|  |  |  |
| --- | --- | --- |
| Лабораторная работа №1  «Изучение инструментов интеллектуальной обработки данных (Google Colab)» | Ф.И.О. | Самороков Н. Н. |
| Группа | СИИ-3 |
| Преподаватель |  |
| Дата сдачи |  |

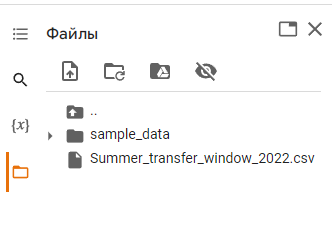
**Ход работы:**

**Этап 1 – загрузка файла**

Датасет - Summer\_transfer\_window\_2022

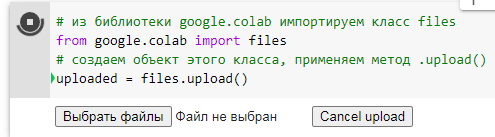
Этап 1 – подгрузка датасета

**Способ 1:**

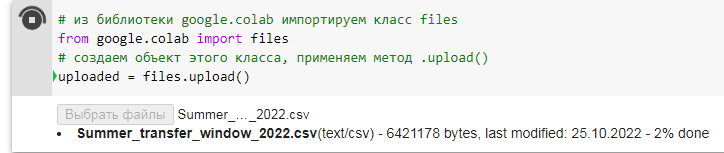


**Способ 2:**

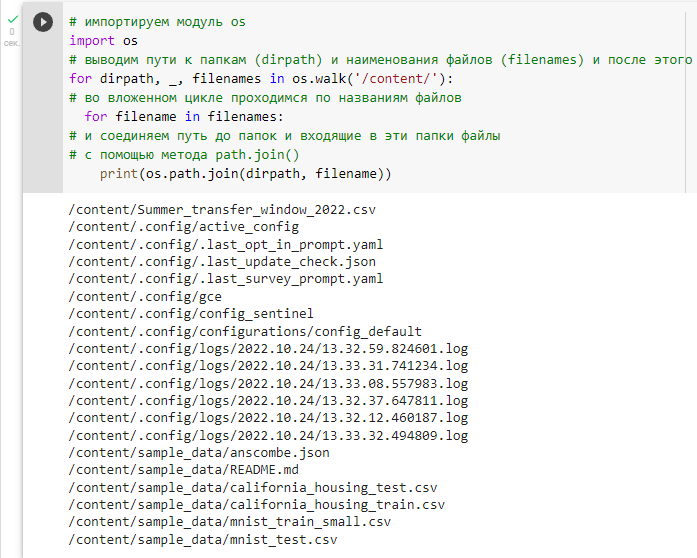
Пишем код для подгрузки файла



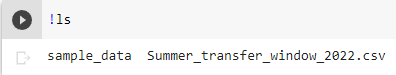
Выбираем файл



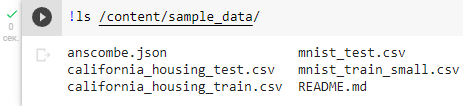
**Этап 2 – просмотр файлов папки /content/ с помощью модуля os и метода .walk/**



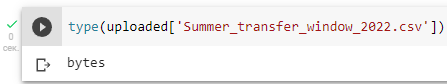
Команда !ls



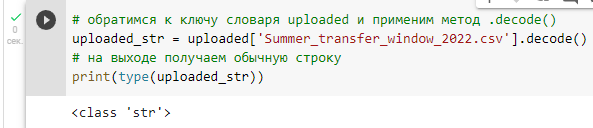
Пробуем заглянуть внутрь папки sampe\_data



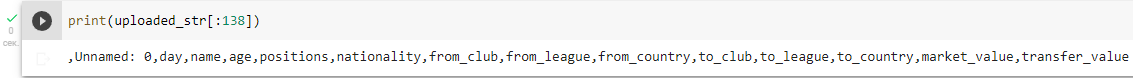
Чтение из переменной uploaded



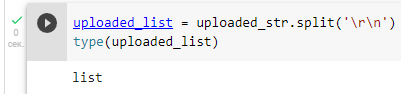
Преобразуем эти данные в обычную строку



Выведем первые 138 значений



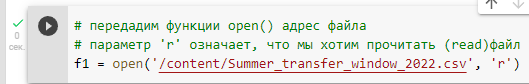
Разбиваем строку методом .split()



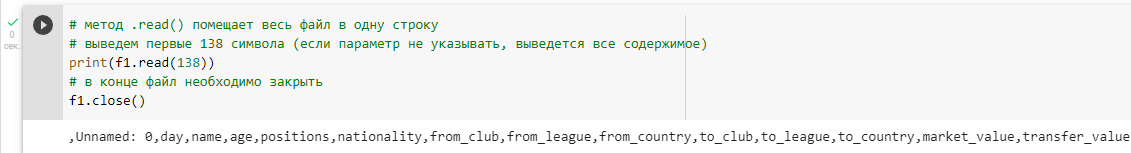
Пройдемся по списку и выведем первые 4 значения



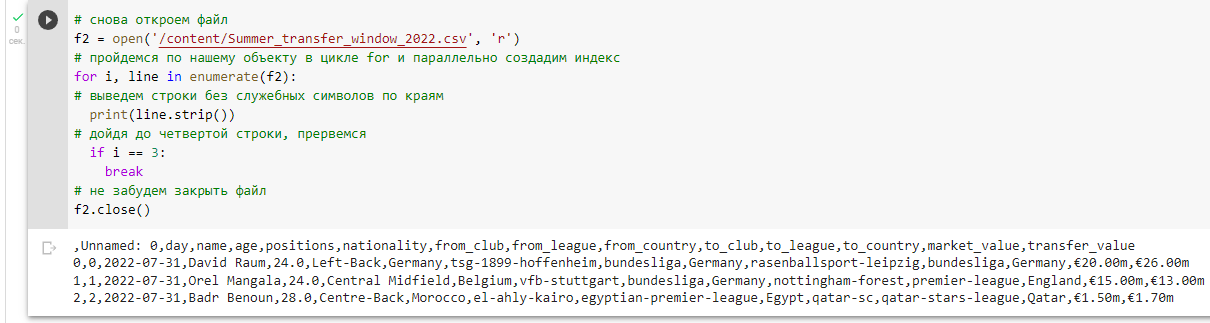
Воспользуемся функцией open()



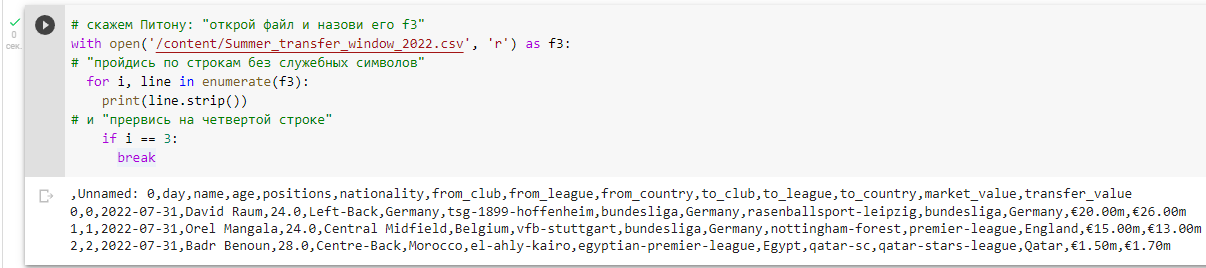
Применим метод .read()



Пройдемся по файлу for



Пройдемся по файлу с конструкцией with open()



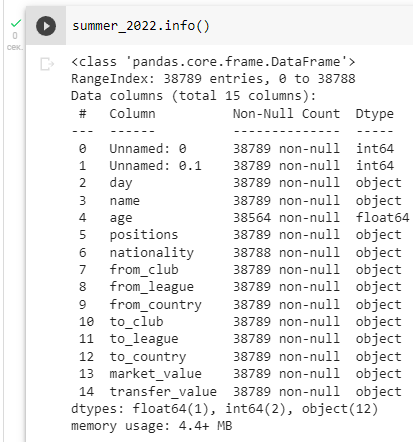
Чтение через библиотеку Pandas

Чтение файла файла через библиотеку Pandas



**Этап 3. Построение модели и прогноз**

Исследовательский анализ данных (EDA)



**Краткое описание каждой из переменных**

**!**

**!**

**!**

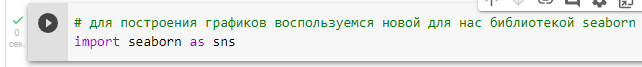
**!**

**!**

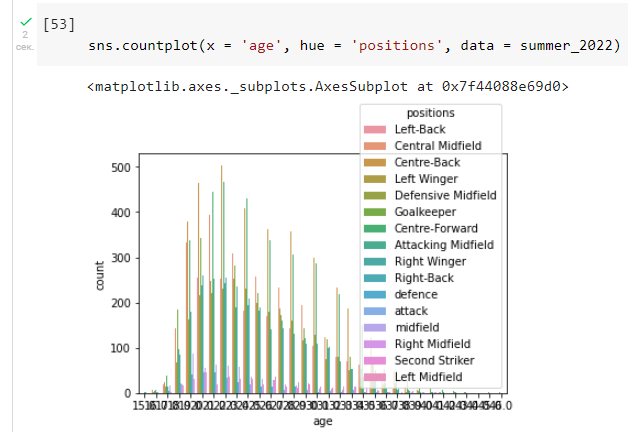
**!**

**!**

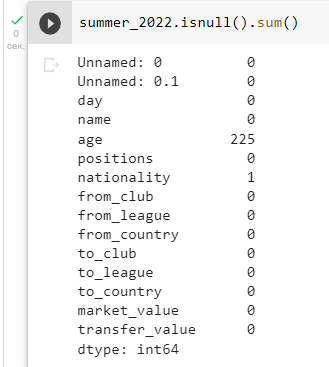
Подгрузим библиотеку seaborn



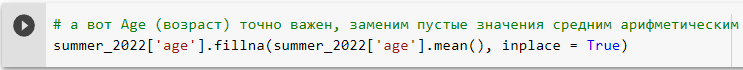
Отрисуем графические диаграммы (зависимость возраста футболиста от его позиции)



Посмотрим сколько каких значений было пропущено



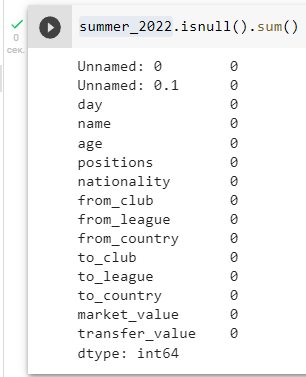
Заменим пустые значения в age средним арифметическим



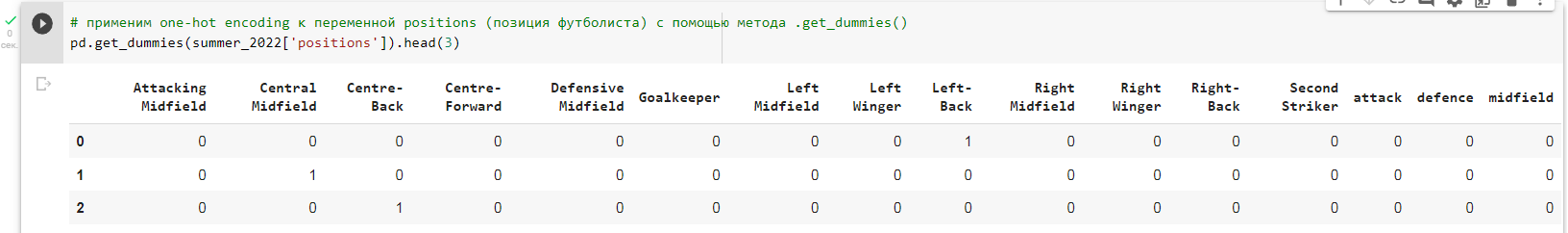
Удалим пустую строку в nationality



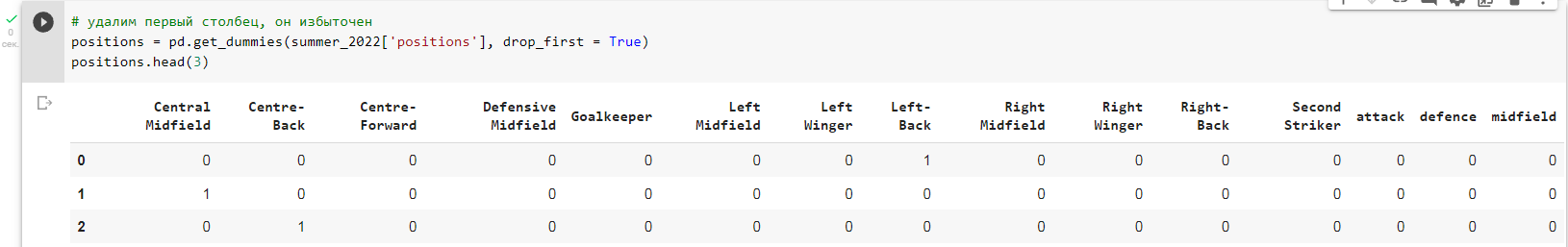
Проверим сколько пустых строк осталось



Применим one-hot с помощью метода .get\_dummies()



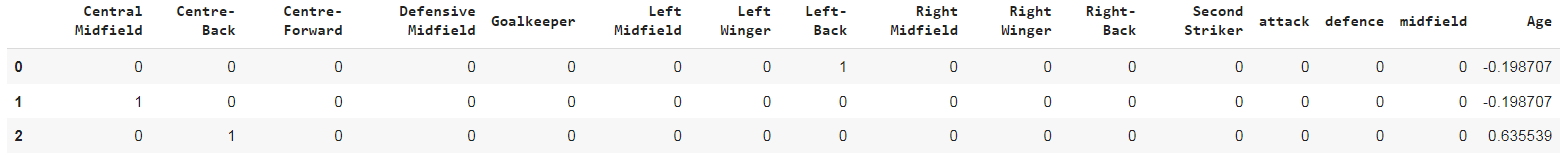
Удалим первый столбец, он избыточный



Добавим в positions значение возраста, который уже «нормализован» (необходим далее, для обучающей выборки)



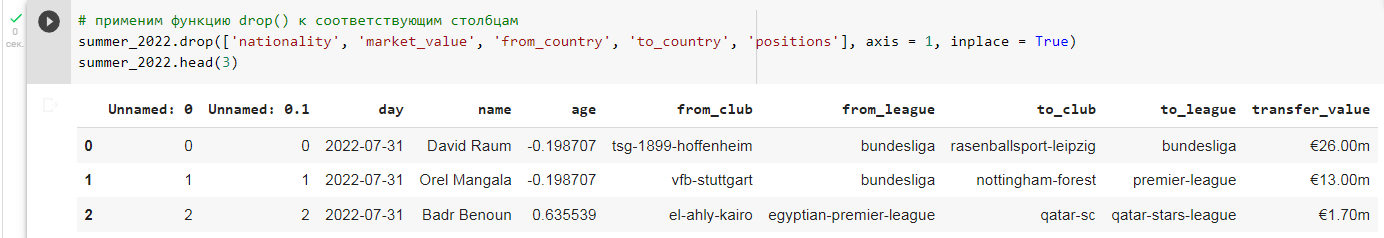
Получим такую таблицу



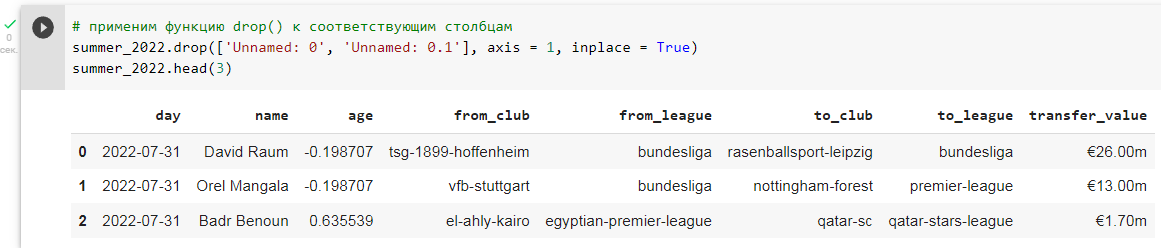
Присоединим новые переменные к исходному датафрейму



Отобрали те признаки, которые будем использовать в модели



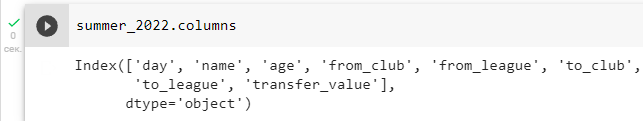
Доудаляем ненужные столбцы, которые не удалили ранее



Нормализуем данные (выбираем столбец “age”)



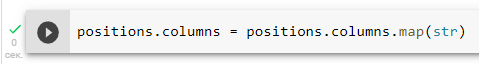
Проверим, что выражено строками, а что числами в нашей новой табличке



Проверим, на-всякий, что выражено строками в positions

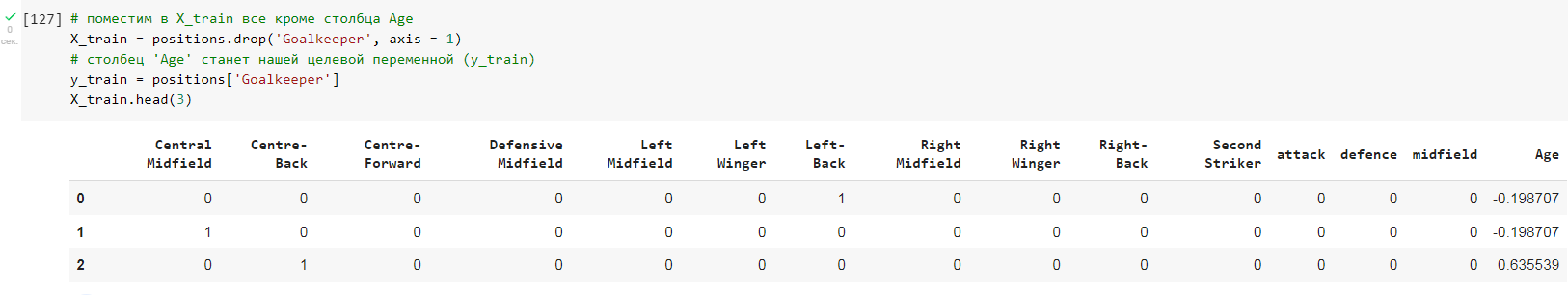


Преобразуем эти переменные в тип str



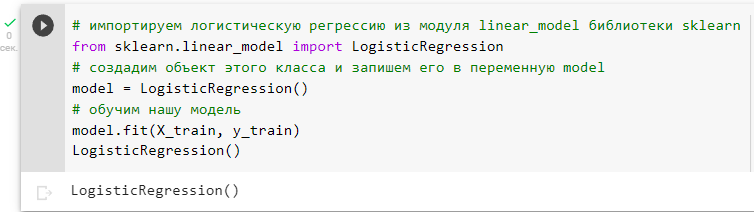
**Шаг 2. Разделение обучающей выборки на признаки и целевую переменную**

Поместили в переменную y\_train – значение роли футболиста = “Goalkeeper”(будем проводить обучающую выборку на примере positions)

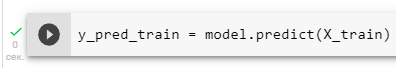


**Шаг 3. Обучение модели логической регрессии**

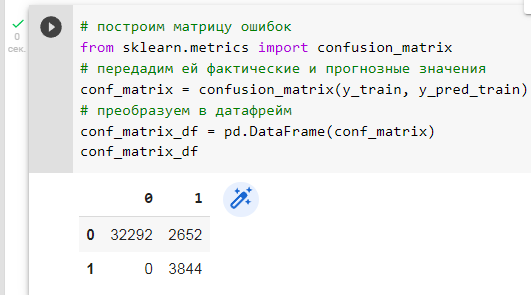
Воспользуемся моделью логистической регрессии из библиотеки sklearn и передадим ей обучающую выборку



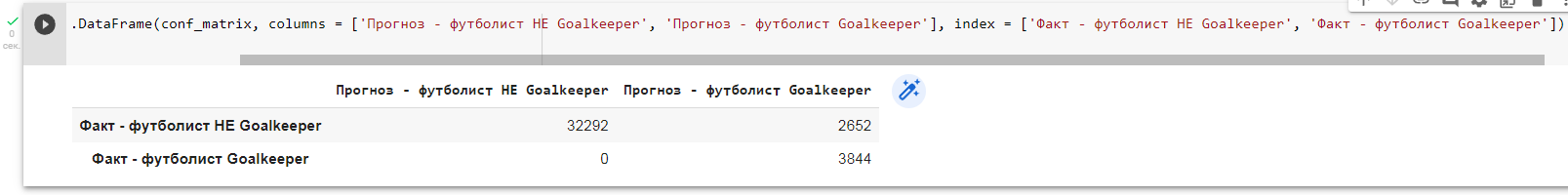
Сделаем предсказание на обучающей выборке



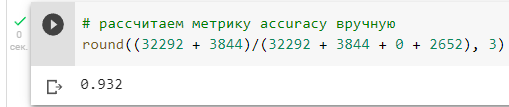
Построим матрицу ошибок



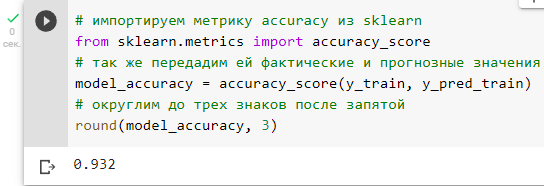
Для удобства интерпретации добавим подписи



Рассчёт метрики accuracy вручную



Используем метрику из sklearn

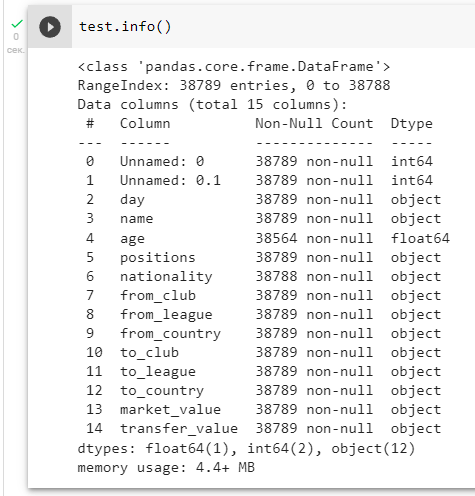


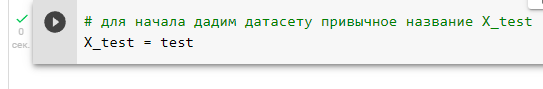
**Шаг 4. Построение прогноза на тестовых данных**

Предварительно, скопируем начальный датасет в test

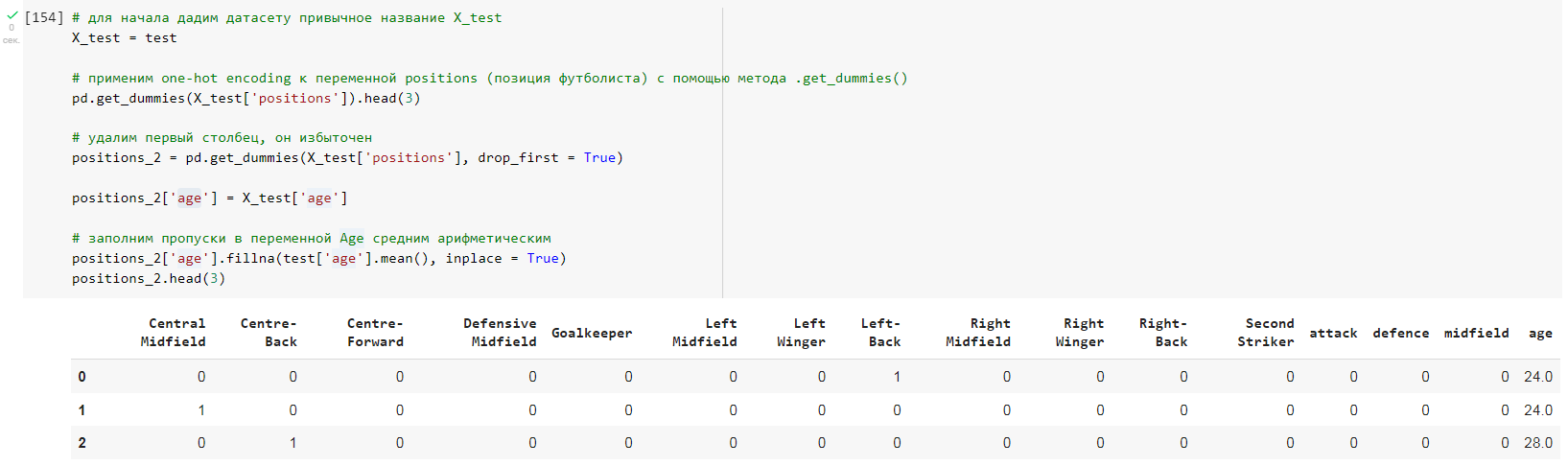


Просмотрим информацию о test

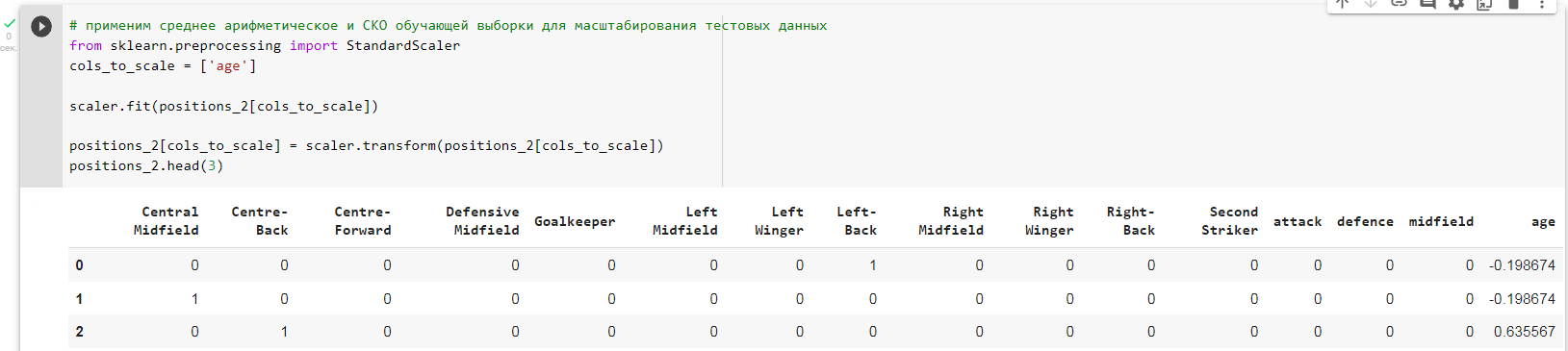




Подготовили тестовые данные к обработке



Применим среднее арифметическое и СКО обучающей выборки для масштабирования тестовых данных



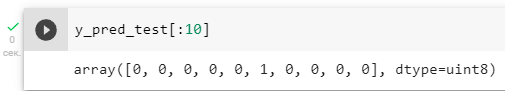
Превращаем название столбцов в строки



Забыли удалить поле «Goalkeeper»

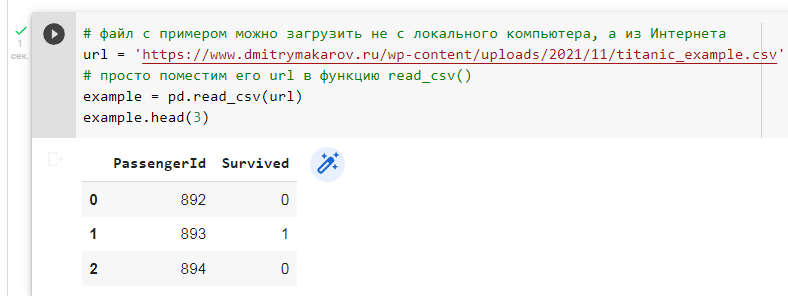


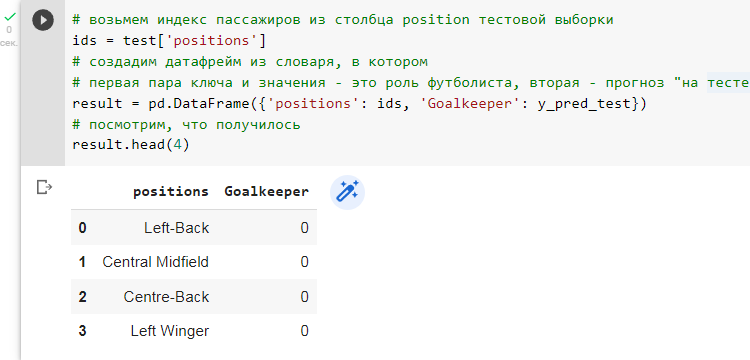
Делаем прогноз тестовой выборки



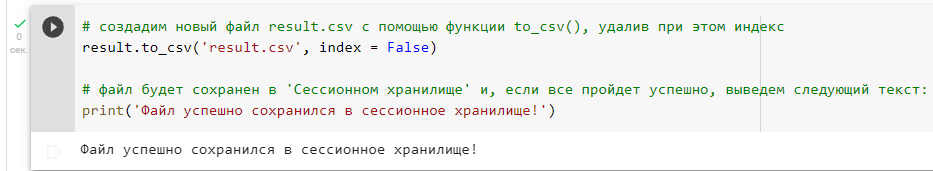
**Этап 4. Сохранение результата в новом файле на сервере.**

Можем загрузить файл из интренета

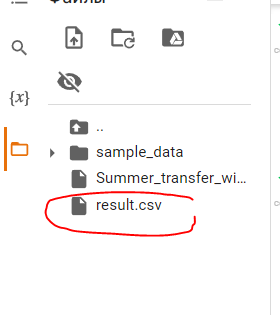




Сохранение файла в резервное хранилище



Результат сохранения



**Этап 5. Скачивание обратно на жесткий диск**

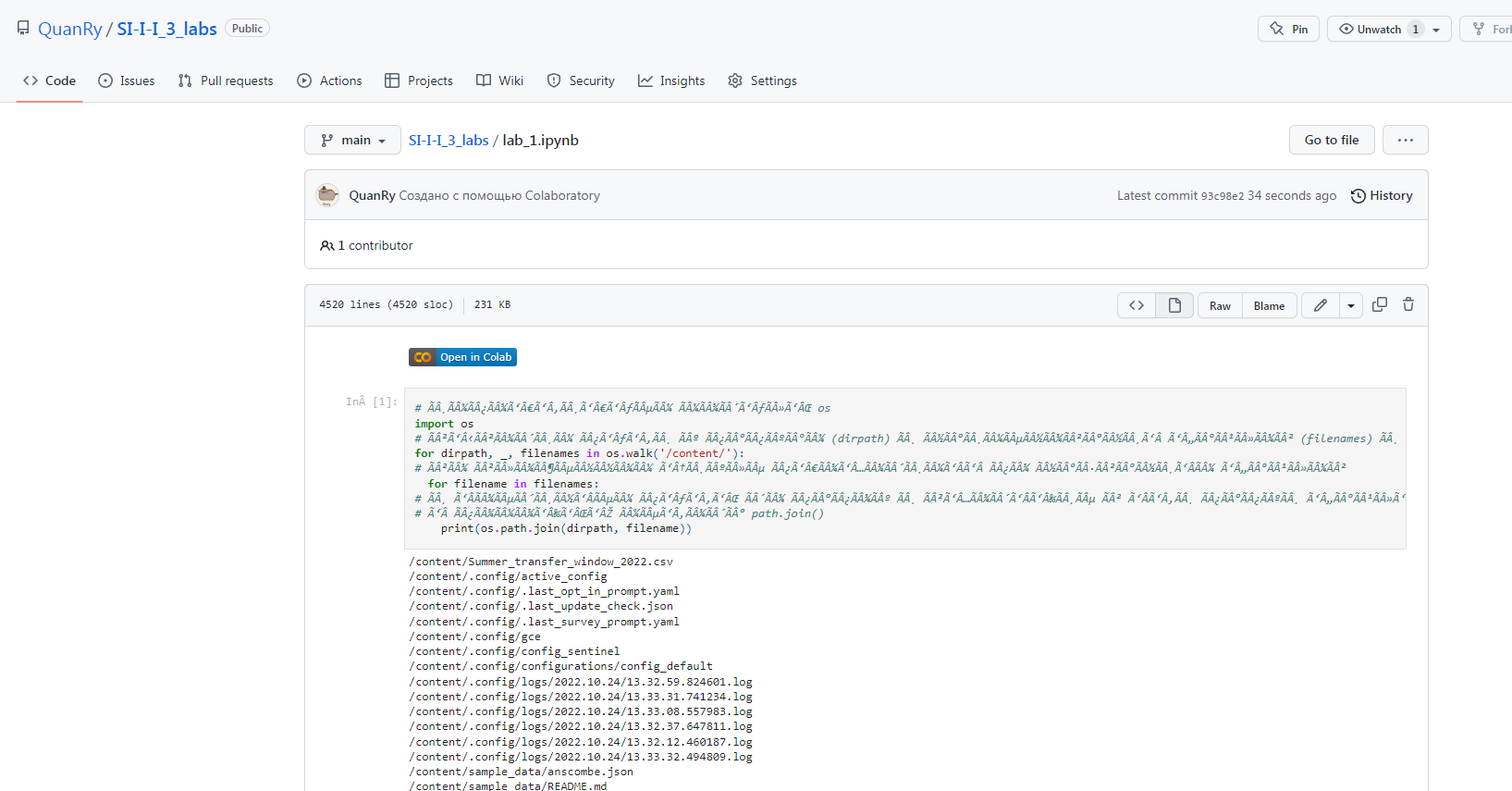
Скачивание результата на диск



Результат в result



Сохранение на GitHub



Ссылка на GitHub

https://github.com/QuanRy/SI-I-I\_3\_labs/blob/main/lab\_1.ipynb