Peer Review Report

Project title: Machine Learning on Hyperspectral Data in Wheat Trials

Author: Nicolas Vuille-dit-Bille (vuillenicolas@gmail.com)

Date: 23.12.24

Reviewer: Galina Glousker (galina.glousker@epfl.ch)

Summary

The research presented by Nicolas Vuille-dit-Bille from Agroscope investigates the application of machine learning techniques on hyperspectral data to analyze and optimize wheat variety performance under varying nitrogen (N) treatments. The study is significant as it bridges advanced data analysis methods with practical agricultural challenges, aiming to reduce nitrogen fertilization costs while maintaining crop yield and quality. Below is a detailed evaluation of the study's context, methodology, findings, and implications.

Research Context and Objectives

The study focuses on winter wheat trials conducted across three Swiss sites (Changins, Goumoens, and Reckenholz) over two years (2021–2022). Five wheat varieties were analyzed under three main nitrogen treatments: none, reduced, and conventional. The primary objective was to identify wheat varieties that can perform well under reduced nitrogen availability, thereby addressing economic and environmental concerns. This research seeks to accurately assess responses of the plants to different environmental factors in the field conditions, with uncontrolled abiotic and biotic stresses.

Methodology

A robust experimental design was implemented, incorporating multiple repetitions and hyperspectral data collection at critical growth stages. Hyperspectral imaging, known for its ability to capture detailed spectral information, was processed and analyzed to extract meaningful insights. Key steps included:

1. *Data Cleaning*: Spectral bands with aberrant or zero values were removed to ensure data quality. The first notebook demonstrates an effective approach to preprocessing, using libraries such as `numpy` and `pandas` for filtering, normalization, and visualization. However, the code could be improved by:

- Encapsulation: Encapsulating repetitive preprocessing logic, such as filtering invalid spectral bands, into reusable functions to reduce redundancy and enhance maintainability.

- Vectorization: Replacing iterative loops with `numpy` vectorized operations to improve performance.
- Documentation: Adding markdown explanations for each preprocessing step, making the rationale behind decisions clearer to collaborators.

2. *Feature Reduction*: Principal Component Analysis (PCA) reduced 38 initial spectral features to 17, focusing on red-edge wavelengths (700–800 nm) due to their importance in characterizing wheat performance. As implemented in the second notebook, PCA was effectively applied with a clear interpretation of variance explained by components. However, testing the stability of PCA outputs across multiple random seeds would further validate its robustness. Additionally, automating feature selection thresholds (e.g., cumulative explained variance) could streamline this process.

3. *Data Exploration*: Techniques like K-means clustering, Gaussian Mixtures, and UMAP were employed to uncover patterns and relationships among treatments and varieties. The clustering workflows in the second notebook are functional but can be improved by:
- Parameter Rationale: Adding markdowns explaining choices for parameters such as the number of clusters, UMAP `n_neighbors`, and `min_dist`.
- Stability Testing: Running clustering algorithms with multiple random seeds to ensure consistent results and identify variability.

4. *Predictive Modeling*: A Random Forest model was developed to predict grain yield, highlighting the 700–710 nm spectral range as a critical predictor linked to nitrogen chlorophyll content. The third notebook implements supervised learning effectively, but the following enhancements are recommended:
- Cross-Validation: Use k-fold cross-validation or repeated random splits to assess model performance more robustly and mitigate overfitting.
- Evaluation Metrics: Beyond accuracy, include precision, recall, F1-score, and ROC-AUC to capture a more nuanced view of model performance.
- Hyperparameter Tuning: Incorporate automated tuning methods such as `GridSearchCV` or `RandomizedSearchCV` to optimize model parameters like `max_depth` and `n_estimators`.
- Feature Importance Consistency: Validate the stability of feature importance rankings by training Random Forest models on multiple train-test splits or using different random seeds.

Key Findings

The study yielded several promising results:

1. *Hyperspectral Data Insights*: The red-edge spectral bands (700–800 nm) were pivotal in distinguishing wheat performance, aligning with their theoretical link to chlorophyll content and nitrogen availability.
2. *Random Forest Predictions*: The model demonstrated strong performance on the training data (R-squared = 0.955) but showed reduced accuracy on the test set (R-squared = 0.685), indicating potential overfitting. Incorporating repeated train-test splits or k-fold cross-validation could mitigate this issue.
3. *Agronomic Implications*: The analysis suggested that reduced nitrogen input could be viable for certain wheat varieties without compromising yield significantly. Additionally, other parameters such as straw yield, grain protein, and plant height were identified as potential factors for further exploration.

Strengths and Contributions

The integration of advanced machine learning techniques with agronomic research is a notable strength of this study. The use of hyperspectral data offers a high-resolution perspective on wheat variety performance, enabling precise identification of spectral features linked to key agronomic traits. By addressing environmental and economic challenges, the study contributes to sustainable agricultural practices.

Challenges and Limitations

Despite its strengths, the study faces challenges common to field-based agricultural research. The noisy nature of hyperspectral data, influenced by weather and other external factors, complicates analysis. The notebooks demonstrate strong methodology, but the inclusion of multiple random seeds for cluster stability and model evaluation would add rigor. Additionally, leveraging cross-validation for supervised learning would strengthen the generalizability of the findings. Suggestions like encapsulating preprocessing logic and incorporating automated hyperparameter tuning would further enhance the code quality and reproducibility of the workflows. The reduced performance of the Random Forest model on the test set raises concerns about overfitting and generalizability. Additionally, while the focus on hyperspectral data is justified, integrating other agronomic parameters like disease resistance and harvest index could enhance the model's explanatory power.

Recommendations

To build on the promising results, the following recommendations are proposed:

1. *Model Optimization*: Address overfitting by exploring alternative algorithms or incorporating regularization techniques. Splitting datasets based on environmental conditions (e.g., site or year) and running models with multiple random seeds could improve generalizability.
2. *Expanded Analysis*: Incorporate additional agronomic parameters, such as plant height and disease incidence, to contextualize hyperspectral insights and improve prediction accuracy.
3. *Cluster Robustness Testing*: In the unsupervised learning step, test cluster stability by running clustering methods (e.g., K-means, Gaussian Mixtures) across multiple random seeds and reporting metrics like silhouette scores.
4. *Cross-Validation in Supervised Learning*: Use k-fold cross-validation or repeated random splits to evaluate model performance, ensuring robustness and reducing reliance on a single train-test split.
5. *Feature Selection Consistency*: Validate feature importance (e.g., spectral bands) identified by Random Forest across multiple random seeds to ensure robustness of conclusions.
6. *Enhanced Documentation*: Improve markdown explanations in the notebooks to provide detailed rationales for parameter choices and preprocessing decisions. Encapsulate repetitive code into functions for efficiency and readability.
7. *Performance Monitoring*: Incorporate logging for long-running operations, such as clustering and model training, to better track progress and debug issues.

Conclusion

Nicolas Vuille-dit-Bille's study effectively demonstrates the potential of machine learning and hyperspectral data in agricultural research. By identifying critical spectral features and variety responses to nitrogen treatments, the research paves the way for sustainable crop management practices. Fixing random seeds during visualization and conducting robustness testing for generalizability are recommended improvements to add rigor to the analysis. While challenges like data noise and model overfitting remain, the study provides a solid foundation for future work integrating advanced analytics with practical agronomy. The findings have significant implications for reducing nitrogen fertilization costs and environmental impact, making this research a valuable contribution to sustainable agriculture.