Анализ данных о пингвинах Антарктики

Кабанов Даниил

май 2025г.

Введение

- **Цель проекта**: Классификация видов пингвинов по морфологическим признакам
- **Актуальность**: Пример применения Data Science в биологии
- **Данные**:
  - 344 особи пингвинов
  - 7 признаков: длина клюва, масса тела и др.
  - 3 вида: Adelie, Chinstrap, Gentoo

Логистическая регрессия

Многоклассовая классификация

Функция Softmax:

$$P(y = k|X) = \frac{e^{w_k^T X}}{\sum_{j=1}^{K} e^{w_j^T X}}$$

где:

- w_k - веса для класса k
- K - количество классов

Функция потерь (кросс-энтропия):

$$\mathcal{L}(W) = -\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik} \log P(y_i = k|x_i)$$

Метод k-ближайших соседей (kNN)

Формальное определение:

$$\hat{y} = \text{argmax}_k \sum_{x_i \in N_k(x)} I(y_i = k)$$

Параметры:

- Метрика расстояния: Евклидова

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Оптимальное k = 3 (подобрано через GridSearch)

Кластеризация K-means

Алгоритм:

1. Инициализация центроидов
2. Назначение точек ближайшему центроиду

$$c_i = \text{argmin}_k ||x_i - \mu_k||^2$$

3. Пересчет центроидов

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$$

Метод локтя для выбора k:

$$W(k) = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

Silhouette Score

Формула:

Для объекта i:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

где:

- $a(i)$ — расстояние до своего кластера
- $b(i)$ — расстояние до ближайшего кластера

Общий score:

$$S = \frac{1}{N} \sum_{i=1}^{N} s(i) \quad \in [-1, 1]$$

$S \approx 1$	Отличная кластеризация
$S \approx 0$	Кластеры пересекаются
$S \approx -1$	Ошибка кластеризации

Распределение признаков

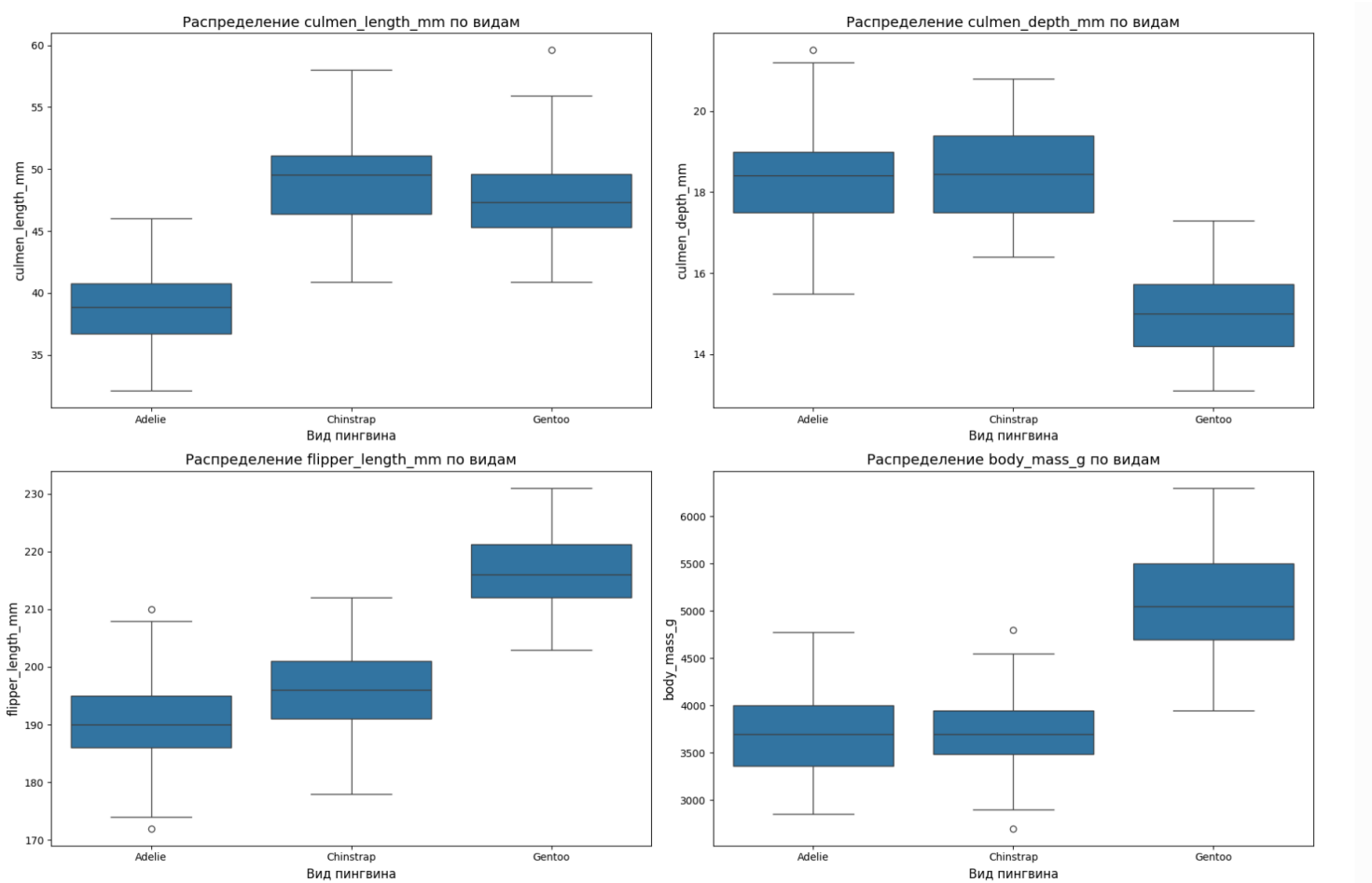


Рис.: Распределение числовых признаков по видам пингвинов

Корреляция признаков

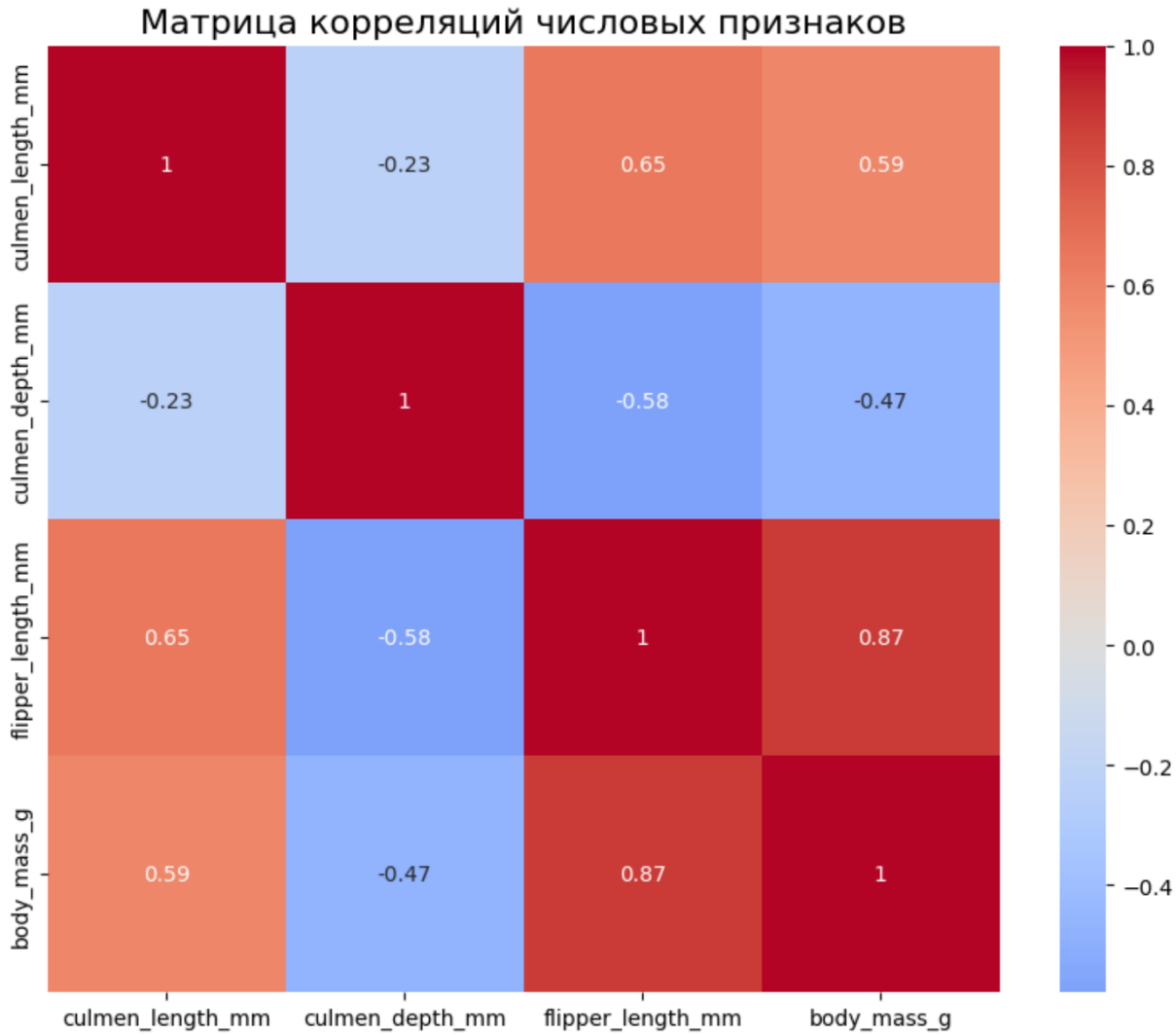


Рис.: Матрица корреляций между признаками

Оценка моделей

Таблица: Сравнение моделей классификации

Модель	Accuracy	Precision	Recall	F1-score
Логистическая регрессия	1.00	1.00	1.00	1.00
kNN (k=9)	1.00	1.00	1.00	1.00

Кластеризация

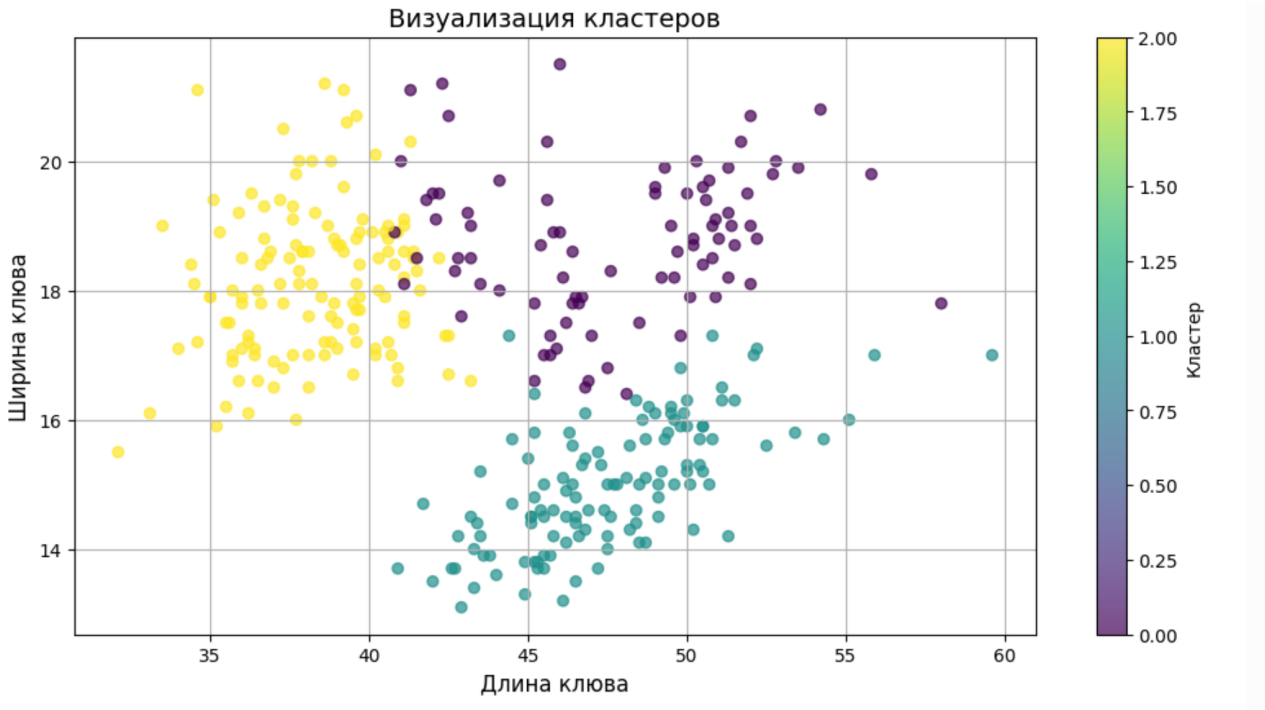


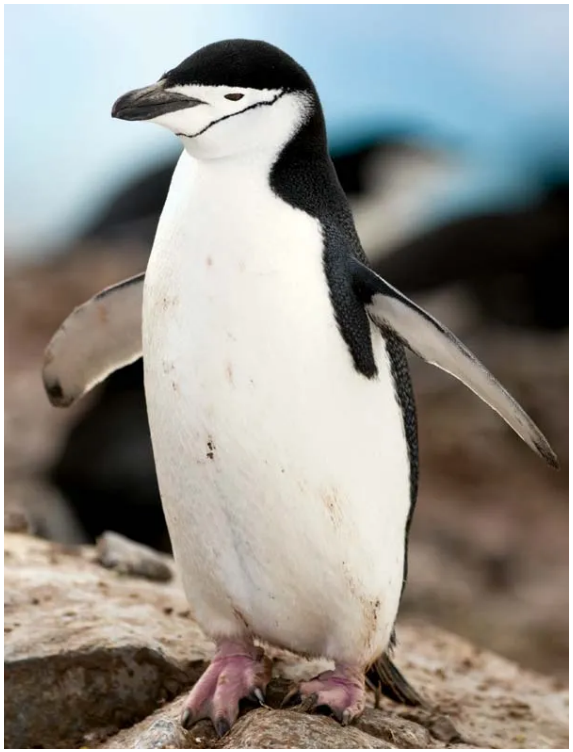
Рис.: Результаты кластеризации

Выводы

- Обе модели классификации показали отличные результаты
- Кластеры почти соответствуют биологическим видам
- Ключевые различия:
  - Gentoo: большая масса тела
  - Chinstrap: длинные узкие клювы

Эти самые пингвины

(a) Пингвин Адели
(Pygoscelis adeliae)



(b) Антарктический
(Pygoscelis antarcticus)

(c) Папуанский
(Pygoscelis papua)

- Адели: Самые распространённые, чёрная голова
- Антарктические: Чёрная полоса под подбородком
- Папуанские: Ярко-оранжевый клюв, крупные