Decision Supporting System for Enhancing Upcountry Vegetable Cultivation in Sri Lanka

U. R. Samaratunge

129168 K

Faculty of Information Technology

University of Moratuwa

March 2015
February 2015

Decision Supporting System for Enhancing Upcountry Vegetable Cultivation in Sri Lanka

U. R. Samaratunge

Index Number : 129168 K

Dissertation submitted to the Faculty of Information Technology,
University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of
the M. Sc. in Information Technology.

March 2015

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

U.R. Samaratunge
Name of Student   Signature of Student

Date:

Supervised by:

Mr. S.C. Premaratne
Name of Supervisor   Signature of Supervisor

Date:

i

Acknowledgements

From the initiation there were many people behind me to support and encourage the research, development and the evaluation of this project. This research on agro based data mining and content search and the implemented tool for upcountry vegetables was not an isolated achievement of just me but an outcome of many people who helped me in many other ways and means.

First and foremost the Dean of the faculty of Information Technology, Department Heads, the kind and friendly Lecturers and Instructors were a big help for me to successfully complete this project. I must also thank my Supervisor for being very helpful and inspective throughout the project and helping me every time I was in trouble. Also there were many friends and batch mates and friends who helped me to complete this project both from their knowledge and encouragement.

.

Last but not least I also thank the editors of the various contents I used to gain knowledge to complete this project. Without their knowledge and generous thoughts I would have found it much difficult to find knowledge this easy and complete this task this much comprehensively. So, all in all I would like to thank all the people who helped this attempt become fruitful in many ways and means.

Abstract

Agriculture is considered to be the most widely spread and most commonly practiced form of employment in Sri Lanka. The Sri Lankan society has been dependent on agriculture since the beginning of its early civilization. Agriculture is the most important sector of the economy which fulfills the basic food needs of the citizens. This project considers upcountry vegetables; particularly selected four crops of which the information related to these are so important for development of this sector. The areas where upcountry vegetables have been cultivated fall into different agro-ecological zones varying from rain fall, elevation, land type, height of the top soil, slope of the land and soil erosion etc. The potential of information technology (IT) can be assessed as a tool for direct contribution to agricultural productivity and as an indirect tool for empowering farmers to take informed and quality decisions which will have positive impact on the agricultural activities. In this project, it is attempted to build an indirect tool for all interested parties including farmers aiming to develop an information system to that facilitates accurate and reliable forecasting through data mining to enhance upcountry vegetable cultivation in Sri Lanka. This information system going to use two sets of data; particularly one from Data Bank of Hector Kobbekaduwa Agrarian Research and Training Institute and the other from a sample survey conducted by this tools web interface developed using a content management system and using the sample data that recorded harvest will be forecast through data mining and creating appropriate modules and finally these modules will be incorporated in to the information system. The final phase of the project is the evaluation of the web-based information system with several different data mining algorithms. The Project has successfully implemented dynamic filtering, classification and prediction of data using the latest technologies and tools available in the field of IT to uplift the agricultural field in Sri Lanka and the evaluation of its performance looks promising for practical implementation.

iii

Contents

Chapter 9 – Conclusion & Further Work

List of Figures

List of Tables

Introduction

1.1 Introduction

The Sri Lankan society has been dependent on agriculture since the beginning of its early civilization. Agriculture is the most important sector of the economy which fulfills the basic food needs of the citizens. It is therefore essential that the technology thrust should lay greater emphasis on the transfer of scientific and technological information from the sources of information to its actual users. Even though agricultural contribution to the gross domestic product declined substantially during the past three decades it is the most important source of employment for the majority of the Sri Lankan workforce. Sri Lanka's main goal is to achieve an equitable sustainable agricultural development through development and dissemination of improved agricultural technology.

This project is to consider upcountry vegetables, particularly selected four crops of which the information related to these are so important for development of this sector. The areas where upcountry vegetables have been cultivated fall into different agro-ecological zones varying from rain fall, elevation, land type, height of the top soil, slope of the land and soil erosion etc. Due to this situation, it can be observed that farmers may harvest different quantities from areas with similar extent even though they practiced similar process in cultivating of vegetables. The farmers engaged in this sector are really interested to be aware of price fluctuations, annual requirement and contribution from major cultivation areas, future trends, land utilization and how much they could produce individually for a crop.

Although the surveys related to agriculture information are conducted at different locations and levels by many researchers, government and non-government institutions, unfortunately farmers are reluctant to use this information due to various reasons. The annual vegetable demand is rapidly increasing and during the period of last ten years. The demand for potatoes, Beans and Carrot has increased by 14.49, 6.84 and 56.98 percent respectively and no such an increase is shown for Cabbage compared to the year 2000 [1] . Even though the

production of these crops has also increased by 7.28, 23.22 and 31.70 and 17.62 percent, it is not sufficient for the fulfillment of the requirement at all except for Cabbage. Since the major cultivating areas fall in to districts such as Matale, Nuwara-Eliya, Badulla and Kandy, expanding the land area for cultivating is limited and therefore, utilization of land is a very important factor to be considered. According to the review information, there is a significant difference in average vegetable production per hectare between these areas. Since the form of a land and weather are main issues relating to varying production, utilization of land with all possible measures would be very much important for the enhancing average production. In the other hand, awareness about farming and related activities is another issue that has to be implemented. Moreover, in the context of agriculture, the potential of information technology (IT) can be assessed as a tool for direct contribution to agricultural productivity and as an indirect tool for empowering farmers to take informed and quality decisions which will have positive impact on the agricultural activities. In this project, it is attempted to build an indirect tool for all interested parties including farmers.

1.2 Background and Motivation

As a developing country in the South-Asian region where agriculture is the main way of living and income in Sri Lankan economy and approximately 33% of the total labor force is engaged in agriculture. Production of rice is one of the most important activities and cultivating of other cereals, vegetables and fruits are also prominent in the economy of Sri Lanka. Therefore it is very important to analyze the patterns or deviations of the agriculture with currently available information for the purpose of formulating policies and also to enhance the knowledge of farmers to encourage them to continue to be in the agriculture. In this context, production, price fluctuations, seasonal indexes, requirement and the factors could affect these variables are highly important. With the latest information technology (IT), although the government bodies disseminate timely information on their publications and web sites, still there is a lack of information for the purpose of decision making.

Data bank of Hector Kobbekaduwa agrarian research and training institute is a national center for secondary source of information which has been collecting primarily collected information from many government institutions since 50's. Currently there is a huge volume of agricultural information in the institute data bank. In this project, for this selected four upcountry vegetables namely potatoes, beans, carrot and cabbage an attempt is made to construct an information system, which can be used for effective decision making on expected harvest of selected crops with the utilization of lands according to Agro-ecological measures and also on seasonal indexes and time series forecasting to educate to solve the questions that had farmers, consumers and other interested parties as well.

Thus the motivation to start this project primarily comes from this huge potential that is brought in to light with availability of collected data over the years from the institute and the cutting edge information technology advancements such as data mining and data warehousing.

1.3 Aim and objectives

Aim

The aim of this project is to develop an information system that facilitates accurate and reliable forecasting of harvests for individual farmers through data mining to enhance upcountry vegetable cultivation in Sri Lanka.

Objectives

- To Study on the problem domain, to propose a solution based on Data mining and to forecast very veritable information to deliver to the user.
- To Study data mining, web technologies that support the implementation of it.
- To design and develop an information system to incorporate above technologies.
- Evaluate the System.
- Prepare the final Report with evaluation results.

1.4 Purpose of the document

This report mainly delivers the information about the general idea and progress of this project. The project mainly focuses on making predictions for specially selected four vegetables using data from government collected sources and surveys which would help farmers for their agricultural activities. This document contains the initial ideas, others work and the progress of the project through methodology and implementations and their evaluations. The various technologies and methodologies used are described in the technologies adapted chapter. Furthermore, the document provides a list of references, and gives an overview of our findings and a brief description of the implementation of the tool using appropriate diagrams in the Implementation chapter. Finally the progress so far and work to be completed are disused in discussion and work to be completed chapters. By referring to this document it is possible to have a complete understanding of the project.

1.5 Structure of the thesis

This chapter consists of the following parts. The introduction to the project, background and motivation to it which shows the importance of the problem with the support from high potential of the subject area are described first. Then the aim and objectives of the project are under a separate heading and in the last section, briefly stated are the work outputs and the structure of the report. The next chapter will be discussing the various other approaches followed by other parties on similar topics

Literature Review

2.1 Introduction

The preceding chapter described background information related to the project including the aim and objectives and the structure of this report. This chapter will discuss about some information of the completed research work on the field of study of this project and some tools developed by various parties.

2.2 A Prediction Model Framework for Crop Yield Prediction

This is a study done over the yield of corn on a selected area and in it the researchers have selected two types of variables namely climatic variables such as rainfall, temperature, solar radiation, humidity, and data about weather disturbance and corn production related fourteen agronomic variables such as soil type, seed type and plant spacing. The data has been collected from different organizations one being weather bureau of the state, while agro-related data has been collected from Bureau of Agricultural Statistics (BAS) and the Office of Provincial Agriculturist (OPA).

The problem they have addresses have been described as [3] " *need to develop an crop prediction model framework that can (1) pre-process and fuse potential predictor raw data from multiple sources, (2) provide an accurate prediction of crop yield, (3) identify significant variables that affect crop yield, and (4) learn useful prediction policies for decision planners, particularly for Provincial Agriculturists.*"

The methodology that has been used to do this study can be explained as below. Initially the selection of inputs consisting of selected two types of climatic variables such as rainfall, temperature and data about weather disturbance and corn production related fourteen agronomic variables such as soil type, and seed type has been done.

Then the data preprocessing has taken place through data cleaning, transformation and data discretization. The classification of the data is mainly achieved through rule induction and attribute selection, evaluating using specific algorithms such as JRip Algorithm. For the study of corn the following figure shows a comparison of their models for the task.

Performance Criteria	MODEL 1 (Full Attribute Set)	MODEL 2 (Reduced Attribute Set)
Correctly Classified Instances	89.7541%	89.0369%
Incorrectly Classified Instances	10.2459%	10.9631%
Kappa statistic	0.6332	0.6138
MAE	0.0986	0.1061
RMSE	0.2371	0.2438
RAE	49.886%	53.6932%
RRSE	75.5908%	77.7198%
Time to build the model	0.42 seconds	0.73 seconds

Table 2.1 - Performance Evaluation of the Models [3]

2.3 Early Prediction of Crop Yield

In the study neural networks have been used to evaluate and predict the harvest of fruits from one weak ahead up to eight weeks ahead. The dataset used here includes environmental data such as Radiation Level, Day Temperature, Night Temperature, Average 24h Temperature, Average 24h Humidity, Day / Night Humidity levels , CO_2 Levels Inside green houses and Physiological data like New Fruit Set. The datasets used were obtained through 5 years starting from 2007 to 2011[4].

6

Artificial models used in this study were constructed using the software Matlab version R2010a and Back-propagation algorithm, one of the most popular learning methods under neural networks, has been used to do predictions.

Five types of prediction models has been done in the study such as Predicting the expected yield of the current week through this week's environmental data, Predicting expected yield of the following week using environmental data of this week, Predicting expected yield of following week using this week's environmental data added no. of new fruit sets six weeks before the current week, predicting of the expected yield of week 8 using current week data and new fruit set in current week and finally creating NN (Neural Network) model to that is able to update the users weekly to increase reliability of its prediction in the expected yield on eighth week.
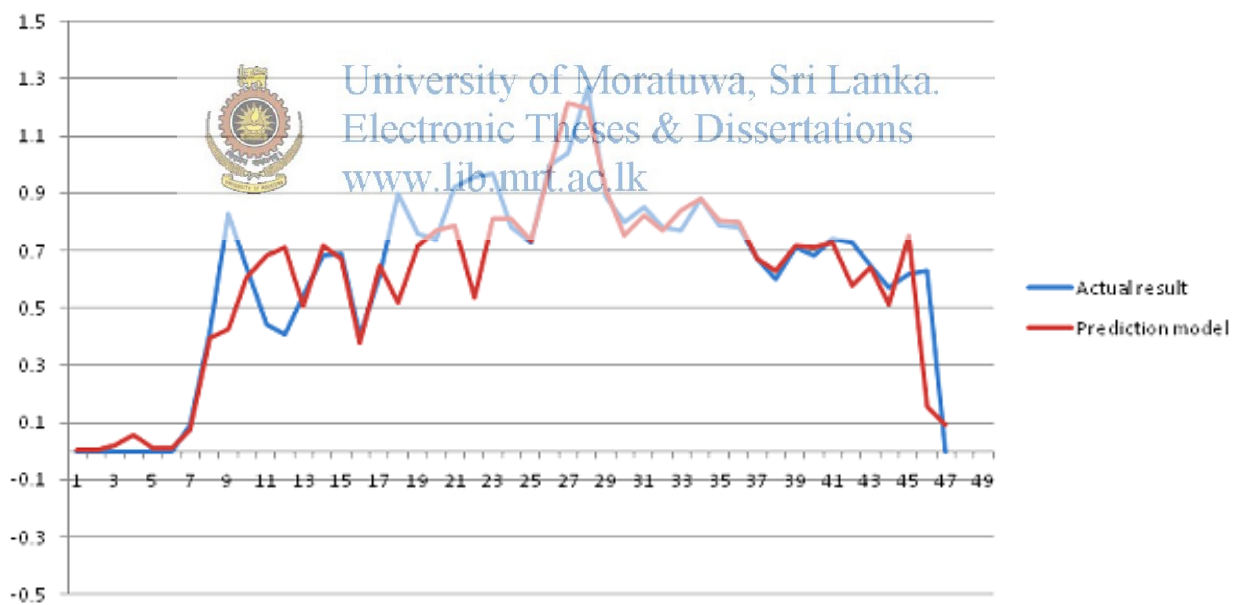

Figure 2.1 - Actual vs. Predicted Data [4]

(Current week data- next week yield prediction)

2.4 Operational Prediction of crop yields using MODIS

This is a study which uses MODIS algorithm to predict crop yields in the United States and this algorithm includes features like crop-specific classification and yield prediction before crop harvest. In the initial phase, MODIS data and products for states of Illinois and Iowa has been used to develop an operational assessment of crop yield predictions specialized for corn and soybeans [5].
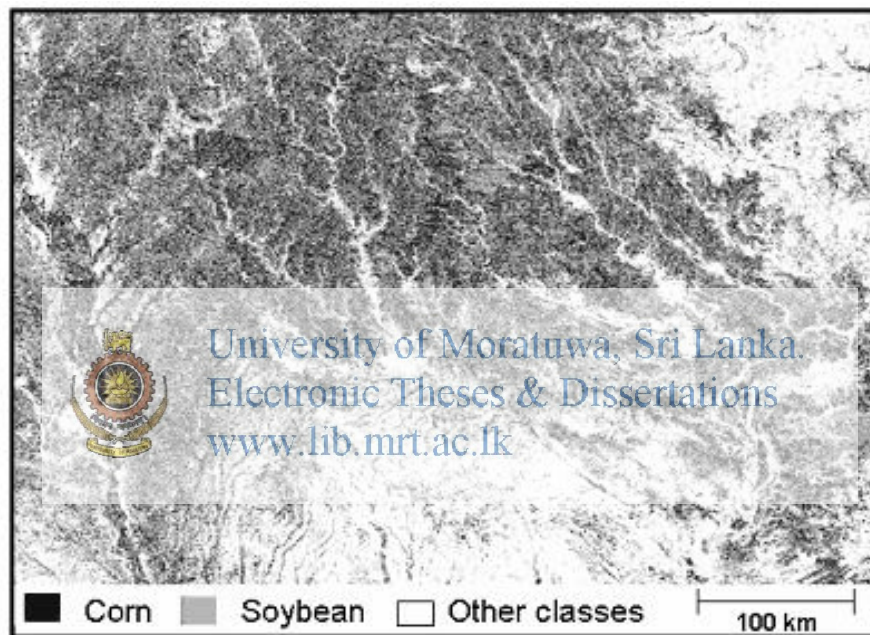
Figure 2.2 – Classification for Iowa with the 8-day MODIS 250 m composite imagery [5]

The methodology for the study includes the study of the given geographical area and development of a plan to do data collection and re-processing and afterwards crop classification and forecasting is given priority. The MODIS data acquired for research had data of 5 crop seasons of 2002 to 2006. The results from the research are compared with USDA official estimates from NASS.

The final results of the study were able to predict yields are all within 20% of the NASS estimates, and the better majority is within 10%. RMSE for predicted yield for corn and soybeans in Iowa are 10.1b/ac and 3.6 b/ac and in Illinois are 19.3 b/ac and 5.6 b/ac respectively.

2.5 Summary

This chapter described about various approaches other people have tried and tested in the field of data mining to make predictions accurately for selected crops. It includes the details of these different approaches and what technologies they used in achieving their goals. This chapter also summarizes the weaknesses and incompleteness's of the approaches considered above. In the next chapter the adopted technologies will be taken in to consideration.
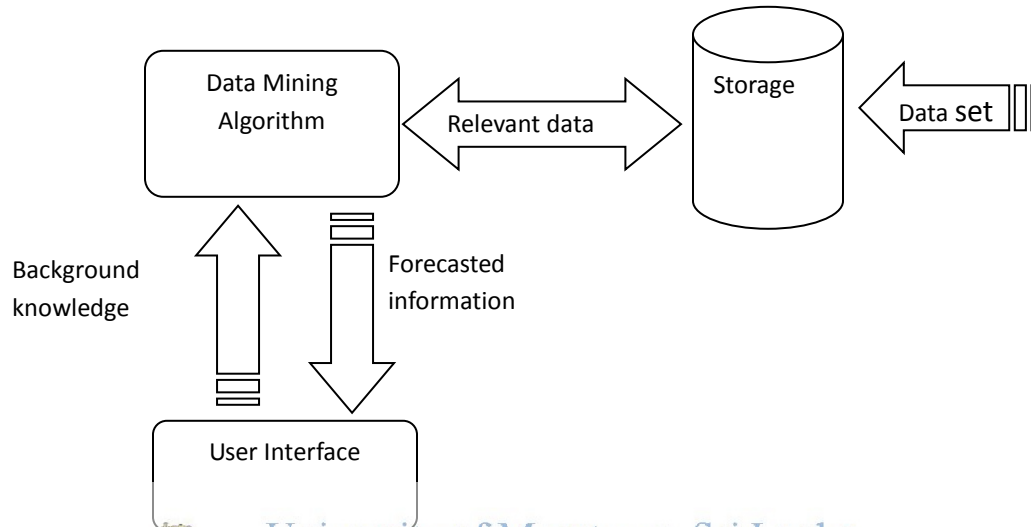
Methodology

3.1 Introduction

The previous chapter explained about various other attempts that were used to do mining and extraction of data to use for predictions. This chapter describes the all process of project activities being carried out. In the initial stage of the project, available information pertaining to selected crops was compiled. In this information system, the data set of vegetable crops information for the period of 10 years duration commencing from 2000 from Data Bank of Hector Kobbekaduwa Agrarian Research and Training Institute was collected and stored in the " .arff " format. Furthermore this section of the report describes about the approach taken to complete the project using the technologies mentioned in the previous chapters. Also it gives an outline idea of what kind of inputs, outputs and processes will be used for the process of the system that will ultimately be developed.

3.1 Outline of the Methodology

In the agricultural sector, there is growing need for information exchange at the national, regional, grass root level. This requires bridging the information gap and brings improved Information and Communication Technologies (ICT) driven solutions to forecast the agriculture sector information system. The proposed solution is a web-based information system and in this solution, communication has been considered as the key and ICT are being used effectively and efficiently for the overcome the existing situation. The main goal of this approach is to provide agricultural and related information on upcountry vegetable

cultivation and the system will be constructed to get the information of expected harvest of any selected crops based on the measures considered here. The system could be used not only by decision makers but also by farmers and other interested parties as an information system for the purpose of enhancing vegetable production.

Figure 3.1 – Architectural Design of proposed System

3.3 Methodology for Implementing the System

This information system going to use two sets of data; particularly one from Data Bank of Hector Kobbekaduwa Agrarian Research and Training Institute and the other from a sample survey conducted by this institution would be stored in a SQL database. The data bank has data from several sources mainly from department of census and statistics, central bank and department of agriculture and also from other reliable sources. The monthly prices of selected crops for the period of last ten years would be collected into a SQL Database for generating seasonal indexes and forecasting modules. To predict the harvest, data from a

sample survey would be formulated. For this purpose conduct a sample survey among over 200 farmers who were engaged in cultivating of upcountry vegetables; covering several Agro-ecological zones particularly in Matale, Kandy, Nuwara-Eliya and Badulla districts will be considered. This questionnaire-based data set and consists of demographic information, pattern of vegetable cultivation, land utilization, conservation of lands, farmer's knowledge regarding Agro-ecological problems, use of protection measures relating to vegetable farming, use of fertilizer , awareness on soil acidity and other cultivation methods aimed at protected farming systems and knowledge regarding environmental conservation. It would be basically considered that the vegetable production could be a function of the measures coming under following broadly classified groups.

- Demographic characteristics of the farmers
- Awareness/Knowledge regarding farming and technologies
- Location by district wise
- Crop establishment

Using the sample data that recorded will be forecast through data mining and creating appropriate modules and finally these modules will be incorporated in to the information system. The final phase of the project is development of the web-based information system with the content management tool 'Joomla'.

3.4 Summary

This section described about the approach expected to be followed within the scope of the project and discussed of the procedure that is to be followed and the figures showed the flow of the system being developed. In the next chapter the design and the analysis of technologies will be discussed

Data Analysis and System Design

4.1 Introduction

The previous chapter described about the approach towards researching and implementing the findings as a prediction tool. This chapter will discuss and illustrate the way data is analyzed and how the findings of different types are utilized for the tool design. Furthermore this chapter examines the conditions of the sample population with special focus on attributes such as demographic pattern, Ageing, Level of education, Primary occupation and income of the sample household.

4.2 Analysis of the Optimal Technologies

In order to develop the tool from the research knowledge of data mining and warehousing technologies gained, it is needed to define a set of optimal technologies and techniques for each component and each procedure. For an example which type of classifier to be used in classifying the dataset etc.

Initially it was required to see how to extract the needed information from the Data Bank of Hector Kobbekaduwa Agrarian Research and Training Institute can be pre-processed and used for the task of revealing patterns associated with each type of vegetable and its associated attributes. There is a need to evaluate what is the optimal way to classify the data sets and the task is achieved by evaluating several classifiers on the same data set.

Finally to interact with the users of the system to get inputs and to provide detailed

predictions based on the inputs to the system the role of the content management system is essential and Joomla was selected as the best option for its ease of use and computability with number of database types. So after much consideration the optimal set of tools and technologies were selected for the implementation of the forecasting information system for enhancing upcountry vegetable cultivation in Sri Lanka.

4.3 Architecture of the Prediction Tool

The Figure 4.1 shows the design of the tool that implemented through this project. It consists of the following main modules.

1. Training Module

2. Prediction Module

3. User Interface through Content Management System
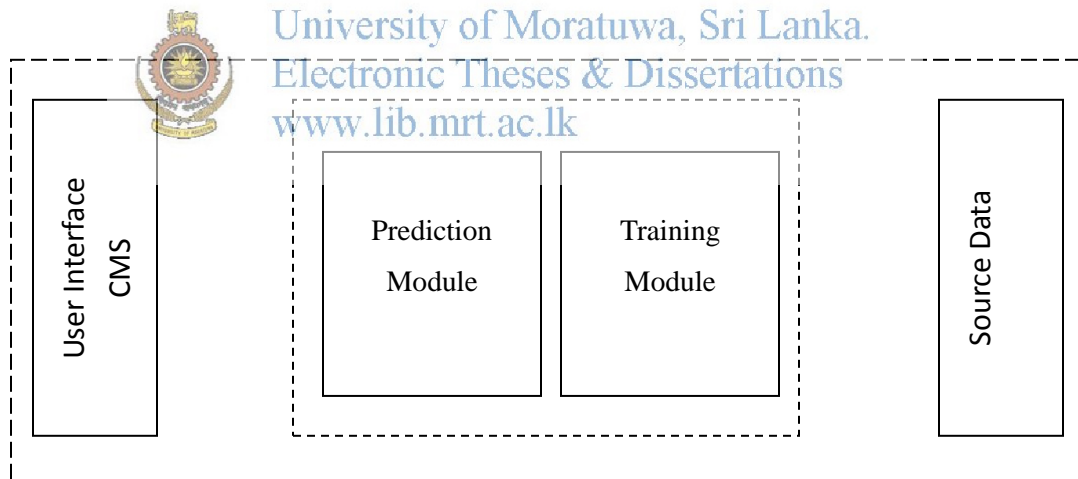
Figure 4.1 – Component diagram of the System

4.3.1 Training Module

This is the module responsible for collecting only the necessary data from the data bank and storing them on the system after extracting the necessary fields only and preprocessing the

data. The classification will also be done through this module for classifiers such as decision trees. The module mainly consists of the following layers.

1. Extraction Layer

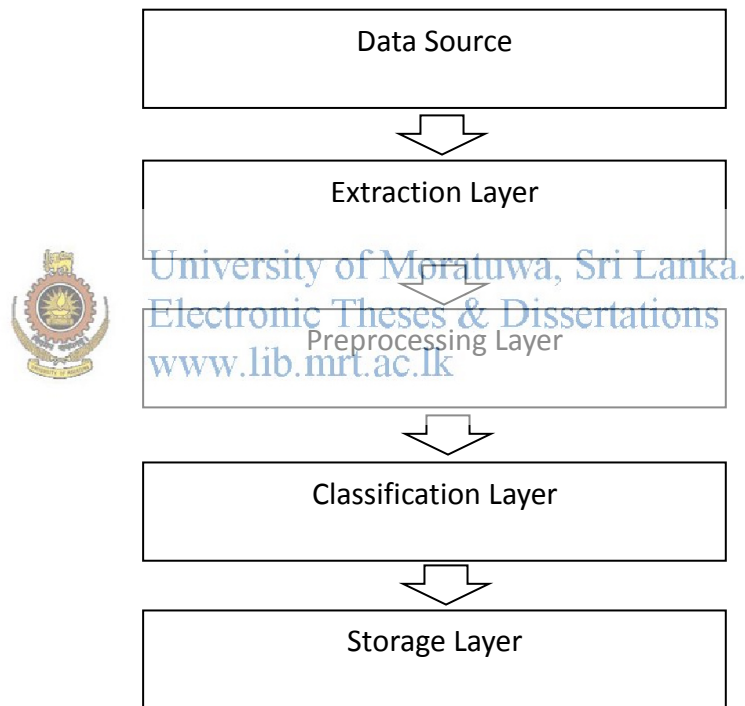2. Preprocessing Layer

3. Classification Layer

4. Storage Layer

Figure 4.2 – Layered Architecture of the Training Module

As shown in the above diagram, first the data goes through the data store and then consecutively through extraction, preprocessing and classification processes. The output of this module will reside in the Storage layer in the form of classified data.

The data store represents where the raw data for the system resides. In this instance it will be

the data bank obtained findings and those will go through the extraction and preprocessing layers to be classified and stored as labeled and classified data. For an instance if the mining algorithm is a decision tree the generated tree will be finally on the storage.

4.3.2 Prediction Module

The main task performed by the prediction module is obtaining the user input to the system through a web interface and providing the necessary and predicted outputs of each correct input. The presentation layer mainly consists of the content management system and the This module consists of the following layers.
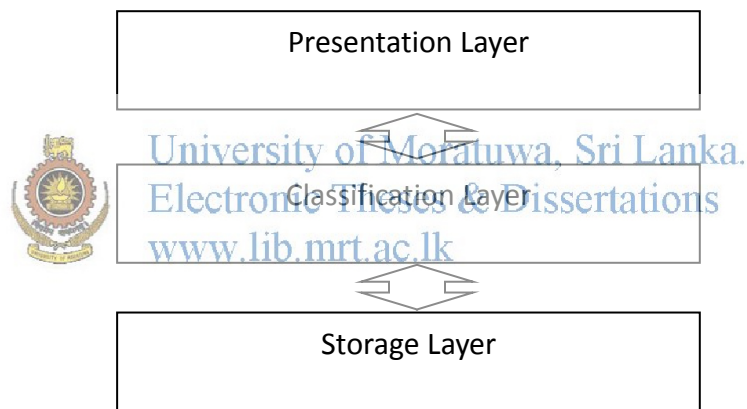
Figure 4.3 – Layered Architecture of the Prediction Module

This module will take inputs from farmers with different demographic information, pattern of vegetable cultivation, land utilization, conservation of lands, farmer's knowledge regarding Agro-ecological problems, use of protection measures relating to vegetable farming, use of fertilizer , awareness on soil acidity and other cultivation methods and find a methodology using the most suitable data mining technologies to predict results.

4.3.3 User Interface through Content Management System

The main tools that were used to develop the client website of the project were Joomla (3.0) and some of its associated plugins. The main reason to choose Joomla was its easy to use and fast responsive features and a wide variety of plugins to easily implement custom features that on the other hand would consume much time developing testing and debugging.

The purpose of adding the plugins for the Joomla based web interface was to add functionality such as dynamic forms and Graphs. The result generation is aided by these plugins as well as Joomlas user friendly and neat design. The following is a overview of the architecture of the web interface.
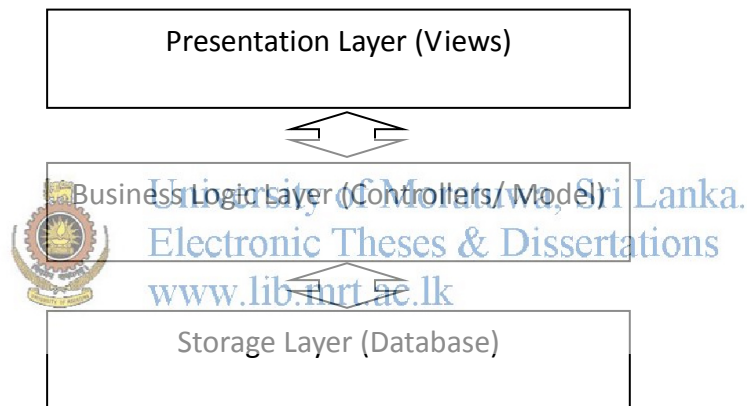
Figure 4.4 – Layered Architecture of the Prediction Module

The database associated with the Web interface is cropinfo_web (See AppindixD) and it contains data tables in order to store the information related to the website and its related data. Most of these tables are managed by Joomla itself and can be accessed through Joomlas configuration panel through the admin login.

The Business Logic Layer is also inbuilt to the Joomla CMS and some of the functionalities such as forms and dynamic menus and banners are managed by using external plugins. Mainly these plugins are used to validate forms and add graph drawing facility to the final

crop prediction result generation section. These plugins also access the storage layer of the web site through Joomla.

The presentation layer was handled also using Joomla and a theme package which was available for Joomla was used as the main theme for the entire website without manually editing or writing CSS and HTML pages. This helps to make the Website to look more professional and easy to manipulate and debug.

4.4 Summary

This chapter presented the architectural design, which will be used to implement the "Forecasting information system for enhancing upcountry vegetable cultivation in Sri Lanka" application. In the next chapter the implementation phase of the project will be discussed with more detail.

Requirement Analysis

5.1 Introduction

The objectives of this chapter are to analyze and describe software system requirements and also express them in meaningful way. The term requirement is not used in software industry in a consistent way. It is now widely understood that successful systems and products begin with an understanding of the needs and requirements of the users. A fundamental challenge for technology-oriented enterprises is to meet customer and/or user expectations, when developing applications and services. Due to strong competition, a successful approach should be the design of applications on the basis of the current state-of-the-art. In some cases, a requirement is simply a high level abstract statement of a service that the system should provide. In this project the requirements were expressed in terms of user, system, functional and non-functional requirements.

5.2 User Requirements

In this direction, user requirements analysis is the foundation of a user-centered approach. Understanding user requirement is an integral part of information systems design and is critical to the success of interactive systems. Essentially, user requirements analysis is about setting the base-line for developing the application and refining it as the implementation is in progress so that it meets customer needs. It is now widely understood that successful systems and products begin with an understanding of the users. Improved user satisfaction, reductions in training costs, enhances the quality of work and increase the productivity is benefits of applying requirements of the user. The first step in user requirements analysis which I applied

in this project was to analyze the stakeholders who may influence or be impacted by the system. This helps ensure that the needs of all those involved are taken into account. In order to specify user requirements, desires, needs and expectations of user were to be gathered. By default, developer is a user but they are not final users. The project was basically addressed to farming community and then policy makers, researchers and other interested parties. Accordingly user requirements were grouped as in table 5.1.

Stakeholder	Requirements
Developer or Administrator	The user shall be able to login into the system as an administrator level
	The shall be able to alter the back end information
	The user shall be able to add new classification models or replace existing models as per the more accurate models proposed.
	The user shall be able to add new forecasting variables or replace existing variables as per the mode accurate models found with more yearly information gathered.
	The user shall be able to add new index for any more vegetable crops to be in the system environment.
Farming Community	The user shall be able to enter into the system with no barriers.
	The user shall be able to select any option created in the system with minimum effort.
	The user shall be able to select any of control variables to calculate and visualize expected harvest
	The user shall be able to view models in an easily understandable way for any of selected crop.
	The user shall be able to visualize behavior of actual and expected yeilds of any selected crops.

	The user shall be able to manipulate all requirements mention in farming community
	The user shall be able to compare classification models with changing measures of selected variables
Policy makers, Researchers and Other interested Parties	The user shall be able to understand highest and lowest yields thast may report during a given scenario
	The user shall be able to compare price fluctuations within a given Time
	The user shall be able to identify easily manageable control variable for enhancing vegetable production.
	The user shall be able to identify the high priority research areas which should be addressed.
	The user shall be able to compare performance of the system by actual values with expected yields.

Table 5.1 – User requirements

5.3 Functional Requirements

Other than the user requirements, the functional requirements of a system are the services the system should provide. Further, it is the behavior of the system saying how it reacts to particular inputs and behavior in specific situation. The proposed system has been developed in the following manner.

- The system shall search user name and password with the stored information in the login process.

21

- System shall enable specific interfaces for administrator for manipulating of back end information.

- Once the identification variables have been selected by the user, system goes through the model and locates appropriate result. Then get the values stored in the table and complete the calculation process.

- The system shall do the indexing of tables once the new cases have been added or existing cases have been removed.

- The system shall be able to update predefined all existing models.

- The administrator shall be able to search either all of the initial set of data tables or subset from it.

- The system shall provide appropriate viewers for the user to read related information.

5.4 Non-functional Requirements

These are other types of requirements that are not directly concerned with the system functions. These are emergent system properties such as reliability, security and response time. The system should be reliable for any of stake holder who wishes to use the system. High security is another requirement which should not allow unauthorized access to stored information. The system should be easy to use by any user with minimum technical knowledge. The figure 5.1 shows the business activity diagram. The use cases and activity diagrams (*Appendix G*) show how the system should interact with stakeholders.

Figure 5.1 - Business Activity Diagram of the Home page

23
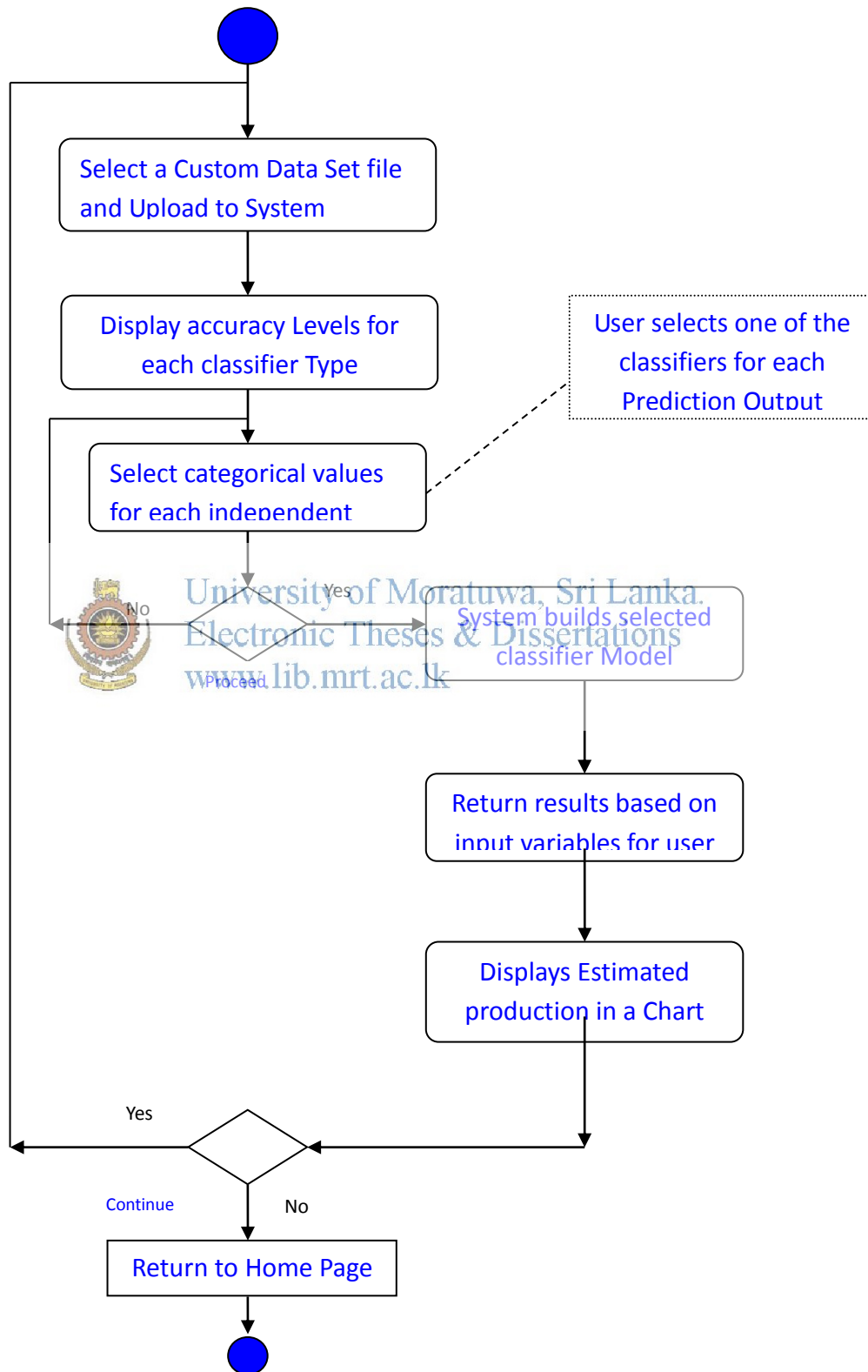
5.5 Summary

The developer or administrator, general farming community and policymakers together with researchers interested in similar topics, were considered as stake holders of the proposed system. The use-cases were drawn to indicate user interactions with the system. The next chapter describes the technology adopted and reasoning for the use of such technology for the implementation process of the project.

Technology Adopted for the System

6.1 Introduction

The previous chapter explained about various other attempts that were used to do mining and extraction of data to use for predictions. This chapter will discuss about the technologies that will be involved in the research and implementation phase of the project.

6.2 Reasons for selecting Proposed Technology

In this project, target groups were not limited to a single person or a single community located simple geographical coverage. It is expected to cover farming community who are currently or supposed to engage in upcountry vegetable cultivation and interested parties on research aspects to promote vegetable cultivation in Sri Lanka and also policy makers who are in the position of top level decision making. As the large number of users spread over the country or outside should give access to use the system, more suitable solution, particularly web based technology has to be used. Further, this technology facilitates to administrate, modify and manage the system from the centralized administration, and minimize the administration cost. Since we are really keen on cost saving, we host the system on a web belong to web hosting company. Since we don't have very sensitive data, it is not expected to adopt secure technologies into our solutions.

6.3 Background of the Technology Used

In order to develop the web based information system, two major tasks had to be concerned; particularly data mining and analysis tools and web development tools. The first process was completely a data mining and pattern identification process from the data gathered from the research center. A comprehensive data mining library – weka - was used to get the analysis for selecting appropriate models which had to be incorporated with the information system.

To initially get the data preprocessed before the actual classification process there needs to be a methodology for extracting necessary data from the given datasets such as the data set from the Hector Kobbekaduwa agrarian research and training institute and data collected through the survey. The following are the data extraction technologies which were used in the implementation of this project.

6.3.1 WEKA

The Weka workbench [2] contains a collection of visualization tools and algorithms for data analysis and predictive modeling with GUI for easy access. The original program was a non-Java version which was a modeling algorithm that was implemented in other programming languages, with data preprocessing utilities in C, and a Makefile-based system for the purpose of running machine learning experiments. This first version was mainly designed as a tool for analyzing the data from agricultural domains,[8] and now the recent Java-based versions (Weka 3), are now used in different application areas such as educational purposes and research.

There are many advantages of using Weka including the free availability under the GNU General Public License, portability, (fully implemented in Java and thus runs on almost any platform), comprehensive collection of the modeling and data preprocessing techniques and also user friendliness due to its GU interfaces.

26

Weka is supporting standard data mining tasks such as data preprocessing, regression, visualization, clustering, classification, and feature selection. These techniques are predicated under the assumption that all data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query.

Considering with the nature of this project and the amount of data preprocessing and classifications to be done along with the flat data provided by the resources it is very clear that Weka is one of the best tools suitable for the task.
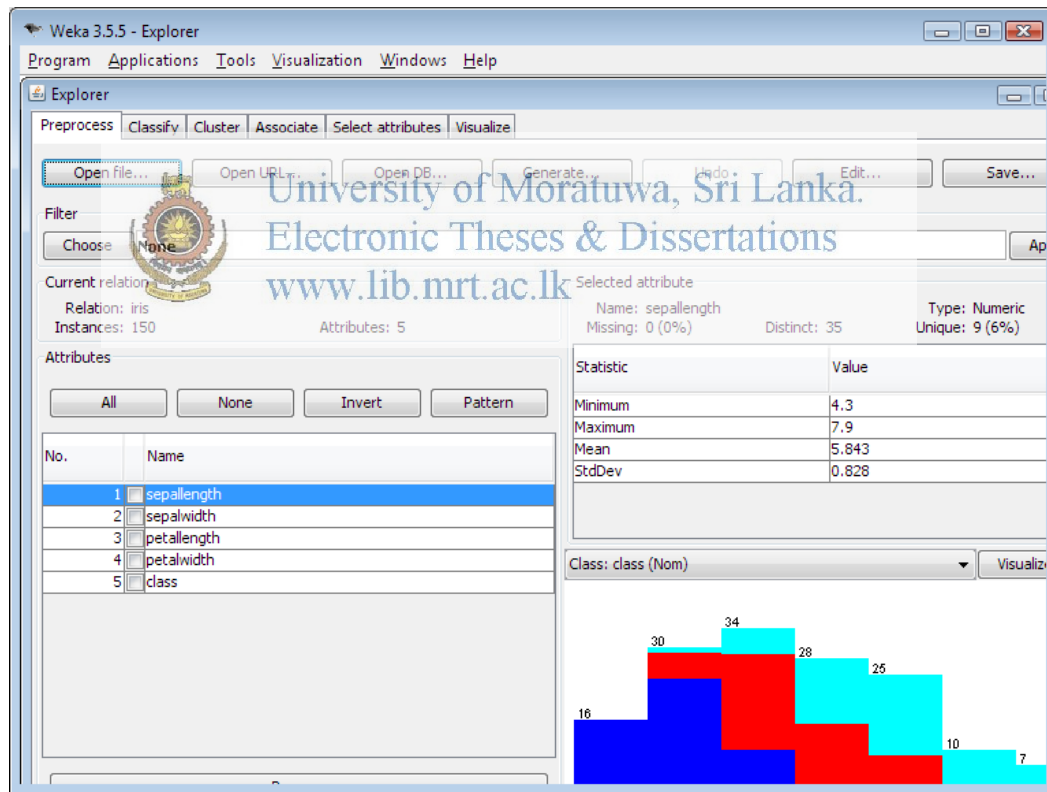
Figure 6.1 – User Interface Snapshot of Weka

6.4 Data Mining Technologies

Once the data is extracted with different attributes and values, we need to classify the data for the purpose of decision making and result generation. For this considering the following classification techniques in data mining is done.

6.4.1 Artificial Neural Networks

Artificial neural networks (ANNs) are computational models inspired by an animal's central nervous systems (in particular the brain) which are capable of machine learning as well as pattern recognition. Artificial neural networks are generally presented as systems of interconnected "neurons" which can compute values from inputs. For example, a such for handwriting recognition is defined by a set of input neurons which may be activated by the pixels of an input image. After being weighted and transformed by a function, the activations of these neurons are then passed on to other neurons. This process is repeated until finally, an output neuron is activated. This determines which character was read.

Like other machine learning methods - systems that learn from data - neural networks have been used to solve a wide variety of tasks that are hard to solve using ordinary rule-based programming, including computer vision and speech recognition.

6.4.2 Entropy based decision trees

A decision tree is a simple representation for classifying examples. Decision tree learning is one of the most successful techniques for supervised classification learning. For this section, assume that all of the features have finite discrete domains, and there is a single target feature

called the classification. Each element of the domain of the classification is called a class. A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions. This process of top-down induction of decision trees (TDIDT) [6] is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees from data.

6.4.3 Bayes Classifier

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

Naive Bayes is a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate preprocessing, it is competitive in this domain with more advanced methods including support vector machines.[7]Although they seem to be very simple and easy naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers. Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests.

29

An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

6.4.4 J48 Classifier

J48 is an open source Java implementation of the C4.5 algorithm in the weka data mining tool.C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set $S = \{s\_1, s\_2, ...\}$ of already classified samples. Each sample s_i consists of a p-dimensional vector $(x\_\{1,i\}, x\_\{2,i\}, ...,x\_\{p,i\})$, where the x_j represent attributes or features of the sample, as well as the class in which s_i falls.

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub lists.

This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.

- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.

- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

30

6.4.5 SMO Classifier

SMO is an iterative algorithm for solving the optimization problem described above. SMO breaks this problem into a series of smallest possible sub-problems, which are then solved analytically. Because of the linear equality constraint involving the Lagrange multipliers α_i, the smallest possible problem involves two such multipliers. Then, for any two multipliers α_1 and α_2, the constraints are reduced to:

$$0 \leq \alpha_1, \alpha_2 \leq C,$$

$$y_1 \alpha_1 + y_2 \alpha_2 = k,$$

and this reduced problem can be solved analytically: one needs to find a minimum of a one-dimensional quadratic function. k is the negative of the sum over the rest of terms in the equality constraint, which is fixed in each iteration.

The algorithm proceeds as follows:

- Find a Lagrange multiplier α_1 that violates the Karush–Kuhn–Tucker (KKT) conditions for the optimization problem.

- Pick a second multiplier α_2 and optimize the pair (α_1, α_2).

- Repeat steps 1 and 2 until convergence.

- When all the Lagrange multipliers satisfy the KKT conditions (within a user-defined tolerance), the problem has been solved. Although this algorithm is guaranteed to converge, heuristics are used to choose the pair of multipliers so as to accelerate the rate of convergence.

6.4.5 IBK Classifier (K-nearest neighbors classifier)

In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

- In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

- In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

6.5 Data Representation Technologies

In this project the results generated are intended to be presented in a user friendly manner so that any generic user can easily understand and interpret the results. Also the system will be designed in a way to easily input and configure data.

6.5.1 Joomla
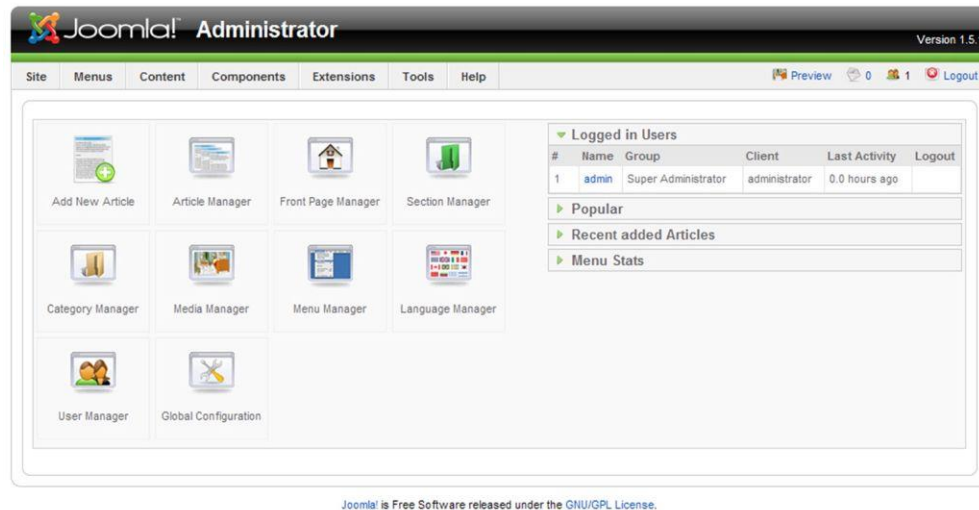
Figure 6.2 – Admin Interface Snapshot of Joomla

Joomla is a free and open-source content management system (CMS) for publishing web content. It is built on a model–view–controller web application framework that can be used independently of the CMS.

Joomla is written in PHP, uses object-oriented programming (OOP) techniques and software design patterns, stores data in a MySQL, MS SQL , or PostgreSQL database and includes features such as page caching, RSS feeds, printable versions of pages, news flashes, blogs, polls, search, and support for language internationalization.

In this project Joomla will be used to handle the input and output processes of data in the system and the users will be able to observe and interact with the activities such as prediction based on selected criteria and selected attributes only using the content management system.

The xampp server-win32-1.7.7-vc9-installer, the open source software was installed as a server which comes with apache, mysql and php web developing tools. Since this is open source, no cost incurred for installations or maintenance. Since the xampp is a collection of three different applications, no compatibility problems would involve as in individual installations of such applications and was used to deploy Joomla.

6.5.2 PHP and MySQL's place in development

PHP is a programming language, designing to generate web pages interactively on the computer serving them, which is called on web server. Unlike HTML, where the web browser uses tags and markup to generate a page, PHP code runs between the requested page and the web server, adding to and changing the basic HTML output.

While PHP is great for web application development, it does not store information by itself. For that, it requires a database. The database of choice for PHP developers is MySQL, which acts like a filing clerk for PHP-proceed user information. MySQL is easily accessed from PHP, and they work well together. An added benefit is that PHP and MySQL run on various computer operating systems. The following advantages were highly considered for selecting of PHP with MySQL.

- PHP and MySQL work well together.
- PHP and MySQL have open source power
- PHP and MySQL have community support
- PHP and MySQL are fast.

6.6 Summary

This chapter described about the technologies involved in this project. The proceeding chapter will discuss about how we took the approach to research and develop a tool to demonstrate our findings.

34

Implementation

7.1 Introduction

This chapter discusses the methodology of implementation with the aim of achieving the objectives declared in the project proposal. The information that is stored in the system, the way how these have been stored, together their descriptions and relationships made on key information have also been discussed. Further, components and features of the web side have also been included for easy accessing and for locating of the information required. Key interfaces have also been incorporated into this chapter and the detailed java codes written are given in annexure.

7.2 Data Preprocessing Implementation.

The project implementation is mainly done through two modules namely training and prediction modules. In the training module the source data is extracted by considering the relevancy to the task and the given four vegetables. Then the data is preprocessed to minimize the redundancy and to avoid error.

Data goes through number of steps in the pre-processing component:

1. Data Cleaning: The data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data. Data with missing values and extreme values are omitted from the dataset.

35

2. Data Transformation: The data is normalized, aggregated and generalized.   Categorical data are transformed using the one-of-N method and numerical data are transformed using the numeric scaling.

3. Data Discretization: This step involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals. The output variable, which is the yield per given area, is discretized in three classes as to low yield, average yield, and high yield. Figure 6.1 shows the discretization process.
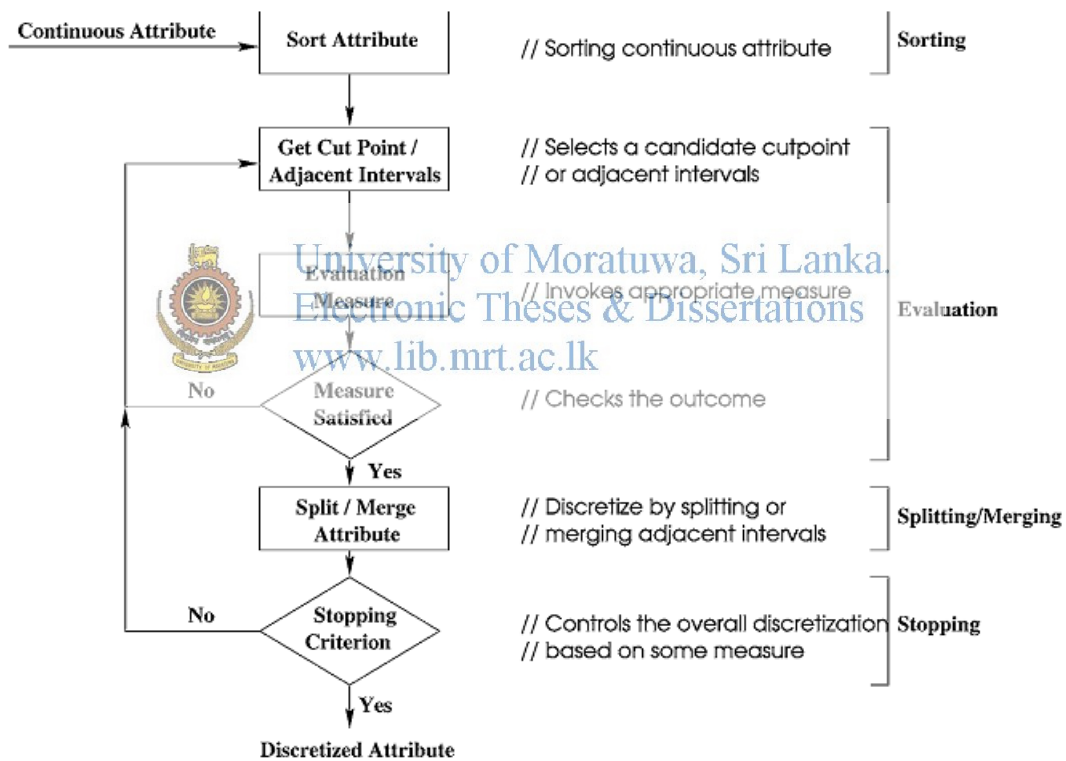
Figure 7.1 – Discretization Process in Data Preprocessing

7.3 Classification Implementation using Weka and Java

Throughout the implementation of the project the data mining tool Weka is used as the main classification and mining resource. The tool is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.
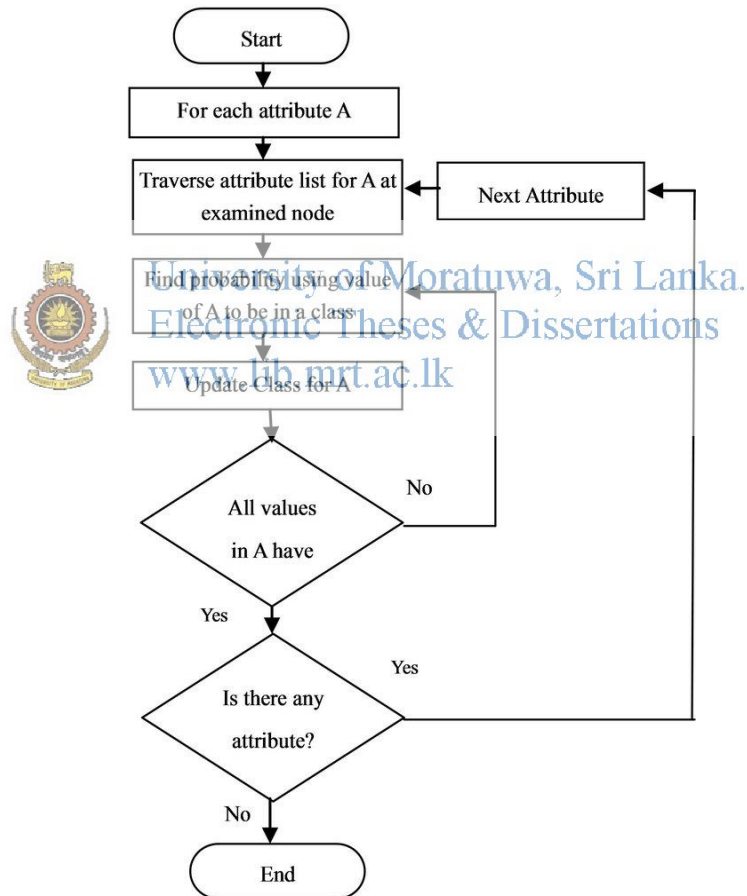
Figure 7.2 – Flowchart of Naïve Bayes decision tree algorithm.

In this project Weka will is used as the main tool to perform tasks in the Training Module for the purpose of data extraction, preprocessing and classification. Weka is also useful in modeling all the classifiers that are to be implemented in the Tool. There are four main classifiers used in the tool.

- Naive Bayes classifier

- C4.5 is an algorithm (implemented in Weka as j48)

- Sequential minimal optimization algorithm (implemented in Weka as smo)

- K-nearest neighbour's classifier (implemented in Weka as ibk)

7.3.1 Dynamic classification using Weka libraries

In the classification module there are two main work flows as per two types of user interactions. If a guest user is requesting a prediction, the system uses a default model generated by the system and outputs a prediction using the default model generated using naïve Bayes classifier.

The second work flow is when a registered user requests a custom prediction based on a selected classification algorithm and his/her inputs. In this scenario the system uses the user uploaded data file or the default data file and generates a classification model dynamically to suit the user request. Then that model is used to make predictions relevant to the user input feature vector.

7.3.2 Data Storage and Generation

There are two main types of data related to this module of the project. They are mainly the user data uploaded and sent through the RESTful web service and the data generated by the system related to the first category of data. The storage of both these types of data is managed by the system using a folder hierarchy.

The user data is stored in a folder named "userData" in the web server while the default data set is stored in a folder named "defaultData" in the web server. Both file types will be used to build models and the serialized objects of these models will be stored in "build" folders inside the respective folders.

The "cropinfo_data" database in the other hand will store the results for each dynamic generation of classifiers and their accuracy levels in the "dynamics" table and the user query requests will be stored in the "results" table in the same database. (See Appendix D for more information on the data storage and Appendix E for the java implementation)

The RESTful web service is responsible for returning a string type result for every user query for results irrespective of whether it is a default prediction request or a custom prediction request. Also it will handle the file uploads and dynamic classifier generation request through separate parameters and methods in the code.

7.4 Joomla Based Web Interface

Joomla was used as the main Content Management System tool to develop the web interface of the system. I had to use some additional plugins to implement additional functionality to the system such as dynamic forms and adding graphs to the results page.

The plugins used for these functionalities are "Chronoforms Joomla Plugin" for adding dynamic forms and "Plotalot" for adding dynamic graph drawing to the system. The site was developed using a Joomla default template and the database used to keep data related to the website is "cropinfo_web" (See Appendix D for more information on the databases) and it is managed by Joomla itself.

The site was decorated by adding some features such as dynamic banners and colorful custom templates. The result generation is further simplified by presenting the prediction output as a gauge chart in the results page.

39

7.5 Design Challenges and Resolutions

There were a lot of challenges and difficulties to face from the research phase up to the implementation of this project. Many of them were related to the data extraction phase/layer and manipulating the data sources of the project.

One major issue encountered is extracting data from a huge set of unclassified details of a data bank with a lot of unwanted attributes. There were no direct methods or guidelines to follow on which feature to use on determining different values. Although there are available tools for higher level data extraction the process has to be done manually for this specific set of data related to Sri Lankan agriculture.

Also to do a lot of experiments using different data mining algorithms which sometimes did not prove to give any useful output related to the interests of the project. For an instance the neural network approach seemed to be not useful and was never used in the implementation of the tool although a lot of research on it was done.

7.6 Summary

This chapter has discussed the components of the system, how those have been stored and incorporated with the system. The implementation phase given attention to each module using the pseudo code. The next chapter describes the testing and evaluation of the system.

Evaluation

8.1 Introduction.

This chapter discusses how the proposed solution meets the objectives of the project and accordingly to achieve the ultimate goal of the project. Further, how far and up to which extend, the solution has been successful. Also this chapter discusses the various evaluations conducted to select an approach for default prediction of the main system and to evaluate the success of the system implemented.

8.2 Achievements of the objectives

Basically there are three main objectives to be achieved in the process of the project

1. To Study on the problem domain, to propose a solution based on Data mining and to forecast very veritable information to deliver to the user was achieved through obtaining data from the Data Bank of Hector Kobbekaduwa Agrarian Research and Training Institute and developing a model to use these data in this system.

2. To Study data mining, web technologies that support the implementation of it has been achieved by conducting research on various data mining techniques and tools , content management tools and classification models and algorithms.

3. To design and develop an information system to incorporate above technologies was achieved as the main objective of this project by the overall implementation of this CIP System and adding functionality to it.

8.3 Classifier Algorithm Evaluations

One of the problems that were faced in the project was selecting a better classification algorithm for the default prediction flow and this was achieved through performing evaluations for each of the four classification algorithms selected suitable for this type of multi value nominal class attribute classification.

- Naive Bayes classifier Evaluation

```
Scheme:weka.classifiers.bayes.NaiveBayes
Relation:     eightlevel-weka.filters.unsupervised.attribute.Remove-R11
Instances:    862
Time taken to build model: 0 seconds


=== Evaluation on test split ===
=== Summary ===


Correctly Classified Instances        189            74.5051 %
Incorrectly Classified Instances      104            25.4949 %
Kappa statistic                                      0.2261
Mean absolute error                                  0.1761
Root mean squared error                              0.3118
Relative absolute error                              63.0294 %
Root relative squared error                          55.9401 %
Total Number of Instances                            293
```

- C4.5 is an algorithm (implemented in Weka as j48) Evaluation

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: eightlevel-weka.filters.unsupervised.attribute.Remove-R11

Time taken to build model: 0.03 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances	206	70.3072 %
Incorrectly Classified Instances	87	29.6928 %
Kappa statistic		0.6447
Mean absolute error		0.0773
Root mean squared error		0.2265
Relative absolute error		36.4608 %
Root relative squared error		69.6772 %
Total Number of Instances		293

- Sequential minimal optimization algorithm (implemented in Weka as smo) Evaluation

Scheme:weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K

"weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"

Relation: eightlevel-weka.filters.unsupervised.attribute.Remove-R11

Time taken to build model: 0.16 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances	130	44.3686 %
Incorrectly Classified Instances	163	55.6314 %
Kappa statistic		0.3364
Mean absolute error		0.2015
Root mean squared error		0.3138
Relative absolute error		94.9773 %
Root relative squared error		96.5433 %
Total Number of Instances		293

- K-nearest neighbour's classifier (implemented in Weka as ibk) Evaluation

Scheme:weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

Relation: eightlevel-weka.filters.unsupervised.attribute.Remove-R11

Time taken to build model: 0 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances	223	76.1092 %
Incorrectly Classified Instances	70	23.8908 %
Kappa statistic		0.716
Mean absolute error		0.0584
Root mean squared error		0.2136
Relative absolute error		57.5431 %
Root relative squared error		65.7259 %
Total Number of Instances		293
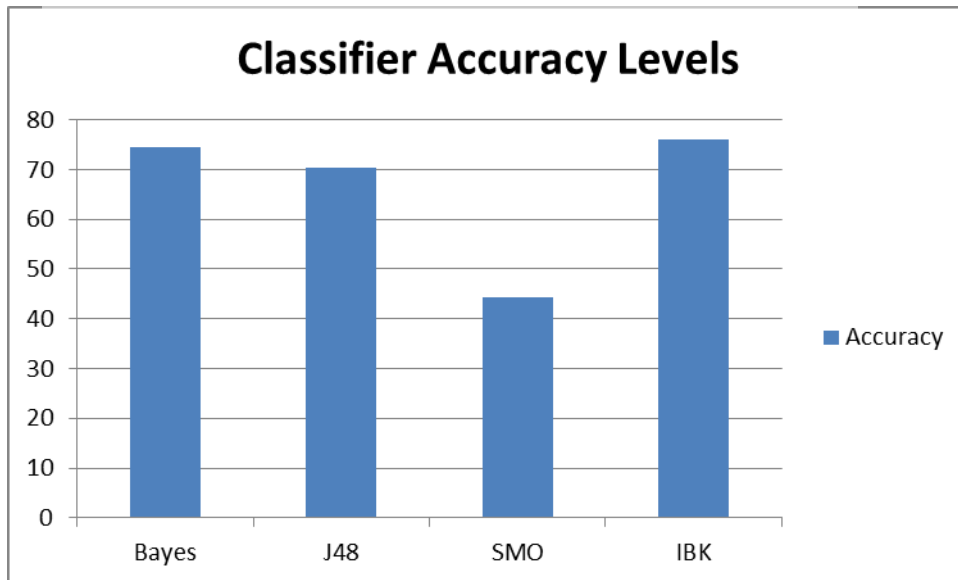
Figure 8.1 – Classifier Accuracy Levels
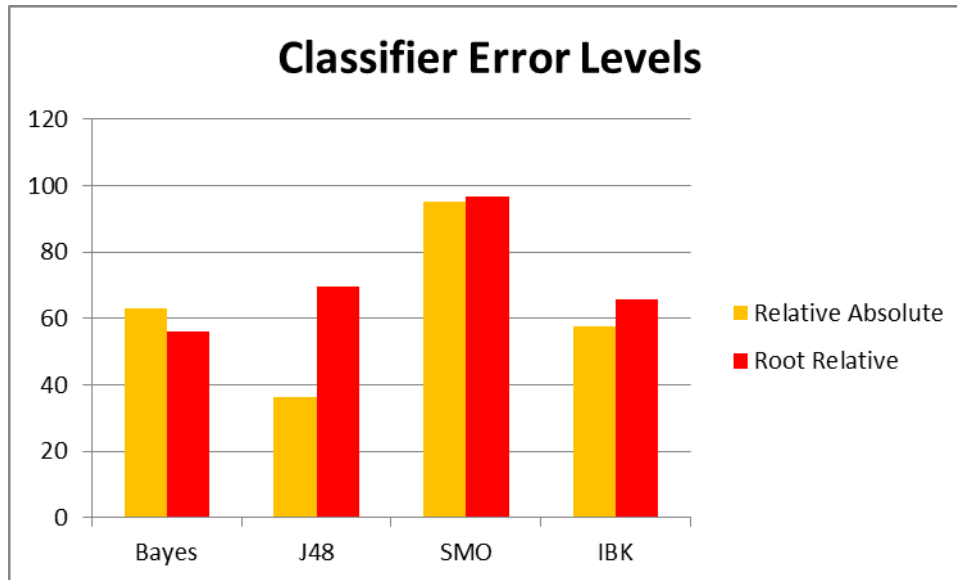
Figure 8.2 – Classifier Error Levels

Considering the accuracy levels as well as the error levels of each approach the Naïve Bayes classifier was selected as the best common approach for a default classifier and model builder for this crop prediction system.

8.4 Web-based Interface for the System

(a) Overview of the Software

The development of web interface was the final objective of the project. This software was basically developed for catering the requirement of decision making authorities as decision support system, especially in the context of up country vegetable cultivation in Sri Lanka. At the final stage of development, comments were obtained from supervisors, colleagues and farming community for investigating drawbacks and further implementation.

45

(b) General Usability

The usability is one of key issues of any interactive system and quality of interaction describes in terms of time taken to perform a task, number of errors could made and time to become a competent user. Further the system should be easy to user, easy to learn and flexible. However, the goals of usability are now primarily seen as concerned with efficiency and effectiveness of system. Therefore in these criteria, main concentration is drawn to the area of easiness of using and learning the system and flexibility of the system. Hence in the discussion with the above users, they are satisfied with the system in the context of easiness of using. The stake holders who tested the system accepted that the developed system is efficient. Effectiveness of system has been achieved in the way that things are arranged in appropriate manner in the software. Since the system has been developed using widely used standards and symbols, anyone can easily learn and remember.

(c) Accessibility

Accessibility is also one of the key issues concerned about a system and ultimate goal is removing barriers. By the other hand, readability, predictability and input assistance is the main area to be tested with the users under accessibility. As text content on the system can be readable, users conclude that system have met that criteria. Since web pages appear and operate and in predictable way, it is accepted that predictability has been achieved in the system. Moreover system provide facilities to avoid and correct mistakes, it is seen that requirement of input assistance has been met. Accordingly accessibility exclusions such as physically, conceptually, economically, culturally and socially are avoided as much as possible. The system was developed with the features of universal Design such as equitable use, flexibility in use, simple intuitive use and perceptible information as much as possible.

(d) Acceptability

This is a key issue to be considered when interface designing. This is about fitting technologies into people's lives. Acceptability is also a key issue concerned about a system and ultimate goal is fitness in the context of use. In the context of new technologies, many communities aware about the technology are preferable to accept the product.

(e) Interface and product design

In the context of obtaining user's attention on designing interfaces and their feedback interface designing is so important. Therefore, user's opinions were obtained on design principles such as;

Visibility   -How users to perceive systems functionalities and state of system

Consistency   -Be Consistent in the appearance, function and structure

Familiarity   - Use language and symbols familiar to the intended audience

Affordance   - Design things so it's clear what they are for buttons look press able

Navigation   - Provide support for finding and accessing the system

Control   - Make it clear who or what is in control

Feedback   - Provide constant and consistent feed back

Recovery   - Enable quick and effective recovery from actions especially errors

Constraints   - Prevent inappropriate actions and serious errors

Flexibility   - Allow multiple ways of doing things

Style   - Designs should be stylish and attractive

Conviviality   - Be polite, friendly and generally pleasant

Users identified the system is consistent in the appearance, function, and structure to the similar systems and standard way of working. Further it ensures consistency in behaviors, colors, names, and layout and so on. As this system uses languages and symbols familiar to the intended users, users are feel more familiarity. Affordance is seen throughout the system through designing the things in a manner that it is clear what purpose they are for. For examples buttons are designed in a way its look press able.

Navigation provides support for finding and accessing the system. Here menus are often used for navigation. Further this system provides constant and consistent feedback which allows users to know what effect their actions have had. Recovery is also important aspect available in the system which enables quick and effective recovery from action especially mistakes and errors. Principle of constraints helps to prevent inappropriate actions and serious errors. So system asks confirmation massages before going for such dangerous and irrecoverable operations. System provides the facilities to do various operations in multiple ways achieving the flexibility. Further users accept that system is designed in stylish and attractive manner.

(f) Guidelines

Help file provides the assistance to the users on various matters relating to the system. So assistance given by the help menu is evaluated by the users and they are satisfied with same. The system guide lines were highly appreciated by the users who worked with the system.

(g) Security

The password protection is only the security measure concerned in the system. Highly security measures were not taken, since there is no highly confidence information. Accessibility for general public is also important for further implementation of this system and possibility of investigating to go for another method applicable. However, daily back ups are taken for data security and avoiding the data losses. This back up process is automatically

run and back up is done to the internal network. Though the system is hosted the data base is stored on internal network as security measure. The sensitive data is transmitted using the method of encryption and decryption ensuring that no one can hack the data in the process of transmission. Moreover data can be entered into the system using sql procedure from the administrator. This is also effective security measure implemented in the system.

8.5 Assessment on Results

Results coming from classification models are based on the efficiency and accuracy of each classification algorithm. In some of models performances are highly accepted while some of them are showing little deviations from actual. However from a further analysis, this deviation values could be changed, (e.g. by cleaning data more thoroughly) but the system limits to existing number of variables and no more variables may introduce to system since it will further reduce the accuracy. All the variables have limitations such as several categories only. There can be unavoidable situation like weather conditions and disease. The harvest forecasting model discusses general trend of variation and it has also shown over 80% of performance.

8.6 Summary.

This chapter has mainly concerned about overall evaluation of objects that were achieved throughout the project. To make a reasonable evaluation, the system was tested with stakeholders and altered on their comments. Therefore, in summary, it has been proved that system full fill the evaluation criteria satisfactorily. Further, next chapter discusses conclusion of further work.

Conclusion & Further work.

9.1 Introduction

The previous chapter discussed about how the tool is evaluated in a variety of ways and means. In this chapter it will be briefly described about the progress of the project, the conclusions from the evaluations and the work and the improvements and the development that that is expected to carryout in the next phase of the project. Also, attention is drawn to assess the achievements of objectives stated in the project proposal and the further work, which could be carried out to improve the derived models qualitatively.

9.2 Assessment of the system.

The first objective namely 'data gathering and domain analysis' had been done by obtaining data from the Data bank of Hector Kobbekaduwa agrarian research and training institute , a national center for secondary source of information which has been collecting primary information from many government institutions since 50's. Currently there is a huge volume of agricultural information in the institute data bank. In this project, detailed records of four selected upcountry vegetables namely potatoes, beans, carrot and cabbage were queried from their databases and analyzed with their attributes for the usage in the this project and therefore it can be concluded that data gathering have been done successfully.

Next objective was the forming classification models to predict the vegetable production, when independent variables present. Even though there can be many hidden variables for varying of production, it was attempted to make relationship with about 10 variables which

could affect the production of vegetable. The developing classification models were so critical, since in the process of calculating the production; there can be high deviations from the actual. Even though it was possible to form one model for entire sample by taking key variables such as districts, seasons and type of vegetable as independent variables, it was attempt to build varying models depending on districts, type of vegetable and the season. The agro-ecological zones were not considered as farming community is not aware about such zones rather than districts. However, with the use of weka application, various methods were used to build up classification models dynamically and finally came to a conclusion that Naïve Bayes was the most suitable common model for the used dataset.

The development of web-based information system was the final objective, which is a collection of all above objectives stated. The web site was tested with colleagues and supervisors and they accepted that its functionalities work properly. However, accuracy of estimation of production has to be verified with similar data collected from farming community on the basics of equal situation.

Farming community can go through the system and farmers can plan to obtain not only the higher yield but also higher price for the marketing. Since this is as a decision making system, by studying harvest fluctuation and crop forecasting, relevant authorities can take most appropriate decisions for the minimizing of harvest variations. According to the expected harvest, possible most appropriate steps can be taken to maximize the expected harvest.

The farmers, through the system can be educated to apply varying methods and from that the expected harvest could be compared with actual and performance could be verified. A monitoring process could be implemented by field officers attached to agriculture sector in order to compare performance of the system from that suggestions could be made for further implementation.

9.3 Problems encountered and Limitations in development of project.

There were a lot of challenges to face from the research phase up to the implementation of the finding so far in this project. Many of them were related to the data extraction phase/layer

and manipulating the data sources of the project.

One major issue encountered is extracting data from a huge set of unclassified details of a data bank with a lot of unwanted attributes. There were no direct methods or guidelines to follow on which feature to use on determining different values. Although there are available tools for higher level data extraction the process has to be done manually for this specific set of data related to Sri Lankan agriculture.

Also to do a lot of experiments using different data mining algorithms which sometimes did not prove to give any useful output related to the interests of the project. For an instance the neural network approach seemed to be not useful and was never used in the implementation of the tool although a lot of research on it was done.

Further in Sri Lanka computer literacy is so low in the agricultural districts, it has become difficult situation to popularize this web base software among farmers. Since now some training programs are going on for farmers, it is hopefully expected to increase the usage of this system. The Hector Kobbekaduwa Agrarian Research & Training institute is in the process of conducting awareness programs in various discipline in agriculture and socio-economic aspects and hence dissemination of knowledge may not be much difficult problems, if necessary financial grant was found. In fact, farmers could really practice with the system for better knowledge.

9.4 Future Developments.

This web-based software has been developed basically as a decision support system which could be used to enhance the up country vegetable cultivation for the satisfaction of country requirements. Similar decision support systems could be developed for the applying of other highly consumable vegetables too. Further, such systems could also be expanded by conducting sample surveys for demanded crops in relevant areas. Based on findings of this system, effectiveness of similar variables could also be carefully measured through well-

organized farming practices. It can be assumed that rain falls and climate changes also important factors to be considered when cultivating of vegetables. Therefore, the systems could also implemented covering agro-ecological zones where factors like condition of soil, soil erosion and rain falls and slope of lands are highly differs.

Even though government and private organizations and many researchers have been conducting numerous studies on price variations, marketing, cost of productions and in various discipline, many problems have been created during the harvesting seasons. Furthermore in Sri Lanka there is no derivative market for any product and financial market should be upgraded and strengthen for such market. This is mainly affected to agricultural products as in other countries derivatives market is more powerful for agricultural market and benefitted to the farmers. Now discussion are going on the development of derivative market in Sri Lanka.Therefore as future development system should provide facility to farmers to search information on related discipline.

9.5. Summary.

This chapter discussed the assessment of the final product and how the product is use full for the parties like farming community and research and decision making authorities and further discussed the limitation and problems encountered regarding the system and possible future developments. As a system in a changing environment, this has the full potential to expand for accommodation of such changes. Further it was suggested future development to run as central unit of information dissemination for enhancing agricultural improvement in Sri Lanka. Therefore fulfilling those further developments too, an effective system for agricultural sector can be materialized. Things proposed to be done in research basics will influence the economic conditions of the country. The proceeding chapter will contain the list of references that we took the use of when implementing the project.

References

[1] Central Bank of Sri Lanka, economic and social statistics of sri lanka 2012,
http://www.cbsl.gov.lk/pics_n_docs/10_pub/_docs/statistics/other/econ_&_ss_2012.pdf

[2] Ian H. Witten; Eibe Frank; Mark A. Hall (2011). "Data Mining: Practical machine
learning tools and techniques, 3rd Edition".

[3] Maria Rossana C. de Leon, Eugene Rex L. Jalao, A Prediction Model Framework for
Crop Yield Prediction

[4] Mohammadhossein Hajiyan, Early Prediction of Crop Yield, School of Engineering ,
University of Guelph

[5] Paul C. Doraiswamya , Bakhyt Akhmedovb , Larry Beardc , Alan Sterna and Richard
Muellerc, OPERATIONAL PREDICTION OF CROP YIELDS USING MODIS DATA AND
PRODUCTS, ISPRS Archives XXXVI-8/W48 Workshop proceedings: Remote sensing
support to crop yield forecast and area estimates

[6] Quinlan, J. R., (1986). Induction of Decision Trees. Machine Learning 1: 81-106, Kluwer
Academic Publishers

[7] Rennie, J.; Shih, L.; Teevan, J.; Karger, D. (2003). "Tackling the poor assumptions of
Naive Bayes classifiers", Artificial Intelligence Laboratory; Massachusetts Institute of
Technology; Cambridge, MA 02139

[8] S.R. Garner; S.J. Cunningham, G. Holmes, C.G. Nevill-Manning, and I.H. Witten (1995).
"Applying a machine learning workbench: Experience with agricultural databases".
Proc Machine Learning in Practice Workshop, Machine Learning Conference, Tahoe City,
CA, USA. pp. 14–21.

Project Plan

ID	Task	Task Name	Duration	Start	Finish	Pred's
1	**1**	**Proposal Writing**	**17 days**	**01/01**	**01/17**	
2	1.1	Information Collecting	06 days	01/01	01/06	
3	1.2	Problem Identification	03 days	01/07	01/09	1
4	1.3	Writing Draft Proposal	04 days	01/10	01/13	2
5	1.4	Obtaining Supervisors Ideas	02 days	01/14	01/15	3
6	1.5	Finalizing Proposal	02 days	01/16	01/17	4
7	**2**	**Information Gathering**	**43 days**	**01/18**	**03/02**	
8	2.1	Gathering of Sample Data	07 days	01/18	01/25	
9	2.3	Survey Data Analysis	13 days	01/26	02/08	
10	2.4	Building Data Structures	18 days	02/09	02/28	9
11	2.5	Storing Data in Format	03 days	02/29	03/02	10
12	**3**	**Project Analysis and Design**	**52 days**	**03/03**	**04/26**	
13	3.1	Requirement Analysis	10 days	03/03	03/13	
14	3.2	Drawing System Diagrams	10 days	03/14	03/24	13
15	3.3	Designing	32 days	03/25	04/26	14
16	**4**	**Implementation**	**97 days**	**04/27**	**08/10**	
17	4.1	Coding the simple Web Service	28 days	04/27	05/29	
18	4.2	Initial Site Design Using Joomla	14 days	05/30	06/13	17
19	4.3	Building Models using Weka	35 days	06/14	07/19	18
20	4.4	Testing the Initial System	20 days	07/20	08/10	19
21	**5**	**Submission of Interim Report**	**38 days**	**08/10**	**09/20**	
22	5.1	Writing of interim Report	14 days	08/10	08/24	

23	5.2	Preparing of Presentation	05 days	08/25	08/30	22
24	5.3	Conducting the Presentation	01 day	09/20	09/20	23
25	5.4	Obtaining vital Feedback	01 day	09/20	09/20	24
26	**6**	**Implementation Cont.**	**101 days**	**09/21**	**01/01**	
27	6.1	Finalizing the Technologies	14 days	09/21	10/04	
28	6.2	Installation of the VM	02 days	10/05	10/06	27
29	6.3	Making Databases	10 days	10/07	10/16	28
30	6.4	Joomla Site Completion	08 days	10/17	10/24	29
31	6.5	Implementing Dynamic Modeling	35 days	10/25	11/30	30
32	6.6	Implementing Dynamic Cleaning	10 days	12/01	12/10	31
33	6.7	User Interface Improvements	08 days	12/11	12/17	32
34	6.8	Testing and Debugging	14 days	12/18	01/01	33
35	**7**	Evaluation	27 days	01/02	01/29	
36	7.1	Running System Evaluations	10 days	01/02	01/11	
37	7.2	Comments from Supervisor	01 day	01/12	01/12	36
38	7.3	Testing for Dynamic Behavior	08 days	01/13	01/20	
39	7.4	Adding Improvements	08 days	01/21	01/29	38
40	**8**	**Completion**	**16 days**	**01/30**		
41	8.1	Writing the Final Report	14 days	01/30	02/12	
42	8.2	Submission	01 day	02/16	02/16	41
43	8.3	Final Presentation	01 day			42

List of coding used in the Dataset

1. District (dist)

   1 - Matale   2 – Nuwara-eliya

   3 - Badulla   4 – Kandy

2. Age group (age)

   1 - Age < = 35   2 – 35 < Age <=50   3 - 50< Age

3. Farmer's knowledge regarding agro ecological problems (trn) Soil erosion/excess use of fertilizer and agro chemicals

   1 – Not known   2 – Insufficient   3 – Sufficient

4. Farmers knowledge about use of fertilizer(fert) Mixtures of fertilizers/ pure fertilizer

   1 – No   2 – Some time   3 – Yes

5. Type of plot(plty)

   1 – High land   2 – Low land

6. Slope of the plot(slty)

   1 – Flat land   2 – Less than 45 degrees

   3 – 45 Degrees   4 – Above 45 Degrees

7. Soil erosion Intensity(ints)

   1 – Excess   2 – Average   3 – No erosion

8. Season (sesn)

    1 – Maha   2 – Yala   3- Intermediate

9. Cultivated crop (crop)

    1 – Potatoes   2 – Beans

    3 – Carrot   4 – Cabbage

10. Extent cultivated in Ac (extn)

    1 – Extent < 0.25 2 – $0.25 =<$ Extent < 0.5

    3 – $0.5 =<$ Extent < 0.75 4 – $0.75 =<$ Extent < 1

    5 – 1.0 or above

Use cases and Activity diagrams

Figure G.1 – Production Estimation Process
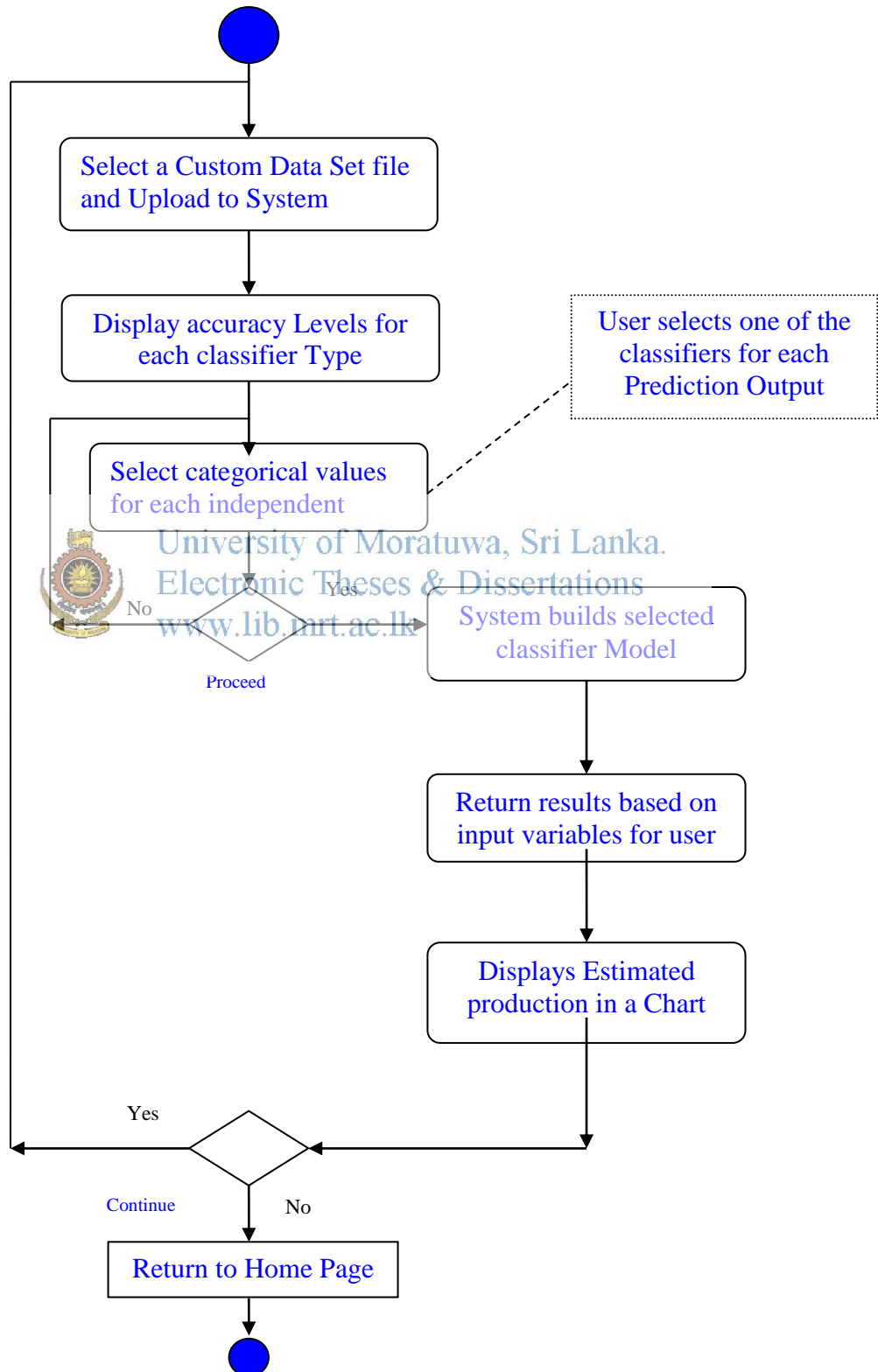
Figure G.2 – Administrator functionalities Use Case diagram

View Home page

View Dynamic information

Login

Visualize default Expected harvest

Manipulate Classification Models

Administrator

Visualize estimated production

Manipulate Harvest forecasting

Manipulate Classifier

Calculate expected Harvest

Display each Classifiers Results
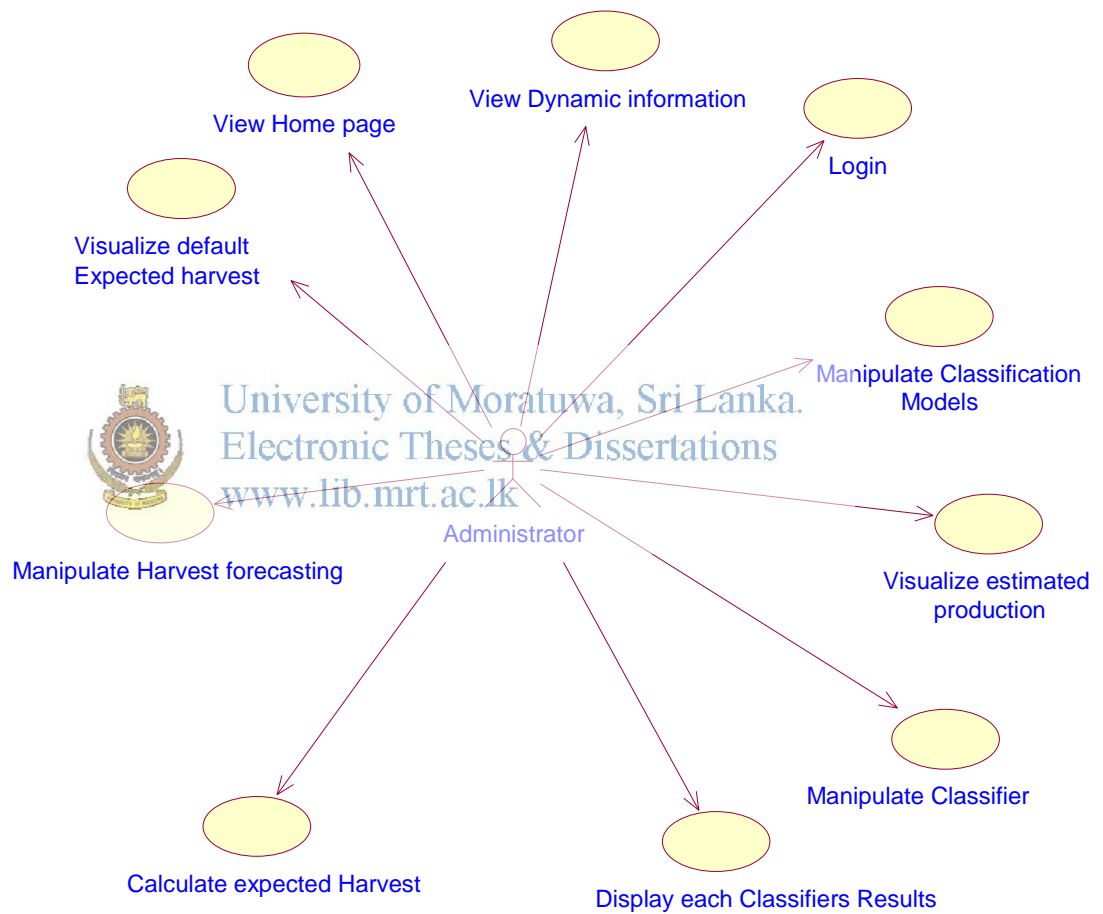
Figure G.3 -Farming community functionalities Use Case diagram

View Home page

View Dynamic information

Farming
User

Visualize Default Harvest

Visualize estimated production

Calculate expected Harvest

Display Harvest
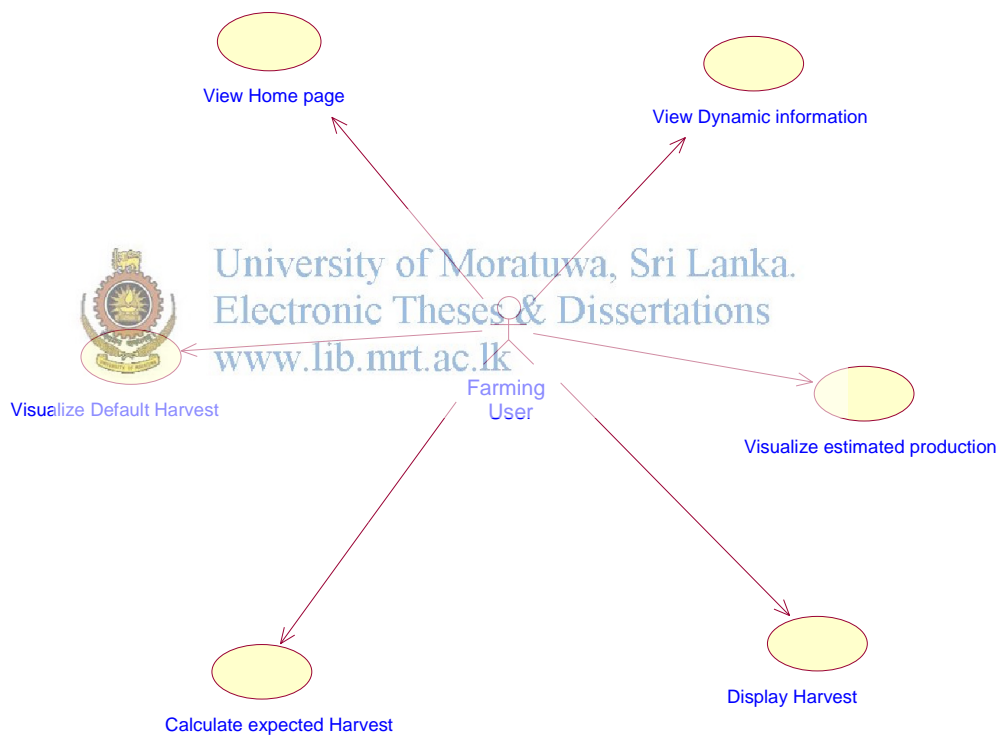
Figure G.4 -General client functionalities Use Case diagram

Login as a specific user

View Home page

View Dynamic information

Register as a User

Guest

Visualize estimated production

Visualize default Prediction

Display Harvest

Calculate expected Harvest
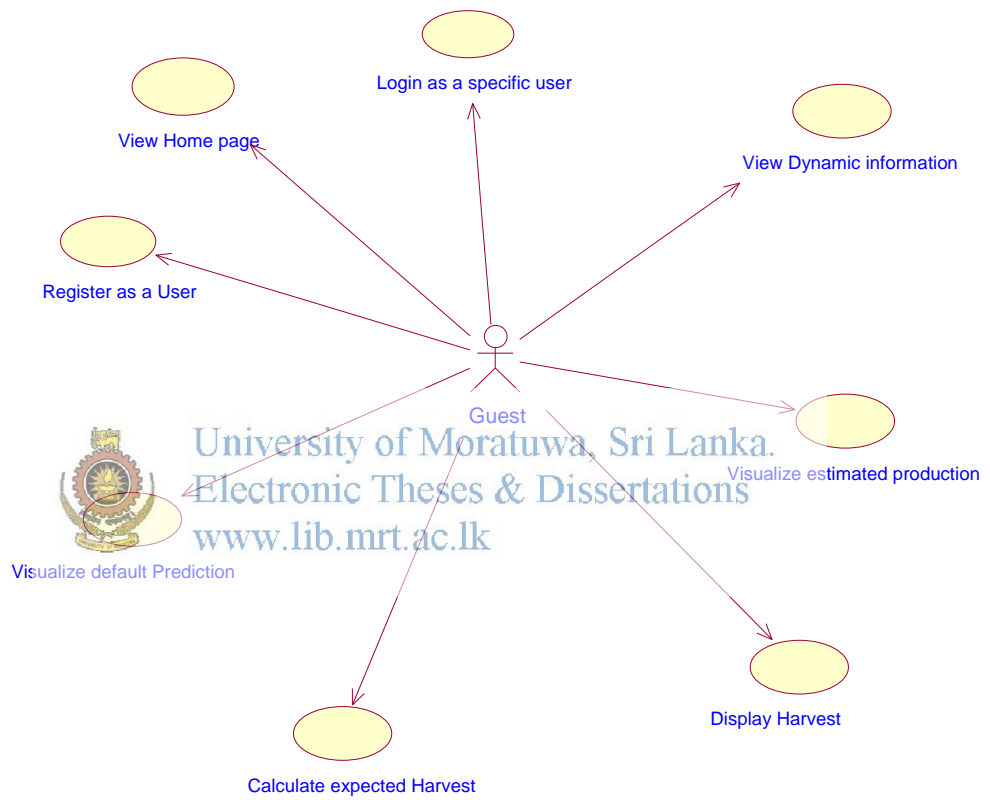
Description of Use Cases

The administrator functionalities use cases have been described below.

Use Case 01 – View Home Page

Use Case Name	View Home Page		Administrator Functionalities
Priority	High		
Primary Business Actor	Administrator		
Brief Description	User Navigate to the web site's home page		
Pre-Condition	User should have the Internet facility		
Basic Flow	1	User type URL of home page	
	2	View home page	
Alternative Flows	1	User type URL of home page	
	2	URL is incorrect	
	3	Navigate to an incorrect page	
Relationships	N/A		
Use-Case Diagram	Administrator — View home page		
Special Requirements	N/A		
Post-Conditions	N/A		
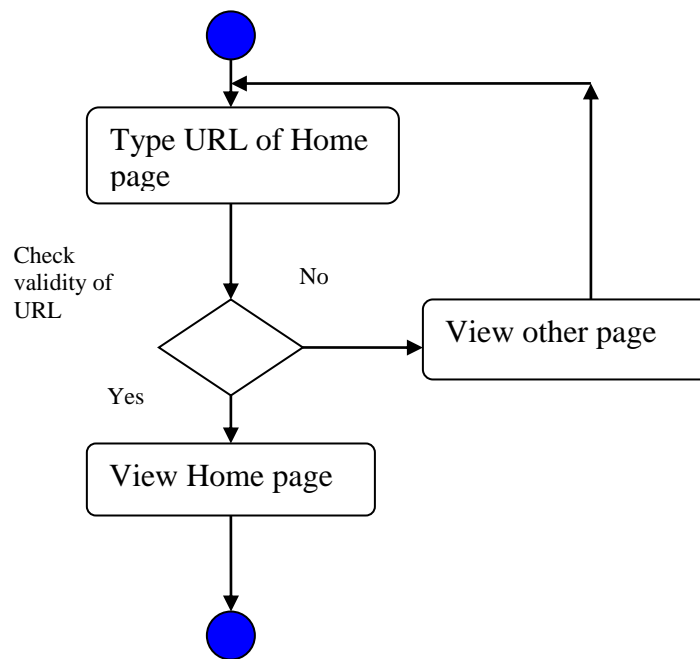
| Activity Diagram | |

Use Case 02 – View Dynamic Information

Use Case Name	View Dynamic Information	Administrator Functionalities
Priority	High	
Primary Business Actor	Administrator	
Brief Description	User Navigate to the dynamic information page	
Pre-Condition	User should be in the site	
Basic Flow	1	Click the Dynamic information link
	2	View dynamic information
Alternative Flows	N/A	
Relationships	N/A	
Use-Case Diagram		
Special Requirements	N/A	
Post-Conditions	N/A	
Activity Diagram		

Use Case 03 – Login

Use Case Name	Login as Administrator	Administrator Functionalities
Priority	High	
Primary Business Actor	Administrator	
Brief Description	User Navigate to the web site's login screen	
Pre-Condition	User should be in the home page	
Basic Flow	1	Click the Administrator login link
	2	Type user name and password
	3	Successfully login to the administrator view
Alternative Flows	1	Click the Administrator login link
	2	Type user name and password
	3	Login is a failure
Relationships	N/A	
Use-Case Diagram	Administrator Login as Administrator	
Special Requirements	Lower cases be used	
Post-Conditions	N/A	
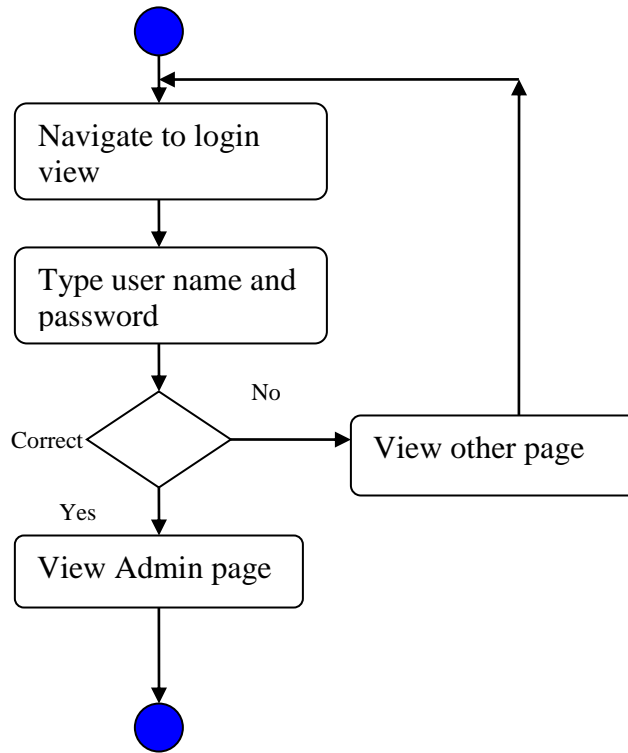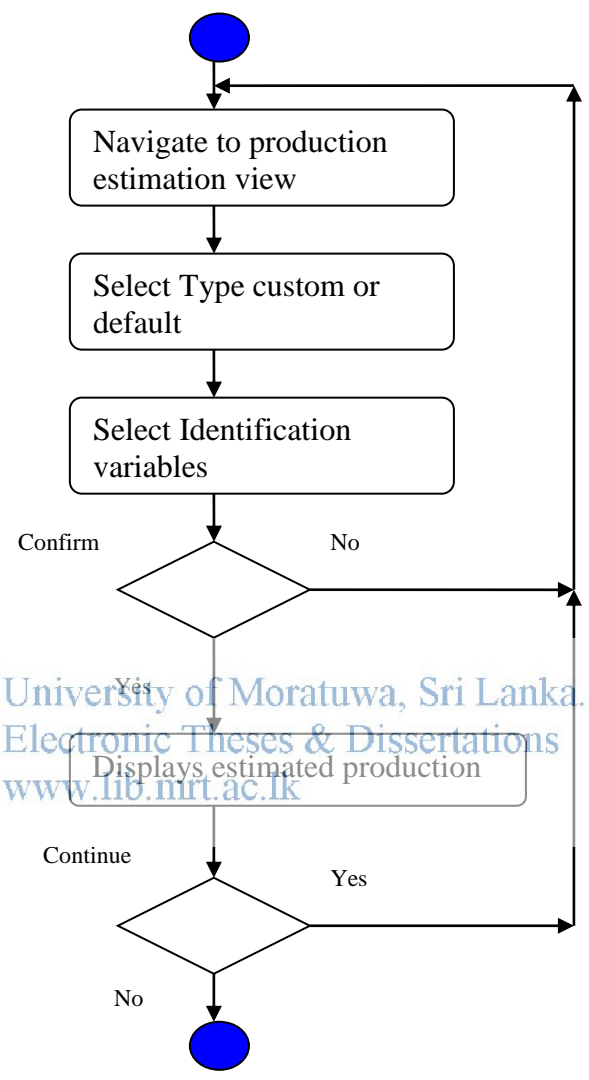
Activity Diagram	

Use Case 04 – Calculate and visualize estimated production

Use Case Name	Calculate and visualize estimated production	Administrator Functionalities
Priority	High	
Primary Business Actor	Administrator	
Brief Description	User Navigate to production estimated Regression models view	
Pre-Condition	User should be in the home page	
Basic Flow	1	Click the Production estimated view
	2	Select identification variables
	3	Select classifier type
	4	View the estimated production
Alternative Flows	1	Click the Upload File
	2	Select classifier type
	3	Select categories for independent variables
	4	Failure, unsuccessful selection
Relationships	N/A	
Use-Case Diagram	Administrator — Calculate and visualize estimated production	
Special Requirements	User may have understanding of categorical variables	

Post-Conditions	User may be verified with practical situation
Activity Diagram	

Database and tables

Cropinfo_Data DataBase

This is the main database that stores the information generated and modeled by the system. There are two main tables in this database.

1. Results Table (Model Builder Web Service)

This table stores the results for each users request for a crop prediction separately with a time stamp and a separate column for the particular users land extent.

2. Dynamic Table (Model Builder Web Service)

In this table the accuracy levels for each classification model is stored separately with a column for a timestamp to uniquely identify the request to build models (or a particular file upload)

Cropinfo_Web DataBase

1. User Registration Table (Web Interface)

Mainly stores the data relevant to the users registered in the website. There are three main user categories namely registered, super user and guest (public).

Java Classes and Methods

The project mainly consists of a web service written in Java and the following are the main classes in the project and their methods. The figure below shows an instance of the NetBeans IDE with the web service project (CropInfoWS) and the structure of the project and its packages.
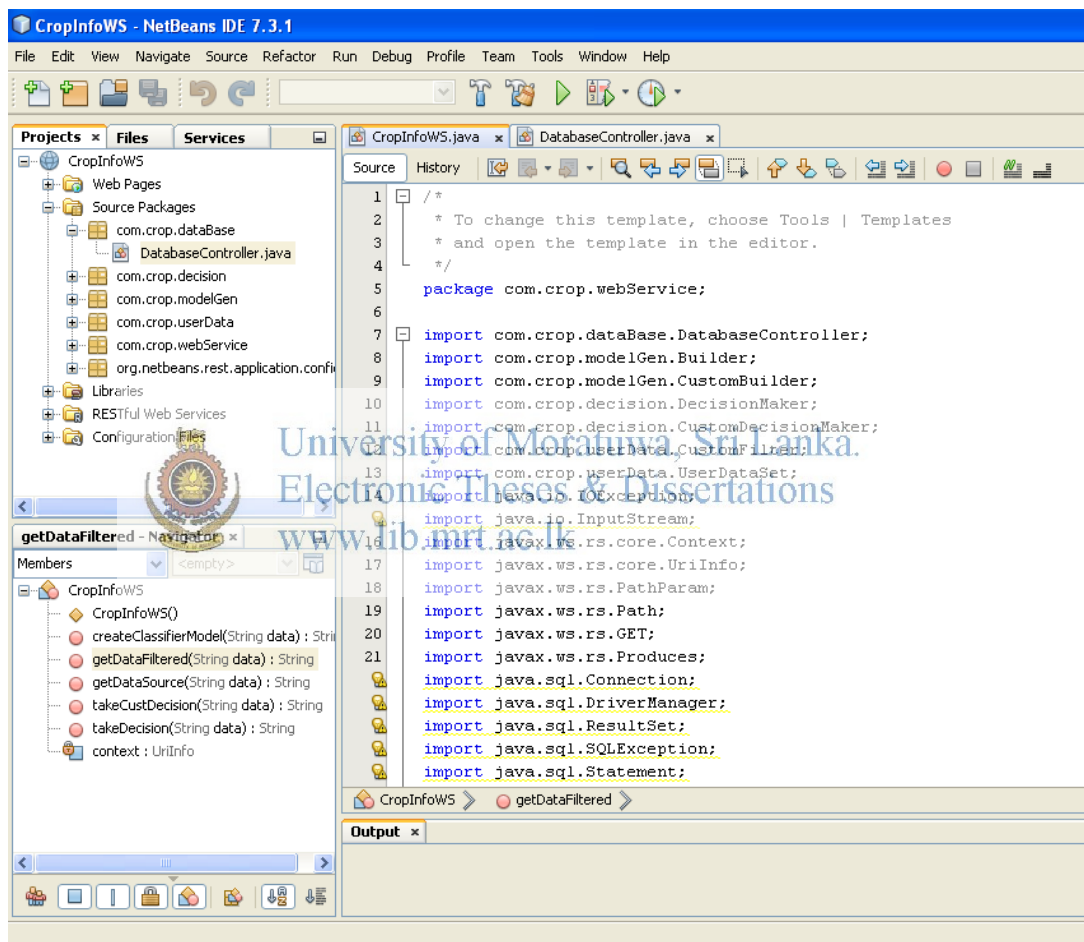
The figure below shows the class and package hierarchy and the project resources of the web service in tree from.

73
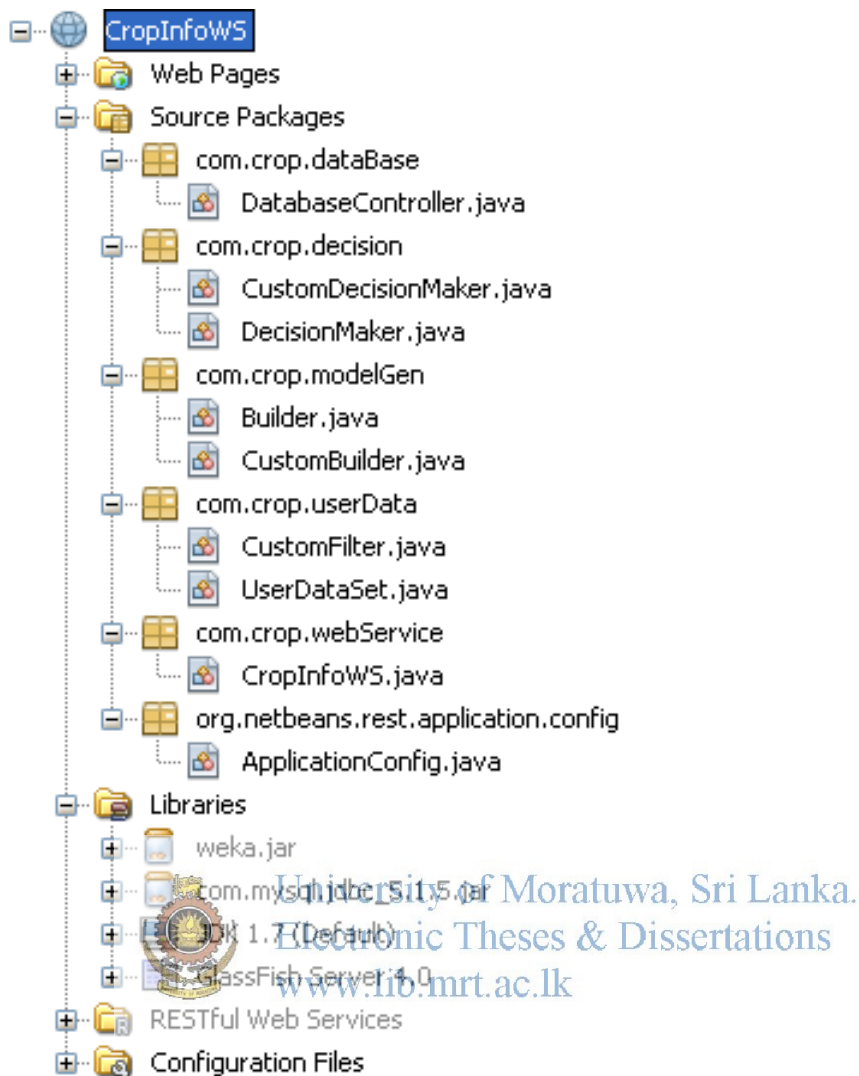
List of Java Classes

01	ApplicationConfig.java	Configuration settings of the Web service.
02	DataBaseController.java	Connect to the System Database(s) and make Database Instances and Connections for the use of other Classes and Methods.
03	DecisionMaker.java	Returns a crop prediction Level based on user inputs and default classifier model.
04	CustomDecisionMaker.java	Returns a crop prediction Level based on user inputs and user selected classifier model.
05	CustomFilter.java	Creates a filtered instance of a given dataset(uploaded file). The filtering options are set inside the class

06	Builder.java		To create a model of a classifier
	1	buildObj()	Builds a Bayes classifier for the default data file
07	CustomBuilder.java		To create a classifier model selected by the user
	1	buildObj()	Builds a Custom classifier for the default data file
08	UserDataSet.java		Dynamically creates classifier models for a data file uploaded by the user
	1	readDataFile()	Read the uploaded data file
	2	simpleClassify()	Build Classifier Models
	3	calculateAccuracy()	Calculate the accuracy of each model
	4	Results()	Output the results for each model
09	CropInfoWS.java		Communicate with the web application through a RESTful Web Service
	1	takeDecision ()	Generate prediction from Default Classifier Model
	2	CustomDecision ()	Generate prediction from user selected dataset and its build Classifier Model
	3	getDataSource ()	Upload and use user input custom Dataset
	4	getClassifierName ()	Generate classifier for user selected Type
10	Resources and Libraries used		weka.jar (weka libraries) Mysql_jdbc.jar (JDBC drivers library)

Web service Class

```java
/*
 * The main web service providing the predictions
 * with connected DB
 */
package com.crop.webService;

import com.crop.dataBase.DatabaseController;
import com.crop.modelGen.Builder;
import com.crop.modelGen.CustomBuilder;
import com.crop.decision.DecisionMaker;
import com.crop.decision.CustomDecisionMaker;
import com.crop.userData.CustomFilter;
import com.crop.userData.UserDataSet;
import java.io.IOException;
import javax.ws.rs.core.Context;
import javax.ws.rs.core.UriInfo;
import javax.ws.rs.PathParam;
import javax.ws.rs.Path;
import javax.ws.rs.GET;
import javax.ws.rs.Produces;
import java.text.DateFormat;
import java.text.SimpleDateFormat;
import java.util.Date;

@Path("cropinforoot")
public class CropInfoWS {

    @Context
    private UriInfo context;

    public CropInfoWS() {
    }

    @GET
    @Produces("text/plain")
    @Path("filter/{data}")
    // Method for filtering the provided Data Set
    public String getDataFiltered(@PathParam("data") String data) throws IOException,
Exception {

// copyright Code

    }
```

```java
@GET
@Produces("text/plain")
@Path("datasrc/{data}")
// Use the uploaded DataSet and Evaluate each classification model
public String getDataSource(@PathParam("data") String data) throws IOException,
Exception {

// copyright Code


}


@GET
@Produces("text/plain")
@Path("cremodel/{data}")
// Creates a new classifier model and Saves it in the disk for result prediction
public String createClassifierModel(@PathParam("data") String data) throws IOException,
Exception {

// copyright Code


}


@GET
@Produces("text/plain")
@Path("makedec/{data}")
// This Method retrives the decision for submitted params
public String takeDecision(@PathParam("data") String data) throws IOException,
Exception {

    // copyright Code


}


@GET
@Produces("text/plain")
@Path("custdec/{data}")
// This Method retrives the decision for submitted params
public String takeCustDecision(@PathParam("data") String data) throws IOException,
Exception {

// copyright Code

}
```

DecisionMaker Class

```
package com.crop.decision;

import java.io.BufferedInputStream;
import java.io.FileInputStream;
import java.io.InputStream;
import java.io.ObjectInput;
import java.io.ObjectInputStream;
import weka.classifiers.bayes.NaiveBayes;
import weka.core.Attribute;
import weka.core.FastVector;
import weka.core.Instance;
import weka.core.Instances;

/**
 *
 * @author Uditha
 */
public class DecisionMaker {
    private static final String SEPARATOR = "_";

    /*
     * This method takes the params submitted by the user
     * puts them in to a sample instance of the same data set and evaluates
     * the sample data set with the trained model of a given classifier
     *
     */
    public static String makeDecision(String data){

// copyright Code

    }
}
```

Builder Class

```
package com.crop.modelGen;

import java.io.BufferedReader;
import java.io.FileOutputStream;
import java.io.FileReader;
import java.io.IOException;
import java.io.ObjectOutputStream;
import weka.classifiers.Classifier;
import weka.classifiers.bayes.NaiveBayes;
import weka.core.Instances;

/**
 *
```

```
 * @author Uditha
 */
public class Builder {

    /*
     * This class is building a model for the given data file with a predetermined
     * classification algorithm  eg. Bayes
     * and then writes the file as a javaObject file ( serialized model)
     *
     */
    public static String buildObj (String s) throws IOException, Exception{
```

// copyright Code

```
    }
}
```

Filter Class

```
package com.crop.userData;

import java.io.BufferedReader;
import java.io.FileOutputStream;
import java.io.FileReader;
import java.io.IOException;
import java.io.ObjectOutputStream;
import weka.core.Instances;
import weka.filters.Filter;
import weka.filters.unsupervised.attribute.Discretize;

/**
 *
 * @author Uditha
 */
public class CustomFilter {

    public static String filterData(String s) throws IOException, Exception{
```

// copyright Code

```
    }

    }

}
```

Page Structure of the Web Interface

Home Page

User Registration

Default Prediction Page

Custom File Uploads Page

Classifier Selection Page

Prediction Data Input Page

Results Page

Joomla Admin Panels