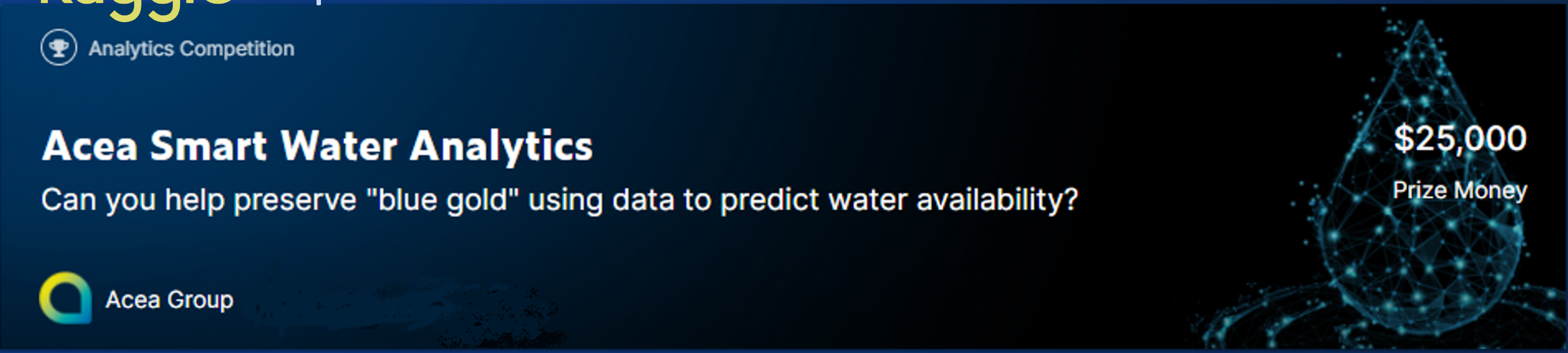
kaggle Competition 2021



한 국 I T 교 육 원

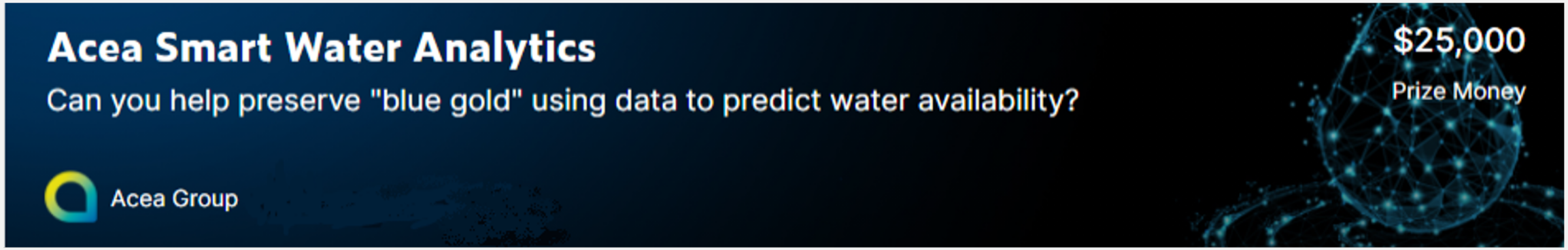
Team : 김규태, 김태윤, 전고은

Date : 2021. 02. 26

CONTENTS

PART I. INTRODUCTION

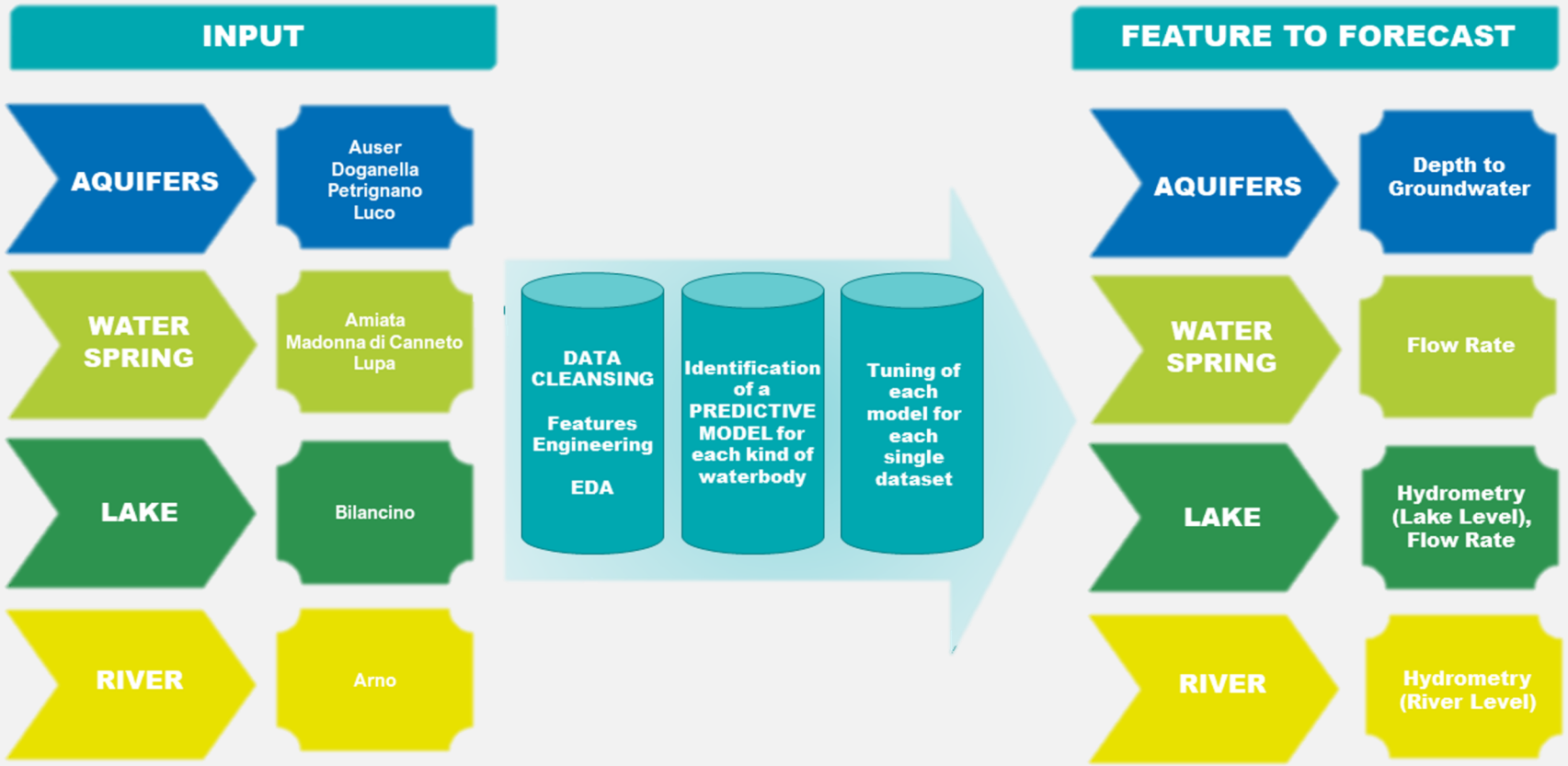
OVERVIEW OF COMPETITION



Competition Timeline :	December 10th, 2020 ~ February 17th, 2021 11:59 PM UTC
Total prizes available :	$25,000
Duration of participation :	10 days (February 8th, 2021 ~ February 17th, 2021)
Participants :	3 persons
Ranking :	announced on 10th March 2021
Kaggle Notebook :	https://www.kaggle.com/arbor0701/notebookb276af2beb

GOAL

To select convincing features and build model which can predict each type of water body level in Italy
To forecast the water level in a waterbody (water spring, lake, river, or aquifer) to handle daily consumption for help Acea Group preserve precious water bodies.



https://www.kaggle.com/c/acea-water-prediction/overview/description

EVALUATION CRITERIA

Methodology	Presentation of logic flow	Application (Prediction)
Statistical models appropriate given the data	**Compelling and coherent narrative**	**Predictive model** able to forecast water availability in terms of level or water flow in a time interval of the year
Developing one or more **machine learning models**	**Data visualizations**	
Way of assessing the performance and accuracy of their solution	Discussion on **the intersection between features and their prediction**	**Methodology applicable** also on new datasets belong to another waterbody
the **Mean Absolute Error** (MAE) of the models	Discussion of **automated insight generation**	
the **Root Mean Square Error** (RMSE) of the models	**Code to easily understand and reproduce**	
	External sources of data made **public and cited appropriately**	

MAE? RMSE?

INDICATOR	FOURMULA	DEFINITION	EVALUATION
MAE (Mean absolute error)	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - y_i'\rvert$	**The mean value** by converting the error value(the difference between the actual and predicted values) to absolute value	Low value is good
$RMSE$ (root-mean-square deviation)	$\sqrt{\sum_{i=1}^{n}\frac{(y_i - y_i')^2}{n}}$	The mean value by squared the difference between the actual value and the predicted value(=error)	Low value is good

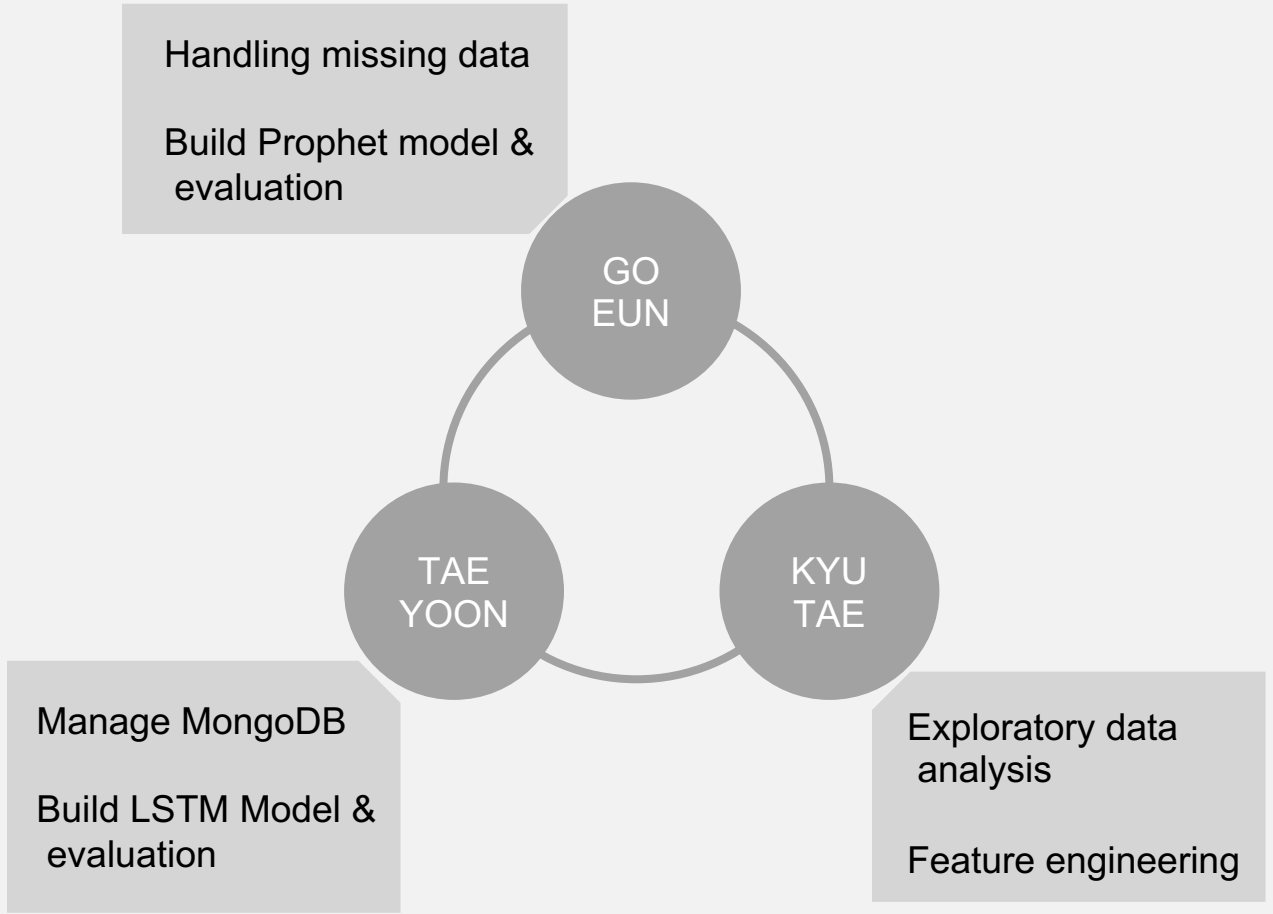
* $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2$

ABOUT OUR TEAM

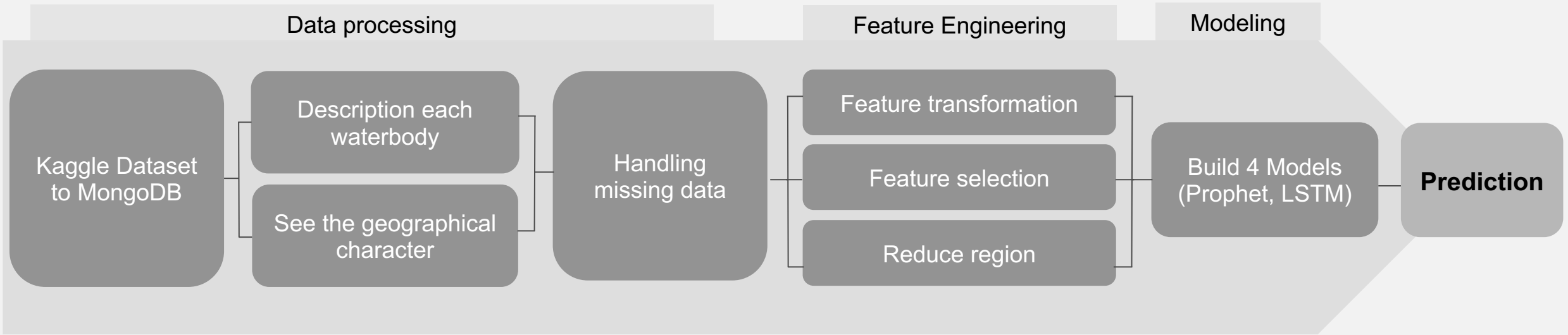
Purpose of participation

- Analyzing Structured Data
- Processing Structured Data
- Understanding the time series data
- Applying Deep learning algorithms

Role

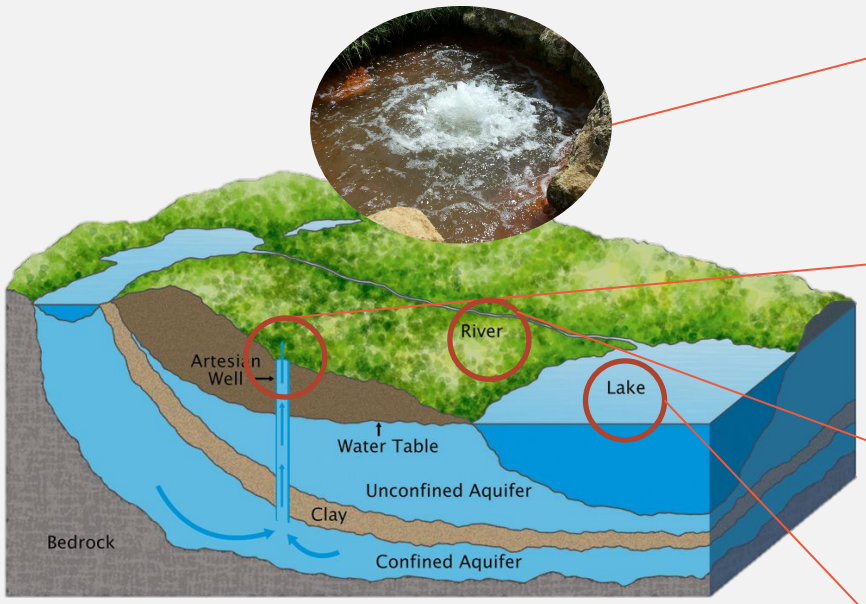


PROCESS


Data processing
Feature Engineering
Modeling
Kaggle Dataset to MongoDB
Description each waterbody
See the geographical character
Handling missing data
Feature transformation
Feature selection
Reduce region
Build 4 Models (Prophet, LSTM)
Prediction

PART Ⅱ. EXPLORATORY DATA ANALYSIS

DATA DESCRIPTION



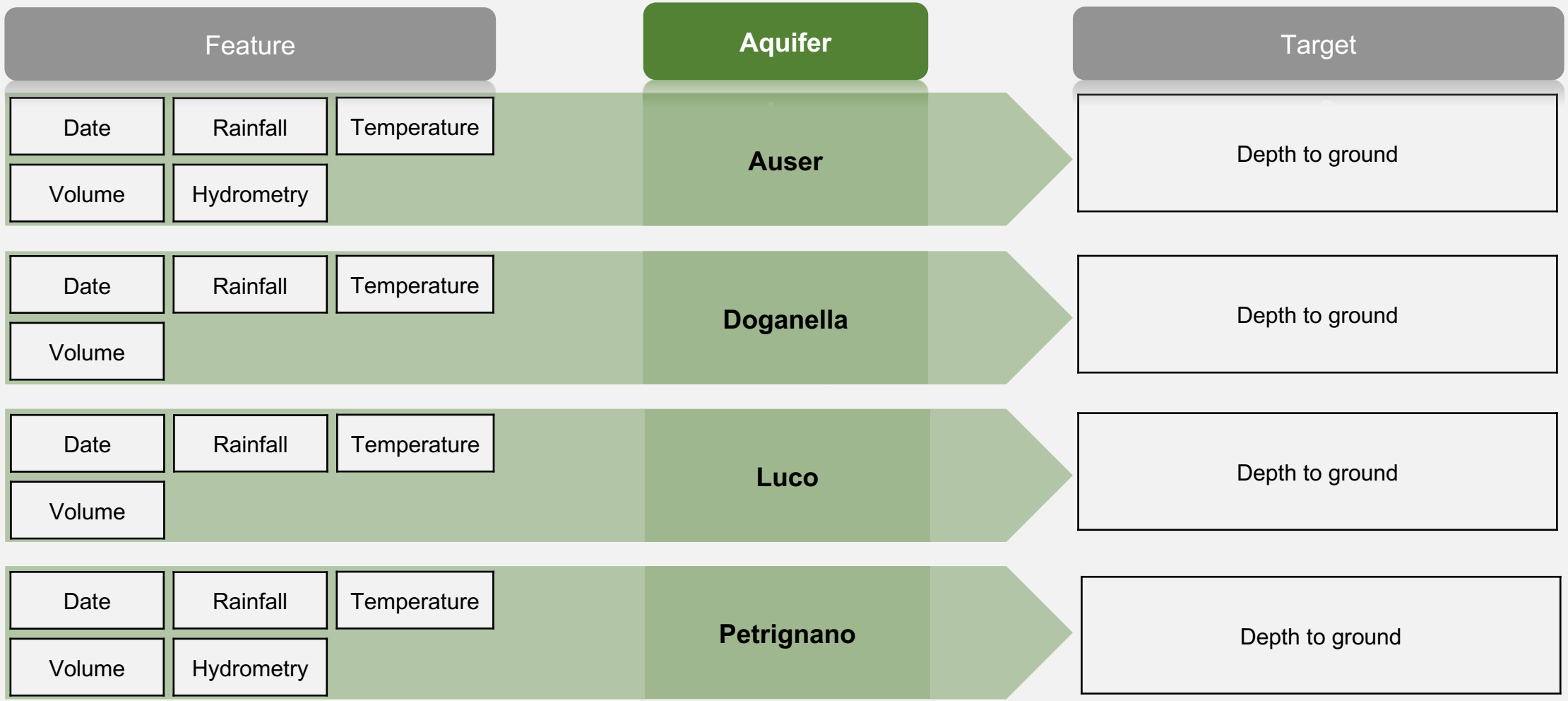
https://www.nationalgeographic.org/media/aquifer-illo/

https://imaggeo.egu.eu/view/1533/

Water body type
Water Spring
Aquifer
River
Lake

Feature	
Date	A day
Rainfall	Quantity of rain falling in the area (mm)
Temperature	Temperature detected by the thermometric station (°C)
Hydrometry	Groundwater level detected by the hydrometric station (m)
Volume	water volume taken from the drinking water treatment plant (m^3)
Flow Rate	flow rate (l/s)(m^3/s)
Lake Level	River level (m)
Depth to groundwater	Groundwater level from the ground floor (m)

DATA DESCRIPTION

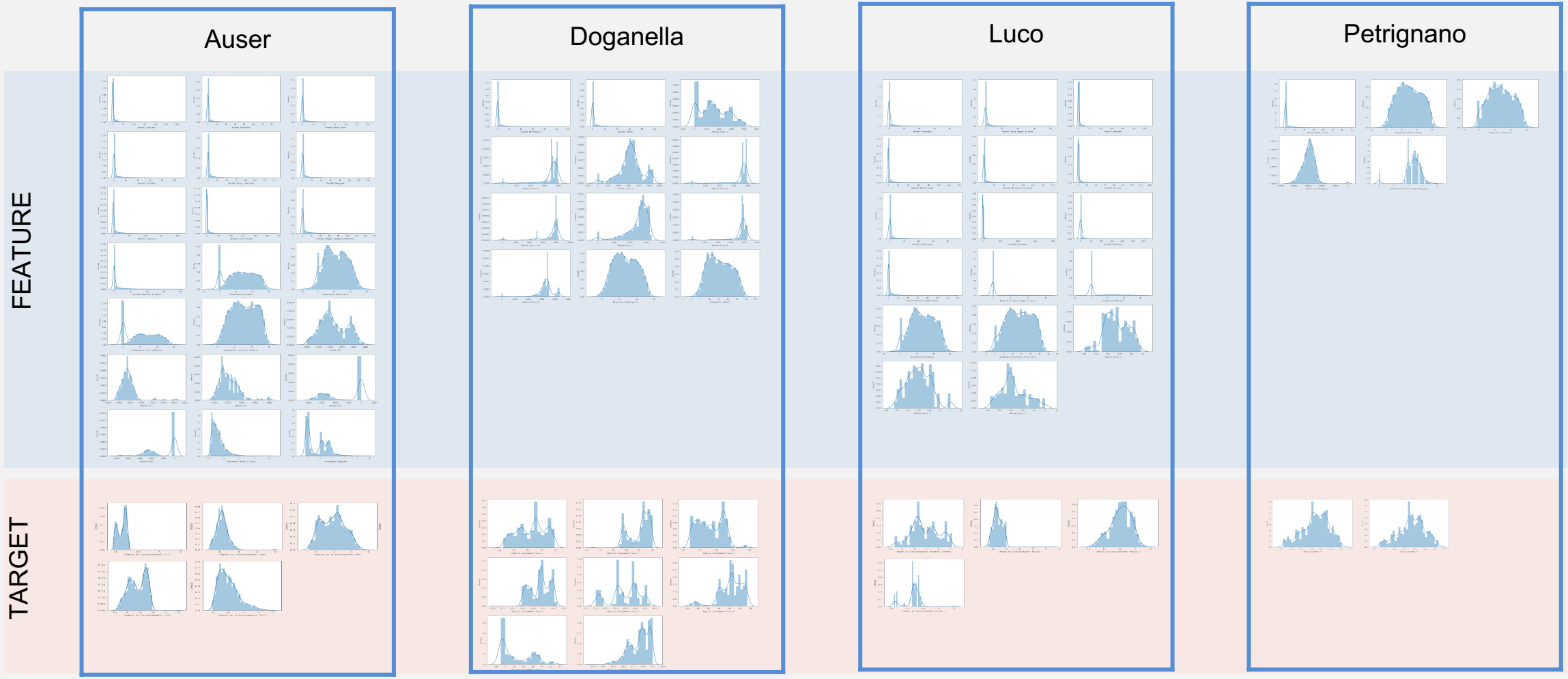


DATA DESCRIPTION

There are a lot of outliers and they don't follow the normal distribution.

Variable transformation required

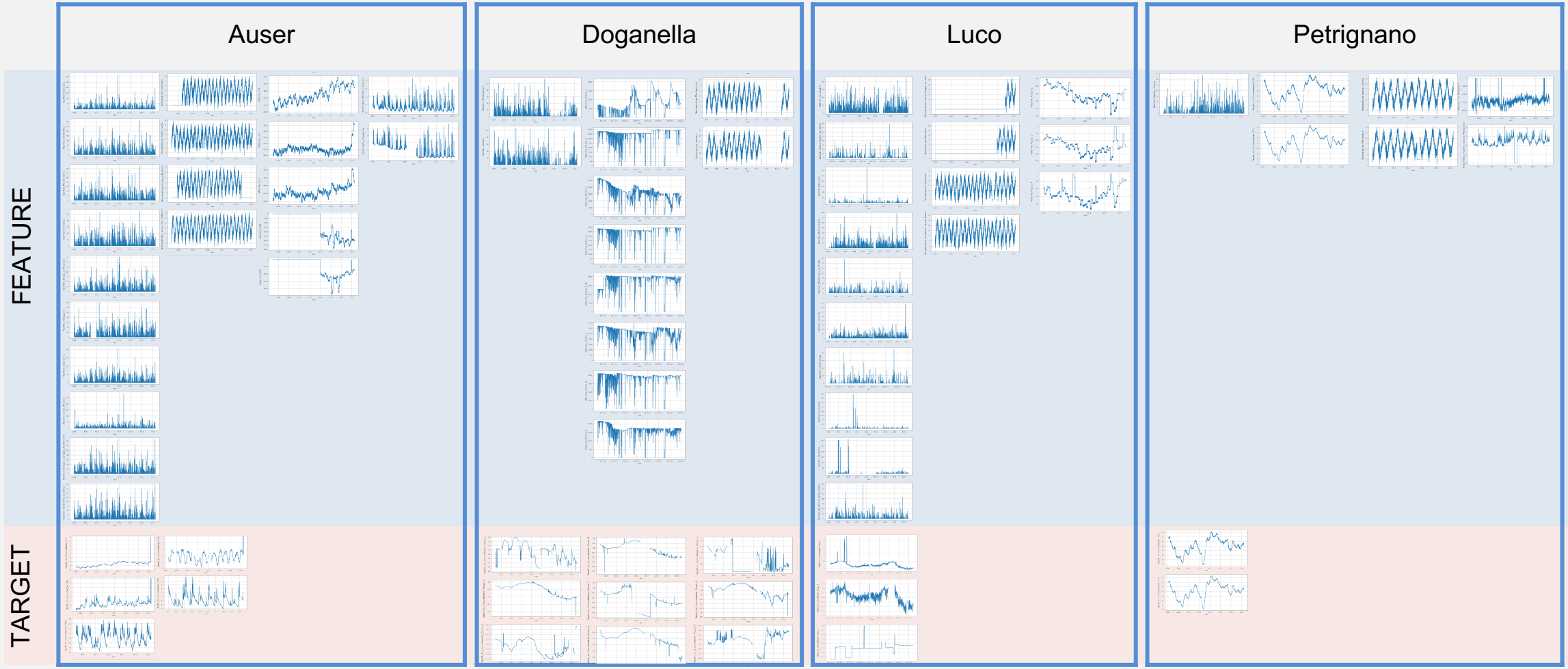
Histogram



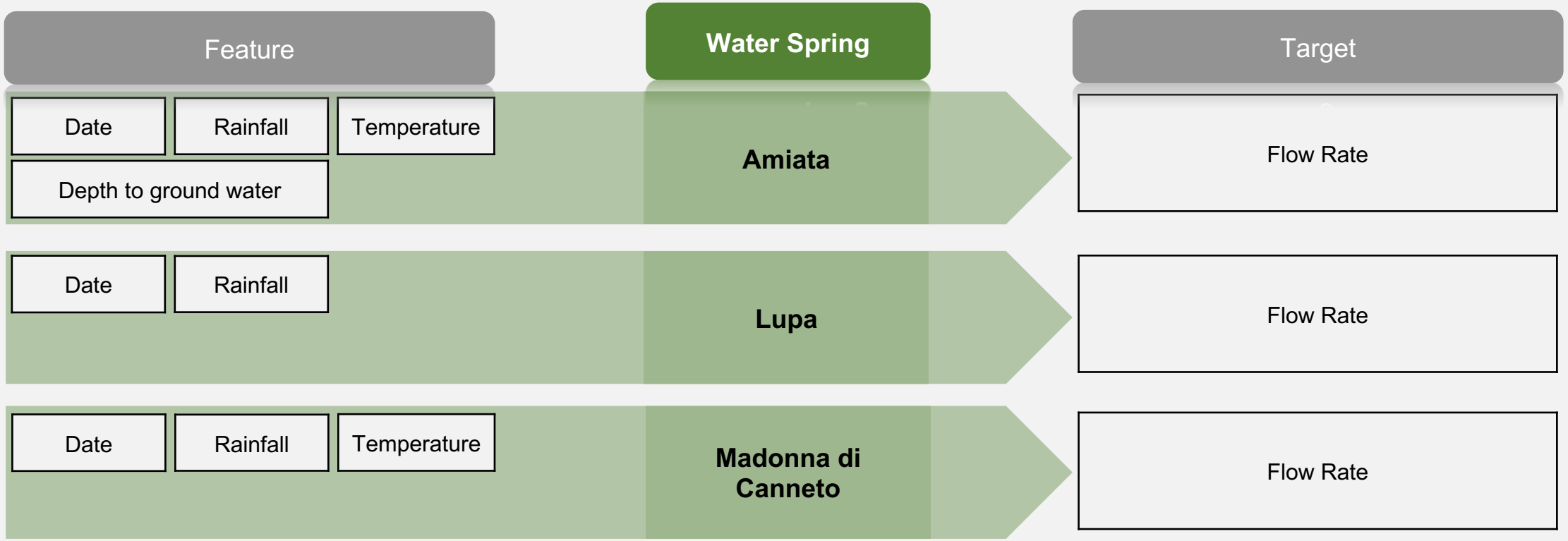
DATA DESCRIPTION

Seasonality and time-dependent patterns, it is a **time series dataset.**

Plotting

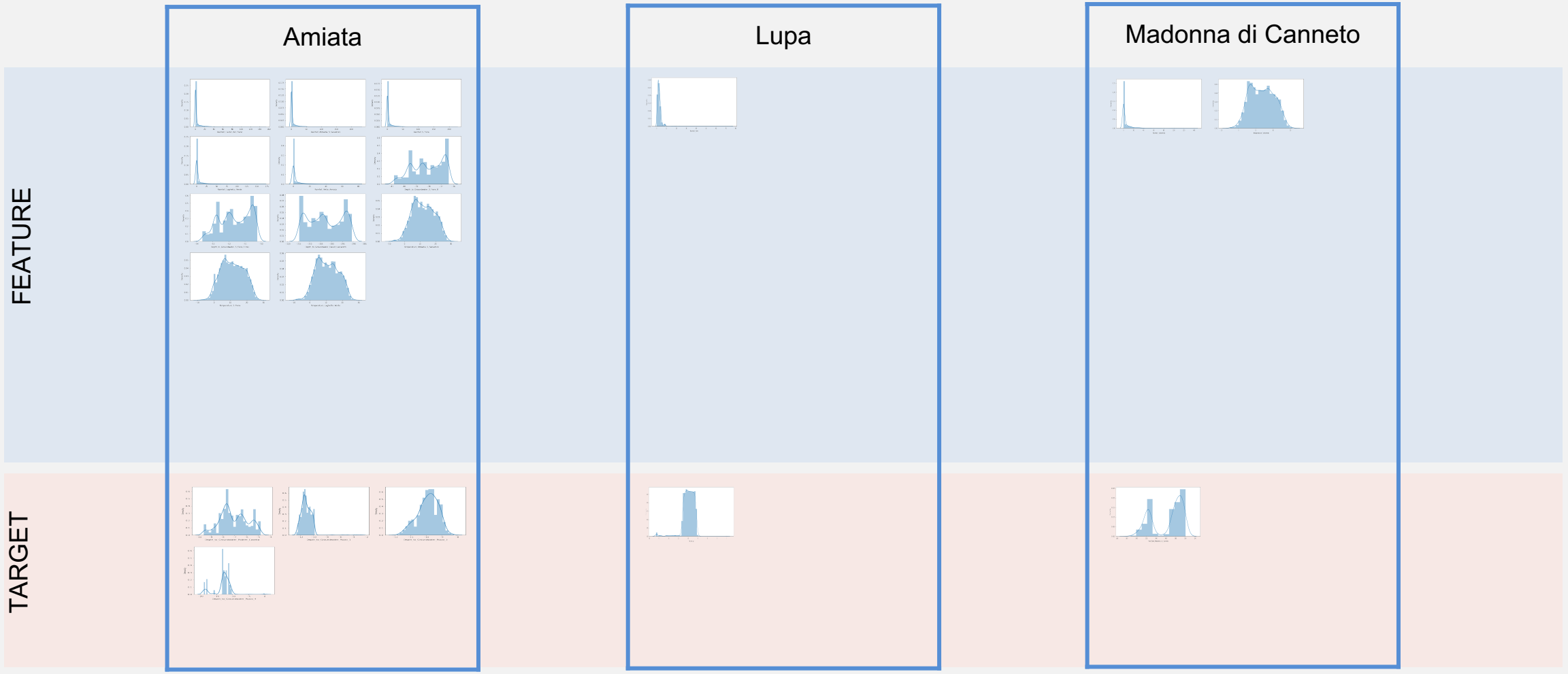


DATA DESCRIPTION



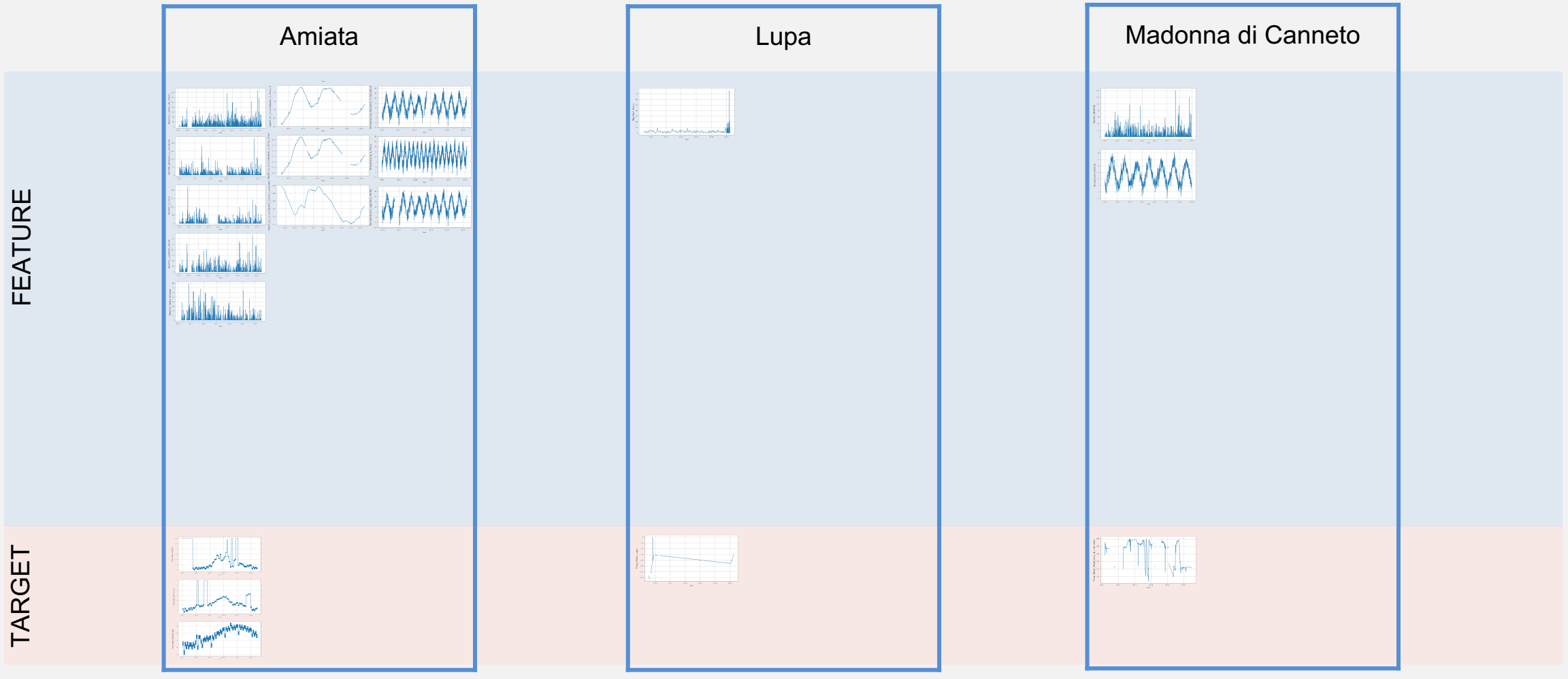
DATA DESCRIPTION

Histogram



DATA DESCRIPTION

Plotting



DATA DESCRIPTION

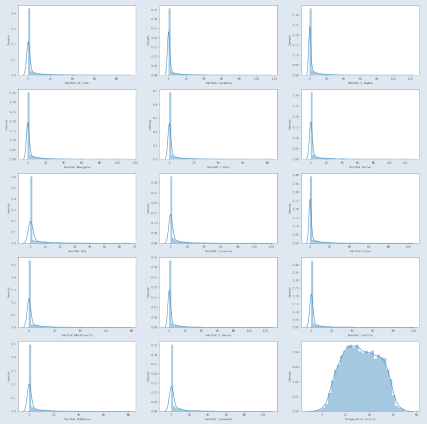
DATA DESCRIPTION

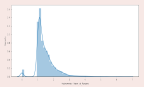
Histogram

FEATURE

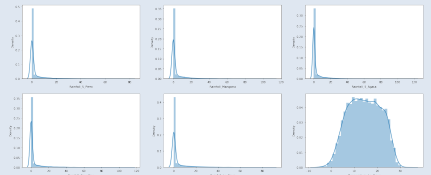
TARGET

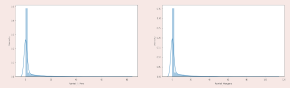
Arno





Bilancino





DATA DESCRIPTION

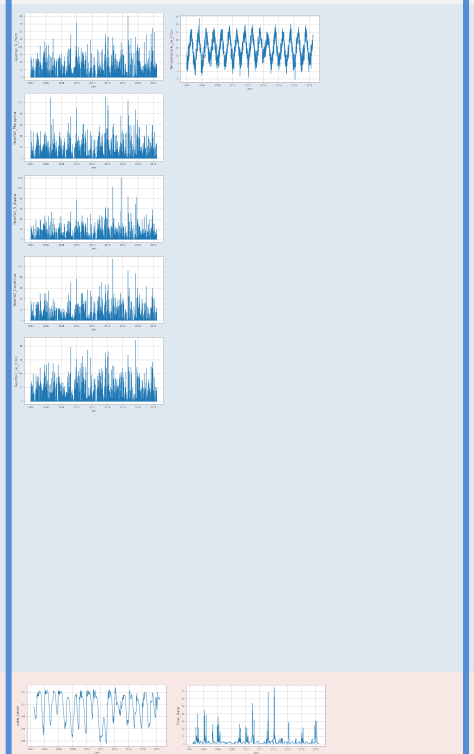
Plotting

FEATURE

TARGET

Arno

Bilancino



DATA DESCRIPTION

According to temperature and rainfall information, we see the relationship beween Lake Bilancino and River Arno

Geographical character

Lake Bilancino & River Arno

Vernio, Mangona, Cavallina, Le_Croci, S_Agata and S_Piero are to the north of Firenze, near to **Lake Bilancino**, these precipitations reach **the river Arno** through the River Sieve. The other points are located to the southeast of Firenze and fall directly into the Arno riverbed.

Relationship between the **lake Bilancino** and **the river Arno**. Because Lake Bilancino is an artificial lake made with a dam on the Sieve river. And the Sieve River is the most important tributary of the Arno river.

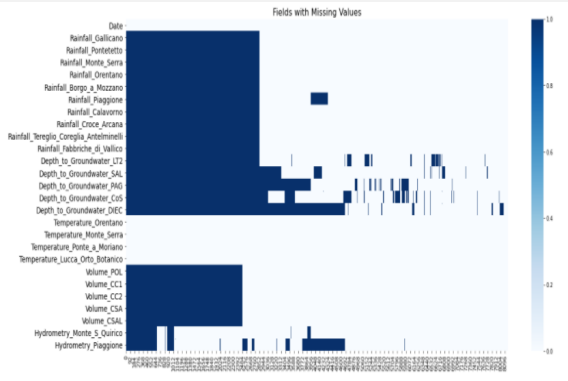
PART Ⅲ. DATA PROCESSING

HANDLING MISSING DATA

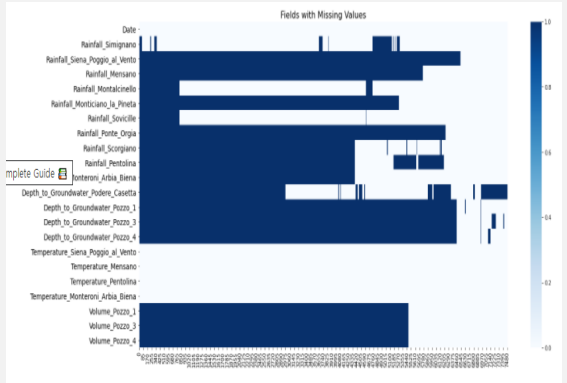
- Several data is missing from 2000 to 2014
- Missing data in particular year
- A lot of data with many discontinuities

AQUIFER

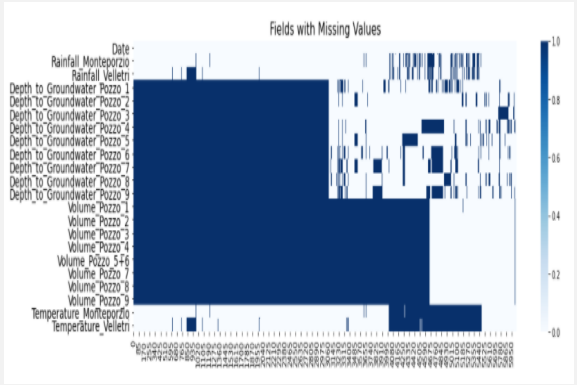
Auser



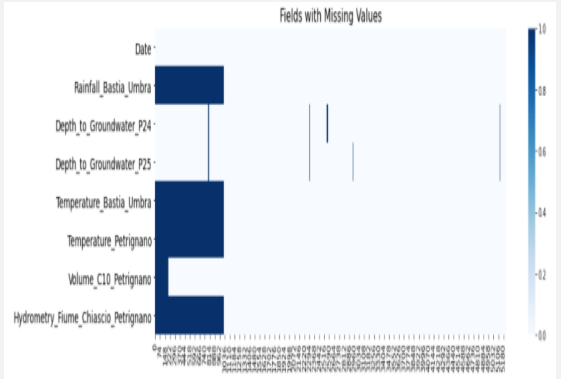
Luco



Doganella

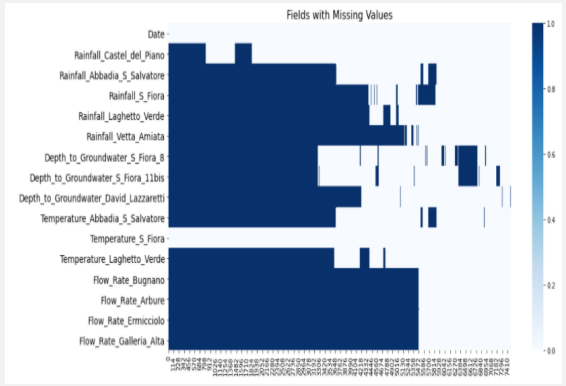


Petrignano

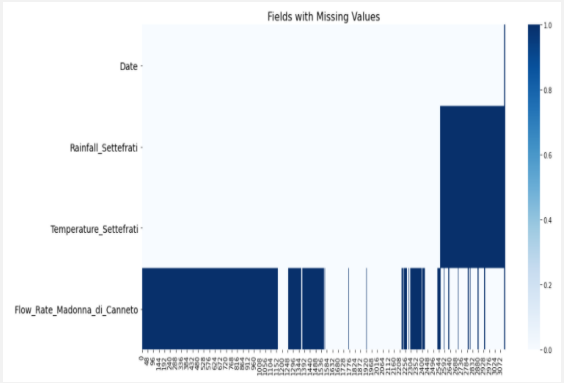


WATER SPRING

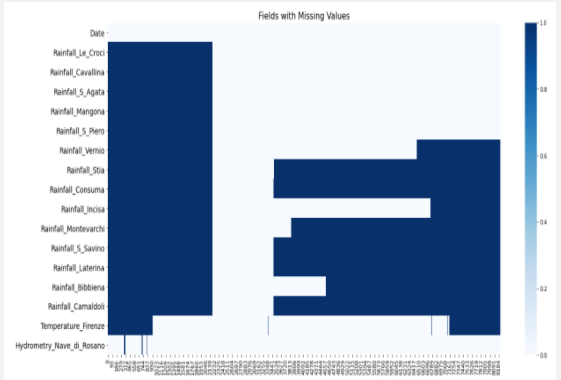
Amiata



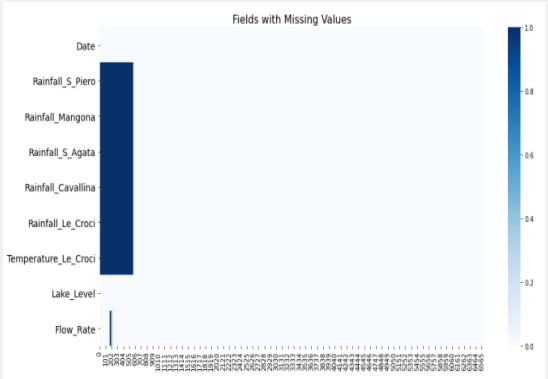
Madonna di Canneto



RIVER Arno



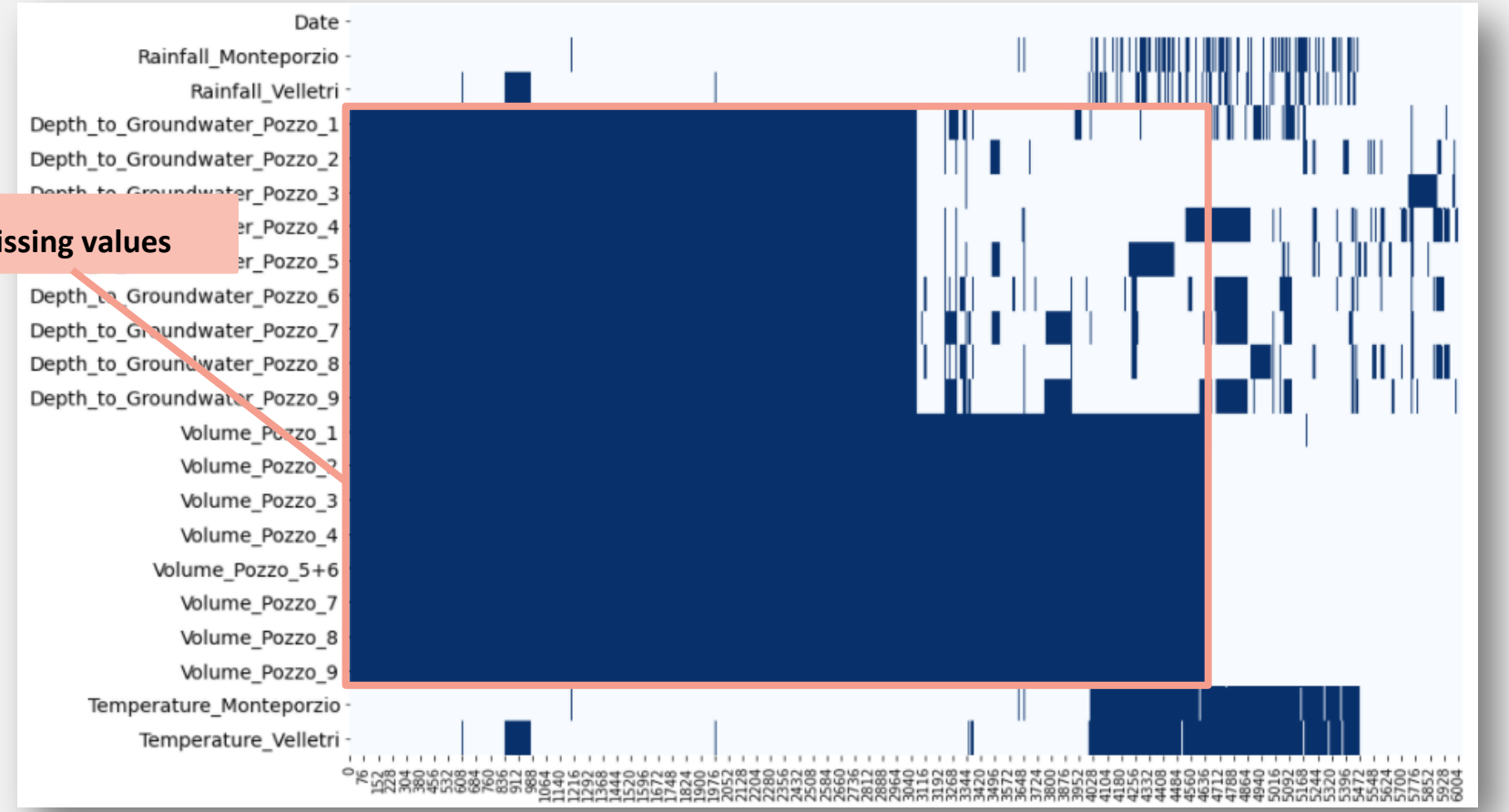
LAKE Bilancino



HANDLING MISSING DATA

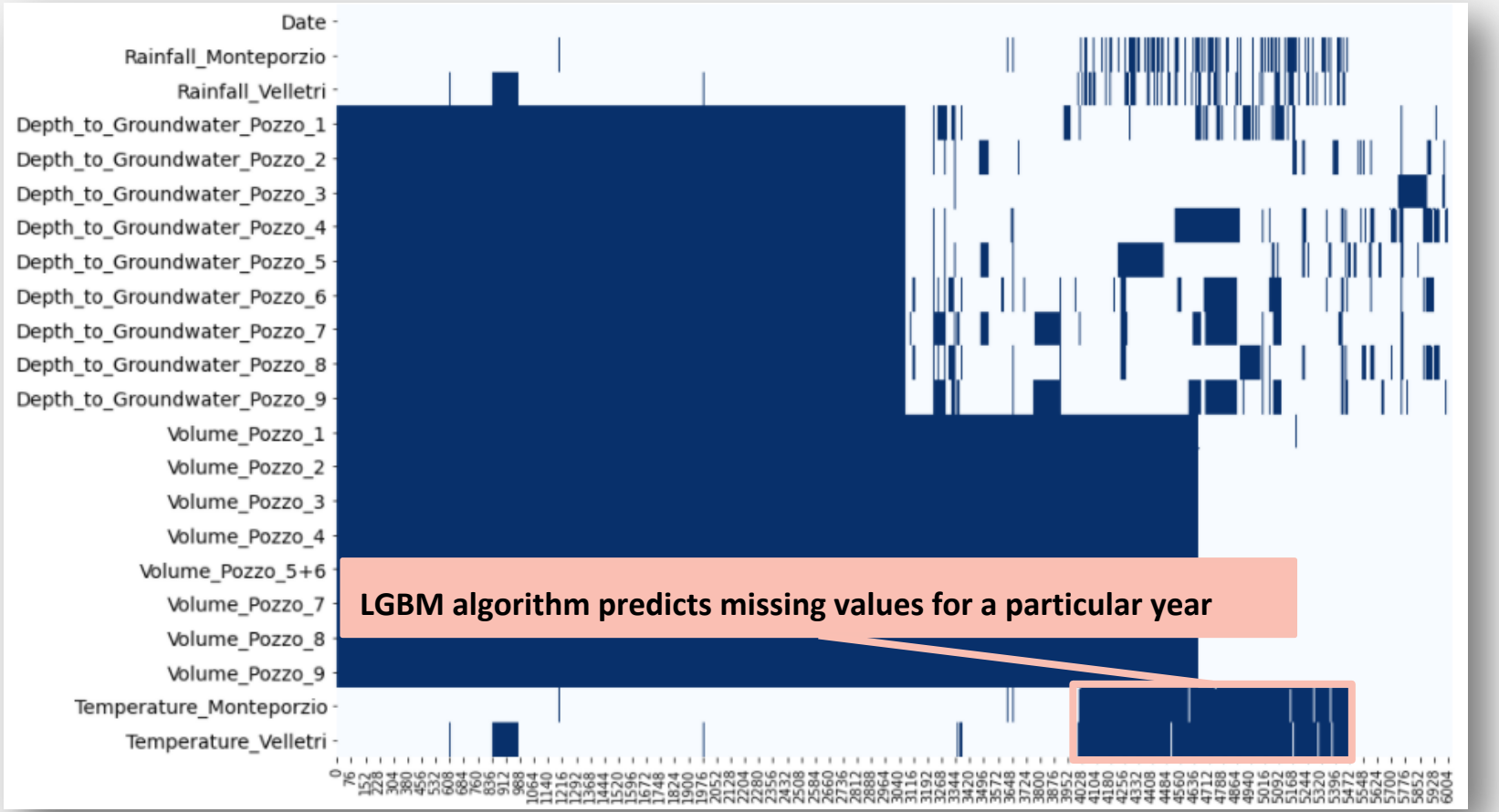
Handling long-term missing values

Remove long-term missing values



HANDLING MISSING DATA

Handling particular year



LGBM ?

L GBM

Light Gradient Boosting Machine

Gradient Boosting Framework with Tree-Based Learning Algorithms

How LightGBM works

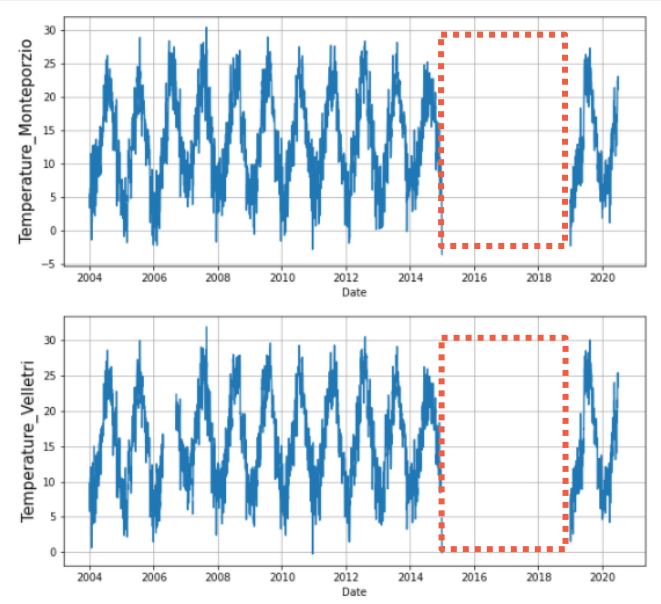
Leaf-wise tree growth

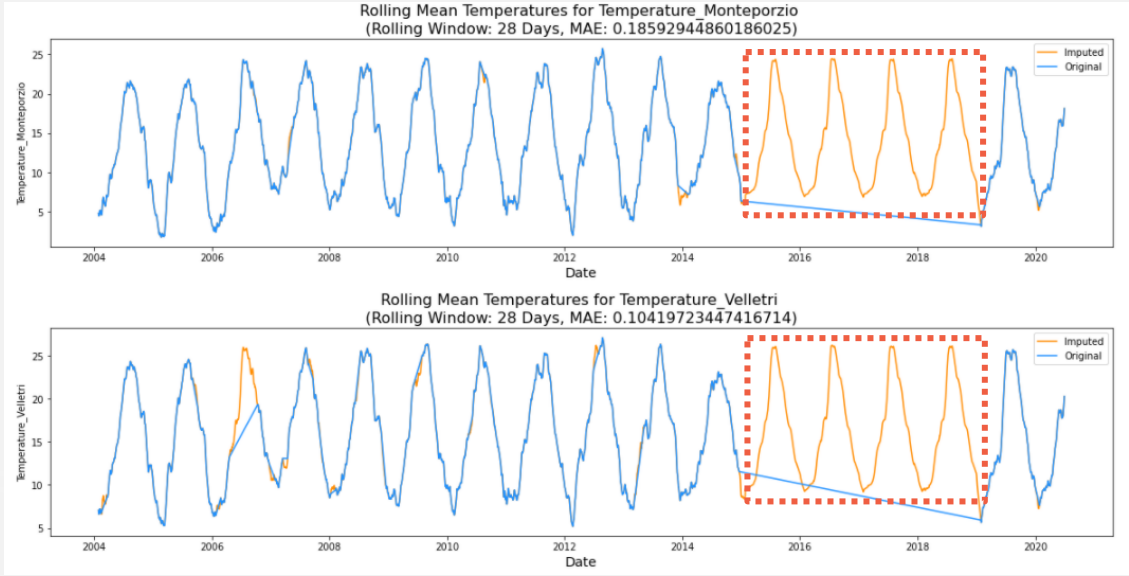
How other boosting algorithm works

Level-wise tree growth

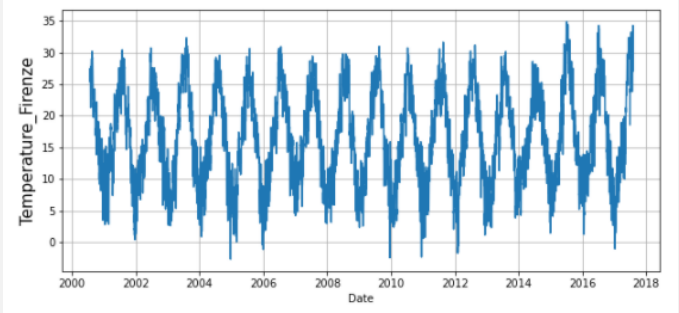
LGBM Algorithm?

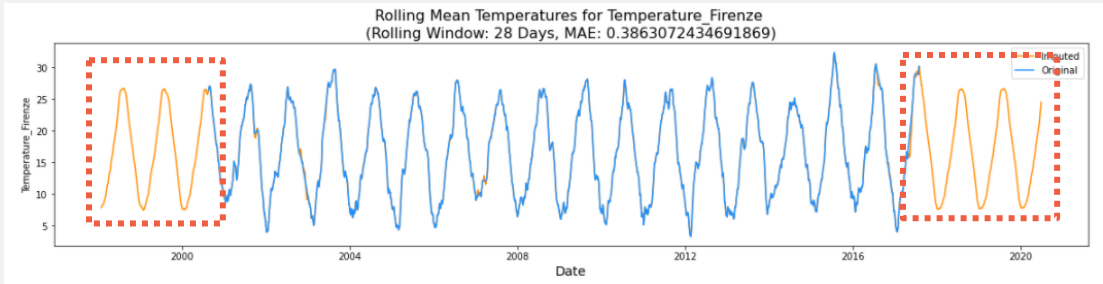
AQUIFER : *Doganella*





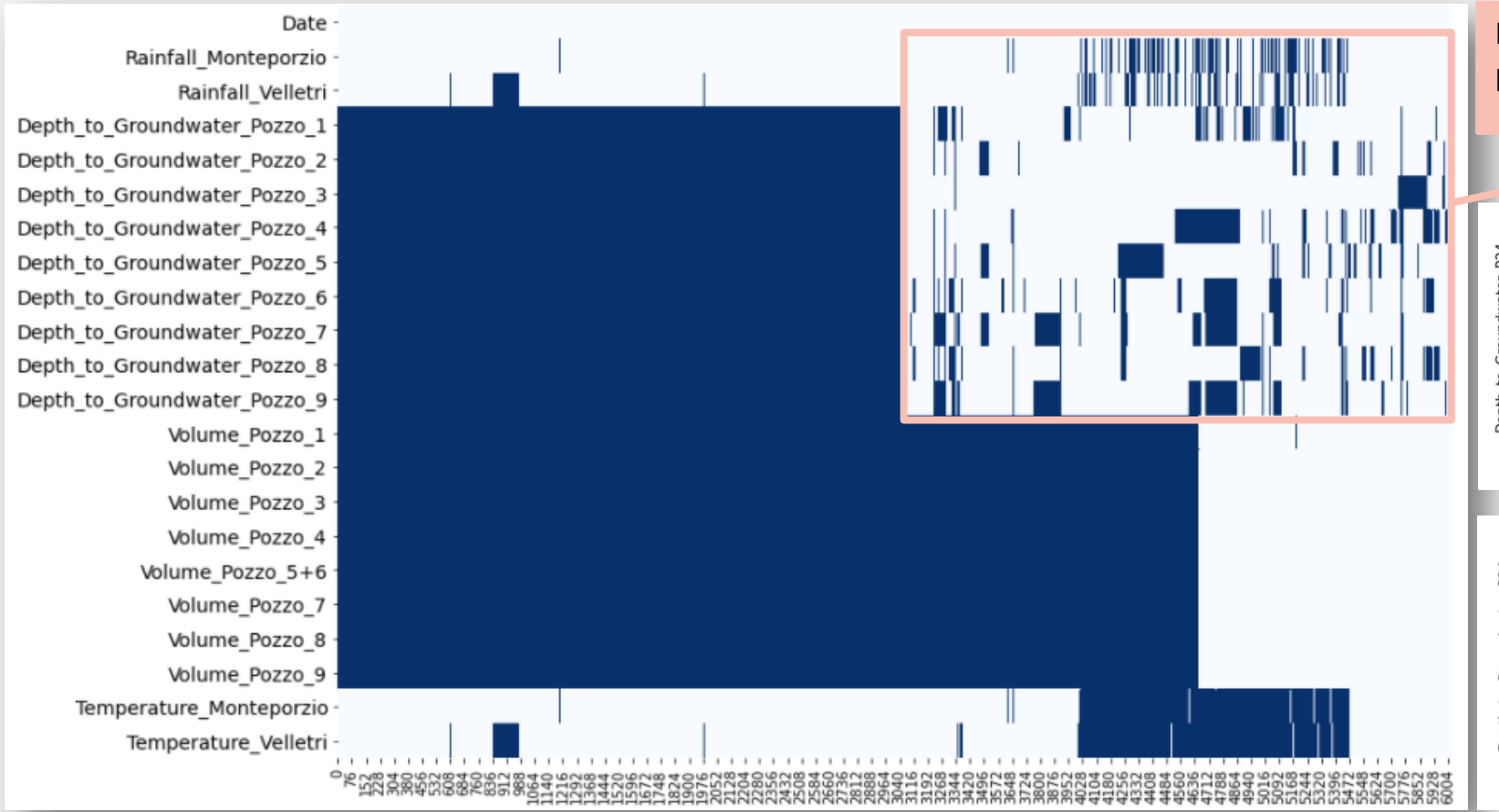
RIVER : *Arno*



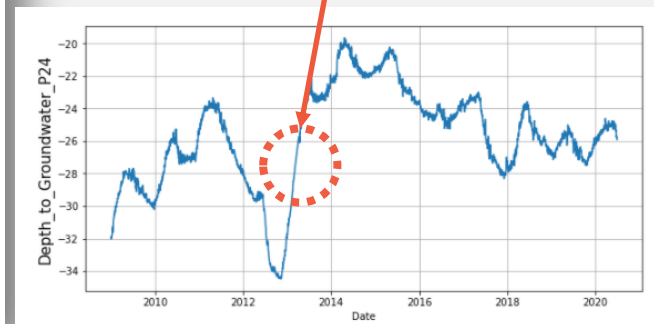


HANDLING MISSING DATA

Linear Interpolation A technique for estimating the value of a value between two values



Interpolating missing values in a linear proportion



TRANSFORMATION FEATURES

Take Absolute Value of Flow Rate

The minus sign (-) was meant to indicate the output water from the waterbody while the plus sign (+) was meant to indicate the input water to the waterbody. As not each company respected this convention, it is **necessary to consider only the absolute value.**

	Date	Flow_Rate_Bugnano	Flow_Rate_Arbure
0	2016-01-26	-0.327773	-2.872341
1	2016-01-27	-0.327773	-2.872341
2	2016-01-28	-0.327773	-2.872341
3	2016-01-29	-0.327773	-2.872341
4	2016-01-30	-0.327773	-2.872341

Unify Unit for Flow Rate

Unit of the lake's flow rate is m^3/s while the unit for aquifers and waters' flow rate is l/s. So, we transformed the lake's flow rate to l/s. (1000L in $1m^3$)

	Date	Flow_Rate_Bugnano	Flow_Rate_Arbure
0	2016-01-26	327.773153	2872.341311
1	2016-01-27	327.773153	2872.341311
2	2016-01-28	327.773153	2872.341311
3	2016-01-29	327.773153	2872.341311
4	2016-01-30	327.773153	2872.341311

Do not take Log transformation

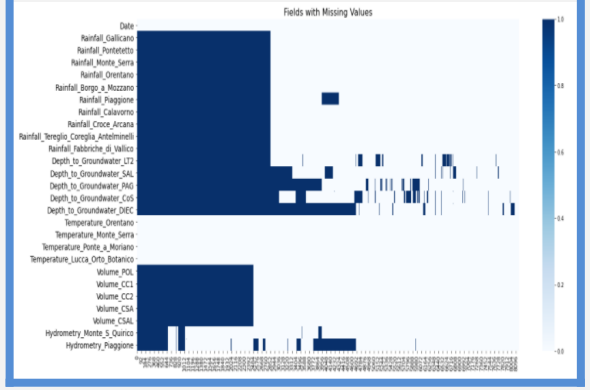
In this stage, it is not necessary.
There are variables **negative values**, therefore, it is **not possible to take log transformation**. And when building **deep learning model**, there is **a standardization process**.

SELECTION REGION

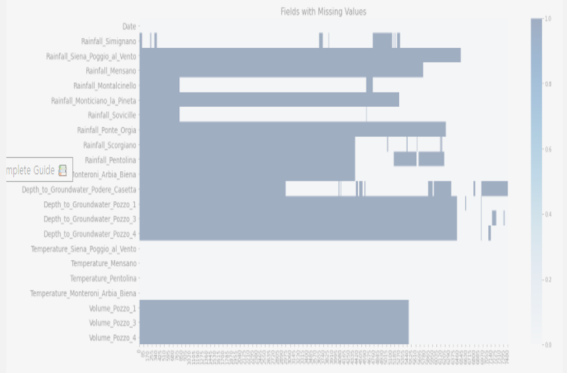
The region is selected based on the low missing values while having various feature variables

AQUIFER

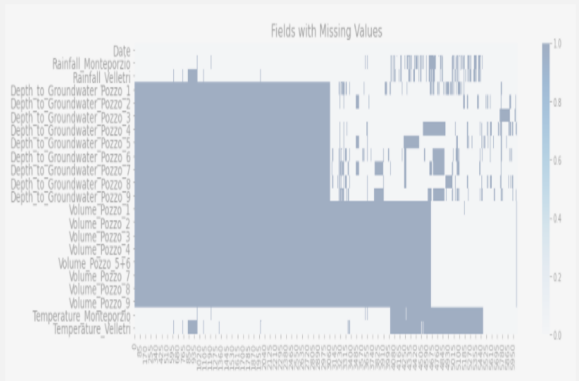
Auser



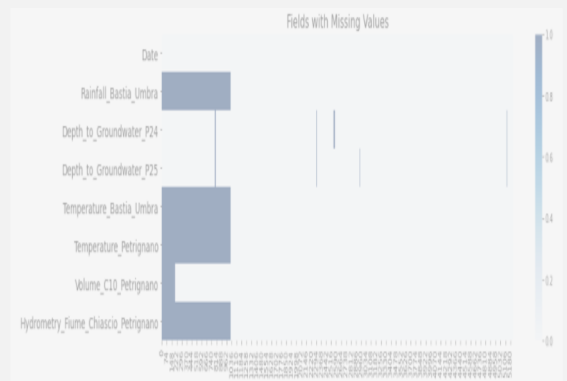
Luco



Doganella

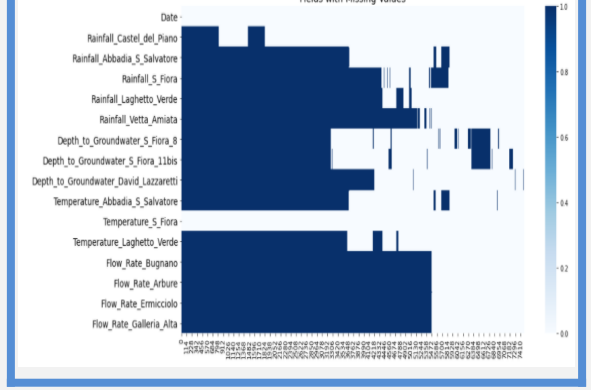


Petrignano

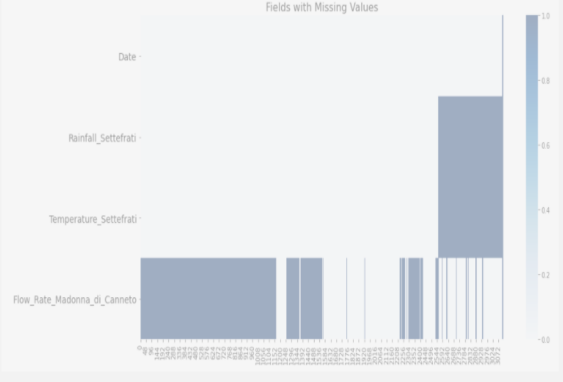


WATER SPRING

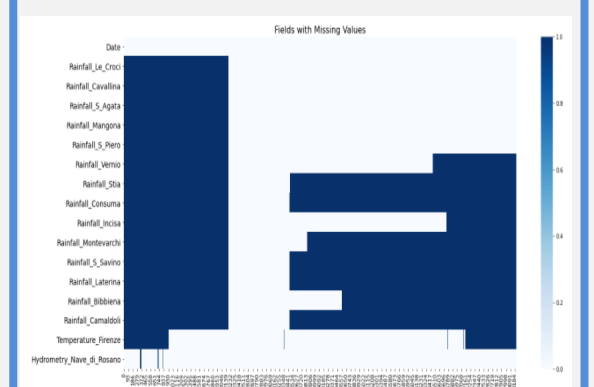
Amiata



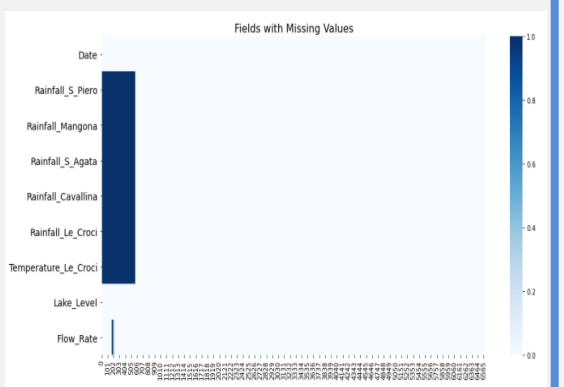
Madonna di Canneto



River Arno



LAKE Bilancino



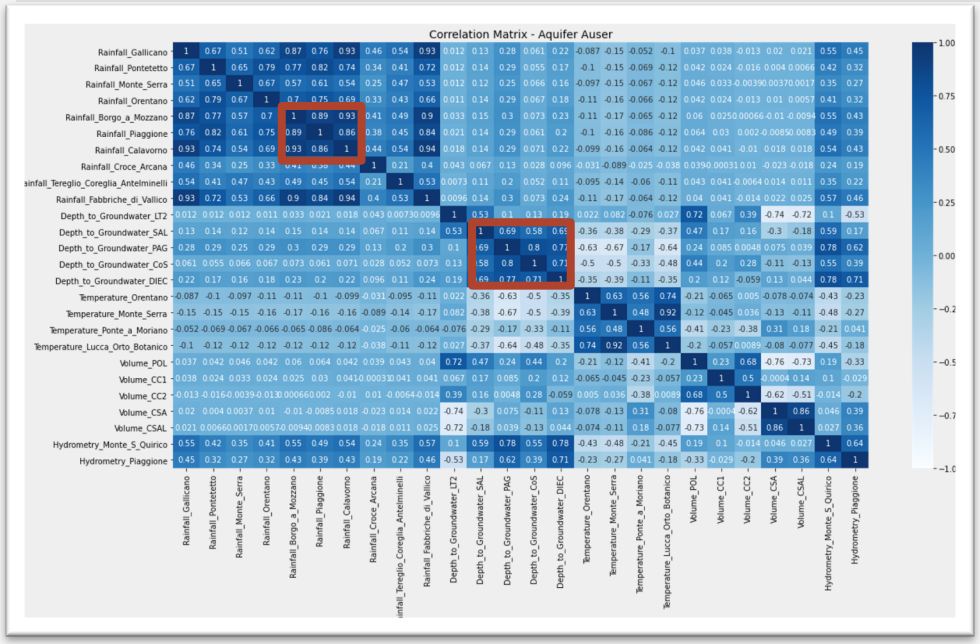
REDUCE FEATURES

Analyze correlation between variables with filled data (after handling missing data)

AQUIFER

Reduce all target variables to only a representative variable

Reduce 11 rainfall variables to 8 rainfall variables



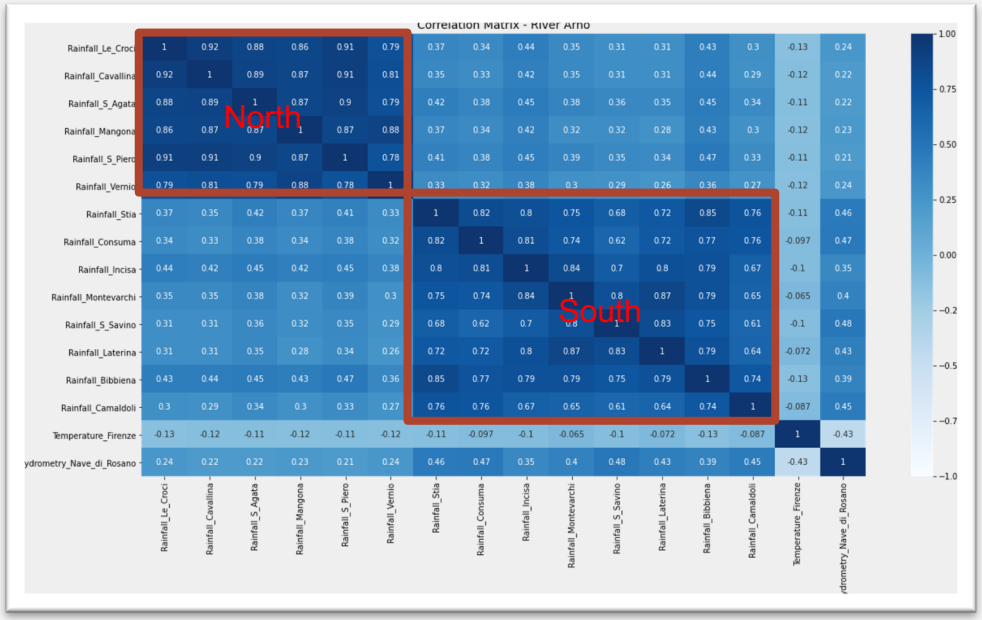
Water Spring

Reduce 3 type of feature variables to only 3 representative variables.

REDUCE FEATURES

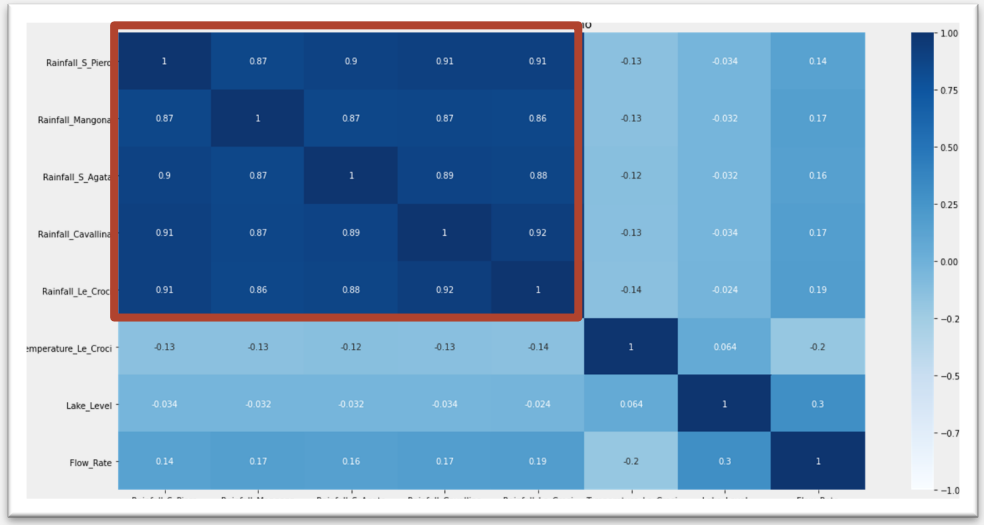
River

Two correlated groups of rainfall. with geograpical analysis, we have found why there are **two rainfall groups**.We chose **Rainfall_S_Piero** to represent precipitation in the **north** and **Rainfall_Incisa** for precipitation in the **south**.



Lake

In the case of **Lake Bilancino**, we select the **Rainfall_S_Piero** which has a strong correlation with the other precipitation variables.



SELECTION FEATURES

Use the target variable of Lake Bilancino as the feature variable of River Arno.

River

River Arno Model

Hydrometry of **River Arno**	=	Feature variable of **River Arno**	+	Target variable of **Lake Bilancino**

Earlier, in the analysis of geographical data features, we confirmed that there is **a geographical relationship between Lake Bilancino and River Arno**. And the Lake level of Bilancino is expected to have a direct effect on River Arno. Therefore, we will include **the target variable of Lake Bilancino** when we make a model of River Arno. We will use the feature variable of Lake Bilancino as **the feature variable of River Arno**.

PART IV. MODELS

Facebook Prophet

Characteristics

- 페이스북에서 만든 시계열 예측 라이브러리
- 통계적 지식이 없어도 직관적 파라미터를 통해 모형을 조정할 수 있음
- 일반적인 경우 기본값만을 사용해도 높은 성능을 보여줌

Model

- **Growth** : 특정 지점이 change point인지 여부를 확률적으로 결정
- **Seasonality** : 사용자들의 행동 양식으로 주기적으로 나타나는 패턴
- **Holidays** : 주기성을 가지진 않지만 전체 추이에 큰 영향을 주는 이벤트가 미치는 영향의 범위를 설정할 수 있음

Data Processing for Prophet modeling

Change column name

	Date	Depth_to_Groundwater_SAL
0	2014-03-20	-4.991429
1	2014-03-27	-5.045714
2	2014-04-03	-5.158571
3	2014-04-10	
4	2014-04-17	

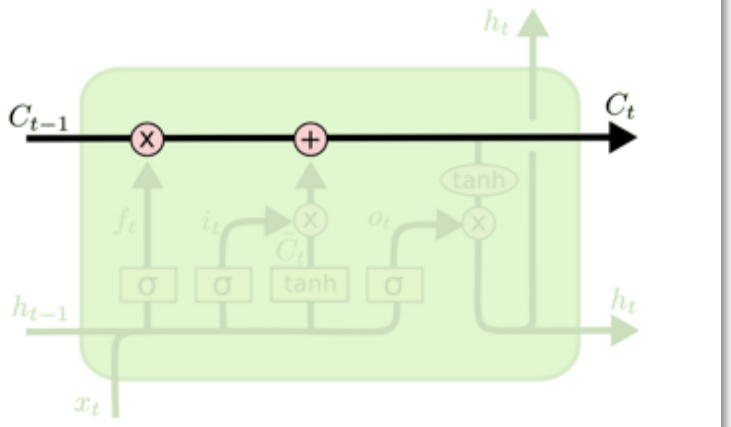
	ds	y
0	2014-03-20	-4.991429
1	2014-03-27	-5.045714
2	2014-04-03	-5.158571
3	2014-04-10	-5.300000
4	2014-04-17	-5.328571

Division Dataset

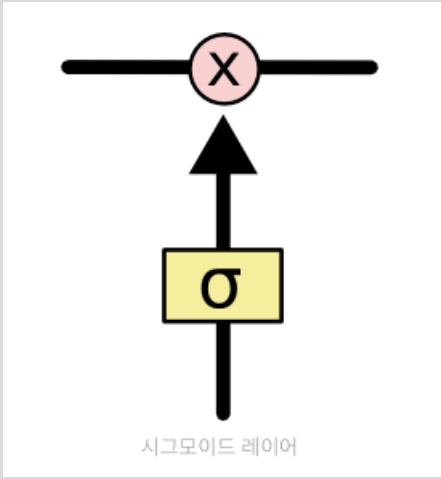
Train data	Test data
(All days – 365)	(365 days)

LSTM-Basic theory

LSTM의 핵심은 **셀 스테이트**입니다.
아래 다이어그램 상단에 있는 수평선입니다.



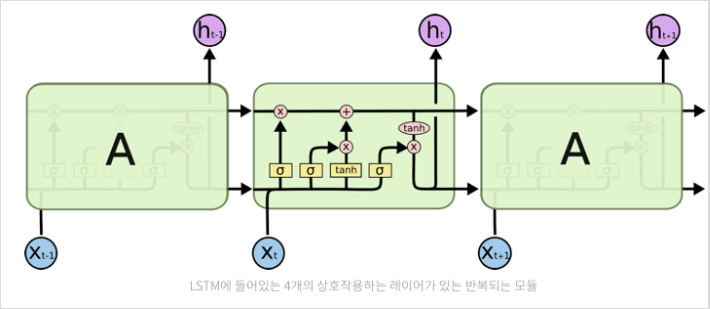
셀 스테이트는 아주 마이너한 선형 연산을 거치고 전체 체인을 관통합니다. 이 구조로 인해 정보는 큰 변함없이 계속적으로 다음 단계에 전달되게 됩니다.



시그모이드 레이어

게이트들은 선택적으로 정보들이 흘러들어 갈 수 있도록 만드는 장치이며, 시그모이드 뉴럴 넷과 점단위 곱하기 연산으로 이루어져 있습니다.

시그모이드 레이는 0 혹은 1의 값을 출력합니다. 그리고, 각 구성요소가 얼마만큼의 영향을 주게 될지를 결정해주는 역할을 합니다. 0이라는 값을 가지게 된다면, 미래의 결과에 영향을 주지 않게 만듭니다. 반면 1이라는 값은 해당 구성요소가 확실히 미래의 예측결과에 영향을 주도록 데이터가 흘러가게 만듭니다.

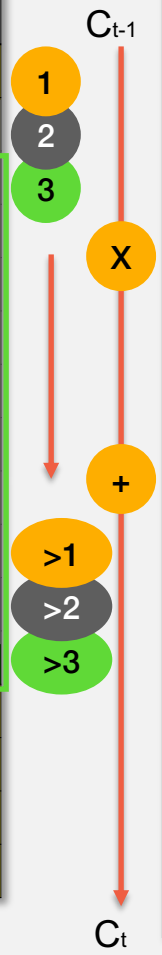


LSTM에 들어있는 4개의 상호작용하는 레이어가 있는 반복되는 모듈

LSTM-dataSet

Date	Rainfall_Gallicano	Rainfall_Pontetetto	Rainfall_Monte_Serra	Depth_to_Groundwater_LT2
04/01/2006	0	0.2	0	-15.02
05/01/2006	0	0	0	-15.02
06/01/2006	0	0	0	-15.04
07/01/2006	0	0	0	-15.06
08/01/2006	0	0.2	0	-15.07
10/01/2006	0	0	0	-15.08
11/01/2006	0	0	0	-15.08
12/01/2006	0	0	0	-15.07
14/01/2006	0	0	0	-15.06
15/01/2006	0	0.2	0	-15.05
16/01/2006	0	0	0.2	-15.04
20/01/2006	0	0	0.2	-14.96
22/01/2006	0	0.4	0.2	-14.9
23/01/2006	0	0	0	-14.93
24/01/2006	0	0	0	-14.97
25/01/2006	0	0	0	-14.95

{ X } { y }




```
def split_dataSet(data,n_steps_in,n_steps_out):
    X,y = list(),list()
    for i in range(len(data)):
        end_ix = i + n_steps_in
        out_end_ix = end_ix + n_steps_out-1
        if out_end_ix > len(data):
            break
        data_x,data_y = data[i:end_ix, 1:],data[end_ix-1:out_end_ix,0]
        X.append(data_x)
        y.append(data_y)
    return np.array(X), np.array(y)
```

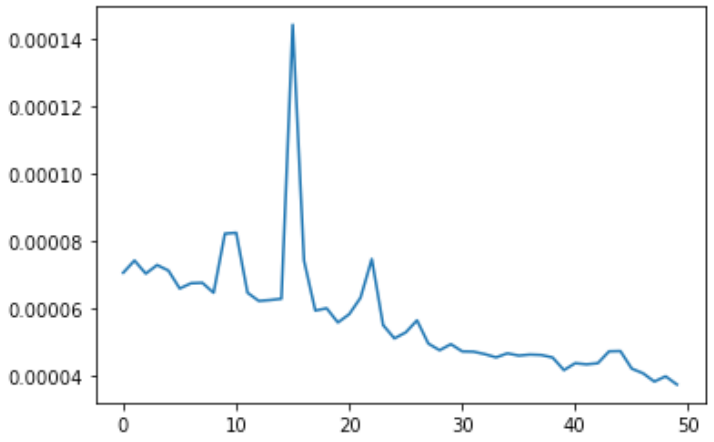
LSTM의 핵심은
셀 스테이트

셀 스테이트는
데이타 체인을 통
과

데이타 체인을 위
한
Data set 함수

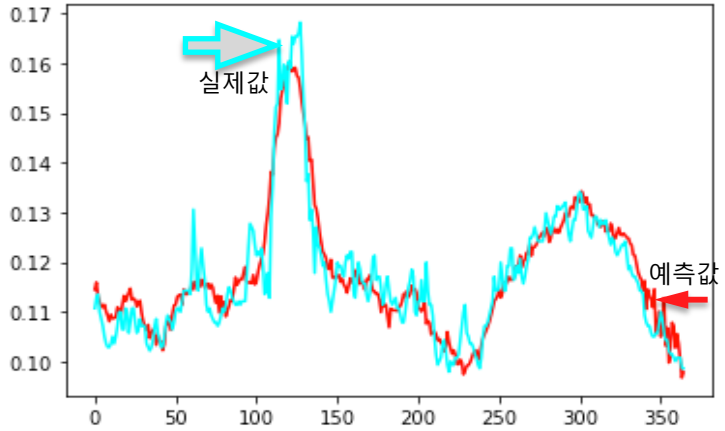
LSTM- Basic setting

epoch = 50 에서 **LOSS RATE** 는 이미 충분히 낮은 값을 보여줌



Multivariate Multi-Step LSTM Models
Optimizer: Adam
중간 LSTM:64

2년치의 dataSet으로 다가올 1년치를 예상



1. Univariate LSTM Models
   1. Vanilla LSTM
   2. Stacked LSTM
   3. Bidirectional LSTM
   4. CNN LSTM
   5. ConvLSTM
2. Multivariate LSTM Models
3. Multi-Step LSTM Models
   1. Vector Output Model
   2. Encoder-Decoder Model
4. Multivariate Multi-Step LSTM Models
   1. Multiple Input Multi-Step Output.
   2. Multiple Parallel Input and Multi-Step Output.

PART V . MODEL EVALUATION

PROPHET & LSTM

AQUIFER

Feature

Rain	Gallicano
	Pontetetto
	Monte_Serra
	Orentano
	Borgo_a_Mozzano
	Croce_Arcana
	Tereglio_Coreglia_Antelminelli
	Fabbriche_di_Vallico
Temperature	Orentano
	Monte_Serra
	Ponte_a_Moriano
	Lucca_Orto_Botanico
Voulume	POL
	CC1
	CC2
	CSA
Hydrometry	Monte_S_Quirico
	Piaggione

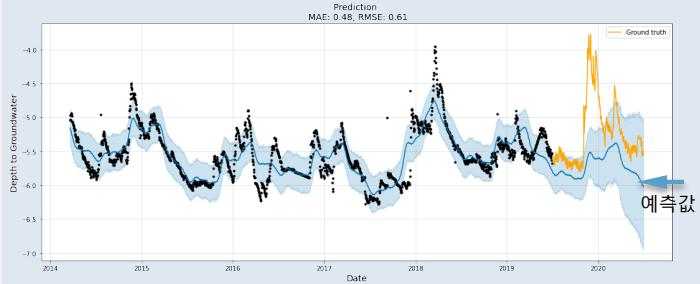
Target

Depth	to_Groundwater_SA

PROPHET

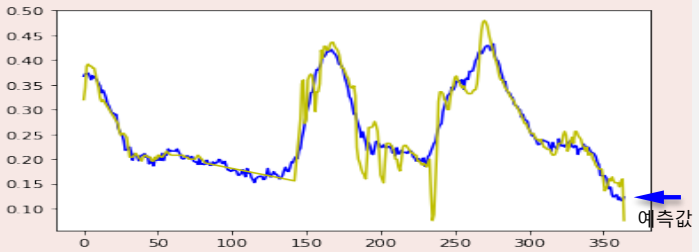
Depth

MAE : 0.48 / RMSE : 0.61



LSTM

MAE : 0.02 / RMSE : 0.04



PROPHET & LSTM

WATER SPRING

Feature

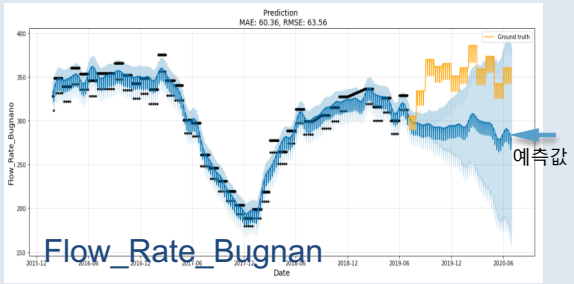
Rainfall	S_Fiora
Depth	to_Groundwater_S_Fiora_8
Temperature	S_Fiora

Target

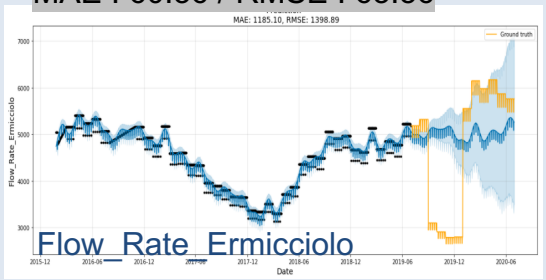
Flow Rate	Bugnano
	Ermicciolo
	Galleria_Alta

Flow Rate

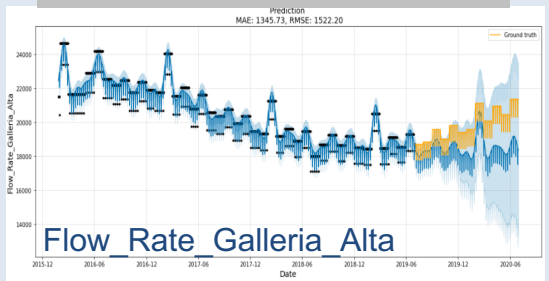
PROPHET



MAE : 60.36 / RMSE : 63.56

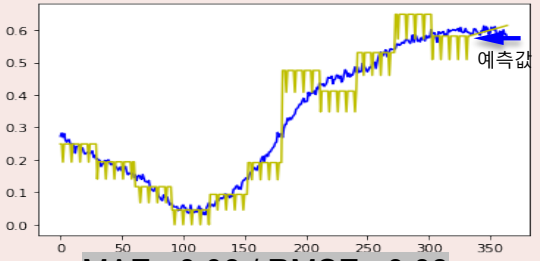


MAE : 1185.10 / RMSE : 1389.89

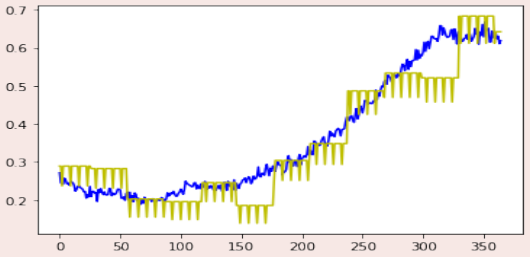


MAE : 1345.73 / RMSE : 1522.20

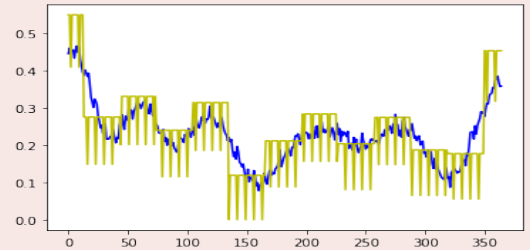
LSTM



MAE : 0.06 / RMSE : 0.08



MAE : 0.07 / RMSE : 0.09



MAE : 0.07 / RMSE : 0.9

PROPHET & LSTM

RIVER

Feature

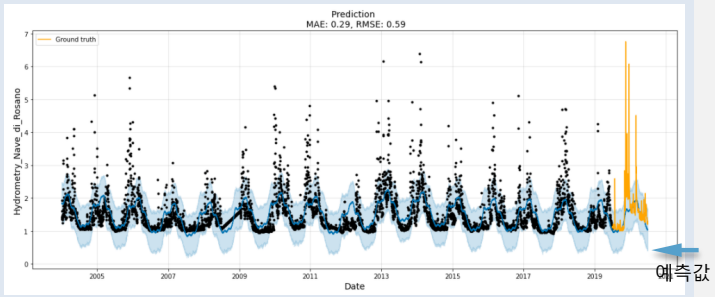
Rainfall	S_Piero
Temperature	Firenze
Lake Level	Bilancino
Flow Rate	Bilancino

Target

Hydrometry	Nave_di_Rosano

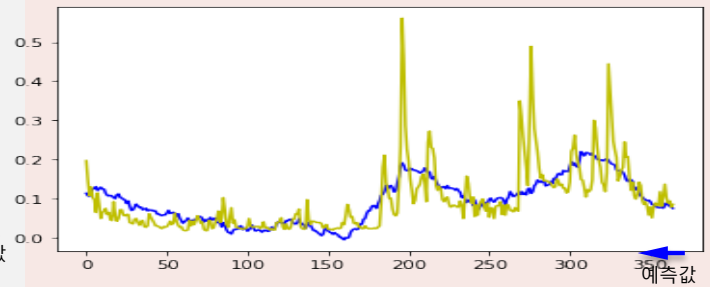
Hydrometry

PROPHET



MAE : 0.29 / RMSE : 0.59

LSTM



MAE : 0.04 / RMSE : 0.07

PROPHET & LSTM

LAKE

Feature

Rainfall	S_Piero
	Le_Croci
Temperature	Le_Crocia

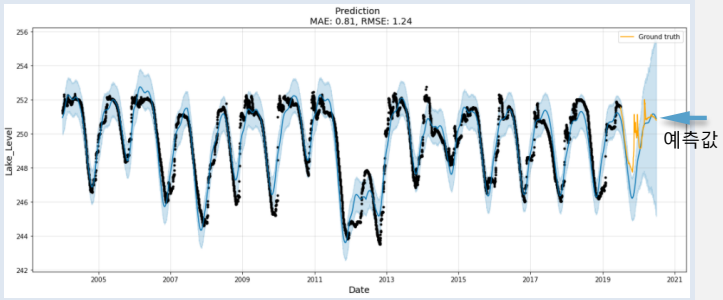
Target

Lake_Level	Bilancino
Flow_Rate	Bilancino

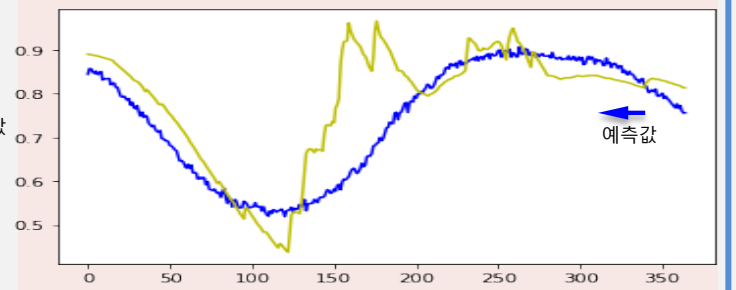
PROPHET

LSTM

Lake Level

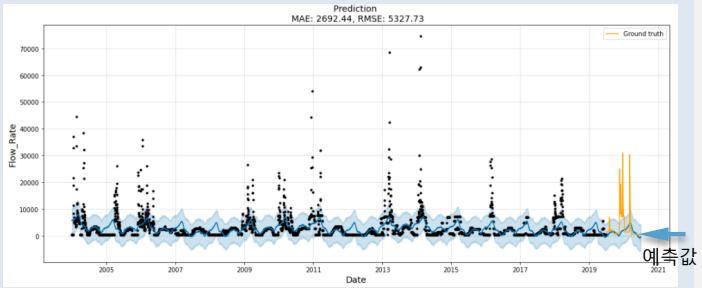


MAE : 0.18 / RMSE : 1.24

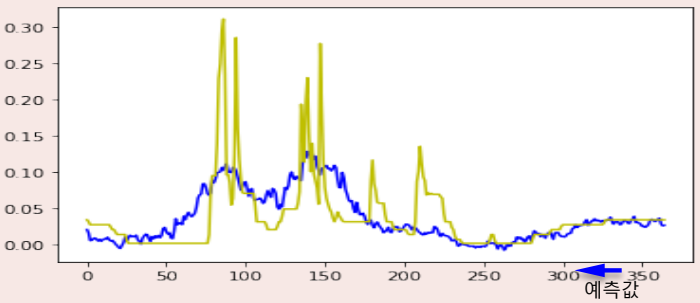


MAE : 0.14 / RMSE : 0.20

Flow Rate



MAE : 2692.44 / RMSE : 5327.73

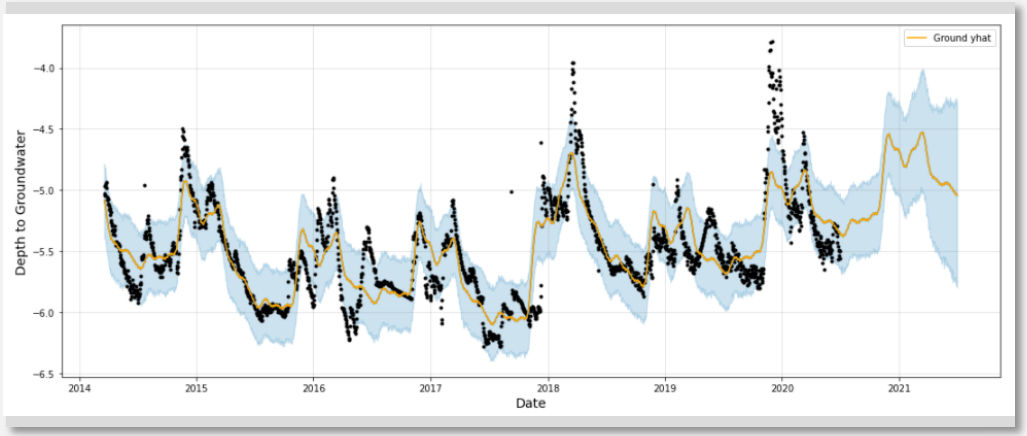


MAE : 0.02 / RMSE : 0.05

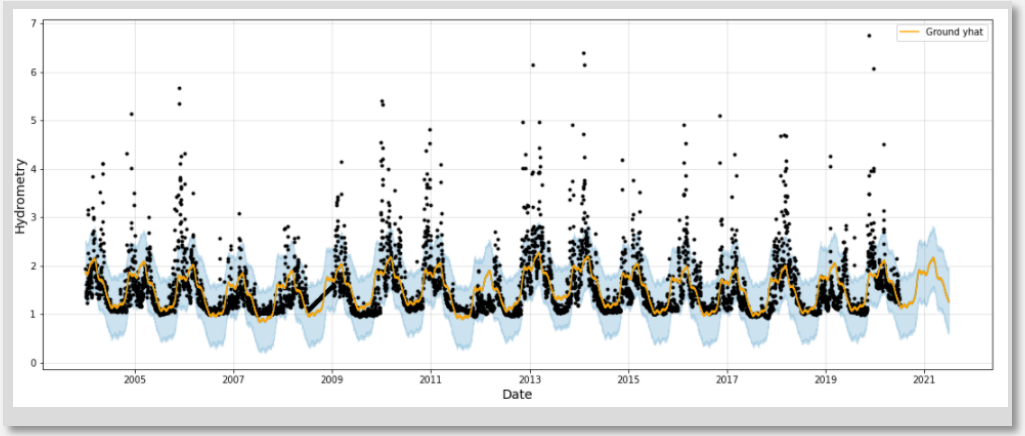
PART VI . PREDICTION

1 YEAR PREDICTION

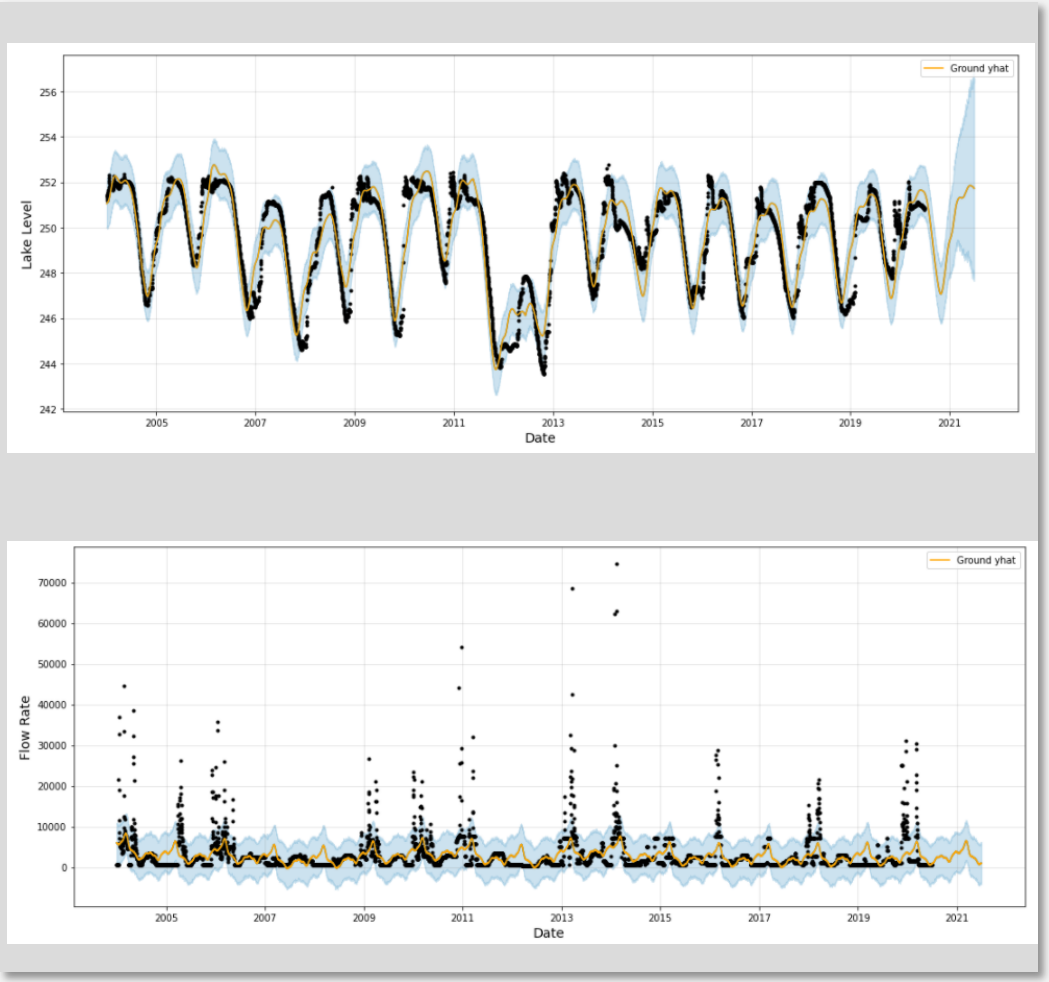
AQUIFER



RIVER

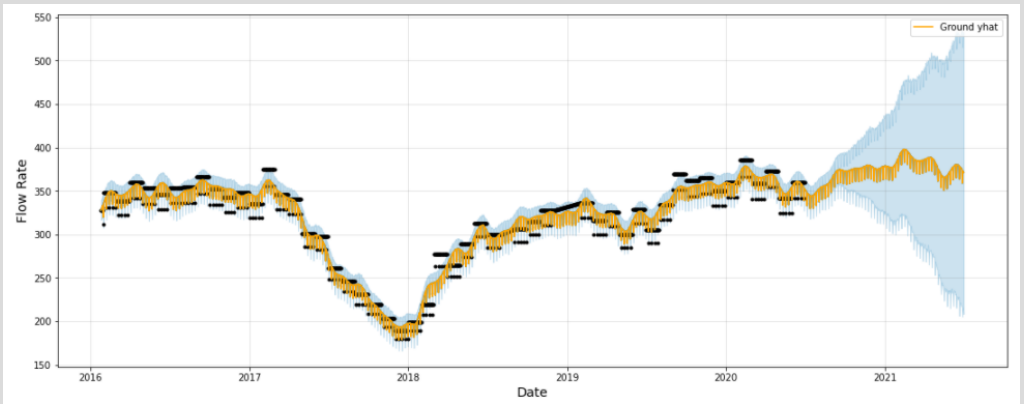


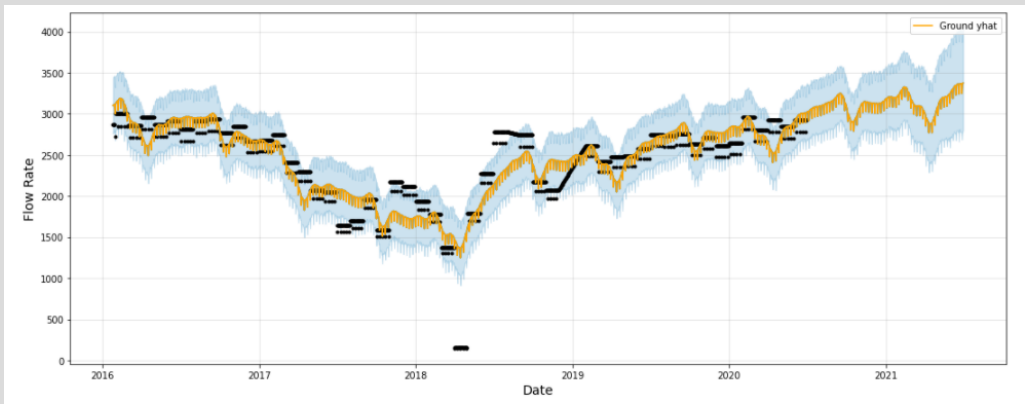
LAKE

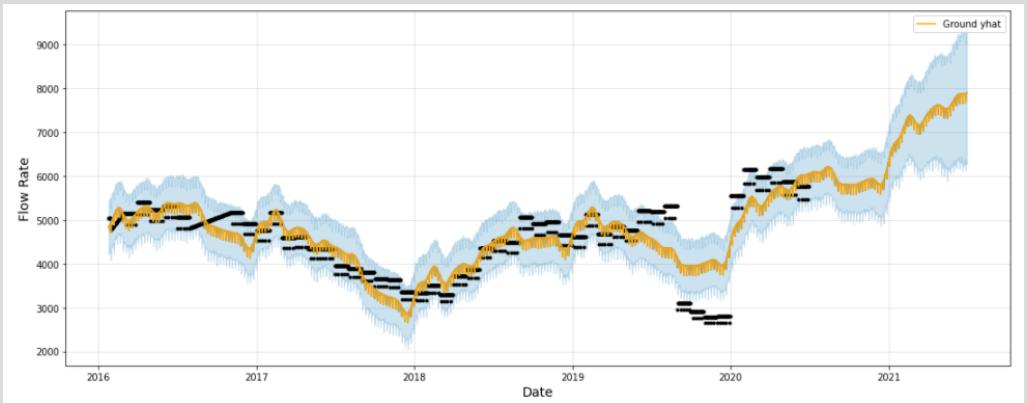


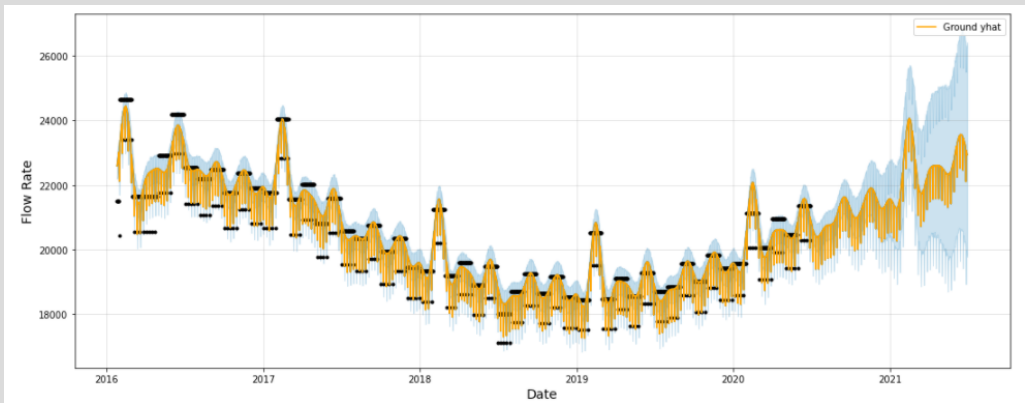
1 YEAR PREDICTION

WATER SPRING









PART Ⅶ. EVALUATION

SUMMARY

Limits

Handling missing data

- 실제 기온데이터를 수집하여 사용하지 않고 통계적 예측값을 사용

Deficient point of Prophet Model

- 갑작스런 데이터의 변동이 있을 경우 정확도가 떨어짐
- 과거의 장기적인 추세를 보고 이후의 대략적인 추세를 확인하는 용도로 제한적 사용

Abour LSTM Model

- Prophet에 비해 정확도가 높음
- 필요에 따라 어느정도 셋팅이 가능하기 때문에 예측의 개선여지가 있음

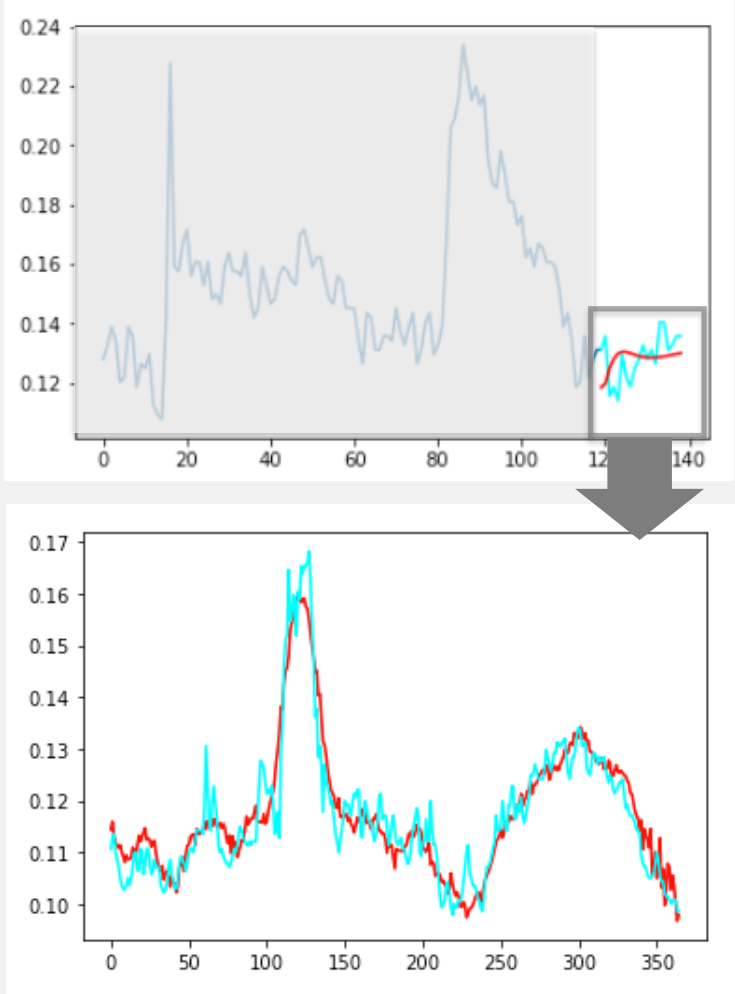
GENERAL REVIEW

What we learned

- Kaggle을 통한 실전 데이터에 대한 경험
- 데이터 전처리에 대한 중요성 인지
- 통계적 방법을 통한 시계열 데이타 분석에 대한 기본 이해
- LSTM 및 Prophet을 적용한 모델을 이용해 시계열 예측

Unsatisfied things

- 실 프로젝트 기간 : 10일
- ITALY 지역에 대한 정보 부족
- 지구 과학적 이론에 대한 충분한 논문 검토 부족



PART Ⅷ . REFERENCES

andreshg. (2020). Time Series Analysis A Complete Guide [Source code].
https://www.kaggle.com/andreshg/timeseries-analysis-a-complete-guide

iamleonie. (2020). EDA: Quenching the Thirst for Insights [Source Code].
https://www.kaggle.com/iamleonie/eda-quenching-the-thirst-for-insights

luca31394. (2020). Acea - EDA and Data Cleaning: Deep water [Source code].
https://www.kaggle.com/luca31394/acea-eda-and-data-cleaning-deep-water/output

동건. (2018, April 10). Long Short-Term Memory (LSTM) 이해하기. *개발새발로그.*
https://dgkim5360.tistory.com/entry/understanding-long-short-term-memory-lstm-kr

박창희, 정일문. (2020). LSTM 모형을 이용한 지하수위 예측 평가. *Journal of Korea Water Resources Association*, *53*(4), 273–283.
https://doi.org/10.3741/JKWRA.2020.53.4.273

AndresHG.(2020). TimeSeries Analysis A Complete Guide[Basic notion]
https://www.kaggle.com/andreshg/timeseries-analysis-a-complete-guide

Jason Brownlee, Phd(2020.11.28) - Machine learning Mastery webpage[Source code]
https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting

Marcomarchetti(2020) Acea Smart Water (Baseline Models Comparison)
https://www.kaggle.com/marcomarchetti/acea-smart-water-baseline-models-comparison