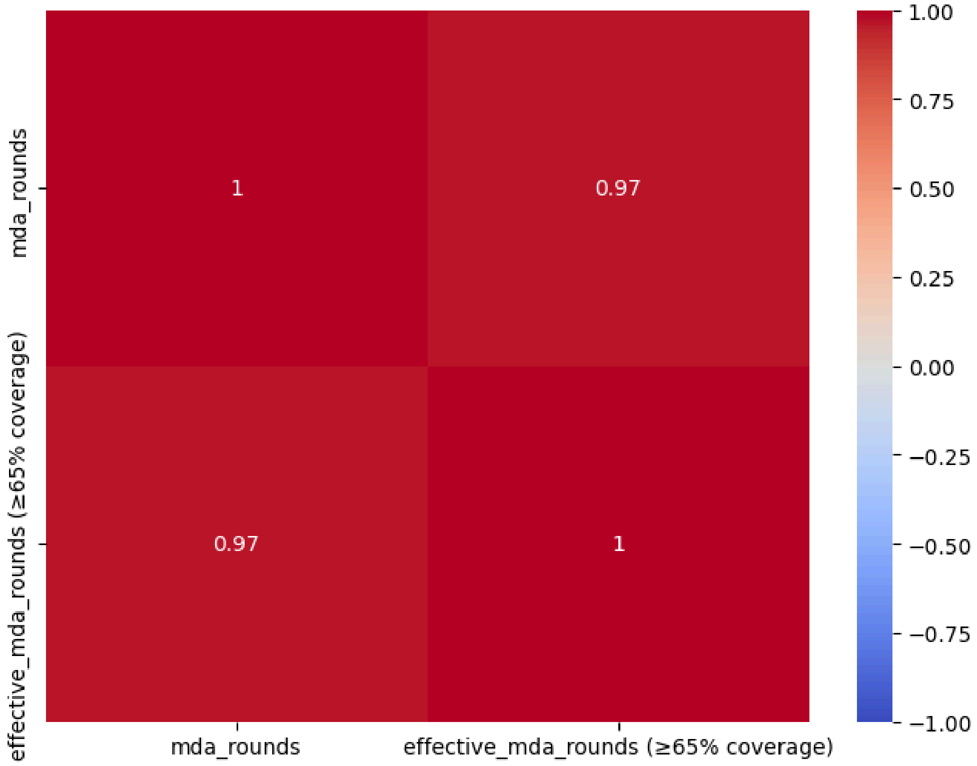
data report: wania

description of data sources

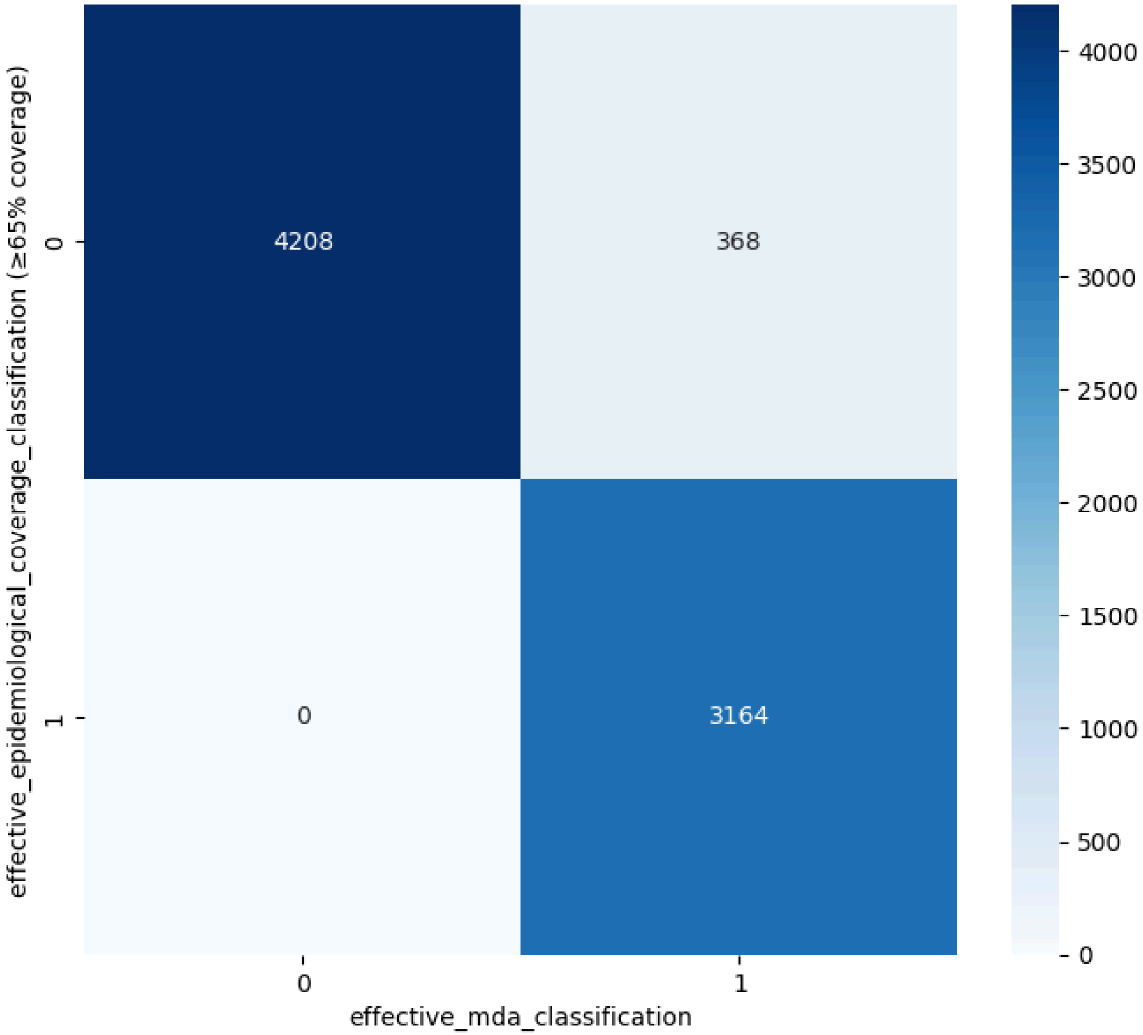
- endemicity of each region for each region, for each year (2014 — 2023)
  - non-endemic
  - endemic
  - undergoing post-intervention surveillance which was **renamed** to post-intervention surveillance
  - **merged** the endemicity of each year into one
  - **source:** https://espen.afro.who.int/api/espen-dashboard/maps/data-download/iu/58/NG/**{year}** **(change 'year' based on the data you're looking for)**
- mda/pca therapeutic coverage for each region, for each year (2014 — 2023)
  - iu_code
  - iu_region
  - therapeutic_coverage
  - **merged** the endemicity of each year into one
  - **source:** https://espen.afro.who.int/api/espen-dashboard/maps/data-download/iu/58/NG/**{year}** **(change 'year' based on the data you're looking for)**
- iu level data for each region, for each year (2014 — 2023)
  - iu_code
  - **iu_region and state were not dropped, because this made it easier to merge this dataset with the geographical coordinates (will be dropped later in the merging section).**
  - year
  - population_requiring_treatment
  - population_treated

- epidemiological_coverage (% of the total population who received treatment)
- effective_epidemiological_coverage_classification (≥65% coverage)
- total_population
- mda_rounds
- effective_mda_rounds (≥65% coverage) **(ask amir whether including mda_rounds and effective_mda_rounds is a good idea, based on their correlation)**



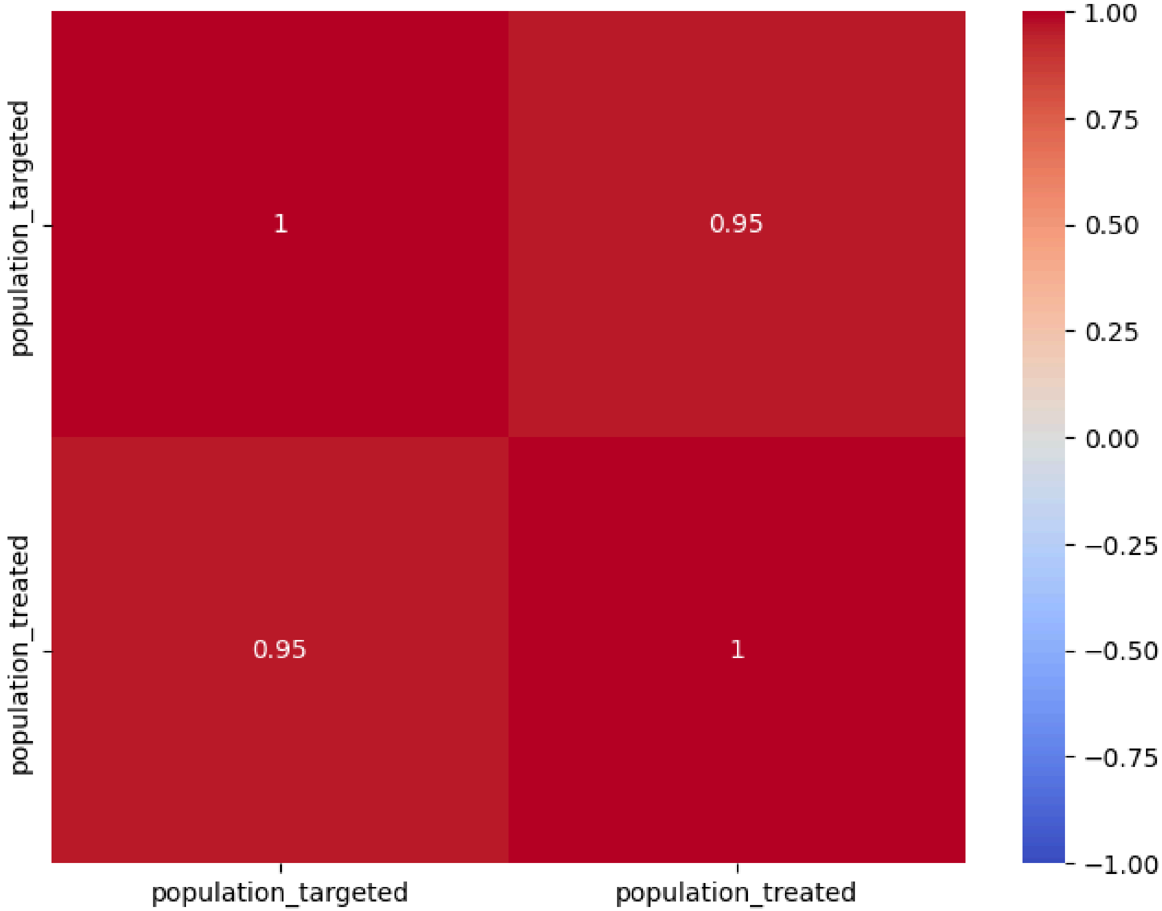
- epidemiological_coverage_classification (≥65%) was **dropped** because we have the contentious value for this (epidemiological_coverage).
- effective_mda_classification (≥65% coverage) was **dropped** because there is strong agreement between this and effective_epidemiological_coverage_classification (≥65%). when mda classification is high, epidemiological coverage is high. however, high effective epidemiological coverage never happens without high effective mda, even though high effective mda can happen with low effective epidemiological coverage.

        - so effective epidemiological classification can be seen as a subset or downstream result of effective mda classification.
        - so we end up using effective_epidemiological_coverage_classification instead of effective_mda_classification, because sometimes programs can report high mda coverage, but if they miss a large part of the at-risk population, the epidemiological coverage will be lower = epidemiological coverage would tell you if the population at risk overall is actually protected = it is a better indicator of actual impact on the disease.
        - **the relationship is shown with the following cross-tabulation:**

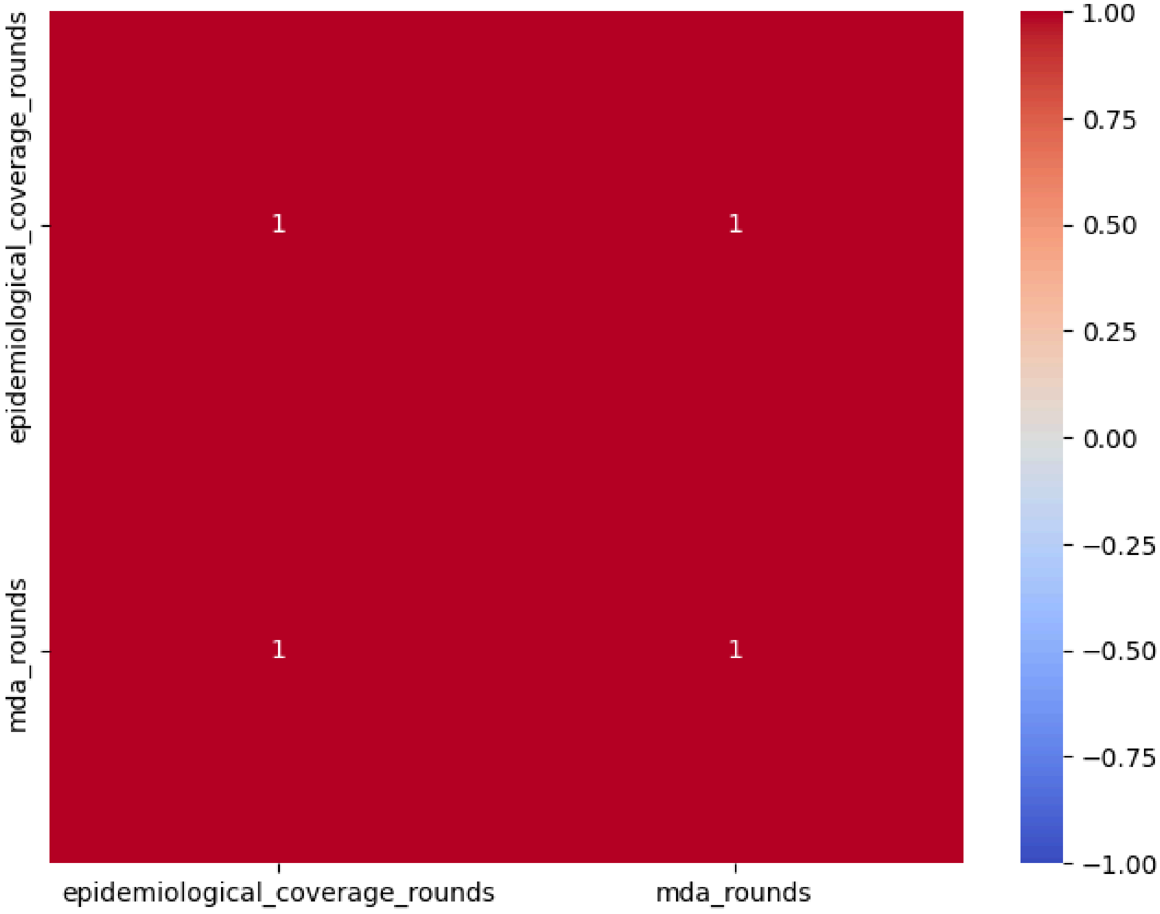


    - province which **dropped** because iu_code already uniquely identifies locations consistently across all data sources.

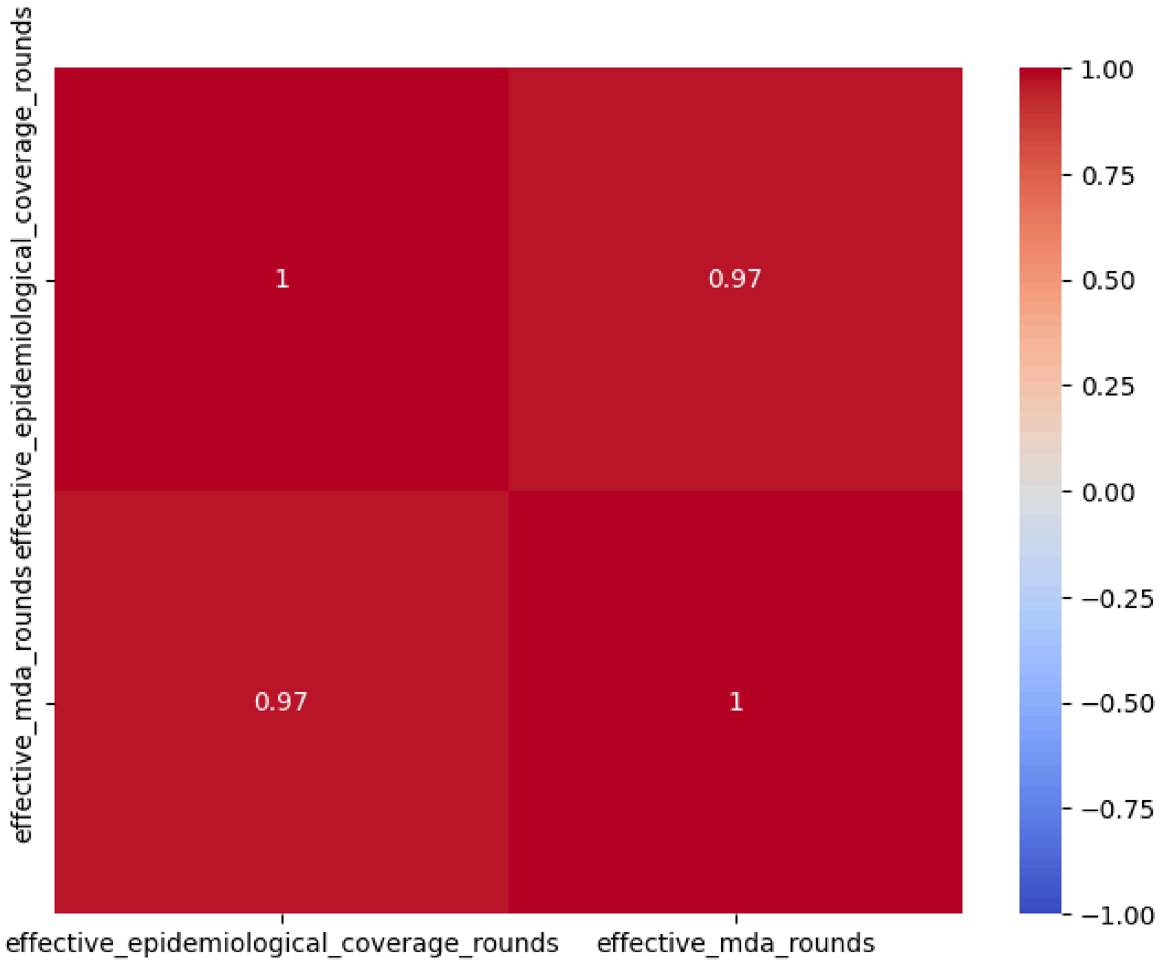
- population_targeted was **dropped** because of strong collinearity with population_treated **(as seen below):**



- progmmatic_coverage (% of the targeted or eligible population that got treated) was **dropped** because this is the same as therapeutic_coverage.
- mda_classification was **dropped** because this is less descriptive then mda_rounds, which is a continuous measure.
- cumulative_mda_rounds was **dropped** because this is just mda_rounds aggregated.
- epidemiological_coverage_rounds was **dropped** because of strong collinearity with mda_rounds) **(as seen below):**



- effective_epidemiological_coverage_rounds (≥65% coverage) was **dropped** because of strong collinearity with effective_mda_rounds (≥65% coverage) **(as seen below):**



- popPreSAC, popSAC, and popAdult were **dropped** because our model does not need to predict on the basis of population categories, only region categories.
- continent, region, whoRegion, admin0, admin0Id, admin0Fip, admin0Iso2, admin0Iso3, iusAdm, admin1Id, ius_adm, and admin2Id were **dropped** because they are all noise, as we chose iu_code to differentiate among regions.
- mda_scheme was **dropped** because knowing which treatment was used is irrelevant to our model.
- endemicity was **dropped** because we have that through our therapeutic_coverage data source.
- endemicity_id was **dropped** because it is just a different way of classifying endemicity, which we have decided to drop from above.
- **source for download:**

    iu_level_data_2014-2023.csv

- **geographical coordinates for each region**
    - iu_region (for making merging easier with iu_level_data)
    - state (for making merging easier with iu_level_data)
    - latitude
    - longitude
    - id, state_id, state_code, country_id, country_code, country_name, and wiki_data_id were **dropped** because these are just different ways of saying the location.
    - **source for download:**

      LGA.csv

description of merging process

- **one:** the mda/pca therapeutic coverage for each region (called df_0) was merged with the iu level data for each region (called df_1). this was through merging on the iu_code and the year, as they both shared this. each row had a match, and nothing had to be dropped. the merged data source was called df_3.
- **two:** the geographical coordinates (called df_2) and df_3, were merged together. this allowed for the longitude and latitude from df_2 to be incorporated into df_3 based on whether they had the same iu_region and state. several rows had missing values because the spelling of some regions was different (though similar) in both data sources. this was solved through iterating over the data sources, figuring out which ones had similar spelling, cross-referencing google, and then pasting in the coordinates from the LGA file.

description of each feature (without normalization)

feature	description	categories
year	year	[2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023]
therapeutic_coverage	the percentage of the targeted or eligible population who actually received treatment.	[‘non-endemic’, ‘no mda coverage in endemic area', 'ineffective coverage (<65%)', 'effective coverage (>=65%)', 'post-intervention surveillance']
total_population	the total population	number
population_requiring_treatment	the population requiring lymphatic filariasis treatment	number
population_treated	the population treated for lymphatic filariasis	number
epidemiological_coverage	the percentage of the total population at risk (entire population in the endemic area) who actually received treatment.	%
mda_rounds	the total number of times mass drug administration campaigns that been conducted in the region.	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
effective_mda_rounds (≥ 65% coverage)	the number of mda rounds that were successfully implemented with	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

feature	description	categories
	sufficient coverage (≥ 65%)	
effective_epidemiological_coverage_classification (≥65% coverage)	a binary classification on whether the epidemiological_coverage was effective.	[0, 1]

description of label

- i was thinking of using **prevalence (%)** as our label, the percentage of individuals infected within a population sample. however, there are several challenges:
    - **sparse data coverage:** our data has prevalence percentage for only 447 out of 774 regions (LGAs). that means nearly 42% of regions lack prevalence labels.
    - **sampling bias and representativeness:** the prevalence is derived from a **small sample size** relative to total population (~600–2000 tested out of ~100,000 population per region). this would mean that the prevalence may not accurately represent the true disease burden across the whole region, and local heterogeneity can cause noisy or biased prevalence estimates.
- therefore, we could then consider using **endemicity**, which is a categorical classification assigned by WHO and ESPEN based on multiple variables including prevalence, historical data, intervention status, and expert evaluation.
    - the categories include:
        - **non-endemic:** no sustained transmission detected.
        - **endemic:** active transmission ongoing.
        - **post-intervention surveillance:** areas where active interventions such as mda have ceased but surveillance continues to confirm elimination.
    - potential advantages:
        - **complete coverage:** endemicity status is available for all 774 regions.
        - **policy relevance:** reflective of practical elimination status recognized by WHO and ESPEN.
        - **meaningful:** predicting whether a region remains or becomes endemic / post-intervention is a directly actionable output for program planning and

        resource allocation.

    - potential disadvantages:
        - **post-intervention surveillance as a "gray zone":** this category may sometimes hide uncertainty or limited data. some regions labeled post-intervention could potentially experience resurgence (return to endemicity). our model needs to learn these dynamics from feature trends (such as treatment coverage or climate change).
        - **granularity loss:** moving from continuous prevalence to categorical endemicity loses some fine detail, but given data sparsity and noise, the tradeoff improves reliability.

handling NAN and missing values

- **thank you god, there were no NAN or missing values for the features or label** 🤲🏼

handling encoding text features

feature	method	description
year	one-hot encoding	year is categorical; one-hot encoding would allow model to learn separate effects for each year; would prevent false assumptions about “closeness” between years.
therapeutic_coverage	one-hot encoding	one-hot encoding would keep categories independent; would avoid incorrectly treating categories as ordered or continuous; would enable the model to detect unique

feature	method	description
		effects of each coverage status.

year	encoding	therapeutic coverage	encoding
2014	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0]	**non-endemic**	[1, 0, 0, 0, 0]
2015	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0]	**no mda coverage in endemic area**	[0, 1, 0, 0, 0]
2016	[0, 0, 1, 0, 0, 0, 0, 0, 0, 0]	**ineffective coverage (<65%)**	[0, 0, 1, 0, 0]
2017	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0]	**effective coverage (>=65%)**	[0, 0, 0, 1, 0]
2018	[0, 0, 0, 0, 1, 0, 0, 0, 0, 0]	**post-intervention surveillance**	[0, 0, 0, 0, 1]
2019	[0, 0, 0, 0, 0, 1, 0, 0, 0, 0]		
2020	[0, 0, 0, 0, 0, 0, 1, 0, 0, 0]		
2021	[0, 0, 0, 0, 0, 0, 0, 1, 0, 0]		
2022	[0, 0, 0, 0, 0, 0, 0, 0, 1, 0]		
2023	[10 0, 0, 0, 0, 0, 0, 0, 0, 1]		

encode it fr (wania)

normalization (all) and why we did it this way

add weather (Alexa)

the input and output for the model (all)

train val and test split (all)