Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

SICAIET
2024

Research Article
Collection: Artificial Intelligence and Emerging Technologies

Data Driven Yield Predictive Analytics of Major Crops in Punjab

Sharaiz Shahid[1], Abdul Razzaq[1*], Irfan Ahmad Baig[2], Zulqurnain Ali[1], Muhammad Aziz Ur Rehman[1], and Zaid Sarfraz[1]

[1]Department of Computer Science MNS University of Agriculture, Multan, Pakistan.
[2]Department of Agribusiness and Applied Economics MNS University of Agriculture, Multan, Pakistan.
[*]Corresponding Author: Abdul Razzaq. Email: abdul.razzaq@mnsuam.edu.pk

Academic Editor: Salman Qadri **Published:** February 01, 2024

Abstract: The economy of Pakistan is primarily dependent on the production of agricultural and agro-based products. Many factors including climatic conditions, soil fertility, topography and water quality reduce agricultural productivity that leads a threat to food security and financial loss to farmers. Many factors affect crop yield, including rainfall, temperature, humidity, soil, and PH. These factors pose significant risk to farms leading to yield reduction when they are not properly monitored and managed. To ensure professionalism, transparency, and integrity, Pakistani regulatory authorities like the Bureau of Statistics (BOS) and the Meteorological Department provide accurate, relevant, timely, and user-friendly data of crops. In this research, two different time series datasets were arranged in systematic order, consisting of crop production (Crop type, district, area, yield) and environmental factors (rainfall, temperature). The aim was to analyze the large collected of data of major crops in south Punjab (Wheat, Rice, Cotton) to identify patterns in different locations and extract main features that were used to predict the yield of crops. Farmers could use this approach to determine which crops are most suitable for their location on the basis of historical data. In the proposed study, data mining techniques were used for processing data, identifying patterns and extracting useful features to predict crop yields using these features.

Keywords: Crop yield; Datamining; Machine Learning; SVM; KNN.

1. Introduction

Providing food for the growing population is the key to human civilization, and farming is the only way to do so. Food supplies increased as human population and settlements increased, and farmland area decreased. In recent years, agricultural technologies have proven helpful for meeting increased needs and reducing crop losses. Fertilizers, weed control methods, and plant genetics are used to overcome the decreasing soil fertility. Analysis of the weather and soil has provided insights into the parameters affecting growth and yield, helping identify suitable crops for a particular location and climate [3].

In view of the importance of soil and farmland monitoring in terms of yield estimation and contributing towards economic development and food security, it is imperative to conduct soil and farmland monitoring.

The use of remote sensing has been proven to be an effective tool in agricultural monitoring, yield forecasting, and irrigation scheduling [12]. A number of these techniques are used to provide flood warnings and warnings about other disasters, as well as to monitor crop growth and to model disease spread. Depending on the crop production and yield prediction, the economics and food management of the country are significantly affected [12]. There is a great deal of interest among various stakeholders regarding the estimation of crop yield, including farmers, national governments, and international institutions, in order to increase food security [13]. In terms of productivity, the crop yield is defined as the weight of the harvest on an area divided by the number of units. Typically, small farmers focus their efforts on increasing yields at the farm scale in order to maximize the amount of profit they make. In contrast, governments, international organizations and commodity traders tend to focus more on producing estimates at higher scales, for example, at the administrative unit or national level. The crop insurance industry focuses on both the estimation of small-scale yields and the estimation of large-scale yields.

The crop yield estimation is beneficial for the farmers as they are getting to know the yield rate of crops well before cultivation so they can use their land effectively, calculate profits, avoid losses and use their expenditure on the crops efficiently [2].

With the advancement in machine learning and data science field we are able to use technology in agriculture to improve the productivity of crops. The Data mining is the process of retrieving required information from a large dataset. This includes the following stages such as Extracting, Transforming loading data in a repository and Data management [4]. The data collection is the first step of data mining in which the desired data is collected from different sources and arranged in systematic order.

In proposed study the dataset was collected from Pakistan the regulatory authorities like Crop reporting services (CRS), Pakistan Meteorological Department (PMD) providing accurate, relevant, timely and user-friendly data of crops to ensure the professionalism, transparency, and integrity. The wheat cotton and rice crop data were collected from CRS website and temperature; rainfall data was collected from PMD. After the collection of raw data from these sources some data preprocessing techniques were applied to convert the data into useful format.

The purpose of this research work is the development of a prediction model that can protect farmers against agricultural risks. Farmers recognize suitable traits combination such as temperature, availability of water, weather to find which perform better and how much yield can be expected on base of these values. The farmers can visualize the historic yields in the last decade on google map according to the districts in Punjab. Using machine learning, those attributes selected that will be more important and have positive effect on crop growth. Based on those selected attribute yield was estimated. yield was estimated on multiple features like Area Year Production Temperature in April may June July Aug Sep Rain fall in Apr may June July Aug and Sep. Crop managers may use forecasts to reduce losses when conditions are unfavorable. Furthermore, predictions can also be used to enhance crop forecasting whenever potential for favorable growing conditions exists. This model is proposed to help farmers in decision making. Farmers can use this model to yield estimation well before time. Predicting crop yields in both developing and developed countries is becoming increasingly important [5].

As part of artificial intelligence, machine learning is an important component. By using historical data as input, its algorithms can predict new outcomes without being explicitly programmed. These algorithms are able to predict new outcomes by utilizing historical data. The focus of this course is on creating systems that can learn or improve their performance based on the data they consume [18].

2. Related Work

Automatic yield monitoring is primarily accomplished by establishing multi-information comprehensive prediction systems or by using intelligent equipment. A review of research in the field of orchard yield prediction and estimation has been conducted over the past 12 years (from 2010 to 2021) [1]. According to the investigation, vegetation indexes and machine learning are the most used input features in yield prediction systems.

Cotton rice and wheat are the major crops with rich nutrients that have industrial and medical importance. However, it reports some industrial, marketing, and cultivation problems [6]. Therefore, the effective components need to be optimized to enhance the yield. Literature showed that machine learning has been utilized to optimize yield of different crops.

An artificial neural network was applied to an experiment on wheat fields. ANN is a process where different parameters of a crop are used as input and yield is the output. Regression was used to find significant positively affected or statistically correlated parameters with grain yield. An ANN optimizes the yield using this regressor as input. Optimization will result in a greater grain yield than that observed from cultivation. Predicted yield accuracy obtained was 72% [7].

Artificial neural networks and genetic algorithms were used in place of statistical methods to improve barley yield productivity. Regression was applied on certain traits as well as on yield components. Two field experiments were conducted with ten genotypes. SAS software used to find correlation of all regressors with yield. In regressions with direct effects, the yield coefficient was considered. In ANNs and GAs, participant regressors were used. Optimization increased the potential yield to 579kg/ha, that was higher than observed for the genotypes 3527 to 5163kg/ha [8].

For prediction of rice crop yield, ANN was applied using WEKA tool. Four-year dataset was collected from 27 district and various parameters of crop were measured. Using Weka, the data was accurately pre-processed for the application of a multilayer perceptron (MLP) neural network. ROC curves were used to visualize the accuracy of the predicted yield using the cross-validation method. The proposed system achieved an accuracy rate of 97.54% [9].

Rice production was predicted using ANN using AEZ model. In the analysis, electrical conductivity, organic matter percentage, and soil pH were found to be the main determinants of rice production potential. Models based on artificial neural networks have shown that they are effective techniques that are capable of monitoring rice fields and forecasting yields in selected fields. [10].

Using data mining techniques, a predictive model was developed to set cultivation plans to ensure high yields of paddy crops. The dataset was real time collected data with attributes of seed quality, land type, soil fertility and crop varieties from paddy crop along the Thamira-barani river. K-mean clustering and classifiers of decision tree were used for prediction. Performances of classifiers were evaluated through classification accuracy and root means squared error values. Result showed that classifier of random forest has high predictive power then other methods of classification, outperform with an accuracy rate of 97.5%, precision of 0.97, ROC of 0.99 and recall of 0.98. Weka was used for implemented of proposed predictive model [11].

The yield prediction of crop includes predicting yield of the crop from the available past data like climate weather, and historic crop [15].

Applying recurrent Q-Network deep learning to forecast paddy yield, the Q-Learning algorithm determines crop yield based on input parameters, and the RNN algorithm works with the output parameters. Q-values were mapped to RNN values via a linear layer, and the reinforcement learning agent incorporated a combination of parametric features with thresholds to predict crop yields. Finally, by minimizing error and maximizing forecast accuracy, the agent receives an aggregate score for the actions performed [31].

3. Proposed methodology

The methodology of proposed system includes the flowing steps and illustrated in Figure 1.

- Data collection
- Features extraction
- Model training & testing
- Model evaluation and prediction
- Clustering and classification
- Visualize the yield on google map district wise.

3.1 Data preprocessing

After obtaining all attributes from sources an important step of machine learning is performed to prepare it for model training. Data pre-processing is crucial to enhance the quality of the data and extract meaningful information from the data. It renders dataset to machine learning model by building and training. It was necessary to transform data into an understandable and readable form as quality of training data determined the quality of model. For practice, the program encounters data that has been cleaned, is not uniformly formatted, contains missing values, outliers, and features that span a wide range of values. The preprocessing involves data munging, data cleaning, wrangling, and removing the null values. The dataset contains weather data, climate data, area and yield through the time period of 2010 to 2020.

3.2 Machine learning models

The machine learning models are labeled as learning target function (f) that best maps input variable (X) to an output variable (Y).

$$\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{X}) \qquad \textbf{(1)}$$

This is a general learning task where the future prediction (y) is made on the base of input variable (X). in this study the machine learning models KNN and SVM models and for classification the prediction of yield is divided into three clusters (Good, Average and Low Yield) through K-mean clustering were used to make crop yield prediction in this the KNN performs better.

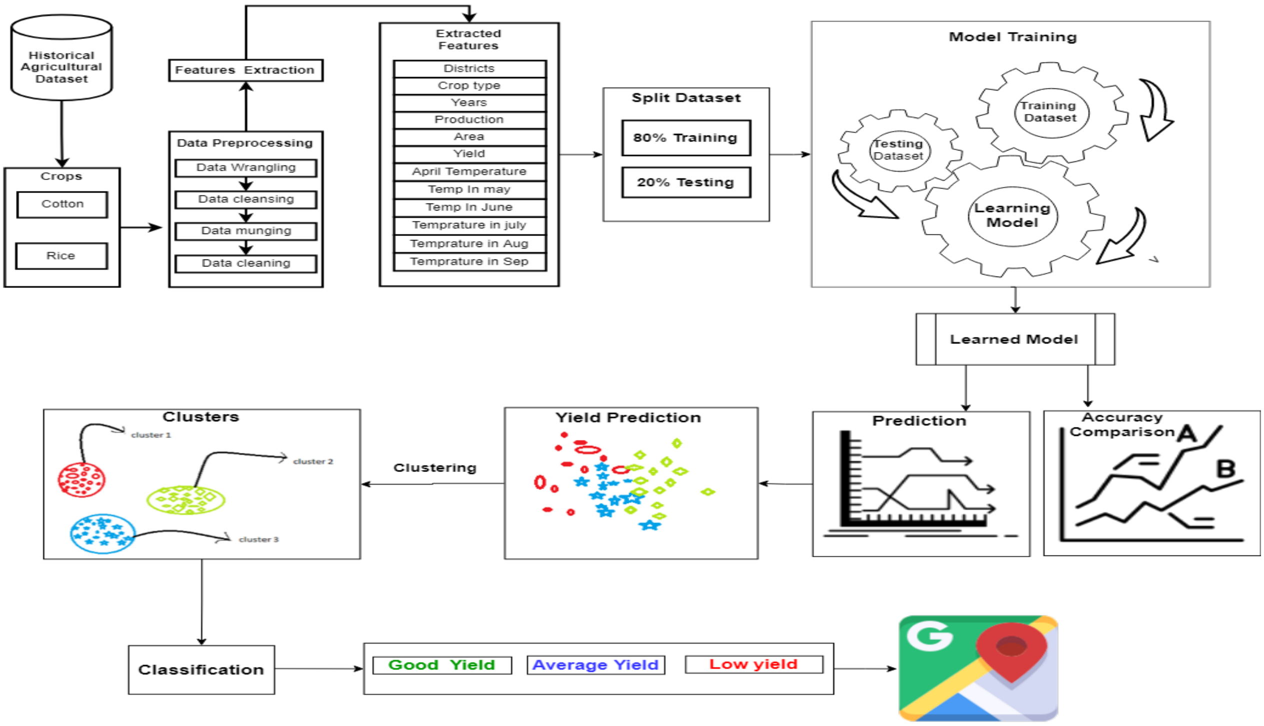


Figure 1. Proposed Methodology

3.3 K- Nearest Neighbor (KNN)

The parameters like Crop, District, Year, Area, Yield, Production, Temperature in April, Temperature in May, Temperature in June, Temperature in July, Temperature in August, Temperature in September, Rain Fall in April, Rainfall in May, Rainfall in June, Rainfall in July, Rainfall in August, Rainfall in September were considered as input variable to make yield prediction. The yield parameter is selected as target variables with continues values. The yield is predicted using the nearest known neighbor's values. This is possible by calculation Euclidian distance between those points. The KNN algorithm finds the nearest neighbor of unseen data point using K-value and assigns the class to unseen data point by having the class which has the highest number of data points out of all classes of K neighbors. The equation-1 is used to calculate the distance between data points.

$$\boldsymbol{Distance(a,b) = \sqrt{(b1 - a1)^2 + (b2 - a2)^2}} \qquad (2)$$

After distance calculation the highest probability of class is assign to the input value X.

$$\boldsymbol{P(y = j \mid X = x) = 1 \div K \sum_{i\text{€}A} I(y^{(i)} = j)} \qquad (3)$$

3.4 Support vector Machine (SVM)

The Support vector machine algorithm is mostly used for the classification approaches; however, it can be employed in regression problem. It can handle multiple continues and categorical variables like in our case the yield parameter. The SVM construct the hyperplane in multidimensional space to separate different classes. It generates the optimal hyperplane in an iterative manner that is used to minimize the error. In SVM the support vectors are the data points, which are contagious to the hyperplane. these support vectors points define the separating line by measuring the margins further these points are used to construct the classifier. The hyperplane is a decision plane which separates the objects. The SVM model is implanted using the kernel which transforms an input data space into the required form. The technique kernel trick is used by SVM in which the kernel takes a low dimensional input space and transforms it into a higher dimensional space. The SVM has three different kernels, linear kernels, polynomial kernel, and radial basis function kernel. In this study we use linear kernel which uses the normal dot product of any two given observations. The product between to vectors is the sum of the multiplication of each pair of input. The linear kernel formula is given in equation 3.

$$\boldsymbol{K(x\,, xi) = sum(x \times\ xi\ )} \qquad (3)$$

3.5 K-mean Clustering

The yield prediction by the models are the continues values for the upcoming year to make it clear we divide the yield in to three clusters (C1, C2, and C3). To perform the classification of crop yield into three classes (Good average, and low) to understand the yield pattern across the years in Punjab districts. The proposed study divides the clusters based on per hectares yield of crop according to the regional agricultural research institute Bahawalpur, Pakistan on average the rice yield in Punjab is 6.9 tons per hectare and for wheat 2.41 tones. The clusters were labelled according to the average yield outcome from the land the figure-2 show the clusters of predicted yield of models. After the clustering the classification of yield is performed through the different machine learning models Adaboost, Naïve Bayes, Neural Network, Random Forest, SVM, and KNN. Result indicates that kNN and Adaboost give accurate results with the accuracy of 98% and 96% respectively.

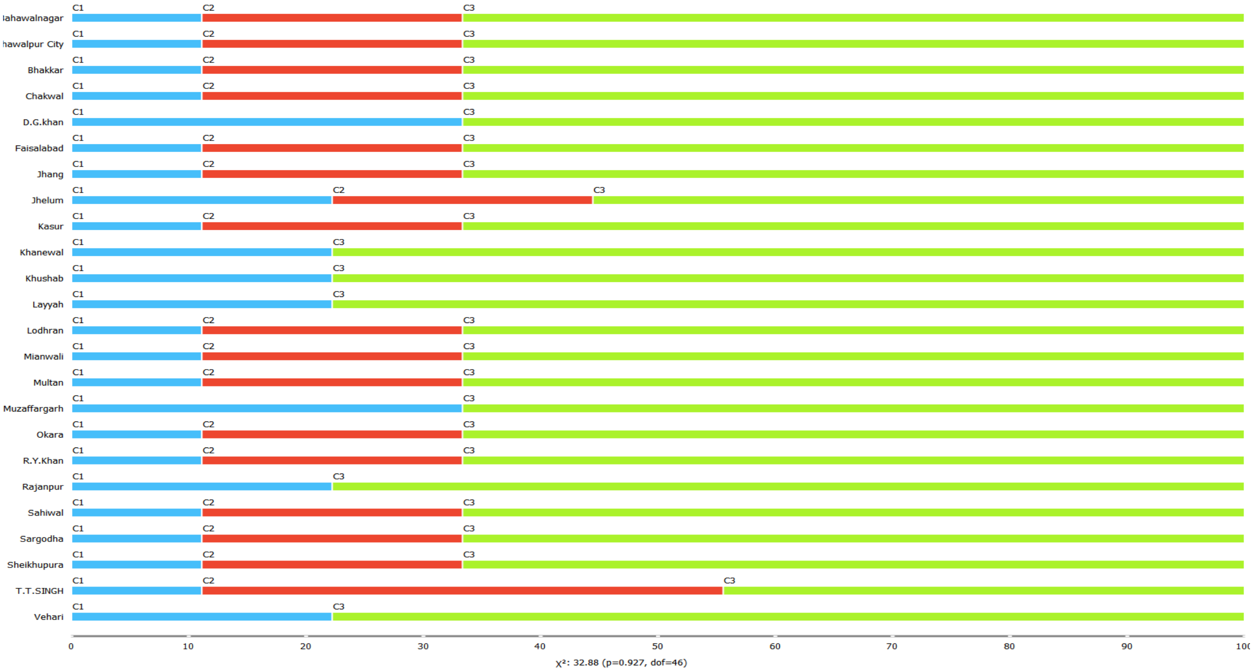


Figure 2. Yield Clusters using K-mean Clustering.

4. Results and Discussion

Yield prediction of any crop is always beneficial not only for farmers but also for exporters and crop users. Crop yield prediction is the estimation of a crop's yield based on historical data like weather and crop attributes [14]. Besides providing information about future yields, it also provides parameters whose values are important for yield and to whom farmer has to pay attention for maximum yield.

4.1 Predicted Yield based on Features

The features that have high inference with yield on the basis of which yield was predicted. To get accurate yield predictions, SVM, KNN machine learning algorithms are used. Based on these algorithms, a predicted yield is presented in figure 3 along with their accuracy. Several factors, such as temperature and rainfall, were taken into account for the cotton and rice crops in 25 districts in Punjab. Crop yield is strongly influenced by these features. Figure-3 and Figure-4 shows the actual and predicted yield of cotton and rice crops on validation datasets.

4.2 Yield Classification after Clustering

After k-mean clusters the 3 clusters are divided into three classes good average and low yield. After labeling the cluster in 3 classes we perform classification. The figure 4 presented the classification results. Result indicates that kNN and Adaboost give accurate result with the accuracy of 98% and 96% respectively.

4.3 Confusion Metrix of Classification models

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing. The Confusion Metrix of Adaboost, Naïve byes and random forest classifier are presented in figure 7, 8 and 9 respectively

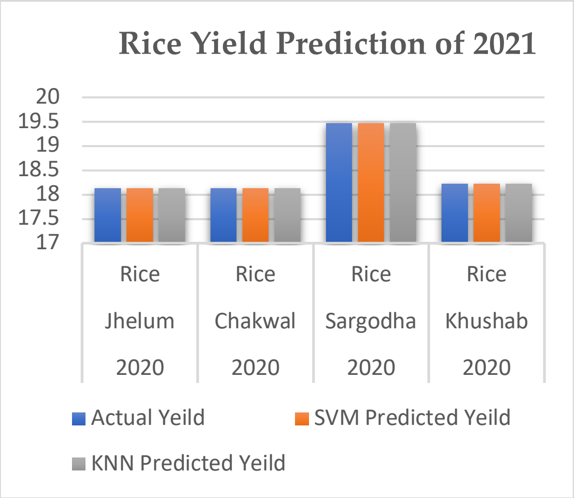


Figure 3a. Rice yield Prediction and Actual yield

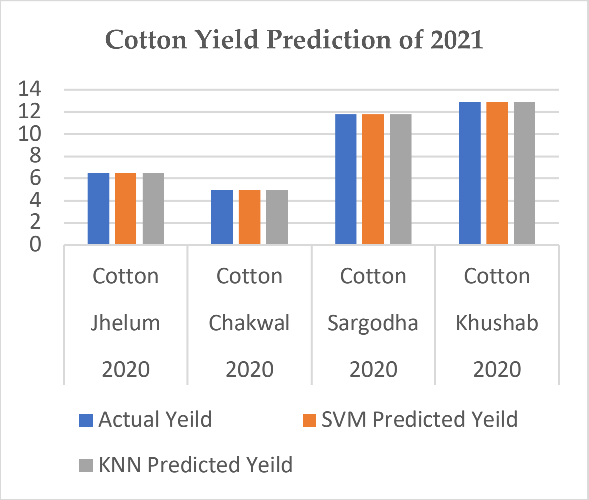


Figure 3b. Rice yield Prediction and Actual yield

Table 1. Confusion Metrix of Adaboost

		Predicted			
		Average	**Good**	**Low**	**∑**
Actual	Average	12	0	1`	13
	Good	0	58	0	58
	Low	0	0	15	15
	∑	11	60	15	86

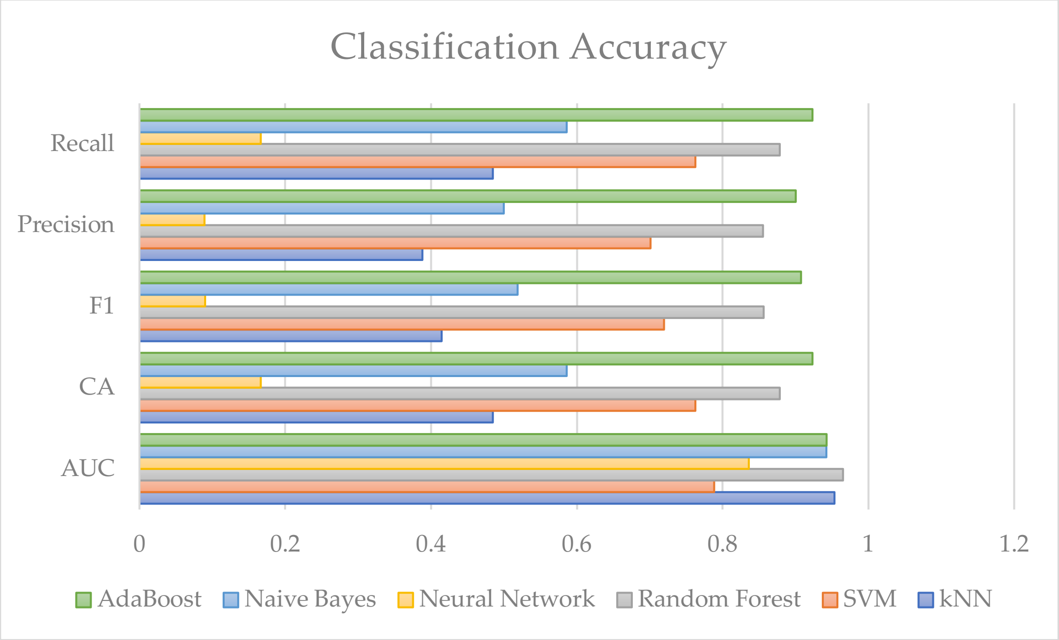


Figure 4. Classification Accuracy

4.4 Google map yield visualization

The yield visualization on the map is shown in figure 7. with simple interface the framers can visualize the historical yield district wise and can make smart move.

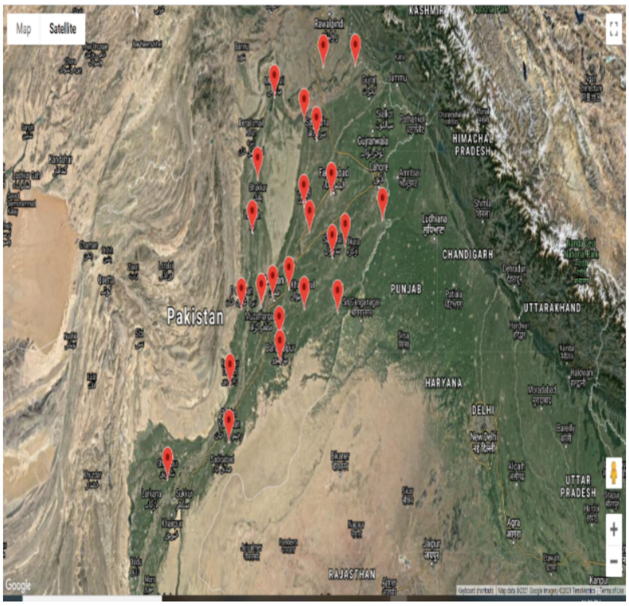
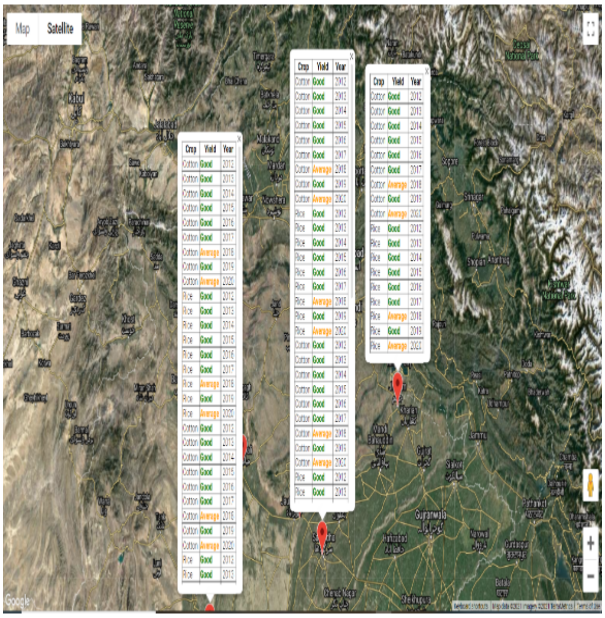


Figure 5a. Yield Punjab Districts



Crop	Yield	Year
Cotton	Good	2012
Cotton	Good	2013
Cotton	Good	2014
Cotton	Good	2015
Cotton	Good	2016
Cotton	Good	2017
Cotton	Average	2018
Cotton	Good	2019
Cotton	Average	2020
Rice	Good	2012
Rice	Good	2013
Rice	Good	2014
Rice	Good	2015
Rice	Good	2016
Rice	Good	2017
Rice	Average	2018
Rice	Good	2019
Rice	Average	2020
Cotton	Good	2012
Cotton	Good	2013
Cotton	Good	2014
Cotton	Good	2015
Cotton	Good	2016
Cotton	Good	2017
Cotton	Average	2018
Cotton	Good	2019
Cotton	Average	2020
Rice	Good	2012
Rice	Good	2013

Crop	Yield	Year
Cotton	Good	2012
Cotton	Good	2013
Cotton	Good	2014
Cotton	Good	2015
Cotton	Good	2016
Cotton	Good	2017
Cotton	Average	2018
Cotton	Good	2019
Cotton	Average	2020
Rice	Good	2012
Rice	Good	2013
Rice	Good	2014
Rice	Good	2015
Rice	Good	2016
Rice	Good	2017
Rice	Average	2018
Rice	Good	2019
Rice	Average	2020
Cotton	Good	2012
Cotton	Good	2013
Cotton	Good	2014
Cotton	Good	2015
Cotton	Good	2016
Cotton	Good	2017
Cotton	Average	2018
Cotton	Good	2019
Cotton	Average	2020
Rice	Good	2012
Rice	Good	2013

Crop	Yield	Year
Cotton	Good	2012
Cotton	Good	2013
Cotton	Good	2014
Cotton	Good	2015
Cotton	Good	2016
Cotton	Good	2017
Cotton	Average	2018
Cotton	Good	2019
Cotton	Average	2020
Rice	Good	2012
Rice	Good	2013
Rice	Good	2014
Rice	Good	2015
Rice	Good	2016
Rice	Good	2017
Rice	Average	2018
Rice	Good	2019
Rice	Average	2020

Figure 5b. Yield Different Districts

5. Conclusion

Using AI algorithms, this model is constructed to help farmers prevent losses in their farms due to a lack of knowledge of how to cultivate in different soil and weather conditions. The model is created by using machine learning (SVM) and KNN techniques. Analyzing the prediction parameters, the model predicts the best crops for growing on land with the least expenses out of a number of crops available. To the best of our knowledge, no work exists that uses the same methods to predict crops. Hence, it is concluded that there is an enhancement in the accuracy of this research work when compared to the existing work that used another technique for prediction of crops. The accuracy is calculated as 97%. It has a vast extension in future and can be actualized and interfaced with a flexible and multi-skilled application. The farmers need to be educated and hence, will get a clear information regarding best crop yield on their mobiles. With this, even if the rancher is at home, the work can be managed at that particular instant of time, without facing any kind of loss ahead. The progress in the agribusiness field will be extremely appreciable which will further result in helping the farmers in production of crops.

References

1. Leilei He, Wentai Fang, Guanao Zhao, Zhenchao Wu, Longsheng Fu, Rui Li, Yaqoob Majeed, Jaspreet Dhupia, Fruit yield prediction and estimation in orchards: A state-of-the-art comprehensive review for both direct and indirect methods,Computers and Electronics in Agriculture,Volume,195,2022,106812,ISSN-0168-699,https://doi.org/10.1016/j.compag.2022.106812.
2. Ratnmala Nivrutti Bhimanpallewar And Manda Rama Narasingarao, "Alternative Approaches Of Machine Learning For Agriculture Advisory System", 10th International Conference On Cloud Computing , Data Science & Engineering (Confluence) IEEE 2020
3. Nisar Ahmed1, H. M. Shahzad Asif2 , Gulshan Saleem3, M. Usman Younus4 Development of Crop Yield Estimation Model using Soil and Environmental Parameters
4. Aruvansh Nigam, Saksham Garg, Archit Agrawal and Parul Agrawal "Crop Yield Prediction Using Machine Learning Algorithms", 2019 Fifth International Conference on Image Information Processing (ICIIP)
5. B. Basso and L. Liu. 2019. Seasonal crop yield forecast: Methods, applications, and accuracies. Adv. Agron, 154: 201-255.
6. Bewal, S., S. K. Sharma, A. Parida, S. Shivam, S. R. Rao and A. Kumar. 2009. Utilization of RAPD marker to analyze natural genetic variation in Calligonum polygonoides L. A key stone species of Thar desert. Int. J. Integr. Biol. 5: 148-151.
7. Arjmand, H. S., M. Ghorbanpour, S. Sharafi and G. Hussein. 2014. Optimization of wheat grain yield by artificial neural network. Intl. J. Farm. Alli. Sci. 3: 806-810.
8. Gholipoor, M., A. Rohani and S. Torani. 2013. Optimization of traits to increasing barley grain yield using an artificial neural network. Int. J. Plant. Prod. 7: 1-18.
9. Gholipoor, M., A. Rohani and S. Torani. 2013. Optimization of traits to increasing barley grain yield using an artificial neural network. Int. J. Plant. Prod. 7: 1-18.
10. Mahabadi, Y. N. 2015. Use of the Intelligent Models to Predict the Rice Potential Production. Int. J. Innov. Res. 2: 20-31.
11. Arumugam, A. 2017. A predictive modeling approach for improving paddy crop productivity using data mining techniques. Turkish J. Electr. Eng. Comput. Sci. 25: 4777-4787.
12. Weiss, M., F. Jacob and G. Duveiller (2020). "Remote sensing for agricultural applications: A meta-review." Remote Sensing of Environment 236: 111402.
13. Hayes, M. and W. Decker (1996). "Using NOAA AVHRR data to estimate maize production in the United States Corn Belt." Remote Sensing 17(16): 3189-3200.
14. Filippi, P., E. J. Jones, N. S. Wimalathunge, P. D. Somarathna, L. E. Pozza, S. U. Ugbaje, T. G. Jephcott, S. E. Paterson, B. M. Whelan and T. F. Bishop (2019). "An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning." Precision Agriculture: 1-15.
15. Priya, P., U. Muthaiah and M. Balamurugan. 2018. Predicting yield of the crop using machine learning algorithm. Int. J. Eng. Sci. Res. Technol. 7: 1-7.
16. Whetton, R. L., T. W. Waine and A. M. Mouazen. 2017. Optimising configuration of a hyperspectral imager for on-line field measurement of wheat canopy. Biosyst. Eng. 155: 84-95.
17. You, L., M. W. Rosegrant, S. Wood and D. Sun. 2009. Impact of Growing Season Temperature on Wheat Productivity in China. Agric. For. Meteorol. 149: 1009-1014.
18. Elavarasan, D., and P. M. D. Vincent. 2020. Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. IEEE Access 8:86886–901. doi:10.1109/ ACCESS.2020.2992480
19. S.C. Agu, F.U. Onu, U.K. Ezemagu, D. Oden,Predicting gross domestic product to macroeconomic indicators, Intelligent Systems with Applications,Volume 14,2022,200082,ISSN 2667-3053,https://doi.org/10.1016/j.iswa.2022.200082.