NDTALGO: Enhancement of Decision Tree Algorithm Based on Mobile Services

A Dissertation Submitted

By

Sukhchain Singh

41200434

To

Department of Computer Science & Engineering

In the Partial Fulfillment of the Requirement for the Degree of

Master of Technology

In

Computer Science & Engineering

Under the guidance of

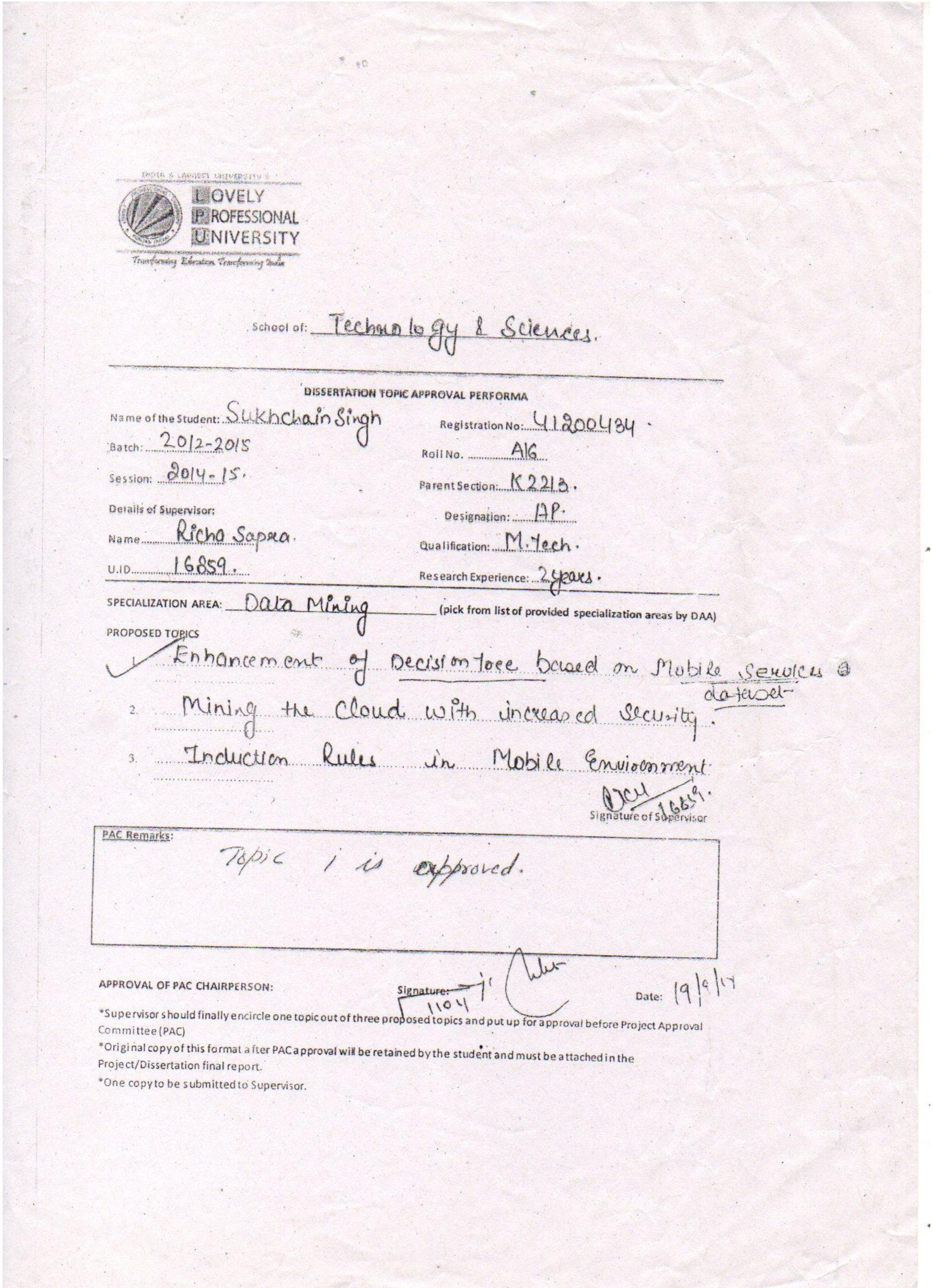
Manjit Kaur

UID 12438

Assistant Professor

Lovely Professional University, Phagwara, Punjab

May, 2015



INDIA'S LARGEST UNIVERSITY

LOVELY
PROFESSIONAL
UNIVERSITY

Transforming Education Transforming India

School of: Technology & Sciences.

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the Student: Sukhchain Singh

Registration No: 41200434.

Batch: 2012-2015

Roll No. A16

Session: 2014-15.

Parent Section: K2213.

Details of Supervisor:

Designation: AP.

Name Richa Sapra.

Qualification: M.Tech.

U.ID 16859.

Research Experience: 2 years.

SPECIALIZATION AREA: Data Mining (pick from list of provided specialization areas by DAA)

PROPOSED TOPICS

✓ 1. Enhancement of Decision tree based on Mobile services dataset

2. Mining the Cloud with increased security.

3. Induction Rules in Mobile Environment

16859.
Signature of Supervisor

PAC Remarks:

Topic 1 is approved.

APPROVAL OF PAC CHAIRPERSON:

Signature: 1104

Date: 19/9/14

*Supervisor should finally encircle one topic out of three proposed topics and put up for approval before Project Approval Committee (PAC)

*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.

*One copy to be submitted to Supervisor.

ABSTRACT

Data Mining is most effective approach for extracting patterns, information and knowledge from large amount of distributed raw facts and figures. Data mining contains list of techniques for retrieving information from these vast dataset. Decision tree is also one technique from these available techniques. Decision tree works on constructing nodes starting from root node to leaf node. Decision tree is analysis techniques. There are many decision trees available in data mining approach. Each has its own advantage and disadvantages. The goal of my research is to create a decision tree which is enhanced version of all these decision trees. The Dataset is based on mobile services such as at particular location which service is available? The data is collected by using wirelessMon software. Apply already existing decision tree algorithm and enhanced decision tree on same dataset. The enhanced version of decision tree has greater accuracy as compared to C4.5 and CART decision tree.

CERTIFICATE

This is to certify that Sukhchain Singh has completed M.TECH Dissertation entitled "**Enhancement of Decision Tree based on Mobile Services**" under my guidance and supervision. To the best of my knowledge, the present work is the result of his original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma. The dissertation proposal is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech Computer Science & Engineering.

Date:

Signature of Advisor
Name: Manjit Kaur
UID: 12438

ACKNOWLEDGEMENT

I would like to thank my Almighty God, who has always blessed me & guided me to work on the right path of life and for giving me strength to do this work. A Warm Thanks to my mentor Miss. Manjit Kaur, Assistant Professor in Department of Computer Science & Engineering, Lovely Professional University, Phagwara, Punjab who helped me to give good ideas and support to complete my task. Without her unbreakable support, it is impossible to complete.

I would like to thanks my father who also helped me to complete my program by providing all multimedia facilities and creating peaceful environment at home during my report.

A special thanks to all faculty members of Computer Science & Engineering Department at Lovely Professional University who always ready to give assistance.

Student Declaration

I hereby declare that my Definition report entitled "**Enhancement of Decision Tree Algorithm based on Mobile Services**" Submitted for M.Tech Degree is entirely my original and bona fide work. No content in my report is copied from any sources. All Ideas and references are duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date:______________ __________________________

Sukhchain Singh

41200434

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER-1

INTRODUCTION

1.1. Need of Data Mining

We are human beings living in modern societies, where terrific amount of data stored as well as transferred from one place to another place. Now days, data storage reaches up to petabytes. All fields such as medical, Science, Agriculture, Engineering etc. are all need data to progress in their respective fields. Due to computerization, amounts of data increases very rapidly. Analysis tools to analyze data are also developed very fast. Every field has its own dataset, which includes list of transactions related to sales, records of stocks, description of equipment, products and items, information of companies and promoting sales guidelines. Consider an example of PVR Cinemas, where thousands of transactions are handled at one movie including ticket booking, availability of seats and advertising companies at one district. Generally these transactions are leads to billions. To fulfill these requirements leads to invention of Data Mining.

Search Engine processes millions of user queries in a day. These queries related multimedia such as list of hit movies in a year, list of albums released in a month. Medical field's queries such as symptoms of new disease and procurement of that disease. During processing of these queries some new patterns are generated. For example, a query related to AIDS executed in BING search engine. It results a list of people who are researching on AIDS as well as who as symptoms of AIDS. When all queries entered to search engine are executed together, a new pattern is generated.

There are lots of cases where new patterns are generated by aggregating queries. Agriculture fields use data mining to know which yield needs fertilizers more, in which yield has more production and comparison of seed samples queries.

Shopkeepers also take advantages of data mining techniques to test which schemes is better than other schemes, new methods of discount and coupon schemes (Jiawei Han, 2012).

1.2. What is Data Mining

Data mining seems to retrieving information from bulky volume of datasets. This information then can be used for decision making process. Data mining is an automated retrieving information process. Data mining assist to mine patterns from datasets. Generally row fact and data are not stored on same file format. Data are stored on different format such as Text File, Excel File, RDBMS, MS Access, Doc file and ppt files. Interesting thing is that data mining tool can extract patterns from all these files by comparing data from all file formats. Data mining tends is to extract meaningful information and transformed this information into more organized form that can be used for further process. Data mining is automotive procedure. Data Mining does not reduce the requirement of components to know your business, to interpret your data (Jiawei Han, 2012).

Many Researchers views data mining to be one stair in knowledge discovery from huge databases or KDD. To see where data mining lies in KDD see the figure shown below:

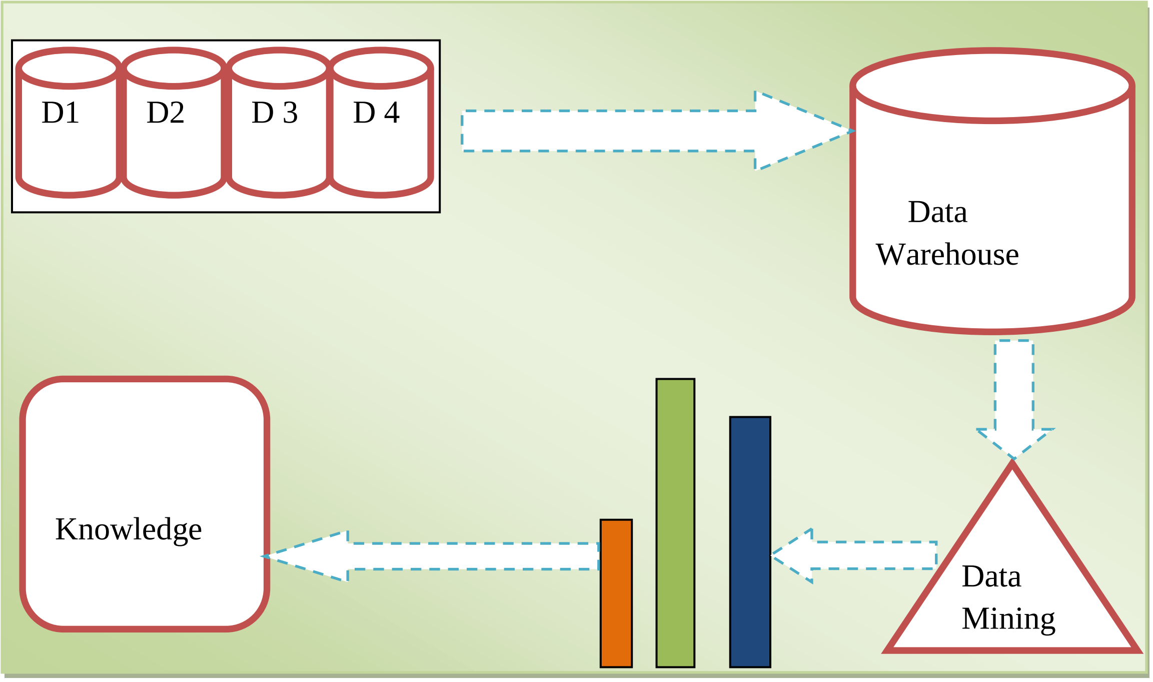


Figure 1.1: Data Mining Process in Knowledge Discovery

Other phases in KDD Process are:

- Elimination noise and inconsistency from data.
- Combination of various data source.
- Relevant data retrieved from database.
- Transformation of data.
- Extraction of data Patterns.
- Evaluation of patterns.
- Representation of patterns to user.

Types of Data can be mined:

- Data Warehouse Data
- Transactional Data
- Spatial Data
- Multimedia Data
- Text Data
- Database Data

Table 1.1: Example of Transactional Data

Transaction_ID	**List_of_Items_ID**
T100	I1, I4, I5, I9
T200	I2, I3, I4
T300	I1, I2, I3

To generate hidden patterns, data mining contains the following passes:

- Modeling the problem in simple words.
- Collect required data for problem solving and then prepare them.
- Create Model and evaluate this model.
- Deliver knowledge obtained from model.

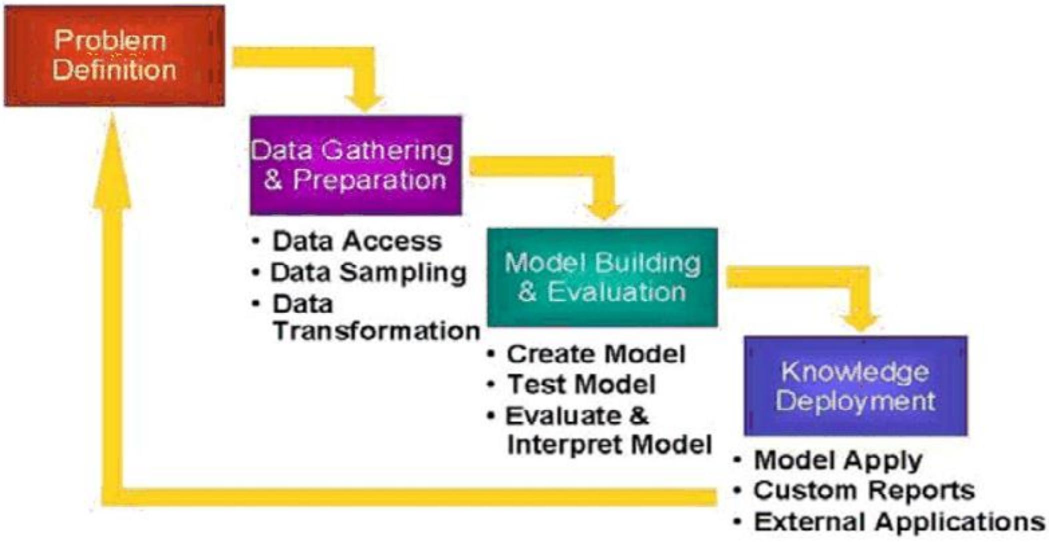


Figure 2.2: Data Mining Process

From Data to Information progress, many new achievements are added one step more from last one(Thearling, 2012).

Table 1.2: Evolution of Data to information

Evolutionary Step	Technology	Provider	Characteristic
Data Collection (1960s)	Tapes, Disks	CDC, IBM	Retrospective, static data delivery
Data Access (1980s)	SQL, RDBMS	Sybase, IBM, Microsoft	Retrospective, dynamic data delivery at record level

Data Warehousing & Decision Support (1990s)	OLAP, Multidimensional database, datawarehouses	Comshare, Arbor, Cognos, Pilot	Retrospective, dynamic data delivery at multiple level
Data Mining (Today)	Advance Algorithm, multiprocessor computers, massive databases	IBM, SGI, Pilot	Prospective, Proactive information delivery

1.3. Data Mining Techniques

Data Mining offers many techniques such as machine learning, statistics, extracting pattern, visualizing. Classification and clustering are two important aspects in machine learning are accomplished by data mining are:

Classification: Classification is a technique in which datasets splits into sub sets and belongs to specified target class.

For example: A customer goes to cloth house, ensure demand for some cloth whether jeans, shirts etc. He belongs to predefined class.

Clustering: In Clustering datasets are also splits and same data are grouped into many different classes according to its properties.

For example: Students with above 90 percents are grouped into same class regardless in subject they obtained (Jiawei Han, 2012).

1.4. Decision Tree

The most important data mining techniques is Decision Tree. It is based on Divide and conquers technique. This technique states divide the input into multiple pieces and test one by one until required outcome is obtained. Decision tree technique is a process of managing the set of rules into such a way that final outcome is obtained from leaf node by following above rules starting from root node. It is the simplest technique of data mining. Decision tree focuses on developing a model that estimates or guesses the value target attribute based on nominal and numeric data. Greedy Algorithm is an illustration of creating decision trees from top to down. It is the most general method for creating decision tree from available data sets, but we cannot say that it is only this one strategy for creating tree, many other strategy are also existing. Recently many other strategies are also come into existence for creating decision tree from bottom-up rules based (Jiawei Han, 2012).

Decision tree is constructed by creating root node from top and move downward by creating intermediate nodes until final leaf node is met. At every non-leaf node, a test condition is performed, depending upon its value, again sub list is created. This process is recursive process. At each node, it performs all the operations iteratively. It is very easy to convert decision tree into classification rule. Greedy algorithm is an example of creating decision tree from top to bottom (Lior Rokach, 2005).

Advantages of Decision Tree:

- It is the simplest method for determining target value.
- It is more readable.
- It is very easy to understand.
- It operates only on single table as well as one attribute at a time.

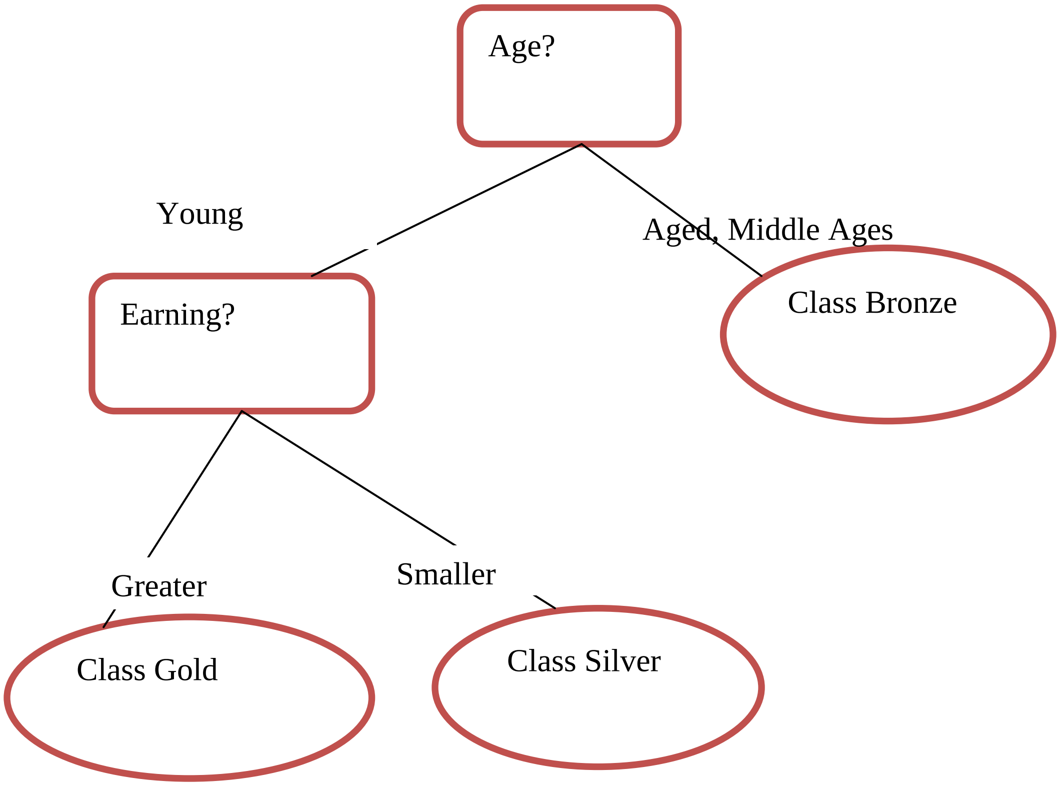
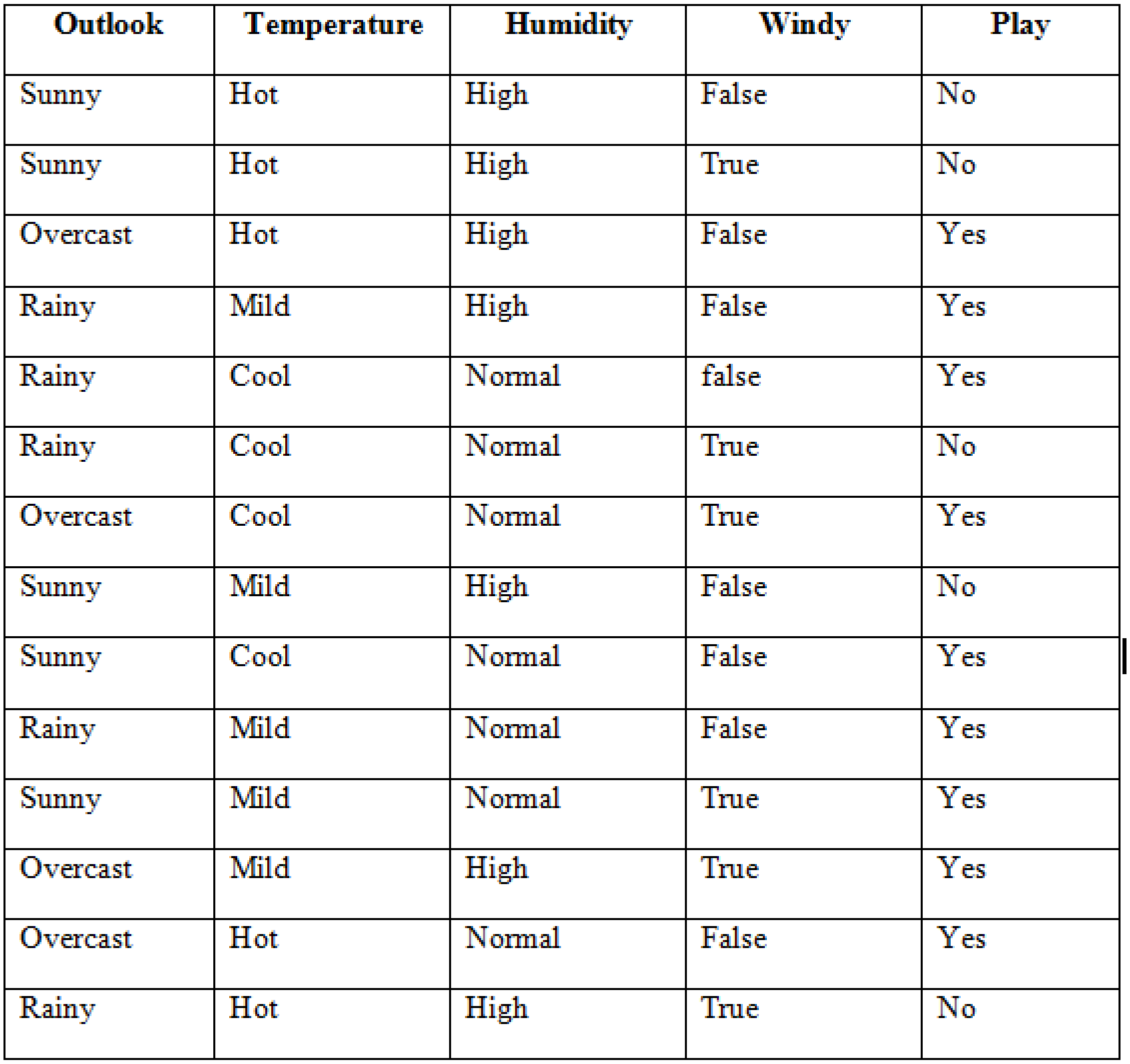


Figure 3.3: Example of Decision Tree

Creation of Decision Tree from available datasets:

I have a dataset obtained from weather report. This decision tree is based on when to play a match depends on weather of a day.

Table 1.2: Dataset of Weather



Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	false	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Hot	High	True	No

Play is a dependent variable. I have to test condition at each node whether play will be equal to Yes or No(Otero, 2005).

PLAY is Dependent Variable

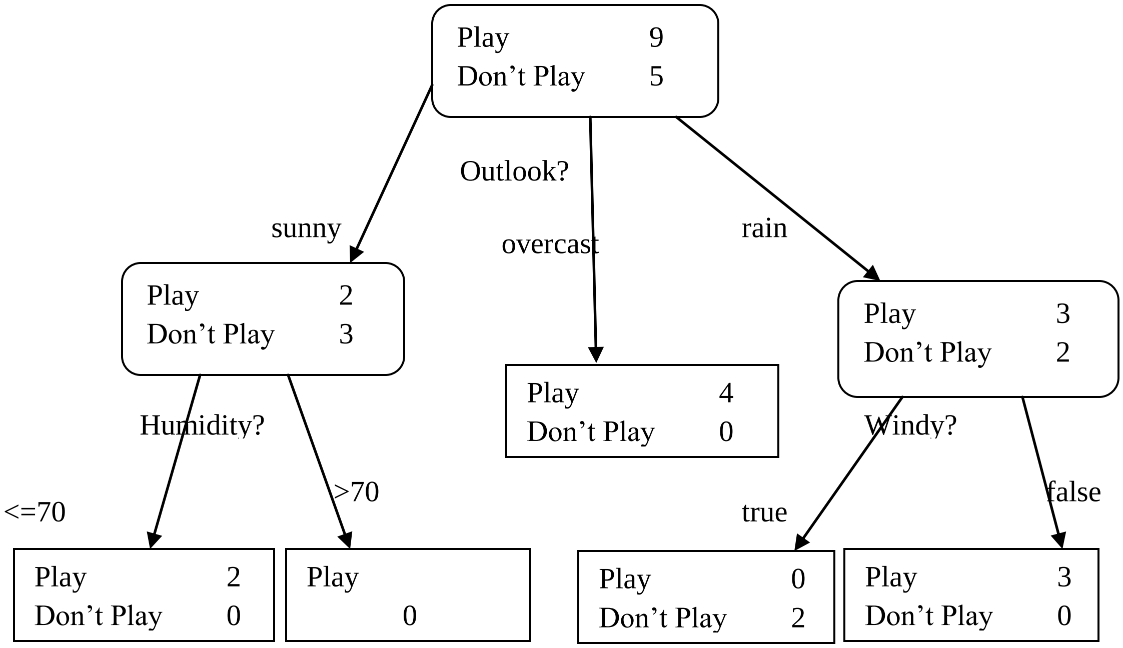


Figure 4.4: Decision Tree of Weather Dataset

Types of Decision Trees:

- ID3: stands for Iterative Dichotomiser 3.
- C4.5: It is an extension of ID3.
- CART: refers to Classification and Regression Tree analysis
- CHAID: Abbreviated for Chi-squared Automatic Interaction Detector.

MARS: It leads decision tree to further numerical data processing

ID3 and CART are discovered independently of each other; however both are using same approach for creating decision tree from available training data.

CHAID stands Chi-Squared Automatic Interaction Detector. CHAID is the oldest method among all the decision tree methods. It is developed by Kass in 1980. It was very accepted in selling research. It is used to create non-binary tree. Non-binary tree is a tree in which more than two arms can be attached to root node as well as single intermediate node. It uses very simple algorithm used to analyze larger datasets. CHAID is used to portioning in a recursive manner.

Algorithm for CHAID Decision Tree:

- Preparing Predictors.
- Merging categories.
- Selecting the split variable.

A variation to CHAID algorithm is known as Exhaustive CHAID Algorithm. It needs more calculating time than CHAID method (Lior Rokach, 2005).

C4.5decision tree is the advanced version ID3. C4.5 was developed by Quinlan. It is used to suggest Classification rules from available examples in the form of decision tree. The training data set is represent as S=s1, s2,….. . Vector C= c1, c2, ..are used to enhance the training data.C4.5 selects one attribute of the data at each non-node of the tree which breaks down samples into subsets. Its decisive factor is the information gain that results from selecting an attribute for splitting the data. Decision can be made by selecting the peek information gain attribute (Kalpesh Adhatrao, 2013).

Base cases:

- Sample Data must relevant to the identical class.
- No information gain is provided by any of the features.

Pseudo code

```
1. Evaluate effectively Base Cases
2. At attribute ccc
      a) Compute Information gain breaking on ccc.
3. Assume highest information gain attribute is ccc_best
4. Construct a node that breaks on ccc_optimal.
5. Recurs on the sub lists got by breaking on ccc_optimal, and insert those nodes as
   children.
```

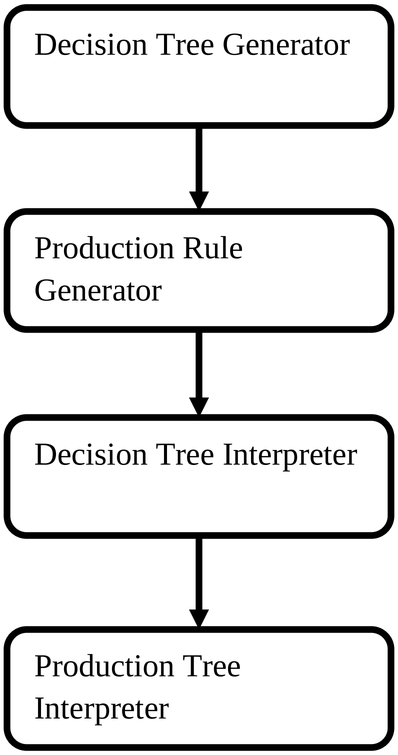

Decision Tree Generator
Production Rule Generator
Decision Tree Interpreter
Production Tree Interpreter

Figure 5.5: C 4.5 Phases

CHAPTER - 2
LITERATURE REVIEW

(Duong Van Hieu, 2014): This Paper states that Facebook have become the major source of information sharing from previous few years. Many of Children, Students and Employees busy in using Facebook for their satisfaction. In this Paper, Various factors are analyzed which influenced Facebook's users attitude and behaviors. These factors are Age, Sex, Income, Moving Devices and Tablets. It investigates how much times users spent on Facebook in this year as compared to previous year on the behalf of various factors using decision tree Techniques.

(Kimberly Carpenter, 2014): This research work aims to study from distinct Preschool Age Psychiatric Assessment (PAPA) the inter-correlations between diverse questions to offer consistent risk analysis by varying decision tree method with boosting to build small and interpretable decision rules. This altered algorithm gives reading of confidence as output. It is beneficial for clinics.

(Xue-Fei Zhang, 2013): This paper states the reasons of accident happened in Canada. The data is collected through primary source. To find out reasons Data mining techniques such as ID3and C4.5 trees are used. The resulted is obtained by implemented with WEKA Tool. The study is conducted to find out reasons for season, age and gender related accidents. There are mainly 11 factors on which research will identify the accidents percentage. The research identified that small age people will used to drink, avoid the rule and generally repeating their mistakes. People having age between 30 and 50 are got accident by not careful regarding traffic rules. They also have drunk during the accident time. The people having age more than 65 years are over confident which lead to them accident. One more reason in senior citizen's accident is their physical fitness. They get tired driving short time.

Regarding Season Accident, In Winter Season, There are more chances of accident of people who disobey the traffic rule and used drink before driving. In summer (non-winter) season accident are occurred due to drink, rule and mistakes. In summer, due to lack of experience, there are accidents.

In Gender Accidents, Drink, rule and mistakes are major factors for contributing accident in male group in Canada. Males are mostly disobeying the rule and get accident. In female, drink and rule are same as male category contribution, but due to weather, there are chances of accidents.

(Pui K. Fong, 2012): Solitude preservation is significant for machine knowledge and data mining, but test planned to defend private in order often result in a trade-off: abridged utility of the preparation sample. It proposes a new solitude preserving come up to by using decision tree, without linked loss of correctness. A different more privacy of protection technique of calm data samples are introduced where proof from the model database has been missing unique data sets has been changed to illusive data sets by this method, from which reversevely unique sample cannot be wrought without whole group of fantasy data sets. In the meantime, a correct decision tree can be built in a straight line from those unbelievable data sets. This new move toward can be useful openly to the data sets as soon as the main example is gathered. The move toward is compatible with other solitude preserving approaches, such as cryptography etc.

(Soldic-Aleksic, 2012): Clustering is similar to Segmentation. This research paper represents the final outcomes of clustering problem in the locale of marketing after applying the CHAID Decision tree model followed by Self Organizing Map (SOM). For Clustering and Visualization of market data K-SOM is applied. The results obtained after applying K-SOM are input to CHAID decision tree to interpret the Concrete Clustering.

(Sivaraman Sriram, 2012): Human Emotions can be classified by Joy, fear, love, anger, surprise and sadness. This Paper investigates enhanced version of text based human emotions classification and prediction with help of modified decision tree method. Decision Tree is broadly used technique in research in the field of data mining. The human emotions can be extracted from online chatting. A modified decision tree is used to classified available dataset by assigning it to two known class labels. Before this research, there are many algorithm developed to classified emotions. But the objective of this research is to modified decision tree to reduce the complexity, make simple and efficient in order to extract emotions and classified it. The efficiency of the proposed approach is determined by comparing it with already existing technique.

(Ji Dan, 2010): With increasing the information Technology field, Large Volume of data is available in our daily life. Decision tree and clustering are two techniques of data mining are used to extract meaningful information. The enhanced version of CURE and C4.5 algorithm is a CA Algorithm which is a synthesized data mining algorithm. CA is used to focus on Principle component processing and parallel processing whose goal is to obtain feature reduction and scale reduction from large datasets. This New algorithm is functional to Maize Seed Breeding. After obtaining the results we can judge CA algorithm is better than previous methods.

(Shiow-yang Wu, 2010): In contemporary days, many users are touching within mobile setting utilize wireless services for the winning act, activity, execution of daily wants and online shopping. We plan the idea of multifaceted activity meant for characterize the unremittingly altering fused behavior pattern of mobile users. A merged activity is treating as uninterrupted movement, service request of location for the object of data management. An activity may be additional consists of number of little activities. Diverse behavior may reveal dependencies that alter behaviors of users. We quarrel that the multifaceted activity provides a more precise, affluent, and state report of user behavioral patterns which are valuable for data organization in mobile environments. Suitable exploration of client activities has the possible of provide much senior advantage and modified services to person user at the correct place on the correct time.

We, therefore, advocate new methods for multifaceted activity mining, incremental preservation, online discovery and practical data management based on user behavior. In exacting, we plan pre-fetching and approaching techniques with cost-sensitive manage to make easy predictive data sharing. Introduction implementation and simulation results exemplify that the future framework and techniques can noticeably amplify local convenience, save carrying out cost, decrease response time, and growth cache use. We have argued and established the value of motion based hands-on data management in mobile environments.

(Minas A. Karaolis, 2010): Coronary heart ailment becomes the major source of disability in humans and in the fit traditional countries it is main basis of death in the well urbanized countries. There is no doubt that chief progress has been made in the finding and action of CHD, but there is requiring of further study. The target of this paper was to make bigger a data-mining scheme for ruling of heart event-related danger factors target in the lessening CHD events. The investigated danger factors were:

1. Factor before the event:
   - Changeable—smoking before the event, hypertension record, and diabetes record.
   - Non unstable—age, gender, and family environment details for impetuous CHD.
2. Factor after the event:
   - Adaptable—smoking after the event, blood force, by and large cholesterol, lipoprotein and glucose.

In misery district of Cyprus, there are total approximately of 528 cases bulk of cases are lies in both proceedings. Data-mining analysis was approved out using the C4.5 choice tree algorithm for the mention two events using diverse splitting criterion. By applying classification rules in C4.5, the extracted main important risk factors were: 1) for Gender, age, smoking. Many other researchers also extract these risk factors.

(Xindong Wu, 2008): Mostly Data mining activities extract data from noisy source of information which leads to inaccuracy and imprecise results. Data cleansing method is used to improve the consistency o data. This paper states that statistical error information aids the error-aware system to increase accuracy of results. It is supposed that noisy information is already available to influence it into data mining process.

(Qiang Yang, 2007): The size of data mining algorithms and gear stop at bare client models, producing supply knowledge on client profiles. These methods are used to pointing out who are more attractors and who are faithful, these techniques comprise expert to post procedure the discovered knowledge. Though many post dispensation technique have been limited generating dream, but they did not counsel actions that would lead helpful. In this paper, we bring in novel algorithms that advise actions to alter consumers from an undesired rank to a preferred. The come near integrates data mining and conclusion making neatly by formulate the result making problems openly on top of the data mining fallout in a post giving out step.

Moreover we pioneer band of decision trees which is given away to be more strong when the training data changes to improve the efficiency. Practical tests are conduct on both a practical insurance application domain and UCI yardstick data. EXTENSIVE investigate in data mining has been finished to find out find out knowledge about information. Models such as Bayesian models, decision trees, and association rules have been practical to various manufacturing application such as customer relationship management. While these techniques are very important to move the data mining result to the ultimate applications, they even so involve a huge deal of human physical work by expert.

(Ruoming Jin, 2005): By Large volume of datasets in real world applications like banking sector, Shopping Malls, it is important to parallelize the data mining activities. By applying parallelism, data mining tasks are performed concurrently to speed up the process. Shared memory Machines are available with more than one processor. Shared Memory machines have ability to scalable. Full replication, full locking and optimized full locking are techniques for parallelism of data mining activities. These techniques are applicable to almost all the data mining techniques instead of any specific algorithm as in case of past parallelization. Parallelization techniques basically applied on Artificial Neural network, Bayesian Network and decision tree algorithm etc.

(Gosta Grahne, 2005): The FP-tree is a compressed form of all repeated related information in a dataset. Every Branch of this tree shows the repeated itemset and nodes alongside the twigs are stored in descending order. Leaves nodes represents smallest amount repeated items. This frequent pattern tree is used by FP growth method which is based on Depth First Search technique. The latest novel FP array method eliminates the requirement of traversing FP-trees and hence increases performance of FP-trees algorithm. This method is most suited in case of sparse data sets. This method is fastest method but problem is that in sparse data set it takes more memory but not in the case dense data sets.
(Dan Braha, 2003)**:** This research paper axis on winning wide applications of decision tree up to 1054 minutes of building lots taken from a lithographic procedure by using 45 steps. Composite relations belongings among manufacturing gear that lead to augmented product inconsistency have been detected. The retrieve knowledge has been habitual by the succession engineers, and assists to enhance the lithographic process. The paper tells decision tree is very rich in case of data is in viewed in many sizes. It also suggests that by using data mining online monitoring of built-up process may be helpful. New Technology semiconductor constructing, (e.g. IC) is very complex, may take of least 60 steps. A revise into the practice of important fabs has exposed that

They gather vast data from databases, generated from a variety of sensors track manufacture data, and information correlated to major actions of give way losses.

- The developed catalog of end yields is included with parametric ability taken at the very last length of the existing line for end-of-line yield analysis.
- Statistical method is adopted as a reserve of visioning developed problems and civilizing the series performance.

Knowledge discovery in database (KDD) is the need for understanding characteristic effects and extracting significant information from volumetric raw data. The short form KDD is useful for transform low-level data into high –level information. Data mining is viewed as a demanding step in an overall process that comprises the request of specific algorithms for extracting patterns.

(Neal Lesh, 2000): This paper concludes that without proper knowledge about the attribute, Several Data sets includes irrelevant attributes. Due to Irrelevancy in datasets, there is degradation in accuracy of Clustering and Classification techniques. If we select right and relevant attribute, then accuracy must be increased as well as response time of the particular algorithm also minimized. FeatureMine is a scalable feature mining approach that joins Sequence mining and classification technique. This proposed algorithm has ability to handle large volumes of records efficiently.

(Venkatesh Ganti, 1999): This Conference Paper differentiate among clustering, Regression and Classification. Clustering is used to arrange similar objects in one group. There are two types of attributes in Training Dataset. First one is Dependent Attribute and second is Predictor Attribute. The Aim is to develop a model that accepts input with Predictor Attribute and output with Dependent Attribute. If the value of Dependent Attribute is Numeric then it is called Regression Technique, otherwise it is known as classification technique. In Classification Technique, Dependent Attribute is used as Class Labels. Researchers proposed many Classification models: genetic algorithms, decision tables and tree structure models also called classification tree. Classification tree is also known as Decision Tree Model.

CHAPTER -3

PRESENT WORK

3.1. Problem Formulation

Decision tree technique is a process of managing the set of rules into such a way that final outcome is obtained from leaf node by following above rules starting from root node. It is the simplest technique of data mining. ID3, CHAID, CART, C 4.5 are all types of Decision Tree. Decision tree is used for classification. In our Research, We used NDTALGO algorithm which is enhanced version of C4.5. NDTALGO is an optimized version of C4.5 algorithm. This NDTALGO is applied to mobile services dataset created by wirelessMon software. NDTALGO has same working as other previous decision trees. Create Tree from root to leaf nodes. Lead node is treating as label.

3.2. Objective

The problems taken for this research work is divided into some objectives which are as follows.

1. Plan to perform complex activity mining to recognize the fluctuating pattern of mobile users.
2. Different activities may exhibit dependencies that affect user behaviors.
3. Full granularity of user activities facilitates high quality, full accuracy services to each user at different places on right time. Propose enhanced methods for complex activity mining.

I am ensuring that Proposed Algorithm will better than existing decision tree algorithms in the following manners:

1. Enhance the efficiency for constructing the decision tree.
2. Reduced Computation Time.
3. Eliminate Error rate.
4. Increased Accuracy.

3.3. Methodology

Data mining techniques are used to classify various services into different classes regarding their characteristics and behaviors. Services such as Bank Services, Medical Facilities, Online Shopping and Agriculture Production are used to assist users. In the Proposed methodology, dataset is obtained by using wirelessMon software is input to the NDTALGO. This algorithm creates tree by performing optimal calculations. Same dataset is also applied to C4.5 for comparison. Tools need for complete our research are WirelessMon, Netbeans and WEKA.

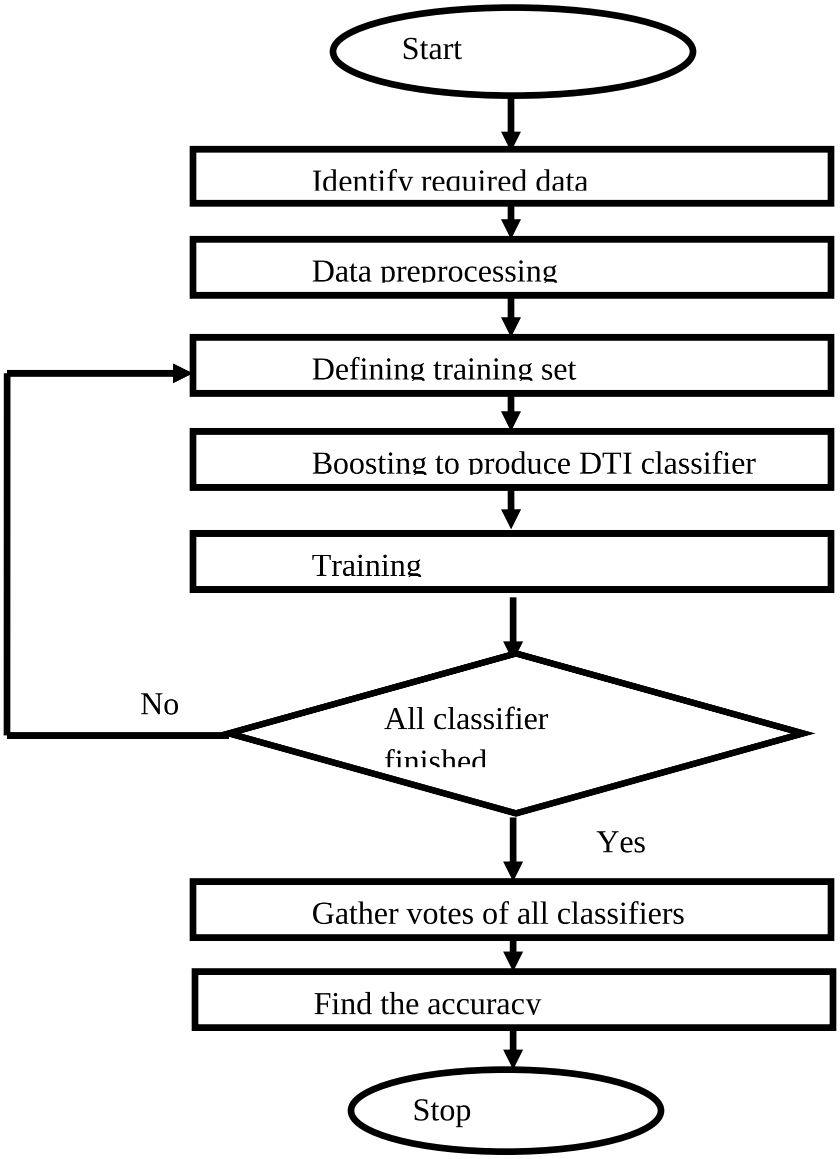


Figure 3.1: Research Methodology

3.4.Sources of Dataset

Data set is a collection of data gathered from different data sources directly and indirectly. Basically there are two source of data collection (Ramamurthy, 2014):

- Primary Source
- Secondary Source

Primary Source: The data are collected by user under direct supervision. It is latest data and being used for first time.

Methods of Primary Source:

- Observations
- Survey
- Questionnaire
- Interview

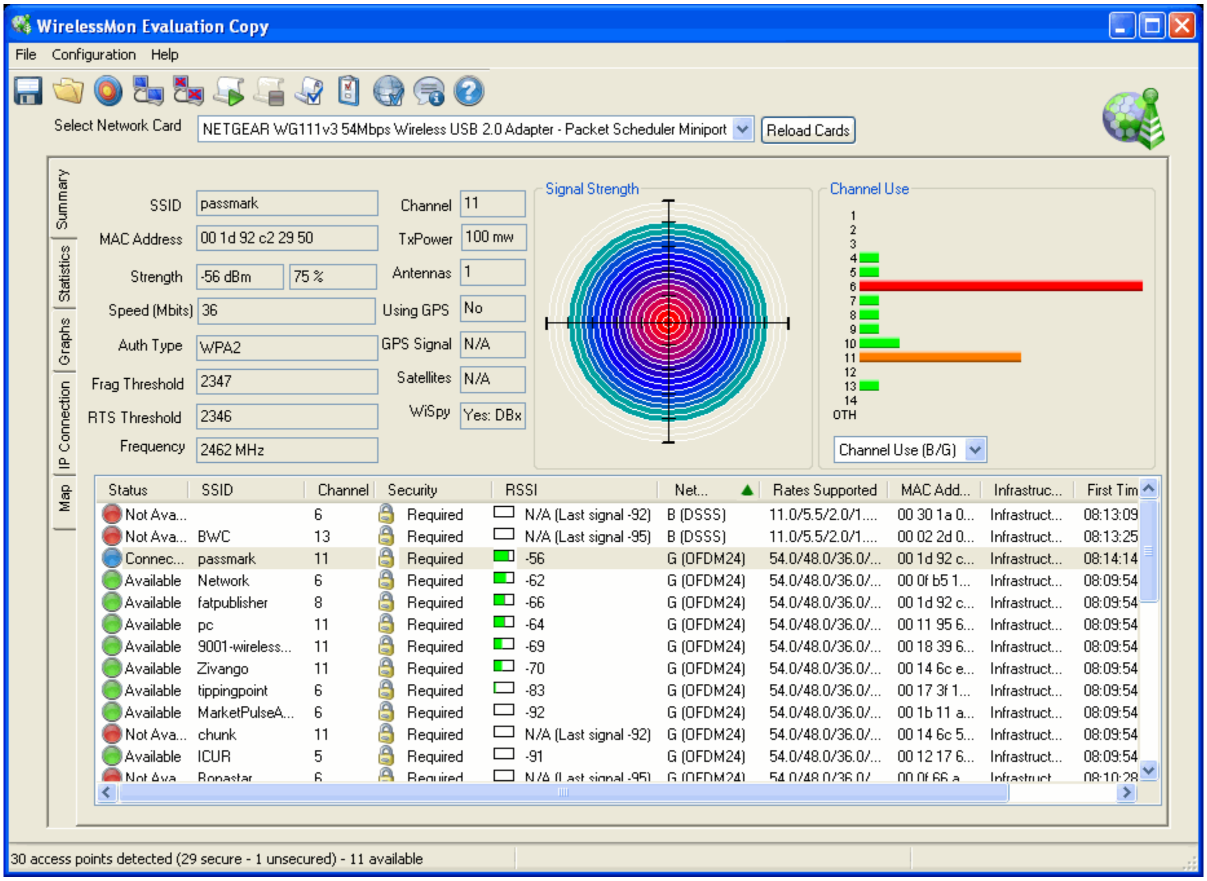
Secondary Source: Data that is collected for other purpose. Secondary data is reused data.

Methods of Secondary Source:

- Sales Record
- Customer Feedback
- Books
- Journals
- Internet.

3.5. WirelessMon

WirelessMon tool allows us to observe the rank of wireless WiFi adapter(s) and collect information regarding local wireless access points and hot spots in real time. WirelessMon can record the information into a file.



Select Network Card: NETGEAR WG111v3 54Mbps Wireless USB 2.0 Adapter - Packet Scheduler Miniport

SSID: passmark
MAC Address: 00 1d 92 c2 29 50
Strength: -56 dBm 75 %
Speed (Mbits): 36
Auth Type: WPA2
Frag Threshold: 2347
RTS Threshold: 2346
Frequency: 2462 MHz

Channel: 11
TxPower: 100 mw
Antennas: 1
Using GPS: No
GPS Signal: N/A
Satellites: N/A
WiSpy: Yes: DBx

Channel Use (B/G)

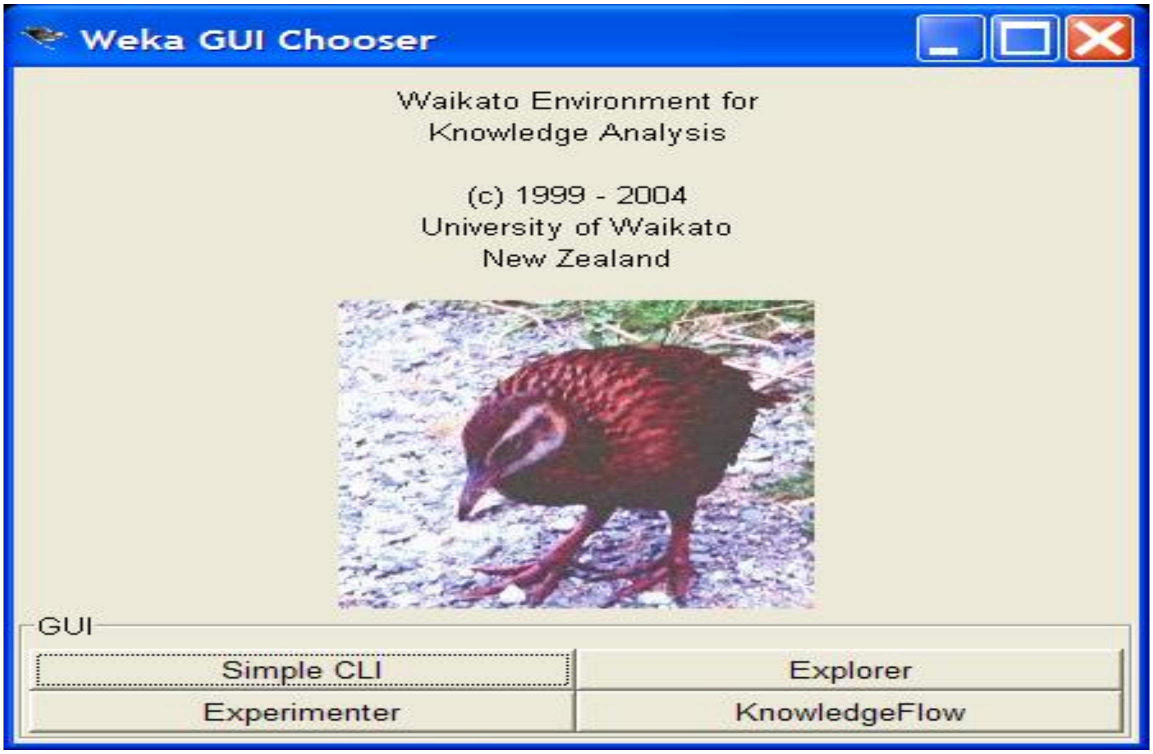
Status	SSID	Channel	Security	RSSI	Net...	Rates Supported	MAC Add...	Infrastruc...	First Tim
Not Ava...		6	Required	N/A (Last signal -92)	B (DSSS)	11.0/5.5/2.0/1....	00 30 1a 0...	Infrastruct...	08:13:09
Not Ava...	BWC	13	Required	N/A (Last signal -95)	B (DSSS)	11.0/5.5/2.0/1....	00 02 2d 0...	Infrastruct...	08:13:25
Connec...	passmark	11	Required	-56	G (OFDM24)	54.0/48.0/36.0/...	00 1d 92 c...	Infrastruct...	08:14:14
Available	Network	6	Required	-62	G (OFDM24)	54.0/48.0/36.0/...	00 0f b5 1...	Infrastruct...	08:09:54
Available	fatpublisher	8	Required	-66	G (OFDM24)	54.0/48.0/36.0/...	00 1d 92 c...	Infrastruct...	08:09:54
Available	pc	11	Required	-64	G (OFDM24)	54.0/48.0/36.0/...	00 11 95 6...	Infrastruct...	08:09:54
Available	9001-wireless...	11	Required	-69	G (OFDM24)	54.0/48.0/36.0/...	00 18 39 6...	Infrastruct...	08:09:54
Available	Zivango	11	Required	-70	G (OFDM24)	54.0/48.0/36.0/...	00 14 6c e...	Infrastruct...	08:09:54
Available	tippingpoint	6	Required	-83	G (OFDM24)	54.0/48.0/36.0/...	00 17 3f 1...	Infrastruct...	08:09:54
Available	MarketPulseA...	6	Required	-92	G (OFDM24)	54.0/48.0/36.0/...	00 1b 11 a...	Infrastruct...	08:09:54
Not Ava...	chunk	11	Required	N/A (Last signal -92)	G (OFDM24)	54.0/48.0/36.0/...	00 14 6c 5...	Infrastruct...	08:09:54
Available	ICUR	5	Required	-91	G (OFDM24)	54.0/48.0/36.0/...	00 12 17 6...	Infrastruct...	08:09:54
Not Ava...	Ronastar	6	Required	N/A (Last signal -95)	G (OFDM24)	54.0/48.0/36.0/...	00 0f 66 a...	Infrastruct...	08:10:28

30 access points detected (29 secure - 1 unsecured) - 11 available

Figure 3.2: Interface of wirelessMon software tool.

3.6. WEKA Tool

I will little discuss about WEKA tool because this tool will be helpful throughout my research. WEKA is data mining software tool developed by the University of Waikato in New Zealand. It consists of group of algorithms for data mining. The algorithms are functional directly to a data. WEKA executes different algorithms for data preprocessing, classification of data, regression, clustering, association rules; it also facilitates visualization tools (Svetlana S. Aksenova, 2004). Its name is derived from WEKA bird which is only living in New Zealand.



Waikato Environment for
Knowledge Analysis

(c) 1999 - 2004
University of Waikato
New Zealand

Figure 3.3: Starting Screen of WEKA TOOL

3.6.1. Windows of WEKA TOOL

Four windows are available in WEKA Tool:

- **Simple CLI** facilitates simple command line.
- **Explorer** provides a way to explore the data.
- **Experimenter** performs experiments and conducting statistical
- **Knowledge Flow** Perform machine learning algorithm.

3.6.2. Features of WEKA TOOL

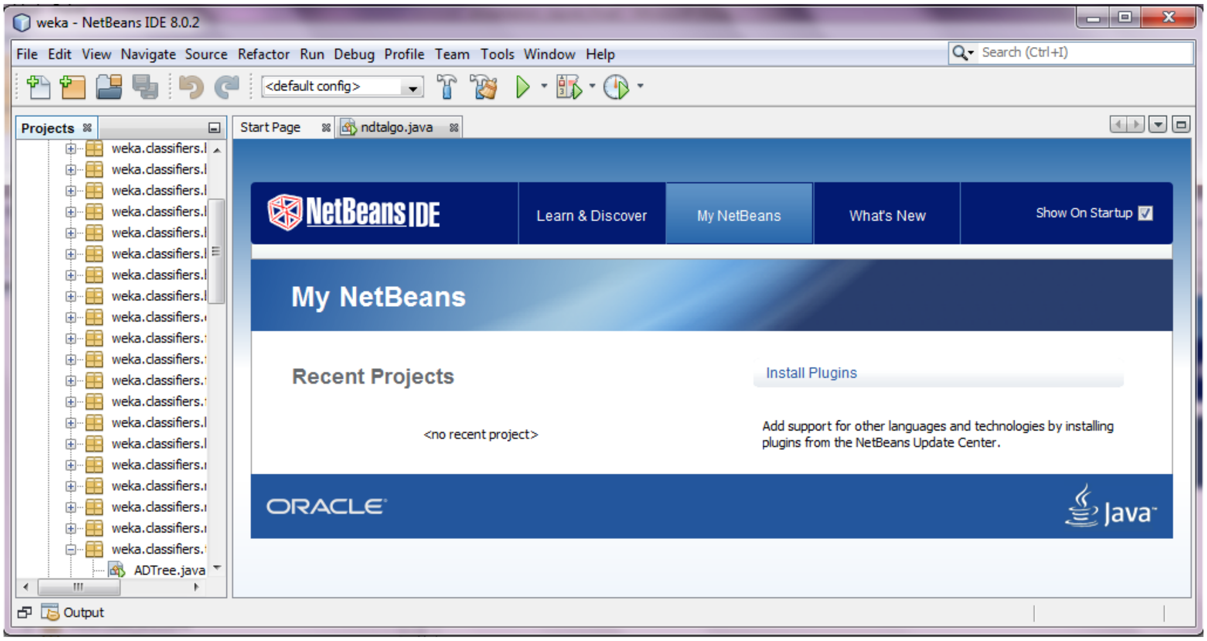
We choose WEKA Tool because of its several features in single tool. They are:

- More than 40 data preprocessing tool are supported in it.
- It supports classification and regression algorithms.
- Clustering algorithms are also working in WEKA.
- Association rules are implemented.
- WEKA supports CSV, ARFF and SQL file formats.
- 3D visualization of objects.
- Single attribute is also visualized.

- Class labels are represented in color coding form.
- Magnify Option.
- It is open source software.

3.7. Netbeans

Netbeans is a faster way of developing web applications. It is open source software. It is very easy to use. It provides smarter way to build software. It is intelligent software. It contains many inbuilt libraries for developing modules. Netbeans is free of cost as well as easily available.



Add support for other languages and technologies by installing plugins from the NetBeans Update Center.

Figure 3.4: Interface of Netbeans

3.8. Algorithm of Proposed Technique

```
(1) Choose base node N;
(2) if all the samples belong to one class CLS then
(3) N is labeled with CLS and returned as a leaf node;
(4) if list of attribute is vacant then
(5) More frequent class is attached to returned node N;
(6) Calculate maximum information gain and select test-attribute from it;
(7) label node N with test-attribute;
(8) for each value $v_i$ of test-attribute;
(9) for test-attribute = $v_{i,}$ build a branch from node N;
```

```
(10)  let $s_i$ be the group of sample from samples for which test-attribute = vi;
(11)  if $s_i$ is blank then
(12)  most frequent class is attached to leaf node;
```

Information Gain is calculated from Entropy. Entropy is amount of impurity in dataset. So to calculate information gain, we have to first calculate entropy.

$$Entropy(\mathrm{s}) = -\mathrm{p}(+\mathrm{ve}) \log_2 \mathrm{p}(+\mathrm{ve}) - \mathrm{p}(-\mathrm{ve}) \log_2 \mathrm{p}(-\mathrm{ve})$$

P(+ve) denotes positive proportions from sample S, whereas P(-ve) denotes negative proportions from Sample S.

Assume that Sample S contains 14 items. Nine Items are positive and five items are negative in sample. We denote this statement as:

Entropy([9+,5-])= -(9/14) $\log_2$(9/14) -(5/14) $\log_2$(5/14)

= 0.940

Remember that if all the items are related to one class then entropy will always equal to Zero. Consider scenario in which all the items are belong +ve class, then number of items belong to –ve class is equal to zero, then entropy will also equal to zero. If both classes' labels have equal number of items, then entropy is equal to one. Entropy will ranging between 0 and 1 if number items has unequal class labels.

We can write Entropy in compact form as:

$$\mathrm{Entropy}(S) = \sum_{i=1}^{\mathrm{c}} -\mathrm{Pi}\ \log_2 \mathrm{Pi}$$

Where Pi is the proportion of number of items in a sample S. This Sample S belong to Class i. The base of Logarithm is 2 because entropy is used in terms of bits. The value of entropy is depend on values of c. Greater the entropy as greater the $\log_2$c.

From Entropy, we can calculate Information gain of an attribute. Basically IG is a quantity of decline in entropy. We can define Information Gain as

$$\mathrm{IG}(S, A) = \mathrm{Entropy}\ (S) - \sum_{v \in Values(A)} \frac{|\mathrm{Sv}|}{|\mathrm{S}|}\ \mathrm{Entropy}\ (Sv)$$

Where Sv is the subset of given Samples. Possible values of attribute A if denoted by Values(A).

Suppose S is Sample contaning 14 attributes, one of these 14 attribute is wind which have two values Weak and Strong.

Positive values for wind = weak is 6.

Negative values for wind = weak is 2.

Positive value for wind= strong is 3.

Negative value for wind=weak is 3.

Information Gain of Wind attribute may be calculated as

S ←[9+, 5-]

S_{weak}←[6+, 2-]

S_{strong}←[3+, 3-]

IG(S,Wind)=Entropy (S)-(8/14) Entropy(S_{weak})-(6/14) Entropy (S_{strong})

=0.940-(8/14)0.811-(6/14)1.00

IG(S, Wind) = 0.048

Similarly, we have one more attribute named Humidity which have two possible values are high and normal.

Calculate IG(S, Humidity) as

Entropy (Humidity_{high})=0.985

Entropty (Humidity_{normal})=0.592

IG(S,Humidity)=0.940 – (7/14)0.985 – (7/14)0.592

IG(S, Humidity) = 0.151

Compare IG of both Wind and Humidity.

IG(S, Humidity)> IG(S, Wind)

Select Humidity attribute for further growing tree.

Consider an example in which PalyTennis attribute is target attribute which has two possible values are Yes and No. We have four attributes(Outlook, Wind, Temperature, Humidity) to decide whether PlayTennis attribute has labeled with Yes or No. Assume IG of four given attribute are as

Table 3.1: List of Information Gain

Name of attribute	**Information Gain**
Outlook	0.246
Wind	0.048
Temperature	0.029
Humidity	0.151

From the above table, it is shown that attribute Outlook has greatest information gain as compared to other attributes. Therefore, outlook attribute is selected as root node. Its branches are created by its possible values Sunny, Rainy and overcast(Mingers, 1989).

CHAPTER -4

RESULTS AND DISCUSSIONS

In this Chapter, we have presented various outcomes by applying our proposed algorithm. The resultant outcomes are represented by snapshot. These outcomes are compared with existing approaches and mentioned our proposed algorithm has more efficiency and accuracy than previous ones.

4.1. Data Set

Data set is generated by using wirelessMon software. This software has facility to generate data file in various file format such as CSV, ARFF and binary file. We create CSV file from wirelessMon software.

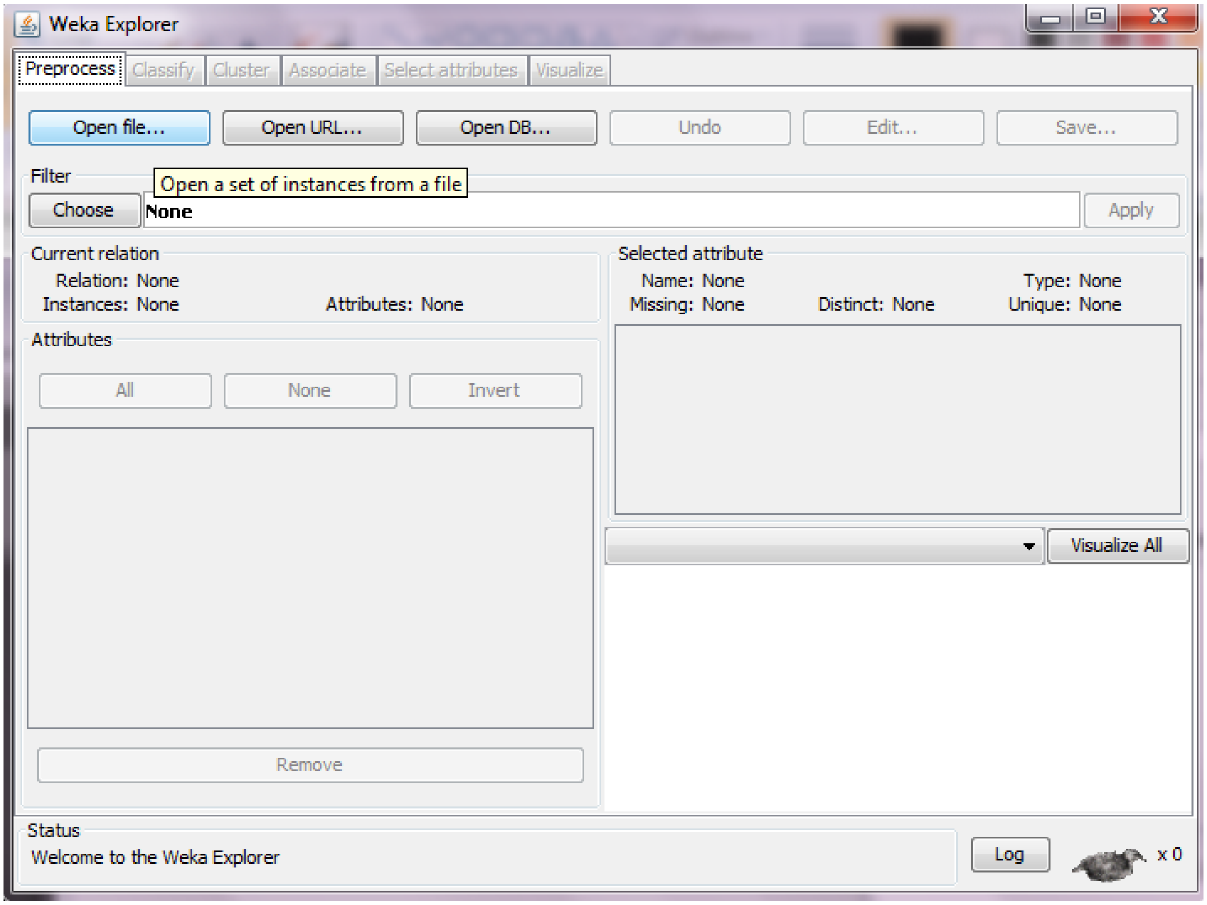
Table 4.1: List of attributes in data set

Strength	97.2	95.7	95.1	104.3	110	95.9	101.2
Authentication Type	173.4	158.7	162.4	188.8	189	190.9	157.3
Threshold	65.2	63.6	63.8	67.2	70.3	66.3	65.4
Frequency	54.7	54.5	53.3	56.2	45	56	55.5
Antennas	2324	2015	2008	2710	2024	2023	3031
GPS Signals	120	122	97	90	164	141	136
Satellites	3.33	3.05	3.15	3.58	3.6	3.62	2.97
Transmitted Frame	3.47	3.03	3.29	3.644	3.67	3.47	3.33
Multicast Frame	8.5	9	9.4	7.5	21.5	9	4.5
ACK Failure Count	97	62	69	162	123	145	160
Services	Airport	Store	Movie	Station	Restaurant	Movie	Station

Services attribute is our target attribute. This target attribute has some possible values are Strength, Authentication Type, Threshold, Frequency, Antennas, GPS Signal, Satellites, Transmitted Frame, Multicast Frame, ACK Failure Count and Services.

4.2. WEKA TOOL

I will little discuss about WEKA tool because this tool will be helpful throughout my research. WEKA is data mining software tool developed by the University of Waikato in New Zealand. It consists of group of algorithms for data mining. The algorithms are functional directly to a data.



Open a set of instances from a file

Filter: None

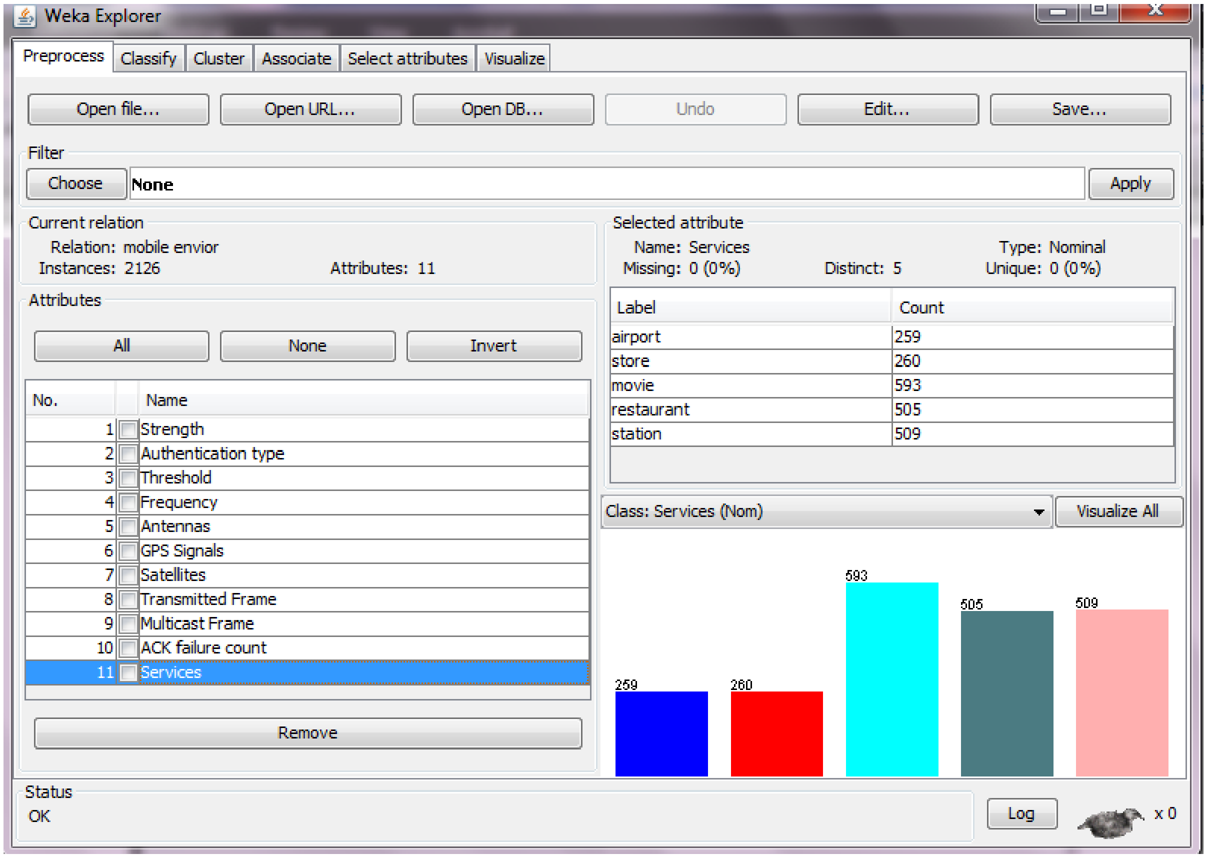
Relation: None
Instances: None
Attributes: None

Name: None
Type: None
Missing: None
Distinct: None
Unique: None

Welcome to the Weka Explorer

Figure 4.1: Select dataset file to open

In figure 4.1, an interface of WEKA Tool that provides a way to open dataset file by clicking on Open file button. After clicked on Open file button, it will ask you to browse your dataset file, select the desired file. In this case, I select CSV format file. BY default no file is selected. Open file option is shown by selecting the preprocessing tab from the main menu.



Filter: None

Current relation

Relation: mobile envior
Instances: 2126 Attributes: 11

Selected attribute

Name: Services Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

Label	Count
airport	259
store	260
movie	593
restaurant	505
station	509

Attributes

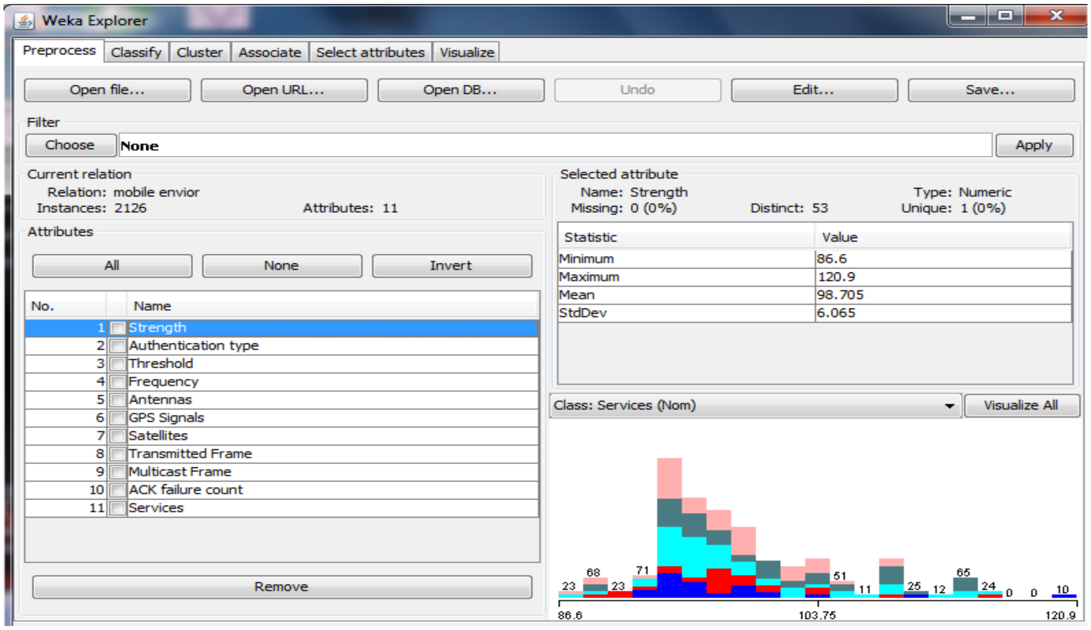
No.		Name
1	□	Strength
2	□	Authentication type
3	□	Threshold
4	□	Frequency
5	□	Antennas
6	□	GPS Signals
7	□	Satellites
8	□	Transmitted Frame
9	□	Multicast Frame
10	□	ACK failure count
11	□	Services

Class: Services (Nom)

Status
OK

Figure 4.2: Display the graph which shows all instances of target attribute services

By opening dataset file, the entire attribute available in the file are shown left pane of the WEKA Tool and numbered sequentially. The values of Services attributes are denoted by color coding. Five different values are denoted by different colors.



Filter: None

Current relation
Relation: mobile envior
Instances: 2126
Attributes: 11

Selected attribute
Name: Strength
Missing: 0 (0%)
Distinct: 53
Type: Numeric
Unique: 1 (0%)

Statistic	Value
Minimum	86.6
Maximum	120.9
Mean	98.705
StdDev	6.065

Attributes

No.		Name
1	□	Strength
2	□	Authentication type
3	□	Threshold
4	□	Frequency
5	□	Antennas
6	□	GPS Signals
7	□	Satellites
8	□	Transmitted Frame
9	□	Multicast Frame
10	□	ACK failure count
11	□	Services

Class: Services (Nom)

Figure 4.3: shows the instances of attribute Strength based on class attribute services

Strength attribute is used as minimum value is 86.6 and maximum value is 120.9.

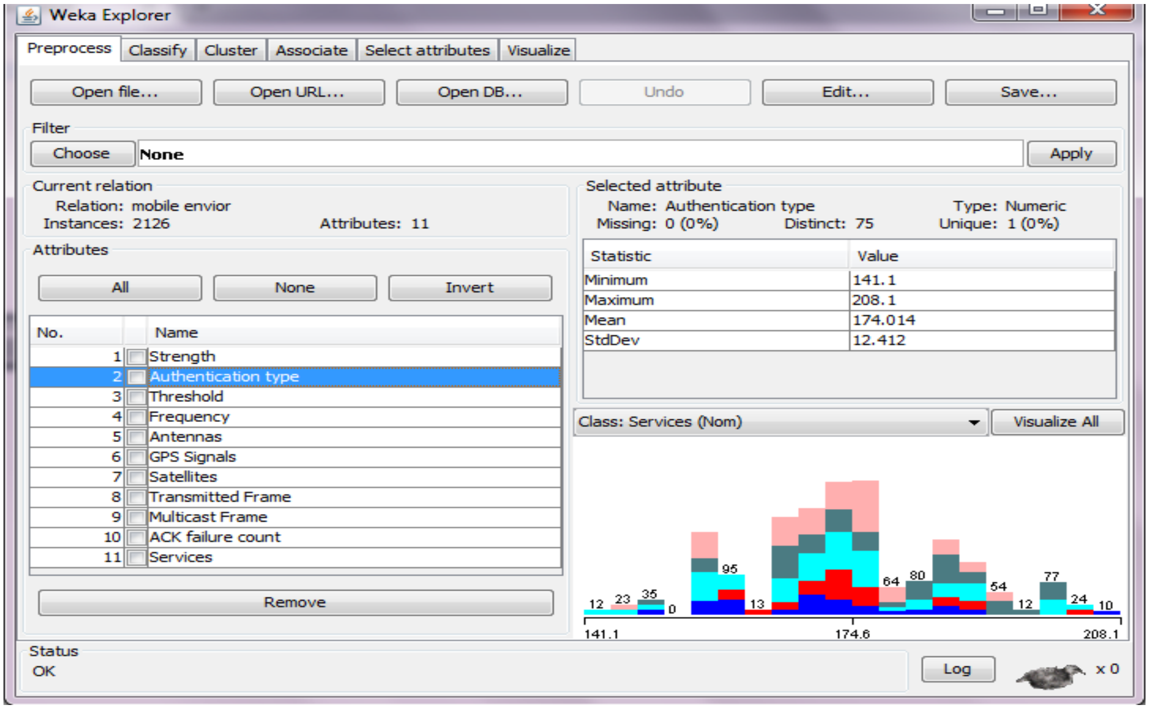
Mean Value is used to show average value and is calculated as

$$Mean = \frac{SumofObservationvalues}{TotalnumberofObservations}$$

Standard Deviation is used to measure how much data is scatter from its mean. Standard Deviation is derived from Greek letter sigma.

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})}\quad ^{2}$$

Where N is total number of items, i denote for from starting value 1 to N. x_i denote particular value of N.



Filter: None

Current relation
Relation: mobile envior
Instances: 2126 Attributes: 11

Selected attribute
Name: Authentication type Type: Numeric
Missing: 0 (0%) Distinct: 75 Unique: 1 (0%)

Statistic	Value
Minimum	141.1
Maximum	208.1
Mean	174.014
StdDev	12.412

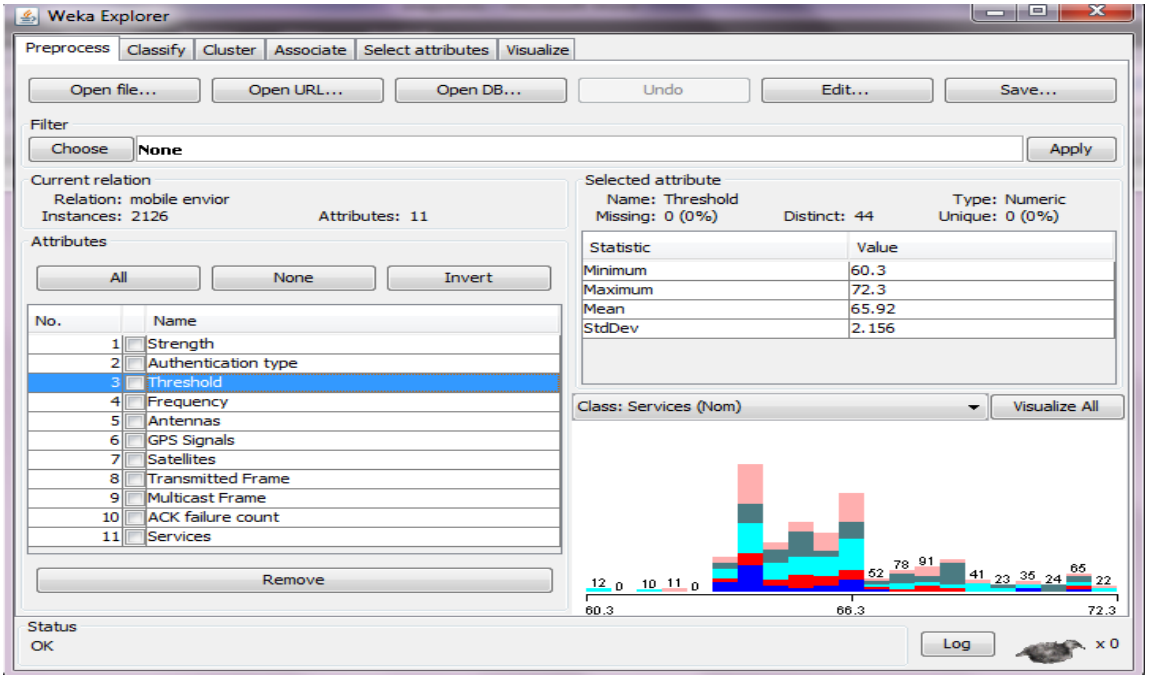
Attributes

No.	Name
1	Strength
2	Authentication type
3	Threshold
4	Frequency
5	Antennas
6	GPS Signals
7	Satellites
8	Transmitted Frame
9	Multicast Frame
10	ACK failure count
11	Services

Class: Services (Nom)

Status
OK

Figure 4.4: shows the instances of Authentication Type based on class attribute services



Filter: None

Current relation
Relation: mobile envior
Instances: 2126 Attributes: 11

Selected attribute
Name: Threshold Type: Numeric
Missing: 0 (0%) Distinct: 44 Unique: 0 (0%)

Statistic	Value
Minimum	60.3
Maximum	72.3
Mean	65.92
StdDev	2.156

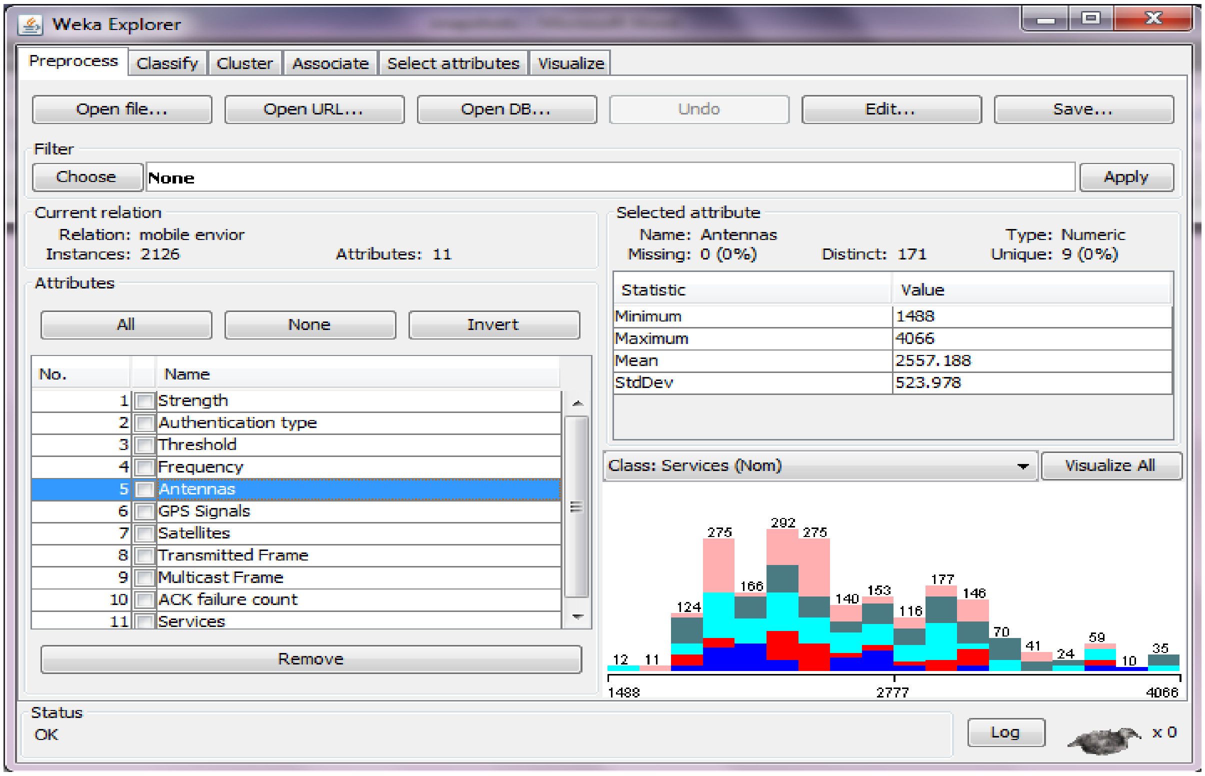
Attributes

No.	Name
1	Strength
2	Authentication type
3	Threshold
4	Frequency
5	Antennas
6	GPS Signals
7	Satellites
8	Transmitted Frame
9	Multicast Frame
10	ACK failure count
11	Services

Class: Services (Nom)

Status
OK

Figure 4.5: shows the instances of Threshold based on class attribute services



Filter

None

Current relation

Relation: mobile envior
Instances: 2126 Attributes: 11

Selected attribute

Name: Antennas Type: Numeric
Missing: 0 (0%) Distinct: 171 Unique: 9 (0%)

Statistic	Value
Minimum	1488
Maximum	4066
Mean	2557.188
StdDev	523.978

Attributes

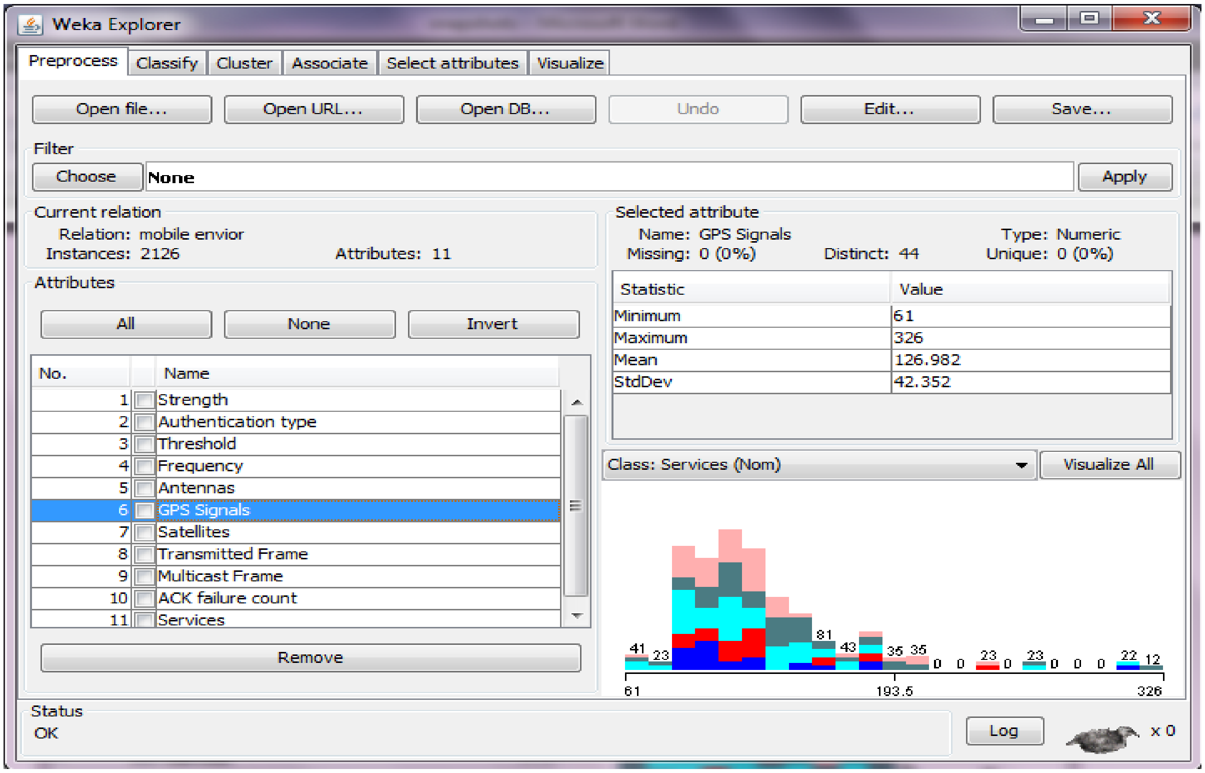
No.		Name
1	□	Strength
2	□	Authentication type
3	□	Threshold
4	□	Frequency
5	□	Antennas
6	□	GPS Signals
7	□	Satellites
8	□	Transmitted Frame
9	□	Multicast Frame
10	□	ACK failure count
11	□	Services

Class: Services (Nom)

Status

OK

Figure 4.6: shows the instances of Antennas based on class attribute services



Filter

None

Current relation

Relation: mobile envior
Instances: 2126 Attributes: 11

Selected attribute

Name: GPS Signals Type: Numeric
Missing: 0 (0%) Distinct: 44 Unique: 0 (0%)

Statistic	Value
Minimum	61
Maximum	326
Mean	126.982
StdDev	42.352

Attributes

No.		Name
1	□	Strength
2	□	Authentication type
3	□	Threshold
4	□	Frequency
5	□	Antennas
6	□	GPS Signals
7	□	Satellites
8	□	Transmitted Frame
9	□	Multicast Frame
10	□	ACK failure count
11	□	Services

Class: Services (Nom)

Status

OK

Figure 4.7: shows the instances of GPS Signal based on class attribute services

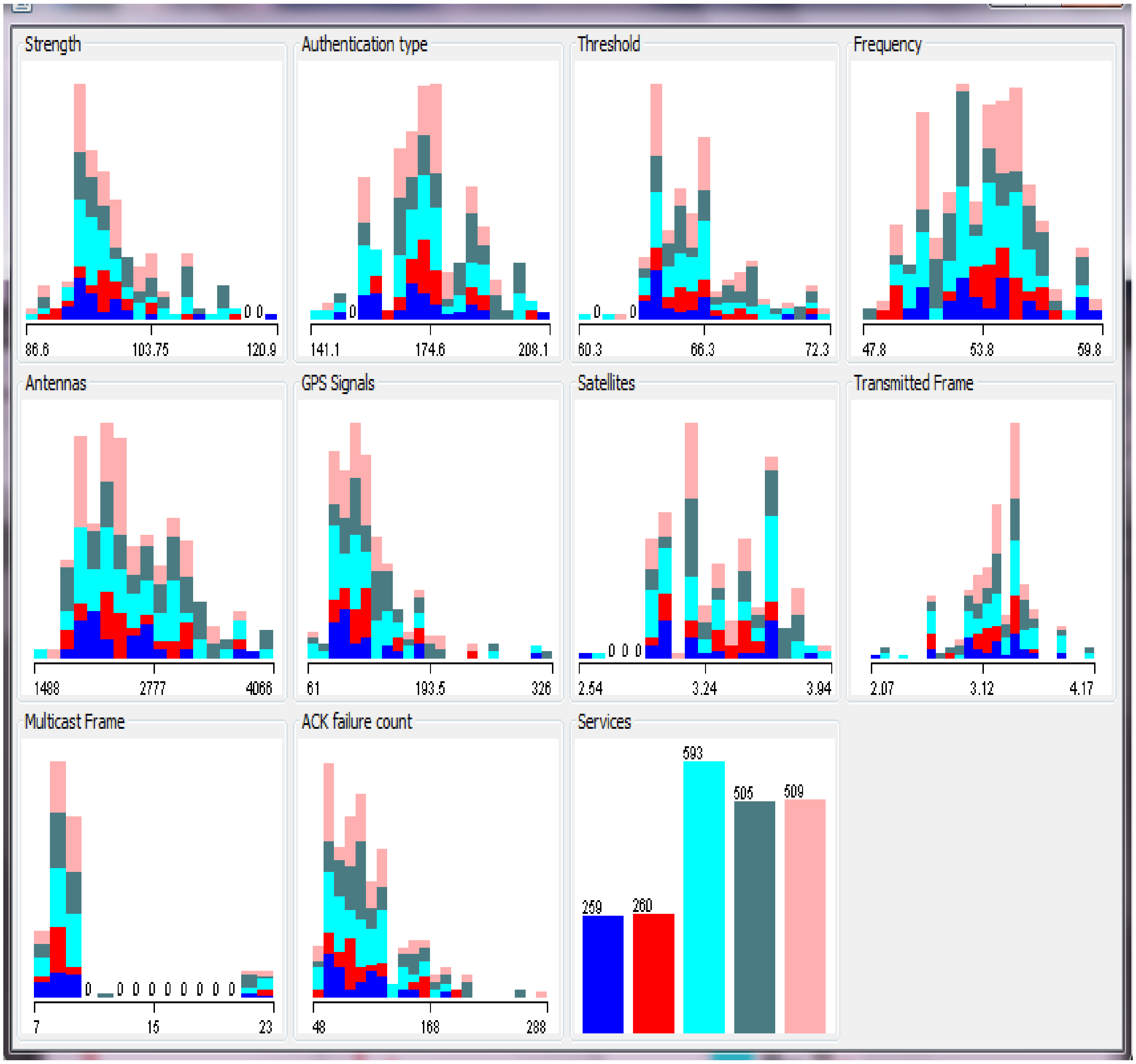
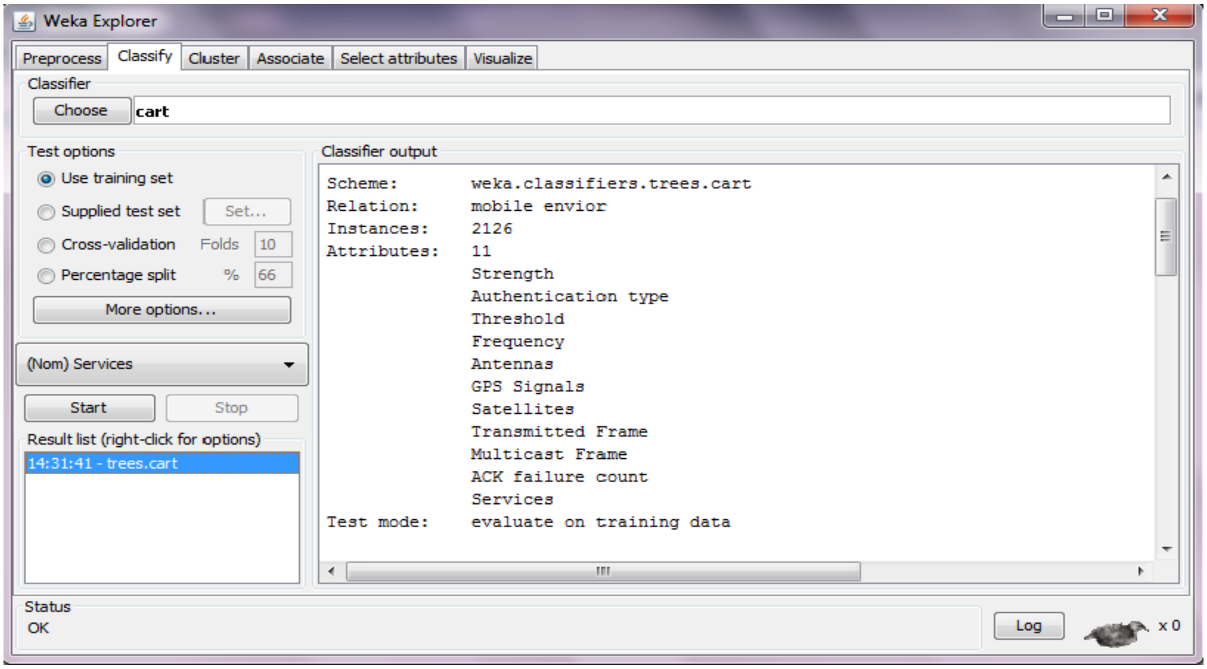


Figure 4.8: shows the graphs of instances of all attributes

Figure 4.8 shows graphs of all the instances of all attributes in a single window. It is a powerful feature of WEKA Tool.

4.3. Classification by WEKA Tool



Classifier: cart

Test options:
● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Services

Result list (right-click for options)
14:31:41 - trees.cart

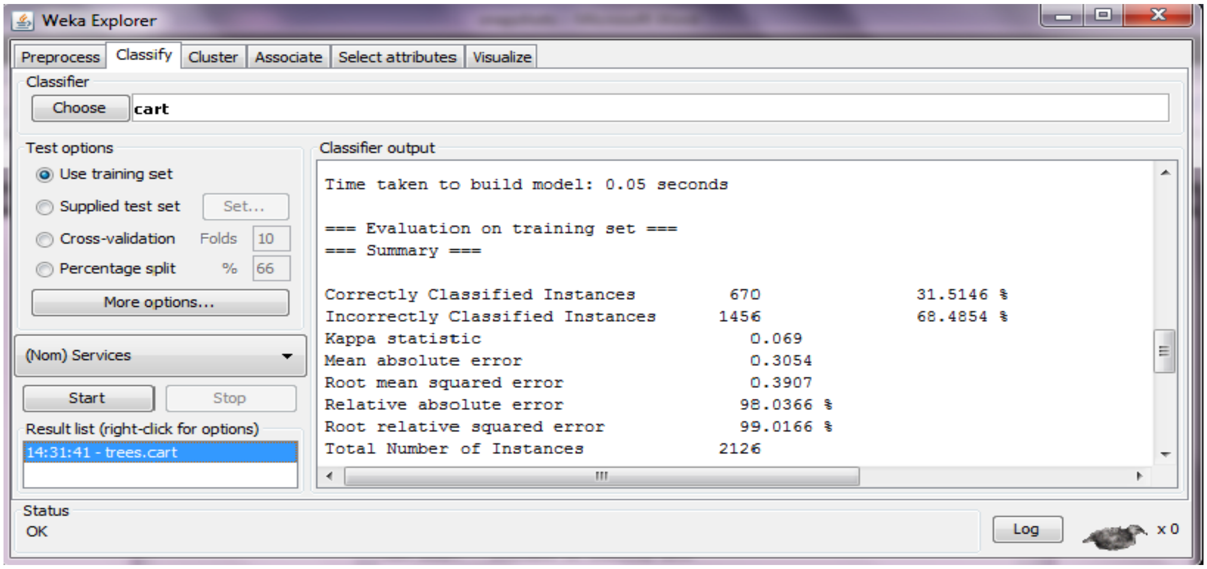
Classifier output

```
Scheme:       weka.classifiers.trees.cart
Relation:     mobile envior
Instances:    2126
Attributes:   11
              Strength
              Authentication type
              Threshold
              Frequency
              Antennas
              GPS Signals
              Satellites
              Transmitted Frame
              Multicast Frame
              ACK failure count
              Services
Test mode:    evaluate on training data
```

Status
OK

Figure 4.9: Display the result of CART algorithm in WEKA Tool

In Above figure, CART (Classification and Regression Tree) algorithm is applied to given dataset. It tests 2126 instances.



Classifier: cart

Test options:
● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Services

Result list (right-click for options)
14:31:41 - trees.cart

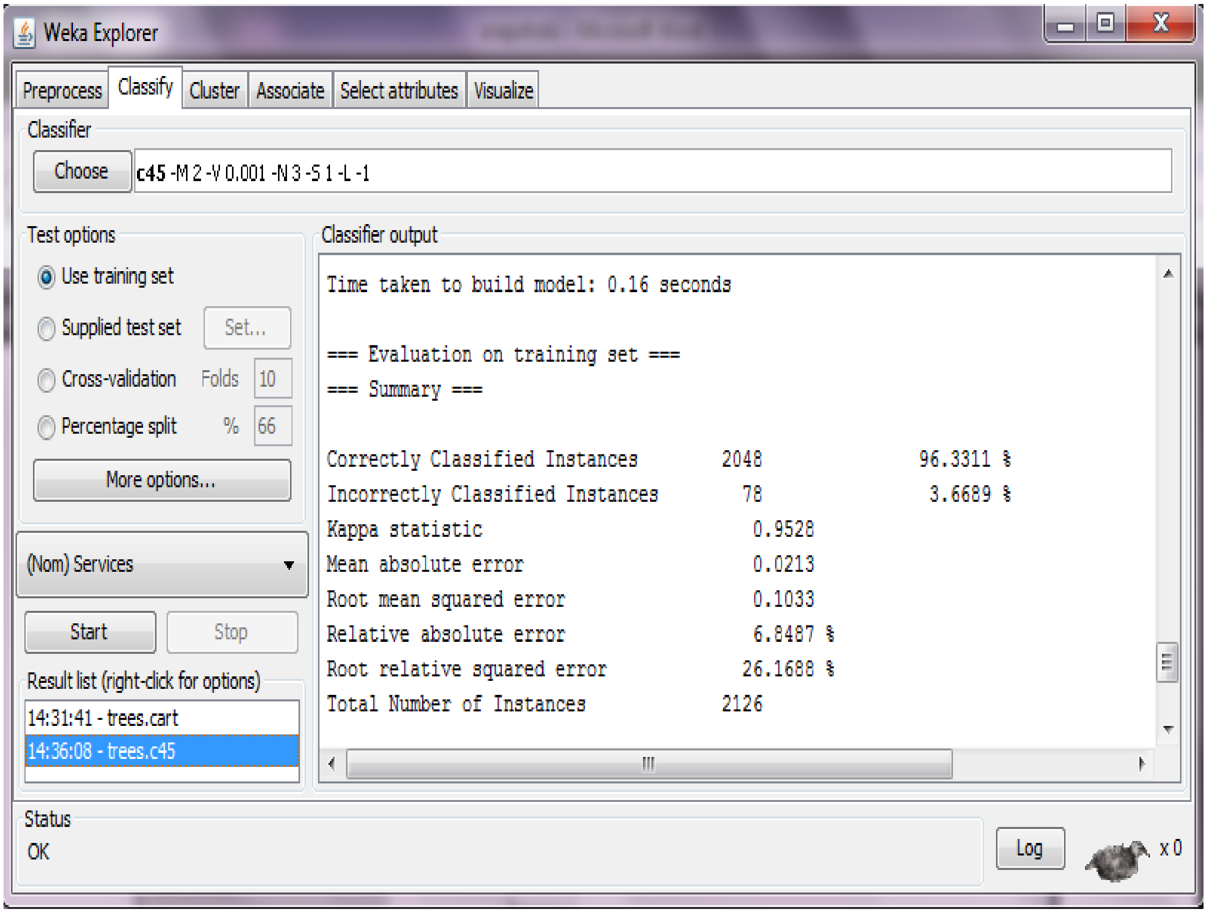
Classifier output

```
Time taken to build model: 0.05 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         670               31.5146 %
Incorrectly Classified Instances      1456               68.4854 %
Kappa statistic                          0.069
Mean absolute error                      0.3054
Root mean squared error                  0.3907
Relative absolute error                 98.0366 %
Root relative squared error             99.0166 %
Total Number of Instances             2126
```

Status
OK

Figure 4.10: shows the correctly classified instances, incorrectly classified instances and error rate using CART Algorithm.

This figure shows that CART algorithm 670 correctly classified instances and 1456 incorrectly classified instances. CART takes time to execute dataset is 0.06 seconds. Above figure shows that only 32 percent of total instances are correctly classified by CART algorithm. Percentage of Incorrectly classified instances is 68. Figure also various types of error rates.



c45 -M 2 -V 0.001 -N 3 -S 1 -L -1

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Services

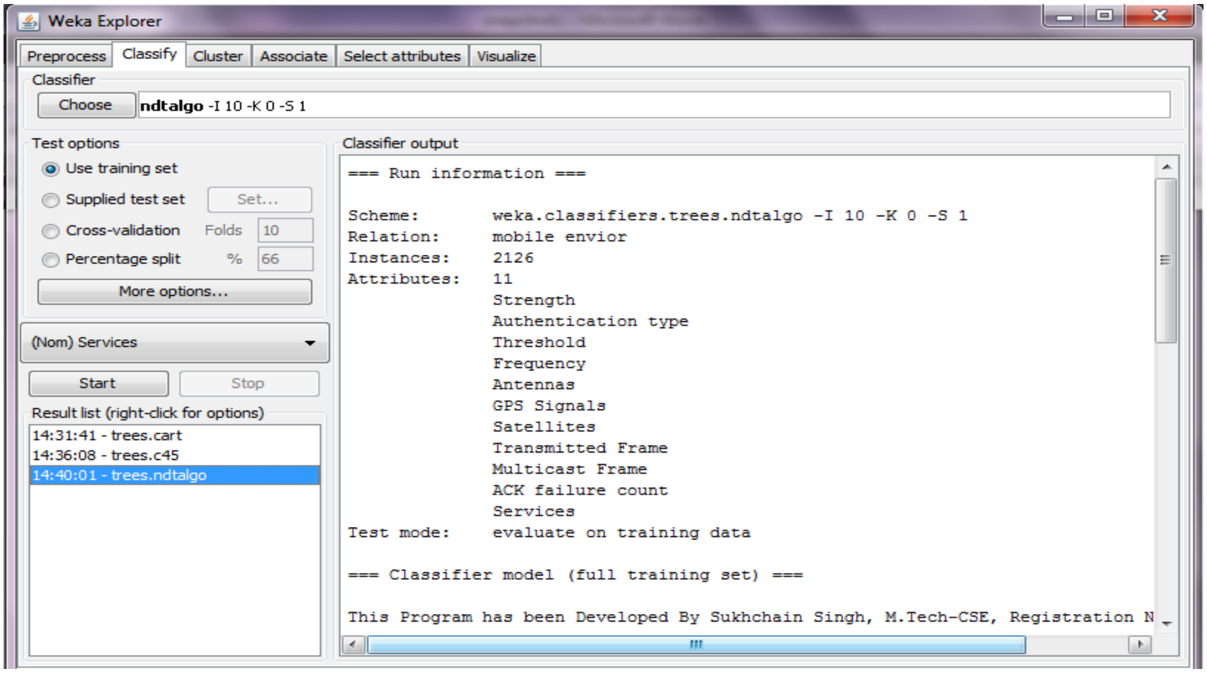
Result list (right-click for options)
14:31:41 - trees.cart
14:36:08 - trees.c45

```
Time taken to build model: 0.16 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        2048               96.3311 %
Incorrectly Classified Instances        78                3.6689 %
Kappa statistic                          0.9528
Mean absolute error                      0.0213
Root mean squared error                  0.1033
Relative absolute error                  6.8487 %
Root relative squared error             26.1688 %
Total Number of Instances             2126
```

Status
OK

Figure 4.11: shows the correctly classified instances, incorrectly classified instances and error rate using C4.5 Algorithm

In Figure 4.11, C4.5 Algorithm is applied to given dataset. C 4.5 algorithm is improved version of CART algorithm. C4.5 algorithm takes 0.16 seconds for computations. The percentage of correctly and incorrectly classified instances is 96 and 4 respectively. C4.5 correctly classified instances are 2048. Incorrectly classified instances by C4.5 are 78. Given figure shows various error rates are less than CART algorithm.



Classifier: ndtalgo -I 10 -K 0 -S 1

Test options: ● Use training set ○ Supplied test set ○ Cross-validation Folds 10 ○ Percentage split % 66

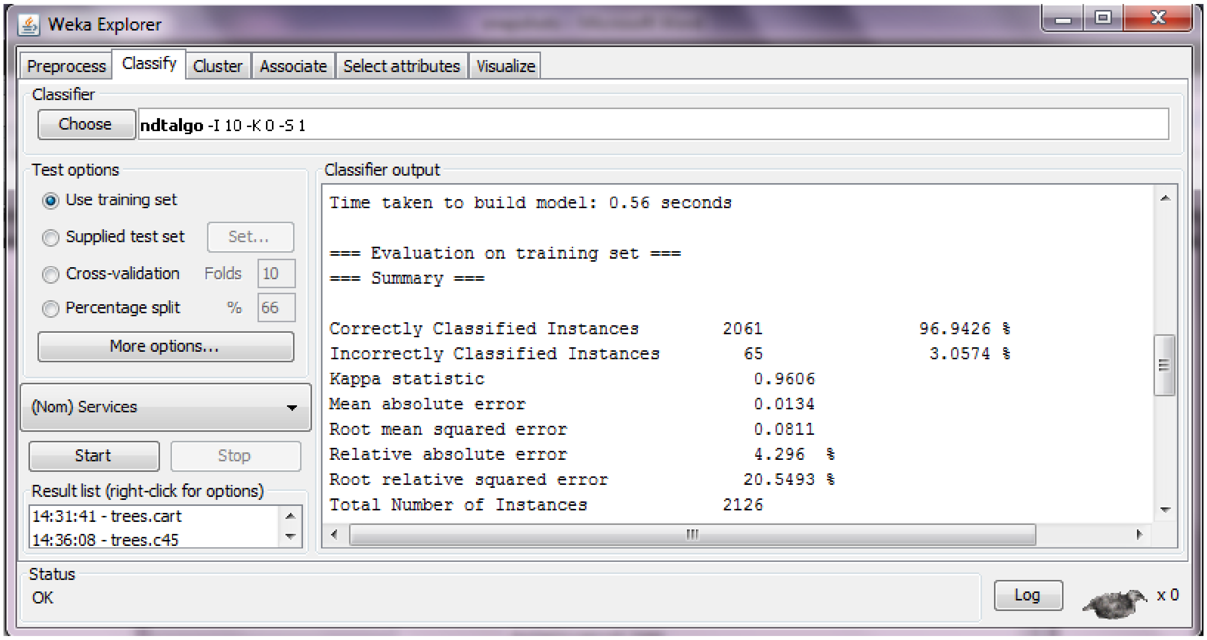
(Nom) Services

Result list:
14:31:41 - trees.cart
14:36:08 - trees.c45
14:40:01 - trees.ndtalgo

```
=== Run information ===

Scheme:       weka.classifiers.trees.ndtalgo -I 10 -K 0 -S 1
Relation:     mobile envior
Instances:    2126
Attributes:   11
              Strength
              Authentication type
              Threshold
              Frequency
              Antennas
              GPS Signals
              Satellites
              Transmitted Frame
              Multicast Frame
              ACK failure count
              Services
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

This Program has been Developed By Sukhchain Singh, M.Tech-CSE, Registration N
```

Figure 4.12: Dataset is tested by Proposed Algorithm (NDTALGO)

Figure 4.12 shows mobile services dataset is tested by proposed algorithm named NDTALGO. It shows run information about the algorithm. 2126 instances are classified by NDTALGO.



Classifier: ndtalgo -I 10 -K 0 -S 1

Test options: ● Use training set ○ Supplied test set ○ Cross-validation Folds 10 ○ Percentage split % 66

(Nom) Services

Result list:
14:31:41 - trees.cart
14:36:08 - trees.c45

```
Time taken to build model: 0.56 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        2061               96.9426 %
Incorrectly Classified Instances        65                3.0574 %
Kappa statistic                          0.9606
Mean absolute error                      0.0134
Root mean squared error                  0.0811
Relative absolute error                  4.296  %
Root relative squared error             20.5493 %
Total Number of Instances             2126
```

Status
OK

Figure 4.13: NDTALGO applied to Mobile services dataset

Proposed algorithm classified instances correctly are 2061. Incorrectly classified instance tested are 65. The percentage of correctly classified instances is 97. The percentage of incorrectly classified instances is 3. The error rate of Root relative squared is 20.5. The error rate of NDTALGO is less than CART and C 4.5 algorithm.

4.4. Comparison among Decision Tree Algorithms

We present in this section, various types of comparison such as correctly classified instances, incorrectly classified instances and error rate among CART, C4.5 and NDTALGO.

Table 4.2: Comparison of correctly and incorrectly classified instances, error rate

Types of comparison	**CART**	**C4.5**	**NDTALGO**
Correctly Classified instances	31.51	96.33	96.94
Incorrectly Classified Instances	68.48	3.66	3.05
Error rate	99.01	26.16	20.54

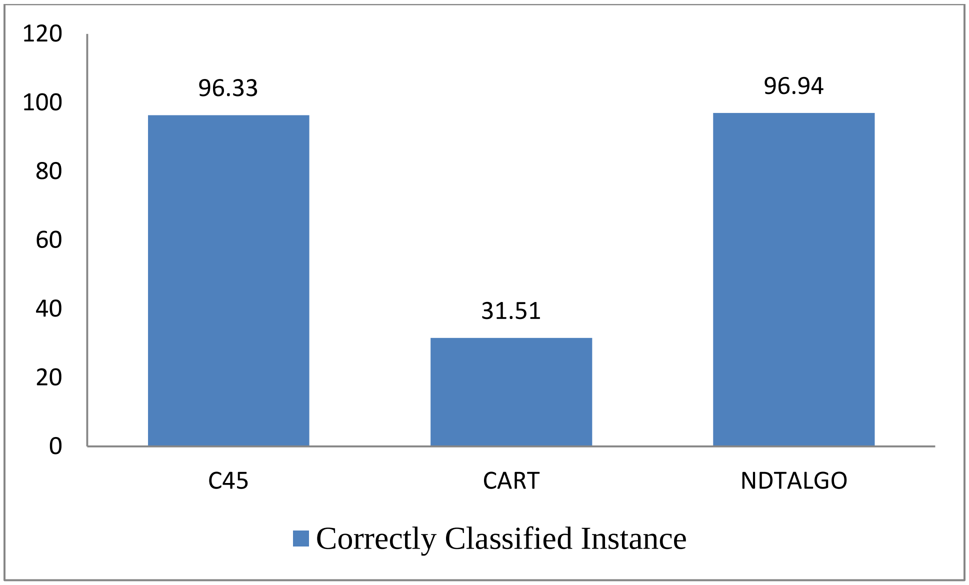


Figure 4.14: Graphical analyses of Correctly Classified Instances

Above Figure visualize the result of correctly classified instances. It shows proposed algorithm has highest percentage of correctly classified instances.

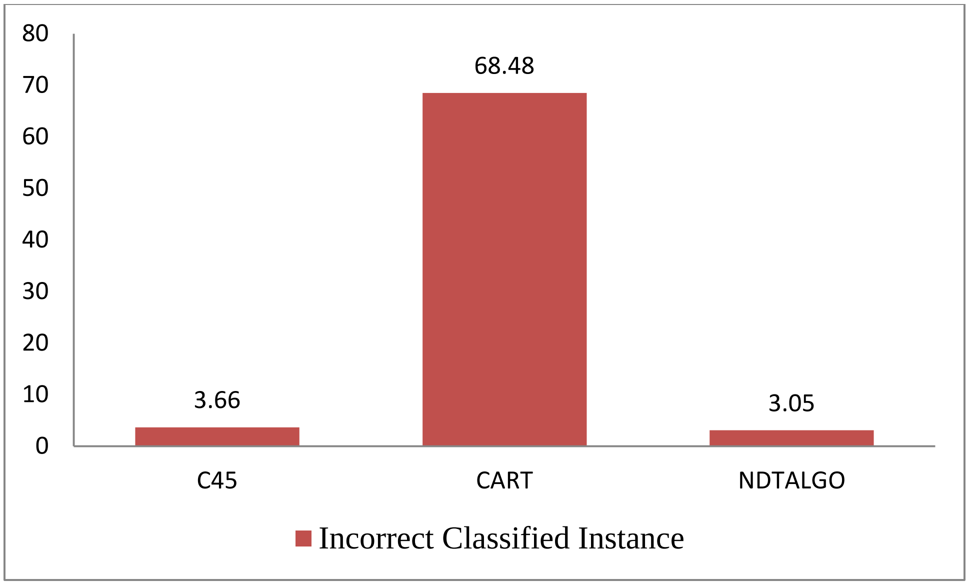


Figure 4.15: Graphical analyses of Incorrectly Classified Instances

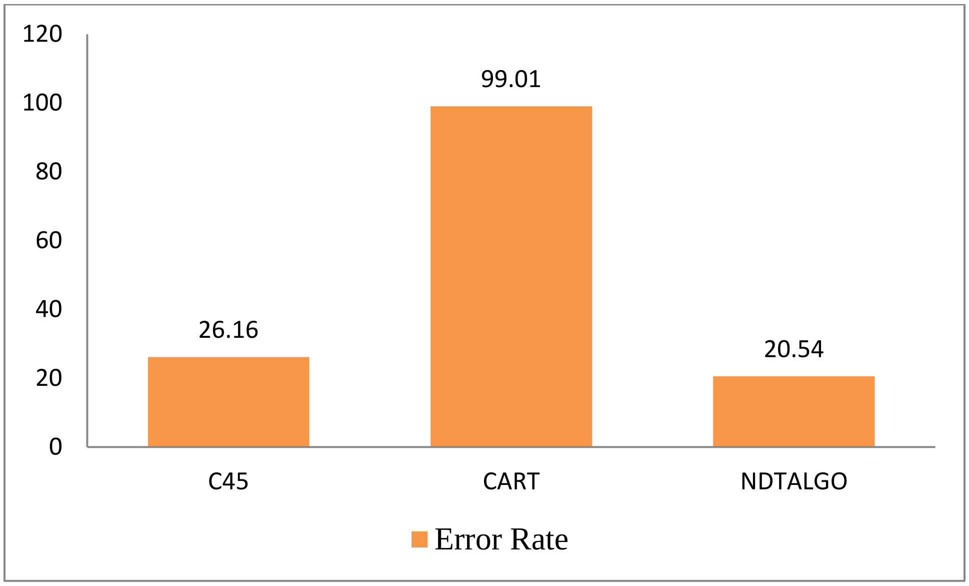


Figure 4.16: Graphical analyses of Error rate among decision tree algorithms

Figure 4.15graphically shows incorrectly classified instances. From comparison, it shows that NDTALGO algorithm has very less incorrectly classified instances.

Figure 4.16 Visualize that proposed algorithm has minimum error rate than CART and C4.5 algorithms.

CHAPTER -5
CONCLUSION AND FUTURE SCOPE

5.1. Conclusion

From the whole discussion, I want to conclude and highlight the approaches for creating a decision tree. CHAID method is used for non-binary tree. After examining ID3 and CART, enhanced algorithm is more preferable to find accuracy and consume less access time to mine data with smallest amount of error rate. Therefore, enhanced algorithm NDTALGO is a optimal algorithm for mining a data on mobile environment data set. NDTALGO is best algorithm than previous algorithm in terms of classified instance and error rate. But one problem of NDTALGO is that it takes more time to computations as compared to CART and C4.5.

In this study, I reviewed the initial data mining task of collecting and cleaning the data prior to the use of the data with the algorithm.

5.2. Future Scope

In future I am thinking to develop complicated software that consists of more data mining techniques capable to operate on large volume of data in order to classify and predict that particular data in the favor of my decision. It is estimated that in coming years data mining will become as famous as facebook, whatsapp etc. It will also very user-friendly. It is estimated that we will be able to work on these advanced techniques to catch out which cinema tickets are best, to track school mate’s contact detail. I will eliminate the computation time of proposed algorithm in future.

References

References to a Book

Jiawei Han, M. K. (2012). *Data Mining Concepts and Techniques.* Morgan Kaufmann.

Ramamurthy, G. (2014). *"Research Methodology".* New Delhi: DreamTech.

References for an Article

Dan Braha, A. S. (2003). ," On the Use of Decision Tree Induction for Discovery of Interactions in a Photolithographic Process",. *IEEE* .

Duong Van Hieu, N. W. (2014). " Analysis of Factors which Impact Facebook User's Attitudes and Behaviours using Decision Tree Techniques ", . *JCSSE* .

Gosta Grahne, M. I. (2005). " Fast Algorithms for Frequent Itemset Mining Using FP- Trees ". *IEEE* .

Ji Dan, Q. J. (2010). " A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree ". *IEEE* .

Kalpesh Adhatrao, A. G. (2013). "Predicting Student's Performance using ID3 and C4.5 Classification Algorithms". International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5,September .

Kimberly Carpenter, P. S. (2014). " Questionnaire Simplification For Fast Risk Analysis Of Children's Mental Health ". *IEEE* .

Lior Rokach, O. M. (2005). Decision Tree. Department Of Industrial Engineering.

Minas A. Karaolis, J. A. (2010). " Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees ",. *IEEE* .

Mingers. (1989). " An empirical comparison of pruning methods for decision-tree induction, Machine Learning ".

Neal Lesh, M. J. (2000). "Scalable Feature Mining for Sequential Data". *IEEE* .

Otero, A. (2005). "Data Mining Techniques: decision trees". Data Mining .

Pui K. Fong, J. H.-j. (2012). " Privacy Preserving Decision Tree Learning Using Unrealized Data Sets",. *IEEE* .

Qiang Yang, S. M. (2007). " Extracting Actionable Knowledge from Decision Trees ". *IEEE* .

Ruoming Jin, G. Y. (2005). "Shared Memory Parallelization of Data Mining Algorithm: Techniques, Programming Interface and Performance". *IEEE* .

Shiow-yang Wu, M. I.-H. (2010). “ Activity-Based Proactive Data Management in Mobile Environments ”,. *IEEE* .

Sivaraman Sriram, X. Y. (2012). " An Enhanced Approach for Classifying Emotions using Customized Decision Tree Algorithm ". *IEEE* .

Soldic-Aleksic, J. (2012). " Combined Approach Of Kohonen Som And Chaid Decision Tree Model To Clustering Problem:A Market Segmentation Example ". *Journal of Economics and Engineering, ISSN: 2078-0346, Vol. 3. No.1,* .

Svetlana S. Aksenova. (2004). “Machine Learning with WEKA”, . California State University, Sacramento California.

Venkatesh Ganti, J. G. (1999). "Mining Very Large Databases",. *IEEE* .

Xindong Wu, X. Z. (2008). " Mining with Noise Knowledge: Error-Aware Data Mining ". *IEEE* .

Xue-Fei Zhang, L. F. (2013). ” A Decision Tree Approach for Traffic Accident Anaysis of Saskatchewan Highways ”, . *IEEE,* .

References to Web page

HTTP://www.thearling.com/text/dmwhite/dmwhite.htm

Appendix

List of Abbreviations

AIDS:- Acquired Immune Deficiency Syndrome

ARFF:- Attribute- Relation File Format

CART:- Classification and Regression Tree

CHAID:- Chi-squared Automatic Interaction Detector

CHD:- Coronary Heart Disease

CSV:-Comma Separated Value

ID:-Iterative Dichotomiser

IG:- Information Gain

KDD:- Knowledge Discovery in Database

MARS:- Multivariate Adaptive Regression Splines

NDTALGO: - New Decision Tree Algorithm

OLAP:- Online Analytical Processing

RDBMS:-Relational Database Management System

SOM:- Self Organizing Map

SQL:- Structured Query Language