2025

**Mark Heuvelink 1638701**

**Prachi Sanghvi 2161320**

**Dani Berends**  **2110881**

**Taskin Ali** **1662355**

**Abducadir Aligure 2155940**

16-6-2025

AI-Driven Sustainability for Dairy Farming



# Preface

This project has been a valuable and enriching experience, made possible through the guidance, support, and collaboration of several individuals and organizations.

First and foremost, we would like to express our sincere gratitude to **Impact Smart Solutions (ISS)** for initiating this project and entrusting us with the challenge of exploring the feasibility of predicting nitrate levels in groundwater through machine learning. Their vision and commitment to sustainable innovation provided the foundation and motivation for this work.

We also extend our thanks to the **HAN University of Applied Sciences** for facilitating this collaboration and providing the academic environment, resources, and mentorship required to successfully carry out this project. Special appreciation goes to our project supervisors and lecturers, whose feedback, encouragement, and expertise helped us refine our approach and deepen our understanding.

We would like to acknowledge the researchers and data scientists whose publicly available environmental and agricultural studies enabled us to construct and enrich our dataset with realistic and context-specific variables relevant to the Netherlands.

Finally, we are grateful to our fellow students, peers, and everyone who provided insight, critique, or support along the way. Their contributions helped shape this project into a meaningful step toward sustainable agricultural monitoring.

This project represents a collective effort, and we are proud to have contributed to a topic of such environmental and societal relevance.

# Contents

[Preface 1](#_Toc200987669)

[Contents 2](#_Toc200987670)

[Business Understanding 4](#_Toc200987671)

[Project Objectives 4](#_Toc200987672)

[The problems 4](#_Toc200987673)

[Success Criteria 4](#_Toc200987674)

[The information 4](#_Toc200987675)

[The want 5](#_Toc200987676)

[Ethics 6](#_Toc200987677)

[Key Stakeholders 6](#_Toc200987678)

[Stakeholder Interests 6](#_Toc200987679)

[Organizational Responsibilities 7](#_Toc200987680)

[Ethical Considerations in Land Cultivation 7](#_Toc200987681)

[Data Understanding 8](#_Toc200987682)

[Understanding nitrate 8](#_Toc200987683)

[Nitrate contamination in Dutch groundwater 8](#_Toc200987684)

[Agricultural sources of nitrate pollution 8](#_Toc200987685)

[Key influencing factors 8](#_Toc200987686)

[Electrical Conductivity (EC) as a Supporting Indicator 9](#_Toc200987687)

[Regulatory and policy context 10](#_Toc200987688)

[Effectiveness of mitigation measures and best practices 10](#_Toc200987689)

[Regional hotspots and case studies 11](#_Toc200987690)

[Data Collection Process 12](#_Toc200987691)

[Stakeholder Engagement and Data Requests 12](#_Toc200987692)

[Synthetic Dataset Creation 12](#_Toc200987693)

[Structure and Completeness 12](#_Toc200987694)

[Data Preparation & Collection 14](#_Toc200987695)

[The meaning of the data 14](#_Toc200987696)

[Metadata and Location 14](#_Toc200987697)

[Well and Geological Information 14](#_Toc200987698)

[Sample Timing 14](#_Toc200987699)

[Chemical Parameters (Laboratory Measurements) 15](#_Toc200987700)

[Field Measurements (VELD) 15](#_Toc200987701)

[Environmental and Land Use Context 15](#_Toc200987702)

[Target Variable 16](#_Toc200987703)

[Data Integration 16](#_Toc200987704)

[Feature Engineering 17](#_Toc200987705)

[Data Cleaning 18](#_Toc200987706)

[Dropped columns 18](#_Toc200987707)

[1. Metadata Columns 18](#_Toc200987708)

[2. General Descriptive Data 18](#_Toc200987709)

[3. Mineral Concentration Columns 18](#_Toc200987710)

[Data Analysing 20](#_Toc200987711)

[Correlation matrix 20](#_Toc200987712)

[Histogram nitrate concentration 21](#_Toc200987713)

[Decision 21](#_Toc200987714)

[Modelling 22](#_Toc200987715)

[Parameter Tuning 22](#_Toc200987716)

[Training & Validation 22](#_Toc200987717)

[Evaluation Metrics 23](#_Toc200987718)

[Key Advantages Observed 23](#_Toc200987719)

[The interface 24](#_Toc200987720)

[Evaluation 26](#_Toc200987721)

[Deployment 26](#_Toc200987722)

[Conclusion 27](#_Toc200987723)

[Bibliography 28](#_Toc200987724)

# Business Understanding

This project was given by the HAN University of Applied Sciences from the company Impact Smart Solutions, also known as ISS. They had approached the University for a collaboration for a research project on a machine learning model to predict nitrate levels in groundwater.

## Project Objectives

The goal for this project is to create a machine learning model that can predict the nitrate levels in groundwater. The reason why this is needed is because at this moment in time it is very expensive and difficult to measure this. This is specifically needed for the farmlands in the Netherlands. Over the years the sustainability goals have gotten stricter and stricter across Europe. This indirectly means the Netherlands too. So, to govern the sustainability goals it is important to get easy and accurate data.

## The problems

That is where the problem lies, it is possible for farmers and government to get accurate data. But it is not easy and especially not cheap. To get accurate insight on the nitrate levels it costs a lot of money. This is because of the use of sensors. For farmers it is possible to install sensors to analyse the farms and report real time data. This is then used to monitor nitrate levels, but it is not sustainable on a monetary basis.

Besides this the Netherlands must adhere to the Paris agreement (United Nations Framework Convention on Climate Change [UNFCCC], n.d.). This means taking control of the nitrate levels. For the Netherlands this has been a long standing issue (Government of the Netherlands, Ministry of Agriculture, Fisheries, Food Security and Nature, n.d.). The government wants better quality of nature, water and the soil, and to fulfil the Netherlands climate obligations.

So that’s why it's important to make machine learning model that can predict this. Especially for the nitrate levels in ground water.

## Success Criteria

To make this project a success it is important to know what information is available and expected.

### The information

All around the world there is more research being done on nitrate levels in ground water. This research details how much nitrate levels impacts nature and what exactly influences it. This research is done by lots of different parties across the world, the problem lies now with getting access to that information. Because different parties are researching this problem not everyone is willing to give this information away. So that’s why it is especially important to get reliable information that is our scope. For example, the effects of climate on nitrate level, must be specific to the Netherlands.

### The want

This project is a cooperation together with Impact Smart Solutions. Before taking this project on it was important to know what exactly we expected from them and what they expected from us. One thing came up repeatedly.

The machine learning model doesn’t need to be fully accurate. But it is important for ISS to know if it is possible to predict nitrate levels in ground water without the use of expensive sensors.

So, the end goal for this project will be making a machine learning model that will prove or disprove the possibility of a machine learning model predicting nitrate levels in groundwater.

# Ethics

This project addresses the issue of nitrate surplus per hectare in groundwater on dairy farms. We propose the development of a machine learning model capable of analyzing groundwater quality without relying on expensive sensors or laboratory testing. The goal is to offer cost-effective and actionable recommendations to improve the environmental impact of dairy farming, particularly concerning groundwater sustainability.

### Key Stakeholders

* Global Citizens
* Government
* Sensor Manufacturers
* Farmers
* Impact Smart Solutions
* HAN University Students
* Agricultural Labor Providers

## Stakeholder Interests

|  |  |
| --- | --- |
| **Stakeholder** | **Interests** |
| Global Citizens | Combatting climate change, ensuring clean water and food security |
| Government | Formulating clear, evidence-based environmental regulations & data |
| Sensor Manufacturers | Developing affordable, accurate nitrate measurement tools |
| Farmers | Maintaining productivity while complying with environmental regulations |
| Impact Smart Solutions | Delivering innovative, effective sustainability tools |
| HAN Students | Developing a successful, real-world applicable model; meeting academic goals |
| Agricultural Labor | Working under safe, compliant, and sustainable farm conditions |

## Organizational Responsibilities

* **Government:** Support transparency and cooperation in nitrate research. Ensure laws are science-based and understandable.
* **Sensor Manufacturers:** Provide precise, affordable nitrate sensors to expand accessibility.
* **Farmers:** Be willing to adopt new technologies and sustainable practices.
* **Impact Smart Solutions & HAN Students:** Collaborate with all stakeholders to create a functional, ethical, and scalable machine learning model.

## Ethical Considerations in Land Cultivation

It is essential that farming activities maintain nitrate concentrations within legally and environmentally safe limits. This includes:

* Applying fertilizers responsibly
* Regular monitoring using affordable tools
* Using predictive models to inform decisions
* Being transparent with data to enhance collaboration and compliance

# Data Understanding

## Understanding nitrate

### Nitrate contamination in Dutch groundwater

Dutch farmland covers much of the country, and intensive nutrient use in agriculture has left many groundwater bodies with elevated nitrate. In fact, RIVM and WUR report that “in many places… groundwater… still contains too much nitrogen,” often exceeding the EU threshold of 50 g NO₃/m³. Nationwide monitoring (e.g. the RIVM/WUR Nitrate Reports and LMM/LML networks) shows that about 14% of wells (2016–2019) had mean nitrate above 50 g/m³ (European Commission, Directorate‑General for Environment, 2022). This pollution pressures aquatic ecosystems and drinking-water sources: roughly 20% of public groundwater intakes are already approaching problematic nitrate levels. Contamination is worst under sandy loess soils in the south and east (e.g. Limburg and North Brabant) and much lower under clay/peat in the west (European Commission, Directorate‑General for Environment, 2022), reflecting natural denitrification differences and land use.

### Agricultural sources of nitrate pollution

Dutch agriculture is the dominant nitrate source. Nitrogen (N) from livestock manure and synthetic fertiliser is applied to crops, and any surplus not taken up becomes nitrate that percolates to groundwater. The Netherlands produces more manure-N than its soils can assimilate about half of the manure produced must be moved off-farm (exported or processed). Even under tight regulations, intensive farms spread very high N rates. For example, in 2022 grassland farms with a derogation applied on average ~228 kg N/ha from manure, well above the EU “good” limit. Synthetic fertilisers further add to this surplus. Crop type also matters: specialized vegetable and arable farms (e.g. intensive vegetable fields or maize for silage) generally use the most fertiliser and leave the most N in residues, so they show the highest groundwater nitrate concentrations. By contrast, permanent grassland (especially with deep-rooted forage) typically retains more N and leaches less nitrate.

### Key influencing factors

Multiple factors modulate how much applied N reaches groundwater:

**- Weather/Climate**: Rainfall and drought strongly affect leaching. Heavy rainfall and wet years tend to flush nitrates down and promote denitrification, whereas multi-year drought (like 2018–2020) left more N in the soil (crops grew less, and there was less water to dilute or transport nitrate). RIVM notes that increased precipitation after the dry 2018–2020 period helped break down stored nitrates, reducing groundwater concentrations by 2023.

**- Soil and hydrogeology:** Denitrification (microbial N removal) is very soil-dependent. Peat and clay soils (common in western NL) have high denitrification capacity, leading to low nitrate in groundwater. By contrast, well-drained sandy and loess soils (south/east) have poor denitrification and high permeability, so they transport nitrates rapidly to the aquifer.

**- Land use and crops:** Lands left bare or with high-N crops in winter (e.g. autumn-ploughed fields, maize stubble) allow more leaching. Catch crops or permanent cover can uptake leftover N. Grassland renewal or fallowing (turning grassland to arable) can flush a pulse of nitrates if not managed carefully.

**- Farm management:** Practices like timing of manure spreading (winter bans reduce losses), application methods (injection vs. surface), irrigation and drainage (tile drains speed leaching), and buffer zones along ditches all influence nitrate transport. The EU Nitrates Directive’s “good farming practices” explicitly call for measures like balanced fertilisation, catch crops, and 5–7 m unfertilised buffer strips by waterways. In practice, farms vary in how they implement these measures, contributing to regional variability.

### Electrical Conductivity (EC) as a Supporting Indicator

Electrical conductivity (EC) is routinely measured alongside nitrate in Dutch groundwater monitoring (e.g. LMM, RIVM, and WUR studies). EC reflects the total concentration of dissolved ions in water — including nitrate, chloride, potassium, calcium, and sulphate — many of which originate from manure or fertiliser.

Elevated EC levels often co-occur with high nitrate values, particularly in sandy regions with intensive livestock farming, like Noord-Brabant and Limburg. Background EC levels in Dutch groundwater typically range from 200–600 µS/cm, but agriculturally impacted sites frequently exceed 1000 µS/cm.

Monitoring studies (e.g., RIVM 2020-0184; Broers & Van der Grift 2004) show that EC can be a useful early-warning tool. It helps detect diffuse pollution trends, identify over-fertilised areas, and differentiate between natural vs. anthropogenic nitrate sources. In denitrifying environments (like clay/peat soils), EC may remain elevated even when nitrate is reduced, signaling residual pollution potential.

Including EC in long-term monitoring enhances the reliability of nitrate trend assessments and supports more accurate source attribution.

### Regulatory and policy context

Dutch nitrate limits are set by EU law and national rules. The EU Nitrates Directive (1991) requires Member States to designate all farmland that drains into vulnerable waters as a Nitrate Vulnerable Zone. Under this directive, the Netherlands must keep agricultural N inputs below 170 kg/ha (total 4 nitrogen) on average (Statistics Netherlands, 2024). Domestically, the Dutch Manure and Fertilisers Act (Meststoffenwet) implements these rules: it caps fertiliser and manure application by crop and soil type, mandates manure storage, and bans spreading on wet soils or in winter. Since its 2006 reforms, the Act has driven large cuts in N use, which helped bring nearly all monitoring sites below 50 g/m³ on average. For example, after 2006 nitrate in the upper water was falling steadily (though primarily due to past measures and soil N depletion).

Historically the Netherlands negotiated a derogation to exceed the 170 kg/ha limit on high-grass farms. In practice, farms with ≥80% grass were allowed up to ~230–250 kg N/ha of manure per year (European Commission, Directorate‑General for Environment, 2022). This exception covered roughly 45–50% of Dutch agricultural land in 2016–2019. However, the EU decided in 2022 to phase out the derogation: no more extra manure use will be permitted after 2026. To support this, the EU set national manure production ceilings (489.4×10^6 kg N for 2022) and is reducing them (~440×10^6 kg N by 2025), effectively forcing a 6% cut in total manure N by 2025.

Additional Dutch policies include enforcing dietary changes (lowering P in feed to reduce manure P), subsidising manure processing and export, and strict checks on compliance. Some regions also use water boards’ discharges permitting to enforce nutrient limits. Progress is tracked in the four-year Nitrate Reports published by RIVM and partners. As of the latest report, the Netherlands remains off target: EU assessments note that NL still fails the Directive’s water-quality goals.

### Effectiveness of mitigation measures and best practices

Over 30+ years, Dutch farmers and regulators have adopted many mitigation measures. By law and subsidy, farms must follow “good agricultural practice” (GAP) codes: balanced fertilization (matching N to crop needs), catch crops (covering soil in winter), and buffer zones (typically 5 m unfertilised strips) beside all waterways. Manure must be stored and applied at times when crops can use it (no spreading on frozen or waterlogged fields, no manure on fallow in winter). On the farm, precision techniques are increasingly used: e.g. precision spreading, manure injection, nitrification inhibitors (slowing nitrate formation), and acidification of slurry (reducing ammonia losses).

These efforts have yielded benefits. After 2006 the national N surplus fell substantially, and many groundwater sites improved accordingly. For instance, average nitrates in the main sandy regions declined by 2015. However, the marginal gains have diminished. RIVM explicitly reports that “water quality has not structurally improved since 2012”, indicating that existing measures alone are not yet sufficient to meet targets. The 2020–2023 data even showed a brief rebound in nitrate under farms. In practice, this means that additional or stricter measures are now being explored. Dutch pilots include controlled drainage systems (to reduce leaching), planting perennial bioenergy crops, constructing field-scale bioreactors or wetlands to denitrify drainage water, and converting marginal farmland to nature. The new 2023 regulations (e.g. compulsory 5 m buffer strips along all ditches, and permits tied to water quality) reflect this next step, but their impact will only be visible in coming years.

In summary, best practices in the Netherlands involve both regulatory limits and on-farm techniques. Official documents list balanced N use, catch crops, buffer strips and fixed no-spread periods as key measures. These practices are widely recommended, but RIVM’s monitoring shows that, to fully protect groundwater, all farms must implement them effectively and consistently.

### Regional hotspots and case studies

Nitrate pollution is spatially uneven. EU reporting highlights the worst provinces: Limburg (southern loess) and Noord-Brabant (sandy/meadow) top the list. In 2016–2019, about 36% of groundwater stations in Limburg exceeded 50 g/m³ and 26% in Noord-Brabant (European Commission, Directorate‑General for Environment, 2022). Zeeland (mixed sand/clay) also had local exceedances. In contrast, northern provinces with mainly clay soils (Friesland, Groningen) had far fewer high readings. Data from 2020–2021 show that average nitrate in the southern/eastern sandy region reached ~67 g/m³ (double the 2017 level) and ~57 g/m³ in the loess region, whereas northern sandy areas averaged only ~37 g/m³.

These patterns correspond to farming intensity. Limburg’s 25 g/m³ sandy loess (Maas basin) and Brabants’ 50 g/m³ sandy polder/peat areas coincide with heavy dairy and maize production. By contrast, lowland Fens (e.g. West Frisia) have many grass farms but also good drainage control. Individual water boards and researchers confirm such local hotspots: for example, some Gelderland wells under intensive dairy measure ~60 g/m³. On the other hand, many clay-delta (Rijnmond, IJsselmeer polders) routinely show <10 g/m³ in deep groundwater. No single study covers all regions, but RIVM’s Landelijk Meetnet and PBL reports clearly identify the south and east as priority areas.

## Data Collection Process

To develop a machine learning model that predicts nitrate concentrations in Dutch groundwater, we followed a two-phase data strategy combining both synthetic and real-world sources.

### Stakeholder Engagement and Data Requests

We initiated contact with various experts and institutions involved in groundwater quality monitoring and agricultural research, including Wageningen University, RIVM (National Institute for Public Health and the Environment), and researchers such as Arno Hooijboer and Koos Verloop. Through this outreach, we were directed to several relevant datasets:

* The Landelijk Meetnet Grondwaterkwaliteit (LMG) dataset, which includes long-term monitoring of groundwater quality (Naus & Valster, 2025).
* The Nitraatkaart and insights from the Landelijk Meetnet effecten Mestbeleid, although the latter contains confidential data not publicly available.

### Synthetic Dataset Creation

While awaiting access to certain real-world datasets, we created a synthetic dataset based on scientific literature, expert input, and public domain knowledge. This dataset included key variables influencing nitrate leaching, such as:

* Field measurements (e.g., pH, electrical conductivity)
* Chemical inputs (e.g., nitrate, ammonium, fertilizers, manure)
* Land use indicators (e.g., crop type, livestock density, urbanization)
* Soil and hydrology features (e.g., soil type, aquifer type, permeability)
* Weather conditions (e.g., precipitation, evaporation, drought periods)
* Groundwater depth and environmental conditions (e.g., chloride levels, redox)

This synthetic data was used to build and test early versions of the model, allowing us to experiment with variable interactions and model behaviour in a controlled setting.

After compiling the initial synthetic dataset—based on expert input, domain literature, and expected environmental ranges—we began exploring the data to understand patterns, detect anomalies, and guide model design.

#### Structure and Completeness

We first verified that the dataset included all the critical variables related to nitrate leaching as identified in the business understanding phase. These included time-based indicators (year, month), environmental parameters (pH, EC, ammonium, chloride), land use characteristics (crop type, livestock density, urbanization), and soil/hydrological features (e.g., soil type, groundwater level, permeability). Synthetic data was checked for:

* **Missing values** — intentionally introduced in some columns to simulate real-world gaps
* **Data types** — ensuring categorical vs. numerical consistency
* **Value ranges** — all features were checked against realistic Dutch agricultural and climatic values (e.g., precipitation between 40–120 mm/month, nitrate levels between 10–80 g/m³)

# Data Preparation & Collection

## The meaning of the data

### Metadata and Location

* **well\_no**: Unique identifier for the monitoring well.
* **original\_wellno**: Code for the original well if this one is a replacement.
* **well\_code**: DINO code assigned to the well for geological tracking.
* **well\_objectcode**: BRO registry code for the groundwater well.
* **network\_name**: The specific network the well belongs to (e.g., LMG).
* **filter\_no**: Screen or filter number used in multi-depth wells.
* **x\_coord / y\_coord**: RD New coordinate system for mapping well location in the Netherlands.
* **province**: Dutch province where the well is located.
* **local\_placename**: Nearest town or locality for the well.

### Well and Geological Information

* **ground\_surface\_elevation**: Height of the land surface above NAP (Dutch sea level reference).
* **depth\_topfilter / depth\_bottomfilter**: Depths of the top and bottom of the well filter below the surface.
* **height\_wellcasing**: Height of well casing above the ground.
* **diam\_wellcasing**: Diameter of the well casing in cm.
* **well\_installationyear**: Year the well was installed.
* **aquitard\_code**: Code representing presence and position of low-permeability layers.
* **tritium\_code**: Indicates water age (T1 = young, T2 = old, T0 = unknown).

### Sample Timing

* **sample\_year / plan\_year**: Actual and planning year of the sample collection.
* **sample\_day / sample\_month**: Day and month of the groundwater sample.

### Chemical Parameters (Laboratory Measurements)

* **aluminium** to **zinc** (e.g. arsenic, cadmium, chloride, copper, iron, etc.): Concentration of individual chemical elements in groundwater.
  + **Unit**: mg/m³ or g/m³ (≈ g/m³)
  + **Note**: Includes trace metals and nutrients like nitrate, phosphate.
* **ec\_lab**: Electrical conductivity at lab conditions (25°C).
  + **Unit**: mS/m — higher values mean more ions/salts.
* **ph\_lab**: pH of groundwater. <7 = acidic; >7 = basic.
  + **Typical range**: 4.5–8.5
* **nh4n**: Ammonium-N concentration.
* **no3n**: Nitrate-N concentration (important target variable).
* **ptopp / po4p**: Total and inorganic phosphorus.
* **so4**: Sulfate ion concentration.
* **corg**: Dissolved organic carbon.

### Field Measurements (VELD)

* **O2\_5\_field**: Dissolved oxygen measured in the field.
* **T\_5\_field**: Temperature in °C at time of sampling.
* **EC\_5\_field**: Field EC measurement.
* **H\_5\_field**: Field pH.
* **SH\_field**: Hydraulic head relative to surface.
* **HCO3\_field**: Field-determined bicarbonate (alkalinity).

### Environmental and Land Use Context

* **Land\_use\_type**: Dominant land use at sampling location (e.g., dairy, arable, mixed).
* **Livestock\_density\_LSU\_per\_ha**: Number of livestock units per hectare by province.
* **Sunlight\_hours**: Monthly total sun hours.
* **Rel\_humidity\_percent**: Mean monthly relative humidity.
* **Avg\_temperature\_C**: Monthly mean air temperature.
* **Precipitation\_mm :** Monthly rainfall (may be duplicate; consolidate).
* **nitrate\_input\_mg\_L**: Estimated nitrate added from fertilizers or manure.
* **fertilizer\_kg\_ha**: Synthetic nitrogen fertilizer use per hectare.
* **manure\_kg\_ha**: Organic nitrogen from manure per hectare.
* **crop\_type**: Crop grown (e.g., maize, potatoes, wheat, grass).
* **livestock\_density\_LU\_ha**: Another measure of livestock intensity.
* **urbanization**: Categorical indicator (Low, Medium, High urban development).
* **soil\_type**: Textural class (Sand, Clay, Peat, Loess).
* **aquifer\_type**: Confined or unconfined aquifer.
* **permeability\_m\_day**: Hydraulic conductivity (m/day) — water movement through soil.
* **groundwater\_level\_cm**: Water table depth below surface.
* **evaporation\_mm**: Monthly evaporation.
* **drought\_days**: Dry days in the sample month.

### Target Variable

* **no3n**: Final nitrate concentration in groundwater (g/m³).

Used as the prediction target. EU safe limit: 50 g/m³.

## Data Integration

**How data from multiple sources was combined:**

* **Primary Source**: The original dataset (LMG\_data\_NIR\_2024.csv) contained groundwater quality measurements from the Netherlands, including pH, nitrate, ammonium, and various trace elements.
* **Supplemental Synthetic Data**: Since some essential environmental and agricultural variables were missing, we **simulated data** based on **known Dutch climate and agricultural patterns**, effectively integrating:
  + **Climatic data** (monthly precipitation, evaporation, sunlight, temperature, humidity, drought days)
  + **Land use and farming practice variables** (nitrate input, fertilizer/manure application, livestock density, crop types)
  + **Soil and aquifer characteristics** (soil type, aquifer type, permeability)

This was done by conducting research about Average Dutch soil and climate data.

## Feature Engineering

**New features created from original or missing data:**

1. **Climatic Monthly Averages (based on 'month'):**
   1. precipitation\_mm: Estimated monthly rainfall using average Dutch climatology.
   2. evaporation\_mm: Estimated average monthly evaporation.
   3. drought\_days: Estimated number of dry days per month.
   4. avg\_sunlight\_hours: Based on latitude-driven seasonal variation.
   5. avg\_humidity\_percent and avg\_temp\_C: Monthly climatic conditions used to simulate plant and leaching dynamics.
2. **Agricultural Inputs:**
   1. nitrate\_input\_mg\_L: Synthetic nitrate loading to soil (not measured in the dataset, but crucial for modeling).
   2. fertilizer\_kg\_ha and manure\_kg\_ha: Added to capture nitrogen input from synthetic and organic sources.
   3. livestock\_density\_LU\_ha: Estimate of nitrogen loading from animals.
3. **Categorical Classifications:**
   1. crop\_type: Simulated based on typical Dutch crop distribution.
   2. urbanization: Proxy for land use intensity.
   3. soil\_type and aquifer\_type: Simulated based on typical distributions in Dutch groundwater studies.
4. **Hydrogeological Properties:**
   1. permeability\_m\_day: Affects water movement through soil, influencing leaching.
   2. groundwater\_level\_cm: Simulated water table depth (important for contamination risk).

## Data Cleaning

### Dropped columns

#### 1. Metadata Columns

**Dropped:**  
 well\_no, original\_wellno, well\_code, well\_objectcode, network\_name, filter\_no, x\_coord, y\_coord, province, local\_placename

**Reason:**  
 These columns contain identifiers, coordinates, or location names that do not carry predictive information themselves. Such metadata is useful for recordkeeping but not suitable for machine learning input.

#### 2. General Descriptive Data

**Dropped:**  
 ground\_surface\_elevation, depth\_topfilter, depth\_bottomfilter, height\_wellcasing, diam\_wellcasing,  
 well\_installationyear, aquitard\_code, tritium\_code, sample\_year, plan\_year, sample\_day, sample\_month,  
 O2\_5\_field, T\_5\_field, EC\_5\_field, H\_5\_field, SH\_field, HCO3\_field, Land\_use\_type

**Reason:**  
 These features are:

* **Highly specific to individual wells** (e.g., depth or installation year),
* **Time-based measurements** not directly relevant to nitrate levels,

These features also do not provide any predictive value.

#### 3. Mineral Concentration Columns

**Dropped:**  
 aluminium, arsenic, barium, calcium, cadmium, chloride, chromium,  
 copper, fluoride, iron, lead, magnesium, manganese, nickel,  
 potassium, sodium, so4, stronium, zinc

**Reason:**  
While mineral concentrations are chemically related to water quality, they were removed because we scaled it under the term ''electrical conductivity'', since EC reflects the total concentration of dissolved ions in water.

Additionally, we aimed to create a model that relies more on **e**asily observed agricultural and environmental variables, rather than lab-based water chemistry data that may not be available at prediction time.

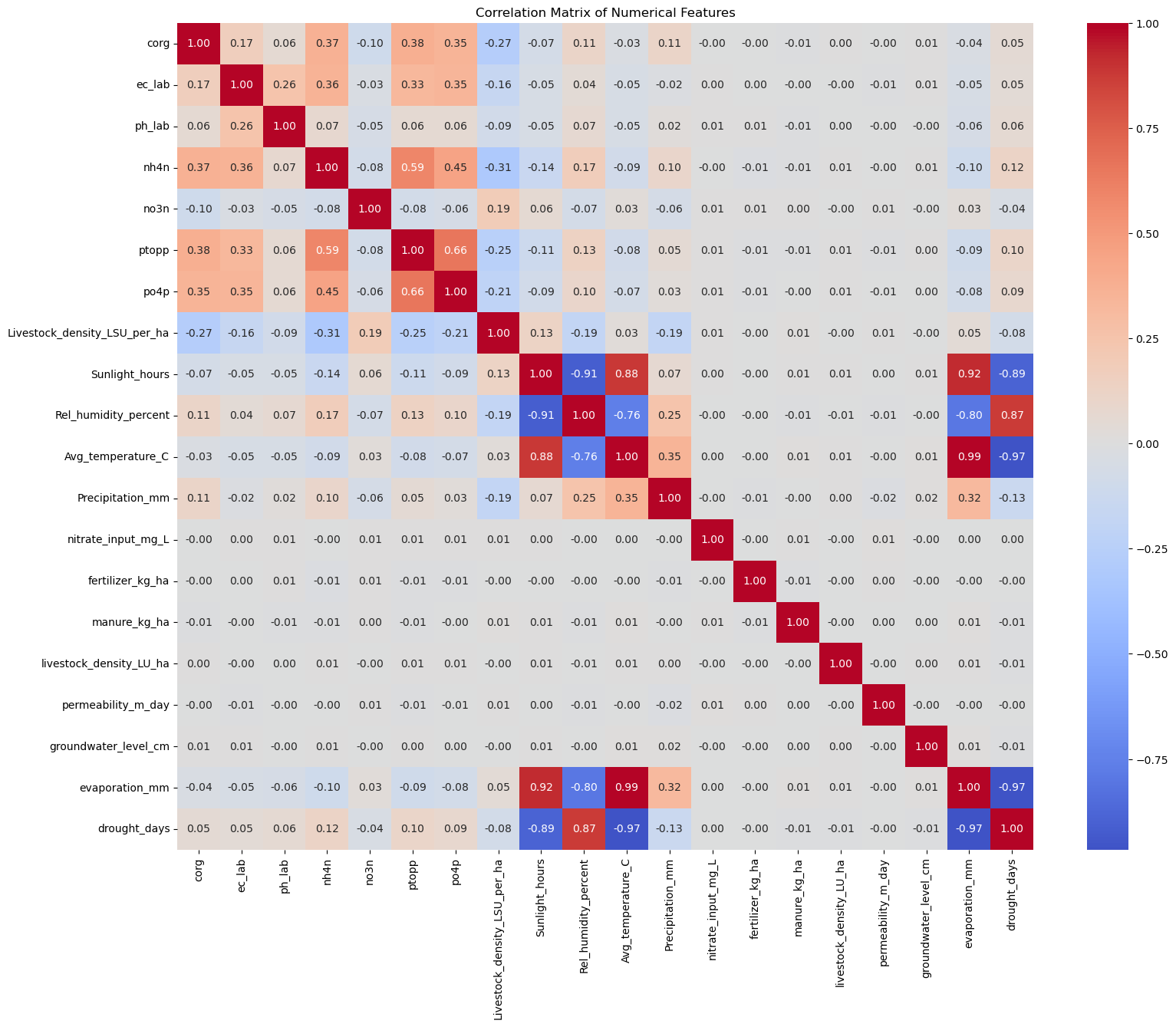
## Data Analysing

After the initial data exploration, there was some analysis of the data. This was very ground level.

### Correlation matrix

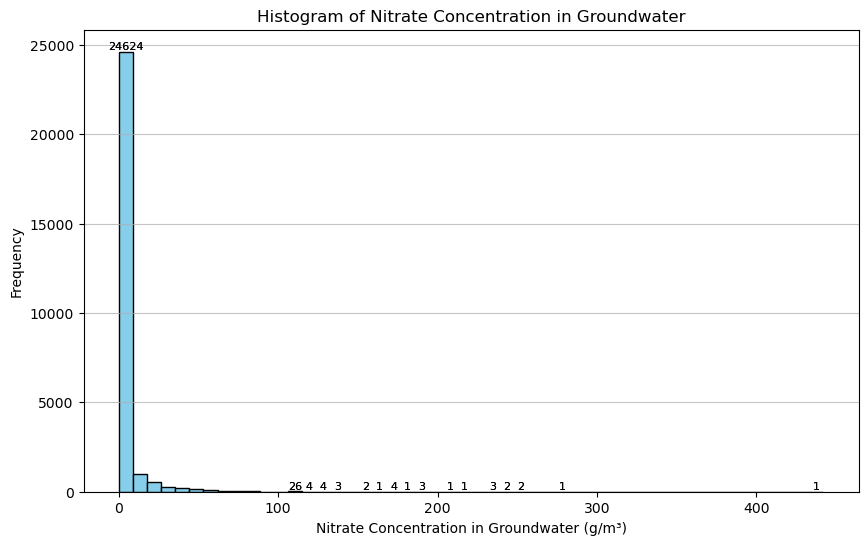
As seen on the correlation matrix below there were some correlations in the data set. When looking at the matrix it shows that there are some obvious correlations. For example, the weather data all correlates with each other.

Besides this the RIVM data shows that the lab values correlate with each other and with a bit off the weather data. Although the correlations are not that high. But this gives further conformation on the fact that weather has influence on some elements. Which on their turn have influence on the nitrate in the ground water.

*Figure 1, Correlation model, Own work*

### Histogram nitrate concentration

When looking at the nitrate concentration amounts it shows something interesting. It shows that for the most part the nitrate concentration is 50 and mostly even lower. Although there are times where the nitrate is way higher then allowed. As shown in the graph below.

*Figure 2, Histogram nitrate concentration count, Own work*

### Decision

It shows in these columns that some data has correlation and others do not. This is because we used synthetic data, but our research has shown that these factors affect nitrate levels in groundwater. Better quality data would prove more useful during the deployment of this model.

# Modelling

To explore the feasibility of predicting nitrate concentrations in groundwater, we tested several supervised regression models. These included:

* **Linear Regression** – as a baseline to establish a minimum performance benchmark.
* **Random Forest Regressor** – known for robustness to overfitting and handling mixed data types.
* **XGBoost Regressor** – selected for its high predictive accuracy, efficiency with structured data, and ability to model complex non-linear relationships through gradient boosting.
* **CatBoost** - known for its strengths in handling categorical features and its overall efficiency and accuracy

### Parameter Tuning

We implemented systematic hyperparameter optimization for XGBoost using GridSearchCV with 5-fold cross-validation. The search space included:

**n\_estimators:** [100, 200] (number of boosting rounds)

**max\_depth:** [3, 6, 10] (tree complexity control)

**learning\_rate:** [0.01, 0.1, 0.2] (shrinkage weights)

**subsample:** [0.8, 1.0] (stochastic sampling ratio)

Parallel computation was enabled via n\_jobs=-1, and models were evaluated using negative mean squared error during tuning.

### Training & Validation

The dataset was split into training (80%) and testing (20%) sets using train\_test\_split with random\_state=42 for reproducibility. Preprocessing included:

* Numerical features: Mean imputation and standard scaling
* Categorical features: One-hot encoding with unknown category handling

These operations were encapsulated in a Pipeline to ensure consistent transformations during training/inference.

### Evaluation Metrics

Performance was assessed on the test set using:

**Root Mean Squared Error (RMSE)**: Prediction error magnitude

**Mean Absolute Error (MAE)**: Interpretable average deviation

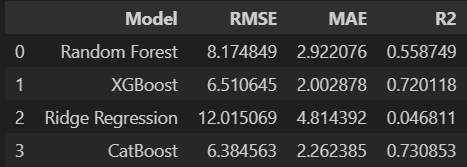
**R² Score**: Proportion of variance explained

Results on the test set:

* **RMSE**: *≈ 6.51 g/m³*
* **MAE**: *≈ 2* g/m³
* **R²**: *≈* 0.720

These metrics indicate that the XGBoost model was able to reasonably predict nitrate levels, though further improvements could be achieved with real-world data, hyperparameter tuning, and feature refinement.

Visual validation was performed through a predicted vs. actual scatter plot with reference line.



*Figure 3, modelling results, Own work*

### Key Advantages Observed

XGBoost demonstrated superior predictive capability compared to baseline models, efficiently capturing non-linear interactions (e.g., soil type × rainfall effects on nitrate leaching) while maintaining computational efficiency through parallelization.

## The interface

This interface allows users to estimate the concentration of nitrate (NO₃-N in mg/L) in groundwater based on a range of environmental and chemical parameters. It was designed to be accessible and intuitive, enabling users to interact with a trained machine learning model without needing to write code.

At the top of the interface, a title section provides a brief explanation of the tool and its purpose. Directly below that, users are presented with a set of input fields corresponding to various parameters relevant to groundwater quality. These parameters include both numerical values, such as electrical conductivity or pH, and categorical features, such as soil type or land use classification.

Each input field is accompanied by additional information to guide the user. For numerical inputs, hovering over the field reveals statistical context, including the range of observed values and the median. This helps users input values that are realistic and within expected limits. For categorical inputs, dropdown menus are provided with predefined options that reflect the values used during model training. Importantly, all input fields can be left blank if the relevant information is unavailable. The system is designed to handle partial input and will make a prediction using whatever data is provided.

After entering the desired parameters, the user can click the "Predict Nitrate Level" button. The model will process the inputs and produce a prediction of the nitrate concentration in the groundwater sample. The result is shown clearly on the screen, along with a qualitative interpretation of the risk level associated with the predicted concentration. This interpretation is based on established thresholds, such as the European Union’s limit of 50 mg/L for drinking water.

In addition to displaying the predicted nitrate value, the interface also provides a confidence score. This score estimates the reliability of the prediction based on how many of the most important input features were filled in. The confidence is displayed as a visual bar and accompanied by a textual label (e.g., “Low Confidence” or “High Confidence”), allowing users to better judge the robustness of the output.

To assist with transparency and reproducibility, the interface also summarizes which parameters were provided by the user and which were left blank. This input summary helps users identify whether additional information could improve the prediction accuracy.

A second button, labeled "Clear All," allows users to reset the interface quickly. This is useful when performing multiple predictions or starting over with a different set of input values.

A screenshot of a computer

AI-generated content may be incorrect.In summary, the interface provides an accessible way to use a machine learning model for predicting nitrate concentrations in groundwater. It accommodates incomplete data, supports decision-making through risk and confidence interpretation, and includes user-friendly guidance to ensure meaningful input.

*Figure 4, Groundwater nitrate prediction interface, Own work*

# Evaluation

Since we built our model using synthetic data but similar to real-life situations, our main goal was to check if the model works well rather than the exact result. We used RootMean Squared Error (RMSE**)** and R-squared (R²) to assess prediction accuracy and the proportion of variance explained by the model. After trying different models, the XG Boost Model worked perfectly with the RMSE: 6.51 g/m³ and R²: 0.72, which means that the model can predict nitrate levels accurately, and it can be a useful option instead of traditional lab tests.

This model will have significant business value for farmers and government companies like RIVM. One of the benefits of this model is that it is cost-effective compared to traditional lab tests. Usually, the laboratory test results take a few days to publish, but with this model, it is possible to predict nitrate leaching instantly based on available data. This model will make organizations work faster and more efficiently. It will help them quickly spot areas that are at risk, make better rules for fertilizer use, plan land use wisely, and enforce environmental laws more effectively. In this way, the model will support clean water, protect public health, and reduce the need for frequent field testing.

Besides all the benefits, there is also a limitation to this model. As the model was built based on synthetic data, there might be limitations when applying the real data. As the current data set is artificial, conclusions are provisional. A detailed evaluation with actual RIVM data from approximately 450 farms, which requires a Non-Disclosure Agreement (NDA), is recommended as the next step for validation and improvement.

# Deployment

The model is built specifically for farmers, government organizations, and local water authorities. It will help them monitor nitrate levels in groundwater without needing expensive lab testing. By inputting some important data like environmental, agricultural, and water-related data, the model will quickly predict the nitrate levels in groundwater. This will help the authorities to make quicker and wider decisions about fertilizer rules, how land will be used, and how to keep water safe.

The model will have a very simple and user-friendly interface. It is designed in such a way that the farmers can easily understand how the model will work and which data they will need to input. There will be clear instructions and helpful tips for filling out key data fields, such as crop type, amount of manure, fertilizer use, and local environmental conditions. The model will give farmers practical tips to reduce nitrate levels, especially when they are too high. It will suggest useful changes—like using less fertilizer, changing irrigation methods, rotating crops, or managing manure better—and explain how each change can help and why it matters.

To keep the model accurate over time, it will need to be updated regularly using new, real-world data, like recent measurements from groundwater wells or updated climate and land use information. As more data is added, the model settings might also need to be adjusted to keep it working well. In our modelling, we dropped the chemical compositions as analyzing and validating the chemical composition data is expensive and time-consuming, but in the future, to be more accurate they can use those chemical compositions.

The person in charge of data or the IT team should be in charge of looking after the model. They should track how well it performs, manage different versions of it, and make sure all data is handled properly and follows the rules.

# Conclusion

The purpose of this project was to check whether machine learning can be used to predict nitrate levels in Dutch groundwater because, usually, the traditional lab testing is expensive and time-consuming. We used synthetic data based on real environmental and agricultural conditions. We tried different models and decided to work with the XG Boost model because of its high accuracy rate.

Though the model worked well but we cannot yet say how well it will work with real data. To take the next step, the model needs to be tested with the RIVM data. By implementing the model, the farmers and the government organisations can predict the nitrate level easily and cost-effectively. Further development and a proper validation of the model can significantly reduce the cost and time which were required to monitor nitrate pollution. They can also provide more timely information, enabling quicker responses to potential environmental risks, better land management strategies, and more effective implementation of nitrate regulations. It can also help farmers decide how much fertilizer they should use on the ground.

The climate is changing very frequently, and the rules are also getting stricter day by day. Farmers need to be careful to be safe from getting fines. Machine learning can make farmers’ lives easier and affordable as it can predict the nitrate level in a shorter time and cheaply. Machine learning is a game-changer across many industries.

# Bibliography

**European Commission, Directorate‑General for Environment. (2022, September 30).** *Commission Implementing Decision (EU) 2022/2069 on granting a derogation requested by the Netherlands pursuant to Council Directive 91/676/EEC concerning the protection of waters against pollution caused by nitrates from agricultural sources* (Official Journal L 277, pp. 195–207) [Decision]. EUR‑Lex. [https://eur‑lex.europa.eu/eli/dec\_impl/2022/2069/oj](https://xn--eurlex-dg0c.europa.eu/eli/dec_impl/2022/2069/oj)

Government of the Netherlands, Ministry of Agriculture, Fisheries, Food Security and Nature. (n.d.). *The nitrogen strategy and the transformation of the rural areas*. Government.nl. Retrieved June 16, 2025, from <https://www.government.nl/topics/nature-and-biodiversity/the-nitrogen-strategy-and-the-transformation-of-the-rural-areas>

Naus, F., & Valster, N. (2025, January 7). *Landelijk Meetnet Grondwaterkwaliteit (LMG) data NIR 2024* [Data set]. National Institute for Public Health and the Environment (RIVM). <https://doi.org/10.5281/zenodo.14923733>

Statistics Netherlands (CBS). (2024, February 28). *Nitrogen and phosphate in animal manure within limits for 2023* [News]. CBS. [https://www.cbs.nl/en‑gb/news/2024/09/nitrogen-and-phosphate-in-animal-manure-within-limits-for-2023/production-ceiling--nitrogen-and-phosphate--](https://www.cbs.nl/en%E2%80%91gb/news/2024/09/nitrogen-and-phosphate-in-animal-manure-within-limits-for-2023/production-ceiling--nitrogen-and-phosphate--)

United Nations Framework Convention on Climate Change. (n.d.). *The Paris Agreement*. UNFCCC. Retrieved June 16, 2025, from <https://unfccc.int/process-and-meetings/the-paris-agreement>