ISSN 2278-3091

Volume 10, No.2, March - April 2021

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse061022021.pdf

https://doi.org/10.30534/ijatcse/2021/061022021

Applying Logistic Regression Data mining techniques for Ethiopian Government Agricultural Open Data Sets

Gizealew Alazie Dagnaw[1], SisayEbabyeTsigie[2]

Department of Information Science, College of Informatics, University of Gondar, Gondar, Ethiopia

gizeinstra@gmail.com,sisebe2007@gmail.com

ABSTRACT

Ethiopia has a great agricultural potential because of its vast areas of fertile land, diverse climate, generally adequate rainfall, and large labor force. With its verified importance to the Ethiopian economy, there is sufficient evidence to show that the potential of the agricultural sector can be expanded considerably by attracting investors towards the sector. This study aims at applying classification techniques in developing a predictive model that can estimate yield production of vegetable crops and the correlation of crops based on their class. In the process of building a model, different steps were undertaken. Among the steps, data collection, data preprocessing and model building and validation were the major ones. Different tasks performed in each step are mentioned as follows. The data were collected Food and Agriculture Organization of the United Nations (FAO). Under preprocessing, data cleaning, discretization and attribute selection were done. The final step was model building and validation and it was performed using the selected tools and techniques. The data mining tool used in this research was Weka. In this software the logistic regression algorithm was selected since it is capable to score more accuracy. After successive experiments were done using this software, a model that can classify crop yield as high, medium and low with better accuracy to the extent of 88.6%. Experimental results show that logistic regression is a very helpful tool to depict the contribution of yield estimation and crop correlation. The reported findings are optimistic, making the proposed model a useful tool in the decision making process. Eventually, the whole research process can be a good input for further in-depth research.

Key words: Data mining, predictive model, logistic regression, vegetable crop, Ethiopia

1. INTRODUCTION

Agriculture is the corner stone of the development policy of the Government of Ethiopia. The country's economic development will depend, in large part on sustainable improvements in agriculture. Agriculture remains by far the most important sector in the Ethiopian economy for the following reasons[1]: (i) It directly supports about 85% of the population in terms of employment and livelihood; (ii) It contributes about 50% of the country's gross domestic product (GDP); (iii) It generates about 90% of the export earnings. Agriculture is also the major source of food for the population and hence the prime contributing sector to food security .In addition, agriculture is expected to play a key role in generating surplus capital to speed up the overall socio economic development of the country. A high rate of agricultural growth has far reaching positive implications for economic development of low income countries in terms of increasing employment and accelerating poverty reduction .The majority of the Ethiopian populations live in rural areas where agriculture is the main occupation and source of livelihood. It contributes for about 47.3% of GDP and 90% of export earnings[2] .Ethiopia is one of the developing countries with high population and food insecurity. The country has been implementing different strategies to achieve food security. Diversification of crops, increasing the availability of food production, and encouraging the production of early maturing and high yielding crops in different agro-ecologies of the country are some of such strategies . Food security is one of the most important problems for the rural population of Ethiopia, whose life is almost entirely dependent on agricultural products. Ethiopia is characterized by having different agro-ecological zones and it accounts about a total area of 1.13 million km2 5. A variety of fruit crops has been growing in different agro ecological Zones by small farmers, for subsistence and income generation[3].

Data Mining (DM) is the process of analyzing data from different perspectives and summarizing it into useful information. There are different DM algorithms exist, including the predictive Data Mining algorithms, which result in classifiers that can be used for prediction and classification, and descriptive data mining algorithm that serve other purposes like finding of associations and clusters . Data mining application has been recently gained much attention of every application fields like industry, economics, medicine, CRM, trade, etc., due to the existence of large collections of data in different formats, and the increasing need of data analysis and comprehension. Since data mining is the most important tool to discovery knowledge from large database. It is a process of semi-automatically analyzing large databases to find valid, novel, useful and understandable patterns. In addition, Data mining has paid attention to modeling as much as possible. Since Agriculture is the backbone of the Ethiopian economy, As such in the context of Ethiopia crops are cultivated between two cropping seasons i.e. during belg and meher. Based on the researcher preliminary discussion with experts currently the productivity of crop prediction has been done using farmers past experience, through field observations and the production output also predicted using statistically estimation of the crops with field observation during pre and post harvesting. In addition, the statistical prediction of crops production is not sufficient to predict the determinant factors for crops productivity. Nowadays, crop Productivity prediction is essential to identify the cause for low or high productivity factors and used to enhancing the productivity and production of smallholder farmers mainly by reducing the traditional ways of estimating productivity. As a result, it used to strengthen the implementation of effective cropping strategies for national development program and it has been benefited to make data mining based decision making system for decision makers and experts[4].

Crop agriculture in Ethiopia continues to be dominated by the country's numerous smallholder farms that cultivate mainly cereal crops for both own-consumption and sales. This research presents an applied study using data mining to discover some factors affecting agricultural vegetable production and predicting the yield production in Ethiopia. In this research, we are interested in finding some factors that will influence the agricultural production to increase the amount of production to benefit the farmers in particular and individual, society in general[5].

1.1 Problem Statement

Agriculture is one of the most important inventions of human civilization. The development of human civilization and development of agriculture technology were the two wheels of the cart. Unfortunately, it has been witnessed that the development of agriculture technology is not in the same ratio as human civilization is developed. Traditional tools and techniques used for forming are neither sufficient to predict nor, to optimize production results of yield[4]. The agricultural data is diversified, complex and non-standard and information available about agriculture is in the form of static maps or tables or reports. Agriculture and plantation is an important and interesting research area everywhere in the world and Ethiopia is no exception. Nowadays available land area for a plantation is becoming scarce. This scarce resource is frequently wasted through our bad practices and improper management. Cultivation is a more economical but complex process. These tasks and the whole process need a lot of expert knowledge and experience. But unfortunately, people having this type of knowledge are very limited. Their assistance is not available when the person who is going to cultivate needs it. Agriculture, being a very vast and varied domain of knowledge with over a hundred crops distributed in different geographic regions having varied climatic conditions, building such a team in every domain of knowledge of agriculture is itself a challenging and huge task The purpose of human to do cultivation is to meet the needs of growing food along with the increase of population. In addition the plant also serves as a provider of oxygen to the human respiratory system as well as the aesthetic and the beauty that can be enjoyed by humans[3].

Ethiopia's wide range of agro climatic conditions and soil types make it suitable for the production of diverse verities of fruits and vegetables including temperate, tropical and subtropical vegetables. Cabbages, onions, passion fruits, bananas, avocados, citrus fruits, mangoes, tomato, papayas, etc., are produced in Ethiopia. The main fruits produced and exported are banana, citrus fruits, mango, avocado, papaya and grape fruits .Tropical and sub-tropical fruit can make a significant direct contribution to the subsistence of small-scale farmers by providing locally generate nutritious food that is often available when other agricultural crops have not yet been harvested. Vegetables are a versatile product that, depending on need, can be consumed within the household or sold. Marketing fresh and processed fruit products generates income which can act as an economic buffer and seasonal safety net for poor farm households. Diversification into fruit production can generate employment and enable small-scale farmers to embark on a range of production, processing and marketing activities to complement existing income-generating activities.

Vegetables and Fruits in different forms such as whole fruit, fruit juice, fruit pulp, and fruit concentrate have a vital role for health. They are dietary sources of nutrients, micronutrients and vitamins for human and are thus vital for health and well-being. Well balanced diet rich in fruits and vegetables are especially valuable for their ability to prevent deficiency diseases and are also reported to reduce the risk of several diseases. For many countries fruit products have become valuable, making a substantial contribution to the economy as well as to the health of country population. Vegetables and fruits contain vitamin C, foliate and dietary fibers and other bioactive components such as carotenoids and flavonoids which are suggested to be responsible for the prevention of degenerative diseases. Studies have shown that if fruits are consumed daily in sufficient amount, it could help to prevent major diseases such as cardio vascular and certain cancers. According to WHO report, low fruit and vegetable intake is estimated to cause about 31 % of heart disease and 11 % of stroke worldwide and around 2.7 million lives could potentially be saved each year if fruits and vegetable consumption was sufficiently increased[6].What data mining algorithm is more appropriate for predicting the vegetable crop?, how to evaluate the model? And how to design and develop a model?

1.2 Objective

The general objective of this study is to explore the potential applicability of data mining technology in developing a model that can support agricultural researchers in predicting the amount of adequate vegetable crops based on FAO data set

- ✓ To collect the valuable data needed to do the research FAO website
- ✓ To prepare the data for analysis using different preprocessing techniques, this involves extracting the data, dealing with missing values and transforming into the format required for the data mining algorithm
- ✓ To select the data mining tools and algorithms to develop the model
- ✓ To design and develop the model
- ✓ To evaluate (test) the model
- ✓ To analyze the outcome of the research and make recommendations based on findings

1.3 Scope

The scope of this research is limited to assess the possible application of data mining technology in Vegetable crops in Ethiopia. Regarding the limitation that the researcher faced, there was time constraint to incorporate as many experiments as possible in the process of building models. Besides, there was absence of sufficient data for analysis. Above all, it wasn’t easy to meet and discuss with agricultural experts. Anyhow, we tried to communicate through telephone, e-mails to handle the problem.

1.4 Significance

The immediate beneficiaries of the study are primary agriculture workers and agriculture professionals or agronomist. Particularly, the research will have great significance to teach primary agriculture extension workers, general agronomy experts in order to have well understanding about vegetable crops future yield. Additionally, the research can be used for agriculture professionals as a guide. Even though those professionals are highly qualified persons, they may get difficulty of remembering all the critical production of each vegetables and fruits.

2. RESEARCH METHODOLOGY

Agricultural production in the world and in Ethiopia are great importance in terms of satisfaction and high quality production, especially in terms of meeting farmers and population needs. Increasing agricultural products has become a necessary need in our time. This research applies predictive data mining techniques in agriculture to predict the amount of yield for vegetable crops and determinate factors. In DM there are four process models. These are KDD, SEMMA, CRISP and the newly emerging hybrid DM process model. For the purpose of conducting this research the six-step process hybrid DM process model is selected in order to estimate yield production and correlation analysis. The main reason hybrid DM process model selected was it combines the main aspects of both models of academic and research. The data source for this research is obtained from the Food and Agriculture Organization of the United Nations (FAO)

2.1.1 Understanding agricultural data

Initial data set for this research work, which is the last twenty five years' survey data of vegetable crop production ranging from 1993 up to 2018 G.C, was collected from the FAO's database. The surveys were conducted to provide data on farmland area and production of crops on the private peasant holdings for all season. The empirical data from the FAO's statistical reports, which provide basic quantitative information on total cultivated land area and vegetable crop production, contain major attributes that have relation to crop production for vegetables crops in Ethiopia.

2.1.1.1 Data collection method

Crop production and land use data sets are taken from the Food and Agriculture Organization of the United Nations (FAO) collects the data at different areas of the country and this data is stored centrally at the on the organization database with their data processing software or (excel). Therefore, to conduct this study, the researcher was taken the data set from the Food and Agriculture Organization of the United Nations (FAO).

2.1.1.2 Description of the Data

As described above, the initial vegetable crop production dataset was taken from the Food and Agriculture Organization of the United Nations (FAO) database contains many attributes together with their instances. According to[31]data quality can be verified in terms of its accuracy, completeness, consistency, timeliness, believability and interpretability which are helpful to examine the quality of the data before conducting the experimentation. In this regard, the initial dataset has been statistically described and visualized using SPSS package and Microsoft excel to examine the properties of the whole dataset records and to obtain high level information regarding the data mining questions. Simple statistical analysis has been performed to verify the quality of the dataset, addressing questions such as: do the data cover all cases required? Is the data correct or does it contains errors? Are there missing values in the data?

2.1.1.3 Exploration of the data

Exploration of the data includes surveying the data that is taking a high-level overview to discover what is contained in the dataset and to gain over all insight into the nature of the data. Surveying the data, therefore, looks at the general structure of the data and reports to identify whether there is a useful information contained in the dataset about various areas of the business or not. The particular purpose of the data survey is to find out if the answer to the problem to be modeled actually exists in the dataset prior to investing much time, money, and resource in building the model. Consequently, data surveying and analysis address directs the data mining goals[32].

According to[33],exploring the nature and the relationships of the information contained in a dataset is the task of the data survey. In addition, finding the places, defining the limits, and understanding the structures of the dataset is the purpose of data surveying. Thus, the whole purpose of the data survey is to help the miner to draw a high-level map of the business territory. With the help of this map, a data miner discovers the general nature or characteristics of the data, as well as area of gaps, limitation, and usefulness of the data. As a result, the researcher has performed basic statistical data analysis on the initial dataset to clarify the data mining goals or to make them more precise. In this task, basic statistical exercise has been conducted to identify the characteristics of interesting sub-populations using SPSS and MS Excel software. Then, we have analyzed the properties of major attributes that indicate the data characteristics or lead to interesting data subsets for further examination. Accordingly, in the following section we describe the first findings of the primary data analysis and evaluate this information regarding their impact on the remainder of the study.

2.1.1.4 Data Preparation

The main objective of data preparation is to get a prepared dataset (or datasets) that is of maximum use for modeling, in which the natural order of the data is least disturbed and best enhanced for the particular purposes of the miner. The best way to actually make the changes in the data depends on two key decisions: what the solution requires and what the mining tool requires, since these decisions affect how the data is prepared, while the inputs to and outputs from the process are not affected[33].

In this study, the major activities done during data preparation phase included *data selection*, *data cleaning, attribute* or *feature selection, data transformation and aggregation, data integration* and *formatting* of the dataset. The purpose of these activities was to produce best model that can predict vegetable crops production in efficient and cost-effective ways. The following sub sections elaborate on these tasks in detail.

2.1.1.5 Data Selection

According to[34], one of the major activities that would be carried out during data preparation phase is data selection, which deals with decision on the target data set, by focusing on a subset of variables and data samples, on which the knowledge discovery task is to be performed. The criteria used for data selection include: relevance of the data items to the data mining goals, data quality and technical constraints such as limitations on data size or data types. Besides, the criteria for excluding data may include resource constraints, cost, restrictions on data use, or other data quality problems. Sometimes, the whole collected dataset may not be taken for the experiments. Thus, the relevancy of each data to the overall research goals and objectives need to be checked.

2.1.1.6 Data Cleaning

Usually, real world databases contain incomplete, noisy and inconsistent data and such uncleanness data may cause confusion in the data mining process. Thus, data cleaning has become a very important activity during data preparation phase in order to assure the quality of data so as to improve the accuracy and efficiency of the data mining techniques.

Data cleaning task deals with all the data quality issues until the targeted dataset reaches the level required by the selected analysis techniques. This may involve selection of clean subsets of the data, insertion of suitable defaults or more ambitious techniques such as the estimation of missing data by modeling. Generally, data cleaning is a process which fills missing values, removes noise (invalid data), and corrects data inconsistency[33].

Accordingly, the researcher took actions such as *missing value handling*, and *outlier detection and removal* on the selected dataset to improve the quality of the targeted data set. The data cleaning report describes the decisions and actions that were taken to address the data quality problems of the initial data that were reported during data quality verification tasks. The report also addresses outstanding data quality issues and possible effects it may have on the final results.

2.1.1.7 Missing Value Handling

Missing value refers to the values of one or more attributes in a data that do not exist. Sometimes missing value may be significant by itself and has to be properly handled. Theoretically, there are several methods suggested in the data mining literature to handle missing values, such as: calculating the average of continuous attribute values and filling this mean value for missing attribute values, removing the tuple, using the global constant value i.e. question mark (?), and filling in the missing value manually[33].

The statistical summary of the initial dataset shows that some attributes in the original dataset contain missing values, ranging from 0.1% to as high as 100%, which are difficult to predict, replace or fill. Thus, the researcher has decided to ignore those attributes that have large amount of missing values from the dataset to assure the quality of the data. This is because, the research believes that trying to fill these missing values using any accepted method will result in changing the original input data with artificial data.

As stated in the data description and data quality verification tasks, that are shown above in tables out of the total 12 attributes of the original dataset, 4 attributes were ignored because of their large amount of missing values and hold area identification information i.e., Code of District, Farmers Association, Enumeration Area, etc., which have no relevance for the research since the scope of the study is limited to regional level,

Therefore, for this study 8 attributes that have less missing values, from 0% to 0.3%, were selected and the remaining attributes were omitted because of the above mentioned reasons. The missing values found in the selected attributes are insignificant when compared with the initial data set. Then, these missing values were handled by removing all the records since they are few in numbers. The basic reason is that these missing values are few in number and contain values that cannot be predicted, replaced or filled with mean value, and if we do so, it will create biasness on the output of the experiment or lead to false/wrong interpretation.

2.1.1.8 Outlier Detection and Removal

The data stored in a database may reflect outlier noise, exceptional case, incomplete data object and random error in a measure of attribute values. These incorrect attribute values may occur due to data entry problems, faulty data collection, inconsistency in naming convention or technology limitation. According to[31], there are four basic noise handling methods for a given dataset such as: *binning, clustering, regression and combined computer and human inspection* methods.

There are also other outliers handling methods that have been developed to handle noise for a given dataset. Some of these methods are sensitive to extreme values, like the standard deviation (SD) method, and others are resistant to extreme values. The SD method is a simple classical approach, which uses less robust measures such as the mean values, to identify outliers in a dataset. Mean value is the most common labeling method that detects how much distance the data has from the average, since mean value describes the average value of the global data[35].

Data Construction

Data construction tasks include data preparation operations such as the computation of derived attributes that are derived from one or more existing attributes in the same record, or constructing completely new records or transformed values for existing attributes. According to[33],the basic reasons for constructing derived attributes during the course of data because:

- Background knowledge convinces us that some facts are important and should be represented, even though there is no attribute which currently represent it.
- The modeling algorithm currently in use handles only certain types of data. In this study, we are using logistic regression and J48 algorithm for developing predictive model of vegetable crop correlation between the crop
- When doing the experiments, the outcome of the rule phase may suggest that certain facts are not being covered.

2.1.1.9 Attribute or Feature Selection

The ideal practice for variable selection is to take all the variables in the database, feed them to the data mining tool and let it find those which are the best descriptor. But, in practice this doesn’t work very well. One reason is that the time it takes to build association model increases with the number of variables. The second reason is that blindly including unnecessary columns can lead to incorrect result. Although, in principle some data mining algorithms will automatically ignore irrelevant variables and properly account for related (covariant) columns, in practice it is wise to avoid depending solely on the tool. Often, knowledge of the problem domain helps us to make attributes selection correctly[34].

Consequently, after consultation with the domain experts at FAO about the meaning of the attributes, the researcher decided to eliminate some attributes from the target dataset, since their instances are irrelevant for the analysis of this data mining problem domain. Thus, 11attributes that are irrelevant or redundant or already represented by other attributes in the database were excluded from the target dataset. Besides, those attributes with no variation in their value throughout the dataset and attributes which serve to assign sequence number for the records were also eliminated. The selected attributes are indicated below.

Table 1: Selected attributes

No	Attribute name	Data type	Description
1	Country	Nominal	The district of crop growth
2.	Year	Numeric	Year of crop production
4.	Harvested area	Numeric	Area of the crop which cover in hectare
5	Production	Numeric	The amount of product in tone
6	Flag	Nominal	The measurement of the data from FAO
7.	Crop type	Nominal	The type of crop
8.	Yield	Numeric	The amount of yield in hectogram per hectare

3. RESULT AND DISCUSSION

The process of building model was supported by Weka software. Weka software has different classification packages. Among others, logistic regression, and J48algorithms are selected for the purpose of this research. Weka can read files saved as .csv extensions. Hence, the Excel data of this research was transported to csv format. And, it has preprocessing facilities by the name filter that is found in the 'choose' menu of Weka user interface and under this, one can get various techniques of preprocessing like discretization, replacing missing value etc. In order to accomplish this study, two data mining techniques are used. These are J48 decision tree algorithm and logistic regression algorithm. Weka data mining tool has the facility to generate rule sets using decision tree and rule induction techniques that can help to easily interpret the results of the model. To build a model, the first task performed was importing the cleaned and prepared dataset of CSV format into Weka software.

Results of both models from WEKA is recorded and an analysis is carried out to compare the prediction power of two competing models in accordance with three important measures Accuracy, Root Mean Squared Error and Area under ROC. Accuracy is the percentage of total number of instances correctly classified. RMSE measures the square root of average of squares of errors i.e. the difference between the actual class and the predicted class. ROC, receiver operating characteristic, is a graphical representation to measure the performance of a classifier system; it plots the true positive rate against the false positive rate. The area under ROC curve ranges from 0 to 1, with 1 implies a perfect test and 0 implies a useless test. The analysis also includes confusion matrix, which is a table layout to visualize the performance of a model. A typical confusion matrix consists of rows and columns where each column represents the number of instances in the predicted class and each row represents the number of instances in an actual class. In predictive analytics, a confusion matrix represents the total number of true positives, false positives, false negatives and true negatives

Table 2: Confusion matrix

	Predicted class	
Actual class	True positives	False Positives
	False Negatives	True Negatives

In order to better understand the terminologies, consider a scenario where a test is conducted that on FAO data set. Crop has high yield, medium yield and low yield does have. Test result can be either positive (meaning the crop has high yield) or negative (meaning the crop does not have high yield) In this case, True Positive means the yield with high value is correctly classified with Based on their level; False

Positive means the crop with low yield is incorrectly classified with high yield, True Negative means the crop without high yield is correctly classified with high yield and False Negative means the crop with high yield is incorrectly classified with low yield. True positive rate is also known as Sensitivity and true negative rate is also known as Specificity.

3.1 Analysis of classification Results obtained for Yield

Table shows an overview of the result obtained from all the 2 models when both models are trained on 378 instances. It has been found that logistic regression resulted into highest accuracy.

Table 3: analysis of classification result obtained for yield.

	Logistic regression	J48
Total number of instance	378	378
Correctly classified instance	335	332
Incorrectly classified instance	43	46
Accuracy	88.6	87.8
Root mean squared error	22%	22%
ROC area	0.94	0.9

3.1.1 Confusion matrix of the model

Table 4: confusion matrix of logistic regression result

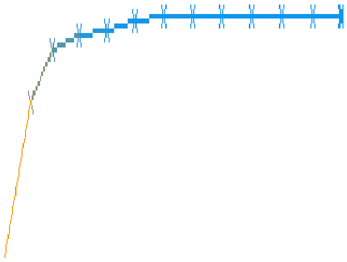
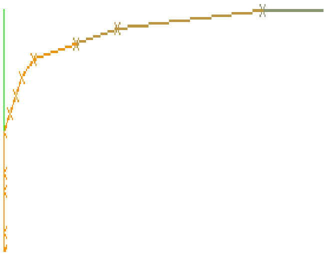
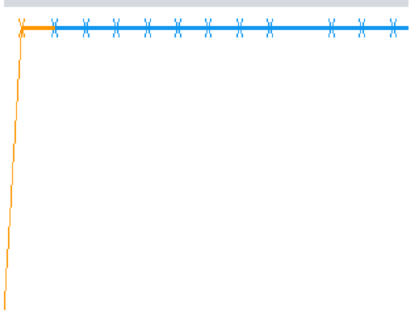
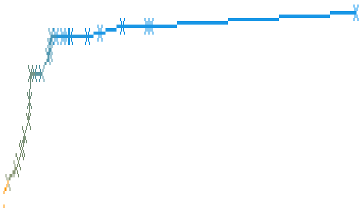
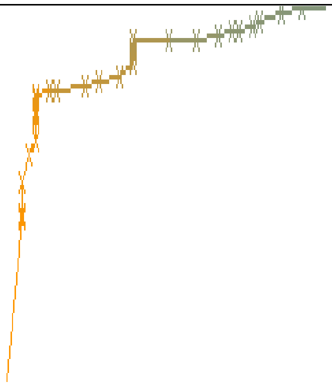
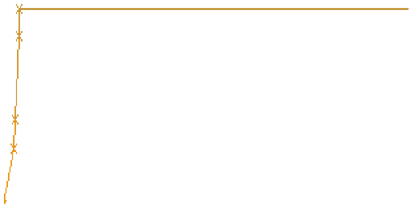
Logistic regression		Predicted class		
		High	Medium	Low
Actual Class	High	310	0	0
	Medium	36	25	
	Low	7		0

Table 5: confusion matrix of J48 result

J48		Predicted class		
		High	Medium	Low
Actual Class	High	304	6	0
	Medium	33	28	
	Low	7		0

3.1.2 ROC for logistic regression and J48 yield prediction

Table 6: ROC for logistic regression and J48 yield prediction

Algorithm	**High**	**Medium**	**Low**
Logistic regression			
J 48			

An incredibly useful tool in evaluating and comparing predictive models is the ROC curve. When we do logistic regression, which are given two classes coded as 1 and 0. Now, compute probabilities that given some explanatory variables an individual belongs to the class coded as 1. If we now choose a probability threshold and classify all individuals with a probability greater than this threshold as class 1and below as 0, in the most cases make some errors because usually two groups cannot be discriminated perfectly. For this threshold we can now compute errors and the so-called sensitivity and specificity.

When do this thresholds, we construct a ROC curve by plotting sensitivity against 1-Specificity for many possible thresholds. The area under the curve comes in to compare different methods that try to discriminate between two classes. On the above construct the ROC curve for both these models and the one with the highest area under the curve can be seen as the best model so logistic regression is the best model for this study. As shown in the above results for Yield prediction, models work differently for different performance measures; logistic regression models provide better accuracy withhigh Area under ROC, J48 models resulted into High root mean squared error but less accuracy and ROC. Hence the study selected the best model that is logistic regression model.

Table 7: yield production level based on odd ratio

Crop type	High yield Value	Medium yield Value	Low yield value
Carrot	✓		
Garlic	✓		
Orange		✓	
Papaya	✓		
Potato	✓		
Sugar cane		✓	
Fruit fresh		✓	

The above tables are shown that the crop which is repeatedly occur on the class because of their odd ratio of the regression shows more than one then OR >1 indicates increased occurrence of event. OR <1 indicates decreased occurrence of event (protective exposure) Look at CI and P value for statistical significance of value. In rare outcomes OR = RR (RR = Relative Risk)[36].

3.2 Discussion of the developed model result

We have discussed the design and implementation of a predictive analytics based system to predict the productivity of having a yield status of each crop. In order to give better results and build a powerful system we used two machine Learning approaches to generate the predictions those are logistic regression and J48. When see the result of the model logistic regression is score more accuracy and ROC then the study take the logistic regression result .when using more than one techniques for overcoming the disadvantages of one model with the help of other models. The system is successfully able to generate the Predictions based on the data provided by the FAO. The research has shown that identify the potential crops for the next time of agricultural investment. Ethiopia has a great agricultural potential because of its vast areas of fertile land, diverse climate, generally adequate rainfall, and large labor force[37]. With its verified importance to the Ethiopian economy, there is sufficient evidence to show that the potential of the agricultural sector can be expanded considerably by attracting investors towards the sector because the research should identify right crop for the right time so Ethiopian potential crops are shown on the table for the next time to get sufficient yields. Agriculture, being a very vast and varied domain of knowledge with over a hundred crops distributed in different geographic regions having varied climatic conditions[38],.This study aims at applying classification techniques in developing a predictive model that can support by identifying and predicting potential vegetable crop for agricultural investment activities. Data obtained from Food and Agriculture Organization of the United Nations (FAO) was preprocessed and made suitable for Weka data mining tool. Yield estimation models are utilized in preciseness Agriculture to extend yield production to satisfy demand and to recommend to the government of Ethiopia in regard to prediction crop yield for sustainable development. As per the information obtained from the document in the five year development strategic plan for sustainable development to end poverty (GTP), the Ethiopian government has been giving significant focus and attention for agriculture and rural development. This is accomplished by offering over 8 million acres of land to commercial farming investors. Expansion will open up opportunities for advanced farming technology, high value crops, progressive irrigation techniques, improved seeds, increased fertilizer use, and strategies to yield multiple harvests each year. Furthermore, the productivity of the sub-sector is affected by poor management system and shortage of skilled experts who provide advice for farmers at Woreda level. Despite the importance of agriculture in its economy, Ethiopia has been a food deficit country since the early 1970s. A closer look at the performance of the Ethiopian agriculture reveals that over the last three decades it has been unable to produce sufficient quantity to feed the country's rapidly growing human population.

4. CONCLUSION AND RECOMMENDATION

Today many agriculture facilities are supported by IT applications resulting in increased agricultural sector services by reducing time and cost of providing agricultural sector services. Data mining is relatively a novel research field and it is expected to grow in the future. This emerging, multidisciplinary and interesting research field, there is a lot of work to be done. Yield production predictive model have been found to be very useful in our today's world driven by technology. Regression logistic model is currently attracting a great deal of interest in the business community including agriculture. Almost all the business contemporary depends on knowledge. So it's always a good idea to find a way to keep and preserve knowledge. In Ethiopia most of the peoples are farmers, it's believed that more than 80% of the population depends on agriculture; therefore it's logical to assume that there exists shortage of agricultural experts in the country. This predictive model can be used as an additional or source of information when good yield are available and the type of crop to score maximum yield. It will help in forecasting/ managing agricultural crop effectively. This research focuses on applying data mining techniques regarding vegetable data to extract knowledge from these data and estimate vegetable crop with 88.6% accuracy. Also, it aims to predict the main factors that impact the vegetable' production to satisfy farmers and citizens' needs for the upcoming years using logistic regression algorithm. The obtained results could help decision makers for achieving food security and the country's productivity for the upcoming years continuously. Logistic regression is useful for predicting the main factors that affect the vegetables production. To reach meaningful outputs and predictions, various experiments conducted through modifications of the attributes and the use of different numbers of these attributes to reach meaningful outputs and prediction.The experiments conducted in this paper showed that farmers attempt to increase vegetable crop by using reliable crop at reliable time in the reliable districts. For the future using the discovered knowledge the researcher develops a prototype using java NetBeans IDE for implementing a decision support system, and that can able to identify the determinant factors that affect crops productivity. Hence, it is essential to make data mining based decisions for policy makers in the area of agriculture which is focused on crop production.

REFERENCES

[1] M. Gates, "UNITED NATIONS DEVELOPMENT PROGRAMME Agricultural Growth and Transformation Strengthening National Capacity through Economic Growth & Poverty Reduction," 2016.

[2] W. F. P. Cropfood, S. Assessment, and M. To, "FAO / WFP CROP AND FOOD SECURITY ASSESSMENT MISSION," no. January, pp. 1–44, 2009.

[3] A. Chekole and T. Beshah, "Application of Data Mining Tools for Identifying Determinant Factors for Crop Productivity," *Int. J. Comput. Appl.*, vol. 181, no. 42, pp. 16–21, 2019, doi: 10.5120/ijca2019918497.

[4] FAO, "Ethiopia Country Programming Framework 2012-2015," *Off. FAO Represent. Ethiop. to AU ECA*, p. 60, 2011.

[5] R. B. Palepu and R. R. Muley, "An Analysis of Agricultural Soils by using Data Mining Techniques," vol. 7, no. 10, p. Angeles, L., Advocacy, S., Location, O. (2002)., 2017.

[6] N. Rehman, "Data Mining Techniques Methods Algorithms and Tools," *Int. J. Comput. Sci. Mob. Comput.*, vol. 6, no. 7, pp. 227–231, 2017.

[7] "Mining Frequent Patterns , Associations and Correlations – Mining Methods – Mining Various Kinds of Association Rules – Constraint Based Association Mining – Classification and Prediction – Basic Concepts – Decision Tree Induction – Bayesian Classificatio."

[8] A. S. Rao, A. V. Ramana, and S. Ramakrishna, "Implementing the data mining approaches to classify the images with visual words," *Int. J. Recent Technol. Eng.*, vol. 7, no. 6, pp. 901–909, 2019.

[9] J. Majumdar, S. Naraseeyappa, and S. Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data," *J. Big Data*, vol. 4, no. 1, 2017, doi: 10.1186/s40537-017-0077-4.

[10] G. N. Fathima, "Agriculture Crop Pattern Using Data Mining Techniques," vol. 4, no. 5, pp. 781–786, 2014.

[11] S. Chouhan, "A Survey and Analysis of Various Agricultural Crops Classification Techniques," vol. 136, no. 11, pp. 25–30, 2016.

[12] H. Fetanat, L. Mortazavifar, and N. Zarshenas, "The Application of Data Mining Techniques in Agricultural Science," pp. 108–116, 2015.

[13] S. A. R. Kumar, “A STUDY ON PADDY CROPS DISEASE PREDICTION USING DATA MINING TECHNIQUES Department of Information Technology,” vol. 7, no. 1, pp. 336–347, 2015.

[14] N. Neelaveni, “DATA MINING IN AGRICULTURE- A Survey,” vol. 4, no. 4, pp. 104–107, 2016.

[15] B. H. Dhivya, R. Manjula, S. B. S, and R. Madhumathi, “A Survey on Crop Yield Prediction based on Agricultural Data,” pp. 4177–4183, 2017, doi: 10.15680/IJIRSET.2017.0603053.

[16] A. A. Chaudhari and H. K. Khanuja, “Database transformation to build data-set for data mining analysis - A review,” *Proc. - 1st Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2015*, no. July 2015, pp. 386–389, 2015, doi: 10.1109/ICCUBEA.2015.81.

[17] C. Science, “HiLCoE Journal of Computer,” vol. 2, no. 1, 2013.

[18] A. S. Taffesse, P. Dorosh, and S. Asrat, “Crop Production in Ethiopia : Regional Patterns and Trends,” 2011.

[19] “School of Natural and Computational Science Information Technology Department,” no. October, 2015.

[20] S. Singhal and G. N. Singh, “Classification using Association Rule Mining,” *Int. J. Comput. Sci. Commun.*, vol. 3, no. 2, pp. 973–7391, 2012.

[21] A. L. Prichard, “da,” no. May, 2012.

[22] K. Raghuveer, “Data Mining in Agriculture: A Review,” *AE Int. J. Multidiscip. Res.*, vol. 2, no. 9, pp. 1682–1690, 2014, doi: 10.1007/s13398-014-0173-7.2.

[23] V. N, “A Survey on Data Mining Techniques in Image Mining,” pp. 296–299, 2018, doi: 10.21467/proceedings.1.47.

[24] A. A. Body, “K s n d m c,” 2016.

[25] R. J. McQueen, S. R. Garner, C. G. Nevill-Manning, and I. H. Witten, “Applying machine learning to agricultural data,” *Comput. Electron. Agric.*, vol. 12, no. 4, pp. 275–293, 1995, doi: 10.1016/0168-1699(95)98601-9.

[26] R. Jahan, “Applying Naive Bayes Classification Technique for Classification of Improved Agricultural Land soils,” *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 6, no. 5, pp. 189–193, 2018, doi: 10.22214/ijraset.2018.5030.

[27] K. H. Coble, A. K. Mishra, S. Ferrell, and T. Griffin, “Big data in agriculture: A challenge for the future,” *Appl. Econ. Perspect. Policy*, vol. 40, no. 1, pp. 79–96, 2018, doi: 10.1093/aepp/ppx056.

[28] M. J. Zaki and L. Wong, “Data mining techniques,” Publisher Springer Science+Business Media 2012.

[29] J. M. Jethva, N. Gondaliya, and V. Shah, “A Review on Data Mining Techniques for Fertilizer Recommendation,” vol. 3, no. 1, pp. 1386–1390, 2018.

[30] A. Mucherino, P. Papajorgji, and P. Pardalos, *Data mining in agriculture*, vol. 34. 2009.

[31] M. Kamber, M. Kaufmann, and P. All, "Note : This manuscript is based on a forthcoming book by Jiawei Han Jiawei Han and Micheline Kamber," 2000.

[32] B. Sanjeewa and R. Kalupahana, "An investigation into automated processes for generating focus maps," no. April, 2015.

[33] J. Swierzowicz, "Analysis of Current Data Mining Standards," pp. 764–766, 2003.

[34] A. Palmer, R. Jiménez, and E. Gervilla, "Data Mining : Machine Learning and Statistical Techniques," 2006.

[35] C. Priyadharsini, "An Improved Novel Index Measured Segmentation Based Imputation Algorithm for Missing Data Imputation," no. 6, pp. 283–286, 2017, doi: 10.23956/ijarcsse/V7I6/0217.

[36] S. Mulik, "Analysis of Crop Yield Prediction of Kharif & Rabi Jowar Crops Using Data Mining Techniques," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 7, no. 11, p. 79, 2017, doi: 10.23956/ijarcsse.v7i11.468.

[37] S. K. Devalkar, S. Seshadri, C. Ghosh, and A. Mathias, "Data Science Applications in Indian Agriculture," *Prod. Oper. Manag.*, vol. 27, no. 9, pp. 1701–1708, 2018, doi: 10.1111/poms.12834.

[38] D. Ramesh and B. V. Vardhan, "Data Mining Techniques and Applications to Agricultural Yield Data," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 2, no. 9, pp. 3477–3480, 2013, doi: http://dx.doi.org/10.5120/16620-6472.