Machine Learning to better understand and optimize cheese production

Manon Perrignon, Mathieu Emily, Romain Jeantet, Thomas Croguennec

Machine Learning to better understand and optimize cheese production

Manon Perrignon[1], Mathieu Emily[2], Romain Jeantet[1] and Thomas Croguennec[1]

[1] L'Institut Agro, INRAE, STLO (Science et Technologie du Lait et de l'œuf), Rennes, France
[2] L'Institut Agro, Université de Rennes, CNRS, IRMAR (Institut de Recherche Mathématique de Rennes)-UMR 6625, Rennes, France

CONTEXT

Cheese production and monitoring:

→ Milk standardisation → Cheese making → Cheese ripening →

Dry matter
Yield
Quality
...

Cheese production and monitoring:

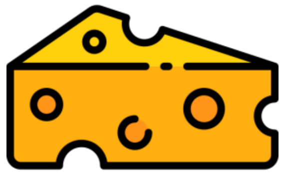
Milk standardisation → Cheese making → Cheese ripening →

Dry matter
Yield
Quality
...

- **Complex** process
  - Many sources of **variability** (process , ingredient,..)
  - Many process **parameters** to monitor (manual, automatic)
- **Large amount of data** collected during daily cheese process

Cheese production and monitoring:

Milk standardisation → Cheese making → Cheese ripening →



Dry matter
Yield
Quality
…

- **Complex** process
  - Many sources of **variability** (process , ingredient,..)
  - Many process **parameters** to monitor (manual, automatic)
- **Large amount of data** collected during daily cheese process

How to improve dry matter by adopting a holistic view of the process and associated data ?

Dry matter optimization (target value) at present:

Guinee TP (2021) International Dairy Journal **121**: 105095
Kern C et al. (2019) Food Research International **121**: 471–478

→ Modification of standardisation parameters (casein micelle content,...)

→ Modification of process parameters (stirring time/speed,...)

Dry matter optimization (target value) at present:

Guinee TP (2021) International Dairy Journal **121**: 105095
Kern C et al. (2019) Food Research International **121**: 471–478

→ Modification of standardisation parameters (casein micelle content,...)

→ Modification of process parameters (stirring time/speed,...)

Optimization tools

- Expert knowledge
- Model using classical statistical methods (linear regression,...)

→ **No consideration of all process and manufacturing parameters**

Methodology for modelling dry matter:

- Complex process
- No global equation
- Huge amount of data

Need an appropriate and data-driven method

Methodology for modelling dry matter:

Dagan DT, Wilkins EJ (2023) Journal of Outdoor Recreation and Tourism 100668
Wang Z et al. (2022) Computers & Chemical Engineering **165**: 107945

- Complex process
- No global equation
- Huge amount of data

Need an appropriate and data-driven method

(Generalized) Linear Regression

+	−
→ Known function	→ Additivity of effects
→ Easy interpretation	

Methodology for modelling dry matter:

Dagan DT, Wilkins EJ (2023) Journal of Outdoor Recreation and Tourism 100668
Wang Z et al. (2022) Computers & Chemical Engineering **165**: 107945

- Complex process
- No global equation
- Huge amount of data

} Need an appropriate and data-driven method

(Generalized) Linear Regression

+
→ Known function
→ Easy interpretation

−
→ Additivity of effects

Machine Learning

+
→ Ability to detect complex relationships
→ High prediction power

−
→ Black box: difficult interpretation

Methodology for modelling dry matter:

Dagan DT, Wilkins EJ (2023) Journal of Outdoor Recreation and Tourism 100668
Wang Z et al. (2022) Computers & Chemical Engineering **165**: 107945

- Complex process
- No global equation
- Huge amount of data

} Need an appropriate and data-driven method

(Generalized) Linear Regression

+

→ Known function

→ Easy interpretation

−

→ Additivity of effects

Machine Learning

+

→ Ability to detect complex relationships

→ High prediction power

−

→ Black box: difficult interpretation

How to implement a Machine Learning approach to optimize dry matter ?

METHOD

Data obtained from one cheese company over a one year period:

Classical pre-processing of data = obtain the database suitable for analysis

⚠ In collaboration with industrial experts

→ Remove **redundant variables**
→ Remove **outliers**
→ Remove **missing data**

Data obtained from one cheese company over a one year period:

Classical pre-processing of data = obtain the database suitable for analysis

⚠ In collaboration with industrial experts

→ Remove **redundant variables**
→ Remove **outliers**
→ Remove **missing data**

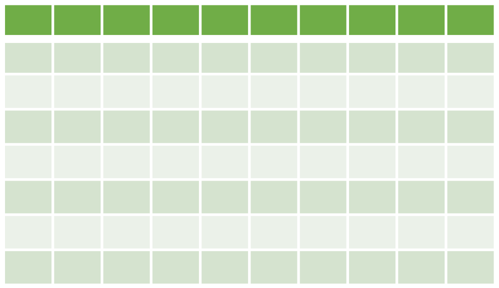
After pre-processing:

Nb. individuals (production vat) : **~ 3000**

Nb. variables : **~ 100**

100 columns

3000 rows



Selection of Machine Learning methods:

Breiman L (2001)
Friedman JH (1999)
Boser et al. (1992)

RANDOM FOREST (2001)

➔ Uses a set of decision trees built on random sub-samples of the training data

GRADIENT BOOSTING (1999)

➔ Builds decision trees sequentially, with each new tree correcting the errors of the previous ones

Selection of Machine Learning methods:

Breiman L (2001)
Friedman JH (1999)
Boser et al. (1992)

RANDOM FOREST (2001)

➔ Uses a set of decision trees built on random sub-samples of the training data

GRADIENT BOOSTING (1999)

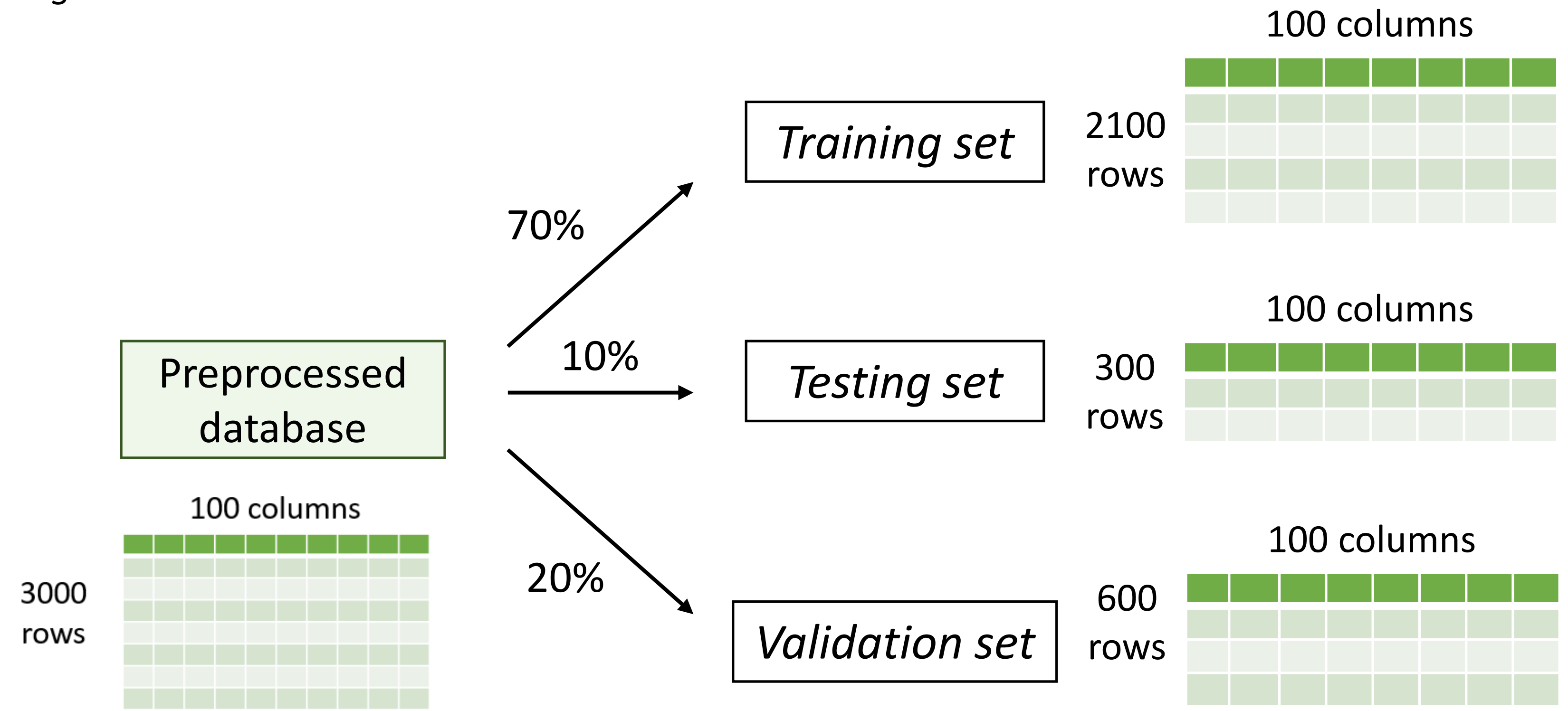
➔ Builds decision trees sequentially, with each new tree correcting the errors of the previous ones

SUPPORT VECTOR MACHINE (SVM) (1992)

➔ Find the optimal hyperplane that separates or fits the data

Comparison of Machine Learning methods:

Using R



Comparison of Machine Learning methods:

*RMSE (Root mean square error) = Standard deviation of the residuals (prediction error)

Using R

Preprocessed database

100 columns

3000 rows

70% → *Training set* — 100 columns, 2100 rows → Training model

10% → *Testing set* — 100 columns, 300 rows → Optimizing hyperparameters

20% → *Validation set* — 100 columns, 600 rows → Predicting and evaluating model with Root Mean Square Error (RMSE)*

→ Resampling techniques (cross-validation, bootstrap, out-of-bag) can be used to optimize hyperparameters

Interpretation of Machine Learning models:

Breiman L (2001) Machine Learning
Lundberg SM et al. (2018) Computer Science, Mathematics

Importance of variables in model:

Using R

Principle: Calculate the importance of variables in the model for predicting dry matter

→ Rank variables according to their importance in predicting the variability of the target

Interpretation of Machine Learning models:

Breiman L (2001) Machine Learning
Lundberg SM et al. (2018) Computer Science, Mathematics

Importance of variables in model:

Using R

Principle: Calculate the importance of variables in the model for predicting dry matter

→ Rank variables according to their importance in predicting the variability of the target

Shapley value:

Using Python

Principle: modify one variable at a time, keeping all others constant, to assess its impact on dry matter

→ Assess the single effect of a variable

Interpretation of Machine Learning models:

Breiman L (2001) Machine Learning
Lundberg SM et al. (2018) Computer Science, Mathematics

Importance of variables in model:

Using R

Principle: Calculate the importance of variables in the model for predicting dry matter

➜ Rank variables according to their importance in predicting the variability of the target

Shapley value:

Using Python

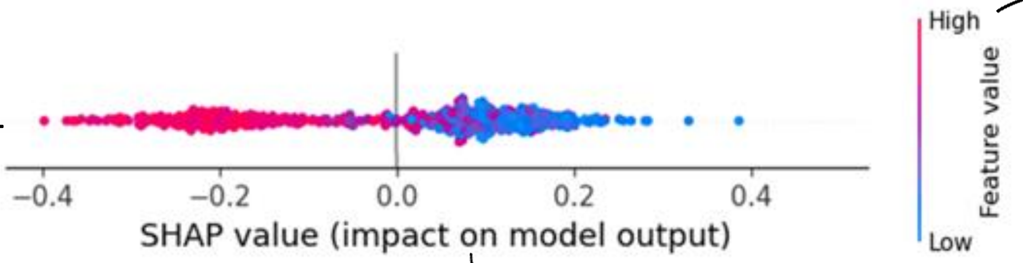
Principle: modify one variable at a time, keeping all others constant, to assess its impact on dry matter

➜ Assess the single effect of a variable

Results example:

One point = one prediction

Variable1



Colour scale to represent variable value

Output prediction

RESULTS

Comparison of methods:

- 3 machine learning methods and 1 classical statistical method
- For each method: model training, hyperparameter optimization, model evaluation

Comparison of methods:

- 3 machine learning methods and 1 classical statistical method
- For each method: model training, hyperparameter optimization, model evaluation

Results of the four methods:

Method	RMSE
Random Forest	0.27
Gradient Boosting	0.35
Linear Regression	0.37
SVM	0.37

Comparison of methods:

- 3 machine learning methods and 1 classical statistical method
- For each method: model training, hyperparameter optimization, model evaluation

Results of the four methods:

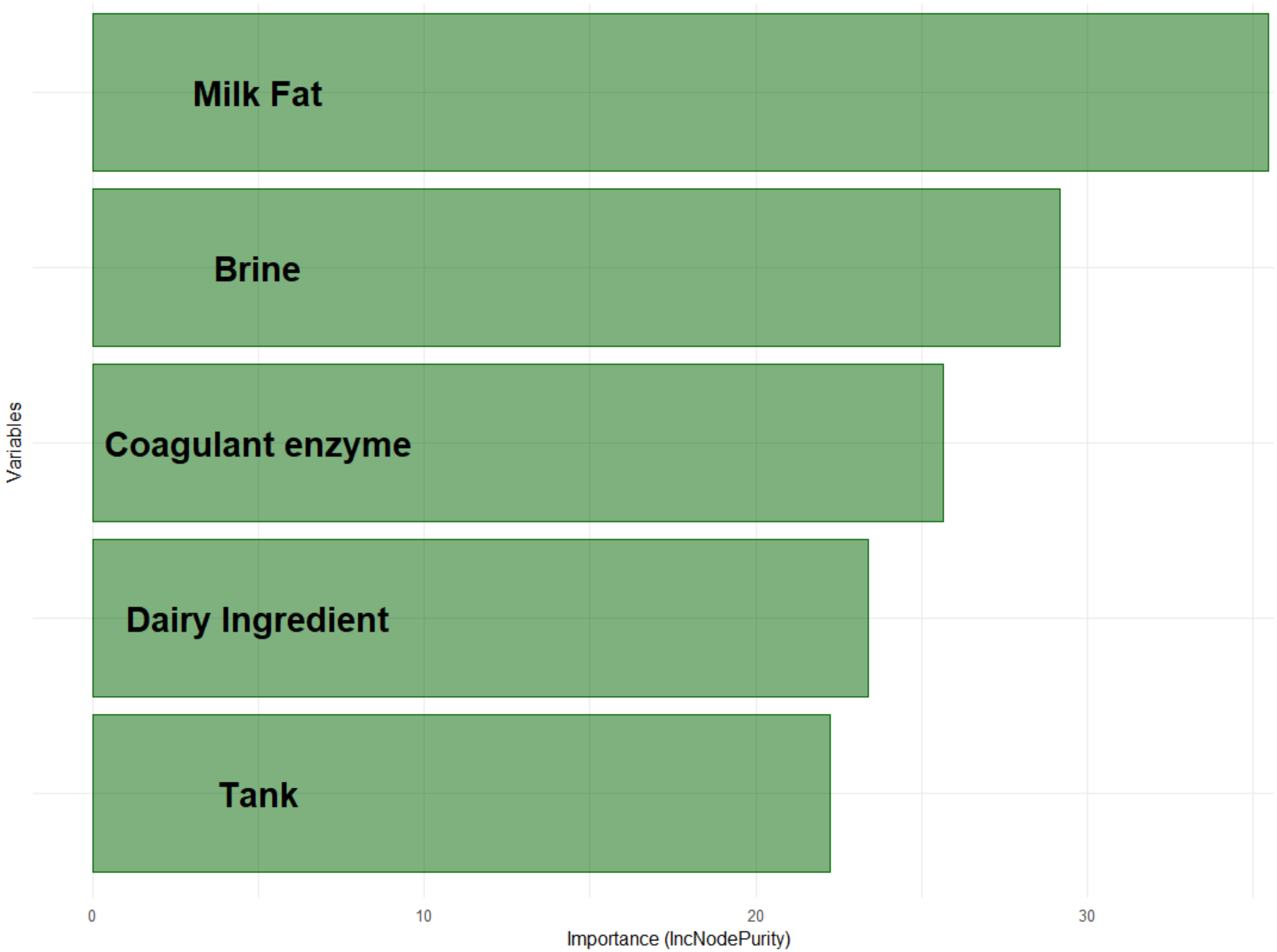
Method	RMSE
Random Forest	0.27
Gradient Boosting	0.35
Linear Regression	0.37
SVM	0.37

➡ Selection of Random Forest to model dry matter

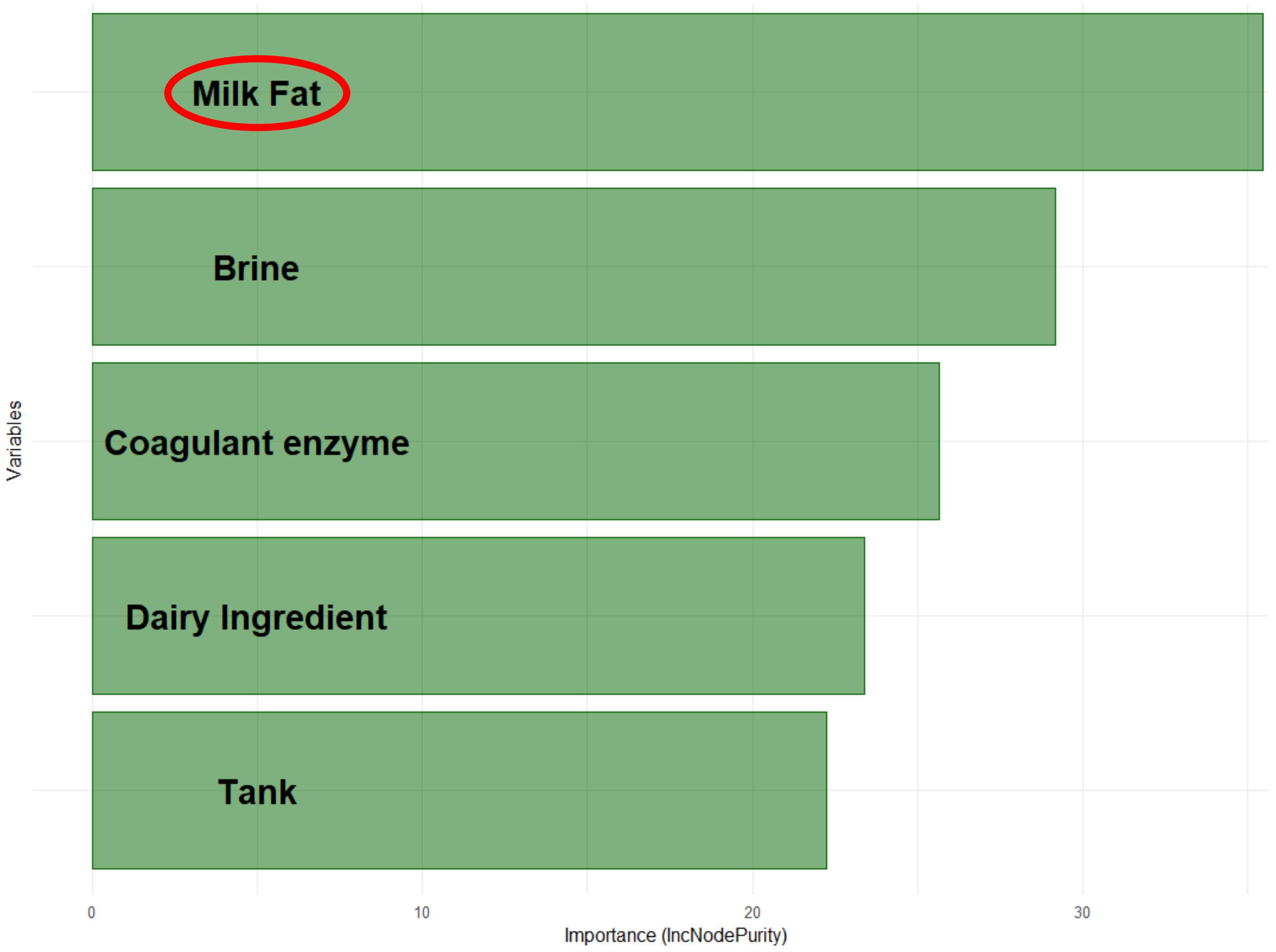
% of variability explained
66.6

- Additional data could enhance this measurement and the accuracy of the model

Importance of variables on dry matter with Random Forest:



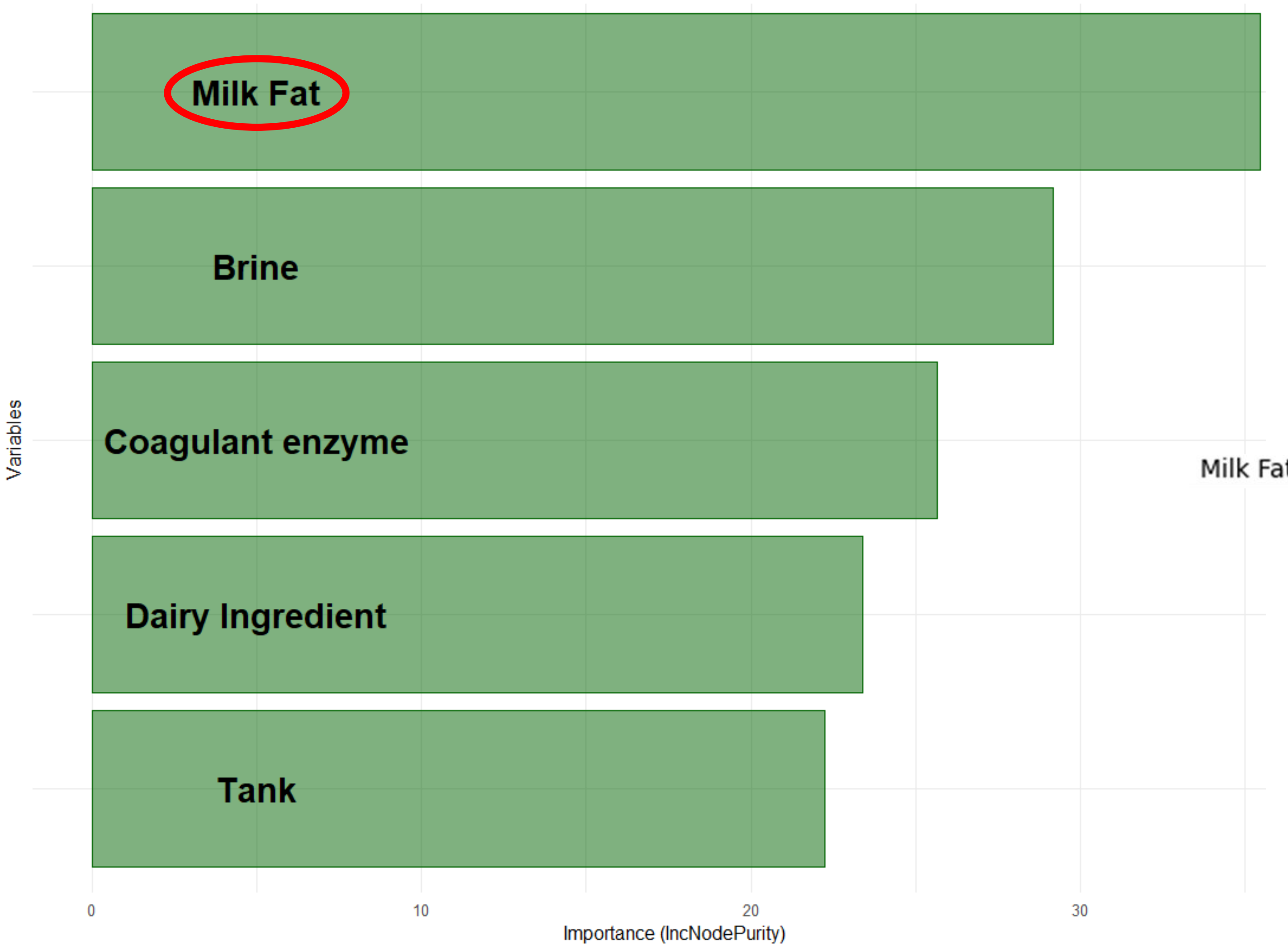
Importance of variables on dry matter with Random Forest:



Milk fat:

Informative variable

Importance of variables on dry matter with Random Forest:



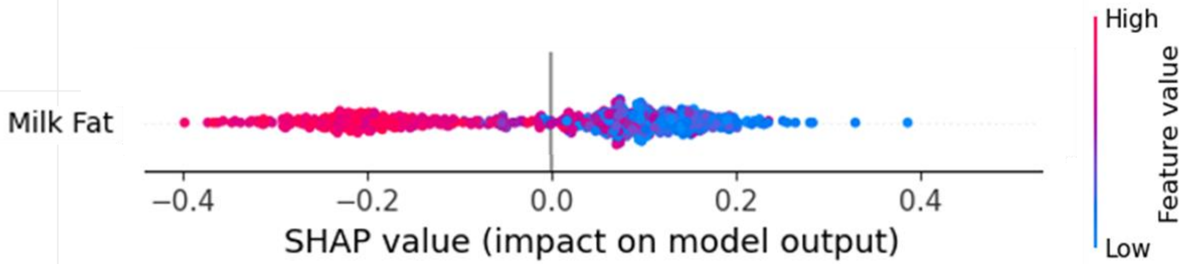
Milk fat:

Informative variable

How does this variable affect dry matter ?

⇩

SHAPLEY VALUE

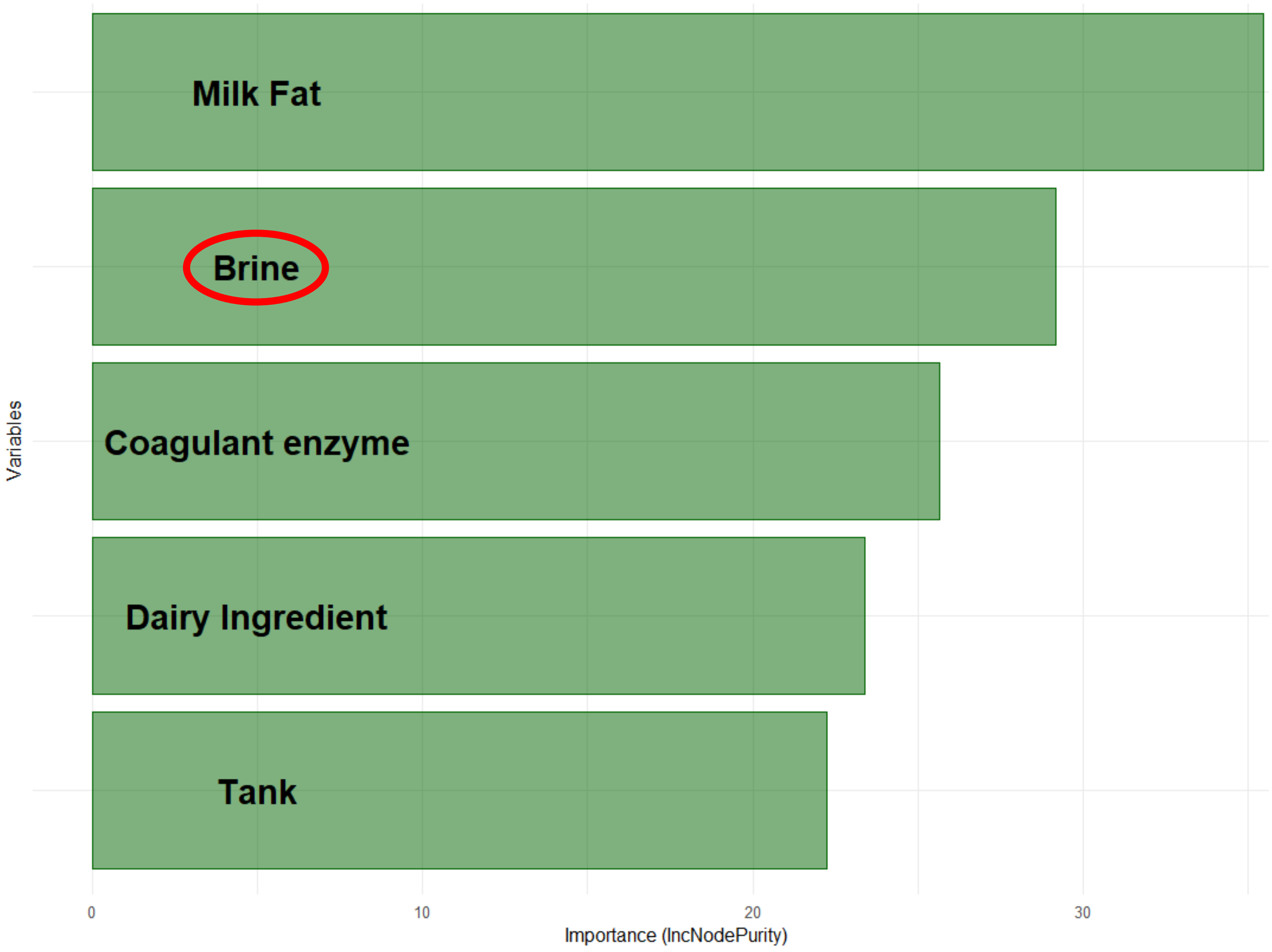


→ Higher milk fat leads to lower dry matter

Potential challenge for industrial process:

- How to adapt process to a given milk composition

Importance of variables on dry matter:

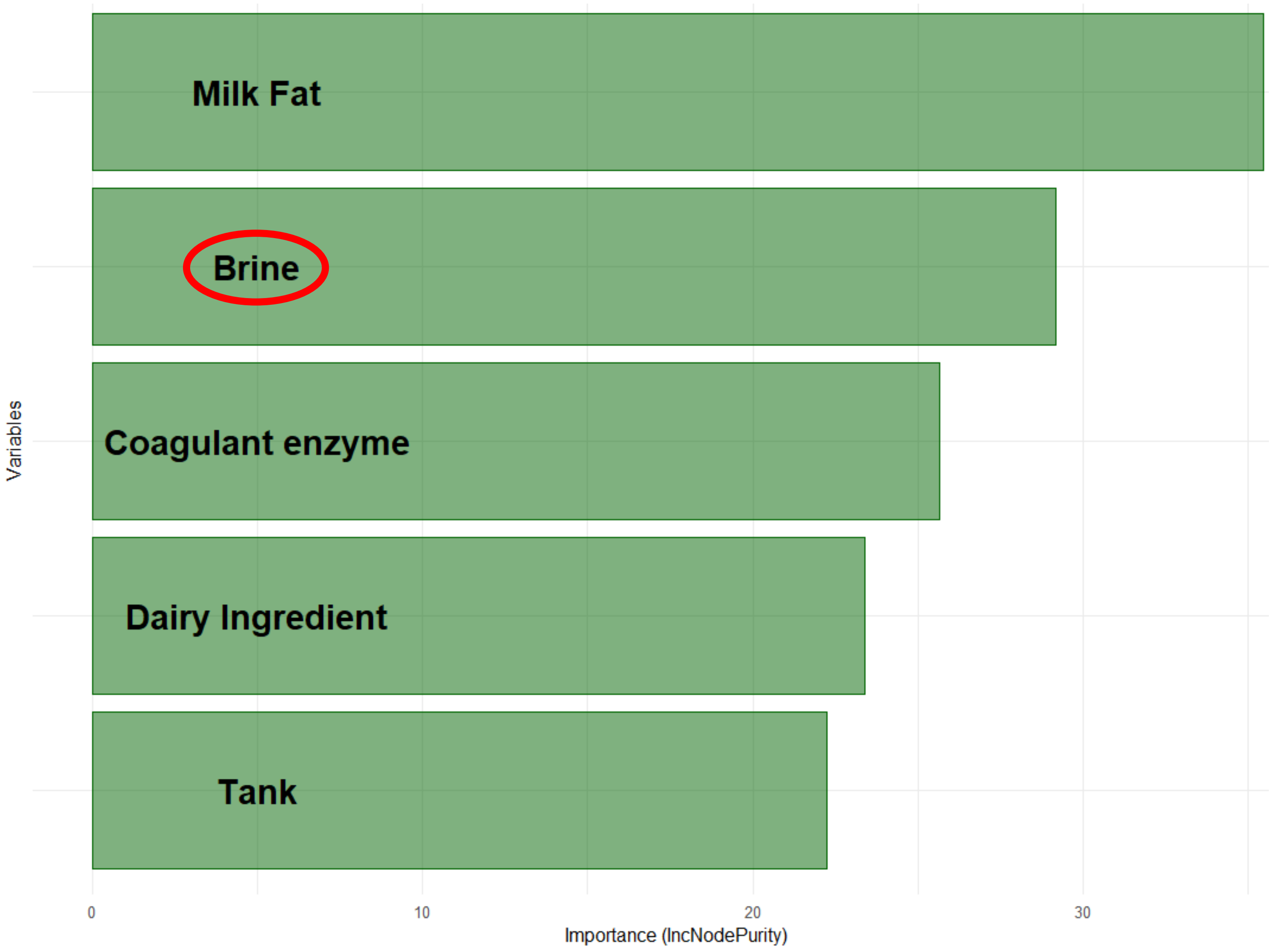


Brine:

Cheese position in the brine pool

Informative variable

Importance of variables on dry matter:



Brine:

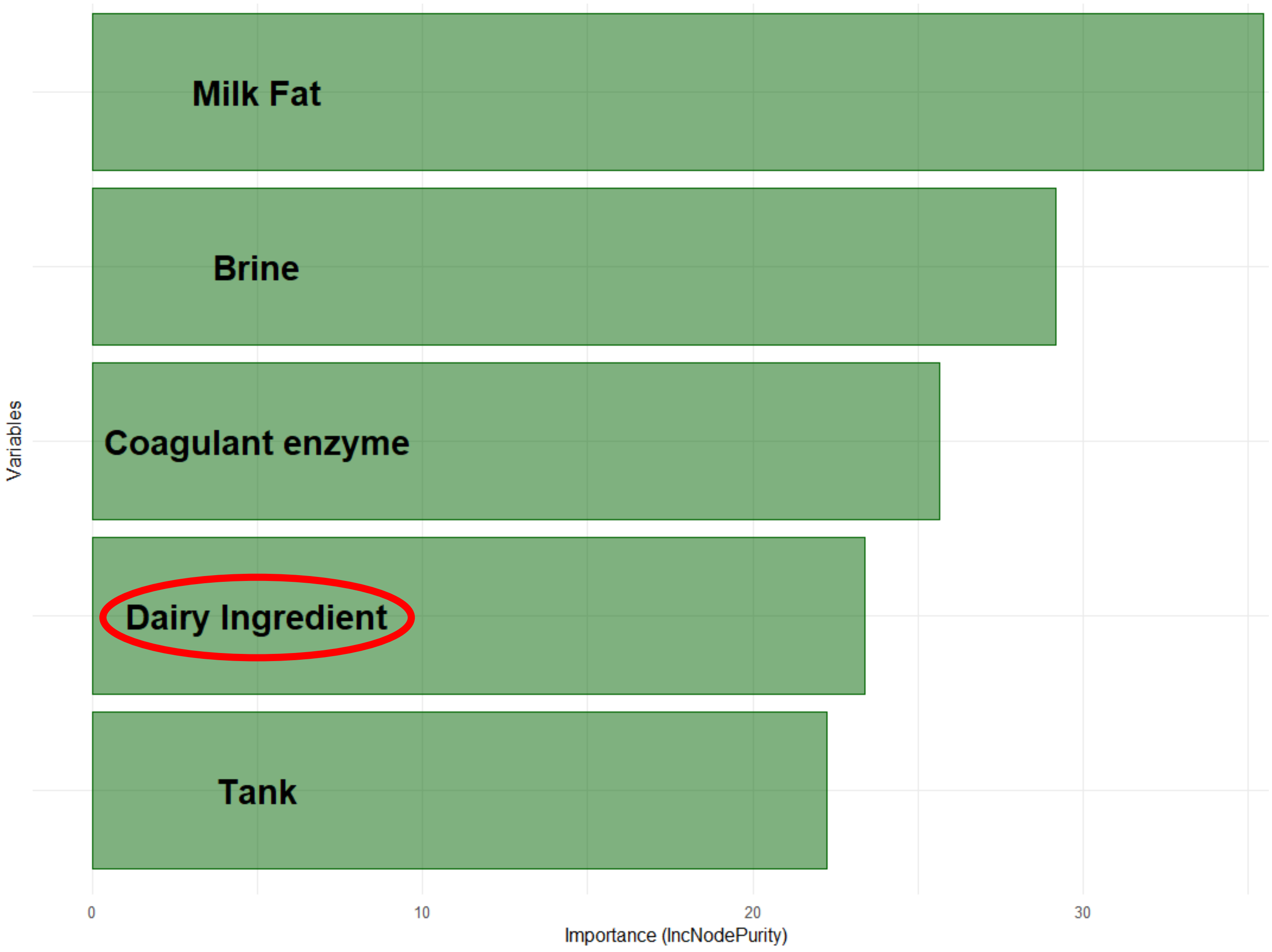
Cheese position in the brine pool

Informative variable

Potential challenge for industrial process:

- Checking information with experts
- Measuring new data

Importance of variables on dry matter:

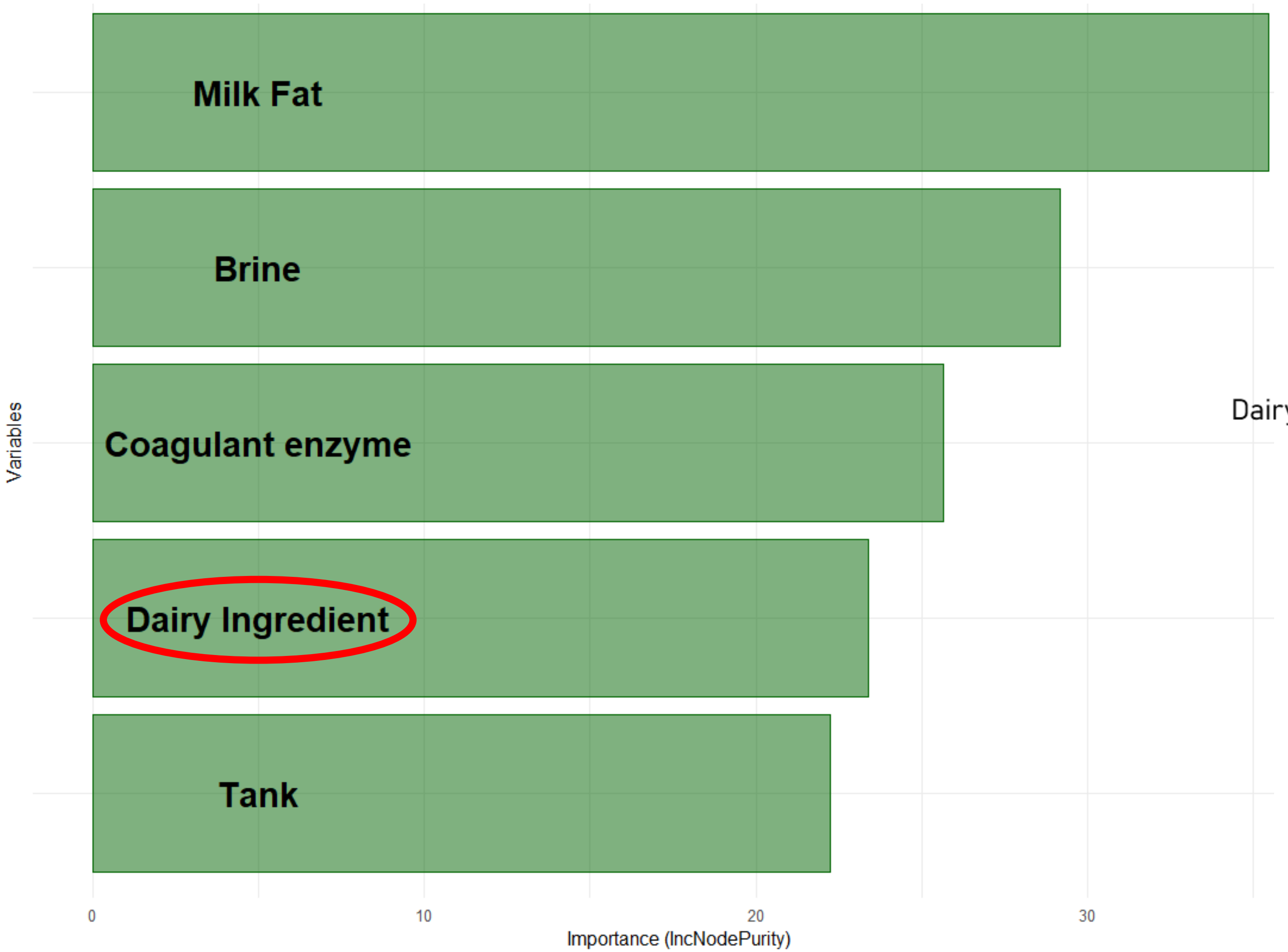


Dairy ingredient:

Quantity of dairy ingredient incorporated for standardization

Actionable variable

Importance of variables on dry matter:

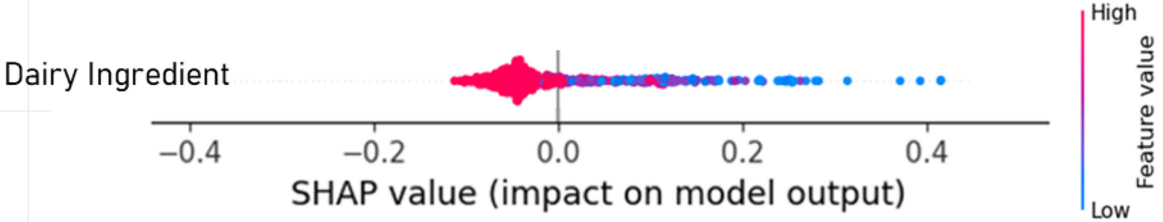


Dairy ingredient:

Quantity of dairy ingredient incorporated for standardization

Actionable variable

SHAPLEY VALUE:



→ Higher amount of dairy ingredient leads to lower dry matter

Potential challenge for industrial process:

- Understand the impact of this ingredient to better adapt standardisation

CONCLUSION
AND
PERSPECTIVES

CONCLUSION

- Machine Learning establish **complex relationships** between process parameters and dry matter
- Essential **collaboration** with experts to understand output of the model and overall data
- Need for a large database to implement machine learning methods

CONCLUSION

- Machine Learning establish **complex relationships** between process parameters and dry matter
- Essential **collaboration** with experts to understand output of the model and overall data
- Need for a large database to implement machine learning methods

PERSPECTIVES

- Cheese production defined by **several performance indicators**
- Machine learning methods learn from data: **industrial trials** can provide new information
- **Known equation** could be integrated into modelling: hybrid model

Thanks for your attention !

Contact : manon.perrignon@agrocampus-ouest.fr