Predictive Modeling of Methane CH4 Emissions from Dairy Barnyards Using Surface and Feeding Data

Group Barnyard Emission   Sabrina Song   Alice Wang

Abstract

This study aims to predict CH4 emissions in dairy barnyards using surface and animal feed characteristics. We analyzed a dataset from 9 experimental barnyards and evaluated the performance of different machine learning models. Results showed that the random forest model had the best performance with an RMSE of 6.969 and an R-squared value of 0.373. The top five predictors of CH4 emissions were ambient temperature, dry matter content of the feed, measurement month, soil nitrate concentration, and the previous year's CH4 emissions. These findings can help inform strategies to mitigate CH4 emissions in dairy barnyards and contribute to the development of sustainable agricultural practices.

1 Introduction

Greenhouse gas(GHG) emissions have become a major contributor to global temperature rise. The Intergovernmental Panel on Climate Change (IPCC) projected different warming scenarios to demonstrate the need for reducing GHG emissions. According to statistics from the World Greenhouse Gas Emissions Report, the agriculture sector accounts for about 18.3% of global emissions [1]. Within agriculture, the management of livestock and soil are major sources of GHG emissions, particularly methane and nitrous oxide [2]. Therefore, accurately predicting GHG emissions from animal husbandry systems is crucial for developing effective mitigation strategies. In this project, we aim to predict GHG emissions from experimental barnyards. Specifically, we investigate how animal feed and soil surface affect GHG emissions using machine learning models. By understanding the relationships between these variables and GHG emissions, we can develop prediction models that inform sustainable farming practices and reduce the environmental impact of livestock production.

2 Background

Previous studies have demonstrated that agriculture significantly impacts the environment and human health [3]. Researchers like Alan Rotz have estimated GHG emissions from dairy farms [4], while others like Manuel R. Rodriguez have used evolutionary algorithms and artificial neural networks to predict CO2 emissions from weaned piglet farms [5].

In our project, we also use predictive models to estimate CH4 emissions with data from the Dairy Coordinated Agricultural Project (Dairy CAP). The project constructed 9 rectangular barnyards to investigate the effects of surface materials and cattle corralling on GHG fluxes. From October 2010 to October 2015, dairy heifers were placed in these barnyards for approximately seven-day periods, four times per year, and received the same feed across all 9 barnyards. During the study, surface samples were collected from different barnyards, and GHG fluxes were measured for two days before and after heifers were corralled in the barnyard [6]. The objective of our study is to develop a prediction model for CH4 emissions using surface data and heifer feeding data.

3 Data Processing

3.1 Datasets and features

To prevent overfitting the model with excessive features, we only use three datasets here since they cover our predictors and outcomes comprehensively. This section presents an overview of the datasets and their features, with detailed descriptions of each feature provided in the appendix.

Surface Dataset
The Surface dataset contains information about randomly collected surface samples in each barnyard for a given sampling date. It includes the surface type (SOIL, SAND, or BARK), water content, pH value, and more.

TMR Dataset
The TMR (Total Mixed Ration) dataset provides information about heifer feeds for a given feeding date, including Dry Matter, Total Nitrogen Content (TNC), Neutral Detergent Fiber (NDF), and other nutritional components. Heifers were fed once per day in the morning during the placement period and received the same feeds for all barnyards. Refusals were also recorded for each barnyard in the Intake dataset. However, as most observations in the Intake dataset have zero refusals, intake does not vary much across different barnyards. Therefore, we only use the TMR dataset for the following analysis as it reflects similar information about feeds as the Intake dataset.

GHG Dataset
The GHG dataset contains the target output. GHG emissions were measured around two days before heifers were placed in the barnyard and two days after they were moved off. On each measurement day, samples were collected at two randomly selected locations within each barnyard[1], and the measurement

[1] Frame feature in GHG dataset

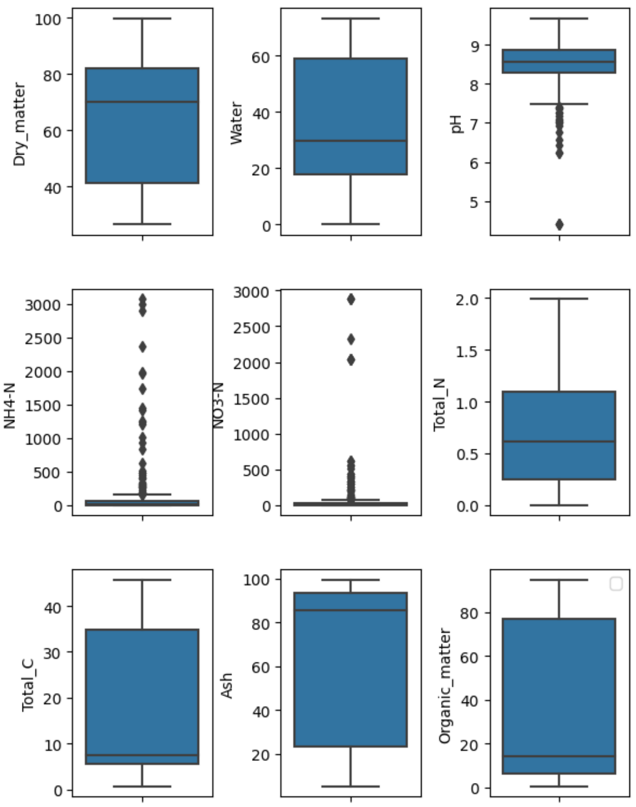


Figure 1: Box-plot for numeric features in Surface Dataset

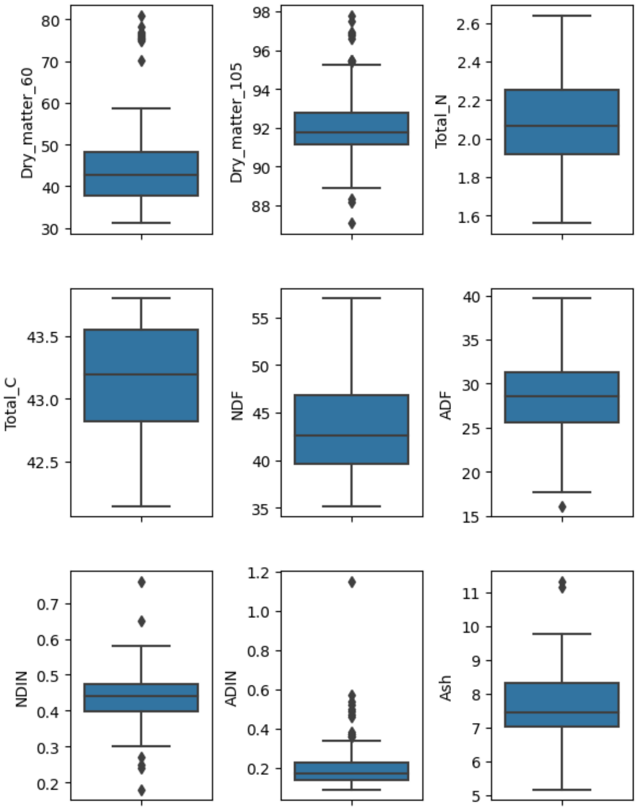


Figure 2: Box-plot for numeric features in TMR Dataset

took place both in the morning (M) and afternoon (A).

3.2 Preprocessing Datasets

This section explains the methods used to address missing values in each dataset.

Surface

We plot a box-plot for each numeric feature except sampling date and barnyard number in Figure 1. Based on the plot, the median for features Total_C, Ash, and Organic Matter is closer to one end of the box. Furthermore, there are significant outliers for features pH, NH4-N, and NO3-N. Most features in the surface dataset are skewed or have significant outliers, so we use the median to impute missing values. Specifically, we group the dataset by the sampling month and barnyard number to calculate a median value per month per barnyard.

After median imputing, we still find that NaN values exist, which turns out we have no data for features Total_N and Total_C during all of December. Additionally, we checked there are no corresponding GHG emission data in December as well. Therefore, we dropped the December surface sample from the dataset. The remaining missing values are addressed later in section 3.3.

TMR

We also plot box plots for each numeric feature in the TMR dataset (Figure 2). The median for most features is close to the center of the box, and there are some outliers, but not as significant as in the surface dataset. This suggests that the TMR data is more symmetric and appropriate to use the mean imputing.

As we mentioned in the background section, heifers were placed into barnyards at regular intervals, and the TMR feeding data reflect the heifer placement period. During September and October 2010, heifers were placed into barnyards for 33 days. Starting in April 2012, heifers were placed in each barnyard for 7-day periods, four times per year. In total, we have 19 heifer placement periods from 2010 to 2015.

Since all heifers received the same feeds regardless of the barnyard number, we group the data by heifer feeding period to calculate the average feeding data. For instance, we compute the mean feeding data for the period from September 22, 2010, to October 24, 2010, rather than grouping by month, which we believe is a more reasonable approach.

After analyzing the mean TMR table, we observed that ‘Total_C‘ had 18 NaN values out of the 19 placement periods, indicating that there were no records of ‘Total_C‘ on any day during these placement periods. Thus, we have decided to drop the ‘Total_C‘ feature as it does not provide much information. Similarly, we will also drop the ‘ADIN‘ and ‘NDIN‘ columns. After imputing the missing values with mean data averaged by heifer placement periods, the TMR dataset is now ready for further analysis.

GHG

The original GHG dataset consists of 2,167 observations, but many of the gas measurements, such as NH3_ugN, NH3_mg, and CO2_eq, have more than half of their observations as NaN. Therefore, we have decided to limit our target variable to CH4_mg only, as methane is the major GHG emission in the agriculture sector. We will drop the features CH4_mgC as well, as they are correlated with our target outcome CH4_mg. We plot boxplots for the remaining numeric features, Temp and CH4_mg, and find that the target outcome has many extreme values. Therefore, we use the median to impute missing values.

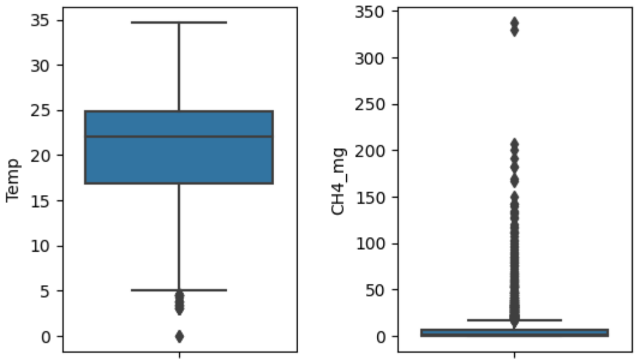


Figure 3: Box-plot for numeric features in GHG Dataset

3.3 Merging Datasets

Initially, we planned to calculate the variation in GHG by subtracting the BEFORE values [2] from the AFTER values[3]. This also allows us to merge the surface and TMR with the GHG dataset if their measurement dates fall within the same GHG measurement period. However, upon examining the dataset, we discover that some GHG measurements only have BEFORE data, while others only have AFTER data. Even when we keep measurements with both BEFORE and AFTER data, the GHG measurement period does not align with the TMR feeding period. For example, GHG were measured in both July and August 2011, but we have no TMR records during these two months. It is unclear whether the researcher did not record the feeding data during this time or if they measured the GHG regardless of the presence of heifers. Additionally, after removing all no-paired BEFORE-AFTER data, the dataset size reduces significantly to just 181 rows. Therefore, we decide to use the direct measurements of GHG emissions without calculating variations.

Merging Surface with GHG

Compared to GHG, surface data has a lower sampling frequency, usually only collected once per month with one sample per barnyard. In contrast, GHG was measured at different times and frames within each barnyard[4]. As a result, while the GHG varies due to different measurement times and frames, the surface data may remain constant. To merge the two datasets accurately, we follow these principles:

1. Whenever the GHG and Surface datasets have exactly the same date and barnyard number, we merge them together. In other words, researchers both collected surface samples and measured the gas fluxes on the same day.

2. If the GHG and Surface datasets have samples collected in the same month and year(different dates), we merge them based on the barnyard number. We assume that the surface data collected during that month is representative of the general soil condition for the entire month.

3. If there is a surface sample collected only in the same month as the GHG measurement(different date and year), we use the median surface data in the preprocessing step (section 3.2) to merge based on month and the barnyard number.

Merging TMR with GHG

The TMR data at any given date is the same across barnyard for all heifers, but GHG measurement varies with barnyard, frame, and sub_frame. Thus, we can find varying GHG emissions across different barnyards on a given day, but they all with the same feeds. This might lead to errors in subsequent predictions. However, since the feeding data is all we have, we make efforts to merge the datasets as accurately as we can. Our first criteria for merging are:

Merge GHG and TMR data if there is the same measurement date.[5]

We then try two methods to merge the remaining TMR data:

1. Merge the nearest TMR into the GHG dataset for each GHG measurement date. If there is no nearest feeding date (date difference > 10 days), merge with the mean feeding data averaged over heifer placement periods.

2. Directly use the mean feeding data averaged over heifer placement periods if it has the same date and month as the GHG date.

The result shows Method 2 has better R-square and RMSE in a simple linear regression, so we will use this method going forward.

Aggregated Dataset

We observe 207 NaNs in the aggregated dataset for Ash and Organic_matter. Upon investigation, we find that no surface sample data for these variables were collected on any sampling day in August. Similarly, we found 286 for TMR data because there was no feeding data collected at all during certain months such as 2011-07, 2011-08, 2011-09, etc. To clean the dataset, we removed these NaN values.

4 Models and Results

4.1 Baseline

With the cleaned dataset, we first build a simple linear regression using a 20-80 test split as the baseline model. However, the results are poor, with an R-square of only 0.06748 and an

[2] GHG measured two days before heifers were placed in barnyards

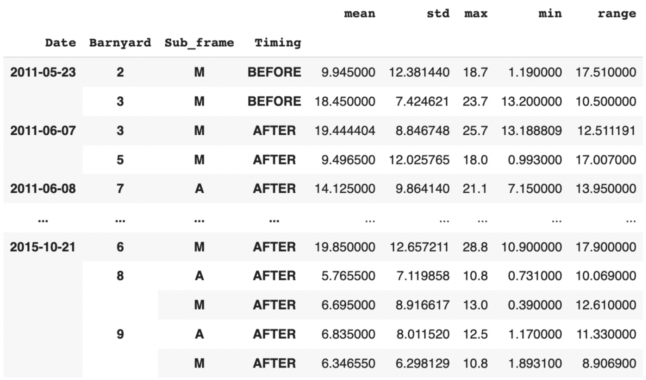
[3] GHG measured two days after heifers were moved off from barnyards

[4] In each barnyard, researchers measured the gas fluxes in different frames(A, B, C, or D), and different times(Morning or Afternoon)

[5] Since GHG is measured before and after the heifer placement period, theoretically there should not be any feeding date that is exactly the same as the GHG measurement date. However, we did find several dates that are the same in this case.

RMSE of 29.73. To improve our model, we take the following steps:

1. Adding a previous year feature. As shown in Figure 4, we analyze the statistics of our target variable CH4. Even when grouping by measurement date, barnyard number, sub_frame, and timing, there are still many observations with a standard deviation larger than 5. We believe there might exist significant predictors that haven't been covered within our datasets. We therefore decide to add an extra feature: the measurement results from the previous year. By doing so, we start our predictions from 2012. This led to a significant improvement in our model performance, with an R-square of 0.2805 and an RMSE of 22.53 (See Figure 7 and Figure 8).



				mean	std	max	min	range
Date	Barnyard	Sub_frame	Timing					
2011-05-23	2	M	BEFORE	9.945000	12.381440	18.7	1.190000	17.510000
	3	M	BEFORE	18.450000	7.424621	23.7	13.200000	10.500000
2011-06-07	3	M	AFTER	19.444404	8.846748	25.7	13.188809	12.511191
	5	M	AFTER	9.496500	12.025765	18.0	0.993000	17.007000
2011-06-08	7	A	AFTER	14.125000	9.864140	21.1	7.150000	13.950000
...	...	...	...	...	...	...	...	...
2015-10-21	6	M	AFTER	19.850000	12.657211	28.8	10.900000	17.900000
	8	A	AFTER	5.765500	7.119858	10.8	0.731000	10.069000
		M	AFTER	6.695000	8.916617	13.0	0.390000	12.610000
	9	A	AFTER	6.835000	8.011520	12.5	1.170000	11.330000
		M	AFTER	6.346550	6.298129	10.8	1.893100	8.906900

Figure 4: CH4 statistics group by Date, Barnyard,Sub_frame, Timing

2. Removing outliers. We use cat-plots, box-plots, and histograms to check the distribution of our target output CH4. (See Figure 5 and Figure 6). Barnyard 1 and barnyard 7 have some extreme values as high as 350, while the median of CH4 is between 0 and 10. We decide to restrict our target output to those below 50 and remove outliers. This resulted in a significant improvement in our model performance, with an R-square of 0.3455 and an RMSE of 7.1214 (See Figure 7 and Figure 8).

3. Averaging over frames. We also consider averaging the dataset over frames (A, B, C, D), as researchers have measured GHG in different frames per barnyard to reduce variation. However, this does not lead to better results. Therefore, we decide to stick with the results of part 2.

4.2 Models

In this section, our goal is to develop and compare seven different predictive models. These models are selected based on their ability to handle specific challenges associated with prediction, including high dimensionality, non-linearity, and multicollinearity. By comparing the performance of these models, we hope to identify the most accurate and interpretable approach for predicting CH4 emissions. Our seven predictive models are as follows:

1. Ridge regression - this model is suitable for handling high-dimensional data and multicollinearity. For example, in our surface dataset, water content and dry matter may be highly correlated.

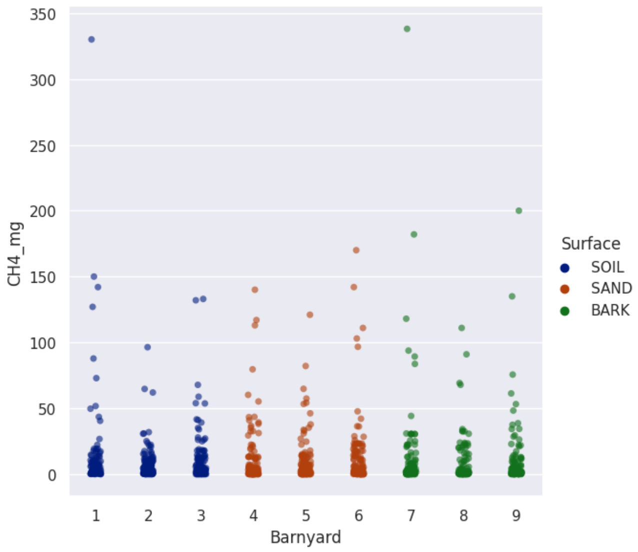


Figure 5: Box-plot for numeric features in GHG Dataset

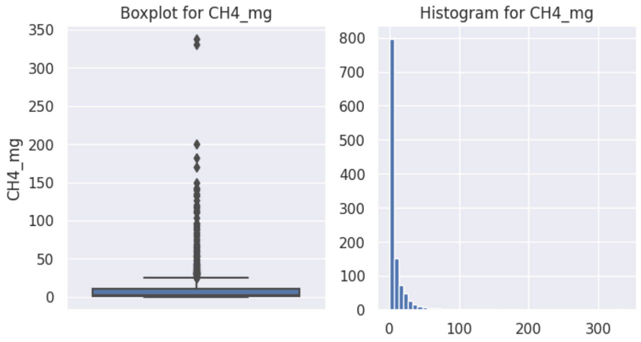


Figure 6: Box-plot for numeric features in GHG Dataset

2. Lasso regression - this model is also suitable for high-dimensional data and multicollinearity. Additionally, it performs feature selection by shrinking the coefficients of less important variables to zero.

3. Decision tree - this model can capture non-linear relationships between features and provide a graphical representation of all possible solutions.

4. Random forest - this model is a bagging method based on decision trees. It can capture complex nonlinear relationships among features and improve predictive performance.

5. Adaboost - this model is an ensemble method based on decision trees, which combines multiple weak models to improve predictive accuracy.

6. XGboost - this model is an ensemble method similar to Adaboost but uses a more powerful gradient-boosting algorithm.

7. Deep neural network - our model consists of multilayer perceptrons with the ReLU activation function. DNNs can capture more complex non-linear relationships between variables, making them suitable for datasets with high complexity.

4.3 Evaluation Metric and Results

To assess the performance of each model, we utilize three different evaluation metrics: R-squared, RMSE, and MAE. R-squared is a measure of how well a model fits the data and predicts new data. RMSE represents the average distance between predicted and actual values, while MAE measures the average magnitude of prediction errors.

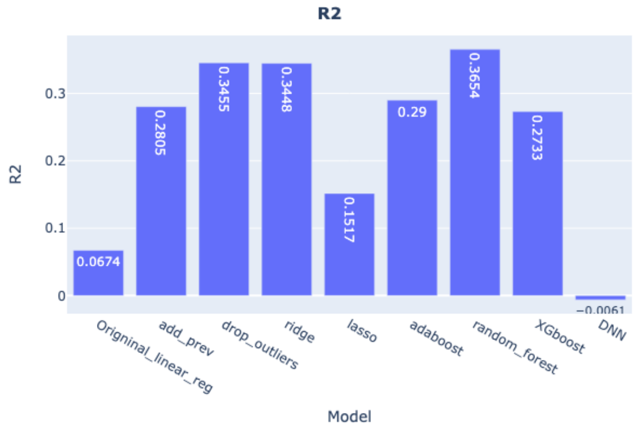


Figure 7: Box-plot for numeric features in GHG Dataset

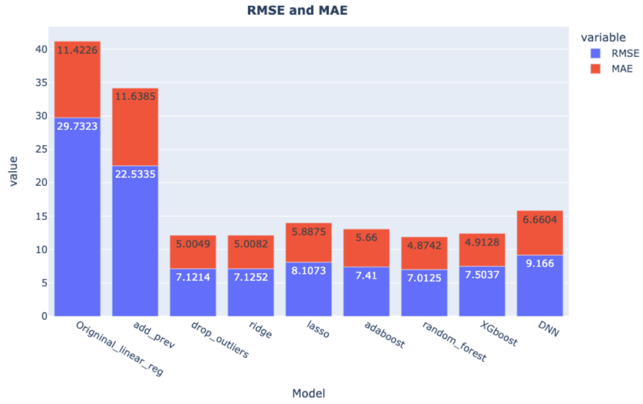


Figure 8: Box-plot for numeric features in GHG Dataset

According to the performance metrics displayed in Figure 7 and Figure 8, the random forest model demonstrates the best performance among all models with the highest R-squared value of 0.3654, the lowest RMSE value of 7.0125, and an MAE value of 4.8742. The random forest model might have performed the best because:

1. Non-linearity: random forest model captures the non-linear relationships between our features and CH4, making it more flexible than linear models like Ridge and Lasso regression.

2. Robustness to outliers: random forest model is less sensitive to outliers. Even though we have already dropped some outliers in section 4.1, we still have some outliers indicated by the box plot in Figure 6.

3. Ensemble learning: random forest employs an ensemble of decision trees, which can enhance its accuracy by reducing overfitting and increasing generalization.

It's worth noting that the DNN model has a negative R-squared and higher RMSE, indicating it might overfit the training data. DNN might be so complicated that captures noise or irrelevant information, resulting in poor generalization performance on the test data.

4.4 Hyperparameters

We use the best-performing model, the random forest, to optimize its hyperparameters with RandomizedSearchCV. The hyperparameters that give the best results are: n_estimators=200 , min_samples_split=2, min_samples_leaf=5 max_features= 'sqrt', max_depth=20.

We use the negative MSE as the scoring metric for RandomizedSearchCV, as the algorithm aims to maximize the score and we want to minimize the loss function.

After tuning the hyperparameters, the random forest model's performance improves with an RMSE of 6.969 and an R-squared value of 0.373, indicating that it explains a high percentage of the variance in the target variable and provides more accurate predictions. The MAE value is also improved to 4.830.

4.5 Feature Importance

To identify the most influential predictors for CH4 emissions, we conduct a feature importance analysis using the tuned random forest model. The top 5 important features for predicting CH4 emissions are:

1. Temp (0.111594): Ambient temperature at measurement time. It is a key factor affecting the rate of microbial activity and methane production in anaerobic digestion processes.

2. tmr_Dry_matter_60 (0.096992): The dry matter content of the feed after drying at 60 degrees Celsius. This measure of the total solids in the feed affects the digestion efficiency and CH4 production.

3. Month (0.077539): Measurement month is the third most important feature, indicating a possible seasonal effect on CH4 emissions.

4. NO3-N (0.066352): Nitrate concentration in the surface sample, as it can act as a substrate for methane-producing bacteria and enhance CH4 production.

5. Prev (0.052168): The CH4 emissions in the same month, same barnyard in the previous year. It suggests that there are some hidden features that contribute to the CH4 emission pattern but haven't been covered in our dataset.

In addition, other important features (Figure 9) include the dry matter content of the feed at 105 degrees Celsius, the water content of the soil, pH, and various nutrient concentrations such as surface Total C, surface Total N, and NH4-N. These features may be related to feed digestion efficiency and soil microbial activity.

5 Discussions

Overall, we find that the random forest model performs best in predicting CH4 emissions, and we have identified the top 5 im-

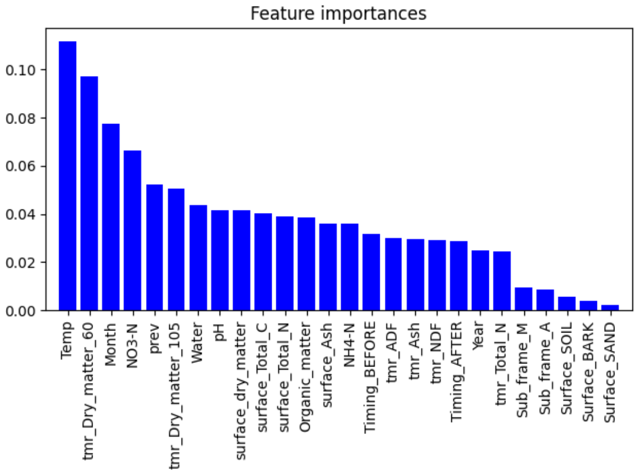


Figure 9: Feature importance

portant features using a tuned random forest model. However, there are some limitations to our study:

1. The presence of hidden features not captured in our study and experiment may impact CH4 emissions. As mentioned in section 4.1, GHG measurement results have large standard deviations. Factors such as microbial community composition, heifer quantities and weights, and manure storage may affect GHG emissions. Even with our tuned random forest model, only 37.3% of the variance in CH4 emissions is explained, indicating the presence of other important predictors.

2. Our dataset has a large number of missing values due to incomplete measurements during certain weather conditions, such as heavy rain, wind, and heat. Imputing missing values with mean or median values may introduce bias into the model, leading to less accurate predictions.

3. Our cleaned dataset is a merged dataset from three different subsets, with varying measurement frequencies and dates. As described in Section 3, we have used group mean or median values to represent the surface conditions or feed content in a given barnyard on a particular day. However, this assumption may not fully capture the true barnyard conditions, and the merged dataset may differ from the actual barnyard conditions.

In addition to the limitations of the dataset, there are several ways we can further investigate:

1. Hyperparameter tuning. We have only tuned the hyperparameters for our best-performing model. To ensure that our models are optimized, we can tune the hyperparameters for all models.

2. Predict other GHG emissions. While our current model predicts CH4 only, the GHG dataset also provides measurements for CO2, NH3, and N2O emissions. We can try to predict other greenhouse gases and compare the performance of the models.

3. Apply our model to real barnyards. While our data was collected from experimental barnyards, we can check whether our prediction model is applicable to other barnyards by measuring the feature data in real barnyards and comparing the predictions to actual emissions.

References

1. ”Climate Watch.” Climate Watch, World Resources Institute, 2023, https://www.climatewatchdata.org/.

2. Agriculture and greenhouse gas emissions. University of Missouri Extension. (n.d.). Retrieved May 5, 2023, from https://extension.missouri.edu/publications/g310

3. Lake, Iain R et al. “Climate change and food security: health impacts in developed countries.” Environmental health perspectives vol. 120,11 (2012): 1520-6. doi:10.1289/ehp.1104424.

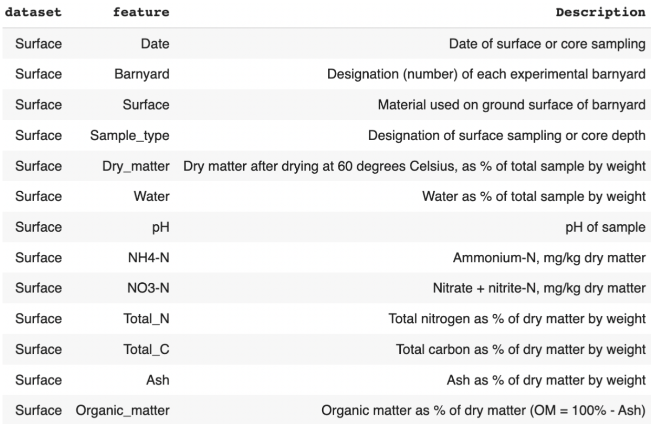
4. Rotz, C. A. (2018). Modeling greenhouse gas emissions from dairy farms. Journal of Dairy Science, 101(7), 6675–6690. https://doi.org/10.3168/jds.2017-13272

5. Rodriguez, Manuel R., Roberto Besteiro, Juan A. Ortega, Maria D. Fernandez, and Tamara Arango. 2022. ”Evolution and Neural Network Prediction of CO2 Emissions in Weaned Piglet Farms” Sensors 22, no. 8: 2910. https://doi.org/10.3390/s22082910

6. Powell, J. Mark; Vadas, Peter A.; Barford, Carol (2017). Data from: Gas emissions from dairy barnyards. Ag Data Commons. https://doi.org/10.15482/USDA.ADC/1401976. Accessed 2023-03-22

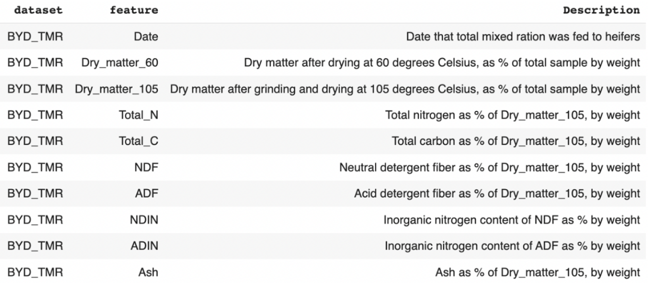
7. Powell, J. M., & Vadas, P. A. (2016). Gas emissions from dairy barnyards. Animal Production Science, 56(3), 355. https://doi.org/10.1071/an15598

Appendix



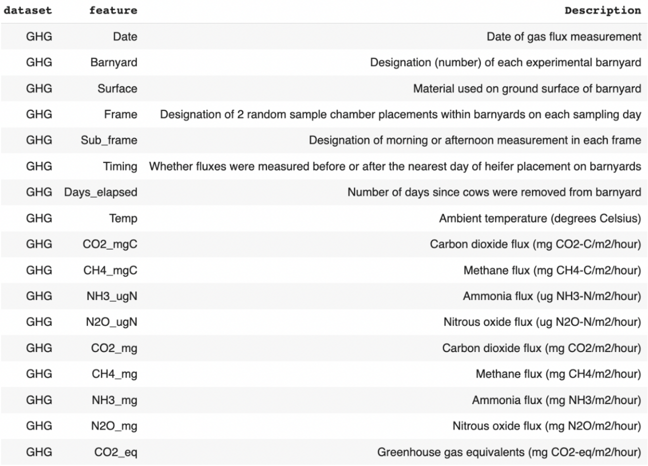
dataset	feature	Description
Surface	Date	Date of surface or core sampling
Surface	Barnyard	Designation (number) of each experimental barnyard
Surface	Surface	Material used on ground surface of barnyard
Surface	Sample_type	Designation of surface sampling or core depth
Surface	Dry_matter	Dry matter after drying at 60 degrees Celsius, as % of total sample by weight
Surface	Water	Water as % of total sample by weight
Surface	pH	pH of sample
Surface	NH4-N	Ammonium-N, mg/kg dry matter
Surface	NO3-N	Nitrate + nitrite-N, mg/kg dry matter
Surface	Total_N	Total nitrogen as % of dry matter by weight
Surface	Total_C	Total carbon as % of dry matter by weight
Surface	Ash	Ash as % of dry matter by weight
Surface	Organic_matter	Organic matter as % of dry matter (OM = 100% - Ash)

Figure 10: Surface Dataset Feature Description



dataset	feature	Description
BYD_TMR	Date	Date that total mixed ration was fed to heifers
BYD_TMR	Dry_matter_60	Dry matter after drying at 60 degrees Celsius, as % of total sample by weight
BYD_TMR	Dry_matter_105	Dry matter after grinding and drying at 105 degrees Celsius, as % of total sample by weight
BYD_TMR	Total_N	Total nitrogen as % of Dry_matter_105, by weight
BYD_TMR	Total_C	Total carbon as % of Dry_matter_105, by weight
BYD_TMR	NDF	Neutral detergent fiber as % of Dry_matter_105, by weight
BYD_TMR	ADF	Acid detergent fiber as % of Dry_matter_105, by weight
BYD_TMR	NDIN	Inorganic nitrogen content of NDF as % by weight
BYD_TMR	ADIN	Inorganic nitrogen content of ADF as % by weight
BYD_TMR	Ash	Ash as % of Dry_matter_105, by weight

Figure 11: TMR Dataset Feature Description



dataset	feature	Description
GHG	Date	Date of gas flux measurement
GHG	Barnyard	Designation (number) of each experimental barnyard
GHG	Surface	Material used on ground surface of barnyard
GHG	Frame	Designation of 2 random sample chamber placements within barnyards on each sampling day
GHG	Sub_frame	Designation of morning or afternoon measurement in each frame
GHG	Timing	Whether fluxes were measured before or after the nearest day of heifer placement on barnyards
GHG	Days_elapsed	Number of days since cows were removed from barnyard
GHG	Temp	Ambient temperature (degrees Celsius)
GHG	CO2_mgC	Carbon dioxide flux (mg CO2-C/m2/hour)
GHG	CH4_mgC	Methane flux (mg CH4-C/m2/hour)
GHG	NH3_ugN	Ammonia flux (ug NH3-N/m2/hour)
GHG	N2O_ugN	Nitrous oxide flux (ug N2O-N/m2/hour)
GHG	CO2_mg	Carbon dioxide flux (mg CO2/m2/hour)
GHG	CH4_mg	Methane flux (mg CH4/m2/hour)
GHG	NH3_mg	Ammonia flux (mg NH3/m2/hour)
GHG	N2O_mg	Nitrous oxide flux (mg N2O/m2/hour)
GHG	CO2_eq	Greenhouse gas equivalents (mg CO2-eq/m2/hour)

Figure 12: GHG Dataset Feature Description