MACHINE LEARNING FOR PREDICTING ENVIRONMENTAL IMPACTS IN INTENSIVE LIVESTOCK FARMING, OVERCOMING DATA CHALLENGES

Authors: L. Foschi, D. Pennino, M. Marconi, M. Barbanera, G. Grossi, F. Cesarini, C. Rossi, L. Bava, A. Vitali, N. Lacetera

Ms. Lorena Foschi
Research fellow at University of Tuscia
The need for more accurate Life Cycle Assessment (LCA) predictions is driving its integration with Machine Learning (ML) techniques. Current research focuses on using ML algorithms to improve LCA models, enhancing the accuracy and efficiency of environmental impact assessments.

**Context:**
- **Data Quality:** Accurate LCA predictions depend on high-quality data.
- **Missing Data Challenges:** A major challenge in LCA, caused by incompleteness, measurement errors, or data collection limitations.
- **Traditional Methods:** Mean- or median-based approaches for handling missing data are often limited and can result in biased or unreliable outcomes.

**Problem:**
- **ML Potential:** Machine learning techniques offer sophisticated solutions for data completion and imputation, enhancing the efficiency and reliability of LCA analyses.

**Focus of the study:**
Selection between ML regression techniques to improve the accuracy and reliability of data predictions.
METHODOLOGY

Approach Overview

- The methodology helps select the optimal machine learning algorithm for predictions and handling missing values.
- 8 regression algorithms were considered
- Presented as a decision tree for simplicity and usability.
METHODOLOGY

Methodology application

• Considered dataset: data from ~100 dairy farms in Italy/Europe, focused on Carbon Footprint analysis
• Key inputs for CF analysis: Fat and Protein Corrected Milk (FPCM) produced, herd composition, cultivated land area, and diet details (amount and origin of forage and concentrates), energy and fuel.
• Dataset was decomposed in 3 subdataset: Soy meal, herd size, total feed.

Regression model evaluation

Root Mean Square Error (RMSE) was used.
• Standard deviation of prediction errors, indicating the spread of residual values.
• RMSE normalized between 0 and 1 for comparison.
# RESULTS AND DISCUSSION

### Table 1. Normalized RMSE of three different datasets.

<table>
<thead>
<tr>
<th>Dataset/ML model</th>
<th>Linear</th>
<th>Ridge</th>
<th>Lasso</th>
<th>Polynomial</th>
<th>Decision tree</th>
<th>Random forest</th>
<th>Support vector</th>
<th>Gradient boosting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soy Meal</td>
<td>0.118</td>
<td>0.116</td>
<td>0.116</td>
<td>0.114</td>
<td>0.114</td>
<td>0.105</td>
<td>0.121</td>
<td>0.115</td>
</tr>
<tr>
<td>Heard Size</td>
<td>0.023</td>
<td>0.019</td>
<td><strong>0.020</strong></td>
<td>0.033</td>
<td>0.088</td>
<td>0.091</td>
<td>0.023</td>
<td>0.081</td>
</tr>
<tr>
<td>Total Feed</td>
<td>0.066</td>
<td>0.058</td>
<td>0.061</td>
<td>0.086</td>
<td>0.09</td>
<td>0.08</td>
<td>0.059</td>
<td>0.076</td>
</tr>
</tbody>
</table>

**Soy meal**

- **Dataset characteristics:**
  - Many heterogeneous inputs, presence of noise due to data collection limitations.
- **Model Selection:**
  - Random Forest Regression.

**Herd size**

- **Dataset characteristics:**
  - Simpler data collection, no significant noise, no complex input selection required.
- **Model Selection:**
  - Lasso Regression.

**Total feed**

- **Dataset characteristics:**
  - Correlated inputs, some multicollinearity, complex input selection required.
- **Model Selection:**
  - Ridge Regression.
CONCLUSIONS

**Key Findings:**

**Novel Approach Introduced:** Developed a systematic method to predict missing values in datasets

**Algorithm Selection:** The approach selects the optimal ML algorithm based on dataset characteristics, enhancing prediction accuracy and reliability.

**Effectiveness Demonstrated:**

**Testing and Validation:** Conducted rigorous testing on three distinct datasets, each influencing the environmental impacts of cattle milk farming.

**Performance:** Showcased the versatility and effectiveness of our approach in predicting missing values and improving dataset completeness.

**Practical Applicability:**

**Enhanced Utility:** Significantly increased the size of the dataset, improving its utility for robust environmental impact assessments.

**Implications:**

**ML and LCA Integration:** Demonstrated the potential of integrating ML techniques with LCA methodologies to address data-related challenges.

**Improved Assessments:** Enabled more robust and comprehensive environmental impact assessments, facilitating informed decision-making and sustainable practices.
THANK YOU
FOR YOUR ATTENTION

lorena.foschi@unitus.it