Федеральное государственное автономное

образовательное учреждение

Высшего образования

«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

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

**ОТЧЕТ О ПРАКТИЧЕСКОЙ РАБОТЕ №3**

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Преподаватель | |  |  |  | Е.О. Пересунько |
|  | |  | подпись, дата |  | инициалы, фамилия |
| Студент | КИ20-17/1Б, 032052203 |  |  |  | Н.С. Черных |
|  | номер группы, зачётной книжки |  | подпись, дата |  | инициалы, фамилия |

Красноярск 2023

**СОДЕРЖАНИЕ**

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

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

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

3.1 Выбор варианта…........................................................................................ 3

3.2 Создание веток моделей.............................................................................. 3

3.3 Архитектура пайплайна............................................................................... 4

3.4 Итоговое сравнение....................................................................................... 5

4 Заключение........................................................................................................ 5

**1 Цель работы**

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:

1. Создать удаленный репозиторий для проекта;
2. Разработать решение задачи в соответствии с вариантом;
   1. Построить автоматический ML-пайплайн, состоящий из всех необходимых шагов для решения задачи (предобработка данных, обучение моделей, оценка качества и т.д.);
   2. 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;
3. Выполнить аугментацию, сгенерировав дополнительные данные с помощью различных преобразований, и сохранить новый датасет;
4. Повторить пункт 2, используя все данные (исходные и аугментированные);
5. В отчете привести сравнение результатов, получаемых при различных условиях экспериментов на разных наборах данных (с аугментацией и без).

**3 Ход работы**

**3.1 Выбор варианта**

Для выполнения данной работы был взят второй вариант, включающий в себя датасет “Food vs Non-Food Image”. Данные в нём разделены на тренировочные, валидационные и тестовые выборки по 3000, 1000 и 1000 изображений соответственно.

**3.2 Создание веток моделей**

Чтобы организовать версионирование моделей машинного обучения были использованы GitLab, Git и DVC. Иерархия веток в проекте, следующая:

1. Main. Главная ветка, где хранится основная версия продукта;
2. Dvc-test. Ветка для тестирования возможностей dvc;
3. Dvc-stupid-model. Ветка для обучения и сохранения плохой модели;
4. Dvc-middle-model. Ветка для обучения и сохранения средней модели;
5. Dvc-smart-model. Ветка для обучения и сохранения умной модели;
6. Augment. Ветка для аугментации данных;
7. Aug\_first\_model. Ветка для обучения и сохранения плохой модели при новом наборе данных;
8. Aug\_second\_model. Ветка для обучения и сохранения средней модели при новом наборе данных;
9. Aug\_third\_model. Ветка для обучения и сохранения умной модели при новом наборе данных.

**3.3 Архитектура пайплайна**

На рисунке 1 представлена структура разрабатываемого пайплайна, который дублируется в каждой ветке.

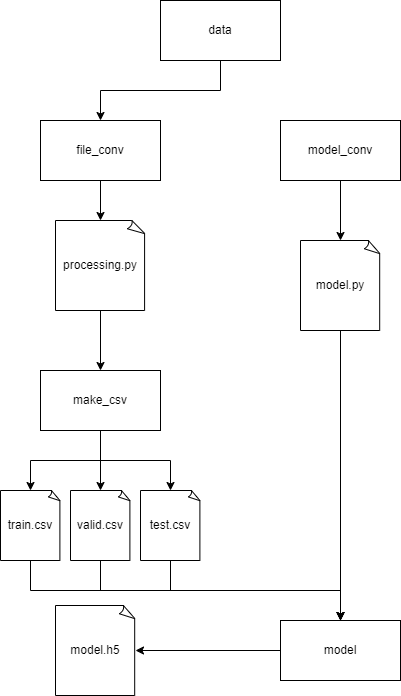


Рисунок 1 – Структура пайплайна

**3.4 Итоговое сравнение**

Решение данной задачи показано на рисунках 2 и 3. Как видно из рисунков, последняя модель лучше всего справляется с данными при обычном наборе данных, ещё до аугментации.

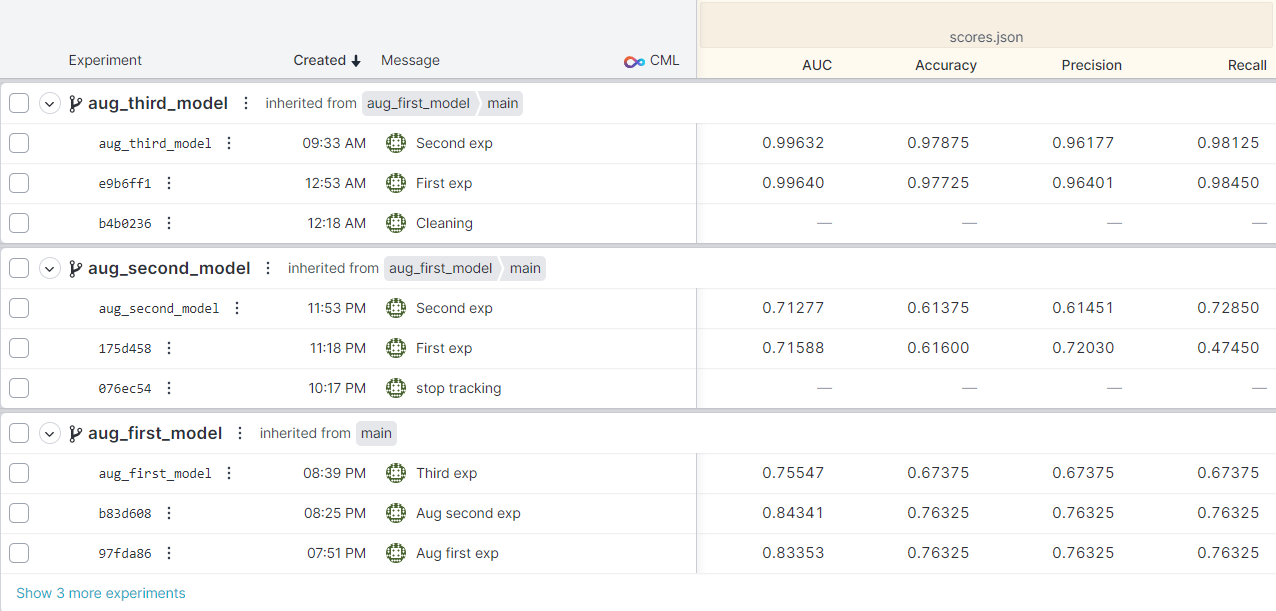
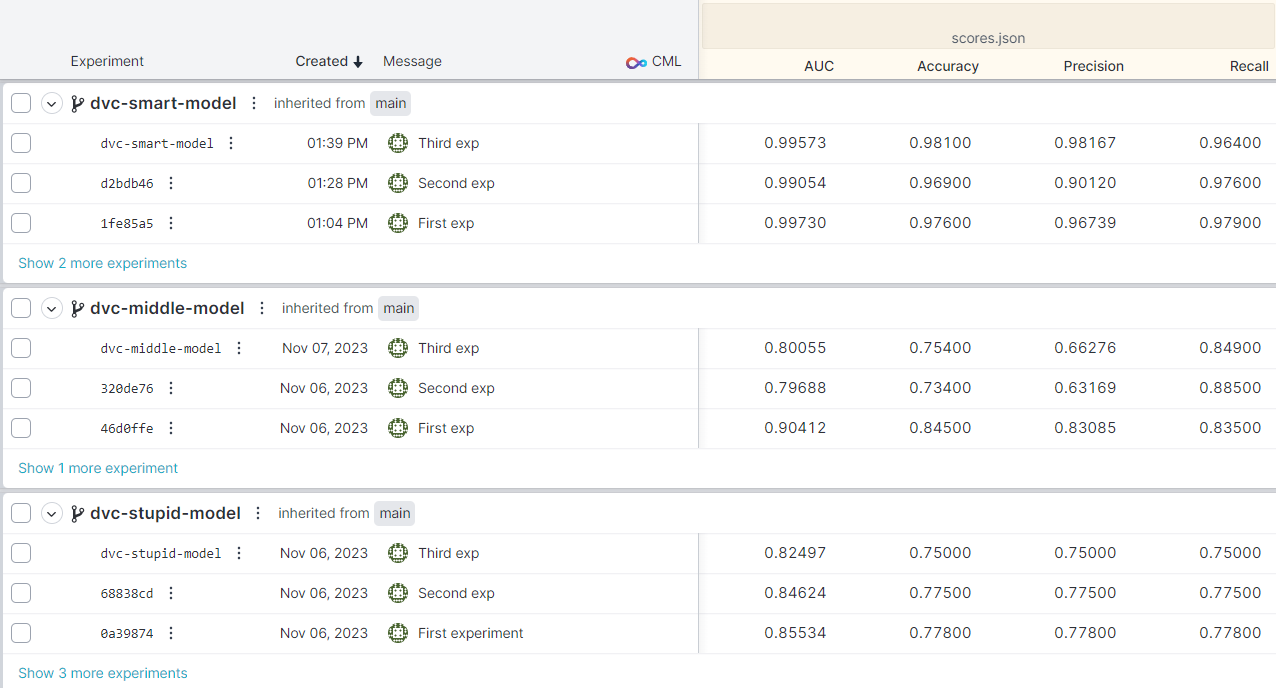
Рисунок 2 – Сводная статистика до аугментации

Рисунок 3 – Сводная статистика после аугментации

**4 Заключение**

В ходе выполнения практической работы были освоены и использованы технологии версионирования данных для создания различных моделей машинного обучения, аугментации данных, контроля версий и проведения серии экспериментов.