Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования

«ВОЛГОГРАДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Цифровая кафедра

Лабораторная работа №3

по дисциплине «Системы искусственного интеллекта»

Выполнил:

студент

группы АУЗ-561

Жолобов А.В.

Проверил:

преподаватель

Фокин Р.О.

Волгоград, 2023 г.

|  |  |  |
| --- | --- | --- |
| Лабораторная работа №3 | Ф.И.О. | Жолобов А.В. |
| Группа | АУЗ-561 |
| Преподаватель | Фокин Р.О. |
| Дата сдачи | 12.11.2023 |

**Цель работы:**

Целью данной работы является инструментов интеллектуальной обработки данных (Google Colab).

**Задачи:**

1. Изучение Google Colab
2. Загрузка данных из датасета.

Датасеты:

Kaggle – https://www.kaggle.com/datasets

Dataset Search – https://datasetsearch.research.google.com

GitHub – https://github.com/awesomedata/awesome-public-datasets

1. Изучение
   1. Построение статистики
   2. Построение графиков
2. Сохранение проекта

**Ход работы:**

Пошаговое выполнение примера из презентации:

Лабораторная работа №3.

**Выполнение индивидуального задания:**

1. Подзагрузка файлов в Google Colab.

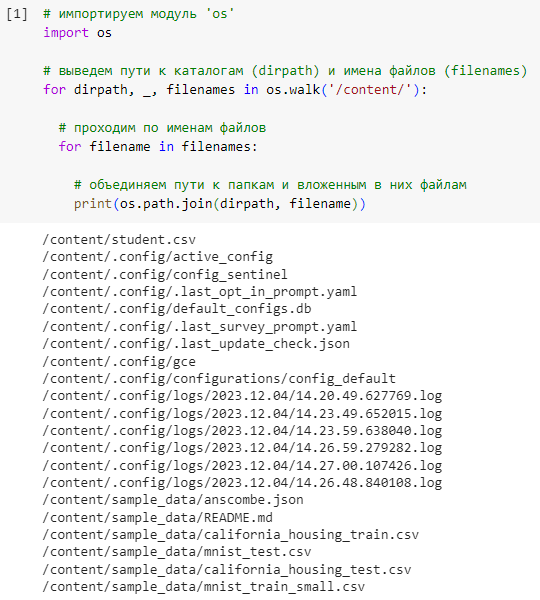
В ходе выполнения контрольной работы был использован Student Classification Dataset:

<https://www.kaggle.com/datasets/jacksondivakarr/student-classification-dataset>.

Этот набор данных охватывает различные аспекты, связанные с успеваемостью учащихся.

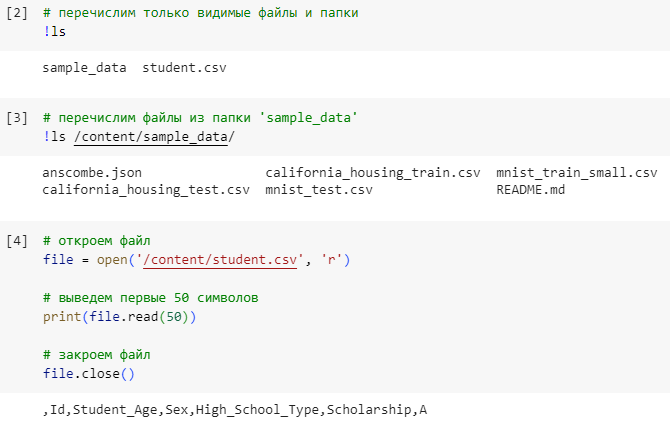
2. Чтение файла.

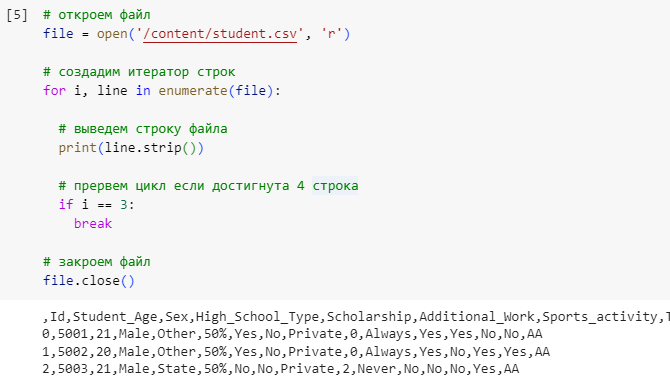
Просмотрим содержимое в папке **/content/**.

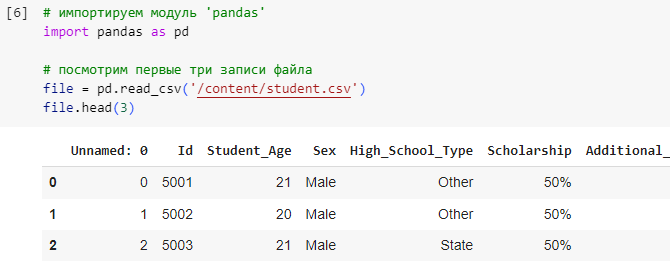


Код успешно обходит все файлы в директории «/content/» и её поддиректориях, выводя полные пути к каждому из них. Одним из файлов в этой директории является «/content/superstore.csv», который представляет собой загруженный файл Global Superstore Dataset.

Рассмотрим различные виды предоставления информации о файлах.







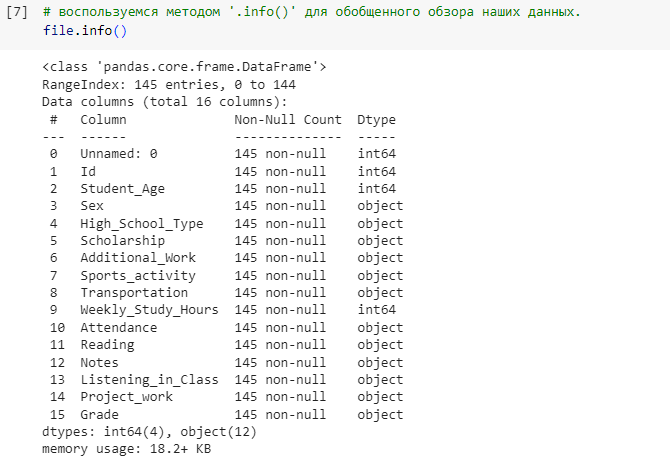
3. Построение модели и прогноз.

Разработаем небольшую модель, которая будет предсказывать оценку и обеспечит всесторонний обзор успехов учащихся.

2.1. Обработка и анализ данных.

2.1.1. Исследовательский анализ данных (EDA).

Выявим взаимосвязь между признаками и целевой переменной.

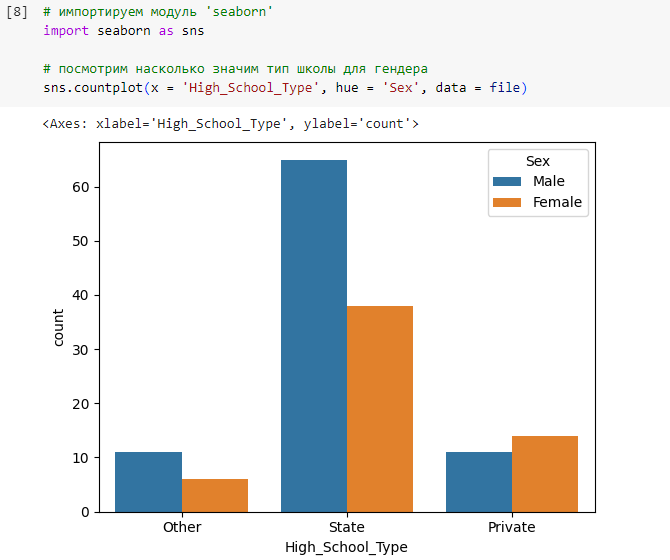


Dataset содержит 145 записей и 16 столбцов.

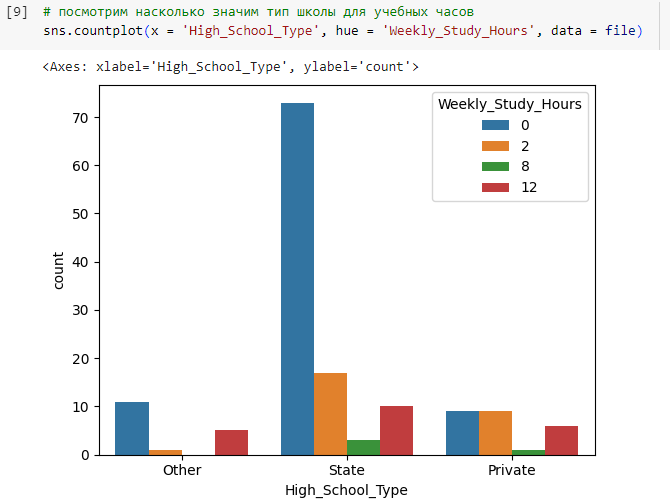
Каждая запись уникально идентифицируется по 'Id'. В набор данных включена демографическая информация, такая как 'Student\_Age' и 'Sex'. 'High\_School\_Type' категоризирует тип посещенной старшей школы, а 'Scholarship' указывает, есть ли у студента стипендия. Детали о 'Additional\_Work' и участии в 'Sports\_activity' предоставляют представление о внеклассных обязательствах.

'Transportation' описывает способ коммутирования каждого студента. Академические аспекты учитываются через 'Weekly\_Study\_Hours', 'Attendance' и оценки по 'Reading', 'Notes', 'Listening\_in\_Class' и 'Project\_work'. Итоговый результат этих факторов отражен в столбце 'Grade', предоставляя всесторонний обзор успеваемости студента. Этот набор данных служит ценным ресурсом для изучения многогранных динамик, влияющих на академические результаты.

Проведем визуальный анализ данных.

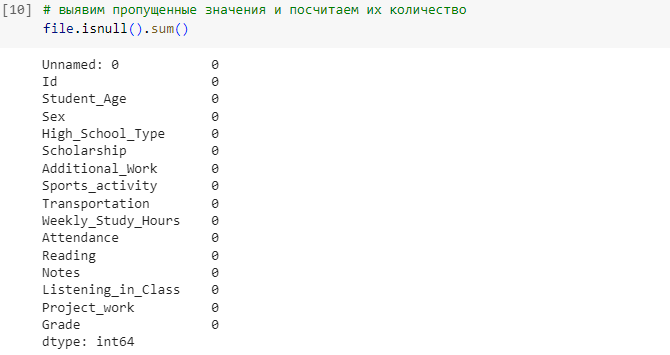


В результате выполнения кода видно, что в обычной старшей школе количество мальчиков превышает количество девочек.

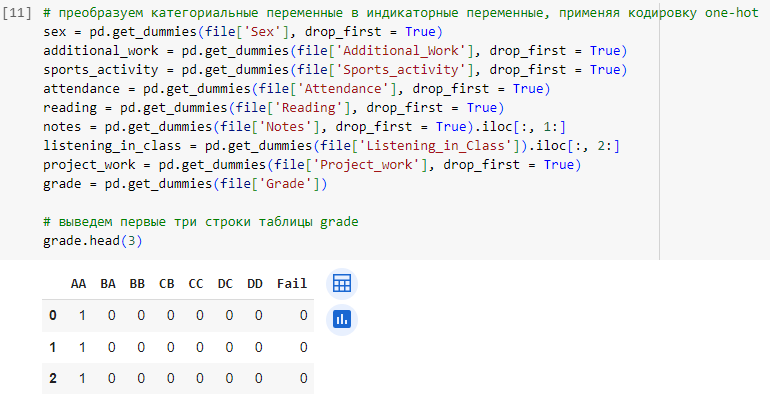


В результате выполнения кода видно, что количество прогулов в обычных старших школах оказалось больше, чем в других типах школ.

2.1.2. Пропущенные значения.



2.1.3. Категориальные переменные.

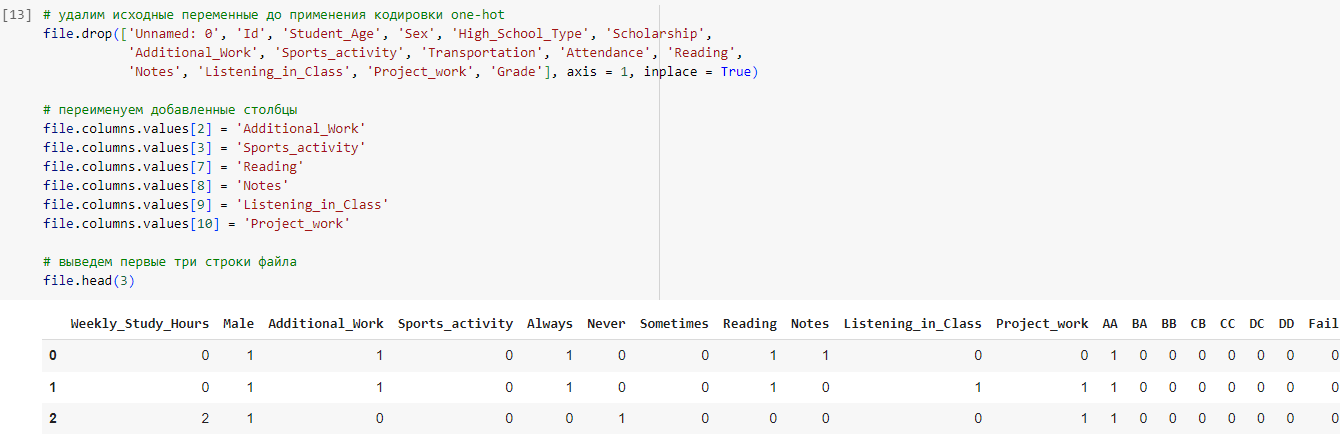


В результате выполнения кода видно, что уникальные категории превращаются в отдельный столбец, и в каждой строке ставится 1, если наблюдение принадлежит данной категории, и 0 в противном случае.



2.1.4. Отбор признаков.

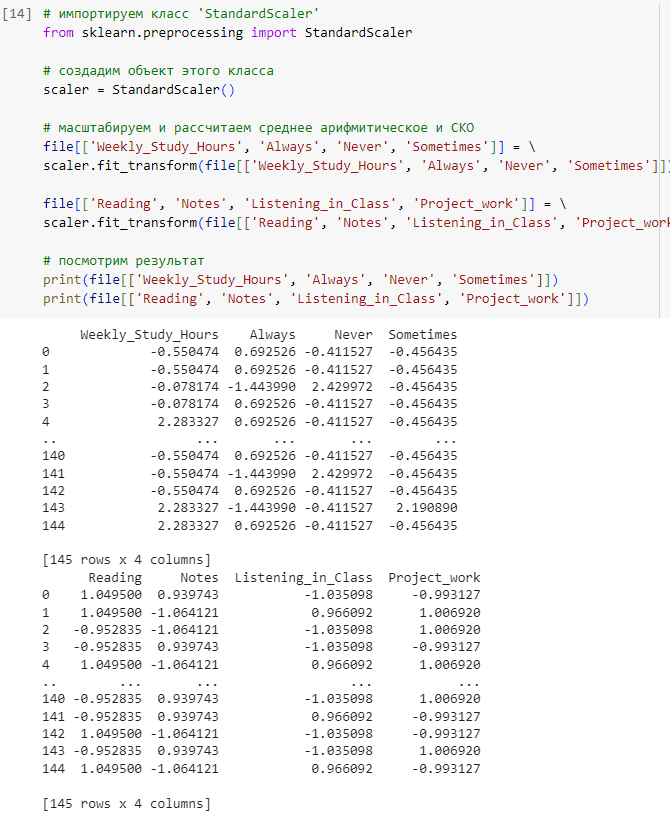
Удалим исходные переменные до применения кодировки one-hot.



В результате выполнения кода видно, что были добавлены таблицы с индикаторными переменными для уникальных значений в переменных. Одновременно были удалены исходные столбцы, оставляя только новые таблицы с данными.

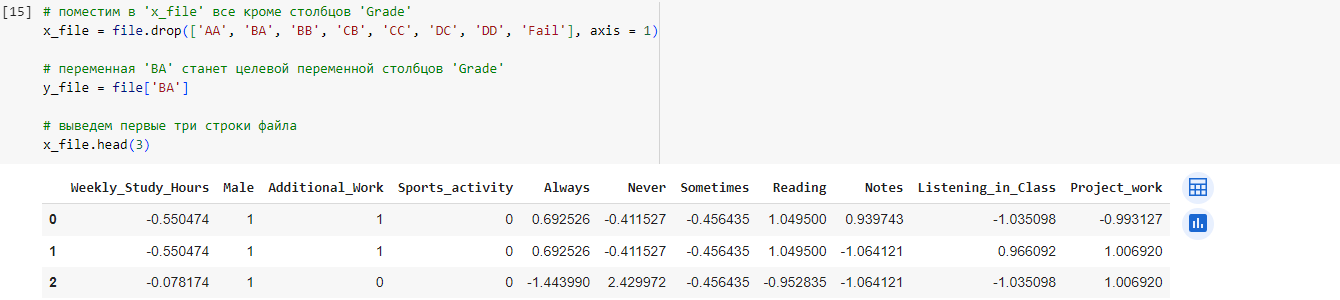
2.1.5. Нормализация данных.

Приведем количественные переменные к одному масштабу.



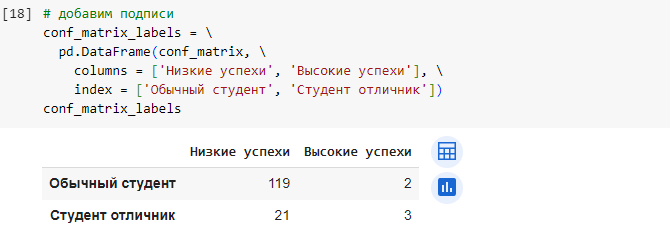
В результате выполнения кода видно, что переменные были масштабированы с использованием StandardScaler. Обратим внимание, что они теперь имеют среднее значение 0 и стандартное отклонение 1, что помогает модели лучше интерпретировать вклад каждой переменной.

2.2. Разделение выборки на признаки и целевую переменную.

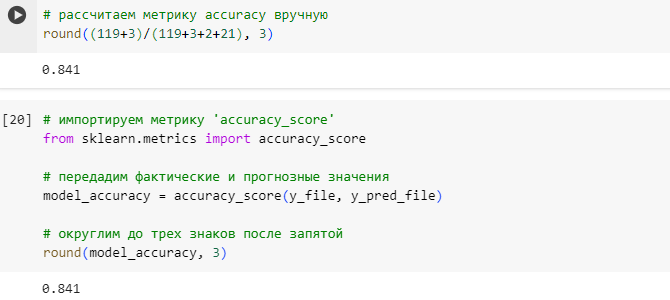


2.3. Обучение модели логической регрессии.



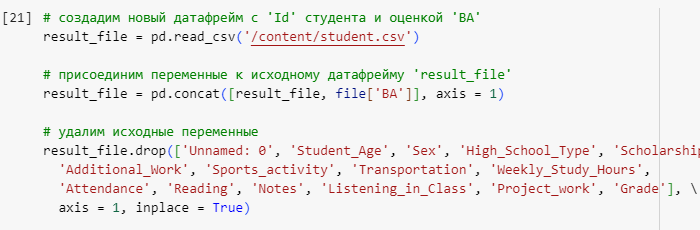


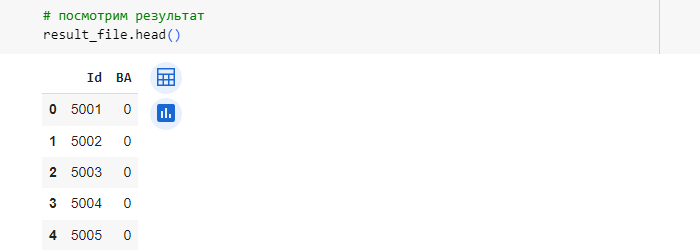
Рассмотрим метрику точности (accuracy). Она отображает долю правильно предсказанных значений.



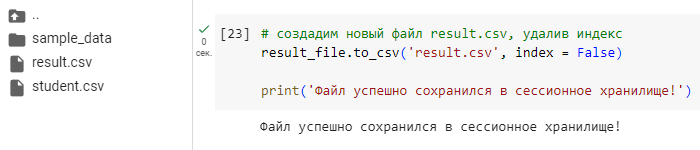
На выборке наша модель показала результат в 84,1%.

3. Сохранение результата в новом файле на сервере.





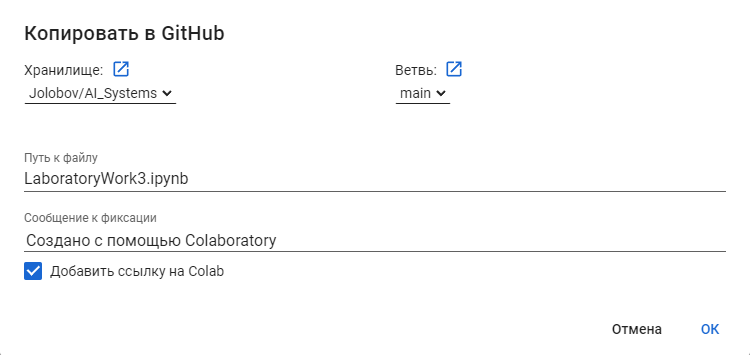
Сохраним новый файл.



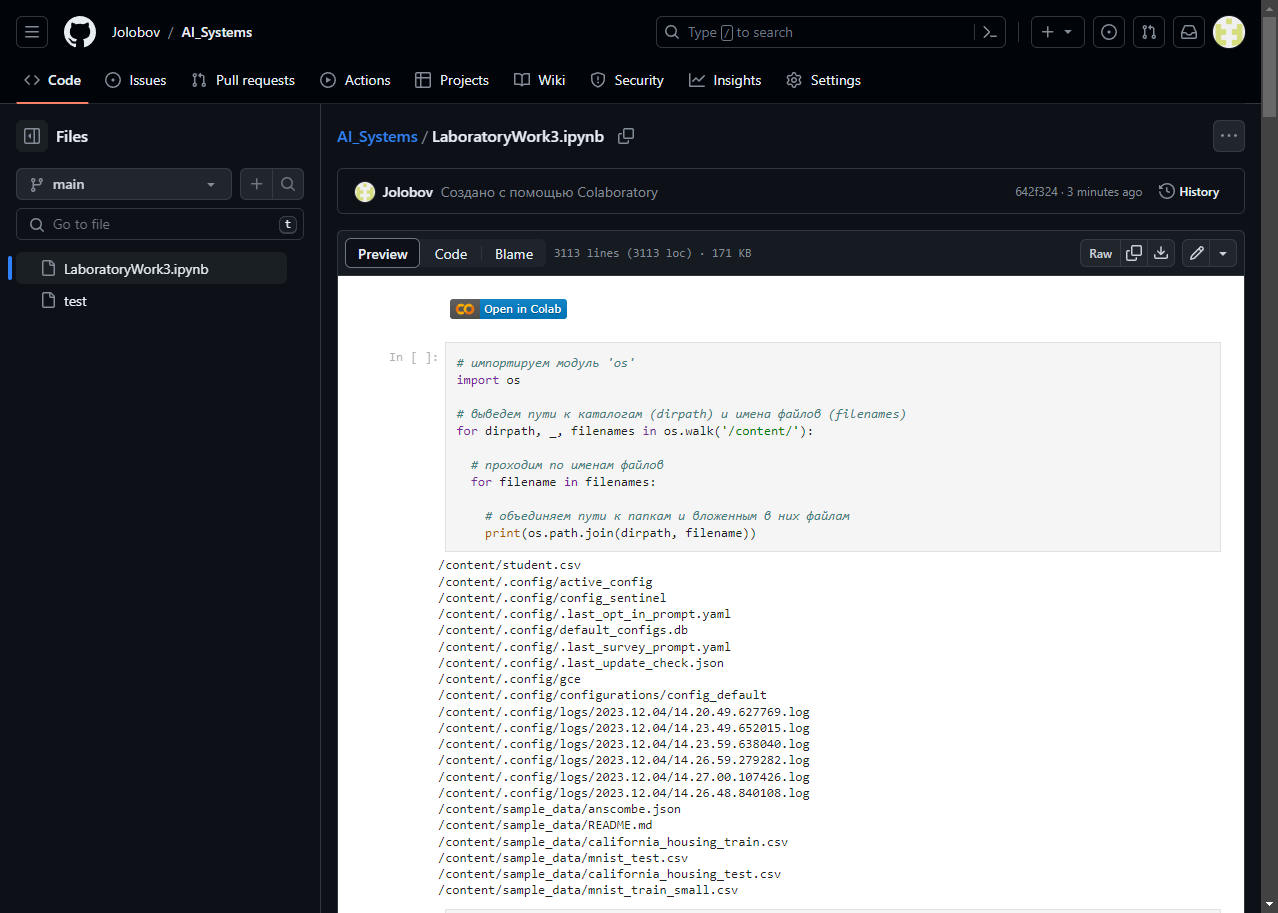
В результате выполнения кода видно, что новый файл ‘result.csv’ появился в ‘Сессионном хранилище’.

4. Связывание проекта с GitHub.

Скопируем проект на GitHub.



Посмотрим на скопированный проект на GitHub’е.



GitHub – [LaboratoryWork3.ipynb](https://github.com/Jolobov/AI_Systems/blob/main/LaboratoryWork3.ipynb)

**Вывод:**

В ходе выполнения данной работы были успешно достигнуты поставленные цели и выполнены поставленные задачи.

В первую очередь, осуществлено изучение инструмента интеллектуальной обработки данных – Google Colab. Это включало в себя ознакомление с основными функциональными возможностями и инструментами предоставляемого сервиса.

Далее был проведен этап загрузки данных из различных источников, таких как Kaggle, Dataset Search и GitHub, что позволило ознакомиться с разнообразием датасетов, представленных в сети.

Основной фокус работы был на изучении инструментов анализа данных. Были проведены следующие действия:

Построение статистики: проанализированы основные характеристики данных, вычислены средние значения, медианы, стандартные отклонения и прочие статистические показатели, что позволило получить более глубокое представление о распределении данных.

Построение графиков: были созданы разнообразные визуализации, такие как гистограммы, диаграммы рассеяния и линейные графики, для визуального анализа данных. Это способствовало выявлению закономерностей, трендов и аномалий в данных.

В завершении работы был осуществлен этап сохранения проекта. Это важный этап, позволяющий сохранить все проделанные настройки, код и результаты анализа для последующего использования или демонстрации.

Таким образом, данная работа с Google Colab позволила успешно освоить инструменты интеллектуальной обработки данных, загрузить данные из различных источников, провести анализ данных с использованием статистических методов и визуализаций, а также эффективно сохранить результаты работы.