Tribhuvan University
Institute of Science and Technology

Performance Analysis of Attribute Selection Methods in Decision Tree Induction

Dissertation

Submitted to

Central Department of Computer Science & Information Technology
Kirtipur, Kathmandu, Nepal

In partial fulfillment of the requirements
for the Master's Degree in Computer Science and Information Technology

By
Ganesh Yogi
April 3, 2018

Tribhuvan University
Institute of Science and Technology

Performance Analysis of Attribute Selection Methods in Decision Tree Induction

Dissertation

Submitted to

Central Department of Computer Science & Information Technology
Kirtipur, Kathmandu, Nepal

In partial fulfillment of the requirements
For the Master's Degree in Computer Science & Information Technology

By
Ganesh Yogi
April 3, 2018

Supervisor
Mr. Nawaraj Paudel

Tribhuvan University
Institute of Science and Technology

Central Department of Computer Science & Information Technology

Student's Declaration

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

... ...
Mr. Ganesh Yogi

April 3, 2018

Tribhuvan University
Institute of Science and Technology

Central Department of Computer Science & Information Technology

Supervisor's Recommendation

I hereby recommend that this dissertation prepared under my supervision by Mr. Ganesh Yogi, entitled "**Performance Analysis of Attribute Selection Methods in Decision Tree Induction**" in partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Information Technology be processed for the evaluation.

... ...
Mr. Nawaraj Paudel (Head of the Department)
Central Department of Computer Science and Information Technology,
Kirtipur, Nepal
(Supervisor)

Date: ……..

Tribhuvan University
Institute of Science and Technology

Central Department of Computer Science & 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 the partial fulfillment for the requirement of Master's Degree in Computer Science and Information Technology.

Evaluation Committee

... ...

Mr. Nawaraj Paudel
Central Department of Computer Science
and Information Technology
Tribhuvan University
Kathmandu, Nepal
(Head)

....

Mr. Nawaraj Paudel
Central Department of Computer Science
and Information Technology
Tribhuvan University
Kathmandu, Nepal
(Supervisor)

... ...
(External Examiner)

. ...
(Internal Examiner)

Date: April 12, 2018

ACKNOWLEDGEMENTS

Above all, I thank God for his blessing and submit my graduate to my almighty for providing strength and confidence in me to complete this work. Secondly, I would like to extent my, gratitude and sincerest thanks to my respected Supervisor Head of the Department, Asst. Prof. Nawaraj Paudel, Central Department of Computer Science and Information Technology, for his impressive guidance, constructive criticism and intellectual support best owed for me sacrificing his invaluable time .

I would like to express my gratitude to respected teachers Prof. Dr. Shashidhar Ram Joshi, Prof.Dr. Subarna Shakya, Prof. Sudarshan Karanjeet, Mr. Sarbin Sayami, Mr. Min Bahadur Khati, Mr.Bishnu Gautam, Mr. Jagdish Bhatt, Mr. Bikash Balami, Mr. Dhiraj Pandey, Mr. Arjun Singh Saud , Mrs. Lalita Sthapit and others staffs of CDCSIT for granting me broad knowledge and inspirations within the time of period of two years.

Thanks to all my close friends for their supports. From the beginning to end in all count, I would like to thank my family members for their love, support and encouragement.

As we know that, there won't be 100% accuracy and efficiency in any work done by both machine and human, so there may be some errors in my project. But, I have done my best to complete this dissertation, so any suggestion regarding the mistakes of this work will be always welcomed.

Abstract

Decision tree learning algorithm has been successfully used in expert systems in capturing knowledge. The main task performed in these systems is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. It is one of the most effective forms to represent and evaluate the performance of algorithms, due to its various eye catching features: simplicity, comprehensibility, no parameters, and being able to handle mixed-type data. There are many decision tree algorithm available named ID3, C4.5, CART, CHAID, QUEST, GUIDE, CRUISE, and CTREE. In this paper, I have used attribute Selection Methods: ID3, C4.5 and CART, and meteorological data collected between 2004 and 2008 from the city of Kathmandu, Nepal, for Decision Tree algorithm. A data model for the meteorological data was developed and this was used to train the Decision Tree with these different attribute selection methods. The performances of these methods were compared using standard performance metrics.

Cross fold validation is performed to test the built model i.e. Decision Tree. 10-fold cross validation is performed which partitions the dataset into 10 partitions and uses 90% data as training and 10% as testing. This testing is performed for ten repetitions.

Experimentation results show, CART Decision tree has slightly more accuracy with large volume of dataset than that of other algorithms ID3 and C4.5. From the view of speed, C4.5 is better than other two algorithms. CART Decision tree has the average system accuracy rate of 80.9315%, system error rate of 19.0685%, precision rate of 83.1%, and recall rate of 83.1%. Similarly, C4.5 Decision Tree has the average system accuracy rate of 80.6849%, system error rate of 19.3151%, and precision rate of 82% recall rate of 84.4%. And ID3 Decision Tree has the average system accuracy rate of 28.08%, system error rate of 4.08%, and precision rate of 89.4% recall rate of 91.3%. From the time to complete perspective C4.5 completes in 0.05 seconds, ID3 completes in 0.32 seconds where as CART completes in 251.82 seconds.

Keywords: ***Data Mining, Classification, Classifier, ID3, C4.5, CART, Supervised Learning, Unsupervised Learning, Decision Tree, Information Gain, Gain Ratio, Gini Index.***

Table of Content

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
DT	Decision Tree
C	Class
D	Dataset
N	Node
DNA	Data Not Available
ID3	Iterative Dichotomize 3
C4.5	Classifier 4.5
CART	Classification and Regression Tree
UCI	Unique Client Identifier
A.D	Anno Domini
WEKA	Waikato Environment for Knowledge Analysis
CSV	Comma Separated Value
BSI	Binary Serialized Instances
ARFF	Attribute Relation File Format
TIA	Tribhuvan International Airport

Chapter 1

Introduction

Decision trees are a very effective method of supervised learning. It aims is the partition of a dataset into groups as homogeneous as possible in terms of the variable to be predicted. It takes as input a set of classified data, and outputs a tree that resembles to an orientation diagram where each end node (leaf) is a decision (a class) and each non- final (internal) node represents a test. Each leaf represents the decision of belonging to a class of data verifying all tests path from the root to the leaf.

The tree is simpler, and technically it seems easy to use. In fact, it is more interesting to get a tree that is adapted to the probabilities of variables to be tested. Mostly balanced tree will be a good result. If a sub-tree can only lead to a unique solution, then all sub-tree can be reduced to the simple conclusion, this simplifies the process and does not change the final result. Ross Quinlan worked on this kind of decision trees. [1]

Rainfall prediction is estimate of future condition of rainfall. It is a state of atmosphere at given time in terms of weather variables like temperature, humidity, pressure, wind speed, wind direction etc. It is a nonlinear and dynamic process. So, it varies day to day even minute to minute [2].

The important hydrological event rainfall is the quantity of water falling in drops from vapor condensed in the atmosphere. When water droplets in clouds become too heavy to stay in the air, they fall out towards the ground. Making reliable prediction about rainfall is very important in many areas of human activities. Rainfall supplies water for crops production. Crop plants use a huge amount of water and since Nepal is an agro-based country, accurate prediction of rainfall will be useful for proper planning of cultivation. That's why those who are involved in agriculture, they will be interested to know whether the next days (or months) will be rainy/non-rainy. Although water is vital to life, yet water can be extremely destructive. Thus, rainfall forecasting can warn of happening flood or drought so that peoples can save their lives and properties. Rainfall forecasting is also important for engineering applications, mainly for the design of hydroelectric power projects, because this system requires prior information about average rainfall, maximum/minimum rainfall for a year/each month. In urban areas, rainfall also has a strong influence on traffic control. Rainfall is one of the most important and challenging

operational tasks carried out by the meteorological services all over the world. It is furthermore a complicated procedure that includes multiple specialized fields. The most widespread techniques used for rainfall prediction are the numerical and statistical methods [3]. Even thought researches in these fields are being conducted for a long time, successes of these models are rarely visible. Even though statistical model can predict accuracies in short term rainfall, it is difficult to predict the long term prediction of the rainfall due to nonlinear character of rainfall process. So, statistical method cannot generate good results [4].

1.1 Motivation

In Data Mining data sets will be explored to yield hidden and unknown predictions which can be used in future for the efficient decision making. Now a day's companies use different techniques of Data Mining. Building a decision tree is one of the most important tasks in data mining since it is very simple approach to classify future data and it is also considered as a basis of building other model like "If-Then Rule". So, decision tree is one of the data mining techniques which can be used to provide the accurate result. Weather data has been taken for the training of the model and its accuracy is measured by classifying the data. As Nepal is the agricultural country and most of Nepali people depend on agriculture, weather forecasting plays vital role in agriculture. Research on weather forecasting may help directly or indirectly in agriculture by using decision tree algorithm.

1.2 Problem definition

Decision Tree Induction is one of the mostly used model training methods. It can be used with different nature of data like numeric or discrete. The method of choosing the node in decision tree makes different during the training phase and it impact in the performance of decision tree during the classification of future data. So it is an important task of choosing the right method based on the nature of dataset available. If so, it will partition the dataset more correctly and classify the future data more correctly with high accuracy.

Building a right model using decision tree algorithm is really a big challenge. It concerns with different factors like nature of data, attribute selection methods, number of attributes in the dataset etc.

Rainfall prediction has always been a challenging task. This is due to the fact that the data sets are highly nonlinear in nature. The accuracy of the class prediction system is also highly dominated by the actual parameters chosen.

1.3 Objective

The objectives of this research work are as follows:

a. To perform the analysis of attribute selection methods in Decision Tree Induction.
b. To suggest the suitable method of attribute selection for building Decision Tree.

1.4 Outline of document

The remaining part of the document is organized as follows:

- Chapter 2 describes necessary background information and related work in Decision Tree Induction.
- Chapter 3 describes detail system model and the theoretical approaches for building a Decision Tree Induction problem. It includes data normalization, data transformation, Training and Testing Approaches.
- Chapter 4 describes the implementation details of the system. It includes description about the tools used and fundamental methods.
- Chapter 5 includes analysis and experimentation results about Performance of DT based on Accuracy and speed.
- Chapter 6 includes conclusion and future works.

Chapter 2

Literature Review

Data mining was introduced in the 1990s and it is traced back along three categories i.e. classical statistics, artificial intelligence and machine learning. Data mining is the process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. It is also known as knowledge discovery i.e. detecting something new from large–scale or information processing [5]. Its objective is to extract knowledge or discovering of new information from large volumes of raw data for further use [6]. It is mainly related to database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post processing of discovered structures, visualization, and online updating. The data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. A Decision Support System is a computer-based information system that supports business or organizational decision making activities [5]. It serves the management, operations, and planning levels of an organization and help to make management decisions, which may be rapidly changing and not easily specified in advance .Hence, the actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis i.e. grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups), unusual records (anomaly detection i.e. detection of outliers, noise, deviations or exceptions in large data sets) and dependencies (association rule mining i.e. detecting interesting relations between the variables in large databases).

2.1 Classification and prediction

Classification is the technique in which set of items are classified in the predefined category. Prediction is the process of predicting categorical class labels, constructing a model based on the training set.

Classification techniques are supervised learning techniques that classify data item into predefined class label. It is one of the most useful techniques in data mining to build

classification models from an input data set. The used classification techniques commonly build models that are used to predict future data trends. There are several algorithms for data classification such as decision tree and Naïve Bayes classifiers. There are two main kinds of models in data mining which are as follow:

2.1.1 Predictive model: In this model, known data results are used to develop a model and that can be used to explicitly predict values. The purpose of Predictive model is mainly to predict the future outcome than current behavior. The prediction output can be numeric value or in categorized form. The predictive models are the supervised learning functions which predict the target value.

2.1.2 Descriptive model: In this model, patterns are described from existing data and models are abstract representation of reality which can be reflected to understand business and suggest actions. The second approach for mining data from large datasets is known as Descriptive data mining. It is normally used to generate correlation, frequency, etc. This Descriptive method can be defined as to discover regularities in the data and to uncover patterns. This is also used to find interesting subgroups in the bulk of data

2.2 Association rule

In this technique, interesting association between attributes that are contained in a database is discovered which are based on the frequency counts of the number of items occur in the event (i.e. a combination of items), association rule tells if item X is a part of the event, then what is the percentage of item Y is also the part of event. Maket Basket Analysis and Apriori Algorithm are used in Association rule mining.

2.3 Clustering

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. [5]

Clustering is a technique used to discover appropriate groupings of the elements for a set of data. It is undirected knowledge discovery or unsupervised learning i.e. there is no target field and relationship among the data is identified by bottom-up approach. A cluster is a subset of objects which are "similar". A subset of objects such that the distance between any two objects in the cluster is less than the distance between any object in the cluster and any object not located inside it. It is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. It helps users understand the natural grouping or structure in a data set. It is unsupervised classification that means there are no predefined classes.

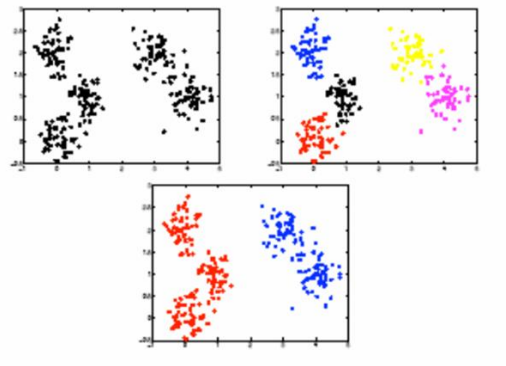


Fig: 2.1 Clustering of Objects

2.4 Supervised Learning

Supervised learning is the task of machine learning where the system can learn from the given available data. This type of learning is same as the human learning from the past experience to gain a new knowledge. Classification and prediction are supervised learning.

A data set used in the learning task consists of a set of data records, which are described by a set of attributes, A = {A1, A2, … $A_{|A|}$}, where |A| denotes the number of attributes or the size of the set A. The data set also has a special target attribute C, which is called the class attribute [7].

2.5 Unsupervised Learning

Unsupervised learning is the form of learning by observation, rather than learning by examples. In unsupervised learning the class label information is not present. Clustering is an example of unsupervised learning.

In unsupervised or undirected data mining, however, variable is singled out as the target as like the descriptive mining technique. The goal is rather to establish some relationship among all the variables in the data. The user asks the computer to identify patterns in the data that may be significant. Undirected modeling is used to explain those patters and relationships one they have been found.

2.6 Decision Tree

Decision tree induction is the learning of decision trees from class-labeled training tuples. 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.

Decision tree is method used in data mining. It is used to predict the value of target based on several input parameter. Tree can be constructed by splitting the source data set into subsets based on an attribute value. In decision tree dependent variable is predicted from the independent variable. [8]

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target parameter based on several input parameter. A tree can be made to learn by splitting the source data set into subsets based on an attribute value test [6].

2.6.1 Algorithm: Generate decision tree. Generate a decision tree from the training tuples of data partition, D.

Input:

- Data partition, D, which is a set of training tuples and their associated class labels;
- Attribute_list, the set of candidate attributes;
- Attribute_selection_method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting attribute and, possibly, either a split-point or splitting subset.

Output: A decision tree.

2.6.2 Method:

(1) create a node N ;

(2) if tuples in D are all of the same class, C, then

(3) return N as a leaf node labeled with the class C;

(4) if attribute_list is empty then

(5) return N as a leaf node labeled with the majority class in D; // majority voting

(6) Attribute_selection method(D, attribute_list) to find the "best" splitting_criterion;

(7) label node N with splitting_criterion;

(8) If splitting_attribute is discrete-valued and

multiway splits allowed then // not restricted to binary trees

(9) attribute_list ← attribute_list – splitting_attribute; // remove splitting attribute

(10) for each outcome j of splitting criterion

// partition the tuples and grow subtrees for each partition

(11) let Dj be the set of data tuples in D satisfying outcome j; // a partition

(12) if Dj is empty then

(13) attach a leaf labeled with the majority class in D to node N ;

(14) else attach the node returned by Generate_decision_tree(Dj,attribute_list) to node N ;

endfor

(15) return N ;

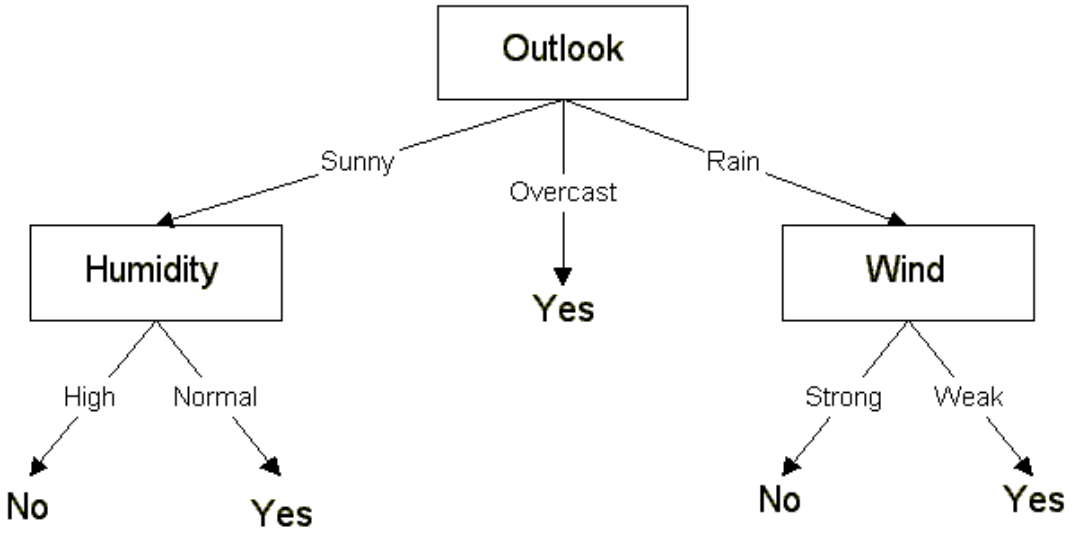


Fig: 3.6

Figure 2.2 Example of DT.

2.7 Attribute Selection Measures

Two approaches that enable standard machine learning algorithms to be applied to large databases are feature selection and sampling. Both reduce the size of the database. Recently lot of research work is going on feature selection, a process that can benefit learning algorithms regardless of the amount of data available to learn from. [9]

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. If the splitting attribute is continuous-valued or if we are restricted to binary trees, then, respectively, either a split point or a splitting subset must also be determined as part of the splitting criterion. Three popular attribute selection measures- information gain, gain ratio and Gini index.[7]

2.7.1 Information Gain or ID3

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. Let node N represent or hold the tuples of partition D. The attribute with the highest information gain is chosen as the splitting attribute for node N . This attribute

minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or "impurity" in these parti-tions. Such an approach minimizes the expected number of tests needed to classify a given tuple and guarantees that a simple (but not necessarily the simplest) tree is found. [7]

The expected information needed to classify a tuple in D is given by

Entropy:

It measures homogeneity of a node and it is denoted by formula

Entropy(D) = Info(D)= $\sum_{i=1}^{m} p_{i} log_2(p_i)$

Where, p_i is the proportion of D belonging to class i (i.e. Ci).

D is entire dataset and $p_i = |C_{i,D}| / |D|$

The amount of information needed to arrive at an exact classification is measured by

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

The term $\frac{|D_j|}{|D|}$ acts as the weight of the jth partition. $Info_A(D)$ is the expected information required to classify a tuple from D based on the partitioning by A. The smaller the expected information required, the greater the purity of the partitions. Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e.,obtained after partitioning on A). [10]

That is,

$$Gain(A) = Info(D) - Info_A(D)$$

The attribute A with the highest information gain, Gain(A), is chosen as the splitting attribute at node N.

2.7.2 Gain Ratio or C4.5

The information gain measure is biased toward tests with many outcomes. C4.5, a successor of ID3, uses an extension to information gain known as gain ratio, which attempts to overcome this bias. It applies a kind of normalization to information gain using a "split information" value defined analogously with Info(D) as:

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2(\frac{|D_j|}{|D|})$$

This value represents the potential information generated by splitting the training data set, D, into v partitions, corresponding to the v outcomes of a test on attribute A.

The gain ratio is defined as

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

The attribute with the maximum gain ratio is selected as the splitting attribute.

It was found that the performance of C4.5 (J4.8) decision tree algorithm was far better than that of Naïve Bayes. [6] This is one of the reasons that have motivated me to compare C4.5 with other attribute selection methods.

2.7.3 Gini index or CART

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

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

The Gini index considers a binary split for each attribute. Considering the case where A is a discrete-valued attribute having v distinct values, {a1, a2, :::, av}, we examine all the possible subsets that can be formed using known values of A. If A has v possible values, then there are 2^v possible subsets. 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 a binary split, we compute a weighted sum of the impurity of each resulting partition. For example, if a binary split on A partitions D into D1 and D2, the Gini index of D given that partitioning is

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

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

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

The attribute that maximizes the reduction in impurity (or, equivalently, has the minimum Gini index) is selected as the splitting attribute. This attribute and either its splitting subset (for a discrete-valued splitting attribute) or split-point (for a continuous-valued splitting attribute) together form the splitting criterion.[7]

2.8 Previous work

True accurate model building is really a big challenge due to the different nature of attribute selection methods and the nature of dataset itself in which the model is trained. Different research works have been conducted before my thesis and they have different analysis in the performance of different attribute selection methods in decision tree induction.

In 2013, three existing decision tree algorithms (ID3, C4.5, and CART) have been applied on the educational data for predicting the student's performance in examination. All the algorithms are applied on student's internal assessment data to predict their performance in the final exam. The efficiency of various decision tree algorithms can be analyzed based on their accuracy and time taken to derive the tree. The predictions obtained from the system have helped the tutor to identify the weak students and improve their performance. Table 2 shows the accuracy of ID3, C4.5 and CART algorithms for classification applied on some data sets using 10-fold cross

validation is observed. It shows that a C4.5 technique has highest accuracy of 67.7778% compared to other methods. ID3 and CART algorithms also showed an acceptable level of accuracy. [11]

In the year of 2015, a research has been performed to find the "Impact of Evaluation Methods on Decision Tree Accuracy" based on the available datasets. Different algorithms have been implemented for their accuracy with different validation techniques and the analysis report shows that the best algorithm that performed well in overall is CART. [12]

In 2015, another research has been performed based on the accuracy and speed of different attribute selection measures used in Decision tree induction. For comparison they have used four datasets, viz., Weighting, Yeast, Deals and Car from open source UCI datasets. And the analysis from accuracy point of view, accuracy graph of "Yeast" data set clearly shows the difference between the accuracy of the three algorithms. It can be seen from the graph that CART is somewhat superior to ID3 and C4.5. Comparing ID3 and C4.5, ID3 looks inferior to C4.5. Analyzing all the graphs, the general trend is that CART>C4.5>ID3. For smaller data sets ID3 has the least execution time, followed by CART and then C4.5. As the dataset size increases, ID3 takes higher execution time than CART and C4.5. CART and C4.5 take almost same execution time but CART has an edge over C4.5.Hence, to generalize the observations CART>ID3>C4.5 in terms of execution time. [13]

A research in a comparative study in decision tree ID3 and C4.5 is made. In the article, they have focused on the key elements of the decision tree construction from a set of data and presented the algorithm ID3 and C4.5 that respond to these specifications. Also they have compared ID3/C4.5, C4.5/C5.0 and C5.0/CART, which has led to the confirmation that the most powerful and preferred method in machine learning is certainly C4.5.[14]

In 2017, it has been found that both C4.5 and CART are better than ID3 when missing values are to be handled whereas ID3 cannot handle missing or noisy data. But it is also analyzed that ID3 produces faster results. This paper has used the database of an Electronic store to see whether a person buys a laptop or not. [15]

Different researcher has research on the decision tree for rain fall prediction. Here it has mentioned the some research on the decision tree in the context of the classification of the rain fall.

Prasad proposed to employ Supervised Learning decision tree using Gini index for the prediction of the precipitation which resulted in an accuracy of 72.3% [16].

E. G. Petre [17] presented a small application of CART decision tree algorithm for weather prediction. The data collected is registered over Hong Kong. The data is recorded between 2002 and 2005. The data used for creating the dataset includes parameters year, month, average pressure, relative humidity, clouds quantity, precipitation and average temperature. The decision tree, results and statistical information about the data are used to generate the decision model for prediction of weather.

F. Oliya and A. B. Adeyemo [18] investigated the use of data mining techniques in predicting maximum temperature, rainfall, evaporation and wind speed. C4.5, ID3 decision tree algorithms and artificial neural networks are used for prediction. The meteorological data is collected between 2000 and 2009 from the city Ibadan, Nigeria. A data model for the meteorological data is developed and is used to train the classifier algorithms. The performance of each algorithm is compared with the standard performance metrics and the algorithm with the best result is used to generate classification rules for the mean weather variables.

Data mining methods was implemented for guiding the path of the ships during sailing. Global Positioning System is used for identifying the area in which the ship is currently navigating. The attributes of weather data includes climate, humidity, temperature, stormy [19]. The weather report of the area traced is compared with the existing database. The analyzed dataset is provided to the decision tree algorithm, C4.5 and ID3. The decision obtained regarding the weather condition is instructed to the ship and the path is chosen accordingly.

Soo-YeonJi [20] predicted the hourly rainfall in any geographical regions. Rainfall, the hourly rainfall prediction is performed. CART and C4.5, ID3 are used to provide outcomes, which may provide hidden and important patterns with transparent reasons. Result obtained from both algorithms was satisfactory.

S. Kannan and S. Ghosh [21] contributed towards developing methodology for predicting state of rainfall at local or regional scale for a river basin from large scale climatological data. A model based on decision tree algorithm, CART, ID3, is used for the generation of rainfall states from large scale atmospheric variables in a river basin.

Although there is a lots of AI model for prediction of rainfall but there is no comparative study of decision tree algorithm (ID3) and neural network algorithm (back propagation) for rainfall prediction. So in this thesis we are doing the comparative study of these three models for the decision tree induction used in weather data of Nepal at airport location.

Chapter 3

Research Methodology

The Top level Decision Tree Induction System as shown in Fig 3.1 is divided into four sub-systems, data acquisition, Data selection, data preprocessing and data transformation. Each stages of this theoretical model are briefly described in this section. Detail of each subsystem is given in later sections.

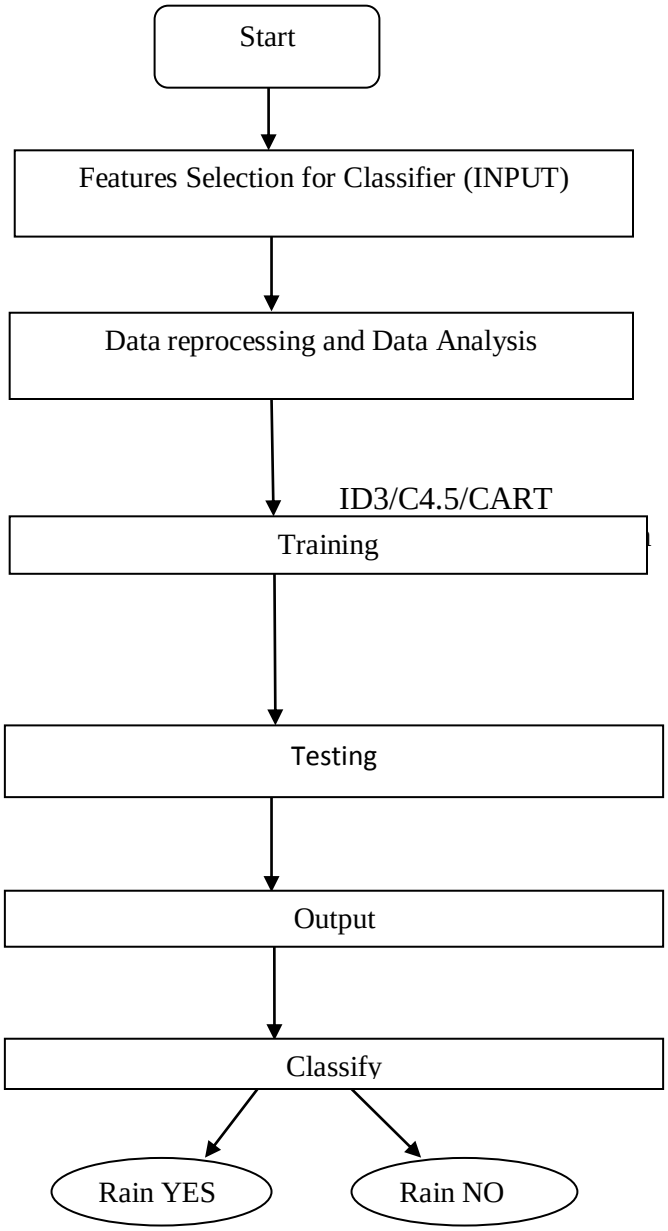


Fig: 3.1 Process Flow-chart

In this thesis I am using three different attribute selection methods for Decision Tree Induction and they are ID3 that uses Gain Information, C4.5 that uses Gain Ratio and CART that uses Gini Index. So there are different attribute selection measures for decision tree induction and they have influence in building the model process that can be analyzed in terms of their accuracy and speed.

3.1 Process for Decision Tree Induction

3.1.1 Data Collection

Dataset, as secondary data for Decision Tree Inducion is collected from the Department of meteorology and Hydrology department of Nepal. Here I have used different input parameters like wind speed, min temperature, max temperature, humidity, evaporation and rainfall of the Kathmandu valley satiation" Tribhuvan International Airport " from the year 1999 to 2008 A.D (data of 10 years). All the data is collected of daily base.

3.1.2Data Selection

At this stage, data relevant to the analysis was decided and retrieved from the dataset. The meteorological dataset has five attributes which are wind speed, humidity precipitation, minimum temperature, maximum temperature, their type and description is presented in Table 3.1.

Table: 3.1 sample data set with 7 parameters.

Morning Humidity	Evening Humidity	Evaporation	Wind	Min Temperature	Max Temperature	Precipitation
0.972414	0.584601	DNA	0.823241	0.130435	0.433824	NO
1	0.587452	0.585366	0.824059	0.150198	0.463235	NO
1	0.569392	0.414634	0.827332	0.44664	0.279412	YES
1	0.520913	0.390244	0.824059	0.185771	0.470588	YES
1	0.518061	DNA	0.822422	0.43083	0.680147	NO
0.781034	0.461027	0.658537	0.828151	0.462451	0.613971	NO
0.72069	0.664449	4.341463	0.828969	0.521739	0.636029	NO
0.922414	0.474335	DNA	0	0.474308	0.617647	YES

3.1.3 Data preprocessing

There are many methods for data normalization: min-max normalization, z-score normalization, and normalization by decimal scaling. Let A be a numeric attribute with n observed values, v1, v2, : : : , vn.

Min-max normalization performs a linear transformation on the original data. Suppose that min_A and max_A are the minimum and maximum values of an attribute, A. Min-max normalization maps a value, v_i , of A to v'_i in the range [new minA, new maxA] by computing

$$v'_i = ((v_i - min_A) / (max_A - min_A))*(new\ max_A - new\ min_A) + new\ min_A$$

Min-max normalization preserves the relationships among the original data values. It will encounter an "out-of-bounds" error if a future input case for normalization falls outside of the original data range for A.[7]

The obtained input and the output data have to be normalized because they are of different units and otherwise there will be no correlation between input and output values. First the mean of all the data separately was taken for humidity, wind speed, rainfall, and temperature.

Z-Score Normalization

Let M be the mean.

M=sum of all entries/number of entries.

Then the standard deviations, SD, for each of these parameters were calculated individually. Now after having the values of mean and SD for every parameter, the values for each parameters were normalized by using

Normalized value = (x-M)/SD. [22]

3.1.4 Data Transformation

This is also known as data consolidation. It is the stage in which the selected data is transformed into forms appropriate for data mining. Here, instead of non-zero and zero value of precipitation, i have used “YES” or “NO” as class label value. For temperature, Min Temperature and Max Temperature and for Humidity, Morning Humidity and Evening Humidity are used as attributes. Similarly, initially data are in numeric values and during preprocessing they are changed into nominal values by using 1-R discretize method. Unwanted attributes are removed from the consideration during preprocessing. The data file is saved in CSV file format.

3.2 Formation of DT

To build a tree, it uses a humidity, minimum temperature, maximum temperature and wind speed as a attributes. As mentioned above decision tree generate some rules and based on the generated rules, it classifies the values. To build a tree, it selects the attribute as a root. This attribute selection can be performed based on the value provided by the Information Gain, Gain Ratio or Gini Index. ID3, C4.5 and CART use Information Gain, Gain Ratio and Gini Index respectively. The attribute which has the highest value of information gain or Gain Ratio or Gini Index is chosen as the root node.

3.2.1 Information Gain or ID3

The expected information needed to classify a tuple in D is given by

Entropy:

It measures homogeneity of a node and it is denoted by formula

Entropy(D) = Info(D)= $\sum_{i=1}^{m} p_{i log_2(p_i)}$

Where p_i is the proportion of D belonging to class i(Ci).

D is entire dataset and $p_i = |C_{i,D}| / |D|$

The amount of information needed to arrive at an exact classification is measured by

$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$

The gain for every attribute is calculated as:

$$Gain(A) = Info(D) - Info_A(D)$$

The attribute A with the highest information gain, Gain(A), is chosen as the splitting attribute at node N.

3.2.2 Gain Ratio or C4.5

C4.5 applies a kind of normalization to information gain using a "split information" value defined analogously with Info(D) as:

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2(\frac{|D_j|}{|D|})$$

The gain ratio for each attribute is defined as:

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

The attribute with the maximum gain ratio is selected as the splitting attribute.

3.2.3Gini Index or CART

The Gini Index measures the impurity of D (set of training tuples) as:

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

The Gini index considers a binary split for each attribute as described in previous chapter. When considering a binary split, we compute a weighted sum of the impurity of each resulting partition. For example, if a binary split on Attribute (A) partitions Dataset (D) into D1 and D2, the Gini index of D is given by:

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

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

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

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

Here tree is generated having humidity as root, temperature and wind are the sub tree as shown in Fig 3.2.

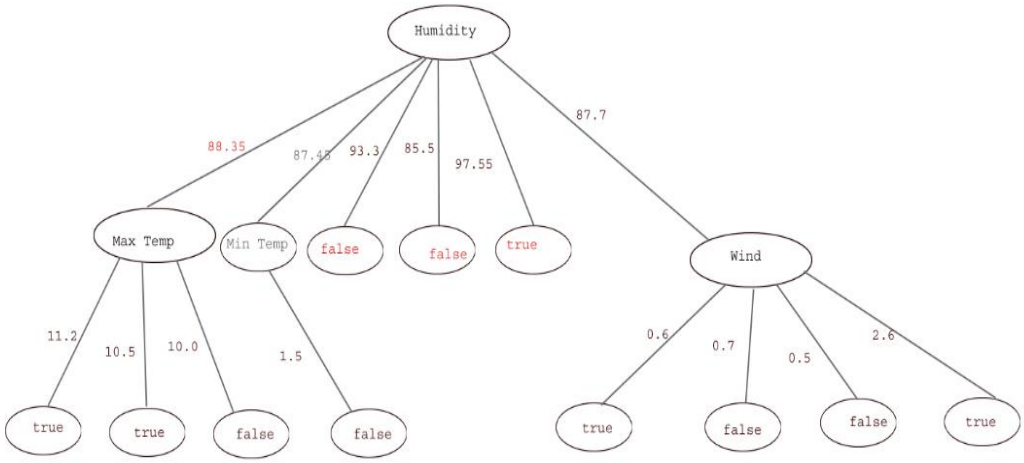


Fig: 3.2 DT for weather data.

3.3 System Evaluation Measures

The correctness of the build model can be evaluated by computing the number of correctly recognized class examples (true positives), the number of correctly recognized examples that do not belong to the class (true negatives), examples that either were incorrectly assigned to the class (false positives) and examples that were not recognized as class examples (false negatives).

Measures for multi-class classification based on a generalization of the measures of binary classification for many classes Ci are given below. Where, tpi represent true positive for class Ci, fpi represent false positive for class Ci, fni represent false negative for class Ci, tni, representtrue negative for class Ci.

3.3.1 Average System Accuracy

Average system accuracy evaluates the average per-class effectiveness of a classification system.

Average Accuracy=$\sum_{i=1}^{l} \frac{tpi+tni}{tpi+tni+fpi+fni}$

3.3.2 System Error

System error is the average per-class classification error of the system.

Error rate = $\sum_{i=1}^{l} \frac{fpi+fni}{tpi+tni+fpi+fni}$

3.3.3 Precision

Precision (also called positive predictive value) is the number of correctly classified positive examples divided by the number of examples labeled by the system as positive. Micro precision is the agreement of the data class labels with those of classifiers calculated from sums of per-test decisions.

$$\text{Precision} = \frac{\sum_{i=1}^{l} tpi}{\sum_{i=1}^{l} tpi + fpi}$$

3.3.4 Recall

Recall is the effectiveness of a classifier to identify class labels if calculated from sums of per-test decisions.

$$\text{Recall} = \frac{\sum_{i=1}^{l} tpi}{\sum_{i=1}^{l} tpi + fni}$$

Chapter 4

Implementation Tools and Techniques

All the algorithms of purposed Decision Tree Induction are implemented in popular data mining tool called WEKA. WEKA is installed on an Intel(R) Core(TM)2 Duo CPU T5470 @ 1.60 GHz, 1.60 GHz processor. The Computer has total main memory of 1 Gigabyte and 32-bit Operating system, x86-based processor and Microsoft Windows7 ultimate operating system installed in it.

4.1 Weka

WEKA was developed at the University of Waikato in New Zealand, and the name stands for Waikato Environment for Knowledge Analysis.[23] WEKA is machine learning/ data mining software written in Java Language (distributed under the GNU Public License). WEKA is a collection of machine learning algorithms for data mining tasks. It is used for research, education and applications. Main features of WEKA are:

- Comprehensive set of data pre-processing tools, learning algorithms and evaluation methods.
- Graphical user interfaces (including data visualization).
- Environment for comparing learning algorithms.

WEKA contains tools for developing new machine learning schemes. It can be used for

- Preprocessing
- Classification
- Clustering
- Association
- Visualization

Input to WEKA is given as a dataset. WEKA permits the input data set to be in numerous file formats like CSV (comma separated values:*.csv). Binary Serialized Instances (*.bsi) etc, However, the most preferred and the most convenient input file format is the attribute relation file format (arff). Here I have used the dataset in .csv file format.[24]

4.2 Methods Used in Implementation

Steps to apply classification techniques on dataset and get result in WEKA:

Step 1: Take the input dataset.

Step 2: Apply the classifier algorithm on the whole data set.

Step 3: Note the accuracy given by it and time required for execution.

Step 4: Repeat step 2 and 3 for different classification algorithms on different datasets.

Step 5: Compare the different accuracy provided by the dataset with different classification algorithms and identify the significant classification algorithm for particular dataset.[4]

Chapter 5

Experiments and Results

Decision tree is experimented by creating one training data set and one testing dataset. This chapter describes the datasets used in experiment and empirical results. Training and testing dataset are described below.

5.1 Cross fold Validation

In *k*-fold cross-validation, the initial data are randomly partitioned into *k* mutually exclusive subsets or "folds," *D*1, *D*2, : : : , *Dk*, each of approximately equal size. Training and testing is performed *k* times. In iteration *i*, partition *Di* is reserved as the test set, and the remaining partitions are collectively used to train the model. That is, in the first iteration, subsets *D*2, : : : , *Dk* collectively serve as the training set to obtain a first model, which is tested on *D*1; the second iteration is trained on subsets *D*1, *D*3, : : : , *Dk* and tested on *D*2; and so on. Unlike the holdout and random subsampling methods, here each sample is used the same number of times for training and once for testing. For classification, the accuracy estimate is the overall number of correct classifications from the *k* iterations, divided by the total number of tuples in the initial data.

Leave-one-out is a special case of *k*-fold cross-validation where *k* is set to the number of initial tuples. That is, only one sample is "left out" at a time for the test set. In stratified cross-validation, the folds are stratified so that the class distribution of the tuples in each fold is approximately the same as that in the initial data.

In general, stratified 10-fold cross-validation is recommended for estimating accuracy (even if computation power allows using more folds) due to its relatively low bias and variance.[7]

5.2 Training Dataset for DT

As shown in Table 5.1 dataset comprises of attributes like Minimum-Temperature, Maximum-Temperature, Morning-Humidity, Evening-Humidity, Wind Speed and Precipitation. This data

set is used for training DT models using ID3, C4.5 and CART respectively. Further dataset are given in appendix A.

Table: 5.1 Sample dataset for training

Morning Humidity	Evening Humidity	Evaporation	Wind	Min Temperature	Max Temperature	Precipitation
98.4	56.3	DNA	0.7	0.3	20.2	NO
100	56.6	2.4	0.8	0.8	21	NO
100	54.7	1.7	1.2	8.3	16	YES
100	49.6	1.6	0.8	1.7	21.2	YES
100	52.5	1.6	1	1.4	22.5	NO
100	51.2	2.4	1	1	22	NO
98.6	54.6	2.3	1.1	3.8	16.2	NO
95.5	44.7	DNA	-99.9	9	25.2	YES

5.3 Training of DT

This dataset has been collected from the department of meteorology and Hydrology of Nepal. Here we use different input parameters like wind speed, min temperature, max temperature, humidity (For morning and evening) and precipitation of the Kathmandu valley satiation "Tribhuvan International Airport" from the year 1999 to 2008 A.D. All the data is collected of daily base. Cross Fold 10 validation technique is used for the model validation which partitions dataset into 10 partitions. Randomly 9 partitions are chosen as training dataset and remaining single partition is used for testing. So here, 90% of total dataset is used for the training of decision tree and 10% as testing. This model building process is repeated for 10 times since it is 10 fold cross validation. In order to train in decision tree, we calculate the entropy and information gain. Information gain is used to select the root node.

Here, we choose the ID3, C4.5 and CART separately which uses Gain Information, Gain Ratio and Gini Index respectively to train the decision tree model.

Here it is shown for calculation of entropy and information gain, Gain Ratio and Gini Index for selection of root node.

Gain Information

Entropy(D) or Info(D)= $-\sum_{i=1}^{m} p_i \log_2(p_i)$ where m represents number of class labels.

$$= -\frac{2058}{3650}\log_2(\frac{2058}{3650}) - \frac{1592}{3650}\log_2(\frac{1592}{3650})$$

= 0.988209924 bits.

$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$ Where v is discrete values of A.

$\text{Info}_{\text{Min-Temp}}(D) = \sum_{j=1}^{3} \frac{|D_j|}{|D|} \times Info(D_j)$

$$= \frac{|D_1|}{|D|} \times Info(D_1) + \frac{|D_2|}{|D|} \times Info(D_2) + \frac{|D_3|}{|D|} \times Info(D_3)$$

$$= \frac{1639}{3650}\left(-\frac{188}{1639}\times\log_2(\frac{188}{1639}) - \frac{1451}{1639}\times\log_2(\frac{1451}{1639})\right) + \frac{674}{3650}\left(-\frac{291}{674}\times\log_2(\frac{291}{674}) - \frac{383}{674}\times\log_2(\frac{383}{674})\right)$$

$$+ \frac{1337}{3650}\left(-\frac{1113}{1337}\times\log_2(\frac{1113}{1337}) - \frac{224}{1337}\times\log_2(\frac{224}{1337})\right)$$

= 0.651794703 bits

Gain(Min Temp) = Info(D)-$\text{Info}_{\text{Min-Temp}}$(D)

= 0.336415221 bits.

Similarly Gain of every other attributes can be calculated and the attribute with the highest gain is selected as a root node. This node best partitions the dataset.

In the context here, Min Temperature is selected as root node. Further we calculate the information gain with the help of entropy for different values of Min Temperature. And we select

Evening Humidity as second root node as shown in figure 5.1. In this similar way the entire tree is created using WEKA.

Gain Ratio

$$\text{SplitInfo}_{\text{Min\_Temp}}(\text{D}) = -\sum_{j=1}^{3} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

$$= -\frac{1639}{3650} \times \log_2\left(\frac{1639}{3650}\right) - \frac{674}{3650} \times \log_2\left(\frac{674}{3650}\right) - \frac{1337}{3650} \times \log_2\left(\frac{1337}{3650}\right)$$

= 1.12847 bits.

$$\text{Now, Gain Ratio} = \frac{Gain(Min\_Temp)}{SplitInfo(Min\_Temp)}$$

= 0.298116324

Similarly, Gain Ratio for every other attributes can be calculated and the attribute having the highest gain ratio is selected as a root node. This process is repeated for each partitioned dataset to select the next node in next level of decision tree. In this similar way, the entire tree is created using WEKA.

Gini Index

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^{\,2}$$

Here, value of m is 2.

$$= 1 - (p_1^{\,2} + p_2^{\,2})$$

$$= 1 - \left(\left(\frac{1592}{3650}\right)^2 + \left(\frac{2058}{3650}\right)^2\right)$$

= 0.49185

Gini Index for attribute Min_Temperature is calculated for each binary partition of attribute values. Here for example, we take {-inf to 11.95, 11.95 to 17.35} and {17.35 to inf}that partition the dataset D in to D1 and D2 respectively. Then the Gini Index value based on this partition is:

$$\text{Gini}_{\{\text{-inf to } 11.95,\ 11.95 \text{ to } 17.35\}\varepsilon\{17.35 \text{ to inf}\}} = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

$$= \frac{2313}{3650}\left(1 - \left(\frac{479}{2313}\right)^2 - \left(\frac{1834}{2313}\right)^2\right) + \frac{1337}{3650}\left(1 - \left(\frac{1113}{1337}\right)^2 - \left(\frac{224}{1337}\right)^2\right)$$

$$= 0.310287629$$

Similarly, the Gini index values for splits on the remaining subsets can be calculated. And the Gini index of subset with highest value is selected because it minimizes the Gini index. And this process is repeated for every other attribute in order to select the best partition node in each level of decision tree. In this similar way, the entire tree is created using WEKA.

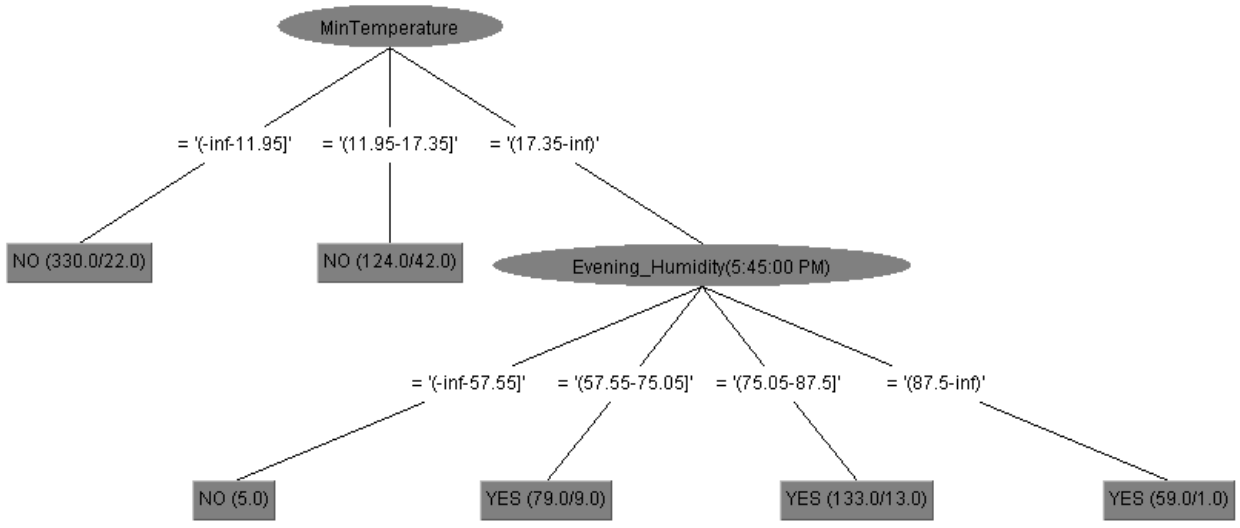


Fig: 5.1 Final DT.

5.4 Sample of Rule Generated by DT

After the training for DT using the above sample dataset, it generates a certain rule as shown in Table 5.1.

Table: 5.2 Rule generated by DT

Min Temperature	**Max Temperatue**	**Morning Humidity**	**Evening Humidity**	**Evaporation**	**Wind speed**	**Precipi-Tation**
17.35-inf			57.55-75.05			YES
17.35-inf			75.05-87.5			YES
17.35-inf			87.5-inf			YES
-inf - 11.95						NO
11.95-17.35						NO
17.35-inf			-inf-57.55			NO

5.5 Testing Dataset

In order to test DT model, 10% of the total dataset is selected in every fold and this process is repeated for 10 times since it is 10-fold cross validation. Table 5.2 shows the sample dataset for testing the DT.

Table: 5.3 Sample dataset for testing.

Morning Humidity	Evening Humidity	Evaporation	Wind	Min Temperature	Max Temperature	Precipitation
0.481034	0.714829	DNA	0	0.644269	0.764706	NO
0.051724	0.460076	0.97561	0.839607	0.592885	0.702206	YES
0.67931	0.577947	1.073171	0.828969	0.703557	0.742647	NO
0.562069	0.594106	0.926829	0.828969	0.695652	0.790441	NO
0.67069	0.571293	DNA	0	0.758893	0.794118	NO
0.632759	0.5827	DNA	0	0.865613	0.886029	NO

5.6 Testing of DT

The decision tree was tested on the basis of dataset shown in Table 5.1. The data were fed as input to the trained Decision tree. Cross fold validation generates the testing dataset by itself by partitioning the dataset in the ratio of 1:10 since 10-fold cross validation technique is used. The predicted output was verified against the actual output.

5.7 Result Analysis

The analysis was carried out on the basis of precision, recall, average accuracy, average speed and also based on the volume (small and large) and nature (nominal and numeric) of dataset. The parameters required for carrying out analysis like True Positive, True Negative, False Positive and False Negative were as provided by confusion matrix output. And they are resented as shown in Table 5.4, Table 5.5 and Table 5.6 for ID3, C4.5 and CART used in Decision Tree Induction respectively.

Table: 5.4 Analysis Parameters for DT using ID3.

Dataset	TP		TN		FP		FN	
	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric
Small	327	No result	207	No result	52	No result	49	No result
Large	697	No Result	328	No result	83	No result	66	No result

Table: 5.5 Analysis Parameters for DT using C4.5.

Dataset	TP		TN		FP		FN	
	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric
Small	390	384	248	255	64	57	28	34
Large	1736	1736	1210	1209	382	383	322	322

Table: 5.6 Analysis Parameters for DT using CART.

Dataset	TP		TN		FP		FN	
	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric
Small	390	390	248	248	64	64	28	28
Large	1711		1243		349		347	

After applying all three attribute selection measures: Information Gain, Gain Ratio and Gini Index used by ID3, C4.5 and CART respectively for 10-fold cross validation, the result for precision, recall, average system accuracy, system error and system speed calculated is given in the Tables.

Table 5.7 has values for small dataset and Table 5.8 has values for large dataset.

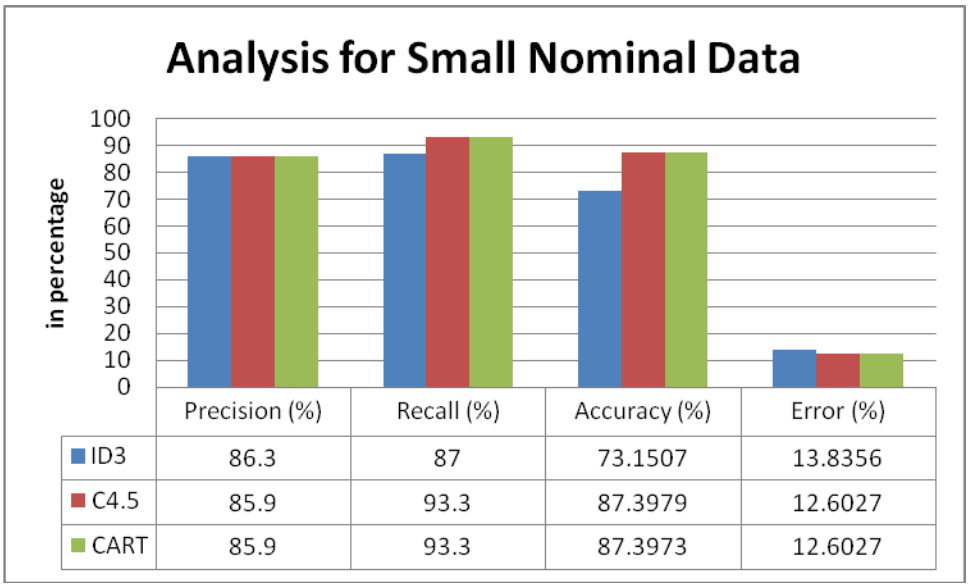
Table: 5.7 Experimentation Results for small dataset.

Algorithm	Performance In Second		Precision		Recall		Average System Accuracy (%)		Error (%)	
	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric
ID3	0.01	No	0.863	No	0.87	No	73.1507	No	13.8356	No
C4.5	0.5	0.3	0.859	0.871	0.933	0.919	87.3979	87.5342	12.6027	12.4658
CART	0.56	0.6	0.859	0.859	0.933	0.933	87.3973	87.3973	12.6027	12.6027

Table: 5.8 Experimentation Results for large dataset.

Algorithm	Performance In Second		Precision		Recall		Average System Accuracy (%)		Error (%)	
	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric
ID3	0.32	No	0.894	No	0.913	No	28.0822	No	4.0822	No
C4.5	0.05	0.6	0.82	0.819	0.844	0.844	80.7123	80.6849	19.2877	19.3151
CART	251.82	253.73	0.831	0.831	0.831	0.831	80.9315	80.9315	19.0685	19.0685

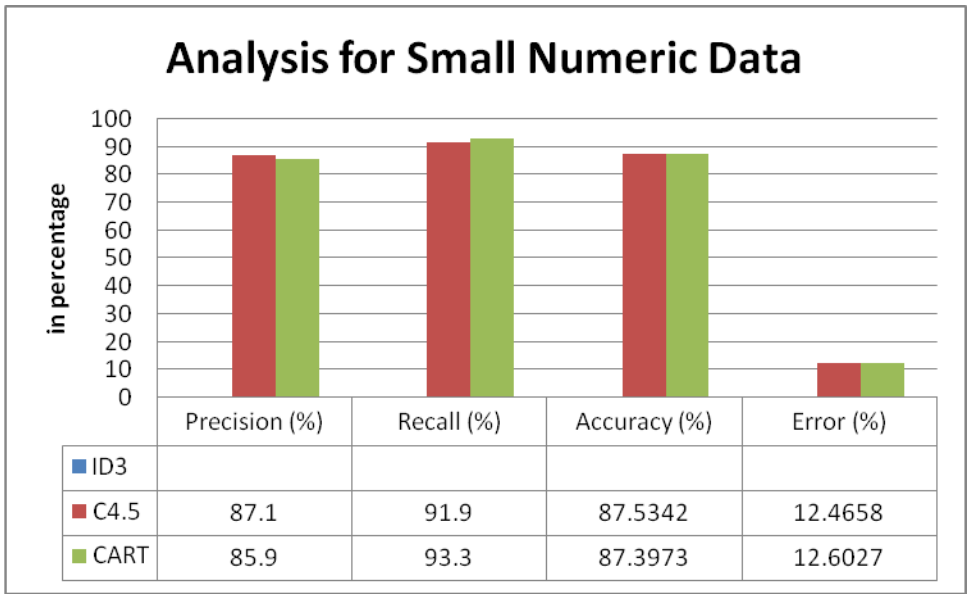
As shown in Fig 5.2, precision, recall, average system accuracy and error for ID3, C4.5 and CART for two years of nominal data is represented in graphical way for the comparison.



	Precision (%)	Recall (%)	Accuracy (%)	Error (%)
ID3	86.3	87	73.1507	13.8356
C4.5	85.9	93.3	87.3979	12.6027
CART	85.9	93.3	87.3973	12.6027

Figure 5.2 Graph of Experimentation

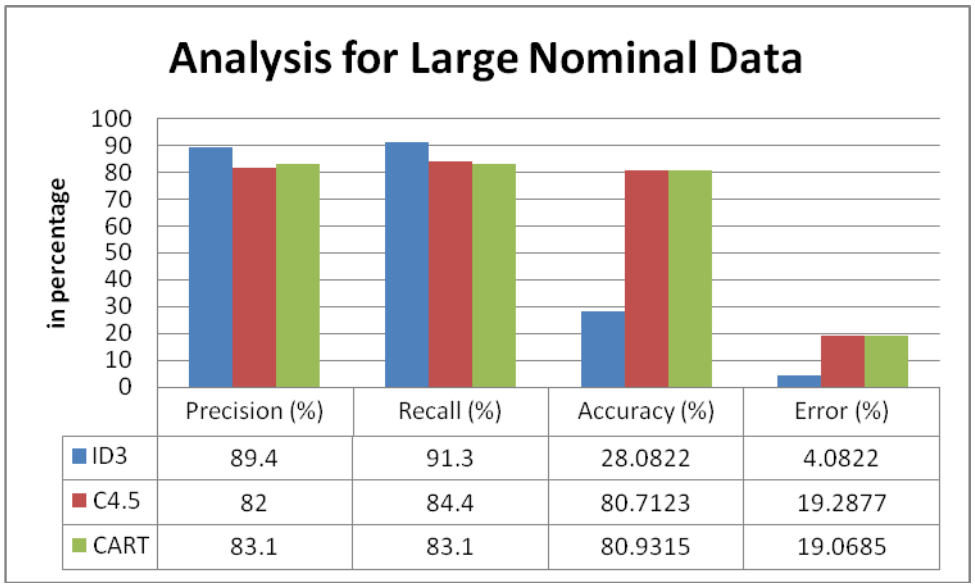
As shown in Fig 5.3, precision, recall, average system accuracy and error for ID3, C4.5 and CART for two years of numeric data is represented in graphical way for the comparison



	Precision (%)	Recall (%)	Accuracy (%)	Error (%)
ID3				
C4.5	87.1	91.9	87.5342	12.4658
CART	85.9	93.3	87.3973	12.6027

Figure 5.3 Graph of Experimentation.

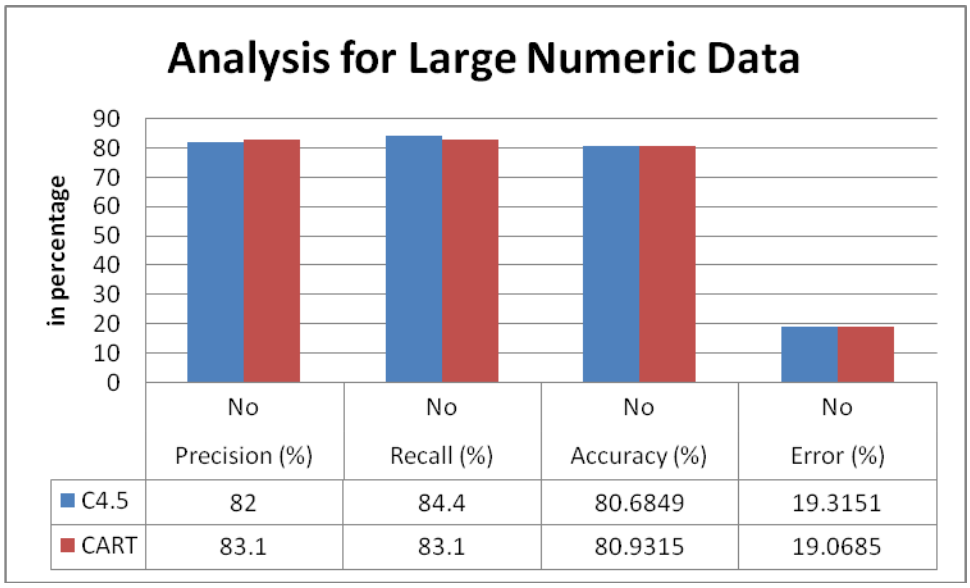
As shown in Fig 5.4, precision, recall, average system accuracy and error for ID3, C4.5 and CART for ten years of nominal data is represented in graphical way for the comparison.



	Precision (%)	Recall (%)	Accuracy (%)	Error (%)
ID3	89.4	91.3	28.0822	4.0822
C4.5	82	84.4	80.7123	19.2877
CART	83.1	83.1	80.9315	19.0685

Figure 5.4 Graph of Experimentation

As shown in Fig 5.4, precision, recall, average system accuracy and error for ID3, C4.5 and CART for ten years of numeric data is represented in graphical way for the comparison.



	No Precision (%)	No Recall (%)	No Accuracy (%)	No Error (%)
C4.5	82	84.4	80.6849	19.3151
CART	83.1	83.1	80.9315	19.0685

Figure 5.5 Graph of Experimentation.

As shown in Fig 5.6, the performance in the form of speed for ID3, C4.5 and CART for two years of nominal data is represented in graphical way for the comparison.

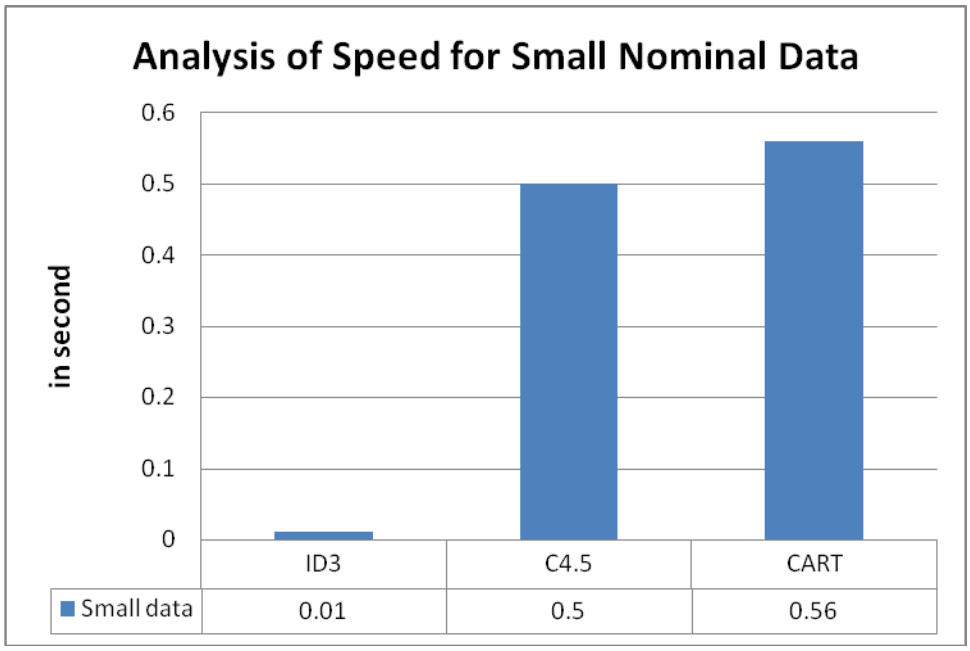


Figure 5.6 Graph of Experimentation

As shown in Fig 5.7, the performance in the form of speed for ID3, C4.5 and CART for ten years of nominal data is represented in graphical way for the comparison.

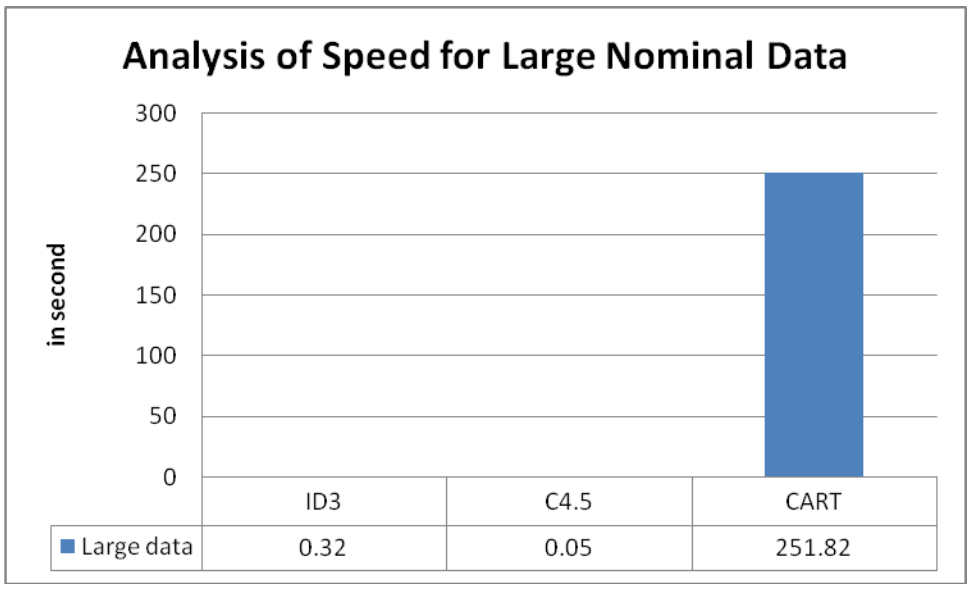


Figure 5.7 Graph of Experimentation

In the above, results show that the performance of different algorithms i.e. ID3, C4.5 and CART are different with the size and nature of data for building Decision Tree. From accuracy point of view, it shows that with the growth of data size, CART algorithm is better than C4.5 and C4.5 is better than ID3. And similarly from the speed point of view, with small size of data ID3 (with slightly low accuracy) is better than all other but with growth of data size, the C4.5 (with slightly low accuracy than CART) has better speed than other two algorithms. With larger size of data, CART has the worst speed though it has good accuracy. So generalizing the result, ID3 does not work with numerical attributes. For small size of nominal data, C4.5 and CART perform equally well. And with larger dataset, CART has good accuracy than other two algorithms (but it has very poor speed).

Result of system is influence by the number of training and testing data and extracted features. Number of parameters also plays important roles for better learning the system. So far- so good, results are promising and can be enhanced further.

Chapter 6

Conclusion Limitation and Future Work

6.1 Conclusion

Performance measure of attribute selection for decision tree induction is addressed in this dissertation work. Here it is used three algorithms for the decision tree classifier. These three algorithms or attribute selection measures are ID3, C4.5 and CART. Here, secondary data is used that is collected from the department of Hydrology and meteorology and is used of airport satiation of Kathmandu valley. For all the algorithms same dataset is used and dataset is used of ten years. For large dataset it is of ten years and for small dataset it is of two years. All the attributes are numerical values initially and missing values are labeled as DNA.

Before training the system, data is preprocessed. In the given dataset we have used 6 input parameters namely Min Temperature, Max Temperature, Morning Humidity, Evening Humidity, Evaporation and wind speed. All three algorithms use same input parameters.

From accuracy point of view empirical results shows that, with growing dataset, CART Decision Tree performs slightly better than C4.5 Decision tree and its predecessor ID3 Decision Tree. CART Decision Tree has the average system accuracy rate of 80.9315%, system error rate of 19.0685%, and precision rate of 83.1% recall rate of 83.1%. C4.5 Decision Tree has the average system accuracy rate of 80.6849%, system error rate of 19.3151%, and precision rate of 82% recall rate of 84.4%. Similarly, ID3 Decision Tree has the average system accuracy rate of 28.0822%, system error rate of 4.0822% for nominal data, and precision rate of 89.4% recall rate of 91.3%. With smaller dataset ID3 seems to have higher accuracy (missing values should be processed), here it has 73.1507% of accuracy in 2 years of data and it can be increased with decreasing size of input dataset.

From speed point of view empirical results show that, with growing dataset, CART Decision Tree performs very poor. It takes almost 250 seconds to build a model and for 10-fold validation, it is almost ten times longer. For this scenario, C4.5 is better than all other since it takes almost 0.05 seconds to build the model. With growing dataset ID3 cannot perform faster. Here it takes almost 0.32 seconds. But for smaller dataset ID3 can be very fast. Here, with particular dataset of two years, it has provided result in 0.01seconds with the least time than other two algorithms

6.2 Limitation and Future work

- The accuracy of the system can be increased by using other parameters like sea level, pressure, dew point and global warming factors. And in order to have the role of global warming, it would be better to perform test on current data as far as possible. These parameters could not be incorporated in the current thesis work due to unavailability of data.
- Thesis doesn't work on the calculation of intensity of rainfall. Precipitation has either zero or non-zero values. Zero value has been considered as class label "NO" and non-zero value has been considered as class label "Yes" for representing rainfall.

References

[1] Badr HSSINA, Abdelkarim MERBOUHA, Hanane EZZIKOURI, Mohammed ERRITALI TIAD, "*A comparative study of decision tree ID3 and C4.5*" (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications

[2] Kamlesh Dhayal, Role of Feature Selection in Data Filtering: A Comparative Analysis (Master Thesis, Computer Science and Engineering Department, Thapar University, Patiala-147004, 2009).

[3] Nikita Patel, Saurabh Upadhyay, "*Study of Various Decision Tree Pruning Methods with their Empirical Comparison in WEKA*" International Journal of Computer Applications (0975 – 8887) Volume 60– No.12, December 2012.

[4] V. Vaithiyanathan, K. Rajeswari, Kapil Tajane, Rahul Pitale, "*Comparison of Different Classification Techniques Using Different Datasets*" International Journal of Advances in Engineering & Technology, May 2013. ©IJAET, ISSN: 2231-1963.

[5] J. Han and M. Kamber, Data Mining: Concepts and Techniques.

[6] Rajesh Kumar, Ph.D "*Decision Tree for the Weather Forecasting*" International Journal of Computer Applications (0975 – 8887) Volume 76– No.2, August 2013 31.

[7] Jiawei Han | MichelineKamber | jian Pei "Data Mining Concepts and Techniques" Third Edition Chapter 8: Classification: Basic Concepts.

[8] Asst. Prof. Rajesh Kumar, "*Decision Tree for the Weather Forecasting*" International Journal of Computer Applications (0975 – 8887) Volume 76– No.2, August 2013.

[9] Bharat, V., Shelale, B., Khandelwal, K., Navsare, S., *"A Review Paper on Data Mining Techniques*", International Journal of Engineering Science and Computing (IJESC), Volume. 6, Issue.5, May 2016, pp: 6268-6271.

[10] Nishant Mathur, Sumit Kumar, Santosh Kumar, Rajni Jindal, "*The Base Strategy for ID3 Algorithm of Data Mining Using Havrda and Charvat Entropy Based on Decision Tree*", IJIEE 2012 Vol.2(2): 253-258 ISSN: 2010-3719 DOI: 10.7763/IJIEE.2012.V2.93.

[11] Anuja Priyam, Abhijeet, Rahul Gupta, Anju Rathee, and Saurabh Srivastava, "*Comparative Analysis of Decision Tree Classification Algorithms*" International Journal of Current Engineering and Technology ISSN 2277 - 4106 © 2013 INPRESSCO.

[12] Batuhan Baykara, Impact of Evaluation Methods on Decision Tree Accuracy

(M.Sc. Thesis, University of Tampere School of Information Sciences Computer Science, April 2015).

[13] Ashish Kumar, Pranav Bhatia, Anshul Goel, Silica Kole, "*Implementation and Comparison of Decision Tree Based Algorithms*" International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 4, Special Issue May 2015.

[14] Badr HSSINA, Abdelkarim MERBOUHA, Hanane EZZIKOURI, Mohammed ERRITALI, "*A comparative study of decision tree ID3 and C4.5*" (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications.

[15] Bhumika Gupta, Aditya Rawat, Akshay Jain, Arpit Arora, Naresh Dhami, "*Analysis of Various Decision Tree Algorithms for Classification in Data Mining*" International Journal of Computer Applications (0975 – 8887) Volume 163 – No 8, April 2017.

[16] Narasimha Prasad, Prudhvi Kumar Reddy, Naidu MM, "*An Approach to Prediction of Precipitation Using Gini Index in SLIQ Decision Tree*", 4th International Conference on Intelligent Systems, Modeling & Simulation, Bangkok, pp.56-60, 2013.

[17] Elia Georgiana Petre "A Decision Tree for Weather Prediction", Buletinul, Vol. LXI No. 1, 77-82, 2009.

[18] Kaya, E.; Barutçu, B.; Menteş, S. "A method based on the van der Hoven spectrum for performance evaluation in prediction of wind speed". Turk. J. Earth Science, 22, 1–9, 2013.

[19] P.Hemalatha, "*Implementation of Data Mining Techniques for Weather Report Guidance for Ships Using Global Positioning System*", International Journal Of Computational Engineering Research Vol. 3 Issue. 3 , march 2013.

[20] Soo-YeonJi, Sharad Sharma, Byunggu Yu, Dong Hyun Jeong, "*Designing a Rule Based Hourly Rainfall Prediction Model*", IEEE IRI 2012, August – 2012.

[21] S. Kannan , SubimalGhosh, "Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output", Springer-Verlag, July- 2010.

[22] Kumar Abhishek1, Abhay Kumar2, Rajeev Ranjan, Sarthak Kumar, "*A Rainfall Prediction Model using Artificial Neural Network*" 2012 IEEE Control and System Graduate Research Colloquium (ICSGRC 2012)

[23] Ian H. Witten, Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques (ELSEVIER, Second Edition).

[24] NishantMathur, Sumit Kumar, Santosh Kumar, and Rajni Jindal, "*The Base Strategy for ID3 Algorithm of Data Mining Using Havrda and Charvat Entropy Based on Decision Tree*" International Journal of Information and Electronics Engineering, Vol. 2, No. 2, March 2012.

Appendix A

Sample dataset before normalization

Morning Humidity	Evening Humidity	Evaporation	Wind	Min Temperature	Max Temperature	Precipitation
98.4	56.3	DNA	0.7	0.3	20.2	NO
100	56.6	2.4	0.8	0.8	21	NO
100	46.8	0.6	0.7	0.5	23.6	NO
98.4	47	1.9	0.7	1.2	22.2	NO
96.9	54.9	DNA	-99.9	1	21.7	NO
96.8	58.3	DNA	-99.9	0.2	21	NO
100	75	2.8	0.7	0.8	13.6	NO
100	52.3	DNA	-99.9	1	20.1	NO
100	61.6	DNA	-99.9	2.3	14.8	NO
100	58.8	0.4	0.6	1.2	17.1	NO
100	57.3	7.3	1	-2	20.4	NO
100	42.9	2.3	0.8	-1.4	21	NO
100	48.6	DNA	0.9	-0.2	20.2	NO
98.4	49.3	1.1	0.9	-0.9	21	NO
98.4	50.2	2.9	1	0	21.1	NO
100	50.2	8.2	3.9	0.5	21	NO
96.9	45.6	10.3	-99.9	1.1	21	NO
82.6	49.7	DNA	-99.9	1.4	21.7	NO
91.1	49.2	DNA	-99.9	1.2	21.5	NO
100	48.2	DNA	-99.9	1.2	21.5	NO
98.5	45.3	2.2	1	2.2	23	NO
98.6	41	2.2	1	2.1	23	NO
98.4	41.3	1.3	1.4	1.4	24.4	NO
81.4	41.2	3.1	1.4	2.6	23.5	NO
100	46.2	2.3	1.9	2.7	24.3	NO
100	46.3	3.4	1.2	3.2	22.6	NO
91.4	54.5	0.6	1.3	2.2	23.2	NO
100	59.1	DNA	1.2	2.6	22.2	NO
100	54.7	1.7	1.2	8.3	16	YES
100	49.6	1.6	0.8	1.7	21.2	YES
100	52.5	1.6	1	1.4	22.5	NO
100	51.2	2.4	1	1	22	NO
98.6	54.6	2.3	1.1	3.8	16.2	NO
100	52.3	0.7	0.3	2.5	20.4	NO
91.4	39.3	DNA	-99.9	2.5	24	NO
76.1	31.3	DNA	-99.9	7	25.2	NO
100	29.3	DNA	-99.9	3.6	23.8	NO

98.6	48.4	2.9	1.4	2.8	24.5	NO
100	55.3	2.4	1.2	4.3	24.4	NO
95.3	51.8	DNA	0.9	8.1	25.6	NO
88.5	49.1	DNA	0.8	7.7	26.6	NO
97.6	38.6	2	1	6.1	27	NO
93.8	38.6	1.8	1.1	5.9	28.7	NO
97.4	26.9	8.3	1.2	5.2	28.5	NO
93.8	38.8	0.8	1.3	5.4	28.4	NO
90.3	53	DNA	-99.9	6.7	26.6	NO
100	68.8	DNA	-99.9	7.3	28.2	NO
98.7	24.5	DNA	-99.9	5	29	NO
92.4	39.4	2.8	1.5	5.2	28.2	NO
97.5	40.3	3.2	1.4	5.7	27.2	NO
100	35.8	3.9	1.3	5.2	27	NO
90.6	54.7	DNA	2.1	8.9	26.4	NO
88.4	58.1	DNA	-99.9	7	27.1	NO
93	58.1	DNA	-99.9	7.9	25.5	NO
100	49.3	DNA	0.6	7.9	26.9	NO
87.3	43.3	2.7	1.3	8.7	25.1	NO
83.8	64.7	17.8	1.4	10.2	25.7	NO
95.5	44.7	DNA	-99.9	9	25.2	YES
87.9	36.9	DNA	-99.9	6	26	NO
87.2	32.5	DNA	-99.9	3.6	24.8	NO
87.4	50.5	DNA	-99.9	4.1	25.8	NO
81.9	46.1	DNA	-99.9	6.8	27.2	NO
95.2	47.4	DNA	-99.9	7.6	28.2	NO
82.9	54.3	DNA	-99.9	8.2	28.1	NO
97.6	47.6	3.4	1.6	8.8	29.2	NO
98.9	50	3	1.6	9.5	29	NO
91.5	37	DNA	1.5	10.3	31	NO
71.4	30.7	DNA	-99.9	8.7	31.1	NO
72.8	16.6	DNA	-99.9	8.2	31	NO
50.6	20.9	DNA	-99.9	7.1	29.5	NO
52	26.5	DNA	-99.9	6.3	29.1	NO
63.1	26.4	DNA	-99.9	5.5	28.1	NO
64.3	38.4	4.5	1.9	4.1	28	NO
73.9	48.9	DNA	-99.9	6.4	25	NO
78.4	48	DNA	-99.9	7.5	26.4	NO
76.9	25.6	3.1	1.5	7.4	29.5	NO
54	21.8	4.5	2.5	6.8	30	NO
62.4	37.6	2.2	2.1	6.3	30.2	NO
68.1	27.6	DNA	-99.9	6.6	29.7	NO
69.9	36.5	DNA	-99.9	4.8	28.7	NO

71	35.5	3.8	1.4	6	27	NO
77.1	40.3	1.9	1.3	5.7	27	NO
73.1	44.7	3.9	1.7	7.2	27.5	NO
76	30.4	5	1.7	8.5	28.5	NO
74.9	31.4	DNA	-99.9	8.6	28.6	NO
61.4	55	DNA	-99.9	7.7	27	NO
69.7	43.9	DNA	-99.9	8	27.5	NO
74.3	45	DNA	-99.9	11.2	28	NO
71.1	51.7	DNA	1.2	10.5	26.6	NO
73.8	54.1	2.6	0.9	13	29	NO
73.3	48.3	2.7	1.4	12.3	32	NO
76.9	30.1	5.3	1.8	12.8	32	NO
73.3	22.7	5.4	1.7	10.5	33	NO
57.2	20.6	7	2.7	12	33.2	NO
60.2	29.5	6.8	2.1	10.5	32.3	NO
42	35.6	4.5	1.6	8.6	30.3	NO
66.7	42.5	DNA	-99.9	10.6	29.8	NO
69.9	70	DNA	-99.9	13.3	29.2	NO
45	43.2	4	2.7	12	27.5	YES
81.4	55.6	4.4	1.4	14.8	28.6	NO
74.6	57.3	3.8	1.4	14.6	29.9	NO
80.9	54.9	DNA	-99.9	16.2	30	NO
78.7	56.1	DNA	-99.9	18.9	32.5	NO
75.8	40.9	5	2.8	16.6	33.8	NO
59.9	32.9	DNA	-99.9	14.5	34.4	NO
64.6	24	DNA	-99.9	11.9	32.1	NO
55.8	24.8	4.2	1.4	9.7	31.1	NO
61.4	28.5	5.8	1.6	10.3	30	NO
53.7	38.2	3	1.1	10.2	29.5	NO
64.8	42	DNA	-99.9	12.7	30.6	NO
70.4	50.4	DNA	-99.9	15.9	31.8	NO
69.6	58.9	3.9	1.2	17.6	32.7	NO
73.9	58.7	4.2	1.5	17.6	33.5	NO
62.5	52	15.1	1.8	16.2	34.5	NO
61.1	41.4	DNA	1.7	15.6	34	NO
58.3	41.5	4.8	1.8	15.6	35.6	NO
57	45.1	4.6	1.7	16.5	34.7	NO
67.1	39.8	5.1	1.5	16.6	35	NO
57.3	59.9	1	2	15.7	34.5	NO
72.9	61.5	10.1	1.6	16.7	32.5	NO