Geo2025 – Hackathon Assignment
==========

Theme: Flood-Proof Rotterdam – Subsurface Lithology Prediction

Overview

Rotterdam is one of Europe's lowest and most flood-prone cities. With rising groundwater, increasing rainfall, and climate change, the municipality is designing new underground flood-storage basins ("waterbuffers").
But these basins can only be built where the **subsurface material is suitable**.

Different soil types behave very differently:

- **zand_grof (coarse sand)** – excellent drainage
- **zand_matig_grof (medium sand)** – good drainage
- **zand_fijn (fine sand)** – acceptable
- **kleiig_zand (sandy clay)** – uncertain
- **klei (clay)** – unsuitable
- **veen (peat)** – unstable
- **grind (gravel)** – rare in Rotterdam
- **schelpen (shell deposits)** – coastal, variable
- **vervallen** – invalid voxel (ignore)

Your mission is to **predict the lithology (soil class)** at **18 public Rotterdam locations** using the datasets provided.

You will:

1. Wrangle large geospatial datasets
2. Do reverse-geocoding with the Nominatim API
3. Filter training voxels using OSM land-use metadata
4. Train ML classification models using coordinate-based features
5. Generate predictions for the test set
6. Submit your CSV + short presentation

Provided Data Sources

You receive a `/data/` folder containing:

1. `geotop_south_holland_student.parquet`
A voxel-based 3D geological model. Columns include:

- `x`, `y` – RD coordinates (meters)
- `z` – depth (meters, negative)
- `lithoklasse` – soil class
- `kans_*` columns – probabilities for different lithologies

2. `mapping_geo_coordinates_to_RD_coordinates.json`
A mapping generated using the official Dutch coordinate transformation (pyproj):

- `latitude`, `longitude` (WGS84)
- `RD_east`, `RD_north`

3. `test_dataset.csv`
The 18 Rotterdam points you must predict.

Columns:

- `ID` – from 0 to 17
- `cityName`
- `displayName`
- `latitude`, `longitude`
- `class`, `type` – OSM metadata
- `RD_east`, `RD_north` – coordinates used for prediction

4. `sample_submission.csv`
Format reference for your final leaderboard submission.

Dutch Geological Glossary

- **lithoklasse** – lithology class (target variable)
- **veen** – peat (unstable)
- **klei** – clay (impermeable)
- **kleiig_zand** – sandy clay (mixed)
- **zand_fijn** – fine sand
- **zand_matig_grof** – medium sand
- **zand_grof** – coarse sand
- **grind** – gravel
- **schelpen** – shell deposits

Your Tasks

Task 1 — Reverse geocode voxel locations
Use:

`NOMINATIM_URL = "https://nominatim.openstreetmap.org/reverse"`

You must:

- Derive latitude/longitude for training voxels using the mapping file
- Call the Nominatim API
- Extract:
  - `class`
  - `type`
- Cache your results to avoid rate-limit issues

Task 2 — Filter training voxels
Remove voxels where:

- OSM `type == "construction"`
- or `type` is missing

This ensures training data is based on stable land-use categories.

Task 3 — Build training dataset
Use only voxels at depth window:

−20 m ± 0.5 m

Features:
- `x`, `y`

Target:
- `lithoklasse`

Task 4 — Train allowed ML models
You may use only models covered in the course:

- Logistic Regression

- Decision Tree
- Random Forest

Evaluate models using **Macro-F1** (SLU10 + SLU11 material).

Task 5 — Predict test_dataset.csv
Use:

- Features: `RD_east`, `RD_north`
- Model: your best model from validation

Create:

```
ID,prediction
0,zand_fijn
1,zand_grof
...
```

Task 6 — Submit:
You must deliver:

1. `prediction.csv`
2. Team presentation (Google Slides template provided)
3. Short description of wrangling pipeline

Evaluation
---

1. Leaderboard Score — 30%
Macro-F1 evaluated on hidden ground truth.

2. Presentation Score — 70%

Schedule

Time	Activity
08:00	Arrival & Setup
08:30	Hackathon Prompt & Teams
09:00	Start hacking
14:00	Goal: first submission
15:00	Goal: improved submission
16:00	Work on presentation
17:00	Submissions close
17:30	Team presentations
18:30	Instructor Baseline
19:00	Winners & Closing

Hackathon Rules

- Teams assigned randomly
- Max. **5 leaderboard submissions**
- Only use provided datasets
- All predictions must come from *your model*
- No manual guessing
- No external datasets
- Respect the Nominatim user policy
- Use your own email in API requests

Presentation Guidelines

Max **5 minutes** per team.

Must include:

1. Problem framing
2. Workflow overview
3. Data wrangling steps
4. OSM filtering logic
5. Model choice & evaluation
6. Final predictions
7. Funny pun

Resources

Resource	Link / Path
Data Folder	`/data/`
Submission Format	`sample_submission.csv`
Presentation Template	**[To be supplied]**

Good luck — and help keep Rotterdam dry! ⛈️🌊