TRIBHUVAN UNIVERSITY

INSTITUTE OF SCIENCE AND TECHNOLOGY

COMPARATIVE ANALYSIS OF DECISION TREE METHODS FOR THE PREDICTION OF PADDY PRODUCTIVITY

DISSERTATION

Submitted to

Central Department of Computer Science and Information Technology

Kirtipur, Kathmandu, Nepal

In Partial Fulfillment of the Requirements for

Master's Degree in Computer Science & Information Technology

Submitted by

Mr. Chaturbhuj Bhatt

Under the Supervision of

Prof. Dr. Subarna Shakya

Department of Electronics and Computer Engineering

Institute of Engineering, Pulchowk, Nepal

Under the co-supervision of

Asst. Prof. Tej Bahadur Shahi

Central Department of Computer Science and Information Technology

Kirtipur, Kathmandu, Nepal

November, 2019

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

STUDENT'S DECLARATION

I hereby declare that I am the only author of this work and that no sources other than listed here have been used in this work.

_____

Mr. Chaturbhuj Bhatt

Date: ……

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

SUPERVISOR'S RECOMMENDATION

We hereby recommend that this dissertation prepared under our supervision by Mr. Chaturbhuj Bhatt titled "**Comparative Analysis of Decision Tree Methods for the Prediction of Paddy Productivity**" in partial fulfillment of the requirement for the degree of M. Sc. in Computer Science and Information Technology be processed for the evaluation.

_____

Prof. Dr. Subarna Shakya

Department of Electronics and Computer Engineering

Institute of Engineering, Pulchowk, Kathmandu, Nepal

(Supervisor)

Date: ……

_____

Asst. Prof. Tej Bahadur Shahi

Central Department of Computer Science and Information Technology

Kirtipur, Kathmandu, Nepal

(Co-supervisor)

Date: ……

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

LETTER OF APPROVAL

We certify that, we have read this dissertation and, in our opinion, it is satisfactory in the scope and quality as a dissertation in partial fulfillment of the requirement for the degree of M. Sc. in Computer Science and Information Technology.

Evaluation Committee

Asst. Prof. Nawaraj Poudel	**Prof. Dr. Subarna Shakya**
Central Department of CSIT,	Department of Electronics and
Tribhuvan University	Computer Engineering
Kathmandu, Nepal	IOE, Pulchowk, Kathmandu, Nepal
(Head of Department)	**(Supervisor)**

(External Examiner)	**(Internal Examiner)**

ACKNOWLEDGEMENT

I am highly indebted to my thesis supervisor, Prof. Dr. Subarna Shakya, Department of Electronics and Computer Engineering Institute of Engineering, Pulchowk, and Co-supervisor Asst. Prof. Tej Bahadur Shahi, Central Department of Computer Science and Technology, Kirtipur, Kathmandu, for their valuable and constructive suggestion during the planning and development of this research. Otherwise it would have never seen the light of the day. Their willingness to give their time so generously has been very much appreciated.

Also, my gratitude goes to Asst. Prof. Nawaraj Poudel, Head of Central Department of Computer Science and Technology, Kirtipur, Kathmandu and Asst. Prof. Jagdish Bhatta, Central Department of Computer Science and Technology, Kirtipur, Kathmandu.

My special thanks to Mrs. Sanju Rimal Joshi (Senior Agriculture Officer, Government of Nepal) of Pradhanmantri Krishi Adhunikaran Pariyojana (Paddy Super Zone), Kanchanpur, Nepal, and all the farmers of Kanchanpur district for participation during data collection.

I would also like to thank Mr. Bhupendra Singh Saud, Mr. Deepak Bhatt, Mr. Indra Chaudhary, all my colleagues and best wishers who exhorted me for the initiation.

Finally, I would like to thank my family members for their love and blessings. Without whom I may not make any sense to this research work. What I am today is because of their love, guidance and support.

ABSTRACT

Data mining applications has got rich focus due to its significance of classification algorithms. The agricultural data is difficult to study. The challenge from a research perspective is to identify the key attributes that determine paddy performance across different farming situations such as geographic location, soil types, and seasonal conditions. This study aims to survey on the two different decision tree algorithms with primary dataset collected in Kanchanpur district and to implement as well as assist by comparing J48 and SimpleCart decision tree methods to predict the production of paddy. From the result analysis it was seen that SimpleCart was able to classify 80.198% of the data correctly which was better than J48 in comparison to results of evaluation metrics (Accuracy, Precision, Recall and F-Measure). In a nut shell, the experiment result showed that J48 has got smaller tree size than SimpleCart but SimpleCart has got 1.9802% better accuracy than J48 for the prediction of paddy productivity.

Keywords: CART, Classification, C4.5, Data Mining, Decision Tree, J48, Paddy productivity, SimpleCart, Supervised Machine Learning

TABLE OF CONTENTS

LIST OF TABLE

LIST OF FIGURE

LIST OF ABBREVIATION

API	:	Application Programming Interface
ARFF	:	Attribute Relation File Format
CART	:	Classification and Regression Tree
CSV	:	Comma-Separated Value
DM	:	Data Mining
FN	:	False Negative
FP	:	False Positive
GNU	:	General Public License
GPS	:	Global Positioning System
IBLE	:	Identity-Based Lossy Encryption
ID3	:	Iterative Dichotomiser
KDD	:	Knowledge Discovery from Data
PCU	:	Primary Care Unit
RD	:	Reduced Table
RNR	:	Recursion Noise Removal
SVM	:	Support Vector Machine
TN	:	True Negative
TP	:	True Positive
WEKA	:	Waikato Environment for Knowledge Analysis

CHAPTER 1
INTRODUCTION

1.1. Background of the Study

In our daily life there are lots of data in different fields. Whenever there is data, we can have lots of information, patterns, meaning etc. and information is an important asset for an organization during this competitive global market. The information can be stored in computer in the form file, database or data warehouse. Moreover, this information helps us to extract knowledge for decision making. Good decision-making process helps us for identifying, selecting, and implementing alternatives. The right information, in the right form, at the right time is needed to make good decisions. The process of extracting or "mining" knowledge from large amount of data is called Data Mining [1]. Data mining also can be defined as exploration and analysis of large quantities of data to discover meaningful pattern from data and is also known as "Knowledge Discovery from Data (KDD)" [1].

Decision Tree is also the most widely applied supervised machine learning or classification technique. The learning and classification steps of decision tree induction are simple and fast and it can be applied to any domain [2].

Information on new paddy varieties is important to farmers when assessing whether to adopt these varieties. This information can be used as part of the farmer's decision-making process to help to improve paddy production. Often changing to a newer paddy variety will result in greater yields with little or no change in farm resource outlays. Thereby, it is important to both the farmer and seed marketer that this variety information is accurate. Data mining is the extraction of hidden diagnostic information from huge databases to help companies focus on the most significant information in their data warehouses. Data mining tasks predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, potential analysis presented by data mining move past the investigation of past procedures provided by retrospective tools typical of decision support systems [3].

1.2. Statement of Problem

Data mining applications have got rich focus due to its significance of classification algorithms. The comparison of classification algorithm is a complex task and it is an open problem. First, the notion of the performance can be defined in many ways: accuracy, speed, cost, reliability etc. Second, an appropriate tool is necessary to quantify this performance. Third, a consistent method must be selected to compare with the measured values.

The agricultural data is difficult to study. The challenge from a research perspective is to identify the key attributes that determine paddy performance across different farming situations such as geographic location, soil types, and seasonal conditions. The key attributes include nutrition and soil type, grain yield and quality, sowing and harvest dates, tolerance to environmental stresses, measurement inaccuracy, sampling discrepancy, outdated data sources and other errors. Due to vagaries of climate factors the agricultural productivities are continuously decreasing over a decade. Little research work had addressed the issue of mining uncertain data. The traditional system had some drawbacks. Irrelevance of the delivered information, inability of the system to cover all farmers, lack of avenues to improve performance, unaccountability regarding advice given by the system, etc. are some of the problems.

1.3. Objectives of the Study

The objectives of this research are:

- To prepare the paddy data for the Prediction of Paddy Productivity.
- To compare two decision tree algorithms (J48 and SimpleCart) for the Prediction of Paddy Productivity.

1.4. Limitations of the Study

Limitations of this research is that it focused only on agriculture data of Kanchanpur district with sampling of 101 farmers. Each farmer has at least 5 Kattha land area and maximum of 60 Kattha (3 Bigha) land area productivity data had been collectively – which are the basis of all experiment done in this study.

1.5. Structure of the Report

This report is organized in five chapters and is enlisted below:

- Chapter 1 **"Introduction"** explains the background of the study, statement of problems, objectives of the study as well as limitations of the study.
- Chapter 2 **"Literature Review"** describes the various concepts of data mining, and related works in our domain.
- Chapter 3 **"Research Methodology"** explains the framework of the research and implemented algorithms.
- Chapter 4 **"Experiment and Result Analysis"** explains about experiments, results analysis.
- Chapter 5 **"Conclusion and Future Works"** describes the conclusion and future works for the upcoming researcher.

CHAPTER 2
LITERATURE REVIEW

2.1. Data Mining

In our daily life; there are lots of data in different fields. Whenever there is data, we can have lots of information, patterns, meaning etc. and information is an important asset for an organization during this competitive global market. Moreover, this information helps us to extract knowledge for decision making. Good decision-making process helps us for identifying, selecting, and implementing alternatives. The right information, in the right form, at the right time is needed to make good decisions. The process of extracting or "mining" knowledge from large amount of data is called data mining [1]. Data mining also can be defined as Exploration and analysis of large quantities of data to discover meaningful pattern from data and is also known as "Knowledge Discovery from Data (KDD)" [1].

In data mining [1] there are lots of techniques to mine the knowledge from data which are recently used widely in different fields such as Business, Scientific Research, Computer Science, Machine Learning, Information Science, Statistics, and Database Technology etc. Most commonly used data mining techniques are Classification,Dependencies and Associations,Regression and Clustering. These above-mentioned techniques are effectively used in different fields separately.

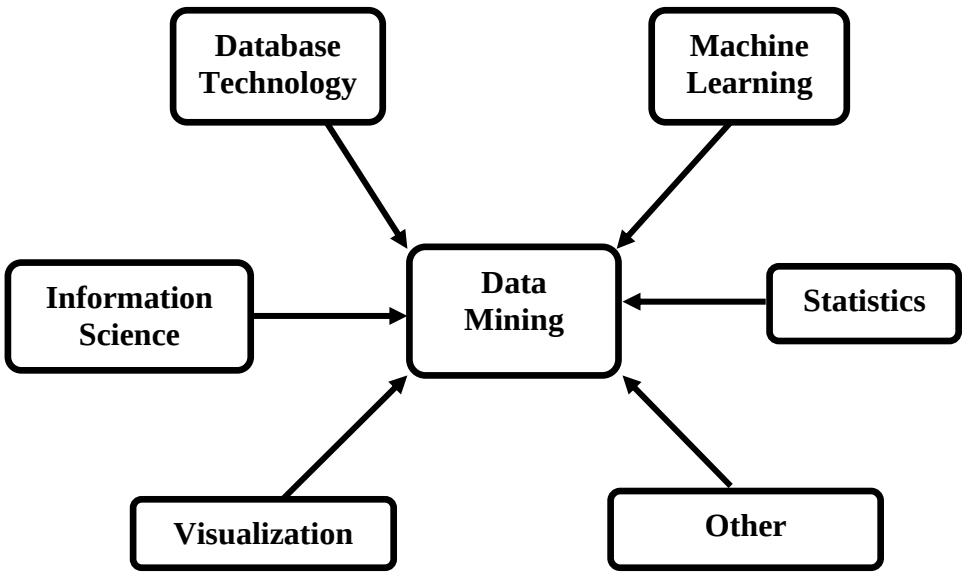

Figure 2.1: Data mining as confluence of multiple disciplines

2.1.1. Machine Learning

Machine learning investigates how computers can learn or improve their performance based on data. It is the main research area is for computer programs to automatically learn to recognize complex patterns and make intelligent decision based on the data. For example: a system that can automatically recognize hand written postal codes on mail after learning from a set of examples. Machine learning are sub divided into two parts i.e. supervised learning and unsupervised learning.

- **Supervised Learning**

Supervised learning is fairly common in classification problems because the goal is often to get the computer to learn a classification system that we have created. Digit recognition, once again, is a common example of classification learning. More generally, classification learning is appropriate for any problem where deducing a classification is useful and the classification is easy to determine. Supervised learning is the most common technique for training neural networks and decision trees. Both of these techniques are highly dependent on the information given by the pre-determined classifications [4].

- **Unsupervised Learning**

Unsupervised learning seems much harder: the goal is to have the computer learn how to do something that we don't tell it how to do! There are actually two approaches to unsupervised learning. The first approach is to teach the agent not by giving explicit categorizations, but by using some sort of reward system to indicate success. Note that this type of training will generally fit into the decision problem framework because the goal is not to produce a classification but to make decisions that maximize rewards. This approach nicely generalizes to the real world, where agents might be rewarded for doing certain actions and punished for doing others. A second approach is called clustering. In this type of learning, the goal is not to maximize a utility function, but simply to find similarities in the training data [4].

2.1.2. Classification

Classification or prediction is the most widely used data mining task. Classification algorithms are supervised methods that uncover the hidden relationship between the target class and the independent variables [5]. Supervised learning algorithms allow

labels to be assigned to the observations so that new data can be classified based on training data [1, 5]. Examples of classification tasks are image and pattern recognition, medical diagnosis, loan approval, detecting faults or financial trends [5].

Classification is a two-step process and they are:

1. **Model Construction (Learning step or Training Phase):** This step builds a model to explain the target concept and is represented as classification rules, decision trees, or mathematical formulae

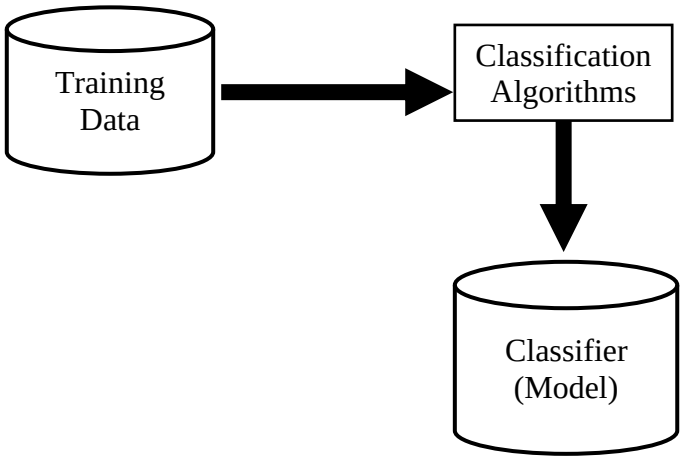

Figure 2.2: Model Construction step for classification process [1]

2. **Using the Model in Prediction (Testing Phase):** This step is used for classifying future or unknown cases and estimate the accuracy of the model.

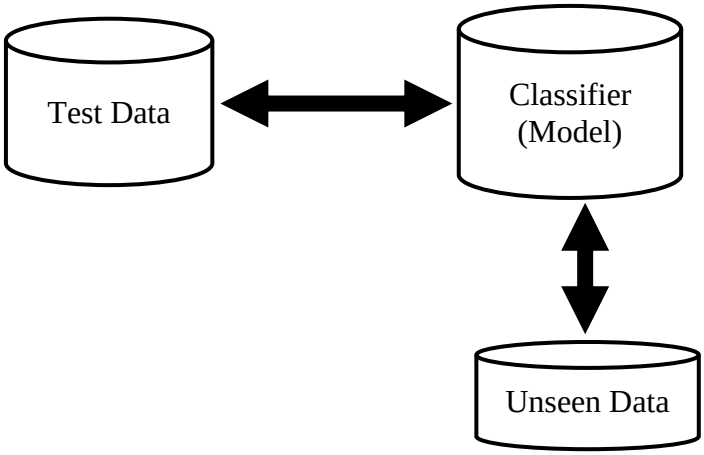

Figure 2.3: Using the Model in prediction step for classification process [1]

Decision Tree

A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and

each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. A typical decision tree is shown in figure 2.4.

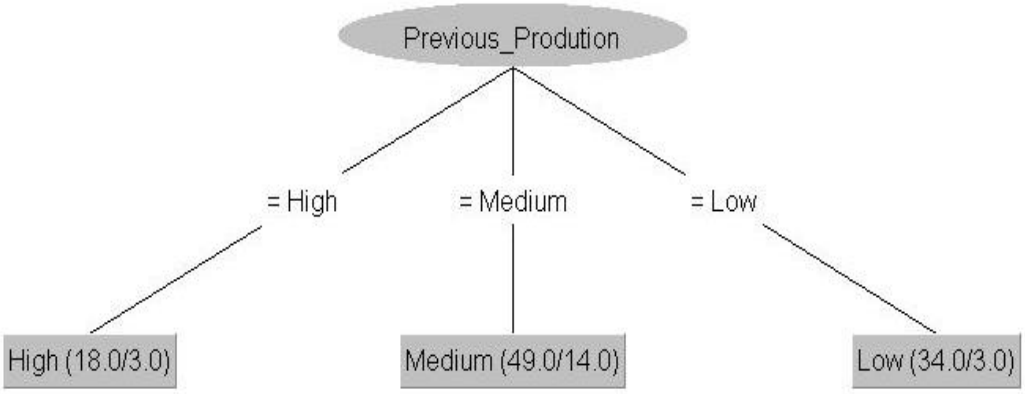

Figure 2.4: Decision tree example

During the late 1970s and early 1980 J. Ross Quinlan, a researcher in machine learning, developed a decision tree algorithm known as ID3 (Iterative Dichotomiser). Quinlan later presented C4.5[6, 7] (a successor of ID3), which become a benchmark to which newer supervised learning algorithms are often compared. In 1984, a group of statisticians published the book classification and regression trees (CART)[7], which described the generation of binary decision trees. ID3 and CART were invented independently of one another at around the same time, yet follow a similar approach for learning decision trees from training tuples. These two cornerstone algorithms spawned a flurry of work on decision tree induction. The basic decision tree algorithm is summarized as below:

- **Decision Tree Construction Algorithm**

Input: A data set, D

Output: A decision tree

- If all the instances have the same value for the target attribute then return a decision tree that is simply this value (not really a tree - more of a stump).
- Else
    1. Compute Gain values for all attributes and select an attribute with the highest value and create a node for that attribute.
    2. Make a branch from this node for every value of the attribute
    3. Assign all possible values of the attribute to branches.

4. Follow each branch by partitioning the dataset to be only instances whereby the value of the branch is present and then go back to 1.

- **Attribute Selection Measures**

In a data set there are lots of attributes and we do have problem on selection of attribute as node and attribute as leaf. There arise questions which attribute first? attribute selection measure [1] is a heuristic for selecting the splitting criterion that "best" separates given data partition, D, of class-labeled training tuples into individual classes. Attribute selection measures are also known as splitting rules because they determine how the tuples at a given node are to be split. The attribute selection measure provides a ranking for each attribute describing the given training tuples. The attribute having the best score for the measure is chosen as the splitting attribute for the given tuples.

- **Information Gain**

ID3 uses information gain as its attribute selection measure. This measure is based on pioneering work by Claude Shannon on information theory, which studied the value or "information content" of messages [1].

Information gain = (information before split) – (information after split) bits

$$Gain(A) = Info(D) - Info_A(D)\ bits$$ ---------- Equation 2.1

Where,

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)\ bits$$ ---------- Equation 2.2

- $P_i = |C_{i,D}| / |D|$
- A having v distinct value, $\{a_1, a_2, \ldots, a_v\}$
- $D_1, D_2, \ldots, D_v$ then,

$$Info_A(D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} \times Info(Dj)\ bits$$ ---------- Equation 2.3

- **Gain Ratio:**

The information gain [1] measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values. C4.5[6, 7], a successor

of ID3, uses an extension to information gain known as gain ratio, which attempts a overcome this bias. It applies a kind of normalization to information gain using a "Split information" value define analogously with $Info(D)$ as

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} \times \log_2\left(\frac{|Dj|}{|D|}\right) bits$$ ---------- Equation 2.4

The gain ratio is defined as

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)}$$ ---------- Equation 2.5

The attribute with the maximum gain ratio is selected as the splitting attribute. Note, however, that are the split information approaches 0, the ratio becomes unstable. A constraint is added to avoid this, whereby the information gain of the test selected must be large-at last as great as the average gain over all tests examined.

- **Gini Index**

The Gini index [1] is used in CART [7]. Using the notation previously described, the Gini index measures the impurity of D, a data partition or set of training tuples, as

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \; bits$$ ---------- Equation 2.6

Where, P_i is the probability that a tuple in D belong to class C_i and is estimated by $|C_{i,D}| / |D|$. The sum is computed over m classes. The Gini index considers a binary split for each attribute. Let's first consider the case where A is a discrete-valued attribute having *v* distinct values, $\{a_1, a_2, ..., a_v\}$, occurring in D. If A has v possible values, then there are 2^v possible subsets but we exclude the power set, and the empty set from consideration since, conceptually, they do not represent a split. Therefore, there are 2^v-2 possible ways to form two partitions of the data, D, based on a binary split on A.

When considering split, we compute a weighted sum of the impurity of each resulting partition. For example, if a binary split on A partitions D into D_1 and D_2, the Gini index of D given that partitioning is

$$Gini_A(D) = \left\{\frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)\right\} bits$$ ---------- Equation 2.7

For each attribute, each of these possible binary splits is considered. For discrete-valued attribute, the subset that gives the minimum Gini index for that attribute is selected as its splitting subset.

The reduction in impurity that would be incurred by a binary split on a discrete-or continuous-valued attribute A is

$$\Delta Gini(A) = \{Gini(D) - Gini_A(D)\} bits$$ ---------- Equation 2.8

The attribute that maximizes the reduction in impurity (or, equivalently, has the minimum Gini index) is selected as the splitting attribute.

2.2. Application Programming Interface for Data Mining

Application Programming Interface (API) [8] is a set of routines used by an application program to direct the performance of procedures by the computer's operating system. To achieve the goal of this research, the most commonly used API for data mining, WEKA is used.

2.2.1. WEKA

WEKA [9] is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your 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 [10].

WEKA was developed at the University of Waikato in New Zealand; the name stands for *Waikato Environment for Knowledge Analysis*. (Outside the university, the WEKA, pronounced to rhyme with *Mecca*, is a flightless bird with an inquisitive nature found only on the islands of New Zealand). The system is written in Java and distributed under the terms of the General Public License (GNU). It runs in almost any platform and has been tested under Linux, Windows and Macintosh operating systems- and even on a personal digital assistant [11]. WEKA's native data storage method is *Attribute-Relation File Format* (ARFF) [12].

2.3.Related Works

In recent years, several models for the simulation of soil dynamics have been developed. Data mining techniques are often used to study soil characteristics. As an example, the k-means approach is used for classifying soils in combination with GPS-based technologies, k-means approach to classify soils and plants and SVMs to classify crops [13].

Studies conducted by agricultural researchers in Pakistan have shown that attempts of crop yield maximization through pro-pesticide state policies have led to a dangerously high pesticide usage. These studies have reported a negative correlation between pesticide usage and crop yield. Hence excessive use of pesticides is harming the farmers with adverse financial, environmental and social impacts. Study had shown that how data mining integrated agricultural data including pests counting, pesticide usage and meteorological recordings is useful for optimization of pesticide usage. Unsupervised clustering of the data was performed first through Recursive Noise Removal (RNR). These clusters reveal interesting patterns of farmer practices along with pesticide usage dynamics and hence help identify the reasons for this pesticide abuse [14].

Influence of climatic factors on major Kharif and Rabi crops production in Bhopal District of Madhya Pradesh State was studied. The findings of the study revealed that the decision tree analysis indicated that the productivity of soybean crop was mostly influenced by Relative humidity followed by rainfall and temperature. The decision tree analysis indicated that the productivity of paddy crop was mostly influenced by Rain fall followed by Relative humidity and Evaporation. A web based expert information system based on ID3 algorithm was studied in which an expert system provides advisory services to Tomato growers regarding pests, diseases and their control measures[15].

The web-based system has also provision for the growers to interact with other growers on the management practices of tomato crop cultivation. An advanced version of decision making tree algorithm IBLE, it mainly used in the information theory [16]. The channel capacity concept to take chooses the important characteristic

to the entity in the measure. Combine the rule with many characteristics the point to distinguish the example can effectively the correct distinction. They applied this algorithm in the oral cavity disease diagnosis; the experimental result indicated this algorithm has the very strong recognition capability to agriculture case diagnosis to very good assistance diagnosis function.

CHAPTER 3

RESEARCH METHODOLOGY

3.1. Background

This chapter deals with the framework of research and used algorithms.

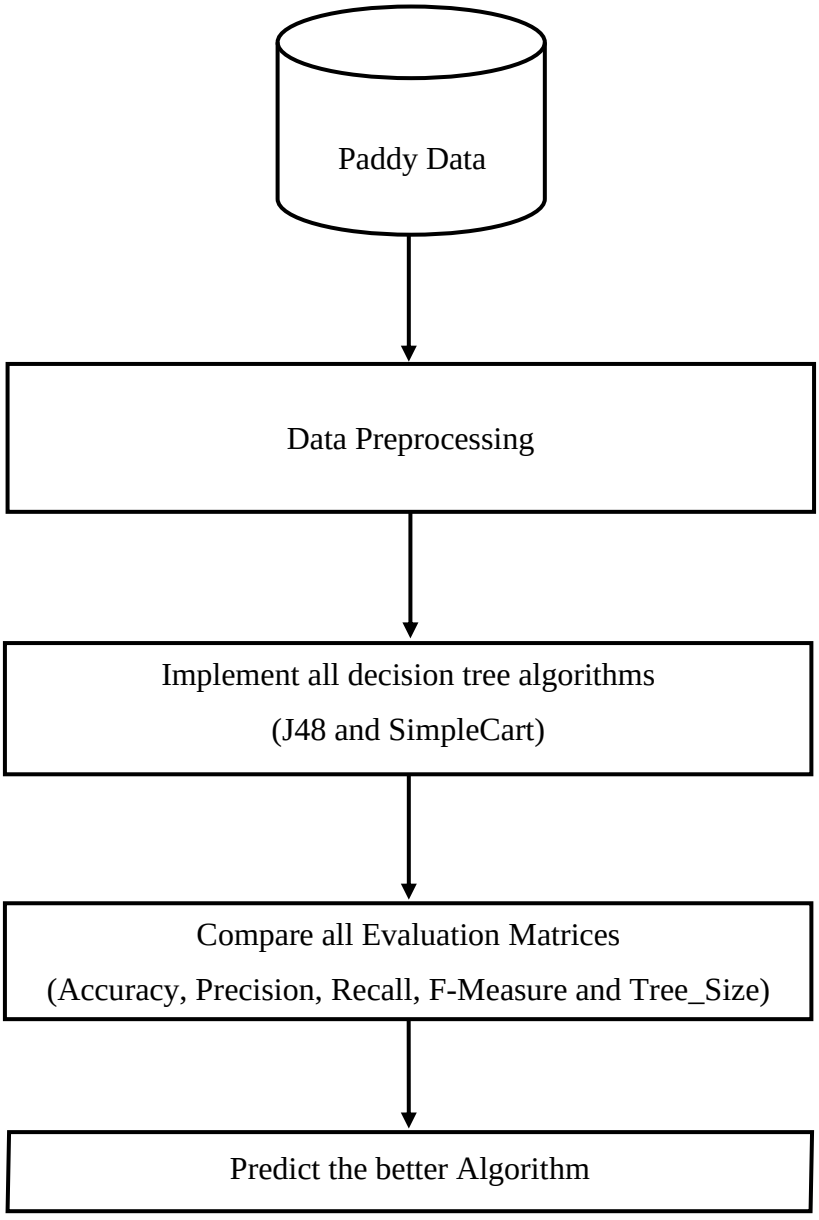

Figure 3.1: Implementation Model

3.2. Algorithms

In this research, two decision tree algorithms were implemented and they are

a) J48 b) SimpleCart

3.2.1. J48

J48 is a WEKA implementation of C4.5 [6] decision tree classifier algorithm, which uses Gain ratio for the attribute selection.

- **Algorithm**

Input: A data set, D

Output: A decision tree by J48

- If all the instances have the same value for the target attribute then return a decision tree that is simply this value (not really a tree - more of a stump).
- Else
    1. Compute Gain ratios for all attributes and select an attribute with the highest value and create a node for that attribute.
    2. Make a branch from this node for every value of the attribute
    3. Assign all possible values of the attribute to branches.
    4. Follow each branch by partitioning the dataset to be only instances whereby the value of the branch is present and then go back to 1.

3.2.2. SimpleCart

SimpleCart is a WEKA implementation of CART [7] (Classification and Regression Tree) algorithm, which uses Gini index as attribute selection metric and employs the minimal cost-complexity pruning strategy [11].

- **Algorithm**

1. Establish Classification Attribute (in Table D)
2. Compute Gini index for all the attributes of the table.
3. Select Attribute with the minimum Gini Index to be the next Node in the tree (Starting from the Root node).
4. Remove Node Attribute, creating reduced table (RD).
5. Repeat steps 3-5 until all attributes have been used, or the same classification value remains for all rows in the reduced table.

3.3. Source of Data

Source of data was primary source which was collected from the survey among 101-farmers of Bheemdatta Municipality, Ward No. 2, 10, 18, and Krishnapur Municipality, Ward No. 3 of Kanchanpur district, Mahakali Zone, Nepal.

3.4. Experimental Setup

Experiment had been done in WEKA version 3.8.3.

Table 3.1: Experimental Parameters of J48 and SimpleCart

Scheme1: weka.classifiers.trees.J48 *Scheme2: weka.classifiers.trees.SimpleCart*
Relation: paddy_data_V3.csv
Test mode:4-fold cross-validation

CHAPTER 4
EXPERIMENT AND RESULT ANALYSIS

4.1. Background

This section deals with the successful implementation and comparative analysis of J48 and SimpleCart for the prediction of paddy productivity. The experiments were performed in WEKA version 3.8.3 installed in system consist of 1.8 GHz AMD A9-9420e RADEON R5, 5 COMPUTE CORES 2C+3G with 4 GB RAM in Windows 10 (64-bit) Operating System.

4.2. Tools

Algorithms can be compared using many data mining tools. However in this research WEKA version 3.8.3 had been used for simulation.

4.3. Data Structure

The main data structures used in this study are enlisted below:

Table 4.1: Dataset Description

S. No.	**Attributes**	**Attribute Type**	**Description**
1	Seed_Name	Nominal	Name of Seed used by framer for paddy cropping
2	Seed_Source	Nominal	Source type of seed taken by farmer and they are: local, agrovet, agricultural and mixed
3	Seed_Amount	Numerical	Quantity (Kg) of seed taken by farmer.
4	Prod_Area	Numerical	Production Area (Kattha) used for cropping
5	Soil_Type	Nominal	Nature of Soil on production area and they are: pango, sandy and both
6	Irrigation_used	Nominal	Source of irrigation used by farmer during cropping and they are: motor (M),

			Rain (R), Canal (C), Motor + Rain (MR)
7	Fertilizer_used	Nominal	Types of fertilizer used are: Organic, Chemical, and Both
8	Labour_used	Nominal	Types of labour used are: Human, Machine, Both
9	Disease_Insect	Nominal	Disease or Insect found in crop are: Khaire, Gabero, Gandi, Suke, Paterol, etc.
10	Pestiside_used	Nominal	Types of pesticides used in paddy are: organic, chemical and no
11	Harvest_duration	Nominal	Types of harvest duration of paddy are: Long (above 130 days), Medium (between 110 and 131 days), short (below 110 days)
12	Expenses	Numeric	Total expense (NRs) during cropping
13	Previous_Production	Nominal	Types are: Low (below 1001 Kg, Medium (between 1000 and 2000 Kg), and High (above 1999 Kg).
14	Current_Production	Nominal	Types are: Low (below 1001 Kg, Medium (between 1000 and 2000 Kg), and High (above 1999 Kg).

4.4. Data Samples

Sample of dataset used in the research is shown below:

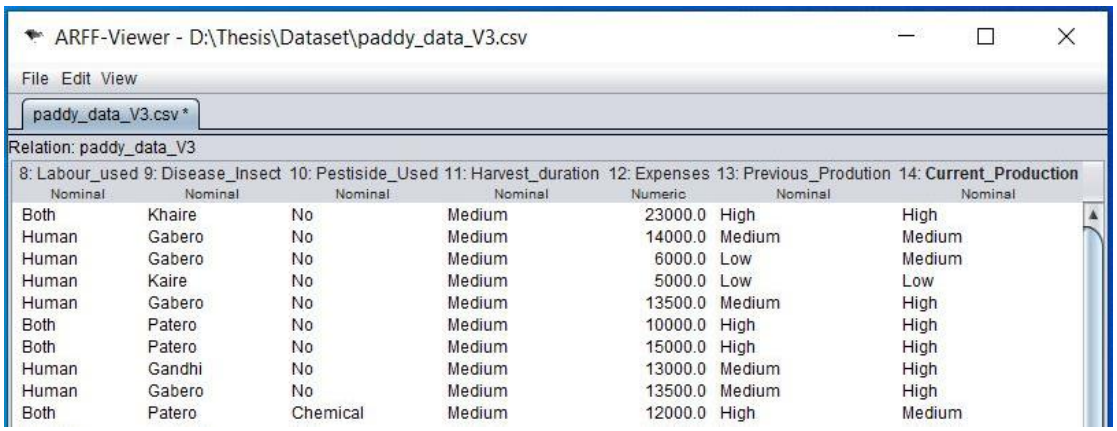

Figure 4.1: Portion of dataset paddy_dataV3.csv viewed in WEKA ARFF-Viewer

4.5. Experiments and Results

In this section, each steps of the methodology was implemented for simulation and results were described.

4.5.1. Experiments

The output showed in figure 4.2 and figure 4.3 explains the run information (Scheme, Relation, Instances, Attributes and Test mode).

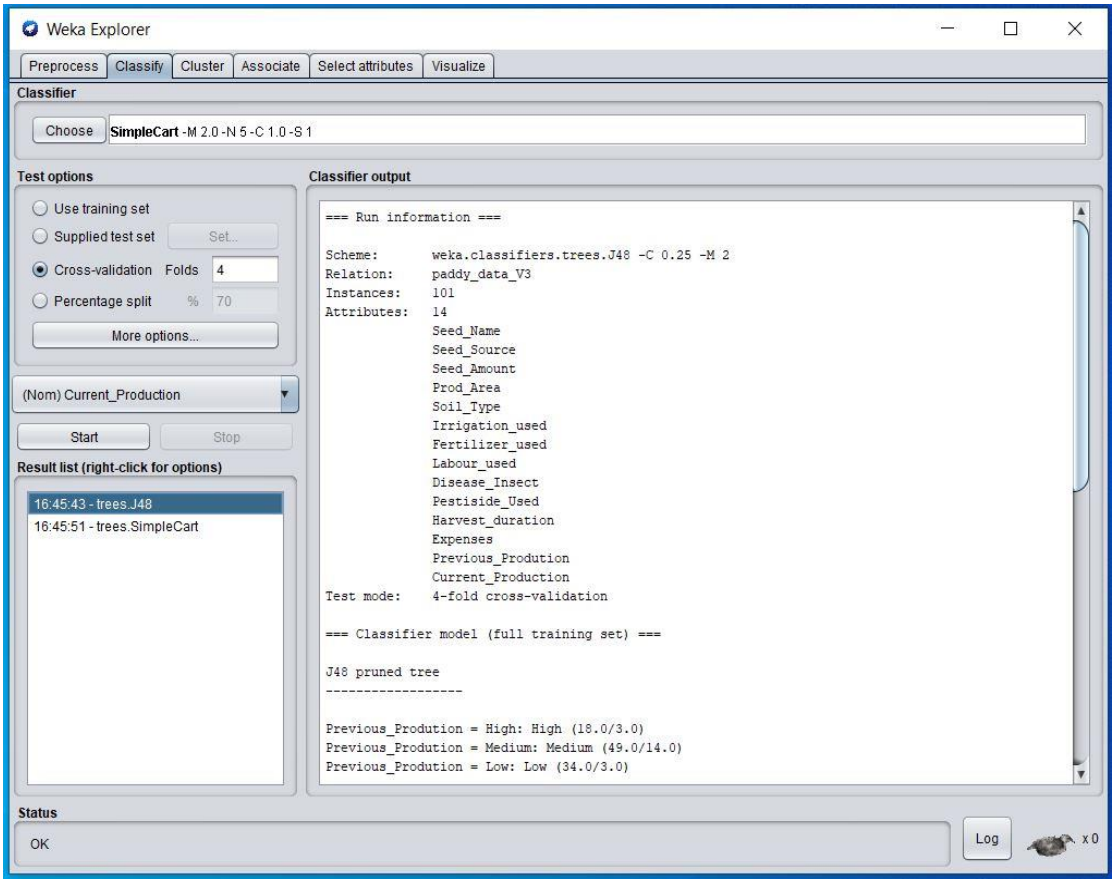

Figure 4.2: Result of J48 algorithm

For the classification the most important figures here are the number of correctly and incorrectly classified instances. The output from the WEKA program is shown in the figure 4.4 and figure 4.5. In these outputs, J48 was able to classify 78.2178 % of the data correctly and SimpleCart was able to classify 80.198 % of the data correctly.

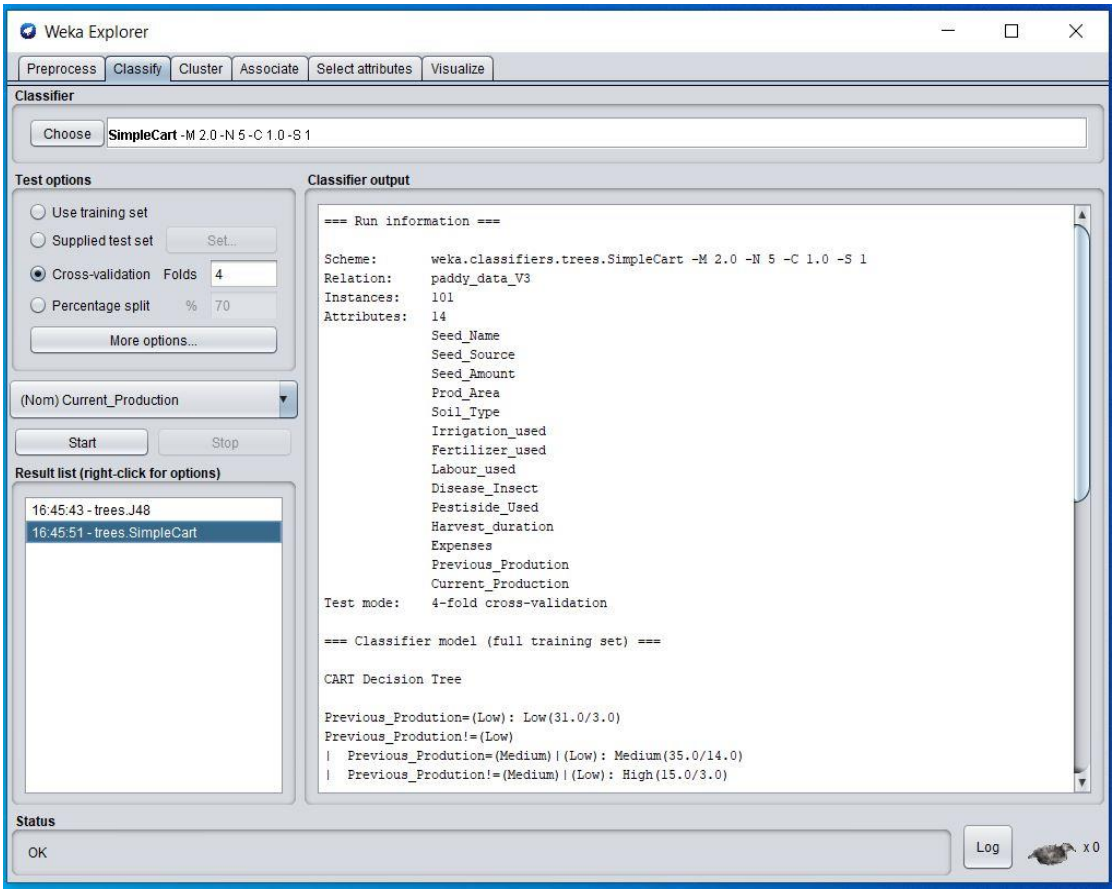

Figure 4.3: Result of SimpleCart algorithm

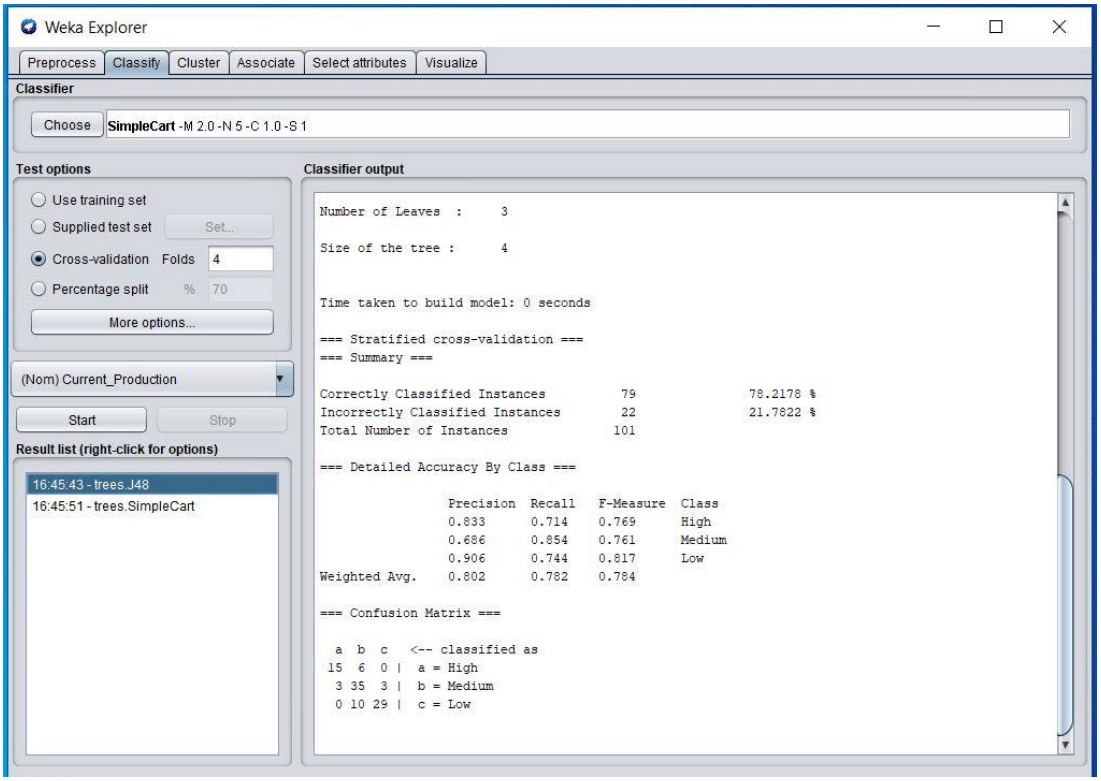

Figure 4.4: Classified instances by J48 algorithm

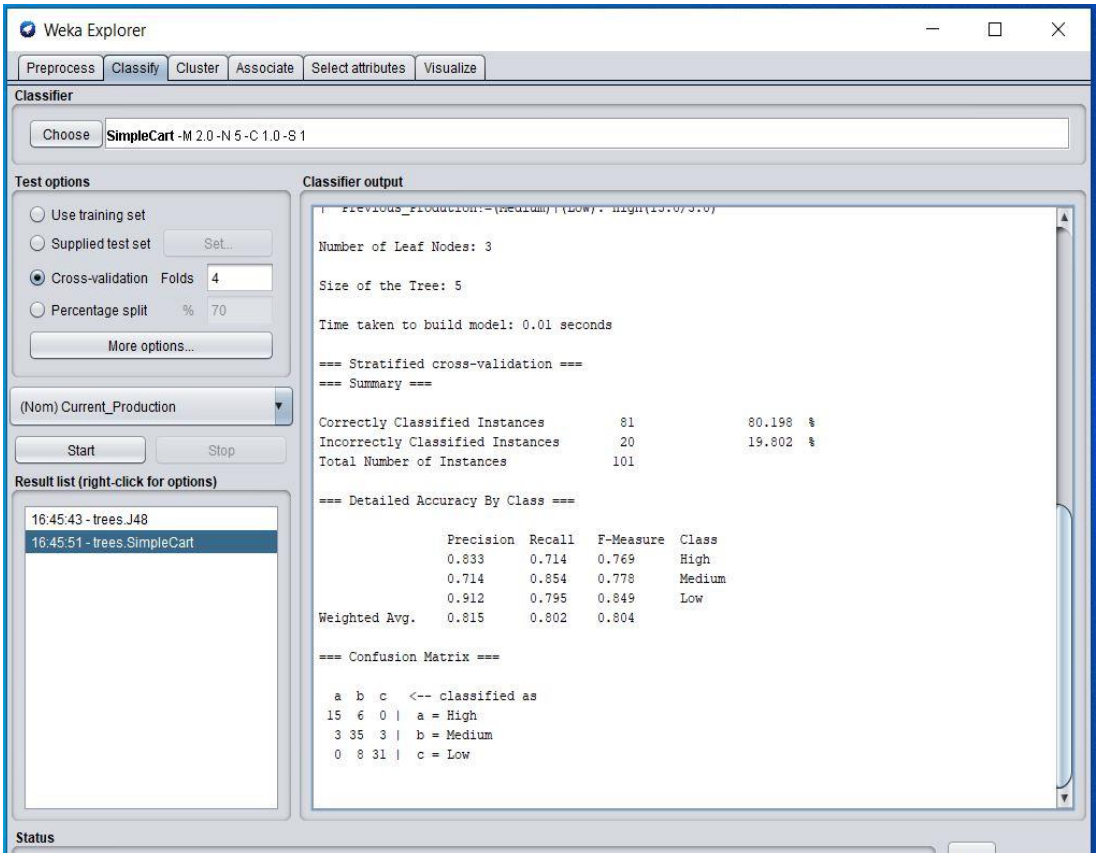

Figure 4.5: Classified instances by SimpleCart algorithm

4.5.2. Evaluation

For the comparison of two different classification algorithms followings are evaluation metrics.

- **4-fold Cross Validation**

In **4-fold cross-validation**, the initial data are randomly partitioned into 4 mutually exclusive subsets or "folds" i.e. D_1, D_2, D_3, D_4 each of approximately equal size. Training and testing is performed 4 times in the ratio of 3:1 means to say 3 fold as Training and 1 fold as Testing.

- **Confusion Matrix**

A confusion matrix is a table for analyzing the result of the classifiers. It deals with how classifier can recognize tuples of different classes. In order to develop the confusion matrix, the following terms should be considered:

- **True Positive (TP):** Positive tuples that are correctively labelled by the classifier.
- **True Negative (TN):** Negative tuples that are correctly labelled by the classifier.

- **False Positive (FP):** Negative tuples that are incorrectly labelled as positive.
- **False Negative (FN):** Positive tuples that are mislabelled as negative.

		Predicted Class		
		Yes	No	Total
Actual Class	Yes	**TP**	**FN**	**P**
	No	**FP**	**TN**	**N**
	Total	**P′**	**N ′**	**P+N**

Figure 4.6: Confusion Matrix

➢ **Accuracy**

Accuracy of a classifiers on a given test set is the percentage of test set tuples that are correctly classified by the classifiers. It also refers to the recognition rate of the classifier that means how the classifier recognizes tuples of the various classes.

$$\text{Accuracy} = \frac{TP + TN}{P + N}$$ ---------- Equation 4.1

➢ **Precision**

Precision refers to the measure of exactness that means what percentage of tuples labeled as positive are actually such.

$$\text{Precision} = \frac{TP}{TP + FP}$$ ---------- Equation 4.2

➢ **Recall**

Recall refers to the true positive rate that means the proportion of positive tuples that are correctly identified. It is also known as sensitivity of the classifier.

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{TP}{P}$$ ---------- Equation 4.3

➢ **F-Measure**

The F-Measure also refers to F_1-scorewhich combines both the measures i.e. Precision and Recall as the harmonic mean

$$\text{F} - \text{Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precesion} + \text{Recall}}$$ ---------- Equation 4.4

4.5.3. Results

Table 4.2: Results of all algorithms

S.No	Algorithms	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)	Tree_Size
1	J48	78.2178	80.2	78.2	78.4	4
2	SimpleCart	80.198	81.5	80.2	80.4	5

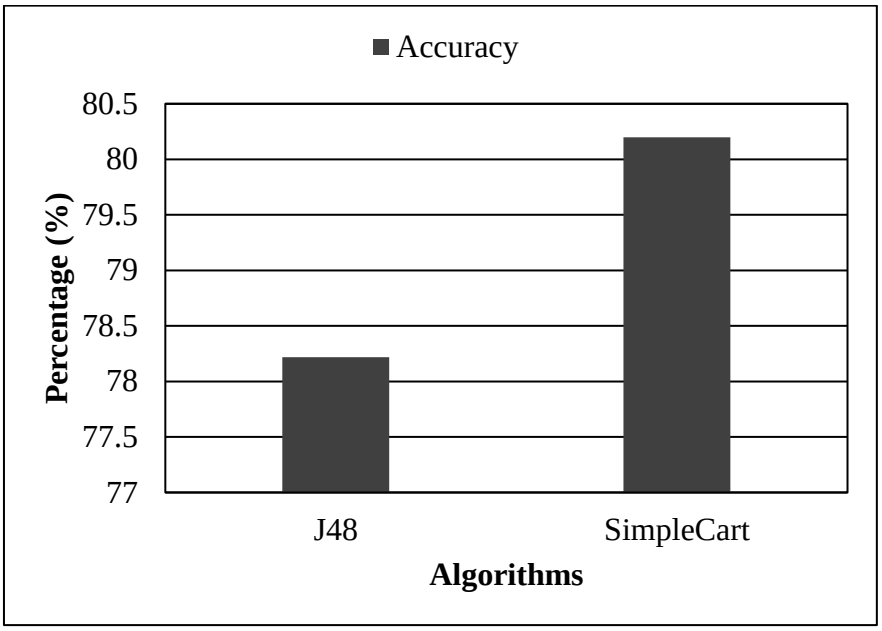

Figure 4.7: Graph of table 4.2 taking Accuracy

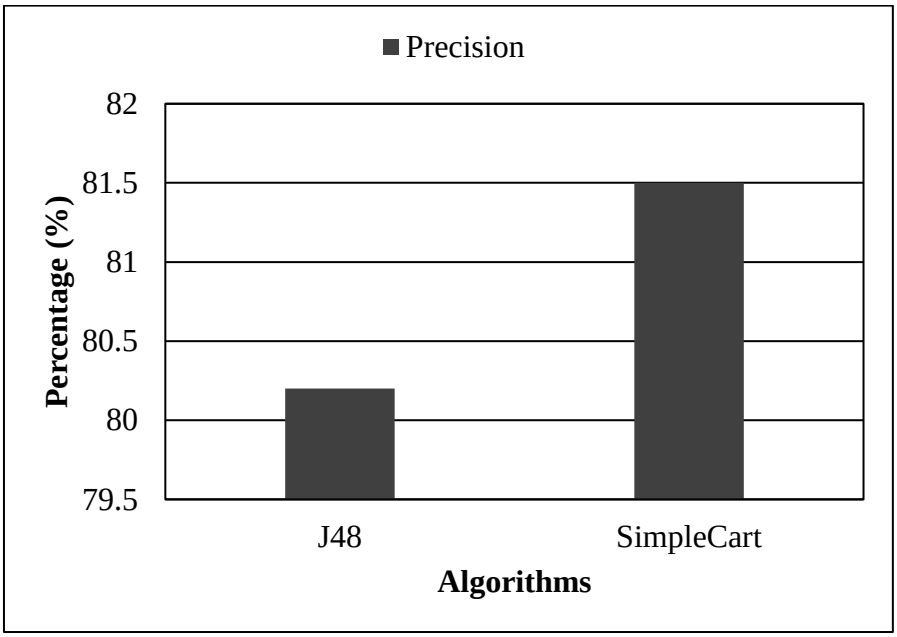

Figure 4.8: Graph of table 4.2 taking Precision

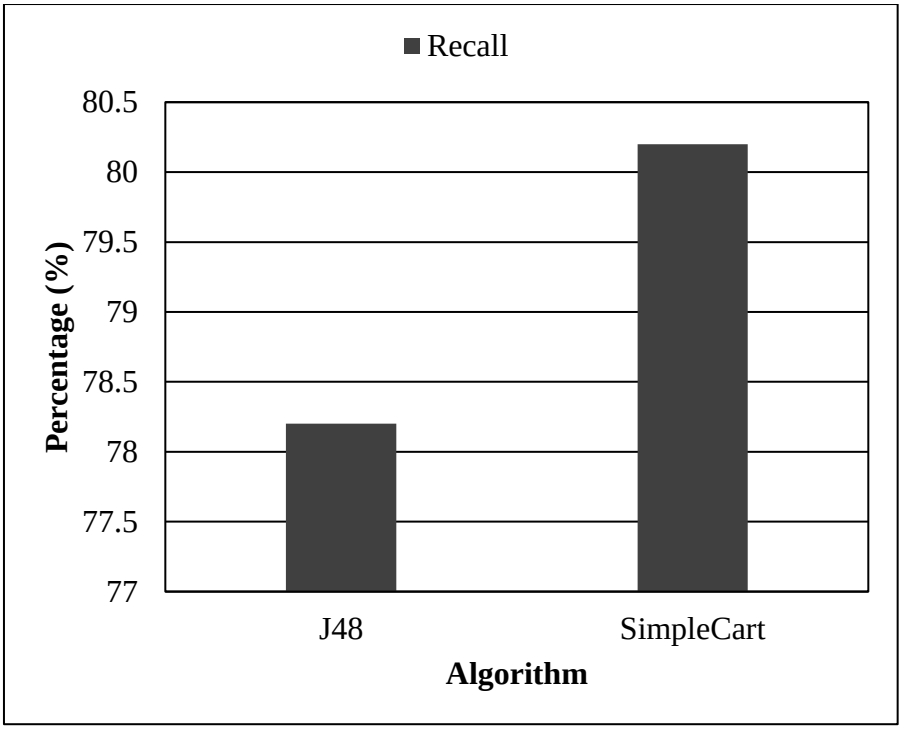

Figure 4.9: Graph of table 4.2 taking Recall

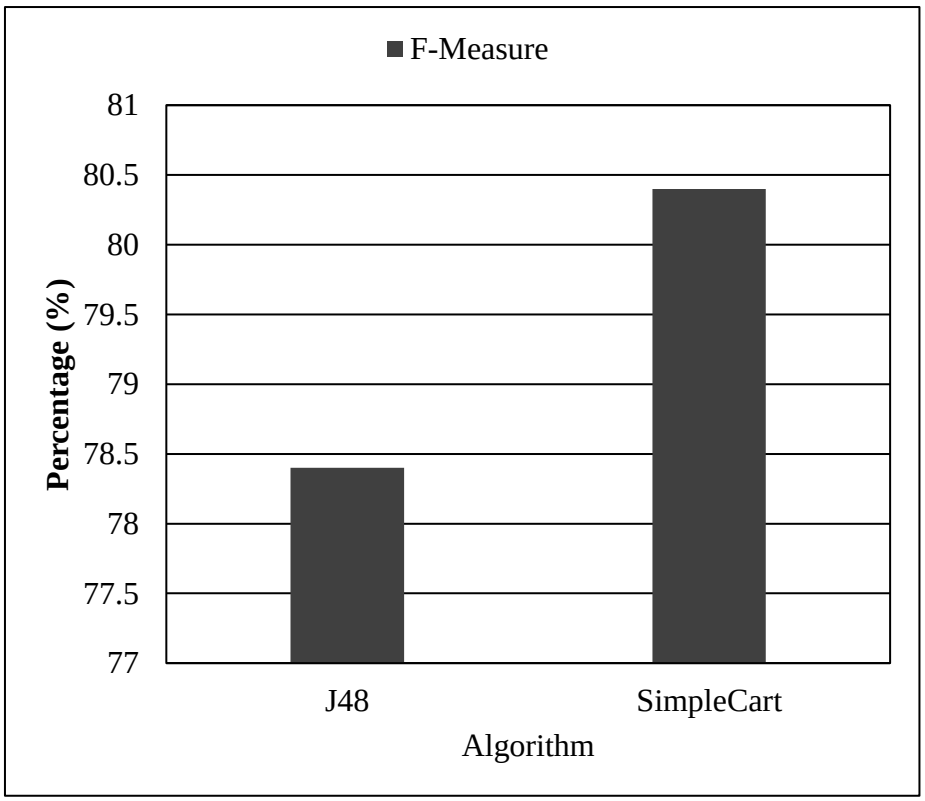

Figure 4.10: Graph of table 4.2 taking F-Measure

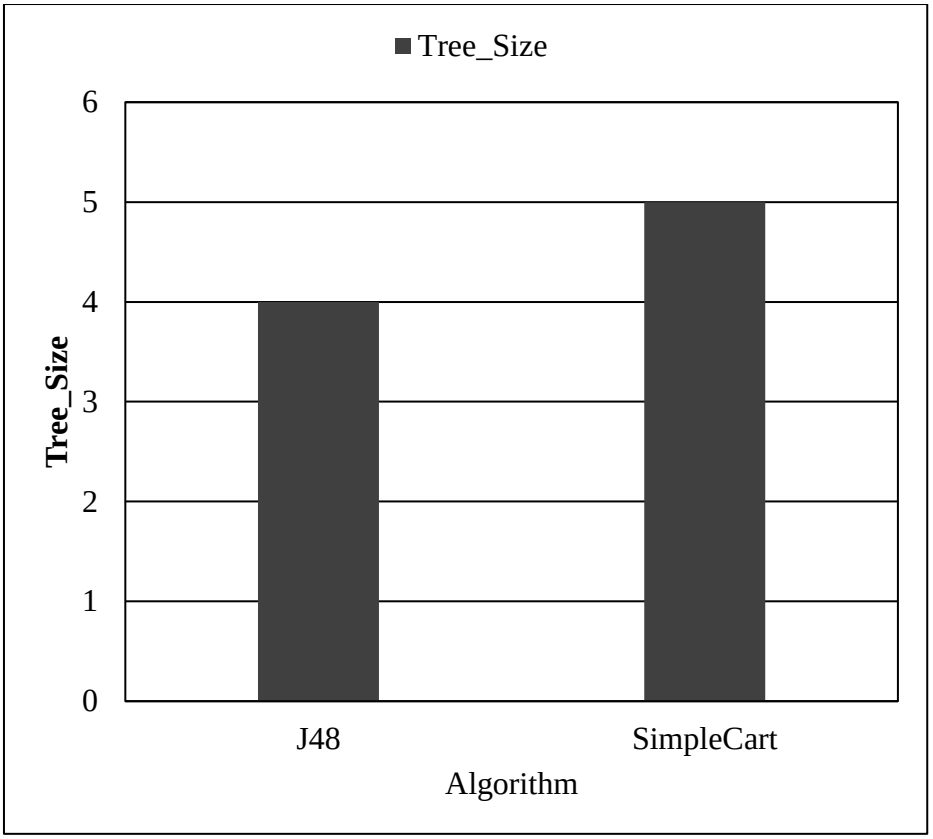

Figure 4.11: Graph of table 4.2 taking Tree_Size

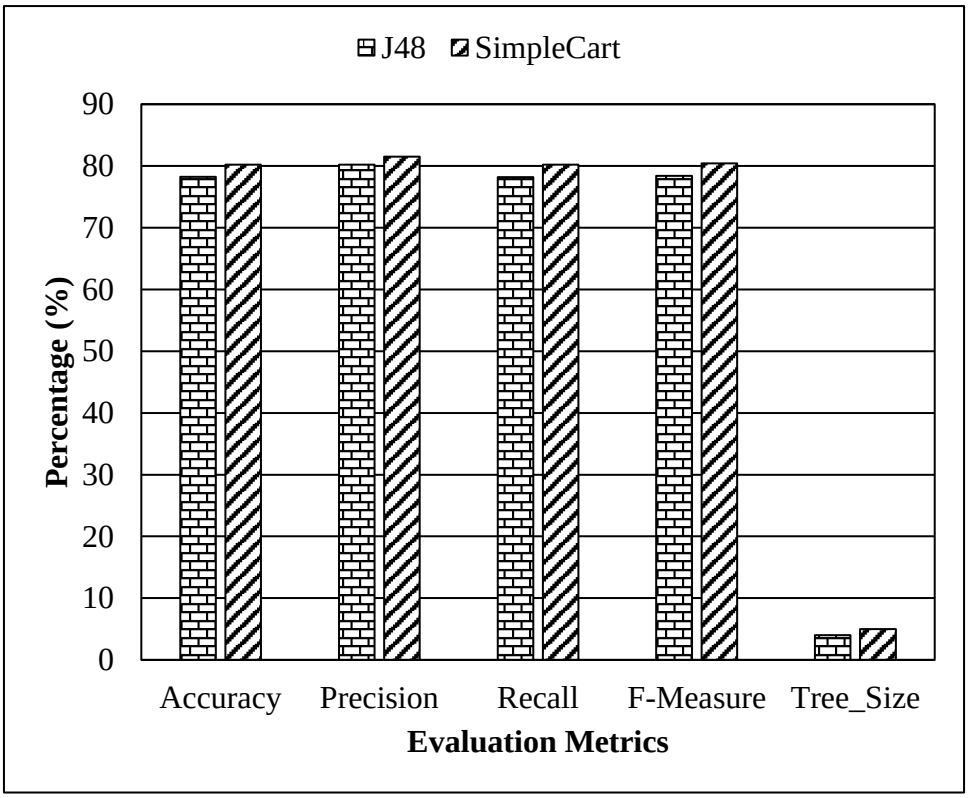

Figure 4.12: Graph of table 4.2 taking all evaluation metrics

4.6. Result Analysis

The table 4.2 and figures 4.7, 4.8, 4.9, 4.10, 4.11, and 4.12 were results of the simulations, which demonstrated the performance of classification algorithm for the comparative analysis of decision tree methods for the prediction of paddy productivity.

Figure 4.7 showed that accuracy observed by implemented classification algorithms where it ranged from 78.2178% to 80.198%. Among the algorithms SimpleCart had got rich as well as motivating and encouraging result with 80.198% and J48 was less capable to classify with accuracy of 78.2178%.

Figure 4.8 showed that precision observed by implemented classification algorithms where it ranged from 80.2% to 81.5%. SimpleCart had got better precision level of 81.5% whereas J48 got less precision level of 80.2%.

Figure 4.9 showed that recall observed by implemented classification algorithms where it ranged from 78.2 to 80.2%. SimpleCart had got again encouraging recall of 80.2% whereas J48 got minimum recall of 78.2%.

Figure 4.10 showed that F-measure observed by implemented classification algorithms where it ranged from 78.4% to 80.4%. Again, SimpleCart had got victory over J48 with the value 80.4%.

Figure 4.11 showed that Tree_Size observed by implemented classification algorithms where J48 had got size 4 and had smaller decision tree size then SimpleCart with tree size 5.

Figure 4.12 showed that the comparison between all the evaluation metrics of the implemented algorithms and from that comparison; SimpleCart had got rich as well as motivating and encouraging performance in every aspect.

CHAPTER 5
CONCLUSION AND FUTURE WORKS

5.1. Conclusion

The comparison of classification algorithm is a complex task and it is an open problem. First, the notion of the performance can be defined in many ways: accuracy, speed, cost, reliability, etc. Second, an appropriate tool is necessary to quantify this performance. Third, a consistent method must be selected to compare with the measured values. The selection of the best classification algorithm for a given dataset is a very widespread problem. In this sense it requires to make several methodological choices. So, this research focused in the comparative analysis of decision tree methods for the prediction of paddy productivity.

From the result analysis it was seen that SimpleCart was able to classify 80.198% of the data correctly which was better than J48 in comparison to results of evaluation metrics (Accuracy, Precision, Recall and F-Measure). In a nut shell, the experiment result showed that J48 has got smaller tree size than SimpleCart but SimpleCart has got 1.9802% better accuracy than J48 for the prediction of paddy productivity.

5.2. Future Works

Directions for future works are:

- One important area for improvement is performance (Accuracy).
- Another is enhancing the performance (Accuracy) more by implementing other classification algorithms.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei. "Data mining: concepts and techniques, Waltham, MA." *Morgan Kaufman Publishers* 10 (2012): 978-1.

[2] T. M. Lakshmi, A. Martin, R. M. Begum, and V. P. Venkatesan. "An analysis on performance of decision tree algorithms using student's qualitative data." *International Journal of Modern Education and Computer Science* 5, no. 5 (2013): 18.

[3] P. Revathi, R. Revathi, and M. Hemalatha. "Comparative Study of Knowledge in Crop Diseases Using Machine Learning Techniques." Int. J. Comput. Sci. Inf. Technol 2 (2011): 2180-2182.

[4] "AI Horizon: Introduction to Machine Learning", *Aihorizon.com*, 2019. [Online]. Available: http://www.aihorizon.com/essays/generalai/supervised_unsupervised_machine_learning.htm [Accessed: 02- Nov- 2019].

[5] A. Papagelis, D. Kalles, "Breeding Decision Trees Using Evolutionary Techniques." In *ICML*, vol 1, pp. 393-400, 2001.

[6] J. R. Quinlan, "C4. 5: programs for machine learning." *Mach. Learn* 16, no. 3 (1993): 235-240.

[7] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan et. al., "Top 10 algorithms in data mining." *Knowledge and Information Systems* 14, no. 1 (2008): 1-37.

[8] Microsoft Press, "Microsoft® Computer Dictionary Fifth Edition", Microsoft Press a division of Microsoft corporation, one Microsoft way, Redmond, Washington 980852-6399, ISBN: 0-7356-1495-4,May 01, 2002.

[9] "WEKA 3 - Data Mining with Open Source Machine Learning Software in Java" Internet: http://www.cs.waikato.ac.nz/ml/WEKA/ [Oct. 16, 2019].

[10] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, K. R. K. Murthy. "Improvements to Platt's SMO algorithm for SVM classifier design." Neural computation 13, no. 3 (2001): 637-649.

[11] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[12] “Attribute-Relation File Format (ARFF)” Internet: http://www.cs.waikato.ac.nz/ml/WEKA/arff.html , Apr. 1, 2002 [Oct. 16, 2019]

[13] A. A. Raorane, and R. V. Kulkarni. "Data Mining: An effective tool for yield estimation in the agricultural sector." International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) 1, no. 2 (2012): 1-4.

[14] A. Nithya, and V. Sundaram. "Identifying the rice diseases using classification andbiosensor techniques." International Journal of Advanced Research in Technology 1, no. 1 (2011): 76-81.

[15] S. Veenadhari, Bharat Mishra, and C. D. Singh. "Soybean productivity modelling using decision tree algorithms." International Journal of Computer Applications 27, no. 7 (2011): 11-15.

[16] J. HaiYue, and S. Kai. "IBLE Algorithm in Agricultural Disease Diagnosis." In 2010 Third International Conference on Intelligent Networks and Intelligent Systems.

APPENDIX

A1. Questionnaire used for primary data collection.

कृषकको नाम:- .....

कृषकको ठेगाना:- ..... नगरपालिका, वडा न:-..... टोल:-.....

१)धानको विउको जात .....

२) धानको विउको स्रोत ?

क) Agrovet ख) Local ग) कृषि विउ

३) प्रति कट्टा लाग्ने विउको औसत मात्रा .....

४) कुल धान खेती गरिएको जमिनको क्षेत्रफल

क) कट्टा ..... ख) विघा .....

५) माटोको प्रकार

क) पाँगो माटो ख) बलौटे माटो

६) सिचाइको स्रोत

क) नहर ख) बोरिङ्ग ग) वर्षा

७) धानमा प्रयोग हुने मलखाद ?

क) जैविक ख) रसायनिक ग) दुवै

८) खेतीकालागि प्रयोग हुने श्रमिक ?

क) Human Labour ख) Machine Labour

९) धानमा लाग्ने मुख्य रोग र किरा

क) ..... ख)..... ग).....

१०)धानमा लाग्ने हानिकारक जीवहरुको नास गर्न कुन विषादी प्रयोग गर्नु हुन्छ ?

क) रसायनिक..... ख) जैविक .....

११) धानवाली पाक्ने समय .....

१२) धान खेतीमा लाग्ने सम्पूर्ण लागत .....

१३) धानको उत्पादन (कि.ग्रा.)

क) गत वर्ष ..... ख) यो वर्ष .....

.....

कृषकको हस्ताक्षर

A2. Instances of Paddy_data_V3.arff

```
@relation paddy_data_V3

@attribute Seed_Name
{Radha,mixed,Dy69,SamaMansulit,Sabitri,Sarzu,Silky,Sukka,Santi
,Shankar,Gold,Dev,Mahak,Gorakhnath}
@attribute Seed_Source
{Agriculture,'Agriculture',Agrovet,mixed,Local,'Local '}
@attribute Seed_Amount numeric
@attribute Prod_Area numeric
@attribute Soil_Type {'Pango  ',Both,Pango,'Pango
',Sandy,'Sandy '}
@attribute Irrigation_used {M,C,MR}
@attribute Fertilizer_used {Both,Chemical,Organic}
@attribute Labour_used {Both,Human,Machine}
@attribute Disease_Insect
{Khaire,Gandhai,Patero,Gabaro,khaire,'Gabero ','jhuse
','Putala ',Kaire,Gabero,Kahire,Sanke,Gandhi,'Gandhai
','Gandhi ','Not Known','Khire ','Kalo
',Jhuse,sanke,Khashare,'Bhusina ',Fategra,'Suke
',Suke,Fauji,jhuse,Khire,Ranya,'Not known'}
@attribute Pestiside_Used {No,Chemical,Organic}
@attribute Harvest_duration {Medium,Long,Short}
@attribute Expenses numeric
@attribute Previous_Prodution {High,Medium,Low}
@attribute Current_Production {High,Medium,Low}

@data
Radha,Agriculture,40,20,'Pango
',M,Both,Both,Khaire,No,Medium,23000,High,High
mixed,Agriculture,16,16,Both,M,Both,Both,Gandhai,No,Long,13500
,Medium,Medium
```

mixed,Agriculture,40,20,Pango,M,Chemical,Both,Patero,Chemical,Medium,12000,High,Medium
mixed,Agriculture,27,9,'Pango
',C,Both,Both,Gabaro,No,Medium,18000,Low,Low
Dy69,Agriculture,12,8,Pango,M,Both,Both,khaire,No,Long,12500,Medium,Low
SamaMansulit,'Agriculture ',20,10,Pango,C,Both,Both,'Gabero
',Organic,Short,14000,Low,Low
Sabitri,'Agriculture ',36,18,Pango,C,Both,Both,'jhuse
',No,Medium,16000,Medium,Medium
mixed,'Agriculture ',10,5,Sandy,C,Both,Both,'Putala
',No,Medium,6000,Low,Low
Sarzu,Agrovet,10,5,Pango,MR,Chemical,Both,Patero,No,Long,10000,Low,Low
Radha,Agrovet,5,5,Sandy,M,Both,Human,Kaire,No,Medium,5000,Low,Low
Radha,Agrovet,16,8,'Sandy
',M,Chemical,Human,Gabero,No,Medium,6000,Low,Medium
Radha,Agrovet,11,7,Sandy,M,Both,Human,Gabero,No,Medium,14000,Medium,Medium
Silky,Agrovet,6,6,Pango,MR,Chemical,Both,Patero,No,Short,6000,Low,Low
Silky,Agrovet,30,15,Pango,MR,Both,Both,Patero,No,Short,15000,Medium,Medium
Sukka,Agrovet,20,10,Sandy,M,Chemical,Human,Gabero,No,Medium,13500,Medium,Medium
Sukka,Agrovet,40,20,Both,M,Both,Human,Kahire,No,Medium,14000,High,High
Santi,Agrovet,48,16,Pango,C,Organic,Both,Sanke,No,Medium,15000,Medium,Low
Shankar,Agrovet,11,7,Sandy,M,Both,Human,Gabero,No,Medium,8000,Low,Low

mixed,Agrovet,10,5,Both,M,Both,Human,Gandhi,No,Medium,3500,Low
,Low
mixed,Agrovet,20,10,Both,M,Both,Both,'Gandhai
',No,Medium,13300,Medium,Medium
mixed,Agrovet,16,8,Both,M,Both,Both,Khaire,No,Medium,13300,Med
ium,Medium
mixed,Agrovet,40,20,Both,M,Both,Human,Gabero,No,Medium,14000,H
igh,Medium
mixed,Agrovet,30,15,Both,M,Both,Human,'Gandhi
',No,Medium,13000,Medium,High
mixed,Agrovet,12,12,Pango,MR,Chemical,Both,Patero,No,Short,150
00,Medium,Medium
mixed,Agrovet,28,14,'Pango
',MR,Both,Both,Patero,No,Short,16000,High,High
mixed,Agrovet,20,10,Both,M,Both,Human,khaire,No,Medium,6000,Me
dium,Medium
mixed,Agrovet,58,29,'Pango ',C,Chemical,Machine,'Not
Known',No,Medium,22000,Medium,Medium
mixed,Agrovet,60,30,'Pango
',C,Chemical,Both,Gabaro,No,Medium,20000,High,High
mixed,Agrovet,30,15,'Sandy
',MR,Both,Both,Patero,No,Short,8000,Medium,Medium
mixed,Agrovet,32,16,Both,MR,Both,Both,Patero,Organic,Short,160
00,High,High
mixed,Agrovet,33,13,'Pango
',C,Both,Both,Gabaro,No,Medium,15000,Medium,Low
mixed,Agrovet,11,7,Sandy,M,Both,Human,Gabero,No,Short,8000,Med
ium,Low
mixed,Agrovet,8,5,Sandy,M,Both,Human,Khaire,No,Short,8000,Low,
Low
Gold,Agrovet,15,10,Sandy,M,Both,Human,Gabero,No,Short,5000,Low
,Medium

mixed,Agrovet,8,5,Pango,M,Both,Human,Gabero,No,Short,7000,Low,Low

Dev,Agrovet,15,10,'Pango

',M,Both,Human,Gabero,No,Medium,10000,High,High

Mahak,Agrovet,20,20,Pango,C,Both,Machine,'Khire

',No,Long,31000,Low,Medium

mixed,mixed,50,25,'Sandy

',MR,Both,Both,Patero,No,Short,20000,High,High

Sarzu,Local,10,5,Pango,C,Chemical,Both,'Kalo

',No,Long,7000,Low,Low

Sarzu,Local,10,5,Pango,C,Both,Both,Jhuse,No,Medium,12000,Low,Low

Sarzu,Local,16,8,'Pango

',C,Chemical,Machine,Gabaro,No,Medium,7000,Low,Low

Sarzu,Local,10,5,'Pango

',C,Both,Both,sanke,No,Medium,15000,Low,Low

Sarzu,Local,20,10,'Pango

',C,Both,Both,Gabaro,No,Medium,11000,Low,Low

Sarzu,Local,15,5,Sandy,C,Both,Both,Khashare,No,Medium,5000,Low,Low

Sarzu,Local,35,10,'Pango

',C,Both,Both,Gabaro,No,Medium,13000,Low,Low

Sarzu,Local,39,13,Pango,C,Both,Both,Gabero,No,Medium,13000,Medium,Low

Sarzu,Local,26,13,'Pango

',C,Both,Both,Gabaro,No,Medium,9000,Low,Low

Sarzu,Local,20,10,Pango,C,Both,Both,Khaire,No,Medium,11000,Medium,Medium

Sarzu,Local,16,8,Pango,M,Both,Human,Gabero,No,Long,6000,Medium,Medium

Sarzu,Local,32,16,'Pango

',C,Both,Both,Gabaro,No,Medium,20000,Medium,Medium

Sarzu,Local,30,15,'Pango
',C,Both,Both,Gabaro,No,Medium,15000,Medium,Medium
Sarzu,Local,30,15,Pango,M,Both,Both,Khaire,Organic,Long,13000,Medium,Medium
Sarzu,Local,50,25,'Pango
',C,Both,Both,Khaire,No,Medium,19000,Medium,Medium
Radha,Local,30,15,Both,MR,Both,Human,Gabero,No,Medium,13500,Medium,High
Silky,Local,10,5,Pango,C,Chemical,Both,'Bhusina
',No,Long,6000,Low,Low
Silky,Local,10,5,Pango,C,Chemical,Both,Fategra,No,Short,10000,Low,Low
Sukka,Local,10,5,Sandy,M,Both,Both,khaire,No,Long,12500,Low,Low
Santi,Local,20,5,Pango,C,Both,Both,'Suke
',No,Medium,13000,Low,Low
Santi,Local,24,8,Pango,C,Both,Both,Khaire,No,Medium,9000,Low,Low
Santi,Local,40,20,Pango,C,Both,Both,Suke,No,Medium,16000,Medium,Medium
Santi,Local,40,20,Pango,C,Both,Both,Khaire,No,Medium,16000,Medium,Medium
Santi,Local,60,20,Pango,C,Chemical,Both,Fauji,No,Medium,18000,Medium,Medium
Santi,Local,40,20,Pango,C,Both,Both,Khaire,No,Medium,18000,Medium,Medium
Santi,Local,18,9,Pango,C,Both,Both,Fategra,No,Long,15000,Medium,Medium
Santi,Local,80,40,Pango,C,Both,Both,Khaire,No,Medium,25000,High,Medium
Gorakhnath,Local,24,8,'Pango
',C,Both,Both,Gabaro,No,Medium,20000,Low,Low

mixed,Local,23,15,Both,M,Both,Human,Gabero,No,Medium,13500,Med
ium,Low
mixed,Local,25,10,Pango,M,Both,Human,Khaire,No,Medium,13400,Me
dium,Medium
mixed,Local,23,15,Both,M,Both,Human,Gabero,No,Medium,13500,Med
ium,High
mixed,Local,102,51,Both,MR,Both,Both,Patero,Organic,Short,3500
0,High,High
mixed,Local,14,7,'Sandy
',MR,Both,Both,Patero,No,Short,5000,Low,Low
mixed,Local,60,30,Sandy,C,Both,Both,'Not
Known',No,Medium,10000,High,High
mixed,Local,14,7,Sandy,MR,Both,Both,Patero,No,Short,4000,Low,L
ow
mixed,Local,24,12,'Sandy
',MR,Both,Both,Patero,Organic,Short,12000,Medium,Medium
mixed,Local,75,35,Pango,M,Chemical,Both,Patero,No,Medium,10000
,High,High
mixed,Local,14,7,Pango,M,Chemical,Both,Patero,Chemical,Medium,
4000,Low,Low
mixed,Local,24,12,Pango,MR,Chemical,Both,Patero,Organic,Medium
,12000,Medium,Medium
mixed,Local,24,12,Pango,MR,Chemical,Both,Patero,No,Medium,6000
,Medium,Medium
mixed,Local,90,60,Sandy,MR,Chemical,Human,Patero,No,Medium,100
00,Medium,Medium
mixed,Local,45,15,Pango,C,Both,Both,jhuse,No,Medium,14000,Low,
Low
mixed,Local,20,10,Pango,C,Both,Both,Khire,No,Long,15000,Medium
,Medium
mixed,Local,30,15,'Pango
',C,Both,Both,Gabaro,Chemical,Medium,16000,Medium,Medium

mixed,Local,40,16,'Pango
',C,Both,Both,Gabaro,No,Medium,18000,Medium,Medium
mixed,Local,18,9,'Pango
',C,Both,Both,Gabaro,No,Medium,10000,Low,Low
mixed,Local,36,18,'Pango
',C,Chemical,Machine,Gabero,No,Long,16000,Medium,Medium
mixed,Local,32,16,'Pango
',C,Both,Both,Gabaro,No,Medium,15000,Medium,Medium
mixed,Local,22,11,'Pango
',C,Both,Both,Fategra,No,Medium,9000,Medium,High
mixed,Local,12,8,Pango,M,Chemical,Both,Patero,No,Medium,10000,
Medium,Low
mixed,Local,100,40,Pango,C,Both,Both,Jhuse,No,Medium,24000,Hig
h,High
mixed,Local,35,20,Pango,MR,Chemical,Both,Patero,No,Short,8000,
Medium,High
mixed,Local,40,20,Pango,C,Both,Both,Ranya,Organic,Long,21000,M
edium,High
Santi,'Local ',10,5,Pango,C,Both,Both,'jhuse
',No,Medium,8000,Low,Low
Santi,'Local ',32,16,Pango,C,Organic,Both,'Not
known',No,Medium,15000,Medium,Medium
Santi,'Local
',20,10,Pango,C,Both,Both,Fauji,No,Medium,15000,Medium,Medium
Santi,'Local ',80,40,Pango,C,Both,Both,'jhuse
',No,Medium,35000,High,High
mixed,'Local ',24,12,Pango,C,Both,Both,'jhuse
',Organic,Medium,20000,Low,Low
mixed,mixed,30,15,Sandy,MR,Both,Both,Patero,No,Short,10000,Med
ium,Low
mixed,mixed,60,30,Both,MR,Both,Both,Patero,Chemical,Long,15000
,High,High

mixed,mixed,40,20,Pango,M,Chemical,Both,Patero,No,Medium,15000,High,High

mixed,mixed,48,24,Pango,MR,Both,Both,Patero,Organic,Short,20000,High,High

mixed,mixed,16,8,Sandy,MR,Both,Both,Patero,No,Short,12000,Medium,Medium

A3. Data visualization of all attributes of dataset paddy_data_V3.csv in WEKA.