Федеральное государственное автономное

образовательное учреждение

Высшего образования

«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

|  |
| --- |
| Институт Космических и Информационных Технологий |
| институт |
| Программная инженерия |
| кафедра |

**ОТЧЕТ О ПРАКТИЧЕСКОЙ РАБОТЕ**

|  |
| --- |
| Версионирование данных и моделей |
| тема |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Преподаватель | |  |  |  | Пересунько Е.О. |
|  | |  | подпись, дата |  | инициалы, фамилия |
| Студент | КИ20-17/1Б, 032050911 |  |  |  | Смыков А.Д. |
|  | номер группы, зачётной книжки |  | подпись, дата |  | инициалы, фамилия |

Красноярск 2024

**СОДЕРЖАНИЕ**

1 Цель.................................................................................................................... 3

2 Задачи................................................................................................................ 3

3 Описание варианта задания............................................................................. 3

4 Ход работы........................................................................................................ 3

5 Вывод................................................................................................................. 7

**1 Цель**

Знакомство с инструментами организации процесса MLOps, получение  
практических навыков по работе с версионированием экспериментов и  
автоматизации пайплайнов машинного обучения.

**2 Задачи**

Для успешного выполнения практической работы необходимо выполнить  
следующие задачи:

- создать удаленный репозиторий для проекта;

- Разработать решение задачи в соответствии с вариантом; ;

- построить автоматический ML-пайплайн, состоящий из всех  
необходимых шагов для решения задачи (предобработка данных, обучение  
моделей, оценка качества и т.д.);

- провести серию экспериментов, используя как минимум три различные  
модели машинного обучения для решения поставленной задачи. При защите  
работы необходимо продемонстрировать возможность перемещения по  
истории экспериментов с целью выбора конкретной модели;

- выполнить аугментацию, сгенерировав дополнительные данные с  
помощью различных преобразований, и сохранить новый датасет;

- повторить пункт 2, используя все данные (исходные и  
аугментированные);

- в отчете привести сравнение результатов, получаемых при различных  
условиях экспериментов на разных наборах данных (с аугментацией и без)..

**3 Описание варианта задания**

Для выполнения данной работы был взят второй вариант, включающий в  
себя датасет “Food vs Non-Food Image”.

**4 Ход работы**

За основу была взята работа по этому датасету с предмета "Теория  
вычислительного обучения".

Был создан гитхаб репозиторий, куда была загружена текущая  
директория, скриншот представлен на рисунке №1.

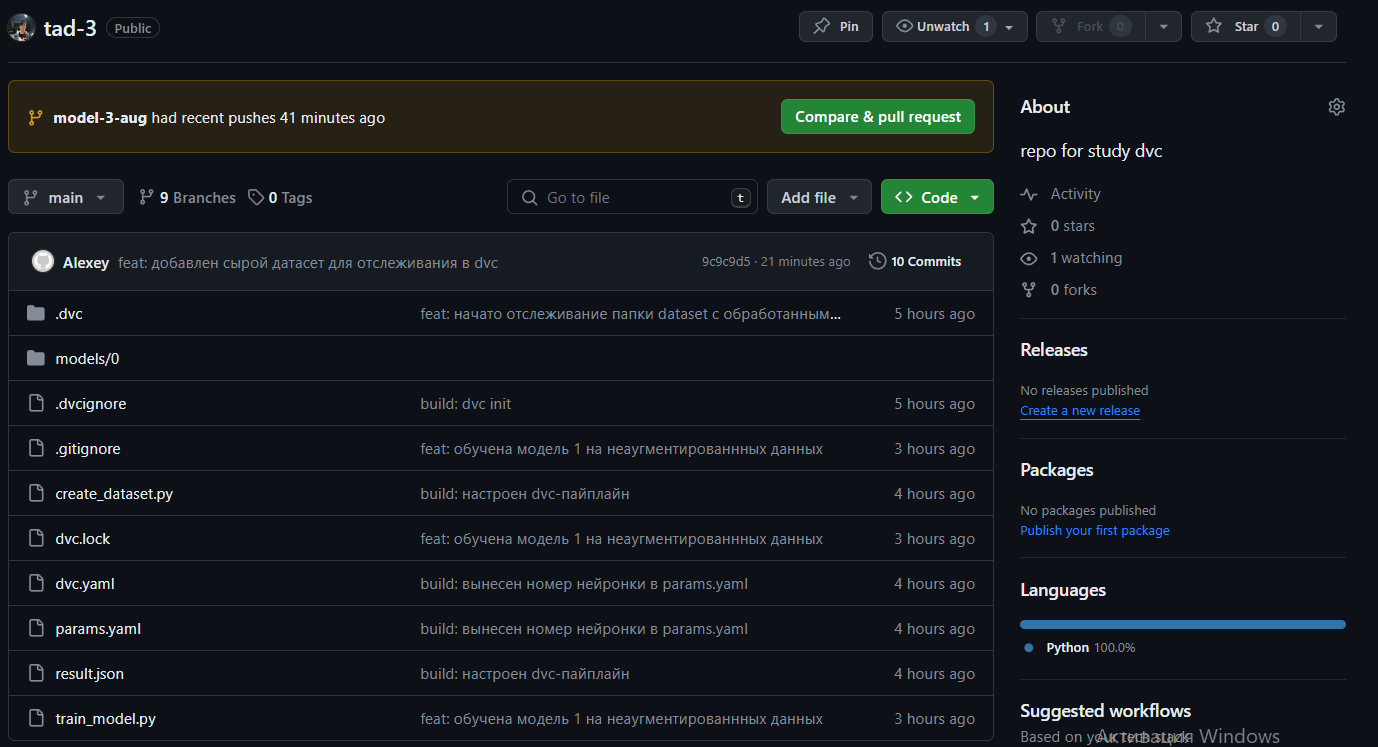


Рисунок 1 – Скриншот гитхаб репозитория

Следующим шагом был настроен dvc-пайплайн с помощью команд "dvc  
init", "dvc add", "dvc stage add", "dvc params", "dvc remote add"

Для параметризации моделей и данных были добавлены параметры для  
пайплайна.

А для удалённого dvc-репозитория была назначена локальная директория.

Список файлов, содержащихся в репозитории после этих команд  
представлен на рисунке №2.

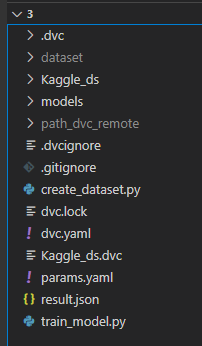


Рисунок 2 – Список файлов в репозитории

Для обучения параметризированных моделей были созданы отдельные  
ветки для каждой модели, их список представлен на рисунке №3.

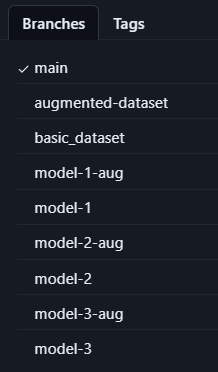


Рисунок 3 – Список веток репозитория

Обучение моделей происходило через команду "dvc exp run", которая  
запускала обработку данных только если они отличались (первый запуск  
данные обрабатывались, а последующие нет, так как обработанные данные уже  
существовали).

В результате, результат обучения моделей хранится в собственной ветке.  
Общий результат обучения моделей представлен на рисунке №4.

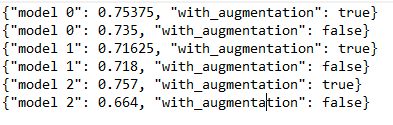


Рисунок 4 – Результат обучения моделей

**5 Вывод**

В результате работы, был создан удалённый репозиторий, содержащий  
ветки для каждой модели, был настроен пайплайн, для автоматизации обучения  
моделей и обработки данных, отслеживаются исходные данные и изучена  
работа с dvc.