YieldPredict: A Crop Yield Prediction Framework for Smart Farms

Nitu Kedarmal Choudhary*, Sai Sree Laya Chukkapalli*, Sudip Mittal†,
Maanak Gupta‡, Mahmoud Abdelsalam§, Anupam Joshi*
* *University of Maryland Baltimore County*, Baltimore, MD, USA
† *University of North Carolina Wilmington*, Wilmington, NC, USA
‡ *Tennessee Technological University*, Cookeville, TN, USA
§ *Manhattan College*, Riverdale, NY, USA
*{nituc1, saisree1, joshi}@umbc.edu, †mittals@uncw.edu, ‡mgupta@tntech.edu, §mabdelsalam01@manhattan.edu

***Abstract*—In recent years, machine learning approaches are gaining popularity with the advent of big data. The massive amount of data generated, when served as an input to machine learning approaches, provides useful insights. Adoption of these approaches in the agricultural sector has immense potential to increase crop productivity and quality. In this paper, we analyze the crop data collected from an agriculture site in Rajasthan, India, that includes both Rabi and Kharif cropping patterns. In addition, we utilize a smart farm ontology that contains concepts and properties related to the agricultural domain. We link the collected data and our smart farm ontology to populate a knowledge graph. We utilize the generated knowledge graph to provide structural information and aggregate data by using SPARQL queries. The aggregated data is further used by our machine learning models to predict the crop yield to benefit farmers and various stakeholders. We also analyze and compare our results obtained for various machine learning models used.**

I. Introduction

With the increasing population, the demand for food has risen across the world. According to the United Nations [1], the world's population will be around 8.5 billion by 2030. Therefore, the need for technological advancements in the agriculture sector is deemed necessary in order to increase the crop productivity and prevent imminent food crises. Particularly, monitoring the yield of different crops for every season is an important attribute in estimating crop productivity. Analysis of various factors like soil, climate, temperature, etc. can contribute to crop yield prediction. Incorporation of machine learning (ML) models to analyze these factors present in the agriculture data is an appealing solution, since these models have shown accurate performance for prediction and forecasting problems in various applications such as retail [2], finance [3], healthcare [4], etc. At the same time cybersecurity concerns [5], [6] have also been raised with the deployment of smart faming technologies, which needs proper attention.

In this work, we focused on the agriculture sector in India, since it has the second largest population in the world and food security is of utmost importance, specially in the state of Rajasthan. Agriculture is the backbone of Rajasthan's economy contributing 25.56% of the state's total GDP in 2019-2020 [7]. Further, it provides livelihood to a vast majority of population in rural India. This means that many people depend on the seasonal yield and, hence, crop yield prediction is extremely important for planning and storage purposes. Evidently, it is difficult for the Indian farmers to obtain remote sensed images, vegetation indices, crop genotype data for crop yield prediction. Therefore, insights like crop yield and what factors contribute to increasing the crop yield can help the farmers in making decisions such as the use of right quantity of fertilizers, proper irrigation, etc. For example, an important use case of crop yield prediction is the strategic planning to maximize the crop productivity by using the forecast insights [8]. These obtained metrics can help in protecting the crop, as necessary steps can be taken by the government proactively to increase crop yield and also maintain the food security.

In our research, we utilized crop data obtained from the Open Government Data (OGD) platform, India as well as publicly available data from the agricultural website of Rajasthan [9], [10]. The data collected has two different sets, representing seasons namely Rabi and Kharif. We integrated the collected data set with an existing smart farm ontology [11] which captures the relationship between different entities of data. Further, we populated a knowledge graph that can be queried using SPARQL query language [12]. The results obtained after querying serve as input to the ML models. The reason for incorporating ML models is to provide meaningful insights like predicting crop yield, etc. which in turn aids the farmers in making informed decisions that pertain to high crop yield. Farmers can also use the knowledge in various aspects while planning for the next crop season depending on the predicted yield. Simultaneously, farmers can estimate the crops that give maximum yield for a particular season and can plan accordingly depending on the weather parameters. For example, if the crop needs particular soil nutrients for improving productivity, farmers can plan to use proper fertilizers and try to maximize their crop yield. This way farmers can predict crop yield based on a well trained ML model and use those insights in strategic planning for the next season.

The rest of the paper is organized as follows: Section II discusses the related work. Section III describes architecture of our framework. Section IV explains the performance of machine learning models by comparing the results obtained. Finally, we conclude this paper in Section V.

II. Related Work

Increasing the quality and quantity of the agriculture crop has become a crucial task for the food industry in order to ensure food sustainability for the growing population. Therefore, farmers have started utilizing machine learning (ML) models for managing their farms efficiently by making key decisions with the help of data-driven insights. Earlier researchers mostly implemented non-linear iterative multivariate optimization approaches and empirical piecewise linear crop yield prediction equations that make use of vegetation index and meteorological parameters for predicting crop yield [13].

However, considering the capabilities of ML algorithms such as data-driven insights in real-time scenarios widely used in various CPS domains, researchers have started implementing these algorithms in agriculture for better management and efficiency of the agricultural farms. For example, Behmann et al. described how ML models can be used for protecting the crop in advance by detecting the biotic stress [14]. Another example is irrigation recommendations provided by the ML algorithms by utilizing farm data such as soil moisture, weather, etc [15]. Similarly, Khaki and Wang developed a deep neural network model that significantly outperformed methods such as Lasso, shallow neural networks and regression tree [16] as part of the 2018 Syngenta Crop Challenge where participants had to predict yield performance in 2018. In order to predict the rice yield, data from the Indian state of Maharashtra which had parameters like precipitation, minimum temperature, average temperature, maximum temperature, etc. was utilized as an input to the multilayer perceptron neural network where the model showed an accuracy of 97.5% with a sensitivity of 96.3 and specificity of 98.1 [17].

Olivera et al. developed a novel yield forecast system that has fewer data requirements compared to existing solutions that depend on large amounts of remote sensing data. Further, the developed system works on large regions and provides forecasts at a resolution compatible with best input data resolution. The system consists of a Recurrent Neural Network (RNN) trained with precipitation, temperature, and soil properties as features and historical observed soybean and/or maize yield at municipality level for 1500+ cities in Brazil and USA as labels [18]. Another study evaluated the effectiveness of the Random Forest (RF) machine learning model for its crop yield prediction ability in comparison with Multiple Linear Regressions (MLR) serving as a benchmark [13]. RF was found to outperform MLR benchmarks in all performance statistics compared. Lately, other sophisticated methods like empirical analysis have also been applied to crop yield forecasting. Dharmaraja et al. described linear regression and time-series models to predict crop yield using Bajra yield data of Alwar district in Rajasthan, India [19].

Recently, various concepts of smart farms have also gained wide popularity as there is a lot of research being done in this field that exploits the prowess of Big Data, Cyber Physical Systems (CPS), Machine Learning (ML), blockchain, etc. to develop their application in agriculture [5]. Knowledge Graphs have also played an important role in this domain. By incorporating these technologies the farmers are maximizing their benefits as they can radically improve the crop yield prediction, detecting crop diseases,etc. Chukkapalli et al. [11], [20] have discussed how the developed ontologies support various AI applications in a single and co-op smart farm that benefits the farmers in various aspects such as labor, supply, marketing and distribution.

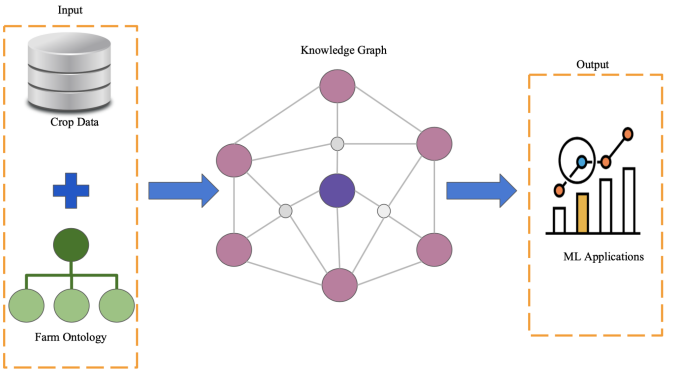

Fig. 1. System Architecture

III. Architecture

Estimating the crop yield ahead of time will be a great boon to farmers and the agriculture industry. As it enables proper planning and better decision making related to the crop management. Therefore, we develop a novel framework named *YieldPredict* that could be incorporated in a real-world agricultural environment for predicting the crop yield. Our YieldPredict framework contains three major parts shown in Figure 1. The first part integrates agricultural data with the existing smart farm ontology [11], [20]. In the second part, we populate a knowledge graph once the data is added to the ontology in the form of RDF triples. Further, we query the populated knowledge graph with the help of SPARQL [21] query language. The third part predicts the crop yield by utilizing machine learning algorithms where input is the data obtained after querying the knowledge graph. We will further describe various details and functionalities of our framework in the next few subsections.

A. Dataset Description

Our YieldPredict framework collects the data from multiple sources such as Open Government Data (OGD) Platform India and also the agricultural site of Rajasthan [9], [10]. The data acquired has entries for 33 districts of Rajasthan and is from the year 2007 to 2010. This data is divided into two data sets that represents Kharif crops and Rabi crops based on the rainfall pattern in the country. These patterns are adopted in many Asian countries where July to October is the Kharif cropping season and Rabi cropping season is during the winter season which is from October to March. The Kharif crops are Bajra, Jowar and Maize, and the Rabi crops are Barley and Wheat. Number of entries in Kharif crops data set are 335

State_Nan	District_Na	Crop_Year	Season	Crop	Area	Productior	Seasonal_	Nitrogen	Phosphoru	Potassium	SalineSoil	SodicSoil
Rajasthan	AJMER	2007	Rabi	Barley	7225	14843	0.1	L	M	H	16712	19830
Rajasthan	ALWAR	2007	Rabi	Barley	14246	41485	1.1	L	M	M	15976	97625
Rajasthan	BANSWAR	2007	Rabi	Barley	1257	2067	0.1	L	M	H	2131	2130
Rajasthan	BARAN	2007	Rabi	Barley	396	857	0	M	M	H	1008	1584

Fig. 2. Snapshot of Rabi crops data set.

and Rabi crops data set are 226 where the collected data set has a set of attributes monitored during the cropping season such as State Name, District Name, Crop Year, Season, Crop, Area, Production, Seasonal Rainfall, Nitrogen, Phosphorus, Potassium, Saline Soil (ha), Sodic Soil (ha). A snapshot of Rabi crop data set is shown in Figure 2.

We have chosen the above attributes since they play an important role in crop yield prediction. For example, area used in plantation of a specific crop is described by the Area attribute and production attribute gives the total produce of crop measured in tonnes. Seasonal Rainfall attribute provides the average amount of rainfall registered for Rabi and Kharif season measured in millimeters (mm). Attributes such as nitrogen, phosphorus and potassium are considered since they represent the three major soil nutrients. Saline soil is another important attribute which describes excessive levels of soluble salts in the soil water (soil solution). As high levels of saline soil can negatively affect plant growth, resulting in reduced crop yields and even plant death under severe conditions. Sodic soil attribute indicates the levels of exchangeable salts which impact the soil structure affecting the crop yield negatively.

B. Dataset Pre-processing

In our YieldPredict framework, we select attributes for both our data sets that potentially contribute towards crop yield prediction. Therefore, number of columns has been reduced to 8 from the existing 13 columns present in the initial data set. For example, 'State_Name' column is dropped since we have considered only one state, Rajasthan for our work. Similarly, column named 'Season' is also dropped as Rabi crops data set has data only for crops grown in Rabi season and Kharif crops data set has data of crops grown in Kharif season. Other columns that have been dropped include 'District_Name' and 'Crop_Year' which doesn't play any significant role in yield prediction. Simultaneously, we calculate the crop yield by dividing the production data per unit of harvested area obtained from the data present in columns such as 'Production' and 'Area'. The calculated crop yield data is added as a new column named 'Yield' and existing columns like 'Area' and 'Production' are dropped from the data sets. The 'Yield' column determines the amount of crop harvested per area of land measured in tonnes per hectare. Therefore, it plays a vital role in estimating the yield for a sector of agricultural land.

Data pre-processing techniques like removing the rows with null values and integer-encoding the categorical values in order to transform the input data to a suitable format for machine learning models have been applied. For example in the Rabi crops data set, the crop values for Barley and Wheat are integer encoded as 0 and 1, respectively. Similarly in the Kharif crops data set, the crop values for Bajra, Jowar and Maize are integer encoded as 0, 1 and 2, respectively. The other columns having categorical values present in both the data sets are 'Nitrogen', 'Phosphorus', and 'Sulphur'. The recorded values indicate their levels in the soil as very low (VL), low (L), medium (M) or high (H). For example, if a data point has value recorded as 'L' for phosphorus then it means the phosphorus level in soil is low. The input shape for Rabi crops data set is (226,7) and for Kharif crops data set is (335,7) where the target attribute is the 'Yield' column.

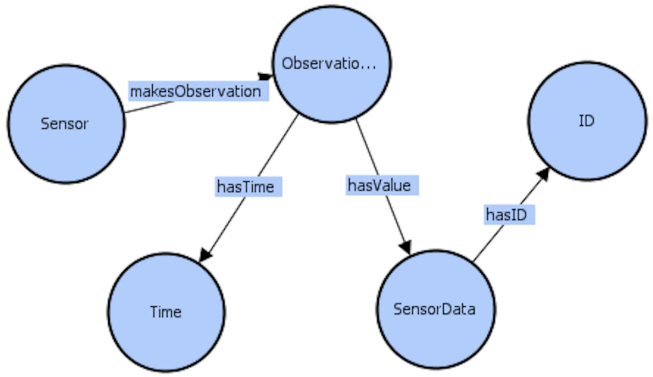

Fig. 3. Visual Notation for OWL Ontologies (VOWL) showing different classes from our ontology.

C. Integration of Smart Farm Ontology and Crop Data

In this work, we add the pre-processed data described in the above section by extending our existing smart farm ontology [11] since it supports information integration. In our smart farm ontology shown in Figure 3, we have class named Sensor where we add the attributes such as Crop, Nitrogen, Yield, Potassium, Saline Soil, Phosphorous, Seasonal_Rainfall, SodicSoil present in the Rabi and Kharif crop data set as instances of this class. Each attribute represents a physical sensor where the SensorData class represents the recorded values and Time class represents the time at which the data was recorded.

Data to the ontology is added in the form of Resource Description Framework (RDF) triples using RDF triple store which is a type of graph database. The linked data set determines the entities and their relationships. Simultaneously, Uniform Resource Identifier (URI) references for every relation and entities are created. The URI references are then added to the graph in the form of RDF triples along with their types. The RDF triples represent the relationship between two

entities, i.e subject, predicate, object where predicate gives the relation between subject and predicate.

We have also added an ID column to both Rabi and Kharif crop data sets, which represents the subject. Therefore, we created class named ID in our smart farm ontology. Each row has a unique identifier. The column values for a particular row are the data properties of the ID which are the objects in the triple store. The column fields Season, Crop_year, Crop,etc are stored as predicates which shows the relationship between the ID and the column values. For example, a single row from the data set as shown below can be represented in terms of rdf triples as follows: ID for the first row is 1, so 1 is added as an entity in the graph which is the subject in the RDF triple store. The values for all the column fields for the row with ID 1 are the objects. And, the column fields namely State_Name, District_Name, Crop_Year, Season, Crop, etc are the predicates which show the relationship between the ID and the column values. Below is the representation of one row in terms of subject, predicate, and object in triple store:

(1, has_State_Name, Rajasthan)
(1, has_District_Name, AJMER)
(1, has_Crop_Year, 2007)
(1, has_Season, Rabi)
(1, has_Crop, Barley)
(1, has_Area, 7225)
(1, has_Production, 14843)
(1, has_Seasonal_Rainfall, 0.1)
(1, has_Nitrogen, L)
(1, has_Phosphorus, M)
(1, has_Potassium, H)
(1, has_SalineSoil, 16712)
(1, has_SodicSoil, 19830).

D. Knowledge Graph Population

We populate a knowledge graph, once the crop data set is linked with the ontology described in the above section. A graphical representation of the generated knowledge graph is shown in Figure 4 where a single row of data present in the form of RDF triple is populated. In which '574' is represented as an instance of class named ID. The edge labels indicate the relation between the nodes. For example, ID '574' present in the data set provides us with information regarding the Maize crop grown during Kharif season, in the year 2010 had low concentration of nitrogen, phosphorous and medium concentration of potassium. The seasonal rainfall during that period was 794.4 mm where the value 5636 represents the salinity and the value 9403 represents the sodicity present in the soil. As a result, crop yield for that particular area was 1.36 tonnes per hectare.

In our next step, we query the populated knowledge graph utilizing a SPARQL interface. The goal for querying the knowledge graph in our work is to provide machine learning models with training data based on the necessary requirements that are important for yield prediction. When we query the knowledge graph using SPARQL, retrieved data is separated into two data sets namely Rabi crops data set and Kharif crops data set based on the growing season of the crops.

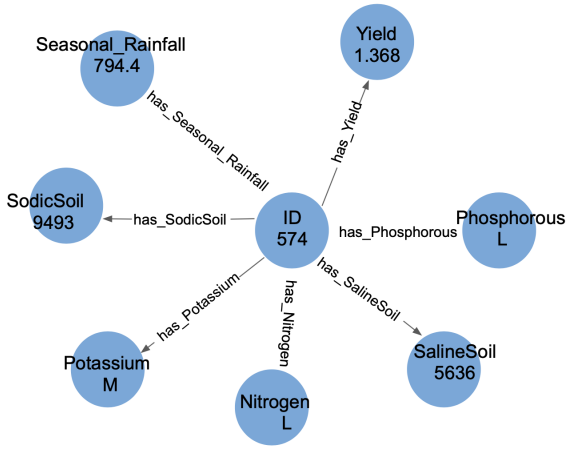

Fig. 4. Graphical representation of the Knowledge Graph populated from a single RDF triple

E. Implementation of Machine Learning Models

Several applications in different areas include machine learning techniques for predictions based on their past observations. In the agriculture sector, yield estimation of crops is said to be quite important as it provides insights for improving the agriculture statistics. Therefore, we implement ten machine learning algorithms on our Rabi and Kharif crops data sets in order to predict the crop yield for the state of Rajasthan.

1) K-Nearest Neighbors: K-nearest neighbors (KNN) [22] is a supervised machine learning algorithm used for both classification and regression problems. KNN classification algorithm is used to determine the membership of the class whereas KNN regression algorithm is used to calculate property value for the object. In our case, we use KNN regression to predict the crop yield by calculating the average of the values of k nearest neighbors. Input to the algorithm is the data obtained after querying the knowledge graph which is split randomly into training and test sets in 1:4 ratio. We train the model twice with two separate data sets, namely Rabi crops data set and Kharif crops data set. Information about wheat and barley are present in the Rabi crops data set while Kharif crops data set has details about Maize, Jowar and Bajra crops. The model performance is measured by calculating the Root Mean Square Error (RMSE) value and the R2 score.

We observe that RMSE is 0.46 and the R2 score is 0.59 when model is trained on Rabi crops data set. Figure 5 provides us with a visual representation of predicted yield and actual yield for Rabi crop data set. The best results are observed when value of k is 4, i.e 4 neighbors, and weights being uniform as all points in the neighborhood are weighted equally. In case of Kharif crops data set, RMSE is 0.56 and the R2 score for this model is 0.04. The R2 score for the Kharif crops data set is very low when compared to the above model using Rabi crops data set. This could be because Rabi crops data set has data for only two crops whereas Kharif crops data set had data for three crops which may be one of the reasons

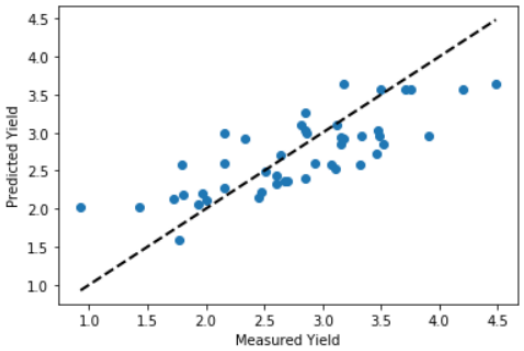

Fig. 5. Predicted Yield vs Measured Yield plot for KNN model for Rabi crops data set

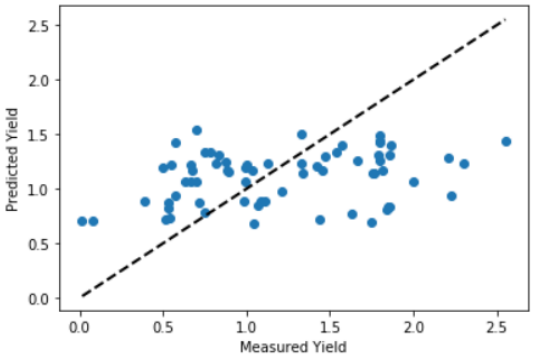

Fig. 6. Predicted Yield vs Measured Yield plot for KNN model for Kharif crops data set

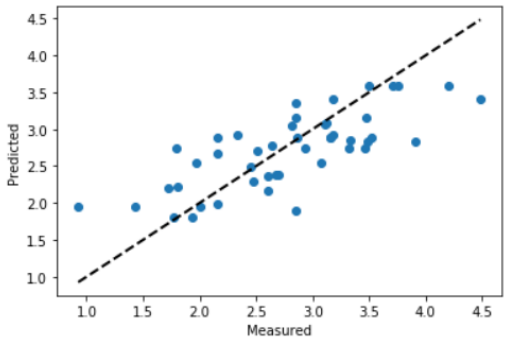

Fig. 7. Predicted Yield vs Measured Yield plot for Support Vector Regression model for Rabi crops data set

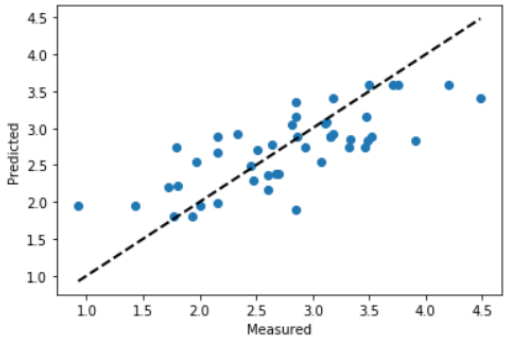

Fig. 8. Predicted Yield vs Measured Yield plot for Support Vector Regression model for Kharif crops data set

for the model not being a good fit on the data. Figure 6 shows a scattered plot for predicted yield vs measured yield in Kharif crop data set.

2) Support Vector Regressor: Support Vector Regression(SVR) [23] is another supervised machine learning algorithm used for regression. There are multiple kernels such as linear, polynomial and radial basis function kernel (RBF). We utilize RBF kernel in our SVR algorithm since it is faster and provides better accuracy. Therefore, we train the SVR algorithm on both Rabi and Kharif crops data set. Firstly, we utilize Rabi crop data set as input to machine learning algorithm. The model predicts the crop yield as output. Performance of the model is measured by calculating the Root Mean Square Error (RMSE) value and the R2 score where the RMSE is 0.547 and the R2 score for this model is 0.496. Then we train SVR algorithm on the Kharif crops data set where we observe an RMSE value of 0.678 and R2 score of -0.399. This model doesn't perform well for Kharif crops data set, as the fit is worse than the null hypothesis. This is not the case for Rabi crops data set. Figure 7 and Figure 8 show the plots for predicted yield versus measured yield for Support Vector Regression model trained using Rabi crops dataset and Kharif crops data set.

3) XGBoost Regressor: Extreme Gradient Boosting(XGBoost) [24], [25] is a decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework. Decision tree based algorithms are considered best in case of small-to-medium structured/tabular data. XGBoost regressor gives good results for both Rabi and Kharif crop data sets when compared to other machine learning algorithms used here. This model performs the best for Kharif crops data set with an R2 value of 0.572 and RMSE 0.37. The performance of the model for Rabi crops data set is better than most of the other machine learning models used in our work. The R2 value for the model trained and tested using Rabi crops data set is 0.538 and RMSE is 0.529. Figure 9 and Figure 10 show the plots for predicted Yield vs measured Yield for XGBoost Regressor model for both Rabi crops and Kharif crop data set.

4) Gradient Boosting Regressor : Boosting [24], [26] is a way of combining multiple simpler models to create one composite model. We combine multiple weak learners(generally decision trees) such that overall loss is minimized. This is why it is also known as an additive model. Trees are constructed in a greedy manner, choosing the best split points based on purity scores like Gini is done to minimize the loss. We have used the standard gradient boosting regressor which uses 'least squares regression' as a loss function with 100 boosting stages.

This model is the best performing model for Rabi crops data set. The model obtained by training Gradient boosting Regressor algorithm on Rabi crops data set has R2 value of

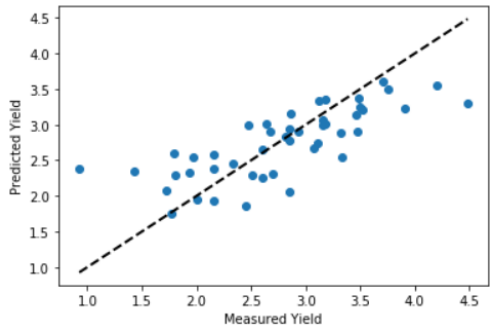

Fig. 9. Predicted Yield vs Measured Yield plot for XGBoost Regressor model for Rabi crops data set

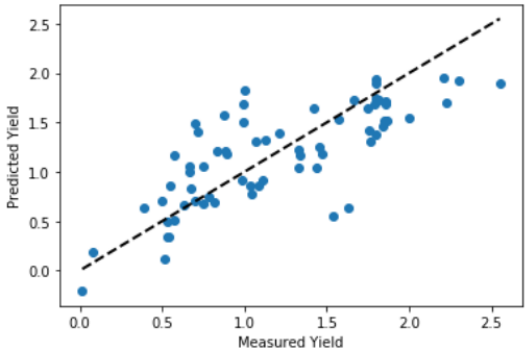

Fig. 10. Predicted Yield vs Measured Yield plot for XGBoost Regressor model for Kharif crops data set

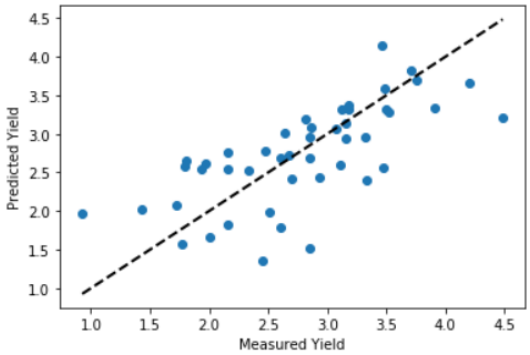

Fig. 11. Predicted Yield vs Measured Yield plot for Gradient Boosting Regressor model for Rabi crops data set

0.616 and RMSE value of 0.482. R2 value indicates that 61.6% of the observed variation in data can be explained by the model's inputs and is the highest among all the other models. The model is obtained from an ensemble of weak predictive models. The other model trained on Kharif crops data set using the same algorithm also performs well where R2 is 0.468 and RMSE is 0.400. The fit of the models can be seen by observing the Figure 11 and Figure 12, where the data points are close to the x=y line, indicating a good fit.

5) Light GBM: LightGBM [27] is a variant of tree based boosting algorithms with one of the key differences is growing

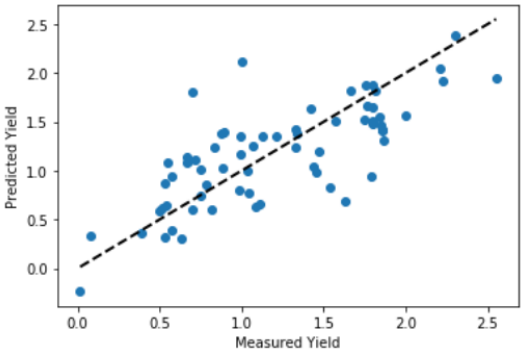

Fig. 12. Predicted Yield vs Measured Yield plot for Gradient Boosting Regressor model for Kharif crops data set

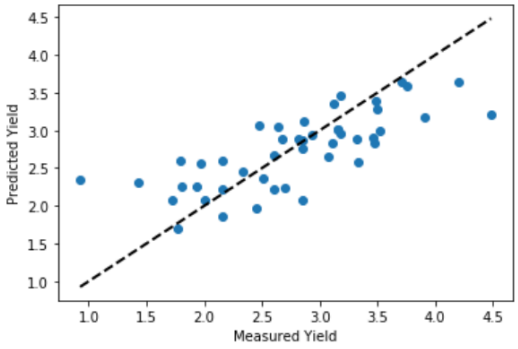

Fig. 13. Predicted Yield vs Measured Yield plot for LightGBM Regressor model for Rabi crops data set

its tree structure leaf wise rather than level wise. It uses Gradient-Based One-Side Sampling(GOSS) technique which inspects the most informative samples while skipping the less informative samples. It can train very quickly compared to XGBoost while maintaining comparable accuracy. However, it is very sensitive to overfitting if the data set is small.

The R2 value for the LightGBM model trained on Rabi-Crops dataset is 0.499 and the RMSE value is 0.551. Plot for predicted yield versus measured yield for Rabi crop data set is shown in Figure 13. The R2 value for the LightGBM model trained on Kharif-Crops dataset is 0.407 and RMSE value is 0.23. Figure 14 shows the plot for predicted yield versus measured yield for Kharif crop data set. As our data set is small in size making the model very sensitive to overfitting.

6) Random Forest Regressor: Random forest [28] is a supervised learning algorithm which uses ensemble learning method for classification and regression. Random forest is a bagging technique and not a boosting technique like XGBoost regressor. The trees in random forests are run in parallel. There is no interaction between these trees while building the trees. The output prediction is a mean prediction of the individual trees. The performance for the both models trained on Rabi crops data set and Kharif crops data set are fairly similar to other models like Light GBM, AdaBoost Regressor model which uses ensemble learning method. It is observed that most

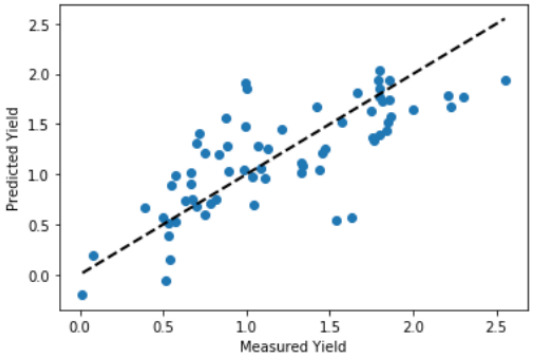

Fig. 14. Predicted Yield vs Measured Yield plot for LightGBM Regressor model for Kharif crops data set

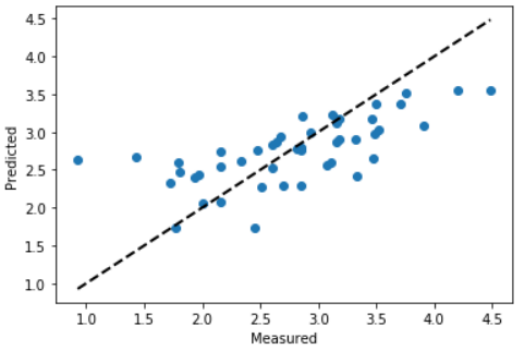

Fig. 15. Predicted Yield vs Measured Yield plot for Random Forest Regression Model trained using Rabi crops data set

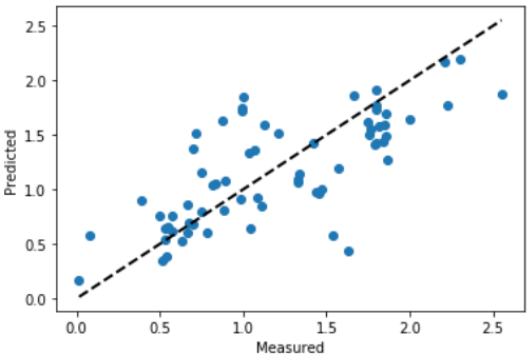

Fig. 16. Predicted Yield vs Measured Yield plot for Random Forest Regression Model trained using Kharif crops data set

ensemble learning models perform better than other individual algorithms like Decision Tree. The R2 value when trained on Rabi-Crops dataset is 0.501 with RMSE value of 0.550. While the R2 value when trained on Kharif-Crops dataset is 0.409 with RMSE value of 0.422. Figure 15 and Figure 16 show the plots for predicted yield versus measured yield for Random Forest Regression Model trained using Rabi crops dataset and kharif crops data set.

7) Decision Trees: Decision tree [29] algorithm is similar to the tree data structure where each node represents a question on the given data and based on each answer it branches further.

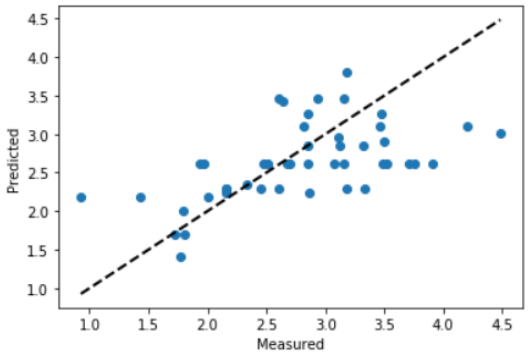

Fig. 17. Predicted Yield vs Measured Yield plot for Decision Tree Model trained using Rabi crops data set

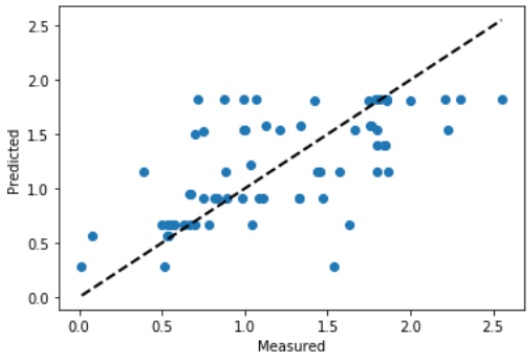

Fig. 18. Predicted Yield vs Measured Yield plot for Decision Tree Model trained using Kharif crops data set

The algorithm identifies the ways to split the data based on features and conditions. Each split section is measured for its effectiveness by calculating information gained(a measure of change in entropy after splitting). Another method is using the Gini index which tells us how many times a randomly selected data point will give incorrect prediction. Naturally we want to select attributes with lower Gini index. Decision tree model doesn't perform very well for both both our data sets. As R2 and RMSE value for Rabi crop data set are 0.07 and 0.75 respectively. Similarly, R2 and RMSE value for Kharif crop data set are 0.24 and 0.478. Plots comparing predicted yield and measured yield for Decision Tree Models trained using Rabi crops data set and Kharif crops data set are shown in Figure 17 and Figure 18.

8) AdaBoostRegressor : AdaBoost(Adaptive Boost) [30] is another algorithm that falls under the ensemble learning category. It first fits the regressor on the complete data then based on the error, additional copies are fit on a subset of data based on a certain condition known as a stump. Decision Tree Regressor is used as the base estimator since it is very fast to train and simple to execute. AdaBoost captures non linear relationships in data which something like a logistic regressor will miss resulting in better accuracy.

The AdaBoost Regressor models for both data sets perform better than the Decision Tree, Linear Regressor and Extra

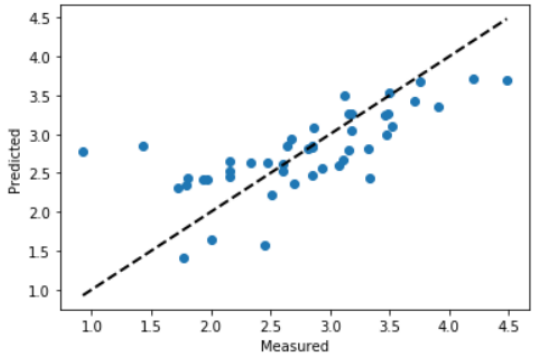

Fig. 19. Predicted Yield vs Measured Yield plot for AdaBoost Regressor trained using Rabi crops data set

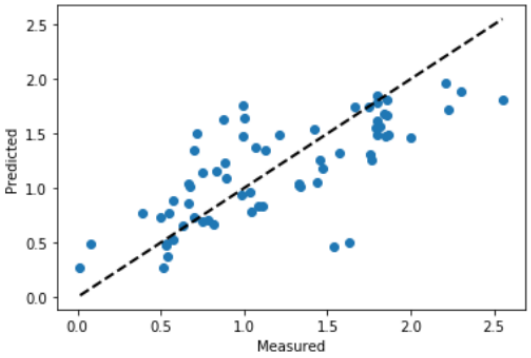

Fig. 20. Predicted Yield vs Measured Yield plot for AdaBoost Regressor trained using Kharif crops data set

Trees Regressor models. As R squared value for Rabi-Crops data set indicates that 48.4% of the observed variation in data can be explained by the model's inputs, the RMSE value for the same model is 0.559. The R2 value for the model trained and tested on Kharif crops data set is 0.450 and RMSE value is 0.407. The RMSE value for the second model is less than the model trained on Rabi crops data set. Predicted yield versus measured yield for AdaBoost Regressor trained using Rabi crops data set and kharif crop data set are shown in Figure 19 and Figure 20. Moreover, AdaBoost Regressor models perform better for both data sets than other models. Therefore, AdaBoost Regressor algorithm can be considered as one of the models to provide better insights regarding crop yield prediction.

9) Linear Regressor: Linear regression [23] is one of the simplest statistical algorithms. It tries to find a relationship between dependent and independent variables. A linear regression line is of the type Y = mx+c where m is the regression coefficient and c is a constant. In real-time scenarios, there are many more dependent variables present in the data sets and we try to find the coefficients such that we minimize the residual sum of squares between predicted and actual values. The drawback of this approach is that it does not work well with complex data sets which are bound to have non-linear relationships among the features. In our work,

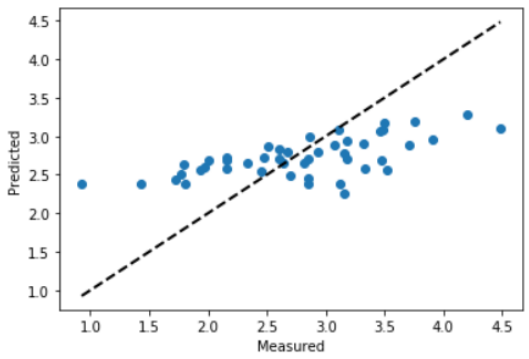

Fig. 21. Predicted Yield vs Measured Yield plot for Linear Regression Model trained using Rabi crops data set

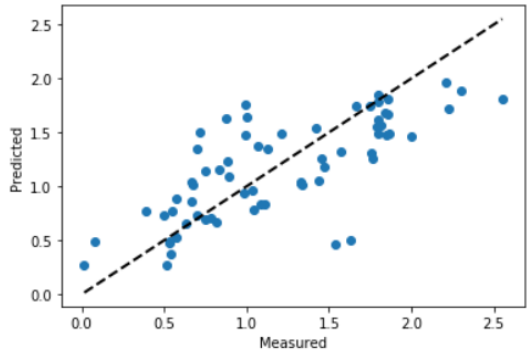

Fig. 22. Predicted Yield vs Measured Yield plot for Linear Regression Model trained using Kharif crops data set

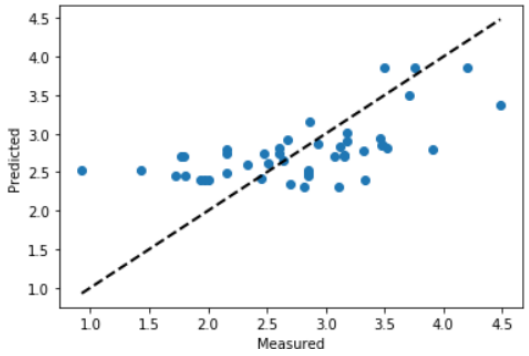

Fig. 23. Predicted Yield vs Measured Yield plot for Extra Trees Regressor Model trained using Rabi crops data set

linear regression doesn't give us a good fit for any of the data sets due to non-linearity of the data. The R2 value for the model trained and tested on Rabi crops data set is 0.217 and RMSE is 0.689. The model trained and tested on Kharif crops dat aset has R2 value of -0.002 and RMSE of 0.550. Plots comparing the predicted yield versus the measured yield for Rabi and Kharif crops data sets are shown in Figure 21 and Figure 22 respectively. The bad performance of model for both the data sets was expected since the data is non-linear and linear models cannot give a good fit for this data.

TABLE I
PERFORMANCE COMPARISON OF ALL THE MACHINE LEARNING MODELS FOR RABI CROPS AND KHARIF CROPS DATASET.

Model	Rabi-Crops Dataset		Kharif-Crops Dataset	
	R^2	*RMSE*	R^2	*RMSE*
K Nearest Neighbors Regressor	0.595	0.46	0.04	0.56
Support Vector Regressor (rbf kernel)	0.547	0.496	-0.399	0.678
XGBoost Regressor	0.538	0.529	0.572	0.37
Gradient Boosting Regressor	0.616	0.482	0.468	0.400
Light GBM	0.499	0.551	0.407	0.23
Random Forest Regressor	0.501	0.550	0.409	0.422
Decision Trees	0.072	0.750	0.244	0.478
AdaBoost Regressor	0.484	0.559	0.450	0.407
Linear Regressor	0.217	0.689	-0.002	0.550
Extra Trees Regressor	0.185	0.703	0.341	0.446

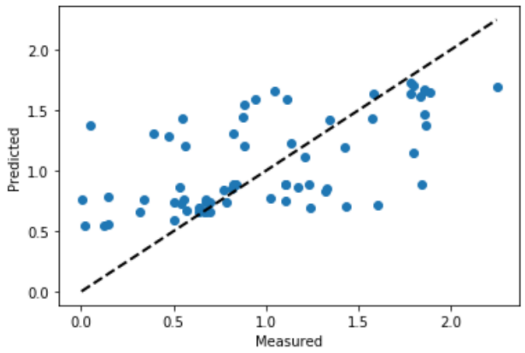

Fig. 24. Predicted Yield vs Measured Yield plot for Extra Trees Regressor Model trained using Kharif crops data set

10) Extra Trees Regressor: Extra trees regressor [31] is similar to random forest but it chooses a random value for the selection of split points unlike random forests that compute the locally optimal split condition. Other difference is that Extra tree uses the entire input sample whereas random forests uses bootstrap replicas. However, it is less computationally expensive while providing similar accuracy. This model doesn't perform very well for both the data sets. The model trained on Rabi crops data set has low R squared value which indicates that only 18% of observed variations of the data can be explained by the model inputs and only 34% of observed variations of the data can be explained by the model inputs for Kharif crops data set. The RMSE value for the model trained and tested on Rabi crops data set is 0.703 as this model is not a very good fit for the data and size of the data set being very small is another drawback. The RMSE value for the model trained on Kharif-Crops dataset is 0.446, which is lower when compared to the model trained on Rabi crops data set. Also, predicted yield versus measured yield plotted for both Rabi crop data set and Kharif crop data set are shown in Figure 23 and Figure 24.

IV. RESULTS

In this section, we evaluate the performance of machine learning models on both Rabi crop and Kharif crop data set as shown in TABLE I .We considered RMSE and R2 as metrics for evaluating the performance [32]. The Gradient Boosting Regressor model performed the best for Rabi crops data set with an R2 value of 0.616 which is the highest value when compared to other machine learning models. Also, RMSE value for this model is 0.482 which is quite low.

The best performing model for Kharif-Crops Dataset is XGBoost Regressor with an R squared value of 0.572 and RMSE 0.37. This is an ensemble algorithm which provides us with a composite model obtained using multiple weak models. XGBoost has two advantages over other models. First, it is faster than other implemented ensemble models and second one is due to the performance of the model. At present, we have a smaller data set but even if the size of the data set is increased, it can still perform faster compared to other models. Since it is an implementation of gradient boosting decision trees which is designed for increasing speed and performance. Subsequently, K-Nearest Neighbors Regressor is the second best model for Rabi crops data set, but that's not the case for Kharif crops data set. For the Kharif crops data set, the ensemble models performed better considering the inclusion of three different types of crop data present in this data set.

Another observation, we noticed from the results is that the models with Rabi crops data set as input performs well when compared to models with Kharif crops data set as input. As Kharif crops data set has data for three crops whereas Rabi crops data set has only two crops of data. Therefore, it was easier to get a good fit with only two crop data (Rabi crops data set) due to less variability in data when compared to Kharif crops data set. The same models might show better performance with an inclusion of larger data set.

V. CONCLUSION AND FUTURE WORK

Agriculture plays an important role in the economic development of a country. In India, rural population primarily depend on agriculture as their primary source of livelihood. In this paper, we describe our YieldPredict framework which provide farmers with insights that can further help them in decision making and agricultural planning to maximize their crop production. We link an existing smart farm ontology with the pre-processed crop data that consists of soil attributes, area, production and seasonal rainfall of the crop. The reason for choosing these attributes is that the data can be acquired easily when compared to remote sensing data, crop genotype

data, or vegetation index which is not very easy to obtain. We populate a knowledge graph from the linked data set. The knowledge graph can further be queried with the help of SPARQL to generate an input data set for the machine learning models. Finally, we have trained ten different machine learning algorithms to predict the crop yield and evaluated them by comparing their predictive accuracy. The best overall results were obtained from Gradient Boosting Regressor model on Rabi crops data set with R2 value of 0.616 and RMSE value of 0.482.

In the future, we plan to collect more data and also consider other environment variables that have an impact on crop yield prediction. Once a larger data set is available, neural networks based techniques can be used for predicting the crop yield.

REFERENCES

[1] UN projects world population to reach 8.5 billion by 2030, driven by growth in developing countries. https://www.un.org/sustainabledevelopment/blog/2015/07/un-projects-world-population-to-reach-8-5-billion-by-2030-driven-by-growth-in-developing-countries/#:~:text=The%20world's%20population%20is%20projected,around%2035%20years%20from%20now%2C. [Online].

[2] Frida Femling, Adam Olsson, and Fernando Alonso-Fernandez. Fruit and vegetable identification using machine learning for retail applications. In *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 9–15. IEEE, 2018.

[3] Bruno Miranda Henrique et al. Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, 124:226–251, 2019.

[4] M Amin and Amir Ali. Performance evaluation of supervised machine learning classifiers for predicting healthcare operational decisions. *Wavy AI Research Foundation: Lahore, Pakistan*, 2018.

[5] Maanak Gupta, Mahmoud Abdelsalam, Sajad Khorsandroo, and Sudip Mittal. Security and privacy in smart farming: Challenges and opportunities. *IEEE Access*, 8:34564–34584, 2020.

[6] Sina Sontowski, Maanak Gupta, Sai Sree Laya Chukkapalli, Mahmoud Abdelsalam, Sudip Mittal, Anupam Joshi, and Ravi Sandhu. Cyber attacks on smart farming infrastructure. *UMBC Student Collection*, 2020.

[7] Agriculture in Rajasthan, India. https://www.rajras.in/index.php/rajasthan/economy/agriculture/. [Online].

[8] Rajiv Singh et al. Note on the crop yield forecasting methods. *Asian Journal of Agricultural Research*, 13:1–5, 01 2019.

[9] District-wise, season-wise crop production statistics from 1997. https://data.gov.in/catalog/district-wise-season-wise-crop-production-statistics?filters%5Bfield_catalog_reference%5D=87631&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc. [Online].

[10] Agriculture Statistics. http://www.agriculture.rajasthan.gov.in/content/agriculture/en/Agriculture-Department-dep/agriculture-statistics.html. [Online].

[11] Sai Sree Laya Chukkapalli, Aritran Piplai, Sudip Mittal, Maanak Gupta, and Anupam Joshi. A smart-farming ontology for attribute based access control. In *6th IEEE International Conference on Big Data Security on Cloud (BigDataSecurity 2020)*, 2020.

[12] Ilianna Kollia, Birte Glimm, and Ian Horrocks. Sparql query answering over owl ontologies. In *Extended Semantic Web Conference*, pages 382–396. Springer, 2011.

[13] AK Prasad et al. Use of vegetation index and meteorological parameters for the prediction of crop yield in India. *International Journal of Remote Sensing*, 28(23):5207–5235, 2007.

[14] Jan Behmann, Anne-Katrin Mahlein, Till Rumpf, Christoph Römer, and Lutz Plümer. A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. *Precision Agriculture*, 16(3):239–260, 2015.

[15] Anat Goldstein, Lior Fink, Amit Meitin, Shiran Bohadana, Oscar Lutenberg, and Gilad Ravid. Applying machine learning on sensor data for irrigation recommendations: revealing the agronomist's tacit knowledge. *Precision agriculture*, 19(3):421–444, 2018.

[16] Saeed Khaki and Lizhi Wang. Crop yield prediction using deep neural networks. *Frontiers in plant science*, 10:621, 2019.

[17] Niketa Gandhi, Owaiz Petkar, and Leisa J Armstrong. Rice crop yield prediction using artificial neural networks. In *2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, pages 105–110. IEEE, 2016.

[18] Renato LF Cunha, Bruno Silva, and Marco AS Netto. A scalable machine learning system for pre-season agriculture yield forecast. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pages 423–430. IEEE, 2018.

[19] S Dharmaraja, Vidyottama Jain, Priyanka Anjoy, and Hukum Chandra. Empirical analysis for crop yield forecasting in india. *Agricultural Research*, 9(1):132–138, 2020.

[20] Sai Sree Laya Chukkapalli, Sudip Mittal, Maanak Gupta, Mahmoud Abdelsalam, Anupam Joshi, Ravi Sandhu, and Karuna Joshi. Ontologies and artificial intelligence systems for the cooperative smart farming ecosystem. *IEEE Access*, 8:164045–164064, 2020.

[21] Evren Sirin and Bijan Parsia. Sparql-dl: Sparql query for owl-dl. In *OWLED*, volume 258, 2007.

[22] Zhenyun Deng, Xiaoshu Zhu, Debo Cheng, Ming Zong, and Shichao Zhang. Efficient knn classification algorithm for big data. *Neurocomputing*, 195:143–148, 2016.

[23] S Athmaja, M Hanumanthappa, and Vasantha Kavitha. A survey of machine learning algorithms for big data analytics. In *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pages 1–4. IEEE, 2017.

[24] David Forsyth. Boosting. In *Applied Machine Learning*, pages 275–302. Springer, 2019.

[25] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, and Yuan Tang. Xgboost: extreme gradient boosting. *R package version 0.4-2*, pages 1–4, 2015.

[26] Peter Prettenhofer and Gilles Louppe. Gradient boosted regression trees in scikit-learn. 2014.

[27] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154, 2017.

[28] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[29] Soham Pathak, Indivar Mishra, and Aleena Swetapadma. An assessment of decision tree based classification and regression algorithms. In *2018 3rd International Conference on Inventive Computation Technologies (ICICT)*, pages 92–95. IEEE, 2018.

[30] Durga L Shrestha and Dimitri P Solomatine. Experiments with adaboost. rt, an improved boosting scheme for regression. *Neural computation*, 18(7):1678–1710, 2006.

[31] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[32] Nathalie Japkowicz and Mohak Shah. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.