Токарев Александр Васильевич

АМО лабораторная работа №4

В практическом задании Вам необходимо продемонстрировать навыки практического использования утилиты dvc для работы с данными.

В постановке задачи используется датасет из соревнования Kaggle “Titanic Disaster” однако Вы можете использовать свои наборы данных, в этом случае в п.п.7-10 необходимо использовать информацию и признаки из вашего датасета.

Этапы задания:

1. Создайте отчет, который снабдите скриншотами последующих действий, в который также поместите ссылку на репозиторий со всеми артефактами, которые возникнут в процессе выполнения задания.

2. Установите git и dvc.

3. Создайте папку проекта.

4. Настройте папку проекта для работы с git и dvc.

5. Настройте git репозиторий.

6. Настройте удаленное хранилище файлов, например на Google Disk.

7. Создайте датасет о пассажирах ‘Титаника”, например, catboost.titanicO-

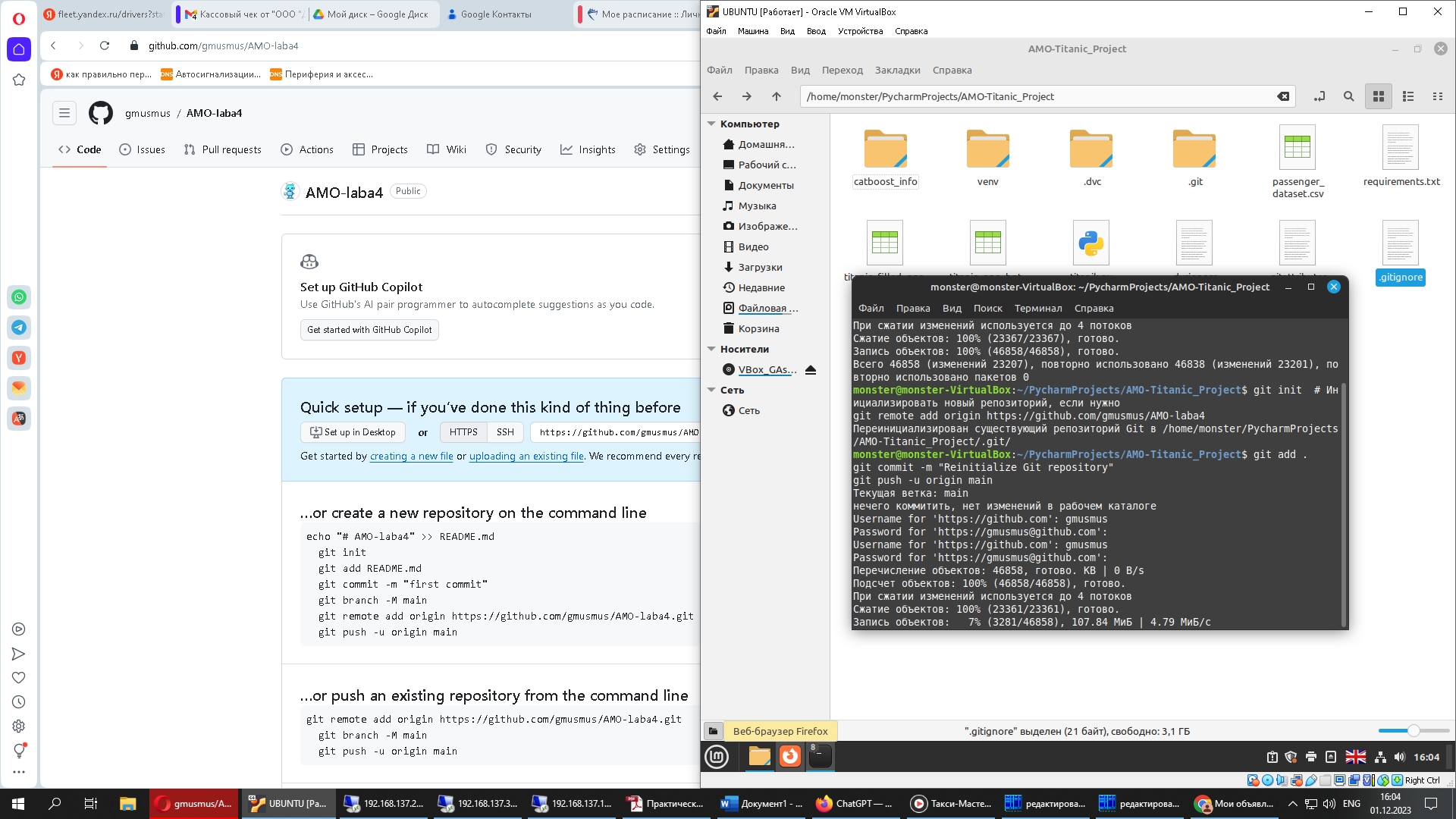
8. Создайте датасет, в котором содержится информация о классе (“Pclass”), поле (“Sex”) и возрасте (“Age”) пассажира. Сделайте коммит в git и push в dvc.

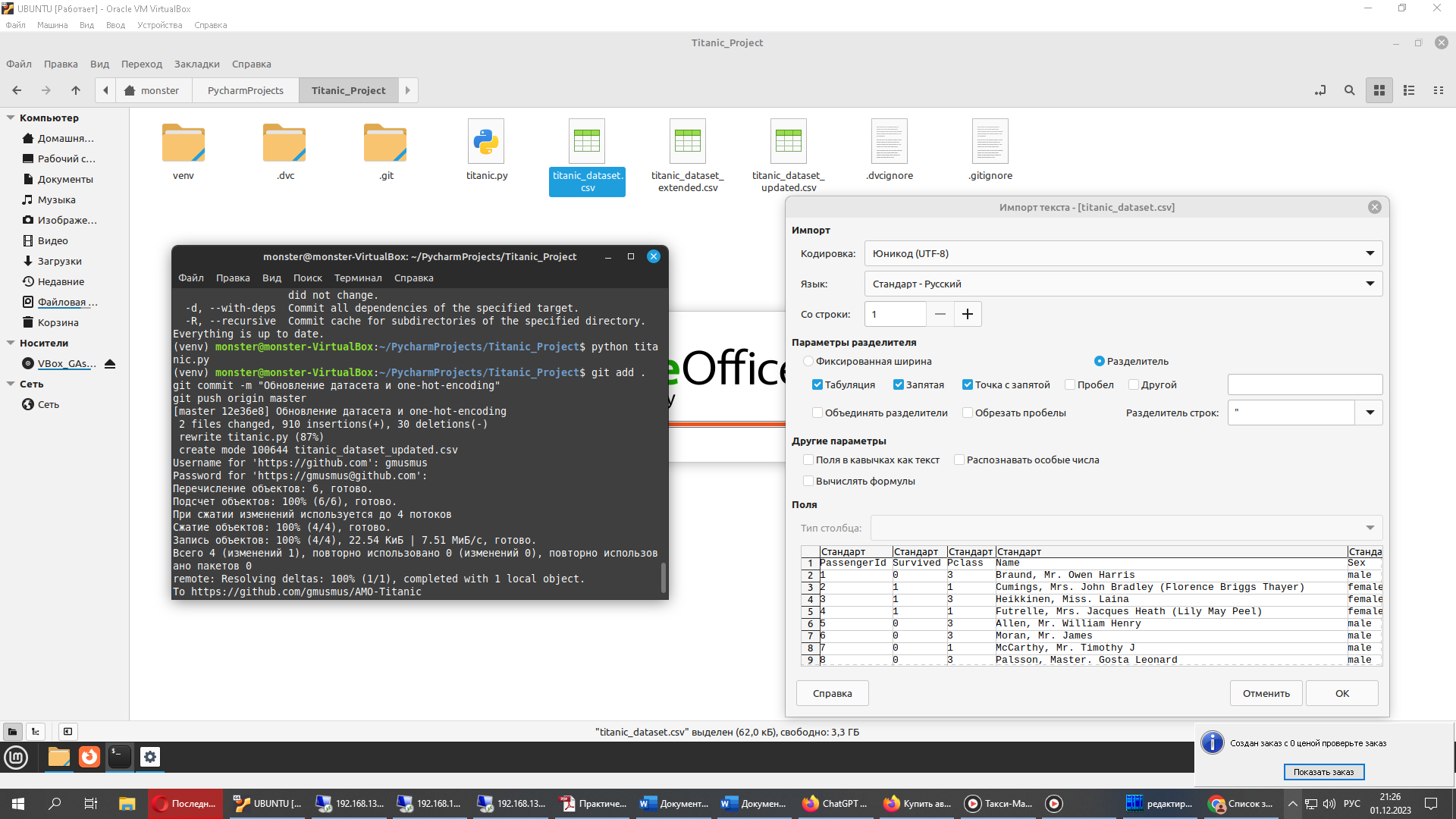
9. Создайте новую версию датасета, в котором пропущенные (пап) значения в поле “Age” будут заполнены средним значением. Сделайте коммит в git и push в dvc.

10. Создайте новый признак с использованием one-hot-encoding для строкового признака “Пол” (“Sex”). Сделайте коммит в git и push в dvc.

11. Выполните переключение между всеми созданными версиями датасета.

При правильном выполнении задания у вас появится git-репозиторий с опубликованной метаинформацией и папка на Google Disk, в которой хранятся различные версии датасетов.





<https://github.com/gmusmus/AMO-Titanic>

