|  |  |
| --- | --- |
| 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 РК2

«*Методы построения моделей машинного обучения*»

по дисциплине «Технологии машинного обучения»

Выполнила:

студентка группы № ИУ5-62Б Кулькина Дарья Александровна

30.05.2023

Проверил:

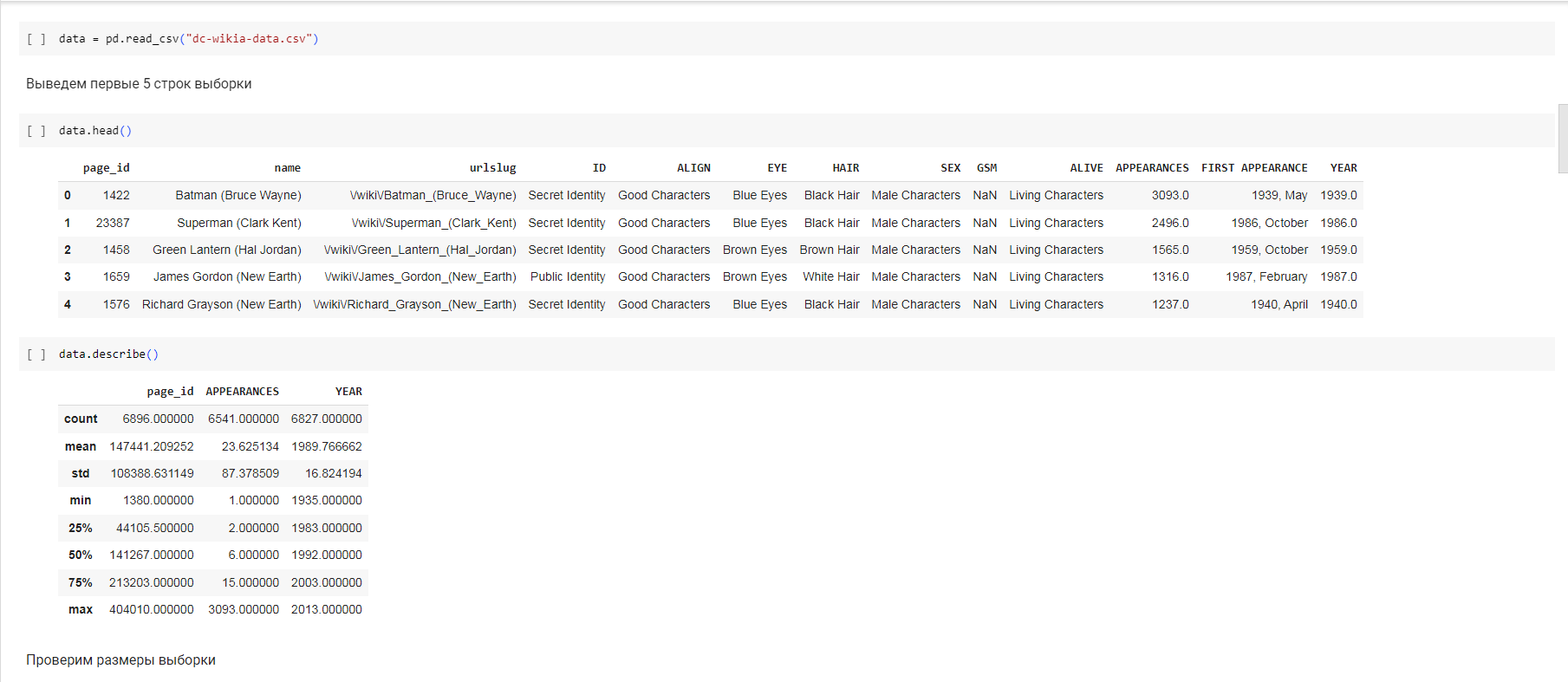
к.т.н., доц., Ю. Е. Гапанюк

2023 г.

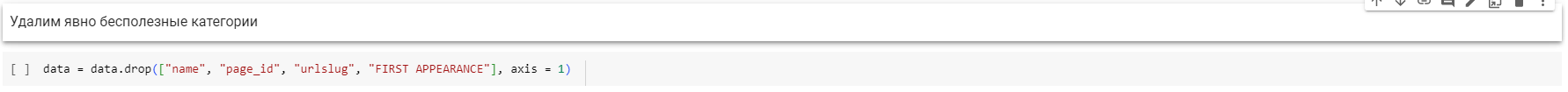
**Описание задачи.** Для заданного набора данных (по Вашему варианту) постройте модели классификации или регрессии (в зависимости от конкретной задачи, рассматриваемой в наборе данных). Для построения моделей используйте методы 1 и 2 (по варианту для Вашей группы). Оцените качество моделей на основе подходящих метрик качества (не менее двух метрик). Какие метрики качества Вы использовали и почему? Какие выводы Вы можете сделать о качестве построенных моделей? Для построения моделей необходимо выполнить требуемую предобработку данных: заполнение пропусков, кодирование категориальных признаков, и т.д. **Вариант № 12**

**Описание датасета.** В работе был использован датасет FiveThirtyEight Comic Characters Dataset. Датасет состоит из двух файлов: dc-wikia-data.csv и marvel-wikia-data.csv. Согласно заданию, берём файл dc-wikia-data.csv. Файл содержит следующие колонки:

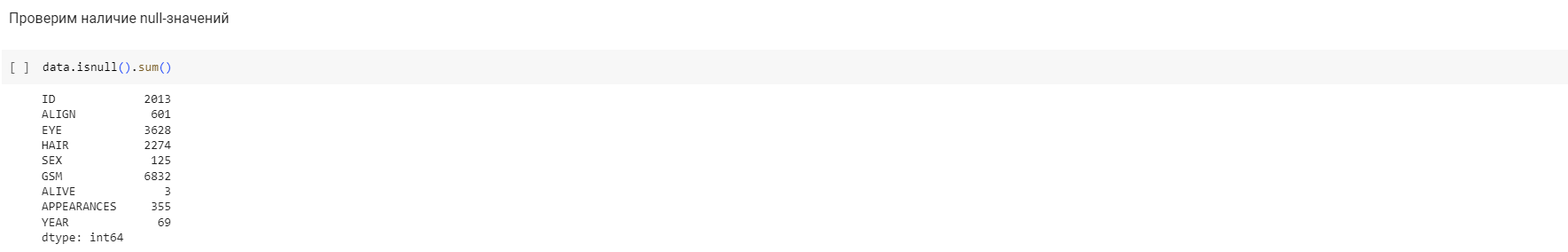
1. page\_id (int64) - Уникальный идентификатор для этой страницы символов в wikia
2. name (string) - Имя персонажа
3. urlslug (string) - Уникальный URL-адрес в викии, который приведет вас к персонажу
4. ID (string) - Идентификационный статус персонажа (Секретная личность, публичная личность)
5. ALIGN (string) - Если персонаж хороший, плохой или нейтральный
6. EYE (string) - Цвет глаз персонажа
7. HAIR (string) - Цвет волос персонажа
8. SEX (string) - Пол персонажа (например, мужчина, женщина и т.д.)
9. GSM (string) - Если персонаж принадлежит к гендерному или сексуальному меньшинству (например, гомосексуалисты, бисексуалы)
10. ALIVE (string) - Если персонаж жив или умер
11. APPEARANCES (int64) - Количество появлений персонажа в комиксах (по состоянию на 2 сентября 2014 года. С течением времени это число будет становиться все более устаревшим.)
12. FIRST APPEARANCE (string) - Месяц и год первого появления персонажа в комиксе, если таковые имеются
13. YEAR (int64) - Год первого появления персонажа в комиксе, если таковой имеется

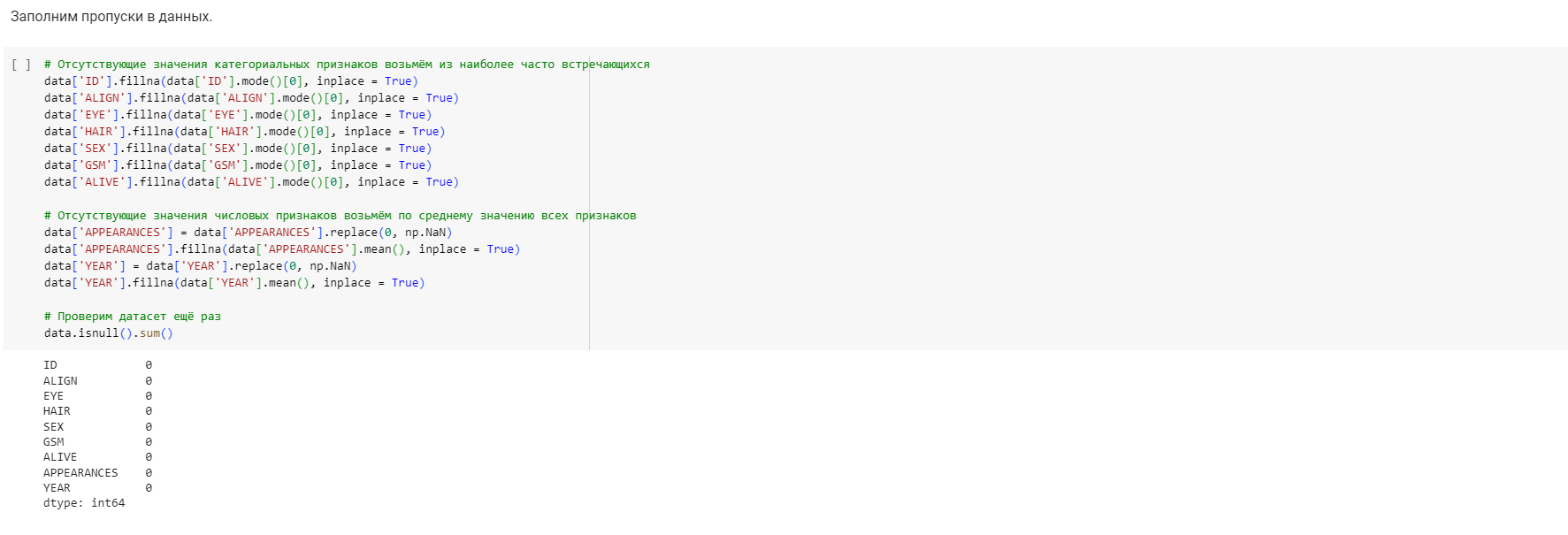


**Подготовка датасета к обучению модели.** Колонки name, page\_id, urlslug, FIRST APPEARANCE были удалены.



После удаления колонок выполним заполнение пропусков в данных.

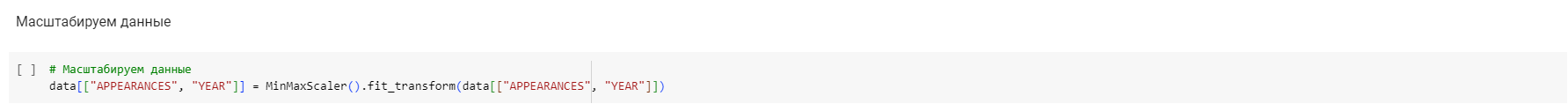




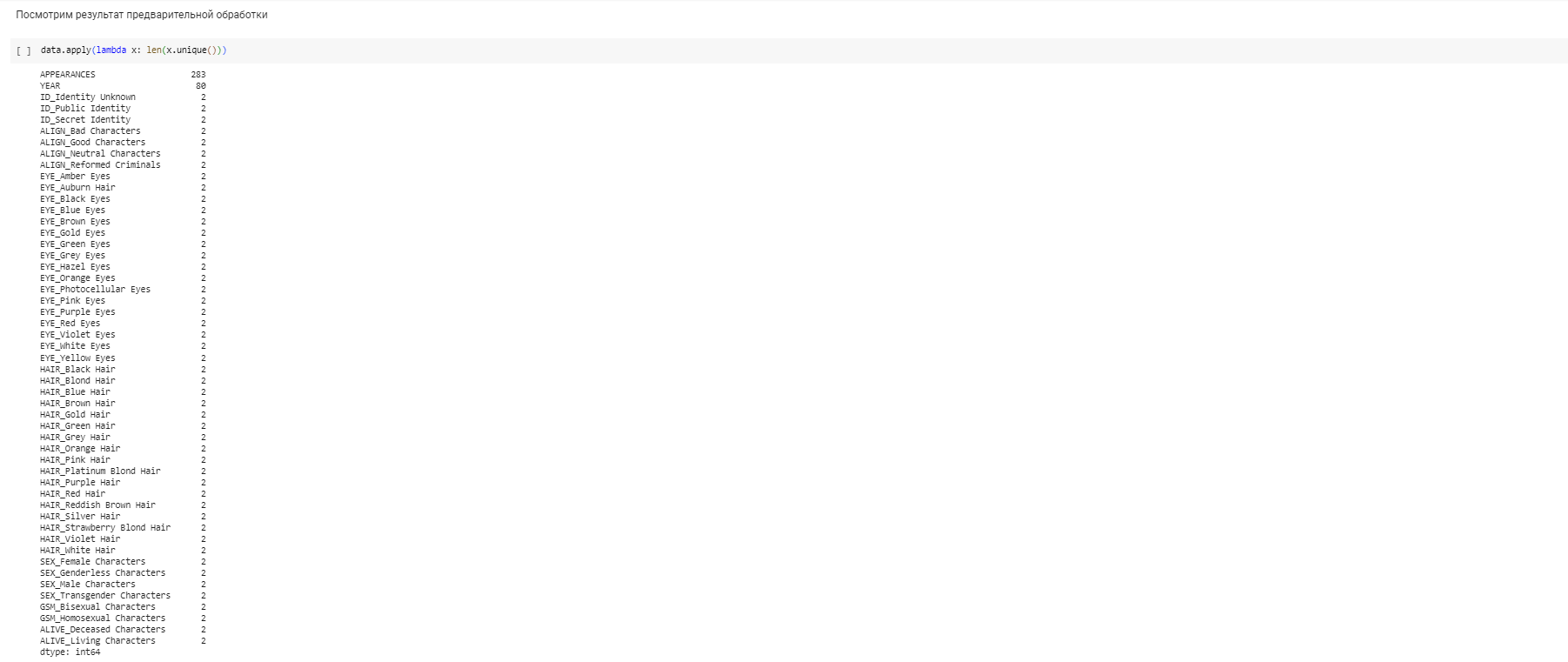
Далее кодируем категориальные признаки.



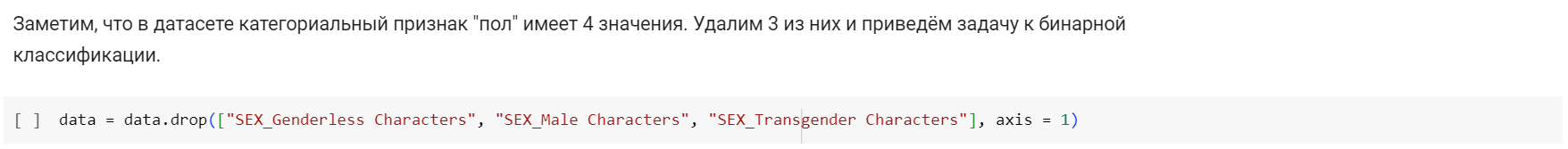
После кодирования категориальных признаков выполним масштабирование данных в колонках APPEARANCES и YEAR.



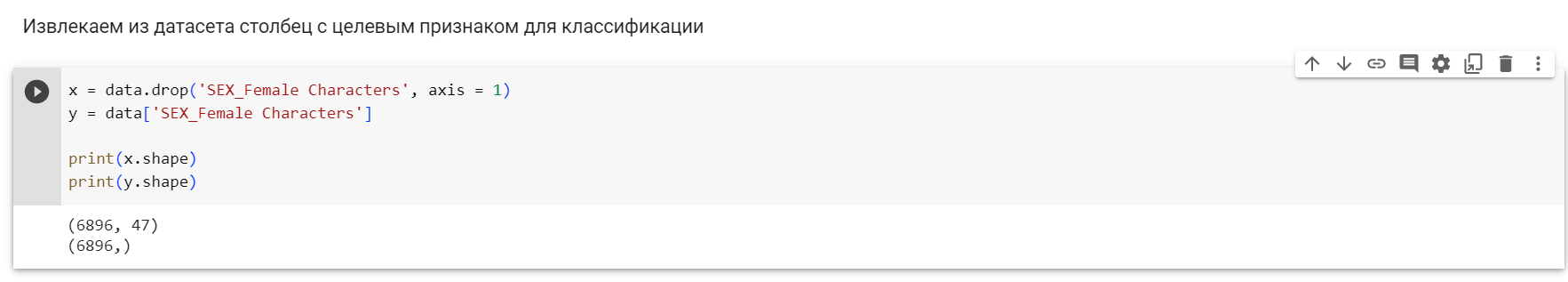
Проверим уникальные значения по столбцам датасета.



Удалим лишние категории в признаке SEX.



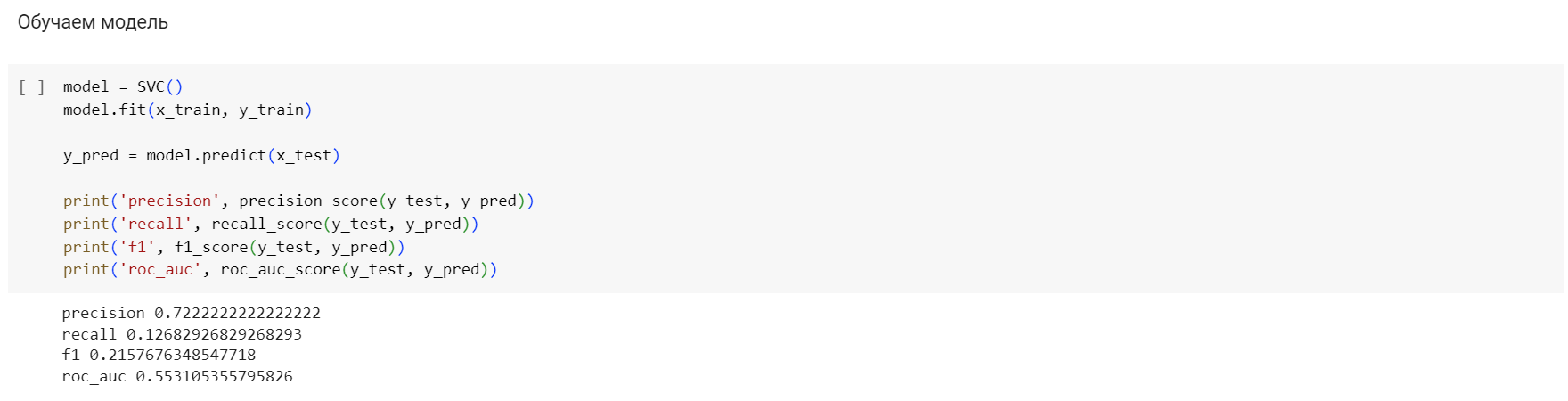
Извлечём столбец целевого признака для классификации (SEX\_Female Characters).



Делим датасет на обучающую и тестовую выборки.



**Обучаем модель классификатора (SVC).**



В качестве метрик качества будем использовать precision, recall, f1 и roc\_auc. Совокупность этих метрик даёт возможность достаточно точно определить качество классификатора.

Заметим, что значения метрик recall и f1 говорят о низкой точности модели, несмотря на достаточно хороший результат по метрике precision. Возможно, SVC не является оптимальной моделью для поставленной задачи.

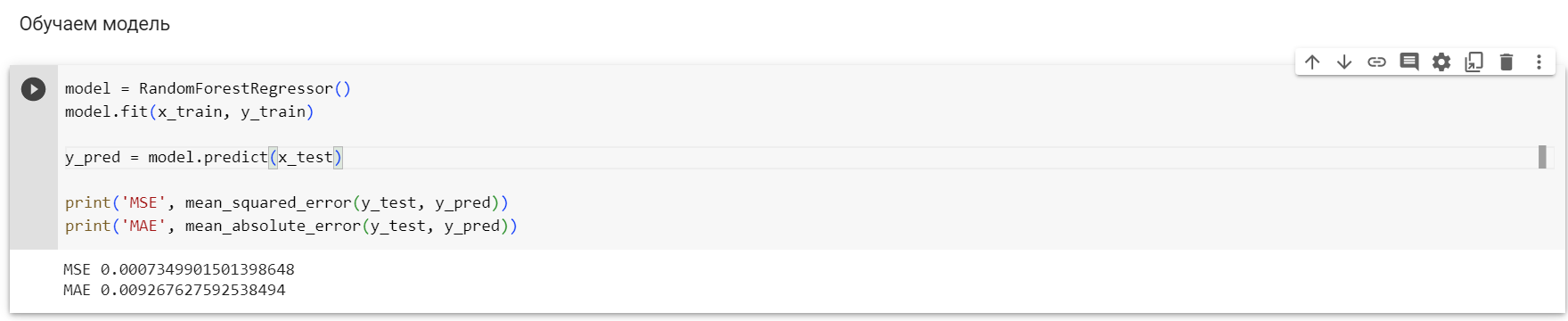
**Обучаем регрессионную модель (RandomForestRegressor).** Выделим из датасета целевой признак APPEARANCES.



Разделим датасет на тестовую и обучающую выборки.



Обучаем модель и выводим значения метрик для модели на тестовой выборке.



В качестве метрик качества для регрессионной модели будем использовать MSE и MAE. Значения метрик позволяют сделать вывод, что полученная модель работает достаточно точно.