Домашние задания

Описание проекта

Цель: Создать полноценный ML проект с использованием современных инженерных практик. Фокус на инженерных аспектах, а не на сложности ML задач.

Проект: Система классификации/регрессии на простых датасетах с полным MLOps workflow.

Рекомендуемые датасеты

Простые датасеты для быстрого старта:

1. Iris Dataset - классификация цветов ириса
2. Wine Quality - классификация качества вина
3. Boston Housing - регрессия цен на недвижимость
4. Breast Cancer - классификация рака груди
5. Titanic - классификация выживания пассажиров

Средние датасеты для более сложных задач:

6. Heart Disease - классификация болезней сердца
7. Customer Churn - классификация оттока клиентов
8. Car Price - регрессия цен на автомобили

Рекомендация: Выберите один датасет в начале курса и используйте его для всех ДЗ. Это позволит сосредоточиться на инженерных практиках, а не на решении ML задач.

Расписание ДЗ

Дедлайн всех ДЗ: 26 декабря

ДЗ	Тема курса	Баллы	Срок сдачи
ДЗ 1	Рабочее место DS	8 баллов	24 ноября
ДЗ 2	Версионирование данных	10 баллов	8 декабря
ДЗ 3	Трекинг экспериментов	12 баллов	15 декабря
ДЗ 4	Автоматизация	10 баллов	22 декабря
ДЗ 5	ClearML	12 баллов	26 декабря
ДЗ 6	Документация	8 баллов	26 декабря

ДЗ 1: Настройка рабочего места Data Scientist

- **Баллы:** 8 баллов
- **Срок сдачи:** 24 ноября

Описание

Настройте полноценное рабочее место для Data Science с использованием современных инженерных практик.

Требования

1. **Структура проекта (2 балла):**
   - Создать структуру папок с помощью Cookiecutter или Copier
   - Настроить шаблоны для новых проектов
   - Создать README с описанием проекта
2. **Качество кода (2 балла):**
   - Настроить pre-commit hooks
   - Настроить форматирование кода (Black, isort, Ruff)
   - Настроить линтеры (Ruff, MyPy, Bandit)
   - Создать конфигурационные файлы
3. **Управление зависимостями (2 балла):**
   - Настроить Poetry или pipenv для управления зависимостями
   - Создать requirements.txt с точными версиями
   - Настроить виртуальное окружение
   - Создать Dockerfile для контейнеризации
4. **Git workflow (1 балл):**
   - Настроить Git репозиторий
   - Создать .gitignore для ML проекта
   - Настроить ветки для разных этапов работы
5. **Отчет о проделанной работе (1 балл):**
   - Создать отчет в формате Markdown
   - Описать настройку каждого инструмента
   - Добавить скриншоты результатов
   - Сохранить отчет в Git репозитории

Критерии оценки

- **Отлично (8 баллов):** Все требования выполнены, код качественный
- **Хорошо (6-7 баллов):** Основные требования выполнены
- **Удовлетворительно (4-5 баллов):** Большинство требований выполнено
- **Неудовлетворительно (0-3 балла):** Требования не выполнены

⚠ **ВАЖНО:** Менторы будут воспроизводить ваши результаты, поэтому постарайтесь все автоматизировать. Если что-то не совпадет при воспроизведении, можно потерять баллы.

ДЗ 2: Версионирование данных и моделей

- **Баллы:** 10 баллов
- **Срок сдачи:** 8 декабря

Описание

Настройте систему версионирования данных и моделей для ML проекта.

Требования

Выберите ОДИН из инструментов для версионирования данных:

- **DVC** - Data Version Control
- **LakeFS** - Git-like data versioning
- **Git LFS** - Git Large File Storage

Выберите ОДИН из инструментов для версионирования моделей:

- **MLflow** - Model registry
- **DVC** - Model versioning

1. **Настройка выбранного инструмента для данных (4 балла):**
   - Установить и настроить выбранный инструмент
   - Настроить remote storage (S3/Local)
   - Создать систему версионирования данных
   - Настроить автоматическое создание версий
2. **Настройка выбранного инструмента для моделей (3 балла):**
   - Настроить выбранный инструмент для моделей
   - Создать систему версионирования моделей
   - Настроить метаданные для моделей
   - Создать систему сравнения версий
3. **Воспроизводимость (2 балла):**
   - Создать инструкции по воспроизведению
   - Настроить фиксацию версий зависимостей
   - Протестировать воспроизводимость
   - Создать Docker контейнер
4. **Отчет о проделанной работе (1 балл):**
   - Создать отчет в формате Markdown
   - Описать настройку выбранных инструментов
   - Добавить скриншоты результатов
   - Сохранить отчет в Git репозитории

Критерии оценки

- **Отлично (10 баллов):** Полная настройка, качественное версионирование
- **Хорошо (8-9 баллов):** Хорошая настройка, базовое версионирование
- **Удовлетворительно (6-7 баллов):** Базовая настройка
- **Неудовлетворительно (0-5 баллов):** Требования не выполнены

⚠ **ВАЖНО:** Менторы будут воспроизводить ваши результаты, поэтому постарайтесь все автоматизировать. Если что-то не совпадет при воспроизведении, можно потерять баллы.

ДЗ 3: Трекинг экспериментов

- **Баллы:** 12 баллов
- **Срок сдачи:** 8 декабря

Описание

Настройте систему трекинга экспериментов и проведите серию ML экспериментов.

Требования

Выберите ОДИН из инструментов для трекинга экспериментов:

- **MLflow** - Open source ML platform
- **Weights & Biases** - Cloud-based experiment tracking
- **Neptune** - Team collaboration platform
- **TensorBoard** - TensorFlow visualization toolkit
- **CleaML** - MLOPS platform

- **DVC** - Data Version Control

1. **Настройка выбранного инструмента (4 балла):**
   - Установить и настроить выбранный инструмент
   - Настроить базу данных/облачное хранилище
   - Создать проект и эксперименты
   - Настроить аутентификацию и доступ
2. **Проведение экспериментов (4 балла):**
   - Провести 15+ экспериментов с разными алгоритмами
   - Настроить логирование метрик, параметров и артефактов
   - Создать систему сравнения экспериментов
   - Настроить фильтрацию и поиск экспериментов
3. **Интеграция с кодом (2 балла):**
   - Интегрировать выбранный инструмент в Python код
   - Создать декораторы для автоматического логирования
   - Настроить контекстные менеджеры
   - Создать утилиты для работы с экспериментами
4. **Отчет о проделанной работе (2 балла):**
   - Создать отчет в формате Markdown
   - Описать настройку выбранного инструмента
   - Добавить скриншоты результатов
   - Сохранить отчет в Git репозитории

Критерии оценки

- **Отлично (12 баллов):** Полная настройка, качественные эксперименты
- **Хорошо (10-11 баллов):** Хорошая настройка, базовые эксперименты
- **Удовлетворительно (8-9 баллов):** Базовая настройка
- **Неудовлетворительно (0-7 баллов):** Требования не выполнены

⚠ **ВАЖНО:** Менторы будут воспроизводить ваши результаты, поэтому постарайтесь все автоматизировать. Если что-то не совпадет при воспроизведении, можно потерять баллы.

ДЗ 4: Автоматизация ML пайплайнов

- **Баллы:** 10 баллов
- **Срок сдачи:** 22 декабря

Описание

Создайте автоматизированные ML пайплайны с использованием современных инструментов оркестрации.

Требования

Выберите ОДИН из инструментов для оркестрации пайплайнов:

- **Snakemake** - Workflow management system
- **DVC Pipelines** - Data versioning pipelines
- **Apache Airflow** - Workflow orchestration platform
- **Luigi** - Python workflow management

Выберите ОДИН из инструментов для управления конфигурациями:

- **Hydra** - Configuration management framework
- **OmegaConf** - YAML configuration library
- **Pydantic** - Data validation and settings

1. **Настройка выбранного инструмента оркестрации (4 балла):**

    - Установить и настроить выбранный инструмент
    - Создать workflow для ML пайплайна
    - Настроить зависимости между этапами
    - Реализовать кэширование и параллельное выполнение
2. **Настройка выбранного инструмента конфигураций (3 балла):**
    - Настроить выбранный инструмент для управления конфигурациями
    - Создать конфигурации для разных алгоритмов
    - Настроить валидацию конфигураций
    - Создать систему композиции конфигураций
3. **Интеграция и тестирование (2 балла):**
    - Интегрировать выбранные инструменты
    - Создать систему мониторинга выполнения
    - Настроить уведомления о результатах
    - Протестировать воспроизводимость
4. **Отчет о проделанной работе (1 балл):**
    - Создать отчет в формате Markdown
    - Описать настройку выбранных инструментов
    - Добавить скриншоты результатов
    - Сохранить отчет в Git репозитории

Критерии оценки

- **Отлично (10 баллов):** Полная автоматизация, надежные пайплайны
- **Хорошо (8-9 баллов):** Хорошая автоматизация, базовые пайплайны
- **Удовлетворительно (6-7 баллов):** Базовая автоматизация
- **Неудовлетворительно (0-5 баллов):** Требования не выполнены

⚠ **ВАЖНО:** Менторы будут воспроизводить ваши результаты, поэтому постарайтесь все автоматизировать. Если что-то не совпадет при воспроизведении, можно потерять баллы.

ДЗ 5: ClearML для MLOps

- **Баллы:** 12 баллов
- **Срок сдачи:** 26 декабря

Описание

Настройте ClearML для комплексного MLOps workflow и управления экспериментами.

Требования

1. **Настройка ClearML (3 балла):**
    - Установить и настроить ClearML Server
    - Настроить базу данных и хранилище
    - Создать проект и эксперименты
    - Настроить аутентификацию
2. **Трекинг экспериментов (3 балла):**
    - Настроить автоматическое логирование
    - Создать систему сравнения экспериментов
    - Настроить логирование метрик и параметров
    - Создать дашборды для анализа
3. **Управление моделями (3 балла):**
    - Настроить регистрацию и версионирование моделей
    - Создать систему метаданных для моделей
    - Настроить автоматическое создание версий
    - Создать систему сравнения моделей

4. **Пайплайны (2 балла):**
   - Создать ClearML пайплайны для ML workflow
   - Настроить автоматический запуск пайплайнов
   - Создать систему мониторинга выполнения
   - Настроить уведомления
5. **Отчет о проделанной работе (1 балл):**
   - Создать отчет в формате Markdown
   - Описать настройку каждого инструмента
   - Добавить скриншоты результатов
   - Сохранить отчет в Git репозитории

Критерии оценки

- **Отлично (12 баллов):** Полная настройка ClearML, качественный MLOps
- **Хорошо (10-11 баллов):** Хорошая настройка, базовый MLOps
- **Удовлетворительно (8-9 баллов):** Базовая настройка
- **Неудовлетворительно (0-7 баллов):** Требования не выполнены

⚠ **ВАЖНО:** Менторы будут воспроизводить ваши результаты, поэтому постарайтесь все автоматизировать. Если что-то не совпадет при воспроизведении, можно потерять баллы.

ДЗ 6: Документация и отчеты

- **Баллы:** 8 баллов
- **Срок сдачи:** 26 декабря

Описание

Создайте полную документацию проекта и систему генерации отчетов.

Требования

1. **Техническая документация (2 балла):**
   - Создать документацию с помощью Sphinx или MkDocs
   - Создать руководство по развертыванию
   - Настроить автоматическую генерацию документации
   - Создать примеры использования
2. **Публикация в Git Pages (3 балла):**
   - Настроить GitHub Actions для автоматической публикации
   - Создать сайт с документацией на Git Pages
   - Настроить автоматическое обновление при изменениях
3. **Отчеты об экспериментах (2 балла):**
   - Создать отчеты об экспериментах в формате Markdown
   - Добавить графики и визуализации результатов
   - Создать сравнительные таблицы экспериментов
   - Настроить автоматическую генерацию отчетов
4. **Воспроизводимость (1 балл):**
   - Создать инструкции по воспроизведению
   - Создать README с полным описанием
   - Настроить автоматическую сборку документации

Критерии оценки

- **Отлично (8 баллов):** Полная документация, качественные отчеты
- **Хорошо (6-7 баллов):** Хорошая документация, базовые отчеты
- **Удовлетворительно (4-5 баллов):** Базовая документация
- **Неудовлетворительно (0-3 балла):** Требования не выполнены

⚠ **ВАЖНО:** Менторы будут воспроизводить ваши результаты, поэтому постарайтесь все автоматизировать. Если что-то не совпадет при воспроизведении, можно потерять баллы.

Общие требования к проекту

Обязательные компоненты

- **Git репозиторий** с правильной структурой и историей коммитов
- **Выбранный инструмент версионирования данных** (DVC/LakeFS/Git LFS)
- **Выбранный инструмент трекинга экспериментов** (MLflow/W&B/Neptune/TensorBoard/ClearML)
- **Выбранный инструмент оркестрации** (Snakemake/DVC Pipelines/Airflow/Luigi)
- **Выбранный инструмент конфигураций** (Hydra/OmegaConf/Pydantic)
- **Docker** для контейнеризации и воспроизводимости
- **Документация проекта** с автоматической публикацией в Git Pages
- **Отчеты об экспериментах** в формате Markdown

Критерии качества

- **Код** должен быть читаемым, хорошо документированным и следовать стандартам
- **Воспроизводимость** - все компоненты должны быть воспроизводимы на разных машинах
- **Автоматизация** - максимальная автоматизация процессов сборки
- **Документация** должна быть полной, актуальной и доступной онлайн
- **Отчетность** - четкие отчеты о проделанной работе с визуализациями