Crop Yield (Wild BlueBerries) Prediction Using Machine Learning and XAI
BRAC University

Mohammad Waseq-ul Islam 21366001,
Department of Computer Science and Engineering, M.Sc. Computer Science and Engineering
BRAC University, Dhaka, Bangladesh
Email: mdwaseq225@gmail.com,

***Abstract*—Machine Learning is a branch of Artificial intelligence (AI) and it can be used in different sectors like the health sector, transportation, share market, education, agriculture, and others. In this paper, we are not only focusing on Machine Learning, we also focus on Explainable AI (XAI). Machine Learning models or algorithms help us to predict the value or outcomes but it has no transparency and interpretability and for those reasons, AI researchers and practitioners have focused their attention on XAI to help humans to better trust and understand the models [1]. Several Machine Learning Models were employed to evaluate the relative importance of the factors that regulate agroecoculture. Random forest (RF), and extreme gradient boosting (XGBoost), Gradient Boosting(GB) were evaluated as predictive tools. A dimension reduction algorithm helped to reduce the input predefined in the dataset without dropping precision. As a result, clone size, honeybee, bumblebee, Andrena bee species, Osmia bee species, maximum of upper-temperature ranges, the number of days with precipitation, and the fruitset, fruitmass, and seeds were chosen as the best predictor variable subset. The results showed that the Gradient Boosting outperformed other algorithms in all measures of model performance for predicting the yield of wild blueberry by achieving a coefficient of determination (R2) of 0.989, root means square error (RMSE) of 123.971, mean absolute error (MAE) of 97.261. We also included visualization through XAI called SHAP (SHapley Additive exPlanations) for better explainability of the black box Machine Learning models. This study showed that crop yield predictions can be based on computer modeling, using machine learning. Therefore, this research could have a significant impact when there is a scarcity of resources.**

***Index Terms*—Predict Crop Yield, Wild Blueberry, Random Forest, XGBoost, Gradient Boosting**

I. Introduction

The greatest and the most challenging activities in precision agriculture are accurate predictions of crop yield [2]. Extensive research is underway in agriculture to better predict crop yield using machine learning algorithms [3]. Many machine learning models require large amounts of data to provide any reliable result. One of the major challenges in training and experimenting with data is finding sufficient quality samples for both testing and training. The blueberry yields predictive models require data that sufficiently characterize the influence of plant spatial traits, bee species composition, and weather conditions on production.

Our research objective was to develop a predictive model with the assistance machine learning algorithms and eXplainable Artificial Intelligence (XAI) [4] tools such as SHapley Additive exPlainable (SHAP) [5]. Once we determined the best model to predict yield, our study aimed to address how to understand the black box using explainability such as SHAP.

Furthermore, our goal was to determine the most robust model for yield prediction by comparing black box machine learning algorithms while at the same time using a minimal number of features.

This article contributes to the usage of black box models such as XGBoost Regressor to predict the yield of wild blue berries as well as giving a visual representation as to how the machine learning model is performing its task.

II. Method

A. Data Acquired

Generated from Simulation Modelling of Wild Blueberry Pollination by an open source GAMA simulation platform V1.7 (http://gama-platform.org), using GAML (Gamification Modelling Language).

B. Data Description

A total of 77,700 simulations were conducted to achieve both an extensive and intensive sampling effort and this resulted in a dataset consisting of 777 records, each of which is an average of 100 simulation runs e.g., "Obsie, E et al" [2]. Some feature descriptions include the clonesize- The average blueberry clone size in the field per meter square, MinOfUpperTRange - The lowest record of the upper band daily air temperature, AverageRainingDays- The average of raining days of the entire bloom season.

C. Preprocessing and EDA

For machine learning model development and analysis, the calibrated version of the simulation model was used and performed a set of simulation experiments to develop a simulated dataset. The simulation from the dataset was used as experiments that aimed to characterize the influence of wild blueberry spatial arrangement, bee species composition in the field, and weather conditions on yield, as well as the fruitmass and seed amount.

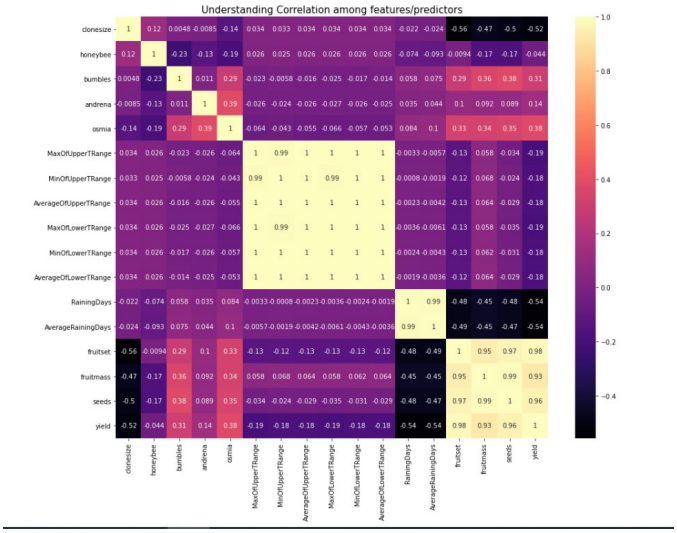

Figure 1. This heatmap generated gives a good picture how strongly connected the features are to each other

Our dataset consisted of 777 records and 17 features, including the target value yield. There was no missing value after thorough inspection. All the entries were numerical and continuous values. So the idea was to use regression machine learning algorithm to conduct our research and predict the production value.

A database analysis library showed a negatively skewed graph on the yield and positive correlation between fruitmass, seeds, fruitset in the exploratory data analysis. The rest of of the features showed ambiguity. For a better understanding of our dataset, we approached for a more informed visualization. A heatmap, see FigureII-C, of all the feature was drawn to show a clear picture of the correlation among all the extracted and selected features in the dataset. Ignoring the diagonal, the map draws a clear picture that the yield is strongly correlated to the 3 features, namely fruitset, fruitmass and seeds. We will further investigate this analyis using XAI.

The data was then checked for any outliers before dividing it into a training and testing dataset. There are loads of ways to removing outliers but the most efficient one that we employed was finding the Interquartile Range or H-Spread. The middle 50% of the data was retained and the outliers below the 1st quantile and above the 3rd quantile are removed. As a result 752 of the total 777 records remain for training and testing.

D. Predictive model development

In a multi-step process, we developed predictive models of wild blueberry yield using 3 machine learning algorithms from the dataset generated by the simulation model. The data used for model development included wild blueberry yield as the dependent variable, bee species composition, weather conditions, fruit data as independent predictor variables.

The dataset was split into training data included 601 randomly selected records, comprising 80% of the total dataset, and testing data included 151 records, comprising the remaining 20% using the "traintestsplit" function, part of the *sklearn* [6] package.

1) Random Forest Regression: The Random Forest algorithm is a type of ensemble method that makes predictions by averaging over predictions of several independent base models [7] . It consists of many decision trees and makes use of bagging, which reduces the variance of a statistical learning model. A majority vote is taken for the predicted class of the decision tree (via bagging and feature randomness), and an average prediction is returned. In our study, we trained and applied RF to predict yields of wild blueberry. Since RF can be used for classification and regression purposes we use the scope of this study to use the random forest as a regression tool.

2) XGBoost: Lately, XGboost has become the most used machine learning model by researchers to predict crop yield. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable [8]. It is an open-source library that implements gradient boosted decision trees that are efficient and highly optimized. XGBoost models require more knowledge and model tuning to get the best accuracy than techniques like Random Forest.

3) Gradient Boosting: Gradient boosting is one of the most powerful techniques for building predictive models. It is a machine learning technique for regression, classification and other tasks, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees [9]. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error.

E. Model Evaluation

We used three metrics to order to evaluate our regression models. First, we used the coefficient of determination (R^2),defined as the proportion of the variance in the response variable that is explained by independent variables. Most of the applications in default uses R squared as a metric for the regression problems. R squared gives us the proportion of the target variable is explained by the feature(s). It provides an indication of the goodness of fit of a set of predictions to the actual values. The following equation is how we determine the coefficient of determination.

$$R^2 = \sqrt{1 - \frac{\sum_{i=1}^{n} (Y_i - X_i)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}} \tag{1}$$

Second, we used the Root Mean Squared Error (RMSE),a measure of the difference between predicted and actual values. RMSE is basically the square root of the MSE and we can see it as follows.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=0}^{n} (Y_i - X_i)^2} \tag{2}$$

Third, the mean absolute error (MAE) was used and is defined as the absolute mean difference between actual and predicted yield values. MAE is easily interpretable. Lower the MAE, better the prediction. The following equation shows how to calculate the MAE.

$$MAE = \frac{\sum_{i=2}^{n} |Y_i - X_i|}{n} \tag{3}$$

F. Principal Component Analysis

We tried to implement Principal Component Analysis in our research to get better accuracy on our models. Using just 2 PCA, the variation of almost the whole dataset was covered with 56% variation for the 1st PC and 32% for the 2nd one. However, using the PCA did not increase the accuracy of the models.

G. K-fold Cross Validation

The cross validation allows us to compare different Machine Learning methods and get a sense of how well they will work in practice and avoid overfitting of the model even if there is a risk of dropping the accuracy of the model. We used 5 fold Cross validation for this study in each model. In practice, it is common to divide the data into 10 blocks.

H. XAI (eXplainable AI)

Over the last decades, great advances in the field of affective computing and affect recognition have been made. Computational models constantly improved to provide more accurate approximations for highly complex human behaviours. However, with their increasing accuracy they gained ever growing attention from companies and non-research facilities. That is why it is strongly emphasized on the fact that those computational models and the application of such systems have to be carefully revised and scrutinized. Furthermore, we argue that when classification or regression results may even lead to harmful events for individuals it is important to fully comprehend the underlying process and analysis that leads to classified classes or regression. And to make the machine models more transparent and understandable for the user is the focus of an XAI [10]. Generally, machine learning models can be distinguished between two types: Inherently interpreteble models and black box models. The former one includes models like Bayesian or decision trees, whereas the black box models include a typical neural network. Our work is to make the black box model more interpretable using XAI. Further XAI approaches include model techniques like model-agnostic (generic explanations irrespective of the model) and model-specific (exploits the underlying inherent structures of the model and its learning mechanism, which in return bounds them to one specific type of model). We used the model agnostic approach, SHAP.

SHAP values interpret the impact of getting a definite value for a given feature as compared to the prediction we'd build if that feature took some baseline value. Various methods and a summary of the SHAP library hava been used in this study.

III. Results

We carried out experiments to evaluate the strength of the predictive models (modern machine learning techniques), identify important factors that affect yield most using Exploratory Data Analysis, seek optimal bee composition, weather conditions and fruit amount that achieve the highest predicted yield.

Models	R Squared	RMSE	MAE
RandomForest	0.988	140.908	110.392
XGBoost	0.989	137.237	109.777
GradientBoosting	0.989	123.971	97.261

Table I
THE TABLE WITH THE EVALUATION METRICS OF THE 3 MODELS

1) Model Evaluation: A comparison of the model was conducted, see Table I. The models were compared based on the coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE) evaluation metrics. The Table I shows that the R^2 values for each model were more or less the same. This meant our models fitted quite well with the observed data. The RMSE value for RF, XGBoost and GB was 140.908kg/ha, 137.908kg/ha and 123.971kg/ha respectively which is 10% less than the average yield (6079.902kg/ha). This shows the models could effectively predict the yield values. The last metric showed the lowest figure in Gradient Boosting. The lower the value of MAE, the better the model. Hence considering all the values in the Table I, we found Gradient Boosting performed the best in terms of predicting the yield.
Other factors, such as accuracy, were also calculated but the values were not very distinguishable among the models.

2) SHAP: Using the Shapley values and SHAP library, we were able to visualize the outcome of the models that were employed. The SHAP library in *Python* has inbuilt functions to use Shapley values for interpreting machine learning models [11]. The library consists of many optimized functions for interpreting tree-based models and a model agnostic explainer function for interpreting any black-box model for which the predictions are known and thus can be used to visualize. A prediction can be explained by assuming that each feature value of the instance is a *player* in a game where the prediction is the payout. Shapley values – a method from coalition game theory – tells us how to fairly distribute the *payout* among the features.
We begin by plotting a bar graph using the Shapley values of the first model, Random Forest. We sort the features by decreasing importance and plot them.

The figure III-2 shows the SHAP feature importance measured as the mean absolute Shapley values. The variable fruitset was the most important feature, changing the predicted yield on average by 1000 on x-axis using Random Forest. Followed by Seeds changing the predicted yield on average by 200 on x-axis.

The visualization was plotted for the other two machine learning models as well, and both showed similar result to

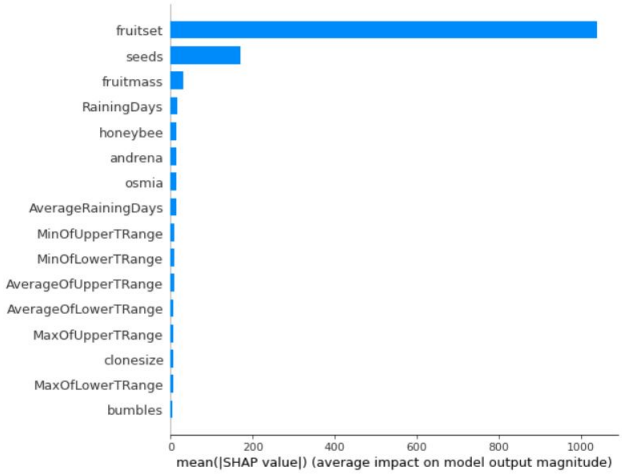

Figure 2. SHAP feature importance measured as the mean absolute Shapley values in a decreasing importance using Random Forest

that of Random Forest, i.e, the variable fruitset was the most important feature. Figure III-2 show the results using XGBoost and Gradient Boosting.

(a)

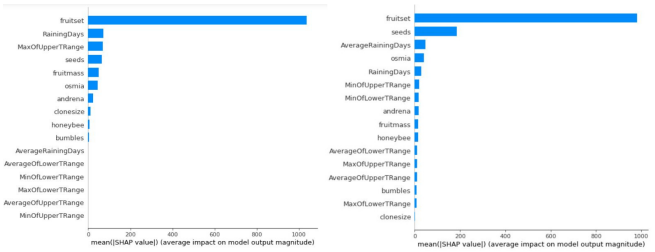

Figure 3. The left graph shows the result using XGBoost. The right graph is after using the Gradient Boosting

For a more detailed explanation using the shapley values, we further developed graphs to summarize a *violin* plot using Gradient Boosting only. Figure III-2 shows the result.

The figure III-2 of the summary plot combines feature importance along with the feature effects. Each point on the summary plot could be a Shapley value for a feature and an instance. The position on the coordinate y-axis is decided by the feature and on the coordinate x-axis by the Shapley value. The color represents the value of the feature from low to high. Overlapping points are plotted in y-axis direction, so we get a sense of the distribution of the Shapley values per feature. The features are ordered in accordance to their importance. . This plot is made of all the dots in the train data. It demonstrates the following information:

- Feature importance: Variables are ranked in negative ascending order.
- Impact: The horizontal location shows whether the effect of that value is associated with a higher or lower prediction.

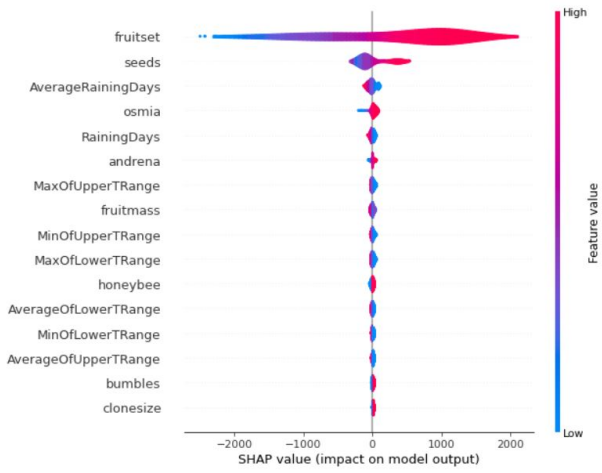

Figure 4. SHAP summary plot using Gradient Boosting

- Original value: Color indicates whether that variable is high (shown in red) or low (in blue) for that observation.
- Correlation: A high level of the fruitset has a high and positive impact on the yield. The "high" comes from the red color, and the "positive" impact is shown on the X-axis.

Similarly, seeds is positively correlated with the target variable yield.

A deeper dive into the XAI SHAP model included the following figure using XGBoost for one instance in the training set.

Figure 5. SHAP explanation force plot using XGBoost

The figure III-2 shows features each contributing to push the model output from the base value (the average model output over the training dataset we passed) to the model output. Features pushing the prediction higher are shown in red and those pushing the prediction lower are in blue. So for this one instance in the training set, *osmia* is pushing the prediction higher, while fruitset, RainingDays, fruitmass push the prediction lower. The base value is 6051kg/ha while the predicted value is 5433.29kg/ha.

Another instance using Gradient Boosting showed a different result which is consistent with our bar plot of the Shapley value.

Figure 6. SHAP explanation force plot using Gradient Boosting

Lastly, we plotted a dependence plot using the most important Shapley value feature. The SHAP Dependence plot shows the marginal effect one or two features have on the predicted outcome of a machine learning model. It gives us an idea whether the relationship between the target and a feature is linear, monotonic or more complex. For the machine learning model, we used the Shapley values attained from Gradient Boosting.

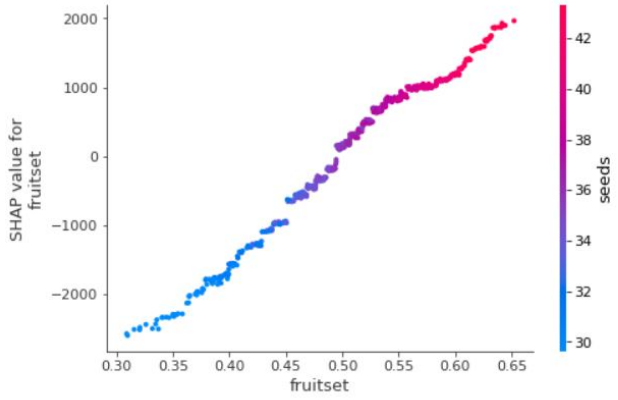

Figure 7. SHAP dependance plot

The figure III-2 shows a dependence plot using fruitset as the key feature. The function automatically includes another variable that the chosen variable interacts most with. The plot shows there is an approximately linear and positive trend between fruitset and the target variable, and fruitset interacts with seeds frequently.

IV. Discussion

We found in our modeling approach that prediction of yield is extremely complex. The black box showed results that were not easily understood. So we used XAI SHAP to get a clear insight as to how the features are interacting with the target variable. Lower yields resulted due to the weathering impacts on bee foraging, while the impact of seeding and fruit amass increased the yield. Our yield prediction model demonstrates that wet rainy springs will greatly reduce yield for farmers in the future, if the trend continues.

V. Conclusion

We investigated three machine learning algorithms: random forest, extreme gradient boosting and gradient boosting algorithms to develop predictive models for wild blueberry yield. The input dataset was produced from a simulation model of wild blueberry pollination [2]. Gradient Boosting performed the best among the three models, followed by XGBoost and then Random Forest.

In general, our study demonstrated that crop yields can be predicted effectively by using data generated with a validated simulation model and XAI. Thus, modeling of agricultural production systems is of paramount importance, especially when there is scarcity of quality data. Also the visualization of such models also holds priority when understanding the model becomes a monument task itself. We hope to further our study using data regarding crops such as paddy, jute etc. to get more sense as to how to effectively increase production.

References

[1] O. Gillath, T. Ai, M. S. Branicky, S. Keshmiri, R. B. Davison, and R. Spaulding, "Attachment and trust in artificial intelligence," *Computers in Human Behavior*, vol. 115, p. 106607, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S074756322030354X

[2] Q. H. . D. F. Obsie, E., "Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms," *Computers And Electronics In Agriculture*, vol. 178, p. 105778, 2020.

[3] M. D. Johnson, W. W. Hsieh, A. J. Cannon, A. Davidson, and F. Bédard, "Crop yield forecasting on the canadian prairies by remotely sensed vegetation indices and machine learning methods," *Agricultural and Forest Meteorology*, vol. 218-219, p. 74–84, 2016.

[4] . F. A. Benk, M., "Explaining interpretable machine learning: Theory, methods and applications," in *SSRN Electronic Journal*, 2020.

[5] . B. J. Rodríguez-Pérez, R., "Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions," *Journal Of Computer-Aided Molecular Design*, vol. 34, 2020.

[6] "learn." [Online]. Available: https://scikit-learn.org/stable/

[7] O. Mbaabu, "Introduction to random forest in machine learning." [Online]. Available: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

[8] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[9] J. Brownlee, "A gentle introduction to the gradient boosting algorithm for machine learning," Aug 2020. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

[10] T. Spinner, U. Schlegel, H. Schafer, and M. El-Assady, "Explainer: A visual analytics framework for interactive and explainable machine learning," *IEEE Transactions on Visualization and Computer Graphics*, p. 1–1, 2019.

[11] P. Banerjee, "Explain your model predictions with shapley values," Feb 2020. [Online]. Available: https://www.kaggle.com/prashant111/explain-your-model-predictions-with-shapley-values