Welcome everyone to our presentation, I am Ramon and together with Julien, Adrien, Albert, Michael and Jesus we are team IMP. Today I will give you some insight in the work we have done for our project Imputation.

Today I will be discussing the next topics :
Imputation methods, where I will talk about the methods we have studied, learned and tried out so far.
Creating a pipeline, where we make sure all methods get tested the same way
Neural networks, here I will talk about different neural networks that we have studied to decide which NN we should use for our project.
Research paper, where I will show what our final product is about.

What is our project about?
Lectorate Energy in Transition is a group that conducts their research in the field of technology and economics. They gave a task to us to write a guideline on which imputation methods are best used in a time series building management systems.
Why is this important?
Missing data in databases can reduce the value of studies and can produce biased estimates, leading to invalid conclusions for companies. Imputing data has high potential for making data more valuable, meaningful and gaining new insights

In short what we have done so far is study which imputation methods there are and have been used in similar cases to ours. Written a program that can create artificial gaps in datasets so we can simulate incomplete data and use imputation methods on it to see how effective they can be. We have used open data from KNMI (the dutch weather institute) for this and a dataset from FactoryZero, a company that works on green houses that make energy transition feasible and more affordable.

The goal is to have studied, tested, compared and ranked different imputation methods and write guidelines about it to help future projects or companies with imputation for Building management system time series data.

Imputation Methods
Many datasets may contain missing values for various reasons. They are often encoded as NaNs, blanks or any other placeholders. Training a model with a dataset that has a lot of missing values can impact the machine learning model's quality. One way to handle this problem is to get rid of the observations that have missing data. However we would risk losing data points with valuable information. A better strategy would be to impute the missing values. In other words, we need to infer those missing values from the existing part of the data.

Mode median mean: This works by calculating the mean/median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others. It can only be used with numeric data. This is an easy and fast method and works well with small numerical datasets. However it doesn't factor the correlations between features. It only works on the column level. Not very accurate. Doesn't account for the uncertainty in the imputations.

K-Nearest Neighbour: How does it work? It creates a basic mean impute then uses the resulting complete list to construct a KDTree (This class provides an index into a set of k-dimensional points which can be used to rapidly look up the nearest neighbors of any point.). Then, it uses the resulting KDTree to compute nearest neighbours (NN). After it finds the k-NNs, it takes the weighted average of them. This can be much more accurate than the mean, median or most frequent imputation

methods but depends on the dataset. The downside is that K-NN is quite sensitive to outliers in the data.

Hot Deck Imputation:
What is Hot Deck? Hot deck imputation involves replacing missing values of one or more variables for a recipient, with observed values from a respondent which is called the donor that are similar to each other. In some versions, the donor is selected randomly from a set of potential donors, which we call the donor pool. In other versions a single donor is identified and values are imputed from that case, usually the "nearest neighbour". We have used this method on the FactoryZero dataset. In the dataset we work with data from 20 green houses and try to impute missing data with the Hot Deck Method. The overall goal here is to find a way to identify 2 houses most similar to each other to impute missing data in the gaps most similar to each other. The main challenge that remains is the donor pool selection. A good donor means : no missing data around the timestamps of the gaps that we are trying to impute Must follow, the trends of the "recipient" around the gap we are trying to impute.

Pipeline:
What is a pipeline? It is a uniform workflow allowing to load data from ,in our case KNMI or FactoryZero, visualize it, create gaps of different sizes, impute it and evaluate the imputation method using our evaluation criteria. In general the main objective of having a proper machine learning pipeline for our project is to exercise control over it. A well-organized pipeline makes the implementation more flexible for us.

Evaluation methods:
We studied evaluation methods used in similar researches and came up with the following methods:

- Variance of error: In the context of linear regression, or of any other model that can yield predictions on one variable (response) from values of other variables (predictors), we usually have a set of observations, that is, points where we observed the actual response and the predictors. Given a model, for each observation we can compute the predicted value (from model and predictors) and the actual value. The error is the difference between predicted and observed value.
- Skewness: Skewness is a measure of symmetry, or more precisely, the lack of symmetry.
- Kurtosis: Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case.
- RMSE: ***RMSE*** is the most common metric for evaluating a *regression* model. It is the *squared root* of *Mean Squared Error*. RMSE takes the squared root to scale the value back to its original units. If RMSE is high, then there is a large deviation in *actual* and *predicted* values. RMSE is *sensitive* to large errors. It is quite useful whenever you want to detect large differences in predicted and actual values.

Neural Networks:
We wanted to train a neural network for our research and have studied different approaches to this. Those are :

CNN: Convolutional neural networks are distinguished from other neural networks by their superior performance with image, speech, or audio signal inputs.

RNN: Recurrent Neural Networks. Type of neural network classified as trained, feedback. In ordinary neural networks, data points are considered independently from each other but RNN are able to incorporate the dependencies between data points and remember the previous inputs to generate new outputs. RNNs for missing data have been studied in earlier works and examples of where they have been applied are speech recognition and blood-glucose prediction. Those studies tried to handle missingness in RNNs by fusing missing entries or timestamps with the input or performing simple imputations. However, as far as we did research there hasn't been works which design RNN structures incorporating the patterns of missingness for time series classification problems.

GAN: Generative Adversarial Networks. Classified as unsupervised, it consists of two different models: the generator model that we train to generate new examples, and the discriminator model that tries to classify examples as either real (from the domain) or fake (generated).

GAIN stands for Generative Adversarial Imputation Networks. It seems to be the most popular GAN architecture to handle missing data. The idea behind it is straightforward: Generator takes the quantity of real data which has some missing values and imputes them accordingly. The imputed data is fed back to the Discriminator whose job is to figure out which data was originally missing.

As far as training the neural networks we haven't touched enough upon this yet to show here but will be working on that in the coming weeks.

Research paper:
Our final product will be a research paper/guidelines that describe the different imputation methods we studied, used and ranked. In the guidelines there will be reasons why the methods should be used and in which situations they can be used most effective. This will be published in CLIMA 2022. What is CLIMA? This is a congress is organized every 3 years where CLIMA 2022 brings professional and researchers together to discuss and find answers to the research questions in the field of energy, digitization and Health & comfort.