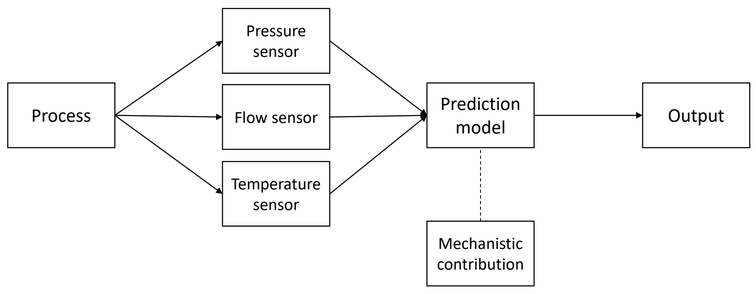
Design of a soft sensor for the insulin purification process

Daniel Lopez Alonso, Tanmay Chaturvedi (AAU), Kasper Juhl Græsbøll Ottesen (NN)

INTRODUCTION

What is a soft sensor?

A soft sensor can be defined as a prediction model that delivers a measurement-like output. Process data is used as an input to predict a target, and therefore works as an indirect measurement system. The input data can be obtained from other hardware sensors in a data-driven soft sensor, or it can be based on process variables transformed by first principle formulas in a mechanistic soft sensor. A hybrid approach can also take place as a combination of the above.



The process at Novo Nordisk

The drying of insulin crystals is an important part of the purification process at Novo Nordisk. Tight control over the drying parameters as well as the moisture percentage is important for quality control. The two main focus of this thesis are:

1. To gain insight into the relationships between the parameters involved in drying through data analysis.
2. To use this knowledge for designing a soft sensor that can substitute offline moisture measurement by an in-line predictive model.

Why use soft sensors?

Soft sensors may be used in several applications:

- Measuring system back-up
- In What-if analysis
- Real-time prediction

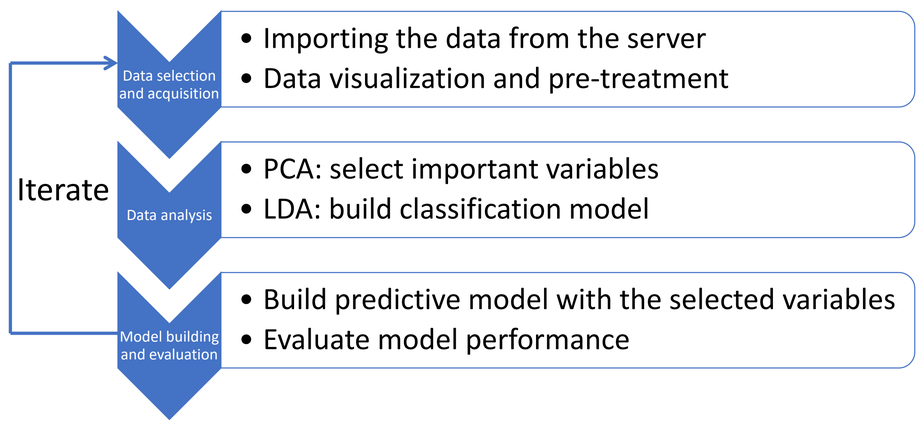
They offer several advantages with regards to hardware sensors as they are a low-cost alternative, can work in hostile environments and for hard to measure process parameters. They can also be used in combination rather than substitution of hardware sensors in processes that have tight control requirements.

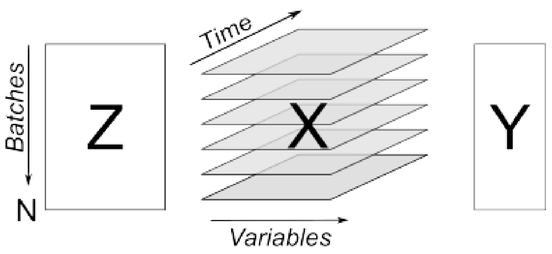
METHODS

1. Data selection

The relevant data is selected by understanding and screning the process with help of data and process specialists. Two different types of data are available: batch-dependent and time-dependent data. The image below gives an idea of the problem with multidimensionality of time-dependent data which needs to be unfolded.

Data pre-processing includes the identification and elimination of outliers and the data centering and standardization. Batch alignment is also necessary in specific cases and can be done through a maturity variable or through dynamic time warping (DTW).

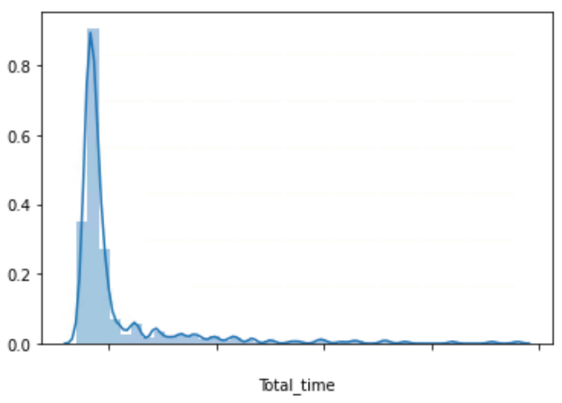




2. Data analysis

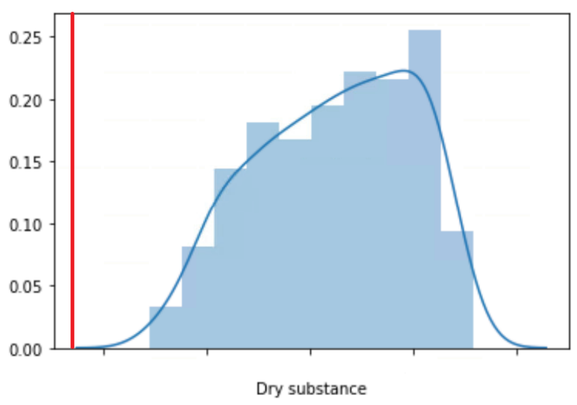
The main purpose of the data analysis is to gain insight about the factors that may be responsible for the existence of elongated batches. This would allow to change the process accordingly in order to avoid longer times and increase the overall stability.

There are two main methods utilized in this work: Principal component analysis (PCA) and Linear discriminant analysis (LDA).



Principal component analysis (PCA)

PCA is an unsupervised analysis that allows to summarize the information from several variables by projecting them into principal components. These principal components can be classified according to the amount of variance in the data they explain. It is possible to find the loadings, which represent the importance of individual batches within the component. The scores represent the individual observations projection to the latent space.



Linear discriminant analysis (LDA)

LDA is a supervised classification tool that can also be used for data analysis and dimensionality reduction. Similarly to PCA, it utilizes the projection to a latent space with the constraint of maximizing the separation between known categories rather than maximizing the variance.

3. Soft sensor

The main purpose of the soft sensor is to elaborate a predictive model that can deliver a value of the final dry substance in-line. This would prevent the waiting time caused by the off-line analysis at QC labs.

Three different approaches are attepted at the design of a soft sensor: mechanistic, hybrid and data-driven. The focus here will be given to the Data driven approach

Multiple linear regression (MLR)

The MLR model assumes a linear relationship between the predictor variables and the response variable. The coefficients for each variable contribution to the model are found through the least squares method.

Projection on latent structures(PLS)

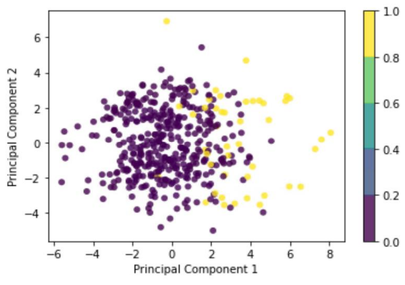
The PLS also uses a Linear regression assumption but it is carried out on the principal components resulting from dimensionality reduction. Furthermore, the components are oriented in the direction that maximizes the variance within predictor(X) and response(Y) variables, but also the covariance among them. For this reason PLS models are preferred for a predictive purpose.

RESULTS

Data analysis results

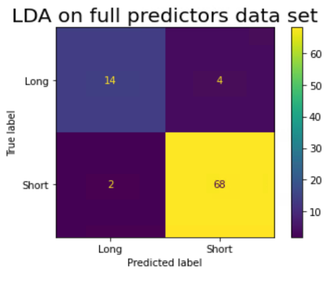
PCA

PCA was carried out with 3 components and a respective explained variance of 16%, 13% and 9%. Despite the low explained variance, some form of clustering can be observed along the axis of PC1. The variables correlating with PC1 are then considered possibly significant. Nevertheless, a large amount of the variance is still unaccounted for as it can be observed from the huge overlap in the Figure below.



LDA

LDA developed a classifier model that could classify the batches with an accuracy of 93% with the chosen predictors. The specificity corresponding to 77% shows the model limitations when classifying true long batches. Furthermore, this classifier uses all data available until the end of the process and therefore cannot be used for its predictive capabilities.



Hybrid approach

The crystal coefficient obtained from the derivation of Kozeny-Carman equation was found to have no correlation with the response variables. For this reason, the coefficient was not used in the predictive models. There are 3 possible explanations for the lack of correlation:

1. The coefficient does not have an effect on the drying process
2. The equation has been over-simplified and the coefficient is not calculated accurately
3. The filtration step does not experience laminar flow, making the equation invalid.

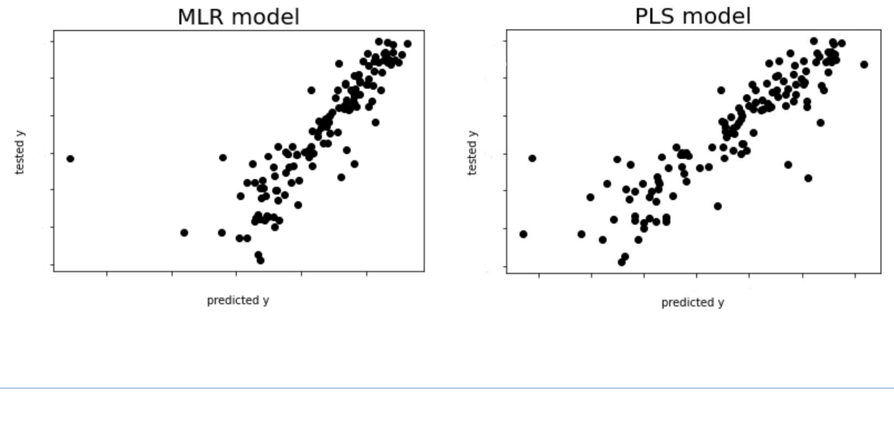
Model performance results

MLR

The multiple linear regression model yielded a coefficient of determination of 0.66 and mean error of 0.87. The predicted vs measured plot can be seen below.

PLS

The PLS model yielded a coefficient of determination of 0.78 and mean error of 0.70. This is until now the best performing model, utilizing all available variables as predictors. The predicted vs measured plot can be seen below.



PERSPECTIVES

Perspectives for the project

From this work, the company is provided with a list of variables that may have a significant effect on the drying process as concluded by the PCA. A well-performing PLS model is provided which may be used as a foundation to build a better model that can be validated as a soft sensor. Designing experiments to evaluate the crystal parameters may be a useful path to explain a greater part of the variance observed between batches. Finally, a python script for importing time-series data from the server as well as the code for the different types of models is provided, together with proposals on how to continue moving forward on this project.

Personal perspectives

On a personal note, doing my thesis with Novo Nordisk has been a rich experience which has prepared me in several skills necessary to join the workforce in the future. I have learned abilities that will help me become a competent worker, but also to know myself better. I now know the topics where I would like to focus my career as well as my areas of expertise and the ones that have room for self-improvement. The friendly environment created by Helix Lab as well as their professional advice on technical matters have been an essential part of succesfully completing this thesis. The great work environment at the PP Future Factory played a big role in having a positive first experience in the workplace. The support from my Aalborg university supervisor and professor has been equally valuable in moving forward with this project.

Overall, doing this thesis collaboration has been the richest experience experience of my student life and I am grateful to Aalborg university, Helix Lab and Novo nordisk for presenting me with this opportunity.