**УДК 004**

# ОЦЕНКА ВЛИЯНИЯ СЛУЧАЙНОСТИ ПРИ РАЗДЕЛЕНИИ ДАННЫХ НА ТРЕНИРОВОЧНУЮ И ТЕСТОВЫЕ ВЫБОРКИ

Смирнов А.П., студент гр. М20126, I курс
Научный руководитель: Соколова С.С., преподаватель кафедры киберфизических систем СПБГМТУ
Федеральное государственное бюджетное образовательное учреждение высшего образования
Санкт-Петербургский государственный морской технический университет
г. Санкт-Петербург

В задачах машинного обучения зачастую исследуются реально существующие данные. Грамотная их интерпретация и анализ могут существенно повлиять на результаты исследования, а также на качество и надежность построенных моделей. Одним из ключевых этапов при работе с данными является их разбиение на тренировочную и тестовую выборки, что необходимо для корректной оценки качества алгоритмов. Однако процесс этого разбиения не всегда очевиден и может оказывать значительное влияние на итоговую производительность модели.

Цель данной работы — исследовать, действительно ли случайное разделение датасета на тренировочную и тестовые выборки способно повлиять на итоги работы алгоритмов машинного обучения, а также определить, насколько велики различия между разными стратегиями разбиения данных. Данная проблема особенно актуальна при работе с небольшими выборками, где случайные флуктуации в распределении данных могут приводить к значительным колебаниям метрик качества модели.

Для достижения этой цели были поставлены следующие задачи:

1. Выбрать реально существующий, но относительно простой датасет, который позволит провести сравнение нескольких методов разбиения входных данных. В качестве рассматриваемых методов выбраны: простое разбиение, стратифицированное разбиение, кросс-валидация со стратификацией.
2. Оценить точность (accuracy) алгоритма в зависимости от метода разбиения данных, а также вычислить стандартное отклонение точности для разных случаев, чтобы выявить степень вариативности полученных результатов.
3. Проанализировать причины расхождений в результатах, объяснить, почему одни методы обеспечивают более стабильные оценки качества модели, и определить, какой метод рациональнее использовать в реальных задачах.

Результаты исследования быть использованы в образовательных целях при изучении дисциплин, связанных с машинным обучением, а также при

разработке более надежных стратегий разбиения выборок для практических задач. Наглядная демонстрация влияния метода разбиения данных на конечные результаты позволит избежать распространенных ошибок и повысить достоверность проводимых экспериментов. Выводы могут наглядно продемонстрировать, почему вопросу случайного разбиения выборок все же следует уделять внимание.

В качестве датасета для исследования был выбран один из самых известных и широко применяемых наборов данных — Ирисы Фишера (Iris Dataset)**.** Это классический датасет, который активно используется в учебных и научных целях, а также при тестировании различных алгоритмов машинного обучения, особенно в задачах классификации. Данный набор содержит 150 наблюдений (строк), каждая из которых представляет собой параметры одного цветка ириса, включая такие характеристики, как длина и ширина лепестков и чашелистиков. На основе этих признаков можно классифицировать цветок как один из трех видов ирисов. Датасет был импортирован из библиотеки Scikit-learn (sklearn)**.**

В качестве базового алгоритма для классификации данных использовался метод логистической регрессии. Его основная цель — предсказать, к какому классу принадлежит конкретный цветок, используя входные признаки. Перед обучением модели данные были разделены на тренировочную и тестовую выборки в стандартном соотношении 80:20. Однако важно понимать, что это разбиение не является абсолютно случайным в математическом смысле. Оно реализуется с помощью псевдослучайного генератора, где результаты разбиения зависят от фиксированного начального значения (random_state). Чтобы оценить влияние случайности, в исследовании рассматривались различные значения random_state в диапазоне от 0 до 45, что позволило проанализировать устойчивость модели к различным вариантам разделения данных.

Далее в методе простого разбиения обучение применяется со всеми вариантами random_state.

В методе стратифицированного разбиения проводится стратификация данных, то есть сохранение пропорций классов такими же в выборках, как и в изначальном датасете. Потенциально, это может уменьшить дисбаланс в данных, ведь присутствие одного класса может превалировать.

А в методе кросс-валидации со стратификацией сначала проводится стратификация данных, затем кросс-валидация на одном из популярных методов – 5-блочной кросс валидации(5-fold cross-validation). То есть данные делятся на 5 блоков, 4 из них определяются в обучающую выборку, 1 в тестовую. Модель обучается 5 раз, пока каждый из блоков не попадет и в одну, и в другую выборки. Как итоговая точность – берется среднее значение всех точностей. Схематично процесс разбиения отображен на рисунке 1.

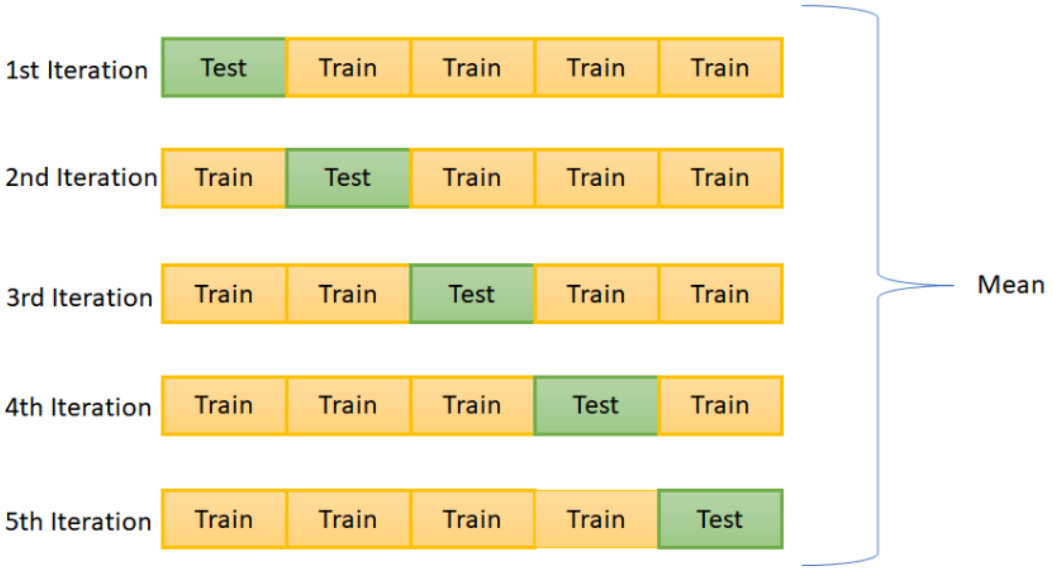


Рисунок 1. Процесс разбиения данных при кросс-валидации.

Результаты исследования отображены в таблице 1.

| random_state | Точность (Стратифицированное разбиение) | Точность (Кросс-валидация со стратификацией) | Точность (Простое разбиение) |
|---|---|---|---|
| 0 | 1.0 | 0.95 | 1.0 |
| 5 | 0.966667 | 0.966667 | 0.966667 |
| 10 | 1.0 | 0.95 | 1.0 |
| 15 | 0.933333 | 0.958333 | 1.0 |
| 20 | 1.0 | 0.95 | 0.933333 |
| 25 | 0.9 | 0.975 | 0.966667 |
| 30 | 0.933333 | 0.966667 | 0.966667 |
| 35 | 0.9 | 0.966667 | 1.0 |
| 40 | 0.966667 | 0.966667 | 1.0 |
| 45 | 0.9 | 0.966667 | 0.966667 |

Таблица 1. Точность для методов при различном random_state

Средние и стандартные отклонения точности:

- Стратифицированное разбиение: 0.9500 ± 0.0401
- Кросс-валидация со стратификацией: 0.9617 ± 0.0085
- Простое разбиение: 0.9800 ± 0.0221

Графики с результатами точности изображены на рисунке 2.

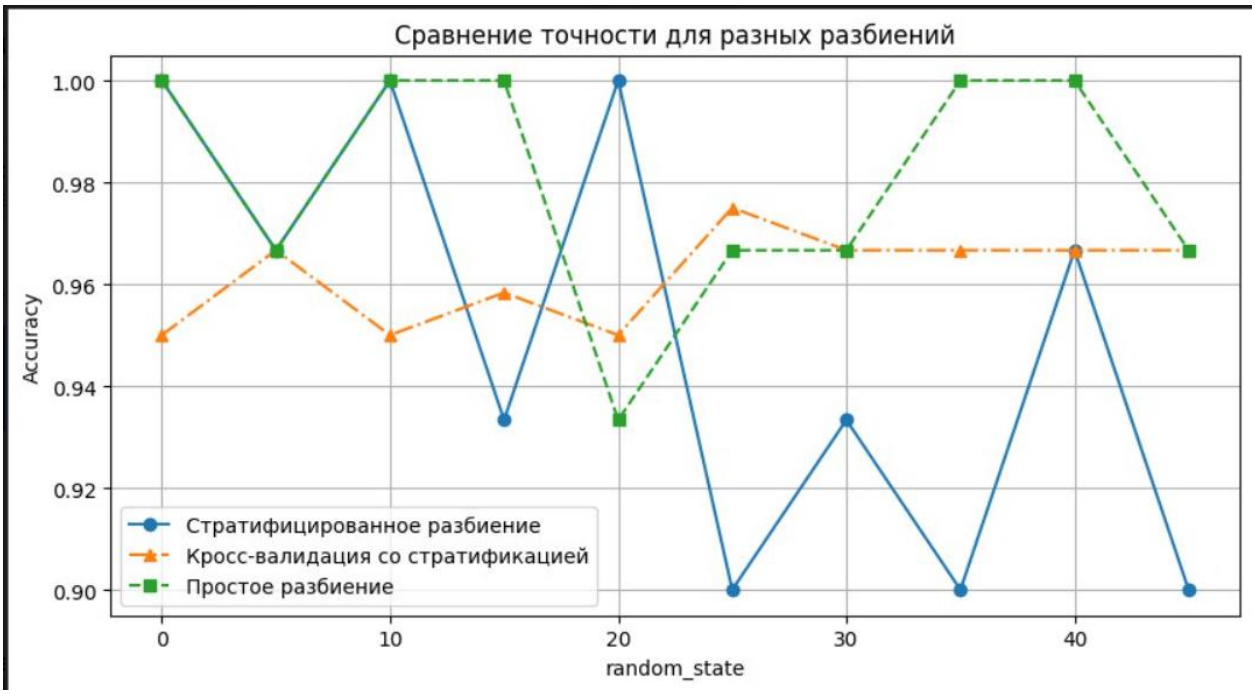


Рисунок 2. Сравнение точности для разных разбиений.

По полученным результатам можно сделать выводы, что хотя стратификация сохраняет пропорции классов в обучающей и тестовой выборках, это не гарантирует равномерное представление всех признаков в каждом разбиении. При малом размере датасета, как в данном случае (150 наблюдений), небольшие изменения в распределении объектов по обучающей и тестовой выборкам могут значительно повлиять на результат. Также если в классе, который мало представлен, будет находиться большое количество выбросов, только стратификация может усугубить ситуацию.

Простое разбиение также продемонстрировало свою нестабильность.

А вот кросс-валидация уменьшает зависимость от случайности разбиения данных, так как модель оценивается несколько раз на разных подвыборках. Это делает метод более надежным, особенно при небольших датасетах. Стратификация в кросс-валидации помогает гарантировать равномерное распределение классов во всех разбиениях. Стандартное отклонение самое низкое среди всех методов, что говорит о стабильности результатов.

**Список литературы:**

1. Cross-Validation. [Электронный ресурс]. - URL: https://scikit-learn.ru/stable/modules/cross_validation.html (дата обращения: 20.02.2025).

2. LinearRegression. [Электронный ресурс]. - URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html (дата обращения: 20.02.2025).

3. Load_Iris. [Электронный ресурс]. - URL: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html (дата обращения: 20.02.2025).

4. StratifiedKFold. [Электронный ресурс]. - URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html (дата обращения: 20.02.2025).