Performance Evaluation in Education System using Sprint Decision tree Classification Algorithm

Thesis submitted in partial fulfillment of the requirements
for the award of degree of

Master of Engineering
in
Computer Science and Engineering

Submitted By
Savy Chandna
(801732044)

Under the supervision of:
Dr. Karun Verma
Assistant Professor

Dr. Ravinder Kumar
Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY
PATIALA – 147004

AUGUST 2019

CERTIFICATE

I hereby certify that the work which is being presented in the thesis titled, *"Performance Evaluation in Education System using Sprint Decision tree Classification Algorithm"*, in partial fulfillment of the requirements for the award of degree of Master of Engineering in Computer Science and Engineering submitted in *Computer Science and Engineering Department* of Thapar Institute of Engineering and Technology, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Karun Verma* and *Dr. Ravinder Kumar* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for the award of any other degree of this or any other University.

Savy 15/10/19
Signature:

(Savy Chandna)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Dr. Karun Verma
Assistant Professor
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
Patiala

Ravinder.
Dr. Ravinder Kumar
Assistant Professor
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
Patiala

ACKNOWLEDGEMENT

No volume of words is enough to express my gratitude towards my guides **Dr. Karun Verma, Dr. Ravinder Kumar** Department of Computer Science & Engineering, Thapar Institute of Engineering and Technology, Patiala who has been very concerned and has aided for all the materials essentials for the preparation of this thesis report. They helped me to explore this vast topic in an organized manner and provided me all the ideas on how to work towards a research-oriented venture.

I am also thankful to **Dr. S.S. Bhatia,** Dean of Academic Affairs, **Dr. Maninder Singh,** Head of Computer Science & Engineering Department and **Dr. Ashutosh Mishra,** P.G. Coordinator, for the motivation and inspiration that triggered me for the thesis work.

I would also like to thank the staff members and my colleagues who were always there at the need of hour and provided with all the help and facilities, which I required, for the completion of my thesis work.

Most importantly, I would like to thank my parents and the almighty for showing me the right direction out of the blue, to help me stay calm in the oddest of the times and keep moving even at times when there was no hope.

Savy 16/10/19

Savy Chandna

(801732044)

ABSTRACT

At the present time, the amount of data stored in educational database is increasing rapidly. These databases contain hidden information for improvement of student's performance. Decision tree is the most useful classification algorithm in educational data mining because of its ease of execution and easier to understand compared to other algorithms. We can get more accurate and valuable results with the help of decision tree algorithm which can be useful for instructors to improve the student learning outcomes.

The ID3, C4.5 and CART decision tree algorithms has been applied on the data of students to predict their performance. But all these algorithms are used only for small database. For large database, we are using a new algorithm i.e. SPRINT which removes all the memory restriction and accuracy problem comes in other algorithms. It is fast and scalable than others because it can be implemented in both serial and parallel fashion for good data placement and load balancing.

In this work, SPRINT decision tree algorithm is used to solve the problem of classification in education system. Most of the current classification algorithms require that all or a portion of the entire dataset remain permanently in memory. This limits their suitability for mining over large databases. Accuracy and time complexity of SPRINT algorithm is much lesser than other decision tree algorithms.

Table of Contents

Certificate ii

Acknowledgment iii

Abstract iv

Table of Contents v

List of Figures viii

List of Tables x

1 Introduction **1-13**

1.1 Basic Classes 1

1.2 Advantages of Data Mining 2

1.3 Applications 2

1.4 Association of Data Mining 3

1.5 Educational Data Mining 3

1.6 Objectives of Data Mining 4

1.7 Phases of Data Mining 6

1.8 Fundamental Methodologies 6

1.9 Applications of EDM 7

1.10 Distribution Settings 8

1.11 Classification 8

1.11.1 Model construction 9

1.11.2 Model Usage 10

1.12 Decision tree 10

1.12.1 Advantages of decision tree 12

1.12.2 Limitations of decision tree 12

2	**Literature Survey**		**14-29**
	2.1	ID3	14
	2.2	C4.5	17
		2.2.1 C4.5 Procedure	18
	2.3	CART	18
		2.3.1 Features of Cart	19
	2.4	SLIQ	20
		2.4.1 SLIQ algo	21
	2.5	PUBLIC	21
		2.5.1 The Building stage	23
		2.5.2 DT pruning stage	24
	2.6	RAINFOREST	25
	2.7	SPRINT	26
		2.7.1 Calculation	26
3	**Current Study**		**30-38**
	3.1	Problem Formulation	30
	3.2	Objective of Problem	30
	3.3	Methodology	31
		3.3.1 Calculation	26
4	**Result and Discussion**		**39-54**
	4.1	Comparison	43
	4.2	Output	45
		4.2.1 Classifier accuracy	54

5 Conclusion and future scope **55**

5.1 Conclusion 55

5.2 Future Scope 55

References **56**

Appendix **59**

List Of Figures

Figure	1.1	Educational Data Mining	4
Figure	1.2	The sequence of applying Data Mining	5
Figure	1.3	Model construction	9
Figure	1.4	Model usage	10
Figure	1.5	Example of DT	12
Figure	2.1	SLIQ Methodology	22
Figure	3.1	DT of Data Set	35
Figure	4.1	Preview of Data Set imported in weka	39
Figure	4.2	Visualizing All Attributes used in Classification	40
Figure	4.3	Classification by J48 Decision Tree	41
Figure	4.4	Classification by CART	41
Figure	4.5	Classification by Decision Stump	42
Figure	4.6	Classification by Sprint Decision Tree	43
Figure	4.7	Sprint Decision Tree	45
Figure	4.8	Classifier Model	46
Figure	4.8.1	Spin Decision Tree	46
Figure	4.8.2	Spin Decision Tree	46
Figure	4.8.3	Spin Decision Tree	47
Figure	4.8.4	Spin Decision Tree	47
Figure	4.8.5	Spin Decision Tree	48
Figure	4.8.6	Spin Decision Tree	48
Figure	4.8.7	Spin Decision Tree	49

Figure 4.8.9 Spin Decision Tree 49

Figure 4.8.10 Spin Decision Tree 50

Figure 4.8.11 Spin Decision Tree 50

Figure 4.8.12 Spin Decision Tree 51

Figure 4.8.13 Spin Decision Tree 51

Figure 4.8.14 Spin Decision Tree 52

Figure 4.8.15 Spin Decision Tree 52

Figure 4.8.16 Spin Decision Tree 53

List OF Tables

1.1	The Training Data Set	11
2.1	Example of Pre Sorting in SLIQ	20
3.1	Example of Attribute List of Data Set	32
3.2	Data Set after presorting	33
3.3	Attribute List after splitting	34
3.4	Evaluating continuous spilt points	35
4.1	Parameter Comparison of DT algorithm	43
4.2	Classifiers Accuracy	54

Chapter -1
INTRODUCTION

Data mining is a new & growing zone of innovative work, both in scholarly & in trades. Data mining is alternately referred as knowledge discovery in database (KDD) that's a rising technique utilized as a part of the informative field to achieve the desired data & to excavate the veiled relationships helpful in decision making [2]. As we know decision making is an important aspect of life. It is required in everyone's life to decide something. So, the role of Artificial intelligence along with data mining is increased. At present huge amounts of data is being collected. Data mining is the mix of measurable demonstrating like linear regression, database storage & artificial intelligence technologies like marketing & manufacturing applications. The usual objectives of the info mining process deal with the separation of useful info from a large data & transform it into a reasonable arrangement for facilitation usage [1]. Data mining is the advance analysis of the "knowledge discovery in databases" methodology or KDD. In data mining, we first store large data sets & on basic of the algorithm, then we find particular features & their association with the problem [5].

1.1 Basic Classes

Data mining includes six basic classes of undertakings:

1.1.1 Anomaly location (anomaly/change/deviation recognition) – The recognizable proof of odd data records, that may premium or data mistakes that require facilitate examination.

1.1.2 Association run learning (reliance demonstrating) – Scans for associations among elements. As an instance, the market may amass data on customer gaining affinities. Using alliance control taking in, the market can make sense of which things are a significant part of the time acquired together & use this data for display purposes. This is sometimes implied as market compartment examination.

1.1.3 Clustering – The method of identifying & distributing the population or useful info nodes into several sets such that info nodes of the similar sets are more likely to be the same group to info nodes in the similar set & different from the info nodes in different sets. In simple terms, it is a class of data on the grounds of resemblance & variation among them.

1.1.4 Classification – This method is cast-off to sort various 'discrete' raw data that ensures the methodology of amassing the known structure or labeled class of data & then sort the unknown data sets on the grounds of the resemblance of the properties of unlabeled data groups & labeled data groups or sets.

1.1.5 Regression – This specific algorithm endeavors in order to sort several 'continuous' raw data variables into the labeled data on the grounds of the resemblance of the parameters of labeled data classes.

1.2 Advantages of Data Mining

Data mining provides many merits to administrations, communities, managements similar to a being. Data mining techniques are cost-effective & secure. These can also be used in identifying criminal suspects, in the medical field like predicting the symptoms & finding disease on the grounds of symptom's [5]. These can be used for the financial purpose also for example in cryptocurrency & analyzing the stock market [7]. Protection, notwithstanding, security, & abuse of data are huge subjects on the likelihood of resolving it appropriately.

1.3 Applications

Data mining can be applied for different domains like:

1.3.1 Banking - In managing an account, data mining is utilized as a part of MasterCard endorsement, customer prediction based on features of old clients & view the debt & income changes by month, districst, area & by different components.

1.3.2 Telecommunication industry - Data mining is utilized as a part of the telecommunication industry for recognizing possibly fraud clients & their utilization designs. It can be used to distinguish good & fraud customers to increase deceitful section to clients' records, find examples which may require exceptional consideration. It also maintains the condition of services to clients.

1.3.3 DNA examination - It analyzes every new & old pattern of each class in DNA investigation. It can be utilized as a part of distinguishing grouping patterns that may be parts of different infections.

1.3.4. Retail Market – The role of data mining in retail market is to find an association rule between different products like bread & butter. Besides these, it is also used to predict & analysis the purchasing behavior of the customer's. It will help in more customer satisfaction & hence it will enhance the particular retail market.

1.3.5 Science & Engineering - Data mining has been widely used in the sections of science & engineering, for example, bioinformatics, genetics, medicine, instruction & electrical power designing. It is the demand of the modern-day to use data mining with a neural network.

1.3.6 Intrusion Detection - As we know that everything has good & bad sides to it. Similarly, in the case of internet & technology, their are various grey hat hackers who try to hack our system either by spam email, phishing attack, etc. So, data mining can help it to reduce by mining out the desired negative features of such kind of actions. However, these can be further used with the neural network so that this intrusion can be detected at the initial phase.

There are certain disadvantages of data mining like privacy & accuracy problem. Privacy problem means that certain companies give their customer data to their technical team or any other third party which increases the risk of using their personal data in an inappropriate way. The other problem is accuracy because for this we have to select a model which must give the correct result. For this, we have to select various data mining methods like classification, clustering rules, etc.

1.4 Association of Data Mining

Data may contain qualities created & recorded at different conditions. For this circumstance finding significant associations in the data may require contemplating the brief solicitation of the characteristics. A common link may demonstration a causal relationship, or basically an association [8]. For ex- We have seen many customers who used to purchase some things in an association like bread & butter, bread & jam, etc. If we analyze these we will find a close relationship between these products [9]. So, these are basically used in the supermarket with the aid of association data mining.

1.5 Educational Data Mining

It is important about creating strategies for investigating the data that originate from educational database & by utilizing data mining methods; we can forecast student's scholarly performance & their conduct towards training (Yadav, 2012). Development (Education) theatres a vital part of a person's growth as well as the nation. It ensures the

nation's population growth. Mining in the scholastic situation is known as educational data mining (Ramachandram, 2010).

As a matter of fact mindful & huge information is secured in scholastic databank; useful information mining is the route toward unraveling & gathering a great deal of useful info set in databanks, data inventories or other info sources. As shown in figure 1.1 educational datamining is a combination of e-learning, datamining AI and learning analytics.

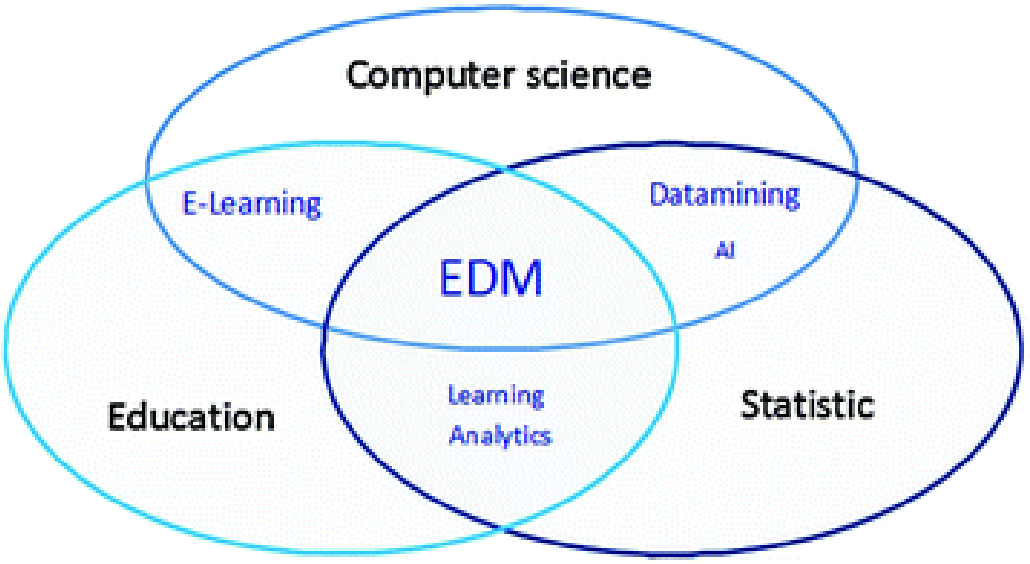

Figure 1.1 Educational Data mining

1.6 Objectives of Educational Datamining

1.6.1 Predicting students future learning conduct – With the use of students showing, this goal can be cultivated by making students models that contains the students' traits, including data, for instance, their knowledge, practices & motivation to learn[13]. The experience of the students & their satisfaction with learning is also evaluated with the help of EDM.

1.6.2 Discovering or enhancing area models – Through the various techniques & usages of EDM, exposure to existing models is possible. These help in enhancing new area models which can play a crucial role in EDM.

Considering the effects of educational assistance that can be practiced through learning structures [12]. Progressing genuine data about learning to students by applying & joining models, the field of EDM report & the development & programming.

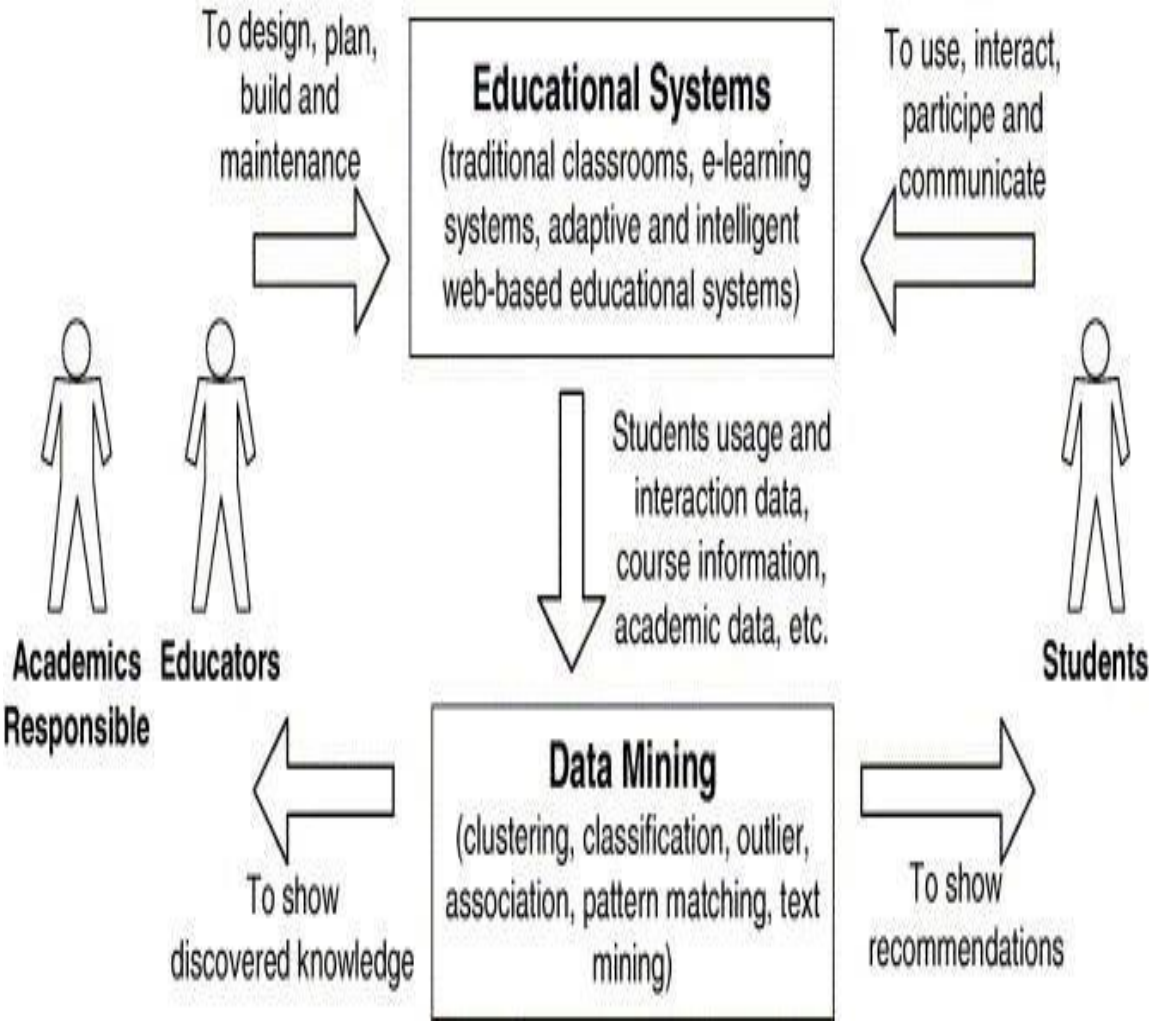

Figure 1.2: The sequence of applying data mining in the development system of students

As aforementioned in figure 1.2, we can see that teachers & scholastics dependable are responsible for outlining, arranging, formulating & keeping up the instructive frameworks. Students utilize & collaborate with them (Anyanwu, 2011).

So the utilization of data mining in training frameworks can be situated to various on-screen characters with every specific perspective (Anyanwu, 2011).

1.7 Phases of Educational Datamining

As research proceeds on the grounds of EDM have grown, numerous data mining frameworks have been associated. For every circumstance, the goal is to make an understanding of rough data into significant data about the learning method with a particular ultimate aim to choose better decisions about the blueprint & heading of a learning space. Along these lines, EDM, for the most part, comprises of following stages:

- The first period of the EDM procedure (not including pre-preparing) is finding connections in data [15]. This includes seeking through an archive of data from an educational situation with the objective of finding steady connections between factors. A few methods for distinguishing such connections have been used, including classification, clustering, association, regression, decision trees, artificial intelligence & many more. These connections should then be approved with a specific end goal to abstain from overfitting.
- Validated connections are connected to sort forecasts around upcoming proceedings in the learning condition.
- Extrapolations are utilized in aid for basic leadership procedures & policy decisions.

In above stages, data is frequently checked or in some other path refined for human judgment. A vast measure of research has been led for visualizing data.

1.8 Fundamental methodologies

The general classifications techniques specified, forecast, grouping & relationship mining are viewed as common strategies over a wide range of data mining.

1.8.1 Disclosure with models

In the Discovery with the Model system, a model is delivered by methods for desire, gathering or by human reasoning data planning & a short time later used as a section in another examination, to be explicit in estimation & relationship mining. In the prediction methodology usage, the pre-made model's desires are used to predict another variable [17]. For the utilization of relationship mining, the pre-made model sanctions the investigation between new expectations & extra factors in the study. In

numerous cases, an explore with models utilizes approved expectation models that had demonstrated generalizability crosswise over settings.

The key utilization of this technique includes finding connections between student's practices, attributes & relevant factors in the learning environment [18]. Further disclosure of wide & particular research done over an extensive variety of settings can likewise be investigated utilizing this strategy.

1.8.2 Refining of data for human judgment

Individuals can make inferences about data that may be past the expansion where an automated data mining strategy gives. For the use of preparing data mining, data is refined for human judgment for two key purposes, distinctive confirmation & order.

With the end goal of ID, data is refined to empower people to recognize surely understood examples, which may some way or another be hard to interpret [19].

Data is additionally purified for the purpose behind ordering highlights of data, which for educational data mining is utilized to help the advancement of the forecast shows. Grouping speeds up the improvement of the forecast demonstrate, enormously. The objective of this strategy is to condense & display the data in a valuable, intelligent & outwardly engaging path with the aim to comprehend the educational information & improve decision-making skills.

1.9 Applications

A run over of fundamental usages of EDM is given by Cristobal Romero Sebastian Ventura. In their logical order, the uses of EDM application are:

- Recommendations for students
- Predicting students execution
- Student illustrating
- Market Basket Analysis for MBA students.
- Grouping students.
- Social sort out examination
- Weather Forecasting

EDM can be connected to course administration frameworks, for example, open-source Module. New research on compact learning conditions moreover prescribes that data mining can be important. Data mining can be used to help give arbitrary substance

to convenient customers, paying little mind to the qualifications in supervising content between phones & standard PCs & web programs [21].

New EDM applications will revolve around allowing non-specific customers use & dealing with progressively accessible for all customers of EDM. Cases fuse computable & observation contraptions that investigate social frameworks & their effect on learning results & gainfulness.

1.10 Distribution settings

Significant measures of EDM work are distributed at the associate explored International Conference on EDM.

These are as per the following:

- Classification.
- Clustering.
- Regression.
- Artificial Neural Network.
- Convolutional Neural systems.
- Decision trees(DT).
- Genetic Procedure(GP).
- Association rules & etc.

These system assist the customers for examination of data from vaious provisions, sort them & distinguish the relationships among the mining practice (Yadav, 2012). Unconventional terms for data mining are KDD (Knowledge disclosure database), learning, excavation, design examination & so on.

1.11 Classification

Classification is a data mining technique utilized for execution change in the education system. It depends on predefined data of the items utilized as a part of collection comparable data questions together. Classification is among the several algorithm that are recognized as an essential field in the developing arena residing in data mining. It maps data into predefined groups of classes.

The classification has played a vital part in data mining manner. It is considered largely by the machine learning enthusiasts as a conceivable answer for the useful information procurement issue [22]. Properties with distinct spheres are considered to be a categorical

value, whereas ones with the well-ordered spheres are denoted as numeric. The goal of this is to build a prototype model which will follow the properties of EDM. This aims a prototype model for every group of set keeping in mind the properties. The prototype model is further utilized by the classifier to categorize future records whose group of the set is anonymous.

Classification consists of two types:

1.11.1 Model development

It encompasses a set of predefined classes. In this, a pre-existing arrangement of data logs is partitioned between preparation & assessment data indexes. The preparation dataset aids in creating the prototype machine learning model whereas the validating dataset aids in the substantiation of the same. Following is the model in figure 1.2.

The machine learning prototype model is utilized for grouping & anticipating the novel arrangement of data logs which are not the same as the preparation & test data records. Supervised learning is utilized by this model due to its pre-existing information of the categorized data records which ensures characteristic determination simple & this prompts great classification accuracy.

The classifier i.e. model will take data from the training set. This training set will be applied on classification algorithm & based on those it will be decided which professor will get tenure. When this preparation is done, we have a prepared model that can be utilized for this.

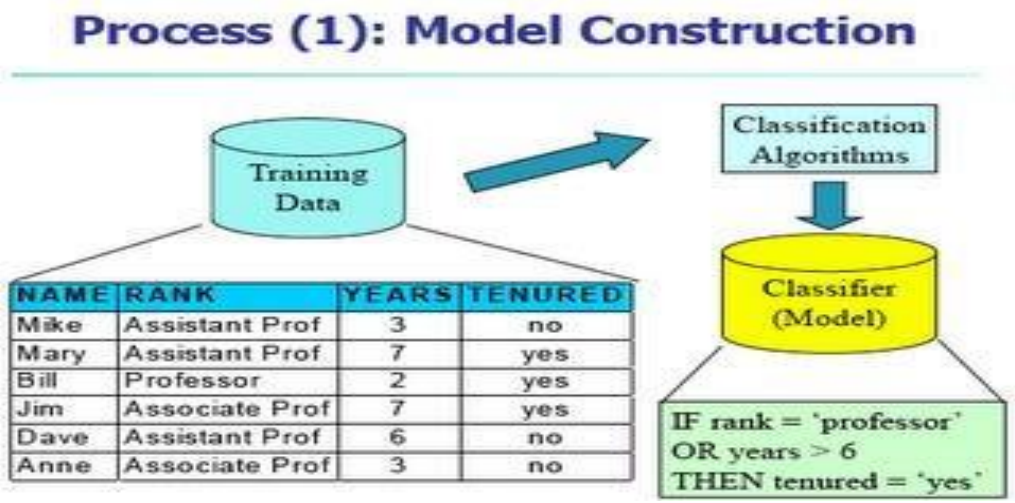

Figure 1.2 Model Construction

1.11.2 Model usage

This model is applied for characterizing imminent articles. Accuracy degree refers to the amount of preparation data set examples that are accurately ordered by the machine learning prototype model (Bhardwaj, 2011)

The test set is self-governing & doesn't bound with the preparation data set, in case if it not independent then over-fitting will occur. Basically, in this, we use unknown data for classification.

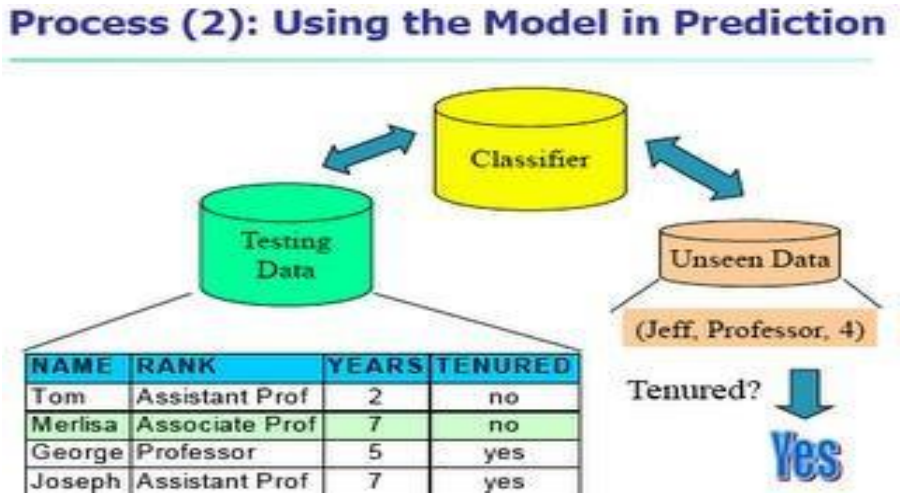

Figure 1.3 Model Usage

As shown in figure 1.3, initially a model is develop for precision. Once the exactness is worthy, we utilize the model to characterize new data. So when another personnel arrives, the model would appoint a class name contingent upon regulations that was found out from the model development stage.

1.12 Decision Tree

Decision tree (DT) looks like a flow chart tree arrangement that is used for making a decision based on a certain set of training data. DT procedure revolves around repetitively also that segments uses information logs via a depth-first search approach. It is widely accepted as one of the most used data mining systems. A DT assembly is consisting of root nodes, inward nodes & at the bottom of the tree – leaf nodes which classify the outcome [23]. The tree arrangement is employed as a part of grouping obscure data records. The consequence of the DT in a problem of identifying the results of candidates forecasts the result of several applicants of an examination in terms of which candidates will be passing, getting failed or endorsed to afterward year [23].

DT constructional methodology is implemented in two phases: tree construction & tree pruning. These are basically a top-down approach. In this, a tree is divided again & again till we reach at the end [22]. It is exceptionally entrusting & computationally concentrated as the preparation data index is crossed more than once. A DT portrays regulations in order to isolate the records into classes.

Trees can be smoothly transformed to SQL database which can further be utilized for reaching the databanks proficiently. DT classifiers get comparative & at times improved precision whenever contrasted & other algorithm. The three generally utilized DT cultured algorithm s are CART, ID3, & C4.5. These algorithms are utilized just for the little dataset as shown in table 1.1.

Table 1.1 : The training data set

AGE	**CAR TYPE**	**RISK**
23	Family	HIGH
17	Sports	HIGH
43	Sports	HIGH
68	Family	LOW
32	Truck	LOW
20	Family	HIGH

A DT is a set differentiator that repetitively segments - preparation dataset till the point that every one of the segment comprises of completely or predominantly of cases from one class [14].

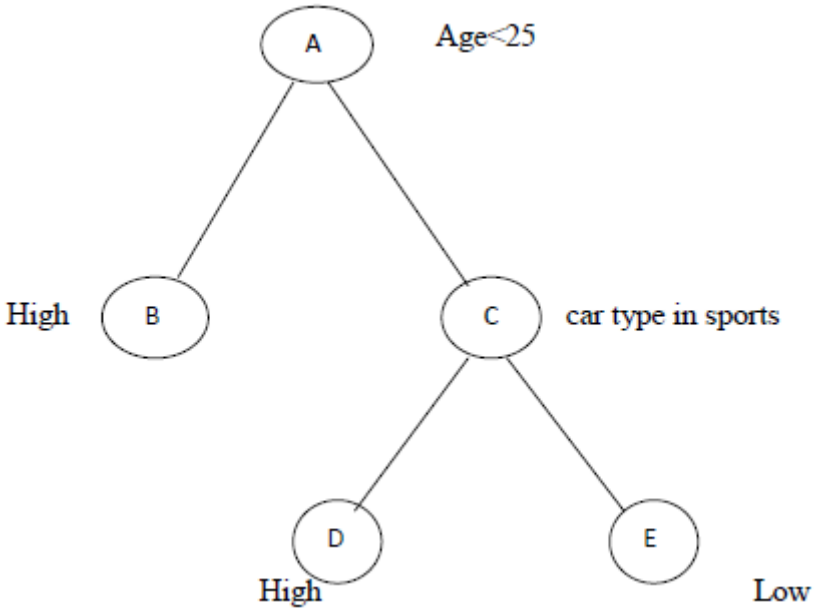

Figure 1.4: Example of DT

Figure 1.4 demonstrates an example of DT classifier in view of the preparation set appeared in table 1.1. There are 2 part focuses that segment the records into high & generally safe classes.

1.12.1 Advantages of the DT are:

- Less mistake.
- Its speed is considerably more than other models.
- It can be changed over into basic & simple techniques to comprehend grouping rules.
- Decay is less demanding as contrasted & other techniques. It speaks to learning as If-then standards. Tenets are less demanding for people to get it.

1.12.2 Limitation of Decision Tree:

- It requires certain data of quantitative or factual experience to finish the procedure precisely.
- There are excessive data likewise an issue, while the inadequate data can make challenges in the DT process.
- Factors are hard to incorporate into DT.

The three generally utilized DT knowledge algorithms are CART, C4.5, & ID3. These algorithm s are utilized just for the little dataset. All the 3 algorithm depend on Hunt's technique & are executed in the serial request. One of the drawbacks of these algorithm is low arrangement precision when the preparation data index is substantial [21]. This deficiency of serial DT execution is tended to by SPRINT algorithm. In the run, the preparation data index isn't memory occupant however circle inhabitant. This approach enhances the grouping of exactness & lessens misclassification mistakes. In this examination, we concentrated on serial usage of SPRINT DT algorithm that is space (memory) inhabitant, quick & simple to actualize contrasted with parallel tree execution that is intricate to actualize.

Chapter 2
LITERATURE SURVEY

Various data mining procedures are prepared for educational data mining (EDM) in order to enhance the throughput of pupils such as “Bayes classification, Regression, Genetic algorithm k-means clustering, prediction & so on”. Data mining strategies have keenly taken in part of Education domain to upgrade the acceptance of “learning process” while stressing on detecting, exploring & analyzing data linked to the “learning process” of pupils.

DT algorithm is often actualized in a sequential or corresponding form in view of the capacity of data, space (in terms of memory) accessible over PC & versatility of the algorithm [3]. The CART, C4.5 & ID3 DT algorithm s are linked on the data of students to anticipate their throughput. In any case, these are helpful for only those data set whose training data collection is small. This algorithm is clarified below:

2.1 ID3

Iterative Dichotomiser 3 is a DT algorithm is proposed by Quinlan Ross. It formulates on the grounds Hunt's algorithm. ID3 utilizes “information gain” technique to select the “splitting property”. ID3 extract the data & concedes all qualities while constructing a tree. ID3 do not offer exact outcome whenever noise is presented & it is consecutively executed. In this manner, a concentrated pre-preparing of records is completed afore creating a DT prototype with ID3 [23].

Therefore a pre-handling of data is done before building a DT show with ID3. The fundamental thought utilized as a part of ID3 states that it uses different algorithms which are utilized to builds DTs in a top-down recursive partition.

The tree begins as root with all training datasets (test or tree split decision attribute) & branch is made for each estimation of the root property & the examples are divided in the same way [5]. The algorithm utilizes a similar procedure recursively to frame a DT at each segment. Once a property has happened at root, it will not be considered in some other of the root's descendant’s. The recursive dividing stops just when any of the conditions become valid.

2.1.1 ID3 Algorithm

ID3 (Instances, goal property, Properties)

- Make a "root" of DT.
- When entirely the instances are in affirmation, "output single- node" DT, with tag=+.
- When entirely the instances are in negation, "output single-node" DT, with tag =+.
- When the quantity of forecasting properties is null, forwards "return the single-node" DT, with tag= "most similar value" of the goal property among the instances.
- Or else initiate
  - Ai = property that groups the instances best.
  - DT property for "root" = Ai.
  - For every affirmative node, vi of Ai.
  - For every potential node vi of Ai,
    1. Include a novel DT branch under root, adjacent to the check Ai= vi.
    2. Instances (vi) is the subgroup of instances holding the cost vi for Ai.
    3. When instances(vi) is null,
       - Include a leaf node with tag = "most similar goal property" value among the instances under the novel branch of DT.
       - Otherwise include the "subtree ID3 (instances(vi), goal property, Property-{Ai})" under the novel branch
  - Finish
  - Outputs "root"

In order to discover the best way for categorizing a "learning set" is either the height of DT needs to be minimalized or minimalize the query requested. Therefore, a methodology is required that can quantifies among the whole set – which queries offers the well-poised "splitting".

- **Estimating polluting influence:**

We have a database comprising of properties & labels for the corresponding property, the similarity & dissimilarity for records present in the database can be identified on the grounds of their classes.

A database is clean or similar when the database is consists of one & only set. When the database is comprised of multiple sets then in succession the database is called polluted or homogeneous.

However there are a lot of alternatives to quantify the pollution inside the database & widely used practices are "Entropy, Gini index & Classification error".

$$Entropy = \sum_j -p_{j \log p_j} \quad (2.1)$$

"Entropy" for the pure table is nil & the reason behind this conclusion is that the "Entropy" achieves the possible max value whenever the sets in the database have equivalent probability, thus probability becomes 1 & log (1) = nil. For data set S

$$\text{Gini Index} = 1 - \Sigma p_{ji}^2 \quad (2.2)$$

In the above formula, Pj resembles the occurrence of set j_i in Si. When a "split" distributes Si into two subsets S1 & S2, the index of the divided data Gini split(S) is given by the following formula 2.3, where n is number of data set.

$$Gini\ split(S) = {}^{n_1}/_{n}\ Gini(S1) + {}^{n_2}/_{n}\ Gini(S2) \quad (2.3)$$

The merits for indices are that its estimation needs just the circulation of the set values for every segment. In order to find the finest "split point" for a nodule, scanning of every nodule's property contents is done & assess the "splits" on the grounds of the same property.

The property comprising the "split point" having the minimum worth for "Gini index" is further utilized to distribute the nodule. "Gini index" of the homogeneous database comprises of only one set is nil since the probability is 1 & $1\text{-}1^2$ =0. "Akin to entropy, the Gini index also reaches the maximum value when all classes in the table have equal probability."

$$\text{Classification error} = 1 - \max\{Pj\} \quad (2.4)$$

"Classification error index" is akin to "Entropy & Gini index", as when this method is applied to the homogeneous database its output is zero since the probability is 1 & 1-max (1) = 0. The classification error-index output mostly lies in the range of 0 &1. For a

matter of fact the highest “Gini index” for a proposed sum of sets is permanently equivalent with the maximum of “classification error index” since, for the majority of sets n_i, the probability is set to $p_i = {}^{1}/_{N_i}$

- **Splitting Criteria:**

To decide the best trait for a specific hub in the tree we utilize the measure called Info gain. The Info gain, gain(S_i, A_i) of a characteristic A_i, in respect to a gathering of models Si, is characterized as

$$Gain\ Ratio = {}^{Gain(S_i A_i)}/_{Split\ Info} \tag{2.5}$$

The way toward choosing another trait and dividing the dataset is currently rehashed for each non-terminal relative hub. Traits that have been joined higher in the tree are prohibited, with the goal that any given property can show up all things considered once along any way.

2.2 C4.5

C4.5 is an upgraded version of ID3 computation made by “Quilan Ross”. C4.5 relies upon “Hunt's computation” & moreover alike ID3, it is consecutively executed. Pruning occurs in C4.5 by exchanging the inward nodule with a leaf center point thusly diminishing the botch rate. C4.5 recognizes both diligent & obvious characteristics in structuring the DT. It has an improved methodology for “tree pruning” that diminishes misclassification faults in light of bustle & such countless nuances in the readiness enlightening gathering [2]. Alike ID3 the records are orchestrated at each nodule of the tree for choosing the finest “splitting attribute”. C4.5 utilizes “Gain-ratio heterogeneity” methodology to survey the split property (Bhardwaj, 2011).

The C4.5 methodology has the following focal points:

a) Managing qualities with various expenses.

b) Managing preparing data with missing trait values C4.5 permits property estimated to be tagged as '?' for lost. Missing trait esteems are basically not utilized as a part of “Gain & Entropy” computations.

c) Dealing with both "consistent and discrete qualities" in order to manage predictable properties, C4.5 makes a farthest point and a while later parts the once-over into those

whose trademark worth is over as far as possible and those that are not actually or comparable to it.

d) Pruning tree a short time later definition C4.5 return to the tree once it's made and tries to remove branches that don't help by substituting them with leaves (leaf knobs) [23].

2.2.1 C4.5 Procedure

• If all instances are of a similar class, the tree is a leaf thus the leaf is returned marked with this class.

• For every characteristic, compute the possible useful data given by a test on the characteristics (in light of the probabilities of every instance having a specific incentive for the characteristic). Additionally, figure the increase in the data that would provide output from an experiment on the characteristic (in the light of the probabilities of every instance with a specific incentive for the trait related to a specific class).

• Reliant on the present decision premise, discover the best possible characteristic to branch on.

2.3 CART

CART represents "**classification & regression trees"** & was presented by Breiman in the year 1984. CART forms "**classification & regression trees"**. The classification tree development via CART depends on the twofold split of the properties. It is likewise founded on Hunt's calculation & can be actualized sequentially. It utilizes the gini file part extent for choosing the split property. "CART" is one of a kind from other Hunt's based calculation as it is additionally utilized for 'regression' investigation with the assistance of the "regression trees" [3]. The regression assessment highlight is utilized in determining a needy variable given a lot of indicator factors over a provided timeframe. It utilizes many "single-variable split measures - *gini* list, symgini & so on & one multi-variable" in deciding the finest split node & information is put away at each nodule to decide the best split index [8].

"SALFORD SYSTEMS" executed a variant of CART called CART utilizing the first program of Breiman (1984). CART has upgraded highlights & abilities that acknowledges the weaknesses of CART offering ascend to a cutting edge DT classifier

with high arrangement & forecast precision. The CART DT is a parallel repetitive dividing technique equipped for handling continuous & ostensible traits both as targets & indicators. Information is dealt with in their crude structure; discarding is not required or prescribed. Trees are developed to the greatest size without the utilization of a ceasing guideline & after that pruned back to the root through cost-multifaceted nature pruning. The strategy yields trees that are immutable under any request protecting change of the indicator traits. The CART component is planned to create not one, rather a succession of settled pruned trees, which are all applicant ideal trees.

The privilege measured or pure tree is distinguished by assessing the prescient exhibition of each tree in the pruning arrangement. CART provides no inner exhibition dealings for tree choice dependent on the preparation information all things considered measures are regarded, suspect. Rather, tree throughput is constantly estimated on autonomous test information (or by means of approval) & tree determination continues simply after test-information based assessment. On the off chance that no test information is available & cross approval has not been executed, CART will stay skeptic in regards to which tree in the grouping is ideal.

This is in sharp difference to strategies, for example, C4.5 that produce favored models based on preparing information measures. The CART system incorporates programmed class adjusting, automated values management, & takes into account cost-delicate learning, dynamic attribute creation, & likelihood of tree estimation. The last reports incorporate a new property significance standing.

The CART creators additionally kicked off something new in indicating how authentication can be utilized to evaluate execution for each tree in the pruning serialization provided trees in various CV folds often not adjust on the magnitude of terminal nodules.

2.3.1 Features of Classification & Regression Trees (CART)

- Tests in CART are always binary.
- CART utilizes the Gini assorted variety file to grade tests.
- CART prunes trees utilizing a cost-multifaceted nature model whose attributes are assessed by authentication.
- CART searches for alternate tests that surmised the results when the tried parameter has an obscure worth.

2.4 SLIQ

SLIQ means "supervised learning in the quest". It was presented by Mehta et al. SLIQ is a quick adaptable choice tree calculation that can be executed in a sequential & parallel example. It did not depend on HUNT'S Algorithm for choice tree grouping. It segments a preparation informational index recursively utilizing a broadness first insatiable methodology that is incorporated with the pre-arranging procedure during the tree building stage. The principal procedure utilized in SLIQ is to execute a plan that disposes of the need to sort the information at every hub of the choice tree. In structure, a choice tree model SLIQ handles both numeric & unmitigated traits (Rissanem, 2010). Arranging of information is required to locate the split for numeric properties. The table 2.1 represents the condition of the information structures when pre-arranging.

Table 2.1 Example of pre-sorting in SLIQ

TRAINING DATA

Age	salary	Class
30	65	G
23	15	B
40	75	G
55	40	B
55	100	G
45	60	G

AFTER PRE-SORTING

Age	Class list Index
23	2
30	1
40	3
45	6
55	5
55	4

AGE LIST

salary	Class List index
15	2
40	4
60	6
65	1
75	3
100	5

SALARY LIST

Class	leaf
G	N1
B	N1
G	N1
B	N1
G	N1
G	N1

CLASS LIST

It utilizes “minimum description principle” (MDL) in pruning the tree subsequent to developing it MDL is a costly procedure in tree pruning which utilizes the minimal sum of programmed code in delivering a tree that is little in size utilizing base up technique[11]. For numeric traits, arranging time is the overwhelming element when finding the best split at a DT node. In this manner, the principal method utilized in SLIQ is to execute a plan that dispenses with the need to arrange the information at every node of the DT.

Past calculations sort information at each nodule in the tree. Utilizing the different rundown information structure, SLIQ just sorts information once toward the start of the tree constructing stage. We utilize the accompanying information structure to accomplish this pre-arranging. A different rundown is made for each property of preparing information.

Moreover, a different rundown, called "class list", is made for the class marks connected to the instances. A section in both the property rundown list & the class rundown list has two categories. In property rundown list the primary one comprises of characteristic estimation & the second one a file into the class rundown list.

In the class rundown list, the primary value comprises a class name, & the second one mention to a leaf node of the DT. The i^{th} item of the class rundown list compares to the i^{th} instance in the preparation information. Each leaf nodule of the DT speaks to a parcel of the preparation information, the segment being characterized by the combination of the predicates on the way from the leaf to the root.

2.4.1 SLIQ Algorithm:

- Compute Split()
- for every characteristic A_i do
- go across a characteristic rundown list of A_i
- for every point v_i in the characteristic rundown, list do
- search the adjacent item in the class rundown list, &
- Therefore the adjacent class & the left node(supposedly l)
- apprise the class histogram at leaf l
- when A_i will be a numeric characteristic then
- at that point evaluate split key for assessment (A <= v) for leaf l
- when A_i will be a continuous property
- for every leaf nodule tree do
- search subclass of A_i with the finest split point.

The leaf mention entries of the considerable number of passages of the class rundown list are set to direct to the 'root' of DT. At that point pass is made over the preparation information, dispersing the estimation of the properties for every model over every one of the rundowns. Every characteristic record is likewise labeled with the relating class rundown file. The characteristic records for the numeric highlights are then arranged autonomously.

This sort of deficiency of sequential DT execution is tended by “SLIQ & SPRINT calculations”. The preparation informational index isn't memory dependable but hard disk-resident. The SLIQ methodology as shown in figure 2.1 develops grouping precision & decreases classification errors by partitioning. SLIQ DT calculations are memory occupier, quick & simple to actualize contrasted with parallel tree execution which is complicated to perform. One of the weaknesses of SLIQ is that it utilizes a class rundown list information structure that is residing in the memory in this way forcing memory limitations on the information.

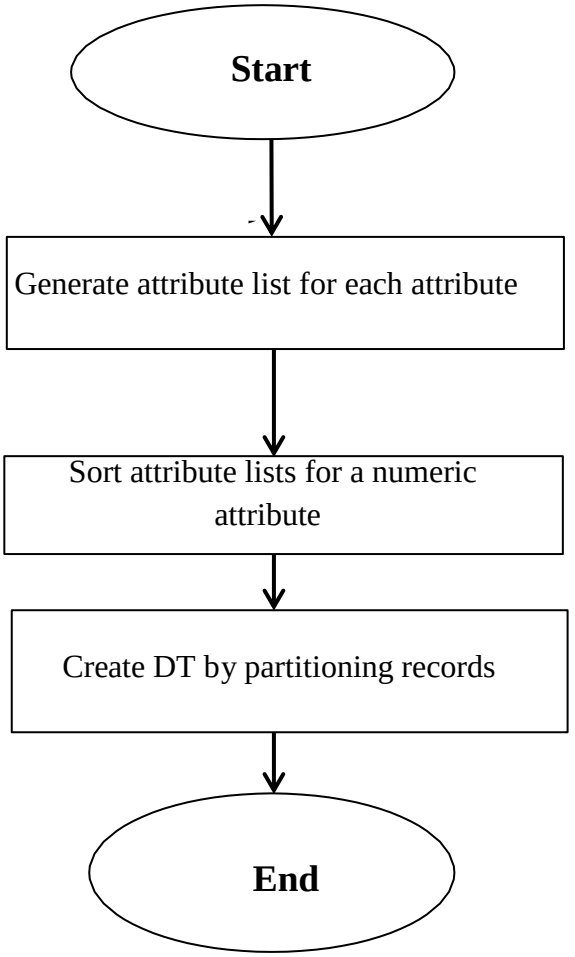

Figure 2.1 SLIQ methodology

2.5 PUBLIC

It represents "Pruning & building incorporated with classification". PUBLIC is a DT classifier that during the developing stage, first decides whether a nodule will be pruned during the subsequent pruning stages, & quits extending such nodules. Henceforth, PUBLIC incorporates the pruning stage into the structuring stage as opposed to executing them in a steady progression.

Customary DT classifiers, for example, ID3, C4.5, & CART, for the most part, develop a DT in two unmistakable stages. In the primary structure stage, a DT is primarily created by the recursive database, while in the subsequent pruning stage, nodules in the assembled tree are pruned to increase exactness & counteract overfitting [4].

So as to decide, during the creating stage, nodules that are sure to be pruned, we have to understand the expense of encrypting the subtrees at the nodules. For this, we created three procedures for figuring a minimal limit on the expense of a subtree at a "yet to be expanded" leaf nodule. By playing out an extra calculation, each successive strategy can produce increasingly precise appraisals for the base expense subtree.

2.5.1 Tree building stage:

In open DT calculation, the tree is manufactured breadth-first by respectively splitting the information until each segment is unadulterated. It implies each part consists of data having a place with a similar class. The parting condition for parceling the information is both of the structure A< v if A will be a numeric property. Henceforth, each split is paired. Its structure calculation is:

The technique to build tree (T_i):

- Introduce root nodule utilizing informational collection set T_i
- Introduce queue D_i to contain root hub
- While D_i isn't void do {
- De queue the main nodule M_i in D_i
- On the off chance that M_i isn't pure {
- For each property P_i
- Assess parts on property P_i
- Utilize the finest spit to part node P_i into $P1_i$ & $P2_i$
- Add $P1_i$ & $P2_i$ to D_i
- }

- }

A disadvantage with playing out the structure & pruning activities in independent stages is that it brings about a considerable lot of squandered exertion since a critical bit of the tree produced during the structure stage may thusly be pruned during the pruning stage.

For this, we projected a classifier, PUBLIC, that incorporates the pruning stage onto the construction stage. In particular, nodules that are sure to be pruned aren't extended throughout the construction stage—therefore, less number of nodules are extended throughout the construction stage, & in this way the measure of work done (e.g., plate I/O) needed to build the DT is diminished.

2.5.2 DT pruning stage:

We utilize this stage to keep the tree safe from the overfitting issue. To make sure this happens, MDL rule is connected to "prune the tree" worked in the developing stage & make it more casual. At that point, we exhibit a pruning algorithm that, with regards to our encoding plan, finds the "base" cost subtree of the tree built in the developing stage.

Subsequent to doing this, we are registering the base cost subtree of the tree developed in the building stage. The straightforward recursive calculation figures the base cost sub tree established at a subjective hub N_i & returns its expense. Given S_i an arrangement of records related to N_i. On the off chance that N_i is a leaf, at that point the base cost sub tree established at N_i is essentially Ni itself. The expense of the least expensive subtree established at N_i is $C_i(S_i)$ +1.

Pruning algorithm for public algorithm is:

Procedure compute cost & prune (Node N)

- If N is a leaf return (C(S)+1)
- Mincost1= compute cost & prune(N1);
- Mincost2= compute cost & prune(N2);
- MincostN= min{C(S)+1,Csplit(N)+1+mincost1+mincost2};
- If mincostN= C(S)+1
- Prune child nodes N1 & N2 from tree
- Return minCostN

On the off chance that N is an inner hub inside the tree with N_i1 & N_i2 as children, at that point, two decisions for the base cost subtree came into being:

- The hub N_i itself without any further node.
- Hub N alongside N_i1 & N_i2 as children & the base cost subtrees established at N_i1 & N_i2.

Out of the aforementioned decisions, the primary with the less expensive value brings about the minimal cost subtree for N_i.

In PUBLIC algorithm, creating the DT in two stages will bring about a considerable measure of squandered exertion from the time when a whole subtree developed in the principal stage may be pruned later in the following stage. “PUBLIC” is a better DT classifier that coordinates the second pruning stage with the underlying “building” stage. However, a hub isn't extended during the construction stage, in the event that it is resolved that it will be pruned throughout the pruning stage. Afore it is extended, PUBLIC processes a lower bound on the minimal cost subtree established at the nodule.

This approximation is then utilized by the PUBLIC to recognize the nodules that are “sure to be pruned” & for those hubs which are not consuming exertion on splits.

2.6 Rainforest

It gives a system for quick DT developments of huge datasets. In this algorithm, we have a binding system for DT classifiers that isolates the versatility parts of algorithm for developing a DT from the focal highlights that decide the nature of the tree. This conventional calculation is anything but difficult to substantiate with explicit algorithm s from the recent study (counting CHART, C4.5, QUEST, ID3 & expansions, SLIQ, Sprint, & CHAID).

Rainforest is a usual structure which is utilized to minimize the gap between the confinements to primary memory datasets of algorithm s in the AI & research work & the versatility necessities of an information mining [12].

In the rainforest structure, we utilize a voracious top-down DT acceptance scheme. At that point, we illustrate how this mapping can be cultured to the nonexclusive “Rainforest Tree Induction Schema” & detail how the partition of adaptability problems from quality concerns is accomplished.

DT algorithm fabricates the tree top-down in an accompanying manner: At the root hub r_i, the databank is analyzed & the finest split condition criti(r_i) is registered.

Repetitively, at a non-root hub n_i, F (n_i) is inspected & from it, criti (n_i) is figured. An intensive assessment of the algorithm s in the research work demonstrates that the covetous (greedy) pattern can be cultured to the nonexclusive "Rain Forest Tree Induction Schema."

2.7 SPRINT Algorithm

It represents the "Scalable Parallelizable Induction of Decision Tree" calculation It is quick, adaptable DT classifier. It doesn't depend on Hunt's algorithm in building the DT, but it segments the preparation informational collection repetitively utilizing BFS a greedy approach until every segment has a place with a similar leaf hub or class. It tends to be executed in both sequential & parallel manner for good information locality & load adjustment [3].

SPRINT algorithm is intended to be effectively parallelized, enabling numerous processors to cooperate to fabricate a solitary predictable model. This parallelization shows incredible versatility to the clients.

It gives fantastic speedup, size up & scale-up properties. The blend of these properties or attributes makes Sprint a perfect device for information mining.

2.7.1 Calculation

- Splitting (information S_i)
- When all nodes are in S_i are in a similar class at that point
- Return;
- For every characteristic A_i do assess parts on characteristic A_i;
- Utilize the finest split discovered to segment S_i into S_i1 &S_i2;
- Splitting (S_i1);
- Splitting (S_i2);
- first call: Splitting (Practice information)

The noteworthy issues that have basic execution suggestions in the tree-development stage have been identified:

- Instructions to discover split focuses that characterize hub tests.
- Having picked a split point, instructions to segment the information.

Attribute rundown list & histogram are two data structure that has been utilized that isn't memory residing making SPRINT appropriate for enormous informational collections, in this way it eradicates every one of the information memory confinements on the information. It manages both persistent & classified characteristics. Data structures of SPRINT are clarified underneath:

- **Attribute list**: SPRINT at first makes a trait list for every trait in the records. Sections in these lists, which we call attribute records, consist of an attribute value, a class label & the index of the record from which these values were obtained. Initial list for continuous attributes is sorted by attribute value once when first created. On the off chance that the whole information doesn't adjust in main memory, attribute rundown list is kept up on hard drives. The underlying rundown list made from the preparation set is related to the foundation of the ordered tree. As the tree is developed & nodules are part to make new children, the characteristic records having a place with every nodule are divided & connected with the child nodes.
- **Histograms**: Two histograms are related with every DT hub that is under thought for the split. These histograms represent as C_{below} which handles information that has been prepared & C_{above} which handles information that hasn't been managed. Classified properties additionally have a histogram related with a nodule.

  Be that as it may, just a single histogram is required & it comprises of a class dispersion for each estimation of the given characteristic named as check network.

  SPRINT has likewise been intended to be effectively parallelized. Estimations of this parallel execution on a mutual nothing IBM POWER parallel framework SP2. SPRINT has magnificent scale-up, speed up & size up characteristics. The blend of these attributes makes SPRINT a perfect device for information mining [10].

Endogen & timer in the year 2005 utilized EDM to detect & upgrade instructive procedure which can enhance the basic judgment process [9].

Han & Kamber in the year 2006 depicts information mining programming that enables the clients to dissect information from various measurements, arrange it & outline the relationship which is recognized throughout the procedure [9].

Al-Radaideh, et al connected a DT prototype to anticipate the last grade of students who concentrated the c++ course in Yarmouk college, Jordan (2005). Three diverse order

techniques in particular ID3, C4.5 & the Naives Bayes were utilized. The result of the outcome shows that DT model would have wiser forecast than different models.
Z.J.Kovacic introduced a contextual investigation on instructive information mining to recognize at what degree the enrolment information can be utilized to anticipate student's prosperity. The algorithm CHAID & CART were connected on student enrolment information of data framework students of open polytechnic of New Zealand to get two DT grouping effective & ineffective students. The precision got with CHAID & CART was 59.4 & 60.5 individually [18].

Hijazi & Naqvi directed as concentrate on the student execution by choosing an example of 300 students from a gathering of schools partnered to Punjab college of Pakistan. The theory that was expressed as "students mentality towards participation in class, hours spent in the concentrate on a regular schedule after school, students family pay, students mother's age & mom's scholastic background are altogether associated with student performance." was confined. By methods for straightforward direct regression investigation, it was discovered that elements like mother's scholastic background & student's family salary were very connected with the student scholastic execution [18].

Galit provided a contextual analysis that utilizes students information to examine their learning conduct to anticipate the outcomes & to caution students in danger before their last, most important tests [9].

Bharadwaj & Pal [3] got the college students information like participation, class test, course & task marks from the student's past database, to anticipate the exhibition toward the finish of the semester with the assistance of three DTs. It was seen that C4.5 is the best algorithm between ID3, CART & C4.5 [23].

Bray in his investigation on private mentoring & its suggestions, saw that the level of students accepting private coaching in India was generally higher than in third world countries. It was likewise seen that there was an improvement of scholastic execution with the force of private mentoring & this variety of power of private coaching relies upon the aggregate factor in particular financial conditions.

Khan led a presentation to consider on 400 students including 200 young men & 200 young ladies chosen from India's Aligarh established - Aligarh Muslim University with a primary target to build up the predictive estimation of various proportions of comprehension, character & statistic factors for progress at a higher auxiliary level in science branch. The choice depended on bunch testing method in which the whole populace of intrigue was separated into gatherings or groups, & an irregular example of these groups was chosen for further investigation. It was discovered that young ladies with high financial status had moderately higher scholarly accomplishment.

Chapter -3

CURRENT STUDY

3.1 Problem Formulation

DT arrangement algorithm can be executed in a sequential or parallel style dependent on the amount of information, memory accessible on the PC asset & adaptability of the calculation. The primary burdens of sequential choice tree calculation (ID3, C4.5 & CART) are low characterization precision when the preparation information is huge. The C4.5, ID3 & CART DT algorithm s are as of now connected on student's information to anticipate their exhibition in the last test of the year. However, everything is utilized uniquely for little informational index & necessitated that whole or a segment of the data remain forever in main memory. This restricts their appropriateness for mining over huge databanks. This issue is fathomed by SPRINT DT algorithm . In sequential execution of SPRINT, the preparation informational collection is repetitively segmented utilizing BFS method.

Under the current research study, the dataset of 600 students has been taking from B.Tech. (CSE/IT) by considering the info parameters as - name, reg. no., marks, their open elective subject in fourth sem., midterm marks, end term marks, number of study hours, number the decision of Open elective subject, surveying ought to be there? Truly or no, recommendation with respect to surveying: if yes then why & if no then why? There are 9 OE subjects in B.tech. (CSE/IT) & as a result of constrained sheets, the greater part of the students don't get their own particular decision of subject. It could impact on their execution in an exam. So the yield would turn out to be the way students are performing as indicated by the decision of their inclination.

3.2 Objectives of Problem

The objectives of the present investigation are framed so as to assist the low academic achievers in higher education & they are:

- Distinguishing proof of the selection of students in surveying framework which influences an student's performance during the scholastic profession.
- Authentication of the created prototype for advanced education students concentrating in different colleges or foundations.
- Prediction of students' performance in their last exam of the year.

In this proposed work, the actualizing of SPRINT DT algorithm for enhanced order precision & decrease misclassification mistakes & execution time is done. This algorithm is clarified & after that serial execution is applied on it to discover the actual outcomes. Differentiating it & other existing algorithm to find which will be dynamically viable to the extent the decisively visualizing the consequence of the student & time taken to construe the tree.

3.3 Methodology

Scalable Parallelizable Induction of Decision Tree or SPRINT was introduced by Shafer et al. It's a quick & accessible DT classifier. It did not depend on Hunt's algorithm in developing the DT, instead, it segments the preparation informational collection data repetitively by utilizing bfs insatiable procedure until each segment has a place with a similar leaf nodule or class. It tends to be executed in both sequential & parallel example for good information situation & load adjusting (Shafer).

Sprint algorithm is intended to be effectively parallelized, enabling numerous processors to cooperate to construct a solitary steady model. This parallelization shows phenomenal adaptability to the clients. It gives astounding speedup, size up & scale-up characteristics. The blend of these attributes makes Sprint a perfect device for information mining. It utilizes two information structure i.e. characteristics rundown list & histogram which isn't memory occupant making SPRINT appropriate for huge informational collections, in this way it evacuates every one of the information memory limitations on the information. It manages both ceaseless & unmitigated properties. In Sprint algorithm, for tree pruning, we utilize the calculation which depends on the MDL rule

Algorithm for SPRINT DT is:

- Split (dataset Si)
- If (all concentrations in S_i are of similar class) at a particular point
- Return;
- For every characteristic A_i do evaluate parts on property A_i;
- Utilize finest split achieved to segment S_i into S_i1 &S_i2;
- Split (Si1);
- Split (Si2);

- First call: segment (Preparation data)

Following are the imperative issues that have fundamental execution recommendations in the tree-development stage:

- Step by step instructions to find split centers that portray root tests.
- Having picked a split point, instructions toward fragmenting the information are passed.

3.3.1 Data structure

• **Attribute list:** SPRINT makes a characteristic list for each characteristic in the information. Records in these rundowns are called quality records. These records comprise of a property estimation for example imprints, evaluation & alternative of the open elective subject & the list of the record from which these qualities are acquired as shown in table 3.1. In the sprint, on the off chance that every one of the information does not get into the main memory, at that point, the property records are kept up it on the secondary memory [10].

The underlying rundown list made from the testing set is related to the base of the classification tree. Then the tree is developed & hubs are part to make new child nodes, the characteristic records having a place with every hub are apportioned & connected with the child nodes.

Table 3.1: Example of attribute list of dataset

Marks	**Grade**	**Rid**
75	Good	0
86	Good	1
75	Good	2
93	Good	3
62	Average	4
56	Average	5
48	Average	6

Option	**Grade**	**Rid**
B	Average	0
C	Average	1
A	Good	2
A	Good	3
A	Good	4
B	Average	5
A	Good	6

- **After Presorting:**

In sprint algorithm , the arrangement of information is needed to identify the split for numeric qualities. It utilizes gini-splitting index for assess split. Sprint just sorts information once toward the start of the tree creation stage by utilizing various data structure. Every node has its own attribute list & to identify the finest split point for a hub, we examine every one of the hub's characteristic records & assess splits dependent on that characteristic. This is well described in table 3.2

Table 3.2 Dataset after applying presorting splits based on that attribute.

Marks	**Grade**	**Rid**
43	Average	6
52	Average	5
65	Average	4
72	Good	0
78	Good	2
83	Good	1
91	Good	3

Option	**Grade**	**Rid**
B	Average	0
C	Average	1
A	Good	2
A	Good	3
A	Good	4
B	Average	5
A	Good	6

Histogram: Histograms are utilized to catch the class distribution of the characteristic records at every hub. Two histograms are related with every DT hub for consistent characteristic. These histograms signified as C_{below} which keep up information that has been handled & C_{above} which keep up information that hasn't been prepared. Clear cut traits additionally have a histogram related with a hub. A count framework histogram consists of the class distribution for every estimation of the given characteristic [10].

Table 3.3 Attribute list after splitting

Marks	Grade	Rid
43	Average	6
52	Average	5
65	Average	4

Attribute list for node 1

Marks	Grade	Rid
72	Good	0
78	Good	2
83	Good	1
91	Good	3

Attribute list for node 2

Option	Grade	Rid
B	Average	0
C	Average	1
B	Average	5

Marks	Grade	Rid
A	V.Good	3
A	V.Good	4
A	V.Good	6
A	V.Good	2

The table 3.3 shows partitioned list & when this process occurs, the sequence of record in the rundown is conserved [10].

Table 3.4: Evaluating continuous split points

Attribute list

Marks	Grade	Rid
43	Average	6
52	Average	5
65	Average	4
72	Good	0
78	Good	2
83	Good	1
91	Good	3

→ Position 0

→ Position 3

→ Position 6

State of class histogram

C_{below} / C_{above}

0	0
4	3

C_{below} / C_{above}

0	3
4	0

C_{below} / C_{above}

4	3
0	0

Table 3.4 shows the schematic for the histogram update. For deciding the partition for the numeric attribute, the histogram C_{below} is set to 0 & C_{above} is set with the class dissemination at that node. Root node - dissemination is acquired during sorting in figure 3.1. Characteristic data are read one by one & C_{below} & C_{above} are informed for every data delivered.

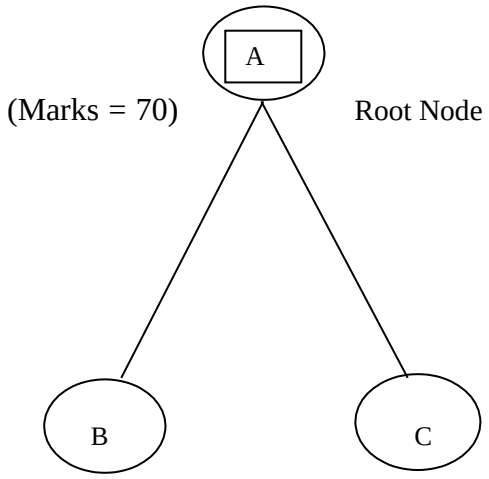

Figure 3.1 DT of data set.

- **Performing the Split:**

When the best split point has been found for a node, we execute the split by creating child nodes & dividing the attribute records between them. We can perform this by splitting the node‘s list into two as shown in figure 3.1. In our example, the attribute used in the winning split point is Marks. After this, we scan the list & apply the split test on it. Then we move the records to two new attribute list i.e. one for each new child. We have no test that we can apply to the attribute values for the remaining attribute lists of the node to decide how to divide the records. To solve this problem, we work with rids (Shafer).

As we partition the list of the splitting attribute i.e. marks, we insert rids of each record into a hash table to notify that the record was moved in which child. We can scan the list of the remaining attributes & probe the hash table after collected rids. The output then tells us with which child to place the record. Splitting process is done in more than one step if the hash table is large for memory. The attribute list is partitioned up to the attribute records for which the hash table will fit in the memory. Only a portion of the attribute lists of non-splitting attributes is partitioned. This process is repeated for the remainder of the attribute list of the splitting attribute.

- **Finding split points**

    - During the process of making a decision tree, the goal at each node is to determine the split point that best divides the dataset belonging to that node. The value of a split point depends upon how well it separates the classes. Many splitting has been proposed in the past to evaluate the goodness of the split. We need some function which can measure which questions provide the most balanced splitting. The information gain metric is such a function.
    - Impurity measure: we have an information table that consists of attributes & set of those attributes, we can gauge similarity or dissimilarity of the table depending on the sets. A framework is unadulterated or similar in the event that it contains just a solitary class. On the off chance that it comprises a few classes, at that point the table

is homogeneous. There are such a large number of records to quantify the level of debasement. Most usual lists are entropy, gin record & misclassification.

$$Entropy = \sum_j (-) p_j \log p_j \qquad (3.1)$$

The entropy of an unadulterated table is 0 in light of the fact that the probability is 1 & log (1) equals zero. Entropy achieves most extreme esteem when all classes in the table have measured up to likelihood. For a data collection S_i

$$Gini\ Index = 1 - \sum_j p_j^2 \qquad (3.2)$$

In the above equation, Pj is the relative recurrence of set j in S_i. On the off chance that a split partitions S_i into two subsets S_i1 & S_i2,where n is the number of data set, the record of the isolated data Gini split(S_i) is given by:

$$\text{Gini split(S)} = {}^{n1}/_{n}\ \text{Gini (S1)} + {}^{n2}/_{n}\ \text{Gini (S2)} \qquad (3.3)$$

Gini list of an unadulterated table comprises of a solitary class is null value on the grounds that the probability is 1 & $1\text{-}1^2 = 0$. Like entropy, Gini list additionally achieves the greatest value when all groups/labels in the table have broken even with likelihood.

$$\text{Classification Error} = 1 - \max\{Pj\} \qquad (3.4)$$

Like "entropy & Gini record" misclassification indices of an unadulterated table is zero in light of the fact that the probability is 1 & 1-max(1) equals zero. The estimation of misclassification is dependably in the vicinity of 0 & 1. For a matter of fact the most extreme gini indices for a given number of sets is constantly equivalent to the greatest of misclassification file on the grounds that for various classes n_i, we set likelihood is equivalent to $p = {}^{1}/_{N}$.

Splitting criteria:

To determine the best attribute for a particular node in the tree we use the measure called information gain. The information gain, gain(S, A) of an attribute A, relative to a collection of examples S, is defined as

$$\text{Gain ratio} = \text{Gain}(S_i, A_i)/\text{Split data} \qquad (3.5)$$

The course of choosing a new attribute & dividing the data set is now iterated for every nonterminal descendant nodes. Characteristics that incorporated higher in the tree are eliminated so that any provided characteristics can occur only once along any path.

Chapter-4

RESULTS & DISCUSSION

The proposed SPRINT DT algorithm is executed in WEKA tool. This tool contains the number of data analysis algorithms , visualization tools & predictive modeling for forecasting the results. Apart from this, it provides a very interactive User Interface (UI) to easily access all its functionalities. In this, the dataset can be imported in multiple formats like Arff, CSV, parallel & it is also capable of reading data from URL or from a database by utilizing SQL. There are different models for classifiers like Naïve Bayes, DTs & so on. We have used classifiers for our experimental reasons. In this, the classify board enables the client to apply grouping SPRINT decision tree & other existing calculations to the informational collection gauge the precision of the subsequent model.

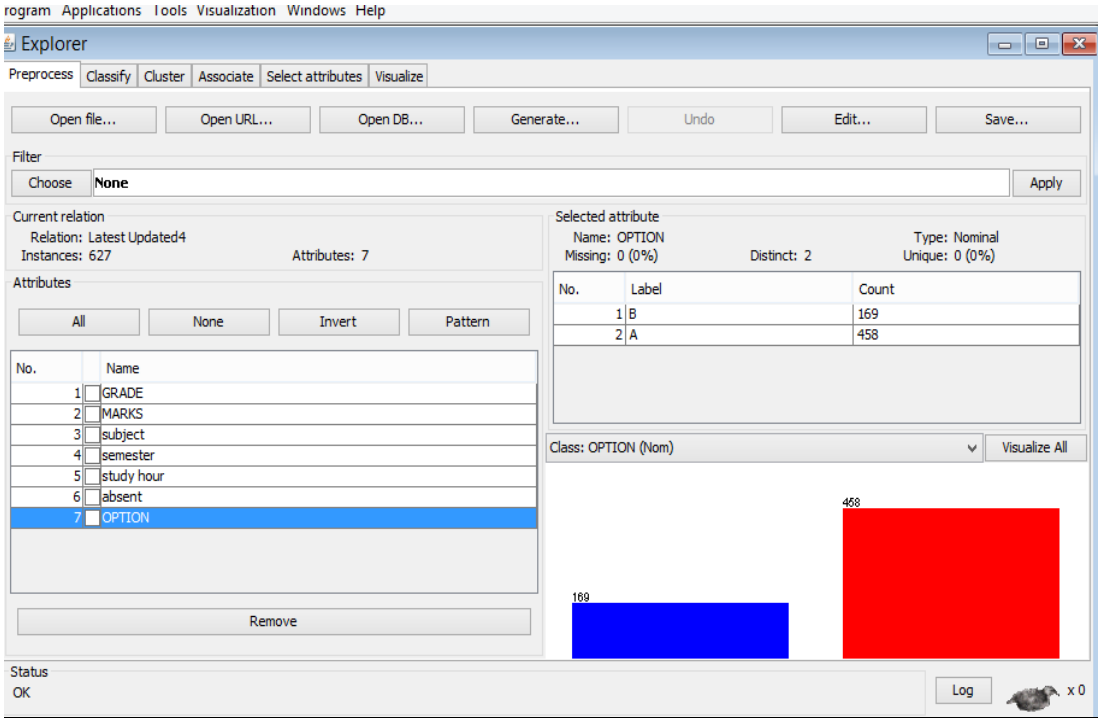

Figure 4.1 Preview after data set imported in Weka

In figure 4.1 Red color implies that these attributes belong to option A, Blue color implies that these attributes belong to option B .

In figure 4.2 the visualization of all the attributes are done . Red color implies to option A and blue implies to option B .The first block shows grades, then followed by marks, subject, semester, study hour, absent and option .

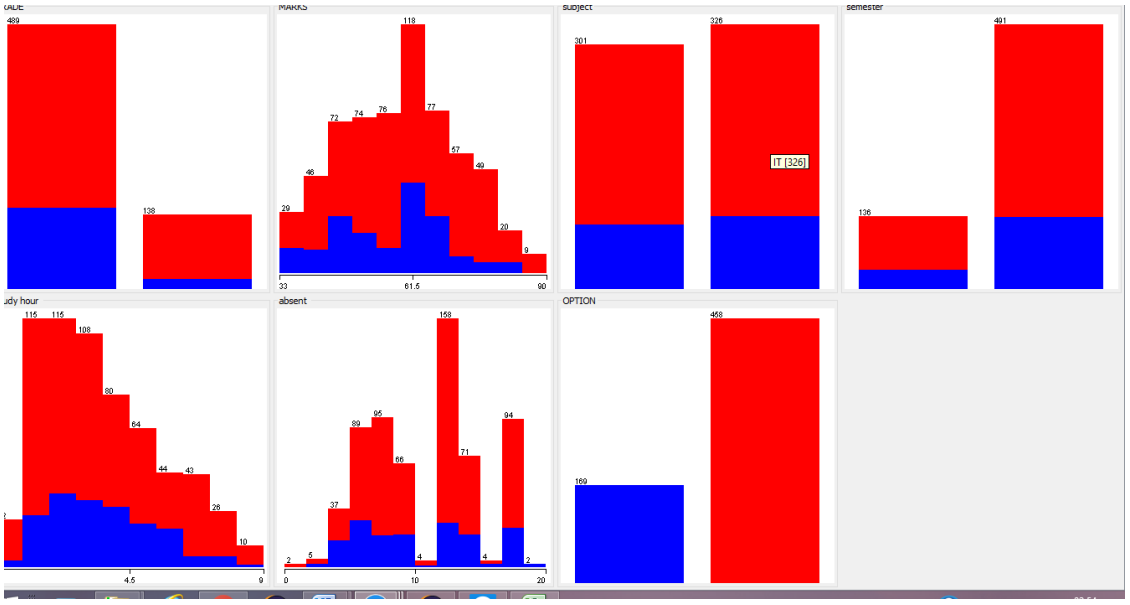

Figure 4.2 Visualizing all Attributes used in Classification

Now after applying all data set, the next step is to apply different classifiers on a given dataset .

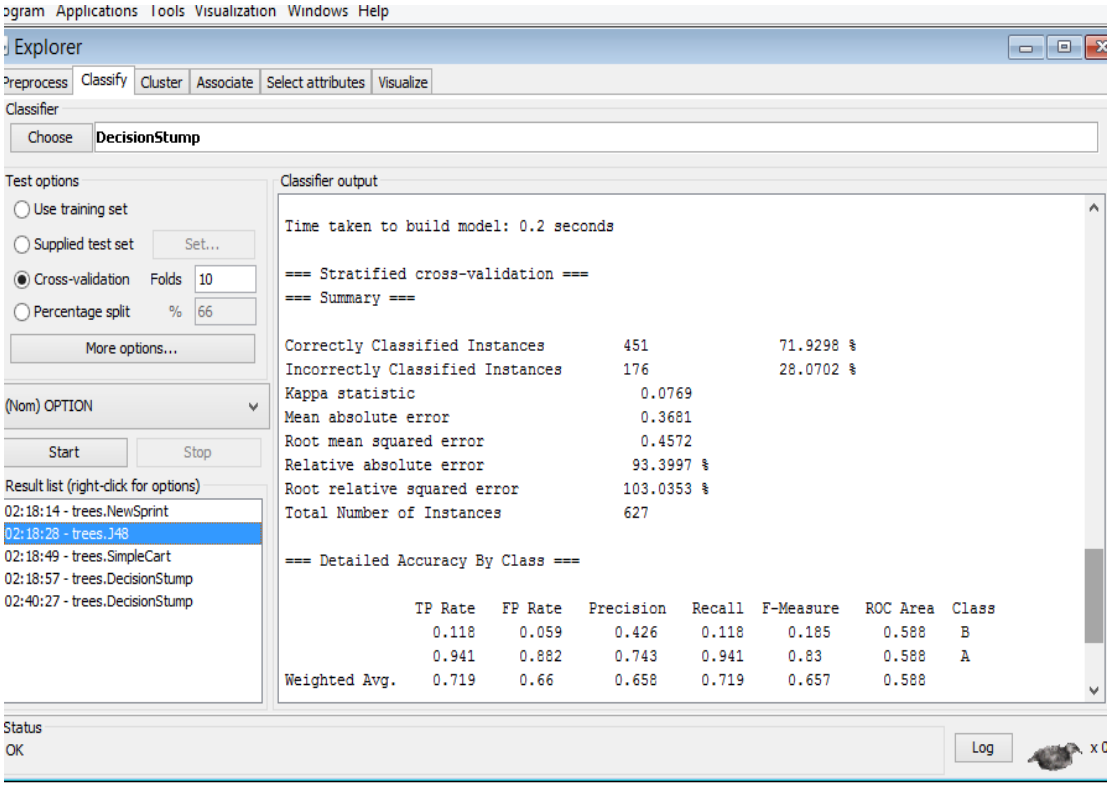

Figure 4.3 Classification by J48 DT

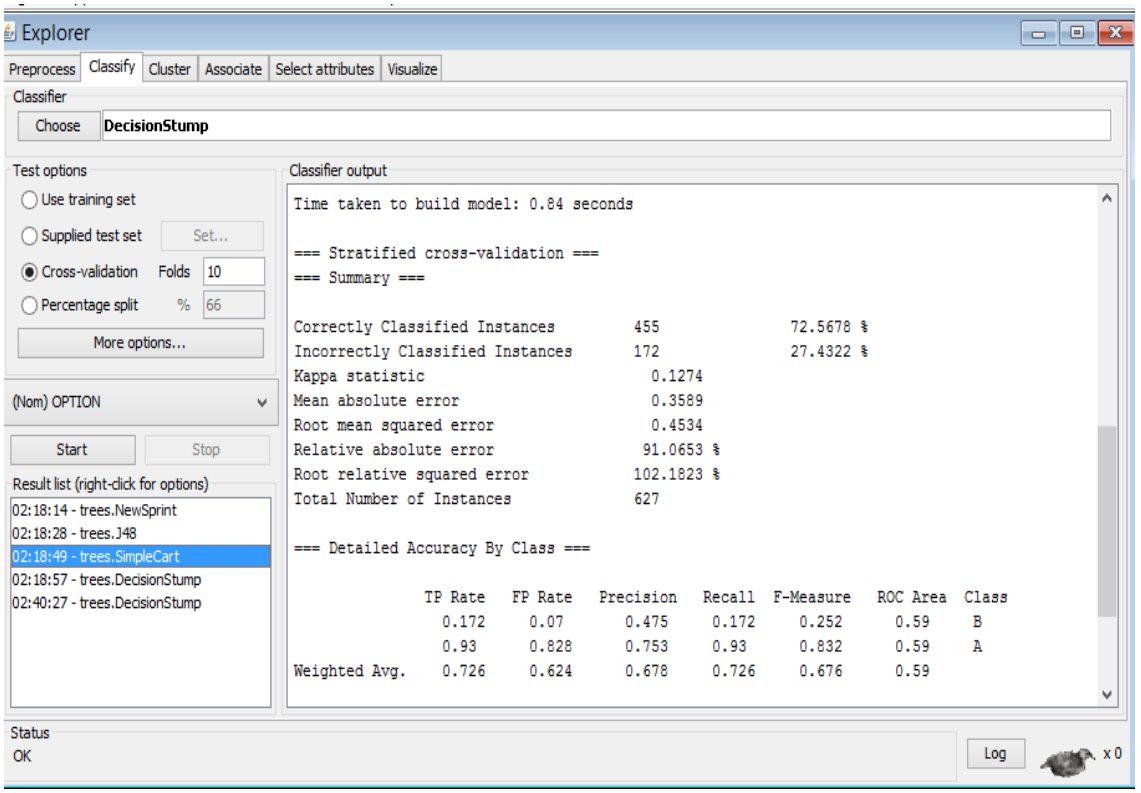

Figure 4.4 Classification by Simple CART

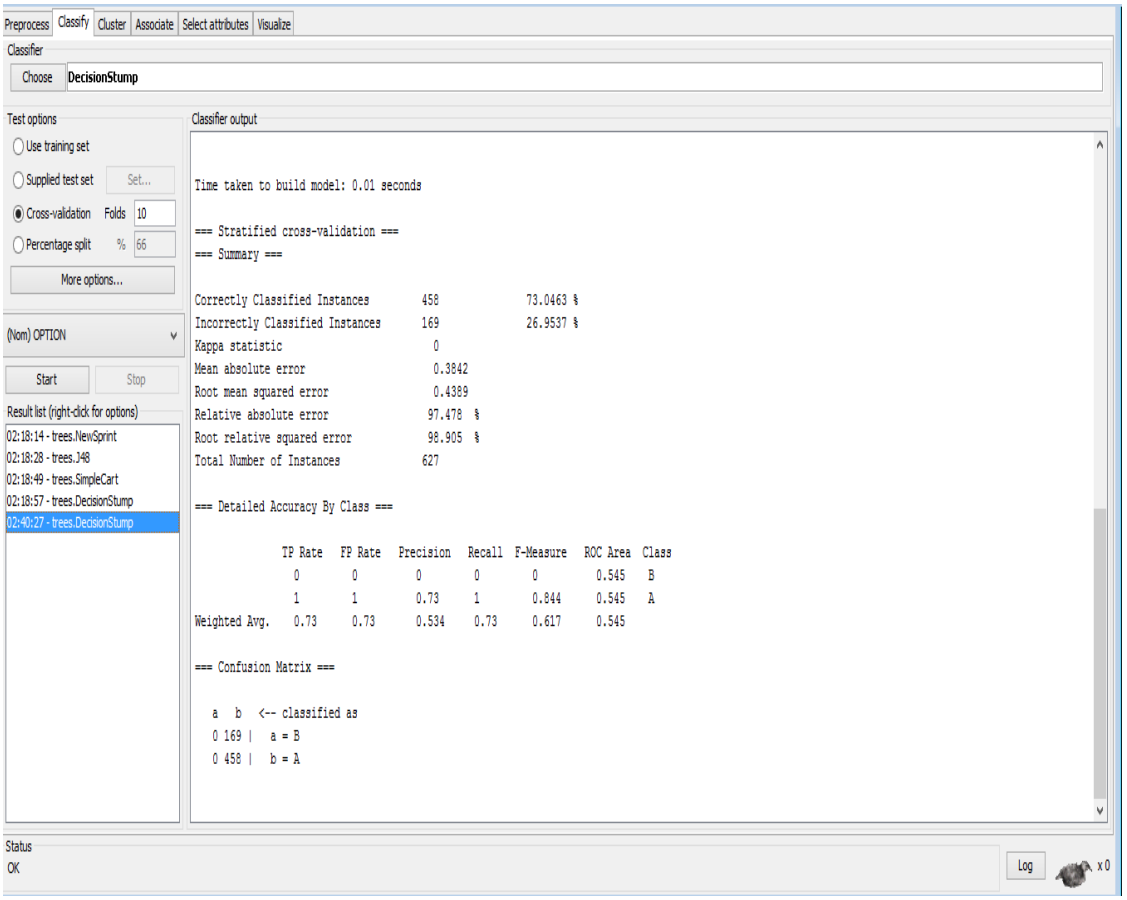

Figure 4.5 Classification by Decision Stump

In figure 4.3, figure 4.4, figure 4.5, a new classifier name as Scalable Parallelizable Induction of Decision Tree (SPRINT) is been applied, which is not used for classification task before and it provides better results than previous classifiers. It is an advancement over J48 DT algorithm because it provides better execution time than J48, Simple CART, Decision Stump(ID3).

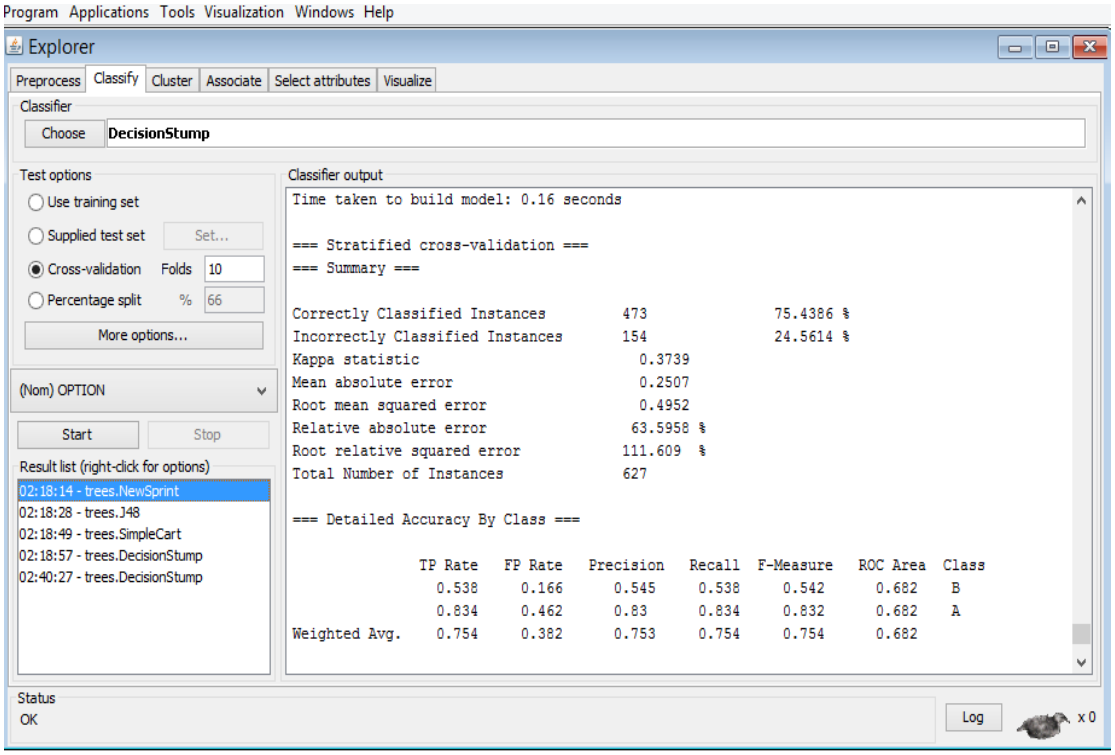

Figure 4.6 Classification by Sprint DT

Figure 4.6 shows the difference between all attributes on parameters such as accuracy, true positive rate, and false-positive rate. The definitions of these terms are explained below:

- **Accuracy:** It's the proportion of an overall number of forecasts that were precise.
- **True Positive Rate:** It's proportion of examples that are classified to be x class, among a total number of examples which contains x class, i.e. how much class part is occupied. It's correspondent to recall.
- **False Positive Rate:** It's the portion of examples which actually contains x class among all classes which are classified as X class.
- **F-measure:** It's a combined measure for recall & precision. It can be calculated by the following formula:

$$F_{measure} = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4.1)$$

4.1 COMPARISION

The table 4.1 represents the comparison among working of different decision algorithm on the basis of different parameters.

Table 4.1 Parameter Comparison of DT algorithms

ALGORITHMS	**ID3 & C4.5**	**CART**	**SPRINT**
Measure	Entropy information gain	Gini diversity index	Gini Index
Procedure	Construction of Top-down decision tree	Constructs binary DT	Breadth-first manner based decision tree construction
Pruning	Pre-pruning using a single-pass algorithm	Post pruning based on cost-complexity measure	Post pruning based on MDL principle

4.2 OUTPUT

The DT's as instances of forecast prototypes attained from the data index of 600 students by these ML algorithm: CART DT algorithm,J48 and Decision stump. Spint tree is dense as shown in figure 4.7 we have a large amount of data and the window size of Weka is small bydefault. We are classifying accuracy of correctly classified instance and execution time of data set through sprint tree.

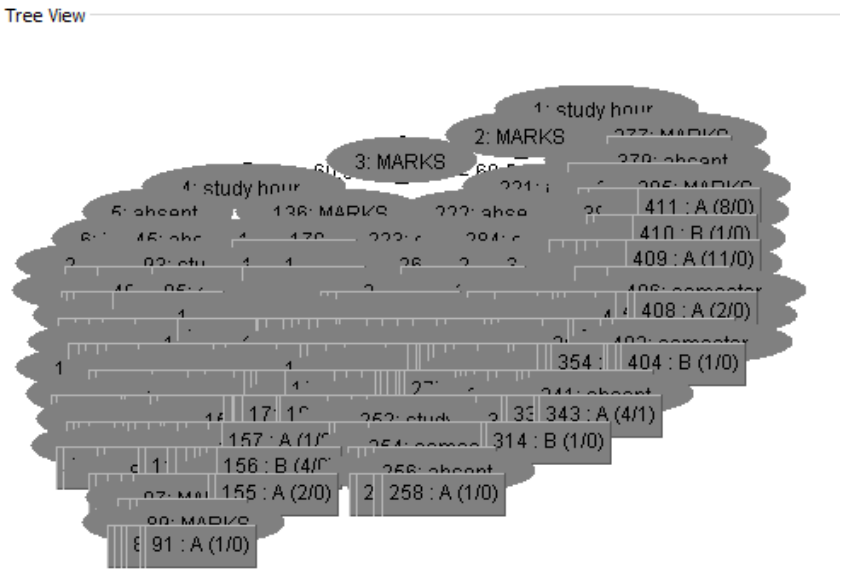

Figure 4.7 Sprint DT

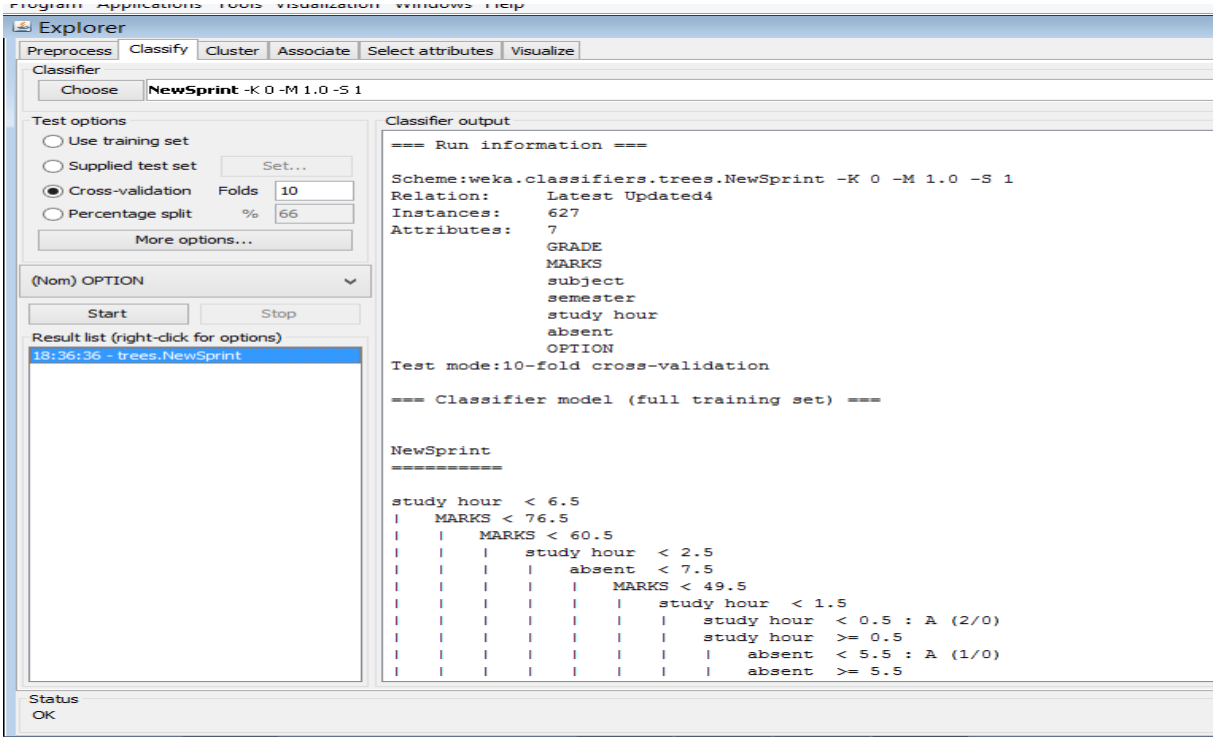

Figure 4.8.1 Sprint decision tree (classifier model)

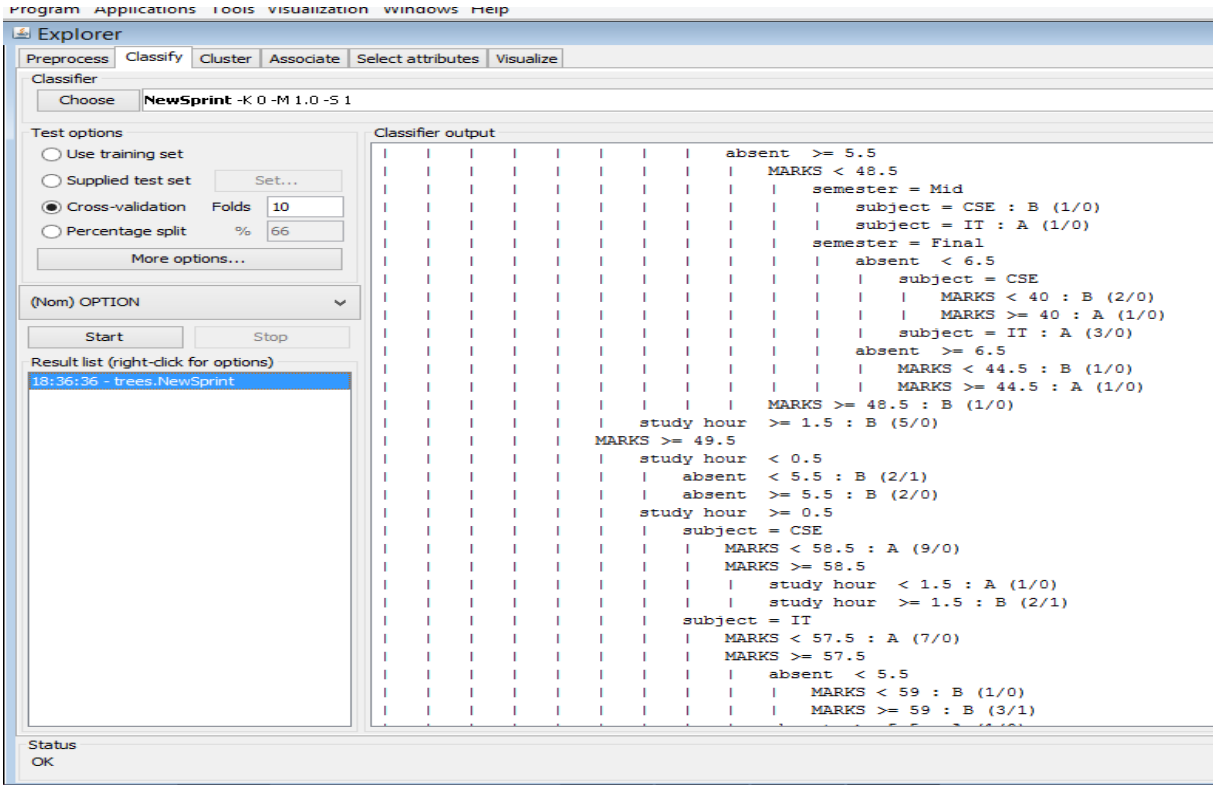

Figure 4.8.2 Sprint decision tree (classifier model)

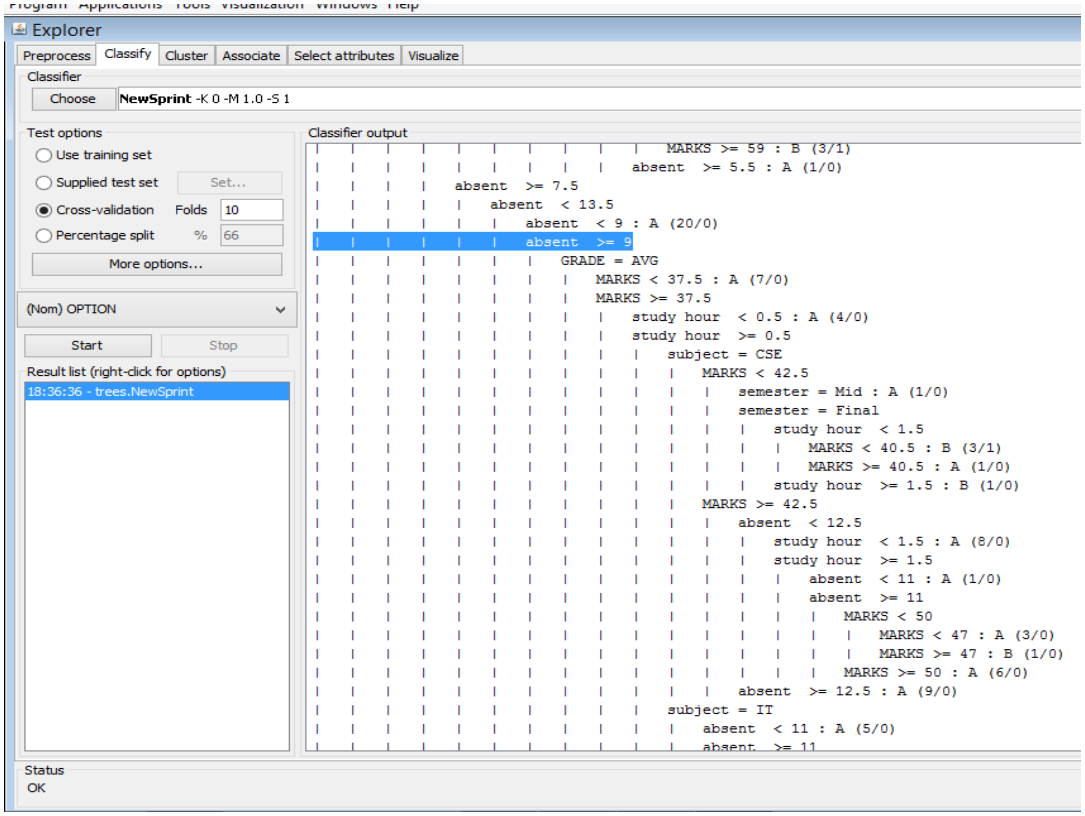

Figure 4.8.3 Sprint decision tree (classifier model)

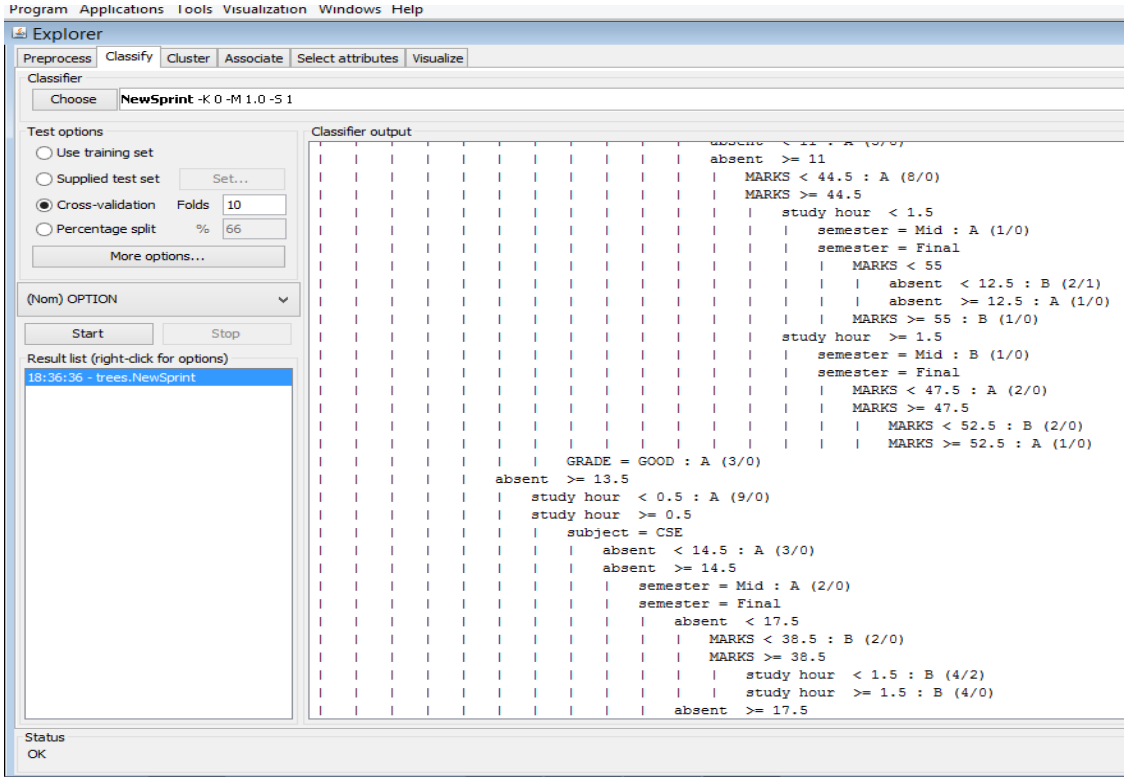

Figure 4.8.4 Sprint decision tree (classifier model)

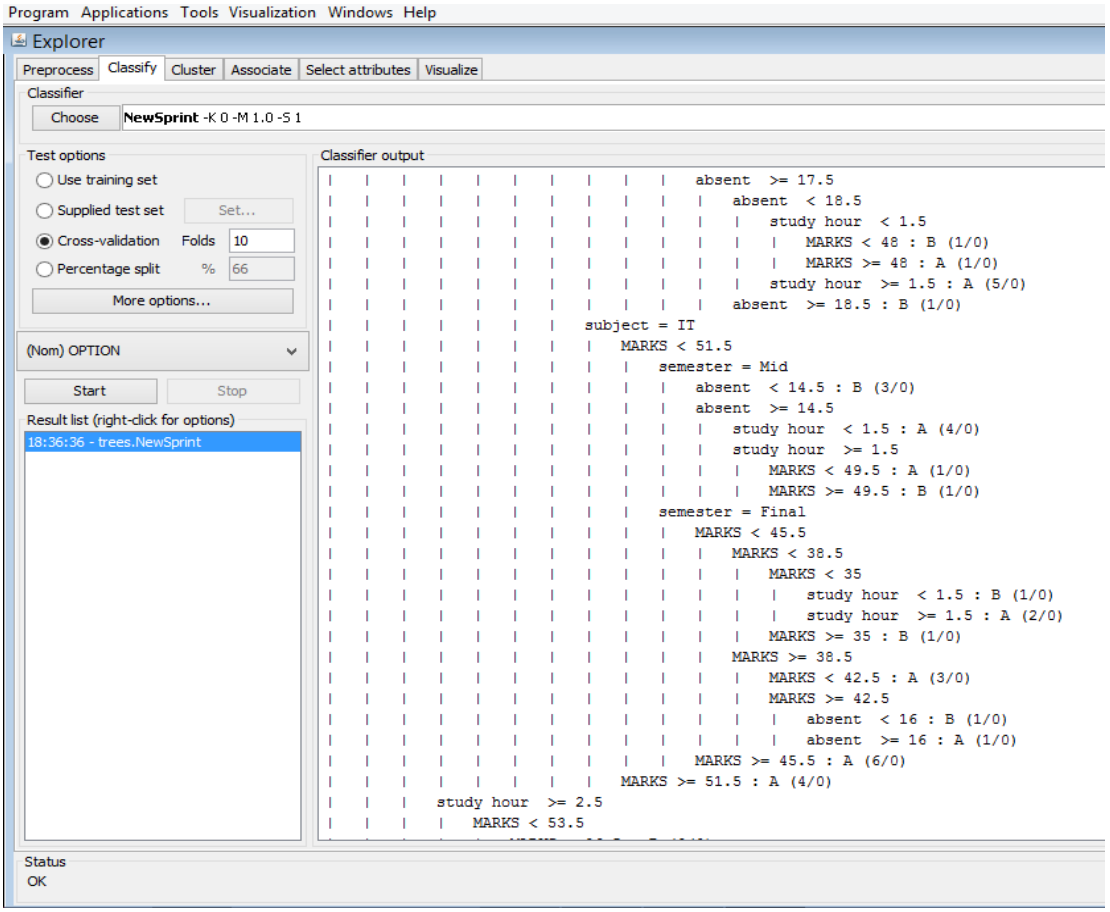

Figure 4.8.5 Sprint decision tree (classifier model)

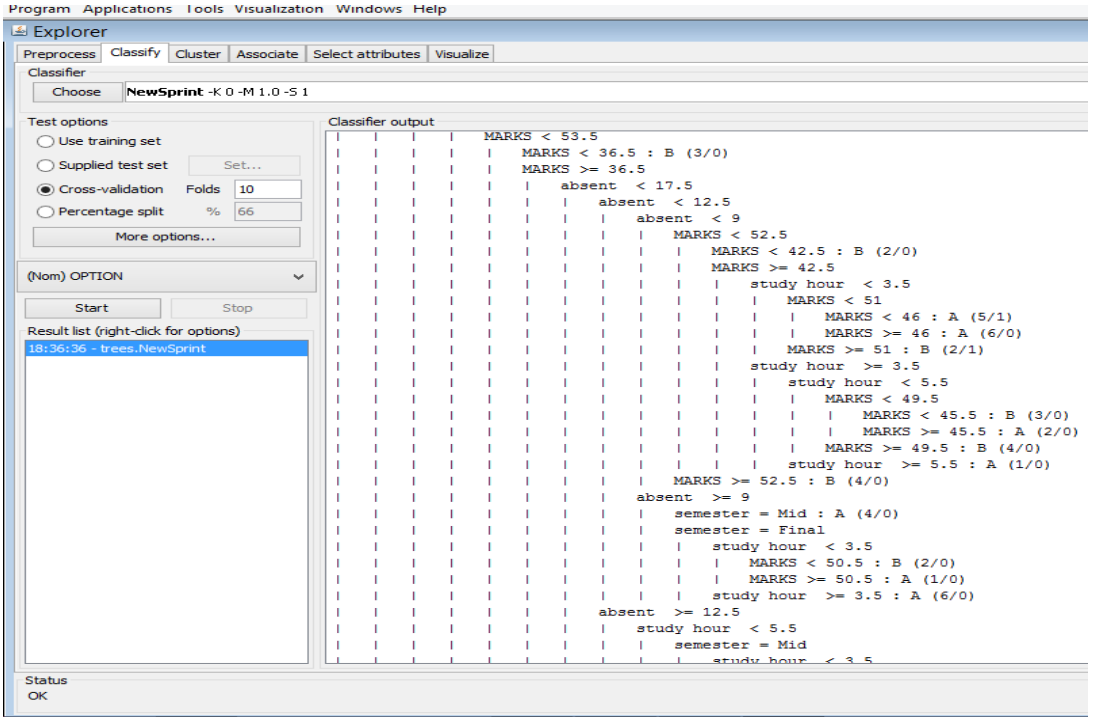

Figure 4.8.6 Sprint decision tree (classifier model)

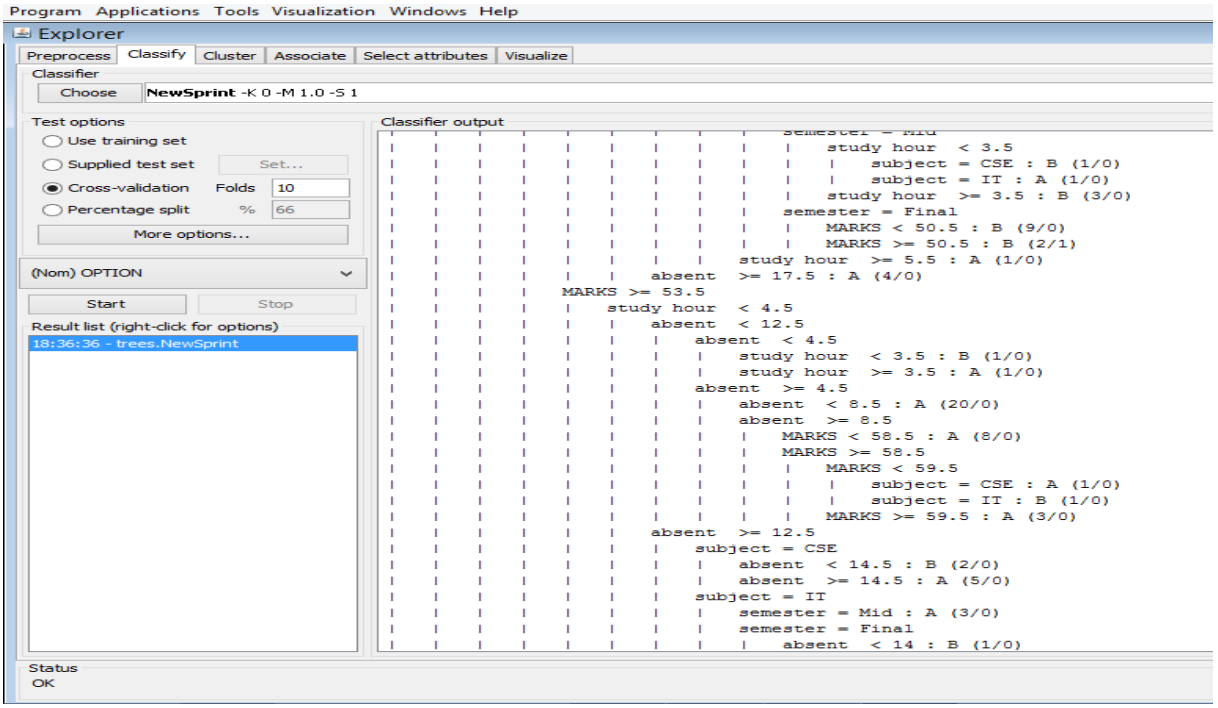

Figure 4.8.7 Sprint decision tree (classifier model)

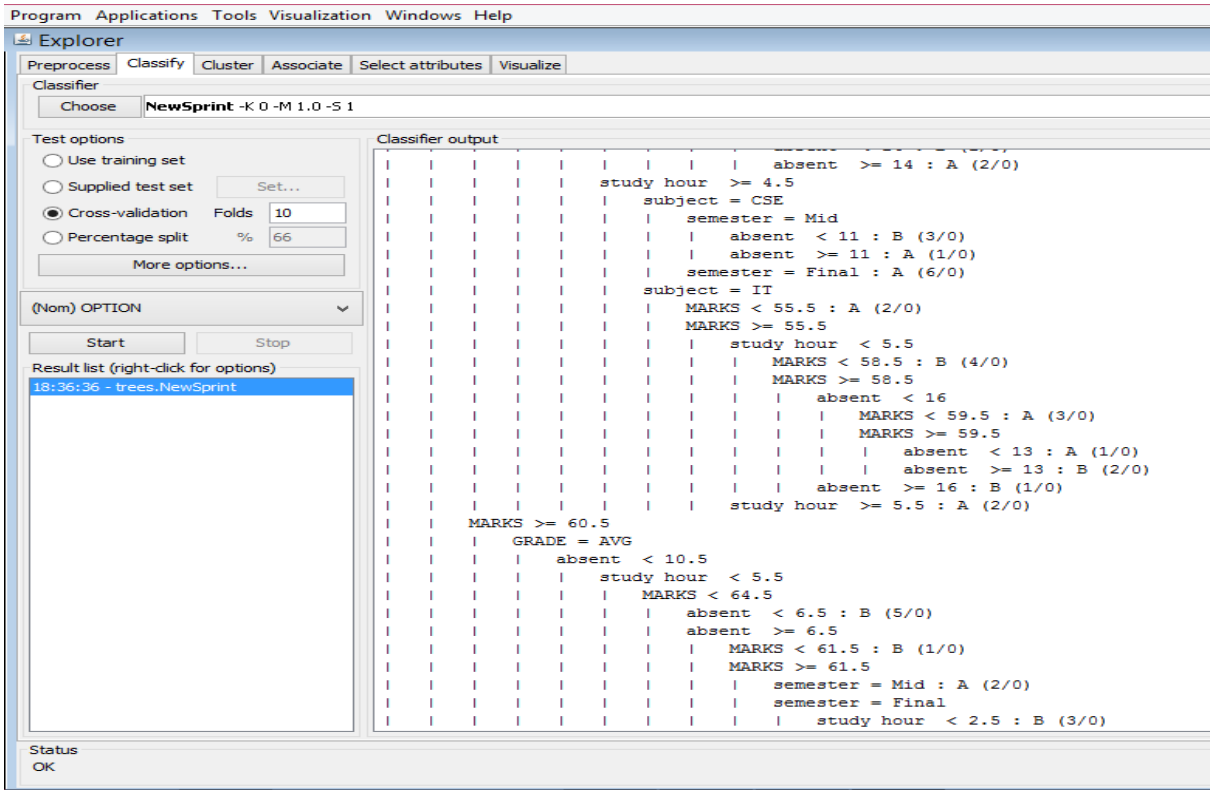

Figure 4.8.8 Sprint decision tree (classifier model)

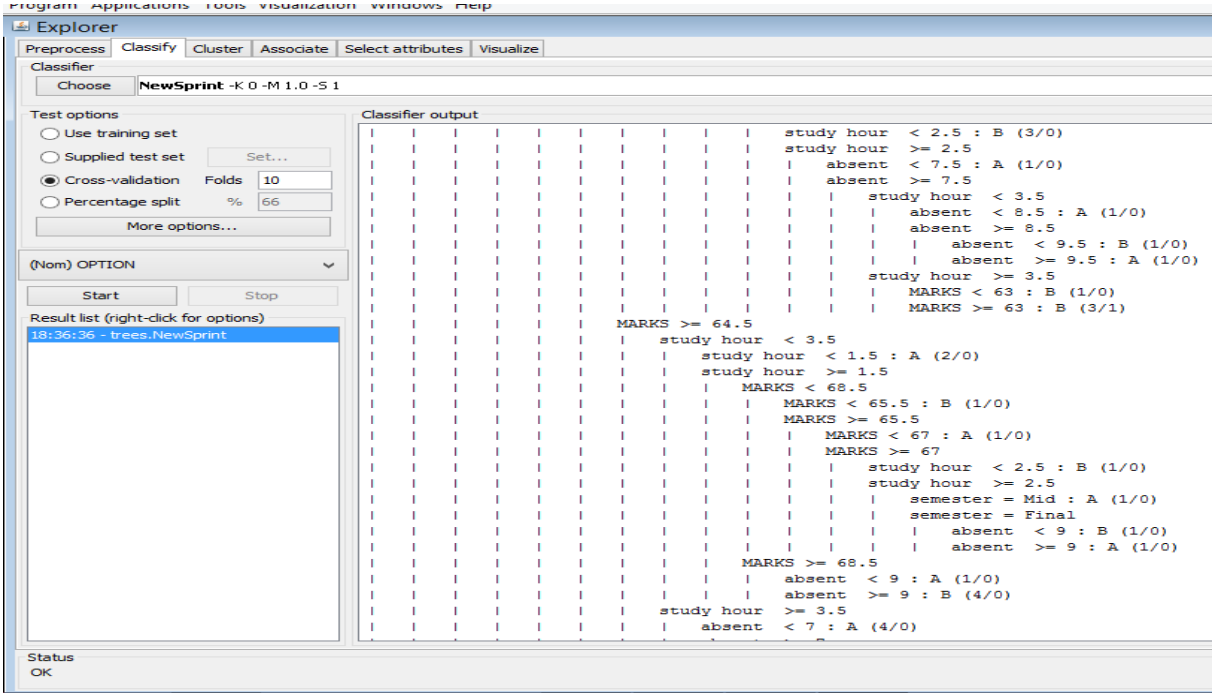

Figure 4.8.9 Sprint decision tree (classifier model)

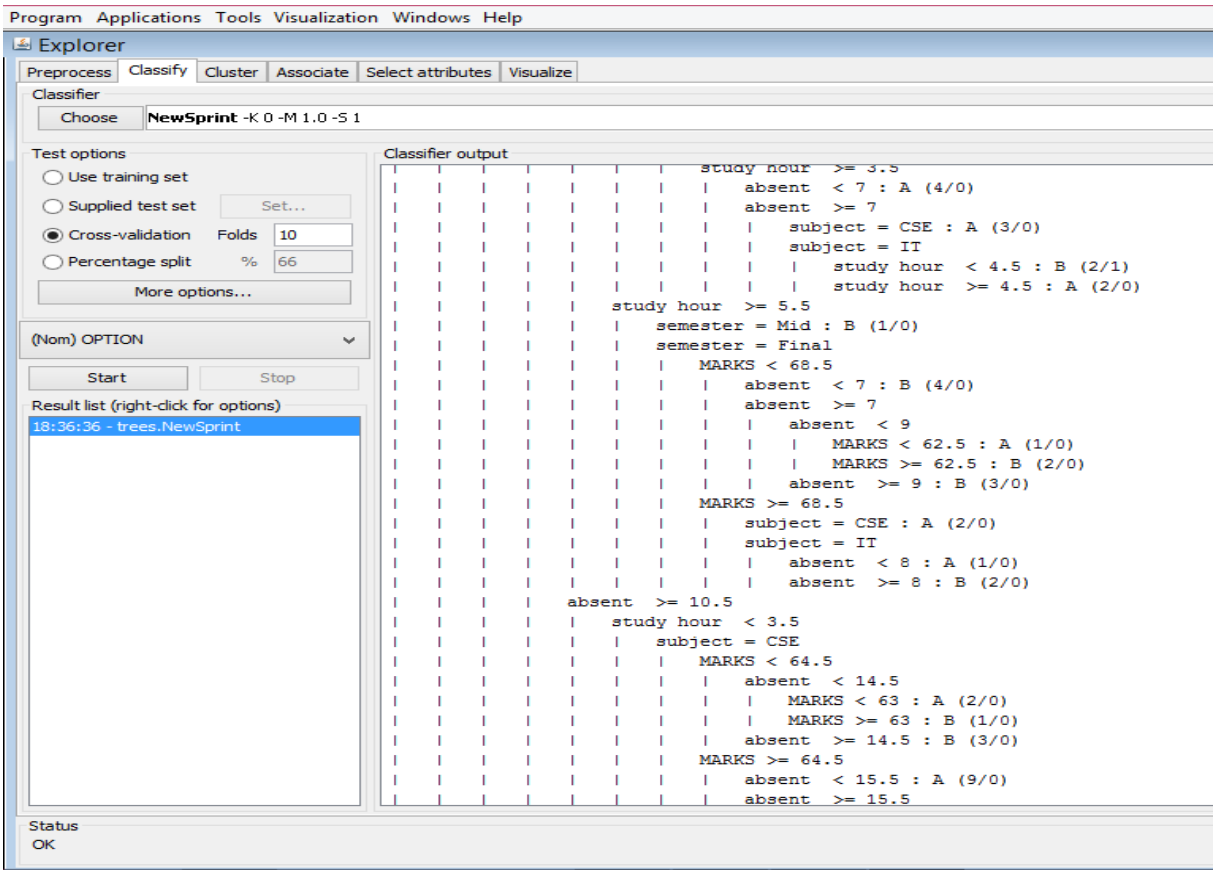

Figure 4.8.10 Sprint decision tree (classifier model)

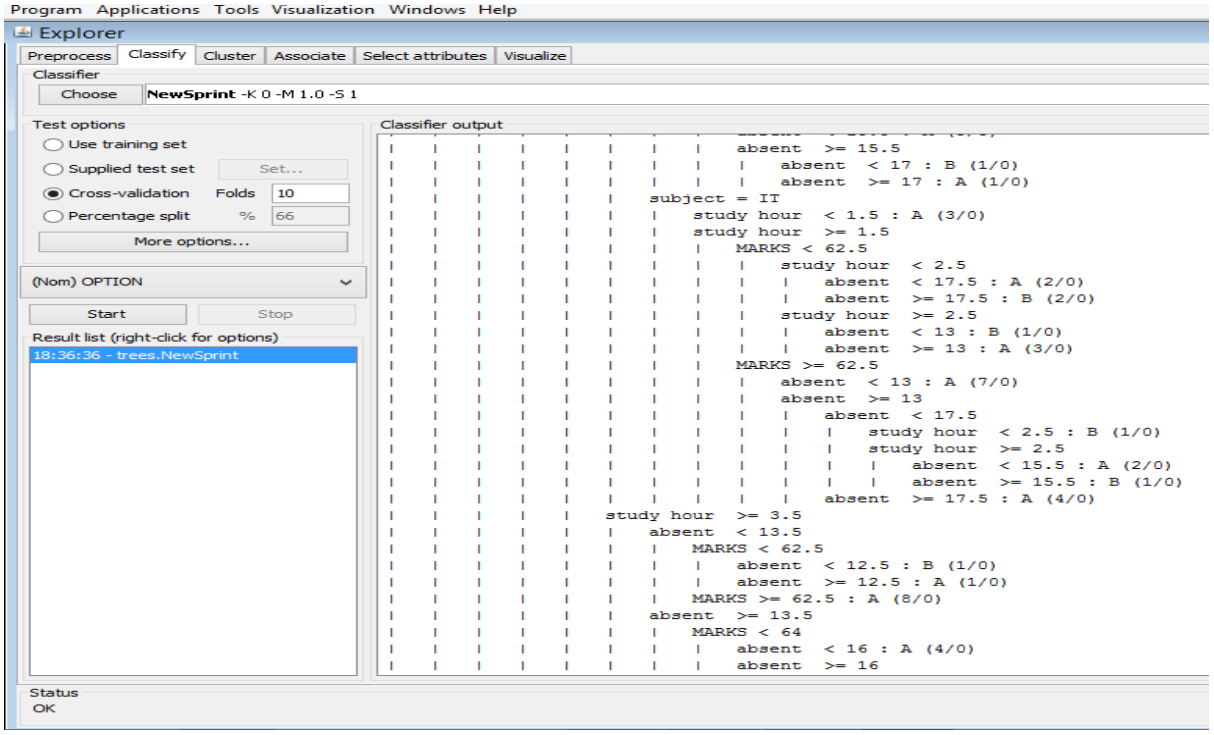

Figure 4.8.11 Sprint decision tree (classifier model)

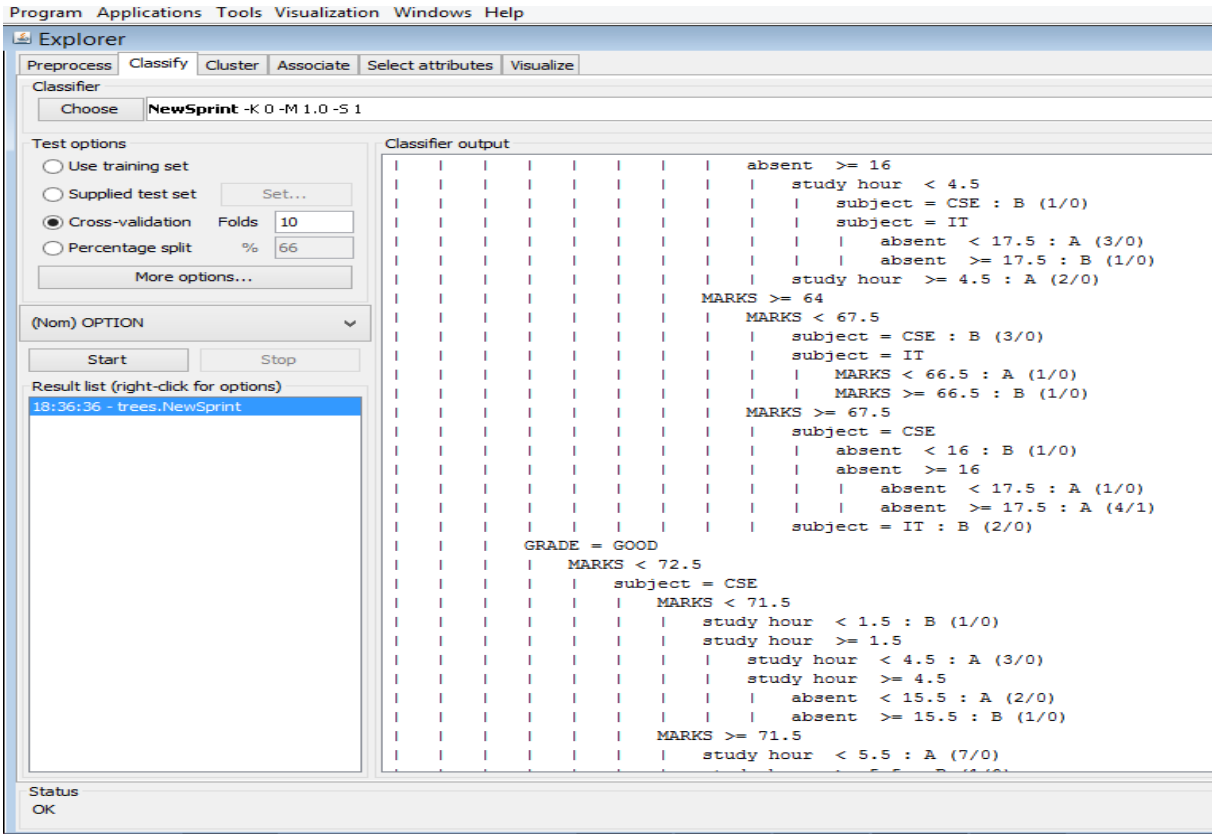

Figure 4.8.12 Sprint decision tree (classifier model)

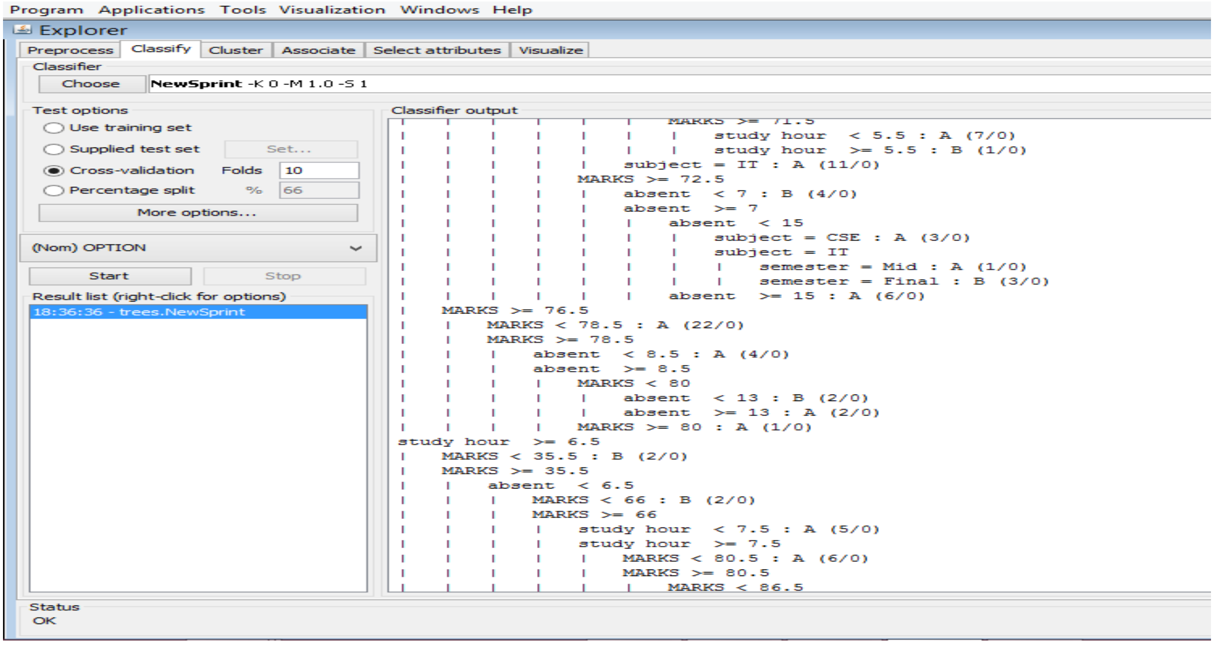

Figure 4.8.13 Sprint decision tree (classifier model)

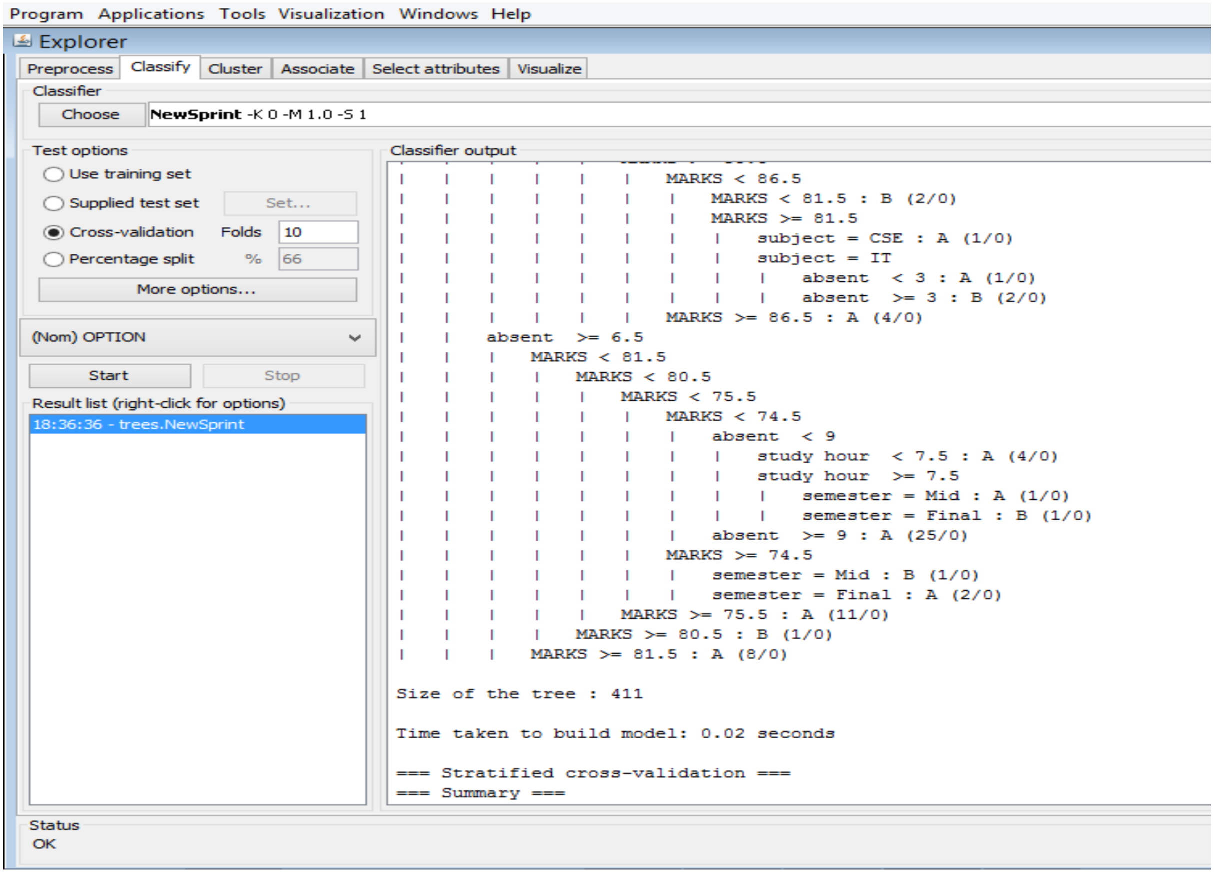

Figure 4.8.14 Sprint decision tree (classifier model)

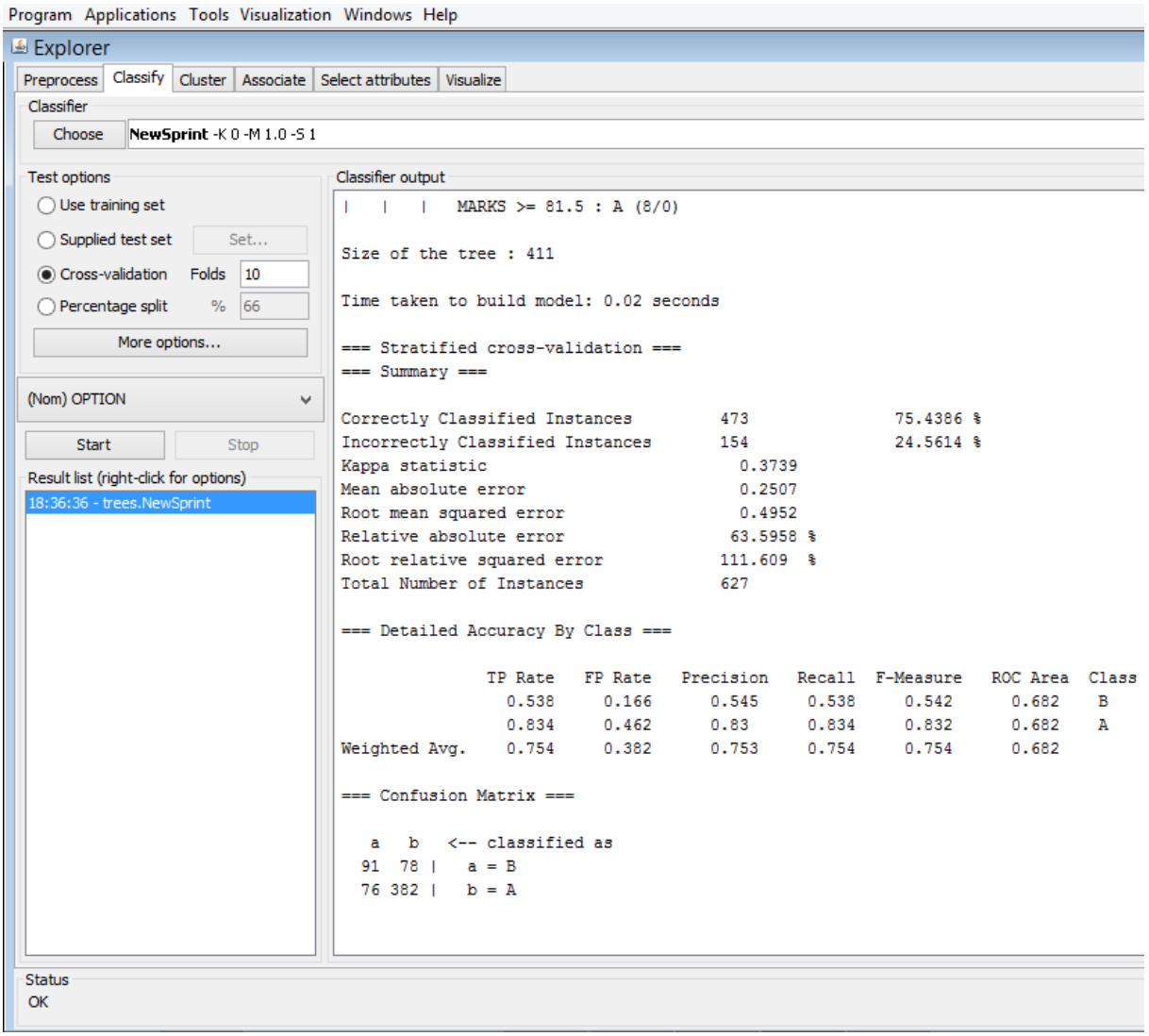

Figure 4.8.15 Sprint decision tree (classifier model)

The figure 4.8 shows the preprocessing of dataset by Sprint Algorithm. On the basis of these calculations the final model is build.

4.2.1 Classifier Accuracy

Table 4.2 demonstrates the reenactment aftereffect of every algorithm. From this table, we can see that a Sprint algorithm has the most superior efficiency of 75.4386 % contrasted with different algorithm. It additionally demonstrates the time multifaceted nature in seconds of different classifiers to assemble the model for preparing data. By this test correlation, plainly Sprint is the best algorithm among four as it is increasingly precise & less tedious.

Table 4.2 Classifiers Accuracy

Algorithm	**Correctly classified instances**	**Incorrectly classified instances**	**Execution Time (sec)**	**Mean absolute error**	**RMS error**	**Relative absolute error**
SPRINT	75.4386%	24.5614%	0.16	0.2507%	0.4962%	63.5958%
Simple CART	72.5678%	27.4322%	0.84	0.3589%	0.4534%	91.0653%
Decision Stump (id3)	73.0463%	26.9537%	0.18	0.3842%	0.4389%	97.478%
J48	71.9298%	28.0702%	0.2	0.3681%	0.4572%	93.3997%

Chapter -5

CONCLUSION & FUTURE SCOPE

5.1 Conclusion

The productivity of all the DT algorithm can be broke down dependent on their precision & time taken to infer the tree. Using our predictive model we came to the result that those students who have opted their subject themselves have a good grade as compared to those who opted subject by their parents or friends. The principal weaknesses of sequential DT algorithm (ID3, C4.5 & CART) are low arrangement precision when the preparation data is huge. This issue is explained by SPRINT DT algorithm. Run evacuates all the memory limitation & exactness issue which comes in other existing algorithm. It is quick & adaptable than others since it tends to execute in both sequential & parallel style for good data situation & burden adjusting.

In this work, the SPRINT DT algorithm has been connected on the dataset of 600 students for foreseeing their exhibition in the test based on their decision in surveying framework. We used this project for more number of students & its outcome will not change as per our new improved algorithm. This outcome helps us to find that the students who are selected their own.

5.2 Future Scope

The outcomes of this thesis work can be used for further improvement in the future like:

- We will be able to discover out the performance of a student on the ground of their decision of subject also.
- Parallel implementation of sprint environment can be used to expanding the searchspace.

REFERENCES

[1] Wei Zhang, Shiming Qin, "A brief analysis of the key technologies & applications of educational data mining on online learning platform", Big Data Analysis (ICBDA) 2018 IEEE 3rd International Conference on, pp. 83-86, 2018

[2] Shaojie Qu, Kan Li, Shuhui Zhang, Yongchao Wang, "Predicting Achievement of Students in Smart Campus", Access IEEE, vol. 6, pp. 60264-60273, 2018

[3] British Kumar Bhardwaj & Saurabh Pal "Data mining: a prediction for performance improvement using classification", International journal of computer science & data security, vol. 9, no. 4, 2011.

[4] Md. Hasibur Rahman, Md. Rabiul Islam, "Predict Student's Academic Performance & Evaluate the Impact of Different Attributes on the Performance Using Data Mining Techniques", Electrical & Electronic Engineering (ICEEE) 2017 2nd International Conference on, pp. 1-4, 2017

[5] Arshad Ahmad, Kan Li, Chong Feng, Syed Mohammad Asim, Abdallah Yousif, Shi Ge, "An Empirical Study of Investigating Mobile Applications Development Challenges", Access IEEE, vol. 6, pp. 17711-17728, 2018

[6] Marcell Nagy, Roland Molontay, "Predicting Dropout in Higher Education Based on Secondary School Performance", Intelligent Engineering Systems (INES) 2018 IEEE 22nd International Conference on, pp. 000389-000394, 2018

[7] C.Romero & S.Ventra "Educational data mining: A survey from 1995 to 2005", 2006 Elsevier ltd.

[8] Dorina kabakchieva," Student performance prediction by using data mining classification algorithm rithms", International journal of computer science & management research, Vol 1 issue 4 November 2012

[9] Devi Prasad Shukla & S. Ramachandram, "Decision tree induction- An Approach for data classification using AVL –Tree", International Journal of computer & electrical engineering, Vol. 2, no. 4, August 2010.

[10] John Shafer, Rakesh Agrawal, Manish Mehta "SPRINT: A scalable parallel classifier for data mining" IBM Almaden Research Center, 650 Harry road, San Jose, CA 95120.

[11] Jorma Rissanen, Rakesh Agrawal, Manish Mehta "SLIQ: A scalable parallel classifier for data mining" IBM Almaden Research Center, 650 Harry road, San Jose, CA 95120.

[12] Johannes Gehrke & Raghu Ramakrishna, "Rainforest- A Framework for fast decision tree construction of large data sets", Proceedings of the 24th VLDB conference, New York, USA

[13] Matthew N. Anyanwu & Sajjan G.shiva "Comparative analysis of serial decision tree classification algorithm rithms", International journal of computer science & security, (IJCSS) Volume 3: Issue (3).

[14] M. Sukanya, S. Biruntha, Dr. S. Karthik & T.Kalaikumaran "Data mining: Performance Improvement in Education Sector using Classification & Clustering Algorithm rithm", International conference on computing & control engineering (ICCCE 2012) 12 & 13 April 2012.

[15] R.R.Kabra & R.S.Bichkar," Performance prediction of engineering students using decision tree", International Journal of computer applications (0975-8887), volume 36-no, December 2011

[16] Rajeev Rastogi & Kyuseok shim, "PUBLIC: A Decision Tree Classifiers that integrates building & pruning", data mining & knowledge discovery,4, 315-344, 2000.

[17] Surjeet Kumar Yadav & Saurabh Pal "Data mining: a prediction for performance improvement of engineering students using classification", World of science & data technology journal (WCSIT) ISSN: 2221-0741, Vol 2, no. 2, 2012.

[18] S.Anupama Kumar & Dr. Vijayalakshmi M.N. (2011) "Efficiency of decision trees in predicting student's academic performance", D.C. Wyld, et al. (Eds): CCSEA 2011, CS & IT 02, pp. 335-343, 2011.

[19] Saurabh pal & Brijesh Kumar Baradwaj "Mining educational data to analyze students performance", (IJACSA) International Journal of Advanced computer science & applications. Vol. 2 no. 6,2011

[20] Shaeela Ayesha, Tasleem Mustafa, M.Inayat Khan & Ahsan Raza Sattar "Data mining model for higher education system", European journal of scientific research, ISSN 1450-216X Vol. 43 no. 1(2010), pp.24-29.EuroJournals Publishing, inc. 2010.

[21] Sunita Bahre, Mr. LOBO L.M.R.J " Data mining in an educational system using weka tool", International conference on emerging technology trends (ICETT) 2011.

[22] Smitha T & V. Sundaram, "Comparative study of data mining algorithm rithms for high dimensional data analysis" International Journal of advances in engineering & technology, Sept. 2012.

[23] Tarun Verma, Sweety raj, Mohammad Asif Khan, Palak Modi, "Literacy Rate Analysis", International journal of science & engineering research volume 3, issue 7, July-2012, ISSN 2229-5518.

[24] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda ,Geoffrey J. McLachlan, Angus Ng, Bing Liu,Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand ,Dan Steinberg, "Top 10 algorithm rithms in data mining", Springer- Verlag London limited 2007

APPENDIX

GLOSSARY

WEKA TOOL:

It is a tool, which has a collection of machine learning algorithms for data mining. It stands for the Waikato Environment for Knowledge Analysis. It is basically an open-source & this tool is programmed in Java language. In this data can be imported in any format like CSV, Arff, binary, etc. data can also read from URL or database using SQL. There are preprocessing tools which are known as filters. They are used to remove noise or disturbance from data. There are various models for classifiers like Naïve Bayes, Decision Trees, etc. We have used classifiers for our experiment purpose.

ECLIPSE JUNO IDE:

We haven't run Weka directly with a GUI instead we have added all its Jar files in Eclipse IDE & then we have applied different algorithm rithms & our data set after running Weka main class in Eclipse.

The main advantage of this software is that we can add our own algorithm rithm with the help of Eclipse IDE in the Weka tool & analysis the output of that particular algorithm rithm so

that we can enhance its output. The other advantage of using it through Eclipse that it reduces dependency on already built classifiers in the Weka tool.

The next step in the classification process is the addition of data set in Weka. The Steps are as follows:

- Click the Programs tab, then on Explorer tab.
- Now click on the first tab in the explorer tab i.e. Preprocess Tab.
- Third & last step is to click on the Open File button & import data set where data saved in the system.

LIST OF ABBREVIATIONS

- SLIQ- Supervised learning in ques
- SPRINT- Scalable parallelizable induction of decision tree algorithm rithm
- ID3- Iterative dichotomiser 3
- CART- Classification & regression trees
- PUBLIC- Pruning & Building Integrated in classification
- OE- Open Elective
- CSE – Computer Science and engineering
- IT- Information Technology
- DT- Decision Tree

ORIGINALITY REPORT

11%	9%	5%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	inpressco.com Internet Source	2%
2	www.cs.sfu.ca Internet Source	2%
3	Submitted to Banaras Hindu University Student Paper	2%
4	docplayer.net Internet Source	1%
5	Monelli Ayyavaraiah. "Analysis of a Data Mining Based Student Data Collection", 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019 Publication	1%
6	imtuoradea.ro Internet Source	1%
7	cirworld.com Internet Source	1%
	mientayvn.com	

8	Internet Source	<1%
9	arxiv.org Internet Source	<1%
10	Submitted to Eastern Mediterranean University Student Paper	<1%
11	www.lido.dist.unige.it Internet Source	<1%
12	Submitted to Staffordshire University Student Paper	<1%
13	Submitted to UT, Dallas Student Paper	<1%
14	itlab.uta.edu Internet Source	<1%

Exclude quotes	On	Exclude matches	< 20 words
Exclude bibliography	On		