ABSTRACT

The demand for data mining is increasing day by day, with the enormous increase in the amount of data generated by the industries, market, and other sectors. As per the current trends, when most of the industries have already shifted or have planned to shift towards digitization, we require a data mining technique or algorithm which can handle such large data. The issues with the existing data mining algorithms are that they are very complex to understand and also take comparatively high time in execution. The primary issue is that association rule mining from such a large data with the existing data mining algorithms is having some constraints. These constraints include the complexity of existing algorithms, slow execution time, less accuracy, and low efficiency with large datasets.

The association rule mining techniques is been helpful in predicting the behaviouristic pattern of consumers buying pattern that can be used in effective marketing techniques. It performs vital role in the essential areas of data mining. It poses many exciting problems for the enlargement of well-organized and effective techniques. After taking a closer look at its applications, we discover that the application of association rules needs much more investigations in order to aid in more unambiguous objectives. Applications of association rule mining are: In Large database - logistics, marketing, defence, public safety, Industries, Medical diagnosis, Hospital, Medical shops, scan centres.

With the prodigious proliferation of ginormous-scale data depots, a need for incorporating the empirical techniques of data mining with the effectiveness of commensurate systems to intuitively manage cosmic volumes of data has now risen. To quell these obstructions to managing data efficiently, in this document we present a new algorithm name as DAEM based on ANN for such mining activities, which overcomes the problems in the existing algorithms of data mining apropos of their execution time and interestingness and to prove its efficiency the new algorithm will be compared to the two popular data mining algorithms that are Apriori algorithm and FP Growth. The Apriori algorithm first considers every single item as an itemset and counts the support from their frequency in the database, and then captures those who have support equal to or more than the minimum support threshold. In this process extraction of each frequent itemset requires the scanning of the entire database by the algorithm until no more itemsets are left with more than the minimum threshold support. In the FP-Growth algorithm, the algorithm represents the data in a tree structure. It is a lexicographic tree structure that we call the FP-tree.

Which is responsible for maintaining the association information between the frequent items. After making the FP-Tree, it is segregated into the set of conditional FP-Trees for every frequent item. A set of conditional FP-Trees further can be mined and measured separately. For that, the table of every conditional FP-Tree, Based on that conditional table will be created and Based on that the rules for frequent itemsets are been generated.

The DAEM algorithm provides faster execution time and more accurate results as per the observation shown in interestingness compare to Apriori and FP Growth. It majorly used the concept of ANN for finding the rules between the itemsets. Although it generates interesting pattern for meaningful information from such a vast dataset and a wide number of transactions rules. In today's world almost every shops, companies, industries are digitization, so they need some sort of a mining mechanism to extract meaningful information and draw the pattern of their consumer's behaviour. This proposes a technique that will generate strong association rules from a large dataset with the help of DAE (Denoising Autoencoder) techniques.

LIST OF ABBREVIATIONS

AI: Artificial Intelligence

ANN: Artificial Neural Network

ANS: Artificial Neural Network System

ARM: Association Rule Mining

ARs: Association Rules

DAE: Denoising Autoencoder

DAEM: Denoising Autoencoder Mining

DM: Data Mining

FP-stream: Frequent Pattern-Stream

KDD: Knowledge Discovery in Database

OLAP: On-Line Analytical Processing

Minconf: Minimal Confidence

Minsupp: Minimal Support

ML: Machine Learning

PDP: Parallel Distributed Processing

PR: Pattern Recognition

WMFP-SW: Weighted Sliding Window Mining

TABLE OF CONTENTS

DECLARATION……..i

CERTIFICATE…….....ii

ACKNOWLEDGEMENT……iii

ABSTRACT……...……...iv

LIST OF ABBREVIATIONS……..vi

TABLE OF CONTENTS…….……vi

LIST OF FIGURES……..ix

LIST OF TABLES……....x

		Page No.
Chapter 1	**INTRODUCTION**	**1-24**
1.1	Evolution of Data Mining	2
1.2	Data Mining	5
1.2.1	Data Mining Methods	6
1.3	Knowledge Discovery in Database	6
1.3.1	KDD vs DM	7
1.3.2	Stages of KDD	7
1.4	DBMS vs DM	9
1.5	Learning Types	10
1.6	Data Mining Tasks	13
1.6.1	Descriptive Model	13
1.6.2	Predictive Model	15
1.7	Measures for Association Rule Mining	17
1.8	Quality measures and range of feasible values	20
1.9	Comparative study of different DM algorithms	21
1.10	Artificial Neural Network	22
Chapter 2	**LITERATURE REVIEW**	**25-39**
Chapter 3	**PROBLEM DEFINITION**	**40-41**

3.1	Problem Formation	40
3.2	Objective	40
3.3	Requirement	40
Chapter 4	**PROPOSED MODEL**	**42-53**
4.1	Apriori algorithm	43
4.1.1	Steps to find out the Strong Rules in Apriori algorithm	44
4.2	FP Growth algorithm	46
4.2.1	Steps to find out Strong Rules in FP Growth algorithm	48
4.3	DAEM algorithm	50
4.3.1	Principle of DAEM algorithm	50
4.3.2	Steps to implement DAEM algorithm	53
Chapter 5	**RESULTS AND ANALYSIS**	**54-64**
5.1	Data Visualization	54
5.1.1	Data Preview	54
5.1.2	Most frequent items in Grocery dataset	56
5.1.3	Most frequent items in Supermarket dataset	57
5.2	Association Rules obtained by different algorithms	57
5.3	Runtime Analysis	59
5.4	Interestingness Analysis	62
Chapter 6	**CONCLUSION AND FUTURE SCOPE**	**65-66**
6.1	Conclusion	65
6.2	Future Scope	66
	References	66
	List Of Publications	71
	Curriculum Vitae Of The Student	72
	Student Research Paper	77

LIST OF FIGURES

Fig 1.1: Stages of KDD process……..7

Fig 1.2: Overview of Supervised Learning……11

Fig 1.3: Overview of Unsupervised Learning……11

Fig 1.4: Overview of Unsupervised Learning……12

Fig 1.5: DM Model and its Tasks……...12

Fig 1.6: Artificial Neuron,

Fig 1.7: Multilayered Artificial Neural Network……23

Fig 1.8: Architecture of Artificial Neural Network……....24

Fig 4.1: FP Tree for the above transaction table…….....48

Fig 4.2: Working of an Autoencoder…….....50

Fig 4.3: Architecture of Denoising Autoencoder……50

Fig 5.1: Most frequent items in Grocery dataset…….....53

Fig 5.2: Most frequent items in Supermarket dataset……..54

Fig 5.3: Run time analysis for different Support values of Grocery dataset……...57

Fig 5.4: Run time analysis for different Support values of Supermarket dataset……58

Fig 5.5: Performance analysis of the algorithms based on the number of rules for the Grocery dataset……59

Fig 5.6: Performance analysis of the algorithms based on a number of rules for the Supermarket dataset……59

Fig 5.7: Average Interestingness analysis of rules for different Support values of Grocery Dataset……....60

Fig 5.8: Average Interestingness analysis of rules for different Support values for Supermarket Dataset……60

LIST OF TABLES

Table 1.1: The Transactional dataset containing ten transactions and eight items…….17

Table 1.2: Quality measures and range of feasible values…….....21

Table 1.3: A comparative study of algorithms used in DM……...22

Table 4.1: Transactional dataset of supermarket…….....44

Table 5.1: Parameters of Selected Datasets……52

Table 5.2: Grocery dataset……..52

Table 5.3: Supermarket dataset…….....53

Table 5.4: Describing the grocery dataset using describe() function…….....53

Table 5.5: Describing the Supermarket dataset using describe() function……..53

Table 5.6: Rules obtained by Apriori algorithm on Grocery dataset……...54

Table 5.7: Rules obtained by FP Growth algorithm on Grocery dataset…….55

Table 5.8: Rules obtained by DAEM algorithm on Grocery dataset……...55

Table 5.9: Rules obtained by Apriori algorithm on Supermarket dataset……55

Table 5.10: Rules obtained by FP Growth algorithm on Supermarket dataset……56

Table 5.11: Rules obtained by DAEM algorithm on Supermarket data……..56

Table 5.12: Run time (second) analysis of the algorithms for different Support values for both datasets.....57

CHAPTER 1

INTRODUCTION

CHAPTER 1
INTRODUCTION

The inception of computing technology has conspicuously enticed our lives and, the two consequential impacts of this effect were: Business Data Processing and Scientific Computing. In the initial years, of the enhancement of computer techniques for business, computer professionals were apprehensive mainly about the pretension of circumscribing files to store the data so that information could be easily recuperated. There were limitations on ambry size for stowing data and on the speed of gleaning the data? It's extraneous to say, that the happenings were demarcated to a very few, highly qualified professionals. Then came longevity when the task was foundational by a DBMS. The accountability of byzantine tasks, such as declarative aspects of the retrospectives was laid down on to the database administrator and the user could pretence his query in more facile languages such as the query languages. Thus, almost any business whether small, medium, or immense scale, started using computers for everyday tasks. The scenario on the hardware side was encouraging too. The storage cost and machine cost were drastically reduced so that almost any business house could afford a standard machine. Thus, due to comprehensive computerization and also due to bargainous ambry facilities, there was a prodigious amount of wealth of information embedded in massive databases belonging to different enterprises. Every organization was now escalating an immense volume of everyday transaction details and keeping them as chronicled files. As a result, proles and proles of data in MBs, GBs, and TBs, were accumulating in the electronic vaults of companies, governments, and research institutions.

Business is instinctually antagonistic and in this antagonistic world of business, one is unrelentingly on the citadel for ways to beat the competition. A question that undeniably arose is whether the immense amount of data that is prompted and accumulated as repositories can be used for elevating the productiveness of business performance.

In the area of scientific computing, the ponderous problem was to induce some valuable information from the observed data. Thus the use of computing technology has bucked up researchers to collect very enormous expanses of data. Furthermore, the development of other scientific disciplines leads the community to collect such enormous expanses of data effortlessly. Some typical examples were remote-sensing data, as satellites were streaming in an enormous amount of remote sensing data every day. In the area of health science, the repository of protein data and genome data was an invaluable source of scientific

experiments. On the other side, this leads to drowning in oceans of text data and data of other media that are generated from the web.
The biggest that arose was- “What use is all this data?”

1.1 Evolution of Data Mining

Up to the early 1990s, the answer to the last question was ‘not much. No one was interested in utilizing data that was accumulated during the process of daily activities. Once the transaction processing is over, those data are dumped into archival files.

These collections of data, whether their origin is a business enterprise or a scientific experiment, have recently spurred a tremendous interest in the areas of knowledge discovery and Data Mining (DM). Statisticians and data miners now have faster analysis tools that can help sift and analyse the stockpiles of data, turning up valuable and often surprising information. As a result, a new discipline in Computer Science, DM, gradually evolved. DM is the exploration and analysis of large data sets, to discover meaningful patterns and rules. The key idea behind this was to find effective ways to combine the computer’s power to process data with the human eye's ability to detect patterns. The techniques of DM were designed for, and work best with large data sets.

The evolution of DM began when business data was first stored in computers and technologies were generated to allow users to navigate through the data in real-time. DM takes this evolutionary process beyond retrospective data access and navigation, to prospective and proactive information delivery. This evolution is due to the support of three sufficiently mature technologies massive data collection, high-performance computing, and DM algorithms. The core components of the DM technologies have been under development for decades, and today the maturity of these techniques coupled with the high-performance relational database engines and broad data integration efforts have made these techniques practically applicable.

DM is a component of a wider process called knowledge discovery from databases. It involves scientists from a wide range of disciplines, including mathematicians, computer scientists, and statisticians, as well as those working in fields such as machine learning, artificial intelligence, information retrieval, and pattern recognition. Before a data set can be mined, it first has to be cleaned. This removes errors, ensures consistency, and takes missing values into account. DM may use quite simple statistical techniques or it may use highly

sophisticated data analysis. What is new for data miners is the employment of these techniques on vast quantities of data.

Business enterprises are beginning to realize that information on customers and their buying patterns is the most valuable information. Competitiveness increasingly depends on the quality of decision-making and learning to make quality decisions from past transactions and decisions. The large databases maintained (generated during the routine operations) by retailers, telecom service providers, and credit card companies contain very valuable information related to customers. These enterprises could benefit immensely in the areas of marketing, advertising, and sales if interesting and previously unknown customer buying and calling patterns can be discovered from the large volumes of data. Several factors have contributed to bringing DM to the forefront. To identify a few of these factors:

- The untapped value in large databases.
- The concept of Data Warehousing.
- The dramatic drop in the cost/performance ratio of hardware systems.

The most important of these three is the development of Data Warehousing technology. Let us take a few minutes to review the layered architecture of the modern database management systems and the modern applications of these DBMSs. The different layers can be viewed as follows:

1. **External Interface-** This layer deals with the concepts of data definition language, data manipulation language and host-language interfaces.
2. **Language Processing-** Query processing, algebraic optimization, and plan creation are the components of this layer.
3. **Code Generation-** The procedural part is handled in this layer in the form of index structures, and the implementation of algebraic operations.
4. **Transaction Management-** The important aspects of modern DBMS are concurrency control, logging and recovery.
5. **Storage Management-** This layer is concerned with memory and buffer management, secondary storage and device management.

A major change in the above approach is Data Integration. Data integration of data from multiple sources is the key term in modem database applications. Modern applications of

database technology demand integrated access to multiple databases. Typical applications where data integration is relevant, arise from the merging of different companies, accessing healthcare data from different institutions, and data access through web browsing. Recent developments in e-commerce also demand data integration. Data integration can be carried out in two modes, virtual or materialized. The virtual modes of data integration adopt a search engine to collect data from multiple sources, or a mediated model to decompose the queries to each of the sources and uniformly present the result. The materialized mode of data integration can be either through a universal DBMS or a data warehouse. In a universal DBMS, data are integrated by physically migrating data from different sources to a new database in OO-DBMS or OR-DBMS, The data warehouse collects data from different sources for OLAP and DM applications. Among all these modes of integrations, data warehousing is considered to be the most efficient for decision support applications. This is particularly relevant in the contexts of DM and On-Line Analytical Processing (OLAP).

As the terms “DM” and “Data Warehousing” are becoming popular, the last few years have witnessed the emergence of extremely innovative and elegant techniques for DM and warehousing. Most of these results are presented at different international fora. However, more and more researchers from other areas are interested in knowing about the subject. The present work aims at providing in one place the most recent and important techniques of DM and warehousing.

When an enterprise has made substantial investments in Capturing, cleaning, storing, and managing data for decision support and has built a data warehouse, it is in a position to generate a range of reporting applications through ad hoc query capabilities, visualization, and drill-down functionalities via OLAP tools. DM complements all of these, while also adding value to existing investments by enabling people who understand the warehousing environment (and others) to exploit data assets and gain deeper insights into key aspects of business performance. For example, a marketing department might have well-established models for modelling customer profiles. DM helps the marketing analyst to break free of established customer models and construct detailed customer profiles reflecting actual behaviour in different situations. In a marketing situation, such profiles can be used as a basis for enabling promotion costs to be cut and response ratios to be increased. The net impact of an effective DM techniques can be very high as it reduce the time delay of generating interesting patterns and ARs in dataset.

1.2 Data Mining

DM is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. With the widespread use of databases and the explosive growth in their sizes, organizations are faced with the problem of information overload. The problem of effectively utilizing these massive volumes of data is becoming a major problem for all enterprises. Traditionally, we have been using data for querying a reliable database repository via some well-circumscribed application or canned report-generating utility. While this mode of interaction is satisfactory for a large class of applications, there exist many other applications which demand exploratory data analyses. We have seen that in a data warehouse, the OLAP engine provides an adequate interface for querying summarized and aggregated information across different dimension hierarchies. Though such methods are relevant for decision support systems, these lack the exploratory characteristics of querying. The OLAP engine for a data warehouse (and query languages for DBMS) supports query-triggered usage of data, in the sense that the analysis is based on a query posed by a human analyst. On the other hand, DM techniques support the automatic exploration of data. DM attempts to source out patterns and trends in the data and infers rules from these patterns. With these rules, the user will be able to support, review and examine decisions in some related business or scientific area. This opens up the possibility of a new way of interacting with databases and data warehouses. Consider, for example, a banking application where the manager wants to know whether there is a specific pattern followed by defaulters. It is hard to formulate a SQL query for such information. It is generally accepted that if we know the query precisely we can turn to query language to formulate the query. But if we have some vague idea and we do not know the precise query, then we can resort to DM techniques.

The evolution of DM began when business data was first stored in computers, and technologies were generated to allow users to navigate through the data in real-time. DM takes this evolutionary process beyond retrospective data access and navigation, to prospective and proactive information delivery. This evolution is due to the support of three sufficiently mature technologies, massive data collection, high-performance computing, and DM algorithms.

So in most simple words, this can be described as, it is the search for the relationships and global patterns that exist in large databases but are hidden among vast amounts of data, such as the relationship between patient data and their medical diagnosis. This relationship

represents valuable knowledge about the database, and the objects in the database, if the database is a faithful mirror of the real world registered by the database.

1.2.1 Data Mining Methods

DM, defined to be the method that extracts usable insights from the set of larger data sets, of collection of raw information, entails the analysis of the patterns hidden in the information. So it is used in the wider applications of belonging to different areas like science, research, business, industries, etc. So DM can be viewed as the optimum method of knowledge discovery and warehousing along with the computing, to provide better insight leading to improved decision making.

Some of the key features that are involved in DM methods are listed below:

- **i.** Flow and attitude analysis by pattern recognition,
- **ii.** Development of decision-adapted information,
- **iii.** Prediction from the similar outcomes,
- **iv.** Analysis of large data sets,
- **v.** Detection of hidden facts using the clustering.

1.3 Knowledge Discovery in Database

Knowledge Discovery in Database (KDD) was formalized in 1989, concerning the general concept of being broad and high level in the pursuit of seeking knowledge from data. The term DM was then coined; this high-level application technique is used to present and analyse data for decision-makers. DM is only one of the many steps involved in knowledge discovery in databases. The various steps in the knowledge discovery process include data selection, data cleaning and pre-processing, data transformation and reduction, DM algorithm selection, and finally post-processing and the interpretation of the discovered knowledge. The KDD process tends to be highly iterative and interactive. DM analysis tends to work up from the data and the best techniques are developed with an orientation towards large volumes of data, making use of as much data as possible to arrive at reliable conclusions and decisions. The analysis process starts with a set of data, and as a methodology to develop an optimal representation of the structure of data, during which knowledge is acquired. Once knowledge is acquired, this can be extended to large sets of data on the assumption that the large data set has a structure similar to the simple dataset.

1.3.2 KDD vs. DM

Fayyad et al. distinguish between KDD and DM by giving the following definitions.

"KDD is the process of identifying a valid, potentially useful, and ultimately understandable structure in data. This process involves selecting or sampling data from a data warehouse, cleaning or pre-processing it, transforming or reducing it (if needed), applying a DM component to produce a structure, and then evaluating the derived structure.

DM is a step in the KDD process concerned with the algorithmic means by which patterns or structures are enumerated from the data under acceptable computational efficiency limitations."

Thus, the structures that are the outcome of the DM process must meet certain conditions so that these can be considered knowledge. These conditions are validity, understand ability, utility, novelty, and interestingness.

1.3.3 Stages of KDD

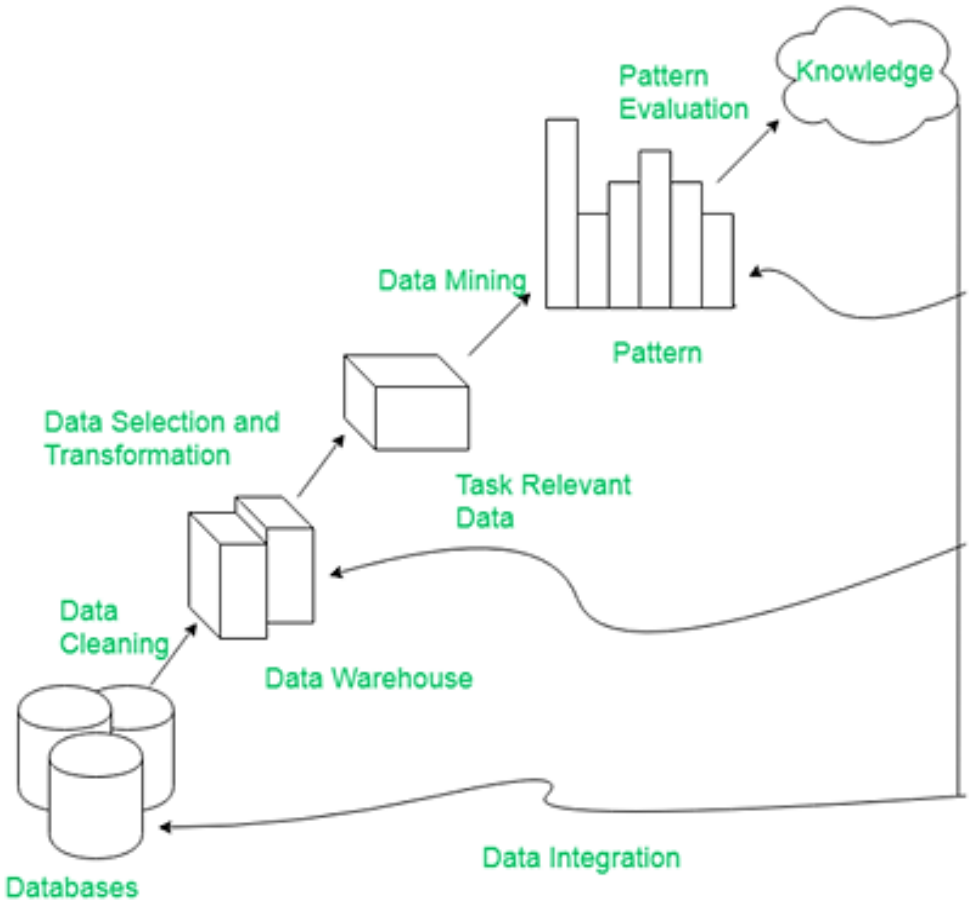

Fig 1.1: Stages of KDD process

The stages of KDD as shown in Fig 1.1, starting with the raw data and finishing with the extracted knowledge, are discuss below.

i. Selection

This stage is concerned with selecting or segmenting the data that are relevant to some criteria. For example, for credit card customer profiling, we extract the type of transactions for each type of customer and we may not be interested in the details of the shop where the transaction takes place.

ii. Pre-processing

Pre-processing in the data cleaning stage where unnecessary information is removed. When the data is drawn from several sources, it is possible the same information is represented in different sources in different formats. This stage reconfigures the data to ensure a consistent format, as there is a possibility of inconsistent formats

iii. Transformation

The data is not merely transferred across, but transformed in order to be suitable for the task of DM. In this stage, the data is made usable and navigable.

iv. Data Mining

This stage is concerned with the process of discovering meaningful, new correlation patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition techniques as well as statistical and mathematical techniques.

v. Interpretation and Evaluation

The patterns obtained in the DM stage are converted into knowledge, which in turn, is used to support decision-making.

vi. Data Visualization

Data visualization makes it possible for the analyst to gain a deeper, more intuitive understanding of the data and as such can work well alongside DM allowing the analyst to focus on certain patterns and trends and explore them in depth using visualization. On its own, data visualization can be overwhelmed by the volume of data in a database but in conjunction with DM can help with exploration.

Data visualization helps users to examine large volumes of data and detect patterns visually. Visual displays of data such as maps, charts, and other graphical representations allow data to be presented compactly to the users. A single graphical screen can encode as much information as can a far larger number of text screens. For example, if a user wants to find out whether the production problems at a plant are correlated to the location of the plants, the problem locations can be encoded in a special colour, say red, on a map. The user can then discover locations in which the problems are occurring. He may then form a hypothesis about why problems are occurring in those locations and may verify the hypothesis against the database.

1.4 DBMS vs DM

We have seen that DBMS supports query languages which are useful for query-triggered data exploration, whereas DM supports automatic data exploration. If we know exactly what information we are seeking, a DBMS query would suffice; whereas if we vaguely know the possible correlations patterns, then DM techniques are useful. One of the tasks of DM is hypothesis testing, wherein we formulate a hypothesis and test it by sifting through the database. This task can be handled by a DBMS query. Thus, in these series, DBMS supports some primitive DM tasks.

From the architectural perspective, let us look at the way a DM system makes use of an existing database.

There are three different ways in which DM systems use a relational DBMS. They may not use it at all, be loosely coupled or tightly coupled. A majority of DM systems do not use any DBMS and have their memory and storage management. They treat the database simply as a data repository from which data is expected to be downloaded into their memory structures before the DM algorithm starts. The advantage of such an approach is that one can optimize the memory management specific to the DM algorithm. On the contrary, these systems ignore the field-proven technologies of DBMS, such as recovery, concurrency, etc.

The second approach is to have a loosely-coupled DBMS. In this case, DBMS is used only for the storage and retrieval of data. For instance, one can use a loosely coupled SQL to fetch data records as required by the mining algorithm. The front-end of the application is implemented in a host programming language, with embedded SQL statements in it. The applications use a SQL select statement to retrieve the set of records of interest from the

database. A loop in the application program copies records in the result set one by one from the database address space to the application address space, where computation is performed on them. This loosely-coupled approach does not use the querying capability provided by the DBMS In the tightly-coupled approach, the portions of the application programs are selectively pushed to the database system to perform the necessary computation. Data is stored in the database and all processing is done at the database end. It is different from bringing the data from the database to the DM area. On the other hand, the DM application goes where the data naturally resides. This avoids performance degradation and takes full advantage of database technology. The performance of this approach depends on the way to optimize the DM process while mapping it to a query. There are two suggested approaches. We can leave the optimization task to a built-in query optimizer of the DBMS or we can have an external optimizer.

1.5 Learning Types

Learning is the automation of a learning process and learning is tantamount to the construction of rules based on observations. This is a broad field that includes not only learning from examples but also reinforcement learning, learning with a teacher, etc. A learning algorithm takes the data set and its accompanying information as the input and returns a statement, e.g., a concept representing the results of learning as output.

Inductive learning, where the system infers knowledge itself by observing its environment, has three main strategies. Supervised Learning, Unsupervised Learning, and Reinforcement Learning. The model produced by inductive learning methods can be used to predict the outcome of future situations; in other words, not only for states encountered but rather for unseen states that could occur. There can be many possible models from all given sets of examples. In such situations, the principle of Occam's razor is very appropriate. It states that if there are multiple explanations for a particular phenomenon, it makes sense to choose the simplest because it is more likely to capture the nature of the phenomenon.

i. Supervised Learning

Supervised learning means learning from examples, where a training set is given that acts as examples for the classes. The system finds a description of each class. Once the description (and hence a classification rule) has been formulated, it is used to predict the class of previously unseen objects. This is similar to discriminate analysis which occurs in statistics.

Input examples are categorized into a known set of classes.

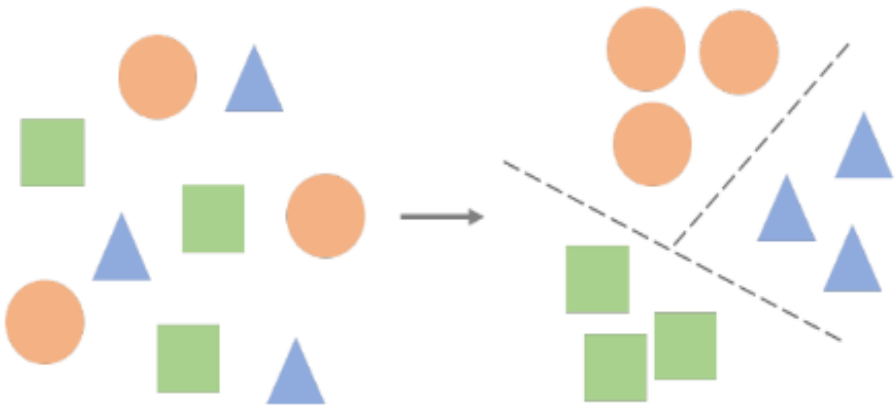

Fig 1.2: Overview of supervised learning

ii. Unsupervised Learning

Unsupervised learning is learning from observation and discovery. In this mode of learning, there is no training set or prior knowledge of the classes. The system analyses the given set of data to observe similarities emerging out of the subsets of the data. The outcome is a set of class descriptions, one for each class, discovered in the environment. This is similar to cluster analysis in statistics. Input samples are grouped into clusters based on the underlying patterns.

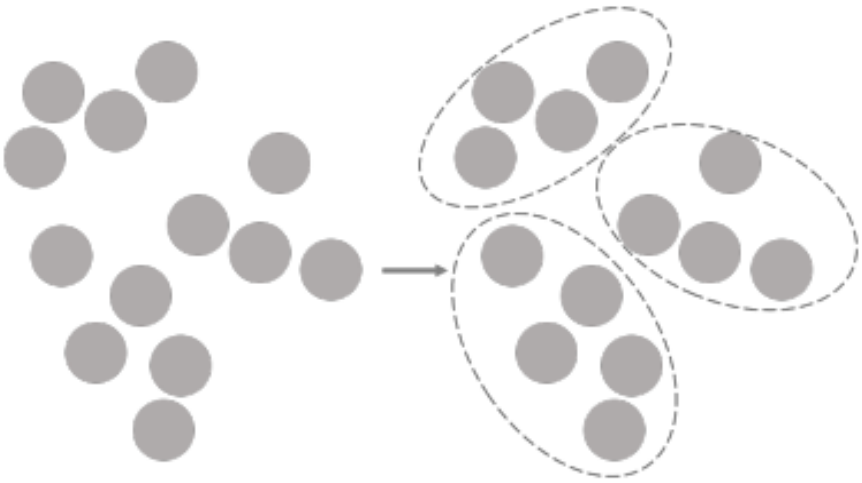

Fig 1.3: Overview of unsupervised learning

iii. Reinforcement Learning

Reinforcement learning is applied when the task at hand is to make a sequence of decisions towards a final reward. During the learning process, an artificial agent gets either rewards or penalties for the actions it performs. Its goal is to maximize the total reward.

Example - It includes learning agents to play computer games or performing robotics tasks with an end goal.

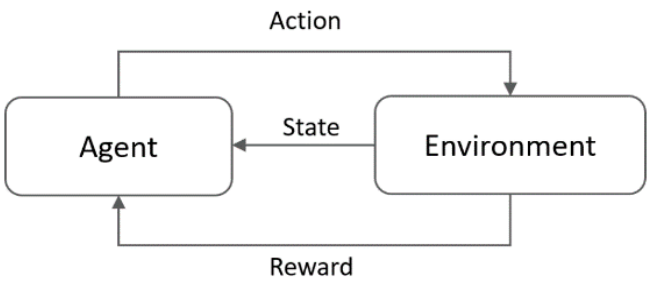

Fig 1.4: Overview of reinforcement learning

An agent observes the environment state and performs actions to maximize an overall reward.

1.6 Data Mining Tasks

DM can often be categorized into two categories based on what a particular project is trying to accomplish. Those two categories are the descriptive model and the predictive model. There are several DM functions, such as ARM, Time Series Analysis, prediction, Neural Network, etc. Either of these functions falls under the predictive model or descriptive model. The DM system can perform one or more of the above functions as part of the DM.

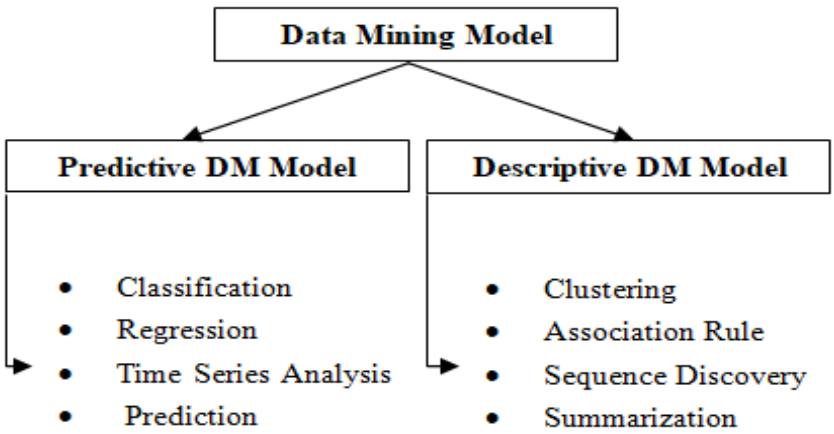

Fig 1.5: DM Model and its Tasks

1.6.1 Descriptive Model

The descriptive model defines a domain that represents a way that can be described or find the relationship among data. It can be used for many purposes. It may include ethical, structural, and other definitions that establish reasonable relationships about the system, such as its component relationships, the interactions between its components, and the distribution of its ethical properties to structural elements. Descriptive models are not usually constructed in a way that directly supports imitation, animation, or performance, but can be considered in compliance with grammatical rules, and logical relationships can be considered for them. This model relies upon an unsupervised learning approach. Some of the major tasks of descriptive models are as follows

i) Association Rule,
ii) Clustering,
iii) Sequence Discovery, and
iv) Summarization.

i) Clustering

Clustering is the method of converting a group of abstract objects into classes of similar objects. It is a method of partitioning a set of data or objects into a set of significant subclasses called clusters. A cluster is a subset of similar objects. A subset of objects such that the distance between any of the two objects in the cluster is less than the distance between any object in the cluster and any object that is not located inside it. It helps users to understand the structure or natural grouping in a data set and is used either as a stand-alone instrument to get a better insight into data distribution or as a pre-processing step for other algorithms.

ii) Association Rule Mining

Association Rule Mining (ARM) is an event for determining organizations, patterns, and relationships between sets of objects on a dataset. The law of association is of the form M ⟶N [support, confidence]. The support and confidence are the primary measures that are used for assurance of the rule. Association Rule is said to be strong if it satisfies both the minsupp (minimal support) and the minconf (minimal confidence) that is defined by the user. These Association Rules are easy to establish due to the small database but become more

complex as the database transforms. Some general values and concepts are required in order to better understand DM in large data sets. A set contains n items and is called an n-item set. So set {A, B} is a set of 2- item set. Based on the frequency of item sets, the number of active functions is calculated.

The dilemma of mining Association Rules can be fragmented into two sub-dilemmas.

a) Find all sets of item sets whose support is greater thanσ. These itemsets are known as frequent item sets.

b) Use these frequent item sets to develop the desired rules. The conventional ideology is that if, say p, q, r, s are frequent item sets, than we can decide on the rule $pq \longrightarrow rs$ that holds by checking the following inequality

$$\frac{s(\{p,q,r,s\})}{n} \geq \sigma$$

where, n is the total number of transactions and $\sigma$is a minsupp.

iii. Sequence Discovery

Sequence discovery, or sequential pattern mining, is a DM technique that discovers statistically relevant patterns in sequential data. This mining program evaluates certain criteria, such as occurrence frequency, duration, or values in a set of sequences to find interesting hidden patterns. It uncovers relationships among data. This technique defines a sequential pattern of events and actions. For example, a customer who buys more than twice in the first quarter of the year may be likely to buy at least once during the second quarter. The data miner simply provides a sequence database and chooses a search parameter called the minimum support threshold. This specifies the minimum number of sub-sequences (called support) in which a pattern must appear to be considered relevant. The results are presented in a column format, with the specific pattern on the left and the number of subsequent instances (supports) on the right.

iv. Summarization

Summarization is a mining technique that is completely based on the extraction of data. It is obtained by identifying attributes such as customer name, contact, etc. with many different values and by removing them or performing a threatening function. Also, we can use standard statistics in data to represent its summary. It is a key DM concept that involves techniques for finding a compact description of a dataset. Simple summarization methods such as tabulating

the mean and standard deviations are often applied for data analysis, data visualization, and automated report generation.

1.6.2 Predictive Model

The Predictive-model is a mathematical method that is generally used by DM technologies, to define modelling prediction as the process of predicting the subsequent time behaviour of an analysis of factual and recent data. A predictive model makes theories based on what has previously occurred and what is currently occurring. If new data shows that the current situation has changed, the likelihood of the outcome must be recalculated. This model relies upon the supervised learning approach. Some major tasks of predictive models are as follows

- **i.** Regression,
- **ii.** Classification,
- **iii.** Prediction, and
- **iv.** Time Series Analysis.

i. Classification

It is the process of DM that assigns objects to a group into specific categories or classes. The main goal of this mining technique is to accurately predict the target category in each case data. The principle of classification is to accurately predict the target class in each case data. It is used to separate each item by a set of data into a predefined set of classes or groups. The division of the data analysis function is where the model or separator is designed for predicting category labels. Sometimes there can be more than two classes to classify. It allows you to organize data sets of all sorts, including complex and huge datasets as well as small and simple ones. This technique also aims to connect a variable of interest with the required variables.

ii. Regression

Regression is a mining technique that is used in the prediction of a range of numeric values (also called continuous values) from a particular dataset. It refers to a type of learning technique that can be used to predict any continuous-valued attribute. Regression is used in various industries for business and marketing behaviour, trend analysis, and financial forecast. In a business organization, it helps by analysing the target variable and predictor variable relationships. Regression involves the technique of fitting a straight line or a curve

on numerous data points and happens in such a way that the distance between the data points and cures comes out to be the lowest. The most popular types of regression are linear and logistic regressions. Other than that, many other types of regression can be performed depending on their performance on an individual data set.

iii. Time Series Analysis

In a dataset where the observations are ordered by time, this time-ordered sequence of observations is called time-series. Examples of such datasets are numerous, such as daily closing stock prices, monthly unemployment figures, quarterly crime rates, and annual birth rates. The Fundamental characteristic of a time series is that its observations are correlated. And the most standard statistical methods based on random samples are not applicable, and therefore different methods are needed. Thus the body of statistical methods for analysing time series is referred to as time series analysis.

Time series analysis includes the time domain approach and frequency domain approach. The time-domain approach uses time functions such as the autocorrelation function (ACF) and the partial autocorrelation function (PACF) to describe the characteristics of a time series process whose evolution is represented through various time-lag relationships. Whereas, the frequency domain approach, uses a spectral function to study how the variation of a time series may be accounted for by the mixture of sines and cosines at various frequencies in the dataset.

iv. Prediction

Prediction is used to find a numerical output same as in classification, the training dataset contains the inputs and corresponding numerical output values. The algorithm derives the model or a predictor according to the training dataset. The model should find a numerical output when the new data is given. Unlike in classification, this method does not have a class label. The model predicts a continuous-valued function or ordered value. Regression is generally used for prediction. Predicting the value of a house depending on the facts such as the number of rooms, the total area, etc., is an example of prediction. Another example is; that suppose the marketing manager needs to predict how much a particular customer will spend at his company during a sale. We are bothered to forecast a numerical value in this case. Therefore, an example of numeric prediction is the data processing activity.

1.7 Measures for Association Rule Mining

Frequent Set is a transactional database and σ is the "minimum support threshold" specified by the user or domain experts. An itemset is frequent if they satisfy the min support threshold,

$$s(M)_T \geq \sigma$$

Maximal Frequent Set In order to be a "maximal frequent set", a frequent set must be recurring and no superordinate of it must be recurring.

So now, let us assume the value of minsupp = 0.25 and minconf = 0.5.

Table 1.1: The Transactional dataset containing ten transactions and eight items

TID	Items
T_1	A_1, A_2, A_3
T_2	A_1, A_2, A_4
T_3	A_1, A_3, A_5
T_4	A_2, A_4, A_6
T_5	A_3, A_6, A_7
T_6	A_1, A_2, A_3, A_4
T_7	A_2, A_3, A_4
T_8	A_3, A_7
T_9	A_2, A_3, A_4, A_8
T_{10}	A_6, A_7

i. Support(s)

Support is measured as the amount of logs that contain M∪N from the entire logs in the database. The proportion for each item is altercation by one, when so ever the item is crossover in dissimilar transactions in a database across the period of scanning.

It is also used to measure the quantity or frequency of an itemset in a database. This measure gives an idea of how common itemset is in all activities. Support can be measured as:

$$Supp(M \longrightarrow N) = P(MN) = Freq\frac{(M,N)}{n}$$

Where- Freq (M, N) = Transaction containing M and N, and n = Total number of transitions.

It helps us to identify the rules that need to be considered for further analysis or not. If support of the rule is greater than minsupp, than find the confidence of the rule.

$$Support\ (A2, A4\ \longrightarrow A3)\ =\ 3/10 = 0.3$$

Thus the support value of items A_2, A_4 & A_3 is 0.3.

ii. Confidence(c)

Confidence is defined as the amount of the numbers of transactions that contains M∪N to the entire logs that contain M, where, if the fraction exceeds the kick-off of confidence, an association rules M⟶N can be obtained. Confidence is an indication of how often the rule is true.

Confidence (M⟶N) in relation to the set of functions n, the part of the functions containing M and N.

$$conf(M \longrightarrow N) = \frac{supp(M \cup N)}{supp\ M}$$

Confidence explains how N is frequently occurring when already buying M. This describes the link between two things. For example, if a person buys jam there is a good chance to buy bread. It is calculated as part of the number of operations where both M and N occur to support the M object.

$$Supp(A_2, A_4 \longrightarrow A_3)\ =\ 3/10\ =\ 0.3$$

$$Supp(A_2, A_4)\ =\ 5/10\ =\ 0.5$$

$$conf\ (A_2, A_4 \longrightarrow A_3)\ =\ 0.3/0.5\ =\ 0.6$$

An association rule M⟶N will be strong only if$conf(M \longrightarrow N) \geq minconf$. Here the value of $conf(A_2, A_4 \longrightarrow A_3)$ is 0.6, which is greater than the minconf. Therefore, the rule can be reasoned as a strong rule because it meet with minconf conditions. But we need to check it further as support and confidence alone cannot be sufficient to find a strong rule. In the above association rule$(A_2, A_4 \longrightarrow A_3)$, support of the consequent $(s(A_3) = \frac{7}{10} = 0.7)$ is greater than the confidence of the rule i.e. 0.6, which is not feasible. Therefore, this may be a case of misleading rule as these are the ones which can be generated from irrelevant datasets.

Furthermore, two additional measures will be needed to solve the situation of misleading rules.

iii. Lift

The Lift can be defined as one of the measures of ARM which defines how far the inter-dependence is in between M and N. The measure, lift is not sensitive to rule i.e. $lift(M \longrightarrow N) = lift(N \longrightarrow M)$. A Lift could be formulated as

$$lift(M \longrightarrow N) = \frac{conf(M \longrightarrow N)}{supp(N)} = \frac{supp(M \cup N)}{supp(M) * supp(N)}$$

So,

$$lift(A2, A4 \longrightarrow A3) = \frac{0.3}{0.5 * 0.7} = 0.35 < 1$$

The fraction of the observance support and the predicted support if M and N are free for each other is known as lift. It has three possible values

- If Lift = 1, the probability of occurrence and outcome are independent of each other.
- If Lift >1, than the itemsets are dependent on each other.
- If Lift < l, tells us that one thing replaces other things, which means one thing has a negative effect on another thing.

  An association rules $M \longrightarrow N$ is engaging if it is strong and$lift(M \longrightarrow N) > 1$.

Here as$lift(M \longrightarrow N) < 1$, the rule is not valid to consider as a strong rule.

iv. Conviction

The conviction is defined as one of the measures of ARM which undertake to analyse the magnitude of execution of the rule, to evaluate the conviction.

Unlike Lift, Conviction is tactful to rule direction i.e. $(conv\ M \longrightarrow N \neq conv\ M \longrightarrow N)$.

A Large conviction value shows that the obtained result is largely relying on the predecessor. The conviction can be formulated as

$$conv(M \longrightarrow Y) = \frac{1 - supp(N)}{1 - \mathrm{conf(M \longrightarrow N)}} = (1 - supp\ N)/(1 - conf(M \longrightarrow N))$$

It correlates the likeliness that M exists without N when they are relying on the factual frequentness of the existence of M without N. In that scenario, it is similar to lift. However, conviction have monotonousness in confidence and lift.

So,

$$conv(A2, A4 \longrightarrow A3) = \frac{1 - 0.7}{1 - 0.6} = 0.75$$

$$conv(A3 \longrightarrow A2, A4) = \frac{1 - 0.5}{1 - 0.42} = 0.86$$

Here, as the value of $conv(A2, A4 \longrightarrow A3)$ has a value less than the value of $conv(A3 \longrightarrow A2, A4)$, therefore the $conv(A3 \longrightarrow A2, A4)$ can be considered as a strong rule.

1.8 Quality measures and range of feasible values

The mining of valuable and previously unknown association rules in a database might produce a large number of different solutions, which makes the process hard to analyse each single rule. Thus besides, a large percentage of this set of solutions which might not be interesting and useful, being the user responsible for reducing the output by means of a process of quantifying the association rules. To solve this issue, different quality measures (see Table 2) based on the analysis of the statistical properties of data have been proposed by different authors.

Table 1.2: Quality measures and range of feasible values

Name	Equation	Feasible values
Support	P_{MN}	[0,1]
Confidence	$\frac{P_{MN}}{P_M}$	[0,1]
Lift	$\frac{P_{MN}}{P_M * P_N}$	[0,1]
Conviction	$\frac{P_M * P_{\bar{N}}}{P_{M\bar{N}}}$	$\left[\frac{1}{n}, \frac{n}{4}\right]$

This gives us an idea of the minimum and maximum acceptable values of the following quality measures of ARM, and if any rule having value outside the given range shall not be considered for further processing.

1.9 Comparative study of different DM algorithms

DM is the process of analysing data from different perspectives and summarizing it into useful information. It is an analytical tool for analysing data, which allows users to analyse data, categorize it, and summarize the relationships among data. Technically, DM is the process of finding correlations or patterns in large relational databases. Thus, to perform DM tasks the user has to choose a suitable algorithm.

In the table below, there is some comparison between the most common DM algorithms such as AIS, SETM, Apriori, AprioriTID, Apriori Hybrid, Eclat, Recursive Elimination, and FP-growth. The comparison is done solely based on the performance criteria like load carrying capacity, qualitativeness, and efficiency. Along with comparison, the basic advantages and disadvantages, and some suitability factors are also mentioned below.

Table 1.3: A comparative study of algorithms used in DM

S. NO.	Algorithm Name	Application	Advantages	Disadvantages	Year
1.	AIS	Not frequently used, but when used is used for small problems.	1. Better than SETM. 2. Easy to use	1. Candidate sets generated on the fly. 2. Size of candidate set large.	1994
2.	SETM	Not frequently used.	1. Separates generation from counting.	1. Very large execution time and the size of candidate set is large.	1994
3.	Apriori	Best for closed item sets.	1. Less candidate sets. 2. Generates candidate sets from only those items that were found large.	1. Takes a lot of memory.	2003
4.	AprioriTID	Used for smaller problems.	1. Better than SETM, Apriori for small databases. 2. Fast & Time saving	--	2013
5.	Apriori Hybrid	Used where Apriori and AprioriTID used.	1. Better than both Apriori and AprioriTID.	--	2013
6.	Eclat	Best used for free item sets.	1. Less memory usage. 2. Lower minimum support.	1. Apriori wins in cases where candidate sets are more.	2004
7.	Recursive Elimination	Effectively select the most relevant itemset.	1. Better than Apriori in all cases.	1. Less than éclat in all cases.	2005
8.	FP Growth	Used in cases of large problems as it doesn't require generation of candidate sets.	1. Only 2 passes requires. 2. No candidate set generation required.	1. Using tree structure creates complexity.	2003

As per Table 1.3, different types of DM algorithms are available in the marketplace, each with its advantages and disadvantages. Internal auditors must be aware of the different kinds of algorithms available and recommend the purchase of a tool that matches the organization's needs. As each of these algorithms has some suitability factors, and an algorithm that suits one type of dataset may not be suitable for another type of dataset, therefore the user needs an algorithm that can be suitable for almost all types of datasets, and it must fulfil the needs of industries. Existing DM algorithms sometimes fail due to the complex structure, long training time, and poor interoperability. But neural networks are a powerful technique to solve many real-world problems. Thus, ANN can play a vital role in DM.

1.10 Artificial Neural Network

An artificial Neural Network is a system loosely modelled based on the human brain. The field goes by many names, such as connectionism, parallel distributed processing, neuro-computing, natural intelligent systems, machine learning algorithms, and artificial neural networks. It can account for any functional dependency. The network discovers (learns, models) the nature of the dependency without needing to be prompted. It is a powerful data modelling tool that can represent complex input /output relationships. The motivation for the development of neural network technology stemmed from the desire to implement an artificial system that could perform intelligent tasks similar to those performed by the neurons of the human brain.

In this project we will be discussing the Potential use of Artificial Neural Network in DM.

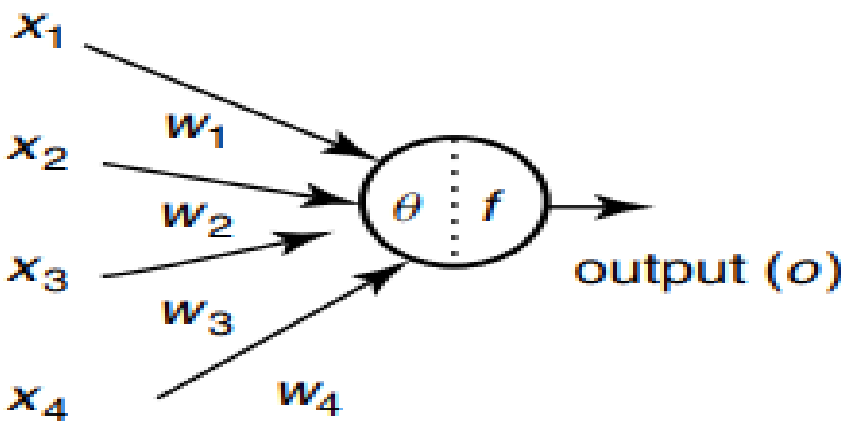

Fig 1.6: Artificial Neuron

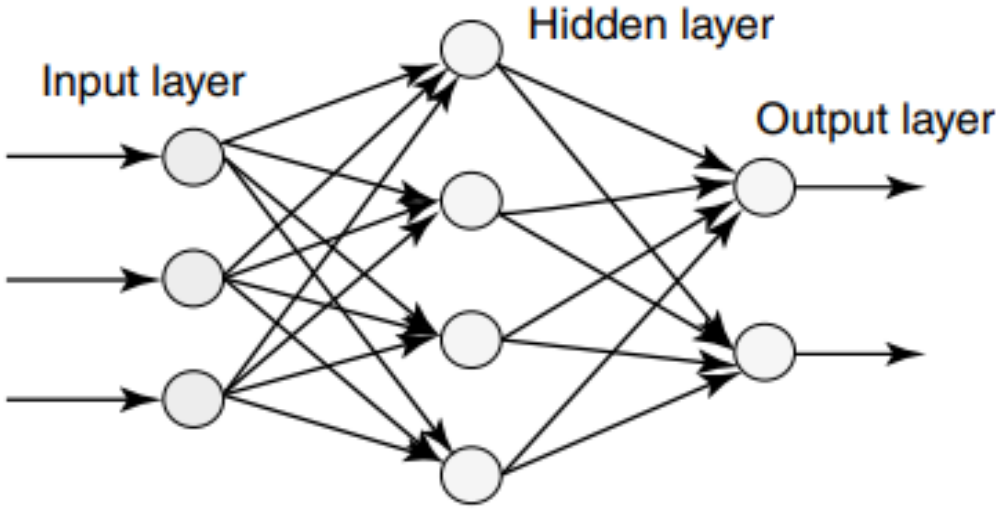

Fig 1.7: Multi-layered Artificial Neural Network

An artificial neural network has to be configured such that the application of a set of inputs produces the desired set of outputs. Various methods to set the strengths of the connections exist. One way is to set the weights explicitly, using a priori knowledge. Another way is to train the neural network by feeding it teaching patterns and letting it change its weights according to some learning rule. The learning situations in neural networks may be classified into three distinct sorts. These are supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, an input vector is presented at the inputs together with a set of desired responses, one for each node, at the output layer. A forward pass is done, and the errors or discrepancies between the desired and actual response for each node in the output layer are found. These are then used to determine weight changes in the net according to the prevailing learning rule. The term supervised originates from the fact that the desired signals on individual output nodes are provided by an external teacher.

Artificial Neural Networks possess large numbers of interconnected nodes or neurons. They perform computation in parallel but are configured in a regular architecture. ANN's collective behaviour is recognized by their ability to learn, recall and generalize the training patterns to that of a human brain. One of the methods of the Artificial Neural Network System (ANS) named Backpropagation can be used to solve text categorization problems. There are good reasons for using it for text categorization. ANS provides a better solution for problems that cannot be solved sequentially or by sequential algorithms. Other applications of artificial neural networks include the matching of images. They result in better performance. The backpropagation algorithm is widely used among the rest of the algorithms given by ANS.

Backpropagation is very useful in recognizing complex patterns and performing nontrivial mapping functions. Neurons are processing elements in the neural network.

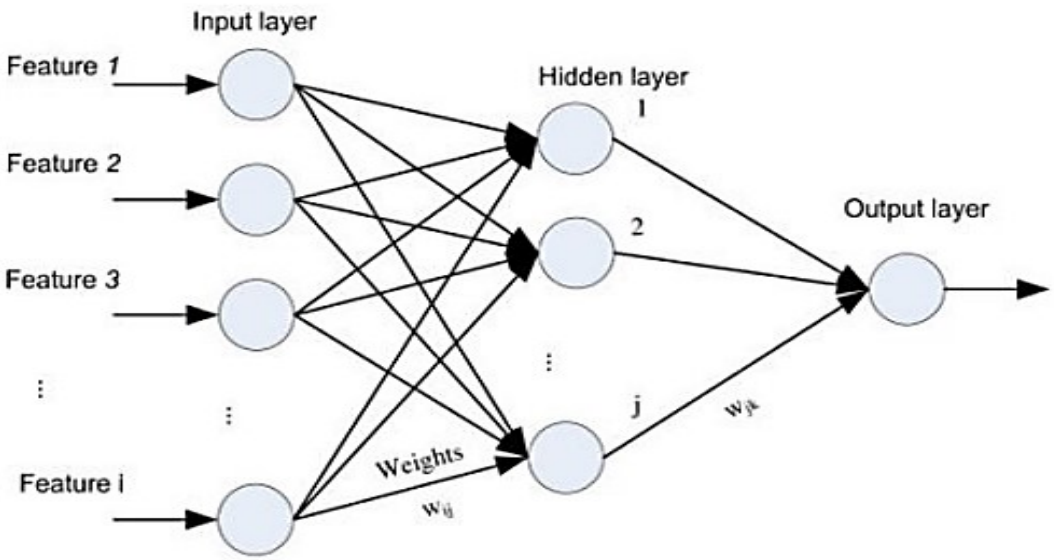

Fig 1.8: Architecture of Artificial Neural Network

The figure 1.8 presents the architecture of an Artificial Neural Network. The lines that connect the neurons are called weights. Every line of processing elements is a layer of a network; though there can be multiple layers present in ANS in Back Propagation network (BPN) generally there are three.

CHAPTER 2

LITERATURE REVIEW

CHAPTER 2

LITERATURE REVIEW

In the field of data mining, interesting rules are generated based on manifold criteria for the classification of itemsets. With this in reference, **Tasnia Rahman et al**., in [1] compare the performance of different fuzzy ARM algorithms based on their execution time and interestingness. The main comparison was between the Apriori algorithm, Fuzzy Apriori, and Genetic Fuzzy Apriori DC algorithm for two different transactional datasets. The Fuzzy Apriori and Genetic Fuzzy Apriori DC algorithms belong to the category of fuzzy ARM algorithms whereas the Apriori algorithm belongs to the category of classical algorithms. In the primordial phase, the algorithms were completely compared based on their Execution time, and later on, the two most superior algorithms were compared based on their interestingness. Finally, after analysing the results, it was concluded that the Apriori algorithm was one of the most efficient among the three compared algorithms.

ARM algorithms are one of the most suitable algorithms for DM when it is to mining the most frequent and potentially interesting patterns. In this series, **Vlad Robu and Vitor Duarte dos Santos** [2] compared the two most known and used ARM algorithms in the industry. Apriori and Eclat. This comparison led to the fact that how many possibilities it has to highlight how computationally expensive it becomes for Apriori when compared to Eclat to mine frequent patterns at a very low support threshold level and this might result in a critical difference and factor of choice when the research or the business goal is focused on identifying very rare and interesting patterns. Furthermore, it was also noticed that at very low thresholds Apriori tends to generate significantly more rules than Eclat, which is a real-world scenario that can increase the complexity of identifying interesting and meaningful rules. Despite these elements that might lead one to choose Eclat over Apriori, when using the R open-source software Apriori shows the great advantage of allowing for the calculation of additional interestingness measures which might help to explore and objectively choose and navigate through the vast universe of generated association rules. Considering the limitations of the research described, the differences between the two algorithms are based on the quality of the generated rules. Being aware of the tendency of Apriori to generate redundant rules, a new dimension of analysis can be added to the comparison by eliminating the redundancy within the rules and observing the impact on the output and algorithm performance. In conclusion, given the framework and used tools, there is no definitive verdict

over which algorithm results are better than the other. Given a specific real-world scenario and goal to be achieved, one will have to carefully evaluate the dimensions that differentiate Apriori from Eclat and choose appropriately based on the advantages and disadvantages each method is characterized by.

Fault diagnosis of power systems is an important method to realize self-repair and iteration of the power grid. In this series, **Liya Ma et al**. [3] presented a mining technique of SCADA frequent itemsets based on the FP growth algorithm. Based on the event collapse window and FP-growth algorithm, this system analyses a large number of data, and solves the problem that the Apriori algorithm can form due to a large number of candidate itemsets, finally resulting in low efficiency, and a large amount of SCADA data is difficult to form frequent itemsets. Experimental analysis shows that the FP growth algorithm is used to mine the frequent itemsets of data, which is consistent with the actual situation. Therefore, the algorithm is applied to the SCADA system dispatching centre, which can effectively improve the auxiliary analysis ability of the on-site dispatchers, to achieve the effect of effective and rapid accident handling.

The collection of transactional data on the sale of goods of a minimarket, which is increasing every day, is often treated only as a record, causing data to accumulate in a database. DM can help the decision-making process quickly, making it possible to manage the information contained in transaction data into new knowledge. So in this context, **Islamiyah et al**., in [4] forecasted a study to analyse the data of sales transactions of goods by comparing the Apriori algorithm and the FP-Growth an algorithm, to determine the association rules based on consumer purchase patterns with association techniques that seek several frequent itemsets and proceed with the establishment of Association Rules. As a result, although the number of association rules produced is different, this association rule has the same final association value (support × confidence). The final three highest association values in the sequence are 2.38% if you buy a Biscuit, you will buy a Snack; 2.35% if you buy a Snack, you will buy Biscuits; and 2.09% if you buy a Snack, you will buy Milk. The execution time needed to execute 979 item categories in the application states that the FP-Growth algorithm is faster than the Apriori algorithm. This result concludes the fact that the greater the minimum support specified, the less Association Rule formed.

Decision Making is one of the hardest and most crucial skills and there are many tools to support the leader of big companies to make error-free decisions, but ever since DM has

arrived in the world of business the job of such leaders has become much more crucial to make the best decision. So in this series, **J Sivapriya et al**., in [5] proposed some important contributions in this area filling the gap in the process of decision making and how decision-makers should work with the ever fast-changing data sets that contribute to the descriptive and prescriptive decision theory and for this, the two most known and popular ARM algorithms i.e., Apriori and FP growth are taken into account to get those frequent itemsets that are majorly observed in the database. The findings from the different algorithms when summarized lead to multiple decisions like making a reference point to divide the objectives of the problem carefully and doing well-research to make an initial choice Each step in business should take keeping all the parameters in the mind which is given in the problem statement and reference points should be created from time to time to get a gist out of each step of the process that helps in making a crystal clear decisions and providing support to get a bigger impact and desired results. Moreover detected tendencies will help data scientists and BIA on how to better design systems to support decision-makers.

In several places, it is seen that individuals are seen getting into trouble while using the frequent pattern mining strategies. With this remark, **K. Solanki Surbhi and T. Patel Jalpa** [6], focus on a mining strategy that works by examining the database several times, which eventually results in higher process costs. Not only this, this strategy leads to the production of candidate itemsets, which ultimately requires more memory and becomes more sophisticated in handling when the database is outsized. So, to reduce the downside of this problem of candidate set generation, a tree-based strategy came for mining periodic patterns. However, the tree-based strategy produces numerous conditional trees. So, to vanquish this problem of producing numerous based -trees, DAG for normal pattern mining is enhanced, whereas -a tree is constructed as a DAG. In addition, to find effectual mining, fuzzy deviation could be applied to the assayable database to provide the optimal patterns.

Many algorithms have been proposed to find association rules in various fields. In General, values associated with data items need not be the same type in all the databases. Applying a single algorithm for all types of data items may not be feasible. To solve this problem applying an individual algorithm for each type of data is one solution and the conversion of data item values to algorithm compatible types is another solution. With this in mind, **P. Amaranatha Reddy and MHM Krishna Prasad** in [7], presented ways how to obtain ARs for the differing types of data objects and requisitions involved in them. Some of the ways of data sets are reasoned in this document are Boolean, Quantitative, weighted, time-series,

stream data, infrequent, Diversified, and fuzzy itemsets. So the idea of differentiating ARs amongst correlated items like milk and banana will not be considered enterprise comprehension but differentiating unknown ARs between distinct items such as liquor and diapers will be beneficial in enterprises expansion and finding such an unknown category of ARs requires in-depth information about the data.

In today's scenario of the internet world, a cloud computing and data centre, a lot of importance has to be given to the security of information while sharing it with different organizations. Present advanced DM techniques give knowledge about various relationships of itemsets which will be useful to a group of people while enhancing their businesses. However, in many cases, the organizations having the source of information may not intend to share such relationships (association rules) with any other organization or they want to have rights of information. Therefore, there is a need to hide the sensitive information in the database before sharing it among the organizations. The focus of this work is to understand the hiding methodologies of association rules in the database while maintaining the disturbance in the original database as little as possible. So with these remarks, **Mrs. Geeta S. Navale and Dr. Suresh N. Mali** in [8] proposed a variety of methods to hide association rules on the database and to develop support and confidence measures. This paper intends to conceal the critical association rules of the DM under the following conditions, no production of false rules, no loss of information, Modification Degree, and robustness against intentional or unintentional attacks. Except for the three conditions above, the proposed method will be compatible with a measurable database. Therefore, the proposed method of data encryption of the operating system will be evaluated and verified concerning the various parameters as set out in the conditions and it is necessary to assess the satisfaction of the above-mentioned conditions.

Itemset mining identifies groups of frequent itemsets that signify possibly relevant information. Unique constraints are usually forced to emphasize the analysis of the most interesting itemsets. With these remarks, **L Greeshma and Dr. G Pradeepini** [9] focused on developing the latest Apriori-based algorithm, which satisfies positive aspects of constrained itemset-based mining such as anti-monotonicity. The problem of ARM is to retrieve relevant itemsets for which it presents a new constraint, called relation-based constraints, applicable to relevant data. In the CIM algorithm, it helps us to recognize the main components of a candidate's key itemsets and generate frequent itemsets, which satisfy the anti-monotonicity properties, which means small coverage and cardinal size limited to a particular dataset.

Association rule mining is a widely used approach in DM to discover trends from databases. There are a plethora of ARM algorithms available for frequent itemset mining and generating association rules. With this knowledge, **P. Naresh and Dr. R. Suguna** [10] explain a relative study of four ARM algorithms, namely Apriori, FPGrowth, LCM, and FIN. These algorithms are discovered in terms of their purpose, the way they instigate recurring itemsets, and their appearance on the organization's data. The performances of all these algorithms are evaluated using time. The algorithms presented in the DM configuration, namely "Sequential Pattern Mining Framework" are used to make an investigational assessment of the algorithms. These investigational results on these algorithms disclosed that the amount of data has a great significance on the implementation time and at last these led to the conclusion that, the FIN algorithm displays the least implementation time while FPGrowth shows minimum memory utilization.

It is particularly important to effectively mine potential customers and provide differentiated services for them to improve the competitiveness of banks with the strong competition among banks. With these remarks, **Guo Junfeng and Hou Handan** in [11] centralized the idea of putting customer relationships at the core of bank operation and made it to be the actual marketing concept of the bank. Apart from this, the usage of DM techniques can further help in bringing significant profits for enterprises by analysing customer data, mastering customer characteristics, and mining customer value. This led to the use case of the inductive algorithm of decision tree based on information entropy to conduct DM based on the survey results of all activities market of Australian banking institutions from May 2008 to November 2010, and explore the influencing factors of customer subscription for long-term deposit. The study showed that only employees are the factors that significantly affect customer subscription for long-term deposits. The 3 indicators including the number of employees, duration, and month have greatly reduced the scope of customers pushing customers to subscribe for long-term deposits, and significantly increased the rate of return on investment of banks, also further increasing the operating interest rate of banks, and to a certain extent, providing better services to customers. It is of great practical significance for the bank to expand its business and improve its core competitiveness.

DM may be viewed as the extraction of interesting patterns and models from observed data or a method used for analytical processes designed to explore data. With this **Mr. Nilesh Kumar Dokania and Ms. Navneet Kaur** in [12] presented a comparative analysis of various DM techniques. It is already known that different mining techniques can be implemented on

data to predict their future performance. But here comes the toughest thing, which is to determine which mining technique is most suitable for any particular instance. Most of the time, the techniques to be used are determined by trial and error. There are certain differences in the kinds of problems that are most favorable to each technique but the authenticity of real-world data and the dynamic way in which markets, customers, and data that represents them have formed means that the data is continuously changing. These dynamics mean that it is not always possible to build a "perfect" model on the historical data since whatever was known in the past cannot adequately predict the future. But sometimes the situation is very crucial for a business person who is waiting for all the information to come in before they make their decision. Since business, economy, and even the world are changing in unpredictable and even chaotic ways DM techniques are not always predictable. So it's safe to select a robust model that may under-perform when compared to the best DM tools for analysis and execution at the earliest to take business decisions before it's too late.

Crime prevention has always been one of the basic and important issues in human life that have been applied in various ways throughout history. Due to the development of information and communication technologies and the launch of comprehensive information systems in the police force and the registration of criminals' information in databases, the use of DM techniques and knowledge discovery to analyse and track down crimes, including theft, is one of the necessities of the Iranian police and the judiciary. With this in focus, **Nima Norouzi and Elham Ataei** [13] use DM and its algorithms, to perform an accurate analysis of the mass data in the database and help the relevant managers make decisions using the data's hidden patterns. Here using the two methods of coordination rules and clustering method, an attempt has been made to discover and make available to these tribes the requirements of managers. For example, some of the extracted patterns show that villas are more vulnerable to encroachment than apartments or that some of these rules indicate that women are more likely to be robbed. These patterns also clear the fact that the car contents are more likely to be stolen in the spring than in the winter. Here it has been focused on four useful rules out of fifty rules that were interpreted and approved by experts. Of course, what is clear is that crime DM is far more extensive than can be comprehensively presented in an article. Security is a national and sometimes regional and global issue, so the perpetrators of crime DM are generally the claimants of security (the judiciary, the Ministry of Intelligence, the National Security Council, the police, and other relevant organizations). Depending on the type of crime, crime DM can be discussed in the following contexts.

Once someone said - “Predicting the future isn’t magic, its artificial intelligence”. With this in mind, **Yasmeen Alhamdan et al**., in [14] presented a system that integrates Artificial Intelligence (AI) methods with Augmented Reality (AR) techniques to enhance the grocery shopping experience through the use of smart glasses. The framework deployed in this was a Convolutional Neural Network (CNN) object detection model that allows for item identification. By simultaneously retrieving data from a large nutrition database, personal medical reports, and other grocery store-related datasets, the intelligent system was able to provide user-centric nutrition facts, health and wellness tips, and unhealthy selection warnings that are augmented in real-time broadcasting of the smart glasses. Our state-of-the-art framework Co-Shopper demonstrates high accuracy in detecting grocery items, improves product selection, increases cost efficiency, and reduces the time spent in the process.

In the ARM field many different quality measures have been proposed over time to quantify the interestingness of each discovered rule. In evolutionary computation, many of these metrics have been used as functions to be optimized, but the selection of a set of suitable quality measures for each specific problem is not a trivial task. Taking this issue as a serious concern **J. M. Luna et al**., in [15] carried out an empirical study to facilitate the decision about which measure should be selected for each specific goal. For this, extensive analysis has been performed by considering a set of thirty varied well-known datasets from which association rules are mined over a set of twenty quality measures widely used in literature. A set of exhaustive search models, as well as highly promising evolutionary algorithms, have been also used to perform this study, considering either single-objective algorithms or multi-objective optimization approaches. For each algorithm and each quality measure, more than 57,000 executions were done. The obtained results were of high interest for future research in the field, determining which single quality measure (or pairs of them) should be used if other different metrics are required to be optimized.

In today’s big data environment, ARM has to be extended to big data. Whereas traditional ARM algorithms, like Apriori, mostly mine positive association rules. Positive ARM finds items that are positively related to one another, that is, if one item goes up, the related item also goes up. Though the classic application of positive ARM is market basket analysis, applications of positive rule mining have been extended to a wide range of areas like biological datasets, web-log data, fraud detection, census data, etc. Negative association rules can be defined as items that are negatively correlated, that is, if one item goes up, the other item goes down. Negative ARM also has many applications, including the building of

efficient decision support systems, crime data analysis, in the health care sector, etc. So with these concepts in the background, **Sikha Bagui and Probal Chandra Dhar** in [16] presented a Hadoop implementation of the Apriori algorithm to mine positive as well as negative association rules. The experiments showed that for the Web Docs dataset, there were more negative association rules than positive association rules, so if we mine just for positive association rules, we could be losing some information. Also, for this dataset, 25 slave nodes with a block size of 256 MB gave the best runtime performance. So, for large datasets, a larger amount of parallelization, that is, more number of slave nodes with higher block size, is more efficient.

At present, association rules have been widely used in prediction, personalized recommendation, risk analysis, and other fields. However, it has been pointed out that the traditional framework to evaluate association rules, based on Support and Confidence as measures of importance and accuracy, has several drawbacks. With these concerns, **Fuguang Bao et al**., [17] present several key indicators for analysing and comparing the various measures developed in the statistics and DM literature. Currently given differences in some of their indicators, a significant number of these measures may provide conflicting information about the interestingness of rules. Therefore, how to estimate the reliability of the obtained association rules has become a research hotspot. Here it has demonstrated the benefits of using Support in eliminating irrelevant and poorly correlated patterns. Studying the traditional ARM relies on the Support-confidence framework, and only rules that meet both the Support-Confidence threshold can be called strong association rules. However, sometimes strong association rules are of no interest to users, and even worse, these association rules can be misleading to some extent. Thus, to find the most valuable association rules, it is necessary to further analyse and evaluate the mined rules. Finally, we build up a reasonable measure framework (Bi-support and Bi-confidence framework) for mining appropriate association rules.

In recent trends, information and science technology has been developed which makes most of the research in ARM to be focused on enhancing the computational efficiency. But the searching for the most appropriate data in large databases becomes more difficult. The major difficulty in this is determining the threshold values for confidence and support which causes the effect on the quality of the ARM. As a solution to this, **Dr. K. Kala** in [18] has represented a novel optimization-based ARM algorithm that could be used in deriving sensitive and efficient information from large databases. Initially, a pre-processing procedure

was taken where the input raw data obtained from a large database, got converted into a binary data type. Secondly, a novel MBPSO algorithm determines the threshold values for minimal support and confidence effectively. Here, fitness function estimation has been done and the best particles are determined. Finally, the most legitimate ARM rules are obtained. In this process, various experimental and comparative analyses have been done to measure the effectiveness of the proposed system. The proposed system has been compared with certain existing methods like FP-Growth, and Eclat concerning processing time for various datasets like Mushroom dataset, letter recognition, etc. From the observations, it has been obvious that the proposed system outperforms other compared methods. Moreover, the proposed MBPSO algorithm has been compared with certain algorithms like Apriori, SAR, SARM SRC, etc. concerning accuracy-index, time-index, etc. From those analyses, it has been confirmed and proved that the proposed algorithm has shown more remarkable behaviour than many other algorithms.

Higher education institutions are often very curious to know about the success rate of the students throughout their studies. For this reason, they need to use several methods like physical examination, Statistical methods and currently prevailing DM techniques for the prediction of a student's performance. An upcoming area of research that uses techniques of DM is known as Educational DM. It involves machine learning algorithms and statistical techniques to help the user for interpretation of students' learning habits, their academic performance, and further improvement if required. Under this series, **Chitra Jalota and Rashmi Agrawal** [19] presented the significance of DM in educational institutions. The knowledge acquired by the usage of DM techniques can be further utilized to cast successful and effective decisions that will improve and progress the student's performance in education. The data set contains 163 instances and sixteen attributes. Here, five classifiers have been used under weka and the comparisons are made based on the accuracy among these classifiers, and different error measures are used to determine the best classifier. Experiment results show that Multilayer Perceptron has the best performance among other classifiers.

The evaluation system of students is to find a way to solve the status of students according to the exact needs of students and the teaching requirements of teachers, to improve the teaching level of teachers and improve the quality of school education. With this base, **Yuzi Dou et al**., in [20] focus on how the association rule can be applied to the student evaluation teaching system. Thus, by improving the Eclat algorithm, the complexity of sorting and analysing teaching data is reduced, and the efficiency of the improved algorithm is improved (fast).

However, it is necessary to take into account the preparatory work in the early stage that is, the initial working time required to apply the breadth priority. So the Eclat algorithm still has some shortcomings. At the same time, Weka can be used to implement the Eclat algorithm and further explore its application. In a conclusion, to facilitate the integration and analysis of evaluation information, the Eclat algorithm can be proposed. Moreover, for the effectiveness of information collection, some of the following measures can be implemented to improve the student evaluation system, May strengthen the student to evaluate the teaching propaganda work. For the evaluation system indicators, students are mobilized before the evaluation and special training to clearly define the purpose of the evaluation. According to the classification design index system of different disciplines, the evaluation and evaluation form is scientifically and reasonably improved to make it more flexible and humanized. At the same time organize experts, students, teachers, and other people to discuss. The overall setting should pay attention to teaching attitude, teaching content, teaching methods, teaching methods, and teaching effects, and need to be divided according to the emphasis. The related functional departments of the organization and management evaluation shall fully implement the evaluation rules, deal with the evaluation results fairly and equitably, and formulate rigorous solutions to the problems. Focus on outstanding achievements at home and abroad, and learn from each other's strengths.

By aiming at the shortcomings of the traditional "support confidence" association rules mining framework and the problems of mining negative association rules, the concept of interestingness measure is introduced. **Junrui Yang and Lin Xu** [21], analysed the commonly used interestingness measures and merged the interestingness measure based on the different ideas and the cosine measure to obtain a new interestingness measure model. By using this model in mining association rules, the positive and negative association rules can be effectively mined and segregated. Moreover, the algorithm is further verified by using a certain grocery store transaction dataset in comparison with the existing positive and negative association rules mining algorithm and the mining results using the traditional "support-confidence" framework. The results verify the effectiveness of the improved interest model and the algorithm using the model and contribute to the overall improvement of the mining accuracy of the algorithm.

DM is the process of extracting hidden and useful patterns and information from data. DM is a new technology that helps businesses to predict future trends and behaviours, allowing them to make proactive, knowledge-driven decisions. In this series, **Anshu** [22] shows the process

of DM and how it can help decision-makers to make better decisions. It also presents a detailed description of DM techniques and their applications in various fields. Some of the major mining techniques such as classification, clustering, etc., help in finding the patterns to decide upon the future trends in businesses to grow. Different mining techniques can be used for different purposes as per their pros and cons.

DM has been widely used in the business field, and machine learning can perform data analysis and pattern discovery, thus playing a key role in DM applications. **Teng Xiuyi and Gong Yuxia** in [23] expound on the definition, model, development stage, classification, and commercial application of machine learning, and emphasizes the role of machine learning in DM. As per them, the machine learning technology in DM can be applied in many industries, including the financial industry, retail industry, insurance industry, telecommunication industry, and so on. For example, in the financial industry, financial analysts use DM to build prediction models to identify the patterns that have caused market volatility in history, thereby improving the ability to predict market volatility. In the retail industry, salespeople can build predictive models through data mining to understand who is most likely to respond to correspondence, thereby increasing sales. When enterprises apply data mining technologies, they should fully understand the advantages and disadvantages of various technologies and methods, and select appropriate technologies for specific environments and tasks.

Frequent pattern mining is an important topic that is needed for Internet of Things (IoT) applications frequently. Many IoT applications have been developed in which continuous streaming data is used. In continuation with this, **Halil İbrahim DEDE et al**., in [24] performed a comparison study based on performance evaluation of FP-stream (Frequent Pattern-Stream) algorithm and weighted sliding window mining (WMFP-SW) algorithms, which are developed for frequent pattern analysis in IoT, using a novel real-world dataset. The dataset was created by collecting 6 different sensor values over 3 months in an established testbed. As per the earlier studies, both the algorithms are efficient enough in extracting FPs but the experiments done shows that the WMFP-SW algorithm extracts more FPs than the FP-stream algorithm when minimum support is increasing. Regarding the results of execution time, the FP stream is always faster than the WMFP-SW. Moreover, the number of common FPs and the total number of FPs per algorithm decrease as the minimum support value increases.

Generally, the power systems are disturbed by the presence of electromagnetic interference and crosstalk between the transmission link layers in the transmission and distribution process. Though it is easy to produce transmission distribution faults, in sincere efforts to improve the efficiency of fault diagnosis, a method of fault diagnosis for power systems based on neural network algorithms was proposed by **Kai Xu** in [25]. In this, the multi-sensor quantization fusion method is used to carry out the electricity. The transmission distribution signal in the power transmission link layer is extracted from the power system, the transmission distribution signal is decomposed and the association rules are excavated. Furthermore, the spectral analysis model is used to extract the spectral characteristics of the transmission information of the power system, and the fault diagnosis and fault type identification are carried out according to the spectrum difference. The power system fault features are then further classified and identified by a neural network learning algorithm to realize the optimal diagnosis of power system fault. The simulation results show that the method was more accurate and more efficient in the fault diagnosis of the power system.

Embedded systems are finding their way into almost every aspect of our daily life from mp3 players and console games to mobile phones. Different Artificial Intelligence (AI) based applications are commonly utilized in embedded systems from which computer vision-based approaches are included. The demand for higher accuracy in computer vision applications is associated with the increased complexity of convolutional neural networks and the storage requirement for saving pre-trained networks. Different factors can lead to data corruption in the storage units of the embedded systems, which can result in drastic failures due to the propagation of the errors. Hence, the development of software-based algorithms for the detection and recovery of data corruption is crucial for the improvement and failure-prevention of embedded systems. So with this motivation, **Mohammadreza Ramzanpour and Simone A. Ludwig** in [26], proposed a new recovery algorithm for the recovery of the data in the case of a single event upset (SEU) error. In this process, the ResNet32 and ResNet110 CNN were trained on the CIFAR10 dataset, with two different accuracy levels for each of them. The number of parameters involved with ResNet110 was even 3.7 times greater than that of ResNet110. Furthermore, the Apriori and FP-Growth DM algorithms were also used in finding the frequent patterns and the associated confidence level with the obtained rules. The confidence levels were used later to define a probabilistic measure for identifying the corruption possibility. The pre-trained parameters were stored in the format of IEEE754, which uses a 32-bit binary. Therefore, the bits closer to the far-right side of the binary

representation will have a marginal and insignificant effect on the final value of that parameter. Hence, the number of bits was reduced with each test to investigate its effect on our recovery algorithm on 9, 11, and 13 bits. The recovery algorithm was noticed to drop when the number of most significant bits was increased. In the case of using 9 significant bits, the recovery rate was observed to be in the range of 61 to 66% when the errors were introduced for a random bit. Moreover, in some bits, more than 99% of all the parameters are shown to have one specific value. The performance of the algorithm was assessed in the case of corruption for those bits. For some bits, the recovery rate was noticed to be in the range of 98 to 100%, however, for some other frequent bits and especially for ResNet110 the performance declined as low as 74%.

The field of Solar Energy Potential Forecasting and Optimization using Artificial Neural Network is based on the comparison of different artificial neural network (ANN) training functions for modelling the daily average solar irradiance (DASI) forecaster. With this reference, **Sempe Leholo et al**., [27] aimed to establish a comparison study concerning the performances of four training functions namely, the Resilient Backpropagation (TRAIN-RP), Levenberg–Marquardt (TRAIN-LM), Scaled Conjugate Gradient (TRAIN-SCG), and Quasi-Newton (TRAIN-BFGS) implemented in the training of the ANN for forecasting the DASI. The final results demonstrate how the training algorithm selected in the course of ANN training has a tremendous effect on the predictions of the ANN. Using the Levenberg–Marquardt's learning rule, the feed-forward network with backpropagation architecture is implemented in the approach of neural network design. The comparison is done by looking at the Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of correlation (R) of the models. The training function TRAIN-LM was probably found the most appropriate function having the experimental data of daily average solar irradiance amid the four functions that have been chosen for the evaluation as it achieved a 99.742% coefficient of correlation with a mean absolute percentage error (MAPE) of 2.394%. Subsequently to TRAIN-LM, the training function TRAIN-BFGS is discovered to be second in performance, however, TRAINRP is discovered to be the very least in performance among all of the four training functions. The training, testing, and validation of the neural network are done by utilizing the actual data from six different locations in the North West cities, and these data are mined from the NASA database. These findings will be useful in the selection of training functions in the course of the modelling of the ANN for daily average solar irradiance prediction in areas around South Africa. The chosen feed-

forward backpropagation ANN that has the TRAIN-LM training function is a great instrument for modelling daily average solar irradiance studies.

The increase in the rate of technological evolution is resulting in a reduction in the cost of various storage devices and as a consequence, enormous amounts of data are deposited from heterogeneous sources in raw forms. Therefore, some efficient DM techniques are required that can process those data and retrieve useful information from them. Recently, machine learning algorithms are becoming very popular for doing various DM tasks. A neural network is one of them, which has fascinated a lot of researchers due to its efficacy and fruitfulness in doing many tasks, especially classification. But the main problem with neural networks is their nature as a black box, i.e., explaining the decision generated by a neural network is a daunting task. As a solution to this problem, **Manomita Chakraborty et al**., [28] proposed a rule extraction technique that expresses the knowledge hidden in a learned network in the guise of understandable classification rules. So, here it is just being surveyed how a Neural Network can be converted into a white box by the rule extraction process. Rule extraction is already a deep-rooted and well-studied topic, therefore, it is important to know the areas which need improvement to enrich the literature and fulfil the needs of emerging applications.

DM plays a key role in Crime Analysis. Over the years, many algorithms are being extensively used in this field, like the virtual identifier, pruning strategy, support vector machines, and apriori algorithms. Among which VID is to find the relation between record and vid, while the apriori algorithm helps the fuzzy association rules algorithm and it takes around six hundred seconds to detect a mail bomb attack. In this field, **Biswajit Panja et al**. [29] identified Crime mapping analysis based on KNN (K – Nearest Neighbour) and ANN (Artificial Neural Network) algorithms to simplify this process. Crime Mapping is conducted and Funded by the Office of Community Oriented Policing Services (COPS). Evidence-based research helps in this field by analysing the crimes. The crime rate is calculated based on the previous data using DM techniques. Crime Analysis uses quantitative and qualitative data in combination with analytic techniques and is majorly used in resolving the cases. With the assistance of these technologies, the wrongdoing information will be nourished to the information-digging device for investigation and afterward, two unique models will be recorded. With the assistance of the SAM instrument/tools, a strategic distance will be maintained from the distinction in the outcome and after that, the subsequent information will be utilized for finding the relations amongst those et cetera. Along these lines, we will lessen

false positives and false negatives in the field of the interruption identification framework utilizing the information mining in the field of wrongdoing information examination. For public safety purposes, crime mapping is an essential research area to concentrate on. We can identify the most frequent crime-occurring zones with the help of DM techniques.

Given the problems existing in traditional DM, such as unstable execution efficiency and poor adaptability of data scale change and to achieve real-time and accurate DM, **Tianyu Xia et al**., in [30] proposed a data association rules mining method based on RBF neural network optimization algorithm. Based on considering the constraint association rules, the data frequent itemsets are reduced to get the corresponding candidate data sets that meet the rules, and the update of the data frequent itemsets for user needs is completed. The candidate itemset is input to RBF neural network for training, the network output is optimized by combining with rough set theory, the user demand attributes are described, and the user demand data is obtained by using the upper bound pruning method, the user demand data is obtained and real-time mining is realized. RBFNN is used to optimize data in candidate data sets, combined with rough set theory, and the upper bound pruning method is used to get user demand data. When compared with other DM algorithms, the simulation results show that the performance of the DM algorithm proposed is less constrained by the proportion of redundant data and the size of the data, the accuracy of DM is higher, and it has a better stability.

CHAPTER 3

PROBLEM DEFINITION

CHAPTER 3

PROBLEM DEFINITION

3.1 Problem Formation

The main issue with the existing ARM algorithm is that it follows complex procedures to generate frequent patterns, those it is very hard to understand for a non-IT person. Also, the existing association rules mining algorithms more or less fails when it needs to handle large dataset and it is also a takes longer processing time for large dataset.

Currently, we observe with the enormous demands in technology the data size is also increase exponentially. Thus, for a normal human being or the existing algorithm it is almost impossible to extract interesting information, or to draw a frequent pattern over it.

3.2 Objective

1. The algorithm will predict whether a user will end up buying a product or not based on the behaviour of user it draw a pattern.
2. The algorithm must draw a pattern for the items that are likely to be frequently purchased together.
3. To compare the execution time of the DAEM (De-noising Auto Encoder Mining) algorithm model with the two well-known ARM algorithm Apriori and FP-Growth model.
4. It must fulfil the current industrial needs of handling large data in DM.

3.3 Requirement

- For this experiment, an Intel(R) Core(TM) i3-8130U CPU @2.20GHz with 8.00 GB RAM or above has been used. The operating system was Microsoft Windows 10 Enterprise. While comparing the performance of the three algorithms Apriori, FP Growth, and Optimised algorithm using ANN, no other application was being run for feasible analysis of the Run Time and the Interestingness.
- The platform of python IDE with python version 3.6.* along with tensorflow version 1.13.1 is being used in this experiment. DM is an interdisciplinary field with the contribution of so many fields such as statistics, machine learning, information

retrieval, etc. Python is an open-source general purpose programming language which can merge all these libraries in a few lines of code.

- The datasets used for the experiment are the Grocery shop dataset and the supermarket dataset. These datasets are collected from the Data world.

CHAPTER 4

PROPOSED MODEL

CHAPTER 4

PROPOSED MODEL

The main goal of this project is to analyse and test the existing DM algorithms, and to explore their applications, merits, demerits, and further to develop an DM algorithm using ANN technique like deep auto encoder (DAE) and compare it with some of popular and widely used DM algorithms. By using the ANN techniques, the project aims to build an algorithm which has the capabilities to generate strong association rules and patterns more efficiently than compared to Apriori and FP-Growth Algorithms. The comparison of these algorithms has been performed on the basis of their execution time, and interestingness.

This project focuses on the DM algorithms techniques and ARM. ARM is commonly used for market basket analysis (DM) problems to identify combinations of items more likely to appear together in the transaction of a dataset. It helps to identify interesting rules, frequent patterns, correlations, or just casual data structures in transactional datasets. An association rule can be seen as an expression of type$X \longrightarrow Y$, where X and Y are sets of items. In a real-world scenario, such a rule can be that, for example, customers that have purchased X also purchased Y. The interestingness of such a rule is represented by its support and confidence. Where support of a rule $X \longrightarrow Y$ is the percentage of transactions that contain both X and Y. While the confidence of the rule represents the percentage of transactions of X that also contain Y. Here, some of the well-known methodologies such as Apriori, FP growth, and Optimised Algorithm using ANN are considered for the experiments of analysing the performance of mining association rules.

Existing ARM algorithms are not capable of finding the most efficient and strong rules which are suitable for any dataset due to the presence of numerous rules. Here, we have taken two popular algorithms; Apriori, one of the most popular DM algorithms, and FP-Growth, another popular DM algorithm. Both algorithms have followed different techniques and parameters but both are having some limitations. Which we are trying to solve using the concept of ANNs. A deep analysis of all three algorithms are carried out which focuses on run-time, efficiency and interestingness measures by the algorithm.

In this project, we had designed and developed an Optimized algorithm for DM, in order to achieve the more efficient, and refined mining methods by using the concepts of Artificial Neural networks (ANN). In this firstly we used Association rules mining techniques such as

support, lift, confidence, and conviction. These techniques or measures are being used to filter out the strong rules in the dataset. Therefore, we had applied concepts of ANN in order to achieve more efficiency to obtain strong rules for the further analysis.

These strong rules will further be used in applications like market basket analysis, data analysis, stock price prediction, biological data analysis, and many other industries that work in large data.

4.1 Apriori Algorithm

In the year of 1994, R. Agrawal and R. Srikant introduced a fast way of ARM which is said to be the Apriori algorithm. It can be regarded that the Apriori algorithm is the pioneer of the ARM technique. It follows the breadth-first search and a Hash tree structure is created to count candidate itemsets. It generates candidate itemsets of different lengths from itemsets up to the length of the itemset. Than the candidates are pruned according to the frequency, infrequent sub-patterns get pruned off. According to the downward closure, the candidate set is contained by all frequent-length item sets. After that, it scans the transaction database several times to determine frequent item sets among the candidates of the transactions. It is a classical algorithm that finds out the rules, using specified thresholds, Support, Confidence, Conviction and Lift. Here the value of minimum support and confidence is user-defined and for this reason, it requires the user to have profound knowledge about the corresponding dataset.

Apriori algorithm refers to an algorithm that is used in mining frequent product sets and relevant association rules. Generally, the Apriori algorithm operates on a database containing a huge number of transactions. For example; the items customers buy at a Grocery Shop. This algorithm is easy to execute and very simple, and is used to mine all frequent itemsets in the database. This algorithm was developed in 2003. The algorithm makes many searches in the database to find frequent itemsets where k-itemsets are used to generate k+1-itemsets. Each k-itemset must be greater than or equal to minimum support threshold to be frequency. Otherwise, it is called candidate itemsets. In the first, the algorithm scans the database to find the frequency of 1-itemsets that contains only one item by counting each item in the database. The frequency of 1-itemsets is used to find the itemsets in 2- itemsets which in turn is used to find 3-itemsets and so on until there are not any more k-itemsets. If an itemset is not frequent, any large subset from it is also non-frequent; this condition prunes from search space in the database.

So, here are some of the basic measures to generate frequent item sets

A. **Support($X \longrightarrow Y$) -** frequency of relationship.

B. **Confidence($X \longrightarrow Y$) -** Probability of occurrence of $\{Y\}$ given $\{X\}$ is present.

C. **Lift -** Ratio of *confidence* to baseline probability of occurrence of {right hand side Y}

$$Lift(X \longrightarrow Y) = confidence\ (X \longrightarrow Y)/support(Y)$$

- If $lift = 0$, means no relation w/n items
- If $lift > 1$, means positive relation w/n items
- If $lift > 1$, means negative relation w/n items

D. **Conviction -** The conviction is defined as one of the measures of ARM which undertake to analyse the magnitude of execution of the rule, to evaluate the conviction.

$$conv(M \longrightarrow Y) = \frac{1 - supp(N)}{1 - conf(M \longrightarrow N)} = (1 - supp\ (N))/(1 - conf(M \longrightarrow N)$$

A Large conviction value shows $t$hat the obtained result is largely relying on the predecessor.

4.1.1 Steps to find out the Strong Rules in Apriori Algorithm

Following are the steps required to find out the strong rules in the case of a dataset

Step 1: Choose the dataset, on which you have to perform the tasks.

Table 4.1: Transactional dataset of supermarket

T_{ID}	List of Items
T_1	I_1,I_2,I_5
T_2	I_2,I_4
T_3	I_2,I_3
T_4	I_1,I_2,I_4
T_5	I_1,I_3
T_6	I_2,I_3
T_7	I_1,I_3
T_8	I_1,I_2,I_3,I_5
T_9	I_1,I_2,I_3

We had assumed the value of min support = 2 and the value of confidence to be 70% for the above transactional dataset.

Step 2: Itemset generated by Intersection of two itemset

Table: L1

Itemset	Frequency
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

Table: C2 = L1 * L1

Itemset	Frequency
I_1,I_2	4
I_1,I_3	3
I_1,I_4	1
I_1,I_5	2
I_2,I_3	4
I_2,I_4	2
I_2,I_5	2
I_3,I_4	0
I_3,I_5	1
I_4,I_5	0

Table: Largest datasets (L2)

Itemset	Frequency
I_1,I_2	4
I_1,I_3	3
I_1,I_5	2
I_2,I_3	4
I_2,I_4	2
I_2,I_5	2

Step3: Itemset generated by intersection of two itemset

Table: L3

Itemset	Frequency
I_1,I_2	4
I_1,I_3	3
I_1,I_5	2
I_2,I_3	4
I_2,I_4	2
I_2,I_5	2

Table: C2 = L1 * L2

Itemset	Frequency
I_1,I_2,I_3	2
I_1,I_2,I_5	2
I_2,I_3,I_4	0
I_2,I_3,I_5	1
I_2,I_4,I_5	0

Table: Largest datasets (L3)

Itemset	Frequency
I_1,I_2,I_3	2
I_1,I_2,I_5	2

Step 4: Itemset generated by intersection of two itemset

Table: L3

Itemset	Frequency
I_1,I_2,I_3	2
I_1,I_2,I_5	2

Table: C4 = L1 * L3

Itemset	Frequency
I_1,I_2,I_3,I_5	1

The frequency of the itemset present in C4 i.e., $\{I_1, I_2, I_3, I_5\}$ is 1, which is less than the minimum support value which is 2.

As C4 is not satisfying the MinSupp () value, therefore, we will consider L3 as having the largest frequent itemset which is $\{I_1, I_2, I_3\}$ and$\{I_1, I_2, I_5\}$.

Step 4: Generation of the following rules obtained from L3

Here $X_1 = \{I_1, I_2, I_3\}$ and $X_2 = \{I_1, I_2, I_5\}$ are the two most frequent itemsets obtained by L3 with a minimum support value equal to 2.

Association Rule		Confidence	Association Rule		Confidence
$I_1 \longrightarrow I_2, I_3$= 2/6	=	33%	$I_1 \longrightarrow I_2, I_5$= 2/6	=	33%
$I_2 \longrightarrow I_1, I_3$= 2/7	=	28%	$I_2 \longrightarrow I_1, I_5$= 2/7	=	28%
$I_3 \longrightarrow I_1, I_2$= 2/6	=	33%	$I_5 \longrightarrow I_1, I_2$= 2/2	=	100%
$I_2, I_3 \longrightarrow I_1$= 2/4	=	50%	$I_2, I_5 \longrightarrow I_1$= 2/2	=	100%
$I_1, I_3 \longrightarrow I_2$= 2/3	=	66%	$I_1, I_5 \longrightarrow I_2$= 2/2	=	100%
$I_1, I_2 \longrightarrow I_3$= 2/4	=	50%	$I_1, I_2 \longrightarrow I_5$= 2/4	=	50%

Therefore, the strong association rules obtained through the apriori algorithm are:

$\{I_5 \longrightarrow I_1, I_2\}, \{I_2, I_5 \longrightarrow I_1\}$ and$\{I_1, I_5 \longrightarrow I_2\}$, as these three rules satisfy all the measures of Apriori algorithm which is minimum Support and minimum Confidence values.

4.2 FP Growth Algorithm

FP-growth algorithm is an association analysis algorithm proposed by Han Jiawei and others in 2000. The main idea of the algorithm is to take the following divide and conquer strategy, compress the database which provide frequent sets to a frequent-pattern tree (or FP-tree), but still retain relational information of itemsets; then divide this compressed database into a set of conditional database (a special type of project database), each associated with a frequent set, and do DM on each database. FP-Growth (Frequent-Pattern Growth) algorithm is based on Apriori's improved frequent pattern growth algorithm. Compared to Apriori, it only traverses the database twice, which improves search efficiency. Apriori is an algorithm for frequent pattern mining that focuses on generating itemsets and discovering the most frequent itemset. It was developed in 2003. It greatly reduces the size of the itemset in the database, however, Apriori has its own shortcomings as well. This algorithm is an improvement to the Apriori method. A frequent pattern is generated without the need for candidate generation.

The workflow of the FP-growth algorithm is as follows: First, it scans the transaction set for the first time and lists the number of times each item appears in the transaction set. Then according to the experience, the minimum support is determined. Then the process starts as items that do not meet the minimum support are deleted from each transaction of the transaction set, and the items meeting the minimum support are arranged in reverse order according to the number of times, to obtain a frequent item set. Secondly, the root node of FP-Trees are set to null.[13] Then the modified transaction set is scanned again, and a path is established from the root node to the leaf nodes as per the order of frequent itemsets. If the items in the transactions are already in the FP-Tree, then the corresponding number of items is uplift by 1, and different items are added after it, that is, the same items are shared and the coefficient is uplift by 1. FP growth algorithm represents the database in the form of a tree called a frequent pattern tree or FP tree. Its advantages are, only 2 passes required, no candidate set generation required. Its disadvantages are, it uses tree structure and creates complexity. It is used in large cases of problems as it doesn't require generation of candidate sets.

Related concepts for FP Growth algorithm are as follow:

- **FP-Tree-** Sort data items in the transaction data table by support, than insert the data items in each transaction into a tree with NULL as its root by descending turn and record the support of each node.

- **Conditional pattern base-** Contains the set of prefix path which appears together with the suffix pattern set in the FP-Tree.
- **Condition tree-** Construct the conditional pattern base into a new FP-Tree according to the principles of the formation of FP-Tree.

4.2.1 Steps to find out Strong Rules in FP Growth Algorithm

There are three steps to find out strong rules in the case of fp growth algorithm

Step 1: Construct a new table with the shorted value frequency of items.

Step 2: Compress database into FP tree, which retain the itemset association information.

Step 3: Divide the comparess database into set of conditional databases

Note- We have considered the same dataset as discussed earlier table 4.1.

Table: Frequency of items in transactional dataset

Itemset	**Frequency**
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

Table: Sorted Transactional dataset of supermarket based on frequency of items.

T_{ID}	**List of Items**
T_1	I_2,I_1,I_5
T_2	I_2,I_4
T_3	I_2,I_3
T_4	I_2,I_1,I_4
T_5	I_1,I_3
T_6	I_2,I_3
T_7	I_1,I_3
T_8	I_2,I_1,I_5,I_3
T_9	I_2,I_1,I_3

FP GROWTH TREE -

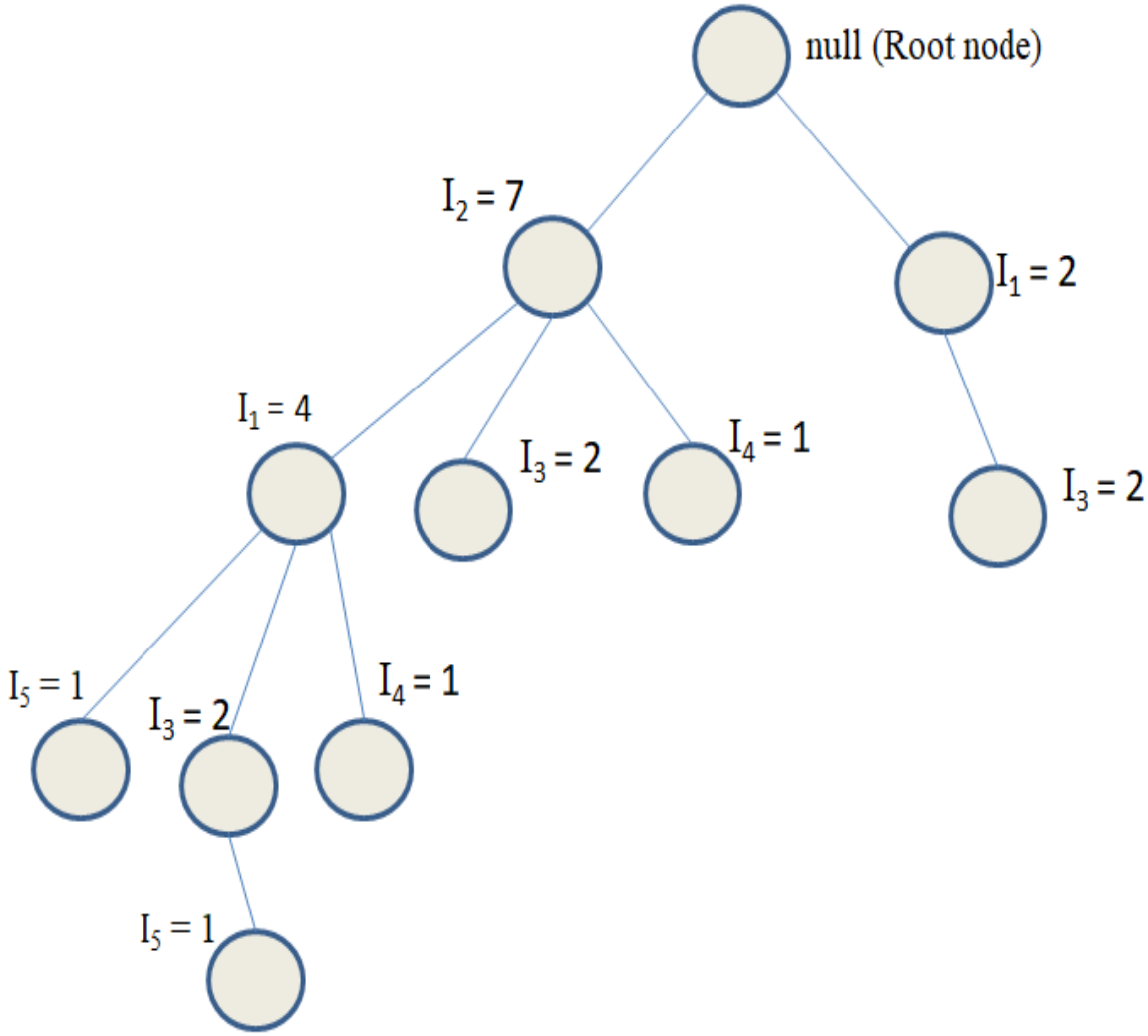

Fig 4.1: FP Tree for the above transaction table

Item	**Conditional pattern base**	**Conditional FP-Tree**	**Frequent Pattern Generated**
I_5	$\{\{I_2,I_1:1\},\{I_2,I_1,I_3:1\}\}$	$\{I_2:2,I_1:2\}$	$\{I_2,I_5:2\},\{I_1,I_5:2\},\{I_2,I_1,I_5:2\}$
I_4	$\{\{I_2,I_1:1\},\{I_2:1\}$	$\{I_2:2\}$	$\{I_2,I_4:2\}$
I_3	$\{\{I_2,I_1:2\},\{I_2:2\},\{I_1:2\}\}$	$\{I_2:4,I_1:2\},\{I_1:2\}$	$\{I_2,I_3:4\},\{I_1,I_3:4\},\{I_2,I_1,I_3:2\}$
I_1	$\{\{I_2:4\}\}$	$\{I_2:4\}$	$\{I_2,I_1:4\}$

Here $X_1 = \{I_2, I_1, I_3\}$ and $X_2 = \{I_2, I_1, I_5\}$ are the two most frequent itemsets obtained from the above table having a minimum support value equal to 2.

Association Rule		Confidence	Association Rule		Confidence
$I_1 \longrightarrow I_2, I_3$= 2/6	=	33%	$I_1 \longrightarrow I_2, I_5$= 2/6	=	33%
$I_2 \longrightarrow I_1, I_3$= 2/7	=	28%	$I_2 \longrightarrow I_1, I_5$= 2/7	=	28%
$I_3 \longrightarrow I_1, I_2$= 2/6	=	33%	$I_5 \longrightarrow I_1, I_2$= 2/2	=	100%
$I_2, I_3 \longrightarrow I_1$= 2/4	=	50%	$I_2, I_5 \longrightarrow I_1$= 2/2	=	100%
$I_1, I_3 \longrightarrow I_2$= 2/3	=	66%	$I_1, I_5 \longrightarrow I_2$= 2/2	=	100%
$I_1, I_3 \longrightarrow I_2$= 2/4	=	50%	$I_1, I_2 \longrightarrow I_5$= 2/4	=	50%

Hence, the strong association rules obtained through the FP growth algorithm are$\{I_5 \longrightarrow I_1, I_2\}, \{I_2, I_5 \longrightarrow I_1\}$ and$\{I_1, I_5 \longrightarrow I_2\}$, as these three rules satisfy both minimum Support and minimum Confidence values.

4.3 DAEM Algorithm

DAEM stands for Denoising Auto Encoder Mining algorithm. It is an unsupervised learning algorithm in which artificial neural network (ANN) is designed in such a way that it can perform tasks such as data encoding along with data decoding to reconstruct input. Data encoding is the process of mapping (sensory) input data to a different (often lower dimensional, compressed) feature representation. While, Data decoding is the process to map the feature representation back into the input data.

Denoising autoencoder is a complex mathematical model which trains on unlabelled as well as unclassified data and is used to map the input data to another compressed feature representation and from that feature representation reconstructing back the input data.

4.3.1 Principle of DAEM algorithm

In an autoencoder, there are two basic parts, an encoder, and a decoder.

First, the encoder takes the input and encodes it. For example, let the input data be x. Then, we can define the encoded function as f(x). Between the encoder and the decoder, there is

also an internal hidden layer. Let's call this hidden layer 'h'. This hidden layer learns the coding of the input that is defined by the encoder.

So, basically after the encoding, we get$h = f(x)$.

Finally, the decoder function tries to reconstruct the input data from the hidden layer coding. If we consider the decoder function as 'g', than the reconstruction can be defined as,

$$r = g\big(f(x)\big), r = g(h)$$

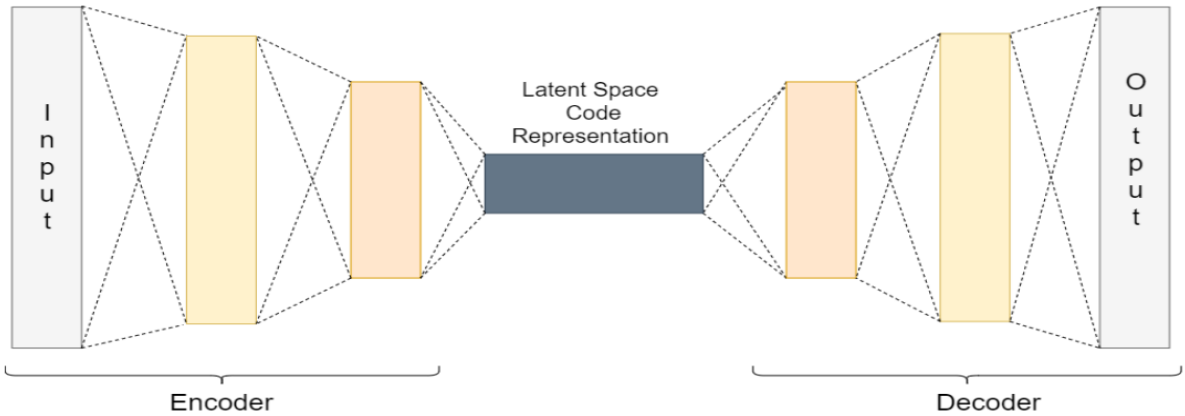

Fig 4.2: Working of an Auto-encoder

Denoising Autoencoder is an alternative to the concept of regular Autoencoder, which is prone to a high risk of overfitting. In the case of a **Denoising Autoencoder**, the data is **partially corrupted by noises** added to the input vector in a stochastic manner. Than, the model is trained to predict the original, **uncorrupted data point as its output.**

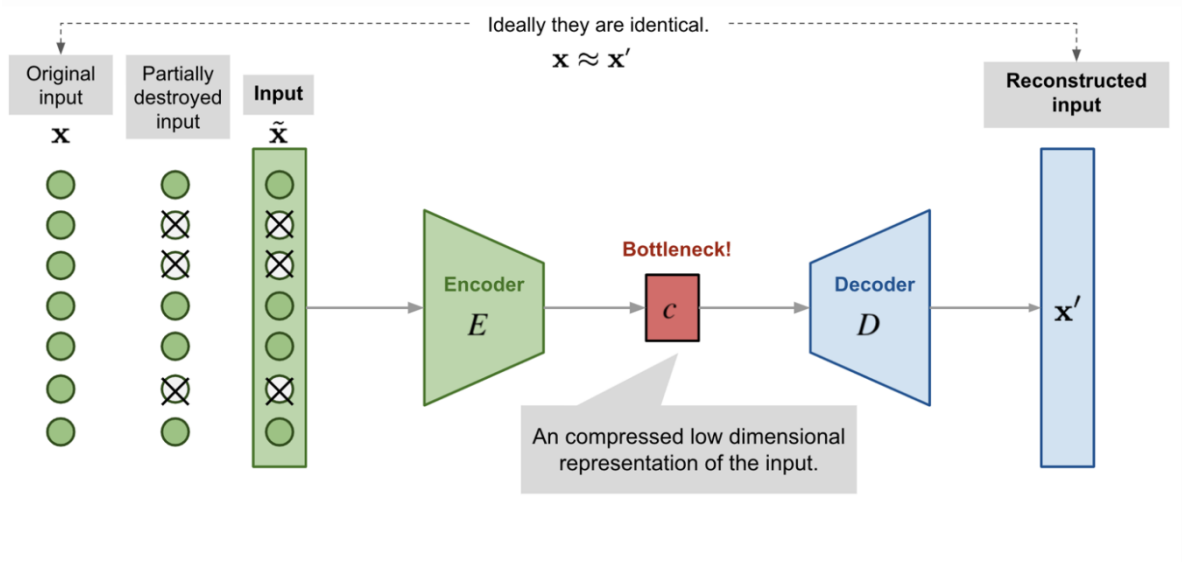

Fig 4.3: Architecture of Denoising autoencoder

This algorithm will follow a Denoising autoencoder technique for training to build compact representations of obtained association rules. Using this compact representation as mapping features, it will than train an ANN to predict target strong association rules from set of obtained association rules. Finally, it will help us to construct a deep neural network from the trained Denoising autoencoder and ANN weights, which were than fine-tuned using back-propagation. We compare the proposed method to existing methods using Gaussian mixture models and frame-selection. We evaluated the methods objectively and also conducted perceptual experiments to measure the execution time of algorithm. The results showed that, for 70 training sentences, frame-selection performed best, regarding both accuracy and quality. When using only two training sentences, the pre-trained deep neural network performed best, regarding both accuracy and quality.

- An input is sampled from our dataset.
- A corrupted version of this input is sampled from a stochastic mapping $m(x\backslash\widetilde{x})$.
- $(x, \widetilde{x})$is used as a training example

Just like regular AE, our DAE is a feedforward network which can be trained with a gradient-based approximate minimization on the negative log-likelihood

$$-\log_{p_{DAE}}(x, \tilde{x})$$

Here, p_{DAE} is the Denoising autoencoder reconstruction distribution.

As now we have defined a loss function, than during training.
We will aim to minimize the negative log-likelihood cost function.
This optimization leads to minimizing the distance between the corrupted input and the black manifold which characterizes our inputs. Thus, our model learns a reconstruction vector field$d(e(x)) - x$. Again, the goal here is not to make a *formal demonstration* but rather to give a first understanding of how we can train a denoising autoencoder.
So now with the image and the explanations above, we have a better understanding of a DAE's training. To conclude this theoretical part let us recall the three main advantages of this architecture.

- Learns more robust filters
- Prevents from learning a simple identify function
- Decreases the risk of overfitting that can be problematic with regular ae

4.3.2 Steps to implement DAEM algorithm

The steps to implement the DAEM algorithm are given below:

Step 1: Selection of the dataset.

Step 2: Create a validation set from the training set and visualize some data.

Step 2: To add a layer of Gaussian noise to implement our encoder.

Step 3: Than reshape our inputs and add some Gaussian noise

Step 5: Assemble the encoder and decoder and train the autoencoder end-to-end. (The optimization will be done based on a binary cross-entropy).

Step 6: Visualize the reconstructions for association rules.

CHAPTER 5

RESULT AND ANALYSIS

CHAPTER 5

RESULTS AND ANALYSIS

The datasets used for the experiment are the Grocery shop dataset and the supermarket dataset. These datasets are collected from the Data world. The parameters of the selected datasets are specified below.

Table 5.1: Parameters of Selected Datasets

Parameters	**Name of the Dataset**	
	Grocery Dataset	**Supermarket dataset**
No. of unique Itemsets	169	124
No. of transaction	9835	7501

In this project we do perform certain tasks for analysing the performance of different DM algorithms on the above datasets, this section is classified into two parts as Run-Time Analysis and Interestingness Analysis.

5.1.1 Data Visualization

In this section, we will look at the datasets that we have used to analyse the performance of different DM algorithms like Apriori algorithm, FP Growth algorithm, and DAEM algorithm (Optimised algorithm).

5.1.1 Data Preview

In this project, we take two different transactional dataset to compare different algorithm on various aspects. One Grocery dataset as shown in table 5.2, and other one is supermarket dataset as shown is table 5.3. In grocery dataset we have 9835 number of transaations with total 169 numbers of items and 7501 number of transactions in supermarket dataset with 124 numbers of items as shown in table no. 5.1.

Table number 5.4 and table number 5.5 will describe the grocery dataset and supermarket dataset respectively.

Table 5.2: Grocery dataset

	1	2	3	4	5	6	7	8	9	10
0	citrus fruit	semi-finished bread	margarine	ready soups	None	None	None	None	None	None
1	tropical fruit	yogurt	coffee	None	None	None	None	None	None	None
2	whole milk	None	None	None	None	None	None	None	None	None
3	pip fruit	yogurt	cream cheese	meat spreads	None	None	None	None	None	None
4	other vegetables	whole milk	condensed milk	long life bakery product	None	None	None	None	None	None

Table 5.3: Supermarket dataset

	1	2	3	4	5	6	7	8	9	10
0	shrimp	almonds	avocado	vegetables mix	green grapes	whole weat flour	yams	cottage cheese	energy drink	tomato juice
1	burgers	meatballs	eggs	None	None	None	None	None	None	None
2	chutney	None	None	None	None	None	None	None	None	None
3	turkey	avocado	None	None	None	None	None	None	None	None
4	mineral water	milk	energy bar	whole wheat rice	green tea	None	None	None	None	None

Table 5.4: Describing the grocery dataset using describe() function

	1	2	3	4	5	6	7	8	9	10
count	9835	7676	6033	4734	3729	2874	2229	1684	1246	896
unique	158	151	155	153	150	137	138	140	128	120
top	sausage	whole milk	whole milk	whole milk	rolls/buns	soda	soda	shopping bags	soda	shopping bags
freq	825	654	506	315	176	150	120	76	61	49

Table 5.5: Describing the Supermarket dataset using describe() function

	1	2	3	4	5	6	7	8	9	10
count	7501	5747	4389	3345	2529	1864	1369	981	654	395
unique	115	117	115	114	110	106	102	98	88	80
top	mineral water	mineral water	mineral water	mineral water	green tea	french fries	green tea	green tea	green tea	green tea
freq	577	484	375	201	153	107	96	67	57	31

5.1.2 Most frequent items in grocery dataset

The below graph fig 5.2 has shown the items that are most frequently occurs in grocery dataset, using the apriori algorithm.

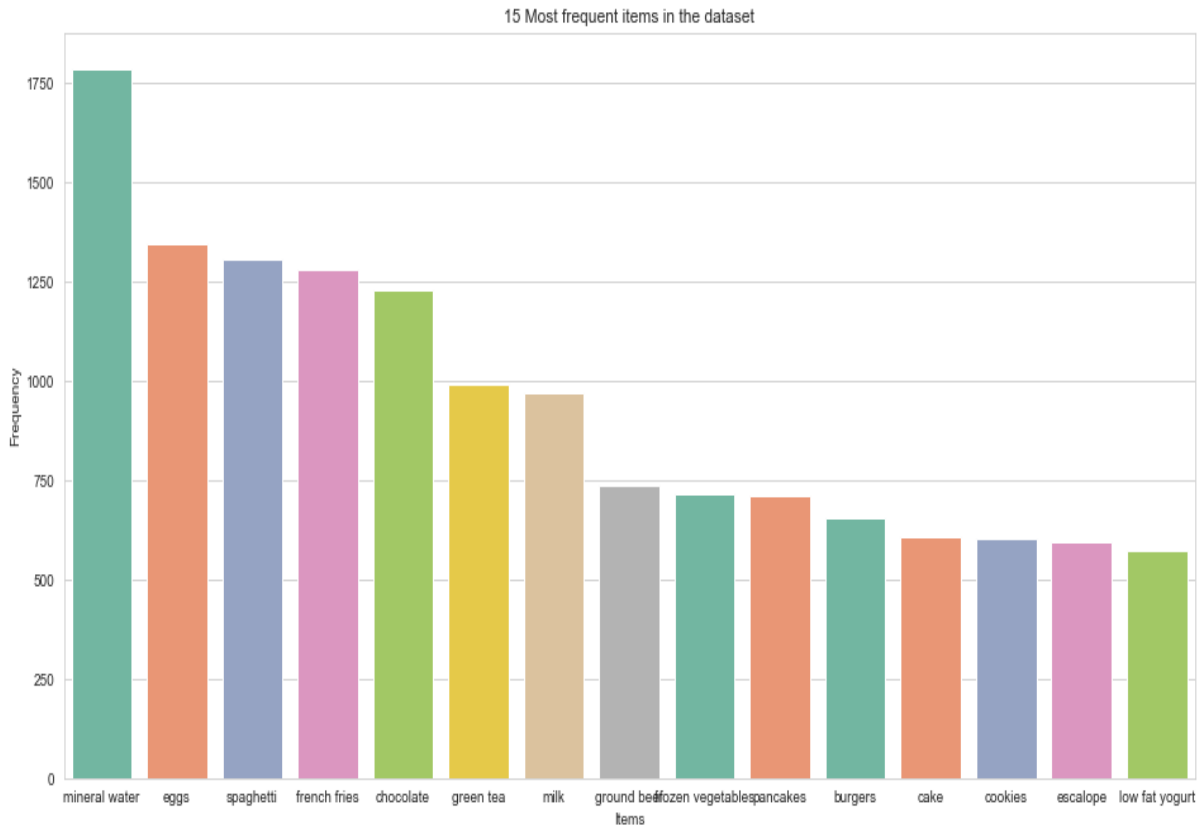

Fig 5.1: Most frequent items in Grocery dataset

5.1.3 Most frequent items in Supermarket dataset

The below graph fig 5.2 has shown the items that are most frequently occurs in grocery dataset, using Apriori algorithm.

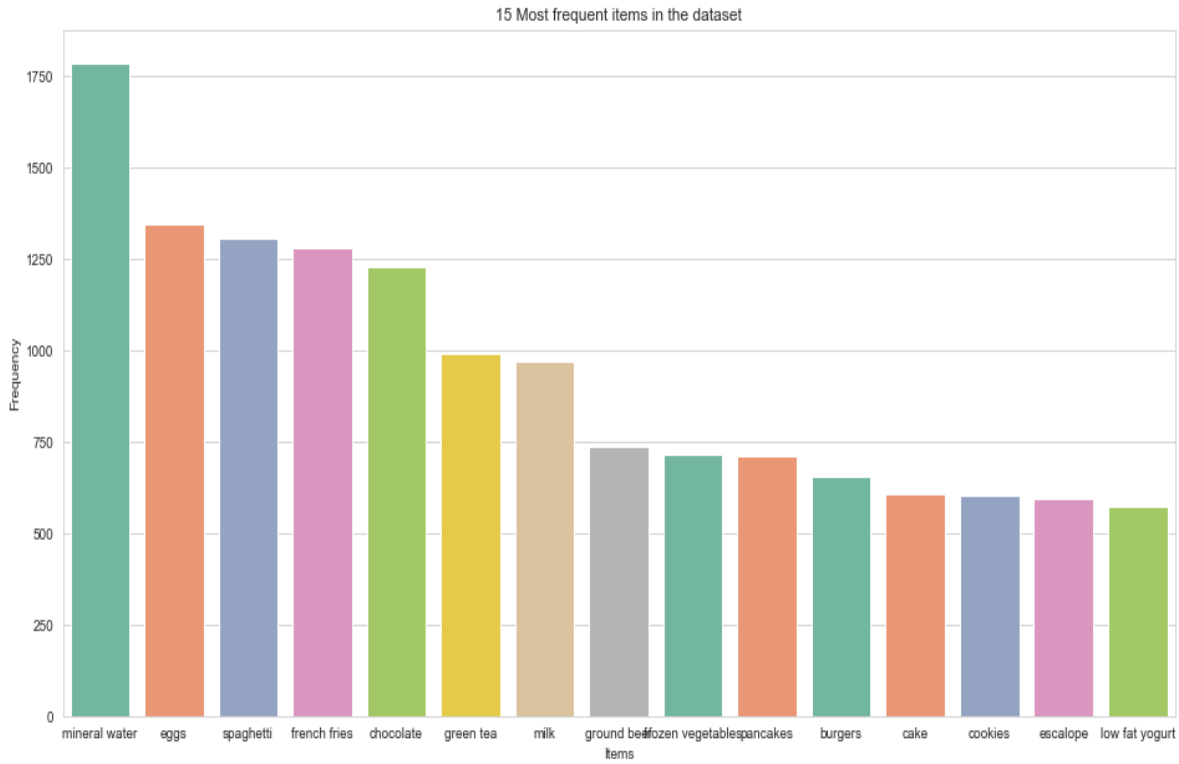

Fig 5.2: Most frequent items in Supermarket dataset

5.2 Association Rules obtained by different algorithms:

Table 5.6: Rules obtained by Apriori algorithm on Grocery dataset

	item1	item2	Support	Confidence	Lift	Conviction
0	['other vegetables', 'beef']	['pork']	0.004169	0.211340	3.665839	1.262688
1	['other vegetables', 'beef']	['root vegetables']	0.007931	0.402062	3.688692	1.659150
2	['beef', 'tropical fruit']	['other vegetables']	0.004474	0.586667	3.031985	2.408531
3	['rolls/buns', 'beef']	['root vegetables']	0.004982	0.365672	3.354833	1.568616
4	['beef', 'whole milk']	['root vegetables']	0.008033	0.377990	3.467851	1.594778
5	['beef', 'yogurt']	['root vegetables']	0.004575	0.391304	3.589998	1.635340

Table 5.7: Rules obtained by FP Growth algorithm on Grocery dataset

	Unnamed: 0	antecedents	consequents	support	confidence	lift	leverage	conviction
0	143	frozenset({'citrus fruit', 'root vegetables', ...	frozenset({'other vegetables'})	0.004474	0.785714	4.060694	0.003372	3.763701
1	1676	frozenset({'domestic eggs', 'curd'})	frozenset({'whole milk'})	0.004779	0.734375	2.874086	0.003116	2.802763
2	1616	frozenset({'butter', 'curd'})	frozenset({'whole milk'})	0.004881	0.716418	2.803808	0.003140	2.625286
3	3513	frozenset({'tropical fruit', 'yogurt', 'whippe...	frozenset({'whole milk'})	0.004372	0.704918	2.758802	0.002787	2.522974
4	581	frozenset({'root vegetables', 'yogurt', 'tropi...	frozenset({'whole milk'})	0.005694	0.700000	2.739554	0.003616	2.481613
5	1070	frozenset({'other vegetables', 'yogurt', 'butt...	frozenset({'whole milk'})	0.004372	0.682540	2.671221	0.002735	2.345125

Table 5.8: Rules obtained by DAEM algorithm on Grocery dataset

	Item 1	Item 2	support	confidence	lift	Unnamed: 8	conviction
0	{yogurt}	{whole milk}	0.600000	0.470000	1.570000	NaN	1.240000
1	{yogurt}	{whole milk}	0.600000	0.470000	1.570000	NaN	1.240000
2	{yogurt}	{whole milk}	0.600000	0.470000	1.500000	NaN	1.240000
3	{yogurt}	{whole milk}	0.600000	0.470000	1.570000	NaN	1.240000
4	{whole milk}	{soda}	0.400000	0.160000	1.570000	NaN	1.240000
5	{'zwieback'}	{whole milk}	0.090000	0.250000	0.980000	NaN	0.990000

The tables above from 5.6 to table 5.8 shows the rules obtained using Ariori, FP Growth and DAEM algorithm respectively on grocery dataset.

Table 5.9: Rules obtained by Apriori algorithm on Supermarket dataset

	item1	item2	Support	Confidence	Lift	Conviction
0	['chocolate', 'frozen vegetables']	['shrimp']	0.005333	0.232558	3.254512	1.296082
1	['ground beef', 'cooking oil']	['spaghetti']	0.004799	0.571429	3.281995	2.322135
2	['ground beef', 'eggs']	['herb & pepper']	0.004133	0.206667	4.178455	1.255295
3	['spaghetti', 'frozen vegetables']	['ground beef']	0.008666	0.311005	3.165328	1.438812
4	['milk', 'frozen vegetables']	['olive oil']	0.004799	0.203390	3.088314	1.249294
5	['tomatoes', 'milk']	['frozen vegetables']	0.004133	0.295238	3.097316	1.413055

Table 5.10: Rules obtained by FP Growth algorithm on Supermarket dataset

	Unnamed: 0	antecedents	consequents	support	confidence	lift	leverage	conviction
0	1234	frozenset({'frozen vegetables', 'soup'})	frozenset({'mineral water'})	0.005066	0.633333	2.656954	0.003159	2.077178
1	1210	frozenset({'soup', 'pancakes'})	frozenset({'mineral water'})	0.004266	0.627451	2.632276	0.002645	2.044380
2	386	frozenset({'olive oil', 'tomatoes'})	frozenset({'spaghetti'})	0.004399	0.611111	3.509912	0.003146	2.123717
3	1101	frozenset({'whole wheat rice', 'pancakes'})	frozenset({'mineral water'})	0.004133	0.596154	2.500979	0.002480	1.885945
4	1570	frozenset({'cooking oil', 'pancakes'})	frozenset({'mineral water'})	0.004666	0.593220	2.488672	0.002791	1.872345
5	1228	frozenset({'olive oil', 'soup'})	frozenset({'mineral water'})	0.005199	0.582090	2.441976	0.003070	1.822476

Table 5.11: Rules obtained by DAEM algorithm on Supermarket dataset

	Item 1	Item 2	support	confidence	lift	Unnamed: 8	conviction
0	{yogurt cake}	{whole wheat rice}	0.130000	0.060000	0.010000	NaN	3.620000
1	{yogurt cake}	{whole wheat rice}	0.130000	0.060000	0.010000	NaN	3.620000
2	{whole wheat rice}	{turkey}	0.530000	0.120000	0.030000	NaN	4.200000
3	{whole wheat rice}	{turkey}	0.530000	0.120000	0.030000	NaN	4.200000
4	{yogurt cake}	{whole wheat rice}	0.130000	0.060000	0.010000	NaN	3.620000
5	{yogurt cake}	{whole wheat rice}	0.130000	0.060000	0.010000	NaN	3.620000

The tables above from 5.9 to table 5.11 shows the rules obtained by Ariori, FP Growth, and DAEM algorithm respectively on Supermarket dataset.

5.3 Runtime Analysis

In this section, difference in the runtime is observed for different support values of the three algorithms, along with their performance. The lower the Run time, the higher the performance is counted here. The Runtime analyses of the datasets are shown in the table 5.12.

Table 5.12: Run time (second) analysis of the algorithms for different Support values for both datasets.

Dataset	Min Support Value	Apriori Algorithm	Fp Growth Algorithm	Optimized Algorithm
Grocery Dataset	0.002	112.18	78.4	60.00
	0.004	90.98	70.63	54.39
	0.006	85.70	61.8	49.20
	0.008	76.24	55.73	42.70
	0.010	72.45	52.31	41.15
Supermarket Dataset	0.002	90.03	57.85	43.60
	0.004	67.82	49.28	37.43
	0.006	60.79	45.41	35.30
	0.008	57.43	42.60	33.50
	0.010	54.52	43.86	31.15

The table above shows the execution time of different algorithms such as Apriori, FP Growth, and DAEM algorithm (Optimised algorithm) for different values of minimum support.

From the above table, it is clearly visible that the Apriori algorithm takes a comparatively higher execution time as compared to FP Growth algorithm and DAEM Algorithm. This observation can be shown as in Fig 5.3 for grocery dataset, and in fig 5.4 for the Supermarket dataset

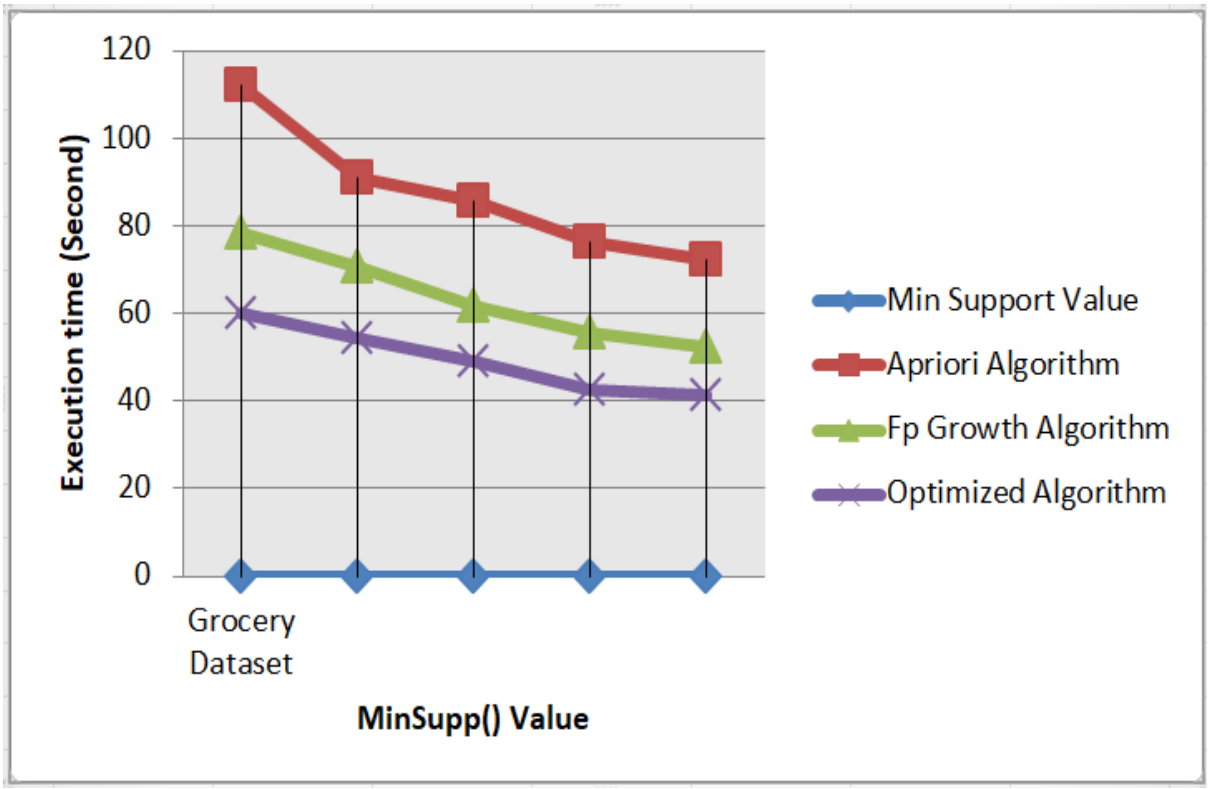

Fig 5.3: Run time analysis for different Support values of Grocery dataset

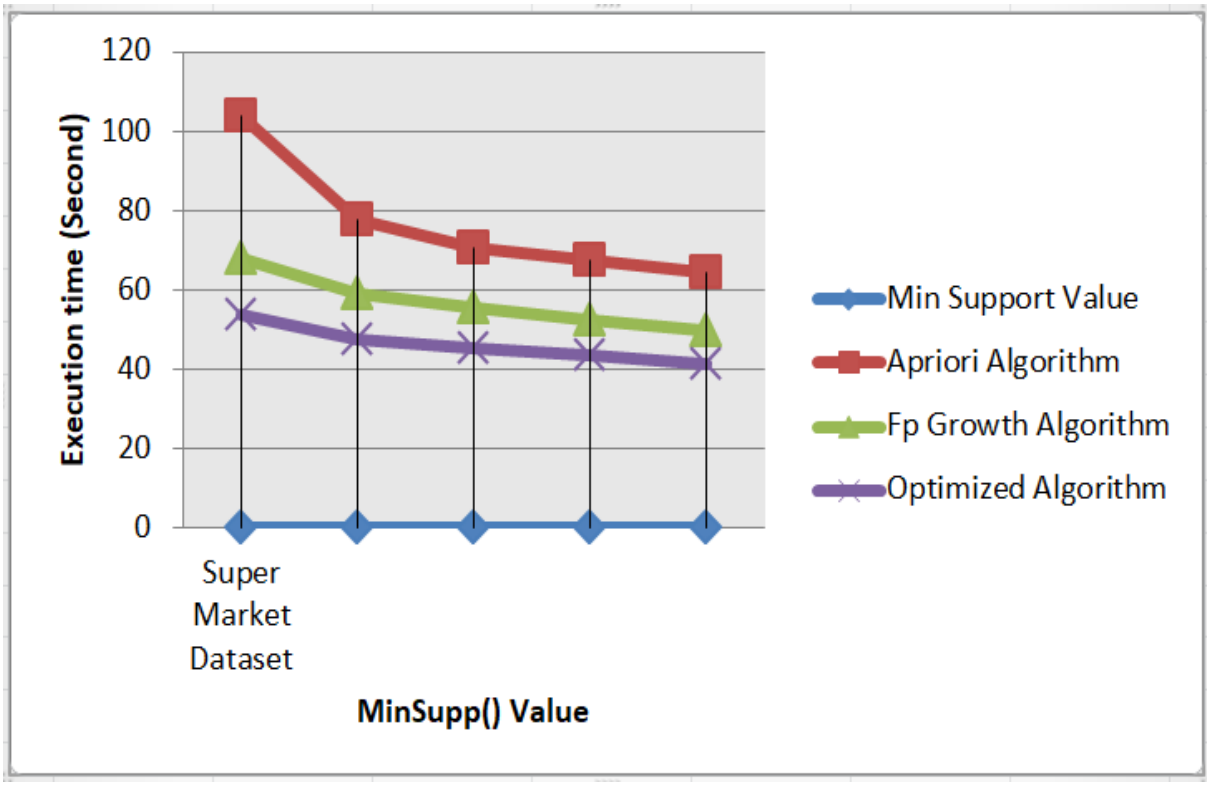

Fig 5.4: Run time analysis for different Support values of Supermarket dataset

We can observe from Fig 5.3 and Fig 5.4 that Apriori algorithm has a higher Runtime than the FP Growth and DAEM algorithm. The ANN DM algorithm takes the least execution time for both datasets. But in each case, runtime is decreasing rapidly with the increase in the value of support. As the Apriori algorithm scans the whole dataset repeatedly, it requires more execution time while FP Growth algorithm and Optimized algorithm using ANN does not do so. From here, it can be concluded that the algorithms using the ANN concepts perform better according to runtime.

In the next part, we are analysing the interestingness of the rules generated by algorithms that have comparatively better performance based on runtime.

5.4 Interestingness Analysis

Interestingness defines the acceptability of a rule. For this measurement, there are several types of Interestingness measures. We are measuring the Average Interestingness of the rules of the ARM algorithms – FP Growth and optimised ANN algorithm as they performed better according to runtime in the previous section.

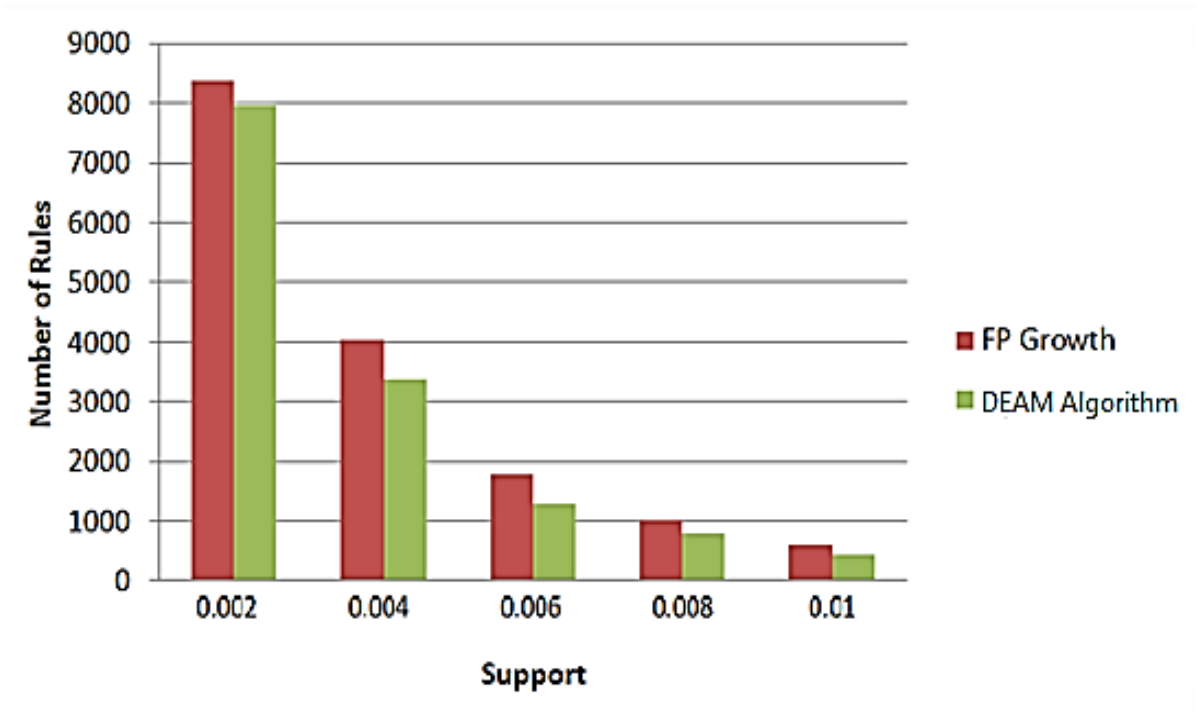

Fig 5.5: Performance analysis of the algorithms based on the number of rules for the Grocery dataset

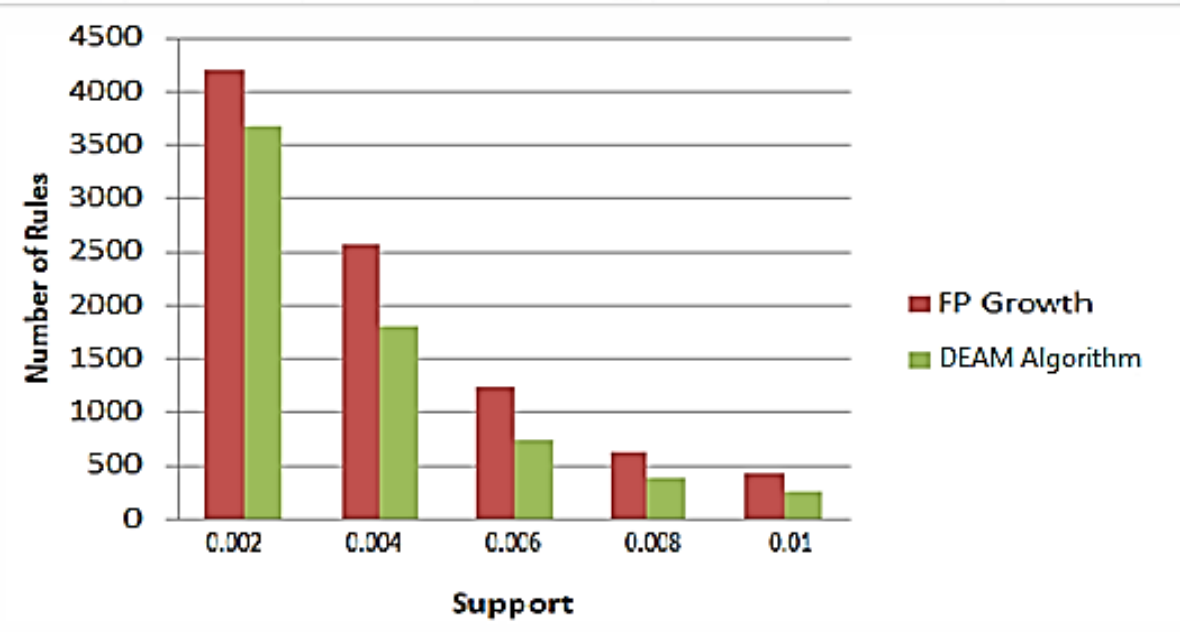

Fig 5.6**:** Performance analysis of the algorithms based on a number of rules for the Supermarket dataset

From Fig 5.5 and Fig 5.6, we can observe that the Optimized algorithm generates fewer rules than the FP-Growth algorithm in the case of both datasets. It is regarded as less efficient if the number of rules generated by any algorithm is high. As the Optimized algorithm generates fewer rules than the FP-Growth algorithm, it is regarded as the better one in performance. The following table describes the Average Interestingness of two different algorithms.

From the vast experimental analysis, it can be concluded that –

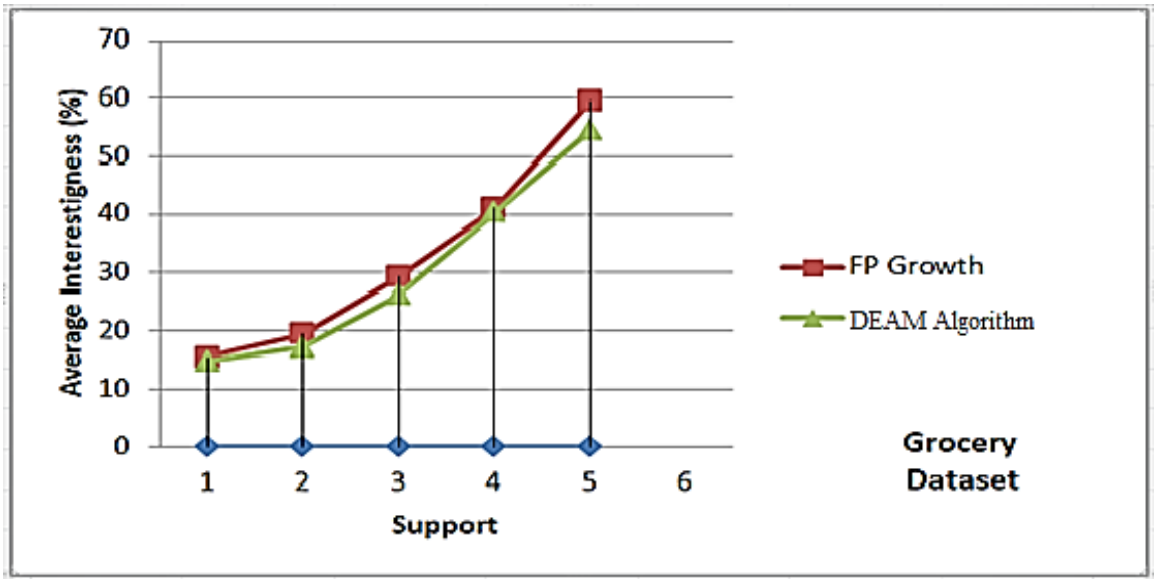

Fig 5.7: Average Interestingness analysis of rules for different Support values of Grocery Dataset

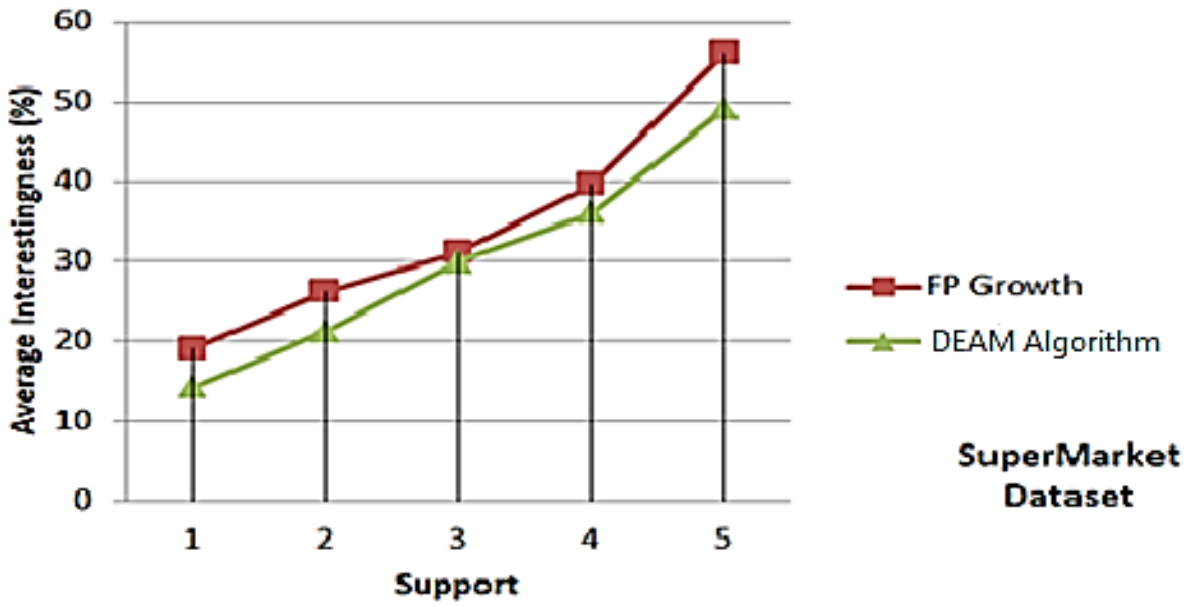

Fig 5.8: Average Interestingness analysis of rules for different Support values for Supermarket Dataset

From Fig 5.7 and Fig 5.8 we can depict that in both cases Average Interestingness is higher for the Optimised algorithm. This algorithm performs some complex steps for rule mining and so in spite of requiring much more run time than the FP Growth, it provides rules with better interestingness. So, it can be assured from the analysing results that the Apriori algorithm outperforms the FP-Growth algorithm.

- The FP Growth algorithm performs better than the Apriori and the Optimized algorithm, based on runtime.
- The Optimized algorithm has better accuracy than the FP Growth algorithm in case of average interestingness.
- Superseding runtime, focusing on the interestingness of the rules of FP Growth algorithms observed, it can be said that the Optimized algorithm performs better than the FP Growth algorithm.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

CHAPTER 6
CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

The observation in this project shows; how applications of ANNs in DM perform better than the existing DM algorithms. While the focusing on ARM, and its key measures: support, confidence, lift, and conviction to find strong association rules in a large transactional dataset. There is a complete comparison between the two most popular DM algorithms that is Apriori and FP Growth. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important to develop a powerful means for analysis and perhaps the interpretation of such a huge data that is exponentially increasing day by day and for the extraction of interesting knowledge that could help in decision-making. Thus DM is the process of automated extraction of hidden patterns, and predictive information from large databases. DM includes extraction, transformation, and a load of transaction data onto the data warehouse system.

The DAEM algorithm will follow a Denoising autoencoder (DAE) technique for training to build compact representations of obtained association rules. Using this compact representation as mapping features, we then trained an ANN to predict target strong association rules from the set of obtained association rules. Finally, we constructed a DNN from the trained DAE and NNs weights, which were then fine-tuned using back-propagation. We compared the proposed method to existing methods using Gaussian mixture models and frame selection. We evaluated the methods objectively, and also conducted perceptual experiments to measure the execution time of the algorithm. The results showed that the DAEM algorithm performed best, in both execution time and efficiency. To further simplify the process, DAE has been implemented as a pre-training method for rule generation and to reduce the training time. Along with this, the activation function and some sparse rules are also used and added to pre-train the models, which will eventually result in the efficient extraction of strong association rules and patterns. The observation is made to explore the application of ANNs in DM techniques, and the key technology and ways to achieve the DM based on neural networks are also researched. Given the current state of the art, ANN deserves a place in the toolboxes of data-mining specialists and it has better in terms of run-time, efficiency, and interestingness than the Apriori and FP-Growth the two popular DM algorithms.

6.2 Future work

- This project will focus on improving the DM techniques by incorporating it with latest technology like ANN, and other such algorithm.
- Integrates it with other algorithm in order to boost up its application in technology like GAN algorithm. Generative Adversarial Networks (GANs) are a powerful class of neural networks that are used for unsupervised learning. It is very useful in retail market when it is integrated with DM.
- Increasing Automation in DM as earlier incarnations of DM involved manual coding by specialists with a deep background in statistics and programming. Modern techniques are highly automated, with AI/ML replacing most of these previously manual processes for developing pattern-discovering algorithms.
- Embedded DM features are increasingly finding their way into a myriad of enterprise software use cases, from sales forecasting in CRM SaaS platforms to cyber threat detection in intrusion detection/prevention systems. The embedding of DM into vertical market software applications enables prediction capabilities for any number of industries and opens up new realms of possibilities for unique value creation.

REFERENCES

1. Tasnia Rahman et al., "Performance Evaluation of Fuzzy Association Rule Mining Algorithms", 4th International Conference on Electrical Information and Communication Technology (EICT), pp.272-275, 2019.
2. Vlad Robu and Vitor Duarte dos Santos, "Mining Frequent Patterns in Data Using Apriori and Eclat", 6th International Conference on Systems and Informatics (ICSAI), pp.1478-1481, 2019.
3. Liya Ma et al., "Frequent Itemsets Mining of SCADA Data based on FP-Growth Algorithm", 4th IEEE Conference on Energy Internet and Energy System Integration, pp.3057-3061, 2020.
4. Islamiyah et al., "Comparison of Apriori and FP-Growth Algorithms in Determining Association Rules", International Conference on Electrical, Electronics and Information Engineering (ICEEIE), vol.4, pp.320-323, 2020.
5. J. Sivapriya et al., "Comparative Study of APRIORI and FP Algorithm for Decision Making", 3rd International Conference on Trends in Electronics and Informatics (ICOEI), pp. 1058-1061, 2019.
6. Surbhi K. Solanki and Jalpa T. Patel, “A Survey on Association Rule Mining”, Fifth International Conference on Advanced Computing & Communication Technologies, Volume 5, pp. 212-216, 2015.
7. P. Amaranatha Reddy and MHM Krishna Prasad, “Challenges to find Association Rules over various types of data items: a Survey”, International Conference on Computing, Communication and Automation, pp. 180-184, 2017.
8. Mrs. Geeta S. Navale and Drs. Suresh N. Mali, “A Survey on Sensitive Association Rules Hiding Methods”, Third International Conference on Computing, Communication, Control And Automation (ICCUBEA), IEEE, 2017.
9. L Greeshma and Dr. G Pradeepini, “Unique Constraint Frequent ItemSet Mining”, 6th International Conference on Advanced Computing, IEEE, pp. 68-72, 2016.
10. P.Naresh and Dr.R.Suguna, "Association Rule Mining Algorithms on Large and Small Datasets: A Comparative Study", Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 587-592, 2019.
11. Guo Junfeng and Hou Handan, "Statistical Decision Research of long-term deposit subscription in banks based on Decision Tree", International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), pp. 614-617, 2019.

12. Mr. Nilesh Kumar Dokania and Ms. Navneet Kaur, "Comparative Study of Various Techniques in Data Mining", International Journal of Engineering Sciences & Research Technology, pp. 202-209, 2018.

13. Nima Norouzi and Elham Ataei, "Application of data mining in identifying and discovering hidden patterns of theft", International Journal of Innovative Research in the Humanities, Vol.1, No.1 , pp. 29-42, 2021.

14. Yasmeen Alhamdan, et al., "CoShopper - Leveraging Artificial Intelligence for an Enhanced Augmented Reality Grocery Shopping Experience", IEEE International Conference on Artificial Intelligence and Virtual Reality(AIVR), pp. 337-338, 2020.

15. J. M. Luna, et al., "Optimization of quality measures in association rule mining: An empirical study", International Journal of Computational Intelligence Systems, November 2018.

16. Sikha Bagui and Probal Chandra Dhar, "Positive and negative association rule mining in Hadoop's MapReduce environment", Journal of Big Data, pp. 1-16, 2019.

17. Fuguang Bao, et al., "An Improved Evaluation Methodology for Mining Association Rules", Axioms by MDPI, vol. 11, pp. 1-16, 2022.

18. K.Kala, "An Optimization based Algorithm for Association Rule Mining in Large Databases", International Journal of Recent Technology and Engineering (IJRTE), vol. 8, issue 3, pp. 3746-3754, September 2019.

19. Chitra Jalota and Rashmi Agrawal, "Analysis of Educational Data Mining using Classification", International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), pp. 243-247, February 2019.

20. Yuzi Doua, et al., "Application of Improved Eclat Algorithm in Student's Evaluation of Teaching", MATEC Web of Conferences, pp. 1-5, 2018.

21. Junrui Yang and Lin Xu, "A novel interestingness measure based on a fusion model for association rules mining", MATEC Web of Conferences, pp. 1-6, 2021.

22. Anshu, "Review Paper on Data Mining Techniques and Applications", International Journal of Innovative Research in Computer Science & Technology (IJIRCST), vol. 7, issue 2, pp. 22-26, March 2019

23. Teng Xiuyi and Gong Yuxia, "Research on Application of Machine Learning in Data Mining", IOP Conference Series: Materials Science and Engineering, pp. 1-5, 2018.

24. Halil İbrahim Dede et al., "Comparison of Frequent Pattern Mining Algorithms in Internet of Things", 28th Signal Processing and Communications Applications Conference (SIU), pp.1057-1060, 2021.

25. Kai Xu, "Fault Diagnosis Method of Power System Based on Neural Network", International Conference on Virtual Reality and Intelligent Systems, pp. 172-175, 2018.

26. Mohammadreza Ramzanpour and Simone A. Ludwig, "Association Rule Mining Based Algorithm for Recovery of Silent Data Corruption in Convolutional Neural Network Data Storage", IEEE Symposium Series on Computational Intelligence (SSCI), pp. 3057-3064, 2020.

27. Sempe Leholo et al., "Solar Energy Potential Forecasting and Optimization using Artificial Neural Network: South Africa Case Study", Institute of Electrical and Electronics Engineers, pp. 533-536, 2019.

28. Manomita Chakraborty et al., "Data Mining Using Neural Networks in the form of Classification Rules: A Review", 4th International Conference on Computational Intelligence and Networks (CINE), pp. 332-337, 2020.

29. Biswajit Panja, et al., "Crime Analysis Mapping, Intrusion Detection - using Data Mining", IEEE Technology and Engineering Management Conference (TEMSCON), pp. 255-259, 2020.

30. Tianyu Xia et al., "Data Association Rules Mining Method Based on RBF Neural Network Optimization Algorithm", International Conference on Communications, Information System and Computer Engineering (CISCE), pp. 433-436, 2020.

31. X. Y. Kong et al., “Fault diagnosis method of a distribution network based on cloud-edge computing architecture and wavelet neural network”, Distribution & Utilization, vol.37, pp. 17-23, 2020.

32. Y. H. Huang et al., "Research on weak point analysis of distribution networks based on FP-Growth algorithm", Electrical Measurement & Instrumentation, pp. 1-7, 2020.

33. Hossein Riahi-Madvar et al., "Derivation of optimized equations for estimation of dispersion coefficient in natural streams using hybridized ANN with PSO and CSO algorithms", IEEE Access, vol.8, pp.156582-156599, 2020.

34. Jin Bo Chen et al., "A Log Analysis Technology Based on FP-growth Improved Algorithm", International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA), pp.1096-1101, 2021.

35. Wisnu Arya Dipa and Wikan Danar Sunindyo, "Software Defect Prediction Using SMOTE and Artificial Neural Network", International Conference on Data and Software Engineering (ICoDSE), pp. 1049-1054, 2021.

36. Emny Harna Yossy et al., “Comparison of Data Mining Classification Algorithms for Student Performance”, IEEE International Conference on Engineering, Technology and Education (TALE), pp.351-354, 2019.

37. Ahmad Ari Aldino et al., "Comparison Of Market Basket Analysis To Determine Consumer Purchasing Patterns Using Fp-Growth And Apriori Algorithm", International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), pp.1432-1437, 2021.

38. Sanket Sandip Khedkar and Sangeeta Kumari, "Market Basket Analysis using A-Priori Algorithm and FP-Tree Algorithm", International Conference on Artificial Intelligence and Machine Vision (AIMV), pp.2030-2034, 2022.

39. Xiaoxia Lin et al., "Analysis and research on hidden danger of coal mine safety based on improved FP-growth algorithm", IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), vol.4, pp.1397-1400, 2021.

LIST OF PUBLICATION

1. Aditya Shukla, Pankaj Kumar Gond and Dr. Harvendra Kumar Patel, "A Comparative Study of Association Rule Mining and Its Preliminaries", Intelligent Systems - Proceedings of International Conference on Intelligent Systems, Article 9, 2022.

2. Satish Sahani, Neha Gond, Dr. Harvendra Kumar Patel, "An In depth study of the different Data Mining And Machine Learning Techniques", International Journal of Advances in Engineering and Management, Volume 4, Issue 5, May 2022.

3. Pankaj Kumar Gond, Aditya Shukla, Satish Sahani, Neha Gond, Dr. Harvendra Kumar Patel, "Association Rule Mining using FP-Growth and An Innovative Artificial Neural Network Techniques", International Journal for Research in Applied Science and Engineering Technology, Volume 10, May 2022.

Pankaj Kumar Gond

Bachelor of Technology
(Information Technology)

ABOUT ME

I am very passionate about my works, especially in the field of Information Technology and Engineering. I like to learn different tools and technology and try to integrate them with other technology. I am interested in Web application development, Software development, Programming, and Data Science.

EDUCATION

Bachelor of Technology (2018-Present)
Information Technology Agg. 69.72%
From: Institute of Technology and Management, Gorakhpur

Intermediate (10+2) (2016-2018)
Maharana Pratap Senior Secondary School, Gorakhpur (UP Board) Grade A 66.66%

High School (2014-2016)
From: Mahatma Gandhi Inter College, Gorakhpur (UP Board) Grade A 79%

INDUSTRIAL TRAINING

Data Science (Training) (June'2020-July'2020)
Internshala
The training consisted of Introduction to Data Science, Python for Data Science, Statistics for Data Science and Predictive Modeling and Basics of Machine Learning.

Android Programming and Information Security Internship (June-July'19)
NIELIT Gorakhpur
1. Build an Android application that used proper encryption to save users data and secure its data base from the hackers.
2. Debugging and Testing of Android application.

PROJECTS

Virtual Lab Development (August-2020)
Virtual Labs, MHRD Govt. of India
Awarded as **Gold Developer** in eBothathon-3 by AKTU Lucknow.
Under the guidance of IIT-Kanpur and AKTU-Lucknow, for building labs of Data Structure and Engineering Chemistry.

Analysis of Covid-19 (April-2021)
Data Science and Machine Learning Project
Analysis of Covid19 data from March to October and building a model to predict cases for next 15 days.

ANN Approach for Association Rule Mining in transactional Dataset (Aug'21-May'22)
Data Mining and Machine Learning Project
An Approch to explore the application of ANNs in the field of data mining with the help of Denoising Algorithm, it reduce the execution time to generate association rules compare to existing algorithms.

PUBLICATIONS

1. Aditya Shukla, Pankaj Kumar Gond and Dr. Harvendra Kumar, "A Comparative Study of Association Rule Mining and Its Preliminaries", Intelligent Systems - Proceedings of ICIS-2022, Article 9, 2022.
2. Pankaj Kumar Gond, et.al., "Association Rule Mining using FP-Growth and An Innovative Artificial Neural Network Techniques", IJRASET, Volume 10, May 2022.

CONTACT INFO.

pankajkumargond79@gmail.com
https://bit.ly/3fP9Dxa
https://github.com/pankaj-creator
www.codechef.com/users/pankaj_gkp

Address: Mirzapur, Gorakhpur, UP, India (273005)

SKILLS

WEB DEVELOPMENT
SOFTWARE DEVLOPMENT
DATA ANALYTICS
PYTHON PROGRAMING
C++ PROGRAMMING

EXTRACURRICULAR ACTIVITIES

International Conference on Intelligent System (ICIS-2022) Mar'22
Uttaranchal University

CNSS Certified Network Security Specialist July'20
ICSI – United Kingdom

International Conference on Contemporary and Application (IC3A 2020) Feb'20
AKTU Lucknow

Master Diploma in Information Technology (2016-2017)
From: Education World, Gorakhpur Grade A+ 80%

ACHIEVEMENTS AND AWARDS

1. **Gold Developer Certificate** for virtual lab development in eBothathon03, by AKTU Lucknow and IIT-Kanpur.
2. Secured **242 global ranks** in Codechef August Long challenge.
3. **3 star rating in Codechef**.
4. **2nd Rank** in Code Monk in Parampara'22.

HOBBIES

Coding, Testing tools, Reading tech-blog, listening podcasts and travelling.

LANGUAGE KNOWN

1. ENGLISH
2. HINDI

Aditya Shukla

Bachelor of Technology (Information Technology)

Rustampur, Gorakhpur, UP, India (273016)
E-mail: aditya11111999@gmail.com

OBJECTIVE

To become a successful engineer in the field of Information Technology by channelizing my technical knowledge and skills to ensure personal and professional growth and to contribute to the prosperity of the organization. Interested in Web application development, Software development, Programming, and Machine Learning.

EDUCATIONAL QUALIFICATION

CLASS	YEAR	NAME OF SCHOOL / COLLEGE	NAME OF BOARD/ UNIVERSITY	% OF MARKS
10th	**2016**	St. Paul's School	ICSE Board	**81%**
12th	**2018**	Gorakhpur Public School	CBSE Board	**66.66%**
B.TECH. 1st Year	**2019**	Institute of Technology and Management, Gida, Gorakhpur	AKTU Lucknow	**71.85%**
B.TECH. 2nd Year	**2020**	Institute of Technology and Management, Gida, Gorakhpur	AKTU Lucknow	**69.40%**
B.TECH. 3rd Year	**2021**	Institute of Technology and Management, Gida, Gorakhpur	AKTU Lucknow	**74.43%**
B.TECH. 4th Year (Till 7th Sem)	**2022**	Institute of Technology and Management, Gida, Gorakhpur	AKTU Lucknow	**76.94%**

COMPUTER PROFECIENCY

PLATFORMS KNOWN: OS – Window 7/10, Linux Ubuntu and Kali.

SOFTWARE – VS CODE, ANACONDA, MS OFFICE, and other.

LANGUAGES: C, C++, Python, Java, PHP and JavaScript.
FRAMEWORKS: Django and Flask

AREA OF INTEREST

Software Development Engineering, Web Development and Machine Learning.

CONFERENCES, WORKSHOPS and SEMINAR

Conference: International Conference on Intelligent Systems (ICIS 2022) – Uttaranchal University, Dehradun.
Workshops: Ethical Hacking Workshop (MMMUT)
Seminars: Virtual &Augmented Reality (ITM), and many other.

PROJECT

1. **Artificial Neural Network Approach For Association Rule Mining In Transactional Dataset –** This project is an initiative to utilize ANN techniques in the field of Data mining.
2. **IPL Prediction Master** – This project predicts name of winning team and other details with regards to current match.
3. **Periodic Table** – This project is used to showcase the periodic chart of Chemistry along with details of and about each particular element.

PUBLICATIONS

Conference Papers in Review

Aditya Shukla, Pankaj Kumar Gond and Dr. Harvendra Kumar, "A Comparative Study of Association Rule Mining and Its Preliminaries", Intelligent Systems - Proceedings of ICIS-2022, Article 9, 2022.

Journal Papers in Research

Pankaj Kumar Gond et al., "Association Rule Mining using FP-Growth and An Innovative Artificial Neural Network Techniques", IJRASET, Volume 10, May 2022.

INDUSTRIAL TRAINING

C and C++: NIELIT GORAKHPUR **(June-July`19)**

1. Developed a program on periodic table using the advanced concepts of these languages.
2. Further improved the program by even adding all the possible details about the elements in it.

Python with Machine Learning: EduFruit **(June-July`20)**

The training consisted of Introduction to Python, Machine Learning, Understanding the Statistics for Data Science and Predictive Modeling and Basics of Machine Learning.

1. Using regression analysis, and python statistics, a model which analyze IPL performance.

SPECIAL ACHIEVEMENTS

1. Awarded Certificate of Merit for securing third rank in my first year.
2. 2 star rating in codechef.
3. Received a chance to present my presentation in ICIS-2022.
4. Secured 3rd Rank in Code Monk challenge in Parampara 2022.
5. Secured 1st Rank in Code Debugging challenge in Parampara 2022.

PERSONAL INFORMATION

Father's Name:	Mr. Vinay Kumar Shukla
Date of Birth:	11/11/2000
Professional Status:	Student
Nationality:	Indian
Gender:	Male
Marital Status:	Single
Mobile no:	6306482868
E-Mail Address:	aditya11111999@gmail.com

I hereby declare that all the above-mentioned information is true to the best of my knowledge.

Neha Gond

Student

Hardworking, Goal oriented, determined and focused with a keen eye for details. Skilled in Frontend & Website development & programming language including C, Java, Python, PHP & SQL. **I am always energetic and eager to learn new skills.**

nehag63061@gmail.com | 6306133757 | Gorakhpur, India | linkedin.com/in/neha-gond-71b96721b

EDUCATION

UG

Institute of Technology & Management Gida Gorakhpur

08/2018 - Present, *76%*

Courses

- Bachelor of technology
- Information Technology

Class 12th

SVMBIC Arya Nagar North Gorakhnath Gorakhpur

2018, *82.2%*

Courses

- PCM

Class 10th

SVMBIC Arya Nagar North Gorakhnath Gorakhpur

2016, *88.5%*

INTERNSHIP AND TRAINING

Website Design & Development

Gorakhpur/EASY CAD SOLUTIONS

07/2019 - 08/2019,

Achievements/Tasks

- Basic Knowledge of HTML,CSS,PHP,Wordpress

Python with Machine Learning

Gorakhpur/edufruit

08/2020 - 09/2020,

Achievements/Tasks

- basic knowledge of python & machine learning

EXTRA CARRICULUM

A Research on an In Depth Study of the Different Data Mining and Machine Learning Techniques, Published in IJAEM

A Research on Association Rule Mining using FP-Growth and An Innovation Artificial Neural Network Techniques, Published in IJRASET

SKILLS

C | Java | basic Python | HTML | CSS | MS-Word | MS-Excel | MS-PowerPoint | MYSQL | JavaScript

PERSONAL PROJECTS

College Website (07/2019 - 08/2019)

- Using HTML & CSS

Iris Flower classification & Prediction (08/2020 - 09/2020)

- Using Machine Learning

ANN Approach for Association Rule Mining in Transactional Dataset (09/2021 - 05/2022)

Using Association Rule Mining & Data Mining Technique

ATTENDED SEMINAR

Artificial Intelligence Workshop at IIT BHU, Virtual & Augmented Reality Workshop at ITM (2019)

5G Technology 2days Seminar at ITM Gida by Airtel Company Expert (2019)

CERTIFICATES

1. Certificate of 'Indian Culture Knowledge Test'at SVM
2. Certificate of participated in Jagran Sanskarshala Programme 2014-2015
3. Certificate of 'Bharat scout and guide' at Gorakhpur (08/2013)
4. Certificate of 'Respect for meritorious students' by world Hindu Federation

LANGUAGES

HINDI
Full Professional Proficiency

ENGLISH
Full Professional Proficiency

Satish sahani

Student

Hardworking ,Goal oriented ,determined and focused with a keen eye for details. Sklled in Frontend & Website developement & programming language including *C, Java,Python,PHP & SQL*. **I am always energetic and eager to learn new skills.**

satishsahani951@gmail.com 7268803341 Gorakhpur linkedin.com/in/satish-sahani-b058061b5

EDUCATION

UG
Institute Of Technology & Management Gida
09/2021 - Present, *71.47%*
Courses
- Bachelor Of Technology

Intermediate (10+2)
Alma Mater The School Maniram Gorakhpur
04/2017 - 04/2018, 58.6% Cources

- PCM

High - School
P.N.National Public School Jungle Dhusad Pipraich Road Gorakhpur
04/2015 - 06/2016, 84.4% Courses

- Science

INTERNSHIP AND TRAINING

WEBSITE DESIGN &DEVELOPMENT | Gorakhpur/EASY CAD SOLUTIONS
-Using HTML & CSS
06/2019 - 07/2019,

PYTHON WITH MACHINE LEARNING | Gorakhpur/ edufruit
-Using python with Machine Learning
08/2020 - 09/2020,

COMPLETION COURSES
Eduonix Learning Solutions
08/2021 - 09/2021,

Eduonix Learning Solutions
”Learn PHP & MYSQL Development By Building Projects”

Eduonix Learning Solutions
”Linux For Absolute Beginners”

Eduonix Learning Solutions
”R programming guide for beginners”

Eduonix Learning Solutions
”Learn HTML5 Programming By Building Projects”

SKILLS

C | Basic Java | Basic python | Machin Learning | HTML | CSS | SQL | Javascript | MS-Word | MS-Excel | MS-Power Point | PHP | Wordpress

PERSONAL PROJECTS

Hotel Website (07/2019 - 08/2019)
- Using HTML % CSS

Iris Flower Classification (08/2020 - 09/2020)
- Using python with Machine Learning

EXTRA CARRY-CULAR ACTIVITIES

Winner of 100m &400 m race at Intermediate Level (2017-2018)

Finalist of khokho& kabaddi at Intermediate Level (2017-2018)

Runner-up of Vollyball at Intermediate Level (2017-2018)

Participation in Photography Challenge “PARAMPARA”at College Level (2020)

CERTIFICATES

Artificial Intelligence Workshop at IIT BHU, Virtual & Augmented Reality Workshop at ITM (2019)

5G Technology 2days Seminar at ITM Gida by Airtel Company Expert (2019)

Machine learning with Python AT MMMUT conducted under Sanchaar programme by Wissenaire ”20” The Annual Techno-Management Fest if Indian Institute of Technology ,Bhubaneswar (2019)

LANGUAGES

English
Full Professional Proficiency

Hindi
Native or Bilingual Proficiency

INTERESTS

Surfing Through Internet | Travelling | Yoga | Coding | Participating in Social Activities | Gaming

A Comparative Study on Association Rule Mining and Its Preliminaries

Aditya Shukla[1], Pankaj Kumar Gond[2], Dr. Harvendra Kumar[3]

[1]B.Tech 4th Year, Dept. of Information Technology, ITM Gorakhpur, U.P, India

[2]B.Tech 4th Year, Dept. of Information Technology, ITM Gorakhpur, U.P, India

[3]Associate Professor, Dept. of Computer Science and Engineering, ITM Gorakhpur, U.P, India

harvendra.patel81@gmail.com

Abstract.

The process of extracting and identifying nuggets of findings from outsized amounts of records is known as "data mining (DM)". It consists of several approaches, such as clustering, data summarization, association mining, and classification. Association Rule Mining (ARM), in particular, aims to extract common patterns, interesting connections, associations, or structures that can be adjusted between sets of objects or other statistics. This technique plays a significant part in the route of refining strong rules to demonstrate the stable association between several itemset present in the database. With combined rule mines, different types of techniques and measures have been designed, but it is important to know which way is the best to extract appropriate association rules. Therefore, in this document, we assess the procedures used in an ARM to test the ability to extract high-dimensional data. The paper contains the following sections: section first is the introduction, section second is literature survey, section third is preliminary concepts, section fourth is DM and DM tasks, section fifth is association rule mining, and the last section is the conclusion.

Keywords Data Mining, Frequent Itemset, Data mining Model, Association rule mining

1. INTRODUCTION

DM is the methodology of finding sensible, new patterns related to trends by filtering large amounts of archived data using Pattern Recognition (PR) techniques and mathematical techniques. According to researchers, the two important DM models are predictive and descriptive. The predictive uses a variety of available databases to predict unknown results, while the descriptive focus on finding patterns that define data. Each model for further classification has four functions, as represented in figure 1. Out of these eight functions, the ARM is the most commonly used DM function for studying trends or patterns in a dataset. It also provides rules which help in understanding customers' behavior. Nevertheless, these document sprits around a relative study of different measures of ARM. The measures are support, lift, confidence, and conviction.

This document presents a relative study of the principle and methods used in each measure. The representations provided in this document provide a further understanding of the effectiveness of the measures. Tests are performed on these scales and the results are seen in the provision of time to act and use memory in addition to the patterns produced by it.

2. LITERATURE SURVEY

K. Solanki Surbhi and T. Patel Jalpa [1], write about the trouble faced by individuals while using the frequent pattern mining strategies. This Mining strategy work by examining the database several times, which eventually results in higher process costs. Not only this, this strategy leads to the production of candidate itemsets, which ultimately requires more memory and becomes more sophisticated in handling when the database is outsized. So, to reduce the downside of this problem of candidate $s$et generation, a tree-based strategy came for mining periodic patterns. However, the tree-based strategy produces numerous conditional fp trees. So, in progression to vanquish this problem of producing numerous based fp-trees, fp is DAG for normal pattern mining is enhanced, whereas fp-tree is constructed as a DAG. In addition, to find effectual mining, fuzzy deviation could be applied to the assayable database to provide the optimal patterns.

P. Amaranatha Reddy and MHM Krishna Prasad [2], on how to obtain ARs for the differing types of data-objects and requisitions involved in them. Some of the ways of data sets reasoned in this document are Boolean, Quantitative, weighted, time-series, stream data, infrequent, Diversified, and fuzzy item-sets. So the idea of differentiating ARs amongst correlated items like milk and banana will not be considered enterprise comprehension but differentiating unknown ARs between distinct items such as liquor and diapers will be beneficent in enterprises expansion and finding such an unknown category of ARs requires in-depth information about the data.

Mrs. Geeta S. Navale and Drs. Suresh N. Mali [3] proposed a variety of methods to hide association rules on the database and to develop support and confidence measures. The intention of this paper is to conceal the critical association rules of the DM on the following conditions: no production of false rules, no loss of information, Modification Degree and robustness against intentional or unintentional attacks. Except for the three conditions above, the proposed method will be compatible with a measurable database. Therefore, the proposed method of data encryption of the operating system will be evaluated and verified with regard to the various parameters as set out in the conditions and it is necessary to assess the satisfaction of the above-mentioned conditions.

L Greeshma and Dr. G Pradeepini [4] focused on developing the latest Apriori-*b*ased algorithm, which satisfies positive aspects of constrained *i*temset based mining such as anti-*m*onotonicity. The problem of ARM is to retrieve relevant itemsets for which it presents a new constraint, called relation-based constraints, applicable to relevant data. In the CIM algorithm, it helps us to recognize the main components of a candidates' key itemsets and generate frequent itemsets, which satisfy the anti-*m*onotonicity properties, which means small coverage and cardinal size limited to a particular dataset.

Table 1: A comparative study of algorithms used in DM:

S. NO.	Algorithm Name	Application	Advantages	Disadvantages	Year
1.	AIS	Not frequently used, but when used is used for small problems.	1. Better than SETM. 2. Easy to use	1. Candidate sets generated on the fly. 2. Size of candidate set large.	1994
2.	SETM	Not frequently used.	1. Separates generation from counting.	1. Very large execution time and the size of candidate set is large.	1994
3.	Apriori	Best for closed item sets.	1. Less candidate sets. 2. Generates candidate sets from only those items that were found large.	1. Takes a lot of memory.	2003
4.	AprioriTID	Used for smaller problems.	1. Better than SETM, Apriori for small databases. 2. Fast & Time saving	--	2013
5.	Apriori Hybrid	Used where Apriori and AprioriTID used.	1. Better than both Apriori and AprioriTID.	--	2013
6.	Eclat	Best used for free item sets.	1. Less memory usage. 2. Lower minimum support.	1. Apriori wins in cases where candidate sets are more.	2004
7.	Recursive Elimination	Effectively select the most relevant itemset.	1. Better than Apriori in all cases.	1. Less than éclat in all cases.	2005
8.	FP Growth	Used in cases of large problems as it doesn't require generation of candidate sets.	1. Only 2 passes requires. 2. No candidate set generation required.	1. Using tree structure creates complexity.	2003

P. Naresh and Dr.R. Suguna [5] explain a relative study of four ARM algorithms, namely $Apriori$, $FPGrowth$, LCM and FIN. These algorithms are discovered in terms of their purpose, the way they instigate recurring itemsets, and their appearance on the organization's data. The performances of all these algorithms are evaluated by means of time. The algorithms presented in the DM configuration, namely "Sequential Pattern Mining Framework" are used to make an investigational assessment of the algorithms. These investigational results on these algorithms disclosed that the amount of data has a great significance on the implementation time and at last these led to the conclusion that: the FIN algorithm displays the least implementation time while $FPGrowth$ shows minimum memory utilization.

3. PRELIMINARIES

1) Frequent Set: T is a transactional database and σ is the "minimum support threshold" specified by the user or domain experts. An itemset is frequent if they satisfy the min support threshold,

$$s(M)_{T \geq \sigma}$$

2) Maximal Frequent Set: In order to be a "$maximal\ frequent\ set$", a frequent set must be recurring and no superordinate of it must be recurring.

3) Support(s): In a database D, s is the proportions of agreements that comprise both M and N itemsets. In s of an ARs $M \longrightarrow N$, there is,

$$supp(M \longrightarrow N) = supp(M \cup N) = P(M \cup N)$$

4) Confidence(c): In database D, this is the percentage of agreements that contain itemsets M and N. 'c' is calculated by considering the conditional probability as well as the itemset support. Confidence can be calculated using the equation,

$$conf\ (M \longrightarrow N) = P(N/M) = supp\ (M\ U\ N)\ /\ supp\ (M)$$

Here, $supp(M \cup N)$ indicates the number of agreements considering both sets of items M and N, and $supp(M)$ indicates the number of agreements considering just set M.

5) Lift: It is used to analyze the frequency M and N together, if both are precisely in different. The lift of rule $M \longrightarrow N$ is defined as,

$$lift\ (M \longrightarrow N) = \frac{confidence}{expected\ confidence} = \frac{conf(M \longrightarrow N)}{supp(N)}$$

6) Conviction: Conviction analyses the implication stability of the rule from statistical independent. Conviction is defined as,

$$conv\ (M \longrightarrow N) = \frac{1 - supp(N)}{1 - conf(M \longrightarrow N)} = \frac{P(M) * P(\overline{N})}{P(M \cup \overline{N})}$$

Where $P(\overline{N})$ is the probability that N doesn't appear in an agreement.
It compares the probability that M appears without N if they were dependent on the actual frequency of the appearance of M without N [6, 7].

4. DM AND DM TASKS

DM can often be categorized into two categories based on what a particular project is trying to accomplish. Those two categories are descriptive model and predictive model. There are a number of DM functions, such as ARM, Time Series Analysis, prediction, Neural Network, etc. Either of these functions falls under the predictive model or descriptive model. The DM system can perform one or more of the above functions as part of the DM.

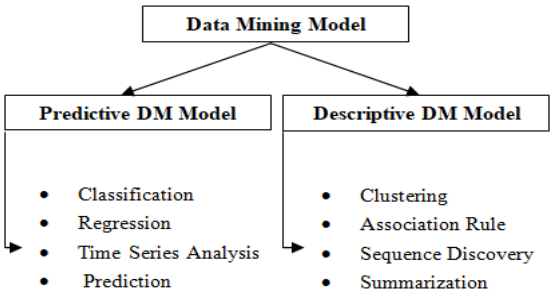

Fig 1: DM Model and its Tasks

4.1 Descriptive Model: The descriptive model defines a domain that represents a way that can be described or find the relationship among data. It can be used for many purposes. It may include ethical, structural, and other definitions that establish reasonable relationships about the system, such as its component relationships, the interactions between its components, and the distribution of its ethical properties to structural elements. Descriptive models are not usually constructed in a way that directly supports imitation, animation or performance, but can be considered compliance with grammatical rules, and logical relationships can be considered for them. This model basically relies upon an unsupervised learning approach. Some of the major tasks of descriptive models are as follows: Association Rule, Clustering, Sequence Discovery, and Summarization.

4.2 Predictive Model: The Predictive-model is a mathematical method that is generally used by DM technologies to define $m$odeling prediction as to the process of predicting the subsequent time behavior of an analysis of factual and recent data. A predictive model makes theories based on what has previously occurred and what is currently occurring. If new data shows that the current situation has changed, the likelihood of the outcome must be recalculated. This model relies upon the supervised learning approach. Some major tasks of predictive models are regression, classification, prediction, and Time Series Analysis.

5. ASSOCIATION RULE MINING

ARM [3] is an event for determining organizations, patterns, and relationships between sets of objects on dataset. The law of association is form $M \longrightarrow N$ [support, confidence]. The support and confidence are two measures that are used for assurance of the rule. AR is said to be strong if it satisfies both the minsupp(minimal support) and the minconf(minimal confidence) that is defined by the user. These ARs are facile to establish due to the $s$mall database but become more complex as the database transforms. Some general values and concepts are required in order to better understand DM in large data sets. A set contains n items and is called an n-itemset. So set {A, B} is a set of 2- itemset. Based on the frequency of itemsets, the number of active functions is calculated. So now, let us assume the value of minsupp=0.25 and minconf=0.10.

The dilemma of mining ARs can be fragmented into two sub-dilemmas:

1) Find all sets of itemsets whose support is greater thanσ. These itemsets are known as $frequent\ itemsets$.
2) Use these $frequent\ itemsets$ to develop the desired rules. The conventional ideology is that if, say p, q, r, s are $frequent\ itemsets$, then we can decide on the rule $pq \rightarrow rs$ that holds by checking the following inequality

$$\frac{s(\{p, q, r, s\})}{n} \geq \sigma$$

Where n is the total number of transactions and σ is a minsupp.

Measures for Association Rule Mining

Table2 presents Transactional Super Market Data

TID	Items
T_1	A_1, A_2, A_3
T_2	A_1, A_2, A_4
T_3	A_1, A_3, A_5
T_4	A_2, A_4, A_6
T_5	A_3, A_6, A_7
T_6	A_1, A_2, A_3, A_4
T_7	A_2, A_3, A_4
T_8	A_3, A_7
T_9	A_2, A_3, A_4, A_8
T_{10}	A_6, A_7

Here: Ten transactions and eight items in a transactional dataset are shown in table 1.

5.1 $\boldsymbol{Support(s)}$: $Support$ is measured as the amount of logs that contain $M \cup N$ from the entire logs in the database. The proportion for each item is altercation by one, when so ever the item is crossover in dissimilar transaction in a database across the period of scanning.

$Support$ is used to measure the quantity or frequency of an itemset in a database. This measure gives an idea of how common itemset is in all activities. Support can be measured as:

$$Supp(M \longrightarrow N) = P(MN) = Freq\frac{(M, N)}{n}$$

Where: Freq (M, N) = Transaction containing M and N, and n = Total number of transitions.

It helps us to identify the rules that need to be considered for further analysis or not. If support of the rule is greater than minsupp, then find the confidence of the rule.

$$Support\ (A_2, A_4 \longrightarrow A_3) = 3/10 = 0.3$$

Thus the support value of items A_2, A_4 & A_3 is 0.3.

5.2 **Confidence(c):** Confidence is defined as the [9] amount of the numbers of transactions that contains $M \cup N$ to the entire logs that contain M, where, if the fraction exceeds the kick-off of $confidence$, an ARs $M \longrightarrow N$ can be obtained. Confidence is an indication of how often the rule is true.

Confidence ($M \longrightarrow N$) in relation to the set of functions n, the part of the functions containing M and N.

$$conf\ (M \longrightarrow N) = \frac{supp(M\ U\ N)}{supp\ (M)}$$

Confidence explains how N is frequently occur when already buying M. This describes the link between two things. For example, if a person buys jam there is a good chance to buy bread. It is calculated as part of the number of operations where both M and N occur to support the M object.

$$Supp(A_2, A_4 \longrightarrow A_3) = 3/10 = 0.3$$
$$Supp(A_2, A_4) = 5/10 = 0.5$$
$$conf\ (A_2, A_4 \longrightarrow A_3) = 0.3/0.5 = 0.6$$

An association rule $M \longrightarrow N$ will be strong if, $conf(M \longrightarrow N) \geq minconf)$ and here $conf(A_2,\ A_4 \longrightarrow A_3)$ is 0.6 which is greater than the minconf, therefore the rule can be reasoned as a strong rule because it met with minsupp and minconf conditions. But we need to check it further than support and confidence alone cannot be sufficient to find a strong rule. In the above association rule$(A_2,\ A_4 \longrightarrow A_3)$, support of the consequent $(s(A_3) = \frac{7}{10} = 0.7)$ is greater than the confidence of the rule (0.6). This is not feasible. Therefore, this may be a misleading rule. Misleading rules can be generated from irrelevant datasets. Therefore, additional steps are needed to avoid misleading rules.

So to solve this problem of misleading rules another two measures can be used i.e. lift and conviction.

5.3 **Lift:** The Lift [10] is defined as one of the measures of ARM which define how far the inter-dependence in between M and N. The measure lift is not sensitive to rule i.e. $(lift(M \longrightarrow N) = lift(N \longrightarrow M))$. A Lift could be formulated as:

$$lift\ (\mathrm{M} \longrightarrow \mathrm{N}) = \frac{conf(M \longrightarrow N)}{supp(N)} = \frac{supp(MUN)}{\mathrm{supp}(M) * supp(N)}$$

So,

$$lift(A_2, A_4 \longrightarrow A_3) = \frac{0.3}{0.5 * 0.7} = 0.35 < 1$$

The fraction of the observance $support$ and the predicted $support$ if M and N are free for each other is known as lift. It has three possible values:

- If Lift = 1, the probability of occurrence and outcome are independent of each other.
- If Lift >1, then the itemsets are dependent on each other.
- If Lift < l, tells us that one thing replaces other things, which means one thing has a negative effect on another thing.

An ARs $M \longrightarrow N$ is engaging if it is $strong$ and $lift(M \longrightarrow N) > 1$.

Here as $lift(M \longrightarrow N) < 1$, thus the rule is not valid to consider it as a strong rule.

5.4 **Conviction:** The conviction is defined as one of the measures of ARM which undertake to analyze the magnitude of execution of the rule, to evaluate the conviction. [8, 11].

Unlike Lift, Conviction is tactful to rule direction i.e. $(conv\ (M \longrightarrow N) \neq conv\ (M \longrightarrow N))$.

A Large conviction value shows that the obtained result is largely relying on the predecessor. The conviction can be formulated as:

$$conv(M \longrightarrow Y) = \frac{1-supp(N)}{1-conf(M \longrightarrow N)} = (1 - supp\ (N))/(1 - conf(M \longrightarrow N))$$

It correlates the likeliness that M exists without N when they are relying on the factual frequentness of the existence of M without N. In that scenario, it is similar to $lift$. However, $conviction$ have monotonousness in $confidence$ and $lift$.

So,

$$conv(A_2, A_4 \longrightarrow A_3) = \frac{1 - 0.7}{1 - 0.6} = 0.75$$

$$conv(A_3 \longrightarrow A_2, A_4) = \frac{1 - 0.5}{1 - 0.42} = 0.86$$

Here, as the value of $conv(A_2, A_4 \longrightarrow A_3)$ has a value less than the value of $conv(A_3 \longrightarrow A_2,\ A_4)$, therefore the $conv(A_3 \longrightarrow A_2, A_4)$ can be considered as a strong rule.

Table 3: Quality measures and range of feasible values:

Name	Equation	Feasible values
Support	P_{MN}	[0,1]
Confidence	$\frac{P_{MN}}{P_M}$	[0,1]
Lift	$\frac{P_{MN}}{P_M * P_N}$	[0,1]
Conviction	$\frac{P_M * P_{\bar{N}}}{P_{M\bar{N}}}$	$\left[\frac{1}{n}, \frac{n}{4}\right]$

6. CONCLUSION

In this paper, there is a preliminary of the DM and a detailed discussion on ARM. ARM is facing the problem of finding the most efficient and strong rules which are suitable for any dataset due to the presence of numerous rules. In most literature, the fascinating steps of governance in ARM algorithms are based on $support$ and $confidence$. Depending on the types of application, different measures could be used to compute the interestingrules. As described in the above sections of the paper, the main issues are to trust the method for support is low forecasting capability and similar support issues. Whereas, the previous work retains presuming solutions to these issues, as an add-on the rate of promotion or sentencing and the use of an unusual support barrier, where it remains without guidance in defining support. Without such an order, users may set the wrong support limit and suffer from combinatorial explosion or loss of new cognitive arrangement. Our plan for constructing the solution for this issue is to exclude clients from determining a support limit. Therefore, we need to refine the rules in sort to obtain strong rules that must satisfy the following parameters.
Association rules will be strong if they satisfy the following conditions: (i) $supp(M \longrightarrow N) \geq minsupp,$ (ii) $conf(M \longrightarrow N) \geq minconf$ (iii) $lift(M \longrightarrow N) > 1$ and higher $conv(M \longrightarrow N)$ value.

7. REFERENCES

1. Surbhi K. Solanki and Jalpa T. Patel, “A Survey on Association Rule Mining”, Fifth International Conference on Advanced Computing & Communication Technologies, Volume 5, pp. 212-216, 2015.
2. P. Amaranatha Reddy and MHM Krishna Prasad, “Challenges to find Association Rules over various types of data items: a Survey”, International Conference on Computing, Communication and Automation, pp. 180-184, 2017.
3. Mrs. Geeta S. Navale and Drs. Suresh N. Mali, “A Survey on Sensitive Association Rules Hiding Methods”, Third International Conference on Computing, Communication, Control And Automation (ICCUBEA), IEEE, 2017.
4. L Greeshma and Dr. G Pradeepini, “Unique Constraint Frequent Item Set Mining”, 6th International Conference on Advanced Computing, IEEE, pp. 68-72, 2016.
5. P.Naresh and Dr.R.Suguna, "Association Rule Mining Algorithms on Large and Small Datasets: A Comparative Study", Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 587-592, 2019.
6. Hemant Kumar Soni, “Multi-objective Association Rule Mining using Evolutionary Algorithm”, IJARCSSE, Volume 7, Issue 5, May 2017.
7. H. K. Soni et al., “Frequent Pattern Generation Algorithms for Association Rule Mining: Strength and Challenges”, IEEE International Conference on Electrical, Electronics and Optimization Techniques (ICEEOT), pp. 3744-3747, 2016.
8. Dinesh J. Prajapati et al., “Interesting association rule mining with consistent and inconsistent rule detection from big sales data in distributed environment”, Future Computing and Informatics Journal 2, pp. 19-30, 2017.
9. J. M. Luna et al., “Optimization of quality measures in association rule mining: an empirical study”, International Journal of Computational Intelligence Systems, Volume 12, pp. 59-78, 2018.
10. H. K. Soni et al., “Association Rule Mining: A data profiling and prospective approach”, International Journal of Current Engineering and Scientific Research, Volume 3, pp. 57-60, 2016.
11. Memoona Khanum and Tahira Mahboob, “A Survey on Unsupervised Machine Learning Algorithms for automation, Classification and Maintenance”, International Journal of Computer Applications, volume 119-No.13, June 2015.
12. Sikha Bagui and Probal Chandra Dhar, "Positive and negative association rule mining in Hadoop’s MapReduce environment", Journal of Big Data, pp.1-16, 2019.
13. Lichun Li et al., “Privacy-Preserving-Outsourced Association Rule Mining on Vertically Partitioned Databases”, IEEE Transactions on Information Forensics AND Security, pp. 1-15, 2016.
14. Mandeep Mittala et al., “Loss profit estimation using association rule mining with clustering”, Management Science Letters, pp. 167–174, 2015.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 1118-1123 www.ijaem.net ISSN: 2395-5252

An In-Depth Study of the Different Data Mining and Machine Learning Techniques

Satish Sahani[1], Neha Gond[2], Dr. Harvendra Kumar[3]
[1]B.Tech 4th Year, Dept. of Information Technology and Engineering, ITM Gorakhpur, U.P, India
[2]B.Tech 4th Year, Dept. of Information Technology and Engineering, ITM Gorakhpur, U.P, India
[3]Associate Professor, Dept. of Computer Science and Engineering, ITM Gorakhpur, U.P, India

Submitted: 10-05-2022 Revised: 17-05-2022 Accepted: 20-05-2022

ABSTRACT: Data mining is the process of extracting meaningful patterns and information from large amounts of data. It is a fast expanding research topic, whose appeal and demand have risen in tandem with the advent of tools that may assist in revealing and comprehending hidden information in massive amounts of data. Several entities, such as banks, insurance firms, and grocery stores, produce such data on a daily basis. This paper describes the various DM approaches in depth, with illustrations, and compares them in terms of performance parameters such as execution time and efficiency. This paper covers the following sections: Introduction, Data Mining, Data Mining Model, Performance Measure and Conclusion.

Keyword - Data Mining, Data mining Model, Performance Measure, Association Rule Mining.

I. INTRODUCTION

The growing interest in DM and the use of historical data to find irregularities that refine future decisions follows the convergence of some recent trends, like the falling amount of large facts storage gadgets and the growing ease of assembling data over networks. The enlargement of robust and well-organised (ML) algorithms to operate this data and the falling cost of calculation power are enabling the utilisation of computationally exhaustive and intensive data analysis techniques. Data mining is the practise of identifying well-organized, new, potentially helpful, and intelligible patterns in data. With the worldwide utilisation of databases and the explosive widening of their dimensions, organisations are looking towards the difficulty of facts overload. The issue of effectively using these huge volumes of data is becoming a major problem for all enterprises. As a result, while some techniques such as clustering, decision trees, and so on exist in many fields, they are not well suited to DM for association rule mining. As a result, in order to use statistical and mathematical tools, we must adapt these techniques so that they can efficiently sift through enormous amounts of data stored in secondary memory.

II. MACHINE LEARNING

As stated by Arthur Samuel, machine learning is defined as the research-based study of algorithms and analytical models that provide computers with the capability to learn in the absence of explicitly programmed instructions. It's used to trained machine how to manage the data even more efficiently, e.g., a concept representing the outcome of learning as output. Inductive learning, where the arrangement infers grasp itself by observing its environment, has three leading strategies: supervised learning, unsupervised learning, and reinforcement learning. Inductive learning methods can be used to forecast the results of future situations; in other words, not only for states encountered but rather for unseen states that could occur. There can be many possible models based on having handsets as examples. In such situations, the principle of Occam's razor is very appropriate. It states that if there are many explanations for a particular phenomenon, it makes sense to choose the simplest one because it is more likely to capture the nature of the phenomenon.

1.1. Supervised Learning

Supervised learning means learning from examples, where an instruction set is given that acts as an example for the classes. The system then finds a description of each category. Once the description (and hence the classification rule) has been formulated, it's used to forecast the class of previously unseen objects. This is analogous to the distinction between analysis and interpretation in statistics. It uses levelled data. Classification is a

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 1118-1123 www.ijaem.net ISSN: 2395-5252

powerful learning technique. It is a way of categorising a piece of information for already defined groups. The classification method provides mathematical algorithms such as linear programming, decision trees, ANNs, and statistics. For example, classification is applied in an application that makes up for all records of company members who have left or who will probably plan to leave the company in the future. Here, employees may be classified as on leave or staying. Or a bank loan manager analyses his/her customer's data to determine which loan applicants are either safe/risky.

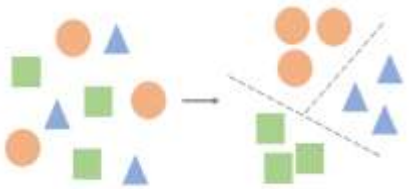

Fig -1: Overview of Supervised Learning
(Input examples are categorized into a known set of classes)

1.2. Unsupervised Learning

Unsupervised learning is study from observation and discovery. In this mode of learning, there is no training set or prior knowledge of the classes. The system examines the given set of facts to observe similarities emerging out of the subsets of the facts. The outcome is the acquisition of class descriptions, one for each class, discovered in the environment. Clustering is an example of this technique. Clustering is a technique for grouping similar types of data. Cluster analysis refers to clusters that are similar, but other clusters have different objects. The cluster technique and the classification technique are almost the same kinds of techniques. The classification technique uses predefined classes, while clustering does not use predefined classes. Groups or classes are obtained in clustering by using similarity between records according to the characteristics of real data. The clustering method describes groups and objects in a dataset, while the classification technique is used in predefined classes to assign objects. A book in a library is an example of clustering techniques. The algorithms used in DM and the segment of ML dealing with learning from examples overlap, as does the problem addressed. DM is concerned with discovering understandable knowledge, while ML is concerned with improving the performance of an intelligent system or agent for problem-solving tasks.

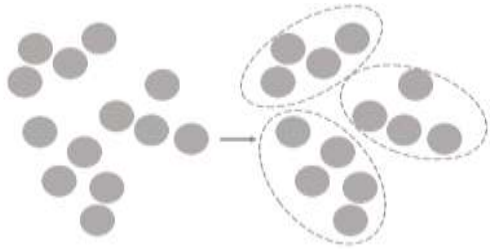

Fig -2: Overview of Unsupervised Learning
(Input samples are grouped into clusters based on the underlying patterns)

1.3. Reinforcement Learning

Reinforcement learning is applied when the task at hand is to make a sequence of decisions towards a final reward. During the learning process, an artificial agent gets either rewards or penalties for the actions it performs. Its target is to enhance the total reward. Examples include learning a negotiator to play computer games or perform robotics tasks with the end goal in mind.

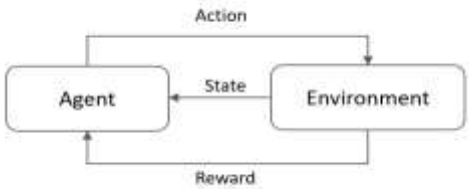

Fig -3: Overview of reinforcement learning
(An agent observes the environment state and performs actions to maximize an overall reward)

III. DATA MINING

DM is the procedure of extracting useful information and patterns from vast data. Research in databases and information technology has given rise to an approach to storing and manipulating this precious data for further decision-making. This is an important step in obtaining information from a website that produces useful patterns or models. The goal of this process is to find previously unknown patterns. DM uses statistical algorithms to separate data and evaluate the likelihood of upcoming events. DM is also known as KDD. Example: Mobile service suppliers use DM to outline their marketing campaigns and keep customers from migrating to different vendors. From huge amounts of facts such as payment information, email, text messages, web data shipping, and customer service, DM tools can predict 'churn' by telling customers they want to switch vendors. An opportunity result is provided by these outcomes. Mobile service providers are now free to dispense compensation, and specials to

customers at high risk of withdrawal. This type of mining is frequently used by major service suppliers such as broadband, telephone, gas suppliers, etc. DM can be categorised into two dimensions: predictive DM model and descriptive DM model.

3.1 Predictive Model

Predictive-modeling is a mathematical method which aims to forecast future incidents based on past behavior. There are a number of DM functions, such as ARM, Time Series Analysis, Prediction, etc. It works by examining facts to recognise patterns and then using those patterns to create algorithms that data scientists train. It is the fundamental function of predictive analytics applications and is broadly used in a variety of applications, from meteorology to fraud detection, to security management, etc. In short, predictive fashion is a statistical technique using ML and DM to predict and foresee likely future outcomes with the aid of factual and existing data. It acts by analysing current and factual data and projecting what it learns onto a model generated to forecast likely outcomes. This data mining model is broken into four subcategories, namely

Classification:- Classification is a kind of SL that aims to accurately forecast the target category in each case of data. It is the way of DM that assigns objects to a group into specific groups or classes. The principle of separation is to exactly predict the target class in each case of data. It's used to separate each item by a set of data into a predefined set of classes or groups. The division of a data analysis function is where the model or separator is designed for predicting category labels. It allows you to arrange data sets of all sorts, including rigid and huge datasets as well as small and easy ones. The primary aim of classification is to connect a variable of interest with the required variables. For example, a bank loan manager identifies whether the loan applicant is risky or safe.

Regression:- Regression is an approach that is used to predict a range of numeric values (also called continuous values) given in a specific dataset. It refers to a kind of SL that is used to predict some continuous-valued attribute. Regression helps some business institutions analyse the desired variable and predictor variable relations. It is a significant device to analyse the facts that can be used for commercial forecasting and time series modeling. Regression involves the method of fitting a straight line or a curve to numerous data points. It happens in such a way that the distance between the data points and cures comes out to be small. The most popular kinds of regression are linear and logistic regressions. Other than that, many other kinds of regression can be performed depending on their performance on a separate data set.

Time Series Analysis:- In simple words, "a time series is nothing but a sequence of many data points that occur in a successive order for a given period." It represents a series of time-based orders that would be in years, months, weeks, days, hours, minutes, and seconds. It is an observation from the sequence of discrete-time successive intervals. It typically requires a huge number of data points to ensure consistency and reliability. It is a particular way of analysing a series of data points collected over an interval of time. In this analysis, analysts record data points at compatible intervals over a set period rather than recording the data points periodically or randomly.

Prediction:- Prediction is used to observe a numerical result same as in classification, the training dataset contains the inputs and corresponding numerical output values. The algorithm derives the model or a predictor according to the training dataset. The model should observe a numerical output when the new data is given. Unlike in classification, this method does not have a class label. The model predicts a continuous-valued function or sequence value. Predicting the value of a house depending on facts such as the number of rooms, the total area, etc., is an example of prediction.

3.2 Descriptive Model

Non-trivial datasets are studied in the descriptive DM model. It provides information to understand what is going on inside the data. In general, this is the DM model that identifies the patterns or/and relationships in the data. This model uses UL techniques. An example of a descriptive DM model is the retailer. They try to recognise the relationship between the purchased items. On further classification, the predictive model breaks into four subcategories, namely:

Clustering:- Clustering is the way of converting a group of abstract objects into classes of identical objects. It is the procedure of splitting a set of data or objects into a set of significant subclasses called clusters. A subset of objects such that the gap between any of the two objects in the cluster is less than the gap between any object in the cluster and any object that is not located inside it. It helps users to understand the systematic or natural grouping in

a data set's aims and to add it either as a stand-alone instrument to get a better perception of data distribution or as a pre-processing step for further algorithms.

Association Rule:- The Association Rule is used to identify frequently used items in a vast data set. It associates the existence of a set of items with another price range in another set of variables. This rule strives to obtain patterns in data based on interpersonal relationships in the same practice. This kind of strategy helps businesses make certain decisions, such as catalogue design, advertising, and customer purchasing behaviour analysis. These are used to discover correlations and co-occurrences between data sets. They are ideally used to describe patterns in data from seemingly unconventional information repositories, such as relational databases and transactional databases. The act of using association rules is sometimes referred to as "association rule mining".

Sequence Discovery:- Sequence discovery, or successive pattern mining, is a DM method that discovers statistically applicable patterns in sequential data. This mining programme assesses definite criteria, such as happening frequency, duration, or values in a set of sequences to find interesting hidden patterns and uncover relationships among data. This technique defines a sequential pattern of events and actions. For example, a customer who buys more than twice in the first phase of the year may be likely to buy at least once during the second quarter.

Summarization:- Summarization is a kinds of UL technique that focuses on the compression of data. It is obtained by identifying attributes such as customer name, contact, etc. with many different values. and by removing them or performing a threatening function. Also, we can use standard statistics on the data to represent its summary. This is the key DM concept that involves techniques for finding a compact description of a dataset. Straightforward summarization techniques such as tabulating the mean and standard deviations are often applied for data analysis, data visualization, and automated report generation.

IV. LITERATURE SURVEY

Ting Witten and Yosef Hasan Jbara [1] write about ways to improve efficiency and scalability by executing several learning processes and combining the collective results. The most well-known learning algorithm used in estimating the values of the weights of a neural network —is back propagation, which stores more information during training and then uses this elaborate information to generate better predictions about the test data. Freund and Schapire and Yu and Liu [2], focused on refining and overcoming the classification correctness by scaling down the required classification time. Furthermore, choosing a more suitable distance metric for the specific dataset can improve the correctness of instance-based classifiers. One other choice in designing a training set reduction algorithm is to change the instances using a new representation such as prototypes. They do this independently of all the other classes in the training set. For this reason, for small datasets, it may be better to use a divide-and-conquer algorithm that considers the entire set at once. Villada Drissi and Breiman [3], give the important observation about the issue faced by users while using the frequent patterns of DM. The combining of clustering and classification is known as the prediction analysis clustering algorithm groups the data according to their similarity and the classification algorithm assigns the class to the data. Therefore, the observed problem of data security in the OS will be evaluated and verified against the many parameters set out in the conditions, and it is necessary to access the satisfaction of the above-mentioned conditions. Hassan M. Najadatet al. [4], focused on developing the latest apriori-based algorithm, which satisfies positive aspects of constraints itemsets based mining as anti-monotonicity. The trouble with ARM is that it tries to retrieve matching itemsets for which they are present and for which they represent new constraints, known as relation-based constraints, applicable to the relevant data. In the CIM algorithm, this helps us to recognise the main components of a candidate's key itemsets and generate frequent itemsets, which satisfies anti-monotonicity properties, which means a little coverage cardinal size restricted to a specific dataset.

R. Srikant et al. [5] proposed the reason for selecting FP-Growth: FP-Growth is a popular, convenient, and efficient enough algorithm for many applications and has replaced LCM, which is faster and uses more memory. It is important to promote SSD quality with an advanced mining algorithm. Significant contributions result in novel results. The support methodology is an efficient and convenient procedure to promote SSD quality. This represents an important contribution for many researchers and companies in vast data analysis. This study is also the first trial with a novel application in terms of using the association rule for SSD quality from an industrial perspective,

thereby providing the rationale for its novelty. The study output justifies the capability of the association rule for DM implicational areas. G Nandi and A Das [14] introduced the reversed ARM as one of the most dominant ways to extract confidential information. Aprion was considered the first algorithm for mining organizations. This algorithm is based on complete search and dividing the ARM problem into two sub-problems. The first one is to find all available common sets of data, and the second one is outputs organizational rules from the step of pre-acquired item sets. Then, with the growth of FP, another mining algorithm often overcomes the obstacles of Apriori. Rupam Some [15] propose a variety of ways to hide association rules on the database and to develop support, confidence, and measures. Every method has its own advantages and disadvantages. Its advantage is that it does not produce false rules. The proposed techniques will be compatible with a measurable database. Therefore, the proposed methods of data encryption for the operating system will be evaluated and verified with respect to the many parameters as set out in the conditions.

Table -1 Total Study of Algorithm: [8]

S. No.	Algorithm Name	Application	Disadvantages	Disadvantages	Year
1.	LCM	For enumerating frequent closed item.	It features stable performance for both computation time and memory usage.	LCM Algorithm requires frequent itemsets.	2004
2.	Pascal	To identify the frequent itemsets which are generators in a transaction database.	It is utilized pruning strategy	It takes a lot of memory and space.	2003
3.	Éclat	Find frequent items.	Scan the whole database only once. Less parameter required.	Required more memory space for intersecting long transactional sets.	2004
4.	AIS	Not frequently used, but when use is used for small problems.	Better than SETM. Easy to use.	Candidate sets generated on the fly. Size of candidate set large.	1994
5.	SETM	Not frequently used.	Separates generation from counting.	Very large execution time and the size of candidate set is large.	1994

V. CONCLUSIONS

This paper discussed the various learning techniques and gave a brief overview of various kinds of data mining, and their models. DM is a concept of drawing knowledge or patterns from large amounts of data, and DM can be categorised into two categories based on what a particular project is trying to accomplish. Those two methods are called "descriptive models" and "predictive models." Depending on the type of application, different models could be used to calculate the interesting rules. The study shows that in sequence to achieve the efficiency in DM. It is important to choose an almost acceptable algorithm for the application. The study also showed that more efficient and refined mining methods needed to be developed.

REFERENCES

[1]. Ting Witten and Yosef Hasan Jbara, “Predictingpatterns Using Data Mining Techniques”, Washington DC 20052, vol. 6, 2016.

[2]. Freund et al., “Learning technique: association rule,” Middle-East Journal of scientific research, vol. 5, 2019.

[3]. VilladaDrissi andBreiman, “Learning technique:Mining Association Rules in Student’s Assessment Data”, IJCSI International Journal of Computer Science Issues, vol. 9, Issue 5, 2012.

[4]. Hassan M. Najadat et al., "An Improved Apriori Algorithm for Association Rules", International Research Journal of Computer Science and Application, vol. 1, 2019.
[5]. R. Srikant et al.,"Algorithms for Mining Association Rules", 20th International Conference on Very Large Data Bases (VLDB), pp. 487–499, 2018.
[6]. Mrs. Bharati and M. Ramageri, "Data Mining Techniques and Applications", Indian Journal of Computer Science and Engineering, vol.1, pp. 301-305, 2010.
[7]. ShitalBhojani and Nirav Bhatt, "Data Mining Techniques and Trends", Indian Journal of Computer Science and Engineering, vol.5, 2016.
[8]. Anshu at al., "Review Paper on Data Mining Techniques and Applications", International Journal of Innovative Research in Computer Science & Technology (IJIRCST), vo.7, 2019.
[9]. G.Kesavaraj and Dr.S.Sukumaran, "A Study on Classification Techniques in Data Mining", Fourth International Conference on Computer Communication and Networking Technologies, 2013.
[10]. Kabra. R, Bichkar. R, "Performance Prediction of Engineering Student using Decision Tree", International Journal of Computer Application December 2011.
[11]. Han. J et al., "Data Mining Concepts and Techniques", Third Edition the Morgan Kaufmann Series in Data Management Systems Morgan Kaufmann Publishers, 2011.
[12]. Kwame Boakye Agyapong et al., "An Overview of Data Mining Models (Descriptive and Predictive)", International Journal of Software & Hardware Research in Engineering, vol. 4, 2016.
[13]. B. Nigam et al., "The Comparative Study of Top 10 Algorithm for Association Rule Mining", International Journal of Computer Sciences and Engineering, vol. 5, 2017.
[14]. G Nandi, A Das, "A Survey of Data Mining Techniques ", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 6, No 2, November 2013.
[15]. Rupam Some, "Mining Technique and its Future Trends", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 6, June 2013.

Association Rule Mining using FP-Growth and An Innovative Artificial Neural Network Techniques

Pankaj Kumar Gond[1], Aditya Shukla[2], Satish Sahani[3], Neha Gond[4], Dr. Harvendra Kumar[5]

[1]*B.Tech 4th Year, Dept. of Information Technology, ITM Gorakhpur, U.P, India*

[2]*B.Tech 4th Year, Dept. of Information Technology, ITM Gorakhpur, U.P, India*

[3]*B.Tech 4th Year, Dept. of Information Technology, ITM Gorakhpur, U.P, India*

[4]*B.Tech 4th Year, Dept. of Information Technology, ITM Gorakhpur, U.P, India*

[5]*Associate Professor, Dept. of Computer Science and Engineering, ITM Gorakhpur, U.P, India*

Abstract— With the prodigious proliferation of ginormous-scale data depots, a need for incorporating the empirical techniques of Data Mining (DM) with the effectiveness of commensurate systems to intuitively manage cosmic volumes of data has now risen. To quell these obstructions to managing data efficiently, in this document we present a new algorithm based on ANN for DM activities, which overcomes the problems in the currently available algorithms of mining apropos of their execution time and interestingness and to prove its efficiency the new algorithm will be compared to a popular mining algorithm: FP Growth.
The following sections of the paper are sorted as – section I exhibits the introduction, section II exhibits the previously done related works, section III is exemplifying the methodologies, section IV exhibits the experimental setup of the analogous work, section V inculcates the experimental result and finally section VI depicts the conclusion of the respective work.

Keywords— Data Mining, Association Rule Mining, FP Growth algorithm, Artificial Neural Networks.

I. INTRODUCTION

DM is a component of a wider process called KDD or knowledge discovery from databases. It involves researchers from a large assortment of approaches, including *c*omputer scientists and *s*tatisticians, as well as those working in fields such as *m*achine learning (ML), *a*rtificial intelligence (AI), *i*nformation retrieval and *p*attern recognition (PR). In this process, before a data set can be mined it first has to be cleaned. This method supplants fallacies, ensures resoluteness, and takes obscured values into account. It may use convert, edit, and evaluate simple statistical techniques or it may use highly sophisticated data analysis. What is new for data miners is the employment of these techniques on vast quantities of data. In general, it uses quite simple statistical techniques, though it may use highly sophisticated data analysis. This paper focuses on an *a*ssociation rule mining (ARM) algorithm named FP-growth, which is being juxtaposed with an optimized algorithm which is based on the ANN concepts. The comparisons are carried out between the different algorithms based on their execution time. Even, interestingness is obtained from the comparison of the FP Growth algorithm, and Optimized algorithm. Thus to test the impactfulness of the ANN approach on the DM algorithm on its execution time and the interestingness is observed and juxtaposed with the FP growth algorithm.
ANN is a loosely modelled system based on the human *b*rain. The field goes by many names, such as connectionism, parallel *d*istributed processing (PDP), neuro-*c*omputing, natural intelligent systems, ML algorithms, and ANN. It has the flair to account for any functional dependency. An optimized algorithm follows a UL algorithm in which an ANN is volitional in a way to perpetuate the task of data encoding along with data *d*ecoding to reconstitute input.

II. RELATED WORK

By concentrating on the foibles of the stereotyped "support-confidence" ARM framework and the worriment of mining negative ARs, the concept of *i*nterestingness measure is introduced. Junrui Yang and Lin Xu [1], analzsed the commonly used *i*nterestingness measures and merged the *i*nterestingness measure based on the discrepancy idea and the cosine measure to obtain a new interestingness measure model. By using this model in mining ARM-based rules, the positive and negative ARs can be effectively mined and segregated. Moreover, the algorithm is further verified by using a certain grocery store transaction dataset in comparison with the existing positive and negative ARM algorithm and the mining results using the stereotyped

"support-confidence" framework. The results verify the potencies of the improved interest model and the algorithm using the model, and contribute to the overall improvement of the mining accuracy of the algorithm.

DM is the process of extracting hidden and useful patterns and information from data. It is a technology that helps businesses to prognosticate future trends and deportments, allowing them to make proactive and knowledge-driven decisions. In this series, Anshu in [2] shows the procedure of DM and how it can help decision-makers to make better decisions. It also presents a detailed description of DM techniques and their applications in various fields. Some of the major techniques such as classification, clustering etc., helps in finding the patterns to decide upon the future trends in businesses to grow. Different techniques can be used for different purposes as per their pros and cons.

DM has been widely used in the business field, and ML can perform data analysis and pattern discovery, thus playing a key role in DM application. Teng Xiuyi and Gong Yuxia in [3] expound on the definition, model, development stage, types and commercial application of ML, and emphasizes the role of machine learning in DM. As per them, the ML technology in DM can be applied in many industries, including the financial industry, retail industry, insurance industry, telecommunication industry and so on. For example, in financial industry, the financial analysts use DM to build prediction models to intuit the patterns that have caused market fluidity in history, thereby improving the ability to predict market fluidity. Similarly, in the retail industry, salespeople can build predictive models through DM to understand who is most likely to respond to commonality, thereby increasing sales. When enterprises apply DM technologies, it is believed that enterprises fully cognize the advantages and disadvantages of various technologies and methods, and select appropriate technologies for specific environments and tasks.

Frequent pattern mining is an important topic that is needed for the Internet of Things (IoT) applications frequently. Many IoT applications have been refined in which continuous streaming data is used. In continuation with this, Halil İbrahim DEDE et al., in [4] performed a comparison study based on pursuance evaluation of the FP-stream algorithm and the WMFP-SW algorithms, which are developed for frequent pattern analysis in IoT, using a novel real-world dataset. The dataset was created by collecting 6 different sensor values over 3 months in an established testbed. As per the earlier studies, both the algorithms are efficient enough in extracting FPs but the experiments done shows that the WMFP-SW algorithm extracts more FPs the FP-stream algorithm when minimum support is increasing. Regarding the results of execution time, the FP stream is always faster than the WMFP-SW. Moreover, the number of common FPs and the total number of FPs per algorithm decrease as the minimum support value increases.

Generally, the power systems are disturbed by the presence of electromagnetic interference and crosstalk between the transmission link layers in the transmission and dissipating process. Though it is easy to produce transmission distribution faults, insincere graft to embellish the productiveness of fault diagnosis, a method of fault diagnosis for power systems based on ANN algorithms was proposed by Kai Xu in [5]. In this, the multi-sensor quantization fusion method is used to carry out the electricity. The transmission dissipating signal in the power transmission link layer is hauled out from the power system, and the transmission dissipating signal is decomposed and the ARs are excavated. Furthermore, the spectral analysis model is used to citrate the spectral characteristics of the transmission information of the power system, and the fault diagnosis and fault type identification are carried out according to the spectrum difference. The power system fault features are then further tiered and identified by an ANN algorithm to realize the optimal diagnosis of power system fault. The simulation results show that the method was more accurate and more streamlined in the fault diagnosis of the power system.

III. METHODOLOGIES

ARM is an event for determining organizations, patterns, and relationships between sets of objects on a dataset. The law of association is of the form M ⟶N [support, confidence]. The support and confidence are the primary measures that are used for assurance of the rule. AR is said to be strong if it satisfies both the minsupp (minimal support) and the minconf (minimal confidence) that is defined by the user. These ARs are easy to establish due to the small database but become more complex as the database transforms. Some general values and concepts are required to better understand DM in large data sets. A set contains n items and is called an n-itemset. So set {A, B} is a set of 2- itemset. Based on the frequency of itemsets, the number of active functions is calculated.

So here, a well-known mining algorithm i.e., FP growth is taken along with an Optimised algorithm using ANN is considered for the experiments of analysing the performance of mining ARs.

A. Artificial Neural Networks

ANNs are biologically inspired that is, they are made of neurons that perform in a manner that is analogous to the most elementary functions of the biological neurons. These elements are then organized in a way that may be related to the anatomy of the brain.

ANN was introduced as early as 1960. After a lapse of nearly two decades, interests in ANN have grown rapidly over the past few years. Professionals from diverse fields like engineering, physiology, physics and others are intrigued by the potential offered by this technology. Researchers are moving ahead on the theoretical as well as the applications front. Suddenly, this

technology has raised the possibility of realizing several attributes of human learning and intelligence. The models suggested can now capture some features of the brain such as high fault tolerance, dexterity to recognize objects under varied positions, high degree of robustness and ability to learn. [7][9]

The fields which are revived in the light of ANN, include:

1. Cognitive sciences, Vision.
2. Pattern recognition.
3. Learning paradigms.
4. Artificial intelligence.

Intensive research is going on in all the fields mentioned above and many models of ANN have been proposed. Though many of the algorithms presented do not present rigorous proofs of convergence, their utility has been demonstrated by the applications. Since the models have an internal distributed representation of knowledge in terms of connection strength it becomes difficult to decipher the knowledge in conventional norms. Models are either of self-organizing type or learn under supervision.

Generally, a neuron can realize the basic logic function and hence we can think of it as the switching device. Though the parallel architecture of ANN has an appeal for parallel processing, the hardware development in this field is very limited.

B. FP Growth Algorithm

FP-growth algorithm is an association analysis algorithm proposed by Han Jiawei and others in 2000. The algorithm uses a divide and conquers strategy: compress the database providing frequent itemsets into a FP-Tree, and keep the association information between itemsets. When compared with one of the most common mining algorithms Apriori, its advantages are as follows: first, it does not produce candidate itemsets; and second, it does not need to traverse data sets frequently. FP-growth algorithm generates FP-Tree data structure to find frequent itemsets, reduce the number of scanning transaction sets, and can efficiently mine data. [11]

The workflow of the FP-growth algorithm is as follows: First, it scans the transaction set for the first time and lists the number of times each item appears in the transaction set. Then according to the experience, the minimum support is determined. Then the process starts as items that do not meet the minimum support are deleted from each transaction of the transaction set, and the items meeting the minimum support are arranged in reverse order according to the number of times, to obtain a frequent item set. Secondly, the root node of FP-Trees are set to null.[13] Then the modified transaction set is scanned again, and a path is established from the root node to the leaf nodes as per the order of frequent itemsets. If the items in the transactions are already in the FP-Tree, then the corresponding number of items is uplift by 1, and different items are added after it, that is, the same items are shared and the coefficient is uplift by 1. Finally, after this the step 2 is repeated again to get the FP-Tree, header table and node linked list. According to the item header table, we can find the conditional pattern base of the FP-Tree, and then we can find the conditional FP-Tree. Finally, the above process is repeated to find the frequent itemsets of the corresponding sub conditional pattern base, as shown in Fig 1.

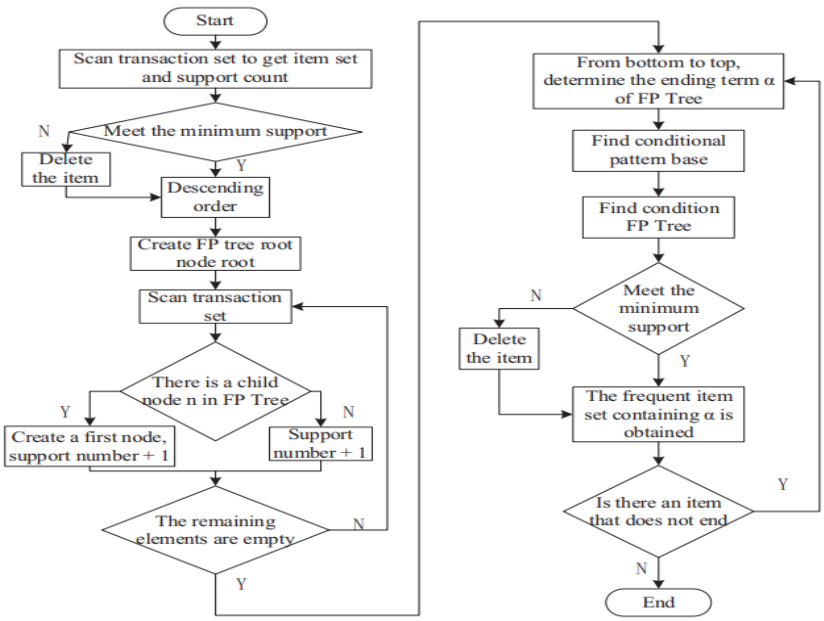

Fig 1: Establishment of FP-Tree

IV. EXPERIMENTAL SETUP

For this experiment, an Intel(R) Core(TM) i3-8130U CPU @2.20GHz with 8.00 GB RAM or above has been used. The operating system was Microsoft Windows 10 Enterprise. While comparing the performance of the two algorithms FPGrowth, and Optimised algorithm using ANN, no other application has been used implementation for feasible analysis of the implementation time of an algorithm.

The datasets used for the experiment are the shop transactions dataset, and the parameters of the selected datasets is having 120 unique datasets in a total of 520 transactions.

V. EXPERIMENTAL RESULTS

For analysing the performance of different algorithms this section is classified into two parts as implementation-time analysis.

A. Data Visualization

Shop transactional dataset:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
frankfurter	rolls/buns	soda	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
chicken	tropical fruit	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
butter	sugar	fruit/vegetable juice	newspapers	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
fruit/vegetable juice	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
packaged fruit/vegetables	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

B. Rules Generation

	Unnamed: 0	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	140	frozenset({'citrus fruit', 'tropical fruit', '...	frozenset({'other vegetables'})	0.005694	0.193493	0.004474	0.785714	4.060694	0.003372	3.763701
1	1675	frozenset({'curd', 'domestic eggs'})	frozenset({'whole milk'})	0.006507	0.255516	0.004779	0.734375	2.874086	0.003116	2.802763
2	1615	frozenset({'butter', 'curd'})	frozenset({'whole milk'})	0.006812	0.255516	0.004881	0.716418	2.803808	0.003140	2.625286
3	3510	frozenset({'yogurt', 'tropical fruit', 'whippe...	frozenset({'whole milk'})	0.006202	0.255516	0.004372	0.704918	2.758802	0.002787	2.522974
4	578	frozenset({'yogurt', 'tropical fruit', 'root v...	frozenset({'whole milk'})	0.008134	0.255516	0.005694	0.700000	2.739554	0.003616	2.481613
5	1071	frozenset({'yogurt', 'butter', 'other vegetabl...	frozenset({'whole milk'})	0.006406	0.255516	0.004372	0.682540	2.671221	0.002735	2.345125

C. Implementation time Analysis

In this section, different Implementation Time is observed for different support values of two algorithms and to observe their performance. The lower the Implementation time, means the higher the performance. Implementation time analyses of the datasets are shown in the following tables.

TABLE I
Implementation time (second) analysis of the algorithms for different Support values for both datasets.

Dataset	Min Support Value	FP Growth Algorithm	Optimized Algorithm
Shop Transactions Dataset	0.002	48.21	43.09
	0.004	46.63	40.2
	0.006	41.8	37.20
	0.008	35.73	31.13
	0.010	32.31	26.35

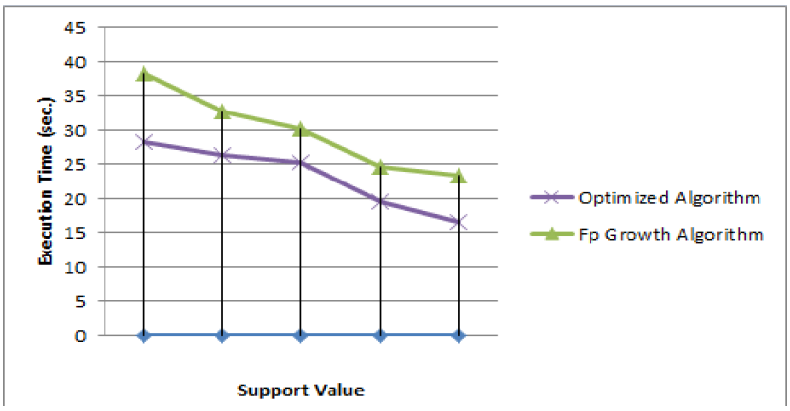

Fig. 2: Implementation time analysis for different Support values of Shop transactions dataset.

The observation from fig 2, shows that the implementation time of the FP-Growth algorithm takes more time compared to optimised algorithm. Here algorithms using the ANNs concepts perform better according to implementation time. In the next part, we are analysing the interestingness of rules generated by FP Growth and optimised algorithms.

D. Interestingness Analysis

Interestingness defines the acceptability of a rule.

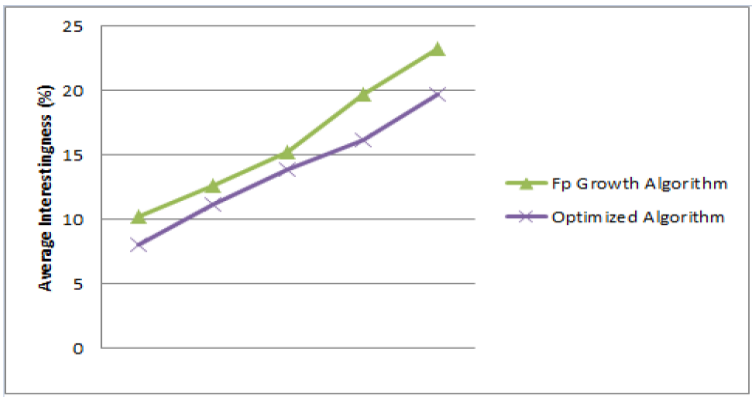

Fig. 3: Average Interestingness analysis of rules for different Support values for shop transaction dataset

Observation from Fig 3, categories the average interestingness of shop transaction dataset using FP growth and optimised algorithm. The average interestingness is higher for the FP Growth compare to optimised algorithm.

VI. CONCLUSION

The observation in this paper focuses on the implementation time of the FP-Growth algorithm, and the optimised algorithm which is made to explore the application of ANNs in DM. The observation shows that the optimised algorithm takes lesser implementation time as compared to FP Growth. The average Interestingness value obtained by FP growth is more than the optimised algorithm, this result shows that the optimised algorithm generates a lesser number of association rules, which means optimised algorithm generates only the potentially strong association rules. The results show that the optimised algorithm has better in terms of implementation time, efficiency and interestingness then FP-Growth algorithms.

REFERENCES

[1] S Junrui Yang and Lin Xu, "A novel interestingness measure based on a fusion model for association rules mining", MATEC Web of Conferences, pp. 1-6, 2021.

[2] Anshu, "Review Paper on Data Mining Techniques and Applications", International Journal of Innovative Research in Computer Science & Technology (IJIRCST), vol. 7, issue 2, pp. 22-26, March 2019.

[3] Teng Xiuyi and Gong Yuxia, "Research on Application of Machine Learning in Data Mining", IOP Conference Series: Materials Science and Engineering, pp. 1-5, 2018.

[4] Kai Xu, "Fault Diagnosis Method of Power System Based on Neural Network", International Conference on Virtual Reality and Intelligent Systems, pp. 172-175, 2018.

[5] Mohammadreza Ramzanpour and Simone A. Ludwig, "Association Rule Mining Based Algorithm for Recovery of Silent Data Corruption in Convolutional Neural Network Data Storage", IEEE Symposium Series on Computational Intelligence (SSCI), pp. 3057-3064, 2020.
[6] Sempe Leholo et al., "Solar Energy Potential Forecasting and Optimization using Artificial Neural Network: South Africa Case Study", Institute of Electrical and Electronics Engineers, pp. 533-536, 2019.
[7] Manomita Chakraborty et al., "Data Mining Using Neural Networks in the form of Classification Rules: A Review", 4th International Conference on Computational Intelligence and Networks (CINE), pp. 332-337, 2020.
[8] Biswajit Panja, et al., "Crime Analysis Mapping, Intrusion Detection - using Data Mining", IEEE Technology and Engineering Management Conference (TEMSCON), pp. 255-259, 2020.
[9] Tianyu Xia et al., "Data Association Rules Mining Method Based on RBF Neural Network Optimization Algorithm", International Conference on Communications, Information System and Computer Engineering (CISCE), pp. 433-436, 2020.
[10] X. Y. Kong et al., “Fault diagnosis method of a distribution network based on cloud-edge computing architecture and wavelet neural network”, Distribution & Utilization, vol.37, pp. 17-23, 2020.
[11] Y. H. Huang et al., "Research on weak point analysis of distribution networks based on FP-Growth algorithm", Electrical Measurement & Instrumentation, pp. 1-7, 2020.
[12] Hossein Riahi-Madvar et al., "Derivation of optimized equations for estimation of dispersion coefficient in natural streams using hybridized ANN with PSO and CSO algorithms", IEEE Access, vol.8, pp.156582-156599, 2020.
[13] Jin Bo Chen et al., "A Log Analysis Technology Based on FP-growth Improved Algorithm", International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA), pp.1096-1101, 2021.
[14] Wisnu Arya Dipa and Wikan Danar Sunindyo, "Software Defect Prediction Using SMOTE and Artificial Neural Network", International Conference on Data and Software Engineering (ICoDSE), pp. 1049-1054, 2021.