Контрольное задание

Компетенция «Машинное обучение и большие данные»

**Общее описание**

В эти три для мы с вами познакомимся с задачами машинного обучения, с которыми могут столкнуться специалисты в реальной жизни: от загрузки и предобработки данных и до вопросов комментирования кода.

**Модуль 4. Геоданнные.**

В данном модуле участники должны загрузить, обработать и отобразить геоданные, построить несколько моделей машинного обучения, а также расширить исходный датасет.

Входные данные:

1. В архиве находятся следующие файлы: **train.parquet**. Данный файл содержит наши геоданные для кукурузного поля где-то в аргентине

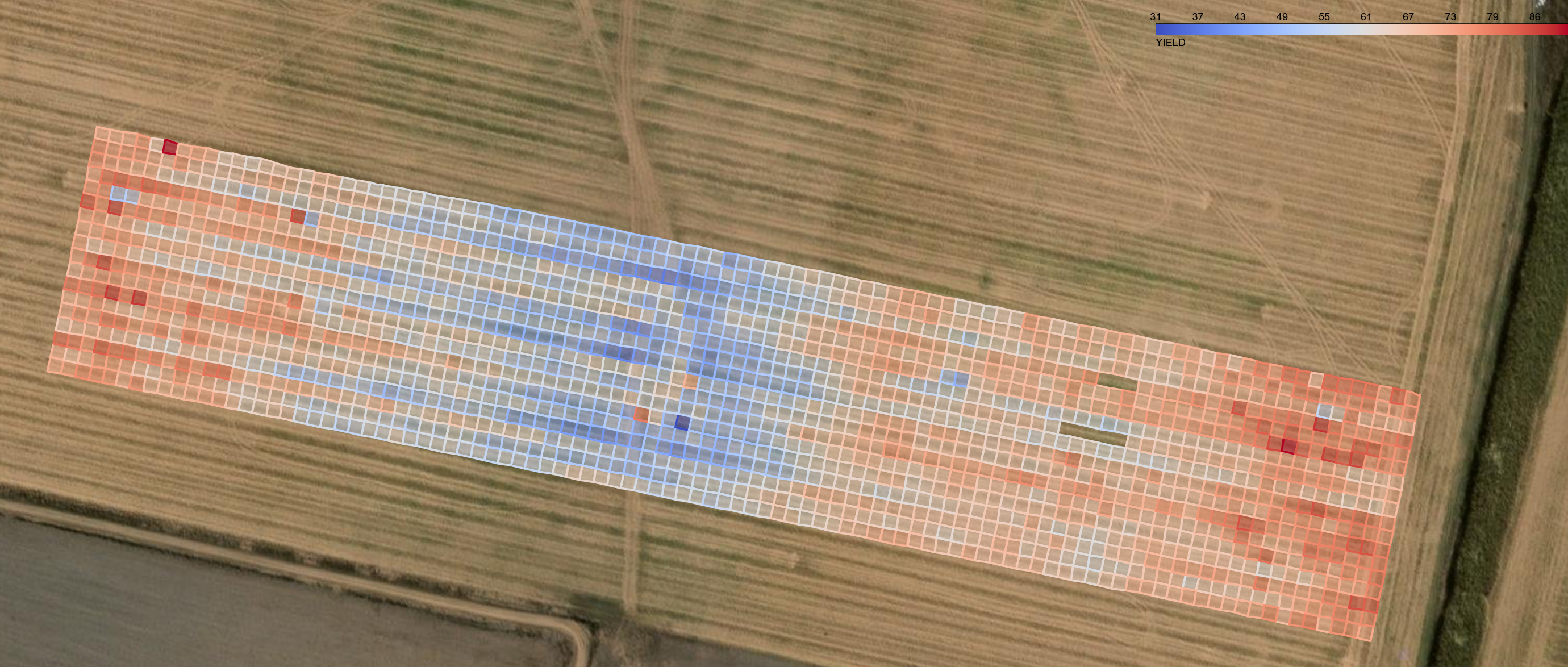
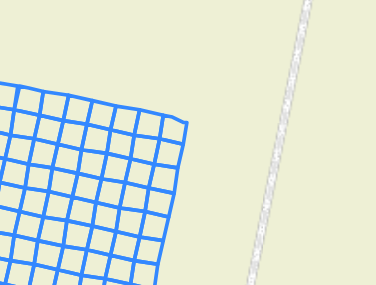


Таблица 1. Описание столбцов в датасете

|  |  |
| --- | --- |
| **Переменная** | **Описание** |
| ID | Identifier |
| TOP2 | Topography dummy, Slope E |
| TOP3 | Topography dummy, Hilltop |
| TOP4 | Topography dummy, Slope W |
| NXTOPz | Interaction Nitrogen and topography zone (z: from 2 to 4) |
| LONGITUDE | Longitude |
| LATITUDE | Latitude |
| OBS | Observation number |
| YIELD | Corn yield (quintals per hectare) |
| N | Nitrogen fertilizer application (kg per hectare) |
| N2 | Nitrogen squared |
| TOPO | Zone: Low E (1), Slope E (2), Hilltop (3), Slope W (4) |
| BV | Brightness value (proxy for low organic matter content) |
| BV2 | Brightness squared |
| NXBV | Interaction Nitrogen - brightness |
| BVXT2z | Interaction brightness and topographic zone (z: from 2 to 4) |
| BV2XTz | Interaction squared brightness and topographic zone (z: from 2 to 4) |
| SAT (99) | Red to NIR ratio (proxy for low organic matter content) |
| SAT2 (99) | SAT squared |
| NXSAT (99) | Interaction Nitrogen - Red to NIR ratio |
| SATXTz (99) | Interaction Sat - topographic zone (z: from 2 to 4) |
| SAT2XTz (99) | Interaction Sat squared - topographic zone (z: from 2 to 4) |

Подробное описание задания:

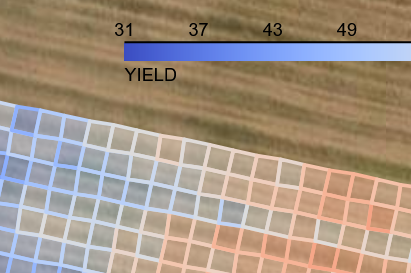
1. Загрузите датасет из файла train.parquet.
2. Отрисуйте границы участков на карте с подложкой в виде карт OpenStreetMap.



1. Отрисуйте границы участков на карте с подложкой в виде любых спутниковых карт.



1. Отрисуйте уровень урожая кукурузы (YIELD) на карте с подложкой в виде любых спутниковых карт.



1. Отрисуйте на карте границы всех районов. Районам, где не вносили азотные удобрения (N = 0), границы отметьте зеленым цветом.



1. Отрисуйте на карте только те районы, координаты которых содержат пересечения с линией latitude = -33.051.



1. Выполните корреляционный анализ для валового урожая (YIELD) и координат участка. Сделайте выводы о том, равнозначны ли условия на участках без учета удобрений.
2. Выполните кластерный анализ участков. Определите оптимальное количество кластеров. Используя техники снижения размерности, отобразите все данные на одном графике. В качестве цвета для раскраски точек используйте номера полученных кластеров.

Постарайтесь описать особенности каждого кластера с точки зрения их характеристик в исходном датасете.

1. Создайте новые фичи. Для каждого участка найдите три ближайших к нему (по расстоянию между центрами районов) и в качестве нового столбца добавьте средний урожай этих трех соседних участков.
2. Создайте новые фичи. Рассчитайте коэффициент эффективности внесенных уобрений

Coef = (YIELD / (N+1)) – min(YIELD).

1. Разделите ваш датасет на обучающий набор и таргет (YIELD).
2. Разделите выборку на обучающую и тестовую части.
3. Обучите модель машинного обучения для предсказания урожая на участке.
4. Выполните оценку точности по метрикам r2 и RMSE.
5. Предложите способы по дополнительному сбору данных для повышения качества модели.
6. Дополните датасет за счет скачивания спутниковых снимков для данного региона в форматах NDVI, Moisture index, SWIR. Сопоставьте районы на изображении с полигонами районов в датасете и получите аггрегированные показатели из обрезанных изображений, после чего добавьте их в датасет.
7. Обучите модель на расширенных данных. Выполните сравнение точности с базовой моделью.
8. Подготовьте отчет по модулю (отчет должен быть в формате HTML, отчет должен быть структурирован, описывать методы, разработанные для REST API, содержать комментарии и пояснения к каждому методу).

Выходные данные по модулю:

1. Отчет в формате HTML.
2. Файл .ipynb.

Отчет должен иметь строго определенные название:

M4\_{lastname}\_{first name}

(например, M4\_Ivanov\_Ivan)