Start-end date: **06-02-2023 until 30-06-2023**	HZ	UNIVERSIDADE FEDERAL DE MINAS GERAIS INCIPIT VITA NOVA 7 DE SETEMBRO DE 1927
Student name: **Laura Batista Cwienk**		
Course: **Chemical Engineering**		
Internship Department/Company: **Water Technology**		
Brazilian Professor/Supervisor: **Érika Cren**		
Dutch Professor/Supervisor: **Iarima Mendonça**		
Project : **Digital Twin**		

Problem/assignment

My research assignment was guided through the following research questions :

Primary research question :

How to assembly a data-driven model of a wastewater drainage system for predicting the arriving flow in a wastewater treatment plant – taking EDE wastewater treatment plant as case study?

Secondary research question :

- *What are the most important features (independent variables) for predicting the water flow at pumping and treatment stations*?
- *What are the impacts of precipitation on Ede WWTP's transport system*? *How does precipitation affect water flow* ?
- *Which machine learning algorithm is the most suitable for predicting the flow at the pumping and treatment stations*?

Used methods/project phases

For the construction of the digital model, the research was divided in the following phases :

- ***Literature review and analysis of available data :*** *All the information regarding Ede plant's drainage system was provided by the Vallei en Veluwe Water Board. For the model's construction, all of treatment plant and nearby pumping station's flow data were analyzed, as well as supplementary precipitation data obtained from METEOBASE (2021).*

- ***Feature Engineering :*** *this process consists on selecting, manipulating, designing, transforming and training raw data into features (independent variables) to be used in supervised learning. In that sense, the proper input dataset is prepared in order to improve the general performance of the digital model.*

- ***Modelling :*** *consists on training the machine learning model with the target system's variables - in this research case, Ede plant's flow - in order to predict the future variables of interest. In this step, three regressors models were tested in order to chose the best model for Ede's system : Lasso, Random Forest and Gradient Boosting.*

Used methods/project phases

- ***Preliminary evaluation :*** *consists on the first evaluation of the created model, in order to determine its accuracy. For that, two metrics were chosen : Root – mean squared error (RMSE) and Mean absolute percentage error (MAPE).*

- ***Hyperparameters tuning :*** *consists on searching the best model's architecture for the given target's variables, based on each of the model's own parameters.*

- ***Final evaluation :*** *after improving internally the model, a final evaluation, with the same metrics, is performed to measure its accuracy.*

Results

For all forecasting horizon, the Gradient Boosting regressor presented the lowest errors for Ede's flow predictions with 11% MAPE for the first hour of predictions as an example. However, for more distant forecasting horizons, the predictions lose increasingly accuracy. For example, for 24 hours forecasting horizon, the predictions have 37% error comparing to the original flow's values.

Therefore, as the results are highly dependent on feature creation and selection, future investigations are expected for features that presented high correlation with the target values. In addition to that, after the best selection of features, the hyperparameters tuning routine proposed can be adopted to decrease, even more, the prediction errors.

Extra info/advice/link to final document and presentation

The final report is in attachment in this factsheet's e-mail.

LinkedIn's page : laura-cwienk-36bb