Forecasting of Mannosylerythritol Lipids Production using Artificial Intelligence Methods

Carolina A. Vares[1,2], Sofia P. Agostinho[1,3], Ana L. N. Fred[1,3], Nuno T. Faria[1,2], Carlos A.V. Rodrigues[1,2,4]

[1] *Department of Bioengineering, IBB—Institute for Biotechnology and Bioengineering, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal*

[2] *Associate Laboratory i4HB – Institute for Health and Bioeconomy at Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal*

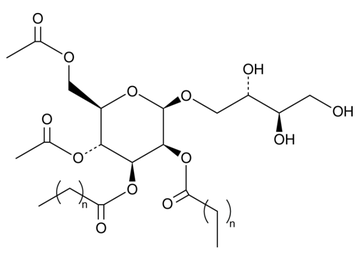
[3] *Instituto de Telecomunicações (IT), Av. Rovisco Pais 1, Torre Norte Piso 10, 1049-001 Lisbon, Portugal*

[4] *Cell4Food, Avenida General Norton de Matos, 4450-208 Matosinhos, Portugal*

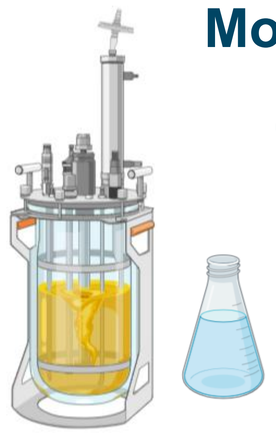
Motivation

Case-study:
Mannosylerythritol lipids (MELs) fermentation

MELs are biosurfactants that can be used in the pharmaceutical and food industries



Production is not yet cost-effective to be economically viable at larger scales

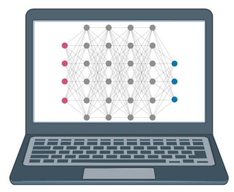


Fermentations are **complex** and **unpredictable** processes

Generates

Large volumes of **data** are generated during the process

Allows



The development of **data-driven** models, such as **machine learning**

- Better knowledge of the process
- Improve predictability

Objectives

Predict MELs concentration at the end of the **fourth** day of fermentation

Predict MELs concentration at the end of the **seventh** day of fermentation

Develop a **pipeline** that can be applied to future MELs datasets

Methodology

Dataset

Constructed from 4 different experiments

Day 0 Day 2 Day 4 Day 7

Offline measurements (glucose, biomass,…) were taken every 2 days and then at the seventh day

47 shake flasks samples

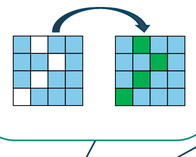
9 categorical features

42 features

33 numerical features

Data preprocessing

Value imputation



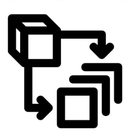
Encoding

Data scaling

$$z = \frac{x - u}{\sigma}$$

Feature engineering

Feature extraction



Feature selection

All features

Feature selection

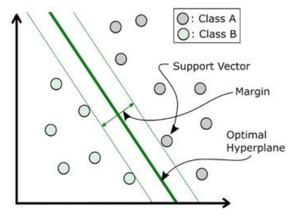
Final features

- PCA
- ANOVA
- Correlation
- LASSO
- RFE

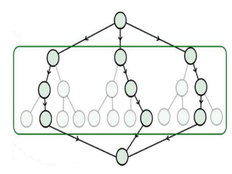
Clean dataset

Machine learning models

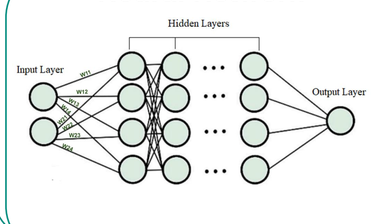
Support vector machine



Random Forest



Neural Network



Hyperparameter tuning using grid search

For each machine learning model used, different **hyperparameters** combinations were tested in the **training** set

Validation set

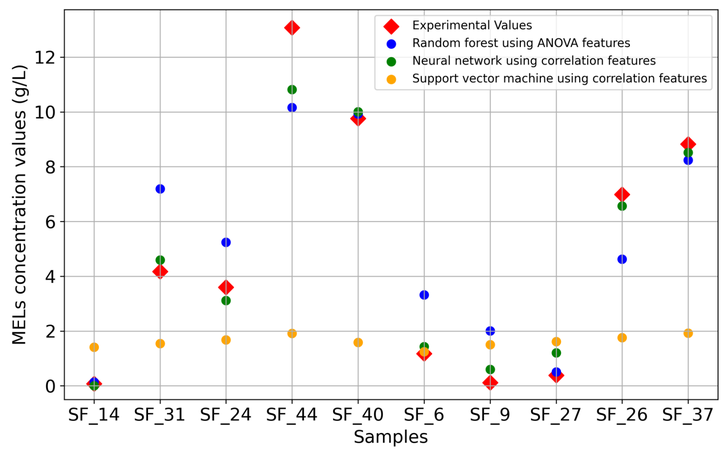
During training, a **validation set** was created and used to validate the models and the hyperparameters

Testing

Forecast MELs production at the end of the seventh day, using data **unseen** by the models

Results

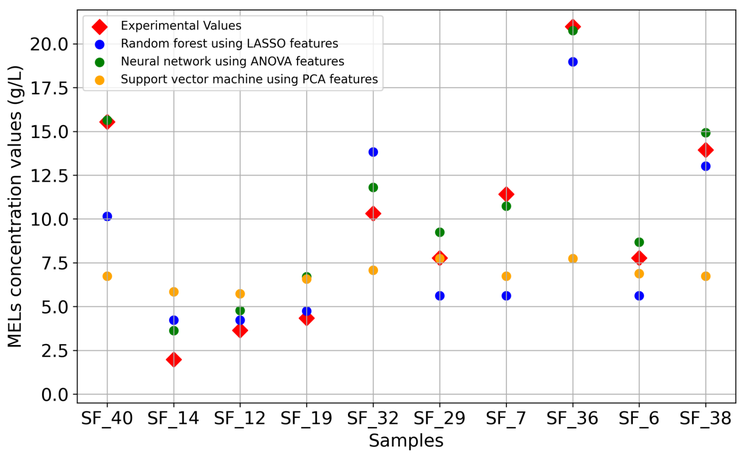
Day 4



Forecasting of test samples of MELs production at the end of day 4 using different models and feature subsets

Models: Feature subset	Mean Squared Error	Mean Absolute Error
Random Forest: ANOVA	3.45	1.49
Support Vector Machine: Correlation	28.24	4.08
Neural Network: Correlation	0.69	0.58

Day 7



Forecasting of test samples of MELs production at the end of day 7 using different models and feature subsets

Models: Feature subset	Mean Squared Error	Mean Absolute Error
Random Forest: LASSO	9.48	2.52
Support Vector Machine: PCA	36.25	4.63
Neural Network: ANOVA	1.63	1.10

Conclusions

Predictions for day 4 were more accurate than those for day 7, likely because the MEL production in the early stages of fermentation is more strongly influenced by a specific feature or set of features.
The ability to forecast the MEL concentration for day 7 with data from the first 4 days of fermentation is quite promising as an **early batch quality assessment metric**.

Both results are promising as they show that it is possible to have reliable predictions for **day 4** and **day 7**. This could in turn help **reduce** the **costs** associated with the process, making it more **economically viable**.

Want to learn more about this work? Check out the article! →

Acknowledgments

This work was supported by FCT through the PhD grant (2025.02674.BDANA). The authors acknowledge Bluebio Alliance, Portugal, for the networking, industrial contacts and project support. Thanks to institutional funding through iBB-Institute for Bioengineering and Biosciences (UID/04565/2025), i4HB—Associate Laboratory Institute for Health and Bioeconomy (LA/P/0140/2020). This work was carried out within the SURFs UP project. The project is supported by the Circular Bio-based Europe Joint Undertaking and its members under grant agreement No 101157586 and UKRI grant agreement No 10115826. Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or CBE JU, or UKRI. Neither the European Union, CBE JU, nor the UKRI can be held responsible for them.