Machine Learning Approach for the Wheat Production Prediction

[1]Manjot Kaur, [2]Er. Amandeep Kaur Sohal, [3]Er. Amanpreet Singh Brar
[1]Student, [2]Assistant Professor, [3]Associate Professor
[123]Department of CSE, Guru Nanak Dev Engineering College, Ludhiana, India

Abstract: The prediction analysis can be done with three steps (a) pre-process (b) feature extraction and (c) classification. The predication analysis models are designed according to application type such as crop production analysis is one of the applications of prediction analysis. The model is trained using the sample data that includes known attributes which are High level production, Medium level production and Low level production. New data is analyzed and its behavior is determined using the trained model. In this paper, the KNN classifier is applied for the crop prediction in India. To improve accuracy of the existing algorithm the KNN classifier is replaced with the naïve bayes classifier for the wheat production prediction. The proposed and existing work is implemented in python and compared with each state of art. Simulation results show that proposed work improves the accuracy with reduction in execution time.

Index Terms **- KNN, Naives Bayes, Wheat production**

I. INTRODUCTION

The data mining is the approach which can extract useful information based on the current data. From the huge amount of collected data, extracting only the necessary information is important. Huge amount of data is available today in every field. It is very difficult as well as time consuming to analyze such large amount of data [1]. Raw form of data is not of any use and so extracting the necessary information from it is important due to which data mining is designed.A suitable crop should be selected for planting to maximize the production of crop. The selection of suitable crop is extremely important. Different aspects such as soil type, its composition, weather, region topography, crop produce, market prices and so on play an important role in crop selection [2]. Different algorithms such as artificial neural networks, K-nearest neighbors and Decision Trees have fixed a place for themselves with respect to the selection of crop and it depends on several factors [3].

In machine learning, the effect of natural disasters such as food shortage influences the selection of crop. The researchers have depicted the utilization of artificial neural networks for electing crops on the basis of soil and type of weather. In this study, a plant nutrient management model has been proposed for fulfilling soil requirements, to maintain its productiveness and thereby improves the production of crop [4]. This model is based on machine learning techniques. A crop selection method called CSM has been presented in this study. This method assists in the selection of crop on the basis of its produce forecasting and other dynamics. Since the ancient times, agriculture is known to be an important culture and source of income for humans. In earlier times, the humans used to cultivate crops in their own lands. Fulfilling the requirements of their own was their prior objective. So, ever since, the cultivation is being followed. This culture is being adopted by all the living beings [5]. The animals, birds and humans are directly dependent on the natural crops being cultivated. A very health and welfare life is led on by the individuals that survive on greenish products being grown on lands. Agricultural field is degrading slowly due to the development of new innovative technologies. So, today the people trend to generate hybrid products that include artificial products as well. This is very unhealthy for individuals.

Today, the awareness related to cultivating the crops at appropriate time and place is less in the modern people. There is also a change in the seasonal climatic conditions against the fundament assets such as air, soil and water based on the cultivating techniques. Thus, the food is very insecure due to these changes. Evident solutions and technologies for handling the problems have not been designed yet even with the rise in such problems [6]. Increasing the economical growth in agriculture can be possible in India through different measures. For improving and increasing the crop yield and crop quality, several methods have been used. Also, the crop yield production can be predicted through data mining.
The process through which data can be analyzed from various perspectives and summarized into important information is called data mining [7]. For analyzing the data from various angles and summarizing the recognized relationships, this is considered as an analytical tool. It helps in identifying the patterns among huge amount of data being generated in large relational databases of different applications. Information can be provided here using the associations and patterns. The knowledge related to historical patterns and future trends can be generated by converting the information [8].

For instance, the farmers can recognize the loss of crops and future problems can be prevented from occurring with the help of summary information related to crop production from previous years. An important agricultural issue that is being researched today is crop yield prediction. Amount of yield one can get as per his expectations is a general query that any farmer would like to predict [9]. The previous experience of a farmer on one specific crop was used to analyze the yield prediction previously. The plan of harvest operation, pests and weather conditions are some of the factors that affect the agricultural yield. To make decisions that are related to the agricultural risk management, it is important to have appropriate information related to crop yielding history. Evolution of a prediction model is the aim of this research through which the production of crop yield can be predicted [10].

II. LITERATURE SURVEY

Sahu, et al. (2017) presented a study related to the utilization of big data scheme in farming. Farming was the major source of human living. In faming, the scrutiny of crop data was considered an extremely imperative dynamic. In this study, with the help of big data theory, the accurateness of farming knowledge was used to identify the experiences of farmers [11]. Therefore, a structure was borrowed to provide huge computational challenges in crop analysis through the efficient gathering of valuable data. From the distant applications, this information was collected to do crop scrutiny. In this study, Hadoop structure was utilized for the storage of this large volume of farming data. For determining the sorts of crops to be sowed as per the soil content, an improved forecasting approach was developed for farmers. With the help of this approach, the productivity could be increased. In this structure, the random forest algorithm was combined with MapReduce programming system.

Yan, et al. (2017) proposed a novel scalable and private continual and private continual geo-distance assessment system known as SPRIDE. This system was proposed to provide geographic based services. The proposed approach computed remoteness amid sensors and fields in a confidential as well as incessant way. Without discovering any other information regarding locations, the distance among servers was concluded [12]. The evaluation of competent remoteness upon the encoded locations across a sphere by using a homomorphic cryptosystem was the major objective of SPRIDE. New and realistic modifications dependent on data segmentation and distance forecasting methodologies were presented for the scaling of big user base. This was done to minimize the cost of information sharing and calculation. A real-time private distance assessment was attained on the big network of fields because of the implementation of SPRIDE. The proposed approach showed seventeen times better improvement in runtime performance in comparison with earlier approaches according to simulation outcomes.

Ponce-Guevara, et al. (2017) presented a study related to some most imperative factors of farming. These factors included humidity, soil dampness, carbon dioxide and intensity level. These factors influenced the photosynthesis process of plants which affected the production of crop in a greenhouse [13]. For decision making within financial and business applications, these two fields provided huge assistance. In presence of massive data, the pattern identification was the prime focus in this approach. There was no particular control of data analytics through a standard with the help of these technologies and methods. Nevertheless, this study provided a set of algorithms for the generation of descriptive models and a set of data for information classification and forecasting.

Luminto, et al. (2017) proposed a novel approach for predicting the farming time of rice crop. This approach was identified as multiple linear regression models. This model provided uppermost exchange rate of cultivator at two season areas for the year 2016-2017. The important variables utilized in this approach were standard temperature and solar energy [14]. This model just utilized these two variables but these two variables were not sufficient for prediction. Through the testing of all variable grouping that caused less RMSE values, the forecasting could be done in some particular areas. The issue related to multiple dependent variables could be predicted using multiple linear regression technique. The implementation of this approach was extremely simple. In comparison with other machine learning algorithms, this approach provided high speed outcomes.

Yolanda. M. et al. (2017) presented a study related to the estimation of maize crop production using remote sensing and practical models. Therefore, as per the experiential true field values, improved accuracy was estimated. The LAI borrowed forecasting system overvalued production by 14%. On the other hand, 97% of accuracy was achieved by NDVI model [15]. The disparity found within the field data gathering that occurred at various intervals during day time was the major reason behind the behavior of LAI based model. The angle of incidence of sun light on the plant top was directly affected by it. This affected the foliar response estimated by the tool. For estimation of crop and amount of corn being produced, this method was utilized in various areas e.g. State of Mexico. For the implementation of grain import strategies in relevance to domestic requirement, the government officials utilized these measurements.

III. RESEARCH METHODOLOGY

Following are the various research gaps of this study

1. The prediction approach is based on the predicted future possibilities based on the current information. The features are extracted for the generation of final predictions. The feature extraction approach is required which predict features accurately.

2.The technique which are proposed in the previous study for the prediction are based on the classification. The classification methods which are used in previous methods give low accuracy which is to be improved.

The data is classified into several classes using KNN classifier. The arithmetic mean of complete dataset is used to calculate the centered points using k-mean clustering algorithm. Thus, the accuracy of prediction analysis is reduced here. Establishing the relationship among attributes of dataset is not easy due to their huge complexity. For classifying the wheat production among several numbers of classes, this research uses the KNN classifier. If the accuracy of classification needs to be improved, it is possible to replace the KNN classifier with some other classifiers.

The different phases of this proposed research are explained below:

Phase1: Pre-Processing:- To load the dataset that is extracted from UCI repository, pre-processing phase is performed which is the initial step of this research. Any kinds of missing values are removed such that the input data can be cleaned.

Phase2: Feature Extraction:- For establishing relation among each of the attribute of data and the target set, the feature extraction method is applied in the second step. Recognizing the valuable attributes is easy with the help of feature extraction.

Phase3: Classification:- Naïve Bayes classifier is applied in the final step. Here, the output of prediction will be generated through the classification results. The popularity of Naïve Bayes is growing since it is a subset of the Bayesian decision theory. Due to the need of minimal storage and high speed training process, this algorithm is applied in highly critical applications. Generating a rule that can allocate the future objects from a set of objects to a class is the major objective of this algorithm. It is very common to find a supervised classification problem. The rules are constructed using various methods. Even the datasets of huge size can use this algorithm since there is no need to include any complex repetitive parameter estimating method. Mainly due to its easy interpretation, even the unskilled users can use this classifier.

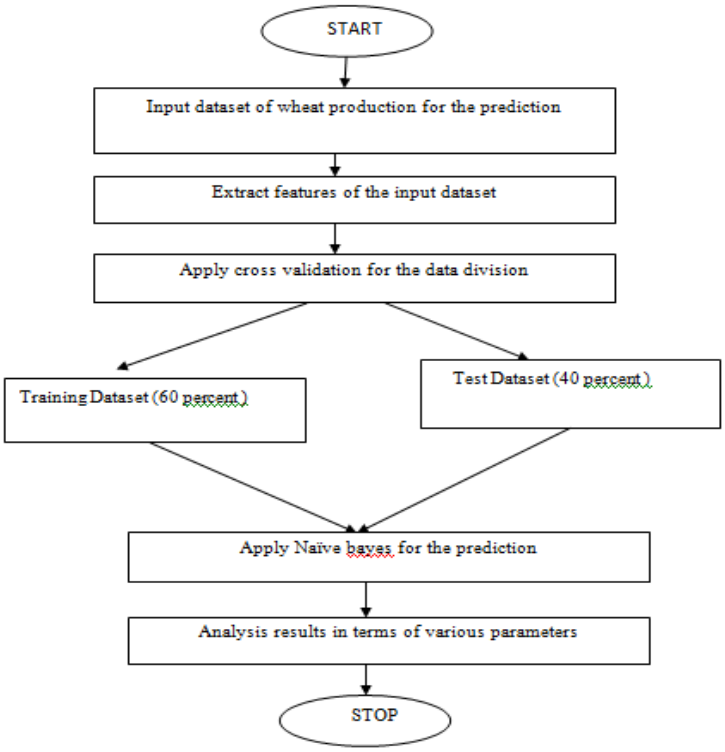

Fig 1: Proposed Flowchart

IV. RESULTS AND DISCUSSIONS

In order to get productive results first the data is used to execute the KNN code so as to predict wheat yield in India. The data is then applied as input and is used for prediction analysis. The data then extracts the features of the input and divide the data into training and test set. The Naïve Bayes classifier is used for the wheat prediction. The results are implemented in python and analysed in terms of various parameters which are accuracy, f measure, precision-recall and execution time for performance analysis.

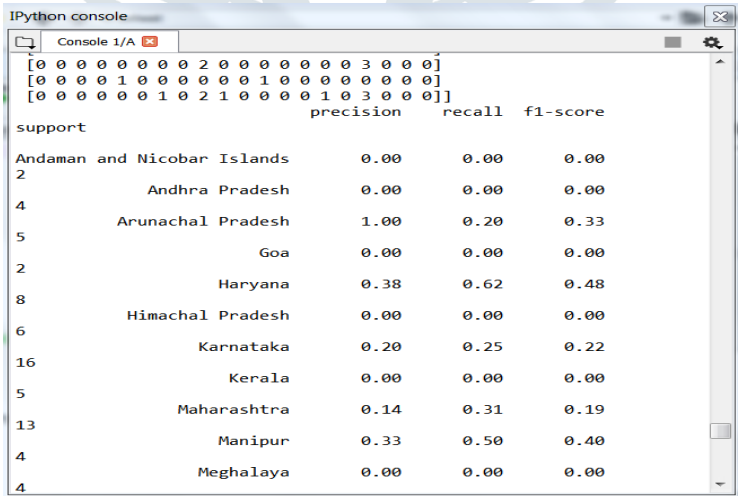

Fig 2: Calculation of Confusion Value

The Figure 2 depicts the implementation of Naïve bayse classifier. This classifier can divide entire dataset into training set and test set. The confusion matrix is implemented for the classification of state wise data.

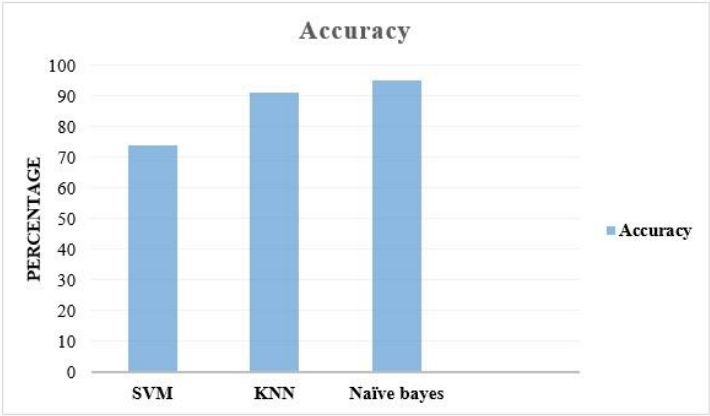

Fig 3: Accuracy Comparison

The Figure 3 depicts that the accuracy of the three classifiers such as SVM, KNN and naïve bayes are compared for the forecasting of wheat yield. The naïve bayes classifier shows maximal accuracy level in comparison with other classifiers as per analysis.

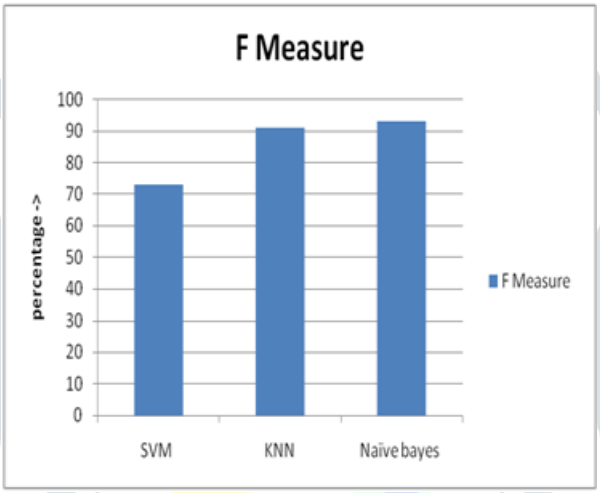

Fig 4: F Measure Comparison

The Figure 4 depicts that the F measure of three classifiers such as SVM, KNN and naïve bayes is compared for the forecasting of wheat yield. The naïve bayes classifier has maximum f measure as compared to other classifiers

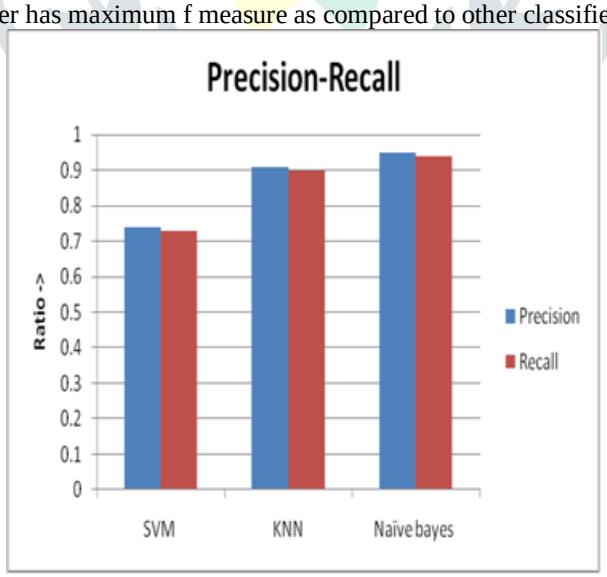

Fig 5: Precision-Recall Comparison

The Figure 5 depicts that the precision-recall value of three classifiers such as SVM, KNN and naïve bayes is compared for the analysis of performance. The naïve bayes classifier provides high precision-recall value in comparison with other classifiers as per analysis.

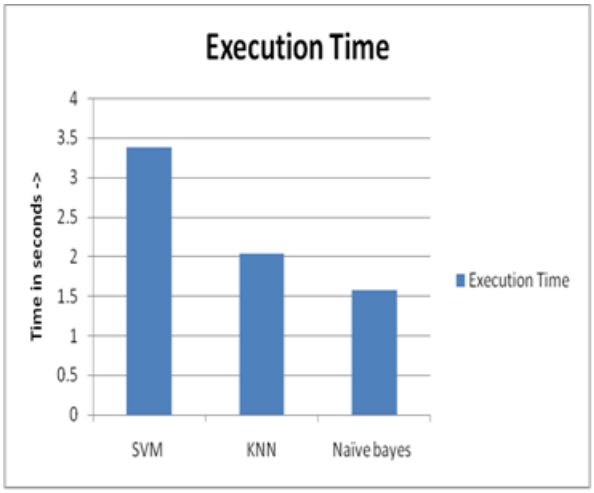

Fig 6: Execution Time Comparison

The Figure 6 depicts that the execution of three classifiers such as SVM, KNN and naïve bayes is compared for the forecasting of wheat yield. The naïve bayes classifier has minimal execution time in comparison with other classifiers as per analysis.

Table 1: Performance Analysis

Parameter	SVM	KNN	Naïve bayes
Accuracy	74.5%	91.53%	95.34%
Precision	0.74	0.91	0.95
Recall	0.73	0.90	0.94
F Measure	73%	91%	93%
Execution Time	3.38 second	2.03 second	1.57 second

V. Conclusion

The greenish products produced in the land used by the creature leads to a strong and wellbeing life. Data mining is the procedure to analyze data from different viewpoints and summarize it into valuable information. In this study, the naïve bayes classification model is implemented for the forecasting of wheat yield. The performance of naïve bayes classifier is compared with KNN and SVM classifier. The naïve bayes shows highest accuracy of about 95.34 approx for wheat yield forecasting.

REFERENCES

[1] S. Sunita, B. J. Chandrakanta, and R. Chinmayee, "A Hybrid Approach of Intrusion Detection using ANN and FCM," Eur. J. Adv. Engg. Tech. , vol. 3, no. 2, pp. 6–14, 2016.

[2] T. Gladkykh, T. Hnot, and V. Solskyy, "Fuzzy Logic Inference for Unsupervised Anomaly Detection," IEEE 1st Int. Conf. Data Stream Mining and Proc., vol.9, pp. 42–47, 2016.

[3] M. M. Ali, "Framework for Surveillance of Instant Messages in Instant messengers and Social neworking sites using Data Mining and Ontology," IEEE- Student's Technology Symposium, vol. 11, pp. 297–302, 2014.

[4] M. Durairaj and G. R. J. Farzana, "Criminal behavior analysis by using data mining techniques," Proc - Int. Conf. Adv. Eng. Sci. Manag ICAESM 2012, vol. 11, pp. 30-31, January 2012

[5] P. K. Khobragade and L. G. Malik, "Data Generation and Analysis for Digital Forensic Application using Data Mining," Proc - 4 Int. Conf. Comm. Syst. Netw. Techno. ICCSNT 2014, vol. 11 pp. 12-23, 2014

[6] S. Bharti and A. Mishra, "Prediction of Future Possible Offender's Network and Role of Offenders," Proc. - 2015 5th Int. Conf. Adv. Comput. Commun. ICACC 2015, pp. 159–162, 2016.

[7] D. A. A. Zainaddin and Z. M. Hanapi, "Hybrid of fuzzy clustering neural network over NSL dataset for intrusion detection system," J. Comput. Sci., vol. 9, no. 3, pp. 391–403, 2013.

[8] J. Quentin-Trautvetter, P. Devos, A. Duhamel, and R. Beuscart, "Assessing association rules and decision trees on analysis of diabetes data from the DiabCare program in France," Stud. Health Technol. Inform., vol. 90, pp. 557–561, 2002.

[9] J.M. Simeon and R. J. Hilderman, "Exploratory quantitative contrast set mining: A discretization approach," Proc. - Int.

Conf. Tools with Artif. Intell. ICTAI, vol. 2, pp. 124–131, 2007.
[10] J. P. Jiawei Han, Micheline Kamber, Data Mining – Concepts & Techniques, vol. 11, pp. 47-94, 2006.
[11] S. Sahu, M. Chawla, and N. Khare, "An efficient analysis of crop yield prediction using Hadoop framework based on random forest approach," Proceeding - IEEE Int. Conf. Comput. Commun. Autom. ICCCA 2017, vol. 2017–Janua, pp. 53–57, 2017.
[12] Q. Yan, H. Yang, M. C. Vuran, and S. Irmak, "SPRIDE: Scalable and private continual geo-distance evaluation for precision agriculture," 2017 IEEE Conf. Commun. Netw. Secur. CNS 2017, vol. 2017–Janua, pp. 1–9, 2017.
[13] K. L. Ponce-Guevara *et al.*, "GreenFarm-DM: A tool for analyzing vegetable crops data from a greenhouse using data mining techniques (First trial)," 2017 IEEE 2nd Ecuador Tech. Chapters Meet. ETCM 2017, vol. 2017–Janua, pp. 1–6, 2018.
[14] Luminto and Harlili, "Weather analysis to predict rice cultivation time using multiple linear regression to escalate farmer's exchange rate," Proc. - 2017 Int. Conf. Adv. Informatics Concepts, Theory Appl. ICAICTA 2017, pp. 0–3, 2017.
[15] Y. M. Fernandez-ordoñez, M. Ieee, M. Ieee, I. Nacional, and D. I. Forestales, "Maize Crop Yield Estimation With Remote Sensing And Empirical Models Agrícolas y Pecuarias , Zinacantepec , Mexico .," pp. 3035–3038, 2017.