Spring 3-28-2022

SEABEM: An Artificial Intelligence Powered Web Application To Predict Cover Crop Biomass

Aime Christian Tuyishime
University of Nebraska - Lincoln

Andrea Basche
University of Nebraska - Lincoln

SEABEM:

AN ARTIFICIAL INTELLIGENCE POWERED WEB APPLICATION TO PREDICT COVER CROP BIOMASS

An Undergraduate Honors Thesis

Submitted in Partial fulfillment of

University Honors Program Requirements

University of Nebraska-Lincoln

by

Aime Christian Tuyishime, BS

Integrated Science

College of Agricultural Sciences and Natural Resources

March 28, 2022

Faculty Mentors:

Andrea Basche, PhD, Agronomy and Horticulture

ABSTRACT

SEABEM, the Stacked Ensemble Algorithms Biomass Estimator Model, is a web application with a stacked ensemble of Machine Learning (ML) algorithms running on the backend to predict cover crop biomass for locations in Sub-Saharan. The SEABEM model was developed using a previously developed database of crop growth and yield that included site characteristics such as latitude, longitude, soil texture (sand, silt, and clay percentages), temperature, and precipitation. The goal of SEABEM is to provide global farmers, mainly small-scale African farmers, the knowledge they need before practicing and benefiting from cover crops while avoiding the expensive and time-consuming operations that come with blind on-site experimentation. The results were derived from comparing ten different ML algorithms, demonstrating the dominance of ensemble models. The top-performing models - Gradient Boost Regressor, Extra Trees Regressor, and Random Forest Regressor - were stacked together into one model to power the SEABEM web application. As the project is open-sourced on a GitHub repository, the GitHub community is available for others to improve the project. The SEABEM web application is also accessible and valuable to anyone worldwide as its development came from global data.

Key Words: SEABEM, Machine Learning, Cover Crops, Sub-Saharan Africa, Ensemble Algorithms, Biomass Prediction, Random Forest, Gradient Boost, Extra Trees.

SEABEM: AN ARTIFICIAL INTELLIGENCE POWERED WEB APPLICATION TO PREDICT COVER CROP BIOMASS

INTRODUCTION

Healthy and fresh food are essential to humans. Increasing risks to agriculture from climate change necessitate data-driven approaches to food production now more than ever as the world's population grows. Under these conditions, agriculture ultimately requires increases in both efficiency and accuracy as we strive for increased food production in less space. Due to massive soil degradation, food security, poverty alleviation, and social development are all limited in most of Sub-Saharan Africa (SSA). As of 2020, upwards of 1.14 billion people live in SSA, and the population is growing at roughly 3% annually (O'Neill, 2021).

According to the World Bank, as of 2018, 40% of the population lives on less than $1.90 a day, with Sub-Saharan Africa accounting for two-thirds of the world's severely poor. While the poverty rate has dropped from 56% in 1990 to 40% in 2018, poor people continue to climb. In other words, the rate of poverty in Sub-Saharan Africa has not kept pace with population growth, and 433 million Africans are predicted to be living in extreme poverty in 2018, up from 284 million in 1990 (Schoch & Lakner, 2020). Most of Africa's inhabitants living in poverty live in villages and rely on subsistence farming for their livelihood. The majority of rural farming is done by women, and combined these rural women farmers produce the majority of the world's food. However, they suffer significantly from financial insecurity and lack of advanced technologies and skills, limiting them from sustaining quality production on their farms (Schoch & Lakner, 2020).

Sub-Saharan Africa's per capita agricultural productivity keeps falling, contrasting it to the majority of the rest of the world, and soil depletion is a significant contributor to this decrease.

Land degradation in SSA is largely unabated. According to the Global Assessment of Soil Degradation, 65% of African agricultural land has been degraded, 31 percent of permanent grazing land has been damaged, and 19 percent of forest and woodland has been degraded. According to the survey, water and wind erosion account for 46 percent and 38 percent of overall soil deterioration in Africa, respectively (Sivakumar and Wills, 1995). Chemical degradation accounts for 12% of total deterioration, while physical degradation accounts for 4%. Overgrazing (49%) was cited as a cause of soil deterioration in Africa, as was poor agricultural management (28%), deforestation (14%), and overexploitation of vegetation for domestic and industrial purposes (13 percent) (Sivakumar and Wills, 1995).

Agricultural production must be significantly enhanced over the next decade to avoid a catastrophic food shortage in Sub-Saharan Africa (SSA). Food supply across most areas has not managed to keep up with population increase since the 1970s (Boserup, 1965 and Ruthenberg, 1980). As a result, an increase in land pressure caused infertile soils and desertification of soils. Efforts to restore food production in Sub-Saharan Africa require addressing the continent's damaged lands. Soil fertility and organic matter (OM) content has declined, and soil acidity increased due to a long time of farming with no break and no proper soil amendments.

As 80 percent of small-scale farmers in SSA today have fewer than 5 acres of land (approximately 2 hectares). African farmers have used fallowing for almost three millennia to keep their lands productive. However, as the population increases and more land is degraded, it is increasingly difficult for farmers to leave significant portions of their farms fallow in order to support their families (Bunch, 2016). Fallowing has now been forgotten in almost all SSA because of increased pressure of food production to meet the rising population. This decline causes disastrous repercussions for millions worldwide since as fallowing declines, the soil health, fertility, and production decreases, causing food insecurity. Over 150,000,000 rural residents of African countries prone to droughts and other agricultural climatic challenges are at serious risk of starvation in the future as climate change increases. Sub-Saharan Africa has experienced severe famines and hunger crises (Bunch, 2016).

For instance, the famine in East Africa in 2011-2012 was one of the greatest food crises in the last 25 years. Due to tremendous poverty, SSA is a hotbed of chronic hunger. According to the FAO, chronic hunger occurs when a person's daily calorie intake falls below what they require for a healthy and active lifestyle for an extended length of time. The lower limit is 1,800 calories per day on average. About 226.7 million people in Africa are hungry. One in every four people in Sub-Saharan Africa is hungry, making it the continent with the most significant proportion of the world's hungry. (SOS Children's Villages, n.d.).

As a specific example, stunting is a severe form of poor nutrition in quantity and/or quality that irreversibly damages the child physically and mentally. Throughout 25 countries across the world have child stunting levels of above 40% 14 of which are the African countries susceptible to droughts and other climate calamities (Bunch, 2016). It is a significant and long-term health problem. Other places such as Asia and Latin America passed through the same process of ending fallowing. However, they had an advantage of already existing large major industries and highly developed, with resources that enabled people to switch from agriculture in reasonably significant masses while still securing a reasonably good life from those other industries. Unfortunately, Africa still relies on agriculture as a considerable part of the economy.

For instance, smallholder farmers account for more than 60% of the population in Sub-Saharan Africa, and agriculture accounts for roughly 23% of the continent's GDP (Goedde et al., 2019). Farmers can try to boost their crop production by using synthetic fertilizers. However, they are costly, particularly for areas with degraded soils, as they require a significant external input and may be only beneficial in the short term and not after their use has ceased. Even where fertilizers have been subsidized via policies, some producers and policymakers are losing interest because it is not sustainable and ineffective in the long term. Another option is utilizing livestock manure; however, this may not be enough to sustain the whole field fertility. Manufacturing compost is another option, but it is hard to get enough plant biomass in most drought-stricken locations, and this may also be expensive and inaccessible. Cover crops offer an alternative that could be a more viable, sustainable, and long-term option for farmers. Cover crops are crops that would be grown when the soil would be otherwise be bare that can provide erosion prevention, nutrients, weed suppression, and other benefits. Overall, intercropping cover crops provide fallowing and crop production benefits at the same time. On 5 acres of land, over 100 tons of cover crop biomass can be generated, whereas it would be tough to produce that amount of compost (Bunch, 2016). This biomass can sustain farm fertility and restore degraded lands to highly fertile soils. After harvesting, cover crops serve as nutrient replenishment for the soils. They add organic matter to the soil, which improves the soil structure and retains the majority of the rain and irrigated water (Findlay & Manson, 2011). They also maintain the topsoil and increase the soil carbon level, aiding the growth and development of beneficial soil microbiome (Findlay & Manson, 2011). However, growing certain non-commercial crops looks like a bygone era. Modern agricultural tactics such as spraying synthetic fertilizers and pesticides have boosted crop yields and lowered labor costs. Unfortunately, they also inadvertently altered the soil microbiome, which is essential to plant productivity soil health and develops soil resilience. Cover crops, when optimized for particular benefits, can enable a significant reduction of fertilizer reliance. They are an excellent option for soil nutrient amendments, especially in places where other options like fertilizers are expensive and scarce, and where temperature and precipitation do not typically limit plant growing seasons. It is crucial that farmers also focus on the biological component of agriculture, not just the chemical and physical components.

The ability of cover crops to suppress invasive weeds such as cogongrass (*Imperata cylindrica)*, which brings competition to crops in many areas, is a great benefit. The biomass produced by cover crops can also be utilized as livestock feed, assisting in creating mechanisms to limit cropland-damaging wildfires (Buckles, 1998). Moreover, cover crops remove some expenses like transportation as organic matter is created throughout the field. It removes work like using animals to pack, carrying on the back, or spreading with hands that other strategies require. Also, most cover crops are a great source of nutrients for humans; hence farmers can either use them as a source of food or earn additional income by selling them on local markets.

The evidence given by Gaofu Qi et al. (2020) suggests that cover crops aid in restoring degraded land and preserving intensive agricultural operations. Farmers are experimenting with different cover crops in the developing world for several decades. In Central America, Mucuna spp., an invasive Asian legume that is now common in the tropical regions, was successful and is noteworthy. Mucuna spp. has been used worldwide since the 1990s by many small-scale farmers, mainly from Guatemala, Honduras, and Mexico. (Bunch 1993; Buckles 1995; Arteaga et al. 1997; Flores 1997; Buckles et al. 1998). Many farmers in south Brazil employ a variety of cover crops to increase soil health and fertility, manage weeds, and produce livestock forage feed (Calegari et al., 1997). Many Southeast Asian countries have also adopted traditional and new enhanced agricultural systems, which occasionally involve cover crops since the 1990s (Cairns 1997). The performance of these methods shows the potential and benefits of beginning cover-crop research practices, especially in places like Africa, where it is still not highly practiced.

The biggest challenge for African farmers is determining the best places with the most cover crop growing potential. This study aimed to develop a tool usable by an average farmer to estimate the expected cover crop performance in their specific field. This can allow farmers to understand some basic performance information and therefore not practicing them blindly, which might cause losses or decide not to practice cover crops at all due to uncertainty. Currently, the available tools for crop yield prediction are either expensive for an average African small-scale farmer or require tech skills that local farmers might not have. SEABEM is a free and easy-to-use web application with user-friendly interphase that can be accessed here.

The goal of this study was to see if machine learning techniques could be used to estimate cover crop biomass from various sites across Africa. The specific research questions included: (a) Can machine learning techniques be utilized to estimate cover crop biomass? (b) How well can ML algorithms predict cover crop biomass in Africa? The specific goals were to (a) develop a tangible ML-based model for predicting cover crop biomass expected from farms (b) develop an ML-based application that farmers can use to estimate the performance of cover crops in their field.

METHODS

A variety of crop yield prediction systems have been deployed for various reasons. These include APSIM, Aqua-Crop, DSSAT, Patched-Thirst, and SOYGRO, which have been utilized to analyze the implications of various agricultural management approaches, climate change, and other factors on crop production and farm sustainability (Ciscar, 2018). Some of the systems commonly used to analyze small-scale farming operations in Sub-Saharan Africa (SSA) (Matthews, 2002). However, these models are limited by several issues, such as the unavailability of data to set, test, and evaluate them before deploying them and insufficient training to integrate computational methods in small-scale farming operations in SSA properly (Luedeling, 2016; Matthews, 2002). The use of several of these technologies in various regions of SSA is hampered by a lack of good quality climate, crop, and soil data (Waongo, 2014; Zinyengere, 2015).

Machine learning (ML) methods have evolved as a viable complement to traditional crop yield modeling (Crane-Droesch, 2018; Liakos, 2018; Mishra, 2016). In comparison to other agriculture research areas like livestock, water, and soil, ML techniques are rapidly being used in yield prediction studies (Liakos, 2018). Crop production forecasting using machine learning approaches can be done locally, regionally, and nationally (Kaul, 2005). Machine learning is a branch of artificial intelligence that allows a model to train from samples of experience data without programming and to build it explicitly. A class of algorithms enables computer programs to improve their forecasting accuracy for research model results (Crane-Droesch, 2018; Arthur, 1959). The core idea behind machine learning is to create models that can process data to make predictions using statistical analysis and automatically recalibrating on newly available data. The significant strengths of machine learning are deriving new information and detecting patterns from large data sets.

Furthermore, ML models can be refined over time, increasing prediction accuracy, and ML may be used in various industries, particularly agriculture (Balducci, 2018). In an extensive data collection, machine learning technologies enable the discovery of valuable complicated correlations (Gorni, 1997). Moreover, when essential data are missing, the ML technique fills in the gaps (Wolfert, 2017), a circumstance that is all too prevalent in SSA (Waongo, 2014; Zinyengere, 2015).

SEABEM sorts and selects the most successful cover crops for a given set of instances using datasets of labeled information from successful farms, including crop, soil, and climate data. SEABEM was created by combining datasets from successful farms with supervised ML to create a model that works on both aspects. It was crucial to use data variables that are generally accessible to a small-scale farmer. The climate data used to train the model includes temperature and precipitation. The soil data used in training includes textural information: sand, silt, and clay soil percentages. The crop data used in training is the crop biomass, which was the dependent variable. By converting the obtained data into a vector form understandable by the ML algorithm, statistical and mathematical techniques were used to normalize and clean the inputs data by removing incomplete data and outliers. After training and feeding the vector into the algorithm, the model predicts cover crop biomass expected from a farming site in tons per hectare. More on the python code of the dataset cleaning and model development can be found in the project GitHub repository linked here.

Model Problems Characteristics

Various essential elements of the challenge of estimating crop production significantly impacted the immediate design decisions made while building the ML model. The majority of these characteristics are related to data issues. Data quantity was the primary challenge. The initial method was developing a dataset containing past cover crop research studies in Africa. These data were exact and high-quality, but the size of the dataset was limited, with just about ten relevant research papers, which totaled to about 40 sites (rows). There is little research on cover crops in Africa; hence, it is hard to find enough data to train an ML algorithm with a slight bias. Also, the data content was inconsistent as they contained different variables depending on the paper's research subject.

Then the alternative method was looking at publicly available datasets in this field. The selected paper was "A global experimental dataset for assessing grain legume production" from all of the materials I looked at due to its thoroughness and simplicity to use. This dataset was built by using one hundred seventy-three (173) articles that span five continents, comprising about 200 variables, and include more than 30 cover crop species (Hestie, 2009; Cernay, 2016). Other datasets found had data formats that are hard to use, such as images and unique file formats. On the other hand, this selected dataset included a user-friendly, appropriately detailed MySQL database, hence extremely accessible and understandable. Also, using a global dataset enabled developing an ML model that has a general scope of cover crop biomass production with a global scale and usability.

It is essential to increase the data quality and filter out the invaluable parts. Categorical data was turned into numerical data, which was also cleaned before using the ML model. For instance, the soil texture classification name was used to fill missing sand, silt, and clay soil texture percentages using the soil texture triangle. Univariate analysis was used to understand the data structure and then used a confusion matrix and plots to evaluate correlations and distributions. No outliers were found. There was a considerable quantity of incomplete data in several columns, despite the efforts in the initial phase being solely focused on filling the data. The data lacking was mostly for soil, such as pH, which is hard to point to a particular location.

In such circumstances, the remaining option was to use the distribution average in interest to replace incomplete data. Since there were no outliers in the dataset, the average was adequate for this purpose. The data must be normalized afterward using the z-score, a popular statistics strategy. It defines a data point in relation to the mean and standard deviation of the whole set by mapping the data to a distribution with a zero mean and a standard deviation of one (Curtis et al., 2016).

Machine Learning

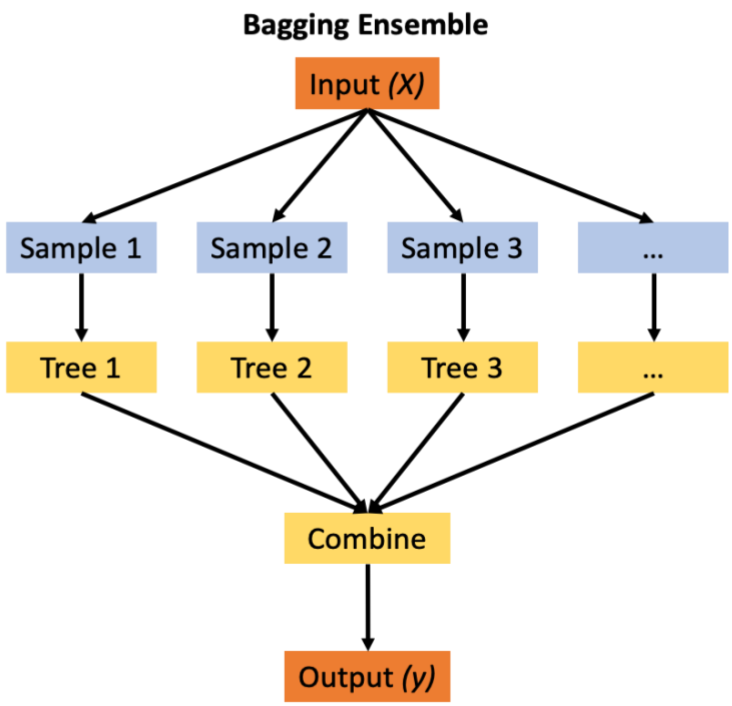
Regression analysis was employed to predict outcomes based on pairings of outputs and inputs from sampling data. In other words, the intention was to convert the agricultural parameters into a continuous function. It is also a supervised ML problem that uses the dataset for training and tries to produce a relevant prediction. Each sample is a pillar in supervised learning, comprising a vector quantity input value and an imposed output, also known as a supervisory signal. A supervised learning algorithm first accomplishes the analytical process using training samples, then creates a contingent function for future use in predicting novel cases (Praveena, 2017).

Ensemble Learning

Ensemble learning is a broad integrative method to machine learning that combines estimates from different models to improve forecasting performance. Although there appears to be no limit to the number of ensembles that can be used to solve the predictive modeling challenge, the area of ensemble learning is dominated by three approaches, namely bagging, stacking, and boosting.

Bagging

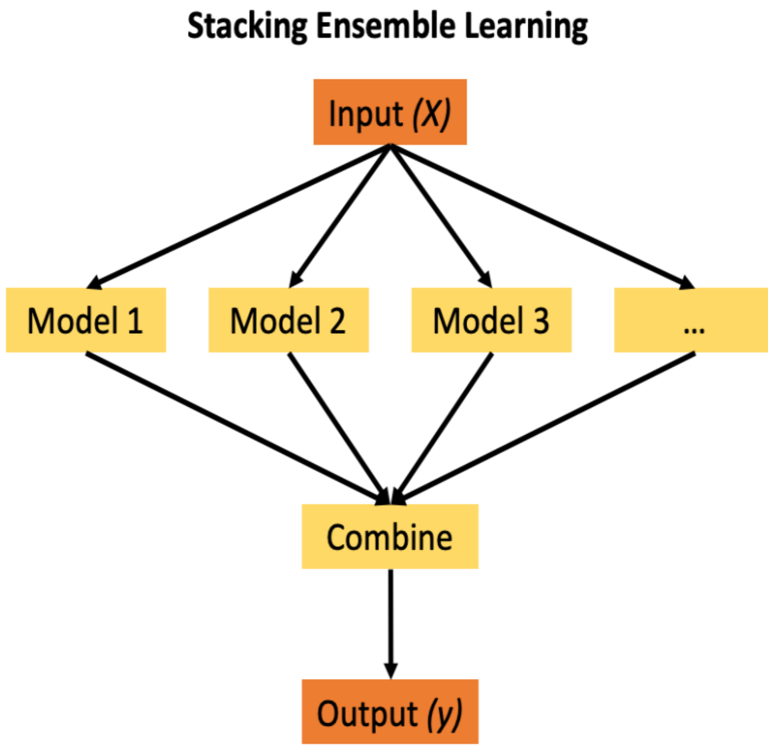
The algorithms tested in this research using bagging are the Random Forest Regressor and Extra Trees Regressor. Bagging is an ensemble learning algorithm that varies the train data to find a varied collection of ensemble members. Bootstrap AGGregatING inspired the term Bagging indicating its two main parts, bootstrap and aggregation. Bagging usually entails training each model on a distinct sample of the same training dataset using one algorithm, mostly just an unpruned decision tree. The ensemble members' forecasts are then merged using simple statistics like voting or average (Zhou, 2012). The way each data instance is handled to train ensemble members is crucial to the process. Each algorithm is given a different sample of the data. The instances (rows) are randomly chosen from the dataset in the bootstrap sampling method but with replacement. The rows are returned in the data for possible multiple re-sampling for training. This strategy for estimating the significance level of a data sample is commonly used in statistics with small datasets. A significantly better prediction can be accomplished by producing numerous separate iterations, estimating a statistical value, and finding the mean of the estimates rather than predicting directly from the dataset. Multiple independent training datasets can be created simultaneously and then utilized to build an estimator. Averaging the forecasts across the algorithms usually yields better results than fitting a single algorithm straight to the training dataset (Zhou, 2012).



Stacking Ensemble Learning

Stacking, also known as “stacked generalization,” is an ensemble approach for finding a varied group of algorithms by shifting their fit on the training data and combining results with a model

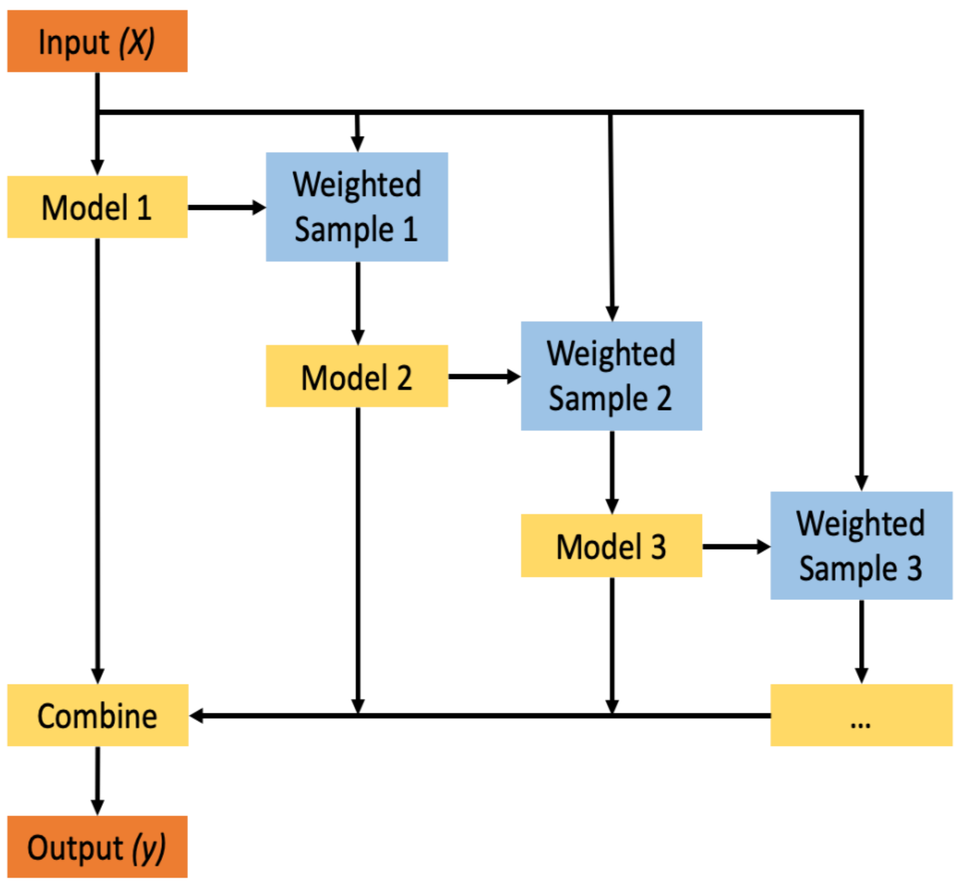
(Zhou, 2012). It is generally used when a model is taught to mix many models. The models are classified into two groups. The individual models to be combined are level-0 or first-level models, and the combined model is a level-1 model or second-level model (Rokach, 2010). This methodology was used to stack the best three models and develop the model running on the backend of the SEABEM web application.



Boosting Ensemble Learning

The algorithms tested in this research using boosting are the Ada Boost Regressor and Gradient Boost Regressor. Boosting is an ensemble strategy that attempts to alter the training data to focus attention on cases that prior fit algorithms on the training dataset have incorrectly identified (Zhang, 2012). The idea of rectifying forecast errors is an essential feature of boosting ensembles. The estimators are fitted and added to the ensemble in order, with the second estimator attempting to correct the first estimator's predictions, the third estimator correcting the second estimator, and so on. Voting or averaging is used to integrate the forecasts of the poor learners, though their contributions are weighted proportionally to their performance or competence. The goal is to use estimators that might be weak to develop one stronger estimator (Zhang, 2012). The training dataset is usually left alone, and the estimator is tweaked to focus on certain instances based on whether previous models correctly or wrongly forecasted them. For example, the instances can be graded to indicate how much attention an estimator should devote to the model while learning (Zhou, 2012).

Boosting Ensemble



Linear Models

The algorithms tested in this research based on linear models are Linear Regression, Ridge CV, and Elastic Net. Linear models are a type of model that describes a dependent variable as a linear combination of independent variables. The response should be a continuous variable with a distribution that is at least somewhat normal (Brownlee, 2020). However, these models have many uses; they cannot handle discrete or skewed continuous responses. A linear model is created using linear regression, a statistical procedure. The model depicts the relationship between one or more predictor variables, Xi (also known as independent variables) and a response variable y (also known as the dependent variable). Below is the general formula for a linear model:

$$y = \beta_0 + \sum \beta_0 X_i + \epsilon_i$$

where β is the linear estimated coefficients to be computed and ϵ denotes the error terms.

Creation of a validation data set and test harness

Cross-validation was used to assess the model's capacity to make predictions for first seen data different from the data used in the estimation process, in order to identify flaws like overfitting or selection bias (Dangeti, 2017), as well as to provide understanding into how the model will perform to a different first seen dataset. To measure accuracy, a tenfold cross-validation method was applied. The approach divided the dataset into ten segments, with train and test on one, and was done for all possible train–test splits. This was calculated as a percentage of accurately predicted cases to the total dataset cases.

Building algorithms

The analysis and plotting were carried out in Python3 using the Scikit-learn and Matplotlib libraries. Before every run, the random number seed was reset to its original position to guarantee that each method was evaluated using identical data splits. It assured that the outcomes were comparable. The complete code can be found in the GitHub repository.

Metrics

The next step was to establish metrics to rate the model once trained and tested (Brownlee 2015). However, it is pointless to establish a measure without properly contextualizing the model. Continuous variables are used in a regression model; hence the objective is to make predictions extremely near the actual values. Therefore, Mean Square Error (MSE) was chosen as the metric. The MSE reflects how far the model's results vary from the true parameter values. MSE is never negative, and the closer it is to zero, the more accurate the estimator is. The MSE is calculated from the following equation:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

y is the vector of actual values of the parameters being estimated, producing a vector ŷ of n predicted values of the whole data sample.

The models' accuracy was also rated using the coefficient of correlation (R2), which ranges from -1 to 1, indicating the precision of the model. The further the R2 is from 0, the better the model fits the data.

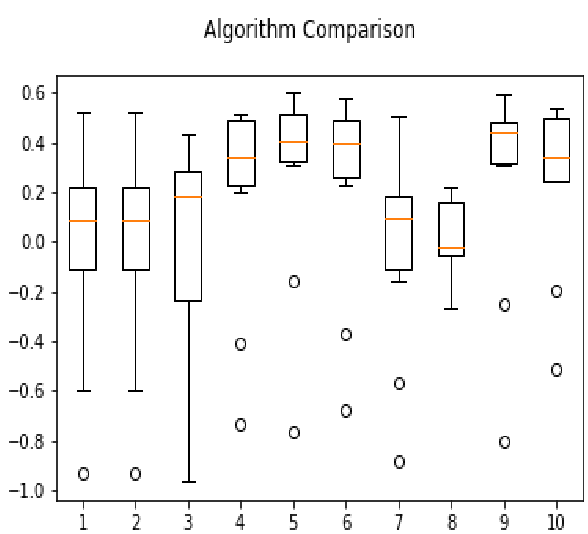
RESULTS

The findings are based on comparing ten different algorithms, demonstrating the superiority of algorithms based on an ensemble of decision trees, particularly boosted trees, where the Gradient Boost Regressor was our top estimator. It was followed by Extra Trees Regressor in second place and Random Forest Regressor in third place. They had an accuracy of 0.52406, 0.51894, and 0.51716 and a mean square root of 2.17229, 2.18395, and 2.18798, respectively. Generally, linear models had the worst performance. Linear Regression, Ridge CV, and Elastic Net had a test accuracy of 0.22128, 0.22125, and 0.22032 and a mean square root of 2.77865, 2.77870, and 2.78036, respectively, as shown by Table 1.

The cross-validation accuracy in Figure 1 also supports these results. The Gradient Boost, Extra Trees, Random Forest regressors (Algorithm 4, 5, and 6) had the least standard deviations. In contrast, Linear Regression, Ridge CV, and Elastic Net (Algorithm 1, 2, and 7) had massive standard deviations, as shown by Figure 1.

A stacked and a voting regressor combining multiple algorithms also had performances close to top ensemble algorithms. They had a test accuracy of 0.50850 and 0.45848 and a mean square error of 2.20751 and 2.31713, respectively, as shown in Table 1. They also had the least standard deviation among all ten algorithms, as shown in Figure 1.

Figure 1: Cross-Validation Accuracy



1: Linear Regression

2: Ridge CV

3: Ada Boost Regressor

4: Extra Trees Regressor

5: Random Forest Regressor

6: Gradient Boost Regressor

7: Elastic Net

8: Supper Vector Regression

9: Stacking Regressor

10: Voting Regressor

Table 1. Algorithm performance on the test dataset (Accuracy on the test dataset)

Algorithm	**RSQ**	**RMSE**
Linear Regression	0.22128	2.77865
Ridge CV	0.22125	2.77870
Ada Boost Regressor	0.35418	2.53046
Extra Trees Regressor	**0.51894**	**2.18395**
Random Forest Regressor	**0.51716**	**2.18798**
Gradient Boost Regressor	**0.52406**	**2.17229**
Elastic Net	0.22032	2.78036
Supper Vector Regression	0.12808	2.94023
Stacking Regressor	0.50850	2.20751
Voting Regressor	0.45848	2.31713

DISCUSSION

The use of machine learning was prompted by the need for the system to understand input data from different users swiftly. Despite the small training dataset and only seven variables (longitude, latitude, sand%, silt%, clay%, temperature, and precipitation) used, the model could reasonably estimate cover crop biomass from various farm sites with varying degrees of accuracy. Identifying the variance patterns in the output data is crucial since it exposes concerns that need to be addressed when using ML predictions. For instance, the models can be improved with additional data for training in new environmental features.

The ensemble models generally performed better than the linear models because ensemble models handle inconsistency in the data, like skewness, than linear models (Brownlee, 2020).

The SEABEM system was developed using a stacked algorithm of the best three models, namely Gradient Boost, Extra Trees, and Random Forest regressors, which are all ensemble algorithms. Small-scale farmers can depend on it to make biomass predictions that can be included in future cover crop introduction and management decisions, mainly in Africa. The SEABEM system is free and accessible to an average African farmer and is easy to use with no prior knowledge of prediction models. It will boost the sustainability of small-scale agricultural systems in Africa by helping farmers make informed and data-based decisions.

CONCLUSION

The SEABEM is a machine learning-powered, open-source application that allows anyone to enter field location (latitude and longitude), soil data (texture percentages), and climate data (temperature and precipitation) and gets instant cover crop biomass predictions. This approach gives global farmers, mainly small-scale African farmers, the knowledge they need before practicing cover crops while avoiding the expensive and time-consuming operations that come with blind on-site experimentation. This prior insight on expected performance enables farmers to conduct effective, profitable food production using their field's primary, easily accessible data. The main challenge faced while developing the model running on the backend of the application is the shortage of publicly available data on cover crops. However, the stacked estimator combining three regressor algorithms, Gradient Boost, Extra Trees, and Random Forest, reasonably estimated the cover crop biomass. A more extensive dataset can significantly produce better results. As the project is open-sourced on a GitHub repository, the GitHub community is welcome and encouraged to improve the project.

Web App Link: https://share.streamlit.io/aimechristian/biomasspredictor/main/streamlit.py

GitHub Repository: https://github.com/aimechristian/BiomassPredictor

REFERENCES

Arteaga, L.; Carranza, T.; Elitta, M.; González, M.; Guerrero, C; Guevara, F.; Herrera, B.; Lopez, A.; Martinez, F.; Mendoza, A.; Narváez, G.; Puentes, R.; Reyes, H.; Robles, C.; Sohn, I.; Triomphe, B. 1997. El uso de sistemas de cultivo con plantas de cobertura en algunas comunidades del sureste mexicano: contexto, resultados y lecciones aprendidas.

Arthur, S. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229.

Balducci, F., Impedovo, D., & Pirlo, G. (2018). Machine learning applications on agricultural datasets for smart farm enhancement. Machines, 6(3), 38.

Boserup, E., & Chambers, R. (2014). The conditions of agricultural growth: The economics of agrarian change under population pressure. Routledge.

Brownlee, J. (2016). Machine learning mastery with Python: understand your data, create accurate models, and work projects end-to-end. Machine Learning Mastery.

Brownlee, J. (2020, August 14). Linear regression for machine learning. Machine Learning Mastery. https://machinelearningmastery.com/linear-regression-for-machine-learning/

Buckles, D. (Ed.). (1998). Cover crops in West Africa: Contributing to sustainable agriculture.
IDRC.

Buckles, D. 1995. Velvetbean: a “new” plant with a history. Economic Botany, 49(1), 13–25.

Buckles, D.; Triomphe, B.; Sain, G. 1998. Cover crops in hillside agriculture: farmer innovation with Mucuna. International Maize and Wheat Improvement Center; International Development Research Centre, Ottawa, Canada. 218 pp.

Bunch, R. (2016, February 23). Green manure crops in Africa: A report from the Field. Food First. https://foodfirst.org/green-manure-crops-in-africa-a-report-from-the-field/

Bunch, R. 1993. What we have learned to date about green manure crops for small farmers (2nd ed.). International Cover Crop Clearing House, Tegucigalpa, Honduras. Technical Report 3. 8 pp.

Cairns, M. 1997. Indigenous fallow management in Southeast Asia, Rural Extension and Agricultural Research Institute of Santa Catarina, Santa Catarina, Brazil. 14 pp.

Calegari, A.; do Prado Wildner, L.; de Freitas, V. 1997. Adubaçâo verde e sistemas de cobertura do solo na regiâo sul do brazil. Rural Extension and Agricultural Research Institute of Santa Catarina, Santa Catarina, Brazil.

Cernay, C., Pelzer, E., & Makowski, D. (2016). A global experimental dataset for assessing grain legume production. Scientific data, 3(1), 1-20.

Ciscar, J. C., Fisher-Vanden, K., & Lobell, D. B. (2018). Synthesis and review: an inter-method comparison of climate change impacts on agriculture. Environmental Research Letters, 13(7), 070401.

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters, 13(11), 114003.

Curtis, A. E., Smith, T. A., Ziganshin, B. A., & Elefteriades, J. A. (2016). The mystery of the Z-score. Aorta, 4(04), 124-130.

Dangeti P (2017) Statistics for Machine Learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R. Packt Publishing

Findlay, N., & Manson, A. (2011, February). COVER CROPS: WHAT ARE THEY AND WHY ARE THEY USED. KZN Agriculture & Rural Development.

Flores, M. 1997. El uso de cultivos de cobertura en centroamerica: mas allá del entusiasmo: retos y oportunidades. Rural Extension and Agricultural Research Institute of Santa Catarina, Santa (Catarina, Brazil, 1997)

Goedde, L., Ooko-Ombaka, A., & Pais, P. (2019, February 15). Winning in Africa's agricultural market. McKinsey & Company. https://www.mckinsey.com/industries/agriculture/our-insights/winning-in-africas-agricultural-market

Gorni, A. A. (1997). The application of neural networks in the modeling of plate rolling processes. JOM-e, 49(4), 252-260.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

Kaul, M., Hill, R. L., & Walthall, C. (2005). Artificial neural networks for corn and soybean yield prediction. Agricultural Systems, 85(1), 1-18.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674.

Luedeling, E., Smethurst, P. J., Baudron, F., Bayala, J., Huth, N. I., Van Noordwijk, M., ... & Sinclair, F. L. (2016). Field-scale modeling of tree–crop interactions: Challenges and development needs. Agricultural Systems, 142, 51-69.

Matthews, R., Stephens, W., Hess, T., Middleton, T., & Graves, A. (2002). Applications of crop/soil simulation models in tropical agricultural systems.

Mishra, S., Mishra, D., & Santra, G. H. (2016). Applications of machine learning techniques in agricultural crop production: a review paper. Indian J. Sci. Technol, 9(38), 1-14.

Muchena, F. N., Onduru, D. D., Gachini, G. N., & De Jager, A. (2005). Turning the tides of soil degradation in Africa: capturing the reality and exploring opportunities. Land Use Policy, 22(1), 23-31.

O'Neill, A. (2021, July 21). Sub-Saharan Africa - population growth 2010-2020 | Statista. Statista. https://www.statista.com/statistics/805619/population-growth-in-sub-saharan-africa/

Praveena, M., & Jaiganesh, V. (2017). A literature review on supervised machine learning algorithms and boosting process. International Journal of Computer Applications, 169(8), 32-35.

Qi, G., Chen, S., Ke, L., Ma, G., & Zhao, X. (2020). Cover crops restore declining soil properties and suppress bacterial wilt by regulating rhizosphere bacterial communities and improving soil nutrient contents. Microbiological Research, 238, 126505.

Rokach, L. (2010). Pattern classification using ensemble methods (Vol. 75). World Scientific.

Ruthenberg, H., MacArthur, J. D., Zandstra, H. D., & Collinson, M. P. (1980). Farming systems in the tropics (Vol. 97). Oxford: Clarendon Press.

Schoch, M., & Lakner, C. (2020, December 16). The number of poor people continues to rise in sub-Saharan Africa, despite a slow decline in the poverty rate. World Bank Blogs. https://blogs.worldbank.org/opendata/number-poor-people-continues-rise-sub-saharan-africa-despite-slow-decline-poverty-rate

Sivakumar, M. V. K., & Wills, J. B. (1995). Combating Land Degradation in Sub-Saharan Africa Summary Proceedings of the International Planning Workshop for a Desert Margins Initiative 23-26 Jan 1995, Nairobi, Kenya. International Crops Research Institute for the Semi-Arid Tropics.

SOS Children’s Villages. (n.d.). Hunger in Africa: Facts, causes, consequences. SOS Children's Villages. https://www.sos-usa.org/about-us/where-we-work/africa/hunger-in-africa?

Waongo, M., Laux, P., Traoré, S. B., Sanon, M., & Kunstmann, H. (2014). A crop model and fuzzy rule based approach for optimizing maize planting dates in Burkina Faso, West Africa. Journal of Applied Meteorology and Climatology, 53(3), 598-613.

Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. J. (2017). Big data in smart farming–a review. Agricultural systems, 153, 69-80.

Zhang, C., & Ma, Y. (Eds.). (2012). Ensemble machine learning: methods and applications. Springer Science & Business Media.

Zhou, Z. H. (2012). Ensemble methods: foundations and algorithms. CRC press.

Zinyengere, N., Crespo, O., Hachigonta, S., & Tadross, M. U. N. D. P. (2015). Crop model usefulness in drylands of southern Africa: an application of DSSAT. South African Journal of Plant and Soil, 32(2), 95-104.

APPENDIX I: Cover Crop Species

1. *Cicer arietinum:* Chickpea
2. *Vicia faba:* Fababean
3. *Lens culinaris:* Lentil
4. *Pisum sativum:* Garden pea
5. *Vigna radiata:* Mungbean
6. *Arachis hypogaea:* Peanut
7. *Glycine max:* Soybean
8. *Triticum aestivum:* Common wheat
9. *Lupinus albus:* White lupine
10. *Lupinus angustifolius:* Narrowleaf lupine
11. *Lupinus atlanticus:* Lupinus atlanticus
12. *Lupinus pilosus:* Blue lupine
13. *Vicia narbonensis:* Purple broad vetch
14. *Lathyrus cicera:* Red pea
15. *Lathyrus ochrus:* Cyprus vetch

16. *Lathyrus sativus:* White pea

17. *Vicia benghalensis:* Purple vetch

18. *Vicia sativa:* Garden vetch

19. *Lupinus luteus:* Yellow lupine

20. *Avena sativa:* Common oat

21. *Brassica napus:* Rape

22. *Brassica rapa:* Field mustard

23. *Brassica juncea:* Brown mustard

24. *Triticum turgidum:* Rivet wheat

25. *Crambe abyssinica:* Crambe

26. *Carthamus tinctorius:* Safflower

27. *Helianthus annuus:* Common sunflower

28. *Phaseolus vulgaris:* Kidney bean

29. *Hordeum vulgare:* Common barley

30. *Triticum sativum:* Triticum sativum

31. *Vigna subterranea:* Bambarra groundnut

32. *Vigna unguiculata:* Cowpea

33. *Cajanus cajan:* Pigeonpea

34. *Sorghum bicolor:* Sorghum

35. *Linum usitatissimum:* Common flax

36. *Fallow:* Fallow

37. *Panicum miliaceum:* Proso millet

38. *Triticum durum:* Durum wheat

39. *Triticosecale:* Triticale

40. *Vigna mungo:* Black gram

41. *Vigna aconitifolia:* Moth bean

42. *Cyamopsis tetragonoloba:* Guar

43. *Lablab purpureus:* Hyacinthbean

44. *Phaseolus lunatus:* Sieva bean

45. *Fagopyrum esculentum:* Buckwheat

46. *Brassica chinensis:* Pak choi

47. *Vicia ervilia:* Blister vetch

48. *Brassica campestris:* Brassica campestris

49. *Pennisetum glaucum:* Pearl millet

50. *Vigna angularis:* Adzuki bean

51. *Vicia pannonica:* Hungarian vetch

52. *Secale cereale:* Cereal rye

53. *Lathyrus aphaca:* Yellow pea

54. *Vicia articulata:* Oneflower vetch

55. *Vicia villosa:* Winter vetch

56. *Vicia hybrida:* Hairy yellow vetch

57. *Ricinus communis:* Castorbean

58. *Trifolium repens:* White clover

59. *Sinapis alba:* White mustard

APPENDIX II: Benefits of Cover Crops (Findlay & Manson, 2011)

Soil Fertility Management

Increased soil fertility is among the most common uses of cover crops. They produce between 45 and 220 kg ha-1 N per year of nitrogen. The quantity of accessible nitrogen is determined by the species planted, crop biomass generated, and N proportion in crop tissues. The amount of nitrogen remaining to a subsequent crop from green manure is typically 40-60% of the total nitrogen present in legumes. Also, cover crops acquire other nutrients other than nitrogen like phosphate, potassium, calcium, magnesium, and sulfur. Green manure or mulch replenishes the soil nutrients through their decomposition by microbes. After organic acids, produced through decomposition, interact with other minerals, exchangeable nutrients like phosphate get discharged into the soil.

Soil Quality Management

Cover crops can significantly boost soil quality by gradually increasing the organic matter content in the soil. Substances difficult to decompose like waxes are generated during the breakdown of organic matter in the soil, improving soil structure by holding the soil particles together and forming aggregates. Aggregates increase water infiltration in the soil and aeration. Compacted soil is broken down by strong and long cover crops roots like radish roots, while shallow roots break compacted surfaces.

Erosion Management

Soil productivity can be irreversibly reduced by erosion. Cover crops can be grown to reduce the raindrop impact on the soil by directly decelerating the speed of precipitation prior to reaching the topsoil, hence avoiding soil runoff. Root channels also help to keep the soil structure, limiting soil movement. According to studies, farms with a cover crop during the winter have 55 percent less soil erosion and 50 percent less soil loss than farms without a cover crop. Covering the soil with a combination of a legume and a grass cover crop promotes surface cover while also delivering nutrients, especially nitrogen, beneficial to the next crop.

Water Management

Cover crops operate as a protective border at the topsoil, slowing the amount of water that escapes from a farm. Organic matter from decomposed cover crops enhances the strength of soil aggregates by preventing them from tearing due to rainfall pressure, which lowers pore blockage and the production of soil crusts. Soil water infiltration and retention rises, while erosion during periods of high rainfall decreases. The creation of pores in the soil by roots enables proper drainage in the soil instead of just water flow on the soil surface. Mulch also protects the topsoil from losing moisture and water through evaporation. Water infiltration also enables fast recharge of groundwater.

Weed, Disease & Pest Management

Cover crops shade the land and make it harder for weeds to develop and grow. Some allelopathic cover crops like sorghum and rye can also produce toxins that help inhibit surrounding crops like weeds. Cover crops also act like "trap crops" to lure insects and other pests away from valuable crops and into what the pest perceives as a better hospitable environment. Trap crop regions can be created inside crops, fields, or landscapes. Most of the time, trap crops are planted in the same growing season as the main crop. Several types of cover crops are utilized to lure wanted organisms for biological pest control. In no-tillage farming practices, a mulch cover crop can reduce weeds significantly. Cover crops have been linked to a rise in the population of wanted insects. Since these insects operate as biological controls to restore the balance of pest populations, they help prevent severe pest outbreaks. Mulch offers a stable haven for beneficial insects in habitat augmentation until the food crop emerges. When only a food crop is planted, this results in a rise of wanted insects being established at the beginning of the growing season. Farmers save money usually spent on herbicides by using cover crops and at the same time protecting the environment from herbicide toxins.