Hybrid Feature Selection Method and Random Forest for Predicting Crop Yield Final Report

Jordana Izquierdo
a1876235

April 18, 2024

Report submitted for **Data Science Research Project B** at the School of Mathematical Sciences, University of Adelaide

THE UNIVERSITY of ADELAIDE

Project Area: **Data Science**
Project Supervisor: **Vince Wang**

In submitting this work I am indicating that I have read the University's Academic Integrity Policy. I declare that all material in this assessment is my own work except where there is clear acknowledgement and reference to the work of others.

I give permission for this work to be reproduced and submitted to other academic staff for educational purposes.

OPTIONAL: I give permission this work to be reproduced and provided to future students as an exemplar report.

Abstract

The prediction of crop yield remains a significant concern for farmers, governments, and society, as the volume of crop production impacts food availability. Traditional methods of predicting crop yields have limitations, needing the exploration of more efficient techniques. In recent years, the use of Machine Learning, particularly the Random Forest algorithm, has increased. However, a primary challenge in applying machine learning is the process of feature selection. To address this issue, this project proposes a hybrid method that combines causal discovery and backward feature selection for selecting features, along with the Random Forest algorithm for training the dataset to develop a predictive model capable of estimating crop yields. The proposed methodology has been implemented on a dataset comprising soil and weather attributes from four locations in Queensland, Australia: Emerald, Dalby, Mungindi, and Taroom. Initially, the methodology is applied to the entire dataset to develop what is referred to as the global model, which is then compared to the performance of the Random Forest alone without the hybrid feature selection method. The results show that the proposed method outperforms the latter. Subsequently, the method is applied individually to each location to construct separate models, resulting in a total of four models, designated as local models 1, 2, 3, and 4, respectively. A comparative performance analysis between the global model and the local models is conducted, resulting that the local models surpass the global model in terms of performance.

1 Introduction

Crop yield prediction is still a major concern for farmers, governments, and markets the volume of crop production directly affects food availability. Due to environmental change events, crop yield prediction remains a challenge and requires mechanisms for accurate predictions.

Usually, crop yield predictions have leveraged two primary methodologies: process-based modeling and purely statistical modeling [7]. Process-based crop models simulate the physiological processes that influence crop growth and development, taking into account environmental conditions and management practices [9]. However, these models require substantial data for calibration and are demanding in terms of data input [9]. Conversely, purely statistical models derive direct correlations between climatic and soil characteristics and crop yields [9]. These models rely less on extensive field calibration data and offer measures for evaluating performance [9]. Nevertheless, they face challenges related to multicollinearity among attributes, presuppositions of stationarity, and levels of accuracy that may not be satisfactory [9].

In recent years, Random Forest and Support Vector Machine have been proposed as alternatives for crop yield prediction that can overcome limitations of other approaches. Jeong et al. (2016) evaluate Random Forests (RF) in comparison with multiple linear regression (MLR) models for predicting wheat, maize, and potato yields at global and regional scales, considering climate and biophysical attributes [7]. Basha et al. (2020) use Random Forest to forecast crop yield considering only weather information [1]. Prasad, Patel & Danodia (2020) apply random forest to predict cotton yield in Maharashtra India, taking into account long-term agromet-spectral variables to train the model [12]. Faisal, Sreelakshmi & Chandra (2020) employed Random Forest to predict crop yield, assessing the performance of the model in 22 different crops [3]. Suresh et al. (2021) suggests applying Random Forest for predicting crop yields based on climate and soil parameters [15]. Elavarasan & Vincent (2021) propose the prediction of crop yield through Reinforcement Random Forest by incorporating reinforcement learning into the feature selection process during node splitting [2]. Kuradusenge et al. (2023) proposes the use of Random Forest, Polynomial Regression, and Support Vector Regression to predict crop yields in Rwanda. The results of the study suggest that Random Forest is the most effective model among the three [8]. Mujawar & Patil (2023) compare the performance of Deep Learning to Random Forest, Support Vector machine and K-Nearest Neighbour concluding

that Deep Learning underperforms for crop yield prediction [11]. T. & Sinha (2023) present an approach to crop yield prediction using Random Forest and Random Grid Search to tune the model's hyperparameters [16]. The Random Forest algorithm emerged as the most promising model for predicting crop yields.

Nonetheless, a significant challenge in model deployment involves selecting appropriate features. This step is essential in every machine learning process. Without it, nearly all machine learning techniques might struggle to operate effectively on datasets with high dimensionality [17]. Traditional feature selection methods can be categorized into wrapper, embedded, and filter methods. Filter methods rank features using a variety of statistical tests, such as T-tests, F-tests, and Chi-squared tests [5]. Wrapper methods utilize a learning machine as a black box to score subsets of features based on their predictive power [5]. Embedded methods integrate feature selection as part of the training process and are typically tailored to specific learning machines [5]. These methods are often used. However, these techniques do not optimally differentiate between correlation and causation and might inadvertently include confounders, leading to biased predictions and incorrect conclusions.A hybrid feature selection method emerges as an alternative, combining causal discovery and backward feature selection. This hybrid technique is less prone to overfitting and is computationally less demanding. By incorporating causal discovery as filter method, causal relationships between the features and the target are taken into account, and the explanatory capability of predictive model improves [17]. By considering backward feature selection method, the performance of the features selected is evaluated in a machine learning model.

This project proposes the adoption of Random Forest to develop a predictive model for crop yield estimation and a hybrid feature selection method consisting of causal discovery and backward feature selection. This approach has been applied to a dataset containing soil and weather data from four locations in Queensland, Australia: Emerald, Dalby, Mungindi, and Taroom. Initially, the methodology is applied to the entire dataset to develop what is referred to as the global model, which is then compared to the performance of the Random Forest alone without the hybrid feature selection method. Subsequently, four distinct models, referred to as local models 1, 2, 3, and 4 respectively, are trained using its corresponding data. Finally, a performance comparison between the global model and the local models is undertaken. The results shows that the local models outperform the global model in performance.

2 Background

Feature Selection

Guyon (2006) defines feature selection as a critical step in the data analysis process, significantly influencing the success of any subsequent statistical or machine learning endeavors [5].

Yu et al. (2020) separate existing feature selection methods into three categories: filter, wrapper, and embedded methods [17].

Filters establish a complete order of the features using a variety of statistical tests such as T-test, F-test, and Chi-squared test [5].

Wrappers use a learning machine as a black box to score subsets of features based on their predictive power [5].

Embedded methods integrate feature selection as part of the training process and are typically specific to certain learning machines [5].

Wrappers and embedded methods may produce vastly different feature subsets when minor perturbations occur in the dataset [5].

Hybrid Feature Selection Method

Hybrid methods combine causal discovery and backward feature selection approaches [5]. Initially, causal discovery is used to generate a reduced list of causal features. Based on this list, bachward feature selection is evaluated using a OLS. This project propose a hybrid method applying causal discovery and backward feature selection as method.

Causal Discovery

A family of methods that are able to discover a causal graph by analysing statistical properties of observational data [4]. Observational data refers to data collected by observing and recording information in their natural setting without any manipulation or intervention. Recent causal discovery advances allow encoding expert knowledge into the graph from interventional data [10]. Additionally, according to Molak (2023) [10] causal discovery methods are based in the following assumptions:

- Sufficiency: Implies that all common causes of the observed variables are included in the dataset.
- Faithfulness: Implies that the statistical relationships observed in the data accurately reflect the underlying causal relationships, and that any independence present in the data corresponds to a lack of direct causal connection in the underlying causal structure.
- Minimality: Implies that the causal model is as simple as possible.

PC Algorithm

This algorithm is part of constraint-based and score-based family methods. It finds causal structure from the data by taking into account rules of three basic graphical structures: chains, forks, and colliders [10]. It also assumes no latent confounder variables [14]. A confounding variable influences two or more other variables and produces a spurious association between them. A non-causal relationship is a spurious association. The PC algorithm works as follow:

- The PC algorithm begins by forming a fully connected graph based on the set of variables. It tests for independence between each pair of variables, using a statistical test (Pearson correlation for continuous data or the Chi-square test for categorical data).
- If the statistical test shows independence between two variables, the edge connecting them is removed. This process continues with increasingly larger sets of conditioning variables until no more edges can be removed through these conditional independence tests.
- After identifying the undirected graph, the algorithm then orients the edges using a set of orientation rules based on the conditional independence tests. These rules are applied iteratively to determine the direction of causality to the extent possible, resulting in a partially directed acyclic graph (PDAG).

More details can be found in Spirtes et al. (2012) [14].

Random Forest

It is a machine learning technique that used with both categorical and quantitative variables. It's a type of ensemble learning, where a group of trees work together to improve results. James et al. (2023) [6] and Gerón (2022) [7] explain Random Forests works as follow:

When building each tree in a Random Forest, the algorithm randomly selects a subset of m features out of the total features P available in the training set. This selection is done independently for every single tree. The size of the subset, m, is much smaller than the total number of features P (usually, m is chosen to be the square root of P for classification tasks, and around one-third of P for regression tasks).

For each node of the tree, only the randomly selected m features are considered for determining the best split. The algorithm evaluates the potential splits based only on these m features rather than all P features. By limiting each tree to only consider a random subset of features at

each split, Random Forest ensures that the trees are less correlated with each other.

This diversity among the trees in a Random Forest generally leads to improved model performance on unseen data, as it reduces the variance part of the error without significantly increasing the bias. The process of randomly selecting features for splits also allows Random Forest to more accurately measure the importance of each feature. Since each tree is built using different subsets of features, the performance impact of including or excluding a feature can be assessed across multiple trees, giving insights into which features are most important for predicting the target variable.

3 Methods

This project proposes the adoption of hybrid feature selection method as strategy and Random Forest as the algorithm to build a model capable of prediction crop yield. This approach has been applied to a dataset containing 100 years of soil and weather data from four locations in Queensland, Australia: Emerald, Dalby, Mungindi, and Taroom.

Initially, the methodology is applied to the entire dataset to develop what is referred to as the global model, which is then compared to the performance of the Random Forest alone without the hybrid feature selection method.

Subsequently, four distinct models, referred to as local model 1, 2, 3, and 4 respectively, are trained and tested using its corresponding data.

Finally, a performance comparison between the global model and the local models is undertaken. The results shows that the local models outperform the global model in performance.

3.1 Global Model

Initially, the entire dataset, which contains data from four locations, is trained using the approach detailed in Figure 1, with the sole difference that it does not employ the hybrid feature selection method. After, the entire dataset from the same four locations is trained using the proposed approach detailed in Figure 1 to develop what is referred to as the global model. The workflow of the Global Model is as follows:

Feature selection-Hybrid feature selection method

Before feature selection, zero-variance features are removed from the dataset, and the type of each feature (categorical or quantitative) is verified.

Subsequently, feature selection is conducted using causal discovery methods, specifically the PC (Peter-Clark) algorithm. The PC algorithm offers insights to eliminate all predictors that are (conditionally) independent of the target (given the other predictors) and which, therefore, do not contribute additional information for predicting the target [13]. For this process, only variables that are time-lagged with respect to the target variables are considered as potential predictors. The PC algorithm selects six features for this particular situation.

Following the causal discovery phase, the process continues with backward feature selection using Ordinary Least Squares (OLS) Regression. In backward feature selection, the process starts with all features suggested by the PC algorithm and iteratively remove the least significant

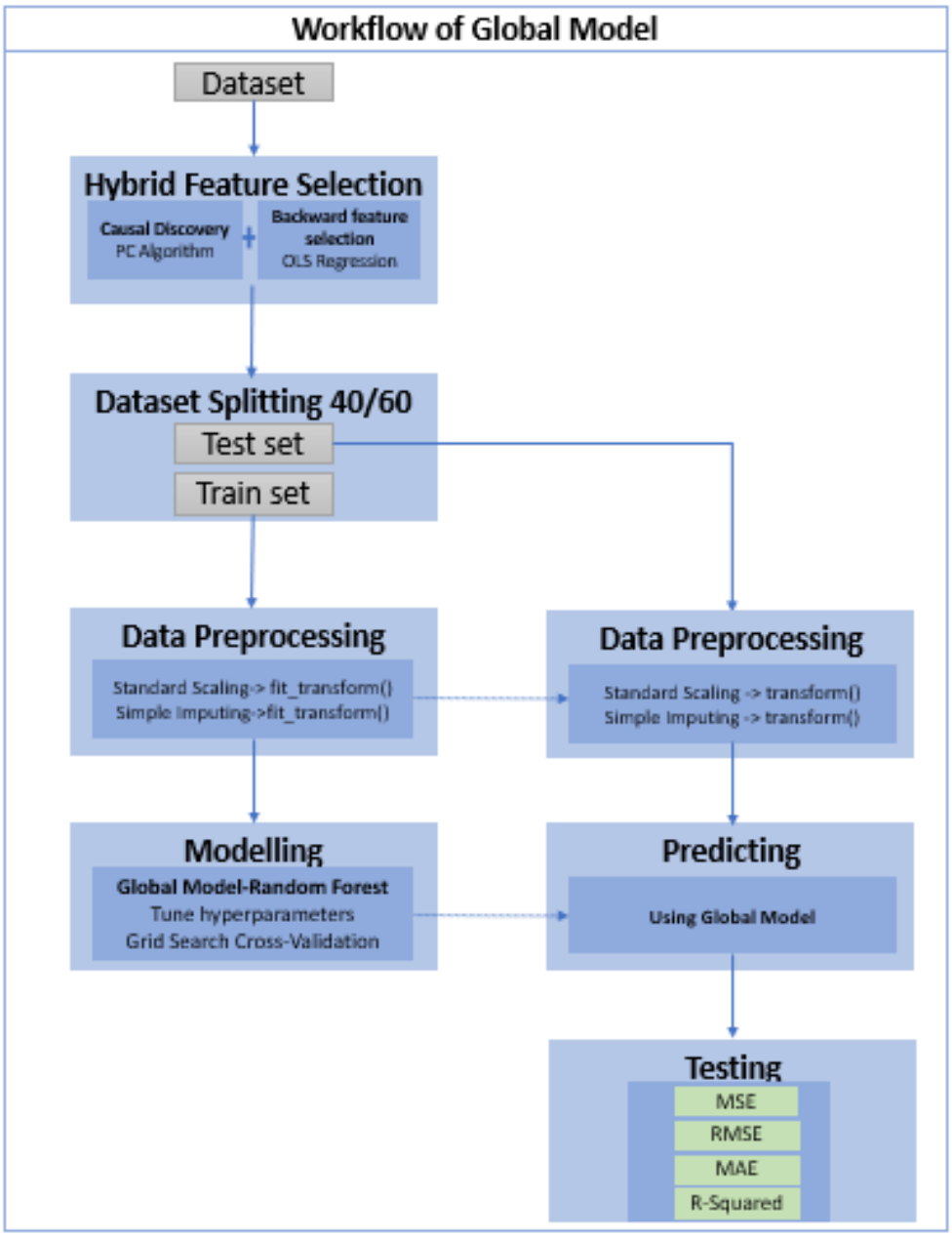

Figure 1: Workflow of Global model including a) Feature selection, b) Data Splitting, c) Data Preprocessing, d) Modelling, e) Predicting, and f) Testing

one according to a p-value. This step is done until the best subset of features is identified that provides sufficient predictive power for the model. The process finalize with five features for this particular situation.

Data Splitting

The dataset is then split into a training set and a test set, with the test set comprising 40% of the data and the training set 60%.

Data Preporcessing

This step includes standard scaling using fit_transform() method and simple imputing (in case of missing values), also using fit_transform().

Modelling

A model that is referred as Global Model is constructed using a Random Forest algorithm. Hyperparameters are tuned, and grid search cross-validation is performed to find the best model parameters.

Prediction

The test set is also preprocessed, but it is important to note that only the 'transform()' method is used for both standard scaling and simple imputing, indicating that the parameters from the training set preprocessing are applied to the test set. Subsequently, predictions are made using the Global Model that was trained in the modelling step.

Testing

The model's performance is evaluated using various metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-Squared (R^2), which collectively measure the accuracy and goodness of fit of the model.

3.2 Local Models

Four distinct models, referred to as Local Model 1, Local Model 2, Local Model 3, and Local Model 4, are each trained and tested using their corresponding data. The workflow for training and testing each model follows the same steps as the Global Model workflow, with the sole difference being that the data used to train each local model corresponds to a specific location. Additionally, the four models were trained using the features selectedfor the Global Model. Table 1 lists the locations corresponding to Datasets 1, 2, 3, and 4. Furthermore, the workflow is detailed in Figure 2.

Dataset-Model	Location
Dataset 1 - Local Model 1	Emerald
Dataset 2 - Local Model 2	Dalby
Dataset 3 - Local Model 3	Mungindi
Dataset 4 - Local Model 4	Taroom

Table 1: Location corresponding to Dataset-Local Model 1, 2, 3, and 4

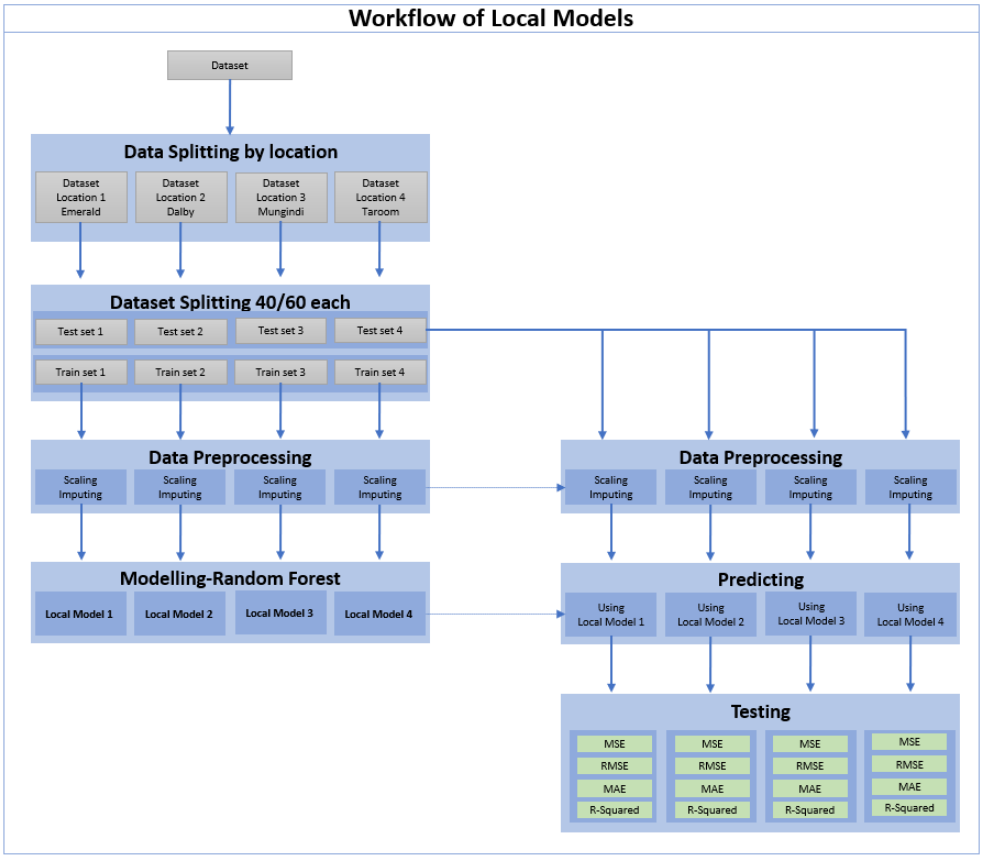

Figure 2: Workflow of Local models 1, 2, 3, and 4 including a) Data Splitting, b) Data Preprocessing, c) Modelling, d) Predicting, and e) Testing

4 Results

4.1 Feature selection

4.1.1 Traditional Feature Selection

Feature selection is conducted using backward feature selection. With this methodology, we do not know the causal relationships between the features and the target. We ended up with five features as follow:

1. EpBeforeFlowering:The period just before flowering.
2. EpAroundFlowering:The period just before, during, and immediately after flowering.
3. EpAfterFlowering:The period after flowering.
4. MinTAVG: Minimum average temperatures.
5. MaxTAVG: Maximum average temperatures.

4.1.2 Hybrid Feature Selection

Firstly, feature selection is conducted using the PC (Peter-Clark) algorithm. This algorithm provides insights to eliminate all predictors that are conditionally independent of the target and thus do not contribute additional information for predicting it. During this process, only variables that are time-lagged with respect to the target variables are considered potential predictors. The PC algorithm selects six features for this particular case scenario.

Following causal discovery, the process proceeds with backward feature selection using Ordinary Least Squares (OLS) Regression. This process starts with all the features suggested by the PC algorithm and iteratively removes the least significant one based on the p-value. This step continues until the optimal subset of features is determined. The process concludes with five features presented as follow.

1. EpBeforeFlowering:The period just before flowering.
2. EpAroundFlowering:The period just before, during, and immediately after flowering.
3. EpAfterFlowering:The period after flowering.
4. pawc: Possibly Plant Available Water Capacity.
5. CGRAroundFlowering: Crop Growth Rate around the flowering stage.

4.2 Modelling

During the modeling process, six models were evaluated: Linear Regression, Ridge Regression, Lasso Regression, Random Forest, Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM). The Random Forest model outperformed the other five models. Consequently, the Random Forest has been employed for further experimentation in this research project, specifically for the Global Model and Local Models 1, 2, 3, and 4.

Global Model

The Random Forest Regression model was trained using GridSearchCV from the sklearn library to conduct hyperparameter tuning and cross-validation. A KFold cross-validation object with five splits was configured to shuffle the data and was set with a consistent 'random_state' of 1 to ensure reproducibility. The model's final parameters are presented in Table 2.

Parameter	**Value**
max_depth	8
max_features	None
min_samples_leaf	10
min_samples_split	10
n_estimators	300

Table 2: Optimal hyperparameters for the Random Forest Global model

Local Models

For each local model, a Random Forest Regression model is used. Moreover, GridSearchCV from the sklearn library is employed to perform hyperparameter tuning and cross-validation. A KFold cross-validation object with five splits is specified to shuffle the data, and it is configured with a consistent 'random_state' of one to ensure reproducibility in each case. The models' optimal parameters are presented in Table 3.

4.3 Testing

Model Without Hybrid Feature Selection shows a lower performance compared to the model with hybrid feature selection method. A Local Model shows a better fit and lower errors compared to a Global Model.

Parameter	Local Model 1	Local Model 2	Local Model 3	Local Model 4
max_depth	8	8	8	10
max_features	None	None	None	None
min_samples_leaf	10	8	8	8
min_samples_split	10	8	8	8
n_estimators	300	300	300	300

Table 3: Optimal hyperparameters for Local Models 1, 2, 3, and 4.

Model Without Hybrid Feature Selection

The metrics resulting from the testing of the model Without Hybrid Feature Selection are presented in Table 4, and the performance scatter plot of the model is depicted in Figure 3.

Model Without Hybrid Feature Selection	Value
RMSE	180.032
MAE	112
R-Squared	0.96

Table 4: Metrics - Model Without Hybrid Feature Selection

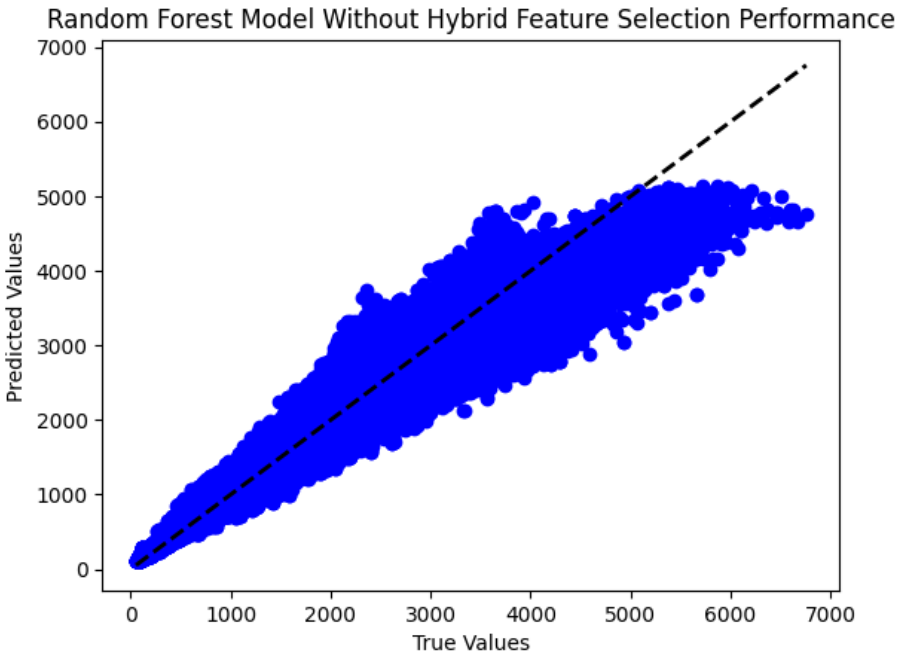

Figure 3: Model Without Hybrid Feature Selection scatter plot

Global Models

The metrics resulting from the testing of the global model are presented in Table 5, and the performance scatter plot of the model is depicted in Figure 3.

Global Model Metrics	Value
RMSE	178.67
MAE	117.48
R-Squared	0.964

Table 5: Metrics - Global model

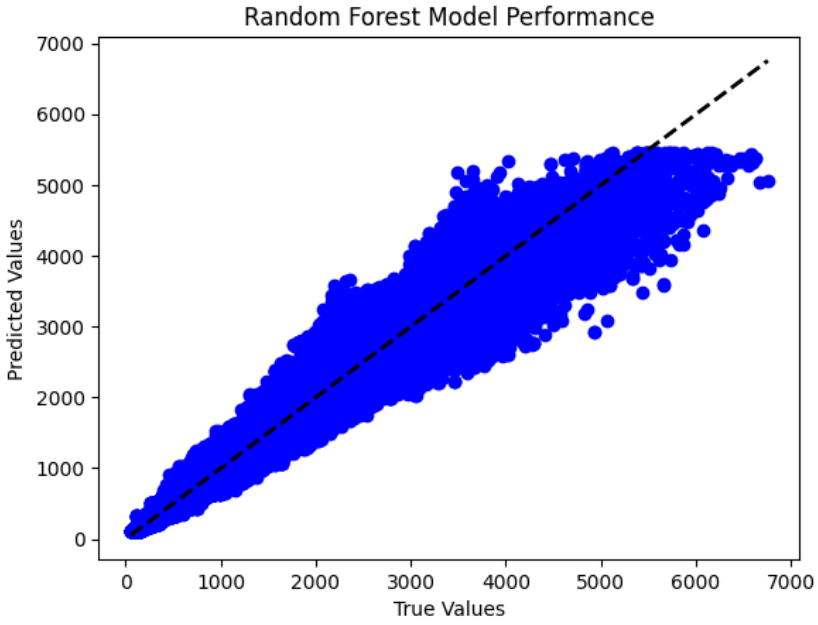

Figure 4: Global Model Performance Scatter Plot

Local Models

The metrics resulting from the testing of the global model are presented in Table 6, and the performance scatter plot of the model is depicted in Figure 4.

Local Models Metrics	**Local Model 1**	**Local Model 2**	**Local Model 3**	**Local Model 4**
RMSE	122.5	140	168.15	150.98
MAE	81.28	88	110.273	99.691
R-Squared	0.98	0.98	0.968	0.971

Table 6: Metrics from Local Models 1, 2, 3, and 4.

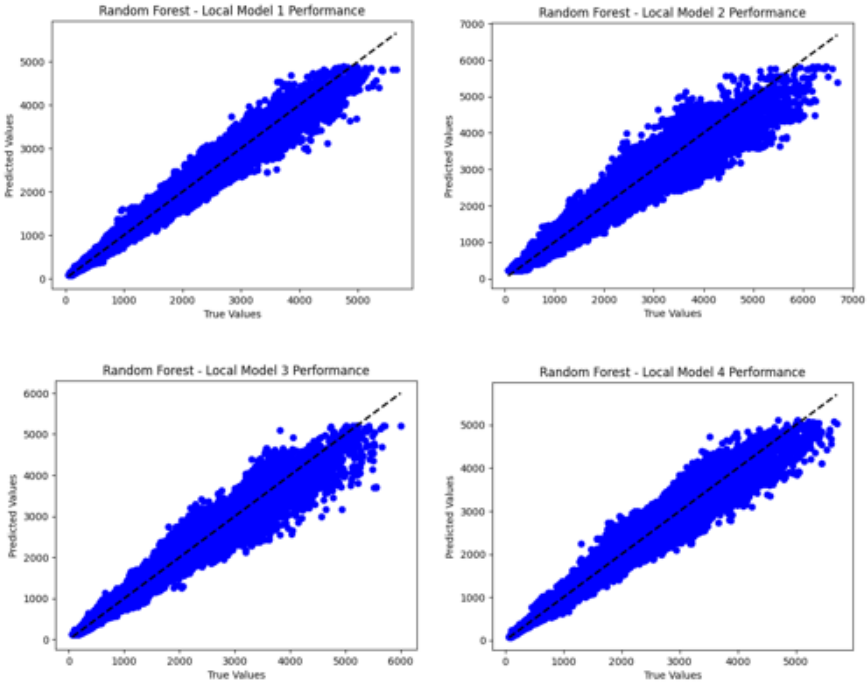

Figure 5: Local Models Performance Scatter Plot

5 Conclusion

The proposed methodology was implemented on a dataset comprising soil and weather attributes from four locations in Queensland, Australia: Emerald, Dalby, Mungindi, and Taroom. Initially, the methodology was applied to the entire dataset to develop what is known as the Global Model. Subsequently, it was applied individually to each location to construct four distinct models, referred to as Local Models 1, 2, 3, and 4. A comparative performance analysis between the Global Model and the Local Models was conducted, revealing that the Local Models surpass the Global Model in performance metrics.

The approach proposes a hybrid feature selection process that combines causal discovery (PC Algorithm) with backward elimination to construct models that are predictive and also capture the underlying causal relationships within the data, which is crucial for accurate interpretation. A hybrid feature selection methodology outperforms a traditional feature selection process.

The Random Forest model was selected for further experimentation because it surpassed other models, including Linear Regression, Ridge Regression, Lasso Regression, Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM), in terms of performance.

While a Local Model exhibits a better fit and lower errors compared to the Global Model, there is still potential for further improvement.

Acknowledgements

I extend my sincere gratitude to the research professionals and data scientists whose continuous contributions have significantly propelled the field of Data Science forward. The collective effort and dedication are truly commendable and serve as the foundation of our ever-growing knowledge and technological advancements.

A Appendices

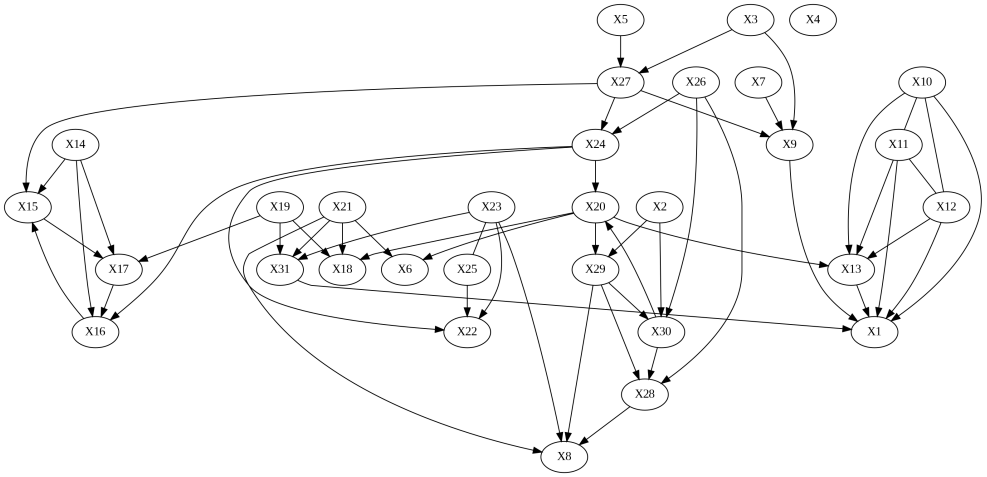

Figure 6: Causal Graph obtained from the PC Algorithm applied to the dataset

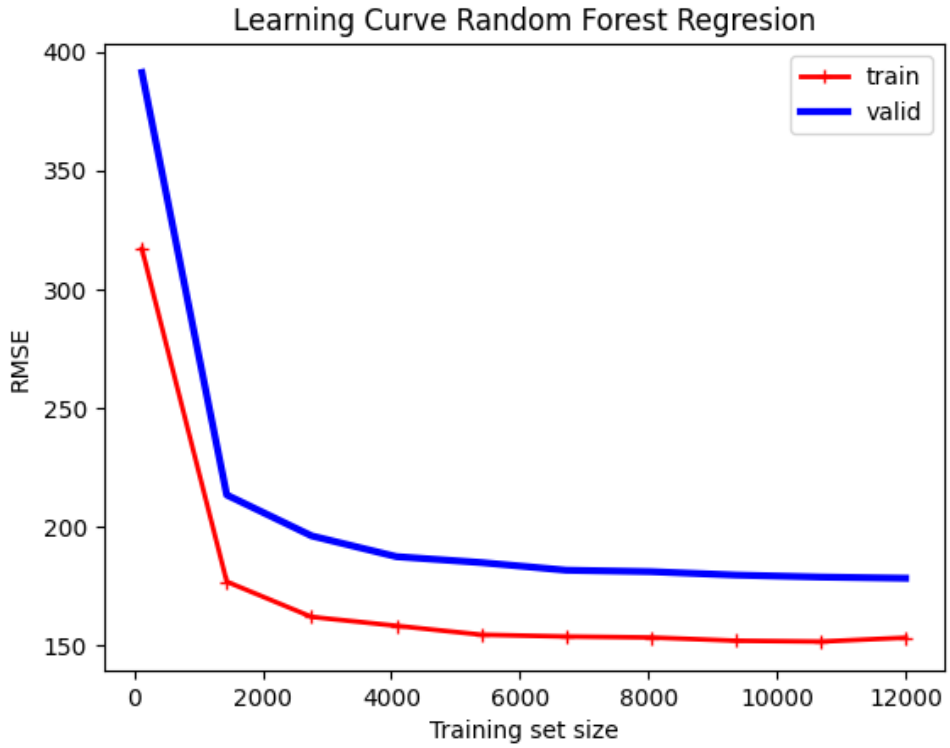

Figure 7: Global Model learning curve using Random Forest

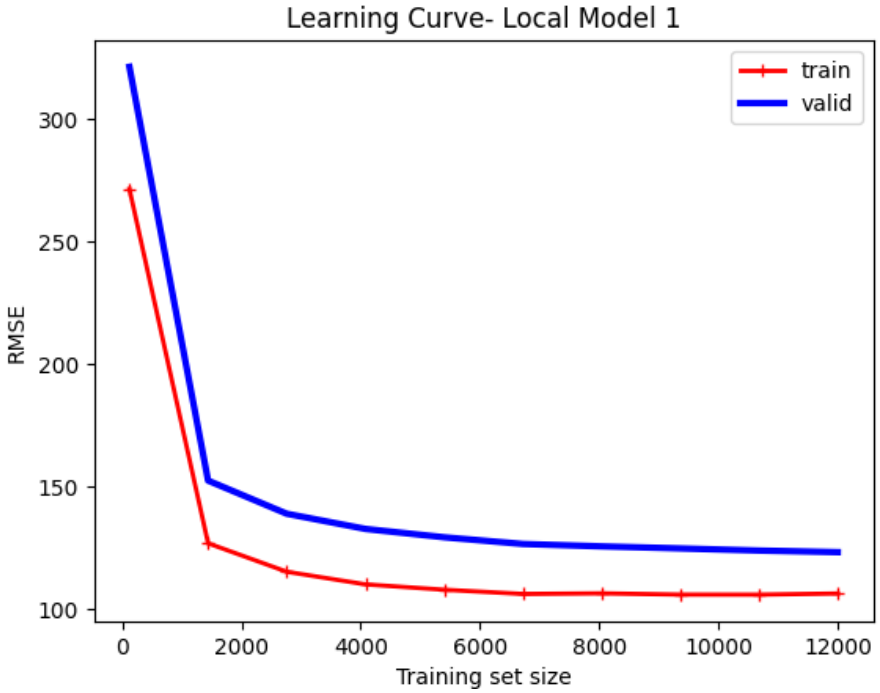

Figure 8: Local Model 1 learning curve using Random Forest

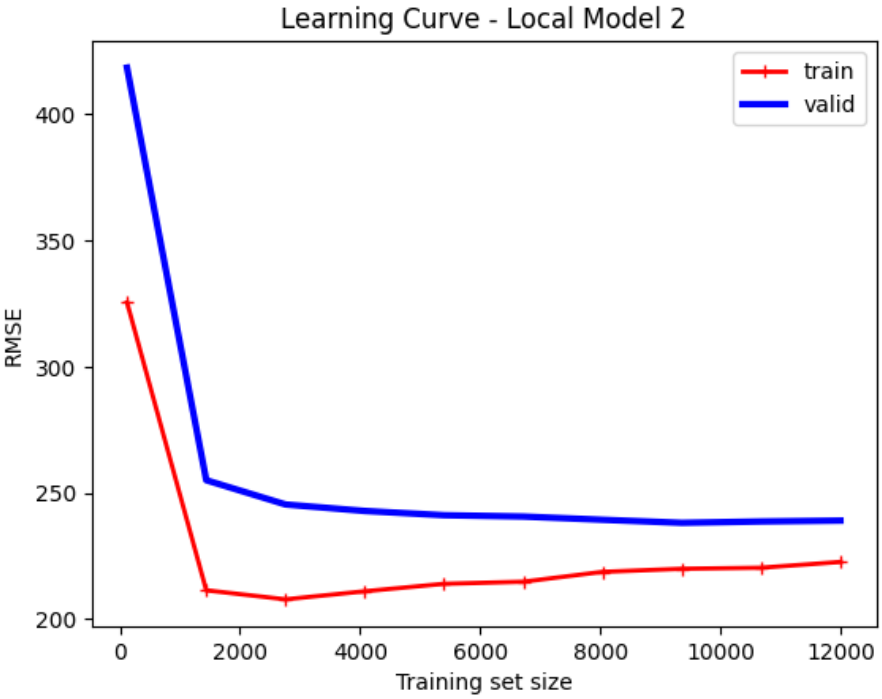

Figure 9: Local Model 2 learning curve using Random Forest

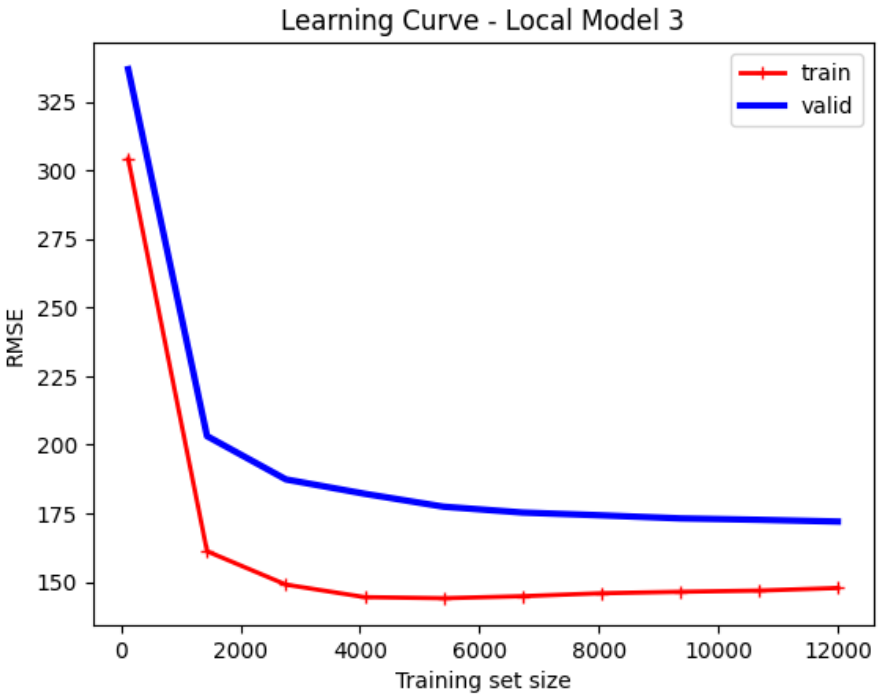

Figure 10: Local Model 3 learning curve using Random Forest

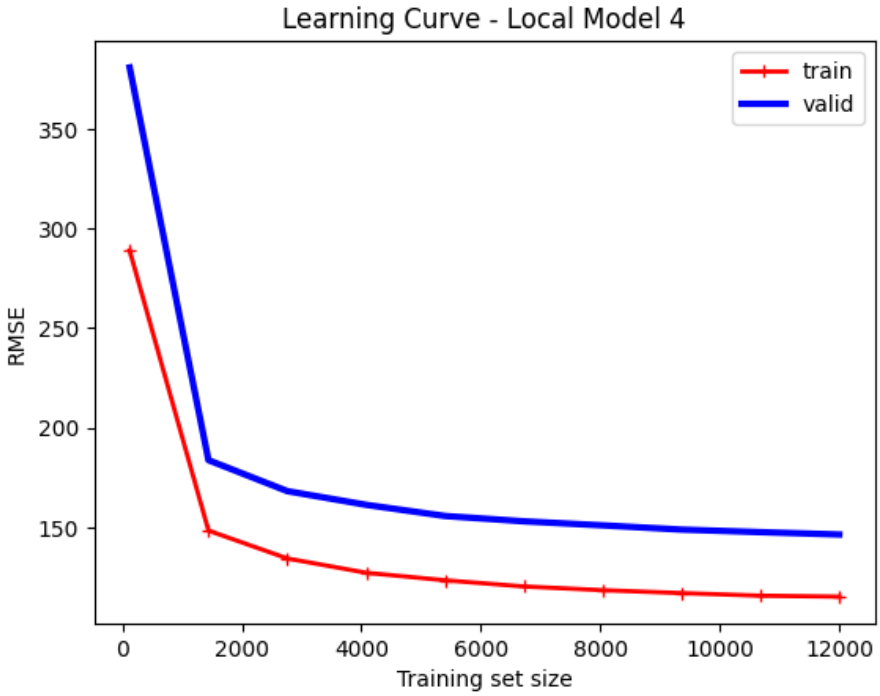

Figure 11: Local Model 4 learning curve using Random Forest

References

[1] S. M. Basha, D. S. Rajput, J. Janet, S. Ramasubbareddy, and S. Ram. Principles and practices of making agriculture sustainable: Crop yield prediction using random forest. *Scalable Computing. Practice and Experience*, 21(4):591–599, 2020.

[2] D. Elavarasan and P. M. D. R. Vincent. A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters. *Journal of Ambient Intelligence and Humanized Computing*, 12(11):10009–10022, 2021.

[3] G. Faisal, S. Sreelakshmi, and V. Chandra S. S. Crop yield prediction for smart agriculture with climatic parameters using random forest. In *Advances in Computing and Data Sciences*, volume 1848, pages 367–376. Springer Nature Switzerland, Cham, 2023.

[4] C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019.

[5] I. Guyon, editor. *Feature Extraction: Foundations and Applications*. Springer, Berlin, Germany, 1 edition, 2006.

[6] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. *An Introduction to Statistical Learning with Applications in Python*. Springer International Publishing, Cham, 1 edition, 2023.

[7] J. H. Jeong, J. P. Resop, N. D. Mueller, D. H. Fleisher, K. Yun, E. E. Butler, and S. H. Kim. Random forests for global and regional crop yield predictions. *PloS One*, 11(6):e0156571, 2016.

[8] M. Kuradusenge, E. Hitimana, D. Hanyurwimfura, P. Rukundo, K. Mtonga, A. Mukasine, and A. Uwamahoro. Crop yield prediction using machine learning models: Case of irish potato and maize. *Agriculture (Basel)*, 13(1):225–, 2023.

[9] D. B. Lobell and M. B. Burke. On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, 150(11):1443–1452, 2010.

[10] A. P. Molak. *Causal Inference and Discovery in Python: Unlock the secrets of modern causal machine learning with DoWhy, EconML, PyTorch and more*. Packt Publishing Ltd., United Kingdom, 1 edition, 2023.

[11] A. S. Mujawar and S. R. Patil. A review on crop yield prediction using random forest, svm & knn. *International Journal of Computer Science and Mobile Computing*, 12(6):41–44, 2023.

[12] N. R. Prasad, N. R. Patel, and A. Danodia. Crop yield prediction in cotton for regional level using random forest approach. *Spatial Information Research (Online)*, 29(2):195–206, 2021.

[13] G. S. Saranya, T. Beucler, F. I-H. Tam, M. S. Gomez, J. Runge, and A. Gerhardus. Selecting robust features for machine learning applications using multidata causal discovery. arXiv.Org, 2023.

[14] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*, volume 81. Springer, 2012.

[15] N. Suresh, N. V. K. Ramesh, S. Inthiyaz, P. P. Priya, K. Nagasowmika, KotaVNH. Kumar, and B. N. K. Reddy. Crop yield prediction using random forest algorithm. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 279–282. IEEE, 2021.

[16] P. T. and D. Sinha. Crop yield prediction using improved random forest. In *ITM Web of Conferences*, volume 56, pages 2007–, 2023.

[17] K. Yu, X. Guo, L. Liu, J. Li, H. Wang, Z. Ling, and X. Wu. Causality-based feature selection: Methods and evaluations. *ACM Computing Surveys*, 53(5):1–36, 2020.