|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

**Факультет «Информатика и системы управления»**

**Кафедра «Системы обработки информации и управления»**

Отчет по лабораторной работе №3

**«Подготовка обучающей и тестовой выборки, кросс-валидация и подбор гиперпараметров на примере метода ближайших соседей»**

по дисциплине «Технологии машинного обучения»

Выполнил:

студент группы ИУ5Ц-84Б

Падалко К.Р.

подпись, дата

Проверил:

к.т.н., доц., Ю.Е. Гапанюк

подпись, дата

2025 г.

**СОДЕРЖАНИЕ ОТЧЕТА**

[1. Цель лабораторной работы 3](#_Toc161592050)

[2. Задание 3](#_Toc161592051)

[3. Основные характеристики датасета 3](#_Toc161592052)

[4. Листинг 4](#_Toc161592053)

[4.1. Изучение данных 4](#_Toc161592054)

[4.2. Преобразование данных 6](#_Toc161592055)

[4.3. Описательная статистика 7](#_Toc161592056)

[4.4. Предобработка данных 8](#_Toc161592057)

[4.4.1. Пропущенные значения 8](#_Toc161592058)

[4.4.2. Дубликаты 9](#_Toc161592059)

[4.5. Отсев до определенного количества уникальных значений 9](#_Toc161592060)

[4.6. Машинное обучение 10](#_Toc161592061)

[4.6.1. Деление на обучающей и тестовой выборки 10](#_Toc161592062)

[4.7. Кодирование признаков – прямое кодирование (One-Hot Encoding) 10](#_Toc161592063)

[4.8. Обучение модели 11](#_Toc161592064)

[4.8.1. KNeighborsRegressor 11](#_Toc161592065)

[4.9. Итог 12](#_Toc161592066)

[4.9.1. Анализ моделей 12](#_Toc161592067)

[5. Вывод 12](#_Toc161592068)

## Цель лабораторной работы

Изучение способов подготовки выборки и подбора гиперпараметров на примере метода ближайших соседей.

## Задание

1. Выберите набор данных (датасет) для решения задачи классификации или регрессии.
2. В случае необходимости проведите удаление или заполнение пропусков и кодирование категориальных признаков.
3. С использованием метода train\_test\_split разделите выборку на обучающую и тестовую.
4. Обучите модель ближайших соседей для произвольно заданного гиперпараметра K. Оцените качество модели с помощью подходящих для задачи метрик.
5. Произведите подбор гиперпараметра K с использованием GridSearchCV и RandomizedSearchCV и кросс-валидации, оцените качество оптимальной модели. Используйте не менее двух стратегий кросс-валидации.
6. Сравните метрики качества исходной и оптимальной моделей.

## Основные характеристики датасета

Название датасета: Набор данных о видах ирисов.

Ссылка: https://www.kaggle.com/datasets/uciml/iris

**О датасетах**

Этот набор данных содержит информацию о различных аспектах ирисов (цветков) из трех видов: Setosa, Versicolor и Virginica. В наборе представлены характеристики, такие как длина и ширина чашелистика и лепестка для 150 образцов ирисов. Данные используются для классификации видов ирисов на основе этих характеристик.

Набор данных включает 150 строк, каждая из которых представляет один ирис, и 5 столбцов.

Этот датасет использован для задач классификации и обучения моделей машинного обучения, таких как k-ближайших соседей, дерева решений, логистической регрессии и других методов классификации.

**Структура данных**

sepal length (длина чашелистика) — измеряется в сантиметрах.

sepal width (ширина чашелистика) — измеряется в сантиметрах.

petal length (длина лепестка) — измеряется в сантиметрах.

petal width (ширина лепестка) — измеряется в сантиметрах.

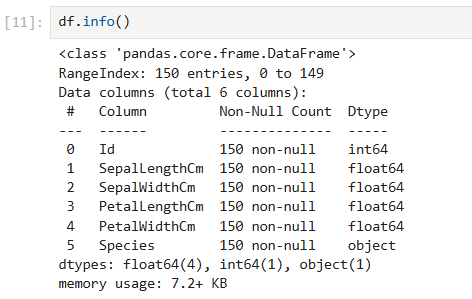
species (вид) — категориальная переменная, указывающая на вид ириса, который представлен в строке (Setosa, Versicolor или Virginica).

**Выбор признаков для машинного обучения**

Для машинного обучения выберем целевой признак – вид ирисов. Сопоставим с остальными признаками, а именно, характеристики цветов, вывялим примерный вид ириса.

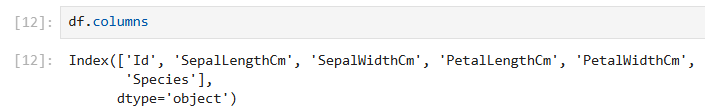
## Листинг

## Изучение данных

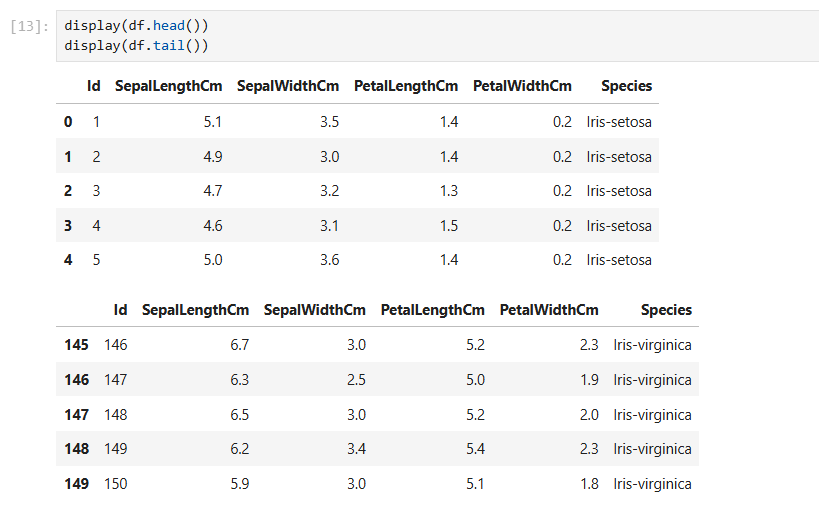


В датасете содержатся 150 строки, имеются 3 различные типы: object, int64 и float64.

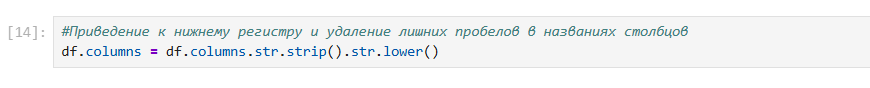
Просмотр названий столбцов.



Первые и последние пять строк датасета.

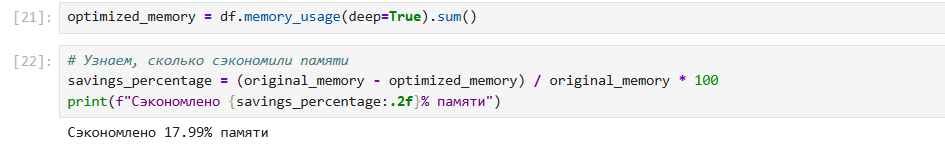
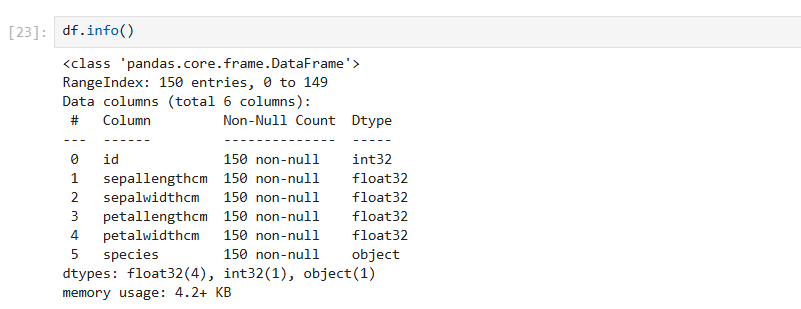


Приведем к нижнему регистру и удаление лишних пробелов в названиях столбцов.

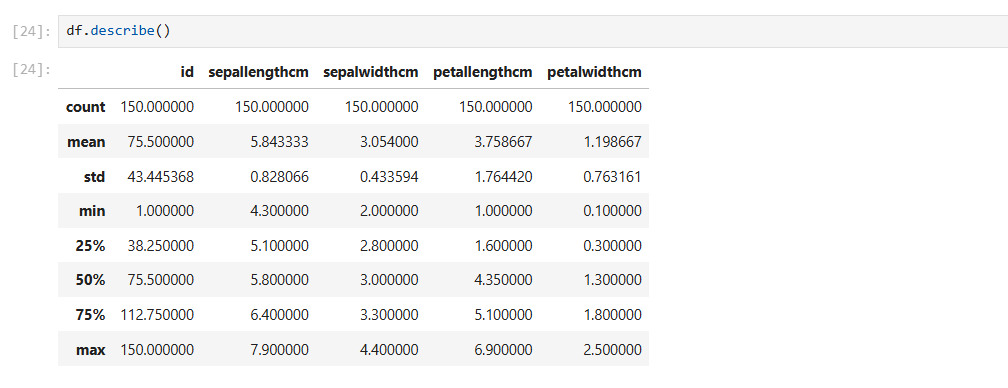


## Преобразование данных



## Описательная статистика



**id:** Это уникальные идентификаторы записей в данных. Каждая строка представляет отдельный образец ириса, и этот столбец не несет дополнительной информации о характеристиках цветов. Он служит лишь для идентификации строки в наборе данных.

**sepallengthcm:** Это длина чашелистика цветка ириса, измеренная в сантиметрах. Среднее значение длины чашелистика составляет 5.84 см. Диапазон значений от 4.3 см до 7.9 см, что указывает на разнообразие в длине чашелистика среди разных видов ирисов.

**sepalwidthcm:** Это ширина чашелистика цветка ириса, измеренная в сантиметрах. Средняя ширина чашелистика составляет 3.05 см. Значения варьируются от 2.0 см до 4.4 см, что показывает, что ширина чашелистика также имеет значительные колебания среди ирисов.

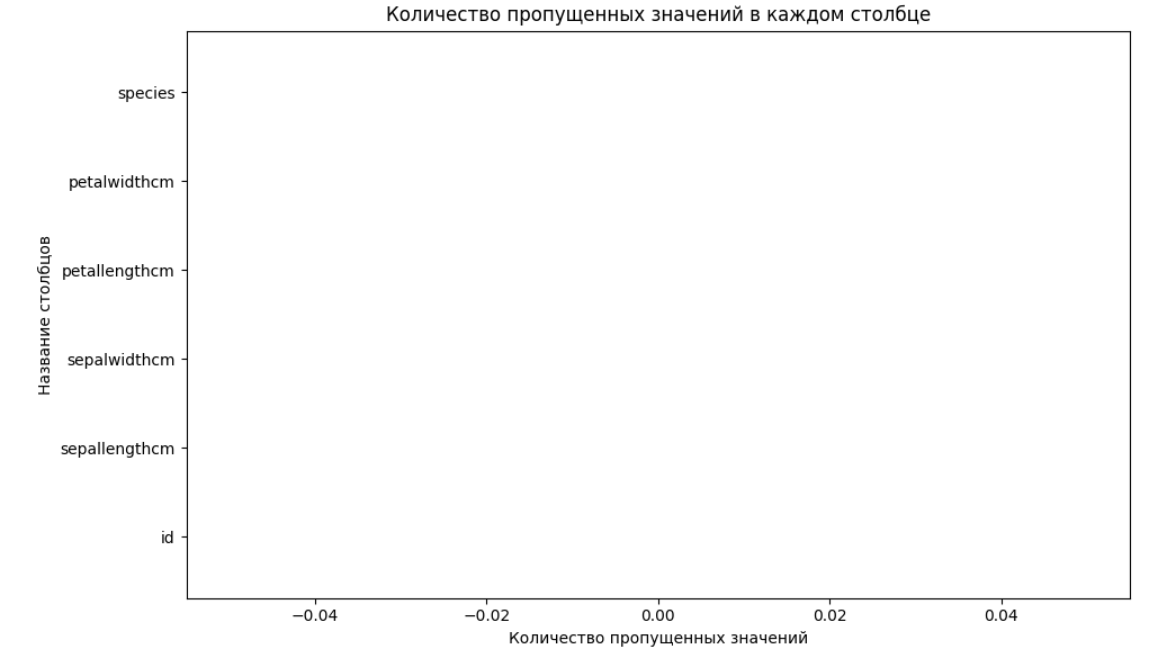
**petallengthcm:** Это длина лепестка цветка ириса, измеренная в сантиметрах. Средняя длина лепестка составляет 3.76 см. Длина лепестков варьируется от 1.0 см до 6.9 см, с большими различиями между образцами, что может указывать на разнообразие форм лепестков в зависимости от вида ириса.

**petalwidthcm:** Это ширина лепестка цветка ириса, измеренная в сантиметрах. Средняя ширина лепестка составляет 1.20 см. Значения варьируются от 0.1 см до 2.5 см, что также указывает на значительный разброс в характеристиках лепестков среди разных видов ирисов.

## Предобработка данных

## Пропущенные значения

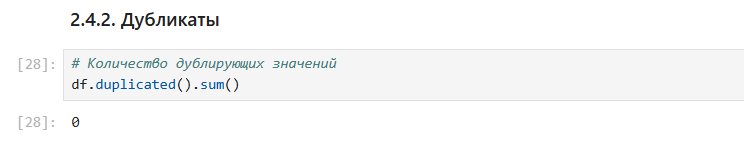




В диаграмме отсутствуют пропуски в столбцах.



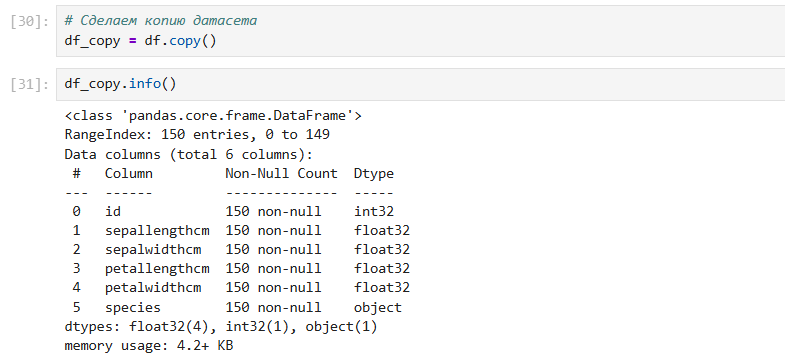
## Дубликаты

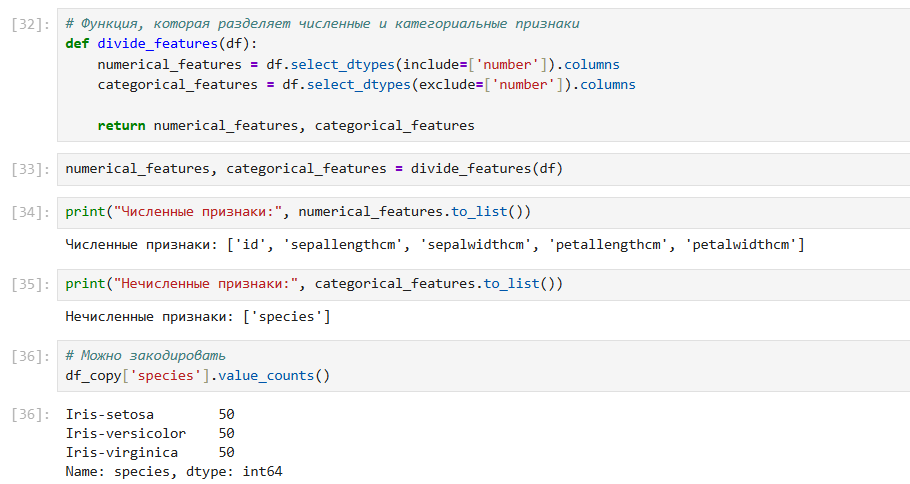


Дубликатов нет, это говорит о том, что датасет был хорошо сделан.

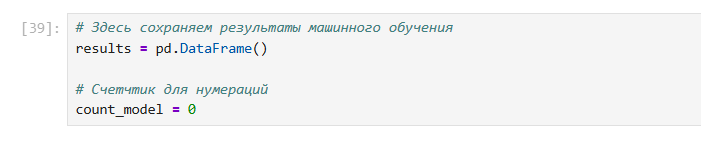
## Отсев до определенного количества уникальных значений

Для кодирования признаков OHE или OH будет черевато, если оставить много уникальных названий, т.к. это приведет к созданию много новых признаков. Поэтому отсеиваем до небольших количеств, так чтобы создали максимум небольших новых закодированных признаков.

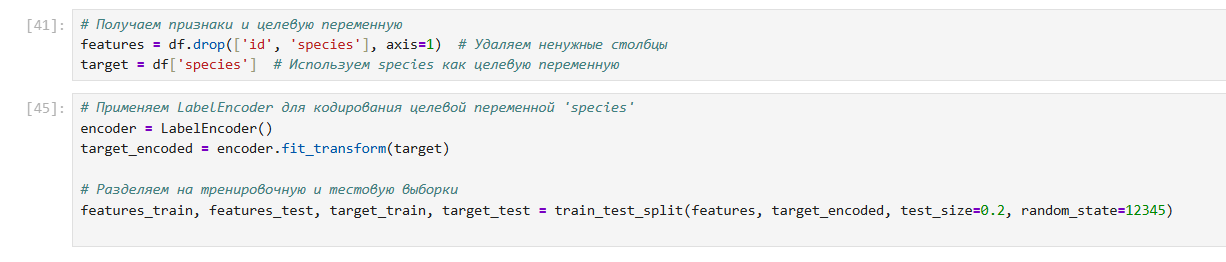




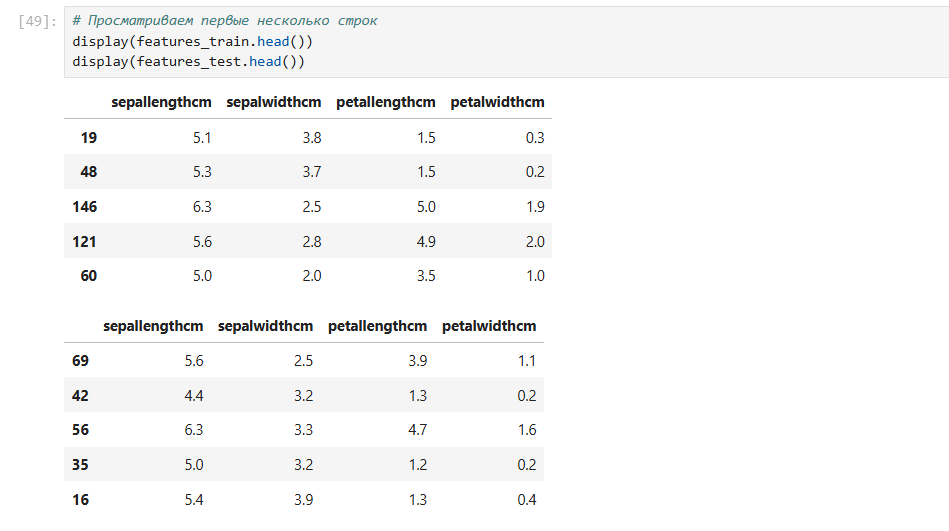
## Машинное обучение



## Деление на обучающей и тестовой выборки



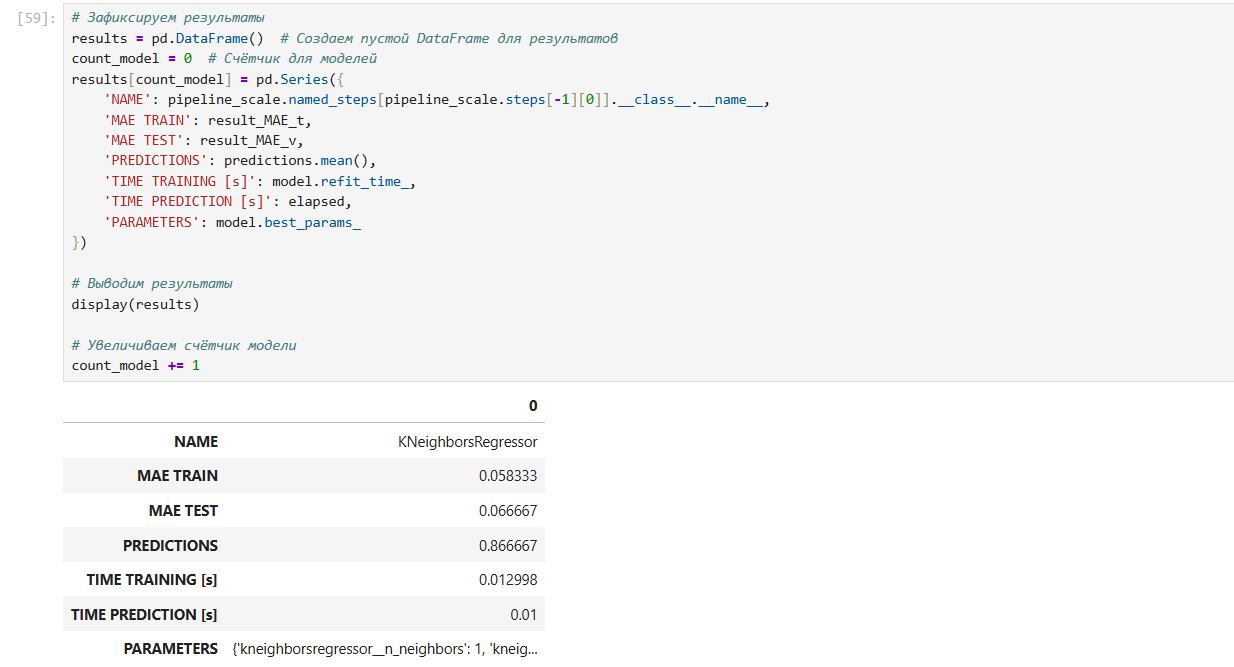
## Кодирование признаков – прямое кодирование (One-Hot Encoding)



## Обучение модели

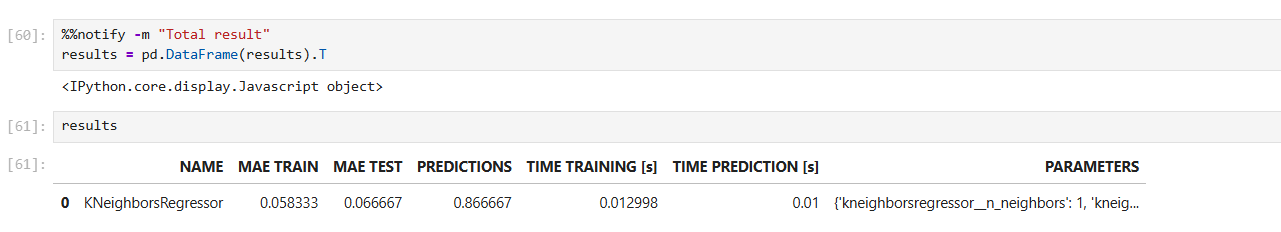
## KNeighborsRegressor



## Итог

## Анализ моделей



## Вывод

В ходе лабораторной работы изучили способы подготовки выборки и подбора гиперпараметров на примере метода ближайших соседей.