**Abstract:**

The aim of our work is to study different types of data mining algorithms by applying them on a proposed dataset to compare and analyze their performances. We’ve shown different data mining techniques and designed a basic data warehouse to simplify the dataset. Then we’ve applied the algorithms on the warehouse data and discussed different application issues. We’ve concluded our work by visualizing the comparative performance of the algorithms and made a discussion on further works in this field.

**Introduction:**

From the very beginning of human evolution, people started to observe and learn the environment. Human brain saved all the observations into its memory system. With the passage of time they adapted themselves to the environment based on their previous observations. Each distinct observation can be analogized as data. From then to now the increasing exponentiality of data generation has been continued. After the industrial revolution, all the industrial operations started to generate data at high scales. These huge amounts of unstructured data became valuable to the authorities. From then data management become a part of study. With the help of increased computational ability, different complex data management and maintenance techniques were invented and became popular to the computer science community. During 90s internet became popular and internet based applications & users started to generate huge data than before. The importance of modern data management and analysis have become more essential than before. Modern business processes make billions of transactions per seconds and create enormous amount of data which are important to understand customer mode, business growth, sustainability etc. To deal with such data and perform efficient analysis, a new branch of computation is created which is called data mining.

**Introduction to Our Work:**

We’ve aimed to study on different data mining algorithms and apply them to understand their performance. For this we’ve selected a dataset from 1[abcd]. The dataset is about global superstore transactions over the period of 2011 to 2015. The dataset contains 26 attributes and more than fifty-one thousands of record. We plan to predict different measures based on the interrelated data and perform the actions via different algorithms. For this we’ve selected **python** as programming language and **Scikit Learn,** a python based data manipulation library and different editor and IDEs.

**Problem Definition:**

The first and foremost step of our work is the problem definition. There are some interesting quotes regarding this step which will give us a clear view why this step is important.  
***"There are no right answers to wrong questions."* - Ursula K. Le Guin.** So, at first, we need to state what we really want to do. But that is not enough. More precisely, we must ask the right question with proper details. Then we can find a solution.

In our work, initially we’ve to analyze the dataset and create a structured format of it. To perform this, we’ve to design a data warehouse to store organized data. then we can to apply mining algorithm on the data warehouse. But there are different variance of each algorithm and performance of a particular algorithms varies in accordance to its application. So we will test the performance of the algorithms by finding the accuracy rate on different applications and then show a comparative performance chart.

For example, we want to predict the order priority of a product from our data set. First, we must ensure that the dataset has enough information by which we can predict the order priority. To do that, we need to find the factors which are important to our problem. Then we need to define a goal. By goal here we mean we must obtain more that 50% accuracy. If we have 50% accuracy that means the probability that our prediction is correct is 50 % which is not seems good. Half of our prediction will be wrong. So our goal will be to gain more than 70 % accuracy. So now we have the complete definition of our problem. Now we can say “We want to predict a product’s Order priority using our dataset at more than 70% accuracy”

**Process of Solution:**

We’ve segmented our solutions into different processes which are given below:

1. Data Collection
2. Data pre-processing
3. Algorithm Selection
4. Important Feature selection for each algorithm
5. Model Creation
6. Performance Testing
7. Comparison and visualization of algorithms.

**Data Collection:**

First and foremost step of our solution is to collect the data to perform all the operation. For this, we’ve collected a dataset. An overview of data collection is given on the introduction section of this report.

**Data Pre-processing:**

When the data collection is completed the second step is to process the dataset to perform further operations. We’ve designed a data basic data warehouse to process the dataset. Where the dataset is normalized and segmented into different dimension tables and a fact table. Details description of this step will be discussed on the data warehousing section of this report.

**Algorithm Selection:**

Most important process of our work is to select appropriate algorithm to perform operations on the dataset. We’ve selected different algorithms for different operations.

**Important Feature selection:**

For a specific algorithm we’ve to select relevant features to perform the operation. There are many approaches for selecting best features. Some of the techniques are discussed later in this report.

**Model Creation:**

When the features/ parameters are ready, the next step is to make the mathematical model for the algorithm. In this step, the mathematical expression of the algorithms are developed.

**Performance Testing:**

After the mathematical model is created, we’ve test the model whether it works or not and provides desired solution.

**Comparison and visualization of algorithms:**

Finally, we’ve to compare the results of the algorithms for a specific operation and determine which algorithm works better for that specific operation and visualize the comparative diagram.

**Our Goal:**

The goal of our work is simple. We will study on few data mining algorithms and understand their scopes and limitations. For this purpose, we’ve designed data warehouse and performed other operations which will be discussed later in this report.

**Outline of the Report: aikhane likhbi kon chapeter a ki discus kora hobe.. full report lekha ses a likhbi.**

**Literature Review:**

To understand and gain knowledge on any topic, literature is the best reliable source to start with. It gives an in-depth overview on any topic. There are many researchers who have worked on data mining before and their works are recognized globally. To start and continue our work we’ve to study few books and references of such researchers.

**Historical background of Data Mining:**

### According to ***Data Mining: Concepts and Techniques, 3rd ed by* Jiawei Han, Micheline Kamber and Jian Pei[1]** Since 1960, systematic database and information technology has been evolved from the primitive file processing system to sophisticated database system. In 1970 the database system progressed from hierarchical and network database system to relational database. From then users gained convenient flexible data access through query languages and user interface. After the establishment of database management systems, database technology moved toward the development of advanced database systems, data warehousing, and data mining for advanced data analysis and web-based databases during mid-1980. Advance data analysis started from 1980. Data warehousing techniques were introduced and from 1990, when internet started, huge volumes of data have been accumulated beyond databases and data warehouses and the trend of data mining started to popular to the community.

**Data Mining:**

According to ‘***Data Mining with Predictive Analytics for Financial Applications’ by S.B. Soumya1, N. Deepika*** *[2] ‘Data mining, also popularly referred to as knowledge discovery from data (KDD), is the extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams.’*

From other Two references we can define data mining as: ‘**Data mining** is the computing process of discovering patterns in large data set involving methods at the intersection of machine learning, statistics and database system [3]. It is an essential process where intelligent methods are applied to extract data patterns. [1][3]

**Data mining involves six common classes of tasks: [4]**

**Anomaly Detection:**

 (outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.

**Association Rule Learning:** (dependency modelling) –

Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

**Clustering:**

It is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

**Classification:**

Itis the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

**Regression:**

It attempts to find a function which models the data with the least error that is, for estimating the relationships among data or datasets.

**Summarization:**

providing a more compact representation of the data set, including visualization and report generation.

**Data Warehouse:**

**Tor ager report theke aikhane lekha suru kor. Sob lekhar reference hobe han n kamber.**

**References**

1. Han, Kamber, Pei, Jaiwei, Micheline, Jian (June 9, 2011). [*Data Mining: Concepts and Techniques*](http://proquest.safaribooksonline.com.proxy.library.carleton.ca/book/databases/data-warehouses/9780123814791) (3rd ed.). Morgan Kaufmann. [*ISBN*](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [*978-0-12-381479-1*](https://en.wikipedia.org/wiki/Special:BookSources/978-0-12-381479-1).
2. Deg
3. [*"Data Mining Curriculum"*](http://www.kdd.org/curriculum/index.html). [*ACM*](https://en.wikipedia.org/wiki/Association_for_Computing_Machinery) [*SIGKDD*](https://en.wikipedia.org/wiki/SIGKDD). 2006-04-30*. Retrieved 2014-01-27*.
4. [*Fayyad, Usama*](https://en.wikipedia.org/wiki/Usama_Fayyad); *[Piatetsky-Shapiro, Gregory](https://en.wikipedia.org/wiki/Gregory_Piatetsky-Shapiro" \o "Gregory Piatetsky-Shapiro)*; Smyth, Padhraic (1996). [*"From Data Mining to Knowledge Discovery in Databases"*](http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf) *.*