
Data learning

Курс “Машинное обучение”

Лабораторная работа


# ECOC coding design schemes

Леонов В.В., М23-524

Вариант 2-06

2023

# Исходные данные

Исходные данные представлены в виде файла **data_ml_v2-06.csv**, который содержит переменные *x1*, *x2, label*

Объем выборки составляет **500 записей**.
Решается задача классификации:

Признаков – 2.

Классов – 4.

# Используемые методы и формулы

Кросс-валидация – это метод оценки производительности модели машинного обучения, который помогает уменьшить влияние случайности в процессе разделения данных на обучающую и тестовую выборки. Основная идея заключается в разделении данных на несколько подмножеств и последующем обучении и тестировании модели на их разных комбинациях. Кросс-валидация позволяет более точно оценить обобщающую способность модели, уменьшает риск переобучения и обнаруживает стабильность модели на различных подмножествах данных.

Этапы:

*Разбиение данных*: Исходные данные разделяются на K подмножеств.

*Обучение и тестирование*: Модель обучается K раз, каждый раз используя K-1 фолдов в качестве обучающего набора данных и оставшийся 1 фолд в качестве тестового набора данных.

*Оценка производительности*: За каждую итерацию вычисляются метрики производительности модели, и в конце процесса получается усредненная оценка.

# Используемые методы и формулы

Байесовский бинарный классификатор на основе кодирования избыточности один против всех (ECOC OVA) представляет собой метод, который использует байесовский подход для решения задачи бинарной классификации для каждого из классов в многоклассовой задаче. Кодирование избыточности один против всех (OVA) означает, что каждый класс сравнивается с остальными классами как бинарная задача.

Этот метод позволяет строить модели байесовских классификаторов для каждого класса, рассматривая их в отдельности. Применение кодирования избыточности один против всех обеспечивает масштабируемость метода для задач с большим числом классов.

# Используемые методы и формулы

Байесовский бинарный классификатор на основе кодирования избыточности один против одного (ECOC OvO) представляет собой метод, который решает задачу бинарной классификации для каждой пары классов в многоклассовой задаче. В данном контексте "один против одного" означает, что каждая пара классов рассматривается как отдельная бинарная задача.

Преимущества этого метода включают возможность использования различных байесовских классификаторов для каждой пары классов, что может быть полезным в случаях, когда предположения о распределении данных могут различаться для разных пар классов.

# Используемые методы и формулы

Байесовский бинарный классификатор с использованием кодирования избыточности полного бинарного кода (ECOC Full Binary Code) представляет собой метод, который использует полный набор всех возможных двоичных кодов для представления каждого класса в многоклассовой задаче. Каждый класс ассоциируется с уникальным бинарным кодом, и для каждой пары классов создается бинарная задача.

Использование полного бинарного кода обеспечивает уникальный код для каждого класса, что может улучшить различимость между классами в контексте бинарных задач.

# Используемые методы и формулы

Байесовский бинарный классификатор с использованием кодирования избыточности полного тернарного кода (ECOC Full Ternary Code) - это метод, который использует полный набор всех возможных тернарных кодов для представления каждого класса в многоклассовой задаче. Каждый класс ассоциируется с уникальным тернарным кодом (состоящим из троичных значений: -1, 0, 1), и для каждой пары классов создается бинарная задача.

Использование полного тернарного кода обеспечивает уникальный код для каждого класса, что может улучшить различимость между классами в контексте бинарных задач.

# Используемые методы и формулы

В контексте кодирования избыточности (ECOC) декодирование является процессом преобразования выхода отдельных бинарных классификаторов в прогноз для многоклассовой задачи.

**Невзвешенное декодирование** (Unweighted Decoding):
1) В невзвешенном декодировании каждый бинарный классификатор вносит одинаковый вклад в окончательное решение для многоклассовой задачи.
2) Простой подсчет голосов: класс, который набирает наибольшее количество "положительных голосов" от бинарных классификаторов, выбирается как окончательный прогноз.

# Используемые методы и формулы

**Взвешенное декодирование** (Weighted Decoding):

1) В взвешенном декодировании каждый бинарный классификатор вносит вклад с весом, отражающим его "уверенность" или надежность.

2) Каждый бинарный классификатор может выдавать оценку вероятности принадлежности к классу, и взвешенное декодирование учитывает эти вероятности.

3) Веса могут быть заданы заранее или могут быть определены на основе уверенности каждого классификатора в своем прогнозе.

# Используемые методы и формулы

Вычисление очков бинарного классификатора

$$\phi_k = 2p_k - 1,$$

где $p_k$ – апостериорная вероятность *k-го* класса

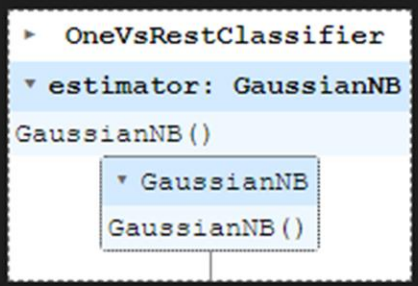
Вычисление точности классификатора

$$accuracy\ (y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y} = y_i)$$

# Результаты исследований

Построение и обучение классификатора OVA

```
1  # Классификатор OVA - OneVsAll
2  ova_classifier = OneVsRestClassifier(GaussianNB())
3  ova_classifier.fit(X_train, Y_train)
```



# Результаты исследований

Построение и обучение классификатора OVO

# Результаты исследований

Построение и обучение классификатора «Полное бинарное кодирование»

# Результаты исследований

Построение и обучение классификатора «Полное тернарное кодирование»

# Результаты исследований

Расчёт классификационных очков

Оценка результатов с использованием невзвешенного ECOC-декодирования (OVA):

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.79 | 0.97 | 0.87 | 34 |
| 2 | 0.82 | 0.96 | 0.88 | 47 |
| 3 | 0.67 | 0.60 | 0.63 | 43 |
| 4 | 0.86 | 0.46 | 0.60 | 26 |
| accuracy | | | 0.77 | 150 |
| macro avg | 0.78 | 0.75 | 0.75 | 150 |
| weighted avg | 0.77 | 0.77 | 0.76 | 150 |

# Результаты исследований

Расчёт классификационных очков

Оценка результатов с использованием взвешенного ECOC-декодирования (OVA):

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.79 | 0.97 | 0.87 | 34 |
| 2 | 0.82 | 0.96 | 0.88 | 47 |
| 3 | 0.67 | 0.60 | 0.63 | 43 |
| 4 | 0.86 | 0.46 | 0.60 | 26 |
| accuracy | | | 0.77 | 150 |
| macro avg | 0.78 | 0.75 | 0.75 | 150 |
| weighted avg | 0.77 | 0.77 | 0.76 | 150 |

# Результаты исследований

Расчёт классификационных очков

Оценка результатов с использованием невзвешенного ECOC-декодирования (OVO):

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.78 | 0.91 | 0.84 | 34 |
| 2 | 0.82 | 0.96 | 0.88 | 47 |
| 3 | 0.62 | 0.60 | 0.61 | 43 |
| 4 | 0.85 | 0.42 | 0.56 | 26 |
| accuracy | | | 0.75 | 150 |
| macro avg | 0.76 | 0.72 | 0.72 | 150 |
| weighted avg | 0.76 | 0.75 | 0.74 | 150 |

# Результаты исследований

Расчёт классификационных очков

Оценка результатов с использованием взвешенного ECOC-декодирования (OVO):

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.78 | 0.91 | 0.84 | 34 |
| 2 | 0.82 | 0.96 | 0.88 | 47 |
| 3 | 0.62 | 0.60 | 0.61 | 43 |
| 4 | 0.85 | 0.42 | 0.56 | 26 |
| accuracy | | | 0.75 | 150 |
| macro avg | 0.76 | 0.72 | 0.72 | 150 |
| weighted avg | 0.76 | 0.75 | 0.74 | 150 |

# Результаты исследований

Расчёт классификационных очков

Оценка результатов с использованием невзвешенного ECOC-декодирования (Полное бинарное кодирование):

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.76 | 0.94 | 0.84 | 34 |
| 2 | 0.61 | 1.00 | 0.76 | 47 |
| 3 | 0.77 | 0.23 | 0.36 | 43 |
| 4 | 0.72 | 0.50 | 0.59 | 26 |
| accuracy | | | 0.68 | 150 |
| macro avg | 0.72 | 0.67 | 0.64 | 150 |
| weighted avg | 0.71 | 0.68 | 0.63 | 150 |

# Результаты исследований

Расчёт классификационных очков

Оценка результатов с использованием взвешенного ECOC-декодирования (Полное бинарное кодирование):

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.76 | 0.94 | 0.84 | 34 |
| 2 | 0.61 | 1.00 | 0.76 | 47 |
| 3 | 0.77 | 0.23 | 0.36 | 43 |
| 4 | 0.72 | 0.50 | 0.59 | 26 |
| accuracy | | | 0.68 | 150 |
| macro avg | 0.72 | 0.67 | 0.64 | 150 |
| weighted avg | 0.71 | 0.68 | 0.63 | 150 |

# Результаты исследований

Расчёт классификационных очков

Оценка результатов с использованием невзвешенного ECOC-декодирования (Полное тернарное кодирование):

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.79 | 0.97 | 0.87 | 34 |
| 2 | 0.74 | 0.98 | 0.84 | 47 |
| 3 | 0.69 | 0.51 | 0.59 | 43 |
| 4 | 0.86 | 0.46 | 0.60 | 26 |
| accuracy | | | 0.75 | 150 |
| macro avg | 0.77 | 0.73 | 0.72 | 150 |
| weighted avg | 0.76 | 0.75 | 0.73 | 150 |

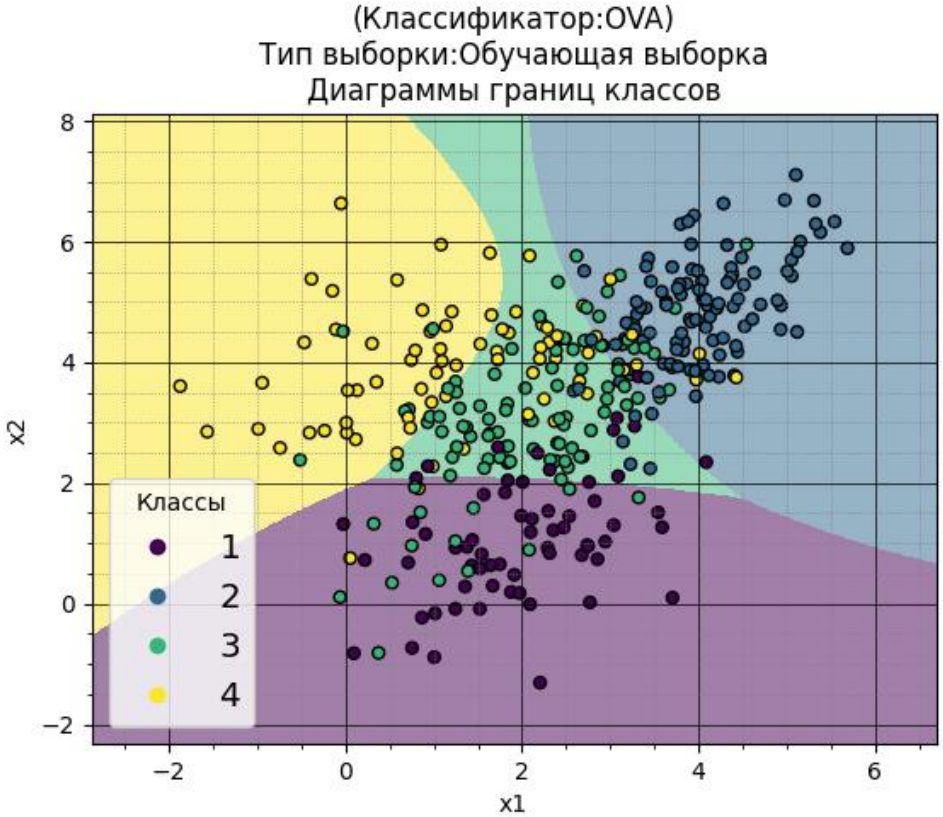
# Результаты исследований

Расчёт классификационных очков

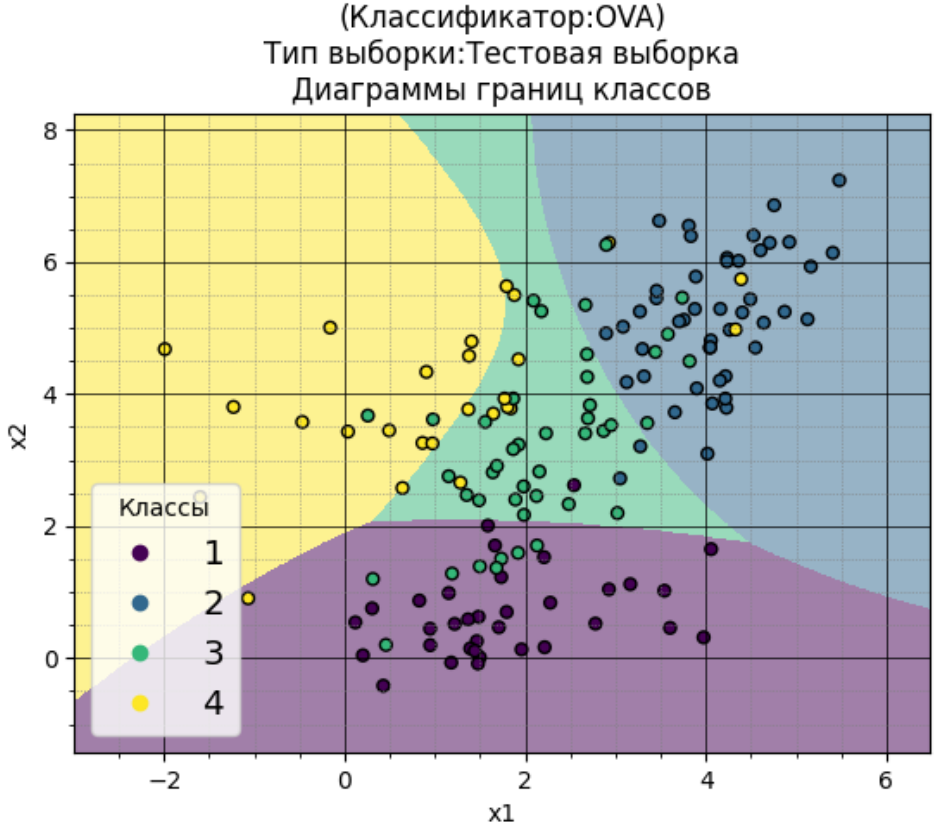
Оценка результатов с использованием взвешенного ECOC-декодирования (Полное тернарное кодирование):

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.79 | 0.97 | 0.87 | 34 |
| 2 | 0.74 | 0.98 | 0.84 | 47 |
| 3 | 0.69 | 0.51 | 0.59 | 43 |
| 4 | 0.86 | 0.46 | 0.60 | 26 |
| accuracy | | | 0.75 | 150 |
| macro avg | 0.77 | 0.73 | 0.72 | 150 |
| weighted avg | 0.76 | 0.75 | 0.73 | 150 |

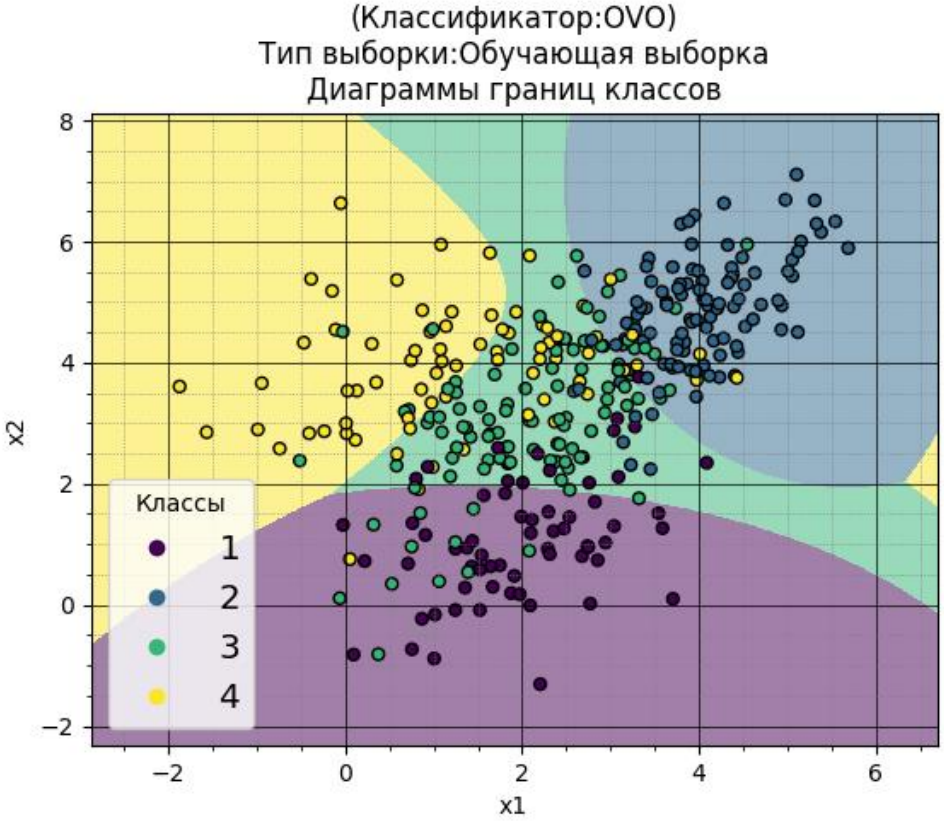
# Результаты исследований



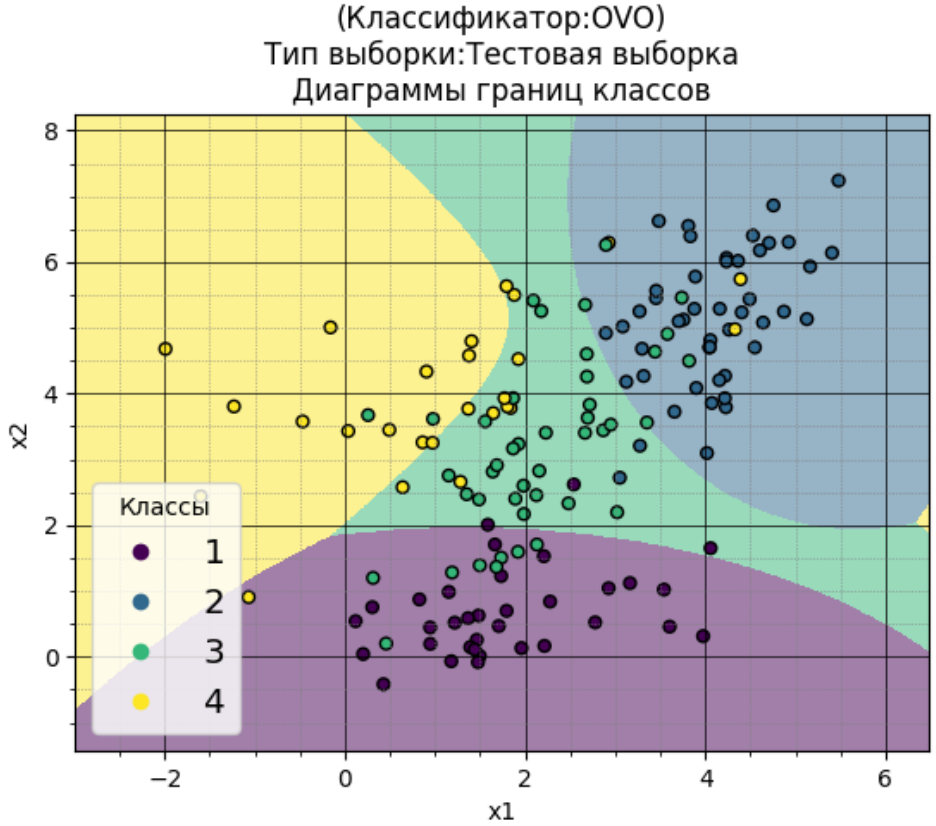
# Результаты исследований



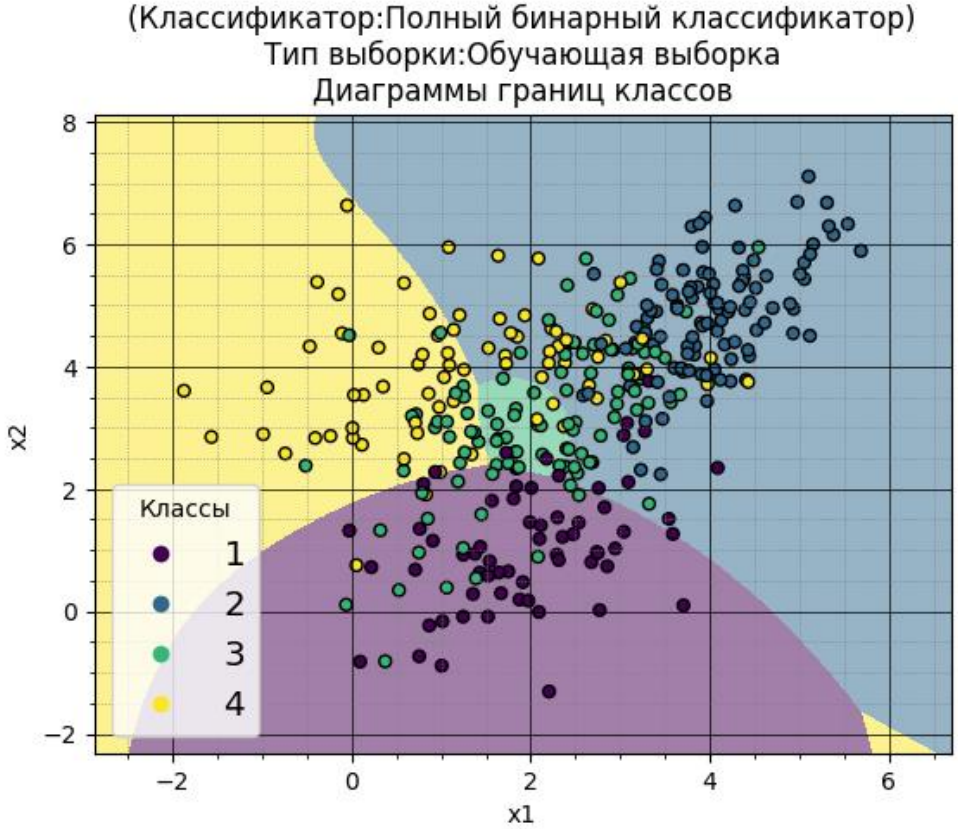
# Результаты исследований



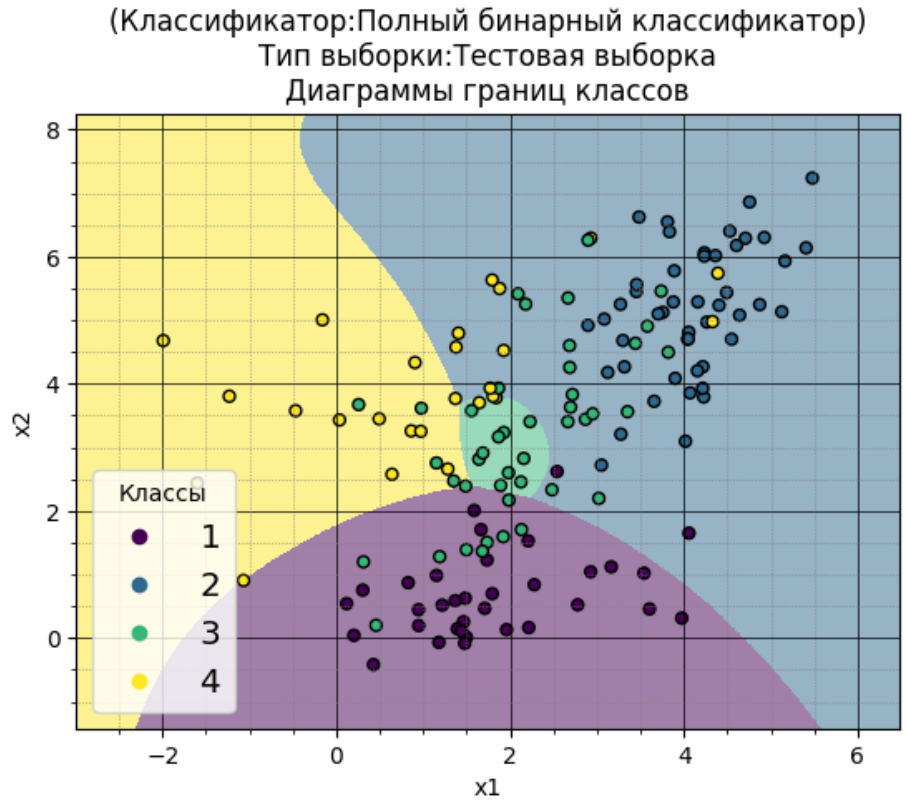
# Результаты исследований



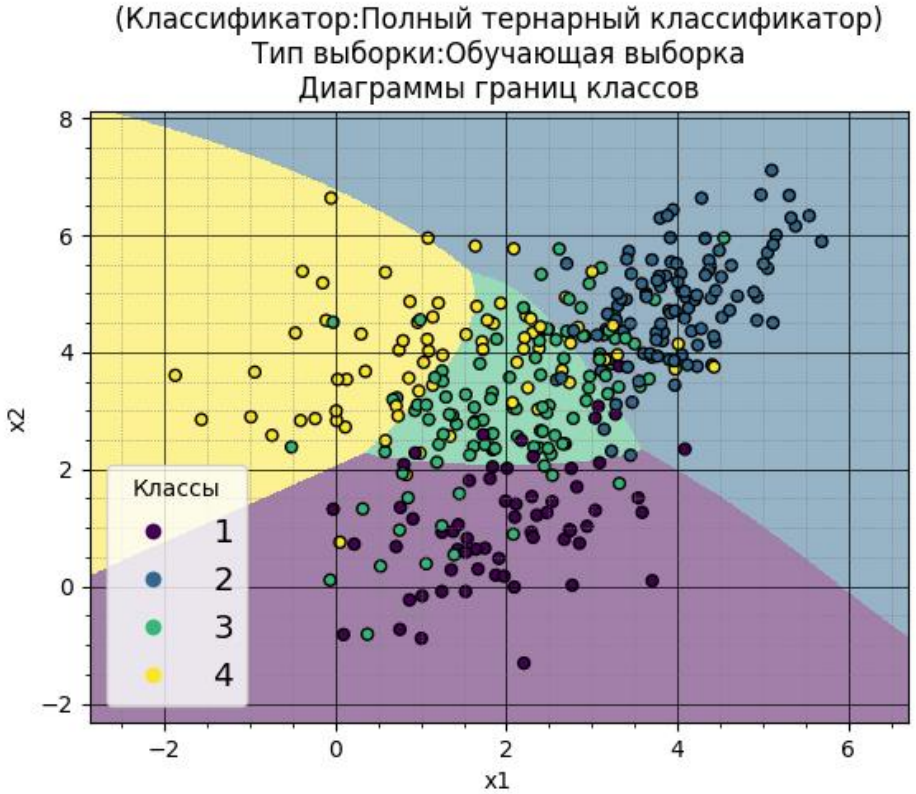
# Результаты исследований



# Результаты исследований

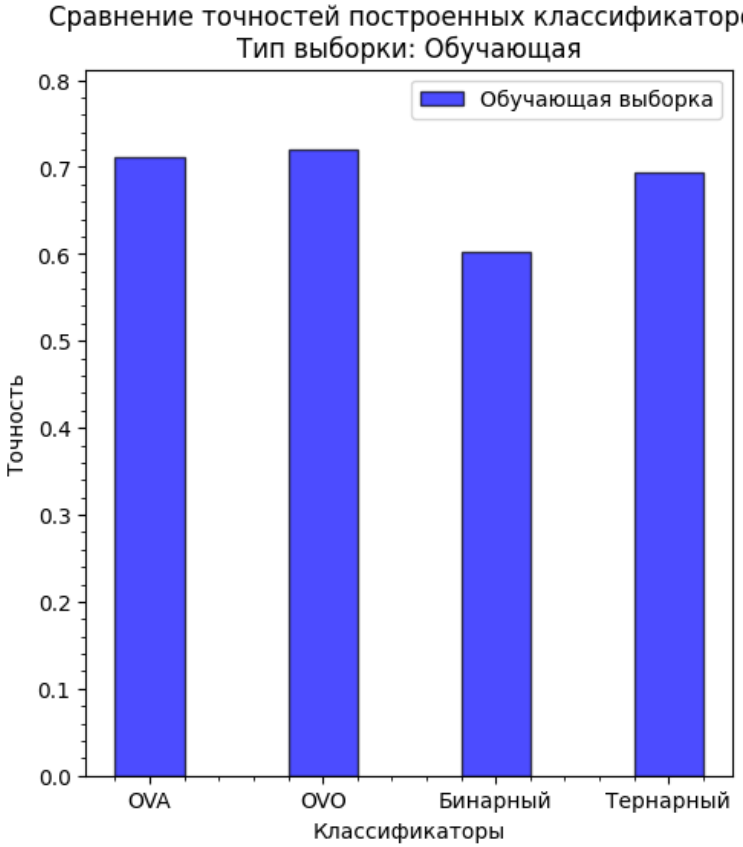


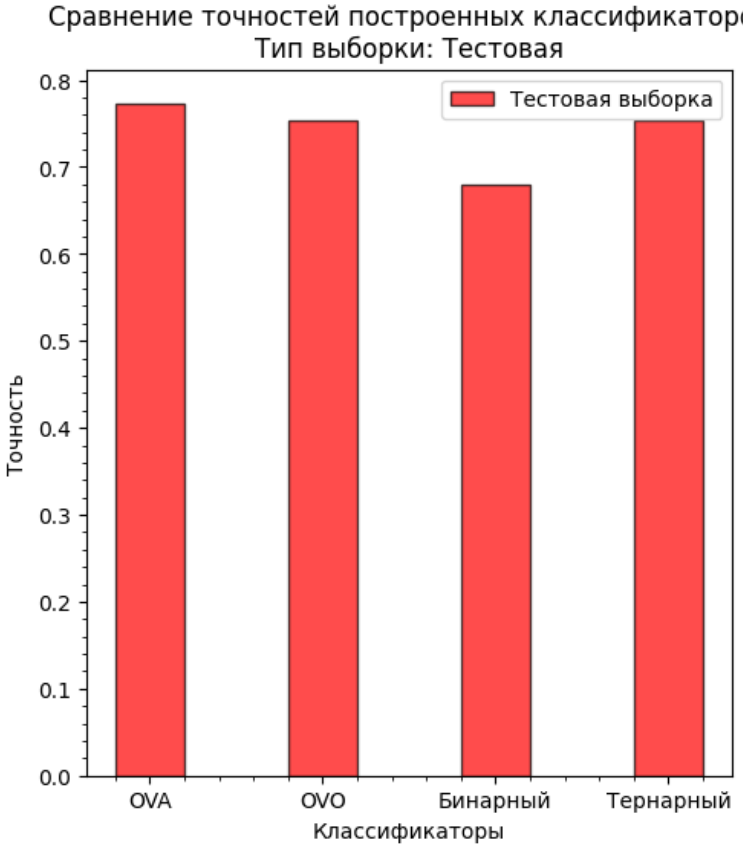
# Результаты исследований



# Результаты исследований

# Результаты исследований



# Выводы

**Одно против всех (OVA)**

Преимущества: Простота реализации, хорошо работает с большим количеством классов.

Недостатки: Может столкнуться с проблемой несбалансированных классов, так как каждый бинарный классификатор обучается на различном количестве положительных и отрицательных примеров.

# Выводы

**Один против одного (OVO)**

Преимущества: Решает проблему несбалансированных классов, более эффективен в случае небольшого количества классов.

Недостатки: Требует больше бинарных классификаторов, что может быть проблемой при большом числе классов.

# Выводы

**Полный бинарный кодификатор**

Преимущества: Уникальный код для каждого класса, что может улучшить различимость между классами.

Недостатки: Требует больше бинарных классификаторов, чем OVA и OVO, что может увеличить вычислительную сложность.

# Выводы

**Полный тернарный кодификатор**

Преимущества: Аналогично полному бинарному кодификатору, но с использованием тернарных значений.

Недостатки: По аналогии с полным бинарным кодификатором может потребовать больше ресурсов.

# Выводы

**Невзвешенное декодирование (Unweighted Decoding)**

Преимущества: Простота реализации, равномерное влияние каждого бинарного классификатора на решение.

Недостатки: Не учитывает разную уверенность бинарных классификаторов.

# Выводы

**Взвешенное декодирование (Weighted Decoding)**

Преимущества: Учет разной уверенности бинарных классификаторов, что может улучшить качество решения.

Недостатки: Требует оценки уверенности от каждого бинарного классификатора.