




Submission

-  My Files
-  My Files
-  University

Document Details

Submission ID**trn:oid:::17268:78370639****Submission Date****Jan 9, 2025, 1:41 AM GMT+5:30****Download Date****Jan 9, 2025, 1:42 AM GMT+5:30****File Name****Part 6_x23217677_Answers to Examiner Questions.docx****File Size****132.2 KB****6 Pages****1,984 Words****11,354 Characters**

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.





A Hybrid Machine Learning Approach for Crop Classification, Yield and Fertilizer Prediction for Sustainable Agriculture

MSc Research Project
MSCDAD

Lynnet Grace Nakagiri
Student ID: X23217677

School of Computing
National College of Ireland

Supervisor: Harshani Nagahamulla

QUESTION 1: What are the key machine learning models used in agricultural prediction, and why did you choose SVR, XGBoost, and MLP for your study?

Machine learning has emerged as a cornerstone of modern agricultural prediction, offering powerful tools to analyze diverse and complex datasets. These models excel in yield forecasting, pest detection, and resource optimization, among other tasks, that are transforming agricultural decision-making. The Key models used include:

- a) Ensemble Methods such as Random Forest (RF) and Gradient Boosting Machines (GBM) enhance the prediction accuracy by combining multiple models. RF builds a multitude of decision trees for handling both numerical and categorical data and thus, is appropriate for crop classification and soil analysis. GBM, on the other hand, decreases the error in a sequential manner and hence is appropriate for regression tasks like yield prediction.
- b) Neural Networks and Deep Learning: Models in these categories perform very well in finding complex patterns in agriculture data. Some of the commonly used ones include:
 - Convolutional Neural Networks: These are designed for spatial data analysis. CNNs find wide applications in satellite imagery to monitor crop health, detect pests, and enable precision agriculture.
 - Long Short-Term Memory Networks-LSTMs: This is a form of RNN which learns from sequence dependencies found in time-series data, and that can be quite ideal for seasonal trend forecasting, and resource demands.
- c) Support Vector Machines: SVM is useful for classification and regression problems where either the relationship between the data is complicated or the sample size is small.
- d) Hybrid Models: The hybrid models combine the power of several approaches and integrate spatial, temporal, and other agricultural data for robust and comprehensive analysis.

These models lead the way in agricultural innovation, along with improvements in ensembling and deep learning, by allowing precise, data-driven insights into sustainable and efficient farming.

In this research, SVR, XGBoost, and MLP were chosen for the meta-model layer due to the following reasons:

- Support Vector Regressor is a kind of kernel-based method that can easily capture the nonlinear relationship in the data. Thus, it has robust predictions with less overfitting in smaller to medium-sized datasets. The inclusion of SVR was because it provides a good balance between complexity and accuracy, especially in regression tasks. Its ability to handle data variability makes it well-suited for agricultural prediction.
- XGBoost is the most well-known gradient-boosted decision tree algorithm; it is mainly appreciated for its speed, scalability, and handling huge datasets. It incorporates regularization techniques that enhance performance on regression tasks of yield and

fertilizer predictions. In this way, it can model complex nonlinear relationships and hence assures robust performance in precision agriculture.

- Multi-Layer Perceptron (MLP) is a feed-forward neural network that can model complex and nonlinear patterns. The structure and adaptability in multi-layers make it suitable for high-accuracy tasks such as the classification of crops and the prediction of fertilizers. The ability of MLP to learn intricate dependencies enables it to be efficient for agricultural applications.

These models further complement the basic models, that is, Random Forest, Gradient Boosting Regressor, and LSTM, using their outputs and incorporating them into a robust hybrid framework. Among the experimental results found in this paper, the highest accuracy in crop classification was that of the SVR meta-model and very good performance of XGBoost and MLP with high precision in regression tasks such as yield and fertilizer prediction.

QUESTION 2: What is a meta model in the context of your project?

In the context of this project, a meta-model is considered as a high-level model that combines the predictions of base models to achieve higher overall accuracy and robustness. Instead of using raw data directly, the meta-model takes the output from base models like Random Forest, Gradient Boosting Regressor, and LSTM, analyzes the pattern or relationship between these predictions, and develops refined predictions to feed into the final neural network layer. This last layer takes the output from the meta-model and give very optimized and accurate results for target tasks.

In this study, each of the base models was designed to suit a particular task: Random Forest was used for crop classification, considering its efficiency in handling categorical data as well as numerical data, Gradient Boosting Regressor (GBR) was chosen for yield prediction since it possesses some useful strengths for minimizing iteration errors arising in regression tasks. Meanwhile, the Long Short-Term Memory network(LSTM) was used for predicting fertilizer variables because of their excellent performance on sequential data to capture temporal information in input features.

For the next step, the SVR, XGBoost, and MLP meta-models were trained using the base model predictions as their input features. Given this architecture, each meta-model learns to aggregate the strengths of various base models while canceling their individual weaknesses for more robust predictions.

This is further developed in a hierarchical manner so that mistakes or biases from individual base models get minimized. For example, if some pattern in the data keeps one base model performing worse, the meta-model learns to balance and refine the outputs, leveraging the complementary capabilities of all base models.

The overall performance is improved by using a meta-model that has integrated various approaches to learn effectively. For example, the meta-model in this study outperformed others in classifying crops using RF with an accuracy of more than 99%, yield prediction using GBR with an RMSE of 0.15, and modelling sequential data using LSTM with an RMSE of 0.05 for

fertilizer prediction. That is important because the integration made sure that the system leveraged the best predictive capabilities of each base model while minimizing the errors. For example, the meta-model in this work outperformed the task of crop classification and yield prediction by effectively fusing temporal insights from LSTM, classification power of Random Forest, and regression strength of Gradient Boosting.

This is a hierarchical structure that not only ensures superior predictive accuracy but also caters for various agricultural tasks with great flexibility, altogether making the whole framework highly flexible and effective. Its modular nature enables scalability of the architecture, accommodating new datasets or agricultural tasks that may be used in future scenarios and hence ensure long applicability in different farming scenarios.

QUESTION 3: How do you predict the yield of the next year do you use variables from the current year?

This study predicts next year's yield using a combination of historical data and expected environmental and management inputs based on historical patterns. In this way, the models capture long-term trends and relationships to provide accurate and reliable predictions.

The methodology begins by using historical data to understand the pattern. Models were trained on historical yields, rainfall, fertilizer usage, and soil characteristics. These variables enable the models to learn the developing pattern and the different factors that come into play in determining the crop performance of a particular region over time. The temporal features of Crop Year and Season were also crucial to account for year-on-year variability that enables the model to identify and adapt to seasonal trends and changes.

The other key factors of prediction included the forecasted variables. These variables including rainfall, temperature, humidity, and fertilizer application were not taken from the current year but estimated from historical averages and trends. In this way, models use a realistic set of inputs for the upcoming year and this is reflective of expected environmental and management conditions.

The Gradient Boosting Regressor was used as the base model in the machine learning framework for the prediction of yield. It was selected because it is very efficient in handling the complexity in relationships between data. Further refinement of outputs from the base model was performed through a meta-model, which in turn was optimized by the final neural network layer. In this layered architecture, strengths of various models are combined together for more accurate yield forecasting.

The experimental results proved the reliability of this approach. For example, the overall best model (the Grid Search Optimized MLP meta-model) achieved an R^2 score of 0.9769 and a RMSE of 0.1628 while predicting yield, hence proving to be effective in capturing both long-term trends and projected variability.

In conclusion the prediction process does not use the current year's variables directly but is based on historical data and projected trends. Through this method, the study gives the most accurate and reliable yield estimates for the next year.

QUESTION 4: Is there a formula relating the area to the yield?

Yes, there is a direct formula that relates area to yield, given by:

$$\text{Yield} = \frac{\text{Production}}{\text{Area}}$$

In this formula Yield is defined as a quantity of production per unit area usually measured in metric tons per hectare. Production in this context is the total quantity of crops harvested, while Area is the total land under cultivation.

$$\text{Yield} = \frac{100 \text{ metric tons}}{50 \text{ hectares}} = 2 \text{ metric tons per hectare.}$$

For example, if 100 metric tons are produced on land measuring 50 hectares, then the crop yield would be 2 metric tons per hectare. These are standardized calculations that allow the comparison of crop performance across regions or seasons or even farming practices.

However, even though the formulae gives a simple and direct relationship of area to yield, it does not consider many other variables that will determine crop productivity. For this project, Yield was a key target variable and predictors included Area, Fertilizer, Rainfall, Pesticide Usage, Soil Nutrients and many others. These environmental and management variables that were incorporated in the machine learning models of this study go ahead to capture interactions that were not considered by the formula hence giving more precise yield predictions.

QUESTION 5: In some tasks, hyperparameter tuning had minimal impact. Why do you think this occurred?

In this project, hyperparameter tuning was mostly used to further refine the performance of the meta-models (SVR, XGBoost, and MLP), but it had a varying degree of impact for different tasks, mainly because:

1. The meta-models showed high baseline performance.

For tasks like crop classification, the default configurations of SVR and XGBoost already yielded very good results. For instance, SVR achieved 100% accuracy with default parameters, which means that the model captured relationships in the dataset so well that further optimization was not necessary. Since the baseline performance was already high, there was little room for improvement with hyperparameter tuning.

2. Ease of Certain Tasks

Crop classification is a relatively simple task, and its input-output relationships were strong and clearly distinguishable. Therefore, meta-models like SVR and XGBoost learned such patterns easily with their default settings. Hyperparameter tuning in this case did not make much difference, as the complexity of the task did not require further adjustments of model configurations.

3. Model Robustness

Both XGBoost and SVR are intrinsically robust models. XGBoost has regularization and smart optimization techniques, while SVR is good at handling nonlinear relationships due to its kernel functions. Their intrinsic design let them generalize effectively across tasks with little tuning.

4. Task-Specific Complexity

The simple tasks such as crop classification hardly benefited much from the tuning of hyperparameters, but the effect was immense in fertilizer prediction and to some extent, in yield prediction:

- In the tuned case, MLP obtained the minimum RMSE and MAPE in predicting fertilizers, proving its ability in modeling complicated relationships.
- XGBoost with optimized hyperparameters by Grid Search greatly improved yield prediction, suggesting the necessity of the tuning for a more variable regression task.

In conclusion, the influence of hyperparameter tuning on simpler tasks like crop classification was very limited due to the inherent robustness of the meta-models, the clear relationships in the data, and the high baseline performance of default configurations. However, for more complex tasks, such as fertilizer and yield prediction, where variability is higher and intricate dependencies exist in the data, hyperparameter tuning became significant. This approach ensured that resources were well utilized, focusing on tasks where tuning had the most impact while maintaining consistently high performance across all predictions.