Analysis of Wheat Production using Naïve Bayes Classifier

Simrat Kaur Bains
Student, M-TECH CSE
D.A.V. Institute of Engineering & Technology,
Jalandhar, Punjab 1440022, India

Shaveta Kalsi
Assistant Professor, CSE
D.A.V. Institute of Engineering & Technology,
Jalandhar, Punjab 1440022, India

ABSTRACT
Data mining is defined as the process in which useful information is extracted from the raw data. In order to acquire essential knowledge it is essential to extract large amount of data. This process of extraction is also known as misnomer. Currently in every field, large amount of data is present and analyzing whole data is very difficult as well as it consumes a lot of time. The prediction analysis is most useful type of data which is performed today. To perform the prediction analysis the patterns needs to generate from the dataset with the machine learning. The prediction analysis can be done by gathering historical information to generate future trends. So, the knowledge of what has happened previously is used to provide the best valuation of what will happen in future with predictive analysis. Crop production analysis is one of the applications of prediction analysis. In this research work, the Naïve Bayes classifier is applied for the wheat production prediction. The Naïve Bayes classifier is compared with SVM and KNN. The Naïve Bayes performs well for the wheat production analysis.

Keywords
Classification, Prediction Analysis, Wheat Production, Naïve Bayes.

1. INTRODUCTION
Data mining is defined as the process in which useful information is extracted from the raw data. In order to acquire essential knowledge it is essential to extract large amount of data. Currently in every field, large amount of data is present and analyzing whole data is very difficult as well as it consumes a lot of time. This present data is in raw form that is of no use hence a proper data mining process is necessary to extract knowledge [1]. This is a world where having a lot of information leads to power and success and this is possible only because of sophisticated technologies such as satellites, computers. With the advent in the technology in the mass digital storage and computers it becomes easy to handle large amount of information by which different types of data is stored. The most important culture being followed in India since ancient times is agriculture. The crops were cultivated by the people in ancient times within their own land areas such that they could fulfill their own requirements. Thus, cultivation has been followed ever since and all the living beings have been dependent on this culture. Therefore, the natural crops are cultivated and have been used by many creatures such as human beings, animals and birds [2]. The greenish goods produced in the land which have been taken by the creature leads to a healthy and welfare life. Since the invention of new innovative technologies and techniques the agriculture field is slowly degrading. Due to these, abundant invention people are been concentrated on cultivating artificial products that is hybrid products where there leads to an unhealthy life. Nowadays, modern people don't have awareness about the cultivation of the crops in a right time and at a right place. Because of these cultivating techniques the seasonal climatic conditions are also being changed against the fundamental assets like soil, water and air which lead to insecurity of food. By analyzing all these issues and problems like weather, temperature and several factors, there is no proper solution and technologies to overcome the situation faced by us. In India there are several ways to increase the economical growth in the field of agriculture [3]. There are multiple ways to increase and improve the crop yield and the quality of the crops. Data mining is also useful for predicting the crop yield production. Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is an analytical tool that allows users to analyze data from many different dimensions or angles, categorize, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. The patterns, associations, or relationships among all this data can provide information. Information can be converted into knowledge about historical patterns and future trends [4]. For example, summary information about crop production can help farmers identify the crop losses and prevent it in future. Crop yield prediction is an important agricultural problem. Each and Every farmer tries to know, how much yield one will get from his expectation. In the past, yield prediction was calculated by analyzing farmer's previous experience on a particular crop. The Agricultural yield primarily depends on weather conditions, pests and planning of harvest operation. Accurate information about history of crop yield is an important thing for making decisions related to agricultural risk management [5]. This research focuses on evolution of a prediction model which may be used to predict crop yield production. There are several applications in the field of agriculture. To maximize the crop yield, selection of the appropriate crop that will be sown plays a vital role. It depends on various factors like the type of soil and its composition, climate, geography of the region, crop yield, market prices etc. Techniques like Artificial neural networks, K-nearest neighbors and Decision Trees have carved a niche for themselves in the context of crop selection which is based on various factors. Crop selection based on the effect of natural calamities like famines has been done based on machine learning. The use of artificial neural networks to choose the crops based on soil and climate has been shown by researchers [6]. A plant nutrient management system has been proposed based on machine learning methods to meet the needs of soil, maintain its fertility levels, and hence improve the crop yield. A crop selection method called CSM has been proposed which helps in crop selection based on its yield prediction and other factors. Indian agriculture mainly relies

on seasonal rains for irrigation. Therefore, an accurate forecast of weather can reduce the enormous toil faced by farmers in India including crop selection, watering and harvesting. As the farmers have poor access to the Internet as a result of digital-divide, they have to rely on the little information available regarding weather reports [7]. Up-to-date as well as accurate weather information is still not available as the weather changes dynamically over time. Researchers have been working on improving the accuracy of weather predictions by using a variety of algorithms. Artificial Neural networks have been adopted extensively for this purpose. Likewise, weather prediction based on machine learning technique called Support Vector Machines had been proposed. These algorithms have shown better results over the conventional algorithms. Farming sector consumes a huge portion of water in India. The levels of ground water are dropping down day-by-day and global warming has resulted in climate change. The river water for irrigation is a big issue of dispute among many states in India [8]. To combat the scarcity of water, many companies have come up with sensor based technology for smart farming which uses sensors to monitor the water level, nutrient content, weather forecast reports and soil temperature. EDYN Garden sensor is another example.

2. LITERATURE REVIEW

Jharna Majumdar, et.al (2017) proposed a technique through which the optimal climate requirements which were necessary for increasing the production of wheat were obtained by applying various data mining techniques such as PAM, CLARA and DBSCAN. Particular quality metrics were utilized to compare the clustering techniques [9]. It was seen through the performed analyses of clustering quality metrics that in comparison to PAM and CLARA, the quality of clustering was better when DBSCAN was applied. However, the clustering quality of PAM was even less than CLARA. Soil and various other factors that defined the growth of crop could be analyzed in future to extend the proposed work. Also, the future work could focus on increasing the production of crop during the changes in climatic conditions of surrounding regions.

Shriya Sahu, et.al (2017) presented a study related to big data in this paper. Agriculture is the main source of human survival in which the crop data analysis is a very important factor to be considered. From the accuracy of agriculture information, the identification of experiences is done with the help of big data in this paper [10]. In order to determine which types of crops are to be planted on the basis of the content of soil, a better prediction mechanism is generated for farmers through which the productivity can be enhanced. In Hadoop framework, the random forest algorithm is integrated along with Map Reduce programming model.

Qiben Yan, et.al (2017) proposed a scalable and private continual geo-distance evaluation system known as SPRIDE such that geographic based services can be provided. In a private as well as continuous manner, the distances amongst sensors and farms are calculated here. Without learning any additional information related to the locations, the distances of servers are determined [11]. There is real-time private distance evaluation achieved on the large network of farms due to the utilization of SPRIDE such that there is enhancement of up to 17 times of runtime performance in comparison to the existing techniques as per the simulation results.

K. L. Ponce-Guevara, et.al (2017) presented a study related the most important factors such as humidity, soil moisture, carbon dioxide and lighting level, that influence the photosynthesis of plants such that the crop growth in a greenhouse can be affected. There will be rise in presence of nutrients with the establishment of correct values and the quality of fruits will also rise with the help of this approach [12]. With the help of huge data, the pattern recognition is focused in this approach. There is no specific governance of data analytics through a standard with the help of these tools and techniques. However, a set of algorithms in which the descriptive models can be generated on a set of data such that the information can be classified and predicted is provided here.

Luminto, et.al (2017) proposed a novel multiple linear regression model for predicting the rice cultivation time. Here, for the session 2016-17, the highest Farmer's Exchange Rate at 2 season regions is achieved. The significant variables used here are Average Temperature and Solar Radiation [13]. Only these two variables are utilized here which are not enough for prediction. Through testing of all variable combinations that cause less RMSE values, the prediction can be made within particular regions. In order to predict the issue that is based on multiple dependent variables, an appropriate method known as multiple linear regression technique is utilized. The implementation of this approach is very easy and in comparison to other machine learning techniques, this technique provides high speed results.

Yolanda. M. et.al (2017) presented a study in which for the various flowering stage of maize, the yield was estimated. Thus, on the basis of observed true field values, higher accuracy was calculated. Around 14% of yield for LAI based prediction model and 97% of accuracy for NDVI based predictive model was estimated [14]. The variation found within the field data collection that occurred at various hours of the day is the major cause of the behavior of LAI based model. The angle of incidence of sun light within the plant canopy is directly affected by it. For the implementation of grain imports policies in relevance to domestic demand, the government officers utilize these estimates.

Anshul Garg, et.al (2017) proposed the establishment of relationships amongst fuzzy intervals through the utilization of Frequency based Partitioning which helps in fuzzily of data. By using the Years as well as the values achieved from the Fuzzy Logical Relationships a graph is plotted to perform regression analysis [15]. In order to assess, appraise as well as estimate the rice production, this proposed fuzzy approach is known to be infallible and economical. In order to handle the multidimensional time series data, the planned model is extended for future work. Also, more advanced algorithms are optimized with this proposed approach. Thus, on the data various degrees of Fuzzy Logical Relationship are applied such that high order FLR results can be achieved. Further, a different and more efficient mechanism is to be selected such that the Universe of Discourse can be partitioned in future.

Susanto B. Sulistyo, et.al (2017) proposed a novel approach through which the nutrient content present within wheat leaves is estimated. This novel technique is basically a computational intelligence vision sensing approach [16]. For the normalization of plant images and for the minimization of color variability because of the variation of sunlight intensities a deep sparse extreme learning machines (DSELM) fusion and genetic algorithm (GA) is proposed. Along with committee machine, a number of DSELMs are integrated and GA is utilized for their optimization such that the nitrogen

content within the wheat leaves can be estimated. With respect to quality and processing speeds within all the steps, this approach shows enhance results as per the conducted simulations.

Abishek.B, et.al (2017) presented that it is very difficult to predict the effective rainfall and crop water. There are certain factors such as temperature and humidity that need to be known in order to provide a meticulous and scrupulous analysis [17]. In order to define the effective amount of rainfall that has been within a designated region in simple manner, several factors such as humidity, groundwater, and temperature have been considered. For predicting the amount of rainfall and predicting the crop water requirements of specific region, this technique has been utilized. For the determination of effective rainfall and crop water requirements within particular region such that the crop yield can be maximized, the proposed approach is applied. Various issues that have come forth during irrigation of crops in a region are also avoided through this approach.

3. PROBLEM FORMULATION

The prediction analysis is the technique which can predict the future possibilities from the existing data. The prediction analysis techniques are based on the clustering and classification. In the base paper, the approach for the wheat production is analyzed and the dataset is collected from different internet sources. The base paper modal for the prediction analysis is based on the neural networks. The clustered data is given as input to the classification algorithm which can divide the dataset into two parts testing and training. The SVM classifier is used to classify the data into certain number of classes. In the k-mean clustering algorithm, the centered points are calculated by taking arithmetic mean of the whole dataset which can reduce accuracy of prediction analysis. When the dataset is complex, it is difficult to establish relationship between the attributes of the dataset. In this research work, the KNN classifier is applied which can classify the wheat production in certain number of classes. The KNN classifier can be replaced with some other classifiers to improve accuracy of classification.

4. RESEARCH METHODOLOGY

Following are the various phases of research process:-

1. Pre-Processing:- The first phase of the research process is the pre-processing in which dataset is loaded which is collected from the UCI repository. The input data is cleaned in this phase means missing values are removed from the dataset.

2. Feature Extraction: - In the second phase, the technique of feature extraction is applied which establishes relationship between every attribute of the data with the target set. When the technique of feature extraction is applied it becomes easy to identify important attributes.

3. Classification: - In the last phase, the Naïve Bayes classifier is applied on the data. The classification results generate the output of the prediction. A subset of Bayesian decision theory which is growing popularity lately is Naïve Bayes. This algorithm is applied in critical applications since it requires minimal storage and has a fast training process. The objective of this algorithm is to create a rule through which the future objects can be allocated from a set of objects to a class given that the vectors of variables mark out the future objects. There is a supervised classification issue found very commonly and several methods have been designed to construct such rules. Any complicated repetitive parameter estimation mechanism is not required in this algorithm which means that huge data sets can apply this algorithm. Even the unskilled users using the classifier technique can understand the classification process since it is easy to be interpreted.

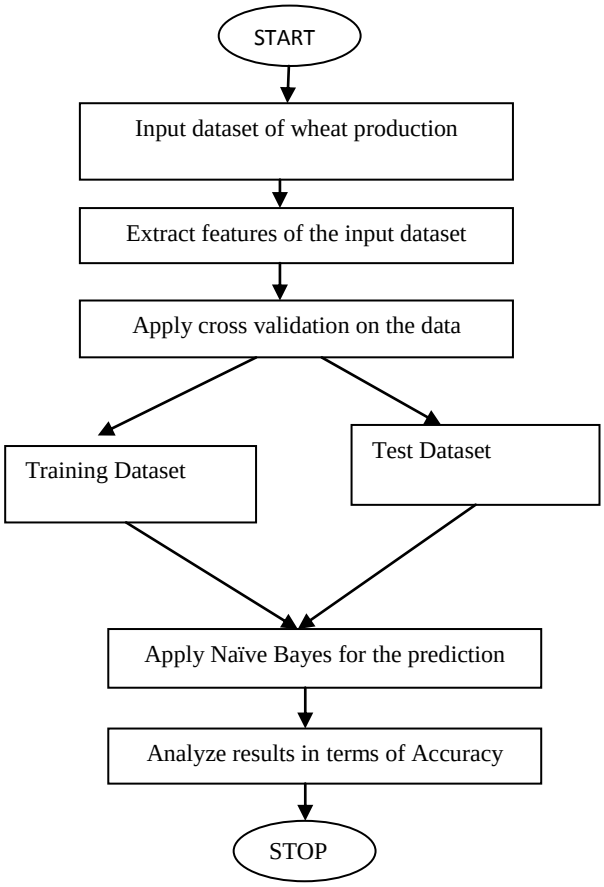

Fig 1: Proposed Flowchart

5. RESULT AND DISCUSSION

This research work is related to wheat production prediction in data mining. The technique of Naïve Bayes is applied in this work for the wheat production prediction. In the existing work, the technique of KNN is used for the wheat production analysis. The dataset of wheat production is collected from the different internet sources. The performance of SVM, KNN and Naïve Bayes are compared in terms of accuracy.

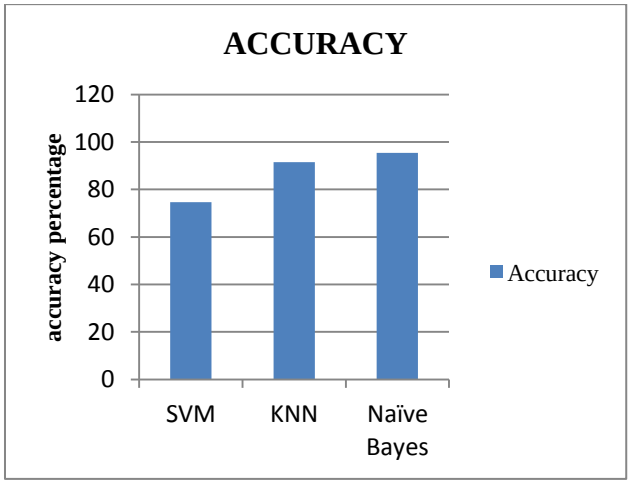

Fig 2: Accuracy comparison

As shown in figure 2, the accuracy of the three classifiers which are SVM, KNN and Naïve Bayes are compared for the wheat production prediction. It is analyzed that Naïve Bayes classifier has maximum accuracy as compared to other classifiers which are SVM and KNN.

Table 1: Accuracy comparison

Parameter	SVM	KNN	Naïve Bayes
Accuracy	74.58	91.53	95.34

As shown in table 1, the accuracy of the SVM, KNN and Naïve Bayes is compared numerically. The Naïve Bayes gives maximum accuracy as compared to other classifiers for the wheat production prediction.

6. CONCLUSION

The crops were cultivated by the people in ancient times within their own land areas such that they could fulfill their own requirements. Thus, cultivation has been followed ever since and all the living beings have been dependent on this culture. Therefore, the natural crops are cultivated and have been used by many creatures such as human beings, animals and birds. The greenish goods produced in the land which have been taken by the creature leads to a healthy and welfare life. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. In this work, the Naïve Bayes classifier is applied for the wheat production prediction. The Naïve Bayes classifier is compared with KNN and SVM classifier. The Naïve Bayes give maximum accuracy of 95.34 for the wheat production prediction.

7. REFERENCES

[1] Tetiana Gladkykh, Taras Hnot and Volodymyr Solskyy, Fuzzy Logic Inference for Unsupervised Anomaly Detection, IEEE First International Conference on Data Stream Mining & Processing , vol. 9, issue 4, pp. 42-47, 2016.

[2] Mohammed Mahmood Ali, Khaja Moizuddin Mohammed and Lakshmi Rajamani, "Framework for Surveillance of Instant Messages in Instant messengers and Social networking sites using Data Mining and Ontology", IEEE- Students' Technology Symposium, Vol. 11, issue 3, pp. 12-23, 2014.

[3] K. Zakir Hussain, M. Durairaj and G. Rabialahani Farzana. "Criminal Behavior Analysis By Using Data Mining Techniques", IEEE-International Conference on Advances in Engineering, Science and Management (ICAESM -2012), Vol. 11, issue 3, pp. 30-31, 2012.

[4] Prashant K. Khobragade and Latesh G. Malik, "Data Generation and Analysis for Digital Forensic Application using Data mining", Fourth International Conference on Communication Systems and Network Technologies, Vol. 11, issue 3, pp. 12-23, 2014.

[5] Sushant Bharti, Ashutosh Mishra. "Prediction of Future possible offender's network and role of offender's", Fifth International Conference on Advances in Computing and Communications, Vol. 11, issue 3, pp. 12-23, 2015.

[6] Dahlia Asyiqin Ahmad Zainaddin and Zurina Mohd Hanapi, Hybrid of Fuzzy Clustering Neural Network over Nsl Dataset for Intrusion Detection System, Journal of Computer Science, Volume 9, No. 3, pp. 12-44, 2013.

[7] Ashwani Kumar Kushwaha, Sweta Bhattachrya, "Crop yield Prediction using Agro Algorithm in Hadoop", published in International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol. 5, No2, April 2015.

[8] Gaurav Sharma, Rudrakshi, "Enhancing Back Propagation Neural N/w Algorithm for crop prediction", published in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6, June 2014.

[9] Jharna Majumdar, Sneha Naraseeyappa and Shilpa Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data", 2017, Journal of Big Data.

[10] Shriya Sahu, Meenu Chawla, Nilay Khare, "An Efficient Analysis Of Crop Yield Prediction Using Hadoop Framework Based On Random Forest Approach", International Conference on Computing, Communication and Automation (ICCCA2017).

[11] Qiben Yan, Hao Yang, Mehmet C. Vuran, Suat Irmak, "SPRIDE: Scalable and Private Continual Geo-Distance Evaluation for Precision Agriculture", IEEE Conference on Communications and Network Security (CNS), 2017.

[12] K. L. Ponce-Guevara, J. A. Palacios-Echeverrıa, E. Maya-Olalla, H. M. Dom´ınguez-Limaico, "GreenFarm-DM: A tool for analyzing vegetable crops data from a greenhouse using data mining techniques (First trial)", IEEE , 2017.

[13] Luminto, Harlili, "Weather Analysis to Predict Rice Cultivation Time Using Multiple Linear Regression to Escalate Farmer's Exchange Rate", IEEE, 2017.

[14] Yolanda. M. Fernandez-Ordoñez, J. Soria-Ruiz, "MAIZE CROP YIELD ESTIMATION WITH REMOTE SENSING AND EMPIRICAL MODELS", IEEE, 2017.

[15] Anshul Garg, Bindu Garg, A Robust and Novel Regression Based Fuzzy Time Series Algorithm for Prediction of Rice Yield", International Conference on Intelligent Communication and Computational Techniques (ICCT), 2017.

[16] Susanto B. Sulistyo, Di Wu, Wai Lok Woo, S. S. Dlay, and Bin Gao, "Computational Deep Intelligence Vision Sensing for Nutrient Content Estimation in Agricultural Automation", IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, 2017.

[17] Abishek.B, R.Priyatharshini, Akash Eswar M, P.Deepika, "Prediction of Effective Rainfall and Crop Water Needs using Data Mining Techniques", 2017 IEEE International Conference on Technological Innovations in ICT For Agriculture and Rural Development (TIAR 2017).