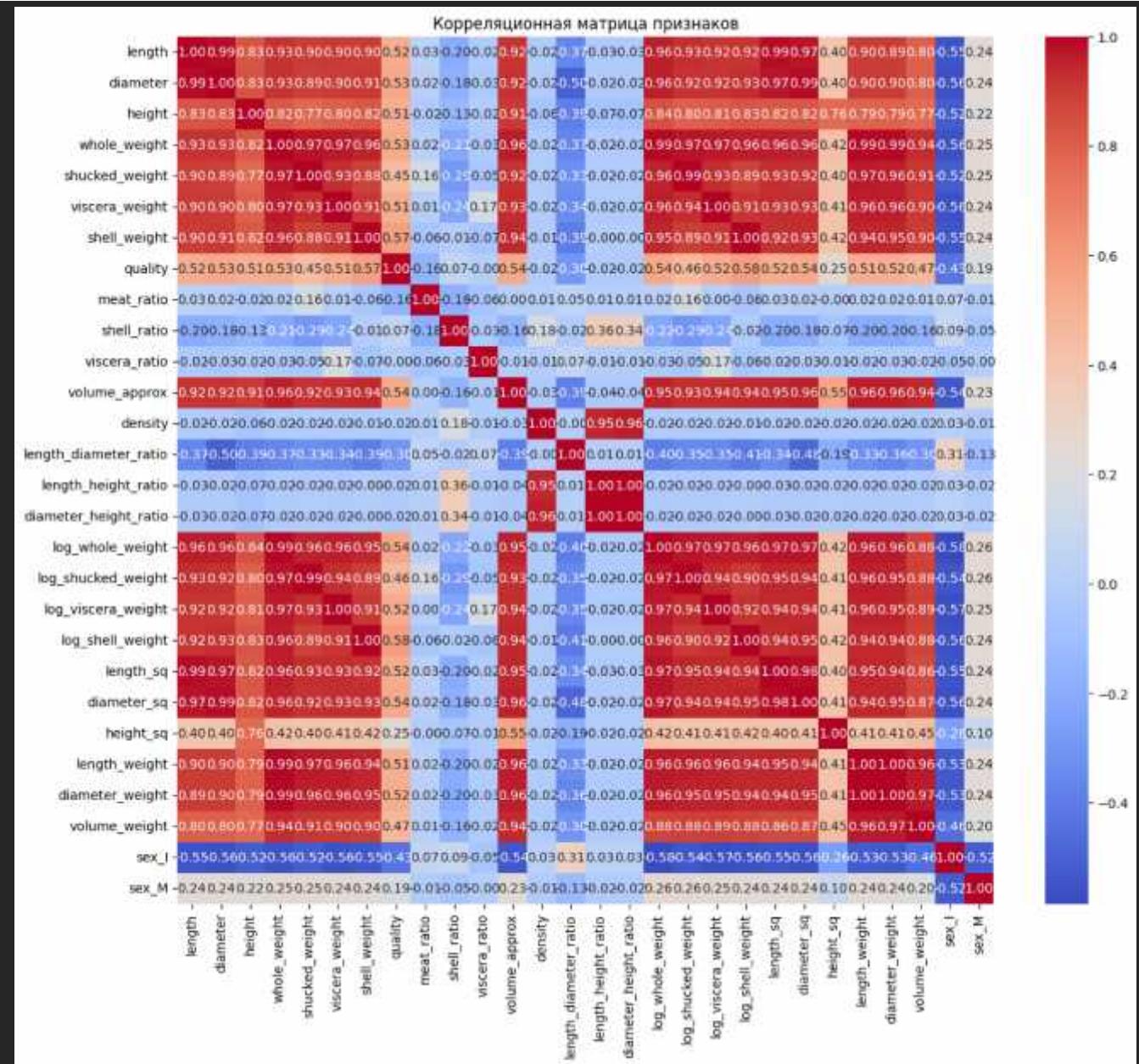
Класс 1: Зрелый моллюск ($\text{rings} \geq 10$)

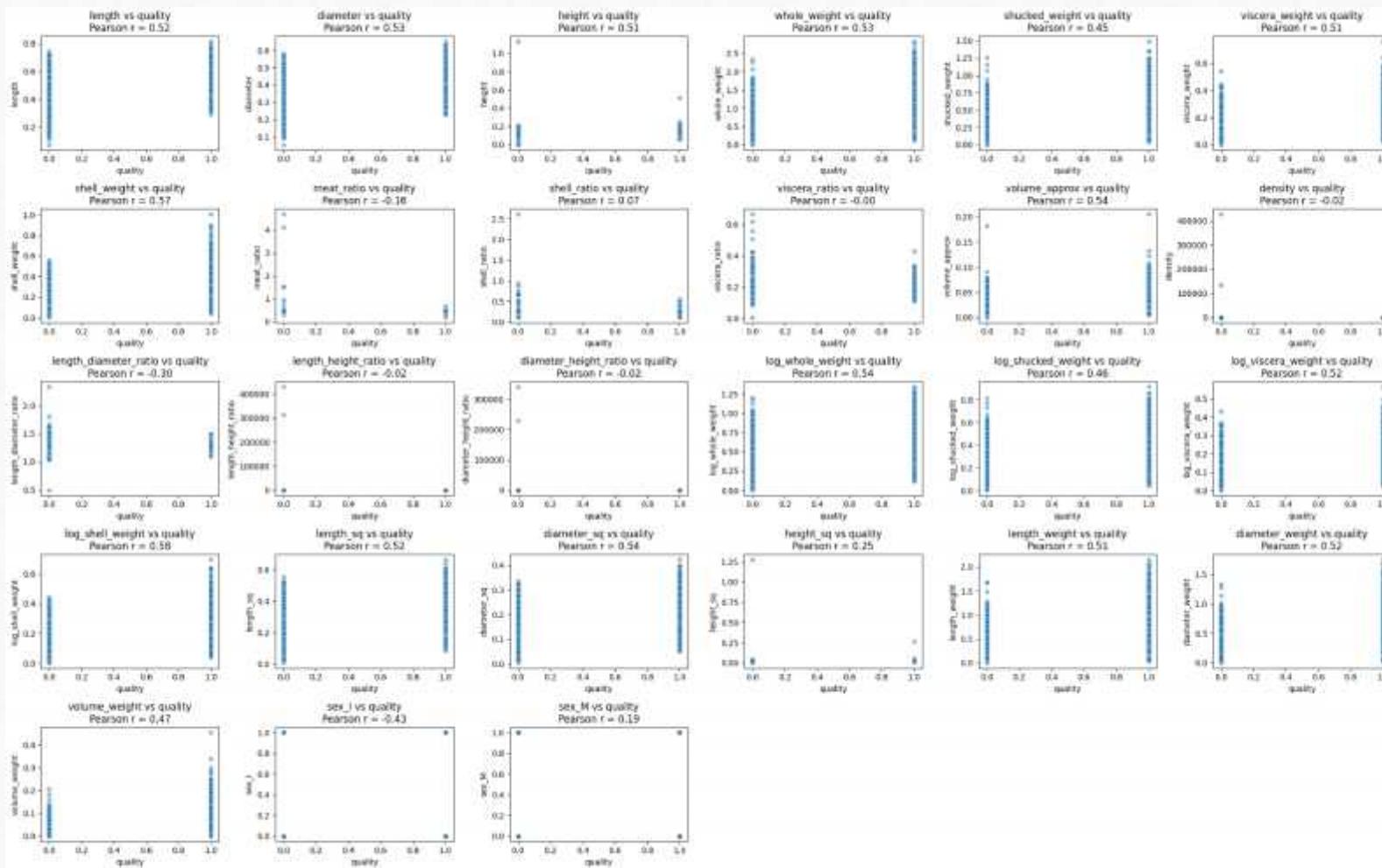
Шаг 3

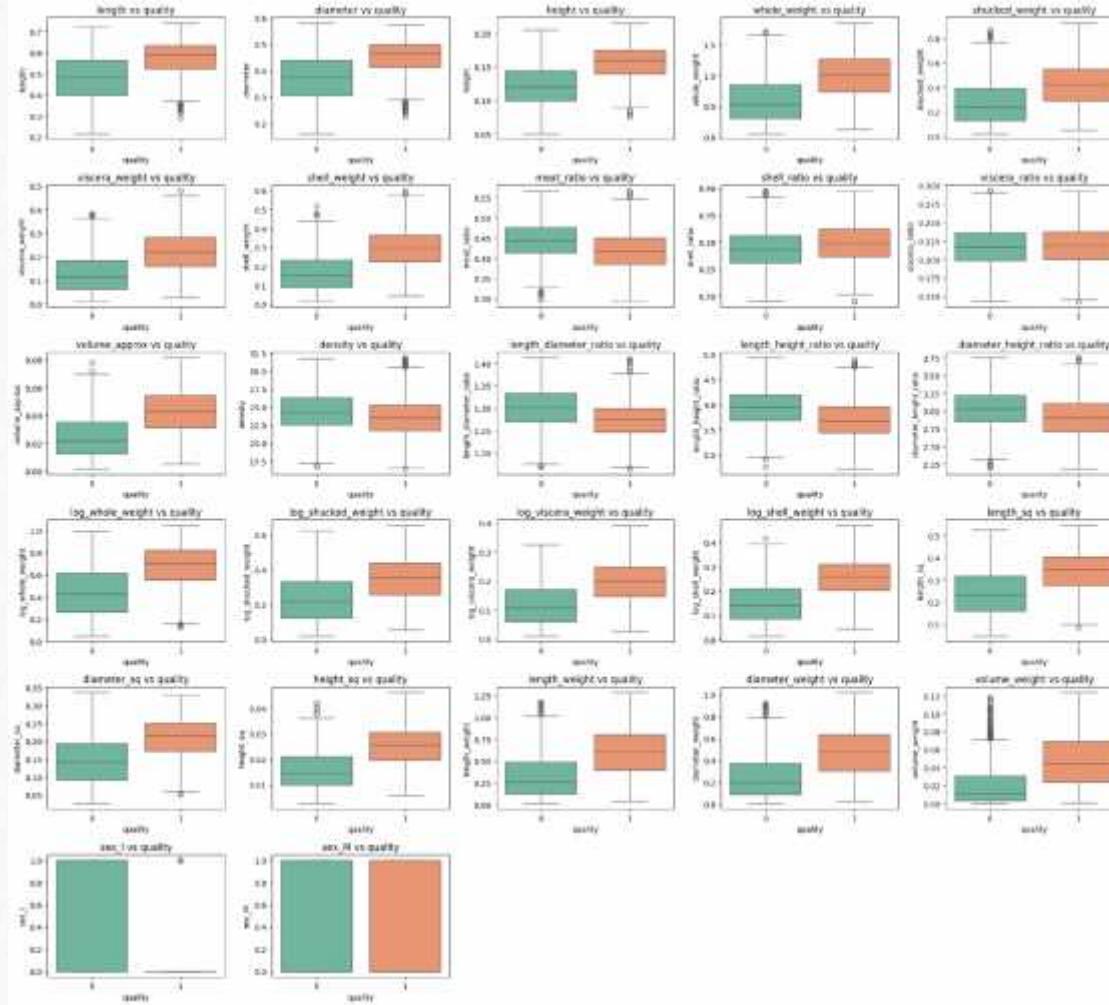
Цель: определить зрелость моллюска по физическим характеристикам

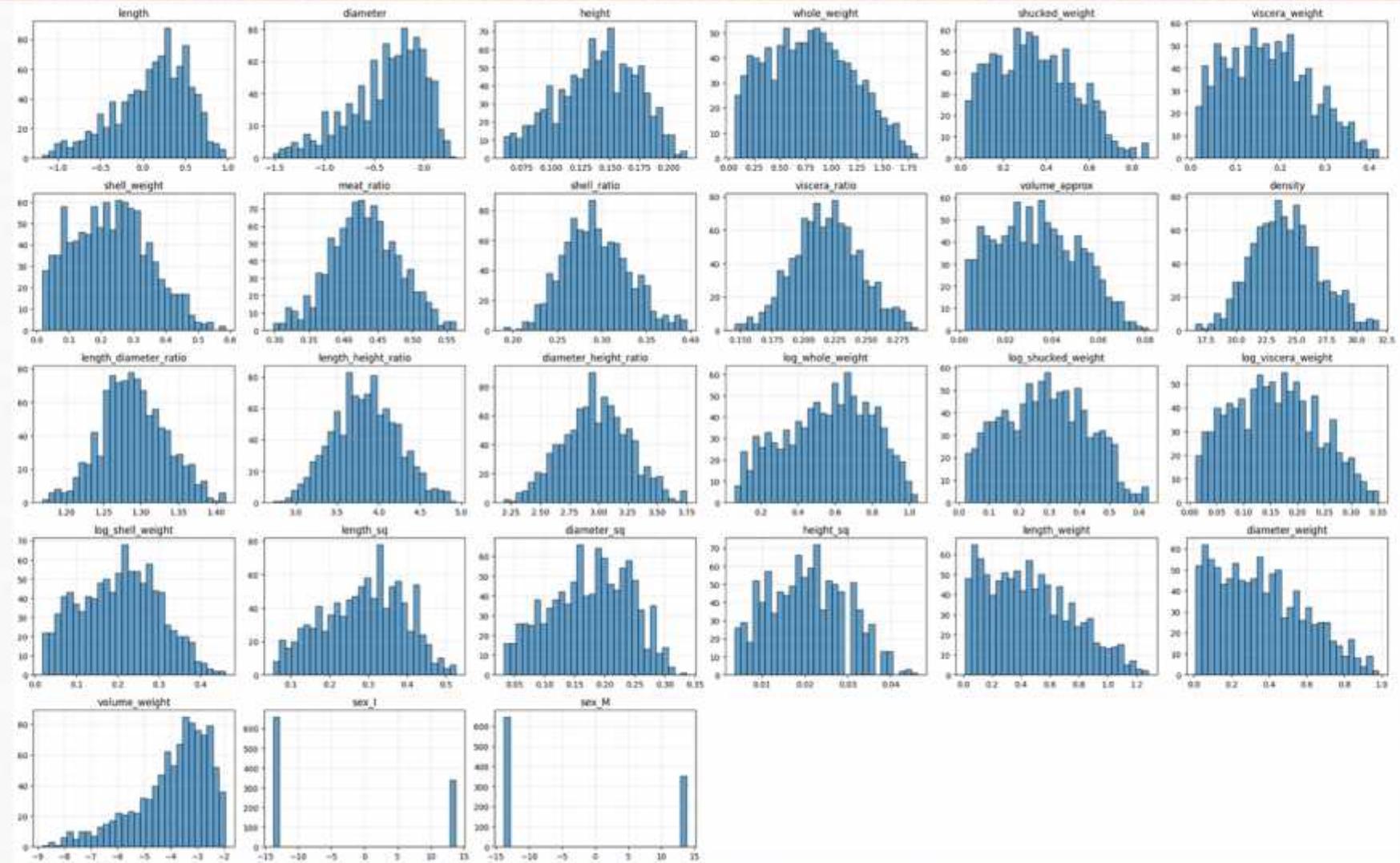
Исходные признаки:

- Категориальный: `sex` (пол)
- Числовые: `length`,
`diameter`, `height` (размеры)
- Числовые: `whole_weight`,
`shucked_weight`,
`viscera_weight`, `shell_weight` (веса)
- Целевая переменная: `rings` (преобразована в `quality`)





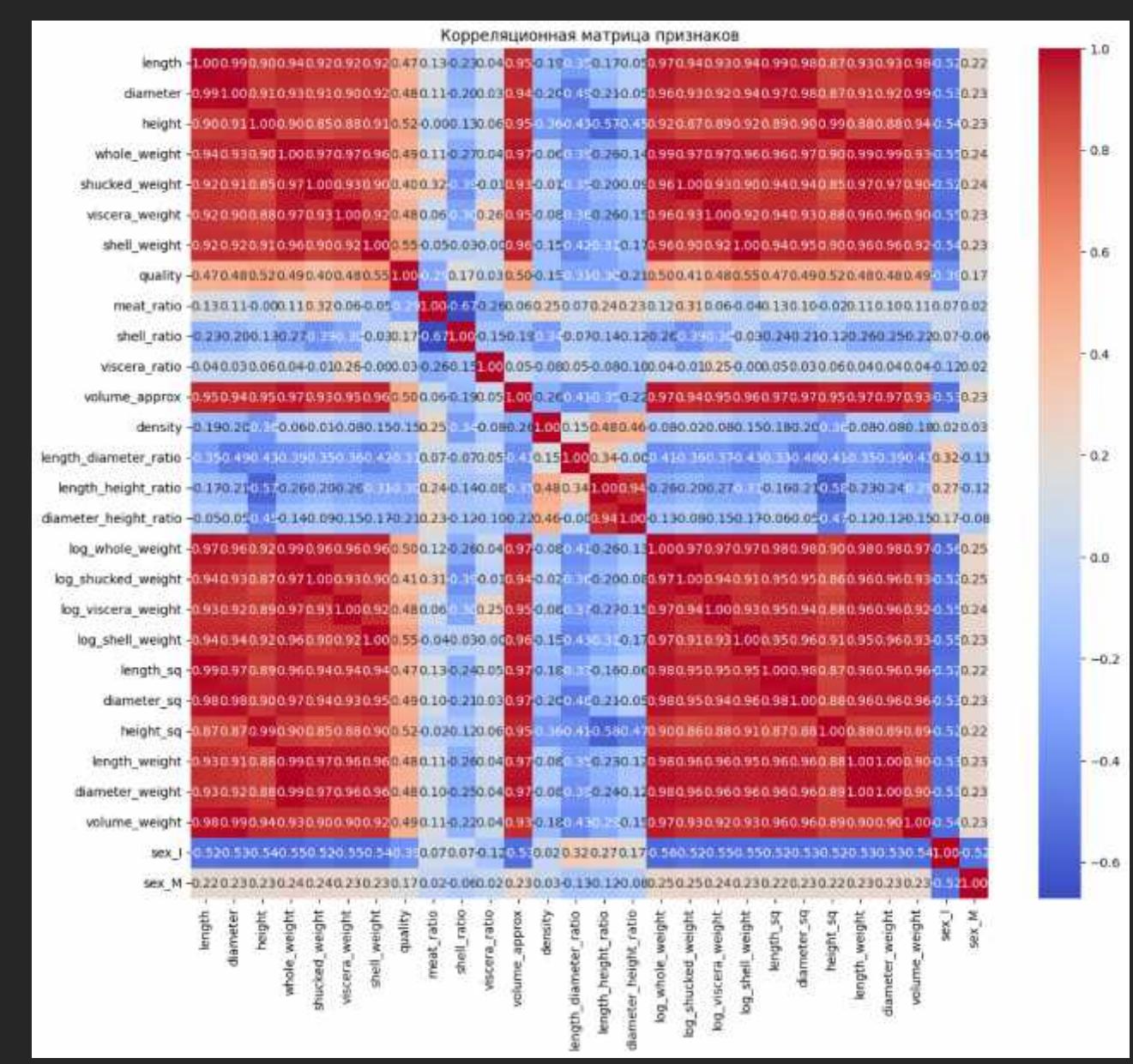




Итоговая корреляционная матрица

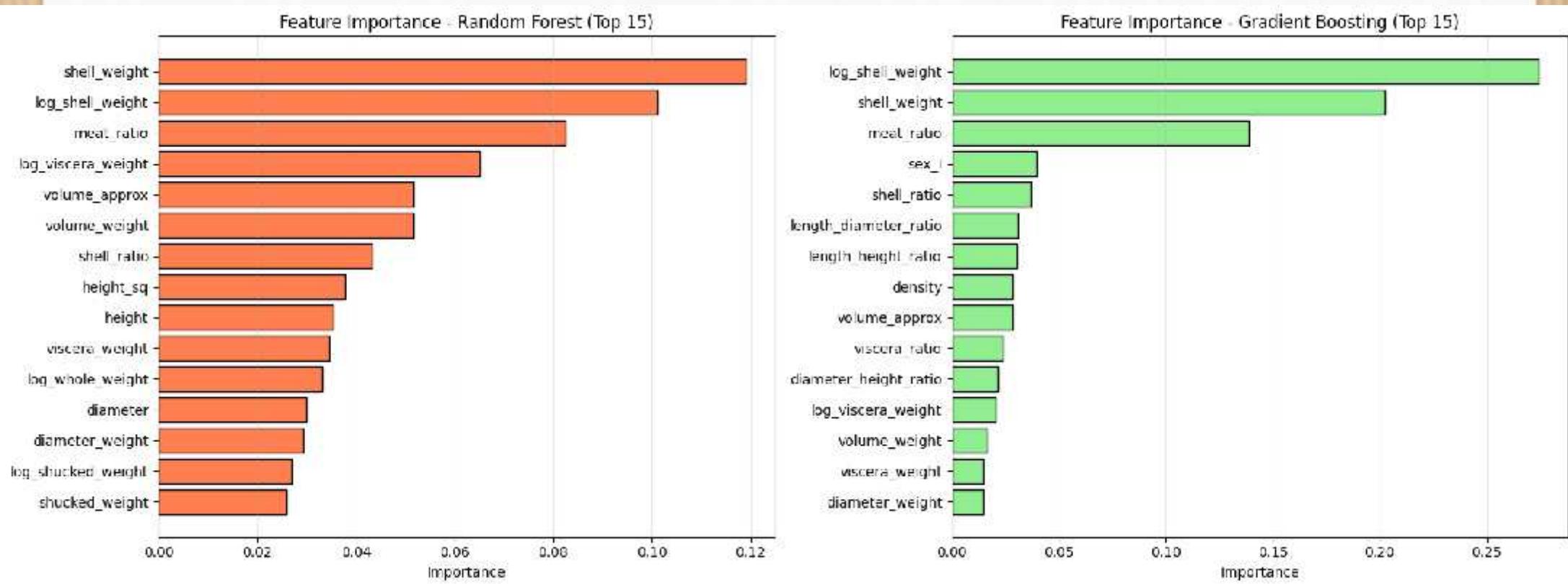
Исходно: 9 признаков

- После Feature Engineering: 28 признаков
- Созданные признаки:
 - Соотношения весов: `meat_ratio`, `shell_ratio`, `viscera_ratio`
 - Объемные признаки: `volume_approx`, `density`
 - Соотношения размеров: `length_diameter_ratio`, `length_height_ratio`, `diameter_height_ratio`
 - Логарифмические преобразования:** `log_whole_weight`, `log_shucked_weight`, и т.д.
 - Квадратные признаки: `length_sq`, `diameter_sq`, `height_sq`
 - Взаимодействия: `length_weight`, `diameter_weight`, `volume_weight`



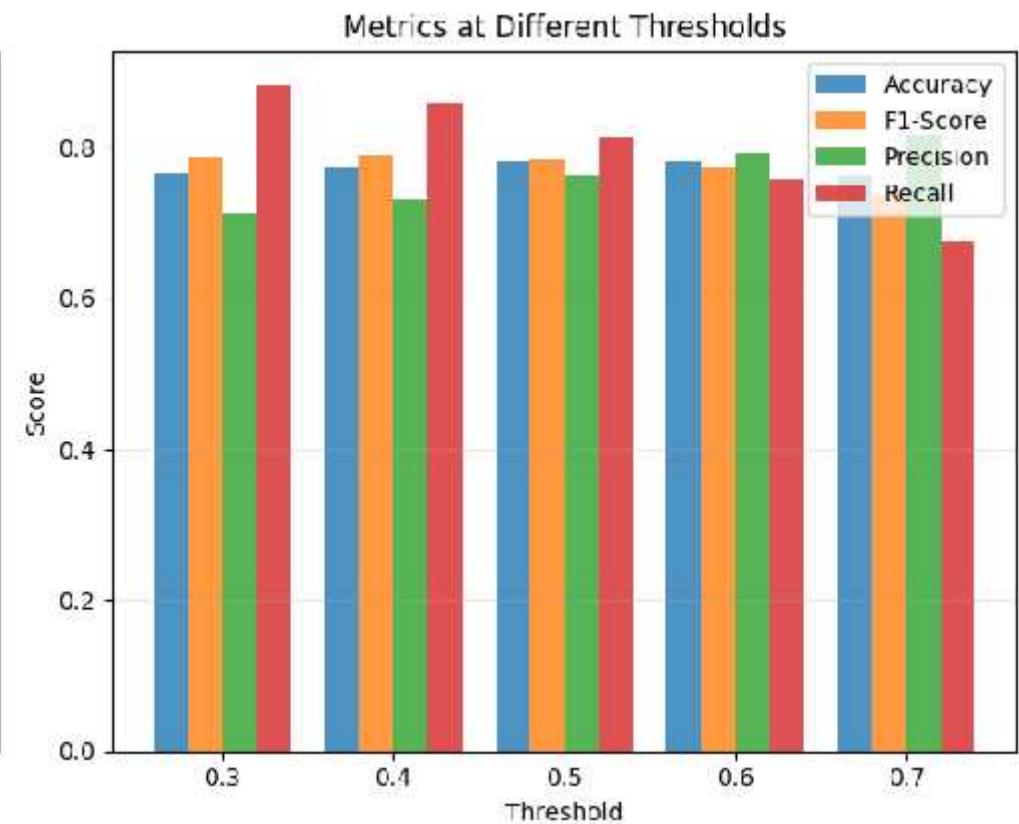
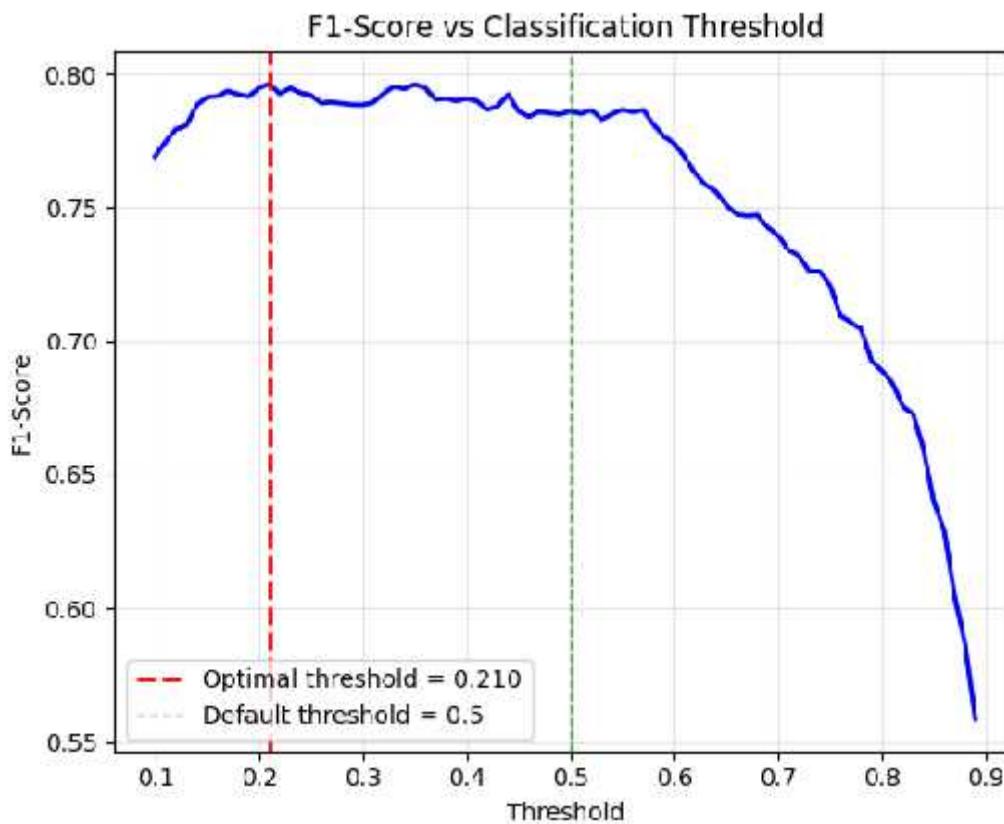
Важность признаков (Feature Importance)

- Какие признаки наиболее важны для модели?
- Топ-10 наиболее важных признаков для Gradient Boosting



Уникальная фича - Оптимальный порог классификации

- Проблема: стандартный порог 0.5 может быть неоптимальным
- Решение: поиск оптимального порога для максимизации F1-Score
- Результат: найден оптимальный порог 0.210 (вместо 0.5)



Основные достижения

1. Успешно преобразована задача регрессии в классификацию
2. Создано 19 новых признаков (Feature Engineering)
3. Gradient Boosting показал лучшие результаты
4. Найден оптимальный порог классификации (0.210)
5. Достигнуто высокое качество: ROC-AUC = 0.8730

Практическая значимость:

- Модель может использоваться для быстрой оценки зрелости моллюсков
- Снижение затрат в аквакультуре



Спасибо за
внимание!
