Predicting Nitrogen Excretion of Dairy Cattle with Machine Learning

Herman Mollenhorst[1(✉)], Yamine Bouzembrak[2], Michel de Haan[1],
Hans J. P. Marvin[2], Roel F. Veerkamp[3], and Claudia Kamphuis[3]

[1] Livestock & Environment, Wageningen Livestock Research,
Wageningen University and Research, P.O. Box 338,
6700 AH Wageningen, The Netherlands
{erwin.mollenhorst,michel.dehaan}@wur.nl

[2] Toxicology, Novel Foods & Agro Chains, Wageningen Food Safety Research,
Wageningen University and Research, P.O. Box 230, 6700 AE Wageningen,
The Netherlands
{yamine.bouzembrak,hans.marvin}@wur.nl

[3] Animal Breeding and Genomics, Wageningen University and Research,
P.O. Box 338, 6700 AH Wageningen, The Netherlands
{roel.veerkamp,claudia.kamphuis}@wur.nl

Abstract. Several tools were developed during the past decades to support farmers in nutrient management and to meet legal requirements such as the farm specific excretion tool. This tool is used by dairy farmers to estimate the farm specific nitrogen (N) excretion of their animals, which is calculated from farm specific data and some normative values. Some variables, like intake of grazed grass or roughage, are hard to measure. A data driven approach could help finding structures in data, and identifying key factors determining N excretion. The aim of this study was to benchmark machine learning methods such as Bayesian Network (BN) and boosted regression trees (BRT) in predicting N excretion, and to assess how sensitive both approaches are on the absence of hard-to-measure input variables. Data were collected from 25 Dutch dairy farms. In the period 2006–2018, detailed recordings of N intake and output were made during 6–10 weeks distributed over each year. Variables included milk production, feed intake and their composition. Calculated N excretion was categorized as low, medium, and high, with limits of 300 and 450 g/day/animal. Accuracy of prediction of the farm specific N excretion, and distinguishing the low and high cases from the medium ones, was slightly better with BRT than with BN. Leaving out information on intake during grazing did not negatively influence validation performance of both models, which opens opportunities to diminish data collection efforts on this aspect. Further analyses are required to confirm these results, such as cross-validation.

Keywords: Bayesian networks · Boosted regression trees · Nitrogen excretion · Dairy cows

1   Introduction

Agricultural policy, at least in the Netherlands, is in a transition. Since the second world war, the focus has been on producing as much food as possible in a highly efficient way. Favorable conditions combined with high-level knowledge and expertise, have put Dutch agriculture in a world leading position. However, drawbacks of this policy, become clear in regards to environmental impacts of agriculture and mineral surpluses due to factors such as high imports of animal feed. In 2018, the Dutch minister of agriculture presented a new policy [1], positioning Dutch agriculture as the front runner in circular agriculture, with minimal losses and minimal inputs of artificial inputs (fertilizer, pesticides) and scarce resources (e.g., phosphate, potassium).

Wageningen UR researchers have developed several tools and collected several datasets during the past decades to support farmers in nutrient management to meet the legal requirements. Examples of tools are the Annual Nutrient Cycling Assessment [2], which includes the whole nutrient cycle, or the Excretion Assessment [3], focusing on nutrients excreted in the manure by the animals. The Excretion Assessment is a tool that farmers can use to proof the actual excretion of their animals, in comparison to the excretion standards. The government accepts the results of this tool in order to calculate the manure export of a farm. This tool is based on farm specific data and some normative values modelled in a mechanistic way. However, some variables including intake of grazed grass or roughage, are hard to measure directly, and therefore are measured indirectly or estimated. A data driven approach could help finding structures in data, and identifying key factors determining N excretion. This could lead to possible exclusion of estimates and normative values for hard-to-measure input variables, or to focus on more detailed assessments, on field, group, or individual animal level. This would make the process of monitoring nutrient management more efficient and could be used to elaborate tools with prediction of the effects of possible management interventions. Better monitoring and prediction could support farmers and advisors, in making informed decisions on nutrient management, such as changes in feed ration or manure application.

An earlier study has proven that machine learning can help to predict grass and crop yields at field level, to support decisions on manure application rates [4]. However, it considered only one part of the nutrient circle, but decisions have to be made in more parts of the circle, such as animal nutrition and management. Therefore, in this study we will focus on the excretion of N from the animals, predicted from feed intake and production variables. To make models more explainable, a Bayesian network (BN) approach was applied, which delivers a graphical model that presents probabilistic relationships among a set of variables [5–8]. Next to BN, a less explainable model, boosted regression trees (BRT), was applied, which was also used in the previous study [4].

The aim of this study was to compare BN and BRT in predicting N excretion of dairy cows, and to assess how sensitive both approaches are on the absence of hard-to-measure input variables.

2 Materials and Methods

2.1 Dataset

Data used for this study originated from the project called 'Cows and Opportunities'. In this project, a group of 16 Dutch dairy farmers, together with researchers, searched for opportunities for sustainable and socially accepted agriculture [9]. The farms varied in intensity, scale, soil type, and style of farming. The main aim was to implement expected environmental legislation to monitor environmental, technical, and economic effects at farm level. In this way, each of these farms are a kind of pilot farm.

Data from years 2006 to 2018 were used, and originated from 25 different farms that were part of the project for different periods of time. On each farm detailed recordings were made during 6 to 10 weeks, distributed over the year. Between 14 to 16 farms were present in the dataset per year, which resulted in 108 to 151 weeks with detailed recordings per year. The final dataset contained 1,640 records.

Variables used concerned; milk production, milk composition, feed intake, and feed composition, for milk producing animals, on an average per cow basis. Milk production and milk composition were measured as part of the national milk production registration, according to ICAR regulations. Feed intake was measured through weighing of offered feed and refusals. Feed composition was determined by routine laboratory analyses of feed composition of all feeds. These variables are also used for calculating the N excretion based on formulas described by Oenema et al. [10], which were used as 'observed' values to be predicted with machine learning models. N excretion was categorized as low, medium, or high with limits of 300 and 450 g N per day per animal.

2.2 Machine Learning Models

In this study, two machine learning techniques were applied and benchmarked, namely BN and BRT. Decision tree induction is one of the basic machine learning techniques, is robust against irrelevant input variables, and is able to handle missing values [11, 12]. To alleviate the main disadvantage of decision trees, namely its inaccuracy in prediction, we used the iterative method called boosting. Generalized boosted regression using the Gradient Boosting Machine (h2o.gbm function (h2o version 3.20.0.2)), was applied in this study. BRT, lacked the possibility to clearly show relationships between variables in the model and, subsequently some interpretability. Therefore, BN was applied, which is a class of probabilistic model originating from Bayesian statistics and decision theory combined with graph theory. BN has the ability to integrate different data sources and types such as expert knowledge, measurement data, and feedback experience via Bayes theorem [7, 8]. Hugin software (version 8.7), was used to develop the BN model. Both models were applied on the same training and validation datasets. The validation dataset was a 20% random sample from the original dataset, and the remaining 80% was used for training. Model performance was evaluated by reporting confusion matrices and accuracy indicators. Accuracy was calculated in the conventional method, as well as for the low and high categories only (LHacc). Furthermore, a false alert rate (FAR) was calculated, representing the number of false alerts where an actual low case was predicted to be high or vice versa, over the total number of

low and high predictions. All data processing and analyses, except for the BN development, were performed in RStudio (version 1.1.463 running R version 3.6.1).

3   Results and Discussion

Categorization of records concerning N excretion was rather imbalanced with over 88% of cases being medium. Randomization of records over training and validation sets, resulted in a prevalence of 89% medium in the training set and just 85% in the validation set. Especially notable was records categorized high, were over-represented in the validation dataset (Table 1).

**Table 1.** Number of low, medium, and high cases, and percentage of medium in training and validation datasets

Dataset	# low	# medium	# high	Percentage medium
Training	80	1171	61	89.3
Validation	22	280	26	85.4

Performance of BRT models was better than for BN models, with an overall accuracy for BRT models, just above the prevalence of medium in the validation dataset. The accuracy of predicting for high and low was 8.3% and 10.4%, for BN and BRT respectively (Table 2), both with 0% false alerts for low and high. For both models, this means that they have difficulty distinguishing the low and high cases from the medium ones, and predicting some of these cases corresponds with classifying slightly more (BN) or fewer (BRT) true medium cases incorrectly. When the first hard-to-measure variable, dry matter (DM) intake from grazing, was excluded from the validation set, performance improved slightly for BRT and stayed equal for BN, which means that there was no additional value in this variable for the model (Table 2). Additionally leaving out DM intake from silages and hay, decreased the performance of BRT compared to leaving out only DM intake from grazing, but improved the overall accuracy of the BN model. However, for both BRT and BN, the accuracy of predicting high and low decreased, which means that models tend to predict all records as medium (overall accuracy approaching prevalence of medium in validation dataset of 85.4%, see Table 1), when information on DM intake is left out.

Table 2. Overall accuracy (Accur.), and accuracy for low and high categories (LHacc) of Bayesian Networks (BN) and Boosted Regression Tree (BRT) models, for predicting N excretion with all variables present in the validation dataset or when excluding variables; dry matter intake (DMint) information of grazing, (grass and maize) silages, and hay.

Model description	BN		BRT	
	Accur.	LHacc	Accur.	LHacc
All variables	82.9	8.3	85.7	10.4
All, except DM_{int} grazing	82.9	8.3	86.9	10.4
All, except DM_{int} grazing, silages	84.1	2.1	86.3	6.3
All, except DM_{int} grazing, silages, hay	84.5	2.1	86.0	6.3

Protein content of grazed grass consistently appeared to be the most important variable in all models, often closely followed by variables concerning DM intake of grazed grass, grass or maize silage or their protein content, and milk production. For BRT models, only variable importance could be reported, whereas for BN models also relationships could be made visible (Fig. 1). A remarkable point that should be investigated further is that the DM intake from mineral supplements, which constitutes a minor part of the diet, was rather important in the BN model, with many connections to other variables.
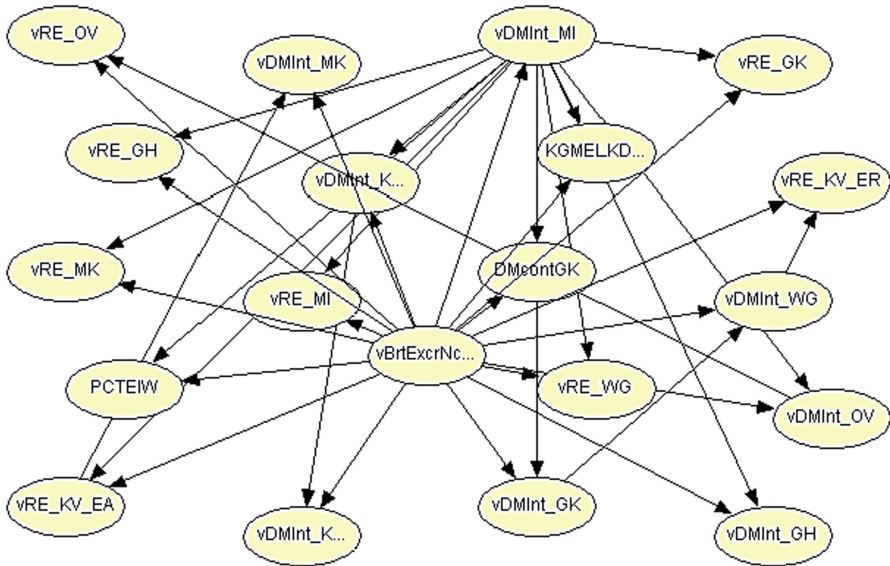
Fig. 1. Graphical representation of Bayesian Network model predicting N excretion (vBrtExcrNc…). (vDMInt = DM intake; vRE = protein content; DMcont = dry matter content; GH = grass hay; GK = grass silage; MI = minerals; MK = maize silage; KV_EA = low-protein concentrates; KV_ER = high-protein concentrates; OV = other feeds; WG = grazing; KGMELKD… = milk production level; PCTEIW = protein content of milk)

Additional analyses were performed to check the consistency of the results. We checked with BRT the effect of leaving out DM intake from both the training and validation dataset. When leaving out DM intake from grazing, this resulted in an increased overall accuracy (87.8%), as well as an increased accuracy for the low and high classes (18.8%), probably indicating that data on DM intake from grazing were noisy data and hampered the development of the prediction model. When also DM intake from grass and maize silage were excluded, accuracy dropped below that of the model containing all variables.

Results in this paper were based on a single split of the dataset and only two machine learning techniques were applied, with their default settings. These results, therefore, need to be confirmed with additional analyses, for example by applying cross-validation or different machine learning techniques.

4   Conclusions

In this study, we developed and benchmarked two machine learning methods (BN and BRT), to predict N excretion of dairy cows. Based on the initial evaluation of models for predicting N excretion from dairy cattle, BRT performed slightly better than BN, both with respect to overall accuracy as well as distinguishing high and low cases from medium ones. We conclude that both BN and BRT have difficulty distinguishing the low and high cases from the medium ones. Further analyses of the relationship between variables in the model, is better facilitated by the graphical representation of the model for BN, whereas for BRT only variable importance is available. Leaving out information on DM intake of grazed grass did not negatively influence the validation performance of both models, which opens opportunities to diminish data collection efforts on this aspect.

References

1. MinLNV: Landbouw, natuur en voedsel: waardevol en verbonden - Nederland als koploper in kringlooplandbouw (in Dutch). Ministry of Agriculture, Nature and Food Quality, The Hague, The Netherlands (2018)
2. Aarts, H.F.M., et al.: Quantifying the environmental performance of individual dairy farms - the Annual Nutrient Cycling Assessment (ANCA). Grassl. Sci. Eur. **20**, 377–380 (2015)
3. RVO: Handreiking bedrijfsspecifieke excretie melkvee. Rijksdienst voor Ondernemend Nederland, 60 p. (2019)
4. Mollenhorst, H., et al.: Field and crop specific manure application on a dairy farm based on historical data and machine learning (2019, submitted)
5. Marvin, H.J., et al.: Application of Bayesian networks for hazard ranking of nanomaterials to support human health risk assessment. Nanotoxicology **11**(1), 123–133 (2017)
6. Cheng, J., et al.: Learning Bayesian networks from data: an information-theory based approach. Artif. Intell. **137**(1–2), 43–90 (2002)
7. Bouzembrak, Y., et al.: Application of Bayesian Networks in the development of herbs and spices sampling monitoring system. Food Control **83**, 38–44 (2018)

8. Bouzembrak, Y., Marvin, H.J.P.: Impact of drivers of change, including climatic factors, on the occurrence of chemical food safety hazards in fruits and vegetables: a Bayesian Network approach. Food Control **97**, 67–76 (2019)
9. Cows and Opportunities homepage (in Dutch). www.koeienenkansen.nl. Accessed 22 Oct 2019
10. Oenema, J., et al.: Toetsing van de Kringloopwijzer - Gemeten en voorspelde stikstof- en fosfaatproducties van mest en gewas. Wageningen University and Research, Wageningen, The Netherlands. p. 84 (2017)
11. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29** (5), 1189–1232 (2001)
12. Witten, I.H. Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Series in Data Management Systems, 2nd edn. Elsevier/Morgan Kaufmann, San Fransisco, CA (2005)