Analysis and Implementation of Data Mining Algorithms for Deploying ID3, CHAID and Naive Bayes for Random Dataset

Amit Verma¹*, Asha Gill¹ and Iqbaldeep Kaur²

¹Department of Computer Science and Engineering, CGC Landran, Mohali - 140307, Punjab, India; dramitverma.cu@gmail.com, iqbaldeepkaur.cu@gmail.com
²Department of Computer Science and Engineering, Chandigarh University, Mohali - 140413, Punjab, India; er.ashagill@gmail.com

Abstract

Objectives: Effective data processing for fast retrieval of information has become a burning issue. A modern document contains not only text but images, video, audio as well. In this paper, a brief history of storage devices from the Vedic period to the world of digitization with some important inventions has been presented. Method/Statistical Analysis: It also includes the discussion on how data is transformed for the decision making process along with preprocessing techniques. Findings: A comparative analysis has been done of various techniques, their specific algorithms, uses, pros, limitations and applications where these can be implemented. It helps to give us an insight about these techniques. Finally experimental results on three different algorithms (Id3, CHAID, Naive Bayes) using Rapid Miner have been evaluated to compare their performance based on three parameters (accuracy, precision and recall). Applications/Improvements: The empirical results show the ID3 as more accurate than others with 95.95% accuracy while CHAID shows 89.11% and Naive Bayes classified 81.77% data accurately.

Keywords: CHAID, Development of Storage Devices, ID3, Information Retrieval, Rapid Miner, Retrieval Techniques, Visualization

1. Introduction

In the 21st century the whole world is connected though the world wide web and there is data and data everywhere just like blood in veins. A real instance of this is the mobile phone which is a carrier and storehouse of huge data including messages, images, audio, video. On social media a lot of data is being uploaded and the challenge is how to manage it. Data plays a very significant role because data comprises every thing. In traditional time, people used to communicate by gestures and drawing symbols on the rocks. They used to have bounded resources to express their views, but in the era of digitization there are numerous ways to symbolize the data. However, the associated overhead and complexities have also increased to obtain the right information. We can obtain knowledge and information from data but the associated unstructured raw facts need to be related, processed and organized precisely to compute with accuracy and relevance. Several storage devices were conceived according to the need of storage of data and extaction of data. From past records to future trends, the process

*Author for correspondence
to extract, transform and storing of data has undergone a massive change. In Section 2, a discussion has been presented on the evolution of the data storage and inventions associated with it. Section 3 introduces basic operations performed for preprocessing like tokenization, stemming, stop word removal. These steps are related to the cleaning or removal of noisy data. Section 4 is a review of comparative analysis of different techniques used for data mining along with their benefits, confines and uses. The empirical results and discussions have been reported in Section 5. Three different algorithms (ID3, CHAID and Naive Bayes) have been compared on the basis of precision, recall and accuracy using Rapid Miner. Also, an outline on confusion matrix with some vital terms has been presented for the better understanding of experiment and the paper ends with the conclusion under Section 6.

2. History Of Storage Devices

The term data is not new in 21st century; only the way of representation and accessing is different now a days. In yester years, the fact that data was limited had made the retrieval of data a simple process. However in today's time when the whole world has become a well-connected global village generating lots and lots of data with every passing minute, the collection and retrieval of data has become the biggest challenge. Moreover, in today's competitive world of information, data is considered as the most precious resource pertaining to potential advantages for the business and corporate world.

2.1 Development of Storage Devices in 1915-1940

In the epic of Vedic, common methods used for storage and accessibility of data are either by making lists on papers or by keeping it in human memory. The Table 1 shows a detail on invention of various mechanical devices used for storage of information. More details of Table 1 can be found in appendix. In the beginning of world of technology the data is stored or kept into different devices these developments started from perforated stencils which were used in the period of Alexandria.

| Year    | Developer | Inventions In Mechanical Devices |
|---------|-----------|----------------------------------|
| 1915    | $T_s$     | $P_s$                            |
| 1918-20 | $S_p$     | $S_{pc}$                         |
| 1928    | $F_p$     | $M_i$                            |
| 1931    | $G$       | $SSP$                            |
| 1932    | $G_T$     | $M_D$                            |
| 1938-39 | $B$       | $M_{SC}$                         |

$T_s$ = Taylor’s, $P_s$ = Perforated Stencils, $S_p$ = Soper, $G$ = Goldberg

After this invention¹, a British inventor worked on improvement of Taylor's stencils System and invented a searching device which retrieves information from the given catalogue. Then² invented a device based on scanning and film strips. At the same time some students from American Institute worked on the feasibility of microfilm scanner using decimal codes. With an extension of this project, first prototype machine on microfilm scanners which was based on decimal codes for each text frame has been proposed. After his Invention, it came into noticed that for efficient information retrieval special devices are required.

2.2 Development of Storage Devices in 1940-1970

The management of storage memory itself was a challenging task in that day. In the earlier day's punch cards, magnetic drums were used for storing the instructions. Table 2 shows a history of different inventions in the area of storage devices for storing the data. More details of Table 2 can be found in appendix.
Table 2. Major Inventions in 1915-1940

| Year | Developer | Inventions in Mechanical Devices |
|------|-----------|---------------------------------|
| 1946 | F_w       | RAM                             |
| 1948 | (Rca)     | S_{TB}                          |
| 1949 | Na        | DLM                             |
| 1950 | Na        | M_{c}                           |
| 1956 | Na        | H_{d}                           |
| 1951 | L'        | PC_{LP}                         |
| 1960 | P'        | M_{r}                           |
| 1966 | R_{d}     | Dram                            |
| 1968 | B_{L}     | T_{M}                           |
| 1970 | A_{b}     | B_{M}                           |

F_w = Fredrick C. Williams, RAM = Random Access Memory, Radio Corporation of America

Later on some other technologies like hard disk, bubble memory, dynamic ram played an important role. These were able to store a large collection of data.

2.3 Development of Storage Devices in 1970-2000

In the early days of computer storage, a huge space was covered by the storage devices and it also seemed as a problem. In 1971, IBM launched its 8” floppy drive that was capable to store large data and much faster than others and after this some floppy drives with compact in size but had great speed and hard rigid cover for protection came into market i.e. 5” and 3.5” floppy disks. Although, Russel used light to store and replay the music during 1960’s and this approach turns into optical digital television, it was in 1980 that Sony aided him and they came up with a new invention i.e. compact disk which now a days has been used by everyone. Table 3 shows a history of different developments. More details of Table 3 can be found in appendix.

Table 3. Major Inventions in 1970-2000

| Year | Developer | Inventions in Mechanical Devices |
|------|-----------|---------------------------------|
| 1971 | IBM       | 8"Fp                            |
| 1976 | A_{s}     | 5.25” Fp                        |
| 1980 | J_{r}     | CD                              |
| 1987 | S_{N}     | DAT                             |
| 1989 | S$H       | DDS                             |
| 1992 | S         | Md                              |
| 1993 | Dec       | DLt                             |
| 1994 | Lm        | Zip                             |
| 1995 | Ts        | Sm                              |
| 1995 | Ps        | P_{D}                           |
| 1997 | Si$S      | M_{c}                           |
| 1999 | Na        | Md                              |

A_{s} = Allan Shugart, J_{r} = James T. Russel, Fp = Floppy, CD = Compact Disk, S_{N} = Sony

Later on zip files, flash drives, minidisc came into existence. DVD’s were used to store multimedia data and much faster than CD’s. In 1997 micro cards which are now called as ‘memory cards’ were introduced. In 1997
USB cards were developed. They were very much easy to carry and provide a faster transferring and storage speed as compared to CD’s. These USB cards now called as ‘pen drives’ are used by most of the people on regular basis.

2.4 Development of Storage Devices in 2000-Today

In the contemporary era of technology where the whole world is well connected as a global village, everyday new advancements are facilitating our lives to make it convenient and easier. The commencement of 21st century brought about a revolution in technological inventions. In 2000, major technologies that found headlines were blue rays, pdf and HD-DVD. Databases are now used to store the huge amount of data in the systems. Table 4 shows major inventions in today’s era. More details of Table 4 can be found in appendix. With the growth of web, complexities for storing the data have also increased. Most of the users now use Gmail, drop box and other cloud services for storing their data.

Table 4. Major inventions in 2000-Today

| Year      | Developer | Inventions In Mechanical Devices |
|-----------|-----------|----------------------------------|
| 2002      | O$F       | X
| 2003      | NA        | Br                               |
| 2004      | Na        | Wmvhd                            |
| 2006      | T, N, S   | HdDvd                            |
| 2009      | NA        | Sm                               |
| 2010-today| NA        | C$_5$                            |

O$F = Olympus and Fujifilm, X = Xdpicture Card, Br = Blue ra, T, N, S = Toshiba, Nec, C$_5$ = Cloud Storage

It provides free software, platform and infrastructure as a service to the users. The latest storage mechanisms support fast conversion of data into knowledge for quick decision making.

3. Transformation of Knowledge to Data

The word data can be viewed as an acronym expanded as ‘Do Analysis Thoroughly and then obtain information’. It means that data is the collection of the co-related facts which when inspected, cleaned and transformed produce results to give us specific information and support for choosing the right alternate for decision making. Data is that core part of information which gives us better and expressive output when we process or transfer it into one form to another. Dispensation of data means deep analysis of data followed by operations which reduce its complexity and the resultant output becomes more useful and convenient. Then knowledge leads to decision making. The process of converting data into knowledge is called as “mining of data”. Figure 1 describes the transformation of data into suitable form.

Figure 1. Flow of data to decision.

For the alteration of data, several steps are performed and these steps are known as “preprocessing steps”. Before the implementation of any technique these steps are necessary to make the data into a suitable form. These steps are:

3.1 Selection of Data

In this step, the target data is selected from sources. These can be databases, doc, pdf, excel files, data warehouse, webpage URL, image, audio/video files.

3.2 Preprocessing of Data

This step leads to removal of unnecessary data. It is also
called as cleaning phase for data. It leads to eradicate the noisy data and outlier detection. The most commonly used steps in preprocessing are tokenization, stemming and stop word removal. Some of tools that support preprocessing steps can be named as Rapid miner, Weka, R, Orange. Table 5 shows the steps performed in preprocessing.

### 3.3 Transformation of Data

After preprocessing, the data is ready for analysis. But it might be not in suitable format. It may be a .doc file from one source and .pdf from the other. Then, it is transformed into a suitable form for the implementation of techniques e.g. a database file or doc files needs to be converted into .csv file, .arff format or replacement of tokens, generation of n-grams be done.

### 3.4 Mode Analysis

In this step, various techniques like classification, clustering, regression, association, summarization and ranking are analyzed for efficient results.

### 3.5 Visualization

This step leads to visualizing the results after implementation of the techniques. Some of the important parameters that can be used for performance evaluation are Precision, Recall and Accuracy etc. These steps play a very important role and are very necessary for obtaining the final resultant output.

### 4. Related Work

Information retrieval illuminates the science of discovering the satisfactory relevant information from a large amount of available data. From the period of 1990s till today, a number of mechanisms have been used for getting the convenient information. While the mechanisms may not be new, but the way of locating information has thoroughly changed. In Vedic period, data was stored in different mechanical devices but the challenges commonly faced included storage and accessibility of information. In six discussed the use of computers for information retrieval along with various classification techniques and focused on two searching techniques i.e. localization and abstraction. The use of ENIVAC and UNIVAC computers at that time for processing the input output operation was also mentioned.

Along with the storage and speed issues, another important challenge in data mining is the clear classification of words. In seven discussed the problem related to manual operations of libraries and put forward the problem of searching of information. In this, the author described retrieval of information in systems using punch cards. The main challenge found in retrieval of information is their accuracy and time taken. In most of the historical

---

**Table 5. Preprocessing steps**

| Steps            | User query                                           | Output                      |
|------------------|------------------------------------------------------|-----------------------------|
| Tokenization     | "John and Maria are playing"                         | 'John', 'and', 'Maria', 'are', 'playing' |
| Stemming         | “Differntiate between Reliability and roubstness”    | 'differe', 'reliab', 'robustn' |
| Stop word removal| "John and Maria are playing"                         | John, Maria, Playing         |
records, the data is found to be arranged in alphabetically order, either by numbers or by symbols. There was no other mechanism which would make the system more efficient for organizing the data. In\textsuperscript{4} offered an “indexing mechanism” by making a coordinate index for each word which is entered into the system and is explained by a table of words. Two major problems that occurred in indexing are synonyms and homonyms. These problems are related to same meaning of words e.g. defined, define, defining are similar kind of words.

In\textsuperscript{9} discussed about “recall” and “pertinency factor”, to define proportion of relevant documents that are retrieved and documents that are relevant. Around 1965 it was recognized as “precision ratio”. Computers were being used to perform complex calculations so researchers intended to make the retrieval automatic with lesser user intervention. This introduced the concept of “synthetically matching documents” which now is called as pattern matching. After these techniques the artificial intelligence proposals were developed by various researchers working on not only text files but with database files also. In indexing, the main issue is the complexity in finding out the index terms. Author also discussed Switzer’s idea of finding query based on similarity and Salton’s approach to measure the cosine angle between documents for similarity measure.

In\textsuperscript{10} defined a literature searching mechanism, which deals with automatically finding the abstracts of papers. The work focused on analysis of words and sentences as an important factor and using this a new term “frequency weighting” was defined. It gauges the number of occurrence of a word in the document. The author worked on around 2300 words in a document and a word frequency diagram has been generated from them.

Based on Luhn’s approach on “frequency of words” using frequently occurring words, another author in\textsuperscript{11} emphasized on Luhn’s “indexing and searching mechanism” which was based on index words and relative frequency of documents. In the period of 1960’s many authors worked on statistical approaches like rank based retrieval, probabilistic indexing. The retrieval of information depends on these common factors i.e. how good it is in finding the relevant documents and how efficiently it rejects the irrelevant data.

In\textsuperscript{12} worked on vector model and discovered the “vector space model”. The measurement on space density using 424 documents has been done and the effect of performance has been shown using two different cluster algorithms. CRAN, MED, TIME data collections has been used to evaluate recall and precision for these. The evaluation of performance between three collections is shown in Table 6.

|                | CRAN 424 | MED 450 | TIME 425 |
|----------------|----------|---------|----------|
| A\textsubscript{TF} | +32%     | +39%    | +17%     |
| A\textsubscript{TP} | +33%     | +50%    | +18%     |
| I\textsubscript{R}   | 0.89     | 0.88    | 0.85     |
| M\textsubscript{R}   | 0.43     | 0.61    | 0.70     |
| H\textsubscript{R}   | 0.13     | 0.23    | 0.45     |

A major development of Porter Stemming rule\textsuperscript{13} was related to focusing on removal of noisy words and finding out the root model. This helps to solve the problem of synonymy. In the period of 1980-90 the advancements in vector space model resulted into “Latent Semantic Analysis” (LSI). A discussion on “Inverse Document Frequency” has been presented.

For analyzing and visualizing of data, In\textsuperscript{14} defined three classes of database mining problems involving classification, associations and sequences and described a model and five basic operations for the process of rule discovery. The author has also thrown light on how these database mining problems map to this model and proposed an algorithm which is not only efficient in discovering classification rules but also has accuracy comparable to Id3, one of the current best classifiers. Id3 stands for iterative diconomer and is one of the most used algorithms in decision tree. With the rapid development of data, new
algorithms were proposed by various authors’ id3, j48, c4.5. Classification itself is a broad area of research that includes many techniques like neural networks, Naive bays, regression, k-mean, rough set and Associative classification.

In\textsuperscript{15} prominence the techniques used for data mining i.e. classification and found that extracting pattern information is very difficult using database queries. An approach to discover information using neural networks has been proposed. In this paper the information is divided extraction algorithm into three phases: The first is training of data i.e. to train the data to achieve accuracy for future data. The second phase describes the network pruning algorithm to remove the noisy and redundant connection from the network. The experimental result was performed on 30 networks and an average of accuracy, conditions and number of rules was calculated as represented in the Table 7.

The third and final phase is the overall extraction of information using previous data. In the implementation, C4.5 decision tree algorithm and three fold Cross validation has been used to obtain estimated accuracy. In\textsuperscript{16} elaborated the various techniques and steps involved in Knowledge Discovery in Databases (KDD). A clear difference between KDD and data mining is given. According to author, KDD process is set of various activities for making sense of data. The term KDD used first in 1989 as discussed by the author. While he describes the data mining as a process to extract the knowledge through different algorithms and he defines nine steps to discover the knowledge from the given data. These steps are called “Preprocessing Steps” in the field of data mining.

In\textsuperscript{17} has discussed about the statistical approach towards the vector space model. This technique was based on weightage of words in document. He clarified this by giving a formula,

\[ w_{i,j} = tf_{i,j} \cdot idf_{j} = tf_{i,j} \cdot \log \frac{N}{df_{j}} \]  

Where ‘i’ is the document and ‘j’ is the weighted term, N is the number of documents. It assigns the high weightage to those terms which are occurring frequently in the

| Function   | Accuracy      | Number of rules | Number of Conditions |
|------------|---------------|-----------------|----------------------|
| Salary     | 99.91 (0.36)  | 2.03 (0.18)     | 2.23 (0.50)          |
| Commission | 98.13 (0.78)  | 7.13 (1.22)     | 4.37 (0.66)          |
| Age        | 98.18 (1.56)  | 6.70 (1.15)     | 3.18 (0.28)          |
| Elevel     | 95.45 (0.94)  | 13.37 (2.39)    | 4.17 (0.88)          |
| Car        | 97.16 (0.86)  | 24.40 (10.18)   | 4.68 (0.87)          |
| zipcode    | 90.78 (0.43)  | 13.13 (3.72)    | 4.61 (1.02)          |
| Hvalue     | 90.50 (0.92)  | 7.43 (1.76)     | 2.94 (0.32)          |
| Loan       | 90.86 (0.60)  | 9.03 (1.65)     | 3.46 (0.36)          |
document set and discussed different models; these were based on normalization of terms, efficiency of normalization factor and computing the projection length of document instead of computing cosine angle between document vectors. At this time Web was also evolving and with the rapid development of websites, the amount of files and data was also increasing. Now the next challenge was the processing and management of this data and such big databases. Then a process named as “Knowledge Discovery of Database” came into existence. At the same time the term “Data Mining” is also adopted that relates to extraction of relevant information from the given data.

Social network is also a challenge now-a-days. Several algorithms have been proposed for ranking, summarization of data. Another approach which is very useful in prediction of results and feature weighting is nearest neighborhood search. In emphasized on algorithms based on elimination rules. These algorithms used the distance measuring properties e.g. partial distance measuring, triangle inequality. This paper puts a light towards the elimination techniques in nearest neighbor search. Other discussed algorithms were elimination based on nearest neighbor search, Approximation based, hypercube method.

In proposed five evolutionary-based search methods were presented. Based on this comparative analysis, the performance of EA s is discussed along with some guidelines for determining the best operators for each algorithm. Table 8 shows a discrete optimization based on success rate, processing time and average cost.

| Algorithm | Average cost | Processing time | Success Rate |
|-----------|--------------|-----------------|--------------|
| GA        | 164,772      | 16              | 0            |
| MA        | 162,495      | 21              | 20           |
| ACO       | 166,675      | 10              | 20           |
| SFL       | 166,675      | 15              | 0            |
| PSO       | 161,940      | 15              | 60           |

The PSO (Particle Swarm Optimization) method was generally found better to perform than other algorithms in terms of success rate and solution quality, while being second best in terms of processing time. In concentrated on effective retrieval of queries based on keyword search on web. A focus on how weights are assigned to keywords along with how assignment of edges and how to calculate the cost of graph has been done. For the proposed model edge weights has been assigned on RDFS and introduce a keyword to node function to simplify the representation. A conversation on disjunctive and conjunctive semantics for the queries has been done. An approximate algorithm for novel IR-style keyword search for semantic web data retrieval has been presented. In this, to answer a query is defined as a minimal connected sub graph that contains all the query keywords. A special ranking mechanism has also developed. In the results a focus on performance and effectiveness and a comparison with Random and optimum models has been performed their performance is slower than Random but provides good results comparable to optimum i.e. 0.93 in 10s.

In discussed the storage of data in the form of records, documents etc. and the process of extracting the information and converting it into useful manner is called KDD. It stands for knowledge and data discovery. After this the various fields of data mining e.g. statistics, database, machine learning, pattern discovery and artificial intelligence have been discussed along with it, elaboration of various fields have been done like customer relation management, web application, search engine optimization. In current trends, different data mining techniques like hypermedia data mining, ubiquitous data mining, multimedia and spatial data mining and time series data mining have been included. In future trends the focus is the difficulties with complex objects, web mining and more efficient information retrieval.

In proposed an oblique decision tree algorithm named as BUTIA. An oblique tree can generate polygonal partitioning, while univariate trees can generate hyper rectangle. While most of the authors worked on top down approach but here a bottom up approach has been discussed by obtaining prior knowledge about the path of tree and before each split, a hyper plane is produced. While in top down the generation of tree depends on purity of each leaf. A comparison analysis of proposed algorithms with other decision tree algorithms (C4.5, CART, OC and FT)
has been performed. The result has been performed on 35 datasets that were related to different types of cancers. The proposed algorithm gives advantages over robustness and data over fitting and it has been divided into five steps.

- In first step, the training data is divided into subsets and a label is assigned to the same.
- In second step, 10-fold cross validation has been performed and EM (Expectation Maximization) is applied on each subset to define the clusters. In third step, the centroid of each cluster has been calculated to act as a leaf node.
- In last step, the internal node is generated by mar-gining two nodes having closest centroid number and calculating the new node centroid. The support vector machine hyper plane for internal nodes has also been discussed.

In\textsuperscript{26} discussed about the different techniques like generalization, characterization, pattern matching and applications of data mining in various fields. Also neural network along with the applications like fuzzy sets, regression, flow networks and Bayesian networks has also been discussed. In next section there is a descrip-tion of a mathematical model for stochastic dynamics for modeling molecules used in stock market, vehicle fault diagnosis and fault prediction, forcasting and anomaly detection. Some suggestions to work upon the method-ologies of social science and integration of methodologies have been discussed to enhance the understanding of data mining.

In\textsuperscript{27} discussed various techniques of Text Mining like summarization of text, visualization, nearest neighbor-hood, clustering. Graphical Visualization is used to provide illustrative information for mining the documents. A discussion on various applications such as identifying news stories, junk e-mails and analysis of the market trends has been done. In\textsuperscript{28} emphasized on social media’s effect on customer relation management, tracking of user feedback and also emphasis on some other applications. Social network an became very important medium of communication and it has a major influence on public opinion behaviors like blogging, stock market analysis and sentiment analysis. According to author, the analysis of web is a kind of natural language processing technique which aims to find out the hidden patterns in a large number of reviews and blogs.

As most of the times the users\textsuperscript{29} in such informal communication use slang words instead of exact words and sentences. 3600 verbs and 6800 adjectives have been used to conduct this analysis. The author divides this analysis into four modules. In the first module, he discusses about product reviews for example camera, phones, food in the hotels. For the experimental analysis 268 reviews has been taken from the opinion of users. In the second module, an analysis on movie reviews like ‘Slum Dog Millionaire’, ‘American Gangster’ has been performed. In\textsuperscript{30} different verbs and adjectives were used to analyze the user feedback. In third module an examine on political orientation using 10,000 tweet messages investigated in 2011 has been performed. In third they discuss about the stock market analysis using machine learning to predict the industrial average closing values. In the sentiment score evaluation of different brands data such as IBM, Nokia, Air India etc. have been used and then discussed about different methods like twitter sampling, lexicon.in the last they discuss the results by using different pie charts. The libraries of twitter, the plyer and ggplot have been used with R software package to conduct the sentiment score and a Stem graph software package has been used to visualize to tweets on all brands. A talk on further use large training data sets to detect of has been discussed.

In\textsuperscript{31} worked on visualization of relevant words. In this three document of small (tutorial document of ‘sna’), medium (H.P. Luhn’s paper) and large (W3C RFC 2616) size has been selected. It was based on two main val-ues i.e. number of same or relevant words appearing in documents and means distance between them to link the chain of same words. In this paper NumPy and Natural Language Toolkit has been used for visualizing the results. After performing the preprocessing steps (stemming, stop word removal) they have created a network of relevant words from these three documents.

In\textsuperscript{32} emphasized on an alternative approach for relevance feature discovery in text documents. It presents a method to find and classify low-level features based on
both their appearances in the higher-level patterns and their specificity. A method to select irrelevant documents for weighting features also discussed in the paper. Two algorithms named fclustering () and Wfeature has been presented. In this paper, RFD model has been presented and experimentally prove that the proposed Wfeature provides better results specificity function is reasonable and classification can be effective approximated by a feature clustering method. The experimental result has been performed on RCV1, (TREC topics), Reuters-21578. These data sets are open source and freely available. In a novel approach has been proposed using neural network. In this paper the concept of neural networks has been discussed in detail for the speech reorganization application in information retrieval. A comparison analysis of different techniques has been done in Table 9. More details of Table 9 can be found in appendix. Along with advantages and limitations, specific algorithms are also discussed. Some of the techniques like probability, vector based, Boolean were commonly used some years ago and they are beneficial for some specific applications with the advancement of technology and growth of web, these techniques do not provide efficient results. The most commonly used techniques now days like neural networks, regression, decision tree, Bayesian, Support vector machine are also elaborated in it.

Table 9. Comparative analysis of techniques

| Sr. no | Techniques | Algorithm Used | Specific Application | Pros | Cons |
|--------|------------|----------------|----------------------|------|------|
|        |            |                |                      | (i)  | (ii) |
| 1.     | P_{rank}   | P_{tp}, P_{zm}, B_{25} | Tpc_{dr} | (i) Rnk_{dc,Prob} | Coo\_df | P_{kn} |
| 2.     | I_{index}  | P_{rt}, B_{dm}, NA | S'_{ng,ish} | (ii) A_{uc} | Sig_{tr,df} | na |
| 3.     | B_{index}  | Use_{as,lp}, NA | S'_{rch} | L_{rs,arch} | na | P$S_{ric}$ | (ii) B_{qinc} |
| 4.     | V_{sh}     | B_{em}, E_{im}, I_{ai} | D_{pg}, S_{mrt}, Pr_{mch} | C_{r,df}, SSL_{mog} | Trm_{ia} |
| 5.     | Dc_{vec}   | Id3 | M_{rs} | C4.5 | IBM_{am} | M_{sql2} | Hn_{ms,Cg} | Lt_{pp} | !O.R_{np-hard} | C_{Pod} |
| 6.     | S'          | O_{F}, S_{lm} | H_{lp}, L_{b} | SMO | P_{rec} | P_{ct,th,l} | A_{pl,of} | Pi_{df} | Al_{c5mm} |
| 7.     | N_{G}      | GN_{B}, MN_{B}, BN_{G} | T_{x_{ct}}, S_{B}, Fs_{TIC} | Hl_{RSDd} | As_{if} | 0c_{pib} |
Table 9 Continued

| 9. | Rg | Rg | SRg | LRg | Binf | Taul | RprAss | HdRSCf | lnDf | MdpV |
|---|---|---|---|---|---|---|---|---|---|---|
| 10. | Rgt | Lt | Lt2 | Lem | Rse | Lsex | Dtg | USF, | §Ri | HNd | §Ms | Exdm | Dp |
| 11. | Ft | Mpar | Mpar | Na | Bdh | Dsk | SlT | Opexp | Sx |
| 12. | Acf | CAEP | CMAR | ADT | Gpg | RvNd | AlgCS | RAp | ChlmgR | NA |
| 13. | Kn | Iw | Fw | NA | Ner | Rs | EffTSP | na | BvLM | ChcsLA |
| 14. | An | Rdr | Pr | Emt | Srec | Hrec | ElO | HAI | ExmImp | NA |
| 15. | Rbcx | Hmany | JEP | Rpr | Bmed | DSC | UsLT0 | RprAss | Dpc | NA |
| 16. | Exc | Bun | Bgng | Bmc | Mtrng | NA | Uspp | Cqpt | lnS | |
| 17. | Pbcx | Rem | Fsa | NA | Rcmd | Bmt | HIR3P | PCA | SnNf | Bl |
| 18. | Nn | Ppt | Almp | Wsa | Rst | AI | Abocr | HAI | NEppt | Diffimg |
| 19. | Pm | Pm | Cr | Br | Crms | NA | EsNic | Sutcs | IboclCl | SnNIO |
| 20. | Hm | DIANA | AGNES | Fs | Xcl | Re | Pclt1 | Ap2af | Ibm | Alscs |

*S* = search, *w* = of, *df* = difficult, *probm* = probability method, *indxm* = Indexing method, *porter* = porter algorithm, *bdm* = bdm algorithm, *Sng,hsh,sm* = Used by search engines for n grams, *A-lang* = Language ambiguity, *B-bool* = Boolean model, *Use* = Uses AND, OR and other logical operations, *B-queCx* = Complexity in making Boolean queries, *VSM* = Vector space model, *B-vM* = Basic vector model, *E-eM* = extended Boolean model, *L-LSI* = LSI, *D-pig* = Dynamic programming, *S-mart* = SMART

5. Results and Discussions

This section discusses the experimental results and discussion of comparative analysis of three algorithms i.e. Id3, CHAID, naïve bays. The evaluation of results has been performed using Rapid Miner5 on windows 64 bit operating system. Table 10 shows the sample data set.
| Table 10. Sample dataset |
|--------------------------|
| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 |
5.1 Data Set

The dataset has 30 attributes and it is based on performance of two schools “GP” - Gabriel Pereira or “MS” - Mousinho da Silveira. It is open source and available on UCI repository. Each attributes is further sub-divided into sub attribute. The corresponding mapping of attributes of sample data set is given below.

5.2 Important Terms

While doing the literature or during performing any experiments results following terms are very much important to understand. Table 10 shows the sample dataset. A brief discussion on various parameters on which we evaluate the performance or accuracy of any algorithm has been discussed below.

SC = school(MS,GP), SE = sex (M, F)
AG = age AD = address (Rural, urban)
FS = Familysize (Gt3, Lt3)
PS = Pstatus (Apart,Together)
ME, FE = Mother, father Education
MJ, FJ = Mother, father job
R = reason to adopt school (H,R,O)
G = guardia, H = higherstudy
TT = traveltime
ST = studytime
F = failures
SU = schoolsup (Y,N)
FSP = famsup (Y,N)
P = paid (Y,N)
A = activities,
I = internet (Y,N)

Confusion matrix: A confusion matrix is used to describe the performance of classification, usually represented in table form. Table 11 shows a tabular structure of confusion matrix and some other important terms have been discussed for the better understanding of how to calculate them.

Table 11. Structure of confusion matrix

|       | Negative | Positive |
|-------|----------|----------|
| Negative | A        | B        |
| Positive | C        | D        |

From the given confusion matrix we can evaluate the precision and recall values. Following example will help to understand these in a good manner. Suppose a user put a query for document ‘A1’. On the server side there correctly classified or relevant, rest of the document are total 50 documents are relevant to the user’s query. But as a result, the search engine classifies 42 documents that are correct.

- Precision: It is the fraction of those retrieved instances that are relevant.
  \[ P = \frac{35}{42} \]
- Recall: It is fraction of relevant instances that are retrieved.
  \[ R = \frac{35}{50} \]

In case we have both training and testing data sets. When one of the classes is providing true response and other class is providing false result, that is called type 1 error and type 2 error. Type 1 is incorrect rejection of true null hypothesis. On the other hand rejection of false null hypothesis is called type 2 error. Table 12 shows some other important terms related to confusion matrix.

Table 12. Calculation of various parameters

| Terms | Formulae |
|-------|----------|
| TP    | \( \frac{(D)}{(C+D)} \) |
| FP    | \( \frac{(B)}{(C+D)} \) |
| TN    | \( \frac{(A)}{(A+B)} \) |
| FN    | \( \frac{(C)}{(C+D)} \) |

\[ TP = \text{truepositive}, FP = \text{falsepositive}, TN = \text{truenegative}, FN = \text{false negative} \]
The precision can be calculated from formulae (2)

\[ P = \frac{TP}{TP + FP} \] (2)

The Recall can be calculated from formulae (3)

\[ R = \frac{TP}{TP + FN} \] (3)

The accuracy can be calculated by formulae (4)

\[ A = \frac{B+C}{A+B+C+D} \] (4)

5.3 Algorithm Used

5.3.1 ID3 Algorithm

ID3 stands for Iterative Dichotomiser 3 and invented by Ross Quinlan33. It builds a decision tree based on fixed example set. Each leaf node acts as class and non-leaf node acts as a decision. It worked according to the selection of criterion parameter (information gain, gini index). Both training and testing sets has been provided. It first searches out the attributes of the training set and extracts those attribute that closely separates the provided data set. If the attributes perfectly classifies the training sets.

**Algorithm_ID3**

```
Begin
Initialize D_set

For (D_set = 1, D_set <= D_set 1, D_set n)

    Select l_value n

    For (l_value = 1, l_value <= l_label 1, l_label n)

        Select attribute A_b

        If (A_b = nominal)

            Apply id3
            //id3 is a algorithm for decision tree

        else

            Change A_b

        M_model N = float C_Index, I_Gdx // where M_model N € Id3

        Generate M

        //where M_model M is clm of M_model N

    End

End
```

**Figure 2.** Algorithm ID3 for decision tree.
then algorithm stops, otherwise it performs n number of iterations. The algorithm starts with initialization of \( D_{set} \) and for each value of data set a \( l_{value} \) has been selected from the Attributes \( (A_b) \). If the data set is nominal the ID3 will be applied on it otherwise, \( A_b \) needs to be changed. Figure 2 shows an algorithm for ID3.

In next step, criterion parameter will be selected and a \( M_{model N} \) will be generated based on that parameters. Then a model \( M_{model M} \) has been generated and it will undergo from testing and training module. If it validates successfully then \( P_T \) will be generated. An implementation of this algorithm using 10 fold cross validation has been performed in Rapid Miner. Table 13 shows the confusion matrix for the provided data set using ID3 algorithm.

**Table 13.** Confusion matrix for ID3

|          | true GP | true MS |
|----------|---------|---------|
| pred. GP | 340     | 7       |
| pred. MS | 9       | 39      |

The precision, recall and accuracy can be calculated for this algorithm on the basis of confusion matrix. In Table 14 a calculation of three parameters for their performance analysis has been discussed. In this table the (✓) belongs to class ‘GP’ and (❖) belongs to class ‘MS’. The values of TP, TN, FP, FN are related to the correctly and incorrectly classified instances form the algorithm. The values of TP, TN, FP, FN are calculated by the algorithm using the formulae discussed in previous section.

**Table 14.** Results of precision, recall and accuracy

|       | TP | TN | FP | FN | GP | MS |
|-------|----|----|----|----|----|----|
|       | 340| 39 | 7  | 9  |    |    |
| R     | ✓  | ❖  | ❖  | ✓  | 97.42% | 84.78% |
| P     | ✓  | ❖  | ✓  | ❖  | 97.98% | 81.25% |

R = Recall, P = Precision

Table 15 shows mean recall, precision and accuracy values using ID3 algorithm.

**Table 15.** Mean recall, precision and accuracy

| Parameter      | Formulae                                      | Mean value |
|----------------|-----------------------------------------------|------------|
| Mean recall    | \( \frac{(97.42 + 84.78\%)}{2} \)            | 91.10%     |
| Mean precision | \( \frac{(97.98\% + 81.25\%)}{2} \)          | 89.62%     |
| Accuracy       | \( \frac{(340 + 39)}{(340+7+9+39)} \)       | 95.95%     |

5.3.2 CHAID Algorithm

CHAID stands for Chi-squared Automatic Interaction Detection. Chi-square is a statistical technique used to determine if a distribution of observed frequencies differs from expected frequencies. It performs results on nominal data. It only uses nominal data. We cannot apply it on data set having numerical attributes. In Figure 3 an algorithm has been presented for CHAID. The algorithm starts with initialization of \( D_{set} \) and for each value of data set as \( l_{value} \) has been selected from the Attributes \( (A_b) \). If the data set is nominal the ID3 will be applied on it otherwise, \( A_b \) needs to be changed. In this algorithm each leaf node represents the label value which we set while importing the data set into rapid miner. In next step, an \( E_{epoch\ n} \) value will be selected, where \( E_{epoch} \) represents to number of iterations and a \( M_{model N} \) will be generated based on that parameters. \( N \) can vary from 1 to \( N \). During the evalu-
ation of results \( n = 3 \) has been selected to make the data set into more general form.

Then model \( M_{\text{model M}} \) has been generated and it will undergo from testing and training module. If it validates successfully then \( P_I \) will be generated. Table 16 shows the confusion matrix for CHAID algorithm.

### Table 16. Confusion matrix for CHAID

|       | true GP | true MS |
|-------|---------|---------|
| pred. GP | 328     | 22      |
| pred. MS  | 21      | 24      |

---

**Figure 3.** Algorithm for CHAID.
Table 17. Calculation of precision, recall and accuracy

|   | TP | TN | FP | FN | GP | MS |
|---|----|----|----|----|----|----|
|   | 328| 24 | 22 | 21 | 93.71% | 53.33% |
| R | ✓  | ❖  | ✓  | ❖  |     |     |
| P | ✓  | ❖  | ✓  | ❖  | 93.98% | 52.17% |

Table 17 shows the empirical results using CHAID algorithm. The values of TP, TN, FP, FN are calculated by the algorithm.

Table 18 shows the results of mean recall, precision and accuracy using CHAID algorithm.

Table 18. Mean recall, precision and accuracy

| Parameter   | Formulae                                      | Mean value |
|-------------|-----------------------------------------------|------------|
| Mean recall | \((93.98\% + 52.17\%) / 2\)                  | 73.08%     |
| Mean precision | \((93.71 + 53.33\%) / 2\)                    | 73.52%     |
| Accuracy    | \((328 + 24) / (328 + 22 + 21 + 24)\)         | 89.11%     |

An algorithm have presented and results has been discussed which has been performed in Rapid Miner.

5.3.3 Naive Bayes Algorithm

A Naive Bayes classifier is a simple probabilistic classifier based on Bayes. It deals with strong independence assumptions. In Figure 4 an algorithm for naïve bayes has been presented.

Another term for this model would be ‘independent feature model’. Naive Bayes classifier adopts that the presence or absence of a certain feature of a class is not associated to the presence or absence of any other feature. For example, a fruit may be considered to be an orange if it is orange, round and about 5 inches in diameter. Even if these features depend on some other or upon the presence of the other classes. It considers all of the properties independently. The advantage of the Naive Bayes classifier is that it requires a small amount of training dataset for assessment of means and variances of the variables essential for classification. As independent variables are assumed, only the variances of the variables for each label is required to be determined and not the complete covariance matrix. The confusion matrix for Naïve Bayes algorithm is shown in Table 19.

Table 19. Confusion matrix for Naïve Bayes

|            | true GP | true MS |
|------------|---------|---------|
| pred. GP   | 302     | 25      |
| pred. MS   | 47      | 21      |

Table 20. Calculation of precision, recall and accuracy

|   | TP | TN | FP | FN | GP | MS |
|---|----|----|----|----|----|----|
|   | 302| 21 | 25 | 47 |    |    |
| R | ✓  | ❖  | ✓  | ❖  | 92.35% | 30.88% |
| P | ✓  | ❖  | ✓  | ❖  | 86.53% | 45.68% |
Algorithm Naïve Bayes

```
Begin
Initialize D_set
{
    For (D_set = 1, D_set <= D_set 1, D_set n)
    {
        Select l_value_n
    }
    For (l_value = 1, l_value <= l_label l, l_label n)
    {
        Select attribute A_b
        If (A_b = nominal)
        {
            Apply Naive Bayes
        }
        else
        {
            Change A_b
            M_model_N = Select L_C Where L_C € each value N_B
            Generate M_model_M //where M_model_M is clm of M_model_N
        }
        if (M_model_M = P_T) //P_T = Performance Index
    }
End
```

Table 21 shows a weighted mean, recall and accuracy using naïve bays algorithm.

Table 21. Mean recall, precision and accuracy

| Parameter     | Formulae                        | Mean value |
|---------------|---------------------------------|------------|
| Mean recall   | ( 92.35% + 30.88 %) / 2         | 61.62%     |
| Mean precision| ( 86.53 + 45.68%) / 2           | 66.09%     |
| Accuracy      | ( 302 + 21) / ( 302+ 21 + 25 +47) | 81.77%     |

A discussion on advantages and limitation of three algorithms has been presented i.e. CHAID is having limitation over working on nominal data only whereas naïve bays can handle real and discrete data, but it suffers from the Problem of zero condition probability. For analysis of which performs better a comparison of these three algorithm has been shown in Table 22.

Table 22. Comparative analysis of algorithm

| Algorithm Used | Accuracy | Precision | Recall  |
|----------------|----------|-----------|---------|
| Id3            | 95.95%   | 91.10%    | 89.62%  |
| CHAID          | 89.11%   | 73.08%    | 73.52%  |
| Naïve Bayes    | 81.77%   | 66.09%    | 61.62%  |
But each one performs better according to their requirements. As per our experimental results ID3 classifies more accurately as compare to other two algorithms.

5. Conclusion

In this work, invention of different storage devices and vast comparative analysis of data mining techniques have been presented. The basic operations associated with data mining also discussed. Three widely used algorithms with respect to the parameters precision, recall and accuracy showed the efficiency and output can be improved if one can deploy ID3, CHAID and Naïve Bayes with random data set. Experimental set up is done in Rapid Miner and the output clearly depicts the performance level has raised up to 95.95% in case of ID3, 89.11% in case of CHAID and 81.77% in case of Naïve Bayes algorithm. In the last section results have been shown. Thus, verifying the performance of proposed algorithm.

6. References

1. Soper HE. Means for compiling tabular and statistical data. U.S. Patent US0135169231; 1920.
2. Goldberg E. Statistical Machine. U.S. Patent 183838929; 1931.
3. Modern computer system. 2016. Available from: http://history-computer.com /Modern Computer/ Basis/compact_disc.html
4. History of computer storage devices. 2016. Available from: http://www.zetta.net/history-of-computer-storage
5. History of all usb devices. 2016. Available from: http://www.allusb.com/usb-history
6. Nanus B. The use of electronic computers for information retrieval. Bulletin of the Medical Library Association. 1960; 48(3):278.
7. Calvin NM. The theory of digital handling of non-numerical information and its implications to machine economics. Zator Co.1950; 48:48–52.
8. Taube M, Gull CD, Wachtel IS. Unit terms icordinate indexing. American Documentation. 1952; 3(4):213–8.
9. Kent BA. Operational criteria for designing information retrieval systems. American Documentation. 1955; 6(2):93–101.
10. Luhn HP. The automatic creation of literature abstracts. IBM Journal of Research and Development. 1958; 2(2):159–65.
11. Bar-Hillel Y. The mechanization of literature searching. in: mechanization of thought processes. Proceedings of a Symposium held at the National Physical Laboratory, HM Stationery Office; 1959. p. 791–807.
12. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of the ACM. 1975; 18(11):613–20.
13. Porter MF. An algorithm for suffix stripping. Program: Electronic Library and Information Systems. 1980; 14(3):130–7.
14. Agrawal R, Imielinski T, Swami A. Database mining: A performance perspective. IEEE Transactions on Knowledge and Data Engineering. 1993; 5(6):914–25.
15. Hongjun L, Setiono R, Liu H. Effective data mining using neural networks. IEEE Transactions on. Knowledge and Data Engineering. 1996; 8(6):957–61.
16. Usama MF, Pietetsky-Shapiro G, Smyth P. Knowledge discovery and data mining: Towards a Unifying Framework. KDD. 1996; 96:82–8.
17. Lee DL, Chuang H, Seamons K. Document ranking and the vector-space model. Software, IEEE. 2000; 14(2):67–75.
18. Ramasubramanian V, Paliwal KK. Fast nearest-neighbor search algorithms based on approximation-elimination search. Pattern Recognition. 2000; 33(9):1497–510.
19. Sahu SK, Mohapatra DP, Balabantaray RC. Information retrieval in the context of checking semantic similarity in web: Vision of future web. Indian Journal of Science and Technology. 2016 Aug; 9(32):1–12.
20. Elbeltagi E, Hegazy T, Grierson D. Comparison among five evolutionary-based optimization algorithms. Advanced Engineering Informatics. 2005; 19(1):43–53.
21. Ning X, Jin H, Jia W, Yuan P. Practical and effective IR-style keyword search over semantic web. Information Processing and Management. 2009; 45(2):263–71.
22. Meenakshi A, Suganthi P, Aghila R, Nirmala S. Information retrieval using dynamic decision quadtree in soil database. Indian Journal of Science and Technology. 2016 Mar; 9(10):1–7.
23. Reddy LC. A review on data mining from past to the future. International Journal of Computer Applications. 2011; 7(4):19–22.
24. Barros RC, Cerri R, Jaskowiak P, de Carvalho ACPLF. A bottom-up oblique decision tree induction algorithm. 2011 IEEE 11th International Conference on Intelligent Systems Design and Applications (ISIDA); 2011. p. 450–6.
25. Liao SH, Chu PH, Hsiao PY. Data mining techniques and applications – A decade review from 2000 to 2011. Expert Systems with Applications. 2012; 39(12):11303–11.
26. Sukanya M, Biruntha S. Techniques on text mining. 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCCT); 2012. p. 269–71.
27. Mohamed MM. More than words: Social networks text mining for consumer brand sentiments. Expert Systems with Applications. 2013; 40(10):4241–51.
28. Silva LCS, Sampaio RR. Using Luhn’s automatic abstract method to create graphs of words for document visualization. Social Networking. 2014; 3(2):65–70.
29. Vigneshwari S, Aramudhan M. Social information retrieval based on semantic annotation and hashing upon the multiple ontologies. Indian Journal of Science and Technology. 2015 Jan; 8(2):103–7.
30. Venkatesan N, Arunmozhi K, Arasan AA, Muthukumaran S. An ID3 algorithm for performance of decision tree in predicting student’s absenteeism in an academic year using categorical datasets. Indian Journal of Science and Technology. 2015 Jul; 8(14):1–5.
31. Sajeer K, Rodrigues P. Novel approach of implementing speech recognition using Neural Networks for information retrieval. Indian Journal of Science and Technology. 2015 Dec; 8(33):1–5.
32. Abdulmohsen A, Mubarak A. Relevance feature discovery for text mining. IEEE. 2015 Jun; 27(6):237–43.
33. History of ID3. 2016. Available from: https://Id3www.cse.unsw.edu.au/~billw/ cs9414/ notes/ ml/06prop/ id3/id3.html

Appendix

T_s = Taylor’s, PS = Perforated Stencils, SP = Soper, S’C = Device for Searching Information from Catalogue, FP = Fritz Pfleumer, G = Goldberg, S$P = Scanning And Photographic Device, GT = G. Taushek, MD = Magnecy Drum, Mt = Magnetic Tapes, B = V. Bush, MSC = Microfilm Scanner, FW = Fredrick C. Williams, RAM = Random Access Memory, Radio Corporation Of America (Rca), STB Selecton Tube, DLM = Delay Line Memory, MC = Magnetic Core, HD = Hard Disk, L’= Luhn, PCL,P = Selector Based On Punch Cards, Light And Photocells, P’ = Philips, MT = Music Tape, RD = Robert H. Dennard, Dram = Dynamic random access memory, BL = Bell Labs, TM = Twistor Memory, AB = Andrew Bobek, BM = Bubble MemoryAS = Allan Shugart, JT = James T. Russel, Fp = Floppy, CD = Compact Disk, SN = Sony, DAT = Digital Audio Tape, S$H = Sony And Hp, DDS = Digital Data Storage, MD = Magneto optical Disc, Md = Minidisk, S = Sony, Dec = Digital Equipment Corporation, DLT = Digital Linear Tape (Dlt), CF = Compact Flash, Lm = Lomega, Sm = Smartmedia, Ts = Toshiba, Ps = Panasonic, S$S = Siemens And Sandisk, PD = Phasewriter Dual,MC = Multimedia Cad, Md = Microdrive, O$F = Olympus And Fujifilm, T,N,S = Toshiba, Nec, And Sanyo, XC = Xdpicture Card, Br = Blueryar, Wmvhd = Windows Media

High Definition Video ,CS = Cloud Storage, S” = search .’ = of, df = difficult, P rob . = Probability method, P rob 2m. = 2-poissen model, B 25 = BM25, Rnk = Rank documents based on the estimated probability, A uc = Address uncertainty in query representation, Tp df = Topic detection and ranking, Coo df = difficulty in control over output, P kn = Prior knowledge is needed, I ndx = Indexing method, P ri = Porter algorithm, B dm = BDM algorithm, S ngr,hsh,sm = Used by search engines for n grams, Sig tf.idf = hash table and sparse matrix = useful in significance of document using term frequency, inverse document frequency, A ing = Language ambiguity, B in = Boolean model, Use as,lp = Uses AND, OR and other logical operations, Srch as dc = Used by search engines for finding similar kind of documents, L rs,arch = limit the number of results and search found, P$R sic = Precision and recall usually have strong inverse correlation, B qucx = Complexity in making Boolean queries, V sm = Vector space model, B shi = Basic vector model, E bm = extended Boolean model, L si = LSI, D prg = Dynamic programming, S nart = SMART, Dc tree = Decision tree, C hrt = CHART, C had = CHAID, M rs = MARS, IBM sm = IBMspaas modeler, M sql2 = Microsoft sql server 2, H n = handle both numerical and categorical data, L tpp = Little data preprocessing is needed, !OR np-hard = Does not provide optimal results in NP-hard problems,
Choosing size of area = 
Helpful to remove redundant 
useful to obtain better 
learning -

= Useful in if then and optimization problems, C_{opt} = complexity in computation, E_{ex} = Ensemble classifier, In_{ss} = increased storage space, U_{pp} = useful to obtain better predictive performance, M_{lr} = Machine learning, B_{sm} = Boosting, B_{bg} = bagging, B_{bs} = bayesian model classifier, R_{em} = Regular expression matching, F_{es} = feature extraction algorithm, R_{cm} = Recommended Systems, B_{mu} = biometics, Hl_{RSP} = Helpful to remove redundant and pruned data, PCA = Principal component Analysis, Sn_{n} = Sensitive to noisy features, B_{b} = Biased towards the class with larger samples, N_{n} = Neural network, P_{pl} = Perceptron, A_{nlp} = AutoMLP, Wsa = Wake sleep algorithm, R_{bt} = Robotics, AI = Artificial intelligence, Ab_{dc} = ability to detect complex relationships, Hl_{Al} = Helpful in artificial intelligence, N_{lpe} = Needs expertise, Diff_{beg} = providing training is a challenging task, P_{m} = Partition methods, B_{r} = BIRCH, C_{r} = CURE, PAM = Partition Around Medoids, C_{Rosa} = CLARANS, ScI = easily scalable, Su_{mf} = Suitable for datasets with compact spherical clusters that are well-separated, Ib_{sc} = In Inability to deal with non-convex clusters of varying size and density, Sn_{ADO} = High sensitivity to initialization phase, noise and outlier, H_{m} = Hierarchal method, DIANA, AGNES, F_{s} = Feature selectionL_{dh} = Landuse detection, W_{up} = Web user profiles, Hist_{col} = color histograms, Ib_{mr} = Inability to make corrections once the splitting/merging decision is made, D_{p} = Density based methods, DBSCAN, DENCLUE, LOF, OPTICS, Res_{rLO} = Resistance to noise outliers, UC_{cs} = Useful only in convex clusters if similar size and density, !Det_{lcs} = Can't detect intrinsic cluster structure.
VISUALIZATION OF RESULTS

Figure 5 shows the reading of data set. At this step, a ‘csv’ file of data set has been imported. Another file formats can be arff, excel etc.

Figure 6 shows how to remove noisy values. At this step, A filter of select attributes has been applied. In our experiment nominal data is used.

Figure 5. Reading the data set.

Figure 6. Applying the ‘selection of attributes’.
Figure 7. Implementing the Id3 algorithm.

Figure 8. Applying the model.

Figure 7 shows the Implementing The Id3 Algorithm. After selecting the required values id3 algorithm has been implemented.

Figure 8 shows the model building process. If the user wants to generate a tree from the provided data set then a model building is required. At this stage a model has been prepared.
Figure 9. Implementing cross fold validation.

Figure 10. Training and testing modeling.

Figure 9 shows the cross fold validation. At this stage A 10 fold validation which actually refers to number of iterations used for the building of tree.

Figure 10 shows the training and testing modeling. At this stage training and testing module is generated. These come under the validation part.
Figure 11. Running of the process.

Figure 12. Tree generation of ID3 algorithm.

Figure 11 shows the running of the process. At this stage we will connect the port to resultant port. Figure 12 shows the tree generation of algorithm. This tree is generated from id3 from the given dataset and a resultant node shows the two different schools.
Figure 13. Performance evaluation.

Figure 14. Performance vector for ID3.

Figure 13 shows the evaluation of precision, recall and accuracy performance filter has been applied. Figure 14 shows the performance of id3 algorithm.
Figure 15. Modeling the CHAID Algorithm.

Figure 16. Tree generation using CHAID.

Figure 15 shows the modeling of CHAID algorithm. Figure 16 shows the resultant tree generation from the CHAID algorithm.
Figure 17. Performance vector for CHAID.

Figure 18. Building Naïve Bays model.

Figure 17 shows the performance parameters of algorithm.  
Figure 18 shows the modeling of naïve bays algorithm.
Figure 19. Performance vector for Naïve Bays.

Figure 20. Visualization of results using parameter 'Mjob'.

Figure 19 shows the performance vector of naïve bays algorithm. Figure 20 shows an Andrew curve form the parameter 'Mjob'. It shows the profession of those mothers whom children are studying in the school.
Figure 21. Parameters 'reason.'

Figure 22. Parameter 'Fjob.'

Figure 21 shows a distribution curve of reason behind choosing the school. Figure 22 shows the status of father's occupation.
Figure 23. Parameter ‘Reason’ (mother).

Figure 24. Admission status between schools.

Figure 23 shows the reasons behind mothers are opting school for their Children.

Figure 24 Shows a comparison of admission in the schools.
Figure 25. Parameter 'higher studies'.

Figure 26. Parameter 'performance'.

Figure 25 Predicts how many students wants to go for higher studies form both schools.  
Figure 26 shows the prediction of which school is performing better.