Отчет о проделанной работе

Студенты: Серков Александр, Скрипко Николай

Тип: соревнование на Kaggle

https://www.kaggle.com/competitions/geoplant-at-paiss/overview

Задача: предсказать, какие растения можно найти в конкретной точке в Европе.
На входе: временные ряды (климат в этой местности за последние 20 лет), картинки (спутниковые фотографии местности), табулярные данные (широта, долгота, страна, регион) и много другого.

Важно: данные делятся на две категории: PO + PA.

- PO (presence only) это единичные наблюдения – где-то что-то есть
- PA (presence-absence) это полноценные исследования территорий (список ВСЕГО, что растет где-то)

На выходе: список ID растений (всего примерно 11255 классов).

Метрика: f1 macro

Что было проделано:

- Разобраны и воспроизведены открытые baseline решения
- Написано обучение и инференс PyTorch моделей с использованием предоставленных данных. Обучен каскад моделей на разных архитектурах (все архитектуры можно найти в src.models
- Кластеризация PO данных для использования вместе с PA данными
- Обработка данных PA: encoding текстовых фичей
- Обучена модель для adaptive k (предсказывает, СКОЛЬКО растений нужно предсказать, дальше берется top k)

Что будет сделано

- Объединение каскада моделей в одну e2e модель
- Использование PO данных в обучении (сейчас использовались только PA, PO были кластеризованы лишь подготовлены – кластеризованы для использования)