Application of Data Mining Tools for Identifying Determinant Factors for Crop Productivity

Assefa Chekole
Department of Information Science
University of Gondar, Ethiopia

Tibebe Beshah, PhD
School of Information Science
Addis Ababa University, Ethiopia

ABSTRACT
Agriculture is the backbone of the Ethiopian economy and it contributes the highest GDP of the country. Among this, crop production takes the highest level of income for most smallholder farmers in all regions of Ethiopia.

The objective of this research is to build a model that can predict crops productivity and implement a decision support system. In order to conduct this research, a hybrid Knowledge Discovery Process model was adopted. For the purpose of this research, the datasets were taken from Central Statistical Agency of Ethiopia database, and the researcher used a total of 25,000 instances for training and building a model. Hence, for building a model and implementing decision support system for predicting crop productivity, WEKA data mining tool and java NetBeansIDE was used respectively. To achieve the objective of these research different experiments were conducted using J48, HoeffdingTree decision tree and PART rule based classifiers. In addition, the predictive performances of the classifiers are evaluated and compared using accuracy rate, confusion matrix and ROC curve. Based on this, out of the three classifiers PART rule based classifier performs best accuracy and ROC rate which is 95.44 % and 0.992 respectively. As a result PART rule based classifier were selected for implementing the model to predict crop productivity. In this thesis, the experimental result shows that, the main determinant factors for crop productivity are main season (season type), use of extension program, fertilizer used and fertilizer type. Therefore, the outcome of this research is essential to make data mining based decisions for policy makers and for experts in the area of crop agriculture to give an attention on the factors affecting crop productivity and to take corrective measures.

Keywords
Data mining, predictive model, decision support system, crop production, Ethiopia

1. INTRODUCTION
Data Mining (DM) is the process of analyzing data from different perspectives and summarizing it into useful information [1, 2]. 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 [1, 2]. 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 [1, 2].

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 [1, 2]. In addition, Data mining has paid attention to modeling as much as preprocessing and cleaning data to gain best results [3, 4].

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.

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 [5]. The major cereal crops which are mostly harvested by smallholder farmers are Teff, wheat, maize, sorghum, and barley.

Decision support systems (DSS) is an interactive computer-based systems intended to help decision makers utilize data and models in order to identify problems, solve problems and make decisions [6]. They incorporate both data and models and they are designed to assist decision makers in semi-structured and unstructured decision making processes. Also they provide support for decision making, they do not replace it. The goal of decision support systems is to improve effectiveness, rather than the efficiency of decisions [6].

The use of data mining to facilitate decision support can lead to an improved performance of decision making and can enable the tackling of new types of problems that have not been addressed before [7]. The integration of data mining and decision support can significantly improve current approaches and create new approaches to problem solving, by enabling the fusion of knowledge from experts and knowledge extracted from data [7].

In order to conduct this research, the researcher used crop production sample survey datasets acquired from central statistical agency (CSA) of Ethiopia. For that purpose, predictive data mining approach were employed.

Therefore, this study were addressed the following specific objectives:

- To conduct data preparation and preprocessing.
- To select algorithms for experimentation and modeling
- To develop and evaluate a prototype

1.2 Objectives

The main objective of this study were to identify the main determinant factors for crop productivity by using data mining tools and implementing a decision support system.

2. METHODOLOGY

In order to attain the stated objective that is building a model that can predict the production of crops productivity and to implement a decision support system, the researcher conducts datasets collection, selection of data mining tools, selection of algorithms that are undertaken for modeling and selection of data mining methodology that are adopted in this research.

2.1 Data source and datasets collection

Central statistical agency (CSA) of Ethiopia database is the main source of data for conducting this research. Because the agency collects various annual agricultural crops production and land use data each year during the two main cropping seasons i.e. meher and belg of Ethiopia. For that purpose 25,000 instances were selected from all regions of Ethiopia using stratified random sampling technique.

In addition, the datasets were extracted to Microsoft office excel worksheet for preparing the data and preprocessing with WEKA. Finally, it was converted to a standard and supported data format for WEKA data mining software i.e. saved with attribute relation file format (.ARFF) file type.

In order to develop a model that can predict crop production using data mining techniques based on crop production sample survey datasets acquired from CSA of Ethiopia database a hybrid Knowledge Discovery Process (KDP) model were adapted [8].

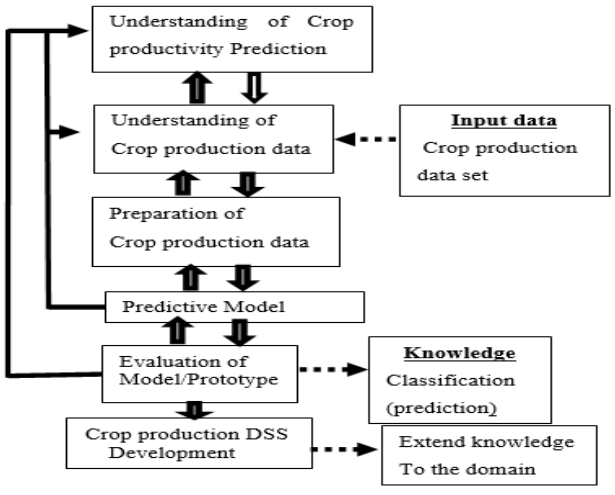

Figure 1: knowledge discovery process (KDP) model

2.2 Evaluation methods

In this research work, in order to measure the accuracy of the classifier the researcher were conducts the following testing options such as training and test set data, cross-validation and percentage split. In addition, the models used in this research were compared using performance evaluation metrics like accuracy, TP rate, TN rate, F-measure, ROC area and confusion matrix. The time taken by each classifier to build the models was also considered

3. RELATED WORKS

In this research, an attempt has been made to review different literatures and related works regarding, application of data mining for crop production prediction on the area of agriculture.

The research conducted on crop productivity mapping based on decision tree and Bayesian classification [9]. According to his study, Influence of climatic factors on major kharif and rabbi crops production in Bhopal District of Madhya Pradesh State was considered. Based on this finding the decision tree analysis indicated that the productivity of soybean crop was mostly influenced by comparative humidity followed by temperature and rainfall.

The study conducted on [10], which is Classification and Prediction of Future Weather by using Back Propagation Algorithm Approach. These were done using Neural Networks data mining techniques. Since, their study was focused on the information about weather and area observed and stored. As a result, according to their study the recorded parameters are used to forecast weather. Accordingly, in their finding they conclude that, if there is a change in any one of the recorded parameters like wind speed, wind direction, temperature, rainfall, humidity, then the upcoming climatic condition can be predicted using artificial neural networks, back propagation techniques.

In addition, another study on analysis of crop yields prediction using data mining techniques [11]. The data used for their study are obtained for the years from 1955 to 2009 for East Godavari district of Andhra Pradesh in India. In their study they analyze crop yield prediction using Multiple Linear Regression (MLR) technique and Density based clustering technique for the selected region. Accordingly, the main objective of the study was to create a user friendly interface for farmers, which gives the analysis of rice production based on available data.

Finally, the statistical model Multiple Linear Regression technique is applied on existing data. And the results obtained were verified and analyzed using the Data Mining technique namely Density-based clustering technique. Furthermore comparison between exact production and estimated values using multiple linear regression technique and density-based clustering techniques are discussed.

The study conducted on [12] building a predictive model for annual cereal crops production using data mining techniques. the main objective of the study was explore factors that determine crop productivity and build a predictive model for annual cereal crop production and identify important and interesting rules from the generated model by applying data mining techniques. The datasets for the study was taken from CSA, Ethiopia from two regions that are Oromia and Amhara agricultural survey of crop production of MEHER season. J48, REPTree and PART algorithms were applied for the experimentation in the study. Based on the finding improved seed, optimum fertilizer use and percentage of crop damage are the most determinant factors to predict annual major cereal crops production. Therefore, as per the findings of this study PART rule induction algorithm was performing best accuracy which is 82.37%.

Application of Data Mining Techniques for Crop Productivity Prediction [13], in their study the main objective was to assess predictive data mining applications that can be applied on Ethiopian agricultural crop productivity by focusing on small

holder farmers. The datasets used for their study was taken from the Ethiopian Economic Association (EEA). Additionally as a methodology for their study CRISP-DM was employed. And also, they used J48, Random Forest, and REPTree classification algorithms among them J48 algorithm has shown more predictive power. Based on their finding fertilizer use has the highest predictable power than the other factors. Additionally, they used small number of datasets when we compared to the present research and the datasets they used is only one year data which is 8540 records or instances.

4. RESULTS AND DISCUSSIONS

4.1 Experimental setting

The experiments were conducted using J48, PART and HoeffdingTree algorithms.

4.1.1 Experiment 1: Building a Model using J48 Decision Tree Classifier

```
=== Summary ===

Correctly Classified Instances       23599               94.396  %
Incorrectly Classified Instances      1401                5.604  %
Kappa statistic                          0.8185
Mean absolute error                      0.0569
Root mean squared error                  0.1686
Relative absolute error                 29.0459 %
Root relative squared error             53.8983 %
Coverage of cases (0.95 level)          99.968  %
Mean rel. region size (0.95 level)      46.4107 %
Total Number of Instances            25000
```

Figure 2: Summary of outputs from J48 classifier

Based on the above figure 2, the experiment result indicates that out of 25,000 instances of the datasets 23,599 instances/ cases were classified correctly and the model performs an accuracy of 94.396 %. And also the remaining 1401 records were wrongly classified. Additionally, the confusion matrix in this experiment shows that all 3,522 instances has beenpredicted as low, all 19,927 instances are predicted as medium and all 150 instances has been predicted as high.

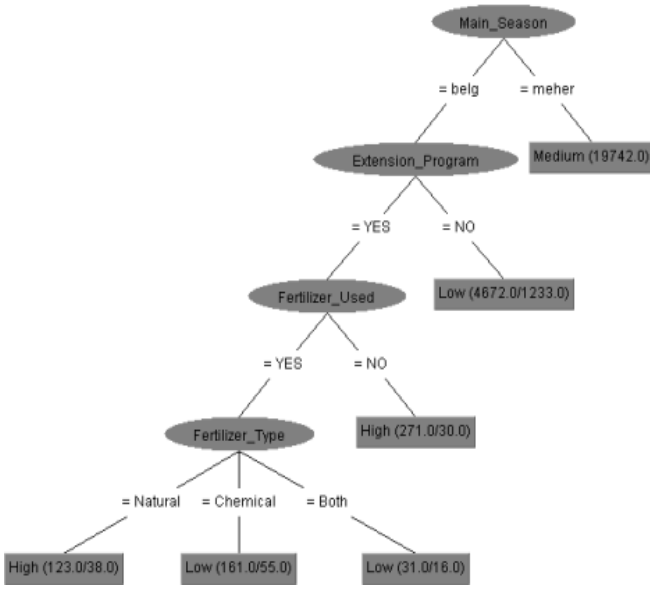

Figure 3: Visualized tree from J48 classifier

From the above figure, Crop Yield can be predicted with 94% of accuracy using the attribute Main Season which is determined as most relevant by J48 classifier. Rules that can be formed from the above visualized decision tree are:

- If Main Season = belg and extension program = Yes and fertilizer used = Yes and fertilizer type = Natural then Crop Yield = High.
- If Main Season = meher then Crop Yield = Medium.
- If Main Season = belg and extension program = No then Crop Yield = Low

4.1.2 Experiment 2: building a model using PART rule based classifier

```
=== Summary ===

Correctly Classified Instances       23860               95.44   %
Incorrectly Classified Instances      1140                4.56   %
Kappa statistic                          0.8486
Mean absolute error                      0.0436
Root mean squared error                  0.1475
Relative absolute error                 22.2923 %
Root relative squared error             47.1465 %
Coverage of cases (0.95 level)          99.884  %
Mean rel. region size (0.95 level)      41.7973 %
Total Number of Instances            25000
```

Figure 4: Summary of outputs from PART rule classifier

the experiment indicated with figure 4, above shows the result that out of 25,000 instances of the datasets 23,860 instances/ cases were classified correctly and the model performs an accuracy of 95.44 %. And also the remaining 1,140 records were wrongly classified. In addition, the confusion matrix in this experiment shows that all 3,386 instances has been predicted as low, all 20,156 instances are predicted as medium and all 318 instances has been predicted as high.

4.1.3 Experiment 3: Building a Model Using HoeffdingTree Algorithm

In order to measure the predictive performance of the HoeffdingTree classifier, Accuracy rate, TP Rate, FP Rate and confusion matrix performance metrics were employed.

```
=== Summary ===

Correctly Classified Instances       23348               93.392  %
Incorrectly Classified Instances      1652                6.608  %
Kappa statistic                          0.791
Mean absolute error                      0.0673
Root mean squared error                  0.1833
Relative absolute error                 34.3607 %
Root relative squared error             58.5927 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      47.3547 %
Total Number of Instances            25000
```

Figure 5: Summary of outputs from HoeffdingTree Classifier

The above figure 5, shows the result that can be obtained from HoeffdingTree classifier. So far, out of 25,000 instances of the datasets 23,348 instances/ cases were classified correctly and the model performs an accuracy of 93.392%. As well the remaining 1,652 records were wrongly classified these accounts 6.608 %. In addition, the confusion matrix in this

experiment shows that all 3,606 instances has been predicted as low, all 19,742 instances are predicted as medium and none of the instances has been predicted as high in this classifier.

4.2 Model Comparison

In order to achieve the objective of this research, the researcher conducted different experiments to compare the performance of classifiers. So far, the comparison was made in terms of accuracy rate which is calculated based on training and testing set prediction accuracy or performances. As a result, this helps to decide which model performs better accuracy and used to deploy the best model for decision making purpose. Here the comparison was conducted on the same datasets and the same evaluation metrics. Therefore, the experiment is conducted on a total of 25,000 instances and 17 selected attributes including the target or the predicted class.

Table 1. Performance comparison of classifiers

Classifier	Classified instances			
	Correctly classified	Prediction Accuracy in (%)	Incorrectly classified	Wrongly classified (error rate)
J48	23599	94.396	1401	5.604
PART	23860	95.44	1140	4.56
HoeffdingTree	23348	93.392	1652	6.608

The above table 1 shows, the accuracy of classifiers on crop production datasets. Therefore, the result shows that PART is the best classifier compared to the other which performs 95.44 % of accuracy, followed by J48 and HoeffdingTree they perform an accuracy of 94.396 % and 93.392% respectively.

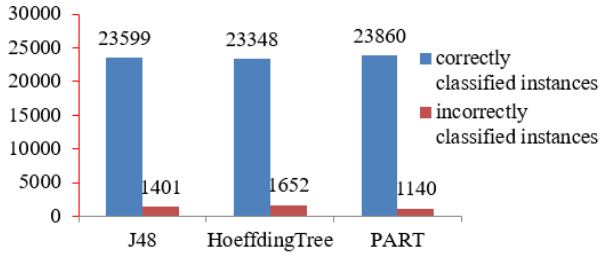

Figure 6: Classifiers error rate

Figure 6 above represents the classification errors of the classifiers, correct classified instances and incorrect classified instances of the training datasets employed in experimentation. Therefore, PART shows best performance compared to other classifiers and it is also employed on crop production datasets for predicting crops productivity.

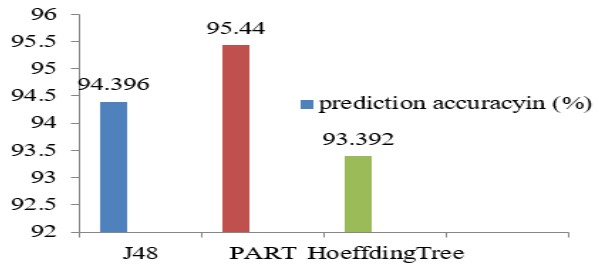

Figure 7: prediction Accuracy of Classifiers on Crop production datasets

The above figure 7, shows that PART rule based classifier performs best accuracy compared to other classifiers used in this research.

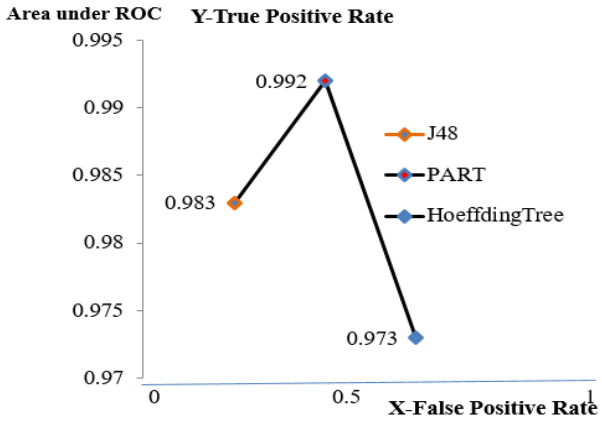

Figure 8: Plot area under ROC Values of the Classifiers/Models

The above figure 8, shows that the values of the ROC area which is obtained from J48, PART and HoeffdingTree classifiers. Additionally the figure shows that the Y-axis indicates the value of true positive rate that the model has better predictive performance and the X-axis indicates that the false positive rate in which the model is worse to predict the dataset. Therefore, PART model is better than the other since it scores the ROC values 0.992 which is slightly close to 1 which means that when the model performs the highest value and the graph approached to the vertical line that shows the model is the best predictive performance.

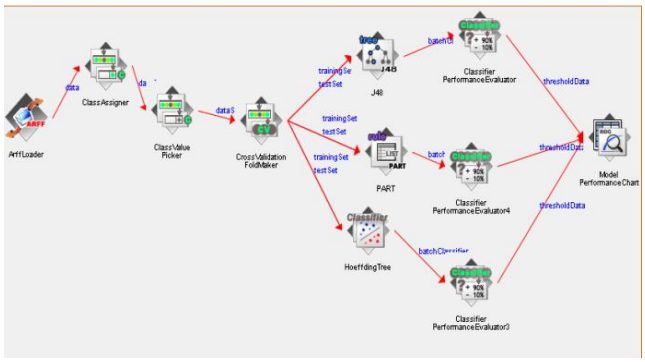

Figure 9: knowledge flow layout

The above figure 9, shows performance Evaluation of models/ classifiers Using Knowledge Flow Configuration Environment. To conduct this evaluation different components are interconnected components like Arff data Loader, class assigner, class value picker, cross validation folder maker, classifier performance evaluator and the last component model performance chart. In addition, component connectors like data sets, training set and test set, batch classifier and threshold data which are used to connect the components. As a result the outcome of the evaluation is used to select the best model.

4.3 Crop Production DSS Development

The discovered knowledge with this research was applied to agricultural crop production for agriculture domain. As well, in order to implement the discovered knowledge on the domain which is used for decision making, the researcher develops a prototype or user interface using JAVA programming language integrated with WEKA workbench. Finally, the discovered knowledge is deployed on CSA of Ethiopia server and integrated with agriculture office.

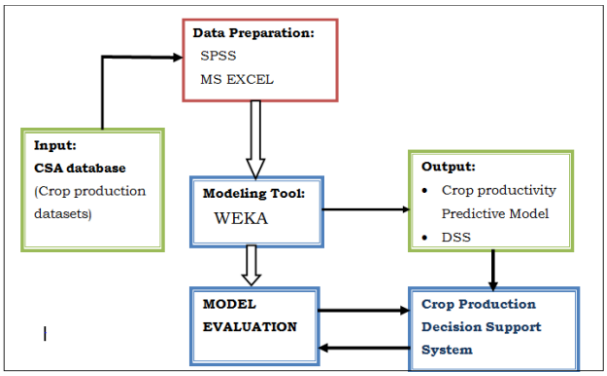

Figure 10: proposed design for crop productivity prediction

4.4 Prototype for Crop Production Prediction

In this context, the farmers necessarily require a timely advice to predict the future crop productivity and an analysis is to be made in order to help the farmers to maximize the crop production. This prototype was mainly used for experts and decision makers for identifying the determinant factors for cropping and predicting productivity. The user interface is designed with NetBeans using WEKA Model results, so as the selected attributes are used to design the user interface form that helps for predicting crops productivity.

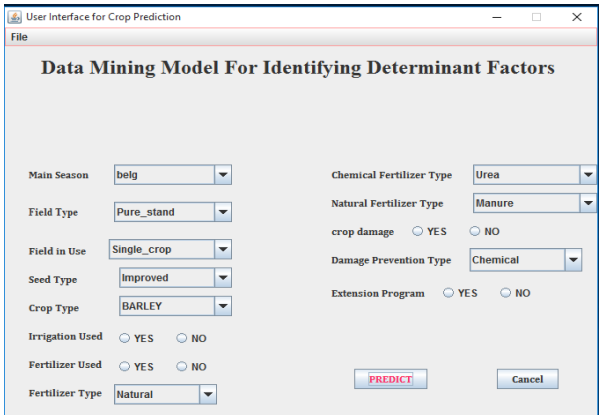

Figure 11: Model for Identifying Determinant Factors

Based on the experiments conducted, PART rule based classifier were selected for implementing the model or for deployment. So far, out of the total selected attributes the determinant factors for crop productivity are main season (season type), use of extension program, fertilizer used and fertilizer type. Finally, the present study is better than the previous studies with datasets (it uses two season datasets nationwide), with the no of instances used, In terms of model performance comparison techniques, model performance or accuracy and implementing of Decision Support System. Therefore, to achieve this research objective, the researcher implemented prototype developed with java programming language integrated with WEKA workbench for predicting the levels of crops productivity. On the other hand, the results of this research were used for Implementing data mining based decision support for policy and decision making purpose.

5. CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

Data mining is a process of extracting relevant information from large database/ data warehouse for organizational use.

Nowadays, application of data mining technology has been applied in different business issues and health services for medical diagnosis and patient treatment. Additionally, it also applied in the area of agriculture to analyze and predict farming outputs and used to forecast metrological weather conditions using different data mining techniques. Among the two major goals of data mining technology, this research followed a prediction approach to build a model and predicting crops productivity using crop production datasets.

The main purpose of this study were to identify the main determinant factors for crop productivity using data mining tools. For that purpose, the hybrid data mining modeling approach were applied. In order to achieve the objective of this research, decision tree and rule based classifier techniques were employed. In addition the researcher conducts data preprocessing techniques like data cleaning, missing values handling and data selection in order to improve the efficiency of the algorithms to classify the data correctly.

In order to build a predictive model for predicting crop productivity, three classification algorithms are applied on crop production datasets and their performances are evaluated using accuracy rate and ROC curve. Based on this reason, PART rule based classifier is the best model for predicting crops productivity. In this research, the researcher find out, Main season, extension program, fertilizer used and fertilizer type are the most determinant factors that affect crops productivity.

For 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.

5.2 Recommendations

This research was focused on build a model that can predict crop productivity using existing data collected by CSA of Ethiopia. Based on this, the main objective of the study was achieved. Accordingly, based on the challenges we faced during the progress of this research work, the researcher recommends the following for the domain knowledge i.e. for agriculture office:

For the purpose of this research work, the data was collected from CSA database, due to the reason that the required data was not obtained from agriculture office. So far, it is better when the data should be collected and organized on the hand of the agriculture office.

The rules extracted from this research shows that there are different factors affecting crops productivity. As a result, the government should give an attention on the implementation of these determinant factors for proper utilization.

The discovered knowledge was deployed on the Central Statistical Agency and applied or used by agriculture office of Ethiopia. The government uses this Decision support system as a support for decision makers that help to enhance crops productivity.

The rules extracted from this classification technique were used for implementing knowledge base to the agriculture domain. Rather than the traditional statistical prediction of crop production using this model is better for predicting the levels of crops productivity.

5.3 Future Research Directions

As per the findings of this study, further investigation was required to identify the relevant features from metrological or weather condition data and mixing the determinant factors that affects crops productivity.

Since, this research is focused on identifying the determinant factors that affect the production of cereal crops and building a predictive model used as a decision support system. So far, additional research is needed to be done for other crop type. Lastly, further investigation is required to improve the classification accuracy of the models to reduce classification errors.

5.4 Acknowledgments

When doing this research many people have contributed their supportive ideas. So, would like to thanks to all of them. Additionally, would like to express sincere gratitude to University of Gondar for the financial support and the excellent facilities to complete this research.

6. REFERENCES

[1] K. M, Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons Inc., 2003.

[2] H. J and K. M, Data mining, concepts and techniques, 2nd ed. Morgan Kaufmann Pub, 2006.

[3] Z. Yan-li and Z. Jia, "Research on Data Preprocessing In Credit Card Consuming Behavior Mining", Energy Procedia, vol. 17, pp. 638-643, 2012.

[4] L. Xiang-wei and Q. Yian-fang, "A Data Preprocessing Algorithm for Classification Model Based On Rough Sets", Physics Procedia, vol. 25, pp. 2025-2029, 2012.

[5] A. Seyoum, P. Dorosh and S. Asrat, "Crop Production in Ethiopia: Regional Patterns and Trends", Ethiopian development research institute, 2011. [Online]. Available: http://reliefweb.int/sites/reliefweb.int/files/resources/essprn11 pdf. [Accessed: 09- Jan 2016].

[6] D. ladeni , Data mining and decision support. Boston: Kluwer Academic ublishers, 2003.

[7] D. ladeni , Data mining and decision support. Boston: Kluwer Academic Publishers, 2003.

[8] K. Cios, L. Kurgan, W. Pedrycz and R. Swiniarski, Data mining. New York, NY: Springer, 2007.

[9] V. S, "Crop productivity mapping based on decision tree and Bayesian classification", 62 Unpublished, 2007.

[10] S. Sawaitul, P. Wagh and D. Chatur, "Classification and Prediction of Future Weather by using Back Propagation Algorithm- An Approach", International Journal of EmergingTechnology and Advanced Engineering, vol. 2, no. 1, 2012

[11] D. . and B. Vardhan, "ANALYSIS OF CROP YIELD PREDICTION USING DATA MINING TECHNIQUES", International Journal of Research in Engineering and Technology, vol. 04, no. 01, pp. 470-473, 2015.

[12] T. Bekele, "building a predictive model for annual cereal crops production using data mining techniques", 2014

[13] Z. Diriba and B. Borena, "Application of Data Mining Techniques for Crop Productivity Prediction", HiLCoE Journal of Computer Science and Technology, vol. 1, no. 2.