Lehrstuhl Steuerung,
Regelung und Systemdynamik

Bachelor Thesis

Programming

Preprocessing of consolidated bioprocessing datasets for data-driven and hybrid modeling

Keywords: feature selection, data quality; categorical encoding microbial consortia

Conditions:

Duration:	3 months
Requirements:	MATLAB or Python knowledge (demonstrated)
Language:	English
Target group:	Bachelor students

Contents:

Consolidated bioprocessing (CBP) is a promising approach in bioprocess engineering, integrating enzyme production, biomass hydrolysis, and fermentation into a single step to simplify biofuel production and reduce costs. This integration yields high efficiency but also complex process behavior with many interdependent variables. In modern research, data-driven modeling techniques (e.g. machine learning) are increasingly used to analyze and simulate such complex bioprocesses. However, high-quality datasets are essential for reliable modeling of CBP dynamics.

In this thesis, a systematic data engineering pipeline will be developed for CBP. First, relevant data sources must be identified and high-quality secondary datasets gathered. The collected data will be preprocessed and cleaned: this involves handling missing or inconsistent values (using imputation strategies or data filtering), correcting any errors, and standardizing units and formats across different sources. Categorical variables (for instance, different microbial strains and pretreatment method) will be encoded appropriately so that they can be utilized in modeling. The dataset may also undergo feature selection or basic feature engineering to highlight the most relevant process variables, reducing dimensionality and focusing on key factors that influence CBP performance. Finally, the curated data will be organized into a structured format (such as a cleaned spreadsheet or database table) that is directly suitable for downstream analysis and machine learning model development.

The goals of this work are:

- Identification and vetting of high-quality secondary CBP data sources
- Acquisition, integration, and harmonization of datasets (input and output)
- Preprocessing of data (cleaning, missing values, outliers) and categorical encoding
- Feature scaling and selection to produce modeling-ready inputs
- Complete and detailed documentation/presentation of the research results

Supervisors:	M. K. Yeboah, MPhil.	Univ.-Prof. Dr.-Ing. D. Söffker
Office:	MB 361	MB 341
Telephone:	01779355391	0203/379-3429
E-Mail	mark.korangyeboah@uni-due.de	soeffker@uni-due.de