CONTENTS

LIST OF FIGURES

ABSTRACT

Now-a-days customers prefer online shopping rather than offline shopping. The crucial challenge of online shopping is to analyze the behavior of the customers. Many of the people are visiting the online shopping sites and spending their time by surfing either to buy or for window shopping. Customers behavior varies from person to person based on their buying behavior patterns. Our main aim is to analyze the customer behavior like who buys what. The result of this analysis is suggesting some techniques for improving the sales. The success of business is to know the requirement of the customer and providing the good offers in right time. Data mining is used to extract the important information from the bulk of data to save it and summarize it in effective manner. Different approaches for customer behavior analysis in data mining are: Classification. We use one of these appropriate Data Mining techniques for behavioral analysis of customers, therefore to optimize the business outcome in online shopping.

INTRODUCTION

1.1 INTRODUCTION

Online shopping is the easy solution for busy life in today's world. Earlier, consumers buy the goods and products by physically at the stores. But now-a-days, the people want to save the time for their professional or personal sake, so that they are willing to go for online shopping. This online shopping saves crucial time for modern people. In the 21st century, trade and commerce have been so diversified that multichannel has taken place and online shopping has increased significantly throughout the world. Unlike physical store, all the goods in the online stores were described through text, with photos, with multimedia files. The online stores also provide the links for much detailed information of the product. The online consumers are adventurous explorer, shopping lover and some are technology muddler, hate waiting for the product to ship. The online consumer behavior like the action during searching, buying and using products became a contemporary research area for an increasing number of researchers to understand this unique nature of online shopping. Thus the purpose of this study is to understand the consumer behavior towards online shopping, their liking, disliking and satisfaction level. This system uses classification or association techniques of data mining to analyze the behavior of customer.

Data analytics- data analytics is the process of evaluating digital information and converting it into information useful for business.

Data warehousing- is the component of the foundational importance of most huge-scale data mining efforts with a large collection of data, that is used for decision making in organizations.

Machine learning- is a computer programmed technique, that makes use of statistical probabilities that gives the computer the capacity to 'learn' even without being clearly programmed.

Regression- is a technique that is made use of to predict a variety of numeric values, including sales, price of a stock, temperatures, that are based on a precise dataset.

1.1.3 WORKING OF DATA MINING

Exploring and analyzing large quantities of information to derive relevant patterns and trends is involved in data mining. Data mining has many uses such as credit risk management, database marketing, spam email filtering, fraud detection, and also to fathom the opinion and sentiment of users. The data mining process is further divided into five steps. First data is collected and loaded into the data warehouse.

Then the data is managed and stored either in the cloud or in the in-house servers. The data is assessed by management teams, business analysts, and information technology professionals and they determine how to organize the data. Then based on the results of the users, it is sorted by the application software. Finally, the data is presented by the end-user in a format like a graph or a table that is easy to share.

The first step in data mining is almost always data collection. Today's organizations can collect records, logs, website visitors' data, application data, sales data, and more every day. Collecting and mapping data is a good first step in understanding the limits of what can be done with and asked of the data in question.

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is an excellent guideline for starting the data mining process. This standard was created decades ago and is still a popular paradigm for organizations that are just starting.

The 6 CRISP-DM phases

The CRISP-DM comprises a six-phase workflow. It was designed to be flexible; data teams are allowed and encouraged to move back to a previous stage if needed. The model also provides opportunities for software platforms that help perform or augment some of these tasks.

1. Business understanding

Comprehensive data mining projects start by first identifying project objectives and scope. The business stakeholders will ask a question or state a problem that data mining can answer or solve.

Why we have choosen data mining for our project

Data mining is mainly used for the purpose of finding patterns and correlations with large datasets to analyze outcomes.

In our project we have used this data mining because it is totally based on analysis and prediction of behavior of a customer

Data mining explores and analyzer large blocks of information by using some techniques to present the output data in the form of graph or a table.

1.1.7.TECHNIQUES USED IN EARLIER PROJECTS

-APRIORI ALGORITHM:

Apriori algorithm is for finding frequent itemsets in a dataset for boolean association rule. Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties. We apply an iterative approach or level-wise search where k-frequent itemsets are used to find k+1 itemsets.

To improve the efficiency of level-wise generation of frequent itemsets, an important property is used called *Apriori property* which helps by reducing the search space.

Apriori Property –

All non-empty subset of frequent itemset must be frequent. The key concept of Apriori algorithm is its anti-monotonicity of support measure. Apriori assumes that

All subsets of a frequent itemset must be frequent(Apriori propertry).
If an itemset is infrequent, all its supersets will be infrequent.

P(I) < minimum support threshold, then I is not frequent.

P (I+A) < minimum support threshold, then I+A is not frequent, where A also belongs to itemset.

If an itemset set has value less than minimum support then all of its supersets will also fall below min support, and thus can be ignored. This property is called the Antimonotone property.

APRIORI algorithm uses association rule mining which needs to calculate confidence of each rule.

Confidence –
A confidence of 60% means that 60% of the customers, who purchased milk and bread also bought butter.

Confidence(A->B)=Support_count(A∪B)/Support_count(A)

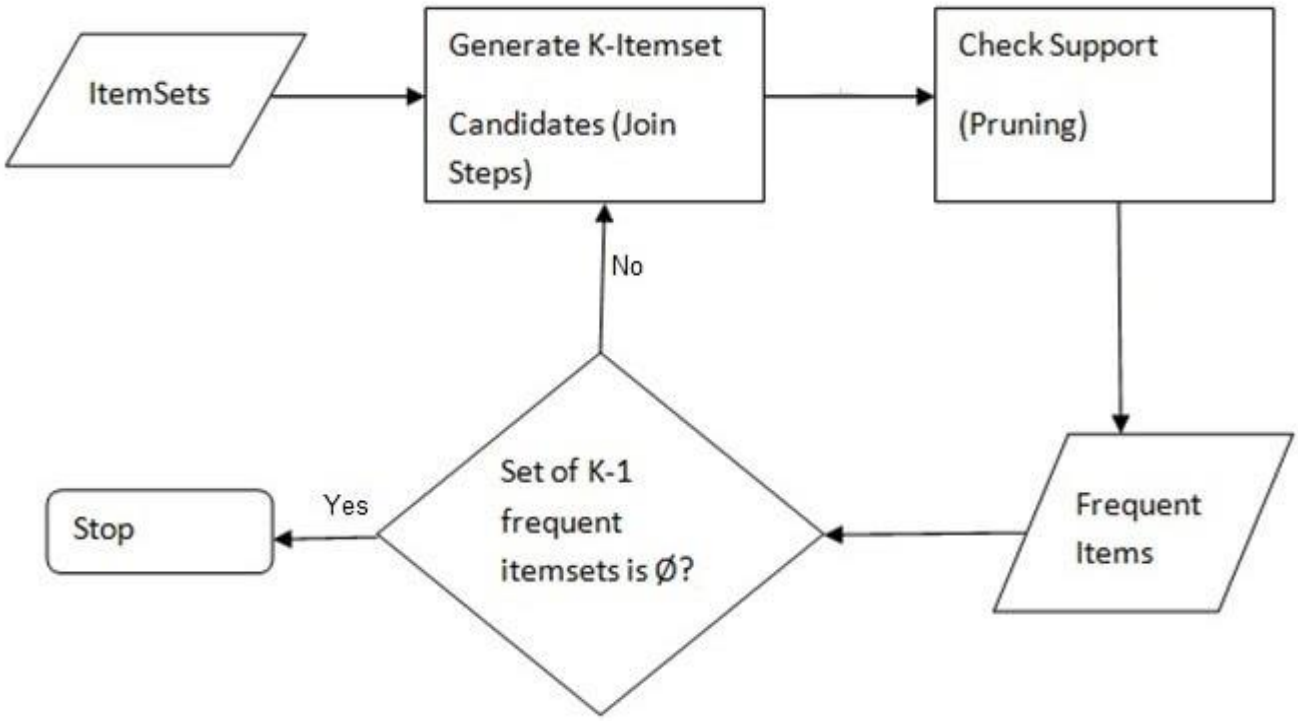


FIG: APRIORI ALGORITHM IN DATA MINING

-ASSOCIATION RULE MINING:

Association (or relation) is probably the better known and most familiar and straightforward data mining technique. Here, you make a simple correlation between two or more items, often of the same type to identify patterns. For example, when tracking people's buying habits, you might identify that a customer always buys cream when they buy strawberries, and therefore suggest that the next time that they buy strawberries they might also want to buy cream.

BASIC DEFINITIONS –

1.Support Count( ) – Frequency of occurrence of a itemset.

2.Frequent Itemset – An itemset whose support is greater than or equal to minsup threshold.

3.Association Rule – An implication expression of the form X -> Y, where X and Y are any 2 itemsets.

SKLEARN

Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

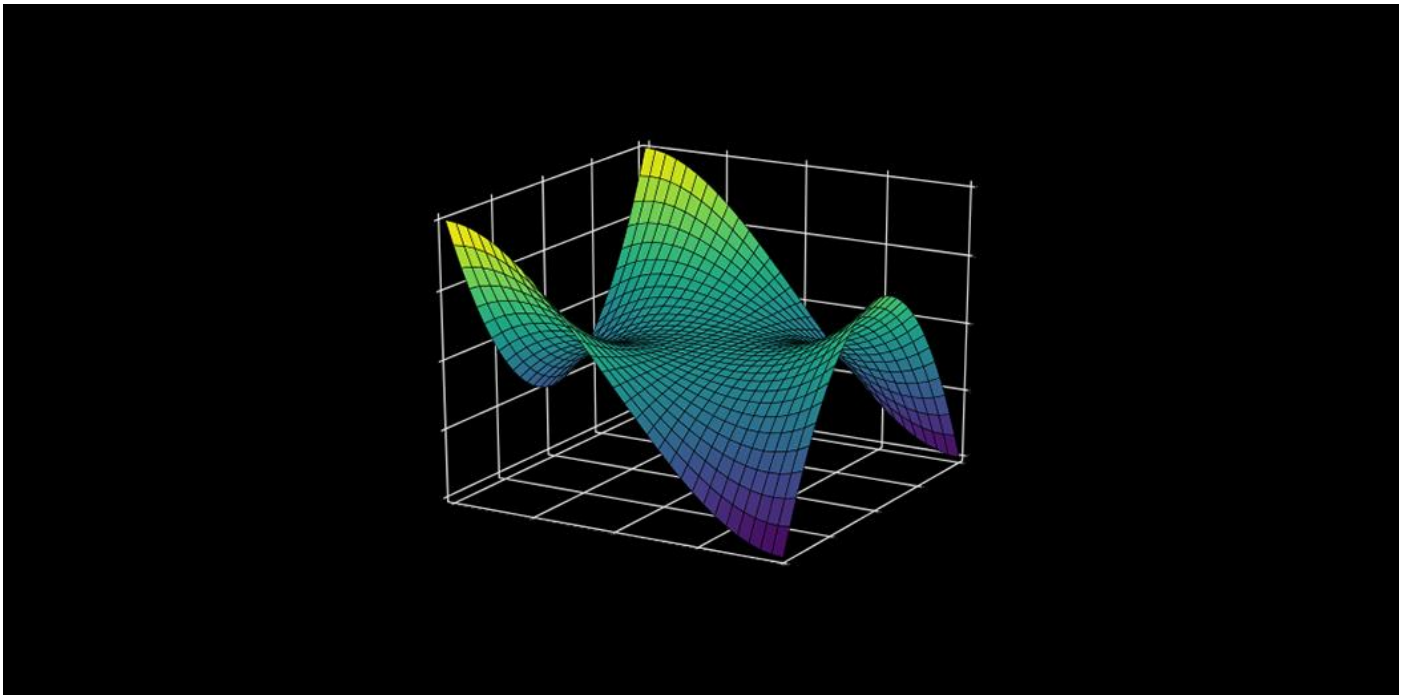
Components of scikit-learn:

- **Supervised learning algorithms**
- **Cross-validation**
- **Unsupervised learning algorithms**
- **Various toy datasets**
- **Feature extraction**

MATPLOTLIB

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. Matplotlib is one of the most popular Python packages used for data visualization. It is a cross-platform library for making 2D plots from data in arrays. It provides an object-oriented API that helps in embedding plots in applications using Python GUI toolkits such as PyQt, WxPythonotTkinter. It can be used in Python and IPython shells, Jupyter notebook and web application servers also.

matplotlib.**pyplot** is a collection of functions that make **matplotlib** work like MATLAB. Each **pyplot** function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels.



1.2 PROBLEM STATEMENT

The main aim of the proposed system is to find the maximum sold product category based on the customer purchasing behaviour using datamining algorithm. So we can find the maximum sold product category by using the graphs and its accuracy and we also analyse its future sales.

	conducted on 7-Eleven convenience store. The factors considered in this regard are cultural, social, personal and psychological factor. After applying multiple regression analysis and hypothesis testing the coefficient of determination (R2) is derived that describes the influence of all the factors that affects consumer inclination to buy the products. All these factors are also independently discussed and analyzed statistically .The online reviews help the customers establishing an opinion regarding online shopping. These reviews vary in both quality and quantity. They can have both positive and negative kind of effect on customers and as well as on business. The data in this regard is collected via questionnaire and the results are compiled after various complicated statistical methods. These results interpret that reviews do contribute to decision making in online shopping .There have been various techniques of data mining for the identification of frequent item sets. As the data retrieved after processing is very large and requires some efficient technique to discover some useful pattern. The paper discusses those techniques that can aid in the formation of any such pattern. Association rule has been considered as one of the basic data mining tools. There exist many algorithms like Apriori algorithm, AIS, SETM, Apriorihybrid, FP-growth for pattern discovery. Apart from the pros and cons of these algorithms, any of these can be used along with association rule mining for data analysis .