Ahliya Journal of Business Technology and MEAN Economies [Published by Palestine Ahliya University]
Vol. 02 Issue 01 (2025)
ISSN: 3007-9691

Crop Yield Prediction Using Supervised Machine Learning Algorithms

Fouad Sleiby[*1], Mutaz Rasmi Abu Sara[*2], Mohammad Shakarnah[**1], Emma Qumsiyeh[**2] and Murad Zeer[***2]
[1]Palestine Ahliya University (Palestine)
[*]✉ fuadsleibi@gmail.com
[**]✉ moh.shakarnah@gmail.com
[2] Faculty of Engineering and Information Technology, Palestine Ahliya University (Palestine)
[*]✉ moutaz.a@paluniv.edu.ps
[**]✉ e.qumsiyeh@paluniv.edu.ps
[***]✉ muradzeer@paluniv.edu.ps

Received:14/02/2025 Accepted:17/03/2025 Published:31/05/2025

Abstract:

This paper discusses different supervised machine learning models in order to come up with a predictive model of crop yield, which utilizes the soil and environmental parameters. The data is a creation of the Kaggle in the project of Samrudha hackathon and is intended to support AI-based applications to smart and sustainable farming. We have done strict data preprocessing, which includes the elimination of outliers, the processes of duplicate and missing data. Various regression algorithms, such as K-Nearest Neighbors (KNN), Linear Regression, Ridge and Lasso Regression, Support Vector Regression (SVR), Decision Trees, and Random Forests were used and tested. R 2 and Mean Squared Error (MSE) were used as performance measurements. The Random Forest Regressor was the best performing model out of all the models tested with a test R 2= 0.9394 and a test MSE= 4.0840. This indicates the strength and capability of generalizing of ensemble approaches of agricultural yield forecasting activities. The originality of this study lies in its systematic and rigorous comparison of multiple supervised machine learning models for crop yield prediction using carefully preprocessed soil and environmental data. It further contributes by demonstrating the superior generalization capability of ensemble methods, particularly Random Forests, in supporting accurate and sustainable smart farming applications.

Keywords: *Crop Yield Prediction; Supervised Machine Learning; Data Preprocessing.*

Ahliya Journal of Business Technology and MEAN Economies Vol. 02 Issue 01 (2025)
[Published by Palestine Ahliya University] ISSN: 3007-9691

التنبؤ بمحصول المحاصيل باستخدام خوارزميات التعلم الآلي الخاضعة للإشراف

فؤاد صليبي[*1]، معتز رسمي أبو سارة[*2]، محمد شكارنة[**1]، إيما قمصية[**2] ومراد زير[***2]

[1] جامعة أهلية فلسطين (فلسطين)

fuadsleibi@gmail.com ✉[*]

moh.shakarnah@gmail.com ✉[**]

[2] كلية الهندسة وتكنولوجيا المعلومات، جامعة أهلية فلسطين (فلسطين)

moutaz.a@paluniv.edu.ps ✉[*]

e.qumsiyeh@paluniv.edu.ps ✉[**]

muradzeer@paluniv.edu.ps ✉[***]

تاريخ الاستلام:2025/02/14 تاريخ القبول:2025/03/17 تاريخ النشر:2025/05/31

ملخص:

تتناول هذه الورقة البحثية نماذج مختلفة للتعلم الآلي الخاضع للإشراف بهدف تطوير نموذج تنبؤي لمحصول المحاصيل، بالاستفادة من خصائص التربة والبيئة. البيانات المستخدمة هي نتاج مشروع هاكاثون سامرودها على منصة كاجل، وتهدف إلى دعم تطبيقات الذكاء الاصطناعي في الزراعة الذكية والمستدامة. خضعت البيانات لمعالجة مسبقة دقيقة، شملت استبعاد القيم الشاذة، ومعالجة البيانات المكررة والمفقودة. تم استخدام واختبار خوارزميات انحدار متنوعة، مثل خوارزمية أقرب الجيران (KNN)، والانحدار الخطي، وانحدار ريدج ولاسو، وانحدار متجه الدعم (SVR)، وأشجار القرار، والغابات العشوائية. استُخدم معامل التحديد (R^2) ومتوسط مربع الخطأ (MSE) كمقياسين للأداء. وقد حقق نموذج الغابات العشوائية أفضل أداء بين جميع النماذج المختبرة، حيث بلغ معامل التحديد (R^2) 0.9394 ومتوسط مربع الخطأ (MSE) 4.0840. يدل هذا على قوة وفاعلية تعميم أساليب التنبؤ الجماعي في أنشطة التنبؤ بمحاصيل الزراعة. تتجلى أصالة هذا البحث في المقارنة المنهجية والشاملة بين عدة نماذج تعلم آلي خاضعة للإشراف لتنبؤ إنتاجية المحاصيل اعتمادًا على بيانات بيئية وتربوية حقيقية بعد معالجتها بدقة عالية. كما يبرز البحث القيمة التطبيقية لاستخدام نماذج التجميع، وبخاصة الغابات العشوائية، في دعم الزراعة الذكية والمستدامة بنماذج تنبؤية عالية الدقة وقابلة للتعميم.

الكلمات المفتاحية: *التنبؤ بمحصول المحاصيل؛ التعلم الآلي الخاضع للإشراف؛ معالجة البيانات المسبقة.*

1. Introduction

This is because agriculture plays a fundamental role in the provision of food, jobs, and other supplies to the different sectors. Crop productivity is however, being decreased by many factors, which include changing climate, soil characteristics, irrigation practices, and farming practices. The conventional approaches of yield forecasting cannot adequately account all the interplay of these factors resulting in poor planning and resource exploitation (Iniyan et al., 2023).

Machine learning (ML) is now a useful technology to find and describe nonlinear relationships in large data, discovering some latent relationships in agricultural data. In the case of predictive tasks, including modeling crop yields, supervised machine learning (ML) seems to shine indeed since the utilization of labeled data allows making more precise and versatile models (Qumsiyeh & Sabha, 2023).

This paper will compare the different supervised learning algorithms, such as K-Nearest Neighbors (KNN), Linear Regression, Ridge and Lasso Regression, Support Vector Regression (SVR), Decision Trees and Random Forests to predict crop yields by examining a combination of environmental and soil related variables. Comparison of the performance of these models assists researchers in deciding which of them is most effective in terms of yield prediction with the purpose of making a new contribution to AI in smart agriculture.

The study uses the data on the project of the Kaggle Samrudha Hackathon that tries to promote the creation of AI solutions to make farming more sustainable. An extensive series of actions which included preprocessing, model development and optimization, hyperparameter optimization, and performance analysis were performed to make sure that the results were not weak and could be applied in different cases.

Although machine learning is increasingly being used in agriculture, a large number of existing agricultural applications have been hampered by the inability to generalize, lack of clear insight into the relationship between features, and the challenge of generating mixed and noisy data. Moreover, even though deep learning methods are sought after, they are too costly and require large-scale and labeled training data, which is unhelpful with small agricultural data. This research paper helps in filling this gap as it looks at the performance of different lightweight and simple regression models under stringent preprocessing and validation procedures. Thus, it contrasts conventional algorithms with contemporary ensemble-based approaches and gives recommendations on how to make the yield prediction instruments, which are transparent, scalable, and efficient, in the area.

2. Literature Review

The use of machine learning (ML) in agriculture is already the main solution to the complicated issue of the crop yield prediction. The artificially intelligent approach to the problem proposed by machine learning techniques can be successfully applied to address the degree of complexity since agricultural output is conditioned by infinitely variable factors like soil, weather, and cultivation methods.

The authors of (Iniyan et al., 2023) offer a solid architecture of a set of regression-based machine learning (ML) algorithms to predict crop yield, and the most successful model is a feature-engineered Long Short-Term Memory (LSTM) model, which attains an accuracy of 86.3%. They highlight the real-world applicability of ML by an end-to-end web application, and they qualify its current limitation to a local offline system. The authors also propose the inclusion of cloud platforms and other parameters, such as plant genotype, in future performance improvements.

Medar et al. (2019) emphasize the economic significance of agriculture in India and advocate for the application of machine learning (ML) as a transformative technology in achieving higher yields through the optimal selection of crops. The research is based on the application of machine

learning (ML) in assisting the resolution of underlying agricultural issues, including unstable market patterns and ineffective policy planning. While there is no lengthy debate of algorithms in the research, the study highlights the greater socio-economic impact that predictive analytics has to offer to agriculture.

Jhajharia et al. (2023) empirically compared some machine learning (ML) and deep learning algorithms for predicting the yield of five prominent crops in Rajasthan. RF performed best among RF, SVM, Gradient Descent, LSTM, and Lasso regression, achieving the maximum coefficient of determination ($R^2 = 0.963$), the minimum root mean squared error (RMSE = 0.035), and the lowest mean absolute error (MAE = 0.0251). Their validation process justifies the correctness of their result and proves RF as a potential algorithm in actual crop yield scenarios.

Reddy and Kumar (2021) conducted a systematic review of machine learning (ML) methods for predicting crop yield. They identified essential flaws in the neural network model, including decreased efficiency and an inability to reduce error. The authors indicate that the standard supervised learning architecture is not able to identify the underlying nonlinearities in the agricultural data and is not premised on the idea of pure and simple heterogeneous input-output relations. The paper ends with a discussion of other research on hybrid and ensemble approaches in order to overcome the existing limitations.

A hybrid machine learning (ML) framework That Combines Decision Trees, random forests, and SVMs with deep learning algorithms (RNN, LSTM) was proposed by (Agarwal & Tarar, 2021) to predict the maximum crop yield and estimated the necessary soil nutrients. The model additionally enhanced accuracy in prediction as well as assists in cost calculation which is very economical to farmers. The deep learning techniques were combined with the conventional machine learning models successfully, which proved the suitability of the hybrid models to the agricultural prediction tasks.

To provide farmers with real-time support during their decision-making, (Champaneri et al., 2016) created an ML-prediction interactive system. The authors used the Random Forest to deal with uncertainty relating to climate during the yield results through the combination of variables, including rain, temperature, humidity, and soil water. The article demonstrates that Random Forest is a suitable model for regression in agriculture, owing to its ability to handle heterogeneous features and non-linear relationships.

Lastly, Abbas et al. (2020) utilized the crop yield prediction problem based on proximal sensing data and four machine learning models: linear regression, Elastic Net, k-nearest neighbors (k-NN), and Support Vector Regression (SVR). (Abbas et al., 2020) research applied this configuration to Canadian farm potato yield records, providing evidence that enhancing spatial and temporal forecasting capabilities can be achieved by blending sensor-based data (e.g., NDVI, conductivity) with machine learning. The multi-disciplinary and multi-season character of the design thus lends their findings higher rigour and generalizability.

Overall, the literature presented here collectively shows that although simple classical models, such as Linear Regression and k-NN, are easy to interpret and comprehend, ensemble models, such as Random Forest, and deep architectures, such as LSTM, tend to be more accurate and better handle sophisticated agricultural data. Hybrid models and cloud-based platforms are also emerging trends that aim to provide scalable and accessible predictive systems.

3. Methodology

3.1 Dataset:

This Kaggle dataset is designed for predicting the compressive strength of concrete based on its ingredients and curing age. It contains continuous numerical features for various concrete mix materials with the amounts of cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate, all in kilograms per cubic meter (kg/m^3). The data also gives age in days of the concrete which is a curing time, which is one of the significant determinants of strength gain. Concrete Strength, which is the target variable, will be measured in megapascals (MPa) and it will show the compressive strength of the concrete. The data can be applied in diverse uses such as regression modeling to project the strength, ratios of mixes to be used in high performance mixes, and maintaining quality controls during the production of concrete.

3.2 Data Preprocessing

Data analysis is one of the most important tasks in preprocessing because it guarantees data quality and integrity before the process of machine learning algorithms is used. Preprocessing refers to a collection of techniques applied to the raw data to give them a clean and filtered appearance and bring more usability to the data and model performance. In this research, we have carried out several preprocessing operations, such as analysis of unique values, detection of duplicates, missing value processing, and outlier processing. The unique value analysis is used in all columns so as to detect one or few-repeated values which can result to rows or potential outliers. We used nunique function in our data to study all the variables and ensured that all the variables have enough variations and none of the columns have very unique or equal values. The duplication of records would distort the outcome of any analysis process because of redundancy or bias. In our work we filtered our data to make sure that we did not have duplicate records by making sure there were no duplicates with simple panda's functions. This makes the data redundancy free and prepared to be processed. Virtue of the prediction and inferences made against a dataset is tainted by missing values. The easiest method of dealing with missing values is deletion or imputation. To ensure the data had no missing values, we performed tedious verification to verify the missing values with the help of functions like isnull and sum and ensured that there were no missing values in the data. This makes imputation unnecessary and ensures that data is complete. A very important process in ensuring that the dataset is clean is outlier detection. Outliers distort statistical measurements and blur the predicting ability of the model. The Interquartile Range (IQR) technique was used to determine any outliers and it revealed that there were 337 rows, which had values which were not within the normal range. We therefore discounted these records so as not to influence the performance of the model. Original dataset shape:(1625, 40). Potential shape of final dataset upon removal of outliers:(1288, 40).

4. Results

4.1 k-Nearest Neighbors Regression (KNN):

KNN is a non-parametric, instance, based learning algorithm that is applied when there is a regression problem. This can be used to predict the output value of a particular input by finding the nearest k training sample in the feature space and averaging their respective output values. Euclidean distance is usually used to calculate the distance and the importance of the selection of k has a direct bearing on the accuracy and generalization of the model. The KNN is also highly beneficial in those applications where the data follows non-linear trends and does not presuppose a particular underlying distribution. It is however sensitive to outliers and computationally infeasible with large data sets because it requires storage and comparison with all examples of training at prediction time.

Table1: KNN Without Scaling

K	Train R^2	Test R^2
1	1.00	0.83
3	0.93	0.85
9	0.88	0.85

Table 2: KNN With Standardscaler

K	Train R^2	Test R^2
1	1.00	0.72
3	0.89	0.74
9	0.68	0.56

Best Result (Unscaled): K=3, Test $R^2 = 0.85$

4.2 Linear Regression:

It is a basic statistical technique that is employed to model the relationship between a dependent variable and an independent variable or independent variables by fitting a straight line to the observed data. It operates under the assumption of linearity, constant variance, and error-independence and it is very popular due to its simplicity and interpretability (Su et al., 2012).

4.3 Ridge Regression:

It is a regularized form of linear regression, which involves inclusion of a penalty term into the loss function to reduce the coefficients. This also lowers the multicollinearity and overfitting of the model by regulating the complexity of the model particularly when the predictors are highly correlated. It is more predictive and can be interpreted(McDonald, 2009).

4.5 Lasso Regression

Lasso It is a regularization method that introduces an L1 penalty on the linear regression cost functional. It does both shrinkage of coefficients and selection of variables, and is applied with many features in a model since it causes some coefficients to shrink to zero(Ranstam & Cook, 2018).

Table 3: Without Scaling

Model	Train R^2	Test R^2	Test MSE
Linear Regression	0.859	0.845	9.58
Ridge Regression	0.858	0.846	9.49
Lasso Regression	0.856	0.846	9.54

Table 4: With Standardscaler

Model	Train R^2	Test R^2	Test MSE
Linear Regression	0.859	0.845	9.58
Ridge Regression	0.859	0.846	9.49
Lasso Regression	0.859	0.847	9.47

Best Result (Scaled): Lasso Regression, Test $R^2 = 0.847$

4.6 Support Vector Regression (SVR)

It is an algorithmic supervised machine learning and it is a support vector machine. It tries to identify a function that is close to the target within a target error within keeping model complexity. SVR can solve both the linear and the non-linear regression problems using the help of the kernel functions(Awad & Khanna, 2015).

The feature scaling with the StandardScaler was very effective in improving the performance of the model. The unscaled version of the model with the default values of the parameters resulted in a training R 2 of 0.5484, a test R2 of 0.5368 and a test Mean Squared error (MSE) of 30.22. But the effectiveness of the model was greatly enhanced after the input features were scaled by StandardScaler where the training R 2 value was 0.7885, the test R 2 value was 0.7718, and the test MSE was reduced to 14.89. This shows that feature scaling contributes highly towards the ability of

the model to make predictions in a generalized manner which results in higher prediction accuracy and performance in general.

4.7 SVR Grid Search (27 models)

The result of the best Grid Search using (kernel, C, Gamma)

Table5: Best Grid Search using (Kernel, C, Gamma)

Kernel	Best C	Best Gamma	Best Train R^2	Best Test R^2	Best Test MSE
Linear	0.1	ALL	0.8132	0.8298	11.53
Sigmoid	10	0.01	0.7926	0.8323	11.36
Poly	1.0	0.10	0.9030	0.8169	12.41

Sigmoid kernel (C=10, gamma=0.01) gave the highest Test $R^2 = 0.8323$

4.8 Random Forest Regression

It is a model of ensemble learning, which forms various decision trees in training and yields the average of the prediction of the trees in a regression problem. It averages variations, enhances the robustness of the model and both linear and non-linear relationships are adequately addressed. The approach is characterized by high accuracy, resistance to overfitting and the capability of estimating the importance of the features(Segal, 2004).

Table 6: Results for Random Forest Regressor

Configuration	**Train R^2**	**Test R^2**	**Train MSE**	**Test MSE**
Default (No Scaling)	0.9802	0.9386	0.6377	4.1395
Default (StandardScaler)	0.9896	0.9386	0.6377	4.1397
n_estimators=500 (Scaled)	0.9902	**0.9394**	0.6006	4.0843
max_depth=5 (Scaled)	0.9652	0.9244	2.1446	5.0967
max_depth=9 (Scaled)	0.9885	0.9371	0.7107	4.2403
Optimized (n=500, depth=30, max_features=1.0)	0.9902	**0.9394**	0.6006	4.084

The optimized model (nestimators = 500, maxdepth = 30) gave the best results and all the other hyperparameters were kept at default. The resulting Train R 2, Test R 2, Train MSE and Test MSE are 0.9902, 0.9394, 0.6006 and 4.0840 respectively. These values were almost the same as the default model, which proves that the initial set was already bordering on being optimal. But we were able to limit overfitting by changing max depth. As an example, a depth of 5 reduced the Train R 2 to 0.9652 but contributed to avoiding overfitting and improving the model generalization. In general, the best fit in terms of predicting the unseen data and the best fit in terms of fitting the training data was achieved with the optimized model.

4.9 Tree Regression

It is a non-parametric supervised learning algorithm which is used to model data by dividing it into branches according to specific feature values. To reduce the prediction error, the tree structure recursively splits the dataset into smaller subsets by the use of simple metrics such as mean squared error (MSE). The approach is easy to interpret, captures non-linear relationships, and is simple to overfit unless it is pruned correctly(Kushwah et al., 2022).

Table7: Decision Tree Regression Result

Configuration	**Train R^2**	**Test R^2**	**Train MSE**	**Test MSE**
Without StandardScaler	1.0000	**0.9012**	0.0000	**6.5005**
With StandardScaler	1.0000	**0.9012**	0.0000	**6.5005**
With StandardScaler, max_depth = 4	0.9279	0.8759	4.5742	8.1661
With StandardScaler, max_depth = 9	0.9930	0.8924	0.4466	7.0777

Our We got the following results: With the application of the Random Forest Classifier, the original model had a training accuracy of 1.0000 and a test accuracy of 97.12 without scaling and

97.81 with scaling. Nevertheless, the ideal training score indicated overfitting of the model. To minimize this and to be able to generalize, we optimized the model by changing important parameters. The limiting of tree depth was done using the max depth parameter, first, and this was to make the model simple. Indicatively, we found that max depth=4 decreased training accuracy by a minor percentage of 98.35 but test accuracy by a minor percentage of 97.95. We then used n estimate trees and with only four trees the model still managed to maintain a high performance with training accuracy of 99.35 and test accuracy of 97.95. The rationale of these changes was to balance the model, minimize overfitting and enhance test accuracy, which proved that parameter tuning is important in the Random Forest to attain the best results.

Table 8: Model Comparison

Model	Params	Scaler	Train RÂ²	Test RÂ²	Test MSE
KNN	K=1	No	1	0.83	
KNN	K=3	No	0.93	0.85	
KNN	K=9	No	0.88	0.85	
KNN	K=1	Yes	1	0.72	
KNN	K=3	Yes	0.89	0.74	
KNN	K=9	Yes	0.68	0.56	
Linear Regression	-	No	0.859	0.845	9.58
Ridge Regression	-	No	0.858	0.846	9.49
Lasso Regression	-	No	0.856	0.846	9.54
Linear Regression	-	Yes	0.859	0.845	9.58
Ridge Regression	-	Yes	0.859	0.846	9.49
Lasso Regression	-	Yes	0.859	0.847	9.47
SVR	Default	No	0.5484	0.5368	30.22
SVR	Default	Yes	0.7885	0.7718	14.89
SVR	Sigmoid, C=10, Î³=0.01	Yes	0.7926	0.8323	11.36
SVR	Linear, C=0.1	Yes	0.8132	0.8298	11.53
SVR	Poly, C=1.0, Î³=0.10	Yes	0.903	0.8169	12.41
Decision Tree	Default	No	1	0.9012	6.5005
Decision Tree	max_depth=4	Yes	0.9279	0.8759	8.1661
Decision Tree	max_depth=9	Yes	0.993	0.8924	7.0777
Random Forest	Default	No	0.9902	0.9394	4.084
Random Forest	Default	Yes	0.9896	0.9386	4.1397
Random Forest	n=500	Yes	0.9902	0.9394	4.0843
Random Forest	max_depth=5	Yes	0.9652	0.9244	5.0967
Random Forest	max_depth=9	Yes	0.9885	0.9371	4.2403
Random Forest	**Optimized (n=500, depth=30 , M_f=1.0 , n_T = 500)**	**No**	**0.9902**	**0.9394**	**4.084**

The optimal model after trying various regression algorithms was the random forest Regressor (depth = 30trees = 500, number of trees = 500, and max features = 1.0) with: Test R 2 =0.9394 and Test MSE =4.0840. The next step involved cleaning the data after which we applied and tested multiple supervised regression models. All the algorithms were identified with and without feature scaling by use of StandardScaler and some of the models were optimized further with hyper parameter tuning. We tested the models on the training and testing sets of R 2(coefficient of determination) and MSE (Mean squared error). The Random Forest Regressor is the only model that performed better than others in various configurations, and the optimized version of the Regressor had a Test R 2 = 0.9394 and a Test MSE = 4.0840. The Decision Tree Regressor also produced good results but it exhibited overfitting characteristics when using default settings. Unscaled data gave KNN the highest Test R 2 which was 0.85. The scaling of SVR was very high and optimally, the sigmoid kernel (C=10, gamma=0.01) gave the highest Test R 2 (0.8323). Linear, Ridge, and Lasso regressions all generated dependable and reasonable outcomes, but the Lasso Regression (scaled) generated the most successful of them(Test R 2 = 0.847).

These findings demonstrate the effects of scaling, regularization, and hyperparameter optimization on the model performance, which indicates the benefits of ensemble learning methods in complicated predictive problems, including crop yield forecasting.

5. Discussion

After a long series of trial and error with the various supervised regression models, the best performing model was the Random Forest Regressor, which was selected to predict the target variable. Having completed a detailed data preprocessing process, including cleaning and outlier management, and feature scaling, we tested different regression models, with training and testing data, in terms of R 2 and Mean Squared Error (MSE). The performance of each model was tested under different settings, that is, default, standardized input (with StandardScaler), and hyperparameter-tuned settings.

Random Forest Regressor using the following parameters: tree depth = 30, 500 estimators, and max features = 1.0, was found to be very good with a Test R2 = 0.9394 and Test MSE = 4.0840, which was very good in comparison to other models. It implies that the ensemble learning algorithms and especially tree-based algorithms are effective in modeling complex non-linear patterns of structured data.

The Decision Tree Regressor also performed well and the only method that did not overfit closely regularized. The K-Nearest Neighbors classification model was better working with unscaled data, as it produced a decent Test R 2 of 0.85. The feature scaling also played a significant part in Support Vector Regression (SVR) with the greatest output of feature scaling was done with the sigmoid kernel (C = 10, gamma = 0.01) with a Test R-square value of 0.8323. The Lasso Regression was the best linear model that scaled since it had the highest Test R- 0.847 which shows that regularization is highly influential in minimizing the variance of the model and enhancing generalization.

The findings confirm the previous studies in machine learning used in predictive analytics in agriculture and concrete strength prediction, where the interpretability of the model, and the predictive accuracy are the most significant. As an example, (Jhajharia et al., 2023) explained why Random Forest works better than any other model in the prediction of crop yields with R 2 = 0.963, which is frighteningly similar to our model (R 2 = 0.9394). (Champaneri et al., 2016) also supported our model, showing how the Random Forests were more effective than any other model when it comes to linear relationships in farm data. Also, (Agarwal & Tarar, 2021) reported the operation of hybrid models that are based on the use of tree-based and deep learning methods. They proposed a future direction of the research adding the LSTM or RNN layers to determine the time-varying behavior on a similar dataset. This is consistent with (Iniyan et al., 2023) results, which indicate that it is possible to establish a high prediction accuracy using a feature-engineered LSTM model, which in turn proves that deep learning can be used to improve the prediction accuracy of more complex processes which are based on time or sequence. We also support the findings of (Reddy & Kumar, 2021) who illustrated the shortcomings of classical neural networks and the need to have ensemble or hybrid networks in expressing the high-dimensionality and non-linearity of agricultural and material data.

6. Conclusion

Here, we have illustrated how powerful monitored machine learning models can be as far as crop yield prediction based on environmental and soil-based variables is concerned. Our intensive preprocessing and testing revealed that ensemble models particularly the Random Forest Regressor performance and stability is superior. Random Forest fitted with the best trade-off between bias and variance with Test R 2=0.9394 and Test MSE=4.0840 was obtained by proper tuning

(nestimators=500, maxdepth=30). KNN which is simple, scaled well, and regression algorithms such as Ridge and Lasso were assisted by regularization. Decision Trees provided a straightforward model modelling style although they needed depth restrictions to generalize. The Support Vector Regression worked much better when judiciously scaled as well as with a suitable choice of the kernel.

References:

Abbas, F., Afzaal, H., Farooque, A. A., & Tang, S. (2020). Crop yield prediction through proximal sensing and machine learning algorithms. *Agronomy*, *10*(7), 1046.

Agarwal, S., & Tarar, S. (2021). A hybrid approach for crop yield prediction using machine learning and deep learning algorithms. *Journal of Physics: Conference Series*, *1714*(1), 012012.

Awad, M., & Khanna, R. (2015). Support vector machines for classification. In *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (pp. 39–66). Springer.

Champaneri, M., Chachpara, D., Chandvidkar, C., & Rathod, M. (2016). Crop yield prediction using machine learning. *Technology*, *9*(38).

Iniyan, S., Varma, V. A., & Naidu, C. T. (2023). Crop yield prediction using machine learning techniques. *Advances in Engineering Software*, *175*, 103326.

Jhajharia, K., Mathur, P., Jain, S., & Nijhawan, S. (2023). Crop yield prediction using machine learning and deep learning techniques. *Procedia Computer Science*, *218*, 406–417.

Kushwah, J. S., Kumar, A., Patel, S., Soni, R., Gawande, A., & Gupta, S. (2022). Comparative study of regressor and classifier with decision tree using modern tools. *Materials Today: Proceedings*, *56*, 3571–3576.

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 93–100.

Medar, R., Rajpurohit, V. S., & Shweta, S. (2019). Crop yield prediction using machine learning techniques. *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, 1–5.

Qumsiyeh, E., & Sabha, M. (2023). Utilizing Convolutional Neural Networks and KMeans Clustering for Efficient Plant Leaf Disease Detection. *2023 2nd International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI)*, 1–7.

Ranstam, J., & Cook, J. A. (2018). LASSO regression. *Journal of British Surgery*, *105*(10), 1348–1348.

Reddy, D. J., & Kumar, M. R. (2021). Crop yield prediction using machine learning algorithm. *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 1466–1470.

Segal, M. R. (2004). *Machine learning benchmarks and random forest regression*.

Su, X., Yan, X., & Tsai, C.-L. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(3), 275–294.